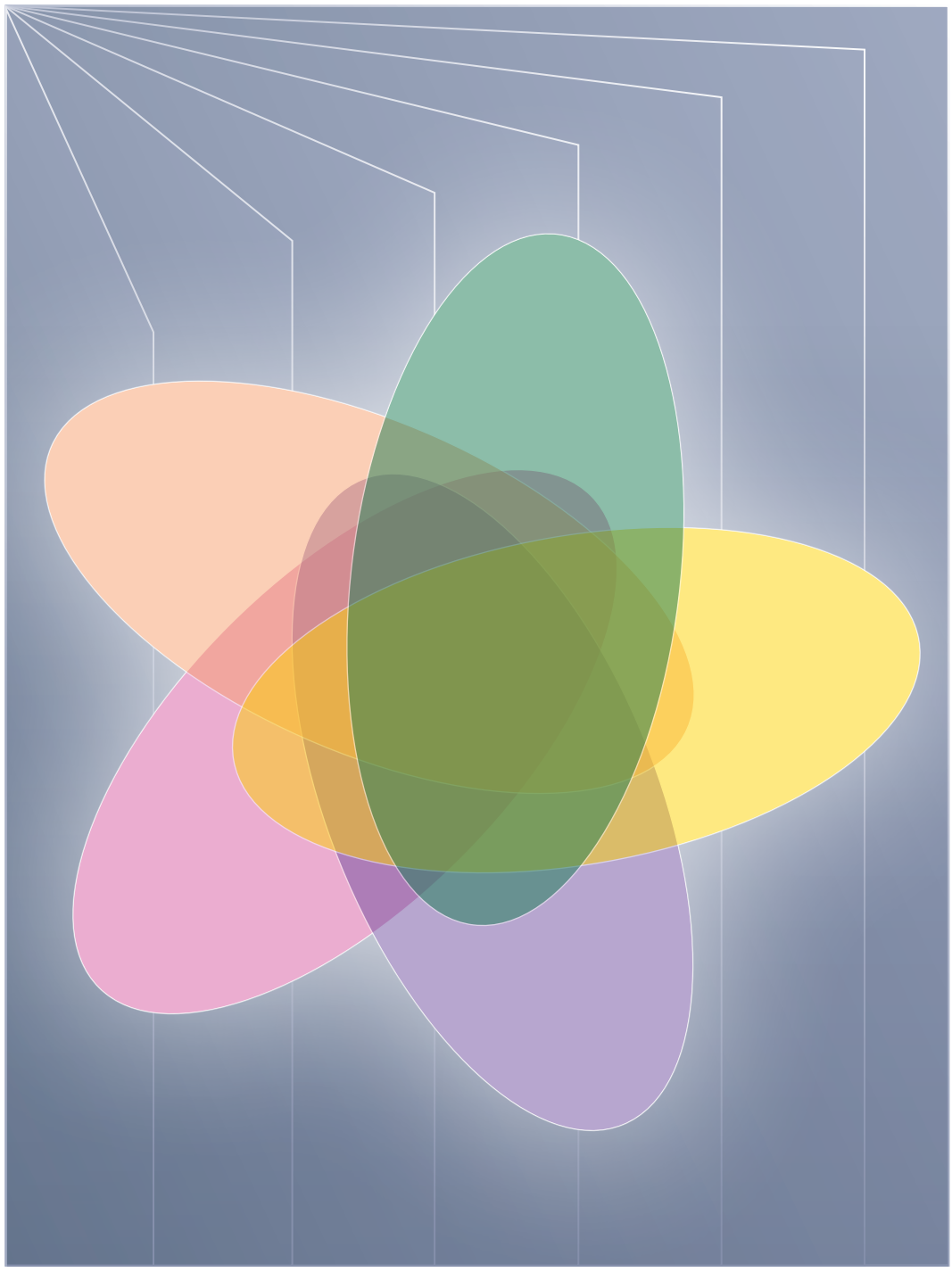
Online ISSN 2500-3259

ВАВИЛОВСКИЙ
ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ
VAVILOV JOURNAL OF GENETICS AND BREEDING

2025 • 29 • 7



Компьютерная геномика / Системная компьютерная биология / Структурная компьютерная биология / Экологическая и популяционная генетика / Эволюционная биоинформатика / Медицинская биоинформатика / Инструментальные средства биоинформатики

vavilov.elpub.ru
vavilovj-icg.ru
vavilov_journal@bionet.nsc.ru

2025 • 29 • 7

Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

VAVILOV JOURNAL OF GENETICS AND BREEDING

Основан в 1997 г.
Периодичность 8 выпусков в год
doi 10.18699/vjgb-25-120

Учредители

Сибирское отделение Российской академии наук

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Межрегиональная общественная организация Вавиловское общество генетиков и селекционеров

Главный редактор

А.В. Кочетов – академик РАН, д-р биол. наук, профессор РАН (Россия)

Заместители главного редактора

Н.А. Колчанов – академик РАН, д-р биол. наук, профессор (Россия)
И.Н. Леонова – д-р биол. наук (Россия)
Н.Б. Рубцов – д-р биол. наук, профессор (Россия)
В.К. Шумный – академик РАН, д-р биол. наук, профессор (Россия)

Ответственный секретарь

Г.В. Орлова – канд. биол. наук (Россия)

Редакционная коллегия

Е.Е. Андронов – канд. биол. наук (Россия)
Ю.С. Аульченко – д-р биол. наук (Нидерланды)
О.С. Афанасенко – академик РАН, д-р биол. наук (Россия)
Д.А. Афонников – д-р биол. наук, доцент (Россия)
Л.И. Афтанас – академик РАН, д-р мед. наук (Россия)
Л.А. Беспалова – академик РАН, д-р с.-х. наук (Россия)
А. Бёрнер – д-р наук (Германия)
Н.П. Бондарь – канд. биол. наук (Россия)
С.А. Боринская – д-р биол. наук (Россия)
П.М. Бородин – д-р биол. наук, проф. (Россия)
А.В. Васильев – чл.-кор. РАН, д-р биол. наук (Россия)
М.И. Воевода – академик РАН, д-р мед. наук (Россия)
Т.А. Гавриленко – д-р биол. наук (Россия)
Н.Е. Грунтенко – д-р биол. наук (Россия)
С.А. Демаков – д-р биол. наук (Россия)
И.К. Захаров – д-р биол. наук, проф. (Россия)
И.А. Захаров-Гезехус – чл.-кор. РАН, д-р биол. наук (Россия)
С.Г. Инге-Вечтомов – академик РАН, д-р биол. наук (Россия)
А.В. Кильчевский – чл.-кор. НАНБ, д-р биол. наук (Беларусь)
А.М. Кудрявцев – чл.-кор. РАН, д-р биол. наук (Россия)
Д.М. Ларкин – канд. биол. наук (Великобритания)
С.А. Лашин – д-р биол. наук (Россия)
Ж. Ле Гуи – д-р наук (Франция)
И.Н. Лебедев – чл.-кор. РАН, д-р биол. наук, проф. (Россия)
Л.А. Лутова – д-р биол. наук, проф. (Россия)
Б. Люгтенберг – д-р наук, проф. (Нидерланды)
В.Ю. Макеев – чл.-кор. РАН, д-р физ.-мат. наук (Россия)
И.В. Максимов – д-р биол. наук (Россия)
Б.А. Малярчук – д-р биол. наук (Россия)
Ю.Г. Матушкин – канд. биол. наук (Россия)
В.И. Молодин – академик РАН, д-р ист. наук (Россия)
М.П. Мошкин – д-р биол. наук, проф. (Россия)
Л.Ю. Новикова – д-р с.-х. наук (Россия)
Е.К. Потокина – д-р биол. наук (Россия)
В.П. Пузырев – академик РАН, д-р мед. наук (Россия)
Д.В. Пышный – чл.-кор. РАН, д-р хим. наук (Россия)
Е.Ю. Рыкова – д-р биол. наук (Россия)
Е.А. Салина – чл.-кор. РАН, д-р биол. наук, проф. (Россия)
В.А. Степанов – академик РАН, д-р биол. наук (Россия)
И.А. Тихонович – академик РАН, д-р биол. наук (Россия)
Е.К. Хлесткина – чл.-кор. РАН, д-р биол. наук, проф. РАН (Россия)
Э.К. Хуснутдинова – д-р биол. наук, проф. (Россия)
М. Чен – д-р биол. наук (Китайская Народная Республика)
Е.В. Шахтшнейдер – д-р мед. наук (Россия)
С.В. Шестаков – академик РАН, д-р биол. наук (Россия)
С.В. Шеховцов – д-р биол. наук (Россия)
Н.С. Юдин – канд. биол. наук (Россия)
Н.К. Янковский – академик РАН, д-р биол. наук (Россия)

2025 • 29 • 7

Online ISSN 2500-3259

Online edition

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

VAVILOV JOURNAL OF GENETICS AND BREEDING

Founded in 1997
Publication frequency: 8 issues a year
doi 10.18699/vjgb-25-120

Founders

Siberian Branch of the Russian Academy of Sciences

Federal Research Center Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences

The Vavilov Society of Geneticists and Breeders

Editor-in-Chief

A.V. Kochetov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Professor of the RAS, Russia

Deputy Editor-in-Chief

N.A. Kolchanov, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia
I.N. Leonova, Dr. Sci. (Biology), Russia
N.B. Rubtsov, Professor, Dr. Sci. (Biology), Russia
V.K. Shumny, Full Member of the Russian Academy of Sciences, Dr. Sci. (Biology), Russia

Executive Secretary

G.V. Orlova, Cand. Sci. (Biology), Russia

Editorial board

O.S. Afanasenko, Full Member of the RAS, Dr. Sci. (Biology), Russia
D.A. Afonnikov, Associate Professor, Dr. Sci. (Biology), Russia
L.I. Aftanas, Full Member of the RAS, Dr. Sci. (Medicine), Russia
E.E. Andronov, Cand. Sci. (Biology), Russia
Yu.S. Aulchenko, Dr. Sci. (Biology), The Netherlands
L.A. Bespalova, Full Member of the RAS, Dr. Sci. (Agricul.), Russia
N.P. Bondar, Cand. Sci. (Biology), Russia
S.A. Borinskaya, Dr. Sci. (Biology), Russia
P.M. Borodin, Professor, Dr. Sci. (Biology), Russia
A. Börner, Dr. Sci., Germany
M. Chen, Dr. Sci. (Biology), People's Republic of China
S.A. Demakov, Dr. Sci. (Biology), Russia
T.A. Gavrilenko, Dr. Sci. (Biology), Russia
N.E. Gruntenko, Dr. Sci. (Biology), Russia
S.G. Inge-Vechtomov, Full Member of the RAS, Dr. Sci. (Biology), Russia
E.K. Khlestkina, Corr. Member of the RAS, Professor of the RAS, Dr. Sci. (Biology), Russia
E.K. Khusnutdinova, Professor, Dr. Sci. (Biology), Russia
A.V. Kilchevsky, Corr. Member of the NAS of Belarus, Dr. Sci. (Biology), Belarus
A.M. Kudryavtsev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
D.M. Larkin, Cand. Sci. (Biology), Great Britain
S.A. Lashin, Dr. Sci. (Biology), Russia
J. Le Gouis, Dr. Sci., France
I.N. Lebedev, Corr. Member of the RAS, Professor, Dr. Sci. (Biology), Russia
B. Lugtenberg, Professor, Dr. Sci., Netherlands
L.A. Lutova, Professor, Dr. Sci. (Biology), Russia
V.Yu. Makeev, Corr. Member of the RAS, Dr. Sci. (Physics and Mathem.), Russia
I.V. Maksimov, Dr. Sci. (Biology), Russia
B.A. Malyarchuk, Dr. Sci. (Biology), Russia
Yu.G. Matushkin, Cand. Sci. (Biology), Russia
V.I. Molodin, Full Member of the RAS, Dr. Sci. (History), Russia
M.P. Moshkin, Professor, Dr. Sci. (Biology), Russia
L.Yu. Novikova, Dr. Sci. (Agricul.), Russia
E.K. Potokina, Dr. Sci. (Biology), Russia
V.P. Puzyrev, Full Member of the RAS, Dr. Sci. (Medicine), Russia
D.V. Pyshnyi, Corr. Member of the RAS, Dr. Sci. (Chemistry), Russia
E.Y. Rykova, Dr. Sci. (Biology), Russia
E.A. Salina, Corr. Member of the RAS, Professor, Dr. Sci. (Biology), Russia
E.V. Shakhtshneider, Dr. Sci. (Medicine), Russia
S.V. Shekhovtsov, Dr. Sci. (Biology), Russia
S.V. Shestakov, Full Member of the RAS, Dr. Sci. (Biology), Russia
V.A. Stepanov, Full Member of the RAS, Dr. Sci. (Biology), Russia
I.A. Tikhonovich, Full Member of the RAS, Dr. Sci. (Biology), Russia
A.V. Vasiliev, Corr. Member of the RAS, Dr. Sci. (Biology), Russia
M.I. Voevoda, Full Member of the RAS, Dr. Sci. (Medicine), Russia
N.K. Yankovsky, Full Member of the RAS, Dr. Sci. (Biology), Russia
N.S. Yudin, Cand. Sci. (Biology), Russia
I.K. Zakharov, Professor, Dr. Sci. (Biology), Russia
I.A. Zakharov-Gezekhus, Corr. Member of the RAS, Dr. Sci. (Biology), Russia

СОДЕРЖАНИЕ • 2025 • 29 • 7

911 ОТ РЕДАКТОРА
Н.А. Колчанов, Т.А. Бухарина

Компьютерная геномика

913 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Отражение процессов повреждения ДНК в эволюции G-трактов в геномах.
И.Р. Грин, Д.О. Жарков

925 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Связь иерархической классификации транскрипционных факторов по структуре ДНК-связывающего домена и вариабельности мотивов сайтов связывания этих факторов.
В.Г. Левицкий, Т.Ю. Ватолина, В.В. Радица

940 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Программа PlantReg 1.1: анализ взаимного расположения сайтов связывания транскрипционных факторов в промоторах генов-мишеней для уточнения молекулярных механизмов их активности в регуляторных сетях.
В.В. Лавреха, Н.А. Омельянчук, А.Г. Богомолов, Ю.А. Рябов, П.К. Мукебенова, Е.В. Землянская

952 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
База знаний FlyDEGdb по дифференциально экспрессирующимся генам *Drosophila melanogaster* – модельного объекта биомедицины.
О.А. Подколодная, М.А. Дерюженко, Н.Н. Твердохлеб, К.А. Золотарева, Ю.В. Маковка, Н.Л. Подколодный, В.В. Суслов, И.В. Чадаева, Л.А. Федосеева, А.А. Серяпина, Д.Ю. Ощепков, А.Г. Богомолов, Е.Ю. Кондратюк, О.Е. Редина, А.Л. Маркель, Н.Е. Грунтенко, М.П. Пономаренко

Системная компьютерная биология

963 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Реконструкция и анализ генной сети регуляции апоптоза при гепатоцеллюлярной карциноме на основе данных scRNA-seq и базы знаний ANDSystem.
А.В. Адамовская, И.В. Яцык, М.А. Клещев, П.С. Деменков, Т.В. Иванисенко, В.А. Иванисенко

978 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Сигнальный путь Hedgehog у человека: описание в базе знаний HH_Signal_pathway_db.
Т.А. Бухарина, А.М. Бондаренко, Д.П. Фурман

990 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Выявление белков, регулирующих фенотип-ассоциированные гены макрофагов группы М2: биоинформатический анализ.
Е.А. Антропова, И.В. Яцык, П.С. Деменков, Т.В. Иванисенко, В.А. Иванисенко

1000 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Компьютерная реконструкция генной сети цитокиновой регуляции генов и белков, ассоциированных с РАС.
Н.М. Леванова, Е.Г. Вергунов, А.Н. Савостьянов, И.В. Яцык, В.А. Иванисенко

1009 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Генная сеть и база знаний по терморегуляции организма человека.
Е.В. Игнатьева, П.С. Деменков, А.Г. Богомолов, Р.А. Иванов, С.А. Лашин, А.Д. Михайлова, А.Е. Алексеева, Н.С. Юдин

1020 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Применение программно-информационной системы ANDSystem для поиска мишеней таргетной терапии ревматоидного артрита на основе анализа биологических процессов.
Е.Л. Мищенко, И.В. Яцык, П.С. Деменков, А.В. Адамовская, Т.В. Иванисенко, М.А. Клещев, В.А. Иванисенко

1031 ОБЗОР
Математические модели метаболизма железа: структура и функции.
Н.И. Мельченко, И.Р. Акбердин

1041 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Идентификация и анализ сетевой структуры связей между компонентами иммунной системы у детей.
Д.С. Гребенников, А.П. Топтыгина, Г.А. Бочаров

1051 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Самообучающиеся виртуальные организмы в физическом симуляторе: об оптимальном разрешении их зрительной системы, архитектуре нервной системы и вычислительной сложности задачи.
М.С. Зенин, А.П. Девятериков, А.Ю. Пальянов

Структурная компьютерная биология

1062 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Молекулярно-динамический анализ функциональной роли аминокислотных остатков V99, F124 и S125 ДНК-диоксигеназы человека ABH2.
М. Чжао, Т.Е. Тюгашев, А.Т. Давлетгильдеева, Н.А. Кузнецов

1073 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Структурные основы влияния фосфорамидной *N*-бензимидазольной группы на эффективность удлинения модифицированного праймера Taq ДНК-полимеразой.
А.А. Бердюгин, В.М. Голышев, А.А. Ломзов

1084 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Предсказание взаимодействий белка ORF3a SARS-CoV-2 с низкомолекулярными лигандами с использованием когнитивной платформы ANDSystem, графовых нейронных сетей и молекулярного моделирования.
Т.В. Иванисенко, П.С. Деменков, М.А. Клещев, В.А. Иванисенко

1097 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Влияние димерных бисбензимидазолов на активность ферментов репарации ДНК тирозил-ДНК-фосфодиэстераз 1 и 2 и поли(АДФ-рибоза)полимераз 1 и 2.
Н.С. Дырхеева, И.А. Чернышова, А.Ф. Арутюнян, А.Л. Захаренко, М.М. Кутузов, К.Н. Науменко, А.С. Вензель, В.А. Иванисенко, С.М. Деев, А.Л. Жузе, О.И. Лаврик

Экологическая и популяционная генетика

1109 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Компьютерное моделирование пространственной динамики и первичной генетической дивергенции в системе популяций на кольцевом ареале.
М.П. Кулаков, О.Л. Жданова, Е.Я. Фрисман

Эволюционная биоинформатика

1122 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Асимметрия нуклеотидных замен в тРНК свидетельствует об общем происхождении современных организмов от термофильного предка.
И.И. Титов

Медицинская биоинформатика

1129 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Оценка зависимости показателей мозговой активности от индивидуальной однонуклеотидной вариабельности генетических маркеров большого депрессивного расстройства с использованием анализа главных компонент. *К.А. Зорина, А.А. Кривецкий, В.С. Карманов, А.Н. Савостьянов*

1137 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
Старение кожи связано с локальным дисбалансом в Т-клеточном иммунитете.
К.С. Матвеева, С.К. Колмыков, Т.С. Соколова, Д.Р. Салимов, Д.В. Шевырев

Инструментальные средства биоинформатики

1145 ОРИГИНАЛЬНОЕ ИССЛЕДОВАНИЕ
OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения. *Е.В. Малюгин, Д.А. Афонников*

© Сибирское отделение Российской академии наук, 2025
© Институт цитологии и генетики СО РАН, 2025
Вавиловский журнал генетики и селекции, 2025

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII • VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2025 • 29 • 7

911 FROM THE EDITOR
N.A. Kolchanov, T.A. Bukharina

Computational genomics

913 ORIGINAL ARTICLE
DNA damage reflected in the evolution of G-runs in genomes. *I.R. Grin, D.O. Zharkov*

925 ORIGINAL ARTICLE
Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs.
V.G. Levitsky, T.Yu. Vatolina, V.V. Raditsa

940 ORIGINAL ARTICLE
PlantReg 1.1 identifies the mutual arrangement of transcription factor binding sites in the target promoters for the elucidation of molecular mechanisms within regulatory networks. *V.V. Lavrekha, N.A. Omelyanchuk, A.G. Bogomolov, Y.A. Ryabov, P.K. Mukebenova, E.V. Zemlyanskaya*

952 ORIGINAL ARTICLE
FlyDEGdb knowledge base on differentially expressed genes of *Drosophila melanogaster*, a model object in biomedicine.
O.A. Podkolodnaya, M.A. Deryuzhenko, N.N. Tverdokhleb, K.A. Zolotareva, Yu.V. Makovka, N.L. Podkolodny, V.V. Suslov, I.V. Chadaeva, L.A. Fedoseeva, A.A. Seryapina, D.Yu. Oshchepkov, A.G. Bogomolov, E.Yu. Kondratyuk, O.E. Redina, A.L. Markel, N.E. Gruntenko, M.P. Ponomarenko

Systems computational biology

963 ORIGINAL ARTICLE
Reconstruction and analysis of the gene network regulating apoptosis in hepatocellular carcinoma based on scRNA-seq data and the ANDSystem knowledge base.
A.V. Adamovskaya, I.V. Yatsyk, M.A. Kleshchev, P.S. Demenkov, T.V. Ivanisenko, V.A. Ivanisenko

978 ORIGINAL ARTICLE
Hedgehog signaling in humans: the HH_Signal_pathway_db knowledge base.
T.A. Bukharina, A.M. Bondarenko, D.P. Furman

990 ORIGINAL ARTICLE
Identification of proteins regulating phenotype-associated genes of M2 macrophages: a bioinformatic analysis.
E.A. Antropova, I.V. Yatsyk, P.S. Demenkov, T.V. Ivanisenko, V.A. Ivanisenko

1000 ORIGINAL ARTICLE
In silico reconstruction of the gene network for cytokine regulation of ASD-associated genes and proteins. *N.M. Levanova, E.G. Vergunov, A.N. Savostyanov, I.V. Yatsyk, V.A. Ivanisenko*

1009 ORIGINAL ARTICLE
The gene network and knowledge base on human thermoregulation.
E.V. Ignatieva, P.S. Demenkov, A.G. Bogomolov, R.A. Ivanov, S.A. Lashin, A.D. Mikhailova, A.E. Alekseeva, N.S. Yudin

1020 ORIGINAL ARTICLE
Searching for biological processes as targets for rheumatoid arthritis targeted therapy with ANDSystem, an integrated software and information platform.
E.L. Mishchenko, I.V. Yatsyk, P.S. Demenkov, A.V. Adamovskaya, T.V. Ivanisenko, M.A. Kleshchev, V.A. Ivanisenko

1031 REVIEW
Mathematical models of iron metabolism: structure and functions. *N.I. Melchenko, I.R. Akberdin*

1041 ORIGINAL ARTICLE
Identification and analysis of the connection network structure between the components of the immune system in children.
D.S. Grebennikov, A.P. Toptygina, G.A. Bocharov

1051 ORIGINAL ARTICLE
Self-learning virtual organisms in a physics simulator: on the optimal resolution of their visual system, the architecture of the nervous system and the computational complexity of the problem.
M.S. Zenin, A.P. Devyaterikov, A.Yu. Palyanov

Structural computational biology

1062 ORIGINAL ARTICLE
Molecular dynamic analysis of the functional role of amino acid residues V99, F124 and S125 of human DNA dioxygenase ABH2.
M. Zhao, T.E. Tyugashev, A.T. Davletgildeeva, N.A. Kuznetsov

1073 ORIGINAL ARTICLE
Structural basis of the phosphoramidate *N*-benzimidazole group's influence on modified primer extension efficiency by Taq DNA polymerase.
A.A. Berdugin, V.M. Golyshev, A.A. Lomzov

1084 ORIGINAL ARTICLE
Prediction of interactions between the SARS-CoV-2 ORF3a protein and small-molecule ligands using the ANDSystem cognitive platform, graph neural networks, and molecular modeling. *T.V. Ivanisenko, P.S. Demenkov, M.A. Kleshchev, V.A. Ivanisenko*

1097 ORIGINAL ARTICLE
The effect of dimeric bisbenzimidazoles on the activity of DNA repair enzymes TDP1, TDP2, PARP1 and PARP2.
N.S. Dyrkheeva, I.A. Chernyshova, A.F. Arutyunyan, A.L. Zakharenko, M.M. Kutuzov, K.N. Naumenko, A.S. Venzel, V.A. Ivanisenko, S.M. Deyev, A.L. Zhuze, O.I. Lavrik

Ecological and population genetics

1109 ORIGINAL ARTICLE
Computer modeling of spatial dynamics and primary genetic divergence for a population system in a ring areal.
M.P. Kulakov, O.L. Zhdanova, E.Ya. Frisman

Evolutionary bioinformatics

1122 ORIGINAL ARTICLE
Asymmetry of nucleotide substitutions in tRNAs indicates common descent of modern organisms from a thermophilic ancestor. *I.I. Titov*

Medical bioinformatics

1129 ORIGINAL ARTICLE
Assessing the dependence of brain activity on individual single-nucleotide variability of genetic markers of major depressive disorder using principal component analysis. *K.A. Zorina, A.A. Kriveckiy, V.S. Karmanov, A.N. Savostyanov*

1137 ORIGINAL ARTICLE
Senescent cell accumulation is associated with T-cell imbalance in the skin.
K.S. Matveeva, S.K. Kolmykov, T.S. Sokolova, D.R. Salimov, D.V. Shevyrev

Bioinformatics tools

1145 ORIGINAL ARTICLE
OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning.
E.V. Malyugin, D.A. Afonnikov

© Siberian Branch RAS, 2025
© Institute of Cytology and Genetics, SB RAS, 2025

Вавиловский журнал генетики и селекции. 2025;29(7):911-912
doi 10.18699/vjgb-25-97

Н.А. Колчанов

Т.А. Бухарина

Уважаемые коллеги!

Предлагаем вашему вниманию очередной выпуск «Вавиловского журнала генетики и селекции», посвященный вопросам компьютерной биологии.

В последние 20 лет происходило стремительное развитие методов секвенирования геномов. Стоимость их расшифровки упала почти на пять порядков (для генома человека, например, со 100 000 до 500 долларов). Стремительно развивались транскриптомика, протеомика, метаболомика и другие омиксные техологии, возникло новое поколение методов визуализации биологических объектов на геномном, клеточном, тканевом, организменном уровнях организации живых систем. В результате в генетике произошел информационный взрыв: она стала главным источником больших данных, обогнав по темпам накопления информации не только другие науки, но и мировые социальные сети. Действительно, сейчас за год в науках о жизни продуцируется до 40 эксабайт геномных данных, тогда как крупнейшая социальная платформа YouTube дает только 2 эксабайта данных, т. е. в 20 раз меньше.

Анализ больших генетических данных привел к формированию новой парадигмы современной генетики. Ее центральным объектом становятся генные сети – группы координированно функционирующих генов, которые взаимодействуют друг с другом через продукты – РНК, белки, метаболиты, обеспечивая на основе информации, закодированной в геномах, формирование фенотипических характеристик организмов (молекулярных, биохимических, клеточных, физиологических, морфологических, поведенческих и т. д.). Была выявлена огромная сложность регуляции генных сетей. Проявляется она в том, что функционирование любого отдельного элемента генной сети может контролироваться десятками и сотнями элементарных регуляторных процессов. Это относится как к регуляции транскрипции генов, которую контролируют десятки транскрипционных факторов, взаимодействующих с сайтами связывания в промоторах генов, функционированию белков, активность которых модулируется взаимодействием с большим количеством лигандов – аллостерических регуляторов, так и к метаболическим путям, в которых количество элементарных управляющих процессов может быть на порядок больше, чем число биохимических реакций. Еще одна фундаментальная особенность живых систем, выявленная при анализе больших генетических данных, – исключительно высокий уровень геномной изменчивости в популяциях человека, животных, растений и микроорганизмов.

Для анализа больших генетических данных необходима разработка нового поколения методов, способных работать со сверхбольшими объемами информации. Это новые методы биоинформатики для реконструкции, анализа и моделирования структурной организации и молекулярно-генетических механизмов функционирования геномов, генов и кодируемых ими генетических макромолекул – РНК и белков. Это также новые методы системной компьютерной биологии для реконструкции, анализа и

моделирования генных сетей и молекулярно-генетических систем, функционирующих на уровне клеток, тканей, органов и целостных организмов.

Новая эра больших генетических данных, в которую входят науки о жизни, требует трансформации базовых подходов биоинформатики и системной компьютерной биологии. Каковы же здесь базовые тренды? Прежде всего, это интеграция классических методов биоинформатики и системной компьютерной биологии с методами искусственного интеллекта и глубокого машинного обучения. Разработка на этой основе нового поколения гибридных информационно-программных систем для интерпретации больших генетических данных, а также, что особенно важно, для планирования экспериментов по проверке результатов компьютерных предсказаний, полученных из анализа больших генетических данных. Движение в этом направлении будет означать принципиальную трансформацию базовой парадигмы современных исследований: дополнение науки, направляемой гипотезами, новой наукой, направляемой результатами анализа больших данных.

И такое движение идет во всех науках, но именно биоинформатика и системная компьютерная биология находятся на его острие.

Научные редакторы выпуска:

академик Н.А. Колчанов,
научный руководитель ФИЦ ИЦиГ СО РАН

Т.А. Бухарина, ученый секретарь
отдела системной биологии ФИЦ ИЦиГ СО РАН

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):913-924

doi 10.18699/vjgb-25-98

Отражение процессов повреждения ДНК в эволюции G-трактов в геномах

И.Р. Грин [1], Д.О. Жарков [1, 2]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
dzharkov@niboch.nsc.ru

Аннотация. Окисление ДНК представляет собой один из главных видов повреждения генетического материала живых организмов. Из многих десятков продуктов окислительного повреждения ДНК в наибольшем количестве встречается 8-оксогуанин (8-oxoG) – предмутагенное основание, приводящее при репликации к трансверсиям G→T. Двуцепочечная ДНК обладает способностью к проводимости положительных зарядов, связанных с дефицитом электронов в π-системе азотистых оснований. Такие заряды в конечном итоге локализуются на 5′-концевом нуклеотиде полигуаниновых трактов (G-трактов). В связи с этим 5′-концевые нуклеотиды G-трактов служат характерными местами образования 8-oxoG. Эти свойства G-трактов хорошо изучены *in vitro* на уровне реакционной способности, но остается неясным, насколько они могут отражаться в спектрах мутагенеза *in vivo*. В работе проанализирован нуклеотидный контекст G-трактов в репрезентативном наборе из 62 полных геномов прокариот и в геноме человека с покрытием «от теломеры до теломеры». Показано, что G-тракты в среднем короче полиадениновых трактов (A-трактов) и вероятность удлинения G-трактов на один нуклеотид ниже, чем в случае A-трактов. Установлено, что представленность T в положении, примыкающем к G-трактам с 5′-стороны, повышена, в особенности у организмов с аэробным метаболизмом, что согласуется с моделью преимущественных мутаций G→T в 5′-положении с 8-oxoG как предшественником. В то же время в положении, примыкающем к A-трактам, повышена частота встречаемости G и C и снижена частота встречаемости T. В геноме человека наблюдается двухфазный характер разрастания G-трактов: начиная с длины 8–9 нуклеотидов вероятность их удлинения на один нуклеотид заметно увеличивается. Выявлена повышенная представленность C с 5′-стороны от длинных G-трактов и A при заменах в теломерных повторах, что может свидетельствовать о существовании мутагенных процессов, механизм которых пока не охарактеризован, но может быть связан с ошибками ДНК-полимераз при репликации продуктов дальнейшего окисления 8-oxoG.

Ключевые слова: повреждение ДНК; мутагенез; 8-оксогуанин; G-тракты; теломеры

Для цитирования: Грин И.Р., Жарков Д.О. Отражение процессов повреждения ДНК в эволюции G-трактов в геномах. *Вавиловский журнал генетики и селекции*. 2025;29(7):913-924. doi 10.18699/vjgb-25-98

Финансирование. Исследование выполнено при финансовой поддержке РНФ в рамках проекта № 24-14-00285 (анализ генома человека). Часть работы, посвященная анализу геномов прокариот, поддержана государственным заданием ИХБФМ СО РАН 125012300657-2.

DNA damage reflected in the evolution of G-runs in genomes

I.R. Grin [1], D.O. Zharkov [1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
dzharkov@niboch.nsc.ru

Abstract. DNA oxidation is one of the main types of damage to the genetic material of living organisms. Of the many dozens of oxidative lesions, the most abundant is 8-oxoguanine (8-oxoG), a premutagenic base that leads to G→T transversions during replication. Double-stranded DNA can conduct holes through the π system of stacked nucleobases. Such electron vacancies are ultimately localized at the 5′-terminal nucleotides of polyguanine runs (G-runs), making these positions characteristic sites of 8-oxoG formation. While such properties of G-runs have been studied *in vitro* at the level of chemical reactivity, the extent to which they can influence mutagenesis spectra *in vivo* remains unclear. Here, we have analyzed the nucleotide context of G-runs in a representative set of 62 high-quality prokaryotic genomes and in the human telomere-to-telomere genome. G-runs were, on average, shorter than polyadenine runs (A-runs), and the probability of a G-run being elongated by one nucleotide is lower than in the case of A-runs. The representation of T in the position 5′-flanking G-runs is increased, especially in organisms with aerobic metabolism, which is consistent with the model of preferential G→T substitutions at the 5′-position with 8-oxoG as a precursor. Conversely, the frequency of G and C is increased and the frequency of T is decreased in the position 5′-flanking A-runs. A biphasic pattern of G-run expansion is observed in the human genome: the probability of sequences longer than 8–9 nucleotides being

© Грин И.Р., Жарков Д.О., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

elongated by one nucleotide increases significantly. An increased representation of C in the 5′-flanking position to long G-runs was found, together with an elevated frequency of 5′-G→A substitutions in telomere repeats. This may indicate the existence of mutagenic processes whose mechanism has not yet been characterized but may be associated with DNA polymerase errors during replication of the products of further oxidation of 8-oxoG.
Key words: DNA damage; mutagenesis; 8-oxogianine; G-runs; telomeres

For citation: Grin I.R., Zharkov D.O. DNA damage reflected in the evolution of G-runs in genomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):913-924. doi 10.18699/vjgb-25-98

Введение

Окислительное повреждение ДНК – неизбежный спутник метаболизма, использующего окисление органических соединений молекулярным кислородом, который составляет основу энергетического обмена подавляющего большинства живых организмов уже более двух миллиардов лет (Prorok et al., 2021). Поврежденные нуклеотиды обычно достаточно быстро подвергаются репарации, однако некоторая их часть может оставаться в составе ДНК до момента репликации, что служит одной из главных причин мутаций (Liu et al., 2016; Chatterjee, Walker, 2017; Tubbs, Nussenzweig, 2017). Благодаря пониманию механизмов возникновения ошибок при работе ДНК-полимераз на молекулярном уровне в настоящее время стало возможным выделять характерные паттерны мутаций, вызванных тем или иным видом генотоксического стресса или даже конкретными поврежденными азотистыми основаниями (Alexandrov et al., 2013; Koh et al., 2021).

Из всех структурных элементов ДНК самым низким окислительно-восстановительным потенциалом характеризуется гуанин (Cadet et al., 2008, 2017; Fleming, Burrows, 2022). Самый распространенный продукт его окисления – 7,8-дигидро-8-оксогуанин (8-oxoG) – встречается в ДНК на фоновом уровне ~$1/10^6$ гуанинов, и этот уровень значительно повышается при окислительном стрессе разного происхождения (ESCODD et al., 2005; Dizdaroglu et al., 2015; Chiorcea-Paquim, 2022; рис. 1, *а*, *б*). Ввиду того, что при атоме C8 в 8-oxoG находится объемный атом кислорода, в соответствующем нуклеозиде, 8-оксо-2′-дезоксигуанозине (8-oxodG), стандартная *анти*-конформация стерически затруднена, а *син*-конформация оказывается более выгодной (Cho et al., 1990; см. рис. 1, *в*, *г*). Это приводит к тому, что в отсутствие уотсон-криковских связей с цитозином, дополнительно стабилизирующих *анти*-конформацию, 8-оксо-2′-дезоксигуанозин преимущественно переходит в *син*-конформацию, в которой может образовывать пару хугстеновского типа с аденином (Kouchakdjian et al., 1991; McAuley-Hecht et al., 1994; Lipscomb et al., 1995). Из-за этого ДНК-полимеразы с высокой частотой включают dAMP напротив 8-oxoG в ДНК-матрице (Shibutani et al., 1991; Miller, Grollman, 1997; Maga et al., 2007; Yudkina et al., 2019).

В живых клетках на исход первичных событий окисления ДНК может оказывать влияние множество дополнительных факторов, и существуют системы репарации ДНК, удаляющие из генома поврежденные основания. Даже с учетом этого 8-oxoG проявляет заметную мутагенность *in vivo*, которая характеризуется спектром с доминирующими трансверсиями G→T, слабо зависящими от окружающего однонуклеотидного контекста (Wood et al., 1992; Moriya, 1993). Такие мутации часто встречаются в опухолях человека и составляют основу мутационных подписей SBS18 и SBS36 (Alexandrov et al., 2013; Pilati et al., 2017; Viel et al., 2017; Kucab et al., 2019). Заметный вклад в мутагенез вносят и продукты дальнейшего окисления 8-oxoG – гуанидиногидантоин и спироиминодигидантоин, которые преимущественно вызывают трансверсии G→C (Fleming, Burrows, 2017; Kino et al., 2020).

Сопряженная π-система ДНК обладает достаточно высокой дырочной проводимостью (Giese, 2002; Genereux, Barton, 2010). Многочисленные эксперименты и квантовомеханические расчеты показывают, что при одноэлектронном окислении ДНК образовавшийся положительный заряд мигрирует по π-системе на значительные расстоя-

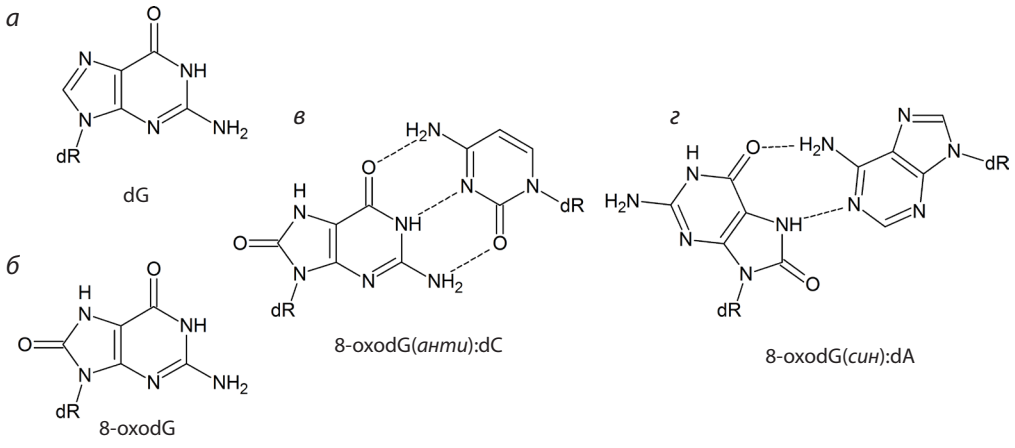


Рис. 1. Структуры 2′-дезоксигуанозина (*а*), 8-оксо-2′-дезоксигуанозина (*б*), пары уотсон-криковского типа 8-oxodG(*анти*):dC (*в*) и пары хугстеновского типа 8-oxodG(*син*):dA (*г*).

ния, а его конечными акцепторами служат основания G, которые главным образом окисляются до 8-oxoG. При этом особенно чувствительны к окислению гуанины, находящиеся в крайнем 5′-положении в трактах из нескольких G (Sugiyama, Saito, 1996; Saito et al., 1998; Kurbanyan et al., 2003; Adhikary et al., 2009).

Несмотря на то что механизм миграции положительного заряда и преимущественного окисления гуанинов на 5′-конце G-трактов сегодня считается общепринятым, все экспериментальные данные в его поддержку получены в простых системах *in vitro*. Исследований спектров мутагенеза, вызванного появлением 8-oxoG в таком контексте, до сих пор не проводилось. Если на 5′-конце G-трактов действительно происходит преимущественное превращение G в 8-oxoG, можно ожидать, что из-за мутагенных свойств 8-oxoG в таких позициях будет повышена частота мутаций G→T, что должно отражаться в увеличении встречаемости T перед трактами из нескольких G. В данной работе с целью проверки этой гипотезы проведен анализ встречаемости нуклеотидов, примыкающих с 5′-стороны к G-трактам (5′-фланговых нуклеотидов), в геномах прокариот и человека.

Материалы и методы

Для анализа были использованы сборка генома человека T2T-CHM13v2.0, включающая полноразмерные теломеры и высокоповторные участки (Nurk et al., 2022), и геномы прокариот, перечисленные в табл. 1.

Для работы с последовательностями геномов использовали программный пакет UGENE v37.0 (Okonechnikov et al., 2012) и специально написанные скрипты на языке bash. Ожидаемую частоту нуклеотидов во фланговых позициях перед повторами G_n (либо A_n) и после них в геномах прокариот рассчитывали, исходя из общего количества A, C и T (либо C, G и T) в данном геноме как $p_A = N_A/(N_A+N_C+N_T)$, где p_A – ожидаемая представленность (в данном случае для A), N_A, N_C и N_T – число A, C и T в обеих цепях генома соответственно. Для генома человека, ввиду известной недопредставленности динуклеотида CG, ожидаемую частоту рассчитывали аналогичным образом, исходя из представленности динуклеотидов AG, CG и TG. Статистический анализ проводили в программах SigmaPlot v11.0 (Grafiti, США), DATAPLOT (Национальный институт стандартов и технологий США) и RStudio v1.2 (Posit PBC, США). При всех множественных сравнениях и сериях тестов для корректировки уровня значимости использовали поправку Данн.

Результаты и обсуждение

Для анализа распределения нуклеотидов в геномах прокариот была составлена выборка из 54 геномов бактерий и 8 геномов архей, максимально отражающая таксономическое разнообразие в этих доменах живых организмов (см. табл. 1). В нее вошли геномы высокого качества, отнесенные в базе данных RefSeq (O'Leary et al., 2016) к референтным. Репрезентативность выборки составляла один геном на таксон уровня типа, за исключением типов Methanobacteriota и Thermoproteota для архей и Actinomycetota, Bacteroidota и Thermodesulfobacteriota для бактерий с репрезентативностью два генома из разных порядков на тип, а также Bacillota и Pseudomonadota – три генома из разных порядков на тип. Содержание G+C в исследуемых геномах варьировало от 23.5 до 69 % (см. табл. 1). Геномы архей по своим параметрам достоверно не отличались от геномов бактерий, поэтому представители обоих доменов рассматривались как одна группа прокариот.

Поскольку большую часть прокариотических геномов составляют белок-кодирующие последовательности, мутации в которых могут подвергаться действию естественного отбора, предварительно было оценено возможное влияние всех 16 потенциальных замен аминокислотных остатков, происходящих в результате нуклеотидных замен G→A, G→C и G→T в первой позиции G-тракта (замены кодонов HHG→HHH, HGG→HHG, GGG→HGG, где H – A, C или T). Для этого использовали две независимые метрики: индекс консервативности C_n, рассчитанный на основе условных расстояний в наборе физико-химических свойств аминокислотных остатков (Taylor, 1986; Livingstone, Barton, 1993), и веса аминокислотных замен в матрице BLOSUM62, составленной по набору нескольких сотен групп гомологичных белков (Henikoff S., Henikoff J.G., 1992). Хотя замены G→A в целом вызывали меньшие изменения в свойствах и встречаемости аминокислотных остатков, как и следовало ожидать для точковых мутаций с сохранением класса нуклеотида, отличие от замен G→C и G→T не было статистически значимым (критерий Краскела–Уоллиса с поправкой Данн на множественные сравнения, $p > 0.05$).

Во всех геномных последовательностях был проведен поиск трактов HG_nH, BA_nB и соответствующих им по комплементарной цепи трактов DC_nD и VT_nV (H – A, C или T; B – C, G или T; D – A, G или T; V – A, C или G) с длиной $n \geq 2$. Встречаемость полипуриновых трактов в геномах была выше, чем можно ожидать из случайного распределения нуклеотидов при том же составе G+C (одновыборочный критерий Уилкоксона, $p < 0.001$), что говорит о функциональной важности таких последовательностей. Повышенная частота замен в первой позиции G-трактов должна постепенно приводить к их укорочению. Действительно, при сравнении длин G- и A-трактов в прокариотических геномах с поправкой на содержание соответствующих пуриновых нуклеотидов оказалось, что G-тракты в среднем короче (рис. 2, *а*). При этом тринуклеотиды HGG встречались чаще, чем BAA, однако при большей длине повтора встречаемость A-трактов была выше (см. рис. 2, *б*).

Для более подробного анализа распределения длин трактов была изучена вариабельность их длин в каждом геноме. Количество G- и A-трактов в составе каждого генома убывало практически строго экспоненциально в диапазоне длин от 2 до 5–6. При больших длинах в некоторых случаях наблюдались отклонения в любую сторону из-за небольшого числа таких повторов, в особенности для геномов малого размера (рис. 3, *а*, *б*). По линейной части зависимости логарифма количества повторов от длины тракта можно определить «коэффициент прирастания» k_{inc}, который показывает, насколько легко в геноме данного нуклеотидного состава удлинить тракт на один нуклеотид: чем выше k_{inc}, тем больше в геноме пропорция относительно более длинных трактов. При сравнении

Таблица 1. Геномы прокариот, использованные для анализа

Вид	Тип	Сборка генома	G+C, %	O_2
Домен Archaea				
Methanobacterium formicicum	Methanobacteriota	GCF_001458655.1	41.0	AN
Methanosarcina barkeri		GCF_000970025.1	39.0	AN
Nanobdella aerobiophila	Nanobdellota	GCF_023169545.1	24.5	A
Nitrososphaera viennensis	Nitrososphaerota	GCF_000698785.1	52.5	A
Promethearchaeum syntrophicum	Promethearchaeota	GCF_008000775.2	31.0	AN
Sulfolobus acidocaldarius	Thermoproteota	GCF_000012285.1	36.5	A
Thermoproteus tenax		GCF_000253055.1	55.0	AN
Cand. *Nanohalobium constans*	Cand. Nanohalarchaeota	GCF_009617975.1	43.0	A
Домен Bacteria				
Acidobacterium capsulatum	Acidobacteriota	GCF_000022565.1	60.5	A
Bifidobacterium longum	Actinomycetota	GCF_000196555.1	60.5	AN
Mycobacterium tuberculosis		GCF_000195955.2	65.5	A
Aquifex aeolicus	Aquificota	GCF_000008625.1	43.5	A
Fimbriimonas ginsengisoli	Armatimonadota	GCF_000724625.1	61.0	A
Atribacter laminatus	Atribacterota	GCF_015775515.1	38.5	AN
Bacillus subtilis	Bacillota	GCF_000009045.1	43.5	A
Clostridioides difficile		GCF_018885085.1	28.5	AN
Lactococcus lactis		GCF_003176835.1	35.0	A
Bacteroides fragilis	Bacteroidota	GCF_000025985.1	43.0	AN
Saprospira grandis		GCF_000250635.1	46.5	A
Cyclonatronum proteinivorum	Balneolota	GCF_003353065.1	51.5	A
Bdellovibrio bacteriovorus	Bdellovibrionota	GCF_000196175.1	50.5	A
Caldisericum exile	Caldisericota	GCF_000284335.1	35.5	AN
Calditrix abyssi	Calditrichota	GCF_001886815.1	45.0	AN
Campylobacter jejuni	Campylobacterota	GCF_000009085.1	30.5	A
Chlamydia trachomatis	Chlamydiota	GCF_000008725.1	41.5	AN
Chlorobium limicola	Chlorobiota	GCF_000020465.1	51.5	AN
Chloroflexus aurantiacus	Chloroflexota	GCF_000018865.1	56.5	A
Desulfurispirillum indicum	Chrysiogenota	GCF_000177635.2	56.0	AN
Coprothermobacter proteolyticus	Coprothermobacterota	GCF_000020945.1	45.0	AN
Synechococcus elongatus	Cyanobacteriota	GCF_022984195.1	55.5	A
Deferribacter thermophilus	Deferribacterota	GCF_049472675.1	30.5	AN
Deinococcus radiodurans	Deinococcota	GCF_020546685.1	66.5	A
Dictyoglomus thermophilum	Dictyoglomota	GCF_000020965.1	33.5	AN
Elusimicrobium minutum	Elusimicrobiota	GCF_000020145.1	40.0	AN
Fibrobacter succinogenes	Fibrobacterota	GCF_000146505.1	48.0	AN
Fidelibacter multiformis	Fidelibacterota	GCF_041154365.1	45.5	AN
Fusobacterium nucleatum	Fusobacteriota	GCF_003019295.1	27.0	AN
Gemmatimonas aurantiaca	Gemmatimonadota	GCF_000010305.1	64.5	A
Ignavibacterium album	Ignavibacteriota	GCF_000258405.1	34.0	A
Kiritimatiella glycovorans	Kiritimatiellota	GCF_001017655.1	63.5	AN
Lentisphaera profundi	Lentisphaerota	GCF_028728065.1	40.5	A
Mycoplasma mycoides	Mycoplasmatota	GCF_018389705.1	23.5	A
Myxococcus xanthus	Myxococcota	GCF_000012685.1	69.0	A
Nitrospina watsonii	Nitrospinota	GCF_946900835.1	57.0	A
Nitrospira moscoviensis	Nitrospirota	GCF_001273775.1	62.0	A
Planctopirus limnophila	Planctomycetota	GCF_000092105.1	53.5	A

Окончание табл. 1

Вид	Тип	Сборка генома	G+C, %	O_2
Escherichia coli	Pseudomonadota	GCF_000005845.2	51.0	A
Pseudomonas aeruginosa		GCF_000006765.1	66.5	A
Sphingomonas paucimobilis		GCF_016027095.1	65.5	A
Rhodothermus marinus	Rhodothermota	GCF_000024845.1	64.5	A
Spirochaeta thermophila	Spirochaetota	GCF_000184345.1	61.0	AN
Thermanaerovibrio acidaminovorans	Synergistota	GCF_000024905.1	64.0	AN
Desulfovibrio desulfuricans	Thermodesulfobacteriota	GCF_017815575.1	57.0	AN
Thermodesulfobacterium commune		GCF_000734015.1	37.0	AN
Thermodesulfobium narugense	Thermodesulfobiota	GCF_000212395.1	34.0	AN
Thermomicrobium roseum	Thermomicrobiota	GCF_000021685.1	64.5	A
Thermosulfidibacter takaii	Thermosulfidibacterota	GCF_001547735.1	43.0	AN
Thermotoga maritima	Thermotogota	GCF_000230655.2	46.0	AN
Verrucomicrobium spinosum	Verrucomicrobiota	GCF_000172155.1	60.5	A
Vulcanimicrobium alpinum	Vulcanimicrobiota	GCF_027923555.1	68.5	A
Cand. *Cloacimonas acidaminovorans*	Cand. Cloacimonadota	GCF_000146065.2	38.0	AN
Cand. *Velamenicoccus archaeovorus*	Cand. Omnitrophota	GCF_004102945.1	53.0	AN

Примечание. Номер сборки генома взят из базы данных RefSeq (O'Leary et al., 2016). A – аэробы и факультативные анаэробы; AN – анаэробы.

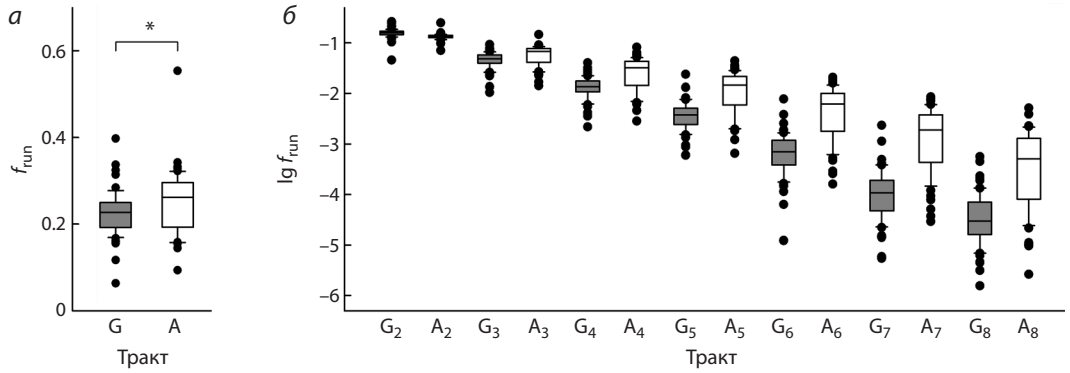


Рис. 2. Длина полипуриновых трактов в геномах прокариот. *а* – общая доля G или A в составе трактов любой длины среди числа соответствующих пуриновых нуклеотидов в геноме (f_{run}), * $p < 0.05$ (критерий Манна–Уитни); *б* – доля G или A в составе трактов длины 2–8 среди числа соответствующих пуриновых нуклеотидов в геноме. Во всех случаях разница между G- и A-трактами значима на уровне $p < 0.001$ (критерий Манна–Уитни).

Здесь и на последующих рисунках линия в «ящике» соответствует медиане, границы «ящика» – первому и третьему квартилям, «усы» – 10-му и 90-му перцентилям, точки – значениям выбросов.

зависимости значений k_{inc} для G- и A-трактов в геномах разного состава выяснилось, что G-тракты с возрастанием содержания G+C удлиняются медленнее, чем удлиняются A-тракты с возрастанием содержания A+T в геноме (см. рис. 3, *в*). Таким образом, баланс удлинения и укорочения G-трактов, определяемый многими факторами, в геномах прокариот смещен в сторону укорочения по сравнению с A-трактами.

Длины полипуриновых трактов могут изменяться в любую сторону из-за проскальзывания ДНК-полимераз при синтезе ДНК (Kunkel, Bebenek, 2000) или в результате отбора по физико-химическим свойствам полипуриновых участков (Bansal et al., 2022), однако эти процессы не зависят от окружающих тракт нуклеотидов. Напротив, укорочение G-трактов вследствие повреждения 5′-концевого основания должно сопровождаться характерным мутационным спектром, который определяется свойствами репликативных ДНК-полимераз. В связи с этим было интересно установить, насколько представленности 5′-фланговых нуклеотидов отличаются друг от друга и от их общей представленности по геному. Для количественной характеристики этих отличий был введен параметр Δrep – разница между наблюдаемой и ожидаемой представленностью каждого нуклеотида. Оказалось, что представленность нуклеотида T в первой позиции перед G-трактами статистически значимо повышена по сравнению с ожидаемой и по сравнению с представленностью A и C (рис. 4, *а*). Встречаемость нуклеотидов A и C в этой позиции была несколько ниже ожидаемой, но это отличие не достигало значимости; также по своей представленности

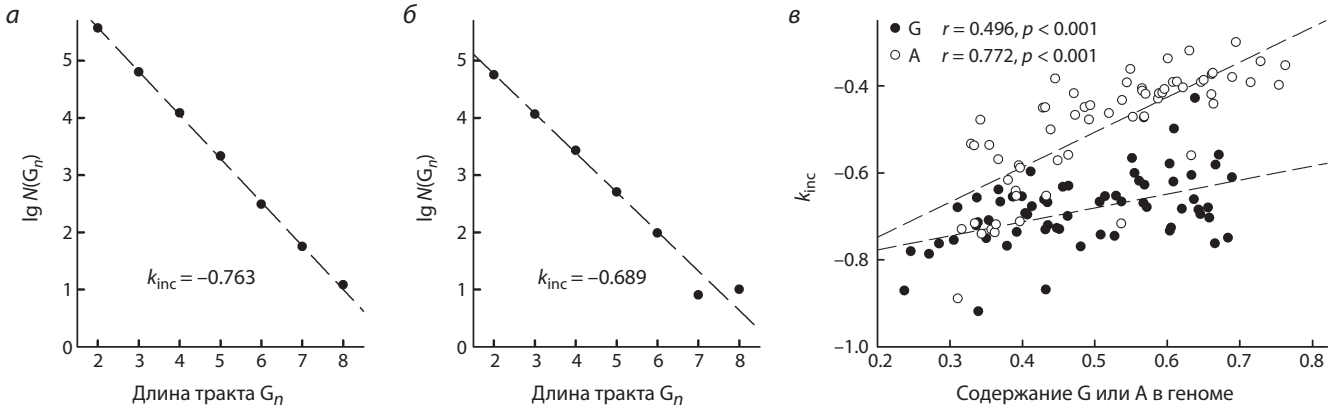


Рис. 3. Зависимость количества полипуриновых трактов в геномах прокариот от длины тракта и нуклеотидного состава генома.

а, *б* – примеры зависимости числа G-трактов $N(G_n)$ от их длины для геномов *E. coli* (*а*), размер генома 4.64×10^6 п.н., содержание G+C 51.0 %, и *Ch. trachomatis* (*б*), размер генома 1.04×10^6 п.н., содержание G+C 41.5 %; *в* – зависимость k_{inc} от нуклеотидного состава генома (содержания G+C для G-трактов, содержания A+T для A-трактов). Черные точки – G-тракты, белые точки – A-тракты; штриховые линии соответствуют линейной регрессии, коэффициенты регрессии приведены на рисунке.

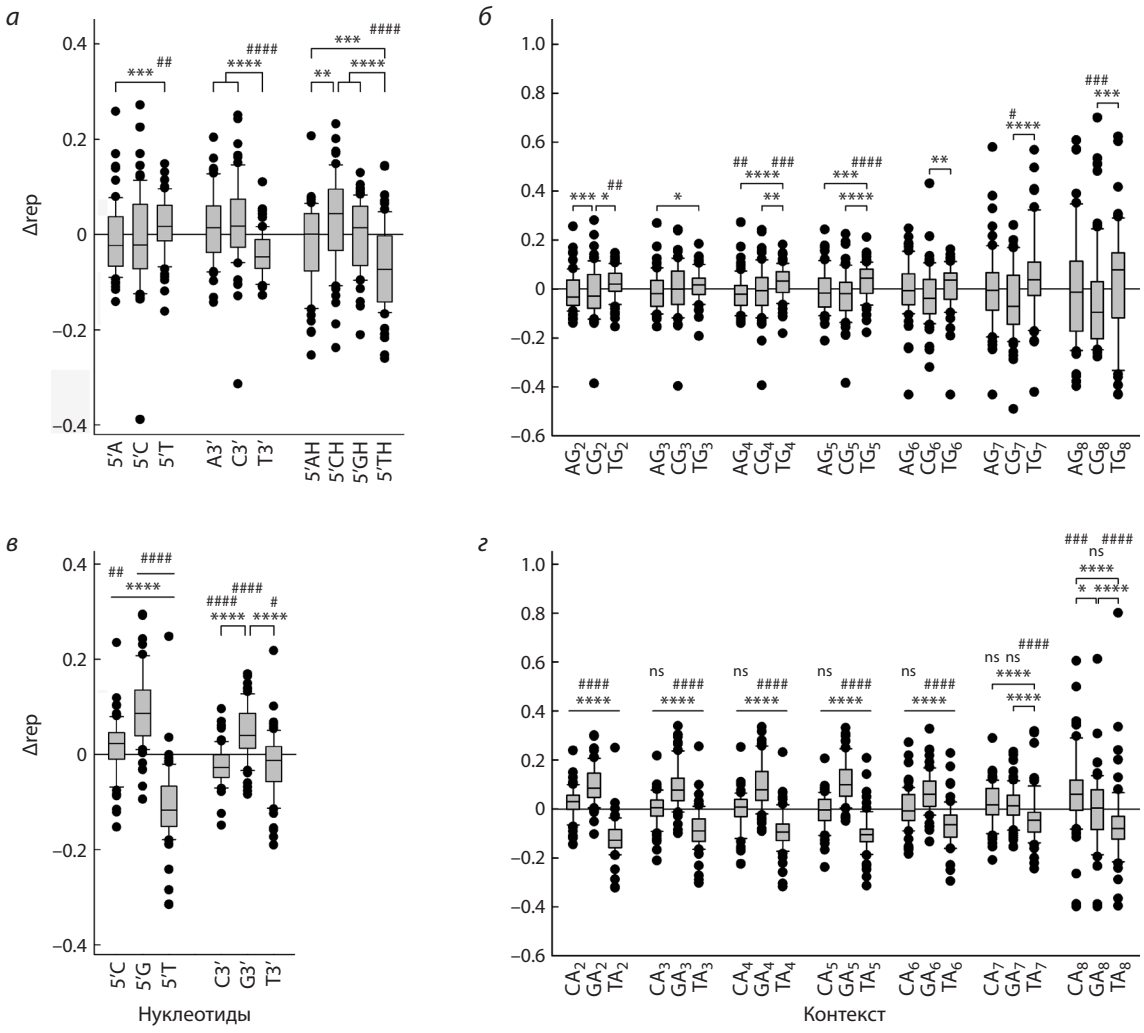


Рис. 4. Представленность разных 5′- и 3′-фланговых нуклеотидов при полипуриновых трактах.

а, *в* – отклонение от представленности 5′- и 3′-фланговых нуклеотидов для G-трактов (*а*) и A-трактов (*в*), ожидаемой на основе содержания соответствующего нуклеотида в геноме, в трактах любой длины. *б*, *г* – отклонение от представленности 5′-фланговых нуклеотидов в G-трактах (*б*) и A-трактах (*г*) длины 2–8. Отличие от ожидаемого: $^{\#}p<0.05$, $^{\#\#}p<0.01$, $^{\#\#\#}p<0.005$, $^{\#\#\#\#}p<0.001$ (одновыборочный критерий Уилкоксона с поправкой Данн на множественные сравнения); ns – нет достоверных отличий. Различия между группами: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.005$, $^{****}p<0.001$ (критерий Краскела–Уоллиса с поправкой Данн на множественные сравнения).

они не отличались друг от друга. T встречался чаще, чем хотя бы один из нуклеотидов A или C, при любой длине G-тракта, при этом его представленность по сравнению с ожидаемой была выше перед трактами G_2, G_4, G_5 и G_6 (см. рис. 4, *б*). Нуклеотид A был недопредставлен в этой позиции лишь перед трактами G_4, а нуклеотид C – перед G-трактами длиной 5 и более. Напротив, как с 3′-стороны от G-трактов, так и во второй позиции с 5′-стороны от них T был представлен меньше ожидаемого (см. рис. 4, *а*). В целом эти данные свидетельствуют в пользу модели преимущественного окисления первого G в трактах до 8-oxoG с последующей трансверсией G→T.

Достаточно неожиданным оказалось то, что перед A-трактами распределение нуклеотидов было неравномерным в еще большей степени, чем перед G-трактами. В этой позиции T был недопредставлен, а C и G, напротив, перепредставлены (см. рис. 4, *в*). Для C данное отклонение объяснялось в основном перепредставленностью в тринуклеотидах CAA, тогда как для G наблюдалась повышенная частота встречаемости вплоть до длины тракта в 6 нуклеотидов (см. рис. 4, *г*). Снижение доли T также имело место в трактах любой длины (см. рис. 4, *г*). После A-трактов встречаемость C и T была ниже ожидаемой, а G – выше ожидаемой (см. рис. 4, *в*). Не исключено, что эти отклонения тоже могут быть объяснены повреждением ДНК и последующими ошибками ДНК-полимераз, однако в настоящий момент физико-химическая природа таких событий остается неясной.

Количество возникающего в геноме 8-oxoG напрямую зависит от наличия во внутриклеточной среде активных форм кислорода (Halliwell, Gutteridge, 2015). Прокариоты отличаются исключительным многообразием вариантов энергетического метаболизма: некоторые ведут строго анаэробный образ жизни, в то время как другие относятся к облигатным аэробам или факультативным анаэробам и подвержены окислительному стрессу более высокой интенсивности. Было проведено сравнение статистики встречаемости 5′-фланговых нуклеотидов в геномах двух рассматриваемых групп (см. табл. 1). В геномах аэробных прокариот в этой позиции при G-трактах с повышенной частотой по сравнению с ожидаемой встречался нуклеотид T и с пониженной – A (рис. 5). Для анаэробных микроорганизмов значимой разницы во встречаемости 5′-фланговых нуклеотидов не обнаружено (см. рис. 5). Однако при сравнении представленности A, C и T непосредственно между группами аэробов и анаэробов различия не достигали статистической значимости, что, скорее всего, связано с недостаточным объемом выборки. При A-трактах разница во встречаемости 5′-фланговых нуклеотидов в геномах аэробов и анаэробов была такой же, как и в объединенной группе (ср. рис. 4, *в* и рис. 5). Таким образом, сниженный уровень окислительного стресса у анаэробных микроорганизмов может быть ассоциирован с менее выраженным преобладанием T в положении, примыкающем с 5′-стороны к G-трактам. Для более определенного ответа на этот вопрос требуются дополнительные исследования.

В отличие от прокариот, геномы эукариот характеризуются большим количеством повторяющихся элементов – транспозонов, сателлитной и теломерной ДНК, точная последовательность которых плохо определяется традиционными методами высокопроизводительного секвенирования (Richard et al., 2008; Liao et al., 2023). Появление методов сверхдлинного секвенирования (Oxford Nanopore, PacBio HiFi) позволило заполнить эти пробелы. Недавно опубликованный геном человека, прочитанный при помощи комбинации нескольких методов с покрытием «от теломеры до теломеры» (T2T) и высоким качеством (оценка частоты ошибок в теломерных областях $\sim 4\times10^{-8}$) (Nurk et al., 2022), предоставляет возможность анализа окружения G-трактов без искажений, вызванных более высокой представленностью уникальных последовательностей.

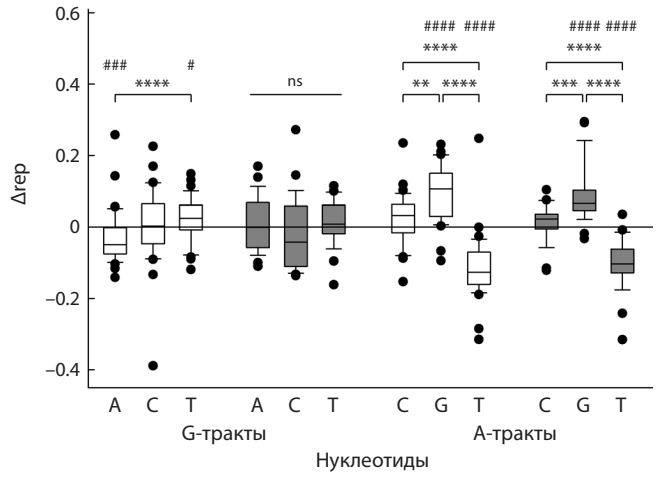


Рис. 5. Представленность 5′-фланговых нуклеотидов при полипуриновых трактах в геномах аэробных (белые прямоугольники) и анаэробных (серые прямоугольники) микроорганизмов.

Показано отклонение от ожидаемой представленности для G- и A-трактов любой длины. Отличие от ожидаемого: $^{\#}p < 0.05$, $^{\#\#\#}p < 0.005$, $^{\#\#\#\#}p < 0.001$ (одновыборочный критерий Уилкоксона с поправкой Данн на множественные сравнения). Различия между группами: $^{**}p < 0.01$, $^{***}p < 0.005$, $^{****}p < 0.001$ (критерий Краскела–Уоллиса с поправкой Данн на множественные сравнения); ns – нет достоверных отличий.

Значительно больший размер генома человека по сравнению с прокариотическими геномами позволил выявить интересные закономерности распределения размеров G_n-трактов. При n = 2–8 их число убывало экспоненциально и описывалось коэффициентом прирастания $k_{inc} = -0.674$, очень близким к центру распределения значений k_{inc} для G-трактов у прокариот (ср. рис. 6, *а* и рис. 3, *в*; $z = 0.141$). При n = 9–16 экспоненциальная зависимость сохранялась, но скорость уменьшения числа повторов замедлялась: значение k_{inc} возрастало до −0.198, что лежит далеко за пределами диапазона значений k_{inc} для прокариотических геномов (ср. рис. 6, *а* и рис. 3, *в*; $z = 5.97$). Тракты такого размера в геномах прокариот отсутствовали или присутствовали в единичных случаях, поэтому выявить этот переход было невозможно. Дальнейший рост длины G-трактов сопровождался еще большим падением скорости уменьшения их числа (см. рис. 6, *а*). Очевидно, что в районе n = 8–9 (значение точки излома, определенное методом сегментированной регрессии: $n = 8.72 \pm 0.04$) происходит изменение баланса процессов укорочения и удлинения G-трактов в пользу последнего; рост тракта за счет проскальзывания ДНК-полимераз при репликации

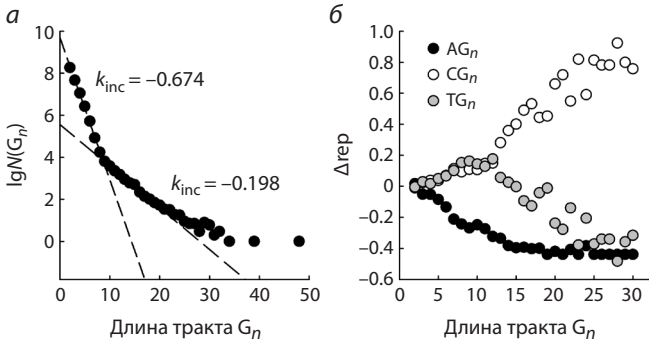


Рис. 6. Зависимость количества G-трактов и представленности 5′-фланговой нуклеотида в геноме человека от длины тракта.

а – зависимость числа G-трактов $N(G_n)$ от их длины. Штриховые линии соответствуют линейной регрессии, значения k_{inc} для n = 2–8 и n = 9–16 приведены на фрагменте *б* – зависимость Δrep 5′-флангового нуклеотида от длины тракта. Черные точки – фланговый нуклеотид A, белые точки – C, серые точки – T.

или репарации становится «самоподдерживающимся», как в хорошо изученном случае трактов тринуклеотидных повторов (Mirkin, 2007; McMurray, 2010).

Еще более неожиданная картина выявилась при анализе представленности 5′-фланговых нуклеотидов. Поскольку хорошо известно, что количество динуклеотидов CG в геноме человека снижено из-за их роли в эпигенетической регуляции (Fazzari, Greally, 2004), ожидаемую представленность рассчитывали исходя не из общего содержания нуклеотидов, а из содержания динуклеотидов. При n = 2 представленность нуклеотидов близко соответствовала ожидаемой, но далее значения Δrep для A стабильно падали, а представленность C и T, напротив, росла практически с одной и той же скоростью (см. рис. 6, *б*). Однако начиная с n = 8–11 (значение точки излома для Δrep(C)−Δrep(T), определенное методом сегментированной регрессии: $n = 9.28 \pm 1.10$) зависимости для C и T резко расходились: представленность T падала, а представленность C росла. Одно из возможных объяснений такого явления может заключаться в том, что с ростом длины G-тракты служат более эффективными ловушками для мигрирующих по ДНК-дуплексу дырок, за счет чего 8-oxoG в 5′-концевом положении доокисляется до гуанидиногидантоина и спироиминодигидантоина с соответствующим переключением преимущественных нуклеотидных замен с G→T на G→C.

Теломерную ДНК выделяют в отдельный класс высокоповторной части генома, представленный у человека множеством копий гексануклеотида TTAGGG. Известно, что теломерные повторы служат «горячими точками» окисления ДНК с образованием 8-oxoG (Billard, Poncet, 2019; Opresko et al., 2025). Концы теломер в клетках зародышевой линии удлиняются за счет фермента теломеразы, который в качестве матрицы использует теломеразную РНК, поэтому изменения в этих областях никак не связаны с повреждением геномной ДНК. Однако бóльшая часть теломер даже при наличии активной теломеразы реплицируется по обычному полуконсервативному механизму (Pfeiffer, Lingner, 2013; Higa et al., 2017; Bonnell et al., 2021), что может приводить к накоплению в них мутаций. Таким образом, конкретная последовательность теломер в соматических клетках человека (в случае T2T-генома – иммортализованной клеточной линии хорионаденомного происхождения CHM13hTERT, экспрессирующей теломеразу) отражает как их недавнее удлинение теломеразой в клетках зародышевой линии, так и события мутагенеза в прошлых поколениях и в индивидуальном онтогенезе.

Распределение повторов TTAGGG в хромосомах, рассчитанное по обеим цепям ДНК, имело стандартный вид с пиками частот на концах хромосом и провалом в околоцентромерном участке (рис. 7, *а*). Единственным исключением была хромосома 8, для которой в районе центромеры наблюдался, напротив, небольшой подъем числа этих повторов. На хромосоме 2 был отчетливо виден пик частоты теломерных повторов в области слияния двух предковых хромосом гоминид, образовавших эволюционно молодую хромосому 2 человека (Ijdo et al., 1991; см. рис. 7, *а*). Но при более детальном анализе этого участка видно, что он уже достаточно сильно деградировал и повторов TTAGGG в нем осталось гораздо меньше, чем в истинных теломерах (см. рис. 7, *б*). Интересно, что аналогичные пики обнаруживались на хромосомах 15 и 22 в интронах активных белок-кодирующих генов *ATP10A* и *MICAL3*; возможно, они представляют собой остатки событий транслокации части теломеры.

Повторы TTAAGG, TTACGG и TTATGG были распределены по хромосомам без теломерных пиков. Общая частота повторов TTACGG при этом была заметно ниже, чем TTAAGG и TTATGG, что согласуется со сниженным количеством динуклеотидов CG в геноме человека (см. рис. 7, *а*). Отдельные пики частоты повторов наблюдались на хромосоме 2 для TTAAGG, хромосомах 8, 12, 17 и Y для TTACGG, хромосомах 4 и 22 для TTATGG (см. примеры на рис. 7, *а*). Характерная картина распределения повторов в околоцентромерной области с провалами во всех вариантах TTANGG имела место для хромосом 1–5, 7, 10–12, 16, 19 и 21. В других случаях в районе центромеры преобладал один тип повторов, а другие были обеднены, причем их суммарный недостаток компенсировал избыток преобладающего типа, как показано на рис. 7, *а* для хромосомы 6. В хромосомах 6, 13–15, 22 и X в качестве преобладающего повтора выступал TTATGG, в хромосоме 8 – TTAGGG, в хромосоме 17 – TTACGG. Хромосома 18 отличалась совпадающими пиками распределения двух повторов – TTACGG и TTATGG (см. рис. 7, *а*). В длинном плече хромосомы 9 в области конститутивного гетерохроматина, прилегающей к околоцентромерному участку с избытком TTATGG, находился длинный участок с преобладанием TTACGG.

Очевидно, что описанные выше случаи совместной или взаимно противофазной локализации повторов TTANGG в нетеломерных областях объясняются не точковыми мутациями в повторе TTAGGG, а наличием в этих локусах большого числа повторяющихся элементов, содержащих какой-то один или два из этих гексануклеотидов. Напротив, особенности точковых мутаций в первой позиции G_3-тракта теломерного повтора должны быть наиболее очевидны в областях, состоящих по большей части из TTAGGG, т. е. в самих теломерах и внутрихромосомных блоках теломероподобных повторов. Для анализа частоты

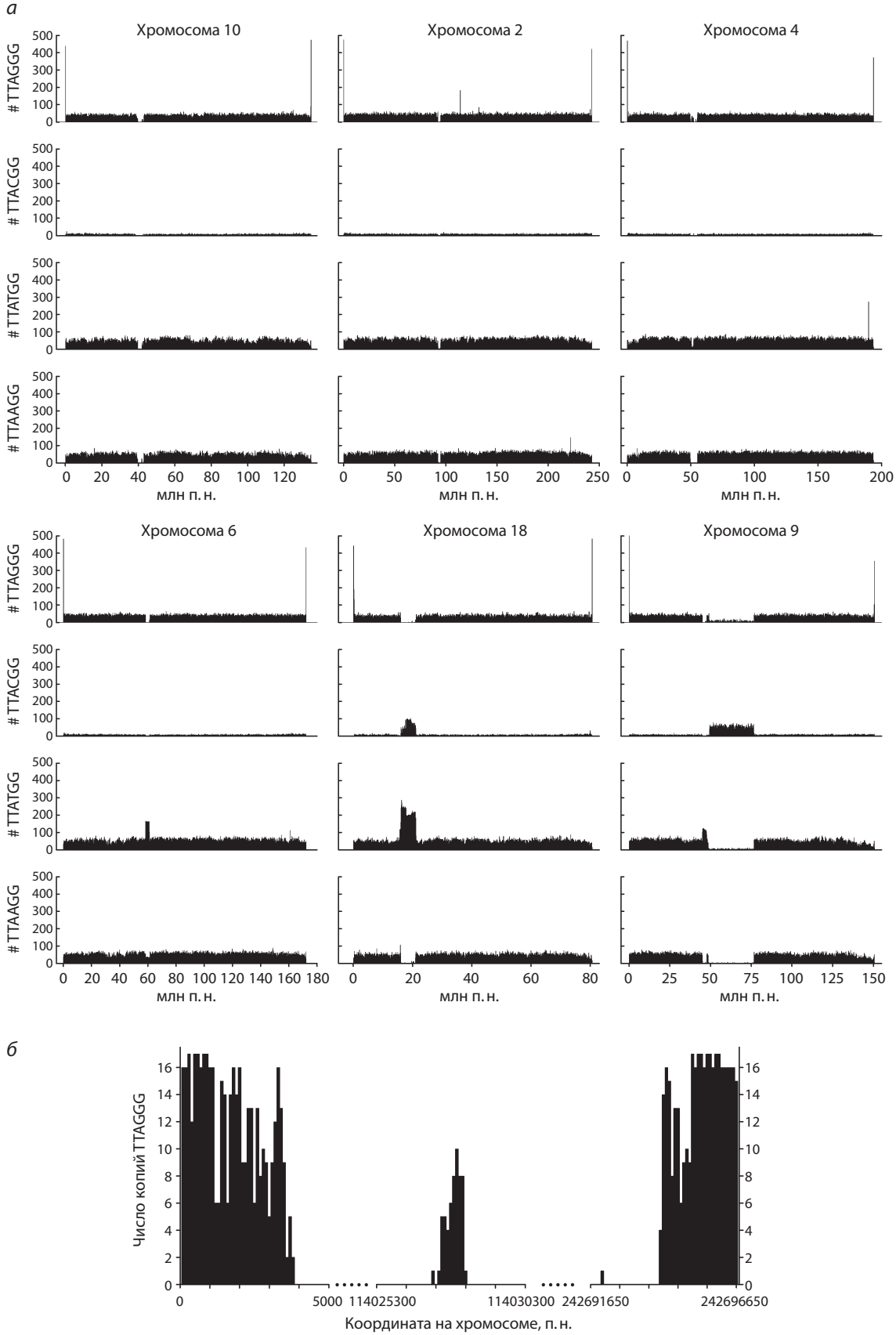


Рис. 7. Примеры распределения повторов TTAGGG, TTACGG, TTATGG и TTAAGG в хромосомах человека.

а – распределение повторов по всей длине хромосом 10, 2, 4, 6, 18 и 9. Количество повторов рассчитано в интервалах размером 100 т. п. н.; *б* – распределение повторов TTAGGG в теломерных участках и в области слияния предковых теломер хромосомы 2. Количество повторов рассчитано в интервалах размером 100 п. н.

Таблица 2. Представленность гексануклеотидов TTANGG в теломерах и внутрихромосомных блоках теломероподобных повторов

Число копий TTAGGG в 100 нт	TTAGGG	TTAAGG	TTACGG	TTATGG	$\chi^2_{(A=C=T)}$*	$\chi^2_{(A=T)}$**
Теломеры						
17	7361	0	0	0	–	–
16	6512	3	0	0	0.0498	0.0833
15	1935	9	1	1	0.00297	0.0114
14	980	26	0	3	8.12×10^{-10}	1.95×10^{-5}
13	715	23	1	2	1.85×10^{-8}	2.67×10^{-5}
12	468	14	2	2	3.35×10^{-4}	0.00270
11	341	6	0	0	0.00248	0.0143
10	490	16	1	0	6.97×10^{-7}	6.33×10^{-5}
9	327	8	0	0	3.35×10^{-4}	0.00468
Внутрихромосомные участки						
16–17	16	0	0	0	–	–
15	45	0	1	0	0.368	–
11–14	221	0	0	0	–	–
10	90	2	1	2	0.818	1.000
9	99	10	0	4	0.00439	0.109
Объединенная статистика						
17	7361	0	0	0	–	–
16	6528	3	0	0	0.0498	0.0833
15	1980	9	2	1	0.00865	0.0114
14	1022	26	0	3	8.12×10^{-10}	1.95×10^{-5}
13	767	23	1	2	1.85×10^{-8}	2.67×10^{-5}
12	540	14	2	2	3.35×10^{-4}	0.00270
11	396	6	0	0	0.00248	0.0143
10	580	18	2	2	8.84×10^{-6}	3.35×10^{-4}
9	423	18	0	4	5.12×10^{-6}	0.00284

* Значения критерия χ^2 для нулевой гипотезы равной представленности A, C и T.
** Значения критерия χ^2 для нулевой гипотезы равной представленности A и T.

замен в таких участках были выделены собственно теломерные области, а также все внутрихромосомные блоки, где в интервалах размером 100 т. п. н. встречалось не менее 100 копий повтора TTAGGG. Они были разбиты на более короткие интервалы длиной 100 п. н. Заполнение такого интервала только повторами TTAGGG соответствует 16 или 17 копиям (в зависимости от позиции первого полного гексануклеотида в интервале). Для анализа были взяты интервалы, в которых число копий TTAGGG было не менее 9, что составляет более половины от длины интервала.

Подсчет встречаемости гексануклеотидов TTAAGG, TTACGG и TTATGG в исследуемых участках выявил однозначную картину значительного обогащения по заменам G→A в первой позиции G-тракта по сравнению с заменами G→C и G→T (табл. 2). Суммарное число последних было в пять раз ниже, а частоты достоверно не различались между собой. Таким образом, несмотря на то, что теломерные повторы служат преимущественными местами окисления гуанина, это не фиксируется в повышенной частоте появления точковых мутаций G→T. Разница между представленностью A и C+T в 5′-фланговой позиции при динуклеотиде GG между теломерными повторами и остальной частью генома может свидетельствовать о существовании в теломерах мутационного процесса, отличного от окисления G в 5′-положении GGG.

Заключение

Таким образом, анализ нуклеотидного контекста G-трактов в наборе из 62 полных геномов прокариот и в T2T-геноме человека показал, что представленность T в положении, примыкающем к G-трактам, в целом повышена, что согласуется с моделью окисления G в 5′-положении трактов с последующими мутациями G→T. Одновременно с этим выявлены и другие закономерности распределения 5′-фланговых нуклеотидов: неравномерность частоты нуклеотидов в положении, примыкающем к A-трактам, повышенная представленность C с 5′-стороны от длинных G-трактов в геноме человека, преобладание A при заменах

G в 5′-положении в теломерных повторах. Гипотеза о том, что удлинение G-тракта может вести к замене специфичности однонуклеотидных мутаций с G→T на G→C из-за изменения природы повреждения-предшественника может быть проверена экспериментально. Характерный спектр мутаций в теломерных повторах может быть вызван их тенденцией складываться в G-квадруплексные структуры, которые затрудняют движение ДНК-полимераз (Pfeiffer, Lingner, 2013; Higa et al., 2017; Bonnell et al., 2021), однако это предположение требует детального изучения точности действия репликативных ДНК-полимераз человека на неповрежденных и поврежденных матрицах такого строения. Для A-трактов неизвестно о существовании преимущественных положений повреждения ДНК; учитывая, что по сравнению с G-трактами A-тракты более длинные (см. рис. 2), разница в относительной представленности C, G и T в 5′-фланговом положении может быть не связана с мутационным процессом. Объяснение всех этих выявленных закономерностей требует дальнейших исследований.

Список литературы / References

Adhikary A., Khanduri D., Sevilla M.D. Direct observation of the hole protonation state and hole localization site in DNA-oligomers. *J Am Chem Soc.* 2009;131(24):8614-8619. doi 10.1021/ja9014869

Alexandrov L.B., Nik-Zainal S., Wedge D.C., Aparicio S.A.J.R., Behjati S., Biankin A.V., Bignell G.R., … Campo E., Shibata T., Pfister S.M., Campbell P.J., Stratton M.R. Signatures of mutational processes in human cancer. *Nature.* 2013;500(7463):415-421. doi 10.1038/nature12477

Bansal A., Kaushik S., Kukreti S. Non-canonical DNA structures: diversity and disease association. *Front Genet.* 2022;13:959258. doi 10.3389/fgene.2022.959258

Billard P., Poncet D.A. Replication stress at telomeric and mitochondrial DNA: common origins and consequences on ageing. *Int J Mol Sci.* 2019;20(19):4959. doi 10.3390/ijms20194959

Bonnell E., Pasquier E., Wellinger R.J. Telomere replication: solving multiple end replication problems. *Front Cell Dev Biol.* 2021;9:668171. doi 10.3389/fcell.2021.668171

Cadet J., Douki T., Ravanat J.-L. Oxidatively generated damage to the guanine moiety of DNA: mechanistic aspects and formation in cells. *Acc Chem Res.* 2008;41(8):1075-1083. doi 10.1021/ar700245e

Cadet J., Davies K.J.A., Medeiros M.H.G., Di Mascio P., Wagner J.R. Formation and repair of oxidatively generated damage in cellular DNA. *Free Radic Biol Med.* 2017;107:13-34. doi 10.1016/j.freeradbiomed.2016.12.049

Chatterjee N., Walker G.C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen.* 2017;58(5):235-263. doi 10.1002/em.22087

Chiorcea-Paquim A.-M. 8-oxoguanine and 8-oxodeoxyguanosine biomarkers of oxidative DNA damage: a review on HPLC–ECD determination. *Molecules.* 2022;27(5):1620. doi 10.3390/molecules27051620

Cho B.P., Kadlubar F.F., Culp S.J., Evans F.E. ^{15}N nuclear magnetic resonance studies on the tautomerism of 8-hydroxy-2′-deoxyguanosine, 8-hydroxyguanosine, and other C8-substituted guanine nucleosides. *Chem Res Toxicol.* 1990;3(5):445-452. doi 10.1021/tx00017a010

Dizdaroglu M., Coskun E., Jaruga P. Measurement of oxidatively induced DNA damage and its repair, by mass spectrometric techniques. *Free Radic Res.* 2015;49(5):525-548. doi 10.3109/10715762.2015.1014814

ESCODD (European Standards Committee on Oxidative DNA Damage), Gedik C.M., Collins A. Establishing the background level of base oxidation in human lymphocyte DNA: results of an interlaboratory validation study. *FASEB J.* 2005;19(1):82-84. doi 10.1096/fj.04-1767fje

Fazzari M.J., Greally J.M. Epigenomics: beyond CpG islands. *Nat Rev Genet.* 2004;5(6):446-455. doi 10.1038/nrg1349

Fleming A.M., Burrows C.J. Formation and processing of DNA damage substrates for the hNEIL enzymes. *Free Radic Biol Med.* 2017;107:35-52. doi 10.1016/j.freeradbiomed.2016.11.030

Fleming A.M., Burrows C.J. Chemistry of ROS-mediated oxidation to the guanine base in DNA and its biological consequences. *Int J Radiat Biol.* 2022;98(3):452-460. doi 10.1080/09553002.2021.2003464

Genereux J.C., Barton J.K. Mechanisms for DNA charge transport. *Chem Rev.* 2010;110(3):1642-1662. doi 10.1021/cr900228f

Giese B. Long-distance electron transfer through DNA. *Annu Rev Biochem.* 2002;71:51-70. doi 10.1146/annurev.biochem.71.083101.134037

Halliwell B., Gutteridge J.M.C. Free Radicals in Biology and Medicine. Oxford Univ. Press, 2015

Henikoff S., Henikoff J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(22):10915-10919. doi 10.1073/pnas.89.22.10915

Higa M., Fujita M., Yoshida K. DNA replication origins and fork progression at mammalian telomeres. *Genes.* 2017;8(4):112. doi 10.3390/genes8040112

Ijdo J.W., Baldini A., Ward D.C., Reeders S.T., Wells R.A. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc Natl Acad Sci USA.* 1991;88(20):9051-9055. doi 10.1073/pnas.88.20.9051

Kino K., Kawada T., Hirao-Suzuki M., Morikawa M., Miyazawa H. Products of oxidative guanine damage form base pairs with guanine. *Int J Mol Sci.* 2020;21(20):7645. doi 10.3390/ijms21207645

Koh G., Degasperi A., Zou X., Momen S., Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer.* 2021;21(10):619-637. doi 10.1038/s41568-021-00377-7

Kouchakdjian M., Bodepudi V., Shibutani S., Eisenberg M., Johnson F., Grollman A.P., Patel D.J. NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7*H*-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7*H*-dG(*syn*)·dA(*anti*) alignment at lesion site. *Biochemistry.* 1991;30(5):1403-1412. doi 10.1021/bi00219a034

Kucab J.E., Zou X., Morganella S., Joel M., Nanda A.S., Nagy E., Gomez C., Degasperi A., Harris R., Jackson S.P., Arlt V.M., Phillips D.H., Nik-Zainal S. A compendium of mutational signatures of environmental agents. *Cell.* 2019;177(4):821-836.e816. doi 10.1016/j.cell.2019.03.001

Kunkel T.A., Bebenek K. DNA replication fidelity. *Annu Rev Biochem.* 2000;69:497-529. doi 10.1146/annurev.biochem.69.1.497

Kurbanyan K., Nguyen K.L., To P., Rivas E.V., Lueras A.M.K., Kosinski C., Steryo M., González A., Mah D.A., Stemp E.D.A. DNA-protein cross-linking via guanine oxidation: dependence upon protein and photosensitizer. *Biochemistry.* 2003;42(34):10269-10281. doi 10.1021/bi020713p

Liao X., Zhu W., Zhou J., Li H., Xu X., Zhang B., Gao X. Repetitive DNA sequence detection and its role in the human genome. *Commun Biol.* 2023;6:954. doi 10.1038/s42003-023-05322-y

Lipscomb L.A., Peek M.E., Morningstar M.L., Verghis S.M., Miller E.M., Rich A., Essigmann J.M., Williams L.D. X-ray structure of a DNA decamer containing 7,8-dihydro-8-oxoguanine. *Proc Natl Acad Sci USA.* 1995;92(3):719-723. doi 10.1073/pnas.92.3.719

Liu B., Xue Q., Tang Y., Cao J., Guengerich F.P., Zhang H. Mechanisms of mutagenesis: DNA replication in the presence of DNA damage. *Mutat Res.* 2016;768:53-67. doi 10.1016/j.mrrev.2016.03.006

Livingstone C.D., Barton G.J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput Appl Biosci.* 1993;9(6):745-756. doi 10.1093/bioinformatics/9.6.745

Maga G., Villani G., Crespan E., Wimmer U., Ferrari E., Bertocci B., Hübscher U. 8-oxo-guanine bypass by human DNA polymerases in

the presence of auxiliary proteins. *Nature.* 2007;447(7144):606-608. doi 10.1038/nature05843

McAuley-Hecht K.E., Leonard G.A., Gibson N.J., Thomson J.B., Watson W.P., Hunter W.N., Brown T. Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs. *Biochemistry.* 1994;33(34):10266-10270. doi 10.1021/bi00200a006

McMurray C.T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet.* 2010;11(11):786-799. doi 10.1038/nrg2828

Miller H., Grollman A.P. Kinetics of DNA polymerase I (Klenow fragment exo–) activity on damaged DNA templates: effect of proximal and distal template damage on DNA synthesis. *Biochemistry.* 1997; 36(49):15336-15342. doi 10.1021/bi971927n

Mirkin S.M. Expandable DNA repeats and human disease. *Nature.* 2007;447(7147):932-940. doi 10.1038/nature05977

Moriya M. Single-stranded shuttle phagemid for mutagenesis studies in mammalian cells: 8-oxoguanine in DNA induces targeted G·C→T·A transversions in simian kidney cells. *Proc Natl Acad Sci USA.* 1993;90(3):1122-1126. doi 10.1073/pnas.90.3.1122

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., … Zook J.M., Schatz M.C., Eichler E.E., Miga K.H., Phillippy A.M. The complete sequence of a human genome. *Science.* 2022;376(6588):44-53. doi 10.1126/science.abj6987

Okonechnikov K., Golosova O., Fursov M.; UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics.* 2012; 28(8):1166-1167. doi 10.1093/bioinformatics/bts091

O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., … Tatusova T., DiCuccio M., Kitts P., Murphy T.D., Pruitt K.D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44(D1):D733-D745. doi 10.1093/nar/gkv1189

Opresko P.L., Sanford S.L., De Rosa M. Oxidative stress and DNA damage at telomeres. *Cold Spring Harb Perspect Biol.* 2025;17(6): a041707. doi 10.1101/cshperspect.a041707

Pfeiffer V., Lingner J. Replication of telomeres and the regulation of telomerase. *Cold Spring Harb Perspect Biol.* 2013;5(5):a010405. doi 10.1101/cshperspect.a010405

Pilati C., Shinde J., Alexandrov L.B., Assié G., André T., Hélias-Rodzewicz Z., Ducoudray R., Le Corre D., Zucman-Rossi J., Emile J.-F., Bertherat J., Letouzé E., Laurent-Puig P. Mutational signature analysis identifies *MUTYH* deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol.* 2017;242(1):10-15. doi 10.1002/path.4880

Prorok P., Grin I.R., Matkarimov B.T., Ishchenko A.A., Laval J., Zharkov D.O., Saparbaev M. Evolutionary origins of DNA repair pathways: role of oxygen catastrophe in the emergence of DNA glycosylases. *Cells.* 2021;10(7):1591. doi 10.3390/cells10071591

Richard G.-F., Kerrest A., Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev.* 2008;72(4):686-727. doi 10.1128/MMBR.00011-08

Saito I., Nakamura T., Nakatani K., Yoshioka Y., Yamaguchi K., Sugiyama H. Mapping of the hot spots for DNA damage by one-electron oxidation: efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. *J Am Chem Soc.* 1998;120(48):12686-12687. doi 10.1021/ja981888i

Shibutani S., Takeshita M., Grollman A.P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature.* 1991;349(6308):431-434. doi 10.1038/349431a0

Sugiyama H., Saito I. Theoretical studies of GG-specific photocleavage of DNA via electron transfer: significant lowering of ionization potential and 5′-localization of HOMO of stacked GG bases in B-form DNA. *J Am Chem Soc.* 1996;118(30):7063-7068. doi 10.1021/ja9609821

Taylor W.R. The classification of amino acid conservation. *J Theor Biol.* 1986;119(2):205-218. doi 10.1016/S0022-5193(86)80075-3

Tubbs A., Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell.* 2017;168(4):644-656. doi 10.1016/j.cell.2017.01.002

Viel A., Bruselles A., Meccia E., Fornasarig M., Quaia M., Canzonieri V., Policicchio E., … Maestro R., Giannini G., Tartaglia M., Alexandrov L.B., Bignami M. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine*. 2017;20:39-49. doi 10.1016/j.ebiom.2017.04.022

Wood M.L., Esteve A., Morningstar M.L., Kuziemko G.M., Essigmann J.M. Genetic effects of oxidative DNA damage: comparative mutagenesis of 7,8-dihydro-8-oxoguanine and 7,8-dihydro-8-oxoadenine in *Escherichia coli*. *Nucleic Acids Res.* 1992;20(22): 6023-6032. doi 10.1093/nar/20.22.6023

Yudkina A.V., Shilkin E.S., Endutkin A.V., Makarova A.V., Zharkov D.O. Reading and misreading 8-oxoguanine, a paradigmatic ambiguous nucleobase. *Crystals.* 2019;9(5):269. doi 10.3390/cryst9050269

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 03.08.2025. После доработки 01.10.2025. Принята к публикации 17.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):925-939

doi 10.18699/vjgb-25-99

Связь иерархической классификации транскрипционных факторов по структуре ДНК-связывающего домена и вариабельности мотивов сайтов связывания этих факторов

В.Г. Левицкий [1, 2], Т.Ю. Ватолина[2], В.В. Радица[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
levitsky@bionet.nsc.ru

Аннотация. Поиск мотивов *de novo* – базовый подход определения нуклеотидной специфичности связывания важнейших регуляторов транскрипции генов, транскрипционных факторов (ТФ), на основе данных массового полногеномного секвенирования районов их сайтов связывания *in vivo*, таких как ChIP-seq. Количество известных мотивов сайтов связывания ТФ (ССТФ) возросло в несколько раз в последние годы. Из-за сходства структуры ДНК-связывающих доменов ТФ многие структурно родственные ТФ имеют сходные или даже неразличимые мотивы сайтов связывания. Классификация ТФ по структуре ДНК-связывающих доменов из базы данных TFClass определяет верхние уровни иерархии (суперклассы и классы ТФ) по структуре этих доменов, а следующие уровни (семейства и подсемейства ТФ) по выравниваниям аминокислотных последовательностей доменов. Однако эта классификация не учитывает сходство мотивов ССТФ, а для идентификации действующих ТФ по данным массового секвенирования ССТФ ChIP-seq приходится иметь дело с мотивами ССТФ, а не с самими ТФ. Поэтому в данной работе мы взяли из баз данных Hocomoco/Jaspar мотивы ССТФ человека/плодовой мушки *Drosophila melanogaster* и рассмотрели сходство мотивов сайтов связывания в парах родственных ТФ согласно их классификации в базе данных TFClass. Показано, что общее дерево иерархии ТФ по структуре ДНК-связывающих доменов можно разделить на отдельные неперекрывающиеся множества ТФ – ветви. В пределах каждой ветви большинство пар ТФ имеет значимо похожие мотивы сайтов связывания. Каждая ветвь включает одну или несколько сестринских элементарных единиц иерархии и все более низкие ее/их уровни: один или несколько ТФ одного подсемейства или целое подсемейство, одно или несколько подсемейств одного семейства, целое семейство и т.д. до целого класса. Анализ семи крупнейших классов ТФ человека и двух плодовой мушки показал, что сходство ТФ по мотивам ССТФ для разных соответствующих уровней (классов, семейств) заметно отличается. Дополнение иерархической классификации ТФ ветвями, объединяющими значимо сходные мотивы ССТФ, может повысить эффективность идентификации ТФ, вовлеченных в регуляцию транскрипции, по результатам *de novo* поиска обогащенных мотивов для данных массового секвенирования ССТФ с помощью технологии ChIP-seq.

Ключевые слова: *de novo* поиск мотивов; мотивы сайтов связывания транскрипционных факторов; структурные варианты мотивов сайтов связывания транскрипционных факторов; сходство мотивов сайтов связывания транскрипционных факторов; кооперативное действие транскрипционных факторов; массовое полногеномное секвенирование сайтов связывания транскрипционных факторов

Для цитирования: Левицкий В.Г., Ватолина Т.Ю., Радица В.В. Связь иерархической классификации транскрипционных факторов по структуре ДНК-связывающего домена и вариабельности мотивов сайтов связывания этих факторов. *Вавиловский журнал генетики и селекции*. 2025;29(7):925-939. doi 10.18699/vjgb-25-99

Финансирование. Работа поддержана проектом Российского научного фонда, проект № 24-14-00133.

Благодарности. Разработка программного пакета и анализ данных проведены с использованием вычислительных ресурсов ЦКП «Биоинформатика» при поддержке бюджетного проекта № FWNR-2022-0020.

Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs

V.G. Levitsky [1, 2], T.Yu. Vatolina[2], V.V. Raditsa[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
levitsky@bionet.nsc.ru

© Левицкий В.Г., Ватолина Т.Ю., Радица В.В., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. *De novo* motif search is the main approach for determining the nucleotide specificity of binding of the key regulators of gene transcription, transcription factors (TFs), based on data from massive genome-wide sequencing of their binding site regions *in vivo*, such as ChIP-seq. The number of motifs of known TF binding sites (TFBSs) has increased several times in recent years. Due to the similarity in the structure of the DNA-binding domains of TFs, many structurally cognate TFs have similar and sometimes almost indistinguishable binding site motifs. The classification of TFs by the structure of the DNA-binding domains from the TFClass database defines the top levels of the hierarchy (superclasses and classes of TFs) by the structure of these domains, and the next levels (families and subfamilies of TFs) by the alignments of amino acid sequences of domains. However, this classification does not take into account the similarity of TFBS motifs, whereas identification of valid TFs from massive sequencing data of TFBSs, such as ChIP-seq, requires working with TFBS motifs rather than TFs themselves. Therefore, in this study we extracted from the Hocomoco and Jaspar databases the TFBS motifs for human and fruit fly *Drosophila melanogaster*, and considered the pairwise similarity of binding site motifs of cognate TFs according to their classification from the TFClass database. We have shown that the common tree of the TF hierarchy by the structure of DNA-binding domains can be split into separate branches representing non-overlapping sets of TFs. Within each branch, the majority of TF pairs have significantly similar binding site motifs. Each branch can include one or more sister elementary units of the hierarchy and all its/their lower levels: one or more TFs of the same subfamily, or the whole subfamily, one or several subfamilies of the same family, an entire family, etc., up to the entire class. Analysis of the seven largest human and two largest Drosophila TF classes showed that the similarity of TFs in terms of TFBS motifs for different corresponding levels (classes, families) is noticeably different. Supplementing the hierarchical classification of TFs with branches combining significantly similar motifs of TFBSs can increase the efficiency of identifying involved TFs through enriched motifs detected by *de novo* motif search for massive sequencing data of TFBSs from the ChIP-seq technology.

Key words: *de novo* motif search; motifs of transcription factor binding sites; structural variants of motifs of transcription factor binding sites; similarity of motifs of transcription factor binding sites; cooperative action of transcription factors; massive whole-genome sequencing of transcription factor binding sites

For citation: Levitsky V.G., Vatolina T.Yu., Raditsa V.V. Linking hierarchical classification of transcription factors by the structure of their DNA-binding domains to the variability of their binding site motifs. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):925-939. doi 10.18699/vjgb-25-99

Введение

Исследование механизмов регуляции транскрипции генов эукариот необходимо для понимания молекулярно-генетических процессов в клетке. Транскрипция генов производится под контролем специальных белков, транскрипционных факторов (ТФ), регулирующих ее за счет специфического по нуклеотидному контексту связывания с геномной ДНК (Lambert et al., 2018). Эта специфичность обусловлена нуклеотидными последовательностями сайтов связывания, распознаваемых отдельными ТФ (ССТФ). Изменчивость сайтов связывания отражает способность каждого ТФ связываться с разными последовательностями ДНК, поэтому совокупность сходных последовательностей сайтов связывания, с которыми взаимодействует ТФ, называют мотивом его сайтов связывания (D'haeseleer, 2006). Длина участка геномной ДНК, непосредственно взаимодействующего с отдельным ТФ, а также длина мотива ССТФ обычно варьируют от 6 до 20 пар оснований (п. о.) (Spitz, Furlong, 2012; Zambelli et al., 2013; Vorontsov et al., 2024). Один ТФ может иметь несколько различных по структуре мотивов сайтов связывания. Самой популярной моделью мотива ССТФ является позиционная весовая матрица (ПВМ). Для построения модели мотива ПВМ необходимо по данному выравниванию ССТФ, представляющему этот мотив, подсчитать частоты встреч нуклеотидов в каждой из позиций, и по этим частотам для каждого из четырех нуклеотидов в каждой позиции подсчитать их вклады в общую оценку аффинности (или вес). Общая оценка аффинности потенциального сайта в последовательности ДНК равна сумме весов, соответствующих встречаемым нуклеотидам, по всем его позициям (Wasserman, Sandelin, 2004).

Экспериментальная технология ChIP-seq основана на иммунопреципитации хроматина (ChIP), то есть применении антител к исследуемому целевому белку, например ТФ. Эта технология используется для выявления взаимодействий целевых белков с геномной ДНК *in vivo*. Суть этой технологии заключается в проведении иммунопреципитации хроматина и последующем картировании геномных локусов, с которыми взаимодействует целевой белок. Транскрипционные факторы *in vivo*, как правило, действуют в составе мультибелковых комплексов, сформированных белок-белковыми взаимодействиями нескольких ТФ, что позволяет этим ТФ совместно регулировать транскрипцию генов, даже без прямых связей каждого из ТФ с геномной ДНК. Поэтому ТФ *in vivo* могут связываться с ДНК разными способами:

- напрямую, в ДНК есть сайт связывания целевого ТФ;
- при помощи другого «партнерского» ТФ, в ДНК есть сайты связывания целевого и партнерского ТФ, встречаются совместно (рядом), с некоторым спейсером или перекрываются (Levitsky et al., 2019);
- не напрямую, в ДНК есть сайт связывания партнерского ТФ и нет сайта связывания целевого ТФ (Slattery et al., 2014).

Картированные в эксперименте ChIP-seq отдельные локусы генома называются пиками и имеют длину от нескольких сотен до тысяч п. о. (Johnson et al., 2007, Nakato, Shirahige, 2017; Lloyd, Bao, 2019). Каждый из пиков не обязательно содержит сайт связывания целевого ТФ, прямое связывание может совершать один из возможных партнерских ТФ. Массовое применение других экспериментальных технологий секвенирования *in vivo* помимо ChIP-seq, например CUT&RUN (Sken, Henikoff, 2017),

а также технологий *in vitro* (PBM, HT-SELEX) (Stormo, Zhao, 2010; Jolma et al., 2013; Franco-Zorrilla et al., 2014) дало возможность накопления данных о нуклеотидной специфичности сайтов связывания сотен ТФ основных модельных видов эукариот. Созданы базы данных (БД), нацеленные на единообразную первичную обработку данных массового полногеномного секвенирования ССТФ, включающих и данные ChIP-seq (GTRD, Kolmykov et al., 2021; ReMap, Hammal et al., 2022; Cistrome DB, Taing et al., 2024).

Анализ обогащения мотивов ССТФ, в частности поиск мотивов *de novo* (Zambelli et al., 2013; Liu et al., 2018; Bailey, 2021), изначально использовался для подтверждения корректности данных эксперимента ChIP-seq (набора последовательностей ДНК или пиков). Затем поиск мотивов *de novo* стал стандартным подходом анализа набора пиков, позволяющим определить обогащенные мотивы, предположительно соответствующие мотивам сайтов связывания целевого ТФ и нескольких партнерских ТФ, кооперативно действующих в регуляции транскрипции генов (Spitz, Furlong, 2012; Slattery et al., 2014; Morgunova, Taipale, 2017).

К настоящему времени для нескольких сотен ТФ основных таксонов эукариот, таких как млекопитающие, насекомые и растения, мотивы ССТФ модели ПВМ (матрицы частот нуклеотидов) содержатся в ряде БД, JASPAR (Rauluseviciute et al., 2024), Hocomoco (Vorontsov et al., 2024) и Cis-BP (Weirauch et al., 2014). Например, версия 12 БД Hocomoco (Vorontsov et al., 2024) представляет 1443 мотива сайтов связывания для 949 ТФ человека. Особенности конвейера анализа, примененного в БД Hocomoco для мотивов ССТФ человека и мыши, позволили выявить более одного структурного типа мотива для нескольких сотен аннотированных транскрипционных факторов.

Для отдельно взятого ТФ как число разных мотивов сайтов связывания, так и структура, и изменчивость каждого из мотивов определяются структурой ДНК-связывающего домена (ДСД) этого ТФ (Wingender, 1997, 2013). На основе анализа сходства структуры ДСД ТФ и выравнивания аминокислотных последовательностей ДСД ТФ была разработана иерархическая классификация TFClass сначала для ТФ человека, а затем и для их ортологов у грызунов и млекопитающих (Wingender et al., 2013, 2015, 2018). Эта классификация определяет шесть уровней иерархии. Верхние уровни иерархии, суперкласс и класс, определены согласно общей топологии и структурным особенностям ДСД ТФ. Следующие ниже уровни, семейство и подсемейство, объединены по сходству аминокислотных последовательностей ДСД ТФ на основе их выравниваний. Нижние уровни – это ген ТФ и структурный вариант его белка. Всего у млекопитающих определено девять суперклассов. Анализ структуры ДСД ТФ растений не выявил дополнительных суперклассов, однако около половины классов ТФ оказались специфичными для растений (БД Plant-TFClass, Blanc-Mathieu et al., 2024).

Важнейшей функцией ТФ *in vivo* является способность их специфичного связывания с ДНК. Однако классификация TFClass не учитывает сходство мотивов ССТФ на отдельных уровнях иерархии, в конкретных классах, семействах и т. д. Сходство мотивов ССТФ может очень сильно варьировать в различных классах ТФ. Например, самый крупный класс ТФ млекопитающих, C2H2 zinc finger factors {2.3} отличается наиболее заметной изменчивостью мотивов ССТФ (Najafabadi et al., 2015; Lambert et al., 2018). Здесь и далее цифры в фигурных скобках обозначают номенклатуру классификации ТФ согласно TFClass (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018). Например, ТФ JUN относится к суперклассу Basic domains {1}, классу Basic leucine zipper factors (bZIP) {1.1}, семейству Jun-related {1.1.1} и подсемейству Jun {1.1.1.1}. Определение функционирующего ТФ по заданному обогащенному мотиву его сайтов связывания как результату поиска мотивов *de novo* может опираться не только на классификацию ТФ по структуре их ДСД, но и на их классификацию по сходству мотивов ССТФ.

Важным шагом в анализе результатов поиска обогащенных мотивов *de novo*, применяемого для данных ChIP-seq, является максимально точное указание целевых и партнерских ТФ по полученным обогащенным мотивам. Общепринятый способ ограничить список предполагаемых ТФ для каждого обогащенного мотива – это оценка значимости его сходства с мотивами сайтов связывания известных ТФ из БД мотивов ССТФ (Weirauch et al., 2014; Rauluseviciute et al., 2024; Vorontsov et al., 2024). Для оценки сходства в парах мотивов модели ПВМ можно использовать стандартные инструменты, например TomTom (Gupta et al., 2007).

Оценка общего числа ТФ человека составляет 1659 (Shen et al., 2023), однако как число структурно различных ДСД ТФ, так и число ТФ с разными мотивами сайтов связывания гораздо меньше, так как родственные по структуре ДСД ТФ обычно имеют схожие мотивы сайтов связывания (Ambrosini et al., 2020). Самым явным исключением из этого общего правила являются ТФ класса C2H2 zinc finger {2.3} (Lambert et al., 2018).

Наличие для одного ТФ двух и более структурно различных мотивов сайтов связывания широко распространено для разных классов ТФ (Vorontsov et al., 2024). Причиной этого может быть способность ТФ связываться только в составе димера из родственных ТФ (например, пары ТФ классов Basic helix-loop-helix factors (bHLH) {1.2}, или Basic leucine zipper factors (bZIP) {1.1}), или как димер или мономер (например, пара ТФ класса Nuclear receptors with C4 zinc fingers {2.1}) (Amoutzias et al., 2008). В целом, обычно мотивы сайтов связывания родственных ТФ из одного класса или семейства демонстрируют высокую или умеренную степень сходства, зависящую от положения класса, семейства или подсемейства в иерархии TFClass/Plant-TFClass. Вместе с тем даже среди мотивов сайтов связывания, принадлежащих к одному и тому же ТФ, может наблюдаться некоторое разнообразие структурных вариантов. Например, для ТФ CDX2 (класс Homeo domain factors {3.1}) и THB (класс Nuclear receptors with C4 zinc fingers {2.1}) в БД Hocomoco версии 12 есть соответственно два и четыре мотива. Два мотива ССТФ CDX2 значимо не сходны между собой (p-value > 0.001, Gupta et al., 2007) (рис. 1, *а*), такого значимого сходства также нет в трех из шести возможных пар, сформированных с участием четырех мотивов ССТФ THB (см. рис. 1, *б*, *в*). Можно предположить, что чаще именно семейства или

Рис. 1. Сходство разных мотивов сайтов связывания, представляющих отдельные ТФ.

а, *б* – лого двух/четырех мотивов ССТФ CDX2/THB из классов Homeo domain factors {3.1}/Nuclear receptors with C4 zinc fingers {2.1}. Для каждого мотива указан идентификатор БД Hocomoco (Vorontsov et al., 2024). Лого для модели мотива ПВМ представляет частоты встреч нуклеотидов в позициях высотами букв (Schneider, Stephens, 1990); *в* – оценки сходства четырех мотивов ССТФ THB, рассчитанные инструментом TomTom (Gupta et al., 2007), цвет отражает значимость сходства, $-\mathrm{Log}_{10}[p\text{-value}]$.

подсемейства, а не классы ТФ представляют значимо сходные мотивы (Nagy G., Nagy L., 2020; de Martin et al., 2021; Zenker et al., 2025). Этот вопрос мы подробнее изучаем в настоящей работе.

Важнейший этап анализа данных ChIP-seq, *de novo* поиск мотивов, выявляет список обогащенных мотивов для пиков ChIP-seq. Для модели мотива ПВМ каждый мотив – это матрица частот нуклеотидов, и необходимо определить список известных ТФ из БД, таких как Jaspar (Rauluseviciute et al., 2024), Hocomoco (Vorontsov et al., 2024) или Cis-BP (Weirauch et al., 2014), имеющих значимо похожие мотивы сайтов связывания известных ТФ. Однако, помимо того, что число мотивов сайтов связывания зависит от структуры ДСД ТФ, ТФ крайне неравномерно распределены по суперклассам, классам, и даже семействам. В наиболее полной для человека/мыши БД мотивов ССТФ (Hocomoco, версия 12, Vorontsov et al., 2024) пять крупнейших классов ТФ представляют около 75 % всех мотивов (1082 из 1443): C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1}, Basic helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1} и Basic leucine zipper factors (bZIP) {1.1}. Десять крупнейших классов – это около 90 % всех мотивов (1303 из 1443). Восемь крупнейших семейств ТФ из всего четырех классов представляют более 51 % (742 из 1443) всех мотивов ССТФ: More than 3 adjacent zinc fingers {2.3.3}, HOX-related {3.1.1}, Multiple dispersed zinc fingers {2.3.4}, Paired-related HD {3.1.3}, NK-related {3.1.2}, Three-zinc finger Kruppel-related {2.3.1}, Tal-related {1.2.3} и Ets-related {3.5.2}. Недавний анализ 1725 ТФ модельного растения *Arabidopsis thaliana* выявил около 40 % из них (686) с имеющимися мотивами ССТФ; включение в анализ мотивов сайтов связывания 92 ТФ других растений показало крайне ограниченный словарь из всего 74 отличных мотивов ССТФ растений (Zenker et al., 2025).

Очень часто обогащенный мотив из результатов поиска мотивов *de novo* имеет высокое сходство с мотивами сайтов связывания известных ТФ из одного или нескольких семейств одного класса, или даже целый класс попадает в список ТФ-кандидатов. В результате получается список из нескольких десятков ТФ, и выбрать среди них конкретный ТФ – не самая простая задача. Такие протяженные списки ТФ-кандидатов могут осложнить идентификацию ТФ, наиболее вероятно связанных с обогащенными мотивами. Однако эту сложность можно уменьшить проведением систематического анализа сходства мотивов сайтов связывания у ТФ, классифицированных по уровням иерархии БД TFClass. На сегодняшний день для родственных ТФ определенной структуры ДСД (класс, семейство и подсемейство) не определено, какой из этих уровней достаточен для идентификации набора ТФ со значимо схожими мотивами сайтов связывания. Для решения этой задачи нужно найти набор неких объединений (или ветвей) нескольких последовательных уровней иерархической классификации TFClass, для которых мотивы ССТФ являются значимо схожими. Такой подход способен дополнительно систематизировать иерархическую классификацию ТФ, адаптировать ее для применения к результатам *de novo* поиска мотивов. Полученная в результате уточненная иерархия ТФ будет отражать сходство ДСД ТФ и сходство мотивов ССТФ.

Мы предлагаем включить аннотацию ветвей сходных мотивов сайтов связывания известных ТФ в стандартный протокол поиска мотивов *de novo*, применяемый к результатам полногеномного картирования ССТФ *in vivo*, например, с помощью технологии ChIP-seq. Применение ветвей может заметно упростить анализ обогащенных мотивов ССТФ. Ветви ТФ привязывают общепринятые единицы иерархической классификации ТФ по ДСД, а именно суперклассы, классы, семейства, подсемейства (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018) к сходству мотивов ССТФ (Gupta et al., 2007).

Материалы и методы

Входные данные и параметры. Входными данными являются наборы мотивов ССТФ, каждый мотив представляется матрицей частот нуклеотидов, идентификатором и именем ТФ; для каждого ТФ указаны его суперкласс, класс, семейство и подсемейство (при наличии), согласно информации БД TFClass (Wingender et al., 2013, 2015, 2018). Мотивы ССТФ человека *Homo sapiens* и дрозофилы *Drosophila melanogaster* были извлечены из БД Hocomoco (версия 12, https://hocomoco.autosome.org/) (Vorontsov et al., 2024). и Jaspar https://jaspar.elixir.no/ (Rauluseviciute et al., 2024). Обе БД строят мотивы ССТФ на основе данных массового секвенирования *in vivo*, например ChIP-seq,

Наборы мотивов ССТФ из БД Hocomoco и Jaspar, использованные в анализе

Таксон: вид организма	Класс ТФ	Число мотивов	Число ТФ
Млекопитающие: *H. sapiens*	Basic leucine zipper factors (bZIP) {1.1}	86	47
	Basic helix-loop-helix factors (bHLH) {1.2}	115	76
	Nuclear receptors with C4 zinc fingers {2.1}	93	44
	C2H2 zinc finger factors {2.3}	479	373
	Homeo domain factors {3.1}	309	184
	Fork head/winged helix factors {3.3}	65	43
	Tryptophan cluster factors {3.5}	67	38
	Всего	1214	805
Насекомые: *D. melanogaster*	C2H2 zinc finger factors {2.3}	79	57
	Homeo domain factors {3.1}	106	90
	Всего	185	147

и *in vitro*, например HT-SELEX. Мотивы ССТФ представляют собой матрицы частот нуклеотидов, что соответствует традиционной модели ПВМ. В обеих БД применена классификация ТФ по структуре ДСД по уровням иерархии суперкласс, класс, семейство, подсемейство и ТФ (БД TFClass, Wingender, 2013; Wingender et al., 2013, 2015, 2018). В анализ мы взяли классы как минимум с 50 мотивами ССТФ: семь/два класса ТФ человека/плодовой мушки, см. таблицу.

Метрика сходства двух транскрипционных факторов. Для оценки значимости *p*-value сходства в парах мотивов ССТФ был использован инструмент TomTom (Gupta et al., 2007), параметр функции сравнения мотивов – коэффициент корреляции Пирсона. Два мотива ССТФ считались похожими, если значимость достигает порога $-\mathrm{Log}_{10}[p\text{-value}] > \mathrm{Thr} = 3$.

Определим метрику сходства для пары ТФ по их мотивам сайтов связывания на основе распределения сходства во всех возможных парах мотивов сайтов связывания одного и другого ТФ, так как ТФ может иметь один и более мотив сайтов связывания. Пусть два ТФ X/Y имеют N_X/N_Y мотивов, $\{M_i\}$, $1 \le i \le N_X$ и $\{M_j\}$, $1 \le j \le N_Y$ соответственно. Распределение оценок сходства в паре этих ТФ по их мотивам сайтов связывания включает $N_X \times N_Y$ пар мотивов. Пусть сходство $\mathrm{Score}(M_i, M_j)$ мотивов M_i и M_j дается оценкой инструмента TomTom (Gupta et al., 2007) как логарифм значимости *p*-value:

$$\mathrm{Score}(M_i, M_j) = -\mathrm{Log}_{10}[p\text{-value}(M_i, M_j)]. \quad (1)$$

Тогда для двух ТФ X и Y метрика сходства $\mathrm{Score}_{X,Y}$ рассчитывается по формуле:

$$\mathrm{Score}_{X,Y} = \max_{1\le i\le N_X,\ 1\le j\le N_Y} \{\mathrm{Score}(M_i, M_j)\}. \quad (2)$$

Если оценка сходства $\mathrm{Score}_{X,Y}$ (2) превышает заданный порог Thr, то ТФ X и Y можно считать значимо сходными по их мотивам сайтов связывания. Для одного ТФ гетерогенность мотивов сайтов связывания оценивается как медиана или второй квартиль (Q2) распределения по всем возможным парам мотивов сайтов связывания этого ТФ:

$$\mathrm{Score}_X = \mathrm{Median}_{1\le i<N_X,\ i<j\le N_X} \{\mathrm{Score}(M_i, M_j)\}. \quad (3)$$

Метрика сходства двух наборов транскрипционных факторов. Пусть некоторый класс имеет семейство A с N_A ТФ. Распределение всех возможных в семействе пар ТФ включает $N_A \times (N_A - 1)/2$ варианта. Пусть в этом же классе семейство B имеет N_B ТФ. Распределение всех возможных в пар ТФ семейств A и B включает $N_A \times N_B$ варианта. Как для случая внутри семейства, так и для случая между семействами, для всех пар ТФ рассчитываются оценки сходства по формуле (2). Аналогично рассматриваются пары подсемейств одного семейства и пара классов одного суперкласса.

Для полученного распределения оценок сходства возможен расчет пяти метрик сходства двух наборов ТФ: минимум (Min), квартили Q1, Q2 (медиана) и Q3 и максимум (Max). Метрики Min/Max означают выбор по распределению сходства минимального/максимального значений, а метрики квартилей – величины соответствующей доли всего распределения. Например, метрика Q2 (медиана) для двух наборов ТФ отражает уровень сходства 50 % от всех возможных пар ТФ из этих наборов. Пусть первый {X} и второй {Y} наборы имеют K и T ТФ, $1 \le k \le K$, $1 \le t \le T$, тогда на основе распределений значений сходства в парах ТФ, рассчитанных по формуле (2) $\{\mathrm{Score}_{X(k),Y(t)}\}$, вычисляется метрика сходства $\mathrm{Score}_{\{X\},\{Y\}}$ двух наборов ТФ по формуле:

$$\mathrm{Score}_{\{X\},\{Y\}} = \mathrm{Median}_{1\le k\le K,\ 1\le t\le T} \{\mathrm{Score}_{X(k),Y(t)}\}. \quad (4)$$

Определение ветвей в иерархической классификации транскрипционных факторов. Если сходство двух наборов ТФ по мотивам их сайтов связывания превышает заданный порог Thr, то эти ТФ можно относить к одной ветви. Далее рассмотрим метрику медиана (4). Например, целый класс может относиться к одной ветви, если для него больше половины от всех возможных пар ТФ оказались сходными по мотивам сайтов связывания. Однако при этом возможно, что отдельные семейства класса не показывают значимого сходства, тем не менее с вероятностью более 50 % произвольная пара ТФ класса покажет значимое сходство мотивов сайтов связывания.

Для проведения кластерного анализа и построения деревьев, отражающих сходство ТФ по мотивам ССТФ сестринских классов одного суперкласса, сестринских

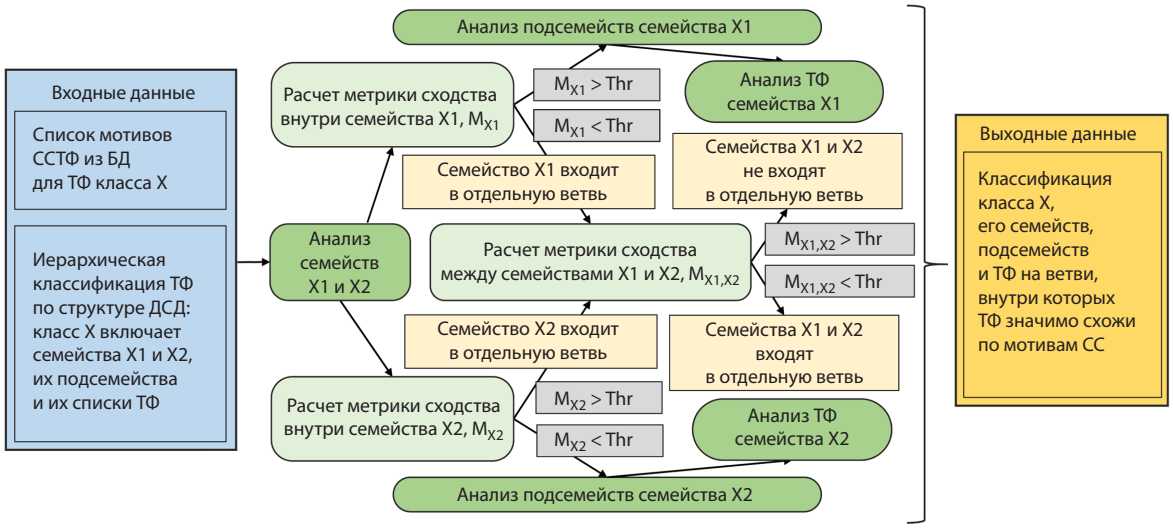


Рис. 2. Схема анализа для определения ветвей схожих мотивов ССТФ. На схеме детально представлен этап анализа одного класса X из двух семейств X1 и X2. Голубой цвет показывает входные данные, темно-зеленый – этапы анализа, светло-зеленый – расчеты метрик сходства, серый – проверки условий сходства мотивов, светло-желтый – промежуточные результаты, темно-желтый – окончательные результаты. Схема раскрывает анализ двух семейств X1 и X2 класса X. Анализ подсемейств этих семейств и анализ ТФ в каждом из подсемейств производится аналогично анализу семейств X1 и X2, согласно описанию в тексте.

семейств одного класса и т. д., мы использовали схему алгоритма UPGMA (unweighted pair group method with arithmetic mean, метод невзвешенной попарной группировки с усреднением) (Sokal, Michener, 1958). В ходе классификации для оценки в любой паре объектов мы применили описанную выше метрику медиана (Q2), формула (4).

Для поиска ветвей анализ начинается с уровня суперкласса и продолжается на более низких уровнях иерархии класса, семейства, подсемейства или ТФ. Сначала производится расчет метрики сходства ТФ внутри заданного уровня иерархии, например класса, а также для всех семейств этого класса. Это дает список семейств со сходством, превышающим заданный порог Thr. Все такие семейства изначально относятся к разным ветвям; для анализа остальных семейств следует переходить на уровень ниже. Затем рассчитываются метрики сходства ТФ по всем возможным парам сестринских семейств этого класса, это дает матрицу сходства класса по семействам, на диагонали которой – значения сходства внутри каждого семейства, а выше диагонали – значения сходства по всем парам разных семейств. Выбирается пара семейств с наибольшим сходством. Если это сходство превышает порог, то пара таких семейств (ветвей) объединяется в одну ветвь. При этом сходство во всех парах обновленных ветвей пересчитывается. Расчеты продолжаются до тех пор, пока есть пары ветвей, допускающие объединение на основе сходства. Таким образом, можно постепенно спуститься на более низкий уровень и дойти до уровня транскрипционных факторов.

Отдельно выполняется анализ сходства мотивов сайтов связывания одного ТФ (см. формулу (3)), хотя, очевидно, этот анализ происходит внутри одной ветви, так как согласно формулам (2) и (4) каждая ветвь для любого ТФ содержит все мотивы его сайтов связывания, мы лишь можем отметить ТФ (см. рис. 1), имеющие значимо различные мотивы сайтов связывания.

Цель всего анализа состоит в последовательном нахождении таких наборов ТФ (например, для класса это список кластеров семейств), для которых метрика (4) превышает заданный порог Thr, а список по каждой из ветвей включает как можно большее число элементарных единиц классификации.

Суперклассы ТФ достаточно гетерогенны по сходству мотивов сайтов связывания, так что каждый суперкласс разбивается на несколько ветвей. Ветвь в иерархии TFClass определяется как максимально возможный набор ТФ от высшего уровня класса до такого максимально низкого уровня (на практике это класс, семейство, подсемейство, ТФ), что в этом наборе для большинства пар ТФ есть значимое сходство ТФ по мотивам их сайтов связывания согласно метрике сходства (4).

Ветвь может включать одну или несколько сестринских единиц классификации:

- целый класс,
- одно или несколько семейств одного класса,
- одно или несколько подсемейств одного семейства,
- один или несколько ТФ одного подсемейства.

Конечным результатом анализа является определение набора всех ветвей, внутри каждой из ветвей метрика (4) указывает на значимое сходство ТФ по мотивам сайтов связывания. Рисунок 2 представляет схему использованного в работе анализа.

Результаты

Сходство транскрипционных факторов в сестринских подсемействах одних семейств

Для того чтобы приступить к массовому анализу сходства мотивов сайтов связывания в разной степени родственных ТФ согласно иерархической классификации TFClass, мы проверим сходство мотивов ССТФ для подсемейств отдельных семейств, принадлежащих к различным классам

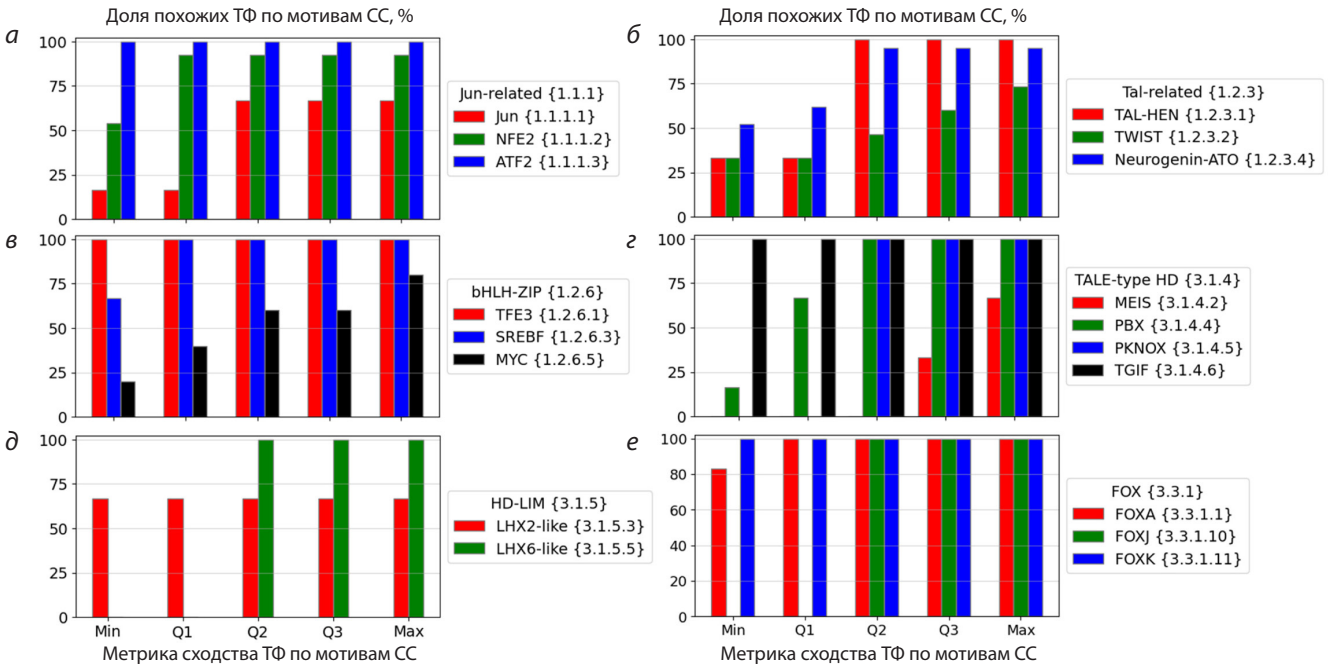


Рис. 3. Доля значимо похожих ТФ по мотивам сайтов связывания для подсемейств разных семейств при использовании пяти метрик сходства: Min, Q1, Q2, Q3 и Max.

а–е – семейства Jun-related {1.1.1}, Tal-related {1.2.3}, bHLH-ZIP {1.2.6}, TALE-type HD {3.1.4}, HD-LIM {3.1.5} и FOX {3.3.1} соответственно. Цвет отмечает подсемейства. Ось *X* – метрики сходства ТФ, ось *Y* – доля значимо похожих ТФ по мотивам сайтов связывания в подсемействе. Значимое сходство предполагает условие $-\mathrm{Log}_{10}[p\text{-value}] > 3$ (инструмент Tomtom, Gupta et al., 2007).

транскрипционных факторов. На рис. 3 изображены значения доли похожих ТФ по мотивам сайтов связывания внутри подсемейств разных семейств при использовании пяти метрик: Min, Q1, Q2, Q3 и Max. Метрика Q2 (медиана) рассчитана по формуле (4), другие – аналогично. По построению в ряду этих метрик от Min к Max доля похожих мотивов ССТФ растет. Однако независимо от выбора метрики некоторые подсемейства демонстрируют более низкое сходство или даже полное отсутствие сходных мотивов ССТФ, по сравнению с другими подсемействами. Например, для трех подсемейств Fox {3.3.1} семейства значения метрики Q2 близки к 100 % (см. рис. 3, *е*), а для подсемейств TWIST {1.2.3.2}/MEIS {3.1.4.2} семейств Tal-related {1.2.3}/TALE-type HD {3.1.4} соответственно эти значения меньше 50 % (см. рис. 3, *б*, *г*).

Таким образом, сходство ТФ по мотивам сайтов связывания способно значительно варьировать по подсемействам одних семейств. Очевидно, этот же вывод можно сделать для семейств одних классов. Далее в анализе для оценки сходства двух наборов ТФ была использована метрика медиана (Q2) (4), так как смысл ее применения наиболее прозрачен, по сравнению с метриками Min, Q1, Q3 и Max. Далее значение метрики Q2 называется «сходство».

Анализ сходства транскрипционных факторов человека

Рисунок 4 показывает деревья сходства ТФ человека по мотивам сайтов связывания для основных классов трех самых крупных суперклассов: Basic domain {1}, Zinc-coordinating DNA-binding domains {2} и Helix-turn-helix domains {3}. Из всех классов только один класс Tryptophan cluster factors {3.5} показывает значимое сходство ТФ по мотивам их сайтов связывания (сходство 3.68), классы Basic leucine zipper factors (bZIP) {1.1} и Nuclear receptors with C4 zinc fingers {2.1} достигают значений сходства 2.51 и 2.68 соответственно, что указывает на тренд к значимому сходству. Классы Fork head/winged helix factors {3.3}, Homeo domain factors {3.1} и Basic helix-loop-helix

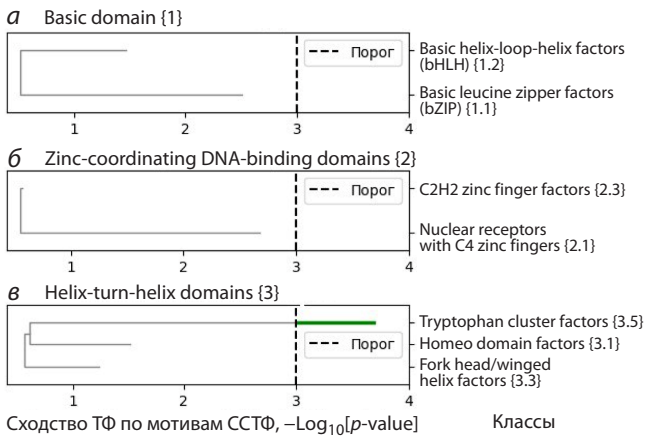


Рис. 4. Сходство ТФ по мотивам сайтов связывания в крупнейших классах трех крупнейших суперклассов человека.

а–в – деревья классов ТФ из суперклассов Basic domain {1}, Zinc-coordinating DNA-binding domains {2} и Helix-turn-helix domains {3}. Ось *X* отражает значение метрики Q2, пунктир обозначает ее пороговое значение 3. Зеленый цвет показывает класс Tryptophan cluster factors {3.5}, формирующий отдельную ветвь, а серым цветом выделены пути, значение метрики Q2 для которых меньше порога. Обрыв горизонтальной линии отмечает значение метрики Q2.

factors (bHLH) {1.2} дают более низкие значения сходства – 1.14, 1.42 и 1.47. Самое низкое сходство ТФ по мотивам сайтов связывания выявляется для класса C2H2 zinc finger factors {2.3} (0.44), этот класс самый крупный у человека, он допускает самую большую вариабельность структуры ТФ (Najafabadi et al., 2015; Lambert et al., 2018, 2019).

Следовательно, для выявления ветвей в пределах всех классов, кроме Tryptophan cluster factors {3.5}, необходимо перейти к анализу их семейств. Далее мы отдельно рассмотрим каждый из трех суперклассов подробнее.

Первый суперкласс имеет два крупных класса Basic leucine zipper factors (bZIP) {1.1} и Basic helix-loop-helix factors (bHLH) {1.2}, сходство ТФ по мотивам сайтов связывания между этими классами очень низко (0.523, рис. 5, *а*). Сходство ТФ внутри каждого класса заметно выше, но класс Basic leucine zipper factors (bZIP) {1.1} имеет значительно более сходные ТФ (2.51), чем класс Basic helix-loop-helix factors (bHLH) {1.2} (1.47).

В классе Basic leucine zipper factors (bZIP) {1.1} всего восемь семейств (см. рис. 5, *б*, *д*): от Jun-related {1.1.1} до C/EBP-related {1.1.8}. Каждое семейство класса имеет одно или несколько других семейств со значимо похожими ТФ по мотивам сайтов связывания. В результате все семейства распадаются на четыре ветви (см. рис. 5, *д*), в

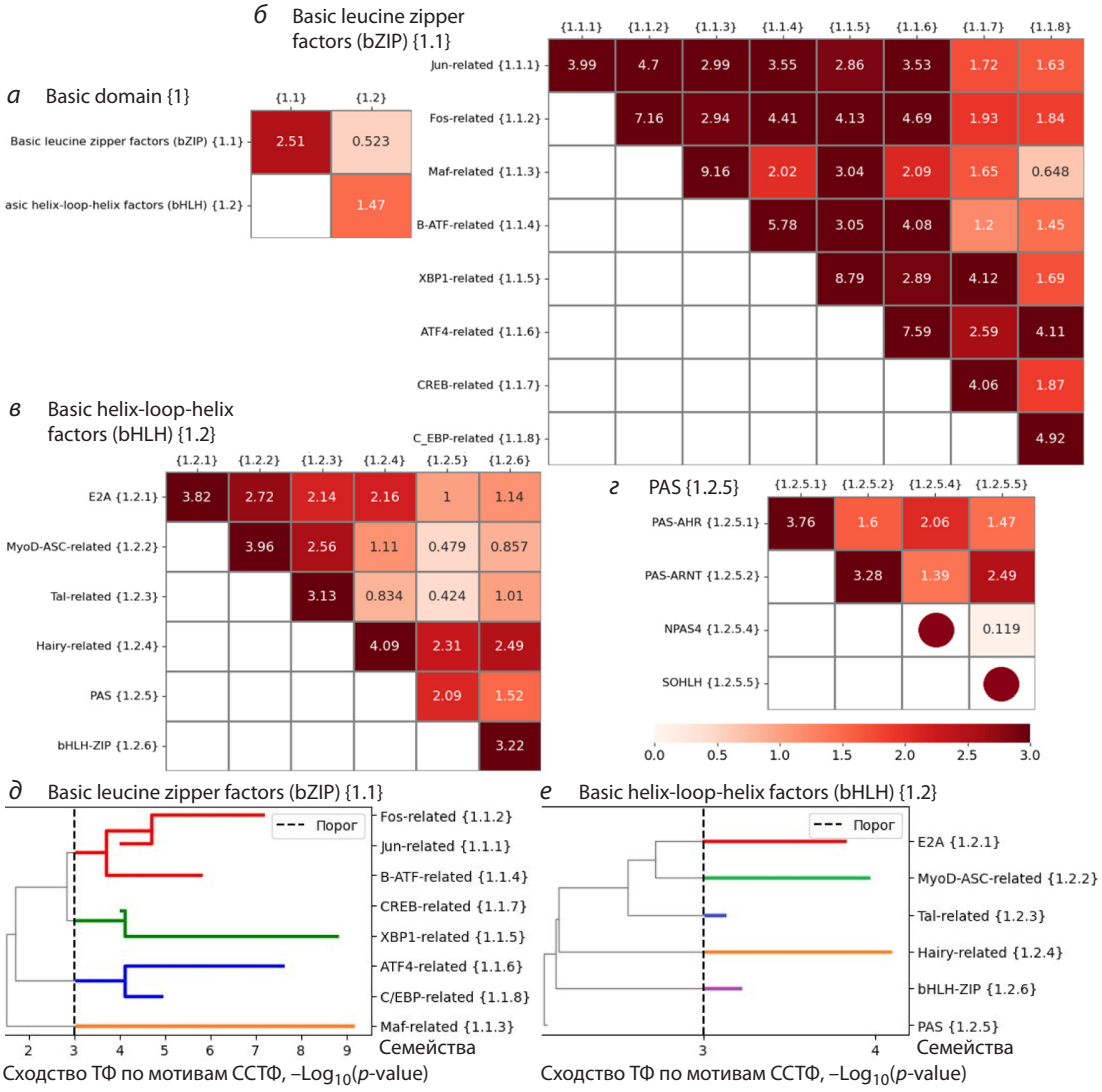


Рис. 5. Сходство ТФ по мотивам сайтов связывания для суперкласса Basic domain {1}.

а–г – тепловые карты для классов суперкласса, для семейств классов Basic leucine zipper factors (bZIP) {1.1}/Basic helix-loop-helix factors (bHLH) {1.2} и для подсемейств семейства PAS {1.2.5} класса Basic helix-loop-helix factors (bHLH) {1.2}. Коричневый круг на диагонали тепловой карты означает, что в подсемействе всего один ТФ с одним мотивом ССТФ. Цвет отражает значение метрики сходства Q2. Здесь и далее справа каждой тепловой карты указаны названия классов/семейств/подсемейств вместе с их цифровыми обозначениями, а сверху – только цифровые обозначения; *д* и *е* – деревья по семействам для классов Basic leucine zipper factors (bZIP) {1.1} и Basic helix-loop-helix factors (bHLH) {1.2}. Ось *Y* отражает значение метрики Q2, пунктир показывает ее пороговое значение 3. Все цвета, кроме серого, отражают отдельные ветви, а серым цветом выделены пути, значение метрики Q2 для которых меньше порога. Обрыв горизонтальной линии отмечает значение метрики Q2 для семейства. Семейство Jun-related {1.1.1} (*д*) имеет более низкое сходство 3.99 (*б*), чем сходство объединения семейств Jun-related {1.1.1} и Fos-related {1.1.2}, поэтому направление пути семейства Jun-related {1.1.1} от точки объединения этих двух семейств меняется на противоположное.

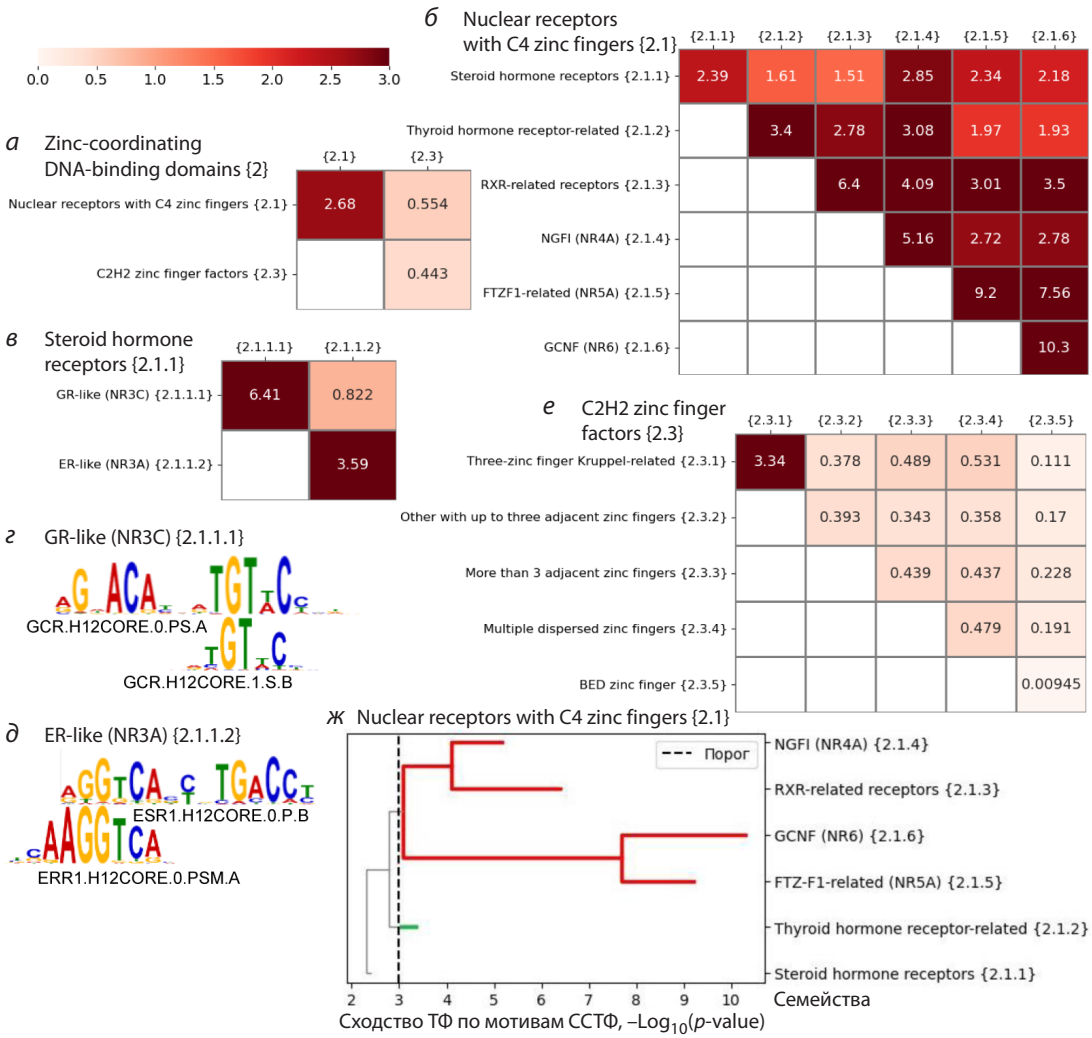


Рис. 6. Сходство ТФ по мотивам сайтов связывания для суперкласса Zinc-coordinating DNA-binding domains {2}.

а–в и *е* – тепловые карты для классов суперкласса, для семейств класса Nuclear receptors with C4 zinc fingers {2.1}, для подсемейств семейства Steroid hormone receptors {2.1.1} класса Nuclear receptors with C4 zinc fingers {2.1} и для семейств класса C2H2 zinc finger factors {2.3}; *г*, *д* – примеры мотивов сайтов связывания ТФ из подсемейств GR-like (NR3C) {2.1.1.1}/ER-like (NR3A) {2.1.1.2} семейства Steroid hormone receptors {2.1.1}; *ж* – дерево семейств класса Nuclear receptors with C4 zinc fingers {2.1}. Ось *Y* – значение метрики Q2, пунктир – ее пороговое значение 3. Красный и зеленый цвета отражают отдельные ветви, а серым цветом выделены пути, значение метрики Q2 для которых меньше порога ветви. Обрыв горизонтальной линии отмечает значение метрики Q2.

двух из которых по два семейства (XBP1-related {1.1.5} и CREB-related {1.1.7}, ATF4-related {1.1.6} и C/EBP-related {1.1.8}), а еще в двух – одно Maf-related {1.1.3} и три (Jun-related {1.1.1}, Fos-related {1.1.2}, B-ATF-related {1.1.4}).

В классе Basic helix-loop-helix factors (bHLH) {1.2} внутри каждого из семейств, за исключением одного PAS {1.2.5}, ТФ имеют значимое сходство по мотивам сайтов связывания (см. рис. 5, *б*, значения на диагонали), но между семействами значимого сходства ТФ по мотивам сайтов связывания не наблюдается. Поэтому каждое из семейств, за исключением семейства PAS {1.2.5}, образует отдельную ветвь (см. рис. 5, *е*). Семейство PAS {1.2.5} разбивается на четыре ветви {1.2.5.1}, {1.2.5.2}, {1.2.5.3} и {1.2.5.4} по четырем подсемействам (см. рис. 5, *г*).

Второй суперкласс имеет два крупных класса Nuclear receptors with C4 zinc fingers {2.1} и C2H2 zinc finger factors {2.3}, сходство ТФ по мотивам сайтов связывания между классами очень низко (0.554, рис. 6, *а*). В классе Nuclear receptors with C4 zinc fingers {2.1} ТФ имеют сходство лишь немного ниже порога (2.68), а сходство ТФ класса C2H2 zinc finger factors {2.3} очень низкое (0.443).

В классе Nuclear receptors with C4 zinc fingers {2.1} (см. рис. 6, *б*) только одно семейство Steroid hormone receptors {2.1.1} имеет сходство ТФ 2.39 ниже порога. Это семейство разбивается на две ветви по двум подсемействам GR-like (NR3C) {2.1.1.1} и ER-like (NR3A) {2.1.1.2} (см. рис. 6, *в*). Сходство ТФ между этими подсемействами низкое (0.822), а внутри каждого подсемейства оно высокое (6.41 и 3.59). Мотивы ССТФ из этих родственных подсемейств имеют сходную структуру: ТФ обоих подсемейств могут связаться как мономеры или как димеры, образованные инвертированным повтором (Nagy G., Nagy L., 2020), но независимо от этого именно моно-

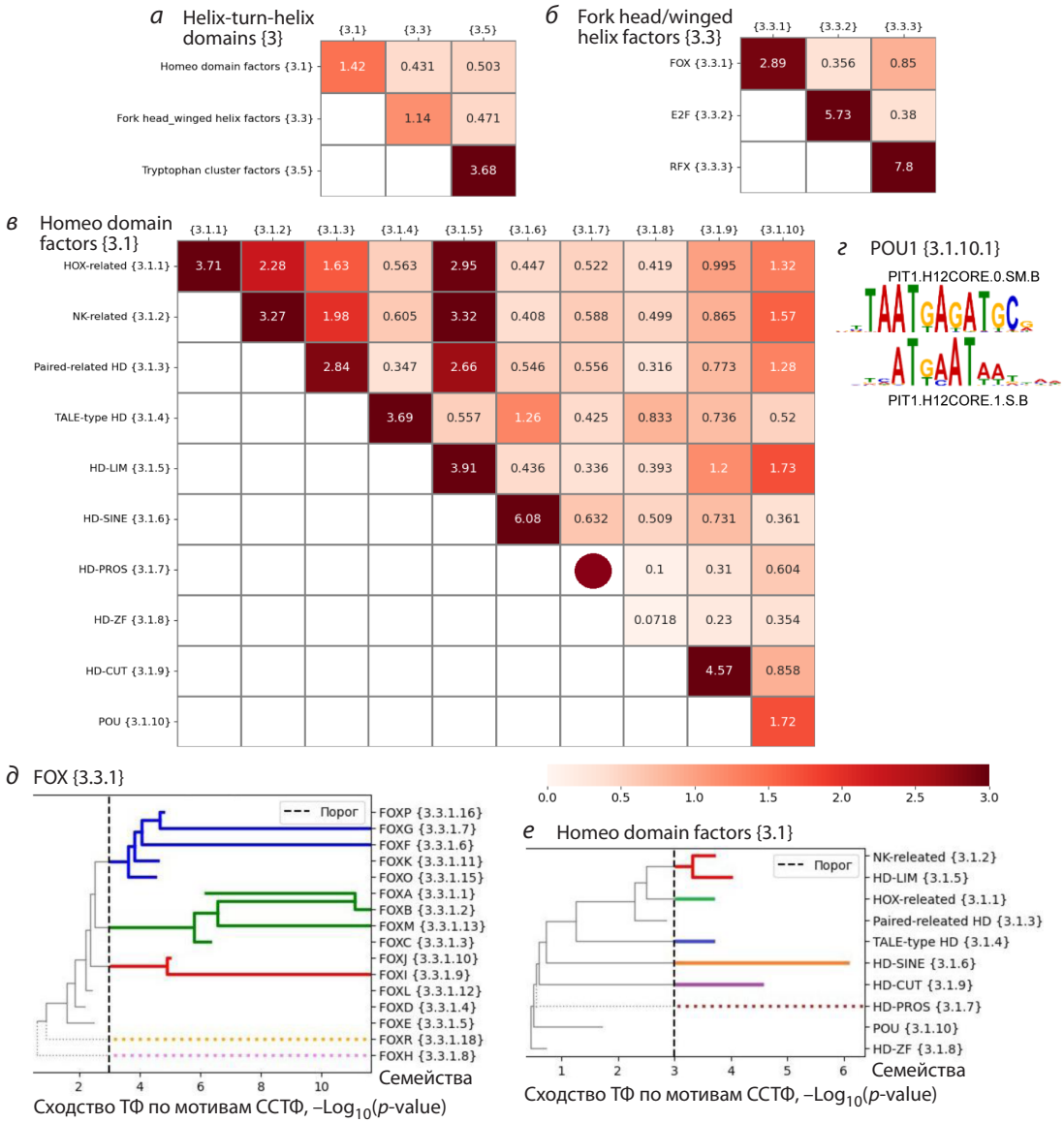


Рис. 7. Сходство ТФ по мотивам сайтов связывания для суперкласса Helix-turn-helix domains {3}.

а–в – тепловые карты для классов суперкласса, для семейств классов Fork head/winged helix factors {3.3} и Homeo domain factors {3.1}. Коричневый круг на диагонали тепловой карты означает, что в семействе всего один ТФ с одним мотивом сайтов связывания. Цвет отражает значение метрики сходства Q2; *г* – лого двух мотивов ССТФ PIT1 из подсемейства POU1 {3.1.10.1}; *д* и *е* – деревья для подсемейств семейства FOX {3.3.1} и семейств класса Homeo domain factors {3.1}. Ось *Y* отражает значение метрики Q2, пунктир показывает ее пороговое значение 3. Все цвета, кроме серого, отражают отдельные ветви, а серым цветом выделены пути, значение метрики Q2 для которых меньше порога ветви. Обрыв горизонтальной линии отмечает значение метрики Q2. Подсемейство FOXA {3.3.1.1} (*д*) имеет более низкое сходство 6.22 (см. рис. S1), чем сходство объединения подсемейств FOXA {3.3.1.1} и FOXB {3.3.1.2}, поэтому направление пути подсемейства FOXA {3.3.1.1} от точки объединения этих двух подсемейств меняется на противоположное.

мерные субъединицы в мотивах ССТФ в подсемействах GR-like (NR3C) {2.1.1.1} (см. рис. 6, *г*) и ER-like (NR3A) {2.1.1.2} (см. рис. 6, *д*) четко различаются. Семейство Thyroid hormone receptor-related {2.1.2} формирует отдельную ветвь, так как сходство его ТФ с ТФ четырех из пяти других семейств ниже порога 3 (см. рис. 6, *б*, *ж*). Четыре семейства от RXR-related receptors {2.1.3} до GCNF (NR6) {2.1.6} формируют одну ветвь: рис. 6, *ж* показывает дерево разделения класса Nuclear receptors with C4 zinc fingers {2.1} по семействам на ветви.

В классе C2H2 zinc finger factors {2.3} (см. рис. 6, *е*) только одно семейство Three-zinc finger Kruppel-related {2.3.1} образует отдельную ветвь. Для определения ветвей по остальным четырем семействам класса следует спуститься до уровней подсемейств или ТФ, см. список всех ветвей класса C2H2 zinc finger factors {2.3} в табл. S1[1].

Третий суперкласс включает три крупных класса Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} и Tryptophan cluster factors {3.5}. Сходство между ТФ разных классов по мотивам сайтов связывания очень низко во всех трех возможных парах классов (рис. 7, *а*, клетки выше диагонали). Сходство ТФ внутри каждого из классов

[1] Табл. S1, а также рис. S1 и S2 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx30.pdf

Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} среднее, 1.42 и 1.12. ТФ класса Tryptophan cluster factors {3.5} формируют одну ветвь (см. рис. 4).

В классе Fork head/winged helix factors {3.3} два семейства E2F {3.3.2} и RFX {3.3.3} представляют две отдельные ветви, а сходство ТФ семейства FOX {3.3.1} почти достигает порога (величина сходства 2.89, см. рис. 7, *б*). Яркой иллюстрацией верности разделения класса Fork head/winged helix factors {3.3} на три семейства (см. рис. 7, *б*) является заметное превышение сходства ТФ внутри семейств (три значения на диагонали) по отношению к сходству ТФ между семействами (три значения выше диагонали).

Среди 16 подсемейств семейства FOX {3.3.1} (см. рис. 7, *д*) только три подсемейства FOXD {3.3.1.4}, FOXH {3.3.1.5} и FOXL {3.3.1.12} достигли сходства ТФ ниже порога 3: 2.19, 2.48 и 2.17 соответственно. Четыре, пять и два подсемейства формируют отдельные ветви (см. рис. 7, *д*). Выделяются два подсемейства FOXH {3.3.1.8}, FOXR {3.3.1.18} с низким сходством ТФ по мотивам сайтов связывания с другими подсемействами и между собой (рис. S1).

Два семейства NK-related {3.1.2} и HD-LIM {3.1.5} класса Homeo domain factors {3.1} сливаются в одну ветвь; каждое из пяти семейств HOX-related {3.1.1}, TALE-type HD {3.1.4}, HD-SINE {3.1.6}, HD-PROS {3.1.7}, HD-CUT {3.1.9} представляет отдельную ветвь (см. рис. 7, *в*, *е*). Для нахождения ветвей по остальным семействам Paired-related HD {3.1.3}, HD-ZF {3.1.8} и POU {3.1.10} необходим переход на уровень подсемейств (рис. S2 и табл. S1). Семейство Paired-related HD {3.1.3} разбито на две отдельные ветви, объединяющие 12 и 6 подсемейств (см. рис. S2, *а* и табл. S1). Семейство HD-ZF {3.1.8} распадается на две ветви по двум подсемействам, ZEB {3.1.8.3} и ZHX {3.1.8.5} (см. рис. S2, *б*). Три подсемейства POU2 {3.1.10.2}, POU3 {3.1.10.3} и POU5 {3.1.10.5} сливаются в одну ветвь. Подсемейство POU1 {3.1.10.1} представлено одним ТФ с двумя значимо не схожими мотивами ССТФ PIT1 PIT1.H12CORE.0.SM.B и PIT1.H12CORE.1.S.B (см. рис. 7, *г*). Оставшиеся три подсемейства POU4 {3.1.10.4}, POU6 {3.1.10.6} и HNF1-like {3.1.10.7} семейства POU {3.1.10} формируют отдельные ветви (см. рис. S2, *в*).

Полный список ветвей для семи крупнейших классов ТФ Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1}, Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} и Tryptophan cluster factors {3.5} приведен в табл. S1.

В целом, на основе результатов, представленных на рис. 5–7, S1 и S2, а также в табл. S1, можно заключить, что часто именно ТФ одного семейства уже имеют несхожие мотивы сайтов связывания. Однако эта общая тенденция нарушается для некоторых классов и семейств. Наиболее отчетливо она нарушается для самого крупного класса ТФ человека C2H2 zinc finger factors {2.3} (см. рис. 6, *е*), для которого для определения ветвей необходимо опускаться на уровень подсемейств или даже на уровень транскрипционных факторов.

Анализ сходства транскрипционных факторов дрозофилы

Для определения того, насколько обнаруженные закономерности сходства по разным классам ТФ зависят от выбора таксона, мы провели анализ, аналогичный проведенному выше, для достаточно удаленного от таксона млекопитающих таксона насекомых. По данным БД Jaspar, есть всего два класса ТФ насекомых с числом мотивов сайтов связывания более 50 (см. таблицу). Все эти ТФ относятся к виду *D. melanogaster*. Результаты, полученные для ТФ насекомых из этих двух классов, C2H2 zinc finger factors {2.3} и Homeo domain factors {3.1}, хорошо согласуются с полученными выше результатами для ТФ человека из семи классов (см. рис. 4–7).

В классе C2H2 zinc finger factors {2.3} дрозофилы (рис. 8, *а*), также, как и в этом же классе человека (см. рис. 6, *е*), есть только одно семейство Three-zinc finger Kruppel-related {2.3.1} со значимо схожими ТФ по мотивам сайтов связывания. Только у ТФ одного другого семейства BED zinc finger {2.3.5} сходство мотивов сайтов связывания весьма различается (человек 0.001, дрозофила 6.32). Однако это семейство очень малочисленное: у дрозофилы оно содержит два почти не отличимых мотива сайтов связывания одного ТФ Dref, а у человека два ТФ ZBED1 и ZBED5 имеют отчетливо непохожие между собой мотива сайтов связывания. Остальные три общие семейства у обоих таксонов, Other with up to three adjacent zinc fingers {2.3.2}, More than 3 adjacent zinc fingers {2.3.3}, Multiple dispersed zinc fingers {2.3.4}, а также все оставшиеся ТФ дрозофилы с неуказанными семействами, отнесенные в семейство Unclassified {2.3.0}, показывают крайне низкое сходство ТФ по мотивам сайтов связывания. В целом, как для ТФ человека, так и для ТФ дрозофилы, класс C2H2 zinc finger factors {2.3} имеет ТФ с очень низким сходством мотивов сайтов связывания дрозофилы (см. рис. 6, *е* и 8, *а*).

Транскрипционные факторы дрозофилы из класса Homeo domain factors {3.1} (см. рис. 8, *б*) показывают несколько меньшее сходство по мотивам сайтов связывания, чем ТФ из этого же класса человека (см. рис. 7, *в*). Однако в каждом из этих двух таксонов среди восьми общих семейств выделяются семейства с большим и меньшим сходством ТФ по мотивам сайтов связывания. А именно у ТФ обоих таксонов наибольшим сходством как внутри семейств, так и между семействами обладают четыре семейства HOX-related {3.1.1}, NK-related {3.1.2}, Paired-related HD {3.1.3} и HD-LIM {3.1.5} (см. рис. 7, *в*, *е*), однако при этом само сходство у ТФ дрозофилы превышает величину 2, но не достигает порога 3 (см. рис. 8, *б*). Остальные семейства имеют ТФ, не похожие как между собой, так и на ТФ указанных выше семейств класса. В целом, гораздо меньшее сходство мотивов ССТФ класса Homeo domain factors {3.1} дрозофилы (см. рис. 8, *б*) по сравнению с ССТФ этого же класса человека (см. рис. 8, *в*) может быть объяснено как заметно меньшим числом доступных данных массового секвенирования по мотивам ССТФ дрозофилы (см. таблицу), а также различием методов получения мотивов ССТФ в БД Hocomoco и Jaspar.

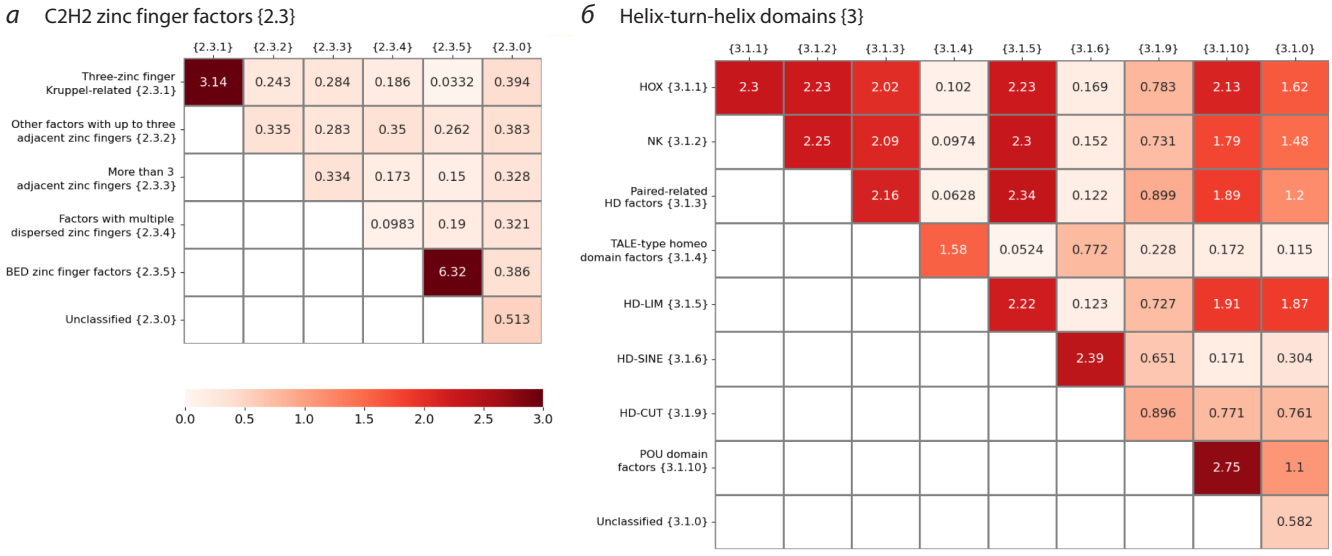


Рис. 8. Сходство ТФ дрозофилы из двух крупных классов по мотивам сайтов связывания.

а и *б* – семейства класса C2H2 zinc finger factors {2.3} и Homeo domain factors {3.1}. Цвет отражает значение метрики сходства Q2.

Обсуждение

Мы предлагаем новый систематический подход для уточнения иерархической классификации ТФ по структуре ДСД набором ветвей, объединяющих ТФ со сходными мотивами сайтов связывания. Сходство мотивов сайтов связывания известных ТФ сегодня можно оценить по данным разных экспериментальных технологий массового секвенирования, включая данные *in vitro* HT-SELEX и *in vivo* ChIP-seq, например, результаты экспериментов для разных условий ткани и стадии развития.

Оценки общего числа ТФ человека/дрозофилы составляют 1659/651 (БД AnimalTFDB, Shen et al., 2023). БД Hocomoco (версия 12) для человека и БД Jaspar для дрозофилы аннотировали 1443 мотивов ССТФ для 949 ТФ и 334 мотивов ССТФ для 273 ТФ. Следовательно, хотя отношение числа ТФ с известными мотивами ССТФ к оценке общего числа ТФ у человека и дрозофилы близки (57 и 51 %), но в среднем на один ТФ приходится 1.52/1.22 аннотированных мотивов сайтов связывания у человека (Hocomoco)/дрозофилы (Jaspar). В согласии с этим БД GTRD (Kolmykov et al., 2021) представляет данные по 21988/3027 ChIP-seq экспериментам для 1531/595 ТФ человека/дрозофилы. Следовательно, разнообразие структурных типов мотивов ССТФ изучено уже заметно лучше у человека, чем у дрозофилы.

Задача анализа данных связывания ТФ *in vivo* осложняется тем, что обогащенные мотивы из результатов поиска мотива *de novo* могут соответствовать сайтам связывания целевых или партнерских ТФ. Данные массового секвенирования *in vitro*, такие как HT-SELEX или DAP-seq, отражают только прямое связывание целевых ТФ и полностью исключают совместное связывание целевых ТФ с любыми партнерскими ТФ и связывание целевых ТФ не напрямую. Следовательно, нуклеотидная специфичность связывания целевых ТФ *in vitro* может определить лишь часть локусов их связывания *in vivo*. Данные секвенирования ССТФ *in vivo* отражают основной кооперативный механизм связывания целевого ТФ с геномной ДНК, включающий его взаимодействия с различными партнерскими ТФ (Morgunova, Taipale, 2017). Это осложняет привязку обогащенных мотивов *de novo* к конкретным партнерским транскрипционным факторам.

Наблюдаемая на основе систематизации современных данных массового секвенирования ССТФ вариабельность их мотивов отражает разнообразие структуры ДСД ТФ. Последние важны для функции прямого связывания целевого и партнерских ТФ. Например, способностью действовать в составе димеров близкородственных ТФ обладают лишь ТФ некоторых классов (Amoutzias et al., 2008); из изученных в нашей работе (см. таблицу) это классы Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2} и Nuclear receptors with C4 zinc fingers {2.1}. Главная функция конкретного ТФ, его способность взаимодействовать с геномной ДНК, зависит от места конкретного ТФ в общей иерархии структуры ДСД всех ТФ, то есть от суперкласса, класса, семейства и подсемейства этого ТФ. Ранее эти уровни иерархической классификации ТФ были определены по структуре ДСД и выравниванию аминокислотных последовательностей ДСД ТФ (БД TFClass, Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018), сходство мотивов ССТФ при этом не учитывалось. Систематический анализ сходства мотивов ССТФ может сделать классификацию ТФ более эффективной для практического применения на этапе интерпретации обогащенных мотивов, результатов *de novo* поиска мотивов по данным массового картирования ССТФ *in vivo*, таким как ChIP-seq.

Определение общей топологии ветвей значимо схожих мотивов ССТФ заключается в выборе для каждого ТФ такого уровня иерархии среди опций одного класса, одного или нескольких сестринских семейств (или подсемейств) или отдельного ТФ, так что для ТФ всей ветви боль-

шинство пар ТФ имеет значимо похожие мотивы сайтов связывания. Для определения списка ветвей необходимы: иерархическая классификация ТФ по структуре ДСД из БД TFClass/Plant-TFclass; наборы мотивов ССТФ из БД; формула расчета сходства двух наборов ТФ по мотивам их сайтов связывания (4). Выявление всех ветвей по иерархии TFClass/Plant-TFclass поможет избежать избыточной детализации в выходных данных *de novo* поиска мотивов. Этот информационный шум и избыточная информация возникают из-за того, что для любой из отдельных классификационных единиц, таких как конкретный класс или семейство/подсемейство, не определены допустимые пределы вариабельности сходства мотивов ССТФ. Таких ограничений также не было изначально и для ДСД ТФ (Wingender, 1997, 2013; Wingender et al., 2013, 2015, 2018).

Мы включили в анализ классы с числом мотивов ССТФ более 50 (см. таблицу). Из семи крупнейших классов человека (см. рис. 4) только один класс Tryptophan cluster factors {3.5} показал значимое сходство мотивов ССТФ. Для классов Basic leucine zipper factors (bZIP) {1.1} и Nuclear receptors with C4 zinc fingers {2.1} сходство ниже порога значимости (значения 3), но является все еще заметным (значения в интервале от 2 до 3), еще ниже сходство для классов Basic helix-loop-helix factors (bHLH) {1.2}, Homeo domain factors {3.1} и Fork head/winged helix factors {3.3} (значения в интервале от 1 до 2). Однако для класса C2H2 zinc finger factors {2.3} значение сходства меньше 1. Такое низкое значение отражает наличие в классе большинства пар ТФ с совершенно непохожими мотивами сайтов связывания, примерно такие же значения сходства наблюдаются между мотивами сайтов связывания в любой паре ТФ из разных классов одного суперкласса (см. значения в клетках вне диагонали на рис. 5, *а*, 6, *а* и 7, *а*). Подобные расхождения наблюдаются и на более низком уровне семейств ТФ.

Для каждого из классов Basic leucine zipper factors (bZIP) {1.1} и Nuclear receptors with C4 zinc fingers {2.1} в большинстве случаев несколько сестринских семейств объединяются в одну ветвь (см. рис. 5, *д* и 6, *ж*), а в классах Basic helix-loop-helix factors (bHLH) {1.2}, Homeo domain factors {3.1} и Fork head/winged helix factors {3.3} (см. рис. 5, *е* и 7, *б*, *е*) разделение на ветви ближе к разделению по семействам. Уровня семейств явно недостаточно для выделения ветвей в классе C2H2 zinc finger factors {2.3} (см. рис. 6, *е*). Таким образом, наш анализ подтверждает четкие отличия изменчивости мотивов сайтов связывания для крупнейших классов ТФ человека (см. рис. 4–7, Lambert et al., 2018; Ambrosini et al., 2020). Согласующаяся с этим тенденция наблюдается и для мотивов сайтов связывания из двух крупнейших классов ТФ насекомых (см. рис. 8). Этот вывод хорошо коррелирует с результатами массового сравнения нуклеотидной специфичности ортологичных ТФ человека и дрозофилы, где установлено, что в целом мотивы ССТФ человека и дрозофилы имеют высокий уровень консервативности (Nitta et al., 2015). Однако позже этот вывод был уточнен детальным анализом сходства мотивов сайтов связывания разных классов ТФ у разных таксонов эукариот в линиях многоклеточных животных и высших растений (Lambert et al., 2019). В этой работе было показано, что консервативность как в линии животных, так и в линии растений очень сильно зависит от класса ТФ. Например, почти половина непохожих мотивов сайтов связывания ортологичных ТФ человека и дрозофилы относится к классу C2H2 zinc finger factors {2.3}, что согласуется с результатами нашего анализа (см. рис. 6, *е* и 8, *а*). Проведенный анализ (Lambert et al., 2019) также показал, что для отдельных ортологичных ТФ дрозофилы и человека сходство распространялось даже на уровень едва различимых предпочтений частот динуклеотидов в мотивах ССТФ.

Мы также пришли к заключению, что среди крупных классов ТФ класс C2H2 zinc finger factors {2.3} обладает ТФ с наиболее изменчивыми мотивами сайтов связывания у человека и дрозофилы (см. рис. 6, *е* и 8, *а*). По сравнению с классом C2H2 zinc finger factors {2.3} у обоих таксонов менее изменчивы мотивы ССТФ класса Homeo domain factors {3.1}. Однако для ТФ класса Homeo domain factors {3.1} обнаруживается большая изменчивость мотивов сайтов связывания у дрозофилы по сравнению с человеком (см. рис. 7, *в* и 8, *б*). Этот результат может отражать различия конвейеров для подготовки мотивов ССТФ в БД Hocomoco и Jaspar.

В БД Hocomoco мотивы сайтов связывания по каждому отдельному ТФ отражают данные по нескольким экспериментам массового секвенирования для этого ТФ (Kolmykov et al., 2021; Vorontsov et al., 2024), таким как ChIP-seq и HT-SELEX, например, часто даже сочетаются доступные данные по видам человек и мышь. Задача анализа в БД Hocomoco состояла в интеграции всех доступных данных по сайтам связывания отдельных ТФ так, чтобы по возможности выявить разные структурные типы мотивов сайтов связывания каждого ТФ. В БД Jaspar реализован более простой способ представления каждого из мотивов отдельным экспериментом, что можно считать оправданным, пока данных по отдельным ТФ все еще мало. Для мотивов ССТФ насекомых анализ, подобный проведенному для получения мотивов ССТФ БД Hocomoco, еще не проводился, что отчасти объясняется существенно меньшим объемом доступных данных массового секвенирования (Kolmykov et al., 2021; Rauluseviciute et al., 2024). Можно предполагать, что подход БД Hocomoco по сравнению с подходом БД Jaspar, скорее всего, отражает большее число минорных мотивов сайтов связывания по каждому из ТФ, что может способствовать большему сходству мотивов, определяемому с помощью разработанного нами подхода, согласно формулам (2) и (4). Тем не менее регулярные обновления и рост количества данных по известным мотивам ССТФ в обеих БД Hocomoco и Jaspar в последние годы (Rauluseviciute et al., 2024; Vorontsov et al., 2024) указывает на то, что в скором будущем классификация мотивов ССТФ может быть уточнена.

В целом, на основе полученных нами результатов можно заключить, что как для таксона млекопитающие, так и для таксона насекомые заметные расхождения сходства мотивов сайтов связывания по ТФ крупных классов и их семейств затрудняют применение стандартной терминологии БД TFClass, включающей классы, семейства и подсемейства ТФ для описания вариабельности мотивов ССТФ. Следовательно, для более эффективного выявления ТФ

по результатам массового секвенирования ССТФ *in vivo* необходим систематический анализ сходства мотивов сайтов связывания известных ТФ с целью упорядочивания изменчивости мотивов ССТФ в пределах различных элементарных единиц классификации от классов до отдельных транскрипционных факторов.

В дальнейшем более масштабный анализ сходства мотивов сайтов связывания в пределах всех классов, семейств, подсемейств ТФ и отдельных ТФ у модельных видов млекопитающих, насекомых и высших растений может быть прочным основанием для более эффективного определения мотивов ССТФ по данным массового секвенирования ChIP-seq. На основе проведенного нами массового анализа мы предлагаем в результатах поиска *de novo* мотивов для выявленных обогащенных мотивов указывать не только имена ТФ с привязанными к ним именами класса/семейства/подсемейства, но и определенные нами ветви иерархической классификации ТФ. Эти ветви являются составными единицами классификации, интегрирующими несколько последовательных уровней иерархии. Каждая ветвь представляет в рамках единой многоуровневой классификации ТФ по сходству и выравниванию ДСД множество ТФ со значимо похожими мотивами сайтов связывания.

Заключение

В данной работе нами представлен подход для систематического анализа сходства мотивов сайтов связывания известных ТФ на основе многоуровневой иерархии ТФ по структуре ДСД из базы данных TFClass, включающей уровни суперклассов, классов, семейств, подсемейств и отдельных ТФ. В общей иерархии мы определили для крупнейших классов ТФ млекопитающих (человек) и насекомых (плодовая мушка) общие деревья ветвей со значимо похожими по мотивам сайтов связывания ТФ. В наш анализ вошли семь классов ТФ млекопитающих: Basic leucine zipper factors (bZIP) {1.1}, Basic helix-loop-helix factors (bHLH) {1.2}, Nuclear receptors with C4 zinc fingers {2.1}, C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1}, Fork head/winged helix factors {3.3} и Tryptophan cluster factors {3.5} – и два класса ТФ насекомых: C2H2 zinc finger factors {2.3} и Homeo domain factors {3.1}. Показано, что как для таксона млекопитающих, так и для таксона насекомых сходство мотивов сайтов связывания заметно отличается среди ТФ из разных классов. Систематический анализ сходства мотивов сайтов связывания структурно-родственных ТФ, определенных согласно иерархической классификации, позволил определить уровни иерархии (классы, семейства, подсемейства, ТФ), начиная с которых и ниже по иерархии мотивы сайтов связывания известных ТФ становятся значимо похожими. В дополнение к улучшению идентификации вовлеченных ТФ по результатам поиска мотивов *de novo* и, следовательно, к более эффективному выявлению механизмов регуляции генов наши результаты могут уточнить иерархическую классификацию ТФ по их ДСД. Мы не переопределяем классификацию ТФ по элементарным единицам от класса, семейства и ниже по иерархии, мы вносим дополнительную информацию о сходстве мотивов ССТФ, которая отражает главную функцию ТФ, функцию специфичного по отношению к последовательности ДНК связывания, что, безусловно, должно более точно отличать разные транскрипционные факторы.

Список литературы / References

Ambrosini G., Vorontsov I., Penzar D., Groux R., Fornes O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Bucher P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol.* 2020;21(1):114. doi 10.1186/s13059-020-01996-3

Amoutzias G.D., Robertson D.L., Van de Peer Y., Oliver S.G. Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem Sci.* 2008;33(5):220-229. doi 10.1016/j.tibs.2008.02.002

Bailey T.L. STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* 2021;37(18):2834-2840. doi 10.1093/bioinformatics/btab203

Blanc-Mathieu R., Dumas R., Turchi L., Lucas J., Parcy F. Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.* 2024;29(1):40-51. doi 10.1016/j.tplants.2023.06.023

D'haeseleer P. What are DNA sequence motifs? *Nat Biotechnol.* 2006; 24(4):423-425. doi 10.1038/nbt0406-423

de Martin X., Sodaei R., Santpere G. Mechanisms of binding specificity among bHLH transcription factors. *Int J Mol Sci.* 2021;22(17): 9150. doi 10.3390/ijms22179150

Franco-Zorrilla J.M., López-Vidriero I., Carrasco J.L., Godoy M., Vera P., Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci USA.* 2014;111(6):2367-2372. doi 10.1073/pnas.1316278111

Gupta S., Stamatoyannopolous J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol.* 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24

Hammal F., de Langen P., Bergon A., Lopez F., Ballester B. ReMap 2022: A database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* 2022;50(D1):D316-D325. doi 10.1093/nar/gkab996

Johnson D.S., Mortazavi A., Myers R.M., Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science.* 2007;316(5830): 1497-1502. doi 10.1126/science.1141319

Jolma A., Yan J., Whitington T., Toivonen J., Nitta K.R., Rastas P., Morgunova E., … Hughes T.R., Lemaire P., Ukkonen E., Kivioja T., Taipale J. DNA-binding specificities of human transcription factors. *Cell.* 2013;152(1-2):327-339. doi 10.1016/j.cell.2012.12.009

Kolmykov S., Yevshin I., Kulyashov M., Sharipov R., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., Kolpakov F. GTRD: An integrated view of transcription regulation. *Nucleic Acids Res.* 2021; 49(D1):D104-D111. doi 10.1093/nar/gkaa1057

Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell.* 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029

Lambert S.A., Yan A.W.H., Sasse A., Cowley G., Albu M., Caddick M.X., Morris Q.D., Weirauch M.T., Hughes T.R. Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet.* 2019;51(6):981-989. doi 10.1038/s41588-019-0411-1

Levitsky V., Zemlyanskaya E., Oshchepkov D., Podkolodnaya O., Ignatieva E., Grosse I., Mironova V., Merkulova T. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. *Nucleic Acids Res.* 2019;47(21):e139. doi 10.1093/nar/gkz800

Liu B., Yang J., Li Y., McDermaid A., Ma Q. An algorithmic perspective of *de novo* cis-regulatory motif finding based on ChIP-seq data. *Brief Bioinform.* 2018;19(5):1069-1081. doi 10.1093/bib/bbx026

Lloyd S.M., Bao X. Pinpointing the genomic localizations of chromatin-associated proteins: the yesterday, today, and tomorrow of ChIP-seq. *Curr Protoc Cell Biol.* 2019;84(1):e89. doi 10.1002/cpcb.89

Morgunova E., Taipale J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol.* 2017;47:1-8. doi 10.1016/j.sbi.2017.03.006

Nagy G., Nagy L. Motif grammar: The basis of the language of gene expression. *Comput Struct Biotechnol J.* 2020;18:2026-2032. doi 10.1016/j.csbj.2020.07.007

Najafabadi H.S., Mnaimneh S., Schmitges F.W., Garton M., Lam K.N., Yang A., Albu M., Weirauch M.T., Radovani E., Kim P.M., Greenblatt J., Frey B.J., Hughes T.R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol.* 2015;33(5):555-562. doi 10.1038/nbt.3128

Nakato R., Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform.* 2017;18(2):279-290. doi 10.1093/bib/bbw023

Nitta K.R., Jolma A., Yin Y., Morgunova E., Kivioja T., Akhtar J., Hens K., Toivonen J., Deplancke B., Furlong E.E., Taipale J. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife.* 2015;4:e04837. doi 10.7554/eLife.04837

Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., … Lenhard B., Sandelin A., Wasserman W.W., Parcy F., Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059

Schneider T.D., Stephens R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990;18(20):6097-6100. doi 10.1093/nar/18.20.6097

Shen W.K., Chen S.Y., Gan Z.Q., Zhang Y.Z., Yue T., Chen M.M., Xue Y., Hu H., Guo A.Y. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* 2023;51(D1):D39-D45. doi 10.1093/nar/gkac907

Skene P.J., Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife.* 2017;6:e21856. doi 10.7554/eLife.21856

Slattery M., Zhou T., Yang L., Dantas Machado A.C., Gordân R., Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.* 2014;39(9):381-399. doi 10.1016/j.tibs.2014.07.002

Sokal R.R., Michener C.D. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958;38:1409-1438. Available: https://archive.org/details/cbarchive_33927_astatisticalmethodforevaluatin1902/page/n1/mode/2up

Spitz F., Furlong E.E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet.* 2012;13(9):613-626. doi 10.1038/nrg3207

Stormo G.D., Zhao Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet.* 2010;11(11):751-760. doi 10.1038/nrg2845

Taing L., Dandawate A., L'Yi S., Gehlenborg N., Brown M., Meyer C.A. Cistrome Data Browser: integrated search, analysis and visualization of chromatin data. *Nucleic Acids Res.* 2024;52(D1):D61-D66. doi 10.1093/nar/gkad1069

Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A., Kamenets V., … Medvedeva Y.A., Jolma A., Kolpakov F., Makeev V.J., Kulakovskiy I.V. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* 2024;52(D1):D154-D163. doi 10.1093/nar/gkad1077

Wasserman W.W., Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet.* 2004;5(4):276-287. doi 10.1038/nrg1315

Weirauch M.T., Yang A., Albu M., Cote A.G., Montenegro-Monter A., Drewe P., Najafabadi H.S., … Bouget F.Y., Ratsch G., Larrondo L.F., Ecker J.R., Hughes T.R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell.* 2014;158(6):1431-1443. doi 10.1016/j.cell.2014.08.009

Wingender E. Classification scheme of eukaryotic transcription factors. *Mol Biol.* 1997:31(4):483-497. (translated from Вингендер Э. Классификация транскрипционных факторов эукариот. *Молекулярная биология*. 1997;31(4):584-600. Russian)

Wingender E. Criteria for an updated classification of human transcription factor DNA-binding domains. *J Bioinform Comput Biol.* 2013;11(1):1340007. doi 10.1142/S0219720013400076

Wingender E., Schoeps T., Dönitz J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* 2013;41(D1):D165-D170. doi 10.1093/nar/gks1123

Wingender E., Schoeps T., Haubrock M., Dönitz J. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.* 2015;43(D1):D97-D102. doi 10.1093/nar/gku1064

Wingender E., Schoeps T., Haubrock M., Krull M., Dönitz J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* 2018;46(D1):D343-D347. doi 10.1093/nar/gkx987

Zambelli F., Pesole G., Pavesi G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform.* 2013;14(2):225-237. doi 10.1093/bib/bbs016

Zenker S., Wulf D., Meierhenrich A., Viehöver P., Becker S., Eisenhut M., Stracke R., Weisshaar B., Bräutigam A. Many transcription factor families have evolutionarily conserved binding motifs in plants. *Plant Physiol.* 2025;198(2):kiaf205. doi 10.1093/plphys/kiaf205

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 09.07.2025. После доработки 09.09.2025. Принята к публикации 10.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):940-951

doi 10.18699/vjgb-25-100

Программа PlantReg 1.1: анализ взаимного расположения сайтов связывания транскрипционных факторов в промоторах генов-мишеней для уточнения молекулярных механизмов их активности в регуляторных сетях

В.В. Лавреха [1, 2#], Н.А. Омельянчук [1#], А.Г. Богомолов [1], Ю.А. Рябов [1], П.К. Мукебенова [1, 2], Е.В. Землянская [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
ezemlyanskaya@bionet.nsc.ru

Аннотация. Развитие высокопроизводительного секвенирования расширило возможности изучения регуляции экспрессии генов, в том числе для реконструкции генных регуляторных сетей и регуляторных сетей транскрипционных факторов (РСТФ). Актуальной задачей остается выявление молекулярных аспектов регуляции данными сетями биологических процессов. Решение этой задачи для растений позволит существенно продвинуться в понимании механизмов формирования хозяйственно важных признаков. Ранее мы разработали программу PlantReg для реконструкции транскрипционной регуляции биологических процессов у модельного вида *Arabidopsis thaliana* L. Воспроизводимые этой программой связи между РСТФ и генами, обеспечивающими протекание биологических процессов, охарактеризованы по типу регуляции (активация/подавление). Однако программа не позволяла определять, в каких случаях активация/подавление экспрессии гена-мишени обусловлены кооперативным или конкурентным взаимодействием транскрипционных факторов (ТФ). Мы предложили использовать информацию о взаимном расположении сайтов связывания (СС) ТФ в промоторе гена-мишени, а также данные о типе активности трансактивационных доменов ТФ для выявления кооперативного/конкурентного действия ТФ. Мы усовершенствовали программу, создав версию PlantReg 1.1, где обеспечили возможность точной локализации СС ТФ в протяженных районах связывания ТФ, устанавливаемых на основании полногеномных профилей DAP-seq (https://plamorph.sysbio.ru/fannotf/). Для демонстрации возможностей программы была исследована регуляция генов-мишеней ранее реконструированных нами РСТФ ответа на ауксин и солевой стресс у *A. thaliana*. В фокусе изучения были гены, кодирующие белки, участвующие в процессах биосинтеза хлорофилла и лигнина, биогенеза рибосом и в передаче сигнала абсцизовой кислоты. В данной работе установлено, что частота случаев конкурентной регуляции под влиянием ауксина и солевого стресса может быть достаточно высока (около 30 %). Показано, что конкуренция ТФ семейства bZIP за общие СС является значимым механизмом подавления транскрипции в ответ на ауксин, и что ауксин и солевой стресс могут задействовать общие механизмы конкурентной регуляции для модуляции экспрессии некоторых генов сигнального пути абсцизовой кислоты.
Ключевые слова: генная онтология; биологические процессы; генные регуляторные сети; сайт связывания; транскрипционный фактор; *Arabidopsis thaliana*

Для цитирования: Лавреха В.В., Омельянчук Н.А., Богомолов А.Г., Рябов Ю.А., Мукебенова П.К., Землянская Е.В. Программа PlantReg 1.1: анализ взаимного расположения сайтов связывания транскрипционных факторов в промоторах генов-мишеней для уточнения молекулярных механизмов их активности в регуляторных сетях. *Вавиловский журнал генетики и селекции*. 2025;29(7):940-951. doi 10.18699/vjgb-25-100

Финансирование. Работа выполнена в рамках бюджетного проекта FWNR-2022-0020.

PlantReg 1.1 identifies the mutual arrangement of transcription factor binding sites in the target promoters for the elucidation of molecular mechanisms within regulatory networks

V.V. Lavrekha [1, 2#], N.A. Omelyanchuk [1#], A.G. Bogomolov [1], Y.A. Ryabov [1], P.K. Mukebenova [1, 2], E.V. Zemlyanskaya [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
ezemlyanskaya@bionet.nsc.ru

© Лавреха В.В., Омельянчук Н.А., Богомолов А.Г., Рябов Ю.А., Мукебенова П.К., Землянская Е.В., 2025

[#] Равный вклад авторов

Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. The development of high-throughput sequencing has expanded the possibilities for studying the regulation of gene expression, including the reconstruction of gene regulatory networks and transcription factor regulatory networks (TFRNs). Identifying the molecular aspects for regulation of biological processes via these networks remains a challenge. Solving this problem for plants will significantly advance the understanding of the mechanisms shaping agronomically important traits. Previously, we developed the PlantReg program to reconstruct the transcriptional regulation of biological processes in the model species *Arabidopsis thaliana* L. The links established by this program between TFRNs and the genes regulating biological processes specify the type of regulation (activation/suppression). However, the program does not determine whether activation/suppression of the target gene is due to the cooperative or competitive interaction of transcription factors (TFs). We assumed that using information on the mutual arrangement of TF binding sites (BSs) in the target gene promoter as well as data on the activity type of TF effector domains would help to identify the cooperative/competitive action of TFs. We improved the program and created PlantReg 1.1, which enables precise localization of TF BSs in extended TF binding regions identified from genome-wide DAP-seq profiles (https://plamorph.sysbio.ru/fannotf/). To demonstrate the capabilities of the program, we used it to investigate the regulation of target genes in previously reconstructed TFRNs for auxin response and early reaction to salt stress in *A. thaliana*. The study focused on genes encoding proteins involved in chlorophyll and lignin biosynthesis, ribosome biogenesis, and abscisic acid (ABA) signaling. We revealed that the frequency of competitive regulation under the influence of auxin or salt stress could be quite high (approximately 30 %). We demonstrated that competition between bZIP family TFs for common BS is a significant mechanism of transcriptional repression in response to auxin, and that auxin and salt stress can engage common competitive regulatory mechanisms to modulate the expression of some genes in the ABA signaling pathway.

Key words: gene ontology; biological processes; gene regulatory networks; binding site; transcription factor; *Arabidopsis thaliana*

For citation: Lavrekha V.V., Omelyanchuk N.A., Bogomolov A.G., Ryabov Y.A., Mukebenova P.K., Zemlyanskaya E.V. PlantReg 1.1 identifies the mutual arrangement of transcription factor binding sites in the target promoters for the elucidation of molecular mechanisms within regulatory networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):940-951. doi 10.18699/vjgb-25-100

Введение

Становление и развитие полногеномных методов анализа, таких как RNA-seq (Deshpande et al., 2023), ChIP-seq (Park, 2009), DAP-seq (O'Malley et al., 2016), открыли широкие возможности для системно-биологического исследования механизмов, обеспечивающих транскрипционную регуляцию биологических процессов и формирование фенотипических признаков (Marshall-Colón, Kliebenstein, 2019; Zemlyanskaya et al., 2021). На основе анализа геномных и транскриптомных данных активно развиваются подходы к реконструкции генных регуляторных сетей, а также регуляторных сетей транскрипционных факторов (РСТФ) (Ko, Brandizzi, 2020; Rybakov et al., 2024). РСТФ – это представляемый в виде графа набор регуляторных взаимодействий (связей) между генами, кодирующими транскрипционные факторы (ТФ). Узлы графа соответствуют генам, а направленные ребра – регуляторным взаимодействиям ТФ, кодируемого одним геном, с другим геном. Реконструкция РСТФ и выявление взаимосвязей между этими сетями и биологическими процессами (или конкретными признаками) имеют большое значение для установления центральных регуляторных контуров, управляющих биологическими процессами, а также для создания предсказательных моделей регуляции этих процессов (Huang et al., 2025; Leong et al., 2025; Sun Y. et al., 2025).

К настоящему времени исследователям доступно несколько программных средств для реконструкции РСТФ у разных видов. Например, R-пакет NetAct (Su et al., 2022) позволяет воссоздавать РСТФ млекопитающих на основании транскриптомных данных и компьютерной базы генов-мишеней ТФ, курируемой авторами этого пакета. Ранее нами разработаны программы CisCross-FindTFnet для реконструкции РСТФ у модельного растения *Arabidopsis thaliana* (Omelyanchuk et al., 2024) и PlantReg для установления регуляторных связей между РСТФ и генами, обеспечивающими протекание биологических процессов, контролируемых этими сетями (Lavrekha et al., 2024). Обе программы осуществляют анализ транскриптомных данных с привлечением репрезентативной коллекции полногеномных DAP-seq-профилей связывания ТФ, при этом PlantReg использует результаты CisCross-FindTFnet в качестве входных данных.

Важным этапом реконструкции РСТФ является определение типа активности составляющих сеть ТФ (активаторы или репрессоры), так как именно эта характеристика обуславливает топологию и динамику функционирования регуляторных сетей (Dhatterwal et al., 2024). Широкомасштабное определение типов активности трансактивационных доменов (белковых доменов, обеспечивающих активаторную или репрессорную функцию ТФ) у более 400 ТФ *A. thaliana* (Hummel et al., 2023) способствовало решению данной задачи. Однако этого недостаточно для корректной классификации связей в сети, поскольку многие ТФ могут работать и как активаторы, и как репрессоры в зависимости от типа клеток, условий, белковых изоформ ТФ, конкретных промоторов и др. (Boyle, Després, 2010; Martínez et al., 2018; Nagahage et al., 2018; Wang et al., 2020). Именно поэтому при реконструкции РСТФ по транскриптомным данным активность ТФ, как правило, устанавливают на основании анализа профилей экспрессии их генов-мишеней, выделяя их среди дифференциально-экспрессирующихся генов (ДЭГ) (Su et al., 2022; Omelyanchuk et al., 2024).

Ранее мы реконструировали две РСТФ: первая, РСТФ-А, контролирует транскрипционный ответ генов на ауксин, вторая, РСТФ-С, – ранний ответ на солевой стресс у *A. thaliana* (Lavrekha et al., 2024; Omelyanchuk et al., 2024). В этих же работах с помощью алгоритма PlantReg показано, каким образом РСТФ-А вовлечена в регуляцию ауксином четырех различных биологических процессов

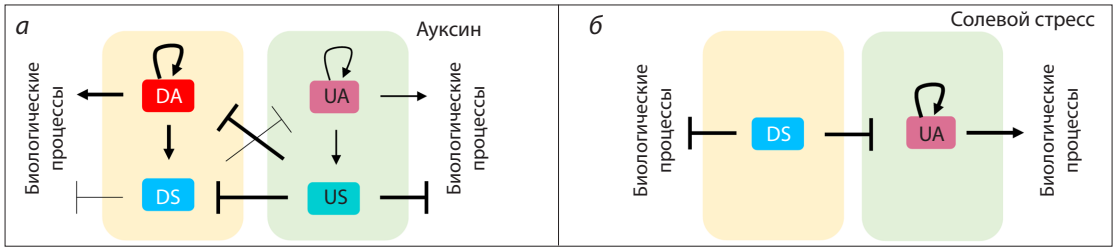


Рис. 1. Принципы регуляции биологических процессов ТФ из РСТФ-А (*а*) и РСТФ-С (*б*).

Желтые и зеленые прямоугольники – подавляемая и активируемая подсети РСТФ. Толщина стрелок отражает количество соответствующих связей в РСТФ. UA – активируемый активатор; US – активируемый супрессор; DA – подавляемый активатор; DS – подавляемый супрессор.

(подавление биосинтеза хлорофилла и лигнина, ответа на абсцизовую кислоту (АБК) и активацию биогенеза рибосом) и как контролируется усиление ответа на АБК при раннем солевом стрессе посредством РСТФ-С. В этих сетях ТФ разбиты на четыре класса: активируемый активатор (upregulated activator, UA), активируемый супрессор (upregulated suppressor, US), подавляемый активатор (downregulated activator, DA) и подавляемый супрессор (downregulated suppressor, DS). DA и DS формируют П-подсеть (активна в отсутствие стимула, подавляется под его воздействием), UA и US формируют А-подсеть (активируется стимулом).

И в транскрипционном ответе на ауксин, и в ответе на солевой стресс выявлена важная роль подавления транскрипции. Так, для ответа на ауксин характерно масштабное перепрограммирование обширной П-подсети, функционирующей в норме (до обработки гормоном), путем ее подавления ТФ типа US, входящими в состав А-подсети (рис. 1, *а*) (Omelyanchuk et al., 2024). В ответе на солевой стресс, напротив, запускается обширная А-подсеть, в том числе благодаря подавлению ее репрессоров типа DS из П-подсети (см. рис. 1, *б*) (Lavrekha et al., 2024).

Большинство супрессоров обеих РСТФ вовлечено также в регуляцию перечисленных выше биологических процессов в ответ на действие ауксина и солевого стресса (Lavrekha et al., 2024; Omelyanchuk et al., 2024). Однако, согласно литературным данным, основная часть предсказанных супрессоров в обеих РСТФ имеет трансактивационный домен активаторного типа (Hummel et al., 2023; Omelyanchuk et al., 2024). Это может быть следствием того, что подавление экспрессии гена-мишени этими ТФ происходит в результате их кооперативного или конкурентного взаимодействия с другими ТФ. Программа PlantReg позволяет установить регуляторные связи между РСТФ и генами, обеспечивающими протекание биологических процессов, но не выявляет случаи кооперации или конкуренции ТФ в контроле активности генов. В то же время понимание механизмов взаимодействия ТФ в регуляции транскрипции исключительно важно для эффективного использования реконструированных РСТФ и их связей с биологическими процессами в проектах по биоинженерии признаков растений.

Информация о взаимном расположении сайтов связывания (СС) различных ТФ в промоторе конкретного гена-мишени с привлечением данных о типе активности трансактивационных доменов этих ТФ (Hummel et al., 2023) может быть использована для выявления и характеристики кооперативного или конкурентного действия ТФ в промоторе конкретного гена. Например, в тех случаях, когда СС двух предсказанных репрессоров, работающих в одной подсети, располагаются в непосредственной близости друг от друга, при этом трансактивационный домен только одного из них обладает репрессорной активностью, тогда как другой по своей природе – активатор транскрипции, можно говорить о потенциальном кооперативном взаимодействии ТФ, превращающем ТФ-активатор в ТФ-репрессор (рис. 2, *а*). Такие примеры широко распространены и подробно описаны в научной литературе (Hanna-Rose, Hansen, 1996; Ahn et al., 2006; Veerabagu et al., 2014; Martínez et al., 2018; Wang et al., 2020).

Аналогично, если СС предсказанного активатора из одной подсети и предсказанного репрессора из другой подсети накладываются друг на друга в промоторе гена-мишени, при этом предсказанная активность одного из ТФ не соответствует установленной активности его трансактивационного домена, можно говорить о потенциальном конкурентном действии ТФ: замещение сильного активатора более слабым проявляется как подавление транскрипции гена-мишени, а замещение сильного репрессора более слабым – как активация гена-мишени (см. рис. 2, *б*). Снижение промоторной активности при увеличении концентрации слабого активатора по сравнению с сильным, как и обратный переход, показаны в ряде экспериментов (Tamura et al., 2004; Zhang et al., 2006; Chupreta et al., 2007; Ren et al., 2015; Selvaraj et al., 2015; Brackmann et al., 2018).

Для надежной идентификации генов-мишеней ТФ программа PlantReg использует пики DAP-seq, что, однако, не позволяет точно локализовать СС ТФ, так как размер пиков (более 150 п. н.) существенно превышает длину последовательностей, узнаваемых ТФ (менее 20 п. н.). В настоящей работе мы усовершенствовали программу, создав версию PlantReg 1.1, где обеспечили возможность точной локализации СС ТФ в протяженных районах связывания, устанавливаемых на основании полногеномных профилей DAP-seq (https://plamorph.sysbio.ru/fannotf/). Новая версия программы, PlantReg 1.1, была использована нами для выявления генов биосинтеза хлорофилла и лигнина, биогенеза рибосом и сигнального пути АБК, экспрессия которых может подавляться под контролем РСТФ-А или РСТФ-С в результате конкуренции ТФ-активаторов за общие сайты связывания.

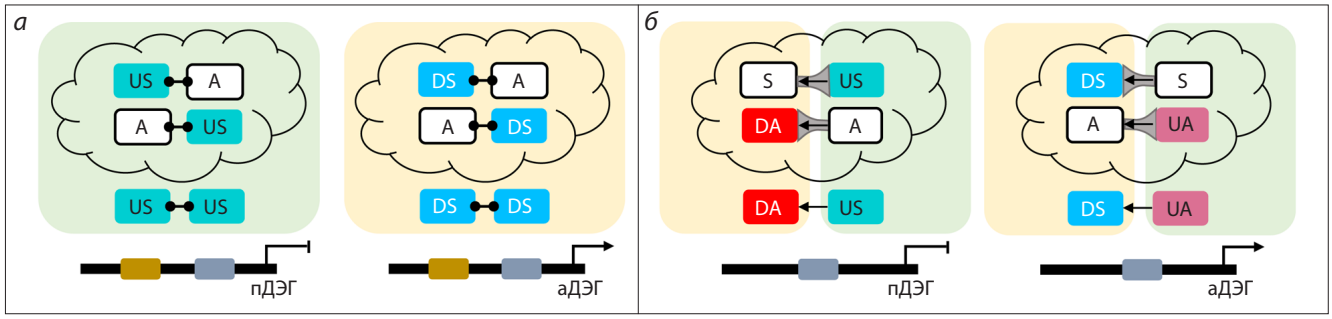


Рис. 2. Кооперативная (*а*) и конкурентная (*б*) регуляция гена-мишени парой ТФ из РСТФ.

Желтые и зеленые прямоугольники – подавляемая и активируемая подсети РСТФ. Предсказанные типы активности ТФ представлены внизу, альтернативные типы активности, возможные в реальности, представлены вверху (в облаке). Соединенные линиями точки между прямоугольниками, обозначающими ТФ, на (*а*) – взаимодействие белков; стрелки на (*б*) – замещение одного ТФ другим после действия стимула; серые воронки на (*б*) – соотношение активностей ТФ (большее основание соответствует более высокой активности); аДЭГ – активируемые ДЭГ; пДЭГ – подавляемые ДЭГ; UA – активируемый активатор; US – активируемый супрессор; DA – подавляемый активатор; DS – подавляемый супрессор.

Проведенный анализ позволил установить, что частота случаев конкурентной регуляции под влиянием ауксина и солевого стресса может быть достаточно высока. Помимо этого, мы показали, что конкуренция ТФ семейства bZIP за общие СС – значимый механизм подавления транскрипции в ответ на ауксин у *A. thaliana*, а также, что ауксин и солевой стресс могут задействовать общие механизмы конкурентной регуляции для модуляции экспрессии некоторых генов сигнального пути абсцизовой кислоты.

Материалы и методы

Интеграция данных о СС ТФ в 5′-регуляторных районах генов в программу PlantReg 1.1. Разработанная нами ранее программа PlantReg (Lavrekha et al., 2024) предназначена для реконструкции механизмов транскрипционной регуляции биологических процессов у *A. thaliana* на основании анализа списка ДЭГ и списка ТФ – известных или предполагаемых регуляторов транскрипции этих генов. PlantReg осуществляет функциональную аннотацию входного списка ДЭГ в терминах генной онтологии (ГО), после чего среди ДЭГ, ассоциированных с обогащенными биологическими процессами, выявляет потенциальные гены-мишени для ТФ из входного списка, используя полногеномные профили связывания ТФ, доступные в веб-версии программы (рис. 3, *а*). Выходные данные PlantReg представлены пятью блоками, в которых отражены отношения между биологическими процессами, участвующими в них ДЭГ и ТФ, регулирующими экспрессию этих ДЭГ.

Принципиальная схема работы обновленной версии PlantReg 1.1 представлена на рис. 3. Помимо исходного функционала, в ней доступны данные о распознанных СС ТФ в 5′-регуляторных районах генов (см. рис. 3, *а*), которые добавлены в блоки выдачи 1 и 4, что необходимо для исследования взаимного расположения их в промоторах генов. Исходно в блоке выдачи 1 представлен подсписок ДЭГ, ассоциированных с обогащенными биологическими процессами (см. рис. 3, *б*). Каждый ген подсписка охарактеризован набором терминов ГО (биологические процессы), которыми аннотирован ген, с кодами обоснования

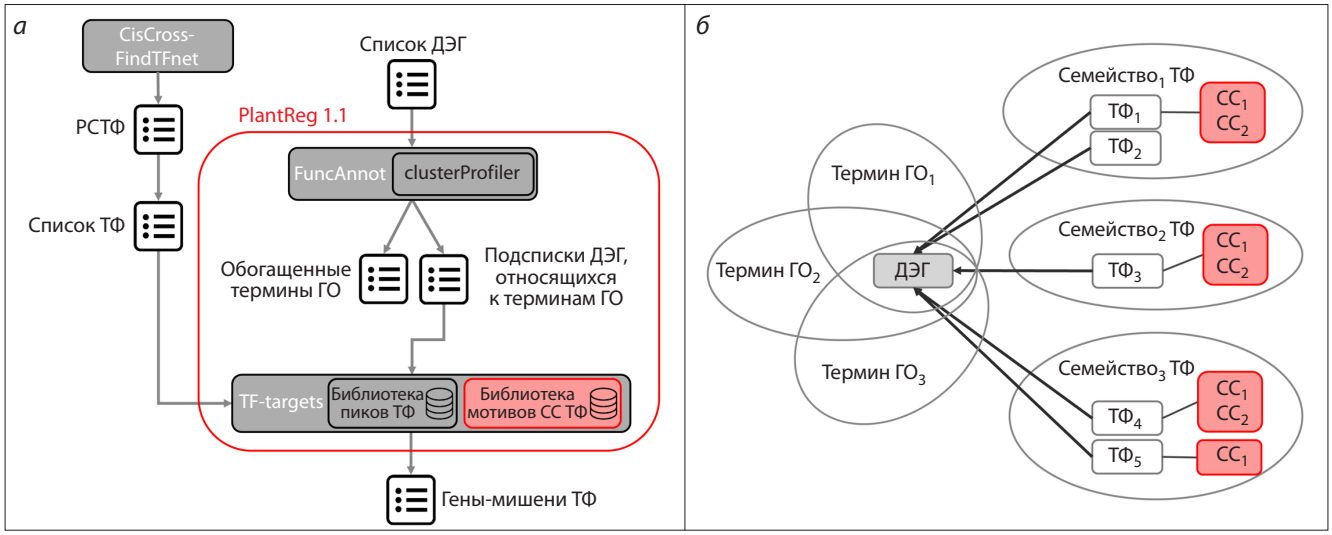


Рис. 3. Схема работы программы PlantReg 1.1 (*а*) и структура блока выдачи 1 (*б*).

Розовым цветом выделены обновления в версии 1.1 программы PlantReg по сравнению с исходной версией.

этой функциональной аннотации гена, количеством этих терминов ГО, списком потенциальных транскрипционных регуляторов гена с указанием семейства для каждого ТФ, а также количества транскрипционных факторов.

В блоке выдачи 4 эта же информация представлена в альтернативном формате, в котором списки терминов ГО и транскрипционных регуляторов каждого гена разбиты построчно. В программе PlantReg 1.1 к характеристике каждого гена добавлены нуклеотидная последовательность СС ТФ, распознанного в районе связывания соответствующего ТФ, геномные координаты СС ТФ (блок 4) или координаты СС ТФ относительно точки старта транскрипции (блок 1) и указание цепи ДНК, в которой распознан СС ТФ (см. рис. 3, *б*). Информация по локализации СС ТФ доступна только при выборе в качестве параметра коллекции полногеномных профилей CisCross-MACS2.

Распознавание СС ТФ в 5′-регуляторных районах генов *A. thaliana*. Частотные матрицы, описывающие СС ТФ *A. thaliana*, реконструировали путем *de novo* поиска мотивов в пиках DAP-seq из коллекции CisCross-MACS2, представленной в веб-версии программы PlantReg (Lavrekha et al., 2024). Коллекция наборов пиков CisCross-MACS2 была получена ранее (Lavrekha et al., 2022) в результате обработки сырых данных полногеномного профилирования сайтов связывания 403 ТФ *A. thaliana* методом DAP-seq (O'Malley et al., 2016). В каждом наборе пиков выбирали лучшие по высоте 2000 пиков, которые использовали для *de novo* поиска мотивов с помощью программы STREME (Bailey, 2021).

Контрольную выборку генерировали с помощью программы AntiNoise (Raditsa et al., 2024). Полагали, что мотив с наилучшей значимостью обогащения в пиках (при *p*-value не более 0.05) описывает сайт связывания целевого ТФ. Для проверки этого предположения устанавливали соответствие выявленных мотивов известным СС ТФ путем их сравнения с мотивами из баз данных JASPAR2024 CORE (Rauluseviciute et al., 2024), CisBP (Weirauch et al., 2014) и ArabidopsisDAPv1 (O'Malley et al., 2016) с помощью программы Tomtom (Gupta et al., 2007).

Поиск потенциальных СС ТФ в 5′-регуляторных областях генов *A. thaliana* (район [–2500; +1) относительно старта транскрипции) осуществляли методом позиционных весовых матриц с помощью функции scan_sequence пакета universalmotif на языке программирования R (Tremblay, 2024). Для извлечения нуклеотидных последовательностей 5′-регуляторных районов использовали версию генома *A. thaliana* TAIR10 (Lamesch et al., 2012) и геномную аннотацию Araport11 (Cheng et al., 2017).

Поиск примеров конкурентного подавления или активации генов в выдаче программы PlantReg 1.1. Установление регуляторных связей между элементами РСТФ-А/С (Lavrekha et al., 2024; Omelyanchuk et al., 2024) и генами, обеспечивающими протекание биологических процессов, изменяющихся под влиянием ауксина и солевого стресса, а также последующее выявление конкурентного подавления или активации генов под действием ауксина и солевого стресса осуществляли с помощью программы PlantReg 1.1. В качестве исходных данных были взяты списки ТФ, которые входили в состав РСТФ-А (39 элементов) и РСТФ-С (19 элементов) (на рис. 4 списки обозначены

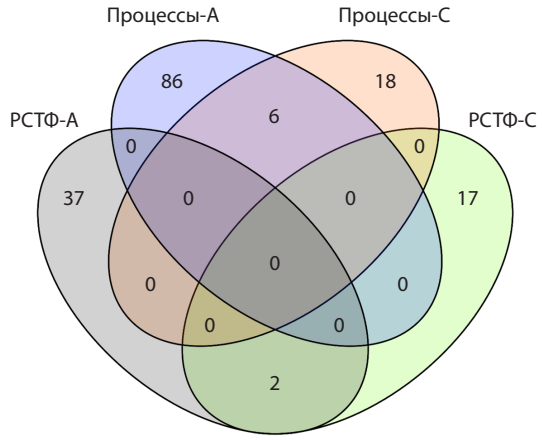


Рис. 4. Количество генов, кодирующих ТФ, которые формируют РСТФ-А и РСТФ-С, а также генов-мишеней, регулируемых этими РСТФ, которые обеспечивают протекание биологических процессов, изменяющихся под влиянием ауксина (Процессы-А) или солевого стресса (Процессы-С).

Множество «Процессы-А» включает в себя гены биосинтеза хлорофилла, биосинтеза лигнина, транспорта, конъюгации и сигнального пути АБК, биогенеза рибосом. Множество «Процессы-С» включает в себя гены транспорта, конъюгации и сигнального пути АБК.

как «РСТФ-А» и «РСТФ-С») (Приложение, табл. S1)[1], а также списки ДЭГ, активируемых (аДЭГ) и подавляемых (пДЭГ) под действием ауксина (5201 аДЭГ и 6704 пДЭГ) или солевого стресса (1476 аДЭГ и 944 пДЭГ), которые ранее использовались для реконструкции РСТФ (De Rybel et al., 2012; Wu et al., 2021; Omelyanchuk et al., 2024).

Списки аДЭГ и пДЭГ из двух транскриптомных экспериментов по отдельности подавали на вход программы PlantReg 1.1 вместе с соответствующим списком ТФ из РСТФ-А или РСТФ-С. Порог на обогащение терминами ГО устанавливали на уровне 0.001, для локализации районов связывания ТФ выбирали коллекцию полногеномных профилей CisCross-MACS2 и длину 5′-регуляторного района 1000 п. н. В результате получали пары «ТФ-регулятор–ген-мишень», где в качестве ТФ-регуляторов выступали ТФ из РСТФ-А или РСТФ-С, а в качестве генов-мишеней – аДЭГ и пДЭГ из соответствующего транскриптомного эксперимента.

Данные DAP-seq, используемые в PlantReg 1.1 для картирования районов связывания ТФ в геноме *A. thaliana*, содержат два типа наборов пиков: в первом случае (наборы пиков "col") для приготовления библиотек использовалась нативная геномная ДНК из листьев, во втором случае (наборы пиков "colamp") – геномная ДНК, в которой метилцитозиновые эпигенетические метки были удалены путем ПЦР-амплификации. Из выдачи программы PlantReg 1.1 отбирали пары «ТФ-регулятор–ген-мишень», реконструированные с использованием DAP-seq профилей связывания ТФ группы "col". Далее среди регулируемых РСТФ-А генов-мишеней были отобраны гены, аннотированные терминами ГО, связанными с биосинтезом хлорофилла (16 генов), биосинтезом лигнина (14 генов), транспортом, конъюгацией и сигнальным путем АБК (34 генов) и био-

[1] Табл. S1–S7 Приложения см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx31.xlsx

генезом рибосом (28 генов) – процессами, которые мы рассматривали ранее в работе (Omelyanchuk et al., 2024).

Среди генов-мишеней, регулируемых РСТФ-С, были отобраны гены, аннотированные терминами ГО, связанными с сигнальным путем АБК (24 гена), который мы рассматривали ранее в работе (Lavrekha et al., 2024). В результате было выбрано 110 генов (на рис. 4 обозначены как «Процессы-А» и «Процессы-С») (см. табл. S1–S3). Из них в качестве потенциальных объектов конкурентного подавления или активации мы выбирали гены, удовлетворяющие следующим требованиям: а) в регуляции таких генов участвовало более одного ТФ; б) СС этих ТФ существенно (не менее 80 %) перекрывались; в) гены, кодирующие эти ТФ, характеризовались противоположным изменением экспрессии в транскриптомном эксперименте.

Результаты

Коллекция предсказанных СС ТФ в 5′-регуляторных районах генов *A. thaliana*, интегрированная в программу PlantReg 1.1

Чтобы обеспечить возможность прогнозирования кооперативного и конкурентного взаимодействия ТФ в транскрипционной регуляции биологических процессов, в программе PlantReg 1.1 была реализована автоматическая локализация СС ТФ в 5′-регуляторных районах генов. Для этого результаты распознавания СС ТФ в промоторах генов методом позиционной весовой матрицы (см. раздел «Материалы и методы») были систематизированы и интегрированы в программу PlantReg 1.1. Для 300 ТФ (74 %) мотив, реконструированный *de novo* хотя бы по одному набору пиков (“col” или “colamp”), был сходен с известным СС этого ТФ, доступным в базах данных JASPAR, CisBP или ArabidopsisDAPv1 (рис. 5, *а*). Доля ТФ, у которых СС распознаны более чем в 90 % пиков, картирующихся в 5′-регуляторные районы генов, достаточно высока и варьирует от 42 % (при длине 5′-регуляторного района 500 п. н.) до 74 % (при длине 5′-регуляторного района 2000 п. н.) (см. рис. 5, *б*).

В последующих разделах мы проиллюстрировали возможности применения нового функционала PlantReg 1.1 для решения конкретных биологических задач.

Конкурентная регуляция экспрессии генов в ответе на ауксин и солевой стресс у *A. thaliana*

Выше сформулировано предположение о том, что подавление транскрипции генов-мишеней при повышении уровня ТФ типа US, а также их активация на фоне снижения уровня ТФ типа DS в ответе на ауксин и под действием солевого стресса могут происходить путем конкурентной регуляции парой ТФ-активаторов. Чтобы проверить это предположение, с помощью программы PlantReg 1.1 были установлены регуляторные связи между РСТФ-А/С и генами, регулирующими биосинтез хлорофилла и лигнина, биогенез рибосом и ответ на АБК. Показано, что 14 генов являются потенциальными объектами конкурентной регуляции транскрипционными факторами из РСТФ-А и РСТФ-С (см. табл. S1, S6, S7).

Примечательно, что 11 генов, кодирующих ТФ в составе РСТФ-А и РСТФ-С, также могут быть подвержены конкурентной регуляции (см. табл. S4, S5). Выявленные 25 генов (12 пДЭГ, 10 аДЭГ и 3 гена, *ABCG25* (*ATP-binding cassette family G25*), *GBF3* (*G-box binding factor 3*) и *PYL7/RCAR2* (*PYR1-like 7/Regulatory components of ABA receptor 2*), экспрессия которых под действием ауксина и солевого стресса изменялась разнонаправленно), составляют 32 % от всей выборки генов, регулируемых репрессорами (79 генов, см. табл. S1, S6, S7). Таким образом, частота случаев конкурентной регуляции посредством РСТФ-А и РСТФ-С может быть достаточно высока.

Транскрипционные факторы объединяются в семейства, классы, суперклассы на основании сходства их доменов

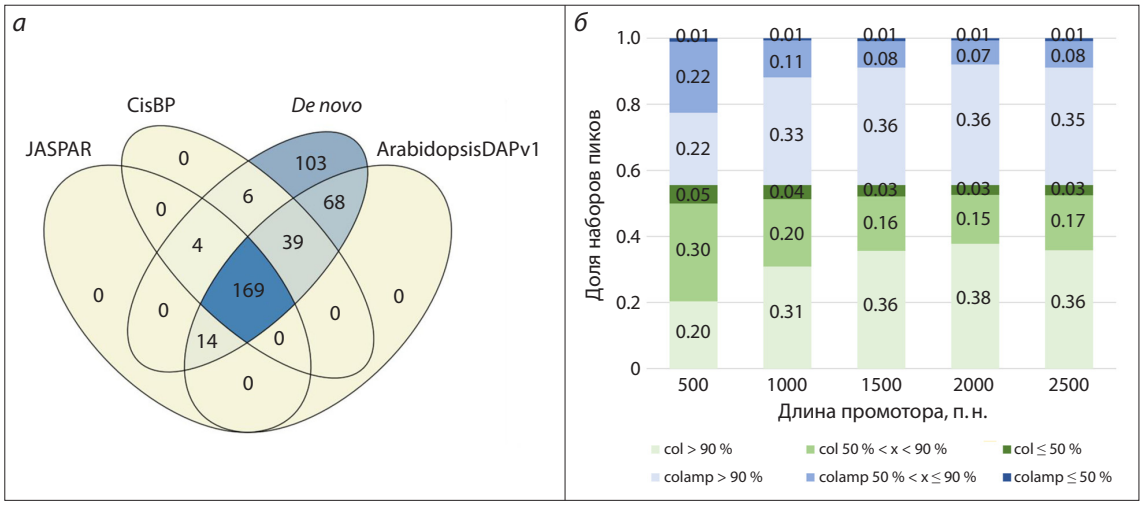


Рис. 5. Характеристика интегрированной в PlantReg 1.1 коллекции предсказанных СС ТФ в 5′-регуляторных районах генов *A. thaliana*.

а – результаты сравнения мотивов, найденных *de novo* в пиках DAP-seq, с известными СС целевых ТФ в базах данных JASPAR, CisBP и ArabidopsisDAPv1; *б* – доли наборов пиков DAP-seq (col – оттенки зеленого/colamp – оттенки синего), в которых мотив распознан более чем в 90 % пиков (светлый оттенок), в 50–90 % пиков и менее чем в 50 % пиков (темный оттенок), картирующихся в 5′-регуляторных районах генов.

связывания с ДНК (Blanc-Mathieu et al., 2024). ТФ из одного семейства часто распознают схожие последовательности ДНК и, следовательно, могут конкурировать за эти сайты связывания. В 5′-регуляторных областях выявленных нами 25 генов могут конкурировать ТФ внутри шести семейств: AP2/ERF (APETALA2/ETHYLENE RESPONSIVE FACTOR), bZIP (BASIC LEUCINE-ZIPPER), BZR1/BES1 (BRASSINAZOLE RESISTANT 1/BRI1 EMS SUPPRESSOR 1), HD-ZIP (HOMEODOMAIN LEUCINE ZIPPER), MYB (V-MYB AVIAN MYELOBLASTOSIS VIRAL ONCOGENE HOMOLOG), WRKY (см. табл. S6). Кроме этого, выявлены случаи возможной конкуренции ТФ из разных семейств одного суперкласса, а именно: “basic domains” и “Helix-Turn-Helix domains” (см. табл. S6).

Более того, в промоторах аДЭГ *MAPKKK18* (*Mitogen-activated protein kinase kinase kinase 18*) и *RRP47* (*Sas10/Utp3/C1D family*) один и тот же СС могут занимать ТФ из семейств двух разных суперклассов: AP2/ERF (суперкласс “Beta-hairpin exposed by an alpha/beta-scaffold”) и bZIP (суперкласс “basic domains”) в первом случае и AP2/ERF и LBD (суперкласс “Zinc-coordinating DNA binding domains”) во втором (см. табл. S6). В дистальном промоторе пДЭГ *GBF3* за СС могут конкурировать ТФ из семейств двух других суперклассов, bZIP (“basic domains”) и MYB (“Helix-Turn-Helix domains”). Были обнаружены случаи конкуренции за СС среди ТФ из трех разных семейств, принадлежащих двум (в промоторах *AFP1* (*ABI five binding protein*), *MYB73* и *PYL7*) и даже трем разным суперклассам (в дистальном промоторе *GBF3*) (см. табл. S6).

Конкуренция ТФ семейства bZIP в промоторах генов, регулируемых РСТФ-А

Чтобы выявить комбинации ТФ-активаторов, которые систематически задействуют РСТФ-А или РСТФ-С для подавления экспрессии генов-мишеней, был проведен сравнительный анализ пар «ТФ-регулятор–ген-мишень», определенных с помощью PlantReg 1.1. Показано, что три ТФ типа DA (bZIP3, bZIP68 и GBF3) и ТФ типа US (bZIP53) имеют общие СС в промоторах целого ряда генов, регулируемых РСТФ-А. К ним относятся гены *CHLG* (*Chlorophyll G*) (рис. 6, *а*, *б*), *HEME2* (*AT5G14220*) и *CH1* (*Chlorina 1*), кодирующие ферменты биосинтеза хлорофилла, а также *ABCG25*, кодирующий белок, транспортирующий АБК через клеточную мембрану (см. табл. S6 и S7).

Поскольку в литературе bZIP53 описывается как транскрипционный активатор (Alonso et al., 2009; Weltmeier et al., 2009), логично предположить, что подавление экспрессии упомянутых выше генов может быть следствием конкуренции ТФ семейства bZIP за общие СС в промоторах, в результате которой сильный активатор заменяется более слабым. Действительно, активность трансактивационных

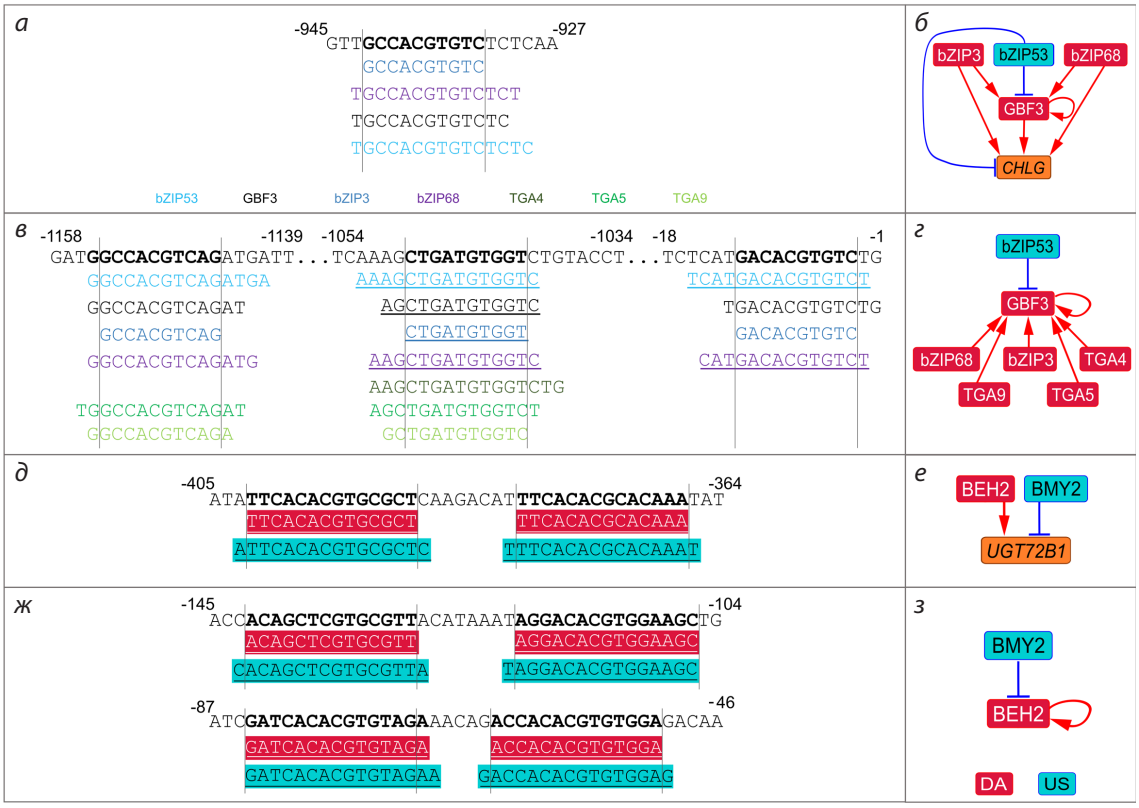


Рис. 6. Перекрывание СС ТФ в промоторах генов-мишеней, выявленное с помощью PlantReg 1.1.

а – дистальная область промотора *CHLG* с перекрыванием СС ТФ семейства bZIP (bZIP3, bZIP53, bZIP68 и GBF3); *в* – дистальная и коровая области промотора *GBF3* с перекрыванием СС ТФ семейства bZIP (bZIP3, bZIP53, bZIP68, TGA4/5/9 и GBF3), цветовая кодировка ТФ для (*в*) такая же, как для (*а*); *д*, *ж* – проксимальные части промоторов *UGT72B1* и *BEH2* соответственно, с перекрыванием СС ТФ BEH2 (красный цвет заливки) и BMY2 (синий цвет заливки); *б*, *г*, *е*, *з* – схемы транскрипционной регуляции генов *CHLG*, *GBF3*, *UGT72B1* и *BEH2* соответственно. Нижнее подчеркивание – СС лежит на антисмысловой цепи по отношению к цепи гена. Координаты приведены относительно точки старта транскрипции. US – активируемый супрессор; DA – подавляемый активатор.

доменов этих ТФ исследована ранее в работе (Hummel et al., 2023), где показано, что bZIP53 – транскрипционный активатор, однако гораздо более слабый, чем представители этого же семейства bZIP3, bZIP68 и GBF3.

Примечательно, что аналогичная комбинация транскрипционных регуляторов, конкурирующих за общие CC (bZIP3, bZIP68 и GBF3 в качестве DA, bZIP53 в качестве US), обнаружена нами в промоторах подавляемых ауксином генов *ERF15*, *GBF3* (см. рис. 6, *в*, *г*) и *AT1G19000*, кодирующих ТФ из РСТФ-А (см. табл. S6 и S7). Таким образом, конкуренция ТФ семейства bZIP за общие CC, вероятно, является значимым механизмом подавления транскрипции в ответ на ауксин.

Заслуживает внимания тот факт, что замена активаторов bZIP3, bZIP68 и GBF3 более слабым bZIP53 обнаружена нами и в промоторе гена *GBF3*, кодирующего одного из участников конкурентной регуляции (см. рис. 6, *в*, *г*). Сходная ситуация выявлена для пары ТФ BEH2 (BES1/BZR1 HOMOLOG 2) (DA) и BMY2 (BETA-AMYLASE 2, известен также как BETA-AMYLASE 8/BAM8) (US), принадлежащих семейству BZR1/BES1. Эти ТФ регулируют экспрессию не только *DFB* и *UGT72B1* – генов, контролирующих уровень лигнина, но и *BEH2* (см. рис. 6, *д–з*).

Теоретически такая обратная связь может выступать «спусковым механизмом» для широкомасштабного конкурентного подавления экспрессии генов-мишеней парой ТФ-активаторов: повышение представленности более слабого транскрипционного активатора ведет к конкурентному подавлению экспрессии гена, кодирующего более сильный активатор транскрипции (и являющегося общей мишенью обоих ТФ), в результате чего ингибирование остальных общих мишеней будет увеличиваться. Следует также отметить, что некоторые ТФ типа DA могут конкурировать между собой за общий CC в рамках функционирования П-подсети до обработки ауксином (см. рис. 6, *а–г*), что, вероятно, обеспечивает дополнительную регуляторную гибкость РСТФ-А.

Конкурентная регуляция генов сигнального пути АБК в рамках РСТФ-А и РСТФ-С

В рассматриваемых нами примерах и ауксин, и солевой стресс модулируют ответ на АБК, в первом случае он ослабляется, во втором – усиливается (Lavrekha et al., 2024; Omelyanchuk et al., 2024). Сопоставление регуляторных связей, реконструированных на основании данных из разных экспериментов, позволяет более полно охарактеризовать потенциальные механизмы транскрипционной регуляции отдельных генов. С помощью программы PlantReg 1.1 мы показали, что три гена пути передачи сигнала АБК (*PYL7*, *AFP1* и *ABCG25*) находятся под контролем обеих регуляторных сетей транскрипционных факторов.

В снижение активности гена *PYL7* ауксином и повышение его активности солевым стрессом вовлечены разные ТФ, которые связываются с разными сайтами в его промоторе (рис. 7, *а*, *б*). Это позволяет предположить, что ответы на ауксин и солевой стресс используют различные молекулярные механизмы конкурентной регуляции для модуляции экспрессии *PYL7*. В регуляции *AFP1* и *ABCG25*, напротив, оба стимула могут задействовать один и тот же блок конкурирующих активаторов, но по-разному. Так, в норме экспрессию гена *AFP1* обеспечивает среди прочих bZIP68. Под действием ауксина он замещается BMY2 (вероятно, более слабым активатором), в условиях солевого стресса, напротив, его заменяет BIM2 (BES1-interacting Myc-like protein 2), который, согласно работе (Hummel et al., 2023), является более сильным активатором, чем bZIP68 (см. рис. 7, *в*, *г*). В промоторе *ABCG25* под действием ауксина слабый активатор bZIP53 замещает более сильные активаторы семейства bZIP, что вносит вклад в уменьшение количества транскриптов (см. рис. 7, *д*, *е*). Солевой стресс модулирует соотношение этих же активаторов, но в этом случае на фоне падения экспрессии *bZIP3* увеличивается количество транскриптов *GBF3*.

Интересно, что в промоторе гена *GBF3*, кодирующего ТФ, который входит в состав обеих РСТФ, наблюдается аналогичная картина. При солевом стрессе, в результате которого *GBF3* активируется, ТФ GBF3 замещает bZIP3 на трех сайтах в проксимальном (–116; +1) (рис. 8, *а*, *б*) и семи сайтах связывания в дистальном (–1312; –701) районах промотора *GBF3* (см. рис. 8, *в–е*), тем самым, судя по всему, усиливая свою самоактивацию. Под действием ауксина происходит другое перераспределение ТФ семейства bZIP на этих же сайтах (см. рис. 8). Эти результаты хорошо согласуются с представлениями о важной роли конкуренции ТФ семейства bZIP за сайты связывания в модуляции экспрессии генов (Schindler et al., 1992; Foster et al., 1994; Ko, Brandizzi, 2022). В то же время ответ на ауксин включает и специфические механизмы регуляции *GBF3*, которые не задействованы в ответе на солевой стресс. Так, MYB3R1 после обработки ауксином может замещать сайты MYB70 и MYB73.

Обсуждение

В нашей работе собрана, систематизирована и интегрирована в программу PlantReg 1.1 информация о потенциальных CC ТФ в промоторах генов *A. thaliana*. Это позволяет, опираясь на данные о типе активности трансактивационных доменов ТФ (Hummel et al., 2023), прогнозировать кооперативное и конкурентное взаимодействие ТФ, входящих в состав РСТФ, при транскрипционной регуляции биологических процессов. Ранее мы реконструировали две РСТФ, контролирующие ответ на ауксин и солевой стресс у *A. thaliana*, и показали, что в обоих случаях важную роль играет подавление транскрипции (Lavrekha et al., 2024; Omelyanchuk et al., 2024). Однако, согласно литературным данным, большинство предсказанных супрессоров в составе РСТФ имеет трансактивационный домен активаторного типа (Hummel et al., 2023; Omelyanchuk et al., 2024). Мы применили PlantReg 1.1 для определения молекулярных механизмов, обеспечивающих функционирование ТФ-активаторов в качестве транскрипционных репрессоров.

Удалось обнаружить, что более трети генов-мишеней ТФ, предсказанных в РСТФ как репрессоры, потенциально могут быть объектами конкурентной регуляции парой ТФ, один из которых является сильным активатором транскрипции, другой – слабым. Таким образом, конкурентная регуляция экспрессии генов, вероятно, – универсальный механизм, позволяющий модулировать уровень экспрес-

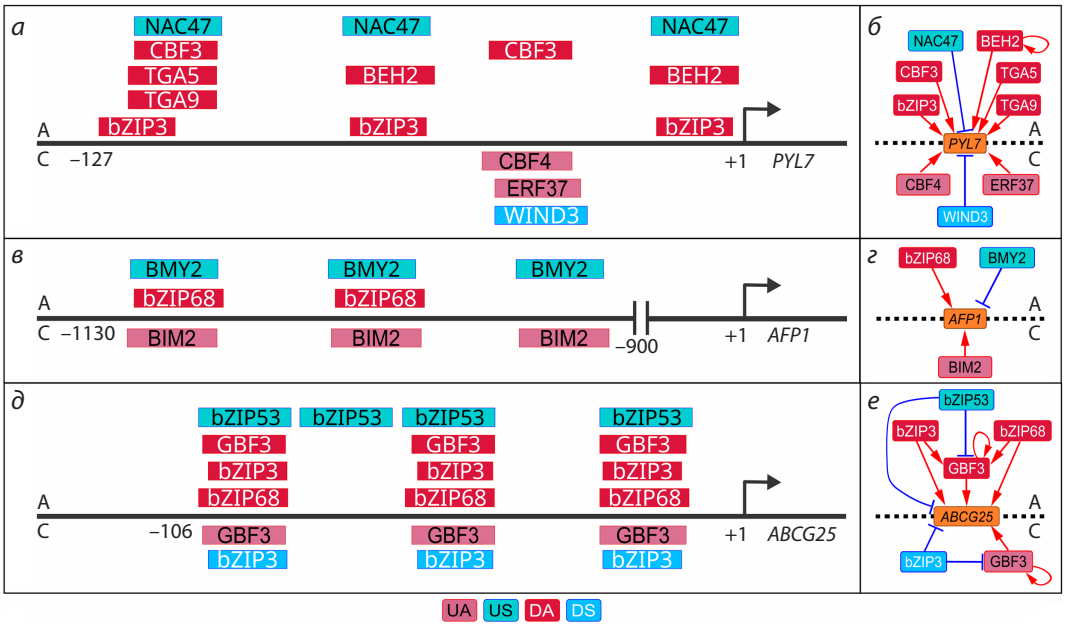


Рис. 7. Перекрывание СС ТФ в промоторах генов-мишеней при обработке ауксином и при раннем солевом стрессе, выявленное с помощью PlantReg 1.1.

а – проксимальная часть промотора *PYL7*; *в* – дистальная часть промотора *AFP1*; *д* – проксимальная часть промотора *ABCG25*; *б*, *г*, *е* – схемы транскрипционной регуляции генов *PYL7*, *ATF1* и *ABCG25* соответственно. Для каждой панели детали регуляции в ответ на ауксин (A) расположены вверху, а реакции на ранний солевой стресс (C) – внизу. СС ТФ представлены прямоугольниками в соответствии с цветовой кодировкой типа регуляции: UA – активируемый активатор; US – активируемый супрессор; DA – подавляемый активатор; DS – подавляемый супрессор.

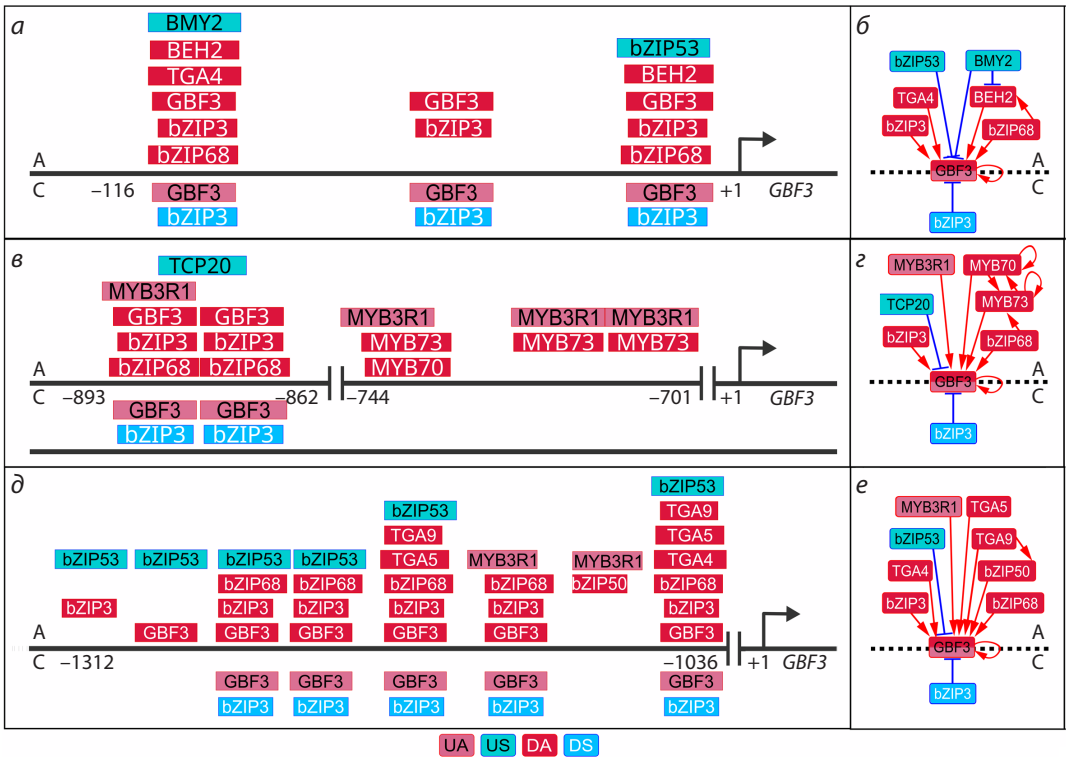


Рис. 8. Перекрывание СС ТФ в промоторе *GBF3* при обработке ауксином и при раннем солевом стрессе, выявленное с помощью PlantReg 1.1.

а, *в*, *д* – проксимальная (–116; +1) и дистальная (–1312; –701) части промотора *GBF3* с перекрыванием СС ТФ; *б*, *г*, *е* – схемы транскрипционной регуляции *GBF3* в проксимальной и дистальной областях промотора. Для каждой панели детали регуляции в ответ на ауксин (A) расположены вверху, а реакции на ранний солевой стресс (C) – внизу. СС ТФ представлены прямоугольниками в соответствии с цветовой кодировкой типа регуляции: UA – активируемый активатор, US – активируемый супрессор, DA – подавляемый активатор, DS – подавляемый супрессор.

сии генов при реализации ответа на ауксин и солевой стресс у *A. thaliana*.

Ауксин – важнейший регулятор большинства процессов в растении, вовлечен в регуляцию переключения программ развития (Vanneste et al., 2025). Наиболее стандартным представлениям о регуляции переключения процессов соответствует замещение репрессора активатором, как, например, замещение репрессорного комплекса E2F ТФ на E2F активаторы перед началом клеточного цикла в промоторах многих генов растений и животных (van den Heuvel, Dyson, 2008; Sánchez-Camargo et al., 2021) или наоборот замещение транскрипционного активатора репрессором (Berlow et al., 2017). Однако результаты, полученные с помощью PlantReg 1.1, свидетельствуют о том, что в ответе на ауксин для подавления транскрипции вместо канонической смены активатор–репрессор может активно использоваться замена сильного активатора более слабым.

Как минимум некоторые предсказанные с помощью PlantReg 1.1 случаи замены сильного активатора более слабым поддерживаются ранее опубликованными результатами научных исследований. К ним относится, например, замещение трех активаторов, bZIP3, bZIP68 и GBF3, более слабым активатором bZIP53 при подавлении ауксином экспрессии генов биосинтеза хлорофилла *CHLG*, *HEME2* и *CH1* (Hummel et al., 2023). Конкуренция между ТФ семейства bZIP за общий сайт связывания и ее влияние на экспрессию гена-мишени ранее описаны для многих ТФ из этого семейства (Schindler et al., 1992; Foster et al., 1994; Ko, Brandizzi, 2022). Известно также, что ряд ТФ семейства bZIP взаимозаменяемо и комплексно регулирует биосинтез хлорофилла. В частности, у тройного мутанта *gbf1 gbf2 gbf3* биосинтез хлорофилла нарушен, что демонстрирует важную роль GBF факторов в этом процессе (Sun T. et al., 2025). Сверхэкспрессия другого представителя семейства, *bZIP1*, приводит к уменьшению уровня хлорофилла, а менее выраженное снижение уровня хлорофилла и снижение уровня транскрипта *CHLG* по сравнению с диким типом (p_{adj} = 0.03) наблюдаются у двойного мутанта *bzip1 bzip53* (Wildenhain et al., 2025).

Специфичное для растений семейство ТФ BZR1/BES1 опосредует транскрипционный ответ на брассиностероиды (стероидные гормоны растений). В этом семействе наряду с ТФ BZR1 и BES1 представлены еще четыре ТФ, называемые гомологами BES1: BEH1, BEH2, BEH3, BEH4 (Shi et al., 2022). Сравнительно недавно к семейству BZR1/BES1 добавились еще два необычных ТФ, BAM7 и BMY2, сходных с β-амилазами, но демонстрирующих также очень слабую гомологию с BES1 (Reinhold et al., 2011). Эти ТФ не обладают амилазной каталитической активностью, но имеют домены типа BZR1, связывающиеся с сайтами, узнаваемыми ТФ семейства BZR1/BES1. BMY2 является активатором транскрипции, в то время как BAM7 регулирует его активность.

Ранее было высказано предположение, что BMY2 контролирует транскрипцию генов-мишеней, противодействуя их регуляции другими ТФ семейства BZR1/BES1 посредством конкуренции с ними за сайты связывания (Reinhold et al., 2011). Согласно результатам, полученным с помощью PlantReg 1.1, именно это может происходить в промоторах ряда подавляемых ауксином генов, в том числе гена *UGT72B1* (*UDP-glucose-dependent glucosyltransferase 72 B1*), кодирующего фермент, конъюгирующий монолигнолы, где BMY2 (вероятно, как более слабый активатор) конкурирует с BEH2 (см. рис. 6, *д*, *е*; табл. S6 и S7).

Более детальный анализ генов-мишеней пары ТФ BEH2 и BMY2, предсказанных с помощью PlantReg 1.1, позволяет добавить в описание механизма регуляции ауксином биосинтеза лигнина несколько важных деталей. Ауксин, активируя *BMY2*, препятствует самоактивации *BEH2* путем конкурентного замещения ТФ BEH2 менее активным BMY2 на их общих сайтах связывания (см. рис. 5, *ж*, *з*). Это приводит к снижению уровня BEH2 в ядре, что в свою очередь способствует замещению этого ТФ на его сайтах в промоторе гена *UGT72B1* более слабым активатором BMY2 (см. рис. 5, *д*, *е*) и, как следствие, обусловливает снижение количества транскрипта *UGT72B1*. То, что BMY2 активирует экспрессию *UGT72B1*, было показано ранее по повышению уровня его транскрипта при сверхэкспрессии *BMY2* и его снижению в двойном мутанте *bmy2 bam7* (Reinhold et al., 2011).

Примечательно, что ауксин подавляет транскрипцию большинства генов, кодирующих ферменты биосинтеза лигнина (Omelyanchuk et al., 2024), снижая таким образом биосинтез монолигнолов. Одновременно ауксин подавляет экспрессию *UGT72B1* и, как следствие, ингибирует конъюгацию монолигнолов, частично компенсируя таким образом снижение уровня монолигнолов. Интересно отметить, что брассиностероиды также модулируют уровень лигнина через ТФ BEH2. Брассиностероиды усиливают биосинтез лигнина через активацию большинства ферментов его биосинтеза (Percio et al., 2025). Одновременно они подавляют BEH2 с помощью GSK3 (GLYCOGEN SYNTHASE KINASE 3) подобных киназ и BES1 (Otani et al., 2022). Поскольку BEH2 стимулирует транскрпцию *UGT72B1*, который конъюгирует монолигнолы, брассиностероиды препятствуют оттоку монолигнолов из биосинтеза лигнина, тем самым дополнительно повышая уровень лигнина.

Данные, получаемые с помощью PlantReg 1.1, позволяют сформулировать конкретные гипотезы для планирования дальнейших экспериментальных исследований. Необходимо отметить, что прогноз может содержать ложноположительные результаты. Например, в паре ТФ HB21 (DA) и HB40 (US) из РСТФ-А, связывающих одни и те же сайты в промоторе подавляемого ауксином гена *bZIP50*, именно HB40 – более сильный активатор. Это значит, что конкуренция за СС с HB21 не может объяснить подавление экспрессии гена-мишени при повышении представленности HB40. Не исключено, что *HB21* и *HB40* экспрессируются в разных тканях или на разных стадиях развития. Чтобы объяснить, почему HB40, являющийся по природе активатором, может функционировать как репрессор, нужно исследовать возможности привлечения корепрессоров этим транскрипционным фактором.

Заключение

Программа PlantReg 1.1 предназначена для установления регуляторных связей между РСТФ и генами, которые обеспечивают протекание биологических процессов,

контролируемых этими сетями. В обновленной версии программа имеет функционал, позволяющий точно локализовать СС ТФ в промоторах генов-мишеней. Благодаря этому можно осуществлять анализ взаимного расположения СС ТФ и, привлекая данные о типе активности трансактивационных доменов этих ТФ, выявлять потенциальные случаи кооперативного или конкурентного действия ТФ в промоторе конкретного гена.

Программа PlantReg 1.1 была успешно использована для реконструкции транскрипционных механизмов регуляции биосинтеза хлорофилла и лигнина, биогенеза рибосом и ответа на АБК под действием ауксина и солевого стресса. Анализ взаимного расположения СС ТФ показал, что активность целого ряда генов, регулирующих эти процессы, может подавляться в результате конкуренции пары ТФ-активаторов за общий СС, когда более слабый активатор приходит на место более сильного. Часть полученных результатов поддержана литературными данными.

Таким образом, результаты, получаемые с помощью PlantReg 1.1, позволяют сформулировать гипотезы для планирования экспериментальных исследований. Следует, однако, подчеркнуть, что прогноз может содержать ложноположительные результаты. Сокращение их количества – одно из возможных направлений дальнейшего развития программы.

Список литературы / References

Ahn J.H., Miller D., Winter V.J., Banfield M.J., Jeong H.L., So Y.Y., Henz S.R., Brady R.L., Weigel D. A divergent external loop confers antagonistic activity on floral regulators FT and TFL1. *EMBO J.* 2006;25(3):605-614. doi 10.1038/sj.emboj.7600950

Alonso R., Oñate-Sánehez L., Weltmeier F., Ehlert A., Diaz I., Dietrich K., Vicente-Carbajosa J., Dröge-Laser W. A pivotal role of the basic leucine zipper transcription factor bZIP53 in the regulation of *Arabidopsis* seed maturation gene expression based on heterodimerization and protein complex formation. *Plant Cell.* 2009;21(6): 1747-1761. doi 10.1105/tpc.108.062968

Bailey T.L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*. 2021;37(18):2834-2840. doi 10.1093/bioinformatics/btab203

Berlow R.B., Dyson H.J., Wright P.E. Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature.* 2017; 543(7645):447-451. doi 10.1038/nature21705

Blanc-Mathieu R., Dumas R., Turchi L., Lucas J., Parcy F. Plant-TFClass: a structural classification for plant transcription factors. *Trends Plant Sci.* 2024;29(1):40-51. doi 10.1016/j.tplants.2023.06.023

Boyle P., Després C. Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. *Plant Signal Behav.* 2010;5(6):629-634. doi 10.4161/psb.5.6.11570

Brackmann K., Qi J., Gebert M., Jouannet V., Schlamp T., Grünwald K., Wallner E.-S., Novikova D.D., Levitsky V.G., Agustí J., Sanchez P., Lohmann J.U., Greb T. Spatial specificity of auxin responses coordinates wood formation. *Nat Commun.* 2018;9(1):875. doi 10.1038/s41467-018-03256-2

Cheng C.Y., Krishnakumar V., Chan A.P., Thibaud-Nissen F., Schobel S., Town C.D. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. 2017;89(4):789-804. doi 10.1111/tpj.13415

Chupreta S., Brevig H., Bai L., Merchant J.L., Iñiguez-Lluhí J.A. Sumoylation-dependent control of homotypic and heterotypic synergy by the Krüppel-type zinc finger protein ZBP-89. *J Biol Chem.* 2007;282(50):36155-36166. doi 10.1074/jbc.M708130200

De Rybel B., Audenaert D., Xuan W., Overvoorde P., Strader L.C., Kepinski S., Hoye R., Brisbois R., Parizot B., Vanneste S., Liu X. A role for the root cap in root branching revealed by the non-auxin probe naxillin. *Nat Chem Biol*. 2012;8(9):798-805. doi 10.1038/nchembio.1044

Deshpande D., Chhugani K., Chang Y., Karlsberg A., Loeffler C., Zhang J., Muszyńska A., … Eskin E., Zhao F., Mohammadi P., Łabaj P., Mangul S. RNA-seq data science: from raw data to effective interpretation. *Front Genet.* 2023;14:997383. doi 10.3389/fgene.2023.997383

Dhatterwal P., Sharma N., Prasad M. Decoding the functionality of plant transcription factors. *J Exp Bot.* 2024;75(16):4745-4759. doi 10.1093/jxb/erae231

Foster R., Izawa T., Chua N.H. Plant bZIP proteins gather at ACGT elements. *FASEB J.* 1994;8(2):192-200. doi 10.1096/fasebj.8.2.8119490

Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. Quantifying similarity between motifs. *Genome Biol*. 2007;8(2):R24. doi 10.1186/gb-2007-8-2-r24

Hanna-Rose W., Hansen U. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* 1996;12(6):229-234. doi 10.1016/0168-9525(96)10022-6

Huang W., Quan M., Qi W., Xiao L., Fang Y., Zhou J., Ren T., Li P., Chen Y., El-Kassaby Y.A., Du F., Zhang D. Phylostratigraphic analysis revealed that ancient ohnologue *PtoWRKY53* innovated a vascular transcription regulatory network in *Populus*. *New Phytol*. 2025;248:2295-2315. doi 10.1111/nph.70403

Hummel N.F.C., Zhou A., Li B., Markel K., Ornelas I.J., Shih P.M. The trans-regulatory landscape of gene networks in plants. *Cell Syst.* 2023;14(6):501-511.e4. doi 10.1016/j.cels.2023.05.002

Ko D.K., Brandizzi F. Network-based approaches for understanding gene regulation and function in plants. *Plant J.* 2020;104(2):302-317. doi 10.1111/tpj.14940

Ko D.K., Brandizzi F. Transcriptional competition shapes proteotoxic ER stress resolution. *Nat Plants.* 2022;8(5):481-490. doi 10.1038/s41477-022-01150-w

Lamesch P., Berardini T.Z., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., … Nelson W.D., Ploetz L., Singh S., Wensel A., Huala E. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 2012;40(D1): D1202-D1210. doi 10.1093/nar/gkr1090

Lavrekha V.V., Levitsky V.G., Tsukanov A.V., Bogomolov A.G., Grigorovich D.A., Omelyanchuk N., Ubogoeva E.V., Zemlyanskaya E.V., Mironova V. CisCross: a gene list enrichment analysis to predict upstream regulators in *Arabidopsis thaliana*. *Front Plant Sci*. 2022;13: 942710. doi 10.3389/fpls.2022.942710

Lavrekha V.V., Omelyanchuk N.A., Bogomolov A.G., Zemlyanskaya E.V. PlantReg: the reconstruction of links between transcription factor regulatory networks and biological processes under their control. *Vavilov J Genet Breed*. 2024;28(8):950-959. doi 10.18699/vjgb-24-102

Leong R., He X., Beijen B.S., Sakai T., Goncalves J., Ding P. Unlocking gene regulatory networks for crop resilience and sustainable agriculture. *Nat Biotechnol.* 2025;43(8):1254-1265. doi 10.1038/s41587-025-02727-4

Marshall-Colón A., Kliebenstein D.J. Plant networks as traits and hypotheses: moving beyond description. *Trends Plant Sci.* 2019;24(9): 840-852. doi 10.1016/j.tplants.2019.06.003

Martínez C., Espinosa-Ruíz A., de Lucas M., Bernardo-García S., Franco-Zorrilla J.M., Prat S. PIF 4-induced BR synthesis is critical to diurnal and thermomorphogenic growth. *EMBO J.* 2018;37(23): e99552. doi 10.15252/embj.201899552

Nagahage I.S.P., Sakamoto S., Nagano M., Ishikawa T., Kawai-Yamada M., Mitsuda N., Yamaguchi M. An NAC domain transcription factor ATAF2 acts as transcriptional activator or repressor dependent on promoter context. *Plant Biotechnol.* 2018;35(3):285-289. doi 10.5511/plantbiotechnology.18.0507a

O'Malley R.C., Huang S.C., Song L., Lewsey M.G., Bartlett A., Nery J.R., Galli M., Gallavotti A., Ecker J.R. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*. 2016; 165(5):1280-1292. doi 10.1016/j.cell.2016.04.038

Omelyanchuk N.A., Lavrekha V.V., Bogomolov A.G., Dolgikh V.A., Sidorenko A.D., Zemlyanskaya E.V. Computational reconstruction of the transcription factor regulatory network induced by auxin in *Arabidopsis thaliana* L. *Plants*. 2024;13(14):1905. doi 10.3390/plants13141905

Otani Y., Kawanishi M., Kamimura M., Sasaki A., Nakamura Y., Nakamura T., Okamoto S. Behavior and possible function of *Arabidopsis* BES1/BZR1 homolog 2 in brassinosteroid signaling. *Plant Signal Behav.* 2022;17(1):2084277. doi 10.1080/15592324.2022.2084277

Park P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet.* 2009;10(10):669-680. doi 10.1038/nrg2641

Percio F., Rubio L., Amorim-Silva V., Botella M.A. Crucial roles of brassinosteroids in cell wall composition and structure across species: new insights and biotechnological applications. *Plant Cell Environ.* 2025;48(3):1751-1767. doi 10.1111/pce.15258

Raditsa V.V., Tsukanov A.V., Bogomolov A.G., Levitsky V.G. Genomic background sequences systematically outperform synthetic ones in de novo motif discovery for ChIP-seq data. *NAR Genomics Bioinform*. 2024;6(3):lqae090. doi 10.1093/nargab/lqae090

Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., … Lenhard B., Sandelin A., Wasserman W.W., Parcy F., Mathelier A. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2024;52(D1):D174-D182. doi 10.1093/nar/gkad1059

Reinhold H., Soyk S., Simková K., Hostettler C., Marafino J., Mainiero S., Vaughan C.K., Monroe J.D., Zeeman S.C. β-amylase-like proteins function as transcription factors in *Arabidopsis*, controlling shoot growth and development. *Plant Cell.* 2011;23(4):1391-1403. doi 10.1105/tpc.110.081950

Ren G., Cui K., Zhang Z., Zhao K. Division of labor between IRF1 and IRF2 in regulating different stages of transcriptional activation in cellular antiviral activities. *Cell Biosci.* 2015;5(1):17. doi 10.1186/s13578-015-0007-0

Rybakov M.A., Omelyanchuk N.A., Zemlyanskaya E.V. Reconstruction of gene regulatory networks from single cell transcriptomic data. *Vavilov J Genet Breed*. 2024;28(8):974-981. doi 10.18699/vjgb-24-104

Sánchez-Camargo V.A., Romero-Rodríguez S., Vázquez-Ramos J.M. Non-canonical functions of the E2F/DP pathway with emphasis in plants. *Phyton.* 2021;90(2):307-330. doi 10.32604/phyton.2021.014967

Schindler U., Beckmann H., Cashmore A.R. TGA1 and G-box binding factors: two distinct classes of Arabidopsis leucine zipper proteins compete for the G-box-like element TGACGTGG. *Plant Cell.* 1992;4(10):1309-1319. doi 10.1105/tpc.4.10.1309

Selvaraj N., Budka J.A., Ferris M.W., Plotnik J.P., Hollenhorst P.C. Extracellular signal-regulated kinase signaling regulates the opposing roles of JUN family transcription factors at ETS/AP-1 sites and in cell migration. *Mol Cell Biol.* 2015;35(1):88-100. doi 10.1128/mcb.00982-14

Shi H., Li X., Lv M., Li J. BES1/BZR1 family transcription factors regulate plant development via brassinosteroid-dependent and independent pathways. *Int J Mol Sci.* 2022;23(17):10149. doi 10.3390/ijms231710149

Su K., Katebi A., Kohar V., Clauss B., Gordin D., Qin Z.S., Karuturi R.K.M., Li S., Lu M. NetAct: a computational platform to construct core transcription factor regulatory networks using gene activity. *Genome Biol.* 2022;23(1):270. doi 10.1186/s13059-022-02835-3

Sun T., Hazra A., Lui A., Zeng S., Wang X., Rao S., Owens L.A., Fei Z., Zhao Y., Mazourek M., Giovannoni J.G., Li L. GLKs directly regulate carotenoid biosynthesis via interacting with GBFs in plants. *New Phytol.* 2025;246(2):645-665. doi 10.1111/nph.20457

Sun Y., Li J., Huang J., Li S., Li Y., Lu B., Deng X. Architecture of genome-wide transcriptional regulatory network reveals dynamic functions and evolutionary trajectories in *Pseudomonas syringae*. *eLife.* 2025;13:RP96172. doi 10.7554/eLife.96172.3

Tamura T., Sakata T., Igarashi H., Okumura K. Transcription factor HUB1 represses SP1-mediated gene expression through the CACCC box of HTLV-I U5RE but not the GC box. *J Health Sci*. 2004;50(4):417-422. doi 10.1248/jhs.50.417

Tremblay B.J.M. universalmotif: an R package for biological motif analysis. *J Open Source Software*. 2024;9(100):7012. doi 10.21105/joss.07012

van den Heuvel S., Dyson N.J. Conserved functions of the pRB and E2F families. *Nat Rev Mol Cell Biol.* 2008;9(9):713-724. doi 10.1038/nrm2469

Vanneste S., Pei Y., Friml J. Mechanisms of auxin action in plant growth and development. *Nat Rev Mol Cell Biol.* 2025;26(9):648-666. doi 10.1038/s41580-025-00851-2

Veerabagu M., Kirchler T., Elgass K., Stadelhofer B., Stahl M., Harter K., Mira-Rodado V., Chaban C. The interaction of the *Arabidopsis* response regulator ARR18 with bZIP63 mediates the regulation of *PROLINE DEHYDROGENASE* expression. *Mol Plant.* 2014;7(10):1560-1577. doi 10.1093/mp/ssu074

Wang L., Ko E.E., Tran J., Qiao H. TREE1-EIN3-mediated transcriptional repression inhibits shoot growth in response to ethylene. *Proc Natl Acad Sci USA*. 2020;117(46):29178-29189. doi 10.1073/pnas.2018735117

Weirauch M.T., Yang A., Albu M., Cote A.G., Montenegro-Montero A., Drewe P., Najafabadi H.S., … Bouget F.Y., Ratsch G., Larrondo L.F., Ecker J.R., Hughes T.R. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014;158(6):1431-1443. doi 10.1016/j.cell.2014.08.009

Weltmeier F., Rahmani F., Ehlert A., Dietrich K., Schütze K., Wang X., Chaban C., Hanson J., Teige M., Harter K., Vicente-Carbajosa J., Smeekens S., Dröge-Laser W. Expression patterns within the *Arabidopsis* C/S1 bZIP transcription factor network: availability of heterodimerization partners controls gene expression during stress response and development. *Plant Mol Biol.* 2009;69(1):107-119. doi 10.1007/s11103-008-9410-9

Wildenhain T., Smaczniak C., Marsell A., Draken J., Maag D., Kreisz P., Krischke M., Müller M.J., Kaufmann K., Weiste C., Dröge-Laser W. A subset of group S_1 bZIP transcription factors controls resource management during starvation and recovery in Arabidopsis. *Plant Cell.* 2025;37(7):koaf149. doi 10.1093/plcell/koaf149

Wu T., Goh H., Azodi C.B., Krishnamoorthi S., Liu M.J., Urano D. Evolutionarily conserved hierarchical gene regulatory networks for plant salt stress response. *Nat Plants*. 2021;7(6):787-799. doi 10.1038/s41477-021-00929-7

Zhang X., Li L., Fourie J., Davie J.R., Guarcello V., Diasio R.B. The role of Sp1 and Sp3 in the constitutive *DPYD* gene expression. *Biochim Biophys Acta.* 2006;1759(5):247-256. doi 10.1016/j.bbaexp.2006.05.001

Zemlyanskaya E.V., Dolgikh V.A., Levitsky V.G., Mironova V. Transcriptional regulation in plants: using omics data to crack the cis-regulatory code. *Curr Opin Plant Biol.* 2021;63:102058. doi 10.1016/j.pbi.2021.102058

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 29.08.2025. После доработки 23.09.2025. Принята к публикации 23.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):952-962

doi 10.18699/vjgb-25-101

База знаний FlyDEGdb по дифференциально экспрессирующимся генам *Drosophila melanogaster* – модельного объекта биомедицины

О.А. Подколодная [1$], М.А. Дерюженко [1$], Н.Н. Твердохлеб [1], К.А. Золотарева [1], Ю.В. Маковка [1], Н.Л. Подколодный [1, 2], В.В. Суслов [1], И.В. Чадаева [1], Л.А. Федосеева [1], А.А. Серяпина [1], Д.Ю. Ощепков [1], А.Г. Богомолов [1], Е.Ю. Кондратюк [1, 3], О.Е. Редина [1], А.Л. Маркель [1, 4], Н.Е. Грунтенко [1], М.П. Пономаренко [1] ✉

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Сибирский федеральный научный центр агробиотехнологий Российской академии наук, р. п. Краснообск, Новосибирская область, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

✉ pon@bionet.nsc.ru

Аннотация. С 1909 г. благодаря исследованиям нобелевского лауреата Моргана дрозофила *Drosophila melanogaster* стала одним из самых популярных модельных животных в генетике. Фундаментальные исследования с дрозофилой в качестве модельного объекта неоднократно были отмечены Нобелевской премией: в 1946 г. (Мёллер, мутагенез при рентгеновском излучении), в 1995 (Льюис, Нюсслайн-Фольхард, Вишаус, генетический контроль эмбриогенеза), в 2004 (Эксел и Бак, обонятельная система), в 2011 (Стайнман, дендритные клетки в адаптивном иммунитете; Бётлер и Офман, активация врожденного иммунитета) и в 2017 г. (Холл, Росбаш и Янг, молекулярный механизм циркадного ритма). Столь яркая роль дрозофилы в генетике обусловлена рядом ее ключевых признаков: кратким жизненным циклом, частой сменой поколений, легкостью в содержании, высокой плодовитостью, малым размером, прозрачностью эмбриона, простым строением личинки, возможностью визуальных наблюдений хромосомных перестроек за счет наличия политенных хромосом, доступностью для молекулярно-генетических манипуляций. Кроме того, благодаря высокой консервативности ряда сигнальных путей и генных сетей дрозофилы и их сходству с таковыми у млекопитающих и человека в совокупности с техническим развитием геномного секвенирования стало возможно использование *D. melanogaster* как модельного объекта в биомедицинских исследованиях в области фармакологии, токсикологии, кардиологии, онкологии, иммунологии, геронтологии и радиобиологии для поиска генетической и эпигенетической основ патогенеза болезней человека. В настоящей статье описана созданная нами курируемая база знаний FlyDEGdb (https://www.sysbio.ru/FlyDEGdb), в которой представлена информация о дифференциально экспрессирующихся генах (ДЭГ) дрозофилы, экстрагированная из 50 научных статей с экспериментальными данными об изменении экспрессии 20 058 генов (80 %) из числа всех 25 079 генов дрозофилы согласно базе данных NCBI Gene под действием 52 стрессовых факторов, включая высокую и низкую температуры, обезвоживание, тяжелые металлы, радиацию, голод, яды, бытовую химию, лекарства, удобрения, инсектициды, пестициды и гербициды. Содержание базы знаний FlyDEGdb проиллюстрировано на примере гена *dysf* (*dysfusion*) дрозофилы, который был идентифицирован в качестве ДЭГ при множестве стрессовых воздействий: холодовом шоке и в испытаниях на токсичность гербицида параквата, растворителя толуола, лекарственного препарата менадиона, пищевой добавки E923. В FlyDEGdb представлена информация об изменениях экспрессии гена *dysf* и его гомологов *Clk*, *cyc*, *per* у дрозофилы и генов *NPAS4*, *CLOCK*, *BMAL1*, *PER1* и *PER2* человека, а также информация о биологических процессах, в которые вовлечены эти гены: созревание ооцитов (оогенез), регуляция стресс-ответа и циркадного ритма, канцерогенез, старение и др. Поэтому FlyDEGdb, содержащая информацию о таком модельном организме, как дрозофила, может быть полезна для исследователей, работающих в области молекулярной биологии и генетики человека и животных, физиологии, трансляционной медицины, фармакологии, диетологии, агрохимии, радиобиологии, токсикологии и биоинформатики.

Ключевые слова: человек; заболевание; биомедицина; модельное животное; дрозофила; *Drosophila melanogaster*; дифференциально экспрессирующиеся гены (ДЭГ); RNA-seq; qPCR; микрочип; база знаний

Для цитирования: Подколодная О.А., Дерюженко М.А., Твердохлеб Н.Н., Золотарева К.А., Маковка Ю.В., Подколодный Н.Л., Суслов В.В., Чадаева И.В., Федосеева Л.А., Серяпина А.А., Ощепков Д.Ю., Богомолов А.Г., Кондратюк Е.Ю., Редина О.Е., Маркель А.Л., Грунтенко Н.Е., Пономаренко М.П. База знаний FlyDEGdb по дифференциально экспрессирующимся генам *Drosophila melanogaster* – модельного объекта биомедицины. *Вавиловский журнал генетики и селекции*. 2025;29(7):952-962. doi 10.18699/vjgb-25-101

Финансирование. Работа поддержана бюджетным проектом FWNR-2022-0019.

Благодарности. Авторы выражают благодарность Центру коллективного пользования (ЦКП) «Биоинформатика» за предоставленные вычислительные ресурсы.

© Подколодная О.А., Дерюженко М.А., Твердохлеб Н.Н., Золотарева К.А., Маковка Ю.В., Подколодный Н.Л., Суслов В.В., Чадаева И.В., Федосеева Л.А., Серяпина А.А., Ощепков Д.Ю., Богомолов А.Г., Кондратюк Е.Ю., Редина О.Е., Маркель А.Л., Грунтенко Н.Е., Пономаренко М.П., 2025

$ Равный вклад.

Контент доступен под лицензией Creative Commons Attribution 4.0

FlyDEGdb knowledge base on differentially expressed genes of *Drosophila melanogaster*, a model object in biomedicine

O.A. Podkolodnaya [1$], M.A. Deryuzhenko [1$], N.N. Tverdokhleb[1], K.A. Zolotareva [1], Yu.V. Makovka [1], N.L. Podkolodny [1, 2], V.V. Suslov [1], I.V. Chadaeva [1], L.A. Fedoseeva [1], A.A. Seryapina [1], D.Yu. Oshchepkov [1], A.G. Bogomolov [1], E.Yu. Kondratyuk [1, 3], O.E. Redina [1], A.L. Markel [1, 4], N.E. Gruntenko [1], M.P. Ponomarenko [1] ✉

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, Russia
[4] Novosibirsk State University, Novosibirsk, Russia

✉ pon@bionet.nsc.ru

Abstract. Since the work of Nobel Prize winner Thomas Morgan in 1909, the fruit fly *Drosophila melanogaster* has been one of the most popular model animals in genetics. Research using this fly was honored with the Nobel Prize many times: in 1946 (Muller, X-ray mutagenesis), in 1995 (Lewis, Nüsslein-Volhard, Wieschaus, genetic control of embryogenesis), in 2004 (Axel and Buck, the olfactory system), in 2011 (Steinman, dendritic cells in adaptive immunity; Beutler and Hoffman, activation of innate immunity), and in 2017 (Hall, Rosbash and Young, the molecular mechanism of the circadian rhythm). The prominent role of Drosophila in genetics is due to its key features: short life cycle, frequent generational turnover, ease of maintenance, high fertility, small size, transparent embryos, simple larval structure, the possibility to observe visually chromosomal rearrangements due to the presence of polytene chromosomes, and accessibility to molecular genetic manipulation. Furthermore, the highly conserved nature of several signaling pathways and gene networks in Drosophila and their similarity to those of mammals and humans, taken together with the development of high-throughput genomic sequencing, motivated the use of *D. melanogaster* as a model organism in biomedical fields of inquiry: pharmacology, toxicology, cardiology, oncology, immunology, gerontology, and radiobiology. These studies add to the understanding of the genetic and epigenetic basis of the pathogenesis of human diseases. This paper describes our curated knowledge base, FlyDEGdb (https://www.sysbio.ru/FlyDEGdb), which stores information on differentially expressed genes (DEGs) in Drosophila. This information was extracted from 50 scientific articles containing experimental data on changes in the expression of 20,058 genes (80 %) out of the 25,079 Drosophila genes stored in the NCBI Gene database. The changes were induced by 52 stress factors, including heat and cold exposure, dehydration, heavy metals, radiation, starvation, household chemicals, drugs, fertilizers, insecticides, pesticides, herbicides, and other toxicants. The FlyDEGdb knowledge base is illustrated using the example of the *dysf* (*dysfusion*) Drosophila gene, which had been identified as a DEG under cold shock and in toxicity tests of the herbicide paraquat, the solvent toluene, the drug menadione, and the food additive E923. FlyDEGdb stores information on changes in the expression of the *dysf* gene and its homologues: (a) the *Clk*, *cyc*, and *per* genes in Drosophila, and (b) the *NPAS4*, *CLOCK*, *BMAL1*, *PER1*, and *PER2* genes in humans. These data are supplemented with information on the biological processes in which these genes are involved: oocyte maturation (oogenesis), regulation of stress response and circadian rhythm, carcinogenesis, aging, etc. Therefore, FlyDEGdb, containing information on the widely used model organism, Drosophila, can be helpful for researchers working in the molecular biology and genetics of humans and animals, physiology, translational medicine, pharmacology, dietetics, agricultural chemistry, radiobiology, toxicology, and bioinformatics.

Key words: human; disease; biomedicine; model animal; fruit fly *Drosophila melanogaster*; differentially expressed genes (DEGs); RNA-Seq; qPCR; microarray; knowledge base

For citation: Podkolodnaya O.A., Deryuzhenko M.A., Tverdokhleb N.N., Zolotareva K.A., Makovka Yu.V., Podkolodny N.L., Suslov V.V., Chadaeva I.V., Fedoseeva L.A., Seryapina A.A., Oshchepkov D.Yu., Bogomolov A.G., Kondratyuk E.Yu., Redina O.E., Markel A.L., Gruntenko N.E., Ponomarenko M.P. FlyDEGdb knowledge base on differentially expressed genes of *Drosophila melanogaster*, a model object in biomedicine. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):952-962. doi 10.18699/vjgb-25-101

Введение

В биомедицинских исследованиях при изучении физиологических, генетических и эпигенетических механизмов регуляции эволюционно закрепленных фенотипических признаков человека в норме и при патологиях, а также воздействии внешних и внутренних стрессоров широко используются модельные лабораторные животные (Mukherjee et al., 2022). Их применение основано на строгих критериях соответствия между изучаемыми особенностями фенотипа человека и их эквивалентами у модельных животных (Gryksa et al., 2023). Более века назад Томас Хант Морган (Morgan, 1910), будучи профессором экспериментальной зоологии Колумбийского университета, положил начало многочисленным открытиям в области наследования признаков на новом для того времени биологическом объекте – дрозофиле, *Drosophila melanogaster*. Результаты его исследований были удостоены в 1933 г. Нобелевской премии «За открытия роли хромосом в наследственности». Позднее генетические исследования с использованием дрозофилы еще пять раз отмечались Нобелевской премией: в 1946 г. ее присудили Герману Джозефу Мёллеру «За открытие появления мутаций под влиянием рентгеновского излучения», в 1995 – Эдварду Льюису, Кристиане Нюсляйн-Фольхард и Эрику Вишаусу «За открытия, касающиеся генетического контроля на ранних стадиях эмбрионального развития», в 2004 – Ричар-

ду Экселу и Линде Бак «За исследования обонятельных рецепторов и организации системы органов обоняния», в 2011 – Ральфу Стайнману «За открытие дендритных клеток и изучение их значения для приобретенного иммунитета» вместе с Жюлем Офманом и Брюсом Бётлером «За работы по изучению активации врожденного иммунитета», в 2017 г. – Джеффри Холлу, Майклу Росбашу и Майклу Янгу «За открытия молекулярных механизмов, управляющих циркадным ритмом» (Lakhotia, 2025).

Столь весомый вклад экспериментов с дрозофилой в науку обусловлен низкой стоимостью ее содержания, большим количеством потомства, коротким периодом индивидуального развития, малыми размерами, оптической прозрачностью эмбрионов, простотой строения личинки, коротким жизненным циклом, большим разнообразием природных линий, адаптированных к различным эколого-климатическим условиям (Telonis-Scott et al., 2013; Chen et al., 2015; von Heckel et al., 2016; Mikucki et al., 2024), относительно малым размером генома и простотой молекулярно-генетических манипуляций. При этом особенно важно сходство многих сигнальных путей и генных сетей человека и дрозофилы (Yu et al., 2022). Это обеспечивает широкие возможности проекции на человека результатов, получаемых на дрозофиле в таких областях, как трансляционная медицина, фармакология, токсикология, иммунология, геронтология и др. (De Gregorio et al., 2001; Chatterjee, Perrimon, 2021; Wu K. et al., 2021; Ali et al., 2022; Rand et al., 2023).

В Интитуте цитологии и генетики (ИЦиГ) СО РАН в рамках указанного выше магистрального направления более 40 лет изучаются различные аспекты реакции на стресс у крысы (Маркель, 1985; Oshchepkov et al., 2024) и мыши (Chadaeva et al., 2019; Avgustinovich et al., 2025). Итоги работ представлены во множестве публикаций и содержат ценную информацию о дифференциальной экспрессии генов (ДЭГ), индуцированной разными экспериментальными воздействиями. При этом были получены большие объемы полногеномных экспериментальных данных (Big Data) о ДЭГ крысы и мыши, которые документированы в созданных нами базах знаний RatDEGdb (Чадаева и др., 2023) и MiceDEGdb (Подколодная и др., 2025) соответственно.

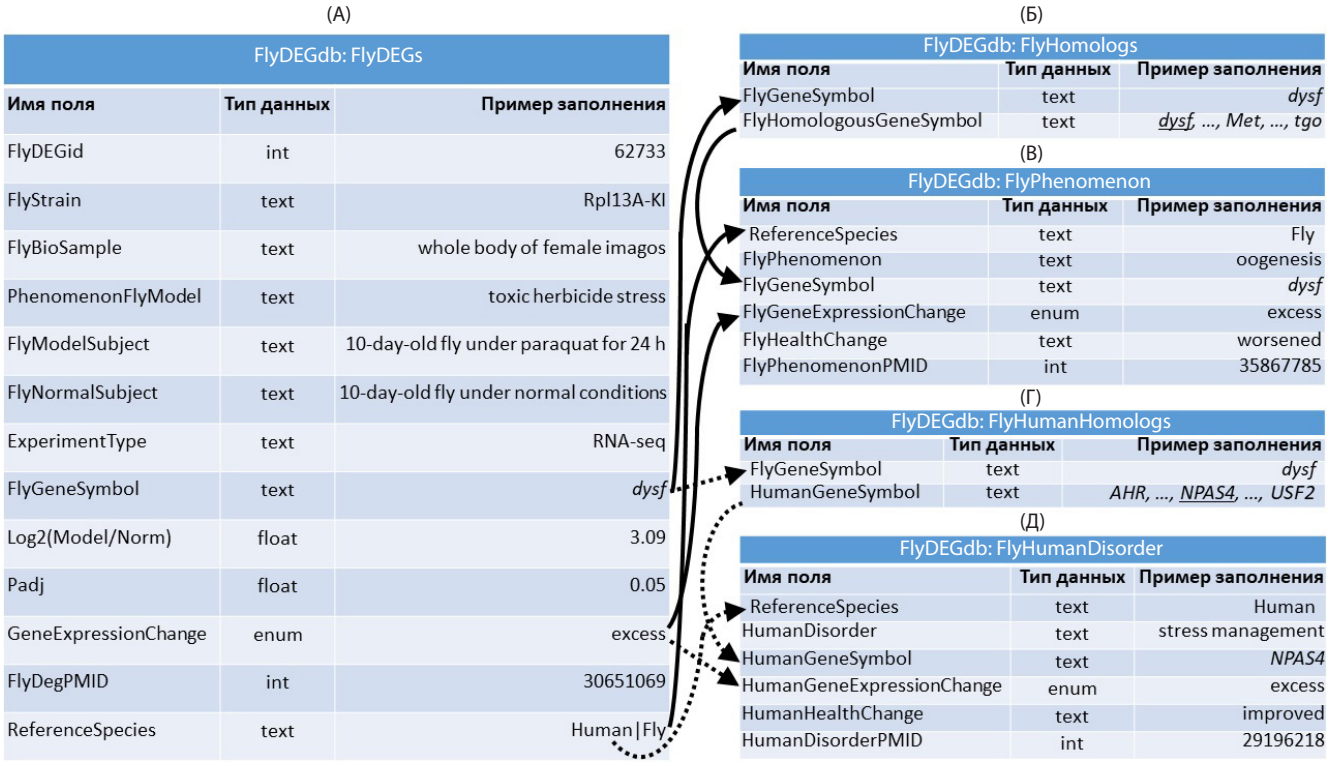
Еще один модельный организм, на котором в ИЦиГ СО РАН более 25 лет проводятся экспериментальные исследования вопросов стресса у животных, это *D. melanogaster* (Gruntenko et al., 1999, 2023). Работа по созданию базы знаний FlyDEGdb, содержащей информацию о ДЭГ дрозофилы, продолжает серию наших исследований в области биомедицинских баз знаний RatDEGdb и MiceDEGdb. FlyDEGdb доступна по адресу https://www.sysbio.ru/FlyDEGdb. Она содержит экспериментальные данные об изменении экспрессии 80 % генов дрозофилы – 20058 генов из их общего числа 25079 в геноме этого организма, согласно базе данных NCBI Gene (Brown et al., 2015). Информация, представленная в базе FlyDEGdb, была экстрагирована из 50 научных статей, содержащих экспериментальные данные о воздействии на 31 линию *D. melanogaster* 52 стрессовых факторов, включая высокую и низкую температуры, обезвоживание, тяжелые металлы, радиацию, голод, яды, бытовую химию, лекарства, удобрения, инсектициды, пестициды и гербициды. Информационное наполнение FlyDEGdb v0.01 проиллюстрировано на примере гена *dysf* (*dysfusion*) дрозофилы, который был идентифицирован в качестве ДЭГ при множестве стрессовых воздействий: холодовом шоке и в испытаниях на токсичность гербицида параквата, растворителя толуола, лекарственного препарата менадиона, пищевой добавки Е923. В FlyDEGdb собрана информация об изменениях экспрессии как гена *dysf* и его гомологов – генов *Clk*, *cyc*, *per* у дрозофилы, так и генов *NPAS4*, *CLOCK*, *BMAL1*, *PER1* и *PER2* человека, а также информация о биологических процессах, в которые вовлечены эти гены: созревание ооцитов (оогенез), регуляция стресс-ответа и циркадного ритма, канцерогенез, старение и др.

Приведены также результаты сравнительного анализа данных по стресс-индуцированным ДЭГ дрозофилы из FlyDEGdb с данными по ДЭГ гипоталамуса крысы линий WAG и НИСАГ в ответ на стрессовое ограничение подвижности животных (Oshchepkov et al., 2024), документированными в RatDEGdb. В качестве общего молекулярного события в ответ на стрессовые воздействия, которым подверглись крысы и дрозофилы, у этих модельных животных обнаружено снижение экспрессии больших групп генов, вовлеченных в формирование плазматической мембраны. База знаний FlyDEGdb, содержащая информацию о таком модельном организме, как дрозофила, может быть полезна для исследователей, работающих в области молекулярной биологии и генетики человека и животных, физиологии, трансляционной медицины, фармакологии, диетологии, агрохимии, радиобиологии, токсикологии и биоинформатики.

Материалы и методы

Стресс-индуцированные ДЭГ дрозофилы. Экспериментально выявленные ДЭГ дрозофилы находили с использованием текстовых запросов из различных комбинаций ключевых слов – “Drosophila melanogaster”, “differentially expressed gene”, “stress response”, “drying”, “heat shock”, “radiation”, “cold shock”, “oxidative stress”, “continuous lighting”, “toxin”, “diet”, “heavy metal”, “drug”, “herbicide”, “pesticide”, “insecticide”, “RNA-seq”, “microarray” и “qPCR” – при обращении к базе данных PubMed (Lu, 2011).

В базу данных FlyDEGdb включали только опубликованные ДЭГ, каждый из которых был охарактеризован авторами соответствующих научных статей величиной $\log_2$(ДЭГ) = $\log_2$([уровень экспрессии ДЭГ при изучаемом стрессовом воздействии на дрозофилу] / [уровень экспрессии ДЭГ в норме]) и опубликованной величиной оценки P_{ADJ} уровня статистической значимости с поправкой на множественные сравнения для стресс-индуцированного изменения экспрессии рассматриваемого ДЭГ. Из числа отобранных по этому критерию ДЭГ дрозофилы дополнительно исключили те, для которых оценки $\log_2$(ДЭГ) были в диапазоне от –0.46 до 0.46, соответствующем недостоверным ($p \geq 0.05$, Z-оценка Фишера) различиям между уровнями экспрессии ДЭГ до и после исследуемого стрессового воздействия при точности ±5 % для экспериментальных измерений величин экспрессии генов.



(А)

FlyDEGdb: FlyDEGs

Имя поля	Тип данных	Пример заполнения
FlyDEGid	int	62733
FlyStrain	text	Rpl13A-KI
FlyBioSample	text	whole body of female imagos
PhenomenonFlyModel	text	toxic herbicide stress
FlyModelSubject	text	10-day-old fly under paraquat for 24 h
FlyNormalSubject	text	10-day-old fly under normal conditions
ExperimentType	text	RNA-seq
FlyGeneSymbol	text	*dysf*
Log2(Model/Norm)	float	3.09
Padj	float	0.05
GeneExpressionChange	enum	excess
FlyDegPMID	int	30651069
ReferenceSpecies	text	Human | Fly

(Б)

FlyDEGdb: FlyHomologs

Имя поля	Тип данных	Пример заполнения
FlyGeneSymbol	text	*dysf*
FlyHomologousGeneSymbol	text	*dysf*, …, *Met*, …, *tgo*

(В)

FlyDEGdb: FlyPhenomenon

Имя поля	Тип данных	Пример заполнения
ReferenceSpecies	text	Fly
FlyPhenomenon	text	oogenesis
FlyGeneSymbol	text	*dysf*
FlyGeneExpressionChange	enum	excess
FlyHealthChange	text	worsened
FlyPhenomenonPMID	int	35867785

(Г)

FlyDEGdb: FlyHumanHomologs

Имя поля	Тип данных	Пример заполнения
FlyGeneSymbol	text	*dysf*
HumanGeneSymbol	text	*AHR*, …, *NPAS4*, …, *USF2*

(Д)

FlyDEGdb: FlyHumanDisorder

Имя поля	Тип данных	Пример заполнения
ReferenceSpecies	text	Human
HumanDisorder	text	stress management
HumanGeneSymbol	text	*NPAS4*
HumanGeneExpressionChange	enum	excess
HumanHealthChange	text	improved
HumanDisorderPMID	int	29196218

Рис. 1. Информационная структура базы знаний FlyDEGdb по дифференциально экспрессирующимся генам (ДЭГ) *Drosophila melanogaster*. Реляционные таблицы: (А) FlyDEGs – экспериментальные данные о ДЭГ в образцах тканей дрозофилы определенных линий в ответ на стрессовое воздействие (в сравнении с нормой согласно цитируемой статье); (Б) FlyHomologs – списки генов дрозофилы, гомологичных конкретным генам дрозофилы согласно базе данных FlyBase (Ozturk-Colak et al., 2024); (В) FlyPhenomenon – фенотипические особенности, ассоциированные с отклонениями экспрессии генов дрозофилы (в сравнении с нормой согласно цитируемой статье); (Г) FlyHumanHomologs – гены человека, гомологичные конкретному гену дрозофилы согласно базе данных FlyBase (Ozturk-Colak et al., 2024); (Д) HumanDisorder – заболевания человека, ассоциированные с отклонениями экспрессии определенных генов человека (в сравнении с нормой согласно цитируемой статье).

Имена реляционных таблиц и их полей выбраны согласно руководству по практическому созданию дружественных интерфейсов (Wade, 1984). Типы данных: int – целое число, float – вещественное число, enum – бинарный индикатор, text – строка символов; PMID – идентификатор цитируемой статьи в базе данных PubMed (Lu, 2011). Стрелки (→) – реляционные ссылки, показывающие, с одной стороны, аннотацию экспериментальных данных о направлении изменения экспрессии ДЭГ дрозофилы (реляционная таблица FlyDEGs), а с другой – сведения об однонаправленном изменении экспрессии гомологичных ДЭГ дрозофилы (непрерывные линии) или человека (пунктирные линии), перечисленных в таблицах FlyHomologs и FlyHumanHomologs, в независимых экспериментах согласно научным статьям, которые цитируются в реляционных таблицах FlyPhenomenon и HumanDisorder.

База знаний FlyDEGdb. Информационная структура базы знаний FlyDEGdb представлена на рис. 1. Она включает пять реляционных таблиц, первая из которых, “FlyDEGs” (см. рис. 1, А), содержит экспериментальные данные об определенном ДЭГ дрозофилы, которому в рамках FlyDEGdb присвоен уникальный номер (поле “FlyDEGid”). В поле “FlyStrain” этой таблицы указана линия дрозофил, в экспериментах с которой был выявлен рассматриваемый ДЭГ, в поле “FlyBioSample” – образец тканей дрозофилы, исследованной в этом эксперименте, в поле “PhenomenonFlyModel” – изучаемое стрессовое воздействие на дрозофилу, в полях “FlyModelSubject” и “FlyNormalSubject” – модельные и контрольные особи соответственно в рамках указанного эксперимента, тип которого – “RNA-seq”, “Microarray” или “RT-qPCR” – приведен в поле “ExperimentType”. Далее, поле “FlyGeneSymbol” содержит символьное обозначение рассматриваемого ДЭГ дрозофилы согласно базе данных NCBI Gene (Brown et al., 2015), поля “Log2(Model/Norm)” и “Padj” – величины стресс-индуцированного изменения экспрессии ДЭГ в сравнении с нормой и уровня его статистической значимости с поправкой на множественные сравнения соответственно, как они были опубликованы в научной статье, которая цитируется в поле “FlyDegPMID” с использованием ее идентификатора в базе данных PubMed (Lu, 2011).

Наконец, поле “ReferenceSpecies” указывает на референсный биологический вид (в пилотной версии FlyDEGdb v0.1: дрозофила как “Fly” или человек как “Human”), независимые экспериментальные данные о котором использованы для аннотации рассматриваемого ДЭГ, или на отсутствие такой аннотации в случае “ND”.

В рамках настоящей работы термином «аннотация» обозначено дополнение экспериментальных данных о стресс-индуцированном изменении экспрессии определенного ДЭГ дрозофилы, описанных в конкретной научной статье, экспериментальными данными из независимых источников по фенотипическим проявлениям однонаправленных изменений экспрессии гомологичных генов у человека и дрозофилы. Детали процедуры аннотации показаны в табл. S1 Приложения[1].

[1] Табл. S1–S3 и рис. S1 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx32.pdf

Завершим описание информационной структуры FlyDEGdb (см. рис. 1) информацией об используемых типах данных, а именно: целое число (int), вещественное число (float), бинарный индикатор (enum) и строка символов (text).

С помощью MySQL-совместимой среды управления данными MariaDB 10.2.12 (MariaDB Corp AB, Финляндия) реляционные таблицы FlyDEGs, FlyHomologs, FlyPhenomenon, FlyHumanHomologs и HumanDisorder были интегрированы в базу знаний FlyDEGdb (https://www.sysbio.ru/FlyDEGdb).

Статистический анализ ДЭГ дрозофилы провели с использованием приложения Past v.4.04 (Hammer et al., 2001) и пакета STATISTICA (Statsoft™, США).

Результаты и обсуждение

База знаний FlyDEGdb

Для формирования FlyDEGdb был проведен поиск статей о ДЭГ дрозофилы в базе данных PubMed (Lu, 2011) по ключевым словам, перечисленным в разделе «Материалы и методы». В PubMed было найдено 50 статей с описанием 287 экспериментов, проведенных на 31 линии дрозофилы различного географического происхождения, а также на их трансгенных модификациях. Всего в этих статьях описано более 190000 стресс-индуцированных ДЭГ дрозофилы. Результаты поиска представлены в табл. S1–S3. Найденные ДЭГ представляют широкий спектр исследований, выполненных на дрозофиле для изучения возрастных заболеваний человека, испытания лекарственных препаратов, оценки токсичности средств бытовой химии, сельскохозяйственных удобрений, инсектицидов, пестицидов и гербицидов и др.

Иллюстративный пример доступа пользователя к информации, собранной в пилотной версии FlyDEGdb, показан на рис. 2.

Три кнопки в верхней части интерфейса FlyDEGdb обеспечивают доступ пользователя к информации в этой базе знаний:

- “DOWNLOAD DB” позволяет скачать всю информацию текущей версии FlyDEGdb v0.1 в виде текстового файла в Excel-совместимом формате;
- “START (USUAL)” обеспечивает доступ к экспериментальным данным о стресс-индуцированных ДЭГ дрозофилы, описанным в основной реляционной таблице “FlyDEGs” (см. рис. 1, A);
- “START (HOMOLOG)” предоставляет доступ к аннотации ДЭГ дрозофилы, как это описано в разделе «Материалы и методы».

Ниже расположены поля интерфейса для выбора типа информации, интересующей пользователя: экспериментальные данные о ДЭГ дрозофилы и аннотация ДЭГ дрозофилы. Поле “Page Number” позволяет просматривать в алфавитном порядке все ДЭГ, документированные в базе знаний.

В нижней части интерфейса в табличной форме показаны данные о ДЭГ, полученные пользователем согласно вышеописанному запросу, хранение которых в информа-

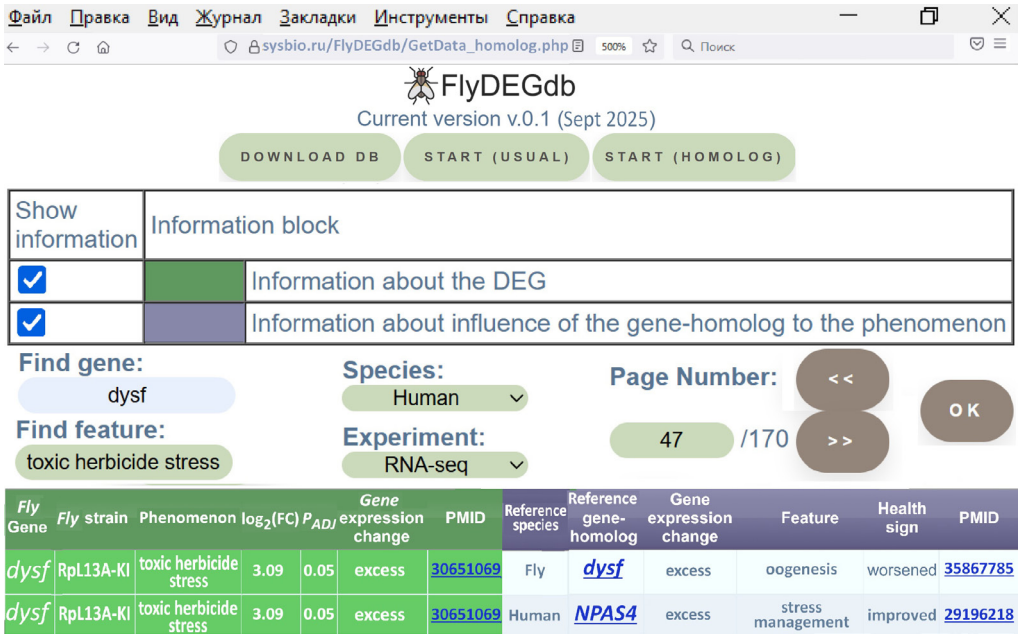


Рис. 2. Интерфейс базы знаний FlyDEGdb по ДЭГ *D. melanogaster*, который обеспечивает режим диалога в реальном времени для доступа пользователя к ее информационному содержанию.

Команды интерфейса: DOWNLOAD DB – скачать всю информацию текущей версии FlyDEGdb v0.1 в виде текстового файла в Excel-совместимом формате; START (HOMOLOG) – доступ к ДЭГ дрозофилы, аннотированным с использованием независимых экспериментальных данных о фенотипических проявлениях однонаправленных изменений экспрессии в сравнении с нормой для гомологичных генов референсных биологических видов дрозофилы и человека; START (USUAL) – доступ к ДЭГ дрозофилы без их аннотации. Левая половина таблицы с информацией о ДЭГ дрозофилы (зеленый фон) – экспериментальные данные о рассматриваемом ДЭГ дрозофилы; правая половина (сиреневый фон) – аннотация рассматриваемого ДЭГ дрозофилы на основе независимых данных о фенотипическом проявлении однонаправленных изменений экспрессии гомологичных генов у референсных видов: дрозофилы и человека.

ционных полях базы знаний FlyDEGdb можно видеть на рис. 1.

Детальное описание ДЭГ *dysf* дрозофилы в ответ на различные стрессовые воздействия, а также информация о ДЭГ гомологичных генов дрозофилы и человека приведены в табл. 1. Семь колонок в левой части табл. 1 содержат экспериментальные данные о ДЭГ *dysf* дрозофилы в ответ на токсичные воздействия гербицида параквата, увеличивающего экспрессию гена, и толуола, снижающего ее. Изменения экспрессии характеризуются величиной $\log_2$(ДЭГ) и уровнем значимости этих изменений P_{ADJ}. Колонка PMID указывает источник информации.

Шесть колонок в правой части табл. 1 – результаты аннотации рассматриваемого гена *dysf* дрозофилы в сравнении с гомологичным геном *Clk* этого организма, а также с гомологичными генами человека *NPAS4* и *CLOCK* на основе четырех независимых статей PMID. Можно видеть, что: а) избыток гена *dysf* ассоциирован с нарушениями оогенеза дрозофилы; б) дефицит гена *Clk*, гомологичного гену *dysf* дрозофилы, нарушает циркадный ритм; в) избыток гена *NPAS4* человека, гомологичного гену *dysf*, повышает эффективность стрессового ответа; г) дефицит гена *CLOCK* человека, гомологичного гену *dysf* дрозофилы, нарушает циркадный ритм. В строках с 5-й по 11-ю даны еще семь аналогичных примеров.

Сравнение стресс-индуцированных ДЭГ гомологичных генов крысы и дрозофилы на основе информации из FlyDEGdb и RatDEGdb

В табл. 2 приведена информация о дифференциально экспрессирующихся генах в гипоталамусе крыс линий WAG и НИСАГ, выявленных в эксперименте с рестрикционным стрессом (ограничение подвижности животных) (Oshchepkov et al., 2024), в сравнении с ДЭГ гомологичных генов дрозофилы, описанных в базе знаний FlyDEGdb.

Рассмотрим структуру представления этой информации на примере первой строки таблицы, соответствующей дифференциально экспрессирующемуся гену *Acr* крысы. В колонке I дано имя этого гена, в колонках II и III – стресс-индуцированные изменения его экспрессии у линий крыс WAG и НИСАГ соответственно. Колонка IV указывает имя гомологичного гена дрозофилы *Jon74E*, колонка V – величину изменения экспрессии этого гена, а в колонке VI приведено суммарное количество таких ДЭГ дрозофилы, гомологичных гену *Acr*.

Колонки VII и VIII содержат значения первой (PC1) и второй (PC2) главных компонент, выявленных в результате анализа описанных выше экспериментальных данных по величине стресс-индуцированного изменения экспрессии ДЭГ дрозофилы из FlyDEGdb и гомологичных генов крысы из RatDEGdb (Oshchepkov et al., 2024) с использованием программы Past v.4.04 (Hammer et al., 2001).

Первая главная компонента (PC1) представляет собой средневзвешенную оценку общего стресс-индуцированного изменения экспрессии гомологичных генов дрозофилы ($ДЭГ_{FLY}$) и крысы ($ДЭГ_{НИСАГ}$ и $ДЭГ_{WAG}$):

$$PC1 = 0.1\log_2(ДЭГ_{FLY}) + \\ + [\log_2(ДЭГ_{НИСАГ}) + \log_2(ДЭГ_{WAG})]. \quad (1)$$

Главная компонента PC1 объясняет 65 % варианса всего набора исследуемых экспериментальных данных по ДЭГ гомологичных генов крысы и дрозофилы.

Вторая главная компонента (PC2) представляет собой средневзвешенную оценку межвидового различия дрозофилы и крысы по стресс-индуцированным изменениям экспрессии (ДЭГ) их гомологичных генов:

$$PC2 = \log_2(ДЭГ_{FLY}) - \\ - 0.1\,[\log_2(ДЭГ_{НИСАГ}) + \log_2(ДЭГ_{WAG})]. \quad (2)$$

Главная компонента PC2 объясняет 33 % варианса исследуемых экспериментальных данных.

Таким образом, нами впервые установлено, что 2/3 (65 %) варианса изменения экспрессии генов при стрессовом воздействии на крысу и дрозофилу были обусловлены общими механизмами ответа на стрессовые воздействия (PC1), тогда как 1/3 (33 %) варианса отражает межвидовые различия крысы и дрозофил (PC2).

Статистическая значимость ($p < 0.05$) выявленных нами главных компонент PC1 и PC2 была установлена на основе 1000 bootstrap-испытаний, реализованных в специальной программе в Past v.4.04 (Hammer et al., 2001) (рис. S1).

Численные значения полученных главных компонент PC1 и PC2 представлены в колонках VII и VIII табл. 2 и на рис. 3. Например, для гена крысы *Acr* и гомологичного ему гена дрозофилы *Jon74E*, описанных в первой строке табл. 2, значения PC1 и PC2 были равны –1.17 и –3.56 соответственно.

Результаты корреляционного анализа зависимости между главными компонентами PC1 и PC2 для экспериментальных данных по парам гомологичных ДЭГ крысы и дрозофилы (см. табл. 1) показаны на рис. 3. Каждая точка на этом рисунке соответствует конкретной паре «ДЭГ дрозофилы из FlyDEGdb и гомологичный ему ДЭГ крысы из RatDEGdb», для которой были вычислены значения PC1 и PC2 по формулам (1) и (2) в качестве координат этой точки на осях ординат и абсцисс. Красная штрихпунктирная линия PC1 = 0 разделяет все ДЭГ на две непересекающиеся группы: группу ДЭГ с величинами PC1 < 0, которая характеризуется стресс-индуцированным понижением экспрессии как у крыс, так и у дрозофилы; группу ДЭГ с величинами PC1 > 0, характеризующимися стресс-индуцированным повышением экспрессии у дрозофилы и у крысы, согласно формуле (1).

Две рассматриваемые группы ДЭГ (выше и ниже красной разделительной линии), выявленные нами при сравнительном анализе стресс-индуцированных изменений экспрессий гомологичных генов дрозофилы и крысы, качественно различаются. Для группы ДЭГ со стресс-индуцированным понижением экспрессии (синий цвет) наблюдается высокодостоверная ($p < 10^{-12}$) положительная зависимость между PC1 и PC2, тогда как для группы ДЭГ со стресс-индуцированным повышением экспрессии (зеленый цвет) зависимость между PC1 и PC2 отсутствует.

В связи с этим отметим неожиданное совпадение наших результатов, полученных на крысе и дрозофиле, с независимыми наблюдениями Д.Ю. Ощепкова и коллег (2025), полученными при анализе изменений экспрессий

Таблица 1. Иллюстративные примеры описания в базе знаний FlyDEGdb экспериментальных данных о ДЭГ *dysf* дрозофилы и их аннотации в сравнении с литературными данными о фенотипических проявлениях однонаправленных изменений экспрессии гомологичных им ДЭГ дрозофилы и человека

Экспериментальные данные о ДЭГ дрозофилы, доступные в FlyDEGdb								Аннотация этих ДЭГ в сравнении с гомологами у дрозофилы и человека					
№ п/п	ДЭГ дрозофилы (Fly Gene)	Линия дрозофилы (Fly strain)	Стрессовое воздействие (Phenomenon)	log_2(ДЭГ)	P_{ADJ}	Изменение экспрессии (Gene expression change)	PMID	Референс-ный вид (Reference species)	Гомологичный ДЭГ (Reference gene-homolog)	Изменение экспрессии (Gene expression change)	Фенотипический признак/ ассоциированный биологический процесс/ заболевание (Feature)	Влияние на этот признак (Health sign)	PMID
1	*dysf*	RpL13A-KI	Гербицид паракват	3.09	0.05	Избыток	30651069	Дрозофила	*dysf*	Избыток	Фертильность	Нарушение	35867785
2	*dysf*	RpL13A-KI	Гербицид паракват	3.09	0.05	Избыток	30651069	Человек	*NPAS4*	Избыток	Стресс-ответ	Улучшение	29196218
3	*dysf*	Canton-S	Растворитель толуол	–1.22	0.005	Дефицит	33484011	Дрозофила	*Clk*	Дефицит	Циркадный ритм	Нарушение	36809369
4	*dysf*	Canton-S	Растворитель толуол	–1.22	0.005	Дефицит	33484011	Человек	*CLOCK*	Дефицит	Циркадный ритм	Нарушение	15950223
5	*dysf*	WT mix #1	Лекарство менадион	0.57	0.001	Избыток	34747443	Дрозофила	*Clk*	Избыток	Старение	Оздоровление	35100266
6	*dysf*	WT mix #1	Лекарство менадион	0.57	0.001	Избыток	34747443	Человек	*BMAL1*	Избыток	Канцерогенез	Осложнение	22510946
7	*dysf*	WT mix #2	Холодовой шок	0.59	0.001	Избыток	27502401	Дрозофила	*cyc*	Избыток	Жизнеспособность потомства	Повышение	33482172
8	*dysf*	WT mix #2	Холодовой шок	0.59	0.001	Избыток	27502401	Человек	*PER1*	Избыток	Апоптоз	Улучшение	16678109
9	*dysf*	white-KO	Пищевая добавка E923	0.95	0.05	Избыток	38249074	Дрозофила	*per*	Избыток	Выживаемость	Повышение	25165772
10	*dysf*	white-KO	Пищевая добавка E923	0.95	0.05	Избыток	38249074	Человек	*PER2*	Избыток	Ку-лихорадка	Предрасположенность	22984121
11	*dysf*	w1118	Пищевая добавка E923	0.98	0.05	Избыток	38249074	Человек	*PER1*	Избыток	Канцерогенез	Отягощает	24144995

Примечание. Линии дрозофилы: RpL13A-KI – трансгенная линия дрозофилы с повышенной экспрессией гена *RpL13A*, кодирующего рибосомальный белок L13A субъединицы 40S-рибосомы; WT mix #1 – смесь коллекций, собранных в Германии, на Кипре, в Малайзии и Замбии; WT mix #2 – смесь коллекций, собранных в Швеции, Дании, Замбии и Зимбабве; white-KO – трансгенная линия дрозофилы с двойным нокаутом гена *white* (*w*). Пищевая добавка E923 – персульфат аммония (АПС). PMID: 15950223 – (Oishi et al., 2005); 16678109 – (Gery et al., 2006); 22510946 – (Elshazley et al., 2012); 22984121 – (Mehraj et al., 2012); 24144995 – (Wang et al., 2013); 25165772 – (Goda et al., 2014); 27502401 – (von Heckel et al., 2016); 29196218 – (Drouet et al., 2018); 30651069 – (Huang et al., 2019); 33482172 – (Fyie et al., 2021); 33484011 – (de Oliveira et al., 2021); 34747443 – (Ramnarine et al., 2022); 35100266 – (Jauregui-Lozano et al., 2022); 35867785 – (Wu J.W. et al., 2022); 36809369 – (Tabuloc et al., 2013); 38249074 – (Balakireva et al., 2024).

Таблица 2. Применение базы знаний FlyDEGdb к анализу ДЭГ гипоталамуса крыс линий WAG и НИСАГ в ответ на ограничение их подвижности (Oshchepkov et al., 2024)

(Oshchepkov et al., 2024)				FlyDEGdb (эта работа)			PC1, 65 %	PC2, 33 %
№ п/п	ДЭГ крысы	log_2(стресс/норма)		ДЭГ дрозофилы	log_2(стресс/норма)	N_{FlyDEG}	Общий стресс-ответ	Межвидовое различие
		WAG	НИСАГ					
	I	II	III	IV	V	VI	VII	VIII
1	*Acr*	–0.75	–0.76	*Jon74E*	–6.28	28	–1.17	–3.56
2	*Alox12**	–0.72	–0.78			0		
3	*Atp2b4*	–1.00	–0.61	*PMCA*	–0.48	7	–1.04	–0.26
4	*Cd180*	–1.19	–1.61	*Toll–7*	–3.60	43	–2.07	–1.96
5	*Cdkn1a*	0.93	1.35	*dap*	–0.79	12	2.01	–0.64
6	*Chrna7*	–0.95	–0.68	*nAChRa6*	–2.23	21	–1.12	–1.25
7	*Creb5*	–0.63	–0.66	*Atf-2*	0.28	3	–0.74	0.16
8	*Cryab*	0.67	0.82	*l(2)efl*	–1.96	22	1.35	–1.26
9	*Cyp26b1*	–0.76	–0.80	*Cyp313a2*	–5.80	714	–1.19	–3.29
10	*Ddit4*	0.61	0.65	*scyl*	–0.81	31	1.22	–0.59
11	*Dhrs9*	–1.60	–1.40	*CG8888*	–1.84	13	–2.18	–0.96
12	*Evi2b**	–0.62	–0.68			0		
13	*Fkbp5*	0.79	1.38	*Fkbp59*	–1.99	28	1.87	–1.32
14	*Flvcr2**	–0.70	–0.59			0		
15	*Fmo2*	0.69	0.96	*Fmo-2*	–2.30	56	1.46	–1.47
16	*Fosb*	1.75	1.23	*kay*	–1.12	29	2.58	–0.85
17	*Fosl1*	1.23	1.70	*kay*	–1.12	22	2.51	–0.86
18	*Fosl2*	0.68	0.74	*kay*	–1.12	22	1.33	–0.78
19	*Gpd1*	1.02	1.68	*Gpdh1*	–1.08	9	2.32	–0.83
20	*Hpd*	0.69	0.62	*Hpd*	–2.97	21	1.18	–1.82
21	*Hspa1b*	2.88	1.01	*Hsc70-4*	–0.56	14	3.37	–0.56
22	*Il17rd**	–0.76	–0.65			0		
23	*Il21r**	–0.60	–0.66			0		
24	*Lims2*	0.59	1.02	*stck*	–1.15	20	1.47	–0.81
25	*Lmod2*	1.07	1.00	*tmod*	–3.56	26	1.75	–2.20
26	*Maff*	0.59	0.82	*maf-S*	–1.72	20	1.29	–1.12
27	*Map3k6*	1.21	1.18	*Ask1*	–0.51	9	2.12	–0.48
28	*Mt2A*	0.65	0.59	*MtnA*	–8.97	156	0.89	–5.24
29	*Npas4*	1.09	–0.67	*dysf*	–1.22	6	0.61	–0.76
30	*P2ry4*	–1.14	–0.76	*PK2-R1*	–4.04	51	–2.07	–2.27
31	*Pcdh11x**	–0.61	–0.69			0		
32	*Pik3ap1*	–0.63	–0.95	*stumps*	–2.81	27	–1.08	–1.58
33	*Pla2g3*	0.78	1.86	*GIIIspla2*	–2.41	24	2.21	–1.59
34	*Plek*	–0.61	–0.90	*kmr*	–2.83	27	–1.03	–1.59
35	*Ptch1*	–0.64	–0.82	*ptc*	–1.75	17	–0.95	–0.98
36	*Rasgrp3*	–0.66	–0.62	*Sos*	–1.04	18	–0.79	–0.59
37	*Rin3*	0.59	1.07	*spri*	–1.13	14	1.51	–0.80
38	*Scrt2*	0.65	0.64	*scrt*	–1.65	24	1.21	–1.07
39	*Tmc7*	–0.87	–0.72	*Tmc*	–2.16	4	–1.08	–1.22
40	*Tnfrsf11a**	0.82	1.18			0		
41	*Ttll10*	–0.68	–0.86	*TTLL1B*	–9.99	27	–1.33	–5.67
42	*Zbtb16*	1.23	2.05	*CG43120*	1.44	36	2.87	0.57
Общее количество ДЭГ дрозофилы, гомологичных ДЭГ крысы						1601		

Примечание. N_{FlyDEG} – количество ДЭГ дрозофилы, гомологичных ДЭГ крысы согласно FlyBase (Ozturk-Colak et al., 2024). * Отмечены гены крысы (*Flvcr2*, *Alox12*, *Evi2b*, *Il17rd*, *Il21r*, *Pcdh11x*, *Tnfrsf11a*), для которых в FlyDEGdb v0.1 отсутствуют гомологичные гены дрозофилы (N_{FlyDEG} = 0).

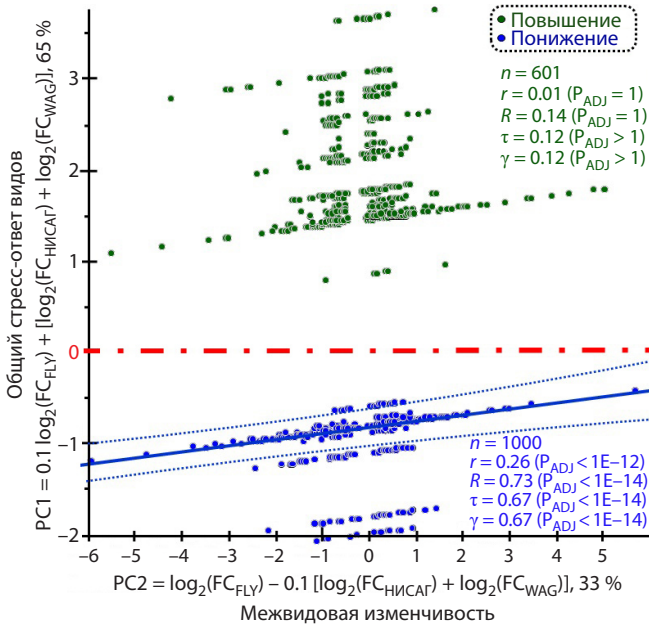


Рис. 3. Результаты корреляционного анализа зависимости между главными компонентами PC1 и PC2 для экспериментальных данных по парам ДЭГ гомологичных генов крысы и дрозофилы (из табл. 1).

Каждая точка на рисунке соответствует значениям величин PC1 и PC2, отложенным по осям ординат и абсцисс соответственно, вычисленным для определенной пары «ДЭГ дрозофилы из FlyDEGdb и гомологичный ему ДЭГ крысы из RatDEGdb». Красная штрихпунктирная линия – граница между областями диаграммы для генов со стресс-индуцированным понижением (синий цвет) и повышением (зеленый цвет) экспрессии, согласно оценке PC1 по формуле (1); непрерывная линия соответствуют линейной зависимости между PC1 и PC2 при PC1 < 0; пунктирные линии – границы 95 % доверительного интервала для нее; r, γ, R, τ, и P_{ADJ} – коэффициенты корреляций (линейной, обобщенной Гудмана–Крускала, ранговых Спирмена и Кендала), а также уровни их статистической значимости с поправкой Бонферрони на множественные сравнения, согласно оценкам пакета Statistica (Statsoft™, США).

гомологичных генов крысы и человека, индуцированных стрессом и гипертонией соответственно. В обоих случаях достоверная корреляция между первой и второй главными компонентами наблюдалась только при стресс-индуцированном понижении экспрессии гомологичных генов.

Наличие зависимости между главными компонентами PC1 и PC2 в области PC1 < 0, которая соответствует стресс-индуцированному снижению экспрессии генов человека, крысы и дрозофилы, может указывать на возможность общего молекулярного механизма подавления экспрессии генов в условиях стрессовых воздействий различной природы.

К числу общеизвестных молекулярных механизмов снижения экспрессии генов при стрессовых воздействиях могут относиться блокирование сплайсинга пре-мРНК генов эукариот (Yost, Lindquist, 1986; Cuesta et al., 2000) и ингибирование трансляции (Bresson et al., 2020).

Для оценки обогащения терминами онтологии генов (Gene Ontology, GO) (Gene Ontology Consortium, 2015) группы генов дрозофилы со стресс-индуцированным снижением экспрессии (см. табл. 2, колонка VII, а также синие точки на рис. 3) был использован информационно-программный инструмент STRING (Szklarczyk et al., 2021). Результаты представлены в табл. 3.

Всего было выявлено пять терминов GO, которыми статистически достоверно ($p < 0.05$) обогащен список исследуемых генов дрозофилы со стресс-индуцированным снижением экспрессии. Четыре из пяти терминов (GO:0005887, GO:0005892, GO:0005886, GO:0120025) непосредственно характеризуют компоненты плазматической мембраны. Пятый термин (GO:0005929, «ресничка (cilium)») тоже относится к этой группе, поскольку реснички-цилии представляют собой специфические органеллы на внешней поверхности мембран эукариотических клеток. Это свидетельствует о том, что одной из универсальных мишеней стрессовых воздействий на дрозофилу, описанных в FlyDEGdb, оказывается плазматическая мембрана ее клеток. Поэтому особо отметим, что стресс-индуцированное снижение экспрессии генов дрозофилы, кодирующих компоненты плазматических мембран клеток, может замедлять их рост при стрессовых воздействиях. Наше предположение согласуется с результатами, приведенными в обзоре (Kassahn et al., 2009) по механизмам ответа животных на стрессовые воздействия. В этой связи обратим внимание на задержку роста клеток *Escherichia coli* при стрессовых воздействиях, выявленную в работе (Haque et al., 2025).

Таблица 3. Результаты оценки обогащения терминами онтологии генов (GO) группы генов дрозофилы со стресс-индуцированным снижением экспрессии

Онтология генов (GO)			Обогащение		P_{ADJ}
№ п/п	Идентификатор онтологии генов GO:ID	Термин онтологии генов	Доля ДЭГ дрозофилы со стресс-индуцируемым снижением экспрессии	Доля GO:ID	
1	GO:0005887	Неотъемлемый компонент плазматической мембраны (integral component of plasma membrane)	12 из 56	12 из 520	0.0025
2	GO:0005892	Комплекс ацетилхолин-зависимых каналов (acetylcholine-gated channel complex)	3 из 56	3 из 7	0.005
3	GO:0005886	Плазматическая мембрана (plasma membrane)	18 из 56	18 из 1485	0.005
4	GO:0120025	Клеточный выступ, ограниченный плазматической мембраной (plasma membrane bounded cell projection)	11 из 56	11 из 717	0.05
5	GO:0005929	Ресничка (cilium)	6 из 56	6 из 188	0.05

Завершая статью, отметим, что в 2023 г. исполнилось восемьдесят лет знаменитому высказыванию Ганса Селье “Stress is the spice of life” (Rochette et al., 2023). Наши результаты являются еще одной иллюстрацией фундаментальной значимости проблематики стресса для наук о жизни.

Заключение

Создана база знаний FlyDEGdb, представляющая собой коллекцию экспериментальных данных о дифференциально экспрессирующихся генах (ДЭГ) дрозофилы при воздействии широкого круга стрессовых факторов, включая высокую и низкую температуры, обезвоживание, тяжелые металлы, радиацию, голод, яды, бытовую химию, лекарства, удобрения, инсектициды, пестициды и гербициды. FlyDEGdb, содержащая информацию о таком модельном организме, как дрозофила, может быть полезна для исследователей, работающих в области трансляционной молекулярной биологии и генетики человека и животных, физиологии, трансляционной медицины, фармакологии, диетологии, агрохимии, радиобиологии, токсикологии и биоинформатики.

Список литературы / References

Маркель А.Л. Генетическая модель индуцированной стрессом артериальной гипертензии. *Известия АН СССР.* 1985;3:466-469 [Markel A.L. Genetic model of stress-induced arterial hypertension. *Izvestiya AN SSSR*. 1985;3:466-469 (in Russian)]

Ощепков Д.Ю., Маковка Ю.В., Чадаева И.В., Богомолов А.Г., Федосеева Л.А., Серяпина А.А., Пономаренко М.П., Маркель А.Л., Редина О.Е. Гены, представляющие стресс-зависимую компоненту при развитии артериальной гипертонии. *Вавиловский журнал генетики и селекции*. 2025;29(8). doi 10.18699/vjgb-25-139 [Oshchepkov D.Yu., Makovka Yu.V., Chadaeva I.V., Bogomolov A.G., Fedoseeva L.A., Seryapina A.A., Ponomarenko M.P., Markel A.L., Redina O.E. Genes representing a stress-dependent component in the development of arterial hypertension. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2025;29(8). doi 10.18699/vjgb-25-139]

Подколодная О.А., Чадаева И.В., Филонов С.В., Подколодный Н.Л., Рассказов Д.А., Твердохлеб Н.Н., Золотарева К.А., Богомолов А.Г., Кондратюк Е.Ю., Ощепков Д.Ю., Пономаренко М.П. База знаний MiceDEGdb по дифференциально экспрессирующимся генам мыши как модельного объекта биомедицинских исследований. *Вавиловский журнал генетики и селекции*. 2025;29(1):153-161. doi 10.18699/vjgb-25-18 [Podkolodnaya O.A., Chadaeva I.V., Filonov S.V., Podkolodnyy N.L., Rasskazov D.A., Tverdokhleb N.N., Zolotareva K.A., Bogomolov A.G., Kondratyuk E.Y., Oshchepkov D.Y., Ponomarenko M.P. MiceDEGdb: a knowledge base on differentially expressed mouse genes as a model object in biomedical research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2025;29(1):153-161. doi 10.18699/vjgb-25-18]

Чадаева И.В., Филонов С.В., Золотарева К.А., Хандаев Б.М., Ершов Н.И., Подколодный Н.Л., Кожемякина Р.В., … Стефанова Н.А., Колосова Н.Г., Маркель А.Л., Пономаренко М.П., Ощепков Д.Ю. База знаний RatDEGdb по дифференциально экспрессирующимся генам крысы как модельного объекта биомедицинских исследований. *Вавиловский журнал генетики и селекции*. 2023;27(7):794-806. doi 10.18699/VJGB-23-92 [Chadaeva I.V., Filonov S.V., Zolotareva K.A., Khandaev B.M., Ershov N.I., Podkolodnyy N.L., Kozhemyakina R.V., … Stefanova N.A., Kolosova N.G., Markel A.L., Ponomarenko M.P., Oshchepkov D.Y. RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):794-806. doi 10.18699/VJGB-23-92]

Ali M.Z., Anushree, Bilgrami A.L., Ahsan J. *Drosophila melanogaster* chemosensory pathways as potential targets to curb the insect menace. *Insects*. 2022;13(2):142. doi 10.3390/insects13020142

Avgustinovich D.F., Chadaeva I.V., Kizimenko A.V., Kovner A.V., Bazovkina D.V., Ponomarev D.V., Evseenko V.I., Naprimerov V.A., Lvova M.N. The liver-brain axis under the influence of chronic *Opisthorchis felineus* infection combined with prolonged alcoholization in mice. *Vavilov J Genet Breed*. 2025;29(1):92-107. doi 10.18699/vjgb-25-11

Balakireva Y., Nikitina M., Makhnovskii P., Kukushkina I., Kuzmin I., Kim A., Nefedova L. The lifespan of *D. melanogaster* depends on the function of the *Gagr* gene, a domesticated *gag* gene of Drosophila LTR retrotransposons. *Insects*. 2024;15(1):68. doi 10.3390/insects15010068

Bresson S., Shchepachev V., Spanos C., Turowski T.W., Rappsilber J., Tollervey D. Stress-induced translation inhibition through rapid displacement of scanning initiation factors. *Mol Cell*. 2020;80(3):470-484.e8. doi 10.1016/j.molcel.2020.09.021

Brown G.R., Hem V., Katz K.S., Ovetsky M., Wallin C., Ermolaeva O., Tolstoy I., Tatusova T., Pruitt K.D., Maglott D.R., Murphy T.D. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015;43(D1):D36-D42. doi 10.1093/nar/gku1055

Chadaeva I., Ponomarenko P., Rasskazov D., Sharypova E., Kashina E., Kleshchev M., Ponomarenko M., Naumenko V., Savinkova L., Kolchanov N., Osadchuk L., Osadchuk A. Natural selection equally supports the human tendencies in subordination and domination: a genome-wide study with *in silico* confirmation and *in vivo* validation in mice. *Front Genet.* 2019;10:73. doi 10.3389/fgene.2019.00073

Chatterjee N., Perrimon N. What fuels the fly: energy metabolism in *Drosophila* and its application to the study of obesity and diabetes. *Sci Adv*. 2021;7(24):eabg4336. doi 10.1126/sciadv.abg4336

Chen J., Nolte V., Schlotterer C. Temperature stress mediates decanalization and dominance of gene expression in *Drosophila melanogaster*. *PLoS Genet.* 2015;11(2):e1004883. doi 10.1371/journal.pgen.1004883

Cuesta R., Laroia G., Schneider R.J. Chaperone Hsp27 inhibits translation during heat shock by binding eIF4G and facilitating dissociation of cap-initiation complexes. *Genes Dev*. 2000;14(12):1460-1470. doi 10.1101/gad.14.12.1460

De Gregorio E., Spellman P.T., Rubin G.M., Lemaitre B. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci USA*. 2001;98(22):12590-12595. doi 10.1073/pnas.221458698

de Oliveira D.S., Rosa M.T., Vieira C., Loreto E.L.S. Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. *J Evol Biol*. 2021;34(4):628-638. doi 10.1111/jeb.13762

Drouet J.B., Peinnequin A., Faure P., Denis J., Fidier N., Maury R., Buguet A., Cespuglio R., Canini F. Stress-induced hippocampus Npas4 mRNA expression relates to specific psychophysiological patterns of stress response. *Brain Res*. 2018;1679:75-83. doi 10.1016/j.brainres.2017.11.024

Elshazley M., Sato M., Hase T., Yamashita R., Yoshida K., Toyokuni S., Ishiguro F., Osada H., Sekido Y., Yokoi K., Usami N., Shames D.S., Kondo M., Gazdar A.F., Minna J.D., Hasegawa Y. The circadian clock gene *BMAL1* is a novel therapeutic target for malignant pleural mesothelioma. *Int J Cancer*. 2012;131(12):2820-2831. doi 10.1002/ijc.27598

Fyie L.R., Gardiner M.M., Meuti M.E. Artificial light at night alters the seasonal responses of biting mosquitoes. *J Insect Physiol*. 2021;129:104194. doi 10.1016/j.jinsphys.2021.104194

Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43(D1):D1049-D1056. doi 10.1093/nar/gku1179

Gery S., Komatsu N., Baldjyan L., Yu A., Koo D., Koeffler H.P. The circadian gene Per1 plays an important role in cell growth and DNA damage control in human cancer cells. *Mol Cell*. 2006;22(3): 375-382. doi 10.1016/j.molcel.2006.03.038

Goda T., Sharp B., Wijnen H. Temperature-dependent resetting of the molecular circadian oscillator in Drosophila. *Proc Biol Sci*. 2014; 281(1793):20141714. doi 10.1098/rspb.2014.1714

Gruntenko N.E., Khlebodarova T.M., Sukhanova M.Jh., Vasenkova I.A., Kaidanov L.Z., Rauschenbach I.Yu. Prolonged negative selection of *Drosophila melanogaster* for a character of adaptive significance disturbs stress reactivity. *Insect Biochem Mol Biol.* 1999;29(5):445-452. doi 10.1016/s0965-1748(99)00021-1

Gruntenko N.E., Deryuzhenko M.A., Andreenkova O.V., Shishkina O.D., Bobrovskikh M.A., Shatskaya N.V., Vasiliev G.V. *Drosophila melanogaster* transcriptome response to different *Wolbachia* strains. *Int J Mol Sci.* 2023;24(24):17411. doi 10.3390/ijms 242417411

Gryksa K., Schmidtner A.K., Masís-Calvo M., Rodriguez-Villagra O.A., Havasi A., Wirobski G., Maloumby R., Jägle H., Bosch O.J., Slattery D.A., Neumann I.D. Selective breeding of rats for high (HAB) and low (LAB) anxiety-related behaviour: a unique model for comorbid depression and social dysfunctions. *Neurosci Biobehav Rev*. 2023;152:105292. doi 10.1016/j.neubiorev.2023.105292

Hammer O., Harper D.A.T., Ryan P.D. PAST: PAleontological STatistics software package for education and data analysis. *Palaeontol Electron*. 2001;4(1):1-9

Haque M.F., Tarusawa T., Ushida C., Ito S., Himeno H. cAMP-CRP-activated *E. coli* causes growth arrest under stress conditions. *Front Microbiol*. 2025;16:1597530. doi 10.3389/fmicb.2025.1597530

Huang K., Chen W., Zhu F., Li P.W., Kapahi P., Bai H. RiboTag translatomic profiling of *Drosophila* oenocytes under aging and induced oxidative stress. *BMC Genomics*. 2019;20(1):50. doi 10.1186/ s12864-018-5404-4

Jauregui-Lozano J., Hall H., Stanhope S.C., Bakhle K., Marlin M.M., Weake V.M. The Clock:Cycle complex is a major transcriptional regulator of *Drosophila* photoreceptors that protects the eye from retinal degeneration and oxidative stress. *PLoS Genet*. 2022;18(1): e1010021. doi 10.1371/journal.pgen.1010021

Kassahn K.S., Crozier R.H., Portner H.O., Caley M.J. Animal performance and stress: responses and tolerance limits at different levels of biological organisation. *Biol Rev Camb Philos Soc*. 2009;84(2): 277-292. doi 10.1111/j.1469-185X.2008.00073.x

Lakhotia S.C. Tracing the roots of molecular biology. Part III: Morgan's Fly Room and emergence of modern genetics. *Reson*. 2025; 29(3):379-409. doi 10.1007/s12045-024-1669-x

Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* (*Oxford*). 2011;2011:baq036. doi 10.1093/database/baq036

Mehraj V., Textoris J., Capo C., Raoult D., Leone M., Mege J.L. Overexpression of the *Per2* gene in male patients with acute Q fever. *J Infect Dis*. 2012;206(11):1768-1770. doi 10.1093/infdis/jis600

Mikucki E.E., O'Leary T.S., Lockwood B.L. Heat tolerance, oxidative stress response tuning and robust gene activation in early-stage *Drosophila melanogaster* embryos. *Proc R Soc B*. 2024;291(2029): 20240973. doi 10.1098/rspb.2024.0973

Morgan T.H. Sex limited inheritance in *Drosophila*. *Science*. 1910; 32(812):120-122. doi 10.1126/science.32.812.120

Mukherjee P., Roy S., Ghosh D., Nandi S.K. Role of animal models in biomedical research: a review. *Lab Anim Res*. 2022;38(1):18. doi 10.1186/s42826-022-00128-1

Oishi K., Ohkura N., Amagai N., Ishida N. Involvement of circadian clock gene *Clock* in diabetes-induced circadian augmentation of plasminogen activator inhibitor-1 (*PAI-1*) expression in the mouse heart. *FEBS Lett*. 2005;579(17):3555-3559. doi 10.1016/j.febslet. 2005.05.027

Oshchepkov D.Y., Makovka Y.V., Fedoseeva L.A., Seryapina A.A., Markel A.L., Redina O.E. Effect of short-term restraint stress on the hypothalamic transcriptome profiles of rats with Inherited Stress-Induced Arterial Hypertension (ISIAH) and normotensive Wistar Albino Glaxo (WAG) rats. *Int J Mol Sci*. 2024;25(12):6680. doi 10.3390/ijms25126680

Ozturk-Colak A., Marygold S.J., Antonazzo G., Attrill H., Goutte-Gattat D., Jenkins V.K., Matthews B.B., Millburn G., Dos Santos G., Tabone C.J.; FlyBase Consortium. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024;227(1):iyad211. doi 10.1093/genetics/iyad211

Ramnarine T.J.S., Grath S., Parsch J. Natural variation in the transcriptional response of *Drosophila melanogaster* to oxidative stress. *G3* (*Bethesda*). 2022;12(1):jkab366. doi 10.1093/g3journal/jkab366

Rand M.D., Tennessen J.M., Mackay T.F.C., Anholt R.R.H. Perspectives on the *Drosophila melanogaster* model for advances in toxicological science. *Curr Protoc.* 2023;3(8):e870. doi 10.1002/cpz1.870

Rochette L., Dogon G., Vergely C. Stress: eight decades after its definition by Hans Selye: "Stress is the spice of life". *Brain Sci*. 2023; 13(2):310. doi 10.3390/brainsci13020310

Szklarczyk D., Gable A.L., Nastou K.C., Lyon D., Kirsch R., Pyysalo S., Doncheva N.T., Legeay M., Fang T., Bork P., Jensen L.J., von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605-D612. doi 10.1093/nar/gkaa1074

Tabuloc C.A., Cai Y.D., Kwok R.S., Chan E.C., Hidalgo S., Chiu J.C. CLOCK and TIMELESS regulate rhythmic occupancy of the BRAHMA chromatin-remodeling protein at clock gene promoters. *PLoS Genet*. 2023;19(2):e1010649. doi 10.1371/journal.pgen. 1010649

Telonis-Scott M., van Heerwaarden B., Johnson T.K., Hoffmann A.A., Sgro C.M. New levels of transcriptome complexity at upper thermal limits in wild *Drosophila* revealed by exon expression analysis. *Genetics.* 2013;195(3):809-830. doi 10.1534/genetics.113.156224

von Heckel K., Stephan W., Hutter S. Canalization of gene expression is a major signature of regulatory cold adaptation in temperate *Drosophila melanogaster*. *BMC Genomics*. 2016;17:574. doi 10.1186/ s12864-016-2866-0

Wade J. Practical guidelines for a user-friendly interface. *ACM SIGAPL APL Quote Quad*. 1984;14(4):365-371. doi 10.1145/384283.801122

Wang T., Yang P., Zhan Y., Xia L., Hua Z., Zhang J. Deletion of circadian gene *Per1* alleviates acute ethanol-induced hepatotoxicity in mice. *Toxicology*. 2013;314(2-3):193-201. doi 10.1016/j.tox.2013. 09.009

Wu J.W., Wang C.W., Chen R.Y., Hung L.Y., Tsai Y.C., Chan Y.T., Chang Y.C., Jang A.C. Spatiotemporal gating of Stat nuclear influx by *Drosophila* Npas4 in collective cell migration. *Sci Adv.* 2022; 8(29):eabm2411. doi 10.1126/sciadv.abm2411

Wu K., Tang Y., Zhang Q., Zhuo Z., Sheng X., Huang J., Ye J., Li X., Liu Z., Chen H. Aging-related upregulation of the homeobox gene *caudal* represses intestinal stem cell differentiation in *Drosophila*. *PLoS Genet.* 2021;17(7):e1009649. doi 10.1371/journal.pgen. 1009649

Yost H.J., Lindquist S. RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis. *Cell.* 1986;45(2):185-193. doi 10.1016/0092-8674(86)90382-x

Yu S., Luo F., Xu Y., Zhang Y., Jin L.H. *Drosophila* innate immunity involves multiple signaling pathways and coordinated communication between different tissues. *Front Immunol.* 2022;13:905370. doi 10.3389/fimmu.2022.905370

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 03.10.2025. После доработки 14.10.2025. Принята к публикации 15.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):963-977

doi 10.18699/vjgb-25-102

Реконструкция и анализ генной сети регуляции апоптоза при гепатоцеллюлярной карциноме на основе данных scRNA-seq и базы знаний ANDSystem

А.В. Адамовская [1, 2] ✉, И.В. Яцык [1, 2], М.А. Клещев [1, 2], П.С. Деменков [1, 2], Т.В. Иванисенко [1, 2], В.А. Иванисенко [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ adamovskayaav@bionet.nsc.ru

Аннотация. Гепатоцеллюлярная карцинома – наиболее распространенный первичный рак печени, характеризующийся быстрым прогрессированием, высокой летальностью и устойчивостью к терапии. Одним из ключевых направлений в изучении молекулярных механизмов развития гепатоцеллюлярной карциномы является анализ нарушений процессов апоптоза в гепатоцитах. На протяжении всей жизни благодаря апоптозу происходит элиминация старых и дефектных клеток, тогда как ослабление апоптотической гибели служит одним из ведущих факторов канцерогенеза. В настоящем исследовании выполнены реконструкция и анализ генной сети регуляции апоптоза гепатоцитов у человека на основе данных секвенирования транскриптома одиночных клеток (scRNA-seq) и базы знаний ANDSystem, использующей методы искусственного интеллекта и компьютерной системной биологии. Сравнительный анализ экспрессии генов показал ослабление транскрипции генов, вовлеченных в регуляцию воспалительных процессов и апоптоза, в опухолевых гепатоцитах по сравнению с гепатоцитами нормальной ткани печени. Реконструированная сеть включала 116 дифференциально экспрессирующихся генов, аннотированных в Gene Ontology как гены, вовлеченные в процесс апоптоза (apoptotic process GO:0006915), 116 соответствующих белков, а также 16 дополнительных белков, не имеющих GO-аннотации, но дифференциально экспрессируемых при гепатоцеллюлярной карциноме и вовлеченных во взаимодействия с генами и белками, участвующими в процессе апоптоза. Компьютерный анализ генной сети выявил ряд ключевых белков – продуктов генов *NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*, демонстрирующих как высокое число связей с другими объектами сети, так и дифференциальную экспрессию при гепатоцеллюлярной карциноме. Особый интерес представляют белки CDKN1A, ERBB2, IL8 и EGR1, не аннотированные в Gene Ontology как участники апоптоза, но обладающие статистически значимым числом взаимодействий с генами, вовлеченными в апоптоз, что указывает на их роль в регуляции программируемой клеточной гибели. Полученные результаты могут найти применение для планирования новых экспериментов по изучению роли апоптоза в канцерогенезе и поиска новых мишеней и подходов для терапии гепатоцеллюлярной карциномы, основанных на модуляции апоптоза в злокачественных гепатоцитах. Предложенный подход к реконструкции и анализу генной сети регуляции апоптоза при гепатоцеллюлярной карциноме может быть использован для анализа других форм опухолей и дает системное представление о нарушениях ключевых регуляторных процессов в онкогенезе и потенциальных мишенях для терапии.

Ключевые слова: гепатоцеллюлярная карцинома; транскриптомика одиночных клеток; апоптоз; генные сети; когнитивная система ANDSystem

Для цитирования: Адамовская А.В., Яцык И.В., Клещев М.А., Деменков П.С., Иванисенко Т.В., Иванисенко В.А. Реконструкция и анализ генной сети регуляции апоптоза при гепатоцеллюлярной карциноме на основе данных scRNA-seq и базы знаний ANDSystem. *Вавиловский журнал генетики и селекции*. 2025;29(7):963-977. doi 10.18699/vjgb-25-102

Финансирование. Работа выполнена при поддержке бюджетного проекта FWNR-2022-0020.

Reconstruction and analysis of the gene network regulating apoptosis in hepatocellular carcinoma based on scRNA-seq data and the ANDSystem knowledge base

A.V. Adamovskaya [1, 2] ✉, I.V. Yatsyk [1, 2], M.A. Kleshchev [1, 2], P.S. Demenkov [1, 2], T.V. Ivanisenko [1, 2], V.A. Ivanisenko [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ adamovskayaav@bionet.nsc.ru

© Адамовская А.В., Яцык И.В., Клещев М.А., Деменков П.С., Иванисенко Т.В., Иванисенко В.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. Hepatocellular Carcinoma (HCC) is the most common primary liver cancer characterized by rapid progression, high mortality rate and therapy resistance. One of the key areas in studying the molecular mechanisms of HCC development is the analysis of disturbances in apoptosis processes in hepatocytes. Throughout life apoptosis ensures the elimination of old and defective cells while the attenuation of this process serves as one of the leading factors in carcinogenesis. In this study we reconstructed and analyzed the gene network regulating hepatocyte apoptosis in humans based on single-cell transcriptome sequencing (scRNA-seq) data and the ANDSystem knowledge base which employs artificial intelligence and computational systems biology methods. Comparative analysis of gene expression revealed weakened transcription of genes involved in the regulation of inflammatory processes and apoptosis in tumor hepatocytes compared to hepatocytes of normal liver tissue. The reconstructed network included 116 differentially expressed genes annotated in Gene Ontology as genes involved in the apoptotic process (apoptotic process GO:0006915), along with their 116 corresponding protein products. It also included 16 additional proteins that, while lacking GO apoptosis annotation, were differentially expressed in HCC and interacting with genes and proteins participating in the apoptosis process. Computational analysis of the gene network identified several key protein products encoded by the genes *NFKB1*, *MMP9*, *BCL2*, *A4*, *CDKN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*. These proteins exhibited both a high degree of connectivity with other network objects and differential expression in HCC. Of particular interest are proteins CDKN1A, ERBB2, IL8, and EGR1, which are not annotated in Gene Ontology as apoptosis participants but have a statistically significant number of interactions with genes involved in apoptosis. This indicates their role in regulating programmed cell death. The obtained results can guide the design of new experiments studying the role of apoptosis in carcinogenesis and aid in the search for novel therapeutic targets and approaches for HCC therapy using apoptosis modulation in malignant hepatocytes. Furthermore, the proposed approach to reconstructing and analyzing the apoptosis regulation gene network in hepatocellular carcinoma can be applied to analyze other tumor forms providing a systemic understanding of disturbances in key regulatory processes in oncogenesis and potential therapy targets.

Key words: hepatocellular carcinoma; single cell transcriptomics; apoptosis; gene networks; cognitive system ANDSystem

For citation: Adamovskaya A.V., Yatsyk I.V., Kleshchev M.A., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Reconstruction and analysis of the gene network regulating apoptosis in hepatocellular carcinoma based on scRNA-seq data and the ANDSystem knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7): 963-977. doi 10.18699/vjgb-25-102

Введение

Гепатоцеллюлярная карцинома (ГЦК) является наиболее распространенным первичным раком печени, возникающим в результате злокачественной трансформации гепатоцитов. От этого заболевания в мире ежегодно умирает около 750 тыс. человек (Ganesan, Kulik, 2023). Данный вид злокачественной опухоли отличается высокой резистентностью к противоопухолевым препаратам и высокой частотой рецидивов (Zou et al., 2025), что обуславливает актуальность изучения как молекулярных механизмов опухолевого процесса, так и возникновения резистентности опухоли и на этой основе – поиска мишеней для противоопухолевых препаратов. К числу факторов риска развития ГЦК относятся, в первую очередь, хроническое инфицирование вирусами гепатита B и C, алкогольный цирроз печени и неалкогольный стеатогепатит; другие известные факторы риска развития ГЦК включают ожирение, сахарный диабет 2 типа и курение табака (Ogunwobi et al., 2019).

Вирусные инфекции и/или неблагоприятные факторы окружающей среды (воздействие гепатотоксических агентов) вызывают изменения в функционировании ряда сигнальных путей в гепатоцитах, что приводит к их злокачественному перерождению и развитию ГЦК. Установлено, что белок HBx (Hepatitis B virus X protein), кодируемый геномом вируса гепатита B, подавляет активность проапоптотического белка p53, репарацию ДНК и активирует ряд сигнальных каскадов: STAT, NF-κB, AP-1 и т. д., вовлеченных в пролиферацию и выживание клеток, стимулируя прогрессирование ГЦК (Jiang Y. et al., 2019). В патогенез ГЦК вовлечены изменения: а) функционирования путей передачи сигналов факторов роста, таких как инсулиноподобный фактор роста (IGF), эпидермальный фактор роста (EGF), тромбоцитарный фактор роста (PDGF), фактор роста фибробластов (FGF) и фактор роста гепатоцитов (HGF/MET); б) сигнальных путей, связанных с дифференцировкой клеток, таких как WNT, Hedgehog и Notch; а также в) сигнальных путей, связанных с ангиогенезом, запускаемых фактором роста эндотелия сосудов (VEGF) и FGF (Dhanasekaran et al., 2016). Кроме того, важнейший вклад в прогрессирование ГЦК вносит нарушение процессов апоптоза – программируемой клеточной гибели (Fabregat, 2009). Хроническое воспаление печени в результате инфицирования вирусами гепатита B и C или воздействия неблагоприятных факторов внешней среды приводит к апоптозу гепатоцитов, который сопровождается компенсаторным увеличением их пролиферации, что в условиях высокого уровня оксидативного стресса, вызываемого воспалением, приводит к накоплению мутаций в ДНК и повышению вероятности злокачественного перерождения гепатоцитов (Yang et al., 2019). Кроме того, апоптоз играет важнейшую роль в элиминации малигнизированных клеток, поэтому активация апоптоза – один из механизмов действия противоопухолевых препаратов при ГЦК (Hajizadeh et al., 2023). Показано, что подавление внешнего и внутреннего путей апоптоза, в частности, регуляторными микроРНК, может быть ассоциировано с развитием ГЦК и неблагоприятным исходом этого заболевания (Khlebodarova et al., 2023). Установлено, что белок вируса гепатита B HBx подавляет активность проапоптотического белка p53, что способствует возникновению и прогрессированию ГЦК (Jiang Y. et al., 2019). Имеющиеся данные свидетельствуют о том, что нарушение баланса про- и антиапоптотических белков в гепатоцитах является

одним из факторов развития ГЦК и формирования лекарственной резистентности опухоли (Ladd et al., 2024; Wu et al., 2024). Это обусловливает необходимость изучения механизмов нарушения путей апоптоза гепатоцитов при развитии ГЦК и поиска ключевых регуляторных звеньев апоптоза, экспрессия которых различается в здоровых и опухолевых гепатоцитах.

Известно, что важную роль в прогрессии заболевания играют нарушения взаимодействий между опухолевыми клетками, стромой и клетками иммунной системы, способствующие развитию ГЦК, формированию лекарственной резистентности и рецидивам (Xue et al., 2022). Отметим также, что ГЦК отличается высокой степенью клеточной гетерогенности, в связи с чем особую значимость приобретают методы исследования молекулярных процессов развития ГЦК на уровне отдельных клеток (Li X. et al., 2022).

Один из таких методов, позволяющий дать ценную информацию об особенностях экспрессии генов в различных типах клеток опухолевой ткани, – секвенирование транскриптомов одиночных клеток. Это особенно актуально при сравнении злокачественно трансформированных гепатоцитов, находящихся внутри опухоли, с нормальными гепатоцитами из гистологически не измененной ткани печени (Zhang et al., 2022). Однако одного лишь анализа дифференциальной экспрессии недостаточно для понимания механизмов опухолевой трансформации. На основе получаемых таким образом экспериментальных данных необходима реконструкция генных сетей – ансамблей координированно функционирующих генов, дающая ценную информацию о нарушениях молекулярных механизмов взаимодействия между генами, ответственных за развитие патологических процессов (Saik et al., 2019; Ivanisenko V.A. et al., 2022; Antropova et al., 2023; Butikova et al., 2025).

Цель нашей работы состояла в реконструкции и анализе генной сети регуляции апоптоза гепатоцитов при гепатоцеллюлярной карциноме человека с использованием интегрированного подхода, основанного на комбинации данных одноклеточного транскриптомного анализа и программно-информационной системы ANDSystem, предназначенной для реконструкции генных сетей на основе автоматического анализа текстов научных публикаций и биомедицинских фактографических баз данных (Demenkov et al., 2011; Ivanisenko V.A. et al., 2015, 2019). Система использует методы искусственного интеллекта и онтологическое описание предметной области, что обеспечивает высокую полноту и точность извлечения знаний из различных источников экспериментальной информации (Ivanisenko T.V. et al., 2020, 2022, 2024).

На основе сравнения транскриптомных данных scRNA-seq для нормальных гепатоцитов и гепатоцитов, злокачественно трансформированных при ГЦК, мы выявили 1853 дифференциально экспрессирующихся гена (ДЭГ). С использованием системы ANDSystem была реконструирована сеть взаимодействий между ДЭГ и генами, аннотированными в Gene Ontology (GO) как участники апоптоза (GO:0006915). Анализ полученной генной сети позволил выделить ряд ДЭГ, продукты которых, включая BCL2, NFKB1, FOXO1, MCL1, CDKN1A, ERBB2, IL8, EGR1, демонстрируют значимую связанность с компонентами сети апоптоза. При этом некоторые из указанных выше белков: CDKN1A, ERBB2, IL8, EGR1 – не были аннотированы в Gene Ontology как участники апоптоза, что подчеркивает их потенциальную новизну и значимость для понимания механизмов программируемой клеточной гибели при ГЦК. Кроме того, на основе анализа данных scRNA-seq было выявлено снижение экспрессии ключевых ингибиторов апоптоза в клетках гепатоцеллюлярной карциномы. Этот результат позволяет предположить, что уклонение от апоптоза при ГЦК может быть обусловлено не усилением антиапоптотических механизмов, а, напротив, нарушением работы проапоптотических сигнальных путей. Полученные результаты могут найти применение при планировании дальнейших экспериментальных исследований, направленных на изучение механизмов регуляции апоптоза в гепатоцитах при ГЦК. Они также представляют интерес для разработки таргетных терапевтических стратегий, направленных на модуляцию апоптотических процессов в опухолевых клетках печени.

Материалы и методы

База данных GEO. Для анализа использованы данные секвенирования транскриптома одиночных клеток, полученные из образцов гепатоцеллюлярной карциномы (первичная опухоль) и парных гистологически нормальных тканей печени, доступные в базе данных Gene Expression Omnibus (GEO) NCBI под идентификатором GSE149614. Были проанализированы данные восьми пациентов (3, 4, 5, 6, 7, 8, 9 и 10-го).

Анализ транскриптомных данных. Обработку данных секвенирования транскриптома одиночных клеток и последующий анализ выполняли с использованием пакета Scanpy (v1.9.3) (Wolf et al., 2018) в среде Python. Начальная фильтрация данных включала: 1) удаление клеток, в которых экспрессия была детектирована менее чем для 100 генов, и 2) удаление генов, экспрессия которых была детектирована менее чем в трех клетках. Нормализацию выполняли с помощью функции scanpy.pp.normalize_total() с последующим преобразованием log1p. Кластеризацию клеток проводили с использованием алгоритма Лейдена (Traag et al., 2019). Дифференциально экспрессирующиеся (маркерные) гены для каждого выявленного кластера клеток идентифицировали с помощью функции scanpy.tl.rank_genes_groups(), использующей критерий суммы рангов Уилкоксона.

На основании экспрессии известных маркерных генов гепатоцитов (ALB, HNF4A, SERPINA1, CYP3A4, TAT, TF) (Si-Tayeb et al., 2010) и результатов кластеризации отбирали клетки, классифицированные как гепатоциты. Для последующего сравнительного анализа экспрессии между опухолевыми и нормальными гепатоцитами формировали псевдообъемные (pseudobulk) (Squair et al., 2021) образцы для каждого пациента путем агрегации значений экспрессии по всем клеткам отдельно для опухолевой и нормальной ткани.

Статистически значимые различия в экспрессии генов между группой образцов pseudobulk опухолевых клеток и группой образцов pseudobulk нормальных гепатоцитов выявляли с помощью пакета DESeq2 (v1.42.0) (Love et

al., 2014) в среде R. Дифференциально экспрессирующиеся гены определяли на основе порога p-value < 0.05 и |logFC| > 0.5.

Построение генных сетей. Реконструкцию и анализ генной сети регуляции апоптоза гепатоцитов при гепатоцеллюлярной карциноме у человека выполняли с использованием программно-информационной системы ANDSystem (Demenkov et al., 2011; Ivanisenko V.A. et al., 2015, 2019). Эффективность ANDSystem была ранее показана в ряде исследований, в том числе при реконструкции сети регуляции апоптоза эндотелия при лимфедеме (Saik et al., 2019), а также при изучении молекулярных механизмов, связанных с гепатоцеллюлярной карциномой (Demenkov et al., 2023; Khlebodarova et al., 2023). Система также использовалась для интерпретации омиксных данных: метаболомных (Ivanisenko V.A. et al., 2022, 2024) и протеомных (Momynaliev et al., 2010; Larina et al., 2015), продемонстрировав свою универсальность и применимость к различным типам данных и заболеваниям.

Реконструкция генной сети включала несколько этапов. На первом этапе с помощью Мастера запросов программного модуля ANDVisio (Demenkov et al., 2011), представляющего собой графический пользовательский интерфейс в системе ANDSystem, была реконструирована ассоциативная генная сеть, включающая гены и их белковые продукты, участвующие в апоптозе. Список белок-кодирующих генов человека, вовлеченных в процесс апоптоза, получили из базы данных (БД) The Gene Ontology Resource (https://geneontology.org/) для термина GO:0006915 apoptotic process.

На втором этапе был выполнен поиск новых белков, участвующих в регуляции апоптоза в гепатоцитах при развитии ГЦК. В качестве таковых мы рассматривали белки, не аннотированные в БД The Gene Ontology Resource как участники апоптоза, но регулирующие экспрессию исходных генов, вовлеченных в процесс апоптоза. Для поиска таких белков с помощью Мастера путей программы ANDVisio были найдены все прямые связи типа *Expression regulation*, *Expression upregulation*, *Expression downregulation*, *Interaction* от белковых продуктов всех ДЭГ, выявленных в эксперименте, к ДЭГ, вовлеченным, по данным Gene Ontology, в процесс апоптоза.

Затем была проведена оценка статистической значимости уровня специфичности связи найденных белков с исходной генной сетью апоптоза, построенной на этапе 1. Показатель специфичности был определен как доля взаимодействий белка с генами сети по отношению к общему количеству взаимодействий этого белка на уровне генома. Статистическая значимость отклонения между наблюдаемым числом взаимодействий анализируемого белка с генами сети и числом взаимодействий, ожидаемым по случайным причинам, была оценена с использованием гипергеометрического распределения:

$$P(X \geq x) = \sum_{k=x}^{\min(N,\, n)} \frac{\left(\frac{n}{k}\right)\left(\frac{M-n}{N-k}\right)}{\frac{M}{N}},$$

где M – общее количество белок-кодирующих генов в базе данных, n – количество генов в анализируемой генной сети, N – общее количество генов человека, взаимодействующих с исследуемым белком, x – количество генов сети, взаимодействующих с исследуемым белком.

Расчет p-значений выполняли с использованием библиотеки Python (scipy.stats.hypergeom). Для коррекции на множественное тестирование применяли поправку Бонферрони (Наркевич и др., 2020), при которой ДЭГ считались статистически значимыми, если их скорректированное по указанной поправке p-значение удовлетворяло условию $p < 0.05$. Все вычисления выполняли с использованием библиотеки statsmodels и других стандартных инструментов Python.

Таким образом, финальная генная сеть регуляции апоптоза при развитии ГЦК включала как ДЭГ и их продукты, аннотированные в Gene Ontology как вовлеченные в процесс апоптоза, так и белковые продукты ДЭГ, статистически значимо связанные с этой генной сетью, но, по данным Gene Ontology, не участвующие в апоптозе.

Анализ генной сети. Для каждого компонента сети (гена или белка) с помощью ANDSystem рассчитывался показатель Network Connectivity, который представляет собой число других объектов сети, с которыми связан данный компонент. Хабами генной сети считались белки и гены, значение Network Connectivity которых превышало критическое значение (квантиль), соответствующее значению p-value = 0.05. Значение квантиля рассчитывалось на основании анализа наблюдаемого распределения Network Connectivity всех вершин генной сети. Таким образом, количество связей в генной сети для хабов было статистически значимым с p-value < 0.05.

Филостратиграфический анализ генной сети. Эволюционный возраст генов определяли с использованием базы данных GenOrigin (http://chenzxlab.hzau.edu.cn/) (Tong et al., 2021), которая содержит информацию об эволюционном возрасте генов различных организмов, установленную с помощью филостратиграфического анализа. Для оценки статистической значимости различий в распределении генов разного возраста между полным набором белок-кодирующих генов человека и генами в реконструированной сети апоптоза гепатоцитов при развитии ГЦК был применен гипергеометрический тест. Вероятность наблюдения m или более генов определенного возрастного интервала среди M генов сети рассчитывали с применением функции hypergeom.pmf из библиотеки SciPy. Анализ проводили для 20 возрастных интервалов, которые представлены в базе данных GenOrigin. При расчетах использовали следующие параметры: N – общее количество белок-кодирующих генов человека, n – количество белок-кодирующих генов человека в определенном возрастном интервале, M – количество генов в реконструированной генной сети, m – количество генов сети в анализируемом возрастном интервале. Различия считали статистически значимыми при $p < 0.05$.

Функциональная аннотация наборов генов. Функциональная аннотация набора генов, представленных в сети, проводилась с использованием веб-инструмента Database for Annotation, Visualization and Integrated Discovery (DAVID 2021) (https://david.ncifcrf.gov/; Sherman et al., 2022) с настройками по умолчанию. Анализ перепред-

Таблица 1. Перепредставленные термины Gene Ontology для генов с повышенной и пониженной экспрессией в опухолевых гепатоцитах по сравнению с гепатоцитами гистологически нормальной ткани печени при гепатоцеллюлярной карциноме

Гены с повышенной экспрессией				Гены с пониженной экспрессией			
№ п/п	Термин Gene Ontology	%*	*p*-value**	№ п/п	Термин Gene Ontology	%*	*p*-value**
1	GO:0051301~cell division (клеточное деление)	7.7	0.000	1	GO:0007165~signal transduction (передача сигналов)	12.4	0.000
2	GO:0007059~chromosome segregation (сегрегация хромосом)	3.6	0.000	2	GO:0035556~intracellular signal transduction (внутриклеточная передача сигналов)	5.5	0.000
3	GO:0006325~chromatin organization (организация хроматина)	3.1	0.020	3	GO:0045944~positive regulation of transcription by RNA polymerase II (позитивная регуляция транскрипции РНК полимеразой II)	10.9	0.000
4	GO:0006281~DNA repair (репарация ДНК)	3.7	0.005	4	GO:0000122~negative regulation of transcription by RNA polymerase II (негативная регуляция транскрипции РНК полимеразой II)	9.4	0.000
5	GO:0006260~DNA replication (репликация ДНК)	2.8	0.000	5	GO:0045893~positive regulation of DNA-templated transcription (позитивная регуляция транскрипции)	6.6	0.021
6	GO:0000398~mRNA splicing, via spliceosome (сплайсинг мРНК в сплайсосомах)	2.8	0.004	6	GO:0006915~apoptotic process (процесс апоптоза)	7.0	0.000
7	GO:0006364~rRNA processing (процессинг рибосомальной РНК)	2.7	0.000	7	GO:0043065~positive regulation of apoptotic process (позитивная регуляция апоптоза)	3.9	0.012
8	GO:0006412~translation (трансляция)	3.4	0.000	8	GO:0043066~negative regulation of apoptotic process (негативная регуляция апоптоза)	5.3	0.017
9	GO:0032543~mitochondrial translation (трансляция в митохондриях)	2.4	0.000	9	GO:0006954~inflammatory response (воспалительный процесс)	4.7	0.019
10	GO:0006457~protein folding (фолдинг белка)	3.3	0.000	10	GO:0016477~cell migration (миграция клеток)	4.0	0.000

* Доля генов, ассоциированных с данным термином, по отношению к общему числу генов с повышенной или сниженной экспрессией; ** *p*-value для статистической значимости перепредставленности термина Gene Ontology с поправкой Бонферрони–Шидака.

ставленности терминов Gene Ontology, описывающих биологические процессы, молекулярные функции, клеточные компоненты, а также пути KEGG (иначе называемый анализом обогащения набора генов), который позволяет определить ключевые биологические процессы, связанные с исследуемыми генами, выполняли для набора всех ДЭГ, выявленных при анализе транскриптома гепатоцитов, и ДЭГ, входящих в генную сеть регуляции апоптоза гепатоцитов. Оценка перепредставленности терминов Gene Ontology и путей KEGG в DAVID выполнялась с использованием точного критерия Фишера (Sherman et al., 2022). Показателем статистической значимости обогащения считался p-value < 0.05 с поправкой Бонферрони–Шидака (Šidák, 1967).

Результаты

Анализ дифференциальной экспрессии генов при ГЦК

В результате сравнения транскриптомов одиночных клеток – злокачественно трансформированных гепатоцитов опухоли и гепатоцитов нормальной ткани печени – было выявлено 1853 дифференциально экспрессирующихся гена. Данные по этим ДЭГ приведены в табл. S1[1]. Среди этих ДЭГ 964 гена характеризовались увеличенной экспрессией, а 889 генов – сниженной экспрессией в опухолевых гепатоцитах по сравнению с нормальными клетками печени. Результаты функциональной аннотации ДЭГ, проведенной с использованием web-ресурса DAVID, а именно список статистически значимо перепредставленных терминов Gene Ontology и путей KEGG, приведены в табл. S2 и S3. Десять наиболее значимых терминов (с наиболее высокой долей ДЭГ, ассоциированных с данным термином, по отношению к общему числу ДЭГ), обозначающих биологические процессы для генов с повышенной и пониженной экспрессией, приведены в табл. 1.

Для генов с повышенной экспрессией в злокачественно трансформированных клетках были достоверно перепредставлены термины, связанные с делением клеток (№ 1, 2 в табл. 1), организацией хроматина (№ 3 в табл. 1), репарацией и репликацией ДНК (№ 4, 5 в табл. 1), сплайсингом мРНК (№ 6 в табл. 1), процессингом рибосомальной РНК (№ 7 в табл. 1), трансляцией белка (№ 8, 9 в табл. 1) и фол-

[1] Табл. S1–S7, а также рис. S1 и S2 см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx33.xlsx

дингом белка (№ 10 в табл. 1). Для генов с повышенной экспрессией были достоверно перепредставлены пути KEGG, связанные с окислительным фосфорилированием (hsa00190:Oxidative phosphorylation) и репликацией ДНК (hsa03030:DNA replication) (см. табл. S2).

Для генов с пониженной экспрессией были достоверно перепредставлены термины, описывающие внутриклеточную передачу сигналов (№ 1, 2 в табл. 1), регуляцию транскрипции (№ 3–5 в табл. 1), позитивную и негативную регуляцию апоптоза (№ 6–8 в табл. 1) воспаление (№ 9 в табл. 1), миграцию клеток (№ 10 в табл. 1), сигнальные пути рецепторов Т клеток (№ 10 в табл. S3), рецепторов тирозин-киназы (№ 11 в табл. S3).

Для генов с повышенной экспрессией были достоверно перепредставлены пути KEGG: сигнальный путь MAPK (hsa04010:MAPK signaling pathway), NF-κB (hsa04064: NF-κB signaling pathway), хемокинов (hsa04062: chemokine signaling pathway) и Т-клеток (hsa04660:T-cell receptor signaling pathway) (см. табл. S3).

Генная сеть ДЭГ, вовлеченных в процесс апоптоза, по данным Gene Ontology

Как указано в разделе «Материалы и методы», реконструкция генной сети регуляции апоптоза в гепатоцитах при развитии ГЦК осуществлялась в два этапа. С учетом известной важности апоптоза при развитии ГЦК (Hajizadeh et al., 2023; Ladd et al., 2024; Wu et al., 2024), а также выявленной в нашем исследовании перепредставленности процессов апоптоза для генов с пониженной экспрессией (см. табл. 1) в малигнизированных гепатоцитах на первом этапе в генную сеть были включены гены и их белковые продукты, которые, по данным Gene Ontology, вовлечены в апоптоз и экспрессия которых в злокачественно трансформированных гепатоцитах отличалась от таковой в гепатоцитах гистологически нормальной ткани печени. Из 746 белок-кодирующих генов (табл. S4), аннотированных в БД The Gene Ontology Resource для термина «процесс апоптоза» (GO:0006915 apoptotic process), 116 (16 % от общего числа генов, аннотированных в БД для данного термина) были дифференциально экспрессирующимися в малигнизированных гепатоцитах. Из них 49 генов характеризовались повышенной экспрессией, а 67 генов – пониженной экспрессией в злокачественно перерожденных гепатоцитах по сравнению со здоровыми клетками печени, что составило 42.2 и 57.8 % соответственно от числа 116 ДЭГ, вовлеченных в апоптоз. Ассоциативная генная сеть, реконструированная с помощью системы ANDSystem (рис. S1), включала 116 ДЭГ, вовлеченных в процесс апоптоза, и 116 их белковых продуктов. Характеристики данной сети приведены в табл. 2 (колонка Генная сеть, этап 1), ее изображение на рис. S1, а список всех ее компонентов (белков и генов) – в табл. S5.

На втором этапе для выявления новых белков-регуляторов апоптоза при малигнизации гепатоцитов генная сеть, реконструированная на первом этапе, была расширена белковыми продуктами всех ДЭГ, выявленных при сравнительном анализе транскриптома злокачественно трансформированных гепатоцитов и гепатоцитов гистологически нормальной ткани печени. При расширении сети были выбраны типы связей, относящиеся к регуляции экспрессии генов, – регуляция экспрессии (*expression regulation*), повышение экспрессии (*expression upregulation*), подавление экспрессии (*expression downregulation*), взаимодействие (*interaction*). Установлено, что из 116 ДЭГ апоптоза экспрессия 68 генов (59 %) регулируется 223 белками, являющимися продуктами генов, которые демонстрируют дифференциальную экспрессию в опухолевых гепатоцитах по сравнению с таковыми нормальной ткани печени, но не аннотированы в Gene Ontology как гены, вовлеченные в процесс апоптоза. Список этих генов приведен в табл. S6. Из них 102 генов характеризовались повышенной экспрессией, а экспрессия 121 гена была понижена.

По данным функциональной аннотации, для генов с пониженной экспрессией достоверно перепредставлены (p-value с поправкой Бонферрони < 0.05) термины, описывающие биологические процессы адгезии лейкоцитов (GO:0007159~leukocyte cell-cell adhesion), хемотаксиса нейтрофилов (GO:0030593~neutrophil chemotaxis), клеточного деления (GO:0051301~cell division), а также регуляцию сигнального пути PI3K/Akt (GO:0051897~positive regulation of phosphatidylinositol 3-kinase/protein kinase B signal transduction).

Далее для 223 найденных белков-кандидатов, потенциально вовлеченных в регуляцию апоптоза гепатоцитов при развитии ГЦК, была проведена оценка статистической значимости уровня их специфической связанности с генной сетью регуляции апоптоза гепатоцитов. Для каждого белка рассчитывали вероятность того, что наблюдаемая доля его взаимодействий с генами сети по отношению к общему числу его взаимодействий с белок-кодирующими генами человека могла возникнуть случайно. В результате анализа было идентифицировано 16 ДЭГ (11 генов с пониженной экспрессией и 5 генов с повышенной экспрессией), статистически значимо (p-value < 0.05 с поправкой Бонферрони) связанных с 43 генами апоптоза (табл. 3). Как видно из табл. 3, наибольшим числом связей с ДЭГ, которые уже аннотированы в Gene Ontology как участники апопптоза, обладают продукты генов *IL8*, *ERBB2*, *EGR1*, *TGFB2*, *CDKN1A*. Белки, кодируемые генами *CDN1A*, *ETS2*, *EGR1*, *BACH2*, *KLF5*, *FEN1*, являются транскрипционными факторами согласно БД The Human Transcription Factors (Lambert et al., 2018; https://humantfs.ccbr.utoronto.ca/).

Итоговая генная сеть апоптоза гепатоцитов при развитии ГЦК представлена на рис. S2, а ее характеристики представлены в табл. 2 (колонка Генная сеть, этап 2). Список всех белков и генов генной сети представлен в табл. S7. Как видно из табл. 2, при расширении первоначальной генной сети апоптоза белками, регулирующими экспрессию генов апоптоза, увеличилось число связей всех типов, за исключением понижающей регуляции (downregulation). Хабы генной сети, т. е. объекты сети (гены или белки), значение Network Connectivity (число остальных объектов сети, связанных с данным объектом) которых превышало критическое значение (квантиль), соответствующее значению p-value = 0.05 (см. раздел «Материалы и методы»), приведены в табл. 4.

Таблица 2. Характеристики ассоциативных сетей генов и белков, вовлеченных в апоптоз гепатоцитов при гепатоцеллюлярной карциноме

Показатель	Генная сеть	
	первый этап	второй этап
Число компонентов сети	238	248
генов	116	116
белков	116	132
Число взаимодействий	1512	1933
Из них следующие типы взаимодействий:		
Экспрессия генов		
Экспрессия (expression)	116	116
Дифференциальная экспрессия (differential expression)	2	2
Коэкспрессия (coexpression)	7	7
Взаимодействия белков		
Взаимодействие (interaction)	259	385
Катализ (catalyze)	21	29
Расщепление (cleavage)	2	5
Модификация (modification)	34	50
Регуляторные взаимодействия		
Регуляция (regulation)	85	95
Повышающая регуляция (upregulation)	69	79
Понижающая регуляция (downregulation)	24	24
Подавление экспрессии (expression downregulation)	122	134
Регуляция экспрессии (expression regulation)	309	385
Повышение экспрессии (expression upregulation)	165	213
Подавление активности (activity downregulation)	35	47
Регуляция активности (activity regulation)	60	79
Повышение активности (activity upregulation)	25	35
Подавление модификации (modification downregulation)	15	17
Регуляция модификации (modification regulation)	54	64
Усиление модификации (modification upregulation)	51	60
Подавление деградации (degradation downregulation)	9	12
Регуляция деградации (degradation regulation)	17	34
Усиление деградации (degradation upregulation)	7	21
Регуляция транспорта (transport regulation)	24	40

Всего выявлено 11 хабов генной сети (см. табл. 4), 10 из которых являются белками, и 1 ген *MMP9*, продукт которого также вошел в число хабов генной сети. Согласно данным scRNA-seq (см. табл. S1) экспрессия генов, кодирующих 3 белка, CDK1, MMP9, G3P, оказалась повышенной в злокачественно трансформированных гепатоцитах по сравнению с гепатоцитами гистологически нормальной ткани печени, а экспрессия генов, кодирующих 7 остальных белков, NFKB1, BCL2, A4, CDN1A, ERBB2, MCL1, FOXO1, была понижена. Гены, кодирующие два хаба генной сети, белки CDN1A и ERBB2, ранее не были аннотированы в Gene Ontology в качестве участников процесса апоптоза.

Сеть регуляции экспрессии генов, вовлеченных в апоптоз гепатоцитов при развитии гепатоцеллюлярной карциномы

Учитывая выявленные при анализе scRNA-seq изменения в экспрессии генов, продукты которых вовлечены в апоптоз гепатоцитов при развитии ГЦК, мы провели анализ

Таблица 3. Список белков, кодируемых ДЭГ малигнизированных гепатоцитов, вовлеченных в регуляцию апоптоза при гепатоцеллюлярной карциноме, которые не аннотированы в Gene Ontology как участники апоптоза (GO:0006915 apoptotic process)

№ п/п	Название белка	Имя белка	Число связей		Экспрессия	*p*-value
			ДЭГ	Все		
1	IL8	Хемокиновый лиганд 8 с мотивом C-X-C (C-X-C motif chemokine ligand 8)	10	25	Понижена	0.00000
2	ERBB2	Рецептор тирозин-киназы erb-b2 (erb-b2 receptor tyrosine kinase 2)	9	32		0.00025
3	**EGR1**	Фактор раннего ответа на рост 1 (early growth response 1)	7	23		0.01576
4	**CDKN1A**	Ингибитор циклин-зависимой киназы 1A (cyclin dependent kinase inhibitor 1A)	6	35		0.00004
5	TGFB2	Трансформирующий фактор роста бета 2 (transforming growth factor beta 2)	6	14		0.00861
6	**ETS2**	Транскрипционный фактор ETS протоонкоген 2 (ETS proto-oncogene 2 transcription factor)	5	8		0.00018
7	KLF5	Транскрипционный фактор KLF5 (KLF transcription factor 5)	5	13	Повышена	0.00196
8	SDF1	Хемокиновый лиганд 12 с мотивом C-X-C (C-X-C motif chemokine ligand 12)	5	15	Понижена	0.04080
9	GELS	Гельсолин (gelsolin)	4	14		0.00120
10	K2C7	Кератин 7 (keratin 7)	3	3		0.00071
11	IMA1	Кариоферин субъединица альфа 2 (karyopherin subunit alpha 2)	3	12	Повышена	0.00198
12	**FEN1**	Эндонуклеаза FEN1 (flap structure-specific endonuclease 1)	3	8		0.00576
13	NEP	Неприлизин (neprilysin)	3	9	Понижена	0.00765
14	CDC20	Белок цикла клеточного деления 20 (cell division cycle 20)	3	19	Повышена	0.01919
15	NEUT	Нейротензин (neurotensin)	3	4		0.02776
16	**BACH2**	Фактор транскрипции BACH2 (BTB domain and CNC homolog 2)	3	7	Понижена	0.03851

Примечание. Число связей ДЭГ – число связей (регуляция экспрессии) белка с генами, участвующими в апоптозе согласно Gene Ontology; Число связей Все – число связей белка со всеми компонентами итоговой генной сети (генами и белками); Экспрессия – направление изменения экспрессии гена в опухолевых гепатоцитах по сравнению с нормальными клетками (Повышена – повышенная экспрессия; Понижена – пониженная экспрессия); *p*-value – статистическая значимость связи белка с генами апоптоза, вычисленная по гипергеометрическому тесту с поправкой Бонферрони. Белки отсортированы в порядке убывания значимости связи с генной сетью апоптоза. Жирным шрифтом выделены транскрипционные факторы, согласно БД The Human Transcription Factors (Lambert et al., 2018; https://humantfs.ccbr.utoronto.ca/).

Таблица 4. Хабы генной сети апоптоза гепатоцитов при гепатоцеллюлярной карциноме у человека

№ п/п	Тип объекта в сети	Имя объекта в сети	Имя белка/гена	Число связанных объектов сети	*p*-value	Экспрессия
1	Белок	NFKB1	Nuclear factor kappa B subunit 1	87	0.004	Понижена
2	Ген	MMP9	Matrix metallopeptidase 9	48	0.008	Повышена
3	Белок	BCL2	BCL2 apoptosis regulator	46	0.012	Понижена
4	Белок	A4	Amyloid beta precursor protein	43	0.016	
5	Белок	**CDN1A**	Cyclin dependent kinase inhibitor 1A	35	0.020	
6	Белок	CDK1	Cyclin dependent kinase 1	33	0.024	Повышена
7	Белок	**ERBB2**	Erb-b2 receptor tyrosine kinase 2	31	0.028	Понижена
8	Белок	MMP9	Matrix metallopeptidase 9	29	0.036	Повышена
9	Белок	G3P	Glyceraldehyde-3-phosphate dehydrogenase	29	0.036	
10	Белок	MCL1	MCL1 apoptosis regulator, BCL2 family member	27	0.044	Понижена
11	Белок	FOXO1	Forkhead box O1	27	0.044	

Примечание. *p*-value – критическое значение (квантиль), рассчитанное на основании анализа наблюдаемого распределения Network Connectivity всех вершин генной сети. Жирным шрифтом отмечены белки, ранее не аннотированные в Gene Ontology как участники процесса апоптоза.

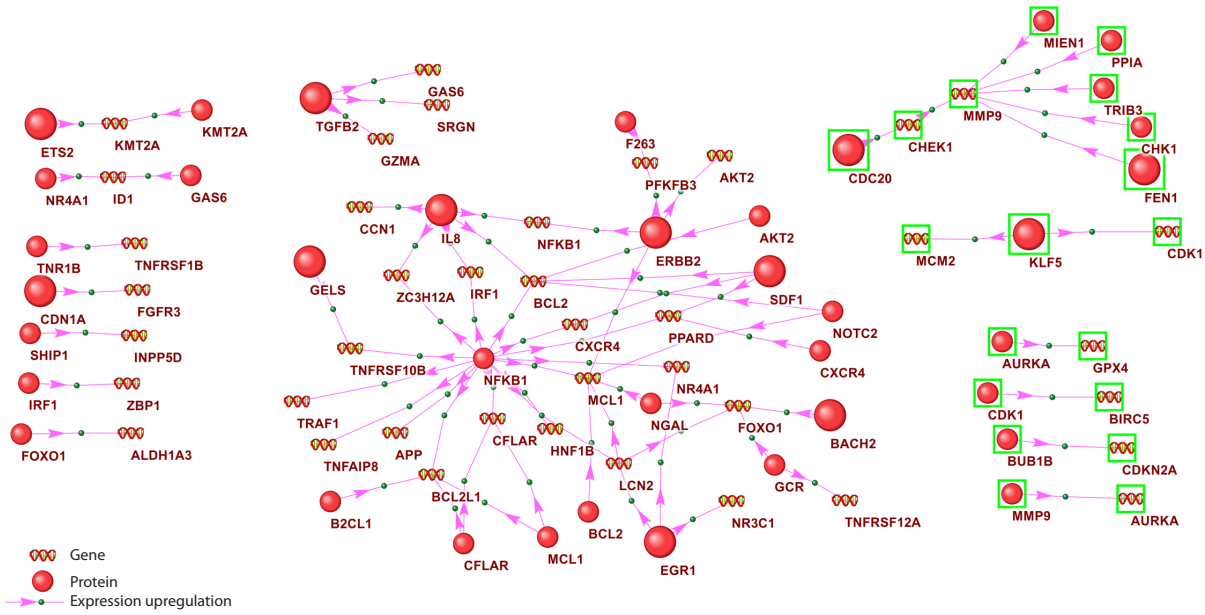


Рис. 1. Генная сеть активации экспрессии генов-компонентов генной сети регуляции апоптоза при развитии ГЦК.

Представлены только белки-компоненты генной сети регуляции апоптоза гепатоцитов при развитии ГЦК (см. рис. S2), активирующие (тип связи expression upregulation) экспрессию генов-компонентов той же сети.
Здесь и на рис. 2: Зеленой рамкой выделены белки и гены, экспрессия которых повышена, белки и гены с пониженной экспрессией не отмечены рамкой. Большими шарами выделены белки, которые ранее не были аннотированы в Gene Ontology как участники апоптоза.

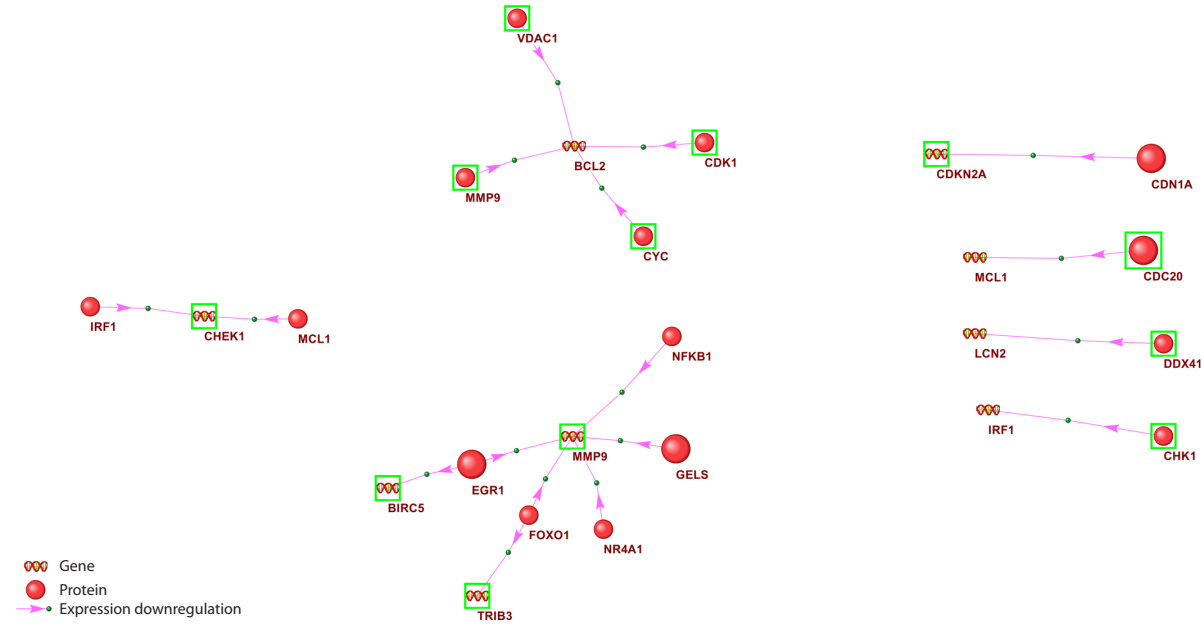


Рис. 2. Генная сеть подавления экспрессии генов-компонентов генной сети регуляции апоптоза при развитии ГЦК.

Представлены только белки-компоненты генной сети регуляции апоптоза гепатоцитов при развитии ГЦК (см. рис. S2), подавляющие (тип связи expression downregulation) экспрессию генов-компонентов той же сети.

регуляции экспрессии генов в итоговой генной сети апоптоза при развитии ГЦК. Для этого мы выполнили фильтрацию связей в реконструированной сети, оставив только те белки, которые усиливают (тип связей expression upregulation, рис. 1) или подавляют (тип связей expression downregulation, рис. 2) экспрессию генов-компонентов итоговой генной сети регуляции апоптоза.

Генная сеть активации экспрессии (см. рис. 1) включала в себя 38 белков, активирующих экспрессию 40 генов-компонентов генной сети апоптоза. По данным ANDSystem, белок NFKB1 активирует экспрессию 15 генов, включая *BCL2*, *MCL1*, *CFLAR* и другие; IL8 – 5 генов; ERBB2 – 4 гена; EGR1, SDF1 и TGFB2 – 3 гена; остальные белки генной сети активации экспрессии регулируют экспрес-

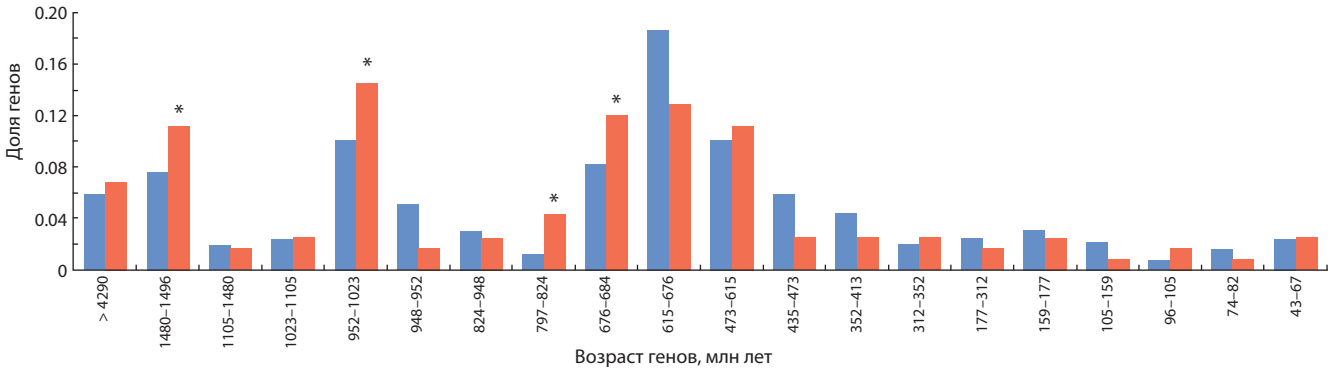


Рис. 3. Распределение эволюционного возраста генов реконструированной генной сети апоптоза гепатоцитов при развитии ГЦК.

По оси *X* показаны возрастные интервалы генов, млн лет, в соответствии с базой данных GenOrigin, по оси *Y* – доля генов в каждом возрастном интервале. Синими столбиками обозначено распределение полного набора генов, кодирующих белки человека, красными – распределение генов в реконструированной сети апоптоза гепатоцитов при ГЦК. * – достоверность различий представленности генов в данном возрастном интервале в наборе всех белок-кодирующих генов человека по сравнению с таковой в реконструированной генной сети.

сию менее трех апоптотических генов. При этом по результатам анализа scRNA-seq в нашем исследовании (см. табл. S1) как перечисленные выше белки, так и их гены-мишени характеризуются сниженной экспрессией в злокачественно трансформированных гепатоцитах по сравнению с гепатоцитами гистологически нормальной ткани печени. Экспрессия гена матриксной металлопротеиназы *MMP9* с повышенной экспрессией активируется, по данным ANDSystem, пятью белками: MEIN1, PPIA, TRIB3, CHK1, FEN1, экспрессия которых также увеличена в гепатоцитах опухоли. Кроме того, была повышена экспрессия белков CDC20, FEN1, KLF5 и их генов-мишеней.

Генная сеть подавления экспрессии (см. рис. 2) генов, участвующих в апоптозе при ГЦК, включала в себя 15 белков, связанных взаимодействиями типа expression downregulation с 9 генами. Экспрессия гена *MMP9*, по данным ANDSystem, может быть подавлена пятью белками: NFKB1, GELS, NR4A1, FOXO1, EGR1, экспрессия генов которых снижена в малигнизированных гепатоцитах, по данным scRNA-seq, что может служить причиной повышенной экспрессии этого гена, выявленной при анализе scRNA-seq. Экспрессия гена *BCL2*, который характеризуется сниженной экспрессией в злокачественно трансформированных гепатоцитах, может быть подавлена четырьмя белками: CDK1, VDAC1, MMP9, CYC, экспрессия которых повышена в малигнизированных гепатоцитах по сравнению с гепатоцитами здоровой ткани печени. Среди белков, вовлеченных в регуляцию апоптоза при ГЦК, но не аннотированных в Gene Ontology как участники этого процесса, в генную сеть подавления экспрессии вошли белки EGR1, CDN1A, GELS, CDC20.

Филостратиграфический анализ генной сети

Анализ эволюционного возрастного распределения генов реконструированной генной сети апоптоза при ГЦК приведен на рис. 3. Доля генов в реконструированной генной сети статистически значимо ($p < 0.05$, гипергеометрический тест) превышала таковую для всех белок-кодирующих генов человека в следующих возрастных интервалах: 1) 1480–1496 млн лет, 13 генов; 2) 952–1023 млн лет 17 генов; 3) 797–824 млн лет, 5 генов; 4) 676–684, 14 генов.

Обсуждение

Апоптоз представляет собой строго регулируемую и эволюционно консервативную программу клеточной гибели, выполняющую ключевые функции в нормальных физиологических процессах, таких как эмбриогенез и гомеостаз тканей во взрослом организме. Устойчивость к апоптозу – известная характеристика раковых клеток, обеспечивающая их выживание и рост опухоли (Kashyap et al., 2021). Однако в литературе имеются данные о том, что процессы апоптоза также могут быть активированы в опухолевых клетках, особенно на поздних стадиях развития новообразования. Таким образом, несмотря на то что уклонение от апоптоза является хорошо известным онкогенным механизмом (Moyer et al., 2025), популяции опухолевых клеток не способны постоянно подавлять программу апоптоза во всех клетках опухоли (обзор Morana et al., 2022). Это свидетельствует об определенных особенностях регуляции апоптоза при развитии злокачественных новообразований в зависимости от стадии формирования опухоли, типа ткани, из которой формируется опухоль, а также типа клеток, учитывая известную клеточную гетерогенность опухолей (Li C. et al., 2020). Поэтому необходимо детальное изучение молекулярно-генетических механизмов апоптоза при развитии различных видов злокачественных новообразований, в частности ГЦК, именно на уровне отдельных клеток.

В настоящем исследовании на основе общедоступных данных scRNA-seq был проведен сравнительный анализ транскриптома злокачественно трансформированных гепатоцитов и гепатоцитов гистологически нормальной ткани печени и реконструирована генная сеть регуляции апоптоза в гепатоцитах при развитии гепатоцеллюлярной карциномы у человека. Анализ данных scRNA-seq и регуляции экспрессии генов в реконструированной генной сети показал, что в малигнизированных гепатоцитах по сравнению со здоровыми клетками снижена экспрессия генов *NFKB1*, *BCL2* и *MCL1*, являющихся хабами генной сети (см. табл. 4). Известно, что белки BCL2 и MCL1 – ключевые ингибиторы апоптоза, поскольку они предотвращают активацию белков BAX/BAK, необходимых для увеличения мембранной проницаемости митохондрий и

последующей активации эффекторных каспаз (Newton et al., 2024). Считается, что повышение экспрессии гена *BCL2* – один из ведущих механизмов возникновения устойчивости клеток к апоптозу при их злокачественной трансформации (Moyer et al., 2025). Однако в нашем исследовании показано снижение экспрессии генов *BCL2* и *MCL1* в гепатоцитах при ГЦК, которое, по данным анализа генной сети регуляции апоптоза, может быть обусловлено как снижением экспрессии белков, активирующих экспрессию *BCL2* и *MCL1*, таких как NF-κB, SDF1, ERBB, IL8 (см. рис. 1), так и повышением экспрессии белков, подавляющих экспрессию *BCL2* (см. рис. 2).

Интересно отметить, что белок NFKB1 представляет собой главный хаб генной сети апоптоза гепатоцитов при развитии ГЦК (см. табл. 4) и важнейший белок генной сети активации экспрессии генов, вовлеченных в апоптоз гепатоцитов (см. рис. 2), который, по данным ANDSystem, способен активировать экспрессию целого ряда антиапоптотических генов, включая *BCL2* и *MCL1*. Известно, что в опухолях активация сигнального пути NF-κB способствует выживанию, ингибируя апоптоз (Gupta et al., 2023), поэтому снижение экспрессии гена *NFKB1*, выявленное в нашем исследовании (см. табл. S1 и 4), вероятно, может повышать готовность гепатоцитов к апоптозу. С другой стороны, показано, что активация белка NFKB1 необходима для протекания апоптоза клеток по внешнему пути, индуцированному хемокинами, в частности IL1b (Wang P. et al., 2023), и опосредованному рецептором TNFR1 (Moyer et al., 2025). Таким образом, снижение экспрессии гена *NFKB1* в злокачественно трансформированных гепатоцитах может, с одной стороны, способствовать апоптозу злокачественно трансформированных гепатоцитов за счет ослабления экспрессии ингибиторов апоптоза, но, с другой стороны, препятствовать индукции апоптоза по внешнему пути, который требует активации белков NF-κB. Кроме того, согласно данным анализа scRNA-seq, в нашем исследовании (см. табл. S1) в гепатоцитах опухоли увеличена экспрессия генов проапоптотических белков, таких как BID, активатор BAX/BAK (Moyer et al., 2025), и FADD (FAS-associated death domain protein), который является важным компонентом внешнего пути апоптоза (Nagata et al., 2017; Kashyap et al., 2021). Один из хабов генной сети регуляции апоптоза – циклин-зависимая киназа 1 (CDK1), экспрессия гена которой в малигнизированных гепатоцитах увеличена (см. табл. S1). В исследовании G. Massacci и ее сотрудников (2023) показано, что CDK1 фосфорилирует белки BCL2L1, BCL2 и MCL1, подавляя их антиапоптотические функции. Однако в этой же работе подчеркивается, что роль CDK1 в регуляции апоптоза может зависеть от различных экспериментальных условий и особенностей клеток. Таким образом, данные анализа scRNA-seq показывают снижение экспрессии ключевых антиапоптотических генов и повышение экспрессии важных проапоптотических генов в злокачественно трансформированных гепатоцитах по сравнению со здоровыми гепатоцитами. Наши результаты свидетельствуют о том, что в условиях ГЦК снижения уровня антиапоптотических белков недостаточно для запуска апоптоза. Это позволяет предположить, что уклонение злокачественных гепатоцитов от апоптоза путем повышения экспрессии ингибиторов апоптоза не является ведущим механизмом прогрессирования ГЦК, которое может быть вызвано иными причинами, связанными, вероятно, с микроокружением гепатоцитов, в частности, с нарушением регуляции процессов воспаления, что подтверждается исследованиями scRNA-seq (Lu et al., 2022; Jiang S. et al., 2024). Мы также считаем, что ключевой мишенью для терапии должна стать активация проапоптотических факторов, таких как каспаза.

Известно, что белки NF-κB относятся к важнейшим регуляторам воспаления и повышение их экспрессии стимулирует воспалительный ответ (Wang P. et al., 2023). Поэтому снижение экспрессии гена *NFKB1*, кодирующего один из членов этого семейства, NFKB1, согласуется с ослаблением в малигнизированных гепатоцитах экспрессии генов, вовлеченных в воспалительный ответ, о чем свидетельствуют результаты функциональной аннотации ДЭГ (см. табл. 1).

Поиск регуляторных связей ДЭГ, вовлеченных в контроль апоптоза гепатоцитов при ГЦК, с белками – продуктами других ДЭГ, выявленных при анализе scRNA-seq, позволил идентифицировать более 200 белков (см. табл. S6), которые потенциально могут изменять экспрессию генов, вовлеченных в контроль апоптоза гепатоцитов при развитии ГЦК, хотя и не аннотированы в Gene Ontology как регуляторы данного процесса. Интересно отметить, что по результатам функциональной аннотации генов, кодирующих эти белки, в опухолевых клетках была снижена экспрессия генов, продукты которых участвуют в обеспечении миграции и адгезии лейкоцитов. Это хемокины (*CCL5*, *CXCL2*, *CXCL8*, *CXCL1*), трансформирующий фактор роста бета 2 (*TGFB2*), тирозинкиназа (*SYK*), интегрин (*ITGA4*). Однако эти же белки, по данным ANDSystem, могут регулировать экспрессию важнейших узлов генной сети регуляции апоптоза гепатоцитов. В частности, CCL5 индуцирует экспрессию металлопротеиназы 9 (*MMP9*) (Sevenich, Joyce, 2014) – одного из основных хабов реконструированной генной сети регуляции апоптоза гепатоцитов при ГЦК. Металлопротеиназа 9 входит в многофункциональное семейство цинк-зависимых эндопептидаз и активируется во время воспаления и при некоторых видах рака. Металлопротеиназы расщепляют белки внеклеточного матрикса и играют важнейшую роль в клеточном апоптозе, ангиогенезе, росте опухоли и метастазировании (Verma et al., 2015). Известно, что MMP9 может индуцировать апоптоз (Liang et al., 2019). Полученные данные свидетельствуют о том, что снижение экспрессии генов, кодирующих важнейшие компоненты иммунной защиты, может способствовать прогрессированию опухоли не только за счет ослабления иммунного ответа на трансформированные клетки, но и за счет влияния на процессы апоптоза в них.

При этом предложенный ранее (Yatsyk et al., 2025) статистический подход для оценки значимости связи определенного белка или гена с генной сетью интереса, в данном случае апоптоза, и анализ реконструированной генной сети позволили приоритизировать ряд белков – потенциальных участников регуляции апоптотического процесса в гепатоцитах, изменение экспрессии которых с большой вероятностью может нарушать регуляцию апоптоза в ге-

патоцитах и тем самым вносить вклад в возникновение и прогрессирование ГЦК. Эти белки, ERBB2, CDN1A, IL8, EGR1, статистически значимо связаны с генной сетью регуляции апоптоза в гепатоцитах при ГЦК и являются центральными регуляторами (хабами) генной сети, влияющими на большое число (более 20) ее узлов.

Семейство гомологов вирусных онкогенов эритробластного лейкоза (ERBB), в которое входит рецептор эпидермального фактора роста (EGFR) и белки ERBB2, ERBB3 и ERBB4, регулирует широкий спектр таких важных клеточных функций, как выживание, рост и миграция клеток опухолей, и поэтому привлекает внимание в качестве мишени для лечения злокачественных новообразований (Chen et al., 2024). Член этого семейства – белок ERBB2, экспрессия которого была снижена в малигнизированных гепатоцитах, по данным scRNA-seq, ранее не был аннотирован как белок, участвующий в регуляции апоптоза, хотя вошел в число статистически значимых хабов реконструированной генной сети регуляции апоптоза (см. табл. 4). Однако было показано, что повышенная экспрессия *ERBB2* ассоциирована с ростом опухоли молочных желез, а подавление *ERBB2* и *ERBB3* приводит к индукции апоптоза в клетках опухоли молочных желез (Xiang et al., 2010). Хотя в литературе нет данных о роли ERBB2 в индукции апоптоза при ГЦК, анализ реконструированной генной сети показывает, что этот белок регулирует ряд белков и генов, вовлеченных в апоптоз при ГЦК, в том числе NFKB1, AKT2, CDK1, MCL1, FOXO1. В частности, показано, что ERBB2 фосфорилирует циклин-зависимую киназу CDK1 и увеличивает устойчивость раковых клеток к апоптозу, вызванному цитостатическим противораковым препаратом Таксол (Vahedi et al., 2015). Также установлено, что ERBB2 является активатором экспрессии антиапоптотических генов *NFKB1*, *AKT2*, *MCL1* (см. рис. 1), экспрессия которых, по данным scRNA-seq, снижена в малигнизированных гепатоцитах. Таким образом, ERBB2 – важное потенциальное звено регуляции апоптоза в гепатоцитах, а изменение его экспрессии может вносить вклад в развитие ГЦК.

IL8, также известный как CXCL8, – провоспалительный хемокин семейства CXC. Повышенный уровень IL8 ассоциирован с неблагоприятным прогнозом при различных опухолях, в том числе при гепатоцеллюлярной карциноме. При ГЦК повышение экспрессии *IL8* также связано с усилением метастатического потенциала опухолевых клеток (Han et al., 2023). Нокдаун *IL8* способствует индукции апоптоза в клетках ГЦК (Choi et al., 2016).

Белок CDN1A, также известный как CDKN1A, – ингибитор циклинзависимой киназы 1A, кодируемый геном *CDKN1A*, не был ранее аннотирован в Gene Ontology как белок, вовлеченный в апоптоз, однако его роль в апоптозе при развитии ГЦК обсуждается в литературе (Thangavelu et al., 2024). В статьях подчеркивается, что роль CDN1A в регуляции апоптоза при развитии новообразований противоречива, поскольку CDKN1A может как подавлять, так и активировать апоптоз (Manu et al., 2019). Согласно экспериментальным данным белок CDKN1A служит мишенью белка p53 и может стимулировать апоптоз в клетках опухоли, активируя рецептор TNF или проапоптотический белок BAX, или регулировать внутренний путь апоптоза, изменяя проницаемость митохондриальной мембраны (Abbas, Dutta, 2009). Природное соединение N-трансферулоилоктопамин может усиливать апоптоз опухолевых клеток при ГЦК посредством его взаимодействия с CDKN1A (Ma et al., 2021). Данные ANDSystem показывают, что этот белок – одно из центральных звеньев генной сети регуляции апоптоза в гепатоцитах при развитии ГЦК. Он взаимодействует с другими хабами сети, в частности, с такими известными регуляторами апоптоза, как NFKB1, BCL2, CDK1. Однако анализ данных scRNA-seq показал, что экспрессия гена *CDKN1A* была снижена в опухолевых гепатоцитах по сравнению нормальными клетками печени (см. табл. 4). Таким образом, полученные данные свидетельствуют о том, что ослабление экспрессии CDKN1A в гепатоцитах может быть важным звеном патогенеза ГЦК, способствуя уклонению опухолевых клеток от апоптоза, но его роль в регуляции апоптоза в гепатоцитах при ГЦК требует дальнейшего экспериментального исследования.

Белок раннего ответа на рост 1 (EGR1) подавляет пролиферацию злокачественно трансформированных клеток и усиливает их апоптоз в опухолях многих тканей и органов, включая печень (обзор Wang B. et al., 2021). Кроме того, установлено, что EGR1 способен подавлять рост ГЦК, подавляя транскрипцию *PFKL* (phosphofructokinase-1, liver type), и ингибирует аэробный гликолиз в клетках опухоли (Pan et al., 2024). В нашем исследовании установлено, что EGR1, экспрессия которого снижена, является активатором экспрессии генов (*LCN2*, *NR3C1*, *NR4A1* (см. рис. 1)), вовлеченных в контроль апоптоза, экспрессия которых также снижена в малигнизированных гепатоцитах. Результаты нашего исследования позволяют предположить, что снижение экспрессии *EGR1* может быть одним из механизмов ослабления апоптоза при злокачественном перерождении клеток.

Использование филостратиграфического анализа для оценки эволюционного возраста генов важно для исследования эволюции генных сетей и идентификации их ключевых компонентов (Mustafin et al., 2021). Следует отметить, что эволюционный возраст большинства генов генной сети апоптоза гепатоцитов и возраст генов перепредставленных возрастных интервалов более 600 млн лет (см. рис. 3), доля же относительно молодых генов невелика, что свидетельствует о консервативности генов генной сети и их важности для жизнедеятельности клетки. В частности, перепредставленная группа генов возрастом 1480–1496 млн лет соответствует времени формирования механизмов симбиоза митохондрий и эукариотической клетки (Raval et al., 2023). На этих этапах симбиоза сформировались многие гены, ответственные за митохондриальные пути программированной клеточной смерти, включая ключевые факторы, регулирующие высвобождение цитохрома C и управление окислительным стрессом, что отражало первоначальные адаптации для поддержания симбиотического баланса (Zmasek, Godzik, 2013). Кроме того, в нашем исследовании выявлено статистически значимое превышение доли генов в генной сети апоптоза гепатоцитов по сравнению с таковой во всем геноме человека в возрастном диапазоне 952–1023 млн лет. В этот возрастной диапазон, в частности, входят такие белки, как BCL2, являющийся хабом генной сети, и

BCL2L1. Известно, что эти белки – ключевые ингибиторы апоптоза (Moyer et al., 2025). Ортологи генов семейства BCL2 обнаружены уже у губок (*Porifera*), плакозой (*Placozoa*) и гидр (*Hydra*) (Banjara et al., 2020), т. е. на ранней стадии эволюции многоклеточных. Критическая роль апоптоза во врожденном и адаптивном иммунном ответе позволяет предположить, что эта функция возникла на ранних этапах эволюции многоклеточности. Вероятно, она предшествовала адаптации апоптоза для других процессов, таких как развитие, гомеостаз и удаление поврежденных клеток у *Metazoa*, что стало важной предпосылкой для возникновения сложных многоклеточных форм жизни (Suraweera et al., 2022). Таким образом, изменения в экспрессии генов гепатоцитов при развитии ГЦК затрагивают достаточно консервативные гены, включая хаб генной сети *BCL2*, которые помимо апоптоза могут регулировать и другие биологические процессы в клетке, обусловливая сложность регуляторных взаимодействий в клетке при злокачественной трансформации.

Таким образом, в нашем исследовании впервые на основе интегрированного подхода, включающего анализ транскриптома гепатоцитов, а также реконструкцию и анализ генной сети ДЭГ, вовлеченных в апоптоз, были получены данные об особенностях регуляции апоптоза гепатоцитов при развитии ГЦК у человека. Полученные результаты, показывающие снижение экспрессии ключевых генов – ингибиторов апоптоза, поддерживают представление о том, что уклонение от апоптоза не всегда характерно для раковых клеток и роль апоптоза в развитии злокачественных новообразований зависит от типа клеток, ткани и микроокружения опухолевых клеток (Morana et al., 2022). Кроме того, ослабление в злокачественно трансформированных гепатоцитах экспрессии генов, вовлеченных в контроль воспаления, а также белка NFKB1 – ключевого регулятора воспалительного процесса (Wang P. et al., 2023), позволяет предположить важную роль взаимодействия гепатоцитов с иммунной системой организма при развитии ГЦК, что требует дальнейшего экспериментального и теоретического исследования. Выявленные хабы генной сети (*NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*) могут быть полезны в качестве мишеней для модуляции апоптоза в гепатоцитах при лечении ГЦК, что в настоящее время является перспективным направлением (Ladd et al., 2024; Wu et al., 2024) терапии этого вида злокачественных новообразований.

Заключение

Анализ данных scRNA-seq нормальных и злокачественно трансформированных гепатоцитов показал изменения в экспрессии генов, вовлеченных в контроль апоптоза гепатоцитов при ГЦК, при этом в злокачественно трансформированных гепатоцитах была снижена экспрессия генов ключевых ингибиторов апоптоза, *BCL2* и *MCL1*, и уменьшена экспрессия генов, вовлеченных в воспалительный ответ. Полученные результаты свидетельствуют о том, что уклонение гепатоцитов от апоптоза путем повышения экспрессии ключевых ингибиторов апоптоза, по-видимому, не является характерной особенностью гепатоцитов при развитии ГЦК. Реконструкция и анализ генной сети регуляции апоптоза гепатоцитов при ГЦК показали, что уменьшение экспрессии гена *NFKB1* может быть важным фактором, обусловливающим уменьшение экспрессии целого ряда генов, вовлеченных в контроль апоптоза, включая *BCL2* и *MCL1*. Кроме того, реконструкция и анализ генной сети апоптоза гепатоцитов при ГЦК позволили выявить ряд ключевых генов: *NFKB1*, *MMP9*, *BCL2*, *A4*, *CDN1A*, *CDK1*, *ERBB2*, *G3P*, *MCL1*, *FOXO1*, которые одновременно демонстрируют дифференциальную экспрессию в малигнизированных гепатоцитах по сравнению со здоровыми гепатоцитами и служат хабами генной сети апоптоза гепатоцитов при развитии ГЦК. Нарушение экспрессии этих генов может приводить к дисрегуляции апоптоза в опухолевых клетках.

Среди ДЭГ также были идентифицированы гены (*CDKN1A*, *ERBB2*, *IL8*, *EGR1*), которые хотя и не были аннотированы в Gene Ontology как участники апоптоза, но число регуляторных взаимодействий их продуктов с генами, вовлеченными в апоптоз, статистически значимо превышало ожидаемое по случайным причинам согласно оценке на основе гипергеометрического распределения. Это позволяет считать, что белки, кодируемые этими генами, играют специфическую роль в регуляции апоптоза гепатоцитов при ГЦК и являются перспективными кандидатами для дальнейшего изучения.

Полученные результаты могут быть использованы при планировании дальнейших экспериментальных исследований по изучению особенностей регуляции апоптоза гепатоцитов при ГЦК. Выдвинутые гипотезы могут способствовать разработке таргетных терапевтических стратегий, направленных на модуляцию программируемой клеточной гибели в злокачественных клетках печени.

Список литературы / References

Наркевич А.Н., Виноградов К.А., Гржибовский А.М. Множественные сравнения в биомедицинских исследованиях: проблема и способы решения. *Экология человека*. 2020;10:55-64. doi 10.33396/1728-0869-2020-10-55-64
[Narkevich A.N., Vinogradov K.A., Grjibovski A.M. Multiple comparisons in biomedical research: the problem and its solutions. *Hum Ecol*. 2020;10:55-64. doi 10.33396/1728-0869-2020-10-55-64 (in Russian)]

Abbas T., Dutta A. p21 in cancer: intricate networks and multiple activities. *Nat Rev Cancer*. 2009;9(6):400-414. doi 10.1038/nrc2657

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Volianskaia A.R., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Kolchanov N.A., Lavrik I.N., Ivanisenko V.A. Reconstruction of the regulatory hypermethylation network controlling hepatocellular carcinoma development during hepatitis C viral infection. *J Integr Bioinform*. 2023;20(3): 20230013. doi 10.1515/jib-2023-0013

Banjara S., Suraweera C.D., Hinds M.G., Kvansakul M. The Bcl-2 family: Ancient origins, conserved structures, and divergent mechanisms. *Biomolecules*. 2020;10(1):128. doi 10.3390/biom10010128

Butikova E.A., Basov N.V., Rogachev A.D., Gaisler E.V., Ivanisenko V.A., Demenkov P.S., Makarova A.A., … Pokrovsky A.G., Vinokurov N.A., Kanygin V.V., Popik V.M., Shevchenko O.A. Metabolomic and gene networks approaches reveal the role of mitochondrial membrane proteins in response of human melanoma cells to THz radiation. *Biochim Biophys Acta Mol Cell Biol Lipids*. 2025; 1870(2):159595. doi 10.1016/j.bbalip.2025.159595

Chen Y., Lu A., Hu Z., Li J., Lu J. ERBB3 targeting: A promising approach to overcoming cancer therapeutic resistance. *Cancer Lett*. 2024;599:217146. doi 10.1016/j.canlet.2024.217146

Choi S.H., Park J.Y., Kang W., Kim S.U., Kim Y., Ahn S.H., Ro S.W., Han K.H. Knockdown of HIF-1α and IL-8 induced apoptosis of hepatocellular carcinoma triggers apoptosis of vascular endothelial cells. *Apoptosis*. 2016;21(1):85-95. doi 10.1007/s10495-015-1185-2

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2011;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Demenkov P.S., Antropova E.A., Adamovskaya A.V., Mishchenko E.L., Khlebodarova T.M., Ivanisenko T.V., Ivanisenko N.V., Venzel A.S., Lavrik I.N., Ivanisenko V.A. Prioritization of potential pharmacological targets for the development of anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway: the reconstruction and analysis of associative gene networks help. *Vavilov J Genet Breed*. 2023;27(7):784-793. doi 10.18699/VJGB-23-91

Dhanasekaran R., Bandoh S., Roberts L.R. Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances. *F1000Res*. 2016;5:879. doi 10.12688/f1000research.6946.1

Fabregat I. Dysregulation of apoptosis in hepatocellular carcinoma cells. *World J Gastroenterol*. 2009;15(5):513-520. doi 10.3748/wjg.15.513

Ganesan P., Kulik L.M. Hepatocellular carcinoma: New developments. *Clin Liver Dis*. 2023;27(1):85-102. doi 10.1016/j.cld.2022.08.004

Gupta R., Kadhim M.M., Turki Jalil A., Obayes A.M., Aminov Z., Alsaikhan F., Ramírez-Coronel A.A., Ramaiah P., Tayyib N.A., Luo X. Multifaceted role of NF-κB in hepatocellular carcinoma therapy: Molecular landscape, therapeutic compounds and nanomaterial approaches. *Environ Res.* 2023;228:115767. doi 10.1016/j.envres.2023.115767

Hajizadeh M., Hajizadeh F., Ghaffarei S., Amin Doustvandi M., Hajizadeh K., Yaghoubi S.M., Mohammadnejad F., Khiabani N.A., Mousavi P., Baradaran B. MicroRNAs and their vital role in apoptosis in hepatocellular carcinoma: miRNA-based diagnostic and treatment methods. *Gene*. 2023;888:147803. doi 10.1016/j.gene.2023.147803

Han X., Wu J., Sha Z., Lai R., Shi J., Mi L., Yin F., Guo Z. Dicer suppresses hepatocellular carcinoma via interleukin-8 pathway. *Clin Med Insights Oncol*. 2023;17:11795549231161212. doi 10.1177/11795549231161212

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved ai-based short names recognition. *Int J Mol Sci*. 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/S12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci Rep*. 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci*. 2024;25(21): 11847. doi 10.3390/ijms252111847

Jiang S., Lu H., Pan Y., Yang A., Aikemu A., Li H., Hao R., Huang Q., Qi X., Tao Z., Wu Y., Quan C., Zhou G., Lu Y. Characterization of the distinct immune microenvironments between hepatocellular carcinoma and intrahepatic cholangiocarcinoma. *Cancer Lett*. 2024;588:216799. doi 10.1016/j.canlet.2024.216799

Jiang Y., Han Q.J., Zhang J. Hepatocellular carcinoma: Mechanisms of progression and immunotherapy. *World J Gastroenterol*. 2019;25(25):3151-3167. doi 10.3748/wjg.v25.i25.3151

Kashyap D., Garg V.K., Goel N. Intrinsic and extrinsic pathways of apoptosis: Role in cancer development and prognosis. *Adv Protein Chem Struct Biol*. 2021;125:73-120. doi 10.1016/bs.apcsb.2021.01.003

Khlebodarova T.M., Demenkov P.S., Ivanisenko T.V., Antropova E.A., Lavrik I.N., Ivanisenko V.A. Primary and secondary micro-RNA modulation the extrinsic pathway of apoptosis in hepatocellular carcinoma. *Mol Biol*. 2023;57(2):165-175. doi 10.1134/S0026893323020103

Ladd A.D., Duarte S., Sahin I., Zarrinpar A. Mechanisms of drug resistance in HCC. *Hepatology*. 2024;79(4):926-940. doi 10.1097/HEP.0000000000000237

Lambert S.A., Jolma A., Campitelli L.F., Das P.K., Yin Y., Albu M., Chen X., Taipale J., Hughes T.R., Weirauch M.T. The human transcription factors. *Cell*. 2018;172(4):650-665. doi 10.1016/j.cell.2018.01.029

Larina I.M., Pastushkova L.Kh., Tiys E.S., Kireev K.S., Kononikhin A.S., Starodubtseva N.L., Popov I.A., Custaud M.A., Dobrokhotov I.V., Nikolaev E.N., Kolchanov N.A., Ivanisenko V.A. Permanent proteins in the urine of healthy humans during the Mars-500 experiment. *J Bioinform Comput Biol*. 2015;13(1):1540001. doi 10.1142/S0219720015400016

Li C., Xu J. Identification of potentially therapeutic target genes of hepatocellular carcinoma. *Int J Environ Res Public Health*. 2020;17(3):1053. doi 10.3390/ijerph17031053

Li X.Y., Shen Y., Zhang L., Guo X., Wu J. Understanding initiation and progression of hepatocellular carcinoma through single cell sequencing. *Biochim Biophys Acta Rev Cancer.* 2022;1877(3):188720. doi 10.1016/j.bbcan.2022.188720

Liang Y., Yang C., Lin Y., Parviz Y., Sun K., Wang W., Ren M., Yan L. Matrix metalloproteinase 9 induces keratinocyte apoptosis through FasL/Fas pathway in diabetic wound. *Apoptosis*. 2019;24(7-8):542-551. doi 10.1007/s10495-019-01536-w

Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi 10.1186/s13059-014-0550-8

Lu Y., Yang A., Quan C., Pan Y., Zhang H., Li Y., Gao C., Lu H., Wang X., Cao P., Chen H., Lu S., Zhou G. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. *Nat Commun*. 2022;13(1):4594. doi 10.1038/s41467-022-32283-3

Ma B., Li J., Yang W.K., Zhang M.G., Xie X.D., Bai Z.T. N-*trans*-feruloyloctopamine wakes up BBC3, DDIT3, CDKN1A, and NOXA signals to accelerate HCC cell apoptosis. *Anal Cell Pathol* (*Amst*). 2021;2021:1560307. doi 10.1155/2021/1560307

Manu K.A., Cao P.H.A., Chai T.F., Casey P.J., Wang M. p21cip1/waf1 coordinate autophagy, proliferation and apoptosis in response to metabolic stress. *Cancers* (*Basel*). 2019;11(8):1112. doi 10.3390/cancers11081112

Massacci G., Perfetto L., Sacco F. The Cyclin-dependent kinase 1: more than a cell cycle regulator. *Br J Cancer.* 2023;129(11):1707-1716. doi 10.1038/s41416-023-02468-8

Momynaliev K.T., Kashin S.V., Chelysheva V.V., Selezneva O.V., Demina I.A., Serebryakova M.V., Alexeev D., Ivanisenko V.A., Aman E., Govorun V.M. Functional divergence of *Helicobacter pylori* related to early gastric cancer. *J Proteome Res*. 2010;9(1): 254-267. doi 10.1021/pr900586w

Morana O., Wood W., Gregory C.D. The apoptosis paradox in cancer. *Int J Mol Sci*. 2022;23(3):1328. doi 10.3390/ijms23031328

Moyer A., Tanaka K., Cheng E.H. Apoptosis in cancer biology and therapy. *Annu Rev Pathol*. 2025;20(1):303-328. doi 10.1146/annurev-pathmechdis-051222-115023

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nagata S., Tanaka M. Programmed cell death and the immune system. *Nat Rev Immunol.* 2017;17(5):333-340. doi 10.1038/nri.2016.153

Newton K., Strasser A., Kayagaki N., Dixit V.M. Cell death. *Cell*. 2024; 187(2):235-256. doi 10.1016/j.cell.2023.11.044

Ogunwobi O.O., Harricharran T., Huaman J., Galuza A., Odumuwagun O., Tan Y., Ma G.X., Nguyen M.T. Mechanisms of hepatocellular carcinoma progression. *World J Gastroenterol*. 2019;25(19): 2279-2293. doi 10.3748/wjg.v25.i19.2279

Pan M., Luo M., Liu L., Chen Y., Cheng Z., Wang K., Huang L., Tang N., Qiu J., Huang A., Xia J. EGR1 suppresses HCC growth and aerobic glycolysis by transcriptionally downregulating PFKL. *J Exp Clin Cancer Res*. 2024;43(1):35. doi 10.1186/s13046-024-02957-5

Raval P.K., Martin W.F., Gould S.B. Mitochondrial evolution: Gene shuffling, endosymbiosis, and signaling. *Sci Adv*. 2023;9(32): p.eadj4493. doi 10.1126/sciadv.adj4493

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics*. 2019;12(Suppl. 2):47. doi 10.1186/s12920-019-0492-9

Sevenich L., Joyce J.A. Pericellular proteolysis in cancer. *Genes Dev*. 2014;28(21):2331-2347. doi 10.1101/gad.250647.114

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Si-Tayeb K., Noto F.K., Nagaoka M., Nagaoka M., Li J., Battle M.A., Duris C., North P.E., Dalton S., Duncan S.A. Highly efficient generation of human hepatocyte-like cells from induced pluripotent stem cells. *Hepatology*. 2010;51(1):297-305. doi 10.1002/hep.23354

Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*. 1967;62(318):626-633. doi 10.2307/2283989

Squair J.W., Gautier M., Kathe C., Anderson M.A., James N.D., Hutson T.H., Hudelle R., … Barraud Q., Levine A.J., La Manno G., Skinnider M.A., Courtine G. Confronting false discoveries in single-cell differential expression. *Nat Commun*. 2021;12(1):5692. doi 10.1038/s41467-021-25960-2

Suraweera C.D., Banjara S., Hinds M.G., Kvansakul M. Metazoans and intrinsic apoptosis: An evolutionary analysis of the Bcl-2 family. *Int J Mol Sci*. 2022;23(7):3691. doi 10.3390/ijms23073691

Thangavelu L., Altamimi A.S.A., Ghaboura N., Babu M.A., Roopashree R., Sharma P., Pal P., Choudhary C., Prasad G.V.S., Sinha A., Balaraman A.K., Rawat S. Targeting the p53-p21 axis in liver cancer: Linking cellular senescence to tumor suppression and progression. *Pathol Res Pract*. 2024;263:155652. doi 10.1016/j.prp.2024.155652

Tong Y.-B., Shi M.-W., Qian S.H., Chen Y.-J., Luo Z.-H., Tu Y.-X., Xiong Y.-L., Geng Y.-J., Chen C., Chen Z.-X. GenOrigin: a comprehensive protein-coding gene origination database on the evolutionary timescale of life. *J Genet Genomics*. 2021;48(12):1122-1129. doi 10.1016/j.jgg.2021.03.018

Traag V.A., Waltman L., van Eck N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*. 2019;9:5233. doi 10.1038/s41598-019-41695-z

Vahedi S., Chueh F.Y., Dutta S., Chandran B., Yu C.L. Nuclear lymphocyte-specific protein tyrosine kinase and its interaction with CR6-interacting factor 1 promote the survival of human leukemic T cells. *Oncol Rep*. 2015;34(1):43-50. doi 10.3892/or.2015.3990

Verma S., Kesh K., Gupta A., Swarnakar S. An overview of matrix metalloproteinase 9 polymorphism and gastric cancer risk. *Asian Pac J Cancer Prev*. 2015;16(17):7393-7400. doi 10.7314/apjcp.2015.16.17.7393

Wang B., Guo H., Yu H., Chen Y., Xu H., Zhao G. The role of the transcription factor EGR1 in cancer. *Front Oncol*. 2021;11:642547. doi 10.3389/fonc.2021.642547

Wang P., Qian H., Xiao M., Lv J. Role of signal transduction pathways in IL-1β-induced apoptosis: Pathological and therapeutic aspects. *Immun Inflamm Dis*. 2023;11(1):e762. doi 10.1002/iid3.762

Wolf F., Angerer P., Theis F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi 10.1186/s13059-017-1382-0

Wu X., Cao J., Wan X., Du S. Programmed cell death in hepatocellular carcinoma: mechanisms and therapeutic prospects. *Cell Death Discov.* 2024;10(1):356. doi 10.1038/s41420-024-02116-x

Xiang S., Sun Z., He Q., Yan F., Wang Y., Zhang J. Aspirin inhibits ErbB2 to induce apoptosis in cervical cancer cells. *Med Oncol*. 2010;27(2):379-387. doi 10.1007/s12032-009-9221-0

Xue R., Zhang Q., Cao Q., Kong R., Xiang X., Liu H., Feng M., … Zhan Q., Deng M., Zhu J., Zhang Z., Zhang N. Liver tumour immune microenvironment subtypes and neutrophil heterogeneity. *Nature*. 2022;612(7938):141-147. doi 10.1038/s41586-022-05400-x

Yang Y.M., Kim S.Y., Seki E. Inflammation and liver cancer: molecular mechanisms and therapeutic targets. *Semin Liver Dis*. 2019;39(1): 26-42. doi 10.1055/s-0038-1676806

Yatsyk I.V., Volyanskaya A.R, Kleshchev M.A., Antropova E.A., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Reconstruction and computational analysis of the insulin response gene network in human type 2 diabetes using transcriptomic data. *Gene Expr*. 2025 (в печати)

Zhang Q.Y., Ho D.W., Tsui Y.M., Ng I.O. Single-cell transcriptomics of liver cancer: hype or insights? *Cell Mol Gastroenterol Hepatol*. 2022;14(3):513-525. doi 10.1016/j.jcmgh.2022.04.014

Zmasek C.M., Godzik A. Evolution of the animal apoptosis network. *Cold Spring Harb Perspect Biol*. 2013;5(3):a008649. doi 10.1101/cshperspect.a008649

Zou Y., Wan X., Zhou Q., Zhu G., Lin S., Tang Q., Yang X., Wang S. Mechanisms of drug resistance in hepatocellular carcinoma. *Biol Proced Online*. 2025;27(1):19. doi 10.1186/s12575-025-00281-6

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 16.07.2025. После доработки 13.09.2025. Принята к публикации 15.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):978-989

doi 10.18699/vjgb-25-103

Сигнальный путь Hedgehog у человека: описание в базе знаний HH_Signal_pathway_db

Т.А. Бухарина [1, 2], А.М. Бондаренко[2], Д.П. Фурман[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
bukharina@bionet.nsc.ru; furman@bionet.nsc.ru

Аннотация. Стремительное развитие омиксных технологий (геномики, транскриптомики, протеомики, метаболомики) и других высокопроизводительных методов экспериментального исследования молекулярно-генетических систем и процессов привело к генерации беспрецедентно огромных объемов разнородных и сложных биологических данных. Эффективное использование этого информационного ресурса требует системных подходов к их анализу. Один из подходов состоит в создании предметно-ориентированных баз знаний/данных – репозиториев, интегрирующих информацию из множества источников, что позволяет не только хранить и структурировать распределенные по различным источникам гетерогенные данные, но и получать новые сведения о биологических системах и процессах. Критически важен системный подход и к решению фундаментальной задачи биологии – выяснению закономерностей морфогенеза. Регуляция морфогенеза осуществляется через эволюционно консервативные сигнальные пути (Hedgehog, Wnt, Notch и др.). Ключевая роль в этом процессе принадлежит пути Hedgehog (HH), поскольку в онтогенезе он начинает функционировать ранее других и детерминирует реализацию каждого этапа индивидуального развития организма: от структурирования эмбриональных зачатков, гисто- и органогенеза до поддержания тканевого гомеостаза и процесса регенерации у взрослых особей. Нами создана база знаний HH_Signal_pathway_db, в которую сведена информация о компонентах и функциях HH сигнального пути у человека. Первый релиз базы (доступен по запросу bukharina@bionet.nsc.ru) содержит информацию о входящих в него 56 генах, их белковых продуктах, сети регуляторных взаимодействий, а также об установленных связях с некоторыми патологическими состояниями человека. HH_Signal_pathway_db предоставляет исследователям инструмент для получения новых знаний о роли пути Hedgehog в норме и при патологии и возможностях применения их в области биологии развития и трансляционной медицины.

Ключевые слова: база знаний; сигнальный путь Hedgehog; морфогенез; эволюция; генные сети; регуляторные контуры

Для цитирования: Бухарина Т.А., Бондаренко А.М., Фурман Д.П. Сигнальный путь Hedgehog у человека: описание в базе знаний HH_Signal_pathway_db. *Вавиловский журнал генетики и селекции*. 2025;29(7):978-989. doi 10.18699/vjgb-25-103

Финансирование. Работа поддержана бюджетным проектом FWNR-2022-0020.

Благодарности. Авторы выражают искреннюю благодарность акад. Н.А. Колчанову за интерес к работе и плодотворную дискуссию, Р.А. Иванову и д. б. н. С.А. Лашину за помощь в определении эволюционных характеристик генов Hedgehog сигнального пути; Н.Л. Подколодному за предоставление данных по аффинности белка TBP к промоторам генов, к. б. н. В.А. Иванисенко и его сотрудникам И.В. Яцыку и А.В. Адамовской за помощь с освоением программы построения генных сетей.

Hedgehog signaling in humans: the HH_Signal_pathway_db knowledge base

T.A. Bukharina [1, 2], A.M. Bondarenko[2], D.P. Furman[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
bukharina@bionet.nsc.ru; furman@bionet.nsc.ru

Abstract. The rapid advancement of omics technologies (genomics, transcriptomics, proteomics, metabolomics) and other high-throughput methods for experimental studies of molecular genetic systems and processes has led to the generation of an unprecedentedly vast amount of heterogeneous and complex biological data. Effective use of this information resource requires systematic approaches to its analysis. One such approach involves the creation of domain-specific knowledge/data repositories that integrate information from multiple sources. This not only enables the storage and structuring of heterogeneous data distributed across various resources but also facilitates the acquisition of new insights into biological systems and processes. A systematic approach is also critical to solving the fundamental

© Бухарина Т.А., Бондаренко А.М., Фурман Д.П., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

problem of biology – clarifying the regularities of morphogenesis. Morphogenesis is regulated through evolutionarily conserved signaling pathways (Hedgehog, Wnt, Notch, etc.). The Hedgehog (HH) pathway plays a key role in this process, as it begins functioning earlier than others in ontogenesis and determines the progression of every stage of an organism's life cycle: from structuring embryonic primordia, histo- and organogenesis, to maintaining tissue homeostasis and regeneration in adults. Our work presents HH_Signal_pathway_db, a knowledge base that integrates curated data on the molecular components and functional roles of the human Hedgehog (HH) signaling pathway. The first release of the database (available upon request at bukharina@bionet.nsc.ru) contains information on 56 genes, their protein products, the regulatory interaction network, and established associations with pathological conditions in humans. HH_Signal_pathway_db provides researchers with a tool for gaining new knowledge about the role of the Hedgehog pathway in health and disease, and its potential applications in developmental biology and translational medicine.
Key words: knowledge base; Hedgehog signaling pathway; morphogenesis; evolution; gene networks; regulatory circuits

For citation: Bukharina T.A., Bondarenko A.M., Furman D.P. Hedgehog signaling in humans: description in the HH_Signal_pathway_db knowledge base. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7): 978-989. doi 10.18699/vjgb-25-103

Введение

Молекулярно-генетические и медико-биологические исследования с использованием современных методов и технологий генерируют огромные объемы разнородной информации (Regev et al., 2017; Schermelleh et al., 2019; Kenneth, 2022), включающей также данные, полученные при изучении различных аспектов такого фундаментального процесса, как морфогенез, результатом которого является создание сложной архитектоники организма. Понимание механизмов морфогенеза необходимо не только для ответа на один из самых глубоких вопросов биологии – как из единственной клетки возникает сложнейший пространственно организованный многоклеточный организм, но и для объяснения механизмов регенерации тканей, причин возникновения врожденных аномалий, патологических состояний различной этиологии, в том числе и онкологических заболеваний.

В регуляцию морфогенеза вовлечено множество генов, белков, миРНК, сигнальных молекул (ENCODE Project Consortium, 2012; Briscoe, Thérond, 2013; Bartel, 2018; Ghafouri-Fard et al., 2022; McIntyre et al., 2024), часть из которых входит в сигнальные пути.

Сигнальный путь (сигнальная трансдукция) – это способ передачи сигнала, поступающего на внешнюю мембрану клетки, внутрь ядра. Сигнальный путь опосредуется каскадом межмолекулярных взаимодействий с участием лиганда, рецептора, распознающего лиганд, и внутриклеточных передатчиков сигнала как белковой, так и небелковой природы, факторов и корегуляторов транскрипции. Результат действия сигнального пути состоит в изменении экспрессии генов-мишеней и соответствующих белков, что ведет к изменению функционального состояния клетки.

Сигнальные пути у животных и человека эволюционно консервативны, и их роли сходны у представителей разных таксономических групп. Пути образуют сложные сети с перекрестными взаимодействиями; для формирования полноценного организма необходима строгая координация функционирования сигнальных путей. Сигнальные пути критически значимы для нормального онтогенеза; мутации входящих в них генов и изменения в экспрессии соответствующих белков могут приводить к катастрофическим нарушениям развития (Artavanis-Tsakonas et al., 1999; Ingham, McMahon, 2001; Logan, Nusse, 2004; Rubin, 2007; Perrimon et al., 2012; Briscoe, Thérond, 2013; Huttlin et al., 2017).

Существенную роль в контроле морфогенеза играет сигнальный путь Hedgehog (HH), обязанный своим названием открытию в начале 1980-х годов у *Drosophila melanogaster* гена *hedgehog* (*hh*). Личинки мух, мутантных по этому гену, покрыты шипиками, что придает им внешний вид, напоминающий ежа (hedgehog) (Nüsslein-Volhard, Wieschaus,1980).

Hedgehog путь передачи сигнала – не просто один из путей, детерминирующих ход развития организмов, а центральный регулятор морфогенеза, определяющий передне-заднюю и дорсо-вентральную оси тела и сегментацию зачатков у животных на стадии эмбриогенеза, гисто- и органогенез, поддержание пула стволовых клеток в тканях взрослого организма и др. С дисфункцией этого сигнального пути связаны многочисленные врожденные аномалии и заболевания человека, в том числе рак различных органов (Ingham, McMahon, 2001; Spinella-Jaegle et al., 2001; Varjosalo, Taipale, 2007; Briscoe, Thérond, 2013; Wu et al., 2017; Skoda et al., 2018; Jamieson et al., 2020; Fitzsimons et al., 2022; Ingham, 2022; Dutta et al., 2023; Jing et al., 2023). Именно этим обстоятельством продиктован неослабевающий интерес к всестороннему изучению молекулярно-генетической организации и механизмов функционирования HH пути. Общая схема сигнального пути Hedgehog представлена на рис. 1.

Для проведения HH сигнала необходимо наличие в клетке-реципиенте определенного набора основных белков – участников процесса, находящихся в определенных функциональных состояниях. К этим белкам относятся трансмембранные рецепторы Patched1 и Patched2 (PTCH1/2), неактивная форма трансмембранного белка Smoothened (SMO), комплексы транскрипционных факторов GLI1/3 с белком-скаффолдом Suppressor of fused homolog (SUFU) и активная протеинкиназа A (PKA), от которой зависит появление репрессорной формы транскрипционного фактора GLI3 (GLI3R).

При неработающем сигнальном пути в отсутствие лигандов HH (см. рис. 1, *а*) рецепторы PTCH1/2 локализованы на первичной ресничке – специализированной внешней органелле клетки, выполняющей роль сенсора внешних сигналов (Ingham, McMahon, 2001; Eggenschwiler, Anderson, 2007; Oro, 2007; Carballo et al., 2018).

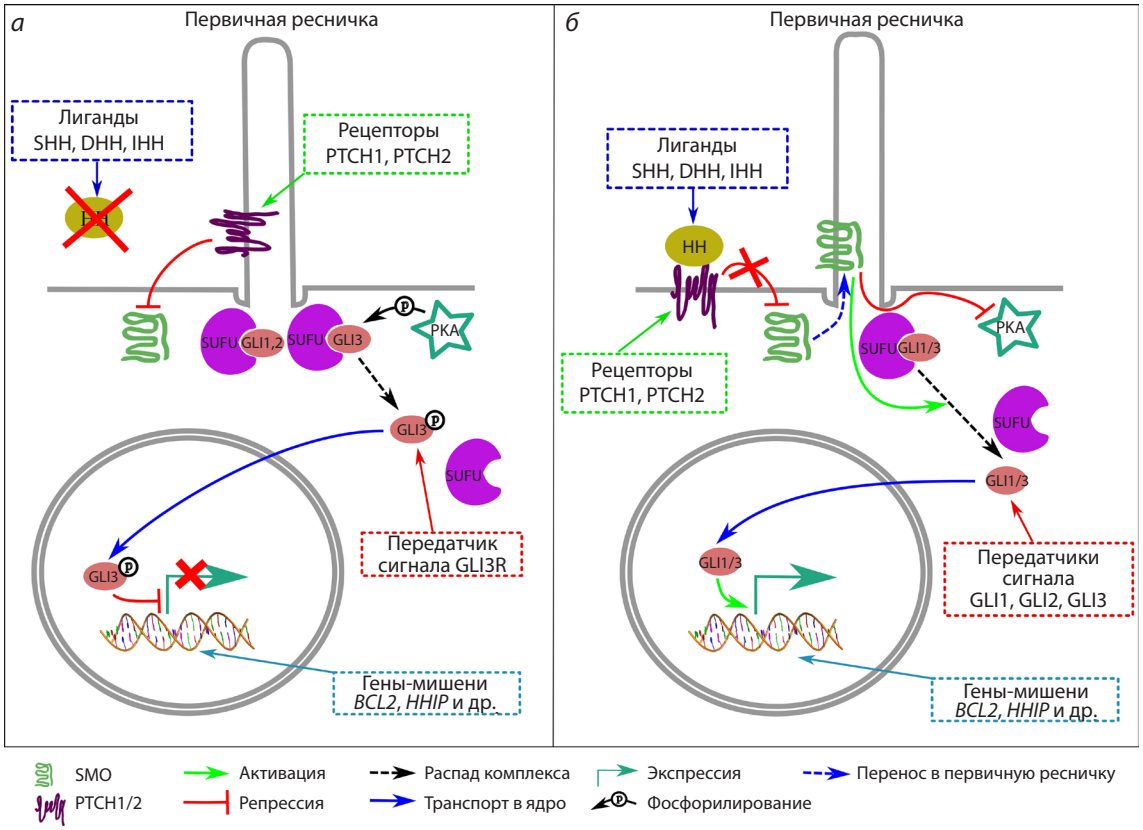


Рис. 1. Общая схема Hedgehog сигнального пути человека.

а – механизм действия сигнального пути в отсутствие лигандов HH; *б* – механизм действия сигнального пути при связывании рецепторов PTCH1/2 с лигандами HH (пояснения см. в тексте).

Рецепторы PTCH1/2 блокируют миграцию белка SMO, находящегося во внутриклеточном пространстве, на мембрану реснички, и SMO не может провзаимодействовать с протеинкиназой А (PKA), чтобы ингибировать ее активность. В результате PKA фосфорилирует комплекс GLI3/SUFU, этот комплекс диссоциирует, и GLI3 подвергается протеолитическому расщеплению с образованием репрессора GLI3R, который затем проникает в ядро и подавляет транскрипцию своих целевых генов, в том числе и некоторых генов HH пути (Gorojankina, 2016; Dilower et al., 2023).

Активация передачи сигнала происходит при взаимодействии с PTCH1/2 внеклеточных лигандов – белков Hedgehog (HH) (у человека их три: Sonic Hedgehog (SHH), Indian Hedgehog (IHH) и Desert Hedgehog (DHH)). Затем комплекс лиганд/рецептор удаляется с мембраны реснички и перемещается во внутриклеточное пространство, где подвергается деградации в лизосоме. Позицию PTCH1/2 занимает SMO, который подавляет активность протеинкиназы А, запрещая фосфорилирование комплекса SUFU/GLI3 и образование GLI3R. Далее в ресничке происходят деградация комплексов SUFU/GLI1/3 и образование активных форм GLI1/3, которые проникают в ядро и активируют транскрипцию генов-мишеней, обеспечивая передачу сигнала (Ingham, McMahon, 2001; Varjosalo, Taipale, 2007; Briscoe, Thérond, 2013; Gorojankina, 2016) (см. рис. 1, *б*).

Существует два варианта HH пути – канонический, представленный на рис. 1, и неканонический, в котором активация транскрипционных факторов GLI1/3 происходит без участия SMO, что меняет путь передачи сигнала (Brennan et al., 2012; Briscoe, Thérond, 2013; Carballo et al., 2018).

В настоящее время информация, касающаяся реализации HH пути у человека, рассеяна по огромному числу источников (на момент написания статьи только в PubMed по запросу “Hedgehog signaling” обнаруживается 15247 публикаций https://pubmed.ncbi.nlm.nih.gov/?term=hedgehog+signaling), и ее объемы продолжают нарастать. Несмотря на экстенсивный рост числа исследований в этой области, полного всестороннего представления об эволюции, структуре и механизмах функционирования HH пути все еще не достигнуто (Ingham et al., 2011; Briscoe, Thérond, 2013; Breeze, 2022).

Для интеграции, структурирования и анализа существующих данных нами создается специализированная база знаний (БЗ) HH_Signal_pathways_db. База наполняется разноплановой информацией, касающейся всех аспектов организации и функционирования HH пути, что дает возможность системного подхода к его изучению.

Биоинформатический анализ структурно-функциональной организации HH пути открывает перспективы для детализации молекулярно-генетических основ морфогенеза, механизмов регенерации органов и тканей, процесса старения, возникновения патологий различной этиологии, а также разработки методов для их диагностики и фармакотерапии.

В рамках данной работы получены новые результаты, касающиеся закономерностей реализации сигнала, – реконструирована ассоциативная генная сеть HH сигнального пути, выявлены регуляторные контуры в ее составе и получены сведения об эволюции входящих в сигнальный путь генов.

Материалы и методы

База знаний HH_Signal_pathway_db и ее наполнение. Блок-схема разработанного формата базы представлена на рис. 2.

Список генов, входящих в Hedgehog сигнальный путь человека, получен из базы данных KEGG (https://www.genome.jp/kegg/) по запросу (Environmental Information Processing→Signal transduction→Hedgehog signaling pathway) (табл. 1).

Для наполнения блоков «информация о гене» и «информация о продукте гена» использовались сведения из баз данных NCBI Gene Base (https://www.ncbi.nlm.nih.gov/gene), UniProt (https://www.uniprot.org), TRRUST database (https://www.grnpedia.org/trrust/).

Информация для блока «аффинность TPB к промотору» (TBP (TATA-связывающий белок) – ключевой регулятор инициации транскрипции генов эукариот), взята из базы данных Human_SNP_TATAdb (Филонов и др., 2023).

Для наполнения блока «эволюционные характеристики» использовали программный комплекс Orthoweb (Ivanov et al., 2024) – специализированный инструмент для расчета двух эволюционных индексов: «филостратиграфического возраста гена» (phylostratigraphic age index, PAI) и «индекса дивергенции» (divergence index, DI) (Mustafin et al., 2021; Ivanov et al., 2024).

Индекс PAI отражает степень удаленности таксона от корня филогенетического дерева и рассчитывается как расстояние от корня до узла, в котором произошло расхождение исследуемого вида с наиболее удаленным родственным таксоном: чем выше PAI, тем «моложе» рассматриваемый ген. Для генов человека значения PAI лежат в диапазоне от 0 (Cellular Organisms-клеточные организмы, корень дерева) до 15 (*Homo sapiens* – человек разумный).

Индекс эволюционной изменчивости гена (DI) оценивает соотношение между несинонимичными заменами, изменяющими кодируемую аминокислоту в последовательностях анализируемого гена и его ортолога (dN), и синонимичными заменами, не изменяющими кодируемую аминокислоту (dS) в нуклеотидных последовательностях генов и их ортологов:

$$DI = \frac{\sum_{i=1}^{n} dnds_i}{n},$$

где $dnds_i$ – значения dN/dS для гена и i-го ортолога, n – количество ортологичных генов.

Индекс DI позволяет установить тип отбора, под действием которого находится данный ген. Значения $DI < 1$ и $DI > 1$ принято интерпретировать как свидетельства стабилизирующего и движущего отбора соответственно, а $DI = 1$ – как свидетельство нейтральной эволюции (Jeffares et al., 2015; Spielman, Wilke, 2015).

Для построения ассоциативной генной сети и поиска регуляторных контуров (генных сетей меньшей размерности) использовали программно-информационную когнитивную платформу ANDSystem, которая с применением методов искусственного интеллекта автоматически извлекает знания из текстов научных публикаций и фактографических баз данных и через модуль ANDVisio осуществляет визуализацию результатов в виде графа генной сети (Demenkov et al., 2011; Ivanisenko et al., 2015, 2019; Иванисенко и др., 2022).

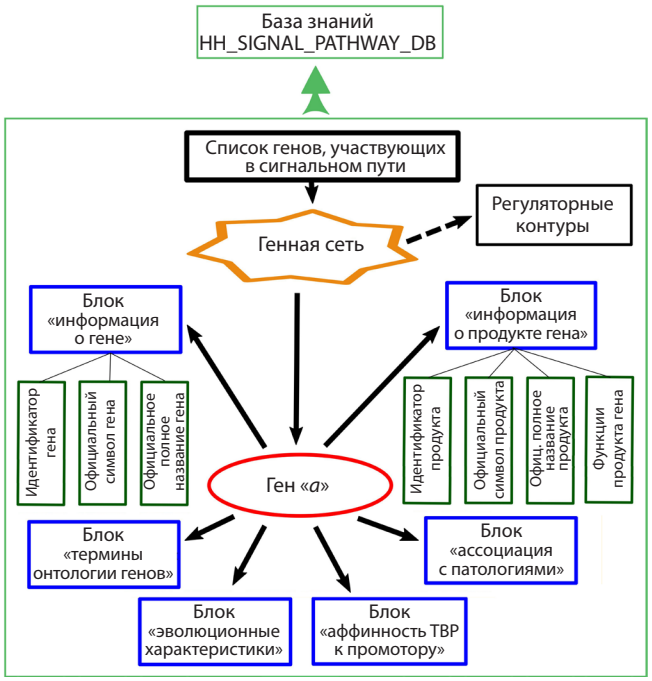


Рис. 2. Блок-схема базы знаний HH_Signal_pathway_db.

Генная сеть реконструирована для 56 генов сигнального пути Hedgehog. В ней отражены связи с белками, которые кодируются этими генами («экспрессия»), с транскрипционными факторами, регулирующими экспрессию генов («регуляция экспрессии»), с белками, регулирующими транспорт белков («регуляция транспорта»), с миРНК, осуществляющими посттранскрипционную регуляцию экспрессии белков («регуляция с помощью миРНК»).

Функциональную аннотацию генов проводили с использованием веб-ресурса DAVID (https://davidbioinformatics.nih.gov/) (Sherman et al., 2022), позволяющего определить, какие биологические процессы статистически чаще ассоциированы с анализируемым набором генов. Рассматривались только процессы, имеющие с учетом поправки на множественное сравнение по Бенджамини–Хохбергу индекс FDR (false discovery rate), отражающий ожидаемую долю ложноположительных результатов, меньше 0.05.

Результаты и обсуждение

База знаний HH_Signal_pathway_db

В текущей версии базы знаний представлена информация о 56 генах человека, вовлеченных в функционирование этого сигнального пути (см. табл. 1). Первый релиз базы содержит следующие блоки: 1) список генов HH сигнального пути со ссылками на литературные источники из

Таблица 1. Гены Hedgehog сигнального пути (по данным базы данных KEGG)

№ п/п	Символ гена	Идентификатор гена	Полное название гена
1	*ARRB1*	408	arrestin beta 1
2	*ARRB2*	409	arrestin beta 2
3	*BCL2*	596	BCL2 apoptosis regulator
4	*BOC*	91653	BOC cell adhesion associated, oncogene regulated
5	*BTRC*	8945	beta-transducin repeat containing E3 ubiquitin protein ligase
6	*CCND1*	595	cyclin D1
7	*CCND2*	894	cyclin D2
8	*CDON*	50937	cell adhesion associated, oncogene regulated
9	*CSNK1A1*	1452	casein kinase 1 alpha 1
10	*CSNK1A1L*	122011	casein kinase 1 alpha 1 like
11	*CSNK1D*	1453	casein kinase 1 delta
12	*CSNK1E*	1454	casein kinase 1 epsilon
13	*CSNK1G1*	53944	casein kinase 1 gamma 1
14	*CSNK1G2*	1455	casein kinase 1 gamma 2
15	*CSNK1G3*	1456	casein kinase 1 gamma 3
16	*CUL1*	8454	cullin 1
17	*CUL3*	8452	cullin 3
18	*DHH*	50846	desert hedgehog signaling molecule
19	*DISP1*	84976	dispatched RND transporter family member 1
20	*EFCAB7*	84455	EF-hand calcium binding domain 7
21	*EVC*	2121	EvC ciliary complex subunit 1
22	*EVC2*	132884	EvC ciliary complex subunit 2
23	*FBXW11*	23291	F-box and WD repeat domain containing 11
24	*GAS1*	2619	growth arrest specific 1
25	*GLI1*	2735	GLI family zinc finger 1
26	*GLI2*	2736	GLI family zinc finger 2
27	*GLI3*	2737	GLI family zinc finger 3
28	*GPR161*	23432	G protein-coupled receptor 161
29	*GRK2*	156	G protein-coupled receptor kinase 2
30	*GRK3*	157	G protein-coupled receptor kinase 3
31	*GSK3B*	2932	glycogen synthase kinase 3 beta
32	*HHAT*	55733	hedgehog acyltransferase
33	*HHATL*	57467	hedgehog acyltransferase like
34	*HHIP*	64399	hedgehog interacting protein
35	*IHH*	3549	Indian hedgehog signaling molecule
36	*IQCE*	23288	IQ motif containing E
37	*KIF3A*	11127	kinesin family member 3A
38	*KIF7*	374654	kinesin family member 7
39	*LRP2*	4036	LDL receptor related protein 2
40	*MEGF8*	1954	multiple EGF like domains 8
41	*MGRN1*	23295	mahogunin ring finger 1
42	*MOSMO*	730094	modulator of smoothened
43	*PRKACA*	5566	protein kinase cAMP-activated catalytic subunit alpha
44	*PRKACB*	5567	protein kinase cAMP-activated catalytic subunit beta
45	*PRKACG*	5568	protein kinase cAMP-activated catalytic subunit gamma
46	*PTCH1*	5727	patched 1
47	*PTCH2*	8643	patched 2
48	*SCUBE2*	57758	signal peptide, CUB domain and EGF like domain containing 2
49	*SHH*	6469	sonic hedgehog signaling molecule
50	*SMO*	6608	smoothened, frizzled class receptor
51	*SMURF1*	57154	SMAD specific E3 ubiquitin protein ligase 1
52	*SMURF2*	64750	SMAD specific E3 ubiquitin protein ligase 2
53	*SPOP*	8405	speckle type BTB/POZ protein
54	*SPOPL*	339745	speckle type BTB/POZ protein like
55	*SUFU*	51684	SUFU negative regulator of hedgehog signaling
56	*TPTEP2-CSNK1E*	102800317	TPTEP2-CSNK1E readthrough

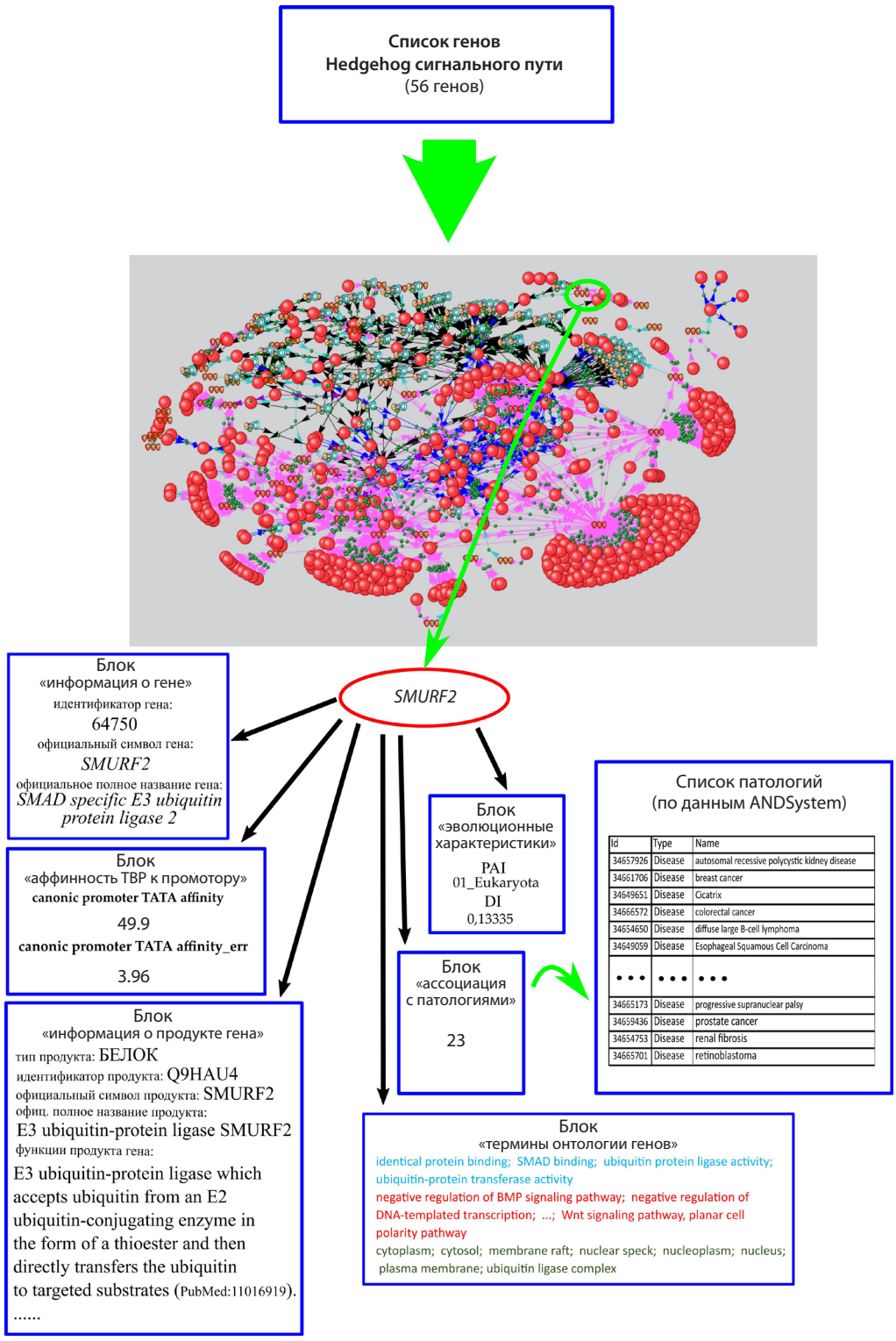


Рис. 3. Пример заполнения блока базы знаний HH_Signal_pathway_db для гена *SMURF2*.

PubMed Base; 2) списки белков, кодируемых генами HH сигнального пути, и их функции; 3) термины генной онтологии; 4) значения индексов эволюционных возрастов генов (PAI); 5) значения индексов эволюционной изменчивости генов (DI); 6) данные по аффинности белка TBP к промоторам генов, определяющей интенсивность их транскрипции; 7) списки патологий, ассоциированных с каждым из генов; 8) реконструированная ассоциативная генная сеть и выявленные в ней регуляторные контуры. Образец заполнения блоков базы знаний для конкретного гена на примере гена *SMURF2* приведен на рис. 3.

Ниже приводятся некоторые результаты биоинформатического анализа информации, представленной в HH_Signal_pathway_db.

Функциональная аннотация генов HH сигнального пути

Анализ терминов биологических процессов в Gene Ontology (GO) для 56 генов, выполненный с привлечением ресурса DAVID, выявил 221 биологический процесс, статистически значимо ассоциированный с сигнальным путем. Список процессов условно можно разделить на три основные группы: морфогенез (94), внутриклеточные процессы (60) и межклеточная коммуникация (67). Для всех процессов индекс значимости (FDR) менее 0.05.

Морфогенез

- GO:0042733~embryonic digit morphogenesis
- GO:0042475~odontogenesis of dentin-containing tooth
- GO:0007507~heart development
- GO:0001658~branching involved in ureteric bud morphogenesis
- GO:0003151~outflow tract morphogenesis
- GO:0030324~lung development
- GO:0003180~aortic valve morphogenesis
- GO:0045766~positive regulation of angiogenesis
- GO:0001501~skeletal system development
- GO:0001942~hair follicle development
- GO:0021983~pituitary gland development
- GO:0001822~kidney development
- GO:0001525~angiogenesis
- GO:0042060~wound healing
- GO:0001889~liver development
- GO:0072091~regulation of stem cell proliferation

и др.

Внутриклеточные процессы

Регуляция транскрипции

- GO:1902895~positive regulation of miRNA transcription
- GO:1902894~negative regulation of miRNA transcription
- GO:0006357~regulation of transcription by RNA polymerase II
- GO:0006338~chromatin remodeling
- GO:0006355~regulation of DNA-templated transcription
- GO:0010468~regulation of gene expression

Ответ на стрессовые воздействия

- GO:0071456~cellular response to hypoxia
- GO:0034599~cellular response to oxidative stress
- GO:0071466~cellular response to xenobiotic stimulus
- GO:0034644~cellular response to UV
- GO:0006974~DNA damage response

Регуляция циклических процессов

- GO:0048511~rhythmic process
- GO:0051726~regulation of cell cycle

Апоптоз

- GO:0043066~negative regulation of apoptotic process
- GO:0043065~positive regulation of apoptotic process

Межклеточная коммуникация

- GO:0042127~regulation of cell population proliferation
- GO:0050673~epithelial cell proliferation
- GO:0010595~positive regulation of endothelial cell migration
- GO:0001938~positive regulation of endothelial cell proliferation
- GO:0042127~regulation of cell population proliferation
- GO:0072089~stem cell proliferation

и др.

Участие в путях передачи сигналов

- GO:0038084~vascular endothelial growth factor signaling pathway
- GO:0007173~epidermal growth factor receptor signaling pathway
- GO:0008543~fibroblast growth factor receptor signaling pathway
- GO:0007224~smoothened signaling pathway
- GO:0060070~canonical Wnt signaling pathway
- GO:0030509~BMP signaling pathway
- GO:0000165~MAPK cascade
- GO:0007219~Notch signaling pathway
- GO:0070371~ERK1 and ERK2 cascade

и др.

Существенная роль Hedhehog сигнального пути состоит в его участии в морфогенетических процессах эмбрио-, гисто- и органогенеза. Гены HH пути вовлечены в закладку нервной системы, формирование хрящевой и скелетной тканей, ангиогенез, развитие почек, печени, легких, сердца, эндокринной части поджелудочной железы, гениталий (Ingham, McMahon, 2001; Roy, Ingham, 2002; Fitzsimons et al., 2022; Ingham, 2022; Dilower et al., 2023).

Среди фундаментальных внутриклеточных процессов, в регуляции которых задействованы гены HH пути, транскрипция (Gao Y. et al., 2023), ответ на стрессовые воздействия (Chung et al., 2022), поддержание геномной стабильности (Ingham, McMahon, 2001). Кроме того, этот сигнальный путь модулирует клеточный ответ на гипоксию, окислительный стресс и другие наблагоприятные факторы, что может быть критически важным для выживания клетки (Kim, Lee, 2023; van der Weele et al., 2024). Показано участие элементов Hedgehog сигнального пути в репарации ДНК (Gao Q. et al., 2019), апоптозе (Harris et al., 2011; Rimkus et al., 2016) и регуляции клеточного цикла, что подтверждает его роль в контроле пролиферации и дифференцировки клеток (Roy, Ingham, 2002).

Согласно существующим данным, HH путь является средством межклеточных коммуникаций не только сам по себе, его компоненты, в частности бета-аррестины (ARRB1/2), киназы (CCND1, CSNK1A1, CSNK1E, CSNK1A1L, GSK3B, PRKACA, PRKACB, PRKACG, TPTEP2-CSNK1E), белки убиквитинизации (BTRC, CUL1, FBXW11) и др. вовлечены в состав других сигнальных каскадов, в том числе MAPK/ERK, Wnt, Notch, VEGF. Участие белков HH пути в других цепях передачи сигналов продемонстрировано и в работах (Rubin, 2007; Butí et al., 2014; Edeling et al., 2016; Luo, 2017; Fang et al., 2023).

Ассоциативная генная сеть Hedgehog сигнального пути

Сеть содержит информацию о 56 генах, 504 белках, 126 миРНК и 1412 взаимодействиях разного типа между ее элементами. Общий вид сети представлен на рис. 4.

Анализ генной сети позволил выявить некоторые закономерности, касающиеся внутрисетевых взаимодействий. В частности, показано, что в рамках сети существует по крайней мере семь регуляторных контуров (рис. 5, 6). Условно их можно разделить на две группы.

С участием контуров первой группы происходит авторегуляция сигнального пути как целого. Вторая группа регламентирует взаимодействие некоторых компонентов

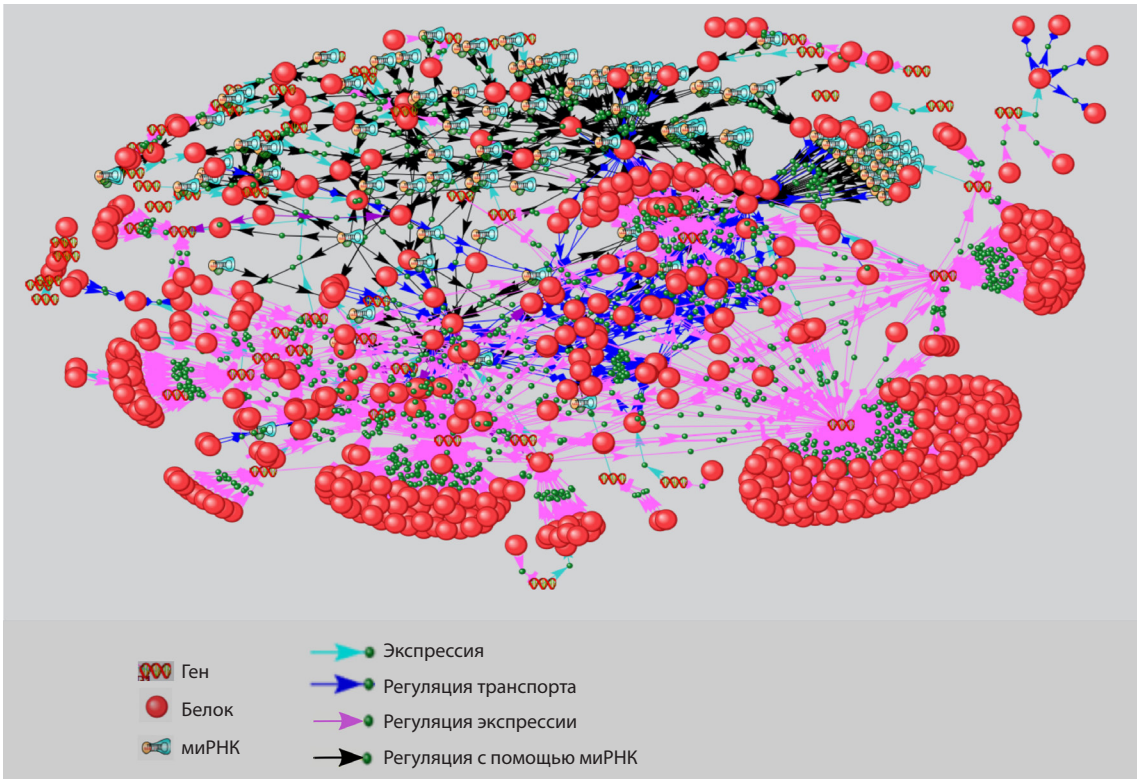


Рис. 4. Общий вид ассоциативной генной сети Hedgehog сигнального пути человека, реконструированной с помощью компьютерного ресурса ANDSystem.

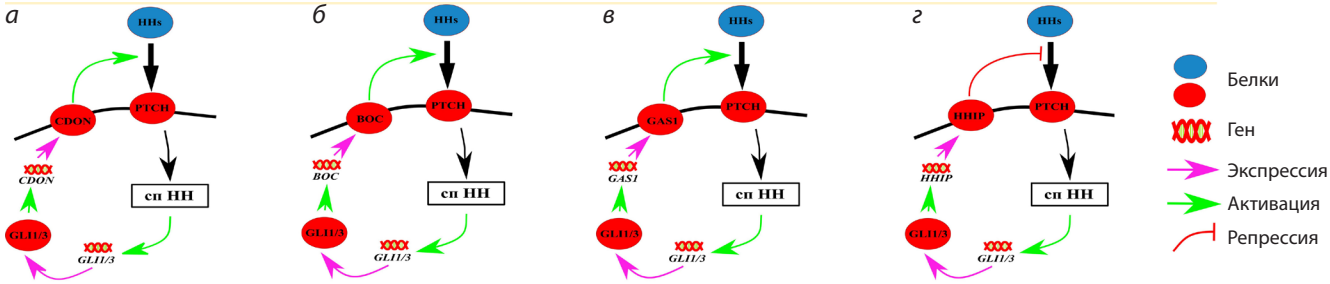


Рис. 5. Авторегуляция HH сигнального пути.

а–в – регуляторные контуры с положительной обратной связью; *г* – регуляторный контур с отрицательной обратной связью; СП – сигнальный путь.

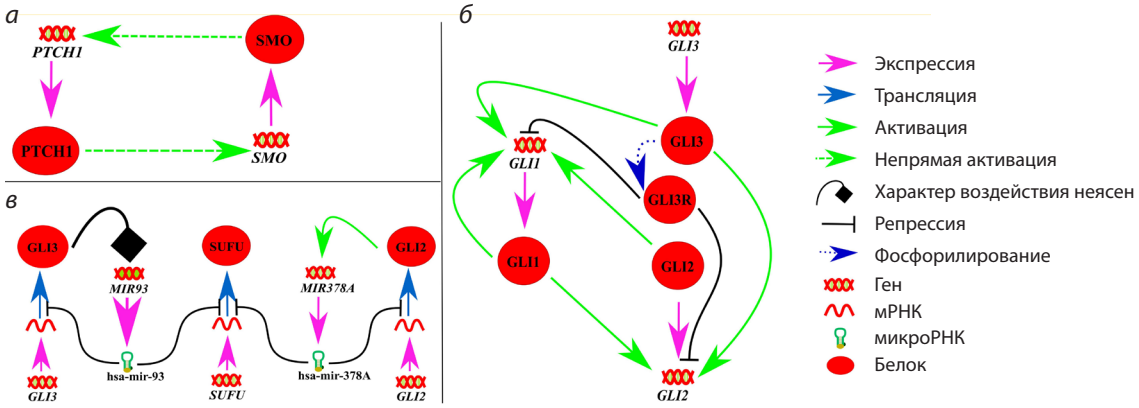


Рис. 6. Схемы взаимной регуляции компонентов трех регуляторных контуров HH сигнального пути.

а – взаимная регуляция PTCH1 и SMO; *б* – авторегуляция GLI1/3; *в* – регуляция GLI2/3 и SUFU.

самого сигнального пути**.** Первая группа включает четыре контура – три с положительными обратными связями, осуществляющими автоактивацию пути (см. рис. 5, *а–в*), и один с отрицательной обратной связью, через который реализуется авторепрессия пути (см. рис. 5, *г*). В состав контуров автоактивации входят мембранные белки GAS1, BOC, CDON, участвующие во взаимодействии рецептора PTCH1/2 с его лигандом HH и тем самым способствующие проведению сигнала. Экспрессия генов, кодирующих названные мембранные белки, контролируется транскрипционными факторами GLI1/3 (Allen et al., 2007; Song et al., 2015; Echevarría-Andino et al., 2023).

Главный компонент четвертого контура – белок HHIP, препятствующий связыванию PTCH1/2 с HH, что запрещает прохождение сигнала. Ген *HHIP* является мишенью GLI1/3 (Chuang, McMahon, 1999; Falkenstein, Vokes, 2014).

Вторая группа, определяющая характер некоторых взаимодействий внутри HH пути, сформирована тремя контурами. Первый контролирует взаимодействие между PTCH1 и SMO через положительную обратную связь (см. рис. 6, *а*); второй – контур взаиморегуляции генов, кодирующих транскрипционные факторы GLI1/3 (см. рис. 6, *б*), в зависимости от функционального статуса пути может находиться в двух состояниях. При наличии HH сигнала контур работает в режиме взаимоактивации генов через положительные обратные связи. При отсутствии сигнала репрессорная форма GLI3R подавляет транскрипцию генов *GLI1/2* и выключает автоактивацию. Таким образом осуществляется поддержание баланса между активаторными и репрессорными формами GLI (Wang et al., 2000; Vokes et al., 2007; Briscoe, Thérond, 2013). Третий контур группы функционирует с участием двух миРНК, hsa-mir-93 и hsa-mir-378A, регулируя содержание GLI2/3 и SUFU через отрицательные обратные связи (см. рис. 6, *в*). Факт участия миРНК, в том числе hsa-mir-93 и hsa-mir-378A, в регуляции экспрессии белков HH пути установлен в работе (Helwak et al., 2013). Анализ реконструированной генной сети HH сигнального пути показал, что гены, кодирующие эти миРНК, являются мишенями для транскрипционных факторов GLI2/3.

Эволюционные характеристики генов Hedgehog сигнального пути человека

Распределение генов по значениям филостратиграфического индекса PAI представлено в табл. 2 и на рис. 7.

Подавляющее большинство генов, формирующих путь, характеризуется индексами PAI = 01 (35 генов) и PAI = 02 (18 генов), что говорит об их появлении на уровне первых одноклеточных эукариот и первых многоклеточных животных. Два гена, *BCL2* и *SUFU*, возникли значительно раньше – на этапе клеточного уровня организации живого (их PAI = 00). Оба гена контролируют численность пула клеток – *BCL2* как регулятор апоптоза, а *SUFU* – как ингибитор роста опухолей, т. е. неконтролируемого размножения клеток (Willis et al., 2003; Cheng, Yue, 2008).

Индекс PAI = 03 имеет только один ген, *HHIP*, возникший на стадии формирования хордовых. Одноименный белок ингибирует сигнальный каскад уже на начальном этапе его функционирования, связываясь с рецептором PTCH1/2 и препятствуя взаимодействию лиганд–рецептор.

Ранее независимые данные о времени появления некоторых компонентов HH сигнального пути человека до дивергенции позвоночных были получены для всех лигандов HH (Kumar et al., 1996) и транскрипционных факторов GLI (Shimeld et al., 2007), что не противоречит приведенным результатам.

Сравнение распределения значений PAI между генами HH каскада и всеми белок-кодирующими генами человека (см. рис. 7) показало статистически значимое смещение в сторону более древних значений у генов HH пути ($p < 0.05$, критерий Манна–Уитни). Это согласуется с тем фактом, что в онтогенезе этот путь начинает функционировать ранее других, и можно полагать, что его основные компоненты поэтому должны были сформироваться на ранних этапах эволюции многоклеточных. Действительно, все формы белков HH, GLI, PTCH и SMO, играющие основную роль в передаче сигнала, характеризуются PAI = 01–02, а их функциональные аналоги присутствуют и у беспозвоночных животных (Ingham, McMahon, 2001; Wilson, Chuang, 2010). Примечательно, что все гены регуляторных контуров, за исключением *HHIP*, имеют древнее происхождение, причем *HHIP* – единственный ген, входящий в регуляторный контур с отрицательной обратной связью.

На рис. 8 показано распределение значений индекса DI для генов HH пути. Учитывая, что весь HH путь ориентирован на реализацию фундаментальных клеточных процессов, задействованных в морфогенезе, в том числе деления, дифференцировки, апоптоза, неудивительно, что 89 % его генов (50) имеют индекс DI < 0.5, причем 12 из них (≈ 21 %) – индекс ниже 0.1. Этот факт подтверждает, что сигнальный путь и в его составе гены регуляторных контуров, регламентирующих функционирование пути, подвергаются давлению стабилизирующего отбора, ограничивающему накопление геномных изменений.

Таблица 2. Распределение 56 генов Hedgehog сигнального пути человека по значениям филостратиграфического индекса PAI

Индекс PAI _Таксон	Гены
00_Cellular Organisms	***SUFU***, *BCL2*
01_Eucaryota	*ARRB1*, *ARRB2*, *BTRC*, *CCND1*, *CCND2*, *CSNK1A1*, *CSNK1A1L*, *CSNK1D*, *CSNK1E*, *CSNK1G1*, *CSNK1G2*, *CSNK1G3*, *CUL1*, *CUL3*, ***DHH***, *DISP1*, *EFCAB7*, *FBXW11*, *GRK2*, *GRK3*, *GSK3B*, ***IHH***, *KIF3A*, *KIF7*, *MOSMO*, *PRKACA*, *PRKACB*, *PRKACG*, ***PTCH1***, ***PTCH2***, *SMURF1*, *SMURF2*, *SPOP*, *SPOPL*, *TPTEP2-CSNK1E*
02_Metazoa	***BOC***, ***CDON***, *EVC*, *EVC2*, ***GAS1***, ***GLI1***, ***GLI2***, ***GLI3***, *GPR161*, *HHAT*, *HHATL*, *IQCE*, *LRP2*, *MEGF8*, *MGRN1*, *SCUBE2*, ***SHH***, ***SMO***
03_Chordata	***HHIP***

Примечание. Жирным шрифтом выделены названия генов, входящих в регуляторные контуры с обратными связями.

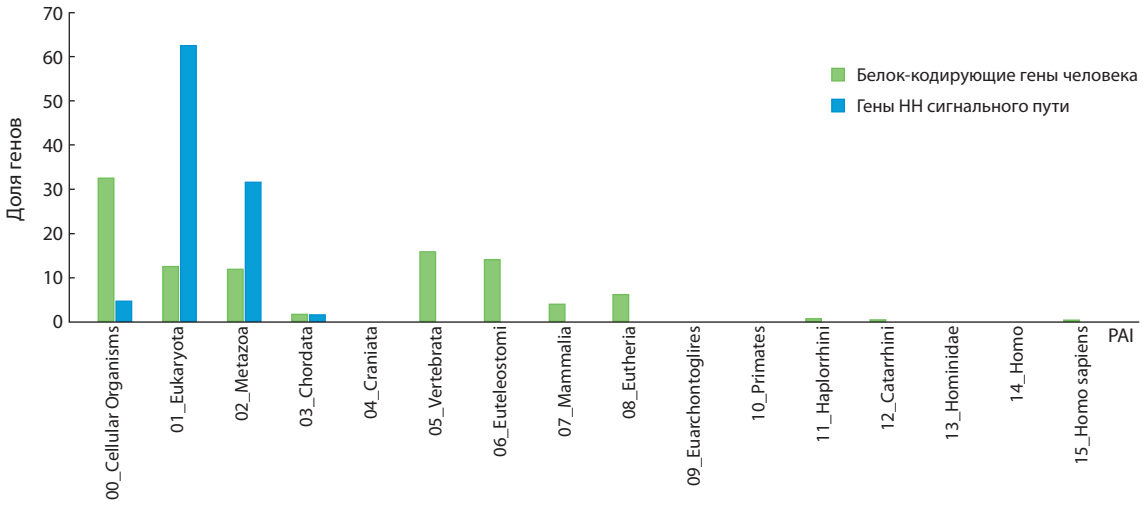


Рис. 7. Распределение значений PAI между генами Hedgehog сигнального пути (56 генов) и всеми белок-кодирующими генами человека (19 491 ген).

Различия в значениях достоверны с $p < 0.05$ по критерию Манна–Уитни.

В анализируемой выборке из 56 генов выделяются только два с DI > 1 – это гены *CSNK1A1L* (1.213) и *EFCAB7* (1.051), что в рамках примененного метода предполагает возможность действия на них движущего отбора. Киназа CSNK1A1L фосфорилирует белки GLI1/3. Согласно данным БД KEGG (hsa04340), в HH сигнальном пути человека в этом процессе наряду с CSNK1A1L участвует еще несколько киназ (CSNK1A1, CSNK1D, CSNK1E, CSNK1G1, CSNK1G2, CSNK1G3, TPTEP2-CSNK1E), кодируемых одноименными генами. Все они попадают в группу с PAI = 02_Eukaryota, однако значения DI для них лежат в диапазоне от 0.0361 для *CSNK1A1* до 0.264 для *CSNK1D*, что свидетельствует о действии на них стабилизирующего отбора. Можно предположить, что *CSNK1A1L*, «встроился» в сигнальный путь эволюционно несколько позже остальных генов киназ, и поэтому пока, возможно, испытывает влияние движущего, а не стабилизирующего, отбора.

Белок EFCAB7 совместно с белками EVC, EVC2 и IQCE участвует в закреплении SMO на первичной ресничке клеток млекопитающих, что отличает механизм трансдукции сигнала от аналогичного процесса у дрозофилы, в клетках которой первичных ресничек нет (Chen et al., 2009; Gorojankina, 2016). Вероятно, слабое давление движущего отбора на ген *EFCAB7*, отраженное в близком к единице значении DI, связано именно с более поздним появлением механизма, вовлекающего первичные реснички в процесс передачи сигнала, по сравнению с другими компонентами пути, выполняющими ту же функцию, – генами *EVC*, *EVC2* и *IQCE* (Chen et al., 2009; Wilson, Chuang 2010), на которые, очевидно, действует стабилизирующий отбор, о чем свидетельствуют значения DI, составляющие 0.298, 0.421 и 0.679 соответственно.

Таким образом, большинство генов сигнального пути Hedgehog можно охарактеризовать как древние, подверженные действию стабилизирующего отбора, препятствующего накоплению генной изменчивости и способствующего функциональной стабильности генов. Их консервативность подтверждает критическую роль HH пути в регуляции фундаментальных онтогенетических процессов.

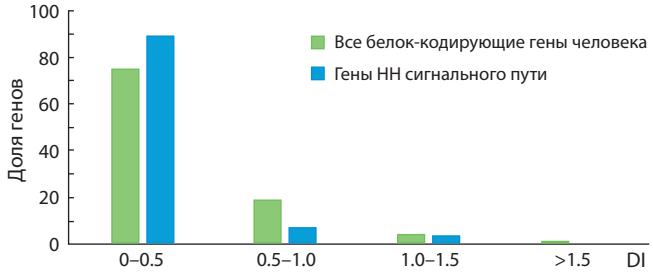


Рис. 8. Распределение значений индекса DI для генов Hedgehog сигнального пути и всех белок-кодирующих генов человека.

Заключение

Разработан прототип базы знаний HH_Signal_pathway_db, аккумулирующей информацию о структурно-функциональной организации эволюционно консервативного Hedgehog сигнального пути у человека, интегрирующей сведения из KEGG, NCBI Gene, UniProt и других источников. База структурирует разрозненные данные о HH сигнальном пути человека и может служить инструментом для системного анализа его роли в онтогенезе, поддержании гомеостаза и развитии патологий.

Проведенный биоинформатический анализ некоторых данных из базы, в частности, показал, что: 1) согласно функциональной аннотации, гены пути ассоциированы с тремя категориями процессов: внутриклеточными, морфогенезом органов и межклеточной коммуникацией, в том числе взаимодействием с другими сигнальными каскадами; 2) значительное большинство генов пути имеет древнее происхождение и подвержено стабилизирующему отбору; 3) реконструированная ассоциативная генная сеть HH пути содержит 56 генов, 504 белка, 126 миРНК

и устанавливает 1412 взаимодействий между ними; 4) функционирование сети регламентируется семью регуляторными контурами – пятью с положительными и двумя с отрицательными обратными связями, причем в один из контуров с отрицательной обратной связью входит две миРНК.

Список литературы / References

Иванисенко В.А., Деменков П.С., Иванисенко Т.В., Колчанов Н.А. ANDSystem: когнитивная система для реконструкции и анализа графа знаний (генных сетей) на основе автоматического извлечения знаний из текстов научных публикаций, патентов и фактографических баз данных. *Наука и технологии Сибири*. 2022;4(7):122-125. Доступно: https://scitech.sb-ras.ru/upload/iblock/010/5ttp14te9uu1r5g0suu7ka05ui8udynq/nit_2022_7.pdf [Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Kolchanov N.A. ANDSystem: a cognitive system for the reconstruction and analysis of knowledge graphs (gene networks) based on the automated extraction of data from scientific publications, patents, and factual databases. *Nauka i Tekhnologii Sibiri*. 2022;4(7):122-125. Available at: https://scitech.sb-ras.ru/upload/iblock/010/5ttp14te9uu1r5g0suu7ka05ui8udynq/nit_2022_7.pdf (in Russian)]

Филонов С.В., Подколодный Н.Л., Подколодная О.А., Твердохлеб Н.Н., Пономаренко П.М., Рассказов Д.А., Богомолов А.Г., Пономаренко М.П. Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования. *Вавиловский журнал генетики и селекции*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85 [Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(7):728-736. doi 10.18699/VJGB-23-85]

Allen B.L., Tenzen T., McMahon A.P. The Hedgehog-binding proteins Gas1 and Cdo cooperate to positively regulate Shh signaling during mouse development. *Genes Dev*. 2007;21(10):1244-1257. doi 10.1101/gad.1543607

Artavanis-Tsakonas S., Rand M.D., Lake R.J. Notch signaling: cell fate control and signal integration in development. *Science*. 1999; 284(5415):770-776. doi 10.1126/science.284.5415.770

Bartel D.P. Metazoan microRNAs. *Cell*. 2018;173(1):20-51. doi 10.1016/j.cell.2018.03.006

Breeze E. Role of Hedgehog signalling pathway in the maintenance and regeneration of adult tissues. *J Cell Signal*. 2022;7:281. doi 10.35248/2576-1471.22.7.281

Brennan D., Chen X., Cheng L., Mahoney M., Riobo N.A. Noncanonical Hedgehog signaling. *Vitam Horm*. 2012;88:55-72. doi 10.1016/B978-0-12-394622-5.00003-1

Briscoe J., Thérond P.P. The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat Rev Mol Cell Biol.* 2013; 14(7):416-429. doi 10.1038/nrm3598

Butí E., Mesquita D., Araújo S.J. Hedgehog is a positive regulator of FGF signalling during embryonic tracheal cell migration. *PLoS One*. 2014;9(3):e92682. doi 10.1371/journal.pone.0092682

Carballo G.B., Honorato J.R., de Lopes G.P.F., Spohr T.C.L.S.E. A highlight on Sonic hedgehog pathway. *Cell Commun Signal*. 2018;16(1):11. doi 10.1186/s12964-018-0220-7

Chen M.H., Wilson C.W., Li Y.J., Law K.K., Lu C.S., Gacayan R., Zhang X., Hui C.C., Chuang P.T. Cilium-independent regulation of Gli protein function by Sufu in Hedgehog signaling is evolutionarily conserved. *Genes Dev*. 2009;23(16):1910-1928. doi 10.1101/gad.1794109

Cheng S.Y., Yue S. Role and regulation of human tumor suppressor SUFU in Hedgehog signaling. *Adv Cancer Res.* 2008;101:29-43. doi 10.1016/S0065-230X(08)00402-8

Chuang P.T., McMahon A.P. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature*. 1999; 397(6720):617-621. doi 10.1038/17611

Chung K.M., Kim H., Roque C.G., McCurdy E.P., Nguyen T.T.T., Siegelin M.D., Hwang J.Y., Hengst U. A systemic cell stress signal confers neuronal resilience toward oxidative stress in a Hedgehog-dependent manner. *Cell Rep*. 2022;41(3):111488. doi 10.1016/j.celrep.2022.111488

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2011;11(3):149-161. doi 10.3233/ISB-2012-0449

Dilower I., Niloy A.J., Kumar V., Kothari A., Lee E.B., Rumi M.A.K. Hedgehog signaling in gonadal development and function. *Cells*. 2023;12(3):358. doi 10.3390/cells12030358

Dutta R.K., Jun J., Du K., Diehl A.M. Hedgehog signaling: implications in liver pathophysiology. *Semin Liver Dis*. 2023;43(4):418-428. doi 10.1055/a-2187-3382

Echevarría-Andino M.L., Franks N.E., Schrader H.E., Hong M., Krauss R.S., Allen B.L. CDON contributes to Hedgehog-dependent patterning and growth of the developing limb. *Dev Biol*. 2023;493: 1-11. doi 10.1016/j.ydbio.2022.09.011

Edeling M., Ragi G., Huang S., Pavenstädt H., Susztak K. Developmental signalling pathways in renal fibrosis: the roles of Notch, Wnt and Hedgehog. *Nat Rev Nephrol*. 2016;12(7):426-439. doi 10.1038/nrneph.2016.54

Eggenschwiler J.T., Anderson K.V. Cilia and developmental signaling. *Annu Rev Cell Dev Biol*. 2007;23:345-373. doi 10.1146/annurev.cellbio.23.090506.123249

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi 10.1038/nature11247

Falkenstein K.N., Vokes S.A. Transcriptional regulation of graded Hedgehog signaling. *Semin Cell Dev Biol*. 2014;33:73-80. 10.1016/j.semcdb.2014.05.010

Fang Z., Meng Q., Xu J., Wang W., Zhang B., Liu J., Liang C., Hua J., Zhao Y., Yu X., Shi S. Signaling pathways in cancer-associated fibroblasts: recent advances and future perspectives. *Cancer Commun* (*Lond*). 2023;43(1):3-41. doi 10.1002/cac2.12392

Fitzsimons L.A., Brewer V.L., Tucker K.L. Hedgehog morphogens act as growth factors critical to pre- and postnatal cardiac development and maturation: how primary cilia mediate their signal transduction. *Cells*. 2022;11(12):1879. doi 10.3390/cells11121879

Gao Q., Zhou G., Lin S.J., Paus R., Yue Z. How chemotherapy and radiotherapy damage the tissue: comparative biology lessons from feather and hair models. *Exp Dermatol*. 2019;28(4):413-418. doi 10.1111/exd.13846

Gao Y., Shan Z., Jian C., Wang Y., Yao X., Li S., Ti X., Zhao G., Liu C., Zhang Q. HIB/SPOP inhibits Ci/Gli-mediated tumorigenesis by modulating the RNA polymerase II components stabilities. *iScience*. 2023;26(8):107334. doi 10.1016/j.isci.2023.107334

Ghafouri-Fard S., Khoshbakht T., Hussen B.M., Taheri M., Samsami M. Emerging role of non-coding RNAs in the regulation of Sonic Hedgehog signaling pathway. *Cancer Cell Int*. 2022;22(1):282. doi 10.1186/s12935-022-02702-y

Gorojankina T. Hedgehog signaling pathway: a novel model and molecular mechanisms of signal transduction. *Cell Mol Life Sci*. 2016; 73(7):1317-1332. doi 10.1007/s00018-015-2127-4

Harris L.G., Samant R.S., Shevde L.A. Hedgehog signaling: networking to nurture a promalignant tumor microenvironment. *Mol Cancer Res*. 2011;9(9):1165-1174. doi 10.1158/1541-7786.MCR-11-0175

Helwak A., Kudla G., Dudnakova T., Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*. 2013;153(3):654-665. doi 10.1016/j.cell.2013.03.043

Huttlin E.L., Bruckner R.J., Paulo J.A., Cannon J.R., Ting L., Baltier K., Colby G., … Guruharsha K.G., Li K., Artavanis-Tsakonas S., Gygi S.P., Harper J.W. Architecture of the human interactome defines protein communities and disease networks. *Nature*. 2017; 545(7655):505-509. doi 10.1038/nature22366

Ingham P.W. Hedgehog signaling. *Curr Top Dev Biol*. 2022;149:1-58. doi 10.1016/bs.ctdb.2022.04.003

Ingham P.W., McMahon A.P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev*. 2001;15(23):3059-3087. doi 10.1101/gad.938601

Ingham P.W., Nakano Y., Seger C. Mechanisms and functions of Hedgehog signalling across the metazoa. *Nature Rev Genet*. 2011;12(6):393-406. doi 10.1038/nrg2984

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9:S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanov R.A., Mukhin A.M., Kazantsev F.V., Mustafin Z.S., Afonnikov D.A., Matushkin Y.G., Lashin S.A. Orthoweb: a software package for evolutionary analysis of gene networks. *Vavilov J Genet Breed*. 2024;28(8):874-881. doi 10.18699/vjgb-24-95

Jamieson C., Martinelli G., Papayannidis C., Cortes J.E. Hedgehog pathway inhibitors: a new therapeutic class for the treatment of acute myeloid leukemia. *Blood Cancer Discov*. 2020;1(2):134-145. doi 10.1158/2643-3230.BCD-20-0007

Jeffares D.C., Tomiczek B., Sojo V., dos Reis M. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol Biol*. 2015;1201:65-90. doi 10.1007/978-1-4939-1438-8_4

Jing J., Wu Z., Wang J., Luo G., Lin H., Fan Y., Zhou C. Hedgehog signaling in tissue homeostasis, cancers, and targeted therapies. *Signal Transduct Target Ther*. 2023;8(1):315. doi 10.1038/s41392-023-01559-5

Kenneth J.H. Big Data among Big Data: Genome Data. 2022. Available at: https://3billion.io/blog/big-data-among-big-data-genome-data/

Kim N.H., Lee A.Y. Oxidative stress induces skin pigmentation in melasma by inhibiting Hedgehog signaling. *Antioxidants* (*Basel*). 2023;12(11):1969. doi 10.3390/antiox12111969

Kumar S., Balczarek K.A., Lai Z.C. Evolution of the *hedgehog* gene family. *Genetics*. 1996;142(3):965-972. doi 10.1093/genetics/142.3.965

Logan C.Y., Nusse R. The Wnt signaling pathway in development and disease. *Ann Rev Cell Dev Biol*. 2004;20:781-810. doi 10.1146/annurev.cellbio.20.010403.113126

Luo K. Signaling cross talk between TGF-β/Smad and other signaling pathways. *Cold Spring Harb Perspect Biol*. 2017;9(1):a022137. doi 10.1101/cshperspect.a022137

McIntyre G., Jackson Z., Colina J., Sekhar S., DiFeo A. *miR-181a*: regulatory roles, cancer-associated signaling pathway disruptions, and therapeutic potential. *Expert Opin Ther Targets*. 2024;28(12):1061-1091. doi 10.1080/14728222.2024.2433687

Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006

Nüsslein-Volhard C., Wieschaus E. Mutations affecting segment number and polarity in Drosophila. *Nature*. 1980;287(5785):795-801. doi 10.1038/287795a0

Oro A.E. The primary cilia, a 'Rab-id' transit system for Hedgehog signaling. *Curr Opin Cell Biol*. 2007;19(6):691-696. doi 10.1016/j.ceb.2007.10.008

Perrimon N., Pitsouli C., Shilo B.Z. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb Perspect Biol*. 2012;4(8):a005975. doi 10.1101/cshperspect.a005975

Regev A., Teichmann S.A., Lander E.S., Amit I., Benoist C., Birney E., Bodenmiller B., … Watt F., Weissman J., Wold B., Xavier R., Yosef N., Human Cell Atlas Meeting Participants. The Human Cell Atlas. *eLife*. 2017;6:e27041. doi 10.7554/eLife.27041

Rimkus T.K., Carpenter R.L., Qasem S., Chan M., Lo H.W. Targeting the sonic Hedgehog signaling pathway: review of Smoothened and GLI inhibitors. *Cancers* (*Basel*). 2016;8(2):22. doi 10.3390/cancers8020022

Roy S., Ingham P.W. Hedgehogs tryst with the cell cycle. *J Cell Sci*. 2002;115(Pt 23):4393-4397. doi 10.1242/jcs.00158

Rubin D.C. Intestinal morphogenesis. *Curr Opin Gastroenterol*. 2007;23(2):111-114. doi 10.1097/MOG.0b013e3280145082

Schermelleh L., Ferrand A., Huser T., Eggeling C., Sauer M., Biehlmaier O., Drummen G.P. Super-resolution microscopy demystified. *Nat Cell Biol*. 2019;21(1):72-84. doi 10.1038/s41556-018-0251-8

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Shimeld S.M., van den Heuvel M., Dawber R., Briscoe J. An amphioxus Gli gene reveals conservation of midline patterning and the evolution of hedgehog signalling diversity in chordates. *PLoS One*. 2007;2(9):e864. doi 10.1371/journal.pone.0000864

Skoda A.M., Simovic D., Karin V., Kardum V., Vranic S., Serman L. The role of the Hedgehog signaling pathway in cancer: a comprehensive review. *Bosn J Basic Med Sci*. 2018;18(1):8-20. doi 10.17305/bjbms.2018.2756

Song J.Y., Holtz A.M., Pinskey J.M., Allen B.L. Distinct structural requirements for CDON and BOC in the promotion of Hedgehog signaling. *Dev Biol*. 2015;402(2):239-252. doi 10.1016/j.ydbio.2015.03.015

Spielman S.J., Wilke C.O. The relationship between dN/dS and scaled selection coefficients. *Mol Biol Evol*. 2015;32(4):1097-1108. doi 10.1093/molbev/msv003

Spinella-Jaegle S., Rawadi G., Kawai S., Gallea S., Faucheu C., Mollat P., Courtois B., Bergaud B., Ramez V., Blanchet A.M., Adelmant G., Baron R., Roman-Roman S. Sonic hedgehog increases the commitment of pluripotent mesenchymal cells into the osteoblastic lineage and abolishes adipocytic differentiation. *J Cell Sci*. 2001;114(Pt. 11):2085-2094. doi 10.1242/jcs.114.11.2085

van der Weele C.M., Hospes K.C., Rowe K.E., Jeffery W.R. Hypoxia-sonic hedgehog axis as a driver of primitive hematopoiesis development and evolution in cavefish. *Dev Biol*. 2024;516:138-147. doi 10.1016/j.ydbio.2024.08.008

Varjosalo M., Taipale J. Hedgehog signaling. *J Cell Sci*. 2007;120(Pt. 1):3-6. doi 10.1242/jcs.03309

Vokes S.A., Ji H., McCuine S., Tenzen T., Giles S., Zhong S., Longabaugh W.J., Davidson E.H., Wong W.H., McMahon A.P. Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development*. 2007;134(10):1977-1989. doi 10.1242/dev.001966

Wang B., Fallon J.F., Beachy P.A. Hedgehog-regulated processing of Gli3 produces an anterior/posterior repressor gradient in the developing vertebrate limb. *Cell*. 2000;100(4):423-434. doi 10.1016/s0092-8674(00)80678-9

Willis S., Day C.L., Hinds M.G., Huang D.C. The Bcl-2-regulated apoptotic pathway. *J Cell Sci*. 2003;116(Pt. 20):4053-4056. doi 10.1242/jcs.00754

Wilson C.W., Chuang P.T. Mechanism and evolution of cytosolic Hedgehog signal transduction. *Development*. 2010;137(13):2079-2094. doi 10.1242/dev.045021

Wu F., Zhang Y., Sun B., McMahon A.P., Wang Y. Hedgehog signaling: from basic biology to cancer therapy. *Cell Chem Biol*. 2017;24(3):252-280. doi 10.1016/j.chembiol.2017.02.010

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 22.07.2025. После доработки 17.09.2025. Принята к публикации 17.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):990-999

doi 10.18699/vjgb-25-104

Выявление белков, регулирующих фенотип-ассоциированные гены макрофагов группы M2: биоинформатический анализ

Е.А. Антропова, И.В. Яцык, П.С. Деменков, Т.В. Иванисенко, В.А. Иванисенко

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
nzhenia@bionet.nsc.ru

Аннотация. Макрофаги – клетки иммунной системы, выполняющие в организме различные, часто противоположные функции в зависимости от поступающих сигналов микроокружения. Это возможно благодаря пластичности макрофагов, позволяющей кардинально менять фенотипические признаки и профили экспрессии генов, а также возвращаться в исходное, неактивированное состояние. В зависимости от действующих на клетку индукторов макрофаги поляризуются в различные функциональные состояния. Принято выделять пять основных фенотипов активированных макрофагов: M1, M2a, M2b, M2c и M2d. Хотя количество полногеномных транскриптомных и протеомных данных, показывающих различия между основными фенотипами макрофагов и неактивированными макрофагами (M0), растет стремительно, все еще остаются вопросы, касающиеся механизмов регуляции профилей экспрессии генов и белков у макрофагов разных фенотипов. Нами были составлены списки белков, ассоциированных с фенотипами макрофагов M1, M2a, M2b, M2c, M2d (фенотип-ассоциированные белки), проанализированы данные о возможных посредниках поляризации макрофагов. Далее с использованием компьютерной системы ANDSystem проведен поиск и анализ связей между потенциальными регуляторными белками и генами, кодирующими белки, ассоциированные с фенотипами группы M2, получены оценки статистической значимости этих связей. Результаты указывают на то, что различия в фенотипах макрофагов M2a, M2b, M2c, M2d могут быть обусловлены регуляторными действиями белков JUN, IL8, NFAC2, CCND1 и YAP1. Уровень их экспрессии варьируется в зависимости от фенотипов группы M2, что в свою очередь приводит к различным уровням экспрессии генов, связанных с конкретными фенотипами.

Ключевые слова: фенотипы макрофагов; регуляция экспрессии; протеомы; система ANDSystem; автоматический анализ текстов

Для цитирования: Антропова Е.А., Яцык И.В., Деменков П.С., Иванисенко Т.В., Иванисенко В.А. Выявление белков, регулирующих фенотип-ассоциированные гены макрофагов группы M2: биоинформатический анализ. *Вавиловский журнал генетики и селекции*. 2025;29(7):990-999. doi 10.18699/vjgb-25-104

Финансирование. Работа выполнена в рамках бюджетного проекта FWNR-2022-0020.

Identification of proteins regulating phenotype-associated genes of M2 macrophages: a bioinformatic analysis

E.A. Antropova, I.V. Yatsyk, P.S. Demenkov, T.V. Ivanisenko, V.A. Ivanisenko

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
nzhenia@bionet.nsc.ru

Abstract. Macrophages are immune system cells that perform various, often opposing, functions in the organism depending on the incoming microenvironment signals. This is possible due to the plasticity of macrophages, which allows them to radically alter their phenotypic characteristics and gene expression profiles, as well as return to their original, non-activated state. Depending on the inductors acting on the cell, macrophages are activated into various functional states. There are five main phenotypes of activated macrophages: M1, M2a, M2b, M2c, and M2d. Although the amount of genome-wide transcriptomic and proteomic data showing differences between major macrophage phenotypes and non-activated macrophages (M0) is rapidly growing, questions regarding the mechanisms regulating gene and protein expression profiles in macrophages of different phenotypes still remain. We compiled lists of proteins associated with the macrophage phenotypes M1, M2a, M2b, M2c, and M2d (phenotype-associated proteins) and analyzed the data on potential mediators of macrophage polarization. Furthermore, using the computational system ANDSystem, we conducted a search and analysis of the relationships between potential regulatory proteins and the genes encoding the proteins associated with the M2 group phenotypes, obtaining estimates of the statistical significance of these relationships. The results indicate that the differences in the M2a, M2b, M2c, and M2d macro-

© Антропова Е.А., Яцык И.В., Деменков П.С., Иванисенко Т.В., Иванисенко В.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

phage phenotypes may be attributed to the regulatory effects of the proteins JUN, IL8, NFAC2, CCND1, and YAP1. The expression levels of these proteins vary among the M2 group phenotypes, which in turn leads to different levels of gene expression associated with specific phenotypes.

Key words: macrophage phenotypes; expression regulation; proteomes; ANDSystem; automated text analysis

For citation: Antropova E.A., Yatsyk I.V., Demenkov P.S., Ivanisenko T.V., Ivanisenko V.A. Identification of proteins regulating phenotype-associated genes of M2 macrophages: a bioinformatic analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):990-999. doi 10.18699/vjgb-25-104

Введение

Макрофаги – клетки иммунной системы, играющие ключевую роль в таких процессах, как: поддержание гомеостаза организма (Mosser et al., 2021), защита от инфекций (Zhang M., Wang, 2014), провоспалительный и противовоспалительный ответы (Xu et al., 2013), регенерация тканей с сопутствующей стимуляцией пролиферации (Wynn, Vannella, 2016) и многих других. Способность макрофагов проявлять разные функции благодаря поляризации – изменению функционального состояния в зависимости от сигналов из микроокружения – связана с их уникальной пластичностью (Mills, 2012; Gurvich et al., 2020). Поляризация приводит к тому, что макрофаги приобретают различные фенотипы – функциональные состояния, характеризующиеся уникальными морфологическими, молекулярными и функциональными особенностями, в зависимости от действовавших индукторов поляризации: белков, пептидов, полисахаридов и других.

Для каждого фенотипа макрофагов известна группа характерных для него белков (Martinez et al., 2008). Эти группы пересекаются между собой, но разные фенотипы макрофагов могут обладать кардинально отличающимися функциями. Так, фенотип М1соответствует провоспалительным макрофагам, необходимым для ответа организма на инфекции. Макрофаги М2a способствуют заживлению ран и очищают организм от апоптотических клеток (Murray et al., 2014). Макрофаги М2b называют регуляторными за их способность регулировать Т-хелперы, что приводит к переключению иммунного ответа из провоспалительного в противовоспалительный. Макрофаги М2c необходимы для ремоделирования тканей, а также фагоцитоза апоптотических клеток. Макрофаги М2d считают опухоль-ассоциированными, так как они сопутствуют опухолевым тканям (Zhang Q., Sioud, 2023).

В ряде работ продемонстрированы связь отдельных фенотипов макрофагов с некоторыми патологиями либо ассоциация исхода заболеваний с определенными фенотипами макрофагов. Например, у пациентов с раком яичников выявлено выраженное преобладание макрофагов фенотипа М1 над М2, что было ассоциировано с улучшением выживаемости (Zhang M. et al., 2014), а изменение фенотипа макрофагов с М2 на М1 подавляло метастазирование опухоли (Yuan et al., 2017). Изучение ювенильного идиопатического артрита в стадии ремиссии показало, что в группе М2-макрофагов преобладали М2b и М2c, а количество макрофагов М2a было существенно снижено (Feng et al., 2021). При этом у детей со стадией активного ювенильного идиопатического артрита преобладали макрофаги фенотипов М2a и М2b, а представленность М2c была снижена. Исследование различий между фенотипами макрофагов имеет не только фундаментальное значение, но представляет также большой практический интерес для ранней диагностики заболеваний, прогноза и корректировки их развития (Zhang M. et al., 2014; Lampiasi, 2023).

Следует отметить, что в публикациях о белках и генах, характеризующих разные фенотипы макрофагов, встречается противоречивая информация. Например, fractalkine receptor (CX3CR1) в публикации (Joerink et al., 2011) указан как маркер фенотипа М2a, а в работе (Chhor et al., 2013) этот белок отмечен как маркер фенотипа М1. Металлопротеиназу MMP12 выделяют как маркер фенотипа М1 (Hirani et al., 2021), но также показано, что этот белок характерен для протеомов фенотипа М2 и дендритных клеток (Lee et al., 2014). Хемокин CXCL13 описан и как маркер М1 (Martinez et al., 2006), и как маркер М2 (van der Lans et al., 2015).

Каким образом при пересечении протеомов в макрофагах различных фенотипов возникают существенные функциональные различия? Какие молекулярно-генетические регуляторные механизмы лежат в основе поляризации макрофагов? Несмотря на стремительное накопление полногеномных транскриптомных и протеомных данных, характеризующих различия между основными фенотипами макрофагов и их различия с неактивированными макрофагами (М0) (Gurvich et al., 2020; Oates et al., 2023), все еще остаются открытыми вопросы о том, каким образом регулируются профили экспрессии генов и белков в макрофагах различных фенотипов.

Цель настоящей работы заключалась в поиске белков-посредников, контролирующих активность фенотип-ассоциированных генов в разных фенотипах макрофагов группы М2. Для этого использовалась программно-информационная система ANDSystem, основанная на применении методов машинного чтения и искусственного интеллекта, включая графовые нейронные сети (Ivanisenko V.A. et al., 2015; Ivanisenko T.V. et al., 2024). ANDSystem обеспечивает автоматический анализ текстов научных публикаций и фактографических баз данных медико-биологической направленности. В настоящее время база знаний системы ANDSystem содержит знания и факты, экстрагированные из более чем 40 млн научных публикаций и патентов, а также фактографических баз данных баз данных, включая информацию о молекулярно-генетических объектах и процессах, значимых для функционирования генных сетей и их базовых компонентов – метаболических сетей, путей передачи сигналов, сетей ДНК-белковых и белок-белковых взаимодействий. Эффективность ANDSystem показана в широком круге исследований: реконструкции молекулярно-генетических механизмов коморбидности астмы и гипертензии (Zolotareva et al., 2019), анализе метаболома плазмы крови пациентов с послеоперацион-

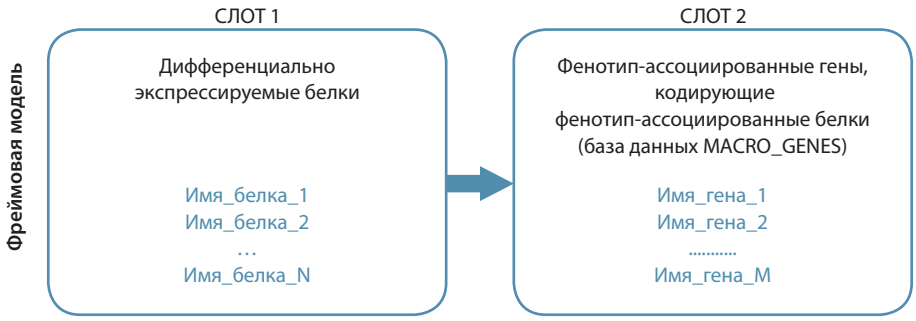


Рис. 1. Схема фреймовой модели для поиска регуляторных связей между дифференциально экспрессируемыми белками и фенотип-ассоциированными генами.

ным делирием (Иванисенко и др., 2023), реконструкции регуляторной сети гиперметилирования, влияющей на развитие гепатоцеллюлярной карциномы при заболевании гепатитом вируса С (Antropova et al., 2023).

В данной работе решали следующие задачи: 1) формирование списков фенотип-ассоциированных белков в макрофагах основных фенотипов: M1, M2a, M2b, M2c, M2d; 2) анализ данных по дифференциальной экспрессии белков в фенотипах группы M2 как потенциальных посредников поляризации макрофагов; 3) анализ регуляторных связей от белков-посредников к генам, кодирующим фенотип-ассоциированные белки с использованием ANDSystem.

Материалы и методы

Протеомные данные по макрофагам различных фенотипов. В работе использовано два типа информации о белках в разных фенотипах макрофагов:

1) Созданная нами курируемая база данных MACRO_GENES, содержащая списки генов и белков, ассоциированных с фенотипами макрофагов (табл. S1)[1]. Она сформирована на основе «ручного» анализа научных публикаций, в которых были описаны характерные белки, позволяющие различать фенотипы макрофагов M1, M2a, M2b, M2c, M2d. В базу данных MACRO_GENES были включены только те белки, присутствие которых в макрофагах определенных фенотипов подтверждено экспериментальными данными.
2) Протеомные данные о дифференциально экспрессируемых белках в макрофагах фенотипов M2a, M2b, M2c и M2d получены из работы (Li et al., 2022): примерно по 200 белков для каждого рассматриваемого фенотипа. Далее такие белки будем называть белками-регуляторами или дифференциально экспрессируемыми белками.

Поиск потенциальных регуляторов, влияющих на активность фенотип-ассоциированных генов. Поиск потенциальных белков-регуляторов, влияющих на активность фенотип-ассоциированных генов, осуществляли с использованием базы знаний программно-информационной системы ANDSystem и программного модуля ANDVisio, входящего в эту систему (Demenkov et al., 2012; Ivanisenko V.A. et al., 2015; Ivanisenko T.V. et al., 2024). База знаний ANDSystem включает информацию о взаимодействиях между молекулярно-биологическими объектами (генами, белками, метаболитами, биологическими процессами и др.), полученную на основе автоматического анализа более чем 40 млн научных публикаций и патентов, а также большого количества фактографических медико-биологических баз данных. В текущей версии эта база знаний содержит информацию о более чем 36 млн белков из разных организмов и примерно столько же генов, 76 тысяч метаболитов, 100 млн взаимодействий, 21 тысячу заболеваний и так далее.

Для поиска регуляторных связей между белками-регуляторами и фенотип-ассоциированными генами использовался программный аппарат фреймовых моделей системы ANDSystem (рис. 1). Шаг 1: первый слот фрейма наполнялся на основе данных протеомного анализа (Li et al., 2022) перечнем дифференциально экспрессируемых белков для каждого из фенотипов – M2a, M2b, M2c и M2d. Шаг 2: второй слот фрейма наполнялся перечнем фенотип-ассоциированных генов для того же фенотипа из сформированной нами курируемой базы данных MACRO_GENES (см. табл. S1). Шаг 3: с помощью программного модуля ANDVisio, использующего наполненный фрейм, в базе знаний системы ANDSystem проводился поиск описанных в ней регуляторных связей в изучаемом фенотипе макрофагов.

Результатами анализа были графы регуляторных процессов, в которых вершины соответствовали дифференциально экспрессируемым белкам из статьи (Li et al., 2022) и фенотип-ассоциированным генам макрофагов из базы данных MACRO_GENES. Ребра, соединяющие вершины графов, соответствовали регуляторным связям между ними.

Поиск функционально значимых белков-регуляторов фенотип-ассоциированных генов макрофагов в графах регуляторных процессов. Ключевой этап анализа регуляторных процессов, связанных с фенотип-ассоциированными генами макрофагов и выявленных с помощью фреймовых моделей, – поиск функционально значимых белков-регуляторов, называемых также центральными вершинами. Центральные вершины играют ключевую роль в передаче сигналов и координации регуляторных процессов. Для оценки центральности разработан широкий спектр методов (Ghasemi et al., 2014; Jalili et al., 2016; Ivanisenko V.A. et al., 2019). В нашей работе центральность вершин оценивали по количеству взаимодействий

[1] Табл. S1–S3 и рис. S1–S6 см. в Приложении по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx34.pdf

рассматриваемого белка с фенотип-ассоциированными генами соответствующего фенотипа.

Высокая степень центральности может наблюдаться как вследствие реальной функциональной значимости между генами и белками, так и по случайным причинам. Чтобы различить эти ситуации, была проведена оценка статистической значимости наблюдаемой степени центральности на основе гипергеометрического теста. В этом контексте гипергеометрический тест оценивает вероятность p того, что наблюдаемое количество связей конкретного рассматриваемого белка с фенотип-ассоциированными генами имеет случайный характер.

Здесь: M – общее число генов, представленных в базе знаний системы ANDSystem; N – общее число генов, с которыми взаимодействует конкретный белок в базе знаний системы ANDSystem; n – число фенотип-ассоциированных генов для конкретного фенотипа в базе данных MACRO_GENES; x – наблюдаемое число взаимодействий рассматриваемого белка с фенотип-ассоциированными генами конкретного фенотипа. Тогда при нулевой гипотезе случайного распределения взаимодействий величина X подчиняется гипергеометрическому закону распределения:

$$X \sim Hypergeom(M, N, n),$$

а величина p для правостороннего теста (обогащения) вычислялась по формуле:

$$p = P(X \geq x) = \sum_{k=x}^{\min(N,n)} \frac{\binom{N}{k}\binom{M-N}{n-k}}{\binom{M}{n}}.$$

Значения p рассчитывали с использованием библиотеки Python SciPy (scipy.stats.hypergeom). Для коррекции на множественные проверки применяли поправку Бонферрони. При $p < 0.05$ наблюдаемая степень центральности считалась статистически значимой и соответствующий белок рассматривался как функционально значимый белок-регулятор, контролирующий экспрессию фенотип-ассоциированных генов.

Результаты и обсуждение

Наша работа была ориентирована на идентификацию регуляторных белков, оказывающих влияние на гены, чья экспрессия различается между фенотипами макрофагов M2a, M2b, M2c и M2d. Понимание регуляторных механизмов, обусловливающих различия между фенотипами макрофагов, имеет не только фундаментальную значимость, но также перспективно для применения в медицине и фармакологии, так как для ряда патологий показано, что преобладание определенного фенотипа макрофагов связано с развитием и исходом этих заболеваний (Zhang M. et al., 2014; Yuan et al., 2017; Feng et al., 2021).

Общая характеристика фенотип-ассоциированных генов и белков макрофагов M1, M2a, M2b, M2c, M2d

Краткая информация по сформированной нами курируемой базе данных MACRO_GENES фенотип-ассоциированных генов, кодирующих фенотип-ассоциированные белки, т.е. белки, специфичные для макрофагов каждого из рассматриваемых фенотипов: M1, M2a, M2b, M2c, M2d, – представлена в табл. 1. Присутствие белков в определенных фенотипах подтверждено экспериментальными данными, приведенными в соответствующих публикациях. Детальное описание информации о генах в MACRO_GENES дано в табл. S1.

Некоторые из фенотип-ассоциированных белков используются в экспериментальных работах как специфические маркеры для разделения фенотипов макрофагов. Гены, кодирующие такие белки, отмечены зеленым цветом

Таблица 1. Списки генов, кодирующих фенотип-ассоциированные белки макрофагов M1, M2a, M2b, M2c и M2d, представленные в базе данных MACRO_GENES

Фенотип макрофагов	Гены, кодирующие фенотип-ассоциированные белки*
M1	*ARGI2, CAHM6, CCL2, CCL3, CCL4, CCL5, CCL7, CCL8, CCL15, CCL19, CCL20, CCR2, CCR7, CD38, CD80, CD86, CSF2, CXCL2, CXCL5, CXCL9, CXCL10, CXL11, FCG2B, FCG3A, FCGR1, GBP2, GBP5, GROA, HLA-DRA, HMGB1, IDO1, IFNA1, IFNB, IFNG, IL1A, IL1B, IL1R1, IL2RA, IL3, IL6, IL7RA, IL8, IL12B, IL15, IL15RA, IL17, IL18, IL23A, IRF1, IRF4, IRF7, ISG20, ITGAX, KCNN2, LAG3, MARCO, MET, MIF, MMP13, NAMPT, NFKB1, NOS2, PGH2, SOCS3, STAT1, TIMP1, TLR2, TLR4, TNFA, TNR5, TSP1, UBD, VEGFA*
M2a	*ALOX15, ARG1, CCL1, CCL2, CCL7, CCL8, CCL13, CCL14, CCL15, CCL17, CCL18, CCL22, CCL23, CCL24, CCL26, CCR2, CD200R1, CD209, CD274, CDH1, CDK11B, CLEC4A, CLEC7A, CLEC10A, CSF1R, CXCR1, CXCR2, EDN1, EGR2, FCER2, FGF2, FLT1, FN1, HAVCR2, HLA-DPA1, HLA-DPB1, HLA-DRA, HLA-DRB1, HLA-DRB3, HRH1, IGF1, IL1R1, IL1R2, IL4, IL10, IL27, IRF4, ITGAX, KLF2, LGALS3, MMP9, MMP14, MRC1, MRC2, MYC, PDCD1LG2, PGF, PPARG, PTGS1, RAMP1, SCN3A, SOCS1, TGFB1, TGM2, TREM2, VEGFA*
M2b	*CCL1, CCL4, CCL20, CD86, CD163, COX2, CXCL3, HLA-DRA, IFNA1, IL1B, IL1R1, IL6, IL10, MRC1, NOS2, PTPRC, SIGLEC10, SPHK1, TNFA, TNFSF14, VEGFA*
M2c	*ARG1, BCL3, C1QA, CCL8, CCL16, CCL18, CCL23, CCR2, CCR3, CCR10, CD14, CD163, CD300E, CDK11A, CDK11B, CSF1R, CX3CR1, CXCL12, CXCL13, CXCR4, EPAS1, F5, FCRLA, FPR1, GAS6, GXYLT2, HIF1A, IL1B, IL1R1, IL4R, IL10, IL21R, IRF3, IRF5, IRF8, ITGAX, JAK3, LIN7A, MAF, MARCO, MCTP2, MERTK, MMP2, MMP8, MMP14, MRC1, MRC2, MSR1, NOS2, PCOLCE2, PGF, PLOD2, SELENOP, SERPINA1, SH3PXD2B, SLAMF1, SOCS3, SPP1, SRPX2, STAT1, STAT3, STAT6, TGFB1, THBS1, TIMP1, TLR1, TLR2, TLR4, TLR5, TLR8, VCAN, VTCN1*
M2d	*ADORA2A, AIF1, C1QA, C1QC, CCL2, CCL3, CCL4, CCL5, CCL7, CD81, CD274, COX2, CSF3R, IL8, CXCL9, CXCL10, CXCL16, EGF, FCRL2, FGF2, FGFR1, GDF15, HLA-DMA, ID3, IDO1, IDO2, IL1A, IL1B, IL6, IL10, IRF7, LILRB4, MIF, MMP2, MMP9, MRC1, MSR1, NCAM1, NOS2, PDCD1LG2, PDGFB, TBX6, TGFBI, TNFA, VEGFA*

* Зеленым цветом отмечены гены, кодирующие белки – маркеры макрофагов различных фенотипов. Оранжевым обозначены гены, которые, по одним источникам, экспрессируются в макрофагах какого-либо фенотипа, но при этом, по другим данным, являются маркерами макрофагов другого фенотипа.

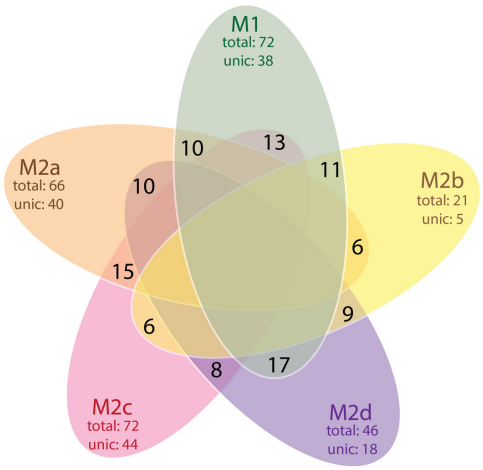


Рис. 2. Диаграмма Венна для сравнения фенотипов макрофагов M1, M2a, M2b, M2c, M2d по спискам генов, представленным в базе MACRO_GENES. Цифрами обозначено количество совпадающих белков.

(см. табл. 1). Если белок характерен для конкретного фенотипа, но при этом считается также специфическим маркером другого фенотипа, то кодирующий его ген обозначен оранжевым цветом (см. табл. 1). Например, белок CCL2 считается маркером фенотипа M1, но в некоторых публикациях указано, что он характерен и для фенотипов M2a и M2d. Можно видеть сложную картину пересечений маркеров между разными фенотипами макрофагов (см. табл. 1).

Диаграмма Венна, построенная на основе информации, приведенной в табл. 1, демонстрирует картину распределения генов, кодирующих фенотип-ассоциированные белки по пяти типам макрофагов: M1, M2a, M2b, M2c, M2d (рис. 2). Отметим, что больше всего совпадающих белков (17) имеют фенотипы M1 и M2d. Пары фенотипов M2a/M2c и M1/M2c имеют 15 и 13 общих белков соответственно. Фенотипы M2b и M1 имеют 11 совпадающих белков. Относительно малое число совпадающих белков (8) можно отметить для пары фенотипов M2c и M2d. Меньше всего пересечений (по 6 белкам) у M2b с M2a.

Общая характеристика дифференциально экспрессируемых белков группы макрофагов M2

Для поиска белков-посредников, передающих сигналы от индукторов поляризации макрофагов M2a, M2b, M2c и M2d к фенотип-ассоциированным генам, мы использовали списки дифференциально экспрессируемых, по сравнению с неактивированными макрофагами, белков из работы (Li et al., 2022). Из опубликованных данных, суммированных нами в табл. S2 и S3, следует, что распределение дифференциально экспрессируемых белков по четырем указанным фенотипам макрофагов характеризуется существенным пересечением, т. е. однозначное соответствие между дифференциально экспрессируемыми белками и фенотипами макрофагов отсутствует. Поэтому для поиска регуляторных связей, обеспечивающих поляризацию макрофагов в разные фенотипы, потребовался биоинформатический анализ больших объемов молекулярно-генетических данных, проведенный с использованием компьютерной системы ANDSystem.

Поиск регуляторных связей от дифференциально экспрессируемых белков к фенотип-ассоциированным генам макрофагов на основе фреймовых моделей

Для анализа больших объемов опубликованных данных о разных фенотипах макрофагов применяли методы и информационные ресурсы компьютерной инженерии знаний, реализованные в системе ANDSystem. С использованием фреймового подхода, реализованного в этой системе, был проведен поиск регуляторных связей между дифференциально экспрессируемыми белками и фенотип-ассоциированным генами макрофагов.

Были реконструированы графы регуляторных процессов, в которых вершины соответствовали дифференциально экспрессируемым белкам из статьи (Li et al., 2022) и фенотип-ассоциированным генам макрофагов из базы данных MACRO_GENES. Ребра, соединяющие вершины графа, соответствовали регуляторным связям между ними. В качестве примера показан граф потенциальных регуляторных связей между дифференциально экспрессируемыми белками и фенотип-ассоциированными генами макрофагов в фенотипе M2b (рис. 3).

Большинство фенотип-ассоциированных генов регулируются не одним, а несколькими белками (рис. 3). Вместе с тем большинство белков, представленных на рисунке, участвует в регуляции нескольких генов. Аналогичные схемы регуляторных связей для макрофагов M2a, M2c и M2d представлены на рис. S1–S3.

Выявление статистически значимых регуляторов фенотип-ассоциированных генов

Количественные характеристики регуляторных связей между дифференциально экспрессируемыми белками и фенотип-ассоциированными генами, выявленных с использованием фреймовых моделей, показаны в табл. 2, А.

На втором этапе анализа оценивали показатели центральности, характеризующие функциональную значимость дифференциально экспрессируемых белков для регуляции фенотип-ассоциированных генов. Оценки центральности позволили выявить белки-регуляторы фенотип-ассоциированных генов с порогом статистической значимости с поправкой Бонферрони $p < 0.05$ (см. табл. 2, Б). Учет статистической значимости дифференциально экспрессируемых белков на основе показателей центральности привел к существенному уменьшению количества вершин, соответствующих фенотип-ассоциированным генам, и количества ребер, соответствующих регуляторным событиям. Например, для фенотипа M2a из 40 дифференциально экспрессируемых белков, которые связаны с фенотип-ассоциированными генами, 16 оказались статистически значимо связаны с этими генами (см. табл. 2, Б). Для фенотипа M2b – 6 из 12 белков. Соответственно уменьшилось количество регуляторных событий и фенотип-ассоциированных генов в реконструированных графах.

На рисунке 4 представлены перечни дифференциально экспрессируемых белков, статистически значимо связанных с фенотип-ассоциированными генами макрофагов M2a, M2b, M2c и M2d. Красным цветом отмечены белки, уровень которых, по данным (Li et al., 2022), повышен в конкретных фенотипах макрофагов. Синим цветом обозначены те белки, уровень которых снижен по сравнению с

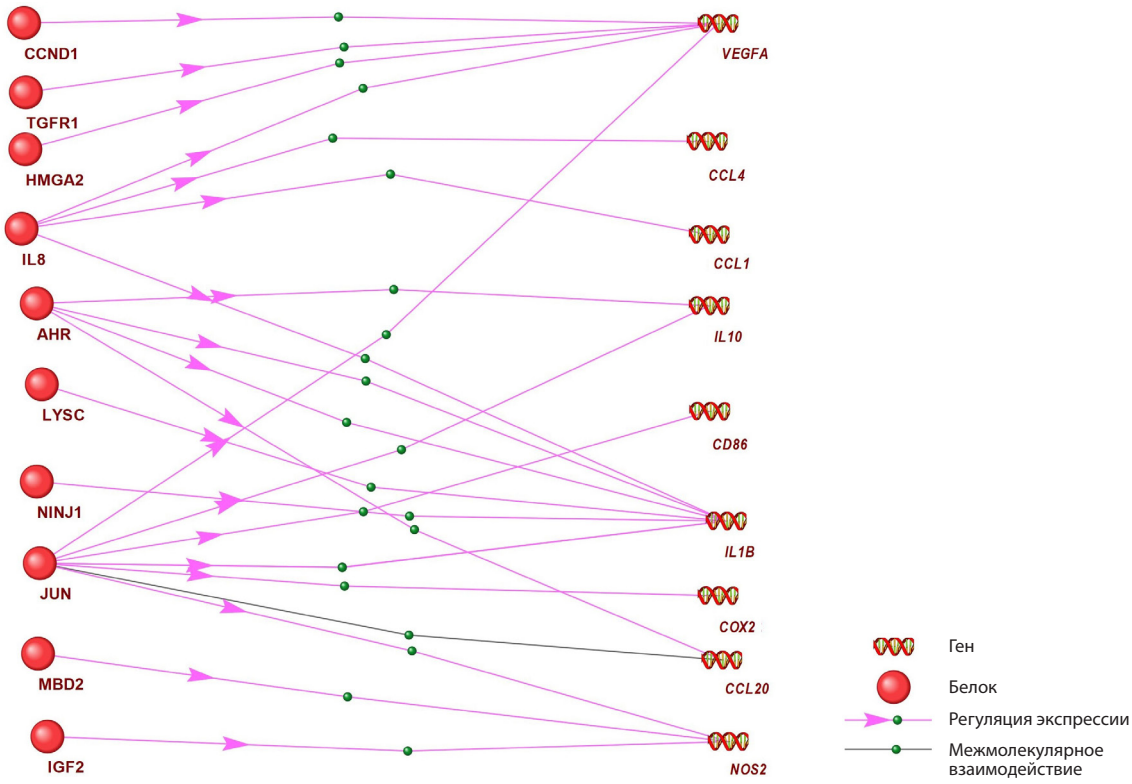


Рис. 3. Граф потенциальных регуляторных связей между дифференциально экспрессируемыми белками (слева) и фенотип-ассоциированными генами (справа) макрофагов фенотипа M2b, представленный в интерфейсе системы ANDSystem. Зеленые шарики на стрелках в интерактивном варианте работы с ANDSystem позволяют пользователям получать дополнительную информацию о конкретных регуляторных связях.

Таблица 2. Количественные характеристики потенциальных регуляторных связей, выявленных на основе коротких фреймовых моделей

Компоненты фреймовых моделей	M2a	M2b	M2c	M2d
А. Результаты первого этапа анализа реконструированного графа регуляторных взаимодействий				
Дифференциально экспрессируемые белки	40	12	41	43
Регуляторные события	127	51	216	252
Фенотип-ассоциированные гены	26	12	41	31
Б. Результаты второго этапа анализа (с учетом статистических оценок центральности регуляторных белков ($p < 0.05$)				
Дифференциально экспрессируемые белки	16	6	10	15
Регуляторные события	85	19	89	133
Фенотип-ассоциированные гены	23	8	28	29

неактивированными макрофагами. Зеленые линии соединяют белки с разнонаправленными изменениями уровней экспрессии в макрофагах разных фенотипов.

На рисунке 5 показаны примеры схем статистически значимых регуляторных взаимодействий между дифференциально экспрессируемыми белками макрофагов фенотипа M2a и фенотип-ассоциированными генами, реконструированные с использованием ANDSystem. Соответствующие схемы для макрофагов фенотипов M2b, M2c и M2d представлены на рис. S4–S6. Рисунок 5 демонстрирует регуляторные связи, использующие два альтернативных варианта контроля молекулярно-генетических процессов в одном и том же фенотипе M2a макрофагов: через повышение (см. рис. 5, *А*) и понижение (см. рис. 5, *Б*) уровня регуляторных белков. Описание реконструированных связей, полученных с помощью фреймовых моделей, приведено в табл. 3.

Для примера рассмотрим связи регуляторного белка TGM2 (нижний белок на рис. 5, *А*, отмеченный синей звездочкой). Протеомные данные (Li et al., 2022) показывают,

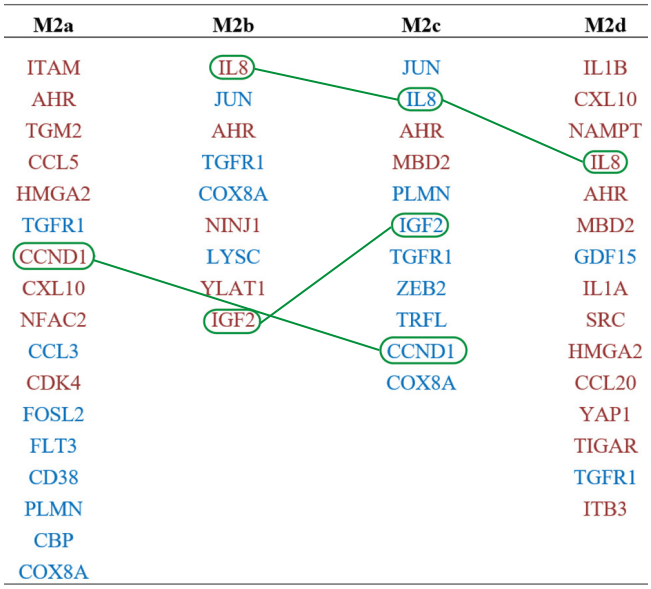


Рис. 4. Дифференциально экспрессируемые белки, статистически значимо ($p < 0.05$) связанные с фенотип-ассоциированными генами макрофагов M2a, M2b, M2c и M2d. Красным цветом отмечены белки, уровень которых согласно (Li et al., 2022) повышен в конкретном фенотипе, синим – белки, уровень которых снижен по сравнению с неактивированными макрофагами. Зеленые линии соединяют белки с разнонаправленными изменениями уровней экспрессии в макрофагах разных фенотипов.

что уровень этого белка в фенотипе M2a повышен по сравнению с неактивированными макрофагами. Согласно информации из базы знаний ANDSystem, полученной через интерфейс этой системы, в макрофагах фенотипа M2a белок TGM2 оказывает активирующее влияние на экспрессию фенотип-ассоциированных с M2a генов: *CD274* и *FN1* (Liu et al., 2021; Sun et al., 2021), что согласуется с данными, представленными в табл. 1 (эти гены отмечены на рис. 5, *А* черными звездочками). TGM2 оказывает также подавляющее влияние на ген *PPARG* (Maiuri et al., 2008), что не согласуется с данными табл. 1 и указывает на то, что экспрессия этого гена активируется также другими факторами, например AHR, CDK4, CCL5, чей уровень в этом фенотипе повышен (см. рис. 5, *А*).

Из группы белков с пониженным уровнем, по сравнению с неактивированными макрофагами, в макрофагах фенотипа M2a в качестве примера рассмотрим белок CBP, регулирующий фенотип-ассоциированные гены *CCL2*, *CD274*, *CDH1* (на рис. 5, *Б* отмечены звездочками). Согласно информации из базы знаний ANDSystem, когда CBP в фенотипе M2a подавлен, экспрессия *CCL2* повышается (Huang et al., 2021), что согласуется с данными, приведенными в табл. 1 (база данных MACRO_GENES). Вместе с тем тот же белок CBP положительно влияет на экспрессию фенотип-ассоциированных генов *CD274* и *CDH1* (Liu et al., 2020; Heng et al., 2021). Можно предположить, что на активность этих генов большее влияние оказывают другие регуляторы. Такими регуляторами для гена *CD274* могут быть белки AHR, CCL5, TGM2, CDK4, чей уровень в фенотипе M2a повышен (см. рис. 5, *А*, двойные зеленые звездочки).

Все выявленные статистически значимые регуляторные взаимодействия в макрофагах группы M2 между дифференциально экспрессируемыми белками и фенотип-ассоциированными генами представлены в табл. 3. Для макрофагов группы M2a – это взаимодействия девяти регуляторных белков с повышенным уровнем (по сравнению с неактивированными макрофагами), отмеченные стрелками (↑), и четырех белков с пониженным уровнем (↓),

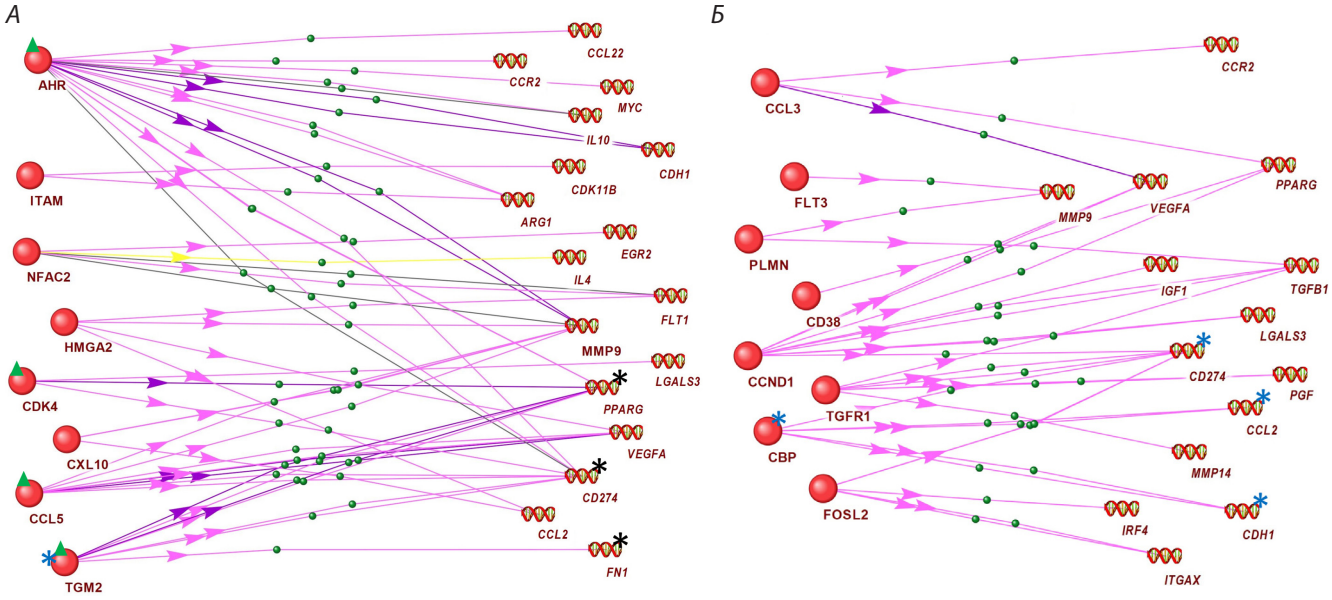


Рис. 5. Граф регуляции экспрессии фенотип-ассоциированных генов (из базы данных MACRO_GENES) в макрофагах фенотипа M2a дифференциально экспрессируемыми белками-регуляторами из статьи (Li et al., 2022), статистически значимо связанными с этими генами: через повышение (*А*) и понижение (*Б*) уровня представленности регуляторных белков в этом фенотипе. Синей звездочкой на *А* отмечен обсуждаемый в тексте белок TGM2, черными звездочками – его гены-мишени; зелеными треугольниками отмечены обсуждаемые белки AHR, CDK4, CCL2, TGM2. Синими звездочками на *Б* обозначены обсуждаемый в тексте белок CBP и его гены-мишени.

Таблица 3. Взаимосвязи между функционально значимыми дифференциально экспрессируемыми регуляторными белками* и регулируемые ими фенотип-ассоциированные гены в макрофагах фенотипов группы М2, выявленные на основе фреймовых моделей

M2a		M2b		M2c		M2d	
Белок	Гены-мишени	Белок	Гены-мишени	Белок	Гены-мишени	Белок	Гены-мишени
AHR (↑)**	*ARG1, MMP9, MYC, PPARG, CCL22, CCR2, CDH1, CD274, IL10*	AHR (↑)	*IL1B, CCL20, IL10*	AHR (↑)	*ARG1, CXCL12, STAT3, SOCS3, CCR2, HIF1A, TLR2, IL1B, IL10*	AHR (↑)	*IDO1, IDO2, IL1B, IL8,MMP9, NCAM1, CD274, IL1A, IL10*
TGM2 (↑)	*PPARG, CD274, FN1, MMP14*	IGF2 (↑)	*NOS2*	MBD2 (↑)	*CXCL12, NOS2, SOCS3*	IL1B (↑)	*FGF2, FGFR1, CXCL10, IL8, MMP9, VEGFA, TGFB1, CCL2, CCL3, CCL4, CCL5, NOS2, COX2, IL10, IL1A, TNF, PDCD1LG2, MIF*
CCL5 (↑)	*VEGFA, CD274, MMP9*	NINJ1 (↑)	*IL1B*				
NFAC2 (↑)	*EGR2, IL4, FLT1, MMP9*	IL8 (↑)	*CCL1, VEGFA*	TGFR1 (↓)	*MMP14, TGFB1, HIF1A, PGF*	MBD2 (↑)	*CXCL9, CXCL10, CD274, CCL5, MMP2, NOS2*
HMGA2 (↑)	*FLT1, MMP9, VEGFA, CCL2*	TGFR1 (↓)	*VEGFA*	ZEB2 (↓)	*MAF, MMP14, TGFB1,*	IL8 (↑)	*IL1B, FGF2, CXCL10, MMP9, CCL2, CD274, MMP2, VEGFA*
ITAM (↑)	*ARG1, CDK11B*	JUN (↓)	*CD86, COX2, NOS2, IL1B, IL10, CCL20, VEGFA*	CCND1 (↓)	*TGFB1, STAT3, HIF1A*	YAP1 (↑)	*IRF7, FGFR1, PDGFB, CD274*
CDK4 (↑)	*PPARG, LGALS3, CD274*			IL8 (↓)	*MMP14, MMP2, MMP8, TLR5, CXCL12, ARG1, CCR2, HIF1A, IL1B, CDK11B*	NAMPT (↑)	*IL1B, FGF2, IL8, MMP9, CCL2, CD274, MMP2*
CXL10 (↑)	*MMP9, CCL2*						
CCND1 (↓)	*VEGFA, TGFB1, CD274*			IGF2 (↓)	*CDK11B, MMP2, HIF1A, NOS2*	IL1A (↑)	*IL6, IL1B, VEGFA, CD274*
						CCL20 (↑)	*MMP9, CD274*
FLT3 (↓) ***	*MMP9*			PLMN (↓)	*TGFB1, CXCL12, CXCR4*	CXL10 (↑)	*IL1B, MMP9, CCL2, MMP2*
CCL3 (↓)	*CCR2, VEGFA, PPARG*			TRFL (↓)	*HIF1A, TLR4*	TIGAR (↑)	*MMP9, MMP2*
PLMN (↓)	*MMP9, TGFB1*			JUN (↓)	*TGFB1, CXCL12, MMP14, MMP2, ARG1, IL1B, NOS2, HIF1A, TLR2, TLR4, IL10, CX3CR1, THBS1, SERPINA1, BCL3, CCL18, ITGAX*	HMGA2 (↑)	*MMP9, CCL2, MMP2, VEFGA*
CD38 (↓)	*VEGFA*					SRC (↑)	*CXCL10, MMP2, VEFGA, NOS2*
TGFR1 (↓)	*TGFB1, CD274, MMP14, PGF*					ITB3 (↑)	*MMP2*
CBP (↓)	*CD274, CCL2, CDH1*					TGFR1 (↓)	*VEGFA, TGFB1, CD274*
FOSL2 (↓)	*CD274, IRF4, ITGAX*					GDF15 (↓)	*MMP9, CCL2, GDF15, CD274, MMP2*

* Белки, отобранные по критерию центральности с $p < 0.05$; ** ↑ – белки с повышенным уровнем экспрессии; *** ↓ – белки с пониженным уровнем экспрессии.

регулирующие 23 фенотип-ассоциированных гена. Для M2b это 4 белка с повышенным и 2 белка с пониженным уровнем, регулирующие 8 фенотип-ассоциированных генов (см. также рис. S4). Для M2c выявлено 2 белка с повышенным и 8 с пониженным уровнем, регулирующие 28 генов (см. также рис. S5). Для M2d найдено 13 белков с повышенным и 2 белка с пониженным уровнем, регулирующих 29 генов (см. также рис. S6).

Таким образом, на основе компьютерного анализа различий в протеомах разных фенотипов макрофагов, а также использования больших объемов информации, накопленной в базе знаний системы ANDSystem, выявлены некоторые регуляторные белки, опосредующие действие индукторов поляризации макрофагов к фенотип-ассоциированным генам макрофагов. В дальнейшем планируется продолжение исследования с помощью фреймовых моделей, содержащих больше слотов, отражающих промежуточные этапы действия индукторов поляризации макрофагов на фенотип-ассоциированные гены макрофагов. Это даст возможность идентификации более тонких особенностей организации регуляторных связей, идущих от индукторов поляризации макрофагов к фенотип-ассоциированным генам через последовательность действий белков-посредников.

Заключение

Изучение опубликованных данных о фенотип-ассоциированных генах и протеомах макрофагов группы М2 и последующий поиск регуляторных связей между ними с использованием фреймового подхода, реализованного в компьютерной системе ANDSystem, дали возможность выявить потенциальные регуляторные белки, которые обеспечивают различия в экспрессии генов в М2 фенотипах макрофагов. Полученные результаты предполагают, что отличия между фенотипами M2a, M2b, M2c и M2d могут быть связаны, в частности, с регуляторными функциями белков JUN, IL8, NFAC2, CCND1 и YAP1, уровень которых варьирует между фенотипами, что приводит к различиям в экспрессии фенотип-ассоциированных генов.

Список литературы / References

Иванисенко В.А., Басов Н.В., Макарова А.А., Вензель А.С., Рогачев А.Д., Деменков П.С., Иванисенко Т.В., Клещев М.А., Гайслер Е.В., Мороз Г.Б., Плеско В.В., Сотникова Ю.С., Патрушев Ю.В., Ломиворотов В.В., Колчанов Н.А., Покровский А.Г. Применение генных сетей к анализу результатов метаболомного скрининга плазмы крови пациентов с послеоперационным делирием. *Вавиловский журнал генетики и селекции.* 2023; 27(7):768-775. doi 10.18699/VJGB-23-89
[Ivanisenko V.A., Basov N.V., Makarova A.A., Venzel A.S., Rogachev A.D., Demenkov P.S., Ivanisenko T.V., Kleshchev M.A., Gaisler E.V., Moroz G.B., Plesko V.V., Sotnikova Y.S., Patrushev Y.V., Lomivorotov V.V., Kolchanov N.A., Pokrovsky A.G. Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2023;27(7):768-775. doi 10.18699/VJGB-23-89]

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Volianskaia A.R., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Kolchanov N.A., Lavrik I.N., Ivanisenko V.A. Reconstruction of the regulatory hypermethylation network controlling hepatocellular carcinoma development during hepatitis C viral infection. *J Integr Bioinform*. 2023;20(3): 20230013. doi 10.1515/jib-2023-0013

Chhor V., Le Charpentier T., Lebon S., Oré M.V., Celador I.L., Josserand J., Degos V., Jacotot E., Hagberg H., Sävman K., Mallard C., Gressens P., Fleiss B. Characterization of phenotype markers and neuronotoxic potential of polarised primary microglia *in vitro*. *Brain Behav Immun*. 2013;32:70-85. doi 10.1016/j.bbi.2013.02.005

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Feng D., Huang W.Y., Niu X.L., Hao S., Zhang L.N., Hu Y.J. Significance of macrophage subtypes in the peripheral blood of children with systemic juvenile idiopathic arthritis. *Rheumatol Ther*. 2021; 8(4):1859-1870. doi 10.1007/s40744-021-00385-x

Ghasemi M., Seidkhani H., Tamimi F., Rahgozar M., Masoudi-Nejad A. Centrality Measures in Biological Networks. *Curr Bioinform*. 2014;9:426-441. doi 10.2174/1574893613086660013

Gurvich O.L., Puttonen K.A., Bailey A., Kailaanmäki A., Skirdenko V., Sivonen M., Pietikäinen S., Parker N.R., Ylä-Herttuala S., Kekarainen T. Transcriptomics uncovers substantial variability associated with alterations in manufacturing processes of macrophage cell therapy products. *Sci Rep*. 2020;10(1):14049. doi 10.1038/s41598-020-70967-2

Heng W.S., Kruyt F.A.E., Cheah S.C. Understanding lung carcinogenesis from a morphostatic perspective: prevention and therapeutic potential of phytochemicals for targeting cancer stem cells. *Int J Mol Sci*. 2021;22(11):5697. doi 10.3390/ijms22115697

Hirani D., Alvira C.M., Danopoulos S., Milla C., Donato M., Tian L., Mohr J., … Seeger W., Khatri P., Al Alam D., Dötsch J., Alejandre Alcazar M.A. Macrophage-derived IL-6 trans-signaling as a novel target in the pathogenesis of bronchopulmonary dysplasia. *Eur Respir J*. 2021;59(2):2002248. doi 10.1183/13993003.02248-2020

Huang Y.H., Cai K., Xu P.P., Wang L., Huang C.X., Fang Y., Cheng S., Sun X.J., Liu F., Huang J.Y., Ji M.M., Zhao W.L. *CREBBP/EP300* mutations promoted tumor progression in diffuse large B-cell lymphoma through altering tumor-associated macrophage polarization via FBXW7-NOTCH-CCL2/CSF1 axis. *Signal Transduct Target Ther*. 2021;6(1):10. doi 10.1038/s41392-020-00437-8

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Syst Biol*. 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019; 20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Jalili M., Salehzadeh-Yazdi A., Gupta S., Wolkenhauer O., Yaghmaie M., Resendis-Antonio O., Alimoghaddam K. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Front Physiol*. 2016;7:375. doi 10.3389/fphys.2016.00375

Joerink M., Rindsjö E., van Riel B., Alm J., Papadogiannakis N. Placental macrophage (Hofbauer cell) polarization is independent of maternal allergen-sensitization and presence of chorioamnionitis. *Placenta*. 2011;32(5):380-385. doi 10.1016/j.placenta.2011.02.003

Lampiasi N. Macrophage polarization: learning to manage it 2.0. *Int J Mol Sci*. 2023;24(24):17409. doi 10.3390/ijms242417409

Lee J.T., Pamir N., Liu N.C., Kirk E.A., Averill M.M., Becker L., Larson I., Hagman D.K., Foster-Schubert K.E., van Yserloo B., Bornfeldt K.E., LeBoeuf R.C., Kratz M., Heinecke J.W. Macrophage metalloelastase (MMP12) regulates adipose tissue expansion, insulin sensitivity, and expression of inducible nitric oxide synthase. *Endocrinology*. 2014;155(9):3409-3420. doi 10.1210/en.2014-1037

Li P., Ma C., Li J., You S., Dang L., Wu J., Hao Z., Li J., Zhi Y., Chen L., Sun S. Proteomic characterization of four subtypes of M2 macrophages derived from human THP-1 cells. *J Zhejiang Univ Sci B*. 2022;23(5):407-422. doi 10.1631/jzus.B2100930

Liu J., He D., Cheng L., Huang C., Zhang Y., Rao X., Kong Y., … Jones K., Napier D., Lee E.Y., Wang C., Liu X. p300/CBP inhibition enhances the efficacy of programmed death-ligand 1 blockade treatment in prostate cancer. *Oncogene*. 2020;39(19):3939-3951. doi 10.1038/s41388-020-1270-z

Liu J., Liu Q., Zhang X., Cui M., Li T., Zhang Y., Liao Q. Immune subtyping for pancreatic cancer with implication in clinical outcomes and improving immunotherapy. *Cancer Cell Int*. 2021;21(1):137. doi 10.1186/s12935-021-01824-z

Maiuri L., Luciani A., Giardino I., Raia V., Villella V.R., D'Apolito M., Pettoello-Mantovani M., Guido S., Ciacci C., Cimmino M., Cexus O.N., Londei M., Quarantino S. Tissue transglutaminase activation modulates inflammation in cystic fibrosis via PPARgamma down-regulation. *J Immunol*. 2008;180(11):7697-7705. doi 10.4049/jimmunol.180.11.7697

Martinez F.O., Gordon S., Locati M., Mantovani A. Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression.

J Immunol. 2006;177(10):7303-7311. doi 10.4049/jimmunol.177.10.7303

Martinez F.O., Sica A., Mantovani A., Locati M. Macrophage activation and polarization. *Front Biosci*. 2008;13:453-461. doi 10.2741/2692

Mills C.D. M1 and M2 macrophages: oracles of health and disease. *Crit Rev Immunol*. 2012;32(6):463-488. doi 10.1615/critrevimmunol.v32.i6.10

Mosser D.M., Hamidzadeh K., Goncalves R. Macrophages and the maintenance of homeostasis. *Cell Mol Immunol*. 2021;18(3):579-587. doi 10.1038/s41423-020-00541-3

Murray P.J., Allen J.E., Biswas S.K., Fisher E.A., Gilroy D.W., Goerdt S., Gordon S., … Suttles J., Udalova I., van Ginderachter J.A., Vogel S.N., Wynn T.A. Macrophage activation and polarization: nomenclature and experimental guidelines. *Immunity*. 2014;41(1):14-20. doi 10.1016/j.immuni.2014.06.008

Oates T.C., Moura P.L., Cross S., Roberts K., Baum H.E., Haydn-Smith K.L., Wilson M.C., Heesom K.J., Severn C.E., Toye A.M. Defining the proteomic landscape of cultured macrophages and their polarization continuum. *Immunol Cell Biol*. 2023;101(10):947-963. doi 10.1111/imcb.12687

Sun W., Qin Y., Wang Z., Dong W., He L., Zhang T., Zhang H. The *NEAT1_2*/miR-491 axis modulates papillary thyroid cancer invasion and metastasis through TGM2/NFκb/FN1 signaling. *Front Oncol*. 2021;11:610547. doi 10.3389/fonc.2021.610547

van der Lans A.A., Boon M.R., Haks M.C., Quinten E., Schaart G., Ottenhoff T.H., van Marken Lichtenbelt W.D. Cold acclimation affects immune composition in skeletal muscle of healthy lean subjects. *Physiol Rep*. 2015;3(7):e12394. doi 10.14814/phy2.12394

Wynn T.A., Vannella K.M. Macrophages in tissue repair, regeneration, and fibrosis. *Immunity*. 2016;44(3):450-462. doi 10.1016/j.immuni.2016.02.015

Xu W., Zhao X., Daha M.R., van Kooten C. Reversible differentiation of pro- and anti-inflammatory macrophages. *Mol Immunol*. 2013;53(3):179-86. doi 10.1016/j.molimm.2012.07.005

Yuan R., Li S., Geng H., Wang X., Guan Q., Li X., Ren C., Yuan X. Reversing the polarization of tumor-associated macrophages inhibits tumor metastasis. *Int Immunopharmacol*. 2017;49:30-37. doi 10.1016/j.intimp.2017.05.014

Zhang M., Wang C.C. Inflammatory response of macrophages in infection. *Hepatobiliary Pancreat Dis Int*. 2014;13(2):138-152. doi 10.1016/s1499-3872(14)60024-2

Zhang M., He Y., Sun X., Li Q., Wang W., Zhao A., Di W. A high M1/M2 ratio of tumor-associated macrophages is associated with extended survival in ovarian cancer patients. *J Ovarian Res*. 2014;7:19. doi 10.1186/1757-2215-7-19

Zhang Q., Sioud M. Tumor-associated macrophage subsets: shaping polarization and targeting. *Int J Mol Sci*. 2023;24:7493. doi 10.3390/ijms24087493

Zolotareva O., Saik O.V., Königs C., Bragina E.Y., Goncharova I.A., Freidin M.B., Dosenko V.E., Ivanisenko V.A., Hofestädt R. Comorbidity of asthma and hypertension may be mediated by shared genetic dysregulation and drug side effects. *Sci Rep.* 2019;9(1):16302. doi 10.1038/s41598-019-52762-w

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 09.08.2025. После доработки 16.10.2025. Принята к публикации 17.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1000-1008

doi 10.18699/vjgb-25-105

Компьютерная реконструкция генной сети цитокиновой регуляции генов и белков, ассоциированных с РАС

Н.М. Леванова[1], Е.Г. Вергунов[1, 2, 3], А.Н. Савостьянов[1, 2, 3], И.В. Яцык[1], В.А. Иванисенко[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
levanova@bionet.nsc.ru

Аннотация. Многочисленные исследования подтверждают связь нарушений цитокиновой регуляции с развитием расстройств аутистического спектра (РАС) на уровне генов, белков и их межмолекулярных взаимодействий. В работе эти данные были систематизированы с применением биоинформатического анализа и методов машинного обучения. Главным инструментом в исследовании являлась когнитивная система ANDSystem, разработанная в Институте цитологии и генетики СО РАН и задействующая методы искусственного интеллекта для автоматического извлечения информации из биомедицинских баз данных и текстов научных публикаций. С использованием ANDSystem была реконструирована ассоциативная генная сеть цитокиновой регуляции генов и белков, ассоциированных с РАС. В результате анализа удалось идентифицировать 110 цитокинов, которые, согласно воссозданной сети, регулируют активность, деградацию и транспорт 58 белков, вовлеченных в развитие РАС, а также экспрессию 91 гена, ассоциированного с этими расстройствами. Анализ перепредставленности биологических процессов Gene Ontology выявил статистически значимые ассоциации этих генов с процессами, связанными с развитием и работой центральной нервной системы. Анализ топологических характеристик сети и оценка функциональной значимости элементов сети через их ассоциацию с биологическими процессами Gene Ontology, связанными с РАС, позволили выделить 21 цитокин, оказывающий наибольшее влияние на элементы сети. Среди них наибольший приоритет имели восемь цитокинов (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α), которые занимали высокое положение по результатам всех использованных методик приоритизации. Кроме того, из 21 приоритетного цитокина выделяются восемь цитокинов (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17), которые являются мишенями препаратов, применяемых в качестве иммуносупрессантов и противоопухолевых средств. Установленная роль этих цитокинов в патогенезе РАС создает предпосылки для потенциального перепрофилирования препаратов, направленных на их ингибирование, для терапии расстройств аутистического спектра.

Ключевые слова: расстройства аутистического спектра (РАС); нарушения нейроразвития; цитокины; автоматический анализ текстов научных публикаций; патогенез РАС; терапия РАС; компьютерная реконструкция генных сетей

Для цитирования: Леванова Н.М., Вергунов Е.Г., Савостьянов А.Н., Яцык И.В., Иванисенко В.А. Компьютерная реконструкция генной сети цитокиновой регуляции генов и белков, ассоциированных с РАС. *Вавиловский журнал генетики и селекции*. 2025;29(7):1000-1008. doi 10.18699/vjgb-25-105

Финансирование. Исследование выполнено в рамках бюджетного проекта № FWNR-2022-0020 в Институте цитологии и генетики СО РАН.

In silico reconstruction of the gene network for cytokine regulation of ASD-associated genes and proteins

N.M. Levanova[1], E.G. Vergunov[1, 2, 3], A.N. Savostyanov[1, 2, 3], I.V. Yatsyk[1], V.A. Ivanisenko[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
levanova@bionet.nsc.ru

Abstract. Accumulated evidence links dysregulated cytokine signaling to the pathogenesis of autism spectrum disorder (ASD), implicating genes, proteins, and their intermolecular networks. This paper systematizes these findings using bioinformatics analysis and machine learning methods. The primary tool employed in the study was the ANDSystem cognitive platform, developed at the Institute of Cytology and Genetics, which utilizes artificial intelligence techniques for automated knowledge extraction from biomedical databases and scientific publications. Using ANDSystem, we reconstructed a gene network of cytokine-mediated regulation of autism spectrum disorder (ASD)-associated genes and proteins. The analysis identified 110 cytokines that regulate the activity, degradation,

© Леванова Н.М., Вергунов Е.Г., Савостьянов А.Н., Яцык И.В., Иванисенко В.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

and transport of 58 proteins involved in ASD pathogenesis, as well as the expression of 91 ASD-associated genes. Gene Ontology (GO) enrichment analysis revealed statistically significant associations of these genes with biological processes related to the development and function of the central nervous system. Furthermore, topological network analysis and functional significance assessment based on association with ASD-related GO biological processes allowed us to identify 21 cytokines exerting the strongest influence on the regulatory network. Among these, eight cytokines (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α) had the highest priority, ranking at the top across all employed metrics. Notably, eight of the 21 prioritized cytokines (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) are known targets of drugs currently used as immunosuppressants and antitumor agents. The pivotal role of these cytokines in ASD pathogenesis provides a rationale for potentially repurposing such inhibitory drugs for the treatment of autism spectrum disorders.

Key words: autism spectrum disorder (ASD); neurodevelopmental disorders; cytokines; automatic text analysis of scientific publications; ASD pathogenesis; ASD treatment; computer reconstruction of gene networks

For citation: Levanova N.M., Vergunov E.G., Savostyanov A.N., Yatsyk I.V., Ivanisenko V.A. *In silico* reconstruction of the gene network for cytokine regulation of ASD-associated genes and proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1000-1008. doi 10.18699/vjgb-25-105

Введение

Согласно DSM-5, диагностическому и статистическому руководству по психическим расстройствам (Diagnostic and Statistical Manual of Mental Disorders, Fifth edition), расстройства аутистического спектра (РАС) представляют собой группу вариантов нейроразвития с выраженной генетической предрасположенностью, диагностика которых основывается на поведенческих критериях. Главные диагностические критерии РАС – это постоянные затруднения людей в социальных коммуникациях и взаимодействиях, а также ограниченность и стереотипность поведения. По современным диагностическим стандартам, описанные симптомы РАС должны обязательно проявляться в период раннего развития пациентов. Их выраженность может варьировать, а у ряда пациентов компенсироваться за счет выученных стратегий поведения (American Psychiatric Association, 2013). Признаки РАС характеризуются значительным варьированием поведенческих фенотипов (Van der Zee, Derksen, 2021). Нейрофизиологические особенности, связанные с аутизмом, наблюдаются как у пациентов с диагнозом, так и в общей популяции (Harms et al., 2010; Tsai et al., 2013; Tseng et al., 2015).

Расстройства аутистического спектра подразделяются на идиопатическую и синдромальную формы. Идиопатическая форма характеризуется отсутствием четко идентифицируемых генетических полиморфизмов, тогда как синдромальная ассоциирована с определенными моногенными мутациями и сопровождается дополнительными симптомами (Ziats et al., 2021). Значительная часть случаев синдромального аутизма обусловлена мутациями в генах, кодирующих белки-компоненты сигнального пути mTOR, приводящими к гиперактивации этого пути (Ganesan et al., 2019). В целом, согласно результатам биоинформатического анализа, проведенного ранее с использованием информации из специализированной базы данных SFARI Gene (Abrahams et al., 2013), приблизительно 58 % генов, мутации в которых ассоциированы с РАС, прямо связаны с сигнальным путем mTOR (Trifonova et al., 2019). Белок mTOR, механистическая или типичная для млекопитающих мишень для рапамицина (mammalian or mechanistic target of rapamycin), – это серин/треонин киназа, которая является центральным компонентом двух мультибелковых комплексов: mTORC1 и mTORC2. mTORC1 чувствителен к рапамицину и реагирует на концентрацию питательных веществ и факторы роста, регулируя рост клеток и метаболизм. mTORC2, напротив, в значительной степени нечувствителен к рапамицину и активируется в ответ стресс и воздействие факторов роста, регулируя процессы выживания и пролиферации клеток (Ragupathi et al., 2024).

Сигнальный путь mTOR вовлечен в регуляцию клеточного и опухолевого роста (Onore et al., 2017), иммунной активности (Liu et al., 2015), а также в формирование памяти и пластичности нейронных связей (Hoeffer, Klann, 2010). Показано, что конститутивная гиперактивация пути mTOR вызывает блокаду аутофагии (McMahon et al., 2012) и нарушение нормального процесса синаптического прунинга (Tang et al., 2014).

Синаптический прунинг – это процесс снижения числа синапсов, происходящий на протяжении всей жизни, в ходе которого микроглия удаляет избыточные синаптические соединения. Он увеличивает нервную проводимость и позволяет перенаправлять ресурсы в значимые области, что обеспечивает эффективность обучения и формирования долговременной памяти (Navlakha et al., 2015). В результате нарушения этого процесса у пациентов с РАС наблюдается увеличение плотности дендритных шипиков и синаптических связей как в поверхностных, так и в глубоких слоях лобной, височной и теменной коры головного мозга (Hutsler, Zhang, 2010).

Микроглия, играющая главную роль в процессе синаптического прунинга, является частью иммунной системы и подвержена регуляторному влиянию цитокинов. Кроме того, микроглия – основной источник провоспалительных цитокинов в мозге, а значит, и ключевой медиатор нейровоспаления, который может индуцировать или модулировать широкий спектр клеточных реакций (Smith et al., 2012). Следы нейровоспаления, обусловленного классической активацией микроглии (M1), обнаружены в аутопсийных препаратах людей с РАС: увеличенные концентрации интерферона IFN-γ и цитокинов IL-1β, IL-6, IL-12p40, TNF-α и CCL2 были зарегистрированы в тканях мозга и в спинномозговой жидкости (Vargas et al., 2005; Li et al., 2009; Morgan et al., 2010).

В целом цитокины обеспечивают регуляцию, необходимую для нормального раннего развития мозга, синаптической пластичности и поддержания процессов гомеостаза

в мозге. Из-за серьезных сдвигов в цитокиновом профиле процессы развития, миграции и дифференцировки нейронов нарушаются, что в конечном итоге приводит к поведенческим отклонениям (Ashwood et al., 2011). Кроме того, показано, что у людей с РАС цитокиновый профиль – количество цитокинов, выделенных из плазмы и сыворотки крови, статистически отличается от такового у нейротипичных людей (Onore et al., 2017). Таким образом, для РАС характерен регуляторный дисбаланс иммунной системы, приводящий к продолжающемуся нейровоспалению.

Для того чтобы выявить и систематизировать регуляторные взаимосвязи между цитокинами, генами и белками, ассоциированными с РАС, была проведена реконструкция ассоциативных генных сетей с применением программных инструментов, основанных на технологиях искусственного интеллекта. В данном исследовании в качестве основного инструмента была использована когнитивная система ANDSystem (Ivanisenko V.A. et al., 2015), предназначенная для автоматизированного анализа научных публикаций и баз данных.

Целью работы стали реконструкция генной сети цитокиновой регуляции генов и белков, ассоциированных с РАС, и анализ полученной генной сети для идентификации цитокинов – перспективных мишеней для иммуномодулирующей терапии РАС.

Анализ реконструированной сети выявил 110 цитокинов, регулирующих активность, деградацию и транспорт 58 белков, а также экспрессию 91 гена, которые ассоциированы с РАС. Изучение перепредставленности биологических процессов Gene Ontology показало значимую ассоциацию этих генов с процессами развития и функционирования ЦНС. Среди 21 цитокина, оказывающего наибольшее влияние на построенную сеть, восемь (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) являются мишенями уже существующих иммуносупрессивных и противоопухолевых препаратов. Выявленная ключевая роль данных цитокинов в патогенезе РАС создает предпосылки для новых исследований, направленных на перепрофилирование ингибирующих их препаратов для терапии РАС.

Материалы и методы

На первом этапе была осуществлена компьютерная реконструкция сети взаимодействий цитокинов с белками и генами, ассоциированными с РАС (итоговая генная сеть). Для максимально полного охвата этих регуляторных взаимодействий сначала были реконструированы пять специализированных генных подсетей, отражающих различные пути воздействия цитокинов (табл. S1 Приложения)[1]. Затем эти подсети были интегрированы в единую итоговую генную сеть.

На втором этапе были проведены биоинформатический анализ структуры итоговой генной сети и оценка вовлеченности ее элементов в биологические процессы (согласно терминологии Gene Ontology), которые ассоциированы с РАС. Затем цитокины были приоритизированы по значимости на основе их потенциального влияния на гены и белки, ассоциированные с РАС.

Последний этап – выявление среди цитокинов, показавших наибольшую значимость в результате выполненного анализа, перспективных мишеней для иммуномодулирующей терапии РАС.

Этап 1. Выборка генов, ассоциированных с РАС (234 гена), была извлечена из базы данных SFARI Gene (Abrahams et al., 2013) (https://gene.sfari.org). В выборку вошли гены, отмеченные в этой базе данных как с высокой достоверностью ассоциированные с РАС (категория 1 в соответствии с внутренней системой оценки базы данных). Список генов цитокинов (186 генов) и список генов рецепторов цитокинов (114 генов) были составлены по данным, извлеченным из Human Protein Atlas (HPA) (https://www.proteinatlas.org/), масштабной базы данных, ориентированной на накопление информации о локализации белков в тканях, клетках и органах человека и их экспрессии (Uhlén et al., 2015).

Для реконструкции генных сетей применялось программное обеспечение ANDVisio (Demenkov et al., 2012), использующее для реконструкции и анализа структуры генных сетей данные из базы знаний когнитивной системы ANDSystem. Когнитивная система ANDSystem предназначена для автоматизированного анализа научных публикаций и баз данных. В системе применяются механизмы онтологического моделирования, анализа графов, обработки естественного языка (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2020, 2022, 2024).

Итоговая генная сеть была получена путем объединения подсетей, для реконструкции которых применялась утилита Pathway Wizard, являющаяся частью программного обеспечения ANDVisio и позволяющая осуществить в базе знаний ANDSystem поиск и визуализацию генных сетей, соответствующих заданным шаблонам-запросам. Для построения итоговой генной сети было реконструировано пять отдельных подсетей с использованием нескольких шаблонов-запросов (см. табл. S1), с последующей интеграцией в единый граф.

Этап 2. Анализ обогащенности итоговой генной сети биологическими процессами Gene Ontology (Ashburner et al., 2000) (база данных GO_BP) выполняли в программном инструменте DAVID (Huang et al., 2009; Sherman et al., 2022) (https://davidbioinformatics.nih.gov/). DAVID позволяет проводить функциональную аннотацию генов и оценивать статистическую значимость обогащения генной сети терминами GO с заданным порогом достоверности.

Для анализа структуры итоговой сети и ранжирования цитокинов по степени посредничества и степени вершин в сети задействовали статистические инструменты программы ANDVisio. Степень посредничества (betweenness centrality) вершины определяется числом кратчайших путей, проходящих через нее, в то время как степень вершины (degree centrality) определяется количеством ее непосредственных связей. Обе метрики являются мерой влияния узла на сеть в целом, более высокие значения свидетельствуют о повышенной значимости элемента. Дополнительно с помощью специально разработанного нами скрипта на Python 3.10 была проведена приорити-

[1] Табл. S1 Приложения см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx35.pdf

зация цитокинов на основе их представленности в биологических путях, ассоциированных с РАС.

Этап 3. С привлечением информации из баз данных DrugBank (Knox et al., 2024) (https://go.drugbank.com/) и GETdb (Zhang et al., 2024) (https://togodb.org/db/getdb) отобранные цитокины рассматривались в качестве кандидатов на роль мишеней для фармакологического воздействия.

Результаты реконструкции генных сетей и их анализа

Реконструкция взаимодействия цитокинов с белками и генами, ассоциированными с РАС

На первом этапе исследования с использованием программы Pathway Wizard были реконструированы пять подсетей (рис. 1). Построение подсетей осуществляли на основе двух наборов данных: списка генов, ассоциированных с расстройствами аутистического спектра (РАС), из базы данных SFARI (https://gene.sfari.org) и перечней цитокинов и их рецепторов из базы данных Human Protein Atlas (HPA) (https://www.proteinatlas.org/).

По завершении автоматической реконструкции были вручную просмотрены все полученные в реконструкции связи и элементы для сравнения их с исходными текстами статей, чтобы исключить ошибки из-за неверного извлечения информации.

В результате интеграции реконструированных подсетей была сформирована итоговая сеть взаимодействий цитокинов с белками и генами, ассоциированными с РАС (рис. 2). Сеть содержит 1112 вершин двух типов и 3679 взаимодействий различных видов между ними (табл. 1).

В ходе анализа было идентифицировано 110 цитокинов (см. рис. 2, I), которые регулируют белки и гены, ассоциированные с РАС, а также 58 белков (см. рис. 2, II) и 91 ген (см. рис. 2, III), вовлеченных в формирование РАС, которые регулируются цитокинами.

Анализ компонентов генной сети

С использованием программного инструмента DAVID был проведен анализ обогащенности итоговой генной сети набором генов, ассоциированных с РАС, экспрессия которых, согласно результатам реконструкции сети, регулируется цитокинами. По материалам базы Gene Ontology этот набор генов статистически значимо связан с биологическими процессами, которые являются специфичными для функционирования и развития нервной системы, такими как формирование дендритных шипиков, развитие гиппокампа, миграция нейронов и др. (табл. 2). Всего в результате анализа было выявлено 56 биологических процессов (FDR < 0.05, false discovery rate – вероятность ложноположительного исхода), в которые вовлечены гены из полученного набора (согласно базе Gene Ontology). В табл. 2 включены только специфичные биологические процессы с наименьшим значением FDR. Неспецифические процессы, такие как регуляция транскрипции, были опущены.

Приоритизация цитокинов

Для выявления цитокинов, оказывающих наиболее сильное влияние на регуляторную сеть, было проведено их ранжирование по трем ключевым критериям: по степени вершины, по степени посредничества и по представленности в биологических процессах, вовлеченных в патогенез РАС.

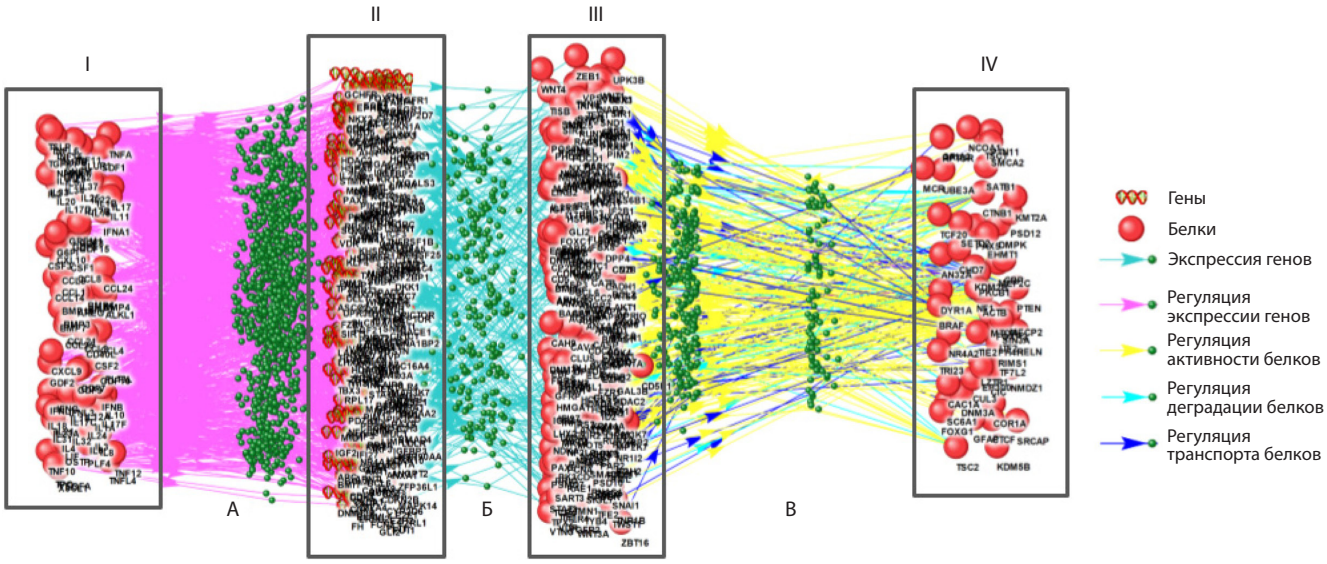


Рис. 1. Пример реконструкции подсети: реконструкция взаимодействия цитокинов с белками, ассоциированными с РАС, в программе Pathway Wizard по шаблону 1 из табл. S1.

Римскими цифрами обозначены: I – цитокины, которые регулируют белки и гены, ассоциированные с РАС; II – гены-посредники; III – белки-посредники; IV – белки, ассоциированные с РАС, которые через регуляторные пути регулируются цитокинами. Буквами обозначены: А – процессы регуляции экспрессии генов; Б – экспрессия генов; В – процессы регуляции активности, транспорта и деградации белков.

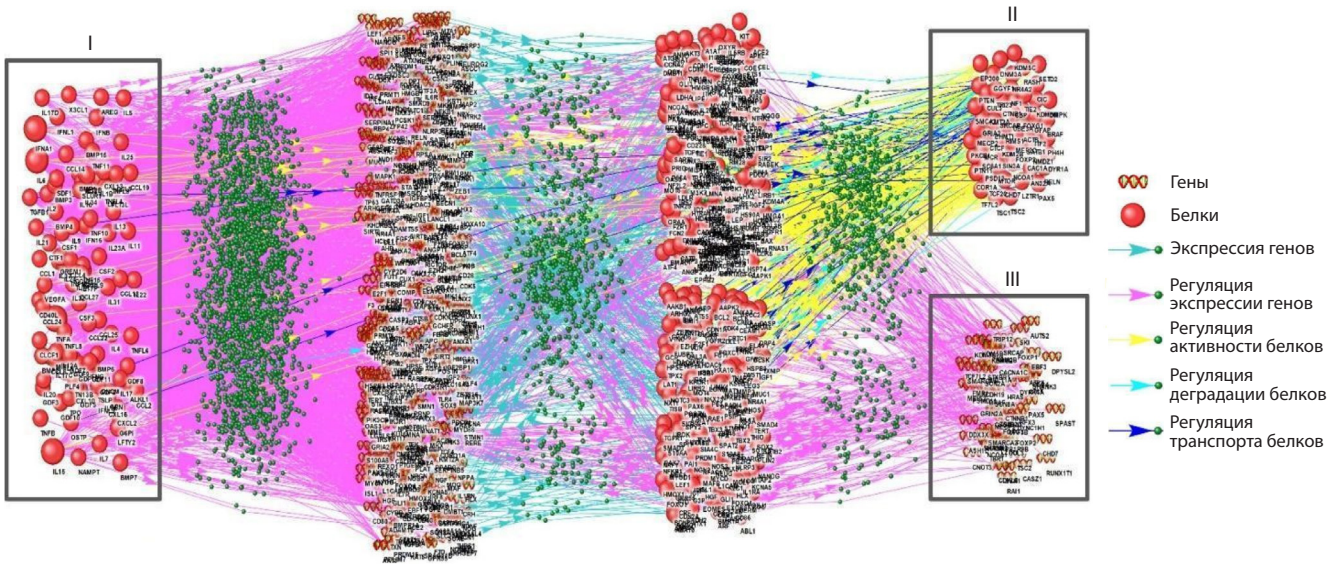


Рис. 2. Результаты реконструкции итоговой сети взаимодействия цитокинов с белками и генами, ассоциированными с РАС.

Римскими цифрами обозначены: I – цитокины, которые регулируют белки и гены, ассоциированные с РАС; II – белки, ассоциированные с РАС, которые регулируются цитокинами; III – гены, ассоциированные с РАС, которые регулируются цитокинами.

Таблица 1. Типы и количество вершин и связей в итоговой генной сети взаимодействия цитокинов с белками и генами, ассоциированными с РАС

Тип связи	Количество	Тип вершины	Количество
Регуляция активности	369	Белок	621
Регуляция деградации	64	Ген	491
Регуляция экспрессии	2772		
Регуляция транспорта	65		
Экспрессия гена	409		

Таблица 2. Обогащенность набора генов итоговой генной сети, ассоциированных с РАС, которые регулируются цитокинами

№ п/п	Биологический процесс	Число генов	FDR
1	Возбуждающий постсинаптический потенциал	6	$9.7 \cdot 10^{-4}$
2	Регуляция развития дендритных шипиков	4	$9.7 \cdot 10^{-4}$
3	Развитие гиппокампа	6	$9.7 \cdot 10^{-4}$
4	Положительная регуляция глутаматергической синаптической передачи	5	$8.5 \cdot 10^{-4}$
5	Миграция нейронов	7	$8.5 \cdot 10^{-4}$
6	Нейроразвитие	12	$7.6 \cdot 10^{-4}$
7	Отрицательная регуляция процесса апоптоза нейронов	8	$4.6 \cdot 10^{-4}$
8	Трансмембранный импорт ионов кальция в цитозоль	4	$4.6 \cdot 10^{-4}$

Примечание. FDR (false discovery rate) – вероятность ложноположительного исхода.

Чтобы оценить вовлеченность цитокинов в биологические процессы, ассоциированные с РАС, был разработан специализированный скрипт, который принимал на вход набор цитокинов, выявленных при реконструкции генной сети, и перечень биологических процессов РАС, полученный из анализа обогащения терминами биологических процессов Gene Ontology генов из базы данных SFARI. Алгоритм последовательно оценивал участие каждого цитокина в биологических процессах из перечня ассоциированных с РАС. Выявлено 13 цитокинов, которые участвуют в биологических процессах, вовлеченных в РАС (FDR < 0.05) (табл. 3).

Для ранжирования цитокинов по влиянию в сети использовались две метрики: степень посредничества (betweenness centrality) и степень вершины (degree centrality). Степень посредничества вершины отражает количество

Таблица 3. Приоритизация цитокинов по представленности в биологических процессах, ассоциированных с РАС

№ п/п	Название цитокина	Число биологических путей, ассоциированных с РАС	PadjValue	FDR
1	BMP2	6	$1.0 \cdot 10^{-6}$	$3.8 \cdot 10^{-5}$
2	IL-4	5	$5.0 \cdot 10^{-7}$	$2.7 \cdot 10^{-5}$
3	TGFB1	8	$1.9 \cdot 10^{-7}$	$2.1 \cdot 10^{-5}$
4	BMP4	6	$8.7 \cdot 10^{-6}$	$2.0 \cdot 10^{-4}$
5	IFN-γ	4	$7.2 \cdot 10^{-5}$	$1.5 \cdot 10^{-3}$
6	VEGFA	5	$8.8 \cdot 10^{-5}$	$1.6 \cdot 10^{-3}$
7	BMP7	4	$1.0 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$
8	TNF-α	5	$3.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-3}$
9	IL-33	3	$4.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-3}$
10	IL-10	3	$1.9 \cdot 10^{-3}$	$2.1 \cdot 10^{-2}$
11	IFNL1	2	$2.4 \cdot 10^{-3}$	$2.5 \cdot 10^{-2}$
12	GPI	2	$5.7 \cdot 10^{-3}$	$4.9 \cdot 10^{-2}$
13	GDF2	2	$5.7 \cdot 10^{-3}$	$4.9 \cdot 10^{-2}$

Примечание. Сортировка выполнена по возрастанию значения FDR; PadjValue – уровень статистической значимости с учетом поправки Бонферрони.

Таблица 4. Приоритизация цитокинов по степени посредничества и степени вершины

№ п/п	Название цитокина	Степень посредничества	№ п/п	Название цитокина	Степень вершины
1	TNF-α	$19.3 \cdot 10^4$	1	TNF-α	100
2	IL-6	$10.4 \cdot 10^4$	2	IL-6	69
3	IL-4	$10.2 \cdot 10^4$	3	IL-4	65
4	TGFB1	$9.8 \cdot 10^4$	4	BMP4	52
5	BMP4	$9.0 \cdot 10^4$	5	TGFB1	51
6	VEGFA	$8.5 \cdot 10^4$	6	VEGFA	51
7	IL-2	$7.7 \cdot 10^4$	7	IL-2	47
8	IL-1β	$6.5 \cdot 10^4$	8	IL-1β	46
9	BMP2	$5.7 \cdot 10^4$	9	IL-10	46
10	OSTP	$5.5 \cdot 10^4$	10	BMP2	37
11	IL-10	$5.4 \cdot 10^4$	11	IFN-γ	34
12	IL-8	$4.0 \cdot 10^4$	12	IL-22	32
13	IFN-γ	$3.9 \cdot 10^4$	13	IL-17	32
14	IL-17	$3.1 \cdot 10^4$	14	OSTP	30
15	IL-15	$2.9 \cdot 10^4$	15	IL-8	29

Примечание. Степень посредничества – количество кратчайших путей в сети, проходящих через данную вершину; степень вершины – число непосредственных связей вершины с другими элементами сети.

кратчайших путей в сети через нее. Степень вершины определяется числом непосредственных ее связей с другими вершинами. Метрики показывают влияние вершины на сеть – чем выше их значения, тем больше роль вершины. В табл. 4 приведены 15 наиболее влиятельных цитокинов по каждой метрике.

Цитокины как потенциальные мишени для фармакологических воздействий

На основе табл. 3 и 4 был сформирован список из 21 потенциально важного регулятора: BMP2, BMP4, BMP7, GDF2, GPI, IFN-γ, IFNL1, IL-10, IL-33, IL-15, IL-17, IL-1β, IL-2, IL-22, IL-4, IL-6, IL-8, OSTP, TGFB1, TNF-α и VEGFA. Проверка списка по базе данных GETdb подтвердила статус цитокинов как перспективных фармакологических мишеней.

На основании информации из базы данных DrugBank из 21 цитокина выделено восемь (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17), которые уже являются мишенями одобренных к применению препаратов. При этом четыре из них (IL-4, VEGFA, TNF-α и IFN-γ) оказались в нашем списке восьми наиболее приоритетных, которые

наиболее высоко оценивались по совокупности методик приоритизации: IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ и TNF-α.

В клинической практике препараты, воздействующие на цитокины TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ и IL-17, применяются преимущественно в качестве иммуносупрессоров и противоопухолевых средств. Воздействие реализуется через блокаду рецепторов цитокинов (антагонисты) или нейтрализацию цитокинов (моноклональные антитела).

Обсуждение

Результатом анализа (см. табл. 3 и 4) стало выделение 21 потенциально важного цитокина (BMP2, BMP4, BMP7, GDF2, GPI, IFN-γ, IFNL1, IL-10, IL-33, IL-15, IL-17, IL-1β, IL-2, IL-22, IL-4, IL-6, IL-8, OSTP, TGFB1, TNF-α и VEGFA) как перспективных фармакологических мишеней (согласно базе данных GETdb), и восемь из них (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ и IL-17) являются мишенями уже используемых в клинической практике препаратов, согласно базе данных DrugBank. По литературным данным, в развитии и функционировании ЦНС непосредственно участвуют провоспалительные цитокины TNF-α, IL-6, IL-2, IL-1β, IFN-γ, VEGFA, IL-17A, экспрессируемые активированной микроглией в состоянии классической активации и способствующие развитию нейровоспаления. Некоторые из этих цитокинов (IL-6, IFN-γ, IL-17A) влияют также на эмбриональное развитие мозга и могут способствовать сдвигу развития в направлении РАС при активации материнского иммунного ответа в период беременности (Fujitani et al., 2022; Majerczyk et al., 2022).

Так, в моделировании на мышах активации материнского иммунитета (MIA) CD4+ Т-лимфоциты пораженного потомства продуцировали более высокие количества IL-17A при активации *in vitro* (Morgan et al., 2010; Parkhurst et al., 2013). Кроме того, установлено, что работа материнских RORγt-экспрессирующих провоспалительных Т-клеток (Th17), основного источника IL-17A, является необходимым условием для индукции подобных РАС фенотипов у потомства. Блокирование активности IL-17A с помощью антител к нему у беременных мышей защищало от развития у их потомства поведенческих аномалий, вызванных MIA. Такой же эффект вызывает и воздействие на клетки Th17 (Choi et al., 2016). С другой стороны, введение IL-17A в мозг зародыша провоцирует нарушения в развитии полушарий головного мозга и проявление ассоциированных с РАС симптомов. На поведенческом уровне симптомы РАС отражают прежде всего измененную активность правого полушария, отвечающего за механизмы адаптации и приспособления (Николаева, Вергунов, 2020). Помимо того, у детей с РАС значимо чаще отмечается леворукость (Николаева, Гайдамакина, 2018).

Несмотря на эффект, который IL-17A оказывает на развитие ЦНС зародыша, существуют исследования, свидетельствующие о том, что терапия этим цитокином может нормализовывать поведение уже взрослых потомков матерей с MIA. В 2020 г. была опубликована работа M. Reed с коллегами, в которой проверялась эффективность терапии липополисахаридами (LPS) на различных мышиных моделях РАС. LPS-терапия нормализовала поведение у модельных грызунов, которые являлись потомками матерей с MIA. Однако такая терапия была неэффективна для модельных мышей с моногенными нарушениями, провоцирующими развитие РАС. Установлено, что такой результат обусловлен различиями в уровнях цитокинов, которые выделяются в ответ на LPS (Reed et al., 2020). По сравнению с мышами, у которых РАС развился вследствие MIA, у мышей с моногенными нарушениями IL-17A продуцировался на более низком уровне.

На фоне активности провоспалительных цитокинов выделяется противовоспалительный цитокин IL-4, который играет ключевую роль в альтернативной активации микроглии (M2) и экспрессируется клетками в этом состоянии. Микроглия M2 проявляет противовоспалительную и заживляющую активность, а также секретирует факторы роста, среди которых инсулиноподобный фактор роста I (IGF-I), фактор роста фибробластов (FGF) и фактор, стимулирующий колонии макрофагов (CSF1), а также нейротрофические факторы (Sica, Mantovani, 2012). Эти факторы задействуют рецепторы Trk (семейство рецепторных тирозинкиназ), которые регулируют синаптическую пластичность.

Значительное повышение уровня IL-4 было выявлено в амниотической жидкости и сыворотке крови беременных, у детей которых впоследствии было диагностировано РАС (Goines, Ashwood, 2013). Патогенетическая роль увеличения концентрации IL-4 остается неясной: оно может как способствовать развитию патологии, так и представлять собой компенсаторный механизм в ответ на воспалительные процессы.

Поскольку речь идет о перепрофилировании уже известных препаратов, то мы предполагаем, что наиболее перспективным направлением может быть проведение исследований, аналогичных работе (Reed et al., 2020), с использованием цитокинов из полученного нами списка и уже применяемых в клинической практике (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17). При планировании последующих экспериментов необходимо учитывать, что эффективность воздействия цитокинов может варьировать в зависимости от этиологии заболевания и стадии онтогенеза организма. Для обеспечения комплексного подхода целесообразно задействовать модельные линии грызунов, которые отличаются способом индуцирования формирования РАС и его форм, с последующим сравнительным анализом полученных данных. Такая методика позволит более глубоко изучить влияние цитокинов на развитие и симптомы РАС различного генеза, а также оценить возможность перепрофилирования соответствующих лекарственных препаратов для терапии и облегчения симптоматики РАС.

Заключение

• С использованием базы знаний когнитивной системы ANDSystem и компонент ANDSystem была выполнена компьютерная реконструкция пяти специализированных генных подсетей, отражающих различные пути воздействия цитокинов на белки и гены, ассоциированные с РАС, для максимально полного охвата регуляторных взаимодействий цитокинов с белками и генами, которые ассоциированы с РАС.

Путем интеграции этих подсетей в единую итоговую генную сеть впервые была реконструирована генная сеть цитокиновой регуляции генов и белков, которые ассоциированы с РАС. Полученная сеть содержит 1112 вершин двух типов (491 ген, 621 белок) и 3675 взаимодействий пяти различных видов между ними.

• Анализ итоговой генной сети позволил идентифицировать 110 цитокинов, регулирующих активность, транспорт и стабильность компонентов сети, которые вовлечены в РАС, 58 белков и 91 ген, вовлеченных в формирование РАС, которые регулируются цитокинами. Определены значимые характеристики итоговой генной сети, свидетельствующие о важной роли цитокиновой регуляции в патогенезе РАС.

Последующий анализ перепредставленности терминов Gene Ontology (GO), относящихся к биологическим процессам, для генов, ассоциированных с РАС, которые, согласно данным реконструированной сети взаимодействий, находятся под регуляторным контролем цитокинов, выявил 56 статистически значимых биологических процессов, связанных с недоразвитием. Среди этих процессов можно выделить формирование дендритных шипиков, развитие гиппокампа, нейрональную миграцию и регуляцию синаптической передачи.

• Для выявления ключевых регуляторов проведена приоритизация цитокинов с использованием анализа метрик генной сети (степень посредничества и степень вершин) и оценки функциональной значимости элементов сети через ассоциацию с биологическими процессами Gene Ontology, связанными с РАС. В результате был выделен 21 цитокин, при этом восемь из них (IL-4, TGF-β1, BMP4, VEGFA, BMP2, IL-10, IFN-γ, TNF-α) показали наивысший приоритет по всем критериям.

Восемь цитокинов (TNF-α, IL-6, IL-4, VEGFA, IL-2, IL-1β, IFN-γ, IL-17) из 21 являются мишенями уже одобренных лекарственных препаратов, которые применяются в иммуносупрессивной и противоопухолевой терапии. Это открывает перспективы для перепрофилирования существующих иммуномодулирующих препаратов для терапии РАС. Остальные 13 цитокинов представляют собой потенциальные мишени для препаратов, находящихся в стадии клинических испытаний. Дальнейшие исследования *in vitro* и *in vivo* позволят уточнить механизмы влияния выявленных цитокинов на нейроразвитие и оценить эффективность воздействия на них.

Список литературы / References

Николаева Е.И., Вергунов Е.Г. Функциональная асимметрия мозга и латеральные предпочтения: перезагрузка. Эволюционный, генетический, психологический и психофизиологический подходы к анализу. СПб., 2020
[Nikolaeva E.I., Vergunov E.G. Functional Asymmetry of the Brain and Lateral Preferences: Reloaded. Evolutionary, Genetic, Psychological and Psychophysiological Approaches to Analysis. St. Petersburg, 2020 (in Russian)]

Николаева Е.И., Гайдамакина М.Д. Специфика латеральных предпочтений мутичных дошкольников с заболеваниями аутистического спектра. *Асимметрия*. 2018;12(4):367-371
[Nikolaeva E.I., Gaidamakina M.D. Specificity of lateral preference of mute preschoolers with autism spectrum disorders. *J Asymmetry*. 2018;12(4):367-371 (in Russian)]

Abrahams B.S., Arking D.E., Campbell D.B., Mefford H.C., Morrow E.M., Weiss L.A., Menashe I., Wadkins T., Banerjee-Basu S., Packer A. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism*. 2013;4(1):36. doi 10.1186/2040-2392-4-36

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorder. American Psychiatric Publ., 2013. doi 10.1176/appi.books.9780890425596

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., … Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi 10.1038/75556

Ashwood P., Krakowiak P., Hertz-Picciotto I., Hansen R., Pessah I., Van de Water J. Elevated plasma cytokines in autism spectrum disorders provide evidence of immune dysfunction and are associated with impaired behavioral outcome. *Brain Behav Immun*. 2011;25(1):40-45. doi 10.1016/j.bbi.2010.08.003

Choi G.B., Yim Y.S., Wong H., Kim S., Kim H., Kim S.V., Hoeffer C.A., Littman D.R., Huh J.R. The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. *Science*. 2016;351(6276):933-939. doi 10.1126/science.aad0314

Demenkov P.S., Ivanisenko T.V., Kolchanov N.A., Ivanisenko V.A. ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem. *In Silico Biol*. 2012;11(3-4):149-161. doi 10.3233/ISB-2012-0449

Fujitani M., Miyajima H., Otani Y., Liu X. Maternal and adult interleukin-17A exposure and autism spectrum disorder. *Front Psychiatry*. 2022;13:836181. doi 10.3389/FPSYT.2022.836181/PDF

Ganesan H., Balasubramanian V., Iyer M., Venugopal A., Subramaniam M.D., Cho S.G., Vellingiri B. mTOR signalling pathway – a root cause for idiopathic autism? *BMB Rep*. 2019;52(7):424-433. doi 10.5483/BMBRep.2019.52.7.137

Goines P.E., Ashwood P. Cytokine dysregulation in autism spectrum disorders (ASD): possible role of the environment. *Neurotoxicol Teratol*. 2013;36:67-81. doi 10.1016/j.ntt.2012.07.006

Harms M.B., Martin A., Wallace G.L. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychol Rev*. 2010;20(3):290-322. doi 10.1007/S11065-010-9138-6

Hoeffer C.A., Klann E. mTOR signaling: at the crossroads of plasticity, memory and disease. *Trends Neurosci*. 2010;33(2):67-75. doi 10.1016/J.TINS.2009.11.003

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi 10.1038/nprot.2008.211

Hutsler J.J., Zhang H. Increased dendritic spine densities on cortical projection neurons in autism spectrum disorders. *Brain Res*. 2010;1309:83-94. doi 10.1016/j.brainres.2009.09.120

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics*. 2020;21(11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int J Mol Sci*. 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an Associative Network Discovery System for automated literature mining in the field of biology. *BMC Systems Biol*. 2015;9(Suppl.2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl.1):34. doi 10.1186/S12859-018-2567-6

Knox C., Wilson M., Klinger C.M., Franklin M., Oler E., Wilson A., Pon A., … Ackerman D., Jewison T., Sajed T., Gautam V., Wishart D.S. DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. doi 10.1093/nar/gkad976

Li X., Chauhan A., Sheikh A.M., Patil S., Chauhan V., Li X.M., Ji L., Brown T., Malik M. Elevated immune response in the brain of autistic patients. *J Neuroimmunol*. 2009;207(1-2):111-116. doi 10.1016/j.jneuroim.2008.12.002

Liu Y., Zhang D.T., Liu X.G. mTOR signaling in T cell immunity and autoimmunity. *Int Rev Immunol*. 2015;34(1):50-66. doi 10.3109/08830185.2014.933957

Majerczyk D., Ayad E.G., Brewton K.L., Saing P., Hart P.C. Systemic maternal inflammation promotes ASD via IL-6 and IFN-γ. *Biosci Rep*. 2022;42(11):BSR20220713. doi 10.1042/BSR20220713

McMahon J., Huang X., Yang J., Komatsu M., Yue Z., Qian J., Zhu X., Huang Y. Impaired autophagy in neurons after disinhibition of mammalian target of rapamycin and its contribution to epileptogenesis. *J Neurosci*. 2012;32(45):15704-15714. doi 10.1523/JNEUROSCI.2392-12.2012

Morgan J.T., Chana G., Pardo C.A., Achim C., Semendeferi K., Buckwalter J., Courchesne E., Everall I.P. Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biol Psychiatry*. 2010;68(4):368-376. doi 10.1016/j.biopsych.2010.05.024

Navlakha S., Barth A.L., Bar-Joseph Z. Decreasing-rate pruning optimizes the construction of efficient and robust distributed networks. *PLoS Comput Biol*. 2015;11(7):e1004347. doi 10.1371/JOURNAL.PCBI.1004347

Onore C., Yang H., Van de Water J., Ashwood P. Dynamic Akt/mTOR signaling in children with autism spectrum disorder. *Front Pediatr*. 2017;5:43. doi 10.3389/fped.2017.00043

Parkhurst C.N., Yang G., Ninan I., Savas J.N., Yates J.R., Lafaille J.J., Hempstead B.L., Littman D.R., Gan W.B. Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. *Cell*. 2013;155(7):1596-1609. doi 10.1016/j.cell.2013.11.030

Ragupathi A., Kim C., Jacinto E. The mTORC2 signaling network: targets and cross-talks. *Biochem J*. 2024;481(2):45-91. doi 10.1042/BCJ20220325

Reed M.D., Yim Y.S., Wimmer R.D., Kim H., Ryu C., Welch G.M., Andina M., King H.O., Waisman A., Halassa M.M., Huh J.R., Choi G.B. IL-17a promotes sociability in mouse models of neurodevelopmental disorders. *Nature*. 2020;577(7789):249-253. doi 10.1038/S41586-019-1843-6

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Sica A., Mantovani A. Macrophage plasticity and polarization: in vivo veritas. *J Clin Invest*. 2012;122(3):787-795. doi 10.1172/JCI59643

Smith J.A., Das A., Ray S.K., Banik N.L. Role of pro-inflammatory cytokines released from microglia in neurodegenerative diseases. *Brain Res Bull*. 2012;87(1):10-20. doi 10.1016/j.brainresbull.2011.10.004

Tang G., Gudsnuk K., Kuo S.H., Cotrina M.L., Rosoklija G., Sosunov A., Sonders M.S., … Peterson B.S., Champagne F., Dwork A.J., Goldman J., Sulzer D. Loss of mTOR-dependent macroautophagy causes autistic-like synaptic pruning deficits. *Neuron*. 2014;83(5):1131-1143. doi 10.1016/j.neuron.2014.07.040

Trifonova E.A., Klimenko A.I., Mustafin Z.S., Lashin S.A., Kochetov A.V. The mTOR signaling pathway activity and vitamin D availability control the expression of most autism predisposition genes. *Int J Mol Sci*. 2019;20(24):6332. doi 10.3390/ijms20246332

Tsai A.C., Savostyanov A.N., Wu A., Evans J.P., Chien V.S.C., Yang H.H., Yang D.Y., Liou M. Recognizing syntactic errors in Chinese and English sentences: brain electrical activity in Asperger's syndrome. *Res Autism Spectr Disord*. 2013;7(7):889-905. doi 10.1016/j.rasd.2013.02.001

Tseng Y.L., Yang H.H., Savostyanov A.N., Chien V.S.C., Liou M. Voluntary attention in Asperger's syndrome: brain electrical oscillation and phase-synchronization during facial emotion recognition. *Res Autism Spectr Disord*. 2015;13-14:32-51. doi 10.1016/j.rasd.2015.01.003

Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., … Johansson F., Zwahlen M., Von Heijne G., Nielsen J., Pontén F. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. doi 10.1126/science.1260419

Van der Zee E., Derksen J.J.L. The power of systemizing in autism. *Child Psychiatry Hum Dev*. 2021;52(2):321-331. doi 10.1007/S10578-020-01014-4

Vargas D.L., Nascimbene C., Krishnan C., Zimmerman A.W., Pardo C.A. Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann Neurol*. 2005;57(1):67-81. doi 10.1002/ana.20315

Zhang Q., He Y., Lu Y.P., Wei Q.H., Zhang H.Y., Quan Y. GETdb: a comprehensive database for genetic and evolutionary features of drug targets. *Comput Struct Biotechnol J.* 2024;23:1429-1438. doi 10.1016/j.csbj.2024.04.006

Ziats C.A., Patterson W.G., Friez M. Syndromic autism revisited: review of the literature and lessons learned. *Pediatr Neurol*. 2021;114:21-25. doi 10.1016/J.PEDIATRNEUROL.2020.06.011

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 11.08.2025. После доработки 12.09.2025. Принята к публикации 15.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(6):1009-1019

doi 10.18699/vjgb-25-106

Генная сеть и база знаний по терморегуляции организма человека

Е.В. Игнатьева[1], П.С. Деменков[1], А.Г. Богомолов[1], Р.А. Иванов[1], С.А. Лашин[1, 2], А.Д. Михайлова[2], А.Е. Алексеева[2], Н.С. Юдин[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
eignat@bionet.nsc.ru

Аннотация. Реконструкция и анализ генных сетей, регулирующих биологические процессы, – один из эффективных подходов к исследованию сложных систем обеспечения жизнедеятельности организмов. Терморегуляция – важное эволюционное приобретение человека и других теплокровных животных. Терморегуляция осуществляется при участии многих физиологических систем организма (нервной, сердечно-сосудистой, эндокринной, дыхательной, мышечной и т.д.), что способствует поддержанию относительно постоянной температуры тела в условиях колебания температуры окружающей среды. Цель работы – компьютерная реконструкция генной сети терморегуляции человека и представление полученных результатов в соответствующей базе знаний Termo_Reg_Human 1.0. Генная сеть реконструирована с использованием программно-информационной системы ANDSystem, предназначенной для автоматизированного извлечения знаний и фактов из текстов научных публикаций и баз данных биомедицинской направленности, основанной на методах машинного обучения и искусственного интеллекта. База знаний Termo_Reg_Human 1.0 (https://www.sysbio.ru/ThermoReg_Human/) содержит информацию о генной сети терморегуляции человека, включая описание 469 генов, 473 белков и 265 микроРНК, значимых для ее функционирования; взаимодействиях между этими объектами, а также эволюционные характеристики генов. С использованием программного инструмента ANDVisio (модуля системы ANDSystem) проведена приоритизация каждого гена, белка и микроРНК, участвующих в терморегуляции организма человека по их функциональной нагруженности – количеству связей с другими объектами реконструированной генной сети. Установлено, что к числу ключевых объектов, имеющих наибольшее количество функциональных связей в генной сети терморегуляции человека, относятся гены *UCP1*, *VEGFA*, *PPARG*, *DDIT3*, белки STAT3, JUN, VEGFA, TLR4, TNFA и микроРНК hsa-mir-335 и hsa-mir-26b. Обнаружено обогащение генной сети терморегуляции генами, предковые варианты которых сформировались на эволюционных этапах появления одноклеточных организмов и дивергенции позвоночных.

Ключевые слова: тепло; холод; генная сеть; база данных; микроРНК; эволюция; филостратиграфия; возраст гена

Для цитирования: Игнатьева Е.В., Деменков П.С., Богомолов А.Г., Иванов Р.А., Лашин С.А., Михайлова А.Д., Алексеева А.Е., Юдин Н.С. Генная сеть и база знаний по терморегуляции организма человека. *Вавиловский журнал генетики и селекции*. 2025;29(7):1009-1019. doi 10.18699/vjgb-25-106

Финансирование. Исследование выполнено за счет средств бюджетного проекта ФИЦ ИЦиГ СО РАН «Системная биология и биоинформатика: реконструкция, анализ и моделирование структурно-функциональной организации и эволюции генных сетей человека, животных, растений и микроорганизмов» № FWNR-2022-0020.

Благодарности. Авторы выражают благодарность В.А. Иванисенко за консультации по использованию ANDSystem.

The gene network and knowledge base on human thermoregulation

E.V. Ignatieva[1], P.S. Demenkov[1], A.G. Bogomolov[1], R.A. Ivanov[1], S.A. Lashin[1, 2], A.D. Mikhailova[2], A.E. Alekseeva[2], N.S. Yudin[1]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
eignat@bionet.nsc.ru

Abstract. Reconstruction and analysis of gene networks regulating biological processes are among the modern methodological approaches for studying complex biological systems that ensure the vital activity of organisms. Thermoregulation is an important evolutionary acquisition of warm-blooded animals. Multiple physiological systems (nervous, cardiovascular, endocrine, respiratory, muscular, etc.) are involved in this process, maintaining stable body temperature despite changes in ambient temperature. This study aims to perform a computer reconstruction of the human thermoregulation gene network and present the results in the Termo_Reg_Human 1.0 knowledge base. The gene network was reconstructed

© Игнатьева Е.В., Деменков П.С., Богомолов А.Г., Иванов Р.А., Лашин С.А., Михайлова А.Д., Алексеева А.Е., Юдин Н.С., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

using the ANDSystem software and information system, designed for the automated extraction of knowledge and facts from scientific publications and biomedical databases based on machine learning and artificial intelligence methods. The Termo_Reg_Human 1.0 knowledge base (https://www.sysbio.ru/ThermoReg_Human/) contains information about the human thermoregulation gene network, including a description of 469 genes, 473 proteins, and 265 microRNAs important for its functioning, interactions between these objects, and the evolutionary characteristics of the genes. Using the ANDVisio software tool (a module of ANDSystem), each gene, protein, and microRNA involved in the thermoregulation of the human body was prioritized according to its functional significance, i.e., the number of interactions with other objects in the reconstructed gene network. It was found that the key objects with the largest number of functional interactions in the human thermoregulation gene network included the *UCP1*, *VEGFA*, *PPARG* and *DDIT3* genes; STAT3, JUN, VEGFA, TLR4 and TNFA proteins; and the microRNAs hsa-mir-335 and hsa-mir-26b. We revealed that the set of 469 human genes from the network was enriched with genes whose ancestral forms originated at an early evolutionary stage (Unicellular organisms, the root of the phylostratigraphic tree) and at the stage of Vertebrata divergence.

Key words: heat; cold; gene network; database; microRNA; evolution; phylostratigraphy; gene age

For citation: Ignatieva E.V., Demenkov P.S., Bogomolov A.G., Ivanov R.A., Lashin S.A., Mikhailova A.D., Alekseeva A.E., Yudin N.S. The gene network and knowledge base on human thermoregulation. *Vavilovskii Zhurnal Genetiki i Selektsii* = *Vavilov J Genet Breed*. 2025;29(7):1009-1019. doi 10.18699/vjgb-25-106

Введение

Человек и другие млекопитающие – гомойотермные животные, способные к поддержанию относительно постоянной температуры тела в условиях колебаний температуры окружающей среды (Osvath et al., 2024). Система терморегуляции человека включает в себя: 1) терморецепторы, расположенные на поверхности тела и во внутренних органах; 2) восходящие нервные пути передачи сигналов; 3) центры терморегуляции в гипоталамусе и других отделах мозга; 4) нисходящие нервные пути, управляющие адаптивными реакциями (Nakamura, 2024). К числу таких реакций относятся: а) сократительный и несократительный термогенез (химические механизмы терморегуляции) (Ikeda, Yamada, 2020; Dumont et al., 2025); б) физическая терморегуляция, включая регуляцию отдачи тепла через испарение и конвекцию, а также теплоизоляцию (Nakamura, 2011; Tattersall et al., 2012); в) поведенческие реакции: избегание открытых участков земной поверхности, характеризующихся экстремальными температурами; скучивание особей и др. (Tattersall et al., 2012; Tansey, Johnson, 2015; McCafferty et al., 2017).

Химическая терморегуляция основана на выработке тепла при мышечных сокращениях (Blondin et al., 2018; Dumont et al., 2025) и несократительном термогенезе в бурой жировой ткани (Tansey, Johnson, 2015; Ikeda, Yamada, 2020) и мышцах (Blondin et al., 2018). Процессы физической терморегуляции осуществляются за счет изменения отдачи тепла с поверхностей тела: теплообмена между кожей и воздушной средой, потоотделения, испарения влаги с поверхности дыхательных путей, термоизоляции за счет подкожного жирового слоя, пилоэрекции (Nakamura, 2011; Tattersall et al., 2012). В управлении как химической, так и физической терморегуляцией активно участвует нейроэндокринная система (Charkoudian et al., 2017; Nakamura, 2024; Mittag, Kolms, 2025).

Кроме того, процессы химической и физической терморегуляции сопряжены с изменениями в функционировании сердечно-сосудистой системы (Tansey, Johnson, 2015). Таким образом, терморегуляция обеспечивается множеством биологических процессов, протекающих в нервной, эндокринной, сердечно-сосудистой, дыхательной, мышечной и других системах организма. Важнейший вклад в терморегуляцию вносят генетические механизмы контроля перечисленных выше процессов (Festuccia et al., 2009; Rehman et al., 2013; Li et al., 2015; Horii et al., 2019; Xiao et al., 2019; Kudsi et al., 2022; Song et al., 2022; Valdivia et al., 2023).

Реконструкция и анализ генных сетей, регулирующих биологические процессы, – один из эффективных подходов к исследованию сложных систем обеспечения жизнедеятельности организмов (Ignatieva et al., 2017; Saik et al., 2018; Mustafin et al., 2019; Мустафин и др., 2021; Михайлова и др., 2024). По проблеме терморегуляции накоплены огромные объемы экспериментальных генетических данных, представленных в десятках тысяч научных публикаций и многих фактографических базах данных (см. KEGG Pathway, WikiPathways, MetaCyc, REACTOME и др.). В связи с этим в нашей работе осуществлена реконструкция генной сети терморегуляции человека с использованием программно-информационной системы ANDSystem, предназначенной для автоматизированного извлечения знаний и фактов из текстов научных публикаций и баз данных биомедицинской направленности и использующей методы машинного обучения и искусственного интеллекта (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2024). Результаты, полученные на основе анализа 30 млн публикаций, накапливаются в специализированной базе знаний системы ANDSystem в виде глобального графа знаний (Ivanisenko T.V. et al., 2024).

Информация о реконструированной генной сети терморегуляции человека представлена в базе знаний Termo_Reg_Human 1.0 (https://www.sysbio.ru/ThermoReg_Human/), включая описание 469 генов, 473 белков, 265 микроРНК, значимых для ее функционирования, а также взаимодействий между ними.

С использованием программного инструмента ANDVisio (модуля системы ANDSystem) проведена приоритизация каждого гена, белка и микроРНК, участвующих в терморегуляции организма человека, по их функциональной нагруженности – количеству связей с другими объектами реконструированной генной сети. Установлено, что к числу ключевых объектов, имеющих наибольшее количество функциональных связей в генной сети терморегуляции человека, относятся гены *UCP1*, *VEGFA*, *PPARG*, *DDIT3*, белки STAT3, JUN, VEGFA, TLR4, TNFA и микроРНК hsa-mir-335 и hsa-mir-26b.

В базе знаний Termo_Reg_Human 1.0 также представлены результаты эволюционного анализа генов, функционирующих в генной сети терморегуляции. Обнаружено обогащение этой генной сети генами, предковые варианты которых сформировались на двух важнейших эволюционных этапах, соответствующих: а) появлению одноклеточных организмов и б) дивергенции позвоночных.

Материалы и методы

Список генов, использованный для построения генной сети. Формирование списка генов человека, вовлеченных в терморегуляцию, осуществляли на основе баз данных Gene Ontology, EntrezGene, а также системы ANDSystem (Ivanisenko V.A. et al., 2019) с использованием ключевых слов, приведенных в Приложении S1[1].

Построение генной сети. Генная сеть терморегуляции была реконструирована с помощью программно-информационной системы ANDSystem (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2024). Система ANDSystem, основанная на методах машинного обучения и искусственного интеллекта, предназначена для автоматизированного извлечения знаний и фактов о структурно-функциональной организации генных сетей из текстов научных публикаций и фактографических баз данных биомедицинской направленности. Получаемая таким образом информация накапливается в специализированной базе знаний системы ANDSystem в виде глобального графа знаний (Ivanisenko T.V. et al., 2024). На основе этой информации осуществляется реконструкция графов целевых генных сетей, вершины которых соответствуют молекулярно-генетическим объектам – генам, РНК, белкам, метаболитам, функционирующим в составе генных сетей, а соединяющие их ребра – функциональным связям между ними. Детали реконструкции генной сети терморегуляции организма человека с использованием системы ANDSystem описаны в Приложении S2.

Приоритизация генов, белков и микроРНК по их функциональной значимости в генной сети терморегуляции человека. Приоритизацию объектов генной сети (генов, микроРНК и белков) проводили с использованием программного инструмента ANDVisio (модуля системы ANDSystem). Для конкретного объекта в графе генной сети терморегуляции человека вычислялось количество его связей с другими объектами. Далее оценивалась вероятность обнаружить по случайным причинам наблюдаемое количество связей каждого из объектов генной сети. Расчет вероятности осуществляли по гипергеометрическому тесту:

$$p\text{-value} = \sum_{i=0}^{k} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}},$$

где k – число связей рассматриваемого объекта в генной сети; n – число объектов, попавших в рассматриваемую генную сеть; K – число связей рассматриваемого объекта в глобальной сети базы знаний ANDSystem; N – общее количество объектов в глобальном графе базы знаний системы ANDSystem (Ivanisenko V.A. et al., 2019).

При вычислении p-value принимались во внимание только объекты того же типа (гены, белки, микроРНК), что и рассматриваемый объект генной сети терморегуляции человека. Далее применялась поправка на множественные сравнения (Benjamini, Yekutieli, 2001), в результате чего была получена величина P-adjusted.

Анализ эволюционных характеристик генов, функционирующих в составе реконструированной генной сети, осуществляли с использованием системы Orthoweb (Иванов и др., 2024), вычисляющей для каждого гена его филостратиграфический индекс PAI (phylostratigraphic age index), характеризующий эволюционный возраст предковой формы гена. Детали процесса вычисления индекса PAI описаны в Приложении S2.

Функциональная аннотация генов. Выявление терминов Gene Ontology, с которыми статистически значимо связаны гены, имеющие определенный филостратиграфический возраст, осуществляли с помощью интернет-инструмента DAVID и его словаря GOTERM_BP_DIRECT (Sherman et al., 2022).

Формирование базы знаний по терморегуляции человека. Данные для информационных таблиц базы знаний получали из текстовых выдач программы ANDVisio (модуля системы ANDSystem) с использованием оригинальных скриптов на языке Python. Интернет-реализацию базы знаний выполняли с помощью MySQL 5.1.73 и PHP 5.3.3. В качестве серверов использовали Apache HTTP Server 2.2.15 и Nginx 1.4.1.

Результаты и обсуждение

Гены, ассоциированные с процессами терморегуляции

Поиск по базам данных Gene Ontology, EntrezGene и ANDCell (информационная компонента ANDSystem) выявил 467 генов, ассоциированных с терморегуляцией, кодирующих белки человека, а также два гена, кодирующих микроРНК.

Генная сеть терморегуляции

На основе списка, включающего упомянутые выше гены человека, вовлеченные в терморегуляцию, с использованием системы ANDSystem была реконструирована генная сеть терморегуляции. Общий вид реконструированной генной сети представлен на рис. 1. Генная сеть включала 469 генов, 473 белка, 265 микроРНК и 7018 взаимодействий между ними. Количество белков превышало количество генов, так как генная сеть содержала шесть генов, каждый из которых кодировал несколько белков за счет альтернативного сплайсинга либо протеолитического расщепления белка-предшественника.

Отметим, что система ANDSystem на основе анализа текстов статей и информации из медико-биологических баз данных выявляет два типа событий, значимых для функционирования генных сетей: прямые молекулярно-генетические взаимодействия между объектами генных сетей и события из разряда «непрямых связей», т. е. ситуации, когда показано влияние одного объекта генной сети на другой, но молекулярно-генетический механизм такого влияния остается неизвестным.

[1] Приложения S1–S7 см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx36.pdf

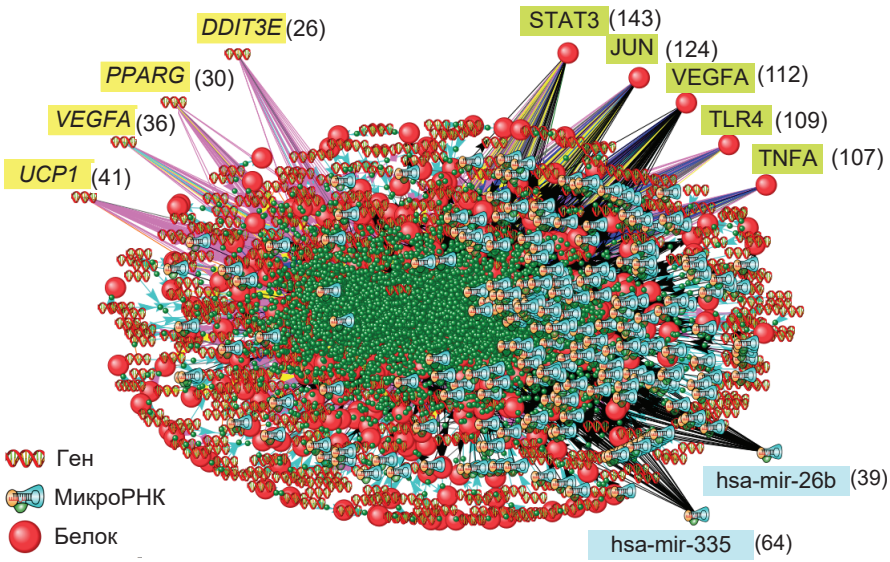


Рис. 1. Общий вид генной сети терморегуляции человека, реконструированной с помощью системы ANDSystem.
Генная сеть включает 469 генов, 473 белка, 265 микроРНК и 7018 взаимодействий между этими объектами. Отдельно от всех объектов сети показаны гены, белки и микроРНК, имеющие наибольшее количество связей в сети. Цифры в скобках обозначают количество связей в сети.

На рис. 2 продемонстрированы детали функционирования двух фрагментов генной сети терморегуляции. Молекулярно-генетические взаимодействия гена *TRPV1*, кодирующего терморецептор, реагирующий на повышение температуры, показаны на рис. 2, *а*. Согласно базе знаний ANDSystem, экспрессия гена *TRPV1* регулируется интерлейкином 13 (IL13) и толл-подобным рецептором 4-го типа (TLR4). И эти регуляторные события, описанные в статьях (Rehman et al., 2013; Li et al., 2015), относятся к классу «непрямых», поскольку речь идет о влиянии цитокина IL13 (внеклеточной сигнальной молекулы) и рецептора TLR4, расположенного на мембране клетки, которые оказывают влияние на экспрессию гена *TRPV1* через пути сигнальной трансдукции. Кроме того, ген *TRPV1* коэкспрессируется с другими генами из генной сети терморегуляции, включая гены терморецепторов (*TRPM8*, *TRPA1*, *TRPV3*, *TRPV4*), а также с геном нейротрофной рецепторной тирозинкиназы 1 (*NTRK1*). Детали экспериментов, в которых выявлена коэкспрессия этих генов, описаны в статьях (Zhu, Oxford, 2007; Cao et al., 2009; Cheng et al., 2012; Gouin et al., 2017; Nguyen et al., 2017).

На рис. 2, *б* представлены регуляторные связи гена *PPARG* и кодируемого им белка. Экспрессия *PPARG* регулируется белками-транскрипционными факторами ZN423, EGR1, CEBPB, которые оказывают действие на транскрипцию, взаимодействуя с участками ДНК в регуляторных районах гена *PPARG*. Экспрессия гена *PPARG* также регулируется транскрипционными кофакторами MECP2 и PRGC1/PGC-1-alpha и белком WN10B, который является участником Wnt-сигнального пути. В регуляции экспрессии гена *PPARG* также участвуют цитокины TNF, IL4, IL1B и LEP. Кодируемый геном *PPARG* белок-транскрипционный фактор контролирует транскрипцию: а) генов-регуляторов метаболических процессов в адипоцитах *LCN2*, *UCP1*, *FABP4*, *PNPLA2*, *SLC27A1*, *LIPE*, *DDIT3*; б) генов транскрипционных факторов *STAT3* и *ARNTL*; а также в) гена *SIRT6*, кодирующего фермент-деацетилазу (ссылки на научные публикации, подтверждающие эти взаимодействия, представлены в Приложении S3).

База знаний Termo_Reg_Human

В рамках исследования разработана база знаний Termo_Reg_Human 1.0 (https://www.sysbio.ru/ThermoReg_Human/). Эта база знаний содержит информацию о 469 генах, 473 белках и 265 микроРНК.

Termo_Reg_Human 1.0 состоит из четырех основных таблиц: *Genes_evol*, *Proteins*, *MicroRNA* и *Genes_all* (схема базы представлена на рис. 3).

Таблица *Genes_evol* содержит описание каждого из 469 генов, функционирующих в составе генной сети терморегуляции человека, включая: уникальный идентификатор NCBI, количество взаимодействий гена с другими генами и белками генной сети, источники экспериментальных данных, подтверждающие связь генов с терморегуляцией (Gene Ontology, ANDSystem, Entrez Gene). В этой таблице для каждого белок-кодирующего гена представлены также такие эволюционные характеристики, как филостратиграфический индекс (PAI) и индекс дивергенции (DI), вычисленные с помощью программного комплекса OrthoWeb (Иванов и др., 2024).

Таблица *Proteins* содержит информацию о белках, кодируемых генами из таблицы *Genes_evol*. Для каждого белка указан его идентификатор из базы UniProtKb, идентификатор гена, кодирующего этот белок из базы NCBI, количество связей белка в генной сети и названия микроРНК, регулирующих экспрессию белка.

Таблица *MicroRNA* содержит сведения о микроРНК, регулирующих экспрессию белков, участвующих в сети. Это две микроРНК, кодируемые генами из исходного

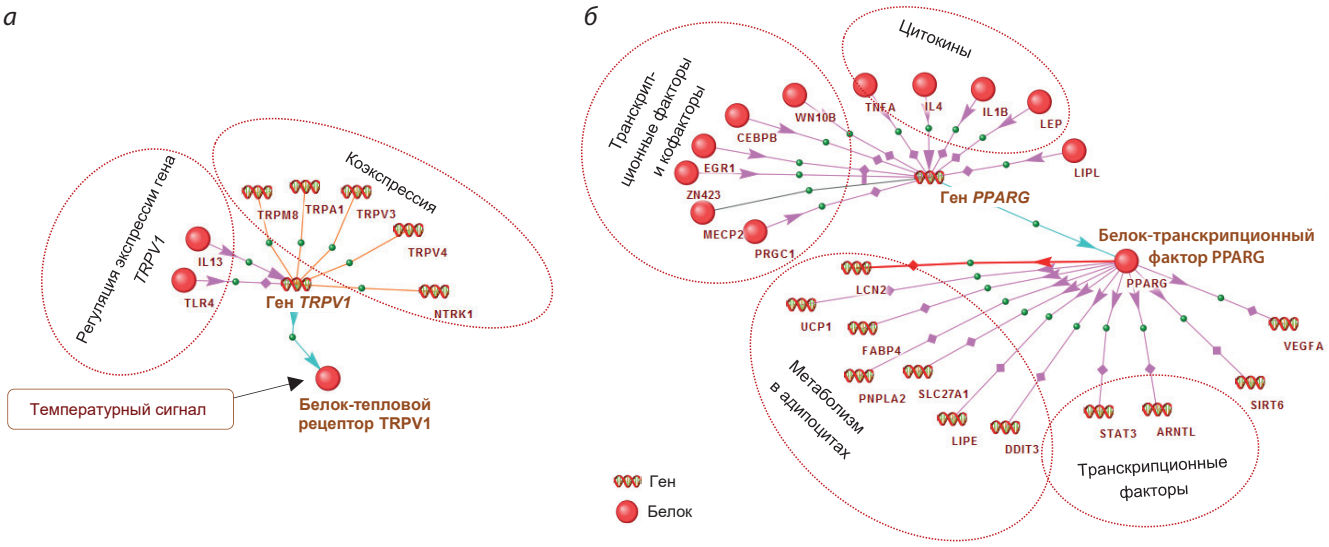


Рис. 2. Фрагменты генной сети терморегуляции, представленной на рис. 1.

а – регуляторные взаимодействия с участием гена, кодирующего тепловой рецептор TRPV1; *б* – регуляторные взаимодействия с участием гена *PPARG* и кодируемого им транскрипционного фактора PPARG.

списка 469 генов, а также другие микроРНК, найденные с помощью программы ANDVisio в ходе построения генной сети. В таблице *MicroRNA* приведены названия микроРНК в сети; идентификаторы NCBI генов, кодирующих данную микроРНК; количество связей микроРНК в сети и перечисление белков, для которых микроРНК выступают в качестве регуляторов экспрессии.

Четвертая таблица, *Genes_all*, содержит дополнительные сведения обо всех 469 генах, охарактеризованных в таблице *Genes_evol*, а также о генах, кодирующих микроРНК, включенных в сеть средствами программы ANDVisio.

Веб-интерфейс позволяет просматривать данные о генах и белках, ассоциированных с процессами терморегуляции, а также проводить поиск генов/белков по идентификаторам или их названиям. Кроме того, доступен поиск объектов (генов, белков, микроРНК) по количеству функциональных взаимодействий в сети – интерфейс отображает только те элементы, у которых количество связей в сети превышает указанное пользователем значение.

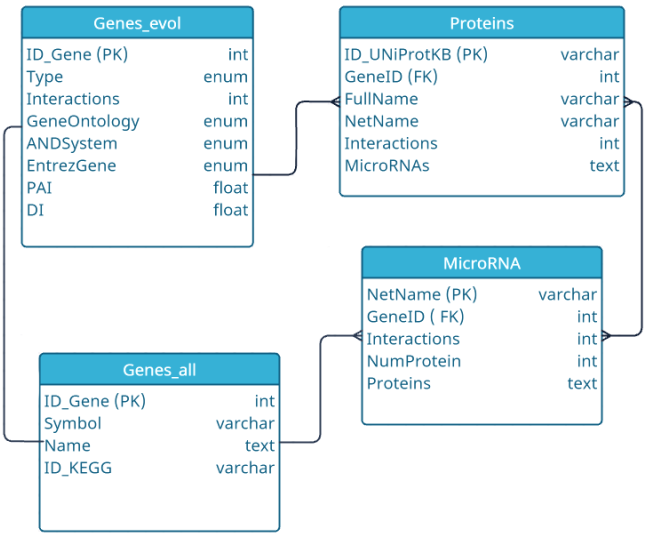


Рис. 3. Структура базы знаний Termo_Reg_Human 1.0.

Применение данных из базы знаний Termo_Reg_Human 1.0 в биоинформатических исследованиях

Приоритизация генов по количеству связей в генной сети терморегуляции. На рис. 4, *а* показано распределение генов по количеству связей с другими объектами генной сети терморегуляции (генами, белками и микроРНК). Большинство генов (373 из 467) связано с остальными объектами сети малым числом связей (пять и менее). Пятая часть, 90 генов, имеет от 6 до 25 связей. Количество связей больше 25 обнаружено у четырех генов: *UCP1* (41), *VEGFA* (36), *PPARG* (30) и *DDIT3* (26). На основе гипергеометрического распределения показано, что эти четыре гена имеют достоверно больше связей с остальными объектами генной сети, чем это можно ожидать по случайным причинам: величина P-adjusted варьирует от $2.44 \cdot 10^{-05}$ для гена *DDIT3* до $1.20 \cdot 10^{-28}$ для гена *UCP1*. Функциональная характеристика этих генов, имеющих наибольшее количество связей, приведена в табл. 1.

Ген *UCP1* кодирует белок термогенин (разобщающий белок 1), вовлеченный в один из ключевых процессов генерации тепла – несократительный термогенез в клетках бурого жира (Wollenberg Valero et al., 2014). Этот белок локализуется на внутренней мембране митохондрий. Он увеличивает проницаемость внутренней митохондриальной мембраны, уменьшая градиент протонов в цепи переноса электронов при дыхании митохондрий. В результате происходит разобщение процессов окислительного фосфорилирования и синтеза АТФ и выделяется тепло (Ikeda, Yamada, 2020).

Ген *VEGFA* кодирует фактор роста эндотелия кровеносных сосудов (Naik et al., 2012). Обусловленная этим

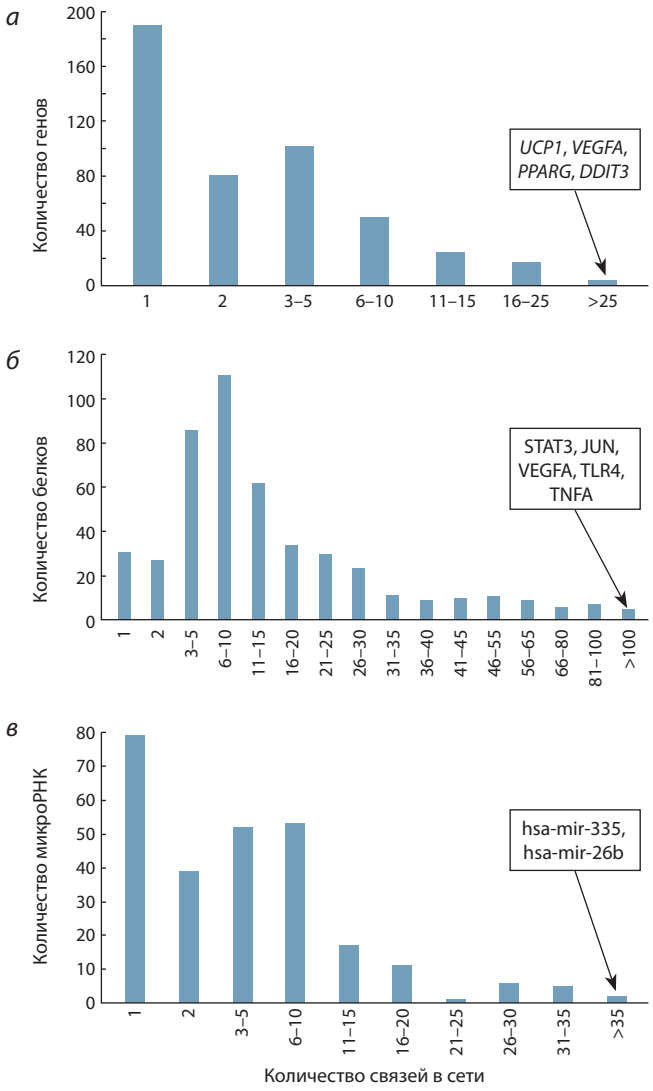


Рис. 4. Распределение генов, белков и микроРНК из генной сети терморегуляции по количеству связей в этой сети (на основе информации из базы знаний Termo_Reg_Human 1.0).

а – распределение генов по числу связей; *б* – распределение белков по числу связей; *в* – распределение микроРНК по числу связей. В прямоугольных панелях приведены названия генов, белков и микроРНК с максимальным количеством связей.

процессом активизация кровоснабжения тканей важна для реализации механизмов физической регуляции: теплообмена между внутренними частями тела и его поверхностью, отдачи тепла через испарение и конвекцию (Tansey, Johnson, 2015).

Ген *PPARG* кодирует транскрипционный фактор PPARG, относящийся к суперсемейству ядерных рецепторов. PPARG контролирует активность генов, управляющих метаболизмом жирных кислот и глюкозы (Festuccia et al., 2009), а также активирует образование белка-термогенина UCP1 в бурых и бежевых адипоцитах (Valdivia et al., 2023).

Ген *DDIT3* кодирует белок CHOP (C/EBP homologous protein), транскрипционный фактор семейства C/EBP, который участвует в регуляции дифференцировки клеток-предшественников в зрелые жировые клетки – адипоциты, играющие важнейшую роль в несократительном термогенезе (Okla et al., 2015).

Приоритизация белков по количеству связей в генной сети терморегуляции. Оказалось, что белки в анализируемой генной сети имеют больше связей, чем гены (см. рис. 4, *б*): доля белков, имеющих пять и менее связей в сети, составляла менее половины от их общего количества (144 из 473); 55 % белков (261 белок) имели от 6 до 30 связей, 13 % белков (63 белка) имели от 31 до 100 связей. Пять белков, STAT3, JUN, VEGFA, TLR4, TNFA, имели более 100 связей с другими объектами сети. На основе гипергеометрического распределения показано, что эти пять белков имеют достоверно большее количество связей с остальными объектами сети, чем можно ожидать по случайным причинам: P-adjusted варьировало от $2.04 \cdot 10^{-18}$ для белка TLR4 до $3.79 \cdot 10^{-43}$ для белка STAT3. Характеристика этих пяти белков приведена в табл. 2.

Белок STAT3 (143 связи в сети) – транскрипционный фактор, конечное звено JAK/STAT3-пути сигнальной трансдукции. STAT3 регулирует дифференцировку адипоцитов в момент индукции этого процесса (induction phase), а последующая инактивация JAK/STAT3 пути в этих клетках индуцирует экспрессию гена *UCP1* и превращение адипоцитов в клетки бурого жира (Song et al., 2022). Кроме того, STAT3 участвует в пути передачи сигналов, активируемых тепловым рецептором TRPV1 в об-

Таблица 1. Функциональные характеристики генов, имеющих наибольшее количество связей в сети терморегуляции

Название гена	Количество связей в генной сети	Роль в терморегуляции	P-adjusted	PAI
UCP1	41	Кодируемый геном *UCP1* белок экспрессируется в бурой жировой ткани и обеспечивает выработку тепла в процессе несократительного термогенеза (Wollenberg Valero et al., 2014; Ikeda, Yamada, 2020)	$1.2 \cdot 10^{-28}$	1
VEGFA	36	Кодируемый этим геном белок VEGFA регулирует васкуляризацию тканей, что обеспечивает теплообмен и теплоотдачу (Naik et al., 2012)	$1.8 \cdot 10^{-6}$	6
PPARG	30	Кодируемый геном *PPARG* белок регулирует дифференцировку адипоцитов, содержание жирных кислот и метаболизм глюкозы в жировых клетках (Festuccia et al., 2009)	$2.66 \cdot 10^{-7}$	6
DDIT3	26	Кодируемый геном *DDIT3* белок CHOP регулирует адипогенез (Okla et al., 2015)	$2.44 \cdot 10^{-6}$	7

Примечание. Гены перечислены в порядке убывания количества связей в генной сети.
Здесь и в табл. 2 и 3: P-adjusted указывает на вероятность наблюдать по случайным причинам данное количество связей в сети, рассчитанную по гипергеометрическому распределению с поправкой на множественные сравнения.

Таблица 2. Функциональные характеристики белков, имеющих наибольшее количество связей в генной сети терморегуляции

Белок	Количество связей в генной сети	Роль в терморегуляции	P-adjusted
STAT3	143	Транскрипционный фактор STAT3 регулирует экспрессию генов в областях мозга, контролирующих терморегуляцию (Yoshida et al., 2016); дифференцировку адипоцитов и их превращение в клетки бурого жира, а также экспрессию гена *UCP1* (Song et al., 2022)	$3.79 \cdot 10^{-43}$
JUN	124	Транскрипционный фактор JUN регулирует экспрессию цитокинов (Schonthaler et al., 2011; Johnson Rowsey, 2013), а также экспрессию гена *UCP1* в буром жире (Xiao et al., 2019)	$3.78 \cdot 10^{-33}$
VEGFA	112	Характеристика белка VEGFA приведена в табл. 1	$6.72 \cdot 10^{-28}$
TLR4	109	TLR4 – рецептор клеточной поверхности, активируемый липополисахаридами, что способствует повышению температуры тела (Roth, Blatteis, 2014), оказывает влияние на процесс дифференцировки клеток бурого жира (Okla et al., 2018)	$2.04 \cdot 10^{-18}$
TNFA*	107	TNFA – фактор некроза опухолей, цитокин, способен индуцировать лихорадку, сопровождающуюся повышением температуры тела (Leon et al., 1998; Netea et al., 2000; Gil et al., 2007), а также оказывает влияние на экспрессию генов в адипоцитах (Valladares et al., 2001; Kim et al., 2006)	$1.78 \cdot 10^{-30}$

Примечание. Белки перечислены в порядке убывания количества связей в генной сети.
* Этот белок является продуктом экспрессии гена *TNF*.

ластях мозга, контролирующих температуру тела (Yoshida et al., 2016).

Белок JUN (124 связи в сети) – субъединица транскрипционного фактора AP1 (гетеродимера JUN/FOS). JUN вовлечен в контроль экспрессии цитокинов, влияя тем самым на процесс воспаления, сопряженный с повышением температуры тела (Schonthaler et al., 2011; Johnson Rowsey, 2013). Показано, что при нарушении экспрессии гена *JUN* в печени мышей (нокаут по гену *JUN*) происходит повышение температуры тела за счет активации симпатической нервной системы и последующей стимуляции экспрессии белка UCP1 в буром жире (Xiao et al., 2019).

Как указано выше, белок VEGFA (vascular endothelial growth factor A), имеющий 112 связей в сети, контролирует рост эндотелия сосудов (Naik et al., 2012), что важно для реализации механизмов физической терморегуляции (Tansey, Johnson, 2015).

TLR4 (109 связей в сети) – трансмембранный белок, толл-подобный рецептор 4. Он может активироваться липополисахаридами клеточной стенки бактерий, что приводит к повышению температуры тела в ответ на инфекцию (Roth, Blatteis, 2014). Показано, что активация TLR4-рецептора липополисахаридами способствует окислительному стрессу и дисфункции митохондрий, вызывая замедление дифференцировки клеток бурого жира (Okla et al., 2018).

Белок TNFA – фактор некроза опухолей, относится к семейству цитокинов (107 связей в сети). Он активирует, в частности, синтез простагландинов в эндотелиальных клетках. Воздействие простагландинов на нейроны, расположенные в преоптической зоне гипоталамуса (центре терморегуляции), способствует повышению температуры тела (Leon et al., 1998; Netea et al., 2000; Gil et al., 2007; Nakamura, 2024). Показано также прямое действие TNFA на адипоциты *in vitro*, снижающее экспрессию белка-термогенина UCP-1 (Valladares et al., 2001) и фермента триглицерид-липазы ATGL/PNPLA2 (Kim et al., 2006). Таким образом, цитокин TNFA играет важную роль в терморегуляции, однако его влияние на температуру тела зависит от типа клеток, на которые этот цитокин оказывает свое действие.

Приоритизация микроРНК по количеству связей в генной сети терморегуляции. МикроРНК регулируют экспрессию генов на уровне трансляции. Эти РНК взаимодействуют с мРНК в составе miRISC комплекса, что тормозит процесс трансляции и может вызвать деградацию мРНК (O'Brien et al., 2018). Согласно данным базы знаний Termo_Reg_Human 1.0, генная сеть терморегуляции включает 265 микроРНК, участвующих в регуляции экспрессии 297 генов. Сведения об этих регуляторных связях получены из базы данных miRTarBase, в которой содержатся экспериментально подтвержденные данные о взаимодействиях между микроРНК и их мРНК-мишенями (Cui et al., 2025). Доля микроРНК, имеющих пять и менее регуляторных связей в сети, составляла 64 % (170 из 265) (см. рис. 4, *в*). 35 % микроРНК (93 из 265) имели от 6 до 30 связей. Две микроРНК обладали максимальным количеством связей (более 35): hsa-mir-335 (64 связи) и hsa-mir-26b (39 связей). Оценка статистической значимости количества связей между этими микроРНК и другими вершинами сети средствами программы ANDVisio показала, что микроРНК hsa-mir-335 и hsa-mir-26b регулируют экспрессию значимо большего количества генов из сети терморегуляции, чем это можно ожидать по случайным причинам (P-adjusted < 0.01).

Две указанные выше микроРНК выполняют важные функции в процессах терморегуляции (табл. 3). Так, hsa-mir-335 регулирует экспрессию терморецепторов TRPM8 и TRPV4, а также белка VEGFA, одного из ключевых для терморегуляции, имеющего 112 связей в сети. МикроРНК

Таблица 3. Функциональные характеристики наиболее значимых микроРНК, участвующих в генной сети терморегуляции

микроРНК	Ген, кодирующий микроРНК	Количество связей в сети	P-adjusted	Регулируемые мРНК*	Примеры функционально значимых белков, кодируемых мРНК, с которыми взаимодействуют микроРНК
			микроРНК с наибольшим количеством связей в сети		
hsa-mir-335	*MIR3*	64	< 0.001	**TRPM8**, **TRPV4**, **VEGFA**, ANO1, ANO3, NPR3, AQP5, ARRDC3, ACVR2B, BAAT, CASQ1, CD14, CD36, CDKN1A, CRNN, DDIT3, DNAJC3, DBH, EIF2AK3, ELOVL6, FABP4, FOS, FOXO1, ABAT, GRB10, HDAC6, HMOX1, HSPA1A, HSPA1B, HSPB3, IGF2BP2, IGF1R, NFKBIA, IL1A, IL4, JAK2, KCNK4, KDM6B, LEPR, MOCOS, AVP, NOS3, NPY, NR1D1, NR2F6, NTSR1, PLA2G7, PTGS2, PPARGC1A, PTGES, RB1, SLC27A1, SCARA5, SCN9A, SQSTM1, STAT6, TCIM, TFE3, PTH2, TAC4, TMEM135, NGFR, TSHR, WNT10B	Терморецепторы TRPV4 и TRPM8, а также ростовой фактор VEGFA, имеющий 112 связей в сети
hsa-mir-26b	*MIR26b*	39	< 0.01	**JUN**, **EDN2**, ACADM, ADRA2A, AGTR1, AKT1, BAG3, CASP8, CASQ1, CAV1, CD36, STUB1, CHORDC1, CRYAA, CXCR4, DNAJA2, DNAJA3, DNAJB4, EIF2AK3, EIF2B1, GRIK2, HADH, HMOX1, HSF1, IER5, NOX3, NRDC, NTSR1, PARK7, PDCL3, PTGS2, RBM3, RRAGC, SLC25A44, SMS, STAT6, VCP, TNFRSF11A, ZNF423	Транскрипционный фактор JUN, имеющий 124 связи в сети, белок-вазоконстриктор EDN2 (Endothelin-2)
			микроРНК, кодируемые генами из исходного списка		
hsa-mir-21	*MIR21*	15	< 0.05	**VEGFA**, PRKAB2, ALMS1, APC, CPEB3, DAXX, DOCK7, EIF2S1, IL1B, PARP1, RB1, RDH11, RRAGC, SMARCA4, STAT3	Ростовой фактор VEGFA, имеющий 112 связей в сети
hsa-let-7c	*MIRLET7c*	5	> 0.05	**MT-CO2/COX2**, **DICER1**, **CNOT3**, IP6K1, QKI	Белок COX2, участвующий в транспорте электронов в митохондриях (Aich et al., 2018), белки DICER1 и CNOT3/NOT, участвующие в микроРНК-опосредованной деградации мРНК (Wingo et al., 2015; Wakiyama, Takimoto, 2022)

* мРНК, трансляция которых регулируется при участии данной микроРНК (подчеркнуты и выделены жирным шрифтом мРНК, кодирующие белки, функции которых охарактеризованы в правой колонке).

hsa-mir-26b регулирует экспрессию белка JUN, имеющего 124 связи в сети, который, как отмечалось выше, влияет на экспрессию белка-термогенина UCP1 в клетках бурого жира (Xiao et al., 2019). Эта микроРНК также регулирует экспрессию белка EDN2 (Endothelin-2), вовлеченного в процесс физической терморегуляции – сужение сосудов (Inoue et al., 1989).

В сформированном нами списке генов терморегуляции присутствуют гены *MIR21* и *MIRLET7c*. Кодируемые этими генами микроРНК hsa-mir-21 и hsa-let-7c регулируют клеточные процессы в ответ на повышение температуры (Jiang et al., 2016; Permenter et al., 2019). В реконструированной нами генной сети выявлено влияние микроРНК hsa-mir-21 и hsa-let-7c на экспрессию 15 и 5 белков соответственно (см. табл. 3).

В числе белков, экспрессия которых регулируется при участии микроРНК hsa-mir-21, найден белок-ростовой фактор VEGFA, имеющий 112 связей в сети (см. табл. 3). Многократное упоминание этого белка в тексте статьи подчеркивает его важный вклад в процессы терморегуляции. А среди белков, экспрессия которых контролируется микроРНК hsa-let-7c, идентифицированы: а) белок COX2 – субъединица цитохром с-оксидазы, осуществляющей транспорт электронов в митохондриях, кодируемая геном *MT-CO2* (Aich et al., 2018); б) белок DICER1, осуществляющий биогенез микроРНК (Wingo et al., 2015); в) белок CNOT3/NOT, участвующий в микроРНК-опосредованной деградации мРНК (Wakiyama, Takimoto, 2022).

Эволюционный возраст генов генной сети терморегуляции (анализ на основе индекса PAI). Оценку эволюционного возраста генов проводили с использованием индекса PAI, данные о котором содержатся в поле PAI информационной таблицы *Genes_evol* из базы знаний Termo_Reg_Human 1.0. Индекс PAI, вычисляемый с помощью системы Orthoweb (Иванов и др., 2024), был предложен в наших предыдущих исследованиях (Mustafin et al., 2017). Мы построили распределение по значениям индексов PAI для 467 белок-кодирующих генов (выборка

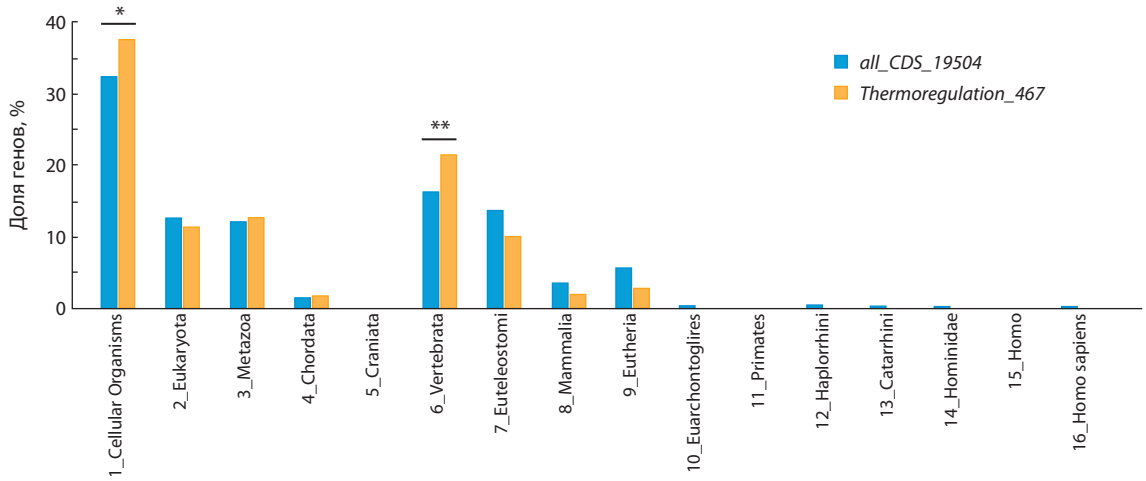


Рис. 5. Распределение значений индекса PAI для группы белок-кодирующих генов, связанных с терморегуляцией (*Thermoregulation_467*) и всех белок-кодирующих генов человека (набор *all_CDS_19504*).

Одна звездочка (*) показывает достоверное ($p < 0.05$) превышение наблюдаемого количества генов терморегуляции, соответствующих PAI = 1 (корень филостратиграфического дерева, одноклеточные организмы), над ожидаемым количеством этих генов, рассчитанным по полной выборке белок-кодирующих генов (*all_CDS_19504*). Две звездочки (**) показывают достоверное ($p < 0.01$) превышение наблюдаемого количества генов терморегуляции, соответствующих PAI = 6 (этап возникновения и дивергенции Vertebrata), над ожидаемым количеством.

Thermoregulation_467), функционирующих в генной сети терморегуляции, описанных в базе знаний Termo_Reg_Human 1.0 (на рис. 5 это распределение отмечено оранжевыми столбиками). Оказалось, что данное распределение имеет два максимума. Первый из них наблюдается при значении PAI = 1 (176 генов терморегуляции, 38 % от их полного списка). Филостратиграфический индекс PAI = 1 соответствует эволюционному этапу возникновения одноклеточных организмов. Второй пик наблюдается при значении PAI = 6 (100 генов терморегуляции, 22 % от их полного списка). Филостратиграфический индекс PAI = 6 соответствует эволюционному этапу возникновения и дивергенции позвоночных.

Для оценки статистической значимости двух описанных выше пиков было построено распределение величины индекса PAI для всех белок-кодирующих генов человека (19 504 гена, выборка *all_CDS_19504*), отмеченное на рис. 5 голубым цветом, как это сделано в работе (Михайлова и др., 2024). Это распределение также имеет два, но менее выраженных, пика. С использованием метода хи-квадрат проведено сравнение количества генов терморегуляции, попадающих в пики 1 и 6, с количеством генов, ожидаемых по случайным причинам в этих пиках. В обоих случаях обнаружено различие сравниваемых наблюдаемых и ожидаемых количеств с уровнями значимости $p < 0.05$ и $p < 0.01$ (Приложения S4 и S5). Таким образом, показано обогащение генной сети терморегуляции генами, предковые варианты которых сформировались на эволюционном этапе появления одноклеточных организмов и этапе дивергенции позвоночных.

Анализ функций генов терморегуляции, проведенный с использованием инструмента DAVID, показал, что группа генов со значением индекса PAI = 1 обогащена связями с терминами Gene Ontology, обозначающими регуляцию транскрипции (см. Приложение S6) – важнейший механизм регуляции экспрессии генов у одноклеточных. Что касается группы генов со значением индекса PAI = 6, то она обогащена генами, вовлеченными в процессы сигнальной трансдукции – важнейшего звена молекулярно-генетических механизмов взаимодействия между клетками (см. Приложение S7). Этот результат согласуется с представлением, что в процессах терморегуляции многоклеточного организма человека важнейшую роль играют взаимодействия большого количества физиологических систем организма: дыхательной, кровеносной, мышечной, нервной и т. д. (Tansey, Johnson, 2015; Nakamura, 2024). При этом процесс транскрипции обеспечивает генетический контроль при формировании клеток и тканей, вовлеченных в терморегуляцию, а координация деятельности физиологических систем, обеспечивающих терморегуляцию, осуществляется на клеточном уровне через процессы сигнальной трансдукции.

Заключение

В настоящей работе реконструирована генная сеть терморегуляции человека, включающая гены, микроРНК и белки человека, вовлеченные в терморегуляцию. Создана база знаний Termo_Reg_Human 1.0, в которой систематизированы данные о молекулярно-генетических механизмах регуляции этого процесса. На основе анализа информации из базы знаний осуществлена приоритизация генов, белков и микроРНК по количеству связей в сети терморегуляции, а также обнаружены особенности эволюции генов терморегуляции.

Выявлено обогащение генной сети терморегуляции генами, предковые варианты которых сформировались на эволюционных этапах появления одноклеточных организмов и дивергенции позвоночных. Обнаруженные нами закономерности в эволюции генов сети терморегуляции следует учитывать при разработке новых концепций возникновения эндотермии у животных разных таксонов (Osvath et al., 2024).

Список литературы / References

Иванов Р.А., Мухин А.М., Казанцев Ф.В., Мустафин З.С., Афонников Д.А., Матушкин Ю.Г., Лашин С.А. Orthoweb: программный комплекс для эволюционного анализа генных сетей. *Вавиловский журнал генетики и селекции.* 2024;28(8):874-881. doi 10.18699/vjgb-24-95
[Ivanov R.A., Mukhin A.M., Kazantsev F.V., Mustafin Z.S., Afonnikov D.A., Matushkin Y.G., Lashin S.A. Orthoweb: a software package for evolutionary analysis of gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2024;28(8): 874-881. doi 10.18699/vjgb-24-95]

Михайлова А.Д., Лашин С.А., Иванисенко В.А., Деменков П.С., Игнатьева Е.В. Реконструкция и компьютерный анализ структурно-функциональной организации генной сети регуляции биосинтеза холестерина у человека и эволюционная характеристика участвующих в ней генов. *Вавиловский журнал генетики и селекции*. 2024;28(8):864-873. doi 10.18699/vjgb-24-94
[Mikhailova A.D., Lashin S.A., Ivanisenko V.A., Demenkov P.S., Ignatieva E.V. Reconstruction and computer analysis of the structural and functional organization of the gene network regulating cholesterol biosynthesis in humans and the evolutionary characteristics of the genes involved in the network. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2024;28(8):864-873. doi 10.18699/vjgb-24-94]

Мустафин З.С., Лашин С.А., Матушкин Ю.Г. Филостратиграфический анализ генных сетей заболеваний человека. *Вавиловский журнал генетики и селекции*. 2021;25(1):46-56. doi 10.18699/VJ21.006
[Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2021;25(1):46-56. doi 10.18699/VJ21.006]

Aich A., Wang C., Chowdhury A., Ronsör C., Pacheu-Grau D., Richter-Dennerlein R., Dennerlein S., Rehling P. COX16 promotes COX2 metallation and assembly during respiratory complex IV biogenesis. *eLife*. 2018;7:e32572. doi 10.7554/eLife.32572

Benjamini Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4)1165-1188. doi 10.1214/aos/1013699998

Blondin D.P., Haman F. Shivering and nonshivering thermogenesis in skeletal muscles. *Handb Clin Neurol*. 2018;156:153-173. doi 10.1016/B978-0-444-63912-7.00010-2

Cao D.S., Yu S.Q., Premkumar L.S. Modulation of transient receptor potential Vanilloid 4-mediated membrane currents and synaptic transmission by protein kinase C. *Mol Pain.* 2009;5:5. doi 10.1186/1744-8069-5-5

Charkoudian N., Hart E.C.J., Barnes J.N., Joyner M.J. Autonomic control of body temperature and blood pressure: influences of female sex hormones. *Clin Auton Res*. 2017;27(3):149-155. doi 10.1007/s10286-017-0420-z

Cheng W., Yang F., Liu S., Colton C.K., Wang C., Cui Y., Cao X., Zhu M.X., Sun C., Wang K., Zheng J. Heteromeric heat-sensitive transient receptor potential channels exhibit distinct temperature and chemical response. *J Biol Chem*. 2012;287(10):7279-7288. doi 10.1074/jbc.M111.305045

Cui S., Yu S., Huang H.Y., Lin Y.C.D., Huang Y., Zhang B., Xiao J., … Chen B., Zhang H., Fu J., Wang L., Huang H.-D. miRTarBase 2025: updates to the collection of experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 2025;53(D1):D147-D156. doi 10.1093/nar/gkae1072

Dumont L., Richard G., Espagnet R., Frisch F., Fortin M., Samson A., Bouchard J., … Dubreuil S., Guérin B., Turcotte É.E., Carpentier A.C., Blondin D.P. Shivering, but not adipose tissue thermogenesis, increases as a function of mean skin temperature in cold-exposed men and women. *Cell Metab*. 2025;37(9):1789-1805.e4. doi 10.1016/j.cmet.2025.06.010

Festuccia W.T., Blanchard P.-G., Turcotte V., Laplante M., Sariahmetoglu M., Brindley D.N., Deshaies Y. Depot-specific effects of the PPARγ agonist rosiglitazone on adipose tissue glucose uptake and metabolism. *J Lipid Res.* 2009;50(6):1185-1194. doi 10.1194/jlr.M800620-JLR200

Gil A., María Aguilera C., Gil-Campos M., Cañete R. Altered signalling and gene expression associated with the immune system and the inflammatory response in obesity. *Br J Nutr.* 2007;98(Suppl. 1): S121-S126. doi 10.1017/S0007114507838050

Gouin O., L'Herondelle K., Lebonvallet N., Le Gall-Ianotto C., Sakka M., Buhé V., Plée-Gautier E., Carré J.L., Lefeuvre L., Misery L., Le Garrec R. TRPV1 and TRPA1 in cutaneous neurogenic and chronic inflammation: pro-inflammatory response induced by their activation and their sensitization. *Protein Cell.* 2017;8(9):644-661. doi 10.1007/s13238-017-0395-5

Horii Y., Shiina T., Uehara S., Nomura K., Shimaoka H., Horii K., Shimizu Y. Hypothermia induces changes in the alternative splicing pattern of cold-inducible RNA-binding protein transcripts in a non-hibernator, the mouse. *Biomed Res.* 2019;40(4):153-161. doi 10.2220/biomedres.40.153

Ignatieva E.V., Igoshin A.V., Yudin N.S. A database of human genes and a gene network involved in response to tick-borne encephalitis virus infection. *BMC Evol Biol.* 2017;17(Suppl. 2):259. doi 10.1186/s12862-017-1107-8

Ikeda K., Yamada T. UCP1 dependent and independent thermogenesis in brown and beige adipocytes. *Front Endocrinol* (*Lausanne*). 2020; 11:498. doi 10.3389/fendo.2020.00498

Inoue A., Yanagisawa M., Kimura S., Kasuya Y., Miyauchi T., Goto K., Masaki T. The human endothelin family: three structurally and pharmacologically distinct isopeptides predicted by three separate genes. *Proc Natl Acad Sci USA*. 1989;(8):2863-2867. doi 10.1073/pnas.86.8.2863

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Jiang T., Wang X., Wu W., Zhang F., Wu S. Let-7c miRNA inhibits the proliferation and migration of heat-denatured dermal fibroblasts through down-regulating HSP70. *Mol Cells*. 2016;39(4):345-351. doi 10.14348/molcells.2016.2336

Johnson Rowsey P. Thermoregulation: cytokines involved in fever and exercise. *Annu Rev Nurs Res.* 2013;31:19-46. doi 10.1891/0739-6686.31.19

Kim J.Y., Tillison K., Lee J.H., Rearick D.A., Smas C.M. The adipose tissue triglyceride lipase ATGL/PNPLA2 is downregulated by insulin and TNF-α in 3T3-L1 adipocytes and is a target for transactivation by PPARγ. *Am J Physiol Endocrinol Metab.* 2006;291(1): E115-E127. doi 10.1152/ajpendo.00317.2005

Kudsi S.Q., Piccoli B.C., Ardisson-Araújo D., Trevisan G. Transcriptional landscape of *TRPV1*, *TRPA1*, *TRPV4*, and *TRPM8* channels throughout human tissues. *Life Sci.* 2022;308:120977. doi 10.1016/j.lfs.2022.120977

Leon L.R., White A.A., Kluger M.J. Role of IL-6 and TNF in thermoregulation and survival during sepsis in mice. *Am J Physiol.* 1998;275(1):R269-R277. doi 10.1152/ajpregu.1998.275.1.R269

Li Y., Adamek P., Zhang H., Tatsui C.E., Rhines L.D., Mrozkova P., Li Q., … Jawad A.B., Ghetti A., Yan J., Palecek J., Dougherty P.M. The cancer chemotherapeutic paclitaxel increases human and rodent sensory neuron responses to TRPV1 by activation of TLR4. *J Neurosci*. 2015;35(39):13487-13500. doi 10.1523/jneurosci.1956-15.2015

McCafferty D.J., Pandraud G., Gilles J., Fabra-Puchol M., Henry P.Y. Animal thermoregulation: a review of insulation, physiology and

behaviour relevant to temperature control in buildings. *Bioinspir Biomim.* 2017;13(1):011001. doi 10.1088/1748-3190/aa9a12

Mittag J., Kolms B. Hypothalamic control of heart rate and body temperature by thyroid hormones. *Rev Endocr Metab Disord*. 2025. doi 10.1007/s11154-025-09966-5

Mustafin Z.S., Lashin S.A., Matushkin Y.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics*. 2017;18(Suppl. 1):1427. doi 10.1186/s12859-016-1427-5

Mustafin Z.S., Zamyatin V.I., Konstantinov D.K., Doroshkov A.V., Lashin S.A., Afonnikov D.A. Phylostratigraphic analysis shows the earliest origination of the abiotic stress associated genes in *A. thaliana*. *Genes*. 2019;10(12):963. doi 10.3390/genes10120963

Naik N.A., Bhat I.A., Afroze D., Rasool R., Mir H., Andrabi S.I., Shah S., Siddiqi M.A., Shah Z.A. Vascular endothelial growth factor A gene (VEGFA) polymorphisms and expression of VEGFA gene in lung cancer patients of Kashmir Valley (India). *Tumour Biol.* 2012;33(3):833-839. doi 10.1007/s13277-011-0306-y

Nakamura K. Central circuitries for body temperature regulation and fever. *Am J Physiol Regul Integr Comp Physiol.* 2011;301(5): R1207-R1228. doi 10.1152/ajpregu.00109.2011

Nakamura K. Central mechanisms of thermoregulation and fever in mammals. *Adv Exp Med Biol.* 2024;1461:141-159. doi 10.1007/978-981-97-4584-5_10

Netea M.G., Kullberg B.J., Van der Meer J.W. Circulating cytokines as mediators of fever. *Clin Infect Dis.* 2000;31(Suppl. 5):S178-S184. doi 10.1086/317513

Nguyen M.Q., Wu Y., Bonilla L.S., von Buchholtz L.J., Ryba N.J.P. Diversity amongst trigeminal neurons revealed by high throughput single cell sequencing. *PLoS One.* 2017;12(9):e0185543. doi 10.1371/journal.pone.0185543

O'Brien J., Hayder H., Zayed Y., Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front Endocrinol* (*Lausanne*). 2018;9:402. doi 10.3389/fendo.2018.00402

Okla M., Wang W., Kang I., Pashaj A., Carr T., Chung S. Activation of Toll-like receptor 4 (TLR4) attenuates adaptive thermogenesis via endoplasmic reticulum stress. *J Biol Chem.* 2015;290(44):26476-26490. doi 10.1074/jbc.M115.677724

Okla M., Zaher W., Alfayez M., Chung S. Inhibitory effects of Toll-like receptor 4, NLRP3 inflammasome, and interleukin-1β on white adipocyte browning. *Inflammation.* 2018;41(2):626-642. doi 10.1007/s10753-017-0718-y

Osvath M., Němec P., Brusatte S.L., Witmer L.M. Thought for food: the endothermic brain hypothesis. *Trends Cogn Sci.* 2024;28(11):998-1010. doi 10.1016/j.tics.2024.08.002

Permenter M.G., McDyre B.C., Ippolito D.L., Stallings J.D. Alterations in tissue microRNA after heat stress in the conscious rat: potential biomarkers of organ-specific injury. *BMC Genomics*. 2019;20(1): 141. doi 10.1186/s12864-019-5515-6

Rehman R., Bhat Y.A., Panda L., Mabalirajan U. TRPV1 inhibition attenuates IL-13 mediated asthma features in mice by reducing airway epithelial injury. *Int Immunopharmacol.* 2013;15(3):597-605. doi 10.1016/j.intimp.2013.02.010

Roth J., Blatteis C.M. Mechanisms of fever production and lysis: lessons from experimental LPS fever. *Compr Physiol.* 2014;4(4): 1563-1604. doi 10.1002/cphy.c130033

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Dosenko V.E., Zolotareva O.I., Choynzonov E.L., Hofestaedt R., Ivanisenko V.A. Search for new candidate genes involved in the comorbidity of asthma and hypertension based on automatic analysis of scientific literature. *J Integr Bioinform.* 2018;15(4):20180054. doi 10.1515/jib-2018-0054

Schonthaler H.B., Guinea-Viniegra J., Wagner E.F. Targeting inflammation by modulating the Jun/AP-1 pathway. *Ann Rheum Dis.* 2011; 70(Suppl. 1):i109-i112. doi 10.1136/ard.2010.140533

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Song L., Cao X., Ji W., Zhao L., Yang W., Lu M., Yang J. Inhibition of STAT3 enhances UCP1 expression and mitochondrial function in brown adipocytes. *Eur J Pharmacol.* 2022;926:175040. doi 10.1016/j.ejphar.2022.175040

Tansey E.A., Johnson C.D. Recent advances in thermoregulation. *Adv Physiol Educ.* 2015;39(3):139-148. doi 10.1152/advan.00126.2014

Tattersall G.J., Sinclair B.J., Withers P.C., Fields P.A., Seebacher F., Cooper C.E., Maloney S.K. Coping with thermal challenges: physiological adaptations to environmental temperatures. *Compr Physiol*. 2012;2(3):2151-2202. doi 10.1002/cphy.c110055

Valdivia L.F.G., Castro É., Eichler R.A.D.S., Moreno M.F., de Sousa É., Jardim G.F.R., Peixoto Á.S., Moraes M.N., Castrucci A.M.L., Nedergaard J., Petrovic N., Festuccia W.T., Reckziegel P. Cold acclimation and pioglitazone combined increase thermogenic capacity of brown and white adipose tissues but this does not translate into higher energy expenditure in mice. *Am J Physiol Endocrinol Metab*. 2023;324(4):E358-E373. doi 10.1152/ajpendo.00217.2022

Valladares A., Roncero C., Benito M., Porras A. TNF-α inhibits UCP-1 expression in brown adipocytes via ERKs. Opposite effect of p38MAPK. *FEBS Lett.* 2001;493(1):6-11. doi 10.1016/s0014-5793(01)02264-5

Wakiyama M., Takimoto K. N-terminal Ago-binding domain of GW182 contains a tryptophan-rich region that confer binding to the CCR4-NOT complex. *Genes Cells*. 2022;27(9):579-585. doi 10.1111/gtc.12974

Wingo A.P., Almli L.M., Stevens J.S., Klengel T., Uddin M., Li Y., Bustamante A.C., … Bradley B., Binder E.B., Jin P., Gibson G., Ressler K.J. DICER1 and microRNA regulation in post-traumatic stress disorder with comorbid depression. *Nat Commun*. 2015;6: 10106. doi 10.1038/ncomms10106

Wollenberg Valero K.C., Pathak R., Prajapati I., Bankston S., Thompson A., Usher J., Isokpehi R.D. A candidate multimodal functional genetic network for thermal adaptation. *PeerJ*. 2014;2:e578. doi 10.7717/peerj.578

Xiao F., Guo Y., Deng J., Yuan F., Xiao Y., Hui L., Li Y., … Chen Y., Ying H., Zhai Q., Chen S., Guo F. Hepatic c-Jun regulates glucose metabolism via FGF21 and modulates body temperature through the neural signals. *Mol Metab.* 2019;20:138-148. doi 10.1016/j.molmet.2018.12.003

Yoshida A., Furube E., Mannari T., Takayama Y., Kittaka H., Tominaga M., Miyata S. TRPV1 is crucial for proinflammatory STAT3 signaling and thermoregulation-associated pathways in the brain during inflammation. *Sci Rep.* 2016;6:26088. doi 10.1038/srep26088

Zhu W., Oxford G.S. Phosphoinositide-3-kinase and mitogen activated protein kinase signaling pathways mediate acute NGF sensitization of TRPV1. *Mol Cell Neurosci.* 2007;34(4):689-700. doi 10.1016/j.mcn.2007.01.005

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 18.09.2025. После доработки 13.10.2025. Принята к публикации 13.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1020-1030

doi 10.18699/vjgb-25-107

Применение программно-информационной системы ANDSystem для поиска мишеней таргетной терапии ревматоидного артрита на основе анализа биологических процессов

Е.Л. Мищенко[1], И.В. Яцык[1] ✉, П.С. Деменков[1], А.В. Адамовская[1], Т.В. Иванисенко[1], М.А. Клещев[1], В.А. Иванисенко[1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ woikin88@mail.ru

Аннотация. Ревматоидный артрит (РА) – системное аутоиммунное заболевание, сопровождающееся поражением преимущественно суставов с прогрессирующей деструкцией хрящевой и костной тканей. До настоящего времени РА остается неизлечимым заболеванием, приводящим к значительному ухудшению качества жизни и инвалидизации пациентов. Несмотря на наличие широкого арсенала базисных противовоспалительных препаратов, около 40 % пациентов демонстрируют недостаточный ответ на стандартное лечение, что подчеркивает острую необходимость поиска новых фармакологических мишеней. Целью настоящей работы был поиск новых биологических процессов, которые могут служить перспективными мишенями для таргетной терапии РА. Для достижения поставленной цели был применен подход, основанный на автоматическом извлечении знаний из текстов научных публикаций и биомедицинских баз данных с помощью программно-информационной системы ANDSystem. Данный подход включал реконструкцию и последующий анализ ассоциативных генных сетей двух типов: а) генные сети, описывающие гены и белки, ассоциированные с развитием РА, и б) генные сети, описывающие гены и белки, вовлеченные в функциональные ответы на действие лекарств, применяемых для терапии заболевания. В результате анализа реконструированных сетей выявлено 11 биологических процессов, играющих значимую роль в патогенезе ревматоидного артрита, но до сих пор не являющихся прямыми мишенями существующих базисных противовоспалительных препаратов. К числу наиболее перспективных относятся следующие процессы, описываемые терминами онтологии генов: а) сигнальный путь Toll-подобных рецепторов; б) активация нейтрофилов; в) регуляция дифференцировки остеобластов; г) регуляция дифференцировки остеокластов; д) биосинтез простагландинов; е) канонический сигнальный путь Wnt. Выявленные биологические процессы и их ключевые регуляторы представляют собой перспективные мишени для разработки новых лекарственных средств, способных повысить эффективность терапии РА, в том числе у пациентов, резистентных к существующим методам лечения. Разработанный подход может быть успешно использован для поиска новых мишеней таргетной терапии и при других заболеваниях.

Ключевые слова: ревматоидный артрит; генные сети; таргетная терапия; ANDSystem

Для цитирования: Мищенко Е.Л., Яцык И.В., Деменков П.С., Адамовская А.В., Иванисенко Т.В., Клещев М.А., Иванисенко В.А. Применение программно-информационной системы ANDSystem для поиска мишеней таргетной терапии ревматоидного артрита на основе анализа биологических процессов. *Вавиловский журнал генетики и селекции*. 2025;29(7):1020-1030. doi 10.18699/vjgb-25-107

Финансирование. Работа выполнена в рамках бюджетного проекта FWNR-2022-0020.

Searching for biological processes as targets for rheumatoid arthritis targeted therapy with ANDSystem, an integrated software and information platform

E.L. Mishchenko[1], I.V. Yatsyk[1] ✉, P.S. Demenkov[1], A.V. Adamovskaya[1], T.V. Ivanisenko[1], M.A. Kleshchev[1], V.A. Ivanisenko[1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ woikin88@mail.ru

Abstract. Rheumatoid arthritis (RA) is a systemic autoimmune disease characterized primarily by joint involvement with progressive destruction of cartilage and bone tissue. To date, RA remains an incurable disease that leads to a significant deterioration in quality of life and patient disability. Despite a wide arsenal of disease-modifying antirheumatic drugs,

© Мищенко Е.Л., Яцык И.В., Деменков П.С., Адамовская А.В., Иванисенко Т.В., Клещев М.А., Иванисенко В.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

approximately 40 % of patients show an insufficient response to standard treatment, highlighting the urgent need to identify new pharmacological targets. The aim of this study was to search for novel biological processes that could serve as promising targets for the targeted therapy of RA. To achieve this goal, we employed an approach based on the automated extraction of knowledge from scientific publications and biomedical databases using the ANDSystem software. This approach involved the reconstruction and subsequent analysis of two types of associative gene networks: a) gene networks describing genes and proteins associated with the development of RA, and b) gene networks describing genes and proteins involved in the functional responses to drugs used for the disease's therapy. The analysis of the reconstructed networks identified 11 biological processes that play a significant role in the pathogenesis of RA but are not yet direct targets of existing disease-modifying antirheumatic drugs. The most promising of these, described by Gene Ontology terms, include: a) the Toll-like receptor signaling pathway; b) neutrophil activation; c) regulation of osteoblast differentiation; d) regulation of osteoclast differentiation; e) the prostaglandin biosynthetic process, and f) the canonical Wnt signaling pathway. The identified biological processes and their key regulators represent promising targets for the development of new drugs capable of improving the efficacy of RA therapy, particularly in patients resistant to existing treatments. The developed approach can also be successfully applied to the search for new targeted therapy targets for other diseases.

Key words: rheumatoid arthritis; gene networks; targeted therapy; ANDSystem

For citation: Mishchenko E.L., Yatsyk I.V., Demenkov P.S., Adamovskaya A.V., Ivanisenko T.V., Kleshchev M.A., Ivanisenko V.A. Searching for biological processes as targets for rheumatoid arthritis targeted therapy with ANDSystem, an integrated software and information platform. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1020-1030. doi 10.18699/vjgb-25-107

Введение

Ревматоидный артрит (РА) – хроническое аутоиммунное заболевание, характеризующееся системным воспалением, преимущественно поражающим суставы и приводящим к прогрессирующей деструкции хрящевой и костной тканей (Guo et al., 2018). По данным Всемирной организации здравоохранения, РА затрагивает около 0.5–0.6 % населения мира, встречаясь в 2–3 раза чаще у женщин, чем у мужчин и являясь одной из ведущих причин инвалидности в трудоспособном возрасте (Kvien et al., 2006; GBD, 2023).

Патогенез РА включает сложные взаимодействия генетических факторов, иммунных нарушений и внешних триггеров, что приводит к активации провоспалительных цитокинов, инфильтрации иммунных клеток в синовиальную оболочку суставов и хроническому воспалению (Firestein, McInnes, 2017). Несмотря на значительный прогресс в понимании молекулярных механизмов РА, достичь полного излечения этого заболевания пока не удается, а существующие терапевтические стратегии направлены в основном на предотвращение прогрессирования заболевания (Smolen et al., 2016).

Современные методы лечения РА основаны на использовании нескольких классов препаратов, оказывающих противовоспалительное действие (Ding et al., 2023; Smolen et al., 2023), включая: а) синтетические базисные противовоспалительные препараты (сБПВП), такие как метотрексат, лефлуномид, сульфасалазин и гидроксихлорохин; б) таргетные синтетические базисные противовоспалительные препараты (тсБПВП), например тофацитиниб и барицитиниб; в) генно-инженерные биологические препараты (ГИБП), включая ингибиторы фактора некроза опухоли (инфликсимаб, адалимумаб), интерлейкина-6 (тоцилизумаб, сарилумаб), интерлейкина 1 (анакинра) и анти-CD20 моноклональные антитела (ритуксимаб); г) нестероидные противовоспалительные препараты (НПВП) для симптоматической терапии; д) глюкокортикоиды (ГК) для быстрого подавления воспаления.

Особое внимание в клинической практике уделяется таким препаратам первой линии, как сБПВП и тсБПВП, способным модулировать иммунный ответ на уровне внутриклеточных сигнальных путей и метаболизма (van der Kooij et al., 2007). При этом действие препаратов группы тсБПВП направлено на определенные гены, кодирующие ключевые компоненты сигнального пути JAK/STAT. В частности, тофацитиниб подавляет воспаление, специфически ингибируя янус-киназу 3 (JAK3), играющую ключевую роль в передаче сигналов цитокинов, регулирующих выживание, пролиферацию, дифференцировку и апоптоз лимфоцитов (Adis Editorial, 2010). Препараты групп сБПВП и тсБПВП эффективны в достижении ремиссии у значительной доли пациентов, однако их применение ограничено побочными эффектами, такими как гепатотоксичность, иммуносупрессия и развитие резистентности (Olivera et al., 2020). Кроме того, около 40 % пациентов с РА демонстрируют слабый ответ на терапию, а у 5–20 % вообще не отмечается улучшения при стандартном лечении (Smolen et al., 2016), что подчеркивает необходимость поиска новых молекулярных мишеней для разработки более эффективных препаратов.

В развитие РА вовлечен ряд путей передачи сигналов – JAK/STAT, Notch, MAPK, Wnt, PI3K, SYK и другие, которые регулируют множество биологических процессов, вовлеченных в патогенез заболевания, включая воспалительный ответ и ремоделирование костной и хрящевой тканей (Ding et al., 2023). Эти и другие биологические процессы и сигнальные пути могут служить мишенями для лекарственной терапии РА. Например, в экспериментах на лабораторных мышах показано, что применение CEP-33779 – высокоселективного ингибитора ключевого компонента сигнального пути JAK/STAT белка JAK2 – может уменьшать проявления воспаления при артрите за счет подавления выработки цитокинов и активации Т- и В-лимфоцитов (Stump et al., 2011).

Целью нашей работы был поиск биологических процессов – новых перспективных фармакологических мишеней для терапии РА на основе реконструкции и анализа особого типа генных сетей, называемых ассоциативными генными сетями (АГС).

Генная сеть – это группа координированно функционирующих генов, контролирующих фенотипические признаки организма (Колчанов и др., 2013). Взаимодействия

между генами в генной сети осуществляются через их первичные и вторичные продукты – РНК, белки и метаболиты. Ассоциативная генная сеть представляет собой расширение генной сети, объединяющее геномные, молекулярные, фенотипические и средовые объекты, и описывает разнообразные типы взаимодействий и ассоциаций между ними (Demenkov et al., 2021).

Для реконструкции АГС использовалась программно-информационная система ANDSystem, обеспечивающая автоматическое извлечение знаний и фактов из текстов научных публикаций и биомедицинских фактографических баз данных (Ivanisenko V.A. et al., 2019). Для достижения поставленной цели решались следующие задачи: а) реконструкция ассоциативной генной сети РА, включающей гены и белки, вовлеченные в развитие этого заболевания; б) реконструкция ассоциативных генных сетей действия лекарств, применяемых для терапии РА, включающих гены и белки, вовлеченные в функциональные ответы на действие препаратов; в) идентификация на основе реконструированных ассоциативных генных сетей биологических процессов – перспективных мишеней для терапии ревматоидного артрита.

На основе описанного выше подхода обнаружено 11 биологических процессов, играющих существенную роль в развитии РА, но до сих пор не идентифицированных как прямые мишени современных базисных противовоспалительных препаратов (БПВП). К их числу относятся следующие процессы, описываемые терминами онтологии генов: а) сигнальный путь Toll-подобных рецепторов; б) активация нейтрофилов; в) регуляция дифференцировки остеобластов и г) остеокластов; д) биосинтез простагландинов, е) канонический сигнальный путь Wnt. Выявленные биологические процессы и их ключевые регуляторы представляют собой перспективные мишени для разработки новых лекарственных средств лечения РА. Реализованный в нашей работе подход может быть использован для поиска новых мишеней таргетной терапии и при других заболеваниях.

Материалы и методы

Перечень базисных противовоспалительных препаратов. Для составления перечня синтетических базисных противовоспалительных препаратов и таргетных синтетических базисных противовоспалительных препаратов, применяемых в лечении РА, использован официальный документ Общероссийской общественной организации «Ассоциация ревматологов России» – «Клинические рекомендации. Ревматоидный артрит (МКБ-10: М05, М06)» (Насонов и др., 2024). Этот документ содержит классификацию препаратов, применяемых для терапии РА, их фармакотерапевтические характеристики и коды анатомо-терапевтическо-химической классификации лекарственных средств. На основе указанных рекомендаций был сформирован список препаратов для дальнейшего исследования: сБПВП (метотрексат, лефлуномид, сульфасалазин, гидроксихлорохин) и тсБПВП (тофацитиниб, барицитиниб).

Реконструкция и анализ ассоциативных генных сетей. Для реконструкции ассоциативных генных сетей использовалась программно-информационная система ANDSystem (Ivanisenko V.A. et al., 2019, 2024; Ivanisenko T.V. et al., 2024). Она базируется на методах машинного чтения и искусственного интеллекта, используемых для автоматического извлечения знаний и фактов из источников больших генетических и биомедицинских данных, таких как тексты научных публикаций, патенты и фактографические базы данных.

На основе анализа текстов более 40 млн научных статей и патентов, а также 150 фактографических баз данных в базе знаний системы ANDSystem в виде графов семантических сетей знаний и фактов накоплена биомедицински значимая информация о 12 типах объектов, функционирующих в составе ассоциативных генных сетей (генах, белках, заболеваниях, биологических процессах, лекарствах и др.), а также о более чем 40 типах функциональных связей между этими объектами, включая регуляцию экспрессии генов, деградацию, модификацию и транспорт белков, физические взаимодействия между объектами («белок–белок», «белок–лиганд» и др.).

Кроме того, в базе знаний системы ANDSystem содержится описание ассоциативных связей между генами, белками, метаболитами и такими объектами, как заболевания, биологические процессы, лекарственные препараты и т. д. (Ivanisenko V.A. et al., 2019, 2022, 2024; Ivanisenko T.V. et al., 2024). В этой базе знаний имеются также связи типа «маркер», показывающие, что какие-либо гены, белки, биологические процессы или фенотипические признаки могут выступать маркерами заболеваний.

Идентификация биологических процессов на основе информации, представленной в реконструированных ассоциативных генных сетях. Анализ перепредставленности биологических процессов в реконструированных ассоциативных генных сетях выполнялся с использованием веб-сервера DAVID, версия 2021 (https://david.ncifcrf.gov/; Sherman et al., 2022) с настройками по умолчанию. DAVID оценивает степень пересечения списка генов, функционирующих в составе каждой из реконструированных генных сетей, со списками генов, соответствующих биологическим процессам, описанным в онтологии генов (GO). На этой основе с использованием гипергеометрического теста вычисляется вероятность получить наблюдаемое пересечение списков по случайным причинам. В нашем случае для идентификации биологических процессов, значимых для функционирования реконструированных генных сетей, использовался порог p-значение < 0.05 с поправкой по Бонферрони. Биологические процессы, удовлетворяющие этому критерию, относились к двум классам: а) БП, значимые для генной сети ревматоидного артрита и б) БП, значимые для генных сетей ответа на препараты сБПВП и тсБПВП при терапии ревматоидного артрита.

Результаты

Реконструкция ассоциативной генной сети ревматоидного артрита

С использованием системы ANDSystem проведена реконструкция ассоциативной генной сети ревматоидного артрита. Для этого использовалась информация, содержащаяся в базе знаний ANDSystem.

Таблица 1. Характеристика связей между центральной и периферическими вершинами в генной сети ревматоидного артрита

№ п/п	Тип связи	Число связей	Доля*, %
	Регуляторные связи	4381	44.4
1	Подавление экспрессии (expression downregulation)	93	0.9
2	Регуляция экспрессии (expression regulation)	472	4.8
3	Повышение экспрессии (expression upregulation)	365	3.7
4	Подавление активности (activity downregulation)	15	0.2
5	Регуляция активности (activity regulation)	26	0.3
6	Повышение активности (activity upregulation)	10	0.1
7	Регуляция (regulation)	1812	18.3
8	Повышающая регуляция (upregulation)	802	8.1
9	Понижающая регуляция (downregulation)	786	8.0
	Прочие связи	5496	55.6
1	Ассоциация (association)	4449	45.0
2	Вовлечение (involvement)	172	1.7
3	Маркер (marker)	338	3.4
4	Фактор риска (risk factor)	274	2.8
5	Лечение (treatment)	263	2.7

* Доля (%) связей конкретного типа от общего числа связей в ассоциативной генной сети ревматоидного артрита.

Граф реконструированной ассоциативной генной сети имел звездообразную структуру: центральная вершина графа, соответствующая термину «ревматоидный артрит», была соединена ребрами с прочими вершинами графа генной сети, соответствующими белкам и генам, связанными, по данным базы знаний ANDSystem, с ревматоидным артритом (рис. S1 Приложения)[1]. Всего этот граф содержал 4685 вершин, соответствующих 2178 генам и 2507 белкам (табл. S1), а также 9877 связей между центральной вершиной (ревматоидным артритом) и прочими вершинами. Отметим, что количество связей в графе было больше, чем количество вершин. Причина в том, что одна и та же вершина, соответствующая гену или белку, могла быть связана с центральной вершиной несколькими ребрами, каждое из которых, согласно базе знаний системы ANDSystem, описывало определенный тип взаимодействий между РА и тем или иным геном или белком.

В табл. S1 представлен список генов и белков, входящих в реконструированную ассоциативную генную сеть РА, которая включает в частности, гены и белки, вовлеченные в воспалительный процесс – интерлейкины (IL1, IL6, IL13 и др.), белки семейства факторов некроза опухоли, ключевой регулятор воспаления белок NF-κB, гены и белки, функционирующие в сигнальных путях Wnt, JAK/STAT, Notch, MAPK, PI3K, SYK, которые, как известно, вносят определяющий вклад в патогенез РА (Ding et al., 2023).

В табл. 1 представлена информация о классификации 14 типов связей между центральной и периферическими вершинами в генной сети ревматоидного артрита. Эти связи подразделяются на два класса. Первый класс (регуляторные связи) объединяет девять типов связей, таких как «подавление экспрессии гена» (expression downregulation), «повышение экспрессии гена» (expression upregulation), «регуляция активности белка» (activity regulation) и другие. Например, экспрессия интерлейкина 1 бета (IL1B) повышается при ревматоидном артрите (Mohd et al., 2019), что отражено в базе знаний ANDSystem как связь типа «повышение экспрессии» между ревматоидным артритом и белком IL1B. Белок интерлейкин-6 (IL6) стимулирует фибробласты в синовиальной оболочке суставов (Singh et al., 2021) и способствует развитию одного из симптомов РА (потере костной ткани), что отражено в ANDSystem как связь типа «повышающая регуляция» между заболеванием «ревматоидный артрит» и белком IL6.

Второй класс «прочие связи» объединяет пять других типов связей, обнаруженных при реконструкции генной сети ревматоидного артрита, описывающих ситуацию, когда ген или белок каким-то образом ассоциирован с РА. Например, это могут быть какие-то структурные или функциональные особенности гена, когда мутация в гене является фактором риска ревматоидного артрита.

На основе информации, содержащейся в ассоциативной генной сети РА и базе знаний системы ANDSystem, возможна реконструкция тонких механизмов участия генов и белков в развитии ревматоидного артрита. На рис. 1 в качестве примера показаны регуляторные взаимодействия между генами и белками, функционирующими в сигнальном пути Wnt, регулируемом провоспалительными цитокинами – интерлейкином 1 бета (IL1B), фактором некроза опухоли альфа (TNFA), интерлейкином 6 (IL6).

Как видно из рис. 1, в регуляции сигнального пути Wnt при ревматоидном артрите участвуют интерлейкин 1 бета (IL1B), фактор некроза опухоли альфа (TNFA), интерлейкин-6 (IL6), которые активируют экспрессию гена *WNT5A*, кодирующего белок WNT5A – лиганд рецепторов FZD,

[1] Рис. S1, S2 и табл. S1–S6 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx37.xlsx

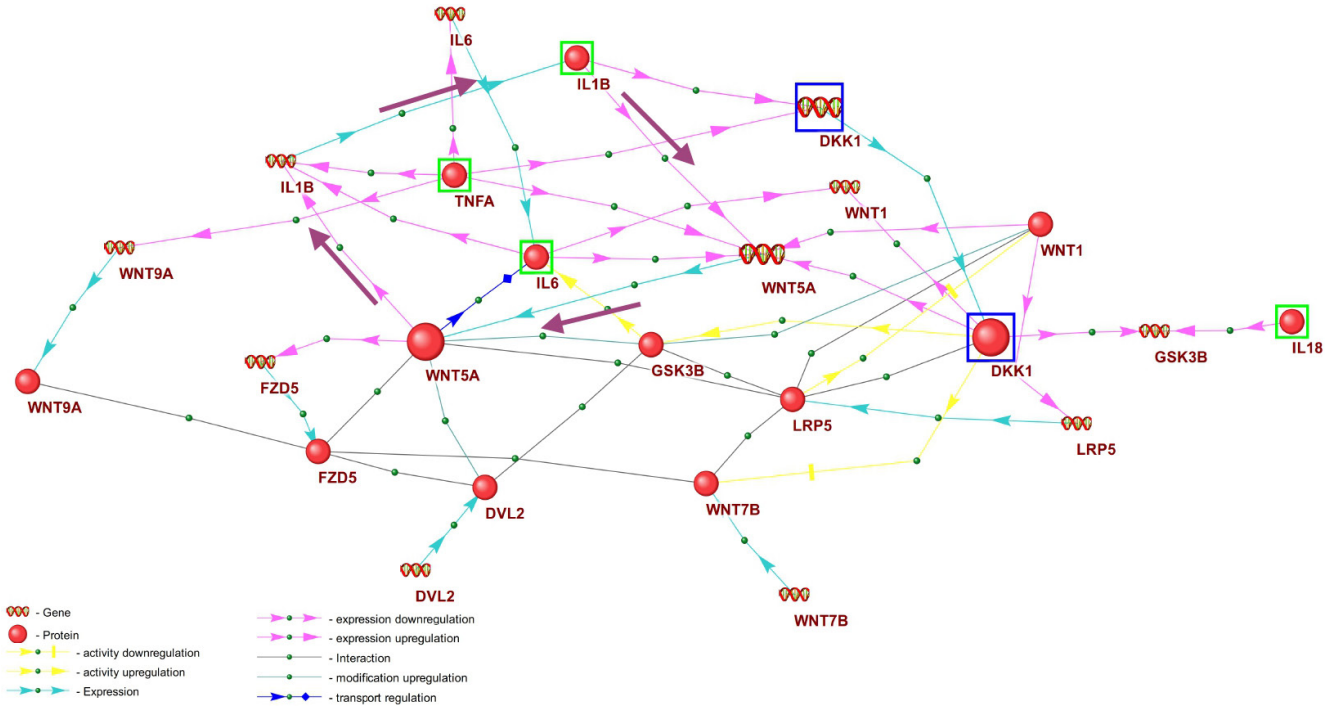


Рис. 1. Механизм регуляции ключевых компонентов сигнального пути Wnt провоспалительными цитокинами, реконструированный на основе генной сети ревматоидного артрита в базе знаний системы ANDSystem.
Зелеными рамками отмечены провоспалительные цитокины; жирными стрелками – компоненты регуляторного контура с положительной обратной связью, синими рамками обозначены ген *DKK1* и кодируемый им белок Dickkopf-1 (DKK1) – ингибитор канонического пути Wnt.

участвующий в неканоническом пути Wnt (Miao et al., 2013). При этом WNT5A, по данным ANDSystem, активирует экспрессию гена *IL1B*, кодирующего интерлейкин 1 бета. Таким образом, *IL1B* и *WNT5A* взаимно активируют экспрессию друг друга и могут образовывать контур с положительной обратной связью (обозначен на рис. 1 жирными стрелками).

Реконструкция ассоциативных генных сетей, вовлеченных в функциональные ответы на действия лекарств, применяемых для терапии ревматоидного артрита

На рис. 2 показан граф АГС ответа на препараты группы тсБПВП (см. также табл. S2). Он содержит две вершины, соответствующие названиям препаратов (тофацитиниб, барицитиниб), и 157 связей этих вершин с другими вершинами графа, соответствующими 22 белкам и 51 гену. Как можно видеть на рис. 2, препарат тофацитиниб, по данным базы знаний ANDSystem, характеризуется существенно большим числом связей с белками и генами (60), по сравнению с препаратом барицитиниб (26). При этом в ответ на оба препарата оказались включены гены, вовлеченные в воспалительный ответ, – *MMP3*, *IL2RA*, *CXCL10* и белки STAT3, STAT5A, JAK1, JAK2 – члены сигнального пути JAK/STAT.

На рис. S2 показан граф АГС ответа на препараты группы сБПВП (метотрексат, лефлуномид, сульфасалазин, гидроксихлорохин). Граф содержит 261 вершину, 4 из которых соответствуют названиям указанных препаратов (см. также табл. S2). С этими четырьмя вершинами 485 связями соединены прочие вершины графа, соответствующие 106 белкам и 151 гену. Наибольшее число связей (160) в АГС ответа на препараты группы сБПВП выявлено для метотрексата. В число белков и генов, связанных с этим препаратом, входят, в частности, рецептор к интерлейкину 1 типа 1 (IL1R1), фактор некроза опухоли альфа (TNFA), транскрипционный фактор, медиатор воспаления NFKB1, каспазы (CASP1, CASP3, CASP9). Второе место по количеству связей (73) занимал гидроксихлорохин, связанный с провоспалительными цитокинами, такими как интерлейкин 1 бета (IL1B), и фактором некроза опухоли альфа (TNFA), а также с каталазой (CAT) и цитохромами, участвующими в метаболизме ксенобиотиков (CP2B6, CYP1B1). Третье и четвертое место делят сульфасалазин и лефлуномид (с количеством связей 26 и 17 соответственно). Отметим, что некоторые белки в АГС ответа на препараты класса сБПВП (например, IL1B, CCL2, TNFA, CASP3) являются мишенями для действия нескольких препаратов.

Распределение связей по типам в АГС ответа на препараты групп сБПВП и тсБПВП приведено в табл. 2. Как можно видеть из табл. 2, в числе связей в АГС ответа на препараты групп сБПВП и тсБПВП преобладали регуляторные связи, в особенности регуляция экспрессии генов.

Идентификация биологических процессов на основе информации, представленной в реконструированных ассоциативных генных сетях

С использованием веб-ресурса DAVID, базирующегося на онтологии генов, выполнен анализ перепредставленности биологических процессов в реконструированных генных сетях: а) генной сети ревматоидного артрита и б) генных

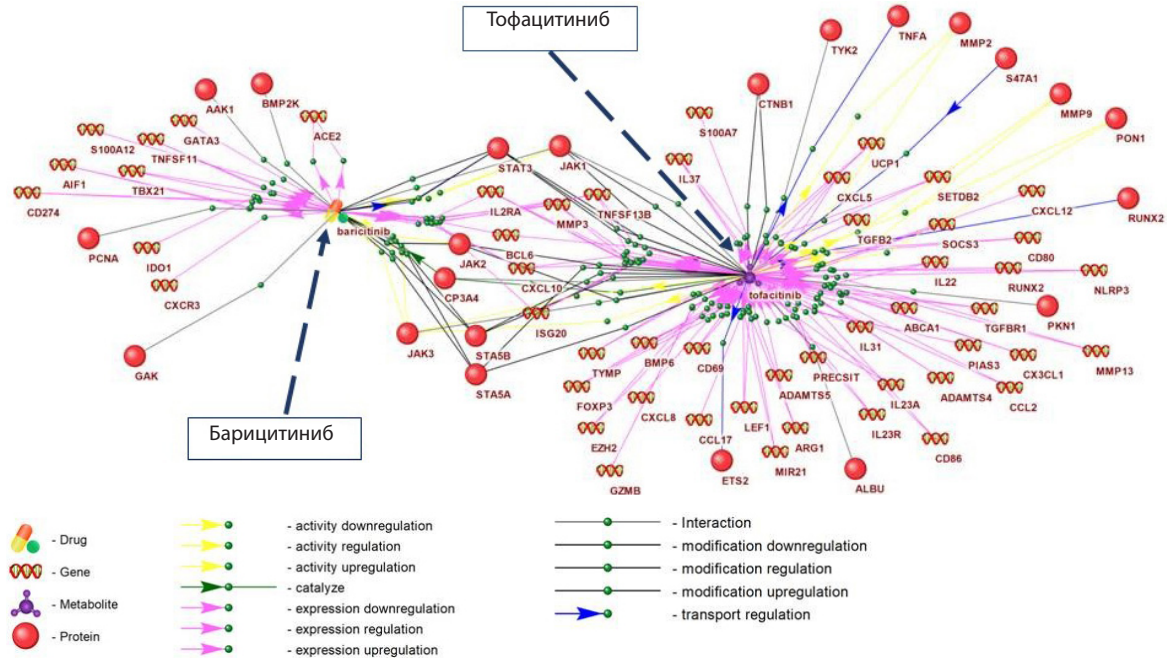


Рис. 2. Реконструированная ассоциативная генная сеть ответа на два таргетных синтетических базисных противовоспалительных препарата тсБПВП, тофацитиниб и барицитиниб.

Таблица 2. Распределение связей по типам в реконструированных ассоциативных генных сетях ответа на синтетические и таргетные синтетические базисные противовоспалительные препараты

№ п/п	Тип связи	сБПВП		тсБПВП	
		Число	Доля, %	Число	Доля, %
	Регуляторные связи	529	87.6	143	91.1
1	Подавление экспрессии (expression downregulation)	73	15.1	35	22.3
2	Регуляция экспрессии (expression regulation)	158	32.6	57	36.3
3	Повышение экспрессии (expression upregulation)	64	13.2	12	7.6
4	Подавление активности (activity downregulation)	18	3.7	6	3.8
5	Регуляция активности (activity regulation)	34	7.0	7	4.5
6	Повышение активности (activity upregulation)	16	3.3	1	0.6
7	Подавление модификации (modification downregulation)	10	2.1	8	5.1
8	Регуляция модификации (modification regulation)	8	1.6	9	5.7
9	Усиление модификации (modification upregulation)	4	0.8	3	1.9
10	Регуляция транспорта (transport regulation)	28	5.8	5	3.2
11	Подавление деградации (degradation downregulation)	5	1.0	Нет	Нет
12	Регуляция деградации (degradation regulation)	6	1.2	Нет	Нет
13	Усиление деградации (degradation upregulation)	1	0.2	Нет	Нет
	Прочие связи	60	12.4	14	7.8
14	Катализ (catalyze)	14	2.4	2	1.1
15	Физическое взаимодействие (interaction)	46	7.8	12	6.7

сетях ответа на два типа противовоспалительных препаратов (сБПВП и тсБПВП).

Для реконструированных ассоциативных генных сетей ревматоидного артрита, ответа на сБПВП и тсБПВП выявлено 381, 64 и 44 перепредставленных биологических процессов, соответственно охарактеризованных в табл. 3 (детали см. в табл. S4–S6). Как видно из табл. 3, воспалительный ответ (идентификатор ГО GO:0006954) оказался статистически значимо перепредставлен как в генной сети РА, так и генных сетях ответа на препараты групп сБПВП и тсБПВП. Интересно отметить, что в списке наиболее значимо перепредставленных процессов в генной сети ответа на сБПВП оказались процессы метаболизма ксенобиотиков, которые не были перепредставлены в генной

Таблица 3. Результаты анализа перепредставленности биологических процессов онтологии генов (GO) для ассоциативных генных сетей ревматоидного артрита, а также генных сетей реакции на сБПВП и тсБПВП

Генная сеть	Количество достоверно перепредставленных биологических процессов	Наиболее статистически значимо перепредставленные биологические процессы		
		Идентификатор	Название	*p*-value*
Генная сеть ревматоидного артрита	381	GO:0006954	Воспалительный ответ (inflammatory response)	$3.7 \cdot 10^{-123}$
		GO:0006955	Иммунный ответ (immune response)	$8.2 \cdot 10^{-103}$
		GO:0007165	Передача сигналов (signal transduction)	$2.9 \cdot 10^{-60}$
Генная сеть реакции на сБПВП	64	GO:0006805	Метаболизм ксенобиотиков (xenobiotic metabolic process)	$5.6 \cdot 10^{-21}$
		GO:0009410	Ответ на ксенобиотики (response to xenobiotic stimulus)	$1.9 \cdot 10^{-19}$
		GO:0006954	Воспалительный ответ (inflammatory response)	$1.4 \cdot 10^{-14}$
Генная сеть реакции на тсБПВП	44	GO:0006954	Воспалительный ответ (inflammatory response)	$4.2 \cdot 10^{-16}$
		GO:0007259	Сигнальный путь рецепторов клеточной поверхности через JAK/STAT (cell surface receptor signaling pathway via JAK/STAT)	$2.2 \cdot 10^{-11}$
		GO:0019221	Сигнальный путь, опосредованный цитокинами (cytokine-mediated signaling pathway)	$3.0 \cdot 10^{-10}$

* $p < 0.05$.

Таблица 4. Биологические процессы, для которых не выявлены регулирующие их препараты групп сБПВП и тсБПВП, применяемые в терапии ревматоидного артрита

№ п/п	Идентификатор онтологии генов (GO)	Биологический процесс онтологии генов	Число генов ревматоидного артрита, вовлеченных в процесс	*p*-value*
1	GO:0034612	Ответы клеток на фактор некроза опухоли (TNF)	58	$9.8 \cdot 10^{-23}$
2	GO:0031295	Костимуляция Т-клеток	29	$3.3 \cdot 10^{-13}$
3	GO:0002224	Сигнальные пути Toll-подобных рецепторов	19	$8.3 \cdot 10^{-8}$
4	GO:0014823	Ответ на физическую активность	26	$1.3 \cdot 10^{-7}$
5	GO:0034097	Реакция на цитокины	24	$8.64 \cdot 10^{-7}$
6	GO:0010468	Регуляция экспрессии генов	53	$2.0 \cdot 10^{-3}$
7	GO:0045668	Негативная регуляция дифференциации остеобластов	27	$8.5 \cdot 10^{-5}$
8	GO:0042119	Активация нейтрофилов	12	$1.53 \cdot 10^{-3}$
9	GO:0045671	Негативная регуляция дифференциации остеокластов	15	$3.27 \cdot 10^{-2}$
10	GO:0001516	Биосинтез простагландинов	12	$1.5 \cdot 10^{-2}$
11	GO:0060070	Канонический сигнальный путь Wnt	30	$2.2 \cdot 10^{-2}$

* *p*-value – уровень значимости перепредставленности терминов онтологии генов для набора генов, ассоциированных с ревматоидным артритом с поправкой Бонферрони.

сети тсБПВП. Для АГС ответа на тсБПВП были наиболее значимо перепредставлены сигнальные пути JAK/STAT (идентификатор ГО GO:0007259, см. табл. 3) и цитокинов (идентификатор ГО GO:0019221, см. табл. 3).

Для дальнейшего анализа из 381 выявленного биологического процесса, перепредставленного в АГС ревматоидного артрита (см. табл. 3), с использованием базы знаний ANDSystem был отобран 71 процесс, характеризующийся типами связей: «регуляция», «понижающая регуляция» и «повышающая регуляция» с заболеванием «ревматоидный артрит». Проведено пересечение списка из 71 биологического процесса, вовлеченного в патогенез РА, со списками биологических процессов, перепредставленных для АГС ответа на препараты групп сБПВП (64 процесса) и тсБПВП (44 процесса). В результате найдено 59 биологических процессов, вовлеченных в патогенез РА, но не входящих в список перепредставленных процессов для АГС ответа на рассматриваемые лекарственные препараты. Из этих 59 процессов было удалено 48 процессов, которые, по данным базы знаний ANDSystem, связаны с рассматриваемыми препаратами групп сБПВП (метотрексатом, лефлуномидом, сульфасалазином, гидроксихлорохином) и тсБПВП (тофацитинибом, барицитинибом) взаимодействиями типов «регуляция», «понижающая регуляция» и «повышающая регуляция».

В результате был получен список из 11 биологических процессов (табл. 4). Выявленные процессы характеризуются следующим: эти процессы вовлечены в патогенез

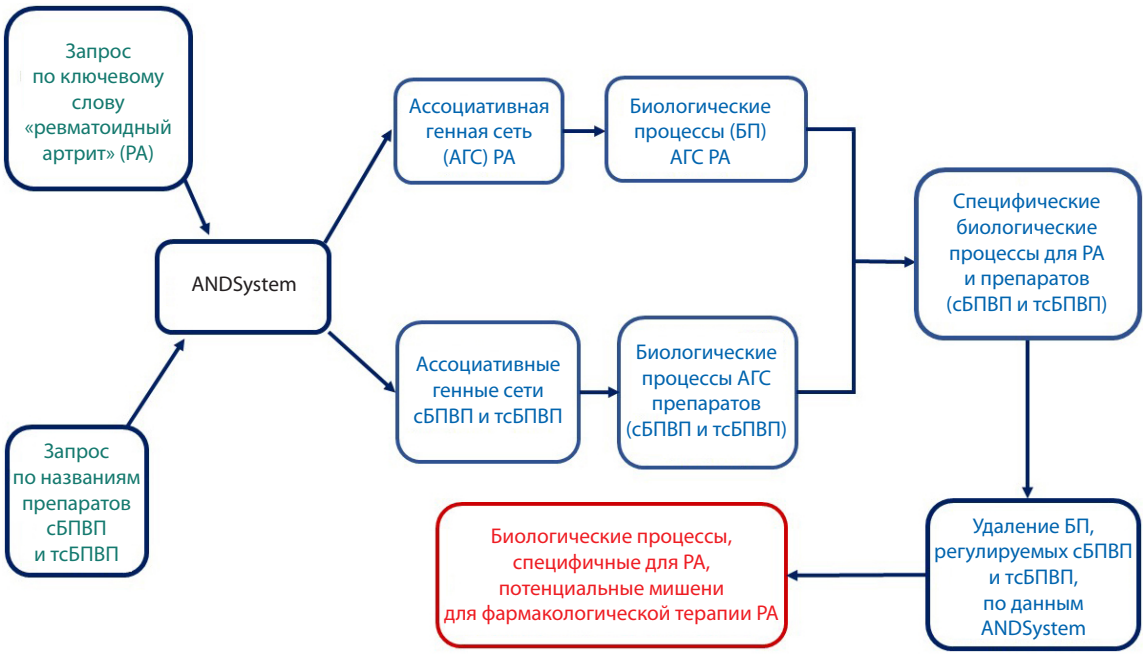


Рис. 3. Основные этапы поиска биологических процессов, перспективных для разработки новых противоревматоидных лекарств.

РА – ревматоидный артрит; АГС – ассоциативная генная сеть; БП – биологический процесс; сБПВП – синтетические базисные противовоспалительные препараты (метотрексат, лефлуномид, сульфасалазин, гидроксихлорохин); тсБПВП – таргетные синтетические базисные противовоспалительные препараты (тофацитиниб, барицитиниб).

ревматоидного артрита. Кроме того, для них не обнаружено регулирующих их препаратов групп сБПВП и тсБПВП. Именно эти процессы представляют особый интерес в качестве мишеней для разработки лекарств терапии ревматоидного артрита.

Как видно из табл. 4, биологические процессы, вовлеченные в патогенез ревматоидного артрита, но не регулируемые базисными противовоспалительными препаратами, включали в себя: а) воспалительные реакции (идентификаторы ГО GO:0034097, GO:0034612, GO:0031295, GO:0002224); б) морфогенез костной ткани (GO:0045668, GO:0045671); в) канонический сигнальный путь Wnt (GO:0060070); г) биосинтез простагландинов (GO:0001516), а также д) ответ на физическую активность (GO:0014823) и регуляцию экспрессии генов (GO:0010468).

Таким образом, нами выполнен поиск биологических процессов – новых перспективных фармакологических мишеней для терапии РА на основе реконструкции и анализа ассоциативных генных сетей.

На рис. 3 представлена реализованная в нашей работе принципиальная схема поиска биологических процессов – новых перспективных мишеней для разработки противоревматоидных препаратов.

Обсуждение

Поиск новых лекарственных мишеней для лечения РА – актуальная задача современной медицины с учетом, что до 40 % пациентов не достигают полного ответа на существующую терапию (Ding et al., 2023). В связи с этим мы предложили метод поиска биологических процессов в качестве мишеней действия новых противоревматических препаратов на основе реконструкции генных сетей и сравнительного анализа биологических процессов, ассоциированных с РА, и процессов, регулируемых базисными противовоспалительными препаратами, используемыми в настоящее время в клинической практике (Насонов и др., 2024).

База знаний ANDSystem, использованная нами для реконструкции генных сетей, интегрирует накопленную в научной литературе информацию о молекулярных механизмах действия лекарственных препаратов и патогенеза заболеваний, что позволяет найти новые мишени для терапии на системном уровне, включая биологические процессы, тем самым увеличить эффективность терапии и диагностики. В нашей работе реконструированы АГС ревматоидного артрита, а также АГС, описывающие взаимодействия синтетических и таргетных противовоспалительных препаратов с генами и белками человека. Анализ показал, что генная сеть ревматоидного артрита обогащена генами, вовлеченными в регуляцию воспалительного ответа, что соответствует хорошо известным данным о ведущем значении системного воспаления в патогенезе этого заболевания (Firestein, McInnes, 2017; Figus et al., 2021). Не случайно поэтому реконструированные генные сети белков и генов, на которые направлено действие препаратов групп сБПВП (см. рис. S2) и тсБПВП (см. рис. 2), включают главным образом гены и белки, вовлеченные в функционирование иммунной системы.

Согласно результатам функциональной аннотации генов-мишеней, для синтетических базисных противовоспалительных препаратов (сБПВП) список статистически значимо перепредставленных биологических процессов включал процессы, связанные не только с воспалением, но

и с метаболизмом ксенобиотиков. Это свидетельствует о значительной нагрузке, оказываемой сБПВП на биохимические системы организма, обеспечивающие утилизацию чужеродных веществ, что может обуславливать серьезные негативные побочные эффекты указанной группы препаратов (Olivera et al., 2020).

С другой стороны, для генов, вовлеченных в ответ на таргетные синтетические базисные противовоспалительные препараты (тсБПВП), биологические процессы метаболизма ксенобиотиков не были значимо перепредставлены. Вместе с тем в списке перепредставленных процессов для АГС ответа на тсБПВП наряду с воспалением оказались процессы, связанные с функционированием сигнального пути JAK/STAT, который является важнейшим звеном патогенеза РА (Ding et al., 2023). Это свидетельствует о более направленном действии препаратов группы тсБПВП на патогенез РА и подчеркивает важность разработки средств таргетной терапии для повышения эффективности лечения и снижения числа побочных эффектов. Однако многообразие и сложность взаимодействий биологических процессов, приводящих к развитию РА, и недостаточная эффективность терапии существующими базисными противовоспалительными препаратами, обуславливают необходимость поиска новых мишеней для лечения РА (Smolen et al., 2016).

Разработанный нами подход, основанный на реконструкции генных сетей, вовлеченных в развитие заболевания и в реакцию на известные лекарственные препараты, а также на сравнительном анализе биологических процессов, регулируемых этими генными сетями, позволил обнаружить 11 биологических процессов (см. табл. 4). Эти процессы относятся к числу ключевых для патогенеза РА, но не являются мишенями применяемых в настоящее время противовоспалительных препаратов. Следует отметить, что регуляция экспрессии (GO:0010468) и ответ на физическую активность (GO:0014823) относятся к группе довольно обширных процессов, охватывающих множество молекулярных механизмов в клетке, что затрудняет разработку таргетных препаратов.

Анализ литературы показал, что для таких процессов, как реакция на цитокины (GO:0034097), реакция на фактор некроза опухоли TNFA (GO:0034612), костимуляция Т-клеток (GO:0031295), имеются данные об их частичной регуляции противовоспалительными препаратами групп сБПВП и тсБПВП, применяемыми в настоящее время. Например, тсБПВП, такие как тофацитиниб и барицитиниб, эффективно блокируют JAK/STAT сигнальные пути, которые находятся ниже по каскаду от рецепторов цитокинов и TNFA, обеспечивая мощное подавление воспалительных ответов (Palmroth et al., 2021).

Однако биологические процессы, такие как функционирование сигнального пути Toll-подобных рецепторов, активация нейтрофилов, негативная регуляция дифференцировки остеобластов, негативная регуляция дифференцировки остеокластов, канонический сигнальный путь Wnt, биосинтез простагландинов, не регулируются напрямую базисными противовоспалительными лекарственными препаратами, которые в настоящее время активно используются ревматологами, в соответствии с клиническими рекомендациями (Насонов и др., 2024). Тем не менее перечисленные выше пути могут быть важными звеньями патогенеза РА. Например, активация нейтрофилов играет важную роль в воспалении при РА, и ингибиторы CXCR2, исследуемые для других воспалительных состояний, могут быть адаптированы для РА (Alam et al., 2020).

Известно, что сигнальный путь Wnt играет важную роль в активации фибробластов и воспалении синовиальной оболочки, а также резорбции костной ткани и разрушении суставов при развитии ревматоидного артрита (Miao et al., 2013). Экспрессия генов, кодирующих белки семейства Wnt, активирующие сигнальный путь Wnt, повышена в синовиальной оболочке при ревматоидном артрите, в том числе благодаря провоспалительным цитокинам (Prajapati, Doshi, 2023). При этом активация неканонического сигнального пути Wnt в свою очередь приводит к повышению экспрессии медиаторов воспаления, включая транскрипционный фактор NF-κB и цитокины (Miao et al., 2013), усиливая воспаление.

По данным базы знаний ANDSystem (см. рис. 1), интерлейкин 1 бета (IL1B) и белок WNT5A взаимно активируют экспрессию друг друга, что может создавать порочный круг в патогенезе ревматоидного артрита. Поэтому модуляция сигнального пути Wnt может быть перспективной для ослабления воспаления суставов при РА. В частности, показано, что белок NAV2 способствует воспалительной реакции фиброцитоподобных синовиоцитов, активируя сигнальный путь Wnt при ревматоидном артрите, а его ингибирование способно ослабить воспаление суставов при этом заболевании (Wang R. et al., 2021).

С другой стороны, провоспалительные цитокины – фактор некроза опухоли альфа (TNFA) и IL1B, по данным ANDSystem (см. рис. 2), могут активировать экспрессию гена *DKK1*, кодирующего белок Dickkopf-1 (DKK1) – важный ингибитор канонического сигнального пути Wnt (Rabelo et al., 2010). Показано, что уровень DKK1 в сыворотке крови повышен у пациентов с РА и коррелирует с уровнем воспаления и степенью разрушения костной ткани в суставах (Wang S.Y. et al., 2011). Активация экспрессии *DKK1* провоспалительными цитокинами при ревматоидном артрите может приводить к подавлению сигнального пути Wnt и, как следствие, к активации сигнального пути RANK/RANKL в остеокластах, увеличению их активности и потере костной массы, характерной для РА (Miao et al., 2013).

Следовательно, нарушение регуляции сигнального пути Wnt может быть причиной изменений в биологических процессах регуляции дифференциации остеобластов и остеокластов при РА, которые, по данным нашего исследования (см. табл. 4), также входят в список биологических процессов – потенциальных мишеней для новых противоревматоидных препаратов. Кроме того, DKK1 стимулирует ангиогенез в синовиальной оболочке и образование паннуса – патологически измененной ткани синовиальной оболочки, играющей важнейшую роль в деструкции суставов при РА (Cici et al., 2019).

Таким образом, сигнальный путь Wnt – перспективная мишень для разработки новых противоревматоидных препаратов, однако регуляция его при РА весьма сложна,

зависит от типа тканей и клеток, поэтому требуются дальнейшие исследования, направленные на реконструкцию генной сети этого пути при РА и анализ ее структурно-функциональных особенностей в различных клетках и тканях.

Простагландины, в частности простагландин Е2, как известно, играют важную роль в развитии как острой воспалительной реакции, так и хронического воспаления (Kawahara et al., 2015), усиливая воспалительные процессы путем активации экспрессии рецепторов цитокинов и белков семейства NFKB, которые являются важнейшими индукторами воспаления (Yao, Narumiya, 2019). Простагландин Е2 – важный медиатор воспаления при РА, служит мишенью действия ряда нестероидных противовоспалительных препаратов (НПВП) при этом заболевании (Park et al., 2006). Биосинтез простагландинов (идентификатор ГО биологического процесса GO:0001516) частично модулируется НПВП, такими как целекоксиб, но разработка более специфичных ингибиторов может улучшить терапевтические исходы (Gong et al., 2012).

Известно, что Toll-подобные рецепторы (TLR) вносят важный вклад в индукцию воспаления, поскольку их активация приводит к повышению активности сигнальных путей и ряда факторов транскрипции, таких как ядерный фактор κB (NF-κB), активаторный белок-1 (AP-1) и факторы регуляции интерферона (IRF), которые индуцируют экспрессию провоспалительных цитокинов, TNF, IL1β, IL6 и других (Kawasaki, Kawai, 2014). Показано, что экспрессия генов Toll-подобных рецепторов увеличена в синовиальной оболочке пациентов с РА, и TLR вносят значительный вклад в развитие воспаления при РА, но терапевтические вмешательства, нацеленные на сигнальные пути TLR, до сих пор не были успешно внедрены в клинику (Unterberger et al., 2021).

Таким образом, все перечисленные выше биологические процессы играют большую роль в развитии РА, но при этом не регулируются базисными противовоспалительными препаратами, применяемыми в настоящее время в клинической практике. Поэтому эти биологические процессы и их ключевые регуляторы могут служить в качестве мишеней для разработки новых лекарственных средств для лечения ревматоидного артрита.

Следует отметить, что ревматоидный артрит характеризуется выраженной коморбидностью с другими заболеваниями, включая сердечно-сосудистые и респираторные заболевания (Figus et al., 2021), тревожно-депрессивные расстройства (Hill et al., 2022), остеопороз (Llorente et al., 2020). В связи с этим в дальнейшей работе планируется провести анализ выявленных биологических процессов как основы для коморбидности РА с другими заболеваниями.

Кроме того, в настоящей работе не проводилась идентификация мишеней на уровне генов, что может послужить предметом дальнейших исследований на основе анализа структурной организации генных сетей.

Заключение

В нашей работе осуществлена компьютерная реконструкция ассоциативных генных сетей ревматоидного артрита, а также АГС, описывающих взаимодействия синтетических и таргетных противовоспалительных препаратов с генами и белками человека. На основе анализа этих генных сетей выполнен поиск биологических процессов – новых перспективных фармакологических мишеней для терапии РА. Предложенный подход может быть использован также для поиска новых мишеней для таргетной терапии других заболеваний, где стандартные методы лечения демонстрируют недостаточный терапевтический эффект.

Список литературы / References

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):833-850
[Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed.* 2013;17(4/2):833-850 (in Russian)]

Насонов Е.Л., Лила А.М., Каратеев Д.Е., Мазуров В.И. и др. Клинические рекомендации. Ревматоидный артрит. Общероссийская общественная организация «Ассоциация ревматологов России», 2024. КР250.
[Nasonov E.L., Lila A.M., Karateev D.E., Mazurov V.I. et al. Clinical Recommendations. Rheumatoid Arthritis. All-Russian Public Organization “Association of Rheumatologists of Russia”, 2024. KR250] (in Russian)

Adis Editorial. Tofacitinib. *Drugs R D*. 2010;10(4):271-284. doi 10.2165/11588080-000000000-00000

Alam M.J., Xie L., Ang C., Fahimi F., Willingham S.B., Kueh A.J., Herold M.J., Mackay C.R., Robert R. Therapeutic blockade of CXCR2 rapidly clears inflammation in arthritis and atopic dermatitis models: demonstration with surrogate and humanized antibodies. *mAbs*. 2020;12(1):1856460. doi 10.1080/19420862.2020.1856460

Cici D., Corrado A., Rotondo C., Cantatore F.P. Wnt signaling and biological therapy in rheumatoid arthritis and spondyloarthritis. *Int J Mol Sci*. 2019;20(22):5552. doi 10.3390/ijms20225552

Demenkov P.S., Oshchepkova E.A., Ivanisenko T.V., Ivanisenko V.A. Prioritization of biological processes based on the reconstruction and analysis of associative gene networks describing the response of plants to adverse environmental factors. *Vavilov J Genet Breed*. 2021;25(5):580-592. doi 10.18699/VJ21.065

Ding Q., Hu W., Wang R., Yang Q., Zhu M., Li M., Cai J., Rose P., Mao J., Zhu Y.Z. Signaling pathways in rheumatoid arthritis: implications for targeted therapy. *Signal Transduct Target Ther*. 2023;8(1):68. doi 10.1038/s41392-023-01331-9

Figus F.A., Piga M., Azzolin I., McConnell R., Iagnocco A. Rheumatoid arthritis: extra-articular manifestations and comorbidities. *Autoimmun Rev*. 2021;20(4):102776. doi 10.1016/j.autrev.2021.102776

Firestein G.S., McInnes I.B. Immunopathogenesis of rheumatoid arthritis. *Immunity*. 2017;46(2):183-196. doi 10.1016/j.immuni.2017.02.006

GBD 2021 Rheumatoid Arthritis Collaborators. Global, regional, and national burden of rheumatoid arthritis, 1990–2020, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol*. 2023;5(10):e594-e610. doi 10.1016/S2665-9913(23)00211-4

Gong L., Thorn C.F., Bertagnolli M.M., Grosser T., Altman R.B., Klein T.E. Celecoxib pathways: pharmacokinetics and pharmacodynamics. *Pharmacogenet Genomics*. 2012;22(4):310-318. doi 10.1097/FPC.0b013e32834f94cb

Guo Q., Wang Y., Xu D., Nossent J., Pavlos N.J., Xu J. Rheumatoid arthritis: pathological mechanisms and modern pharmacologic therapies. *Bone Res*. 2018;6:15. doi 10.1038/s41413-018-0016-9

Hill J., Harrison J., Christian D., Reed J., Clegg A., Duffield S.J., Goodson N., Marson T. The prevalence of comorbidity in rheumatoid arthritis: a systematic review and meta-analysis. *Br J Community Nurs*. 2022;27(5):232-241. doi 10.12968/bjcn.2022.27.5.232

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci*. 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks. *BMC Bioinformatics*. 2019;20(Suppl. 1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., Mishchenko E.L., Khripko O.P., Khripko Y.I., Voevoda S.M. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation. *Sci Rep*. 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci*. 2024;25(21):11847. doi 10.3390/ijms252111847

Kawahara K., Hohjoh H., Inazumi T., Tsuchiya S., Sugimoto Y. Prostaglandin E_2-induced inflammation: relevance of prostaglandin E receptors. *Biochim Biophys Acta*. 2015;1851(4):414-421. doi 10.1016/j.bbalip.2014.07.008

Kawasaki T., Kawai T. Toll-like receptor signaling pathways. *Front Immunol*. 2014;5:461. doi 10.3389/fimmu.2014.00461

Kvien T.K., Uhlig T., Ødegård S., Heiberg M.S. Epidemiological aspects of rheumatoid arthritis: the sex ratio. *Ann NY Acad Sci*. 2006;1069:212-222. doi 10.1196/annals.1351.019

Llorente I., García-Castañeda N., Valero C., González-Álvaro I., Castañeda S. Osteoporosis in rheumatoid arthritis: dangerous liaisons. *Front Med* (*Lausanne*). 2020;7:601618. doi 10.3389/fmed.2020.601618

Miao C.G., Yang Y.Y., He X., Li X.F., Huang C., Huang Y., Zhang L., Lv X.W., Jin Y., Li J. Wnt signaling pathway in rheumatoid arthritis. *Cell Signal*. 2013;25(10):2069-2078. doi 10.1016/j.cellsig.2013.04.002

Mohd Jaya F.N., Garcia S.G., Borràs F.E., Chan G.C.F., Franquesa M. Paradoxical role of Breg-inducing cytokines in autoimmune diseases. *J Transl Autoimmun*. 2019;2:100011. doi 10.1016/j.jtauto.2019.100011

Olivera P.A., Lasa J.S., Bonovas S., Danese S., Peyrin-Biroulet L. Safety of Janus kinase inhibitors in patients with inflammatory bowel diseases or other immune-mediated diseases: a systematic review and meta-analysis. *Gastroenterology*. 2020;158(6):1554-1573. doi 10.1053/j.gastro.2020.01.001

Palmroth M., Kuuliala K., Peltomaa R., Virtanen A., Kuuliala A., Kurttila A., Kinnunen A., Leirisalo-Repo M., Silvennoinen O., Isomäki P. Tofacitinib suppresses several JAK-STAT pathways in rheumatoid arthritis *in vivo* and baseline signaling profile associates with treatment response. *Front Immunol*. 2021;12:738481. doi 10.3389/fimmu.2021.738481

Park J.Y., Pillinger M.H., Abramson S.B. Prostaglandin E_2 synthesis and secretion: the role of PGE_2 synthases. *Clin Immunol*. 2006;119(3):229-240. doi 10.1016/j.clim.2006.01.016

Prajapati P., Doshi G. An update on the emerging role of Wnt/β-catenin, SYK, PI3K/AKT, and GM-CSF signaling pathways in rheumatoid arthritis. *Curr Drug Targets*. 2023;24(17):1298-1316. doi 10.2174/0113894501276093231206064243

Rabelo F.S., da Mota L.M., Lima R.A., Lima F.A., Barra G.B., de Carvalho J.F., Amato A.A. The Wnt signaling pathway and rheumatoid arthritis. *Autoimmun Rev*. 2010;9(4):207-210. doi 10.1016/j.autrev.2009.08.003

Sherman B.T., Hao M., Qiu J., Jiao X., Baseler M.W., Lane H.C., Imamichi T., Chang W. DAVID: a web server for functional enrichment analysis (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi 10.1093/nar/gkac194

Singh A.K., Haque M., Madarampalli B., Shi Y., Wildman B.J., Basit A., Khuder S.A., Prasad B., Hassan Q., Ouseph M.M., Ahmed S. Ets-2 propagates IL-6 trans-signaling mediated osteoclast-like changes in human rheumatoid arthritis synovial fibroblast. *Front Immunol*. 2021;12:746503. doi 10.3389/fimmu.2021.746503

Smolen J.S., Aletaha D., McInnes I.B. Rheumatoid arthritis. *Lancet*. 2016;388(10055):2023-2038. doi 10.1016/S0140-6736(16)30173-8

Smolen J.S., Landewé R.B.M., Bergstra S.A., Kerschbaumer A., Sepriano A., Aletaha D., Caporali R., ... van der Helm-van Mil A., van Duuren E., Vliet Vlieland T.P.M., Westhovens R., van der Heijde D. EULAR recommendations for the management of rheumatoid arthritis: 2022 update. *Ann Rheum Dis*. 2023;82(1):3-18. doi 10.1136/ard-2022-223356

Stump K.L., Lu L.D., Dobrzanski P., Serdikoff C., Gingrich D.E., Dugan B.J., Angeles T.S., Albom M.S., Ator M.A., Dorsey B.D., Ruggeri B.A., Seavey M.M. A highly selective, orally active inhibitor of Janus kinase 2, CEP-33779. *Arthritis Res Ther*. 2011;13(2):R68. doi 10.1186/ar3329

Unterberger S., Davies K.A., Rambhatla S.B., Sacre S. Contribution of toll-like receptors and the NLRP3 inflammasome in rheumatoid arthritis pathophysiology. *Immunotargets Ther*. 2021;10:285-298. doi 10.2147/ITT.S288547

van der Kooij S.M., de Vries-Bouwstra J.K., Goekoop-Ruiterman Y.P., van Zeben D., Kerstens P.J., Gerards A.H., van Groenendael J.H., Hazes J.M., Breedveld F.C., Allaart C.F., Dijkmans B.A. Limited efficacy of conventional DMARDs after initial methotrexate failure. *Ann Rheum Dis*. 2007;66(10):1356-1362. doi 10.1136/ard.2006.066662

Wang R., Li M., Wu W., Qiu Y., Hu W., Li Z., Wang Z., Yu Y., Liao J., Sun W., Mao J., Zhu Y.Z. NAV2 positively modulates inflammatory response through Wnt/β-catenin signaling in rheumatoid arthritis. *Clin Transl Med*. 2021;11(4):e376. doi 10.1002/ctm2.376

Wang S.Y., Liu Y.Y., Ye H., Guo J.P., Li R., Liu X., Li Z.G. Circulating Dickkopf-1 is correlated with bone erosion and inflammation in rheumatoid arthritis. *J Rheumatol*. 2011;38(5):821-827. doi 10.3899/jrheum.100089

Yao C., Narumiya S. Prostaglandin-cytokine crosstalk in chronic inflammation. *Br J Pharmacol*. 2019;176(3):337-354. doi 10.1111/bph.14530

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 03.08.2025. После доработки 01.10.2025. Принята к публикации 17.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1031-1040

doi 10.18699/vjgb-25-108

Математические модели метаболизма железа: структура и функции

Н.И. Мельченко [1], И.Р. Акбердин [1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Научный центр генетики и наук о жизни, Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия

akberdin@bionet.nsc.ru

Аннотация. Математические модели представляют собой мощный теоретический инструмент для изучения сложных биологических систем. Они позволяют прослеживать неочевидные взаимодействия и проводить виртуальные эксперименты для решения практических задач. Железо играет ключевую роль в транспорте кислорода в организме млекопитающих. В то же время высокая концентрация этого микроэлемента может повреждать клеточные структуры за счет продукции активных форм кислорода, а также привести к ферроптозу (программируемая клеточная гибель в связи с железо-зависимым перекисным окислением липидов). Большой вклад в регуляцию метаболизма железа вносит иммунная система: гипоферритинемия (снижение концентрации ферритина в крови) на фоне инфекции является результатом врожденного ответа иммунной системы. В исследовании метаболизма железа многие аспекты регуляции остаются недостаточно изученными; требуется более глубокое понимание структурно-функциональной организации и динамики всех компонентов этого комплексного процесса в норме и при патологии. Важным инструментом, позволяющим выявить наиболее существенные регуляторные взаимодействия и предсказать поведение метаболической системы регуляции железа в организме человека в разных условиях, становится математическое моделирование. Данная работа представляет обзор моделей метаболизма железа, применимых к человеку, в порядке их создания, что позволяет оценить историю развития и приоритеты в моделировании метаболизма железа. Мы акцентировали внимание на постановке численных задач в анализируемых моделях, их структуре и воспроизводимости, на основе чего выделили их недостатки и преимущества. Современные модели способны численно воспроизвести множество экспериментов: гемотрансфузию, нарушение сигнального пути; мутацию в гене ферропортина; хроническое воспаление. Однако существующие математические модели метаболизма железа сложно масштабировать, и они не учитывают работу других органов и систем, в связи с чем их применение остается крайне ограниченным. Для расширения применимости компьютерных моделей в решении практических задач, связанных с метаболизмом железа в организме человека, необходимо создать масштабируемую и верифицируемую математическую модель метаболизма железа с учетом взаимодействия с другими функциональными системами человека (например, иммунной) и современных стандартов представления математических моделей биологических систем.

Ключевые слова: математическое моделирование; метаболизм железа; ферритин; гепсидин; обыкновенные дифференциальные уравнения

Для цитирования: Мельченко Н.И., Акбердин И.Р. Математические модели метаболизма железа: структура и функции. *Вавиловский журнал генетики и селекции*. 2025;29(7):1031-1040. doi 10.18699/vjgb-25-108

Финансирование. Исследование выполнено при поддержке государственного проекта № FWNR-2022-0020 Федерального исследовательского центра ИЦиГ СО РАН.

Mathematical models of iron metabolism: structure and functions

N.I. Melchenko [1], I.R. Akberdin [1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Research Center for Genetics and Life Sciences, Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia

akberdin@bionet.nsc.ru

Abstract. Mathematical models represent a powerful theoretical tool for studying complex biological systems. They provide an opportunity to track non-obvious interactions and conduct *in silico* experiments to address practical problems. Iron plays a key role in oxygen transport in the mammals. However, a high concentration of this microelement can damage cellular structures through the production of reactive oxygen species and can also lead to ferroptosis (programmed cell death associated with iron-dependent lipid peroxidation). The immune system contributes greatly to the regulation of iron metabolism – hypoferritinemia (decreased ferritin concentration in the blood) during infection –

© Мельченко Н.И., Акбердин И.Р., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

which is a result of the innate immune response. In the study of iron metabolism, many aspects of regulation remain insufficiently studied and require a deeper understanding of the structural-functional organization and dynamics of all components of this complex process in both normal and pathological conditions. Consequently, mathematical modeling becomes an important tool to identify key regulatory interactions and predict the behavior of the iron metabolism regulatory system in the human body under various conditions. This article presents a review of iron metabolism models applicable to humans presented in chronological order of their development to illustrate the evolution and priorities in modeling iron metabolism. We focused on the formulation of numerical problems in the analyzed models, their structure and reproducibility, thereby highlighting their advantages and drawbacks. Advanced models can numerically simulate various experimental scenarios: blood transfusion, signaling pathway disruption, mutation in the ferroportin gene, and chronic inflammation. However, existing mathematical models of iron metabolism are difficult to scale and do not account for the functioning of other organs and systems, which severely limits their applicability. Therefore, to enhance the utility of computational models in solving practical problems related to iron metabolism in the human body, it is necessary to develop a scalable and verifiable mathematical model of iron metabolism that considers interactions with other functional human systems (e.g., the immune system) and state-of-the-art standards for representing mathematical models of biological systems.

Key words: mathematical modeling; iron metabolism; ferritin; hepcidin; ordinary differential equations

For citation: Melchenko N.I., Akberdin I.R. Mathematical models of iron metabolism: structure and functions. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1031-1040. doi 10.18699/vjgb-25-108

Введение

Железо играет ключевую роль в транспорте кислорода у позвоночных (Pantopoulos et al., 2012). В организме человека представлено множество форм железа (Vogt et al., 2021). В плазме крови железо транспортируется в свободной, не связанной с трансферрином форме, в связанной с трансферрином форме и в составе гемоглобина. В тканях железо преимущественно находится в свободной форме либо ассоциировано с белком запаса железа – ферритином. Однако большая часть железа в организме присутствует в эритроцитах в виде гемоглобина.

Как избыток, так и недостаток железа ведет к неблагоприятным последствиям. В случае недостатка развивается железодефицитная анемия, избыток приводит к токсичным эффектам свободного железа и развитию программируемой гибели клеток, опосредованной железом, – ферроптозу (Xie et al., 2016). Поэтому в организме позвоночных существует молекулярно-генетическая система, контролирующая постоянство уровня железа. Основным белком, регулирующим обмен железа, является гепсидин. Он связывается с ферропортином (FPN) – белком, который функционирует как единственный экспортер железа у позвоночных. Связывание гепсидина приводит к убиквитинированию, интернализации и деградации FPN, ингибируя экспорт железа. Поскольку FPN высоко экспрессируется в дуоденальных энтероцитах, макрофагах, перерабатывающих железо, и гепатоцитах, то опосредованные гепсидином инактивация и деградация FPN ухудшают всасывание пищевого железа и ограничивают высвобождение запасенного железа, тем самым снижая уровни циркулирующего железа (Xu et al., 2021). Экспрессия гепсидина, в свою очередь, контролируется отрицательной обратной связью с концентрацией железа как в плазме крови, так и в гепатоцитах, а также воспалительным ответом, преимущественно опосредованным активностью IL-6 (Nemeth, Ganz, 2023).

На сегодняшний день многие аспекты метаболизма железа остаются не до конца изученными: например, транспорт негемового железа в энтероцит, аллостерическая регуляция гемоглобина, регуляция гепсидина (Ahmed et al., 2020; Nemeth, Ganz, 2023). Экспериментальные подходы не позволяют полностью раскрыть сложность и иерархическую организацию системы взаимодействий компонентов, регулирующих метаболизм железа в организме человека, поэтому актуально создание полномасштабной модели метаболизма железа, учитывающей молекулярные взаимодействия между различными органами и системами, которая позволит не только объединить в рамках единой концептуальной схемы эти уровни организации молекулярно-генетической системы метаболизма железа, но и станет теоретической основой для *in silico* исследований в области изучения особенностей структурно-функциональной организации и динамики взаимодействий между компонентами системы. Последнее, в свою очередь, послужит базисом для разработки и оценки эффективности действия лекарственных препаратов на различные терапевтические мишени системы метаболизма железа с учетом функциональных взаимодействий с иммунной системой.

В рамках этой публикации мы проводим обзор существующих моделей, оцениваем как их преимущества и недостатки, так и применимость для решения фундаментальных и прикладных аспектов исследования метаболизма железа.

Исходные модели метаболизма железа в организме человека

Математическая модель метаболизма железа (Franzone et al., 1982)

Модель P.C. Franzone с коллегами была разработана с целью численной оценки концентрации железа в различных компартментах организма, а также для изучения воздействия разных методов лечения на пациентов с анемией разного генеза. Метаболические процессы в модели распределены в следующих компартментах: слизистая оболочка кишечника, плазма крови, печень, ретикулоэндотелиальные клетки, костный мозг и эритроциты. Модель описывает поступление железа с пищей, транспорт его в плазму, запас в печени и участие в эритропоэзе. Она

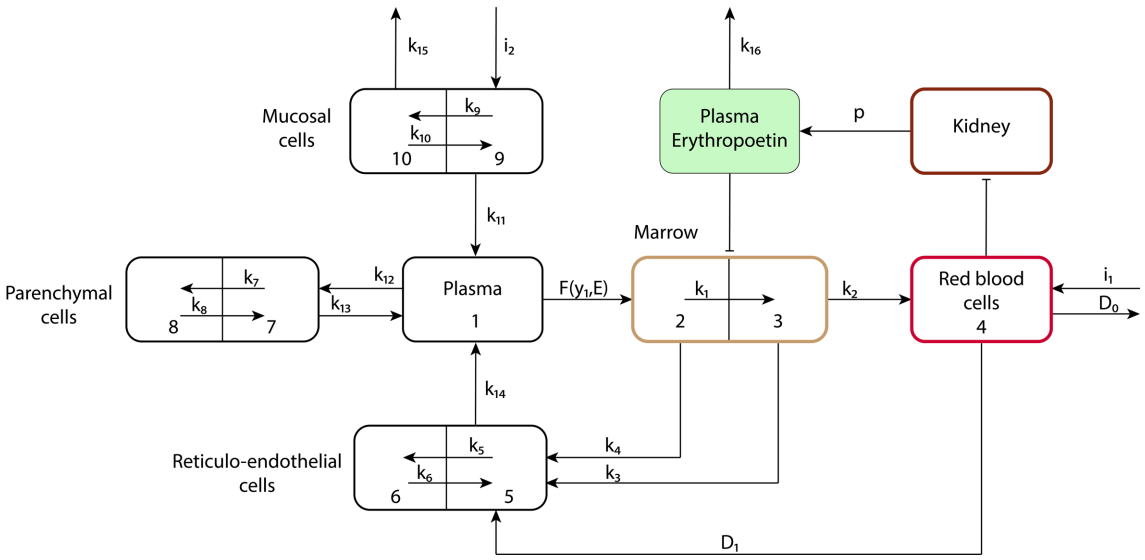


Рис. 1. Схематическое представление модели, адаптировано из (Franzone et al., 1982).

Блоки отражают количество железа в определенном органе или системе, где 1 – плазма крови; 2 – созревающие клетки крови эритроидного ряда; 3 – зрелые клетки крови эритроидного ряда; 4 – эритроциты; 5 – макрофаги; 6 – запас железа в макрофагах; 7 – внесосудистая жидкость; 8 – запас железа в гепатоцитах; 9 – эпителиальные клетки кишечника; 10 – запас железа в эпителиальных клетках кишечника. Стрелки показывают транспорт железа между органами и системами, где k_1, k_2, k_3… k_{16} – скорость транспорта железа между блоками; i_1 – приток железа вследствие донорства крови; i_2 – приток железа с пищей; D_0 – потеря железа в связи с кровотечением; D_1 – переход железа в клетки ретикулоэндотелиальной системы вследствии фагоцитоза; $F(y_1, E)$ – функция перехода железа из плазмы в клетки эритроидного ряда, где y_1 – количество железа в плазме крови, E – количество эритропоэтина; p – функция синтеза эритропоэтина.

учитывает влияние эритропоэтина на пролиферацию и созревание эритроидных клеток, а также позволяет учитывать пополнение пула железа за счет донорской крови и потерю железа вследствие кровотечений. Для контроля процесса возвращения железа из эритроцитов в плазму крови предусмотрен блок, описывающий разрушение эритроцитов ретикуло-эндотелиальными клетками. Модель также включает неэффективный гемопоэз, вследствие которого часть эритроидных клеток не способна к окончательной дифференциации (рис. 1).

С использованием этой модели численно были исследованы такие состояния, как сдача крови у здорового пациента, трансфузия крови после спленэктомии у больного с гемолитической анемией, лечение гипопластической анемии с использованием трансфузии и андрогенов.

При численном эксперименте, описывающем донорство крови у здоровых пациентов, модель показывает полное восстановление уровня гемоглобина примерно через 25–30 дней. В свою очередь, полное восстановление уровня железа в костном мозге занимает 60 дней, а на восстановление уровня железа в запасающем пуле требуется более двух месяцев, что соответствовало данным литературы на момент публикации (Wadsworth, 1955; Liedén et al., 1975), а также согласуется с результатами современных работ (Kiss et al., 2015; Ziegler et al., 2015).

На данной модели также численно исследовалась трансфузия крови после спленэктомии (удаление селезенки). Для этого использовалась модель, адаптированная путем изменения параметров, для учета состояния после спленэктомии, чтобы более точно отразить динамику железа. Однако ее валидация основана на данных лишь одного пациента.

На примере рассматриваемой модели исследовали также эффект лечения гипопластической анемии с использованием трансфузии и андрогенов, однако эти результаты потеряли свою актуальность, поскольку на сегодняшний день такая терапия не применяется (Killick et al., 2016). Авторы разработанной модели указывают, что уравнения системы могут демонстрировать жесткое поведение из-за большой разницы между численными значениями скоростей переноса при моделировании анемических состояний. Учитывая жесткость системы, с целью достижения компромисса между точностью и вычислительными ресурсами авторы использовали неявный трапециевидный метод для численного решения системы (Tavernini, 1973).

Модель Franzone с соавторами – одна из первых моделей, описывающих метаболизм железа; она существенно уступает современным моделям. В ней отсутствует описание ключевых участников метаболизма железа: гепсидина, ферритина, трансферрина и белков, регулирующих экспрессию генов, вовлеченных в метаболизм железа (iron regulatory proteins, IRP). Процесс запасания железа заметно упрощен и представлен линейным коэффициентом. Тем не менее авторам удалось смоделировать такие сложные состояния, как трансфузия крови после спленэктомии у больного с гемолитической анемией и лечение гипопластической анемии с использованием трансфузии и андрогенов. Поскольку для валидации численных расчетов модели при описании каждого из этих состояний использовались данные только по одному пациенту на анализируемое состояние, сложно оценить, насколько результаты численного моделирования применимы к популяционным данным и как они параметрически изменятся при воспроизведении данных на других пациентах.

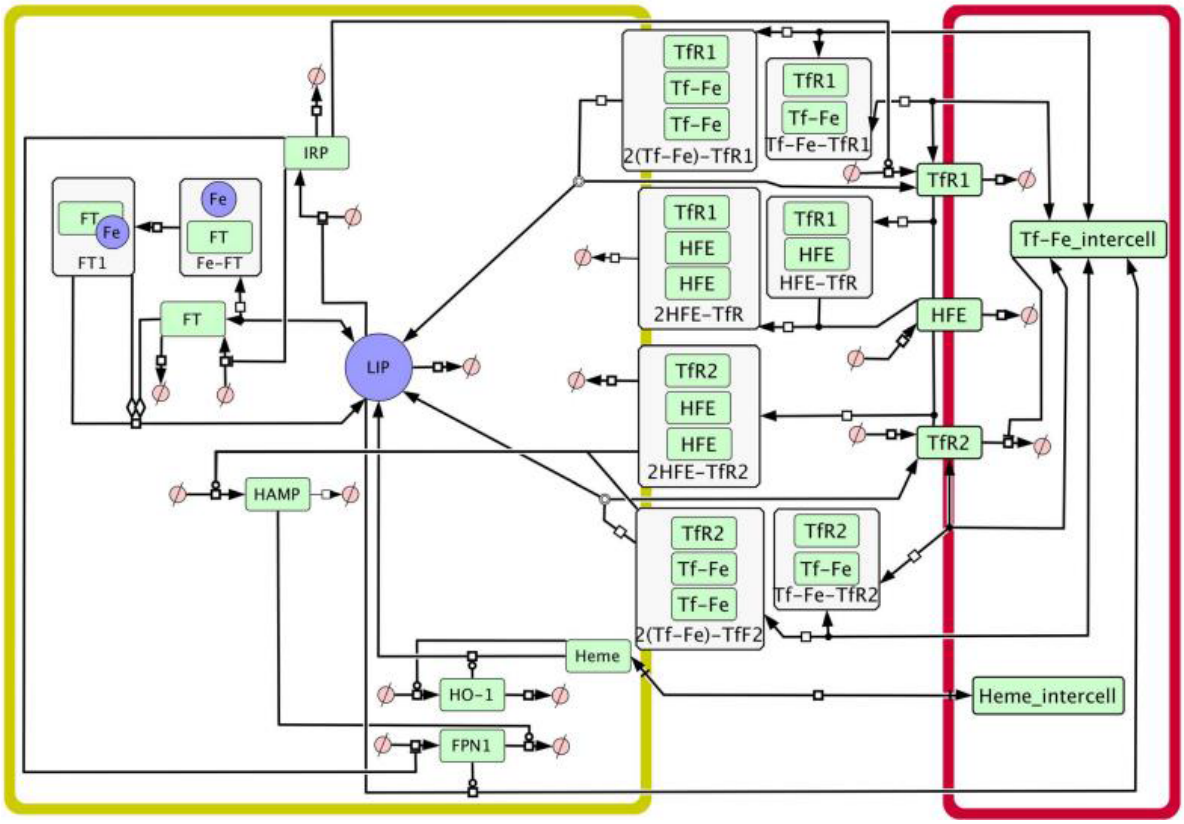


Рис. 2. Графическое представление модели в SBGN стандарте (Le Novère et al., 2009).

Стрелки отражают транспорт веществ. Желтый компартмент – гепатоцит; красный компартмент – плазма; LIP – свободный пул железа; FT – ферритин; Fe – железо; HAMP – гепсидин; Heme – гем; HO-1 – гемоксигеназа 1; IRP – белки, регулирующие железо; FPN1 – ферропортин; TfR1 – рецептор трансферрина 1; TfR2 – рецептор трансферрина 2; Tf-Fe_intercell – железо, связанное с трансферрином плазмы (Mitchell, Mendes, 2013).

Компьютерная модель метаболизма железа в печени (Mitchell, Mendes, 2013)

Математическая модель, предложенная S. Mitchell и P. Mendes в 2013 г., позволяет численно оценить процессы, связанные с транспортом железа в гепатоциты. Она дает возможность количественно предсказывать концентрацию белков, синтезирующихся в печени и регулирующих обмен железа. Модель состоит из 21 обыкновенного дифференциального уравнения и включает в себя два компартмента: гепатоцит и плазму (рис. 2).

С использованием данной модели авторы численно анализировали следующие физиологические состояния: наследственный гемохроматоз первого и третьего типа. Для воспроизведения состояния гемохроматоза первого типа был произведен виртуальный нокдаун белка регулятора гомеостаза железа человека (HFE) путем уменьшения константы синтеза в 100 раз. Количественно модель не смогла точно воспроизвести численный результат эксперимента: у мышей с такой патологией уровень железа в печени в три раза выше нормы. Это обусловлено фиксированной концентрацией межклеточного железа, связанного с трансферрином, в модели, в отличие от мышей, у которых наблюдается увеличение насыщения трансферрином в результате увеличения всасывания железа в кишечнике. Несмотря на фиксированные внеклеточные условия, модель предсказывает внутриклеточную перегрузку гепатоцитов железом. Модель гемохроматоза также смогла воспроизвести динамику, наблюдаемую в экспериментах с изменением содержания железа в пище: повышенное содержание железа в пище двукратно увеличивает экспрессию ферропортина в печени как у здоровых мышей, так и при гемохроматозе. Для воспроизведения состояния гемохроматоза третьего типа был проведен виртуальный нокдаун TfR2, тоже путем уменьшения константы синтеза в 100 раз. В результате численного анализа выявлено увеличение концентрации гепсидина и уменьшение концентрации ферропортина, что соответствует экспериментальным данным (Chua et al., 2010).

Рассматриваемая модель хорошо описывает транспорт железа в гепатоцит, учитывая запас железа, его экспорт и использование для синтеза гема. Также нам удалось полностью воспроизвести результаты авторов как в программном обеспечении COPASI (Hoops et al., 2006), так и в BioUML (Kolpakov et al., 2022). Однако модель имеет некоторые недостатки: 1) отсутствует важное звено в регуляции обмена железа, а именно влияния гепсидина на поступление железа из кишечника; 2) используется фиксированная концентрация гема и межклеточного железа, связанного с трансферрином; 3) из-за ограниченной доступности количественных клинических данных о метаболизме железа у человека для параметризации этой модели были интегрированы различные другие источники

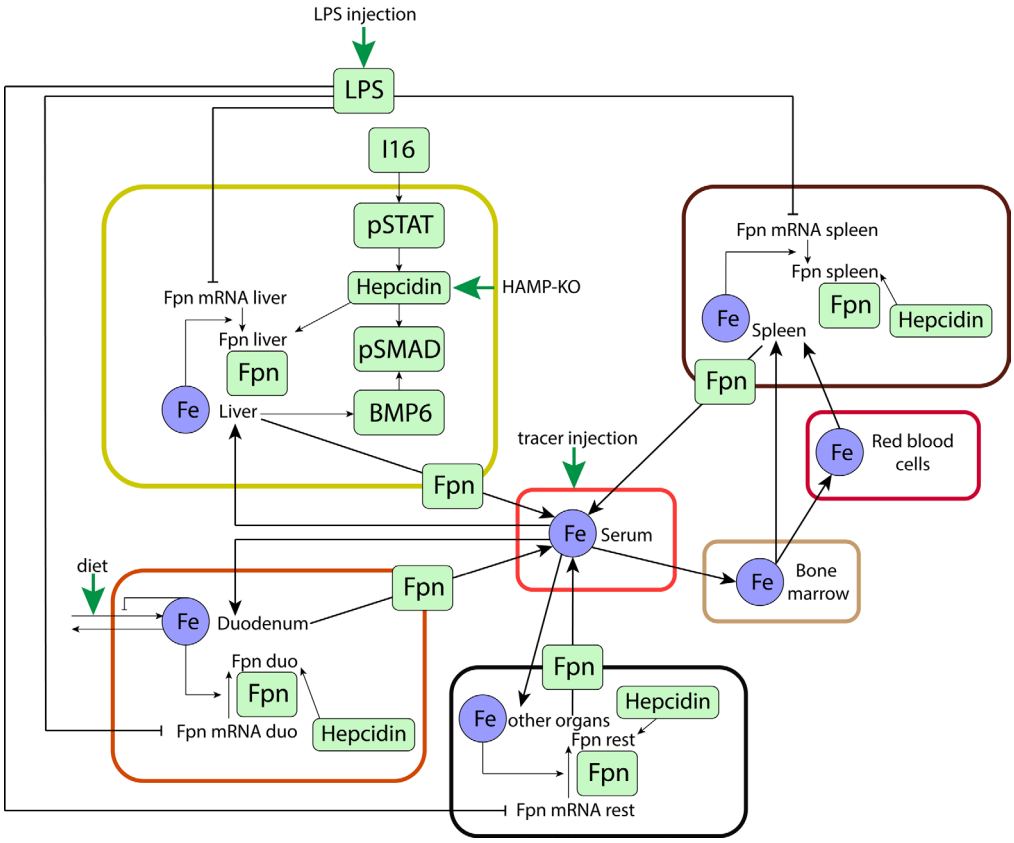


Рис. 3. Графическое представление модели (Enculescu et al., 2017) в SBGN стандарте (Le Novère et al., 2009).

LPS – липополисахариды; Fpn – ферропортин; BMP6 – костный морфогенетический белок (белок-регулятор); pSMAD, pSTAT – факторы транскрипции. Черные стрелки отражают транспорт веществ, зеленые – поступление веществ извне организма.

данных, такие как эксперименты *in vitro* и модели на животных; 4) параметры, указанные в статье, не соответствуют параметрам модели в приложении.

Моделирование системной регуляции железа при различных патологиях с учетом гепсидин-независимых механизмов (Enculescu et al., 2017)

Модель M. Enculescu с коллегами (2017) описывает обмен железа во всем организме человека, учитывая внутри- и внеклеточные механизмы регуляции метаболизма железа. Авторы сосредоточили внимание на системной регуляции метаболизма железа с помощью регуляторной оси гепсидин-ферропортин как основном механизме регуляции. В модели исследуется содержание железа в семи компартментах: сыворотка, печень, селезенка, костный мозг, эритроциты, двенадцатиперстная кишка и «другие органы» (отражает содержание железа в оставшихся тканях организма мыши). Учитываются всасывание и потеря железа в двенадцатиперстной кишке, а также потеря железа в блоке «другие органы». Модель объясняет ингибирование транскрипции ферропортина при воспалении и регуляцию его трансляции внутриклеточным железом, а также опосредованную гепсидином посттрансляционную дестабилизацию ферропортина. Экспорт железа из периферических органов контролируется экспортером железа ферропортином (Fpn), который преимущественно локализован на плазматической мембране трех типов клеток: дуоденальных энтероцитов, макрофагов и гепатоцитов. Экспрессия Fpn описывается отдельно для каждого органа и контролируется тремя регуляторными механизмами: 1) воспалительные сигналы снижают транскрипцию мРНК Fpn; 2) внутриклеточное железо усиливает трансляцию мРНК Fpn; 3) оборот белка Fpn усиливается растворимым полипептидом гепсидином.

Экспрессия гепсидина активируется железо-чувствительным путем BMP6/SMAD и воспалительным сигнальным каскадом, который включает выработку цитокинов (в первую очередь IL-6) и последующее фосфорилирование транскрипционного фактора STAT3 в гепатоцитах (рис. 3).

Для калибровки модели использовались как данные авторского исследования, так и опубликованные ранее. Всего получено 344 экспериментальных измерения. Для параметризации модели были сделаны следующие допущения: в некоторых случаях гомологичные реакции в разных компартментах протекают с одинаковыми кинетическими константами скорости. Кроме того, кинетические параметры модели промотора гена, кодирующего гепсидин, были зафиксированы на значениях, которые авторы определили ранее в системе культуры клеток HuH7.

С помощью этой модели численно были исследованы такие состояния, как: введение липополисахаридов (ЛПС) при перегрузке железом; нарушение сигнального пути BMP6; мутация в гене ферропортина, приводящая к поте-

ре способности ферропортина связываться с гепсидином; хроническое воспаление.

При численном анализе состояния введения ЛПС при перегрузке железом авторы использовали данные собственного эксперимента для валидации модели. В эксперимент были взяты самцы мышей линии C57BL/6, которые находились на диете, богатой железом (содержащей в 100 раз больше железа, чем нормальная диета), в течение четырех недель; затем им вводилась однократная доза ЛПС в дозировке 1 мкг/кг. Экспериментальные данные соответствовали предсказанию модели для большинства переменных: железо в сыворотке, печени и двенадцатиперстной кишке; содержание гепсидина в печени; концентрация мРНК BMP6; содержание pSTAT и pSMAD в печени; содержание мРНК и белка ферропортина в печени. Отклонения аппроксимации модели от экспериментальных данных наблюдаются в следующих показателях: содержание железа в селезенке и эритроцитах, концентрация ферропортина в селезенке.

В работе (Enculescu et al., 2017) также численно анализировалась динамика поведения системы регуляции железа при блокировке обратной связи гепсидина. Для этого были воссозданы две ситуации: нарушение сигнального пути BMP6; мутация в гене ферропортина, приводящая к потере способности ферропортина связываться с гепсидином.

Для воспроизведения первого состояния экспрессия SMAD была приравнена к нулю. Для воспроизведения второго состояния значения параметров, описывающих влияние гепсидина на деградацию ферропортина, тоже были приравнены к нулю. При численном расчете модели в обоих случаях наблюдается увеличение концентрации железа в сыворотке, печени и уменьшение концентрации железа в селезенке, что подтверждено экспериментальными данными. Более того, как и в экспериментах, устойчивость ферропортина к гепсидину приводила к повышенной экспрессии гепсидина, тогда как потеря передачи сигналов SMAD способствовала заметному снижению экспрессии гепсидина.

Затем авторы предположили, что гепсидин влияет на ферропортин тканеспецифичным образом. Для моделирования такой ситуации они поочередно приравнивали значение параметров, описывающих влияние гепсидина на деградацию ферропортина в разных тканях, к нулю. Результаты численного анализа показывают, что только устранение гепсидин-опосредованной регуляции ферропортина в двенадцатиперстной кишке оказывает системное влияние, увеличивая концентрацию железа в остальных органах. В то время как моделирование устойчивого к гепсидину ферропортина в печени или селезенке приводит только к местному эффекту со снижением запаса железа в соответствующем органе и минимальным изменениям в других органах. Мышиные модели с тканеспецифической устойчивостью к гепсидину до сих пор не описаны. Однако тканеспецифическая делеция FPN в клетках кишечника была изучена на мышах. Данное исследование показало, что делеция FPN в клетках кишечника приводит к серьезному дефициту железа в крови, печени и селезенке.

Исследовательский коллектив предложенной модели применил ее также для проведения *in silico* эксперимента по изучению хронического воспаления. Для этого были добавлены уравнения, описывающие кинетику ЛПС и их влияние на гепсидин. Численный анализ модели, описывающей персистирующее воспаление, показал уменьшение концентрации сывороточного железа на 85 %. При исследовании хронического воспаления концентрация железа в эритроцитах уменьшалась более длительное время по сравнению с нормой, пока не стабилизировалась через два месяца на значении, равном 10 % от нормального уровня.

В данной работе рассматриваются два механизма регуляции ферропортина: на уровне транскрипта и регуляция гепсидином. Для оценки вклада каждого звена регуляции авторы моделируют реакции ЛПС, когда было устранено либо транскрипционное, либо посттрансляционное влияние ЛПС на уровни белка ферропортина. Численный анализ показал, что отсутствие влияния гепсидина при воспалении привело к нормальному снижению уровня железа в сыворотке крови (75 % от изначальной величины). Напротив, удаление транскрипционного контроля ферропортина при воспалении уменьшило гипоферремию до 50 %. Авторы пришли к выводу, что удаление транскрипционного контроля ферропортина приводит к бóльшим отклонениям значений сывороточного железа от нормы, чем удаление контроля со стороны гепсидина. Это указывает на то, что гипоферремия возникает в результате сочетания гепсидин-зависимых и независимых механизмов.

Среди недостатков предложенной модели можно отметить различную степень точности ее параметров, отсутствие описания запасания железа, а также его связывание с ферритином.

Модель эритропоэза и метаболизма железа у человека (Schirm, Scholz, 2020)

Группа исследователей из Университета Лейпцига разработала математическую модель (Schirm, Scholz, 2020) с целью прогнозирования эффектов лечения непроверенными вариантами терапии, такими как цитотоксическая химиотерапия, поддерживаемая введением железа и эритропоэтина (ЭПО). В основе модели лежит более ранняя работа этих авторов по моделированию эритропоэза (Schirm et al., 2013), которая была расширена за счет дополнительного модуля обмена железа. Исходный модуль эритропоэза описывает динамику развития эритропоэтических клеток, отражая все основные этапы дифференцировки: стволовую клетку, бурст-образующую единицу, колониеобразующую единицу, пролиферирующие эритробласты, созревающие эритробласты, ретикулоциты. Также в данном модуле учитывается влияние химиотерапии на эритропоэз. Модуль, описывающий обмен железа, включает в себя следующие компартменты: гепсидин, не связанное с трансферрином железо в плазме, гемоглобин-катаболическую систему, запасное железо, трансферрин в связанном состоянии с железом и свободный трансферрин (рис. 4).

В рамках компьютерного моделирования были использованы некоторые упрощения комплексной физиологической системы для уменьшения числа неизвестных параметров модели или из-за отсутствия количественных

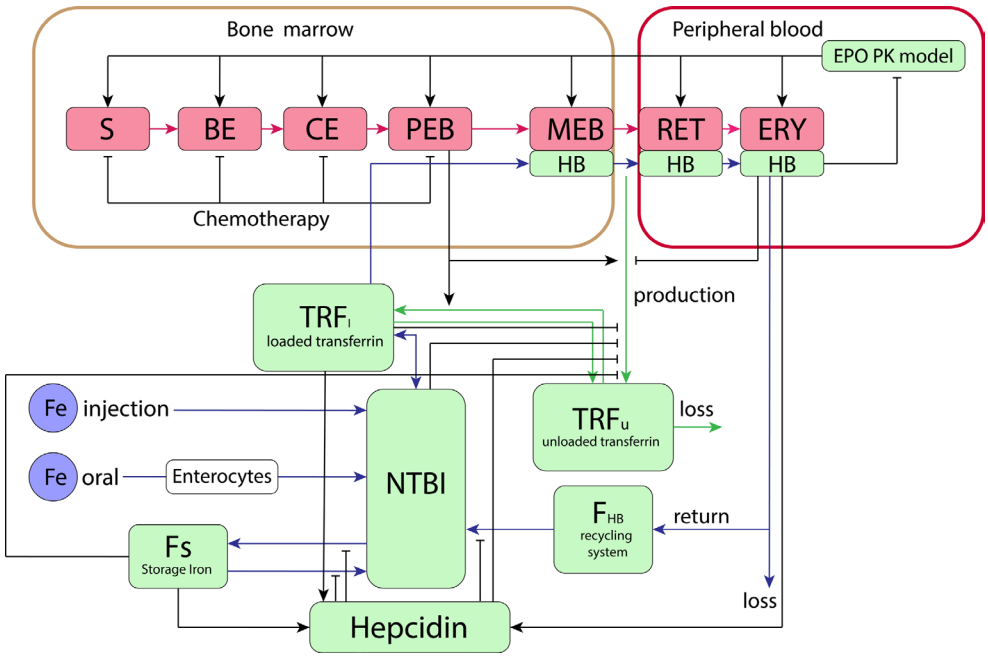


Рис. 4. Графическое представление модели (Schirm, Scholz, 2020) в SBGN стандарте (Le Novère et al., 2009).

S – стволовые клетки; BE – бурст-образующая единица; CE – колониеобразующая единица; PEB – пролиферирующие эритробласты; MEB – созревающие эритробласты; RET – ретикулоциты; ERY – эритроциты; HB – гемоглобин; NTBI – железо, не связанное с трансферрином. Синие стрелки обозначают поток железа, зеленые – поток трансферрина, красные – ход дифференцировки клеток эритроидного ряда, черные – регуляторное влияние.

данных для человека. В модели не рассматриваются отдельные пулы концентраций Fe^{2+} и Fe^{3+} из-за отсутствия данных, как и специфицированные концентрации трансферрина, обогащенного одним или двумя ионами железа.

При численном анализе предложенной модели были исследованы такие состояния, как: оральное применение железа у здоровых людей; внутривенная инъекция ЭПО при пероральном применении железа у здоровых людей; дефицит железа; внутривенное введение железа здоровым людям; кровотечение/флеботомия; хроническое воспаление; гемохроматоз.

Для валидации численных расчетов модели авторы использовали данные нескольких клинических исследований с разными схемами лечения (Rutherford et al., 1994; Souillard et al., 1996; Kiss et al., 2015). Анализировался сценарий эксперимента (Souillard et al., 1996), в котором здоровые спортсмены получали 200 МЕ/кг ЭПО в 0, 2, 4, 7 и 10-й дни без приема препаратов железа. Полученные результаты по количеству или концентрации ретикулоцитов, гемоглобина, эритроцитов, гематокрита и ферритина, как правило, отличаются от данных клинического исследования не более чем на одно стандартное отклонение.

Для подтверждения корректности результатов моделирования, описывающих прием ЭПО с препаратами железа, авторы использовали данные работы (Rutherford et al., 1994). В этом клиническом исследовании пациенты получали ЭПО в дозе 1200 МЕ/кг в неделю с разными режимами дозировки и железо в дозе 300 мг перорально ежедневно в течение 10 дней. В численном эксперименте результаты по концентрациям гематокрита, ретикулоцитов, ферритина и насыщения трансферрина очень хорошо отражали динамику этих показателей в клиническом исследовании, однако значения гемоглобина были занижены.

S. Schirm и M. Scholz провели также численный эксперимент по донорству 500 мл крови, как с поддержкой препаратами железа, так и без нее. Для валидации численных результатов использовали клиническое исследование (Kiss et al., 2015), где были получены количественные измерения динамики ферритина и гемоглобина. Результаты по концентрации ферритина, рассчитываемые в модели, отличаются от данных, полученных в клиническом исследовании, не более чем на одно стандартное отклонение в обоих сценариях, в то время как численные результаты анализа динамики гемоглобина в сценарии с поддержкой препаратами железа на большом интервале отличаются от клинических данных более чем на одно стандартное отклонение.

В работе был проведен также виртуальный эксперимент с целью создания прогноза для неиспользуемой терапии. Группа Шольца смоделировала влияние терапии CHOP-14 с поддержкой препаратами железа и ЭПО на эритропоэз и обмен железа. CHOP-14 – общепринятая терапия для лечения злокачественных форм неходжкинских лимфом, которая включает в себя такие препараты, как доксорубицин, циклофосфамид, винкристин и преднизолон. В настоящее время терапия расширена до R-CHOP и включает также ритуксимаб (Phan et al., 2010). Такая терапия является гематотоксичной, поэтому авторы рассмотрели возможность расширения терапии препаратами железа и ЭПО. Для верификации расчетов модели в *in silico* эксперименте при химиотерапии без включения препаратов железа и ЭПО использовались данные немецкой исследовательской группы по неходжкинской лимфоме высо-

кой степени злокачественности (Pfreundschuh et al., 2004). Согласно *in silico* эксперименту, добавление препаратов железа вместе с ЭПО у пациентов, принимающих терапию CHOP-14, тормозит снижение концентрации гемоглобина. При применении препаратов железа с ЭПО на 3, 7 и 21-й дни концентрация гемоглобина на 80-й день составляет примерно 11.2 г/дл, тогда как без поддерживающей терапии она равна примерно 10.7 г/дл. При еженедельном введении препаратов железа вместе с ЭПО концентрация гемоглобина начиная с 45-го дня восстанавливается, достигая 12.5 г/дл к 80-му дню, в то время как без поддерживающей терапии она падает до 10.7 г/дл. Важно отметить, что значительный вклад в восстановление концентрации гемоглобина вносят препараты ЭПО, поскольку результаты численного расчета поддерживающей терапии лишь препаратами железа практически не отличаются от результатов без нее.

Авторы придерживаются модульного подхода и развивают свою предыдущую работу, добавляя в нее новые блоки. Большим плюсом является валидация с использованием большого количества данных из разных исследований. Рассматриваемая модель показывает хорошую сходимость с клиническими исследованиями, поскольку в большинстве случаев различия между численными данными модели и результатами клинических исследований не превышают одно стандартное отклонение. К минусам можно отнести пониженный уровень гемоглобина, предсказанный моделью, в сравнении с экспериментальными измерениями.

Модель секвестрации железа ферритином (Masison, Mendes, 2023)

J. Masison и P. Mendes разработали модель, описывающую связывание ионов железа белком ферритина. Ферритин состоит из 24 субъединиц и способен связывать около 4300 атомов железа на молекулу ферритина. Ферритин – один из важных участников метаболизма железа, поэтому модели обмена железа должны включать его. Модель J. Masison и P. Mendes позволяет интегрировать взаимодействие ферритина с ионами железа в более комплексные модели.

Модель рассматривает: 1) как железо, связанное с ферритином, влияет на динамику секвестрации железа; 2) как модель секвестрации железа с константами, полученными экспериментально *in vitro*, может численно воспроизводить экспериментальные результаты, полученные на клеточных линиях; 3) влияние состава субъединиц ферритина на скорость секвестрации железа; 4) зависимость динамики высвобождения железа из ферритина от концентрации свободного железа и ферритина в клетке.

В модели учитываются четыре химические сущности: LIP – пул лабильного железа, растворимое или легкорастворимое двухвалентное железо в цитоплазме; DFP – пероксокомплекс, содержащий два атома железа; ядро – железо, входящее в состав минерализованного ядра ферригидрита; FT – 24 субъединицы ферритина. Модель включает четыре реакции, три из которых описывают процесс секвестрации железа ферритином (окисление – преобразует два LIP в один DFP; нуклеация – преобразует два DFP в новое ядро кристалла; минерализация – добавляет

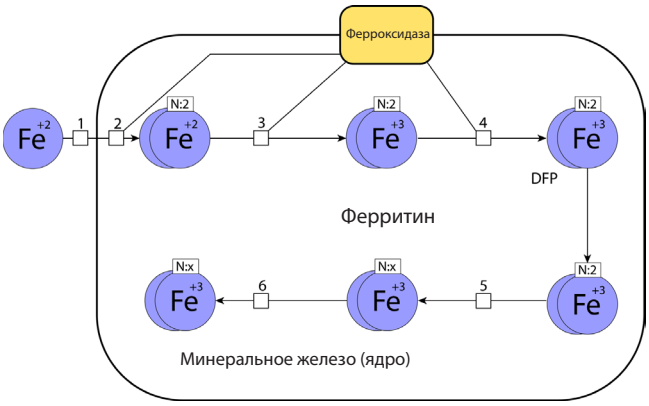


Рис. 5. Адаптировано графическое представление процесса секвестрации железа (Masison, Mendes, 2023) в SBGN стандарте (Le Novère et al., 2009).

Представлены следующие реакции: 1 – транспорт Fe^{2+} в ферритин; 2 – соединение Fe^{2+} с ферроксидазой, формирование DFP; 3, 4 – окисление Fe^{2+}; 5 – нуклеация; 6 – минерализация.

один DFP к существующему ядру), а одна реакция описывает деградацию промежуточного продукта (восстановление – преобразует один DFP обратно в два LIP). Процесс секвестрации графически представлен на рис. 5. Авторы упрощают и объединяют некоторые из его звеньев, для того чтобы составить систему дифференциальных уравнений, отражающую с достаточной точностью данный биохимический процесс. В то же время авторы избегают излишней детализации и не перегружают модель переменными.

Первая реакция описывает окисление LIP до DFP и представлена функцией Хилла:

$$\frac{kcat \times \frac{H + rO}{24 + rO} \times FT \times LIP^n}{Km^n + LIP^n}.$$

Здесь *kcat* – число каталитического оборота; Km – константа Михаэлиса; *n* – коэффициент Хилла. Молекула ферритина состоит из 24 субъединиц двух разных типов – H и L, и только H субъединицы имеют активный центр ферроксидазы, поэтому молекулы с разным содержанием субъединиц имеют разную скорость окисления. Для учета этого используются два дополнительных параметра: H – количество субъединиц H (число от 0 до 24); *rO* – масштабируемый коэффициент, представляющий эффективность окисления L-гомополимера.

Параметр *rO* включен авторами модели, поскольку, несмотря на то что субъединица L не обладает известной ферроксидазной активностью, L-гомополимеры, согласно экспериментальным данным, по-прежнему катализируют образование трехвалентного железа в составе FT, хотя и со снижением скорости более чем на четверть (Carmona et al., 2014). Данные о том, как окисление происходит в отсутствие субъединицы H, и о соответствующем значении *rO* ограничены, поэтому значение *rO* установлено эмпирически равным двум.

Вторая реакция деградация DFP следует из закона действующих масс:

$$k_{deg} \times DFP.$$

Третья реакция нуклеации

$$kcat \times DFP^2 \times FT \times \frac{L + rN}{24 + rN} \times \frac{Ki^n}{Ki^n + core^n}$$

описывает процесс образования нового кристалла из двух молекул DFP внутри молекулы ферритина. Внутри молекулы ферритина с уже существующим ядром также могут образовываться новые ядра. Уравнение получено эмпирически на основе закона действующих масс. Коэффициенты L и *rN* отражают, насколько состав субъединиц ферритина влияет на нуклеацию. В связи с недостатком информации об этом процессе коэффициент *rN* подобран таким образом, чтобы вносить значительно меньший вклад в диапазон скорости нуклеации, чем вносит коэффициент *rO* в диапазон скорости окисления. Последний множитель включает в себя константу ингибирования и коэффициент Хилла и позволяет учитывать уменьшение вероятности образования новых кристаллов с увеличением размера уже существующего ядра.

Четвертая реакция минерализации:

$$\frac{kcat \times DFP \times core}{Km + DFP} \times \frac{Ki^n}{Ki^n + core^n} \times \frac{4300^m - apc^m}{4300^m}.$$

Скорость этой реакции достигает максимума при 1500–2000 атомах железа на ядро и уменьшается при дальнейшем росте ядра (Harrison et al., 1974). Второй множитель нужен, чтобы учесть этот процесс, а третий множитель устремляет скорость к нулю при приближении количества атомов железа на ядро (apc) к максимально допустимому – 4300.

Для валидации модели использовались результаты экспериментов из разных лабораторий с разными условиями. Получены некоторые различия в первых 20 секундах с экспериментальными данными: более сильный кооперативный эффект в скорости минерализации DFP; более быстрое достижение стационарной концентрации. Поскольку модель ориентирована на клеточные процессы, где ключевым временным масштабом выступают минуты (и более), выявленное расхождение с экспериментальными данными нельзя считать существенным.

Авторы провели виртуальный эксперимент с исследованием влияния атомов железа в ядре на скорость минерализации. Моделирование выявило зависимость скорости минерализации с течением времени от начальных значений атомов железа на ядро (atoms per core, apc). Как правило, кривые, показывающие скорость минерализации, делятся на три группы в зависимости от начального apc. В первой группе (<1000 apc) скорость минерализации начинается с низкой, далее увеличивается по мере накопления железа внутри ферритина, а затем падает по мере уменьшения концентрации железа в растворе. Во второй группе (1000–3000 apc) скорость минерализации начинается с высокой, а затем быстро падает в связи со снижением концентрации железа в растворе. В третьей группе (>3000 apc) скорость минерализации уменьшается по мере продолжения моделирования, так как в этом случае накопление железа в ядре ферритина замедляет скорость новой минерализации.

Затем группа P. Mendes исследовала модель при концентрациях ферритина и железа, соответствующих клеткам млекопитающих. Она была добавлена как модульный блок в ранее разработанную ими модель метаболизма железа в гепатоците. Авторы сообщают, что поведение системы качественно аналогично модели до расширения, однако расширенная модель позволяет больше понять и оценить механизмы запасания железа. Из-за повышенной детализации новой модели видно, что пик железа, связанного с ферритином, обусловлен повышением концентрации DFP, а не минерализованного ядра. Это важно учитывать, поскольку DFP легче высвобождается обратно в цитоплазму. Численные результаты моделей отличаются как на временном интервале, так и в состоянии равновесия. Наибольшие различия видны после 1000 секунд моделирования. В исходной модели наблюдается постепенное увеличение содержания ферритин-связанного железа, тогда как в новой модели его концентрация уменьшается. Авторы оригинальной работы выдвигают гипотезу, что это может быть обусловлено новой кинетикой запасания железа, которая способствует уменьшению доступного железа за счет буферизации ферритином, в то время как в исходной модели на кинетику доступного железа основное влияние оказывали другие механизмы.

Заключение

Анализ представленных математических моделей метаболизма железа выявляет тенденцию к прогрессивному увеличению их структурной сложности с течением времени (табл. S1 Приложения)[1]. С развитием исследований возрастает как количество уравнений, так и число параметров в моделях, что свидетельствует о стремлении к более точному и детальному описанию биологических процессов. Новые модели позволяют моделировать более широкий спектр физиологических и патологических состояний, расширяя возможности для проведения *in silico* экспериментов. Исключением является последняя модель секвестрации железа ферритином (Masison, Mendes, 2023), которая реализована согласно модульному принципу и создавалась с целью интеграции в более комплексные системы. Такой подход обеспечивает гибкость и масштабируемость модели, что важно для дальнейшего развития и внедрения в мультифакторные модели метаболизма железа.

Для более глубокого понимания метаболизма железа необходимо учитывать взаимодействие с иммунной системой, играющей одну из ключевых ролей в регуляции обмена железа (Vogt et al., 2021). В то же время снижение доступности железа для патогенов, а также продукция активных форм кислорода могут существенно повлиять на динамику инфекционных заболеваний (Weinberg, 2009). Включение этих факторов в математические модели позволит проводить виртуальные эксперименты, анализирующие влияние различных инфекций на метаболизм железа, и оценивать долгосрочные последствия таких взаимодействий. Это знание может быть критически важным для разработки новых подходов к лечению заболеваний, связанных с нарушением обмена железа, а также для понимания патогенеза таких состояний, как анемия при хронических заболеваниях, гемохроматоз или поствирусные синдромы, например постковидный синдром.

[1] Табл. S1 Приложения см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx38.pdf

Таким образом, интеграция данных о взаимодействии между иммунной системой и метаболизмом железа не только углубит наше понимание этих процессов, но и может открыть новые возможности для клинических исследований и терапевтических стратегий. В связи с этим создание детализированной модели метаболизма железа, учитывающей взаимодействия с иммунной системой, представляет собой актуальную задачу, решение которой позволит лучше понять взаимодействие двух комплексных систем и предположить ключевые звенья в патологии метаболизма железа при различных заболеваниях.

Список литературы / References

Ahmed M.H., Ghatge M.S., Safo M.K. Hemoglobin: structure, function and allostery. In: Hoeger U., Harris J. (Eds) Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins. Subcellular Biochemistry. Vol. 94. Springer, 2020;345-382. doi 10.1007/978-3-030-41769-7_14

Carmona U., Li L., Zhang L., Knez M. Ferritin light-chain subunits: key elements for the electron transfer across the protein cage. *Chem Commun*. 2014;50:15358-15361. doi 10.1039/C4CC07996E

Chua A.C.G., Delima R.D., Morgan E.H., Herbison C.E., Tirnitz-Parker J.E.E., Graham R.M., Fleming R.E., Britton R.S., Bacon B.R., Olynyk J.K., Trinder D. Iron uptake from plasma transferrin by a transferrin receptor 2 mutant mouse model of haemochromatosis. *J Hepatol*. 2010;52(3):425-431. doi 10.1016/j.jhep.2009.12.010

Enculescu M., Metzendorf C., Sparla R., Hahnel M., Bode J., Muckenthaler M.U., Legewie S. Modelling systemic iron regulation during dietary iron overload and acute inflammation: role of hepcidin-independent mechanisms. *PLoS Comput Biol*. 2017;13(1):e1005322. doi 10.1371/journal.pcbi.1005322

Franzone P.C., Paganuzzi A., Stefanelli M. A mathematical model of iron metabolism. *J Math Biol*. 1982;15(2):173-201. doi 10.1007/BF00275072

Harrison P.M., Hoy T.G., Macara I.G., Hoare R.J. Ferritin iron uptake and release. Structure–function relationships. *Biochem J*. 1974;143(2):445-451. doi 10.1042/bj1430445

Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N., Singhal M., Xu L., Mendes P., Kummer U. COPASI – a COmplex PAthway SImulator. *Bioinformatics*. 2006;22(24):3067-3074. doi 10.1093/bioinformatics/btl485

Killick S.B., Bown N., Cavenagh J., Dokal I., Foukaneli T., Hill A., Hillmen P., Ireland R., Kulasekararaj A., Mufti G., Snowden J.A., Samarasinghe S., Wood A., Marsh J.C.W. Guidelines for the diagnosis and management of adult aplastic anaemia. *Br J Haematol*. 2016;172(2):187-207. doi 10.1111/bjh.13853

Kiss J.E., Brambilla D., Glynn S.A., Mast A.E., Spencer B.R., Stone M., Kleinman S.H., Cable R.G.; National Heart, Lung, and Blood Institute (NHLBI) Recipient Epidemiology and Donor Evaluation Study-III (REDS-III). Oral iron supplementation after blood donation: a randomized clinical trial. *JAMA*. 2015;313(6):575-583. doi 10.1001/jama.2015.119

Kolpakov F., Akberdin I., Kiselev I., Kolmykov S., Kondrakhin Y., Kulyashov M., Kutumova E., Pintus S., Ryabova A., Sharipov R., Yevshin I., Zhatchenko S., Kel A. BioUML – towards a universal research platform. *Nucleic Acids Res*. 2022;50(W1):W124-W131. doi 10.1093/nar/gkac286

Le Novère N., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., … Sander C., Sauro H., Snoep J.L., Kohn K., Kitano H. The systems biology graphical notation. *Nat Biotechnol*. 2009;27(8):735-741. doi 10.1038/nbt.1558

Liedén G., Höglund S., Ehn L. Changes in certain iron metabolism variables after a single blood donation. *Acta Med Scand*. 1975;197(1-2):27-30. doi 10.1111/j.0954-6820.1975.tb04873.x

Masison J., Mendes P. Modeling the iron storage protein ferritin reveals how residual ferrihydrite iron determines initial ferritin iron sequestration kinetics. *PLoS One*. 2023;18(2):e0281401. doi 10.1371/journal.pone.0281401

Mitchell S., Mendes P. A computational model of liver iron metabolism. *PLoS Comput Biol*. 2013;9(11):e1003299. doi 10.1371/journal.pcbi.1003299

Nemeth E., Ganz T. Hepcidin and iron in health and disease. *Annu Rev Med*. 2023;74:261-277. doi 10.1146/annurev-med-043021-032816

Pantopoulos K., Porwal S.K., Tartakoff A., Devireddy L. Mechanisms of mammalian iron homeostasis. *Biochemistry*. 2012;51(29):5705-5724. doi 10.1021/bi300752r

Pfreundschuh M., Trümper L., Kloess M., Schmits R., Feller A.C., Rübe C., Rudolph C., Reiser M., Hossfeld D.K., Eimermacher H., Hasenclever D., Schmitz N., Loeffler M.; German High-Grade Non-Hodgkin's Lymphoma Study Group. Two-weekly or 3-weekly CHOP chemotherapy with or without etoposide for the treatment of elderly patients with aggressive lymphomas: results of the NHL-B2 trial of the DSHNHL. *Blood*. 2004;104(3):634-641. doi 10.1182/blood-2003-06-2095

Phan J., Mazloom A., Medeiros L.J., Zreik T.G., Wogan C., Shihadeh F., Rodriguez M.A., Fayad L., Fowler N., Reed V., Horace P., Dabaja B.S. Benefit of consolidative radiation therapy in patients with diffuse large B-cell lymphoma treated with R-CHOP chemotherapy. *J Clin Oncol*. 2010;28(27):4170-4176. doi 10.1200/JCO.2009.27.3441

Rutherford C.J., Schneider T.J., Dempsey H., Kirn D.H., Brugnara C., Goldberg M.A. Efficacy of different dosing regimens for recombinant human erythropoietin in a simulated perisurgical setting: the importance of iron availability in optimizing response. *Am J Med*. 1994;96(2):139-145. doi 10.1016/0002-9343(94)90134-1

Schirm S., Scholz M. A biomathematical model of human erythropoiesis and iron metabolism. *Sci Rep*. 2020;10(1):8602. doi 10.1038/s41598-020-65313-5

Schirm S., Engel C., Loeffler M., Scholz M. A biomathematical model of human erythropoiesis under erythropoietin and chemotherapy administration. *PLoS One*. 2013;8(6):e65630. doi 10.1371/journal.pone.0065630

Souillard A., Audran M., Bressolle F., Gareau R., Duvallet A., Chanal J.L. Pharmacokinetics and pharmacodynamics of recombinant human erythropoietin in athletes. Blood sampling and doping control. *Br J Clin Pharmacol*. 1996;42(3):355-364. doi 10.1046/j.1365-2125.1996.41911.x

Tavernini L. Linear multistep methods for the numerical solution of Volterra functional differential equations. *Appl Anal*. 1973;3(2):169-185. doi 10.1080/00036817308839063

Vogt A.-C.S., Arsiwala T., Mohsen M., Vogel M., Manolova V., Bachmann M.F. On iron metabolism and its regulation. *Int J Mol Sci*. 2021;22(9):4591. doi 10.3390/ijms22094591

Wadsworth G.R. Recovery from acute haemorrhage in normal men and women. *J Physiol*. 1955;129(3):583-593. doi 10.1113/jphysiol.1955.sp005380

Weinberg E.D. Iron availability and infection. *Biochim Biophys Acta*. 2009;1790(7):600-605. doi 10.1016/j.bbagen.2008.07.002

Xie Y., Hou W., Song X., Yu Y., Huang J., Sun X., Kang R., Tang D. Ferroptosis: process and function. *Cell Death Differ*. 2016;23(3):369-379. doi 10.1038/cdd.2015.158

Xu Y., Alfaro-Magallanes V.M., Babitt J.L. Physiological and pathophysiological mechanisms of hepcidin regulation: clinical implications for iron disorders. *Br J Haematol*. 2021;193(5):882-893. doi 10.1111/bjh.17252

Ziegler A.K., Grand J., Stangerup I., Nielsen H.J., Dela F., Magnussen K., Helge J.W. Time course for the recovery of physical performance, blood hemoglobin, and ferritin content after blood donation. *Transfusion*. 2015;55(4):898-905. doi 10.1111/trf.12926

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 26.05.2025. После доработки 08.08.2025. Принята к публикации 15.08.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1041-1050

doi 10.18699/vjgb-25-109

Идентификация и анализ сетевой структуры связей между компонентами иммунной системы у детей

Д.С. Гребенников [1, 2, 3], А.П. Топтыгина[4], Г.А. Бочаров [1, 2, 3]

[1] Институт вычислительной математики им. Г.И. Марчука Российской академии наук (ИВМ РАН), Москва, Россия
[2] Отделение Московского центра фундаментальной и прикладной математики в ИВМ РАН, Москва, Россия
[3] Первый Московский государственный медицинский университет им. И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский университет), Москва, Россия
[4] Московский научно-исследовательский институт эпидемиологии и микробиологии им. Г.Н. Габричевского Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Москва, Россия

g.bocharov@inm.ras.ru

Аннотация. Идентификация связей между различными функциональными компонентами иммунной системы представляет собой чрезвычайно актуальную задачу современной иммунологии. Это необходимо для понимания механизмов динамики и исхода инфекционных и онкологических заболеваний при реализации системно-биологического подхода. Параметры, характеризующие иммунный статус человека, отличаются большой размерностью пространства состояний при малой мощности выборки. Для изучения сетевой топологии иммунной системы нами использованы ранее опубликованные оригинальные данные измерений показателей иммунного статуса у 19 здоровых индивидуумов – детей, 9 мальчиков и 10 девочек, в возрасте от одного до двух лет: популяций иммунных клеток (42 субпопуляции), полученных с помощью проточной цитометрии; уровней цитокинов (13 типов), полученных методами мультиплексного анализа; уровня антител (4 типа), определенных с помощью иммуноферментного анализа. Для корректного (статистически значимого) определения корреляционных связей между измеряемыми переменными и построения графа сетевой топологии может быть использован подход, который учитывает малый размер множества данных. В нашей работе был реализован и исследован подход, в основе которого лежит регуляризированный алгоритм скорректированных разреженных частных корреляций (DSPC) оценивания разреженных частных корреляций и идентификации сетевой структуры взаимосвязей в иммунной системе по данным иммунного статуса здоровых детей, включающего набор показателей субпопуляций клеток иммунной системы, уровня цитокинов и антител. Для разных уровней статистической значимости были построены тепловые карты частных корреляций, выполнена визуализация сетей частных корреляций в виде графов и проведен анализ их топологических характеристик. Получено, что при ограниченной выборке измерений выбор порога для уровня статистической значимости имеет принципиальное значение для формирования матрицы частных корреляций. Окончательная верификация иммунологически корректной структуры связей требует как увеличения размера выборки, так и сопряжения с априорными механизменными представлениями и моделями функционирования компонент иммунной системы. Результаты могут быть использованы для выбора мишеней терапии и формирования комбинированных воздействий.

Ключевые слова: иммунная система; иммунный статус; корреляционный анализ; частные корреляции; сетевая топология; графы; алгоритм DSPC

Для цитирования: Гребенников Д.С., Топтыгина А.П., Бочаров Г.А. Идентификация и анализ сетевой структуры связей между компонентами иммунной системы у детей. *Вавиловский журнал генетики и селекции*. 2025;29(7):1041-1050. doi 10.18699/vjgb-25-109

Финансирование. Работа выполнена при финансовой поддержке Российского научного фонда, проект № 23-11-00116 (построение сетей корреляций и анализ топологии графов связей) и, частично, Отделения Московского центра фундаментальной и прикладной математики в ИВМ РАН (Соглашение с Минобрнауки России № 075-15-2025-347) (базовый статистический анализ данных).

Identification and analysis of the connection network structure between the components of the immune system in children

D.S. Grebennikov [1, 2, 3], A.P. Toptygina[4], G.A. Bocharov [1, 2, 3]

[1] Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences (INM RAS), Moscow, Russia
[2] Moscow Center of Fundamental and Applied Mathematics at INM RAS, Moscow, Russia
[3] Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), Moscow, Russia
[4] Gabrichevsky Research Institute for Epidemiology and Microbiology, Moscow, Russia

g.bocharov@inm.ras.ru

© Гребенников Д.С., Топтыгина А.П., Бочаров Г.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. Identification of the connections between the various functional components of the immune system is a crucial task in modern immunology. It is key to implementing the systems biology approach to understand the mechanisms of dynamic changes and outcomes of infectious and oncological diseases. The data characterizing an individual's immune status typically have a high-dimensional state space and a small sample size. To study the network topology of the immune system, we utilized previously published original data from Toptygina et al. (2023), which included measurements of the immune status in 19 healthy individuals (children, 9 boys and 10 girls, aged 1 to 2 years), i.e., the immune cells (42 subpopulations) obtained by flow cytometry; cytokine levels (13 types) obtained by multiplex analysis; and antibody levels (4 types) determined by using enzyme immunoassay. To correctly identify statistically significant correlations between the measured variables and construct the respective network graph, it is necessary to use an approach that takes into account the small size of the dataset. In this study, we implemented and analyzed an approach based on the regularized debiased sparse partial correlation (DSPC) algorithm to evaluate sparse partial correlations and identify the network structure of relationships in the immune system of healthy individuals (children) based on immune status data, which includes a set of indicators for subpopulations of immune cells, cytokine levels, and antibodies. For different levels of statistical significance, heatmaps of the partial correlations were constructed. The graph visualization of the DSPC networks was performed, and their topological characteristics were analyzed. It is found that with a limited measurements sample, the choice of a statistical significance threshold critically affects the structure of the partial correlations matrix. The final verification of the immunologically correct structure of the correlation-based network requires both an increase in the sample size and consideration of a priori mechanistic views and models of the functioning of the immune system components. The results of this analysis can be used to select the therapy targets and design combination therapies.

Key words: immune system; immune status; correlation analysis; partial correlations; network topology; graphs; DSPC algorithm

For citation: Grebennikov D.S., Toptygina A.P., Bocharov G.A. Identification and analysis of the connection network structure between the components of the immune system in children. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1041-1050. doi 10.18699/vjgb-25-109

Введение

Иммунная система человека, основная функция которой состоит в контроле антигенного гомеостаза внутренней среды организма, представляет собой систему с распределенными параметрами, отражающими пространственную организацию, фенотипическую и клональную структуру составляющих ее клеточных популяций. Клетки иммунной системы непрерывно взаимодействуют друг с другом, и баланс процессов усиления или ослабления их активности лежит в основе развития продуктивных или абортивных реакций (Ng et al., 2013). Идентификация структуры сетей клеточных взаимосвязей необходима для понимания механизмов динамики и исхода инфекционных и онкологических заболеваний при реализации системно-биологического подхода. Примером изучения сети связей (сетевой топологии) между популяциями клеточных компонент иммунной системы является идентификация социальной архитектуры взаимодействий иммунных клеток на основе данных количественной протеомики (Rieckmann et al., 2017). Формирование сетевой топологии связано с построением графа, вершинам которого соответствуют специфические популяции клеток иммунной системы, а ребрам – связи разнообразной природы между соответствующими вершинами.

К настоящему времени разработано большое число (около 100 документированных) методов анализа структурной организации межклеточных взаимодействий по данным разнообразной природы, включая пространственную и клеточную транскриптомику, экспрессию лиганд-рецепторов, а также внутриклеточных сигнальных компонент (Armingol et al., 2024). Результатом их применения будет оценка индексов связи или коммуникации между клетками, по которым могут строиться графы сети связей. Для оценки выраженности межклеточных связей могут быть использованы как биофизические и биохимические принципы, так и методы статистического анализа данных в сопряжении с машинным обучением.

Построению количественного интерактома иммунных клеток на основе рецепторных белков, экспрессируемых на их поверхности, посвящена работа (Shilts et al., 2022), в которой реализованы графы набора физических взаимосвязей между клетками иммунной системы в основных органах человека с использованием технологий мультиплексного иммунного и транскриптомного анализа, генетических баз данных и биохимических методов скрининга взаимодействий между клетками. Визуализация данных транскриптомного анализа в виде графа, отражающего ко-экспрессию генов, является интегративной частью современной системной вакцинологии (Cortese et al., 2025).

Целью данной работы была реализация нового подхода к идентификации сетевой структуры взаимосвязей в иммунной системе здорового человека по результатам корреляционного анализа опубликованных ранее данных иммунного статуса детей в возрасте от одного до двух лет, включающего набор показателей субпопуляций клеток иммунной системы, уровня цитокинов и антител (Toptygina et al., 2023). В задачи исследования входит корреляционный анализ данных иммунного статуса детей для построения тепловых карт частных корреляций, визуализация сетей частных корреляций в виде графов и анализ топологических характеристик.

Настоящая работа состоит из трех разделов. В разделе «Материалы и методы» описаны особенности исходных данных, методы корреляционного анализа данных, подход к формированию на их основе сетевой структуры взаимосвязей и задачи исследования топологических свойств

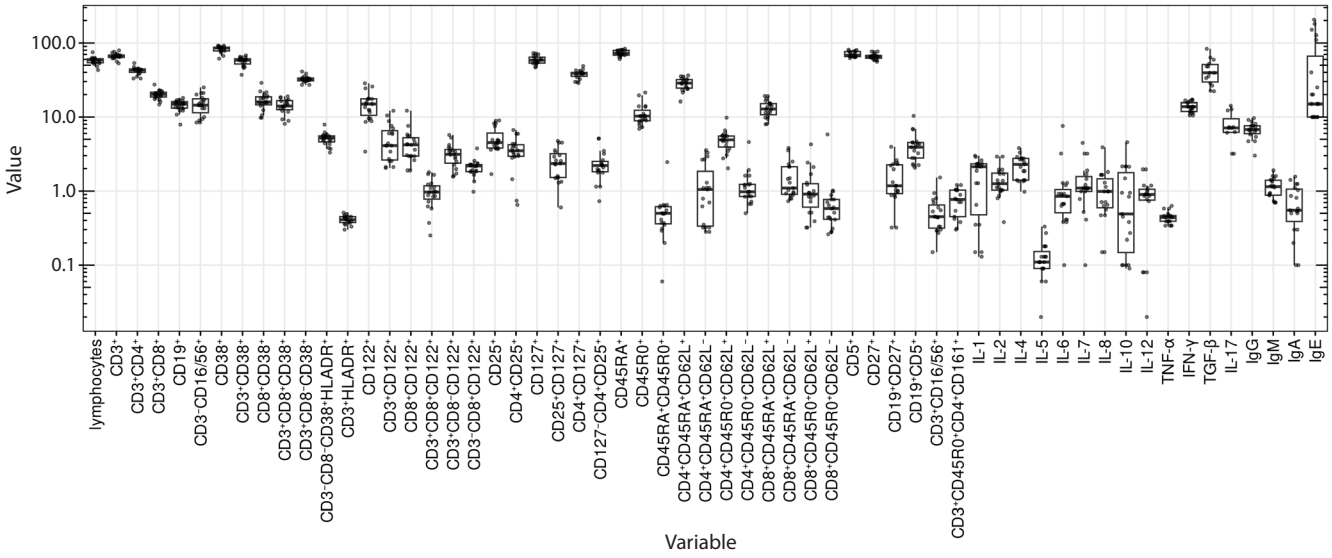


Рис. 1. Данные показателей иммунного статуса у здоровых индивидуумов – детей в возрасте от одного до двух лет (адаптировано из: Toptygina et al., 2023).

Представлены отдельные измерения, медианные значения выборки и 25–75 % квартили. По оси абсцисс – названия показателей иммунного статуса. По оси ординат – доля клеток (%) и уровень цитокинов (пг/мл), иммуноглобулины A, M, G (г/л), IgE (МЕ/мл).

соответствующих графов. Выполнен анализ главных компонент. В разделе «Результаты» приведены результаты построения сетей при различных параметрах статистической значимости корреляций, иммунологическая интерпретация соответствующих сетевых топологий, проведен анализ устойчивости результатов. Обсуждение результатов работы представлено в разделе «Обсуждение».

Материалы и методы

Данные иммунного статуса. Для изучения сетевой топологии иммунной системы нами использованы ранее опубликованные оригинальные данные (Toptygina et al., 2023). Они представляют собой массив измерений показателей иммунного статуса у 19 здоровых индивидуумов – детей в возрасте от одного до двух лет: популяций иммунных клеток (42 субпопуляции), полученных с помощью проточной цитометрии; уровней цитокинов (13 типов), полученных методами мультиплексного анализа; уровня антител (4 типа), определенных с помощью иммуноферментного анализа. Данные представлены в виде отдельных измерений, медианных значений, 25 и 75 % квартилей (рис. 1). Поведение показателей не согласуется ни с нормальным, ни с лог-нормальным законом.

Данные, характеризующие иммунный статус детей, отличаются большой размерностью пространства состояний (59) при малой мощности выборки (19 пациентов), что характерно для системно-биологических исследований (Basu et al., 2017). Когда размер выборки большой, для определения взаимосвязей может быть использован подход на основе частных корреляций. В противном случае для корректного (статистически значимого) определения корреляционных связей между измеряемыми переменными и построения графа сетевой топологии может быть использован подход, который учитывает малый размер множества данных. Следует отметить, что все дети относились к одной возрастной группе, от одного года до двух лет, которую в медицинской практике не принято фрагментировать далее. По причине малочисленности группы (19 человек) дополнительное разбиение по полу (10 девочек и 9 мальчиков) снизило бы мощность выборки ниже критического уровня необходимого для применяемого в работе алгоритма анализа.

Анализ главных компонент. Анализ главных компонент выполнен с помощью функции prcomp на языке R, для визуализации использован R-пакет factoextra (версия 1.0.7). Для анализа главных компонент данные были предварительно стандартизированы, а переменные TGF-β, IL-17 и $CD3^+CD45R0^+CD4^+CD161^+$ были исключены из анализа из-за отсутствия некоторых измерений. Проведенный анализ главных компонент не выявил возможности объяснения дисперсии данных малым числом главных компонент (рис. 2, *а*), а исходные переменные не образуют отдельных кластеров по корреляциям с первыми двумя главными компонентами (см. рис. 2, *б*).

Методы анализа частных корреляций и построения сетей связей. Альтернативу классическому методу оценивания частных корреляций представляет подход с применением методов регуляризации для оценивания матрицы частных корреляций (Epskamp, Fried, 2018). В основе принципа регуляризации лежит предположение о том, что число связей в конструируемой модели сети существенно меньше, чем число наблюдаемых переменных, т. е. реальная сеть является разреженной. Следовательно, в качестве регуляризирующей поправки, позволяющей обнулить несущественные корреляционные связи между переменными (числами ребер в графе), используется метод LASSO (Epskamp, Fried, 2018). Для анализа наших данных мы использовали реализацию этого подхода в алгорит-

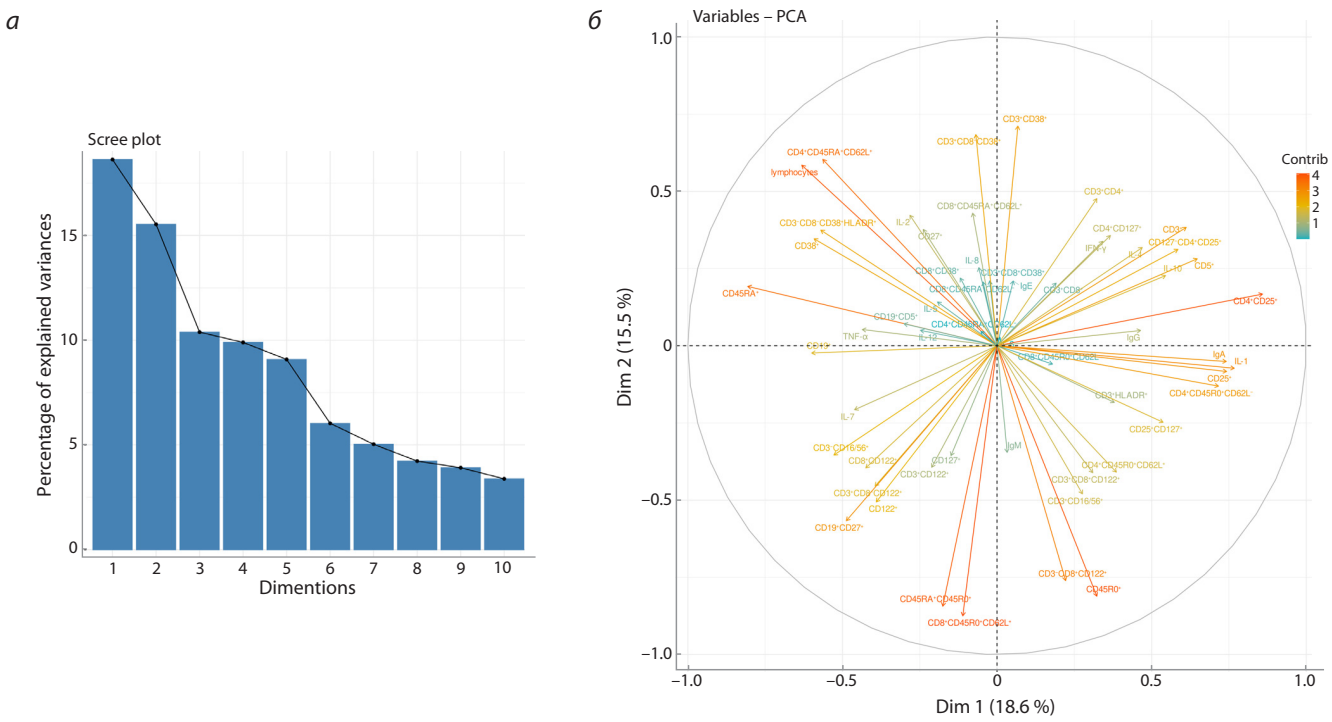


Рис. 2. Анализ главных компонент: *а* – доля объясненной дисперсии; *б* – график первых двух главных компонент.

ме расчета разреженной матрицы скорректированных частных корреляций DSPC (Basu et al., 2017), в котором проводится дополнительная коррекция оценок элементов обратной матрицы к выборочной матрице ковариаций, т. е. элементов матрицы частных корреляционных связей. Оценки элементов матрицы корреляций представлялись в виде тепловых карт и визуализировались в виде взвешенных сетей, где вершины (узлы) – переменные иммунного статуса, а ребра изображают корреляции между ними. Результаты оценивания корреляционных связей существенно зависят от параметров алгоритма: 1) величины коэффициента λ при слагаемом для регуляризации в виде ℓ_1 нормы обратной ковариационной матрицы; 2) выбора уровня статистической значимости предсказанной корреляционной связи p. Ниже мы изучим влияние значения p на сетевую топологию связей в иммунной системе.

Для расчета разреженных частных корреляций методом DSPC было использовано Java-приложение CorrelationCalculator (версия 1.0.1) авторов алгоритма (Basu et al., 2017). Исходные данные были предварительно нормированы: логарифмически преобразованы и стандартизированы. Графическое представление статистически значимых корреляций (при $p < 0.01$; 0.05; 0.1; 0.15) в виде тепловых карт и графов сетей корреляционных связей выполнено с использованием R-пакетов igraph (версия 1.6.0) и ggplot2 (версия 3.5.2). Расчет топологических характеристик графов сетей корреляционных связей выполнен с помощью R-пакета igraph (версия 1.6.0).

Результаты

Ниже мы изучим влияние p-значения на сетевую топологию связей в иммунной системе. Будут рассмотрены общепринятые уровни значимости 0.01, 0.05, 0.1, 0.15.

Тепловая карта и граф связей для $p = 0.01$

Тепловая карта корреляций между показателями иммунного статуса для здоровых детей при пороге статистической значимости $p = 0.01$ представлена на рис. 3, *а*. Соответствующий граф сети имеет 23 вершины и 12 ребер, или связей (см. рис. 3, *б*). Связность в сети, по сути, отсутствует. На рис. 3, *в* дано распределение показателей иммунного ответа по числу выявленных связей между ними. Вершина, имеющая максимальное число (всего две) корреляционных связей, представляет популяцию CD4 Т-клеток ($CD3^+CD4^+$).

Тепловая карта и граф связей для $p = 0.05$

Тепловая карта корреляций между показателями иммунного статуса для здоровых детей при пороге статистической значимости $p = 0.05$ представлена на рис. 4, *а*. Соответствующий граф сети имеет 53 вершины и 44 ребра, или связи (см. рис. 4, *б*). Связность отдельных компонент сети усиливается, но в целом отсутствует. На рис. 4, *в* дано распределение показателей иммунного ответа по числу выявленных связей между ними. Вершины, имеющие наибольшее число корреляционных связей (своеобразные хабы), – это цитокины провоспалительного типа IL-8, IL-12, а также Т-клетки памяти центральные ($CD4^+CD45RA^+CD62L^+$, $CD8^+CD45R0^+CD62L^+$), Th17 ($CD3^+CD45R0^+CD4^+CD161^+$) и активированные NK- клетки ($CD3^-CD8^+CD122^+$). Максимальное число связей увеличивается до трех.

Тепловая карта и граф связей для $p = 0.1$

Тепловая карта корреляций между показателями иммунного статуса для здоровых детей при пороге статистической значимости $p = 0.1$ представлена на рис. 5, *а*.

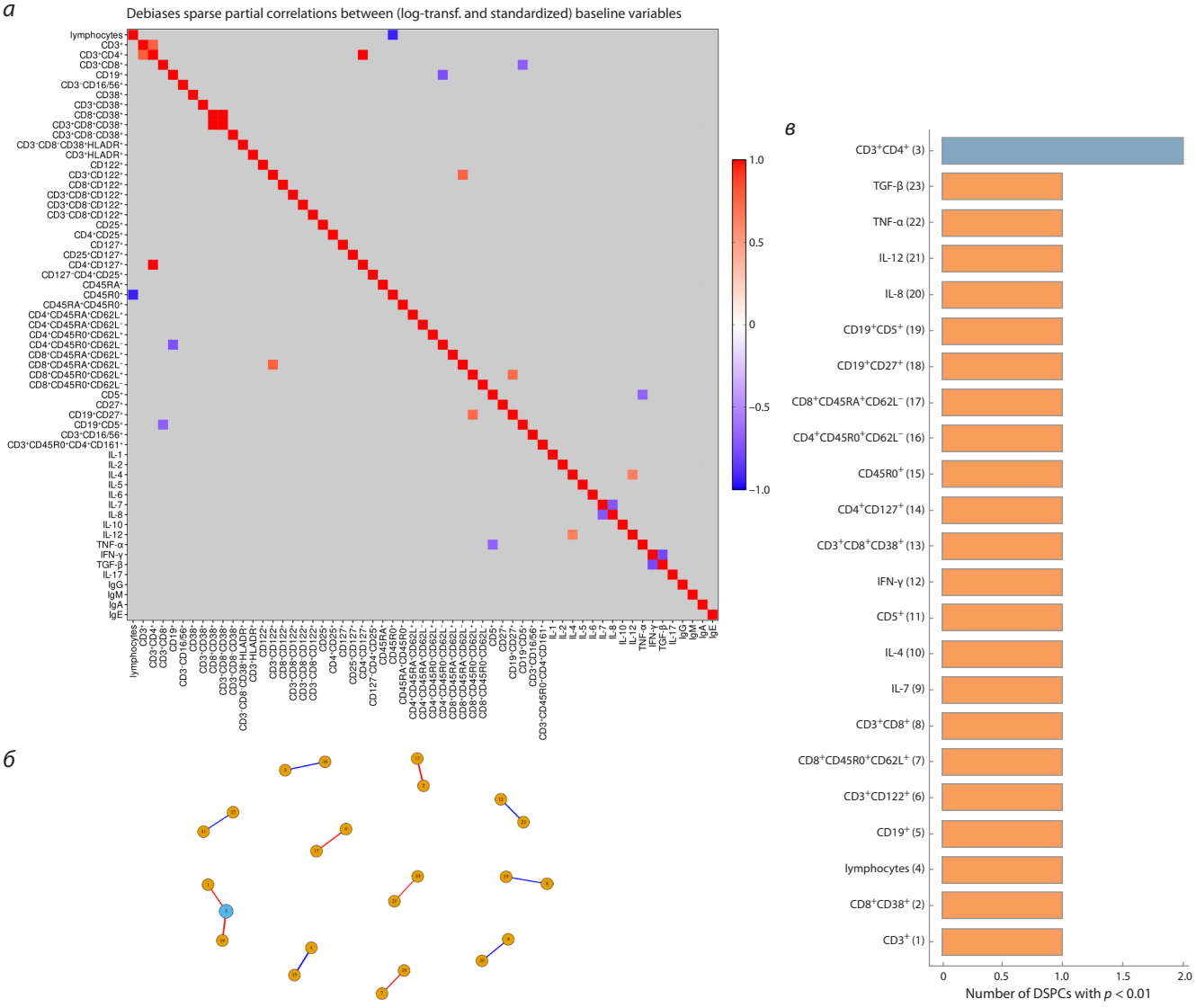


Рис. 3. Тепловая карта и граф сети связей иммунологических параметров у здоровых детей при пороге статистической значимости $p = 0.01$:

а – тепловая карта корреляций между показателями иммунного статуса; *б* – граф сети связей при $p = 0.01$; *в* – характеристики сложности сети связей. Здесь и на рис. 4–6: Номера вершин соответствуют параметрам иммунного статуса, показанным на *в*. По оси ординат – названия показателей иммунного статуса. По оси абсцисс показаны степени вершин графа. Положительные корреляции (красные линии), отрицательные корреляции (синие линии), толщина ребер пропорциональна абсолютным значениям коэффициентов корреляций DSPC. Цвет вершин соответствует индексу вершины, т. е. числу корреляционных связей.

Соответствующий граф сети имеет 59 вершин и 69 ребер, или связей (см. рис. 5, *б*). На рис. 5, *в* дано распределение показателей иммунного ответа по числу выявленных связей между ними. Вершины, имеющие максимальное число корреляционных связей (в данном случае четыре), – это цитокины IL-4, IL-12, активирующие клеточный и гуморальный иммунитет, а также терминально-дифференцированные эффекторные Т-клетки памяти ($CD4^+CD45RA^+CD62L^-$, $CD8^+CD45RA^+CD62L^-$) и Th17 ($CD3^+CD45R0^+CD4^+CD161^+$).

Тепловая карта и граф связей для $p = 0.15$

Тепловая карта корреляций между показателями иммунного статуса для здоровых детей при пороге статистической значимости $p = 0.15$ представлена на рис. 6, *а*. Соответствующий граф сети имеет 59 вершин и 106 ребер, или связей (см. рис. 6, *б*). На рис. 6, *в* дано распределение показателей иммунного ответа по числу выявленных связей между ними. Вершины, имеющие максимальное число корреляционных связей, т. е. хабы сети, – это иммуноглобулины IgM, NK-клетки, активированные цитотоксические Т-клетки, регуляторные Т-клетки, наивные и Т-клетки памяти, плазмоциты ($CD3^-CD8^-CD38^+HLADR^+$), активированные Т-клетки ($CD3^+CD8^-CD38^+$, $CD8^+CD122^+$) и дважды положительные активированные клетки ($CD45RA^+CD45R0^+$), отражающие переход из наивных в клетки памяти. Максимальное число связей равно шести.

Анализ устойчивости оценок корреляционных связей

Для оценки устойчивости полученных коэффициентов корреляции DSPC по отношению к размеру выборки была проведена процедура генерации 10 разных подвыборок по

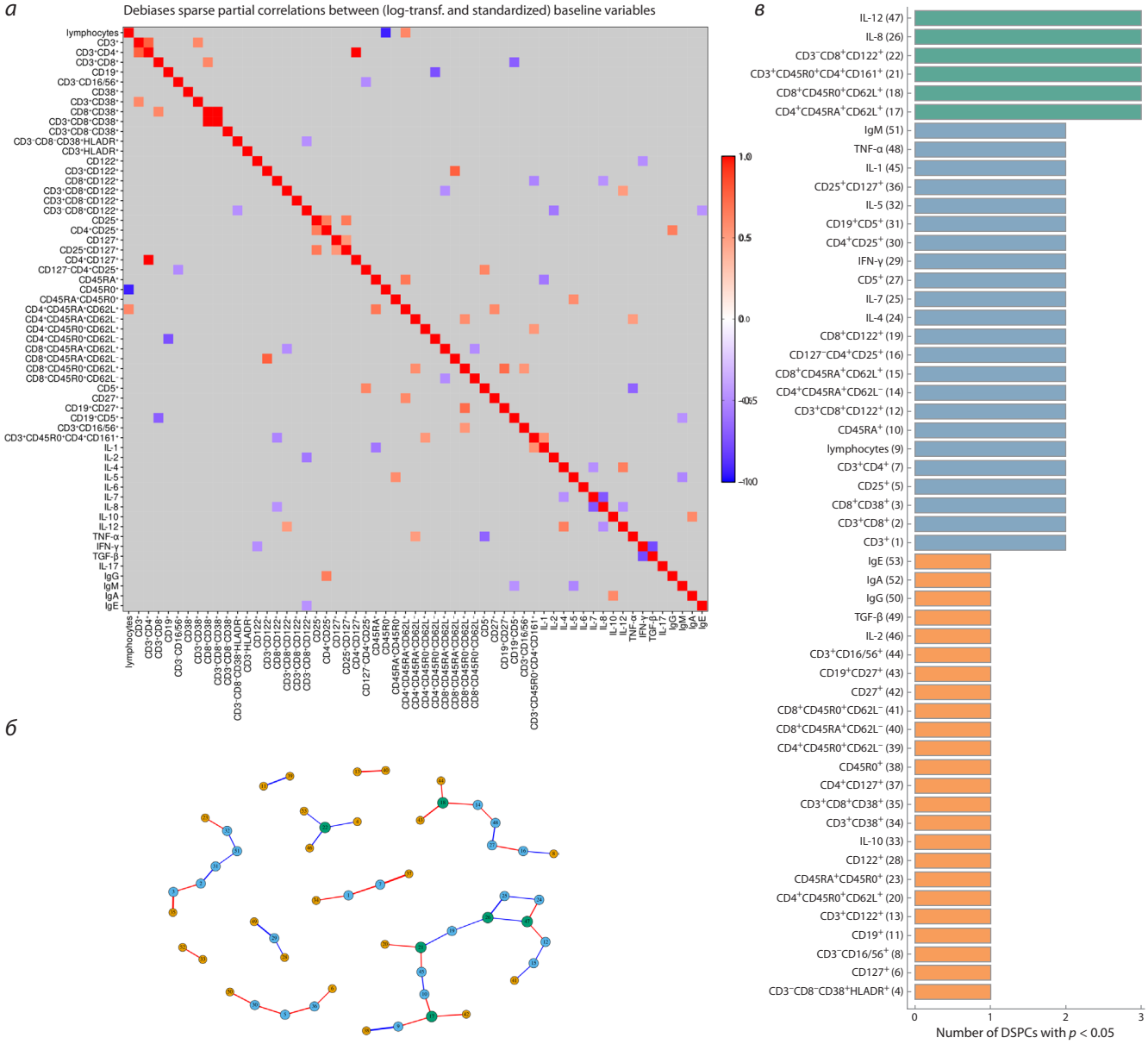


Рис. 4. Тепловая карта и граф сети связей иммунологических параметров у здоровых детей при пороге статистической значимости $p = 0.05$: *а* – тепловая карта корреляций между показателями иммунного статуса; *б* – граф сети связей при $p = 0.05$; *в* – характеристики сложности сети связей.

схеме vfold10, что в большинстве случаев соответствует выбору 17 из 19 измерений. В качестве меры устойчивости выбран коэффициент вариации (отношение стандартного отклонения к среднему значению) коэффициентов DSPC, оцененных на сгенерированных подвыборках. Коэффициенты вариации проиллюстрированы на рис. 7 для четырех порогов статистической значимости в виде тепловых карт. Их значения не превышают по модулю величины 0.1.

Сравнительный анализ топологических свойств графов корреляционных связей между показателями иммунного статуса

В таблице приведены результаты расчета топологических характеристик построенных графов сетей корреляционных связей между показателями иммунного статуса для различных порогов статистической значимости. Рассмотрены базовые характеристики: диаметр графа; радиус графа; обхват графа (girth) – длина наименьшего цикла, содержащегося в графе; средняя длина пути; энергия графа; спектральный радиус; плотность графа; коэффициент кластеризации; среднее по графу разнообразие вершин, определяемое через энтропию, рассчитываемую по весам инцидентных ребер – абсолютным значениям коэффициентов корреляций DSPC; количество разрезающих вершин и ребер и число несвязанных подграфов (см. таблицу).

Число вершин, ребер и максимальных степеней вершин растет с увеличением порога статистической значимости. При этом диаметр графа, радиус, обхват и средняя длина пути проявляют немонотонную зависимость, сначала возрастая, а затем снижаясь, что свидетельствует о трансформации свойств в сторону семейства «малый мир». Энергия графа и спектральный радиус монотонно возрастают с

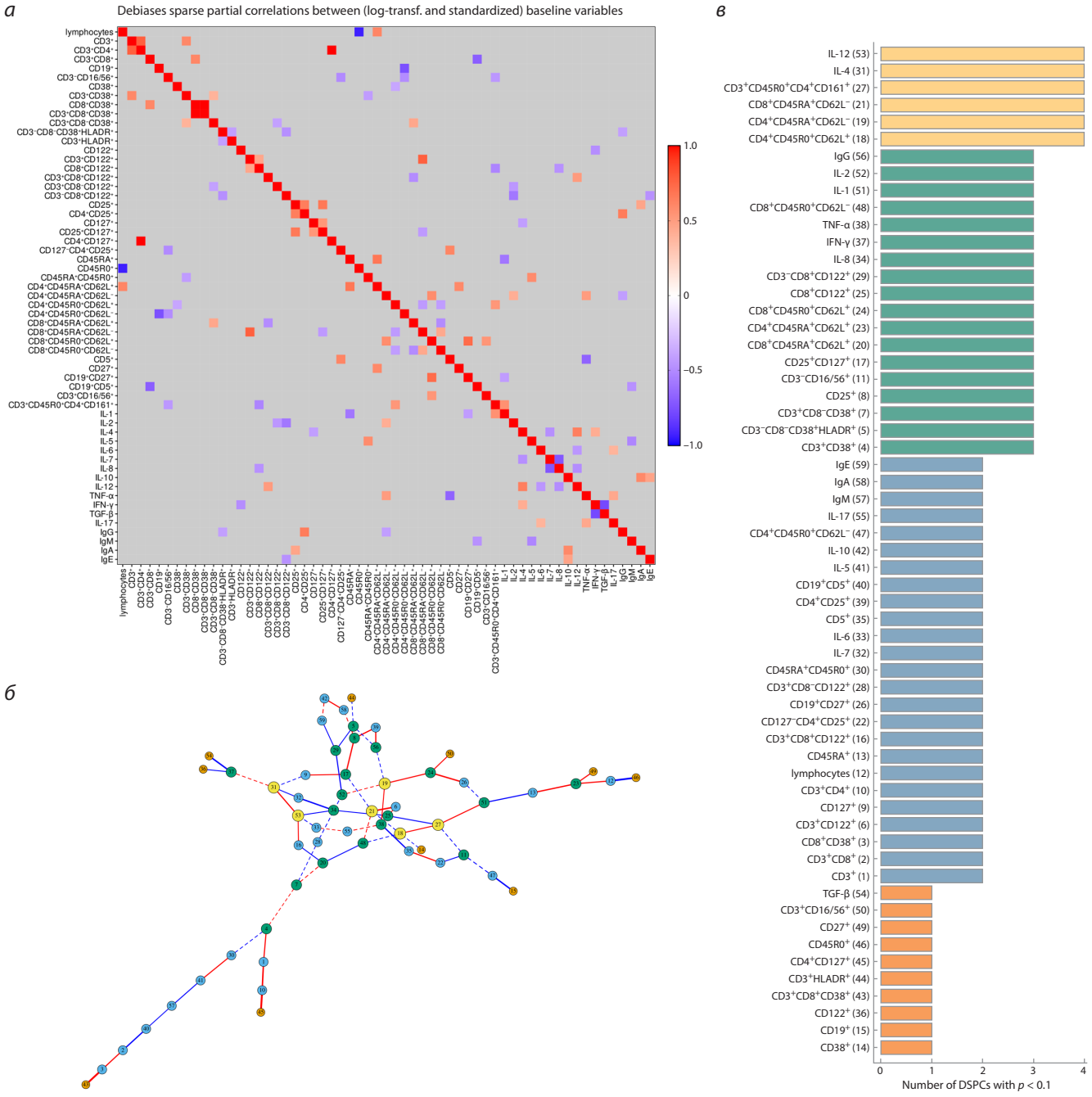


Рис. 5. Тепловая карта и граф сети связей иммунологических параметров у здоровых детей при пороге статистической значимости $p = 0.1$:

а – тепловая карта корреляций между показателями иммунного статуса; *б* – граф сети связей при $p = 0.1$; *в* – характеристики сложности сети связей. Сплошные линии ребер соответствуют корреляциям с уровнем значимости $p < 0.05$, штриховые – $p < 0.1$.

увеличением порога p. Коэффициент кластеризации также увеличивается, т. е. узлы графа имеют тенденцию группироваться. Интересно, что число разрезающих вершин и ребер уменьшается при $p = 0.15$, что может свидетельствовать об увеличении робастности графа связей. Ожидаемо снижается количество несвязанных подграфов.

Обсуждение

Идентификация структур связей между различными функциональными компонентами иммунной системы представляет собой чрезвычайно актуальную задачу современной иммунологии. Это связано с беспрецедентно высоким уровнем числа измеряемых характеристик, при относительно небольшом объеме выборки, отражающих ситуацию в области биоматематики больших данных, называемую «проклятием размерности». Для анализа взаимосвязей между показателями иммунного статуса нами реализован и исследован подход, в основе которого лежит регуляризированный метод оценивания разреженных частных корреляций, минимизирующий число ложных корреляций, реализованный в алгоритме DSPC (Basu et al., 2017). Результаты применения алгоритма могут за-

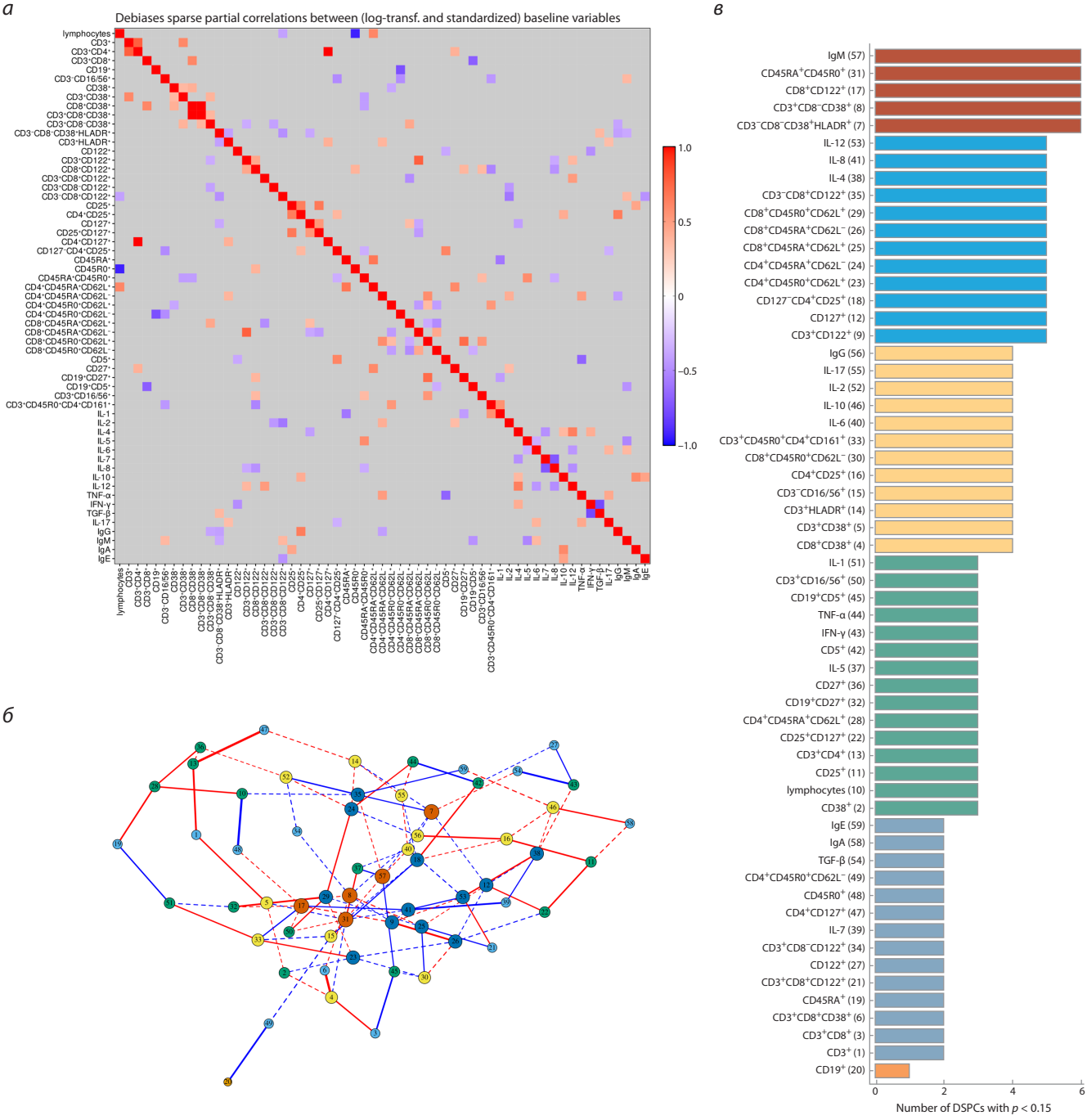


Рис. 6. Тепловая карта и граф сети связей иммунологических параметров у здоровых детей при пороге статистической значимости $p = 0.15$:

а – тепловая карта корреляций между показателями иммунного статуса; *б* – граф сети связей при $p = 0.15$; *в* – характеристики сложности сети связей. Сплошные линии ребер соответствуют корреляциям с уровнем значимости $p < 0.05$, штриховые – $p < 0.15$.

висеть от размера выборки, заполнения пропущенных данных, характера истинной структуры сети и других аспектов. В нашей работе продемонстрировано, что при ограниченном размере выборки измерений априорное задание уровня статистической значимости имеет принципиальное значение для формирования матрицы частных корреляций. Повышение порога статистической значимости увеличивает сложность сетевой топологии, формируемой в рамках подхода на основе машинного обучения. Окончательная верификация иммунологически корректной структуры связей требует как увеличения размера выборки, так и сопряжения с априорными механизменными представлениями и моделями функционирования компонент иммунной системы, т. е. участия клинических иммунологов (Qiao et al., 2025). Важным шагом в этом направлении стала разработка инструмента построения и анализа сети взаимодействий в иммунной системе ImmunoGlobe (Atallah et al., 2020) на основе феноменологической информации из фундаментального учебника «Janeway's Immunobiology" (Murphy, Weaver, 2017).

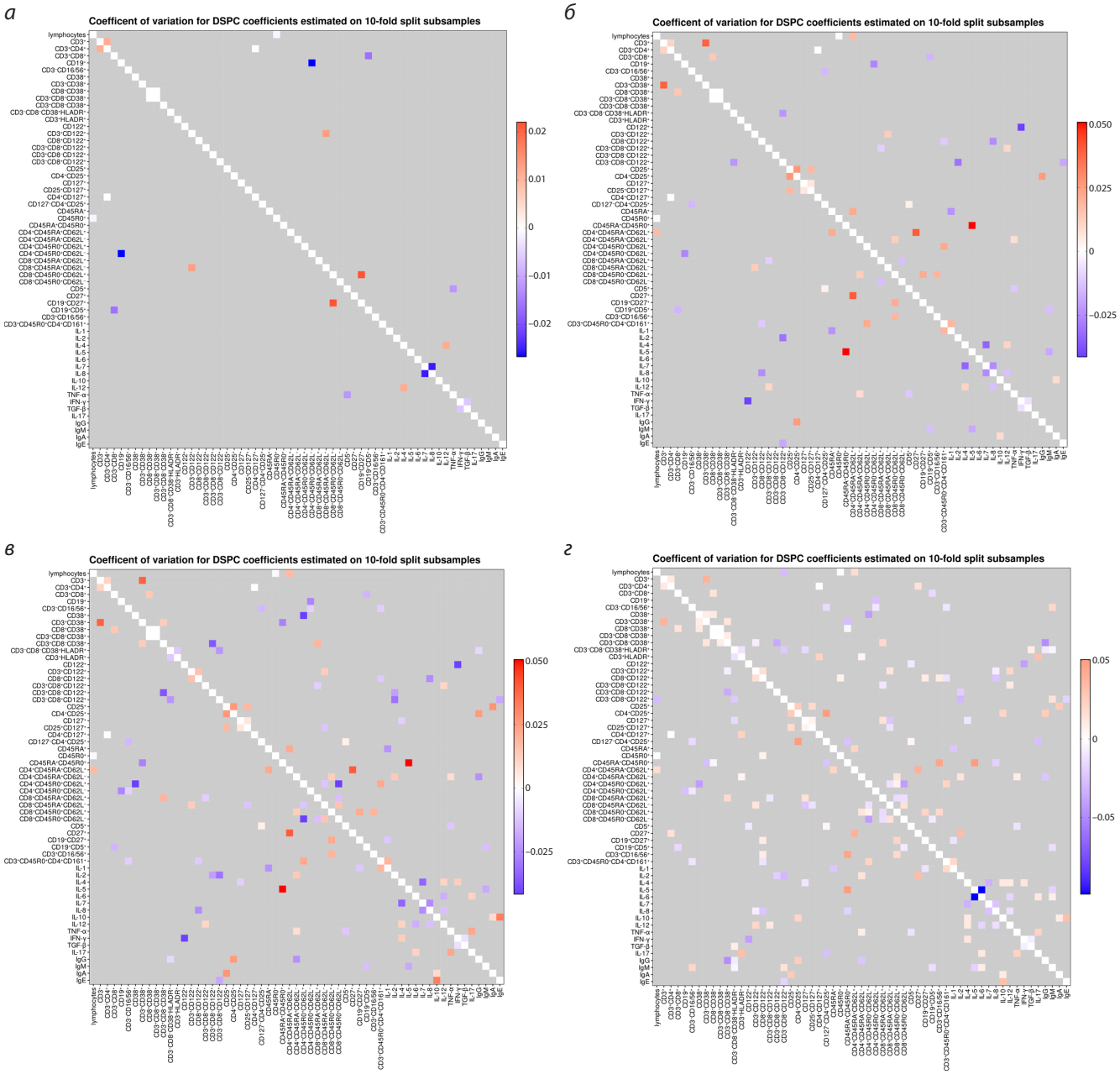


Рис. 7. Матрица оценок коэффициентов вариации для четырех уровней значимости: $p < 0.01$ (*а*); $p < 0.05$ (*б*); $p < 0.1$ (*в*); $p < 0.15$ (*г*).

Цель данной работы – реализация и внедрение нового метода построения взаимосвязей между клеточными и гуморальными компонентами иммунной систем. Задача обнаружения сетевых взаимосвязей между элементами иммунного статуса является центральной в рамках системно-иммунологического подхода, однако соответствующий аналитический инструментарий остается неразвитым. Все существующие к настоящему времени верифицированные представления об иммунных сетях ограничены схемами не более чем с тремя–четырьмя компонентами: антигенпрезентирование, линии дифференцировки, пара- и аутокринные взаимодействия. По этой причине однозначный выбор и верификация одной из представленных сетей не представляется возможным. Если придерживаться общепринятого порогового уровня значимости ($p = 0.05$), то следует отдать предпочтение сети, полученной в разделе «Тепловая карта и граф связей для $p = 0.05$». Выявление сетевой структуры взаимосвязей между компонентами клеточного и гуморального иммунитета – необходимый элемент для перехода от статического описания иммунного статуса к описанию системной динамики поддержания иммунного гомеостаза.

Заключение

Развитие комбинированной терапии хронических заболеваний для воздействия на несколько компонент иммунной системы предполагает определение топологии и силы структурных связей в системе. В нашей работе впервые показано, что для решения подобной задачи в типичной ситуации большого числа измеряемых параметров иммун-

Характеристики сетевой топологии связей в иммунной системе для различных уровней значимости оценивания частных корреляций

Топологические характеристики	$p \le 0.01$	$p \le 0.05$	$p \le 0.1$	$p \le 0.15$
Число вершин, n	23	53	59	59
Число ребер, m	12	44	69	106
Максимальная степень, Δ_G	2	3	4	6
Диаметр, D	2	11	17	7
Радиус, r	1	1	9	4
Охват, g	0	4	3	3
Средняя длина пути, l_G	1.08	3.6	6.0	3.3
Энергия, E_n	22.8	58.7	77.5	94.6
Спектральный радиус, ρ	1.4	2.3	3.0	4.3
Плотность графа, ρ_d	0.05	0.03	0.04	0.06
Коэффициент кластеризации, C	0	0	0.026	0.055
Среднее разнообразие вершин, D_{DSPC}	0.04	0.54	0.82	0.96
Число разрезающих вершин, n_{cut}	1	27	21	2
Число разрезающих ребер, m_{cut}	12	40	24	2
Число несвязанных подграфов	11	10	1	1

ного статуса и небольшой выборки пациентов можно применять методы машинного обучения, позволяющие получать скорректированные оценки разреженных частных корреляций. Трансляция результатов в биомедицинскую практику для решения задач персонализированного лечения и профилактики иммунозависимых патологических процессов предполагает активное участие клиницистов для определения мишеней терапии и количественного прогнозирования ее эффективности.

Список литературы / References

Armingol E., Baghdassarian H.M., Lewis N.E. The diversification of methods for studying cell-cell interactions and communication. *Nat Rev Genet.* 2024;25(6):381-400. doi 10.1038/s41576-023-00685-8

Atallah M.B., Tandon V., Hiam K.J., Boyce H., Hori M., Atallah W., Spitzer M.H., Engleman E., Mallick P. ImmunoGlobe: enabling systems immunology with a manually curated intercellular immune interaction network. *BMC Bioinformatics.* 2020;21(1):346. doi 10.1186/s12859-020-03702-3

Basu S., Duren W., Evans C.R., Burant C.F., Michailidis G., Karnovsky A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. *Bioinformatics.* 2017;33(10):1545-1553. doi 10.1093/bioinformatics/btx012

Cortese M., Hagan T., Rouphael N., Wu S.Y., Xie X., Kazmin D., Wimmers F., … Subramaniam S., Mulligan M.J., Khurana S., Golding H., Pulendran B. System vaccinology analysis of predictors and mechanisms of antibody response durability to multiple vaccines in humans. *Nat Immunol.* 2025;26(1):116-130. doi 10.1038/s41590-024-02036-z

Epskamp S., Fried E.I. A tutorial on regularized partial correlation networks. *Psychol Methods.* 2018;23(4):617-634. doi 10.1037/met0000167

Murphy K., Weaver C. Janeway's Immunobiology. New York, NY: Garland Science/Taylor & Francis Group, 2017. ISBN 978-0-8153-4505-3 Available: https://inmunologos.wordpress.com/wp-content/uploads/2020/08/janeways-immunobiology-9th-ed_booksmedicos.org_.pdf

Ng C.T., Snell L.M., Brooks D.G., Oldstone M.B. Networking at the level of host immunity: immune cell interactions during persistent viral infections. *Cell Host Microbe*. 2013;13(6):652-664. doi 10.1016/j.chom.2013.05.014

Qiao L., Khalilimeybodi A., Linden-Santangeli N.J., Rangamani P. The evolution of systems biology and systems medicine: From mechanistic models to uncertainty quantification. *Annu Rev Biomed Eng*. 2025;27(1):425-447. doi 10.1146/annurev-bioeng-102723-065309

Rieckmann J.C., Geiger R., Hornburg D., Wolf T., Kveler K., Jarrossay D., Sallusto F., Shen-Orr S.S., Lanzavecchia A., Mann M., Meissner F. Social network architecture of human immune cells unveiled by quantitative proteomics. *Nat Immunol.* 2017;18(5):583-593. doi 10.1038/ni.3693

Shilts J., Severin Y., Galaway F., Müller-Sienerth N., Chong Z.S., Pritchard S., Teichmann S., Vento-Tormo R., Snijder B., Wright G.J. A physical wiring diagram for the human immune system. *Nature.* 2022;608(7922):397-404. doi 10.1038/s41586-022-05028-x. Erratum in: *Nature.* 2024;635(8037):E1. doi 10.1038/s41586-024-07928-6

Toptygina A., Grebennikov D., Bocharov G. Prediction of specific antibody- and cell-mediated responses using baseline immune status parameters of individuals received measles-mumps-rubella vaccine. *Viruses.* 2023;15(2):524. doi 10.3390/v15020524

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 30.07.2025. После доработки 15.09.2025. Принята к публикации 19.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1051-1061

doi 10.18699/vjgb-25-110

Самообучающиеся виртуальные организмы в физическом симуляторе: об оптимальном разрешении их зрительной системы, архитектуре нервной системы и вычислительной сложности задачи

М.С. Зенин[1], А.П. Девятериков[2], А.Ю. Пальянов [1, 2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
palyanov@iis.nsk.su

Аннотация. Зрение играет ключевую роль в жизни множества различных организмов, обеспечивая ориентацию в пространстве, поиск пищи, избегание хищников и социальное взаимодействие. У видов с относительно простой зрительной системой, таких как насекомые, эффективная поведенческая стратегия достигается за счет высокой специализации нейронов, адаптации к конкретным условиям среды, а также благодаря дополнительным сенсорным системам – обонянию или слуху. У животных с более сложным зрением и нервной системой, таких как млекопитающие, когнитивные возможности и способности к адаптации значительно выше, однако выше и энергозатраты на работу мозга. Моделирование особенностей таких систем в виртуальной среде позволило бы исследовать фундаментальные принципы функционирования и обучения когнитивных систем, включая механизмы восприятия, памяти, принятия решений и их взаимодействие. В данной работе объектом исследования являются виртуальные организмы, обладающие зрительной системой и функционирующие в трехмерной физической среде на базе Unity ML-Agents – одного из наиболее высокопроизводительных современных симуляторов. Предложенная иерархическая архитектура управления, разделяющая когнитивные задачи между двумя модулями – зрительного восприятия/принятия решений и управления координированным движением конечностей для перемещения в физической среде – показала существенно большую скорость и эффективность обучения по сравнению с единой системой. Проведена серия численных экспериментов, направленных на выявление влияния параметров зрительной системы, конфигурации среды и архитектурных особенностей агентов на успешность их обучения с подкреплением (алгоритм PPO). Показано, что существует диапазон разрешений, обеспечивающий компромисс между вычислительной сложностью и успешностью выполнения задачи, а избыточная размерность сенсорных входных данных или пространства действий приводит к замедлению обучения. Должное внимание уделено также оценке вычислительной сложности системы и профилированию производительности ее основных компонентов. Полученные результаты представляют потенциальный интерес в контексте исследований искусственных агентов, вдохновленных биологическими системами, эволюционного моделирования, включая нейроэволюционные подходы для создания более адаптивных и умных агентов, и изучения когнитивных процессов в них.

Ключевые слова: виртуальный организм; компьютерное моделирование; вычислительная сложность; зрительная система; нейронная сеть; симулятор; PPO; обучение с подкреплением; Unity ML-Agents

Для цитирования: Зенин М.С., Девятериков А.П., Пальянов А.Ю. Самообучающиеся виртуальные организмы в физическом симуляторе: об оптимальном разрешении их зрительной системы, архитектуре нервной системы и вычислительной сложности задачи. *Вавиловский журнал генетики и селекции*. 2025;29(7):1051-1061. doi 10.18699/vjgb-25-110

Self-learning virtual organisms in a physics simulator: on the optimal resolution of their visual system, the architecture of the nervous system and the computational complexity of the problem

M.S. Zenin[1], A.P. Devyaterikov[2], A.Yu. Palyanov [1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
palyanov@iis.nsk.su

© Зенин М.С., Девятериков А.П., Пальянов А.Ю., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. Vision plays a key role in the lives of various organisms, enabling spatial orientation, foraging, predator avoidance and social interaction. In species with relatively simple visual systems, such as insects, effective behavioral strategies are achieved through high neural specialization, adaptation to specific environmental conditions, and the use of additional sensory systems such as olfaction or hearing. Animals with more complex vision and nervous systems, such as mammals, have greater cognitive abilities and flexibility, but this comes with increased demands on the brain's energy costs and computational resources. Modeling the features of such systems in a virtual environment could allow researchers to explore the fundamental principles of sensorimotor integration and the limits of cognitive complexity, as well as test hypotheses about the interaction between perception, memory and decision-making mechanisms. In this work, we implement and investigate a model of virtual organisms with a visual system operating in a three-dimensional physical environment using the Unity ML-Agents software – one of the most high-performance simulation platforms currently available. We propose a hierarchical control architecture that separates locomotion and navigation tasks between two modules: (1) visual perception and decision-making, and (2) coordinated control of limb movement for locomotion in the physical environment. A series of numerical experiments was conducted to examine the influence of visual system parameters (e.g, resolution of the "first-person" view), environmental configuration and agent architectural features on the efficiency and outcomes of reinforcement learning (using the PPO algorithm). The results demonstrate the existence of an optimal range of resolutions that provide a trade-off between computational complexity and success in accomplishing the task, while excessive dimensionality of sensory inputs or action space leads to slower learning. We performed system performance profiling and identified key bottlenecks in large-scale simulations. The discussion considers biological parallels, highlighting cases of high behavioral efficiency in insects with relatively low-resolution visual systems, and the potential of neuroevolutionary approaches for adapting agent architectures. The proposed approach and the results obtained are of potential interest to researchers working on biologically inspired artificial agents, evolutionary modeling, and the study of cognitive processes in artificial systems.

Key words: virtual organism; computational modeling; computational complexity; vision system; neural network; simulator; PPO; reinforcement learning; Unity ML-Agents

For citation: Zenin M.S., Devyaterikov A.P., Palyanov A.Yu. Self-learning virtual organisms in a physics simulator: on the optimal resolution of their visual system, the architecture of the nervous system and the computational complexity of the problem. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1051-1061. doi 10.18699/vjgb-25-110

Введение

Моделирование когнитивной деятельности, поведения и эволюционных механизмов в виртуальных средах представляет собой важный этап на пути к созданию систем искусственного интеллекта, способных к обучению, адаптации и взаимодействию со сложными окружающими средами (Bongard, 2013; Stanley et al., 2019). Развитие таких систем возможно благодаря современным платформам агентного обучения, в частности Unity ML-Agents (Juliani et al., 2018), которые позволяют создавать полноценные трехмерные симуляции с физикой, зрением и множеством обучаемых агентов.

Несмотря на относительно малое количество нейронов ввиду небольших размеров тела по сравнению, скажем, с млекопитающими, нервные системы многих беспозвоночных, включая насекомых, демонстрируют сложное, разнообразное и адаптивное поведение. Например, у муравьев количество нейронов составляет примерно 250000, что на несколько порядков меньше по сравнению даже с млекопитающими (у мыши – около $7.1 \cdot 10^7$), однако эти насекомые способны решать сложные задачи навигации, социальных взаимодействий, координации совместных действий и запоминания маршрутов (Chittka, Niven, 2009). Более того, отдельные виды муравьев способны проходить зеркальный тест – поведенческий индикатор самосознания (Cammaerts M.-C.T., Cammaerts R., 2015). Это делает их уникальными среди насекомых и подчеркивает потенциал минимальных, но эффективно организованных нервных систем, которые представляют огромный интерес для современной науки.

Системы зрения у насекомых также служат источником вдохновения при проектировании искусственных агентов. В частности, фасеточные глаза обеспечивают широкоугольное поле обзора и высокую частотность обновления, что позволяет эффективно реагировать на быстро меняющиеся стимулы (Land, Nilsson, 2012). При этом угловое разрешение значительно уступает человеческому, но это компенсируется высокой чувствительностью к движениям и обучаемостью на уровне всей поведенческой цепочки.

Такие особенности позволяют формулировать принципиально важные исследовательские вопросы: каковы минимальные требования к зрительной системе агента, позволяющие ему успешно адаптироваться к среде? Какая архитектура управления обеспечивает когнитивную модульность при ограниченных вычислительных ресурсах? Иными словами, как сконструировать искусственный организм – агента, обладающего простыми, но функциональными элементами восприятия и принятия решений? Поискам ответов на эти вопросы и посвящено настоящее исследование, в котором виртуальные организмы, обладающие зрением и действующие в 3D-среде, изучаются с позиций предельной когнитивной эффективности, масштабируемости и обучаемости в задачах поиска и навигации.

Интерес к структурам, обеспечивающим движение при минимальной сложности конструкции, проявляется и в инженерных системах. Например, в недавней работе (Song et al., 2022) рассмотрено управление гибридными мягкими конечностями, что отражает общее стремление

к построению структурно простых, но функционально эффективных решений для управления движением. Модель тела виртуального организма, использованная в настоящем исследовании, по числу степеней свободы и составу сегментов близка к тем, что применяются в подобных конструкциях. Это позволяет рассматривать ее как сопоставимую по уровню сложности с физическими аналогами.

В ранее опубликованной нами работе (Devyaterikov, Palyanov, 2022) представлен симулятор эволюции виртуальных организмов в 3D-среде, в которой каждому агенту были даны зрительная система и нейронная сеть для обработки потока сенсорной информации. Система основана на комбинации нейроэволюции и взаимодействия агентов с физической средой, в которой они могли выполнять простейшие когнитивные задачи, требующие использования зрения (поиск «пищи», необходимой для «выживания»), и позволяла оценивать приспособленность агентов в популяции. В новой работе представлены оценки вычислительной сложности расчетов, связанных с физикой (тел агентов, окружающего мира и их взаимодействия), 3D-рендерингом от «первого лица» для каждого из агентов и функционированием их нейронных сетей, а также введена новая иерархическая модель агента и приведены результаты количественного анализа времени, скорости и эффективности обучения в зависимости от разрешения зрительной системы. В работе (Aksoy, Camlitepe, 2018) приведены данные по числу омматидиев (светочувствительных сенсоров) для различных видов муравьев (от 100 до 3000). Грубо аппроксимируя такое зрение квадратной матрицей пикселей, мы можем говорить о разрешении зрения от 10×10 до 55×55.

Текущая работа сочетает методы обучения с подкреплением (PPO (Schulman et al., 2017)), сверточные нейронные сети (O'shea, Nash, 2015), подходы к иерархическому обучению агентов (Vezhnevets et al., 2017), а также практический анализ ресурсосберегающих схем моделирования (Peng et al., 2018). Мы показываем, что подход с иерархическими агентами, например пара Walker/Searcher, позволяет достичь более стабильного и интерпретируемого поведения, снижая время обучения при сходной сложности задачи.

Особое внимание уделено исследованию влияния разрешения зрительной системы на обучаемость агентов, с оценкой минимального размера входного изображения, при котором сохраняется способность к решению задачи визуального поиска и навигации. Подобные исследования актуальны как для биологически вдохновенного моделирования, так и для разработки компактных и эффективных архитектур ИИ-агентов, способных функционировать в условиях ограниченных вычислительных ресурсов (Hassabis, Humaran, 2017; Zador, 2019).

Кроме того, в данной работе рассматривается влияние стратегии разбиения задач на подзадачи (навигация и локомоция) на эффективность обучения. Это позволяет лучше понять принципы, лежащие в основе когнитивной модульности и распределенного управления в сложных агентных системах (Botvinick et al., 2020; Tschantz et al., 2020). Введенный агент Searcher, опирающийся исключительно на визуальное восприятие, взаимодействует с агентом Walker, отвечающим за физическое движение. Такая схема обеспечивает более высокую адаптивность модели и улучшает интерпретируемость поведения агентов.

Таким образом, целью настоящей работы стало системное исследование границ когнитивной сложности агентов с визуальной системой, разработка оптимальных архитектур управления и параметров восприятия, а также оценка производительности и масштабируемости предложенной системы при реализации на платформе Unity ML-Agents.

Материалы и методы

Постановка задачи. Рассматриваемая задача формулируется в терминах марковского процесса принятия решений, где агент взаимодействует с трехмерной физической средой и обучается максимизировать накопленное вознаграждение. Ниже приведено описание задачи, выполняемой агентом:

Среда E: квадратная арена, ограниченная стенами. На арене случайным образом появляются цели радиусом r, которые необходимо собрать. После того как цель достигнута, генерируется новая.

Состояние агента s_t: включает RGB-изображение от камеры от «первого лица» размером $h \times w \times 3$, а также вектор управляющих параметров – углы сочленений конечностей и соответствующие крутящие моменты.

Действие агента a_t: одно скалярное значение, отвечающее за нормализованный угол поворота в интервале $[-1, 1]$. Этот параметр определяет направление движения тела агента. Реальный угол поворота определяется как $\theta = a_t * \theta_{max}$, где θ_{max} – максимально допустимый угол поворота, который задается в параметрах эксперимента. В разных сериях экспериментов использовались различные значения этого параметра, что позволило нам исследовать его влияние на эффективность стратегии; результаты описаны в разделе «Результаты при изменении угла поворота». Ограничение управления одной переменной связано с тем, что задачи низкоуровневой локомоции (координация конечностей и поддержание равновесия) вынесены в отдельный модуль Walker, что позволяет сосредоточиться на когнитивной составляющей задачи, т. е. восприятии и принятии решений.

Функция награды $R(s_t, a_t)$: агент получает положительное вознаграждение за успешное достижение цели.

Цель: максимизировать суммарное вознаграждение за эпизод длиной T, то есть выработать стратегию, позволяющую эффективно ориентироваться в среде и собирать цели на основе зрительной информации.

Одна из задач нашего исследования – поиск наименьшего разрешения входного изображения, при котором агент еще способен успешно обучиться за разумное время. Формальная постановка выглядит так:

Успешность обучения определяется как достижение среднего вознаграждения не менее чем $R_{goal} = 5$ за эпизод (где вознаграждение выдается за сбор агентом целей). Значение R_{goal} выбрано экспериментально. Из графиков результатов обучения (см. раздел «Зависимость обучаемости от разрешения изображения») видно, что не обученный агент в результате случайного блуждания в среднем достигает R_{goal} не более 2.

Время обучения агента до достижения порогового значения: $T(N) \in \mathbb{R}_+$.

Среднее вознаграждение $R(Res, T)$, достигнутое агентом после обучения с входом $Res = h \times w \times 3$ за время T.

Допустимое множество разрешений $Res \in \mathbb{N}$, от $20 \times 20 \times 3$ до $100 \times 100 \times 3$ с шагом 20, а также дополнительный случай $84 \times 84 \times 3$ как разрешение по умолчанию для Unity ML-Agents.

Требуется найти $\min_{r \in Res} T(N)$, где $R(Res, T) \geq R_{goal}$.

Архитектура симулятора. Предлагаемая система использует иерархическую архитектуру управления агентом, разделяя функции восприятия и движения между двумя уровнями. Агент нижнего уровня Walker отвечает за физическое перемещение тела в среде, используя локальные сенсоры и предварительно обученную модель локомоции. Агент верхнего уровня Searcher получает на вход визуальную информацию с камеры и принимает решение о направлении движения, передавая управляющий сигнал агенту Walker в виде нормализованного угла поворота. Такой подход позволяет изолировать сложную задачу сенсомоторного преобразования, из изображения в действие, от задачи стабилизации движения и координации движений конечностей. Обучение Searcher при этом оказывается быстрее и стабильнее, поскольку он управляет только одной переменной. Внутренняя коммуникация между агентами осуществляется внутри среды Unity посредством передачи параметра направления в контроллер Walker. В режиме обучения агент Searcher обрабатывает визуальные данные и генерирует угол поворота, который применяется в качестве управляющего параметра для выбора ориентации тела на следующем шаге. Walker, в свою очередь, реализует заданное направление, обеспечивая движение в нужную сторону.

В ходе симуляции среда динамически обновляется: после сбора цели (единицы «пищи», необходимой для выживания) агентом Searcher вместо нее генерируется новая в случайной позиции (для поддержания числа доступных единиц «пищи» на постоянном уровне), а при падении агента или достижении предельного числа шагов эпизод перезапускается. Архитектура поддерживает параллельный запуск нескольких сред, каждая из которых содержит один Searcher и один Walker, что позволяет масштабировать обучение на уровне Unity ML-Agents.

Среда моделирования. Для проведения экспериментов была выбрана современная платформа Unity ML-Agents, демонстрирующая высокую производительность и предоставляющая удобные средства для создания сложных трехмерных симуляций с возможностью интеграции обучения с подкреплением. Unity также обеспечивает встроенную поддержку параллельных сред, визуальных сенсоров и интеграцию с библиотекой PyTorch.

Каждая среда представляет собой ограниченную квадратную арену (DynamicPlatform) со стенами, полом и случайно размещаемыми целями, которые агенту необходимо собирать. Размер платформы фиксирован, а координаты появления целей выбираются равномерно по доступной площади. При столкновении агента с целью она исчезает и немедленно заменяется новой. Стены непроницаемы и выступают физическим ограничением среды.

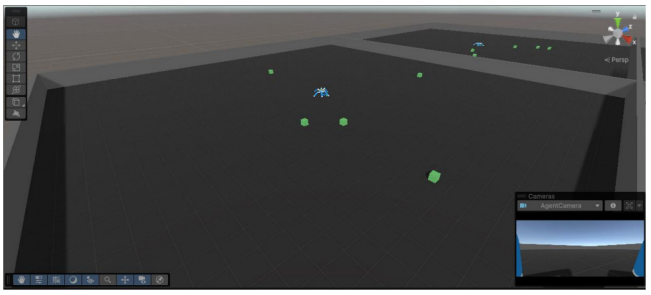


Рис. 1. В правом нижнем углу: вид симуляции со стороны наблюдателя – агента Walker. Показаны также две среды, агенты и цели.

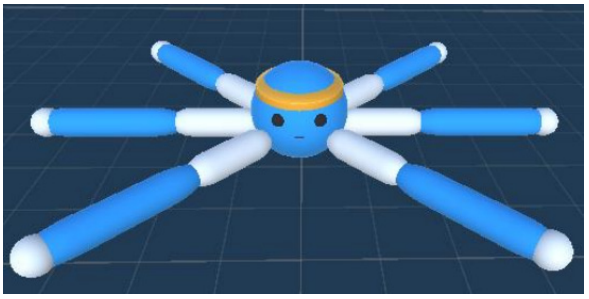


Рис. 2. Модель агента Walker в Unity.

Параметры симуляции задаются через компонент CrawlerSettings и включают «тактовую частоту» физического мира, гравитацию, продолжительность эпизода (max_step – число тактов симуляции, на котором агент получает наблюдения и выполняет действие) и количество параллельных сред. В случае падения агента, определяемого касанием телом пола, происходит автоматический перезапуск среды. В каждой параллельной среде размещается один агент Searcher, содержащий вложенный в него Walker, с индивидуальной камерой, расположенной на передней части головы и обеспечивающей нейросеть агента потоком зрительной информации от «первого лица».

Количество одновременно запущенных сред (num_envs) зависело от типа агента: для агента Walker, не применявшего визуальные данные, использовалось 10 сред, а для агента Searcher – 4. Это позволяло эффективно использовать ресурсы GPU и ускорять сбор данных за счет параллельного взаимодействия со средой. Для каждой среды данные о действиях, наблюдениях и наградах собирались независимо и синхронизировались с обучающей стратегией в Python через gRPC-интерфейс Unity ML-Agents. На рис. 1 представлен вид симуляции от лица наблюдателя; можно видеть две среды, агентов и некоторое число целей.

Модель агента Walker. Агент нижнего уровня Walker – сложная сочлененная модель с шестью конечностями, реализованная в движке Unity на основе компонентов Rigidbody и ConfigurableJoint. Каждая конечность состоит из двух сегментов – верхнего и нижнего, с числом степеней свободы, равным 3, и общим числом степеней свободы для всех ног, равным 18 (рис. 2). Это позволяет агенту осуществлять реалистичную локомоцию и сохранять устойчивость при передвижении.

Система управления реализована через модуль JointDriveController, преобразующий управляющие сигналы в желаемые углы и силы сочленений. Управляющие параметры представлены в виде вектора размерности 30: 18 значений управляют углами сочленений, 12 – крутящими моментами, прикладываемыми к ним. Конкретно для каждой из шести ног верхняя часть управляется по двум углам (вращение по осям X и Y), а нижняя часть – по одному углу (вращение по оси X), что дает 18 управляющих параметров. Дополнительно для каждой из этих 12 частей задается сила воздействия, отвечающая за интенсивность движения, что дает еще 12 параметров. На каждом шаге агент получает наблюдения, включающие информацию о текущих углах, скоростях, контактах с поверхностью, векторе целевого направления и ориентации тела, а также данные трассировки до земли. Модель агента Walker в движении показана на рис. 3.

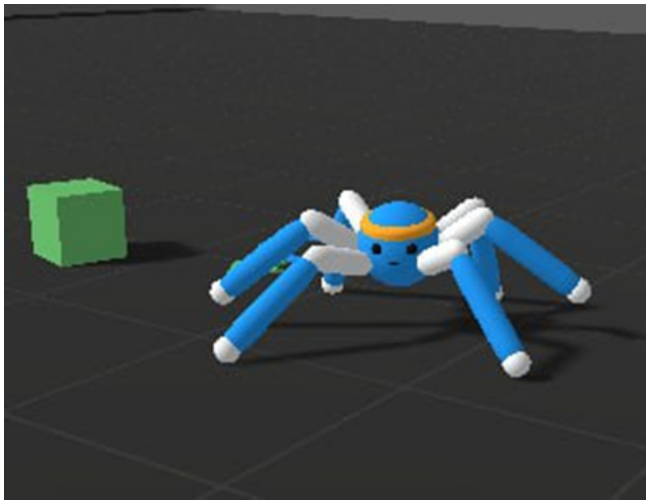


Рис. 3. Модель агента Walker в движении.

Архитектура нейронной сети Walker состоит из трех полносвязных слоев с функциями активации LeakyReLU и выходами actor (30 параметров действий) и critic (одно значение ценности состояния) (рис. 4, *а*). Размерность входного слоя составляет 223 (векторные признаки и параметры сочленений), скрытые слои имеют по 512 нейронов. Общий размер модели составляет 655 903 параметров и 1567 нейронов, что делает ее легкой для обучения в реальном времени.

Функция награды для агента Walker формируется на основе отклонения текущей скорости тела от целевой и совпадения направления движения с заданным вектором. Это позволяет агенту обучаться целенаправленному передвижению в необходимую сторону с учетом физической устойчивости. После завершения обучения агент Walker используется в режиме предсказания в составе агента Searcher, обеспечивая стабильную реализацию движения.

В процессе обучения блок critic получает те же входные данные, что и actor – вектор признаков состояния. На этих данных он учится приближать ожидаемое суммарное вознаграждение, которое агент получит в будущем, если будет продолжать действовать по текущей стратегии. На ранних этапах обучения эта оценка неточная, но постепенно корректируется с помощью обратного распространения ошибки, основанного на фактических наградах, которые получает агент. Таким образом, critic не знает изначально, что хорошо, а что плохо, – он обучается это распознавать, сравнивая прогнозируемые награды с реальными, накопленными в ходе симуляций.

После завершения обучения агент Walker используется в режиме предсказания в составе агента Searcher, обеспечивая стабильную реализацию движения, формирующуюся на основе отклонения текущей скорости тела от целевой и совпадения направления движения с заданным вектором. Это позволяет агенту обучаться целенаправленному передвижению в необходимую сторону при обеспечении физической устойчивости.

Модель агента Searcher. Агент верхнего уровня Searcher отвечает за восприятие окружающей среды и выбор направления движения тела. В отличие от агента Walker, он не взаимодействует напрямую с физическими

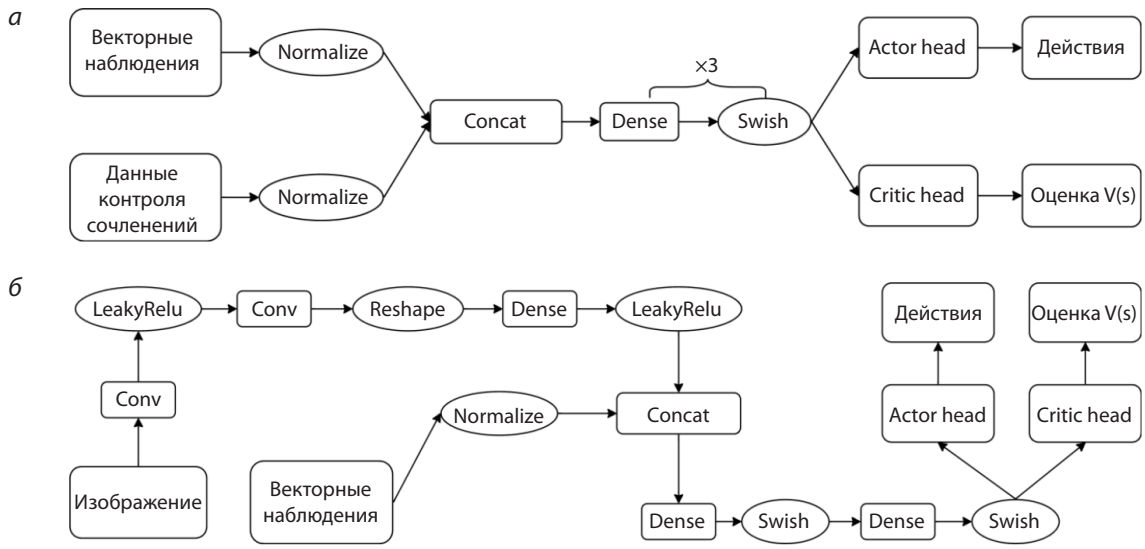


Рис. 4. Архитектура нейронной сети агентов: Walker (*а*) и Searcher (*б*).

компонентами симуляции, а управляет агентом Walker, передавая ему нормализованный угол поворота в интервале [−1, 1]. Таким образом, Searcher выступает в роли когнитивного модуля, реализующего стратегию поиска целей на основе визуальной информации. Основным источником входных данных для агента Searcher служит изображение, полученное с камеры, закрепленной на теле, передней части головы, агента. Камера направлена вперед и расположена на высоте, соответствующей голове виртуального организма. Разрешение изображения варьируется в экспериментах от 20 × 20 до 100 × 100 пикселей, с шагом 20 для каждого измерения (трехканальное RGB), что позволяет анализировать влияние зрительной нагрузки и разрешения кадра видеопотока на обучаемость модели.

Для обработки изображений используется сверточная нейросеть, состоящая из двух сверточных слоев (Conv2D), слоя выравнивания (Flatten) и последующих полносвязных слоев. Результат обработки визуального входа конкатенируется с векторными наблюдениями и подается на два выходных слоя: actor (одно значение – угол поворота) и critic (оценка состояния). Функции активации – LeakyReLU и Swish. Схематичное представление архитектуры нейронной сети агента Searcher показано на рис. 4, *б*.

Обучение агента Searcher происходит по алгоритму Proximal Policy Optimization (PPO) с использованием непрерывного пространства действий. Целевая функция – максимизация накопленного вознаграждения за сбор целей на арене. При столкновении с целью агент получает положительную награду, при столкновении со стеной или отсутствии активности – штраф. При превышении max_step или падении тела симуляция эпизода завершается и начинается новый.

В отличие от агента Walker, который предварительно обучается один раз и затем используется только для выполнения уже выученного поведения (режим inference – применение модели без дообучения), агент Searcher обучается с нуля, а его нейросеть содержит обработку изображений, что увеличивает вычислительные затраты, но позволяет реализовать биологически правдоподобное поведение, основанное исключительно на визуальном восприятии. Это дает возможность моделировать когнитивные ограничения и анализировать влияние разрешающей способности зрения на скорость и устойчивость обучения.

Алгоритмы и гиперпараметры обучения. Алгоритм PPO является градиентным методом оптимизации стратегии, относящимся к семейству Актор-Критик методов. Такие методы сочетают в себе обучение стратегии (policy) и функции ценности (value function). Избегая резких обновлений стратегии, в отличие от классических методов этого типа, PPO ориентирован на повышение стабильности и надежности обучения. Актор – компонент, отвечающий за выбор действия в каждом состоянии, реализует стратегию агента. Критик – компонент, который оценивает, насколько хорошим было принятое действие, используя функцию ценности. Такой подход позволяет совмещать преимущества стохастического выбора действий, важного для исследования среды, и оценку этих действий, основанную на накопленном опыте.

Алгоритм PPO работает в условиях марковского процесса принятия решений (S, A, P, R, γ), где S – множество состояний, A – множество действий, $P(s'|s, a)$ – вероятность перехода, $R(s, a)$ – функция награды, $\gamma \in [0, 1]$ – коэффициент дисконтирования.

Параметризованная стратегия $\pi_\theta(a|s)$ определяет вероятность выбора действия a в состоянии s, где θ – параметры нейросети-актора. Критик $V_\phi(s)$ – приближение функции ценности $V^\pi(S) = E[R_t|s_t = s]$, с параметрами ϕ, где $R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots$ – дисконтированная сумма будущих наград. Целевая функция в PPO вместо прямого градиентного обновления использует так называемую функцию обрезки:

$$L^{CLIP}(\theta) = \hat{E}_t[(r_t(\theta) \cdot \hat{A}_t, clip(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) \cdot \hat{A}_t)],$$

где: $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{Old}}(a_t|s_t)}$ – отношение вероятностей нового и старого действия, $\varepsilon \in (0, 1)$ – параметр обрезки – обычно $\varepsilon = 0.1$ или 0.2, $\hat{A}_t$ – оценка преимущества.

Если новое действие разительно отличается от старого (r_t выходит за границы $[1 - \varepsilon, 1 + \varepsilon]$), то градиент подавляется. Это предотвращает резкие изменения стратегии.

Для оценки $\hat{A}_t$ используется обобщенная оценка преимущества (GAE):

$$\hat{A}_t = \sum_{l=0}^{T-t} (\gamma\lambda)^l \delta_{t+l}, \quad \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t),$$

где $\lambda \in [0, 1]$ – параметр сглаживания. Этот способ улучшает устойчивость обучения, снижая дисперсию.

Функция потерь в PPO включает в себя:

- значение потери стратегии L^{CLIP},
- значение потери критика (MSE между предсказанным $V(s_t)$ и целевым значением),
- энтропийное поощрение для поддержания разнообразия действий:

$$L_t^{CLIP+VF+S} = E_t\left[L^{CLIP}(\theta) - c_1 \cdot (V(s_t) - V_t^{target})^2 + c_2 \cdot H[\pi_\theta](s_t)\right],$$

где $H[\pi]$ – энтропия стратегии, c_1, c_2 – коэффициенты.

Схема алгоритма оптимизации проксимальной политики приведена ниже:

Algorithm: PPO

for iteration = 1, 2, … **do**
  for actor = 1, 2, …, N **do**
    run policy $\pi_{\theta_{Old}}$ in environment for T timesteps
    compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  optimize surrogate L w.r.t. θ, with K epochs and minibatch size $M \leq NT$
  $L_t = E_t\left[L^{CLIP}(\theta) - c_1 \cdot (V(s_t) - V_t^{target})^2 + c_2 \cdot H[\pi_\theta](s_t)\right]$
  $\theta_{Old} \leftarrow \theta$
end for

где N – количество параллельных actors, собравших данные на T временных шагах, K – количество эпох. Нейронные сети применяются для аппроксимации искомой стратегии и функции ценности.

Выбор алгоритма PPO в рамках этой работы обусловлен несколькими причинами: алгоритм поддерживает непре-

рывное пространство действий, что критично для задачи локомоции виртуальных организмов с многосуставчатыми конечностями. Ограничение обновлений позволяет стратегии агента развиваться поступательно, не разрушая ранее выученное поведение. PPO может быть успешно использован в архитектурах, включающих сверточные нейронные сети (CNN), которые обрабатывают изображения с камер агентов. Среда Unity ML-agents имеет также встроенную реализацию PPO, что упрощает настройку и ускоряет цикл вычислительных экспериментов.

Нейросеть-актор получает признаки состояния: скорости, положения суставов, контакт с поверхностью и т. п. – и визуальные данные, обработанные через сверточные слои. Цель агента – максимизировать награду, связанную с передвижением и устойчивостью, а также двигаться в выбранном направлении. PPO позволяет мягко адаптировать стратегию к сложной динамике и шумной обратной связи от среды.

Для агента Walker пространство действий представляет собой вектор из 30 непрерывных значений: 18 – для углов и 12 – для усилий (крутящих моментов), управляющих движением сочленений, тогда как агент Searcher управляет только одним параметром – направлением движения (нормализованный угол в диапазоне [−1, 1]). Обе модели обучаются в асинхронном режиме с использованием нескольких параллельных симуляций среды (от 4 до 10), что позволяет эффективно собирать обучающие данные и ускорять процесс оптимизации.

Основные параметры обучения (подробно описаны в документации к (Juliani et al., 2018)):

- алгоритм: PPO (proximal policy optimization);
- фреймворк: Unity ML-Agents + PyTorch backend;
- learning_rate: 3×10^{-4}. Скорость обучения, коэффициент, определяющий величину шага при обновлении параметров нейронной сети;
- batch_size – размер пакета данных, используемого для одного шага обучения: Searcher 1024, Walker 2048;
- buffer_size: 10240. Объем собранных взаимодействий со средой, используемых для одного цикла обучения. Настраивается кратно batch_size x num_envs;
- num_epochs: 3. Число проходов (эпох) оптимизатора по одному буферу данных перед его обновлением;
- gamma (дисконт): Searcher 0.99, Walker 0.995;
- lambda (GAE): 0.95;
- clip_range: 0.2.

Агент Walker обучался отдельно в изолированной среде до достижения устойчивого и прямолинейного передвижения. Среднее количество шагов до сходимости составляло ~2–3 млн. После завершения этого этапа веса модели фиксировали и агент использовался только в режиме inference.

Агент Searcher обучался отдельно от Walker. Среднее количество шагов на один эксперимент составляло от 5 до 10 млн в зависимости от конфигурации среды – разрешение камеры, max_step, число целевых объектов в окружении и т. д.

Параметры симуляции задавали через YAML-конфигурации ML-Agents. Для достижения стабильных и воспроизводимых результатов применяли фиксированный параметр, который задает начальное значение для генератора случайных чисел, используемых в среде и при обучении (random_seed), и согласованные настройки: при изменении числа сред (num_envs) обязательно пропорционально изменялся buffer_size, такое требование наложено средой ML-Agents.

Все эксперименты проводили на компьютере, оборудованном CUDA-совместимым GPU (см. раздел «Производительность системы и профилирование»). Используемые версии: Unity 2022.3, ML-Agents 21.0, PyTorch 2.0.1, Python 3.10.

Проведенные эксперименты. Экспериментальная часть исследования (численные эксперименты) была направлена на изучение влияния параметров зрительной системы, конфигурации среды и архитектурных ограничений на эффективность обучения агентов. Все эксперименты проводили в изолированных средах с использованием фиксированной модели агента Walker и обучаемого агента Searcher. Основные направления исследований:

1. Влияние разрешения изображения с камеры на обучаемость. Был рассмотрен ряд вариантов разрешений: 20×20, 40×40, 60×60, 80×80, 84×84 (разрешение по умолчанию для Unity ML-Agents) и 100×100 пикселей. Для каждого из них проводилось отдельное обучение Searcher, при прочих равных параметрах. Цель – определить минимальное разрешение, при котором агент стабильно достигает целевого поведения (Reward ≥ 5).
2. Влияние возможности управления скоростью. В одном из экспериментов агенту Searcher дополнительно передавалась возможность управления целевой скоростью движения (второй непрерывный параметр на выходе). Цель – выяснить, приведет ли это к более гибкому поведению или усложнит задачу обучения.
3. Изменение максимального угла поворота. Агент Searcher передает команду поворота тела. В различных экспериментах проверялись максимальные значения допустимого угла поворота: 90, 120, 180, 270°. Проверялась гипотеза о том, что большие углы могут упростить навигацию, но сделать поведение менее точным и стабильным.
4. Влияние длины эпизода (параметр max_step). В ходе экспериментов рассматривались два значения параметра max_step: 5000 и 20000. Значение max_step = 5000 использовалось как базовое, поскольку позволяло агенту получать награду быстро и обеспечивало своевременную обратную связь от алгоритма обучения. Значение 20 000 рассматривалось как альтернатива, применимая для задач с более длительными последовательностями действий и отложенной наградой.
5. Проверка с ручным управлением. Для верификации поведения обученной модели Walker была реализована возможность ручного управления агентом, через клавиши A/D, левее/правее. Это позволило, с одной стороны, подтвердить, что наблюдаемые эффекты, например остановка движения, обусловлены динамикой тела, а не стратегией агента Searcher, а с другой – проверить, способен ли человек, используя такой же тип управления, успешно решать задачу поиска цели – оценка управляемости и восприятия среды).

Все эксперименты записывали через систему логирования Unity ML-Agents и анализировали в Tensorboard –

визуальным инструменте для отслеживания хода обучения, позволяющем строить графики изменения награды, функции потерь, скорости симуляции и других метрик в реальном времени. Критерии успешности описаны в разделе «Постановка задачи».

Результаты

Зависимость обучаемости от разрешения изображения

Результаты серии экспериментов с различным разрешением входного изображения показали, что минимальное разрешение, при котором агент стабильно достигает целевого поведения (средняя награда ≥ 5), составляет 84 × 84 пикселя. При разрешениях 20 × 20, 40 × 40 и 60 × 60 пикселей обучение требовало гораздо больше времени, но с сохранением тенденции на улучшение. Разрешение 100 × 100 пикселей также позволяло добиться целевой награды, но обучение шло немного быстрее при разрешении 84 × 84 пикселя из-за меньшей вычислительной нагрузки. Результат эксперимента представлен в виде графиков из Tensorboard на рис. 5.

Влияние возможности управления скоростью на обучение

Добавление второго управляющего параметра, скорости движения, увеличило размерность пространства действий и заметно усложнило обучение. Агенту требовалось больше времени на сходимость, приблизительно на 33 % дольше при прочих равных условиях, и результирующее поведение было менее стабильным: для решения поставленной задачи управление скоростью является во многом избыточным параметром. Это подтверждает простую гипотезу о том, что увеличение числа степеней свободы требует более сложной стратегии и затрудняет обучение модели. Результат эксперимента представлен в виде графика из Tensorboard (рис. 6).

Результаты при изменении угла поворота

Наилучшие результаты были получены при максимальном угле поворота 90°. При увеличении до 120° наблюдалось небольшое снижение стабильности, а при 180 и 270° агент не достигал целевого уровня награды, обучаясь дольше и менее эффективно. Это указывает на то, что слишком широкое пространство действий мешает выработке устойчивой стратегии навигации.

Влияние длины эпизода (max_step)

При max_step = 5000 агент демонстрировал стабильное обучение, своевременно получая обратную связь о достижении целей. При увеличении длины эпизода до 20000 качество обучения не улучшалось, тогда как время симуляции и нагрузка на ресурсы возрастали. Таким образом, параметр max_step = 5000 использовался в качестве основного, поскольку обеспечивал баланс между эффективностью обучения и вычислительными затратами.

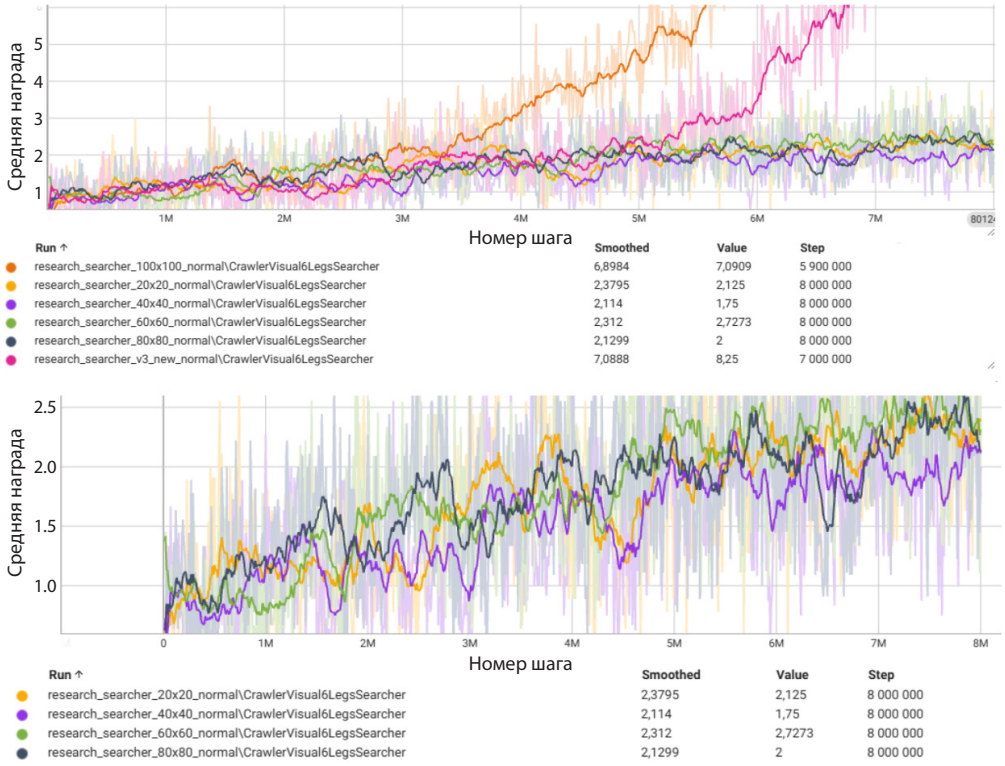


Рис. 5. Результаты обучения агента Searcher при различных разрешениях входного изображения.

В верхней части показаны зависимости средней награды для всех исследованных значений разрешения; малиновая кривая соответствует разрешению 84 × 84 пикселя, оранжевая – 100 × 100 пикселей. В нижней части представлены те же данные с исключением доминирующих графиков, что позволяет более детально рассмотреть оставшиеся варианты: 20 × 20, 40 × 40 и 60 × 60, 80 × 80 пикселей.

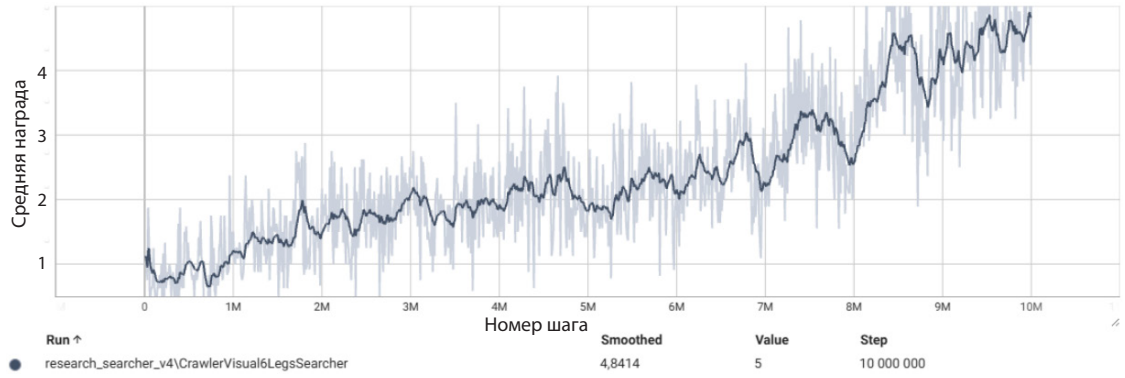


Рис. 6. Результаты обучения агента Searcher при разрешении входного изображения 84×84 пикселя и добавлении второго управляющего параметра, скорости движения, в дополнение к основному, углу поворота.

Производительность системы и профилирование

Для оценки масштабируемости и вычислительной эффективности симулятора было проведено профилирование ключевых компонент системы при различных разрешениях визуального сенсора и числе параллельных сред. Все замеры производили на машине с видеокартой NVIDIA GeForce RTX 3070 и процессором AMD Ryzen 5 7500F (6 ядер, 12 потоков, 3.7 ГГц (5.0 ГГц в турборежиме)).

Вклад основных компонент симуляции в вычислительные затраты системы:

- физика (Physics Engine) – менее 1 мс на шаг, практически не зависит от разрешения;
- графика и сенсоры (Camera.Render, PostProcess) – от 3.2 до 9.5 мс в зависимости от разрешения (почти линейная зависимость);
- нейросеть (PyTorch Inference) – около 35 мс на шаг при использовании сверточной архитектуры Searcher;
- коммуникация Unity и Python (gRPC, сериализация) – от 45 до 60 мс. При увеличении числа агентов этот компонент становится одним из основных узких мест системы, поскольку затраты на коммуникацию: сериализация/десериализация, обмен данными – увеличиваются пропорционально числу агентов;
- прочее (UI, сборка мусора, VSync) – до 20 % времени, при активной отладке может возрастать.

При разрешении 84×84 пикселя и четырех параллельных агентах среднее время одного шага симуляции составляло около 3.6 мс, что эквивалентно приблизительно 278 шагам в секунду. Для разрешения 100×100 пикселей время одного шага увеличивалось до 3.8 мс, что соответствовало снижению производительности примерно до 263 шагов в секунду. Все замеры проводили без визуализации сцены. Число одновременно работающих сред во всех экспериментах с агентом Searcher составляло 4.

Таким образом, основным ограничивающим фактором при масштабировании является не физика или рендеринг, а обмен данными между Unity и Python. Это важно учитывать при планировании больших экспериментов или переходе к популяционному моделированию. Рабочий прототип для воспроизведения результатов доступен в репозитории по ссылке https://github.com/DerpyFox/organism_simulator.

Обсуждение

Интерпретация результатов

Полученные результаты демонстрируют, что успешность обучения агентов с визуальным восприятием напрямую зависит от разрешения входного изображения. Слишком низкое разрешение, до 60×60 пикселей, приводит к потере пространственной структуры сцены и неспособности агента выработать устойчивую стратегию. С другой стороны, разрешение выше 84×84 пикселя не дает ощутимого выигрыша в эффективности, но увеличивает вычислительную нагрузку. Это подтверждает существование оптимального диапазона зрительного восприятия, сопоставимого с эволюционно сформированным у насекомых: их зрение развивалось таким образом, чтобы быть достаточно детализированным для выполнения поведенческих задач (Chittka, Niven, 2009).

Несмотря на наблюдаемую зависимость между разрешением зрительной системы и успешностью обучения агентов, стоит отметить, что в природе существуют организмы, способные к эффективному поведению даже при крайне низком разрешении зрения. Например, у некоторых видов муравьев, как было упомянуто во введении, визуальная система сопоставима по масштабу с разрешением 10×10 пикселей, однако это не мешает им уверенно ориентироваться, находить пищу, взаимодействовать с окружающей средой и даже проходить зеркальный тест (Cammaerts M.-C.T., Cammaerts R., 2015). Такая эффективность обусловлена не только зрением, но и развитой обонятельной системой, которая играет ключевую роль в восприятии окружающего мира. Кроме того, нейронные системы реальных насекомых могут обладать рядом качеств, повышающих их эффективность. Эти системы сформировались в ходе длительной эволюции и адаптированы к конкретным условиям жизни и типичным задачам живого организма, например ориентации в сложной среде, поиску пищи, взаимодействию с сородичами. В них присутствует высокая степень специализации нейронов и механизм адаптации к изменяющимся стимулам. Такая настройка под реальные условия позволяет эффективно обрабатывать даже ограниченные или фрагментарные сенсорные сигналы, включая визуальные, обонятельные и механические.

Добавление управления скоростью и увеличение угла поворота показали, что даже небольшое расширение пространства действий замедляет обучение. Таким образом, важно сохранять баланс между выразительностью модели и ее обучаемостью. Разделение задач восприятия и управления телом между агентами Searcher и Walker оказалось критически важным для достижения стабильного поведения.

Биологические параллели и когнитивная эффективность

Результаты перекликаются с принципами, наблюдаемыми у насекомых: минимальные, но функционально избыточные зрительные системы позволяют успешно ориентироваться и принимать решения в реальном времени. Аналогично предложенная архитектура позволяет агенту достигать целевых стратегий при ограниченном разрешении и небольшой нейросети.

Если рассматривать полученные результаты в контексте реальных биологических систем, можно провести параллель с эволюционными компромиссами, возникающими между сенсорной точностью, вычислительными затратами и адаптивностью поведения. Например, зрительные системы насекомых, таких как мухи дрозофилы (~150 тыс. нейронов) или пчелы (~960 тыс. нейронов), обеспечивают базовое распознавание объектов и ориентацию в пространстве при минимуме нейронов и крайне ограниченной пропускной способности (Menzel, 2012). Эти организмы не обладают высоким разрешением зрительной системы, но достигают высокой эффективности благодаря комбинации быстрого отклика, архитектуры сенсомоторных связей и стратегиям принятия решений (Chittka, Niven, 2009). Подобные соображения хорошо иллюстрируются на примере насекомых с высоким уровнем общественной организации. У муравьев разделение труда и коммуникация формируются не только как врожденные стереотипы, но и как результат гибкой адаптации на уровне отдельных особей. Распределение ролей в колонии может изменяться в зависимости от возраста, физиологического состояния и текущей ситуации, а передача информации между муравьями осуществляется через разнообразные сигналы (Chittka, Muller, 2009). Таким образом, даже простые агенты, обладающие ограниченными когнитивными возможностями, могут достигать высокой эффективности за счет организации взаимодействий и простых правил поведения.

Агенты в нашем симуляторе демонстрируют способность к адаптивному поведению уже при относительно низком зрительном разрешении, например 84 × 84 пикселя, что позволяет проводить дальнейшие аналогии с минимальными когнитивными системами в природе. Подобные модели могут использоваться как искусственные системы, воспроизводящие ключевые поведенческие аспекты простых организмов, и служить основой для построения гипотез о нейрофизиологических механизмах восприятия и поведения у беспозвоночных.

Ограничения системы

Главное ограничение системы – коммуникационная нагрузка между средой Unity и фреймворком обучения PyTorch. Даже при высокой производительности вычислительных блоков именно сериализация и передача данных по gRPC становятся узким местом (англ. bottleneck, «бутылочное горлышко»). Кроме того, на этой стадии среда пока ограничена по сложности: отсутствуют препятствия, изменяющаяся топография и взаимодействие между агентами. Наконец, архитектуры агентов фиксированы и не подвергаются эволюции или адаптации во времени, меняются лишь веса параметров, но не топология сети.

Перспективы развития

Дальнейшее развитие возможно в нескольких направлениях. Внедрение механизмов нейроэволюции, например подхода NEAT – NeuroEvolution of Augmenting Topologies, позволило бы исследовать не только изменение весов нейросети, но и эволюционную оптимизацию ее структуры. Это особенно актуально в контексте энергозатрат: при избыточной сложности мозга возрастает потребление ресурсов, тогда как в более простой среде может быть выгодно уменьшение числа нейронов. Таким образом, агенты могли бы самостоятельно адаптировать размер и, возможно, архитектуру своей нейросети, снижая избыточность в условиях низкой когнитивной нагрузки. В биологических системах даже небольшое усложнение нервной системы может приводить к заметному росту энергозатрат. Так, у мух *Calliphora vicina* только сетчатка глаза потребляет около 8 % всего обмена в состоянии покоя организма (Niven, Laughlin, 2008). При этом у человека мозг составляет лишь 2 % массы тела, но потребляет до 20 % энергии организма (Attwell, Laughlin, 2001). Эти данные показывают, что выигрыш от сокращения числа нейронов или уменьшения сложности сенсорной системы может быть существенным.

Введение элементов среды с конкуренцией за ресурсы: несколько агентов, ограниченное количество целей, более умные агенты могут выбирать и решать более сложные когнитивные задачи из числа различных имеющихся в системе, получая дополнительные преимущества, – даст возможность анализировать стратегии поведения в популяции.

Перспективно направление «добавление обонятельной модели» – сенсорного канала, основанного на кратковременных «следах» в среде, аналогичных феромонным меткам у муравьев. Такие следы могут исчезать со временем, различаться по содержанию, например, сообщать, сыт муравей или голоден, и влиять на маршруты агентов, усиливая элементы косвенной коммуникации и коллективного поведения. Также целесообразно добавление памяти и рекуррентных модулей в модель Searcher для изучения навигации в условиях частичного наблюдения.

Заключение

Настоящее исследование было направлено на количественную и качественную оценку архитектурных и сенсорных параметров в задаче обучения агентов с визуальным восприятием в трехмерной симуляции. Мы предложили и реализовали иерархическую модель управления, в которой агент локомоции (Walker) функционирует как низкоуровневый исполнитель движений, а агент восприятия и

навигации (Searcher) принимает стратегические решения на основе зрительной информации.

Систематический анализ показал, что даже при ограниченном за счет низкого разрешения потоке сенсорной информации агенты способны вырабатывать устойчивые поведенческие стратегии, если модель и условия среды подобраны с учетом когнитивной нагрузки. Было установлено, что разрешение 84×84 пикселя обеспечивает компромисс между вычислительной эффективностью и минимальной когнитивной состоятельностью, а увеличение размерности пространства действий без соответствующего увеличения обучающих ресурсов ведет к ухудшению результатов.

Наши результаты подтверждают гипотезу о том, что минимально сложные нейросетевые агенты могут реализовывать сложные поведенческие паттерны в условиях ограниченного сенсорного восприятия, при котором агент получает лишь частичную информацию о среде. Эти выводы согласуются с наблюдаемыми в природе примерами когнитивной эффективности у беспозвоночных, таких как муравьи или пчелы, и открывают перспективу использования подобных моделей в биологическом моделировании, робототехнике и исследованиях в области нейроэволюции.

В перспективе возможно развитие системы в сторону популяционных симуляций, включающих конкуренцию, коммуникацию между агентами и адаптацию стратегий в изменяющейся среде. Архитектура может быть дополнена модулями памяти, рекуррентными связями или механизмами нейроэволюции, что позволит исследовать более сложные когнитивные явления в виртуальных популяциях.

Также показано, что использование зрительной информации, несмотря на свою выразительность, требует значительных вычислительных ресурсов и в некоторых случаях может уступать по эффективности более простым сенсорным моделям. Эти наблюдения подчеркивают важность выбора сенсорной архитектуры при построении минимально состоятельных когнитивных агентов.

Кроме того, одним из ключевых выводов стало понимание критической роли дизайна среды и структуры обучения в успешности моделирования. Первоначальные попытки обучать поведение через единую нейросетевую модель, объединяющую локомоцию и стратегию, не формировали умение обнаруживать и собирать цели (единицы «пищи») из-за сложности в балансировке наград и формулировке задачи. Введение подхода с разделением функций – поиск и движение – позволило достичь значительного улучшения в обучаемости и стабильности поведения.

Таким образом, полученные результаты демонстрируют потенциал нейроагентных систем в задачах биологически мотивированного моделирования и создают основу для последующих экспериментов, направленных на изучение границ когнитивной сложности при ограниченных ресурсах восприятия и управления.

Список литературы / References

Aksoy V., Camlitepe Y. Spectral sensitivities of ants – a review. *Anim Biol.* 2018;68(1):55-73. doi 10.1163/15707563-17000119

Attwell D., Laughlin S.B. An energy budget for signaling in the grey matter of the brain. *J Cereb Blood Flow Metab.* 2001;21(10):1133-1145. doi 10.1097/00004647-200110000-0000

Bongard J.C. Evolutionary robotics. *Commun ACM*. 2013;56(8):74-83. doi 10.1145/2493883

Botvinick M., Wang J.X., Dabney W., Miller K.J., Kurth-Nelson Z. Deep reinforcement learning and its neuroscientific implications. *Neuron*. 2020;107(4):603-616. doi 10.1016/j.neuron.2020.06.014

Cammaerts M.C. The visual perception of the ant *Myrmica ruginodis* (Hymenoptera: Formicidae). *Biologia.* 2012;67(6):1165-1174. doi 10.2478/s11756-012-0112-z

Cammaerts M.-C.T., Cammaerts R. Are ants (Hymenoptera, Formicidae) capable of self recognition? *J Sci.* 2015;5(7):521-532

Chittka L., Muller H. Learning, specialization, efficiency and task allocation in social insects. *Commun Integr Biol.* 2009;2(2):151-154. doi 10.4161/cib.7600

Chittka L., Niven J. Are bigger brains better? *Curr Biol.* 2009;19(21):R995-R1008. doi 10.1016/j.cub.2009.08.023

Devyaterikov A.P., Palyanov A.Y. A software system for modeling evolution in a population of organisms with vision, interacting with each other in 3D simulator. *Vavilov J Genet Breed*. 2022;26(8):780-786. doi 10.18699/VJGB-22-94

Hassabis D., Humaran D., Summerfield C., Botvinick M. Neuroscience-inspired artificial intelligence. *Neuron.* 2017;95(2):245-258. doi 10.1016/j.neuron.2017.06.011

Juliani A., Berges V.-P., Teng E., Cohen A., Harper J., Elion C., Goy C., Gao Y., Henry H., Matter M., Lange D. Unity: A general platform for intelligent agents. *arXiv*. 2018. doi 10.48550/arXiv.1809.02627

Land M.F., Nilsson D.-E. Animal Eyes. Oxford University Press, 2012. Available: https://www.softouch.on.ca/kb/data/Animal%20Eyes.pdf

Menzel R. The honeybee as a model for understanding the basis of cognition. *Nat Rev Neurosci.* 2012;13(11):758-768. doi 10.1038/nrn3357

Niven J.E., Laughlin S.B. Energy limitation as a selective pressure on the evolution of sensory systems. *J Exp Biol.* 2008;211:1792-1804. doi 10.1242/jeb.017574

O'shea K., Nash R. An introduction to convolutional neural networks. *arXiv*. 2015. doi 10.48550/arXiv.1511.08458

Peng P., Wen Y., Yang Y., Yuan Q., Tang Z., Long H., Wang J. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play StarCraft combat games. *arXiv*. 2017. doi 10.48550/arXiv.1703.10069

Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms. *arXiv*. 2017. doi 10.48550/arXiv.1707.06347

Song K.-Y., Behzadfar M., Zhang W.-J. A dynamic pole motion approach for control of nonlinear hybrid soft legs: A preliminary study. *Machines*. 2022;10(10):875. doi 10.3390/machines10100875

Stanley K.O., Clune J., Lehman J., Miikkulainen R. Designing neural networks through neuroevolution. *Nat Mach Intell*. 2019;1(1):24-35. doi 10.1038/s42256-018-0006-z

Tschantz A., Anil K.S., Christopher L.B. Learning action-oriented models through active inference. *PLoS Comput Biol*. 2020;16(4):e1007805. doi 10.1371/journal.pcbi.1007805

Vezhnevets A.S., Osindero S., Schaul T., Heess N., Jaderberg M., Silver D., Kavukcuoglu K. FeUdal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Vol. 70 (ICML'17). JMLR.org, 2017;3540-3549

Zador A.M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun.* 2019;10(1):3770. doi 10.1038/s41467-019-11786-6

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 30.07.2025. После доработки 23.09.2025. Принята к публикации 23.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1062-1072

doi 10.18699/vjgb-25-111

Молекулярно-динамический анализ функциональной роли аминокислотных остатков V99, F124 и S125 ДНК-диоксигеназы человека ABH2

М. Чжао [1], Т.Е. Тюгашев [2], А.Т. Давлетгильдеева [2], Н.А. Кузнецов [1, 2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
kuznetsov@1bio.ru

Аннотация. ДНК-диоксигеназа человека ABH2 относится к семейству AlkB-подобных негемовых диоксигеназ, которые действуют на широкий спектр субстратов и обладают сложным каталитическим механизмом с участием α-кетоглутарата и иона Fe(II) в качестве кофактора. Представители семейства AlkB катализируют прямое окисление алкильных заместителей в азотистых основаниях ДНК и РНК, обеспечивая защиту от мутагенного воздействия эндогенных и экзогенных алкилирующих агентов, а также участвуя в регуляции уровня метилирования некоторых РНК. Фермент ABH2, локализованный преимущественно в ядре клетки, проявляет специфичность к двуцепочечным ДНК-субстратам и, в отличие от большинства других AlkB-подобных ферментов человека, обладает довольно широким спектром субстратной специфичности, окисляя алкильные группы таких модифицированных азотистых оснований, как, например, N^1-метиладенозин, N^3-метилцитидин, 1,N^6-этеноаденозин и 3,N^4-этеноцитидин. В данной работе с целью анализа механизма, обеспечивающего субстратную специфичность фермента, и выяснения функциональной роли аминокислотных остатков в составе активного центра нами выполнено молекулярно-динамическое моделирование комплексов фермента ABH2 дикого типа и его мутантных форм, содержащих аминокислотные замены V99A, F124A или S125A, с двумя типами ДНК-субстратов, несущих метилированные основания N^1-метиладенин или N^3-метилцитозин. Установлено, что замена V99A приводит к увеличению подвижности белковых петель L1 и L2, участвующих в связывании ДНК-субстрата, и изменяет распределение π-π-контактов боковой цепи остатка F102 с азотистыми основаниями, расположенными рядом с поврежденным нуклеотидом. Замена F124A приводит к потере π-π-стэкинга с поврежденным основанием, что, в свою очередь, дестабилизирует архитектуру активного центра, вызывает нарушение взаимодействия с ионом железа и препятствует оптимальному каталитическому позиционированию α-кетоглутарата в активном центре. Замена S125A приводит к потере прямого взаимодействия петли L2 с 5′-фосфатной группой поврежденного нуклеотида, ослабляя связывание фермента с ДНК-субстратом. Таким образом, полученные данные позволили установить функциональную роль трех аминокислотных остатков активного центра и расширить понимание структурно-функциональных связей в процессах узнавания поврежденного нуклеотида и формирования каталитического комплекса ферментом ABH2 человека.

Ключевые слова: репарация ДНК; метилирование оснований; ДНК-диоксигеназа человека ABH2; МД-моделирование; функциональная роль аминокислотных остатков

Для цитирования: Чжао М., Тюгашев Т.Е., Давлетгильдеева А.Т., Кузнецов Н.А. Молекулярно-динамический анализ функциональной роли аминокислотных остатков V99, F124 и S125 ДНК-диоксигеназы человека ABH2. *Вавиловский журнал генетики и селекции*. 2025;29(7):1062-1072. doi 10.18699/vjgb-25-111

Финансирование. Работа выполнена в рамках государственного задания № 121031300041-4.

Благодарности. Работа проведена с использованием ресурсов ИВЦ НГУ.

Molecular dynamic analysis of the functional role of amino acid residues V99, F124 and S125 of human DNA dioxygenase ABH2

M. Zhao [1], T.E. Tyugashev [2], A.T. Davletgildeeva [2], N.A. Kuznetsov [1, 2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
kuznetsov@1bio.ru

Abstract. The ABH2 enzyme belongs to the AlkB-like family of Fe(II)/α-ketoglutarate-dependent dioxygenases. Various non-heme dioxygenases act on a wide range of substrates and have a complex catalytic mechanism involving α-ketoglutarate and an Fe(II) ion as a cofactor. Representatives of the AlkB family catalyze the direct oxidation of alkyl

© Чжао М., Тюгашев Т.Е., Давлетгильдеева А.Т., Кузнецов Н.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

substituents in the nitrogenous bases of DNA and RNA, providing protection against the mutagenic effects of endogenous and exogenous alkylating agents, and also participate in the regulation of the methylation level of some RNAs. DNA dioxygenase ABH2, localized predominantly in the cell nucleus, is specific for double-stranded DNA substrates and, unlike most other human AlkB-like enzymes, has a fairly broad spectrum of substrate specificity, oxidizing alkyl groups of such modified nitrogenous bases as, for example, N^1-methyladenosine, N^3-methylcytidine, 1,N^6-ethenoadenosine and 3,N^4-ethenocytidine. To analyze the mechanism underlying the enzyme's substrate specificity and to clarify the functional role of key active-site amino acid residues, we performed molecular dynamics simulations of complexes of the wild-type ABH2 enzyme and its mutant forms containing amino acid substitutions V99A, F124A and S125A with two types of DNA substrates carrying methylated bases N^1-methyladenine and N^3-methylcytosine, respectively. It was found that the V99A substitution leads to an increase in the mobility of protein loops L1 and L2 involved in binding the DNA substrate and changes the distribution of π-π contacts between the side chain of residue F102 and nitrogenous bases located near the damaged nucleotide. The F124A substitution leads to the loss of π-π stacking with the damaged base, which in turn destabilizes the architecture of the active site, disrupts the interaction with the iron ion and prevents optimal catalytic positioning of α-ketoglutarate in the active site. The S125A substitution leads to the loss of direct interaction of the L2 loop with the 5′-phosphate group of the damaged nucleotide, weakening the binding of the enzyme to the DNA substrate. Thus, the obtained data revealed the functional role of three amino acid residues of the active site and contributed to the understanding of the structural-functional relationships in the recognition of a damaged nucleotide and the formation of a catalytic complex by the human ABH2 enzyme.
Key words: DNA repair; base methylation; human DNA dioxygenase ABH2; MD modeling; functional role of amino acid residues

For citation: Zhao M., Tyugashev T.E., Davletgildeeva A.T., Kuznetsov N.A. Molecular dynamic analysis of the functional role of amino acid residues V99, F124 and S125 of human DNA dioxygenase ABH2. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1062-1072. doi 10.18699/vjgb-25-111

Введение

Стабильность генетической информации, закодированной в виде последовательности нуклеотидов ДНК, чрезвычайно важна для нормального функционирования и выживания отдельных клеток, организмов и видов в целом (Travers, Muskhelishvili, 2015). При этом клеточная ДНК всех живых организмов регулярно подвергается повреждающему воздействию различных эндогенных и экзогенных факторов, таких как химически активные реагенты и метаболиты, ионизирующее и УФ-излучение и др. (Ougland et al., 2015). У живых организмов в ходе эволюции появилось множество различных путей репарации повреждений, возникающих в геномной ДНК, некоторые из которых представлены только одним ферментом, а другие включают последовательную и согласованную работу целых ферментативных каскадов (Yi et al., 2009; Li et al., 2013; Müller, Hausinger, 2015; Ougland et al., 2015).

Среди ферментов, принимающих участие в узнавании и удалении необъемных единичных повреждений азотистых оснований ДНК, можно выделить следующие: 1) ДНК-гликозилазы, удаляющие поврежденные азотистые основания с образованием в ДНК апуриновых/апиримидиновых сайтов, которые далее процессируются с восстановлением исходной структуры ДНК другими ферментами системы эксцизионной репарации оснований (BER) (Ringvoll et al., 2006; Chen et al., 2010; Li et al., 2013); 2) O^6-алкилгуанин-ДНК-алкилтрансферазы (AGT), переносящие алкильный аддукт на собственный цистеиновый остаток (Ringvoll et al., 2006); 3) фотолиазы, отвечающие за удаление УФ-индуцированных фотоповреждений, таких как циклобутан-пиримидиновые димеры и пиримидин-пиримидиновые фотопродукты (Yi, He, 2013); 4) диоксигеназы семейства AlkB, входящие в суперсемейство Fe(II)/α-кетоглутарат(αKG)-зависимых диоксигеназ, которые используют негемовое железо в качестве кофактора и αKG в качестве косубстрата для прямого окисления алкильных групп в поврежденных основаниях ДНК (Yang et al., 2009; Yi et al., 2009; Kuznetsov et al., 2021). Следует отметить, что разнообразие путей репарации необъемных повреждений ДНК связано с большим разнообразием возможных химических модификаций азотистых оснований.

Представители семейства Fe(II)/αKG-зависимых диоксигеназ AlkB, обнаруженные у человека, в последние годы вызывают особый интерес из-за участия их в репарации алкилированных оснований ДНК. Считается, что ферменты данного семейства могут играть важную роль в прогрессировании некоторых онкологических заболеваний, поскольку они часто сверхэкспрессируются в опухолевых клетках и нивелируют эффект алкилирующих препаратов, используемых в химиотерапии. ABH2 – один из первых идентифицированных у человека представителей семейства AlkB-подобных диоксигеназ (Duncan et al., 2002; Aas et al., 2003). Изменение уровня экспрессии ABH2 оказывает влияние на эффективность удаления некоторых токсичных повреждений ДНК в опухолевых клетках, что делает этот фермент потенциальным маркером для диагностики онкологических заболеваний и возможной терапевтической мишенью (Wilson et al., 2018).

Известно, что ABH2 проявляет активность как минимум по отношению к восьми различным алкилированным основаниям ДНК, а именно: к N^1-метиладенозину (m^1A), N^3-метилцитидину (m^3C), N^3-метилтимидину (m^3T), N^3-этилтимидину (N^3-EtT), 1,N^6-этеноаденозину (εA), 3,N^4-этеноцитидину (εC), 1,N^2-этеногуанозину (1,N^2-εG) и 5-метилцитидину (m^5C) (рис. 1) (Falnes, 2004; Ringvoll et al., 2006, 2008; Bian et al., 2019).

Метилирование – самый распространенный тип повреждений оснований ДНК, вызванных воздействием алкилирующих агентов (Sall et al., 2022), а m^1A и m^3C являются субстратами, наиболее эффективно удаляемыми ABH2 из двуцепочечной ДНК (дцДНК) (Duncan et al., 2002; Aas et al., 2003; Xu et al., 2021). D.H. Lee с коллега-

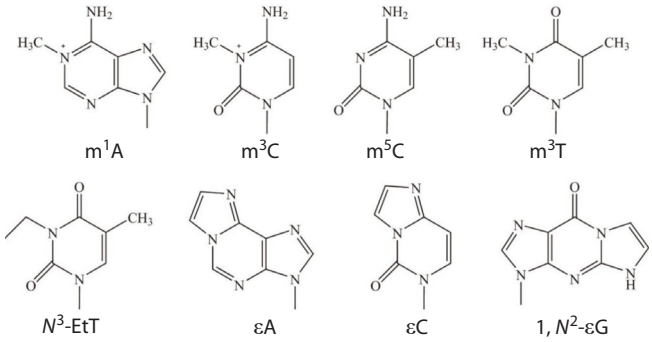


Рис. 1. Алкилированные азотистые основания, являющиеся субстратами для ДНК-диоксигеназы человека ABH2.

ми показали, что ABH2 окисляет m^1A и m^3C в контексте дцДНК как минимум в два раза эффективнее, по сравнению с одноцепочечной ДНК (оцДНК) (Lee et al., 2005).

Известные на сегодняшний день структурные данные позволяют предположить особенности функционирования фермента ABH2 и механизм, обеспечивающий его субстратную специфичность. ABH2 содержит высококонсервативный каталитический домен – двуцепочечный β-спиральный домен (DSBH) суперсемейства Fe(II)/αKG-зависимых диоксигеназ. Неструктурированный N-концевой фрагмент ABH2 также включает мотив связывания ядерного антигена пролиферирующих клеток (PCNA) (Xu et al., 2021). Триплет, состоящий из двух аминокислотных остатков гистидина и одного аспартата (H171, H236 и D173), координирует кофактор Fe(II) в активном центре фермента (Giri et al., 2011; Xu et al., 2021). Аминокислотный остаток D173 посредством взаимодействия с остатком R254 также принимает участие в формировании сети водородных связей, включающей аминокислотные остатки N159, Y161, R248, T252 и R254, координирующие косубстрат αKG в активном центре фермента (Waheed et al., 2020).

Активный центр ABH2 окружен четырьмя функциональными петлями, L1–L4 (рис. 2). Эти петли играют ключевую роль в стабилизации положения ДНК-субстрата в активном центре фермента (Xu et al., 2021). Петля L1, включающая аминокислотные остатки 98–107, содержит гидрофобную шпильку V101-F102-G103, с помощью которой происходит «тестирование» стабильности пар оснований в субстрате. Если поврежденное основание формирует нестабильную пару со своим партнером из комплементарной цепи, остатки V101 и F102 индуцируют выворачивание поврежденного нуклеотида в активный центр. При этом освободившееся пространство в дуплексе ДНК заполняется остатком F102, стабилизирующим вывернутое положение нуклеотида за счет π-π-взаимодействия с окружающими основаниями (Chen et al., 2010, 2014; Yi et al., 2012; Xu et al., 2021).

Петля L2, включающая аминокислотные остатки 122–129, вместе с петлей L1 образует так называемую крышку распознавания нуклеотидов (NRL). Аминокислотный остаток Y122 участвует в сети водородных связей, формирующей каталитически компетентное состояние активного центра фермента (Davletgildeeva et al., 2023); остаток S125 образует водородную связь с 5′-фосфатом вывернутого поврежденного нуклеотида; аминокислотные остатки F124 и H171 образуют π-π-стэкинг с вывернутым азотистым основанием (Chen et al., 2010, 2014;

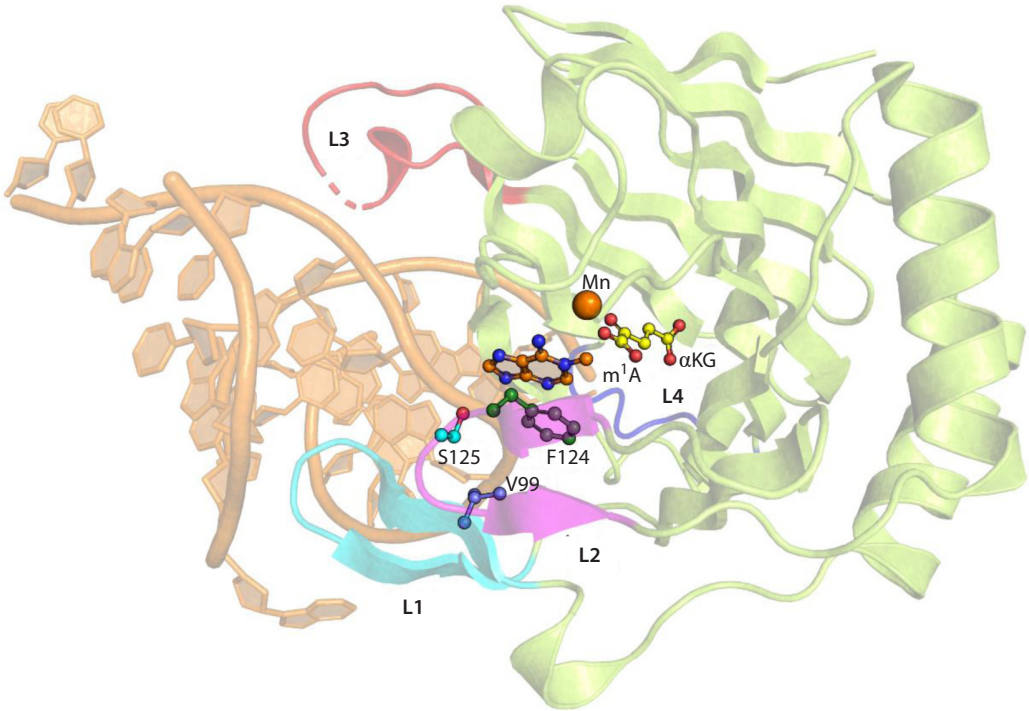


Рис. 2. Кристаллографическая структура комплекса ABH2 с дцДНК, содержащей m^1A (PDB ID 3BUC).

Отмечены петли L1–L4; показаны поврежденное азотистое основание m^1A, αKG и ион Mn^{2+}, а также рассматриваемые в данной работе аминокислотные остатки V99, F124 и S125.

Lenz et al., 2020). Аминокислотный остаток S125 также участвует в формировании стенки кармана, связывающего повреждение, наряду с остатками V99, R110 и I168 (Davletgildeeva et al., 2023).

Следует отметить, что V99 занимает важное положение в сети гидрофобных остатков, сформированных V101, V108, F124 и L127 (Monsen et al., 2010). Петля L3, включающая аминокислотные остатки 198–213, и петля L4, состоящая из аминокислотных остатков 237–247, играют важную роль в связывании с дцДНК субстратом. Аминокислотные остатки R198, R203 и K205 в петле L3 и последовательность RKK (R241-K242-K243) в петле L4 формируют контакты с цепью ДНК, комплементарной поврежденной цепи, тем самым обеспечивая эффективное связывание дцДНК субстрата ферментом ABH2 (Yang et al., 2008; Yi et al., 2009; Waheed et al., 2020).

Молекулярно-динамический анализ структурных данных и экспериментальная проверка активности рекомбинантных препаратов ABH2 дикого типа и ряда его мутантных форм, проведенные нашей группой ранее, позволили установить роль аминокислотных остатков Y122, I168 и D173, которые образуют прямые контакты с основаниями m^1A, m^3C, а также m^5C в кармане активного центра (Davletgildeeva et al., 2023). Сравнительный анализ ферментов выявил влияние замен данных аминокислотных остатков на каталитическую активность фермента и лишь незначительное уменьшение эффективности связывания ДНК. Полученные данные позволили предположить, что эти остатки отвечают за прецизионное позиционирование вывернутого поврежденного нуклеотида в кармане активного центра, которое обеспечивает эффективное протекание каталитической реакции (Davletgildeeva et al., 2023).

Нужно отметить, что широкий спектр субстратной специфичности фермента ABH2 и сложный каталитический механизм действия, включающий кофактор и косубстрат для осуществления реакции, осложняют детальные исследования молекулярного механизма узнавания поврежденной ДНК и формирования каталитически компетентного комплекса, а также локальных конформационных изменений, влияющих на эффективность каталитической реакции. В связи с этим в настоящем исследовании с целью дальнейшего выяснения механизма субстратной специфичности ДНК-диоксигеназы ABH2 с применением метода молекулярной динамики проведен анализ функциональной роли трех аминокислотных остатков, V99, F124 и S125, участвующих в формировании кармана, в котором располагается вывернутый нуклеотид, содержащий поврежденное основание.

Материалы и методы

Модели комплексов были построены на основе рентгенографических структур комплексов ABH2-дцДНК с ионом металла (Mn^{2+}) и αKG: 3BUC (для m^1A) и 3RZJ (для m^3C) (Yang et al., 2008; Yi et al., 2012). Изменение последовательности ДНК, исправление неразрешенных аминокислотных остатков и модификаций фермента выполнены с помощью Chimera и Modeller (Šali, Blundell, 1993), оптимизация протонирования ионизируемых групп – с использованием сервера H++ (Anandakrishnan et al., 2012). МД-моделирование осуществляли в GROMACS (Abraham et al., 2015). Комплекс был помещен в додэкаэдрическую ячейку с TIP3P водой и 50 мМ KCl (Jorgensen et al., 1983; Joung, Cheatham, 2008), для описания комплекса использовали силовое поле AMBER14SB/OL15 (Cornell et al., 1995; Zgarbová et al., 2011, 2015; Maier et al., 2015).

Параметризация для m^1A, m^3C и αKG проведена с использованием модуля Antechamber (пакет AMBER); заряды RESP рассчитаны на сервере REDD; топологии модифицированных остатков преобразованы в формат GROMACS с помощью ACPYPE (Bayly et al., 1993; Wang et al., 2004, 2006; Vanquelef et al., 2011; Sousa da Silva, Vranken, 2012).

С целью сохранения октаэдрической координационной геометрии иона Fe^{2+} при возможных возмущениях активного центра, вносимых заменой аминокислотных остатков, для описания иона применяли модель с разнесенными зарядами (Jiang et al., 2016). Для МД-расчетов применяли следующие параметры: минимизация энергии системы методом наискорейшего спуска; пороговое значение ван-дер-ваальсовых взаимодействий установлено на 10 Å; учет дальнодействующих кулоновских взаимодействий по методу PME (particle mesh Ewald) (Essmann et al., 1995); ограничение колебаний ковалентных связей атомов водорода методом LINCS (Hess et al., 1997).

После минимизации система нагревалась до 310 K в NVT-ансамбле в течение 500 пс с использованием термостата V-rescale (Bussi et al., 2007). Затем проводили уравновешивание в NPT-ансамбле в течение 1 нс, давление поддерживалось на уровне 1 бар с помощью баростата Парринелло–Рахмана (Parrinello, Rahman, 1981).

Расчеты методом классической молекулярной динамики выполняли длительностью 250 нс не менее трех раз. Анализ траекторий проводили с использованием встроенных средств GROMACS и библиотеки MDTraj (McGibbon et al., 2015). Изменение распределения между устойчивыми состояниями комплексов фермента ABH2 дикого типа и его мутантных форм с ДНК-субстратами показано в графиках распределения расстояний между ключевыми атомами при моделировании. Распределение межатомных расстояний в МД-траектории представлено в виде гистограммы с шагом 0.1 Å и высотой шага, равной процентной доле кадров траектории, в которых расстояние попадает в соответствующий диапазон значений. Для каждой траектории сумма долей на всем диапазоне расстояний равна 100 %.

Результаты и обсуждение

Модель фермент-субстратного комплекса ABH2 V99A с поврежденной ДНК

При моделировании фермент-субстратных комплексов и с m^1A-содержащим дцДНК-субстратом (далее m^1A-ДНК, рис. 3, *а*, *б*), и с m^3C-содержащим дцДНК-субстратом (далее m^3C-ДНК, см. рис. 3, *в*, *г*) замена V99A приводила к изменениям как в области петель L1 и L2, взаимодействующих с вывернутым в активный центр фермента нуклеотидом и близлежащим участком дцДНК, так и в области связывания косубстрата. Так, в модели комплекса с m^1A-ДНК боковая цепь аминокислотного остатка F124 утратила π-π-стэкинговые взаимодействия с осно-

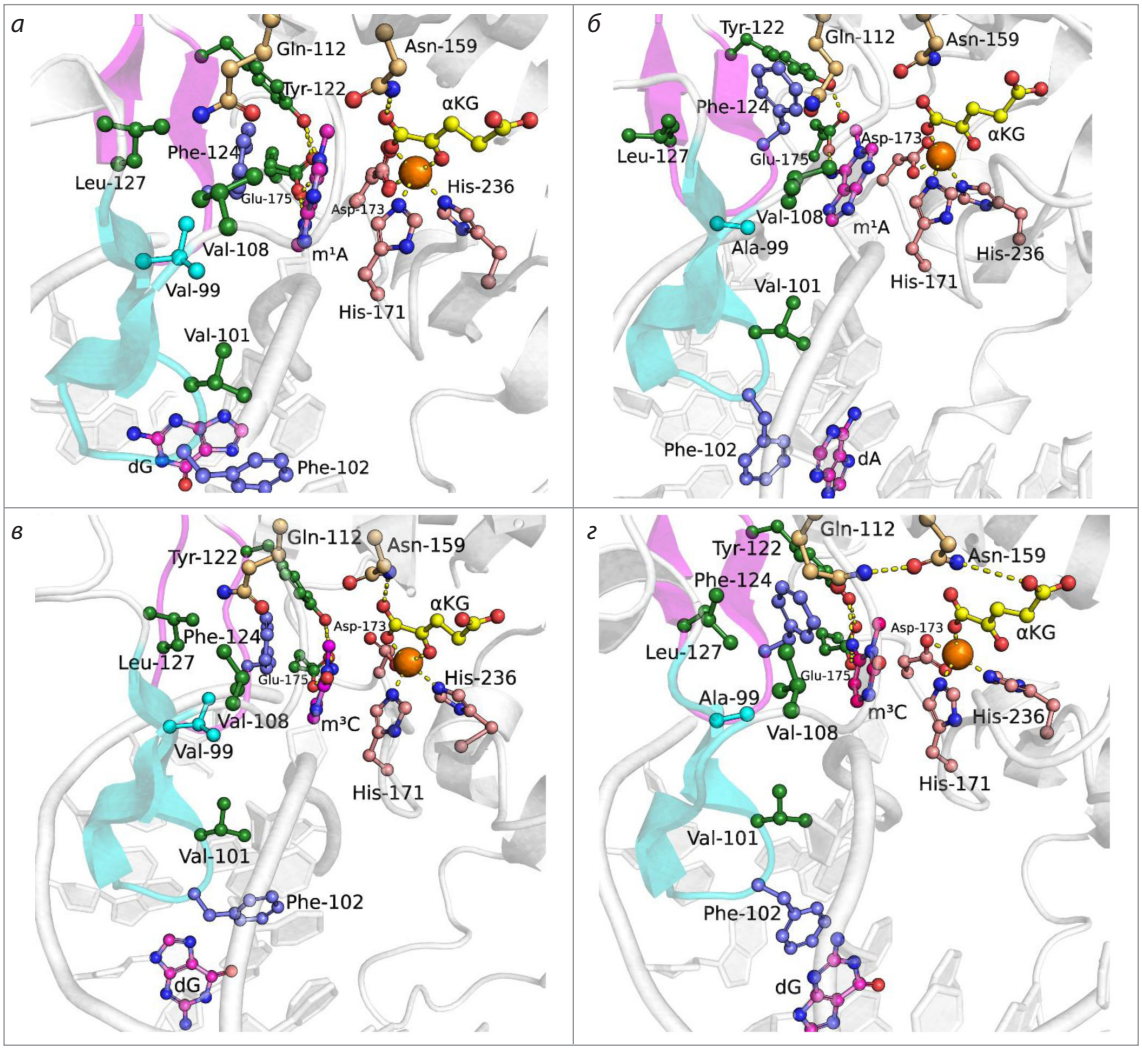


Рис. 3. Репрезентативные МД-структуры ABH2 WT в комплексе с m^1A-ДНК (*а*) и m^3C-ДНК (*в*) и ABH2 V99A в комплексе с m^1A-ДНК (*б*) и m^3C-ДНК (*г*).

Показаны ключевые аминокислотные остатки активного центра, поврежденное азотистое основание, αKG и ион Mn^{2+}. Петли L1 и L2 выделены голубым и розовым цветом соответственно.

ванием вывернутого в активный центр нуклеотида (см. рис. 3, *а*, *б*). При этом уменьшилось время существования водородной связи между гидроксильной группой Y122 и экзоциклической аминогруппой поврежденного основания (рис. 4, *а*). В модели комплекса с m^3C-содержащим дцДНК-субстратом также происходила частичная потеря контакта между гидроксильной группой боковой цепи Y122 и карбоксильной группой боковой цепи E175 (см. рис. 4, *б*), что также нарушает сеть контактов, стабилизирующих вывернутое основание.

Замена V99A индуцировала изменение положения интеркалирующего в ДНК остатка F102, входящего в состав петли L1. При этом в комплексе с m^1A-ДНК происходило перераспределение π-π-контактов, формируемых F102, от азотистого основания комплементарной цепи у WT фермента (см. dG на рис. 3, *а*) к азотистому основанию поврежденной цепи в случае ABH2 V99A (см. dA на рис. 3, *б*).

Значения двугранного угла C-Cα-Cβ-Cγ у остатка F102 составляли 148.1 ± 55.3° для WT фермента и 127.2 ± 47.7° для мутантной формы V99A, что свидетельствует о стабильности данных положений в процессе молекулярной динамики. Тогда как в комплексе с m^3C-ДНК замена V99A индуцировала значительное увеличение подвижности его боковой цепи (двугранный угол C-Cα-Cβ-Cγ равен 135.6 ± 58.6 и 100.2 ± 100.3° для WT и V99A соответственно). Повышение подвижности остатка F102 приводит к тому, что комплементарный m^3C гуанин (см. dG на рис. 3, *в*, *г*), в комплексе фермента дикого типа бывший полностью вывернутым из двойной цепи ДНК, в комплексе мутантной формы фермента имеет возможность вернуться внутрь структуры ДНК, вступая в π-π-контакт с боковой цепью F102.

Замена V99A также индуцировала изменения во взаимодействии с косубстратом, которые привели к тому, что на половине общего времени траектории моделирования αKG принимает каталитически неблагоприятную конформацию. Изменения положения гидрофобных остатков V108, F124, L127 и L129 в петлях L1 и L2 приводят к переориентации аминокислотных остатков Q112 и

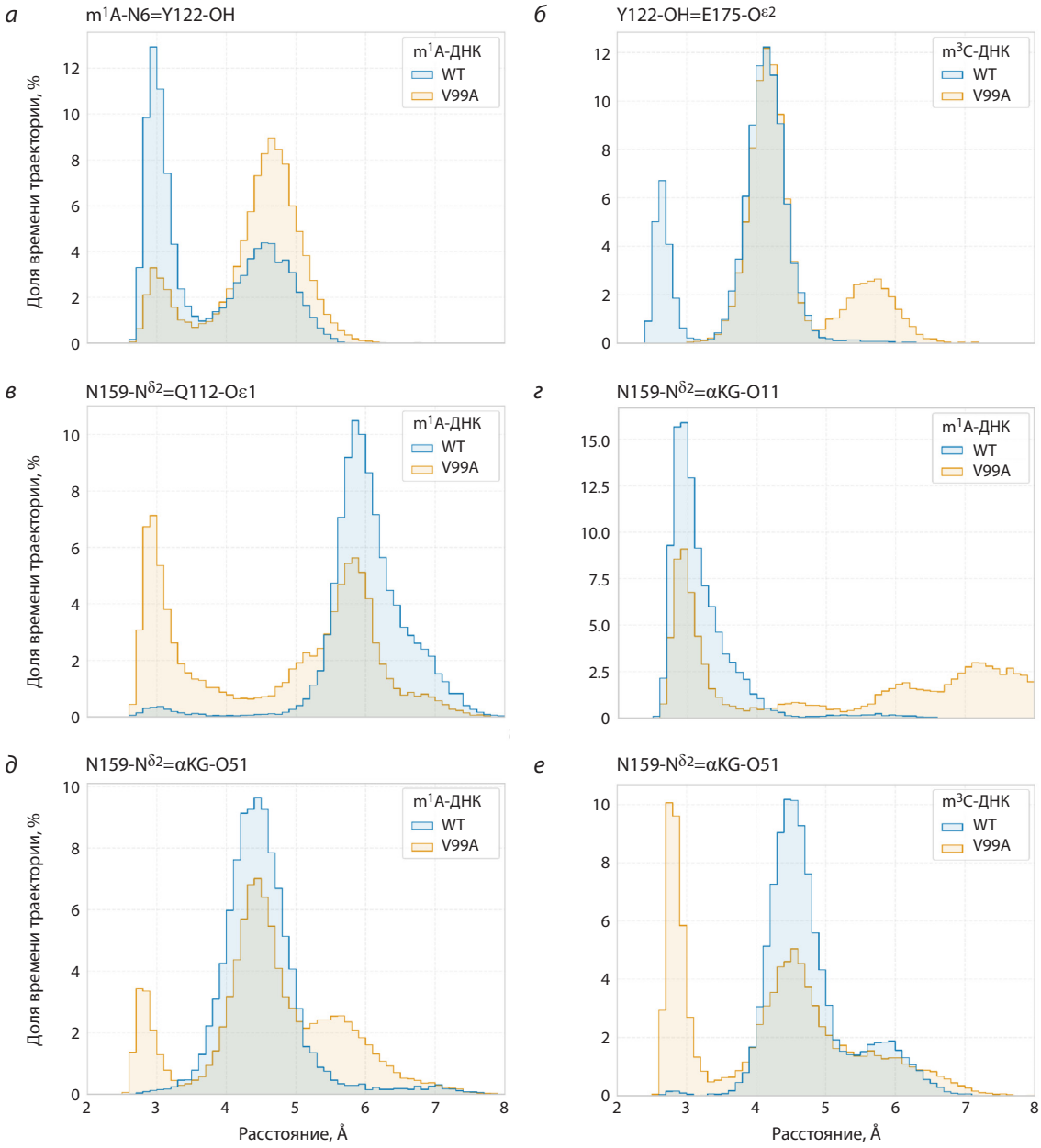


Рис. 4. Распределение расстояний между ключевыми атомами при моделировании комплексов фермента ABH2 дикого типа и его мутантной формы V99A с ДНК-субстратами.

N159. В свою очередь, в ферменте дикого типа боковая цепь N159 является одним из элементов сети контактов, поддерживающих каталитически компетентную ориентацию косубстрата, образуя водородную связь с α-карбоксильной группой αKG. Сближение боковых цепей Q112 и N159 в мутантной форме ABH2 V99A (см. рис. 4, *в*) приводит к переносу водородной связи амидной группы N159 с α-карбоксильной группы αKG (см. рис. 4, *г*) на ω-карбоксильную группу αKG (см. рис. 4, *д*, *е*), провоцируя его смещение из оптимального для катализа положения.

Таким образом, результаты моделирования позволяют предположить, что замена V99A, приводя к нарушениям связывания и субстрата, и косубстрата в активном центре фермента, должна вызвать значительное снижение активности. Эти данные хорошо согласуются с экспериментальными результатами, полученными ранее для мутантной формы V99A и выявившими значительное снижение (Monsen et al., 2010) или полную потерю (Davletgildeeva et al., 2025) каталитической активности ABH2 V99A по отношению к дцДНК-субстратам, содержащим m^1A или m^3C в качестве повреждения.

Модель фермент-субстратного комплекса ABH2 F124A с поврежденной ДНК

Для определения функциональной роли остатка F124 выполнено моделирование комплексов мутантной формы ABH2 F124A с m^1A- и m^3C-содержащими дцДНК (рис. 5). Детальный анализ изменения распределения расстояний между ключевыми атомами активного центра в случае

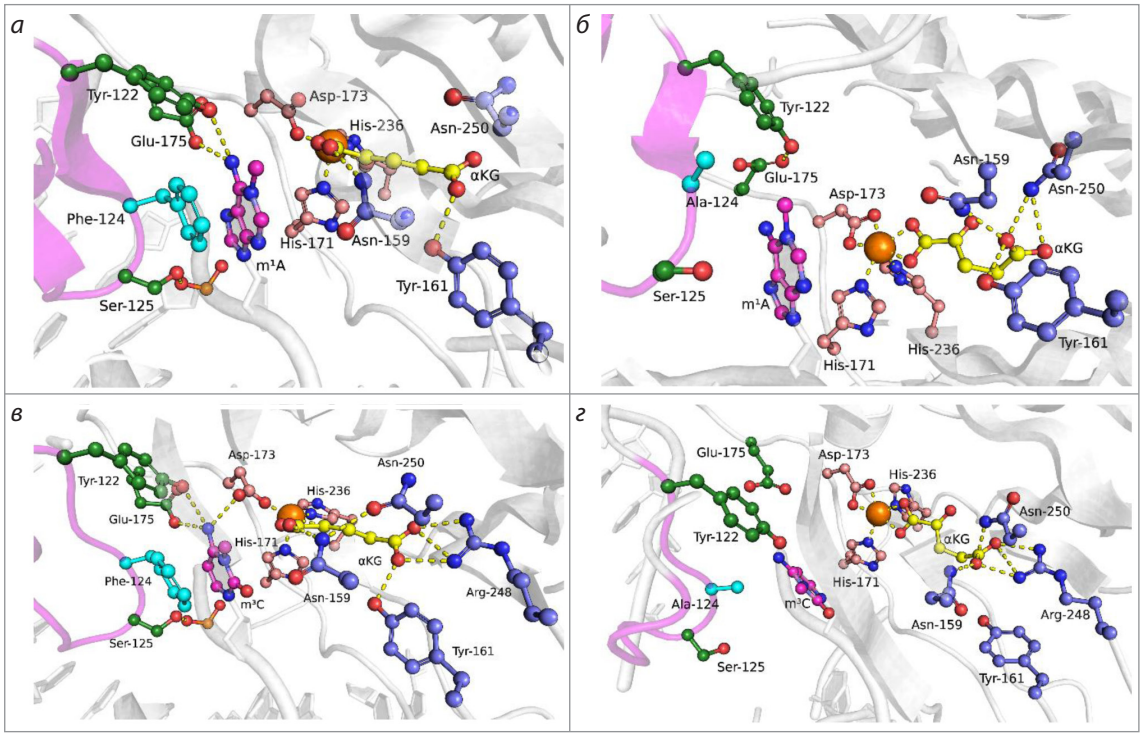


Рис. 5. Репрезентативные МД-структуры комплексов ABH2 WT с m^1A-ДНК (*а*) и m^3C-ДНК (*в*) и ABH2 F124A с m^1A-ДНК (*б*) и m^3C-ДНК (*г*).

Показаны ключевые аминокислотные остатки активного центра, поврежденное азотистое основание, αKG и ион Mn^{2+}. Петля L2 выделена розовым цветом.

замены F124A выявил дестабилизацию и вывернутого метилированного азотистого основания, и αKG в активном центре фермента.

Так, замена F124A, непосредственно приводящая к потере π-π-стэкинга между боковой цепью F124 и азотистым основанием, индуцировала поворот и смещение вывернутого основания из активного центра фермента с сопутствующей потерей водородных связей с боковыми цепями остатков Y122, D173, E175 (рис. 6, *а*, *б*). При этом утрачивается и водородная связь между гидроксильной группой остатка S125 и соответствующей фосфатной группой остова нуклеотида, отражая ухудшение контакта между петлей L2 и ДНК (см. рис. 6, *в*).

Косубстрат также теряет каталитически компетентное положение в результате перестроения сети водородных связей с участием координирующих его аминокислотных остатков. Амидная группа N159 поддерживает водородную связь преимущественно с ω-карбоксильной группой αKG вместо α-карбоксильной группы (см. рис. 6, *г*). Дестабилизация положения косубстрата отражается в изменении характера контактов между боковыми цепями остатков Y161 и R248 и ω-карбоксильной группой αKG. Если в комплексе фермента дикого типа сохраняются устойчивые водородные связи между гуанидиновой группой R248 и атомом O2 ω-карбоксильной группы αKG и между гидроксильной группой Y161 и атомом O1 ω-карбоксильной группы, то в комплексе мутантной формы ABH2 F124A происходит расширение распределения данных расстояний, свидетельствующее о дестабилизации контактов (см. рис. 6, *д*, *е*).

Результаты моделирования указывают на то, что аминокислотный остаток F124 играет важную роль в структуре активного центра фермента ABH2. Этот вывод согласуется с данными работ (Chen et al., 2010; Monsen et al., 2010), а также с результатами, полученными ранее (Davletgildeeva et al., 2025), согласно которым мутантная форма ABH2 F124A полностью утратила каталитическую активность по отношению к m^1A- и m^3C-содержащим ДНК субстратам.

Модель фермент-субстратного комплекса ABH2 S125A с поврежденной ДНК

Замена S125A в ферменте ABH2 вызывает потерю водородной связи между гидроксильной группой аминокислотного остатка и 5′-фосфатной группой поврежденного нуклеотида, что приводит к потере прямого взаимодействия петли L2 с m^1A- (рис. 7, *а*, *б*) и m^3C-ДНК (см. рис. 7, *в*, *г*). Анализ изменения расстояний между ключевыми остатками активного центра показал, что в комплексе фермента и m^1A-ДНК потеря взаимодействия петли L2 с ДНК вызывает потерю водородной связи между гидроксильной группой остатка Y122 из L2 и экзоциклической аминогруппой m^1A (рис. 8, *а*). При этом происходит сближение гуанидиновых групп остатков R110 и R172 с O3′ атомом нуклеотида вывернутого азотистого основания и O5′ атомом нуклеотида, расположенного с 5′-стороны от вывернутого азотистого основания, соответственно (см. рис. 8, *б*, *в*). Таким образом, в случае ДНК-субстрата, содержащего m^1A, замена S125A приводит к тому, что аминокислотные остатки R110 и R172 более прочно связываются с сахарофосфатным остовом ДНК.

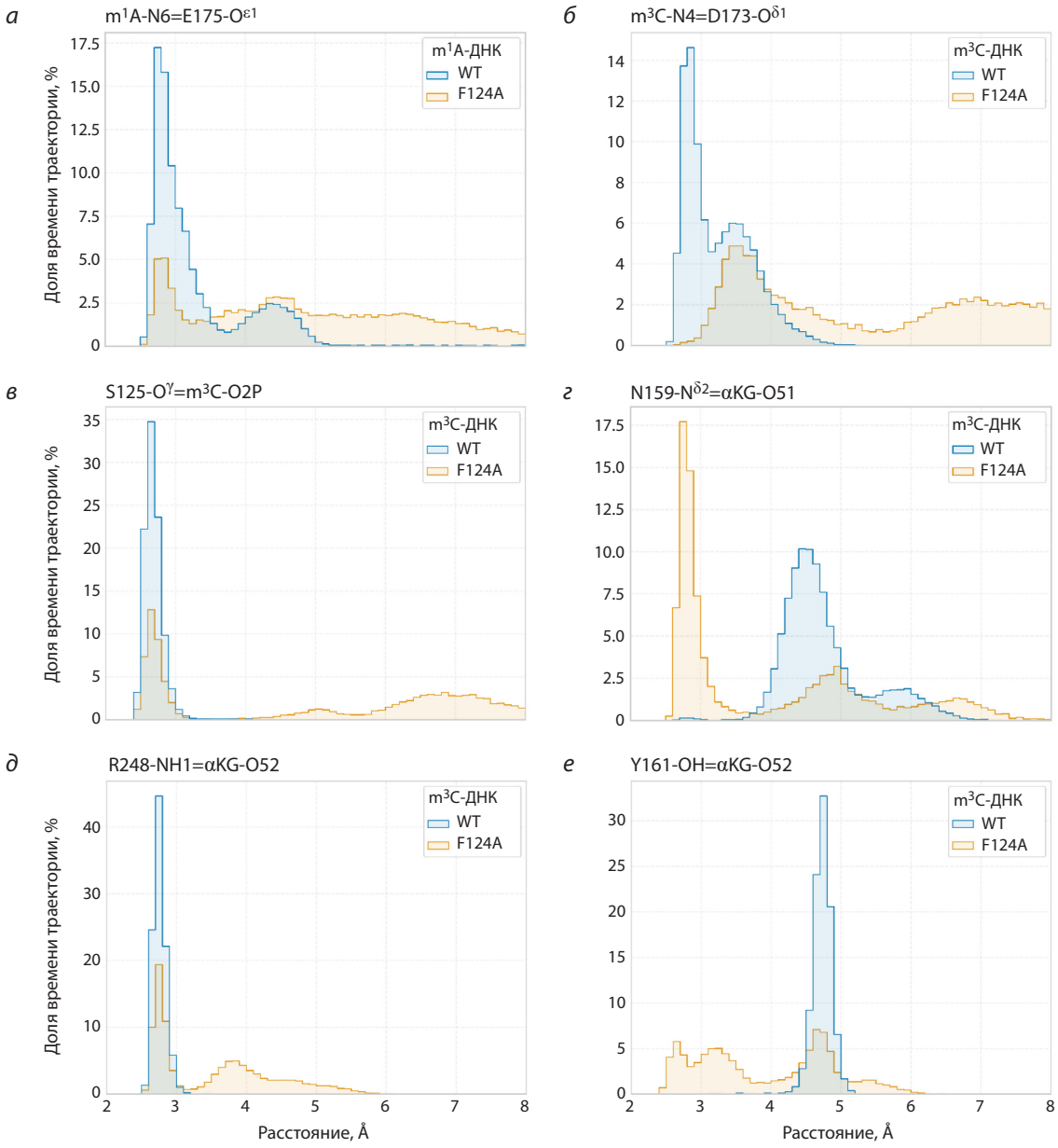


Рис. 6. Распределение расстояний между ключевыми атомами при моделировании комплексов фермента ABH2 дикого типа и его мутантной формы F124A с ДНК-субстратами.

В отличие от комплекса фермента ABH2 S125A с m^1A-ДНК, в модели комплекса с m^3C-ДНК не происходит сближения гуанидиновых групп остатков R110 и R172 с сахорофосфатным остовом (см. рис. 7, *в*, *г*). При этом по сравнению с ферментом WT в случае замены S125A снижается стабильность водородной связи между боковой цепью остатка E175 и экзоциклической аминогруппой m^3C (см. рис. 8, *г*).

Ухудшение прямого контакта с вывернутым основанием и возможная компенсаторная перестройка при замене S125A в активном центре ABH2 согласуются с результатами, полученными B. Chen с коллегами (2010), поскольку в их работе показано, что мутантная форма ABH2 S125A сохраняет каталитическую активность по отношению к дцДНК, содержащей в качестве повреждения m^1A. Однако в более поздней работе (Davletgildeeva et al., 2025) продемонстрировано, что данная замена приводит к потере каталитической активности ABH2 по отношению как к m^3C-, так к m^1A-содержащей ДНК в использованных реакционных условиях. Это позволяет предположить, что компенсаторные перестройки, возникающие, согласно данным моделирования, в структуре ABH2 при замене S125A, не способны в полной степени сохранять каталитическую активность фермента на всех типах ДНК-субстратов.

Заключение

Введение замены V99A в фермент ABH2 оказывало влияние на другие аминокислотные остатки, формирующие гидрофобную сеть, частью которой является заменяемый остаток. Это оказывало негативное влияние на функциональные петли L1 и L2, вызывая дестабилизацию их по-

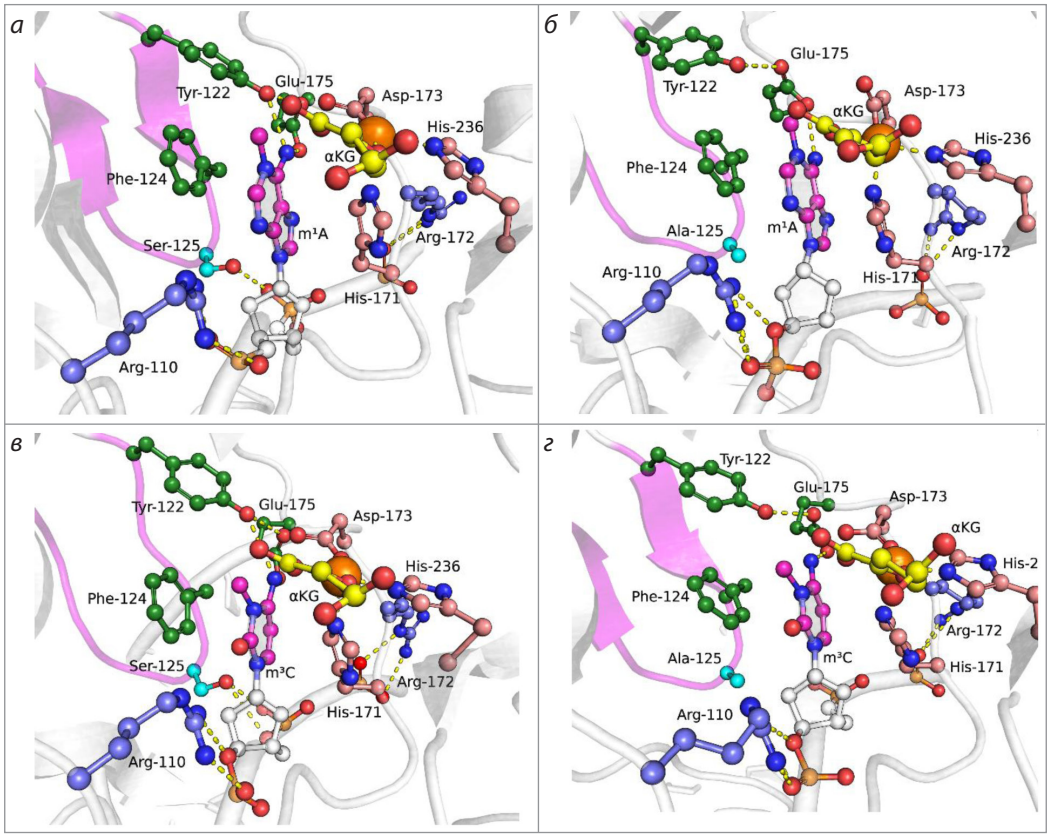


Рис. 7. Репрезентативные МД-структуры комплексов ABH2 WT с m¹A-ДНК (*а*) и m³C-ДНК (*в*) и ABH2 S125A с m¹A-ДНК (*б*) и m³C-ДНК (*г*).

Показаны ключевые аминокислотные остатки активного центра, поврежденное азотистое основание, αKG и ион Mn^{2+}. Петля L2 выделена розовым цветом.

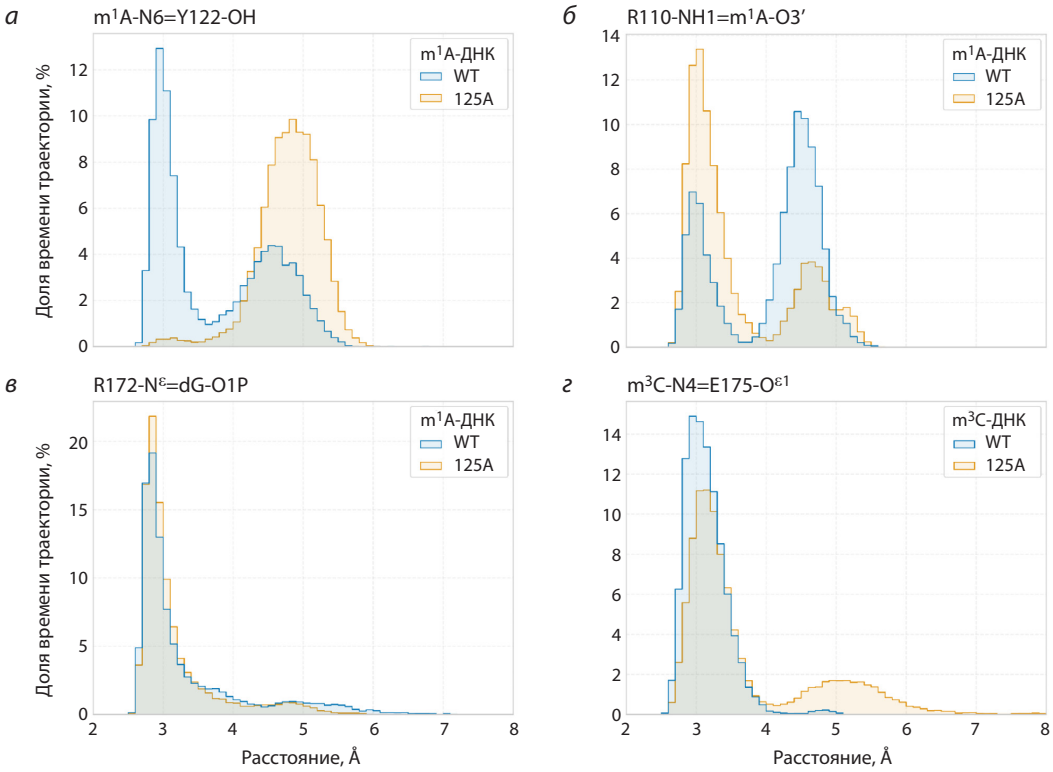


Рис. 8. Распределение расстояний между ключевыми атомами при моделировании комплексов фермента ABH2 дикого типа и его мутантной формы S125A с ДНК-субстратами.

ложения, что, в свою очередь, вызывало переориентацию или сдвиг ключевых аминокислотных остатков Y122, E175 и F102, входящих в эти петли. Кроме того, замена V99A обуславливала неблагоприятную для катализа конформацию αKG в активном центре фермента. Полученные данные подтверждают роль аминокислотного остатка V99 в качестве важного участника внутрибелковой координации, необходимой для эффективного окисления метильных групп в поврежденных основаниях ДНК ферментом ABH2.

Замена аминокислотного остатка F124, локализованного в NRL, приводила к тому, что из-за потери π-π-стэкинга с поврежденным азотистым основанием и петли L1 и L2, и само поврежденное основание значительно смещались относительно друг друга. Также данная замена была причиной изменения координации иона Fe^{2+} как за счет изменения типа координации молекулой αKG, так и возникновения дополнительной координации аминокислотным остатком D173. Полученные данные позволяют предположить чрезвычайную важность аминокислотного остатка F124 в каталитическом процессе, осуществляемом ДНК-диоксигеназой ABH2.

Замена S125A вызывала потерю прямого взаимодействия петли L2 с 5′-фосфатной группой поврежденного нуклеотида, однако, согласно данным МД-моделирования, этот контакт может быть частично компенсирован путем возникновения связей аминокислотных остатков R110 и R172 с сахарофосфатным остовом ДНК. Следует отметить, что такая компенсация контакта была обнаружена лишь в случае комплекса ABH2 S125A с m^1A-содержащим ДНК субстратом, но не в случае m^3C, что косвенно свидетельствует о более сложном механизме, отвечающем за узнавание различных повреждений в активном центре фермента.

Таким образом, полученные в настоящей работе данные МД-моделирования комплексов мутантных форм ДНК-диоксигеназы человека ABH2, содержащих аминокислотные замены V99A, F124A или S125A, с m^1A- и m^3C-содержащими ДНК субстратами указывают на важную роль всех трех аминокислотных остатков для обеспечения формирования каталитически компетентного состояния активного центра при взаимодействии с поврежденной ДНК.

Список литературы / References

Aas P.A., Otterlei M., Falnes P., Vågbø C.B., Skorpen F., Akbari M., Sundheim O., Bjørås M., Slupphaug G., Seeberg E., Krokan H.E. Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature*. 2003;421:859-863. doi 10.1038/nature01363

Abraham M.J., Murtola T., Schulz R., Páll S., Smith J.C., Hess B., Lindahl E. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1-2:19-25. doi 10.1016/j.softx.2015.06.001

Anandakrishnan R., Aguilar B., Onufriev A.V. *H*++ 3.0: automating p*K* prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res*. 2012;40:W537-W541. doi 10.1093/nar/gks375

Bayly C.I., Cieplak P., Cornell W., Kollman P.A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem*. 1993;97:10269-10280. doi 10.1021/j100142a004

Bian K., Lenz S.A.P., Tang Q., Chen F., Qi R., Jost M., Drennan C.L., Essigmann J.M., Wetmore S.D., Li D. DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine *in vitro*. *Nucleic Acids Res*. 2019;47(11):5522-5529. doi 10.1093/nar/gkz395

Bussi G., Donadio D., Parrinello M. Canonical sampling through velocity rescaling. *J Chem Phys*. 2007;126(1):014101. doi 10.1063/1.2408420

Chen B., Liu H., Sun X., Yang C.-G. Mechanistic insight into the recognition of single-stranded and double-stranded DNA substrates by ABH2 and ABH3. *Mol Biosyst*. 2010;6(11):2143-2149. doi 10.1039/c005148a

Chen B., Gan J., Yang C. The complex structures of ALKBH2 mutants cross-linked to dsDNA reveal the conformational swing of β-hairpin. *Sci China Chem*. 2014;57:307-313. doi 10.1007/s11426-013-5029-z

Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M., Ferguson D.M., Spellmeyer D.C., Fox T., Caldwell J.W., Kollman P.A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*. 1995;117(19):5179-5197. doi 10.1021/ja00124a002

Davletgildeeva A.T., Tyugashev T.E., Zhao M., Kuznetsov N.A., Ishchenko A.A., Saparbaev M., Kuznetsova A.A. Individual contributions of amido acid residues Tyr122, Ile168, and Asp173 to the activity and substrate specificity of human DNA dioxygenase ABH2. *Cells*. 2023;12(14):1839. doi 10.3390/cells12141839

Davletgildeeva A.T., Tyugashev T.E., Zhao M., Ishchenko A.A., Saparbaev M., Kuznetsov N.A. Role of individual amino acid residues directly involved in damage recognition in active demethylation by ABH2 dioxygenase. *Int J Mol Sci*. 2025;26:6912. doi 10.3390/ijms26146912

Duncan T., Trewick S.C., Koivisto P., Bates P.A., Lindahl T., Sedgwick B. Reversal of DNA alkylation damage by two human dioxygenases. *Proc Natl Acad Sci USA*. 2002;99(26):16660-16665. doi 10.1073/pnas.262589799

Essmann U., Perera L., Berkowitz M.L., Darden T., Lee H., Pedersen L.G. A smooth particle mesh Ewald method. *J Chem Phys*. 1995;103:8577-8593. doi 10.1063/1.470117

Falnes P. Repair of 3-methylthymine and 1-methylguanine lesions by bacterial and human AlkB proteins. *Nucleic Acids Res*. 2004;32:6260-6267. doi 10.1093/nar/gkh964

Giri N.C., Sun H., Chen H., Costa M., Maroney M.J. X-ray absorption spectroscopy structural investigation of early intermediates in the mechanism of DNA repair by human ABH2. *Biochemistry*. 2011;50(22):5067-5076. doi 10.1021/bi101668x

Hess B., Bekker H., Berendsen H.J.C., Fraaije J.G.E.M. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*. 1997;18(12):1463-1472. doi 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

Jiang Y., Zhang H., Tan T. Rational design of methodology-independent metal parameters using a nonbonded dummy model. *J Chem Theory Comput*. 2016;12(7):3250-3260. doi 10.1021/acs.jctc.6b00223

Jorgensen W.L., Chandrasekhar J., Madura J.D., Impey R.W., Klein M.L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79(2):926-935. doi 10.1063/1.445869

Joung I.S., Cheatham T.E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B*. 2008;112:9020-9041. doi 10.1021/jp8001614

Kuznetsov N.A., Kanazhevskaya L.Y., Fedorova O.S. DNA demethylation in the processes of repair and epigenetic regulation performed by 2-ketoglutarate-dependent DNA dioxygenases. *Int J Mol Sci*. 2021;22:10540. doi 10.3390/ijms221910540

Lee D.H., Jin S.G., Cai S., Chen Y., Pfeifer G.P., O'Connor T.R. Repair of methylation damage in DNA and RNA by mammalian AlkB homologues. *J Biol Chem*. 2005;280(47):39448-39459. doi 10.1074/jbc.M509881200

Lenz S.A.P., Li D., Wetmore S.D. Insights into the direct oxidative repair of etheno lesions: MD and QM/MM study on the substrate scope of ALKBH2 and AlkB. *DNA Repair* (*Amst*). 2020;96:102944. doi 10.1016/j.dnarep.2020.102944

Li P., Gao S., Wang L., Yu F., Li J., Wang C., Li J., Wong J. ABH2 couples regulation of ribosomal DNA transcription with DNA alkylation repair. *Cell Rep*. 2013;4:817-829. doi 10.1016/j.celrep.2013.07.027

Maier J.A., Martinez C., Kasavajhala K., Wickstrom L., Hauser K.E., Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*. 2015;11:3696-3713. doi 10.1021/acs.jctc.5b00255

McGibbon R.T., Beauchamp K.A., Harrigan M.P., Klein C., Swails J.M., Hernández C.X., Schwantes C.R., Wang L.-P., Lane T.J., Pande V.S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys J.* 2015;109:1528-1532. doi 10.1016/j.bpj.2015.08.015

Monsen V.T., Sundheim O., Aas P.A., Westbye M.P., Sousa M.M.L., Slupphaug G., Krokan H.E. Divergent β-hairpins determine double-strand versus single-strand substrate recognition of human AlkB-homologues 2 and 3. *Nucleic Acids Res*. 2010;38:6447-6455. doi 10.1093/nar/gkq518

Müller T.A., Hausinger R.P. AlkB and its homologues. DNA repair and beyond. In: Schofield C., Hausinger R. (Eds) 2-Oxoglutarate-Dependent Oxygenases. Royal Society Chemistry. 2015;246-262. doi 10.1039/9781782621959-00246

Ougland R., Rognes T., Klungland A., Larsen E. Non-homologous functions of the AlkB homologs. *J Mol Cell Biol.* 2015;7(6):494-504. doi 10.1093/jmcb/mjv029

Parrinello M., Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys*. 1981;52(12):7182-7190. doi 10.1063/1.328693

Ringvoll J., Nordstrand L.M., Vagbo C.B., Talstad V., Reite K., Aas P.A., Lauritzen K.H., Liabakk N.B., Bjork A., Doughty R.W., Falnes P.O., Krokan H.E., Klungland A. Repair deficient mice reveal mABH2 as the primary oxidative demethylase for repairing 1meA and 3meC lesions in DNA. *Embo J*. 2006;25:2189-2198. doi 10.1038/sj.emboj.7601109

Ringvoll J., Moen M.N., Nordstrand L.M., Meira L.B., Pang B., Bekkelund A., Dedon P.C., Bjelland S., Samson L.D., Falnes P.Ø., Klungland A. AlkB homologue 2 – mediated repair of ethenoadenine lesions in mammalian DNA. *Cancer Res*. 2008;68(11):4142-4149. doi 10.1158/0008-5472.CAN-08-0796

Šali A., Blundell T.L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234(3):779-815. doi 10.1006/jmbi.1993.1626

Sall S.O., Berens J.T.P., Molinier J. DNA damage and DNA methylation. In: Jasiulionis M.G. (Ed.) Epigenetics and DNA Damage. Academic Press, 2022;3-16. doi 10.1016/B978-0-323-91081-1.00005-4

Sousa da Silva A.W., Vranken W.F. ACPYPE – AnteChamber PYthon Parser interfacE. *BMC Res Notes*. 2012;5:367. doi 10.1186/1756-0500-5-367

Travers A., Muskhelishvili G. DNA structure and function. *FEBS J*. 2015;282(12):2279-2295. doi 10.1111/febs.13307

Vanquelef E., Simon S., Marquant G., Garcia E., Klimerak G., Delepine J.C., Cieplak P., Dupradeau F.-Y. R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res*. 2011;39:W511-W517. doi 10.1093/nar/gkr288

Waheed S.O., Ramanan R., Chaturvedi S.S., Lehnert N., Schofield C.J., Christov C.Z., Karabencheva-Christova T.G. Role of structural dynamics in selectivity and mechanism of non-heme Fe(II) and 2-oxoglutarate-dependent oxygenases involved in DNA repair. *ACS Cent Sci*. 2020;6(5):795-814. doi 10.1021/acscentsci.0c00312

Wang J., Wolf R.M., Caldwell J.W., Kollman P.A., Case D.A. Development and testing of a general amber force field. *J Comput Chem*. 2004;25:1157-1174. doi 10.1002/jcc.20035

Wang J., Wang W., Kollman P.A., Case D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model*. 2006;25(2):247-260. doi 10.1016/j.jmgm.2005.12.005

Wilson D.L., Beharry A.A., Srivastava A., O'Connor T.R., Kool E.T. Fluorescence probes for ALKBH2 allow the measurement of DNA alkylation repair and drug resistance responses. *Angew Chem Int Ed Engl*. 2018;57(39):12896-12900. doi 10.1002/anie.201807593

Xu B., Liu D., Wang Z., Tian R., Zuo Y. Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. *Cell Mol Life Sci.* 2021;78:129-141. doi 10.1007/s00018-020-03594-9

Yang C.G., Yi C., Duguid E.M., Sullivan C.T., Jian X., Rice P.A., He C. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature*. 2008;452:961-965. doi 10.1038/nature06889

Yang C.G., Garcia K., He C. Damage detection and base flipping in direct DNA alkylation repair. *Chembiochem*. 2009;10(3):417-423. doi 10.1002/cbic.200800580

Yi C., He C. DNA repair by reversal of DNA damage. *Cold Spring Harb Perspect Biol*. 2013;5:a012575. doi 10.1101/cshperspect.a012575

Yi C., Yang C.G., He C. A non-heme iron-mediated chemical demethylation in DNA and RNA. *Acc Chem Res*. 2009;42(4):519-529. doi 10.1021/ar800178j

Yi C., Chen B., Qi B., Zhang W., Jia G., Zhang L., Li C.J., Dinner A.R., Yang C.-G., He C. Duplex interrogation by a direct DNA repair protein in search of base damage. *Nat Struct Mol Biol*. 2012;19:671-676. doi 10.1038/nsmb.2320

Zgarbová M., Otyepka M., Sponer J., Mládek A., Banáš P., Cheatham T.E., Jurečka P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theory Comput*. 2011;7(9):2886-2902. doi 10.1021/ct200162x

Zgarbová M., Šponer J., Otyepka M., Cheatham T.E., Galindo-Murillo R., Jurečka P. Refinement of the sugar-phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *J Chem Theory Comput*. 2015;11(12):5723-5736. doi 10.1021/acs.jctc.5b00716

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 06.06.2025. После доработки 12.09.2025. Принята к публикации 12.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1073-1083

doi 10.18699/vjgb-25-112

Структурные основы влияния фосфорамидной *N*-бензимидазольной группы на эффективность удлинения модифицированного праймера Taq ДНК-полимеразой

А.А. Бердюгин [1, 2], В.М. Голышев [1, 2], А.А. Ломзов [1, 2]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

lomzov@1bio.ru

Аннотация. Недавно нами был предложен новый класс производных нуклеиновых кислот – фосфорамидные бензоазольные олигонуклеотиды. В них один из немостиковых атомов кислорода замещен на фосфорамидную *N*-бензоазольную группу: бензимидазольную, диметилбензимидазольную, бензоксазольную или бензотиазольную. Изучение свойств таких производных показало, что их применение в ПЦР увеличивает специфичность и селективность анализа. Данное исследование посвящено изучению влияния фосфорамидной *N*-бензимидазольной модификации ДНК-праймеров на эффективность их удлинения Taq ДНК-полимеразой при помощи метода молекулярной динамики. Мы рассматривали совершенные комплексы нуклеиновых кислот с модификациями в положениях с первого по шестое считая от 3′-конца праймера. Ранее было показано, что степень подавления элонгации зависит от положения модификации: чем ближе к 3′-концу, тем сильнее ингибирование, а максимальное подавление наблюдается при модификации в первом положении, особенно в несовершенных комплексах. Кроме того, в экспериментах наблюдались продукты неполного удлинения праймеров с модификацией в четвертом положении. Проведенные компьютерное моделирование и анализ позволили выявить молекулярные механизмы взаимодействия модифицированных праймеров с ферментом, включая стерические препятствия для продвижения полимеразы по модифицированной цепи и локальные нарушения структуры ДНК, которые объясняют наблюдаемые экспериментально закономерности. Установлено, что как различные стереоизомеры фосфорамидных групп, так и конформеры фосфорамидной *N*-бензимидазольной группы по-разному влияют на структуру фермент-субстратного комплекса и эффективность взаимодействия Taq ДНК-полимеразы с модифицированным ДНК комплексом. Модификация первого и второго межнуклеозидного фосфатного остатка с 3′-конца праймера в наибольшей степени возмущает структуру белково-нуклеинового комплекса, а при расположении модификации в четвертом фосфатном остатке *N*-бензимидазольная модификация располагается в кармане фермента. Полученные результаты открывают перспективы для рационального конструирования специфичных, обладающими заранее заданными свойствами ДНК праймеров с модифицированными *N*-бензимидазольными межнуклеотидными звеньями для использования в ПЦР диагностике.

Ключевые слова: *N*-бензимидазольные олигонуклеотиды; ФАО; молекулярная динамика; структура; Taq ДНК-полимераза; молекулярная диагностика

Для цитирования: Бердюгин А.А., Голышев В.М., Ломзов А.А. Структурные основы влияния фосфорамидной *N*-бензимидазольной группы на эффективность удлинения модифицированного праймера Taq ДНК-полимеразой. *Вавиловский журнал генетики и селекции*. 2025;29(7):1073-1083. doi 10.18699/vjgb-25-112

Финансирование. Исследование выполнено при финансовой поддержке Российского научного фонда (проект № 23-74-01116, https://rscf.ru/project/23-74-01116/) в части построения и первичного анализа модельных систем и проекта государственного задания ИХБФМ СО РАН № 123021600208-7 в части проведения молекулярно-динамического моделирования и анализа полученных данных.

Structural basis of the phosphoramidate *N*-benzimidazole group's influence on modified primer extension efficiency by Taq DNA polymerase

A.A. Berdugin [1, 2], V.M. Golyshev [1, 2], A.A. Lomzov [1, 2]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia

lomzov@1bio.ru

Abstract. We recently proposed a novel class of nucleic acid derivatives – phosphoramidate benzoazole oligonucleotides (PABAOs). In these compounds, one of the non-bridging oxygen atoms is replaced by a phosphoramidate *N*-benzoazole group, such as benzimidazole, dimethylbenzimidazole, benzoxazole, or benzothiazole. Studies of the properties of these derivatives have shown that their use in PCR enhances the specificity and selectivity of the analysis. The

© Бердюгин А.А., Голышев В.М., Ломзов А.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

study investigates the effect of phosphoramide *N*-benzimidazole modification of DNA primers on their elongation by Taq DNA polymerase using molecular dynamics simulations. We examined perfectly matched primer-template complexes with modifications at positions one through six from the 3′-end of the primer. Prior experimental work demonstrated that the degree of elongation suppression depends on the modification position: the closer to the 3′-end, the stronger the inhibition, with maximal suppression observed for the first position, especially in mismatched complexes. Furthermore, incomplete elongation products were experimentally observed for primers modified at the fourth position. Our molecular dynamics simulations and subsequent analysis revealed the molecular mechanisms underlying the interaction of modified primers with the enzyme. These include steric hindrance that impedes polymerase progression along the modified strand and local distortions in the DNA structure, which explain the experimentally observed trends. We established that both different stereoisomers of the phosphoramidate groups and conformers of the phosphoramidate *N*-benzimidazole moiety differentially affect the structure of the enzyme-substrate complex and the efficiency of Taq DNA polymerase interaction with the modified DNA complex. Modification of the first and second internucleoside phosphate from the 3′-end of the primer causes the most significant perturbation to the structure of the protein-nucleic acid complex. When the modification is located at the fourth phosphate group, the *N*-benzimidazole moiety occupies a specific pocket of the enzyme. These findings provide a foundation for the rational design of specific DNA primers bearing modified *N*-benzimidazole moieties with tailored properties for use in PCR diagnostics.

Key words: *N*-benzimidazole oligonucleotides; PABAO; molecular dynamics; structure; Taq DNA polymerase; molecular diagnostics

For citation: Berdugin A.A., Golyshev V.M., Lomzov A.A. Structural basis of the phosphoramidate *N*-benzimidazole group's influence on modified primer extension efficiency by Taq DNA polymerase. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1073-1083. doi 10.18699/vjgb-25-112

Введение

ДНК-зависимая ДНК-полимераза I бактерии *Thermus aquaticus* (ДНК-полимераза Taq) является широко распространенным ферментом, используемым для амплификации нуклеиновых кислот методом ПЦР в разных приложениях. Она обладает ДНК-полимеразной и 5′→3′ экзонуклеазной активностями, но не обладает корректирующей 3′→5′ экзонуклеазной активностью (Terpe, 2013). Данный фермент широко применяют для выявления нуклеиновых кислот и однонуклеотидных замен (точечных мутаций) в них при диагностике различных заболеваний с помощью разных методов ПЦР анализа, например ПЦР в реальном времени, аллель-специфичной ПЦР, цифровой ПЦР (Kalendar et al., 2022; Starza et al., 2022). Метод аллель-специфичной ПЦР основан на ингибировании реакции удлинения праймеров, которые образуют с матричной цепью дуплексы с одним или несколькими мисматчами в области 3′-конца короткой цепи (Rejali et al., 2018). Часто наличие одиночной замены, нарушающей его полную комплементарность с ДНК, не обеспечивает достаточной специфичности выявления полиморфизмов. Для ее повышения в ДНК вводят дополнительные однонуклеотидные несоответствия и/или дополнительные модификации в структуру праймера. Последние вводят как в основание, так и в рибозофосфатный остов и располагают вблизи 3′-конца (Kutyavin, 2011; Ishige et al., 2018; Chubarov et al., 2023). В частности, замещение немостикового атома кислорода в фосфодиэфирном остове влияет на термодинамическую стабильность дуплекса праймер/матрица, а также на координацию концевой 3′-ОН группы в активном центре фермента. Например, введение фосфотиоатной модификации в последний или предпоследний межнуклеозидный фосфат с 3′-конца праймера незначительно снижает эффективность элонгации (5–15 %), но в то же время увеличивает специфичность амплификации (Di Giusto, King, 2003). Введение фосфорилгуанидиновой модификации в структуру праймеров также приводит к изменению эффективности и селективности выявления целевых последовательностей НК (Chubarov et al., 2020).

Недавно в Институте химической биологии и фундаментальной медицины СО РАН был разработан новый класс производных нуклеиновых кислот – фосфорамидные бензоазольные олигонуклеотиды (ФАО), в которых немостиковый атом кислорода фосфатного остатка замещен на *N*-бензоазольную группу (бензимидазольную, бензоксазольную или бензотиазольную) (рис. 1). ФАО можно получать автоматическим твердофазным фосфитамидным методом синтеза (Vasilyeva et al., 2023).

В настоящее время ведется исследование физико-химических свойств ряда *N*-бензоазольных производных (Golyshev et al., 2024; Yushin et al., 2024; Novgorodtseva et al., 2025) и показана возможность их применения в качестве праймеров в ПЦР, в том числе в аллель-специфичной ПЦР (Chubarov et al., 2024). Нами изучена эффективность удлинения тринадцатизвенных праймеров, со-

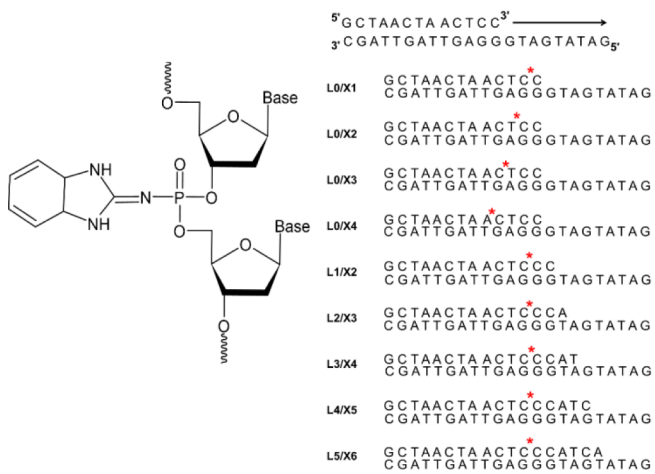


Рис. 1. Структура динуклеотидного шага ФАО, содержащего *N*-бензимидазольную группу, и модельные системы, использованные в исследования.

Положение фосфорамидной *N*-бензимидазольной группы обозначено красной звездочкой.

держащих *N*-бензимидазольную модификацию, по двадцатидвухзвенной ДНК матрице с использованием Taq ДНК-полимеразы (Golyshev et al., 2025). Введение модификации в первый или второй межнуклеозидный фосфат с 3′-конца праймера для совершенных комплексов позволяет им удлиняться до полноразмерного продукта с эффективностью ~50 % и значительно сокращает выход полноразмерного продукта для комплексов с однонуклеотидными несоответствиями в предконцевой паре от 3′-конца короткой цепи. Введение модификации в третье положение типично снижает выход полноразмерного комплекса в наименьшей степени. Кроме того, для всех совершенных комплексов с модификациями отчетливо наблюдался продукт промежуточного удлинения, соответствующий частично удлиненному праймеру, содержащему модификацию в 4-й позиции от 3′-конца праймера.

В данной работе для объяснения закономерностей, наблюдаемых в экспериментах по удлинению ФАО праймеров Taq ДНК-полимеразой, проведено детальное изучение влияния фосфорамидной *N*-бензимидазольной группы в различных положениях праймерной цепи на структуру и динамику фермент-субстратного комплекса с использованием методов молекулярной динамики (МД) и анализа. Для этого мы сконструировали молекулярные модели и провели МД моделирование нативного и набора модифицированных НК субстратов, содержащих одиночные модификации с 1-го по 6-е положение межнуклеозидных фосфатных остатков с 3′-конца праймера, а также их комплексов с Taq ДНК-полимеразой. Полученные данные коррелируют с экспериментальными результатами и объясняют наблюдающиеся *in vitro* эффекты.

Методы

Построение моделей. Структуру комплекса Taq полимеразы с ДНК строили на основании экспериментально полученной структуры белка с идентификатором PDB ID: 1QTM следующим образом. Структуру белка с нуклеозидтрифосфатом и ионами магния брали из этой модели. Структуру ДНК комплекса матрицы с праймером получали путем построения белково-нуклеинового комплекса с применением программы AlphaFold3 (Abramson et al., 2024). В качестве входных данных для расчетов подавали аминокислотную последовательность ДНК-полимеразы I бактерии *Thermus aquaticus* (UniProt ID: P19821), а также нуклеотидные последовательности ДНК матрицы и неудлиненных или удлиненных праймеров, трифосфат и два иона магния. Рассчитанную структуру выравнивали по аминокислотным остаткам и по Cα атомам со структурой 1QTM. Далее в модели 1QTM заменяли нуклеиновые кислоты на построенные с помощью AlphaFold3. Для каждой построенной модели нуклеозидтрифосфат заменяли на комплементарный нуклеотиду в матричной цепи.

Поскольку *N*-бензимидазольная модификация в большинстве случаев требовала дополнительного пространства для геометрического размещения внутри комплекса ДНК/Taq полимераза, использовали библиотеки ротамеров аминокислотных остатков (Shapovalov, Dunbrack, 2011) в программе UCSF Chimera (Pettersen et al., 2004) для снижения пересечения ван-дер-ваальсовых радиусов атомов белка и модификации.

Заряды аминокислотных остатков белка для каждого из комплексов рассчитывали с помощью программы pdb2pqr версии 3.7.1 (Unni et al., 2011). Значение pH устанавливали равным 8.3, для того чтобы условия моделирования соответствовали условиям удлинения праймера (Golyshev et al., 2025). В результате некоторые комплексы имели различающиеся протонированные и депротонированные остатки. Семь из 36 комплексов демонстрировали различия в состоянии протонирования следующих остатков. В комплексах L0/X2/R1, L0/X2/R2, L0/X3/R1, L0/X3/R2 и L1/X2/R2 (обозначения см. ниже) четыре аминокислотных остатка – LYS540, ASP610, LYS663 и ASP785 – находились в протонированной форме. В комплексах L0/X4/R1 и L0/X4/R2 остатки LYS663, LYS762 и GLU786 также были протонированы.

Комплекс праймер/матрица получали из структуры белково-нуклеинового комплекса путем удаления всех остатков, кроме принадлежащих ДНК.

Молекулярно-динамическое моделирование. Исследование структуры комплексов нативных и модифицированных ДНК с Taq ДНК-полимеразой проводили методами молекулярно-динамического (МД) моделирования и анализа, используя программный пакет AMBER20 (Case et al., 2020) с параллельными вычислениями на центральных процессорах (CPU) и графических ускорителях (GPU) с архитектурой CUDA. Все МД расчеты выполняли с использованием силового поля ff19SB (Tian et al., 2020) для Taq полимеразы, OL21 (Zgarbová et al., 2021) для ДНК, параметрами gaff2 для модифицированных *N*-бензимидазолом остатков фосфорной кислоты, для ионов магния и натрия – описанные в работе (Li Z. et al., 2020). Данные силовые поля считаются наиболее современными и рекомендованы разработчиками программного пакета AMBER для проведения достоверного моделирования. Из работы (Meagher et al., 2003) брали параметры для нуклеозидтрифосфатов, которые являются единственными опубликованными для семейства силовых полей AMBER.

Протокол МД-моделирования. Начальные модели релаксировали в неявном растворителе (saltcon = 0.10, igb = 1, T = 1 K) методом сопряженных градиентов в течение 2500 шагов. Далее помещали в водную оболочку (модель воды OPC (Izadi et al., 2014), усеченная октаэдрическая ячейка, расстояние от молекулы до грани ячейки 14 Å) с добавлением ионов Na^+ для нейтрализации заряда периодической ячейки. После этого проводили минимизацию системы с ограничениями на положение атомов белка (позиционные ограничения с потенциалом 1.0 ккал/(моль·Å)) в течение 10000 шагов, из которых первые 200 шагов выполняли методом наискорейшего спуска. Затем комплексы нагревали от 0 до 300 K в течение 2 нс и стабилизировали давление с помощью баростата Монте-Карло в течение 1 нс. Для контроля температуры использовали динамику Ланжевена (ntt = 3, gamma_ln = 1.0). Шаг интегрирования уравнений движения выбирали равным 2 фс. На следующем этапе сольватированные комплексы минимизировали без наложения ограничений в течение 10000 шагов (200 шагов – метод наискорейшего спуска). И на финальном этапе проводили МД моделирование в течение 100 нс с параметрами, аналогичными этапу на-

гревания, но без наложения позиционных ограничений на атомы модельной системы.

Результаты МД моделирования анализировали в программе cpptraj из пакета AMBER20 (Roe, Cheatham, 2013). Наиболее представленные 10 структур в каждой траектории находили методом иерархического кластерного анализа.

Рисунки структур готовили с помощью программы UCSF Chimera версии 1.15 (Pettersen et al., 2004).

Результаты

Выбор и построение молекулярных моделей

Изучение структуры и динамики комплексов ФАО с Taq ДНК-полимеразой проводили с использованием набора модельных комплексов. За основу был взят ДНК комплекс праймера 5′-GCTAACTAACTCC-3′ с матрицей 5′-GATATGATGGGAGTTAGTTAGC-3′, удлинение праймера в котором было изучено нами экспериментально ранее (Golyshev et al., 2025). Было показано, что введение бензоазольных модификаций в различные положения праймера влияет на эффективность и специфичность его удлинения. В рамках данной работы выполнено МД моделирование набора белково-нуклеиновых комплексов, а также отдельных ДНК комплексов. Были изучены как нативные ДНК комплексы, так и комплексы с единичными *N*-бензимидазольными модификациями, введенными в межнуклеотидную фосфатную группу, находящуюся с 1-го по 6-е положение от 3′-конца праймера. Для оценки влияния удлинения праймера и получения более достоверных данных рассматривали олигомеры с неудлиненной праймерной цепью с модификациями в положениях с 1-го по 4-е. Также исследовали системы с удлинением праймера от одного до пяти нуклеотидов, у которого до добавления нуклеотидов модификация находилась в 1-м положении, что соответствует расположению модификации во 2–6-м положениях от 3′-конца. Последовательности модельных олигонуклеотидных комплексов и их обозначения приведены на рис. 1.

Построение модели проводили на основании модели с идентификационным номером PDB (Protein Data Bank, PDB ID) 1QTM, как описано в разделе Методы. Данная модель представляет собой фрагмент ДНК-полимеразы I бактерии *Thermus aquaticus* без эндонуклеазного домена в закрытом состоянии с ddNTP и Mg^{2+}, что, во-первых, является единственной трехмерной структурой ДНК-полимеразы I бактерии *T. aquaticus* в таком состоянии из описанных в литературе, а во-вторых, в нем отсутствует эндонуклеазный домен. Модификацию в структуру праймера вводили путем замены нативного фосфатного остатка на фосфорамидный, содержащий *N*-бензимидазольный остаток (см. рис. 1). Мы рассматривали два стереоизомера фосфорамидного остатка – Sp и Rp.

Анализ построенных молекулярных моделей модифицированной ДНК с Taq полимеразой выявил возможность двух положений бензимидазольной группы для каждого из стереоизомеров. Они соответствуют двугранному углу OP—P—N—C около −100 или +100°. Предварительное моделирование белково-нуклеиновых комплексов показало отсутствие переходов между ними. Поэтому мы изучали эти конформеры (ротамеры) для каждой из исследованных структур. Для модельных ДНК комплексов мы использовали следующие обозначения: L*i*/X*j*/R*k* или L*i*/X*j*/S*k*, где $i = 0–5$ – число нуклеотидов, на которое удлинен праймер; $j = 1–6$ – положение межнуклеозидного фосфатного остатка с 3′ стороны праймера, в которое введена модификация; $k = 1, 2$ – номер конформера для каждого из изомеров Rp или Sp. Для конформеров R1 и S2 двугранный угол, образованный атомами OP2—P—N—C и OP1—P—N—C, составлял −100°, для R2 и S1 – +100°. При этом пространственная ориентация изомеров для структур R1 и S1 соответствует ориентации бензоазольного кольца модифицированного остатка в сторону от большой бороздки, а R2 и S2 – в сторону малой (рис. 2). Для моделирования

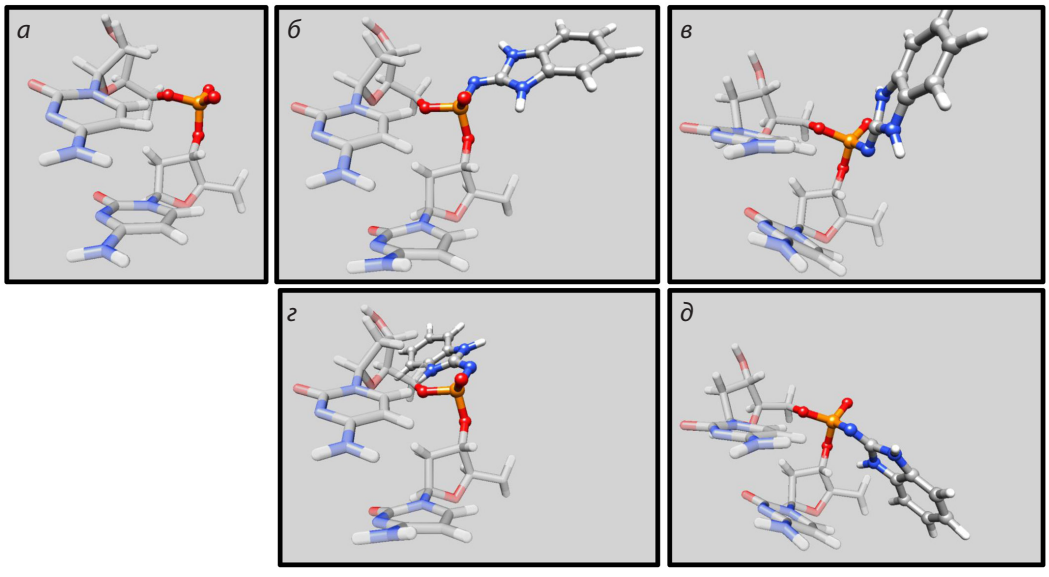


Рис. 2. Пространственная структура динуклеотидных шагов ДНК: нативного (*а*) и модифицированных для изученных стереоизомеров и конформеров R1 (*б*), S1 (*в*), R2 (*г*) и S2 (*д*).

было построено 36 комплексов с модифицированной ДНК и три – с нативной: неудлиненный и два удлиненных на 3 и 5 нт (L0, L3 и L5). Также моделирование было проведено для всех ДНК из этих моделей.

При построении комплексов белка с L3/X4/S2, L0/X4/S2, L0/X2/S2 и L4/X5/S2 наблюдается значительное стерическое перекрывание модифицированного остатка ДНК с белком, что не позволяет провести моделирование, или же при моделировании в течение первых нескольких наносекунд осуществляется переход S2 конформера в S1. В случае изомера S2 мы брали релаксированную структуру комплекса с изомером S1 и выполняли МД моделирование длиной 25 нс с наложением потенциала на двугранный угол OP1—P—N—C с возрастающим значением потенциала (0–0.2 нс, возрастание вклада от потенциала от 0 до 1, плоскодонный потенциал со стенками от −130.0 до −125.0 и от −115.0 до −110.0 с константой 200.0 ккал/моль/рад). После такой релаксации комплекса ротамер модифицированного остатка принимал необходимое положение в комплексе. Далее получали 100 нс МД траекторию, которую анализировали таким же образом, как и остальные.

Анализ конформационной подвижности

Устойчивость структуры белково-нуклеинового комплекса

В ходе МД моделирования структура белка для некоторых моделей претерпевала перестройки, что хорошо видно по увеличению величин среднеквадратичного отклонения (СКО) для белков (рис. S1)[1]. В полученных траекториях СКО для белка варьировало достаточно сильно на первых 50 нс МД траектории. Для достоверного анализа мы продлили траектории таких комплексов на равновесном участке дополнительно на 50 нс. СКО для 150 нс траекторий таких комплексов приведены на рис. S1. Последние 50 нс были взяты для анализа.

Анализ МД траекторий показал, что одноцепочечный фрагмент матричной цепи оказывается чрезвычайно подвижным при МД моделировании и типично не занимает каких-либо предпочтительных положений. Поэтому при анализе структурированных участков белково-нуклеинового комплекса он был исключен из рассмотрения. На рис. S2 показана зависимость СКО вдоль траектории для всех исследованных комплексов. Видно, что на 50 нс участке траектории все структуры стабильны, о чем свидетельствуют постоянство величин СКО после кратковременного роста в первые 1–5 нс, а также малые значения флуктуаций графиков. Среднее значение СКО составляет 2.63 ± 0.29 Å, среднее стандартное отклонение вдоль траектории – 0.39 ± 0.11 Å.

Устойчивость структуры белка

Чтобы оценить изменение структуры белков при МД моделировании, были построены зависимости СКО от времени для белков в течение последних 50 нс траекторий относительно первой структуры в каждом из анализируемых участков траектории (рис. S3). Из представленных данных видно, что структура белка во всех моделях очень стабильна после релаксации в течение первых 50 нс.

Проведенный анализ распределений СКО в траекториях, представленный на рис. S4, показывает, что значения СКО лежат в диапазоне до 3.5 Å, а распределения являются достаточно узкими, что свидетельствует о стабильности конформации белка. Наличие нескольких пиков в распределениях свидетельствует о существовании нескольких устойчивых конформаций в МД траектории, что подтверждается результатами кластерного анализа структуры (см. ниже). Вместе с тем такие структуры близки между собой.

Устойчивость структуры ДНК в составе комплекса

Изменение структуры ДНК при МД моделировании оценивали по величине СКО от времени в течение последних 50 нс траекторий относительно первой структуры в каждой из траекторий (рис. S5). Мы рассматривали структуру как только дуплексного участка, так и комплекса с одноцепочечным нависанием матричной цепи. Это связано с высокой конформационной подвижностью последнего. Из представленных данных видно, что структура дуплексной части ДНК во всех траекториях очень стабильна после релаксации в течении первых 50 нс. Анализ СКО вдоль траекторий для ДНК в комплексе с белком при рассмотрении ДНК с одноцепочечным нависанием матричной цепи и без учета одноцепочечного участка показал, что средние величины СКО и его стандартного отклонения (усреднение по всем моделям) существенно выше в первом случае (3.46 ± 0.97 и 0.84 ± 0.31 Å), чем во втором (1.97 ± 0.77 и 0.39 ± 0.12 Å). Таким образом, для достоверного анализа мы не рассматривали положение одноцепочечного фрагмента ДНК, который вдоль МД траекторий принимал различные положения.

Устойчивость структуры при моделировании свободной ДНК

Среднеквадратичное отклонение вдоль траекторий ДНК без белка показало существенно более высокую подвижность ДНК в сравнении с белковым комплексом (рис. S6). Средние величины СКО и его стандартного отклонения (усреднение по всем моделям) для всей ДНК составили 5.11 ± 1.72 и 1.29 ± 0.61 Å соответственно, а без одноцепочечного фрагмента – 2.45 ± 0.41 и 0.50 ± 0.12 Å. Таким образом, ожидаемо белок значительно ограничивает конформационную подвижность дуплексного фрагмента ДНК и его одноцепочечного фрагмента. Разнообразие принимаемых конформаций в свободной ДНК тоже выше, чем в составе комплекса, о чем говорят более высокие значения средних величин стандартного отклонения.

Анализ структуры белка, ДНК и белково-нуклеинового комплекса

Влияние модификации в ДНК на структуру белка проводили, анализируя величины СКО между Cα атомами белков для наиболее представленных структур в МД траекториях (последние 50 нс), полученных методом кластерного анализа. На основании этого была построена двумерная тепловая карта СКО (рис. S7). Для нативных и модифици-

[1] Рис. S1–S10 и табл. S1–S6 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx39.pdf

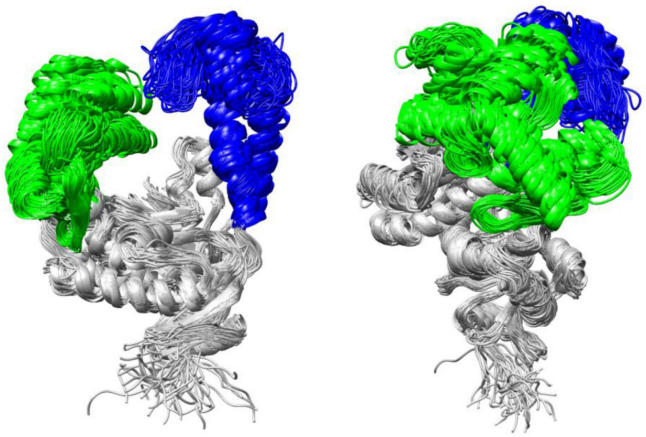


Рис. 3. Сопоставление наиболее представленных структур белков в МД траекториях всех исследованных комплексов, полученных методом кластерного анализа.

Серым цветом показан домен «ладонь», синим – «большой палец», зеленым – «пальцы». Выравнивание структуры белков проведено по домену «ладонь». Справа приведена структура, повернутая относительно левой вокруг вертикальной оси на 90°.

рованных комплексов усредненные значения СКО близки между собой и составляют в среднем 2.60 Å. В случае введения модификаций в различные положения для различных изомеров и ротамеров величины СКО значительно варьируют (от 1.31 до 4.37 Å). Средние значения СКО для каждой из структур в сравнении с другими лежат в диапазоне от 2.33 до 3.26 Å (табл. S1), что свидетельствует о близости всех рассматриваемых структур. Средние величины СКО для каждого из положений модификации в цепи ДНК после усреднения по стереоизомерам и конформерам могут быть выстроены в следующий ряд: X1 < X2 < X6 < X4 < X3 < X5, т. е. в положениях 3 и 5 модификации в наибольшей степени влияют на структуру Taq полимеразы. Анализ значений СКО, усредненных по различным положениям модификации для различных ротамеров и стереоизомеров, показывает, что величины СКО для белков располагаются в ряду S1 > R1 > R2 > S2. Это коррелирует с расположением модификации относительно ДНК цепи: при ориентации *N*-бензимидазольной группы в сторону большой бороздки нарушения в структуре белка оказываются наибольшими.

Сопоставление наиболее представленных в МД траекториях структур белка для всех модельных комплексов показывает, что отличие в их структуре наблюдается главным образом в расположении доменов «пальцы» и «большой палец» и сохранении структуры домена «ладонь», при этом N-конец белка обладает высокой конформационной подвижностью (рис. 3). Такие вариации связаны как с конформационной подвижностью доменов «большой палец» и «пальцы», так и с влиянием модификации на их расположение.

Структура ДНК

Известно, что НК субстрат в комплексе с полимеразой претерпевает значительные конформационные перестройки по сравнению со структурой в растворе (например, (Виноградова, Пышный, 2010)). Основные изменения в ДНК, наблюдаемые в экспериментально определенных структурах их комплексов с полимеразами, – изменение конформации сахара, изменение ширины малой бороздки и появление изгиба цепей. Проведен анализ структур, полученных в ходе моделирования как свободного НК субстрата, так и в комплексе с Taq ДНК-полимеразой. Для этого использовали наиболее представленные в МД моделировании структуры, определенные методом кластерного анализа. Среднеквадратичные отклонения между структурами для дуплексных фрагментов, рассчитанные при моделировании в комплексе с белком и в свободном состоянии для обоих типов стереоизомеров и двух типов конформеров для удлиненных и неудлиненных модельных систем, приведены в табл. S2.

Среднее значение СКО для всех комплексов составило ~2.4 Å. Наибольшее различие между структурой дуплексных участков НК при моделировании с белком и без белка наблюдалось для модификации, введенной по четвертому положению (L0/X4/S1), СКО которой было равно 3.3 Å. Данный факт связан с наибольшей шириной малой бороздки у этой структуры в комплексе с белком, чему способствует положение модификации, которая «экранирует» основания в большой бороздке от взаимодействия с водой и растягивает сахарофосфатный остов. Наименьшие отличия для структур НК с белком и без белка отмечены у структур X5 и X6 (см. табл. S2). СКО между немодифицированным и модифицированным НК субстратами в комплексе с Taq полимеразой составляет ~1.75 Å и в среднем меньше для тех модификаций, которые ориентированы в сторону большой бороздки, в результате чего они в меньшей степени взаимодействуют с белком.

Средняя по различным конформерам и изомерам величина СКО для ДНК в комплексе с белком составляет 2 Å. Меньшие значения соответствуют структурам, у которых модификация ориентирована в одном и том же направлении. Анализ структуры показывает, что даже значительные взаимодействия модификации с белком существенно не нарушают глобальную структуру дуплексной части НК субстрата. В целом структура субстрата с модификацией в большей степени зависит от того, с какими участками белка она взаимодействует, что определяется как изомером, так и конформером *N*-бензимидазольного остатка.

Структурные параметры исследованных НК субстратов в основном типичны для B-формы двойной спирали ДНК, при этом выявлены локальные изменения формы спирали вблизи 3′-конца праймера и места введения модификации. Для неудлиненных модельных систем (L0) показано значительное увеличение значения параметров крена (Roll) и прогиба (Buckle) для AT пар оснований вблизи каталитического центра. Как для удлиненных, так и для неудлиненных комплексов значение угла кручения (Propeller) было в отрицательной области для таких пар (AT), что характерно для A-трактов ДНК, но не последовательностей гетеронуклеотидного состава (Strahs, Schlick, 2000). Отклонение от главной оси спирали пар оснований было тем больше, чем ближе было расположение модификации к каталитическому центру, и значительно снижалось для ДНК с увеличенной дуплексной частью

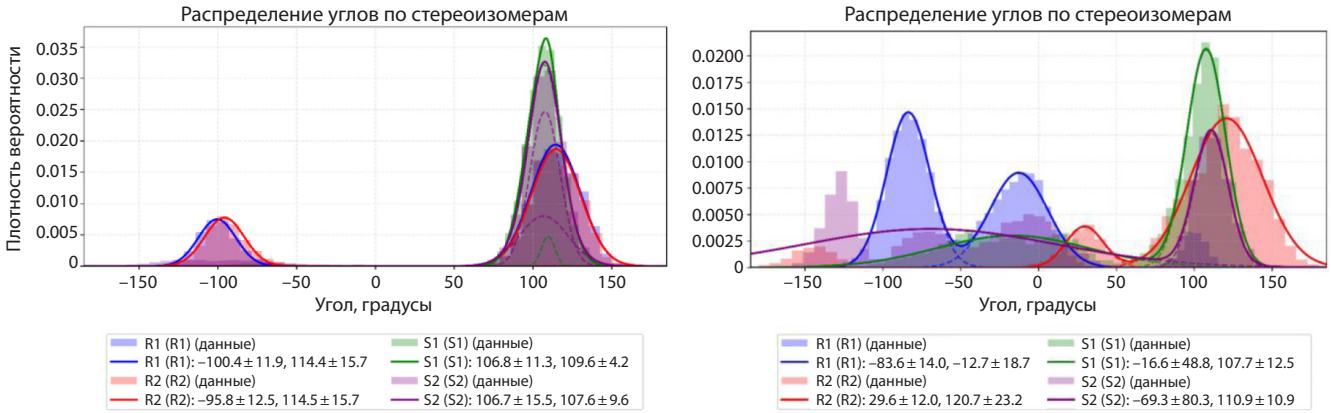


Рис. 4. Величины двугранного угла P—N связи фосфорамидного остатка для конформеров 1 и 2 вдоль МД траектории для свободной ДНК (слева) и ДНК в составе комплекса с белком (справа) для всех исследованных комплексов.

(комплексы L1–L5). Значение параметра поворота (Twist) в среднем составляет 34° и не зависит от длины дуплекса. Это свидетельствует о том, что плотность упаковки ДНК практически не меняется при удлинении дуплексного участка или при введении модификации.

Во всех исследованных комплексах ширина малой бороздки ДНК вблизи каталитического центра значительно увеличена (расстояние между фосфорными атомами противоположных цепей достигает 15–18 Å) и дополнительно возрастает в модифицированных системах по мере увеличения длины дуплекса, что связано с удалением модификации от 3′-конца праймера. У нативных комплексов этот параметр изменялся в меньшей степени. Зависимости структурных параметров НК субстрата от пространственного расположения модификации не выявлено ввиду намного большего влияния положения модификации относительно 3′-конца праймера.

Анализ конформации сахара в ДНК показывает, что в большинстве случаев дезоксирибоза находится в C2′-эндо или близкой к ней конформации, но вблизи 3′-конца праймера наблюдались C1′-экзо и O4′-эндо конформации отдельных нуклеотидов. Это характерно для напряженных состояний, требуемых для акта катализа – присоединения нуклеотида к цепи праймера.

Наличие модификации в цепи ДНК в составе комплекса с Taq полимеразой вызывало значительное изменение канонического плоского спаривания оснований только в случае концевых и предконцевых пар оснований и модификации в первом или втором положении праймера. Структурный анализ показывает, что модификация не влияет на характер спаривания оснований: формируются уотсон-криковские пары со стандартными длинами водородных связей, за исключением концевых пар, что наблюдалось ранее как экспериментально, так и при МД моделировании (Nonin et al., 1995; Zgarbová et al., 2014). Таким образом, наибольшее влияние на локальную структуру ДНК в комплексе с полимеразой оказывала модификация первого межнуклеозидного фосфатного остатка. В целом наличие модификации не сильно изменяет структуру ДНК в составе как свободных дуплексов, так и фермент-субстратного комплекса.

Исследованы расположения *N*-бензимидазольного остатка в составе ДНК дуплекса в свободном состоянии и в комплексе с белком. Для этого анализировали двугранный угол P—N связи, образованный немостиковым атомом кислорода фосфатного остатка (OP1 для Rp изомера и OP2 для Sp изомера), атомами фосфора, азота и атомом углерода бензоазольного кольца. Полученные результаты показали достаточно высокую подвижность остатков и возможность перехода из одного конформера в другой (рис. S8).

Популяционный анализ величин углов в МД траекториях показывает, что в случае свободной ДНК для удлиненных и неудлиненных комплексов в целом наблюдаются одинаковые закономерности (рис. S8 и S9). Из приведенных данных видно, что Rp изомер модифицированного остатка преимущественно ориентирован в сторону малой бороздки, а Sp изомер – в сторону большой бороздки, что соответствует углу около +100°. В ряде случаев происходит выворачивание модификации в сторону от двойной спирали, что соответствует углу около −100° (конформеры R1 и S2). Меньшая доля такой ориентации связана с тем, что гидрофобный остаток стремится минимизировать контакт с водой путем взаимодействия с ДНК цепями. Следует отметить, что в большинстве случаев различий в форме распределений между двумя изомерами не выявляется: как правило, наличие двух пиков у изомера 1 сопровождается их наличием и у изомера 2. Разная амплитуда пиков в распределениях указывает на то, что мы не в достаточной степени покрыли конформационное пространство для положения модификации в рамках 50 нс траекторий для каждой из моделей. Однако если построить вероятностное распределение углов для всех положений в цепи каждого из изомеров, то средние значения для конформеров 1 и 2 каждого из стереоизомеров практически полностью совпадут (рис. 4).

В комплексе с белком положение модификации претерпевает значительные изменения в сравнении со свободной ДНК (см. рис. 4). На графиках величин углов и их вероятностей (см. рис. 4, рис. S9 и S10) в МД траекториях присутствуют углы, как лежащие между пиками, характерными для свободной ДНК (+100° и −100°), так и

смещенные относительно них в бóльшую по абсолютной величине сторону. Это свидетельствует о значительном взаимодействии модифицированного остатка с белком.

Сопоставление средних величин вероятностей для разных типов изомеров в комплексах показывает, что они сильно отличаются между собой как для различных положений модификации, так и для данных, полученных в случае свободной ДНК (см. рис. 4, рис. S8 и S10). Видно, что распределение вероятностей для конформеров R1 и R2 радикально различается. Основной пик для R1 лежит в области −80°, что соответствует ориентации модификации в сторону большой бороздки (в сторону от ДНК). Это связано с тем, что нативный фосфатный остов ДНК при взаимодействии с белком координируется амикислотными остатками и, как следствие, модифицированный остаток вытесняется из малой бороздки ДНК. Основной пик распределения для конформера R2 лежит в районе +100°, что соответствует ориентации модификации в область малой бороздки. В случае изомера S1 угол составляет +100°, и модификация ориентирована преимущественно в сторону большой бороздки. Значительно от нее отличается изомер S2, в графике вероятности которого имеется набор равных по амплитуде пиков практически во всем диапазоне. Это говорит о том, что модификация может занимать в различных положениях цепи (X*j*) разные пространственные положения, что обусловлено ее взаимодействием с белком.

Следует отметить, что для всех рассмотренных изомеров возникает пик в области 0° (см. рис. 4), что соответствует ориентации модификации в сторону от ДНК. При этом одна из аминогрупп пятичленного кольца *N*-бензимидазола формирует водородную связь с немостиковым атомом кислорода фосфатного остатка. Отсутствие таких ориентаций в свободной ДНК свидетельствует о том, что это положение стабилизируется дополнительным взаимодействием с белком.

Анализ взаимодействия модификации и Taq полимеразы

Проведен кластерный анализ последних 50 нс МД траекторий для получения наиболее представленных структур в них. Изучено расположение *N*-бензоазольных групп относительно активного центра и рассчитано количество атомов белка и карты контактов модификации с Taq полимеразой. Все атомы аминокислотных остатков, которые находились ближе 3 Å к модифицированному фосфатному остатку, считали взаимодействующими с ней (табл. S3 и S4). Размер дуплексного участка ДНК, который взаимодействует с Taq полимеразой, составляет 5–8 пар оснований, а во взаимодействии с ним принимает участие около 40 аминокислотных остатков, роль которых состоит узнавании и удержания НК, а также акте катализа (Eom et al., 1996; Li Y. et al., 1998).

Анализ контактов между фосфорамидным *N*-бензимидазольным остатком и Taq полимеразой позволил выявить несколько закономерностей. Во-первых, в комплексах L0/X1/R1, L1/X2/R2, L0/X3/R2, L2/X3/R1 и L4/X5/R2 бензимидазольная модификация находилась в карманах белковой молекулы, а в случае четвертого положения для

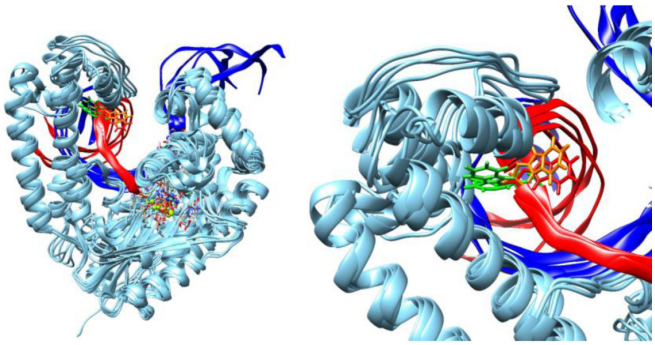


Рис. 5. Сопоставление структуры комплексов L0/X4: общий вид (слева) и приближение области взаимодействия модификации с доменом фермента «большой палец» (справа).

Taq ДНК-полимераза показана голубым цветом, матричная цепь ДНК – синим, праймер – красным. Модифицированные *N*-бензоазольные группы показаны в виде атомарной модели, Sp изомеры – красным и оранжевым цветом, Rp – зеленым.

R стереоизомеров оба конформера (X4/R1 и X4/R2) находились в кармане, причем с образованием стабильных связей между электроотрицательными атомами модификации и положительно заряженными аргининами белка (рис. 5).

В целом модификация в положениях с первого по пятое (X1–X5) формирует обширную сеть водородных связей и ван-дер-ваальсовых контактов с белком, в то время как для шестого положения такие взаимодействия намного менее выражены. Стереоизомерия также значительно влияет на характер связывания: Sp стереоизомеры демонстрировали преимущественное взаимодействие с положительно заряженными остатками, тогда как Rp стереоизомеры чаще образовывали контакты с гидрофобными аминокислотами. Sp изомеры часто ориентированы в сторону большой бороздки и «экранируют» гетероциклические основания в структуре дуплекса от взаимодействия с водой, а Rp изомеры в основном направлены от ДНК в сторону белка. Наличие модификаций нередко приводит к нарушению регулярной структуры НК за счет взаимодействия бензимидазольной группы в кармане белка и возникновения напряжения в сахарофосфатном остове. Введение модификации в первое и второе положение праймера (X1 и X2) приводит к значительному нарушению спаривания концевой и предконцевой пар, а в случае комплексов L1/X2/S2, L0/X3/R1 и L0/X2/R1 наблюдается разрушение уотсон-криковских пар вблизи модификации.

Из литературных данных известно, что остаток Arg660 из домена «пальцы» координирует фосфатную группу в первом положении праймера от 3′-конца, Arg587 из домена «ладонь» координирует второй межнуклеозидный фосфат, а Arg536 из «большого пальца» – четвертый (Виноградова, Пышный, 2010). Наличие *N*-бензимидазольной модификации должно приводить к нейтрализации заряда фосфатной группы, а также к стерическим затруднениям при координирования фосфатной группы аргининами, что должно способствовать снижению скорости катализа. Однако структурный анализ показывает, что в случае Sp изомеров во второй и четвертой позиции (X2 и X4)

для обоих конформеров немостиковый атом кислорода фосфатного остатка все же координируется Arg536, как и один из конформеров Rp изомера при расположении модификации в первом (X1) или втором (X2) положении праймера.

Оба конформера Rp изомера в четвертом положении (X4/R1 и X4/R2) находятся в гидрофобном кармане домена большого пальца, в то время как Sp изомер значительно с белком не взаимодействует. В результате Rp фосфатная группа не позволяет переместить фермент в следующее положение вдоль цепи ДНК, необходимое для присоединения следующего нуклеотида к праймеру. Скорее всего, именно поэтому наблюдаются продукты неполного удлинения при расположении модификации в четвертом положении.

Нами было показано (Golyshev et al., 2025) в экспериментах по удлинению праймеров с *N*-бензимидазольной модификацией Taq ДНК-полимеразой, что введение модификации во вторую позицию демонстрирует наименьшее снижение эффективности удлинения в случае совершенных комплексов, что коррелирует с наименьшим количеством контактов модификации с белком среди первых трех межнуклеозидных фосфатных остатков. Кроме того, для всех совершенных комплексов с модификациями отчетливо наблюдалась полоса, соответствующая частично удлиненному праймеру, содержащему модификацию в четвертой позиции от 3′-конца. Эффект был более выражен в случае праймеров, несущих модификации в первой и третьей позициях. Эти наблюдения хорошо коррелируют с данными о расположении обоих конформеров R стереоизомеров (X4/R1 и X4/R2), которые находятся в кармане белка и эффективно взаимодействуют с ферментом (см. рис. 5).

Таким образом, стерические взаимодействия Rp изомеров с карманами белка способны замедлять или, как в случае четвертого положения модификации, препятствовать перемещению Taq ДНК полимеразы по субстрату. Это находит экспериментальное подтверждение в снижении скорости работы полимеразы и возникновении продуктов абортивного удлинения модифицированного праймера, содержащего фосфорамидную *N*-бензимидазольную группу.

Энергия взаимодействия субстрата с полимеразой

Описанные в предыдущем разделе взаимодействия могут находить отражение в энергии взаимодействия фермента с субстратом. Поэтому мы рассчитали энергию взаимодействия между субстратом и Taq полимеразой при помощи метода молекулярной динамики по обобщенной модели Борна (Molecular Mechanics/Generalized Born Surface Area, MMGBSA), используя только МД траекторию комплекса. Для расчета брали в рассмотрение только дуплексную часть НК субстрата, чтобы снизить вариации рассчитываемой свободной энергии, связанные со свободным движением одноцепочечного участка матрицы. Величины энергий ДНК, белка, их комплекса, а также энергии комплексообразования приведены в табл. S5 и S6. Анализ энергий взаимодействия между модифицированным НК субстратом и Taq полимеразой выявил следующие закономерности. Для нативных комплексов энергия связи увеличивалась по модулю с увеличением длины дуплексного участка, тогда как для модифицированных комплексов не обнаружено корреляции между энергией и длиной НК субстрата. В то же время наличие модификаций в 5-й и 6-й межнуклеозидных фосфатных группах показывало снижение эффективности связывания по сравнению со всеми остальными модельными системами, что коррелировало с уменьшением количества контактов ДНК с белком.

В случае неудлиненных модельных систем (L0) с наименьшей длиной дуплексного участка энергия комплексообразования оказалась в среднем значительно ниже (примерно −200 ккал/моль), чем в случае удлиненных комплексов (примерно −180 ккал/моль). Для большинства комплексов более энергетически выгодное связывание НК субстрата с белком наблюдалось для S стереоизомеров, что, скорее всего, обусловлено большей доступностью немостикового атома кислорода модифицированной фосфатной группы для взаимодействия с белком путем координации фосфатного остатка аминокислотами. Среди двух конформеров наибольшей по модулю энергией связи обладали конформации S1, где *N*-бензимидазольная модификация ориентирована в сторону большой бороздки, позволяя аминокислотным остаткам взаимодействовать с немостиковым атомом кислорода модифицированного фосфата. При этом мы не обнаружили связи между количеством атомов, близлежащих к модификации (атомов, находящихся на расстоянии менее 3 Å от атомов *N*-бензимидазольной модификации, см. табл. S3), и энергией комплексообразования. Однако наибольшая энергия связывания фермента с субстратом наблюдалась для комплексов L0/X4/R1 и L0/X4/R2, для которых модификация располагается в кармане белка, взаимодействуя с наибольшим числом аминокислотных остатков (см. табл. S4).

Заключение

В настоящей работе методами молекулярного моделирования и анализа мы провели исследование структуры, динамики и энергии взаимодействия ДНК субстратов, которые содержат фосфорамидную *N*-бензимидазольную группу в праймерной цепи в различных положениях, с Taq ДНК-полимеразой. Было установлено, что положение модификации вблизи 3′-конца праймера и его стереоизомерия влияют на взаимодействие с ферментом. При этом в составе фермент-субстратного комплекса возникает два устойчивых конформера для каждого из изомеров фосфорамидной *N*-бензимидазольной группы. Анализ влияния стереометрии модификации на взаимодействие праймера с ферментом показывает более сильное взаимодействие Rp изомеров с полимеразой; наиболее сильное взаимодействие наблюдается при расположении модификации в четвертом межнуклеозидном фосфатном остатке с 3′-конца праймера. Структурный анализ ДНК и белка не выявил значимых структурных перестроек в биополимерах, а нарушения структуры при введении модификации в большинстве исследованных модельных систем либо незначительные, либо носят локальный характер. Наибольшее влияние на локальную структуру ДНК в ком-

плексе с полимеразой оказывала модификация первого межнуклеозидного фосфатного остатка.

Эти данные хорошо коррелируют с экспериментальными данными по процессированию ФАО Taq ДНК-полимеразой. В частности, со сниженной скоростью наработки продуктов полноразмерного удлинения ФАО праймеров, образования продуктов неполного удлинения праймера, содержащего модификацию в четвертой позиции от 3′-конца праймера, а также со значительным снижением эффективности удлинения праймера при введении модификации в первое положение (Chubarov et al., 2024; Golyshev et al., 2025).

Полученные нами результаты раскрывают молекулярные основы влияния фосфорамидной *N*-бензимидазольной группы на удлинение фосфорамидных бензоазольных праймеров. Эти данные будут применяться для рационального дизайна структуры ФАО при разработке систем молекулярной диагностики на основе методов ПЦР. Кроме того, существенные различия в эффективности взаимодействия полимеразы с Rp и Sp изомерами ФАО указывают на необходимость разработки методов стереоселективного синтеза таких олигомеров для более точного управления свойствами фосфорамидных бензоазольных олигонуклеотидов.

Список литературы / References

Виноградова О.А., Пышный Д.В. Селективность ферментативного превращения олигонуклеотидных зондов при анализе нуклеотидных полиморфизмов ДНК. *Acta Naturae*. 2010;2(1):40-58 [Vinogradova O.A., Pyshnyi D.V. Selectivity of enzymatic conversion of oligonucleotide probes during nucleotide polymorphism analysis of DNA. *Acta Naturae*. 2010;2(1):40-58. doi 10.32607/20758251-2010-2-1-36-52]

Abramson J., Adler J., Dunger J., Evans R., Green T., Pritzel A., Ronneberger O., … Bapst V., Kohli P., Jaderberg M., Hassabis D., Jumper J.M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500. doi 10.1038/s41586-024-07487-w

Case D.A., Belfon K., Ben-Shalom I.Y., Brozell S.R., Cerutti D.S., Cheatham T.E. III, Cruzeiro V.W.D., … Wu X., Xiong Y., Xue Y., York D.M., Kollman P.A. Amber 20. San Francisco, Univ. of California, 2020. Available at: https://ambermd.org/doc12/Amber20.pdf

Chubarov A.S., Oscorbin I.P., Filipenko M.L., Lomzov A.A., Pyshnyi D.V. Allele-specific PCR for *KRAS* mutation detection using phosphoryl guanidine modified primers. *Diagnostics*. 2020;10(11):872. doi 10.3390/diagnostics10110872

Chubarov A.S., Oscorbin I.P., Novikova L.M., Filipenko M.L., Lomzov A.A., Pyshnyi D.V. Allele-specific PCR for PIK3CA mutation detection using phosphoryl guanidine modified primers. *Diagnostics*. 2023;13(2):250. doi 10.3390/diagnostics13020250/S1

Chubarov A.S., Baranovskaya E.E., Oscorbin I.P., Yushin I.I., Filipenko M.L., Pyshnyi D.V., Vasilyeva S.V., Lomzov A.A. Phosphoramidate azole oligonucleotides for single nucleotide polymorphism detection by PCR. *Int J Mol Sci.* 2024;25(1):617. doi 10.3390/ijms25010617

Di Giusto D., King G.C. Single base extension (SBE) with proofreading polymerases and phosphorothioate primers: improved fidelity in single-substrate assays. *Nucleic Acids Res.* 2003;31(3):e7. doi 10.1093/nar/gng007

Eom S.H., Wang J., Steitz T.A. Structure of Taq polymerase with DNA at the polymerase active site. *Nature*. 1996;382(6588):278-281. doi 10.1038/382278A0

Golyshev V.M., Yushin I.I., Gulyaeva O.A., Baranovskaya E.E., Lomzov A.A. Properties of phosphoramide benzoazole oligonucleotides (PABAOs). I. Structure and hybridization efficiency of *N*-benzimidazole derivatives. *Biochem Biophys Res Commun.* 2024;693:149390. doi 10.1016/j.bbrc.2023.149390

Golyshev V.M., Morozova F.V., Berdugin A.A., Kozyreva E.A., Baranovskaya E.E., Yushin I.I., Lomzov A.A. Structural and thermodynamic insights for enhanced SNP detection using *N*-benzimidazole oligonucleotides. *J Phys Chem B.* 2025;129(44):11409-11420. doi 10.1021/acs.jpcb.5c04047

Ishige T., Itoga S., Matsushita K. Locked nucleic acid technology for highly sensitive detection of somatic mutations in cancer. *Adv Clin Chem*. 2018;83:53-72. doi 10.1016/bs.acc.2017.10.002

Izadi S., Anandakrishnan R., Onufriev A.V. Building water models: a different approach. *J Phys Chem Lett.* 2014;5(21):3863-3871. doi 10.1021/jz501780a

Kalendar R., Baidyussen A., Serikbay D., Zotova L., Khassanova G., Kuzbakova M., Jatayev S., Hu Y.G., Schramm C., Anderson P.A., Jenkins C.L.D., Soole K.L., Shavrukov Y. Modified "Allele-specific qPCR" method for SNP genotyping based on FRET. *Front Plant Sci.* 2022;12:747886. doi 10.3389/fpls.2021.747886

Kutyavin I.V. Use of base modifications in primers and amplicons to improve nucleic acids detection in the real-time snake polymerase chain reaction. *Assay Drug Dev Technol.* 2011;9(1):58-68. doi 10.1089/adt.2010.0303

Li Y., Korolev S., Waksman G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.* 1998;17(24):7514-7525. doi 10.1093/emboj/17.24.7514

Li Z., Song L.F., Li P., Merz K.M. Systematic parametrization of divalent metal ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB water models. *J Chem Theory Comput.* 2020;16(7):4429-4442. doi 10.1021/acs.jctc.0c00194

Meagher K.L., Redman L.T., Carlson H.A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem.* 2003;24(9):1016-1025. doi 10.1002/jcc.10262

Nonin S., Leroy J.L., Guéron M. Terminal base pairs of oligodeoxynucleotides: imino proton exchange and fraying. *Biochemistry*. 1995;34(33):10652-10659. doi 10.1021/bi00033a041

Novgorodtseva A.I., Vorob'ev A.Y., Lomzov A.A., Vasilyeva S.V. Synthesis and physicochemical properties of new phosphoramide oligodeoxyribonucleotides. I. N-caffeine derivatives. *Bioorg Chem.* 2025;157:108313. doi 10.1016/j.bioorg.2025.108313

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25(13):1605-1612. doi 10.1002/jcc.20084

Rejali N.A., Moric E., Wittwer C.T. The effect of single mismatches on primer extension. *Clin Chem.* 2018;64(5):801-809. doi 10.1373/clinchem.2017.282285

Roe D.R., Cheatham T.E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput.* 2013;9(7):3084-3095. doi 10.1021/ct400341p

Shapovalov M.V., Dunbrack R.L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*. 2011;19(6):844-858. doi 10.1016/j.str.2011.03.019

Starza I.D., Eckert C., Drandi D., Cazzaniga G.; EuroMRD Consortium. Minimal residual disease analysis by monitoring immunoglobulin and T-cell receptor gene rearrangements by quantitative PCR and droplet digital PCR. *Methods Mol Biol.* 2022;2453:79-89. doi 10.1007/978-1-0716-2115-8_5

Strahs D., Schlick T. A-tract bending: insights into experimental structures by computational models. *J Mol Biol.* 2000;301(3):643-663. doi 10.1006/jmbi.2000.3863

Terpe K. Overview of thermostable DNA polymerases for classical PCR applications: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol.* 2013;97(24):10243-10254. doi 10.1007/s00253-013-5290-2

Tian C., Kasavajhala K., Belfon K.A.A., Raguette L., Huang H., Migues A.N., Bickel J., Wang Y., Pincay J., Wu Q., Simmerling C. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J Chem Theory Comput.* 2020;16(1):528-552. doi 10.1021/acs.jctc.9b00591

Unni S., Huang Y., Hanson R.M., Tobias M., Krishnan S., Li W.W., Nielsen J.E., Baker N.A. Web servers and services for electrostatics calculations with APBS and PDB2PQR. *J Comput Chem.* 2011; 32(7):1488-1491. doi 10.1002/jcc.21720

Vasilyeva S.V., Baranovskaya E.E., Dyudeeva E.S., Lomzov A.A., Pyshnyi D.V. Synthesis of oligonucleotides carrying inter-nucleotide *N*-(benzoazole)-phosphoramide moieties. *ACS Omega*. 2023; 8(1):1556-1566. doi 10.1021/acsomega.2c07083

Yushin I.I., Golyshev V.M., Novgorodtseva A.I., Lomzov A.A. Properties of phosphoramide benzoazole oligonucleotides (PABAOs). II. Structure and hybridization efficiency of *N*-benzoxazole derivatives. *Biochem Biophys Res Commun.* 2024;740:150997. doi 10.1016/j.bbrc.2024.150997

Zgarbová M., Otyepka M., Šponer J., Lankaš F., Jurečka P. Base pair fraying in molecular dynamics simulations of DNA and RNA. *J Chem Theory Comput.* 2014;10(8):3177-3189. doi 10.1021/ct500120v

Zgarbová M., Šponer J., Jurečka P. Z-DNA as a touchstone for additive empirical force fields and a refinement of the Alpha/Gamma DNA torsions for AMBER. *J Chem Theory Comput.* 2021;17(10):6292-6301. doi 10.1021/acs.jctc.1C00697

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 31.07.2025. После доработки 09.09.2025. Принята к публикации 09.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1084-1096

doi 10.18699/vjgb-25-113

Предсказание взаимодействий белка ORF3a SARS-CoV-2 с низкомолекулярными лигандами с использованием когнитивной платформы ANDSystem, графовых нейронных сетей и молекулярного моделирования

Т.В. Иванисенко, П.С. Деменков, М.А. Клещев, В.А. Иванисенко

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
itv@bionet.nsc.ru

Аннотация. В последние годы методы искусственного интеллекта, основанные на анализе гетерогенных графов биомедицинских сетей, получили широкое распространение для предсказания молекулярных взаимодействий. В частности, графовые нейронные сети (graph neural networks, GNN) позволяют эффективно выявлять отсутствующие ребра в генных сетях, таких как сети белок-белковых взаимодействий, ген–заболевание, лекарство–мишень и др., и тем самым предсказывать новые биологические связи. Для реконструкции генных сетей часто применяют когнитивные системы автоматического анализа текстов научных публикаций и баз данных. Одна из таких платформ, базирующаяся на методах искусственного интеллекта,– ANDSystem, предназначенная для автоматического извлечения знаний о молекулярных взаимодействиях и на этой основе – реконструкции ассоциативных генных сетей. База знаний ANDSystem содержит сведения о более чем 100 млн взаимодействий между различными молекулярно-генетическими объектами (гены, белки, метаболиты, лекарства и др.). Взаимодействия представлены широким спектром типов: регуляторные связи, физические взаимодействия (белок–белок, белок–лиганд), каталитические и химические реакции, ассоциации между генами, фенотипами, заболеваниями и др. В настоящем исследовании мы применили графовые нейронные сети с механизмом внимания, обученные на графе знаний ANDSystem, для предсказания новых ребер между белками и лигандами и поиска потенциальных лигандов для белка ORF3a SARS-CoV-2. Вспомогательный белок ORF3a SARS-CoV-2 играет важную роль в патогенезе вируса за счет ион-канальной активности, индукции апоптоза и способности модулировать эндолизосомальные процессы и врожденный иммунитет хозяина. Несмотря на широкий спектр функций, ORF3a как фармакологическая мишень изучен значительно меньше, чем другие вирусные белки. Применение графовой нейронной сети позволило нам предсказать пять малых молекул разного происхождения (метаболиты и лекарство), потенциально взаимодействующих с ORF3a: N-ацетил-D-глюкозамин, 4-(бензоиламино)бензойная кислота, аустоцистин D, биктегравир и L-треонин. Молекулярный докинг и оценка аффинности методом MM/GBSA подтвердили потенциальную способность этих соединений образовывать комплексы с ORF3a. Анализ локализации показал, что сайты связывания биктегравира и 4-(бензоиламино)бензойной кислоты расположены в цитозольной поверхностной области белка, доступной растворителю; L-треонин связывается в межсубъединичной щели димера, а аустоцистин D и N-ацетил-D-глюкозамин – на границе между цитозольной поверхностью и трансмембранной областью. Доступность этих сайтов связывания может быть снижена из-за влияния липидного бислоя. Энергетические характеристики связывания у биктегравира по сравнению с 4-(бензоиламино)бензойной кислотой оказались более высокими (–7.37 ккал/моль в докинге; –14.71 ± 3.12 ккал/моль по MM/GBSA), что делает его перспективным кандидатом для репозиционирования как ингибитора ORF3a. Взаимодействие биктегравира с ORF3a может нарушать связывание ORF3a с белком хозяина VPS39 – субъединицей комплекса HOPS, участвующего в слиянии аутофагосом и поздних эндосом с лизосомами. Это, в свою очередь, может снимать индуцируемую ORF3a блокаду данного процесса и тем самым способствовать восстановлению аутофагического потока и лизосомной деградации вирусных компонентов.

Ключевые слова: ANDSystem; SARS-CoV-2; ORF3a; генные сети; графовые нейронные сети; предсказание белок–лиганд взаимодействий; биктегравир; 4-(бензоиламино)бензойная кислота; молекулярный докинг; потенциальные лекарства

Для цитирования: Иванисенко Т.В., Деменков П.С., Клещев М.А., Иванисенко В.А. Предсказание взаимодействий белка ORF3a SARS-CoV-2 с низкомолекулярными лигандами с использованием когнитивной платформы ANDSystem, графовых нейронных сетей и молекулярного моделирования. *Вавиловский журнал генетики и селекции*. 2025;29(7):1084-1096. doi 10.18699/vjgb-25-113

Финансирование. Исследование выполнено за счет средств бюджетного проекта ФИЦ ИЦиГ СО РАН «Системная биология и биоинформатика: реконструкция, анализ и моделирование структурно-функциональной организации и эволюции генных сетей человека, животных, растений и микроорганизмов» № FWNR-2022-0020.

© Иванисенко Т.В., Деменков П.С., Клещев М.А., Иванисенко В.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Prediction of interactions between the SARS-CoV-2 ORF3a protein and small-molecule ligands using the ANDSystem cognitive platform, graph neural networks, and molecular modeling

T.V. Ivanisenko, P.S. Demenkov, M.A. Kleshchev, V.A. Ivanisenko

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
itv@bionet.nsc.ru

Abstract. In recent years, artificial intelligence methods based on the analysis of heterogeneous graphs of biomedical networks have become widely used for predicting molecular interactions. In particular, graph neural networks (GNNs) effectively identify missing edges in gene networks – such as protein–protein interaction, gene–disease, drug–target, and other networks – thereby enabling the prediction of new biological relationships. To reconstruct gene networks, cognitive systems for automatic text mining of scientific publications and databases are often employed. One such AI-driven platform, ANDSystem, is designed for automatic knowledge extraction of molecular interactions and, on this basis, the reconstruction of associative gene networks. The ANDSystem knowledge base contains information on more than 100 million interactions among diverse molecular genetic entities (genes, proteins, metabolites, drugs, etc.). The interactions span a wide range of types: regulatory relationships, physical interactions (protein–protein, protein–ligand), catalytic and chemical reactions, and associations among genes, phenotypes, diseases, and more. In the present study, we applied attention-based graph neural networks trained on the ANDSystem knowledge graph to predict new edges between proteins and ligands and to identify potential ligands for the SARS-CoV-2 ORF3a protein. The accessory protein ORF3a plays an important role in viral pathogenesis through ion-channel activity, induction of apoptosis, and the ability to modulate endolysosomal processes and the host innate immune response. Despite this broad functional spectrum, ORF3a has been explored far less as a pharmacological target than other viral proteins. Using a graph neural network, we predicted five small molecules of different origins (metabolites and a drug) that potentially interact with ORF3a: N-acetyl-D-glucosamine, 4-(benzoylamino)benzoic acid, austocystin D, bictegravirum, and L-threonine. Molecular docking and MM/GBSA affinity estimation indicate the potential ability of these compounds to form complexes with ORF3a. Localization analysis showed that the binding sites of bictegravir and 4-(benzoylamino)benzoic acid lie in a cytosolic surface pocket of the protein that is solvent-exposed; L-threonine binds within the intersubunit cleft of the dimer; and austocystin D and N-acetyl-D-glucosamine are positioned at the boundary between the cytosolic surface and the transmembrane region. The accessibility of these binding sites may be reduced by the influence of the lipid bilayer. The binding energetics for bictegravirum were more favorable than for 4-(benzoylamino)benzoic acid (docking score –7.37 kcal/mol; MM/GBSA ΔG –14.71 ± 3.12 kcal/mol), making bictegravirum a promising candidate for repurposing as an ORF3a inhibitor.

Key words: ANDSystem; SARS-CoV-2; ORF3a; gene networks; graph neural networks; protein–ligand interaction prediction; bictegravirum; 4-(benzoylamino)benzoic acid; molecular docking; potential therapeutic agents

For citation: Ivanisenko T.V., Demenkov P.S., Kleshchev M.A., Ivanisenko V.A. Prediction of interactions between the SARS-CoV-2 ORF3a protein and small-molecule ligands using the ANDSystem cognitive platform, graph neural networks, and molecular modeling. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1084-1096. doi 10.18699/vjgb-25-113

Введение

Разработка противовирусных препаратов имеет приоритетное значение из-за риска глобальных пандемий и появления в их ходе новых вариантов патогенных вирусов, что продемонстрировала пандемия COVID-19, вызванная SARS-CoV-2 (Ng et al., 2022). SARS-CoV-2 – оболочечный бетакоронавирус с одноцепочечным геномом РНК положительной полярности длиной около 29.9 тыс. нуклеотидов; геном кодирует структурные (S, E, M, N) белки, а также ряд неструктурных белков, обеспечивающих репликацию и сборку вириона (Naqvi et al., 2020). Поскольку именно эти белки определяют ключевые этапы жизненного цикла вируса, наибольшее внимание при разработке лекарственных средств сосредоточено на трех основных мишенях – главной протеазе (3CLpro/Mpro), РНК-зависимой РНК-полимеразе (RdRp) и шиповидном гликопротеине S (белок Spike) (Boby et al., 2023).

Для поиска и оптимизации ингибиторов этих мишеней применялась совокупность экспериментальных и вычислительных подходов: *de novo* дизайн, высокопроизводительный скрининг и репозиционирование известных лекарств (von Delft et al., 2023). В результате такого подхода получены соединения с подтвержденной противовирусной активностью *in vitro* и *in vivo*, а также реализованы клинические стратегии лечения COVID-19, включая ингибирование протеазы и полимеразы. В частности, противовирусный препарат нирматрелвир/ритонавир (Paxlovid), нацеленный на основную протеазу Nsp5 (nonstructural protein 5) коронавируса SARS-CoV-2, получил полное одобрение FDA (Food and Drug Administration – Управление по санитарному надзору за качеством пищевых продуктов и медикаментов, США) 25 мая 2023 г. для лечения взрослых с COVID-19 (FDA, 2023). Лекарство ремдесивир (Veklury), нацеленное на вирусную РНК-зависимую РНК-полимеразу (RdRp, nsp12), одобрено FDA в октябре 2020 г. (FDA, 2020). Параллельно развиваются альтернативные подходы, направленные на блокирование слияния вирусной и клеточной мембран

при проникновении SARS-CoV-2 в клетку. В частности, пептидные ингибиторы, комплементарные HR1/HR2-доменам S2-субъединицы белка Spike, препятствуют формированию шестиспирального пучка (6-HB) – ключевой структуры, обеспечивающей мембранное слияние, и тем самым предотвращают проникновение вируса (Dong et al., 2024).

К числу перспективных классов фармакологических мишеней относятся вспомогательные вирусные белки, модулирующие взаимодействия SARS-CoV-2 с клеточными системами хозяина. Одним из таких белков является ORF3a. Он преимущественно локализуется в поздних эндосомах и лизосомах, где колокализуется с лизосомальными белками человека LAMP1 и катепсином D (Zhang J. et al., 2021; Hinkle et al., 2025). ORF3a формирует ионные каналы (виропориновая активность) (Zhang J. et al., 2022), индуцирует апоптоз через окислительный стресс и активацию каспаз (Zhang Y. et al., 2021), активирует инфламмасому NLRP3 (каскад ORF3a–NLRP3–ASC) (Zhang J. et al., 2022) и подавляет интерфероновые сигнальные пути, усиливая патогенность вируса (Zhang J. et al., 2022).

ORF3a представляет собой димерный мембранный белок с тремя трансмембранными спиралями и крупным цитозольным С-концевым доменом, что показано методом крио-ЭМ (Kern et al., 2021). Он взаимодействует с человеческим белком VPS39 – компонентом комплекса HOPS, и это взаимодействие приводит к блокированию слияния аутофагосом с лизосомами. Ключевую роль в связывании ORF3a с VPS39 играет короткий тирозин-зависимый сигнальный мотив YXXФ (Y – тирозин, X – произвольная аминокислота, Ф – гидрофобный остаток), представленный в ORF3a последовательностью YNSV (позиции 160–163) (Stephens et al., 2025). Точечная замена Y160A, нарушающая данный мотив, устраняет коиммунопреципитацию с VPS39 и снимает блокаду слияния аутофагосом с лизосомами (Zhang Y. et al., 2021).

В последние годы все более широкое применение в фармакологии и смежных областях находят методы искусственного интеллекта, дающие возможность выявлять скрытые закономерности в больших биомедицинских данных. Перспективным направлением считаются графовые нейронные сети (graph neural networks, GNN), которые позволяют интегрировать разнородную биологическую информацию и предсказывать новые, ранее не представленные в литературе взаимодействия в сложных сетях. Одной из ранних работ, сыгравших заметную роль в формировании данного подхода, было исследование M. Zitnik с коллегами (2018), в котором показано, что графовые сверточные нейронные сети способны моделировать взаимодействия лекарство–заболевание и предсказывать побочные эффекты лекарств.

В дальнейшем этот подход активно развивался: появились исследования по интеграции различных источников данных (внешние базы данных, рефераты и научные публикации, патенты, электронные медицинские карточки и др.), предсказанию белок-лигандных и белок-белковых взаимодействий, а также по поиску мишеней для репозиционирования препаратов с использованием GNN (Stokes et al., 2020; Gaudelet et al., 2021). В частности, обнаружено соединение halicin как кандидат с антибактериальной активностью против устойчивых штаммов; с помощью графовой нейронной сети было показано, что эта молекула обладает бактерицидным действием в отношении *Mycobacterium tuberculosis*, карбапенем-резистентных энтеробактерий, а также мультирезистентных штаммов *Acinetobacter baumannii*, *Pseudomonas aeruginosa* и *Clostridioides difficile* (Stokes et al., 2020).

Для поиска фармакологических мишеней на уровне генома человека все чаще применяют методы реконструкции и анализа генных и ассоциативных сетей (Ali, Alrashid, 2025). На этом фоне активно развиваются когнитивные системы и методы инженерии знаний, автоматизирующие извлечение фактов из литературы и специализированных баз, а также формирующие биомедицинские графы знаний. В таких графах узлы представляют собой гены, белки, метаболиты, заболевания, лекарственные средства и другие биомедицинские сущности, а ребра – их взаимодействия (регуляторные связи, взаимодействия белок–белок, ассоциации с заболеваниями и др.). Среди известных ресурсов, реализующих подобный подход, можно отметить STRING (Nicholson, Greene, 2020; Szklarczyk et al., 2023), QIAGEN Ingenuity Pathway Analysis (Krämer et al., 2014), GeneGo/MetaCore (Clarivate) и др.

Ранее нами была разработана когнитивная платформа ANDSystem, предназначенная для реконструкции ассоциативных генных сетей. Она объединяет два направления: 1) автоматическое извлечение знаний из научных публикаций и биологических баз данных с помощью семантико-лингвистических шаблонов и правил (Ivanisenko V.A. et al., 2015, 2019) и 2) интеграцию статистических и машинных методов, включая графовые нейронные сети, для предсказания и добавления новых белок-белковых взаимодействий в сеть (Ivanisenko T.V. et al., 2024).

База знаний (БЗ) ANDSystem содержит сведения более чем о 100 млн взаимодействий между различными типами молекулярно-генетических объектов (гены, РНК, белки, метаболиты, лекарственные вещества), а также объектами клеточного и организменного уровня, такими как клетки, биологические процессы, заболевания и фенотипические признаки. Взаимодействия классифицированы по 49 типам, включая регуляторные связи (регуляция экспрессии, активности, стабильности, транспорта и др.), физические взаимодействия (белок–белок, белок–лиганд), химические взаимодействия (каталитические реакции, посттрансляционные модификации и др.) и ассоциативные связи (ген–заболевание, ген–фенотип, биологический процесс–заболевание и др.). Отдельного внимания заслуживают связи типа «маркер», показывающие, что ген, биологический процесс или фенотипический признак служат индикатором связанного с ним заболевания или фенотипа. Кроме того, БЗ включает взаимодействия типа «фактор риска», когда ген, процесс, заболевание, фенотипический признак или иной объект рассматриваются как фактор риска для связанной с ним болезни (Ivanisenko V.A. et al., 2019).

Особенностью ANDSystem является наличие веб-ориентированного модуля ANDDigest, предназначенного для поиска и анализа публикаций PubMed с использованием онтологических словарей (Ivanisenko T.V. et al., 2020, 2022). Модуль поддерживает сложные запросы, в которых одновременно учитываются несколько типов объектов

из словарей ANDSystem, а также уточняющие ключевые слова, задаваемые пользователем напрямую. Результаты поиска отображаются в графическом виде с разметкой обнаруженных сущностей в тексте, возможностью сортировки и фильтрации (по дате, числу цитирований источников и другим параметрам), а также с визуализацией динамики числа упоминаний размеченных сущностей по годам и ссылками на внешние базы данных.

База знаний ANDSystem применялась для решения широкого круга задач на основе реконструкции и анализа генных сетей: реконструкции интерактома вируса гепатита С (Saik et al., 2016), приоритизации генов, связанных с восприимчивостью к туберкулезу (Bragina et al., 2016), системных исследований преэклампсии (Glotov et al., 2015), анализа коморбидности астмы и туберкулеза (Bragina et al., 2014), изучения эндотелиального апоптоза при лимфедеме (Saik et al., 2019), анализа экспрессии генов и протеомного профиля клинических штаммов *Helicobacter pylori*, ассоциированных с ранними стадиями рака желудка (Momynaliev et al., 2010), стабильности протеома в проекте «Марс-500» (Larina et al., 2015), интерпретации метаболомных данных в исследованиях послеоперационного делирия (Ivanisenko V.A. et al., 2024) и ответа меланомы на ТГц-излучение (Butikova et al., 2025). Применение ANDSystem для анализа метаболомных данных плазмы крови пациентов с COVID-19 позволило реконструировать генные сети, описывающие молекулярно-генетические пути влияния белков SARS-CoV-2 на метаболические нарушения при инфекции (Ivanisenko V.A. et al., 2022). Показано, что в таких сетях особую роль играют неструктурные белки коронавируса.

В настоящей работе мы использовали графовые нейронные сети с механизмом внимания (Veličković et al., 2017) для предсказания новых лигандов белка ORF3a среди метаболитов и лекарственных веществ, представленных в базе знаний ANDSystem. С помощью модели, обученной нами на графе знаний ANDSystem, предсказано пять малых молекул эндогенного и экзогенного происхождения, потенциально взаимодействующих с ORF3a.

1. N-ацетил-D-глюкозамин – мономер природного полисахарида хитина. Согласно данным молекулярного моделирования, он способен образовывать устойчивые комплексы с четырьмя белками SARS-CoV-2: Spike-белком (PDB ID: 6M0J), нуклеокапсидным фосфопротеином N (PDB ID: 6WKP), S-белком (PDB ID: 6X79) и протеазой 3CLpro (PDB ID: 7JVZ), а также потенциально вызывать иммунный ответ против вируса (Baysal et al., 2021; Tekin, 2023).
2. 4-(Бензоиламино)бензойная кислота – амидное производное бензойной кислоты. Это соединение демонстрирует антивирусную активность против флавивирусов (Gwon et al., 2020) и вируса лихорадки Рифт-Валли (Islam et al., 2018).
3. Аустоцистин D – поликетидный метаболит грибов рода *Aspergillus*, обладающий цитотоксической и антионкогенной активностью (Marks et al., 2011).
4. Биктегравир – низкомолекулярный ингибитор интегразы, применяемый для лечения ВИЧ-инфекции (Sax et al., 2023). Исследования показали его высокое сродство связывания с белком Spike (Ahsan, Sajib, 2021; Sun et al., 2021), а также с главной протеазой SARS-CoV-2 (Mpro, PDB ID: 6LU7) (Oner et al., 2023).
5. L-Треонин – незаменимая аминокислота, участвующая в белковом синтезе, гликозилировании и регуляции иммунного ответа. Имеются данные, что уровни L-треонина изменяются при различных вирусных инфекциях, включая COVID-19, что отражает метаболическую перестройку организма в ответ на инфекционный процесс (Barberis et al., 2020). В ряде работ отмечено, что профили аминокислот, в том числе треонина, могут служить биомаркерами тяжести течения COVID-19 и вовлечены в регуляцию воспалительных реакций и функций слизистых барьеров (Páez-Franco et al., 2021).

Молекулярный докинг и расчеты свободной энергии связывания показали, что наибольший интерес в качестве перспективных кандидатов для экспериментальной проверки представляют биктегравир и 4-(бензоиламино) бензойная кислота. Для биктегравира получены значения энергий связывания –7.37 ккал/моль (AutoDock Vina) и –14.71 ± 3.12 ккал/моль (MM/GBSA), что свидетельствует о более высокой аффинности по сравнению с 4-(бензоиламино)бензойной кислотой (–5.68 ккал/моль и –11.01 ± 3.58 ккал/моль соответственно). При этом биктегравир представляет особый интерес как кандидат для исследований по лекарственному репозиционированию.

Материалы и методы

Когнитивная система ANDSystem – платформа для автоматизированного извлечения фактов и знаний из текстов научных публикаций и фактографических баз данных, их интеграции в единую онтологическую модель (граф знаний) и реконструкции ассоциативных генных сетей (Ivanisenko V.A. et al., 2015, 2019). В графе знаний вершины соответствуют молекулярно-генетическим сущностям (гены, транскрипты РНК, белки, метаболиты, лекарственные вещества), а также объектам клеточного и организменного уровней (типы клеток, биологические процессы, заболевания, фенотипические признаки). Ребра представляют связи между сущностями, включая регуляторные отношения (влияние на экспрессию, активность, стабильность, транспорт и др.), физические контакты (взаимодействия белок–белок, белок–лиганд), химические отношения (каталитические реакции, посттрансляционные модификации и др.), а также ассоциативные связи (ген–заболевание, ген–фенотип, процесс–заболевание и др.). В текущей версии граф знаний ANDSystem содержит более 1.5 млн вершин и свыше 100 млн ребер.

Для распознавания имен биомедицинских сущностей и извлечения контекстно-зависимых связей в ANDSystem используется более 20 000 семантико-лингвистических шаблонов и правил; дополнительно применяются большие языковые модели, что повышает полноту и точность автоматического анализа текстовых источников. Для предсказания новых взаимодействий, в частности белок–белок, используются графовые нейронные сети, обученные на графе знаний ANDSystem, сформированном на основе научной литературы и специализированных баз данных (Ivanisenko T.V. et al., 2024).

В состав ANDSystem входит модуль ANDDigest – специализированная веб-система для поиска и анализа публикаций PubMed с опорой на онтологическую модель ANDSystem, использующая словари 13 типов биомедицинских объектов (Ivanisenko T.V. et al., 2020, 2022). База данных ANDDigest содержит индексированные и размеченные тексты публикаций PubMed, а также вычисленные характеристики и оценки статистической совстречаемости биомедицинских объектов, применяемые на последующих этапах анализа и извлечения знаний.

Получение векторных представлений вершин графа знаний ANDSystem. Для расчета векторных представлений вершин этого графа знаний ANDSystem использовалась графовая нейронная сеть с механизмом внимания (GAT) на основе TransformerConv (пакет PyTorch Geometric версии 2.5.3) (Fey, Lenssen, 2019). Архитектура сети включала четыре скрытых слоя по 256 нейронов в каждом. Каждая вершина графа знаний ANDSystem описывалась двоичным вектором длиной 13, в котором значение «1» указывало на принадлежность объекта к одному из 13 типов словарей, предусмотренных онтологией ANDSystem. Каждое ребро кодировалось вектором длиной 50: первые 49 ячеек вектора соответствовали различным типам взаимодействий и принимали значения «0» или «1» в зависимости от наличия данного типа связи между парой вершин в графе знаний, а последняя компонента содержала числовую оценку их совстречаемости (*p*-мера). Эта мера отражала статистическую значимость совместного упоминания пары объектов в аннотациях PubMed и вычислялась с помощью модуля ANDDigest. Итоговые векторные представления вершин, формируемые нейронной сетью, имели размерность 256.

Механизм внимания в каждом скрытом слое включал четыре независимые «головы» (компоненты), которые вычисляли вклад соседних вершин, – т. е. вершин, соединенных с рассматриваемой вершиной ребрами графа ANDSystem. При этом учитывались признаки как самих соседних вершин, так и ребер, связывающих их (типы связи и значение *p*-меры). В качестве функции потерь применялась логистическая функция (Mao et al., 2023) с температурным коэффициентом $\tau = 0.2$. Параметры оптимизировались методом AdamW (Zhou et al., 2024).

Ввиду больших размеров графа знаний ANDSystem для ускорения машинного обучения модель обучалась не на всем графе сразу, а на подграфах, автоматически формируемых на его основе. Для каждой целевой вершины строился подграф, включавший саму рассматриваемую вершину и ее соседей, находящихся на расстоянии не более чем трех ребер от нее. На каждом «уровне» соседства (т. е. на расстоянии 1, 2 или 3 ребра) число учитываемых соседних вершин ограничивалось: максимум 15 на первом уровне, 10 – на втором и 5 – на третьем. Эти соседи выбирались случайным образом.

Расчеты выполнялись на рабочей станции с шестью графическими ускорителями NVIDIA GeForce RTX 4090 (по 24 ГБ памяти каждый); все программы написаны на языке Python версии 3.12.11.

Полносвязная нейронная сеть. Для предсказания новых взаимодействий (ребер) между белками и метаболитами в графе знаний ANDSystem использована полносвязная нейронная сеть (многослойный перцептрон). Размер входного слоя соответствовал размерности векторного представления пары вершин (512). Архитектура модели включала три последовательных скрытых слоя с количеством нейронов 512, 256 и 128. В каждом скрытом слое применялась выпрямленная линейная функция активации (rectified linear unit, ReLU) (Glorot et al., 2011):

$$f(x) = \max(0, x).$$

Выходной слой содержал один нейрон, значение которого отражало вероятность наличия ребра между двумя вершинами. Для каждой пары вершин «белок–метаболит» нейронная сеть возвращала значение от 0 до 1, интерпретируемое как вероятность наличия взаимодействия между данной парой. Для классификации использовали стандартный порог 0.5: значения выше этого порога интерпретировались как наличие взаимодействия, а ниже – как его отсутствие (Harris, 2021).

Из графа знаний ANDSystem случайным образом было отобрано 250 000 пар объектов, в каждой из которых один участник имел тип «белок», а другой – «метаболит»; эти пары рассматривались как положительные примеры. В качестве отрицательных примеров формировалось равное число пар «белок–метаболит», случайно составленных из множества всех белков и метаболитов при условии отсутствия соответствующего ребра в исходном графе знаний.

Для каждой пары (u, v) строился составной вектор признаков длиной 512 (при размерности векторных представлений вершин $d = 256$), включавший четыре блока: 1) векторное представление белка e_u; 2) векторное представление метаболита e_v; 3) поэлементная абсолютная разность $|e_u - e_v|$; 4) поэлементное произведение $e_u \times e_v$.

Полученный массив векторов делился в соотношении 80, 10, 10 % на обучающую, валидационную и тестовую части соответственно, где обучающая часть использовалась для подбора параметров модели в ходе обучения, тестовая часть – для промежуточной оценки качества и выбора оптимальных гиперпараметров моделью, а валидационная применялась только для оценки точности финальной модели после ее обучения. В каждой части соотношение положительных и отрицательных примеров было 1:1.

Качество модели после каждой эпохи обучения (полной обработкой моделью всей обучающей выборки) оценивалось на тестовом наборе данных с помощью корреляционного коэффициента Мэттьюса (Matthews correlation coefficient, MCC) (Chicco, Jurman, 2020), по формуле:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}},$$

где *TP* (true positives, истинно положительные) – число верно классифицированных моделью пар объектов как взаимодействующие; *TN* (true negatives, истинно ложные) – число правильно классифицированных моделью пар объектов как невзаимодействующие; *FP* (false positive, ложноположительные) – число ошибочно классифицированных моделью пар объектов как взаимодействующие; *FN* (false negative, ложноотрицательные) – число ошибочно классифицированных моделью пар объектов как невзаимодействующие.

Обучение проводилось в течение 83 эпох; достигнутое значение MCC составило 0.9542, что свидетельствует о высокой точности модели. Реализация нейронной сети выполнена с использованием библиотеки PyTorch версии 2.4.1.

Молекулярный докинг применяли для первичной оценки аффинности по баллу докинга (Vina score) и построения моделей комплексов белок–лиганд. Используемый на данном этапе Vina score представляет собой эмпирическую оценку энергии связывания (ккал/моль); более отрицательные значения соответствуют более высокому предсказанному сродству. Расчеты выполняли с использованием AutoDock Vina 1.2.0 (Python API) (Trott, Olson, 2010; Eberhardt et al., 2021). Докинг проводили в слепом режиме (blind docking), задавая поисковую область, охватывающую всю поверхность белка ORF3a.

Наиболее энергетически выгодные конформации комплекса белок–лиганд (минимальные Vina score) использовали как исходные структуры для оценки свободной энергии связывания (ΔG) методом MM/GBSA.

Оценка MM/GBSA. Расчет ΔG выполняли в пакете AmberTools (Case et al., 2023). Метод учитывает энергии молекулярной механики и сольватационные вклады (обобщенная модель Борна) с неполярной составляющей, пропорциональной площади доступной растворителю поверхности, и дает приближенную термодинамическую характеристику устойчивости комплекса. Трехмерная структура белка ORF3a SARS-CoV-2 взята из БД Protein Data Bank (PDB ID: 6XDC).

Результаты

Предсказание новых взаимодействий белок–лиганд методом графовых нейронных сетей

Схема анализа, применяемая в ANDSystem для предсказания новых взаимодействий с использованием графовых нейронных сетей, представлена на рис. 1.

Из базы знаний ANDSystem была выгружена ассоциативная генная сеть человека на уровне полного генома. В сеть входили все 13 типов объектов (включая гены, белки, метаболиты, заболевания и др.) и 49 типов взаимодействий (регуляторные связи – регуляция экспрессии, активности, стабильности, транспорта и др.; физические взаимодействия – белок–белок, белок–лиганд и др.). Всего граф содержал около 310 тыс. вершин, соединенных 48 млн ребер. Для получения векторных представлений вершин графа знаний была обучена графовая нейронная сеть с механизмом внимания; качество по F1-мере = 0.8003 достигалось к 230-й эпохе.

На основе полученных векторных представлений белков и метаболитов графа знаний ANDSystem с целью предсказания отсутствующих в нем ребер проведено обучение многослойного перцептрона в форме двоичного классификатора. Обучение продолжалось 83 эпохи; достигнутое значение MCC составило 0.9542. Обученная модель была использована для предсказания ребер белок–метаболит для белка ORF3a. Всего проанализировано 38172 потенциальных связей данного белка с малыми молекулами эндогенного и экзогенного происхождения, включая метаболиты человека и других организмов, а также лекарственные вещества, неорганические молекулы и ионы. Обнаружено пять новых взаимодействий, отсутствовавших в базе знаний ANDSystem.

На рис. 2 продемонстрирована сеть взаимодействий ORF3a: черным цветом показаны ребра, изначально присутствовавшие в базе знаний ANDSystem, красным – новые связи, предсказанные графовой нейронной сетью и двоичной классификационной моделью. В базе знаний представлено 19 взаимодействий, извлеченных из научных публикаций, включая как прямые физические контакты, так и ассоциативные связи между ORF3a и малыми молекулами. Например, физические взаимодействия, экспериментально подтвержденные методами флуоресцентной и УФ-видимой спектроскопии, описаны для хлорина

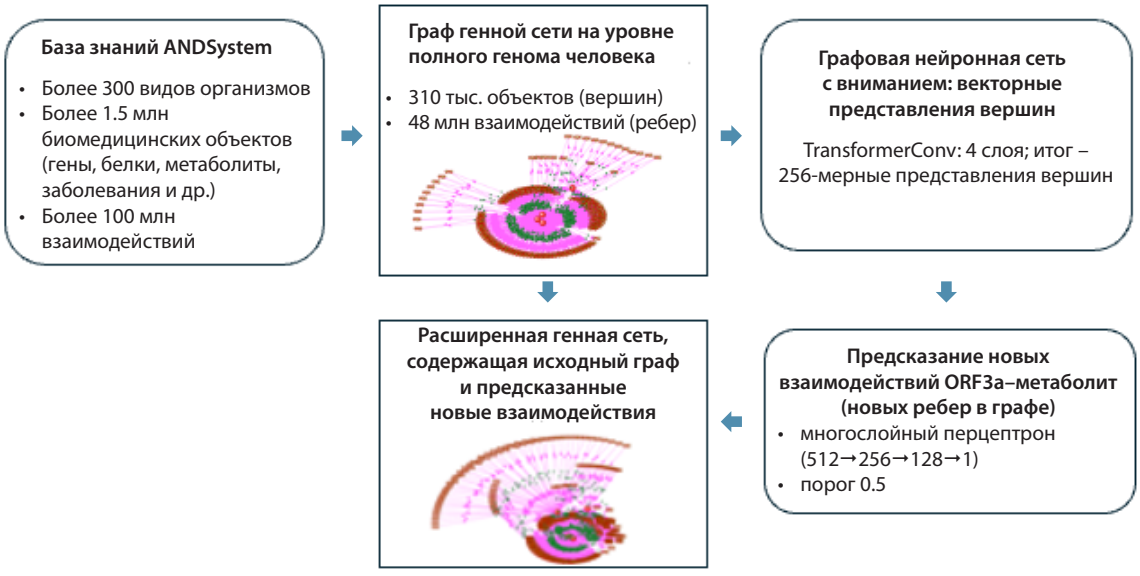


Рис. 1. Схематичное представление расчетного конвейера для предсказания новых взаимодействий между белками человека и метаболитами на основе анализа графа знаний ANDSystem.

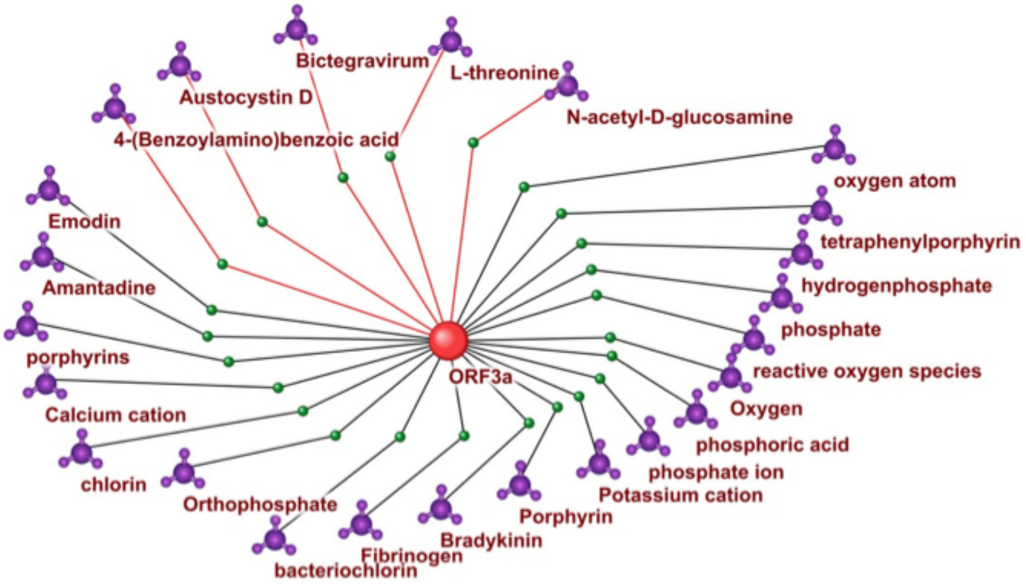


Рис. 2. Сеть взаимодействий ORF3a с малыми молекулами, реконструированная с использованием ANDSystem.

Темными линиями показаны взаимодействия, информация о которых получена из научных публикаций; красными линиями – взаимодействия, предсказанные графовой нейронной сетью: N-ацетил-D-глюкозамин (N-acetyl-D-glucosamine), 4-(бензоиламино)бензойная кислота (4-(benzoylamino)benzoic acid), аустоцистин D (austocystin D), биктегравир (bictegravirum) и L-треонин (L-threonine).

Таблица 1. Расчетные показатели связывания ORF3a–лиганд по данным AutoDock Vina и MM/GBSA

№ п/п	Лиганд	AutoDock Vina (ккал/моль)*	MM/GBSA (ккал/моль)**
1	Аустоцистин D	–8.296	–21.67 ± 2.3
2	Биктегравир	–7.368	–14.71 ± 3.12
3	N-Ацетил-D-глюкозамин	–6.242	–16.76 ± 2.58
4	4-(Бензоиламино)бензойная кислота	–5.682	–11.01 ± 3.58
5	L-Треонин	–4.89	–19.04 ± 2.15

* Расчетный балл AutoDock Vina (ккал/моль); ** ΔG-связывания (ккал/моль), оцененная методом MM/GBSA.

и катионных порфиринов; в той же работе молекулярный докинг показал образование комплексов для родственных порфиринов (bacteriochlorin, tetraphenylporphyrin, TPP) (Lebedeva et al., 2021). В качестве примера ассоциативной связи можно привести связь ORF3a–брадикинин, обсуждаемую в контексте усиления «брадикининового шторма» через взаимодействие ORF3a/NS7b при COVID-19 (Messina et al., 2021).

В группу предсказанных взаимодействий вошло пять кандидатов: N-ацетил-D-глюкозамин (N-acetyl-D-glucosamine; мономер хитина, предшественник гликозилирования); 4-(бензоиламино)бензойная кислота (4-(benzoylamino)benzoic acid; производное бензойной кислоты); аустоцистин D (austocystin D; поликетидный метаболит грибов рода *Aspergillus*); биктегравир (bictegravirum; ингибитор интегразы ВИЧ, лекарственный препарат); L-треонин (L-threonine; незаменимая аминокислота).

Молекулярный докинг и оценка энергии связывания

Для оценки способности пяти предсказанных малых молекул физически взаимодействовать с ORF3a мы выполнили молекулярный докинг с использованием AutoDock Vina и для полученных 3D моделей комплексов пересчитали свободную энергию связывания (ΔG) методом MM/GBSA (табл. 1). Балл докинга (Vina score), характеризующий эмпирическую оценку аффинности, применяли для относительного ранжирования лигандов, тогда как значения ΔG MM/GBSA рассматривали как приближенную термодинамическую характеристику стабильности комплексов.

По оценкам AutoDock Vina, наибольшее расчетное сродство показали аустоцистин D (−8.296 ккал/моль) и биктегравир (−7.368 ккал/моль); промежуточные – N-ацетил-D-глюкозамин (−6.242 ккал/моль) и 4-(бензоиламино) бензойная кислота (−5.682 ккал/моль); наименьшее сродство – L-треонин (−4.89 ккал/моль).

По оценкам MM/GBSA, минимальное (наиболее отрицательное) значение ΔG получено для аустоцистина D (−21.67 ± 2.30 ккал/моль); далее следуют L-треонин (−19.04 ± 2.15) и N-ацетил-D-глюкозамин (−16.76 ± 2.58), тогда как у биктегравира (−14.71 ± 3.12) и 4-(бензоиламино)бензойной кислоты (−11.01 ± 3.58) ΔG имели меньшее значение по модулю.

В совокупности баллы докинга (Vina score) и оценки ΔG методом MM/GBSA указывают на возможность фор-

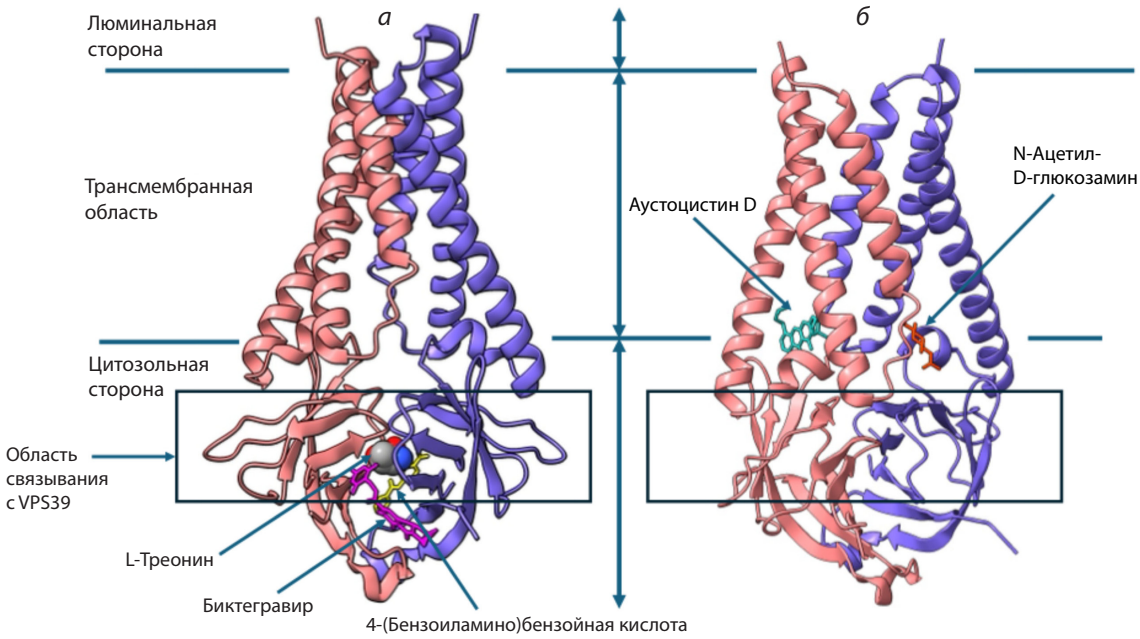


Рис. 3. Пространственные структуры комплексов ORF3a с анализируемыми лигандами.

а – комплекс ORF3a с L-треонином, биктегравиром и 4-(бензоиламино)бензойной кислотой; *б* – комплекс ORF3a с аустоцистином D и N-ацетил-D-глюкозамином. Белок показан в ленточной модели; две субъединицы димера окрашены разными цветами. В панели (*б*) структура белка повернута для более наглядного отображения лигандов. Лиганды показаны в палочковой модели (stick); их положение отмечено стрелками; L-треонин показан в шаровой модели (balls) для улучшения восприятия. Линиями обозначены участки белка, соответствующие его положению в мембране (Kern et al., 2021): цитозольная сторона, трансмембранная область и люминальная сторона (просвет аппарата Гольджи и эндо-/лизосом). Рамкой выделена область, участвующая во взаимодействии с белком VPS39. Изображения созданы в ChimeraX.

мирования комплексов ORF3a с проанализированными малыми молекулами, выступая взаимодополняющими критериями расчетной оценки сродства.

3D модели комплексов ORF3a с исследуемыми лигандами, построенные по результатам молекулярного докинга, представлены на рис. 3. По данным крио-ЭМ, ORF3a образует димер; каждая субъединица содержит три трансмембранные спирали и крупный цитозольный С-концевой домен (Kern et al., 2021). ORF3a преимущественно локализуется на мембранах аппарата Гольджи, эндосом и лизосом, участвуя в регуляции везикулярного транспорта и лизосомного экзоцитоза; также обнаруживается на плазматической мембране (Hinkle et al., 2025).

Известно, что ORF3a взаимодействует с VPS39 (комплекс HOPS) и блокирует слияние аутофагосом с лизосомами, что приводит к накоплению неслившихся аутофагосом и способствует уклонению вируса от деградации (Zhang J. et al., 2021; Miller et al., 2023). Для наглядности соответствующий участок белка, вовлеченный во взаимодействие с VPS39, на рис. 3 выделен рамкой.

Согласно результатам докинга, сайты связывания L-треонина, биктегравира и 4-(бензоиламино)бензойной кислоты расположены на цитозольной поверхности димера и частично перекрываются с участком связывания ORF3a–VPS39 (см. рис. 3, *а*). При этом L-треонин связывается в интерфейсе контакта субъединиц ORF3a (межсубъединичной щели), глубоко погружен в нее и практически недоступен растворителю. Биктегравир и 4-(бензоиламино)бензойная кислота занимают участки поверхности белка, доступные для растворителя (рис. 4). Аустоцистин D и N-ацетил-D-глюкозамин связываются на границе между цитозольной поверхностью и трансмембранным доменом (см. рис. 3, *б*).

Детализация водородных (H-) и гидрофобных контактов между лигандами и аминокислотными остатками белка ORF3a приведена в табл. 2 и иллюстрирована на рис. 5. N-ацетил-D-глюкозамин образует множественные H-связи с остатками Lys61, Ile63, Thr64, Arg126 и др. 4-(бензоиламино)бензойная кислота формирует H-связи с Ser165 и Asp226, а также гидрофобные контакты с Val225 и Val228. Аустоцистин D образует H-связи с Ser165, Glu226, His227 и Asn234 и гидрофобные контакты с His227. Биктегравир образует три H-связи (Ser165, Glu226, Asn234). L-треонин, располагаясь глубоко в межсубъединичной щели на интерфейсе димера, формирует множественные H-связи (с шестью остатками) и гидрофобные контакты с Ile186.

Обсуждение

Опираясь на нашу предыдущую работу с GraphSAGE для предсказания взаимодействий белок–белок (Ivanisenko T.V. et al., 2024), в данном исследовании мы применили графовую нейронную сеть с механизмом внимания для задачи предсказания взаимодействий белка SARS-CoV-2 ORF3a с малыми молекулами на графе знаний ANDSystem и получили пять кандидатов-лигандов: N-ацетил-D-глюкозамин, 4-(бензоиламино)бензойную кислоту, аустоцистин D, биктегравир и L-треонин.

В отличие от архитектуры GraphSAGE, модели с вниманием обновляют представления вершин с учетом их соседей, взвешивая вклад каждого: более информатив-

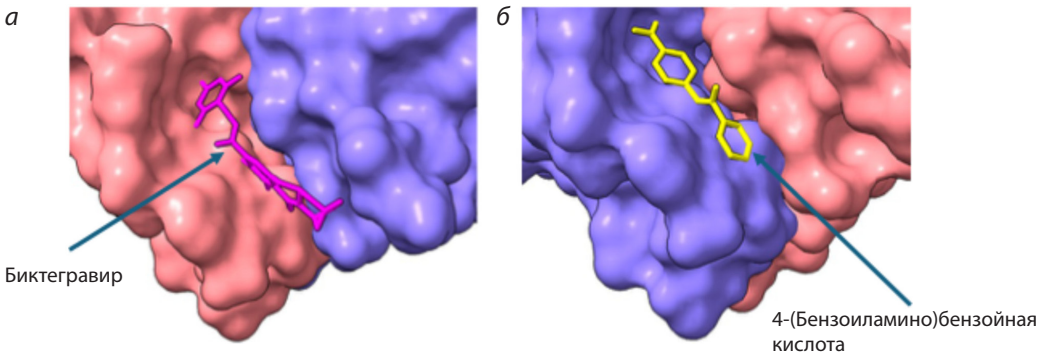


Рис. 4. Поверхность ORF3a при связывании биктегравира (*а*) и 4-(бензоиламино)бензойной кислоты (*б*).
Изображение создано в ChimeraX.

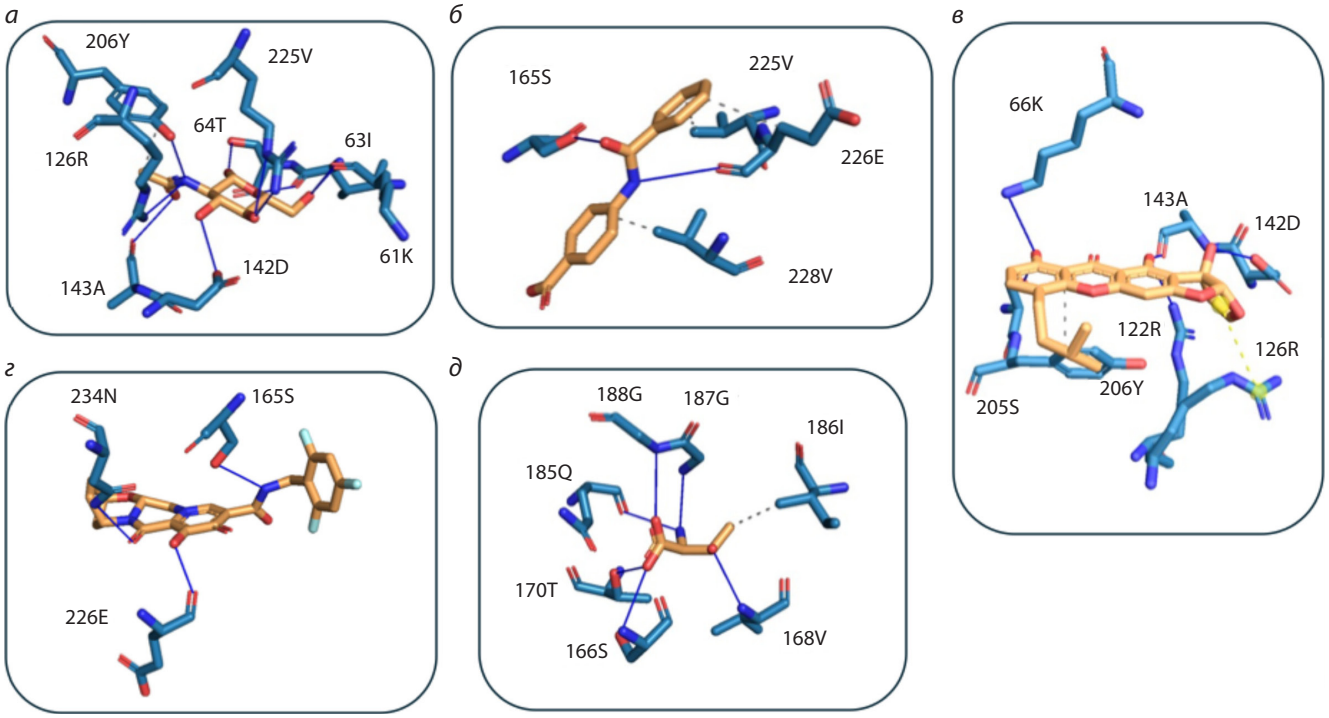


Рис. 5. Детализированное представление взаимодействий анализируемых лигандов с аминокислотными остатками ORF3a.
а – N-ацетил-D-глюкозамин; *б* – 4-(бензоиламино)бензойная кислота; *в* – аустоцистин D; *г* – биктегравир; *д* – L-треонин. Лиганд показан желтым цветом, аминокислотные остатки – синим. Водородные связи обозначены сплошными линиями, гидрофобные взаимодействия – пунктирными. Изображения получены в PyMOL.

ные связи получают больший вес, менее информативные – меньший; несколько «голов внимания» работают параллельно, после чего их результаты агрегируются в итоговый вектор вершины, что позволяет точнее учитывать локальный контекст графа (Wu et al., 2021).

Для проверки этих предсказаний мы выполнили молекулярный докинг и оценили свободную энергию связывания (ΔG) комплексов белок–лиганд методом MM/GBSA. Расчеты показали, что сайты связывания аустоцистина D и N-ацетил-D-глюкозамина располагаются у границы между цитозольной поверхностью и трансмембранным доменом ORF3a, тогда как L-треонин, биктегравир и 4-(бензоиламино)бензойная кислота связываются на цитозольной стороне димера; при этом участки связывания биктегравира и 4-(бензоиламино)бензойной кислоты частично перекрываются с регионом взаимодействия ORF3a–VPS39.

Взаимодействие ORF3a с хозяйским белком VPS39, субъединицей комплекса сортировки белков вакуоли (HOPS), регулирующего поздние этапы слияния эндосомно-лизосомных компартментов, хорошо охарактеризовано (Zhang J. et al., 2021; Miller et al., 2023). Оно препятствует слиянию аутофагосом и поздних эндосом с лизосомами и тем самым подавляет аутофагический поток – ключевой путь деградации вирусных компонентов.

Функциональную значимость интерфейса взаимодействия между белками ORF3a и VPS39 подтверждает наличие в цитозольном домене ORF3a мотива YXXФ (Y – ти-

Таблица 2. Молекулярные взаимодействия белка ORF3a с лигандами, полученные по результатам анализа реконструированных комплексов ORF3a–лиганд с помощью веб-сервера PLIP (Protein–Ligand Interaction Profiler)

Лиганд	Номер аминокислотного остатка*	Аминокислотный остаток**	Расстояние, Å	Тип связи
Аустоцистин D	165B	SER	2.45	H-bond
	226A	GLU	3.70	
	227A	HIS	2.83	
	234B	ASN	2.85	
	227A	HIS	3.59	Hydrophobic
Биктегравир	165A	SER	2.44	H-bond
	226B	GLU	2.20	
	234A	ASN	2.35	
N-Ацетил-D-глюкозамин	61B	LYS	2.07	H-bond
	63B	ILE	2.17	
	63B	ILE	2.69	
	64B	THR	3.15	
	122A	ARG	2.07	
	122A	ARG	2.52	
	126A	ARG	2.36	
	142A	ASP	2.45	
	143A	ALA	3.03	
	206A	TYR	2.87	
	206A	TYR	3.69	Hydrophobic
4-(Бензоиламино) бензойная кислота	165B	SER	2.53	H-bond
	226A	GLU	3.36	
	225A	VAL	3.93	Hydrophobic
	226A	GLU	3.69	
	228A	VAL	3.55	
L-Треонин	166B	SER	3.28	H-bond
	168B	VAL	2.25	
	170A	THR	2.08	
	185A	GLN	2.34	
	187A	GLY	2.36	
	188A	GLY	2.85	
	186B	ILE	3.69	Hydrophobic

* Нумерация аминокислотных остатков приведена по последовательности ORF3a; идентификатор цепи указан по структуре PDB ID: 6XDC.
** Указана аминокислота, участвующая во взаимодействии.

розин, X – произвольная аминокислота, Ф – гидрофобный остаток).

В ORF3a этот мотив представлен последовательностью YNSV (позиции 160–163). В работах (Zhang J. et al., 2021; Miller et al., 2023) показано, что точечная замена Y160A нарушает коиммунопреципитацию ORF3a с VPS39 и снимает блокаду HOPS-зависимого слияния, частично восстанавливая аутофагический поток.

Можно предположить, что предсказанное расположение сайтов для биктегравира и 4-(бензоиламино)бензойной кислоты потенциально способно влиять на формирование/стабильность комплекса ORF3a–VPS39, что делает их перспективными кандидатами для функционального вмешательства в HOPS-зависимый этап слияния аутофагосом и лизосом.

По совокупности метрик (Vina score и ΔG по MM/GBSA) биктегравир демонстрирует более отрицательные значения, указывающие на более высокое расчетное сродство, по сравнению с 4-(бензоиламино)бензойной кислотой (Vina score −7.37 ккал/моль и ΔG по MM/GBSA

−14.71 ± 3.12 ккал/моль против −5.68 ккал/моль и −11.01 ± 3.58 ккал/моль соответственно). Кроме того, это зарегистрированный ингибитор интегразы ВИЧ (препарат Biktarvy) (Gallant et al., 2017), что делает его перспективным кандидатом для репозиционирования. Потенциальный механизм действия биктегравира в качестве лекарственного препарата для лечения COVID-19 может заключаться в ингибировании взаимодействия ORF3a с белком хозяина VPS39, что, в свою очередь, позволит нейтрализовать способность ORF3a блокировать слияние эндосомно-лизосомных компартментов и будет способствовать деградации вирусных компонентов в лизосомах. 4-(Бензоиламино)бензойная кислота может быть интересна как ароматический карбоксамид-фрагмент для таргетирования в структуре ORF3a интерфейсов белок-белковых взаимодействий (Marks et al., 2011).

Заключение

Использованный нами подход, включающий предсказание новых взаимодействий белок–лиганд на графе знаний ANDSystem с последующим молекулярным докингом и оценкой ΔG связывания методом MM/GBSA, позволил выделить перспективные низкомолекулярные лиганды-кандидаты для белка ORF3a SARS-CoV-2. Среди отобранных соединений наибольший интерес представляют биктегравир и 4-(бензоиламино)бензойная кислота: их предсказанные сайты расположены на цитозольной поверхности ORF3a и частично перекрываются с регионом взаимодействия ORF3a–VPS39. По энергетическим оценкам биктегравир демонстрирует более отрицательные значения Vina score и ΔG – AutoDock Vina: −7.37 ккал/моль; MM/GBSA: −14.71 ± 3.12 ккал/моль. Для 4-(бензоиламино)бензойной кислоты получены сопоставимые значения, но меньшие величины по модулю: −5.68 ккал/моль и −11.01 ± 3.58 ккал/моль соответственно.

Ограничением работы является отсутствие явного учета липидного бислоя: расчеты проводились без встраивания белка в мембрану, что может влиять на конформацию ORF3a и на энергетические вклады, связанные с проникновением лиганда в гидрофобную среду. Следующим шагом продолжения работы может быть молекулярная динамика в мембранной модели с пересчетом энергий связывания с последующей экспериментальной проверкой результатов.

В совокупности полученные *in silico* результаты позволяют рассматривать биктегравир в качестве приоритетного кандидата для экспериментальных исследований взаимодействия с ORF3a, в том числе в рамках лекарственного репозиционирования, и служат основой для дальнейшей оптимизации малых молекул, нацеленных на этот белок.

Список литературы / References

Ahsan T., Sajib A.A. Repurposing of approved drugs with potential to interact with SARS-CoV-2 receptor. *Biochem Biophys Rep.* 2021;26:100982. doi 10.1016/j.bbrep.2021.100982

Ali S.I.M., Alrashid S.Z. A review of methods for gene regulatory networks reconstruction and analysis. *Artif Intell Rev.* 2025;58:256. doi 10.1007/s10462-025-11257-z

Barberis E., Timo S., Amede E., Vanella V.V., Puricelli C., Cappellano G., Raineri D., … Rolla R., Chiocchetti A., Baldanzi G., Marengo E., Manfredi M. Large-scale plasma analysis revealed new mechanisms and molecules associated with the host response to SARS-CoV-2. *Int J Mol Sci.* 2020;21(22):8623. doi 10.3390/ijms21228623

Baysal Ö., Abdul Ghafoor N., Silme R.S., Ignatov A.N., Kniazeva V. Molecular dynamics analysis of N-acetyl-D-glucosamine against specific SARS-CoV-2's pathogenicity factors. *PLoS One.* 2021;16(5):e0252571. doi 10.1371/journal.pone.0252571

Boby M.L., Fearon D., Ferla M., Filep M., Koekemoer L., Robinson M.C., COVID Moonshot Consortium, … Zaidmann D., Zhang I., Zidane H., Zitzmann N., Zvornicanin S.N. Open science discovery of potent non-covalent SARS-CoV-2 main protease inhibitors. *Science.* 2023;380(6640):eabo7201. doi 10.1126/science.abo7201

Bragina E.Y., Tiys E.S., Freidin M.B., Koneva L.A., Demenkov P.S., Ivanisenko V.A., Kolchanov N.A., Puzyrev V.P. Insights into pathophysiology of dystropy through the analysis of gene networks: an example of bronchial asthma and tuberculosis. *Immunogenetics.* 2014;66(7-8):457-465. doi 10.1007/s00251-014-0786-1

Bragina E.Y., Tiys E.S., Rudko A.A., Ivanisenko V.A., Freidin M.B. Novel tuberculosis susceptibility candidate genes revealed by the reconstruction and analysis of associative networks. *Infect Genet Evol.* 2016;46:118-123. doi 10.1016/j.meegid.2016.10.030

Butikova E.A., Basov N.V., Rogachev A.D., Gaisler E.V., Ivanisenko V.A., Demenkov P.S., Makarova A.-L.A., … Pokrovsky A.G., Vinokurov N.A., Kanygin V.V., Popik V.M., Shevchenko O.A. Metabolomic and gene networks approaches reveal the role of mitochondrial membrane proteins in response of human melanoma cells to THz radiation. *Biochim Biophys Acta Mol Cell Biol Lipids.* 2025;1870(2):159595. doi 10.1016/j.bbalip.2025.159595

Case D.A., Aktulga H.M., Belfon K., Cerutti D.S., Andrés Cisneros G., Cruzeiro V.W.D., Forouzesh N., … Roitberg A., Simmerling C.S., York D.M., Nagan M.C., Merz K.M. Jr. AmberTools. *J Chem Inf Model.* 2023;63(20):6183-6191. doi 10.1021/acs.jcim.3c01153

Chicco D., Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21:6. doi 10.1186/s12864-019-6413-7

Clarivate. MetaBase & MetaCore: Early Research Intelligence Solutions. Available at: http://clarivate.com/life-sciences-healthcare/research-development/discovery-development/early-research-intelligence-solutions/

Dong M., Galvan Achi J.M., Du R., Rong L., Cui Q. Development of SARS-CoV-2 entry antivirals. *Cell Insight.* 2024;3(1):100144. doi 10.1016/j.cellin.2023.100144

Eberhardt J., Santos-Martins D., Tillack A.F., Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J Chem Inf Model.* 2021;61(8):3891-3898. doi 10.1021/acs.jcim.1c00203

Fey M., Lenssen J.E. Fast graph representation learning with PyTorch Geometric. *arXiv*. 2019. doi 10.48550/arXiv.1903.02428

Gallant J.E., Thompson M., DeJesus E., Voskuhl G.W., Wei X., Zhang H., Martin H. Antiviral activity, safety, and pharmacokinetics of bictegravir as 10-day monotherapy in HIV-1-infected adults. *J Acquir Immune Defic Syndr.* 2017;75(1):61-66. doi 10.1097/QAI.0000000000001306

Gaudelet T., Day B., Jamasb A.R., Soman J., Regep C., Liu G., Hayter J.B.R., Vickers R., Roberts C., Tang J., Roblin D., Blundell T.L., Bronstein M.M., Taylor-King J.P. Utilizing graph machine learning within drug discovery and development. *Brief Bioinform.* 2021;22(6):bbab159. doi 10.1093/bib/bbab159

Glorot X., Bordes A., Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS). 2011;315-323. Available at: https://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf

Glotov A.S., Tiys E.S., Vashukova E.S., Pakin V.S., Demenkov P.S., Saik O.V., Ivanisenko T.V., Arzhanova O.N., Mozgovaya E.V., Zai-

nulina M.S., Kolchanov N.A., Baranov V.S., Ivanisenko V.A. Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). *BMC Syst Biol.* 2015;9(Suppl. 2):S4. doi 10.1186/1752-0509-9-S2-S4

Gwon Y.-D., Strand M., Lindqvist R., Nilsson E., Saleeb M., Elofsson M., Överby A.K., Evander M. Antiviral activity of benzavir-2 against emerging flaviviruses. *Viruses.* 2020;12(3):351. doi 10.3390/v12030351

Harris J.K. Primer on binary logistic regression. *Fam Med Community Health.* 2021;9(Suppl. 1):e001290. doi 10.1136/fmch-2021-001290

Hinkle J.J., Trychta K.A., Wires E.S., Osborn R.M., Leach J.R., Faraz Z.F., Svarcbahs R., Richie C.T., Dewhurst S., Harvey B.K. Subcellular localization of SARS-CoV-2 E and 3a proteins along the secretory pathway. *J Mol Histol.* 2025;56(2):98. doi 10.1007/s10735-025-10375-w

Islam M., Strand M., Saleeb M., Svensson R., Baranczewski P., Artursson P., Wadell G., Ahlm C., Elofsson M., Evander M. Anti-Rift Valley fever virus activity *in vitro*, pre-clinical pharmacokinetics and oral bioavailability of benzavir-2, a broad-acting antiviral compound. *Sci Rep.* 2018;8:1925. doi 10.1038/s41598-018-20362-9

Ivanisenko T.V., Saik O.V., Demenkov P.S., Ivanisenko N.V., Savostianov A.N., Ivanisenko V.A. ANDDigest: a new web-based module of ANDSystem for the search of knowledge in the scientific literature. *BMC Bioinformatics.* 2020;21(Suppl 11):228. doi 10.1186/s12859-020-03557-8

Ivanisenko T.V., Demenkov P.S., Kolchanov N.A., Ivanisenko V.A. The new version of the ANDDigest tool with improved AI-based short names recognition. *Int J Mol Sci.* 2022;23(23):14934. doi 10.3390/ijms232314934

Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *Int J Mol Sci.* 2024;25(21):11811. doi 10.3390/ijms252111811

Ivanisenko V.A., Saik O.V., Ivanisenko N.V., Tiys E.S., Ivanisenko T.V., Demenkov P.S., Kolchanov N.A. ANDSystem: an associative network discovery system for automated literature mining in the field of biology. *BMC Syst Biol.* 2015;9(Suppl. 2):S2. doi 10.1186/1752-0509-9-S2-S2

Ivanisenko V.A., Demenkov P.S., Ivanisenko T.V., Mishchenko E.L., Saik O.V. A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics.* 2019;20(1):34. doi 10.1186/s12859-018-2567-6

Ivanisenko V.A., Gaisler E.V., Basov N.V., Rogachev A.D., Cheresiz S.V., Ivanisenko T.V., Demenkov P.S., … Karpenko T.N., Velichko A.J., Voevoda M.I., Kolchanov N.A., Pokrovsky A.G. Plasma metabolomics and gene regulatory networks analysis reveal the role of nonstructural SARS-CoV-2 viral proteins in metabolic dysregulation in COVID-19 patients. *Sci Rep.* 2022;12(1):19977. doi 10.1038/s41598-022-24170-0

Ivanisenko V.A., Rogachev A.D., Makarova A.A., Basov N.V., Gaisler E.V., Kuzmicheva I.N., Demenkov P.S., … Kolchanov N.A., Plesko V.V., Moroz G.B., Lomivorotov V.V., Pokrovsky A.G. AI-assisted identification of primary and secondary metabolomic markers for postoperative delirium. *Int J Mol Sci.* 2024;25(21):11847. doi 10.3390/ijms252111847

Kern D.M., Sorum B., Mali S.S., Hoel C.M., Sridharan S., Remis J.P., Toso D.B., Kotecha A., Bautista D.M., Brohawn S.G. Cryo-EM structure of SARS-CoV-2 ORF3a in lipid nanodiscs. *Nat Struct Mol Biol.* 2021;28(7):573-582. doi 10.1038/s41594-021-00619-0

Krämer A., Green J., Pollard J. Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics.* 2014;30(4):523-530. doi 10.1093/bioinformatics/btt703

Larina I.M., Pastushkova L.Kh., Tiys E.S., Kireev K.S., Kononikhin A.S., Starodubtseva N.L., Popov I.A., Custaud M.-A., Dobrokhotov I.V., Nikolaev E.N., Kolchanov N.A., Ivanisenko V.A. Permanent proteins in the urine of healthy humans during the Mars-500 experiment. *J Bioinform Comput Biol.* 2015;13(1):1540001. doi 10.1142/S0219720015400016

Lebedeva N.S., Gubarev Y.A., Mamardashvili G.M., Zaitceva S.V., Zdanovich S.A., Malyasova A.S., Romanenko J.V., Koifman M.O., Koifman O.I. Theoretical and experimental study of interaction of macroheterocyclic compounds with ORF3a of SARS-CoV-2. *Sci Rep.* 2021;11:19481. doi 10.1038/s41598-021-99072-8

Mao A., Mohri M., Zhong Y. Cross-entropy loss functions: theoretical analysis and applications. In: International Conference on Machine Learning (ICML). 2023;23803-23828. Available at: https://proceedings.mlr.press/v202/mao23b/mao23b.pdf

Marks K.M., Park E.S., Arefolov A., Russo K., Ishihara K., Ring J.E., Clardy J., Clarke A.S., Pelish E.P. The selectivity of austocystin D arises from cell-line-specific drug activation by cytochrome P450 enzymes. *J Nat Prod.* 2011;74(4):567-573. doi 10.1021/np100429s

Messina F., Giombini E., Montaldo C., Sharma A.A., Zoccoli A., Sekaly R.P., Locatelli F., Zumla A., Maeurer M., Capobianchi M.R., Lauria F.N., Ippolito G. Looking for pathways related to COVID-19: confirmation of pathogenic mechanisms by SARS-CoV-2-host interactome. *Cell Death Dis.* 2021;12(8):788. doi 10.1038/s41419-021-03881-8

Miller A.N., Houlihan P.R., Matamala E., Cabezas-Bratesco D., Lee G.Y., Cristofori-Armstrong B., Dilan T.L., Sanchez-Martinez S., Matthies D., Yan R., Yu Z., Ren D., Brauchi S.E., Clapham D.E. The SARS-CoV-2 accessory protein Orf3a is not an ion channel. *eLife*. 2023;12:e84477. doi 10.7554/eLife.84477

Momynaliev K.T., Kashin S.V., Chelysheva V.V., Selezneva O.V., Demina I.A., Serebryakova M.V., Alexeev D., Ivanisenko V.A., Aman E., Govorun V.M. Functional divergence of *Helicobacter pylori* related to early gastric cancer. *J Proteome Res.* 2010;9(1):254-267. doi 10.1021/pr900586w

Naqvi A.A.T., Fatima K., Muhammad T., Fatima U., Singh I.K., Singh A., Atif S.M., Hariprasad G., Hasan G.M., Hassan M.I. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies. *Int J Biol Sci.* 2020;16(10):1708-1724. doi 10.7150/ijbs.45127

Ng T.I., Correia I., Seagal J., DeGoey D.A., Schrimpf M.R., Hardee D.J., Noey E.L., Kati W.M. Antiviral drug discovery for the treatment of COVID-19 infections. *Viruses.* 2022;14(5):961. doi 10.3390/v14050961

Nicholson D.N., Greene C.S. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J.* 2020;18:1414-1421. doi 10.1016/j.csbj.2020.05.017

Oner E., Demirhan I., Miraloglu M., Yalin S., Kurutas E.B. Investigation of antiviral substances in COVID-19 by molecular docking: in silico study. *Afr Health Sci.* 2023;23(1):23-36. doi 10.4314/ahs.v23i1.4

Páez-Franco J.C., Torres-Ruiz J., Sosa-Hernández V.A., Cervantes-Díaz R., Romero-Ramírez S., Pérez-Fragoso A., Meza-Sánchez D.E., Germán-Acacio J.M., Maravillas-Montero J.L., Mejía-Domínguez N.R., Ponce-de-León A., Ulloa-Aguirre A., Gómez-Martín D., Llorente L. Metabolomics analysis reveals a modified amino acid metabolism that correlates with altered oxygen homeostasis in COVID-19 patients. *Sci Rep.* 2021;11(1):6350. doi 10.1038/s41598-021-85788-0

Saik O.V., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.* 2016;218:40-48. doi 10.1016/j.virusres.2015.12.003

Saik O.V., Nimaev V.V., Usmonov D.B., Demenkov P.S., Ivanisenko T.V., Lavrik I.N., Ivanisenko V.A. Prioritization of genes involved in endothelial cell apoptosis by their implication in lymphedema using an analysis of associative gene networks with ANDSystem. *BMC Med Genomics.* 2019;12(Suppl. 2):117. doi 10.1186/s12920-019-0492-9

Sax P.E., Arribas J.R., Orkin C., Lazzarin A., Pozniak A., DeJesus E., Maggiolo F., … Hindman J.T., Martin H., Baeten J.M., Wohl D.; GS-

US-380-1489 and GS-US-380-1490 study investigators. bictegravir/emtricitabine/tenofovir alafenamide as initial treatment for HIV-1: five-year follow-up from two randomized trials. *EClinicalMedicine.* 2023;59:101991. doi 10.1016/j.eclinm.2023.101991

Stephens E.B., Kunec D., Henke W., Vidal R.M., Greishaber B., Saud R., Kalamvoki M., Singh G., Kafle S., Trujillo J.D., Ferreyra F.M., Morozov I., Richt J.A. The role of the tyrosine-based sorting signals of the ORF3a protein of SARS-CoV-2 in intracellular trafficking and pathogenesis. *Viruses*. 2025;17(4):522. doi 10.3390/v17040522

Stokes J.M., Yang K., Swanson K., Jin W., Cubillos-Ruiz A., Donghia N.M., MacNair C.R., … Church G.M., Brown E.D., Jaakkola T.S., Barzilay R., Collins J.J. A deep learning approach to antibiotic discovery. *Cell.* 2020;180(4):688-702. doi 10.1016/j.cell.2020.01.021

Sun C., Zhang J., Wei J., Zheng X., Zhao X., Fang Z., Xu D., Yuan H., Liu Y. Screening, simulation, and optimization design of small-molecule inhibitors of the SARS-CoV-2 spike glycoprotein. *PLoS One.* 2021;16(1):e0245975. doi 10.1371/journal.pone.0245975

Szklarczyk D., Kirsch R., Koutrouli M., Nastou K., Mehryary F., Hachilif R., Gable A.L., Fang T., Doncheva N.T., Pyysalo S., Bork P., Jensen L.J., von Mering C. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*. 2023; 51(D1):D638-D646. doi 10.1093/nar/gkac1000

Tekin E.D. Investigation of the effects of N-acetylglucosamine on the stability of the spike protein in SARS-CoV-2 by molecular dynamics simulations. *Comput Theor Chem.* 2023;1222:114049. doi 10.1016/j.comptc.2023.114049

Trott O., Olson A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* 2010;31(2):455-461. doi 10.1002/jcc.21334

U.S. Food and Drug Administration (FDA). FDA approves first treatment for COVID-19 (Veklury/remdesivir): press release. 22 Oct 2020. Available at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-covid-19

U.S. Food and Drug Administration. FDA approves first oral antiviral for treatment of COVID-19 in adults: press announcement. 2023-05-25. Available at: https://www.fda.gov/news-events/press-announcements/fda-approves-first-oral-antiviral-treatment-covid-19-adults

Veličković P., Cucurull G., Casanova A., Romero A., Liò P., Bengio Y. Graph attention networks. *arXiv.* 2017. doi 10.48550/arXiv.1710.10903

von Delft A., Hall M.D., Kwong A.D., Purcell L.A., Saikatendu K.S., Schmitz U., Tallarico J.A., Lee A.A. Accelerating antiviral drug discovery: lessons from COVID-19. *Nat Rev Drug Discov.* 2023;22(7):585-603. doi 10.1038/s41573-023-00692-8

Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P.S. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst.* 2021;32(1):4-24. doi 10.1109/TNNLS.2020.2978386

Zhang J., Cruz-Cosme R., Zhuang M.W., Liu D., Liu Y., Teng S., Wang P.-H., Tang Q. A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins. *Signal Transduct Target Ther.* 2021;6(1):192. doi 10.1038/s41392-021-00564-w

Zhang J., Ejikemeuwa A., Gerzanich V., Nasr M., Tang Q., Simard J.M., Zhao R.Y. Understanding the role of SARS-CoV-2 ORF3a in viral pathogenesis and COVID-19. *Front Microbiol.* 2022;13:854567. doi 10.3389/fmicb.2022.854567

Zhang Y., Sun H., Pei R., Mao B., Zhao Z., Li H., Lin Y., Lu K. The SARS-CoV-2 protein ORF3a inhibits fusion of autophagosomes with lysosomes. *Cell Discov.* 2021;7:31. doi 10.1038/s41421-021-00268-z

Zhou P., Xie X., Lin Z., Yan S. Towards understanding convergence and generalization of AdamW. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(9):6486-6493. doi 10.1109/TPAMI.2024.3382294

Zitnik M., Agrawal M., Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics.* 2018; 34(13):i457-i466. doi 10.1093/bioinformatics/bty294

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 25.09.2025. После доработки 24.10.2025. Принята к публикации 27.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1097-1108

doi 10.18699/vjgb-25-114

Влияние димерных бисбензимидазолов на активность ферментов репарации ДНК тирозил-ДНК-фосфодиэстераз 1 и 2 и поли(АДФ-рибоза)полимераз 1 и 2

Н.С. Дырхеева [1], И.А. Чернышова [1], А.Ф. Арутюнян [2], А.Л. Захаренко [1], М.М. Кутузов [1], К.Н. Науменко [1], А.С. Вензель [3], В.А. Иванисенко [3], С.М. Деев [4], А.Л. Жузе [2], О.И. Лаврик [1]

[1] Институт химической биологии и фундаментальной медицины Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт молекулярной биологии им. В.А. Энгельгардта Российской академии наук, Москва, Россия
[3] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук, Москва, Россия
lavrik@1bio.ru

Аннотация. Онкологические заболевания остаются одной из главных причин патологической смертности в мире, что определяет дизайн противораковых препаратов как ключевое направление медицинской химии. Комбинация ингибиторов ферментов репарации ДНК с ингибиторами топоизомераз – перспективный подход для усиления противоракового действия и снижения токсичности химиотерапии. Особый интерес представляют узкобороздочные ДНК-лиганды, способные эффективно ингибировать ДНК-зависимые ферменты, обладая при этом низкой токсичностью и мутагенностью. Ряд исследовательских групп, включая нашу, разрабатывает ингибиторы ферментов репарации ДНК, действующие одновременно на несколько взаимосвязанных мишеней {тирозил-ДНК-фосфодиэстеразы 1/2 (TDP1/TDP2), поли(АДФ-рибоза)полимераза 1 (PARP1)/TDP1, топоизомераза 1 (TOP1)/TDP1}. Такие бифункциональные ингибиторы призваны решить проблему резистентности опухолевых клеток к известным химиопрепаратам и повысить эффективность последних. В настоящем исследовании представлены данные скрининга ингибирующей активности 22 узкобороздочных лигандов, взаимодействующих с ДНК, – бис- и трисбензимидазолов – в отношении четырех ферментов репарации: TDP1, TDP2, PARP1 и PARP2. Изучены четыре серии димерных соединений и их мономерных единиц. Показана разница в ингибирующей активности димерных бисбензимидазолов в зависимости от структуры соединения и фермента. Мономерные и димерные бисбензимидазолы эффективно ингибируют активность TDP1 в микромолярном и субмикромолярном диапазоне IC_{50} (концентрация полумаксимального ингибирования). Димерные соединения групп $DB_2Py(n)$ и $DB_3P(n)$ проявили более значительную ингибирующую активность в отношении ферментативной реакции с участием TDP1 по сравнению с мономерами, входящими в их состав. Для всех исследованных соединений была показана низкая ингибирующая способность в отношении остальных трех ферментов репарации ДНК, т. е. наблюдается их специфическое воздействие именно на TDP1. Следует отметить, что в данной работе в экспериментах с TDP1 и TDP2 было исключено действие исследуемых соединений как узкобороздочных лигандов, связывающихся с ДНК, и исследовано их непосредственное воздействие на фермент. По результатам молекулярного докинга можно предположить возможность прямого взаимодействия изучаемых соединений с активным центром TDP1. Согласно результатам моделирования, ингибиторы располагаются в области связывания 3′-конца ДНК с активным центром TDP1 и могут образовывать устойчивые связи с каталитически значимыми остатками активного центра His263 и His493. Эти взаимодействия, вероятно, обеспечивают высокую ингибирующую активность соединений, наблюдаемую в биохимических экспериментах.

Ключевые слова: тирозил-ДНК фосфодиэстераза 1 (TDP1); ингибитор TDP1; ингибирующая активность; TDP2; PARP1; PARP2; ДНК-лиганды; производные бисбензимидазола

Для цитирования: Дырхеева Н.С., Чернышова И.А., Арутюнян А.Ф., Захаренко А.Л., Кутузов М.М., Науменко К.Н., Вензель А.С., Иванисенко В.А., Деев С.М., Жузе А.Л., Лаврик О.И. Влияние димерных бисбензимидазолов на активность ферментов репарации ДНК тирозил-ДНК-фосфодиэстераз 1 и 2 и поли(АДФ-рибоза)полимераз 1 и 2. *Вавиловский журнал генетики и селекции*. 2025;29(7):1097-1108. doi 10.18699/vjgb-25-114

Финансирование. Работа выполнена при поддержке гранта № 25-74-30006 Российского научного фонда (получение ферментов и определение уровня их активности) и в рамках государственного задания ИХБФМ СО РАН № 125012300658-9 (олигонуклеотидный синтез, использование оборудования общего пользования).

Благодарности. Авторы выражают благодарность Центру коллективного пользования «Биоинформатика» за предоставленные вычислительные ресурсы и программное обеспечение. Авторы благодарят Рашида Октамовича Анарбаева (ИХБФМ СО РАН) за предоставленный препарат фермента TDP2.

© Дырхеева Н.С., Чернышова И.А., Арутюнян А.Ф., Захаренко А.Л., Кутузов М.М., Науменко К.Н., Вензель А.С., Иванисенко В.А., Деев С.М., Жузе А.Л., Лаврик О.И., 2025

Контент доступен под лицензией Creative Commons Attribution 4.0

The effect of dimeric bisbenzimidazoles on the activity of DNA repair enzymes TDP1, TDP2, PARP1 and PARP2

N.S. Dyrkheeva [1], I.A. Chernyshova [1], A.F. Arutyunyan [2], A.L. Zakharenko [1], M.M. Kutuzov [1], K.N. Naumenko [1], A.S. Venzel [3], V.A. Ivanisenko [3], S.M. Deyev [4], A.L. Zhuze [2], O.I. Lavrik [1]

[1] Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] The Engelhardt Institute of Molecular Biology of the Russian Academy of Sciences, Moscow, Russia
[3] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow, Russia

lavrik@1bio.ru

Abstract. Oncological diseases remain a leading cause of pathological mortality worldwide, making the development of anticancer drugs a critical focus in medicinal chemistry. A promising strategy to enhance therapeutic efficacy and reduce chemotherapy-induced toxicity involves the combined inhibition of DNA repair enzymes and topoisomerases. Of particular interest are minor-groove DNA ligands, which exhibit potent inhibition of DNA-dependent enzymes while having low toxicity and mutagenicity. A number of research groups, including ours, are developing inhibitors of DNA repair enzymes that act simultaneously on several targets: tyrosyl-DNA phosphodiesterase 1/2 (TDP1/TDP2), poly(ADP-ribose) polymerase 1 (PARP1)/TDP1, topoisomerase 1 (TOP1)/TDP1. Such bifunctional inhibitors are designed to resolve the problem of tumor cell resistance to known chemotherapy drugs and increase the effectiveness of the latter. In this study, we evaluated the inhibitory activity of 22 minor-groove DNA ligands – bis- and trisbenzimidazoles against four key repair enzymes: TDP1, TDP2, PARP1, and PARP2. Four series of dimeric compounds and their monomeric units were studied. The difference in inhibitory activity of dimeric bisbenzimidazoles depending on the structure of the compound and the enzyme is shown. Our findings reveal distinct structure-activity relationships, with monomeric and dimeric ligands exhibiting potent TDP1 inhibition at micromolar to submicromolar IC_{50} values (half-maximal inhibitory concentration). Notably, dimeric compounds from the $DB_2Py(n)$ and $DB_3P(n)$ series demonstrated superior TDP1 inhibition compared to their monomers. In contrast, all tested compounds showed negligible activity against the other three repair enzymes; so, the compounds demonstrate specificity to TDP1. It should be noted that in this work, in the experiments with TDP1 and TDP2, the effect of the tested compounds as narrow-groove ligands binding to DNA was excluded, and their direct effect on the enzyme was investigated. The results of molecular docking suggest the possibility of direct interaction of active compounds with the active center of TDP1. According to the results of modeling, the inhibitors are located in the binding region of the 3′-end of DNA in the active site of TDP1 and could form stable bonds with the catalytically significant TDP1 residues His263 and His493. These interactions probably provide the high inhibitory activity of the compounds observed in biochemical experiments.

Key words: tyrosyl-DNA phosphodiesterase 1 (TDP1); TDP1 inhibitor; inhibitory activity; TDP2; PARP1; PARP2; DNA-ligands; bisbenzimidazole derivatives

For citation: Dyrkheeva N.S., Chernyshova I.A., Arutyunyan A.F., Zakharenko A.L., Kutuzov M.M., Naumenko K.N., Venzel A.S., Ivanisenko V.A., Deyev S.M., Zhuze A.L., Lavrik O.I. The effect of dimeric bisbenzimidazoles on the activity of DNA repair enzymes TDP1, TDP2, PARP1 and PARP2. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1097-1108. doi 10.18699/vjgb-25-114

Введение

Для понимания механизмов поддержания генетической стабильности и предотвращения развития различных заболеваний ведутся активные исследования ферментов репарации ДНК. Нарушения в работе систем репарации ДНК приводят к накоплению модифицированных оснований, разрывов и других повреждений, что увеличивает риск развития онкологических и других заболеваний. Изучение функционирования систем репарации помогает выявить причины наследственных заболеваний, нейродегенеративных дисфункций, связанных с нарушениями репарации, а также разработать новые методы терапии и профилактики онкологических заболеваний.

В последние годы большое внимание исследователи уделяют ферментам репарации ДНК как мишеням для разработки лекарственных средств. Ведется поиск новых соединений, подавляющих активность ферментов репарации ДНК, для повышения эффективности противораковой терапии. Ингибирование ферментов, участвующих в репарации, приводит к гибели раковых клеток за счет накопления повреждений ДНК, вызванных химиотерапией или лучевой терапией. В качестве перспективных мишеней для создания лекарственных препаратов рассматриваются такие ферменты репарации, как тирозил-ДНК-фосфодиэстеразы 1 и 2 (TDP1 и TDP2) и поли(АДФ-рибоза)полимеразы 1 и 2 (PARP1 и PARP2) (Pommier et al., 2014; Curtin, Szabo, 2020; Zakharenko et al., 2023).

TDP1 – фермент репарации ДНК, который участвует в удалении ковалентных аддуктов топоизомеразы 1 (TOP1) с ДНК, катализируя гидролиз фосфодиэфирной связи между остатком тирозина-723 TOP1 и 3′-фосфатной группой в одноцепочечном разрыве ДНК, генерируемом TOP1. TDP1 способна также удалять другие аддукты ДНК-белков, расположенные на 3′-конце ДНК, и различные другие повреждения на 3′-конце ДНК (Comeaux, van Waardenburg, 2014; Kawale, Povirk, 2018). TDP2 катализирует гидролиз ковалентных аддуктов между ДНК и остатком тирозина-804 активного центра топоизомеразы 2 (TOP2) (Pommier et al., 2010). TDP2 удаляет ковалентные аддукты из ДНК, расположенные на 5′-конце ДНК, путем

гидролиза 5′ фосфодиэфирной связи, что приводит к образованию ДНК со свободным 5′ фосфатом (Pommier et al., 2014). TDP1 и TDP2 способны в некоторой степени брать на себя функции друг друга, поскольку TDP1 обладает низкой активностью в реакции расщепления 5′-фосфотирозильных связей, а TDP2 – в реакции расщепления 3′-фосфотирозильных связей (Zeng et al., 2012; Pommier et al., 2014).

В настоящее время ингибиторы топоизомераз широко применяются в клинической практике в качестве противораковых препаратов. Из них наиболее часто используют топотекан и иринотекан, подавляющие активность топоизомеразы 1, а также этопозид, направленный на топоизомеразу 2 (Pommier et al., 2010). Механизм их действия заключается в образовании ковалентных аддуктов топоизомераз с ДНК, остановке репликации, что в итоге приводит к подавлению клеточной пролиферации. Различными исследователями высказывалось мнение (Pommier et al., 2014; Zakharenko et al., 2015) о том, что применение ингибиторов TDP1 и TDP2, повышающих эффективность ингибиторов топоизомераз, может позволить снизить дозу этих достаточно токсичных лекарств и, как следствие, токсичность терапии. На сегодняшний день идет активный поиск ингибиторов TDP1 (Zakharenko et al., 2023; Zhang M. et al., 2025). Особо выделяются производные природных соединений, таких как усниновая кислота, берберины, кумарины, нуклеозиды и стероиды (Zakharenko et al., 2023), которые эффективно подавляют активность очищенного фермента TDP1 и являются сенсибилизаторами топотекана в экспериментах, проведенных на клеточных и мышиных моделях рака (Zakharenko et al., 2023; Kornienko et al., 2024). Среди ингибиторов TDP2 следует отметить деазафлавины, считающиеся одними из самых активных для этого фермента (Marchand et al., 2016).

Ферменты PARP1 и PARP2 служат ключевыми регуляторами репарации ДНК и других клеточных процессов. Они катализируют ДНК-зависимый синтез разветвленного полимера поли(АДФ-рибозы) (PAR) и последующее АДФ-рибозилирование белков. АДФ-рибозилирование белков – это посттрансляционная модификация, которая индуцируется в ответ на повреждение ДНК. PARP1 участвует в различных путях репарации ДНК (Ray Chaudhuri, Nussenzweig, 2017; Lavrik, 2020). PARP2 также является ДНК-зависимым PAR-илирующим агентом и способна частично заменить PARP1 (Lavrik, 2020; Szanto et al., 2024), поэтому поиск ингибиторов PARP1 и PARP2 считается актуальной задачей современной медицинской химии. В клинической практике для лечения рака яичников, фаллопиевых труб, молочной железы и брюшины одобрены такие ингибиторы PARP1 и PARP2, как олапариб, рукапариб, нирапариб, велипариб и талазопариб (Kim D.-S. et al., 2021). Используемые ингибиторы работают по принципу синтетической летальности для уничтожения раковых клеток с дефектами в системе гомологичной рекомбинации (например, при мутациях BRCA1/2), превращая одноцепочечные разрывы ДНК в двуцепочечные, которые не могут быть эффективно восстановлены, что приводит к гибели раковой клетки. Активные центры PARP1 и PARP2 очень похожи (Schreiber et al., 2006; Hoch, Polo, 2019), поэтому известные на сегодняшний день ингибиторы чаще всего действуют на оба фермента, а также на другие ферменты семейства PARP. Проводится поиск селективных ингибиторов каждого из этих ферментов (Johannes et al., 2024). Одобренные для клинического применения ингибиторы PARP весьма токсичны и вызывают тяжелые побочные эффекты, поэтому поиск новых ингибиторов активно продолжается (Murai et al., 2014; Kim D.-S. et al., 2021; Johannes et al., 2024).

Один из перспективных классов соединений для поиска новых ингибиторов ферментов репарации – низкомолекулярные ДНК-связывающие агенты. Особый интерес представляют узкобороздочные ДНК-лиганды, которые способны ингибировать ДНК-зависимые ферменты, не обладают при этом высокой токсичностью и мутагенностью, а также хорошо растворимы в воде. Такие ДНК-лиганды характеризуются низким уровнем изменения геометрии ДНК и отсутствием образования ковалентных сшивок при формировании комплекса с ДНК (Arutyunyan et al., 2023a).

Наша научная группа имеет значительный опыт как экспериментального исследования потенциальных ингибиторов на уровне индивидуальных белковых мишеней, клеток и животных моделей (Zakharenko et al., 2023), так и применения методов молекулярного докинга и моделирования для изучения механизмов взаимодействия малых молекул с белками-мишенями. Были найдены эффективные ингибиторы TDP1, подавляющие активность рекомбинантного фермента TDP1 в субмикромолярном диапазоне концентраций. Соединения-лидеры являлись сенсибилизаторами топотекана в экспериментах, проведенных на клеточных культурах и на мышиных моделях опухолей (Zakharenko et al., 2023; Kornienko et al., 2024). Нами разработаны и исследованы ингибиторы PARP1, PARP2 и PARP3 на основе конъюгатов АДФ и морфолинонуклеозидов с использованием структурного моделирования активных сайтов этих ферментов (Sherstyuk et al., 2019; Chernyshova et al., 2024).

В настоящей работе приводятся данные скрининга 22 узкобороздочных лигандов в качестве ингибиторов TDP1, TDP2, PARP1 и PARP2. Исследуемые соединения являются производными бис- и трисбензимидазола. Изучены четыре мономерных соединения – MB_2, MB_2(Ac), MB_2Py(Ac), MB_3, а также четыре серии димерных производных. Димерные производные были получены путем конденсации мономерных субъединиц дикарбоновыми кислотами – DB_2P(n), DB_2Py(n) и DB_3P(n), где n – количество метиленовых звеньев в линкере (рис. 1).

Показано, что активность соединений варьирует в зависимости от их структуры и типа ферментативной мишени. Выявлена выраженная ингибирующая активность соединений в отношении TDP1, причем наблюдаемая корреляция говорит об увеличении активности ингибитора при введении в его структуру дополнительных связывающих блоков, таких как пирролкарбоксамидный фрагмент для серии DB_2Py(n), или при использовании комбинации из трех бензимидазольных блоков в мономерной субъединице. Несмотря на то что для серии DB_3(n) наблюдались высокие значения IC_{50}, это явление может быть объяснено высокой склонностью членов данного ряда соединений к агрегации, поскольку введение пиперазинового фрагмента в линкер в серии DB_3P(n) привело к получению

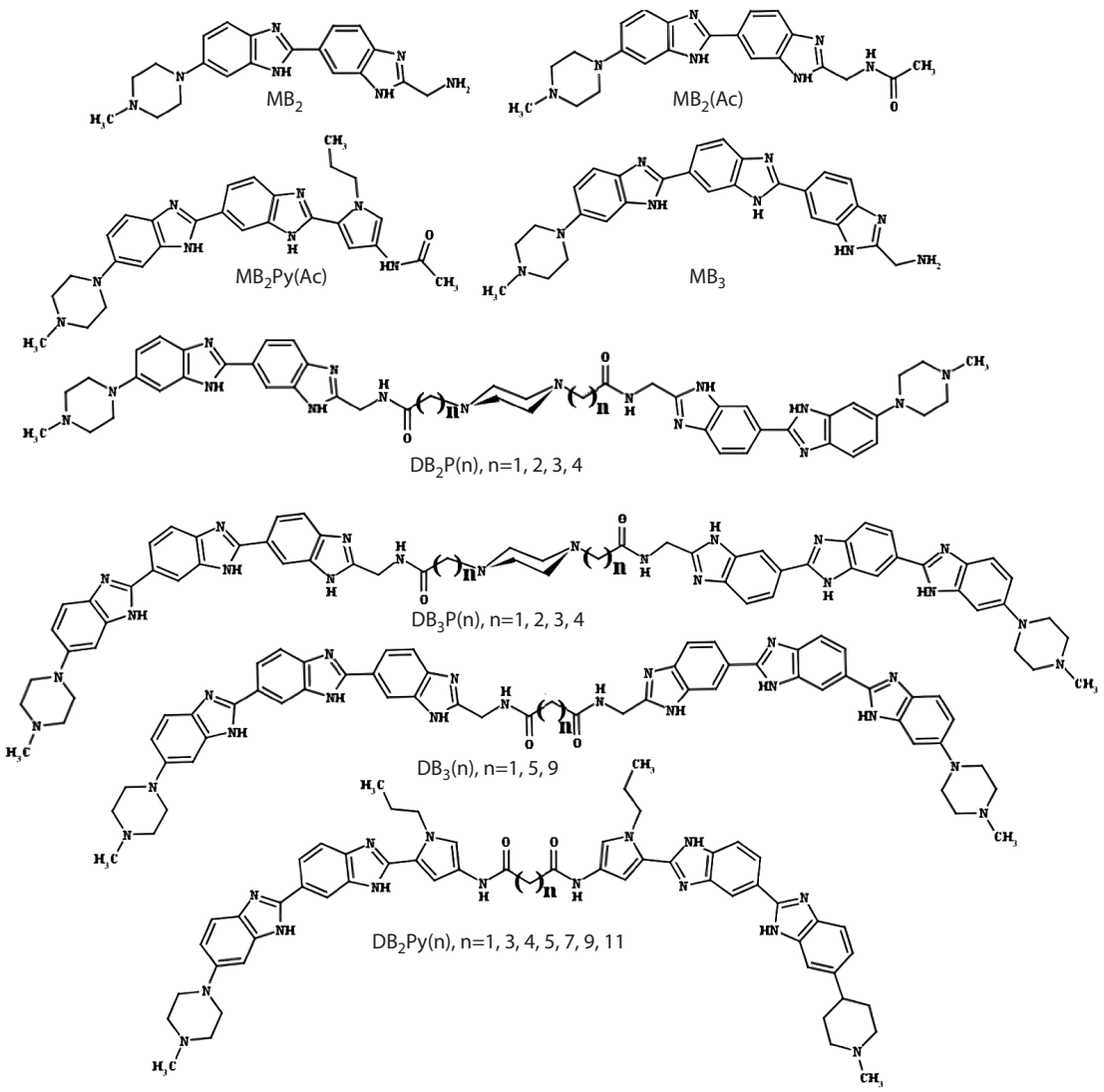


Рис. 1. Структуры исследованных производных бисбензимидазола.

ингибиторов с самыми низкими значениями IC_{50}, что косвенно подтверждает наше предположение. С целью выяснения возможного механизма их ингибирующего действия для этого фермента был проведен молекулярный докинг, результаты которого позволяют предположить наличие прямого взаимодействия между активными соединениями и ферментом TDP1. Согласно построенной модели связывания, ингибиторы располагаются в области ДНК-связывающего кармана TDP1 и способны образовывать устойчивые контакты с каталитически важными аминокислотными остатками His263 и His493. Эффективность действия этих соединений как ингибиторов TDP1 подтверждена экспериментальными данными. Результаты работы могут быть использованы для рационального дизайна новых, еще более эффективных ингибиторов TDP1.

Материалы и методы

Используемые материалы. Исследуемые соединения были синтезированы в Институте молекулярной биологии им. В.А. Энгельгардта РАН в лаборатории ДНК-белковых взаимодействий согласно ранее разработанным методикам (Ivanov et al., 2015; Arutyunyan et al., 2023a, b; Susova et al., 2024). Список названий соединений по IUPAC приведен в Приложении[1].

Рекомбинантные белки человека тирозил-ДНК-фосфодиэстераза 1 (TDP1) и тирозил-ДНК-фосфодиэстераза 2 (TDP2) были экспрессированы в системе *E. coli*, поли(АДФ-рибоза)полимераза 1 (PARP1) и поли(АДФ-рибоза)полимераза 2 (PARP2) – в клетках насекомых с использованием бакуловирусной системы экспрессии и очищены, как описано в работах (Sukhanova et al., 2004; Sherstyuk et al., 2019; Dyrkheeva et al., 2020, 2021).

Олигонуклеотид 5′-FAM-AAC GTC AGG GTC TTC C-BHQ1-3′ синтезирован в лаборатории химии нуклеиновых кислот Института химической биологии и фундаментальной медицины СО РАН (Новосибирск, Россия) согласно (Zakharenko et al., 2015).

Определение активности фермента TDP1. Реакционная смесь (200 мкл) для флуоресцентной детекции активности фермента TDP1 в реальном времени (Zakharenko

[1] Табл. S1, S2 и рис. S1–S4 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx40.pdf

et al., 2015) содержала реакционный буферный раствор TDP1 (50 мМ Трис-HCl, pH 8.0, 50 мМ NaCl и 7 мМ β-меркаптоэтанола), 50 нМ олигонуклеотида 5′-FAM-AAC GTC AGG GTC TTC C-BHQ1-3′, тестируемое в качестве ингибитора соединение в различных концентрациях и TDP1 в конечной концентрации 1.5 нМ. Реакционные смеси инкубировали при постоянной температуре 26 °C в планшетном флуориметре POLARstar OPTIMA (BMG LABTECH, GmbH, Ортенберг, Германия). Интенсивность флуоресценции (Ex485/Em520 нм) измеряли каждую минуту в течение 10 мин. Средние значения полумаксимальной ингибирующей концентрации (IC_{50} – концентрация соединения, которая ингибировала 50 % активности фермента по сравнению с необработанной контрольной лункой, содержащей только фермент и субстрат) определяли с помощью кривой зависимости уровня флуоресцентного сигнала от концентрации ингибитора и рассчитывали с использованием MARS Data Analysis 2.0 (BMG LABTECH). Кинетические кривые были получены минимум в трех независимых экспериментах и статистически обработаны в OriginPro 8.6.0 (OriginLab, Нортгемптон, Массачусетс, США).

Определение активности фермента TDP2. Для определения активности фермента TDP2 в качестве субстрата были использованы олигонуклеотид 5′-тирозин-AAC GTC AGG GTC TTC C-FAM-3′, содержащий метку 6-FAM на 3′-конце, и остаток L-тирозина, присоединенный через фенольную OH-группу к 5′-концевому фосфату, синтезирован в российско-франко-японской лаборатории бионанотехнологии НГУ, как описано в работе (Dyrkheeva et al., 2021). Субстрат в концентрации 100 нМ инкубировали с TDP2 в концентрации 200 нМ в отсутствие или в присутствии ингибитора (500 мкМ) в течение 10 мин при 37 °C в буфере, содержащем 50 мМ Трис-HCl, pH 8.0, 50 мМ NaCl, 7 мМ β-меркаптоэтанола (Dyrkheeva et al., 2021). Реакцию останавливали добавлением буфера для нанесения в ПААГ (TBE, 10 % формамид, 7 М мочевина, 20 мМ ЭДТА, 0.1 % ксиленцианол и 0.1 % бромфеноловый синий). Далее образцы нагревали при 90 °C в течение 5 мин. Продукты ферментативной реакции разделяли электрофорезом в 20 % денатурирующем ПААГ с 7 М мочевиной при соотношении акриламида к бисакриламиду 19:1. Для сканирования геля и визуализации использовали высокоразрешающий лазерный сканер Typhoon FLA 9500 (GE Healthcare, Чикаго, Иллинойс, США). Данные анализировали с помощью программного обеспечения QuantityOne 4.6.7 (Bio-Rad Laboratories, Inc., Геркулес, Калифорния, США). Было проведено минимум три независимых эксперимента, статистическую обработку выполняли в программе OriginPro 8.6.0 (OriginLab, Нортгемптон, Массачусетс, США).

Для определения активности ферментов PARP1 и PARP2 в присутствии и в отсутствие тестируемых соединений радиоактивно меченный [^{32}P]-NAD^+ синтезировали из α-[^{32}P]-АТФ согласно протоколу (Sherstyuk et al., 2019). Реакцию автополи(АДФ-рибозил)ирования проводили в буфере для PARP1: 50 мМ Трис-HCl, pH 8.0, 10 мМ $MgCl_2$, 150 мМ NaCl и 7 мМ β-меркаптоэтанола, а также 2 ое/мл активированной ДНК, 0.3 мМ [^{32}P]-NAD^+ при 37 °C. Инициировали реакцию добавлением PARP1 до 200 нМ и проводили в течение 2 мин.

Буфер для PARP2 содержал: 50 мМ Трис-HCl, pH 8.0, 3 мМ спермина, 150 мМ NaCl и 7 мМ β-меркаптоэтанола, 2 ое/мл активированной ДНК, 0.6 мМ [^{32}P]-NAD^+ при 37 °C. Реакцию инициировали добавлением PARP2 до 600 нМ, реакционные смеси инкубировали в течение 5 мин. Останавливали реакцию помещением аликвот по 5 мкл на бумажные фильтры Whatman 1, пропитанные 5 % трихлоруксусной кислотой (ТХУ). Фильтры промывали 5 % ТХУ четыре раза и высушивали на воздухе после удаления ТХУ с 90 % этанолом. Включение радиоактивной метки в продукт реакции рассчитывали с помощью сканера Typhoon FLA 9500 (GE Healthcare, Чикаго, Иллинойс, США). Проводили минимум три независимых эксперимента.

Молекулярное моделирование. Для оценки взаимодействия исследуемых соединений с ферментом TDP1 нами был проведен молекулярный докинг с последующим анализом полученных комплексов. Исследование включало подготовку структур белка и лигандов, молекулярный докинг, минимизацию энергии соединений в сайте связывания и оценку аффинности ингибиторов с использованием Vinardo, X-Score и REF2015.

В качестве структуры белка-мишени была выбрана кристаллическая структура TDP1 (PDB ID: 8V0B). Отсутствующие в модели петли были восстановлены на основании предсказания AlphaFold2 (Jumper et al., 2021), выполненного в ColabFold (Mirdita et al., 2022) без множественного выравнивания (MSA) и со структуры 8V0B в качестве шаблона.

Далее к полученной модели были добавлены атомы водорода и рассчитаны заряды с применением утилиты DockPrep в UCSF Chimera (Pettersen et al., 2004). Структуры ингибиторов были подготовлены в OpenBabel (O'Boyle et al., 2011): добавлены водороды, рассчитаны частичные заряды и проведена геометрическая минимизация.

Для молекулярного докинга была привлечена программа UCSF DOCK 6.11 (Allen et al., 2015). Использовался полноатомный гибкий докинг по всей поверхности белка. На первом этапе докинга осуществлялось позиционирование базовых фрагментов ингибиторов (MB_2(Ac), MB_2Py(Ac)), после чего проводился докинг полноразмерных молекул с последующей минимизацией их энергии в сайте связывания. Для каждого соединения было запрошено до девяти лучших по GridScore конформаций. Из полученных девяти конформаций для каждого лиганда отбиралась та структура, которая имела минимальный RMSD относительно оптимальной конформации базового фрагмента. В случаях, когда DOCK6 вернул менее девяти уникальных конформаций (вследствие кластеризации, фильтрации по энергии или неудачи генерации дополнительных конформеров), отбор выполнялся из всех доступных конформаций (табл. S1).

Финальную оценку способности связывания ингибиторов с белком проводили с использованием нескольких независимых скоринг-функций: ContinuousScore из DOCK 6, Vinardo (Quiroga, Villarreal, 2016), X-Score (Wang R. et al., 2002), а также REF2015 в среде PyRosetta4 (Chaudhury et

al., 2010; Alford et al., 2017) в соответствии с протоколом (Moretti et al., 2016). ContinuousScore – скоринг-функция в DOCK 6, учитывающая ван-дер-ваальсовы взаимодействия, электростатические взаимодействия, внутреннюю энергию лиганда и штрафы за стерические конфликты при прямом расчете межатомных расстояний. Vinardo – скоринг-функция для докинга, учитывающая вклад водородных связей, гидрофобных и ван-дер-ваальсовых взаимодействий, а также поправки на неоптимальное позиционирование лигандов. Скоринг-функция X-Score состоит из трех составляющих: HPScore, HMScore и HSScore, основанных на различных эмпирических принципах для оценки сродства лиганд–белок. В данном исследовании использовалась усредненная X-Score, отражающая влияние гидрофобных, полярных и электростатических контактов. Полноатомная скоринг-функция REF2015, реализованная в PyRosetta, включает вклад ван-дер-ваальсовых, электростатических, водородных, сольватных и дополнительных атомных парных взаимодействий и позволяет корректно ранжировать близкие по энергии позиции ингибиторов.

Для валидации результатов молекулярного докинга и оценки стабильности предсказанного комплекса во времени было проведено молекулярно-динамическое моделирование комплекса TDP1 с соединением-лидером $DB_2Py(1)$, показавшим наилучшую ингибирующую активность. Моделирование выполнялось с использованием пакета OpenMM 8 (Eastman et al., 2024). Подробный протокол молекулярно-динамического моделирования представлен в Приложении.

Результаты

В настоящей работе была впервые изучена способность четырех серий димерных ДНК-лигандов $DB_2P(n)$, $DB_2Py(n)$, $DB_3(n)$, $DB_3P(n)$, а также их мономерных единиц MB_2, $MB_2(Ac)$, $MB_2Py(Ac)$ и MB_3 (см. рис. 1) ингибировать активность рекомбинантных ферментов репарации ДНК TDP1 и TDP2, PARP1 и PARP2 (см. таблицу).

Первая группа исследуемых соединений представляет собой димерные производные мономерного бисбензимидазольного лиганда MB_2, производного широко изученного узкобороздочного ДНК-лиганда Hoechst 33258, в котором гидроксифенильная группа заменена на более гидрофильный аминометиленовый фрагмент – $DB_2P(n)$. В качестве линкера для соединений этой группы были использованы 1,4-пиперазиндиалкилдикарбоновые кислоты, содержащие метиленовый, этиленовый, пропиленовый или бутиленовый спейсер (см. рис. 1). Также данная серия была дополнена мономерным производным $MB_2(Ac)$, ацилированным по аминометиленовому фрагменту, что

Способность тестируемых соединений ингибировать активность ферментов TDP1, TDP2, PARP1 и PARP2

№ п/п	Соединение	IC_{50} TDP1, мкМ	TDP2	PARP1	PARP2
			Остаточная активность (%) при 500 мкМ соединения		
1	MB_2	2 ± 1	~100	~100	~100
2	$MB_2(Ac)$	1.5 ± 0.5	~100	~100	~100
3	$DB_2P(1)$	6 ± 4	66 ± 7	57 ± 16	~100
4	$DB_2P(2)$	9 ± 3	44 ± 11	51 ± 15	80 ± 20
5	$DB_2P(3)$	4.1 ± 0.6	36 ± 7	37 ± 10	64 ± 16
6	$DB_2P(4)$	2.3 ± 0.3	44 ± 11	33 ± 13	85 ± 13
7	$MB_2Py(Ac)$	5 ± 2	~100	~100	~100
8	$DB_2Py(1)$	0.25 ± 0.05	55 ± 3	~100	~100
9	$DB_2Py(3)$	0.41 ± 0.09	70 ± 11	~100	~100
10	$DB_2Py(4)$	0.4 ± 0.15	~100	~100	~100
11	$DB_2Py(5)$	0.35 ± 0.13	~100	~100	~100
12	$DB_2Py(7)$	0.28 ± 0.01	~100	~100	~100
13	$DB_2Py(9)$	0.30 ± 0.08	~100	~100	~100
14	$DB_2Py(11)$	0.9 ± 0.1	~100	~100	~100
15	MB_3	0.70 ± 0.05	~100	65 ± 15	~100
16	$DB_3(1)$	>50	70 ± 6	55 ± 13	~100
17	$DB_3(5)$	>50	65 ± 10	62 ± 16	~100
18	$DB_3(9)$	>50	~100	~100	~100
19	$DB_3P(1)$	0.10 ± 0.05	~100	70 ± 12	~100
20	$DB_3P(2)$	0.11 ± 0.01	~100	40 ± 5	~100
21	$DB_3P(3)$	0.20 ± 0.05	~100	47 ± 14	~100
22	$DB_3P(4)$	0.15 ± 0.03	~100	48 ± 15	~100

Примечание. Для значений величин IC_{50} и остаточной активности фермента (%) в присутствии ингибитора приведено среднее значение ± стандартное отклонение (не менее трех повторов).

структурно приближает это соединение, по сравнению с MB_2, к половине димерного соединения $DB_2P(n)$ и делает его более корректным препаратом для сравнения. Серия $DB_2P(n)$ отличается от других серий лигандов наличием в структуре линкера положительно заряженного 1,4-пиперазина, что улучшает растворимость лиганда и может увеличивать сродство лиганда к ферменту.

Следующая группа соединений – производные мономерного трисбензимидазольного соединения MB_3, которое можно рассматривать как производное MB_2, содержащее на один бензимидазольный фрагмент больше, что увеличивает количество потенциально возможных водородных связей в комплексе ингибитор–TDP1. Димерные производные MB_3 представлены двумя сериями соединений – $DB_3P(n)$, тоже димеризованных 1,4-пиперазиндиалкилдикарбоновыми кислотами, и $DB_3(n)$, где в качестве линкеров были использованы н-алкилдикарбоновые кислоты. Серии $DB_3(n)$ и $DB_3P(n)$ характеризуются наличием в структуре трисбензимидазолов, а $DB_3P(n)$ – наличием также в линкере 1,4-пиперазина.

К третьей группе соединений относятся производные мономерного соединения $MB_2Py(Ac)$, являющегося изостерным аналогом MB_3, ввиду того, что содержащийся в его структуре пирролкарбоксамидный фрагмент способен выступать донором атома водорода при карбоксамидном азоте для образования водородной связи в аналогичном бензимидазолу положении. Димерные производные представлены серией $DB_2Py(n)$, содержащей в качестве линкера н-алкилдикарбоновые кислоты. Данная серия включает набор соединений, содержащих 1, 3, 4, 5, 7, 9 и 11 метиленовых звеньев, что позволило точнее оценить зависимость ингибирующей активности соединений от длины спейсера. Серия $DB_2Py(n)$ отличается от серии $DB_3(n)$ наличием, помимо бисбензимидазольной структуры, пирролкарбоксамидной структуры, являющейся фрагментом АТ-специфичного антибиотика нетропсина.

Методом флуоресцентного анализа в реальном времени были получены значения концентрации полумаксимального ингибирования (IC_{50}) исследуемых соединений (см. таблицу) в реакции отщепления BHQ1 от 3′-конца олигонуклеотида с помощью TDP1, что приводило к разгоранию флуоресценции FAM на 5′-конце цепи (Zakharenko et al., 2015). Следует также отметить, что в качестве субстрата был использован одноцепочечный олигонуклеотид, чтобы исключить связывание димерных бисбензимидазолов как узкобороздочных лигандов с ДНК-субстратом и направить их воздействие на ферментативную мишень.

Из полученных данных для мономерных соединений MB_2 и $MB_2(Ac)$ и их димерных производных $DB_2P(n)$ при n = 1, 2, 3, 4 значения IC_{50} лежат в микромолярном диапазоне, причем димеризация не привела к повышению ингибиторной активности исследуемых соединений. В то же время для димеров мономерного $MB_2Py(Ac)$, имеющего значение IC_{50} 5 ± 2 мкМ, величина параметра полуингибирующей концентрации значительно снизилась, находясь в диапазоне от 0.25 до 0.90 мкМ. Аналогично переход от мономерного MB_3 к серии димерных $DB_3P(n)$ приводил к повышению ингибиторной активности соединений, хотя и не столь ярко выраженному, однако димерные производные MB_3, не содержащие пиперазинового фрагмента в линкере, – соединения $DB_3(n)$ – показали самый низкий уровень активности среди всех испытанных в работе ингибиторов. Тот факт, что значения IC_{50} для этих соединений (см. таблицу) настолько сильно выбиваются из общего массива данных, по всей видимости, связан с тем, что соединения $DB_3(n)$ обладают протяженной и плоской геометрией, а также жестким линкером, что мешает оптимальному позиционированию соединений этого типа в активном центре фермента (см. рис. 1).

Таким образом, согласно экспериментальным данным, все изученные соединения, кроме группы $DB_3(n)$, эффективно ингибируют активность TDP1 в микромолярных и субмикромолярных концентрациях. Наблюдается корреляция структура–активность, заключающаяся в снижении концентрации для достижения эффекта полумаксимального ингибирования при увеличении в соединении числа блоков, содержащих доноры водородной связи. В частности, димеризация является одним из простых подходов к увеличению таких структур в одной молекуле, что приводит к нелинейному росту константы комплексообразования (Неудачина, Лакиза, 2014). Также наблюдается снижение IC_{50} при введении пиперазинового фрагмента в структуру линкера, что может быть обусловлено увеличением гидрофильности молекул. Полученные результаты позволили построить корреляцию структура–активность и оценить вклад димеризации в увеличение ингибирующей способности исследуемых соединений.

Для изучения влияния рассматриваемых соединений на активность TDP2 мы проверили способность этого фермента удалять остаток тирозина с 5′-конца олигонуклеотидного субстрата в отсутствие и в присутствии ингибиторов, как описано в (Dyrkheeva et al., 2021). Все соединения группы $DB_2P(n)$, а также $DB_2Py(n)$ при n = 1, 3 и $DB_3(n)$ при n = 1, 5 в концентрации 500 мкМ ингибировали активность фермента примерно на 50 %; остальные соединения не проявили ингибирующей активности (см. таблицу). Таким образом, все протестированные соединения показали значительно более низкую склонность ингибировать TDP2 по сравнению с TDP1. Интересно, что группа $DB_2P(n)$ менее эффективно ингибировала TDP1 и более эффективно TDP2, чем соединения других групп.

Следующим этапом нашей работы было тестирование способности исследуемых соединений ингибировать PARP1 и PARP2, т. е. их ферментативной активности в реакции синтеза поли(АДФ-рибозы) (PAR), в довольно высоком диапазоне концентрации соединений. Все соединения показали низкую эффективность ингибирования этих двух ферментов. Наиболее активными оказались соединения группы $DB_2P(n)$, представители которой с n = 2, 3, 4 снижали активность PARP1 и PARP2 в концентрации 500 мкМ. Ингибирующее действие наблюдалось также для соединений серий $DB_3(n)$ и $DB_3P(n)$ в концентрации 500 мкМ, в то же время данные соединения проявили ингибирующую активность только в реакции синтеза PAR, катализируемой PARP1, но не PARP2 (см. таблицу).

Поскольку, согласно экспериментальным данным, все изученные соединения, за исключением группы $DB_3(n)$, эффективно ингибируют активность TDP1, далее нами была оценена *in silico* способность соединений групп $DB_2P(n)$ и $DB_2Py(n)$ связываться с ферментом TDP1 для

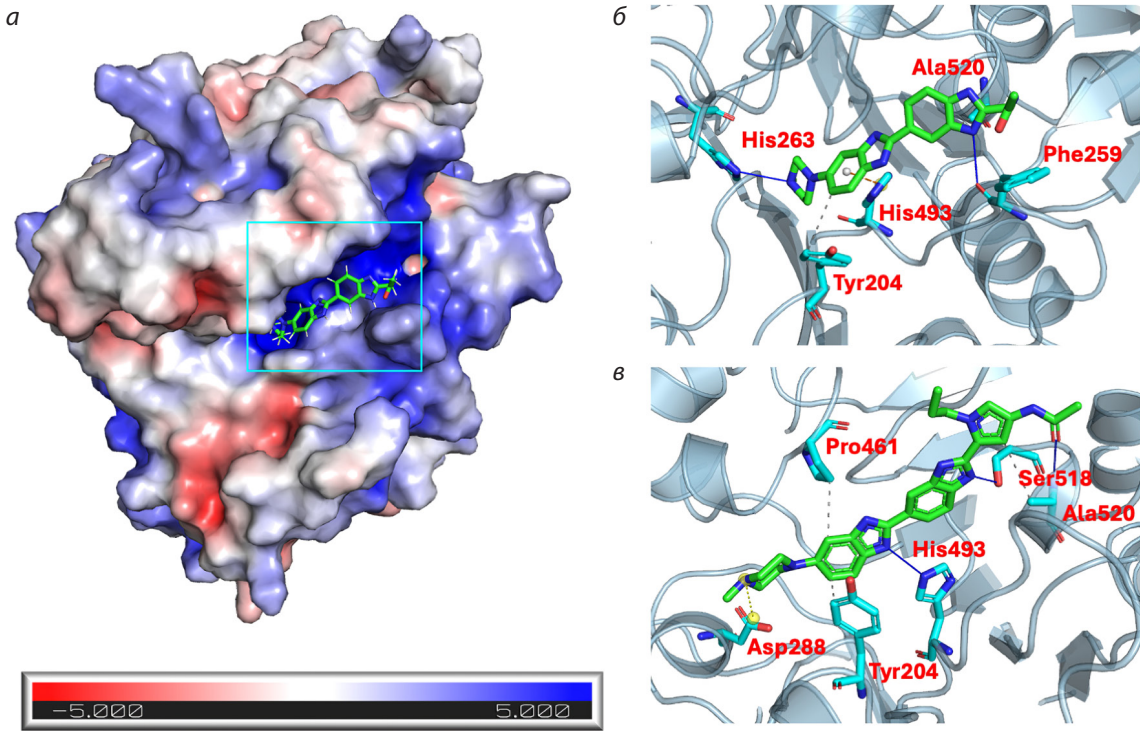


Рис. 2. *а* – структура TDP1 (PDB ID: 8V0 B) с ингибитором $MB_2(Ac)$, расположенным в положительно заряженном участке активного центра TDP1. Поверхность белка окрашена в соответствии с распределением электростатического потенциала, рассчитанного с помощью APBS (Jurrus et al., 2018). Прямоугольной рамкой выделен ДНК-связывающий участок TDP1. Ниже расположена шкала значений электростатического потенциала поверхности TDP1 (в единицах kT/e, где kT/e ≈ 25.7 мВ при 298 K). Цветовая шкала: красный цвет обозначает отрицательный потенциал (–5 kT/e), белый – нейтральный (0 kT/e), синий – положительный потенциал (+5 kT/e). *б*, *в* – предсказанные конформации ингибиторов $MB_2(Ac)$ и $MB_2Py(Ac)$ (зеленый цвет) в комплексе с TDP1 с контактирующими остатками (голубой цвет).

выяснения возможного молекулярного механизма их ингибирующего действия. С этой целью был выполнен полноатомный гибкий молекулярный докинг по всей поверхности белка TDP1 (PDB ID: 8V0B) соединений $DB_2P(n)$ и $DB_2Py(n)$.

Согласно результатам докинга, можно предположить, что конформации с минимальной расчетной энергией для каждого ингибитора формируют взаимодействия в активном центре TDP1, вблизи остатков His263 и His493 (рис. 2, *а*), аналогично соединению $MB_2(Ac)$ (рис. S1). Кроме того, был проведен дополнительный анализ способности связывания димерных соединений с TDP1 с использованием скоринг-функций Vinardo, X-Score и REF2015 в среде PyRosetta (табл. S2). Значения скоринг-функций позволяют предположить высокую аффинность исследованных ингибиторов групп $DB_2P(n)$ и $DB_2Py(n)$ к TDP1. Следует отметить, что полной корреляции параметров, полученных с помощью докинга (см. табл. S2), с величинами IC_{50}, найденными в эксперименте (см. таблицу), не наблюдается, что можно объяснить вкладом гидрофобных линкеров, которые сложно учесть в расчете энергий.

По данным молекулярного моделирования, соединение $MB_2(Ac)$ (см. рис. 2, *б*), являющееся мономерным звеном для димерных производных $DB_2P(n)$, образует водородную связь с His263 и π-катионное взаимодействие с His493, что, вероятно, может приводить к блокировке каталитического акта TDP1. Помимо взаимодействий с каталитически активными остатками, для $MB_2(Ac)$ возможно формирование гидрофобных контактов с Tyr204 и Ala520, а также водородной связи с Phe259, что может усиливать ингибирующее действие этого соединения. В отличие от $MB_2(Ac)$, соединение $MB_2Py(Ac)$ (см. рис. 2, *в*) потенциально может взаимодействовать только с одним каталитическим остатком – His493 – через образование водородной связи. Такая разница во взаимодействиях может быть причиной более высокой ингибирующей активности $MB_2(Ac)$ по сравнению с $MB_2Py(Ac)$, что подтверждается экспериментальными данными (см. таблицу).

Анализ взаимодействий с помощью PLIP (protein–ligand interaction profiler) (Salentin et al., 2015) для комплексов TDP1 с димерными соединениями группы $DB_2Py(n)$ (рис. S2) показал, что эти соединения формируют большее количество контактов белок-лиганд (водородных связей и гидрофобных взаимодействий) по сравнению с мономером $MB_2Py(Ac)$. В частности, соединение $DB_2Py(1)$ образует водородные связи с Ser400 и Ser403, а также гидрофобные взаимодействия с Pro463 – остатки этих аминокислот расположены в сайте связывания лигандов с активным центром TDP1, что, вероятно, способствует стабилизации взаимодействующего фрагмента димера в активном центре фермента. Найденные с помощью анализа результатов докинга данные, характеризующие большую площадь поверхности контакта димерных соединений $DB_2Py(n)$ с TDP1 по сравнению с мономером $MB_2Py(Ac)$, коррелируют с понижением значений IC_{50} для

димеров, что свидетельствует о возрастании сродства этих соединений к активному центру фермента (см. таблицу). Согласно полученным данным, гидрофобные взаимодействия с остатками Pro461 и/или Tyr204, локализованными в активном центре TDP1, также могут вносить вклад в повышение ингибирующей активности соединений группы $DB_2Py(n)$.

Анализ взаимодействий соединений из группы $DB_2P(n)$ с TDP1 показал, что в образовании комплекса участвуют аналогичные аминокислотные остатки, за исключением Tyr204, с остатком которого соединения $DB_2P(n)$, в отличие от $DB_2Py(n)$, по-видимому, не взаимодействуют (рис. S3). Кроме того, отмечены возможные различия в характере взаимодействий с одними и теми же аминокислотами. Например, для остатка Lys519 в случае соединений $DB_2P(n)$ можно предположить образование водородных связей с атомами азота пиперазинового фрагмента через атом N1 боковой цепи. В то же время в соединениях $DB_2Py(n)$ предсказаны два типа взаимодействий с Lys519: водородная связь между атомом азота основной цепи Lys519 и атомом кислорода в пирролкарбоксамидной группе (в $DB_2Py(1)$, $DB_2Py(4)$, $DB_2Py(7)$, $DB_2Py(9)$), а также π-катионное взаимодействие между пирролом и боковой цепью Lys519 (в $DB_2Py(3)$ и $DB_2Py(5)$) (см. рис. S2).

Для соединения $DB_2Py(1)$, продемонстрировавшего наибольшую ингибирующую активность (наименьшее значение IC_{50}) среди исследованных производных, было проведено дополнительное моделирование молекулярной динамики в предсказанном комплексе с TDP1. Анализ траектории МД показал, что комплекс TDP1-$DB_2Py(1)$ сохраняет стабильность на протяжении всего времени моделирования. Значения RMSD лиганда находились в диапазоне 1.5–3.0 Å (рис. S4), что говорит об устойчивом связывании $DB_2Py(1)$ в активном центре белка без признаков диссоциации или значительных конформационных перестроек. Полученные данные подтверждают прочность образованного комплекса и согласуются с высокой биологической активностью этого соединения.

Следует отметить, что выполненный нами анализ молекулярных контактов, а также значения скоринг-функций, полученные по результатам молекулярного докинга, свидетельствуют в пользу способности соединений обеих проанализированных групп – $DB_2P(n)$ с алифатическим линкером и $DB_2Py(n)$ с пиперазиновым фрагментом в линкере – формировать прочный комплекс с TDP1. Тем не менее экспериментальные данные выявляют различия в их ингибирующей активности: соединения с алифатическим линкером демонстрируют более высокую эффективность ингибирования по сравнению с соединениями, содержащими пиперазиновое кольцо. Это различие не может быть полностью объяснено на основании контактного анализа, что указывает на возможную разницу в конформационной подвижности рассматриваемых групп соединений. В частности, включение пиперазинового фрагмента в центральную часть линкера, по-видимому, ограничивает его гибкость, что влияет на динамику взаимодействия ингибитора с активным центром, препятствует оптимальному позиционированию ингибитора в активном центре фермента и, как следствие, снижает его ингибирующую активность.

Обсуждение

TDP1 играет ключевую роль в устранении повреждений ДНК, расположенных на 3′-конце ДНК, стабилизированных противораковыми препаратами, применяемыми в клинической практике, такими как топотекан и иринотекан, которые являются производными природного соединения камптотецина (Comeaux, van Waardenburg, 2014; Kawale, Povirk, 2018). Следовательно, активность TDP1 может быть возможной причиной резистентности опухолей к используемым в клинике ингибиторам TOP1. В настоящее время активно ведутся поиски совместных ингибиторов TOP1 и TDP1 (Conda-Sheridan et al., 2013; Nguyen et al., 2015; Zhang X.-R. et al., 2018; Hu et al., 2021; Yang et al., 2023).

Кроме того, поскольку известно, что активности TDP1 и TDP2 перекрываются, хотя и в незначительной степени (Pommier et al., 2014), их способность выполнять функции друг друга делает весьма перспективным совокупное использование ингибиторов этих двух ферментов или создание агентов, способных одновременно ингибировать как TDP1, так и TDP2. Одновременное подавление активности данных ферментов может быть использовано для повышения эффективности большого набора клинически важных противораковых препаратов, ингибиторов TOP1 и TOP2. Были также обнаружены тройные ингибиторы TOP1/TDP1/TDP2, которые проявляют умеренную активность против TDP1 и слабую активность против TDP2 (Wang P. et al., 2017). Наиболее эффективными ингибиторами TDP2 на сегодняшний день считаются деазафлавины, которые *in vitro* показывают синергию с этопозидом в нетоксичных концентрациях (Marchand et al., 2016). Также были найдены некоторые эффективные ингибиторы TDP2 из других классов соединений (Yang et al., 2021; Zhang Y. et al., 2021).

Известно, что N-концевой домен TDP1 напрямую связывается с С-концевым доменом PARP1, а TDP1 подвергается PARилированию с помощью PARP1 для того, чтобы быть привлеченной к аддукту TOP1-ДНК (Das et al., 2014; Lebedeva et al., 2015). PARилирование TDP1 стимулирует ее посадку на участки с поврежденной ДНК, не влияя на каталитическую активность этого фермента (Chowdhuri, Das, 2021). Показано также, что PARP1 может взаимодействовать с TDP1, формируя белок-белковые контакты (Moor et al., 2015). Установлено, что сочетание нокдауна TDP1 и ингибирования активности PARP1 с помощью рукапариба снижает пролиферацию клеток значительнее, чем те же методы подавления функции ферментов по отдельности (Fam et al., 2013). В литературе встречается предположение, что противораковый эффект ингибиторов TOP1 может быть значительно усилен одновременным ингибированием PARP1 и TDP1 (Smith et al., 2005; Alagoz et al., 2014; Das et al., 2014; Muray et al., 2014; Elsayed et al., 2016; Matsuno et al., 2018; Jing et al., 2020; Kim J.W. et al., 2020; Chowdhuri, Das, 2021; Flörkemeier et al., 2022). Взаимодействие ферментов PARP1 и TDP1 между собой было описано в ряде публикаций (Das et al., 2014; Moor et al., 2015), что делает актуальным поиск двойных ингибиторов TDP1 и PARP1.

Ранее нами были обнаружены двойные ингибиторы TDP1 и TDP2, а также тройные ингибиторы TDP1, TDP2

и PARP1 (Dyrkheeva et al., 2021) – тиоэфиры усниновой кислоты, слабо ингибирующие TDP2 и PARP1, поэтому актуален поиск новых соединений, способных воздействовать сразу на две или три взаимодействующие функционально мишени. В данной работе была проверена способность серии узкобороздочных ДНК-лигандов ингибировать ферменты TDP1, TDP2, PARP1 и PARP2. Эффективных ингибиторов, действующих сразу на четыре фермента, обнаружить не удалось, но было показано, что эти соединения ингибируют TDP1. Исследованные в работе ДНК-лиганды способны ингибировать ДНК-зависимые ферменты за счет связывания с двуцепочечной ДНК. Однако в настоящей работе мы показали, что они способны селективно ингибировать TDP1, поскольку эксперименты проводили в отсутствие двуцепочечной ДНК как альтернативной мишени.

Результаты молекулярного докинга и анализа межмолекулярных взаимодействий говорят в пользу того, что большинство исследованных соединений групп $DB_2P(n)$ и $DB_2Py(n)$ могут обладать высоким сродством к ферменту TDP1 и формировать устойчивые комплексы с его каталитическим центром. Для всех соединений предсказаны взаимодействия с ключевыми каталитическими остатками активного центра белка TDP1.

Заключение

Изучено влияние димерных бис- и трисбензимидазолов на активность ферментов репарации ДНК – TDP1, TDP2, PARP1 и PARP2. Основные результаты показали, что все ингибиторы, кроме соединений серии $DB_3(n)$, эффективно ингибируют TDP1. Наиболее активными оказались соединения $DB_2Py(n)$ и $DB_3P(n)$, способные ингибировать TDP1 в субмикромолярном диапазоне концентраций. Исследуемые соединения демонстрируют высокую селективность, слабо влияя на активность других тестируемых ферментов.

Основываясь на результатах молекулярного докинга, мы сделали предположение, что активные ингибиторы локализуются в области ДНК-связывающего кармана TDP1 и могут формировать стабильные взаимодействия с каталитически важными остатками His263 и His493. Эти взаимодействия, вероятно, лежат в основе наблюдаемой высокой ингибирующей активности.

Важным результатом является также установление зависимости активности от структуры соединений. Димеризация оказала неоднозначное влияние на ингибирующий эффект: соединения серий $DB_2Py(n)$ и $DB_3P(n)$ были существенно (на порядок) более активны, чем соответствующие мономеры; в серии $DB_2P(n)$ на ингибирующую активность повлияла не только димеризация, но и длина линкера и введение 1,4-пиперазина, несущего два положительных заряда в линкер. Серия $DB_3(n)$ оказалась неактивна, в отличие от мономера. Введение пиперазинового фрагмента в линкер в серии $DB_3P(n)$ привело к выраженной ингибирующей активности по сравнению с $DB_3(n)$ без такого фрагмента. Мы предполагаем, что усиление ингибирующего эффекта связано с введением двух положительных зарядов в линкер и увеличением числа возможных контактов лигандов с активным центром фермента.

В целом на основании результатов настоящей работы могут быть предложены новые стратегии для разработки терапии онкологических заболеваний. Полученные данные также подчеркивают потенциал димерных ДНК-лигандов как безопасных и эффективных инструментов для таргетной регуляции ферментов репарации ДНК.

Список литературы / References

Неудачина Л., Лакиза Н. Физико-химические основы применения координационных соединений. Екатеринбург, 2014 [Neudachina L., Lakiza N. Physico-Chemical Principles of the Use of Coordination Compounds. Ekaterinburg, 2014 (in Russian)]

Alagoz M., Wells O.S., El-Khamisy S.F. TDP1 deficiency sensitizes human cells to base damage via distinct topoisomerase I and PARP mechanisms with potential applications for cancer therapy. *Nucleic Acids Res*. 2014;42(5):3089-3103. doi 10.1093/nar/gkt1260

Alford R.F., Leaver-Fay A., Jeliazkov J.R., O'Meara M.J., DiMaio F.P., Park H., Shapovalov M.V., … Das R., Baker D., Kuhlman B., Kortemme T., Gray J.J. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*. 2017; 13(6):3031-3048. doi 10.1021/acs.jctc.7b00125

Allen W.J., Balius T.E., Mukherjee S., Brozell S.R., Moustakas D.T., Lang P.T., Case D.A., Kuntz I.D., Rizzo R.C. DOCK 6: impact of new features and current docking performance. *J Comput Chem*. 2015;36(15):1132-1156. doi 10.1002/jcc.23905

Arutyunyan A.F., Kostyukov A.A., Korolev S.P., Gottikh M.B., Kaluzhny D.N., Susova O.Yu., Zhuze A.L. DNA sequence-specific ligands. 19. Synthesis, spectral properties, virological and biochemical studies of $DB_3(n)$ fluorescent dimeric trisbenzimidazoles. *Mol Biol*. 2023a;57(3):512-521. doi 10.1134/s0026893323030020

Arutyunyan A.F., Kostyukov A.A., Lushpa V.A., Mineev K.S., Korolev S.P., Gottikh M.B., Klimova R.R., Kushch A.A., Kalabina K.V., Susova O.Yu., Zhuze A.L. DNA sequence-specific ligands. XX. Synthesis, spectral properties, virological and biochemical studies of fluorescent dimeric trisbenzimidazoles $DB_3P(n)$. *Med Chem Res*. 2023b;32(3):587-599. doi 10.1007/s00044-023-03017-x

Chaudhury S., Lyskov S., Gray J.J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*. 2010;26(5):689-691. doi 10.1093/bioinformatics/btq007

Chernyshova I., Vasil'eva I., Moor N., Ivanisenko N., Kutuzov M., Abramova T., Zakharenko A., Lavrik O. Aminomethylmorpholino nucleosides as novel inhibitors of PARP1 and PARP2: experimental and molecular modeling analyses of their selectivity and mechanism of action. *Int J Mol Sci*. 2024;25(23):12526. doi 10.3390/ijms252312526

Chowdhuri S.P., Das B.B. Top1-PARP1 association and beyond: from DNA topology to break repair. *NAR Cancer*. 2021;3(1):zcab003. doi 10.1093/narcan/zcab003

Comeaux E.Q., van Waardenburg R.C. Tyrosyl-DNA phosphodiesterase I resolves both naturally and chemically induced DNA adducts and its potential as a therapeutic target. *Drug Metab Rev*. 2014;46(4):494-507. doi 10.3109/03602532.2014.971957

Conda-Sheridan M., Reddy P.V.N., Morrell A., Cobb B.T., Marchand C., Agama K., Chergui A., Renaud A., Stephen A.G., Bindu L.K., Pommier Y., Cushman M. Synthesis and biological evaluation of indenoisoquinolines that inhibit both tyrosyl-DNA phosphodiesterase I (Tdp1) and topoisomerase I (Top1). *J Med Chem*. 2013;56(1):182-200. doi 10.1021/jm3014458

Curtin N.J., Szabo C. Poly(ADP-ribose) polymerase inhibition: past, present and future. *Nat Rev Drug Discov*. 2020;19(10):711-736. doi 10.1038/s41573-020-0076-6

Das B.B., Huang S.N., Murai J., Rehman I., Amé J.-C., Sengupta S., Das S.K., Majumdar P., Zhang H., Biard D., Majumder H.K., Schreiber V., Pommier Y. PARP1-TDP1 coupling for the repair of topoisomerase I-induced DNA damage. *Nucleic Acids Res*. 2014;42(7): 4435-4449. doi 10.1093/nar/gku088

Dyrkheeva N., Anarbaev R., Lebedeva N., Kuprushkin M., Kuznetsova A., Kuznetsov N., Rechkunova N., Lavrik O. Human tyrosyl-DNA phosphodiesterase 1 possesses transphosphooligonucleotidation activity with primary alcohols. *Front Cell Dev Biol*. 2020;8:604732. doi 10.3389/fcell.2020.604732

Dyrkheeva N.S., Filimonov A.S., Luzina O.A., Orlova K.A., Chernyshova I.A., Kornienko T.E., Malakhova A.A., … Burakova E.A., Stetsenko D.A., Zakian S.M., Salakhutdinov N.F., Lavrik O.I. New hybrid compounds combining fragments of usnic acid and thioether are inhibitors of human enzymes TDP1, TDP2 and PARP1. *Int J Mol Sci*. 2021;22(21):11336. doi 10.3390/ijms222111336

Eastman P., Galvelis R., Peláez R.P., Abreu C.R.A., Farr S.E., Gallicchio E., Gorenko A., … Wang Y., Zhang I., Chodera J.D., De Fabritiis G., Markland T.E. OpenMM 8: molecular dynamics simulation with machine learning potentials. *J Phys Chem B*. 2024;128(1):109-116. doi 10.1021/acs.jpcb.3c06662

Elsayed W., El-Shafie L., Hassan M.K., Farag M.A., El-Khamisy S.F. Isoeugenol is a selective potentiator of camptothecin cytotoxicity in vertebrate cells lacking TDP1. *Sci Rep*. 2016;6(1):26626. doi 10.1038/srep26626

Fam H.K., Walton C., Mitra S.A., Chowdhury M., Osborne N., Choi K., Sun G., … Aparicio S., Triche T.J., Bond M., Pallen C.J., Boerkoel C.F. TDP1 and PARP1 deficiency are cytotoxic to rhabdomyosarcoma cells. *Mol Cancer Res*. 2013;11(10):1179-1192. doi 10.1158/1541-7786.mcr-12-0575

Flörkemeier I., Hillmann J.S., Weimer J.P., Hildebrandt J., Hedemann N., Rogmans C., Dempfle A., Arnold N., Clement B., Bauerschlag D.O. Combined PARP and dual topoisomerase inhibition potentiates genome instability and cell death in ovarian cancer. *Int J Mol Sci*. 2022;23(18):10503. doi 10.3390/ijms231810503

Hoch N.C., Polo L.M. ADP-ribosylation: from molecular mechanisms to human disease. *Genet Mol Biol*. 2019;43(Suppl.1):e20190075. doi 10.1590/1678-4685-GMB-2019-0075

Hu D.-X., Tang W.-L., Zhang Y., Yang H., Wang W., Agama K., Pommier Y., An L.-K. Synthesis of methoxy-, methylenedioxy-, hydroxy-, and halo-substituted benzophenanthridinone derivatives as DNA topoisomerase IB (TOP1) and tyrosyl-DNA phosphodiesterase 1 (TDP1) inhibitors and their biological activity for drug-resistant cancer. *J Med Chem*. 2021;64(11):7617-7629. doi 10.1021/acs.jmedchem.1c00318

Ivanov A.A., Koval V.S., Susova O.Yu., Salyanov V.I., Oleinikov V.A., Stomakhin A.A., Shalginskikh N.A., Kvasha M.A., Kirsanova O.V., Gromova E.S., Zhuze A.L. DNA specific fluorescent symmetric dimeric bisbenzimidazoles DBP(n): the synthesis, spectral properties, and biological activity. *Bioorg Med Chem Lett*. 2015;25(13):2634-2638. doi 10.1016/j.bmcl.2015.04.087

Jing C.-B., Fu C., Prutsch N., Wang M., He S., Look A.T. Synthetic lethal targeting of TET2-mutant hematopoietic stem and progenitor cells (HSPCs) with TOP1-targeted drugs and PARP1 inhibitors. *Leukemia*. 2020;34(11):2992-3006. doi 10.1038/s41375-020-0927-5

Johannes J.W., Balazs A.Y.S., Barratt D., Bista M., Chuba M.D., Cosulich S., Critchlow S.E., … Xue L., Yao T., Zhang K., Zhang A.X., Zheng X. Discovery of 6-Fluoro-5-{4-[(5-fluoro-2-methyl-3-oxo-3,4-dihydroquinoxalin-6-yl)methyl]piperazin-1-yl}-*N*-methylpyridine-2-carboxamide (AZD9574): a CNS-penetrant, PARP1-selective inhibitor. *J Med Chem*. 2024;67(24):21717-21728. doi 10.1021/acs.jmedchem.4c01725

Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., … Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi 10.1038/s41586-021-03819-2

Jurrus E., Engel D., Star K., Monson K., Brandi J., Felberg L.E., Brookes D.H., … Krasny R., Wei G., Holst M.J., McCammon J.A., Baker N.A. Improvements to the APBS biomolecular solvation software suite. *Protein Sci*. 2018;27(1):112-128. doi 10.1002/pro.3280

Kawale A.S., Povirk L.F. Tyrosyl-DNA phosphodiesterases: rescuing the genome from the risks of relaxation. *Nucleic Acids Res*. 2018;46(2):520-537. doi 10.1093/nar/gkx1219

Kim D.-S., Camacho C.V., Kraus W.L. Alternate therapeutic pathways for PARP inhibitors and potential mechanisms of resistance. *Exp Mol Med*. 2021;53(1):42-51. doi 10.1038/s12276-021-00557-3

Kim J.W., Min A., Im S.-A., Jang H., Kim Y.J., Kim H.-J., Lee K.-H., Kim T.-Y., Lee K.W., Oh D.-Y., Kim J.-H., Bang Y.-J. TDP1 and TOP1 modulation in olaparib-resistant cancer determines the efficacy of subsequent chemotherapy. *Cancers*. 2020;12(2):334. doi 10.3390/cancers12020334

Kornienko T.E., Chepanova A.A., Zakharenko A.L., Filimonov A.S., Luzina O.A., Dyrkheeva N.S., Nikolin V.P., Popova N.A., Salakhutdinov N.F., Lavrik O.I. Enhancement of the antitumor and antimetastatic effect of topotecan and normalization of blood counts in mice with Lewis carcinoma by Tdp1 inhibitors – new usnic acid derivatives. *Int J Mol Sci*. 2024;25(2):1210. doi 10.3390/ijms25021210

Lavrik O.I. PARPs' impact on base excision DNA repair. *DNA Repair*. 2020;93:102911. doi 10.1016/j.dnarep.2020.102911

Lebedeva N.A., Anarbaev R.O., Sukhanova M., Vasil'eva I.A., Rechkunova N.I., Lavrik O.I. Poly(ADP-ribose)polymerase 1 stimulates the AP-site cleavage activity of tyrosyl-DNA phosphodiesterase 1. *Biosci Rep*. 2015;35(4):e00230. doi 10.1042/BSR20140192

Marchand C., Abdelmalak M., Kankanala J., Huang S.-Y., Kiselev E., Fesen K., Kurahashi K., Sasanuma H., Takeda S., Aihara H., Wang Z., Pommier Y. Deazaflavin inhibitors of tyrosyl-DNA phosphodiesterase 2 (TDP2) specific for the human enzyme and active against cellular TDP2. *ACS Chem Biol*. 2016;11(7):1925-1933. doi 10.1021/acschembio.5b01047

Matsuno Y., Hyodo M., Fujimori H., Shimizu A., Yoshioka K. Sensitization of cancer cells to radiation and topoisomerase I inhibitor camptothecin using inhibitors of PARP and other signaling molecules. *Cancers*. 2018;10(10):364. doi 10.3390/cancers10100364

Mirdita M., Schütze K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi 10.1038/s41592-022-01488-1

Moor N.A., Vasil'eva I.A., Anarbaev R.O., Antson A.A., Lavrik O.I. Quantitative characterization of protein-protein complexes involved in base excision DNA repair. *Nucleic Acids Res*. 2015;43(12):6009-6022. doi 10.1093/nar/gkv569

Moretti R., Bender B.J., Allison B., Meiler J. Rosetta and the design of ligand binding sites. In: Stoddard B. (Ed.) Computational Design of Ligand Binding Proteins. Methods in Molecular Biology. Vol. 1414. New York: Humana Press, 2016;47-62. doi 10.1007/978-1-4939-3569-7_4

Murai J., Marchand C., Shahane S.A., Sun H., Huang R., Zhang Y., Chergui A., Ji J., Doroshow J.H., Jadhav A., Takeda S., Xia M., Pommier Y. Identification of novel PARP inhibitors using a cell-based TDP1 inhibitory assay in a quantitative high-throughput screening platform. *DNA Repair*. 2014;21:177-182. doi 10.1016/j.dnarep.2014.03.006

Nguyen T.X., Abdelmalak M., Marchand C., Agama K., Pommier Y., Cushman M. Synthesis and biological evaluation of nitrated 7-, 8-, 9-, and 10-hydroxyindenoisoquinolines as potential dual topoisomerase I (Top1)–tyrosyl-DNA phosphodiesterase I (TDP1) inhibitors. *J Med Chem*. 2015;58(7):3188-3208. doi 10.1021/acs.jmedchem.5b00136

O'Boyle N.M., Banck M., James C.A., Morley C., Vandermeersch T., Hutchison G.R. Open Babel: an open chemical toolbox. *J Cheminform*. 2011;3(1):33. doi 10.1186/1758-2946-3-33

Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25(13):1605-1612. doi 10.1002/jcc.20084

Pommier Y., Leo E., Zhang H., Marchand C. DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. *Chem Biol*. 2010;17(5):421-433. doi 10.1016/j.chembiol.2010.04.012

Pommier Y., Huang S.N., Gao R., Das B.B., Murai J., Marchand C. Tyrosyl-DNA-phosphodiesterases (TDP1 and TDP2). *DNA Repair*. 2014;19:114-129. doi 10.1016/j.dnarep.2014.03.020

Quiroga R., Villarreal M.A. Vinardo: a scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS One*. 2016;11(5):e0155183. doi 10.1371/journal.pone.0155183

Ray Chaudhuri A., Nussenzweig A. The multifaceted roles of PARP1 in DNA repair and chromatin remodelling. *Nat Rev Mol Cell Biol*. 2017;18(10):610-621. doi 10.1038/nrm.2017.53

Salentin S., Schreiber S., Haupt V.J., Adasme M.F., Schroeder M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res*. 2015;43(W1):W443-W447. doi 10.1093/nar/gkv315

Schreiber V., Dantzer F., Ame J.-C., de Murcia G. Poly(ADP-ribose): novel functions for an old molecule. *Nat Rev Mol Cell Biol*. 2006; 7(7):517-528. doi 10.1038/nrm1963

Sherstyuk Y.V., Ivanisenko N.V., Zakharenko A.L., Sukhanova M.V., Peshkov R.Y., Eltsov I.V., Kutuzov M.M., Kurgina T.A., Belousova E.A., Ivanisenko V.A., Lavrik O.I., Silnikov V.N., Abramova T.V. Design, synthesis and molecular modeling study of conjugates of ADP and morpholino nucleosides as a novel class of inhibitors of PARP-1, PARP-2 and PARP-3. *Int J Mol Sci*. 2019;21(1):214. doi 10.3390/ijms21010214

Smith L.M., Willmore E., Austin C.A., Curtin N.J. The novel poly(ADP-ribose) polymerase inhibitor, AG14361, sensitizes cells to topoisomerase I poisons by increasing the persistence of DNA strand breaks. *Clin Cancer Res*. 2005;11(23):8449-8457. doi 10.1158/1078-0432.ccr-05-1224

Sukhanova M.V., Khodyreva S.N., Lavrik O.I. Poly(ADP-ribose) polymerase-1 inhibits strand-displacement synthesis of DNA catalyzed by DNA polymerase β. *Biochemistry* (*Moscow*). 2004;69(5):558-568. doi 10.1023/b:biry.0000029855.68502.fa

Susova O.Y., Karshieva S.S., Kostyukov A.A., Moiseeva N.I., Zaytseva E.A., Kalabina K.V., Zusinaite E., Gildemann K., Smirnov N.M., Arutyunyan A.F., Zhuze A.L. Dimeric bis-benzimidazole-pyrroles $DB_2Py(n)$ – AT-site-specific ligands: synthesis, physicochemical analysis, and biological activity. *Acta Naturae*. 2024;16(1):86-100. doi 10.32607/actanaturae.27327

Szanto M., Yelamos J., Bai P. Specific and shared biological functions of PARP2 – is PARP2 really a lil' brother of PARP1? *Expert Rev Mol Med*. 2024;26:e13. doi 10.1017/erm.2024.14

Wang P., Elsayed M.S.A., Plescia C.B., Ravji A., Redon C.E., Kiselev E., Marchand C., Zeleznik O., Agama K., Pommier Y., Cushman M. Synthesis and biological evaluation of the first triple inhibitors of human topoisomerase 1, tyrosyl-DNA phosphodiesterase 1 (Tdp1), and tyrosyl-DNA phosphodiesterase 2 (Tdp2). *J Med Chem*. 2017;60(8):3275-3288. doi 10.1021/acs.jmedchem.6b01565

Wang R., Lai L., Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*. 2002;16(1):11-26. doi 10.1023/a:1016357811882

Yang H., Zhu X.-Q., Wang W., Chen Y., Hu Z., Zhang Y., Hu D.-X., Yu L.-M., Agama K., Pommier Y., An L.-K. The synthesis of furoquinolinedione and isoxazoloquinolinedione derivatives as selective Tyrosyl-DNA phosphodiesterase 2 (TDP2) inhibitors. *Bioorg Chem*. 2021;111:104881. doi 10.1016/j.bioorg.2021.104881

Yang H., Qin C., Wu M., Wang F., Wang W., Agama K., Pommier Y., Hu D., An L. Synthesis and biological activities of 11- and 12-substituted benzophenanthridinone derivatives as DNA topoisomerase IB and tyrosyl-DNA phosphodiesterase 1 inhibitors. *ChemMedChem*. 2023;18(10):e202200593. doi 10.1002/cmdc.202200593

Zakharenko A., Khomenko T., Zhukova S., Koval O., Zakharova O., Anarbaev R., Lebedeva N., Korchagina D., Komarova N., Vasiliev V., Reynisson J., Volcho K., Salakhutdinov N., Lavrik O. Synthesis and biological evaluation of novel tyrosyl-DNA phosphodiesterase 1 inhibitors with a benzopentathiepine moiety. *Bioorg Med Chem*. 2015;23(9):2044-2052. doi 10.1016/j.bmc.2015.03.020

Zakharenko A.L., Luzina O.A., Chepanova A.A., Dyrkheeva N.S., Salakhutdinov N.F., Lavrik O.I. Natural products and their derivatives as inhibitors of the DNA repair enzyme tyrosyl-DNA phosphodiesterase 1. *Int J Mol Sci*. 2023;24(6):5781. doi 10.3390/ijms24065781

Zeng Z., Sharma A., Ju L., Murai J., Umans L., Vermeire L., Pommier Y., Takeda S., Huylebroeck D., Caldecott K.W., El-Khamisy S.F. TDP2 promotes repair of topoisomerase I-mediated DNA damage in the absence of TDP1. *Nucleic Acids Res*. 2012;40(17):8371-8380. doi 10.1093/nar/gks622

Zhang M., Wang Z., Su Y., Yan W., Ouyang Y., Fan Y., Huang Y., Yang H. TDP1 represents a promising therapeutic target for overcoming tumor resistance to chemotherapeutic agents: progress and potential. *Bioorg Chem*. 2025;154:108072. doi 10.1016/j.bioorg.2024.108072

Zhang X.-R., Wang H.-W., Tang W.-L., Zhang Y., Yang H., Hu D.-X., Ravji A., Marchand C., Kiselev E., Ofori-Atta K., Agama K., Pommier Y., An L.-K. Discovery, synthesis, and evaluation of oxynitidine derivatives as dual inhibitors of DNA topoisomerase IB (TOP1) and tyrosyl-DNA phosphodiesterase 1 (TDP1), and potential antitumor agents. *J Med Chem*. 2018;61(22):9908-9930. doi 10.1021/acs.jmedchem.8b00639

Zhang Y., He X., Yang H., Liu H., An L. Robustadial A and B from *Eucalyptus globulus* Labill. and their anticancer activity as selective tyrosyl-DNA phosphodiesterase 2 inhibitors. *Phytotherapy Res*. 2021;35(9):5282-5289. doi 10.1002/ptr.7207

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 07.08.2025. После доработки 30.09.2025. Принята к публикации 30.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1109-1121

doi 10.18699/vjgb-25-115

Компьютерное моделирование пространственной динамики и первичной генетической дивергенции в системе популяций на кольцевом ареале

М.П. Кулаков [1], О.Л. Жданова [2], Е.Я. Фрисман [1]

[1] Институт комплексного анализа региональных проблем Дальневосточного отделения Российской академии наук, Биробиджан, Россия
[2] Институт автоматики и процессов управления Дальневосточного отделения Российской академии наук, Владивосток, Россия
k_matvey@mail.ru

Аннотация. Одна из ключевых задач современной эволюционной биологии – изучение процессов, приводящих к первичному разделению (дивергенции) популяций на различающиеся генотипами группы особей. Для дивергенции очевидно необходима репродуктивная изоляция, которая делает невозможным контакт особей или существенно затрудняет обмен генетической информацией между популяциями. Настоящее исследование изучает возможность изоляции, в основе которой лежат не очевидные географические барьеры, удаленность популяций или экологическая специализация, а лишь наследственные механизмы, дрейф и поток генов, а также отбор против гетерозигот. Для этого предложена и исследована динамическая модель с дискретным временем, которая описывает динамику частот и численностей в системе миграционно связанных лимитированных популяций. Рассматривается панмиктичная популяция с менделевскими правилами наследования, монолокусным отбором, действием плотностно-зависимых факторов, лимитирующих рост численности. Особи свободно скрещиваются и перемещаются вдоль одномерного кольцевого ареала. Модель верифицирована с использованием данных эксперимента над ящичной системой популяций дрозофил *Drosophila melanogaster*, проведенного под руководством Ю.П. Алтухова. При достаточно простых предположениях модель описывает некоторые механизмы возникновения и сохранения на однородном ареале существенных генетических различий (первичной генетической дивергенции), сопровождаемых неоднородностью в частотах аллелей и численностях. В этом случае формируется несколько больших групп генетически однородных субпопуляций, развивающихся независимо. В местах их контакта активно идет гибридизация, а полиморфизм сохраняется за счет миграции с сопредельных однородных участков. Обнаружено, что устойчиво поддерживать такое пространственное распределение может только дизруптивный (разрывающий) отбор, направленный против гетерозигот. При движущем отборе дивергенция существует непродолжительное время, как часть переходного процесса. За счет пониженной приспособленности гетерозигот (гибридов) и низкой скорости роста на этих участках (зонах гибридизации) существенно затрудняется обмен генами между смежными участками с противоположными гомозиготными генотипами (фенотипами). В результате зоны гибридизации выполняют функцию географического барьера, который фактически останавливает обмен генов между разными группами в случае смежной симпатрии.

Ключевые слова: метапопуляция; миграция; пространственно-временная динамика; математическое моделирование; генетическая дивергенция; поток генов; гибридные зоны; изоляция

Для цитирования: Кулаков М.П., Жданова О.Л., Фрисман Е.Я. Компьютерное моделирование пространственной динамики и первичной генетической дивергенции в системе популяций на кольцевом ареале. *Вавиловский журнал генетики и селекции*. 2025;29(7):1109-1121. doi 10.18699/vjgb-25-115

Финансирование. Работа выполнена в рамках государственного задания Института комплексного анализа региональных проблем ДВО РАН.

Computer modeling of spatial dynamics and primary genetic divergence for a population system in a ring areal

M.P. Kulakov [1], O.L. Zhdanova [2], E.Ya. Frisman [1]

[1] Institute for Complex Analysis of Regional Problems of the Far Eastern Branch of the Russian Academy of Sciences, Birobidzhan, Russia
[2] Institute of Automation and Control Processes of the Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia
k_matvey@mail.ru

Abstract. One of the main goals of modern evolutionary biology is to understand the mechanisms that lead to the initial differentiation (primary divergence) of populations into groups with genetic traits. This divergence requires reproductive isolation, which prevents or hinders contact and the exchange of genetic material between populations. This study explores the potential for isolation based not on obvious geographical barriers, population distance, or

© Кулаков М.П., Жданова О.Л., Фрисман Е.Я., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

ecological specialization, but rather on hereditary mechanisms, such as gene drift and flow and selection against heterozygous individuals. To this end, we propose and investigate a dynamic discrete-time model that describes the dynamics of frequencies and numbers in a system of limited populations coupled by migrations. We consider a panmictic population with Mendelian inheritance rules, one-locus selection, and density-dependent factors limiting population growth. Individuals freely mate and randomly move around a one-dimensional ring-shaped habitat. The model was verified using data from an experiment on the box population system of *Drosophila melanogaster* performed by Yu.P. Altukhov et al. With rather simple assumptions, the model explains some mechanisms for the emergence and preservation of significant genetic differences between subpopulations (primary genetic divergence), accompanied by heterogeneity in allele frequencies and abundances within a homogeneous area. In this scenario, several large groups of genetically homogeneous subpopulations form and independently develop. Hybridization occurs at contact sites, and polymorphism is maintained through migration from genetically homogeneous nearby sites. It was found that only disruptive selection, directed against heterozygous individuals, can sustainably maintain such a spatial distribution. Under directional selection, divergence may occur for a short time as part of the transitional evolutionary process towards the best-adapted genotype. Because of the reduced adaptability of heterozygous (hybrid) individuals and low growth rates in these sites (hybrid zones), gene flow between adjacent sites with opposite genotypes (phenotypes) is significantly impeded. As a result, the hybrid zones can become effective geographical barriers that prevent the genetic flow between coupled subpopulations.

Key words: metapopulation; migration; spatiotemporal dynamics; mathematical modeling; genetic divergence; gene flow; hybrid zones; isolation

For citation: Kulakov M.P., Zhdanova O.L., Frisman E.Ya. Computer modeling of spatial dynamics and primary genetic divergence for a population system in a ring areal. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1109-1121. doi 10.18699/vjgb-25-115

Введение

Генетическая дивергенция невозможна без репродуктивной изоляции и остановки потока генов между родственными популяциями. В случае пространственной удаленности некоторых участков (аллопатрии) обмен генами может прерваться из-за больших расстояний, которые невозможно преодолеть за время жизни особей, или очевидных географических преград. Однако часто популяции одного вида с существенными различиями в признаках проживают на одной территории (симпатрия) или на примыкающих друг к другу участках (смежная симпатрия), контактируют между собой и образуют жизнеспособных и вполне фертильных гибридов. Тем не менее при этом не происходит размывания признаков. Существует множество механизмов, способствующих репродуктивной изоляции и дивергенции признаков разных форм в этом случае. Одним из таких механизмов является отбор против гибридов, выражающийся в их пониженной приспособленности по сравнению с исходными формами.

Известно множество примеров репродуктивной изоляции, при которой разные субпопуляции даже в случае симпатрии успели накопить достаточно отличий и выработать эффективные меры, препятствующие гибридизации. Например, опознавательные сигналы, связанные с фонетическими признаками и используемые в брачном поведении, способствуют закреплению только крайних форм признака. Так, брачные призывы родственных видов некоторых лягушек (*Microhyla carolinensis* и *M. olivacea*, *Litoria verreauxii* и *L. v. alpina*) сильно различаются в зоне контактов ареалов, но мало отличаются в местах, где они не проживают совместно (Blair, 1955a; Littlejohn, 1965; Smith et al., 2003). Кроме того, размеры тел разных форм лягушек сильнее отличаются в зонах контакта, что усложняет их спаривание (Blair, 1955b).

Презиготическая изоляция симпатричных форм одного вида или подвидов часто сопровождается экологической специализацией, которая не препятствует копулятивному поведению особей с разными признаками и их гибридизации, а лишь делает ее маловероятной. Например, периоды половой активности двух рас пестрокрылки *Rhagoletis pomonella* связаны со временем плодоношения деревьев, на плодах которых они родились и откладывают яйца, – боярышника и яблони (Filchak et al., 2000). Эти две расы мух отличаются инверсией в сенсорной обработке ключевых запахов плодов: если одни особи тянутся к яблокам и сторонятся боярышника, то другие выбирают боярышник и избегают яблок, что существенно затрудняет их контакт (Tait et al., 2021). До конца не ясны брачные предпочтения гибридов. Однако при скрещивании двух рас *R. pomonella* в лабораторных условиях регистрируется более низкая частота зачатий (Yee, Goughnour, 2011), что сигнализирует о некотором отборе против гибридов и устойчивой дивергенции в природе, вызванной специализацией мух.

Можно привести несколько примеров, когда гибридизация не имеет явных негативных последствий в виде пониженной приспособленности или катастрофического снижения репродуктивного успеха гибридов (гетерозигот). Так, внутривидовая изменчивость некоторых видов птиц часто выражается различием окраски перьев. При этом наблюдается четкая дивергенция признаков между разными частями протяженного ареала и длительное время существуют устойчивые зоны гибридизации в местах контактов ареалов. Хорошо известны популяции черной и серой вороны (*Corvus corone*, *C. cornix*) с сибирской (между реками Обь и Енисей) и европейской зонами гибридизации (Haring et al., 2012; Poelstra et al., 2014; Kryukov, 2019; Блинов, Железнова, 2020), гибридизация двух подвидов шилоклювого дятла (*Colaptes auratus cafer* и *C. a. auratus*) в США (Aguillon, Rohwer, 2022). Другой пример – гибридизация большой (*Parus major*) и восточной (*P. minor*) синиц на территории Приамурья (Капитонова и др., 2012).

Для вороны известен генетический механизм, поддерживающий изоляцию на основе врожденных брачных предпочтений: в качестве партнеров они стараются выбрать себе подобных, а не экзотических чужаков. Процесс

формирования фенотипов серой и черной вороны связан с хромосомной инверсией, которая одновременно затронула как окраску перьев, так и зрительное восприятие цвета перьев, а также некоторые аспекты репродуктивного поведения (Poelstra et al., 2014). Однако в зонах гибридизации, возникших, по всей видимости, одновременно с разной окраской, брачные предпочтения оказались более разнообразны, и полной изоляции не происходит. Это связано с тем, что инвертированный участок серой вороны наследуется целиком и не рекомбинирует с гомологичными участками черной вороны.

Одной из простых моделей, позволяющей исследовать генетическую дивергенцию, является линейная цепь или кольцо из частично изолированных субпопуляций, которые обмениваются генами. Результаты изучения таких моделей показывают, что поток генов между связанными или слабосвязанными миграцией популяциями способен привести к устойчивой географической изменчивости признака и сохранению зон гибридизации только в случае дизруптивного (разрывающего) отбора. В случае движущего отбора дивергенция будет невозможной и станет появляться лишь как часть переходного процесса при специально подобранных начальных условиях (Базыкин, 1972; Фрисман, 1986; Yeaman, Otto, 2011; Láruson, Reed, 2016). Для цепей миграционно связанных популяций с разной топологией было обнаружено, что дивергенция чаще возникает в линейных цепях и кольцах и реже – в полносвязанных сетях (с глобальной связью) (Láruson, Reed, 2016; Sundqvist et al., 2016).

Вместе с тем для многих реальных популяций с выраженной дивергенцией признаков, а порой и известными изолирующими механизмами сложно указать конкретный адаптивный признак, по которому действует дизруптивный отбор. Им может оказаться какой-то «скрытый» признак (например, врожденные факторы иммунитета, комплекса гистосовместимости и т. п.), прямо не связанный с фиксируемыми у особей внешними признаками, которые в настоящем демонстрируют географическую изменчивость (окраска перьев у птиц, узор на коже или шкуре, форма и размер клюва, особенности поведения и пр.). Регистрируемая дивергенция признака прямо не указывает на причины и, тем более, на тип отбора, способствовавший дивергенции в прошлом. Но он может оказаться удачно «сцепленным» с наблюдаемым признаком и выступать в качестве некоего индикатора приспособленности или маркера. Особенно это касается видов с обширными ареалами, неоднородными условиями среды, выраженной дивергенцией признаков и высокой степенью полиморфизма (Orsini et al., 2008; Murphy et al., 2010).

Настоящая статья входит в серию работ, исследующих базовые механизмы первичной генетической дивергенции в системах миграционно связанных панмиктичных популяций диплоидных организмов с отбором, направленным против гетерозигот (Жданова, Фрисман, 2023; Кулаков, Фрисман, 2025). Предлагается динамическая модель с дискретным временем, которая учитывает действие плотностно-зависимых факторов, ограничивающих рост численности, дрейф генов (через определенные возмущения начальных условий), естественный отбор и миграцию особей между смежными участками. Модель верифицируется на основе данных лабораторного эксперимента с ящичными популяциями дрозофил (*Drosophila melanogaster*), проведенного под руководством Ю.П. Алтухова (Алтухов и др., 1979; Алтухов, Бернашевская, 1981; Алтухов, 2003). В этих экспериментах была показана выраженная дивергенция аллельного состава по локусу α-глицерофосфатдегидрогеназа (α-Gdph) между группами смежных ящиков.

В нашей работе на основе множественной компьютерной реализации модели анализируются процессы отбора и миграции (поток генов), которые сформировали и поддерживали данное пространственное распределение частот. Исследуется роль зон гибридизации с высокой долей гетерозиготных особей по аллелю гена α-*Gdph*, которые разделяют мономорфные группы ящиков и не дают наиболее приспособленному генотипу закрепиться на всем кольцевом ареале.

Материал и методы

Исследование проведено на основе оригинальной математической модели – системы связанных нелинейных отображений (уравнений с дискретным временем), описывающих динамику частот и численностей генотипов. Описание миграции особей и потока генов между связанными субпопуляциями выполнено с помощью матрицы миграции со случайными коэффициентами. В качестве генератора случайных чисел был выбран генератор MT19937 (Matsumoto et al., 1998), доступный в программной библиотеке численных вычислений GSL. Этот генератор имеет чрезвычайно длинный период ($\sim 10^{6000}$), низкую корреляцию и проходит большинство статистических тестов на случайность генерируемых последовательностей псевдослучайных чисел.

Для верификации модели использованы данные из эксперимента с ящичными популяциями дрозофил (*D. melanogaster*), выполненного коллективом под руководством Ю.П. Алтухова. Данные представлены частотами аллелей локуса, кодирующего фермент α-Gdph, а также значениями численностей мух в разных ящиках на разных этапах эксперимента (Алтухов, 2003). Оценка параметров выполнена методом наименьших квадратов.

Численные эксперименты проведены с применением авторского комплекса программ, включающего компьютерную реализацию математической модели, визуализацию результатов расчетов и анализ типов динамики.

Модель локальной популяции

Рассмотрим популяцию диплоидных организмов, между двумя смежными поколениями которой наблюдается следующая последовательность элементарных популяционных процессов: формирование зигот из гамет, естественный отбор зигот (особей), миграция (расселение) зигот между смежными популяциями, продуцирование новых гамет. Будем рассматривать популяции, все адаптивное разнообразие которых определяется одним диаллельным локусом с аллеломорфами *A* и *a*, наследуемыми кодоминантно. Фенотип особей жестко определяется ее генотипом. Популяция панмиктична, и в ней действуют менделевские правила наследования. Это означает, что в популяции присутствуют особи с генотипами *AA*, *Aa* и *aa*,

которые в момент времени t имеют численности $N_1(t)$, $N_2(t)$ и $N_3(t)$ и частоты $q_1(t) = N_1(t)/N(t)$, $q_2(t) = N_2(t)/N(t)$ и $q_3(t) = N_3(t)/N(t)$, где $N(t) = N_1(t)+N_2(t)+N_3(t)$ – общая численность.

Будем считать, что рассматриваемые генотипы отличаются репродуктивными возможностями, выраженными различием продукции гамет или выживаемости особей. Через g_{AA}, g_{Aa} и g_{aa} обозначим интенсивность производства гамет особей с генотипами *AA*, *Aa* и *aa* соответственно, с учетом гибели части гамет до момента объединения в зиготы следующего поколения. Через W_{AA}, W_{Aa} и W_{aa} обозначим долю зигот (особей) соответствующего генотипа, переживших отбор и потенциально способных к миграции.

В случае, когда интенсивность производства гамет не зависит от генотипов родителей, т. е. $g_{AA} = g_{Aa} = g_{aa} = g$, уравнения частот генотипов локальной панмиктичной популяции имеют вид:

$$\begin{cases} q_1(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_1) q_1^*(t), \\ q_2(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_2) q_2^*(t), \\ q_3(t+1) = \dfrac{1}{\overline{W}(t)} \cdot (1+s_3) q_3^*(t), \end{cases} \quad (1)$$

где $q_1^*(t) = (q_1(t) + q_2(t)/2)^2$,

$q_2^*(t) = 2(q_1(t) + q_2(t)/2)(q_3(t) + q_2(t)/2)$,

$q_3^*(t) = (q_3(t) + q_2(t)/2)^2$ – частоты генотипов сразу после объединения гамет в зиготы, но до действия отбора и переселения (миграции) особей (Жданова, Фрисман, 2023; Кулаков, Фрисман, 2025). Параметр s_k – коэффициент отбора зиготы с соответствующим генотипом, который связывает приспособленность W_k каждого генотипа и скорость продукции гамет g_k следующим образом: $1+s_k = gW_k$ ($k = AA, Aa, aa$). В системе (1) нормировочный множитель

$$\overline{W}(t) = 1 + s_1 q_1^*(t) + s_2 q_2^*(t) + s_3 q_3^*(t) \quad (2)$$

равен средней (обобщенной) приспособленности, его значение определяет скорость роста популяции. При отсутствии факторов, сдерживающих рост популяции, ее численность изменяется согласно уравнению

$$N(t+1) = \overline{W}(t) N(t). \quad (3)$$

Число особей каждого генотипа определяется соотношением $N_k(t+1) = q_k(t+1)N(t+1) = (1+s_k)q_k^*(t)N(t+1)$ ($k = AA, Aa, aa$).

Из всех вариантов отбора, определяемых значениями s_1, s_2 и s_3, наиболее интересен дизруптивный (разрывающий) отбор ($s_2 < s_1$ и $s_2 < s_3$), при котором система (1) демонстрирует бистабильность. Как показывают ранние исследования, именно в этом случае наблюдаются условия возникновения и закрепления генетических различий на разных участках однородного ареала даже без учета экологических и прочих факторов.

Вместе с тем на больших временных масштабах рост численностей реальных эволюционирующих популяций ограничен действием экологических факторов. В модели (1)–(3) ограничение роста может быть описано нелинейной зависимостью параметров отбора или гаметопродукции от численностей генотипов или общей численности (плотности населения). Несложно показать, что в случае равенства скоростей производства гамет у разных генотипов ситуация лимитирования продуцирования гамет (g) или интенсивности отбора (W_{ij}) неотличима при конкуренции за общий ресурс. Таким образом, без нарушения общности можно положить, что

$$W_{ij} = w_{ij} F(N), \quad (4)$$

где w_{ij} – максимально возможная доля особей генотипа *ij* (*AA*, *Aa* или *aa*), оставшихся после естественного отбора в условиях минимальной конкуренции (при низкой плотности); F – функция, описывающая действие плотностно-зависимого лимитирования; N – общая численность. С учетом (4) уравнения динамики частот (1) не меняют своего вида (за исключением замены W_{ij} на w_{ij} и gW_{ij} на $1+s_k$), в то время как уравнения численностей (3) будут содержать нелинейную зависимость от плотности:

$$N(t+1) = \overline{W}(t) N(t) F(N(t)). \quad (5)$$

В популяции диплоидных организмов для обмена гаметами часто необходим контакт особей, вероятность которого значительно падает при низкой плотности, т. е. существует прямая связь между средней индивидуальной приспособленностью и плотностью населения – эффект Олли (Allee, 1958). В результате при численности ниже некоторой критической N_0 рост популяции невозможен, так же как и невозможен эффективно работающий отбор, и действует только дрейф генов. Поэтому для описания плотностно-зависимого лимитирования будем пользоваться функцией вида

$$F(N) = a\varphi(N)\exp(-N/K), \quad (6)$$

где $\varphi(N)$ – сигмоидная функция вида:

$$\varphi(N) = \frac{1}{1+e^{-h(N-N_0)}}, \quad (7)$$

а параметр $h \geq 2$ задает крутизну кривой в точке N_0. Значение N_0 определяет минимальную эффективную численность, необходимую для простого воспроизводства популяции (1 к 1). Параметр K определяет экологическую емкость среды, a – среднее число потомков, приходящихся на одну особь при единичной средней приспособленности. Два этих значения определяют стационарную (равновесную) численность $\overline{N} \approx K\ln(a\overline{W})$. С учетом (7) уравнение динамики численности (5) имеет вид:

$$N(t+1) = rN(t)\varphi(N(t))\exp(-N(t)/K), \quad (8)$$

где $r = a\overline{W}(t)$ – суммарный репродуктивный потенциал всех генотипов.

При $r > 1$ функция (8) имеет три неподвижные точки [$N(t+1) = N(t)$]: 0, N_0 и $\overline{N} \approx K\ln(a\overline{W})$. При $N < N_0$ число выживших детей $N(t+1)$ меньше числа родителей $N(t)$, и численность неизбежно падает, что соответствует сильному эффекту Олли. При $N_0 < N < \overline{N}$ и $r > 1$ нет недостатка производителей, и численность растет. При $N > \overline{N}$ размер популяции превосходит емкость среды, и численность падает до стационарного значения $\overline{N}$.

Перейдем к рассмотрению систем миграционно связанных популяций, каждая из которых развивается описанным образом.

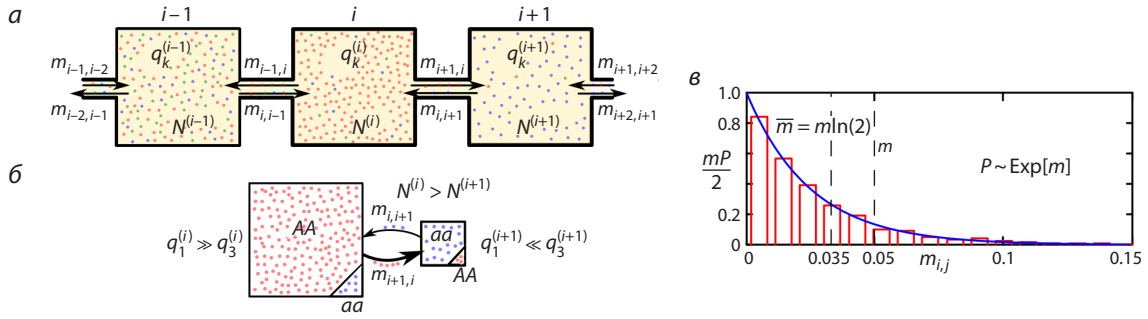


Рис. 1. *а* – схема ящичных популяций, связанных узкими миграционными коридорами; *б* – иллюстрация того, что потоки генов между различающимися по размеру популяциями могут существенно изменить генотипы в маленькой популяции, но не повлиять на большую; *в* – плотность вероятности экспоненциально распределенного случайного значения коэффициента миграции $m_{i,j}$.

Модель динамики с учетом потока генов

Одним из методов исследования динамики и эволюции распределенных популяционных систем (метапопуляций) является проведение лабораторных экспериментов с ящичными популяциями, связанными узкими коридорами. При этом жестко контролируются условия среды, параметры роста, отбора и миграции. Как правило, связанные ящики (камеры) образуют замкнутые в кольцо цепочки субпопуляций, которые обмениваются небольшим числом мигрантов (рис. 1, *а*). Такие популяционные системы воспроизводят в лабораторных условиях, например, для *D. melanogaster* (Алтухов и др., 1979; Алтухов, Бернашевская, 1981; Dey, Joshi, 2006) и *Escherichia coli* (Keymer et al., 2006).

Рассмотрим систему из n ящиков (субпопуляций), каждому из которых присвоен номер i от 1 до n (см. рис. 1, *а*). Обозначим через $0 \le m_{i,j} < 1$ долю особей от общей численности (коэффициент миграции), которые совершают перемещения из ящика с номером j в i. Эмигранты состоят из особей трех рассматриваемых генотипов, т. е. верно $m_{i,j}N^{(j)} = m_{i,j}q_{AA}^{(j)}N^{(j)} + m_{i,j}q_{Aa}^{(j)}N^{(j)} + m_{i,j}q_{aa}^{(j)}N^{(j)}$.

Тогда для системы миграционно связанных субпопуляций уравнения динамики частот (1) и численностей (8) приобретут вид:

$$\begin{cases} q_k^{(i)}(t+1) = \dfrac{1+s_k}{G^{(i)}(t)}\Big(q_k^{(i)^*}(t)N^{(i)}(t)\big(1-m_{i-1,i}-m_{i+1,i}\big) + \\ \qquad + q_k^{(i-1)^*}(t)N^{(i-1)}(t)m_{i,i-1} + q_k^{(i+1)^*}(t)N^{(i+1)}(t)m_{i,i+1}\Big), \\ N^{(i)}(t+1) = N^{(i)^*}(t+1)\big(1-m_{i-1,i}-m_{i+1,i}\big) + \\ \qquad + N^{(i-1)^*}(t+1)m_{i,i-1} + N^{(i+1)^*}(t+1)m_{i,i+1}, \end{cases} \quad (9)$$

где k = 1, 2, 3 – номер группы особей с генотипом *AA*, *Aa* и *aa*; $q_k^{(i)^*}$ – частоты до миграции; $N^{(i)^*}(t+1) = a\overline{W}^{(i)}(t)N^{(i)}(t)F\big(N^{(i)}(t)\big)$ – численность i-й субпопуляции после отбора, но до миграции. Нормировочный коэффициент G равен:

$$G^{(i)}(t) = \overline{W}^{(i)}(t)\big(1-m_{i-1,i}-m_{i+1,i}\big)N^{(i)}(t) + \\ + \overline{W}^{(i-1)}(t)m_{i,i-1}N^{(i-1)}(t) + \overline{W}^{(i+1)}(t)m_{i,i+1}N^{(i+1)}(t), \quad (10)$$

где $\overline{W}^{(i)}(t) = 1 + s_1q_1^{(i)^*}(t) + s_2q_2^{(i)^*}(t) + s_3q_3^{(i)^*}(t)$. Для замыкания цепочки субпопуляций в кольцо мы полагаем, что 1-й ящик связан со 2-м и n-м, n-й – с (n–1)-м и 1-м, т. е. для номера участка применяется следующее отображение: $i \to i \bmod n$. В системе (9) множитель $(1 - m_{i-1,i} - m_{i+1,i})$ – доля особей, оставшихся в i-м ящике после эмиграции в два соседних ящика; $m_{i,i-1}$ и $m_{i,i+1}$ – доля особей (i–1)-й и (i+1)-й субпопуляций, переселяющихся в i-й ящик.

Как видно из уравнений (9), интенсивность потока генов из каждой субпопуляции зависит не только от частот генотипов родного участка, как было в локальном случае, но и от абсолютной численности. Это явно следует из предположения, что мигранты состоят из представителей всех трех генотипов. В результате поток мигрантов из малочисленной популяции, но состоящих, например, из гомозигот *aa*, оказывает слабое влияние на большую популяцию, состоящую преимущественно из гомозигот *AA* (см. рис. 1, *б*). Аналогично поток из многочисленной популяции способен быстро изменить частоты генотипов даже при низком коэффициенте миграции. Отметим, что описанный механизм в ряде случаев явно нарушает условие панмиктичности в масштабе всей метапопуляции. В этом случае изменение частот несопоставимых по размеру субпопуляций в большей степени определяется генетическим составом иммигрантов, а не случайным скрещиванием, дрейфом генов или отбором.

Поток генов и особей из одной субпопуляции в другую может быть строго детерминированным или чисто случайным процессом. Так, в первом случае число мигрантов и их генетический состав зависят от плотности исходящего и принимающего участка и степени перенаселения там или определяются действием внешних факторов среды, например потоками пищи (таксис) или энергии (фототаксис). Во втором случае случайным оказывается как направление, так и доля мигрантов, которая случайно изменяется от поколения к поколению.

Далее будем рассматривать только случайную миграцию. Для ее описания поступим следующим образом. Для каждого номера сезона t выберем два случайных числа $m_{i-1,i}$ и $m_{i+1,i}$, равных долям особей, покидающих участок с номером i и переселяющихся на смежные участки. Будем пренебрегать возможностью более дальней миграции. Каждую пару значений $m_{i-1,i}$ и $m_{i+1,i}$ будем генерировать независимо, используя генератор экспоненциально распределенной случайной величины с математическим ожиданием $m/2$ и медианой $m\ln(2)$.

На рис. 1, *в* показаны гистограмма распределения 200 реплик, каждая из которых состоит из 30 пар независимых случайных значений коэффициентов миграции (n = 30 и m = 0.05), а также теоретическая плотность вероятности. Обе кривые масштабированы на величину интенсивности распределения $\lambda = 2m^{-1}$. Эти значения параметра соответствуют ситуации, при которой примерно половина всех коэффициентов миграции не превосходит величины $m\ln(2) \approx 0.035$, а среднее их значение равно $\overline{m} = m/2 = 0.025$.

Рассмотрим динамические режимы в системе (9)–(10) в случае случайной миграции для значений параметров, полученных из экспериментальных данных.

Верификация модели

Для верификации модели и поиска условий первичной генетической дивергенции можно поступить двумя способами. Во-первых, выполнить множество симуляций и убедиться, что именно при пониженной приспособленности гетерозигот система (9) порождает режимы, соответствующие генетической дивергенции. Во-вторых, сопоставить результаты симуляций с результатами наблюдений, что довольно проблематично, поскольку при всей изученности и обилии данных большинство природных популяций с выраженной дивергенцией признаков по пространству изначально сильно неоднородны.

Идеальным выходом может быть использование данных специально организованного эксперимента над животными. В упомянутом эксперименте, выполненном под руководством Ю.П. Алтухова, исследовались эволюционные процессы в системе из 30 связанных узкими трубками ящиков, заселенных мухами *D. melanogaster* (Алтухов и др., 1979; Алтухов, Бернашевская, 1981). Случайность миграции обеспечивалась равномерными условиями (освещенность и пища), а также случайным вращением кольцевой системы связанных ящиков. В ходе всего эксперимента анализировались пространственное распределение и динамика численностей, а также частоты аллелей аутосомных локусов эстераза-6 (*Est-6*) и α-глицерофосфатдегидрогеназа (α-*Gdph*). К 60-му поколению сформировалась четкая и устойчивая дифференциация аллельного состава по локусу α-*Gdph* между группами смежных ящиков.

Из описания эксперимента сразу известны некоторые параметры – коэффициент миграции ($m \approx 0.03$), число ящиков (n = 30). В начале эксперимента в ящики поместили небольшое число (150 пар особей, в каждом ящике оказалось от 1 до 37) гетерозиготных особей ($q_2^{(i)}(0) = 1$). Параллельно закладывалась большая локальная панмиктичная популяция, которая сопоставима по численности и начальным частотам с распределенной системой связанных ящиков. Из данных о динамике частоты аллеля A локуса α-*Gdph* в панмиктичной популяции легко оценить параметры отбора s_k (см. таблицу). За основу оценок s_k были взяты оценки из работы (Жданова, Фрисман, 2023), вычисленные для одномерного уравнения частоты аллеля A локуса α-*Gdph*. Характер роста экспериментальной частоты аллеля A четко вписывается в типичное решение модели (1) именно при дизруптивном ($s_2 < s_1$ и $s_2 < s_3$), а не движущем отборе ($s_1 > s_2 > s_3$ или $s_3 > s_2 > s_1$).

Зная начальные численности ($N^{(i)}(0) = 1\ldots37$, $\sum N^{(i)}(0) = 300$), характер роста, а также предельную численность ($\overline{N}^{(i)} \approx 135$) в каждом ящике и в панмиктичной популяции, несложно подобрать параметры роста численности (значения a, h, N_0 и K в таблице).

Приведенное в таблице среднее значение коэффициента миграции $\overline{m} = 0.025$ и медианное значение $m\ln(2) \approx 0.035$ указывают на то, что число мигрантов в большинстве случаев не превосходит 4–5 особей, как и говорилось в описании эксперимента.

Самая большая сложность верификации модели (9) заключается в подборе начальных распределений частот аллелей и численностей, которые дадут финальные распределения, похожие на приведенные в гл. 4 книги Ю.П. Алтухова (2003). Для подбора начальных условий будем генерировать множество начальных частот и численностей, используя следующую особенность эксперимента. В некоторые ящики случайно попадали особи одного пола, которые не оставляли потомства. Для описания этого сформируем вектор случайных численностей следующим образом: $N^{(i)}(0) \sim U[0, 37]$ такой, что $\sum N^{(i)}(0) \approx 300$, а некоторые ящики в начале эксперимента оказываются пустыми ($N^{(i)}(0) = 0$). В результате из-за того, что $0 \le N^{(i)}(0) < N_0$ (численность ниже эффективного числа производителей), в последующих поколениях ящики все равно остаются свободными и повторно заселяются мигрантами из соседних камер, чей генетический состав уже мог существенно отличаться от изначального из-за случайного дрейфа генов и отбора. Однако и этого числа мигрантов может оказаться недостаточно для эффективного поддержания численности, и ящик в течение нескольких поколений остается пустым.

Поскольку начальные численности во всех ящиках ниже эффективной величины N_e, при которой действие отбора превосходит дрейф генов, невозможно игнорировать влияние случайных факторов. Авторы описанного эксперимента полагали $N_e \approx 50$. Это указывает на то, что примерно ко 2-му или 3-му поколению действие детерминированных процессов отбора начинает превалировать над случайными процессами изменения частот аллелей. Прямое описание дрейфа в модели (9) довольно проблематично без ее существенной модификации или перехода к имитационной модели. Поэтому «симулируем» результат действия дрейфа подбором наиболее правдоподобного начального распределения частот, которое типичным образом формируется в модели (1). При дизруптивном отборе (значения s_k из таблицы) система (1) прогнозирует, что частоты генотипов потомков ко 2–3-му поколению

Значения параметров модели (9)

n	$\overline{m}$	s_1	s_2	s_3	a	h	N_0	K
30	0.025	0.244	0.069	0.227	3.6	5	5	90

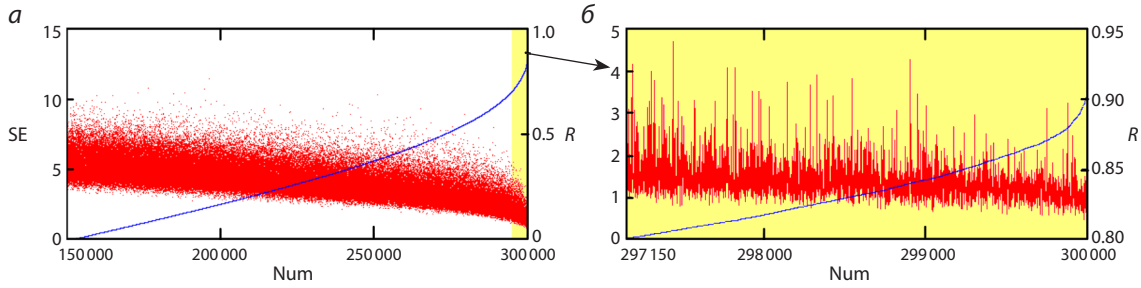


Рис. 2. Квадратичные ошибки SE и коэффициенты корреляции *R* для 300 тыс. начальных условий, отсортированные по возрастанию *R*.

Num – «номер» начальных условий.

от полностью гетерезиготных предков (при $q_2(0) = 1$) составят порядка $q_1 \approx 0.27$, $q_2 \approx 0.46$ и $q_3 \approx 0.27$. В результате можно считать, что действие дрейфа будет случайным образом отклонять частоты от этих значений в первых нескольких поколениях, пока численность будет оставаться ниже эффективной N_e. Сразу можно предположить, что сформированная в ящичной системе популяций дивергенция связана с неоднородностью как начальных численностей, так и частот из-за начального дрейфа генов до достижения эффективной численности в каждой камере.

Для подбора начальных частот сгенерируем два независимых вектора случайных чисел: $q_1^{(i)}(0) \sim U[0,1]$ и $q_2^{(i)}(0) \sim U[0,1]$ $(q_3^{(i)}(0) = 1 - (q_1^{(i)}(0) + q_2^{(i)}(0)))$, и оценим, насколько «истинные» начальные частоты могут отклоняться от теоретических значений 0.27, 0.46 и 0.27 из-за дрейфа, чтобы через 50–60 поколений модель (9) хотя бы отдаленно описывала реальное распределение частот аллелей *A* локуса α-*Gdph*. Рассмотрев 300 тыс. случайных начальных частот и численностей (реплик), мы выяснили, что наиболее правдоподобно описывающими реальное распределение оказались только примерно 100 вариантов со следующим распределением начальных частот:

$$\begin{aligned} q_1^{(i)}(0) &= 0.25 \pm 0.1, \\ q_2^{(i)}(0) &= 0.41 \pm 0.1, \\ q_3^{(i)}(0) &= 0.33 \pm 0.16. \end{aligned} \qquad (11)$$

То есть имеем немного более низкую частоту гетерозигот со смещением в сторону гомозигот с генотипом *aa*, чем предсказывает модель (1). Заметим, что в экспериментальной распределенной популяции также отмечен некоторый перекос средней частоты аллеля *A* в сторону аллеля *a* в 5-м поколении, несмотря на более низкую приспособленность s_3. Следовательно, имеет смысл подбирать начальные частоты примерно из этих интервалов. Из новых 300 тыс. начальных условий вида (11) уже порядка 3 тыс. неплохо описывали реальное распределение частот (рис. 2). В качестве меры качества аппроксимации использовался коэффициент корреляции *R* между реальным $Q^{(i)}(t)$ и модельным $q^{(i)}(t) = q_1^{(i)}(t) + 0.5q_2^{(i)}(t)$ распределением частот аллелей *A* локуса α-*Gdph* в поколении с номером *t*, а также квадратичная ошибка SE:

$$\mathrm{SE}(t) = \sum_{i=1}^{n} \left(Q^{(i)}(t) - (q_1^{(i)}(t) + 0.5q_2^{(i)}(t))\right)^2.$$

Результаты моделирования

Рассмотрим результаты верификации уравнений (9), а также выполним модельный анализ механизмов, которые приводят к устойчивой генетической дивергенции.

На рис. 3, *а* представлены две диаграммы, которые характеризуют тип и характер пространственно-временной динамики системы (9) для приведенных в таблице значений параметров при наиболее удачных начальных условиях (см. рис. 2, *б*).

Цвет пикселя на первой диаграмме кодирует преобладающий генотип на участке с номером *i* в момент времени *t*, а на второй – количество особей. Видно, что на начальных этапах все субпопуляции полиморфны и содержат все три генотипа (зеленый цвет, см. рис. 3, *а*). Со временем из-за отбора и расселения особей в распределенной системе устанавливается равновесный режим, соответствующий устойчивой генетической дивергенции, который сохраняется длительное время (в том числе для $t \gg 200$). В одной части ящиков живут только особи с генотипом *AA* (красный), в другой – *aa* (синий), а между ними расположены полиморфные субпопуляции с высокой частотой гетерозигот (зеленый цвет). Выделенная на диаграмме субпопуляция с номером $i = 16$, так же как ее соседи, сохраняет полиморфизм для $t \gg 200$. На второй диаграмме показано изменение численностей: розовый цвет соответствует максимальным значениям (~135), черный – минимальным. Отмечено также несколько изначально пустых ящиков, их расположение не коррелирует с финальным распределением генотипов.

Как видно из рис. 3, *б*, модель (9) неплохо описывает реальное распределение частот. Однако во всех полученных симуляциях (репликах коэффициентов миграции $m_{i,j}$) выход на распределение, аналогичное тому, что было в опытах с дрозофилами, происходит немного раньше (примерно на 50-м поколении вместо 60-го). С одной стороны, это может быть связано с неточно подобранными параметрами роста – уравнение (9) описывает немного более быстрый рост численности и темпы эволюции, чем в реальности. С другой стороны, процессы дрейфа генов, которые симулировались случайными начальными частотами, в реальности могли преобладать над отбором дольше, чем мы полагаем (2–3 поколения, пока численности не достигли эффективной $N_e \approx 50$). Но вероятнее, существует другое объяснение. При проведении эксперимента над дрозофилами искусственно поддерживался такой

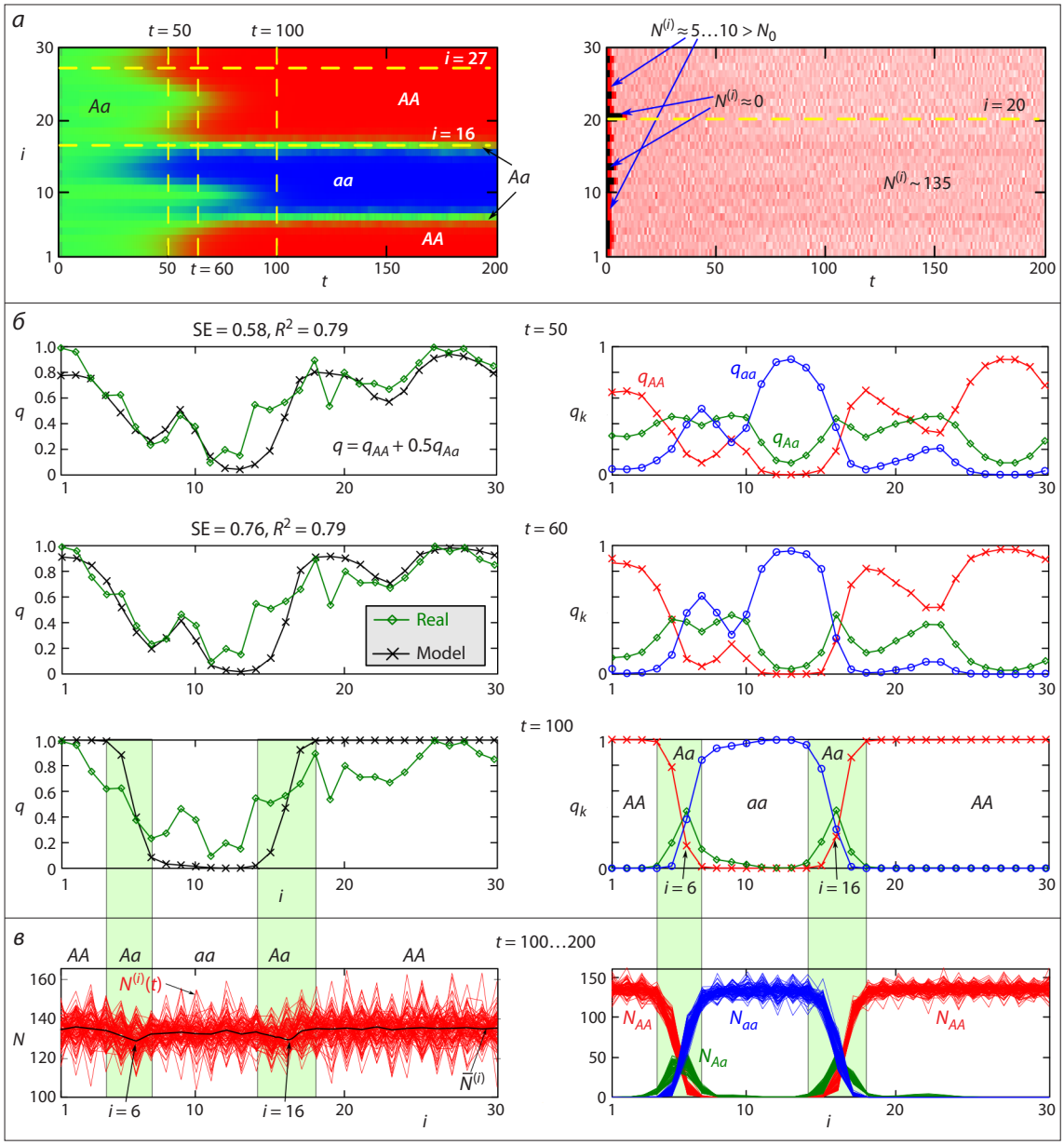


Рис. 3. Пространственно-временная динамика частот генотипов и численностей (*а*) в системе миграционно связанных популяций, описываемых системой (9); *б* – модельное и реальное распределение частот *q* аллеля *A* локуса α-*Gdph* и частоты зигот q_k в 50-, 60- и 100-м поколении; *в* – распределение общей численности по ареалу (слева), а также составляющие в виде численностей особей с генотипами *AA*, *Aa* и *aa*.

половозрастной состав всех субпопуляций, при котором поколения не перекрывались, т. е. из ящиков изымались все взрослые особи, после того как самки откладывали яйца. Однако соотношение полов между ящиками было достаточно неоднородным в ходе всего эксперимента – в некоторых ящиках был значимый недостаток самок, в других – самцов. В результате до момента изъятия не все самки успевали оставить потомство либо некоторые самцы оплодотворяли несколько самок, что нарушало панмиксию и могло исказить данные, поскольку каждое такое тотальное изъятие отбрасывало эволюцию немного назад. Эти процессы не заложены в довольно простую модель (9), и поэтому она дает чуть более быстрый темп эволюции.

На рис. 3, *в* друг на друга наложено 100 последних распределений общей численности каждого генотипа для t = 100…200. Видно, что из-за флуктуаций числа «мигрантов» численность в разных ящиках испытывает нерегулярные несинхронные колебания. Несложно заметить, что полиморфные субпопуляции (i = 6 и 16) из-за значимой частоты гетерозигот имеют более низкую среднюю численность ($\overline{N}^{(i)}$), чем окружающие их мономорфные субпопуляции.

Субпопуляции эволюционируют с разными темпами (см. первую диаграмму на рис. 3, *а*), что, очевидно, определяется тем, насколько их численность изначально близка к эффективной, а аллельный состав – к финальному (q = 1 или 0). На диаграмме выделен ящик с номером i = 27,

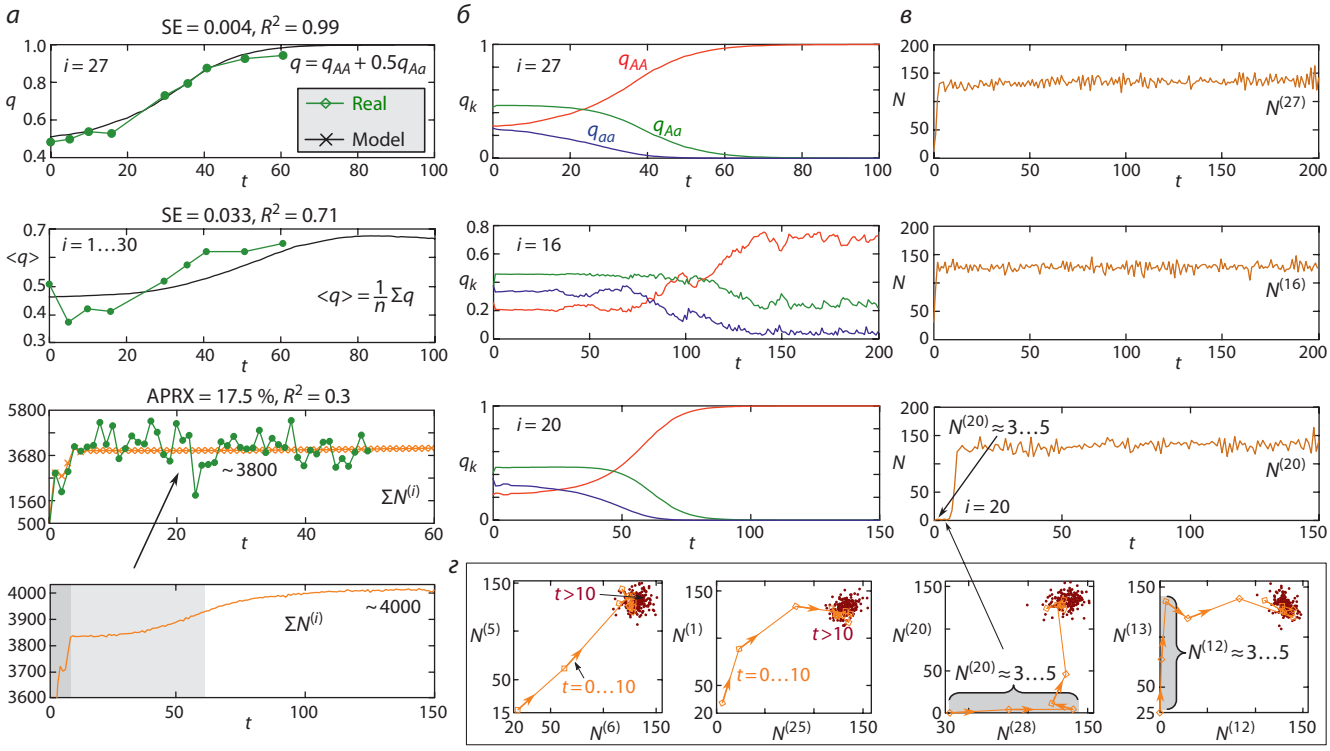


Рис. 4. *а* – модельная и реальная динамика частот *q* аллеля *A* локуса α-*Gdph* и общей численности в системе ящичных популяций *D. melanogaster*. SE – квадратичная ошибка; R^2 – коэффициент детерминации; APRX – ошибка аппроксимации. Модельная динамика частот генотипов (*б*) и численностей (*в*) субпопуляций, выделенных на рис. 3, *а*. *г* – фазовые портреты, демонстрирующие характер совместной динамики двух субпопуляций: светло-коричневый цвет – этап стремительного заполнения ящиков, коричневый – переход к предельной численности.

частота аллеля *A* в котором одной из первых достигает значения $q = 1$. Важно, что именно эта субпопуляция эволюционирует схожим образом с большой панмиктичной популяцией (рис. 4, *а*, первый график). Другие субпопуляции, как правило, развиваются медленнее.

На рис. 4 показано, как динамика частот и численностей по модели (9) соотносится с реальными экспериментальными данными. На втором графике видно, что модельная и экспериментальная средняя частота $<q>$ аллеля *A* во всех 30 ящиках изменяется схожим образом, фиксируясь на значении $q \approx 0.65$. Несоответствие модельной и реальной средней частоты в точке $t = 5$ можно объяснить тем, что модель (9) явно не учитывает дрейф генов, происходивший в экспериментальной популяции, а его действие лишь симулируется случайными возмущениями частот полиморфной популяции.

Третий график на рис. 4, *а* показывает реальную и модельную суммарную численность системы 30 субпопуляций. Из четвертого графика видно, что переход к предельной численности проходит через три этапа: взрывной рост (за 2–3 поколения) от небольшого числа основателей; выход на квазистационарный уровень с суммарной численность $\sum N^{(i)} \sim 3800$ особей, при котором уже отмечается отчетливая дифференциация генотипов по группам ящиков, но их состав все еще достаточно полиморфен (см. рис. 3, *б* при $t = 50$); переход к финальному распределению (см. рис. 3, *б* при $t = 100$) и предельной суммарной численности ~4000 особей. То есть модель (9) описывает лишь общие тенденции роста численности. Поскольку такое поведение численности в модели (9) – это, в принципе, единственно возможный тип динамики при $r = a\overline{W} < e^2 \approx 7.38$. Кроме того, уравнение (8), описывающее динамику численности локальной популяции, не учитывает половозрастной состав и множество других факторов, которые, без сомнения, вызвали нерегулярные колебания численности экспериментальной популяции. Но более важно, что модель (9) описывает только репродуктивное ядро популяционной системы (самки и равное им число самцов) и не учитывает тот факт, что часть самцов могла остаться «холостяками» и составлять большую часть мигрантов. В результате модельная численность будет ниже реальной.

Вместе с тем модельная динамика общей численности $\sum N^{(i)}$ складывается из несинхронных флуктуаций каждой субпопуляции вокруг стационарного значения ~135 особей в каждом ящике (см. рис. 4, *в*, *г*). Суммирование по этим значениям сглаживает все различия численностей субпопуляций. Несмотря на неоднородности в начальных распределениях особей, в первых пяти поколениях рост численности, подгоняемый ростом приспособленности, почти во всех ящиках происходит синхронно (первый и второй график на рис. 4, *г*). За исключением изначально пустых или ящиков с недостаточным числом производителей (третий и четвертый графики на рис. 4, *г*), для которых ненулевая численность (~3–5 особей) поддерживается только за счет мигрантов. Далее численности в разных ящиках медленно достигают предельных значений, флуктуируя около них (темные точки на рис. 4, *г*).

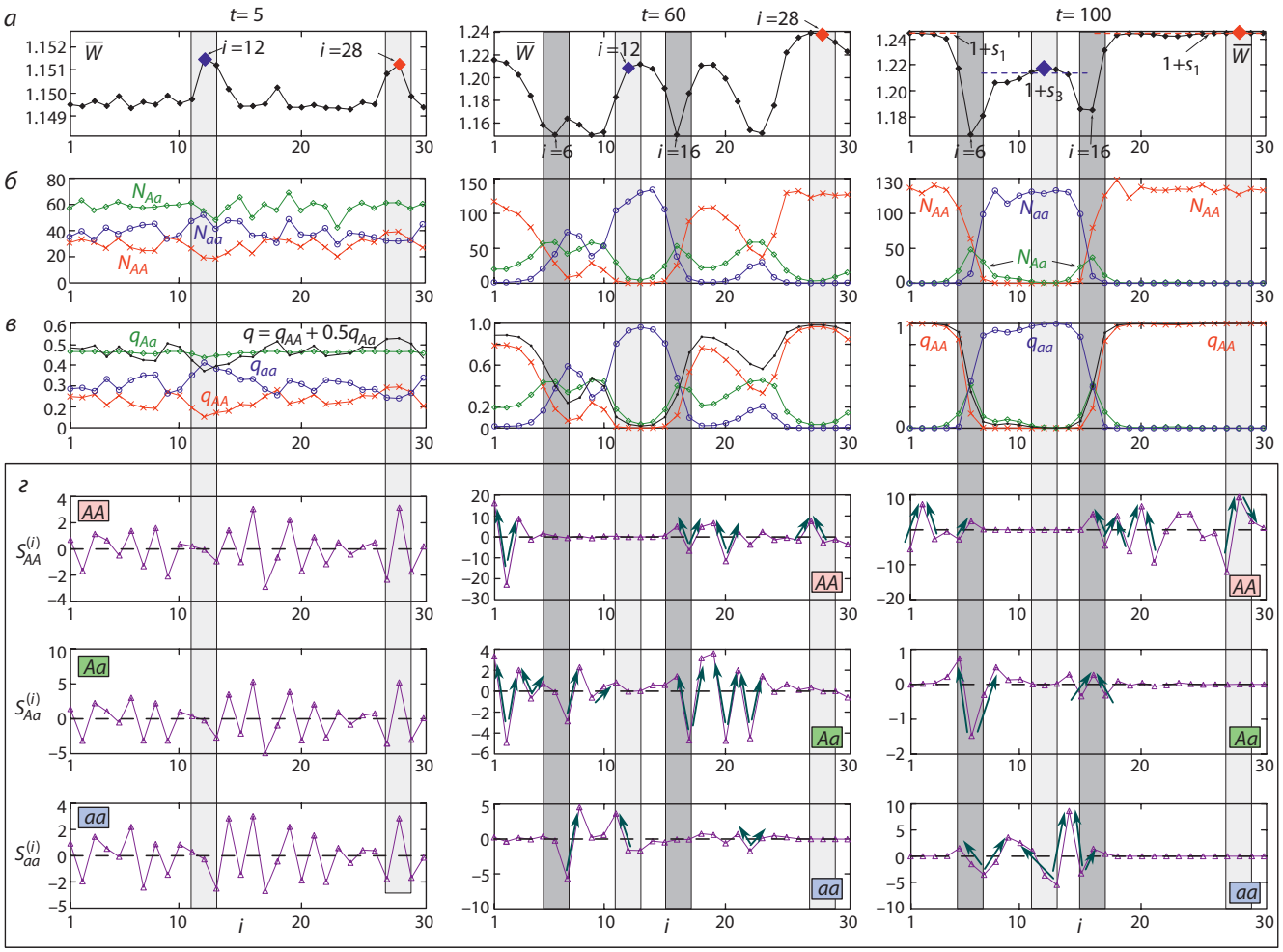


Рис. 5. *Распределение значений средней приспособленности каждой субпопуляции до миграции (а), численности (б) и частоты (в) генотипов AA, Aa и aa; г – распределение значений миграционного баланса (сальдо) каждого участка.*

На графиках выделены участки, вокруг которых формируется группа с генотипами *AA* (i = 28) и *aa* (i = 12), а также участки с активной гибридизацией особей (i = 6 и 16). Стрелки на графиках сальдо показывают направления потоков особей с соответствующим генотипом.

Рассмотрим возможные механизмы формирования и поддержания пространственной дифференциации аллельного состава (дивергенции), которая могла происходить в описанной экспериментальной популяционной системе.

Анализ миграционных потоков

На одну из причин наблюдаемой дифференциации субпопуляций по признакам указывают небольшие провалы численностей ящиков (i = 6 и 16), в которых сохраняется полиморфизм (ящики, обозначенные как *Aa* на рис. 3). Они становятся заметны лишь в финальном распределении, поскольку их окружают группы с противоположными генотипами, имеющие большую численность. Однако присутствие таких субпопуляций указывает только на возможные механизмы поддержания дивергенции, а не причины их возникновения. Такие ящики можно условно назвать зонами гибридизации. Очевидно, что их аллельный состав поддерживается только за счет контакта, миграции и потока генов с участков, где проживают особи с противоположными признаками.

Для изучения механизмов формирования и сохранения дивергенции рассмотрим изменение средней приспособленности в каждом ящике $\overline{W}^{(i)}$ (рис. 5, *а*), численностей генотипов $N_k^{(i)}$ (см. рис. 5, *б*) и частот $q_k^{(i)}$ (см. рис. 5, *в*) со временем, оценивая вклад миграции в процесс естественного отбора и перехода к финальному распределению частот. Миграционный баланс (сальдо) особей генотипа k ($k = AA$, Aa или aa) на участке с номером i будем находить по формуле:

$$S_k^{(i)} = m_{i,i+1} q_k^{(i+1)} N^{(i+1)^*} + m_{i,i-1} q_k^{(i-1)} N^{(i-1)^*} - \\ - (m_{i-1,i} + m_{i+1,i})\, q_k^{(i)} N^{(i)^*}, \quad (12)$$

где $q_k^{(i)} N^{(i)^*}$ – численность особей с генотипом k после отбора, но до миграции. Эта величина равна разнице между числом прибывших (первые два слагаемых) на участок с номером i и числом выбывших оттуда (третье слагаемое) особей. Величина S показывает, увеличилось ($S > 0$) или уменьшилось ($S < 0$) население субпопуляции с номером i за счет миграции (см. рис. 5, *г*). Сравнив между собой три величины $S_k^{(i-1)}$, $S_k^{(i)}$ и $S_k^{(i+1)}$, легко оценить направления миграции (стрелки на рис. 5, *г*).

При подборе начальных условий было установлено, что наблюдаемое в эксперименте распределение частот в модели (9) возникает в случае смещения начальных частот в сторону преобладания гомозигот с генотипом *aa*. Заметим, что генотипы *AA* и *aa* отличаются по приспособленности примерно на 11 %. Значит, чтобы наиболее приспособленный генотип *AA* смог закрепиться, достаточно преодолеть это смещение хотя бы у небольшой части субпопуляций. Однако требуется более редкое стечение обстоятельств, при котором менее приспособленный генотип *aa* не вытесняется полностью и оба признака сохраняются.

На рис. 5, *а* показано, что после стремительного роста к 5-му поколению выделяются два участка, с номерами $i = 12$ и 28, распределение частот в которых дает наибольшие среди остальных значения средней приспособленности $\overline{W}^{(i)}$ и суммарных репродуктивных потенциалов $a\overline{W}^{(i)}$. Хотя это разница невелика (1 % для *aa* и 0.7 % для *AA*), ее оказывается вполне достаточно, чтобы в окрестности данных ящиков началось обособление особей одного генотипа. Для этого, вероятно, потребовалось смещение частот более чем на одном участке. На рис. 5, *б*, *в* показано распределение численностей и частот; видно, что около участка $i = 12$ при $t = 5$ присутствует минимум 6 ящиков с повышенным относительно окружающих числом гомозигот *aa* и $q < 0.5$. Это означает, что поток мигрантов оттуда при любых случайных $m_{i,j}$ преимущественно представлен именно этим генотипом, что способствует его закреплению. Участок $i = 28$ имеет только одного соседа с высокой численностью гомозигот *AA* и $q > 0.5$, но этого вполне достаточно для закрепления наиболее приспособленного генотипа. Примерно до 50-го поколения на участках $i = 12$ и 28 сохраняются наибольшие темпы роста приспособленности, частоты ближе к финальным ($q = 1$ или 0), а численность соответствующего генотипа оказывается заведомо выше, чем у окружающих. В результате мигранты из этих ящиков более однородны, чем из других, и даже случайный характер миграции не меняет общей эволюционной тенденции – гомозиготы вытесняют наименее приспособленных гетерозигот.

На графиках значений миграционного баланса $S_k^{(i)}$ (см. рис. 5, *г*) на начальных этапах ($t = 5$) наблюдалось довольно случайное и сопоставимое для разных генотипов распределение направлений и интенсивностей потока особей между участками. По мере пространственной дифференциации, когда более приспособленные особи вытесняют менее приспособленных, однородные участки с наибольшей численностью ($i = 12$ и 28) начинают вносить больший вклад в миграцию, чем участки с высоким полиморфизмом. К 60-му поколению две мономорфные группы *AA* и *aa* с противоположными признаками достигают наибольшего размера (*AA* – 17 ящиков, *aa* – 8 ящиков) и «соприкасаются» друг с другом. Однако поскольку они уже успели накопить достаточное число особей, а их численности сопоставимы, потоки мигрантов тоже оказываются сопоставимыми, несмотря на 11 % разницу в приспособленности. В результате в зонах гибридизации около участков с номерами $i = 6$ и 16 сходятся два равновеликих потока особей с противоположными генотипами, за счет которых там обеспечивается ненулевое число гетерозигот. Поток из этих ящиков значительно слабее, его едва хватает, чтобы обеспечить небольшой полиморфизм в окрестности данных участков. Но именно они, по всей видимости, «тормозят» потоки гомозиготных особей разных форм, не давая окончательно закрепиться по всему ареалу наиболее приспособленному генотипу *AA*.

Обсуждение

Проведенная верификация модели (9) по данным эксперимента под руководством Ю.П. Алтухова с ящичными популяциями *D. melanogaster*, а также анализ сценариев формирования неоднородного распределения частот и численностей требуют дополнительных пояснений.

Прежде всего, необходимо обсудить причину весьма заметных различий в приспособленностях генотипов с различающимися аллельными наборами фермента α-Gdph, которую выявили полученные оценки коэффициентов отбора s_k. Вполне возможно, что локус α-*Gdph* является маркером дизруптивного отбора, протекающего в системе и действующего не непосредственно на ген α-*Gdph*, а на тесно сцепленные с ним адаптивные гены. Этим можно объяснить некоторые несоответствия между реальным и модельным распределением и динамикой частот, поскольку общий адаптивный эффект и направления отбора даже у сильно сцепленных с α-*Gdph* разных генов не являются их суммой из-за полигенного или комплементарного действия генов, эпистаза или полимерии.

Отметим, что существенная разница приспособленностей не является необходимой для генетической дивергенции в модели (1). Ранее было показано, что пространственная дифференциация возникает и при небольшой разнице в приспособленностях. От того, насколько различаются генотипы, а также от коэффициента миграции зависит, как быстро будет достигнута устойчивая дивергенция и насколько крупными будут группы мономорфных субпопуляций и зоны гибридизации (Кулаков, Фрисман, 2025).

Несмотря на отмеченные ограничения, предложенная модель позволяет проанализировать процессы, которые привели к наблюдаемой в эксперименте первичной генетической дивергенции. Так, установлено, что совместное действие дрейфа генов, плотностно-зависимого лимитирования и потока генов до того, как была достигнута эффективная численность N_e и минимальное число производителей N_0, привело к тому, что в некоторых ящиках менее приспособленных особей *aa* случайно оказалось больше, чем более приспособленных *AA*. В результате субпопуляции, в которых наблюдалось даже незначительное отклонение частот от теоретически ожидаемых (характерных для локальной панмиктичной популяции), раньше других достигают наибольшей средней приспособленности и скорости роста численности. Поскольку «эмигранты» повторяют аллельный состав родного участка, то именно вокруг таких стремительно растущих субпопуляций формируются группы ящиков с генотипами *AA* или *aa*. Постепенно эти генотипы вытесняют наименее приспособленных гетерозиготных особей *Aa* и захватывают наибольшее число участков. Взаимодействие двух потоков особей, с генотипами *AA* и *aa*, поддерживает ненулевое число гетерозиготных особей в некоторых ящиках – зонах гибридизации. С одной стороны, их наличие сохраняет генетическое разнообразие во всей системе популяций,

с другой – эти зоны препятствуют захвату всего ареала наиболее приспособленными особями.

Описанный эволюционный сценарий может быть достаточно универсальным по следующим соображениям. Дивергенции природных популяций предшествует появление на некоторых участках мутантных особей с новым признаком. Для закрепления нового признака необходим мощный механизм изоляции от исходной формы, особенно если он не дает существенных преимуществ по сравнению с исходным. Им может быть дизруптивный отбор, выражающийся не только в пониженной приспособленности гетерозигот (гибридов), но и в положительной ассортативности скрещивания, которая снижает репродуктивный успех малочисленных гибридов. Например, в упомянутых во введении примерах с серой и черной вороной изолирующий механизм, по всей видимости, связан с брачными предпочтениями. Для ворон он достоверно связан с врожденным восприятием окраски перьев потенциальных партнеров, при которой существенно снижается вероятность спаривания непохожих форм, но допускается скрещивание уже гибридных особей или гибридных особей с «чистыми» формами (Poelstra et al., 2014; Kryukov, 2019).

Отметим, что расселение особей или колонизация новых участков, в отличие от сезонных миграций, не происходит быстро, а длится несколько поколений. В результате удаленный участок смогут заселить лишь потомки тех, кто покинул одну часть ареала и устремился в другую. В ходе такой миграции особи по пути будут неизбежно скрещиваться с местными представителями. Предложенная в работе модель показывает, что такое перемещение неизбежно остановится в случае, если принимающий участок заселен особями с иным, чем у мигрантов, признаком, поскольку там может действовать отбор против гибридов. Для ворон ассортативность скрещивания будет ограничивать возможность скрещивания разных форм на новых участках и значительно снизит шансы дальнейшей экспансии. Для дрозофил в кольцевой системе популяций пониженная приспособленность гетерозигот снижает плодовитость гибридов мух и не дает возможности их потомкам расселяться дальше. В результате для видов, у которых расселение – растянутый на поколения процесс, зоны гибридизации существенно препятствуют переходу особей с одним признаком на участок, занятый особями с другим признаком, без необходимости осесть там и с большой вероятностью оставить гибридное потомство. В случае, если возможно более быстрое расселение, ситуация может кардинально измениться.

Заключение

Предложенная в работе динамическая модель позволяет детальнее изучить механизмы формирования первичной генетической дивергенции, которые объясняются различиями приспособленностей разных генотипов, особенностями расселения, миграции и формирования устойчивых гибридных зон. Показана возможность репродуктивной изоляции между разными формами диплоидных организмов, основанная не только на географической изоляции, удаленности участков или экологической специализации, а также на наследственных механизмах, дрейфе и потоке генов, а также отборе против гетерозигот. Такой тип отбора приводит к устойчивой пространственной дифференциации генотипов, поддерживаемой зонами гибридизации, которые исполняют роль труднопреодолимых для разных признаков барьеров.

Таким образом, в работе показана важная роль дизруптивного отбора, действие которого можно зарегистрировать по некоторым генам-маркерам, но не всегда по внешним признакам. В результате он может быть более распространен в природе, чем считалось раньше.

Список литературы / References

Алтухов Ю.П. Генетические процессы в популяциях. М.: Академкнига, 2003
[Altukhov Yu.P. Genetic Processes in Populations. Moscow: Akademkniga Publ., 2003 (in Russian)]

Алтухов Ю.П., Бернашевская А.Г. Экспериментальное моделирование генетических процессов в популяционной системе *Drosophila melanogaster*, соответствующей кольцевой ступенчатой модели. Сообщ. 2. Стабильность аллельного состава и периодическая зависимость изменчивости частот аллелей с расстоянием. *Генетика*. 1981;17(6):1052-1059
[Altukhov Yu.P., Bernashevskaya A.G. Experimental modeling of genetic processes in a population system of *Drosophila melanogaster* corresponding to a circular stepping-stone model: 2. Stability of allelic composition and periodic relationship of allele frequency with distance. *Soviet Genetics*. 1981;17(6):1052-1059 (in Russian)]

Алтухов Ю.П., Бернашевская А.Г., Милишников А.Н. Экспериментальное моделирование генетических процессов в популяционной системе *Drosophila melanogaster*, соответствующей кольцевой ступенчатой модели. *Генетика*. 1979;15(4):646-655
[Altukhov Yu.P., Bernashevskaya A.G., Milishnikov A.N. Experimental modeling of genetic processes in the population system of *Drosophila melanogaster* corresponding to the ring step model. *Soviet Genetics*. 1979;15(4):646-655 (in Russian)]

Базыкин А.Д. Пониженная приспособленность гетерозигот в системе смежных популяций. *Генетика*. 1972;8(11):155-161
[Bazykin A.D. Reduced fitness of heterozygotes in a system of adjacent populations. *Soviet Genetics*. 1972;8(11):155-161 (in Russian)]

Блинов В.Н., Железнова Т.К. Чёрная *Corvus corone* и серая *C. cornix* вороны: спорные вопросы о статусе (расы, полувиды или виды?), происхождении (алло- или симпатрическое?) и феномене стабильных гибридных зон. *Русский орнитологический журнал*. 2020;29(1958):3596-3601
[Blinov V.N., Zheleznova T.K. Black *Corvus corone* and grey *C. cornix* crows: controversial issues about status (races, semispecies or species?), origin (allo- or sympatric?) and the phenomenon of stable hybrid zones. *Russkiy Ornitologicheskiy Zhurnal = Russian Ornithological Journal*. 2020;29(1958):3596-3601 (in Russian)]

Жданова О.Л., Фрисман Е.Я. О генетической дивергенции миграционно-связанных популяций: современное моделирование по результатам экспериментов Ю.П. Алтухова с соавторами. *Генетика*. 2023;59(6):708-717. doi 10.31857/S0016675823060139
[Zhdanova O.L., Frisman E.Y. On the genetic divergence of migration-coupled populations: modern modeling based on the experimental results of Yu.P. Altukhov et al. *Russ J Genet*. 2023;59:614-622. doi 10.1134/S1022795423060133]

Капитонова Л.В., Формозов Н.А., Федоров В.В., Керимов А.Б., Селиванова Д.С. Особенности поведения и экологии большой *Parus major* Linneus, 1758 и восточной *P. minor* Temmink et Schlegel, 1848 синиц, как возможные факторы поддержания устойчивости видоспецифических фенотипов в зоне сообитания и локальной гибридизации в Приамурье. *Дальневосточный орнитологический журнал*. 2012;3:37-46
[Kapitonova L.V., Formozov N.A., Fedorov V.V., Kerimov A.B., Selivanova D.S. Peculiarities of behavior and ecology of the Great

tit *Parus major* Linneus, 1758 and Japanese tit *P. minor* Temmink et Schlegel, 1848 as possible factors of maintaining the stability of speciesspecific phenotypes in the area of sympatry and local hybridization in the Amur Region. *Dal'nevostochnyy Ornitologicheskiy Zhurnal = Far Eastern Journal of Ornithology*. 2012;3:37-46 (in Russian)]

Кулаков М.П., Фрисман Е.Я. Первичная генетическая дивергенция в системе миграционно связанных лимитированных популяций на кольцевом ареале. *Математическая биология и биоинформатика*. 2025;20(1):1-30. doi 10.17537/2025.20.1
[Kulakov M., Frisman E.Ya. Primary genetic divergence in a system of limited population coupled by migration in a ring habitat. *Mathematical Biology and Bioinformatics*. 2025;20(1):1-30. doi 10.17537/2025.20.1 (in Russian)]

Фрисман Е.Я. Первичная генетическая дивергенция (Теоретический анализ и моделирование). Владивосток, 1986
[Frisman E.Y. Primary Genetic Divergence (Theoretical analysis and modeling). Vladivostok, 1986 (in Russian)]

Aguillon S.M., Rohwer V.G. Revisiting a classic hybrid zone: movement of the northern flicker hybrid zone in contemporary times. *Evolution*. 2022;76(5):1082-1090. doi 10.1111/evo.14474

Allee W.C. The Social Life of Animals. Beacon Press, 1958

Blair W.F. Mating call and stage of speciation in the *Microhyla olivacea–M. carolinensis* complex. *Evolution*. 1955a;9(4):469-480. doi 10.1111/j.1558-5646.1955.tb01556

Blair W.F. Size difference as a possible isolation mechanism in Microhyla. *Am Nat*. 1955b;89(848):297-301. doi 10.1086/281894

Dey S., Joshi A. Stability via asynchrony in *Drosophila* metapopulations with low migration rates. *Science*. 2006;312(5772):434-436. doi 10.1126/science.1125317

Filchak K., Roethele J., Feder J. Natural selection and sympatric divergence in the apple maggot *Rhagoletis pomonella*. *Nature*. 2000; 407(6805):739-742. doi 10.1038/35037578

Haring E., Däubl B., Pinsker W., Kryukov A., Gamauf A. Genetic divergences and intraspecific variation in corvids of the genus *Corvus* (Aves: Passeriformes: Corvidae) – a first survey based on museum specimens. *J Zool Syst Evol Res*. 2012;50(3):230-246. doi 10.1111/j.1439-0469.2012.00664.x

Keymer J.E., Galajda P., Muldoon C., Park S., Austin R.H. Bacterial metapopulations in nanofabricated landscapes. *Proc Natl Acad Sci USA*. 2006;103(46):17290-17295. doi 10.1073/pnas.0607971103

Kryukov A.P. Phylogeography and hybridization of corvid birds in the Palearctic Region. *Vavilov J Genet Breed*. 2019;23(2):232-238. doi 10.18699/VJ19.487

Láruson Á.J., Reed F.A. Stability of underdominant genetic polymorphisms in population networks. *J Theor Biol*. 2016;390:156-163. doi 10.1016/j.jtbi.2015.11.023

Littlejohn M.J. Premating isolation in the *Hyla ewingi* complex (Anura: Hylidae). *Evolution*. 1965;19(2):234-243. doi 10.2307/2406376

Matsumoto M., Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans Model Comput Simul*. 1998;8(1):3-30. doi 10.1145/272991.272995

Murphy M.A., Dezzani R., Pilliod D.S., Storfer A. Landscape genetics of high mountain frog metapopulations. *Mol Ecol*. 2010;19(17): 3634-3649. doi 10.1111/j.1365-294X.2010.04723.x

Orsini L., Corander J., Alasentie A., Hanski I. Genetic spatial structure in a butterfly metapopulation correlates better with past than present demographic structure. *Mol Ecol*. 2008;17(11):2629-2642. doi 10.1111/j.1365-294X.2008.03782.x

Poelstra J.W., Vijay N., Bossu C.M., Lantz H., Ryll B., Müller I., Baglione V., Unneberg P., Wikelski M., Grabherr M.G., Wolf J.B.W. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*. 2014;344(6190):1410-1414. doi 10.1126/science.1253226

Smith M.J., Osborne W., Hunter D. Geographic variation in the advertisement call structure of *Litoria verreauxii* (Anura: Hylidae). *Copeia*. 2003;4:750-758. doi 10.1643/HA02-133.1

Sundqvist L., Keenan K., Zackrisson M., Prodöhl P., Kleinhans D. Directional genetic differentiation and relative migration. *Ecol Evol*. 2016;6(11):3461-3475. doi 10.1002/ece3.2096

Tait C., Kharva H., Schubert M., Kritsch D., Sombke A., Rybak J., Feder J.L., Olsson S.B. A reversal in sensory processing accompanies ongoing ecological divergence and speciation in *Rhagoletis pomonella*. *Proc Biol Sci*. 2021;288(1947):20210192. doi 10.1098/rspb.2021.0192

Yeaman S., Otto S.P. Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*. 2011;65(7):2123-2129. doi 10.1111/j.1558-5646.2011.01277.x

Yee W.L., Goughnour R.B. Mating frequencies and production of hybrids by *Rhagoletis pomonella* and *Rhagoletis zephyria* (Diptera: Tephritidae) in the laboratory. *Can Entomol*. 2011;143(1):82-90. doi 10.4039/n10-047

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 29.07.2025. После доработки 02.09.2025. Принята к публикации 05.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1122-1128

doi 10.18699/vjgb-25-116

Асимметрия нуклеотидных замен в тРНК свидетельствует об общем происхождении современных организмов от термофильного предка

И.И. Титов [1, 2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
titov@bionet.nsc.ru

Аннотация. Природа последнего универсального общего предка (last universal common ancestor, LUCA) всех ныне живущих организмов до сих пор остается актуальной проблемой биологии. Существуют свидетельства в пользу того, что LUCA был как термофилом, так и мезофилом. Усложнение клеточного аппарата в ходе эволюции от ранних форм жизни к современным организмам могло проявиться в долговременных эволюционных изменениях нуклеотидного состава генетических последовательностей. Выявлению подобных тенденций в последовательностях тРНК посвящена эта работа. Представлены результаты эволюционного анализа точечных нуклеотидных замен в тРНК 123 видов трех доменов: Bacteria, Archaea и Eukaryota. Обнаружен универсальный вектор направленного эволюционного изменения последовательностей тРНК, при котором замены гуанина (G) и цитозина (C) на аденин (A) и урацил (U) суммарно происходят чаще обратных. Наиболее ярко асимметрия числа замен наблюдается в следующих переходах: а) между пуринами в преобладании числа замен G→A над числом замен A→G; б) между пиримидинами в преобладании C→U над U→C, а также в) при переходе из пурина в пиримидин и наоборот – в преобладании G→U над U→G. В результате эволюционного процесса тРНК могли терять «сильные» комплементарные пары с тремя водородными связями, формируемые гуанином и цитозином, и фиксировать «слабые» комплементарные пары с двумя водородными связями, образуемые аденином и урацилом. Обнаруженному изменению состава последовательностей были подвержены 16 из 20 семейств тРНК, что соответствует уровню статистической значимости $p = 0.006$ согласно одностороннему биномиальному тесту. Выявленная закономерность свидетельствует о высоком GC-содержании в последовательности общего предка современных тРНК и, следовательно, подтверждает предположение о том, что самая молодая из гипотетических общих предковых клеток, от которой произошли все ныне живущие организмы (последний универсальный общий предок, LUCA), обитала в более горячей среде, нежели ныне живущие организмы.

Ключевые слова: эволюция; термофил; мутации; тРНК; матрица перехода; последний универсальный общий предок

Для цитирования: Титов И.И. Асимметрия нуклеотидных замен в тРНК свидетельствует об общем происхождении современных организмов от термофильного предка. *Вавиловский журнал генетики и селекции*. 2025;29(7):1122-1128. doi 10.18699/vjgb-25-116

Финансирование. Работа поддержана бюджетным проектом № FWNR-2022-0020.

Asymmetry of nucleotide substitutions in tRNAs indicates common descent of modern organisms from a thermophilic ancestor

I.I. Titov [1, 2]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
titov@bionet.nsc.ru

Abstract. The nature of the last universal common ancestor (LUCA) of all living organisms remains a controversial issue in biology. There is evidence of both thermophilic and mesophilic LUCA origin. The increasing complexity of the cellular apparatus during the evolution from early life forms to modern organisms could have manifested itself in long-term evolutionary changes in the nucleotide composition of genetic sequences. This work is devoted to the identification of such trends in tRNA sequences. The results of an evolutionary analysis of single-nucleotide substitutions in tRNAs of 123 species from three domains – Bacteria, Archaea and Eukaryota – are presented. A universal vector of directed evolutionary change in tRNA sequences has been discovered, in which substitutions of guanine (G) to adenine (A)

© Титов И.И., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

and cytosine (C) to uracil (U) occur more frequently than the reverse. The most striking asymmetry in the number of substitutions is observed in the following transitions: a) purine-to-purine, where G→A outnumbers A→G, b) pyrimidine-to-pyrimidine, where C→U outnumbers U→C, and c) purine-to-pyrimidine and vice versa, where G→U outnumbers U→G. As a result, tRNAs could lose "strong" three-hydrogen-bond complementary pairs formed by guanine and cytosine and fix "weak" two-hydrogen-bond complementary pairs formed by adenine and uracil. 16 out of 20 tRNA families are susceptible to the detected change in sequence composition, which corresponds to the significance level $p = 0.006$ according to the one-sided binomial test. The identified pattern indicates a high GC content in the common ancestor of modern tRNAs, supporting the hypothesis that the last universal common ancestor (LUCA) lived in a hotter environment than do most contemporary organisms.

Key words: evolution; thermophile; mutations; tRNA; transition matrix; last universal common ancestor

For citation: Titov I.I. Asymmetry of nucleotide substitutions in tRNAs indicates common descent of modern organisms from a thermophilic ancestor. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1122-1128. doi 10.18699/vjgb-25-116

Введение

Несмотря на длительные исследования, природа последнего универсального общего предка (last universal common ancestor, LUCA) всех ныне живущих организмов – актуальная проблема биологии до сих пор. Согласно недавней работе E.R.R. Moody с коллегами (2024), LUCA возник ~ 4.2 млрд лет назад и обладал базовыми элементами клеточного аппарата современных прокариот (генами и молекулярно-генетическими системами транскрипции и трансляции, включая тРНК). Существуют свидетельства в пользу того, что LUCA был как термофилом (Di Giulio, 2000; Weiss et al., 2016; Moody et al., 2024), так и мезофилом (Galtier et al., 1999; Cantine, Fournier, 2017).

Усложнение клеточного аппарата в ходе эволюции от ранних форм жизни к современным организмам могло проявиться в долговременных эволюционных изменениях нуклеотидного состава генетических последовательностей. Так, I.K. Jordan с коллегами (2005), используя метод неукорененной парсимонии (Rickert et al., 2025), обнаружили закономерности систематического однонаправленного изменения аминокислотного состава белков при их эволюции от предковых форм: увеличение содержания аминокислот Cys, Met, His, Ser и Phe за счет снижения содержания аминокислот Pro, Ala, Glu и Gly. В работе (Galtier et al., 1999) проведено сравнение рибосомальных РНК LUCA и современных видов по GC-содержанию, но результаты позднее были подвергнуты критике (Di Giulio, 2000). Интерес представляет исследование Y. Men с коллегами (2022), в котором реконструированы фрагменты рибосомальных РНК LUCA (16S, 5S и 23S rRNA), эволюционно консервативные в современных последовательностях и соответствующие сайтам взаимодействия рРНК с белками рибосомы. Однако в этой работе нуклеотидные последовательности рРНК рассматривались в бинарном пурин-пиримидиновом коде и, следовательно, не содержали оценок GC-состава РНК. Таким образом, эволюционное изменение нуклеотидного состава РНК от LUCA до современных видов не является окончательно установленным.

В связи с этим представлялось интересным исследовать долговременные тренды изменения нуклеотидного состава такого класса нуклеотидных последовательностей, а именно, молекул тРНК, являющихся обязательным элементом систем трансляции тРНК во всех организмах.

В нашей работе изучены особенности молекулярной эволюции 20 изоакцепторных семейств тРНК, каждое из которых в процессе трансляции обеспечивает перенос определенной аминокислоты. Эти семейства тРНК были проанализированы для 123 организмов трех доменов, Bacteria, Archaea и Eukaryota.

Филогенетический анализ проводили с применением метода неукорененной парсимонии (Jordan et al., 2005). Были обнаружены одиночные нуклеотидные замены, фиксировавшиеся в тРНК в ходе их эволюции от предковых последовательностей к современным; показано, что замены гуанина (G) или цитозина (C) на аденин (A) или урацил (U) фиксируются чаще замен A или U на G или C. Это обосновывает картину преимущественно однонаправленного эволюционного изменения последовательностей тРНК, при котором они теряли «сильные» комплементарные пары с тремя водородными связями, формируемые гуанином и цитозином, и фиксировали «слабые» комплементарные пары с двумя водородными связями, образуемыми аденином и урацилом. Эта особенность была характерна для 16 из 20 семейств тРНК, с уровнем статистической значимости $p < 0.006$ согласно одностороннему биномиальному тесту.

Полученные результаты свидетельствуют о высоком GC-содержании в нуклеотидных последовательностях тРНК общего предка современных Bacteria, Archaea и Eukaryota и, следовательно, в пользу предположения о том, что последний универсальный общий предок (LUCA) обитал в более горячей среде, нежели ныне живущие организмы, т. е. он был термофилом или теплолюбивым мезофилом (умеренным термофилом). Этот вывод обосновывается тем, что содержание нуклеотидов G и C в нуклеотидных последовательностях связано с оптимальной температурой обитания организмов, в связи с чем генетические макромолекулы (ДНК, РНК) можно рассматривать как своеобразные молекулярные термометры, а их GC-содержание – как индикатор температуры среды обитания.

Материалы и методы

Нуклеотидные последовательности тРНК трех доменов, Bacteria, Archaea и Eukaryota, взяты из курируемой базы данных, представленной в работе (Sprinzl et al., 1998) (Приложение S1)[1]. В этой базе данных содержалось выравнивание последовательностей тРНК, «наиболее совместимое с филогенией и известными трехмерными струк-

[1] Приложения S1 и S2 см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx41.zip

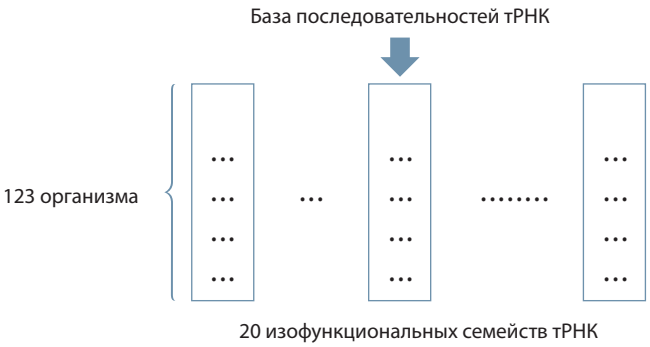


Рис. 1. Схема создания выборки из базы последовательностей тРНК.

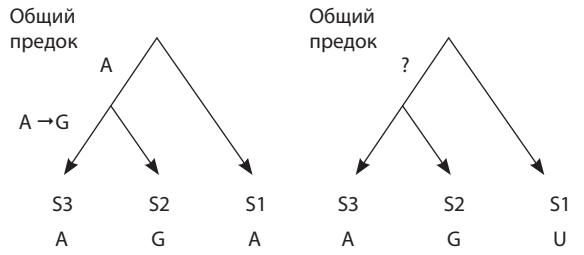


Рис. 2. Поиск нуклеотидных замен методом неукорененной максимальной парсимонии по простейшим деревьям из трех ближайших тРНК.

Слева показана идентифицированная замена A→G внутри группы из двух ближайших родственников S2 и S3, а справа – неинформативная замена.

турами тРНК» (Sprinzl et al., 1998). Каждая из тРНК была сопоставлена авторами базы данных своей аминокислоте.

Процедура формирования выборки нуклеотидных последовательностей для эволюционного анализа состояла в следующем: 1) для каждого из 123 организмов рассматривали 20 групп тРНК. Каждая группа включала тРНК, взаимодействующие с одной из 20 аминокислот. Возможным горизонтальным переносом (Soucy et al., 2015), как и переходами между группами в результате ремолдинга (смены изоакцепторной группы в результате изменения антикодона, для которой к настоящему времени известно лишь около 20 случаев (Bermudez-Santana et al., 2010; Velandia-Huerto et al., 2016; Romanova et al., 2020)), пренебрегали; 2) для каждой позиции нуклеотидных последовательностей этой группы, соответствующих конкретному организму и аминокислоте, подсчитывали частоты четырех нуклеотидов и рассматриваемой позиции приписывали нуклеотид с наибольшей частотой; при рассмотрении всех позиций последовательностей группы конструировали консенсусную последовательность группы тРНК; 3) для консенсусной последовательности, соответствующей определенной группе тРНК, оценивали ее сходство с каждой из нуклеотидных последовательностей множественного выравнивания, входящих в рассматриваемую группу, и из этой группы выбиралась та последовательность, которая была наиболее сходна с консенсусом.

Таким способом была построена выборка нуклеотидных последовательностей тРНК для эволюционного анализа, содержавшая $20 \times 123 = 2460$ типичных последовательностей тРНК (рис. 1). Каждая последовательность этой выборки была наиболее типичной для одного из изофункциональных семейств тРНК конкретного организма (из 123).

Согласно работе (Jordan et al., 2005), идентификацию нуклеотидных замен, фиксировавшихся в ходе эволюции нуклеотидных последовательностей каждого изофункционального семейства тРНК, осуществляли на основе метода неукорененной максимальной парсимонии по филогенетическим деревьям из трех вершин (рис. 2) с помощью программы Dnapars (пакет Phylip, https://phylip.web.github.io/phylip).

При анализе отдельного семейства изоакцепторных тРНК выполняли следующую процедуру: для каждой нуклеотидной последовательности S1 из 123 последовательностей тРНК, входящих в семейство, идентифицировали такую ближайшую (по сходству) нуклеотидную последовательность S2 и далее – ближайшую к S2 последовательность S3 (см. рис. 2), что S2 и S3 образовывали пару ближайших родственников. В результате формировалась филогенетическая тройка, в которой S1 была «внешним родственником» (outgroup) по отношению к паре S2 и S3.

В методе неукорененной максимальной парсимонии предполагается, что, если в определенной позиции последовательности находится нуклеотид, совпадающий у S1, S2 и S3, то этот нуклеотид содержался в той же позиции тРНК у общего предка S1, S2 и S3. Если при этом у S3 наблюдается другой нуклеотид, то вдоль ветви, ведущей к S3, произошла одиночная нуклеотидная замена. Если все три нуклеотида были различны, то, согласно (Jordan et al., 2005), эту позицию считали неинформативной и исключали из рассмотрения. Этот способ не требует стационарности и обратимости эволюционного процесса (Klopfstein et al., 2015).

Результаты

Следуя подходу I.K. Jordan с коллегами (2005) и рассматривая изменения нуклеотидов между последовательностями ближайших предков и потомков, мы построили матрицу мутационных переходов для каждого из 20 выравненных семейств тРНК. В табл. 1 приведен пример такой матрицы для семейства $тРНК^{Cys}$. Недиагональные элементы $M_{i,k}$ ($i, k = 1, \ldots, 4$) характеризуют полное число одиночных замен в последовательностях $тРНК^{Cys}$ нуклеотида i на нуклеотид k. Диагональные элементы $M_{i,k}$ соответствуют консервативным позициям. Строки и столбцы с пропусками в выравниваниях (–) в основном соответствовали району вариабельной петли и были опущены для количественных оценок.

Из табл. 1 видно, что среди нуклеотидных замен, выявленных для семейства $тРНК^{Cys}$, наиболее часто наблюдались транзиции, т. е. замены между пуринами ($N_{G\to A} = 139$ и $N_{A\to G} = 113$) и между пиримидинами ($N_{C\to U} = 177$ и $N_{U\to C} = 138$).

Существенно, что количество замен «сильных» нуклеотидов на «слабые» (G→A, G→U, C→A, C→U), в сумме равное 417, преобладает над количеством замен «слабых» нуклеотидов на «сильные» (A→G, A→C, U→C, U→G), которое равно 340. Это свидетельствует об эволюционной тенденции в направлении снижения GC-состава в тРНК в пользу увеличения AU-состава. Выявленный нами эффект,

Таблица 1. Матрица числа однонуклеотидных замен в последовательностях тРНК^{Cys}

Из\в	A	C	G	U	–
A	1526	38	113	44	27
C	43	2292	74	177	25
G	139	91	2469	58	6
U	45	138	51	1492	23
–	20	27	0	39	3131

Примечание. Здесь и в табл. 2 и 3: зеленым цветом обозначено количество замен из «сильных» нуклеотидов (G и C, образующих комплементарные пары с тремя водородными связями) в «слабые» (A и U, образующих комплементарные пары с двумя водородными связями); желтым – количество замен «слабых» нуклеотидов A и U на «сильные» нуклеотиды (G и C). В колонке, обозначенной знаком «–», указано количество замен в позициях выравнивания, соответствующих делециям.

Таблица 2. Матрица числа однонуклеотидных замен в последовательностях тРНК^{Glu}

Из\в	A	C	G	U	–
A	1353	40	101	57	37
C	52	2526	105	184	35
G	167	105	2389	51	9
U	58	124	37	1608	27
–	30	35	0	23	2956

описанный выше, был назван асимметрией нуклеотидных замен.

К тем же выводам мы приходим после рассмотрения мутационных переходов в семействе тРНК^{Glu} (табл. 2). В этом семействе количество замен «сильных» нуклеотидов на «слабые» равно 454, а количество замен «слабых» нуклеотидов на «сильные» равно 302.

Такой же анализ проведен для всех 20 изоакцепторных семейств тРНК (Приложение S2). Далее была получена оценка эффекта асимметрии по всем семействам изоакцепторных тРНК. Для этого рассчитали общую матрицу замен посредством суммирования соответствующих элементов всех 20 матриц семейств изоакцепторных тРНК (см. Приложение S2). При этом по всем тРНК количество идентифицированных одиночных замен было равно 24653, а неинформативных замен – 2083.

Диагональные элементы полученной матрицы (табл. 3) характеризуют средний нуклеотидный состав тРНК рассмотренных видов организмов – 32.9 % (G), 27.8 % (C), 21.0 % (U), 18.3 % (A), а также содержание «сильных» нуклеотидов G+C (60.7 %) и «слабых» (39.3 %). Транзиции представлены 4 из 12 недиагональных элементов. Доля транзиций в общем числе замен была равна 56 %.

Как и в большинстве частных матриц для отдельных семейств изоакцепторных тРНК (см. для примера табл. 1 и 2), в табл. 3 число замен «сильных» нуклеотидов на «слабые» (показаны зеленым цветом) превышает число замен

Таблица 3. Матрица числа нуклеотидных замен, выявленных методом неукорененной парсимонии тРНК, суммированная по всем изоакцепторным семействам

Из\в	A	C	G	U	–
A	28841	952	2949	1273	853
C	963	43778	1829	3963	951
G	3451	1881	51756	1421	330
U	1272	3438	1261	32994	715
–	666	862	210	867	53981

Таблица 4. Асимметрия нуклеотидных замен в тРНК

$A_{G\to A}$	$A_{C\to U}$	$A_{G\to U}$	$A_{G\to C}$	$A_{A\to U}$	$A_{C\to A}$
0.16	0.14	0.12	0.028	0.008	0.011

«слабых» нуклеотидов на «сильные» (отмечены желтым цветом): ср. $N_{G\to A} = 3451$ и $N_{A\to G} = 2949$, $N_{C\to U} = 3963$ и $N_{U\to C} = 3468$, $N_{G\to U} = 1421$ и $N_{U\to G} = 1261$, $N_{C\to A} = 963$ и $N_{A\to C} = 952$.

Для количественной оценки асимметрии замен вычислим относительную разницу, определяемую как удвоенную разность двух величин, деленную на их сумму, числа замен между нуклеотидами F и Z, где F, Z∈(A, U, G, C):

$$A_{F\to Z} = \frac{2(N_{F\to Z} - N_{Z\to F})}{N_{F\to Z} + N_{Z\to F}}. \quad (1)$$

В табл. 4 представлены результаты расчетов $A_{F\to Z}$ на основе уравнения (1) и табл. 3. Асимметрия числа замен составила: 0.16 для G→A и A→G; 0.14 для C→U и U→C; 0.12 для G→U и U→G. Остальные переходы оказались слабо асимметричными: от 0.008 до 0.028 (см. табл. 4).

На основе табл. 3 мы можем также рассчитать баланс потерь и приобретений B_F для нуклеотида типа F:

$$B_F = \sum_Z (N_{Z\to F} - A_{F\to Z}). \quad (2)$$

В табл. 5 продемонстрировано суммарное снижение числа «сильных» нуклеотидов G/C в исследованных нуклеотидных последовательностях всех проанализированных семейств тРНК на 1198 (714 G+484 C) за счет эволюционного приобретения такого же количества слабых нуклеотидов A/G (512 A+686 U). С учетом суммарного количества нуклеотидов G, C, A и U в исследованных последовательностях тРНК изменения количества этих нуклеотидов в ходе эволюции семейств тРНК, нормированные на их количество, составили: –0.014, –0.011, +0.018 и +0.021 для G, C, A и U соответственно (см. табл. 5).

Матрицы числа нуклеотидных замен для всех 20 изоакцепторных семейств тРНК приведены в Приложении S2. В табл. 6, полученной на основе этих 20 матриц, приведены арифметические разности $N_{F\to Z} - N_{Z\to F}$ (F, Z ∈(A, U, G, C) между количествами всех возможных типов нуклеотидных замен, фиксировавшихся в процессе эволюции 20 изоакцепторных семейств тРНК. Каждому варианту арифметической разности замен F→Z и Z→F со-

Таблица 5. Характеристики состава и эволюционной динамики исследованных нуклеотидных последовательностей всех проанализированных семейств тРНК

Характеристики состава и эволюционной динамики	G	C	A	U
Суммарное количество консервативных нуклеотидов четырех типов в деревьях неукорененной парсимонии для исследованных последовательностей тРНК	51 756	43 778	28 841	32 994
Среднее содержание нуклеотидов четырех типов в исследованных последовательностях тРНК, %	32.9	27.8	18.3	21.0
Изменения количества нуклеотидов четырех типов в ходе эволюции семейств тРНК	–714	–484	+512	+686
Изменения количества нуклеотидов четырех типов в ходе эволюции семейств тРНК, нормированные на их количество	–0.014	–0.011	+0.018	+0.021

Таблица 6. Арифметические разности $N_{F\to Z} - N_{Z\to F}$ (F, Z∈(A, U, G, C)) между количествами всех возможных типов нуклеотидных замен, фиксировавшихся в процессе эволюции 20 изоакцепторных семейств тРНК

tRNA	$N_{G\to A} - N_{A\to G}$	$N_{C\to U} - N_{U\to C}$	$N_{G\to U} - N_{U\to G}$	$N_{G\to C} - N_{C\to G}$	$N_{A\to U} - N_{U\to A}$	$N_{C\to A} - N_{A\to C}$	$A_{S\to W}$*
Ala	–5	36	21	0	–20	–4	0.13
Arg	20	41	4	6	21	21	0.14
Asn	45	30	10	4	–10	–11	0.19
Asp	32	4	20	2	–2	13	0.21
Cys	26	39	7	17	–1	5	0.20
Gln	–4	–2	21	–3	10	31	0.11
Glu	66	60	14	0	–1	12	0.40
His	52	2	–18	10	–4	–17	0.04
Ile	25	–2	5	16	13	8	0.12
Leu	62	89	25	6	–13	25	0.14
Met	34	45	–11	14	7	–9	0.12
Phe	20	44	7	4	19	8	0.24
Pro	29	21	24	–9	2	14	0.20
Ser	50	105	61	–5	–12	–32	0.19
Trp	44	13	6	0	3	–4	0.16
Tyr	44	48	7	–3	–23	5	0.24
Gly	–11	4	–7	9	5	6	–0.04
Thr	–21	–26	–17	–14	–5	0	–0.06
Val	12	12	–23	–5	7	–19	–0.03
Lys	–18	–58	–31	3	5	–41	–0.34

* В последнем столбце приведено значение относительной разности количества замен между «сильными» и «слабыми» нуклеотидами, $A_{S\to W} = 2(N_{S\to W} - N_{W\to S})/(N_{S\to W} + N_{W\to S})$, где S∈(G, C), W∈(A, U).

ответствует определенный столбец табл. 6. Каждая строка этой таблицы соответствует определенному изоакцепторному семейству тРНК. В последней колонке приведена относительная разность количества замен, $A_{S\to W}$, «сильных» нуклеотидов, S∈(G, C), на слабые, W∈(A, U), определяемая уравнением (1).

Из табл. 6 можно видеть, что 16 семейств тРНК характеризуются положительной величиной относительной разности количества замен, $A_{S\to W} > 0$. Вместе с тем для четырех семейств тРНК (нижние строки) характерна отрицательная разность, $A_{S\to W} < 0$. Из этих четырех семейств тРНК для трех тРНК, $тРНК^{Gly}$, $тРНК^{Thr}$ и $тРНК^{Val}$, наблюдаемый отрицательный тренд, т. е. преобладание замен W→S над S→W, незначителен ($-0.06 \le A_{S\to W} \le -0.03$), и только для $тРНК^{Lys}$ преобладание замен W→S над S→W имеет выраженный характер ($A_{S\to W} = -0.34$).

Для оценки статистической значимости преобладания положительных величин , характеризующих относитель-

ную разность между а) количеством замен «сильных» нуклеотидов на «слабые» (S→W) и б) количеством замен «слабых» нуклеотидов на «сильные» (W→S), фиксировавшихся в ходе эволюции 20 семейств тРНК, использовали односторонний биномиальный тест (Lehmann, 2012). В нашем случае уровень значимости вычисляли как вероятность p случайного наблюдения 16 матриц из 20 с заменами в пользу уменьшения количества «сильных» нуклеотидов G/C: см. выражение (3). При этом предполагалось, что количество фиксированных замен типов S→W и W→S в среднем одинаково:

$$p = \sum_{l=16}^{l=20} C_l^{20} 0.5^{20} = 0.0059. \quad (3)$$

С использованием выражения (3) статистическая гипотеза об асимметрии матриц эволюционных замен в направлении потери нуклеотидов G и C и приобретения нуклеотидов A и U была принята с уровнем значимости $p < 0.006$.

Обсуждение

Проведенный нами анализ эволюции 20 семейств изоакцепторных тРНК 123 видов трех доменов (Bacteria, Archaea и Eukaryota) от их предковых форм выявил тенденцию снижения G/C-состава тРНК в пользу увеличения A/U-состава. Этот эффект был назван асимметрией нуклеотидных замен. Он заключался в эволюционной потере «сильных» нуклеотидов G и C, способных формировать энергетически выгодные комплементарные пары с тремя водородными связями и приобретении «слабых» нуклеотидов A и U, формирующих менее стабильные комплементарные пары с двумя водородными связями. Обнаруженному изменению состава последовательностей были подвержены 16 из 20 семейств тРНК, что соответствует уровню статистической значимости $p < 0.006$, согласно одностороннему биномиальному тесту.

Полученные результаты свидетельствуют в пользу предположения о том, что последний универсальный общий предок (LUCA) обитал в более горячей среде, нежели ныне живущие организмы, т. е. был термофилом или теплолюбивым мезофилом (умеренным термофилом). Этот вывод обосновывается тем, что содержание нуклеотидов G и C в нуклеотидных последовательностях связано с оптимальной температурой обитания организмов (Dutta, Chaudhuri, 2010), поэтому генетические макромолекулы (ДНК, РНК) можно рассматривать как своеобразные молекулярные термометры, а их GC-содержание – как индикатор температуры среды обитания.

Условия ранней Земли должны были определить энергетические, метаболические, биохимические и экологические особенности LUCA. Согласно работам (Di Giulio, 2000; Weiss et al., 2016), LUCA жил в горячих источниках, высокая температура которых облегчает протекание биохимических реакций и молекулярно-генетических процессов, но требует термодинамической и кинетической устойчивости биомолекулярных структур, термодинамические флуктуации которых значительно выражены при более высокой температуре среды обитания. Современные термофилы адаптированы к большим температурам благодаря в том числе высокому содержанию в геноме нуклеотидов G и C (Dutta, Chaudhuri, 2010), формирующих более энергетически прочные комплементарные связи друг с другом. И это особенно важно для термостабильности структурных РНК, включая тРНК.

Необходимо отметить, что четыре семейства тРНК из 20 исследованных в нашей работе не следуют общей тенденции потери «сильных» нуклеотидов. Причины, обуславливавшие особенности эволюции этих тРНК, могут быть разнообразны. Например, два семейства, $тРНК^{Gly}$ и $тРНК^{Val}$, соответствуют химически простым так называемым миллеровским аминокислотам. Предположительно, эти аминокислоты входили в самые древние белки, и нуклеотидный состав их тРНК мог успеть достигнуть их индивидуального эволюционного равновесия, пусть и отличного от среднего по всем тРНК. Однако в целом при сравнении GC-состава тРНК организмов, живущих при разных температурах, наши результаты свидетельствуют о том, что современные организмы в среднем живут в более холодной среде, нежели последний универсальный общий предок (LUCA).

Заключение

Обнаружен универсальный вектор направленного эволюционного изменения последовательностей тРНК, при котором замены гуанина (G) и цитозина (C) на аденин (A) и урацил (U) суммарно происходят чаще обратных. В результате эволюционного процесса тРНК могли терять «сильные» комплементарные пары с тремя водородными связями, формируемые гуанином и цитозином, и фиксировать «слабые» комплементарные пары с двумя водородными связями, образуемыми аденином и урацилом. Обнаруженному изменению состава последовательностей были подвержены 16 из 20 семейств тРНК, что соответствует уровню статистической значимости $p = 0.006$ согласно одностороннему биномиальному тесту. Выявленная закономерность свидетельствует о высоком GC-содержании в последовательностях общих предков современных тРНК и, следовательно, в пользу предположения о том, что самая молодая из гипотетических общих предковых клеток, от которой произошли все ныне живущие организмы (последний универсальный общий предок, LUCA), обитала в более горячей среде, нежели ныне живущие организмы.

Список литературы / References

Bermudez-Santana C., Attolini C.S.-O., Kirsten T., Engelhardt J., Prohaska S.J., Steigele S., Stadler P.F. Genomic organization of eukaryotic tRNAs. *BMC Genomics*. 2010;11(1):270. doi 10.1186/1471-2164-11-270

Cantine M.D., Fournier G.P. Environmental adaptation from the origin of life to the last universal common ancestor. *Orig Life Evol Biosph*. 2017;48(1):35-54. doi 10.1007/s11084-017-9542-5

Di Giulio M. The universal ancestor lived in a thermophilic or hyperthermophilic environment. *J Theor Biol*. 2000;203(3):203-213. doi 10.1006/jtbi.2000.1086

Dutta A., Chaudhuri K. Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: indications for thermal adaptation. *FEMS Microbiol Lett*. 2010;305(2):100-108. doi 10.1111/j.1574-6968.2010.01922.x

Galtier N., Tourasse N., Gouy M. A non hyperthermophilic common ancestor to extant life forms. *Science*. 1999;283(5399):220-221. doi 10.1126/science.283.5399.220

Jordan I.K., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., Sunyaev S. A universal trend of amino acid gain and loss in protein evolution. *Nature*. 2005;433(7026):633-638. doi 10.1038/nature03306

Klopfstein S., Vilhelmsen L., Ronquist F. A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Syst Biol.* 2015;64(6):1089-1103. doi 10.1093/sysbio/syv052

Lehmann E.L. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? In: Rojo J. (Ed.) Selected Works of E.L. Lehmann. Selected Works in Probability and Statistics. Boston, MA: Springer, 2012;201-208. doi 10.1007/978-1-4614-1412-4_19

Men Y., Lu G., Wang Y., Lin J., Xie Q. Reconstruction of the rRNA sequences of LUCA, with bioinformatic implication of the local similarities shared by them. *Biology*. 2022;11(6):837. doi 10.3390/biology11060837

Moody E.R.R., Álvarez-Carretero S., Mahendrarajah T.A., Clark J.W., Betts H.C., Dombrowski N., Szánthó L.L., … Spang A., Pisani D., Williams T.A., Lenton T.M., Donoghue P.C.J. The nature of the last universal common ancestor and its impact on the early Earth system. *Nat Ecol Evol.* 2024;8(9):1654-1666. doi 10.1038/s41559-024-02461-1

Rickert D.A., Fan L.W.-T., Hahn M.W. Inconsistency of parsimony under the multispecies coalescent. *Theor Popul Biol*. 2025;166:56-69. doi 10.1016/j.tpb.2025.09.004

Romanova E.V., Bukin Y.S., Mikhailov K.V., Logacheva M.D., Aleoshin V.V., Sherbakov D.Yu. Hidden cases of tRNA gene duplication and remolding in mitochondrial genomes of amphipods. *Mol Phylogenet Evol*. 2020;144:106710. doi 10.1016/j.ympev.2019.106710

Soucy S.M., Huang J., Gogarten J.P. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 2015;16(8):472-482. doi 10.1038/nrg3962

Sprinzl M., Horn C., Brown M., Ioudovitch A., Steinberg S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*. 1998;26(1):148-153. doi 10.1093/nar/26.1.148

Velandia-Huerto C.A., Berkemer S.J., Hoffmann A., Retzlaff N., Romero Marroquín L.C., Hernández-Rosales M., Stadler P.F., Bermúdez-Santana C.I. Orthologs, turn-over, and remolding of tRNAs in primates and fruit flies. *BMC Genomics*. 2016;17(1):617. doi 10.1186/s12864-016-2927-4

Weiss M.C., Sousa F.L., Mrnjavac N., Neukirchen S., Roettger M., Nelson-Sathi S., Martin W.F. The physiology and habitat of the last universal common ancestor. *Nat Microbiol*. 2016;1(9):16116. doi 10.1038/nmicrobiol.2016.116

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию 13.09.2025. После доработки 07.10.2025. Принята к публикации 07.10.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1129-1136

doi 10.18699/vjgb-25-117

Оценка зависимости показателей мозговой активности от индивидуальной однонуклеотидной вариабельности генетических маркеров большого депрессивного расстройства с использованием анализа главных компонент

К.А. Зорина [1], А.А. Кривецкий [4], В.С. Карманов [4], А.Н. Савостьянов [1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[4] Новосибирский государственный технический университет, Новосибирск, Россия
a.savostianov@g.nsu.ru

Аннотация. Большое депрессивное расстройство (БДР) относится к наиболее широко распространенным психическим заболеваниям, что обусловливает необходимость поиска факторов повышенной предрасположенности к этому нарушению. В качестве молекулярно-генетических маркеров БДР часто рассматривают однонуклеотидные полиморфизмы генов нейромедиаторных систем мозга. Показатели индивидуальной однонуклеотидной вариабельности в генах нейромедиаторов применяются для оценки риска появления БДР до проявления его симптоматики на поведенческом уровне. Однако прогностические возможности анализа геномных вариаций для оценки риска депрессии до настоящего времени недостаточно надежны и дополняются поведенческой и нейрофизиологической информацией о пациентах. Нейрофизиологические маркеры БДР дают наиболее надежные оценки выраженности патологической симптоматики, но они отражают состояние человека в момент обследования, а не предрасположенность к возникновению этого патологического состояния и не позволяют выполнить оценку риска его появления в будущем. Большое депрессивное расстройство часто сопровождается отклонениями в способности человека контролировать двигательные реакции, включая возможность произвольно подавлять неадекватное поведение. «Стоп-сигнал парадигма» (ССП) – экспериментальный метод для оценки функционального баланса между тормозными и активационными системами головного мозга в условиях выполнения целенаправленных движений. Объединенный с регистрацией ЭЭГ, этот экспериментальный метод позволяет учитывать не только поведенческие характеристики участников, такие как скорость или точность ответов, но и нейрофизиологические особенности головного мозга, ассоциированные с контролем над поведением. Цель настоящего исследования заключалась в оценке зависимости между особенностями ЭЭГ реакций в условиях парадигмы стоп-сигнал и индивидуальной однонуклеотидной вариабельностью в генах-кандидатах для выявления БДР. Размерность в исходных генетических и нейрофизиологических экспериментальных данных была снижена при помощи анализа главных компонент (PCA) для последующего выявления ассоциации между компонентами ЭЭГ реакций, регистрируемыми в условиях контроля произвольных двигательных реакций, и однонуклеотидными вариациями в генах, изменчивость которых ассоциирована с риском БДР. Установлено, что изменчивость в этих генах ассоциирована с амплитудными показателями мозговых ответов в условиях тестирования испытуемых методом ССП. Полученные результаты могут быть использованы для разработки систем ранней диагностики депрессии, выявления индивидуальных паттернов нарушения в работе головного мозга, подбора методов коррекции заболевания и контроля над эффективностью терапии.

Ключевые слова: стоп-сигнал парадигма; ЭЭГ; вызванные потенциалы; однонуклеотидные полиморфизмы; большое депрессивное расстройство; метод главных компонент; регрессионный анализ

Для цитирования: Зорина К.А., Кривецкий А.А., Карманов В.С., Савостьянов А.Н. Оценка зависимости показателей мозговой активности от индивидуальной однонуклеотидной вариабельности генетических маркеров большого депрессивного расстройства с использованием анализа главных компонент. *Вавиловский журнал генетики и селекции*. 2025;29(7):1129-1136. doi 10.18699/vjgb-25-117

Финансирование. Исследование выполнено в рамках бюджетного проекта ИЦиГ СО РАН № FWNR-2022-0020.

Assessing the dependence of brain activity on individual single-nucleotide variability of genetic markers of major depressive disorder using principal component analysis

K.A. Zorina [1], A.A. Kriveckiy [4], V.S. Karmanov [4], A.N. Savostyanov [1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[4] Novosibirsk State Technical University, Novosibirsk, Russia
a.savostianov@g.nsu.ru

© Зорина К.А., Кривецкий А.А., Карманов В.С., Савостьянов А.Н., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

Abstract. Major depressive disorder (MDD) is one of the most widespread mental illnesses, which necessitates the search for factors of increased predisposition to this disorder. Single nucleotide polymorphisms in genes of the brain's neurotransmitter systems are often considered as molecular genetic markers of MDD. Indicators of individual single nucleotide variability in neurotransmitter genes are used to assess the risk of MDD before its symptomatology at the behavioral level. However, the predictive capabilities of analyzing genomic variations to assess the risk of depression are not yet sufficiently reliable and are complemented by behavioral and neurophysiological information about patients. Neurophysiological markers of MDD provide the most reliable estimates of the severity of pathological symptoms, but they reflect a person's state at the time of examination, and not a predisposition to the occurrence of this pathological state and do not allow assessing the risk of its appearance in the future. Major depressive disorder is often accompanied by abnormalities in a person's ability to control motor responses, including the ability to voluntary suppress inappropriate behavior. The "stop-signal paradigm" (SSP) is an experimental method for assessing the functional balance between the inhibitory and activation systems of the brain during targeted movements. Combined with EEG recording, this experimental method allows for the consideration of not only participants' behavioral characteristics, such as speed or accuracy of responses, but also the brain's neurophysiological features associated with behavior control. The objective of this study was to evaluate the relationship between EEG responses in the stop-signal paradigm and individual single nucleotide variability in candidate genes for MDD detection. Dimensionality in the original genetic and neurophysiological experimental data was reduced by principal component analysis (PCA) to subsequently detect an association between EEG response components recorded during the control of random motor responses and single nucleotide variations in genes, the variability of which is associated with MDD risk. Variability in these genes has been shown to be associated with the amplitude of brain responses under the conditions of test subjects using the PCA method. The results obtained can be used to develop systems for the early diagnosis of depression, identify individual patterns of impairment in the brain, select methods for correcting the disease and control the effectiveness of therapy.

Key words: stop-signal paradigm; EEG; event related potentials; single nucleotide polymorphisms (SNPs); major depressive disorder; principal component analysis; regression analysis

For citation: Zorina K.A., Kriveckiy A.A., Karmanov V.S., Savostyanov A.N. Assessing the dependence of brain activity on individual single-nucleotide variability of genetic markers of major depressive disorder using principal component analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1129-1136. doi 10.18699/vjgb-25-117

Введение

Большое депрессивное расстройство (БДР), или клиническая депрессия, – психическое заболевание, симптомы которого включают: подавленное настроение, потерю интереса к ранее выполнявшимся делам, упадок сил и усталость, изменения сна и аппетита, трудности с концентрацией внимания и памятью, а также чувство вины и низкую самооценку (DSM-5, 2013). Заболевание относится к числу самых распространенных психических расстройств (Wong, Licinio, 2001). Известно, что предрасположенность к разным типам депрессивных расстройств зависит как от генетических факторов, так и от особенностей жизни пациентов, в том числе в период, предшествующий появлению симптомов БД (Cross-Disorder Group, 2013; Northoff, 2013; Haase, Brown, 2015; Ivanov et al., 2019; Whitney et al., 2019). Моноаминовая теория депрессии в течение многих лет считалась наиболее достоверной, и в качестве молекулярных маркеров предрасположенности к депрессии наиболее часто рассматривали аллельные полиморфизмы генов моноаминовых нейромедиаторных систем мозга (Willner et al., 2013). Однако попытки оценить риск появления депрессии только на основе анализа генетических данных обычно оказывались неудовлетворительными (Duncan et al., 2014; Halldorsdottir, Binder, 2017) вследствие того, что депрессия – мультифакториальное заболевание, формирующееся в результате комбинации разных генетических факторов и факторов окружающей среды (Ivanov et al., 2019; Wang et al., 2025). В связи с этим выявление надежных маркеров депрессии предполагает одновременное использование не только генетических, но также и нейрофизиологических показателей, отражающих состояние головного мозга человека.

В качестве нейрофизиологических маркеров депрессии могут быть использованы значения амплитуды и латентности вызванных потенциалов головного мозга (Stone et al., 2025). Хорошо известно, что депрессия часто ассоциирована с нарушениями тормозного контроля, что проявляется как на поведенческом, так и на нейрофизиологическом уровнях (Shetty et al., 2025). Примером метода, позволяющего определить индивидуальную способность к самоконтролю поведения, служит экспериментальная стоп-сигнал парадигма (ССП, stop-signal paradigm) (Band et al., 2003). Этот метод тестирования обеспечивает объективную оценку функционального баланса между работой активационных систем мозга, контролирующих целенаправленные действия, и тормозных систем, ответственных за подавление неадекватного поведения.

Одна из основных проблем комплексного изучения депрессии – необходимость учитывать большое количество факторов, взаимосвязи между которыми изначально не очевидны для исследователя. Решение проблемы может быть достигнуто за счет использования методик снижения размерности экспериментальных данных, применяемых для поиска межфакторных зависимостей. В частности, метод главных компонент (Principal Component Analysis, PCA) широко внедряется для снижения размерности исходных данных и выявления наиболее информативных признаков (Gewers et al., 2021). PCA позволяет преобразовывать исходные характеристики в пространство меньшей размерности для уменьшения числа анализируемых параметров, чтобы избежать избыточности из-за данных высокой размерности (Subasi, Gursoy, 2010).

Целью исследования было выявление зависимости нейрофизиологических показателей, регистрируемых в

парадигме стоп-сигнал, от индивидуальной однонуклеотидной вариабельности в генах, связанных с повышенным риском депрессии.

В нашей работе проведен анализ генетических и нейрофизиологических данных, взятых из открытой базы ICBrainDB, созданной сотрудниками Института цитологии и генетики СО РАН и НИИ нейронаук и медицины и размещенной на сайте ИЦиГ СО РАН (Ivanov et al., 2022). Гены-кандидаты для БДР были отобраны ранее на основе биоинформатического анализа научных публикаций из открытых баз данных, содержащих информацию о наличии у обследованных людей расстройств депрессивного спектра (Ivanov et al., 2019).

Материалы и методы

Выборка участников обследования. Объем выборки, для которой были проанализированы геномные и ЭЭГ данные, составил 212 человек. Среди них 47 участников обследования, проживавших в Новосибирске, имели диагностированное большое депрессивное расстройство; не имели диагностированной депрессии 165 участников, 67 из них проживали в Новосибирске, 50 – в Якутске, 48 – в пос. Хандыга, Республика Саха (Якутия).

Схема эксперимента. Участникам предлагалось выполнить серию заданий в парадигме стоп-сигнал, модифицированной А.Н. Савостьяновым с коллегами (Savostyanov et al., 2009). При выполнении задания участнику демонстрировали одно из двух изображений – целевой стимул, после появления которого необходимо было нажимать кнопку на клавиатуре. В некоторых случаях после появления целевого визуального стимула на экране появлялся стоп-сигнал, после чего участник должен был прервать уже начатое действие. Всего в ходе эксперимента участнику предъявлялось 135 заданий, 35 из которых содержали стоп-сигнал. ЭЭГ регистрировалась с помощью 128-канального усилителя NVX-132. Электроды располагались по международной схеме «10-5» с заземляющим электродом AFz и референтным Cz. Ширина полосы пропускания составляла от 0.3 до 100 Гц, а частота оцифровки сигнала 1000 Гц.

Обработка ЭЭГ сигналов. Исходные ЭЭГ сигналы содержали внемозговой шум, включающий артефакты движений глаз, лицевой мимической мускулатуры, электрической активности сердца и кровеносных сосудов. Все внемозговые артефакты удалялись с помощью метода анализа независимых компонентов ICA (Independent component analysis), реализованного в программе EEGLAB (Delorme, Makeig, 2004). ICA – это алгоритм компьютерной обработки данных, позволяющий выделить несколько групп показателей с высокой степенью статистической независимости друг от друга. В свою очередь, PCA находит показатели с высокой степенью взаимной зависимости.

В исходной записи ЭЭГ были выделены: go-эпизоды – интервалы мозговой активности, содержащие момент нажатия участником клавиши после появления целевого визуального стимула, и stop-эпизоды – интервалы, соответствующие ситуациям торможения после появления стоп-сигнала. Для go-эпизодов эпохирование ЭЭГ выполнено относительно момента появления целевого стимула, а для stop-эпизодов относительно момента предъявления стоп-сигнала. В go-эпизодах были выделены два пика вызванных потенциалов: премоторный (400–600 мс после стимула) и постмоторный (700–800 мс после стимула) пик. Премоторный пик отражает мозговую активность в условиях подготовки к выполнению движения, а постмоторный пик указывает на состояние мозга в момент выполнения движения.

В stop-эпизодах были выделены два пика, предшествовавшие подавлению двигательного ответа или следующие за ним. Описанные пики с соответствующими им временными окнами выявлялись на основе визуального анализа графиков вызванных потенциалов для канала C3, который расположен над моторными отделами коры левого полушария. С помощью программы ERPLAB (Lopez-Calderon, Luck, 2014) для каждого из указанных пиков были вычислены значения максимальной амплитуды, средней амплитуды пика и латентности отдельно для каждого участника исследования и каждого канала ЭЭГ. Поскольку каждый участник во время обследования выполнял 100 go- и 35 stop-эпизодов, показатели мозговых ответов для отдельных проб усреднялись по каждому участнику. Каналы ЭЭГ были объединены в группы с различной локализацией на поверхности головы. Всего было выделено 12 групп электродов: левая лобная, медиальная лобная, правая лобная, левая височная, левая центральная, медиальная центральная, правая центральная, правая лобная, левая теменная, медиальная теменная, правая теменная и объединенная затылочная. Таким образом, анализ был выполнен для 144 параметров исходных ЭЭГ данных: 12 (групп электродов) × 2 (пиков вызванных потенциалов) × 3 (количественные оценки максимальной амплитуды, средней амплитуды и латентности пика) × 2 (экспериментальные условия go или stop).

Генетические данные. Генетический материал в виде крови или буккального эпителия был взят у всех участников обследования. Целевое секвенирование 164 генов было выполнено на основе этого материала. Гены были отобраны в рамках реконструкции и анализа генной сети по фактору наличия влияния на предрасположенность к БДР (Ivanov et al., 2019). Библиотеки целевого секвенирования были подготовлены для этих 164 генов, и NGS секвенирование с высоким покрытием было проведено для всех участников. Для каждого аллеля гена из списка у каждого участника обследования был определен показатель бинарной вариабельности в сравнении с референсным геномом (Ivanov et al., 2022). В случае, если у участника аллель гена полностью соответствовал референсному геному, показатель изменчивости был равен нулю; если имелась хотя бы одна замена – этот показатель был равен единице. При наличии в аллеле нескольких замен этот показатель также приравнивался к единице. Всего у всех обследованных участников было идентифицировано 799 однонуклеотидных замен в 121 из рассмотренных генов. В 43 генах из анализируемого списка ни у одного из участников не было найдено каких-либо нуклеотидных замен. Таким образом, количество анализируемых параметров для входных генетических данных было равно 242 (121 гена по два аллеля на каждый ген).

Таблица 1. Пример пары показателей для оценки линейной регрессии.
В качестве первого показателя выбрана индивидуальная вариабельность гена *ADRA2B*,
а в качестве второго показателя – амплитуда пика премоторного пика вызванного потенциала в правой теменной коре

ID участника	Ген *ADRA2B* (0 – нет вариабельности; 1 – есть вариабельность)	Амплитуда постмоторного пика ЭЭГ в условии "go" в правой теменной коре, мкВ
D_Nov_001	0	1.68
D_Yak_2016_001	1	8.05

Результаты

Как указано ранее, цель нашей работы заключалась в оценке зависимости между особенностями ЭЭГ реакций в условиях парадигмы стоп-сигнал и индивидуальной однонуклеотидной вариабельностью в генах-кандидатах для оценки риска БДР. Для ее достижения проведен многоэтапный анализ полученных экспериментальных данных, результаты которого приведены ниже.

Задача 1. Определение списка генов-кандидатов БДР, для которых выявляется зависимость между однонуклеотидной вариабельностью и показателями ЭЭГ

В рамках решения Задачи 1 был построен линейный модельный ряд, в котором каждый ЭЭГ параметр рассматривался в качестве отдельной зависимой переменной, а бинарный признак наличия или отсутствия однонуклеотидных замен в конкретном гене – в качестве независимой переменной. Под «линейным модельным рядом» имелись в виду отдельные линейные регрессии для каждой пары «ЭЭГ параметр – однонуклеотидная изменчивость» (табл. 1). Поскольку было использовано 144 ЭЭГ параметра и 242 генетических параметра, то исходное число пар параметров, взятых для линейной регрессии, составило 34 848. Отдельная модель линейной регрессии выглядит следующим образом:

$$\text{ЭЭГ параметр} = B_0 + B_1 + e.$$

Здесь B_1 – бинарный признак, равный 0 при отсутствии вариабельности и 1 в противоположном случае.

В этом случае зависимая переменная – количественный показатель ЭЭГ, а предиктор – бинарный признак наличия или отсутствия замены в гене. Если в одном аллеле была хотя бы одна замена, то бинарному признаку присваивалось значение один. Два аллеля одного и того же гена анализировались как два разных бинарных признака. Таким образом проверялось, ассоциируется ли однонуклеотидная вариабельность в каждом из выбранных генов с изменением параметра ЭЭГ.

При решении Задачи 1 коррекция множественных сравнений проводилась с использованием метода Бенджамини–Хохберга (FDR, False discovery rate) для контроля ожидаемой доли ложных отклонений от нулевой гипотезы (Benjamini, Hochberg, 1995). FDR – более мощный метод, чем поправка Бонферрони, и он полезен в тех случаях, когда количество сравниваемых гипотез велико или когда важно минимизировать количество ложных срабатываний, а не строго контролировать вероятность ошибки первого рода для каждой отдельной гипотезы.

Были проверены зависимости для 144 ЭЭГ показателей с вариабельностью по каждому из 121 гена, в которых однонуклеотидные замены были выявлены хотя бы у одного участника. Проведенный анализ показал, что статистически значимые зависимости (скорректированный после FDR уровень достоверности $q < 0.05$) были выявлены только для пяти генов (*ADRA2B*, *TF*, *HCRTR2*, *WFS1*, *PENK*) и четырех ЭЭГ показателей (в go-эпизодах в медиальной фронтальной, правой и левой теменной, объединенной затылочной областях коры головного мозга) (табл. 2). Отметим, что для трех генов, *ADRA2B*, *TF*, *HCRTR2*, достоверные зависимости выявлены в трех областях коры (правой теменной, левой теменной и затылочной), а для двух генов, *WFS1* и *PENK*, только в медиальной лобной коре. Выбранные таким образом гены были включены в последующий анализ.

В табл. 2 представлены результаты линейного регрессионного анализа зависимости ЭЭГ показателей от наличия полиморфизмов в генах-кандидатах БДР. Указаны 11 наиболее значимых пар «ген–показатель ЭЭГ» с наименьшими значениями FDR-корректированных *p*-уровней (*q*-values). Также указаны некорректированные p-значения. Все показанные ассоциации достоверны при FDR < 0.05.

Среднее количество однонуклеотидных замен для каждого из пяти выбранных генов в общей выборке обследованных участников представлено в табл. 3. Распределение носительства замены по выбранным генам варьирует примерно от одной трети до двух третей участников, что обеспечивает достаточную вариативность для статистического анализа.

Задача 2. Снижение размерности нейрофизиологических данных при помощи анализа главных компонент

При решении Задачи 2 для снижения размерности анализируемых данных ЭЭГ был применен метод главных компонент (Principal Component Analysis, PCA) с предварительной стандартизацией признаков (Rokhlin et al., 2010). На исходном наборе из 144 ЭЭГ величин были определены 15 главных компонент. Рисунок показывает, что эти 15 компонент в существенной степени описывают вариативность исходных ЭЭГ параметров, в совокупности объясняя около 80 % общей дисперсии.

Задача 3. Оценка влияния вариабельности в генах-кандидатах БДР на интегральные показатели мозговой активности, определенные при помощи PCA

При решении этой задачи для каждого из пяти выбранных генов, имеющих статистически значимые связи со сходными ЭЭГ показателями (см. табл. 2), был проведен регрессионный анализ между главными компонентами

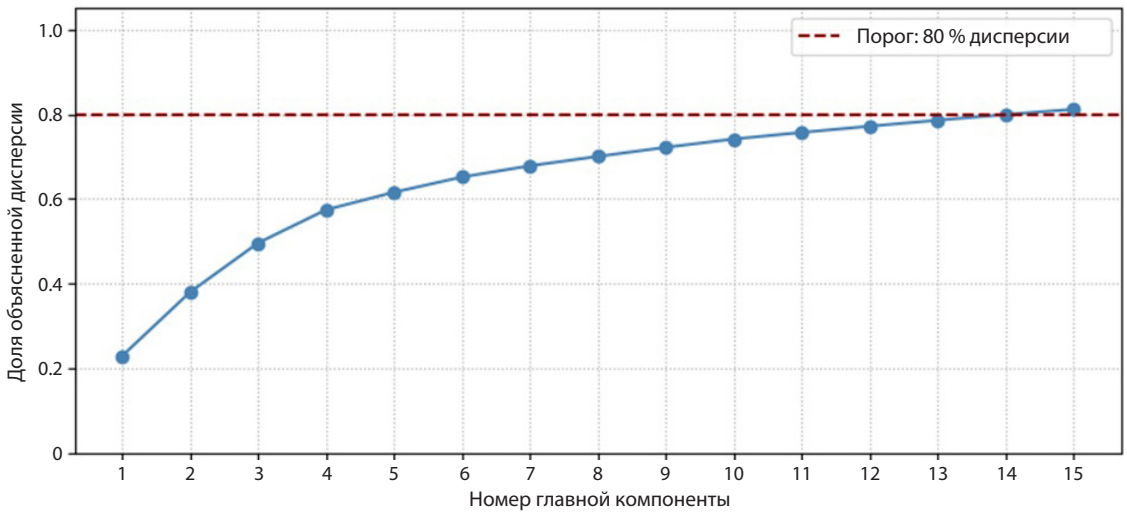
Таблица 2. Результаты анализа зависимости амплитуды пост-моторного пика в go-эпизодах стоп-сигнал парадигмы от наличия бинарной изменчивости в генах-кандидатах БДР

Ген, для которого выявлена вариабельность	Область коры мозга, ЭЭГ ответы в которой зависели от вариабельности гена	Уровень достоверности *p*-value	Скорректированный уровень достоверности *q*-value
ADRA2B	Правая теменная	7.35E-34	1.15E-29
	Левая теменная	9.88E-26	1.03E-22
	Затылочная	2.65E-28	5.91E-25
TF	Правая теменная	1.48E-32	7.70E-29
	Левая теменная	1.66E-26	1.85E-23
	Затылочная	1.34E-29	4.17E-26
HCRTR2	Правая теменная	1.48E-32	7.70E-29
	Левая теменная	1.66E-26	1.85E-23
	Затылочная	1.34E-29	4.17E-26
WFS1	Медиальная лобная	4.93E-27	8.53E-24
PENK	Медиальная лобная	5.44E-28	1.06E-24

Таблица 3. Среднее количество однонуклеотидных замен в выбранных генах-кандидатах БДР*

Ген-кандидат БДР	Значение изменчивости**	Стандартное отклонение (Std)	Доля людей, у которых нет замен в этом гене, %
ADRA2B	0.29	0.45	70.54
TF	0.34	0.47	65.75
HCRTR2	0.34	0.47	65.75
WFS1	0.65	0.48	34.93
PENK	0.63	0.48	36.98

* Представлены данные только по тем генам, для которых выявлены достоверные зависимости между вариабельностью и ЭЭГ показателями. ** Средние значения в данной ситуации отражают долю обследованных в выборке, у которых замена в гене встречалась хотя бы один раз.



Кумулятивное объяснение дисперсии с использованием метода главных компонент для ЭЭГ данных.
Красной штриховой линией отмечен достигнутый порог в 80 %.

и бинарными показателями наличия полиморфизмов. В этом случае, в отличие от Задачи 1, регрессионный анализ проводился не для отдельных ЭЭГ показателей, а для интегральных показателей, которые в совокупности объясняют 80 % общей межиндивидуальной дисперсии ЭЭГ данных (см. рисунок).

Оказалось, что из 15 выделенных с помощью PCA компонент мозговой активности только третья компонента имела достоверную взаимосвязь с генетической изменчивостью в генах-кандидатах БДР, что демонстрирует табл. 4, которая содержит результаты оценки статистической значимости влияния генетической вариабельности

в пяти рассматриваемых генах-кандидатах для трех наиболее значимых компонент PCA.

Чтобы провести нейрофизиологическую интерпретацию полученных зависимостей, для каждого из исходных ЭЭГ показателей была вычислена факторная нагрузка для третьей компоненты PC3. В нашем случае факторная нагрузка для компоненты в методе PCA – величина коэффициента корреляции, который показывает, насколько сильно ЭЭГ показатель связан с главной компонентой. Факторные нагрузки исходных ЭЭГ показателей для компоненты PC3 представлены в табл. 5. Как можно видеть из этих результатов, PC3 наиболее значимо зависит от мозговой активности в затылочно-теменных областях коры и в меньшей степени от активности лобных отделов коры. Такая корковая локализация характерна для функциональных процессов, связанных с контролем внимания при распознавании зрительных стимулов. Также заметно, что в эту компоненту наибольший вклад внесли показатели пре- и постмоторных пиков как для go-, так и для stop-эпизодов.

Задача 4. Предсказание вариабельности в генах-кандидатах на основе комплексных показателей ЭЭГ (решение обратной задачи)

При решении этой задачи для предсказания наличия или отсутствия однонуклеотидных вариаций в генах-кандидатах БДР по главным компонентам ЭЭГ (PC1–PC15) использовалась логистическая регрессия с L1-регуляризацией (Flach, 2016). В отличие от линейной регрессии, использующей количественные величины зависимых переменных, в логистической регрессии зависимая переменная является бинарной. В нашем случае модель логистической регрессии оценивала возможность предсказать вариабельность в генах-кандидатах БДР на основе данных ЭЭГ, что и представляло собой обратную задачу. В качестве признаков использовали первые 15 главных компонент (PC1–PC15), извлеченных из параметров ЭЭГ сигнала, а целевыми переменными выступали бинарные индикаторы наличия или отсутствия отклонения от референсного генома человека в пяти генах, ранее показавших значимые ассоциации с компонентами ЭЭГ, а именно *ADRA2B*, *WFS1*, *PENK*, *TF* и *HCRTR2*. Оценка качества моделей логистической регрессии проводилась с использованием метрики площади под ROC-кривой (AUC), вычисляемой по процедуре кросс-валидации (5 блоков, стратифицированное разбиение). Результаты оценки точности логистического анализа и стандартные отклонения по пятикратной кросс-валидации представле-

Таблица 4. Результаты линейной регрессии между главными компонентами ЭЭГ (PC1–PC3) и вариабельностью пяти генов-кандидатов БДР*

Ген, вариабельность которого влияла на мозговую активность	Уровень достоверности *p*-value
PC3	
WFS1	**0.0055**
TF	**0.0065**
HCRTR2	**0.0065**
PENK	**0.0065**
ADRA2B	**0.0258**
PC1	
ADRA2B	0.3297
TF	0.2844
HCRTR2	0.2844
WFS1	0.2876
PENK	0.2844
PC2	
TF	0.3109
HCRTR2	0.3109
WFS1	0.3028
PENK	0.3109
ADRA2B	0.3933

* Результаты упорядочены по уровню достоверности линейной регрессии.

Таблица 5. Факторные нагрузки для исходных показателей мозговой активности, включенных в компоненту PC3

ЭЭГ показатель	Факторная нагрузка* для PC3 ($p < 0.05$)
Затылочная кора, постмоторный пик, stop_эпизоды	0.24
Правая теменная кора, премоторный пик, stop_эпизоды	0.22
Левая теменная кора, премоторный пик, stop_эпизоды	0.21
Правая лобная кора, постмоторный пик, stop_эпизоды	0.19
Затылочная кора, постмоторный пик, go_эпизоды	0.19
Правая теменная кора, постмоторный пик, go_эпизоды	0.18
Медиальная теменная кора, постмоторный пик, go_эпизоды	0.17
Медиальная теменная кора, премоторный пик, stop_эпизоды	0.17
Левая лобная кора, постмоторный пик, stop_эпизоды	0.17
Левая теменная кора, постмоторный пик, go_эпизоды	0.16

* Факторная нагрузка – коэффициент корреляции между ЭЭГ показателем и интегральным показателем компоненты PC3.

Таблица 6. Результаты логистической регрессии для предсказания наличия мутаций по 15 главным компонентам ЭЭГ

Ген	AUC – показатель точности предсказания изменчивости	Стандартное отклонение
WFS1	0.78	0.15
PENK	0.76	0.18
TF	0.75	0.13
HCRTR2	0.75	0.13
ADRA2B	0.73	0.15

ны в табл. 6. Из этой таблицы можно видеть, что точность предсказания наличия бинарной изменчивости в генах-кандидатах БДР по ЭЭГ для трех рассмотренных генов лежит в пределах 0.73–0.78 со стандартным отклонением 0.13–0.18. Решение обратной задачи свидетельствует о том, что наличие бинарной вариабельности в генах-кандидатах БДР предсказывается с точностью 70–80 % на основании ЭЭГ данных, полученных при выполнении заданий в парадигме стоп-сигнал.

Таким образом, последовательное решение четырех задач нашего исследования позволило: 1) установить список генов-кандидатов БДР, изменчивость которых ассоциирована с показателями мозговой активности в условиях контроля поведения; 2) выявить комплексные характеристики мозговой активности, описывающие 80 % дисперсии ЭЭГ данных; 3) выявить интегральный показатель мозговой активности, наиболее достоверно зависящий от однонуклеотидной вариабельности в генах-кандидатах БДР; 4) провести решение обратной задачи по предсказанию вариабельности в генах-кандидатах БДР на основе ЭЭГ показателей.

Обсуждение

Основная проблема выявления генов-кандидатов для большинства психических заболеваний состоит в том, что поведенческие эффекты однонуклеотидных вариаций в каждом из выбранных генов достаточно слабые (Duncan et al., 2014). Депрессия – пример заболевания, для которого нет прямых и однозначных ассоциаций с конкретными генами (Halldorsdottir, Binder, 2017). Это позволяет предположить, что генетические причины депрессии вариативны для разных людей и не могут быть сведены к небольшому набору генов и мутаций в них.

Отсюда возникла задача перехода от анализа вклада отдельных генов или мутаций к исследованию взаимосвязанных комплексов генов, их продуктов и метаболитов. Такие комплексы генов получили название «генных сетей» (Колчанов и др., 2013). Генная сеть может включать от нескольких десятков до сотен генов, а также множество кодируемых ими белков и метаболитов. Ранее с использованием методов биоинформатики в работе (Ivanov et al., 2019) были реконструированы фрагменты генной сети, определяющей предрасположенность к БДР. Была собрана коллекция психометрических, нейрофизиологических и генетических данных, отражающая в совокупности результаты анализа однонуклеотидных полиморфизмов в 164 генетических локусах, включенных в генную сеть депрессии (Ivanov et al., 2022). Задачей предлагаемого исследования был поиск генов, ассоциированных не только с психометрическими, но и нейрофизиологическими характеристиками мозга, которые также могут рассматриваться как симптомы депрессии.

Способность контролировать поведение – одна из главных когнитивных функций человека, нарушение которой является симптомом многих заболеваний. Нами была проанализирована зависимость параметров вызванных потенциалов головного мозга человека от наличия однонуклеотидных изменений в генах-кандидатах БДР на объединенной выборке, включающей как здоровых людей, так и людей с диагностированным депрессивным расстройством. Согласно результатам, полученным в нашей работе, амплитуда пост-моторного пика в go-эпизодах парадигмы стоп-сигнал зависит от наличия бинарной изменчивости в пяти генах-кандидатах БДР: *ADRA2B*, *TF*, *HCRTR2*, *WFS1* и *PENK* (см. табл. 2).

Зависимости между генетической вариабельностью были выявлены как для нескольких локальных ЭЭГ показателей, отражающих активность отдельных областей коры мозга на краткосрочных этапах выполнения экспериментальных заданий, так и для интегрального показателя обобщенной мозговой активности, выделенного с помощью PCA и отражающего более общие характеристики состояния нервной системы (см. табл. 4). Интегральный показатель мозговой активности, для которого были выявлены достоверные влияния генетической вариабельности, отражает активность областей коры, вовлеченных в процессы восприятия зрительных сигналов и произвольного контроля над вниманием (см. табл. 5). Мы также показали, что интегральные показатели ЭЭГ могут с точностью 70–80 % служить предикторами однонуклеотидной вариабельности в генах-кандидата БДР при выполнении логистической регрессии (см. табл. 6), что говорит о возможности решать обратную задачу предсказания генетической изменчивости по нейрофизиологическим данным.

Из другого нашего исследования следует, что амплитуды вызванных потенциалов при выполнении заданий парадигмы стоп-сигнал позитивно коррелирует с выраженностью депрессивной симптоматики (Zorina et al., 2025). Таким образом, можно проследить связь генов, изменчивость которых ассоциирована как с депрессией на поведенческом уровне, так и с нейрофизиологическим маркером повышенного уровня депрессии. Из информации, представленной в статье (Ivanov et al., 2019), известно, что: а) ген *ADRA2B* кодирует альфа-2B адренорецептор, который является частью семейства G-белок-сопряженных рецепторов; б) ген *TF* кодирует белок трансферри; в) ген *HCRTR2* кодирует белки-рецепторы к окситоцину; г) *WFS1* кодирует белок вольфрамин; д) ген *PENK* кодирует предшественник энкефалина. Результаты нашего нового исследования свидетельствуют, что изменчивость в этих генах-кандидатах БДР ассоциирована с параметрами мозговой активности, отражающими способность человека к самоконтролю поведения, которая нарушается при БДР, что позволяет говорить о наличии комплексного генетико-нейрофизиологического показателя, связанного с риском депрессии.

Заключение

Проведенный анализ продемонстрировал наличие статистически значимых ассоциаций между полиморфизмами в генах *ADRA2B*, *TF*, *HCRTR2*, *WFS1* и *PENK* и характеристиками ЭЭГ сигнала, регистрируемого при выполнении заданий парадигмы стоп-сигнал. Метод главных компонент позволил сократить размерность данных и выявить наиболее информативные показатели интегральной мозговой активности. Построенные модели логистической регрессии показали, что на основе ЭЭГ параметров можно с определенной степенью точности предсказывать наличие однонуклеотидных замен в генах-кандидатах БДР. Полученные результаты могут быть использованы для оценки комплексной зависимости между генетическими и нейрофизиологическими показателями, ассоциированными с депрессией.

Ограничения. В рамках данного исследования мы не оценивали различия между группами клинических пациентов с депрессией и здоровыми участниками. Более детальное сравнение выявленных ассоциаций между нейрофизиологическими и молекулярно-биологическими маркерами депрессии является задачей более детального анализа, который мы планируем выполнить в дальнейших исследованиях.

Список литературы / References

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):833-850
[Kolchanov N.A., Ignatyeva E.V., Podkolodnaya O.A., Likhoshvai V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2013;17(4/2):833-850 (in Russian)]

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Arlington, VA: American Psychiatric Publishing, 2013;87-122. doi 10.1176/appi.books.9780890425596

Band G.P.H., van der Molen M.W., Logan G.D. Horse-race model simulations of the stop-signal procedure. *Acta Psychol.* 2003; 112(2):105-142. doi 10.1016/s0001-6918(02)00079-3

Benjamini Y., Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995; 57(1):289-300. doi 10.1111/j.2517-6161.1995.tb02031.x

Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a Genome-wide analysis. *Lancet*. 2013;381(9875):1371-1379. doi 10.1016/s0140-6736(12)62129-1

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neuroscience Methods.* 2004;134(1):9-21. doi 10.1016/j.jneumeth.2003.10.009

Duncan L.E., Pouastri A.R., Smoller J.W. Mind the gap: Why many geneticists and psychological scientists have discrepant views about gene–environment interaction (G×E) research. *Am Psychol.* 2014; 69(3):249-268. doi 10.1037/a0036320

Flach P.A. ROC Analysis. In: Sammut C., Webb G. (Eds). Encyclopedia of Machine Learning and Data Mining. Springer, 2016. doi 10.1007/978-1-4899-7502-7_739-1

Gewers F.L., Ferreira G.R., de Arruda H.F., Silva F.N., Comin C.H., Amancio D.R., da Costa L.F. Principal component analysis: A natural approach to data exploration. *ACM Comput Surv*. 2021;54(4):70. doi 10.1145/3447755

Haase J., Brown E. Integrating the monoamine, neurotrophin and cytokine hypotheses of depression: A central role for the serotonin transporter? *Pharmacol Ther*. 2015;147:1-11. doi 10.1016/j.pharmthera.2014.10.002

Halldorsdottir T., Binder E.B. Gene × environment interactions: From molecular mechanisms to behavior. *Annu Rev Psychol*. 2017;68: 215-241. doi 10.1146/annurev-psych-010416-044053

Ivanov R., Zamyatin V., Klimenko A., Matushkin Y., Savostyanov A., Lashin S. Reconstruction and analysis of gene networks of human neurotransmitter systems reveal genes with contentious manifestation for anxiety, depression, and intellectual disabilities. *Genes.* 2019;10(9):699. doi 10.3390/genes10090699

Ivanov R., Kazantsev F., Zavarzin E., Klimenko A., Milakhina N., Matushkin Y.G., Savostyanov A., Lashin S. ICBrainDB: An integrated database for finding associations between genetic factors and EEG markers of depressive disorders. *J Pers Med*. 2022;12(1):53. doi 10.3390/jpm12010053

Lopez-Calderon J., Luck S.J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci*. 2014;8: 213. doi 10.3389/fnhum.2014.00213

Northoff G. Gene, brains, and environment – genetic neuroimaging of depression. *Curr Opin Neurobiol*. 2013;23(1):133-142. doi 10.1016/j.conb.2012.08.004

Rokhlin V., Szlam A., Tygert M. A randomized algorithm for principal component analysis. *SIAM J Matrix Anal Appl*. 2010;31(3):1100-1124. doi 10.1137/080736417

Savostyanov A.N., Tsai A.C., Liou M., Levin E.A., Lee J.D., Yurganov A.V., Knyazev G.G. EEG-correlates of trait anxiety in the stop-signal paradigm. *Neurosci Lett*. 2009;449(2):112-116. doi 10.1016/j.neulet.2008.10.084

Shetty T., Kashyap H., Mehta U.M., Binu V.S. Executive function and emotion regulation in depressive and anxiety disorders: A cross-sectional study. *Indian J Psychol Med.* 2025. doi 10.1177/02537176251340586

Stone B., Desrochers P.C., Nateghi M., Chitadze L., Yang Y., Cestero G.I., Bouzid Z., … Bremner J.D., Inan O.T., Sameni R., Lynn S.K., Bracken B.K. Decoding depression: Event related potential dynamics and predictive neural signatures of depression severity *J Affect Disord.* 2025;391:119893. doi 10.1016/j.jad.2025.119893

Subasi A., Gursoy M.I. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Syst Appl*. 2010;37(12): 8659-8666. doi 10.1016/j.eswa.2010.06.065

Wang Z., Zellers S., Piirtola M., Aaltonen S., Salvatore J., Dick D., Kuhn S., Kaprio J. Gene-environment interplay in the relationship between the visibility of the environment and self-reported depression in early midlife: a Finnish twin cohort study. *medRxiv*. 2025. doi 10.1101/2025.05.23.25328215

Whitney D.G., Shapiro D.N., Peterson M.D., Warschausky S.A. Factors associated with depression and anxiety in children with intellectual disabilities. *J Intellect Disabil Res*. 2019;63(5):408-417. doi 10.1111/jir.12583

Willner P., Scheel-Krüger J., Belzung C. The neurobiology of depression and antidepressant action. *Neurosci Biobehav Rev.* 2013; 37(10 Pt. 1):2331-2371. doi 10.1016/j.neubiorev.2012.12.007

Wong M.L., Licinio J. Research and treatment approaches to depression. *Nat Rev Neurosci*. 2001;2(5):343-351. doi 10.1038/35072566

Zorina K.A., Kriveckiy A.A., Klemeshova D.I., Bocharov A.V., Karmanov V.S. Using machine learning methods to search for EEG and genetic markers of depressive disorder. In: IEEE 26th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation. 2025;1790-1793. doi 10.1109/EDM65517.2025.11096763

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила в редакцию 25.06.2025. После доработки 10.09.2025. Принята к публикации 11.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1137-1144

doi 10.18699/vjgb-25-118

Старение кожи связано с локальным дисбалансом в Т-клеточном иммунитете

К.С. Матвеева, С.К. Колмыков, Т.С. Соколова, Д.Р. Салимов, Д.В. Шевырев

Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Россия
dr.daniil25@mail.ru

Аннотация. Старение организма сопровождается накоплением поврежденных нефункциональных клеток, которые называют сенесцентными. Эти клетки находятся в состоянии ареста клеточного цикла, устойчивы к апоптозу, имеют нарушенный метаболизм, а также продуцируют широкий спектр провоспалительных факторов – цитокинов, хемокинов, протеаз, молекул адгезии и продуктов арахидонового каскада. Накопление таких клеток с возрастом связано с нарушением функций тканей, способствует хроническому воспалению (inflammaging) и развитию различных возраст-ассоциированных заболеваний. В свою очередь, элиминация сенесцентных клеток восстанавливает тканевые функции и позитивно сказывается на общем метаболизме. В норме сенесцентные клетки удаляются системой врожденного иммунитета, однако с возрастом эффективность этого процесса падает. При этом участие адаптивного иммунитета и роль Т-лимфоцитов в удалении сенесцентных клеток остаются неизученными. Целью исследования был поиск изменений в локальном Т-клеточном иммунитете, которые связаны с накоплением сенесцентных клеток в коже человека. Анализ проводился на открытых данных РНК секвенирования единичных клеток биоптатов кожи. Сенесцентный статус клеток оценивали при помощи алгоритма SenePy с применением смешанных гауссовских моделей. Было выявлено, что появление клеток с выраженными признаками сенесцентности в пределах ткани происходит неравномерно среди клеточных типов. Накопление этих клеток ассоциировано с изменением соотношения популяций $CD4^+$ и $CD8^+$ лимфоцитов, а также сопряжено с увеличением содержания регуляторных Т-лимфоцитов. В ходе функционального анализа обнаружено, что данные количественные изменения с возрастом сопровождаются более выраженной активацией регуляторных Т-лимфоцитов совместно с анергией и истощением $CD8^+$ лимфоцитов, тогда как функциональные изменения $CD4^+$ лимфоцитов имеют гетерогенный характер. Полученные результаты подчеркивают значение адаптивного иммунитета в поддержании тканевого гомеостаза и указывают на потенциальную дисфункцию эффекторных тканевых Т-лимфоцитов, которая возникает с возрастом. Понимание механизмов взаимодействия адаптивного иммунитета с сенесцентными клетками важно в контексте разработки сенолитических вакцин и других иммунологических подходов, направленных на усиление эндогенной элиминации сенесцентных клеток.

Ключевые слова: сенесцентность; адаптивный иммунитет; регуляторные Т-лимфоциты; транскриптом единичных клеток; старение; генетические сигнатуры; тканерезидентные Т-лимфоциты; элиминация сенесцентных клеток; кожа

Для цитирования: Матвеева К.С., Колмыков С.К., Соколова Т.С., Салимов Д.Р., Шевырев Д.В. Старение кожи связано с локальным дисбалансом в Т-клеточном иммунитете. *Вавиловский журнал генетики и селекции*. 2025;29(7):1137-1144. doi 10.18699/vjgb-25-118

Финансирование. Работа выполнена при поддержке Российского научного фонда, проект № 24-15-20003, https://rscf.ru/project/24-15-20003/

Senescent cell accumulation is associated with T-cell imbalance in the skin

K.S. Matveeva, S.K. Kolmykov, T.S. Sokolova, D.R. Salimov, D.V. Shevyrev

Sirius University of Science and Technology, Sirius Federal Territory, Krasnodar region, Russia
dr.daniil25@mail.ru

Abstract. Organismal aging is accompanied by the accumulation of senescent cells – damaged, non-functional cells that exhibit cell cycle arrest, resistance to apoptosis, metabolic dysfunction, and production of a wide range of pro-inflammatory substances. The age-related accumulation of these cells is associated with impaired tissue function, contributes to chronic inflammation (inflammaging), and promotes the development of various age-associated diseases. Conversely, the elimination of senescent cells restores tissue functions and positively affects overall metabolism. Under normal conditions, senescent cells are removed by the innate immune system; however, the efficiency of this process declines with age. The involvement of adaptive immunity and the role of T cells in

© Матвеева К.С., Колмыков С.К., Соколова Т.С., Салимов Д.Р., Шевырев Д.В., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

the clearance of senescent cells remain poorly understood. The aim of this study was to identify alterations in local T cell immunity associated with the accumulation of senescent cells in human skin. The analysis was performed on publicly available single-cell RNA-sequencing data from skin biopsies, and the senescent status was assessed using the SenePy algorithm with Gaussian mixture models. It was found that the emergence of senescent cells occurs heterogeneously across cell types within the tissue. The accumulation of these cells is associated with alterations in the $CD4^+$ to $CD8^+$ T cell ratio, as well as with an increased abundance of regulatory T cells. Functional analysis revealed that these quantitative age-related shifts were accompanied by more pronounced activation of regulatory T cells together with features of anergy and exhaustion in $CD8^+$ T cells, whereas functional changes in $CD4^+$ T cells were heterogeneous. These findings underscore the importance of adaptive immunity in maintaining tissue homeostasis and suggest potential age-related dysfunction of tissue-resident T cells. Understanding the mechanisms underlying the interaction between adaptive immunity and senescent cells is crucial for the development of senolytic vaccines and other immunological approaches aimed at enhancing endogenous elimination of senescent cells.
Key words: senescence; adaptive immunity; regulatory T cells; single-cell transcriptome; aging; genetic signatures; tissue-resident T cells; senescent cell elimination; skin

For citation: Matveeva K.S., Kolmykov S.K., Sokolova T.S., Salimov D.R., Shevyrev D.V. Senescent cell accumulation is associated with T-cell imbalance in the skin. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7): 1137-1144. doi 10.18699/vjgb-25-118

Введение

Сенесцентность – это состояние необратимой остановки клеточного цикла, которое возникает в ответ на различные стрессовые факторы, включающие репликативное истощение, повреждение ДНК, укорочение теломер, окислительный стресс и активацию онкогенов (Regulski, 2017; Di Micco et al., 2021). Сенесцентные клетки устойчивы к апоптозу, имеют сниженную функцию, нарушенный метаболизм и различные аберрации в машинерии контроля качества белков. Важная особенность таких клеток – постоянная продукция широкого спектра провоспалительных факторов (так называемый SASP фенотип – senescence-associated secretory phenotype), являющаяся одной из главных причин хронического вялотекущего воспаления, которое наблюдается при старении (inflammaging). Хотя сенесцентности отводится роль противоопухолевой защиты, длительное присутствие и накопление сенесцентных клеток приводит к нарушению функций тканей и органов, а также способствует возрастным и дегенеративным заболеваниям (Di Micco et al., 2021; Liao et al., 2021; Witham et al., 2023).

В исследованиях на животных моделях показано, что удаление сенесцентных клеток улучшает функцию тканей и метаболизм, увеличивает продолжительность жизни и замедляет развитие возрастных заболеваний (Yousefzadeh et al., 2019; Yang et al., 2023). В норме сенесцентные клетки удаляются иммунной системой, причем механизмы врожденного иммунитета в этом контексте изучены наиболее подробно. NK-клетки распознают их с помощью рецептора NKG2D, а затем элиминируют перфорин-гранзимовым путем и за счет секреции IFN-γ (Antonangeli et al., 2019). iNKT-клетки также могут устранять сенесцентные клетки при активации гликолипидными антигенами (Arora et al., 2021). Макрофаги, привлеченные факторами SASP, поглощают сенесцентные клетки при ремоделировании тканей (Song P. et al., 2020). С возрастом способность организма удалять сенесцентные клетки снижается, по-видимому, из-за старения самой иммунной системы. Это приводит к увеличению сенесцентной нагрузки, хроническому воспалению, дисфункции тканей и способствует развитию возрастных заболеваний (Song S. et al., 2020; Hense et al., 2024).

Несмотря на обширные исследования в области физиологической элиминации сенесцентных клеток, о роли адаптивного иммунитета в их удалении известно мало (Matveeva et al., 2024). Традиционные методы не позволяют адекватно воссоздать сложную трехмерную архитектуру тканей, где происходят ключевые взаимодействия между компонентами адаптивного иммунитета и сенесцентными клетками. Значительная часть Т-лимфоцитов находится в тканях, не выходит в циркуляцию и обладает функциональными свойствами, отличающимися от периферических Т-лимфоцитов (Li et al., 2025). В свою очередь, сенесцентные клетки локализованы преимущественно в паренхиме и строме органов, где могут формировать микросреду, влияющую на эффективность иммунного ответа (Zhang W. et al., 2024). В этом контексте данные транскриптома единичных клеток (scRNA-seq), полученные непосредственно из тканей, имеют особое значение, поскольку позволяют выявить сенесцентные клетки в разных клеточных типах, а также оценить основные характеристики адаптивного иммунитета по содержанию тех или иных субпопуляций и по их функциональной компетентности. Благодаря сохранению тканевого контекста данные scRNA-seq из различных тканей позволяют соотнести накопление сенесцентных клеток с изменениями (численными и функциональными) в различных популяциях Т-лимфоцитов – основными эффекторами адаптивного иммунитета (Kim S., Kim C., 2021).

В этом исследовании мы использовали данные scRNA-seq из открытых источников, чтобы оценить, связано ли возрастное накопление стареющих клеток в тканях с изменениями в пуле Т-лимфоцитов соответствующей ткани. В настоящее время принято считать, что старение разных типов клеток происходит и проявляется по-разному (Cohn et al., 2023). Кроме того, отсутствуют достоверные и уникальные маркеры сенесцентности, универсальные для всех сенесцентных клеток. Поэтому для определения сенесцентного статуса клеток мы применяли алгоритм SenePy, который вместо традиционного анализа дифференциальной экспрессии находит сетевые кластеры коэкспрессии генов, ассоциированные со старением (Sanborn et al., 2025). Старение кожи – это многогранный процесс, который связан с действием разнообразных повреждаю-

щих факторов в течение жизни. Накопление стареющих клеток, нарушение архитектуры дермального матрикса и деградация эластиновых волокон совместно с нарушением барьерной функции – это основные признаки старения кожи (Shin et al., 2025). В нашем исследовании выявление сенесцентных клеток в каждом клеточном типе кожи человека наряду с оценкой содержания различных субпопуляций Т-лимфоцитов позволило выявить значимые возрастные изменения тканевых Т-лимфоцитов, которые были ассоциированы с накоплением сенесцентных клеток.

Материалы и методы

Для проведения анализа были взяты данные scRNA-seq, опубликованные в базах GEO NCBI и GSA-Human. В автоматическом режиме были собраны данные секвенирования биоптатов кожи, полученных от здоровых доноров ($n = 32$) в возрасте от 18 до 76 лет (табл. S1 Приложения)[1].

Матрицы подсчетов UMI (Unique Molecular Identifiers) были получены из сырых прочтений с использованием программы 10x Genomics Cell Ranger v.9.0.1. Обработку матриц и метаданных проводили преимущественно в пакете Scanpy (Wolf et al., 2018). Предварительно из образцов были исключены клетки с низким числом прочтений (>5 MAD (median absolute deviations) и менее 500 UMI на клетку), низким числом экспрессированных генов (>5 MAD), а также высокой долей экспрессии митохондриальных генов (>4 MAD, или более 15 %). Выявление и фильтрация дуплетов осуществлялись с использованием пакета Scrublet (Wolock et al., 2019).

Далее образцы объединяли в единый набор данных и подготавливали к кластеризации, что включало в себя нормализацию (scanpy.pp.normalize_total(target_sum=1e4)), логарифмирование, масштабирование, снижение размерности с помощью анализа главных компонент и коррекцию батч-эффекта путем применения алгоритма Harmony (Korsunsky et al., 2019). Аннотацию клеточных типов проводили на лог-нормализованных данных с использованием инструмента CellTypist (Domínguez Conde et al., 2022), в основе которого лежат предсуществующие классификаторы логистической регрессии. Для аннотации была выбрана модель “Adult_Human_Skin” (Reynolds et al., 2021), содержащая информацию о различных типах дермальных, эпидермальных и иммунных клеток кожи человека. Для контроля и уточнения автоматической аннотации клетки кластеризовали с помощью алгоритма Лейдена. Кластеры сопоставляли с результатами автоматической аннотации и, при необходимости, корректировали ее. Последовательность обработки данных представлена на рис. 1. Контроль аннотации субпопуляций Т-лимфоцитов выполнялся особо тщательно. Для этого кластер, соответствующий Т-лимфоцитам, выделяли из общего набора данных, предобрабатывали начиная с исходных значений матрицы подсчетов UMI для более точного представления клеток в пространстве сниженной размерности и корректировали аннотацию при необходимости. Образцы с низкой представленностью конкретных клеточных типов исключались из анализа на соответствующих этапах исследования.

[1] Табл. S1–S4 и рис. S1 Приложения см. по адресу: https://vavilovj-icg.ru/download/pict-2025-29/appx42.zip

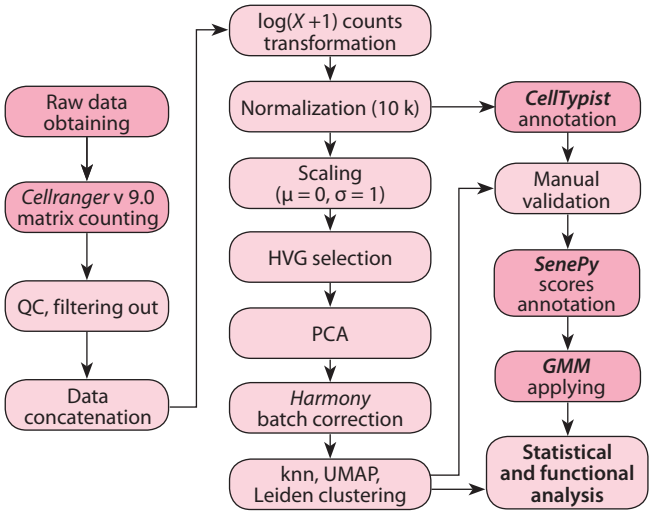


Рис. 1. Схематичное отображение последовательности обработки данных.

Известные маркеры клеточного старения специфичны для конкретных клеточных типов и плохо характеризуют состояние сенесцентности *in vivo*. Поэтому сенесцентный статус клеток оценивали с помощью опубликованного в 2025 г. алгоритма SenePy (Sanborn et al., 2025), позволяющего отличить специфические маркеры старения от генов, экспрессия которых повышается по причинам, не связанным с сенесцентностью. Поиск генов, потенциально связанных с возрастным накоплением сенесцентных клеток, в этом алгоритме происходит таким образом, чтобы ген присутствовал менее чем в 5 % молодых клеток, более чем в 1 % и менее чем в 20 % старых клеток. При этом доля клеток с данным геном в пожилом возрасте должна быть как минимум в 2.5 раза выше, чем в молодом, или абсолютный прирост доли клеток с геном (разница между старыми и молодыми донорами) должен быть больше 5 %. Это позволяет выявить набор генетических сигнатур сенесцентности, специфичных для каждого клеточного типа в ткани, что, в свою очередь, позволяет более точно определять сенесцентные клетки в образцах *ex vivo* по сравнению с традиционными методами. В результате каждой клетке присваивается числовая оценка *SenePy score*, отражающая степень соответствия ее профиля экспрессии конкретной сигнатуре сенесцентности.

После применения алгоритма SenePy модели смешанных гауссовских распределений (GMM) использовались для анализа распределения *SenePy score* клеток в каждом клеточном типе. В зависимости от распределения *SenePy score* в конкретном типе клеток модель включала в себя два или три компонента. Порог назначения клетке статуса сенесцентной определялся как значение между двумя последними компонентами GMM. Такой подход позволил количественно оценить долю клеток, имеющих выраженные признаки сенесцентности.

Корреляционный анализ проводили с помощью функции spearmanr() из модуля scipy.stats для вычисления коэффициента ранговой корреляции Спирмена и p-value. В условиях множественных сравнений применяли поправку Бонферрони.

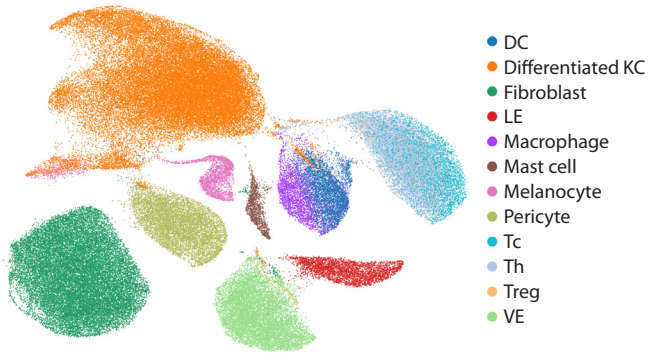


Рис. 2. Аннотация клеточных типов кожи человека с помощью инструмента CellTypist.

DC – дендритные клетки; KC – кератиноциты; LE – клетки лимфоидного эпителия; Tc – цитотоксические (классический фенотип – $CD3^+CD8^+$) Т-лимфоциты; Th – Т-хелперы (классический фенотип – $CD3^+CD4^+$); Treg – регуляторные Т-лимфоциты (классический фенотип – $CD3^+CD4^+FoxP3^+$); VE – эндотелиоциты.

Для анализа дифференциально экспрессированных генов (ДЭГ) в популяциях Т-лимфоцитов молодых и старых доноров применяли функцию rank_genes_groups() пакета Scanpy с использованием U-критерия Манна–Уитни. Дифференциально экспрессированными считали гены с FDR < 0.01, присутствующие в более 10 % клеток внутри группы интереса и менее 50 % клеток группы сравнения. Анализ функционального обогащения ДЭГ производили при помощи функции enricher() пакета clusterProfiler (Yu et al., 2021) на языке программирования R. Для анализа были взяты генные сигнатуры из коллекций C5 (ontology gene sets) и C7 (immunologic signature gene sets) базы данных MSigDB (Subramanian et al., 2005). Сигнатуры, для которых было показано достоверное обогащение, объединяли в функциональные группы вручную.

Результаты

Для выявления сенесцентных клеток в тканях кожи человека мы адаптировали и использовали недавно опубликованный алгоритм SenePy (Sanborn et al., 2025) с последующим применением GMM. Анализ проводился на основных популяциях клеток кожи, которые были аннотированы ранее (рис. 2).

В результате мы выявили значительное возрастное повышение содержания таких клеток в различных клеточных типах образцов кожи (рис. 3). В частности, мы наблюдали увеличение доли сенесцентных клеток с возрастом в тканерезидентных дендритных клетках, макрофагах, Т-лимфоцитах, кератиноцитах, меланоцитах, фибробластах, перицитах и эндотелиальных клетках. При этом ско-

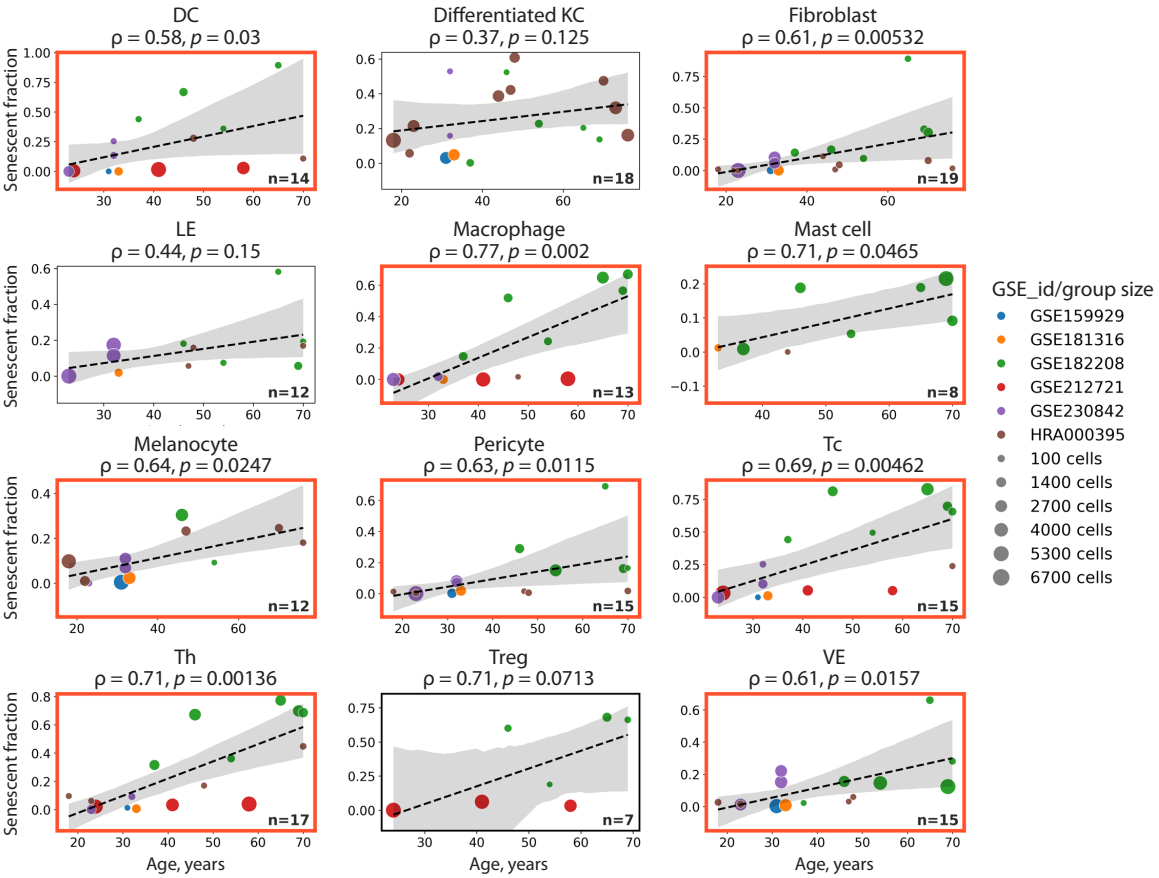


Рис. 3. Корреляции накопления сенесцентных клеток в разных клеточных типах кожи человека в зависимости от возраста.

Для каждого клеточного типа оценивалось число клеток; если оно было ниже 2SD от среднего числа по всем донорам, то образец исключался из анализа. Красными рамками выделены значимые корреляции. DC – дендритные клетки; KC – кератиноциты; LE – клетки лимфоидного эпителия; Tc – цитотоксические Т-лимфоциты; Th – Т-хелперы; Treg – регуляторные Т-лимфоциты; VE – эндотелиоциты.

рость накопления различалась между типами клеток, что отражает гетерогенность старения разных клеточных типов в пределах одной ткани.

Проведенный анализ выявил значительное накопление клеток с признаками сенесцентности в коже с возрастом, что согласуется с предыдущими данными о роли сенесцентности как одного из ключевых признаков старения тканей (Childs et al., 2015). Общее число сенесцентных клеток среди всех типов клеток тоже положительно коррелировало с возрастом (рис. 4), что указывает на прогрессирующее нарушение тканевого гомеостаза. Так как сенесцентные клетки не пролиферируют из-за устойчивого ареста клеточного цикла, их накопление с возрастом может быть связано со снижением эффективности механизмов, способствующих их элиминации.

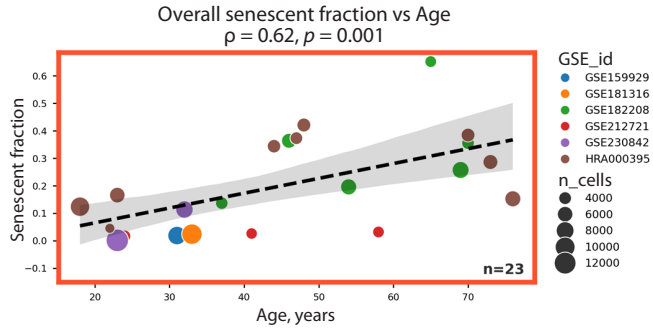


Рис. 4. Содержание сенесцентных клеток по всем клеточным типам в зависимости от возраста.

В связи с этим мы решили проанализировать, как меняется содержание основных субпопуляций Т-лимфоцитов в коже в зависимости от возраста. Корреляционный анализ не выявил достоверного изменения содержания трех анализируемых субпопуляций Т-лимфоцитов, а также основных иммунологических индексов с возрастом (рис. 5). Поскольку мы не выявили возрастных изменений среди тканевых Т-лимфоцитов, дальнейший интерес представляло изучение потенциальной связи между популяциями Т-лимфоцитов и накоплением сенесцентных клеток вне возрастного контекста.

Разные клеточные типы могут отличаться скоростью старения или иммуногенностью сенесцентных клеток, что может объяснить разницу в накоплении таких клеток с возрастом. Сначала мы решили проверить наличие каких-либо изменений в популяциях кожных Т-лимфоцитов, связанных с накоплением сенесцентных клеток. Мы оценили связь накопления сенесцентных клеток в том или ином клеточном типе с содержанием субпопуляций Т-лимфоцитов (рис. S1). Было обнаружено повышение содержания Т-лимфоцитов при накоплении сенесцентных перицитов, а также незначительные тенденции ($p < 0.07$) повышения числа регуляторных Т-лимфоцитов в зависимости от накопления сенесцентных клеток в некоторых клеточных типах.

На следующем этапе мы исследовали, как меняется содержание различных популяций Т-лимфоцитов в зависимости от общего числа сенесцентных клеток. Было выявлено значительное увеличение содержания Т-хелперов и регуляторных Т-лимфоцитов при повышении суммарного числа сенесцентных клеток из всех клеточных типов (рис. 6). Кроме того, следует подчеркнуть значимое повышение «тканевого иммунорегуляторного индекса», который отражает преобладание Т-хелперов над цитотоксическими лимфоцитами.

Таким образом, нами выявлена связь между накоплением сенесцентных клеток в коже человека и дисбалансом Т-клеточного иммунитета. Она выражалась повышением содержания регуляторных лимфоцитов и Т-хелперов, а также снижением числа цитотоксических лимфоцитов. При этом связи с возрастом обнаружено не было, что подчеркивает роль взаимодействия Т-клеточного иммунитета с сенесцентными клетками.

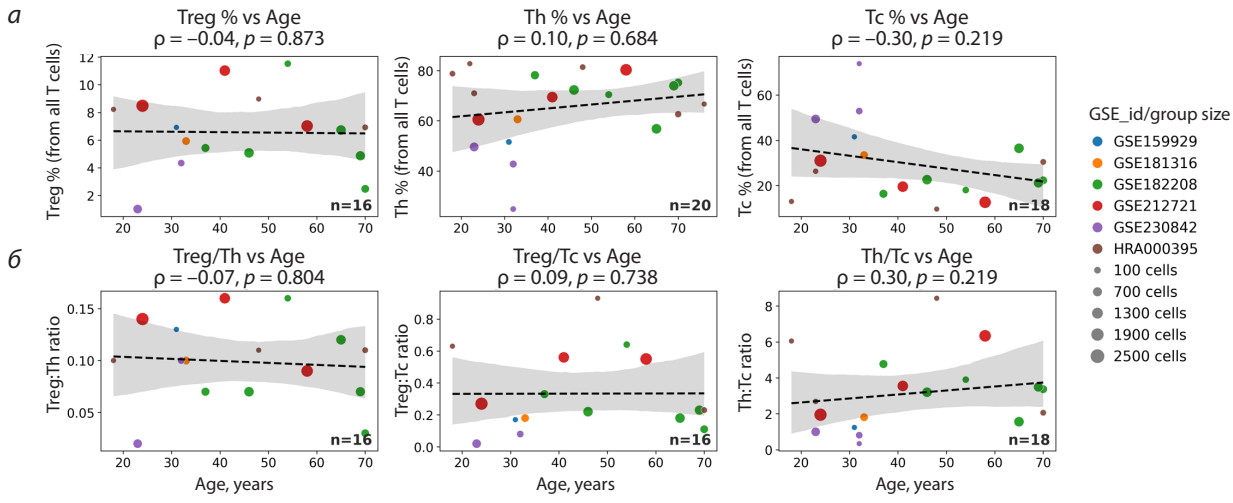


Рис. 5. Возрастное изменение содержания основных популяций Т-лимфоцитов (*а*), а также их соотношений (*б*).

Приведенные иммунологические индексы (соотношения Th/Tc, Treg/Tc и Treg/Th) позволяют с большей точностью и чувствительностью оценить состояние иммунитета в различных компрометирующих состояниях. На рисунке представлено содержание каждой популяции Т-лимфоцитов относительно их общего числа, что отражает баланс субпопуляций внутри всего пула кожных Т-лимфоцитов. Treg – регуляторные Т-лимфоциты; Th – Т-хелперы; Tc – цитотоксические Т-лимфоциты.

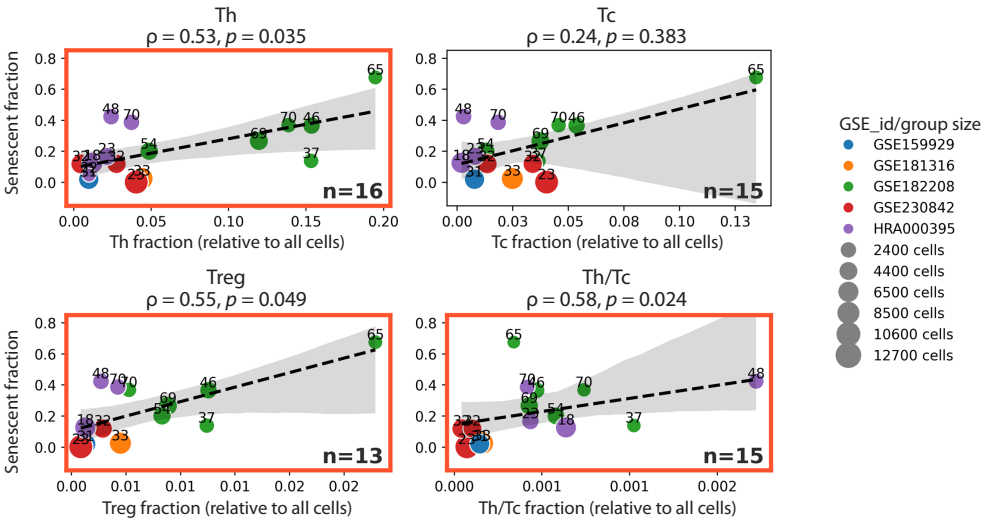


Рис. 6. Содержание основных популяций Т-лимфоцитов в зависимости от суммарного числа сенесцентных клеток.

Отражено содержание каждой популяции Т-лимфоцитов относительно общего числа всех клеток. Такой подход позволяет представить возрастные изменения в содержании субпопуляций Т-лимфоцитов относительно всех клеточных типов кожи, а не только изменения в пределах общей популяции тканевых Т-лимфоцитов. Это имеет физиологический смысл и лучше отражает изменения, связанные с накоплением клеток с признаками сенесцентности. Th – Т-хелперы; Tc – цитотоксические Т-лимфоциты; Treg – регуляторные Т-лимфоциты.

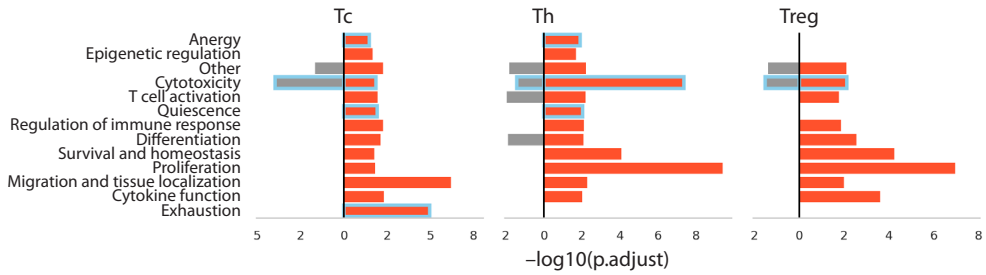


Рис. 7. Результаты анализа функционального обогащения генов, дифференциально экспрессированных у пожилых доноров в сравнении с молодыми, в пуле тканевых Т-лимфоцитов.

Оранжевые столбцы соответствуют обогащению функциональных путей генами с повышенной экспрессией, серые – генами с пониженной экспрессией. Ось *X* отображает логарифмическое преобразование *p*-value после FDR-коррекции, при котором большее значение соответствует более выраженному обогащению. Tc – цитотоксические Т-лимфоциты; Th – Т-хелперы; Treg – регуляторные Т-лимфоциты.

Выявленные нами на предыдущих этапах изменения в пуле тканевых Т-лимфоцитов подчеркивают вовлеченность адаптивного иммунитета в процессы старения тканей. Однако они не отражают функциональные особенности регуляторных Т-лимфоцитов, Т-хелперов и цитотоксических Т-лимфоцитов. Поэтому далее мы отобрали ДЭГ и провели анализ функционального обогащения по популяциям Т-лимфоцитов (см. Материалы и методы) для поиска обогащенных подмножеств генов в клетках пожилых доноров в сравнении с молодыми (рис. 7).

В результате анализа обнаружено статистически значимое обогащение путей, связанных с повышенной функциональной активностью Т-хелперов, тканевой адаптацией, дифференцировкой и ответом на цитокины, участвующие в гомеостазе этих клеток. Также встречалось обогащение путей, характерных для покоящихся (quiescence) клеток и клеток в состоянии анергии (отмечено голубыми рамками). При этом данная популяция Т-лимфоцитов не имела явных признаков истощения. Возрастные изменения, обнаруженные в популяции цитотоксических Т-лимфоцитов, были связаны с обогащением типичных для покоящихся клеток путей, состоянием анергии и истощением. Любопытно, что в этой популяции Т-лимфоцитов тоже наблюдалось значимое подавление путей, ассоциированных с их прямой функцией – цитотоксичностью. Напротив, регуляторные Т-лимфоциты не имели признаков покоящихся клеток (quiescence), анергии и истощения и наряду с Т-хелперами характеризовались повышенной функциональной и пролиферативной активностью. Также они имели обогащение генов, связанных с дифференцировкой и ответом на гомеостатические цитокины IL-2, IL-7 и IL-15, которые необходимы для тканевого поддержания регуляторных Т-лимфоцитов (табл. S2).

Таким образом, был проведен анализ функционального обогащения ДЭГ, идентифицированных с использованием данных scRNA-seq. Это позволило выявить признаки истощения и снижения функциональной активности в популяции цитотоксических лимфоцитов. При этом регу-

ляторные Т-лимфоциты отличались повышенной функциональной активностью и не имели признаков истощения или анергии. Изменения в популяции Т-хелперов носили более гетерогенный характер: наряду с повышением функциональной активности наблюдались признаки, типичные для анергии и покоящихся клеток.

Обсуждение

Накопление сенесцентных клеток – ключевой признак старения тканей, тесно связанный с развитием вялотекущего системного воспаления, которое является фактором риска развития возрастных заболеваний (Franceschi et al., 2018). С помощью современного алгоритма поиска сигнатур сенесцентных клеток мы показали, что в коже человека с возрастом увеличивается число клеток с признаками сенесцентности. При этом старение происходит неравномерно и затрагивает не все клеточные типы, что подчеркивает гетерогенность траекторий старения различных клеточных типов и многогранность процесса тканевого старения (Ge et al., 2022).

Иммунная система играет центральную роль в элиминации сенесцентных клеток. Провоспалительное микроокружение сенесцентных клеток привлекает макрофаги, нейтрофилы, NK и NKT-клетки (natural killer и natural killer T-cells соответственно), а также другие факторы неспецифического иммунитета, которые участвуют в распознавании и уничтожении сенесцентных клеток (Song P. et al., 2020). Хотя накапливается все больше сведений об участии Т-лимфоцитов в перечисленных выше процессах, роль адаптивного иммунитета остается неясной (Matveeva et al., 2024). Согласно нашим данным, накопление сенесцентных клеток ассоциировано с дисбалансом в локальном Т-клеточном иммунитете. Это свидетельствует об участии Т-лимфоцитов в процессе регуляции содержания сенесцентных клеток. Любопытно, что увеличение числа сенесцентных клеток ассоциировалось с увеличением доли регуляторных Т-лимфоцитов, а также с увеличением соотношения Th/Tc. Такой сдвиг указывает на формирование супрессорного микроокружения, которое может способствовать ускользанию сенесцентных клеток от элиминации иммунитетом (Zhang W. et al., 2024). Это предположение подкрепляется результатами анализа функционального состояния популяций Т-лимфоцитов пожилых доноров. Так, в популяции цитотоксических Т-лимфоцитов были выявлены признаки истощения и снижения функционального потенциала. В то же время регуляторные Т-лимфоциты и Т-хелперы имели признаки высокой функциональной активности и тканевой адаптации. Таким образом, в коже пожилых доноров наблюдались количественные и функциональные изменения в пуле тканевых Т-лимфоцитов, которые могут благоприятствовать формированию периферической толерантности к антигенам сенесцентных клеток. Это согласуется с гипотезой, согласно которой иммунная система с возрастом теряет способность распознавать и эффективно элиминировать сенесцентные клетки, что содействует их накоплению (Song P. et al., 2020).

Известно, что сенесцентные клетки могут не только формировать воспалительное микроокружение, но также подавлять функции эффекторных Т-лимфоцитов и ускользать от иммунологического надзора (Lorenzo et al., 2022). Так, продукция некоторых хемокинов привлекает регуляторные Т-лимфоциты, а опосредованная сенесцентными клетками дифференцировка моноцитов в макрофаги 2-го типа приводит к подавлению активации цитотоксических Т-лимфоцитов (Zhang X. et al., 2024). Более того, в стареющих клетках активируются эндогенные ретроэлементы, в частности LINE-1, что запускает IFN-γ-опосредованный ответ (Zhang X. et al., 2020). Этот механизм аналогичен противовирусному иммунному ответу и может способствовать формированию хронического воспаления, а также истощению цитотоксических лимфоцитов, которое мы также наблюдаем в группе пожилых доноров.

Таким образом, наши данные свидетельствуют о том, что пул Т-лимфоцитов кожи претерпевает значительную функциональную перестройку с возрастом. Снижение цитотоксической активности в сочетании с усилением функций регуляторных Т-лимфоцитов может формировать иммунологическую толерантность, поддерживающую накопление сенесцентных клеток и развитие так называемого воспалительного старения (inflammaging). По-видимому, это активный процесс формирования периферической толерантности к ассоциированным с сенесцентностью антигенам, при котором иммунная система теряет способность распознавать и элиминировать сенесцентные клетки. Выявленный дисбаланс в популяциях тканевых Т-лимфоцитов представляет потенциальную мишень для терапевтического вмешательства, направленного на восстановление иммунного надзора и элиминацию стареющих клеток.

Заключение

В настоящем исследовании с помощью методов биоинформатического анализа, проведенного на открытых данных секвенирования РНК единичных клеток, полученных из биоптатов кожи здоровых доноров, нами были выявлены ассоциированные со старением изменения в тканевом адаптивном иммунитете. Так, нами показано, что старение кожи, выражающееся накоплением сенесцентных клеток в разных клеточных типах, связано с изменением баланса между Т-хелперами и цитотоксическими лимфоцитами, а также с накоплением регуляторных Т-лимфоцитов. При этом функциональный анализ выявил общее снижение цитотоксической активности в тканевых Т-лимфоцитах на фоне активации регуляторных функций. По-видимому, такие изменения отражают компенсаторные изменения в пуле тканевых Т-лимфоцитов, связанные с накоплением и персистенцией сенесцентных клеток, формирующих постоянную воспалительную среду в ткани. В этом случае изменения в пуле тканевых Т-лимфоцитов направлены на формирование иммуносупрессорной среды, что, вероятно, вносит вклад в возрастное снижение эффективности элиминации сенесцентных клеток.

Исследование данных scRNA-seq предоставляет мощный инструмент для изучения взаимодействий иммунной системы с сенесцентными клетками на тканевом уровне. Сохранение тканевого контекста позволяет выявлять физиологические сигнатуры старения, а также анализировать наборы генов, ассоциированные с активацией или подавлением тех или иных компонентов иммунного ответа. Однако этот подход имеет ограничения. Утрата инфор-

мации о пространственной архитектуре ткани снижает возможность непосредственной оценки клеточных взаимодействий. Кроме того, различные технические искажения, связанные с пробоподготовкой и объединением данных из разных источников, требуют тщательной предобработки, коррекции батч-эффектов и нормализации, что может вносить значительную погрешность в полученные результаты. Поэтому в будущем для более глубокого понимания роли адаптивного иммунитета в элиминации сенесцентных клеток необходимо сочетать scRNA-seq данные с пространственной транскриптомикой, гистологическими методами и подходами, позволяющими определять антигенную специфичность Т- и В-клеток, а также изучать динамические изменения в репертуарах антигенраспознающих рецепторов.

Список литературы / References

Antonangeli F., Zingoni A., Santoni A., Soriani A. Senescent cells: living or dying is a matter of NK cells. *J Leukoc Biol*. 2019;105(6):1275-1283. doi 10.1002/jlb.mr0718-299r

Arora S., Thompson P.J., Wang Y., Bhattacharyya A., Apostolopoulou H., Hatano R., Naikawadi R.P., Shah A., Wolters P.J., Koliwad S., Bhattacharya M., Bhushan A. Invariant natural killer T cells coordinate removal of senescent cells. *Med*. 2021;2(8):938-950. doi 10.1016/j.medj.2021.04.014

Childs B.G., Durik M., Baker D.J., van Deursen J.M. Cellular senescence in aging and age-related disease: from mechanisms to therapy. *Nat Med*. 2015;21(12):1424-1435. doi 10.1038/nm.4000

Cohn R.L., Gasek N.S., Kuchel G.A., Xu M. The heterogeneity of cellular senescence: insights at the single-cell level. *Trends Cell Biol*. 2023;33(1):9-17. doi 10.1016/j.tcb.2022.04.011

Di Micco R., Krizhanovsky V., Baker D., d'Adda di Fagagna F. Cellular senescence in ageing: from mechanisms to therapeutic opportunities. *Nat Rev Mol Cell Biol*. 2021;22(2):75-95. doi 10.1038/s41580-020-00314-w

Domínguez Conde C., Xu C., Jarvis L.B., Rainbow D.B., Wells S.B., Gomes T., Howlett S.K., … Sims P.A., Farber D.L., Saeb-Parsy K., Jones J.L., Teichmann S.A. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*. 2022;376(6594):eabl5197. doi 10.1126/science.abl5197

Franceschi C., Garagnani P., Parini P., Giuliani C., Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol*. 2018;14(10):576-590. doi 10.1038/s41574-018-0059-4

Ge M.X., Yu Q., Li G.H., Yang L.Q., He Y., Li J., Kong Q.P. Multiple time-series expression trajectories imply dynamic functional changes during cellular senescence. *Comput Struct Biotechnol J*. 2022;20:4131-4137. doi 10.1016/j.csbj.2022.08.005

Hense J.D., Isola J.V.V., Garcia D.N., Magalhães L.S., Masternak M.M., Stout M.B., Schneider A. The role of cellular senescence in ovarian aging. *NPJ Aging*. 2024;10(1):35. doi 10.1038/s41514-024-00157-1

Kim S., Kim C. Transcriptomic analysis of cellular senescence: one step closer to senescence atlas. *Mol Cells*. 2021;44(3):136-145. doi 10.14348/molcells.2021.2239

Korsunsky I., Millard N., Fan J., Slowikowski K., Zhang F., Wei K., Baglaenko Y., Brenner M., Loh P.R., Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289-1296. doi 10.1038/s41592-019-0619-0

Li J., Xiao C., Li C., He J. Tissue-resident immune cells: from defining characteristics to roles in diseases. *Signal Transduct Target Ther*. 2025;10(1):12. doi 10.1038/s41392-024-02050-5

Liao Z., Yeo H.L., Wong S.W., Zhao Y. Cellular senescence: mechanisms and therapeutic potential. *Biomedicines*. 2021;9(12):1769. doi 10.3390/biomedicines9121769

Lorenzo E.C., Torrance B.L., Keilich S.R., Al-Naggar I., Harrison A., Xu M., Bartley J.M., Haynes L. Senescence-induced changes in CD4 T cell differentiation can be alleviated by treatment with senolytics. *Aging Cell*. 2022;21(1):e13525. doi 10.1111/acel.13525

Matveeva K., Vasilieva M., Minskaia E., Rybtsov S., Shevyrev D. T-cell immunity against senescence: potential role and perspectives. *Front Immunol*. 2024;15:1360109. doi 10.3389/fimmu.2024.1360109

Regulski M.J. Cellular senescence: what, why, and how. *Wounds*. 2017;29(6):168-174

Reynolds G., Vegh P., Fletcher J., Poyner E.F.M., Stephenson E., Goh I., Botting R.A., … Rajan N., Reynolds N.J., Teichmann S.A., Watt F.M., Haniffa M. Developmental cell programs are co-opted in inflammatory skin disease. *Science*. 2021;371(6527):eaba6500. doi 10.1126/science.aba6500

Sanborn M.A., Wang X., Gao S., Dai Y., Rehman J. Unveiling the cell-type-specific landscape of cellular senescence through single-cell transcriptomics using *SenePy*. *Nat Commun*. 2025;16:1884. doi 10.1038/s41467-025-57047-7

Shin S.H., Lee Y.H., Rho N.K., Park K.Y. Skin aging from mechanisms to interventions: focusing on dermal aging. *Front Physiol*. 2023;14:1195272. doi 10.3389/fphys.2023.1195272

Song P., An J., Zou M.-H. Immune clearance of senescent cells to combat ageing and chronic diseases. *Cells*. 2020;9(3):671. doi 10.3390/cells9030671

Song S., Kirkland J.L., Sun Y., Tchkonia T., Jiang J. Targeting senescent cells for a healthier aging: challenges and opportunities. *Adv Sci*. 2020;7(23):2002611. doi 10.1002/advs.202002611

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545-15550. doi 10.1073/pnas.0506580102

Witham M.D., Granic A., Miwa S., Passos J.F., Richardson G.D., Sayer A.A. New Horizons in cellular senescence for clinicians. *Age Ageing*. 2023;52(7):afad127. doi 10.1093/ageing/afad127

Wolf F.A., Angerer P., Theis F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. doi 10.1186/s13059-017-1382-0

Wolock S.L., Lopez R., Klein A.M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst*. 2019;8(4):281-291.e9. doi 10.1016/j.cels.2018.11.005

Yang D., Sun B., Li S., Wei W., Liu X., Cui X., Zhang X., Liu N., Yan L., Deng Y., Zhao X. NKG2D-CAR T cells eliminate senescent cells in aged mice and nonhuman primates. *Sci Transl Med*. 2023;15(709):eadd1951. doi 10.1126/scitranslmed.add1951

Yousefzadeh M.J., Melos K.I., Angelini L., Burd C.E., Robbins P.D., Niedernhofer L.J. Mouse models of accelerated cellular senescence. In: Demaria M. (Ed.) Cellular Senescence. Methods in Molecular Biology. Vol. 1896. New York: Humana Press, 2019;203-230. doi 10.1007/978-1-4939-8931-7_17

Yu G., Wang L.G., Han Y., He Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi 10.1089/omi.2011.0118

Zhang W., Zhang K., Shi J., Qiu H., Kan C., Ma Y., Hou N., Han F., Sun X. The impact of the senescent microenvironment on tumorigenesis: insights for cancer therapy. *Aging Cell*. 2024;23(5):e14182. doi 10.1111/acel.14182

Zhang X., Zhang R., Yu J. New understanding of the relevant role of LINE-1 retrotransposition in human disease and immune modulation. *Front Cell Dev Biol*. 2020;8:657. doi 10.3389/fcell.2020.00657

Zhang X., Ng Y.E., Chini L.C.S., Heeren A.A., White T.A., Li H., Huang H., Doolittle M.L., Khosla S., LeBrasseur N.K. Senescent skeletal muscle fibroadipogenic progenitors recruit and promote M2 polarization of macrophages. *Aging Cell*. 2024;23(3):e14069. doi 10.1111/acel.14069

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 30.07.2025. После доработки 25.09.2025. Принята к публикации 26.09.2025.

Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2025;29(7):1145-1154

doi 10.18699/vjgb-25-119

OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения

Е.В. Малюгин[1], Д.А. Афонников[2]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
evgeny.malyugin98@gmail.com

Аннотация. В последние годы быстрый рост объемов данных секвенирования обострил проблему функциональной аннотации белковых последовательностей, поскольку традиционные методы, основанные на гомологии, сталкиваются с ограничениями при работе с отдаленными гомологами, что затрудняет наиболее точное определение функций белков. В нашей работе представлен метод предсказания функций белков OrthoML2GO, который интегрирует поиск гомологичных последовательностей с помощью алгоритма USEARCH, анализ ортогрупп на базе OrthoDB 12-й версии и алгоритм машинного обучения (градиентный бустинг). Ключевая особенность подхода заключается в использовании информации об ортогруппах для учета эволюционного и функционального сходства белков и применения машинного обучения для дальнейшего уточнения терминов Gene Ontology (GO) для анализируемой последовательности. Для выбора оптимального алгоритма аннотации белков были поэтапно применены следующие подходы: метод *k* ближайших соседей (KNN); метод на основе аннотации ортогруппы, наиболее представленной у *k* ближайших гомологов (OG); метод верификации выявленных на предыдущем этапе терминов GO с помощью алгоритмов машинного обучения. Проведено сравнение точности предсказания терминов GO методом OrthoML2GO с программами аннотации Blast2GO и PANNZER2 на выборках последовательностей как отдельных организмов (человек, арабидопсис), так и на комбинированной выборке последовательностей, представленных разными таксонами. Результаты показали, что предложенный метод не уступает, а по некоторым показателям превосходит их по качеству предсказания функций белков, особенно на больших и разнородных выборках организмов, а наибольший прирост точности достигается за счет комбинации информации о ближайших гомологах и ортогруппах в сочетании с верификацией терминов методами машинного обучения. Разработанный подход демонстрирует высокую эффективность для крупномасштабной автоматической аннотации белков. Перспективы дальнейшего развития включают оптимизацию параметров моделей машинного обучения под конкретные биологические задачи и интеграцию дополнительных источников структурно-функциональной информации, что позволит еще больше повысить точность и универсальность метода. Кроме того, внедрение новых инструментов биоинформатики и расширение базы данных аннотированных белков будут способствовать дальнейшему совершенствованию предложенного подхода.

Ключевые слова: предсказание функций белка; генная онтология; гомология; ортогруппа; машинное обучение

Для цитирования: Малюгин Е.В., Афонников Д.А. OrthoML2GO: предсказание функций белков по гомологии с использованием ортогрупп и алгоритмов машинного обучения. *Вавиловский журнал генетики и селекции*. 2025;29(7):1145-1154. doi 10.18699/vjgb-25-119

Финансирование. Исследование поддержано Курчатовским геномным центром ИЦиГ СО РАН, соглашение с Министерством образования и науки Российской Федерации № 075-15-2019-1662 и бюджетным проектом № FWNR-2022-0020.

Благодарности. Исследование выполнено с использованием суперкомпьютерного комплекса ЦКП «Биоинформатика» ИЦиГ СО РАН.

OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning

E.V. Malyugin[1], D.A. Afonnikov[2]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
evgeny.malyugin98@gmail.com

Abstract. In recent years, the rapid growth of sequencing data has exacerbated the problem of functional annotation of protein sequences, as traditional homology-based methods face limitations when working with distant homologs, making it difficult to accurately determine protein functions. This paper introduces the OrthoML2GO method for

© Малюгин Е.В., Афонников Д.А., 2025
Контент доступен под лицензией Creative Commons Attribution 4.0

protein function prediction, which integrates homology searches using the USEARCH algorithm, orthogroup analysis based on OrthoDB version 12.0, and a machine learning algorithm (gradient boosting). A key feature of our approach is the use of orthogroup information to account for the evolutionary and functional similarity of proteins and the application of machine learning to refine the assigned GO terms for the target sequence. To select the optimal algorithm for protein annotation, the following approaches were applied sequentially: the k-nearest neighbors (KNN) method; a method based on the annotation of the orthogroup most represented in the k-nearest homologs (OG); a method of verifying the GO terms identified in the previous stage using machine learning algorithms. A comparison of the prediction accuracy of GO terms using the OrthoML2GO method with the Blast2GO and PANNZER2 annotation programs was performed on sequence samples from both individual organisms (humans, Arabidopsis) and a combined sample represented by different taxa. Our results demonstrate that the proposed method is comparable to, and by some evaluation metrics outperforms, these existing methods in terms of the quality of protein function prediction, especially on large and heterogeneous samples of organisms. The greatest performance improvement is achieved by combining information about the closest homologs and orthogroups with verification of terms using machine learning methods. Our approach demonstrates high performance for large-scale automatic protein annotation, and prospects for further development include optimizing machine learning model parameters for specific biological tasks and integrating additional sources of structural and functional information, which will further improve the method's accuracy and versatility. In addition, the introduction of new bioinformatics tools and the expansion of the annotated protein database will contribute to the further improvement of the proposed approach.
Key words: protein function prediction; gene ontology; homology; orthogroup; machine learning

For citation: Malyugin E.V., Afonnikov D.A. OrthoML2GO: homology-based protein function prediction using orthogroups and machine learning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(7):1145-1154. doi 10.18699/vjgb-25-119

Введение

Внедрение технологий секвенирования нового поколения (NGS) привело к экспоненциальному росту объемов данных о последовательностях ДНК, РНК и белков (Goodwin et al., 2016). Основными источниками этих данных служат масштабные и многочисленные проекты в области геномики, транскриптомики и протеомики (Cheng et al., 2018; Lewin et al., 2018). Однако значительная доля идентифицируемых в рамках таких проектов последовательностей остается с неизвестной функцией (Galperin, Koonin, 2010).

Экспертная аннотация генов требует большого количества времени для поиска информации о функциях генов в литературных источниках и, хотя является наиболее достоверной, использовать ее для огромного количества вновь предсказанных генов невозможно. Поэтому для большинства новых аминокислотных последовательностей (далее для краткости – последовательности) необходима разработка эффективных методов автоматического аннотирования, позволяющих определять их молекулярные функции, роль в клеточных процессах и клеточную локализацию. С учетом широкого использования для аннотации функций базы данных онтологии генов (Gene Ontology, GO) (Ashburner et al., 2000; Du Plessis et al., 2011; Gene Ontology Consortium, 2023) задача сводится к присвоению последовательности этих терминов в автоматическом режиме.

Большинство методов прогнозирования функций белков, основанных на анализе последовательности или трехмерной структуры, опирается на фундаментальный принцип: функция может быть предсказана на основе установления достоверного структурного или эволюционного сходства с белком, функция которого уже известна (Benso et al., 2013). Важнейшей задачей здесь является расшифровка взаимосвязи между обнаруженным сходством структур или последовательностей и фактическим уровнем функционального родства (Pearson, 2013). Среди этих методов к числу наиболее популярных относятся методы предсказания функции по гомологии, благодаря их универсальности и относительной простоте. Методы, основанные на гомологии, присваивают анализируемому белку термины GO на основе сходства его аминокислотной последовательности с первичными структурами белков с известной функцией. Другими словами, функция белка может быть расшифрована путем анализа его сходства с другими белками, для которых надежно определена функция (Eisenberg et al., 2000; Pearson, 2013).

Для сравнения аминокислотных последовательностей двух белков широко используется метод BLAST (Altschul et al., 1990). Однако в последнее время стали появляться новые инструменты поиска гомологичных последовательностей в базах данных, такие как GHOSTX (Suzuki et al., 2014), DIAMOND (Buchfink et al., 2015), MMseqs2 (Steinegger, Söding, 2017) и многие другие. Их характерная особенность – скорость обработки информации, на порядки превышающая скорость работы программ пакета BLAST, достигаемая в большинстве своем за счет более эффективной обработки совпавших фрагментов последовательности.

Концепция гомологии является основополагающей для получения выводов о процессах эволюционного формирования генов. В 1970 г. Уолтер Фитч (Fitch, 1970) предложил классифицировать гомологичные белки на ортологи и паралоги в соответствии с их происхождением. Ортологи происходят в процессе эволюционной дивергенцией генов, относящихся к различным таксонам в ходе их видообразования. Паралоги образуются за счет дупликаций генов. Предполагается, что ортологи сохраняют функцию гена-предшественника предкового вида, в то время как паралогичные гены после дупликаций могут изменить функцию (Fitch, 2000; Kuzniar et al., 2008; Altenhoff et al., 2019). С учетом огромной важности ортологов для сравнительной геномики и функциональной аннотации информация об ортологичных генах и их семействах накапливается в ряде специализированных баз данных, которые предоставляют собой важнейшие источники информации для идентификации и анализа ортологичных

Таблица 1. Список организмов, вошедших в исследование

Организм	Число последовательностей	Источник аннотации
Arabidopsis thaliana	27 655	TAIR (Reiser et al., 2024)
Homo sapiens	19 763	EBI Gene Ontology Annotation Database (Huntley et al., 2015)
Drosophila melanogaster	28 543 (с изоформами)	FlyBase (Öztürk-Çolak et al., 2024)
Solanum tuberosum	40 722 (с изоформами)	SpudDB (Hamilton et al., 2025a)
Danio rerio	33 428 (с изоформами)	ZFIN (Bradford et al., 2022)
Chlamydomonas reinhardtii	16 090	PhycoCosm (Grigoriev et al., 2021)
Oryza sativa	34 226 (с изоформами)	RGAP (Hamilton et al., 2025b)

групп генов (ортогрупп) (Jensen et al., 2008; Kriventseva et al., 2008). Отметим, что для решения задач предсказания функций генов успешно используются методы, включающие алгоритмы машинного обучения, которые позволяют увеличить точность по сравнению с более ранними подходами (Sanderson et al., 2023; Yuan et al., 2023).

В настоящей работе исследуется возможность предсказания функций белков на основе поиска гомологичных последовательностей, учета их ортологов и методов машинного обучения. Проведен поэтапный анализ влияния трех указанных факторов на точность предсказания терминов GO. Показано, что среди методов машинного обучения наибольшую точность предсказания демонстрирует алгоритм градиентного бустинга. На этой основе реализован алгоритм предсказания OrthoML2GO. Его точность сравнили с методами Blast2GO и PANNZER2. Показано, что предложенный метод обеспечивает более высокую точность, особенно на больших и разнородных выборках.

Материалы и методы

Данные об аминокислотных последовательностях. Перечень видов организмов и аминокислотных последовательностей, использованных в работе, дан в табл. 1. Он включает организмы с разной степенью полноты аннотации геномов (табл. S1[1]), представляющих разные таксоны как растений, так и животных: двудольных, однодольных, одноклеточных водорослей, позвоночных, членистоногих (см. табл. 1).

База данных OrthoDB как источник гомологичных последовательностей, аннотации и информации об ортологии. Мы использовали базу данных OrthoDB v 12.0 (https://www.orthodb.org/) (Tegenfeldt et al., 2025) в качестве источника гомологичных последовательностей, их аннотации терминами GO и данных по ортологии. База данных включает информацию по 5827 видам эукариот, 17 551 бактерий, 607 архей и 7962 вирусов. Она содержит более 162 млн последовательностей, классифицированных более чем на 10 млн групп ортологов. База данных также включает GO-аннотацию для части последовательностей и представляет собой удобный источник их классификации на ортологи и GO-аннотации. Кроме того, эта база данных предоставляет классификацию белковых последовательностей на ортологические семейства, для которых также представлена обобщенная аннотация функций белков в терминах GO.

Поиск гомологичных последовательностей. Поиск гомологов проводился с помощью алгоритма USEARCH v 11.0.667 (https://drive5.com/usearch/) (Edgar, 2010) командой usearch_local. Она производит поиск высокоидентичных совпадений на несколько порядков быстрее BLAST. В процессе поиска гомологичных последовательностей неизбежно оказывалось, что список гомологов включал и саму искомую последовательность. Для объективной оценки мы исключали из результатов поиска такие идентичные последовательности.

Общая схема аннотации последовательностей. Поиск терминов GO был реализован на языках bash Linux и R на вычислительных ресурсах центра коллективного пользования «Биоинформатика» ИЦиГ СО РАН. Было разработано три алгоритма аннотации функций белков на основе базы данных OrthoDB (рис. 1).

Слева (см. рис. 1, *а*) в большом овале схематически показана база OrthoDB v 12.0 (Tegenfeldt et al., 2025) с представителями ортологических групп (ортогрупп) OG1…OG3. Последовательности ортологичных семейств показаны на рисунке прямоугольниками одного цвета. Первый, базовый алгоритм предсказания последовательностей основан на поиске k ближайших гомологов и обозначается KNN. С помощью программы USEARCH в базе OrthoDB для анализируемой последовательности ведется поиск гомологичных последовательностей и ранжируются по уровню сходства. Они могут включать как представителей одной ортогруппы, так и других (показаны разными цветами). Анализируемой последовательности присваиваются термины GO k наиболее сходных последовательностей (см. рис. 1, *б*).

Второй метод основан на принципе ортологии и будет обозначен как OG. Для каждого из k ближайших гомологов анализируемой последовательности определяется его ортогруппа в БД OrthoDB. Ортогруппа, к которой принадлежит анализируемая последовательность, определяется методом голосования: это ортогруппа, имеющая наибольшую частоту встречаемости среди всех k ближайших гомологов (см. рис. 1, *в*). Термины GO для последовательностей из этой ортогруппы присваиваются анализируемой последовательности (см. рис. 1, *г*).

Третий подход, обозначенный как KNN+OG (см. рис. 1, *д*), заключается в объединении терминов GO, полученных с помощью предыдущих алгоритмов – KNN

[1] Табл. S1–S12 Приложения см. по адресу:
https://vavilovj-icg.ru/download/pict-2025-29/appx43.pdf

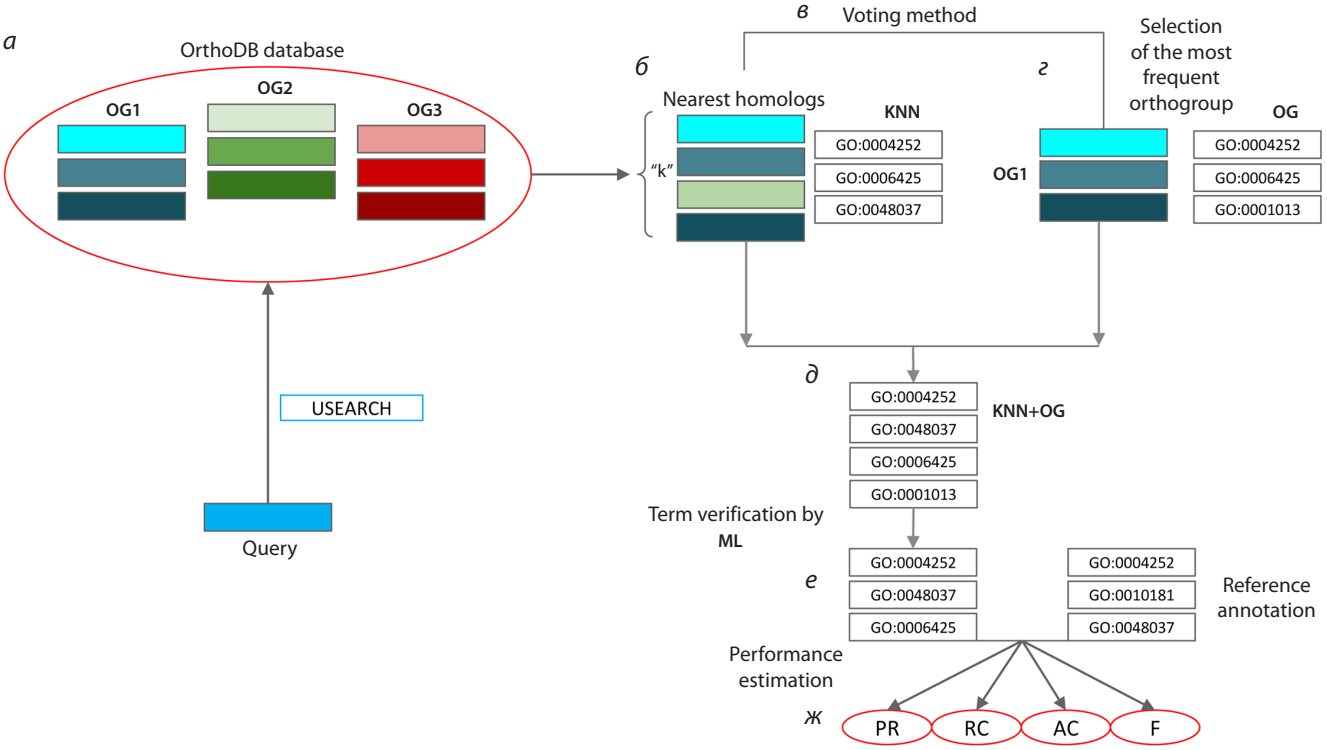


Рис. 1. Общая схема аннотации последовательностей и ее оценки. Последовательности, принадлежащие одной ортогруппе, представлены разными оттенками одного цвета: синего, зеленого или красного.

а – база данных OrthoDB с ортогруппами; *б* – присвоение терминов GO от *k* ближайших гомологов (метод KNN); *в* – определение наиболее представленной ортогруппы методом голосования; *г* – присвоение терминов GO, ассоциированных с выбранной ортогруппой (метод OG); *д* – объединение терминов GO, полученных методами KNN и OG (метод KNN+OG); *е* – верификация объединенного списка терминов с помощью машинного обучения; *ж* – сравнение предсказанных терминов с референсной аннотацией и расчет метрик.

и OG – для анализируемой последовательности (см. рис. 1, *е*). Этот список терминов GO, который сравнивается с референсной аннотацией (истинной) при помощи таких мер, как: precision, recall (sensitivity), accuracy и F-score (F-measure), которая была результирующей (см. рис. 1, *ж* и раздел «Верификация терминов методами машинного обучения»).

Методы аннотации анализируемой последовательности терминами GO. ***Метод k ближайших гомологов*** **(KNN).** *k* ближайших гомологов по уровню сходства определяются в результате поиска по БД OrthoDB программой USEARCH с параметрами: identity (идентичность аминокислотных последовательностей) = 50 %, coverage (покрытие анализируемой последовательности найденным гомологом) = 70 %, *e*-value (статистическая значимость найденного совпадения) = 10^{-6}, что оправдано целью уменьшить ложноположительные находки на этапе поиска гомологов. Анализируемой последовательности присваиваются термины GO *k* наиболее сходных последовательностей, имеющихся в БД OrthoDB. Значения параметра *k* могут варьировать (Kharsikar et al., 2007; Dongardive, Abraham, 2016), поэтому в нашем случае из интервала *k* = 1–30 с шагом 5 мы выбирали значение, при котором точность идентификации терминов с использованием аннотации OrthoDB (табл. S4–S9) была наилучшей.

Использование ортологических групп **(OG).** В этом методе для каждого из *k* ближайших гомологов анализируемой последовательности, определенного методом KNN, согласно аннотации OrthoDB, выбирается ортологическая группа, соответствующая наиболее древнему предковому таксону. После этого определяется наиболее часто встречающаяся среди *k* ортогруппа, которая присваивается анализируемой последовательности. Термины аннотации GO для последовательностей из этой ортогруппы в БД OrthoDB присваиваются анализируемой последовательности. Метод KNN+OG основан на объединении терминов GO (с исключением повторов), полученных отдельно методами KNN и OG, описанными выше.

Верификация терминов методами машинного обучения. Для уточнения списка предсказанных терминов GO на третьем этапе анализа (см. рис. 1, *е*) использовали три алгоритма машинного обучения (ML – machine learning): логистическую регрессию (LR), метод градиентного бустинга (XGB) и метод случайного леса (RF). Отметим, что этот этап не позволяет добавить к аннотации новые термины. Он лишь исключает термины, для которых ряд параметров сходства анализируемой последовательности и гомологов не соответствует заданным критериям.

Метод логистической регрессии (LR – logistic regression) реализован во встроенном пакете stats (R Core Team, 2013), функция glm (family = binomial). Логистическая регрессия предсказывает вероятность принадлежности объекта к классу, например «спам» или «неспам». Он предсказывает вероятность принадлежности объекта к классу на основе взвешенной суммы признаков и пропускает ее через логистическую (сигмоидную) функцию, которая

нормализует результат в число (вероятность) от 0 до 1. Градиентный бустинг (XGB – eXtreme Gradient Boosting) использован в варианте, реализованном в пакете xgboost (Chen, Guestrin, 2016), функция xgb.train. Метод случайного леса (RF – random forest) применен в версии из пакета randomForest (Liaw, Wiener, 2002), функция randomForest. Как градиентный бустинг, так и метод случайного леса относятся к ансамблевым алгоритмам, основанным на деревьях решений (decision trees). Это означает, что итоговое предсказание – результат совокупной работы множества отдельных деревьев решений. Параметры алгоритмов градиентного бустинга и случайного леса указаны в табл. S12.

Параметры для моделей подбирали в процессе обучения, и в каждом методе их набор был одинаков для всех терминов GO, анализируемых последовательностей и их гомологов. Это термины, отражающие уровень сходства, аминокислотный состав и частоту встречаемости терминов GO (табл. S2). Если термин GO у гомолога присутствовал в аннотации анализируемой последовательности в обучающей выборке, значение функции предсказания в методе машинного обучения было равно 1, в противном случае – 0.

Для оценки точности методов машинного обучения использовали аминокислотные последовательности белков *Arabidopsis thaliana* и *Homo sapiens* (см. табл. 1). Набор этих последовательностей каждого из этих двух видов был разделен на две части: 80 % для обучения и 20 % для тестирования. Дополнительно была сформирована комбинированная выборка белков для организмов, представленных в табл. 1: из комбинированной выборки случайным образом для обучения были отобраны 50 000 последовательностей, а 20 000, не совпадающих с ними, – для тестирования моделей машинного обучения (табл. S3).

Метрики качества. Оценку точности аннотации проводили на языке R с использованием пакета dplyr (Wickham et al., 2025). Для этого были сформированы два списка: а) референсный список с аминокислотными последовательностями, аннотированными терминами GO из БД по модельным организмам, подробнее см. в табл. S1, и б) список, полученный путем функциональной аннотации с использованием различных методов аннотирования (см. рис. 1). Для оценки точности аннотации, полученной каждым из описанных выше методов, проводилось их сравнение с референсной аннотацией. В дальнейшем под True Positive (TP) мы понимаем термины GO, присутствующие в обоих списках; к False Positive (FP) относятся термины, присутствующие в списке предсказанной аннотации, но отсутствующие в референсном (истинном) списке; к False Negative (FN) относятся термины, присутствующие в референсном списке, но отсутствующие в списке предсказанной аннотации.

Для оценки аннотации белков использованы следующие метрики: Precision (PR), Recall (RC), Accuracy (AC), а также метрика F-score, которая была результирующей.

Precision (PR) – доля истинно положительных предсказаний среди всех положительных предсказаний метода:

$$\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100. \quad (1)$$

Recall (RC) – доля истинно положительных предсказаний среди всех истинных терминов в референсной аннотации:

$$\mathrm{RC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100. \quad (2)$$

Accuracy (AC) представляет собой среднее арифметическое между Precision и Recall:

$$\mathrm{AC} = \frac{\mathrm{PR}+\mathrm{RC}}{2} \times 100. \quad (3)$$

F-score представляет собой гармоническое среднее между Precision и Recall. Эта метрика стремится к нулю, если значение Precision или Recall стремится к нулю:

$$F1 = 2\,\frac{\mathrm{PR}\times\mathrm{RC}}{\mathrm{PR}+\mathrm{RC}} \times 100. \quad (4)$$

Поскольку алгоритмы машинного обучения (LR, XGB, RF) оценивают вероятность принадлежности термина GO для анализируемой последовательности, а не бинарное решение, необходимо выбрать порог отсечения (t), по которому термин будет считаться предсказанным. Для учета несбалансированности данных и выбора оптимального порога, не зависящего от его конкретного значения, рассчитывалась метрика $F_{\max}$ для порога отсечения $t \in (0; 1)$ с шагом в 0.1. Термин GO считался верно предсказанным (положительным классом), если его прогнозируемая вероятность превышала порог t. $F_{\max}$ определяется как максимальное значение F-score(t) по всем порогам:

$$F_{\max} = \max\left\{2\,\frac{\mathrm{PR}(t)\times\mathrm{RC}(t)}{\mathrm{PR}(t)+\mathrm{RC}(t)}\right\} \times 100. \quad (5)$$

В задачах прогнозирования GO-терминов, где распределение терминов по частоте встречаемости крайне не сбалансировано (некоторые термины очень часты, другие крайне редки), а классификация является многометочной (одному белку может соответствовать множество терминов), часто используется метрика $F_{\max}$. Она рассчитывается для всего набора предсказаний путем варьирования порога отсечения (t), начиная с которого предсказанный ML-моделью термин считается положительным. $F_{\max}$ показывает максимальное качество, которого может достичь модель в идеальном случае выбора порога. В отличие от $F1$, которая рассчитывается для фиксированного порога, $F_{\max}$ оценивает качество ранжирования терминов по вероятности.

Сравнение с другими методами. Для верификации разработанного нами метода OrthoML2GO сравнили его с методами Blast2GO (https://www.blast2go.com/) (Conesa et al., 2005) и PANNZER2 (http://ekhidna2.biocenter.helsinki.fi/sanspanz/) (Törönen et al., 2018). Поиск гомологов BLAST запускали на вычислительном комплексе центра коллективного пользования «Биоинформатика» ИЦиГ СО РАН. Параметры запуска Blast2GO и PANNZER2 были выбраны по умолчанию.

Результаты и обсуждение

Влияние информации об ортогруппах на точность предсказания терминов GO

Для оценки влияния информации об ортогруппах на точность предсказания функции было проведено сравнение

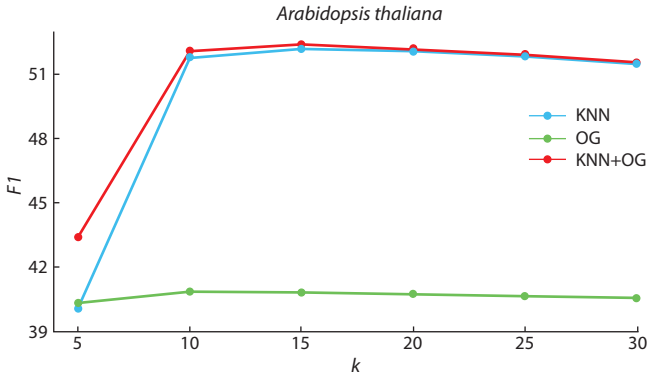


Рис. 2. Зависимость меры *F1* на белках *Arabidopsis thaliana* от параметра *k* (число ближайших гомологов) для трех вариантов аннотации.

По оси *X* отложены значения *k*, по оси *Y* – значения меры *F1*(%). Линии разного цвета соответствуют разным алгоритмам аннотации нашего метода: KNN – синий, OG – зеленый, KNN+OG – красный.

F1 меры на трех методах аннотации терминами GO тремя алгоритмами (KNN, OG и KNN+OG) в зависимости от числа ближайших гомологов для последовательностей *Arabidopsis thaliana* (рис. 2).

Из рисунка 2 видно, что для всех трех вариантов аннотации наблюдается зависимость меры *F1* от числа *k*. Однако характер этих зависимостей различен: OG демонстрирует самую низкую точность (*F1* < 41 %). Для метода OG, как и для остальных методов, наблюдается максимум при *k* = 15. При этом дальнейшее увеличение параметра *k* приводит к монотонному, хотя и незначительному снижению меры *F1*. Мы полагаем, что для наиболее точного предсказания достаточно определить корректную ортологическую группу белка, которая идентифицируется уже при малых значениях *k*. Дальнейшее увеличение *k* приводит лишь к зашумлению предсказания из-за увеличения ложноположительных терминов GO от ортогрупп, к которым рассматриваемый белок на самом деле не относится.

Метод KNN демонстрирует выраженную зависимость точности от параметра *k*. При малых значениях (*k* = 5) мера *F1* самая низкая (~40 %) и ниже метода OG и KNN+OG, что, вероятно, связано с недостаточным количеством гомологов для надежного статистического вывода и высокой чувствительностью к шуму и возможным ошибкам аннотации отдельных последовательностей. При увеличении *k* до 15 наблюдается рост *F1* до максимального значения (~52 %), однако дальнейшее увеличение *k* приводит к постепенному снижению точности, поскольку в выборку начинают попадать отдаленные гомологи, несущие нерелевантную для целевой последовательности функциональную информацию (ложноположительные термины GO).

Объединение методов KNN и OG (KNN+OG) приводит к увеличению *F1*-меры при всех значениях параметра *k*, причем наибольший прирост (более 3 % по абсолютному значению) наблюдается именно при *k* = 5. Вероятно, этот эффект можно объяснить тем, что при малом *k* список гомологов может быть неустойчивым и статистически ненадежным. Добавление информации об ортогруппе, которая агрегирует данные о функции целой группы эволюционно родственных генов, стабилизирует предсказание и позволяет компенсировать недостаточность данных от малого числа ближайших соседей.

Стоит отметить, что значение *F1*-меры в диапазоне 40–52 % является конкурентоспособным результатом для задачи предсказания функций белков, что подтверждается сравнением с другими популярными методами (см. раздел «Сравнение точности методов KNN, KNN+OG и OrthoML2GO с инструментами Blast2GO и PANNZER2»). Это связано со сложной природой задачи: во-первых, как было показано ранее, аннотация GO является множественной, т. е. одному белку соответствует множество терминов, и предсказание считается верным, только если найдены все правильные термины и не добавлены лишние. Во-вторых, распределение терминов GO крайне не сбалансировано: некоторые термины очень часты, другие крайне редки, что дополнительно усложняет достижение высокой точности. Таким образом, абсолютное значение *F1*-меры следует интерпретировать в контексте сложности задачи и в сравнении с альтернативными подходами.

Для других организмов результаты применения трех подходов показаны в приложении (см. табл. S4–S9). Объединение методов KNN и OG (KNN+OG) позволяет получить интегральное предсказание, которое демонстрирует наибольший выигрыш в точности при малых значениях параметра *k* для всех организмов, кроме *Chlamydomonas reinhardtii*. Например, для белков *Danio rerio* при *k* = 5 метод KNN+OG превосходит базовый KNN более чем на 13 % по абсолютному значению *F1*-меры (74.66 против 61.37 %). Это объясняется тем, что при малом *k* список гомологов может быть статистически ненадежным и чувствительным к шуму в аннотациях отдельных последовательностей. Добавление информации об ортогруппе, которая агрегирует данные о функции целой группы эволюционно родственных генов, стабилизирует предсказание и позволяет компенсировать недостаточность данных от малого числа ближайших соседей. Следовательно, гибридный подход KNN+OG не только показывает лучший результат в пике (при *k* = 15), но и существенно снижает зависимость точности предсказания от параметра *k*, делая метод более устойчивым.

Таким образом, объединение вариантов KNN и OG (KNN+OG) позволяет получить интегральное предсказание, давая лучшую оценку по сравнению с каждым из них, для всех значений параметра *k* большинства организмов и для машинного обучения будет использован именно он.

Верификация терминов GO различными алгоритмами машинного обучения

Для верификации ложноположительных терминов GO, полученных на предыдущем этапе, использованы такие алгоритмы машинного обучения, как логистическая регрессия (LR), градиентный бустинг (XGB) и случайного леса (RF) (см. раздел «Верификация терминов методами машинного обучения»). Сравнение точности методов машинного обучения с помощью меры F_{max} (см. раздел «Сравнение с другими методами») на тестовых данных *Arabidopsis thaliana*, *Homo sapiens* и комбинированной выборки из 20000 последовательностей разных организмов представлено в табл. 2.

Логистическая регрессия демонстрирует значительно более низкие значения F_{max} по сравнению с методами градиентного бустинга и случайного леса, причем разница может достигать более 25 %. Вероятно, это связано с тем, что ансамблевые методы (XGB и RF), в отличие от линейной модели LR, способны улавливать сложные нелинейные взаимоотношения между признаками. Кроме того, эти методы более устойчивы к шуму в данных за счет процедур бэггинга (RF) и бустинга (XGB), которые усредняют прогнозы множества отдельных деревьев решений, уменьшая влияние выбросов и некорректных аннотаций отдельных белков. Градиентный бустинг показывает лучшие результаты на последовательностях арабидопсиса и общей выборке всех организмов, но лишь незначительно уступает методу случайного леса на белках человека – разница в величине F_{max} составляет всего 0.1 %. Таким образом, для финальной версии метода OrthoML2GO выбран метод машинного обучения градиентный бустинг (XGB), как показавший лучшие результаты на тестовых выборках.

Сравнение точности методов KNN, KNN+OG и OrthoML2GO с инструментами Blast2GO и PANNZER2

Для всесторонней оценки эффективности разработанного метода было проведено сравнение его точности с двумя широко используемыми инструментами автоматической функциональной аннотации – Blast2GO и PANNZER2. Сравнение выполняли на трех тестовых наборах данных: отдельные протеомы *Arabidopsis thaliana* и *Homo sapiens*, а также комбинированная выборка, включающая последовательности всех организмов, перечисленных в табл. 1. В качестве результирующей метрики для методов, не использующих машинное обучение (KNN, KNN+OG, Blast2GO), применялась *F1*-мера, в то время как для OrthoML2GO и PANNZER2, выдающих вероятностную оценку, использовалась метрика F_{max}, позволяющая оценить максимально достижимое качество модели при идеальном выборе порога отсечения (табл. 3).

Анализ результатов демонстрирует, что разработанный метод OrthoML2GO, интегрирующий поиск гомологов, анализ ортогрупп и верификацию терминов GO с помощью градиентного бустинга, показал статистически значимое преимущество по точности над всеми сравниваемыми методами на всех тестовых выборках. Можно отметить, что разработанный метод OrthoML2GO на основе алгоритмов *k* ближайших соседей, ортогрупп и градиентного бустинга продемонстрировал статистически значимое преимущество по точности над всеми другими рассматриваемыми методами на всех тестовых выборках. Так, для

Таблица 2. Сравнение меры F_{max} на тестовых данных для разных алгоритмов машинного обучения, %

Организм	LR	XGB	RF
Arabidopsis thaliana	53.20	**68.95**	66.86
Homo sapiens	71.92	**83.92**	84.02
Комбинированная выборка	52.25	**79.55**	78.32

Arabidopsis thaliana OrthoML2GO достиг F_{max} = 68.95 %, что на 18.21 % превысило результат PANNZER2 и на 14.65 % – Blast2GO (*F1* = 54.30 %). На белках человека, по сравнению с PANNZER2, OrthoML2GO оказался существенно лучше – 83.92 против 75.14 %, а для метода Blast2GO значения *F1* = 54.95 %. На общей выборке всех организмов наблюдалось улучшение показателя F меры более чем на 30 % по сравнению со всеми другими методами.

Следует отметить, что гибридный подход KNN+OG, лежащий в основе OrthoML2GO, уже демонстрирует небольшое, но стабильное улучшение по сравнению с базовым KNN на всех выборках, что подтверждает полезность интеграции информации об ортогруппах. Однако основной выигрыш в точности обеспечивает этап верификации с помощью градиентного бустинга (XGB), который эффективно фильтрует ложноположительные предсказания, возникающие из-за шума в аннотации.

Одним из возможных факторов успеха метода OrthoML2GO является интеграция эволюционной информации из гомологов и ортогрупп БД OrthoDB с последующей верификацией терминов GO методом градиентного бустинга. В отличие от методов PANNZER2 и Blast2GO, наш метод учитывает информацию об ортогруппах и верифицирует термины GO с помощью ансамблей деревьев решений, адаптивно отбирая наиболее информативные признаки. В итоге это позволило снизить долю ложноположительных аннотаций и повысить точность от 8 % (на белковых последовательностях человека) до 30 % (на комбинированной выборке) по сравнению с аналогами.

Важно отметить, что в основе наших моделей машинного обучения лежит выборка последовательностей из БД OrthoDB. В отличие от этого, методы Blast2GO и PANNZER2 используют для своей работы более широкие наборы данных из базы UniProt, включающие как последовательности, так и аннотации. Поэтому нельзя исключить смещения в оценках точности, присущих двум этим методам в силу указанных обстоятельств.

Таблица 3. Сравнение методов KNN, KNN+OG, OrthoML2GO (XGB), PANNZER2 и Blast2GO на трех наборах данных, %

Набор данных	KNN*	KNN+OG*	OrthoML2GO (XGB)	PANNZER2	Blast2GO*
Arabidopsis thaliana	51.54	51.68	68.95	50.74	54.30
Homo sapiens	71.72	72.18	83.92	75.14	54.95
Комбинированная выборка	47.29	47.35	79.55	49.14	42.11

* Приведено значение *F1*-меры; для OrthoML2GO и PANNZER2 – значение F_{max}.

Таблица 4. Сравнение точности предсказания функции словарей GO на комбинированной выборке, %

Алгоритм	BP	MF	CC
LR	50.9	48.5	56.8
RF	78.4	77.0	82.9
XGB (OrthoML2GO)	78.8	79.8	83.6

Оценка точности идентификации терминов GO для разных словарей

Для проведения более детального анализа работы метода выполнен сравнительный анализ точности предсказания терминов GO для трех основных аспектов (словарей) Gene Ontology: Biological Process (BP, биологические процессы), Molecular Function (MF, молекулярные функции) и Cellular Component (CC, клеточные компоненты). Результаты оценки на комбинированной тестовой выборке для различных алгоритмов машинного обучения, использованных на этапе верификации, представлены в табл. 4.

Результаты показывают, что все алгоритмы машинного обучения демонстрируют схожую тенденцию: наивысшая точность предсказания достигается для аспекта Cellular Component (CC), затем следует Molecular Function (MF), и несколько ниже точность для Biological Process (BP). Это согласуется с общепринятым представлением в биоинформатике: предсказание клеточной локализации (CC) часто является наиболее простой задачей, так как оно сильно коррелирует с наличием специфических сигнальных пептидов и доменов. Предсказание молекулярной функции (MF) также в значительной степени зависит от консервативных функциональных доменов. В свою очередь, предсказание участия в биологических процессах (BP) наиболее проблематично, так как один и тот же белок может участвовать в нескольких процессах, а сами процессы определяются сложными взаимодействиями множества белков, что труднее установить исключительно из данных о гомологии и ортологии.

Метод XGB, выбранный для OrthoML2GO, показал лучшие результаты среди всех протестированных алгоритмов по всем трем аспектам, что дополнительно подтверждает корректность его выбора в качестве финального классификатора. Наши результаты по классификации GO сравнимы с оценками точности других методов, приведенными в литературных источниках (табл. 5). Сравнение выполнено с использованием метрики F_{max} по отдельным аспектам GO: BP – биологические процессы, MF – молекулярные функции, CC – клеточные компоненты.

Метод OrthoML2GO (см. табл. 4) продемонстрировал конкурентоспособные результаты: 78.8 % (BP), 79.8 % (MF) и 83.6 % (CC) на выборке из 20000 последовательностей семи разнородных организмов – как растений, так и животных. При сравнении видно, что OrthoML2GO превосходит большинство исследуемых методов по всем аспектам. Однако PANNZER2 показал более высокие значения для MF (85.8 %) и CC (85.3 %), хотя на меньшей по объему и менее разнообразной выборке (5000 последовательностей из Swiss-Prot).

Следует отметить, что прямое количественное сопоставление с другими методами может быть осложнено методологическими различиями. Во-первых, тестовые выборки существенно разнятся: большинство методов использует БД UniProt/Swiss-Prot, тогда как наша комбинированная выборка включает как растения, так и животных, что может влиять на сравнимость результатов. Во-вторых, критически важна версия Gene Ontology: OrthoML2GO опирается на последнюю версию аннотации OrthoDB v 12.0 (GO 2025), что может вызывать сложности в сопоставлении метрик качества.

Для демонстрации применимости метода OrthoML2GO к слабо изученным организмам был аннотирован протеом зеленой водоросли *Ostreococcus lucimarinus* (табл. S10 и S11). Метод предсказал функции для 5273 из 7603 белковых последовательностей. Анализ выявил преобладание таких биологических процессов, как фосфорилирование (GO:0016310) и трансляция (GO:0006412). Среди молекулярных функций наиболее частыми оказались связывание АТФ (GO:0005524) и нуклеотидов (GO:0000166), а среди клеточных компонентов – мембрана (GO:0016020) и ядро (GO:0005634). Эти результаты демонстрируют способность метода аннотировать слабо изученные протеомы и выявлять функциональные профили, характерные для немодельных организмов.

Заключение

Разработанный нами и представленный в этой работе метод предсказания функций белков OrthoML2GO, интегрирующий поиск гомологов и ортологических групп в базе данных OrthoDB с градиентным бустингом, продемонстрировал высокую эффективность на тестовых выборках.

Таблица 5. Оценки точности аннотации терминами GO для различных словарей разными методами по литературным данным, %

Метод	BP	MF	CC	Литературный источник
PANNZER2	78.4	85.8	85.3	Törönen et al., 2018
DeepGOPlus	58.5	47.4	69.9	Kulmanov, Hoehndorf, 2020
GOLabeler	58.6	37.2	69.1	You et al., 2018
NetGO 2.0	66.6	36.6	66.3	Yao et al., 2021
TALE+	66.7	45.9	67.7	Cao, Shen, 2021

Одним из основных результатов является значительное повышение точности аннотации за счет комбинированного подхода, объединяющего метод *k* ближайших гомологов и информацию об ортологических группах (KNN+OG). Этот гибридный метод превзошел отдельные подходы KNN и OG, особенно при малых значениях параметра *k*. Верификация терминов GO с помощью алгоритмов машинного обучения, в особенности градиентного бустинга (XGB), позволила дополнительно повысить точность за счет эффективного отсева ложноположительных предсказаний, возникающих при использовании отдаленных гомологов и ортогрупп.

Полученные результаты подтверждают, что использование эволюционной информации, заключенной в ортогруппах OrthoDB, в сочетании с алгоритмами машинного обучения – эффективная стратегия для автоматического предсказания функций белковых последовательностей. Предложенный метод OrthoML2GO может стать хорошей альтернативой уже существующим методам. Дальнейшее улучшение точности возможно за счет оптимизации параметров машинного обучения, а также включения дополнительных источников биологической информации. В качестве перспективных исследований намечены следующие направления: оценка переносимости модели на слабо аннотированные протеомы и сравнительный анализ с другими методами, использующими машинное обучение, в том числе нейросетевыми.

Список литературы / References

Altenhoff A.M., Glover N.M., Dessimoz C. Inferring orthology and paralogy. *Methods Mol Biol*. 2019;1910:149-175. doi 10.1007/978-1-4939-9074-0_5

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi 10.1016/S0022-2836(05)80360-2

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., … Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi 10.1038/75556

Benso A., Di Carlo S., Ur Rehman H., Politano G., Savino A., Suravajhala P. A combined approach for genome wide protein function annotation/prediction. *Proteome Sci*. 2013;11(Suppl. 1):S1. doi 10.1186/1477-5956-11-S1-S1

Bradford Y.M., Van Slyke C.E., Ruzicka L., Singer A., Eagle A., Fashena D., Howe D.G., Frazer K., Martin R., Paddock H., Pich C., Ramachandran S., Westerfield M. Zebrafish information network, the knowledgebase for *Danio rerio* research. *Genetics*. 2022;220(4):iyac016. doi 10.1093/genetics/iyac016

Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12(1):59-60. doi 10.1038/nmeth.3176

Cao Y., Shen Y. TALE: Transformer-based protein function Annotation with joint sequence-Label Embedding. *Bioinformatics*. 2021;37(18):2825-2833. doi 10.1093/bioinformatics/btab198

Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. In: KDD ‘16. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016;785-794. doi 10.1145/2939672.2939785

Cheng S., Melkonian M., Smith S.A., Brockington S., Archibald J.M., Delaux P.M., Li F.W., … Graham S.W., Soltis P.S., Liu X., Xu X., Wong G.K. 10KP: A phylodiverse genome sequencing plan. *Gigascience*. 2018;7(3):1-9. doi 10.1093/gigascience/giy013

Conesa A., Götz S., García-Gómez J.M., Terol J., Talón M., Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 2005;21(18):3674-3676. doi 10.1093/bioinformatics/bti610

Dongardive J., Abraham S. Protein Sequence Classification Based on N-Gram and K-Nearest Neighbor Algorithm. In: Behera H., Mohapatra D. (Eds). Computational Intelligence in Data Mining. Vol. 2. Advances in Intelligent Systems and Computing. Vol. 411. Springer, New Delhi, 2016;163-171 doi 10.1007/978-81-322-2731-1_15

du Plessis L., Skunca N., Dessimoz C. The what, where, how and why of gene ontology – a primer for bioinformaticians. *Brief Bioinform*. 2011;12(6):723-735. doi 10.1093/bib/bbr002

Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-2461. doi 10.1093/bioinformatics/btq461

Eisenberg D., Marcotte E.M., Xenarios I., Yeates T.O. Protein function in the post-genomic era. *Nature*. 2000;405(6788):823-826. doi 10.1038/35015694

Fitch W.M. Distinguishing homologous from analogous proteins. *Syst Biol*. 1970;19(2):99-113. doi 10.2307/2412448

Fitch W.M. Homology a personal view on some of the problems. *Trends Genet*. 2000;16(5):227-231. doi 10.1016/s0168-9525(00)02005-9

Galperin M.Y., Koonin E.V. From complete genome sequence to ‘complete’ understanding? *Trends Biotechnol*. 2010;28(8):398-406. doi 10.1016/j.tibtech.2010.05.006

Gene Ontology Consortium; Aleksander S.A., Balhoff J., Carbon S., Cherry J.M., Drabkin H.J., Ebert D., ... Ponferrada V., Zorn A., Ramachandran S., Ruzicka L., Westerfield M. The Gene Ontology knowledgebase in 2023. *Genetics*. 2023;224(1):iyad031. doi 10.1093/genetics/iyad031

Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi 10.1038/nrg.2016.49

Grigoriev I.V., Hayes R.D., Calhoun S., Kamel B., Wang A., Ahrendt S., Dusheyko S., Nikitin R., Mondo S.J., Salamov A., Shabalov I., Kuo A. PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res*. 2021;49(D1):1004-1011. doi 10.1093/nar/gkaa898

Hamilton J.P., Brose J., Buell C.R. SpudDB: a database for accessing potato genomic data. *Genetics*. 2025a;229(3):iyae205. doi 10.1093/genetics/iyae205

Hamilton J.P., Li C., Buell C.R. The rice genome annotation project: an updated database for mining the rice genome. *Nucleic Acids Res*. 2025b;53(1):1614-1622. doi 10.1093/nar/gkae1061

Huntley R.P., Sawford T., Mutowo-Meullenet P., Shypitsyna A., Bonilla C., Martin M.J., O’Donovan C. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Res*. 2015;43(D1):1057-1063. doi 10.1093/nar/gku1113

Jensen L.J., Julien P., Kuhn M., von Mering C., Muller J., Doerks T., Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2008;36(Database issue):250-254. doi 10.1093/nar/gkm796

Kharsikar S., Mugler D., Sheffer D., Moore F., Duan Z.H. A weighted k-nearest neighbor method for gene ontology based protein function prediction. In: Proceedings of the Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS ‘07). IEEE Computer Society, USA, 2007;25-31. doi 10.1109/IMSCCS.2007.13

Kriventseva E.V., Rahman N., Espinosa O., Zdobnov E.M. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res*. 2008;36(Database issue):271-275. doi 10.1093/nar/gkm845

Kulmanov M., Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*. 2020;36(2):422-429. doi 10.1093/bioinformatics/btz595

Kuzniar A., van Ham R.C., Pongor S., Leunissen J.A. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24(11):539-551. doi 10.1016/j.tig.2008.08.009

Lewin H.A., Robinson G.E., Kress W.J., Baker W.J., Coddington J., Crandall K.A., Durbin R., …van Sluys M.A., Soltis P.S., Xu X.,

Yang H., Zhang G. Earth BioGenome project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*. 2018;115(17):4325-4333. doi 10.1073/pnas.1720115115

Liaw A., Wiener M. Classification and Regression by randomForest. *R News*. 2002;2(3):18-22. doi 10.32614/CRAN.package.randomForest

Öztürk-Çolak A., Marygold S.J., Antonazzo G., Attrill H., Goutte-Gattat D., Jenkins V.K., Matthews B.B., Millburn G., Dos Santos G., Tabone C.J.; FlyBase Consortium. FlyBase: updates to the *Drosophila* genes and genomes database. *Genetics*. 2024;227(1):iyad211. doi 10.1093/genetics/iyad211

Pearson W.R. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*. 2013;42(3):3.1.1-3.1.8. doi 10.1002/0471250953.bi0301s42

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2013. Available: http://www.R-project.org/

Reiser L., Bakker E., Subramaniam S., Chen X., Sawant S., Khosa K., Prithvi T., Berardini T.Z. The Arabidopsis Information Resource in 2024. *Genetics*. 2024;227(1):iyae027. doi 10.1093/genetics/iyae027

Sanderson T., Bileschi M.L., Belanger D., Colwell L.J. ProteInfer, deep neural networks for protein functional inference. *eLife*. 2023;12:e80942. doi 10.7554/eLife.80942

Steinegger M., Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. 2017;35(11):1026-1028. doi 10.1038/nbt.3988

Suzuki S., Kakuta M., Ishida T., Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One*. 2014;9(8):e103833. doi 10.1371/journal.pone.0103833

Tegenfeldt F., Kuznetsov D., Manni M., Berkeley M., Zdobnov E.M., Kriventseva E.V. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res*. 2025;53(D1):D516-D522. doi 10.1093/nar/gkae987

Törönen P., Medlar A., Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res*. 2018;46(W1):W84-W88. doi 10.1093/nar/gky350

Wickham H., François R., Henry L., Müller K., Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2025. doi 10.32614/CRAN.package.dplyr

Yao S., You R., Wang S., Xiong Y., Huang X., Zhu S. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res*. 2021;49(W1):W469-W475. doi 10.1093/nar/gkab398

You R., Zhang Z., Xiong Y., Sun F., Mamitsuka H., Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*. 2018;34(14):2465-2473. doi 10.1093/bioinformatics/bty130

Yuan Q., Xie J., Xie J., Zhao H., Yang Y. Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Brief Bioinform*. 2023;24(3):bbad117. doi 10.1093/bib/bbad117

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 24.07.2025. После доработки 10.09.2025. Принята к публикации 15.09.2025.

Прием статей через электронную редакцию на сайте http://vavilov.elpub.ru/index.php/jour
Предварительно нужно зарегистрироваться как автору, затем в правом верхнем углу страницы выбрать «Отправить рукопись». После завершения загрузки материалов обязательно выбрать опцию «Отправить письмо», в этом случае редакция автоматически будет уведомлена о получении новой рукописи.

«Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» до 2011 г. выходил под названием «Информационный вестник ВОГиС»/ "The Herald of Vavilov Society for Geneticists and Breeding Scientists".

Сетевое издание «Вавиловский журнал генетики и селекции (Vavilov Journal of Genetics and Breeding)» – реестровая запись СМИ Эл № ФС77-85772, зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 14 августа 2023 г.

Издание включено ВАК Минобрнауки России в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, Russian Science Citation Index, Российский индекс научного цитирования, ВИНИТИ, Web of Science CC, Scopus, PubMed Central, DOAJ, ROAD, Ulrich's Periodicals Directory, Google Scholar.

Открытый доступ к полным текстам:
русскоязычная версия – на сайте https://vavilovj-icg.ru/
и платформе Научной электронной библиотеки, elibrary.ru/title_about.asp?id=32440

англоязычная версия – на сайте vavilov.elpub.ru/index.php/jour
и платформе PubMed Central, https://www.ncbi.nlm.nih.gov/pmc/journals/3805/

При перепечатке материалов ссылка обязательна.

✉ email: vavilov_journal@bionet.nsc.ru

Издатель: Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр Институт цитологии и генетики
Сибирского отделения Российской академии наук»,
проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Адрес редакции: проспект Академика Лаврентьева, 10, Новосибирск, 630090.
Секретарь по организационным вопросам С.В. Зубова. Тел.: (383)3634977.
Издание подготовлено информационно-издательским отделом ИЦиГ СО РАН. Тел.: (383)3634963*5218.
Начальник отдела: Т.Ф. Чалкова. Редакторы: В.Д. Ахметова, И.Ю. Ануфриева. Дизайн: А.В. Харкевич.
Компьютерная графика и верстка: Т.Б. Коняхина, О.Н. Савватеева.
Дата выхода в свет 01.12.2025. Формат 60 × 84 1/8. Уч.-изд. л. 32.9