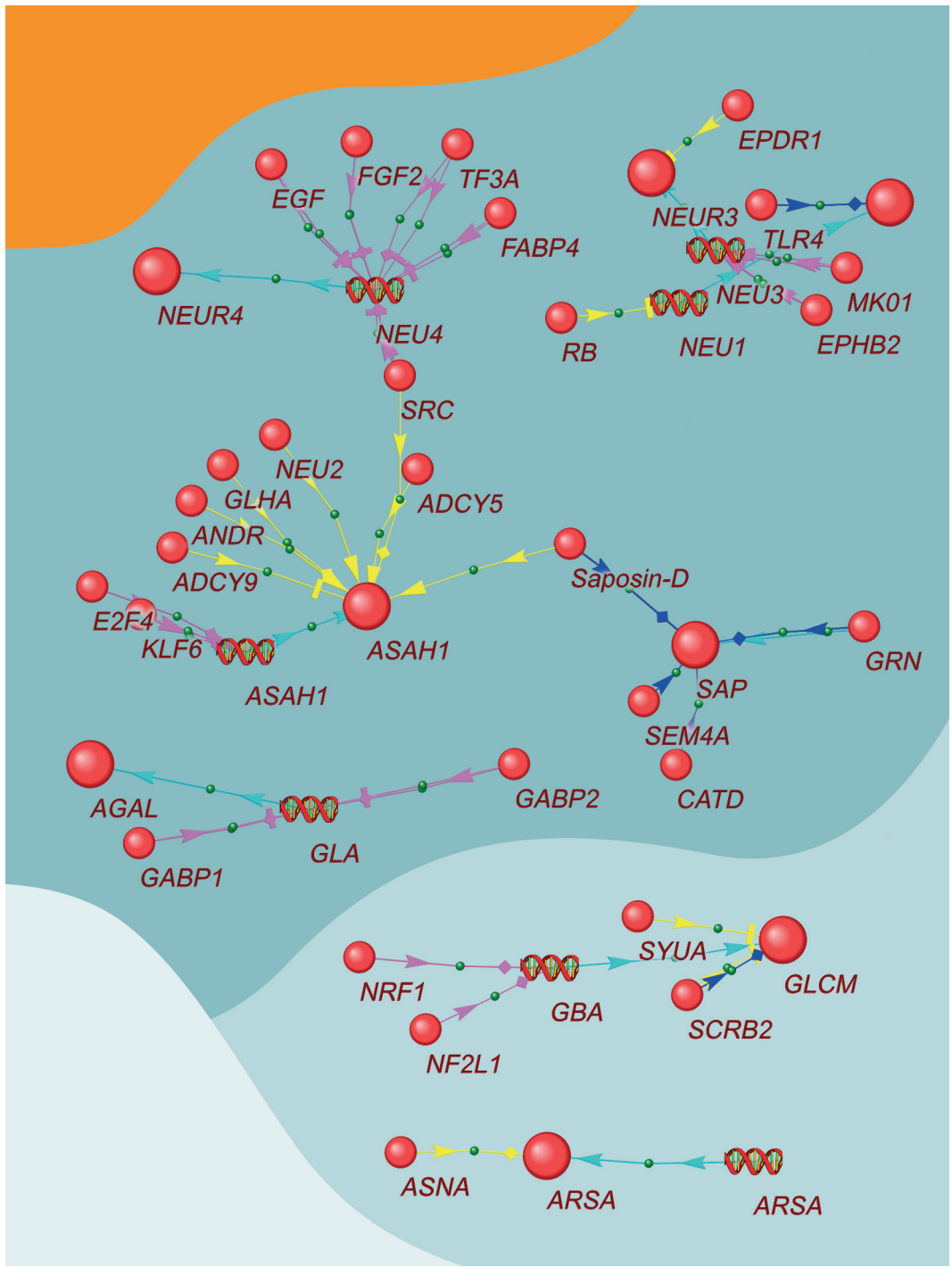
Print ISSN 2500-0462
Online ISSN 2500-3259

ВАВИЛОВСКИЙ
ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ
VAVILOV JOURNAL OF GENETICS AND BREEDING

2023 • 27 • 7



Компьютерная геномика / Системная компьютерная биология / Структурная компьютерная биология и фармакология / Эволюционная компьютерная биология / Методы глубокого машинного обучения для биоинформатики и системной биологии / Экологическая компьютерная биология / Компьютерная биология растений / Техническая биоинформатика / Молекулярная и клеточная биология

vavilov.elpub.ru
vavilovj-icg.ru
vavilov_journal@bionet.nsc.ru
Индекс издания 42153

2023 • 27 • 7

Print ISSN 2500-0462
Online ISSN 2500-3259

Сетевое издание

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

Основан в 1997 г.
Периодичность 8 выпусков в год
DOI 10.18699/VJGB-23-83

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

2023 • 27 • 7

Print ISSN 2500-0462
Online ISSN 2500-3259

Online edition

VAVILOV JOURNAL OF GENETICS AND BREEDING

VAVILOVSKII ZHURNAL GENETIKI I SELEKTSII

Founded in 1997
Published 8 times annually
DOI 10.18699/VJGB-23-83

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

ВАВИЛОВСКИЙ ЖУРНАЛ ГЕНЕТИКИ И СЕЛЕКЦИИ

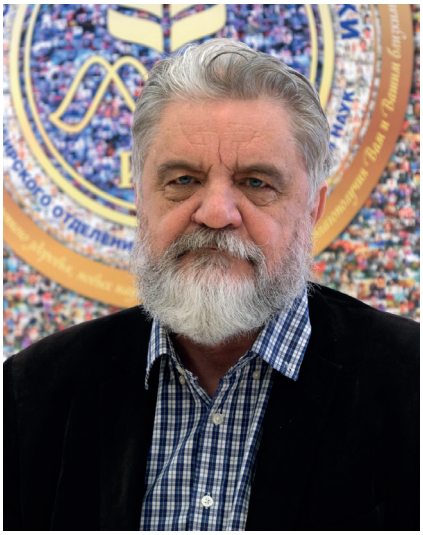
СОДЕРЖАНИЕ • 2023 • 27 • 7

VAVILOV JOURNAL OF GENETICS AND BREEDING

CONTENTS • 2023 • 27 • 7

Вавиловский журнал генетики и селекции. 2023;27(7):725-727
DOI 10.18699/VJGB-23-84

Н.А. Колчанов



Ю.Г. Матушкин

Уважаемые коллеги, дорогие читатели!

Представляем вашему вниманию очередной выпуск «Вавиловского журнала генетики и селекции», посвященный биоинформатике и системной компьютерной биологии. Эти научные направления находятся сейчас в состоянии стремительной трансформации, что обусловлено вступлением наук о жизни в эпоху больших данных. Стремительное развитие омиксных технологий: геномики, транскриптомики, протеомики, метаболомики, а также других высокопроизводительных технологий изучения молекулярно-генетических основ функционирования живых систем привело к информационному взрыву в генетике, которая является основным источником больших данных в мировой науке, обогнав другие науки и технологии по скорости и объемам накопления экспериментальной информации.

Важнейшим результатом анализа, интерпретации, осмысливания больших генетических данных стало формирование новой парадигмы, в рамках которой главными объектами генетики являются не отдельные гены, а генные сети – группы координированно функционирующих генов, взаимодействующих друг с другом через свои продукты, такие как РНК, белки, метаболиты и другие вещества. Именно генные сети обеспечивают формирование всех фенотипических признаков организмов (молекулярных, биохимических, клеточных, физиологических, морфологических, поведенческих, ментальных и т. д.) на основе информации, закодированной в их геномах (Колчанов и др., 2000, 2013; Ananko et al., 2002).

Реконструкция генных сетей – очень сложная задача, требующая поиска, извлечения и интеграции информации, рассеянной среди десятков миллионов научных статей, тысяч фактографических баз данных и миллионов патентов, содержащих биологические, медицинские, фармакологические, химические и другие знания. Решение этой задачи потребовало разработки компьютерных программных систем для автоматизированного извлечения генетических знаний из упомянутых источников, использующих комбинацию традиционного анализа текста и методов машинного обучения (Ivanisenko V.A. et al., 2019; Ivanisenko T.V. et al., 2022). На сегодняшний день более 70 000 генных сетей и их основных компонентов (путей передачи сигналов, сетей белок-белковых, ДНК-белковых, РНК-белковых взаимодействий, метаболических путей) были реконструированы и представлены в базах данных (Pico et al., 2008; Caspi et al., 2020; Kanehisa et al., 2023).

Накопление больших данных привело к пониманию огромной сложности регуляции генных сетей на базовых уровнях их организации, проявляющейся в том, что любой элементарный фундаментальный биохимический или молекулярно-биологический процесс в генной сети контролируется, как правило, десятками, а иногда и сотнями элементарных регуляторных процессов, относится ли это к ферментативной активности белков, регуляции транскрипции генов или к «регуляции сложных метаболических путей» (Колчанов и др., 2008). Указанное обстоятельство создает огромные сложности при реконструкции молекулярных механизмов повреждающего

влияния геномной изменчивости на фенотипические характеристики организмов и клинические симптомы заболеваний, в том числе потому, что регуляторные процессы часто характеризуются высокой степенью нелинейности (Costanzo et al., 2019; Trifonova et al., 2021; Pratap et al., 2022) и динамической неустойчивостью по отношению к изменению начальных данных и констант физико-химических и молекулярно-биологических процессов, лежащих в основе функционирования генных сетей и регуляторных систем (Khlebodarova et al., 2018).

Обработка, анализ и интерпретация потоков больших генетических данных требуют разработки современных методов искусственного интеллекта, ориентированных на живые системы. Одним из ключевых событий, инициировавших в последние годы бурное развитие методов искусственного интеллекта, стала разработка новой архитектуры нейронных сетей, называемых трансформерами, ориентированных на обработку символьных последовательностей, включая тексты на естественных языках (Vaswani et al., 2017). Главная особенность трансформеров состоит в том, что порядок входных последовательностей при обработке не играет никакой роли. Это обеспечивает широкие возможности для распараллеливания, позволяя производить глубокое обучение моделей сразу на терабайтах данных, за гораздо меньшее время, чем было возможно ранее при классической архитектуре нейронных сетей.

Отметим несколько выдающихся достижений данного подхода. Важнейшее значение имеет создание качественных систем машинного перевода с одного естественного языка на другой (Jiao et al., 2023; Wang et al., 2023). Значение этого результата для науки, технологий, культуры, искусства, развития человеческих коммуникаций трудно переоценить.

На основе трансформерных моделей достигнут огромный успех в решении одной из центральных задач молекулярной биологии, над которой бились физики, химики, биологи в течение 60 лет, а именно в предсказании пространственной структуры глобулярных белков по их аминокислотным последовательностям. Для решения этой задачи были разработаны нейронные сети AlphaFold (Thornton et al., 2021) и Rosetta (https://www.rosettacommons.org/), предсказывающие 3D координаты тяжелых атомов белков с точностью, близкой к экспериментальной. Сеть была обучена на сотнях тысяч белков с известной пространственной структурой и десятках миллионов аминокислотных последовательностей.

Благодаря методам машинного обучения, использующим трансформерные подходы, открылась возможность моделирования динамики сложных молекулярно-биологических структур, содержащих очень большое (до 10^9) количество атомов (Pandey et al., 2022). Эти результаты имеют исключительное значение не только для фундаментальной науки, но и для широкого круга областей с громадным потенциалом фактического применения, таких как биотехнологии, генетика, медицина, фармакология, создание новых материалов и множество других.

После 2017 г., когда появились первые публикации по трансформерным технологиям, отмечена экспоненциальная динамика роста количества публикаций с использованием методов искусственного интеллекта (Eraslan et al., 2019; Boudry et al., 2022). Еще один подход к машинному обучению, получивший широкое распространение и развитие в последние годы, – это графовые нейронные сети (GNN), которые на основе векторного представления вершин графов с учетом их локального окружения дают качественно новые возможности для анализа сложных сетевых структур (Hamilton et al., 2017). Применение GNN эффективно для описания, анализа и моделирования широчайшего круга сетевых систем – как природных, так и антропогенных и технических: генных сетей, сетей межмолекулярных взаимодействий, сетей знаний, социальных и др. (Ektefaie et al., 2023).

В заключение следует отметить, что принципиальным ограничением для широкого применения методов искусственного интеллекта в практически значимых областях человеческой деятельности является непрозрачность принимаемых им решений. В ряде работ (Ma et al., 2018) показано, что стратегический путь преодоления этого недостатка – разработка гибридных информационных систем нового поколения, интегрирующих классические методы биоинформатики, системной компьютерной биологии и новые технологии искусственного интеллекта на основе онтологического описания предметных областей исследований. Только такой подход, как нам представляется, может обеспечить как скорость и качество обработки больших генетических данных с помощью методов искусственного интеллекта, так и прозрачность получаемых на его основе результатов.

Список литературы / References

Колчанов Н.А., Ананько Е.А., Колпаков Ф.А., Подколодная О.А., Игнатьева Е.В., Горячковская Т.Н., Степаненко И.Л. Генные сети. *Мол. биология*. 2000;34(4):533-544

[Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaia O.A., Ignatieva E.V., Goriachkovskaia T.N., Stepanenko E.L. Gene networks. *Molekulyarnaya Biologiya = Molecular Biology*. 2000;34(4):533-544 (in Russian)]

Колчанов Н.А., Гончаров С.С., Лихошвай В.А., Иванисенко В.А. Системная компьютерная биология. Новосибирск: Изд-во СО РАН, 2008

[Kolchanov N.A., Goncharov S.S., Likhoshvay V.A., Ivanisenko V.A. Systems Computational Biology. Novosibirsk: Publ. House SB RAS, 2008 (in Russian)]

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции*. 2013;4(2):833-850

[Kolchanov N.A., Ignatieva E.V., Podkolodnaya O.A., Likhoshvay V.A., Matushkin Yu.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vaviliv Journal of Genetics and Breeding.* 2013;4(2):833-850 (in Russian)]

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. GeneNet: a database on structure and functional organisation of gene networks. *Nucleic Acids Res*. 2002;30(1):398-401. DOI 10.1093/nar/30.1.398

Boudry C., Al Hajj H., Arnould L., Mouriaux F. Analysis of international publication trends in artificial intelligence in ophthalmology. *Graefes Arch. Clin. Exp. Ophthalmol*. 2022;260(5):1779-1788. DOI 10.1007/s00417-021-05511-7

Caspi R., Billington R., Keseler I.M., Kothari A., Krummenacker M., Midford P.E., Ong W.K., Paley S., Subhraveti P., Karp P.D. The MetaCyc database of metabolic pathways and enzymes – a 2019

update. *Nucleic Acids Res*. 2020;48(D1):D445-D453. DOI 10.1093/nar/gkz862

Costanzo M., Kuzmin E., van Leeuwen J., Mair B., Moffat J., Boone C., Andrews B. Global genetic networks and the genotype-to-phenotype relationship. *Cell*. 2019;177(1):85-100. DOI 10.1016/j.cell.2019.01.033

Ektefaie Y., Dasoulas G., Noori A., Farhat M., Zitnik M. Multimodal learning with graphs. *Nat. Mach. Intell*. 2023;5:340-350. DOI 10.1038/s42256-023-00624-6

Eraslan G., Avsec Ž., Gagneur J., Theis F.J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet*. 2019;20(7):389-403. DOI 10.1038/s41576-019-0122-6

Hamilton W., Ying Z., Leskovec J. Inductive representation learning on large graphs. *Adv. Neural Inf. Process. Syst*. 2017;30:1024-1034

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 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

Jiao W., Wang W., Huang J.T., Wang X., Tu Z.P. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv*. 2023. DOI 10.48550/arXiv.2301.08745

Kanehisa M., Furumichi M., Sato Y., Kawashima M., Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 2023;51(D1):D587-D592. DOI 10.1093/nar/gkac963

Khlebodarova T.M., Kogai V.V., Trifonova E.A., Likhoshvai V.A. Dynamic landscape of the local translation at activated synapses. *Mol. Psychiatry*. 2018;23(1):107-114. DOI 10.1038/mp.2017.245

Ma J., Yu M.K., Fong S., Ono K., Sage E., Demchak B., Sharan R., Ideker T. Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods*. 2018;15(4):290-298. DOI 10.1038/nmeth.4627

Pandey M., Fernandez M., Gentile F., Isayev O., Tropsha A., Stern A.C., Cherkasov A. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell*. 2022;4(3):211-221. DOI 10.1038/s42256-022-00463-x

Pico A.R., Kelder T., van Iersel M.P., Hanspers K., Conklin B.R., Evelo C. WikiPathways: pathway editing for the people. *PLoS Biol*. 2008;6(7):e184. DOI 10.1371/journal.pbio.0060184

Pratap A., Raja R., Agarwal R.P., Alzabut J., Niezabitowski M., Hincal E. Further results on asymptotic and finite-time stability analysis of fractional-order time-delayed genetic regulatory networks. *Neurocomputing*. 2022;475:26-37. DOI 10.1016/j.neucom.2021.11.088

Thornton J.M., Laskowski R.A., Borkakoti N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat. Med.* 2021;27(10):1666-1669. DOI 10.1038/s41591-021-01533-0

Trifonova E.A., Klimenko A.I., Mustafin Z.S., Lashin S.A., Kochetov A.V. Do autism spectrum and autoimmune disorders share predisposition gene signature due to mTOR signaling pathway controlling expression? *Int. J. Mol. Sci.* 2021;22(10):5248. DOI 10.3390/ijms22105248

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need. *arXiv*. 2017. DOI 10.48550/arXiv.1706.03762

Wang L., Lyu C., Ji T., Zhang Z., Yu D., Shi S., Tu Z. Document-level machine translation with large language models. *arXiv*. 2023. DOI 10.48550/arXiv.2304.02210

Научные редакторы выпуска:

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

канд. биол. наук Ю.Г. Матушкин,
вед. науч. сотрудник ФИЦ ИЦиГ СО РАН

Вавиловский журнал генетики и селекции. 2023;27(7):728-736
DOI 10.18699/VJGB-23-85

Перевод на английский язык https://vavilov.elpub.ru/jour

Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования

С.В. Филонов[1, 2], Н.Л. Подколодный[1, 3] ✉, О.А. Подколодная[1], Н.Н. Твердохлеб[1], П.М. Пономаренко[1], Д.А. Рассказов[1], А.Г. Богомолов[1], М.П. Пономаренко[1]

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

Аннотация. Ранее было показано, что уровень экспрессии генов человека положительно коррелирует с аффинностью TBP к промоторам этих генов. В свою очередь, однонуклеотидные полиморфизмы (SNP) в промоторах генов человека могут влиять на аффинность белка TBP к ДНК и, как следствие, на экспрессию генов. В ИЦиГ СО РАН разработан метод предсказания аффинности TBP к промоторам генов на основе трехшагового механизма связывания, включающего скольжение TBP по ДНК, остановку TBP в месте связывания, фиксацию комплекса TBP–промотор за счет изгиба спирали ДНК. Метод показал высокую корреляцию теоретических предсказаний с измеренными значениями при многократной экспериментальной проверке независимыми группами исследователей. На основе этой модели в ИЦиГ СО РАН ранее были разработаны веб-сервисы SNP_TATA_Z-tester и SNP_TATA_Comparator, позволяющие вычислять статистическую оценку вызванного SNP изменения аффинности связывания TBP с промотором гена человека и прогнозировать изменение экспрессии, которые могут быть связаны с генетической предрасположенностью к заболеваниям или фенотипическими особенностями организма. В настоящей работе проведена интеграция в единой базе данных информации об однонуклеотидных полиморфизмах в промоторах генов человека, полученной путем автоматической экстракции из различных гетерогенных источников данных, а также результатов оценки аффинности TBP к промотору с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом. Показана возможность использования базы данных Human_SNP_TATAdb для аннотации и выявления кандидатных SNP-маркеров заболеваний. Представлены результаты полногеномного анализа данных, включая особенности распределения генов по количеству транскриптов, распределение SNP, влияющих на аффинность TBP к ДНК по позициям внутри промоторов, а также закономерности, связывающие между собой аффинность TBP к промотору, специфичность сайта связывания TBP с промотором и другие характеристики промоторов. Результаты полногеномного анализа показали, что аффинность TBP к промотору и специфичность его сайта связывания статистически связаны с другими характеристиками промоторов, важными для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов.

Ключевые слова: TATA-бокс; аффинность; TBP; однонуклеотидный полиморфизм; база данных; полногеномный анализ.

Для цитирования: 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

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

S.V. Filonov[1, 2], N.L. Podkolodnyy[1, 3] ✉, O.A. Podkolodnaya[1], N.N. Tverdokhleb[1], P.M. Ponomarenko[1], D.A. Rasskazov[1], A.G. Bogomolov[1], M.P. Ponomarenko[1]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pnl@bionet.nsc.ru

Abstract. It was previously shown that the expression levels of human genes positively correlate with TBP affinity for the promoters of these genes. In turn, single nucleotide polymorphisms (SNPs) in human gene promoters can affect TBP affinity for DNA and, as a consequence, gene expression. The Institute of Cytology and Genetics SB RAS (ICG) has developed a method for predicting TBP affinity for gene promoters based on a three-step binding mechanism: (1) TBP slides along DNA, (2) TBP stops at the binding site, and (3) the TBP-promoter complex is fixed due to DNA helix bending. The method showed a high correlation of theoretical predictions with measured values during repeated experimental testing by independent groups of researchers. This model served as a base for other ICG web services, SNP_TATA_Z-tester and SNP_TATA_Comparator, which make a statistical assessment of the SNP-induced change in the affinity of TBP binding to the human gene promoter and help predict changes in expression that may be associated with a genetic predisposition to diseases or phenotypic features of the organism. In this work, we integrated into a single database information about SNPs in human gene promoters obtained by automatic extraction from various heterogeneous data sources, as well as the estimates of TBP affinity for the promoter obtained using the three-step binding model and predicting their effect on gene expression for wild-type promoters and promoters with SNPs. We have shown that Human_SNP_TATAdb can be used for annotation and identification of candidate SNP markers of diseases. The results of a genome-wide data analysis are presented, including the distribution of genes with respect to the number of transcripts, the distribution of SNPs affecting TBP-DNA affinity with respect to positions within promoters, as well as patterns linking TBP affinity for the promoter, the specificity of the TBP binding site for the promoter and other characteristics of promoters. The results of the genome-wide analysis showed that the affinity of TBP for the promoter and the specificity of its binding site are statistically related to other characteristics of promoters important for the functional classification of promoters and the study of the features of differential gene expression.
Key words: TATA box; affinity; TBP; single nucleotide polymorphism; database; genome-wide analysis.

For citation: 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 Journal of Genetics and Breeding*. 2023;27(7):728-736. DOI 10.18699/VJGB-23-85

Введение

Разработка методов предсказания влияния мутаций на уровень экспрессии генов различных организмов имеет важное значение для решения задач в области биотехнологии, селекции растений, медицины и так далее. Мутации в геноме человека могут быть ассоциированы со множеством физиологических особенностей и заболеваний, и знание о наличии и причине их безусловно необходимо для активно развивающегося подхода персонализированной медицины. Самым распространенным типом мутаций в геноме человека являются однонуклеотидные полиморфизмы (Single Nucleotide Polymorphism, SNP) – отличия последовательности ДНК размером в один нуклеотид. Однонуклеотидные полиморфизмы могут локализоваться в различных функциональных районах генома, от чего зависит характер их проявления. Наиболее изучены мутации в кодирующих районах гена, они непосредственно влияют на структуру транскрибируемой мРНК и синтезируемого белка. Однако полногеномные ассоциативные исследования (GWAS) показали, что большинство однонуклеотидных полиморфизмов, которые в значительной степени связаны с предрасположенностью к заболеванию, лежит в некодирующих областях (Hindorff et al., 2009; French, Edwards, 2020; Chandra et al., 2021), а более 90 % из них расположены в регуляторных элементах (Maurano et al., 2012). Одним из наиболее изученных регуляторных районов на данный момент является район TATA-бокса в промоторе, от последовательности которого зависит сродство к нему белка TBP (TATA Binding Protein), – ключевого фактора инициации транскрипции. Мутации в этом районе могут влиять на связывание белка TBP с промотором и, как следствие, на экспрессию гена (Савинкова и др., 2007).

В ИЦиГ СО РАН разработан метод предсказания аффинности TBP к промоторам генов на основе трехшагового механизма связывания (Пономаренко и др., 2008). Метод показал высокую корреляцию теоретических предсказаний с измеренными значениями аффинности при многократной экспериментальной проверке независимыми группами исследователей (Delgadillo et al., 2009; Savinkova et al., 2013; Oshchepkov et al., 2022). На основе этой модели в ИЦиГ СО РАН разработан веб-сервис SNP_TATA_Z-tester (Рассказов и др., 2013), позволяющий вычислять статистическую оценку вызванного SNP изменения аффинности связывания TBP с промотором гена человека и прогнозировать изменение экспрессии. С помощью этого веб-сервиса мы ранее выявили кандидатные SNP-маркеры аутоиммунных заболеваний (Ponomarenko et al., 2016а), поведенческих расстройств (Chadaeva et al., 2016), хронопатологий (Ponomarenko et al., 2016б) и других заболеваний.

В настоящей работе проведена интеграция в единой базе данных информации об однонуклеотидных полиморфизмах в промоторах генов человека, полученной путем автоматической экстракции из различных гетерогенных источников данных, а также результатов оценки аффинности TBP к промотору и специфичности сайта связывания TBP с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов из референсного генома и промоторов с однонуклеотидным полиморфизмом.

Ключевым вариантом использования базы данных Human_SNP_TATAdb является аннотация промоторов и генов с целью поиска кандидатных SNP-маркеров заболеваний. Учитывая, что к настоящему времени уже вы-

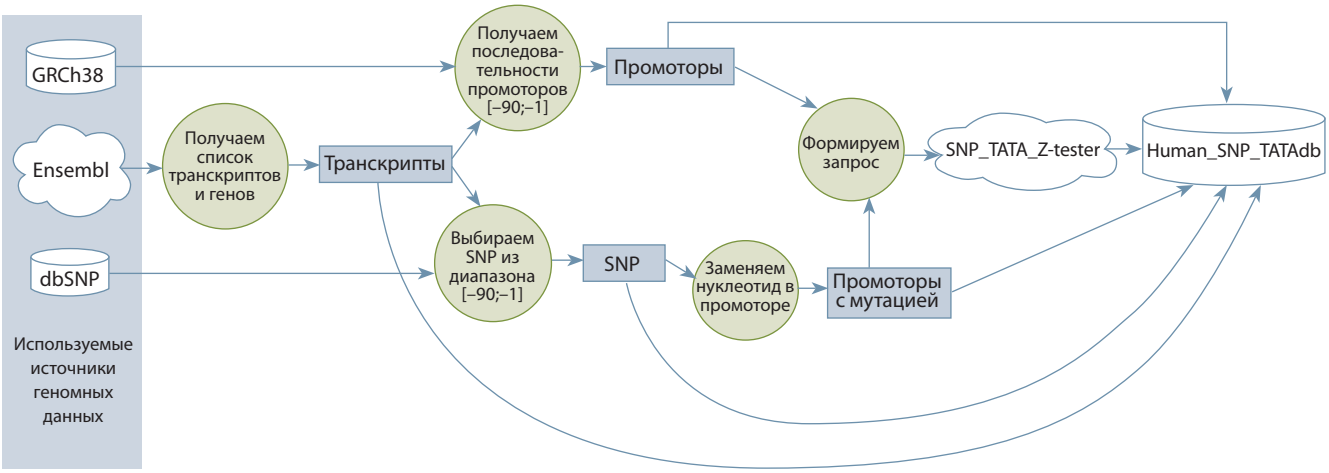


Рис. 1. Схема потока данных для инициализации базы данных Human_SNP_TATAdb.

полнено много исследований, в которых проводилась такого рода аннотация, мы привели в качестве примера один из вариантов.

Представлены результаты полногеномного анализа данных, включая особенности распределения генов по количеству транскриптов, распределение SNP, влияющих на аффинность TBP к ДНК по позициям внутри промоторов, а также закономерности, связывающие между собой аффинность TBP к промотору, специфичность сайта связывания TBP с промотором и другие характеристики промоторов, важные для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов.

Материалы и методы

Ниже представлены этапы работы по интеграции данных и создания базы данных (рис. 1). Данные о генах и их атрибутах, стартах транскрипции и транскриптах получены с веб-сервиса Ensembl (Birney et al., 2004). Для доступа к сервисам и базам данных использована библиотека Bioconductor языка R со следующими пакетами:

1. biomaRt[1] – пакет, который обеспечивает интерфейс для коллекции баз данных Ensembl, позволяя извлекать большие объемы данных унифицированным способом и использовать при анализе данных в Bioconductor.
2. BSgenome.Hsapiens.NCBI.GRCh38[2] – пакет, обеспечивающий доступ к последовательностям генома *Homo sapiens* (Human), предоставленным NCBI (GRCh38.p13).
3. SNPlocs.Hsapiens.dbSNP155.GRCh38[3] – пакет для доступа к dbSNP 155, включающий информацию о 949021448 SNP в хромосомах 1–22, X, Y и MT.

Для выявления старта транскрипции необходимо использовать транскрипты с качественной аннотацией, которая включает эту информацию и для которых доказана их биологическая релевантность. При описании транскриптов в Ensembl для определения наиболее качественно аннотированных ставят специальные метки. Мы включили в базу данных только те транскрипты, качество аннотации которых соответствует метке GENCODE Basic[4]. В соответствии со спецификацией Ensembl GENCODE Basic содержит по крайней мере один транскрипт для каждого гена в генетическом наборе GENCODE независимо от биотипа, т.е. каждый ген представлен в базовом наборе GENCODE. Для генов, кодирующих белок, в базовый набор GENCODE включены только полноразмерные транскрипты, кодирующие белок.

Для заданных координат старта транскрипции определяются координаты и нуклеотидные последовательности соответствующего им промотора ([–90; –1] от старта транскрипции). Данные о SNP получены с использованием базы данных dbSNP[5] (Sherry et al., 2001). Для каждого промотора выделены SNP, локализованные в пределах [–90; –1] от старта транскрипции. Минорные варианты последовательности промотора созданы автоматически путем внесения в основные варианты последовательностей соответствующих замен нуклеотидов из базы данных dbSNP (вып. 155). Для выявления TATA-содержащих промоторов использована весовая матрица Бухера (Bucher, 1990).

Аффинность TBP к ДНК рассчитывали с применением трехшаговой модели связывания, разработанной ранее в ИЦиГ СО РАН (Ponomarenko et al., 2008) и реализованной нами многопоточной высокопроизводительной версии программы SNP_TATA_Z-tester. Эта программа также позволяет оценить статистическую значимость изменения аффинности белка TBP к промотору при точечных заменах нуклеотидов (SNP) в промоторе с использованием z-критерия.

Аффинность, или сродство, TBP описывается константой ассоциации комплекса TBP/ДНК. Однако в настоящее время вместо константы ассоциации обычно используют обратную меру – константу диссоциации K_d. В этом случае аффинность TBP к ДНК, измеренная в наномолях на литр (нМ/л), будет равна $A = 10^9/K_d$. Чем меньше K_d, тем

[1] https://bioconductor.org/packages/release/bioc/html/biomaRt.html
[2] https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.NCBI.GRCh38.html
[3] https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP155.GRCh38.html
[4] https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html
[5] https://www.ncbi.nlm.nih.gov/snp/

выше сродство TBP к промотору и сильнее взаимодействие TBP с промотором.

Второй вариант, представленный в базе данных, – логарифмическая форма аффинности $\alpha = 9*\ln(10) - \ln(K_d)$, которая удобна для сравнения показателей аффинности TBP к промотору, так как имеет близкое к нормальному распределение. При увеличении α возрастают сродство TBP к промотору и сила их взаимодействия.

Расчеты аффинности проведены для референсных последовательностей ДНК всех промоторов и минорных вариантов последовательностей этих промоторов с одним однонуклеотидным полиморфизмом. Для каждой минорной последовательности оценивали отклонение аффинности TBP к промотору от аффинности, полученной для последовательности ДНК промотора из референсного генома. При этом определяли уровень статистической значимости этих изменений.

Ранее показано, что аффинность TBP к промотору статистически достоверно коррелирует с уровнем экспрессии соответствующего транскрипта (Mogno et al., 2010). Поэтому при статистически достоверном увеличении или уменьшении аффинности TBP в базе данных указывается оценка соответствующего изменения уровня экспрессии транскрипта. На основе оценок аффинности белка TBP к промотору введены дополнительные характеристики, например специфичность сайта связывания белка TBP с промотором, который можно использовать для классификации промотора и биологической аннотации групп промоторов или генов.

Специфичность сайта связывания TBP с промотором гена соответствует максимальной нормированной аффинности TBP к промотору гена относительно средней аффинности TBP по каждой позиции скользящего окна (Ponomarenko et al., 2015), не включая 10 позиций ближайших к старту транскрипции (всего 55 значений). Специфичность Z рассчитывали следующим образом:

$$Z = \frac{\alpha_{\max} - \overline{\alpha}}{\sigma_\alpha}, \ \sigma_\alpha = \sqrt{\frac{1}{54}\sum_1^{55}(\alpha_i - \overline{\alpha})},$$

где α_i – оценка аффинности TBP к промотору в позиции i, $\overline{\alpha}$ – среднее значение α_i, σ_α– несмещенная оценка среднеквадратичного отклонения α_i, Z – специфичность сайта связывания белка TBP с промотором.

Еще один важный показатель, описывающий вызванное SNP изменение аффинности TBP к промотору, – натуральный логарифм отношения K_d для референсных (*wt*) и минорных (*mt*) аллелей рассматриваемого SNP:

$$k_{snp} = \ln(K_{d,\,wt}/K_{d,\,mt}).$$

Положительные или отрицательные значения k_{snp} указывают на то, что экспрессия гена для минорного аллеля соответственно выше или ниже, чем для случая референсного варианта. Этот показатель использовался для выявления кандидатных SNP-маркеров, которые могут быть связаны с генетической предрасположенностью к заболеваниям; в частности, сделаны предсказания, которые согласуются с клиническими данными о недостаточной экспрессии этого гена у пациентов с вариабельным иммунодефицитом, инсультом и преэклампсией (Ponomarenko et al., 2017).

Результаты и обсуждение

База данных

В ИЦиГ СО РАН разработана база данных Human_SNP_TATAdb (рис. 2). Базу данных заполняли в соответствии со сценарием интеграции данных и инициализации базы данных (см. рис. 1). База данных реализована на основе СУБД MySQL[6] версии 8.0 и включает 6 основных таблиц (chromosomes, genes, transcripts, snps, promoters, promoters_has_snps), 10 вспомогательных таблиц и словарей (см. рис. 2). Работа с базой данных осуществляется через SQL-запросы.

Таблица chromosomes включает идентификатор хромосомы, длину, количество нуклеотидов и вид организма.

Таблица genes содержит информацию об идентификаторах гена в разных базах данных, в том числе в Ensembl, символьное имя гена, ссылку на хромосому, цепь, биотип гена.

Таблица transcripts включает информацию об идентификаторах транскрипта, координаты транскрипта в геноме, биотип транскрипта и ссылку на промотор и ген.

Таблица snps включает следующую информацию: идентификаторы SNP, позиции SNP в геноме, ссылка на хромосому и аллель. За один SNP здесь и далее принимается однозначный вариант изменения генома. Полиморфизмы, имеющие один rs идентификатор, но допускающие несколько вариантов замены нуклеотида, считаются по количеству таких вариантов.

Необходимо отметить, что одна и та же нуклеотидная замена может попадать в разные промоторы гена и поразному изменять уровень аффинности белка TBP к этим промоторам, и поэтому в базе данных заданы две таблицы для описания промоторов promoters и promoters_has_snps с отношением 1:N (на один промотор может оказывать влияние несколько SNP), а таблицы snps и promoters_has_snps также связаны отношением 1:N (один SNP может входить в несколько промоторов).

В таблицу promoters включена следующая информация: идентификатор промотора, последовательность ДНК, соответствующая району [–90; –1] от старта транскрипции, координаты старта и конца промотора в геноме, аффинность белка TBP к промотору с ошибкой, ссылка на ген.

Таблица promoters_has_snps содержит информацию об идентификаторе промотора, ссылку на SNP, координаты SNP в промоторе и относительно старта транскрипции, последовательность промотора дикого типа и промотора с SNP, аффинность TBP к промотору с ошибкой, характер изменения экспрессии гена при мутации в промоторе, уровень значимости статистического теста.

Таблица source_snp_dbs включает информацию, которая необходима для автоматизированного обновления базы данных Human_SNP_TATAdb: об источниках данных, версии баз данных, ссылки на базы данных.

Типы отношений между таблицами задают ограничения, которые соответствуют природе данных и поэтому важны для сохранения целостности базы данных, а также обеспечивают дополнительный контроль данных и уменьшают возможность ошибок. В частности, у каждого гена

[6] https://www.mysql.com/

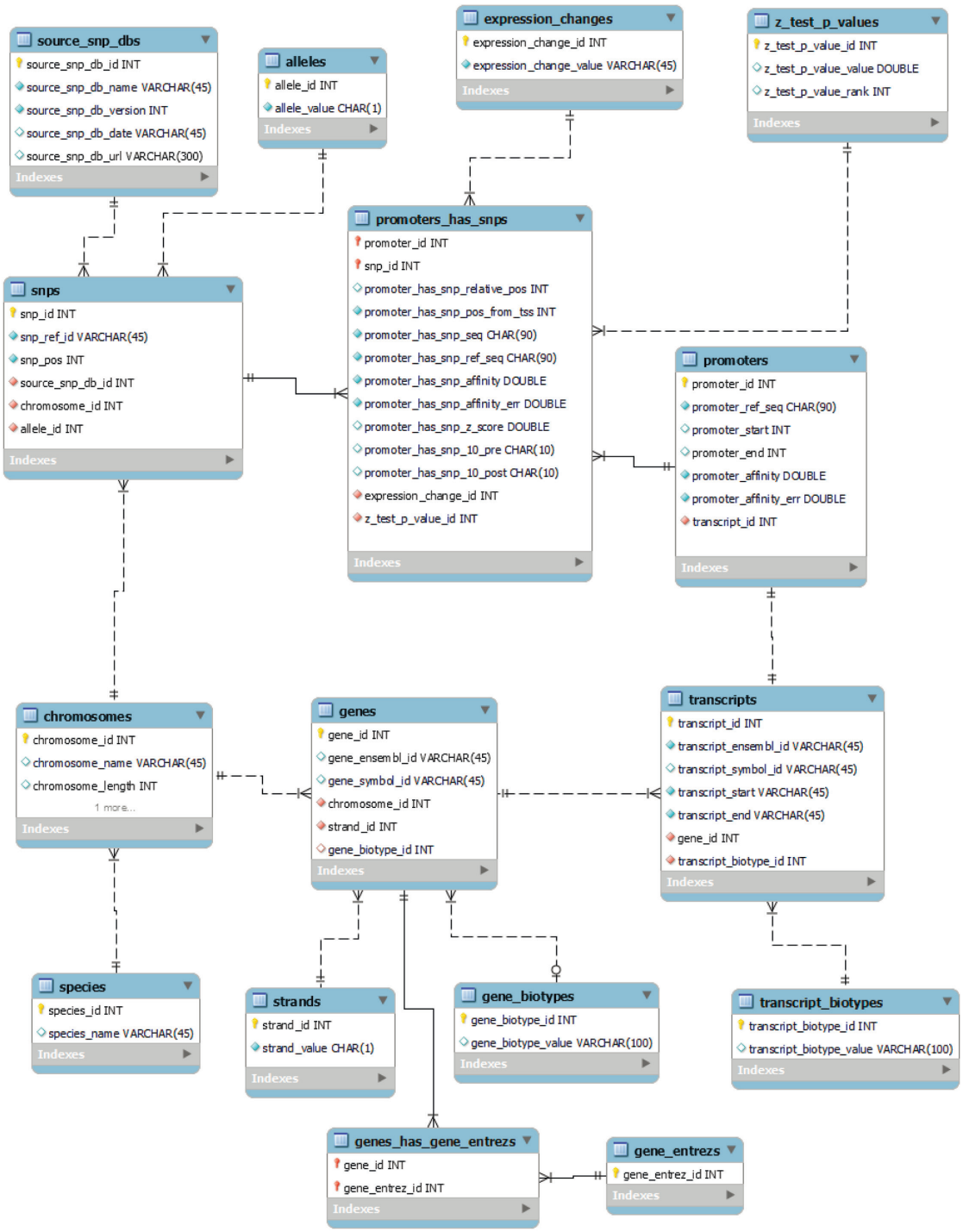


Рис. 2. Схема базы данных Human_SNP_TATAdb.

может быть один или несколько промоторов, каждый промотор может регулировать экспрессию одного или несколько транскриптов.

В итоге база данных содержит информацию о:

- 62603 генах, из которых 19314 кодируют белки;
- 117414 транскриптах, из которых 63141 кодирует белки;
- 5305816 вариантов SNP в промоторах генов в интервале [–90; –1] от старта транскрипции, из них 3199285 в промоторах белок кодирующих генов;
- для 445875 вариантов SNP в промоторе белок кодирующего гена предсказано, что они статистически значимо (p-value < 0.05) изменяют уровень аффинности TBP к этому промотору.

Варианты использования базы данных Human_SNP_TATAdb

Представленные в базе данных аффинность белка TBP к промотору, специфичность сайта связывания TBP с промотором и оценки изменения этих характеристик при

однонуклеотидном полиморфизме важны для поиска маркеров генетической предрасположенности заболеваний, выявления и функциональной интерпретации классов промоторов, схожих по механизму регуляции ранней стадии инициации транскрипции и так далее.

База данных Human_SNP_TATAdb также может быть использована для аннотации генов или группы генов в терминах аффинности TBP к промотору или специфичности сайта связывания TBP с промотором. Чтобы определить характеристику гена, связанную со спецификой связывания TBP с промоторами гена с целью проведения GO-анализа, можно использовать средние значения аффинности TBP к промоторам гена или аффинности TBP к промотору, соответствующему единственному для гена транскрипту, который определен экспертами Ensembl в качестве канонического и задается в базе данных меткой Ensembl Canonical[7], который в целом наиболее консервативен, наиболее экспрессируем, имеет самую длинную кодирующую последовательность и представлен в других ключевых ресурсах, таких как NCBI и UniProt. Мы помечаем соответствующий ему промотор как канонический и используем такие характеристики, как аффинность TBP к каноническому промотору и специфичность сайта связывания TBP с каноническим промотором, для аннотации гена или группы генов.

Корреляционный анализ показал, что между аффинностью TBP к каноническому промотору гена и средней аффинностью промоторов гена наблюдается сильная линейная зависимость (R = 0.88, d.f. = 19308), поэтому оба варианта дают сходные результаты. Однако использование аффинности TBP к каноническому промотору гена, по-видимому, биологически более обосновано. Безусловно, ключевым вариантом применения базы данных Human_SNP_TATAdb является аннотация генов и поиск кандидатных SNP-маркеров предрасположенности к заболеваниям.

Учитывая, что к настоящему времени уже выполнено много исследований, в которых проводилась такого рода аннотация, мы приведем в качестве примера использование базы данных Human_SNP_TATAdb для аннотации и выявления кандидатных SNP-маркеров атерогенеза, атеросклероза и атеропротекции работу (Bogomolov et al., 2023).

Предварительно были отобраны 1068 генов человека, связанных с этими заболеваниями. Информация о SNP в промоторах этих генов человека, результатах оценки аффинности TBP к промоторам и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом получена из базы данных Human_SNP_TATAdb. Эта информация была дополнена аннотацией отобранных генов, подготовленной экспертами, и сформировано представление базы данных, ориентированное на анализ генов, связанных с атерогенезом, атеросклерозом и атеропротекцией, внешний доступ к которой осуществляется через веб-интерфейс[8].

Анализ *in silico* всех 5112 SNP в их промоторах выявил 330 кандидатов в маркеры SNP, статистически значимо изменяющих аффинность TATA-связывающего белка (TBP) к этим промоторам. Далее сравнили соответствующие частоты SNP, которые увеличивают и уменьшают сродство TBP к промоторам одних и тех же генов. Сравнение было сделано для анализа того, находятся ли эти гены под действием естественного отбора или нейтрального дрейфа. Мы обнаружили, что естественный отбор действует против недостаточной экспрессии хаб-генов атерогенеза, атеросклероза и атерозащиты и благодаря усиленной атеропротекции способствует улучшению здоровья человека (Bogomolov et al., 2023).

Примеры использования базы данных Human_SNP_TATAdb для полногеномного анализа

Разработанная база данных позволяет проводить анализ полногеномной статистики и распределения указанных показателей в различных группах промоторов, например TATA-содержащих. Для полногеномного анализа мы использовали белок кодирующие гены и транскрипты, отобранные по значениям полей ‘gene_biotype’ и ‘transcript_biotype’ равными ‘protein_coding’.

Альтернативные промоторы и аффинность TBP/ДНК

Следует отметить, что один ген может иметь несколько транскриптов, инициация транскрипции которых происходит с использованием разных промоторов, для которых оценивается аффинность белка TBP. Как видно из рис. 3, наибольшее число белок кодирующих генов (29.77 % генов) имеет единственный транскрипт и, как следствие, один промотор. Пять процентов белок кодирующих генов имеют не менее 9 белок кодирующих транскриптов. Анализ распределения генов по числу транскриптов показал, что среднее число транскриптов на ген – 3.27, а медиана – 2 транскрипта на ген. Максимальное число (87) белок кодирующих транскриптов наблюдается у гена *Mapk10* (*mitogen-activated protein kinase 10*).

Наш анализ показал, что распределение средней аффинности TBP к каноническим промоторам в группах генов, разбитых по числу транскриптов, близко к равномерному. Таким образом, нет необходимости нивелировать эффекты, обусловленных разным числом транскриптов у гена при проведении анализа данных с использованием аффинности TBP.

Распределение SNP, изменяющих экспрессию генов по позициям промотора

Распределение SNP, статистически значимо изменяющих экспрессию генов по позициям от старта транскрипции, имеет ярко выраженное отклонение от равномерного (рис. 4). В районе [–35; –20], соответствующем обычному расположению TATA-бокса, число таких SNP заметно выше, чем в других районах промотора.

Число SNP, уменьшающих экспрессию генов в районе [–35; –20], соответствующие расположению TATA-бокса, более чем в полтора раза выше, чем в других районах промотора. Это может быть связано с тем, что SNP в этом районе, как правило, разрушают TATA-бокс.

Число SNP, увеличивающих экспрессию генов, выше на флангах наиболее частых локализаций TATA-бокса. Пики локализованы в –24 и –32 позициях от старта транскрипции. Следует отметить, что распределение всех SNP по позициям промоторов белок кодирующих генов равномерно.

[7] https://www.ensembl.org/info/genome/genebuild/canonical.html
[8] http://www.sysbio.ru/Human_SNP_TATAdb

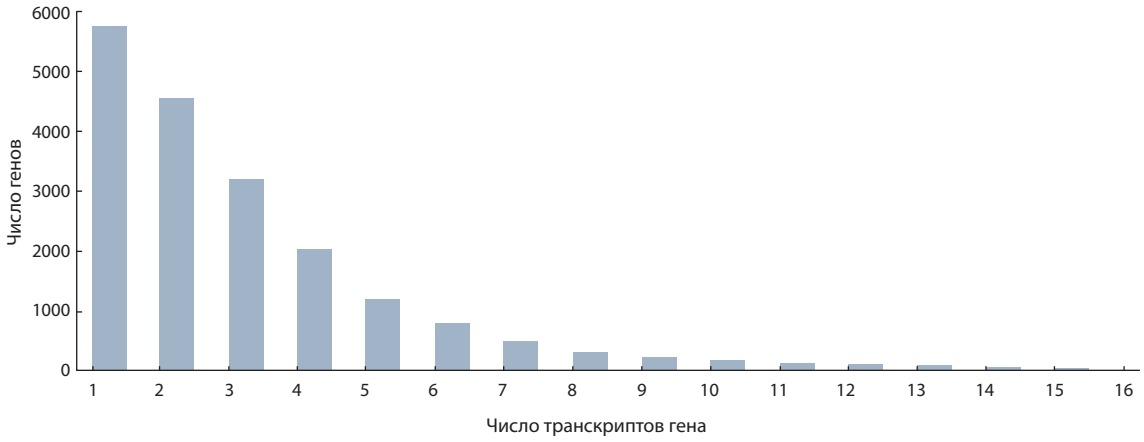


Рис. 3. Распределение белок кодирующих генов по числу транскриптов.

Рис. 4. Распределение числа SNP, увеличивающих (excees) и уменьшающих (deficiency) аффинность TBP к ДНК промоторов белок кодирующих генов в зависимости от позиции SNP относительно старта транскрипции.

Это говорит о том, что увеличение на флангах TATA-бокса числа SNP, увеличивающих экспрессию генов, может иметь функциональное значение.

Аффинность TBP к TATA-содержащим и TATA-не содержащим промоторам белок кодирующих генов

Анализ зависимости показателей аффинности TBP/ДНК, измеренной в логарифмической шкале ($\alpha = 9*\ln(10) - \ln(K_d)$) для TATA-содержащих и TATA-не содержащих промоторов белок кодирующих генов (рис. 5), показал, что в группе TATA-содержащих промоторов наблюдается более высокая аффинность TBP/ДНК, что соответствует более сильному сходству TBP к промотору.

Функциональные SNP, влияющие на аффинность TBP к ДНК промоторов и специфичность сайта связывания белка TBP

Проведен анализ зависимости доли SNP, статистически значимо влияющих на аффинность TBP к ДНК промоторов белок кодирующих генов, от специфичности сайта связывания белка TBP (рис. 6). Показано, что SNP в промоторах с низкой специфичностью сайта связывания TBP с промотором, как правило, приводят к увеличению экспрессии генов, а в промоторах с высокой специфичностью доля SNP, понижающих экспрессию, повышена.

Анализ таблицы сопряженности показал, что низкие значения специфичности сайта связывания TBP с промотором (Spec < 2.5) чаще наблюдаются на промоторах без TATA-бокса (TATA–) ($\chi^2 = 10385$, p-value < 1.0e–228).

Заключение

В настоящей работе описана база данных Human_SNP_TATAdb, которая включает информацию о SNP в промоторах генов человека, полученную путем автоматической экстракции из различных гетерогенных источников данных, результатах оценки аффинности TBP к промотору с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом.

Представленные в базе данных аффинность белка TBP к промотору, специфичность сайта связывания TBP с про-

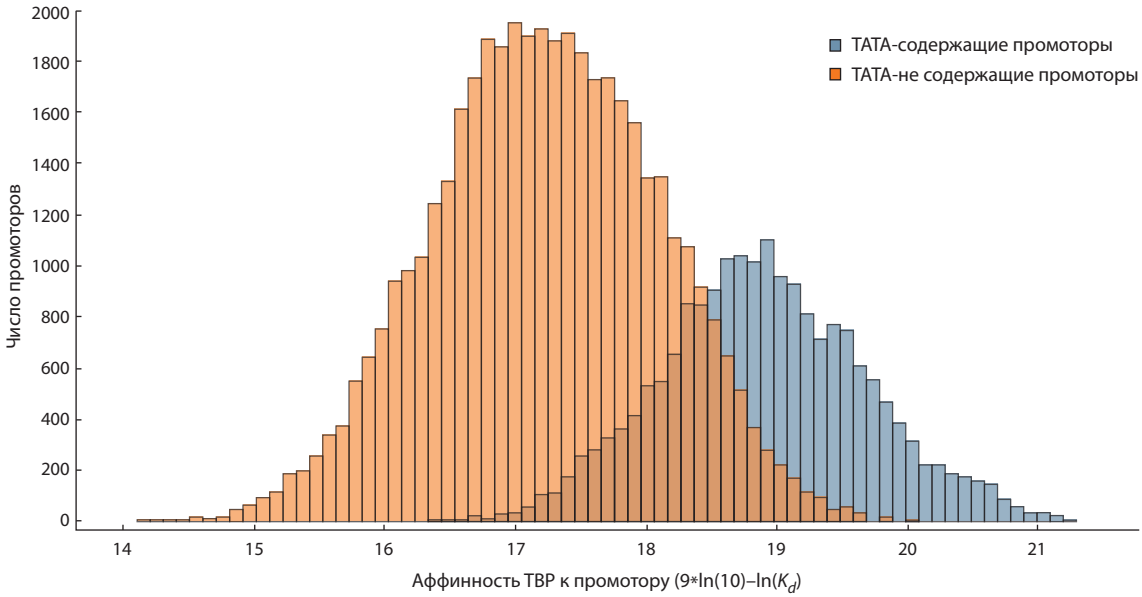


Рис. 5. Распределение промоторов белок кодирующих генов по аффинности TBP в группах TATA-содержащих промоторов и промоторах без TATA-бокса. Оценка аффинности TBP к промотору по оси *x* задается в логарифмической шкале.

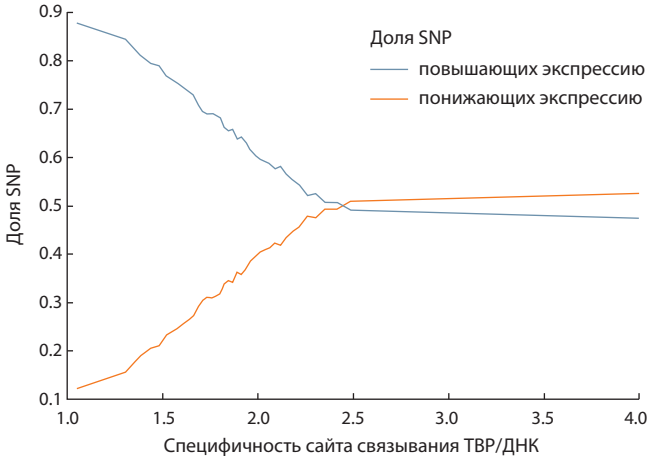


Рис. 6. Доля SNP в промоторах, повышающих и понижающих экспрессию белок кодирующих генов в зависимости от специфичности сайта связывания TBP к ДНК промотору.

Таблица сопряженности специфичности сайта связывания TBP с промотором и наличия TATA-бокса в промоторе

Специфичность	TATA–	TATA+	Всего
Spec < 2.5	**29114**	10379	39493
Spec ≥ 2.5	14538	9109	23647
Всего	43652	19488	63140

мотором и оценки изменения этих характеристик при SNP важны для поиска кандидатных маркеров генетической предрасположенности заболеваний, выявления и функциональной интерпретации классов промоторов, схожих по механизму регуляции ранней стадии инициации транскрипции, и так далее. База данных Human_SNP_TATAdb также может быть использована для аннотации генов или групп генов в терминах аффинности TBP к промотору или специфичности сайта связывания TBP с промотором.

Результаты полногеномного анализа показали, что аффинность TBP к промотору и специфичность его сайта связывания статистически связаны с другими характеристиками промоторов, важными для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов. Использование базы данных Human_SNP_TATAdb для аннотации генов и выявление кандидатных SNP-маркеров атерогенеза, атеросклероза и атеропротекции – один из примеров, в результате которого становятся доступны новые знания о влиянии различных одиночных полиморфизмов на предрасположенность к тем или иным заболеваниям.

Список литературы / References

Рассказов Д.А., Гунбин К.В., Пономаренко П.М., Вишневский О.В., Пономаренко М.П., Афонников Д.А. SNP_TATA_COMPARATOR: web-сервис применения уравнения равновесия TBP/TATA-комплекса в сравнительной оценке SNPS промоторов генов, связанных с болезнями человека. *Вавиловский журнал генетики и селекции.* 2013;17(4/1):599-606
[Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_COMPARATOR: web service for comparison of SNPS within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2013;17(4/1):599-606 (in Russian)]

Савинкова Л.К., Драчкова И.А., Пономаренко М.П., Лысова М.В., Аршинова Т.В., Колчанов Н.А. Взаимодействие рекомбинантного TATA-связывающего белка с TATA-боксами промоторов генов млекопитающих. *Экологическая генетика.* 2007;5(2):44-49. DOI 10.17816/ecogen5244-49

[Savinkova L.K., Drachkova I.A., Ponomarenko M.P., Lysova M.V., Arshinova T.V., Kolchanov N.A. Interaction of recombinant TATA-binding protein with mammals gene promoter TATA boxes. *Ekologicheskaya genetika = Ecological genetics.* 2007;5(2):44-49. DOI 10.17816/ecogen5244-49 (in Russian)]

Birney E., Andrews T.D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., ..., Cox A., Hubbard T., Clamp M. An overview of Ensembl. *Genome Res*. 2004;14(5):925-928. DOI 10.1101/gr.1860604

Bogomolov A., Filonov S., Chadaeva I., Rasskazov D., Khandaev B., Zolotareva K., Kazachek A., … Kolchanov N., Tverdokhleb N., Ponomarenko M. Candidate SNP markers significantly altering the affinity of TATA-binding protein for the promoters of human hub genes for atherogenesis, atherosclerosis and atheroprotection. *Int. J. Mol. Sci.* 2023;24(10):9010. DOI 10.3390/ijms24109010

Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*. 1990;212(4):563-578. DOI 10.1016/0022-2836(90)90223-9

Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A., Sharypova E.B., Kashina E.V., Matveeva M.Yu., Arshinova T.V., Ponomarenko P.M., Arkova O.V., Bondar N.P., Savinkova L.K., Kolchanov N.A. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics.* 2016;17(Suppl. 14):995. DOI 10.1186/s12864-016-3353-3

Chandra V., Bhattacharyya S., Schmiedel B.J., Madrigal A., Gonzalez-Colin C., Fotsing S., Crinklaw A., Seumois G., Mohammadi P., Kronenberg M., Peters B., Ay F., Vijayanand P. Promoter interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat. Genet*. 2021;53(1):110-119. DOI 10.1038/s41588-020-00745-3

Delgadillo R.F., Whittington J.E., Parkhurst L.K., Parkhurst L.J. The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. *Biochemistry.* 2009; 48(8):1801-1809. DOI 10.1021/bi8018724

French J.D., Edwards S.L. The role of noncoding variants in heritable disease. *Trends Genet.* 2020;36(11):880-891. DOI 10.1016/j.tig.2020.07.004

Hindorff L.A., Sethupathy P., Junkins H.A., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*. 2009; 106(23):9362-9367. DOI 10.1073/pnas.0903103106

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., … Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. DOI 10.1126/science.1222794

Mogno I., Vallania F., Mitra R.D., Cohen B.A. TATA is a modular component of synthetic promoters. *Genome Res*. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Oshchepkov D., Chadaeva I., Kozhemyakina R., Zolotareva K., Khandaev B., Sharypova E., Ponomarenko P., Bogomolov A., Klimova N.V., Shikhevich S., Redina O., Kolosova N.G., Nazarenko M., Kolchanov N.A., Markel A., Ponomarenko M. Stress reactivity, susceptibility to hypertension, and differential expression of genes in hypertensive compared to normotensive patients. *Int. J. Mol. Sci*. 2022;23(5):2835. DOI 10.3390/ijms23052835

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl. Biochem. Biophys*. 2008;419:88-92. DOI 10.1134/S1607672908020117

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed Res. Int*. 2015;2015:359835. DOI 10.1155/2015/359835

Ponomarenko M.P., Arkova O., Rasskazov D., Ponomarenko P., Savinkova L., Kolchanov N. Candidate SNP markers of genderbiased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front. Immunol*. 2016a;7:130. DOI 10.3389/fimmu.2016.00130

Ponomarenko P., Rasskazov D., Suslov V., Sharypova E., Savinkova L., Podkolodnaya O., Podkolodny N.L., Tverdokhleb N.N., Chadaeva I., Ponomarenko M., Kolchanov N. Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed Res. Int.* 2016б;2016:8642703. DOI 10.1155/2016/8642703

Ponomarenko M., Rasskazov D., Chadaeva I., Sharypova E., Ponomarenko P., Arkova O., Kashina E., Ivanisenko N., Zhechev D., Savinkova L., Kolchanov N. SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* (*Schol. Ed.*). 2017;9(2):276-306. DOI 10.2741/s488

Savinkova L., Drachkova I., Arshinova T., Ponomarenko P., Ponomarenko M., Kolchanov N. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PLoS One*. 2013;8(2).e54626. DOI 10.1371/journal.pone.0054626

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311. DOI 10.1093/nar/29.1.308

ORCID ID

N.L. Podkolodnyy orcid.org/0000-0001-9132-7997
O.A. Podkolodnaya orcid.org/0000-0003-3247-0114
P.M. Ponomarenko orcid.org/0000-0003-2715-9612
D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
M.P. Ponomarenko orcid.org/0000-0003-1663-318X

Благодарности. Работа выполнена при поддержке бюджетных проектов FWNR-2022-0020, № 0251-2022-0005 и Федеральной научно-технической программы развития генетических технологий России.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 22.08.2023. После доработки 15.09.2023. Принята к публикации 19.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):737-745
DOI 10.18699/VJGB-23-86

Перевод на английский язык https://vavilov.elpub.ru/jour

GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования

А.Ю. Пронозин[1, 2] ✉, Е.А. Салина[1, 2, 3], Д.А. Афонников[1, 2, 4]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский государственный аграрный университет, Новосибирск, Россия
[4] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ pronozinartem95@gmail.com

Аннотация. Развитие технологий секвенирования нового поколения открыло новые возможности для генотипирования различных организмов, включая растения. Метод генотипирования путем секвенирования (GBS) применяется для идентификации генетической изменчивости и более быстрого генотипирования образцов, а также является более экономически эффективным методом в сравнении с полногеномным секвенированием. GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Этот метод был применен для генетического картирования, выявления молекулярных маркеров, геномной селекции, в исследовании генетического разнообразия, идентификации сортов, а также в исследованиях в области биологии охраны природы и эволюционной экологии. Однако сокращение времени и стоимости секвенирования привело к необходимости разработки качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных. Для этих целей были разработаны биоинформатические конвейеры анализа данных, полученных методом GBS. Вследствие схожести этапов обработки существующие конвейеры в основном различаются комбинацией программных пакетов, специфически подобранных для обработки данных как для определенных, так и для любых организмов. Несмотря на качественно подобранные пакеты программ, конвейеры имеют некоторые недостатки, например отсутствие возможности автоматизации процесса расчета (каждый этап нужно запускать вручную), что значительно снижает скорость исследования. В большинстве конвейеров отсутствует возможность автоматической установки всех необходимых программных пакетов, а также нет возможности отключения ненужного или пройденного этапа. В настоящей работе нами был разработан биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер применим для любых видов организмов. Реализация конвейера на платформе Snakemake позволила полностью автоматизировать процесс расчета и установки необходимых программных пакетов. Конвейер позволяет обрабатывать большие объемы данных (более 400 образцов).
Ключевые слова: генотипирование путем секвенирования; биоинформатический конвейер; ячмень.

Для цитирования: Пронозин А.Ю., Салина Е.А., Афонников Д.А. GBS-DP: биоинформатический конвейер для обработки данных, полученных генотипированием путем секвенирования. *Вавиловский журнал генетики и селекции*. 2023;27(7):737-745. DOI 10.18699/VJGB-23-86

GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing

A.Y. Pronozin[1, 2] ✉, E.A. Salina[1, 2, 3], D.A. Afonnikov[1, 2, 4]

[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
[3] Novosibirsk State Agrarian University, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
✉ pronozinartem95@gmail.com

Abstract. The development of next-generation sequencing technologies has provided new opportunities for genotyping various organisms, including plants. Genotyping by sequencing (GBS) is used to identify genetic variability more rapidly, and is more cost-effective than whole-genome sequencing. GBS has demonstrated its reliability and flexibility for a number of plant species and populations. It has been applied to genetic mapping, molecular marker discovery, genomic selection, genetic diversity studies, variety identification, conservation biology and evolutionary studies. However, reduction in sequencing time and cost has led to the need to develop efficient bioinformatics analyses for an ever-expanding amount of sequenced data. Bioinformatics pipelines for GBS data analysis serve the purpose. Due to the similarity of data processing steps, existing pipelines are mainly characterised by a combination of software

packages specifically selected either to process data for certain organisms or to process data from any organisms. However, despite the usage of efficient software packages, these pipelines have some disadvantages. For example, there is a lack of process automation (in some pipelines, each step must be started manually), which significantly reduces the performance of the analysis. In the majority of pipelines, there is no possibility of automatic installation of all necessary software packages; for most of them, it is also impossible to switch off unnecessary or completed steps. In the present work, we have developed a GBS-DP bioinformatics pipeline for GBS data analysis. The pipeline can be applied for various species. The pipeline is implemented using the Snakemake workflow engine. This implementation allows fully automating the process of calculation and installation of the necessary software packages. Our pipeline is able to perform analysis of large datasets (more than 400 samples).
Key words: genotyping by sequencing (GBS); bioinformatic pipeline; hordeum.

For citation: Pronozin A.Y., Salina E.A., Afonnikov D.A. GBS-DP: a bioinformatics pipeline for processing data coming from genotyping by sequencing. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023; 27(7):737-745. DOI 10.18699/VJGB-23-86

Введение

Генетическое разнообразие является важнейшей основой для изучения устойчивости растений к биотическим и абиотическим стрессам и создания новых высокоадаптивных и урожайных сортов сельскохозяйственных культур. Изучение генетического разнообразия осуществляется с использованием различных методов генетического анализа. На сегодняшний день один из наиболее перспективных методов – применение молекулярных маркеров (ДНК-маркеров) (Канукова и др., 2019). Это генетические маркеры, анализируемые на уровне ДНК (Хлесткина, 2013). С их помощью можно выявлять генетическое разнообразие популяций, подвидов, видов, эффективно определять хозяйственно ценные признаки еще на начальном этапе селекции на уровне ДНК (Сухарева, Кулуев, 2018).

Для генетического анализа особенно удобны SNP-маркеры (Хлесткина, 2013). SNP (single-nucleotide polymorphism – однонуклеотидный полиморфизм) – это однонуклеотидная позиция в геномной ДНК, для которой в популяции встречаются различные вариации последовательности (аллелей) (Сухарева, Кулуев, 2018). SNP широко используют для изучения аллельного полиморфизма, тестирования чистоты семян, анализа гаплотипа и родословных, а также для генотипирования и построения генетических карт.

Получить информацию об SNP-маркерах в настоящее время можно для любого растения в масштабе полного генома благодаря технологиям высокопроизводительного секвенирования нового поколения. Идентификация SNP возможна с помощью стратегий полногеномного секвенирования (WGS) и генотипирования путем секвенирования (GBS) (Scheben et al., 2017). Цель полногеномного секвенирования – получить короткие фрагменты последовательности полного генома и на этой основе путем их выравнивания на референсный геном или полногеномной сборки оценить вариации ДНК. Это сложная и дорогостоящая задача, цена за один геном превышает 2000$ и зависит от размера и сложности генома, желаемого уровня полноты и вычислительных ресурсов (Narum et al., 2013). Например, секвенирование полного генома ячменя до уровня хромосом обходится примерно в 60,000$ (Monat et al., 2019). Выделяют также методы полногеномного секвенирования с более низкой глубиной прочтения, стоимость которых в разы меньше: 100–400$ за геном. Однако, как утверждают авторы (Bimber et al., 2016), при этом снижается точность получаемых данных о генотипах.

Метод генотипирования путем секвенирования более быстрый и экономически эффективный, чем WGS. Например, стоимость секвенирования фрагментов генома ячменя в эксперименте GBS не превышает 30$ (Monat et al., 2019). В методе GBS выделяют два подхода секвенирования. В первом для фрагментации образцов ДНК используются ферменты рестрикции, специфичные для конкретных сайтов, после чего производится секвенирование полученных фрагментов (Glaubitz et al., 2014). Во втором к обоим концам фрагментов ДНК лигируются уникальные последовательности адаптеров, один из которых содержит уникальную последовательность «штрихкод», после чего производится секвенирование данных маркированных фрагментов ДНК (Elshire et al., 2011). Поскольку при секвенировании фрагменты ДНК прочитываются только вблизи сайтов рестрикции, в методе GBS не происходит прочтения полногеномной последовательности ДНК. За счет этого процесс секвенирования существенно удешевляется, однако количество SNP, которые можно идентифицировать, оказывается меньше, чем при полногеномном секвенировании. Тем не менее данных, полученных при помощи протокола GBS, оказывается вполне достаточно, чтобы с приемлемой точностью характеризовать генетическое разнообразие популяций сельскохозяйственных растений.

Метод GBS продемонстрировал свою надежность и гибкость для ряда видов и популяций растений. Он был применен для выявления молекулярных маркеров для генетического картирования и геномной селекции (Poland et al., 2012), в исследовании генетического разнообразия (Lu et al., 2013; Peterson et al., 2014), идентификации сортов (Wang et al., 2020; Rajendran et al., 2022), а также исследованиях в области биологии охраны природы и эволюционной экологии (Narum et al., 2013). GBS существенно сокращает как стоимость, так и время, необходимое для секвенирования исследуемых образцов. Это потребовало разработки качественного биоинформатического анализа для постоянно расширяющегося количества секвенированных данных. В результате были разработаны биоинформатические конвейеры анализа данных, полученных методом GBS. Существующие конвейеры имеют схожую схему анализа данных, которая включает: проверку качества сырых прочтений, демультиплексирование, картирование на референсный геном и поиск полиморфизмов.

Этап картирования на геном делится на два типа: на основе референсного генома (Glaubitz et al., 2014; Tor-

kamaneh et al., 2017; Wickland et al., 2017) и на основе «имитации» референсного генома (Mock Reference) (Melo et al., 2016). В первом случае после контроля качества сырых прочтений последовательности картируются на референсный геном с целью выявления полиморфизмов (Torkamaneh et al., 2017). Однако если референсный геном отсутствует или имеет низкое качество сборки, применяют метод «имитации» референсного генома. Этот метод производит кластеризацию исследуемых прочтений для выявления консенсусных последовательностей (центроидов), на основе которых осуществляется сборка генома (Melo et al., 2016). Вследствие схожести этапов обработки данных, существующие конвейеры в основном различаются комбинацией программ. Подобные комбинации программ должны учитывать различные геномные характеристики, такие как количество выявленных полиморфизмов, сложность генома, степень гетерозиготности, доля повторяющихся последовательностей во всем геноме. Также более современные конвейеры позволяют подбирать параметры для исследуемых организмов (Torkamaneh et al., 2017; Wickland et al., 2017), тогда как более ранние конвейеры имеют некоторые ограничения. Например, в программе TASSEL надо указывать ограничение длины последовательностей, что приводит к потере значительного количества коротких сырых прочтений (Glaubitz et al., 2014; Melo et al., 2016). Из-за постоянного роста количества секвенированных библиотек конвейеры должны предоставлять возможность обработки большого объема данных за один запуск. Важным аспектом конвейеров является также автоматизация процесса обработки и простота установки программы.

В настоящей работе мы разработали биоинформатический конвейер GBS-DP для анализа данных, полученных методом GBS. Конвейер включает схему обработки GBS данных, предложенную в работе (Jayakodi et al., 2020), и применим для любых видов организмов. Конвейер позволяет обрабатывать большие объемы данных (более 400 образцов) и реализован с помощью программного менеджера Snakemake (Köster, Rahmann, 2012).

Материалы и методы

Биоинформатический конвейер GBS-DP анализа данных, полученных методом GBS, представлен на рис. 1.

На вход конвейера подается путь к набору библиотек прочтений и путь к референсному геному. Библиотеки прочтений должны быть в формате FASTQ, референсный геном – в формате FASTA. В случае если библиотеки имеют баркодирование, необходимо предварительно их демультиплицировать.

Конвейер состоит из трех основных этапов: предобработка данных, поиск полиморфизмов, анализ генетического разнообразия. Предобработка данных включает проверку качества сырых прочтений, удаление адаптеров и построение индекса референсного генома. Поиск полиморфизмов состоит из картирования предобработанных прочтений на референсный геном, сортировки картированных прочтений и поиска однонуклеотидных полиморфизмов. Анализ генетического разнообразия разделяется на два варианта обработки данных: если полученные данные превышают занимаемый объем памяти в 1 Тб и если полученные данные не превышают занимаемый объем памяти в 1 Тб. Более детальное описание каждого этапа приведено ниже.

Предобработка данных. На этом этапе производится контроль качества, удаление адаптеров сырых прочтений и построение индекса референсного генома. Контроль качества и удаление адаптеров производятся программой cutadapt (Martin, 2011). Для прочтений каждой библиотеки удаляются адаптеры, список которых пользователь должен внести в файл конфигураций.

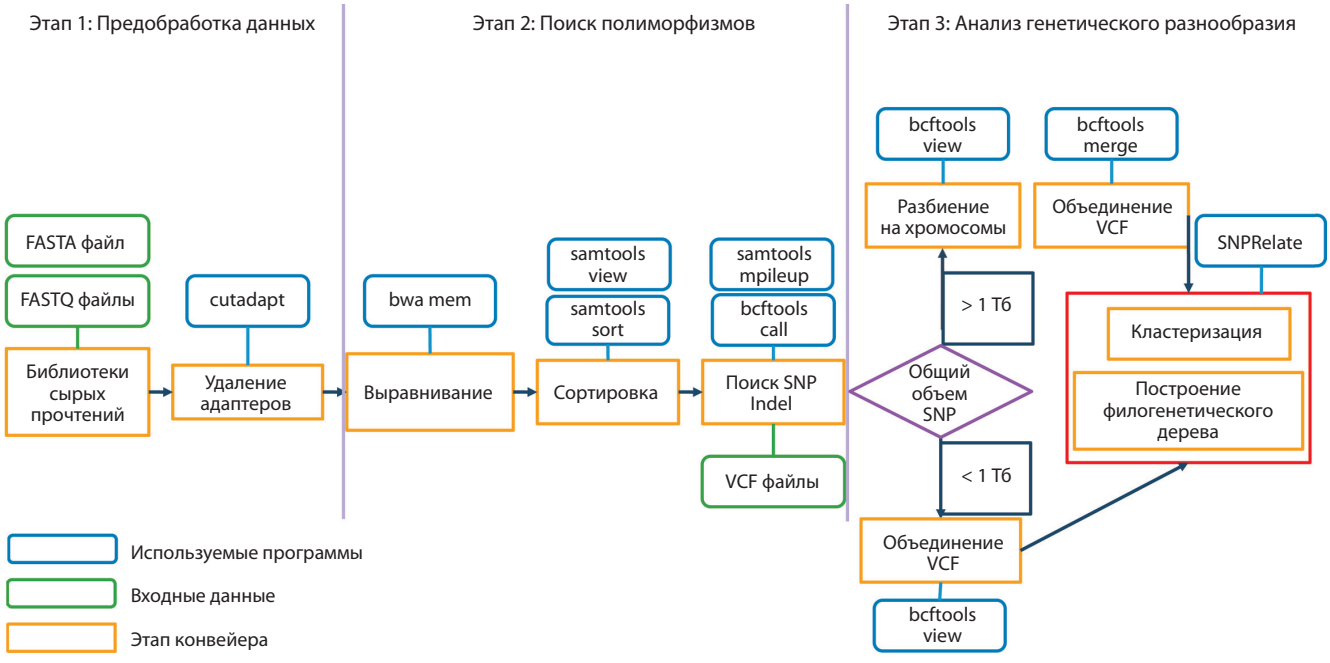


Рис. 1. Блок-схема биоинформатического конвейера GBS-DP обработки данных GBS.

На этом этапе конвейер выполняет построение индекса референсного генома с помощью программы bwa index (Li H., 2013).

Поиск полиморфизмов состоит из картирования предобработанных прочтений на референсный геном, сортировки картированных прочтений и поиска однонуклеотидных полиморфизмов.

Картирование предобработанных прочтений производится программой bwa mem (Li H., 2013) с параметрами "–k 19 –w 100".

Результаты картирования, полученные в формате SAM, переводятся в формат BAM и сортируются комбинацией программ samtools view и samtools sort соответственно (Danecek et al., 2021). В отсортированных файлах производится поиск полиморфизмов (SNP, вставок и делеций (индел)) с помощью комбинации программ samtools mpileup и bcftools call (Danecek et al., 2021). Ранее было показано на примере генома пшеницы (Yao et al., 2020), что комбинация программ "Samtools/mpileup + BWA-mem", которая использована в нашем конвейере, превосходит другие комбинации программ картирования и идентификации полиморфизмов.

Анализ генетического разнообразия разделяется на два варианта обработки данных: если полученные данные превышают занимаемый объем памяти в 1 Тб и если полученные данные не превышают занимаемый объем памяти в 1 Тб.

Выбор соответствующей опции осуществляется автоматически и связан с увеличенной нагрузкой на оперативную память компьютера при работе с большими данными (если суммарный размер полученных файлов VCF превышает 1 Тб). Вариант обработки для данных с общим объемом меньше 1 Тб включает три этапа:

1) результаты поиска полиморфизмов в формате VCF для каждой библиотеки индексируются с помощью программы bcftools index (Danecek et al., 2021);
2) проиндексированные файлы объединяются в общий файл формата VCF в программе bcftools merge (Danecek et al., 2021). Этот файл содержит данные о полиморфизмах всех исследуемых образцов для всех хромосом;
3) полученный общий файл формата VCF конвертируется в формат GDS (Genomic Data Structure) с помощью пакета R – SeqArray (Zheng et al., 2017). Данный формат позволяет значительно сократить объем оперативной памяти, затрачиваемой на обработку результатов поиска полиморфизмов, за счет перевода табличного формата в бинарный.

Вариант обработки для данных с общим занимаемым объемом больше 1 Тб включает четыре этапа:

1) результаты поиска полиморфизмов в формате VCF для каждой библиотеки разбиваются на хромосомы с помощью программы bcftools view (Danecek et al., 2021);
2) полученные файлы с полиморфизмами для каждой хромосомы индексируются с использованием программы bcftools index (Danecek et al., 2021);
3) далее файлы с полиморфизмами объединяются для каждой хромосомы. В результате получаются файлы, содержащие информацию о полиморфизмах во всех библиотеках для отдельной хромосомы;
4) файлы для отдельных хромосом в формате VCF конвертируются в формат GDS. После этого полученные файлы формата GDS для каждой хромосомы объединяются в общий файл с помощью функции snpgdsCombineGeno пакета SNPRelate (Zheng et al., 2017).

Схема построения филогенетического дерева и кластеризации для обоих вариантов идентичная. Следует отметить, что при оценке функционального значения SNP важно также учитывать функциональное значение полиморфных локусов, находящихся с ним в неравновесии по сцеплению (LD) (Пономаренко, 2018). Два аллеля различных локусов находятся в неравновесии по сцеплению, когда частота состоящего из них гаплотипа значимо отличается от частоты, ожидаемой при случайной сегрегации (Gabriel et al., 2002). Величина LD зависит от ряда факторов: величины и скорости дрейфа генов, генетических примесей в популяции, мутаций и рекомбинаций, размера популяции (Аульченко, Аксенович, 2006). Обычно LD оценивается коэффициентом неравновесности сцепления (D), однако эта мера не всегда удобна, поскольку диапазон его возможных значений зависит от частот аллелей, к которым он относится. Это затрудняет сравнение уровня неравновесия по сцеплению между разными парами аллелей. Таким образом, производится нормировка коэффициента D на основе коэффициента корреляции Пирсона r^2, который варьирует от 0 до 1. Чем ближе значение r^2 к 0, тем больше вероятность, что выявленные SNP случайны.

Для полученного общего файла, содержащего информацию о полиморфизмах для всех библиотек по всем хромосомам в формате GDS, анализируется параметр LD. Расчет производится с помощью пакета R – SNPRelate (Zheng et al., 2017), функция snpgdsLDpruning.

Для анализа главных компонент, отфильтрованных SNP, применяется пакет R – SNPRelate, для построения филогенетического дерева – тоже пакет SNPRelate, но с использованием метода иерархической кластеризации.

Системные требования и установка. Конвейер GBS-DP реализован с применением программного менеджера Snakemake v6.0.0 (Köster, Rahmann, 2012), инструмента для создания конвейеров анализа данных, реализованного на языке Python. Созданные в этой среде конвейеры можно легко масштабировать для серверных, кластерных, сетевых и облачных сред без необходимости изменять определение рабочего процесса. Snakemake совместим с системой Conda, что позволяет без труда устанавливать новые программы, необходимые для конвейера. Конвейер разработан для операционной системы Linux. Для запуска требуется минимум 10 Гб оперативной памяти (чем больше данных, тем больше требуется оперативной памяти). Для запуска конвейера в файле конфигураций, необходимо указать путь к сырым прочтениям и путь к референсному геному, после чего можно запускать программу. Код и пошаговая инструкция запуска конвейера доступны по адресу: https://github.com/artempronozin95/GBS-DP-bioinformatics-pipeline-for-genotyping-by-sequencing-data-processing/tree/main.

Данные для тестового анализа. Для тестового применения конвейера GBS-DP в настоящей работе был использован проект PRJEB39633 из базы данных European Nucleotide Archive (ENA) (Leinonen et al., 2011), который

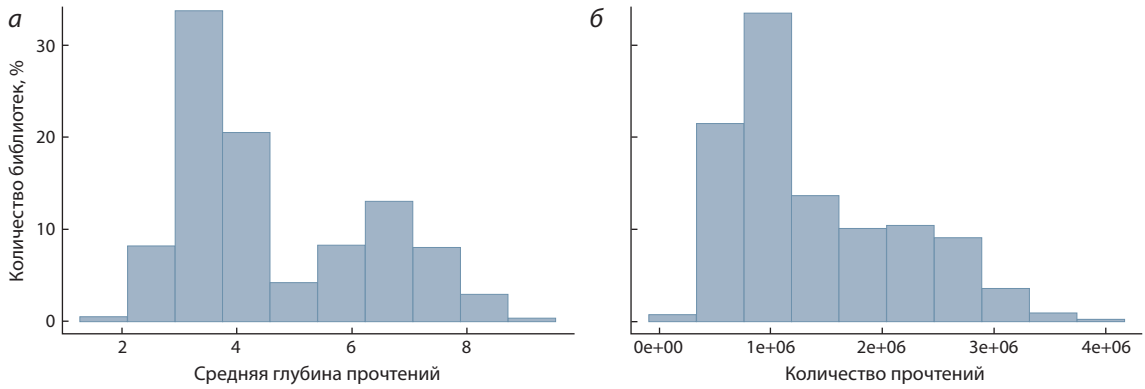


Рис. 2. Распределение средней глубины прочтений (*а*) и количества прочтений (*б*).

содержит библиотеки GBS для популяции ячменя, полученной в результате скрещивания шестирядного ячменя сорта Morex и мутантной линии *luteostrians*-P1 (*lst/LST*) (Li M. et al., 2021). Библиотеки были получены с помощью комбинации ферментов рестрикции *MspI* и *PstI* (Wendler et al., 2015). Всего проект PRJEB39633 содержит 679 библиотек для 272 генотипов; на один генотип в среднем приходится три библиотеки, поэтому перед проведением анализа прочтения библиотеки для одного генотипа объединялись.

Мы использовали референсный геном ячменя 51-й версии (IBSC_v2), загруженный с базы данных Ensembl plants (Bolser et al., 2016).

Результаты

Время, затраченное для обработки данных на различных этапах выполнения конвейера GBS-DP для разного количества библиотек ячменя (10, 50, 100, 150, 200 и 272 шт.), приведено в электронном Приложении[1]. Характеристики вычислительного узла: процессор AMD EPYC 74521, 32 ядра, объем памяти 1 Тб. Для анализа мы использовали 100 Гб оперативной памяти и 20 ядер процессора. Наибольшее время было затрачено на формирование общего файла, содержащего полиморфизмы. Однако можно заметить, что на формирование общего файла для 200 библиотек было затрачено меньше времени, чем для 150 библиотек; это связано с включением режима обработки больших данных, который ускоряет процесс расчета.

Конвейер предоставляет результаты оценки базовых характеристик секвенированных библиотек. Длина прочтения для каждой библиотеки равна 107 нк. Средняя глубина прочтения варьирует в пределах 2–8, что является допустимым значением для метода GBS (рис. 2, *а*). Более 30 % библиотек содержат свыше 1 000 000 прочтений (см. рис. 2, *б*). В среднем для одной библиотеки покрытие референсного генома ячменя (4 225 577 519 нк.) фрагментами ДНК составляет 3 % от общей длины.

Также конвейер предоставляет результаты поиска полиморфизмов между исследуемыми генотипами. Для 272 исследуемых образцов выявлено 447 409 SNP. Общее количество индел 46 557. Медиана значения транзиции/трансверсию = 1.75, что указывает на преобладание транзиций. Параметр LD (r^2) был выбран равным 0.5. После применения фильтра LD осталось 45 402 полиморфных и независимых SNP.

Распределение обнаруженных SNP по хромосомам показало, что больше SNP выявлено для хромосом 3, 6 и 7 (рис. 3). Основными результатами конвейера являются анализ главных компонент генотипов на основе выявленных SNP (рис. 4) и построение филогенетического дерева. Результаты анализа главных компонент на основе 45 402 SNP показывают, что внутри исследованной популяции на диаграмме рассеяния в пространстве двух первых компонент четко выделяется несколько кластеров (см. рис. 4). Однако суммарная доля дисперсии, приходящаяся на две эти компоненты, невелика (20 %), что может свидетельствовать об общем высоком уровне генетического разнообразия в полученной популяции растений.

Филогенетическое дерево, построенное иерархическим методом кластеризации, так же как и при кластеризации методом главных компонент, приведено на рис. 5. На дереве выделяются три больших кластера, что согласуется с данными, представленными на рис. 4.

Обсуждение

Благодаря снижению стоимости и сокращению времени, необходимого для секвенирования методом GBS, появилось множество экспериментов, проведенных этим методом. Например, база данных генетических профилей ячменя IPK Gatersleben (Milner et al., 2019) содержит 22 626 образцов, полученных методом GBS. Такое количество образцов требует быстрого и качественного способа обработки данных. На сегодняшний день уже существуют конвейеры, позволяющие обрабатывать результаты GBS. Однако, несмотря на качественно подобранные пакеты программ и возможность подстраивать параметры под исследуемые организмы, данные конвейеры имеют некоторые недостатки. Так, для GBS-SNP-CROP и TASSEL нельзя автоматизировать процесс расчета (каждый этап нужно запускать вручную), что значительно снижает скорость исследования. GB-eaSy не позволяет одновременно исследовать сразу несколько библиотек сырых прочтений. Во всех существующих конвейерах нет возможности отключения ненужного или пройденного этапа. Например, если нет возможности предоставить данные по бар-кодам для исследуемых библиотек, то ни один из перечислен-

[1] Приложение см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx23.pdf

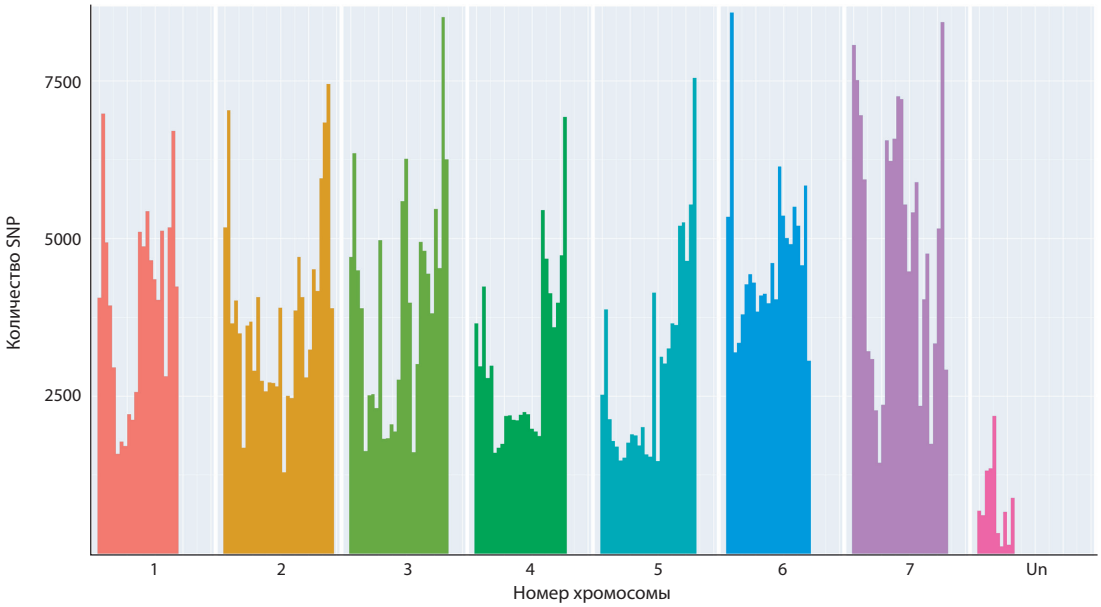


Рис. 3. Распределение выявленных SNP по хромосомам.

Ось *X* – координаты SNP на хромосомах, ось *Y* – количество SNP, соответствующих данным координатам. Шаг 10^8 нк.

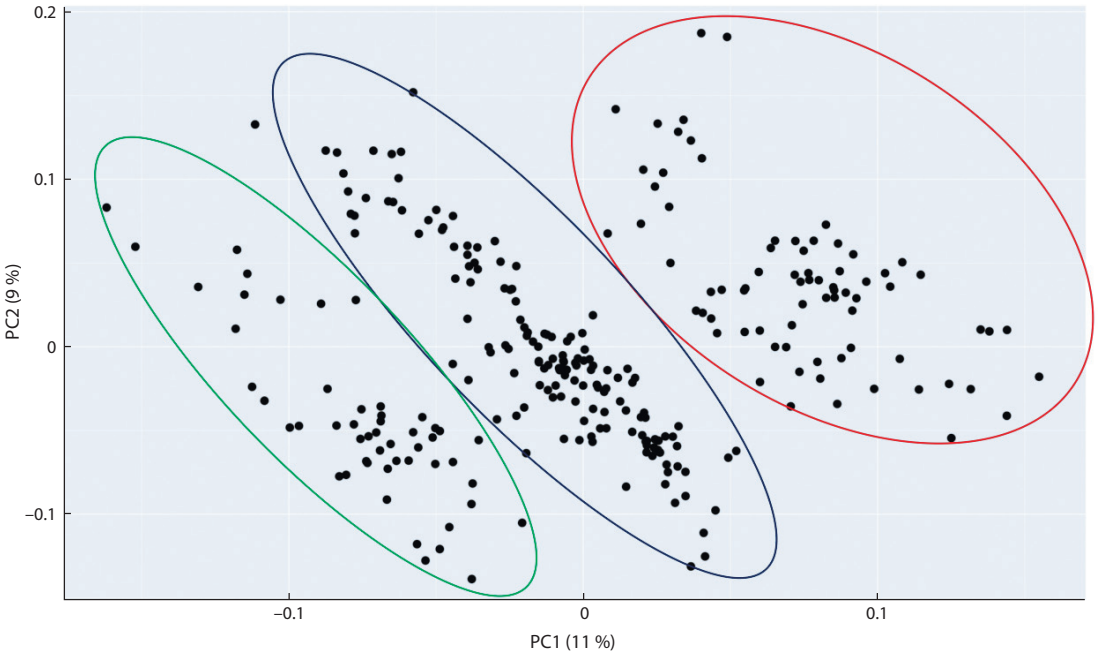


Рис. 4. Диаграмма рассеяния генотипов для популяции ячменя, полученной в результате скрещивания сорта Morex и мутантной линии *luteostrians*-P1 (*lst/LST*) для двух главных компонент, полученных при анализе генетического разнообразия конвейером GBS-DP.

В скобках рядом с названиями компонент указана доля от общей дисперсии.

ных конвейеров работать не будет. Также в большинстве конвейеров отсутствует возможность автоматической установки всех необходимых программных пакетов.

Разработанный нами конвейер основан на методе, предложенном M. Jayakodi с коллегами (Jayakodi et al., 2020). Авторы подобрали программы таким образом, чтобы предоставлять наиболее точный результат по поиску полиморфизмов. Однако этот метод хорошо применим для малых данных – до 50 библиотек. С увеличением количества библиотек увеличивается нагрузка на оперативную память и на занимаемый объем на жестком диске, что приводит к нежелательным ошибкам и прерыванию процесса расчета. Нами был предложен подход для расчета больших данных. Результаты применения данного подхода представлены в электронном Приложении и на рис. 6.

Как видно из рис. 6, предложенный нами подход значительно ускоряет процесс расчета для больших данных, однако для малых данных разница в скорости расчета не-

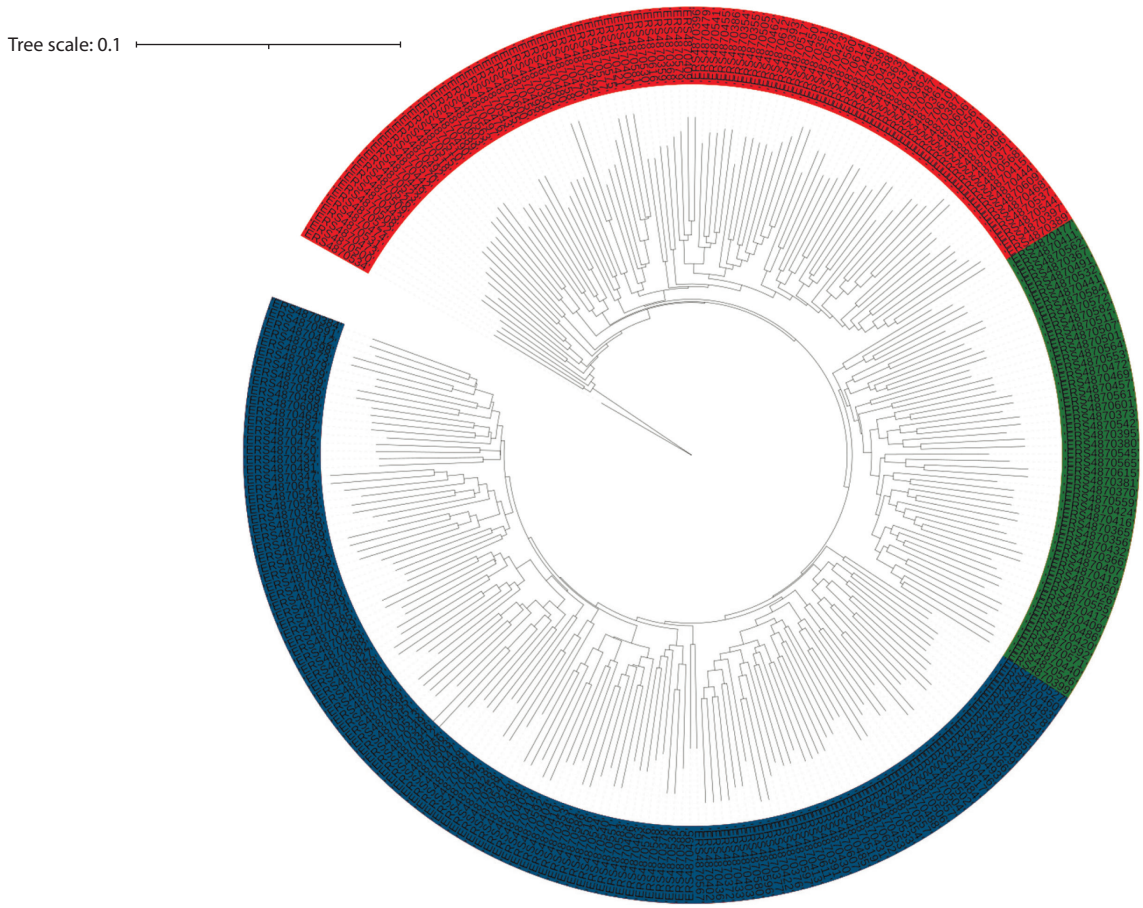


Рис. 5. Филогенетическое дерево 272 библиотек ячменя, построенное методом иерархической кластеризации.

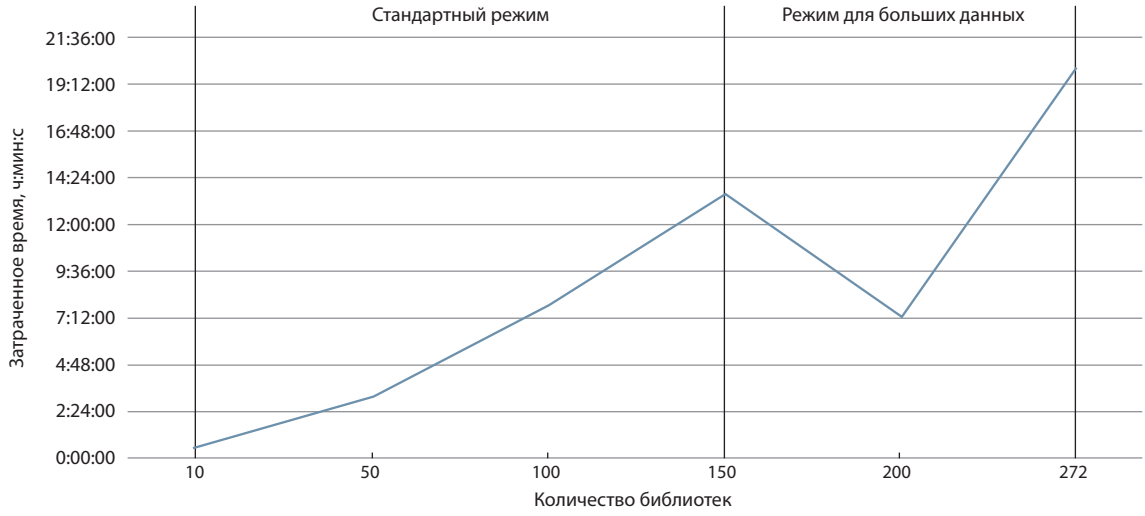


Рис. 6. Зависимость времени, затраченного на работу конвейера, от количества исследуемых библиотек.

велика. Поэтому режим активируется только на данных, общий объем найденных полиморфизмов которых превышает 500 Гб.

Разработанный нами конвейер использует программный менеджер Snakemake. Данный метод реализации автоматически учитывает выполненные задачи для каждого образца, что позволяет исключить дублирование задач, а также дает возможность возобновить процесс расчета с момента его остановки (например, вследствие ошибки). Модульная структура дает более удобный функционал манипуляции этапами конвейера (удаление, добавление, перемещение, отключение). Также Snakemake имеет возможность автоматической установки всех необходимых программ для работы конвейера.

Заключение

Методы генотипирования путем секвенирования продемонстрировали свою надежность и гибкость для ряда видов и популяций растений. Они сократили как стоимость, так и время, необходимое для секвенирования исследуемых образцов, что позволило проводить секвенирование в еще большем объеме. В настоящей работе нами был предложен биоинформатический конвейер GBS-DP, который позволяет обрабатывать данные широкомасштабного секвенирования, проведенного методом GBS. Результаты демонстрируют достаточно высокую скорость работы конвейера как для больших данных (более 400 библиотек), так и для малых (30 библиотек). Конвейер предоставляет также анализ выявленных полиморфизмов.

Список литературы / References

Аульченко Ю.С., Аксенович Т.И. Методологические подходы и стратегии картирования генов, контролирующих комплексные признаки человека. *Информ. вестн. ВОГиС*. 2006;10(1):189-202
[Aulchenko Yu.S., Aksenovich T.I. Methodological approaches and strategies for mapping genes controlling complex human traits. *Informatsionnyy Vestnik VOGiS = The Herald of Vavilov Society for Geneticists and Breeders*. 2006;10(1):189-202 (in Russian)]

Канукова К.Р., Газаев И.Х., Сабанчиева Л.К., Боготова З.И., Аппаев С.П. ДНК-маркеры в растениеводстве. *Изв. Кабардино-Балкарского науч. центра РАН*. 2019;6(92):220-232. DOI 10.35330/1991-6639-2019-6-92-220-232
[Kanukova K.R., Gazaev I.Kh., Sabanchieva L.K., Bogotova Z.I., Appaev S.P. DNA markers in crop production. *Izvestiya Kabardino-Balkarskogo Nauchnogo Tsentra RAN = News of the Kabardin-Balkar Scientific Center of RAS*. 2019;6(92):220-232. DOI 10.35330/1991-6639-2019-6-92-220-232 (in Russian)]

Пономаренко И.В. Отбор полиморфных локусов для анализа ассоциаций при генетико-эпидемиологических исследованиях. *Науч. результаты биомед. исследований*. 2018;4(2):40-54. DOI 10.18413/2313-8955-2018-4-2-0-5
[Ponomarenko I.V. Selection of polymorphic loci for association analysis in genetic-epidemiological studies. *Nauchnye Rezultaty Biomeditsynskikh Issledovaniy = Research Results in Biomedicine*. 2018;4(2):40-54. DOI 10.18413/2313-8955-2018-4-2-0-5 (in Russian)]

Сухарева А.С., Кулуев Б.Р. ДНК-маркеры для генетического анализа сортов культурных растений. *Биомика*. 2018;10(1):69-84. DOI 10.31301/2221-6197.bmcs.2018-15
[Sukhareva A.S., Kuluev B.R. DNA markers for genetic analysis of crops. *Biomika = Biomics*. 2018;10(1):69-84. DOI 10.31301/2221-6197.bmcs.2018-15 (in Russian)]

Хлесткина Е.К. Молекулярные маркеры в генетических исследованиях и в селекции. *Вавиловский журнал генетики и селекции*. 2013;17(4/2):1044-1054
[Khlestkina E.K. Molecular markers in genetic studies and breeding. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):1044-1054 (in Russian)]

Bimber B.N., Raboin M.J., Letaw J., Nevonen K.A., Spindel J.E., McCouch S.R., Cervera-Juanes R., Spindel E., Carbone L., Ferguson B., Vinson A. Whole-genome characterization in pedigreed non-human primates using genotyping-by-sequencing (GBS) and imputation. *BMC Genomics*. 2016;17(1):676. DOI 10.1186/s12864-016-2966-x

Bolser D., Staines D.M., Pritchard E., Kersey P. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In: Edwards D. (Ed.) Plant Bioinformatics. Methods in Molecular Biology. Vol. 1374. New York: Humana Press, 2016;115-140. DOI 10.1007/978-1-4939-3167-5_6

Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M., Li H. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):giab008. DOI 10.1093/gigascience/giab008

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. 2011;6(5):e19379. DOI 10.1371/journal.pone.0019379

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., Altshuler D. The structure of haplotype blocks in the human genome. *Science*. 2002;296(5576):2225-2229. DOI 10.1126/science.1069424

Glaubitz J.C., Casstevens T.M., Lu F., Harriman J., Elshire R.J., Sun Q., Buckler E.S. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*. 2014;9(2):e90346. DOI 10.1371/journal.pone.0090346

Jayakodi M., Padmarasu S., Haberer G., Bonthala V.S., Gundlach H., Monat C., Lux T., Kamal N., Lang D., Himmelbach A., Ens J., Zhang X.Q., Angessa T.T., Zhou G., Tan C., Hill C., Wang P., Schreiber M., Boston L.B., Plott C., Jenkins J., Guo Y., Fiebig A., Budak H., Xu D., Zhang J., Wang C., Grimwood J., Schmutz J., Guo G., Zhang G., Mochida K., Hirayama T., Sato K., Chalmers K.J., Langridge P., Waugh R., Pozniak C.J., Scholz U., Mayer K.F.X., Spannagl M., Li C., Mascher M., Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*. 2020;588(7837):284-289. DOI 10.1038/s41586-020-2947-8

Köster J., Rahmann S. Snakemake – a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520-2522. DOI 10.1093/bioinformatics/bts480

Leinonen R., Akhtar R., Birney E., Bower L., Cerdeno-Tárraga A., Cheng Y., Cleland I., Faruque N., Goodgame N., Gibson R., Hoad G., Jang M., Pakseresht N., Plaister S., Radhakrishnan R., Reddy K., Sobhany S., Ten Hoopen P., Vaughan R., Zalunin V., Cochrane G. The European nucleotide archive. *Nucleic Acids Res.* 2011;39(Database issue):D28-D31. DOI 10.1093/nar/gkq967

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. 2013. DOI 10.48550/arXiv.1303.3997

Li M., Guo G., Pidon H., Melzer M., Prina A.R., Börner T., Stein N. ATP-dependent *Clp* protease subunit *C1*, *HvClpC1*, is a strong candidate gene for barley variegation mutant *luteostrians* as revealed by genetic mapping and genomic re-sequencing. *Front. Plant Sci.* 2021;12:664085. DOI 10.3389/fpls.2021.664085

Lu F., Lipka A.E., Glaubitz J., Elshire R., Cherney J.H., Casler M.D., Buckler E.S., Costich D.E. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 2013;9(1):e1003215. DOI 10.1371/journal.pgen.1003215

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17(1):10-12. DOI 10.14806/ej.17.1.200

Melo A.T., Bartaula R., Hale I. GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics*. 2016;17(1):29. DOI 10.1186/s12859-016-0879-y

Milner S.G., Jost M., Taketa S., Mazón E.R., Himmelbach A., Oppermann M., Weise S., Knüpffer H., Basterrechea M., König P., Schüler D., Sharma R., Pasam R.K., Rutten T., Guo G., Xu D., Zhang J., Herren G., Müller T., Krattinger S.G., Keller B., Jiang Y., González M.Y., Zhao Y., Habekuß A., Färber S., Ordon F., Lange M., Börner A., Graner A., Reif J.C., Scholz U., Mascher M., Stein N. Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* 2019;51(2):319-326. DOI 10.1038/s41588-018-0266-x

Monat C., Schreiber M., Stein N., Mascher M. Prospects of pan-genomics in barley. *Theor. Appl. Genet.* 2019;132(3):785-796. DOI 10.1007/s00122-018-3234-z

Narum S.R., Buerkle C.A., Davey J.W., Miller M.R., Hohenlohe P.A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* 2013;22(11):2841-2847. DOI 10.1111/mec.12350

Peterson G.W., Dong Y., Horbach C., Fu Y.-B. Genotyping-by-sequencing for plant genetic diversity analysis: a lab guide for SNP genotyping. *Diversity*. 2014;6(4):665-680. DOI 10.3390/d6040665

Poland J., Endelman J., Dawson J., Rutkoski J., Wu S., Manes Y., Dreisigacker S., Crossa J., Sánchez-Villeda H., Sorrells M., Jannink J.-L. Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*. 2012;5(3):103-113. DOI 10.3835/plantgenome2012.06.0006

Rajendran N.R., Qureshi N., Pourkheirandish M. Genotyping by sequencing advancements in barley. *Front. Plant Sci.* 2022;13:931423. DOI 10.3389/fpls.2022.931423

Scheben A., Batley J., Edwards D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.* 2017;15(2):149-161. DOI 10.1111/pbi.12645

Torkamaneh D., Laroche J., Bastien M., Abed A., Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. *BMC Bioinformatics*. 2017;18(1):5. DOI 10.1186/s12859-016-1431-9

Wang N., Yuan Y., Wang H., Yu D., Liu Y., Zhang A., Gowda M., Nair S.K., Hao Z., Lu Y., San Vicente F., Prasanna B.M., Li X., Zhang X. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 2020;10(1):16308. DOI 10.1038/s41598-020-73321-8

Wendler N., Mascher M., Himmelbach A., Johnston P., Pickering R., Stein N. Bulbosum to go: a toolbox to utilize *Hordeum vulgare/bulbosum* introgressions for breeding and beyond. *Mol. Plant*. 2015;8(10):1507-1519. DOI 10.1016/j.molp.2015.05.004

Wickland D.P., Battu G., Hudson K.A., Diers B.W., Hudson M.E. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*. 2017;18:586. DOI 10.1186/s12859-017-2000-6

Yao Z., You F.M., N'Diaye A., Knox R.E., McCartney C., Hiebert C.W., Pozniak C., Xu W. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics.* 2020;21(1):360. DOI 10.1186/s12859-020-03704-1

Zheng X., Gogarten S.M., Lawrence M., Stilp A., Conomos M.P., Weir B.S., Laurie C., Levine D. SeqArray – a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*. 2017;33(15):2251-2257. DOI 10.1093/bioinformatics/btx145

ORCID ID

A.Yu. Pronozin orcid.org/0000-0002-3011-6288
E.A. Salina orcid.org/0000-0001-8590-847X
D.A. Afonnikov orcid.org/0000-0001-9738-1409

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

Прозрачность финансовой деятельности. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 21.07.2023. После доработки 08.09.2023. Принята к публикации 09.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):746-754
DOI 10.18699/VJGB-23-87

Перевод на английский язык https://vavilov.elpub.ru/jour

Центральный регуляторный контур генной сети морфогенеза механорецепторов дрозофилы: анализ *in silico*

Т.А. Бухарина[1, 2] ✉, В.П. Голубятников[3], 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
[3] Институт математики им. С.Л. Соболева Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ bukharina@bionet.nsc.ru

Аннотация. Выявление механизмов генетического контроля формирования пространственных структур остается одной из актуальных задач биологии развития. Для ее решения используются как экспериментальные, так и теоретические подходы и методы, в том числе методология генных сетей, а также методы математического и компьютерного моделирования. Реконструкция и анализ генных сетей, обеспечивающих становление признака, позволяют интегрировать существующие экспериментальные данные, выявить ключевые звенья и внутрисетевые связи, обеспечивающие функционирование сетей. Для получения динамических характеристик исследуемых систем, предсказания их состояния и поведения привлекаются методы математического и компьютерного моделирования. Одним из примеров пространственной морфологической структуры является щетиночный рисунок дрозофилы со строго определенным расположением на голове и теле мухи его составляющих – механорецепторов (внешних сенсорных органов). Механорецептор развивается из единственной родительской клетки (РКСО), которая выделяется из клеток эктодермы имагинального диска. Ее отличает от окружения наибольшее содержание пронейральных белков (ASC) – продуктов комплекса пронейральных генов *achaete-scute* (*AS-C*). Статус РКСО обеспечивается реконструированной нами ранее генной сетью, ключевым объектом которой является комплекс генов *AS-C*. Контроль активности комплекса осуществляется ее подсетью – центральным регуляторным контуром в составе семи генов (*AS-C*, *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), *extramacrochaete* (*emc*)) и одноименных белков. Кроме того, в состав центрального регуляторного контура входят вспомогательные белки Daughterless (DA), Groucho (GRO), Ubiquitin (UB) и Seven-in-absentia (SINA). В работе приведены результаты компьютерного моделирования различных режимов функционирования контура. Показано, что клетка детерминируется как РКСО при повышении содержания ASC примерно в два с половиной раза относительно уровня в клетках окружения. Выявлена иерархия влияния мутаций в генах контура на динамику накопления белков ASC. Наиболее значим главный компонент центрального регуляторного контура – *AS-C*. Мутации, снижающие содержание ASC более чем на 40 %, приводят к запрету выделения родительской клетки сенсорного органа.
Ключевые слова: центральный регуляторный контур; генная сеть; математическая модель; компьютерное моделирование; дрозофила; *achaete-scute* комплекс; мутации.

Для цитирования: Бухарина Т.А., Голубятников В.П., Фурман Д.П. Центральный регуляторный контур генной сети морфогенеза механорецепторов дрозофилы: анализ *in silico*. *Вавиловский журнал генетики и селекции*. 2023;27(7):746-754. DOI 10.18699/VJGB-23-87

The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis

T.A. Bukharina[1, 2] ✉, V.P. Golubyatnikov[3], 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
[3] Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ bukharina@bionet.nsc.ru

Abstract. Identification of the mechanisms underlying the genetic control of spatial structure formation is among the relevant tasks of developmental biology. Both experimental and theoretical approaches and methods are used for this purpose, including gene network methodology, as well as mathematical and computer modeling. Reconstruction and analysis of the gene networks that provide the formation of traits allow us to integrate the existing experimental data and to identify the key links and intra-network connections that ensure the function of networks. Mathematical and computer modeling is used to obtain the dynamic characteristics of the studied systems and to predict their state and behavior. An example of the spatial morphological structure is the Drosophila bristle pattern with a strictly defined arrangement of its components – mechanoreceptors (external sensory organs) – on the head and body. The mechanoreceptor develops from a single sensory organ parental cell (SOPC), which is isolated from the ectoderm cells of the

imaginal disk. It is distinguished from its surroundings by the highest content of proneural proteins (ASC), the products of the *achaete-scute* proneural gene complex (*AS-C*). The SOPC status is determined by the gene network we previously reconstructed and the *AS-C* is the key component of this network. *AS-C* activity is controlled by its subnetwork – the central regulatory circuit (CRC) comprising seven genes: *AS-C*, *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), and *extramacrochaete* (*emc*), as well as their respective proteins. In addition, the CRC includes the accessory proteins Daughterless (DA), Groucho (GRO), Ubiquitin (UB), and Seven-in-absentia (SINA). The paper describes the results of computer modeling of different CRC operation modes. As is shown, a cell is determined as an SOPC when the ASC content increases approximately 2.5-fold relative to the level in the surrounding cells. The hierarchy of the effects of mutations in the CRC genes on the dynamics of ASC protein accumulation is clarified. *AS-C* as the main CRC component is the most significant. The mutations that decrease the ASC content by more than 40 % lead to the prohibition of SOPC segregation.

Key words: central regulatory circuit; gene network; mathematical model; computer modeling; drosophila; *achaete-scute* complex; mutations.

For citation: Bukharina T.A., Golubyatnikov V.P., Furman D.P. The central regulatory circuit in the gene network controlling the morphogenesis of Drosophila mechanoreceptors: an *in silico* analysis. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):746-754. DOI 10.18699/VJGB-23-87

Введение

Современные представления о контроле биологических процессов, в том числе дифференцировки клеток, роста и развития организмов, создание пространственных структур и др. объединены в концепции генных сетей. В соответствии с этой концепцией генные сети – это молекулярно-генетические системы, обеспечивающие формирование всех фенотипических характеристик организмов (молекулярных, биохимических, структурных, морфологических, этологических, физиологических, когнитивных и др.) на основе информации, закодированной в их геномах. Согласно определению, данному в работе (Колчанов и др., 2013), генные сети представляют собой группы координированно функционирующих генов, взаимодействующих друг с другом как через свои первичные продукты (РНК и белки), так и через разнообразные метаболиты и другие вторичные продукты функционирования генных сетей.

Реконструкция генных сетей осуществляется на основе анализа экспериментальных данных и дает наиболее полное и систематизированное описание рассматриваемой биологической системы или процесса (Schlitt et al., 2003; Zhu et al., 2007; Emmert-Streib, Glazko, 2011; Chasman et al., 2016). Важным атрибутом генных сетей являются регуляторные контуры, позволяющие обеспечивать правильное их функционирование и выполнение программы формирования фенотипического признака.

Для выяснения структурно-функциональной организации генных сетей, архитектуры внутрисетевых связей, выявления ключевых элементов и модулей, закономерностей динамики их функционирования и эволюции широко используют методы математического и компьютерного моделирования, позволяющие получить наиболее полное представление об устройстве и поведении сетей.

Одним из примеров генных сетей, отвечающих за развитие упорядоченных пространственных структур в ходе онтогенеза, являются реконструированные нами ранее генные сети Neurogenesis:prepattern, Neurogenesis:determination и Neurogenesis:asymmetric division, совместно обеспечивающие определенную композицию механорецепторов – органов периферической нервной системы – на голове и теле дрозофилы (Furman, Bukharina, 2022). Анализ сетей выявил важнейшее связующее звено, управляющее их функционированием, – центральный регуляторный контур (ЦРК). От корректной работы ЦРК в рамках сети Neurogenesis:determination зависит реализация ключевого события морфогенеза каждого механорецептора, состоящего в определении единственной родительской клетки сенсорного органа (РКСО), которая обособляется в пределах пронейрального кластера – группы эпидермальных клеток имагинального диска (Furman, Bukharina, 2022). Родительская клетка отличается от окружения содержанием пронейральных белков ASC, кодируемых одноименным генным комплексом (*achaete-scute complex*, *AS-C*) (Reeves, Posakony, 2005). Повышенное содержание ASC является фактором детерминации нейральной судьбы клетки. Именно регуляцию наработки этих белков до уровня, необходимого для приобретения клеткой статуса РКСО, и осуществляет ЦРК, обеспечивая как развитие отдельного механорецептора, так и формирование их совокупности – так называемого щетиночного рисунка (Furman, Bukharina, 2022).

К сожалению, несмотря на многолетнюю историю изучения, морфогенез механорецепторов все еще далек от исчерпывающего описания и в основном охарактеризован лишь качественно: известны участники процесса (гены и белки) и сформировано общее представление об их взаимодействии, тогда как большинство его количественных параметров, как и относительный вклад участвующих генов, экспериментально не установлены. Следует заметить, что с неполнотой данных исследователи биологических систем встречаются достаточно часто, и решить проблему их восполнения можно с помощью методов математического и компьютерного моделирования. Модель с правильно подобранными параметрами не только позволяет оценить текущее состояние системы или протекания процесса, но и обладает прогностической ценностью. Численные эксперименты, проводимые на основе математических моделей, дают возможность изучить режимы функционирования системы, а также составить прогноз ее будущих состояний и, изменяя параметры или добавляя новые допущения, предсказать ее новые функции. Во многих случаях моделирование становится единственным способом понять происходящие в системе процессы, характеристики которых невозможно измерить непосредственно в биологическом эксперименте.

Моделирование морфогенеза механорецепторов на этапе выделения РКСО из клеток пронейрального кластера предпринималось и ранее, однако при этом авторы ограничивались обобщенными характеристиками и общими схемами внутри- и межклеточных взаимодействий групп генов либо без детализации их состава и конкретизации вклада отдельных участников, либо с минимальным их уточнением (Marnellos, Mjolsness, 1998; Meir et al., 2002; Ghysen, Thomas, 2003; Hsu et al., 2006; Corson et al., 2017; Yasugi, Sato, 2022). До сих пор отсутствует целостное представление о механизмах внутриклеточных взаимодействий при формировании РКСО, не определены количественные характеристики для содержания белков ASC, критичные для детерминации нейральной судьбы клетки, и неясна степень влияния составляющих ЦРК элементов на экспрессию генов *AS-C*.

В задачу настоящей работы входило построение математической модели функционирования ЦРК с учетом роли входящих в него генов, определяющих динамику содержания белков ASC, детально описывающей внутриклеточные события в презумптивной РКСО, и проведение компьютерных экспериментов для проверки устойчивости модели и ее соответствия экспериментальным данным.

Материал и методы

Объектом моделирования являлся ЦРК. Схема ЦРК представлена на рис. 1. В состав контура, помимо пронейральных генов *AS-C* и кодируемых ими белков ASC, входят гены *hairy*, *senseless* (*sens*), *charlatan* (*chn*), *scratch* (*scrt*), *phyllopod* (*phyl*), *extramacrochaete* (*emc*) и соответствующие им белки. Кроме того, ЦРК включает белки Daughterless (DA), Groucho (GRO), Ubiquitin (UB) и Seven-in-absentia (SINA). Все компоненты связаны с *AS-C* отношениями активации–репрессии.

Содержание пронейральных белков ASC в родительской клетке устанавливается через авто- и трансрегуляцию активности генов *AS-C*. Активирующая авторегуляция осуществляется гетеродимерами ASC/DA, а репрессорная – гетеродимерами ASC/EMC. Трансрегуляция генов комплекса с активирующим эффектом осуществляется белками Senseless и Charlatan, а с негативным эффектом – комплексами Hairy/GRO и ASC/EMC (Cabrera, Alonso, 1991; Van Doren et al., 1992, 1994; Cabrera et al., 1994; Vaessin et al., 1994; Nolo et al., 2000; Escudero et al., 2005) (см. рис. 1).

Существуют также дополнительные механизмы, позволяющие исключить репрессорное воздействие Hairy/GRO и ASC/EMC на *AS-C*. Так, активация гена *scratch* гетеродимерами ASC/DA влечет репрессию транскрипционной активности *hairy* (Roark et al., 1995) и, как следствие, усиление экспрессии *AS-C*. Активация гена *chn* приводит к репрессии транскрипции генов *hairy* и *emc* (Yamasaki et al., 2011) и к тому же эффекту – повышению экспрессии *AS-C* (см. рис. 1).

Экспрессия генов *sens*, *scrt* и *chn* и, соответственно, наработка одноименных белков регулируется гетеродимерами ASC/DA, инициирующими их транскрипцию (Cabrera, Alonso, 1991; Vaessin et al., 1994; Nolo et al., 2000; Escudero et al., 2005) (см. рис. 1).

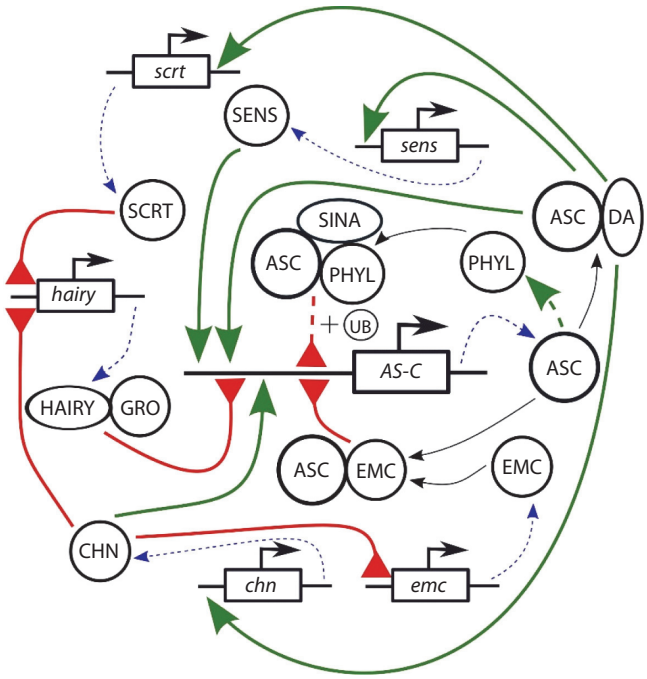


Рис. 1. Схема центрального регуляторного контура генных сетей, поддерживающих развитие макрохет дрозофилы.

AS-C – комплекс генов *achaete-scute*, ASC – белки *achaete-scute* комплекса, da – daughterless, gro – groucho, sens – senseless, emc – extramacrochaete, chn – charlatan, scrt – scratch. Стрелками зеленого цвета показаны активаторные воздействия (сплошная линия – прямые, штриховая – опосредованные), стрелки красного цвета с обрубленными концами обозначают репрессорные воздействия (сплошная линия – прямые, штриховая – опосредованные). Ранее опубликованная схема (Golubyatnikov et al., 2015) была дополнена системой деградации белков ASC.

Для функционирования ЦРК необходимы и участники системы деградации белков – убиквитин (Ubiquitin – UB) и E3 убиквитин лигаза Seven-in-absentia (SINA), а также адапторный белок Phyllopod (PHYL) (Pi et al., 2001; Chang et al., 2008).

Модель. Разработанная динамическая модель регуляции активности комплекса *achaete-scute* представлена в виде системы обыкновенных дифференциальных уравнений кинетического типа (1) (Bukharina et al., 2020):

$$\begin{aligned}
\frac{dx}{dt} &= k_x \frac{\sigma_1(D\cdot x)+\sigma_4(z)+\sigma_6(w)}{(1+G\cdot y)(1+E\cdot x)} - (1+p(t-\tau)\cdot U\cdot S)m_x\cdot x,\\
\frac{dy}{dt} &= k_y \frac{C_y}{(d_1+u)(d_2+w)} - m_y\cdot y,\\
\frac{dE}{dt} &= k_e \frac{C_e}{(d_3+w)(d_2+w)} - m_e\cdot E,\\
\frac{dz}{dt} &= k_z s_4(D\cdot x) - m_z\cdot z, \qquad (1)\\
\frac{du}{dt} &= k_u s_5(D\cdot x) - m_u\cdot u,\\
\frac{dw}{dt} &= k_w s_6(D\cdot x) - m_w\cdot w,\\
\frac{dp}{dt} &= k_p \frac{s_7(D\cdot x)\cdot h(t-\tau)\cdot(t-\tau)^2}{(L+h(t-\tau)\cdot(t-\tau)^2)(1+G\cdot y)(1+E\cdot x)} - m_p\cdot p.
\end{aligned}$$

Функции в этой системе описывают содержание белков ЦРК в клетке: $x(t)$ – содержание ASC, $y(t)$ – Hairy,

$E(t)$ – Extramacrochaete, $z(t)$ – Senseless, $u(t)$ – Scratch, $w(t)$ – Charlatan, $p(t)$ – Phyllopod.

В последнем уравнении системы (1) сглаженная функция Хевисайда $h(t - \tau)$ описывает процесс отложенного появления белка PHYL с временной задержкой τ.

Для учета мутационных изменений генов, входящих в ЦРК, в модели задаются неотрицательные коэффициенты мутаций k_x, k_y, k_e и т. д., отражающие степень влияния мутации на наработку соответствующего белка. Значение каждого из них не может превосходить 1; $k = 1$ соответствует нормальному функционированию гена, тогда как $k = 0$ означает полную инактивацию гена и отсутствие кодируемого им белка.

Параметры x_0, y_0, z_0, u_0, w_0, p_0 и E_0 соответствуют содержанию белков ASC, Hairy, SENS, SCRT, CHN, PHYL и EMC соответственно в исходном состоянии моделируемого ЦРК, когда пронейральный кластер уже сформирован, во всех составляющих его клетках наблюдается экспрессия генов *AS-C*, и все они еще имеют равные нейральные потенции.

Значения параметров D, G, S, U в системе (1) предполагаются постоянными, поскольку концентрации описываемых ими белков Daughterless (DA), Groucho (GRO), Seven-in-absentia (SINA) и Ubiquitin (UB) практически не меняются в процессе формирования родительской клетки. Постоянными предполагаются и параметры C_y и C_e, d_1, d_2, d_3.

Положительные коэффициенты m_x, m_y, m_e, m_z, m_u, m_w, m_p учитывают скорости деградации соответствующих белков.

Положительное слагаемое во втором уравнении системы (1) описывает отрицательные связи SCRT–Hairy и CHN–Hairy (см. рис. 1). Сигмоидные функции σ_l, $l = 1, 4, 6$ в первом уравнении системы (1), и сигмоидные функции s_i, $i = 4, 5, 6, 7$ в соответственно четвертом, пятом, шестом и седьмом уравнениях этой системы отражают положительные связи, изображенные на рис. 1 (зеленые стрелки):

$$\sigma_l(q) = \frac{a_l q^{n_l}}{b_l + q^{n_l}},$$

$$s_i(q) = \frac{\alpha_i q^{\nu_i}}{\beta_i + q^{\nu_i}}.$$

Здесь α_i, β_i, ν_i и a_l, b_l, n_l – положительные параметры, $q \geq 0$ (Bukharina et al., 2015).

В модели предусмотрена возможность выбора временного интервала функционирования ЦРК (T) и времени появления в клетке белка PHYL (τ). Функционирование ЦРК продолжается до деления клетки, поэтому время T напрямую зависит от τ: чем позднее появляется PHYL, тем позднее клетка вступает в деление и тем дольше функционирует ЦРК. Начальное условие для уравнения с запаздыванием для функции $p(t)$ ($0 \leq \tau \leq t$).

Программное обеспечение. Для проведения численных экспериментов с рассматриваемой моделью ЦРК и визуализации результатов был разработан программный комплекс на основе пакета Shiny, предназначенного для создания на базе языка программирования R интерактивных веб-приложений с графическим интерфейсом пользователя (https://shiny.rstudio.com/), и пакета deSolve (http://desolve.r-forge.r-project.org/), включающего большое количество методов интегрирования дифференциальных уравнений, в том числе с запаздывающими аргументами, и метод lsoda, автоматически переключающийся между жесткими и нежесткими системами, что позволяет оптимально сочетать точность вычислений и общее время счета.

Разработанное веб-приложение (https://gene-nets-simulation.shinyapps.io/crc-asc-modeler/) позволяет исследовать режимы функционирования ЦРК при различных параметрах системы (1), выдавая результаты в виде графиков. Параметры системы задаются на основе существующих данных биологических экспериментов.

Результаты и обсуждение

Рассмотрим результаты моделирования различных режимов функционирования ЦРК.

Моделирование функционирования ЦРК в презумптивной родительской клетке механорецептора при отсутствии мутаций входящих в него генов

На рис. 2 представлен результат компьютерного моделирования функционирования ЦРК в будущей РКСО в норме (отсутствие мутаций во всех входящих в ЦРК генах). Параметры системы (1) подбирались с учетом имеющихся в литературе экспериментальных данных (Reeves, Posakony, 2005; Chang et al., 2008, Giri et al., 2022):

$$\begin{aligned}
&D = 1.6;\ G = 1;\ m_x = 0.3;\ U = 1.1;\ S = 5.5;\\
&a_1 = 2.9;\ n_1 = 1;\ b_1 = 1;\ a_4 = 5.8;\ n_4 = 1;\ b_4 = 5.6;\\
&a_6 = 6;\ n_6 = 1;\ b_6 = 5.7;\\
&C_y = 14.1;\ d_1 = 4.1;\ d_2 = 4.7;\ m_y = 0.5;\\
&C_e = 2.9;\ d_3 = 7.5;\ m_e = 0.4;\\
&\alpha_4 = 3;\ \nu_4 = 1.9;\ \beta_4 = 1.2;\ m_z = 1.6;\\
&\alpha_5 = 14.8;\ \nu_5 = 1.1;\ \beta_5 = 14.8;\ m_u = 2.3;\\
&\alpha_6 = 2;\ \nu_6 = 1;\ \beta_6 = 1;\ m_w = 1;\\
&\alpha_7 = 4.5;\ \nu_7 = 3.1;\ \beta_7 = 0.5;\ m_p = 0.6;\ L = 1.1;\\
&x_0 = 0.8;\ y_0 = 1.6;\ E_0 = 1.1;\ z_0 = 0.4;\ u_0 = 0;\ w_0 = 0;\ p_0 = 0;\\
&T = 28;\ \tau = 12.
\end{aligned} \qquad (2)$$

Коэффициенты k во всех уравнениях системы (1) равны 1.

Известно, что детерминация РКСО для механорецепторов различной локализации занимает разное время (Cubas et al., 1991; Huang et al., 1991; Usui, Kimura, 1993). Временной отрезок $T = 28$ ч выбран как близкий к максимальному интервал, необходимый для определения нейральной судьбы клетки (Huang et al., 1991). Предполагается, что функционирование центрального регуляторного контура начинается еще на стадии формирования пронейральных кластеров, за 35–40 ч до образования пупариума, когда в клетках имагинального диска впервые отмечается экспрессия генов *AS-C* (Cubas et al., 1991; Skeath, Carroll, 1991). За начало отсчета принимается момент, когда пронейральный кластер уже сформирован, во всех составляющих его клетках наблюдается экспрессия генов *AS-C*, и все они еще имеют равные нейральные потенции.

Характер изменения содержания белков ASC на графике (см. рис. 2) качественно соответствует картине, наблюдаемой в эксперименте (Reeves, Posakony, 2005; Chang et al., 2008). Известно, что содержание белков ASC, по-

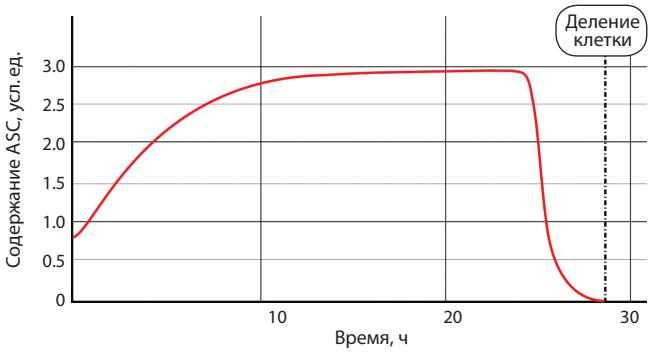


Рис. 2. Динамика содержания белков ASC в презумптивной родительской клетке механорецептора в норме.

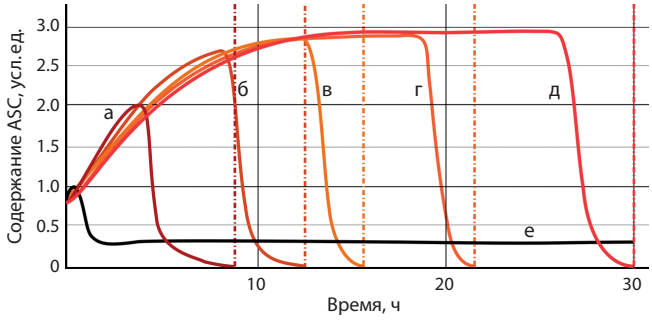


Рис. 3. Динамика содержания белков ASC в презумптивной родительской клетке механорецептора для различных временных параметров.

Значения параметров для графиков а–д приведены в тексте. Для графика е τ = 0, *T* = 30 ч. Вертикальными штриховыми линиями обозначены моменты деления клеток.

степенно нарастая, достигает некоторого критического уровня, после чего судьба клетки определяется однозначно – она становится РКСО. В приведенном численном эксперименте мы получили плавное увеличение содержания белков в течение приблизительно 10 ч до уровня, превышающего исходный примерно в 3.7 раза – от 0.8 до 2.95.

Достигнув максимума, содержание белков ASC через некоторое время начинает снижаться и к моменту начала деления РКСО падает практически до нуля. Этот процесс обусловлен запуском дополнительного механизма регуляции, связанного с деградацией белков ASC (Chang et al., 2008). При выбранных параметрах модель прогнозирует начало резкого снижения содержания ASC примерно через 15 ч с достижением нулевых значений в течение последующих приблизительно 3 ч.

Существенно, что на отрезке времени, ограниченном моментом деления клетки, модель исключает возможность возникновения циклических процессов: это означает, что детерминация нейральной судьбы клетки является необратимой. Это также соответствует известным литературным данным (Reeves, Posakony, 2005; Chang et al., 2008).

Согласно (Huang et al., 1991; Audibert et al., 2005; Kawamori et al., 2013), выделение РКСО из пронейральных кластеров для механорецепторов различной локализации в норме занимает от 9–12 до 28–30 ч. При этом деление РКСО всех механорецепторов происходит более или менее синхронно и приходится на 0–3 ч после окукливания (Huang et al., 1991; Ayeni et al., 2016).

С целью проверки устойчивости модели к изменению временных отрезков, на протяжении которых в клетках пронейрального кластера происходит накопление белков ASC, необходимое для достижения статуса РКСО и ее перехода к делению (от 9 до 30 ч), была проведена серия дополнительных численных экспериментов. При этом значение параметра τ (время появления белка PHYL, критичное для перехода клетки к делению) изменялось таким образом, чтобы значения параметра *T* (время перехода РКСО к делению) находились в промежутке от 9 до 30 ч:

а) $T = 9$; $\tau = 2.1$;
б) $T = 18$; $\tau = 4$;
в) $T = 18$; $\tau = 6$;
г) $T = 22$; $\tau = 9$;
д) $T = 28$; $\tau = 12$.

Дополнительно были проведены эксперименты, в которых значение τ принималось равным 0 (т. е. белок PHYL появлялся одномоментно с белками ASC), а параметр *T* выбирался произвольно и был равен или превышал 30 ч. Прочие параметры модели в экспериментах сохранялись неизменными и соответствовали набору параметров (2).

Графики, иллюстрирующие динамику содержания белков в родительской клетке механорецептора при выбранных временных параметрах, приведены на рис. 3. Видно, что профили графиков а–д (кривые различных оттенков красного цвета) сходны между собой и с графиком на рис. 2. Различаются кривые только продолжительностью фазы, при которой содержание белка ASC находится на максимальном уровне. Существенно, что характер кривой сохраняется в выбранном диапазоне значений τ и соответствующих значений *T*, что свидетельствует об устойчивости предложенной модели функционирования ЦРК. Для случая τ = 0, имитирующего эффект отсутствия задержки появления PHYL – участника деградации белков ASC, вид графика (кривая е черного цвета на рис. 3) значительно отличается от остальных. Вначале наблюдается небольшой рост содержания ASC (всего в пределах 16–17 % от первоначального уровня), а затем падение (примерно вполовину первоначального значения) с выходом на плато на низком уровне, хотя и отличном от 0, но явно недостаточном для детерминации клетки как РКСО.

Полученный результат косвенно подтверждает выдвинутое ранее предположение, что именно отсроченное появление белка PHYL является необходимым условием детерминации родительской клетки (Furman, Bukharina, 2022).

Рассматриваемая модель позволяет получить представление о динамике содержания ASC в презумптивной РКСО. Манипулируя параметром τ, можно оценить минимальное превышение содержания в ней белков ASC относительно уровня в окружающих клетках, необходимое и достаточное для приобретения нейрального статуса. При этом необходимо учитывать экспериментально установленный факт, что на этот процесс отводится не менее 9 ч (Huang et al., 1991; Audibert et al., 2005; Kawamori et al., 2013). На рис. 4 приведены графики, отображающие

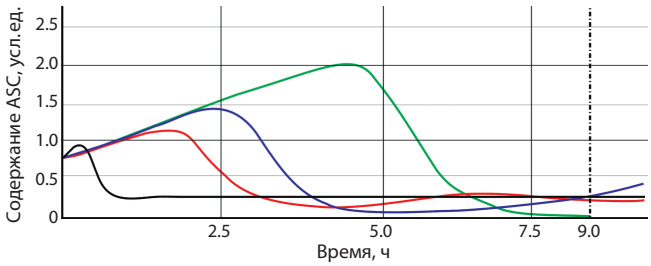


Рис. 4. Оценка минимального уровня содержания белка ASC в презумптивной РКСО, достаточного для приобретения клеткой нейрального статуса.

Значения временных параметров для графиков приведены в тексте.

Таблица 1. Значения параметров k_{xi} и x_{0i} при моделировании влияния мутаций в *AS-C* на содержание одноименных белков в презумптивной РКСО

Параметры	Номер эксперимента								
	1 (норма)	2	3	4	5	6	7	8	9
k_{xi}	1	0.9	0.6	0.5	0.4	0.3	0.2	0.1	0
x_{0i}	0.8	0.72	0.48	0.4	0.32	0.24	0.16	0.08	0

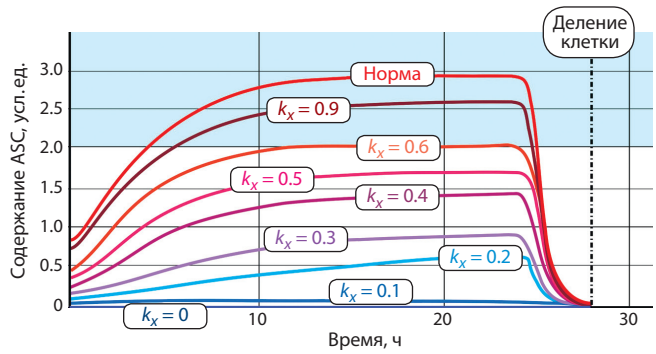


Рис. 5. Динамика содержания белков ASC в презумптивной родительской клетке механорецептора при наличии мутаций в комплексе генов *achaete-scute*.

Цветом показана область, где содержание белка ASC достаточно для детерминации РКСО.

результат моделирования для значений τ, равных 0 ч (кривая черного цвета), 0.5 ч (красная кривая), 1 ч (синяя), 2.1 ч (зеленая кривая).

Значение τ = 2.1 ч – первое, при котором выполняются два условия перехода клетки к делению: 1) произошло падение содержания белков ASC до нулевой отметки и 2) время *T* составило примерно 9 ч. Таким образом, можно предполагать, что повышения содержания ASC в клетке примерно в два с половиной раза от исходного уровня уже достаточно, чтобы клетка пошла по пути нейральной дифференцировки.

Приведенные данные были получены для описания функционирования ЦРК в норме. Но модель позволяет также оценить относительный вклад генов ЦРК в его функционирование через учет мутаций в каждом из них.

Моделирование функционирования ЦРК в родительской клетке механорецептора при наличии мутаций в генах *AS-C*

Из экспериментальных данных известно, что мутации генов *achaete-scute* выражаются в отсутствии части, а в ряде случаев даже всех механорецепторов стандартного набора (Agol, 1931; Dubinin, 1932; Cabrera et al., 1994; Roark et al., 1995; Pi et al., 2001; Escudero et al., 2005; Acar et al., 2006; Usui et al., 2008; Garcıa-Bellido, de Celis, 2009).

Для оценки влияния мутаций генов *AS-C* на функционирование ЦРК был проведен ряд численных экспериментов, в которых использовались следующие значения параметров системы (1):

$D = 1.6$; $G = 1$; $m_x = 0.3$; $U = 1.1$; $S = 5.5$;
$a_1 = 2.9$; $n_1 = 1$; $b_1 = 1$; $a_4 = 5.8$; $n_4 = 1$; $b_4 = 5.6$;
$a_6 = 6$; $n_6 = 1$; $b_6 = 5.7$;
$C_y = 14.1$; $d_1 = 4.1$; $d_2 = 4.7$; $m_y = 0.5$;
$C_e = 2.9$; $d_3 = 7.5$; $m_e = 0.4$;
$\alpha_4 = 3$; $v_4 = 1.9$; $\beta_4 = 1.2$; $m_z = 1.6$;
$\alpha_5 = 14.8$; $v_5 = 1.1$; $\beta_5 = 14.8$; $m_u = 2.3$;
$\alpha_6 = 2$; $v_6 = 1$; $\beta_6 = 1$; $m_w = 1$;
$\alpha_7 = 4.5$; $v_7 = 3.1$; $\beta_7 = 0.5$; $m_p = 0.6$; $L = 1.1$;
$y_0 = 1.6$; $E_0 = 1.1$; $z_0 = 0.4$; $u_0 = 0$; $w_0 = 0$; $p_0 = 0$;
$T = 28$; $\tau = 12$.

Коэффициенты *k* во всех уравнениях системы (1), кроме первого, принимались равными 1.

Значения k_{xi} и x_{0i} приведены в табл. 1. Значение параметра k_{xi} меняется от 0 (полное отсутствие белка) до 1 (содержание белка в норме) и с биологической точки зрения отражает степень влияния мутации в *AS-C* на содержание белков ASC. Чем меньше значение k_{xi}, тем меньше содержание белка в клетке. Параметр x_{0i} задает начальное содержание белков ASC. В численных экспериментах принятое значение для x_{01} составляет 0.8, что соответствует норме ($k_{x1} = 1$) (см. рис. 2). Коэффициенты k_{xi} задают пропорциональное снижение значений содержания белков x_{0i} по формуле $x_{0i} = x_{01} \cdot k_{xi}$.

Результаты численных экспериментов представлены на рис. 5. Согласно приведенным выше данным, при отсутствии мутаций в генах ЦРК детерминация клетки как РКСО становится возможной при увеличении содержания ASC не менее чем в ~2.5 раза относительно начального значения (см. рис. 4). Исходя из этого можно оценить минимальное значение k_{xi}, при котором выполняется указанное условие. На рис. 5 область значений для содержания белков ASC, допускающих детерминацию РКСО, показана бирюзовым цветом. Кривые содержания белков ASC, попадающие в эту область, соответствуют значениям k_{xi}, при которых сохраняется возможность детерминации клетки как РКСО.

Необходимый уровень содержания белков ASC достигается при значениях $k_{xi} \geq 0.6$. Значение параметра $k_{x3} = 0.6$ соответствует снижению содержания белка ASC на 40 % относительно начальных значений для нормы. С биологической точки зрения это означает, что снижение содержания в клетке белков ASC более чем на 40 % запрещает ее дифференцировку по нейральному пути развития и, следовательно, влечет отсутствие механорецептора.

Моделирование функционирования ЦРК в презумптивной РКСО при наличии мутаций входящих в его состав генов

Компоненты ЦРК объединены внутриклеточной системой прямых и обратных связей (см. рис. 1), строго регламентирующей наработку и деградацию белков ASC. Следовательно, мутации в каждом из них должны влиять на содержание этих белков в клетке и иметь определенный фенотипический эффект. Действительно, экспериментально было показано, что мутации в генах ЦРК проявляются в отклонениях от канонической архитектуры щетиночного рисунка – изменениях числа и/или позиционирования механорецепторов. Рассматриваемая модель с учетом мутационных изменений генов ЦРК позволяет оценить степень и характер их влияния на динамику содержания белков ASC. Были проведены численные эксперименты, в которых коэффициенты k_y (для *hairy*), k_e (для *emc*), k_z (для *sens*), k_u (для *scrt*), k_w (для *chn*), k_p (для *phyl*) полагались равными 0, что соответствует ситуации полного отсутствия соответствующих белков.

Ряд параметров сохранялся неизменным:
$D = 1.6$; $G = 1$; $m_x = 0.3$; $U = 1.1$; $S = 5.5$;
$a_1 = 2.9$; $n_1 = 1$; $b_1 = 1$; $a_4 = 5.8$; $n_4 = 1$; $b_4 = 5.6$;
$a_6 = 6$; $n_6 = 1$; $b_6 = 5.7$;
$C_y = 14.1$; $d_1 = 4.1$; $d_2 = 4.7$; $m_y = 0.5$;
$C_e = 2.9$; $d_3 = 7.5$; $m_e = 0.4$;
$\alpha_4 = 3$; $v_4 = 1.9$; $\beta_4 = 1.2$; $m_z = 1.6$;
$\alpha_5 = 14.8$; $v_5 = 1.1$; $\beta_5 = 14.8$; $m_u = 2.3$;
$\alpha_6 = 2$; $v_6 = 1$; $\beta_6 = 1$; $m_w = 1$;
$\alpha_7 = 4.5$; $v_7 = 3.1$; $\beta_7 = 0.5$; $m_p = 0.6$; $L = 1.1$;
$T = 28$; $\tau = 12$;
$k_x = 1$; $x_0 = 0.8$.

Переменные параметры представлены в табл. 2, где значения k, равные 0 или 1, означают наличие или отсутствие мутации в гене, а параметры y_0, z_0, u_0, w_0, p_0 и E_0 задают начальное содержание белков Hairy, SENS, SCRT, CHN, PHYL и EMC соответственно.

Результаты проведенных численных экспериментов представлены на рис. 6. Сравнение профилей кривых на рис. 6 выявляет некую иерархию генов ЦРК по влиянию на содержание белков ASC, что отражается в размахе отклонений от кривой, характеризующей динамику этих белков в норме, т. е. при отсутствии мутаций во всех генах, входящих в ЦРК. Влияние определенного гена тем сильнее, чем больше это отклонение.

Наиболее сильное влияние оказывают гены *emc* и *hairy*, поскольку мутации в них значительно отклоняют

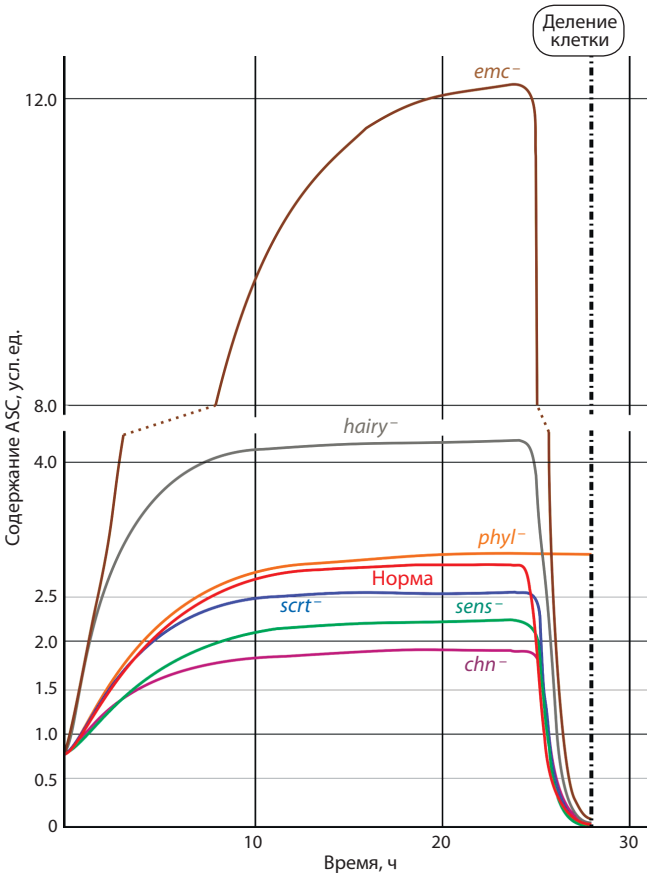


Рис. 6. Динамика содержания белков ASC в презумптивной родительской клетке механорецептора при наличии мутаций в генах ЦРК. Знак «–» после названий генов означает наличие мутаций в этих генах.

уровень ASC относительно его нормальных показателей в бóльшую сторону. Это биологически оправданный результат, так как EMC и Hairy репрессируют *AS-C* (Moscoso del Prado, Garcia-Bellido, 1984), и снятие этой репрессии должно проявляться увеличением содержания ASC. Фенотипическое проявление мутаций состоит в развитии дополнительных механорецепторов (Ingham et al., 1985; de Celis et al., 1991). Возможно, одновременное резкое и быстрое нарастание ASC в клетках пронейрального кластера приводит к рассогласованию межклеточных взаимодействий, опосредованных сигнальными путями, и формированию в пронейральном кластере не одной, как в норме, а нескольких родительских клеток.

Таблица 2. Значения изменяющихся параметров при моделировании влияния мутаций в генах ЦРК на содержание белков ASC

Мутация в гене	k_y	k_e	k_z	k_u	k_w	k_p	y_0	E_0	z_0	u_0	w_0	p_0
hairy⁻	0	1	1	1	1	1	0	1.1	0.4	0	0	0
emc⁻	1	0	1	1	1	1	1.6	0	0.4	0	0	0
sens⁻	1	1	0	1	1	1	1.6	1.1	0	0	0	0
scrt⁻	1	1	1	0	1	1	1.6	1.1	0.4	0	0	0
chn⁻	1	1	1	1	0	1	1.6	1.1	0.4	0	0	0
phyl⁻	1	1	1	1	1	0	1.6	1.1	0.4	0	0	0

Мутация в *chn* проявляется в заметном снижении уровня ASC (соответствующая кривая лежит ниже кривой для нормы). Эффект связан с тем, что мутация в гене *chn* приводит к отсутствию одноименного белка, напрямую активирующего гены *AS-C* и репрессирующего гены *emc* и *hairy* (Escudero et al., 2005; Yamasaki et al., 2011). При этом наработка белков ASC не может достигнуть требуемых значений.

Менее выраженное снижение уровня белков вызывают мутации в генах *sens* и *scrt*, что также согласуется с известными данными о функциях этих генов в системе ЦРК и проявлениях мутаций в этих генах: белок SENS известен как коактиватор активности *AS-C*, следовательно, мутация приведет к некоторому снижению наработки ASC. Белок SCRT репрессирует ген *hairy*, вследствие чего возможно увеличение уровня ASC, которое, тем не менее, не достигает нормальных значений из-за действия других прямых репрессоров активности генов *AS-C* (см. рис. 1) (Roark et al., 1995; Nolo et al., 2000).

При мутации гена *phyl* уровень ASC ожидаемо остается на достигнутом плато, поскольку в этом случае не нарабатывается белок PHYL, ответственный за его деградацию (Chang et al., 2008). РКСО не может перейти к делению, и фенотипически эффект должен проявиться в отсутствии механорецептора в положенной позиции. Этот вывод подтверждается экспериментальными данными (Pi et al., 2001).

Заключение

За десятилетия изучения системы формирования щетиночного узора на голове и теле дрозофилы накоплен огромный фактологический материал и выявлены отдельные механизмы, лежащие в основе ее функционирования. Между тем конкретные детали морфогенеза механорецепторов до конца не ясны до сих пор.

Ранее нами было показано, что развитие отдельного механорецептора и становление рисунка в целом регламентируются центральным регуляторным контуром, определяющим экспрессию генов *AS-C* и наработку одноименных белков в родительской клетке. С учетом всех выявленных компонентов ЦРК и характера связей между ними была разработана математическая модель его функционирования, позволившая отойти от чисто качественного описания системы контроля содержания белков ASC и выявить ее новые количественные характеристики.

В частности, из проведенных численных экспериментов в рамках принятой математической модели следует, что клетка детерминируется как РКСО при повышении содержания ASC примерно в два с половиной раза относительно уровня в клетках пронейрального кластера. Показано, что разные элементы контура по-разному влияют на содержание белков ASC в презумптивной клетке механорецептора. Наиболее значимое влияние оказывает главный компонент ЦРК – комплекс *AS-C*, и мутации, снижающие содержание ASC более чем на 40 %, приводят к запрету выделения РКСО. Мутации в остальных генах контура в разной степени изменяют уровень белков ASC. Наиболее выраженный эффект наблюдается при мутациях в генах *emc* и *hairy*.

Таким образом, модель показывает, что ЦРК как система чувствителен к изменению внутренних взаимодействий и его полноценное функционирование, результатом которого становится определенная динамика изменения в уровне белков ASC, возможно лишь при согласованной работе всех составляющих регуляторного контура.

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 Y.G. Gene networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2013;17(4/2):833-850 (in Russian)]

Acar M., Jafar-Nejad H., Giagtzoglou N., Yallampalli S., David G., He Y., Delidakis C., Bellen H.J. Senseless physically interacts with proneural proteins and functions as a transcriptional co-activator. *Development*. 2006;133(10):1979-1989. DOI 10.1242/dev.02372

Agol I.J. Step allelomorphism in *D. melanogaster*. *Genetics*. 1931; 16(3):254-266. DOI 10.1093/genetics/16.3.254

Audibert A., Simon F., Gho M. Cell cycle diversity involves differential regulation of Cyclin E activity in the *Drosophila* bristle cell lineage. *Development*. 2005;132(10):2287-2297. DOI 10.1242/dev.01797

Ayeni J.O., Audibert A., Fichelson P., Srayko M., Gho M., Campbell S.D. G2 phase arrest prevents bristle progenitor self-renewal and synchronizes cell division with cell fate differentiation. *Development*. 2016;143(7):1160-1169. DOI 10.1242/dev.134270

Bukharina T.A., Akinshin A.A., Golubyatnikov V.P., Furman D.P. Mathematical and numerical models of the central regulatory circuit of the morphogenesis system of *Drosophila*. *J. Appl. Ind. Math.* 2020;14(2):249-255. DOI 10.1134/S1990478920020040

Cabrera C.V., Alonso M.C. Transcriptional activation by heterodimers of the *achaete-scute* and *daughterless* gene products of *Drosophila*. *EMBO J.* 1991;10(10):2965-2973. DOI 10.1002/j.1460-2075.1991.tb07847.x

Cabrera C.V., Alonso M.C., Huikeshoven H. Regulation of *scute* function by *extramacrochaete in vitro* and *in vivo*. *Development*. 1994; 120(12):3595-3603. DOI 10.1242/dev.120.12.3595

Chang P.J., Hsiao Y.L., Tien A.C., Li Y.C., Pi H. Negative-feedback regulation of proneural proteins controls the timing of neural precursor division. *Development*. 2008;135(18):3021-3030. DOI 10.1242/dev.021923

Chasman D., Fotuhi Siahpirani A., Roy S. Network-based approaches for analysis of complex biological systems. *Curr. Opin. Biotechnol*. 2016;39:157-166. DOI 10.1016/j.copbio.2016.04.007

Corson F., Couturier L., Rouault H., Mazouni K., Schweisguth F. Self-organized Notch dynamics generate stereotyped sensory organ patterns in *Drosophila*. *Science*. 2017;356(6337):eaai7407. DOI 10.1126/science.aai7407

Cubas P., de Celis J.F., Campuzano S., Modolell J. Proneural clusters of *achaete-scute* expression and the generation of sensory organs in the *Drosophila* imaginal wing disc. *Genes Dev*. 1991;5(6):996-1008. DOI 10.1101/gad.5.6.996

de Celis J.F., Marí-Beffa M., García-Bellido A. Function of trans-acting genes of the *achaete-scute* complex in sensory organ patterning in the mesonotum of *Drosophila*. *Rouxs Arch. Dev. Biol*. 1991;200(2): 64-76. DOI 10.1007/BF00637186

Dubinin N.P. Step-allelomorphism in *D. melanogaster*. The allelomorphs *achaete2-scute10*, *achaete1-scute11* and *achaete3-scute13*. *J. Genet*. 1932;25(2):163-181. DOI 10.1007/BF02983250

Emmert-Streib F., Glazko G.V. Network biology: a direct approach to study biological function. *Wiley Interdiscip. Rev. Syst. Biol. Med*. 2011;3(4):379-391. DOI 10.1002/wsbm.134

Escudero L.M., Caminero E., Schulze K.L., Bellen H.J., Modolell J. Charlatan, a Zn-finger transcription factor, establishes a novel level

of regulation of the proneural *achaete/scute* genes of *Drosophila*. *Development*. 2005;132(6):1211-1222. DOI 10.1242/dev.01691

Furman D.P., Bukharina T.A. Genetic regulation of morphogenesis of *Drosophila melanogaster* mechanoreceptors. *Russ. J. Dev. Biol*. 2022;53(4):239-251. DOI 10.1134/S1062360422040038

Garcıa-Bellido A., de Celis J.F. The complex tale of the *achaete-scute* complex: a paradigmatic case in the analysis of gene organization and function during development. *Genetics*. 2009;182(3):631-639. DOI 10.1534/genetics.109.104083

Ghysen A., Thomas R. The formation of sense organs in *Drosophila*: a logical approach. *Bioessays*. 2003;25(8):802-807. DOI 10.1002/bies.10311

Giri R., Brady S., Papadopoulos D.K., Carthew R.W. Single-cell Senseless protein analysis reveals metastable states during the transition to a sensory organ fate. *iScience*. 2022;25(10):105097. DOI 10.1016/j.isci.2022.105097

Golubyatnikov V.P., Bukharina T.A., Furman D.P. A model study of the morphogenesis of *D. melanogaster* mechanoreceptors: the central regulatory circuit. *J. Bioinform. Comput. Biol*. 2015;13(1):1540006. DOI 10.1142/S0219720015400065

Hsu C.P., Lee P.H., Chang C.W., Lee C.T. Constructing quantitative models from qualitative mutant phenotypes: preferences in selecting sensory organ precursors. *Bioinformatics*. 2006;22(11):1375-1382. DOI 10.1093/bioinformatics/btl082

Huang F., Dambly-Chaudiere C., Ghysen A. The emergence of sense organs in the wing disc of *Drosophila*. *Development*. 1991;111(4):1087-1095. DOI 10.1242/dev.111.4.1087

Ingham P.W., Pinchin S.M., Howard K.R., Ish-Horowicz D. Genetic analysis of the hairy locus in *Drosophila melanogaster*. *Genetics*. 1985;111(3):463-486. DOI 10.1093/genetics/111.3.463

Kawamori A., Shimaji K., Yamaguchi M. Temporal and spatial pattern of *dref* expression during *Drosophila* bristle development. *Cell Struct. Funct*. 2013;38(2):169-181. DOI 10.1247/csf.13004

Marnellos G., Mjolsness E. A gene network approach to modeling early neurogenesis in *Drosophila*. In: Pacific Symposium on Biocomputing '98, January 4–9, 1998, in Hawaii. World Scientific Pub Co Inc., 1998;30-41

Meir E., von Dassow G., Munro E., Odell G.M. Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Curr. Biol*. 2002;12(10):778-786. DOI 10.1016/s0960-9822(02)00839-4

Moscoso del Prado J., Garcia-Bellido A. Genetic regulation of the *achaete-scute* complex of *Drosophila melanogaster*. *Wilehm Roux Arch. Dev. Biol*. 1984;193(4):242-245. DOI 10.1007/BF01260345

Nolo R., Abbott L.A., Bellen H.J. Senseless, a Zn finger transcription factor, is necessary and sufficient for sensory organ development in *Drosophila*. *Cell*. 2000;102(3):349-362. DOI 10.1016/s0092-8674(00)00040-4

Pi H., Wu H.J., Chien C.T. A dual function of *phyllopod* in *Drosophila* external sensory organ development: cell fate specification of sensory organ precursor and its progeny. *Development*. 2001;128(14):2699-2710. DOI 10.1242/dev.128.14.2699

Reeves N., Posakony J.W. Genetic programs activated by proneural proteins in the developing *Drosophila* PNS. *Dev. Cell*. 2005;8(3):413-425. DOI 10.1016/j.devcel.2005.01.020

Roark M., Sturtevant M.A., Emery J., Vaessin H., Grell E., Bier E. *scratch*, a pan-neural gene encoding a zinc finger protein related to *snail*, promotes neuronal development. *Genes Dev*. 1995;9(19):2384-2398. DOI 10.1101/gad.9.19.2384

Schlitt T., Palin K., Rung J., Dietmann S., Lappe M., Ukkonen E., Brazma A. From gene networks to gene function. *Genome Res*. 2003;13(12):2568-2576. DOI 10.1101/gr.1111403

Skeath J.B., Carroll S.B. Regulation of *achaete-scute* gene expression and sensory organ pattern formation in the *Drosophila* wing. *Genes Dev*. 1991;5(6):984-995. DOI 10.1101/gad.5.6.984

Usui K., Kimura K.I. Sequential emergence of the evenly spaced microchaetes on the notum of *Drosophila*. *Rouxs Arch. Dev. Biol*. 1993;203(3):151-158. DOI 10.1007/BF00365054

Usui K., Goldstone C., Gibert J.M., Simpson P. Redundant mechanisms mediate bristle patterning on the *Drosophila* thorax. *Proc. Natl. Acad. Sci. USA*. 2008;105(51):20112-20117. DOI 10.1073/pnas.0804282105

Vaessin H., Brand M., Jan L.Y., Jan Y.N. *daughterless* is essential for neuronal precursor differentiation but not for initiation of neuronal precursor formation in *Drosophila* embryo. *Development*. 1994;120(4):935-945. DOI 10.1242/dev.120.4.935

Van Doren M., Powell P.A., Pasternak D., Singson A., Posakony J.W. Spatial regulation of proneural gene activity: auto- and cross-activation of *achaete* is antagonized by *extramacrochaetae*. *Genes Dev*. 1992;6(12B):2592-2605. DOI 10.1101/gad.6.12b.2592

Van Doren M., Bailey A.M., Esnayra J., Ede K., Posakony J.W. Negative regulation of proneural gene activity: *hairy* is a direct transcriptional repressor of *achaete. Genes Dev*. 1994;8(22):2729-2749. DOI 10.1101/gad.8.22.2729

Yamasaki Y., Lim Y.M., Niwa N., Hayashi S., Tsuda L. Robust specification of sensory neurons by dual functions of charlatan, a *Drosophila* NRSF/REST-like repressor of *extramacrochaetae* and *hairy*. *Genes Cells*. 2011;16(8):896-909. DOI 10.1111/j.1365-2443.2011.01537.x

Yasugi T., Sato M. Mathematical modeling of Notch dynamics in *Drosophila* neural development. *Fly* (*Austin*). 2022;16(1):24-36. DOI 10.1080/19336934.2021.1953363

Zhu X., Gerstein M., Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev*. 2007;21(9):1010-1024. DOI 10.1101/gad.1528707

ORCID ID

T.A. Bukharina orcid.org/0000-0002-9011-4196
V.P. Golubyatnikov orcid.org/0000-0002-9758-3833

Благодарности. Авторы выражают искреннюю благодарность А.А. Акиньшину за полезные советы и критические замечания. Работа поддержана бюджетными проектами FWNR-2022-0020 (ИЦиГ СО РАН, для Т.А.Б. и Д.П.Ф.) и FWNF-2022-0009 (ИМ СО РАН, для В.П.Г).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 18.07.2023. После доработки 20.09.2023. Принята к публикации 25.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):755-767
DOI 10.18699/VJGB-23-88

Перевод на английский язык https://vavilov.elpub.ru/jour

Бифуркационный анализ мультистабильности и гистерезиса в модели ВИЧ-инфекции

И.В. Миронов[1, 2], М.Ю. Христиченко[1, 3], Ю.М. Нечепуренко[1, 3], Д.С. Гребенников[2, 3], Г.А. Бочаров[2, 3] ✉

[1] Институт прикладной математики им. М.В. Келдыша Российской академии наук, Москва, Россия
[2] Первый Московский государственный медицинский университет им. И.М. Сеченова Министерства здравоохранения Российской Федерации, Москва, Россия
[3] Институт вычислительной математики им. Г.И. Марчука Российской академии наук, Москва, Россия
✉ gbocharov@gmail.com

Аннотация. Инфекционное заболевание, вызванное вирусами иммунодефицита человека первого типа (ВИЧ-1), остается серьезной угрозой здоровью людей. Существующий подход к лечению ВИЧ-1 основан на применении высокоактивной антиретровирусной терапии, имеющей побочные эффекты для здоровья и высокую стоимость. Для практической медицины актуальной является задача поиска методов функционального лечения, связанных с интенсификацией иммунного контроля размножения вирусов и заражения клеток-мишеней с последующим снижением уровня вирусной нагрузки и восстановления иммунного статуса. Исследования в области иммунотерапии ВИЧ-1 находятся на стадии концептуальной разработки в силу сложности совокупности процессов, регулирующих динамику инфекции и иммунного ответа. По этой причине чрезвычайно актуальным является использование методов математического моделирования динамики ВИЧ-1 инфекции для теоретического анализа возможностей снижения вирусной нагрузки путем воздействия на иммунную систему без применения антивирусной терапии. Целью исследования было изучение, во-первых, свойств би-, мультистабильности и гистерезиса на примере содержательной модели ВИЧ-1 инфекции, которая описывает важнейшие блоки процессов взаимодействия вирусов и организма человека, а именно: распространение инфекции в продуктивно и латентно зараженных клетках, появление мутантов и развитие Т-клеточного иммунного ответа, и, во-вторых, возможностей перевода клинической картины заболевания из более тяжелого состояния в более легкое. В данной работе проведен численный анализ условий существования стационарных решений математической модели ВИЧ-1 инфекции для наборов параметров, отвечающих фенотипически различным вариантам течения инфекционного заболевания. Для этого использованы разработанные авторами методы бифуркационного анализа моделей, представляющих собой системы обыкновенных дифференциальных уравнений и дифференциальных уравнений с запаздыванием. В качестве бифуркационного параметра рассматривается константа скорости активации макрофагов. Определены области в пространстве параметров модели, в частности, для скорости активации клеток врожденного иммунитета (макрофагов), при которых имеют место свойства би-, мультистабильности и гистерезиса, и исследованы особенности кинетики перехода между устойчивыми положениями равновесия. В целом результаты бифуркационного анализа модели ВИЧ-1 инфекции формируют теоретическую основу для разработки комбинированных иммунотерапевтических воздействий для лечения ВИЧ-1. Результаты проведенного исследования модели ВИЧ-1 инфекции для параметров процессов, отвечающих разным фенотипам динамики заболевания (типичное, длительно не прогрессирующее и быстро прогрессирующее), указывают на то, что для эффективного функционального лечения больных ВИЧ-инфекцией требуется развитие персонализированного подхода, учитывающего как свойства популяции квазивидов ВИЧ-1, так и иммунный статус пациента.

Ключевые слова: математическая модель; ВИЧ-инфекция; обыкновенные дифференциальные уравнения; бифуркационный анализ; стационарные решения; бистабильность; мультистабильность; гистерезис; оптимальное управление.

Для цитирования: Миронов И.В., Христиченко М.Ю., Нечепуренко Ю.М., Гребенников Д.С., Бочаров Г.А. Бифуркационный анализ мультистабильности и гистерезиса в модели ВИЧ-инфекции. *Вавиловский журнал генетики и селекции*. 2023;27(7):755-767. DOI 10.18699/VJGB-23-88

Bifurcation analysis of multistability and hysteresis in a model of HIV infection

I.V. Mironov[1, 2], M.Yu. Khristichenko[1, 3], Yu.M. Nechepurenko[1, 3], D.S. Grebennikov[2, 3], G.A. Bocharov[2, 3] ✉

[1] Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences, Moscow, Russia
[2] Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation, Moscow, Russia
[3] Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences, Moscow, Russia
✉ gbocharov@gmail.com

Abstract. The infectious disease caused by human immunodeficiency virus type 1 (HIV-1) remains a serious threat to human health. The current approach to HIV-1 treatment is based on the use of highly active antiretroviral therapy, which has side effects and is costly. For clinical practice, it is highly important to create functional cures that can enhance immune

control of viral growth and infection of target cells with a subsequent reduction in viral load and restoration of the immune status. HIV-1 control efforts with reliance on immunotherapy remain at a conceptual stage due to the complexity of a set of processes that regulate the dynamics of infection and immune response. For this reason, it is extremely important to use methods of mathematical modeling of HIV-1 infection dynamics for theoretical analysis of possibilities of reducing the viral load by affecting the immune system without the usage of antiviral therapy. The aim of our study is to examine the existence of bi-, multistability and hysteresis properties with a meaningful mathematical model of HIV-1 infection. The model describes the most important blocks of the processes of interaction between viruses and the human body, namely, the spread of infection in productively and latently infected cells, the appearance of viral mutants and the development of the T cell immune response. Furthermore, our analysis aims to study the possibilities of transferring the clinical pattern of the disease from a more severe state to a milder one. We analyze numerically the conditions for the existence of steady states of the mathematical model of HIV-1 infection for the numerical values of model parameters corresponding to phenotypically different variants of the infectious disease course. To this end, original computational methods of bifurcation analysis of mathematical models formulated with systems of ordinary differential equations and delay differential equations are used. The macrophage activation rate constant is considered as a bifurcation parameter. The regions in the model parameter space, in particular, for the rate of activation of innate immune cells (macrophages), in which the properties of bi-, multistability and hysteresis are expressed, have been identified, and the features characterizing transition kinetics between stable equilibrium states have been explored. Overall, the results of bifurcation analysis of the HIV-1 infection model form a theoretical basis for the development of combination immune-based therapeutic approaches to HIV-1 treatment. In particular, the results of the study of the HIV-1 infection model for parameter sets corresponding to different phenotypes of disease dynamics (typical, long-term non-progressing and rapidly progressing courses) indicate that an effective functional treatment (cure) of HIV-1-infected patients requires the development of a personalized approach that takes into account both the properties of the HIV-1 quasispecies population and the patient's immune status.
Key words: mathematical model; HIV infection; ordinary differential equations; bifurcation analysis; stationary solutions; bistability; multistability; hysteresis; optimal control.

For citation: Mironov I.V., Khristichenko M.Yu., Nechepurenko Yu.M., Grebennikov D.S., Bocharov G.A. Bifurcation analysis of multistability and hysteresis in a model of HIV infection. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):755-767. DOI 10.18699/VJGB-23-88

Введение

Инфекционное заболевание человека, вызванное вирусами иммунодефицита первого типа (ВИЧ-1), остается серьезной угрозой здоровью людей во всем мире, с числом заражений и смертельных исходов от сопутствующих осложнений порядка 1.5×10^6 и 0.65×10^6 соответственно (Landovitz et al., 2023). Существующий подход к лечению ВИЧ-1 связан с постоянным применением средств высокоактивной антиретровирусной терапии (Gandhi et al., 2023), подавляющих различные стадии внутриклеточного цикла размножения вирусов и таким образом снижающих вирусную нагрузку в организме больного. Однако реализация антиретровирусной терапии характеризуется побочными эффектами для здоровья пациентов, прерыванием режима приема препаратов и высокой стоимостью лечения (Trickey et al., 2022). По этой причине актуальной стала задача поиска новых методов лечения (Rasmussen, Søgaard, 2018; Niessl et al., 2020), в том числе связанных с активацией иммунного контроля процессов размножения вирусов и заражения клеток-мишеней, включая физиологические механизмы обновления клеточного гомеостаза (Grossman et al., 2020), в рамках системного подхода в иммунологии (Ludewig et al., 2012, Villani et al., 2018). Исследования в области иммунотерапии ВИЧ-1 находятся на стадии концептуальной разработки в силу сложности совокупности процессов, регулирующих динамику инфекции и иммунного ответа (Landovitz et al., 2023). В этой связи использование методов математического моделирования динамики ВИЧ-1 инфекции является инструментом теоретического анализа возможностей снижения вирусной нагрузки путем воздействия на иммунную систему без применения антивирусной терапии (Bocharov et al., 2022).

Ранее нами было отмечено, что разработка математических моделей для описания и исследования динамики инфекционных заболеваний имеет одной из целей анализ характеристик чувствительности динамики к воздействиям различной природы, например, по отношению к возмущениям параметров регуляторных процессов или состояния системы в фазовом пространстве (Bocharov et al., 2021). Результаты моделирования позволяют перевести в рациональную плоскость проектирование комбинированных управляющих воздействий для коррекции неблагоприятного течения, в частности из области с высокой вирусной нагрузкой в область с низкой вирусной нагрузкой. Реализуемость соответствующих переходов определяется фундаментальными характеристиками моделируемой системы – наличием бистабильности и/или мультистабильности и гистерезиса. Так, бистабильность как возможность системы «вирус–организм человека» сосуществовать в двух устойчивых равновесных состояниях является основанием для поиска режимов функционального лечения вирусной инфекции путем перехода из хронического устойчивого стационарного состояния с более высокой вирусной нагрузкой в более благоприятное устойчивое стационарное состояние с пониженной вирусной нагрузкой за счет активации компонент иммунной системы. Наличие свойства гистерезиса у бифуркационных кривых динамической системы делает значимой предысторию, в частности, ту ветвь, на которой находится стационарное состояние системы при изменении бифуркационных параметров (Христиченко и др., 2022).

Исследования по математическому моделированию динамики ВИЧ-1 инфекции в организме человека активно развиваются в течение последних 30 лет (Perelson, Nel-

son, 1999; Nowak, May, 2000). Направления исследований достаточно полно представлены в обзоре (Bocharov et al., 2012). Главным образом работы сфокусированы на изучении кинетики инфекции при реализации антиретровирусной терапии на основе моделей малой размерности (Akın et al., 2020). Модели ВИЧ-1 инфекции, рассматривающие развитие противовирусного иммунного ответа, связаны также с решением задачи оценивания параметров инфекции по данным отдельных пациентов (Banks et al., 2017). Концептуальные аспекты динамики ВИЧ-1 инфекции, такие как мультистабильность и гистерезис, остаются недостаточно освещенными. Так, изучение стационарных состояний сводится к выяснению условий существования свободного от инфекции положения равновесия и состояния инфицированного организма в зависимости от параметров модели, комбинируемых в виде базового репродуктивного числа (Perelson, Nelson, 1999; Nowak, May, 2000).

Целью данного исследования было проведение математического анализа, во-первых, свойств би-, мультистабильности и гистерезиса для модели ВИЧ-1 инфекции, которая описывает важнейшие блоки процессов взаимодействия вирусов и организма человека для наборов параметров модели, отвечающих разным фенотипам динамики заболевания (типичное, длительно не прогрессирующее и быстро прогрессирующее), и, во-вторых, возможностей перевода режима течения заболевания из более тяжелого состояния в более легкое.

В задачи исследования входили бифуркационный анализ модели течения инфекционного заболевания ВИЧ-1 для определения областей значений параметров, в которых сосуществуют несколько положений равновесия, и изучение переходов между ними, которые характеризуются зависимостью от предыстории состояния системы «вирус–организм человека» (свойство гистерезиса). В качестве математической модели для исследования стационарных режимов динамики ВИЧ-1 инфекции и переходов между ними мы выбрали модель, в рамках которой:

- описывается вся кинетика инфекционного заболевания от заражения ВИЧ-1 до фазы СПИД;
- рассматривается достаточно полный спектр процессов развития инфекции и иммунного ответа;
- выполнена калибровка параметров модели, соответствующих различным фенотипам динамики инфекции;
- введено описание антиретровирусной терапии;
- рассмотрена задача расчета оптимальной антиретровирусной терапии с учетом побочных эффектов.

Ранее данная модель использовалась нами для разработки более полного описания иммунного ответа на ВИЧ-инфекцию, учитывающего нейроэндокринную регуляцию, в частности влияние гормонов (ТТГ, Т3, Т4) на иммунный ответ, и построение на ее основе оптимальной антивирусной терапии (Савинкова и др., 2019).

Настоящая работа состоит из четырех разделов. В разделе «Материалы и методы» описаны рассматриваемая математическая модель инфекции ВИЧ-1 и численные методы, используемые для анализа модели. В разделе «Результаты» приведены результаты исследования стационарных состояний системы путем их трассирования по параметрам модели и результаты анализа изменений положения равновесия модели при терапевтических воздействиях, которые входят в правые части уравнений через слагаемые, описывающие заражение клеток-мишеней и размножение вирусов. Применение результатов работы для теоретической разработки новых подходов к лечению ВИЧ-1 рассмотрено в разделе «Обсуждение».

Материалы и методы

Определим основные понятия, которые будут нами использоваться в дальнейшем.

- Функциональное лечение ВИЧ-1 – подход к терапии хронической инфекции, связанный с активацией иммунного контроля процессов размножения вирусов и заражения клеток-мишеней, который позволяет не использовать антиретровирусные препараты.
- Би-(мульти)стабильность – свойство динамической системы иметь два (или более) устойчивых положения равновесия при одних и тех же значениях параметров.
- Гистерезис – свойство динамической системы, заключающееся в зависимости ее состояния от предыстории при изменении значений параметров, что можно использовать для перехода из одного стационарного состояния в другое путем варьирования параметров.

Математическая модель ВИЧ-инфекции

Рассматриваемая математическая модель ВИЧ-инфекции сформулирована в (Hadjiandreou et al., 2009) в виде системы из одиннадцати обыкновенных дифференциальных уравнений. Она описывает скорость изменения во времени следующих концентраций: вируса дикого типа V_1, мутировавшего вируса V_2, CD4$^+$ Т-клеток T, зараженных CD4$^+$ Т-клеток T_1, зараженных мутировавшим вирусом CD4$^+$ Т-клеток T_2, латентно зараженных Т-клеток T_{L1}, латентно зараженных мутировавшим вирусом Т-клеток T_{L2}, макрофагов M, зараженных макрофагов M_1, зараженных мутировавшим вирусом макрофагов M_2, цитотоксических CD8$^+$ Т-лимфоцитов CTL. Система включает в себя три блока уравнений: блок CD4$^+$ Т-клеток, блок макрофагов и CTL, блок вирусов дикого типа и мутантов.

Блок CD4$^+$ Т-клеток состоит из уравнения

$$\frac{dT}{dt} = s_1 + \frac{p_1(V_1+V_2)T}{V_1+V_2+S_1} - (1-u_1)(k_1V_1+k_2M_1)T - \\ - \varphi(k_1V_2+k_2M_2)T + rT\left(1-\frac{T+T_1+T_2+T_{L1}+T_{L2}}{T_{\max}}\right) - \delta_1 T, \quad (1)$$

где первый член описывает постоянный приток CD4$^+$ Т-клеток из тимуса, второй – антиген-индуцированное деление, третий – убыль вследствие заражения вирусами дикого типа и зараженными ими макрофагами 1-й популяции, четвертый – заражение мутировавшими вирусами и зараженными ими макрофагами 2-й популяции, пятый – гомеостатическую пролиферацию, шестой – апоптоз, следующих двух уравнений

$$\frac{dT_1}{dt} = (1-u_1)\psi(k_1V_1+k_2M_1)T + \alpha_1 T_{L1} - \delta_2 T_1 - k_3 T_1 CTL \quad (2)$$

и

$$\frac{dT_2}{dt} = \psi\varphi\,(k_1V_2+k_2M_2)T + \alpha_1 T_{L2} - \delta_2 T_2 - k_3 T_2 CTL, \quad (3)$$

где первый член описывает прирост популяции за счет заражения вирионами и зараженными вирусами дикого типа

Таблица 1. Биологический смысл параметров модели и их допустимый диапазон значений

Параметр	Биологический смысл	Диапазон
s_1	Константа скорости образования новых неинфицированных $CD4^+$ Т-клеток	5–36 $mm^{-3}d^{-1}$
s_2	Константа скорости образования новых макрофагов	0.03–0.015 $mm^{-3}d^{-1}$
s_3	Константа скорости образования новых цитотоксических Т-лимфоцитов	–
p_1	Константа скорости активации, обеспечивающая прирост $CD4^+$ Т-клеток за счет иммунного ответа	0.01–5 d^{-1}
p_2	Константа скорости активации, обеспечивающая прирост численности макрофагов	–
S_1	Константа насыщения	1–188 mm^{-3}
S_2	Константа насыщения	–
k_1	Константа скорости инфицирования $CD4^+$ Т-клеток	10^{-8}–10^{-2} mm^3d^{-1}
k_2	Константа скорости инфицирования $CD4^+$ Т-клеток	10^{-6} mm^3d^{-1}
k_3	Константа скорости уничтожения инфицированных $CD4^+$ Т-клеток цитотоксическими Т-лимфоцитами	10^{-4} – 1 mm^3d^{-1}
k_4	Константа скорости инфицирования макрофагов вирусами	$4.7 \cdot 10^{-9}$ – 10^{-3} mm^3d^{-1}
k_5	Константа скорости уничтожения инфицированных макрофагов цитотоксическими Т-лимфоцитами	–
k_6	Константа скорости пролиферации цитотоксических Т-лимфоцитов, связанная с текущим количеством инфицированных $CD4^+$ Т-клеток	10^{-6}–10^{-3} mm^3d^{-1}
k_7	Константа скорости пролиферации цитотоксических Т-лимфоцитов, связанная с текущим количеством инфицированных макрофагов	–
k_8	Константа скорости продукции вирусов инфицированными $CD4^+$ Т-клетками	$2.4 \cdot 10^{-1}$ – $5 \cdot 10^2$ d^{-1}
k_9	Константа скорости продукции вирусов инфицированными макрофагами	$5 \cdot 10^{-3}$ – $3 \cdot 10^2$ d^{-1}
k_{10}	Константа скорости уменьшения числа вирусов, связанная с расходом на инфицирование $CD4^+$ Т-клеток	10^{-8} – 10^{-2} mm^3d^{-1}
k_{11}	Константа скорости уменьшения числа вирусов, связанная с инфицированием макрофагов	$4.7 \cdot 10^{-9}$ – 10^{-3} mm^3d^{-1}
k_{12}	Константа скорости элиминации вирусов, связанная с иммунным ответом	–
δ_1	Константа скорости естественной гибели неинфицированных $CD4^+$ Т-клеток	0.01–0.02 d^{-1}
δ_2	Константа скорости естественной гибели инфицированных $CD4^+$ Т-клеток	0.24–0.7 d^{-1}
δ_3	Константа скорости естественной гибели латентно зараженных $CD4^+$ Т-клеток	0.02–0.069 d^{-1}
δ_4	Константа скорости естественной гибели макрофагов	0.005 d^{-1}
δ_5	Константа скорости естественной гибели зараженных макрофагов	0.005 d^{-1}
δ_6	Константа скорости естественной гибели цитотоксических Т-лимфоцитов	0.015–0.05 d^{-1}
δ_7	Константа скорости естественной гибели вирусов	2.39–13 d^{-1}
α_1	Константа активации латентно инфицированных $CD4^+$ Т-клеток	–
ψ	Доля $CD4^+$ Т-клеток, которые становятся продуктивно инфицированными, $(1 - \psi)$ становятся латентно инфицированными	0.93–0.98
φ	Коэффициент, отвечающий за снижение приспособленности мутировавшего вируса к заражению и способности к репликации	0.1–0.9
r	Константа скорости роста количества неинфицированных $CD4^+$ Т-клеток	0.03 d^{-1}
T_{max}	Максимальная концентрация $CD4^+$ Т-клеток	1500–2000 mm^{-3}
μ	Доля вирусов, которая мутирует	$3 \cdot 10^{-5}$ –10^{-3}
f_i	Коэффициент отношения эффективности действия лечения на макрофаги к эффективности действия лечения на $CD4^+$ Т-клетки	0.34

или мутировавшими вирусами макрофагами, второй – переход латентно инфицированных клеток в продуктивно зараженные, третий – апоптоз, четвертый – уничтожение Т-киллерами, и следующих двух уравнений

$$\frac{dT_{L1}}{dt} = (1-u_1)(1-\psi)(k_1V_1+k_2M_1)T-\alpha_1T_{L1}-\delta_3T_{L1} \quad (4)$$

и

$$\frac{dT_{L2}}{dt} = (1-\psi)\varphi(k_1V_2+k_2M_2)T-\alpha_1T_{L2}-\delta_3T_{L2}, \quad (5)$$

где первый член описывает прирост популяции за счет заражения вирионами и зараженными вирусами дикого типа или мутировавшими вирусами макрофагами, второй – переход латентно инфицированных клеток в продуктивно зараженные, третий – апоптоз.

Блок макрофагов и *CTL* состоит из уравнения

$$\frac{dM}{dt} = s_2+\frac{p_2(V_1+V_2)M}{V_1+V_2+S_2}-(1-f_1u_1)k_4V_1M-\varphi k_4V_2M-\delta_4M, \quad (6)$$

где первый член описывает постоянный приток клеток из костного мозга, второй – процесс активации макрофагов с возможностью их последующего деления вследствие хронического воспаления, вызванного ВИЧ-1 инфекцией, третий – заражение макрофагов вирусами дикого типа, четвертый – заражение макрофагов мутировавшими вирусами, пятый – апоптоз, следующих двух уравнений

$$\frac{dM_1}{dt} = (1-f_1u_1)k_4V_1M-\delta_5M_1-k_5M_1CTL \quad (7)$$

и

$$\frac{dM_2}{dt} = \varphi k_4V_2M-\delta_5M_2-k_5M_2CTL, \quad (8)$$

где первый член описывает прирост численности популяции за счет заражения макрофагов вирионами, второй – апоптоз, третий – уничтожение Т-киллерами, и уравнения

$$\frac{dCTL}{dt} = s_3+k_6(T_1+T_2)CTL+k_7(M_1+M_2)CTL-\delta_6CTL, \quad (9)$$

где первый член описывает постоянный приток клеток из тимуса, второй – клональную пролиферацию, индуцированную зараженными $CD4^+$ Т-клетками, третий – клональную пролиферацию, индуцированную зараженными макрофагами, четвертый – апоптоз.

Блок вирусов дикого типа и мутантов состоит из двух уравнений:

$$\frac{dV_1}{dt} = (1-u_2)(1-\mu)k_8T_1+(1-f_2u_2)(1-\mu)k_9M_1+ \\ +\mu\varphi k_8T_2+\mu\varphi k_9M_2-(k_{10}T+k_{11}M)V_1-k_{12}V_1M-\delta_7V_1 \quad (10)$$

и

$$\frac{dV_2}{dt} = (1-\mu)\varphi k_8T_2+(1-\mu)\varphi k_9M_2+(1-u_2)\mu k_8T_1+ \\ +(1-f_2u_2)\mu k_9M_1-(k_{10}T+k_{11}M)V_2-k_{12}V_2M-\delta_7V_2, \quad (11)$$

где первый член описывает продукцию вирусов зараженными Т-клетками, второй – продукцию вирусов зараженными макрофагами, третий – продукцию вирионов зараженными Т-клетками вследствие мутаций, четвертый – продукцию вирионов зараженными макрофагами вследствие мутаций, пятый – поглощение вирусов клетками при заражении клеток-мишеней, шестой – элиминацию вирусов системой врожденного иммунитета, седьмой – естественную гибель вирусов. Биологический смысл параметров системы и их допустимый диапазон из работы (Hadjiandreou et al., 2009) приведены в табл. 1.

Задача оптимального управления

В статье (Hadjiandreou et al., 2009) исследовалась возможность оптимизации режима введения ингибиторов протеазы (RDV) и обратной транскриптазы (3TC, ZDV), концентрации которых описываются следующими уравнениями одинаковой структуры:

$$C_i(t) = C_i(t_l)e^{-k_e^i(t-t_l)}+ \\ +\frac{F_iD_i}{V_c^i}\frac{k_a^i}{k_a^i+k_e^i}\left[e^{-k_e^i(t-t_l)}-e^{-k_a^i(t-t_l)}\right] \quad (i=1,2,3), \quad (12)$$

где i – индекс препарата, t_l – время введения препаратов, D_i – доза введенного препарата, F_i – абсолютная биодоступность препарата, k_a^i – скорость всасывания препарата, $k_e^i = Cl_i/V_c^i$ – константа скорости элиминации препарата (Cl_i – скорость выведения, а V_c^i – объем распределения препарата). Значения всех указанных выше параметров приведены в табл. 2.

Управляющие переменные u_1 и u_2 зависели от концентрации этих препаратов следующим образом:

$$u_1(t) = \frac{(C_2(t)/IC_{50}^2)+(C_3(t)/IC_{50}^3)}{1+(C_2(t)/IC_{50}^2)+(C_3(t)/IC_{50}^3)},$$

$$u_2(t) = \frac{C_1(t)}{C_1(t)+\omega IC_{50}^1},$$

где $C_i(t)$ – концентрация препарата i в плазме в момент времени t, а IC_{50}^i – средняя концентрация этого препарата, обеспечивающая 50 % ингибирование процессов размножения вирусов. Параметр ω является коэффициентом пересчета между значением средней концентрации препарата, обеспечивающей 50 % ингибирование процессов размножения вирусов IC_50, полученным *in vitro*, и ее же значением, полученным *in vivo*. В расчетах использовалось значение ω = 1. Целью оптимизации в исходной работе было достижение максимальной концентрации

Таблица 2. Значения параметров для уравнения фармакокинетики (12)

Параметр	RDV, C_1	3TC, C_2	ZDV, C_3
D [mg]	600	150	300
k_a [d^{-1}]	2.4	12	12
Cl [$L \cdot d^{-1}$]	$1.48 \cdot 10^4$	$5.6 \cdot 10^2$	$2.69 \cdot 10^3$
V_c [L]	28.7	91	112
F	1.0	0.86	0.64
τ [d]	0.5	0.5	0.5
IC_{50} [$mg \cdot L^{-1}$]	0.11	0.34	0.13

CD4$^+$ T-клеток (переменная T системы (1–11)) при минимальном индексе побочного воздействия препаратов (Joly, Pinto, 2006)

$$S_e = \sum_{i=1}^{N} \overline{e}_i \frac{C_i(t)}{\overline{C}_i},$$

где

$$\overline{e}_i = \frac{e_i}{\max\limits_i e_i}, \quad e_i = \sum_{j \in J_i} q_j h_{i,j}.$$

Здесь J_i – множество побочных эффектов от препарата i, $\overline{C}_i$ – средняя концентрация препарата i в стационарном состоянии при стандартной дозировке, т. е. в соответствии с правилами применения антиретровирусной терапии, e_i ($\overline{e}_i$) – величина (нормированная величина) побочного эффекта, вызванного препаратом i при стандартной дозировке, $h_{i,j}$ – частота проявления побочного эффекта j при воздействии препарата i при стандартной дозировке, а q_j – относительная величина побочного эффекта j, т. е. его «нежелательность».

Задача оптимального управления формулировалась как задача максимизации функционала, описывающего концентрацию CD4$^+$ T-лимфоцитов и выраженность побочных эффектов

$$\int_{t_0}^{t_f} [A_1 T - A_2 S_e] dt \to \max_{C_1, C_2, C_3}, \quad T \geq T_{AIDS}, t_0 \leq t \leq t_f,$$

где $A_1 = 1$ и $A_2 = 1000$ – весовые коэффициенты, t_0 и t_f – временные границы оптимизации, а условие $T \geq T_{AIDS}$ не дает концентрации клеток опуститься ниже порога, соответствующего развитию СПИДа (200 клеток в 1 mm^{-3}).

Было рассмотрено три набора значений параметров, соответствующих различным вариантам течения ВИЧ-инфекции: типичному течению (TP), быстро прогрессирующему течению (RP) и долго не прогрессирующему течению (LNTP). Значения параметров в этих наборах приведены в табл. 3 и 4.

В статье (Hadjiandreou et al., 2009) был найден более эффективный режим введения препаратов, основанный на результатах оптимизации, по сравнению с традиционным режимом лечения на примере модели для параметров пациента с типичным течением ВИЧ-инфекции с начальной концентрацией CD4$^+$ T-клеток, равной 350 mm^{-3}. Так, при традиционной методике лечения пациента удавалось удерживать концентрацию CD4$^+$ T-клеток выше порога СПИДа около 2500 дней, а с применением лечения, основанного на оптимизации режима введения антиретровирусных препаратов, – более 10000 дней, с более чем в 4 раза меньшим индексом побочного воздействия S_e.

Численные методы

Для численного интегрирования системы (1–11) мы использовали неявную схему второго порядка BDF2 (Hairer et al., 1987) на достаточно мелкой равномерной сетке, построенной в полуинтервале $t \geq 0$. Сходимость результатов по шагу сетки проверялась в ходе всех экспериментов, требующих интегрирования по времени. Для нахождения стационарных решений при заданных значениях параметров применялись методы символьных вычислений (Geddes

Таблица 3. Значения параметров модели (1–11), соответствующие типичному течению ВИЧ-инфекции (TP)

Параметр	Значение	Параметр	Значение	Параметр	Значение
s_1	10 mm^{-3}d^{-1}	k_5	3·10^{-6} mm^{3}d^{-1}	δ_4	5·10^{-3} d^{-1}
s_2	0.15 mm^{-3}d^{-1}	k_6	3.3·10^{-4} mm^{3}d^{-1}	δ_5	5·10^{-3} d^{-1}
s_3	5 mm^{-3}d^{-1}	k_7	6·10^{-9} mm^{3}d^{-1}	δ_6	0.015 d^{-1}
p_1	0.16 d^{-1}	k_8	5.37·10^{-1} d^{-1}	δ_7	2.39 d^{-1}
p_2	0.15 d^{-1}	k_9	2.85·10^{-1} d^{-1}	α_1	3·10^{-4} d^{-1}
S_1	55.6 mm^{-3}	k_{10}	7.79·10^{-6} mm^{3}d^{-1}	ψ	0.97
S_2	188 mm^{-3}	k_{11}	10^{-6} mm^{3}d^{-1}	φ	0.9
k_1	3.87·10^{-3} mm^{3}d^{-1}	k_{12}	4·10^{-5} mm^{3}d^{-1}	r	0.03 d^{-1}
k_2	10^{-6} mm^{3}d^{-1}	δ_1	0.02 d^{-1}	T_{max}	1500 mm^{-3}
k_3	4.5·10^{-4} mm^{3}d^{-1}	δ_2	0.28 d^{-1}	μ	0.001
k_4	5.22·10^{-4} mm^{3}d^{-1}	δ_3	0.05 d^{-1}	f_i	0.34

Таблица 4. Значения параметров модели (1–11), различных при разных вариантах течения ВИЧ-инфекции

Параметр	RP	TP	LTNP	Параметр	RP	TP	LTNP
p_1	0.13 d^{-1}	0.16 d^{-1}	0.20 d^{-1}	k_5	2.64·10^{-6} mm^{3}d^{-1}	3·10^{-6} mm^{3}d^{-1}	6.6·10^{-6} mm^{3}d^{-1}
p_2	0.1365 d^{-1}	0.15 d^{-1}	0.1638 d^{-1}	k_6	2.9·10^{-4} mm^{3}d^{-1}	3.3·10^{-4} mm^{3}d^{-1}	3.63·10^{-4} mm^{3}d^{-1}
S_1	50.0 mm^{-3}	55.6 mm^{-3}	55.6 mm^{-3}	k_7	5.28·10^{-9} mm^{3}d^{-1}	6·10^{-9} mm^{3}d^{-1}	6.6·10^{-9} mm^{3}d^{-1}
S_2	169.2 mm^{-3}	188 mm^{-3}	188 mm^{-3}	k_{12}	3.52·10^{-5} mm^{3}d^{-1}	4·10^{-5} mm^{3}d^{-1}	4.4·10^{-5} mm^{3}d^{-1}
k_3	3.96·10^{-4} mm^{3}d^{-1}	4.5·10^{-4} mm^{3}d^{-1}	9.9·10^{-4} mm^{3}d^{-1}	r	0.03	0.03	0.072

et al., 1992), реализованные в процедуре NSolve пакета Mathematica. Для трассирования решений по параметрам (т. е. для исследования зависимости стационарных решений системы (1–11) от параметров) использовался оригинальный алгоритм, предложенный нами в (Nechepurenko et al., 2020). Исследование асимптотической устойчивости заданного стационарного состояния сводилось к вычислению собственных значений линеаризованной относительно этого состояния системы и проверке, что все найденные собственные значения лежат строго в левой полуплоскости. Для вычисления собственных значений мы применили стандартный QR-алгоритм (Golub, Van Loan, 1989).

Результаты

Бифуркационный анализ

В этом разделе представлены результаты исследования зависимости стационарных решений модели динамики ВИЧ-инфекции от скорости активации макрофагов p_2, приводящей к их делению, при трех наборах значений остальных параметров (см. Материалы и методы). Ранее на примере математической модели вирусного гепатита В нами была показана ключевая роль в реализации различных режимов динамики гепатита параметра скорости активации врожденного иммунитета (Khristichenko et al., 2023), аналогом которого в данной модели является p_2. Параметр p_2 изменялся в интервале от 0.13 до 0.17. Диапазон варьирования параметра p_2 был выбран таким образом, чтобы содержать значения, которые соответствуют кинетике активации врожденного иммунитета при трех различных режимах течениях заболевания (типичное, длительно не прогрессирующее и быстро прогрессирующее), приведенных в табл. 4.

Результаты трассирования приведены на рис. 1–3. Вертикальная оранжевая пунктирная линия указывает значение параметра p_2, взятое из соответствующего набора параметров, сплошными линиями показаны устойчивые стационарные состояния, штриховой – неустойчивые; разные цвета обозначают различные стационарные состояния. Следует отметить, что ведущие собственные значения линеаризованных уравнений, отвечающих неустойчивым стационарным состояниям, во всех рассмотренных случаях были вещественными. Таким образом, устойчивые периодические решения, которые могли бы быть в окрестности неустойчивых стационарных состояний (Khristichenko, Nechepurenko, 2021), в рассмотренных случаях отсутствовали.

Бистабильность. Для типичного течения (см. рис. 1) видно, что бистабильность имеется при $0.138 < p_2 < 0.144$ (черная и зеленая линии) и при $0.147 < p_2 < 0.17$ (зеленая и фиолетовая линии). Для быстро прогрессирующего тече-

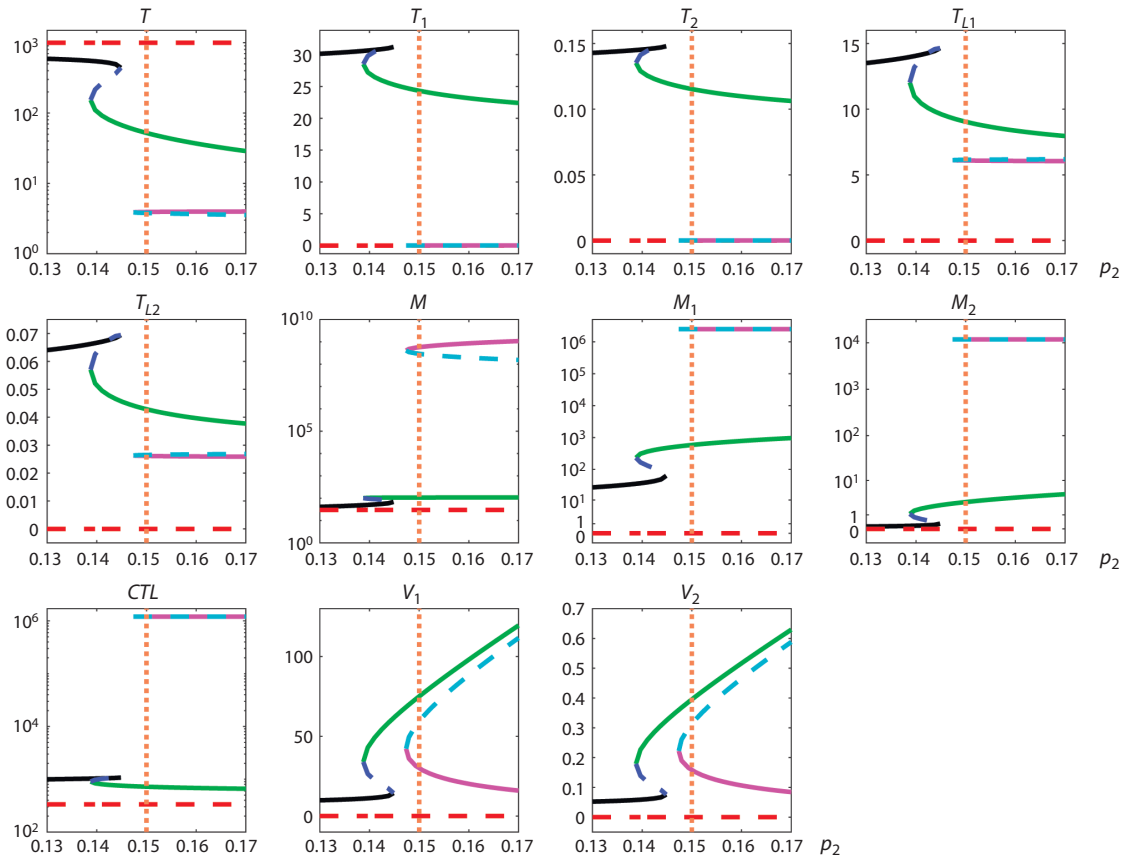


Рис. 1. Стационарные состояния при трассировании по параметру p_2 для типичного течения (ТР). Имеют место бистабильность и гистерезис.

Сплошными линиями обозначены устойчивые стационарные состояния, штриховыми – неустойчивые, разные цвета обозначают различные стационарные состояния. Вертикальная оранжевая пунктирная линия указывает значение параметра p_2, соответствующее типичному прогрессирующему течению инфекции.

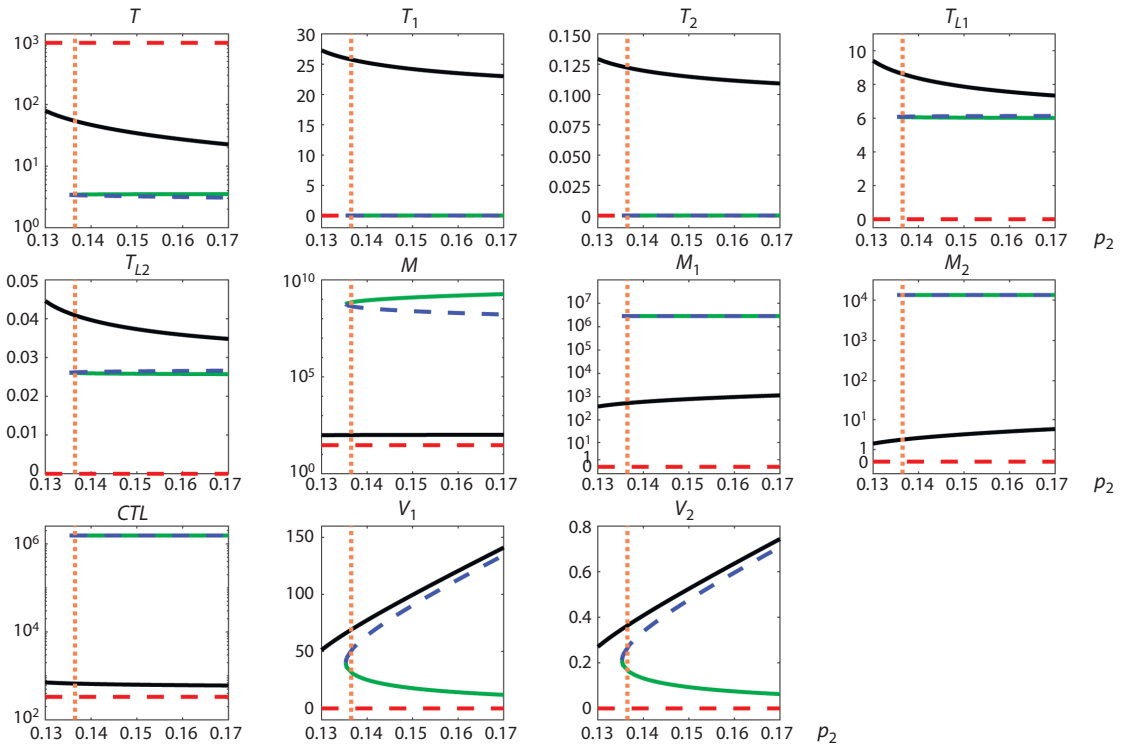


Рис. 2. Стационарные состояния при трассировании по p_2 для быстро прогрессирующего течения (RP). Имеет место бистабильность.

Сплошными линиями обозначены устойчивые стационарные состояния, штриховыми – неустойчивые, разные цвета обозначают различные стационарные состояния. Вертикальная оранжевая пунктирная линия указывает значение параметра p_2, соответствующее быстро прогрессирующему течению инфекции.

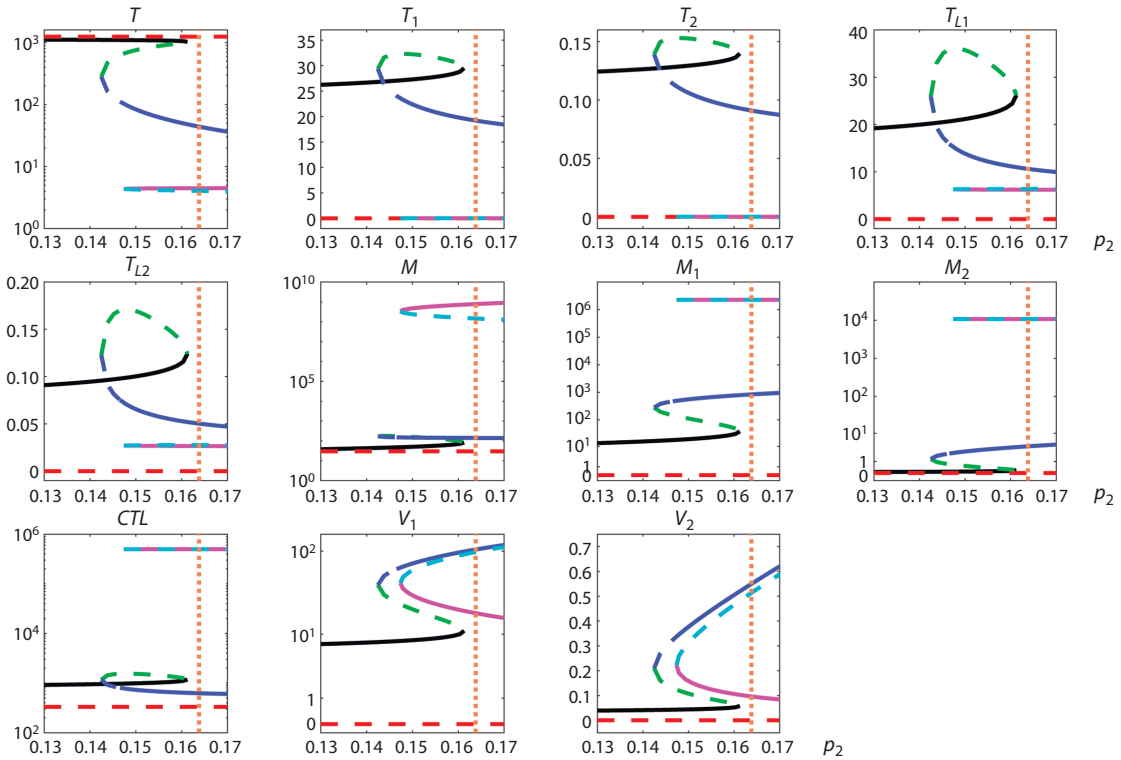


Рис. 3. Стационарные состояния при трассировании по p_2 для длительно не прогрессирующего течения (LTNP). Имеют место мультистабильность и гистерезис.

Сплошными линиями обозначены устойчивые стационарные состояния, штриховыми – неустойчивые, разные цвета обозначают различные стационарные состояния. Вертикальная оранжевая пунктирная линия указывает значение параметра p_2, соответствующее длительно не прогрессирующему течению инфекции.

ния (см. рис. 2) бистабильность имеется при $0.135 < p_2 < 0.17$ (черная и зеленая линии). Для длительно не прогрессирующего течения (см. рис. 3) бистабильность имеется при $0.161 < p_2 < 0.17$ (синяя и фиолетовая линии). Наличие двух различных устойчивых стационарных состояний означает возможность более легкой либо более тяжелой формы заболевания у одного и того же пациента в зависимости от предыстории. Заметим, что для быстро прогрессирующего течения оба положения равновесия характеризуются истощением популяции $CD4^+$ Т-клеток, при этом макрофаги являются доминирующим источником вирусов. Для таких пациентов задача лечения становится более сложной, поскольку необходимо найти изменения параметров системы, при которых появится положение равновесия с более высоким уровнем $CD4^+$ Т-клеток.

В целом полученные оценки областей существования бистабильности вместе с характеристиками бифуркационных диаграмм показывают, что по мере увеличения тяжести инфекционного процесса, т. е. при переходе от длительных непрогрессоров к типичным прогрессорам и далее к быстрым прогрессорам, увеличивается интервал значений параметра активации клеток врожденного иммунитета, в котором имеет место бистабильность. При этом меняется также характер бифуркационных диаграмм. Эти особенности отклика организма ВИЧ-инфицированного пациента следует учитывать и использовать при проектировании режимов иммуномодулирующих воздействий.

Мультистабильность. Свойство мультистабильности, как показано на рис. 3, имеет место в случае длительно не прогрессирующего течения при $0.146 < p_2 < 0.161$ (черная, синяя и фиолетовая линии). Эти устойчивые стационарные состояния соответствуют различным по тяжести и выраженности иммунного ответа формам течения болезни. Таким образом, спектр возможных устойчивых режимов динамики ВИЧ-1 инфекции более разнообразен у длительных непрогрессоров.

Гистерезис. Наличие свойства гистерезиса для данной модели продемонстрировано на рис. 1. В частности, поведение кривых показывает, что, если пациент, относящийся к типичным прогрессорам, изначально находился на нижней зеленой ветви при $p_2 = 0.14$, то для перехода на черную ветвь, характеризующуюся более высокой концентрацией Т-клеток и более низкой вирусной нагрузкой, достаточно уменьшить p_2 до значения, немного меньшего 0.138, что вызовет спонтанный переход в состояние, изображенное черной линией. После этого можно увеличить значение параметра p_2 до исходного, оставаясь на черной ветви.

Гистерезис также имеет место при длительно не прогрессирующем течении, что демонстрирует рис. 3. Состояние, изображенное синей линией, при $p_2 = 0.155$ устойчиво, однако теряет устойчивость при p_2, меньшем 0.146, и при дальнейшем уменьшении значения параметра система из менее благоприятного состояния (зеленая ветвь) перейдет в устойчивое состояние с более высокой концентрацией $CD4^+$ Т-клеток и меньшей вирусной нагрузкой, изображенное черной сплошной линией. После этого можно вернуться к первоначальному значению параметра, оставаясь в этом устойчивом стационарном состоянии.

Практически важным является вопрос о кинетике перехода между различными стационарными состояниями при использовании свойства гистерезиса. Для типичного развития заболевания продемонстрирован переход из менее благоприятного состояния в более благоприятное для системы с гистерезисом (рис. 4). Для реализации пе-

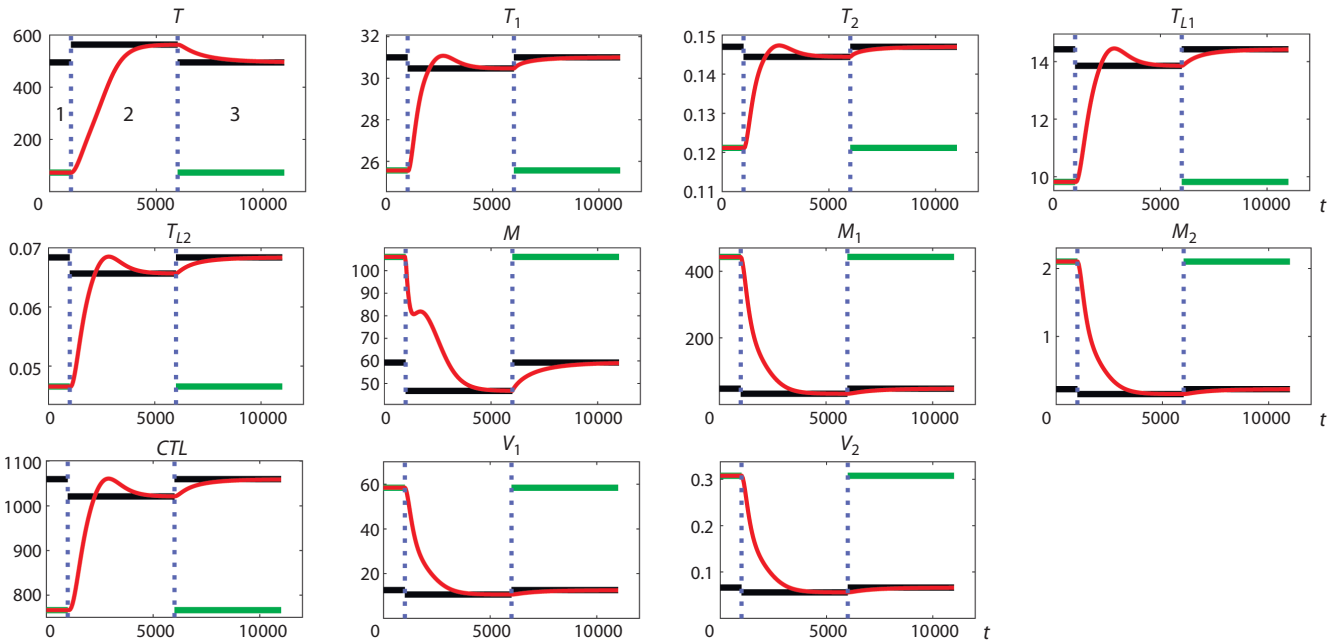


Рис. 4. Демонстрация кинетики перехода от менее благоприятного стационарного состояния к более благоприятному при наличии гистерезиса для типичного течения (ТР), где $p_2 = 0.143$ в областях 1 и 3 и $p_2 = 0.136$ в области 2.

Красной сплошной линией показана динамика переменных модели, синими вертикальными пунктирными – разбиение на области 1–3, горизонтальными сплошными линиями – устойчивые стационарные состояния переменных в этих областях. Здесь и на рис. 5–8 шкала по времени в сутках.

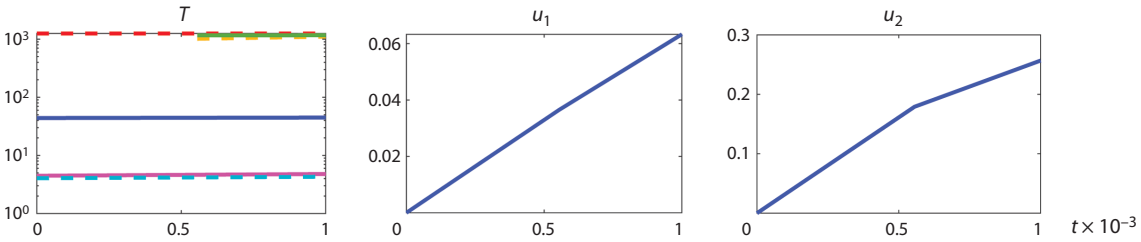


Рис. 5. Зависимости от времени t (в сутках) переменной T стационарного состояния и управляющих воздействий $u_1(t)$ и $u_2(t)$ при $0 \le t \le 0.001$ для длительно не прогрессирующего течения (LTNP).

Сплошные линии на графике $T(t)$ соответствуют устойчивым стационарным состояниям, штриховые – неустойчивым.

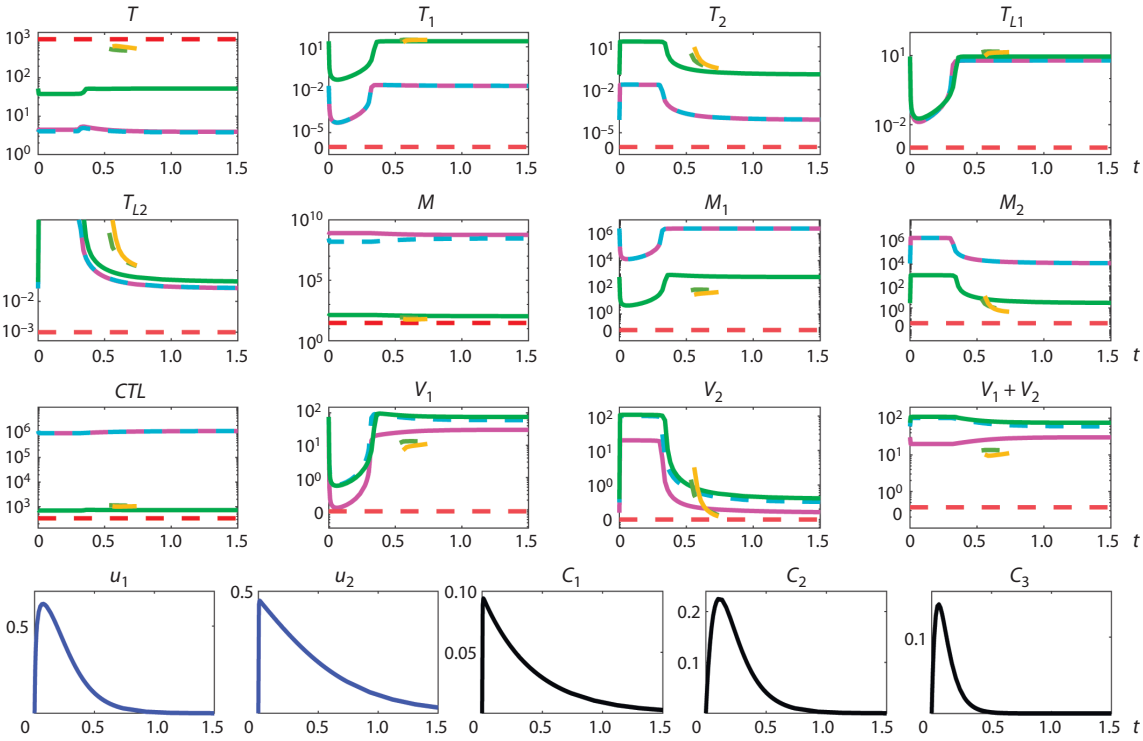


Рис. 6. Стационарные состояния и управляющие воздействия для типичного течения (TP).

Здесь и на рис. 7 и 8: сплошными линиями обозначены устойчивые стационарные состояния, штриховыми – неустойчивые, различные цвета обозначают различные стационарные состояния.

рехода требуется около 5000 суток при неизменных значениях других параметров системы. Эти результаты обосновывают актуальность детального изучения таких переходов.

Изменение стационарных состояний при однократном введении препаратов

Представляют самостоятельный интерес изменения стационарных состояний системы при оптимальном управлении (Hadjiandreou et al., 2009). С этой целью мы исследовали зависимость положений равновесия от времени при терапевтических воздействиях $u_1(t)$, $u_2(t)$, которые входят в правые части уравнений модели через слагаемые, описывающие заражение клеток-мишеней и размножение вирусов. На рис. 5 видно появление двух новых стационарных состояний при $t > 0.0005$, т. е. изменение структуры фазового пространства модели.

Графики, демонстрирующие изменения стационарных состояний при введении препаратов RDV, 3TC и ZDV, влияние которых моделируется с помощью функций $C_1(t)$, $C_2(t)$, $C_3(t)$ через переменные управления u_1 и u_2, представлены на рис. 6–8. Введение препаратов осуществляется единожды в момент времени $t = 0$, сплошной линией обозначены устойчивые стационарные состояния, а штриховой – неустойчивые; разные цвета обозначают различные стационарные состояния. Результаты расчета указывают, что по мере изменения значений управляющих переменных появляются и потом исчезают как устойчивые, так и неустойчивые положения равновесия. То есть применение методов оптимального управления приводит к изменению структуры фазового пространства модели.

Для всех трех вариантов течения вирусной инфекции для одной ветви стационарных решений имеет место снижение на коротком временном интервале значений пере-

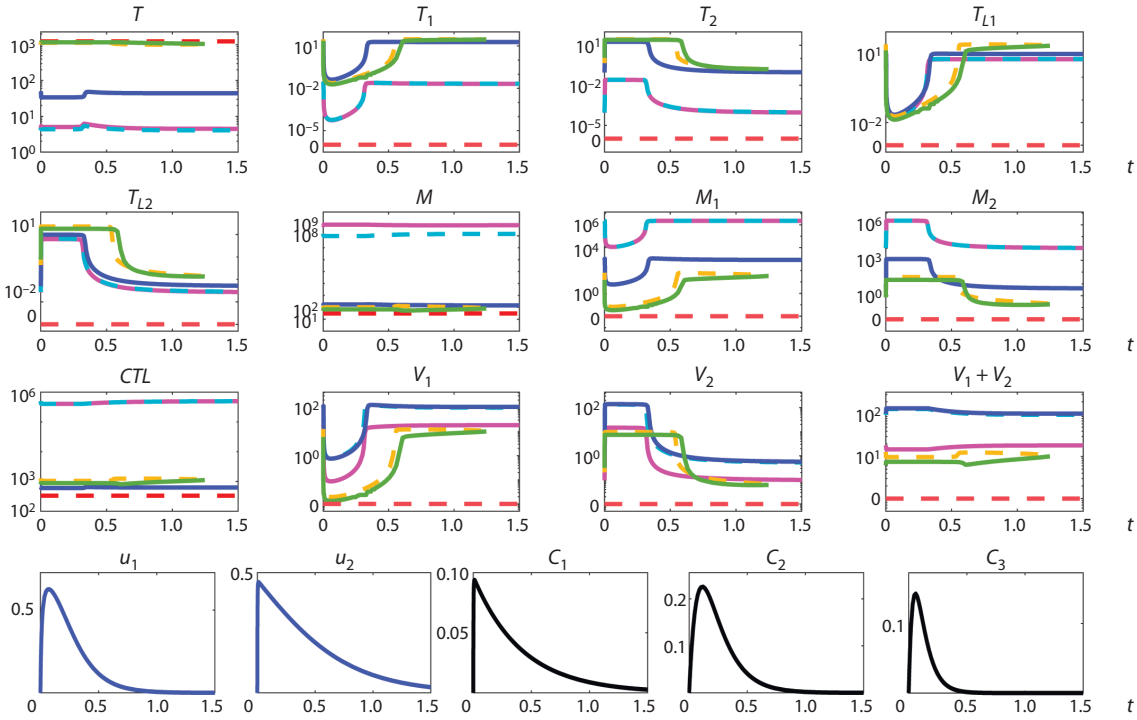


Рис. 7. Стационарные состояния модели и управляющие воздействия для длительно не прогрессирующего течения (LTNP).

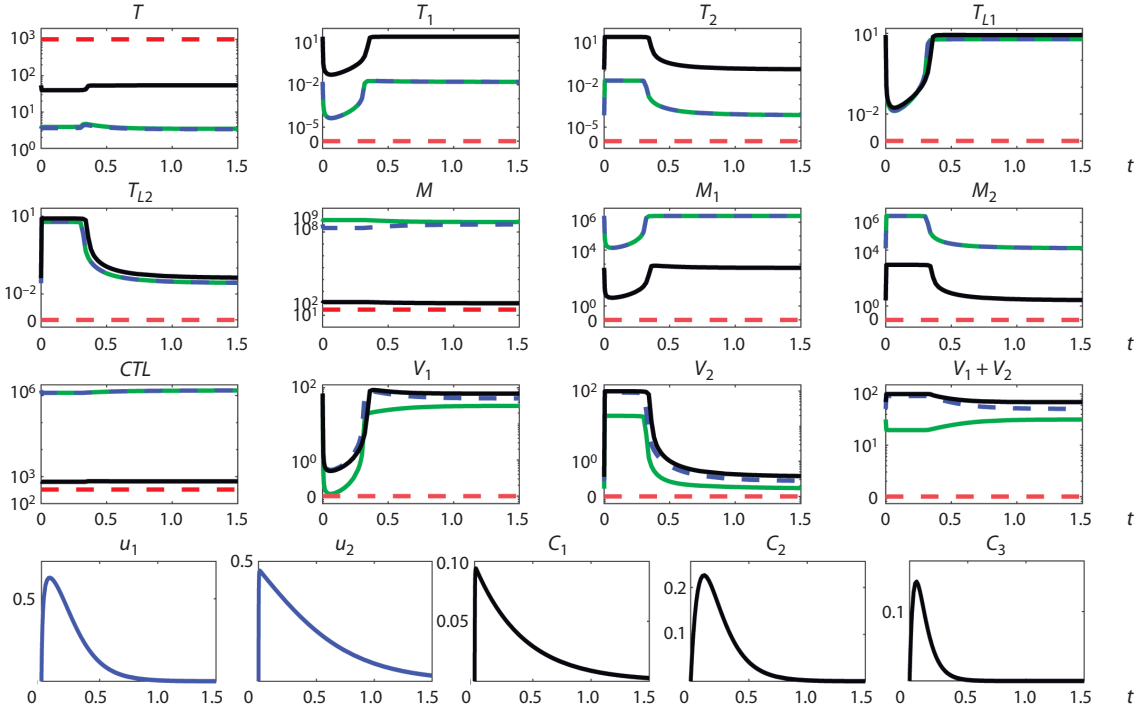


Рис. 8. Стационарные состояния модели и управляющие воздействия для быстро прогрессирующего течения (RP).

менных, характеризующих число CD4$^+$ Т-клеток, и повышение вирусной нагрузки за счет увеличения численности мутантов на фоне снижения стационарных концентраций вирусов дикого типа. На второй устойчивой ветви имеет место обратный процесс. При этом в случае длительно не прогрессирующего течения ВИЧ-1 появляется третья ветвь устойчивого положения равновесия, характеризующаяся низкой вирусной нагрузкой, т. е. отвечающая благоприятной динамике. Таким образом, воздействие оптимального управления на характеристики положений равновесия существенно зависит от варианта течения болезни (параметров системы) и окрестности положения равновесия, в котором находится пациент в случае бистабильности.

Таким образом, имеет место одинаковый качественный характер отклика на возмущение правых частей уравнений. Структура фазового пространства меняется, появляются, а потом, по мере ослабления управляющего воздействия, исчезают как устойчивые, так и неустойчивые стационарные состояния.

Обсуждение

Ситуация устойчивого сосуществования популяции ВИЧ-1 и иммунных процессов в организме человека в различных количественных соотношениях является принципиально важной для разработки новых стратегий терапии ВИЧ-1, относящихся к категории функционального лечения (Bocharov et al., 2022). По существу, речь идет о возможности перевода системы «патоген–организм человека» из клинически более тяжелого состояния в более легкое устойчивое состояние за счет активации механизмов иммунной защиты без дальнейшего применения антиретровирусных препаратов, блокирующих размножение вирусов. Наличие би- или мультистабильности свидетельствует о том, что путем возмущения соответствующей траектории системы в фазовом пространстве можно осуществить перевод инфекционного заболевания в более благоприятный режим. В качестве инструментов для построения соответствующего управления существуют как классические методы оптимального управления (Hadjiandreou et al., 2009; Bocharov et al., 2015), так и предложенные нами ранее методы на основе оптимальных возмущений (Нечепуренко, Христиченко, 2019; Khristichenko, Nechepurenko, 2022). Вместе с тем возможна ситуация, когда для перевода системы в область би- или мультистабильности требуется изменение кинетических параметров биологических и физиологических процессов. Наличие гистерезиса позволяет разрабатывать подходы к лечению, заключающемуся во временном параметрическом сдвиге с последующим возвращением к исходным значениям измененных параметров. Выявленные нами свойства математической модели ВИЧ-1 инфекции, которая имеет достаточно типичную структуру, теоретически подтверждают потенциальную возможность соответствующих комбинированных иммунотерапевтических воздействий (Landovitz et al., 2023).

Полученные оценки областей существования бистабильности вместе с характеристиками бифуркационных диаграмм показывают, что по мере увеличения тяжести инфекционного процесса, т. е. при переходе от длительных непрогрессоров к типичным прогрессорам и далее к быстрым прогрессорам, увеличивается интервал значений параметра активации клеток врожденного иммунитета, в котором имеет место бистабильность. При этом меняется также характер бифуркационных диаграмм. Эти особенности отклика организма ВИЧ-инфицированного пациента следует учитывать и использовать при проектировании режимов иммуномодулирующих воздействий.

Нами показано, что воздействие оптимального управления на характеристики положений равновесия существенно зависит от фенотипа ВИЧ-1 инфекции (параметров системы) и окрестности того положения равновесия, в котором находится пациент в случае би- или мультистабильности.

Заключение

В данной работе проведен расчет и численный анализ стационарных состояний системы уравнений математической модели ВИЧ-1 инфекции для наборов параметров, отвечающих фенотипически различным вариантам течения инфекционного заболевания: типичному, длительно не прогрессирующему и быстро прогрессирующему. Результаты бифуркационного анализа модели ВИЧ-1 инфекции указывают на то, что для эффективного функционального лечения больных ВИЧ-инфекцией требуется развитие персонализированного подхода, учитывающего как свойства популяции квазивидов ВИЧ-1, так и иммунный статус пациента, и формируют теоретическую основу для разработки комбинированных иммунотерапевтических воздействий для лечения ВИЧ-1.

Список литературы / References

Нечепуренко Ю.М., Христиченко М.Ю. Вычисление оптимальных возмущений для систем с запаздыванием. *Журн. вычислит. математики и мат. физики*. 2019;59(5):775-791. DOI 10.1134/S0044466919050120
[Nechepurenko Y.M., Khristichenko M.Y. Computation of optimal disturbances for delay systems. *Comput. Math. and Math. Phys.* 2019;59(5):731-746. DOI 10.1134/S0965542519050129]

Савинкова А.А., Савинков Р.С., Бахметьев Б.А., Бочаров Г.А. Математическое моделирование и управление динамикой ВИЧ-инфекции. *Вестн. Рос. ун-та дружбы народов. Сер. Медицина*. 2019;23(1):79-103. DOI 10.22363/2313-0245-2019-23-1-79-103
[Savinkova A.A., Savinkov R.S., Bakhmetyev B.A., Bocharov G.A. Mathematical modeling and control of HIV infection dynamics taking into account hormonal regulation. *Vestnik Rossiyskogo Universiteta Druzhby Narodov. Seriya Meditsina = RUDN Journal of Medicine*. 2019;23(1):79-103. DOI 10.22363/2313-0245-2019-23-1-79-103 (in Russian)]

Христиченко М.Ю., Нечепуренко Ю.М., Гребенников Д.С., Бочаров Г.А. Численный анализ стационарных решений систем с запаздывающим аргументом в математической иммунологии. *Соврем. математика. Фундам. направления*. 2022;68(4):686-703. DOI 10.22363/2413-3639-2022-68-4-686-703
[Khristichenko M.Yu., Nechepurenko Yu.M., Grebennikov D.S., Bocharov G.A. Numerical analysis of stationary solutions of systems with delayed argument in mathematical immunology. *Sovremennaya Matematika. Fundamental'nye Napravleniya = Contemporary Mathematics. Fundamental Directions*. 2022;68(4):686-703. DOI 10.22363/2413-3639-2022-68-4-686-703 (in Russian)]

Akın E., Yeni G., Perelson A.S. Continuous and discrete modeling of HIV-1 decline on therapy. *J. Math. Biol.* 2020;81(1):1-24. DOI 10.1007/s00285-020-01492-z

Banks H.T., Hu S., Rosenberg E. A dynamical modeling approach for analysis of longitudinal clinical trials in the presence of missing endpoints. *Appl. Math. Lett*. 2017;63:109-117. DOI 10.1016/j.aml.2016.07.002

Bocharov G., Chereshnev V., Gainova I., Bazhan S., Bachmetyev B., Argilaguet J., Martinez J., Meyerhans A. Human immunodeficiency virus infection: from biological observations to mechanistic mathematical modelling. *Math. Model. Nat. Phenom.* 2012;7(5):78-104. DOI 10.1051/mmnp/20127507

Bocharov G., Kim A., Krasovskii A., Chereshnev V., Glushenkova V., Ivanov A. An extremal shift method for control of HIV infection dynamics. *Russ. J. Numer. Anal. Math. Model.* 2015;30(1):11-25. DOI 10.1515/rnam-2015-0002

Bocharov G.A., Nechepurenko Y.M., Khristichenko M.Y., Grebennikov D.S. Optimal perturbations of systems with delayed independent variables for control of dynamics of infectious diseases based on multicomponent actions. *J. Math. Sci*. 2021;253(5):618-641. DOI 10.1007/s10958-021-05258-w

Bocharov G., Grebennikov D., Cebollada Rica P., Domenjo-Vila E., Casella V., Meyerhans A. Functional cure of a chronic virus infection by shifting the virus – host equilibrium state. *Front. Immunol*. 2022;13:904342. DOI 10.3389/fimmu.2022.904342

Gandhi R.T., Bedimo R., Hoy J.F., Landovitz R.J., Smith D.M., Eaton E.F., Lehmann C., Springer S.A., Sax P.E., Thompson M.A., Benson C.A., Buchbinder S.P., Del Rio C., Eron J.J., Jr., Günthard H.F., Molina J.-M., Jacobsen D.M., Saag M.S. Antiretroviral drugs for treatment and prevention of HIV infection in adults: 2022 recommendations of the International Antiviral Society-USA Panel. *JAMA*. 2023;329(1):63-84. DOI 10.1001/jama.2022.22246

Geddes K.O., Czapor S.R., Labahn G. Algorithms for Computer Algebra. Boston: Kluwer Academic, 1992

Golub G.H., Van Loan C.F. Matrix Computations. Baltimore: Johns Hopkins Univ. Press, 1989

Grossman Z., Singh N.J., Simonetti F.R., Lederman M.M., Douek D.C., Deeks S.G., Kawabe T., Bocharov G., Meier-Schellersheim M., Alon H., Chomont N., Grossman Z., Sousa A.E., Margolis L., Maldarelli F. "Rinse and replace": boosting T cell turnover to reduce HIV-1 reservoirs. *Trends Immunol.* 2020;41(6):466-480. DOI 10.1016/j.it.2020.04.003

Hadjiandreou M.M., Conejeros R., Wilson I. HIV treatment planning on a case-by-case basis. *Int. J. Bioeng. Life Sci.* 2009;3(8):387-396

Hairer E., Nørsett S.P., Wanner G. Solving Ordinary Differential Equations I. Springer Series in Computational Mathematics. Vol. 8. Berlin: Springer, 1987. DOI 10.1007/978-3-662-12607-3

Joly M., Pinto J.M. Role of mathematical modeling on the optimal control of HIV-1 pathogenesis. *AIChE J.* 2006;52(3):856-884. DOI 10.1002/aic.10716

Khristichenko M.Y., Nechepurenko Y.M. Computation of periodic solutions to models of infectious disease dynamics and immune response. *Russ. J. Numer. Anal. Math. Model.* 2021;36(2):87-99. DOI 10.1515/rnam-2021-0008

Khristichenko M.Y., Nechepurenko Y.M. Optimal disturbances for periodic solutions of time-delay differential equations. *Russ. J. Numer. Anal. Math. Model.* 2022;37(4):203-212. DOI 10.1515/rnam-2022-0017

Khristichenko M., Nechepurenko Y., Grebennikov D., Bocharov G. Numerical study of chronic hepatitis B infection using Marchuk–Petrov model. *J. Bioinform. Comput. Biol*. 2023;21(2):2340001. DOI 10.1142/S0219720023400012

Landovitz R.J., Scott H., Deeks S.G. Prevention, treatment and cure of HIV infection. *Nat. Rev. Microbiol*. 2023;21(10):657-670. DOI 10.1038/s41579-023-00914-1

Ludewig B., Stein J.V., Sharpe J., Cervantes-Barragan L., Thiel V., Bocharov G. A global "imaging" view on systems approaches in immunology. *Eur. J. Immunol.* 2012;42(12):3116-3125. DOI 10.1002/eji.201242508

Nechepurenko Y., Khristichenko M., Grebennikov D., Bocharov G. Bistability analysis of virus infection models with time delays. *Discrete Cont. Dyn. Syst. - S*. 2020;13(9):2385-2401. DOI 10.3934/dcdss.2020166

Niessl J., Baxter A.E., Mendoza P., Jankovic M., Cohen Y.Z., Butler A.L., Lu C.-L., Dubé M., Shimeliovich I., Gruell H., Klein F., Caskey M., Nussenzweig M.C., Kaufmann D.E. Combination anti-HIV-1 antibody therapy is associated with increased virus-specific T cell immunity. *Nat. Med*. 2020;26(2):222-227. DOI 10.1038/s41591-019-0747-1

Nowak M.A., May R.M. Virus Dynamics: Mathematical Principles of Immunology and Virology. Oxford: Oxford Univ. Press, 2000

Perelson A.S., Nelson P.W. Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev.* 1999;41(1):3-44. DOI 10.1137/S0036144598335107

Rasmussen T.A., Søgaard O.S. Clinical interventions in HIV cure research. In: Zhang L., Lewin S.R. (Eds.) HIV Vaccines and Cure. Advances in Experimental Medicine and Biology. Vol. 1075. Singapore: Springer, 2018;285-318. DOI 10.1007/978-981-13-0484-2_12

Trickey A., Zhang L., Gill M.J., Bonnet F., Burkholder G., Castagna A., Cavassini M., Cichon P., Crane H., Domingo P., Grabar S., Guest J., Obel N., Psichogiou M., Rava M., Reiss P., Rentsch C.T., Riera M., Schuettfort G., Silverberg M.J., Smith C., Stecher M., Sterling T.R., Ingle S.M., Sabin C.A., Sterne J.A.C. Associations of modern initial antiretroviral drug regimens with all-cause mortality in adults with HIV in Europe and North America: a cohort study. *Lancet HIV*. 2022;9(6):e404-e413. DOI 10.1016/S2352-3018(22)00046-7

Villani A.-C., Sarkizova S., Hacohen N. Systems immunology: learning the rules of the immune system. *Annu. Rev. Immunol.* 2018;36(1):813-842. DOI 10.1146/annurev-immunol-042617-053035

ORCID ID

G.A. Bocharov orcid.org/0000-0002-5049-0656

Благодарности. Работа выполнена при финансовой поддержке Российского научного фонда, проект № 22-71-10028.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 14.07.2023. После доработки 15.09.2023. Принята к публикации 19.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):768-775
DOI 10.18699/VJGB-23-89

Перевод на английский язык https://vavilov.elpub.ru/jour

Применение генных сетей к анализу результатов метаболомного скрининга плазмы крови пациентов с послеоперационным делирием

В.А. Иванисенко[1, 2, 7 *], Н.В. Басов[2, 3 *], А.А. Макарова[1], А.С. Вензель[1, 7], А.Д. Рогачев[2, 3] ✉, П.С. Деменков[1, 2, 7], Т.В. Иванисенко[1, 2, 7], М.А. Клещев[1], Е.В. Гайслер[2, 3], Г.Б. Мороз[4], В.В. Плеско[4], Ю.С. Сотникова[2, 3, 5], Ю.В. Патрушев[2, 5], В.В. Ломиворотов[4, 6], Н.А. Колчанов[1, 2], А.Г. Покровский[2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Новосибирский институт органической химии им. Н.Н. Ворожцова Сибирского отделения Российской академии наук, Новосибирск, Россия
[4] Национальный медицинский исследовательский центр им. академика Е.Н. Мешалкина Министерства здравоохранения Российской Федерации, Новосибирск, Россия
[5] Федеральный исследовательский центр «Институт катализа им. Г.К. Борескова Сибирского отделения Российской академии наук», Новосибирск, Россия
[6] Медицинский центр им. Милтона Херши, Херши, Пенсильвания, США
[7] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
✉ rogachev@nioch.nsc.ru

Аннотация. Послеоперационный делирий (ПОД) является серьезным осложнением, приводящим к нарушению когнитивных функций пациентов, увеличению длительности госпитализации, а также повышению расходов на лечение пациента. Проблема ранней диагностики ПОД приобретает особую важность в случае кардиохирургических операций, поскольку частота развития такого осложнения у некоторых категорий пациентов превышает 50 %. Известно, что в развитие ПОД большой вклад вносят нейровоспаление, дисбаланс нейромедиаторов, нарушение нейроэндокринной регуляции и межнейрональных связей, однако молекулярно-генетические механизмы ПОД у пациентов, перенесших кардиохирургические операции, а также метаболомные диагностические маркеры, до сих пор плохо изучены. В данной работе с помощью метода высокоэффективной жидкостной хроматографии с масс-спектрометрической детекцией (ВЭЖХ-МС/МС) был проведен анализ содержания ряда сфингомиелинов в плазме крови пациентов старше 65 лет, взятой после операции на сердце в условиях искусственного кровообращения. Найдено четыре статистически значимо различающихся по содержанию сфингомиелина у пациентов с ПОД по сравнению с пациентами, у которых не развился ПОД (контрольная группа). С помощью реконструкции генных сетей, описывающих генетическую регуляцию пути метаболизма сфинголипидов, определены 82 регуляторных белка, из которых 47 – регуляторы экспрессии генов, кодирующих ферменты метаболического пути, и 35 – регуляторы активности, деградации и транспорта ферментов данного пути. Анализ перепредставленности заболеваний, с которыми ассоциированы эти регуляторные белки, показал, что регуляторы можно разбить на две группы, ассоциированные с сердечно-сосудистыми патологиями и с нервно-психическими заболеваниями соответственно. Регуляторы, ассоциированные с сердечно-сосудистыми патологиями, ожидаемо связаны с воздействием на ткани миокарда во время операции. Сделано предположение, что нарушение функции регуляторов, ассоциированных с нервно-психическими заболеваниями, может специфически обусловливать развитие ПОД после кардиохирургической операции. Таким образом, выявленные регуляторные гены могут представлять основу для планирования дальнейших экспериментов по изучению нарушений на уровне экспрессии данных генов, а также нарушения функции кодируемых ими белков у пациентов с ПОД. Идентифицированные значимые сфинголипиды могут рассматриваться как потенциальные маркеры послеоперационного делирия.
Ключевые слова: ВЭЖХ-МС/МС; метаболомика; липидомика; послеоперационный делирий; кардиохирургия; биомаркеры; сфинголипиды; генные сети; ANDSystem.

Для цитирования: 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

Gene networks for use in metabolomic data analysis of blood plasma from patients with postoperative delirium

V.A. Ivanisenko[1, 2, 7 *], N.V. Basov[2, 3 *], A.A. Makarova[1], A.S. Venzel[1, 7], A.D. Rogachev[2, 3] ✉, P.S. Demenkov[1, 2, 7], T.V. Ivanisenko[1, 2, 7], M.A. Kleshchev[1], E.V. Gaisler[2, 3], G.B. Moroz[4], V.V. Plesko[4], Y.S. Sotnikova[2, 3, 5], Y.V. Patrushev[2, 5], V.V. Lomivorotov[4, 6], N.A. Kolchanov[1, 2], A.G. Pokrovsky[2]

* Авторы внесли равный вклад в работу.

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] N.N. Vorozhtsov Novosibirsk Institute of Organic Chemistry of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] E. Meshalkin National Medical Research Center of the Ministry of Health of Russian Federation, Novosibirsk, Russia
[5] Boreskov Institute of Catalysis of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[6] Penn State Health Milton S. Hershey Medical Center, Hershey, PA, USA
[7] Kurchatov Genomic Center of ICG SB RAS, Novosibirsk, Russia
✉ rogachev@nioch.nsc.ru

Abstract. Postoperative delirium (POD) is considered one of the most severe complications, resulting in impaired cognitive function, extended hospitalization, and higher treatment costs. The challenge of early POD diagnosis becomes particularly significant in cardiac surgery cases, as the incidence of this complication exceeds 50 % in certain patient categories. While it is known that neuroinflammation, neurotransmitter imbalances, disruptions in neuroendocrine regulation, and interneuronal connections contribute significantly to the development of POD, the molecular, genetic mechanisms of POD in cardiac surgery patients, along with potential metabolomic diagnostic markers, remain inadequately understood. In this study, blood plasma was collected from a group of patients over 65 years old after cardiac surgery involving artificial circulation. The collected samples were analyzed for sphingomyelin content and quantity using high-performance liquid chromatography coupled with mass spectrometry (HPLC-MS/MS) methods. The analysis revealed four significantly different sphingomyelin contents in patients with POD compared to those who did not develop POD (control group). Employing gene network reconstruction, we perceived a set of 82 regulatory enzymes affiliated with the genetic coordination of the sphingolipid metabolism pathway. Within this set, 47 are assumed to be regulators of gene expression, governing the transcription of enzymes pivotal to the metabolic cascade. Complementing this, an additional assembly of 35 regulators are considered to be regulators of activity, degradation, and translocation dynamics of enzymes integral to the aforementioned pathway. Analysis of the overrepresentation of diseases with which these regulatory proteins are associated showed that the regulators can be categorized into two groups, associated with cardiovascular pathologies (CVP) and neuropsychiatric diseases (NPD), respectively. The regulators associated with CVP are expectedly related to the effects on myocardial tissue during surgery. It is hypothesized that dysfunction of NPD-associated regulators may specifically account for the development of POD after cardiac surgery. Thus, the identified regulatory genes may provide a basis for planning further experiments, in order to study disorders at the level of expression of these genes, as well as impaired function of proteins encoded by them in patients with POD. The identified significant sphingolipids can be considered as potential markers of POD.
Key words: LC-MS/MS; metabolomics; lipidomics; postoperative delirium; cardiac surgery; biomarkers; sphingolipids; gene networks; ANDSystem.

For citation: 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 Journal of Genetics and Breeding*. 2023;27(7):768-775. DOI 10.18699/VJGB-23-89

Введение

Послеоперационный делирий (ПОД) – серьезное осложнение раннего послеоперационного периода. В сердечно-сосудистой хирургии его частота составляет 52 % (Brown, 2014). Развитие ПОД приводит к ухудшению прогноза, в том числе к увеличению длительности госпитализации, росту количества осложнений и летальности, нарушению когнитивных функций и физического состояния, а также повышению расходов на лечение пациента (Pisani et al., 2009; Gottesman et al., 2010). Делирий и послеоперационные когнитивные нарушения чаще всего возникают у пациентов старше 60 лет (Morimoto et al., 2009). Этому способствуют такие факторы, как гипоксия ЦНС, эмболии, выброс нейротрансмиттеров, системный воспалительный ответ и другие нарушения, включая метаболические (Wimmer-Greinecker et al., 1998; Cerejeira et al., 2010).

Метаболомика – это направление в биоаналитической химии, связанное с идентификацией и количественным определением низкомолекулярных метаболитов (<1500 Да). Метаболомный подход может быть использован для поиска ассоциаций между метаболическими сигнатурами и фенотипами заболеваний. В частности, метаболомные методы позволяют детектировать низкомолекулярные метаболиты, способные пересекать гематоэнцефалический барьер, что делает метаболомный анализ мощным инструментом для выявления маркеров делирия (Ke et al., 2019). Так, в ряде работ было показано, что нарушения энергетического метаболизма, биосинтеза аминокислот, дефицит омега-3 и омега-6 жирных кислот, а также дисфункция глутамат-глутаминового цикла связаны с послеоперационным делирием при несердечных операциях (Guo et al., 2019; Tripp et al., 2021).

Ранее в наших исследованиях применялись методы метаболомного скрининга и реконструкции генных сетей для поиска биомаркеров патологий. Так, с помощью статистического анализа метаболомных профилей цереброспинальной жидкости (ЦСЖ) и плазмы крови пациентов с глиомой высокой степени злокачественности, полученных с применением метода ВЭЖХ-МС/МС, нами обнаружены корреляции метаболомных профилей плазмы крови и ЦСЖ (Rogachev et al., 2021). Метаболомный анализ в сочетании с реконструкцией генных сетей с применением ANDSystem для интерпретации метаболомных данных (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022) позволил установить ключевые белки SARS-CoV-2, взаимодействия которых с белками человека могли приводить к нарушению метаболических процессов у пациентов с COVID-19 (Ivanisenko V.A. et al., 2022).

Сфингомиелины являются одними из основных фосфолипидов, составляющих гидрофобный матрикс плазма-

тических мембран клеток млекопитающих, однако в ответ на стресс сфингомиелины могут расщепляться сфингомиелиназой на фосфатидилхолин и церамид, которые выполняют сигнальную функцию. Изменения в метаболизме сфингомиелинов могут влиять на баланс нейромедиаторов в мозгу, нарушение нейронных связей и индукцию нейровоспаления, что делает их важным объектом для изучения механизмов патогенеза делирия (Wang, Shen, 2018; Xiao et al., 2023).

В настоящей работе с помощью ВЭЖХ-МС/МС был проведен анализ содержания девяти фосфолипидов, относящихся к классу сфингомиелинов, в плазме крови пациентов, перенесших операцию на сердце. Найдено четыре статистически значимо различающихся сфингомиелина по содержанию у пациентов с ПОД по сравнению с пациентами, у которых не развился ПОД (контрольная группа).

Для объяснения возможных механизмов нарушений метаболизма сфинголипидов с помощью ANDSystem нами были реконструированы генные сети, описывающие генетическую регуляцию пути KEGG Sphingolipid metabolism (hsa: 00600). Анализ генных сетей позволил выявить 35 регуляторов транспорта, активности и деградации ферментов данного пути, а также 47 регуляторов экспрессии генов, кодирующих эти ферменты.

Материалы и методы

Пациенты. В исследование были включены пациенты старше 65 лет, которым проводилась кардиохирургическая операция в условиях искусственного кровообращения. Критериями исключения были: экстренное вмешательство, операция на аорте, гемодинамически значимые стенозы сонных артерий, болезнь Паркинсона, цирроз печени (Чайлд В или С), прием антихолинергических препаратов, антидепрессантов, противоэпилептических и химиотерапевтических препаратов. Набор пациентов осуществлялся с июня 2019 г. по январь 2021 г. Всего в исследование было включено 39 пациентов (половозрастная характеристика представлена в табл. 1). В течение 5 дней после операции пациенты оценивались на наличие послеоперационного делирия при помощи теста CAM-ICU (Confusion Assessment Method for the Intensive Care Unit). Первый тест проводился через 6–8 ч после операции, далее – два раза в сутки. Наличием делирия считалось, если тест CAM-ICU был положительным хотя бы один раз.

Исследование одобрено этическим комитетом Национального медицинского исследовательского центра им. Е.Н. Мешалкина (Новосибирск, Россия).

Отбор образцов крови и пробоподготовка. Образцы крови были взяты у пациентов через 24 ч после проведения кардиохирургической операции. Венозную кровь собирали в пробирки BD Vacutainer® KEDTA объемом 10 мл, содержащие ЭДТА калия в качестве антикоагулянта. Плазму отделяли от клеток крови центрифугированием в течение 15 мин при 2000 g и 4 °С, разделяли на аликвоты и хранили в замороженном виде при –80 °С до дальнейшего использования.

Все образцы обрабатывали одновременно в соответствии с протоколом, описанным в работе (Li et al., 2017): к 100 мкл плазмы крови прибавляли 400 мкл охлажденной смеси метанола и ацетонитрила (1:1). Образцы встряхивали на шейкере, затем центрифугировали 15 мин при +4 °С и 16000 об/мин. Супернатант переносили в стеклянную вставку для виалы и анализировали. По той же методике готовили два образца контроля качества, полученных путем смешивания равных объемов образцов плазмы крови от пациентов с ПОД и группы контроля.

Таблица 1. Половозрастная характеристика пациентов

Группа	Пол (М/Ж)	Возраст, лет				
		мин.	макс.	средний	медиана	станд. отклон.
Контроль	11/16	65	75	69.6	70	3.0
ПОД	5/7	65	79	69.7	69.5	4.3

Анализ методом ВЭЖХ-МС/МС проводили на хроматографе Shimadzu LC-20AD Prominence, оснащенном градиентным насосом, автодозатором SIL-20AC (Shimadzu, Япония), термостатируемым при 10 °С, и термостатом для колонок CTO-10ASvp с температурой 35 °С. Хроматографическое разделение выполняли на монолитной колонке с сорбентом на основе 1-винил-1,2,4-триазола (Basov et al., 2024). Монолитный материал синтезировали в стеклянных трубках с внутренним диаметром 2 мм, как описано ранее (Patrushev et al., 2020). В качестве подвижной фазы А был взят водный 20 мМ раствор карбоната аммония, доведенный до pH = 9.8 25%-м водным раствором аммиака и содержащий 5 об. % ацетонитрила; подвижной фазой Б служил чистый ацетонитрил. Градиент элюирования был следующим: 0 мин – 0 % Б, 1 мин – 0 % Б, 6 мин – 98 % Б, 16 мин – 98 % Б, после чего колонка уравновешивалась в течение 3 мин. Скорость потока составляла 300 мкл/мин, объем пробы – 2 мкл.

Детекцию метаболитов проводили на масс-спектрометре API 6500 QTRAP (AB SCIEX, США), оснащенном источником электрораспылительной ионизации, работающим в режиме положительной ионизации. Метаболиты детектировали в режиме мониторинга множественных реакций (multiple reaction monitoring, MRM).

Основные масс-спектрометрические параметры были следующими. Напряжение на источнике ионов 5500 В. Температура газа-осушителя 475 °С, давление газа CAD – «высокое», давление газа GS1, GS2 и газа завесы – 33, 33 и 30 psi соответственно. Потенциал декластеризации (DP) составлял 91 В, потенциал входа (EP) – 10 В, а потенциал выхода из ячейки соударений (CXP) – 10 В. Переходы ионов-предшественников и фрагментных ионов, названия метаболитов, время пребывания и соответствующие энергии столкновений представлены в Приложении S1[1]. Управление прибором и сбор информации осуществлялись с помощью программного обеспечения Analyst 1.6.3 (AB SCIEX, Фремингем, США). Хроматограммы обрабатывали в программе MultiQuant 2.1 (AB SCIEX).

Предварительная обработка и статистический анализ данных. Исходные данные были предварительно обработаны для заполнения пропущенных значений со-

[1] Приложения S1–S10 см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx24.xlsx

держания метаболитов в анализируемых пробах следующим образом. В случае, если число проб с пропусками не превышало 5 % от общего числа по 39 пациентам, в качестве значения содержания метаболита бралась медиана, рассчитанная по остальным пробам. Такой подход обусловлен робастностью медианы к выбросам. Статистические различия по содержанию метаболитов в пробах плазмы крови в группе пациентов с ПОД и группе без ПОД оценивали с использованием непараметрического критерия Манна–Уитни.

Реконструкция и анализ генных сетей. Список генов, кодирующих ферменты, участвующие в метаболическом пути Sphingolipid metabolism (ID: hsa00600), был извлечен из базы данных KEGG (https://www.kegg.jp/kegg/pathway.html, Kanehisa, 2002; Kanehisa et al., 2022). Реконструкцию регуляторной генной сети осуществляли с использованием программно-информационной системы ANDSystem (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022). Работу с базой знаний ANDSystem проводили в программном модуле ANDVisio. Анализ перепредставленности биологических процессов (Gene Ontology) и заболеваний, связанных с белками регуляторной генной сети, выполняли с помощью веб-инструмента DAVID (https://david.ncifcrf.gov/tools.jsp, Huang D.W. et al., 2009).

Результаты

Исследование содержания сфинголипидов в плазме крови пациентов с помощью метода ВЭЖХ-МС/МС

Поскольку нарушение метаболизма сфингомиелинов может вносить вклад в развитие делирия, целью нашего анализа было изучение их роли в осложнении ПОД на основе исследования их содержания в плазме крови пациентов после кардиохирургической операции. В частности, проведен сравнительный анализ экспрессии SM в плазме крови пациентов, перенесших операцию на сердце. Метаболиты данного класса, которые статистически значимо различались по содержанию в пробах, взятых в группе пациентов с ПОД, по сравнению с группой пациентов, у которых не развился ПОД, представлены в табл. 2.

Согласно критерию Манна–Уитни, из девяти анализируемых сфингомиелинов четыре (SM(d18:1/22:2 OH), SM(d18:1/24:0), SM(d18:1/24:1) и SM(d18:1/22:2)) показали статистически значимые (p-value < 0.05) различия между исследуемыми группами пациентов. Мы предположили, что нарушение метаболизма сфинголипидов может быть связано с нарушением метаболического пути их биосинтеза. Для проверки данной гипотезы с помощью программно-информационной системы ANDSystem мы провели реконструкцию и анализ генной сети, описывающей регуляцию экспрессии генов, кодирующих ферменты метаболического пути KEGG Sphingolipid metabolism, а также регуляцию транспорта, активности и деградации данных ферментов.

Реконструкция регуляторной генной сети

Для реконструкции регуляторной генной сети из базы данных KEGG был извлечен список генов, кодирующих ферменты, участвующие в метаболизме сфинголипидов Sphingolipid metabolism (hsa00600). Полученный список содержал 43 гена человека (Приложение S2). Реконструкция графа генной сети проводилась в модуле «Мастер запросов» ANDVisio.

Таблица 2. Статистическая значимость различий между группой пациентов с ПОД и контрольной группой по содержанию метаболитов в образцах плазмы крови при сравнении по критерию Манна–Уитни

Метаболит	p-value
SM(d18:1/22:2 OH)	0.0273
SM(d18:1/24:0)	0.0430
SM(d18:1/24:1)	0.0462
SM(d18:1/22:2)	0.0496
SM(d18:1/18:0)	0.0750
SM(d18:1/22:1)	0.1483
SM(d18:1/20:1)	0.3693
SM(d18:1/20:0)	0.5129
SM(d18:1/24:2)	0.5943

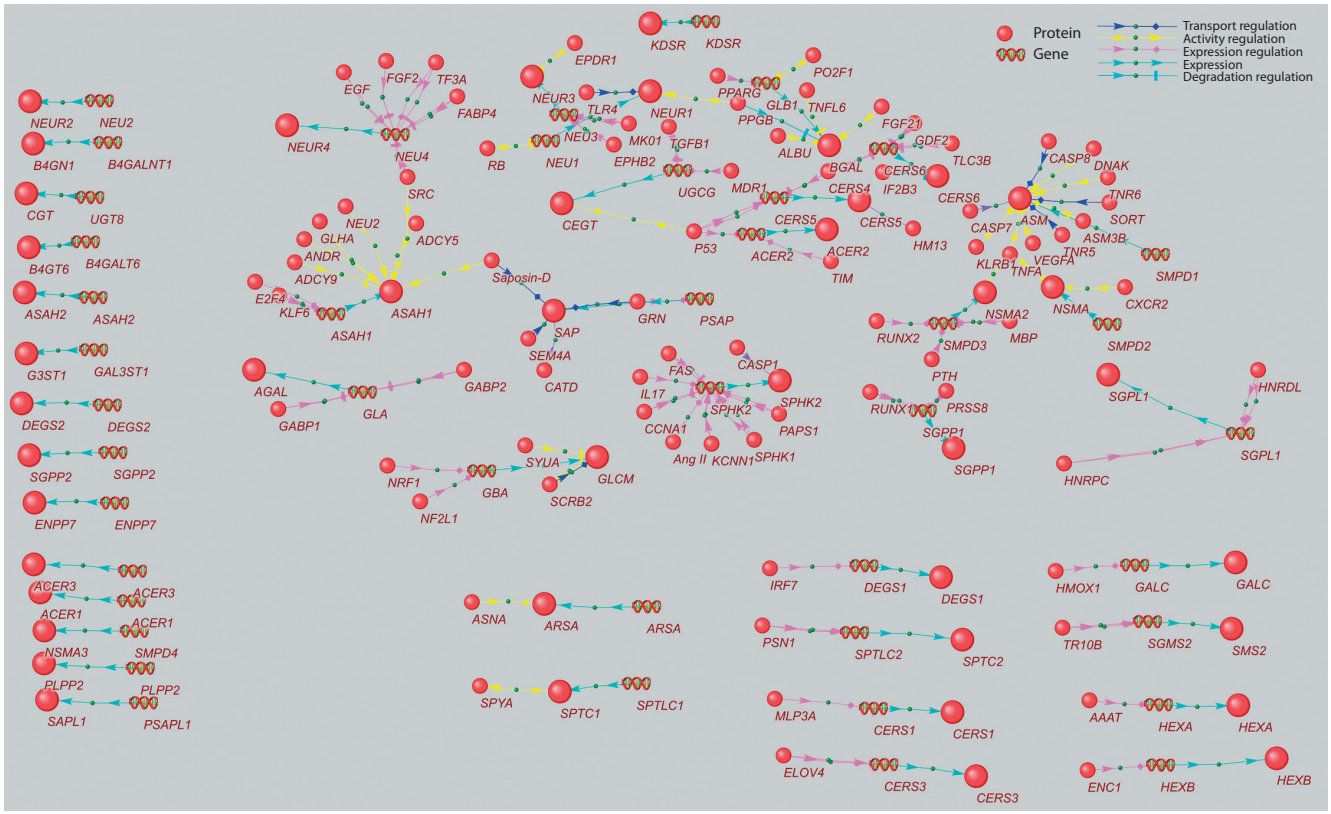
Следует отметить, что в генной сети мы рассматривали только регуляторные связи, направленные от белков-регуляторов к ферментам метаболического пути. Результирующая генная сеть содержала 43 гена человека, 125 белков (43 фермента метаболического пути и 82 регуляторных белка) и 159 взаимодействий между ними (см. рисунок). Различные типы взаимодействий между участниками генной сети были представлены в следующем соотношении: 28 связей, соответствующих типу «регуляция активности», 2 – «регуляция деградации», 4 – «протеолиз», 8 – «регуляция транспорта», 43 – «экспрессия», 74 связи – «регуляция экспрессии».

Чтобы исследовать связи регуляторных белков с патологиями, мы проанализировали перепредставленность заболеваний и биологических процессов Gene Ontology с помощью веб-инструмента DAVID. В качестве входных данных подавался список, состоящих из 82 генов, кодирующих регуляторные белки генной сети. Результаты анализа перепредставленности заболеваний и биологических процессов приведены в Приложениях S3 и S4 соответственно.

Все регуляторные белки, представленные в генной сети (см. рисунок), могут быть разбиты на две группы: 1) регуляторы экспрессии генов и 2) регуляторы активности, стабильности, транспорта и др., которые можно назвать регуляторами функции белков. Для исследования особенностей ассоциированных с ними заболеваний и биологических процессов был проведен анализ перепредставленности отдельно для каждой группы белков (Приложения S5–S8).

Обсуждение

Согласно литературным данным, сфингомиелины (SM) играют важную роль в функционировании нервной системы, и изменение их метаболизма может вносить вклад в развитие делирия путем индукции нейровоспаления, изменения баланса нейромедиаторов и нарушения нейронных связей (Wang, Shen, 2018; Xiao et al., 2023). Про-



Генная сеть регуляции пути метаболизма сфинголипидов.

веденный нами метаболомный анализ с применением ВЭЖХ-МС/МС плазмы крови пациентов, перенесших кардиохирургические операции, позволил выявить четыре из девяти сфингомиелинов, содержание которых статистически значимо отличалось в анализируемых образцах пациентов с ПОД по сравнению с пациентами, у которых ПОД не развился (см. табл. 2).

Для изучения потенциальных механизмов нарушений метаболизма сфинголипидов с помощью ANDSystem была реконструирована генная сеть (см. рисунок), описывающая регуляцию экспрессии генов и функции кодируемых ими ферментов – участников метаболического пути KEGG «Метаболизм сфинголипидов» (Sphingolipid metabolism, hsa: 00600). Анализ сети показал, что в регуляции метаболического пути участвуют 82 регуляторных белка, нарушение функции которых могло оказывать влияние на нарушение метаболизма сфинголипидов. На основе анализа обогащенности списка генов, кодирующих данные белки, генами, ассоциированными с заболеваниями, было определено 168 статистически значимо перепредставленных заболеваний.

Список заболеваний был разделен на пять групп для удобства представления результатов (табл. 3). Наиболее значимым оказалось заболевание из группы патологий сердечно-сосудистой системы, что, вероятно, обусловлено кардиохирургической операцией, которую перенесли пациенты в связи с патологией сердца. Проведенная операция и медицинские процедуры, такие как искусственное кровообращение, также могут объяснять наличие среди выявленных значимых патологий групп «воспаление», «патологии почек» и «оперативное вмешательство» (Stafford-Smith et al., 2008; Squiccimarro et al., 2019). По-видимому, эти патологии связаны с обеими группами пациентов (как с ПОД, так и без него) тем, что каждый из них перенес операцию на сердце.

Особый интерес в контексте развития послеоперационного делирия представляет группа патологий «нервно-психические заболевания». В частности, в работе (Huang H. et al., 2022) рассмотрена роль нейровоспаления в развитии послеоперационного делирия. Авторы выделяют нейровоспаление и нарушение ГЭБ как одни из основных патофизиологических факторов возникновения делирия. Связь нервно-психических патологий с ПОД тоже широко обсуждается в научной литературе. Например, O'Sullivan с коллегами предположили, что связь между делирием и депрессивным расстройством может быть обусловлена общими патофизиологическими механизмами, включающими нарушения стрессовых и воспалительных реакций, моноаминовой и мелатонинергической сигнализации (O'Sullivan et al., 2014). Согласно результатам нашего анализа, на молекулярно-генетическом уровне данные патофизиологические механизмы могут затрагивать генетическую регуляцию пути метаболизма сфинголипидов. Список регуляторных генов из генной сети, ассоциированных с группой «нервно-психические заболевания», приведен в Приложении S9.

Статистический анализ перепредставленности биологических процессов Gene Ontology на основе списка регуляторных генов позволил выявить 67 значимых биологических процессов (БП, см. Приложение S4), которые были

Таблица 3. Статистическая значимость перепредставленности заболеваний на основе анализа списка генов-регуляторов

Группа патологий	Количество патологий в группе	Наиболее значимая патология	FDR	Количество генов
Патологии сердечно-сосудистой системы	23	Гипертония	7.7×10^{-6}	12
Патологии почек	8	Острая почечная недостаточность	7.3×10^{-6}	10
Воспалительные процессы	4	Воспаление	7.3×10^{-6}	11
Оперативное вмешательство	2	Повреждение при реперфузии	1.4×10^{-5}	9
Нервно-психические заболевания	26	Депрессивное расстройство	2.1×10^{-4}	12

Примечание. False Discovery Rate (FDR) и количество генов, ассоциированных с патологией, приведены для наиболее статистически значимой патологии.

Таблица 4. Статистическая значимость перепредставленности биологических процессов на основе анализа списка генов-регуляторов

Группа биологических процессов (БП)	Количество БП в группе	Наиболее значимый БП	FDR	Количество генов
Регуляция апоптоза	7	Положительная регуляция апоптоза	4.1×10^{-7}	14
Ответ на стрессовые факторы	3	Клеточный ответ на механический стимул	4.1×10^{-7}	9
Регуляция клеточных сигнальных путей	9	Положительная регуляция активности MAP-киназ	4.0×10^{-7}	9
Регуляция транскрипции	8	Положительная регуляция транскрипции с промотора RNA-pol. II	3.9×10^{-7}	23
Регуляция пролиферации клеток сердечно-сосудистой системы	7	Положительная регуляция ангиогенеза	6.8×10^{-5}	9
Регуляция клеточной пролиферации	3	Положительная регуляция клеточной пролиферации	0.0014	12
Воспалительные процессы	5	Положительная регуляция воспалительного ответа	0.0046	6

Примечание. False Discovery Rate (FDR) и количество генов, ассоциированных с БП, приведены для наиболее статистически значимого БП.

подразделены на семь групп (табл. 4). Среди значимых оказались фундаментальные регуляторные процессы, включающие регуляцию транскрипции, регуляцию пролиферации, активацию клеточных сигнальных путей и др. Такой результат был ожидаем, поскольку участники генной сети являются регуляторами экспрессии генов и функций кодируемых ими ферментов. Анализируемый нами набор регуляторов оказался обогащен генами, вовлеченными в процесс пролиферации клеток сердечно-сосудистой системы (см. табл. 4 и Приложение S10), что можно объяснить активацией восстановительных процессов после оперативного вмешательства. Помимо указанных, выделим более специфически связанные с делирием БП, такие как воспалительные процессы, ответ на стрессовые факторы и регуляция апоптоза (Steiner, 2011; Vutskits, Xie, 2016).

Заметим, что регуляторные связи в генной сети можно разделить на две группы: регуляция экспрессии генов и регуляция функции (активности, деградации и транспорта) белковых продуктов их экспрессии. По этой причине интересен вопрос, существуют ли характерные особенности, связанные с молекулярными механизмами развития делирия, для регуляторов из этих двух отдельно взятых групп. Для поиска ответа мы проанализировали перепредставленности заболеваний и биологических процессов отдельно для регуляторов экспрессии, а также для регуляторов функции белков (см. Приложения S5–S8).

Неожиданным для нас стало, что среди регуляторов функции белков в десятке наиболее значимых патологий оказались нервно-психические заболевания (например, шизофрения, биполярные расстройства, аутизм), которые, по данным литературы, специфически связаны с делирием (García-Bueno et al., 2016a, b). Интересно, что в литературе обсуждается связь болевого фактора перед операцией с депрессивными симптомами и последующим развитием ПОД (O'Sullivan et al., 2014). При рассмотрении регуляторов экспрессии генов среди значимых патологий преобладала группа патологий сердечно-сосудистой системы, что вполне ожидаемо с учетом анамнеза пациентов. В связи с этим можно предположить особую роль делирия в проявлении патологических механизмов через регуляцию активности белковых продуктов и, в меньшей степени, регуляцию экспрессии генов. Отметим, что в результате анализа перепредставленности БП существенных различий между двумя группами регуляторов не выявлено.

Важной структурной характеристикой графа генных сетей, определяющей особенности их функционирования, является центральность вершин. Один из ее показателей – центральность вершин по степени, которая характеризует отношение количества связей заданной вершины к общему количеству связей в графе и широко применяется в анализе генных сетей. Среди вершин графа, соответствующих ферментам, наибольшим количеством связей (регуляция активности, деградации, транспорта) с регуляторными белками обладал фермент сфингомиелиназа (ASM, см. рисунок). Данный фермент расщепляет сфингомиелины на фосфатидилхолин и церамид, которые выполняют сигнальную функцию. Функция фермента ASM модулировалась десятью регуляторными белками, из которых шесть имели тип связей «регуляция активности»

(ASM3B, Hsp70, KLRB1, TNFA, TNR6, VEGFA), три белка (CASP8, SORT, TNR5) – «регуляция транспорта»; также была представлена одна связь с белком CASP7 с типом «протеолиз». Отметим, что среди регуляторных белков присутствовали каспаза-8 (CASP8) и фактор некроза опухоли альфа (TNFA), которые были ассоциированы с перепредставленными заболеваниями, такими как эпилепсия, депрессия, деменция, и другими нервно-психическими заболеваниями. Согласно литературе, CASP8 осуществляет активацию и транслокацию ASM на поверхность плазматической мембраны. В результате активации ASM происходит расщепление сфингомиелинов и образуется церамид, способствующий повышению активности каспазы-8 и индукции апоптоза (Grassmé et al., 2003). Кроме того, известно, что хирургические вмешательства провоцируют проникновение через ГЭБ провоспалительных факторов, в частности интерлейкинов и TNFA, что способствует нейровоспалению и может быть связано с развитием ПОД (Alam et al., 2018). Согласно реконструированной генной сети, TNFA повышает активность фосфомиелиназы (Corre et al., 2013), а также ассоциирован с перепредставленными нервно-психическими заболеваниями, например депрессией, эпилепсией и др. (см. Приложение S3).

Наибольшим показателем центральности среди вершин графа генной сети, соответствующих генам, обладал ген *SPHK2* (см. рисунок), кодирующий фермент сфингозин киназу 2. В генной сети было представлено семь регуляторов экспрессии данного гена, кодируемых генами *AGT*, *CCNA1*, *FAS*, *IL17A*, *KCNN1*, *SPHK1*, *PAPSS1*. В отличие от вершины, соответствующей белку ASM, среди регуляторов экспрессии *SPHK2* не оказалось таковых, ассоциированных с нервно-психическими заболеваниями. Этот факт еще раз указывает на то, что наиболее важный вклад в нарушение функционирования пути метаболизма сфинголипидов, ассоциированное с послеоперационным делирием, может вносить не регуляция экспрессии генов, кодирующих ферменты метаболического пути, а нарушение транспорта, активности и стабильности продуктов данных генов. Гены, ассоциированные с другими группами заболеваний, были представлены среди регуляторов экспрессии SPHK2 (см. Приложение S5). Например, активность синтазы жирных кислот (FAS) связана с инфарктом миокарда, гипертонией, диабетом 2-го типа и другими заболеваниями (Nosrati-Oskouie et al., 2021).

Заключение

Комплексный подход, заключающийся в метаболомном анализе плазмы крови у пациентов, перенесших кардиохирургические операции, основанный на ВЭЖХ-МС/МС и биоинформатических методах реконструкции генных сетей ANDSystem, позволил выявить потенциальные маркеры класса сфингомиелинов, а также регуляторные гены, нарушение функции которых может лежать в основе механизмов развития послеоперационного делирия. В результате анализа перепредставленности заболеваний обнаружено, что с данными регуляторными белками ассоциированы в первую очередь нервно-психические заболевания, патологии сердца и почек, воспалительные процессы и оперативное вмешательство. Функция регуляторов, ассоциированных с сердечно-сосудистыми заболеваниями, могла быть нарушена у пациентов с ПОД в связи с перенесенной операцией на сердце и медицинскими процедурами, такими как искусственное кровообращение (Gao et al., 2005). В то же время, поскольку операция на сердце была перенесена всеми испытуемыми, можно ожидать, что изменение функции этих регуляторных белков могло в равной степени повлиять на пациентов обеих групп – с послеоперационным делирием и без него. Поэтому можно предположить, что функция группы регуляторов, ассоциированных с нервно-психическими заболеваниями, могла быть специфически нарушена у пациентов с ПОД, что и обусловило снижение содержания сфинголипидов в плазме крови этих пациентов.

Среди вершин графа генной сети наибольшим показателем центральности обладала вершина с 10 регуляторными связями, соответствующая ферменту ASM (фосфомиелиназа). В числе регуляторов активности и транспорта ASM были найдены белки, кодируемые генами *TNFA*, *CASP8*, *TNR5*, *VEGFA*, которые ассоциированы с эпилепсией, депрессией и другими нервно-психическими заболеваниями. Среди вершин, соответствующих генам, наибольшим показателем центральности в графе обладал ген *SPHK2* (сфингозин киназа 2). Его экспрессию регулируют семь белков, кодируемых генами *AGT*, *CCNA1*, *FAS*, *IL17A*, *KCNN1*, *SPHK1*, *PAPSS1*.

Предложенные гипотезы о роли регуляторных генов в развитии послеоперационного делирия могут быть использованы при планировании экспериментов транскриптомного и протеомного анализа для изучения молекулярно-генетических механизмов данного осложнения.

Список литературы / References

Alam A., Hana Z., Jin Z., Suen K.C., Ma D. Surgery, neuroinflammation and cognitive impairment. *EBioMedicine*. 2018;37:547-556. DOI 10.1016/j.ebiom.2018.10.021

Basov N.V., Rogachev A.D., Aleshkova M.A., Gaisler E.V., Sotnikova Y.S., Patrushev Y.V., Tolstikova T.G., Yarovaya O.I., Pokrovsky A.G., Salakhutdinov N.F. Global LC-MS/MS targeted metabolomics using a combination of HILIC and RP LC separation modes on an organic monolithic column based on 1-vinyl-1,2,4-triazole. *Talanta*. 2024;267:125168. DOI 10.1016/j.talanta.2023.125168

Brown C.H. Delirium in the cardiac surgical intensive care unit. *Curr. Opin. Anaesthesiol*. 2014;27(2):117-122. DOI 10.1097/ACO.0000000000000061

Cerejeira J., Firmino H., Vaz-Serra A., Mukaetova-Ladinska E.B. The neuroinflammatory hypothesis of delirium. *Acta Neuropathol*. 2010;119(6):737-775. DOI 10.1007/s00401-010-0674-1

Corre I., Guillonneau M., Paris F. Membrane signaling induced by high doses of ionizing radiation in the endothelial compartment. Relevance in radiation toxicity. *Int. J. Mol. Sci*. 2013;14(11):22678-22696. DOI 10.3390/ijms141122678

Gao L., Taha R., Gauvin D., Othmen L.B., Wang Y., Blaise G. Postoperative cognitive dysfunction after cardiac surgery. *Chest*. 2005;128(5):3664-3670. DOI 10.1378/chest.128.5.3664

García-Bueno B., Gassó P., MacDowell K.S., Callado L.F., Mas S., Bernardo M., Lafuente A., Meana J.J., Leza J.C. Evidence of activation of the Toll-like receptor-4 proinflammatory pathway in patients with schizophrenia. *J. Psychiatry Neurosci.* 2016a;41(3):E46-E55. DOI 10.1503/jpn.150195

García Bueno B., Caso J.R., Madrigal J.L., Leza J.C. Innate immune receptor Toll-like receptor 4 signalling in neuropsychiatric diseases. *Neurosci. Biobehav. Rev.* 2016b;64:134-147. DOI 10.1016/j.neubiorev.2016.02.013

Gottesman R.F., Grega M.A., Bailey M.M., Pham L.D., Zeger S.L., Baumgartner W.A., Selnes O.A., McKhann G.M. Delirium after co-

ronary artery bypass graft surgery and late mortality. *Ann. Neurol.* 2010;67(3):338-344. DOI 10.1002/ana.21899

Grassmé H., Cremesti A., Kolesnick R., Gulbins E. Ceramide-mediated clustering is required for CD95-DISC formation. *Oncogene*. 2003; 22(35):5457-5470. DOI 10.1038/sj.onc.1206540

Guo Y., Li Y., Zhang Y., Fang S., Xu X., Zhao A., Zhang J., Li J.V., Ma D., Jia W., Jiang W. Post-operative delirium associated with metabolic alterations following hemi-arthroplasty in older patients. *Age Ageing*. 2019;49(1):88-95. DOI 10.1093/ageing/afz132

Huang D.W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*. 2009;4(1):44-57. DOI 10.1038/nprot.2008.211

Huang H., Han J., Li Y., Yang Y., Shen J., Fu Q., Chen Y. Early serum metabolism profile of post-operative delirium in elderly patients following cardiac surgery with cardiopulmonary bypass. *Front. Aging Neurosci*. 2022;14:857902. DOI 10.3389/fnagi.2022.857902

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 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 Sys. 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

Kanehisa M. The KEGG Database. In: 'In silico' Simulation of Biological Processes: Novartis Foundation Symposium. Chichester, UK: John Wiley & Sons, 2002;247:91-103. DOI 10.1002/0470857897.ch8

Kanehisa M., Sato Y., Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci*. 2022;31(1): 47-53. DOI 10.1002/pro.4172

Ke C., Pan C.W., Zhang Y., Zhu X., Zhang Y. Metabolomics facilitates the discovery of metabolic biomarkers and pathways for ischemic stroke: a systematic review. *Metabolomics*. 2019;15(12):152. DOI 10.1007/s11306-019-1615-1

Li K., Naviaux J.C., Bright A.T., Wang L., Naviaux R.K. A robust, single-injection method for targeted, broad-spectrum plasma metabolomics. *Metabolomics*. 2017;13(10):122. DOI 10.1007/s11306-017-1264-1

Morimoto Y., Yoshimura M., Utada K., Setoyama K., Matsumoto M., Sakabe T. Prediction of postoperative delirium after abdominal surgery in the elderly. *J. Anesth*. 2009;23(1):51-56. DOI 10.1007/s00540-008-0688-1

Nosrati-Oskouie M., Aghili-Moghaddam N.S., Sathyapalan T., Sahebkar A. Impact of curcumin on fatty acid metabolism. *Phytother. Res*. 2021;35(9):4748-4762. DOI 10.1002/ptr.7105

O'Sullivan R., Inouye S.K., Meagher D. Delirium and depression: inter-relationship and clinical overlap in elderly people. *Lancet Psychiatry*. 2014;1(4):303-311. DOI 10.1016/S2215-0366(14)70281-0

Patrushev Y.V., Sotnikova Y.S., Sidel'nikov V.N. A monolithic column with a sorbent based on 1-vinyl-1,2,4-triazole for hydrophilic HPLC. *Protect. Met. Phys. Chem. Surf*. 2020;56(1):49-53. DOI 10.1134/s2070205119060248

Pisani M.A., Kong S.Y.J., Kasl S.V., Murphy T.E., Araujo K.L.B., Ness P.H.V. Days of delirium are associated with 1-year mortality in an older intensive care unit population. *Am. J. Resp. Crit. Care Med*. 2009;180(11):1092-1097. DOI 10.1164/rccm.200904-0537OC

Rogachev A.D., Alemasov N.A., Ivanisenko V.A., Ivanisenko N.V., Gaisler E.V., Oleshko O.S., Cheresiz S.V., Mishinov S.V., Stupak V.V., Pokrovsky A.G. Correlation of metabolic profiles of plasma and cerebrospinal fluid of high-grade glioma patients. *Metabolites*. 2021;11(3):133. DOI 10.3390/metabo11030133

Squiccimarro E., Labriola C., Malvindi P.G., Margari V., Guida P., Visicchio G., Kounakis G., Favale A., Dambruoso P., Mastrototaro G., Lorusso R., Paparella D. Prevalence and clinical impact of systemic inflammatory reaction after cardiac surgery. *J. Cardiothorac. Vasc. Anesth*. 2019;33(6):1682-1690. DOI 10.1053/j.jvca.2019.01.043

Stafford-Smith M., Patel U.D., Phillips-Bute B.G., Shaw A.D., Swaminathan M. Acute kidney injury and chronic kidney disease after cardiac surgery. *Adv. Chronic Kidney Dis*. 2008;15(3):257-277. DOI 10.1053/j.ackd.2008.04.006

Steiner L.A. Postoperative delirium. Part 1: Pathophysiology and risk factors. *Eur. J. Anaesthesiol*. 2011;28(9):628-636. DOI 10.1097/EJA.0b013e328349b7f5

Tripp B.A., Dillon S.T., Yuan M., Asara J.M., Vasunilashorn S.M., Fong T.G., Metzger E.D., Inouye S.K., Xie Z., Ngo L.H., Marcantonio E.R., Libermann T.A., Otu H.H. Targeted metabolomics analysis of postoperative delirium. *Sci. Rep*. 2021;11(1):1521. DOI 10.1038/s41598-020-80412-z

Vutskits L., Xie Z. Lasting impact of general anaesthesia on the brain: mechanisms and relevance. *Nat. Rev. Neurosci*. 2016;17:705-717. DOI 10.1038/nrn.2016.128

Wang Y., Shen X. Postoperative delirium in the elderly: the potential neuropathogenesis. *Aging Clin. Experim. Res*. 2018;30(11):1287-1295. DOI 10.1007/s40520-018-1008-8

Wimmer-Greinecker G., Matheis G., Brieden M., Dietrich M., Oremek G., Westphal K., Winkelmann B.R., Moritz A. Neuropsychological changes after cardiopulmonary bypass for coronary artery bypass grafting. *Thorac. Cardiovasc. Surg*. 1998;46(4):207-212. DOI 10.1055/s-2007-1010226

Xiao M.Z., Liu C.X., Zhou L.G., Yang Y., Wang Y. Postoperative delirium, neuroinflammation, and influencing factors of postoperative delirium: a review. *Medicine*. 2023;102(8):e32991-e32991. DOI 10.1097/MD.0000000000032991

ORCID ID

V.A. Ivanisenko orcid.org/0000-0002-1859-4631
N.V. Basov orcid.org/0000-0001-6390-5796
A.A. Makarova orcid.org/0009-0005-1844-7921
A.S. Venzel orcid.org/0000-0002-7419-5168
A.D. Rogachev orcid.org/0000-0002-3338-8529
P.S. Demenkov orcid.org/0000-0001-9433-8341
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
M.A. Kleshchev orcid.org/0000-0002-7537-2525
G.B. Moroz orcid.org/0000-0002-0154-4662
Y.S. Sotnikova orcid.org/0000-0002-0545-703X
Y.V. Patrushev orcid.org/0000-0002-2078-5488
V.V. Lomivorotov orcid.org/0000-0001-8591-6461
N.A. Kolchanov orcid.org/0000-0001-6800-8787
A.G. Pokrovsky orcid.org/0000-0001-5982-8580

Благодарности. Работа выполнена при поддержке Российского научного фонда, грант № 22-23-01068. Авторы также выражают благодарность Центру коллективного пользования «Биоинформатика» за вычислительные ресурсы и их программное обеспечение, созданное в рамках бюджетного проекта FWNR-2022-0020.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 21.07.2023. После доработки 12.08.2023. Принята к публикации 24.08.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):776-783
DOI 10.18699/VJGB-23-90

Перевод на английский язык https://vavilov.elpub.ru/jour

Молекулярно-генетические пути регуляции вирусом гепатита С экспрессии клеточных факторов PREB и PLA2G4C, играющих важную роль для репликации вируса

Е.Л. Мищенко[1, 2], А.А. Макарова[1], Е.А. Антропова[1], А.С. Вензель[1, 2], Т.В. Иванисенко[1, 2], П.С. Деменков[1, 2, 3], В.А. Иванисенко[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
salix@bionet.nsc.ru

Аннотация. В репликации генома вируса гепатита С (ВГС) участвуют как вирусные, так и хозяйские белки. Терапевтические подходы, основанные на подавлении активности неструктурных вирусных белков NS3, NS5A, NS5B, проходят клинические испытания разных уровней. Однако быстрые мутационные процессы вирусного генома и приобретение лекарственной устойчивости остаются одними из главных препятствий в борьбе с ВГС. Идентификация и исследование клеточных факторов, участвующих в репликации РНК ВГС, а также регуляция вирусом их экспрессии важны для понимания механизмов репликации вируса и разработки эффективных подходов противовирусной терапии. Известно, что белок PREB, связывающий регуляторный элемент пролактина, и цитозольная фосфолипаза А2 гамма (PLA2G4C) играют важную роль в формировании платформ репликации РНК ВГС, а также в функционировании вирусной репликазы. Экспрессия генов *PREB* и *PLA2G4C* значительно увеличена в присутствии ВГС, но механизмы ее регуляции вирусными белками до сих пор не изучены. В данной работе с применением технологии текст-майнинга, реализованной в программно-информационной системе ANDSystem, реконструированы генные сети регуляции экспрессии генов человека *PREB* и *PLA2G4C* белками ВГС. На основании анализа генных сетей мы выдвинули гипотезы о регуляторных эффектах белков ВГС на функции хозяйских факторов в результате белок-белковых взаимодействий. Среди вирусных белков наибольшее количество регуляторных связей выявлено у вирусной протеазы NS3. Предположительно NS3 в результате белок-белкового взаимодействия подавляет активность транскрипционного фактора NOTCH1, что обусловливает активацию экспрессии *PREB* и *PLA2G4C*. Анализ генных сетей и данных о дифференциальной экспрессии генов в присутствии ВГС позволил нам также выдвинуть гипотезы о регуляции вирусом экспрессии транскрипционных факторов, сайты связывания которых находятся в районах генов *PREB* и *PLA2G4C*, и действии этих транскрипционных факторов на регуляцию транскрипции *PREB* и *PLA2G4C*. Полученные результаты могут быть использованы при планировании исследований по изучению молекулярно-генетических механизмов взаимодействия вирус–хозяин и поиска потенциальных мишеней для разработки лекарств против ВГС.
Ключевые слова: вирус гепатита С; репликация генома ВГС; репликаза ВГС; хозяйские факторы; генные сети; фосфолипаза PLA2G4C; белок PREB.

Для цитирования: Мищенко Е.Л., Макарова А.А., Антропова Е.А., Вензель А.С., Иванисенко Т.В., Деменков П.С., Иванисенко В.А. Молекулярно-генетические пути регуляции вирусом гепатита С экспрессии клеточных факторов PREB и PLA2G4C, играющих важную роль для репликации вируса. *Вавиловский журнал генетики и селекции*. 2023;27(7):776-783. DOI 10.18699/VJGB-23-90

Molecular-genetic pathways of hepatitis C virus regulation of the expression of cellular factors PREB and PLA2G4C, which play an important role in virus replication

E.L. Mishchenko[1, 2], A.A. Makarova[1], E.A. Antropova[1], A.S. Venzel[1, 2], T.V. Ivanisenko[1, 2], P.S. Demenkov[1, 2, 3], V.A. Ivanisenko[1, 2, 3]

[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
[3] Novosibirsk State University, Novosibirsk, Russia
salix@bionet.nsc.ru

Abstract. The participants of Hepatitis C virus (HCV) replication are both viral and host proteins. Therapeutic approaches based on activity inhibition of viral non-structural proteins NS3, NS5A, and NS5B are undergoing clinical trials. However, rapid mutation processes in the viral genome and acquisition of drug resistance to the existing drugs remain the main obstacles to fighting HCV. Identifying the host factors, exploring their role in HCV RNA replication,

and studying viral effects on their expression is essential for understanding the mechanisms of viral replication and developing novel, effective curative approaches. It is known that the host factors *PREB* (prolactin regulatory element binding) and *PLA2G4C* (cytosolic phospholipase A2 gamma) are important for the functioning of the viral replicase complex and the formation of the platforms of HCV genome replication. The expression of *PREB* and *PLA2G4C* was significantly elevated in the presence of the HCV genome. However, the mechanisms of its regulation by HCV remain unknown. In this paper, using a text-mining technology provided by ANDSystem, we reconstructed and analyzed gene networks describing regulatory effects on the expression of *PREB* and *PLA2G4C* by HCV proteins. On the basis of the gene network analysis performed, we put forward hypotheses about the modulation of the host factors functions resulting from protein-protein interaction with HCV proteins. Among the viral proteins, NS3 showed the greatest number of regulatory linkages. We assumed that NS3 could inhibit the function of host transcription factor (TF) NOTCH1 by protein-protein interaction, leading to upregulation of *PREB* and *PLA2G4C*. Analysis of the gene networks and data on differential gene expression in HCV-infected cells allowed us to hypothesize further how HCV could regulate the expression of TFs, the binding sites of which are localized within *PREB* and *PLA2G4C* gene regions. The results obtained can be used for planning studies of the molecular-genetic mechanisms of viral-host interaction and searching for potential targets for anti-HCV therapy.

Key words: hepatitis C virus; HCV gene replication; replicase HCV; host factors; gene networks; phospholipase PLA2G4C; PREB protein.

For citation: Mishchenko E.L., Makarova A.A., Antropova E.A., Venzel A.S., Ivanisenko T.V., Demenkov P.S., Ivanisenko V.A. Molecular-genetic pathways of hepatitis C virus regulation of the expression of cellular factors PREB and PLA2G4C, which play an important role in virus replication. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):776-783. DOI 10.18699/VJGB-23-90

Введение

Вирус гепатита С (ВГС) вызывает опасное заболевание печени, которое, начинаясь бессимптомно, переходит в хроническую форму и может привести к циррозу и гепатоцеллюлярной карциноме (Yamane et al., 2013). Геном ВГС представлен плюс-цепью РНК (~9600 нуклеотидов), кодирующей структурные (Core, E1, E2) и неструктурные (p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B) белки, а также содержит 5′- и 3′-нетранслируемые районы (UTR), которые необходимы для трансляции вирусного полипротеина и репликации вирусного генома (Bartenschlager et al., 2013). Структурные гликопротеины E1 и E2 локализованы на двуслойной липидной оболочке вируса, окружающей нуклеокапсид, состоящий из множества копий корового (Core) белка и РНК-генома. Белок p7 имеет свойства мембранного катионного канала, белки NS2 и NS3/NS4A – протеазы, осуществляющие процессинг вирусного полипротеина, NS3 обладает также хеликазной активностью, NS4B и NS5A способны модифицировать мембраны эндоплазматического ретикулума (ЭР) с образованием везикулярных мембранных структур – платформ репликации генома ВГС, NS5B является РНК-зависимой РНК полимеразой. Комплекс неструктурных белков NS3–NS5B, включающий также хозяйские факторы, выполняет роль вирусной репликазы в хозяйской клетке (Moradpour et al., 2007). Геном вируса высокогетерогенен из-за высокой частоты ошибок функционирования РНК-зависимой РНК полимеразы NS5B. Это свойство NS5B считается основной причиной быстрого приобретения вирусом лекарственной устойчивости (Powdrill et al., 2011).

В настоящее время большое внимание исследователей направлено на идентификацию и изучение свойств клеточных факторов, участвующих в модификации мембран ЭР с образованием кластеров везикул, в которых реплицируется РНК-геном ВГС, а также входящих в состав вирусной репликазы. Так, установлено, что рецептор активированной С киназы 1 (RACK1) ассоциирует с NS5A и комплексом инициации формирования аутофагосом ATG14L-Beclin1-Vps34-Vps15 и стимулирует формирование мембранных везикулярных структур (Lee et al., 2019). Белок ранних эндосом (EE) Rab5, регулирующий эндоцитоз и слияние EE, а также белок поздних эндосом (LE) Rab7, усиливающий транспорт LE к лизосомам, ассоциированы с NS4B и вовлечены в биогенез этих мембранных структур (Manna et al., 2010). Малая ГТФаза Rab18·GTP на мембранах липидных капель (ЛК) взаимодействует с вирусным белком NS5A на мембране ЭР. Ассоциация мембран ЛК и ЭР в результате прямого взаимодействия Rab18·GTP и NS5A приводит к локализации репликазных комплексов ВГС вблизи ЛК и стимуляции репликации РНК ВГС (Salloum et al., 2013).

Важную роль в формировании мембранных везикулярных структур и репликазных комплексов играет фосфатидилинозитол 4-киназа IIIα (PI4KIIIα). Через белок-белковое взаимодействие NS5A стимулирует активность PI4KIIIα с образованием фосфатидилинозитол-4-фосфата (PI4P), который рекрутирует и координирует на мембране вирусные и хозяйские белки, содержащие аффинные к PI4P липид-связывающие домены (Berger et al., 2011; Reiss et al., 2011). Более того, ВГС способен регулировать экспрессию клеточных факторов, играющих важную роль для репликации вируса. Так, у цитозольной фосфолипазы А2 гамма (PLA2G4C), гидролизующей фосфоглицериды мембран с образованием свободной жирной кислоты и лизофосфатида и оказывающей прямой эффект на структуру мембран, их форму, слияние и взаимодействие с белками (Brown et al., 2003), экспрессия в присутствии РНК ВГС на уровне как РНК, так и белка увеличена в несколько раз (Xu et al., 2012).

Экспрессия гена *PREB*, связывающего регуляторный элемент пролактина, также значительно увеличена в присутствии ВГС (Kong et al., 2016). Белок PREB функционирует в качестве регуляторного фактора отпочковывания COPII везикул от мембран ЭР (LaPointe et al., 2004), ассоциирует с NS4B, вовлечен в формирование мембранных везикулярных структур и локализован в активном репликазном комплексе ВГС через взаимодействие с NS4B (Kong et al., 2016). Несмотря на накопленные факты о

повышении экспрессии *PREB* и *PLA2G4C* в присутствии ВГС, молекулярные механизмы регуляции экспрессии этих хозяйских факторов слабо изучены.

Технология текст-майнинга – полезный инструмент для исследования молекулярно-генетических взаимодействий. Ранее нами была разработана программно-информационная система ANDSystem (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022), реализующая полный цикл инженерии знаний, который включает автоматическую экстракцию информации из научных публикаций и фактографических баз данных, интеграцию и представление информации в виде семантических сетей в базе знаний, а также предоставление пользовательского доступа к базе знаний для реконструкции и анализа генных сетей. ANDSystem применялась для решения широкого круга задач, включающих анализ интерактома белков вируса гепатита С с белками человека, интерпретацию результатов метаболомного анализа, задачи приоритизации генов, поиск новых потенциальных мишеней для действия лекарств и др. В частности, анализ белок-белковых взаимодействий белков ВГС и человека позволил реконструировать потенциальные пути регуляции внешнего пути апоптоза вирусными белками (Saik et al., 2016), а также изучить особенности регуляции белками ВГС генов, подверженных аберрантному метилированию при гепатоцеллюлярной карциноме (Antropova et al., 2022). На основе данных метаболомного анализа плазмы крови пациентов с Covid-19 были реконструированы регуляторные пути, описывающие контроль метаболических путей человека белками SARS-Cov-2, и показано, что ряд неструктурных вирусных белков оказывал наибольшее регуляторное воздействие (Ivanisenko V.A. et al., 2022). С помощью реконструкции и анализа генных сетей предложены новые методы приоритизации генов, которые были применены для поиска генов-кандидатов, ассоциированных с лимфедемой, а также с большим депрессивным расстройством (Yankina et al., 2018; Saik et al., 2019). С использованием ANDSystem были предложены новые потенциальные фармакологические мишени для терапии коморбидного состояния астмы и гипертонии (Saik et al., 2018a, b).

В нашей работе с применением программно-информационной системы ANDSystem реконструированы и проанализированы пути регуляции белками ВГС экспрессии генов клеточных факторов *PLA2G4C* и *PREB*, играющих важную роль в формировании мембранных везикулярных структур – платформ репликации вирусной РНК, и в функционировании вирусной репликазы. С помощью компьютерного анализа были найдены 28 транскрипционных факторов (ТФ) человека, находящихся под контролем ВГС, которые могут участвовать в регуляции экспрессии *PLA2G4C* и *PREB*. Оказалось, что из этих ТФ 16 белков участвуют в регуляции *PLA2G4C*, 23 – в регуляции *PREB*, а 11 являются общими. На основе анализа генных сетей и данных о дифференциальной экспрессии генов выдвинуты гипотезы о регуляторных эффектах вирусных белков на функции ТФ, с которыми они образуют комплексы в результате белок-белковых взаимодействий, а также регуляторные эффекты этих ТФ на экспрессию *PLA2G4C* и *PREB*.

Материалы и методы

Получение списка дифференциально экспрессирующихся генов (ДЭГ) человеческих белков в присутствии белков ВГС. С использованием результатов РНК-секвенирования, доступных на ресурсе NCBI GEO (http://www.ncbi.nlm.nih.gov/geo) (Edgar et al., 2002), по идентификатору GSE66842 был получен список человеческих генов, дифференциально экспрессирующихся в гепатоцитах линии Huh7.5.1 в условиях инфекции ВГС. Анализ данных РНК-секвенирования проведен с помощью инструмента GEO2R, который позволяет получить результаты статистической обработки и визуализацию данных о дифференциальной экспрессии генов в экспериментальных условиях. Нами были взяты статистически значимые ДЭГ в контрольной точке «10 дней после заражения ВГС» (GSE66842). В работе использованы также результаты транскриптомного анализа дифференциальной экспрессии генов в гепатоцитах Huh.7.5 в контрольной точке «72 часа после заражения ВГС» (Papic et al., 2012). Для реконструкции генных сетей эти результаты объединены в конечный список ДЭГ.

Идентификация транскрипционных факторов. Транскрипционные факторы, сайты связывания которых локализованы в генах *PREB* и *PLA2G4C*, а также во фланкирующих районах этих генов в диапазоне ±2000 п. н., были извлечены из базы данных GTRD (http://gtrd20-06.biouml.org/) (Yevshin et al., 2017; Kolmykov et al., 2021), которая интегрирует исследования организации геномов. Для построения генных сетей были отобраны те гены ТФ, которые являются дифференциально экспрессирующимися в условиях инфекции вируса гепатита С.

Реконструкция и анализ молекулярно-генетических путей регуляции экспрессии генов *PREB* и *PLA2G4C* белками ВГС с помощью ANDSystem. Молекулярно-генетические пути регуляции экспрессии хозяйских факторов PREB и PLA2G4C белками ВГС были реконструированы с помощью системы ANDSystem и ее графического пользовательского интерфейса ANDVisio. Программа ANDVisio обращается к базе знаний ANDSystem, содержащей более 40 млн фактов о межмолекулярных взаимосвязях, включающих белок-белковые взаимодействия, регуляцию экспрессии генов, регуляцию активности, деградации и транспорта белков.

Построение регуляторных молекулярно-генетических путей, описывающих взаимодействия между белками ВГС и человеческими белками и генами, осуществляли в модуле «Мастер путей» программы ANDVisio. Взаимосвязи между участниками данных путей, включающие белок-белковые взаимодействия и регуляцию экспрессии генов, расположены согласно схеме (рис. 1).

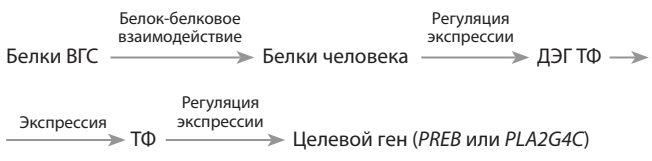


Рис. 1. Схема построения регуляторных молекулярно-генетических путей модуляции экспрессии генов хозяйских факторов белками ВГС.

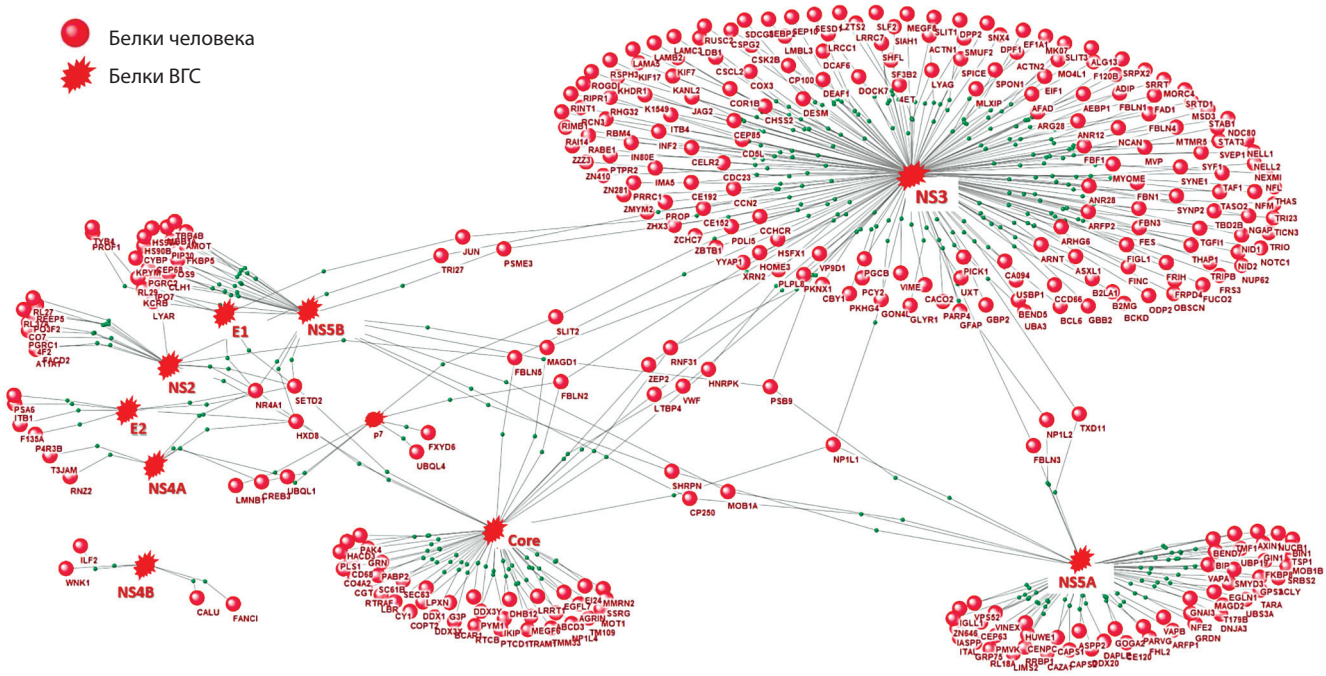


Рис. 2. Граф взаимодействий человеческих белков и белков ВГС, реконструированный с помощью программно-информационной системы ANDSystem.

Белок-белковые взаимодействия обозначены черными линиями.

Результаты и обсуждение

Реконструкция интерактома человеческих белков и белков ВГС

С применением программно-информационной системы ANDSystem был реконструирован интерактом 10 белков ВГС с 333 белками человека (рис. 2). Оказалось, что 195 человеческих белков взаимодействуют с NS3, 59 – с NS5A, 50 – с Core, 26 – с NS5B, 15 – с NS2, 7 – с E2 и p7, 6 – с NS4A, 5 – с E1, 4 белка – с NS4B. Генная сеть иллюстрирует, что лишь небольшое количество человеческих белков взаимодействует более чем с одним белком ВГС. Среди них оказались транскрипционные факторы, потенциально регулирующие экспрессию целевых генов *PREB* и *PLA2G4C*.

Реконструкция молекулярно-генетических путей регуляции экспрессии генов *PREB* и *PLA2G4C* белками ВГС

Опубликованные научные результаты свидетельствуют о том, что в присутствии белков ВГС многократно усиливается экспрессия клеточных факторов PLA2G4C (Xu et al., 2012) и PREB (Kong et al., 2016). Эти хозяйские факторы играют важную роль в репликации ВГС и вовлечены как в формирование мембранных везикулярных структур – компартментов репликации вирусной РНК, так и в функционирование репликазного комплекса ВГС (Xu et al., 2012; Kong et al., 2016). Однако молекулярно-генетические механизмы повышения экспрессии *PREB* и *PLA2G4C* в условиях инфекции ВГС до сих пор не изучены. С помощью информации о дифференциальной экспрессии генов выявлены ТФ, регулируемые вирусными белками. Следует отметить, что в нашей работе мы не рассматривали ТФ, экспрессия которых не изменялась в условиях заражения ВГС. Из базы данных GTRD были извлечены списки, содержащие 432 и 693 ТФ, сайты связывания которых находятся в районах генов *PREB* и *PLA2G4C* соответственно. Среди множества транскрипционных факторов были отобраны 92 ТФ, гены которых дифференциально экспрессируются в присутствии белков ВГС (69 и 63 ТФ для *PREB* и *PLA2G4C* соответственно, 40 ТФ – общие для обоих целевых генов).

С помощью ANDSystem были реконструированы и проанализированы молекулярно-генетические пути регуляции экспрессии *PREB* и *PLA2G4C* белками ВГС (рис. 3 и 4). В составе регуляторных путей, первым звеном которых являлись белки ВГС, а конечным – гены *PREB* и *PLA2G4C*, оказались 28 из 92 ТФ, что свидетельствует о регуляции этих ТФ вирусными белками.

Генная сеть на рис. 3 иллюстрирует регуляторные молекулярно-генетические пути экспрессии *PREB* белками ВГС. Эти пути включают 24 белка, представленных в звене 2, 23 ТФ – участника звена 4 и кодирующих их генов звена 3. Как следует из графа генной сети, только 23 из 69 ТФ вошли в состав регуляторных путей, что может свидетельствовать о том, что именно эти ТФ регулируют транскрипцию гена *PREB* в условиях инфекции вируса гепатита С.

Приведенная на рис. 4 генная сеть иллюстрирует пути регуляции экспрессии *PLA2G4C* белками ВГС. В базе данных GTRD в регуляторных районах гена *PLA2G4C* найдены сайты связывания 63 ТФ, являющихся ДЭГ. Только 16 из 63 ТФ вошли в состав регуляторных путей. Это может свидетельствовать в пользу того, что именно эти ТФ предположительно регулируют транскрипцию гена *PLA2G4C* в условиях инфекции ВГС. Ранее было показано, что белок NS3 вируса гепатита С стимулирует

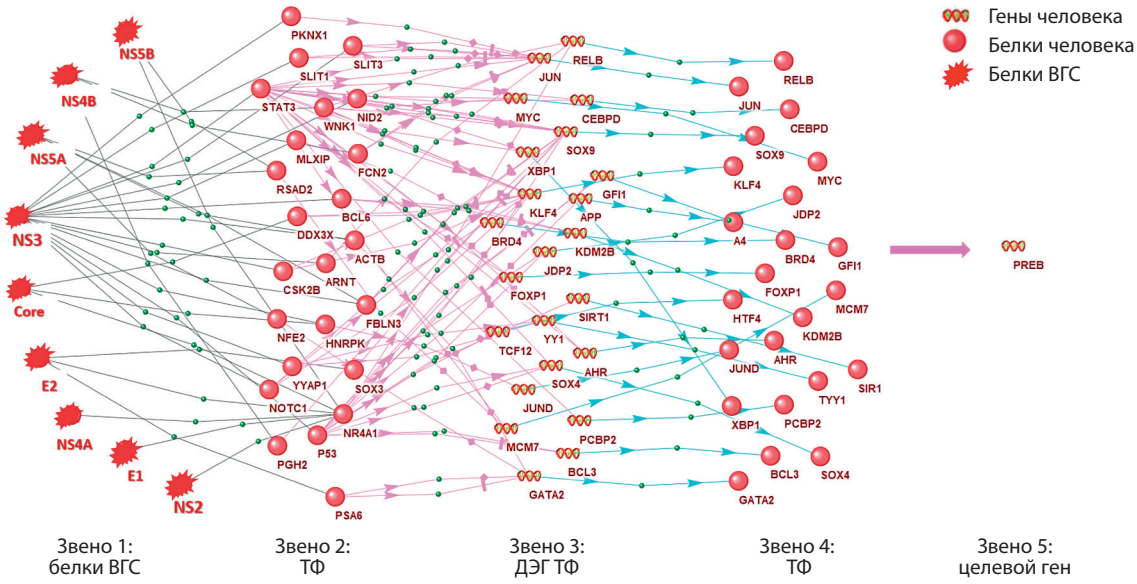


Рис. 3. Генная сеть молекулярно-генетических путей регуляции экспрессии гена *PREB* в условиях инфекции ВГС.

Здесь и на рис. 4: черные линии – белок-белковые взаимодействия; розовые стрелки – регуляция экспрессии; голубые стрелки – экспрессия.

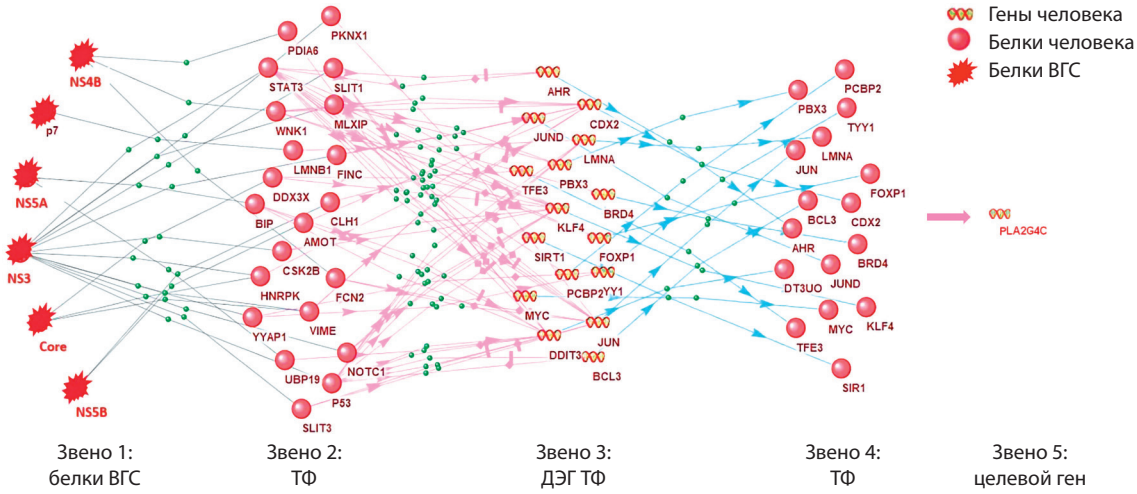


Рис. 4. Генная сеть молекулярно-генетических путей регуляции экспрессии гена *PLA2G4C* в условиях инфекции ВГС.

активность ТФ STAT3 (Machida et al., 2006). При этом STAT3 значительно усиливает транскрипцию гена *MYC* (Kiuchi et al., 1999; Papic et al., 2012). Более того, согласно (Xiong et al., 2017), изменение экспрессии MYC усиливало экспрессию *PLA2G4C*, что согласуется с выявленным нами путем регуляции. Аналогично положительная регуляция экспрессии *XBP1* со стороны STAT3 (Diehl et al., 2008) и повышенная экспрессия *XBP1* (Papic et al., 2012) в присутствии ВГС могут обусловливать активирующий эффект XBP1 на транскрипцию *PLA2G4C*.

Применение ANDSystem позволило нам выдвинуть гипотезы о регуляции белками ВГС экспрессии ТФ, взаимодействующих с сайтами регуляторных районов генов *PREB* и *PLA2G4C* (см. рис. 3 и 4). Следует отметить, что 11 ТФ были одновременно представлены в числе регуляторов как *PREB*, так и *PLA2G4C*. На основании данных о дифференциальной экспрессии генов и характере связей регуляторных молекулярно-генетических путей можно предположить, какой эффект эти ТФ (звено 4) оказывают на транскрипцию *PREB* и *PLA2G4C* (см. таблицу). Например, повышенная экспрессия гена ТФ звена 3 и положительная регуляция со стороны ТФ звена 2 могут обусловливать активацию транскрипции *PREB* и *PLA2G4C*. В частности, из путей регуляции следует, что ТФ CEBPD положительно регулирует экспрессию *PREB*, так как экспрессия *CEBPD* положительно регулируется STAT3 (звено 2) и повышена в присутствии ВГС (Papic et al., 2012). В свою очередь, сниженная в присутствии ВГС экспрессия ТФ из звена 4 и отрицательный знак регуляции экспрессии между участниками звеньев 2 и 3 объясняют подавляющий эффект ТФ на транскрипцию *PREB* и *PLA2G4C*.

Ожидаемый эффект транскрипционных факторов звена 4 на экспрессию *PREB* и *PLA2G4C*

ТФ	Предполагаемый эффект*		ТФ	Предполагаемый эффект		ТФ	Предполагаемый эффект	
	PREB	*PLA2G4C*		*PREB*	*PLA2G4C*		*PREB*	*PLA2G4C*
AHR	↑	↑	JDP2	↑	–	RELB	↑	–
APP	↑	–	JUN	↑	↑	SIRT1	↑	↑
BCL3	↑	↑	JUND	↑	↑	SOX4	↑	–
BRD4	↑	↑	KDM2B	↑	–	SOX9	↓	–
CDX2	–	↑	KLF4	↑	↑	TCF12	↑	–
CEBPD	↑	–	LMNA	–	↑	TFE3	–	↑
DDIT3	↑	↑	MCM7	↑	–	XBP1	↑	–
FOXP1	↓	↓	MYC	↑	↑	YY1	↑	↑
GATA2	↑	–	PBX3	–	↑			
GFI1	↑	–	PCBP2	↑	↑			

* «↑» – положительная регуляция, «↓» – отрицательная регуляция, «–» – нет регуляции.

Из публикаций следует, что белок Core ВГС увеличивает экспрессию *NR4A1* (Tan, Li, 2015), при этом транскрипционный фактор NR4A1 ингибирует экспрессию гена *SOX9* (Hu et al., 2014). В реконструированных нами регуляторных путях NR4A1 является транскрипционным фактором звена 2, связывается с шестью белками ВГС (Core, E1, E2, NS2, NS4A, NS5B) и оказывает отрицательный эффект на *SOX9*. Таким образом, транскрипционный фактор SOX9, ингибируемый на уровне РНК в условиях инфекции ВГС, предположительно снижает экспрессию гена *PREB*. Гипотезы, предложенные нами на основе анализа генных сетей, в дальнейшем следует экспериментально подтвердить.

Анализ реконструированных генных сетей позволил также выдвинуть гипотезы о том, какой эффект могут оказывать вирусные белки на функцию ТФ, с которыми они образуют комплексы в результате белок-белковых взаимодействий. Эти гипотезы строились на основе структуры регуляторных молекулярно-генетических путей и данных о дифференциальной экспрессии генов аналогично гипотезам о регуляции *PREB* и *PLA2G4C* ТФ. Вирусный белок оказывает отрицательный эффект на функцию белка из звена 2 регуляторного пути в результате физического взаимодействия с ним в следующих случаях: 1) участник звена 2 связан с участником из звена 3 по типу положительной регуляции экспрессии, и экспрессия участника звена 3 снижена в присутствии ВГС; 2) участник звена 2 связан с участником из звена 3 по типу отрицательной регуляции экспрессии, и экспрессия участника звена 3 повышена в присутствии ВГС. Вирусный белок оказывает положительный эффект на функцию белка из звена 2 в следующих случаях: 1) участник звена 2 связан с участником звена 3 по типу положительной регуляции экспрессии, и экспрессия участника звена 3 повышена в присутствии ВГС; 2) участник звена 2 связан с участником из звена 3 по типу отрицательной регуляции экспрессии, и экспрессия участника звена 3 снижена в присутствии ВГС.

Согласно реконструированным регуляторным молекулярно-генетическим путям, среди белков ВГС наибольшее число регуляторных связей выявлено у вирусной протеазы NS3. Одним из белков, непосредственно взаимодействующих с NS3, является ТФ NOTCH1. Опубликовано множество научных исследований этого ТФ, но информации об эффекте NS3 на функцию NOTCH1 в результате белок-белковых взаимодействий мы не обнаружили. Из анализа регуляторных путей и данных дифференциальной экспрессии генов мы предположили, что NS3 подавляет активность NOTCH1 в результате белок-белкового взаимодействия. Ранее было показано, что NOTCH1 активирует транскрипцию *SOX9* (Zong et al., 2009) и подавляет *KLF4* (Xue et al., 2016), что привело бы к отрицательному эффекту на транскрипцию *PREB* и *PLA2G4C*. Однако реальное изменение экспрессии целевых генов, а также их ТФ SOX9 и KLF4 согласуется с гипотезой о подавлении активности NOTCH1 вирусным белком NS3.

Заключение

С помощью программно-информационной системы ANDSystem реконструированы и проанализированы молекулярно-генетические пути регуляции экспрессии генов *PLA2G4C* и *PREB* белками вируса гепатита С. Белковые продукты этих генов важны для репликации ВГС, так как они участвуют в модификации мембран с образованием кластеров мембранных везикул, которые являются компартментами репликации генома ВГС, а также вовлечены в состав и функционирование репликазы ВГС. Теоретические данные, полученные в нашей работе, могут быть полезны для планирования исследований по изучению механизмов, посредством которых ВГС использует белки человека для репликации своего генома, а также для поиска потенциальных мишеней противовирусной терапии.

Список литературы / References

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Lavrik I.N., Ivanisenko V.A. Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. DOI 10.18699/VJGB-22-89

Bartenschlager R., Lohmann V., Penin F. The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection. *Nat. Rev. Microbiol*. 2013;11(7):482-496. DOI 10.1038/nrmicro3046

Berger K.L., Kelly S.M., Jordan T.X., Tartell M.A., Randall G. Hepatitis C virus stimulates the phosphatidylinositol 4-kinase III alpha-dependent phosphatidylinositol 4-phosphate production that is essential for its replication. *J. Virol.* 2011;85(17):8870-8883. DOI 10.1128/JVI.00059-11

Brown W.J., Chambers K., Doody A. Phospholipase A2 (PLA2) enzymes in membrane trafficking: mediators of membrane shape and function. *Traffic*. 2003;4(4):214-221. DOI 10.1034/j.1600-0854.2003.00078.x

Diehl S.A., Schmidlin H., Nagasawa M., van Haren S.D., Kwakkenbos M.J., Yasuda E., Beaumont T., Scheeren F.A., Spits H. STAT3-mediated up-regulation of BLIMP1 is coordinated with BCL6 down-regulation to control human plasma cell differentiation. *J. Immunol.* 2008;180(7):4805-4815. DOI 10.4049/jimmunol.180.7.4805

Edgar R., Domrachev M., Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. DOI 10.1093/nar/30.1.207

Hu Y.W., Zhang P., Yang J.Y., Huang J.L., Ma X., Li S.F., Zhao J.Y., Hu Y.R., Wang Y.C., Gao J.J., Sha Y.H., Zheng L., Wang Q. Nur77 decreases atherosclerosis progression in apoE$^{-/-}$ mice fed a high-fat/high-cholesterol diet. *PLoS One*. 2014;9(1):e87313. DOI 10.1371/journal.pone.0087313

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 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

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

Kiuchi N., Nakajima K., Ichiba M., Fukada T., Narimatsu M., Mizuno K., Hibi M., Hirano T. STAT3 is required for the gp130-mediated full activation of the c-*myc* gene. *J. Exp. Med*. 1999;189(1):63-73. DOI 10.1084/jem.189.1.63

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

Kong L., Fujimoto A., Nakamura M., Aoyagi H., Matsuda M., Watashi K., Suzuki R., Arita M., Yamagoe S., Dohmae N., Suzuki T., Sakamaki Y., Ichinose S., Suzuki T., Wakita T., Aizaki H. Prolactin regulatory element binding protein is involved in hepatitis C virus replication by interaction with NS4B. *J. Virol*. 2016;90(6):3093-3111. DOI 10.1128/JVI.01540-15

LaPointe P., Gurkan C., Balch W.E. Mise en place – this bud's for the Golgi. *Mol. Cell.* 2004;14(4):413-414. DOI 10.1016/s1097-2765(04)00267-9

Lee J.S., Tabata K., Twu W.-I., Rahman M.S., Kim H.S., Yu J.B., Jee M.H., Bartenschlager R., Jang S.K. RACK1 mediates rewiring of intracellular networks induced by hepatitis C virus infection. *PLoS Pathog*. 2019;15(9):e1008021. DOI 10.1371/journal.ppat.1008021

Machida K., Cheng K.T., Lai C.K., Jeng K.S., Sung V.M., Lai M.M. Hepatitis C virus triggers mitochondrial permeability transition with production of reactive oxygen species, leading to DNA damage and STAT3 activation. *J. Virol*. 2006;80(14):7199-7207. DOI 10.1128/jvi.00321-06

Manna D., Aligo J., Xu C., Park W.S., Koc H., Heo W.D., Konan K.V. Endocytic Rab proteins are required for hepatitis C virus replication complex formation. *Virology*. 2010;398(1):21-37. DOI 10.1016/j.virol.2009.11.034

Moradpour D., Penin F., Rice C.M. Replication of hepatitis C virus. *Nat. Rev. Microbiol*. 2007;5(6):453-463. DOI 10.1038/nrmicro1645

Papic N., Maxwell C.I., Delker D.A., Liu S., Bret S.E., Heale B.S.E., Hagedorn C.H. RNA-sequencing analysis of 5′ capped RNAs identifies many new differentially expressed genes in acute hepatitis C virus infection. *Viruses*. 2012;4(4):581-612. DOI 10.3390/v4040581

Powdrill M.H., Tchesnokov E.P., Kozak R.A., Russell R.S., Martin R., Svarovskaia E.S., Mo H., Kouyos R.D., Gotte M. Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc. Natl. Acad. Sci. USA*. 2011;108(51):20509-20513. DOI 10.1073/pnas.1105797108

Reiss S., Rebhan I., Backes P., Romero-Brey I., Erfle H., Matula P., Kaderali L., Poenisch M., Blankenburg H., Hiet M.S., Longerich T., Diehl S., Ramirez F., Balla T., Rohr K., Kaul A., Buhler S., Pepperkok R., Lengauer T., Albrecht M., Eils R., Schirmacher P., Lohmann V., Bartenschlager R. Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell Host Microbe*. 2011;9(1):32-45. DOI 10.1016/j.chom.2010.12.002

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., 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*. 2018a;15(4):20180054. DOI 10.1515/jib-2018-0054

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofestaedt R., Lavrik I.N., Rogaev E.I. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018b;11(1):61-76. DOI 10.1186/s12920-018-0331-4

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):117-131. DOI 10.1186/s12920-019-0492-9

Salloum S., Wang H., Ferguson C., Parton R.G., Tai A.W. Rab18 binds to hepatitis C virus NS5A and promotes interaction between sites of viral replication and lipid droplets. *PLoS Pathog*. 2013;9(8):e1003513. DOI 10.1371/journal.ppat.1003513

Tan Y., Li Y. HCV core protein promotes hepatocyte proliferation and chemoresistance by inhibiting NR4A1. *Biochem. Biophys. Res. Commun.* 2015;466(3):592-598. DOI 10.1016/j.bbrc.2015.09.091

Xiong J., Wang L., Fei X.C., Jiang X., Zheng Z., Zhao Y., Wang C., Li B., Chen S., Janin A., Gale R.P., Zhao W. MYC is a positive regulator of choline metabolism and impedes mitophagy-dependent

necroptosis in diffuse large B-cell lymphoma. *Blood Cancer J.* 2017;7(7):e582. DOI 10.1038/bcj.2017.61

Xu S., Pei R., Guo M., Han Q., Lai J., Wang Y., Wu C., Zhou Y., Lu M., Chen X. Cytosolic phospholipase A2 gamma is involved in hepatitis C virus replication and assembly. *J. Virol*. 2012;86(23):13025-13037. DOI 10.1128/JVI.01785-12

Xue Y.K., Tan J., Dou D.W., Chen D., Chen L.J., Ren H.P., Chen L.B., Xiong X.G., Zheng H. Effect of Kruppel-like factor 4 on Notch pathway in hepatic stellate cells. *J. Huazhong Univ. Sci. Technolog. Med. Sci*. 2016;36(6):811-816. DOI 10.1007/s11596-016-1667-7

Yamane D., McGivern D.R., Masaki T., Lemon S.M. Liver injury and disease pathogenesis in chronic hepatitis C. *Curr. Top. Microbiol. Immunol*. 2013;369:263-288. DOI 10.1007/978-3-642-27340-7_11

Yankina M.A., Saik O.V., Ivanisenko V.A., Demenkov P.S., Khusnutdinova E.K. Evaluation of prioritization methods of extrinsic apoptotic signaling pathway genes for retrieval of the new candidates associated with major depressive disorder. *Russ. J. Genet*. 2018; 54(11):1366-1374. DOI 10.1134/S1022795418110170

Yevshin I., Sharipov R., Valeev T., Kel A., Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res*. 2017;45(D1):D61-D67. DOI 10.1093/nar/gkw951

Zong Y., Panikkar A., Xu J., Antoniou A., Raynaud P., Lemaigre F., Stanger B.Z. Notch signaling controls liver development by regulating biliary differentiation. *Development*. 2009;136(10):1727-1739. DOI 10.1242/dev.029140

ORCID ID

A.A. Makarova orcid.org/0009-0005-1844-7921
E.A. Antropova orcid.org/0000-0003-2158-3252
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
P.S. Demenkov orcid.org/0000-0001-9433-8341
V.A. Ivanisenko orcid.org/0000-0002-1859-4631

Благодарности. Работа поддержана бюджетным проектом № FWNR-2022-0020.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 19.07.2023. После доработки 27.08.2023. Принята к публикации 30.08.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):784-793
DOI 10.18699/VJGB-23-91

Перевод на английский язык https://vavilov.elpub.ru/jour

Приоритизация потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза, на основе реконструкции и анализа ассоциативных генных сетей

П.С. Деменков[1, 2, 3], Е.А. Антропова[1] ✉, А.В. Адамовская[1, 3], Е.Л. Мищенко[1, 2], Т.М. Хлебодарова[1, 2], Т.В. Иванисенко[1, 2, 3], Н.В. Иванисенко[1], А.С. Вензель[1, 2, 3], И.Н. Лаврик[4], В.А. Иванисенко[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[4] Медицинский факультет Магдебургского университета им. Отто фон Герике, Магдебург, Германия
✉ nzhenia@bionet.nsc.ru

Аннотация. Гепатоцеллюлярная карцинома (ГЦК) – распространенный тяжелый тип рака печени, характеризующийся крайне агрессивным течением и низкой выживаемостью. Известно, что нарушения регуляции активации апоптоза являются одной из ключевых особенностей, свойственной большинству раковых клеток, что определяет фармакологическую индукцию апоптоза как важную стратегию терапии рака. Компьютерный дизайн химических соединений, способных целевым образом регулировать внешний сигнальный путь индукции апоптоза, представляет перспективный подход для создания новых эффективных средств терапии рака печени и других онкологических заболеваний. Однако в настоящее время большинство исследований посвящено фармакологическим воздействиям на внутренний (митохондриальный) путь апоптоза, тогда как внешний путь, индуцируемый посредством клеточных рецепторов смерти, остается вне поля зрения. Аберрантное метилирование генов наряду с инфекцией вирусом гепатита С считаются важными факторами риска развития ГЦК. Реконструкция генных сетей, описывающих молекулярные механизмы взаимодействия аберрантно метилированных генов с ключевыми участниками внешнего пути апоптоза, а также пути их регуляции белками вируса гепатита С, может дать важную информацию при поиске фармакологических мишеней. В настоящей работе были предложены 13 критериев приоритизации потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза. В основу критериев легли показатели структурно-функциональной организации реконструированных с использованием ANDSystem генных сетей ГЦК, внешнего пути апоптоза и регуляторных путей взаимодействия «вирус – внешний путь апоптоза» и «аберрантное метилирование генов – внешний путь апоптоза». Список наиболее приоритетных 100 генов-мишеней, ранжированных согласно рейтингу приоритизации, оказался статистически значимо (p-value = 0.0002) обогащен известными фармакологическими мишенями, одобренными FDA, что указывает на корректность примененного метода приоритизации. Среди перспективных потенциальных фармакологических мишеней могут быть представлены шесть генов-кандидатов (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4* и *KHDRBS1*), занимающих высокое положение в ранжированном списке согласно результатам приоритизации.
Ключевые слова: генные сети; гепатокарцинома; программируемая клеточная гибель; апоптоз; метилирование.

Для цитирования: Деменков П.С., Антропова Е.А., Адамовская А.В., Мищенко Е.Л., Хлебодарова Т.М., Иванисенко Т.В., Иванисенко Н.В., Вензель А.С., Лаврик И.Н., Иванисенко В.А. Приоритизация потенциальных фармакологических мишеней для создания лекарств против гепатокарциномы, модулирующих внешний путь апоптоза, на основе реконструкции и анализа ассоциативных генных сетей. *Вавиловский журнал генетики и селекции*. 2023;27(7):784-793. DOI 10.18699/VJGB-23-91

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

P.S. Demenkov[1, 2, 3], E.A. Antropova[1] ✉, A.V. Adamovskaya[1, 3], E.L. Mishchenko[1, 2], T.M. Khlebodarova[1, 2], T.V. Ivanisenko[1, 2, 3], N.V. Ivanisenko[1], A.S. Venzel[1, 2, 3], I.N. Lavrik[4], V.A. Ivanisenko[1, 2, 3]

[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
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany
✉ nzhenia@bionet.nsc.ru

Abstract. Hepatocellular carcinoma (HCC) is a common severe type of liver cancer characterized by an extremely aggressive course and low survival rates. It is known that disruptions in the regulation of apoptosis activation are some of the key features inherent in most cancer cells, which determines the pharmacological induction of apoptosis as an important strategy for cancer therapy. The computer design of chemical compounds capable of specifically regulating the external signaling pathway of apoptosis induction represents a promising approach for creating new effective ways of therapy for liver cancer and other oncological diseases. However, at present, most of the studies are devoted to pharmacological effects on the internal (mitochondrial) apoptosis pathway. In contrast, the external pathway induced via cell death receptors remains out of focus. Aberrant gene methylation, along with hepatitis C virus (HCV) infection, are important risk factors for the development of hepatocellular carcinoma. The reconstruction of gene networks describing the molecular mechanisms of interaction of aberrantly methylated genes with key participants of the extrinsic apoptosis pathway and their regulation by HCV proteins can provide important information when searching for pharmacological targets. In the present study, 13 criteria were proposed for prioritizing potential pharmacological targets for developing anti-hepatocarcinoma drugs modulating the extrinsic apoptosis pathway. The criteria are based on indicators of the structural and functional organization of reconstructed gene networks of hepatocarcinoma, the extrinsic apoptosis pathway, and regulatory pathways of virus-extrinsic apoptosis pathway interaction and aberrant gene methylation-extrinsic apoptosis pathway interaction using ANDSystem. The list of the top 100 gene targets ranked according to the prioritization rating was statistically significantly (*p*-value = 0.0002) enriched for known pharmacological targets approved by the FDA, indicating the correctness of the prioritization method. Among the promising potential pharmacological targets, six highly ranked genes (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4*, and *KHDRBS1*) are likely to deserve close attention.
Key words: gene networks; hepatocarcinoma; programmed cell death; apoptosis; methylation.

For citation: 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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):784-793. DOI 10.18699/VJGB-23-91

Введение

Гепатоцеллюлярная карцинома (ГЦК) является наиболее распространенной опухолевой патологией печени, охватывающей более 90 % случаев среди всех злокачественных новообразований печени и внутрипеченочных желчных протоков (Llovet et al., 2018). Она характеризуется крайне агрессивным течением и низкой выживаемостью. В отличие от большинства других видов рака, существуют некоторые зарегистрированные факторы риска возникновения ГЦК, такие как инфекции, вызванные вирусами гепатита С и В, алкоголь, жировая инфильтрация печени, гепатит, аутоиммунные или хронические холестатические заболевания (Forner et al., 2012). Исследования в области гепатоканцерогенеза показали важную роль генетических и эпигенетических механизмов, приводящих к образованию моноклональных популяций аберрантных и диспластических гепатоцитов, у которых наблюдаются эрозия теломер и повторная экспрессия теломераз, микросателлитная нестабильность, а также необратимые структурные изменения в генах и хромосомах (Balogh et al., 2016). Фенотип злокачественных гепатоцитов может быть вызван нарушением ряда генов, которые функционируют в различных регуляторных путях, что вызывает различающиеся молекулярные варианты ГЦК (Thorgeirsson, Grisham, 2002). Данная особенность патологии делает актуальными реконструкцию и анализ генных сетей, описывающих молекулярные механизмы заболевания.

В исследованиях, посвященных поиску терапевтических средств для лечения рака, центральное место занимает проблема подавления клеточной пролиферации и индукции программируемой клеточной гибели. Апоптоз, один из известных механизмов программируемой клеточной гибели, подразделяют на внутренний и внешний, в зависимости от пути индукции сигнала. Сигнал апоптоза, индуцированный клеточными рецепторами смерти, называют внешним путем, а митохондриями – внутренним (Krammer et al., 2007). В обоих случаях сигнал апоптоза инициирует активацию каспаз, ключевых ферментов апоптоза, что приводит к разрушению клетки, однако молекулярные механизмы пути передачи сигнала являются совершенно разными. Представленные в литературе исследования сфокусированы на регуляции внутреннего пути апоптоза, в области которого наметился определенный прогресс по поиску соединений, обладающих фармакологическим потенциалом для терапии ГЦК. Следует отметить, что фармакологическое воздействие на внешний путь апоптоза при ГЦК остается плохо изученным. Однако фармакологическая индукция этого пути может принести существенный, принципиально значимый прогресс для терапии рака.

Индукция апоптоза контролируется рядом белков-ингибиторов, включая c-FLIP, который блокирует активацию каспазы-8, членов антиапоптотического семейства BCl-2, ингибирующих высвобождение цитохрома С из митохондрий, а также белков XIAP, которые блокируют активацию каспазы-3, -7 и -9. Во внешнем пути апоптоза DISC, состоящий из белков PC, FADD, прокаспазы-8, -10 и c-FLIP, служит центральной платформой для активации прокаспазы-8 (Lavrik, Krammer, 2012). c-FLIP может функционировать в составе комплекса DISC как про-, так и антиапоптотически. Предполагается, что проапоптотическая функция c-FLIP опосредуется образованием гетеродимеров прокаспазы-8/c-FLIP. Ранее в совместных исследованиях, проводимых ИЦиГ СО РАН и Университетом Магдебурга, нами впервые в мире был разработан первый в своем классе химический зонд (малое химиче-

ское соединение), способный специфически связываться с c-FLIP в гетеродимерном комплексе каспаза-8/c-FLIP (Hillert et al., 2020). Данная малая молекула была получена путем компьютерного дизайна и обладала биологической активностью – способностью увеличивать активность каспазы-8 (Hillert et al., 2020).

Вирус гепатита С (ВГС) активно изучается в научной литературе как значимый фактор риска ГЦК (Axley et al., 2018). Роль ВГС показана в регуляции апоптоза, а также в аберрантном метилировании генов, которое тесно связано с ГЦК (Zheng et al., 2019; Lee, Ou, 2022).

Генные сети широко применяются для описания молекулярно-генетических механизмов различных процессов. Ранее нами была разработана программно-информационная система ANDSystem (Ivanisenko V.A. et al., 2015, 2019; Ivanisenko T.V. et al., 2020, 2022), предназначенная для реконструкции и анализа ассоциативных генных сетей, на основе автоматической экстракции знаний из научных публикаций и фактографических баз данных. С помощью реконструкции генных сетей, выполненных с использованием ANDSystem, были проведены такие исследования, как анализ взаимодействий белков вируса гепатита С с протеомом человека (Saik et al., 2016), связь ВГС с аберрантным метилированием при ГЦК (Antropova et al., 2022), интерпретация результатов метаболомного анализа пациентов SARS-Cov-2 (Ivanisenko V.A. et al., 2022), задачи приоритизации генов-кандидатов, ассоциированных с лимфедемой, большим депрессивным расстройством (Yankina et al., 2018; Saik et al., 2019), поиск новых потенциальных мишеней для действия лекарств (Saik et al., 2018a, b) и др.

На основе реконструкции и анализа генных сетей ГЦК и внешнего пути апоптоза, а также регуляторных путей, связывающих белки ВГС с аберрантно метилированными генами при ГЦК и ключевыми участниками внешнего пути апоптоза, были предложены критерии приоритизации потенциальных фармакологических мишеней против ГЦК. Анализ обогащенности 100 первых генов-мишеней, упорядоченных по результатам приоритизации, показал значимое содержание (p-value = 0.0002) в списке генов фармакологических мишеней, одобренных FDA, что свидетельствует об эффективности предложенных критериев приоритизации. Мы предполагаем, что механизмом действия лекарств, нацеленных на данные мишени, является модуляция внешнего пути апоптоза с учетом аберрантного метилирования генов, что может быть использовано при создании лекарств нового класса для терапии ГЦК. В качестве перспективных потенциальных фармакологических мишеней, входящих в первые тридцать по рейтингу, можно выделить следующие гены-кандидаты: *JUN*, *IL10*, *STAT3*, *MYC*, *TLR4* и *KHDRBS1*.

Материалы и методы

Программно-информационная система ANDSystem. Реконструкция генных сетей проводилась с использованием программно-информационной системы ANDSystem, которая осуществляет автоматическое извлечение знаний из текстов научных публикаций и фактографических баз данных с помощью методов искусственного интеллекта (Ivanisenko V.A. et al., 2019). Система ANDSystem включает в себя базу знаний, содержащую более 40 млн фактов о молекулярно-генетических взаимодействиях, в том числе физические межмолекулярные взаимодействия, регуляцию экспрессии генов, регуляцию активности, стабильности и транспорта белков. Работа над реконструкцией и анализом генных сетей в ANDSystem выполняется с помощью программы ANDVisio. Для реконструкции регуляторных путей использовалась функция Pathway Wizard, реализованная в ANDVisio, которая по заданному шаблону осуществляет поисковые обращения к базе знаний. Схематическое описание шаблонов, использованных для реконструкции регуляторных путей, приведено в Приложениях 1–4[1].

Данные, специфические для пациентов и тканей, по экспрессии генов и метилированию ДНК. При реконструкции генных сетей применялись пациент-специфические и тканеспецифические данные по экспрессии генов и метилированию ДНК. С использованием данных по тканеспецифической экспрессии генов осуществлялась фильтрация генных сетей встроенными методами ANDSystem. Информация по тканеспецифической экспрессии генов была представлена в системе ANDSystem. Сведения о дифференциальной экспрессии генов взяты из базы GEO (Barrett et al., 2013; https://www.ncbi.nlm.nih.gov/geo/). Были выбраны эксперименты, для которых имелись результаты исследования проб ткани гепатокарциномы, полученных от пациентов с этим заболеванием. Значения статистической значимости дифференциальной экспрессии генов и дифференциального метилирования в образцах опухолевых тканей гепатокарциномы по сравнению с контрольными образцами были рассчитаны в пакете программ GEO2R (Barrett et al., 2013; https://www.ncbi.nlm.nih.gov/geo/geo2r/). Параметры расчетов были выбраны по умолчанию.

FDA одобренные фармакологические мишени. Данные по фармакологическим мишеням, одобренным FDA, извлекались из ресурса Human protein atlas (Uhlén et al., 2015; https://www.proteinatlas.org/).

Метод приоритизации потенциальных фармакологических мишеней. Для приоритизации генов-кандидатов фармакологических мишеней применяли критерии, представленные в табл. 1. Результирующий вес гена оценивался как сумма весов всех критериев.

Результаты и обсуждение

Для приоритизации потенциальных фармакологических мишеней применяли 13 критериев, учитывающих различные характеристики структурно-функциональной организации генных сетей рака печени и программируемой клеточной гибели, в том числе данные, специфические для пациентов и тканей, по метилированию ДНК. В каждом критерии был введен количественный показатель веса. В качестве результирующей характеристики рассчитывали суммарный показатель по всем 13 критериям. Чтобы ранжировать гены по степени приоритетности, их упорядочивали в списке от больших значений суммарного показателя к меньшим. Таким образом, гены, обладающие более высоким приоритетом в качестве кандидатов фар-

[1] Приложения 1–7 см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx25.pdf

Таблица 1. Критерии, разработанные для приоритизации генов-кандидатов фармакологических мишеней

№ п/п	Название критерия	Значение	Характеристика
1	Представленность гена в генной сети ГЦК	score1 = 2	Ген или кодируемый им белок представлен в генной сети
		score1 = 0	Ген или кодируемый им белок не представлен в генной сети
2	Представленность гена в генной сети внешнего апоптоза	score2 = 2	Ген или кодируемый им белок представлен в генной сети
		score2 = 0	Ген или кодируемый им белок не представлен в генной сети
3	Показатель аберрантного метилирования	score3 = 3	Ген гипометилирован при ГЦК (есть данные по повышенной экспрессии)
		score3 = –5	Ген гиперметилирован (есть данные по сниженной экспрессии)
4	Показатель центральности гена в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) генами из генной сети ГЦК (см. Приложение 1)	score4 = 1+ln(Q1)	Ген представлен в регуляторной генной сети. Q1 – количество связей гена с другими объектами сети (показатель центральности по степени)
		score4 = 0	Ген не представлен в регуляторной генной сети
5	Показатель центральности белка в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) генами из генной сети ГЦК (см. Приложение 1)	score5 = 1+ln(Q2)	Белок представлен в регуляторной генной сети. Q2 – количество связей белка с другими объектами сети (показатель центральности по степени)
		score5 = 0	Белок не представлен в регуляторной генной сети
6	Показатель центральности гена в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) белками ВГС (см. Приложение 2)	score6 = 2+ln(Q3)	Ген представлен в регуляторной генной сети. Q3 – количество связей гена с другими объектами сети (показатель центральности по степени)
		score6 = 0	Ген не представлен в регуляторной генной сети
7	Показатель центральности белка в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) белками ВГС (см. Приложение 2)	score7 = 2+ln(Q4)	Белок представлен в регуляторной генной сети. Q4 – количество связей белка с другими объектами сети (показатель центральности по степени)
		score7 = 0	Белок не представлен в регуляторной генной сети
8	Показатель центральности гена в регуляторных путях (см. Приложение 3), описывающих регуляцию гиперметилированных генов белками ВГС	score8 = ln(Q5)	Ген представлен в регуляторной генной сети. Q5 – количество связей гена с другими объектами сети (показатель центральности по степени)
		score8 = 0	Ген не представлен в регуляторной генной сети
9	Показатель центральности белка в регуляторных путях (см. Приложение 3), описывающих регуляцию гиперметилированных генов белками ВГС	score9 = ln(Q6)	Белок представлен в регуляторной генной сети. Q6 – количество связей белка с другими объектами сети (показатель центральности по степени)
		score9 = 0	Белок не представлен в регуляторной генной сети
10	Показатель центральности гена в регуляторных путях (см. Приложение 3), описывающих регуляцию гипометилированных генов белками ВГС	score10 = 1+ln(Q7)	Ген представлен в регуляторной генной сети. Q7 – количество связей гена с другими объектами сети (показатель центральности по степени)
		score10 = 0	Ген не представлен в регуляторной генной сети
11	Показатель центральности белка в регуляторных путях (см. Приложение 3), описывающих регуляцию гипометилированных генов белками ВГС	score11 = 1+ln(Q8)	Белок представлен в генной сети. Q8 – количество связей белка с другими объектами сети (показатель центральности по степени)
		score11 = 0	Белок не представлен в генной сети
12	Показатель центральности гена в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) аберрантно метилированными генами (см. Приложение 4)	score12 = 2+ln(Q9)	Ген представлен в регуляторной генной сети. Q9 – количество связей гена с другими объектами сети (показатель центральности по степени)
		score12 = 0	Ген не представлен в регуляторной генной сети
13	Показатель центральности белка в регуляторных путях, описывающих регуляцию ключевых генов внешнего пути апоптоза (*CFLAR*, *CASP8* и *FADD*) аберрантно метилированными генами (см. Приложение 4)	score13 = 2+ln(Q10)	Белок представлен в регуляторной генной сети. Q10 – количество связей белка с другими объектами сети (показатель центральности по степени)
		score13 = 0	Белок не представлен в регуляторной генной сети

макологических мишеней, находились в верхней части списка (имели меньший ранг).

При расчете показателей веса генов по критериям приоритизации осуществляли реконструкцию генных сетей ГЦК и внешнего пути апоптоза, как описано ниже.

Реконструкция генной сети гепатокарциномы человека

В результате автоматизированного поиска генов, связанных с ГЦК по типу связи association, проводимого с помощью новой версии ANDSystem (Ivanisenko V.A. et al., 2019), найдено более 5100 генов. Далее встроенными методами ANDSystem была выполнена фильтрация генов по тканеспецифичности: оставлены только те гены, которые экспрессируются в печени, – 4905 генов. Затем использовался список из 1211 дифференциально экспрессируемых генов (ДЭГ), взятых на основе анализа RNA-seq в работе (Huang et al., 2011). Данные были получены из тканей десяти пациентов с HBV-ассоциированной ГЦК. В качестве контроля использовали здоровые ткани этих же пациентов.

После этого шага с помощью встроенных функций ANDVisio было проведено пересечение генной сети, реконструированной с помощью ANDSystem, и списка дифференциально экспрессирующихся генов. В результате пересечения в генной сети осталось 584 гена, которые были найдены методами ANDSystem по материалам опубликованных работ и баз данных как связанные с гепатокарциномой и одновременно присутствуют в списке дифференциально экспрессирующихся генов гепатокарциномы человека, полученных из данных RNA в (Huang et al., 2011). Далее был осуществлен поиск белков, которые экспрессируются с этих генов, а также метаболитов, связанных с этими белками прямыми взаимодействиями (связь по типу «катализ»), и реконструирована сеть взаимодействий между всеми объектами генной сети (генами, белками и метаболитами). В генной сети на этом этапе содержалось 584 гена, 580 белков, 1061 метаболит и более 16000 взаимодействий между ними.

На втором этапе генная сеть была расширена данными по пациент- и тканеспецифическому метилированию ДНК (Приложение 5). Они включали 67 генов, метилирование которых было дифференциально изменено (гипер- или гипометилированные гены) в опухолях пациентов по сравнению с контрольными пробами. После добавления в генную сеть аберрантно метилированных генов и их белковых продуктов, а также расширения генной сети метаболитами, взаимодействующими с ними, в итоговой генной сети содержалось 627 генов, 624 белка, 1105 метаболитов, 17387 взаимодействий.

Реконструкция генной сети внешнего пути апоптоза

Проведена реконструкция генной сети внешнего пути апоптоза с учетом данных GeneOntology и ANDSystem (Приложение 6). На первом шаге был сформирован список генов, участников внешнего сигнального пути апоптоза, с помощью запроса к базе данных GeneOntology. Для выполнения запроса использовались следующие ключевые слова: GO термин “extrinsic apoptotic signaling pathway” (внешний сигнальный путь апоптоза), организм “human” (человек). На основе этого запроса получен список из 259 генов. Далее список был загружен в программу ANDVisio для построения генной сети с помощью ANDSystem. С использованием ANDSystem генная сеть была расширена белками, экспрессируемыми с введенных генов, а также метаболитами, связанными с этими генами. В итоге генная сеть внешнего пути апоптоза содержала 259 генов, 260 белков и 513 метаболитов.

Результаты приоритизации генов

Всего проанализировано 1345 генов, включая участников генных сетей ГЦК и внешнего пути апоптоза, а также регуляторных путей. Результаты применения критериев приоритизации для первых 30 наиболее приоритетных генов представлены в табл. 2. Из 1345 генов 137 оказались мишенями FDA подтвержденных лекарств. В список 100 наиболее приоритетных попали 19 генов, являющихся мишенями FDA подтвержденных лекарств. Подробная информация по результатам приоритизации, содержащая количественные значения каждого из критериев, для 100 наиболее приоритетных генов приведена в Приложении 7. Из этих 19 генов-мишеней 17 характеризуются как связанные с раком (cancer-related genes). Согласно гипергеометрическому распределению, при случайном выборе 19 генов из 137 вероятность события, при котором 17 генов и более среди 19 выбранных окажутся ассоциированными с раком, равна $p = 0.0002$. Данный анализ характеризует тот факт, что 100 наиболее приоритетных генов в таблице потенциальных мишеней статистически значимо связаны с раком (уровень значимости $p = 0.0002$).

Расчет показателей критериев приоритизации, основанных на реконструкции регуляторных путей (критерии 4–13), проводился автоматически средствами ANDSystem с помощью шаблонов, приведенных в Приложениях 1–4. Реконструкция и анализ регуляторных путей гиперметированных генов вирусными белками гепатита С, результаты которых использовались в критериях приоритизации 8–11, были описаны нами ранее (Antropova et al., 2022).

Первое место в таблице рангов занимает ген *JUN* (см. табл. 2). Он относится к группе генов-лекарственных мишеней, одобренных FDA, а также связанных с раком (cancer-related genes). В литературе приведены многочисленные данные по его роли в различных видах рака. Так, было показано, что JUN влияет на развитие рака кишечника (Nateri et al., 2005), активированный JUN преимущественно экспрессируется на инвазивном фронте рака молочной железы и связан с пролиферацией и ангиогенезом (Vleugel et al., 2006).

Согласно нашим результатам, этот ген может быть задействован в регуляции внешнего пути апоптоза. Реконструированная нами регуляторная сеть, описывающая молекулярные пути, посредством которых JUN может осуществлять регуляцию маркеров внешнего пути апоптоза *CFLAR*, *CASP8* и *FADD*, представлена на рис. 1. Регуляторная сеть основана на различных выводах экспериментальных работ. Так, например, было показано, что экспрессия *FASLG* зависит от JUN – облучение повышало экспрессию FASLG в клетках ГЦК посредством активации сигнального пути JNK/c-Jun (Dong et al., 2016). Ген *FASLG* кодирует белок TNFL6 – цитокин, который связывается

Таблица 2. Наиболее значимые 30 генов по уровню приоритета

Ранг	Ген	Полное название гена	Наличие одобренных FDA* средств	Суммарный вес
1	*JUN*	Proto-oncogene c-Jun	CR**	37.4
2	*IL10*	Interleukin-10	–	30.9
3	*STAT3*	Signal transducer and activator of transcription 3	–	30.1
4	*CASP8*	Caspase-8	–	29.4
5	*TP53*	Cellular tumor antigen p53	–	28.7
6	*CFLAR*	CASP8 and FADD-like apoptosis regulator	–	28.3
7	*MYC*	Myc proto-oncogene protein	–	23.7
8	*NFKB1*	Nuclear factor NF-kappa-B p105 subunit	CR	23.2
9	*FADD*	FAS-associated death domain protein	–	23.0
10	*IL33*	Interleukin-33	–	23.0
11	*ELAVL1*	ELAV-like protein 1	–	22.9
12	*FASLG*	Tumor necrosis factor ligand superfamily member 6	–	22.8
13	*TERT*	Telomerase reverse transcriptase	–	22.5
14	*TLR4*	Toll-like receptor 4	AR***	22.4
15	*BECN1*	Beclin-1	–	22.3
16	*CLDN1*	Claudin-1	–	22.3
17	*PARP1*	Poly [ADP-ribose] polymerase 1	CR	22.3
18	*TNFRSF1A*	Tumor necrosis factor receptor superfamily member 1A	CR	21.8
19	*CDKN1A*	Cyclin-dependent kinase inhibitor 1	–	21.6
20	*SP1*	Transcription factor Sp1	–	21.1
21	*KHDRBS1*	KH domain-containing, RNA-binding, signal transduction-associated protein 1	–	20.6
22	*MCL1*	Induced myeloid leukemia cell differentiation protein	–	20.6
23	*CLDN7*	Claudin-7	–	20.3
24	*CTSD*	Cathepsin D	–	20.0
25	*FASN*	Fatty acid synthase	CR	19.1
26	*MYCN*	N-myc proto-oncogene protein	–	18.7
27	*DDIT3*	DNA damage-inducible transcript 3 protein	–	18.4
28	*TNFAIP3*	Tumor necrosis factor alpha-induced protein 3	–	18.1
29	*STAT1*	Signal transducer and activator of transcription 1	–	17.6
30	*NLRP3*	NACHT, LRR and PYD domains-containing protein 3	–	17.6

* FDA – Food and Drug Administration, Управление по санитарному надзору за качеством пищевых продуктов и медикаментов, агентство Министерства здравоохранения и социальных служб США; ** CR – гены, связанные с раком (cancer-related genes); *** AR – гены, связанные с заболеванием «возрастная дегенерация желтого пятна» (age-related macular degeneration).

с рецептором TNFRSF6/FAS, передающим сигнал апоптоза в клетки. В другом исследовании (Liu Z. et al., 2019) делеция *FASLG* ингибировала экспрессию *CASP8*, что демонстрирует еще один возможный путь влияния JUN на апоптоз (посредством CASP8).

Особый интерес представляют также одобренные организацией FDA фармакологические мишени, которые не связаны с раком, но могут быть связаны с апоптозом. В частности, в нашей таблице среди таких генов оказался *TLR4*, занимающий 14-ю позицию по рангу. По данным FDA, ген *TLR4* ассоциирован с заболеванием «возрастная дегенерация желтого пятна». Нарушение апоптоза является важным патологическим фактором при этом заболевании (Yi et al., 2012).

Регуляторная сеть, описывающая молекулярные пути, посредством которых TLR4 может осуществлять регуляцию CFLAR, CASP8 и FADD, приведена на рис. 2. Можно, например, увидеть регуляторное воздействие от TLR4 к *TNFAIP3*. Она реконструирована на основе опубликованного исследования, где показано, что TLR4 активирует сигнальный путь, приводящий к активации транскрипционного фактора NF-κB. NF-κB, в свою очередь, индуцирует экспрессию *TNFAIP3*, что продемонстрировано на эндотелиальных клетках (Soni et al., 2018). TNFAIP3

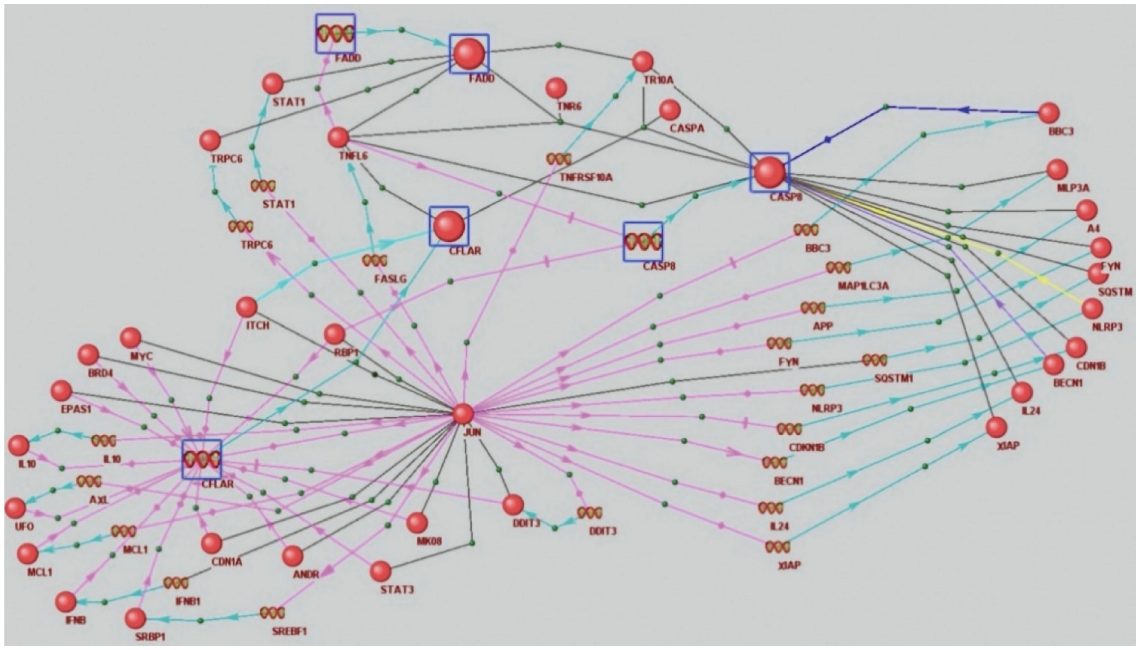


Рис. 1. Реконструированная с помощью ANDSystem сеть взаимодействий, посредством которых JUN может осуществлять регуляцию ключевых белков апоптоза – CFLAR, CASP8 и FADD.

Шарики обозначают белки, спирали обозначают гены. Черные линии – физическое взаимодействие, бирюзовые стрелки – экспрессия, розовые – регуляция экспрессии, синие – регуляция транспорта, желтые стрелки – регуляция активности.

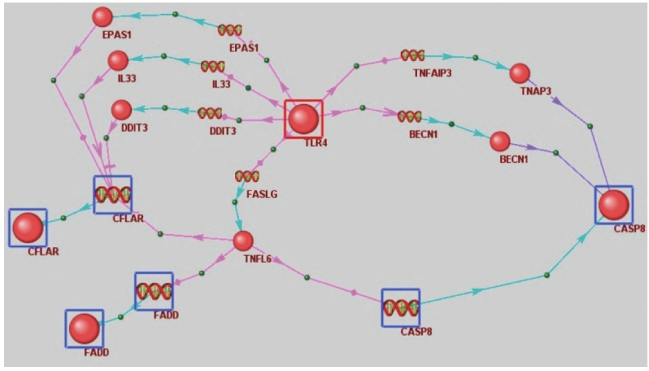


Рис. 2. Реконструированная с помощью ANDSystem сеть взаимодействий, посредством которых TLR4 может осуществлять регуляцию ключевых белков апоптоза – CFLAR, CASP8 и FADD.

Шарики обозначают белки, спирали обозначают гены. Бирюзовые стрелки – экспрессия, фиолетовые – регуляция, розовые стрелки – регуляция экспрессии.

повышает уровень расщепленной каспазы-8, что доказано с помощью нокдауна, в то время как сверхэкспрессия *TNFAIP3* влияла противоположным образом (Liu K. et al., 2018). Аналогично TLR4 через NF-κB может усиливать экспрессию *Beclin-1* (Copetti et al., 2009), который вызывает расщепление каспазы-8, что приводит к аутофагии и апоптозу (Song et al., 2014).

Вторую строчку в таблице рангов занимает ген *IL10*. Его можно отнести к группе генов, которые не входят в список одобренных FDA фармакологических мишеней, но механизмы их влияния на развитие ГЦК широко обсуждаются в литературе. В 2020 г. в работе (Qian et al., 2020) сделано предположение, что комбинация ингибиторов IL10 и PD-L1 может стать основой эффективного лечения. Регуляторная сеть, описывающая молекулярные пути, посредством которых IL10 может осуществлять регуляцию CFLAR, CASP8 и FADD, представлена на рис. 3.

Еще одна группа – гены, для которых в FDA нет указания на одобренные средства, однако механизм действия ряда широко используемых лекарственных препаратов затрагивает эти гены или кодируемые ими белки. К этой группе можно отнести гены *STAT3* и *MYC*, занимающие 3-е и 7-е положение в таблице рангов. Достаточно большое количество публикаций демонстрирует, что STAT3 играет ключевую роль в инициации, прогрессировании, иммуносупрессии и метастазировании ГЦК. Отдельные лекарственные препараты влияют на функционирование STAT3. Например, F.M. Gu с коллегами показали, что ингибирование роста и метастазирования ГЦК противоопухолевым средством направленного действия «сорафениб» опосредовано блокированием STAT3 (Gu et al., 2011). Также известно, что сорафениб индуцирует апоптоз (Xie et al., 2012). L. Wu с соавторами, изучив механизм действия кверцетина (природный флавоноид, входит в состав некоторых биологически активных добавок и препаратов), показали, что он ингибирует прогрессирование ГЦК, влияя на апоптоз, миграцию, инвазию, аутофагию, через сигнальный путь JAK2/STAT3 (по крайней мере частично) (Wu et al., 2019). Механизм действия другого противоопухолевого лекарства – траметиниба (trametinib), применяемого для лечения меланомы, основан на ингибировании белка MEK, входящего в сигнальный каскад. Ингибирование MEK приводит к снижению уровня белка MYC, способствующего выживанию клеток, а также к повышению уровня проапоптозного белка BIM, что, в свою очередь, подавляет рост ГЦК (Zhou et al., 2019).

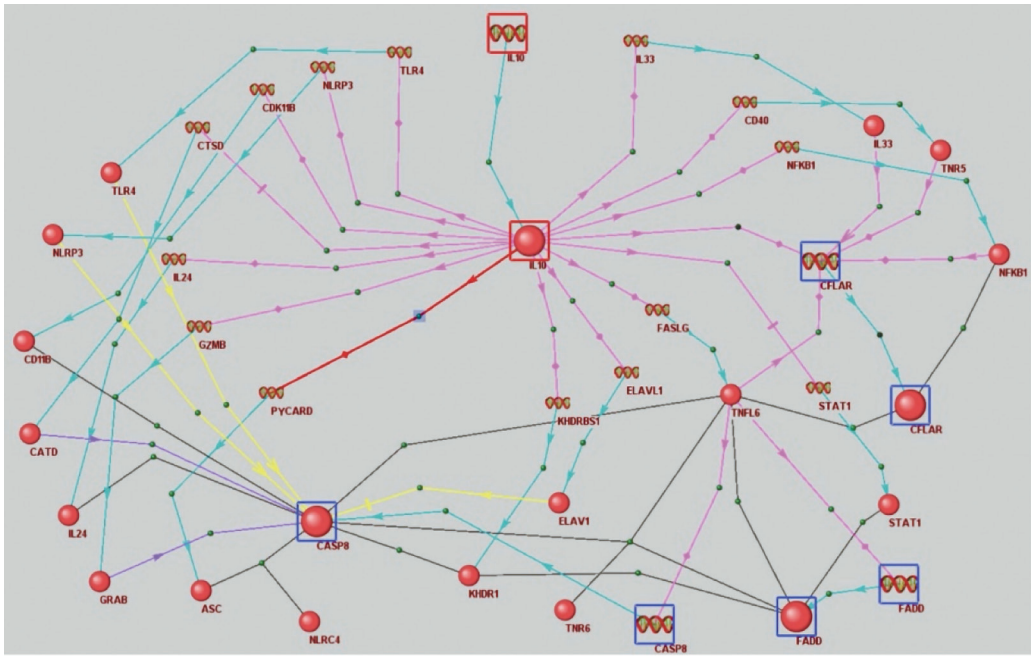


Рис. 3. Реконструированная с помощью ANDSystem сеть взаимодействий, посредством которых IL10 может влиять на CFLAR, CASP8 и FADD.

Шарики обозначают белки, спирали обозначают гены. Черные линии – физическое взаимодействие, бирюзовые стрелки – экспрессия, фиолетовые – регуляция, розовые – регуляция экспрессии, желтые стрелки – регуляция активности.

На 4-й и 6-й позициях в таблице рангов находятся непосредственно маркеры внешнего пути апоптоза *CASP8* и *CFLAR*. Между ними на 5-й позиции расположился ген *TP53*, важность которого для апоптоза хорошо известна. Таким образом, можно сделать вывод, что среди найденных нами потенциальных фармакологических мишеней на верхних строках результатов приоритизации (см. табл. 2) находятся гены, которые действительно являются мишенями лекарств – либо одобренных FDA, либо препаратов, нацеленных на другие мишени, но затрагивающих в механизмах своего действия эти гены и кодируемые ими белки, а также гены, которые еще только обсуждаются как перспективные мишени.

Особый интерес в качестве фармакологических мишеней могут представлять гены, которые к настоящему времени мало изучены по отношению к механизмам развития ГЦК. Такие гены могут быть принципиально новыми фармакологическими мишенями. В частности, в числе таких генов, попавших в список 100 наиболее приоритетных, может быть рассмотрен *KHDRBS1*, занимающий 21-ю позицию в таблице рангов (см. табл. 2). Регуляторная сеть, описывающая молекулярные пути, посредством которых KHDRBS1 может осуществлять регуляцию CFLAR, CASP8 и FADD, представлена на рис. 4.

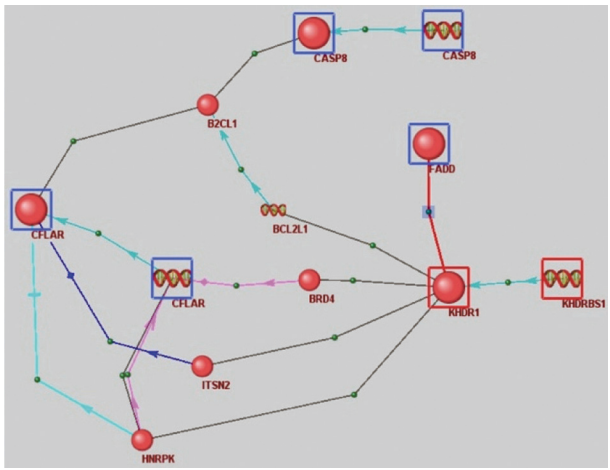


Рис. 4. Реконструированная с помощью ANDSystem сеть взаимодействий, посредством которых KHDRBS1 может осуществлять регуляцию ключевых белков апоптоза – CFLAR, CASP8 и FADD.

Шарики обозначают белки, спирали обозначают гены. Черные линии – физическое взаимодействие, бирюзовые стрелки – экспрессия, розовые стрелки – регуляция экспрессии.

Заключение

Проведена компьютерная реконструкция генных сетей гепатокарциномы и программируемой клеточной гибели (внешнего пути апоптоза), учитывающих данные, специфические для пациентов и тканей, по метилированию ДНК, выполненная с применением программно-информационной системы ANDSystem. На основе разработанных 13 критериев, учитывающих особенности структурно-функциональной организации реконструированных генных сетей, осуществлена приоритизация потенциальных фармакологических мишеней. Наибольший интерес в качестве потенциальных фармакологических мишеней могут представлять шесть генов-кандидатов (*JUN*, *IL10*, *STAT3*, *MYC*, *TLR4* и *KHDRBS1*), занимающих высокое положение в ранжированном списке согласно результатам приоритизации.

Список литературы / References

Antropova E.A., Khlebodarova T.M., Demenkov P.S., Venzel A.S., Ivanisenko N.V., Gavrilenko A.D., Ivanisenko T.V., Adamovskaya A.V., Revva P.M., Lavrik I.N., Ivanisenko V.A. Computer analysis of regulation of hepatocarcinoma marker genes hypermethylated by HCV proteins. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):733-742. DOI 10.18699/VJGB-22-89

Axley P., Ahmed Z., Ravi S., Singal A.K. Hepatitis C virus and hepatocellular carcinoma: a narrative review. *J. Clin. Transl. Hepatol*. 2018;6(1):79-84. DOI 10.14218/JCTH.2017.00067

Balogh J., Victor D., Asham E.H., Burroughs S.G., Boktour M., Saharia A., Li X., Ghobrial R.M., Monsour H.P., Jr. Hepatocellular carcinoma: a review. *J. Hepatocell. Carcinoma*. 2016;3:41-53. DOI 10.2147/JHC.S61146

Barrett T., Wilhite S.E., Ledoux P., Evangelista C., Kim I.F., Tomashevsky M., Marshall K.A., Phillippy K.H., Sherman P.M., Holko M., Yefanov A., Lee H., Zhang N., Robertson C.L., Serova N., Davis S., Soboleva A. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41(D1):D991-D995. DOI 10.1093/nar/gks1193

Copetti T., Bertoli C., Dalla E., Demarchi F., Schneider C. p65/RelA modulates BECN1 transcription and autophagy. *Mol. Cell. Biol*. 2009;29(10):2594-2608. DOI 10.1128/MCB.01396-08

Dong Y., Shen X., He M., Wu Z., Zheng Q., Wang Y., Chen Y., Wu S., Cui J., Zeng Z. Activation of the JNK-c-Jun pathway in response to irradiation facilitates Fas ligand secretion in hepatoma cells and increases hepatocyte injury. *J. Exp. Clin. Cancer Res*. 2016;35(1):114. DOI 10.1186/s13046-016-0394-z

Forner A., Llovet J.M., Bruix J. Hepatocellular carcinoma. *Lancet*. 2012;379(9822):1245-1255. DOI 10.1016/S0140-6736(11)61347-0

Gu F.M., Li Q.L., Gao Q., Jiang J.H., Huang X.Y., Pan J.F., Fan J., Zhou J. Sorafenib inhibits growth and metastasis of hepatocellular carcinoma by blocking STAT3. *World J. Gastroenterol*. 2011;17(34):3922-3932. DOI 10.3748/wjg.v17.i34.3922

Hillert L.K., Ivanisenko N.V., Busse D., Espe J., König C., Peltek S.E., Kolchanov N.A., Ivanisenko V.A., Lavrik I.N. Dissecting DISC regulation via pharmacological targeting of caspase-8/c-FLIP$_L$ heterodimer. *Cell Death Differ*. 2020;27(7):2117-2130. DOI 10.1038/s41418-020-0489-0

Huang Q., Lin B., Liu H., Ma X., Mo F., Yu W., Li L., Li H., Tian T., Wu D., Shen F., Xing J., Chen Z.N. RNA-seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. *PLoS One*. 2011;6(10):e26168. DOI 10.1371/journal.pone.0026168

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 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

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

Krammer P.H., Kamiński M., Kiessling M., Gülow K. No life without death. *Adv. Cancer Res*. 2007;97:111-138. DOI 10.1016/S0065-230X(06)97005-5

Lavrik I.N., Krammer P.H. Regulation of CD95/Fas signaling at the DISC. *Cell Death Differ*. 2012;19(1):36-41. DOI 10.1038/cdd.2011.155

Lee J., Ou J.J. Hepatitis C virus and intracellular antiviral response. *Curr. Opin. Virol*. 2022;52:244-249. DOI 10.1016/j.coviro.2021.12.010

Liu K., Yao H., Wen Y., Zhao H., Zhou N., Lei S., Xiong L. Functional role of a long non-coding RNA LIFR-AS1/miR-29a/TNFAIP3 axis in colorectal cancer resistance to pohotodynamic therapy. *Biochim. Biophys. Acta Mol. Basis Dis*. 2018;1864(9B):2871-2880. DOI 10.1016/j.bbadis.2018.05.020

Liu Z., Fitzgerald M., Meisinger T., Batra R., Suh M., Greene H., Penrice A.J., Sun L., Baxter B.T., Xiong W. CD95-ligand contributes to abdominal aortic aneurysm progression by modulating inflammation. *Cardiovasc. Res*. 2019;115(4):807-818. DOI 10.1093/cvr/cvy264

Llovet J.M., Montal R., Sia D., Finn R.S. Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol*. 2018;15(10):599-616. DOI 10.1038/s41571-018-0073-4

Nateri A.S., Spencer-Dene B., Behrens A. Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. *Nature*. 2005;437(7056):281-285. DOI 10.1038/nature03914

Qian Q., Wu C., Chen J., Wang W. Relationship between IL10 and PD-L1 in liver hepatocellular carcinoma tissue and cell lines. *Biomed. Res. Int*. 2020;2020:8910183. DOI 10.1155/2020/8910183

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., 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*. 2018a;15(4):20180054. DOI 10.1515/jib-2018-0054

Saik O.V., Demenkov P.S., Ivanisenko T.V., Bragina E.Y., Freidin M.B., Goncharova I.A., Dosenko V.E., Zolotareva O.I., Hofestaedt R., Lavrik I.N., Rogaev E.I. Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics*. 2018b;11(1):61-76. DOI 10.1186/s12920-018-0331-4

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):117-131. DOI 10.1186/s12920-019-0492-9

Song X., Kim S.Y., Zhang L., Tang D., Bartlett D.L., Kwon Y.T., Lee Y.J. Role of AMP-activated protein kinase in cross-talk between apoptosis and autophagy in human colon cancer. *Cell Death Dis*. 2014;5(10):e1504. DOI 10.1038/cddis.2014.463

Soni D., Wang D.M., Regmi S.C., Mittal M., Vogel S.M., Schlüter D., Tiruppathi C. Deubiquitinase function of A20 maintains and repairs endothelial barrier after lung vascular injury. *Cell Death Discov*. 2018;4:60. DOI 10.1038/s41420-018-0056-3

Thorgeirsson S.S., Grisham J.W. Molecular pathogenesis of human hepatocellular carcinoma. *Nat. Genet*. 2002;31(4):339-346. DOI 10.1038/ng0802-339

Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å., Kampf C., Sjöstedt E., Asplund A.,

Olsson I., Edlund K., Lundberg E., Navani S., Szigyarto C.A., Odeberg J., Djureinovic D., Takanen J.O., Hober S., Alm T., Edqvist P.H., Berling H., Tegel H., Mulder J., Rockberg J., Nilsson P., Schwenk J.M., Hamsten M., von Feilitzen K., Forsberg M., Persson L., Johansson F., Zwahlen M., von Heijne G., Nielsen J., Pontén F. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419. DOI 10.1126/science.1260419

Vleugel M.M., Greijer A.E., Bos R., van der Wall E., van Diest P.J. c-Jun activation is associated with proliferation and angiogenesis in invasive breast cancer. *Hum. Pathol*. 2006;37(6):668-674. DOI 10.1016/j.humpath.2006.01.022

Wu L., Li J., Liu T., Li S., Feng J., Yu Q., Zhang J., Chen J., Zhou Y., Ji J., Chen K., Mao Y., Wang F., Dai W., Fan X., Wu J., Guo C. Quercetin shows anti-tumor effect in hepatocellular carcinoma LM3 cells by abrogating JAK2/STAT3 signaling pathway. *Cancer Med*. 2019;8(10):4806-4820. DOI 10.1002/cam4.2388

Xie B., Wang D.H., Spechler S.J. Sorafenib for treatment of hepatocellular carcinoma: a systematic review. *Dig. Dis. Sci*. 2012;57(5):1122-1129. DOI 10.1007/s10620-012-2136-1

Yankina M.A., Saik O.V., Ivanisenko V.A., Demenkov P.S., Khusnutdinova E.K. Evaluation of prioritization methods of extrinsic apoptotic signaling pathway genes for retrieval of the new candidates associated with major depressive disorder. *Russ. J. Genet*. 2018;54(11):1366-1374. DOI 10.1134/S1022795418110170

Yi H., Patel A.K., Sodhi C.P., Hackam D.J., Hackam A.S. Novel role for the innate immune receptor Toll-like receptor 4 (TLR4) in the regulation of the Wnt signaling pathway and photoreceptor apoptosis. *PLoS One*. 2012;7(5):e36560. DOI 10.1371/journal.pone.0036560

Zheng Y., Hlady R.A., Joyce B.T., Robertson K.D., He C., Nannini D.R., Kibbe W.A., Achenbach C.J., Murphy R.L., Roberts L.R., Hou L. DNA methylation of individual repetitive elements in hepatitis C virus infection-induced hepatocellular carcinoma. *Clin. Epigenetics*. 2019;11(1):145. DOI 10.1186/s13148-019-0733-y

Zhou X., Zhu A., Gu X., Xie G. Inhibition of MEK suppresses hepatocellular carcinoma growth through independent MYC and BIM regulation. *Cell. Oncol.* (*Dordr*.). 2019;42(3):369-380. DOI 10.1007/s13402-019-00432-4

ORCID ID

P.S. Demenkov orcid.org/0000-0001-9433-8341
E.A. Antropova orcid.org/0000-0003-2158-3252
T.V. Ivanisenko orcid.org/0000-0002-0005-9155
A.S. Venzel orcid.org/0000-0002-7419-5168
V.A. Ivanisenko orcid.org/0000-0002-1859-4631

Благодарности. Исследование выполнено при финансовой поддержке проекта № 075-15-2021-944 Министерства науки и высшего образования РФ в рамках ERA-NET «Идентификация мишеней и разработка лекарственных средств при раке печени (TAIGA)».

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 26.07.2023. После доработки 25.08.2023. Принята к публикации 28.08.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):794-806
DOI 10.18699/VJGB-23-92

Перевод на английский язык https://vavilov.elpub.ru/jour

База знаний RatDEGdb по дифференциально экспрессирующимся генам крысы как модельного объекта биомедицинских исследований

И.В. Чадаева[1], С.В. Филонов[1, 2], К.А. Золотарева[1], Б.М. Хандаев[1, 2], Н.И. Ершов[1], Н.Л. Подколодный[1, 3], Р.В. Кожемякина[1], Д.А. Рассказов[1], А.Г. Богомолов[1], Е.Ю. Кондратюк[1, 4], Н.В. Климова[1], С.Г. Шихевич[1], М.А. Рязанова[1], Л.А. Федосеева[1], О.Е. Редина[1], О.С. Кожевникова[1], Н.А. Стефанова[1], Н.Г. Колосова[1], А.Л. Маркель[1, 2], М.П. Пономаренко[1] ✉, Д.Ю. Ощепков[1]

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

Аннотация. Животные модели, используемые в биомедицинских исследованиях, в настоящее время охватывают практически весь известный спектр заболеваний человека. База знаний RatDEGdb по дифференциально экспрессирующимся генам (ДЭГ) крысы как модельного объекта в биомедицинских исследованиях представляет собой коллекцию опубликованных данных по экспрессии генов у крыс разных линий, предназначенных для изучения артериальной гипертонии, болезней пожилого возраста, психопатологических состояний и других заболеваний человека. Текущий выпуск RatDEGdb содержит 25101 ДЭГ, представляющих 14320 уникальных генов крысы, которые изменяют уровень транскрипции в 21 ткани 10 генетических линий крысы в качестве моделей 11 заболеваний человека согласно 45 оригинальным научным статьям. Новшество RatDEGdb по сравнению с другими биомедицинскими базами данных заключается в курируемой аннотации отклонений ДЭГ крысы как модельного объекта с использованием независимых клинических данных об однонаправленных изменениях экспрессии гомологичных генов, выявленных у людей при различных патологиях. Собранные ДЭГ крыс были аннотированы однонаправленными изменениями экспрессии гомологичных им генов человека у больных людей относительно здоровых. К настоящему времени выпуск RatDEGdb содержит 94873 такие аннотации для 321 гена человека при 836 заболеваниях согласно 959 оригинальным научным статьям, найденным в текущем выпуске базы данных PubMed. Представленная база знаний может быть интересна в первую очередь специалистам по генетике человека, молекулярным биологам, клиницистам и генетическим консультантам, а также специалистам в области биофармацевтики, биоинформатики и персонализированной геномики. RatDEGdb является общедоступной (https://www.sysbio.ru/RatDEGdb).
Ключевые слова: база знаний; ДЭГ; крысы *Rattus norvegicus*; животные модели болезней человека; нейродегенерация; болезнь Альцгеймера; гипертоническая болезнь; преждевременное старение; психопатологические состояния; кататонический синдром; эпилепсия; агрессивность; RNA-seq; ПЦР; микрочипы.

Для цитирования: Чадаева И.В., Филонов С.В., Золотарева К.А., Хандаев Б.М., Ершов Н.И., Подколодный Н.Л., Кожемякина Р.В., Рассказов Д.А., Богомолов А.Г., Кондратюк Е.Ю., Климова Н.В., Шихевич С.Г., Рязанова М.А., Федосеева Л.А., Редина О.Е., Кожевникова О.С., 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

RatDEGdb: a knowledge base of differentially expressed genes in the rat as a model object in biomedical research

I.V. Chadaeva[1], S.V. Filonov[1, 2], K.A. Zolotareva[1], B.M. Khandaev[1], N.I. Ershov[1], N.L. Podkolodnyy[1, 3], R.V. Kozhemyakina[1], D.A. Rasskazov[1], A.G. Bogomolov[1], E.Yu. Kondratyuk[1, 4], N.V. Klimova[1], S.G. Shikhevich[1], M.A. Ryazanova[1], L.A. Fedoseeva[1], O.E. Redina[1], O.S. Kozhevnikova[1], N.A. Stefanova[1], N.G. Kolosova[1], A.L. Markel[1, 2], M.P. Ponomarenko[1] ✉, D.Yu. Oshchepkov[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
[3] Institute of Computational Mathematics and Mathematical Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[4] Siberian Federal Scientific Centre of Agro-BioTechnologies of the Russian Academy of Sciences, Krasnoobsk, Novosibirsk region, Russia
✉ pon@bionet.nsc.ru

Abstract. The animal models used in biomedical research cover virtually every human disease. RatDEGdb, a knowledge base of the differentially expressed genes (DEGs) of the rat as a model object in biomedical research is a collection of published data on gene expression in rat strains simulating arterial hypertension, age-related diseases, psychopatho-

logical conditions and other human afflictions. The current release contains information on 25,101 DEGs representing 14,320 unique rat genes that change transcription levels in 21 tissues of 10 genetic rat strains used as models of 11 human diseases based on 45 original scientific papers. RatDEGdb is novel in that, unlike any other biomedical database, it offers the manually curated annotations of DEGs in model rats with the use of independent clinical data on equal changes in the expression of homologous genes revealed in people with pathologies. The rat DEGs put in RatDEGdb were annotated with equal changes in the expression of their human homologs in affected people. In its current release, RatDEGdb contains 94,873 such annotations for 321 human genes in 836 diseases based on 959 original scientific papers found in the current PubMed. RatDEGdb may be interesting first of all to human geneticists, molecular biologists, clinical physicians, genetic advisors as well as experts in biopharmaceutics, bioinformatics and personalized genomics. RatDEGdb is publicly available at https://www.sysbio.ru/RatDEGdb.

Key words: knowledge base; DEG; *Rattus norvegicus*; animal models of human diseases; neurodegeneration; Alzheimer's disease; hypertension; premature aging; psychopathological states; catatonic syndrome; epilepsy; aggression; RNA-seq; PCR; microarrays.

For citation: Chadaeva I.V., Filonov S.V., Zolotareva K.A., Khandaev B.M., Ershov N.I., Podkolodnyy N.L., Kozhemyakina R.V., Rasskazov D.A., Bogomolov A.G., Kondratyuk E.Yu., Klimova N.V., Shikhevich S.G., Ryazanova M.A., Fedoseeva L.A., Redina O.E., Kozhevnikova O.S., Stefanova N.A., Kolosova N.G., Markel A.L., Ponomarenko M.P., Oshchepkov D.Yu. 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

Введение

Животные модели, необходимые для понимания физиологических, генетических и эпигенетических механизмов, регулирующих эволюционно закрепленные фенотипические признаки организма, должны максимально повторять симптомы изучаемой патологии и соответствовать строгим критериям (Gryksa et al., 2023). Чаще всего в этих исследованиях используют крыс и мышей, лабораторные линии которых насчитывают на сегодняшний день десятки тысяч (Гайдай Е.А., Гайдай Д.С., 2019).

Первая инбредная линия крыс была создана Кингом в 1906 г. (Филадельфия, Институт Вистар), почти одновременно с разведением мышей. Несмотря на это, мышь стала доминирующей моделью для исследований в области генетики млекопитающих, а крыса – в области физиологии и биомедицины. У лабораторных крыс есть определенные преимущества по сравнению с мышами: крысы крупнее, поэтому у них больше ткани для проведения различных анализов. Большие органы упрощают хирургические процедуры и диссекцию небольших анатомических структур.

Благодаря своей неприхотливости и низким затратам при содержании, крысы (*Rattus norvegicus*) стали удобным объектом в многочисленных биомедицинских исследованиях (Carter et al., 2020; Modlinska, Pisula, 2020). Крысы рекомендованы к использованию в качестве модельных животных для изучения старения, гипертонии, каталепсии и др. (Carter et al., 2020; Martín-Carro et al., 2023).

Существуют общепризнанные различия между дикими и лабораторными крысами. Так, лабораторные крысы имеют меньшие надпочечники и препуциальные железы, их отличают более ранняя половая зрелость, отсутствие сезонности репродуктивного цикла и более высокая плодовитость, чем у диких собратьев. Кроме того, геномы крысы и человека совпадают на 90 % (Gibbs et al., 2004). Поэтому были созданы соответствующие генетические линии лабораторных крыс, у которых проявляется моделируемая патология: например, крысы линии Zucker – классическая модель для исследования ожирения, гипертонии, сахарного диабета II типа и нарушений функции сердца (Schmidt, 2002), крысы линии Kawasaki имеют пониженный уровень рилина, который ассоциирован с шизофренией и аутизмом (Aikawa et al., 1988), крысы Brattleboro – модельный объект для исследования несахарного гипоталамического диабета (Ideno et al., 2003). Всего существует около 1000 инбредных линий лабораторных крыс, полученных путем генетической селекции, которые «фиксируют» аллели естественных болезней (Greenhouse et al., 1990), таких как психические расстройства (Taylor et al., 2002), депрессия (Bay et al., 2020), хроническая почечная недостаточность (Zhang H.F. et al., 2019). Линии Wistar и Sprague-Dawley – наиболее широко используемые лабораторные крысы в мире (Sengupta, 2013). На текущий момент база данных PubMed (Lu, 2011) содержит рефераты 19555 оригинальных научных статей, соответствующих набору ключевых слов "rats biomedical model" в качестве свидетельства актуальности этой тематики.

Следуя этому магистральному пути биомедицины, в Институте цитологии и генетики Сибирского отделения РАН (ИЦиГ СО РАН) были созданы несколько линий крыс, моделирующих человеческие заболевания. Так, крысы линии НИСАГ (ISIAH) характеризуются повышенным артериальным давлением и используются для исследований причин возникновения и способов лечения гипертонической болезни у людей (Markel, 1992; Markel et al., 1999; Fedoseeva et al., 2016a, 2019; Klimov et al., 2016; Ryazanova et al., 2016), крысы линии OXYS представляют собой уникальную селекционную модель преждевременного старения и связанных с ним заболеваний (Kozhevnikova et al., 2013; Kolosova et al., 2014; Perepechaeva et al., 2014; Stefanova et al., 2018, 2019; Stefanova, Kolosova, 2023), крысы линии МД (аббревиатура от «маятниковые движения») отличаются стереотипиями и аудиогенной эпилепсией, и, наконец, крысы линии ГК (аббревиатура от «генетическая кататония») выделяются кататоническим синдромом, который встречается при различных психических заболеваниях человека, в том числе при шизофрении (Barykina et al., 1983; Kolpakov et al., 2004; Рязанова и др., 2017; Ryazanova et al., 2023).

У крыс как модельных животных были изучены изменения экспрессии генов, связанных с исследуемым заболеванием, с использованием полуколичественной ПЦР в реальном времени отдельных ключевых генов или про-

филирования транскриптомов методами секвенирования нового поколения либо с помощью микрочипов. В результате было накоплено значительное количество данных о дифференциально экспрессирующихся генах (ДЭГ), которые статистически достоверно связаны с заболеваниями, и стало возможным осуществить сбор, сравнительный анализ и систематизацию результатов таких и подобных экспериментов с использованием биоинформатических технологий для организации баз данных и знаний.

Целью нашей работы была разработка базы знаний по ДЭГ крыс различных линий, созданных, прежде всего, в ИЦиГ СО РАН, а также в ряде других отечественных и зарубежных научных учреждений. База является свободно доступной по веб-адресу (URL=https://www.sysbio.ru/RatDEGdb).

Материалы и методы

Экспериментальные животные. Мы провели эксперименты *in vivo* на 12 взрослых самцах серой крысы (*Rattus norvegicus*) из двух аутбредных линий (более 90 поколений генетической селекции), прошедших отбор в двух направлениях (Belyaev, Borodin, 1982): на повышение и на снижение агрессивной поведенческой реакции на человека (агрессивная и ручная линии соответственно). Животных содержали в стандартных условиях вивария конвенциональных животных ИЦиГ СО РАН (Новосибирск, Россия) в клетках (50 × 33 × 20 см) группами по четыре самца, при регулируемом (12/12) освещении и свободном доступе к воде и полнорационному корму.

В эксперимент были взяты особи в возрасте двух месяцев весом 250–270 г, все животные были из разных неродственных пометов. В течение первых 4 ч световой фазы суточного цикла «свет–темнота» с каждой крысой провели тест «на перчатку» для оценки ее поведенческой реакции на человеческую руку в толстой защитной перчатке, после чего животные были оценены по шкале от «–4» до «+4» диапазона от максимально агрессивного до максимально ручного поведения согласно работе (Plyusnina, Oskina, 1997). После завершения этого теста крыс возвращали в домашние клетки в стандартные условия содержания на 1 неделю, снижая тем самым возможные эффекты теста на перчатку на экспрессию генов, вплоть до эвтаназии и препарирования образцов гипоталамуса согласно атласу (Paxinos, Watson, 2013). Образцы помещали в жидкий азот для транспортировки и дальнейшего хранения при –70 °C до их использования. Протокол экспериментальных работ одобрен Комиссией по биоэтике при ИЦиГ СО РАН (Заключение № 97 от 28.10.2021).

Измерение методом полуколичественной ПЦР уровней мРНК гена *Asmtl* в гипоталамусе у самцов серых крыс ручной и агрессивной линий. Для измерения методом полуколичественной полимеразной цепной реакции в реальном времени (ПЦР-РВ) уровней мРНК генов из полученных образцов, масса которых составляла приблизительно 100 мг каждый, была выделена РНК гипоталамуса шести агрессивных ($n = 6$) и шести ручных ($n = 6$) крыс. Суммарную РНК выделяли с помощью реагента TRIzol™ (Invitrogen, #15596018). Очистку проводили с использованием магнитных шариков Agencourt RNAClean XP Kit (Beckman, #A63987). Количество РНК оценивали с помощью флуориметра Qubit™ 2.0 (Invitrogen/Life Technologies) и набора реагентов (RNA High Sensitivity, Invitrogen #In=Q32852). На основе полученной РНК вместе с набором реагентов для обратной транскрипции (Синтол #ОТ-1) синтезировали кДНК.

Таблица 1. Праймеры для количественной полимеразной цепной реакции в реальном времени (ПЦР-РВ)

Ген	Праймеры: 5′→3′	
	прямой	обратный
Asmtl	CGCACTTCTCGGAGGTCCCGC	ACGGTCGCAGGGCTTCCCCA
Ppia	TTCCAGGATTCATGTGCCAG	CTTGCCATCCAGCCACTC
Rpl30	CATCTTGGCGTCTGATCTTG	TCAGAGTCTGTTTGTACCCC

Примечание. Праймеры отобраны с использованием свободно доступного веб-сервиса PrimerBLAST (Ye et al., 2012). Гены крысы: *Asmtl* – ацетилсеротонин-О-метилтрансфераза-подобный белок, *Ppia* – пептидилпролилизомераза А, *Rpl30* – рибосомный белок L30. ПЦР-РВ – полимеразная цепная реакция в реальном времени с использованием двух референсных генов согласно общепринятой рекомендации (Bustin et al., 2009), в качестве которых были выбраны *Ppia* (Gholami et al., 2017) и *Rpl30* (Penning et al., 2007), что экспериментально обосновано нами ранее (Chadaeva et al., 2021).

С помощью веб-сервиса PrimerBLAST (Ye et al., 2012) для заданного гена конструировали олигонуклеотидные праймеры (табл. 1). ПЦР-РВ проводили с набором EVA Green I в трех технических повторах, которые были выполнены в автоматическом режиме работы на сенсорной системе LightCycler® 96 согласно инструкции производителя (Roche, Швейцария). Эффективность полимеразной реакции определяли серией разбавлений кДНК (стандарты).

Ген *ASMT* человека кодирует фермент ацетилсеротонин-О-метилтрансферазу синтеза мелатонина, одного из гормонов регуляции молекулярно-генетических процессов на уровне функционирования всего организма, включая циркадный ритм, а также онкопротекторные (Lv et al., 2019), противовоспалительные и иммуномедиаторные механизмы (Li G. et al., 2021). Поэтому в качестве пилотного измерения методом полуколичественной ПЦР-РВ было эвристически выбрано содержание мРНК гомологичного ему гена *Asmtl* в гипоталамусе взрослых самцов крыс ручной и агрессивной линий как модельного объекта для биомедицинских исследований повышенной агрессивности. Значения мРНК *Asmtl*, согласно (Bustin et al., 2009), нормировали на значения мРНК двух генов сравнения, *Ppia* (Gholami et al., 2017) и *Rpl30* (Penning et al., 2007). Адекватность выбора *Ppia* и *Rpl30* в качестве генов сравнения при экспериментальном выявлении ДЭГ в гипоталамусе исследуемых линий агрессивных и ручных крыс с использованием ПЦР-РВ была показана нами в предыдущей работе (Chadaeva et al., 2021).

База знаний RatDEGdb. Идентифицированный в нашей работе низкий уровень экспрессии гена *Asmtl* в гипоталамусе взрослых самцов агрессивных крыс в сравнении с таковым для ручных крыс сопоставили с клиническими данными о дефиците белков, кодируемых гомологичными генами *ASMT* и *ASMTL* человека, у пациентов с различ-

ными заболеваниями в сравнении с нормой. Это сравнение привели к плоскому текстовому Excel-совместимому формату и преобразовали в базу знаний RatDEGdb о дифференциальной экспрессии генов крысы как модельного животного для биомедицинских целей (https://www.sysbio.ru/RatDEGdb) с использованием свободно доступного вебсервиса MariaDB 10.2.12 (MariaDB Corp AB, Финляндия).

Таким же образом на основе базы данных PubMed (Lu, 2011) был составлен представительный набор публикаций, отражающих текущее разнообразие как линий лабораторных крыс в качестве биомедицинских моделей заболеваний человека, так и экспериментальных методов оценки дифференциальной экспрессии генов. Затем все ДЭГ крыс из этого набора статей были документированы и загружены в базу знаний RatDEGdb вместе с их курируемыми аннотациями, аналогично описанному выше алгоритму для дефицита *Asmtl* в гипоталамусе агрессивных крыс. При этом списки генов-гомологов брали из базы данных GeneCards, раздел Paralogs (Stelzer et al., 2016). RatDEGdb содержит уровни статистической значимости ДЭГ согласно их авторским оценкам, которые были опубликованы в соответствующих цитируемых статьях.

Статистический анализ дифференциальной экспрессии гена *Asmtl* в гипоталамусе ручных и агрессивных крыс как модели агрессивного поведения человека проводили по пути “Statistics → Nonparametric → Mann–Whitney test” выбора режима работы стандартного пакета STATISTICA (Statsoft™, США), когда для проверки устойчивости результатов оценивают сразу два независимых статистических критерия: непараметрический U-тест Манна–Уитни и параметрический Z-тест Фишера.

Результаты

Дефицит мРНК-*Asmtl* в гипоталамусе агрессивных крыс в сравнении с ручными

Результаты измерения уровня мРНК гена *Asmtl* в гипоталамусе у агрессивных крыс по сравнению с ручными представлены в табл. 2. На рис. 1 можно видеть, что уровень мРНК этого гена у агрессивных крыс статистически значимо ниже, чем у ручных, в условиях рассматриваемого эксперимента ($p < 0.05$; U-тест Манна–Уитни и Z-тест Фишера).

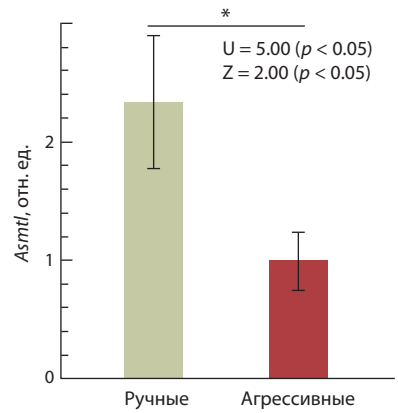


Рис. 1. Статистически достоверное различие между ручными и агрессивными взрослыми самцами крысы по уровню экспрессии гена *Asmtl* в образцах гипоталамуса.

* Уровень значимости $p < 0.05$ согласно двум независимым статистическим критериям: непараметрическому U-тесту Манна–Уитни и параметрическому Z-тесту Фишера, что отражает устойчивость к варьированию статистических критериев оценки *Asmtl* как дифференциально экспрессируемого гена агрессивных *vs* ручных крыс.

Клинические проявления дефицита *ASMTL* и *ASMT* у человека

Результаты поиска по ключевым словам в базе данных PubMed (Lu, 2011) для заболеваний человека, ассоциированных с низкой экспрессией обсуждаемого гена *ASMTL* и его паралога *ASMT* у человека, представлены в табл. 3. Прежде всего, согласно строке 1, в моделях заболеваний человека с использованием мышей с делецией гена *Asmt* (Trent et al., 2013) наблюдали нарушение развития нервной системы в виде поведенческого расстройства «дефицит внимания/гиперактивность», сочетанного с агрессивностью в рамках проявления экстернализации (специфический психический процесс) у детей (Kang et al., 2023).

Кроме того, строка 2 здесь представляет дефицит ASMT в качестве молекулярного маркера аутизма, идентифицированного в исследовании (Melke et al., 2008), тогда как недавний анализ подростков старше 12 лет с расстройствами аутистического спектра и эпилепсией в анамнезе установил их склонность к агрессивному поведению (Gaitanis et al., 2023).

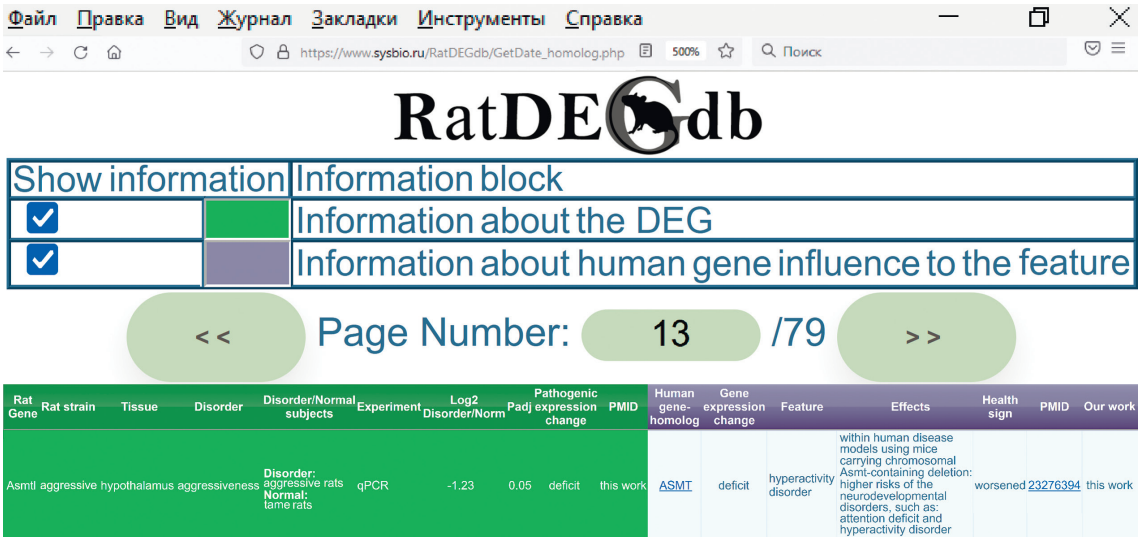
Таблица 2. Экспериментальные данные по поведению в тесте на перчатку и уровню мРНК *Asmtl* для 12 взрослых самцов крысы

Тест	Линия	Аутбредные неродственные взрослые самцы ручных и агрессивных крыс, гипоталамус						$M_0 \pm SEM$
		Крыса 1	Крыса 2	Крыса 3	Крыса 4	Крыса 5	Крыса 6	
ПТ	A	–3	–3	–3	–3	–3	–3	
	P	3	3	3	3	3	3	
ПЦР-РВ (*Asmtl*)	A	1.88 ± 0.67	0.80 ± 1.65	1.56 ± 0.51	0.70 ± 0.04	0.33 ± 0.16	0.73 ± 0.02	1.00 ± 0.24
	P	4.51 ± 0.51	1.21 ± 0.15	1.73 ± 0.63	0.92 ± 0.04	3.30 ± 0.09	2.33 ± 0.13	2.33 ± 0.56

Примечание. Линии крыс: A – агрессивные крысы ($n = 6$); P – ручные ($n = 6$). Тесты: ПТ – «перчаточный тест», где каждая крыса получала индивидуальную оценку по шкале от «–4» до «+4» диапазона между максимально агрессивным и максимально ручным поведением соответственно, согласно (Plyusnina, Oskina, 1997); экспрессия *Asmtl*: $M_0 \pm SEM$ – среднее ± стандартная ошибка среднего по трем техническим повторам при автоматической работе сенсорной системы LightCycler® 96 (Roche, Швейцария).

Таблица 3. Клиническое проявление недостаточности по *ASMTL* или по его паралогу *ASMT* при заболеваниях человека согласно текущему выпуску базы знаний RatDEGdb, созданной в настоящей работе

№ п/п	Заболевание	Клиническое проявление дефицита ASMTL или ASMT	Литературный источник
1	Нарушения развития нервной системы	В моделях заболеваний человека с использованием мышей с делецией гена *Asmt*: нарушения развития нервной системы, дефицит внимания и гиперактивность	Trent et al., 2013
2	Аутизм	В когортном исследовании: низкий уровень мРНК *ASMT* в крови и низкая экспрессия гена *ASMT*, приводящие к дефициту мелатонина, могут быть молекулярными маркерами аутизма	Melke et al., 2008
3	Депрессия с нарушениями речи и обучения	В когортном клиническом исследовании: *ASMT*-дефицит как маркер рекуррентного депрессивного расстройства с нарушением беглости речи и слухо-вербального обучения	Talarowska et al., 2014
4	Депрессия с нарушением сна и циркадного ритма	В моделях поведения человека с использованием мышей с нокаутом *Asmt*: депрессия, нарушения сна и циркадного ритма, обратимые вспять плавательными упражнениями	Liu W. et al., 2022
5	Восстановление после острых травм головного мозга (контузий)	В моделях острых травм головного мозга человека с использованием крыс, подвергнутых сильным сенсорным воздействиям: снижение уровня *Asmt* через 6 ч, 24 ч и даже через 1 месяц после воздействия – проявления контузии мозга в виде нарушения сна	Govindarajulu et al., 2022
6	Гипоксия-ишемия головного мозга	В моделях заболеваний человека с использованием новорожденных крыс: дефицит *Asmt* может быть молекулярным маркером гипоксии-ишемии головного мозга	Yang et al., 2023
7	Нарушения развития	В моделях заболеваний человека с использованием линий индуцированных плюрипотентных стволовых клеток, полученных из фибробластов кожи пациентов с каким-либо нарушением развития: дефицит *ASMTL* может быть одним из самых общих молекулярных маркеров нарушений развития	Li W. et al., 2012
8	Клеточное старение	В моделях старения человека с использованием культур клеток: замедление репликативного старения мезенхимальных стромальных клеток костного мозга человека	Liu X. et al., 2022
9	Глиома	В ретроспективном метаанализе транскриптомов, обобщающем 966 образцов глиомы: *ASMT*-дефицит может быть клиническим молекулярным маркером глиомы	Liu Y. et al., 2022
10	Рак толстой кишки	В моделях заболеваний человека с использованием клеточных линий рака толстой кишки LOVO и HCT116: пролиферация, миграция и инвазия раковых клеток снижались с понижением уровня экспрессии *ASMTL*	Bi et al., 2019
11	Рак простаты	В когортном исследовании пациентов с использованием ПЦР-РВ: повышенная активность *ASMTL* способствует развитию рака простаты	Lau, Zhang, 2000
12	Рак яичников	В когортном клиническом исследовании: пациентки с раком яичников при недостаточности по *ASMT* имели снижение средней выживаемости на несколько месяцев	Cucielo et al., 2022
13	Рак молочной железы	В когортном исследовании: ингибиторы *ASMT* снижают инвазивность клеток рака молочной железы	Xie et al., 2020
14	Бесплодие	В моделях фертильности человека с использованием баранов: ослабленная капацитация сперматозоидов; селекция линии лабораторных мышей с функциональным аллелем *Asmt*: у большинства линий лабораторных мышей есть дефекты этого гена, благодаря чему дефицит мелатонина у них снижает его негативное влияние на сперматогенез	Gonzalez-Arto et al., 2016; Zhang Z. et al., 2018
15	Воспаления дыхательных путей, астма	В моделях заболеваний человека с использованием мышей: *Asmt*-недостаточность способствует воспалению дыхательных путей, такому как астма, из-за дефицита мелатонина	Wu et al., 2020
Итого	19 болезней	24 клинических проявления недостатка *ASMTL* или *ASMT*	16 статей



Rat Gene	Rat strain	Tissue	Disorder	Disorder/Normal subjects	Experiment	Log2 Disorder/Norm	Padj	Pathogenic expression change	PMID	Human gene-homolog	Gene expression change	Feature	Effects	Health sign	PMID	Our work
Asmtl	aggressive	hypothalamus	aggressiveness	Disorder: aggressive rats Normal: tame rats	qPCR	-1.23	0.05	deficit	this work	ASMT	deficit	hyperactivity disorder	within human disease models using mice carrying chromosomal Asmt-containing deletion: higher risks of the neurodevelopmental disorders, such as: attention deficit and hyperactivity disorder	worsened	23276394	this work

Рис. 2. Пример записи в базе знаний RatDEGdb, которая документирует оригинальные экспериментальные данные о пониженном уровне экспрессии гена *Asmtl* в гипоталамусе крыс агрессивной линии в сравнении с ручной линией в качестве биомедицинской модели агрессивности как симптома заболеваний человека (см. рис. 1 и табл. 2). Эти данные аннотированы (см. табл. 3, первая строка) с использованием независимых данных о низкой экспрессии гомологичного гена *ASMT* человека у больных с гиперактивностью согласно модели заболеваний человека с использованием мышей с делецией гена *Asmt* (Trent et al., 2013).

Эти два примера свидетельствуют скорее в пользу низкой экспрессии генов человека *ASMTL* и *ASMT* как по меньшей мере сочетанных молекулярных характеристик предрасположенности к некоторым формам агрессивного поведения.

Наконец, низкую экспрессию этих генов человека выявляли в числе кандидатных молекулярных маркеров широкого спектра заболеваний, не ассоциированных с агрессивностью: депрессии (Talarowska et al., 2014), отклонений в развитии (Li W. et al., 2012), повреждений мозга (Govindarajulu et al., 2022; Yang et al., 2023), клеточного старения (Liu X. et al., 2022) и рака (Bi et al., 2019; Lau, Zhang, 2000; Xie et al., 2020; Cucielo et al., 2022; Liu Y. et al., 2022), а также бесплодия (Gonzalez-Arto et al., 2016; Zhang Z. et al., 2018) и астмы (Wu et al., 2020).

Все перечисленное отражает тот факт, что ген ASMT, кодирующий фермент ацетилсеротонин-О-метилтрансферазу синтеза мелатонина, является одним из ключевых гормонов, вовлеченных в регуляцию молекулярно-генетических процессов во всем организме человека в целом, включая агрессивность (Melke et al., 2008; Trent et al., 2013; Gaitanis et al., 2023; Kang et al., 2023), депрессию (Talarowska et al., 2014), онтогенез (Li W. et al., 2012; Zhang Z. et al., 2018), заживление ран (Govindarajulu et al., 2022; Yang et al., 2023), старение (Liu X. et al., 2022), онкопротекторные (Lv et al., 2019), противовоспалительные и иммуномедиаторные механизмы (Li G. et al., 2021).

База знаний RatDEGdb

В качестве примера работы с базой знаний RatDEGdb на рис. 2 продемонстрировано исследование уровня экспрессии гена *Asmtl* в гипоталамусе крыс агрессивной линии в сравнении с ручной. В данном случае агрессивность рассматривается как коморбидный симптом при таких заболеваниях человека, как талассемия, ожирение, карцинома (Chadaeva et al., 2016). Соответственно в RatDEGdb (см. рис. 1 и табл. 2) интегрированы результаты по низкой экспрессии гена *Asmtl* в гипоталамусе агрессивных крыс и гомологичного ему гена *ASMTL*, которые получены для людей с дефицитом внимания и гиперактивностью при нарушении развития нервной системы согласно модели заболеваний человека с использованием мышей, несущих делецию гена *Asmt* (Trent et al., 2013) (см. табл. 3).

Текущий выпуск RatDEGdb содержит информацию о ДЭГ десяти генетических линий крыс, которые служат модельным объектом для исследований 11 патологий человека (табл. 4–6). Как можно видеть в последних строках этих таблиц, на текущий момент в RatDEGdb документирован 25101 ДЭГ, представляющий 14320 уникальных генов крысы, которые изменяют уровень транскрипции в 21 ткани 10 генетических линий крысы в качестве моделей 11 заболеваний человека согласно 45 оригинальным статьям, цитируемым в последней колонке таблиц 4–6. Эти ДЭГ крысы были аннотированы однонаправленными изменениями экспрессии гомологичных им генов человека у пациентов относительно здоровых людей. Всего текущий выпуск RatDEGdb содержит 94873 такие аннотации для 321 гена человека при 836 заболеваниях согласно 959 публикациям, которые были найдены в базе данных PubMed (Lu, 2011). Таким образом, уникальность RatDEGdb по сравнению с другими биомедицинскими базами данных состоит в том, что для курирования аннотации отклонений ДЭГ крысы как модельного объекта патологии были использованы независимые клинические данные.

Обсуждение

Элементарный шаг наполнения базы данных RatDEGdb был представлен в табл. 1–3 и на рис. 1 и 2 на примере гена *Asmtl* (acetylserotonin O-methyltransferase-like), кодирующего фермент синтеза мелатонина. Анализ экспрессии

Таблица 4. Характеристика ДЭГ крыс как модельных животных в биомедицине, выявленных с использованием технологии ПЦР-РВ и документированных в базе знаний RatDEGdb

№ п/п	Линия	Ткань	Синдром	Модель	Норма	N_{DEG}	Литературный источник
1	Агрессивные	hyp	Агрессия	Агрессивные	Ручные	1	Эта работа
2		hyp	Агрессия	Агрессивные	Ручные	4	Климова и др., 2021
3		fc, hip, hyp, mb	Агрессия	Агрессивные	Ручные	21	Moskaliuk et al., 2023
4		hip, hyp, mb	Агрессия	Агрессивные	Ручные	11	Moskaliuk et al., 2022
5		hyp	Агрессия	Агрессивные	Ручные	8	Климова и др., 2021
6		hyp	Агрессия	Агрессивные	Ручные	3	Gulevich et al., 2019
7		mb, hip, fc	Агрессия	Агрессивные	Ручные	5	Kondaurova et al., 2016
8		hip, hyp, mb	Агрессия	Агрессивные	Ручные	3	Ilchibaeva et al., 2016
9		hyp	Агрессия	Агрессивные	Ручные	7	Oshchepkov et al., 2019
10		hip, hyp, mb	Агрессия	Агрессивные	Ручные	7	Ilchibaeva et al., 2015
11		fc, hip	Агрессия	Агрессивные	Ручные	2	Popova et al., 2010
12		fc	Агрессия	Агрессивные	Ручные	1	Naumenko et al., 2009
13		mb	Агрессия	Агрессивные	Ручные	1	Popova et al., 2007
14		hip	Агрессия	Агрессивные	Ручные	4	Herbeck et al., 2010
15	Ручные	hip	Агрессия	Ручные, метил	Ручные	3	Herbeck et al., 2010
16	SD	mpc, ac, pc, ic	Агрессия	Изоляция	В группе	22	Wall et al., 2012
17		Мозг	Агрессия	Агрессивные	Ручные	5	Suzuki et al., 2010
18		hip	Аутизм	SD, PPA	SD	6	Choi et al., 2018
19	ГК	hip	Кататония	ГК	WAG	1	Плеканчук, Рязанова, 2021
20		mb	Кататония	ГК	WAG	1	Рязанова и др., 2017
21	НИСАГ	Почка, myoc	НТ	НИСАГ	WAG	6	Fedoseeva et al., 2011
22		hyp, mo	НТ	НИСАГ	WAG	3	Klimov et al., 2013
23		Почка	НТ	НИСАГ	WAG	1	Fedoseeva et al., 2009
24		hyp, mo	НТ	НИСАГ, лозартан	НИСАГ	5	Климов и др., 2017
25	OXYS	Сетчатка	AMD	OXYS	Wistar	5	Perepechaeva et al., 2014
26		Сетчатка	AMD	OXYS, SkQ1	OXYS	5	Perepechaeva et al., 2014
27		Сетчатка	AMD	Wistar, SkQ1	Wistar	2	Perepechaeva et al., 2014
Итого	6 линий	14 тканей	5 болезней	11 моделей	7 моделей	143	23 статьи

Примечание. Здесь и в табл. 5 и 6: N_{DEG} – количество ДЭГ. Ткани: ac – передняя поясная кора; ag – надпочечники; bmmscs – мезенхимальные стволовые клетки костного мозга; bmp – микрососудистые перициты головного мозга; bs – ствол головного мозга; fc – лобная кора; hip – гиппокамп; hyp – гипоталамус; ic – инфралимбическая кора; lvcp – сосудистое сплетение бокового желудочка; mb – средний мозг; mo – продолговатый мозг; mpc – медиальная префронтальная кора; mt – покрышка среднего мозга; myoc – миокард; PAG – серое вещество периакведуктума; pc – прелимбическая кора; po – префронтальная кора; rc – корковое вещество почки; rm – мозговое вещество почки. Заболевания: AD – болезнь Альцгеймера; AMD – возрастная макулярная дегенерация; ARBLBD – возрастное развитие, гематоэнцефалический барьер; CRS – репликативное старение клеток; HT – гипертония; PAH – легочная гипертензия.
Модели: PPA – пропионовая кислота; SkQ1 – антиоксидант Скулачева.

гена *Asmtl* показан в гипоталамусе агрессивных крыс в сравнении с ручными, которые были использованы как модельные животные для исследования агрессивности человека. Приведены результаты анализа этого гена с помощью ПЦР в реальном времени. Эти результаты аннотированы с помощью статей об однонаправленных с *Asmtl* изменениях экспрессии гомологичных ему генов *ASMTL* и *ASMT* человека у больных в сравнении со здоровыми людьми, которые удалось найти в текущем выпуске базы данных PubMed (Lu, 2011). Затем как обобщение такой аннотации ДЭГ *Asmtl* в гипоталамусе агрессивных и ручных крыс с учетом клинических данных о гомологичных генах *ASMTL* и *ASMT* человека, прежде всего в рамках PubMed (Lu, 2011), мы собрали все ДЭГ крысы как модельного объекта биомедицинских исследований, выявленные с помощью ПЦР, RNA-seq и микрочипов. Далее отбраковали среди них те гены, которые не охарактеризованы, не аннотированы, предсказаны или не кодируют белки. Нако-

Таблица 5. Характеристика ДЭГ крыс как модельных животных в биомедицине, выявленных с использованием технологии RNA-seq и документированных в базе знаний RatDEGdb

№ п/п	Линия	Ткань	Синдром	Модель	Норма	N_{DEG}	Литературный источник
1	Агрессивные	fc	Агрессия	Агрессивные	Ручные	24	Albert et al., 2012
2		hyp	Агрессия	Агрессивные	Ручные	46	Chadaeva et al., 2021
3		hip	Агрессия	Агрессивные	Ручные	42	Oshchepkov et al., 2022a
4		mt	Агрессия	Агрессивные	Ручные	31	Oshchepkov et al., 2022b
5		PAG	Агрессия	Агрессивные	Ручные	39	Shikhevich et al., 2023
6	НИСАГ	bs	НТ	НИСАГ	WAG	206	Fedoseeva et al., 2019
7		hyp	НТ	НИСАГ	WAG	137	Klimov et al., 2016
8		rm	НТ	НИСАГ	WAG	882	Ryazanova et al., 2016
9		rc	НТ	НИСАГ	WAG	309	Fedoseeva et al., 2016b
10		ag	НТ	НИСАГ	WAG	1020	Fedoseeva et al., 2016a
11	OXYS	hip	AD	OXYS, 20 do	Wistar, 20 do	46	Stefanova et al., 2018
12		hip	AD	OXYS, 5 mo	Wistar, 5 mo	28	Stefanova et al., 2018
13		hip	AD	OXYS, 18 mo	Wistar, 18 mo	85	Stefanova et al., 2018
14		po	AD	OXYS, 20 do	Wistar, 20 do	2	Stefanova et al., 2019
15		po	AD	OXYS, 5 mo	Wistar, 5 mo	7	Stefanova et al., 2019
16		po	AD	OXYS, 18 mo	Wistar, 18 mo	73	Stefanova et al., 2019
17		Сетчатка	AMD	OXYS, 3 mo	Wistar, 3 mo	117	Kozhevnikova et al., 2013
18		Сетчатка	AMD	OXYS, 18 mo	Wistar, 18 mo	85	Kozhevnikova et al., 2013
19		po	AD	OXYS, 5 mo	OXYS, 20 do	52	Stefanova et al., 2019
20		po	AD	OXYS, 18 mo	OXYS, 5 mo	58	Stefanova et al., 2019
21		hip	AD	OXYS, 5 mo	OXYS, 20 do	135	Stefanova et al., 2018
22		hip	AD	OXYS, 18 mo	OXYS, 5 mo	197	Stefanova et al., 2018
23		Сетчатка	AMD	OXYS, 18 mo	OXYS, 3 mo	19	Kozhevnikova et al., 2013
24	Wistar	hip	AD	Wistar, 5 mo	Wistar, 20 do	150	Stefanova et al., 2018
25		hip	AD	Wistar, 18 mo	Wistar, 5 mo	190	Stefanova et al., 2018
26		Сетчатка	AMD	Wistar, 18 mo	Wistar, 3 mo	28	Kozhevnikova et al., 2013
25	SD	bmmscs	CRS	SD, 20 p	SD, 5 p	9167	Liu X. et al., 2022
26		bmmscs	CRS	SD, 5 p	SD, 5 p, ASA	1220	Liu X. et al., 2022
27		bmmscs	CRS	SD, 20 p	SD, 20 p, ASA	446	Liu X. et al., 2022
28		lvcp	ARBLBD	SD, 6 wo	SD, 15 ed	9159	Liddelow et al., 2013
29		Легкое	PAH	SD, MCT	SD	40	Xiao et al., 2020
30		rc	HT	SD, l-NAME	SD	284	Tain et al., 2015
31		rc	HT	SD, DEX	SD	44	Tain et al., 2015
32		rc	HT	SD, hfd	SD	240	Tain et al., 2015
33	SHR	Почка	HT	SHR	WKY	68	Watanabe et al., 2015
34		bmp	HT	SHR	WKY	21	Yuan et al., 2018
35	SHRSP	Почка	Инсульт	SHRSP	WKY	27	Watanabe et al., 2015
36	DSS	Почка	HT	DSS	DSS, QSYQ	13	Du et al., 2021
37		Почка	HT	DSS, $Resp18^{MUT}$	DSS	14	Ashraf et al., 2021
Итого	8 линий	17 тканей	8 болезней	17 моделей	17 моделей	24751	21 статья

Примечание. Модели: ASA – аспирин; do – дни; DEX – дексаметазон; ed – эмбриональные дни; hfd – фруктозная диета; l-NAME – N-нитро-L-аргинин метилэфир; MCT – монокроталин; $Resp18^{MUT}$ – мутантный вариант; mo – месяцы; p – пассажи; QSYQ – рецепт Qi-Shen-Yi-Qi китайской народной медицины; wo – недели.

Таблица 6. Характеристика ДЭГ крыс как модельных животных в биомедицине, которые были выявлены с использованием технологии микрочипов и документированы в базе знаний RatDEGdb

№ п/п	Линия	Ткань	Болезнь	Модель	Норма	N_{DEG}	Литературный источник
1	Wistar	ag	HT	Wistar, DEX	Wistar	93	Tharmalingam et al., 2020
2	SHR	ag	HT	SHR, 3 wo	WKY, 3 wo	12	Yoshida et al., 2014
3		ag	HT	SHR, 6 wo	WKY, 6 wo	42	Yoshida et al., 2014
4		Мозг	HT	SHR, 3 wo	WKY, 3 wo	11	Yoshida et al., 2014
5		Мозг	HT	SHR, 6 wo	WKY, 6 wo	10	Yoshida et al., 2014
6	SHRSP	ag	Инсульт	SHR, 3 wo	SHR, 3 wo	17	Yoshida et al., 2014
7		ag	Инсульт	SHR, 6 wo	SHR, 6 wo	9	Yoshida et al., 2014
8		Мозг	Инсульт	SHR, 6 wo	SHR, 6 wo	11	Yoshida et al., 2014
9		Мозг	Инсульт	SHR, 3 wo	SHR, 3 wo	2	Yoshida et al., 2014
Итого	3 линии	2 ткани	2 болезни	3 модели	5 моделей	207	2 статьи

Примечание. Модели: wo – недели.

нец, аннотировали оставшиеся ДЭГ крысы общедоступными публикациями о клинических проявлениях однонаправленных изменений экспрессии гомологичных им генов человека у больных в сравнении со здоровыми людьми и сделали эту итоговую аннотацию свободно доступной в виде базы знаний RatDEGdb (https://www.sysbio.ru/RatDEGdb).

Предлагаемая база знаний характеризует, прежде всего, дифференциально экспрессирующиеся гены селекционных линий крысы, созданных в ИЦиГ СО РАН (Новосибирск, Россия) (см. рис. 1 и 2). Крысы линии НИСАГ были взяты в качестве модели для биомедицинских исследований стресс-индуцированной артериальной гипертонии (см. табл. 4 и 5). Крыс линии OXYS использовали для изучения возрастных заболеваний и процессов старения, а крыс линии ГК – для изучения психопатологических состояний (см. табл. 4). Кроме того, в ряде экспериментов участвовали крысы ручной и агрессивной линий для изучения как одомашнивания животных (Plyusnina, Oskina, 1997; Gulevich et al., 2019; Chadaeva et al., 2021), так и агрессивности (Popova et al., 2010) в качестве симптома ожирения и талассемии (Chadaeva et al., 2016, 2019). По каждой из этих моделей, за исключением линии ГК, были проведены полногеномные секвенирования (см. табл. 4–6), тогда как у крыс линии ГК измеряли уровни экспрессии генов глутаматных рецепторов и системы катехоламинов.

Существующие базы биомедицинских данных для изучения заболеваний человека обычно сфокусированы на информации о геноме человека (Stenson et al., 2014; Singh et al., 2018; Sun et al., 2022) и содержат первичную информацию по транскриптомам. Предложенная в нашей работе база знаний RatDEGdb впервые дополняет полногеномную информацию независимых клинических данных с целью помочь в решении актуальных задач персонализированной медицины на основе учета однонаправленных изменений экспрессии генов у человека и у крысы в качестве модельного животного в рамках биомедицинских исследований. Таким образом, RatDEGdb может способствовать решению задач системной биологии и клинической медицины за счет новой исследовательской возможности сопоставлять патогенные изменения экспрессии генов у человека и у модельных животных в зависимости от заболевания.

Заключение

База знаний RatDEGdb предоставляет коллекцию экспериментальных данных и набор инструментов для интерактивного анализа в рамках проведения геномных исследований таких заболеваний, как болезнь Альцгеймера, аутизм, гипертоническая болезнь и некоторые другие. В дальнейшем мы планируем добавлять новые опубликованные данные по экспрессии генов крыс как модельных объектов для заболеваний человека, а также аннотировать их с использованием новых публикаций об однонаправленных с ними изменениях экспрессии гомологичных генов человека у больных в сравнении с условно здоровыми добровольцами.

Список литературы / References

Гайдай Е.А., Гайдай Д.С. Генетическое разнообразие экспериментальных мышей и крыс: история возникновения, способы получения и контроля. *Лаб. животные для науч. исследований*. 2019; 4:78-85. DOI 10.29296/2618723X-2019-04-09

[Gayday E.A., Gayday D.S. Genetic diversity of experimental mice and rats: history of origin, methods of production and check. *Laboratornye Zhivotnye Dlya Nauchnykh Issledovaniy = Laboratory Animals for Scientific Research*. 2019;4:78-85. DOI 10.29296/2618723X-2019-04-09 (in Russian)]

Климов Л.О., Рязанова М.А., Федосеева Л.А., Маркель А.Л. Эффекты ингибирования звеньев ренин-ангиотензиновой системы головного мозга у крыс с наследственной индуцированной стрессом артериальной гипертонией (НИСАГ). *Вавиловский журнал генетики и селекции*. 2017;21(6):735-741. DOI 10.18699/VJ17.29-o

[Klimov L.O., Ryazanova M.A., Fedoseeva L.A., Markel A.L. Effects of brain renin-angiotensin system inhibition in ISIAH rats with inherited stress-induced arterial hypertension. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2017;21(6):735-741. DOI 10.18699/VJ17.29-o (in Russian)]

Климова Н.В., Чадаева И.В., Шихевич С.Г., Кожемякина Р.В. Дифференциальная экспрессия 10 генов, ассоциированных с

агрессивным поведением, в гипоталамусе двух поколений крыс, селекционируемых по реакции на человека. *Вавиловский журнал генетики и селекции*. 2021;25(2):208-215. DOI 10.18699/VJ21.50-o
[Klimova N.V., Chadaeva I.V., Shichevich S.G., Kozhemyakina R.V. Differential expression of 10 genes in the hypothalamus of two generations of rats selected for a reaction to humans. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(2):208-215. DOI 10.18699/VJ21.50-o]

Плеканчук В.С., Рязанова М.А. Экспрессия генов глутаматных рецепторов в гиппокампе и лобной коре у крыс линии ГК с генетической каталонией. *Рос. физиол. журн. им. И.М. Сеченова*. 2021;107(2):232-242. DOI 10.31857/S0869813921020060
[Plekanchuk V.S., Ryazanova M.A. Expression of glutamate receptor genes in the hippocampus and frontal cortex in GC rat strain with genetic catatonia. *J. Evol. Biochem. Phys.* 2021;57(1):156-163. DOI 10.1134/S0022093021010154]

Рязанова М.А., Прокудина О.И., Плеканчук В.С., Алехина Т.А. Экспрессия генов системы катехоламинов в среднем мозге и реакция престимульного торможения у крыс с генетической каталонией. *Вавиловский журнал генетики и селекции*. 2017;21(7):798-803. DOI 10.18699/VJ17.296
[Ryazanova M.A., Prokudina O.I., Plekanchuk V.S., Alekhina T.A. Expression of catecholaminergic genes in the midbrain and prepulse inhibition in rats with a genetic catatonia. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2017;21(7):798-803. DOI 10.18699/VJ17.296 (in Russian)]

Aikawa H., Nonaka I., Woo M., Tsugane T., Esaki K. Shaking rat Kawasaki (SRK): a new neurological mutant rat in the Wistar strain. *Acta Neuropathol*. 1988;76:366-372. DOI 10.1007/BF00686973

Albert F.W., Somel M., Carneiro M., Aximu-Petri A., Halbwax M., Thalmann O., Blanco-Aguiar J.A., Plyusnina I.Z., Trut L., Villafuerte R., Ferrand N., Kaiser S., Jensen P., Paabo S. A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet*. 2012;8(9):e1002962. DOI 10.1371/journal.pgen.1002962

Ashraf U.M., Mell B., Jose P.A., Kumarasamy S. Deep transcriptomic profiling of Dahl salt-sensitive rat kidneys with mutant form of *Resp18*. *Biochem. Biophys. Res. Commun*. 2021;572:35-40. DOI 10.1016/j.bbrc.2021.07.071

Barykina N.N., Chepkasov I.L., Alekhina T.A., Kolpakov V.G. Selection of Wistar rats for predisposition to catalepsy. *Genetika*. 1983;19(12):2014-2021

Bay V., Happ D.F., Ardalan M., Quist A., Oggiano F., Chumak T., Hansen K., Ding M., Mallard C., Tasker R.A., Wegener G. Flinders sensitive line rats are resistant to infarction following transient occlusion of the middle cerebral artery. *Brain Res*. 2020;1737:146797. DOI 10.1016/j.brainres.2020.146797

Belyaev D.K., Borodin P.M. The influence of stress on variation and its role in evolution. *Biologisches Zentralblatt.* 1982;101(6):705-714

Bi J., Huang Y., Liu Y. Effect of NOP2 knockdown on colon cancer cell proliferation, migration, and invasion. *Transl. Cancer Res*. 2019;8(6):2274-2283. DOI 10.21037/tcr.2019.09.46

Bustin S.A., Benes V., Garson J.A., Hellemans J., Huggett J., Kubista M., Mueller R., Nolan T., Pfaffl M.W., Shipley G.L., Vandesompele J., Wittwer C.T. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem*. 2009;55(4):611-622. DOI 10.1373/clinchem.2008.112797

Carter C.S., Richardson A., Huffman D.M., Austad S. Bring back the rat! *J. Gerontol. A Biol. Sci. Med. Sci.* 2020;75(3):405-415. DOI 10.1093/gerona/glz298

Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A., Sharypova E.B., Kashina E.V., Matveeva M.Y., Arshinova T.V., Ponomarenko P.M., Arkova O.V., Bondar N.P., Savinkova L.K., Kolchanov N.A. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics*. 2016;17(Suppl.14):995. DOI 10.1186/s12864-016-3353-3

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

Chadaeva I., Ponomarenko P., Kozhemyakina R., Suslov V., Bogomolov A., Klimova N., Shikhevich S., Savinkova L., Oshchepkov D., Kolchanov N., Markel A., Ponomarenko M. Domestication explains two-thirds of differential-gene-expression variance between domestic and wild animals; the remaining one-third reflects intraspecific and interspecific variation. *Animals*. 2021;11(9):2667. DOI 10.3390/ani11092667

Choi J., Lee S., Won J., Jin Y., Hong Y., Hur T.Y., Kim J.H., Lee S.R., Hong Y. Pathophysiological and neurobehavioral characteristics of a propionic acid-mediated autism-like rat model. *PLoS One*. 2018;13(2):e0192925. DOI 10.1371/journal.pone.0192925

Cucielo M.S., Cesario R.C., Silveira H.S., Gaiotte L.B., Dos Santos S.A.A., de Campos Zuccari D.A.P., Seiva F.R.F., Reiter R.J., de Almeida Chuffa L.G. Melatonin reverses the warburg-type metabolism and reduces mitochondrial membrane potential of ovarian cancer cells independent of MT1 receptor activation. *Molecules*. 2022;27(14):4350. DOI 10.3390/molecules27144350

Du H., Xiao G., Xue Z., Li Z., He S., Du X., Zhou Z., Cao L., Wang Y., Yang J., Wang X., Zhu Y. QiShenYiQi ameliorates salt-induced hypertensive nephropathy by balancing ADRA1D and SIK1 expression in Dahl salt-sensitive rats. *Biomed. Pharmacother*. 2021;141:111941. DOI 10.1016/j.biopha.2021.111941

Fedoseeva L.A., Dymshits G.M., Markel A.L., Jakobson G.S. Renin system of the kidney in ISIAH rats with inherited stress-induced arterial hypertension. *Bull. Exp. Biol. Med*. 2009;147(2):177-180. DOI 10.1007/s10517-009-0465-7

Fedoseeva L.A., Riazanova M.A., Antonov E.V., Dymshits G.M., Markel A.L. Expression of the renin angiotensin system genes in the kidney and heart of ISIAH hypertensive rats. *Biochem. Moscow Suppl. Ser. B*. 2011;5(1):37-43. DOI 10.1134/s1990750811010069

Fedoseeva L.A., Klimov L.O., Ershov N.I., Alexandrovich Y.V., Efimov V.M., Markel A.L., Redina O.E. Molecular determinants of the adrenal gland functioning related to stress-sensitive hypertension in ISIAH rats. *BMC Genomics*. 2016a;17(Suppl.14):989. DOI 10.1186/s12864-016-3354-2

Fedoseeva L.A., Ryazanova M.A., Ershov N.I., Markel A.L., Redina O.E. Comparative transcriptional profiling of renal cortex in rats with inherited stress-induced arterial hypertension and normotensive Wistar Albino Glaxo rats. *BMC Genet*. 2016b;17(Suppl.1):12. DOI 10.1186/s12863-015-0306-9

Fedoseeva L.A., Klimov L.O., Ershov N.I., Efimov V.M., Markel A.L., Orlov Y.L., Redina O.E. The differences in brain stem transcriptional profiling in hypertensive ISIAH and normotensive WAG rats. *BMC Genomics*. 2019;20(Suppl.3):297. DOI 10.1186/s12864-019-5540-5

Gaitanis J., Nie D., Hou T., Frye R. Developmental regression followed by epilepsy and aggression: a new syndrome in autism spectrum disorder? *J. Pers. Med*. 2023;13(7):1049. DOI 10.3390/jpm13071049

Gholami K., Loh S.Y., Salleh N., Lam S.K., Hoe S.Z. Selection of suitable endogenous reference genes for qPCR in kidney and hypothalamus of rats under testosterone influence. *PLoS One*. 2017;12(6):e0176368. DOI 10.1371/journal.pone.0176368

Gibbs R.A., Weinstock G.M., Metzker M.L., Muzny D.M., Sodergren E.J., Scherer S., Scott G., Steffen D., Worley K.C., Burch P.E., … Peterson J., Guyer M., Felsenfeld A., Old S., Mockrin S., Collins F; Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature.* 2004;428(6982):493-521. DOI 10.1038/nature02426

Gonzalez-Arto M., Hamilton T.R., Gallego M., Gaspar-Torrubia E., Aguilar D., Serrano-Blesa E., Abecia J.A., Perez-Pe R., Muino-Blanco T., Cebrian-Perez J.A., Casao A. Evidence of melatonin syn-

thesis in the ram reproductive tract. *Andrology*. 2016;4(1):163-171. DOI 10.1111/andr.12117

Govindarajulu M., Patel M.Y., Wilder D.M., Long J.B., Arun P. Blast exposure dysregulates nighttime melatonin synthesis and signaling in the pineal gland: a potential mechanism of blast-induced sleep disruptions. *Brain Sci*. 2022;12(10):1340. DOI 10.3390/brainsci 12101340

Greenhouse D.D., Festing M.F.W., Hasan S., Cohen A.L. Inbred strains of rats and mutants. In: Hedrich H.J. (Ed.) Genetic Monitoring of Inbred Strains of Rats. Stuttgart: Gustav Fischer Verlag, 1990; 410-480

Gryksa K., Schmidtner A.K., Masís-Calvo M., Rodríguez-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

Gulevich R., Kozhemyakina R., Shikhevich S., Konoshenko M., Herbeck Y. Aggressive behavior and stress response after oxytocin administration in male Norway rats selected for different attitudes to humans. *Physiol. Behav*. 2019;199:210-218. DOI 10.1016/j.physbeh. 2018.11.030

Herbeck Yu.E., Os'kina I.N., Gulevich R.G., Plyusnina I.Z. Effects of maternal methyl-supplemented diet on hippocampal glucocorticoid receptor mRNA expression in rats selected for behavior. *Cytol. Genet.* (*Moscow.*). 2010;44(2):108-113. DOI 10.3103/S00954527 10020064

Ideno J., Mizukami H., Honda K., Okada T., Hanazono Y., Kume A., Saito T., Ishibashi S., Ozawa K. Persistent phenotypic correction of central diabetes insipidus using adeno-associated virus vector expressing arginine-vasopressin in Brattleboro rats. *Mol. Ther*. 2003; 8(6):895-902. DOI 10.1016/j.ymthe.2003.08.019

Ilchibaeva T.V., Kondaurova E.M., Tsybko A.S., Kozhemyakina R.V., Popova N.K., Naumenko V.S. Brain-derived neurotrophic factor (BDNF) and its precursor (proBDNF) in genetically defined fear-induced aggression. *Behav. Brain Res*. 2015;290:45-50. DOI 10.1016/j.bbr.2015.04.041

Ilchibaeva T.V., Tsybko A.S., Kozhemyakina R.V., Naumenko V.S. Expression of apoptosis genes in the brain of rats with genetically defined fear-induced aggression. *Mol. Biol.* (*Moscow*). 2016;50(5): 814-820. DOI 10.7868/S0026898416030071

Kang S., Gair S.L., Paton M.J., Harvey E.A. Racial and ethnic differences in the relation between parenting and preschoolers' externalizing behaviors. *Early Educ. Dev*. 2023;34(4):823-841. DOI 10.1080/10409289.2022.2074202

Klimov L.O., Fedoseeva L.A., Ryazanova M.A., Dymshits G.M., Markel A.L. Expression of renin-angiotensin system genes in brain structures of ISIAH rats with stress-induced arterial hypertension. *Bull. Exp. Biol. Med*. 2013;154(3):357-660. DOI 10.1007/s10517-013-1950-6

Klimov L.O., Ershov N.I., Efimov V.M., Markel A.L., Redina O.E. Genome-wide transcriptome analysis of hypothalamus in rats with inherited stress-induced arterial hypertension. *BMC Genet*. 2016; 17(Suppl.1):13. DOI 10.1186/s12863-015-0307-8

Kolosova N.G., Stefanova N.A., Korbolina E.E., Fursova A.Z., Kozhevnikova O.S. Senescence-accelerated OXYS rats: a genetic model of premature aging and age-related diseases. *Adv. Gerontol*. 2014;4:294-298. DOI 10.1134/S2079057014040146

Kolpakov V.G., Kulikov A.V., Alekhina T.A., Chuguy V.F., Petrenko O.I., Barykina N.N. Catatonia or depression: the GC rat strain as an animal model of psychopathology. *Russ. J. Genet*. 2004;40(6): 672-678. DOI 10.1023/B:RUGE.0000033315.79449.d4

Kondaurova E.M., Ilchibaeva T.V., Tsybko A.S., Kozhemyakina R.V., Popova N.K., Naumenko V.S. 5-HT1A receptor gene silencers Freud-1 and Freud-2 are differently expressed in the brain of rats with genetically determined high level of fear-induced aggression or its absence. *Behav. Brain Res*. 2016;310:20-25. DOI 10.1016/j.bbr.2016.04.050

Kozhevnikova O.S., Korbolina E.E., Ershov N.I., Kolosova N.G. Rat retinal transcriptome: effects of aging and AMD-like retinopathy. *Cell Cycle*. 2013;12(11):1745-1761. DOI 10.4161/cc.24825

Lau Y.F., Zhang J. Expression analysis of thirty one Y chromosome genes in human prostate cancer. *Mol. Carcinog*. 2000;27(4):308-321. DOI 10.1002/(sici)1098-2744(200004)27:4<308::aid-mc9>3.0.co;2-r

Li G., Lv D., Yao Y., Wu H., Wang J., Deng S., Song Y., Guan S., Wang L., Ma W., Yang H., Yan L., Zhang J., Ji P., Zhang L., Lian Z., Liu G. Overexpression of ASMT likely enhances the resistance of transgenic sheep to brucellosis by influencing immune-related signaling pathways and gut microbiota. *FASEB J*. 2021;35(9):e21783. DOI 10.1096/fj.202100651r

Li W., Wang X., Fan W., Zhao P., Chan Y.C., Chen S., Zhang S., Guo X., Zhang Y., Li Y., Cai J., Qin D., Li X., Yang J., Peng T., Zychlinski D., Hoffmann D., Zhang R., Deng K., Ng K.M., Menten B., Zhong M., Wu J., Li Z., Chen Y., Schambach A., Tse H.F., Pei D., Esteban M.A. Modeling abnormal early development with induced pluripotent stem cells from aneuploid syndromes. *Hum. Mol. Genet*. 2012;21(1):32-45. DOI 10.1093/hmg/ddr435

Liddelow S.A., Dziegielewska K.M., Ek C.J., Habgood M.D., Bauer H., Bauer H.C., Lindsay H., Wakefield M.J., Strazielle N., Kratzer I., Mollgard K., Ghersi-Egea J.F., Saunders N.R. Mechanisms that determine the internal environment of the developing brain: a transcriptomic, functional and ultrastructural approach. *PLoS One*. 2013;8(7):e65629. DOI 10.1371/journal.pone.0065629

Liu W., Huang Z., Xia J., Cui Z., Li L., Qi Z., Liu W. Gene expression profile associated with Asmt knockout-induced depression-like behaviors and exercise effects in mouse hypothalamus. *Biosci. Rep*. 2022;42(7):bsr20220800. DOI 10.1042/bsr20220800

Liu X., Zhan Y., Xu W., Liu L., Liu X., Da J., Zhang K., Zhang X., Wang J., Liu Z., Jin H., Zhang B., Li Y. Characterization of transcriptional landscape in bone marrow-derived mesenchymal stromal cells treated with aspirin by RNA-seq. *PeerJ*. 2022;10:e12819. DOI 10.7717/peerj.12819

Liu Y., Xiang J., Liao Y., Peng G., Shen C. Identification of tryptophan metabolic gene-related subtypes, development of prognostic models, and characterization of tumor microenvironment infiltration in gliomas. *Front. Mol. Neurosci*. 2022;15:1037835. DOI 10.3389/fnmol.2022.1037835

Lu Z. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. Database (Oxford). 2011;2011:baq036. DOI 10.1093/database/baq036

Lv J.W., Zheng Z.Q., Wang Z.X., Zhou G.Q., Chen L., Mao Y.P., Lin A.H., Reiter R.J., Ma J., Chen Y.P., Sun Y. Pan-cancer genomic analyses reveal prognostic and immunogenic features of the tumor melatonergic microenvironment across 14 solid cancer types. *J. Pineal Res*. 2019;66(3):e12557. DOI 10.1111/jpi.12557

Markel A.L. Development of a new strain of rats with inherited stress-induced arterial hypertension. In: Sassard J. (Ed.) Genetic Hypertension. London: John Libbey Eurotext Ltd., 1992;218:405-407

Markel A.L., Maslova L.N., Shishkina G.T., Mahanova N.A., Jacobson G.S. Developmental influences on blood pressure regulation in ISIAH rats. In: McCarty R., Blizard D.A., Chevalier R.L. (Eds.) Development of the Hypertensive Phenotype: Basic and Clinical Studies. In the series Handbook of Hypertension. Amsterdam: Elsevier, 1999;493-526

Martín-Carro B., Donate-Correa J., Fernández-Villabrille S., Martín-Vírgala J., Panizo S., Carrillo-López N., Martínez-Arias L., Navarro-González J.F., Naves-Díaz M., Fernández-Martín J.L., Alonso-Montes C., Cannata-Andía J.B. Experimental models to study diabetes mellitus and its complications: limitations and new opportunities. *Int. J. Mol. Sci.* 2023;24(12):10309. DOI 10.3390/ijms 241210309

Melke J., Goubran Botros H., Chaste P., Betancur C., Nygren G., Anckarsäter H., Rastam M., Ståhlberg O., Gillberg I.C., Delorme R., Chabane N., Mouren-Simeoni M.C., Fauchereau F., Durand C.M., Chevalier F., Drouot X., Collet C., Launay J.M., Leboyer M., Gillberg C., Bourgeron T. Abnormal melatonin synthesis in autism spec-

trum disorders. *Mol. Psychiatry*. 2008;13(1):90-98. DOI 10.1038/sj.mp.4002016

Modlinska K., Pisula W. The Norway rat, from an obnoxious pest to a laboratory pet. *eLife*. 2020;9:e50651. DOI 10.7554/eLife.50651

Moskaliuk V.S., Kozhemyakina R.V., Bazovkina D.V., Terenina E., Khomenko T.M., Volcho K.P., Salakhutdinov N.F., Kulikov A.V., Naumenko V.S., Kulikova E. On an association between fear-induced aggression and striatal-enriched protein tyrosine phosphatase (STEP) in the brain of Norway rats. *Biomed. Pharmacother*. 2022;147:112667. DOI 10.1016/j.biopha.2022.112667

Moskaliuk V.S., Kozhemyakina R.V., Khomenko T.M., Volcho K.P., Salakhutdinov N.F., Kulikov A.V., Naumenko V.S., Kulikova E.A. On associations between fear-induced aggression, *Bdnf* transcripts, and serotonin receptors in the brains of Norway rats: an influence of antiaggressive drug TC-2153. *Int. J. Mol. Sci*. 2023;24(2):983. DOI 10.3390/ijms24020983

Naumenko V.S., Kozhemjakina R.V., Plyusnina I.Z., Popova N.K. Expression of serotonin transporter gene and startle response in rats with genetically determined fear-induced aggression. *Bull. Exp. Biol. Med*. 2009;147(1):81-83. DOI 10.1007/s10517-009-0441-2

Oshchepkov D., Ponomarenko M., Klimova N., Chadaeva I., Bragin A., Sharypova E., Shikhevich S., Kozhemyakina R. A rat model of human behavior provides evidence of natural selection against underexpression of aggressiveness-related genes in humans. *Front. Genet*. 2019;10:1267. DOI 10.3389/fgene.2019.01267

Oshchepkov D., Chadaeva I., Kozhemyakina R., Zolotareva K., Khandaev B., Sharypova E., Ponomarenko P., Bogomolov A., Klimova N.V., Shikhevich S., Redina O., Kolosova N.G., Nazarenko M., Kolchanov N.A., Markel A., Ponomarenko M. Stress reactivity, susceptibility to hypertension, and differential expression of genes in hypertensive compared to normotensive patients. *Int. J. Mol. Sci*. 2022a;23(5):2835. DOI 10.3390/ijms23052835

Oshchepkov D., Chadaeva I., Kozhemyakina R., Shikhevich S., Sharypova E., Savinkova L., Klimova N.V., Tsukanov A., Levitsky V.G., Markel A.L. Transcription factors as important regulators of changes in behavior through domestication of gray rats: quantitative data from RNA sequencing. *Int. J. Mol. Sci*. 2022b;23(20):12269. DOI 10.3390/ijms232012269

Paxinos G., Watson C. The Rat Brain in Stereotaxic Coordinates. London: Acad. Press, Elsevier Inc., 2013.

Penning L.C., Vrieling H.E., Brinkhof B., Riemers F.M., Rothuizen J., Rutteman G.R., Hazewinkel H.A. A validation of 10 feline reference genes for gene expression measurements in snap-frozen tissues. *Vet. Immunol. Immunopathol*. 2007;120(3-4):212-222. DOI 10.1016/j.vetimm.2007.08.006

Perepechaeva M.L., Grishanova A.Y., Rudnitskaya E.A., Kolosova N.G. The mitochondria-targeted antioxidant SkQ1 downregulates aryl hydrocarbon receptor-dependent genes in the retina of OXYS rats with AMD-like retinopathy. *J. Ophthalmol*. 2014;2014:530943. DOI 10.1155/2014/530943

Plyusnina I., Oskina I. Behavioral and adrenocortical responses to open-field test in rats selected for reduced aggressiveness toward humans. *Physiol. Behav*. 1997;61(3):381-385. DOI 10.1016/S0031-9384(96)00445-3

Popova N.K., Naumenko V.S., Plyusnina I.Z. Involvement of brain serotonin 5-HT1A receptors in genetic predisposition to aggressive behavior. *Neurosci. Behav. Physiol*. 2007;37(6):631-635. DOI 10.1007/s11055-007-0062-z

Popova N.K., Naumenko V.S., Kozhemyakina R.V., Plyusnina I.Z. Functional characteristics of serotonin 5-HT2A and 5-HT2C receptors in the brain and the expression of the 5-HT2A and 5-HT2C receptor genes in aggressive and non-aggressive rats. *Neurosci. Behav. Physiol*. 2010;40(4):357-361. DOI 10.1007/s11055-010-9264-x

Ryazanova M.A., Fedoseeva L.A., Ershov N.I., Efimov V.M., Markel A.L., Redina O.E. The gene-expression profile of renal medulla in ISIAH rats with inherited stress-induced arterial hypertension. *BMC Genet*. 2016;17(Suppl.3):151. DOI 10.1186/s12863-016-0462-6

Ryazanova M.A., Plekanchuk V.S., Prokudina O.I., Makovka Y.V., Alekhina T.A., Redina O.E., Markel A.L. Animal models of hypertension (ISIAH rats), catatonia (GC rats), and audiogenic epilepsy (PM rats) developed by breeding. *Biomedicines*. 2023;11(7):1814. DOI 10.3390/biomedicines11071814

Sengupta P. The laboratory rat: relating its age with human's. *Int. J. Prev. Med*. 2013;4(6):624-630

Schmidt I. Metabolic diseases: the environment determines the odds, even for genes. *News Physiol. Sci*. 2002;17:115-121. DOI 10.1152/nips.01380.2001

Shikhevich S., Chadaeva I., Khandaev B., Kozhemyakina R., Zolotareva K., Kazachek A., Oshchepkov D., Bogomolov A., Klimova N.V., Ivanisenko V.A., Demenkov P., Mustafin Z., Markel A., Savinkova L., Kolchanov N.A., Kozlov V., Ponomarenko M. Differentially expressed genes and molecular susceptibility to human age-related diseases. *Int. J. Mol. Sci*. 2023;24(4):3996. DOI 10.3390/ijms24043996

Singh G., Bhat B., Jayadev M.S.K., Madhusudhan C., Singh A. mutTCPdb: a comprehensive database for genomic variants of a tropical country neglected disease-tropical calcific pancreatitis. *Database* (*Oxford*). 2018;2018:bay043. DOI 10.1093/database/bay043

Stefanova N.A., Kolosova N.G. The rat brain transcriptome: from infancy to aging and sporadic Alzheimer's disease-like pathology. *Int. J. Mol. Sci*. 2023;24(2):1462. DOI 10.3390/ijms24021462

Stefanova N.A., Maksimova K.Y., Rudnitskaya E.A., Muraleva N.A., Kolosova N.G. Association of cerebrovascular dysfunction with the development of Alzheimer's disease-like pathology in OXYS rats. *BMC Genomics*. 2018;19(Suppl.3):75. DOI 10.1186/s12864-018-4480-9

Stefanova N.A., Ershov N.I., Maksimova K.Y., Muraleva N.A., Tyumentsev M.A., Kolosova N.G. The rat prefrontal-cortex transcriptome: effects of aging and sporadic Alzheimer's disease-like pathology. *J. Gerontol. A Biol. Sci. Med. Sci*. 2019;74(1):33-43. DOI 10.1093/gerona/gly198

Stelzer G., Rosen N., Plaschkes I., Zimmerman S., Twik M., Fishilevich S., Stein T.I., Nudel R., Lieder I., Mazor Y., Kaplan S., Dahary D., Warshawsky D., Guan-Golan Y., Kohn A., Rappaport N., Safran M., Lancet D. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*. 2016;54:1.30.1-1.30.33. DOI 10.1002/cpbi.5

Stenson P.D., Mort M., Ball E.V., Shaw K., Phillips A., Cooper D.N. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet*. 2014;133(1):1-9. DOI 10.1007/s00439-013-1358-4

Sun S., Wang Y., Maslov A.Y., Dong X., Vijg J. SomaMutDB: a database of somatic mutations in normal human tissues. *Nucleic Acids Res*. 2022;50(D1):D1100-D1108. DOI 10.1093/nar/gkab914

Suzuki H., Han S.D., Lucas L.R. Increased 5-HT1B receptor density in the basolateral amygdala of passive observer rats exposed to aggression. *Brain Res. Bull*. 2010;83(1-2):38-43. DOI 10.1016/j.brainresbull.2010.06.007

Tain Y.L., Huang L.T., Chan J.Y., Lee C.T. Transcriptome analysis in rat kidneys: importance of genes involved in programmed hypertension. *Int. J. Mol. Sci*. 2015;16(3):4744-4758. DOI 10.3390/ijms16034744

Talarowska M., Szemraj J., Zajączkowska M., Galecki P. ASMT gene expression correlates with cognitive impairment in patients with recurrent depressive disorder. *Med. Sci. Monit*. 2014;20:905-912. DOI 10.12659/MSM.890160

Taylor J.R., Morshed S.A., Parveen S., Mercadante M.T., Scahill L., Peterson B.S., King R.A., Leckman J.F., Lombroso P.J. An animal model of Tourette's syndrome. *Am. J. Psychiatry*. 2002;159(4):657-660. DOI 10.1176/appi.ajp

Tharmalingam S., Khurana S., Murray A., Lamothe J., Tai T.C. Whole transcriptome analysis of adrenal glands from prenatal glucocorticoid programmed hypertensive rodents. *Sci. Rep*. 2020;10(1):18755. DOI 10.1038/s41598-020-75652-y

Trent S., Dean R., Veit B., Cassano T., Bedse G., Ojarikre O.A., Humby T., Davies W. Biological mechanisms associated with increased perseveration and hyperactivity in a genetic mouse model of neurodevelopmental disorder. *Psychoneuroendocrinology*. 2013; 38(8):1370-1380. DOI 10.1016/j.psyneuen.2012.12.002

Wall V.L., Fischer E.K., Bland S.T. Isolation rearing attenuates social interaction-induced expression of immediate early gene protein products in the medial prefrontal cortex of male and female rats. *Physiol. Behav*. 2012;107(3):440-450. DOI 10.1016/j.physbeh.2012.09.002

Watanabe Y., Yoshida M., Yamanishi K., Yamamoto H., Okuzaki D., Nojima H., Yasunaga T., Okamura H., Matsunaga H., Yamanishi H. Genetic analysis of genes causing hypertension and stroke in spontaneously hypertensive rats: gene expression profiles in the kidneys. *Int. J. Mol. Med*. 2015;36(3):712-724. DOI 10.3892/ijmm.2015.2281

Wu H.M., Zhao C.C., Xie Q.M., Xu J., Fei G.H. TLR2-melatonin feedback loop regulates the activation of NLRP3 inflammasome in murine allergic airway inflammation. *Front. Immunol*. 2020;11:172. DOI 10.3389/fimmu.2020.00172

Xiao G., Wang T., Zhuang W., Ye C., Luo L., Wang H., Lian G., Xie L. RNA sequencing analysis of monocrotaline-induced PAH reveals dysregulated chemokine and neuroactive ligand receptor pathways. *Aging* (*Albany NY*). 2020;12(6):4953-4969. DOI 10.18632/aging.102922

Xie F., Wang L., Liu Y., Liu Z., Zhang Z., Pei J., Wu Z., Zhai M., Cao Y. ASMT regulates tumor metastasis through the circadian clock system in triple-negative breast cancer. *Front. Oncol*. 2020;10:537247. DOI 10.3389/fonc.2020.537247

Yang H., Zhang Z., Ding X., Jiang X., Tan L., Lin C., Xu L., Li G., Lu L., Qin Z., Feng X., Li M. RP58 knockdown contributes to hypoxia-ischemia-induced pineal dysfunction and circadian rhythm disruption in neonatal rats. *J. Pineal Res*. 2023;75(1):e12885. DOI 10.1111/jpi.12885

Ye J., Coulouris G., Zaretskaya I., Cutcutache I., Rozen S., Madden T.L. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 2012;13:134. DOI 10.1186/1471-2105-13-134

Yoshida M., Watanabe Y., Yamanishi K., Yamashita A., Yamamoto H., Okuzaki D., Shimada K., Nojima H., Yasunaga T., Okamura H., Matsunaga H., Yamanishi H. Analysis of genes causing hypertension and stroke in spontaneously hypertensive rats: gene expression profiles in the brain. *Int. J. Mol. Med*. 2014;33(4):887-896. DOI 10.3892/ijmm.2014.1631

Yuan X., Wu Q., Liu X., Zhang H., Xiu R. Transcriptomic profile analysis of brain microvascular pericytes in spontaneously hypertensive rats by RNA-Seq. *Am. J. Transl. Res*. 2018;10(8):2372-2386. PMID 30210677

Zhang H.F., Wang J.H., Wang Y.L., Gao C., Gu Y.T., Huang J., Wang J.H., Zhang Z. Salvianolic acid A protects the kidney against oxidative stress by activating the Akt/GSK-3β/Nrf2 signaling pathway and inhibiting the NF-κB signaling pathway in 5/6 nephrectomized rats. *Oxid. Med. Cell. Longev*. 2019;2019:2853534. DOI 10.1155/2019/2853534

Zhang Z., Silveyra E., Jin N., Ribelayga C.P. A congenic line of the C57BL/6J mouse strain that is proficient in melatonin synthesis. *J. Pineal Res*. 2018;65(3):e12509. DOI 10.1111/jpi.12509

ORCID ID

I.V. Chadaeva orcid.org/0000-0002-2724-5441
N.I. Ershov orcid.org/0000-0003-3423-3497
N.L. Podkolodnyy orcid.org/0000-0001-9132-7997
R. Kozhemyakina orcid.org/0000-0001-8948-1127
D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
E.Yu. Kondratyuk orcid.org/0000-0001-8672-7216
O.E. Redina orcid.org/0000-0003-0942-8460
O.S. Kozhevnikova orcid.org/0000-0001-6475-4061
N.A. Stefanova orcid.org/0000-0001-5127-5993
N.G. Kolosova orcid.org/0000-0003-2398-8544
A.L. Markel orcid.org/0000-0002-1550-1647
M.P. Ponomarenko orcid.org/0000-0003-1663-318X
D.Yu. Oshchepkov orcid.org/0000-0002-6097-5155

Благодарности. Авторы выражают благодарность Центру коллективного пользования (ЦКП) «Биоинформатика» за предоставление вычислительных ресурсов по бюджетному проекту FWNR-2022-0020, а также ЦКП «Виварий конвенциональных животных» за предоставление животных в рамках бюджетных проектов FWNR-2022-0019 и FWNR-2022-0015.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 23.08.2023. После доработки 11.09.2023. Принята к публикации 15.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):807-814
DOI 10.18699/VJGB-23-93

Перевод на английский язык https://vavilov.elpub.ru/jour

Применение метода взвешенных гистограмм для расчета термодинамических параметров формирования комплексов олигодезоксирибонуклеотидов

И.И. Юшин[1, 2], В.М. Голышев[1, 2], Д.В. Пышный[1], 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@niboch.nsc.ru

Аннотация. На сегодняшний день разработан широкий спектр производных и аналогов нуклеиновых кислот. Некоторые из них нашли применение при решении научно-исследовательских задач и задач биомедицины. Детальная информация о свойствах таких соединений является основой их эффективного использования. Одну из наиболее значимых физико-химических характеристик олигонуклеотидов – термодинамическую стабильность их дуплексов с ДНК и РНК – можно рассчитывать лишь для некоторых производных нуклеиновых кислот: LNA, мостиковых олигонуклеотидов и PNA. Существующие подходы основаны на анализе экспериментальных данных и построении прогностических моделей. Проводятся пилотные исследования, направленные на разработку методов прогнозирования свойств нуклеиновых кислот с использованием методов компьютерного моделирования, основанные только на знании структуры олигомеров. В данной работе исследована применимость метода взвешенных гистограмм (WHAM) при анализе зонтичной выборки для расчета термодинамических параметров формирования ДНК-дуплексов: изменения энтальпии ΔH°, энтропии ΔS° и свободной энергии Гиббса ΔG°_{37}. Отработана процедура расчета гибридизационных свойств олигодезоксирибонуклеотидов с использованием метода взвешенных гистограмм. Подобраны оптимальные параметры проведения моделирования и расчета термодинамических параметров. На примере представительной выборки из 21 олигонуклеотида длиной от 4 до 16 нт и долей G/C пар от 14 до 100 % показана возможность расчета ΔH°, ΔS° и ΔG°_{37}. Ошибки расчета термодинамических параметров составляют 11.4, 12.9 и 11.8 % соответственно, а температура плавления прогнозируется со средней ошибкой 5.5 °C. Такая высокая точность расчетов сопоставима с экспериментальной и с другими прогностическими методами расчета энергии комплексообразования. В настоящей работе впервые систематически исследовано применение метода WHAM для расчета энергии формирования ДНК-дуплексов. Полученные результаты показывают потенциальную возможность достоверного расчета гибридизационных свойств новых, в том числе еще не синтезированных производных нуклеиновых кислот. Это открывает новые горизонты для рационального дизайна конструкций на основе нуклеиновых кислот для решения задач биомедицины и биотехнологии.
Ключевые слова: ДНК; гибридизация; термодинамические параметры; свободная энергия Гиббса; метод взвешенных гистограмм; WHAM; молекулярная динамика.

Для цитирования: Юшин И.И., Голышев В.М., Пышный Д.В., Ломзов А.А. Применение метода взвешенных гистограмм для расчета термодинамических параметров формирования комплексов олигодезоксирибонуклеотидов. *Вавиловский журнал генетики и селекции*. 2023;27(7):807-814. DOI 10.18699/VJGB-23-93

Application of the weighted histogram method for calculating the thermodynamic parameters of the formation of oligodeoxyribonucleotide duplexes

I.I. Yushin[1, 2], V.M. Golyshev[1, 2], D.V. Pyshnyi[1], 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@niboch.nsc.ru

Abstract. To date, many derivatives and analogs of nucleic acids (NAs) have been developed. Some of them have found uses in scientific research and biomedical applications. Their effective use is based on the data about their properties. Some of the most important physicochemical properties of oligonucleotides are thermodynamic parameters of the formation of their duplexes with DNA and RNA. These parameters can be calculated only for a few NA derivatives: locked NAs, bridged oligonucleotides, and peptide NAs. Existing predictive approaches are based on an analysis of experimental data and the consequent construction of predictive models. The ongoing pilot studies aimed at devising methods for predicting the properties of NAs by computational modeling techniques are based only on knowledge about the structure of oligonucleotides. In this work, we studied the applicability of the weighted

histogram analysis method (WHAM) in combination with umbrella sampling to the calculation of thermodynamic parameters of DNA duplex formation (changes in enthalpy ΔH°, entropy ΔS°, and Gibbs free energy ΔG°_{37}). A procedure was designed involving WHAM for calculating the hybridization properties of oligodeoxyribonucleotides. Optimal parameters for modeling and calculation of thermodynamic parameters were determined. The feasibility of calculation of ΔH°, ΔS°, and ΔG°_{37} was demonstrated using a representative sample of 21 oligonucleotides 4–16 nucleotides long with a GC content of 14–100 %. Error of the calculation of the thermodynamic parameters was 11.4, 12.9, and 11.8 % for ΔH°, ΔS°, and ΔG°_{37}, respectively, and the melting temperature was predicted with an average error of 5.5 °C. Such high accuracy of computations is comparable with the accuracy of the experimental approach and of other methods for calculating the energy of NA duplex formation. In this paper, the use of WHAM for computation of the energy of DNA duplex formation was systematically investigated for the first time. Our results show that a reliable calculation of the hybridization parameters of new NA derivatives is possible, including derivatives not yet synthesized. This work opens up new horizons for a rational design of constructs based on NAs for solving problems in biomedicine and biotechnology.

Key words: DNA; hybridization; thermodynamic parameters; Gibbs free energy; Weighted Histogram Analysis Method; WHAM; molecular dynamics.

For citation: Yushin I.I., Golyshev V.M., Pyshnyi D.V., Lomzov A.A. Application of the weighted histogram method for calculating the thermodynamic parameters of the formation of oligodeoxyribonucleotide duplexes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):807-814. DOI 10.18699/VJGB-23-93

Введение

На сегодняшний день разработан широкий спектр производных и аналогов нуклеиновых кислот. Некоторые из них нашли применение при решении научно-исследовательских задач и задач биомедицины (Wang et al., 2022). Возможность эффективного использования нуклеиновых кислот (НК) обусловлена наличием детальной информации об их физико-химических, молекулярно-биологических и биологических свойствах. Данная информация доступна лишь для ограниченного числа производных НК, таких как замкнутые НК (LNA) (McTigue et al., 2004), пептидные НК (PNA) (Griffin, Smith, 1998), фосфортиоатные производные (PS) (Eckstein, 2014), фосфорамидатные морфолиновые олигомеры (PMO) (Summerton, Weller, 1997), мостиковые олигонуклеотиды (BO) (Lomzov et al., 2006). Разработка подходов прогнозирования свойств нуклеиновых кислот, их аналогов и производных совершенно необходима для рационального дизайна олигонуклеотидных конструкций при решении всех вышеописанных задач. Наличие таких инструментов в существенной мере упростит процесс научных исследований, в которые вовлечены такие соединения, а также создания коммерческих продуктов, например систем молекулярной диагностики или терапевтических нуклеиновых кислот.

Одним из основных физико-химических свойств производных НК является их способность и эффективность формирования комплексов с комплементарными последовательностями ДНК и РНК. Разработаны модели прогностического расчета термодинамических характеристик формирования дуплексных структур ДНК (SantaLucia, Hicks, 2004) и РНК (Xia et al., 1998), гибридных ДНК/РНК (Sugimoto et al., 1995; Banerjee et al., 2020) и некоторых производных нуклеиновых кислот: LNA (McTigue et al., 2004), BO (Lomzov et al., 2006) и PNA (Griffin, Smith, 1998). Они основаны на анализе экспериментальных данных о гибридизационных свойствах этих олигомеров и построении с их использованием прогностических аналитических моделей. Кроме того, проводятся пилотные исследования, направленные на разработку методов достоверной оценки энергии формирования комплексов нуклеиновых кислот с использованием методов компьютерного моделирования. Последние перспективны с точки зрения создания подходов для априорного предсказания свойств производных нуклеиновых кислот, которые еще не были синтезированы. В недавней работе D. Dowerah с коллегами предложен ряд новых аналогов LNA с различными линкерами между O2′ и C4′ атомами (Dowerah et al., 2023). Это указывает на высокий потенциал и востребованность методов предсказания свойств модифицированных нуклеиновых кислот на основании лишь их химической структуры.

Давно зарекомендовавшим себя подходом для расчета свободной энергии Гиббса является метод анализа взвешенных гистограмм (weighted histogram analysis method, WHAM) при анализе зонтичной выборки (umbrella sampling) (Kumar et al., 1992). Общий принцип расчета заключается в проведении молекулярного моделирования методом зонтичной выборки и анализа полученных траекторий методом WHAM (рис. 1). При молекулярном моделировании методом зонтичной выборки на систему накладывается дополнительный (обычно гармонический) потенциал вдоль координаты реакции (ξ), который удерживает систему в положении ξ_i ($i = 1 \ldots i_{max}$) с определенной силой. Для каждого окна зонтичной выборки (i) получают гистограмму, которая представляет собой распределение вероятности по координате реакции, смещенной удерживающим потенциалом. Широко распространенным методом расчета потенциала средней силы по гистограммам является WHAM. В рамках данного подхода оценивают статистическую неопределенность несмещенного распределения вероятностей с учетом зонтичных гистограмм, а затем рассчитывают потенциал средней силы (PMF), который соответствует наименьшей неопределенности (Kumar et al., 1992). Это позволяет вычислить свободную энергию и другие наблюдаемые величины (Grossfield, 2018).

В настоящей работе исследована возможность расчета энергии образования совершенных ДНК-дуплексов различной длины и нуклеотидного состава с использованием WHAM при анализе зонтичной выборки. Проведение расчета свободной энергии Гиббса формирования дуплексов при различных температурах должно позволить рассчитать энтальпийный (ΔH°) и энтропийный (ΔS°) вклады. С помощью величин ΔH° и ΔS° можно рассчитать наи-

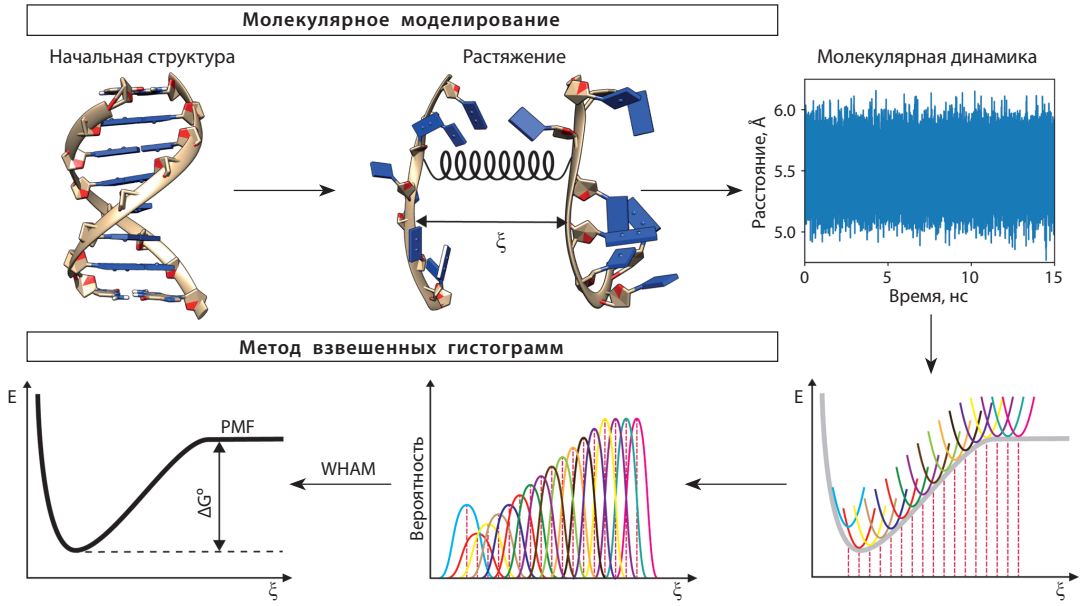


Рис. 1. Протокол проведения расчета свободной энергии Гиббса образования двойной спирали НК с использованием метода взвешенных гистограмм.

более наглядную и широко используемую характеристику для описания термической стабильности комплексов нуклеиновых кислот – температуру плавления ($T_{пл}$).

Методы

Структура ДНК-дуплексов создана в программе NAB пакета программ AmberTools18 (Case et al., 2018). Стартовые структуры имели B-форму двойной спирали.

Моделирование методом молекулярной динамики (МД) выполнено в пакете программ AMBER18 (Case et al., 2018) с параллельными вычислениями на центральных процессорах и графических ускорителях. Моделирование ДНК осуществляли в силовом поле ff99bsc0 (Pérez et al., 2007) в неявной водной оболочке (Tsui, Case, 2000) при фиксированной температуре в диапазоне от 273 до 333 К с шагом 10 градусов. Регулирование температуры проводили с помощью термостата Берендсена и масштабирования скоростей с периодом привязки 10 пс (Omelyan, Kovalenko, 2013). Для возможности использования шага интегрирования уравнений движения 2 фс применяли алгоритм SHAKE.

Процедура моделирования включала восемь шагов:

1) создание структуры ДНК-дуплекса и сохранение в формате PDB (с использованием программы NAB пакета AmberTools18). Сохранение структуры в формате файлов amber (tleap);
2) минимизация структуры в течение 10 000 шагов (pmemd.cuda);
3) ступенчатый нагрев системы от 0 до 100 К за 50 пс и от 100 К до заданной температуры (от 273 до 333 К с шагом 10 К) за 150 пс (pmemd.cuda). Шаг интегрирования 0.5 фс;
4) растяжение двух цепей от 0 до 45 Å в течение 10 нс путем наложения потенциала 10 ккал/моль на расстояние между центрами масс выбранных атомов цепей (pmemd.cuda);
5) извлечение из траектории растяжения двух цепей ДНК структур, для которых расстояние между центрами масс составляет от 0 до 45 Å (или 60) с шагом 0.5 Å (pmemd.cuda);
6) моделирование извлеченных структур методом молекулярной динамики в течение 15 нс с наложением гармонического потенциала 10 ккал/моль на расстояние между центрами масс выбранных атомов цепей (pmemd.cuda);
7) расчет энергии взаимодействия цепей методом взвешенных диаграмм в программе WHAM (Grossfield, 2018). Число точек вдоль координаты реакции для дискретизации профиля свободной энергии было выбрано равным 150 (см. ниже), критерий сходимости WHAM 10^{-6};
8) расчет энергий взаимодействия цепей методом покомпонентного расчета изменения свободной энергии на основе симуляции методом молекулярной динамики по обобщенной модели Борна (Molecular Mechanics/Generalized Born Surface Area, MMGBSA) выполнен с использованием модуля MMPBSA.py пакета AmberTools18.

Структуры молекул визуализировали в программе UCSF Chimera (Pettersen et al., 2004).

Результаты и обсуждение

Для отработки протокола моделирования был выбран набор ДНК олигомеров различной длины (от 4 до 16 п. о.) и GC-состава (от 14 до 100 %). Нуклеотидные последовательности приведены в таблице. Общий протокол моделирования и анализа представлен на рис. 1. Мы выбрали метод проведения исследований, в котором в качестве координаты реакции использовали расстояние между двумя цепями ДНК. То есть осуществляли поэтапное разнесение двух цепей в пространстве и рассчитывали потенциал средней силы (PMF) в зависимости от расстояния между ними. Этот подход с привлечением метода взвешен-

Термодинамические параметры, рассчитанные методами WHAM, MMGBSA и определенные экспериментально

Последовательность олигонуклеотидов от 5′- к 3′-концу	WHAM					MMGBSA	Эксперимент			
	ΔH°	ΔS°	ΔG°_{37}	$T_{пл}$	R^2	ΔH°	ΔH°	ΔS°	ΔG°_{37}	$T_{пл}$
AATTGGAC	–43.7	–115.6	–7.8	36.1	0.918	–77.4 ± 5.1	–56.9	–161	–6.9	31.7
ACGACCTC	–64.0	–169.1	–11.5	55.3	0.959	–85.1 ± 7.3	–59.8	–165	–8.6	40.5
AGAGCTCT	–64.3	–166.1	–12.8	62.5	0.958	–78.4 ± 8.1	–49.8	–134	–8.2	38.8
AGCATTAGACGGACCT	–166.0	–434.1	–31.3	87.8	0.960	–162.8 ± 7.9	–123.9	–335	–19.9	70.4
AGCCG	–39.0	–103.1	–7.0	29.8	0.923	–58.0 ± 5.8	–39.0	–108	–5.5	18.7
AGTTGC	–31.2	–82.0	–5.8	16.8	0.857	–65.4 ± 8.4	–37.0	–101	–5.7	19.0
ATATGGAC	–46.4	–130.6	–5.9	23.9	0.907	–77.6 ± 7.5	–53.8	–153	–6.5	28.0
CAAATAAAG	–67.9	–208.1	–3.4	17.5	0.963	–76.7 ± 8.6	–58.6	–168	–6.5	29.5
CACAG	–26.6	–71.2	–4.6	2.0	0.979	–56.6 ± 6.5	–33.7	–97	–3.6	1.7
CCGCGG	–60.8	–158.3	–11.7	57.2	0.932	–83.9 ± 7.6	–41.4	–106	–8.4	41.4
CGCG	–27.9	–68.3	–6.7	23.7	0.892		–36.3	–103	–4.5	9.1
CGCGCG	–45.5	–113.8	–10.2	52.9	0.905	–79.6 ± 6.7	–46.4	–121	–8.7	43.3
GCACCGAC	–92.3	–249.5	–14.9	62.2	0.986	–87.9 ± 7.6	–71.0	–196	–10.2	47.2
GCATGC	–58.8	–160.4	–9.1	43.1	0.948	–69.5 ± 7.9	–42.2	–117	–6.0	22.7
GCCCGGAC	–69.1	–182.8	–12.4	58.4	0.949	–94.9 ± 6.4	–61.4	–165	–10.3	48.9
GCCTGC	–48.3	–126.6	–9.0	44.0	0.915	–73.6 ± 8.6	–37.5	–100	–6.5	25.3
TACTGGAC	–62.7	–168.9	–10.3	48.9	0.934	–81.0 ± 7.7	–58.5	–165	–7.2	33.7
TCTATGCA	–44.3	–109.8	–10.3	54.4	0.813	–79.5 ± 6.5	–51.7	–145	–6.6	29.8
TGCGCA	–61.2	–162.3	–10.8	52.3	0.980	–76.9 ± 6.9	–42.5	–114	–7.3	31.2
TGTTGC	–41.1	–112.8	–6.1	23.6	0.979	–65.8 ± 7.9	–37.2	–101	–5.8	20.6
ACATTATTATTACA	–148.7	–442.7	–11.4	44.3	0.947	–125.4 ± 13.6	–89.9	–254	–11.1	48.3

Примечание. Размерность величин термодинамических параметров ΔH° и ΔG°_{37} – ккал/моль, ΔS° – кал/моль/К, $T_{пл}$ – °С. R^2 – коэффициент корреляции Пирсона для линейной зависимости ΔG°(*T*). Ошибки экспериментальных значений ΔH°, ΔS° и ΔG°_{37} и $T_{пл}$ составляют 10, 10 и 8 % и 0.5 °С соответственно.

ных гистограмм (WHAM) позволяет определить энергию Гиббса взаимодействия двух цепей непосредственно из компьютерного эксперимента. Проводя такой *in silico* эксперимент при разной температуре, можно вычислить изменение энтальпии и энтропии комплексообразования из линейной температурной зависимости свободной энергии Гиббса. На первом этапе необходимо было определить параметры, оптимальные для проведения расчетов.

В качестве координаты реакции (ξ) мы выбрали расстояние между центрами масс (*r*) C4′ атомов всех нуклеотидов каждой из двух цепей. Начальное расстояние было выбрано равным 0 Å, чтобы рассмотреть возможность «сжатия» двойной спирали и при анализе определить наличие минимума зависимости $\Delta G^\circ_T(r)$, а энергию комплексообразования рассчитывать как разницу между минимумом и максимумом данной зависимости (см. рис. 1). Анализ разнесения центров масс цепей показал, что максимального расстояния в 45 Å достаточно для полной диссоциации комплексов олигонуклеотидов длиной от 4 до 9 п. о., а для комплексов размером 14 и 16 п. о. оно должно быть увеличено до 60 Å. При выбранной таким образом координате реакции диссоциация двух цепей для большинства комплексов олигонуклеотидов происходит в соответствии с моделью «застежка-молния» (Cantor, Schimmel, 1980; Volkov, Solov'yov, 2009), с расплетанием двойной спирали с одного из концов, либо в смешанном режиме «сдвиг/застежка-молния» (Mosayebi et al., 2015; Kurus, Dultsev, 2018). Пример изменения конформаций олигомеров вдоль координаты реакции приведен в Приложении 1[1]. Механизм диссоциации комплексов в проводимом исследовании не является критически важным ввиду того, что рассматриваются только два предельных состояния: релаксированная структура дуплекса и два одноцепочечных невзаимодействующих олигонуклеотида. Соответствие механизма перехода спираль–клубок наблюдаемым экспериментальными методами подтверждает адекватность выбранного подхода для описания диссоциации двойной спирали ДНК.

Согласно общепринятым требованиям для проведения WHAM анализа, необходимо, чтобы перекрывание между гистограммами было не менее 20 %. Анализ показал, что это достигается при ~0.7 Å между соседними окнами моделирования. Пример зависимости гистограмм распределения от расстояния между цепями для дуплекса

[1] Приложения 1–8 см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx26.pdf

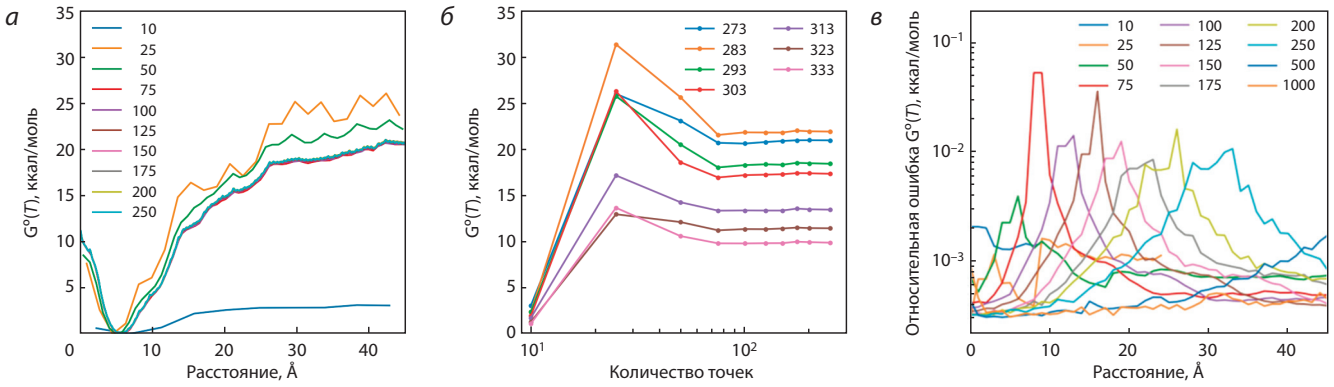


Рис. 2. Поиск параметров моделирования и анализа MD траекторий.

а – зависимость профиля свободной энергии Гиббса от расстояния между центрами масс C4′ атомов углерода двух цепей ДНК при различном числе точек вдоль координаты реакции, выбранных для дискретизации профиля свободной энергии при 273 K; *б* – зависимость энергии Гиббса формирования комплекса от числа точек вдоль координаты реакции, выбранных для дискретизации профиля свободной энергии; *в* – зависимость относительной ошибки расчета свободной энергии Гиббса от расстояния между центрами масс C4′ атомов углерода двух цепей ДНК при различном числе точек вдоль координаты реакции, выбранных для дискретизации профиля свободной энергии.

5′-GCACCGAC-3′/5′-GTCGGTGC-3′ приведен в Приложении 2. Мы выбрали шаг координаты реакции равным 0.5 Å, чтобы заведомо соответствовать данному критерию.

При расчете энергии методом WHAM важным параметром является число точек (bins) вдоль координаты реакции, выбранных для дискретизации профиля свободной энергии. При числе точек 100 и более (до 1000 разбиений) достигается плато как для формы и положения профилей свободной энергии Гиббса (рис. 2, *а*, Приложение 3), так и для величин изменения свободной энергии Гиббса при различных температурах (см. рис. 2, *б*). При этом величины относительных значений ошибок, рассчитанных методом бутстреп (Grossfield, 2018), не превышают 6 % (см. рис. 2, *в*).

Свободную энергию Гиббса при определенной температуре рассчитывали как разницу между минимумом и максимумом на профиле потенциала средней силы: $\Delta G°(T) = PMF_{min} - PMF_{max}$. Для получения температурных зависимостей свободной энергии Гиббса был выбран диапазон от 273 до 333 K с шагом 10 K. Нижнее значение выбрано в соответствии с температурой замерзания воды, а верхнее – чтобы ограничить денатурацию цепей НК при моделировании в течение заданного диапазона времени для коротких олигомеров. Такой диапазон является достаточно широким, чтобы можно было построить зависимость $\Delta G°(T)$ для достоверного определения $\Delta H°$ и $\Delta S°$ методом линейной регрессии с использованием уравнения $\Delta G°(T) = \Delta H° - T\Delta S°$.

Длина траектории при моделировании методом молекулярной динамики при каждом фиксированном расстоянии между выбранными центрами масс и заданной температуре была выбрана равной 15 нс, чтобы получить минимально достоверную траекторию в неявной водной оболочке (Lomzov et al., 2015). Таким образом, для каждого дуплекса были получены траектории длиной 15 нс при 90 (или 120) различных расстояниях между центрами масс при семи значениях температуры. Длина траектории для каждого дуплекса варьировала от 9.45 до 12.6 мкс. Полная продолжительность траекторий для всех комплексов составила более 200 мкс.

Проведен расчет методом WHAM зависимости энергии Гиббса взаимодействия двух олигонуклеотидов от расстояния между центрами масс C4′ атомов каждой из цепей (r) при семи исследованных температурах для 21 исследованного комплекса (см. таблицу). Типичная зависимость энергий Гиббса комплексообразования от r при температурах от 273 до 333 K для комплекса 5′-GCACCGAC-3′/5′-GTCGGTGC-3′ приведена на рис. 3, *а*. Зависимость свободной энергии Гиббса имеет четко выраженный минимум в районе 6 Å и возрастает при сжатии и растяжении двойной спирали. При растяжении график зависимости проходит через максимум и незначительно снижается. Точка максимума соответствует расстоянию, на котором прекращается взаимодействие между цепями.

Для оценки адекватности моделирования мы провели сравнение геометрии двойной спирали релаксированной формы ДНК-дуплекса 5′-GCACCGAC-3′/5′-GTCGGTGC-3′ с литературными данными (Приложение 4). Все структурные параметры хорошо согласуются с данными для структуры додекамера Дрю-Диккерсона (DDD, 5′-CGCGAATTCGCG-3′), определенной экспериментально методами ЯМР спектроскопии (PDB ID: 1NAJ) и рентгеновской кристаллографии (PDB ID: 1BNA).

Проведен расчет свободной энергии Гиббса комплексообразования при различных температурах ($\Delta G°(T)$). Установлено, что данная зависимость имеет линейный характер с высоким коэффициентом корреляции R^2 (более 0.83); среднее значение для всех исследованных комплексов составило 0.93 (см. рис. 3, *б* и таблицу). На основании полученных зависимостей (Приложение 5) рассчитаны величины изменения энтальпии и энтропии комплексообразования (см. таблицу). Сопоставление рассчитанных методом WHAM величин термодинамических параметров с величинами, определенными экспериментально (по данным (Lomzov et al., 2015)), показывает линейную связь между ними с высокими коэффициентами корреляции R^2: 0.87, 0.82, 0.88 и 0.75 для $\Delta H°$, $\Delta S°$, $\Delta G°_{37}$ и $T_{пл}$ соответственно (Приложение 6). В качестве температуры плавления выбирали такое значение, при котором половина олигонуклеотидов находится в двуцепочечном,

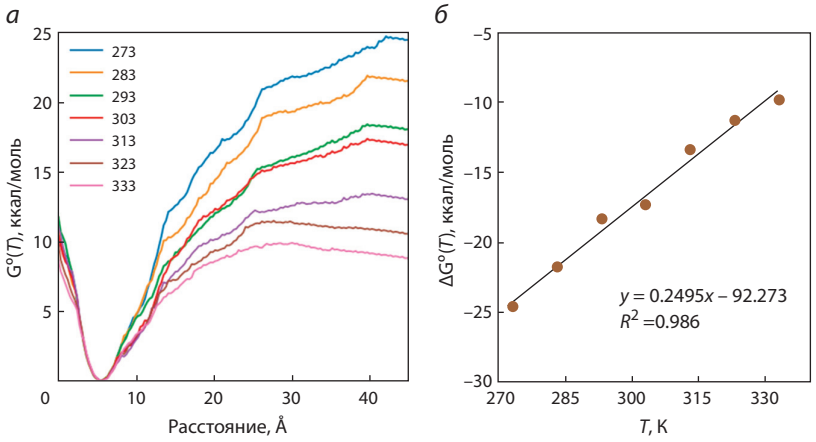


Рис. 3. Зависимость свободной энергии Гиббса: *а* – от расстояния между молекулами при различных температурах (273, 283, 293, 303, 313, 323, 333 K); *б* – от температуры модельного дуплекса 5′-GCACCGAC-3′/5′-GTCGGTGC-3′.

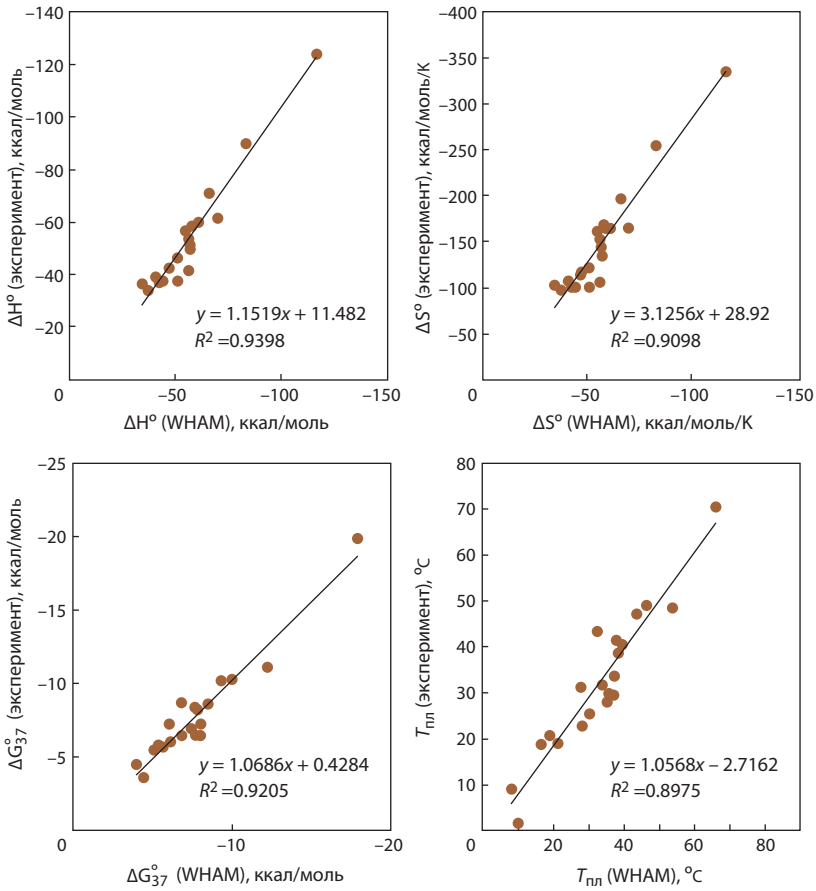


Рис. 4. Корреляция термодинамических параметров ΔH°, ΔS° и ΔG°_{37} и температуры плавления комплексов, рассчитанных методом WHAM с учетом линейных поправок, с величинами, определенными экспериментально (по данным (Lomzov et al., 2015)).

а вторая – в одноцепочечном состоянии. Величина $T_{пл}$ была рассчитана с использованием термодинамических параметров (Lomzov, Pyshnyi, 2012):

$$T_{пл} = \Delta H^{\circ}/(\Delta S^{\circ} + R \ln\left[\frac{Ct}{4}\right]),$$

где R – универсальная газовая постоянная, Ct – полная концентрация олигонуклеотидов в системе. Значение Ct было взято равным 10 мкМ в соответствии с типичными экспериментальными значениями.

Наклон линейной зависимости термодинамических параметров близок к 0.5, а величины свободных членов значительны по сравнению с анализируемыми величинами (см. Приложение 6). Поэтому, как предложено в наших предыдущих работах (Lomzov et al., 2015; Golyshev et al., 2021), можно ввести линейные поправки на значения рассчитанных термодинамических параметров ΔH° и ΔS°. После применения такой коррекции коэффициенты корреляции для свободной энергии Гиббса и температур плавления существенно улучшаются – до 0.94 и 0.86 соответственно (рис. 4). При этом средние абсолютные значения ошибок расчета термодинамических параметров становятся равными 11.4, 12.9 и 11.8 % и 5.5 °C для ΔH°, ΔS° и ΔG°_{37} и $T_{пл}$ соответственно. Для данной выборки аналогичные значения ошибок величин термодинамических характеристик, полученных методом MMGBSA в работах (Lomzov et al., 2015; Golyshev et al., 2021) с учетом линейных поправок, несколько ниже и составляют 7.6, 11.4 и 10.6 % и 4.3 °C. Точность расчета термодинамических параметров в настоящей работе сопоставима с экспериментальной и с точностью при использовании метода ближайших соседей – наиболее распространенного метода расчета эффективности комплексообразования олигонуклеотидов, которая составляет около 10 % для энтальпии и энтропии и около 8 % для свободной энергии Гиббса комплексообразования (SantaLucia, Hicks, 2004; Lomzov et al., 2006).

Для дополнительной проверки качества результатов траектории проанализировали методом MMGBSA и сопоставили рассчитанные величины с данными, рассчитанными методом WHAM, и с ранее полученными методом MMGBSA величинами (Lomzov et al., 2015). Типичный вид зависимости энергии MMGBSA от расстояния между центрами масс C4′ атомов цепей схож с зависимостью свободной энергии Гиббса от расстояния, приведенной на рис. 3, *а* (Приложение 7). При расстояниях, близких к максимальному, значение энергии формирования двойной спирали ДНК выходит на плато, равное нулю, что свидетельствует об отсутствии взаимодействия между цепями при данном способе анализа траекторий. Наблюдается дно потенциальной ямы в области 2–7 Å, что соответствует релаксированной форме двойной спирали ДНК, а ее глобаль-

ный минимум в районе 7 Å близок к минимуму, наблюдающемуся в зависимости энергии Гиббса, определенной в рамках WHAM (см. рис. 3, *а* и Приложение 7). Присутствует слабая зависимость энергии комплексообразования, рассчитанной методом MMGBSA, от температуры, что соответствует малому значению изменения теплоемкости ΔCp. Достоверно определить величину изменения теплоемкости из компьютерных экспериментов не представляется возможным ввиду больших ошибок расчета.

Величины энтальпии комплексообразования, рассчитанные в настоящей работе и определенные ранее, хорошо коррелируют (R^2 = 0.97) с наклоном, близким к единице (0.95), и свободным членом линейной зависимости, близким к нулю (4 ккал/моль) (Приложение 8, *а*). Кроме того, аналогичная линейная корреляция наблюдается для величин энтальпии комплексообразования, рассчитанных методами MMGBSA и WHAM в данной работе. Таким образом, полученные нами МД траектории релевантны.

Одним из важных аспектов в исследованном ранее подходе расчета энергии методом MMGBSA является неопределенность, связанная со структурой одноцепочечного состояния олигонуклеотидов. Ее извлекали из траектории двойной спирали. Тем не менее это позволило вычислить энтальпию комплексообразования с достаточно хорошей точностью. В данной работе при анализе методом взвешенных гистограмм одноцепочечное состояние олигонуклеотидов неплохо представлено в МД траекториях (насколько это возможно сделать в рамках приближения неявной водной оболочки и выбранного силового поля). Такой подход дает хорошие результаты при расчете энергии формирования двойной спирали. Вместе с тем основным преимуществом использования метода WHAM является расчет непосредственно изменения свободной энергии Гиббса формирования двойной спирали ДНК. Тот факт, что эта характеристика оказалась в наших расчетах линейной в широком диапазоне исследованных температур, указывает на то, что выбранные параметры моделирования, а также заложенные в моделирование и анализ модели достаточно хорошо описывают физику как дву-, так и одноцепочечной ДНК. Для последней в пользу этого утверждения свидетельствует конформация олигонуклеотидов при моделировании цепей с большим расстоянием между их центрами масс (см. Приложение 1). Олигомеры не остаются линейными, как в составе дуплекса, что использовали при анализе методом MMGBSA, и не становятся полностью разупорядоченными цепочками, а сохраняют несколько гетероциклических оснований подряд в стэкинге. Это коррелирует с персистентной длиной одноцепочечных олигонуклеотидов, которая составляет несколько нуклеотидов (в зависимости от нуклеотидной композиции и ионной силы раствора) (Chen et al., 2012). Кроме того, наблюдающаяся линейная зависимость $\Delta G°(T)$ в исследованном нами методе позволяет напрямую вычислять энтропию комплексообразования.

Вместе с тем разработанный подход далек от совершенства. В частности, для более достоверного моделирования структуры и динамики ДНК необходимо применять наиболее современные силовые поля и модель явной водной оболочки. Анализ параметров силового поля для такого моделирования является отдельной сложной задачей. Кроме того, сложность и продолжительность расчетов в явной водной оболочке кратно увеличиваются. Так, основные вычислительные затраты происходят на этапе молекулярно-динамического расчета. Для девятизвенного ДНК-дуплекса, детально проанализированного в данной работе, скорость расчета в неявной водной оболочке при использовании современной видеокарты NVIDIA GTX3080 составляет около 800 нс/день. Таким образом, время расчета одного модельного дуплекса составляет 12 суток. В случае явной водной оболочки меделируемая периодическая ячейка будет содержать около 15–20 тыс. молекул ввиду максимального расстояния между цепями 45 Å. Это снизит производительность до ~100 нс/день или составит около 3 месяцев. В явной водной оболочке существенно снизится конформационная подвижность ДНК, что потребует увеличения длины траектории для каждого окна моделирования и, соответственно, увеличения вычислительных затрат. Тем не менее такое усложнение представляется необходимым для увеличения достоверности расчетов.

Еще одним перспективным направлением развития исследованного подхода является его апробация на примерах известных модифицированных нуклеиновых кислот. Это должно ответить на вопрос о его применимости для рационального дизайна химической структуры новых, еще не синтезированных химически производных нуклеиновых кислот, что может быть использовано для решения конкретных задач биомедицины и биотехнологии. Проведенный анализ показывает высокий потенциал и реалистичность применения метода WHAM для расчета энергии формирования комплексов нуклеиновых кислот, их аналогов и производных.

Заключение

Отработана процедура метода взвешенных гистограмм для расчета гибридизационных свойств олигодезоксирибонуклеотидов. Подобраны оптимальные параметры для проведения моделирования и расчета термодинамических параметров формирования ДНК-дуплексов. На примере представительной выборки, содержащей 21 олигонуклеотид длиной от 4 до 16 нт и долей G/C пар от 14 до 100 %, показана возможность расчета энтальпии, энтропии и свободной энергии Гиббса формирования комплексов олигонуклеотидов методом взвешенных гистограмм (WHAM) при анализе МД траекторий с использованием в качестве координаты реакции расстояния между центрами масс C4′ атомов углерода каждой из цепей. Установлена линейная зависимость свободной энергии Гиббса от температуры, при которой проводится моделирование. Это позволяет проводить расчет энтальпии и энтропии комплексообразования из анализа результатов, полученных методом взвешенных гистограмм. Рассчитываемые термодинамические параметры линейно коррелируют с экспериментально определенными величинами с высоким коэффициентом корреляции R^2 (более 0.83). С учетом линейной поправки данной зависимости величины ошибок расчета значений термодинамических параметров сопоставимы с экспериментальными и составляют 11.4, 12.9

и 11.8 % для ΔH°, ΔS° и ΔG°_{37}, а температура плавления прогнозируется со средней ошибкой 5.5 °C. Таким образом, впервые систематически исследовано применение метода WHAM для расчета энергии формирования ДНК-дуплексов. Установлена высокая точность таких расчетов, сопоставимая с экспериментальной и другими методами расчета энергии комплексообразования.

Список литературы / References

Banerjee D., Tateishi-Karimata H., Ohyama T., Ghosh S., Endoh T., Takahashi S., Sugimoto N. Improved nearest-neighbor parameters for the stability of RNA/DNA hybrids under a physiological condition. *Nucleic Acids Res.* 2020;48(21):12042-12054. DOI 10.1093/nar/gkaa572

Cantor C.R., Schimmel P.R. Biophysical Chemistry. Part I: The Conformation of Biological Macromolecules. New York: W.H. Freeman & Company, 1980

Case D.A., Walker R.C., Cheatham T.E., Simmerling C., Roitberg A., Merz K.M., Luo R., Darden T. Amber 18. Reference Manual. San Francisco: Univ. of California, 2018

Chen H., Meisburger S.P., Pabit S.A., Sutton J.L., Webb W.W., Pollack L. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. USA.* 2012;109(3):799-804. DOI 10.1073/pnas.1119057109

Dowerah D., Uppuladinne M.V.N., Sarma P.J., Biswakarma N., Sonavane U.B., Joshi R.R., Ray S.K., Namsa N.D., Deka R.C. Design of LNA analogues using a combined density functional theory and molecular dynamics approach for RNA therapeutics. *ACS Omega*. 2023;8(25):22382-22405. DOI 10.1021/acsomega.2c07860

Eckstein F. Phosphorothioates, essential components of therapeutic oligonucleotides. *Nucleic Acid Ther.* 2014;24(6):374-387. DOI 10.1089/nat.2014.0506

Golyshev V.M., Pyshnyi D.V., Lomzov A.A. Calculation of energy for RNA/RNA and DNA/RNA duplex formation by molecular dynamics simulation. *Mol. Biol.* 2021;55(6):927-940. DOI 10.1134/S002689332105006X

Griffin T.J., Smith L.M. An approach to predicting the stabilities of peptide nucleic acid:DNA duplexes. *Anal. Biochem.* 1998;260(1):56-63. DOI 10.1006/abio.1998.2686

Grossfield A. WHAM: the weighted histogram analysis method. 2018.

Kumar S., Rosenberg J.M., Bouzida D., Swendsen R.H., Kollman P.A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 1992;13(8):1011-1021. DOI 10.1002/jcc.540130812

Kurus N.N., Dultsev F.N. Determination of the thermodynamic parameters of DNA double helix unwinding with the help of mechanical methods. *ACS Omega*. 2018;3(3):2793-2797. DOI 10.1021/acsomega.7b01815

Lomzov A.A., Pyshnyi D.V. Considering the oligonucleotide secondary structures in thermodynamic and kinetic analysis of DNA duplex formation. *Biophysics* (*Oxf*). 2012;57(1):19-34. DOI 10.1134/S0006350912010137

Lomzov A.A., Pyshnaya I.A., Ivanova E.M., Pyshnyi D.V. Thermodynamic parameters for calculating the stability of complexes of bridged oligonucleotides. *Dokl. Biochem. Biophys.* 2006;409(1):211-215. DOI 10.1134/S1607672906040053

Lomzov A.A., Vorobjev Y.N., Pyshnyi D.V. Evaluation of the Gibbs free energy changes and melting temperatures of DNA/DNA duplexes using hybridization enthalpy calculated by molecular dynamics simulation. *J. Phys. Chem. B*. 2015;119(49):15221-15234. DOI 10.1021/acs.jpcb.5b09645

McTigue P.M., Peterson R.J., Kahn J.D. Sequence-dependent thermodynamic parameters for locked nucleic acid (LNA)-DNA duplex formation. *Biochemistry*. 2004;43(18):5388-5405. DOI 10.1021/bi035976d

Mosayebi M., Louis A.A., Doye J.P.K., Ouldridge T.E. Force-induced rupture of a DNA duplex: from fundamentals to force sensors. *ACS Nano*. 2015;9(12):11993-12003. DOI 10.1021/acsnano.5b04726

Omelyan I., Kovalenko A. Generalised canonical-isokinetic ensemble: speeding up multiscale molecular dynamics and coupling with 3D molecular theory of solvation. *Mol. Simul.* 2013;39(1):25-48. DOI 10.1080/08927022.2012.700486

Pérez A., Marchán I., Svozil D., Sponer J., Cheatham T.E., Laughton C.A., Orozco M. Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* 2007;92(11):3817-3829. DOI 10.1529/biophysj.106.097782

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

SantaLucia J., Hicks D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* 2004;33(1):415-440. DOI 10.1146/annurev.biophys.32.110601.141800

Sugimoto N., Nakano S., Katoh M., Matsumura A., Nakamuta H., Ohmichi T., Yoneyama M., Sasaki M. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*. 1995;34(35):11211-11216. DOI 10.1021/bi00035a029

Summerton J., Weller D. Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev.* 1997;7(3):187-195. DOI 10.1089/oli.1.1997.7.187

Tsui V., Case D.A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers*. 2000;56(4):275-291. DOI 10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E

Volkov S.N., Solov'yov A.V. The mechanism of DNA mechanical unzipping. *Eur. Phys. J. D*. 2009;54(3):657-666. DOI 10.1140/epjd/e2009-00194-5

Wang F., Li P., Chu H.C., Lo P.K. Nucleic acids and their analogues for biomedical applications. *Biosensors*. 2022;12(2):93. DOI 10.3390/bios12020093

Xia T., SantaLucia J., Burkard M.E., Kierzek R., Schroeder S.J., Jiao X., Cox C., Turner D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry*. 1998;37(42):14719-14735. DOI 10.1021/bi9809425

ORCID ID

I.I. Yushin orcid.org/0000-0001-5954-641X
V.M. Golyshev orcid.org/0000-0002-0521-6228
D.V. Pyshnyi orcid.org/0000-0002-2587-3719
A.A. Lomzov orcid.org/0000-0003-3889-9464

Благодарности. Исследование поддержано в рамках государственного задания ИХБФМ СО РАН № 121031300042-1.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 15.07.2023. После доработки 20.09.2023. Принята к публикации 21.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):815-819
DOI 10.18699/VJGB-23-94

Перевод на английский язык https://vavilov.elpub.ru/jour

Внутриопухолевая гетерогенность: модели возникновения и эволюции злокачественных опухолей

Р.А. Иванов[1] ✉, С.А. Лашин[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
✉ ivanovromanart@bionet.nsc.ru

Аннотация. Рак – сложное и гетерогенное заболевание, характеризующееся накоплением генетических изменений, которые приводят к неконтролируемому росту и пролиферации клеток. Эволюционная динамика играет решающую роль в возникновении и развитии раковых опухолей, формируя гетерогенность и адаптивность раковых клеток. С точки зрения теории эволюции опухоли представляют собой сложные экосистемы, которые развиваются в процессе микроэволюции под воздействием генетических мутаций, эпигенетических изменений и факторов микроокружения опухолей. Такая динамичная природа опухолей создает значительные проблемы для эффективного лечения рака, и ее понимание необходимо для разработки эффективных и персонализированных методов лечения. Раскрывая механизмы, определяющие гетерогенность опухоли, исследователи могут выявить ключевые генетические и эпигенетические изменения, которые способствуют прогрессированию опухоли и устойчивости к лечению. Эти знания позволяют разрабатывать инновационные стратегии воздействия на конкретные клоны опухоли, минимизируя риск рецидива и улучшая результаты лечения пациентов. Для изучения эволюционной динамики рака ученые используют широкий спектр экспериментальных и вычислительных подходов. Традиционные экспериментальные методы включают в себя геномное профилирование, такое как секвенирование нового поколения и флуоресцентная гибридизация *in situ*, и позволяют выявлять соматические мутации, изменения числа копий генов и структурные перестройки в геномах раковых опухолей. Помимо того, методы одноклеточного секвенирования стали мощным инструментом для изучения внутриопухолевой гетерогенности и отслеживания клональной эволюции. На основании экспериментальных данных разрабатываются вычислительные модели и алгоритмы для моделирования и анализа эволюции рака. Эти модели объединяют данные из различных источников для предсказания закономерностей роста опухоли, выявления драйверных мутаций и построения эволюционных деревьев развития раковых клеток. В настоящей работе мы поставили задачу описать существующие на сегодняшний день подходы к изучению эволюционной динамики развития рака и теории ее возникновения.
Ключевые слова: злокачественные опухоли; эволюция; гетерогенность.

Для цитирования: Иванов Р.А., Лашин С.А. Внутриопухолевая гетерогенность: модели возникновения и эволюции злокачественных опухолей. *Вавиловский журнал генетики и селекции*. 2023;27(7):815-819. DOI 10.18699/VJGB-23-94

Intratumor heterogeneity: models of malignancy emergence and evolution

R.A. Ivanov[1] ✉, S.A. Lashin[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
✉ ivanovromanart@bionet.nsc.ru

Abstract. Cancer is a complex and heterogeneous disease characterized by the accumulation of genetic alterations that drive uncontrolled cell growth and proliferation. Evolutionary dynamics plays a crucial role in the emergence and development of tumors, shaping the heterogeneity and adaptability of cancer cells. From the perspective of evolutionary theory, tumors are complex ecosystems that evolve through a process of microevolution influenced by genetic mutations, epigenetic changes, tumor microenvironment factors, and therapy-induced changes. This dynamic nature of tumors poses significant challenges for effective cancer treatment, and understanding it is essential for developing effective and personalized therapies. By uncovering the mechanisms that determine tumor heterogeneity, researchers can identify key genetic and epigenetic changes that contribute to tumor progression and resistance to treatment. This knowledge enables the development of innovative strategies for targeting specific tumor clones, minimizing the risk of recurrence and improving patient outcomes. To investigate the evolutionary dynamics of cancer, researchers employ a wide range of experimental and computational approaches. Traditional experimental methods involve genomic profiling techniques such as next-generation sequencing and fluorescence *in situ* hybridization. These techniques enable the identification of somatic mutations, copy number alterations, and structural rearrangements within cancer genomes. Furthermore, single-cell sequencing methods have emerged as powerful tools for dissecting intratumoral heteroge-

neity and tracing clonal evolution. In parallel, computational models and algorithms have been developed to simulate and analyze cancer evolution. These models integrate data from multiple sources to predict tumor growth patterns, identify driver mutations, and infer evolutionary trajectories. In this paper, we set out to describe the current approaches to address this evolutionary complexity and theories of its occurrence.
Key words: cancer; evolution; heterogeneity.

For citation: Ivanov R.A., Lashin S.A. Intratumor heterogeneity: models of malignancy emergence and evolution. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):815-819. DOI 10.18699/VJGB-23-94

Модели эволюции злокачественных опухолей

Рак – сложное заболевание, возникающее в результате накопления генетических и эпигенетических изменений в нормальных клетках, что приводит к неконтролируемому росту клеток и образованию опухолей. В последние несколько десятилетий становится все более очевидным, что опухоли – это не статичные образования, а скорее динамические системы, которые подвергаются непрерывной эволюции (Nowell, 1976; Merlo et al., 2006; Besse et al., 2018; Hausser, Alon, 2020; Vendramin et al., 2021). Этот эволюционный процесс формирует гетерогенность и адаптивность раковых клеток, создавая значительные проблемы для эффективного лечения рака. Под гетерогенностью опухолей понимают наличие различных типов клеток в опухоли, которые принято называть клонами. В контексте онкологии и эволюционной биомедицины клональной популяцией считается группа раковых клеток, имеющих общее происхождение и обладающих схожими генетическими изменениями. По мере деления и накопления дополнительных мутаций такие клетки формируют отдельные субпопуляции клонов в опухоли. Гетерогенность может проявляться разными способами, например в различии в морфологии клеток (Meacham, Morrison, 2013; Robertson-Tessi et al., 2015; Haffner et al., 2021), различной экспрессии генов отдельных клонов (Lüönd et al., 2021; Zhao et al., 2022) или их функциональных характеристиках.

Клональные популяции в раке принято рассматривать как различные виды в контексте эволюционной биологии (Vendramin et al., 2021). Подобно тому, как разные виды эволюционируют и адаптируются к окружающей среде с течением времени, клональные популяции в опухоли эволюционируют и адаптируются к своему микроокружению. Генетические изменения, возникающие в этих популяциях, дают преимущества или недостатки с точки зрения роста, выживаемости и ответа на терапию, что приводит к отбору и доминированию определенных клонов в опухоли.

Гетерогенность опухоли представляет собой серьезную проблему в ее лечении, поскольку может способствовать устойчивости к терапии, восстановлению опухоли после операции и развитию метастазов (Morris et al., 2016). В настоящее время существует несколько теорий о механизмах возникновения гетерогенности в опухолях.

Теория клональной эволюции является одной из самых ранних и наиболее широко признанных теорий, объясняющих возникновение гетерогенности рака. Согласно этой теории, опухоли возникают из одной или нескольких трансформированных клеток, потомки которых со временем приобретают дополнительные генетические мутации. Эти мутации способствуют образованию отдельных клонов с уникальными фенотипическими характеристиками. По мере роста опухоли происходит отбор клонов с преимущественными признаками, что приводит, в зависимости от типа рака, к расширению и доминированию этих клонов в опухолевой популяции или их совместному существованию в опухоли.

В концепции клональной эволюции выделяется несколько моделей: линейная, ветвящаяся и прерывистая. В *линейной модели* мутации приобретаются в линейной прогрессии, ведущей к более злокачественным стадиям рака (Fearon, Vogelstein, 1990). Новые драйверные мутации обеспечивают настолько сильное селективное преимущество, что превосходят все предыдущие клоны благодаря селективной чистке, которая происходит во время эволюции опухолей. В *модели ветвящейся эволюции* клоны расходятся от общего предка и развиваются параллельно в ткани опухоли, в результате чего возникает несколько клональных линий (Gawad et al., 2014; Vosberg, Greif, 2019). В отличие от линейной эволюции, в ветвящейся модели селективная зачистка встречается редко, и несколько популяций клонов развиваются одновременно, поскольку все они обладают повышенной приспособленностью. В этой модели значение внутриопухолевой гетерогенности будет колебаться во время прогрессирования опухоли, но ожидается, что на любой момент взятия пробы опухоли будет присутствовать множество клонов.

Модель нейтральной эволюции оспаривает традиционное мнение о том, что все генетические изменения при раке дают селективное преимущество. Согласно данной теории, большинство генетических мутаций в раке являются нейтральными или почти нейтральными, т. е. не оказывают существенного влияния на жизнеспособность опухоли (Williams et al., 2016; Furukawa, Kikuchi, 2020). Возникновение же гетерогенности обусловлено случайным генетическим дрейфом, когда нейтральные мутации случайно накапливаются в различных клонах. Со временем нейтральные мутации могут закрепиться внутри клонов, что вызывает наблюдаемую внутриопухолевую гетерогенность.

Следует отметить, что модель нейтральной эволюции сочетается с другой популярной теорией накопления мутаций – *прерывистой эволюцией*, упомянутой выше. По этой гипотезе раковые клетки представляют собой «многообещающих монстров» Гольдшмидта (Graham, Sottoriva, 2017), в которых постепенные и не проявляющиеся изменения в геноме приводят к резким изменениям в фенотипе. Подобный принцип проявляется, в частности, в новообразованиях – поскольку нет явных промежуточных этапов между здоровой тканью и первичными опухолями. Промежутки между скачками, вероятнее всего, и представляют собой этапы нейтральной эволюции. Согласно

теории прерывистой эволюции, сами популяции могут находиться в некоем равновесии друг с другом, поддерживая существование нескольких популяций клональных линий раковых клеток в опухоли, после чего одна из популяций становится «многообещающим монстром». В случае мутации, повышающей приспособленность, эти клоны занимают бóльшую часть опухоли, вытесняя менее приспособленные и увеличивая размер самой опухоли.

В ряде работ показано, что развитие отдельной опухоли не обязательно соответствует одной модели клональной эволюции и может менять их в течение своего развития. Предположительно, на ранних стадиях опухоли развиваются по модели линейной эволюции, а после того, как опухоль начинает активно увеличиваться, она переключается на ветвящуюся модель (Durrett et al., 2011; Vosberg, Greif, 2019). Более того, в нескольких исследованиях обнаружено, что эволюция опухоли может идти одновременно как по ветвящейся, так и по прерывистой модели – когда клоны с изменением количества копий генов идут по прерывистой модели, а клоны с точечными мутациями – по ветвящейся (Baca et al., 2013; Wang et al., 2014).

Другой распространенной теорией возникновения гетерогенности является **теория раковых стволовых клеток**, которая предполагает, что опухоли иерархически организованы, а небольшая популяция раковых стволовых клеток (РСК) определяет рост и гетерогенность опухоли (Reya et al., 2001; Lee et al., 2022). Раковые стволовые клетки обладают способностью к самообновлению и дифференцировке, подобно нормальным стволовым клеткам. Эти клетки способны генерировать как другие РСК, так и потомков, не являющихся РСК, что в теории способствует клеточному разнообразию, наблюдаемому в опухолях. Важная особенность этой теории заключается в наличии иерархичности раковых клеток: обычные раковые клетки не способны к дифференциации, и соматические мутации в них несут менее значимый клинический эффект из-за более низкой способности к размножению, тогда как основное патологическое значение имеют РСК с разной степенью плюрипотентности. Возникновение гетерогенности в этой модели объясняется асимметричным делением РСК, которое может привести к появлению различных клонов РСК с разными фенотипическими свойствами. Стоит отметить, что на сегодняшний день РСК были найдены лишь в ограниченном количестве типов опухолей, в частности в гематологических опухолях (Bonnet, Dick, 1997; Zarzynska, 2017; Hata et al., 2018; Lee et al., 2022), но в этих случаях они могут быть важным фактором в рецидивах злокачественных опухолей после лечения (Walcher et al., 2020).

Теория отбора по микроокружению предполагает, что важную роль в формировании гетерогенности опухоли играет ее микроокружение. Взаимодействие между раковыми клетками и их микроокружением, включающим в себя иммунные клетки, стромальные клетки и компоненты внеклеточного матрикса, может оказывать селективное давление на опухолевые клетки (Augustin et al., 2020). Факторы микроокружения, такие как гипоксия, воспаление и доступность питательных веществ, могут влиять на рост опухоли, ангиогенез и метастазирование (Mumenthaler et al., 2015; Roma-Rodrigues et al., 2019). Это селективное давление способствует выживанию и размножению определенных клонов с преимущественными признаками, которые позволяют им адаптироваться к условиям микроокружения.

Среди факторов микроокружения особенно важное место занимает иммунная система. Действие иммунных клеток играет двойственную роль в развитии рака: оно может как подавлять рост опухоли, так и способствовать ее прогрессии. Механизмы иммунного надзора распознают и уничтожают раковые клетки, предотвращая образование опухоли. Однако опухоли могут уклоняться от иммунного ответа с помощью различных механизмов, в результате чего иммунный ответ начинает играть роль фактора естественного отбора для популяций клонов и тем самым образом отбирать наиболее устойчивые клональные популяции с измененными антигенами, что напрямую влияет на тяжесть протекания болезни и эффективность иммунотерапии.

Наконец, **теория эпигенетической пластичности** предполагает, что, помимо генетических нарушений, значимую роль в возникновении гетерогенности опухолей играют эпигенетические изменения (Flavahan et al., 2017; Yao et al., 2020). Эпигенетические модификации, такие как метилирование ДНК и модификации гистонов, могут динамически регулировать паттерны экспрессии генов и клеточные фенотипы. По этой теории раковые клетки обладают пластичным эпигенетическим ландшафтом, который позволяет обратимо и динамично изменять экспрессию генов. Такие эпигенетические изменения могут привести к появлению различных клонов с разными фенотипическими характеристиками, способствуя внутриопухолевой гетерогенности.

Методы изучения эволюционных характеристик в гетерогенных опухолях

Для изучения эволюционных особенностей гетерогенных опухолей исследователю необходимо иметь возможность качественной и количественной оценки различных клональных популяций. В данном разделе мы приведем отдельные методы анализа, которые используются в настоящее время для изучения гетерогенности опухолей.

Одним из способов теоретического исследования гетерогенных сообществ опухолей является подход популяционной генетики. С точки зрения популяционной генетики эволюция любой популяции будет зависеть от двух факторов – скорости мутаций и эффективного размера популяции. Скорость мутации определяется как ожидаемое количество генетических мутаций на единичное событие репликации и напрямую влияет на разнообразие в популяции. Эффективный же размер популяции определяет ее способность к поддержанию этого разнообразия. В случае опухолей эффективный размер определяется как общее число раковых клеток, но возможно также выключение некоторых групп раковых клеток из этого числа – если, например, моделируется опухоль, вызванная РСК, которые и будут основной причиной роста опухоли. Разумеется, для применения подобного подхода необходимо использовать одноклеточное секвенирование опухолей. В силу сложности и дороговизны этого метода классический анализ популяционной генетики проводился лишь в нескольких

работах (Navin, 2015; Losic et al., 2020; Heinrich et al., 2021; Deng et al., 2023).

Поскольку методы одноклеточного секвенирования стали доступны сравнительно недавно, большая часть работ посвящена изучению гетерогенности при помощи методов секвенирования нового поколения на образцах из всех клеток опухолей. У этого подхода есть очевидная проблема: в данных, полученных из таких образцов, сложно напрямую выделить клональную архитектуру опухоли. Поэтому при использовании рассматриваемого подхода исследователям приходится применять определенные допущения и модификации экспериментальных методов. Один из них заключается в увеличении глубины секвенирования для оценки частот мутантных аллелей (Koh et al., 2021). При помощи статистических методов эти частоты нормализуются, и на их основе кластеризуются генотипы для определения идентичных клональных популяций. В подобных исследованиях часто используются такие характеристики разнообразия, как индекс разнообразия Шеннона, индекс Симпсона и т. д. Недостаток подхода заключается в том, что он не способен разграничить популяции, если они обладают схожими частотами мутантных аллелей.

Другой модификацией является мультирегиональное секвенирование, в котором отбор образцов идет в нескольких участках опухоли. В частности, такой метод позволяет оценить разницу в гетерогенности у пациентов с несколькими метастатическими опухолями, которых в контексте разнообразия можно воспринимать как популяции клонов с длительной физической изоляцией.

Разумеется, самой перспективной методикой для экспериментальной оценки гетерогенности считаются методы анализа отдельных клеток, поскольку они позволяют судить об индивидуальных различиях клонов на генетическом и фенотипическом уровнях. Один из таких методов – иммунофлюоресцентная *in situ* гибридизация (iFISH). Благодаря использованию флуоресцентно-меченных ДНК-зондов, гибридизирующихся с комплементарными последовательностями мишеней, FISH предоставляет возможность с высокой специфичностью и чувствительностью выявлять генетические изменения, хромосомные перестройки и амплификации генов. *In situ* FISH (iFISH) – это применение FISH непосредственно на срезах ткани, с сохранением пространственной организации клеток в микроокружении опухоли (Gertz et al., 2016). Тем не менее метод iFISH является низкопроизводительным и не позволяет исследовать гетерогенность на полногеномном уровне.

В отличие от вышеописанного метода, одноклеточное секвенирование (scDNA-seq и scRNA-seq) позволяет определить картину генетического разнообразия, экспрессии генов в каждой отдельной клетке и расшифровать ее межклеточные сигнальные сети. Эти методы дают четкое представление не только о механизмах внутриопухолевой гетерогенности, но и о межклеточных взаимодействиях посредством лиганд-рецепторной передачи сигналов.

Заключение

Понимание эволюции и гетерогенности злокачественных опухолей имеет решающее значение для улучшения диагностики и разработки стратегий лечения рака. Для изучения генетических и фенотипических характеристик популяций раковых клонов было разработано множество молекулярно-генетических методов со своими достоинствами и недостатками. С помощью секвенирования нового поколения можно получить полное представление о геномном ландшафте опухоли, однако существует опасность упустить из виду редкие клоны. Одноклеточное секвенирование позволяет выявлять редкие клоны и реконструировать клональные линии, но технически сложно и дорогостояще. Такие методы, как iFISH, позволяют получить пространственную информацию, но имеют ограниченный охват мишеней и низкую производительность.

На основании полученных с помощью подобных методов данных, для объяснения динамического характера эволюции опухолей были предложены различные модели, включая модели клональной эволюции, раковых стволовых клеток, модели влияния микроокружения и эпигенетических факторов. Каждая из них дает ценное представление о механизмах, обусловливающих гетерогенность опухоли и возникновение лекарственной устойчивости.

Кроме того, разработка математических и компьютерных моделей клональной эволюции и алгоритмов анализа крупномасштабных геномных данных могла бы расширить возможности по интерпретации и извлечению значимой информации из сложных наборов данных о злокачественных опухолях. Эти инструменты потенциально позволят исследователям выявлять ключевые драйверные события, отслеживать эволюционную динамику и точнее прогнозировать последствия лечения.

Список литературы / References

Augustin R.C., Delgoffe G.M., Najjar Y.G. Characteristics of the tumor microenvironment that influence immune cell functions: hypoxia, oxidative stress, metabolic alterations. *Cancers* (*Basel*). 2020; 12(12):3802. DOI 10.3390/cancers12123802

Baca S.C., Prandi D., Lawrence M.S., Mosquera J.M., Romanel A., Drier Y., Park K., Kitabayashi N., MacDonald T.Y., Ghandi M., Van Allen E., Kryukov G.V., Sboner A., Theurillat J.-P., Soong T.D., Nickerson E., Auclair D., Tewari A., Beltran H., Onofrio R.C., Boysen G., Guiducci C., Barbieri C.E., Cibulskis K., Sivachenko A., Carter S.L., Saksena G., Voet D., Ramos A.H., Winckler W., Cipicchio M., Ardlie K., Kantoff P.W., Berger M.F., Gabriel S.B., Golub T.R., Meyerson M., Lander E.S., Elemento O., Getz G., Demichelis F., Rubin M.A., Garraway L.A. Punctuated evolution of prostate cancer genomes. *Cell*. 2013;153(3):666-677. DOI 10.1016/j.cell.2013.03.021

Besse A., Clapp G.D., Bernard S., Nicolini F.E., Levy D., Lepoutre T. Stability analysis of a model of interaction between the immune system and cancer cells in chronic myelogenous leukemia. *Bull. Math. Biol.* 2018;80(5):1084-1110. DOI 10.1007/s11538-017-0272-7

Bonnet D., Dick J.E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* 1997;3(7):730-737. DOI 10.1038/nm0797-730

Deng G., Zhang X., Chen Y., Liang S., Liu S., Yu Z., Lü M. Single-cell transcriptome sequencing reveals heterogeneity of gastric cancer: progress and prospects. *Front. Oncol.* 2023;13:1074268. DOI 10.3389/fonc.2023.1074268

Durrett R., Foo J., Leder K., Mayberry J., Michor F. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*. 2011;188(2):461-477. DOI 10.1534/genetics.110.125724

Fearon E.R., Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell*. 1990;61(5):759-767. DOI 10.1016/0092-8674(90)90186-I

Flavahan W.A., Gaskell E., Bernstein B.E. Epigenetic plasticity and the hallmarks of cancer. *Science.* 2017;357(6348):eaal2380. DOI 10.1126/science.aal2380

Furukawa Y., Kikuchi J. Molecular basis of clonal evolution in multiple myeloma. *Int. J. Hematol.* 2020;111(4):496-511. DOI 10.1007/s12185-020-02829-6

Gawad C., Koh W., Quake S.R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl. Acad. Sci. USA.* 2014;111(50):17947-17952. DOI 10.1073/pnas.1420822111

Gertz E.M., Chowdhury S.A., Lee W.-J., Wangsa D., Heselmeyer-Haddad K., Ried T., Schwartz R., Schäffer A.A. FISHtrees 3.0: tumor phylogenetics using a ploidy probe. *PLoS One*. 2016;11(6):e0158569. DOI 10.1371/journal.pone.0158569

Graham T.A., Sottoriva A. Measuring cancer evolution from the genome. *J. Pathol.* 2017;241(2):183-191. DOI 10.1002/path.4821

Haffner M.C., Zwart W., Roudier M.P., True L.D., Nelson W.G., Epstein J.I., De Marzo A.M., Nelson P.S., Yegnasubramanian S. Genomic and phenotypic heterogeneity in prostate cancer. *Nat. Rev. Urol.* 2021;18(2):79-92. DOI 10.1038/s41585-020-00400-w

Hata M., Hayakawa Y., Koike K. Gastric stem cell and cellular origin of cancer. *Biomedicines*. 2018;6(4):100. DOI 10.3390/biomedicines6040100

Hausser J., Alon U. Tumour heterogeneity and the evolutionary trade-offs of cancer. *Nat. Rev. Cancer*. 2020;20(4):247-257. DOI 10.1038/s41568-020-0241-6

Heinrich S., Craig A.J., Ma L., Heinrich B., Greten T.F., Wang X.W. Understanding tumour cell heterogeneity and its implication for immunotherapy in liver cancer using single-cell analysis. *J. Hepatol.* 2021;74(3):700-715. DOI 10.1016/j.jhep.2020.11.036

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

Lee T.K.-W., Guan X.-Y., Ma S. Cancer stem cells in hepatocellular carcinoma – from origin to clinical implications. *Nat. Rev. Gastroenterol. Hepatol.* 2022;19(1):26-44. DOI 10.1038/s41575-021-00508-3

Losic B., Craig A.J., Villacorta-Martin C., Martins-Filho S.N., Akers N., Chen X., Ahsen M.E., von Felden J., Labgaa I., D'Avola D., Allette K., Lira S.A., Furtado G.C., Garcia-Lezana T., Restrepo P., Stueck A., Ward S.C., Fiel M.I., Hiotis S.P., Gunasekaran G., Sia D., Schadt E.E., Sebra R., Schwartz M., Llovet J.M., Thung S., Stolovitzky G., Villanueva A. Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Commun.* 2020;11(1):291. DOI 10.1038/s41467-019-14050-z

Lüönd F., Tiede S., Christofori G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br. J. Cancer*. 2021;125(2):164-175. DOI 10.1038/s41416-021-01328-7

Meacham C.E., Morrison S.J. Tumour heterogeneity and cancer cell plasticity. *Nature*. 2013;501(7467):328-337. DOI 10.1038/nature12624

Merlo L.M.F., Pepper J.W., Reid B.J., Maley C.C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*. 2006;6(12):924-935. DOI 10.1038/nrc2013

Morris L.G.T., Riaz N., Desrichard A., Şenbabaoğlu Y., Hakimi A.A., Makarov V., Reis-Filho J.S., Chan T.A. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*. 2016;7(9):10051-10063. DOI 10.18632/oncotarget.7067

Mumenthaler S.M., Foo J., Choi N.C., Heise N., Leder K., Agus D.B., Pao W., Michor F., Mallick P. The impact of microenvironmental heterogeneity on the evolution of drug resistance in cancer cells. *Cancer Inform.* 2015;14(Suppl.4):19-31. DOI 10.4137/CIN.S19338

Navin N.E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* 2015;25(10):1499-1507. DOI 10.1101/gr.191098.115

Nowell P. The clonal evolution of tumor cell populations. *Science.* 1976;194(4260):23-28. DOI 10.1126/science.959840

Reya T., Morrison S.J., Clarke M.F., Weissman I.L. Stem cells, cancer, and cancer stem cells. *Nature*. 2001;414(6859):105-111. DOI 10.1038/35102167

Robertson-Tessi M., Gillies R.J., Gatenby R.A., Anderson A.R.A. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. *Cancer Res.* 2015;75(8):1567-1579. DOI 10.1158/0008-5472.CAN-14-1428

Roma-Rodrigues C., Mendes R., Baptista P., Fernandes A. Targeting tumor microenvironment for cancer therapy. *Int. J. Mol. Sci.* 2019;20(4):840. DOI 10.3390/ijms20040840

Vendramin R., Litchfield K., Swanton C. Cancer evolution: Darwin and beyond. *EMBO J.* 2021;40(18):e108389. DOI 10.15252/embj.2021108389

Vosberg S., Greif P.A. Clonal evolution of acute myeloid leukemia from diagnosis to relapse. *Genes Chromosomes Cancer*. 2019;58(12):839-849. DOI 10.1002/gcc.22806

Walcher L., Kistenmacher A.-K., Suo H., Kitte R., Dluczek S., Straußa A., Blaudszun A.-R., Yevsa T., Fricke S., Kossatz-Boehlert U. Cancer stem cells-origins and biomarkers: perspectives for targeted personalized therapies. *Front. Immunol.* 2020;11:1280. DOI 10.3389/fimmu.2020.01280

Wang Y., Waters J., Leung M.L., Unruh A., Roh W., Shi X., Chen K., Scheet P., Vattathil S., Liang H., Multani A., Zhang H., Zhao R., Michor F., Meric-Bernstam F., Navin N.E. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*. 2014;512(7513):155-160. DOI 10.1038/nature13600

Williams M.J., Werner B., Barnes C.P., Graham T.A., Sottoriva A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* 2016;48(3):238-244. DOI 10.1038/ng.3489

Yao J., Chen J., Li L.-Y., Wu M. Epigenetic plasticity of enhancers in cancer. *Transcription*. 2020;11(1):26-36. DOI 10.1080/21541264.2020.1713682

Zarzynska J.M. The role of stem cells in breast cancer. In: Breast Cancer – From Biology to Medicine. InTech, 2017. DOI 10.5772/66904

Zhao T., Chiang Z.D., Morriss J.W., LaFave L.M., Murray E.M., Del Priore I., Meli K., Lareau C.A., Nadaf N.M., Li J., Earl A.S., Macosko E.Z., Jacks T., Buenrostro J.D., Chen F. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature*. 2022;601(7891):85-91. DOI 10.1038/s41586-021-04217-4

ORCID ID

R.A. Ivanov orcid.org/0000-0002-4369-356X
S.A. Lashin orcid.org/0000-0003-3138-381X

Благодарности. Работа выполнена при поддержке бюджетного проекта № FWNR-2022-0020.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 13.07.2023. После доработки 07.08.2023. Принята к публикации 17.08.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):820-828
DOI 10.18699/VJGB-23-95

Перевод на английский язык https://vavilov.elpub.ru/jour

Поиск дифференциально метилированных регионов в геномах древних и современных людей

Д.Д. Бородко ✉, С.В. Женило, Ф.С. Шарко

Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Россия
✉ daria.borodko@gmail.com

Аннотация. В настоящее время активно исследуются механизмы, регулирующие развитие различных патологий и их эволюционную динамику. Эпигенетические механизмы, такие как метилирование, играют значимую роль в эволюционных процессах, поскольку их изменения гораздо быстрее отражаются на фенотипе, чем результаты мутагенеза. В данном исследовании мы предприняли попытку разработать алгоритм для выявления дифференциально метилированных областей, связанных с метаболическим синдромом, которые изменили свое метилирование у человека при переходе от охоты и собирательства к оседлой жизни. Применение существующих методов полногеномного бисульфитного секвенирования ограничено для древних образцов из-за их низкого качества и фрагментации, и подход к получению профилей метилирования охотников-собирателей значительно отличается от подходов, используемых для современных тканей. В этой работе мы валидировали DamMet – алгоритм, реконструирующий древние метиломы. Применение DamMet к геномам неандертальца и денисовца показало средний уровень корреляции с профилями метилирования, опубликованными ранее, а также продемонстрировало занижение уровня метилирования реконструированных профилей в среднем на 15–20 %. Также мы разработали новый алгоритм на языке Python, позволяющий сравнивать метиломы в древних и современных образцах, несмотря на отсутствие профилей метилирования современных образцов костной ткани в контексте ожирения. Такой анализ подразумевает двухступенчатую обработку данных, где на первом этапе происходит идентификация тканеспецифичных областей метилирования и их фильтрация, а на втором этапе осуществляется непосредственно поиск дифференциально метилированных регионов в заданных областях, ассоциированных с интересующим исследователя заболеванием. В результате использования алгоритма на тестовых данных мы обнаружили 38 дифференциально метилированных регионов, ассоциированных с ожирением, большая часть которых принадлежала промоторным областям, и разработанный пайплайн показал достаточную эффективность в их поиске. Эти результаты подтверждают возможность восстановления профилей метилирования в древних образцах и их сравнения с современными метиломами. Также обсуждаются возможности дальнейшего развития методологии и внедрения нового шага, позволяющего изучать дифференциально метилированные позиции, связанные с эволюционными процессами.
Ключевые слова: древняя ДНК; метилирование; эпигенетика; DamMet; ДМР.

Для цитирования: Бородко Д.Д., Женило С.В., Шарко Ф.С. Поиск дифференциально метилированных регионов в геномах древних и современных людей. *Вавиловский журнал генетики и селекции*. 2023;27(7):820-828. DOI 10.18699/VJGB-23-95

Search for differentially methylated regions in ancient and modern genomes

D.D. Borodko ✉, S.V. Zhenilo, F.S. Sharko

Federal Research Center "Fundamentals of Biotechnology" of the Russian Academy of Sciences, Moscow, Russia
✉ daria.borodko@gmail.com

Abstract. Currently, active research is focused on investigating the mechanisms that regulate the development of various pathologies and their evolutionary dynamics. Epigenetic mechanisms, such as DNA methylation, play a significant role in evolutionary processes, as their changes have a faster impact on the phenotype compared to mutagenesis. In this study, we attempted to develop an algorithm for identifying differentially methylated regions associated with metabolic syndrome, which have undergone methylation changes in humans during the transition from a hunter-gatherer to a sedentary lifestyle. The application of existing whole-genome bisulfite sequencing methods is limited for ancient samples due to their low quality and fragmentation, and the approach to obtaining DNA methylation profiles differs significantly between ancient hunter-gatherer samples and modern tissues. In this study, we validated DamMet, an algorithm for reconstructing ancient methylomes. Application of DamMet to Neanderthal and Denisovan genomes showed a moderate level of correlation with previously published methylation profiles and demonstrated an underestimation of methylation levels in the reconstructed profiles by an average of 15–20 %. Additionally, we developed a new Python-based algorithm that allows for the comparison of methylomes in ancient and modern samples, despite the absence of methylation profiles in modern bone tissue within the context of obesity. This analysis involves a two-step data processing approach, where the first step involves the identification and

filtration of tissue-specific methylation regions, and the second step focuses on the direct search for differentially methylated regions in specific areas associated with the researcher's target condition. By applying this algorithm to test data, we identified 38 differentially methylated regions associated with obesity, the majority of which were located in promoter regions. The pipeline demonstrated sufficient efficiency in detecting these regions. These results confirm the feasibility of reconstructing DNA methylation profiles in ancient samples and comparing them with modern methylomes. Furthermore, possibilities for further methodological development and the implementation of a new step for studying differentially methylated positions associated with evolutionary processes are discussed.

Key words: ancient DNA; methylation; epigenetics; DamMet; DMR.

For citation: Borodko D.D., Zhenilo S.V., Sharko F.S. Search for differentially methylated regions in ancient and modern genomes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):820-828. DOI 10.18699/VJGB-23-95

Введение

В последнее время все большее внимание уделяется изучению механизмов, регулирующих развитие различных патологий и их эволюционную динамику (Briggs et al., 2009a; Niiranen et al., 2022). Особенно важную роль в этом процессе играют эпигенетические механизмы, такие как метилирование, поскольку они способны вызывать фенотипические изменения гораздо быстрее, чем обычные процессы мутагенеза (Jablonka, Raz, 2009; Feinberg, Irizarry, 2010; Zhur et al., 2021). Главной целью данного исследования было выявление дифференциально метилированных регионов (ДМР), связанных с метаболическим синдромом, которые могли бы стать потенциальными целями эпигенетической терапии метаболического синдрома.

Сейчас ученых нередко останавливает от проведения эволюционных исследований отсутствие подходящих методов для сравнения профилей ДНК древних и современных образцов. Лабораторные протоколы, применяемые для получения этих профилей, значительно отличаются друг от друга и имеют свои особенности и ошибки. Древняя ДНК (дДНК) часто находится во фрагментированном состоянии (Sawyer et al., 2012), и с течением времени происходит естественная деградация молекулы и спонтанное дезаминирование азотистых оснований, что ограничивает доступность качественных данных (Briggs et al., 2007, 2009b). Для решения этой проблемы был предложен специальный протокол обработки образцов с применением урацил-ДНК-гликозилазы (УДГ) и эндонуклеазы, чтобы облегчить извлечение профилей метилирования и повысить их различимость (Briggs et al., 2010), а также несколько программ, позволяющих рассчитать уровни метилирования в древних образцах, последовательности которых были секвенированы с применением УДГ (Gokhman et al., 2014; Orlando et al., 2015; Hanghøj et al., 2019).

Существует два алгоритма реконструкции метилирования древних образцов, которые были разработаны для командной строки и достаточно дружелюбны для пользователя. Более ранний алгоритм epiPALEOMIX основан на исторически первом способе реконструкции метиломов, опубликованном Д. Гохманом в 2014 г., и имеет несколько модулей, в том числе модуль MethylMap, позволяющий получить уровни метилирования в регионах, заданных пользователем (Hanghøj et al., 2016). Однако в этом заключается и ограничение по его применению: пользователю необходимо знать, какие регионы он исследует. Результатом работы алгоритма является рассчитанное количество дезаминированных метилированных цитозинов в CpG контексте и покрытие; соответственно, их отношение и будет уровнем метилирования в данной позиции генома. Алгоритм DamMet более универсален, поскольку, в отличие от epiPALEOMIX, ориентирован на полногеномные исследования. Помимо этого, DamMet способен рассчитывать уровни дезаминирования в метилированных и неметилированных CpG на каждой позиции рида, т. е. использует модель, которая наиболее точно описывает процесс дезаминирования цитозинов во фрагментах дДНК как случайный (Hanghøj et al., 2019).

В настоящее время для исследования метилирования ДНК в современных образцах применяется также метод полногеномного бисульфитного секвенирования (WGBS) (Olova et al., 2018; Suzuki et al., 2018). Для восстановления метилирования из образцов, секвенированных с помощью этой технологии, существует несколько методов (Clark et al., 1994; Bock et al., 2005), из которых самыми известными считаются Bismark, BoostMe и WGBStools. Алгоритм Bismark наиболее часто используется для препроцессинга данных WGBS, что подразумевает картирование ридов на конвертированный референсный геном, а затем подсчет количества метилированных и неметилированных цитозинов на каждой позиции генома (Krueger, Andrews, 2011). Этот метод, как и большинство методов, основанных на подсчете ридов, недостаточно эффективно преодолевает проблему малого покрытия образцов, которая часто появляется при использовании образцов низкого качества или проведении single cell экспериментов. Для решения этой проблемы были разработаны алгоритмы на основе машинного обучения, такие как DeepCPG и BoostMe.

DeepCpG – это алгоритм на основе нейронной сети глубокого обучения, который прогнозирует состояния метилирования малопокрытых сайтов и обнаруживает мотивы, связанные с изменением уровней метилирования и межклеточной изменчивостью (Angermueller et al., 2017). Этот инструмент применяется для улучшения качества данных single cell экспериментов. BoostMe, базирующийся на методе машинного обучения, позволяет решить проблему на этапе препроцессинга генома с помощью импутации (Zou et al., 2018). Градиентный бустинг (XGBoost), имплементированный в этом инструменте, сочетает данные нескольких образцов (более трех) для коррекции пропущенных уровней метилирования в образцах современных тканей, что позволяет использовать для восстановления метилирования образцы с низким покрытием генома. Кроме того, преимуществом BoostMe является способность

восстанавливать не только состояние данного CpG (метилирован/не метилирован), но и уровень метилирования. WGBStools – набор методов, разработанный в рамках проекта по составлению атласа метилирования современных тканей. Используется для сверхкомпактного представления картированных прочтений, статистического анализа, а также для визуализации от небольших фрагментов до целых хромосомных локусов (https://github.com/nloyfer/wgbs_tools).

Несмотря на разнообразие алгоритмов восстановления метилирования, применение WGBS технологии для образцов дДНК ограничено, так как для бисульфитной конверсии требуется большая концентрация хорошо очищенной ДНК. Также при конверсии происходит фрагментация ДНК, что еще сильнее ухудшает качество дДНК, и так значительно фрагментированной из-за деградации (Gu et al., 2011). Следовательно, алгоритмы расчета уровней метилирования, применяемые для современных образцов, не могут использоваться для реконструкции профилей метилирования древних людей. Поэтому мы сосредоточились на разработке нового алгоритма, который позволил бы сравнивать метиломы в древних и современных образцах, учитывая отсутствие доступных образцов костной ткани для проведения полногеномного бисульфитного секвенирования в контексте ожирения.

Материалы и методы

Подготовка данных. Для анализа мы отобрали в базе данных GEO NCBI 11 древних геномов и 12 современных профилей метилирования, полученных с помощью методов Whole Genome Bisulfite Sequencing. При выборе древних образцов мы обратили особое внимание на их возраст, стратегию подготовки библиотек и полученное покрытие по геному: использовались только образцы, полученные с предварительной обработкой урацил-ДНК-гликозилазой (УДГ), не моложе 3 тыс. лет до н. э. и с покрытием не менее 5x. Полные геномы древних образцов были отсеквенированы с обработкой УДГ, кроме образцов Vi33 и PES001 (Peschanitsa), которые не были обработаны УДГ перед секвенированием (табл. 1).

Выбор 12 современных образцов (Loyfer el al., 2023) определен мезодермальным происхождением тканей, из которых были подготовлены библиотеки, а также применением метода полногеномного бисульфитного секвенирования. Данные об образцах приведены в табл. 2.

Препроцессинг ридов древней ДНК. Древние геномы были загружены в формате bam-файлов с индексами. Согласно (Ohm et al., 2010; Gokhman et al., 2014), УДГ недостаточно эффективно работает на концах ДНК. Соответственно, для корректного анализа дДНК при помощи утилиты trimBam мы убрали по два нуклеотида с 3′- и 5′-концов последовательностей (Gansauge, Meyer, 2013; Jun et al., 2015). Для образцов Vi33 и PES001 данная процедура не проводилась, поскольку образцы не были обработаны УДГ при подготовке библиотек. Помимо этого, все риды образцов были отфильтрованы с помощью Trimmomatic (Bolger et al., 2014) по среднему качеству и длине, и в дальнейшем анализе использовались только выровненные на сборку CRCh37 (hg19) последовательности со средним качеством более 20 и длиной более 25 п. о.

Реконструкция профилей метилирования древних людей. Для восстановления метиломов древних образцов мы использовали ПО DamMet (Hanghøj et al., 2019). Пайплайн включал в себя три этапа: фильтрацию однонуклеотидных вариантов, расчет уровней дезаминирования для каждой позиции рида и расчет уровня метилирования. Расчет SNV проводился с помощью GATK HaplotypeCaller v4.3.0.0 (Poplin et al., 2017) с последующей фильтра-

Таблица 1. Древние геномы, отобранные для анализа

Образец	Группа	Возраст образца, тыс. лет	Пол	Ткань	Покрытие	Профиль метилирования	Геномное окно (в CpG)	Литературный источник
Altai Neanderthal	Древний	120	Ж	Фаланга стопы	50	Gokhman et al., 2014, 2020	25	Prüfer et al., 2014
Denisovan		75	Ж	Фаланга стопы	30		25	Meyer et al., 2012
Vindija33		50	Ж	Неопознанная кость	30		50	Prüfer et al., 2017
Ust'-Ishim	ОС	45	М	Бедренная кость	42 (22 XY)	Gokhman et al., 2020	25	Fu et al., 2014
Sunghir		35	М	Бедренная кость + зубы	10.7	Эта работа	38	Sikora et al., 2017
USR1		11.5	Ж	Височная кость	17		50	Moreno-Mayar et al., 2018a
Spirit Cave		11	М	Височная кость + зубы	18		33	Moreno-Mayar et al., 2018b
Peschanitsa		11	М	Зубы	5		50	Saag et al., 2021
SF12		9	Ж	Бедренная кость	57.79		28	Günther et al., 2018
2H10 (France)		3.2	М	Зубы	13.9		33	Seguin-Orlando et al., 2021
2H11 (France)		3.2	М	Зубы	23.9		33	Seguin-Orlando et al., 2021

Примечание. Окно сглаживания – параметр усреднения уровней дезаминирования на следующем этапе анализа; ОС – охотники-собиратели.

Таблица 2. Современные геномы, использованные для поиска тканеспецифично метилированных регионов и ДМР

GEO accession	Пол	Возраст пациента	Орган	Ткань
GSM5652198	М	37	Толстая кишка	Фибробласты
GSM5652202	Ж	35	Сердце	
GSM5652204	М	73	Дерма	
GSM5652205	Ж	59	Скелетная мышца	Гладкие миоциты
GSM5652207	М	22	Аорта	
GSM5652209	Ж	51	Мочевой пузырь	
GSM5652210	М	24	Простата	
GSM5652211	М	57	Легочный бронх	
GSM5652212	М	83	Сердце	Кардиомиоциты
GSM2637888	–	–	Кость	–
GSM2637887	–	–	Кость	–
GSM5652218	Ж	7	Кость	Остеобласты
GSM5652177	Ж	35	Подкожная жировая ткань	Адипоциты
GSM5652176	Ж	53	Подкожная жировая ткань	
GSM5652178	Ж	37	Подкожная жировая ткань	

цией вариантов с покрытием ниже 5 и качеством ниже 30, а также при наличии гомозиготности по альтернативному аллелю или более чем 2 альтернативных аллелей, если на этой позиции находится цитозин. Данный этап был рекомендован автором алгоритма DamMet в публикации (Hanghøj et al., 2019) и приведен в дополнительных материалах источника.

Далее проводилась реконструкция уровней метилирования с исключением найденных вариантов.

```
DamMet estDEAM -b <bam-file> -r <fasta-file> -c
<chromosome> -M <expected-average-methylation>
-O <out-file-prefix> -E <vcf-to-exclude> -L 25
-P 50 -q 20 -Q 20
```

Затем мы рассчитали уровень метилирования с использованием найденных уровней дезаминирования на позициях с метилированными и деметилированными цитозинами. Размер геномного окна для каждого образца указан в соответствующем столбце табл. 1 и был подобран экспериментально.

```
DamMet estF -b <bam-file> -r <fasta-file> -c
<chromosome> -M <expected-average-methylation>
-O <out-file-prefix> -N <genomic-window-size-
in-CpGs>
```

Полученные профили метилирования были дополнительно сглажены с помощью бегущего среднего с окном сглаживания 25 CpG скриптом на языке Python.

Валидация реконструированных метиломов. Сравнение профилей метилирования неандертальца, денисовца и усть-ишимского охотника-собирателя, полученных на предыдущем этапе, производилось с помощью языка R. Для препроцессинга данных, расчета корреляций и построения графиков мы использовали пакеты ggplot, psych, corr.test и семейство tidyverse.

Поиск тканеспецифичных областей метилирования. Для поиска регионов, метилирование которых примерно одинаково во всех тканях мезодермального происхождения, был разработан скрипт на языке Python. На вход подаются значения метилирования, рассчитанные алгоритмом Bismark (Krueger, Andrews, 2011) после картирования образцов, упоминаемых выше. Сравнение значений метилирования проводилось попозиционно с помощью ANOVA для выявления различий в трех группах тканей (фибробласты, миоциты, остеобласты) и исключения из древних костных и современных адипоцитарных профилей метилирования позиций, в которых было обнаружено дифференциальное метилирование ($p < 0.05$).

Поиск ДМР. Подготовленные на предыдущих этапах профили метилирования охотников-собирателей (ОС) и современных людей сравнивались методом ANOVA, как и при поиске тканеспецифичного метилирования. В первой итерации образцы разделялись на три группы: кости охотников-собирателей, адипоциты здоровых людей и адипоциты пациентов с ожирением. При наличии уровня значимости $p < 0.05$ CpG-сайт отбирался для последующего анализа с помощью *post hoc* теста Тьюки. CpG-сайт считался дифференциально метилированным, если изменение метилирования было значимо ($p < 0.05$) при сравнении костей ОС с адипоцитами тучных людей и незначимо при сравнении ОС с контролем. Во второй итерации мы изменили разделение по группам: все образцы были костными, группы представляли собой образцы разного возраста (анатомически древние люди, охотники-собиратели и современные люди), для облегчения нагрузки на вычислительные ресурсы сравнения проводили только в регионах, ассоциированных с ожирением. Для сборки полученных дифференциально метилированных сайтов в регионы мы воспользовались ПО combined-pvalues (https://github.com/brentp/combined-pvalues), в основе которого лежит метод поправок на множественное тестирование Штуффера–Липтака (Pedersen et al., 2012). Статус изменения метилирования определялся сравнением средних значений метилирования региона в группах.

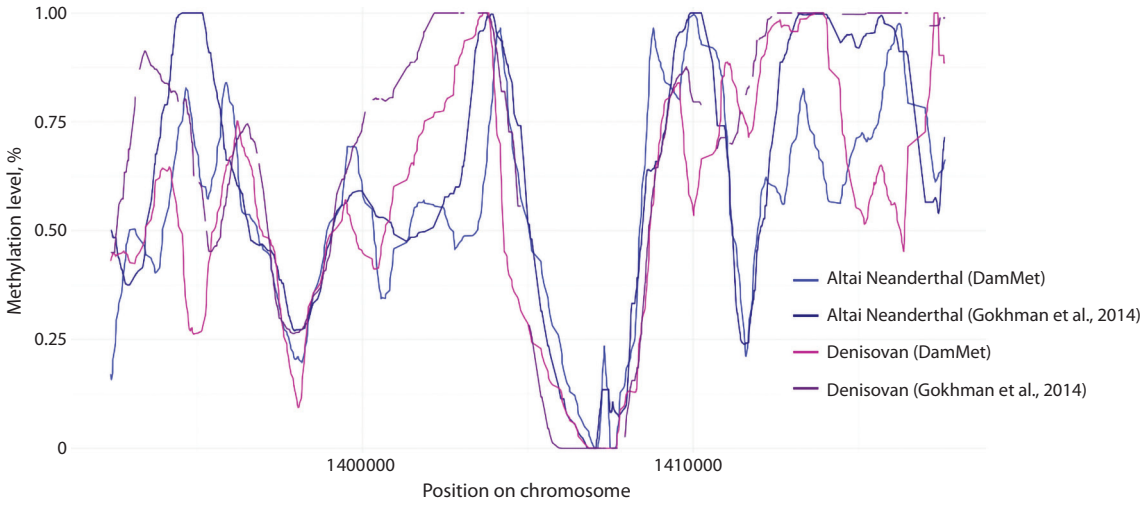


Рис. 1. Профили метилирования денисовца и неандертальца, восстановленные DamMet и опубликованные Д. Гохманом.
В поле зрения – деметилированный CpG-островок chr1:1406845–1407821.

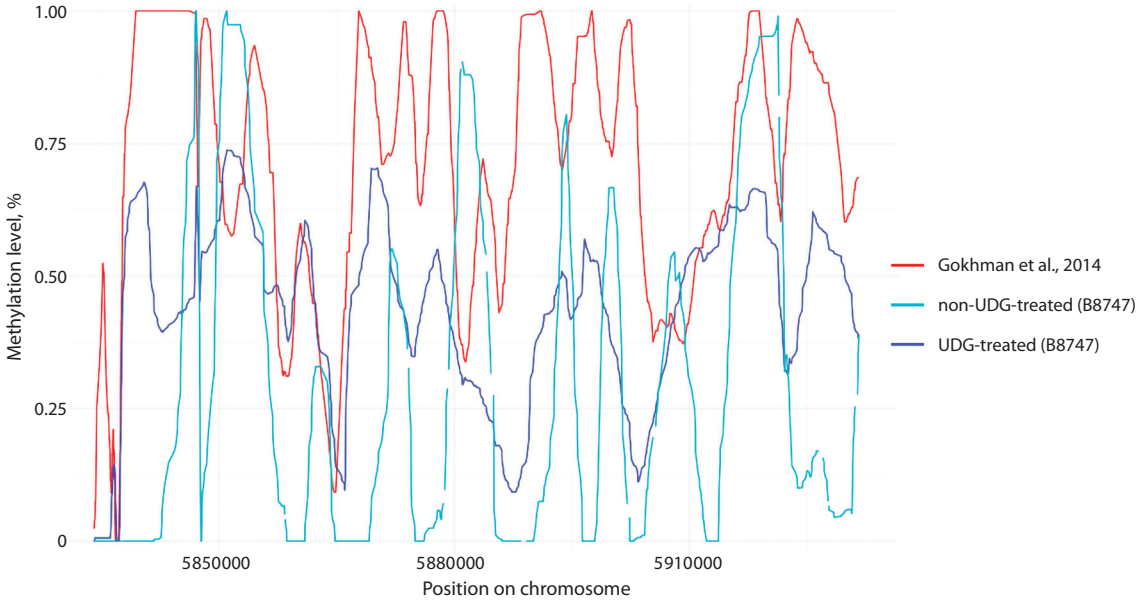


Рис. 2. Сравнение уровней метилирования на участке хромосомы 2 образца Vi33 в присутствии и в отсутствие УДГ обработки библиотек с ранее опубликованными профилями Д. Гохмана.
Уровни метилирования всех образцов были сглажены с помощью бегущего среднего с окном 25 CpG.

Результаты

В данном исследовании мы реконструировали 11 профилей метилирования древних людей с помощью инструмента DamMet. В первую очередь необходимо было разработать пайплайн, который позволял бы восстанавливать метиломы с высокой точностью; для этого мы подали на вход в пайплайн геномы неандертальца и денисовца, прошедшие обработку УДГ. Профили для этих организмов уже были опубликованы ранее (Gokhman et al., 2014, 2020), что позволяет нам провести валидацию пайплайна. Мы обнаружили, что рассчитанное нами метилирование в среднем на 15–20 % ниже, чем ранее опубликованное, при этом в общем профили метилирования схожи (рис. 1). Коэффициент корреляции профилей метилирования в обоих случаях более 85 %: $r_{\text{Denisovan}} = 0.87$, $r_{\text{Neanderthal}} = 0.9$ ($p < 0.05$).

Поскольку у нас было несколько образцов, которые не прошли обработку урацил-ДНК-гликозилазой на этапе подготовки библиотек, мы также хотели удостовериться, что DamMet способен реконструировать профили метилирования и в отсутствие этого этапа пробоподготовки. Для этого мы выбрали образец Vi33, для которого в открытом доступе были опубликованы последовательности, как прошедшие УДГ обработку, так и не прошедшие ее. Параметры пайплайна не изменялись для этих расчетов, т. е. условия были одинаковы при реконструкции мети-

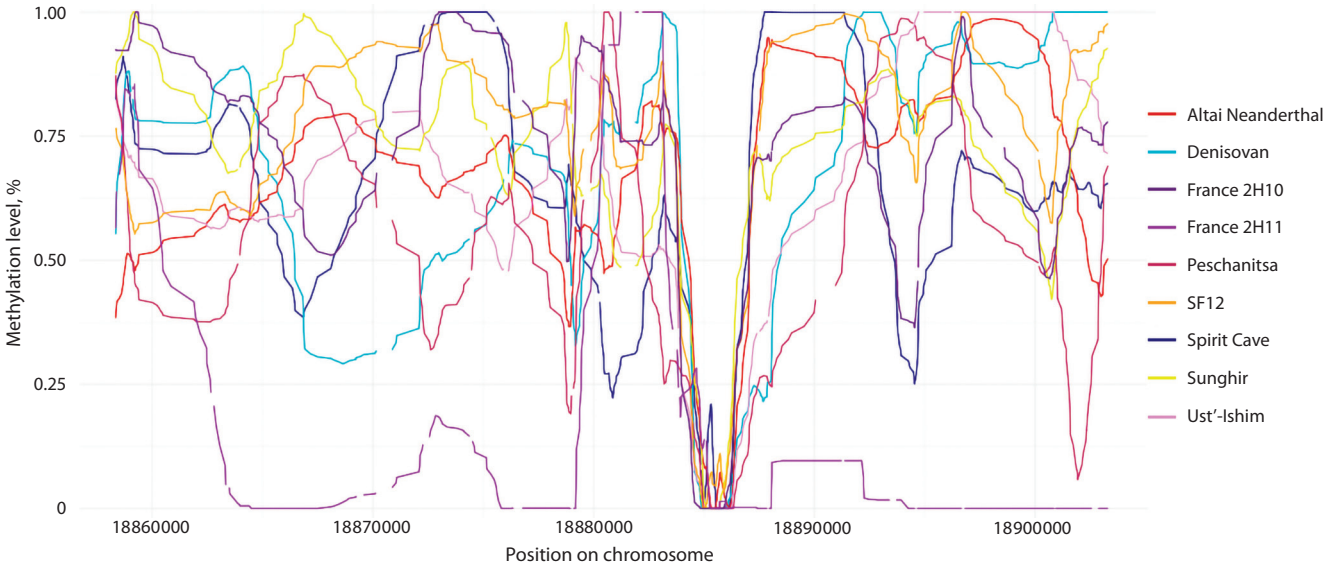


Рис. 3. Профили метилирования охотников-собирателей, реконструированные с помощью DamMet.
Область протяженного деметилирования принадлежит CpG-островку chr21:18884807–18886111 (GRCh37 hg19).

ломов обеих библиотек. Мы обнаружили среднюю корреляцию профиля метилирования, полученного в присутствии УДГ, с рассчитанным Д. Гохманом (r = 0.57); паттерны метилирования представлены на рис. 2. Однако метилом, полученный без обработки УДГ, имеет слабую корреляцию с опубликованным (r = 0.14), и паттерны метилирования совпадают в основном в деметилированных CpG-островках независимо от того, проводим мы последующее сглаживание бегущим средним или нет.

Далее мы обработали с помощью нашего пайплайна восемь геномов охотников-собирателей, для которых ранее реконструкция профилей метилирования не проводилась (см. табл. 1). Полученные профили в целом имели похожую на другие древние метиломы форму, в том числе проявляли полное деметилирование некоторых CpG-островков (рис. 3), схожее с профилем ранее реконструированного усть-ишимского охотника-собирателя (Gokhman et al., 2020). Несмотря на то что образец PES001 не был обработан УДГ при подготовке библиотек, полученный нами профиль метилирования имел тенденции, схожие с прочими профилями охотников-собирателей, поэтому не был исключен из дальнейшего анализа.

Как утверждают авторы метода, реконструированные с помощью DamMet профили метилирования могут быть использованы для прямого сравнения с современными данными, однако метилирование может различаться в клетках разного происхождения, и, следовательно, прямому сравнению подлежат только профили метилирования, полученные из одних тканей. Насколько нам известно, секвенирование костных тканей в контексте ожирения не проводилось, поэтому для конечного сравнения мы выбрали образцы адипоцитов подкожной клетчатки и висцерального жира, которые имеют схожие паттерны метилирования. Однако эти паттерны могут серьезно отличаться от тех, что демонстрируют кости и другие ткани мезодермального происхождения, поэтому мы разработали скрипт на языке Python, который осуществляет поиск дифференциально метилированных позиций в мезодермальных тканях и исключает их из дальнейшего анализа. В основу лег метод дисперсионного анализа в трех группах с последующей проверкой попарными сравнениями и поправкой на множественные сравнения. Образцы мезодермального происхождения были разделены на группы в соответствии с типом ткани: фибробласты, мышечные клетки и остеобласты. Всего исследовали 26.5 миллиона CpG-позиций, из них около 206 тысяч имели дифференциальное метилирование хотя бы в одной из групп, тогда как более 26 миллионов не проявляли значимых различий.

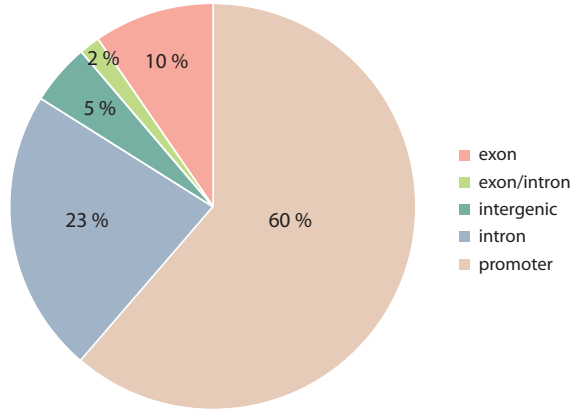


Рис. 4. Процентное соотношение ДМР в различных геномных областях.

Мы провели поиск ДМР для современных образцов костных тканей, но при этом искали изменения только в 642 регионах, для которых в литературе было показано дифференциальное метилирование в контексте ожирения. В данном случае мы проводили попозиционный ANOVA-анализ для групп древних людей, охотников-собирателей

и современных людей (костная ткань) с предварительной фильтрацией нетканеспецифичных CpG-сайтов. Было идентифицировано 38 ДМР, в которых пересечение хотя бы с одним из обозначенных 642 регионов составило бы не менее 20 CpG-сайтов. Как видно на рис. 4, около 60 % найденных ДМР лежат в промоторных областях генов, 35 % принадлежат ДМР в теле гена, и лишь 5 % ДМР находятся в межгенных областях. 94 % этих ДМР гиперметилированы, что способствует подавлению экспрессии генов, ассоциированных с ожирением.

Дополнительные данные и программный код

Профили метилирования древних людей и скрипты на языке Python, использованные для анализа в данной работе, доступны в репозитории GitHub: https://github.com/bor-d/ancDMR

Заключение

В настоящее время существует несколько методов восстановления профилей метилирования древних организмов, из которых в основном используются epiPALEOMIX (Hanghøj et al., 2016) и DamMet (Hanghøj et al., 2019). Оба эти метода, согласно исследованиям, демонстрируют значительную точность, но ограничены качеством древних образцов. Для нашего анализа мы остановились именно на втором методе, так как авторы при разработке учитывали возможность сравнения полученных значений метилирования с профилями, для которых использовались другие технологии секвенирования, что сделало метод универсальным. Однако при валидации пайплайна мы заметили, что значения метилирования, полученные с помощью DamMet, довольно сильно отклоняются от опубликованных ранее (Gokhman et al., 2014, 2020). Разработчики DamMet пишут, что их инструмент имеет свойство занижать значения метилирования по сравнению с профилями epiPALEOMIX, который не учитывает SNV, ошибки секвенирования и дезаминирование неметилированных цитозинов, что мы и наблюдали при реконструкции профилей неандертальца и денисовца. В то же время полученная нами корреляция восстановленных с помощью DamMet значений метилирования с ранее опубликованными имеет положительные значения, как и при сравнении, проведенном авторами DamMet. Соответственно, инструмент эффективно работает и для исследования ранее не восстановленных профилей, которые впоследствии действительно могут использоваться для дальнейшего сравнения с современными метиломами.

В качестве демонстрации разработанного нами пайплайна мы осуществили поиск ДМР в геномах охотников-собирателей и современных людей в контексте ожирения, выявив 38 регионов, среди которых около двух третей находится в промоторных областях, а значит, изменение метилирования промоторов данных генов может быть ассоциировано с их экспрессией. Шаги, на которые мы разделили пайплайн, разрешают проблемы, с которыми может столкнуться исследователь, обнаруживший дефицит опубликованных профилей метилирования нужных тканей в контексте исследуемого им состояния, поскольку они позволяют снизить вероятность получения ложноположительных ДМР, обусловленных тканеспецифичностью. Для применения данного пайплайна при поиске ДМР, ассоциированных с другим заболеванием, исследователю потребуется провести анализ соответствующей литературы и выбрать регионы, метилирование которых связано с интересующим его заболеванием. Однако, несмотря на исключение тканеспецифичных регионов и фильтрацию по регионам, ассоциированным с болезнью, в процессе поиска ДМР все еще будут принимать участие CpG-сайты, метилирование которых было изменено в ходе эволюционного перехода от древних людей (*Homo sapiens neandertaliensis*) к людям современного типа (*Homo sapiens sapiens*).

Список литературы / References

Angermueller C., Lee H.J., Reik W., Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol*. 2017;18(1):67. DOI 10.1186/s13059-017-1189-z

Bock C., Reither S., Mikeska T., Paulsen M., Walter J., Lengauer T. BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*. 2005;21(21):4067-4068. DOI 10.1093/bioinformatics/bti652

Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. DOI 10.1093/bioinformatics/btu170

Briggs A.W., Stenzel U., Johnson P.L.F., Green R.E., Kelso J., Prüfer K., Meyer M., Krause J., Ronan M.T., Lachmann M., Pääbo S. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. USA*. 2007;104(37):14616-14621. DOI 10.1073/pnas.0704665104

Briggs A.W., Good J.M., Green R.E., Krause J., Maricic T., Stenzel U., Lalueza-Fox C., Rudan P., Brajković D., Kućan Ž., Gušić I., Schmitz R., Doronichev V.B., Golovanova L.V., de la Rasilla M., Fortea J., Rosas A., Pääbo S. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science*. 2009a;325(5938):318-321. DOI 10.1126/science.1174462

Briggs A.W., Good J.M., Green R.E., Krause J., Maricic T., Stenzel U., Pääbo S. Primer extension capture: targeted sequence retrieval from heavily degraded DNA sources. *J. Vis. Exp*. 2009b;31:1573. DOI 10.3791/1573

Briggs A.W., Stenzel U., Meyer M., Krause J., Kircher M., Pääbo S. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res*. 2010;38(6):e87. DOI 10.1093/nar/gkp1163

Clark S.J., Harrison J., Paul C.L., Frommer M. High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*. 1994;22(15):2990-2997. DOI 10.1093/nar/22.15.2990

Feinberg A.P., Irizarry R.A. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. USA*. 2010;107(Suppl.1):1757-1764. DOI 10.1073/pnas.0906183107

Fu Q., Li H., Moorjani P., Jay F., Slepchenko S.M., Bondarev A.A., Johnson P.L.F., Aximu-Petri A., Prüfer K., de Filippo C., Meyer M., Zwyns N., Salazar-García D.C., Kuzmin Y.V., Keates S.G., Kosintsev P.A., Razhev D.I., Richards M.P., Peristov N.V., Lachmann M., Douka K., Higham T.F.G., Slatkin M., Hublin J.J., Reich D., Kelso J., Viola T.B., Pääbo S. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*. 2014;514(7523):445-449. DOI 10.1038/nature13810

Gansauge M.-T., Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc*. 2013;8(4):737-748. DOI 10.1038/nprot.2013.038

Gokhman D., Lavi E., Prüfer K., Fraga M.F., Riancho J.A., Kelso J., Pääbo S., Meshorer E., Carmel L. Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*. 2014;344(6183):523-527. DOI 10.1126/science.1250368

Gokhman D., Nissim-Rafinia M., Agranat-Tamir L., Housman G., García-Pérez R., Lizano E., Cheronet O., Mallick S., Nieves-Colón M.A., Li H., Alpaslan-Roodenberg S., Novak M., Gu H., Osinski J.M., Ferrando-Bernal M., Gelabert P., Lipende I., Mjungu D., Kondova I., Bontrop R., Kullmer O., Weber G., Shahar T., Dvir-Ginzberg M., Faerman M., Quillen E.E., Meissner A., Lahav Y., Kandel L., Liebergall M., Prada M.E., Vidal J.M., Gronostajski R.M., Stone A.C., Yakir B., Lalueza-Fox C., Pinhasi R., Reich D., Marques-Bonet T., Meshorer E., Carmel L. Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun*. 2020; 11(1):1189. DOI 10.1038/s41467-020-15020-6

Gu H., Smith Z.D., Bock C., Boyle P., Gnirke A., Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc*. 2011;6(4): 468-481. DOI 10.1038/nprot.2010.190

Günther T., Malmström H., Svensson E.M., Omrak A., Sánchez-Quinto F., Kılınç G.M., Krzewińska M., Eriksson G., Fraser M., Edlund H., Munters A.R., Coutinho A., Simões L.G., Vicente M., Sjölander A., Sellevold B.J., Jørgensen R., Claes P., Shriver M.D., Valdiosera C., Netea M.G., Apel J., Lidén K., Skar B., Storå J., Götherström A., Jakobsson M. Population genomics of Mesolithic Scandinavia: investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol*. 2018;16(1):e2003703. DOI 10.1371/journal.pbio.2003703

Hanghøj K., Seguin-Orlando A., Schubert M., Madsen T., Pedersen J.S., Willerslev E., Orlando L. Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Mol. Biol. Evol*. 2016;33(12):3284-3298. DOI 10.1093/molbev/msw184

Hanghøj K., Renaud G., Albrechtsen A., Orlando L. DamMet: ancient methylome mapping accounting for errors, true variants, and post-mortem DNA damage. *GigaScience*. 2019;8(4):giz025. DOI 10.1093/gigascience/giz025

Jablonka E., Raz G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol*. 2009;84(2):131-176. DOI 10.1086/598822

Jun G., Wing M.K., Abecasis G.R., Kang H.M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res*. 2015;25(6): 918-925. DOI 10.1101/gr.176552.114

Krueger F., Andrews S. RBismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27(11): 1571-1572. DOI 10.1093/bioinformatics/btr167

Loyfer N., Magenheim J., Peretz A., Cann G., Bredno J., Klochendler A., Fox-Fisher I., Shabi-Porat S., Hecht M., Pelet T., Moss J., Drawshy Z., Amini H., Moradi P., Nagaraju S., Bauman D., Shveiky D., Porat S., Dior U., Rivkin G., Or O., Hirshoren N., Carmon E., Pikarsky A., Khalaileh A., Zamir G., Grinbaum R., Gazala M.A., Mizrahi I., Shussman N., Korach A., Wald O., Izhar U., Erez E., Yutkin V., Samet Y., Golinkin D.R., Spalding K.L., Druid H., Arner P., Shapiro A.M.J., Grompe M., Aravanis A., Venn O., Jamshidi A., Shemer R., Dor Y., Glaser B., Kaplan T. A DNA methylation atlas of normal human cell types. *Nature*. 2023;613(7943):355-364. DOI 10.1038/s41586-022-05580-6

Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., de Filippo C., Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. A high-coverage genome sequence from an archaic Denisovan individual. *Science*. 2012;338(6104):222-226. DOI 10.1126/science.1224344

Moreno-Mayar J., Potter B., Vinner L., Steinrücken M., Rasmussen S., Terhorst J., Kamm J., Albrechtsen A., Malaspinas A., Sikora M., Reuther J., Irish J., Malhi R., Orlando L., Song Y., Nielsen R., Meltzer D., Willerslev E. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*. 2018a; 553(7687):203-207. DOI 10.1038/nature25173

Moreno-Mayar J.V., Vinner L., Damgaard P.B., de la Fuente C., Chan J., Spence J.P., Allentoft M.E., Vimala T., Racimo F., Pinotti T., Rasmussen S., Margaryan A., Orbegozo M.I., Mylopotamitaki D., Wooller M., Bataille C., Becerra-Valdivia L., Chivall D., Comeskey D., Devièse T., Grayson D.K., George L., Harry H., Alexandersen V., Primeau C., Erlandson J., Rodrigues-Carvalho C., Reis S., Bastos M.Q.R., Cybulski J., Vullo C., Morello F., Vilar M., Wells S., Gregersen K., Hansen K.L., Lynnerup N., Mirazón Lahr M., Kjær K., Strauss A., Alfonso-Durruty M., Salas A., Schroeder H., Higham T., Malhi R.S., Rasic J.T., Souza L., Santos F.R., Malaspinas A.-S., Sikora M., Nielsen R., Song Y.S., Meltzer D.J., Willerslev E. Early human dispersals within the Americas. *Science.* 2018b;362(6419). DOI 10.1126/science.aav2621

Niiranen L., Leciej D., Edlund H., Bernhardsson C., Fraser M., Sánchez Quinto F., Herzig K.H., Jakobsson M., Walkowiak J., Thalmann O. Epigenomic modifications in modern and ancient genomes. *Genes*. 2022;13(2):178. DOI 10.3390/genes13020178

Ohm J.E., Mali P., Van Neste L., Berman D.M., Liang L., Pandiyan K., Briggs K.J., Zhang W., Argani P., Simons B., Yu W., Matsui W., Van Criekinge W., Rassool F.V., Zambidis E., Schuebel K.E., Cope L., Yen J., Mohammad H.P., Cheng L., Baylin S.B. Cancer-related epigenome changes associated with reprogramming to induced pluripotent stem cells. *Cancer Res*. 2010;70(19):7662-7673. DOI 10.1158/0008-5472.CAN-10-1361

Olova N., Krueger F., Andrews S., Oxley D., Berrens R.V., Branco M.R., Reik W. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* 2018;19(1):33. DOI 10.1186/s13059-018-1408-2

Orlando L., Gilbert M.T.P., Willerslev E. Reconstructing ancient genomes and epigenomes. *Nat. Rev. Genet*. 2015;16(7):395-408. DOI 10.1038/nrg3935

Pedersen B.S., Schwartz D.A., Yang I.V., Kechris K.J. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *P*-values. *Bioinformatics*. 2012;28(22):2986-2988. DOI 10.1093/bioinformatics/bts545

Poplin R., Ruano-Rubio V., DePristo M.A., Fennell T.J., Carneiro M.O., Van der Auwera G.A., Kling D.E., Gauthier L.D., Levy-Moonshine A., Roazen D., Shakir K., Thibault J., Chandran S., Whelan C., Lek M., Gabriel S., Daly M.J., Neale B., MacArthur D.G., Banks E. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2017. DOI 10.1101/201178

Prüfer K., Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P.H., de Filippo C., Li H., Mallick S., Dannemann M., Fu Q., Kircher M., Kuhlwilm M., Lachmann M., Meyer M., Ongyerth M., Siebauer M., Theunert C., Tandon A., Moorjani P., Pickrell J., Mullikin J.C., Vohr S.H., Green R.E., Hellmann I., Blanche H., Cann H., Kitzman J.O., Shendure J., Eichler E.E., Lein E.S., Bakken T.E., Golovanova L.V., Doronichev V.B., Shunkov M.V., Derevianko A.P., Viola B., Slatkin M., Reich D., Kelso J., Pääbo S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 2014;505(7481): 43-49. DOI 10.1038/nature12886

Prüfer K., de Filippo C., Grote S., Mafessoni F., Korlević P., Hajdinjak M., Vernot B., Skov L., Hsieh P., Peyrégne S., Reher D., Hopfe C., Nagel S., Maricic T., Fu Q., Theunert C., Rogers R., Skoglund P., Chintalapati M., Dannemann B., Nelson B.J., Key F.M., Rudan P., Kućan Ž., Gušić I., Golovanova L.V., Doronichev V.B., Patterson N., Reich D., Eichler E.E., Slatkin M., Schierup M.H., Andrés A.M., Kelso J., Meyer M., Pääbo S. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*. 2017;358(6363):655-658. DOI 10.1126/science.aao1887

Saag L., Vasilyev S.V., Varul L., Kosorukova N.V., Gerasimov D.V., Oshibkina S.V., Griffith S.J., Solnik A., Saag L., D'Atanasio E., Metspalu E., Reidla M., Rootsi S., Kivisild T., Scheib C.L., Tambets K., Kriiska A., Metspalu M. Genetic ancestry changes in Stone to Bronze Age transition in the East European plain. *Sci. Adv*. 2021;7:eabd6535. DOI 10.1126/sciadv.abd6535

Sawyer S., Krause J., Guschanski K., Savolainen V., Pääbo S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*. 2012;7(3):e34131. DOI 10.1371/journal.pone.0034131

Seguin-Orlando A., Donat R., Der Sarkissian C., Southon J., Thèves C., Manen C., Tchérémissinoff Y., Crubézy E., Shapiro B., Deleuze J., Dalén L., Guilaine J., Orlando L. Heterogeneous hunter-gatherer and steppe-related ancestries in Late Neolithic and Bell Beaker genomes from present-day France. *Curr. Biol.* 2021;31(5):1072-1083. DOI 10.1016/j.cub.2020.12.015

Sikora M., Seguin-Orlando A., Sousa V.C., Albrechtsen A., Korneliussen T., Ko A., Rasmussen S., Dupanloup I., Nigst P.R., Bosch M.D., Renaud G., Allentoft M.E., Margaryan A., Vasilyev S.V., Veselovskaya E.V., Borutskaya S.B., Deviese T., Comeskey D., Higham T., Manica A., Foley R., Meltzer D.J., Nielsen R., Excoffier L., Lahr M.M., Orlando L., Willerslev E. Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. *Science*. 2017;358(6363):659-662. DOI 10.1126/science.aao1807

Suzuki M., Liao W., Wos F., Johnston A.D., DeGrazia J., Ishii J., Bloom T., Zody M.C., Germer S., Greally J.M. Whole-genome bisulfite sequencing with improved accuracy and cost. *Genome Res.* 2018;28(9):1364-1371. DOI 10.1101/gr.232587.117

Zhur K.V., Trifonov V.A., Prokhortchouk E.B. Progress and prospects in epigenetic studies of ancient DNA. *Biochemistry* (*Mosc.*). 2021; 86(12-13):1563-1571. DOI 10.1134/S0006297921120051

Zou L.S., Erdos M.R., Taylor D.L., Chines P.S., Varshney A., Parker S.C.J., Collins F.S., Didion J.P. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics*. 2018;19(1):390. DOI 10.1186/s12864-018-4766-y

ORCID ID

D.D. Borodko orcid.org/0000-0003-3596-5470
S.V. Zhenilo orcid.org/0000-0003-0874-1594
F.S. Sharko orcid.org/0000-0002-1189-5597

Благодарности. Работа выполнена при финансовой поддержке проекта Минобрнауки России, системный номер № 075-10-2020-116 (номер гранта 13.1902.21.0023).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 15.07.2023. После доработки 04.10.2023. Принята к публикации 05.10.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):829-838
DOI 10.18699/VJGB-23-96

Перевод на английский язык https://vavilov.elpub.ru/jour

Анализ особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, на основе индексов филостратиграфического возраста и микроэволюционной изменчивости

Е.В. Игнатьева ✉, С.А. Лашин, З.С. Мустафин, Н.А. Колчанов

Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ eignat@bionet.nsc.ru

Аннотация. Гены рецепторов клеточной поверхности составляют существенную долю генома человека (более тысячи генов) и выполняют важную роль в генных сетях. Рецепторы клеточной поверхности – это трансмембранные белки, которые взаимодействуют с различными молекулами (лигандами), находящимися во внеклеточном пространстве, что приводит к активации путей сигнальной трансдукции в клетке. Для рецепторов клеточной поверхности известно большое количество экзогенных лигандов различного происхождения, включая лекарственные препараты, что и определяет интерес к их исследованию с точки зрения биомедицины. Аппетит (стремление животного организма потреблять пищу) – один из самых примитивных инстинктов, способствующих выживанию. Однако приобретенный в ходе эволюции механизм приспособления к неблагоприятным факторам в условиях стабильного поступления питательных веществ оказался избыточным, в связи с чем ожирение стало одной из самых серьезных проблем общественного здравоохранения в XXI веке. Патологические состояния человека, характеризующиеся нарушениями аппетита, включают как гиперфагию, неминуемо приводящую к ожирению, так и нервную анорексию, индуцированную психосоциальными стимулами, и снижение аппетита, связанное с воспалительными, нейродегенеративными и онкологическими заболеваниями. Понимание эволюционных механизмов развития болезней человека, особенно связанных с изменениями образа жизни, произошедшими в течение последних 100–200 лет, имеет как фундаментальное, так и прикладное значение. Особенно важно установить взаимосвязи между эволюционными характеристиками генов в генных сетях и устойчивостью этих сетей к изменениям, вызванным мутациями. Цель данной работы – выявление особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, с использованием филостратиграфического индекса PAI (phylostratigraphic age index) и индекса эволюционной изменчивости DI (divergence index). Были проанализированы индексы PAI и DI для 64 генов человека, кодирующих рецепторы клеточной поверхности, ортологи которых участвовали в регуляции аппетита у модельных видов животных. Оказалось, что в рассматриваемом наборе генов содержится повышенное количество генов, имеющих одинаковый филостратиграфический возраст (PAI = 5, этап дивергенции позвоночных), и почти все эти гены (28 из 31) относятся к суперсемейству рецепторов, сопряженных с G-белком. По-видимому, синхронизированное эволюционирование такой многочисленной группы генов (31 из 64 генов) связано с формированием у первых позвоночных мозга как отдельного органа. При исследовании распределения генов из этого же набора по значениям индексов DI была выявлена существенная обогащенность генами с низким DI. При этом восемь генов (*GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1*) характеризовались экстремально низким значением DI (менее 0.05), что указывает на существенную их подверженность стабилизирующему отбору. Обнаружено также, что группа генов с низким DI обогащена генами, тканеспецифически экспрессирующимися в мозге. В частности, к группе генов, тканеспецифически экспрессирующихся в мозге, относится *GPR26*, имеющий самое низкое значение DI. Ввиду того, что эндогенный лиганд для рецептора GPR26 пока не выявлен, этот ген представляется чрезвычайно интересным объектом для дальнейшего теоретического и экспериментального исследования. Выявленные нами особенности распределения генов рецепторов клеточной поверхности по эволюционным индексам PAI и DI являются отправной точкой для дальнейшего анализа эволюционных характеристик генной сети регуляции аппетита в целом.

Ключевые слова: регуляция аппетита; рецепторы клеточной поверхности; чувство голода; эволюция; филостратиграфия; возраст гена; изменчивость генов.

Для цитирования: Игнатьева Е.В., Лашин С.А., Мустафин З.С., Колчанов Н.А. Анализ особенностей эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, на основе индексов филостратиграфического возраста и микроэволюционной изменчивости. *Вавиловский журнал генетики и селекции*. 2023;27(7):829-838. DOI 10.18699/VJGB-23-96

Evolution of human genes encoding cell surface receptors involved in the regulation of appetite: an analysis based on the phylostratigraphic age and divergence indexes

E.V. Ignatieva ✉, S.A. Lashin, Z.S. Mustafin, N.A. Kolchanov

Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ eignat@bionet.nsc.ru

Abstract. Genes encoding cell surface receptors make up a significant portion of the human genome (more than a thousand genes) and play an important role in gene networks. Cell surface receptors are transmembrane proteins that interact with molecules (ligands) located outside the cell. This interaction activates signal transduction pathways in the cell. A large number of exogenous ligands of various origins, including drugs, are known for cell surface receptors, which accounts for interest in them from biomedical researchers. Appetite (the desire of the animal organism to consume food) is one of the most primitive instincts that contribute to survival. However, when the supply of nutrients is stable, the mechanism of adaptation to adverse factors acquired in the course of evolution turned out to be excessive, and therefore obesity has become one of the most serious public health problems of the twenty-first century. Pathological human conditions characterized by appetite violations include both hyperphagia, which inevitably leads to obesity, and anorexia nervosa induced by psychosocial stimuli, as well as decreased appetite caused by neurodegeneration, inflammation or cancer. Understanding the evolutionary mechanisms of human diseases, especially those related to lifestyle changes that have occurred over the past 100–200 years, is of fundamental and applied importance. It is also very important to identify relationships between the evolutionary characteristics of genes in gene networks and the resistance of these networks to changes caused by mutations. The aim of the current study is to identify the distinctive features of human genes encoding cell surface receptors involved in appetite regulation using the phylostratigraphic age index (PAI) and divergence index (DI). The values of PAI and DI were analyzed for 64 human genes encoding cell surface receptors, the orthologs of which were involved in the regulation of appetite in model animal species. It turned out that the set of genes under consideration contains an increased number of genes with the same phylostratigraphic age (PAI = 5, the stage of vertebrate divergence), and almost all of these genes (28 out of 31) belong to the superfamily of G-protein coupled receptors. Apparently, the synchronized evolution of such a large group of genes (31 genes out of 64) is associated with the development of the brain as a separate organ in the first vertebrates. When studying the distribution of genes from the same set by DI values, a significant enrichment with genes having a low DIs was revealed: eight genes (*GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1*) had extremely low DIs (less than 0.05). Such low values of DI indicate that these genes are very likely to be undergoing stabilizing selection. It was also found that the group of genes with low DIs was enriched with genes that had brain-specific patterns of expression. In particular, *GPR26*, which had the lowest DI, is in the group of brain-specific genes. Because the endogenous ligand for the GPR26 receptor has not yet been identified, this gene seems to be an extremely interesting object for further theoretical and experimental research. We believe that the features of the genes encoding cell surface receptors we have identified using the evolutionary metrics PAI and DI can be a starting point for further evolutionary analysis of the gene network regulating appetite.
Key words: regulation of appetite; cell surface receptors; hunger; evolution; phylostratigraphic analysis; gene age; gene variability.

For citation: Ignatieva E.V., Lashin S.A., Mustafin Z.S., Kolchanov N.A. Evolution of human genes encoding cell surface receptors involved in the regulation of appetite: an analysis based on the phylostratigraphic age and divergence indexes. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):829-838. DOI 10.18699/VJGB-23-96

Введение

Аппетит (стремление животного организма потреблять пищу) – это физиологический механизм (ощущение), регулирующий поступление питательных веществ. Стремление потреблять пищу – один из самых примитивных инстинктов, способствующих выживанию. Этот инстинкт формировался на протяжении миллионов лет эволюции живых существ и обеспечил мощные механизмы для адаптации и реагирования на периоды нехватки питательных веществ (Yeo, Heisler, 2012). Способность потреблять избыточные количества пищи в периоды ее доступности существенно влияла на выживание особей как в популяциях человека, так и в популяциях других видов животных.

С развитием человеческой цивилизации население ряда развитых стран столкнулось с проблемой адаптации к изобилию продуктов питания в сочетании со снижением физической активности, в связи с чем ожирение стало серьезной проблемой общественного здравоохранения в XXI веке (Kaidar-Person et al., 2011). Таким образом, приобретенный в ходе эволюции механизм приспособления к неблагоприятным факторам в условиях стабильного поступления питательных веществ оказался избыточным (Yeo, Heisler, 2012).

Система регуляции аппетита человека и других видов животных функционирует при участии белковых продуктов генов, экспрессирующихся как в мозге (Olszewski et al., 2008), так и в периферических органах и тканях:

желудке, кишечнике, поджелудочной железе, жировой ткани. Нейроны, участвующие в регуляции мотивационного стремления для получения пищи, расположены в различных отделах головного мозга (ядрах гипоталамуса, амигдале, дорсальном ядре шва, ядрах солитарного тракта, вентральной тегментальной области, префронтальной коре и т. д.). Они интегрируют сигналы, полученные от органов чувств (обоняние, зрение, вкусовые ощущения), а также различные интероцептивные и гуморальные сигналы и управляют поведением, направленным на поиск и потребление пищи (Yeo, Heisler, 2012; Tremblay, Bellisle, 2015; Heisler, Lam, 2017).

Аппетит может быть вызван дефицитом энергии и питательных веществ, в англоязычной литературе для обозначения этого понятия принят термин «гомеостатический аппетит» (homeostatic appetite). Однако даже в отсутствие недостатка калорий такие факторы, как вид, запах и вкус пищи, сигналы окружающей среды, ожидание ощущений, связанных с едой, могут стимулировать пищевое поведение, т. е. негомеостатический аппетит (non-homeostatic appetite). Системы нейронов, управляющих гомеостатическим и негомеостатическим аппетитом, функционируют в тесной кооперации (Ahn et al., 2022).

Центральным звеном систем регуляции как гомеостатического, так и негомеостатического аппетита являются нейроны аркуатного ядра гипоталамуса, секретирующие нейропептид Y (NPY) и агутиподобный белок (AgRP), а также альфа-меланоцитстимулирующий гормон (α-MSH), который образуется из проопиомеланокортина (POMC) под действием прогормон-конвертаз (PCSK1 и PCSK2) (Yeo, Heisler, 2012). Активность нейронов аркуатного ядра гипоталамуса контролируется гормонами (лептином, инсулином, грелином, полипептидом YY (PYY), глюкокортикоидами, адренокортикотропином, кортикотропин-релизинг гормоном), нейромедиаторными системами мозга (серотонергическая, дофаминергическая, адреналиновая, ГАМК-ергическая), а также нейротрофическими факторами (BDNF и др.) (Maniam, Morris, 2012; Yeo, Heisler, 2012; Heisler, Lam, 2017).

Известны патологические состояния человека, характеризующиеся нарушениями аппетита. Патологическое увеличение массы тела (ожирение) может быть вызвано таким состоянием, как гиперфагия (булимия). Катастрофическое снижение аппетита наблюдается при нервной анорексии, которая чрезвычайно опасна и повышает риск смерти у молодых людей в 10 раз (Fichter, Quadflieg, 2016). Снижение аппетита может наблюдаться при хронических воспалительных и аутоиммунных процессах, онкологических и нейродегенеративных заболеваниях (Grossberg et al., 2010). В этом контексте любые новые знания о системе генов, регулирующих аппетит, приобретают особую значимость.

Ранее нами был проведен функциональный анализ генов, контролирующих пищевое поведение и массу тела (Игнатьева и др., 2014; Ignatieva et al., 2016). При анализе выборки из 105 генов, участвующих в регуляции аппетита, была обнаружена неслучайно частая встречаемость генов, специфически экспрессирующихся в мозге. Также было выявлено, что примерно 45 % выборки составляют гены, кодирующие рецепторы клеточной поверхности. Многие из этих рецепторов относятся к суперсемейству рецепторов, сопряженных с G-белком (G-protein-coupled receptors, GPCR).

Суперсемейство рецепторов, сопряженных с G-белком, представлено белками, имеющими сходное строение (у всех имеется семь трансмембранных доменов) и присутствующими в клетках практически всех эукариот (New, Wong, 1998; Yang et al., 2021). По данным компьютерного анализа, геном человека содержит около 800 генов, кодирующих белки этого суперсемейства (включая 388 генов, кодирующих ольфакторные рецепторы) (Bjarnadóttir et al., 2006). Рецепторы из суперсемейства GPCR опосредуют ответ клеток на внеклеточные сигнальные молекулы различной природы – белковой, пептидной, низкомолекулярные вещества (запаховые и вкусовые стимулы, гормоны), а также фотоны света. В свою очередь, эти рецепторы активируют в клетках пути сигнальной трансдукции, обеспечивая фундаментальные физиологические процессы: зрение, восприятие вкусовых и запаховых сигналов, функционирование клеток нервной системы, эндокринную регуляцию и процессы размножения (Katritch et al., 2013). К числу наиболее известных рецепторов из суперсемейства GPCR, включенных нами ранее в выборку генов, регулирующих аппетит (Igatieva et al., 2016), относятся, например, GHSR (growth hormone secretagogue receptor), MC3R (melanocortin 3 receptor), MC4R (melanocortin 4 receptor), CCKAR (cholecystokinin A receptor), CCKBR (cholecystokinin B receptor) и GCGR (glucagon).

Понимание эволюционных механизмов развития болезней человека, особенно связанных с изменениями образа жизни, произошедшими в течение последних 100–200 лет (а упомянутые выше болезни, обусловленные нарушениями регуляции аппетита, именно таковыми и являются), имеет как фундаментальное, так и прикладное значение. Особенно важно установить взаимосвязи между эволюционными характеристиками генов в генных сетях и устойчивостью этих сетей к изменениям как в самих генах (посредством мутаций), так и в паттернах их экспрессии, обусловленными, например, вариациями в сайтах связывания транскрипционных факторов. Филогенетический и популяционный анализ генов и генных сетей, регулирующих соответствующие биологические процессы, может быть полезным при разработке новых сценариев персонализированной профилактики и таргетной лекарственной терапии заболеваний.

Цель данной работы – выявить особенности эволюции генов рецепторов клеточной поверхности человека, участвующих в регуляции аппетита, с использованием филостратиграфического индекса PAI (phylostratigraphic age index) и индекса эволюционной изменчивости DI (divergence index). Для достижения этой цели на первом этапе на основе анализа научных публикаций была сформирована выборка генов человека, кодирующих рецепторы, ортологи которых участвовали в регуляции аппетита у модельных видов животных. Далее были рассмотрены распределения генов человека по величинам индексов PAI и DI. Характеристические особенности этих распределений выявляли путем сравнения с распределениями, полученными для всех белок-кодирующих генов человека, а также генов, кодирующих белки суперсемейства GPCR.

Таблица 1. Выборки генов человека, для которых проводился анализ распределений индексов PAI и DI

Название выборки	Описание выборки	Количество генов
allCDS_19,566	Все белок-кодирующие гены генома человека, для которых известны значения индексов PAI и DI	19566
Receptors_64	Гены человека, кодирующие рецепторы клеточной поверхности и участвующие в регуляции аппетита*	64
allGPCR_389	Гены человека, кодирующие GPCR (включены все гены, представленные в базе GPCRdb (https://gpcrdb.org), за исключением генов, кодирующих ольфакторные рецепторы)	389
appGPCR_45	Гены из выборки *Receptors_64*, кодирующие GPCR	45
app_not_GPCR_19	Гены из выборки *Receptors_64*, кодирующие рецепторы, не принадлежащие суперсемейству рецепторов, связанных с G-белком	19

* Выборка включает гены человека, ортологичные генам других видов животных, роль которых в регуляции аппетита исследована экспериментально.

Материалы и методы

Формирование списка генов, кодирующих рецепторы клеточной поверхности и участвующих в регуляции аппетита. Список генов был взят из опубликованной ранее работы (Ignatieva et al., 2016) и дополнен на основе поиска по PubMed (https://pubmed.ncbi.nlm.nih.gov/) с использованием ключевых слов, приведенных в Приложении 1[1]. Рассматривались только гены из экспериментальных работ, публикации обзорного характера из рассмотрения исключались. Почти во всех исследованиях роль генов в регуляции количества потребленной пищи была выявлена на модельных животных (мышах, крысах и т. д.). Поэтому список генов человека, контролирующих аппетит, был сформирован из ортологов генов, выявленных в экспериментах на других видах животных. Информацию о том, что продуктом гена является рецептор клеточной поверхности, получали из текстового поля "Summary" базы EntrezGene (https://www.ncbi.nlm.nih.gov/gene).

Контрольные выборки генов. В анализе использовали выборки генов человека, перечисленные в табл. 1. Список генов человека, кодирующих рецепторы и контролирующих аппетит, получил название *Receptors_64.*

Выборка всех белок-кодирующих генов человека (*allCDS_19,566*) была образована из белок-кодирующих генов, для которых известны индексы PAI и DI. Эта выборка включала 19566 генов.

Выборку генов человека, кодирующих рецепторы из суперсемейства GPCR (*allGPCR_389*), формировали на основе данных базы GPCRdb (https://gpcrdb.org) (Pandy-Szekeres et al., 2023). В выборку не вошли гены, кодирующие ольфакторные рецепторы, так как рецепторы этого типа не были представлены в сформированной нами выборке генов рецепторов клеточной поверхности, контролирующих аппетит.

Выборку генов, кодирующих GPCR и контролирующих аппетит (*appGPCR_45*), получали путем пересечения сформированных ранее выборок *Receptors_64* и *allGPCR_389*.

Анализ эволюционных характеристик генов осуществляли с использованием индексов PAI (phylostratigraphic age index) и DI (divergence index).

[1] Приложения 1–13 см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx27.pdf

Таблица 2. Соответствие между значениями индексов PAI и таксономическими единицами, датирующими филостратиграфический возраст генов

PAI	Таксон
0	Cellular Organisms (клеточные организмы, корень дерева)
1	Eukaryota (эукариоты)
2	Metazoa (многоклеточные животные)
3	Chordata (хордовые)
4	Craniata (плеченогие)
5	Vertebrata (позвоночные)
6	Euteleostomi (костные позвоночные)
7	Mammalia (млекопитающие)
8	Eutheria (плацентарные)
9	Euarchontoglires (грызунообразные + эуархонты)
10	Primates (приматы)
11	Haplorrhini (обезьяны)
12	Catarrhini (узконосые обезьяны)
13	Hominidae (гоминиды)
14	Homo (люди)
15	Homo sapiens (человек разумный)

Индекс PAI – филостратиграфический индекс гена. Показывает, в какой степени отдален от корня филогенетического дерева таксон, отражающий возраст гена, т. е. такой таксон, на котором произошло расхождение исследуемого вида с наиболее отдаленным родственным таксоном, в котором обнаружен ортолог рассматриваемого гена (табл. 2). Чем больше PAI, тем моложе исследуемый ген. Величины PAI были рассчитаны в программе Orthoscape на основе сервиса KEGG Orthology, как описано в (Мустафин и др., 2021). Нами использовались индексы PAI, рассчитанные при уровне сходства 0.5.

Индекс DI является индексом эволюционной изменчивости гена. Он вычисляется из отношения dN/dS, где dN – доля несинонимичных замен в последовательностях исследуемого гена и его ортолога; dS – доля синонимичных

замен. Значение данного индекса вычислялось на основе сравнения генов человека с генами близкородственных организмов из семейства гоминид, как описано в работе (Мустафин и др., 2021). Таким образом, DI может быть определен только для белок-кодирующих генов и показывает тип отбора, которому подвержен ген. Значение индекса в диапазоне от 0 до 1 свидетельствует о том, что ген подвержен стабилизирующему отбору, 1 – нейтральной эволюции, а больше 1 – движущему отбору.

Анализ тканеспецифичных характеристик генов. Анализ списка генов на предмет обогащенности генами, экспрессирующимися тканеспецифически в определенном органе либо ткани, осуществляли с помощью информационно-программного ресурса *TSEA tool* (Wells et al., 2015). *TSEA tool* (http://genetics.wustl.edu/jdlab/tsea/) использует данные о тканеспецифической экспрессии генов в 25 различных органах и тканях человека и выявляет группы тканеспецифически экспрессирующихся генов, объем которых значимо превышает ожидаемый по случайным причинам. *TSEA tool* оперирует данными о показателях обогащения тканей продуктами экспрессии генов (SI, specificity index) и соответствующих им значениях параметра *p*-value (pSI), рассчитанных для каждого органа или ткани и для каждого транскрипта в результате анализа данных полнотранскриптомного профилирования (GTEx Consortium, 2015). При значении параметра pSI < 0.01 транскрипт рассматривается как тканеспецифичный (*tissue-specific)* для этой ткани.

Статистический анализ. Оценку статистической значимости различий между количеством генов в подгруппах оценивали с помощью критерия Хи-квадрат.

Результаты

Выборка генов, кодирующих рецепторы клеточной поверхности, и их функциональные характеристики

В результате запросов к PubMed были найдены экспериментальные данные о генах модельных видов животных (мышах, крысах и т. д.), участвующих в регуляции потребления пищи. Используя эти сведения, а также информацию из базы EntrezGene, мы сформировали набор из 64 генов человека, ортологичных генам, выявленным у модельных видов животных, и кодирующих рецепторы клеточной поверхности (выборка *Receptors_64*, см. табл. 1). Полный список приведен в Приложении 2.

При сопоставлении списка из 64 генов с данными базы GPCRdb обнаружено, что белковыми продуктами 45 генов (70.3 %) являлись рецепторы, сопряженные с G-белком (рис. 1, *а*). Подвыборка генов, кодирующих рецепторы из суперсемейства GPCR, поименована в табл. 1 как *appGPCR_45.* Принадлежность рецептора к суперсемейству GPCR указана в Приложении 2. Остальные 19 генов (29.9 %) кодировали рецепторы из других суперсемейств (выборка *app_not_GPCR_19*, см. табл. 1).

Анализ списка генов с помощью информационно-программного ресурса *TSEA tool* (Wells et al., 2015) показал, что выборка *Receptors_64* обогащена генами, специфично экспрессирующимися в мозге. Примерно одна пятая генов (12 генов, или 18.75 %) относится к этой категории (см. рис. 1, *б*, Приложение 3).

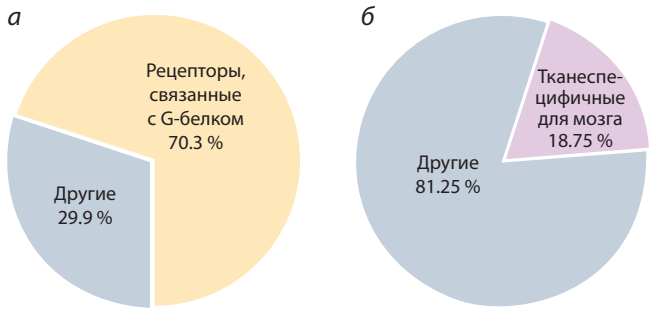


Рис. 1. Функциональные характеристики генов человека, кодирующих рецепторы клеточной поверхности и участвующих в регуляции аппетита (выборка *Receptors_64*).

а – доля генов, кодирующих рецепторы, сопряженные с G-белком; *б* – доля генов, тканеспецифически экспрессирующихся в мозге (тканеспецифичность выявляли с помощью программно-информационного ресурса *TSEA tool*).

Анализ эволюционных характеристик

Филостратиграфический возраст генов (анализ на основе индекса PAI). На первом этапе было определено распределение величин PAI (филостратиграфический возраст) для всех белок-кодирующих генов человека (выборка *allCDS_19,556*). Оказалось, что значения PAI распределены неравномерно (рис. 2, *а*). Третья часть всех генов (33 %) имела PAI, равный нулю (клеточные организмы, корень дерева), а на долю генов с PAI = 5 (этап дивергенции позвоночных) и PAI = 6 (этап дивергенции костных позвоночных) приходилось 17 и 14 % соответственно. При рассмотрении распределения величин PAI для выборки генов человека, кодирующих рецепторы клеточной поверхности и участвующих в регуляции аппетита (выборка *Receptors_64*, Приложение 4), обнаружено, что 31 ген из 64 (т. е. 48 %) имел PAI, равный 5 (этап дивергенции позвоночных) (см. рис. 2, *а*). И это количество статистически значимо отличалось от ожидаемого количества (10.898), которое было рассчитано исходя из распределения, полученного для выборки *allCDS_19,556* ($p < 0.001$, см. рис. 2, *а*, Приложение 5).

Как отмечено выше, большая доля генов из выборки *Receptors_64* (45 из 64 генов) кодирует рецепторы, сопряженные с G-белком (см. рис. 1, *а*). Чтобы выяснить, не связаны ли обнаруженные нами особенности эволюционных характеристик генов из выборки *Receptors_64* с особенностями генов из суперсемейства GPCR, мы проанализировали распределение значений индексов PAI для набора из 389 генов человека, кодирующих GPCR, представленных в базе GPCRdb (https://gpcrdb.org) (выборка *allGPCR_389*). Оказалось, что это распределение также отличается от распределения, полученного для всех белок-кодирующих генов человека (см. рис. 2, *б*). Количество генов в выборке *allGPCR_389*, имеющих PAI, равный 5 (этап дивергенции позвоночных), составляло 39 % (150 генов из 389) и значимо превышало ожидаемое количество, рассчитанное исходя из доли этой группы генов в выборке *allCDS_19,566* (Приложение 6).

Далее мы сравнили распределение по значениям PAI для 45 генов, кодирующих GPCR и регулирующих аппетит (выборка *appGPCR_45*) с распределением для выборки

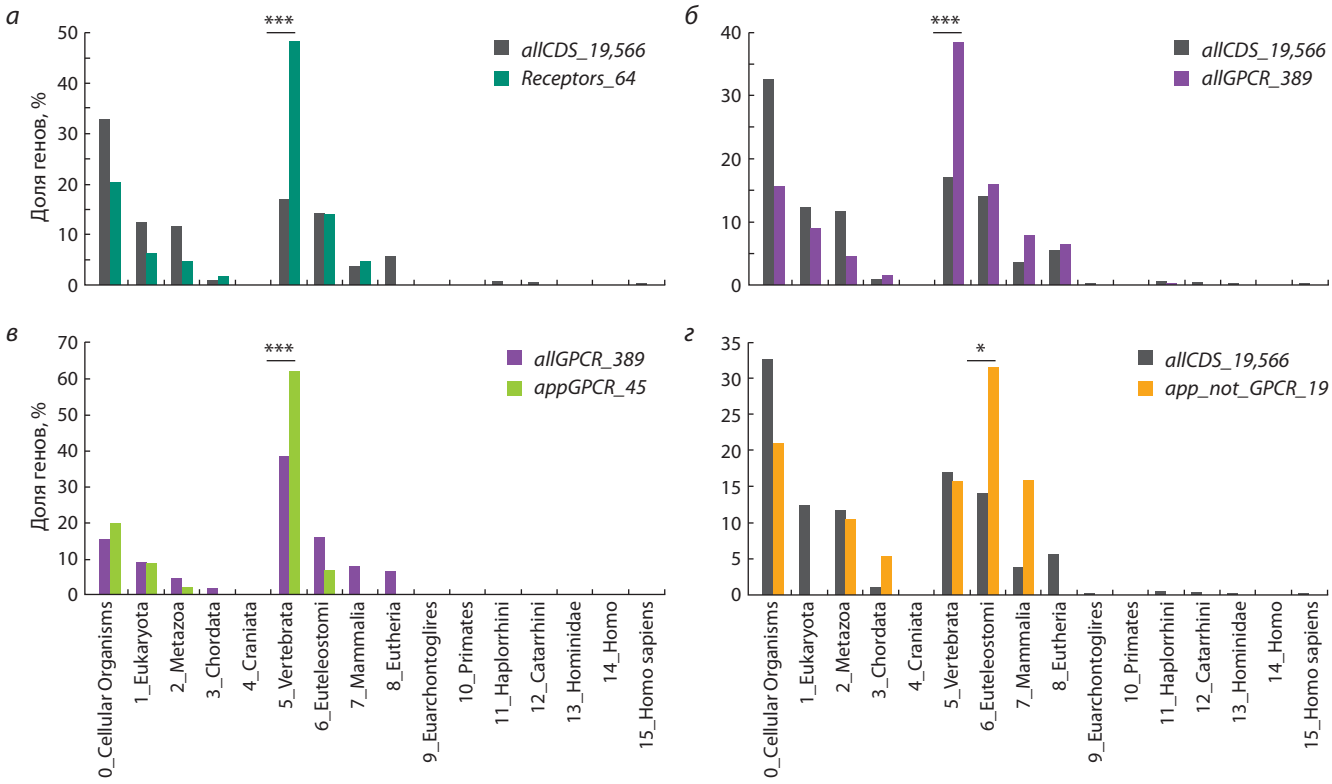


Рис. 2. Распределения белок-кодирующих генов человека из выборок, представленных в табл. 1, по индексу PAI.

а: контрольная выборка – все белок-кодирующие гены человека (*allCDS_19,566*) и гены рецепторов человека, регулирующих аппетит (*Receptors_64*); *б*: контрольная выборка – все белок-кодирующие гены человека (*allCDS_19,566*) и гены рецепторов, связанных с G-белком (*allGPCR_389*); *в*: контрольная выборка – гены рецепторов, связанных с G-белком (*allGPCR_389*) и гены рецепторов, связанных с G-белком и контролирующих аппетит (*appGPCR_45*); *г*: контрольная выборка – все белок-кодирующие гены человека (*allCDS_19,566*) и гены рецепторов, контролирующих аппетит, но не принадлежащих суперсемейству связанных с G-белком (*app_not_GPCR_19*).

Индексы PAI рассчитаны при пороговом значении уровня идентичности последовательностей генов-ортологов 0.5. Звездочками отмечены различия между количеством генов, имеющих PAI = 5 (этап дивергенции позвоночных) (*а–в*) либо PAI = 6 (этап дивергенции костных позвоночных) (*г*), и их ожидаемым количеством, рассчитанным исходя из распределения в контрольной выборке.

*** $p < 0.001$, * $p < 0.05$. Подробнее см. Приложения 5–8.

allGPCR_389 (см. рис. 2, *в*). В группе генов из выборки *appGPCR_45* было обнаружено 28 генов, имеющих PAI, равный 5 (этап дивергенции позвоночных) (т. е. 64 %), что значимо превышало ожидаемое количество (17.35), рассчитанное исходя из доли этой группы генов в выборке *allGPCR_389* (Приложение 7).

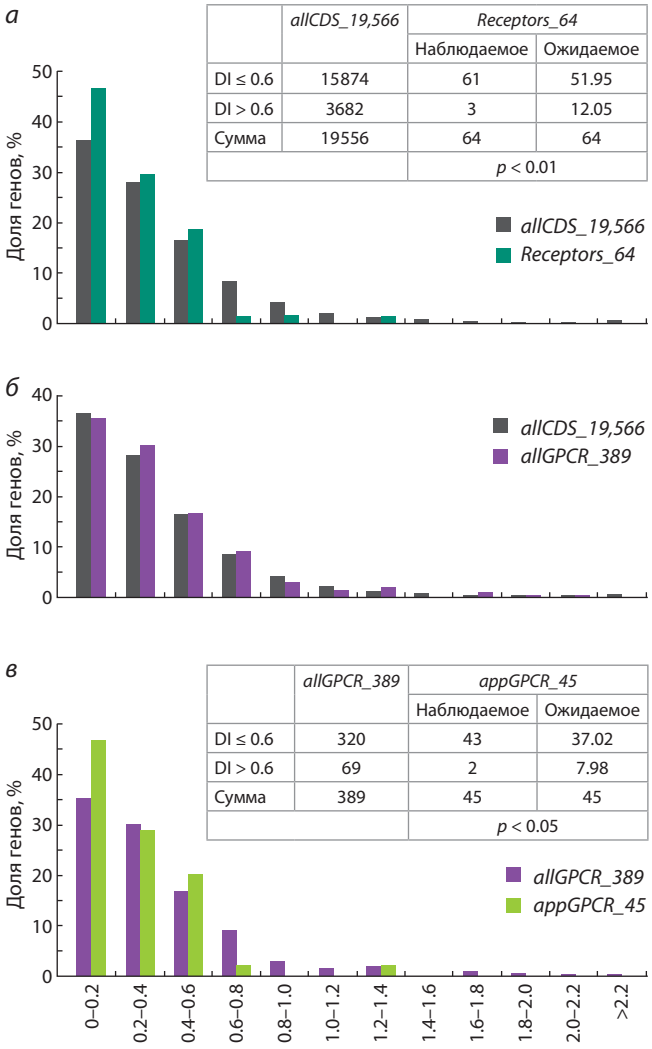
Ранее было указано, что 19 генов, кодирующих рецепторы клеточной поверхности и регулирующих аппетит, не относились к суперсемейству рецепторов, связанных с G-белком (выборка *app_not_GPCR_19*). При рассмотрении распределения по значениям PAI для этой группы генов также были найдены отличия от распределения по значениям PAI для всех белок-кодирующих генов человека (см. рис. 2, *г*). Однако в данном случае значимое ($p < 0.05$) превышение по сравнению с ожидаемым наблюдалось по количеству генов, имеющих PAI, равный 6 (этап дивергенции костных позвоночных), – 6 генов из 19 (32 %), тогда как в выборке *allCDS_19,566* PAI, равный 6, был выявлен у 2769 генов (14 %). Таким образом, ожидаемое количество генов, имеющих PAI, равный 6, составляло 2.69 (Приложение 8).

Эволюционная изменчивость генов (анализ на основе индекса DI). Анализ распределения генов из выборки *Receptors_64* по величинам DI (рис. 3, *а*, Приложение 9) показал, что 47 % генов (30 из 64) имеют DI < 0.2, большинство генов (63 из 64, или ~98 %) имеют DI < 1, и лишь у одного гена (*QRFPR*) DI > 1. Это свидетельствует о том, что большая часть генов подвергается стабилизирующему отбору.

Сравнение распределения генов из выборки *Receptors_64* по значению индекса DI с распределением, полученным для всех белок-кодирующих генов из генома человека (выборка *allCDS_19,566*) показало, что выборка *Receptors_64* характеризуется повышенным содержанием генов с низкими значениями DI (см. рис. 3, *а*). Большая часть генов из выборки *Receptors_64* (61 из 64 генов, или 95 %) имела DI < 0.6. И это количество значимо ($p < 0.01$) превышало ожидаемое (51.95), рассчитанное исходя из распределения, полученного для всех белок-кодирующих генов человека (см. рис. 3, *а*, Приложение 10).

Распределение по величинам DI для выборки всех рецепторов из суперсемейства GPRC (*allGPCR_389*) не отличалось значимым образом от распределения всех белок-кодирующих генов (*allCDS_19,566*) (см. рис. 3, *б*).

Сравнение распределения по величинам DI для выборки *appGPCR_45* с распределением для всех рецепторов из суперсемейства GPCR (*allGPCR_389*) показало, что количество генов, имеющих низкий DI (≤0.6) в выбор-



	allCDS_19,566	Receptors_64	
		Наблюдаемое	Ожидаемое
DI ≤ 0.6	15874	61	51.95
DI > 0.6	3682	3	12.05
Сумма	19556	64	64
		$p < 0.01$	

	allGPCR_389	appGPCR_45	
		Наблюдаемое	Ожидаемое
DI ≤ 0.6	320	43	37.02
DI > 0.6	69	2	7.98
Сумма	389	45	45
		$p < 0.05$	

Рис. 3. Распределения генов из выборок, представленных в табл. 1, по индексу DI.

а: контрольная выборка – все белок-кодирующие гены человека (*allCDS_19,566*) и гены рецепторов (*Receptors_64*). Суммарные наблюдаемые и ожидаемые количества генов, имеющих DI ≤ 0.6 и DI > 0.6, представлены в таблице над графиком, расчет ожидаемых количеств приведен в Приложении 10. *б*: контрольная выборка – все белок-кодирующие гены человека (*allCDS_19,566*) и гены рецепторов, связанных с G-белком (*allGPCR_389*); *в*: контрольная выборка – гены рецепторов, связанных с G-белком (*allGPCR_389*), и гены рецепторов, связанных с G-белком и контролирующих аппетит (*appGPCR_45*). Суммарные наблюдаемые и ожидаемые количества генов, имеющих DI ≤ 0.6 и DI > 0.6, представлены в таблице над графиком, расчет ожидаемых количеств приведен в Приложении 11.

ке *appGPCR_45* (42 гена), значимо ($p < 0.05$) превышает ожидаемое количество генов (37.018), рассчитанное на основе данных о распределении DI для всех генов, кодирующих GPCR (см. рис. 3, *в*, Приложение 11).

Как было указано выше, примерно одна пятая часть (18.75 %) генов из выборки *Receptors_64* тканеспецифически экспрессируется в мозге. Мы определили содержание генов, характеризующихся тканеспецифическим характером экспрессии, в двух подгруппах: подгруппа генов, имеющих низкий DI (DI ≤ 0.2); подгруппа, включающая все остальные гены (0.2 < DI < 1.3). Оказалось, что количества тканеспецифически экспрессирующихся генов в этих

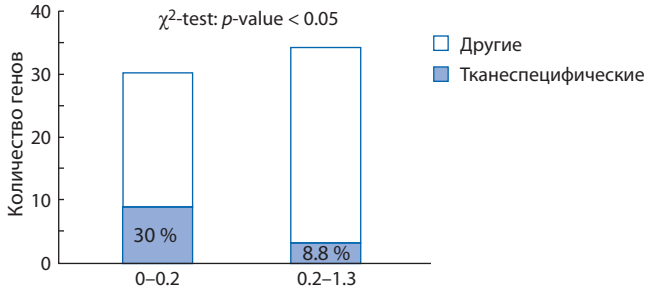


Рис. 4. Распределение по значениям DI генов из выборки *Receptors_64*.

Показаны доли генов, имеющих тканеспецифические паттерны экспрессии в мозге по данным веб-сервиса TSEA. Наблюдаемые количества генов, тканеспецифически экспрессирующихся в мозге, отличаются от ожидаемых количеств, $^*p < 0.05$. (Количества генов в четырех подгруппах приведены в Приложении 12.)

подгруппах значимо отличаются от ожидаемых величин, рассчитанных исходя из случайного распределения: в подгруппе генов с низким DI содержание тканеспецифически экспрессирующихся генов было повышено (рис. 4, Приложение 12).

Обсуждение

Гены рецепторов клеточной поверхности составляют существенную долю (более тысячи генов) в геноме человека (Bausch-Fluck et al., 2018). Интерес к исследованию рецепторов клеточной поверхности связан с важной ролью, которую они выполняют. Это трансмембранные белки, которые взаимодействуют с различными молекулами (лигандами), находящимися во внеклеточном пространстве, и активируют пути сигнальной трансдукции в клетке (Bausch-Fluck et al., 2018; Yang et al., 2021). Для рецепторов клеточной поверхности известно большое количество веществ и биохимических соединений (в частности, лекарственных препаратов), влияющих на их активность (так называемых агонистов и антагонистов). Поэтому рецепторы клеточной поверхности представляют большой интерес и с точки зрения биомедицины – например, эти белки являются мишенями для 66 % лекарственных препаратов, зарегистрированных в DrugBank database (Bausch-Fluck et al., 2018).

В настоящей работе рассмотрена выборка, включающая 64 гена человека, кодирующих рецепторы клеточной поверхности, ортологи которых участвуют в регуляции количества потребленной пищи у модельных видов животных. Выборка создавалась на основе ручного анализа научных публикаций, что свидетельствует о высоком уровне достоверности. Участие такого внушительного количества генов рецепторов в регуляции аппетита хорошо согласуется с представлениями о сложной природе пищевой мотивации. Как было сказано выше, аппетит может удовлетворять базисные потребности организма в пище (гомеостатический аппетит, обеспечивающий компенсацию энергозатрат), а также потребности в ощущениях, связанных с едой (негомеостатический аппетит, направленный на получение положительных эмоций) (Johnson, 2013; Rebello, Greenway, 2016; Ahn et al., 2022). Известно, что интенсивность пищевой мотивации может кор-

ректироваться в зависимости от жизненной ситуации либо психоэмоционального состояния индивида (испуг, депрессия, скука, хронический стресс, для животных – угроза со стороны хищников, охрана территории, брачное поведение и т. д.) (Lindén et al., 1987; Braden et al., 2018, 2023; Hadjieconomou et al., 2020; Siegal et al., 2022). Такая корректировка реализуется за счет того, что мозг обрабатывает информацию, получаемую от органов чувств, и интегрирует ее с сигналами о состоянии различных физиологических систем организма (Tomé et al., 2009; Holtmann, Talley, 2014; Spetter et al., 2014; Tremblay, Bellisle, 2015). И в этот процесс вовлечены нервные клетки с различной специализацией, экспрессирующие на своей поверхности широкий спектр рецепторов (Yeo, Heisler, 2012; Heisler, Lam, 2017).

При рассмотрении распределений генов по величинам PAI было выявлено: 1) выборка *Receptors_64* содержит значимо больше генов, имеющих одинаковый филостратиграфический возраст (PAI = 5, этап дивергенции позвоночных), чем все белок-кодирующие гены; 2) подвыборка генов из суперсемейства GPCR, участвующих в регуляции аппетита (*appGPCR_45*), также содержит повышенное количество генов с одинаковым филостратиграфическим возрастом (PAI = 5, этап дивергенции позвоночных), чем это можно было ожидать, основываясь на данных о распределении значений PAI для всех рецепторов суперсемейства GPCR.

Таким образом, набор генов рецепторов клеточной поверхности, контролирующих аппетит, содержит повышенное количество генов, имеющих одинаковый филостратиграфический возраст (PAI = 5, этап дивергенции позвоночных). По-видимому, синхронизированное эволюционирование такой многочисленной группы генов (31 ген имеет PAI = 5) связано с формированием у первых позвоночных мозга как отдельного органа (Sarnat, Netsky, 2002). Примечательно, что почти все эти гены, имеющие PAI, равный 5 (28 из 31), относятся к суперсемейству GPCR, что хорошо согласуется с тем, что рецепторы этого суперсемейства задействованы в обработке сигналов от органов чувств, а также сигналов, передаваемых гормонами и нейромедиаторами (Pandy-Szekeres et al., 2023). Так, к группе генов рецепторов из суперсемейства GPCR, имеющих PAI, равный 5 (этап дивергенции позвоночных), относятся, в частности, гены, кодирующие рецепторы нейропептида Y (*NPY1R*, *NPY2R*, *NPY4R*, *NPY5R*) и альфа-меланоцитстимулирующего гормона (*MC3R* и *MC4R*). Нейропептид Y и альфа-меланоцитстимулирующий гормон являются сигнальными молекулами, секретируемыми нейронами аркуатного ядра гипоталамуса – структуры мозга, выполняющей функцию центрального регулятора пищевого поведения (Yeo, Heisler, 2012; Heisler, Lam, 2017).

Другая особенность выявлена для подвыборки генов, участвующих в регуляции аппетита, но не принадлежащих суперсемейству GPCR (*app_not_GPCR_19*): здесь содержится повышенное количество генов, имеющих PAI, равный 6 (этап дивергенции костных позвоночных). Примечательно, что четыре гена из этой группы кодируют рецепторы, имеющие отношение к иммунной системе. Это гены *GHR* и *LEPR*, кодирующие белки из семейства цитокиновых рецепторов типа I (type I cytokine receptor family), и гены *TLR2* и *TLR4*, кодирующие белки из семейства толл-подобных рецепторов (Toll-like receptor family).

Рассмотрение значений индексов PAI показало, что в регуляцию потребления пищи вовлечены и так называемые древние гены (т. е. гены, имеющие PAI = 0, – клеточные организмы, корень дерева). К ним относятся, например, ген *INSR*, кодирующий рецептор инсулина и участвующий, в частности, в регуляции секреции нейропептида Y и альфа-меланоцитостимулирующего гормона нейронами аркуатного ядра гипоталамуса (Leibowitz, Wortley, 2004), а также ген *NTRK2*, кодирующий рецептор нейротрофического фактора мозга, опосредующего анорексигенное действие нейротрофического фактора, вырабатываемого в паравентрикулярном ядре (An et al., 2015; Chu et al., 2023). Оба гена экспрессируются в различных тканях и органах (Escandón et al., 1994; Federici et al., 1997). Это указывает на то, что на ранних этапах эволюции живых организмов предковые формы белков INSR и NTRK2 могли быть задействованы в регуляции различных биологических процессов и оказались вовлечены в систему генов, регулирующую потребление пищи, на эволюционном этапе, соответствующем формированию специализированных структур мозга.

Анализ значений индексов PAI позволил выявить группу относительно «молодых» генов (значения PAI равны 6 и 7, этапы дивергенции костных позвоночных и млекопитающих). Пять генов из этой группы кодируют рецепторы, имеющие отношение к иммунной системе: четыре вышеупомянутых гена *GHR*, *LEPR*, *TLR2* и *TLR4*, а также *IL1R1*. Обнаружение этих генов в группе относительно «молодых» хорошо согласуется с известными данными о том, что в ходе эволюции система адаптивного иммунитета начала формироваться сравнительно недавно (Ward, Rosenthal, 2014).

При исследовании распределения генов из выборки *Receptors_64* по значениям индексов DI была выявлена существенная обогащенность этой группы генами, подверженными стабилизирующему отбору. Оказалось, что подгруппа генов, регулирующих аппетит и относящихся к суперсемейству GPCR (*appGPCR_45*), тоже содержала повышенное количество генов с низким индексом DI.

Наиболее низкие значения индекса DI (<0.05) имели восемь генов: *GPR26*, *NPY1R*, *GHSR*, *ADIPOR1*, *DRD1*, *NPY2R*, *GPR171*, *NPBWR1* (см. Приложение 9). Причем семь из этих восьми генов (за исключением *ADIPOR1*) кодируют белки из суперсемейства GPCR.

Экстремально низкое значение DI (<0.005) имел ген *GPR26*. Он кодирует рецептор из суперсемейства GPCR, лиганд которого до сих пор неизвестен. У мышей с нокаутом гена *GPR26* развивается гиперфагия (Chen et al., 2012). Кроме того, согласно поведенческим тестам, такие мыши склонны к депрессии и тревожности, а также в большей степени предпочитают этанол, чем мыши с нормальным генотипом (Zhang et al., 2011). По данным ресурса *TSEA tool*, *GPR26* является тканеспецифичным для мозга. У человека он экспрессируется в амигдале, гиппокампе и таламусе (Jones et al., 2007). Функция гена *GPR26* эволюционно консервативна – в эксперименте по подавлению экспрессии генов с помощью РНК-интерференции у *C. elegans* было показано, что у этого организма

имеется ген *Y5H2B*, обладающий сходством с *GPR26* и регулирующий накопление жира (Ashrafi et al., 2003). Функции других генов с экстремально низкими значениями DI (<0.05) кратко охарактеризованы в Приложении 13.

Среди генов из выборки *Receptors_64* лишь один ген (*QRFPR*) имел DI > 1 (см. Приложение 9). Высокое значение индекса DI указывает на то, что, возможно, этот ген подвержен движущему отбору. Известно, что *QRFPR* кодирует рецептор орексигенного нейропептида QRFP (pyroglutamylated RFamide peptide) (Cook et al., 2022). По данным баз EntrezGene и UniProt, *QRFPR* экспрессируется у человека как в различных отделах мозга, так и в периферических тканях (сердце, почках, желудке, яичках, щитовидной железе). В геномах мыши, крысы и хомяка содержится как минимум два гена, кодирующих рецептор нейропептида QRFP (Cook et al., 2022). Для генома человека таких данных нет, однако можно предполагать, что *QRFPR* не подвержен стабилизирующему отбору, поскольку геном человека тоже содержит несколько генов, кодирующих белки со сходной функцией.

Мы также обнаружили, что группа генов, имеющих низкий DI, т. е. с наибольшей вероятностью подверженных стабилизирующему отбору, обогащена генами, тканеспецифически экспрессирующимися в мозге. Этот результат хорошо согласуется с данными, полученными G. Dumas с коллегами при исследовании почти полного набора белоккодирующих генов человека ($N = 11\,667$). Авторы показали (Dumas et al., 2021), что гены, кодирующие белки, функции которых связаны с мозгом, являются наиболее консервативными в геноме человека. В группе генов, имеющих низкий DI и тканеспецифически экспрессирующихся в мозге, обращает на себя внимание уже упоминавшийся ген *GPR26.* Поскольку *GPR26* обладает экстремально низкой метрикой DI, а также ввиду того, что лиганд для рецептора *GPR26* пока не выявлен (Chen et al., 2012), этот ген представляется чрезвычайно интересным объектом для дальнейшего теоретического и экспериментального исследования.

Заключение

В нашей работе проанализированы распределения индексов PAI и DI для группы генов рецепторов клеточной поверхности человека, ортологи которых участвовали в регуляции аппетита у модельных видов животных. Обнаружено, что в рассматриваемом наборе генов содержится повышенное количество генов, имеющих одинаковый филостратиграфический возраст (PAI = 5, этап дивергенции позвоночных), что, по-видимому, связано с формированием у первых позвоночных мозга как отдельного органа. Выявлена также значительная обогащенность данной группы генами с низким значением индекса DI, что указывает на существенную подверженность этих генов стабилизирующему отбору. При этом группа генов, имеющих низкий DI, обогащена генами, тканеспецифически экспрессирующимися в мозге. Выявленные нами особенности распределения генов рецепторов клеточной поверхности по эволюционным индексам PAI и DI являются отправной точкой для дальнейшего анализа эволюционных характеристик генной сети регуляции аппетита в целом.

Список литературы / References

Игнатьева Е.В., Афонников Д.А., Рогаев Е.И., Колчанов Н.А. Гены, контролирующие пищевое поведение и массу тела человека, и их функциональные и геномные характеристики. *Вавиловский журнал генетики и селекции.* 2014;18(4/2):867-875
[Ignatieva E.V., Afonnikov D.A., Rogaev E.I., Kolchanov N.A. Human genes controlling feeding behavior or body mass and their functional and genomic characteristics: a review. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2014;18(4/2):867-875 (in Russian)]

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 Journal of Genetics and Breeding.* 2021;25(1):46-56. DOI 10.18699/VJ21.006]

Ahn B.H., Kim M., Kim S.Y. Brain circuits for promoting homeostatic and non-homeostatic appetites. *Exp. Mol. Med.* 2022;54(4):349-357. DOI 10.1038/s12276-022-00758-4

An J.J., Liao G.Y., Kinney C.E., Sahibzada N., Xu B. Discrete BDNF neurons in the paraventricular hypothalamus control feeding and energy expenditure. *Cell Metab.* 2015;22(1):175-188. DOI 10.1016/j.cmet.2015.05.008

Ashrafi K., Chang F.Y., Watts J.L., Fraser A.G., Kamath R.S., Ahringer J., Ruvkun G. Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature.* 2003;421(6920):268-272. DOI 10.1038/nature01279

Bausch-Fluck D., Goldmann U., Müller S., van Oostrum M., Müller M., Schubert O.T., Wollscheid B. The in silico human surfaceome. *Proc. Natl. Acad. Sci. USA.* 2018;115(46):E10988-E10997. DOI 10.1073/pnas.1808790115

Bjarnadóttir T.K., Gloriam D.E., Hellstrand S.H., Kristiansson H., Fredriksson R., Schiöth H.B. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics.* 2006;88(3):263-273. DOI 10.1016/j.ygeno.2006.04.001

Braden A., Musher-Eizenman D., Watford T., Emley E. Eating when depressed, anxious, bored, or happy: are emotional eating types associated with unique psychological and physical health correlates? *Appetite*. 2018;125:410-417. DOI 10.1016/j.appet.2018.02.022

Braden A., Barnhart W.R., Kalantzis M., Redondo R., Dauber A., Anderson L., Tilstra-Ferrell E.L. Eating when depressed, anxious, bored, or happy: an examination in treatment-seeking adults with overweight/obesity. *Appetite.* 2023;184:106510. DOI 10.1016/j.appet.2023.106510

Chen D., Liu X., Zhang W., Shi Y. Targeted inactivation of GPR26 leads to hyperphagia and adiposity by activating AMPK in the hypothalamus. *PLoS One.* 2012;7(7):e40764. DOI 10.1371/journal.pone.0040764

Chu P., Guo W., You H., Lu B. Regulation of satiety by *Bdnf-e2*-expressing neurons through TrkB activation in ventromedial hypothalamus. *Biomolecules.* 2023;13(5):822. DOI 10.3390/biom13050822

Cook C., Nunn N., Worth A.A., Bechtold D.A., Suter T., Gackeheimer S., Foltz L., Emmerson P.J., Statnick M.A., Luckman S.M. The hypothalamic RFamide, QRFP, increases feeding and locomotor activity: the role of Gpr103 and orexin receptors. *PLoS One.* 2022;17(10):e0275604. DOI 10.1371/journal.pone.0275604

Dumas G., Malesys S., Bourgeron T. Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition. *Genome Res.* 2021;31(3):484-496. DOI 10.1101/gr.262113.120

Escandón E., Soppet D., Rosenthal A., Mendoza-Ramírez J.L., Szönyi E., Burton L.E., Henderson C.E., Parada L.F., Nikolics K. Regulation of neurotrophin receptor expression during embryonic and postnatal development. *J. Neurosci.* 1994;14(4):2054-2068. DOI 10.1523/JNEUROSCI.14-04-02054.1994

Federici M., Porzio O., Zucaro L., Fusco A., Borboni P., Lauro D., Sesti G. Distribution of insulin/insulin-like growth factor-I hybrid receptors in human tissues. *Mol. Cell. Endocrinol.* 1997;129(2): 121-126. DOI 10.1016/s0303-7207(97)04050-1

Fichter M.M., Quadflieg N. Mortality in eating disorders – results of a large prospective clinical longitudinal study. *Int. J. Eat. Disord*. 2016;49(4):391-401. DOI 10.1002/eat.22501

Grossberg A.J., Scarlett J.M., Marks D.L. Hypothalamic mechanisms in cachexia. *Physiol. Behav.* 2010;100(5):478-489. DOI 10.1016/j.physbeh.2010.03.011

GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660. DOI 10.1126/science.1262110

Hadjieconomou D., King G., Gaspar P., Mineo A., Blackie L., Ameku T., Studd C., de Mendoza A., Diao F., White B.H., Brown A.E.X., Plaçais P.Y., Préat T., Miguel-Aliaga I. Enteric neurons increase maternal food intake during reproduction. *Nature.* 2020;587(7834): 455-459. DOI 10.1038/s41586-020-2866-8

Heisler L.K., Lam D.D. An appetite for life: brain regulation of hunger and satiety. *Curr. Opin. Pharmacol.* 2017;37:100-106. DOI 10.1016/j.coph.2017.09.002

Holtmann G., Talley N.J. The stomach-brain axis. *Best Pract. Res. Clin. Gastroenterol.* 2014;28(6):967-979. DOI 10.1016/j.bpg.2014.10.001

Ignatieva E.V., Afonnikov D.A., Saik O.V., Rogaev E.I., Kolchanov N.A. A compendium of human genes regulating feeding behavior and body weight, its functional characterization and identification of GWAS genes involved in brain-specific PPI network. *BMC Genet.* 2016;17(Suppl.3):158. DOI 10.1186/s12863-016-0466-2

Johnson A.W. Eating beyond metabolic need: how environmental cues influence feeding behavior. *Trends Neurosci.* 2013;36(2):101-109. DOI 10.1016/j.tins.2013.01.002

Jones P.G., Nawoschik S.P., Sreekumar K., Uveges A.J., Tseng E., Zhang L., Johnson J., He L., Paulsen J.E., Bates B., Pausch M.H. Tissue distribution and functional analyses of the constitutively active orphan G protein coupled receptors, GPR26 and GPR78. *Biochim. Biophys. Acta.* 2007;1770(6):890-901. DOI 10.1016/j.bbagen.2007.01.013

Kaidar-Person O., Bar-Sela G., Person B. The two major epidemics of the twenty-first century: obesity and cancer. *Obes. Surg.* 2011; 21(11):1792-1797. DOI 10.1007/s11695-011-0490-2

Katritch V., Cherezov V., Stevens R.C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* 2013;53:531-556. DOI 10.1146/annurev-pharmtox-032112-135923

Leibowitz S.F., Wortley K.E. Hypothalamic control of energy balance: different peptides, different functions. *Peptides.* 2004;25(3):473-504. DOI 10.1016/j.peptides.2004.02.006

Lindén A., Hansen S., Bednar I., Forsberg G., Södersten P., Uvnäs-Moberg K. Sexual activity increases plasma concentrations of cholecystokinin octapeptide and offsets hunger in male rats. *J. Endocrinol.* 1987;115(1):91-95. DOI 10.1677/joe.0.1150091

Maniam J., Morris M.J. The link between stress and feeding behaviour. *Neuropharmacology.* 2012;63(1):97-110. DOI 10.1016/j.neuropharm.2012.04.017

New D.C., Wong J.T. The evidence for G-protein-coupled receptors and heterotrimeric G proteins in protozoa and ancestral metazoa. *Biol. Signals Recept.* 1998;7(2):98-108. DOI 10.1159/000014535

Olszewski P.K., Cedernaes J., Olsson F., Levine A.S., Schiöth H.B. Analysis of the network of feeding neuroregulators using the Allen Brain Atlas. *Neurosci. Biobehav. Rev*. 2008;32(5):945-956. DOI 10.1016/j.neubiorev.2008.01.007

Pandy-Szekeres G., Caroli J., Mamyrbekov A., Kermani A.A., Keseru G.M., Kooistra A.J., Gloriam D.E. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Res.* 2023;51(D1):D395-D402. DOI 10.1093/nar/gkac1013

Rebello C.J., Greenway F.L. Reward-induced eating: therapeutic approaches to addressing food cravings. *Adv. Ther.* 2016;33(11):1853-1866. DOI 10.1007/s12325-016-0414-6

Sarnat H.B., Netsky M.G. When does a ganglion become a brain? Evolutionary origin of the central nervous system. *Semin. Pediatr. Neurol.* 2002;9(4):240-253. DOI 10.1053/spen.2002.32502

Siegal E., Hooker S.K., Isojunno S., Miller P.J.O. Beaked whales and state-dependent decision-making: how does body condition affect the trade-off between foraging and predator avoidance? *Proc. Biol. Sci.* 2022;289(1967):20212539. DOI 10.1098/rspb.2021.2539

Spetter M.S., de Graaf C., Mars M., Viergever M.A., Smeets P.A. The sum of its parts – effects of gastric distention, nutrient content and sensory stimulation on brain activation. *PLoS One.* 2014;9(3): e90872. DOI 10.1371/journal.pone.0090872

Tomé D., Schwarz J., Darcel N., Fromentin G. Protein, amino acids, vagus nerve signaling, and the brain. *Am. J. Clin. Nutr.* 2009;90(3): 838S-843S. DOI 10.3945/ajcn.2009.27462W

Tremblay A., Bellisle F. Nutrients, satiety, and control of energy intake. *Appl. Physiol. Nutr. Metab.* 2015;40(10):971-979. DOI 10.1139/apnm-2014-0549

Ward A.E., Rosenthal B.M. Evolutionary responses of innate immunity to adaptive immunity. *Infect. Genet. Evol.* 2014;21:492-496. DOI 10.1016/j.meegid.2013.12.021

Wells A., Kopp N., Xu X., O'Brien D.R., Yang W., Nehorai A., Adair-Kirk T.L., Kopan R., Dougherty J.D. The anatomical distribution of genetic associations. *Nucleic Acids Res.* 2015;43(22):10804-10820. DOI 10.1093/nar/gkv1262

Yang D., Zhou Q., Labroska V., Qin S., Darbalaei S., Wu Y., Yuliantie E., Xie L., Tao H., Cheng J., Liu Q., Zhao S., Shui W., Jiang Y., Wang M.W. G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct. Target. Ther.* 2021;6(1):7. DOI 10.1038/s41392-020-00435-w

Yeo G.S., Heisler L.K. Unraveling the brain regulation of appetite: lessons from genetics. *Nat. Neurosci.* 2012;15(10):1343-1349. DOI 10.1038/nn.3211

Zhang L.L., Wang J.J., Liu Y., Lu X.B., Kuang Y., Wan Y.H., Chen Y., Yan H.M., Fei J., Wang Z.G. GPR26-deficient mice display increased anxiety- and depression-like behaviors accompanied by reduced phosphorylated cyclic AMP responsive element-binding protein level in central amygdala. *Neuroscience.* 2011;196:203-214. DOI 10.1016/j.neuroscience.2011.08.069

ORCID ID

E.V. Ignatieva orcid.org/0000-0002-8588-6511
S.A. Lashin orcid.org/0000-0003-3138-381X
Z.S. Mustafin orcid.org/0000-0003-2724-4497
N.A. Kolchanov orcid.org/0000-0001-6800-8787

Благодарности. 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.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 11.08.2023. После доработки 05.09.2023. Принята к публикации 07.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):839-850
DOI 10.18699/VJGB-23-97

Перевод на английский язык https://vavilov.elpub.ru/jour

О пространстве вариантов генетических последовательностей SARS-CoV-2

А.Ю. Пальянов[1, 2, 3] ✉, Н.В. Пальянова[2]

[1] Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Научно-исследовательский институт вирусологии, Федеральный исследовательский центр фундаментальной и трансляционной медицины, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ palyanov@iis.nsk.su

Аннотация. Пандемия коронавирусной инфекции, вызванная вирусом SARS-CoV-2, которой человечество противостояло с использованием новейших достижений науки, оставила после себя в том числе обширные генетические данные. Ежедневно начиная с конца 2019 г. в мире собирались образцы геномов вируса, что предоставляет возможность детально проследить его эволюцию с момента возникновения до настоящего времени. Накопленная статистика результатов экспресс-тестирования показала, что число подтвержденных случаев заражения SARS-CoV-2 составило не менее 767.5 млн (9.5 % нынешнего населения Земли без учета бессимптомников), а число секвенированных геномов вируса – более 15.7 млн (что составляет чуть более 2 % от общего числа заразившихся). Эти новые данные потенциально несут в себе информацию о механизмах изменчивости и распространения вируса, его взаимодействия с иммунной системой человека, об основных параметрах, характеризующих механизмы развития пандемии, и многое другое. В этой статье мы анализируем пространство возможных вариантов генетических последовательностей SARS-CoV-2 как с математической точки зрения, так и с учетом биологических ограничений, присущих этой системе (основанных на общебиологических знаниях и учитывающих особенности данного конкретного вируса). Для этого мы разработали программное обеспечение, способное загружать и анализировать нуклеотидные последовательности SARS-CoV-2 в формате FASTA, определять позиции 5′ и 3′ UTR, число и расположение неидентифицированных нуклеотидов (“N”), осуществлять выравнивание относительно референсной последовательности посредством вызова предназначенных для этого программ, определять мутации, делеции и вставки, а также рассчитывать различные характеристики геномов вирусов с заданным шагом по времени (дни, недели, месяцы и т. д.). Полученные данные свидетельствуют о том, что, несмотря на кажущееся математическое многообразие возможных вариантов изменения вируса во времени, коридор эволюционной траектории, которым прошел коронавирус, представляется достаточно узким. Это дает основание полагать, что он в некоторой степени детерминирован, что позволяет надеяться на возможность моделирования эволюции коронавируса.
Ключевые слова: коронавирус; SARS-CoV-2; геном; пространство вариантов; эволюция; изменчивость.

Для цитирования: Пальянов А.Ю., Пальянова Н.В. О пространстве вариантов генетических последовательностей SARS-CoV-2. *Вавиловский журнал генетики и селекции*. 2023;27(7):839-850. DOI 10.18699/VJGB-23-97

On the space of SARS-CoV-2 genetic sequence variants

A.Yu. Palyanov[1, 2, 3] ✉, N.V. Palyanova[2]

[1] A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Research Institute of Virology, Federal Research Center of Fundamental and Translational Medicine of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
✉ palyanov@iis.nsk.su

Abstract. The coronavirus pandemic caused by the SARS-CoV-2 virus, which humanity resisted using the latest advances in science, left behind, among other things, extensive genetic data. Every day since the end of 2019, samples of the virus genomes have been collected around the world, which makes it possible to trace its evolution in detail from its emergence to the present. The accumulated statistics of testing results showed that the number of confirmed cases of SARS-CoV-2 infection was at least 767.5 million (9.5 % of the current world population, excluding asymptomatic people), and the number of sequenced virus genomes is more than 15.7 million (which is over 2 % of the total number of infected people). These new data potentially contain information about the mechanisms of the variability and spread of the virus, its interaction with the human immune system, the main parameters characterizing the mechanisms of the development of a pandemic, and much more. In this article, we analyze the space of possible variants of SARS-CoV-2 genetic sequences both from a mathematical point of view and taking into account the biological limitations inherent in this system, known both from general biological knowledge and from the consideration of the characteristics of this particular virus. We have developed software capable of loading and analyzing

SARS-CoV-2 nucleotide sequences in FASTA format, determining the 5′ and 3′ UTR positions, the number and location of unidentified nucleotides ("N"), performing alignment with the reference sequence by calling the program designed for this, determining mutations, deletions and insertions, as well as calculating various characteristics of virus genomes with a given time step (days, weeks, months, etc.). The data obtained indicate that, despite the apparent mathematical diversity of possible options for changing the virus over time, the corridor of the evolutionary trajectory that the coronavirus has passed through seems to be quite narrow. Thus it can be assumed that it is determined to some extent, which allows us to hope for a possibility of modeling the evolution of the coronavirus.
Key words: coronavirus; SARS-CoV-2; genome; space of variants; evolution; variability.

For citation: Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):839-850. DOI 10.18699/VJGB-23-97

Введение

Возможность компьютерного моделирования эволюции, жизненного цикла и размножения простейшего биологического организма с детализацией до генного уровня стала бы научным прорывом, однако это по-прежнему находится далеко за пределами возможностей современных суперкомпьютеров. Процесс естественного отбора наиболее приспособленных особей происходит с учетом огромного количества факторов как внешней, так и внутренней среды. Особенности организма реализуются через наборы особенностей белков, а влияние изменений каждого белка на приспособленность оценить достаточно трудно в связи с необходимостью учитывать все возникающие изменения взаимодействий этого белка со всеми факторами среды и другими белками, число которых весьма значительно.

В компьютерных моделях эволюционирующих объектов, как правило, внесение изменений в геном потомков осуществляется не явным образом (посредством воспроизведения молекулярных механизмов), а лишь имитируется посредством описания алгоритмов внесения изменений в копию генома предков. Однако и сами механизмы внесения мутаций и горизонтального переноса генов являются субъектами эволюции, и среди возможных изменений, не приводящих к гибели или стерильности особи, встречаются и те, что влияют на скорость и точность репликации генома. Благодаря этому возникает внутривидовая конкуренция, в результате которой, например, для SARS-CoV-2 с момента его появления и до настоящего времени длительность инкубационного периода, напрямую связанная со скоростью репликации вируса, постоянно снижается (Malone et al., 2022).

По сравнению с клеточными формами жизни вирусы представляются значительно более удобными объектами для изучения и компьютерного моделирования эволюции благодаря достаточно простому устройству и значительно меньшему геному при широком спектре взаимодействий с внешней средой и организмом хозяина. До появления технологий быстрого секвенирования геномов эволюцию вирусов можно было рассматривать лишь в рамках моделей «паразит–хозяин», описывающих статистические, но не молекулярные особенности их взаимодействия. С начала пандемии SARS-CoV-2 число подтвержденных случаев заражения SARS-CoV-2 составило не менее 767.5 млн (9.5 % нынешнего населения Земли без учета бессимптомников) (Palyanova et al., 2022). Мировым научным сообществом было получено более 15.7 млн вариантов геномов данного коронавируса (включая дату взятия образца и географическое расположение места его получения), предоставляющих беспрецедентно обширные данные о его эволюции, в таком количестве не имеющиеся ни для какого из других вирусов.

На основе этих данных может быть рассчитана динамика распространения и изменения вируса не только в физическом пространстве и времени, но и в многомерном пространстве возможных жизнеспособных вариантов вирусных геномов. Метрика такого пространства определяется минимальным числом единичных изменений (мутация, делеция или вставка), необходимых для преобразования одного генома в другой (расстояние Левенштейна, или «редакционное расстояние»). При этом вирус изменяется, в том числе в ответ на вакцинацию и формирование иммунитета у переболевших. А значит, изменяется как геном вируса, так и его «фенотипические» проявления при взаимодействии с организмом носителя, т. е. одновременно происходят два процесса: изменение (распространение) множества (облака) точек, представляющих популяцию вируса в тот или иной момент времени в пространстве возможных последовательностей РНК, и изменение самого ландшафта этой многомерной поверхности «функции приспособленности» вируса. Каждая точка в пространстве возможных состояний соответствует определенной нуклеотидной последовательности, более или менее отличающейся от исходного референсного генома (с которого все началось в конце 2019 г. (Wu et al., 2020)) некоторым количеством изменений – мутаций, делеций и вставок.

Между парами точек, каждая из которых соответствует жизнеспособной последовательности, если одна из них получилась из другой вследствие изменений, произошедших с вирусом с момента попадания в организм носителя до появления вирионов следующего поколения (как правило, до этого проходит далеко не один цикл репликации генома вируса), могут и должны существовать переходы. Большинство возможных изменений, возникающих при репликации (у каждого экземпляра вирусной последовательности – свои собственные), приведут к его нежизнеспособности (особенно делеции или вставки, длина которых не кратна трем, т. е. такие, что приведут к сдвигу рамки считывания при трансляции). Однако некоторые изменения могут оставить приспособленность вируса на прежнем уровне или даже улучшить ее – например, повысив скорость синтеза новых вирусных частиц или увеличив их количество, производимое в единицу времени (что увеличит их преимущество перед остальными вариантами, находящимися в это же время в организме, т. е. возникает внутривидовая конкуренция). Под функцией

приспособленности можно понимать число экземпляров вирусной последовательности, существующей в человеческой популяции в данный момент времени (с нормировкой на общее число экземпляров вируса в ней или без таковой).

Таким образом, формируется (проявляется) ландшафт «поверхности» (многомерной) функции приспособленности, который может иметь более или менее обширные «долины», соответствующие множеству сходных последовательностей (появившихся в результате небольших изменений варианта, впервые попавшего в эту долину), «горные хребты» или «плато», разграничивающие «долины» (все точки которых соответствуют нежизнеспособным последовательностям), «перевалы», по которым можно перемещаться между «долинами». Области нежизнеспособных последовательностей соответствуют тем из них, которые, к примеру, не могут создавать свои копии из-за повреждения гена, кодирующего РНК-зависимую РНК-полимеразу (RdRp), осуществляющую репликацию вирусной РНК, либо тем, у которых изменения в структуре соответствующих белков не позволяют вирусу сформировать белковую оболочку, а также в силу множества других разнообразных причин. Также, надо полагать, имеются «долины», для которых ни один из принадлежащих им вариантов последовательностей еще не был реализован, однако в которые все-таки возможно попасть – например, в результате возникновения жизнеспособного рекомбинантного штамма, получившегося при сочетании двух более-менее различных вариантов геномов. Возможно, именно этим путем и возник изначальный вариант коронавируса SARS-CoV-2.

В настоящее время существуют две основные базы данных, предоставляющие пользователям онлайн-доступ к генетическим последовательностям SARS-CoV-2. Крупнейшей из них является созданная в 2006 г. GISAID (Global Initiative on Sharing All Influenza Data – глобальная инициатива по обмену всеми данными о гриппе, https://gisaid.org) (Khare et al., 2021), которая с момента появления коронавируса SARS-CoV-2 в конце 2019 г. стала также репозиторием для накопления секвенированных вариантов этого вируса, полученных лабораториями по всему миру. В июле 2023 г. в ней насчитывалось более 15.7 млн записей. Вторая база данных, NCBI SARS-CoV-2 Data Hub (Sayers et al., 2022) (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?VirusLineage_ss=taxid:2697049), содержит более 7.7 млн образцов геномов SARS-CoV-2. Столь беспрецедентно обширные и детальные данные прежде не были доступны человечеству, поэтому необходимо извлечь как можно больше полезной информации и знаний на основе их всестороннего анализа. Наша работа представляет собой лишь первые шаги на этом пути, и многое еще предстоит сделать.

Высокой значимостью для научного сообщества исследователей вирусных геномов обладает также проект Nextstrain/Nextclade (https://clades.nextstrain.org/) (Aksamentov et al., 2021), предоставляющий онлайн-инструменты для анализа и визуализации генетических данных по различным вирусам, включая SARS-CoV-2. Функциональные возможности Nextstrain выгодно выделяются на общем фоне и включают в том числе графическое отображение карты генома загруженных последовательностей с изображением мутаций, делеций, вставок, неопределенных нуклеотидов (“N”) и ряда других особенностей каждой последовательности, например принадлежность к классу реассортантных (рекомбинантных) вариантов.

Описание пространства вариантов генетических последовательностей SARS-CoV-2 принципиально включает в себя: те, которые мы уже можем наблюдать и изучать благодаря обширному секвенированию; варианты из реального пространства вариантов, которые уже реализовались, однако не попали в поле зрения исследователей; остальные возможные варианты, которые могли бы реализоваться в будущем и представляют особый интерес, поскольку являются потенциально опасными для человечества и было бы хорошо заранее быть готовыми к их возможному появлению (экспресс-тесты для их выявления, вакцины и т. д.).

Рассмотрим теперь наиболее важные характеристики SARS-CoV-2 как системы, основой существования которой является саморепликация в клетках носителя, и которые будут важны в будущем при создании его эволюционного симулятора. Они включают скорость репликации генома (600–700 нт/с, самая большая среди известных скоростей работы вирусных РНК-полимераз) (Shannon et al., 2020), время репликации вирусной РНК ($\frac{3 \cdot 10^4 \text{ нт}}{600 \text{ нт/с}} = 50$ с), время воспроизводства вируса целиком (7–24 ч) (Grebennikov et al., 2021) и частоту возникновения ошибок ($1.3 \cdot 10^{-6} \pm 0.2 \cdot 10^{-6}$ на позицию за инфекционный цикл заражения клетки, т. е. от входа вируса в нее до выхода новых вирионов наружу) (Amicone et al., 2022). При этом скорость эволюционных изменений в геноме SARS-CoV-2 оценивается как как $8.9 \cdot 10^{-4}$ замен в год в каждой позиции (Sonnleitner et al., 2022), что могло бы привести в среднем к 93 заменам за 3.5 года. Это хорошо соотносится с тем, что один из наиболее далеких от референсной последовательности вариантов, относящийся к линии «Омикрон», полученный 20.06.2023, имеет 103 замены (максимальное число мутаций среди вариантов, см. таблицу). У вариантов «Альфа» и «Бета» отличия от первоначальной референсной последовательности составляют более 30 точечных мутаций и более 17 делеций. У вариантов, возникших позднее, отличий больше. Также заметно, что в процессе эволюции вируса увеличивается число делеций, достигая 59 шт. в одной из современных ветвей «Омикрона».

Как уже было сказано, коронавирус SARS-CoV-2 обладает самой быстрой РНК-полимеразой. При этом она имеет еще и один из самых низких для РНК-вирусов показателей частоты возникновения мутаций в процессе репликации, что необходимо ввиду его достаточно большого генома. Это достигается благодаря экзонуклеазе (nsp14-ExoN), корректирующей ошибки, которая встречается только у вирусов с большими геномами (коронавирусов и торовирусов) (Campagnola et al., 2022).

Важными параметрами являются также минимальная инфекционная доза (количество вирионов, необходимое для заражения), составляющая около 100 частиц (Karimzadeh et al., 2021), репродуктивное число (1.8–3.2) (Xu et al., 2021), количество вирусных частиц, которые переносит больной во время пика инфекции ($(1–100) \cdot 10^9$ шт.) и

Наиболее поздние представители различных ветвей филогенетического дерева коронавируса SARS-CoV-2 (https://nextstrain.org/ncov/open/global/all-time)

Имя вируса	Дата получения образца	Идентификатор последовательности в БД	Код классификатора Pangolin	Клада, вариант	Число мутаций	Число делеций	Длина генома
hCoV-19/Wuhan/ WIV04/2019 (референсный геном в GISAID)	30.12.2019	EPI_ISL_402124	B	19A	0	0	29891
Wuhan-Hu-1 (референсный геном в Genbank)	12.2019	NC_045512.2	B	19A	0	0	29903
hCoV-19/Tunisia/ S-1180/2021	29.10.2021	EPI_ISL_11333927	B.1.1.7	20I (Alpha, V1)	37	19	29758
hCoV-19/Madagascar/LA2M-112753/2021	16.01.2021	EPI_ISL_7722749	B.1.351.2	20H (Beta, V2)	31	18	29818
PHL/COVID-74517/2021	01.07.2021	OL629469	B.1.351	20H (Beta, V2)	32	9	29854
hCoV-19/Brazil/AM-IMTSP-CD24003/2021	10.08.2021	EPI_ISL_14800432	P.1.4	20J (Gamma, V3)	42	9	29772
LAO/LOMWRU-0461/2021	24.11.2021	OQ028273	P.1	20J (Gamma, V3)	32	18	29699
hCoV-19/Australia/WA11930/2023	28.02.2023	EPI_ISL_17187319	XBC.1.4	21I (Delta) XBC	77	36	29308
hCoV-19/Yunnan/ YNCDC-1019/2023	23.05.2023	EPI_ISL_17778593	DY.1	22B (Omicron)	89	59	29806
hCoV-19/Japan/TKYmbc38047/2023	06.06.2023	EPI_ISL_17941095	XBB.2.3.11	22F (XBB.2.3)	99	56	29726
hCoV-19/Heilong-jiang/HLJCDC-1665/2023	20.06.2023	EPI_ISL_17850574	XBB.1.5	23A (Omicron) (XBB.1.5)	103	56	29781

Примечание. Представители некоторых ветвей (в основном различных вариантов «Омикрона») продолжают встречаться среди секвенированных последовательностей недавно заболевших людей, а некоторые перестали обнаруживаться вовсе («Альфа», «Бета», «Гамма», «Дельта» и др.). Референсные последовательности в обеих базах различаются только длиной поли-А участка, расположенного в самом в конце, а в остальном совпадают.

число вирионов, в среднем содержащихся в зараженной клетке (10^5 шт.) (Sender et al., 2021), а также другие эпидемиологические характеристики. Вирусные частицы обнаруживаются во многих тканях и органах, от легких до мозга, однако выйдут наружу и могут быть переданы следующим носителям только те, что присутствуют в дыхательных путях или кишечнике. Все остальные вирионы не оставят «потомков», что ощутимо сужает эволюционный коридор. В работах (Day et al., 2020; Markov et al., 2023) рассмотрен ряд важных вопросов, касающихся эпидемиологии и эволюции вируса SARS-CoV-2, включая механизм возникновения рекомбинантных штаммов.

Материалы и методы

Наиболее рациональным с точки зрения как скорости обработки данных, так и обеспечения ничем не ограниченных возможностей (которые при необходимости можно расширять) при их анализе, на наш взгляд, является работа с исходными fasta-файлами с помощью программного комплекса, сочетающего наши собственные разработки со сторонними библиотеками и программами. К настоящему времени реализован прототип, включающий минимально необходимые функциональные возможности. Для разработки использовался язык программирования C++, среда разработки – Microsoft Visual Studio Community 2019. Аппаратное обеспечение – ПК на базе процессора Intel Core i7-10700K, 3.8 ГГц, 8 ядер, 16 Гб оперативной памяти.

Использованные в нашей работе методы в основном могут быть отнесены к следующим двум категориям:

– теоретические оценки и численные расчеты некоторых важных характеристик рассматриваемой системы;
– анализ доступных генетических данных с помощью собственного прикладного программного обеспечения и с помощью существующих программных средств.

Полногеномные генетические последовательности SARS-CoV-2. Базы данных GISAID и Genbank предоставляют посредством веб-интерфейса некоторый набор функциональных возможностей для изучения свойств содержащихся в них последовательностей, однако они недостаточно гибкие для осуществления анализа, который необходим для исследований пространства генетических вариантов SARS-CoV-2, являющихся целью настоящей работы. Для GISAID также существует API (Application Programming Interface – программный интерфейс приложения) (Wirth, Duchene, 2022), реализованный на языке R, однако и для этого способа имеются существенные ограничения (включая скорость работы при значительных объемах обрабатываемой информации) по сравнению с прямым доступом к генетическим последовательностям, хранящимся в виде fasta-файлов на локальной рабочей станции. GISAID существенно ограничивает закачку со своего сайта: за одну загрузку система позволяет скачать не более 2000 последовательностей, что полностью исключает возможность загрузки значимого объема данных «вручную». В NCBI SARS-CoV-2 Data Hub подобных ограничений нет.

Для анализа уже реализовавшихся генетических вариантов SARS-CoV-2 были использованы полногеномные последовательности из баз данных GISAID (https://gisaid.org/) и NCBI Virus SARS-CoV-2 Data Hub (https://www.ncbi.nlm.nih.gov/labs/virus/). Последовательности из Genbank за 2019–2020 гг. были скачаны на локальную вычислительную станцию и проанализированы с помощью

разработанного нами программного обеспечения ParSeq. Последовательности из GISAID, ввиду ограничений на скачивание, мы не загружали, воспользовавшись вместо этого доступом по API для получения лишь некоторых их характеристик (например, полной длины последовательности; при этом возможность определения длины транслируемой части или позиций ее начала и конца не поддерживается).

Для расчета редакционного расстояния между парами коронавирусных последовательностей, включая отдельный расчет количества мутаций, делеций и вставок, использовался веб-ресурс Nextstrain/Nextclade (https://clades.nextstrain.org).

Результаты

Оценка количества реализовавшихся и потенциальных генетических вариантов SARS-CoV-2

Начнем с рассмотрения пространства генетических последовательностей как такового с математической точки зрения в самом общем случае. Любая пара последовательностей характеризуется мерой различия между ними, называемой расстоянием Левенштейна, или редакционным расстоянием – минимальным количеством точечных (одиночных) замен (мутаций, делеций, вставок), которые необходимо сделать в первой последовательности, чтобы преобразовать ее во вторую. Каждый элемент множества последовательностей заданной длины L будет удален от пустой последовательности (Ø) ровно на L. Число вариантов нуклеотидных последовательностей длиной L составляет 4^L. Число возможных точечных мутаций для последовательности длиной L равно $3{\cdot}L$ (нуклеотид в каждой позиции может быть заменен на любой из трех других). Также возможно $3{\cdot}L$ одиночных делеций и $3{\cdot}(L{+}1)$ одиночных вставок. В результате всех возможных одиночных делеций для всех возможных последовательностей длиной L получается множество всех возможных последовательностей длиной $(L{-}1)$, с числом вариантов, равным $4^{(L-1)}$. А в результате всех возможных одиночных вставок – множество всех возможных последовательностей длиной $(L{+}1)$, с числом вариантов $4^{(L+1)}$.

Рассмотрим все возможные варианты нуклеотидных последовательностей длиной $L = 2$ (рис. 1). Последовательности длиной 2 нуклеотида – простой случай, однако даже для него уже необходим гиперкуб в четырехмерном пространстве (тессеракт с 16 вершинами). Для более сложного случая, $L = 4$, аналогичным образом может быть использован 6-мерный гиперкуб (гексеракт) с 64 вершинами, изображение которого вместе с подписями последовательностей и ребер будет перенасыщено деталями и затруднительно для восприятия. Однако его можно в некоторой степени отобразить на двумерной плоскости с помощью одного из вариантов кодов Грея (Mütze, 2023), теория которых тесно связана с гиперкубами, а именно 2D кода, который нам удалось подобрать для данного случая (рис. 2).

Привычная метрика, определяемая как корень из суммы квадратов разностей декартовых координат, в этом случае, как видно, не подходит.

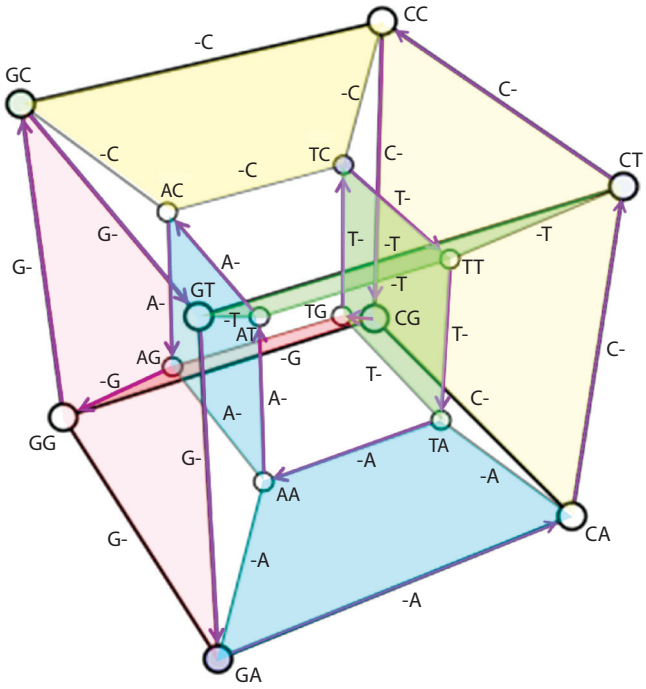


Рис. 1. Пространство вариантов нуклеотидных последовательностей длиной 2, представленное в виде гиперкуба.

Показан один из множества гамильтоновых циклов на гиперкубе (фиолетовые стрелки) – замкнутый путь, проходящий через каждую вершину ровно один раз. Каждый переход соответствует единичному изменению (мутации, делеции или вставке). Также имеются гиперплоскости, которые можно сопоставить подпоследовательностям меньшей длины, получающимся в данном случае, при $L = 2$, посредством делеций слева (-A,-T,-G,-C) и справа (A-,T-,G-,C-), которые для этого простого случая получаются одними и теми же.

Число всех возможных последовательностей такой же длины, как и длина референсного генома SARS-CoV-2, $L = 29\,903$, составляет астрономическое число $4^{29\,903}$, или $\approx 2.511{\cdot}10^{18\,003}$. В этом пространстве вариантов множество последовательностей, соответствующих реализованным вариантам генома SARS-CoV-2, составляет лишь малую часть – точку, соответствующую исходной референсной последовательности, и ее небольшую окрестность, ограниченную на данный момент расстоянием от референсной последовательности до наиболее современного штамма «Омикрона». Можно оценить количество возможных вариантов последовательностей в пределах этой дистанции. Для референсной последовательности с $L = 29\,903$ число ее различных вариаций с одной одиночной мутацией равно $3{\cdot}L$, с двумя мутациями – $(3{\cdot}L)^2 - 3{\cdot}L = 3{\cdot}L{\cdot}(3{\cdot}L{-}1)$ (вычитаем из всех возможных вариантов все те случаи, когда вторая мутация произойдет в той же позиции, что и первая, и тогда получится одна из уже существующих последовательностей – референсная или отличающаяся от нее на 1). Аналогично для третьей мутации: $(3{\cdot}L)^3 - ((3{\cdot}L)^2 - 3{\cdot}L)$, и так далее. Для $L = 29\,903$ получаем, что количество всех вариантов последовательностей с числом мутаций от 0 до n (относительно референсной последовательности) для $n = 103$ составляет $1.387{\cdot}10^{510}$, аналогично для $L = 29\,847$ (56 делеций) – $1.108{\cdot}10^{510}$. Суммируя по всем длинам от 29 903 до 29 847, получаем $7.190{\cdot}10^{511}$.

A = 00	T = 01	G = 10	C = 11		AA	GA	CA	TA	TT	AT	GT	CT	CC	TC	AC	GC	GG	CG	TG	AG
1	0	0	0	0	AA	AAAA	GAAA	CAAA	TAAA	TTAA	ATAA	GTAA	CTAA	CCAA	TCAA	ACAA	GCAA	GGAA	CGAA	TGAA	AGAA
2	1	0	0	0	GA	AAGA	GAGA	CAGA	TAGA	TTGA	ATGA	GTGA	CTGA	CCGA	TCGA	ACGA	GCGA	GGGA	CGGA	TGGA	AGGA
3	1	1	0	0	CA	AACA	GACA	CACA	TACA	TTCA	ATCA	GTCA	CTCA	CCCA	TCCA	ACCA	GCCA	GGCA	CGCA	TGCA	AGCA
4	0	1	0	0	TA	AATA	GATA	CATA	TATA	TTTA	ATTA	GTTA	CTTA	CCTA	TCTA	ACTA	GCTA	GGTA	CGTA	TGTA	AGTA
5	0	1	0	1	TT	AATT	GATT	CATT	TATT	TTTT	ATTT	GTTT	CTTT	CCTT	TCTT	ACTT	GCTT	GGTT	CGTT	TGTT	AGTT
6	0	0	0	1	AT	AAAT	GAAT	CAAT	TAAT	TTAT	ATAT	GTAT	CTAT	CCAT	TCAT	ACAT	GCAT	GGAT	CGAT	TGAT	AGAT
7	1	0	0	1	GT	AAGT	GAGT	CAGT	TAGT	TTGT	ATGT	GTGT	CTGT	CCGT	TCGT	ACGT	GCGT	GGGT	CGGT	TGGT	AGGT
8	1	1	0	1	CT	AACT	GACT	CACT	TACT	TTCT	ATCT	GTCT	CTCT	CCCT	TCCT	ACCT	GCCT	GGCT	CGCT	TGCT	AGCT
9	1	1	1	1	CC	AACC	GACC	CACC	TACC	TTCC	ATCC	GTCC	CTCC	CCCC	TCCC	ACCC	GCCC	GGCC	CGCC	TGCC	AGCC
10	0	1	1	1	TC	AATC	GATC	CATC	TATC	TTTC	ATTC	GTTC	CTTC	CCTC	TCTC	ACTC	GCTC	GGTC	CGTC	TGTC	AGTC
11	0	0	1	1	AC	AAAC	GAAC	CAAC	TAAC	TTAC	ATAC	GTAC	CTAC	CCAC	TCAC	ACAC	GCAC	GGAC	CGAC	TGAC	AGAC
12	1	0	1	1	GC	AAGC	GAGC	CAGC	TAGC	TTGC	ATGC	GTGC	CTGC	CCGC	TCGC	ACGC	GCGC	GGGC	CGGC	TGGC	AGGC
13	1	0	1	0	GG	AAGG	GAGG	CAGG	TAGG	TTGG	ATGG	GTGG	CTGG	CCGG	TCGG	ACGG	GCGG	GGGG	CGGG	TGGG	AGGG
14	1	1	1	0	CG	AACG	GACG	CACG	TACG	TTCG	ATCG	GTCG	CTCG	CCCG	TCCG	ACCG	GCCG	GGCG	CGCG	TGCG	AGCG
15	0	1	1	0	TG	AATG	GATG	CATG	TATG	TTTG	ATTG	GTTG	CTTG	CCTG	TCTG	ACTG	GCTG	GGTG	CGTG	TGTG	AGTG
16	0	0	1	0	AG	AAAG	GAAG	CAAG	TAAG	TTAG	ATAG	GTAG	CTAG	CCAG	TCAG	ACAG	GCAG	GGAG	CGAG	TGAG	AGAG

Рис. 2. Множество вариантов нуклеотидных последовательностей длиной 4, изображенное на плоскости с использованием 2D кодов Грея.

Верхний край таблицы стыкуется с нижним, а левый – с правым, т. е. можно отобразить это множество на поверхность тора. Тогда при движении как по горизонтали, так и по вертикали (в системе координат таблицы), в соответствии со свойствами кодов Грея, каждая пара соседних последовательностей будет отличаться ровно на одну точечную замену.

Последовательности с синонимичными однонуклеотидными мутациями, не приводящими к замене аминокислоты, тоже являются частью полного пространства вариантов. Однако реальное число вариантов в контексте рассмотрения структуры и функций белков, транслируемых с вирусной РНК, существенно меньше из-за вырожденности генетического кода (20 аминокислот кодируются 61 триплетом РНК, т. е. в среднем имеем 3.05 триплета, кодирующего одну и ту же аминокислоту). Также учтем, что белки кодирует не весь геном SARS-CoV-2, 771 из 29903 нт является некодирующим. В результате зависимость, пропорциональная $(3L)^n$, преобразуется в $\approx((L-771)+(3\cdot771))^n$, и, таким образом, скорректированное число вариантов белковых последовательностей может быть оценено как $1.02\cdot10^{465}$. Если же предположить, что когда-нибудь число мутаций превысит вышеупомянутые 103 шт. в 10–11 раз, то последовательность, скорее всего, все еще будет коронавирусной, но уже будет принадлежать другому виду. К примеру, коронавирус летучих мышей RaTG13, ближайший сосед SARS-CoV-2 в пространстве вариантов генетических последовательностей, отличается от него на 1135 точечных мутаций.

Попробуем взглянуть на множество испробованных природой вариантов с биологической точки зрения. Вирус попадает в организм (как правило, воздушно-капельным путем), оказывается в легких и проникает в клетку, где рибосома носителя начинает синтезировать вирусные белки в соответствии с последовательностью нуклеотидов в геноме SARS-CoV-2. Среди этих белков – вирусная РНК-полимераза (RdRp), которая начинает репликацию вирусной РНК. Поначалу, когда в клетке находится одна вирусная РНК и одна RdRp, вероятность их встречи чрезвычайно мала, но затем, по мере накопления тех и других молекул в клетке, она начинает стремительно расти. В итоге концентрация достигает уровня, достаточного для осуществления сборки новых вирионов, и когда их количество в клетке достигает примерно 10^5 шт., клетка разрушается, и эти вирионы начинают заражать как соседние клетки, так и все прочие – если часть вирионов попадет в кровоток и будет разнесена по организму. Учитывая, что количество вирусных частиц, переносимых больным во время пика инфекции, может достигать 10^{11} шт. (Sender et al., 2021), разделим это значение на среднее число вирионов в зараженной клетке и получим 10^6 зараженных клеток в организме. У человека примерно $3\cdot10^{13}$ клеток, т. е. заражено оказывается менее 10^{-4} %.

Частота возникновения ошибок при репликации SARS-CoV-2 составляет, согласно (Amicone et al., 2022), $1.3\cdot10^{-6}\pm0.2\cdot10^{-6}$ замен на позицию за один инфекционный цикл заражения клетки, а по другим данным – $(1–2)\cdot10^{-6}$ на позицию (за цикл репликации) (Markov et al., 2023), т. е. в среднем примерно $1.4\cdot10^{-6}$. С учетом длины последовательности получаем вероятность возникновения одной мутации на всю последовательность за цикл репликации ≈0.04. Даже если все зараженные клетки в организме в какой-то момент будут содержать один и тот же вариант вирусной РНК, то спустя один цикл репликации в организме могут оказаться все возможные варианты одиночных замен (3·29903 шт.) относительно исходной вирусной РНК (до начала этого цикла). Таких будет около 4 %, большинство из которых нежизнеспособны, а 96 % окажутся точными копиями реплицированной последовательности. Какова при этом вероятность возникновения жизнеспособной несинонимичной мутации (изменяющей не только последовательность РНК вируса, но и аминокислотную последовательность одного из его белков), к тому же превосходящей предшественника по приспособленности? Этот вопрос остается открытым, однако искомая вероятность определенно будет весьма незначительной. В подавляющем большинстве случаев

все экземпляры вируса, распространяемые заболевшим во внешнюю среду, являются идентичными, и лишь изредка в одном организме встречаются одновременно два варианта. Каким же образом новые мутантные варианты не просто появляются, но и достаточно быстро вытесняют своих предшественников в масштабах планеты?

Учитывая, что соотношение 4 % : 96 % с каждым последующим циклом репликации будет изменяться в сторону уменьшения доли мутантных последовательностей («эффект основателя» (Ruan et al., 2020)) до полного их вытеснения, возможными сценариями возникновения и распространения мутантных вариантов SARS-CoV-2 представляются следующие довольно маловероятные события.

(а) Иммунитета от SARS-CoV-2 у организма нет, он с ним еще не встречался. В клетку попал единственный экземпляр вирусной РНК, при первом же цикле репликации в нем возникла мутация, и она оказалась жизнеспособной (такое может произойти при заражении коронавирусом хотя и с малой, но ненулевой вероятностью). Тогда все новые вирионы, синтезированные этой клеткой, будут носителями данной мутации, и если она заметно увеличивает их приспособленность, то есть шансы, что в итоге они вытеснят исходный вариант.

(б) У организма уже есть иммунитет от SARS-CoV-2. В него попадает одновременно два варианта вирионов SARS-CoV-2 – доминирующий в популяции и новый, мутантный (возникший по механизму из (а) или рекомбинант). Иммунная система уничтожает знакомый ей «старый» вариант, а новый остается незамеченным, и в результате размножается и передается дальше именно он.

Вероятности возникновения двух этих вариантов еще предстоит оценить, однако и без того видно, что коридор возможных вариантов, по которому прошла эволюция, оказался достаточно узким. Противоположность этой картине представляет, например, вирус гриппа, отличительной особенностью и основой выживания которого является высокая изменчивость, обусловленная механизмами антигенного дрейфа и антигенного сдвига (Kim et al., 2018).

Мы оцениваем моделирование эволюции SARS-CoV-2 как возможное, поскольку, несмотря на большое количество вариантов, которые уже должны были реализоваться и которые могли бы реализоваться с точки зрения математики (теории вероятности) и биологии, в реальности была реализована лишь малая их часть, и мы наблюдаем лишь малую часть возможного пространства вариантов.

Разработка программного обеспечения ParSeq

Для анализа генетических последовательностей SARS-CoV-2 нами было разработано программное обеспечение, названное ParSeq (**Par**ser of **Seq**uences) – парсер и анализатор нуклеотидных последовательностей в формате FASTA, которую мы также использовали при анализе последовательностей SARS-CoV-2 в регионах Сибирского федерального округа (Palyanova et al., 2023). Ниже описаны его основные, уже реализованные на данный момент функциональные возможности.

- Загрузка списка fasta-файлов для анализа и последовательное чтение каждого из них, включая разбор заголовка (содержащего поля Accession ID, Length, Pangolin, Nuc. Completeness, Collection Date, Geo Location, Country) и загрузку нуклеотидной последовательности.
- Первичный анализ нуклеотидной последовательности, включая расчет ее длины, содержания нуклеотидов A, U(T), G, C и неидентифицированных нуклеотидов, обозначаемых буквой “N”. Также в некоторых последовательностях иногда встречаются следующие буквы расширенного алфавита (https://www.bioinformatics.org/sms/iupac.html):

R	Y	S	W	K
A ‖ G	C ‖ T	G ‖ C	A ‖ T	G ‖ T
M	B	D	H	Y
A ‖ C	C ‖ G ‖ T	A ‖ G ‖ T	A ‖ C ‖ T	A ‖ C ‖ G

- Определение позиций начала и конца кодирующей части последовательности. В случае референсной последовательности ее полная длина составляет 29903 нт, длина некодирующего 5′ UTR участка – 265 нт, некодирующего 3′ UTR участка – 229 нт. Для этого используется следующий довольно очевидный алгоритм: в случае 5′ UTR движемся вдоль последовательности от ее начала до 500-го нуклеотида (для удобства выбрано «круглое» значение, при котором 265 находится примерно посередине) окном длиной 17 и считаем число совпадений нуклеотидов в этом окне с фрагментом референсной последовательности той же длины из интервала 266–282 (266 – позиция старта трансляции в референсном геноме). Если из 17 совпадают 14 и более, то считаем, что позиция определена (значения подобраны как достаточные для корректной работы в подавляющем большинстве случаев при малой длине окна, чтобы избежать лишних вычислений). В случае 3′ UTR аналогично: движемся окном длиной 17 от позиции L–500 до конца анализируемой последовательности, сравнивая его содержимое с 17 нуклеотидами, которыми оканчивается кодирующий участок референсной последовательности. Критерий тот же – 14 или более совпадений.
- Расчет длин некодирующих 5′ UTR и 3′ UTR, а также находящегося между ними кодирующего участка, составляющего подавляющую часть генома вирусной последовательности (98.35 % его длины в случае референсной последовательности).
- Расчет распределений этих значений для любой выборки последовательностей геномов SARS-CoV-2 (например, в пределах указанного интервала времени для даты получения образца, который был секвенирован, или содержащих не более заданного числа NNN, или и то и другое одновременно, и т. п.).
- Расчет числа вариантов последовательностей в пределах интервала длин, присутствующих в базах данных.

Результаты, полученные с помощью ParSeq

С помощью разработанного нами программного обеспечения был осуществлен анализ нуклеотидных последовательностей SARS-CoV-2, доступных пользователям по

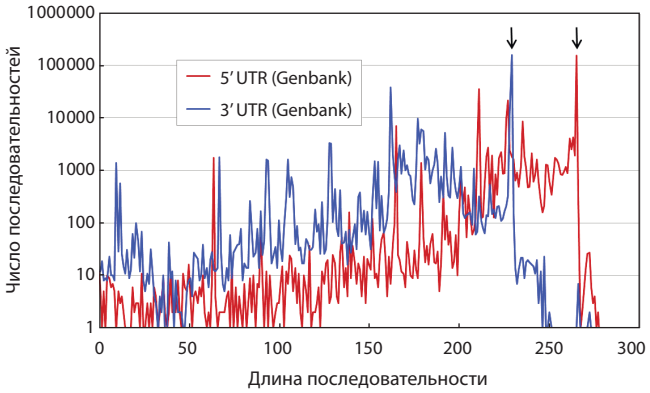


Рис. 3. Распределение длин 5′ UTR и 3′ UTR для последовательностей из базы данных Genbank за период с момента появления SARS-CoV-2 в конце 2019 г. до конца 2020 г.

Длины 5′ UTR и 3′ UTR в референсном геноме SARS-CoV-2 составляют 265 и 229 нт соответственно. Пиковые значения обеих кривых соответствуют именно этим длинам.

всему миру благодаря проектам Genbank и GISAID, их базам данных и веб-ресурсам. В результате были установлены следующие факты.

1. Расчет распределения генетических последовательностей по их полным длинам (5′ UTR + кодирующая последовательность + 3′ UTR) среди последовательностей, имеющих длину ≥ 28000, выявил, что для данных из Genbank (за период с 01.12.2019 по 31.12.2022) минимальное значение длины полной последовательности составило 28784, а максимальное – 29985. Распределение практически полностью расположено левее длины референсной последовательности, составляющей 29903. Таким образом, разница между референсным и минимальным значением длины составила 1119. Скорее всего, такие слишком короткие или слишком длинные последовательности соответствуют данным низкого качества, с ошибками сборки генома, поскольку они плохо соотносятся с данными таблицы, согласно которым максимальная разница между длиной референсной и какой-либо другой последовательности составляет около 159 (103 мутации + 56 делеций). Более того, при таком различии эта последовательность, скорее всего, принадлежала бы уже другому виду вирусов, так как похожую разницу имеют референсная последовательность SARS-CoV-2 и коронавирус летучих мышей RaTG13 (GenBank MN996532.2, collection_date=24-Jul-2013). По данным (Li et al., 2023), они отличаются на 96.2 %, т. е. на 1136 одиночных мутаций (достаточно равномерно рассредоточенных по последовательности). Расчет расстояния между этими же последовательностями, произведенный с помощью веб-сервиса Nextstrain, показал разницу в 1135 одиночных мутаций, а также 20 делеций (в кодирующей части RaTG13 относительно референсной последовательности SARS-CoV-2). Полная длина генома RaTG13 составляет 29855, т. е. число делеций относительно SARS-CoV-2 составляет 48.

Поскольку разница между полной длиной референсного генома SARS-CoV-2 и остальными содержащимися в базе данных последовательностями для некоторых из них существенно превышает число различий (точечных мутаций, делеций и вставок) между референсным геномом SARS-CoV-2 и наиболее отличным от него современным вариантом «Омикрон» (см. таблицу, последняя строка), мы решили исследовать распределение как полных длин геномов, так и их кодирующих и некодирующих участков (рис. 3 и 4). Как видно на рис. 3, участки 5′ UTR и 3′ UTR, встречающиеся в базах данных, обладают длинами от нуля до референсных значений, а также в малом количестве случаев незначительно превосходят их. Последователь-

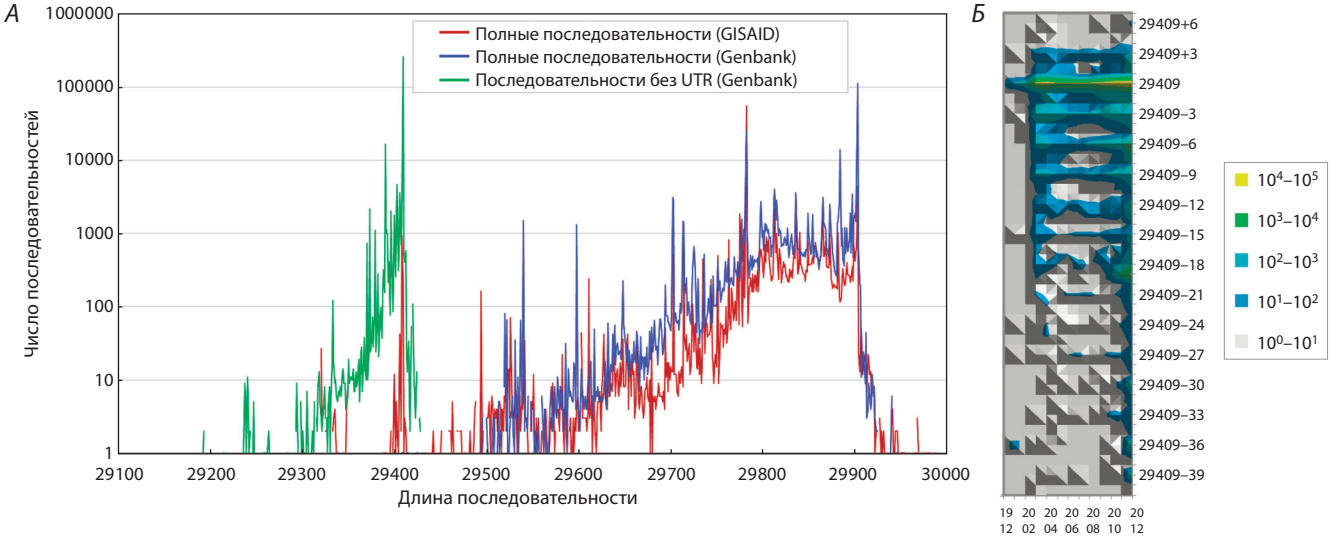


Рис. 4. *А* – распределение полных геномов (GISAID, Genbank) и геномов без UTR (Genbank) за период с момента появления SARS-CoV-2 до конца 2020 г. Длина полного референсного генома SARS-CoV-2 составляет 29903 нт, его же без UTR – 29409 нт. Пиковые значения всех трех кривых соответствуют этим длинам. *Б* – изменение распределения длин геномов без UTR (Genbank) за 2019–2020 гг. по месяцам. По горизонтали – год и месяц, по вертикали – длина генома без UTR, цвета соответствуют числу последовательностей (логарифмическая шкала).

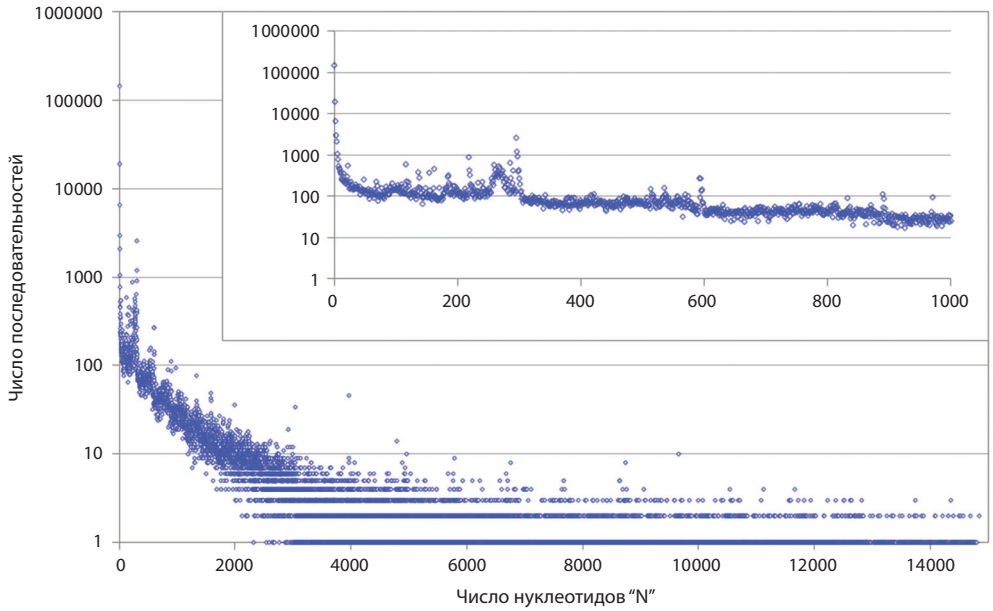


Рис. 5. Соотношение числа неидентифицированных или частично идентифицированных нуклеотидов в транслируемой части геномов SARS-CoV-2, содержащихся в Genbank, полученных в период с конца 2019 г. (начало пандемии) до конца 2020 г.

Врезка содержит часть того же графика, что и на основной картинке, но для области от 0 до 1000 по горизонтали.

ности, длины 5′ UTR и 3′ UTR которых совпадают с референсными, составляют 49.7 и 51.2 % от их полного числа соответственно. Последовательности, длины 5′ UTR и 3′ UTR которых отличаются от референсных не более чем на 10 нт, составляют 55.9 и 55.7 % от их полного числа соответственно.

Из рис. 4, *А* также видно, что основным источником наблюдаемого разброса значений полных длин геномов SARS-CoV-2 действительно были нетранслируемые участки – 5′ и 3′ UTR. Если рассматривать только кодирующую часть, то разброс значительно сокращается: 84.9 % всех последовательностей имеют такую же длину кодирующей части, как и референсный геном, а 90.7 % – длину кодирующей части, отличающуюся от таковой у референсного генома не более чем на 10 нт. Впрочем, среди геномов, длина кодирующей части которых отличается от таковой у референсной последовательности (29409), преобладают те, у которых это отличие кратно трем, – для предотвращения сдвига рамки считывания при трансляции, что обычно приводит к нежизнеспособности (см. рис. 4, *Б*). Таким образом, большинство вирусных последовательностей представляются биологически осмысленными.

Видно, что распределения, полученные на основе данных о полных геномах из GISAID (посредством программных запросов с использованием API) и Genbank (посредством анализа скачанных последовательностей с помощью разработанных нами программных средств) имеют достаточно высокое сходство – вероятно, по причине того, что большинство последовательностей содержатся в обеих базах данных (см. рис. 4). Вопрос о том, сколько последовательностей, отличающихся по длине от референсной, действительно имеют делеции или вставки, а сколько имеют эти отличия из-за ошибок секвенирования и сборки геномов, остается открытым.

2. При изучении генетических последовательностей, представляющих геномы различных вариантов вируса, изменяющиеся с течением времени, часто возникает необходимость сравнения их между собой. Даже если у пары рассматриваемых последовательностей одинаковые длины кодирующих участков, возможность вычислить величину различия между ними (число точечных мутаций) будет зависеть от того, присутствуют ли в этих последовательностях неопределенные нуклеотиды, обычно обозначаемые "N", или другие буквы, помимо стандартных A, T(U), G и C. Используя разработанное нами программное обеспечение и геномы вариантов SARS-CoV-2, полученные в 2019–2020 гг. (содержащиеся в базе данных Genbank), мы рассчитали соотношение между числом последовательностей и числом неидентифицированных или частично идентифицированных нуклеотидов в них (рис. 5).

На большей части графика число последовательностей экспоненциально уменьшается с ростом числа неидентифицированных нуклеотидов, хотя и встречаются участки с некоторыми особенностями. Число последовательностей, в которых все нуклеотиды определены, составляет 47.8 %, а число последовательностей, где неопределенными являются менее 10 нуклеотидов – 58.9 %. Таким образом, для анализа эволюционных изменений, происходящих с вирусом, остается весьма значительная доля от общего числа последовательностей, хранящихся в базе данных.

Обсуждение

Мы осуществили ряд оценок, расчетов и компьютерных вычислений, в том числе с помощью разработанных нами

программных средств, для улучшения понимания того, каким является пространство вариантов генетических последовательностей коронавируса SARS-CoV-2, каковы его основные свойства и особенности, связанные с достаточно длинной геномной последовательностью вируса и низкой для РНК-вирусов вероятностью возникновения мутаций.

Из-за относительно большой длины генома SARS-CoV-2 количество его жизнеспособных вариантов значительно превышает таковое для вирусов, обладающих в несколько раз более коротким геномом. Попробуем определить некоторые другие ориентиры в пространстве вариантов генетических последовательностей. SARS-CoV-2 относится к одноцепочечным РНК(+) вирусам (Modrow et al., 2013). Один из самых маленьких оцРНК(+) вирусов человека – это астровирус 1-го типа (длина генома 6771 нуклеотид) (Lewis et al., 1994). Еще меньшим геномом среди вирусов этого класса обладает нодавирус креветки (*Penaeus vannamei nodavirus*) (Chen et al., 2019) – 4294 нуклеотида. Полное число вариантов различных последовательностей для этих двух длин составляет $3.533 \cdot 10^{4076}$ и $1.760 \cdot 10^{2585}$ соответственно.

Если расширить пространство поиска вирусов с самым малым геномом до ДНК-вирусов, то среди рекордсменов обнаружится цирковирус свиней первого типа, *Porcine circovirus* 1 (PCV1) (Cao et al., 2018), размер генома которого составляет всего 1757–1759 пар нуклеотидов (в 17 раз меньше, чем у SARS-CoV-2). Число возможных вариантов последовательности такой длины составляет $6.597 \cdot 10^{1057}$. Это по-прежнему очень далеко от числа вариантов, которые были потенциально доступны коронавирусу SARS-CoV-2 за период его существования (3.5 года), – $7.985 \cdot 10^{511}$. Весьма близким числом всех возможных вариантов последовательностей, равным $5.636 \cdot 10^{511}$, обладал бы геном размером 850 нт. Впрочем, существуют инфекционные агенты на основе одноцепочечной кольцевой РНК с еще меньшими длинами последовательностей (от 246 до 467 нт) – вироиды (Katsarou et al., 2015). Их РНК не защищена какой-либо оболочкой и не кодирует белков.

Число всех потенциально возможных нуклеотидных последовательностей, которые были бы идентифицированы как варианты SARS-CoV-2, на много порядков превышает число как тех вариантов этого вируса, которые уже были обнаружены и секвенированы, так и тех, которые были опробованы в ходе эволюции, но оказались нежизнеспособными. Рассмотрим коронавирус летучих мышей RaTG13 ($L = 29855$), являющийся ближайшим известным соседом SARS-CoV-2 (но при этом уже другим вирусом) в пространстве вариантов генетических последовательностей, отличающийся от него на 1135 точечных мутаций. Число вариантов последовательностей, отличающихся от референсного генома SARS-CoV-2 не более чем на 1135 мутаций, составляет $\approx 2.943 \cdot 10^{5621}$, что более чем на 1000 порядков превышает полное количество возможных вариантов последовательностей длиной как 4294 ($1.76 \cdot 10^{2585}$), так и 6771 ($3.53 \cdot 10^{4076}$), т. е. может содержать внутри себя объемы информации, которых хватит на огромное количество различных небольших вирусов.

Глобальное филогенетическое древо коронавируса показывает, что вирус с течением времени подвержен изменениям (возможно, вынужденным) – видимо, в связи с тем, что на него действует давление естественного отбора. Еще одной причиной изменений является внутривидовая конкуренция; например, варианты с более быстрыми РНК-полимеразами вытесняют варианты с более медленными (поскольку их число растет быстрее) и тем самым уменьшают инкубационный период вируса, а менее летальные штаммы позволяют вирусу дольше и шире распространяться (носитель не умирает, а является распространителем вируса на протяжении почти всего периода заболевания; легко перенося болезнь, человек остается социально активным и заражает больше других людей в своем окружении). В отличие от упомянутых выше вироидов, изменения в геноме настоящих вирусов, в том числе SARS-CoV-2, могут оказывать различный эффект на внутривидовую конкуренцию в зависимости от функций белков, закодированных в геноме. Этот вопрос остался за рамками данной работы, однако в последующих публикациях мы планируем уделить ему должное внимание.

Кроме того, влияние оказывает и формирование у человечества иммунитета к этому вирусу. Вероятно, имеются и другие механизмы. При этом все эти изменения должны происходить не в ущерб функциональным возможностям вируса. Таким образом, получается, что пространство доступных коронавирусу SARS-CoV-2 вариантов является довольно узким, а траектории его развития, возможно, в некоторой степени детерминированы. И действительно, было показано, что геном SARS-CoV-2 имеет гораздо более низкую частоту мутаций и генетическое разнообразие по сравнению с вирусом SARS-CoV, вызвавшим вспышку атипичной пневмонии в 2002–2003 гг. (Jia et al., 2020; Zhou et al., 2020; Никонова и др., 2021). Так, например, для S-белка значения d_N и d_S для SARS-CoV-2 оказались приблизительно в 12 и 7 раз ниже, чем для SARS-CoV (d_N – доля последовательностей в выборке геномов, в которых присутствуют несинонимичные мутации в определенном гене, d_S – аналогично доля синонимичных мутаций). Для более консервативных генов ORF1a и ORF1b отношения частот мутаций

$$(d_N^{\text{SARS-CoV-2}}/d_N^{\text{SARS-CoV}},\ d_S^{\text{SARS-CoV-2}}/d_S^{\text{SARS-CoV}})$$

имеют меньшую величину, но значения для SARS-CoV-2 ниже, чем для SARS-CoV (лежат в пределах интервала от $\frac{1}{4.8}$ до $\frac{1}{1.5}$). Гипотеза о частичной детерминированности траекторий развития вируса состоит в том, что если бы развитие пандемии SARS-CoV-2, с самого ее начала в декабре 2019 г., в силу случайных факторов пошло бы несколько иначе, то, несмотря на это, раньше или позже, в том же порядке или в ином пространство жизнеспособных вариантов, «посещенных» вирусом, все равно оказалось бы примерно таким же. Вышесказанное позволяет предположить возможность создания эволюционного симулятора на основе анализа траекторий изменения вируса с течением времени, что входит в наши планы на будущее в рамках работы по данной тематике.

Заключение

Изучение пространства вариантов генетических последовательностей – важный этап в разработке подходов к моделированию эволюции вирусов и других организмов. Для построения новой, существенно более реалистичной модели эволюции вируса, способной рассчитывать потенциально возможные, но еще не реализованные в природе варианты новых геномов, чтобы заблаговременно противостоять их появлению, необходимо ответить на такие вопросы как: «Какова вероятность возникновения рекомбинантных вариантов вируса и имеются ли предпочитаемые позиции, по которым происходит обмен частями генома?», «Можем ли мы предположить или рассчитать, какой вариант реализуется, а какой окажется нежизнеспособным?», «Можно ли было предсказать “Дельту” или “Омикрон”?» и, наконец, «Если бы удалось создать реалистичную модель эволюции SARS-CoV-2 и несколько раз рассчитать процесс с самого начала, от исходной референсной последовательности, каждый раз он протекал бы по-разному и приводил к существенно различающимся результатам, или все происходило бы примерно одинаково с небольшими вариациями?»

Список литературы / References

Никонова А.А., Файзулоев Е.Б., Грачева А.В., Исаков И.Ю., Зверев В.В. Генетическое разнообразие и эволюция биологических свойств коронавируса SARS-CoV-2 в условиях глобального распространения. *Acta Naturae*. 2021;13(3):77-89. DOI 10.32607/actanaturae.11337
[Nikonova A.A., Faizuloev E.B., Gracheva A.V., Isakov I.Yu., Zverev V.V. Genetic diversity and evolution of the biological features of the pandemic SARS-CoV-2. *Acta Naturae*. 2021;13(3):77-89. DOI 10.32607/actanaturae.11337]

Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Software*. 2021;6(67):3773. DOI 10.21105/joss.03773

Amicone M., Borges V., Alves M.J., Isidro J., Zé-Zé L., Duarte S., Vieira L., Guiomar R., Gomes J.P., Gordo I. Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health*. 2022;10(1):142-155. DOI 10.1093/emph/eoac010

Campagnola G., Govindarajan V., Pelletier A., Canard B., Peersen O.B. The SARS-CoV nsp12 polymerase active site is tuned for large-genome replication. *J. Virol.* 2022;96(16):e0067122. DOI 10.1128/jvi.00671-22

Cao L., Sun W., Lu H., Tian M., Xie C., Zhao G., Han J., Wang W., Zheng M., Du R., Jin N., Qian A. Genetic variation analysis of PCV1 strains isolated from Guangxi Province of China in 2015. *BMC Vet. Res.* 2018;14(1):43. DOI 10.1186/s12917-018-1345-z

Chen N.C., Yoshimura M., Miyazaki N., Guan H.-H., Chuankhayan P., Lin C.-C., Chen S.-K., Lin P.-J., Huang Y.-C., Iwasaki K., Nakagawa A., Chan S.I., Chen C.J. The atomic structures of shrimp nodaviruses reveal new dimeric spike structures and particle polymorphism. *Commun. Biol.* 2019;2:72. DOI 10.1038/s42003-019-0311-z

Day T., Gandon S., Lion S., Otto S.P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* 2020;30(15):R849-R857. DOI 10.1016/j.cub.2020.06.031

Grebennikov D., Kholodareva E., Sazonov I., Karsonova A., Meyerhans A., Bocharov G. Intracellular life cycle kinetics of SARS-CoV-2 predicted using mathematical modelling. *Viruses*. 2021;13(9):1735. DOI 10.3390/v13091735

Jia Y., Shen G., Nguyen S., Zhang Y., Huang K., Ho H., Hor W., Yang C., Bruning J.B., Li C., Wang W. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. *bioRxiv.* 2020. DOI 10.1101/2020.04.09.034942

Karimzadeh S., Raj B., Nguyen T.H. Review of infective dose, routes of transmission and outcome of COVID-19 caused by the SARS-COV-2: comparison with other respiratory viruses. *Epidemiol. Infect.* 2021;149:e96. DOI 10.1017/S0950268821000790

Katsarou K., Rao A.L.N., Tsagris M., Kalantidis K. Infectious long non-coding RNAs. *Biochimie*. 2015;117:37-47. DOI 10.1016/j.biochi.2015.05.005

Khare S., Gurry C., Freitas L., Schultz M.B., Bach G., Diallo A., Akite N., Ho J., Lee R.T., Yeo W., Curation Team GC, Maurer-Stroh S. GISAID’s role in pandemic response. *China CDC Weekly*. 2021;3(49):1049-1051. DOI 10.46234/ccdcw2021.255

Kim H., Webster R.G., Webby R.J. Influenza virus: dealing with a drifting and shifting pathogen. *Viral Immunol.* 2018;31(2):174-183. DOI 10.1089/vim.2017.0141

Lewis T.L., Greenberg H.B., Herrmann J.E., Smith L.S., Matsui S.M. Analysis of astrovirus serotype 1 RNA, identification of the viral RNA-dependent RNA polymerase motif, and expression of a viral structural protein. *J. Virol.* 1994;68(1):77-83. DOI 10.1128/JVI.68.1.77-83.1994

Li P., Hu J., Liu Y., Ou X., Mu Z., Lu X., Zan F., Cao M., Tan L., Dong S., Zhou Y., Lu J., Jin Q., Wang J., Wu Z., Zhang Y., Qian Z. Effect of polymorphism in *Rhinolophus affinis* ACE2 on entry of SARS-CoV-2 related bat coronaviruses. *PLoS Pathog.* 2023;19(1):e1011116. DOI 10.1371/journal.ppat.1011116

Malone B., Urakova N., Snijder E.J., Campbell E.A. Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat. Rev. Mol. Cell Biol*. 2022;23(1):21-39. DOI 10.1038/s41580-021-00432-z

Markov P.V., Ghafari M., Beer M., Lythgoe K., Simmonds P., Stilianakis N.I., Katzourakis A. The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 2023;21(6):361-379. DOI 10.1038/s41579-023-00878-2

Modrow S., Falke D., Truyen U., Schätzl H. Viruses with single-stranded, positive-sense RNA genomes. In: Molecular Virology. Berlin: Springer, 2013;185-349. DOI 10.1007/978-3-642-20718-1_14

Mütze T. Combinatorial Gray codes – an updated survey. *Electron. J. Comb.* 2023;30(3):DS26. DOI 10.37236/11023

Palyanova N., Sobolev I., Alekseev A., Glushenko A., Kazachkova E., Markhaev A., Kononova Y., Gulyaeva M., Adamenko L., Kurskaya O., Bi Y., Xin Y., Sharshov K., Shestopalov A. Genomic and epidemiological features of COVID-19 in the Novosibirsk region during the beginning of the pandemic. *Viruses.* 2022;14(9):2036. DOI 10.3390/v14092036

Palyanova N.V., Sobolev I.A., Palyanov A.Y., Kurskaya O.G., Komissarov A.B., Danilenko D.M., Fadeev A.V., Shestopalov A.M. The development of the SARS-CoV-2 epidemic in different regions of Siberia in the 2020–2022 period. *Viruses*. 2023;15:2014. DOI 10.3390/v15102014

Ruan Y., Luo Z., Tang X., Li G., Wen H., He X., Lu X., Lu J., Wu C.I. On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? *Natl. Sci. Rev.* 2020;8(1):nwaa246. DOI 10.1093/nsr/nwaa246

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022;50(D1):D20-D26. DOI 10.1093/nar/gkab1112

Sender R., Bar-On Y.M., Gleizer S., Bernshtein B., Flamholz A., Phillips R., Milo R. The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci. USA.* 2021;118(25):e2024815118. DOI 10.1073/pnas.2024815118

Shannon A., Selisko B., Le N.T., Huchting J., Touret F., Piorkowski G., Fattorini V., Ferron F., Decroly E., Meier C., Coutard B., Peersen O.,

Canard B. Rapid incorporation of Favipiravir by the fast and permissive viral RNA polymerase complex results in SARS-CoV-2 lethal mutagenesis. *Nat. Commun*. 2020;11(1):4682. DOI 10.1038/s41467-020-18463-z

Sonnleitner S.T., Prelog M., Sonnleitner S., Hinterbichler E., Halbfurter H., Kopecky D.B.C., Almanzar G., Koblmüller S., Sturmbauer C., Feist L., Horres R., Posch W., Walder G. Cumulative SARS-CoV-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nat. Commun.* 2022;13(1):2560. DOI 10.1038/s41467-022-30163-4

Wirth W., Duchene S. GISAIDR: programmatically interact with the GISAID databases. *Zenodo*. 2022. DOI 10.5281/zenodo.6474693

Wu F., Zhao S., Yu B., Chen Y.M., Wang W., Song Z.G., Hu Y., Tao Z.W., Tian J.H., Pei Y.Y., Yuan M.L., Zhang Y.L., Dai F.H., Liu Y., Wang Q.M., Zheng J.J., Xu L., Holmes E.C., Zhang Y.Z. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. DOI 10.1038/s41586-020-2008-3

Xu H., Zhang Y., Yuan M., Ma L., Liu M., Gan H., Liu W., Lum G.G.A., Tao F. Basic reproduction number of the 2019 novel coronavirus disease in the major endemic areas of China: a latent profile analysis. *Front. Public Health*. 2021;9:575315. DOI 10.3389/fpubh.2021.575315

Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., Si H.R., Zhu Y., Li B., Huang C.L., Chen H.D., Chen J., Luo Y., Guo H., Jiang R.D., Liu M.Q., Chen Y., Shen X.R., Wang X., Zheng X.S., Zhao K., Chen Q.J., Deng F., Liu L.L., Yan B., Zhan F.X., Wang Y.Y., Xiao G.F., Shi Z.L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. DOI 10.1038/s41586-020-2012-7

ORCID ID

A.Yu. Palyanov orcid.org/0000-0003-1108-1486
N.V. Palyanova orcid.org/0000-0002-1783-5798

Финансирование. Исследование выполнено за счет гранта Российского научного фонда, проект № 23-64-00005.

Благодарности. Мы выражаем признательность всем, кто получил и сделал общедоступными генетические последовательности вариантов SARS-CoV-2, которые были задействованы при проведении некоторых расчетов в ходе нашего исследования, – авторам, лабораториям, ответственным за получение образцов, лабораториям, осуществившим секвенирование и ввод метаданных, а также размещение этой информации в базах данных GISAID и Genbank. Мы также благодарны авторам проекта Nextclade, предоставляющего онлайн-инструменты для анализа и визуализации генетических последовательностей различных вирусов.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 16.07.2023. После доработки 14.09.2023. Принята к публикации 18.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):851-858
DOI 10.18699/VJGB-23-98

Перевод на английский язык https://vavilov.elpub.ru/jour

Сверточные нейронные сети для классификации по данным ЭЭГ здоровых людей, практикующих или не практикующих медитацию

С. Фу[1], С.С. Таможников[2], А.Е. Сапрыгин[2, 3], Н.А. Истомина[1], Д.И. Клемешова[3], А.Н. Савостьянов[1, 2, 3]

[1] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[2] Научно-исследовательский институт нейронаук и медицины, Новосибирск, Россия
[3] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
a-sav@mail.ru

Аннотация. В настоящее время разработка объективных методик для оценки уровня стресса является чрезвычайно актуальной задачей прикладной нейронауки. Анализ электроэнцефалограммы (ЭЭГ), записанной в условиях выполнения заданий на самоконтроль поведения, может служить основой для разработки тестовых методик, позволяющих классифицировать людей по уровню стресса. Хорошо известно, что одним из следствий медитационной практики является выработка у участников навыков произвольного контроля над собственным ментальным состоянием за счет повышенной концентрации внимания на самом себе. На фоне медитационной практики часто происходит снижение общего уровня тревожности и стресса. Целью нашего исследования было разработать, обучить и протестировать сверточную нейронную сеть, способную классифицировать людей на группы участвующих или не участвующих в медитационной практике на основе анализа вызванных потенциалов головного мозга, записанных при выполнении заданий парадигмы стоп-сигнал. Были разработаны четыре архитектуры неглубоких сверточных сетей, которые были обучены и протестированы на выборке из 100 человек (51 медитатор и 49 не-медитатор). В дальнейшем все структуры были дополнительно протестированы на независимой выборке в 25 человек. Установлено, что структура, использующая одномерный сверточный слой, который объединяет слой и двуслойную полностью подключенную сеть, показала наилучшие результаты работы в имитационных тестах. Однако эта модель была часто подвержена переобучению из-за ограничения размера отображения набора данных. Явление переобучения было смягчено при помощи изменения структуры и масштаба модели, параметров сети инициализации, регуляризации, случайной деактивации (dropout) и гиперпараметров скрининга перекрестной проверки. В итоге нами получена модель, которая показала 82 % точность в классификации людей на подгруппы. Можно ожидать, что использование таких моделей окажется эффективным методом для оценки уровня стресса и предрасположенности к тревожным и депрессивным расстройствам в других группах испытуемых.
Ключевые слова: сверточные нейронные сети; ЭЭГ; вызванные потенциалы мозга; медитация; парадигма стоп-сигнал.

Для цитирования: Фу С., Таможников С.С., Сапрыгин А.Е., Истомина Н.А., Клемешова Д.И., Савостьянов А.Н. Сверточные нейронные сети для классификации по данным ЭЭГ здоровых людей, практикующих или не практикующих медитацию. *Вавиловский журнал генетики и селекции*. 2023;27(7):851-858. DOI 10.18699/VJGB-23-98

Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data

X. Fu[1], S.S. Tamozhnikov[2], A.E. Saprygin[2, 3], N.A. Istomina[1], D.I. Klemeshova[3], A.N. Savostyanov[1, 2, 3]

[1] Novosibirsk State University, Novosibirsk, Russia
[2] Scientific Research Institute of Neurosciences and Medicine, Novosibirsk, Russia
[3] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
a-sav@mail.ru

Abstract. The development of objective methods for assessing stress levels is an important task of applied neuroscience. Analysis of EEG recorded as part of a behavioral self-control program can serve as the basis for the development of test methods that allow classifying people by stress level. It is well known that participation in meditation practices leads to the development of skills of voluntary self-control over the individual's mental state due to an increased concentration of attention to themselves. As a consequence of meditation practices, participants can reduce overall anxiety and stress levels. The aim of our study was to develop, train and test a convolutional neural network capable of classifying individuals into groups of practitioners and non-practitioners of meditation by analysis of event-related brain potentials recorded during stop-signal paradigm. Four non-deep convolutional network architectures were developed, trained and tested on samples of 100 people (51 meditators and 49 non-meditators). Subsequently,

all structures were additionally tested on an independent sample of 25 people. It was found that a structure using a one-dimensional convolutional layer combining the layer and a two-layer fully connected network showed the best performance in simulation tests. However, this model was often subject to overfitting due to the limitation of the display size of the data set. The phenomenon of overfitting was mitigated by changing the structure and scale of the model, initialization network parameters, regularization, random deactivation (dropout) and hyperparameters of cross-validation screening. The resulting model showed 82 % accuracy in classifying people into subgroups. The use of such models can be expected to be effective in assessing stress levels and inclination to anxiety and depression disorders in other groups of subjects.
Key words: convolutional neural networks; EEG; event-related brain potentials; meditation; stop-signal paradigm.

For citation: Fu X., Tamozhnikov S.S., Saprygin A.E., Istomina N.A., Klemeshova D.I., Savostyanov A.N. Convolutional neural networks for classifying healthy individuals practicing or not practicing meditation according to the EEG data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):851-858. DOI 10.18699/VJGB-23-98

Введение

Стресс – одна из наиболее распространенных проблем в современном обществе; поиск эффективных способов оценки уровня стресса важен для своевременной диагностики риска возникновения психических и психосоматических расстройств (Kuh et al., 2003; Кузнецова и др., 2016). Большинство психологических методов оценки уровня стресса основано на использовании опросников, при заполнении которых респондент отвечает на вопросы, касающиеся его субъективного самочувствия. Слабым звеном такого подхода считается высокая вероятность неверных самооценок, возникающая либо вследствие нежелания человека сообщать о своих проблемах, либо как результат низкой способности распознать изменения в собственном состоянии (Iwata, Higuchi, 2000; McCrae et al., 2000). Возможным решением этой проблемы является разработка объективных подходов к диагностике психических черт или состояний, основанных на анализе мозговых сигналов, таких как фМРТ или ЭЭГ.

Медитация – система особых ментальных практик, направленных на установление произвольного контроля над своими собственными психическими состояниями. Хотя медитация изначально возникает как элемент религиозных практик, особенно распространенных в ориентальных религиях, в настоящее время этот феномен вызывает большой интерес среди ученых. Медитация рассматривается как основа для создания неинвазивных, немедикаментозных техник, позволяющих снизить риск широкого ряда психических или психосоматических заболеваний. В ряде исследований показано, что медитация оказывает множество положительных эффектов на психическое здоровье, включая общее снижение стресса и уровня предрасположенности к депрессии (Chiesa et al., 2011; Saeed et al., 2019). При анализе ЭЭГ, регистрируемой в условиях распознавания эмоциональных стимулов, были выявлены достоверные эффекты медитации на состояние головного мозга человека (Aftanas, Golosheykin, 2005; Atchley et al., 2016; Savostyanov et al., 2020). Поэтому анализ ЭЭГ у людей, занимающихся либо не занимающихся медитационной практикой, может быть рассмотрен как подход, позволяющий разработать методики оценки уровня стресса.

Экспериментальным методом, позволяющим оценить способность человека произвольно управлять собственными движениями в условиях изменяющейся внешней среды, является стоп-сигнал парадигма (ССП) (Logan, Cowan, 1984; Band et al., 2003). ССП позволяет оценить баланс двух процессов – активации и торможения поведения в условиях дефицита времени для принятия решений. В некоторых исследования ССП рассматривается как эффективный метод для диагностики уровня личностной тревожности и предрасположенности к депрессии (Hsieh et al., 2021; Зеленских и др., 2022). Можно предположить, что динамика мозговой активности в условиях ССП будет служить маркером, отличающим друг от друга людей, участвующих либо не участвующих в медитационной практике.

Искусственная нейронная сеть – это развивающаяся технология, основанная на машинном обучении, которая широко используется в различных областях. По сравнению с другими традиционными методами машинной классификации, такими как линейный дискриминантный анализ и алгоритм k-ближайшего соседа, искусственные нейронные сети дают более точные результаты классификации людей в соответствии с их поведенческими и нейрофизиологическими характеристиками (Khosla et al., 2020). Кроме того, по сравнению с машиной опорных векторов искусственная нейронная сеть легче справляется с задачами множественной классификации, обеспечивая удобство для дальнейших исследований, а также более эффективную подгонку нелинейных сложных взаимосвязей.

Целью нашего исследования были разработка, обучение и тестирование искусственной нейронной сети, позволяющей на основе анализа вызванных потенциалов в парадигме стоп-сигнал классифицировать людей по критерию их участия в медитативных практиках. Мы предполагаем, что созданная таким образом нейронная сеть будет в дальнейшем способна оценить индивидуальный уровень стресса и склонности к тревожно-депрессивным расстройствам.

Методы экспериментального исследования

Испытуемые. Группа людей, практикующих медитацию саматха (другое название «медитация осознанности»), была обследована в июле–августе 2018 г. на базе Байкальского медитационного центра (www.geshe.ru/). Экспериментальная группа включала 51 здорового праворукого участника от 25 до 66 лет (32 мужчины; средний возраст 41.0, SD = 8.3), практикующего медитацию в течение 5–15 лет. Контрольная группа была обследована в октябре–ноябре 2019 г. на базе медицинского колледжа пос. Хандыга, Томпонский район Республики Саха (Якутия), и включала 49 здоровых праворуких участников от 22

до 58 лет (22 мужчины; средний возраст 38.0, SD = 8.3), никогда не участвовавших в практиках медитации или йоги.

Протокол исследования одобрен локальным этическим комитетом НИИ нейронаук и медицины в соответствии с Хельсинкской декларацией биомедицинских обследований. Все испытуемые подписывали добровольное согласие на участие в обследованиях.

Экспериментальная процедура. Эксперимент был организован на основе парадигмы стоп-сигнал, предложенной в 1984 г. (Logan, Cowan, 1984) и модифицированной А.Н. Савостьяновым с коллегами (Savostyanov et al., 2009). Эксперимент проходил в форме компьютерной интерактивной игры «Охота». На экране компьютера появлялась одна из двух картинок: олень или танк. Испытуемый должен был нажимать левую кнопку после появления оленя или правую кнопку после появления танка. Время нажатия было ограничено 0.7 с. Если испытуемый нажимал на кнопку правильно и быстрее, чем 0.7 с, его игровой счет увеличивался. Если испытуемый нажимал кнопки неверно или время его ответа было дольше, чем 0.7 с, то игровой счет снижался.

Всего каждому испытуемому было предложено 135 заданий. В 35 случаях после появления целевого сигнала предъявлялся стоп-сигнал (красный квадрат с надписью “Stop”), что означало, что участник должен прервать уже начатое движение. Если участник не нажимал на кнопку после стоп-сигнала, его счет не менялся. Если участник нажимал на кнопку после стоп-сигнала, его счет снижался. Порядок активационных и тормозных заданий был рандомизирован. Также была рандомизирована последовательность заданий «олень» и «танк». Интервал между окончанием предыдущего задания и началом нового варьировал от 3 до 7 с. Общая продолжительность эксперимента составляла примерно 12 мин.

Предобработка экспериментальных данных. Очистка ЭЭГ от артефактов производилась методом ICA (Delorme, Makeig, 2004). Исходный ЭЭГ сигнал был эпохирован относительно метки появления целевого сигнала (олень или танк). Временной интервал от –1 до +3 с был выбран для эпохирования стимулов. Базовый уровень ЭЭГ был установлен в интервале от –1000 до –250 мс. Таким образом, получено от 80 до 90 фрагментов ЭЭГ для каждого участника, содержащих только активационное условие и не содержащих моментов предъявления стоп-сигналов. После исключения артефактов связанные с событиями потенциалы (event-related potential, ERP) вычислялись отдельно для каждого канала ЭЭГ, усредненно по всем испытаниям и всем участникам.

Вычисление ERP проводилось в программе EEGLAB_ toolbox. Для каждого канала ЭЭГ были получены амплитудно-временные графики ERP. Затем выполнен визуальный просмотр графика ERP для отведения C3. В этом отведении максимально четко выделяются моторные пики ERP. В частности, по этому отведению были выбраны два пика – ранний премоторный, амплитуда которого предшествует нажатию на кнопку (так называемый readness potential), и постмоторный пик, амплитуда которого достигает максимума при нажатии на кнопку. В результате визуального просмотра были установлены временные границы как раннего, так и позднего пика, после чего амплитуда в каждом из этих временных окон вычислялась отдельно для каждого человека и каждого ЭЭГ канала, но усредненно по всем испытаниям активационного задания парадигмы стоп-сигнал у каждого участника. Вычисление усредненной амплитуды выполнено при помощи программного пакета ERPLAB (https://erpinfo.org/erplab). Значения амплитуды были откорректированы к базовому уровню отдельно для каждого участника. Полученные значения использовались в качестве обучающих и тестовых данных для искусственных нейронных сетей.

Описание входных данных. Общая структура входных данных показана на рис. 1. У каждого испытуемого анализировалась ЭЭГ для 64 каналов, расположенных на разных участках поверхности головы. В соответствии с международной схемой 10–20 %, название электрода отражает его пространственное положение. Исходный ЭЭГ сигнал для каждого канала представлен в форме непрерывного ряда замеров разности потенциалов между активным и референтным электродами с временным разрешением 1000 замеров в секунду. Каждый участник обследования выполняет серию однотипных заданий (в нашем случае 100 активационных заданий стоп-сигнал парадигмы для каждого участника).

При вычислении амплитуды ERP исследователь выбирает несколько временных окон, в каждом из которых все значения амплитуды суммируются по всем временным

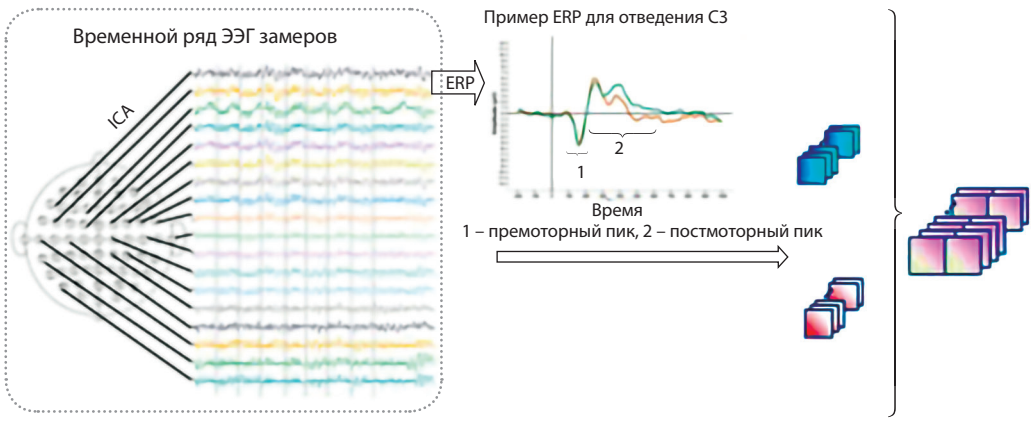


Рис. 1. Схема получения входных данных для нейронной сети.

точкам и усредняются по всем испытаниям. Амплитудные значения в разных окнах отражают временную динамику развития нейрофизиологического процесса. Мы выбрали два временных окна (250–350 и 550–900 мс после целевого сигнала), которые отражали соответственно физиологические процессы, ассоциированные с подготовкой и выполнением движения. У каждого участника было получено численное значение амплитуды ERP отдельно для каждого временного окна и для каждого ЭЭГ канала. В разных участках головы ERP может отклоняться от нулевого значения потенциала как вверх (положительный пик), так и вниз (отрицательный пик), поэтому численные значения амплитуды могут быть как положительными, так и отрицательными. Таким образом, наши данные учитывают как пространственную (название канала, его положение на голове), так и временну́ю (первое или второе окно ERP) характеристику мозгового ответа на задание в парадигме стоп-сигнал, а также электрическую направленность реакции (положительные или отрицательные значения амплитуды пиков).

Для каждого обследованного человека размерность данных составила 2×64 значения. Поскольку в каждой группе было примерно по 50 участников, размер данных для каждой из наших выборок составляет 50×2×64, а общий размер набора данных – 100×2×64.

Проектировка структуры и фреймворка нейронной сети

Поскольку входной набор ERP данных невелик, была спроектирована неглубокая нейронная сеть для предсказания того, участвовал ли человек в долговременных медитациях или нет. Однако исходная ЭЭГ также имеет характеристики временных рядов, поэтому для ее анализа была дополнительно использована сверточная нейронная сеть в качестве глубокой нейронной сети для обучения и прогнозирования. Основные компоненты сверточной нейронной сети включают сверточные слои, объединяющие слои и полностью связанные слои.

В нашем случае на входной слой сверточной сети подаются данные ЭЭГ, преобразованные в виде двумерной матрицы с единичным размером выборки 2×64, где каждая строка представляет отдельный пик ERP, а каждый столбец представляет канал записи ЭЭГ. Скрытый уровень сверточной нейронной сети включает в себя три общие архитектуры: сверточный уровень, объединяющий слой и полностью подключенный слой. В качестве ядра свертки мы применили инструмент Conv1d() в PyTorch, что позволило предотвратить явление переобучения, вызванного использованием более сложных ядер свертки с бóльшим количеством параметров (детальное описание инструмента см.: https://pytorch.org/docs/stable/generated/torch.nn.Conv1d.html#torch.nn.Conv1d, 21.02.2023).

Параметры сверточного слоя включают размер ядра свертки, размер шага и заполнение, которые совместно определяют размер выходной карты объектов сверточного слоя и являются гиперпараметрами сверточной нейронной сети. Из-за особенностей данных ЭЭГ существуют как пространственные, так и временные взаимосвязи, поэтому мы разработали две схемы. Первая схема заключается в использовании в общей сложности двух одномерных сверток. Одна из них извлекает пространственные характеристики, которые представляют собой соединения пиков ERP в различных каналах электродов, а другая извлекает временные характеристики. В этой схеме функция-оболочка PyTorch Conv1d() была использована для завершения соответствующей функции. Второй способ заключается в применении только одной одномерной свертки, но эта свертка может извлекать как временные, так и пространственные объекты, для чего также выбрана функция-оболочка PyTorch Conv1d().

Сверточные слои содержат функции активации, помогающие представлять сложные объекты. В нашем исследовании применялись три функции активации: sigmoid(), relu() и softmax() из PyTorch (https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html, 15.04.2023). После извлечения объектов в сверточном слое выходная карта объектов передавалась на объединяющий слой для выбора объектов и фильтрации информации. Слой объединения выбирает область объединения таким же образом, как и этап карты объектов сканирования ядра свертки, который управляется размером объединения, размером шага и заполнением. Уровень свертки и объединяющий слой в сверточной нейронной сети могут извлекать признаки входных данных. Роль полностью связанного слоя заключается в нелинейной комбинации извлеченных признаков для получения выходных данных. В нашем случае было создано два полностью связанных слоя, чтобы предотвратить переобучение из-за небольшого размера набора данных, для чего применялся инструмент Linear() в PyTorch. Перед выходным слоем в сверточной нейронной сети обычно находится полностью подключенный слой. Мы использовали различные функции потери и активации при обучении на основе этих двух сценариев, чтобы повысить точность и производительность модели двух решений.

В соответствии с описанной выше схемой были спроектированы четыре сетевые структуры, которые использовались для классификации обследованных людей (рис. 2). Единственное различие между этими четырьмя архитектурами заключается в количестве сверточных слоев и количестве выходных нейронов в конце.

В *первой структуре* сверточный слой применяется для извлечения как временных, так и пространственных объектов. Затем берутся два полностью соединенных слоя и выводятся два значения после нормализации с помощью функции активации softmax. В качестве функции потерь используется перекрестная энтропия, в качестве алгоритма градиентного спуска – метод Adam.

Вторая структура применяет сверточный слой для извлечения как временных, так и пространственных объектов. Затем берутся два полностью соединенных слоя, а значение выводится после активации сигмовидной функции. В качестве функции потерь используется двоичная перекрестная энтропия, в качестве алгоритма градиентного спуска – Adam.

Третья структура основана на применении двух видов сверток для извлечения пространственных и временных характеристик данных соответственно. Затем берутся два полностью соединенных слоя и выводятся два значения после нормализации с помощью функции активации

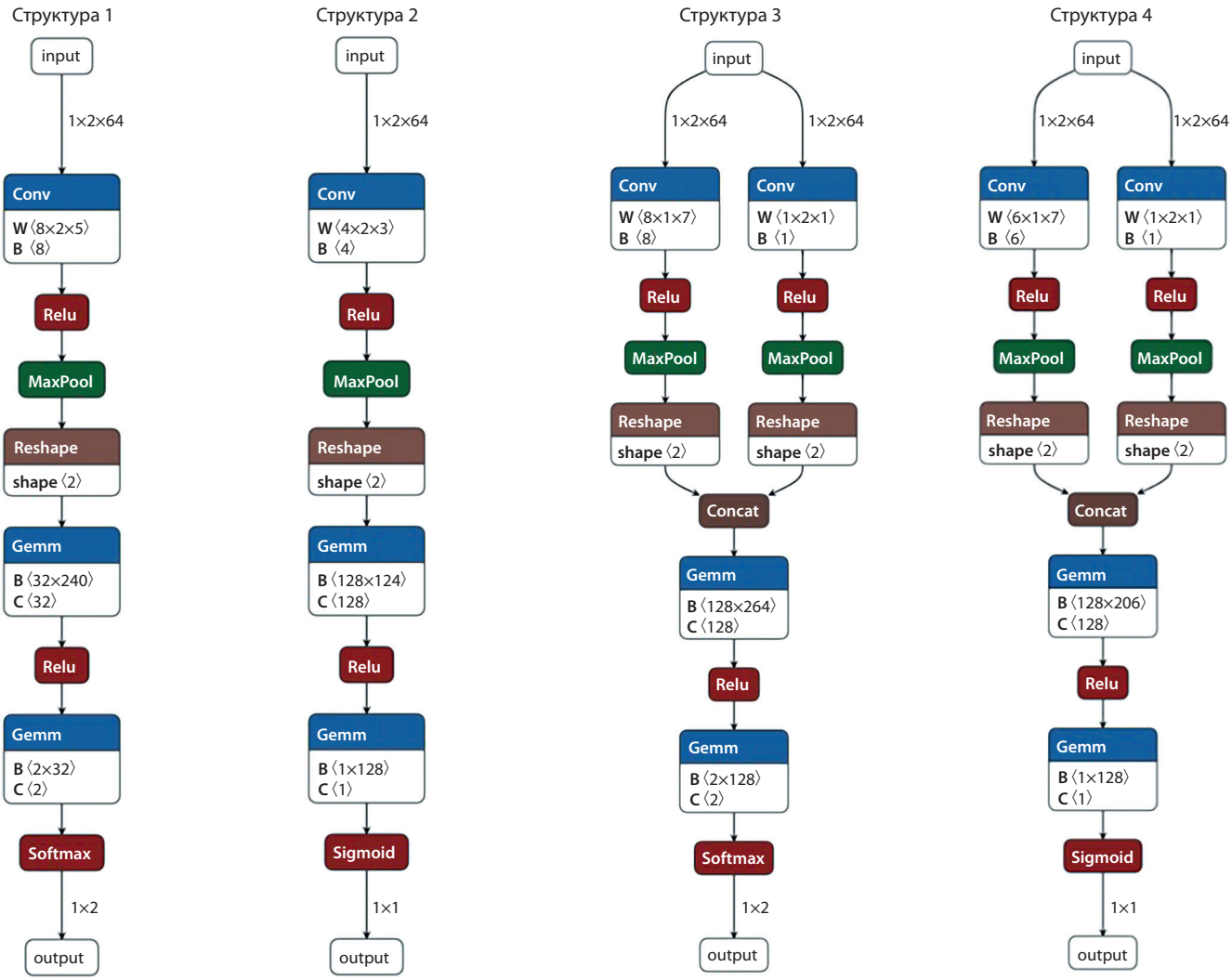


Рис. 2. Блок-схемы четырех моделей (структур) для архитектуры нейронной сети.

softmax. В качестве функции потерь используется перекрестная энтропия, в качестве алгоритма градиентного спуска – Adam.

И наконец, *четвертая структура* применяет два вида сверток для извлечения пространственных и временных характеристик данных соответственно. Затем берутся два полностью соединенных слоя. Значение выводится после активации сигмовидной функции. В качестве функции потерь используется двоичная перекрестная энтропия, в качестве алгоритма градиентного спуска – Adam.

Оптимальные гиперпараметры найдены для каждой структуры и описаны в разделе оценки модели.

Обучение нейронной сети

Процесс обучения искусственной нейронной сети можно разделить на следующие четыре этапа: инициализация, прямое распространение, обратное распространение и обновление веса.

При инициализации мы присвоили случайное начальное значение каждому из параметров (весов и смещений) нейронной сети для нарушения симметрии, чтобы каждый нейрон имел разный градиент и, таким образом, мог изучать различные функции. Позже, при поиске по гиперпараметрам, для каждой архитектуры была определена оптимальная функция инициализации. В процессе прямого распространения обучающие данные (входные и выходные) подавались в нейронную сеть, и значение активации каждого нейрона вычислялось по очереди от входного слоя к скрытому, а затем к выходному слою в соответствии со структурой нейронной сети. Значения активации брались из линейной комбинации входных данных и весов плюс смещение, за которым следовала нелинейная функция, такая как sigmoid или ReLU. Цель прямого распространения состояла в том, чтобы получить прогнозируемый результат нейронной сети и сравнить его с истинным результатом. Целью обратного распространения являлось получение градиента каждого параметра, который может быть применен для обновления параметров. В нашем случае для этого были использованы функция потерь кросс-энтропии и двоичная функция потерь кросс-энтропии (https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html, 20.03.2023). Функция кросс-энтропии применялась нами для измерения расстояния между распределением вероятностей, предсказанным мо-

делью, и истинным распределением вероятностей. С ее помощью мы оценивали производительность модели и выбирали оптимальную модель и параметр путем сравнения значений потерь различных моделей или различных параметров.

В соответствии со своим градиентом каждый параметр обновляется с определенной скоростью обучения (размером шага), так что функция потерь уменьшается. Целью обновления веса является оптимизация параметров нейронной сети, чтобы нейронная сеть могла лучше соответствовать данным обучения. Для этой задачи нами был применен метод Adam. Метод Adam – это алгоритм оптимизации стохастического градиентного спуска с адаптивным импульсом, который был предложен на конференции ICLR в 2015 г. и стал одним из самых популярных и эффективных оптимизаторов в области глубокого обучения. Adam объединяет два классических алгоритма оптимизации: Adagrad и RMSProp, которые способны решать задачи с разреженными градиентами и нестационарными целевыми функциями, а также использовать идею импульса для ускорения сходимости. Adam эквивалентен наличию отдельной скорости обучения для каждого параметра, и эта скорость обучения адаптивно настраивается в соответствии с изменением градиента. В частности, когда градиент велик, оценка второго момента увеличивается, что снижает скорость обучения. Когда градиент мал или разрежен, оценка первого момента увеличивается, что увеличивает скорость обучения. Это позволило эффективно избежать колебаний, вызванных слишком большой скоростью обучения, или увеличения сложности конвергенции, вызванного слишком малой скоростью обучения, или даже попадания в ловушку локального минимума или седловой точки.

Чтобы уменьшить переобучение и лучше обучить модель, мы применили пакетную нормализацию. Batch normalization – это подход, который решает проблему исчезающего градиента за счет улучшения сглаживания потерь, ускоряет конвергенцию сети и повышает точность (Loffe, Szegedy, 2015). Метод нормализует данные в мини-пакете таким образом, чтобы среднее значение было равно 0, а стандартное отклонение – 1. В то же время вводятся два обучаемых параметра, масштаб и сдвиг, чтобы модель могла изучить свое соответствующее распределение при обратном распространении. Для реализации этой функции мы использовали инструмент BatchNorm1d() из PyTorch.

Переобучение – распространенная проблема в процессе обучения искусственной нейронной сети, при возникновении которой модель хорошо работает на обучающем наборе, но плохо – на тестовом наборе или на новых данных, что указывает на то, что она плохо обобщается. В нашем случае проблема заключалась в переобучении из-за небольшого набора исходных данных. Чтобы решить ее, мы применили инициализацию, регуляризацию L2 и случайную деактивацию (dropout), а также перекрестную проверку для оценки модели и выбора гиперпараметров, которые наилучшим образом обучают модель, в некоторой степени уменьшая переобучение. Мы использовали метод регуляризации L2 (уменьшение веса), который включает добавление штрафного члена к функции потерь, пропорциональной сумме квадратов параметров модели. Регуляризация L2 может привести к тому, что параметры модели будут стремиться к меньшим значениям, тем самым снижая чувствительность модели к шуму или выбросам. Случайная деактивация (dropout) означает случайное обнуление определенных нейронов или слоев соединений с определенной вероятностью во время обучения. Это уменьшает количество параметров модели, что повышает надежность и способность к обобщению модели.

Перекрестная проверка – это повторное использование данных, разделение результирующей выборки данных, объединение в различные обучающие и тестовые наборы, обучающий набор для обучения модели и тестовый набор для оценки качества прогнозирования модели. Мы применили метод K-fold умножения в качестве метода перекрестной проверки, чтобы уменьшить переобучение.

Оценка качества работы модели на обучающих данных

В соответствии с характеристиками выборки ЭЭГ данных и показателями эталонной модели классификации мы использовали метрики “F1-score”, “AUC” (area under curve) и «точность» в качестве показателей оценки модели (https://keras.io/api/models/model_training_apis). Чем выше эти показатели, тем выше производительность модели. F1-score и AUC являются комплексными оценочными показателями классификационных моделей, но они имеют разные погрешности. На AUC в меньшей степени влияет соотношение положительных и отрицательных проб в образце. Для целей этой разработки стало ясно, что прогнозирование человека с высоким уровнем стресса как человека с низким уровнем стресса привело бы к принципиально неверным результатам. Поэтому мы выбрали показатель F1-score в качестве наиболее приоритетного для оценки эффективности модели. Гиперпараметры модели оценивали, используя пятикратную перекрестную проверку, чтобы отобрать наиболее подходящие гиперпараметры для предотвращения переобучения и повышения производительности модели.

Результаты оценки модели на обучающей выборке представлены на рис. 3. Рассматривая каждый из выбранных нами показателей, можно увидеть, что наиболее эффективную классификацию показала модель 2. По всем выбранным показателям ее эффективность превысила 80 %. Модели 1 и 4 тоже демонстрируют хорошие результаты классификации, тогда как модель 3 работает хуже всего. Следовательно, мы предполагаем, что выход одного нейрона превосходит использование двух нейронов в задаче

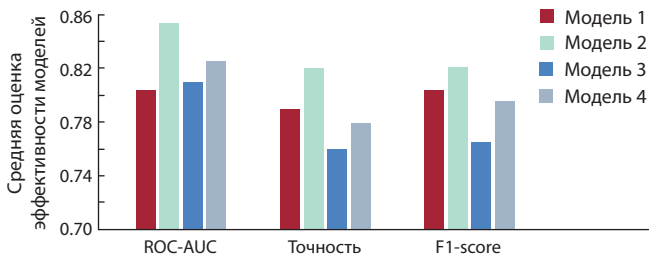


Рис. 3. Результаты тестирования четырех разных моделей нейронной сети на обучающей выборке.

бинарной классификации ЭЭГ. Двоичная перекрестная потеря, очевидно, больше подходит для нашей задачи классификации, основанной на имеющемся наборе данных. При оценке эффективности модели количество выборок составило 100, из них 51 человек занимался медитацией (низкий уровень стресса) и 49 не занимались медитацией. Количество выборок сбалансировано, поэтому это несущественно влияет на обучение и производительность модели. Более того, для данных только с двумя пиками ERP в 64 каналах электродов одна свертка, извлекающая как временные, так и пространственные характеристики, работала лучше, чем две свертки, извлекающие временные и пространственные характеристики по отдельности.

Оценка качества модели на независимых данных. Чтобы оценить качество работы модели на независимой выборке, мы подготовили ЭЭГ данные, записанные у 25 человек, которые не были включены в обучающую выборку. Из этих 25 человек 12 практиковали медитацию, а 13 не практиковали. Оборудование, дизайн эксперимента и предобработка ЭЭГ данных были такими же, как и в случае с обучающей выборкой. В этой части исследования все модели, обученные ранее, были протестированы на новых данных, не включенных в обучающую выборку. В качестве показателей оценки эффективности модели были использованы точность, надежность, отзывчивость, F1-score, ROC-AUC, специфичность, чувствительность. Несмотря на использование функции инициализации параметров, веса по-прежнему инициализировались случайным образом в пределах определенного диапазона, поэтому мы исправили начальное значение случайного числа, чтобы обеспечить стабильность работы модели.

Показатели эффективности для разных моделей на независимой тестовой выборке отражены на рис. 4. Наилучшие результаты по большинству выбранных параметров показала модель 4. Также достаточно хорошие результаты получены для модели 2. Эта структура показала наименьшую чувствительность к переобучению, что говорит о ее большей надежности в сравнении с моделью 4.

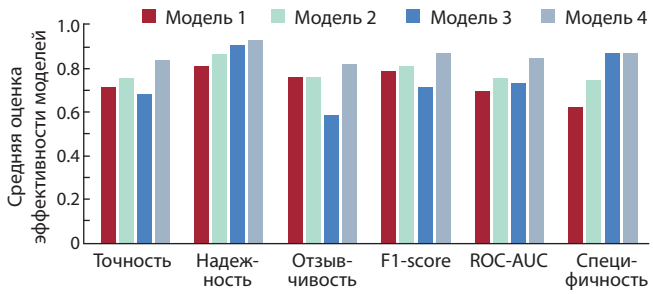


Рис. 4. Результаты тестирования четырех разных моделей нейронной сети на независимой выборке.

Заключение

В нашем исследовании была успешно разработана нейронная сеть, которая классифицирует людей на группы участвующих или не участвующих в медитации на основе анализа их ЭЭГ данных с точностью примерно 80–85 %. Мы использовали набор данных ЭЭГ, собранный и составленный в ходе наших собственных экспериментов, и выбрали амплитуду ERP пика перед нажатием на кнопку 250–350 мс и значение амплитуды пика после нажатия на кнопку 550–900 мс для 64 каналов записи. При этом размер выборки составил 1×2×64.

Были разработаны четыре архитектуры неглубоких сверточных сетей, среди которых структуры 2 и 4 показали себя лучше всего в тестах на независимых выборках данных. Наиболее надежной была структура 2, которая использовала одномерный сверточный слой, объединяющий слой и двуслойную полностью подключенную сеть. Во время разработки этой модели отмечено, что она часто подвержена переобучению из-за ограничения размера отображения набора данных. Явление переобучения было смягчено за счет изменения структуры и масштаба модели, конкретных параметров сети инициализации, регуляризации, случайной деактивации (dropout) и гиперпараметров скрининга перекрестной проверки.

В целом предложенный нами подход был апробирован на двух небольших выборках клинических испытуемых. Похожий метод на экспериментальных данных из парадигмы стоп-сигнал был ранее апробирован нами при классификации на выборках клинических пациентов с депрессивным расстройством и здоровых людей (Зеленских и др., 2022). Результаты нашей новой статьи дополняют предыдущую работу, так как показывают, что, несмотря на небольшие размеры выборок, метод сверточных нейронных сетей позволяет достигать высокого уровня точности в классификации разных, независимых друг от друга групп людей, различающихся по уровню стресса. Взятые вместе, результаты обоих исследований показывают, что применение нейронных сетей к данным, полученным при тестировании людей в рамках парадигмы стоп-сигнал, является перспективным методом для оценки их уровня стресса и степени выраженности симптоматики тревожно-депрессивных расстройств. Необходимо отметить, что результаты М.О. Зеленских с коллегами основаны на применении только поведенческих данных, полученных в парадигме стоп-сигнал, тогда как наши базируются на анализе электрических ответов мозга, полученных в том же эксперименте. Продолжением исследований должно быть применение сверточных нейронных сетей для одновременного совместного анализа и поведенческих, и нейробиологических данных с целью более точной классификации участников по уровню стресса.

Важно отметить, что большинство стандартных методов оценки уровня стресса или предрасположенности к тревожно-депрессивным расстройствам основано на применении психологических опросников или на интервью с психиатром (например, Beck et al., 1988). Однако такие методы обладают недостатком: пациент может не хотеть информировать интервьюера о своем состоянии или неадекватно оценивать самого себя. Неадекватная самооценка пациента часто является причиной неверных выводов, касающихся его предрасположенности к болезни (Nock et al., 2010). Другой подход применяет анализ поведенческих или нейрофизиологических реакций на эмоциональные стимулы. В качестве таких стимулов предъявляют либо фотографии лиц, выражающих эмоциональные состояния самого пациента или других людей (Quevedo et al., 2016), либо эмоциональные сообщения (Bocharov et al., 2020).

Этот метод позволяет объективно оценить степень нарушения аффективных функций головного мозга, но малочувствителен к изменениям общей способности человека к самоконтролю поведения. Предлагаемый нами метод, наоборот, использует неэмоциональные стимулы для индуцирования сложной сенсомоторной реакции, требующей либо активации, либо торможения движения. Наш подход позволяет оценивать общий уровень самоконтроля поведения, но не дает возможности оценить аффективное состояние пациента. Очевидно, что эти три подхода (тестирование при помощи опросников, анализ реакций на аффективное стимулирование и анализ реакций в задачах на моторный контроль) являются взаимодополняющими, т. е. должны применяться вместе для более детальной оценки одного и того же пациента. Хотя предлагаемый нами подход нуждается в дополнительной проверке, в будущем он может дать большие результаты при разработке средств диагностики индуцированных стрессом заболеваний.

Список литературы / References

Зеленских М.О., Сапрыгин А.Е., Таможников С.С., Рудыч П.Д., Лебедкин Д.А., Савостьянов А.Н. Разработка нейронной сети для диагностики риска возникновения депрессии по экспериментальным данным стоп-сигнал парадигмы. *Вавиловский журнал генетики и селекции*. 2022;26(8):773-779. DOI 10.18699/VJGB-22-93

[Zelenskih M.O., Saprygin A.E., Tamozhnikov S.S., Rudych P.D., Lebedkin D.A., Savostyanov A.N. Development of a neural network for diagnosing the risk of depression according to the experimental data of the stop signal paradigm. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2022;26(8):773-779. DOI 10.18699/VJGB-22-93]

Кузнецова В.Б., Князев Г.Г., Дорошева Е.А., Бочаров А.В., Савостьянов А.Н. Роль личности и стресса в развитии депрессивных расстройств у студентов. *Журн. неврологии и психиатрии.* 2016;116(12):114-118. DOI 10.17116/jnevro2016116121114-118

[Kuznetsova V.B., Knyazev G.G., Dorosheva E.A., Bocharov A.V., Savostyanov A.N. A role of personality and stress in the development of depressive symptoms in students. *Zhurnal Nevrologii i Psikhiatrii = Journal of Neurology and Psychiatry*. 2016;116(12):114-118. DOI 10.17116/jnevro2016116121114-118 (in Russian)]

Aftanas L., Golosheykin S. Impact of regular meditation practice on EEG activity at rest and during evoked negative emotions. *Int. J. Neurosci*. 2005;115(6):893-909. DOI 10.1080/00207450590897969

Atchley R., Klee D., Memmott T., Goodrich E., Wahbeh H., Oken B. Event-related potentials correlates of mindfulness meditation competence. *Neuroscience*. 2016;320:83-92. DOI 10.1016/j.neuroscience.2016.01.051

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

Beck A.T., Steer R.A., Garbin M.G. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clin. Psychol. Rev.* 1988;8(1):77-100. DOI 10.1016/0272-7358(88)90050-5

Bocharov A.V., Savostyanov A.N., Tamozhnikov S.S., Merkulova E.A., Saprigyn A.E., Proshina E.A., Knyazev G.G. Oscillatory dynamics of perception of emotional sentences in healthy subjects with different severity of depressive symptoms. *Neurosci. Lett*. 2020;728:134888. DOI 10.1016/j.neulet.2020.134888

Chiesa A., Calati R., Serretti A. Does mindfulness training improve cognitive abilities? A systematic review of neuropsychological findings. *Clin. Psychol. Rev*. 2011;31(3):449-464. DOI 10.1016/j.cpr.2010.11.003

Delorme A., Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*. 2004;134(1):9-21. DOI 10.1016/j.jneumeth.2003.10.009.

Hsieh M.T., Lu H., Lin C.I., Sun T.H., Chen Y.R., Cheng C.H. Effects of trait anxiety on error processing and post-error adjustments: an event-related potential study with stop-signal task. *Front. Hum. Neurosci*. 2021;15:650838. DOI 10.3389/fnhum.2021.650838

Iwata N., Higuchi H.R. Responses of Japanese and American university students to the STAI items that assess the presence or absence of anxiety. *J. Pers. Assess*. 2000;74(1):48-62. DOI 10.1207/S15327752JPA740104

Khosla A., Khandnor P., Chand T. A comparative analysis of signal processing and classification methods for different applications based on EEG signals. *Biocybern. Biomed. Eng*. 2020;40(2):649-690. DOI 10.1016/j.bbe.2020.02.002

Kuh D., Ben-Shlomo Y., Lynch J., Hallqvist J., Power C. Life course epidemiology. *J. Epidemiol. Community Health.* 2003;57(10):778-783. DOI 10.1136/jech.57.10.778

Loffe S., Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv.* 2015;1502.03167. DOI 10.48550/arXiv.1502.03167

Logan G.D., Cowan W.B. On the ability to inhibit thought and action: a theory of an act of control. *Psychol. Rev*. 1984;91(3):295-327. DOI 10.1037/0033-295X.91.3.295

McCrae R.R., Costa P.T., Jr., Ostendorf F., Angleitner A., Hrebícková M., Avia M.D., Sanz J., Sánchez-Bernardos M.L., Kusdil M.E., Woodfield R., Saunders P.R., Smith P.B. Nature over nurture: temperament, personality, and life span development. *J. Pers. Soc. Psychol*. 2000;78(1):173-186. DOI 10.1037//0022-3514.78.1.173

Nock M.K., Park J.M., Finn C.T., Deliberto T.L., Dour H.J., Banaji M.R. Measuring the suicidal mind: implicit cognition predicts suicidal behavior. *Psychol. Sci*. 2010;21(4):511-517. DOI 10.1177/0956797610364762

Quevedo K., Scott R.N.H., Martin J., Smyda G., Keener M., Oppenheimer C.W. The neurobiology of self-face recognition in depressed adolescents with low or high suicidality. *J. Abnorm. Psychol*. 2016;125(8):1185-1200. DOI 10.1037/abn0000200

Saeed S.A., Cunningham K., Bloch R.M. Depression and anxiety disorders: benefits of exercise, yoga, and meditation. *Am. Fam. Physician*. 2019;99(10):620-627

Savostyanov A.N., Tsai A.C., Liou M., Levin A.E., 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

Savostyanov A.N., Tamozhnikov S.S., Bocharov A.V., Saprygin A.E., Matushkin Y., Lashin S., Kolpakova G., Sudobin K., Knyazev G. The effect of meditation on comprehension of statement about oneself and others: a pilot ERP and behavioral study. *Front. Hum. Neurosci*. 2020;13:437. DOI 10.3389/fnhum.2019.00437

ORCID ID

A.E. Saprygin orcid.org/0000-0001-6789-2953
A.N. Savostyanov orcid.org/0000-0002-3514-2901

Благодарности. Разработка и тестирование нейронной сети и коллекция ЭЭГ данных у медитаторов выполнялись в рамках бюджетного проекта ИЦиГ СО РАН FWNR-2022-0020. Сбор ЭЭГ данных у не-медитаторов, а также предпроцессинг всех ЭЭГ данных выполнен в рамках проекта Российского научного фонда № 22-15-00142 «фМРТ и ЭЭГ корреляты фокуса внимания на собственной персоне как фактора предрасположенности к аффективным расстройствам».

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 11.07.2023. После доработки 10.09.2023. Принята к публикации 13.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):859-868
DOI 10.18699/VJGB-23-99

Перевод на английский язык https://vavilov.elpub.ru/jour

Определение содержания меланина и антоцианов в зернах ячменя на основе анализа цифровых изображений методами машинного обучения

Е.Г. Комышев[1] ✉, М.А. Генаев[1, 2, 3], И.Д. Бусов[1, 3], М.В. Кожекин[2], Н.В. Артеменко[2, 3], А.Ю. Глаголева[1], В.С. Коваль[1], Д.А. Афонников[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ komyshev@bionet.nsc.ru

Аннотация. Пигментный состав оболочек семян растений влияет на такие важные их свойства, как устойчивость к действию патогенов, прорастание на корню, а также механическая прочность. У ячменя (*Hordeum vulgare* L.) темная окраска зерен может быть обусловлена синтезом и накоплением двух групп пигментов. Голубая и фиолетовая окраска зерна связана с синтезом антоцианов. Серую и черную окраску придают пигменты меланины. Данные пигменты могут накапливаться в оболочках зерна независимо либо совместно, поэтому визуально определить, накопление каких именно пигментов придает темный цвет зерна, затруднительно. Для точного определения наличия/отсутствия пигментов используются химические и генетические методы, которые дороги и трудоемки. Поэтому создание нового метода для быстрой оценки наличия определенных пигментов в зерновке является актуальной задачей, решение которой поможет при исследовании механизмов генетического контроля пигментного состава зерна. Настоящая работа посвящена разработке метода оценки пигментного состава зерен ячменя на основе анализа цифровых изображений с помощью алгоритмов компьютерного зрения и машинного обучения. Разработан протокол съемки для получения двумерных цифровых цветных изображений зерен. С использованием данного протокола получено 972 изображения для 108 образцов ячменя. Каждый образец мог содержать пигменты антоцианы и/или меланины. Для точного определения содержания пигментного состава образцов применялись химические методы. Для предсказания пигментного состава зерна на основе изображений было разработано четыре модели, основанных на методах компьютерного зрения и сверточных нейронных сетях различной архитектуры. Лучшую производительность на отложенной выборке показала модель сети U-Net, основанная на топологии EfficientNetB0 (значение параметра «точность» составило 0.821).
Ключевые слова: анализ цифровых изображений; машинное обучение; зерна ячменя; пигментный состав.

Для цитирования: Комышев Е.Г., Генаев М.А., Бусов И.Д., Кожекин М.В., Артеменко Н.В., Глаголева А.Ю., Коваль В.С., Афонников Д.А. Определение содержания меланина и антоцианов в зернах ячменя на основе анализа цифровых изображений методами машинного обучения. *Вавиловский журнал генетики и селекции*. 2023;27(7):859-868. DOI 10.18699/VJGB-23-99

Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods

E.G. Komyshev[1] ✉, M.A. Genaev[1, 2, 3], I.D. Busov[1, 3], M.V. Kozhekin[2], N.V. Artemenko[2, 3], A.Y. Glagoleva[1], V.S. Koval[1], D.A. Afonnikov[1, 2, 3]

[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
[3] Novosibirsk State University, Novosibirsk, Russia
✉ komyshev@bionet.nsc.ru

Abstract. The pigment composition of plant seed coat affects important properties such as resistance to pathogens, pre-harvest sprouting, and mechanical hardness. The dark color of barley (*Hordeum vulgare* L.) grain can be attributed to the synthesis and accumulation of two groups of pigments. Blue and purple grain color is associated with the biosynthesis of anthocyanins. Gray and black grain color is caused by melanin. These pigments may accumulate in the grain shells both individually and together. Therefore, it is difficult to visually distinguish which pigments are responsible for the dark color of the grain. Chemical methods are used to accurately determine the presence/absence of pigments; however, they are expensive and labor-intensive. Therefore, the development of a new method for quickly assessing the presence of pigments in the grain would help in investigating the mechanisms

of genetic control of the pigment composition of barley grains. In this work, we developed a method for assessing the presence or absence of anthocyanins and melanin in the barley grain shell based on digital image analysis using computer vision and machine learning algorithms. A protocol was developed to obtain digital RGB images of barley grains. Using this protocol, a total of 972 images were acquired for 108 barley accessions. Seed coat from these accessions may contain anthocyanins, melanins, or pigments of both types. Chemical methods were used to accurately determine the pigment content of the grains. Four models based on computer vision techniques and convolutional neural networks of different architectures were developed to predict grain pigment composition from images. The U-Net network model based on the EfficientNetB0 topology showed the best performance in the holdout set (the value of the "accuracy" parameter was 0.821).
Key words: digital image analysis; machine learning; barley grains; pigment composition.

For citation: Komyshev E.G., Genaev M.A., Busov I.D., Kozhekin M.V., Artemenko N.V., Glagoleva A.Y., Koval V.S., Afonnikov D.A. Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):859-868. DOI 10.18699/VJGB-23-99

Введение

Цвет оболочки зерен злаков – важный признак, характеризующий содержащиеся в ней пигменты и метаболиты. Наличие пигментов в оболочке влияет на различные технологические свойства зерна (Souza, Marcos-Filho, 2001; Flintham et al., 2002). Образцы с темной окраской зерновки являются более холодо- и засухоустойчивыми, а также обладают повышенной устойчивостью к действию патогенов (Ceccarelli et al., 1987; Choo et al., 2005). Такие свойства окрашенных растений связаны с высоким содержанием антиоксидантов, а также с дополнительной механической прочностью оболочек зерна (Ferdinando et al., 2012; Jana, Mukherjee, 2014). Темная окраска зерен ячменя может быть обусловлена синтезом и накоплением двух групп пигментов. Голубая и фиолетовая окраска зерна связана с синтезом антоцианов. Серую и черную окраску зернам ячменя придают пигменты меланины. Данные пигменты могут накапливаться в оболочках зерна растений в зависимости от генотипа как по отдельности, так и совместно. Поэтому визуально определить, накоплением каких именно пигментов вызван темный цвет зерна, затруднительно.

Генетический контроль формирования окраски зерен и других органов растений осуществляется генами, кодирующими ферменты, вовлеченные в биосинтез пигментов, и регуляторными генами. На данный момент путь биосинтеза антоцианов исследован достаточно хорошо, однако молекулярные механизмы биосинтеза меланина все еще слабо изучены (Шоева и др., 2018; Glagoleva et al., 2020). При исследовании механизмов генетического контроля окраски зерен селекционеры и генетики постоянно сталкиваются с необходимостью оценки цветовых характеристик их оболочки. К техническим средствам решения этой задачи относятся спектрофотометры, спектрометры, гиперспектральные камеры. Однако это дорогостоящие камеры, в особенности с высоким разрешением, как пространственным, так и спектральным. Альтернативой является использование цифровых фотокамер, позволяющих получать высококачественные изображения с высоким пространственным и цветовым разрешением (Afonnikov et al., 2016; Li et al., 2020; Kolhar, Jagtap, 2023). В связи с этим в последнее время в области фенотипирования растений интенсивно развиваются методы оценки цветовых и текстурных характеристик зерен злаков, основанные на анализе двумерных изображений, полученных цифровыми камерами или сканерами (Комышев и др., 2020; Sharma et al., 2021; Afonnikov et al., 2022; Arif et al., 2022; Khojastehnazhand, Roostaei, 2022; Wang, Su, 2022).

Цель данной работы – разработка метода оценки пигментного состава зерна ячменя на основе анализа цифровых изображений с помощью алгоритмов компьютерного зрения и машинного обучения.

Материалы и методы

Растительный материал. Для исследования были выбраны семена 39 образцов ячменя с темной окраской оболочек зерна, а также 40 образцов с неокрашенным зерном. Материал получен из коллекции ячменя Всероссийского института генетических ресурсов растений им. Н.И. Вавилова (ВИР, https://www.vir.nw.ru), коллекции ячменя Института цитологии и генетики СО РАН (ИЦиГ, https://www.icgbio.ru/). Также был использован материал популяции Oregon Wolfe Barleys (OWB, https://barleyworld.org/owb). Данные о каждом образце приведены в Приложении 1[1]. Отдельно были выбраны 29 образцов ячменя из коллекции ВИР с различными комбинациями пигментов в зерне (Приложение 2). Дополнительно образцы были охарактеризованы по пленчатости/голозерности. В обучающей и тестовой выборке было 58 пленчатых и 21 голозерный, а в отложенной выборке – 22 пленчатых и 7 голозерных образцов.

Химические и генетические методы определения пигментного состава зерен. Для определения качественного присутствия антоцианов в зерне была проведена экстракция в 1 % растворе HCl в метаноле с последующей детекцией окрашивания раствора в розовый цвет (Abdel-Aal, Hucl, 1999). Присутствие меланина определялось при помощи 2 % NaOH, в котором происходит экстрагирование меланина и окрашивание раствора в темный цвет (Downie et al., 2003). На основе этих методов каждому из образцов был присвоен тип пигментации «антоцианы», «меланины» (по наличию пигментов) либо «без пигментации», если оба пигмента в образцах отсутствовали. Данные о наличии пигментов определенного типа приведены в Приложениях 1 и 2.

Получение изображений. Цветные изображения зерен получены с помощью цифровой фотокамеры Canon EOS 600D, объектив Canon EF 100mm f/2.8 Macro USM с

[1] Приложения 1–8 см. по адресу:
https://vavilovj-icg.ru/download/pict-2023-27/appx28.pdf

Рис. 1. Пример изображения, полученного в результате выполнения протокола для фенотипирования образцов ячменя по интенсивности окраски зерна.

разрешением 18 Мп. На белый матовый лист бумаги формата А3 помещали пластиковую чашку Петри диаметром 55 мм, заполненную зернами без промежутков. По бокам располагался рассеивающий свет, камеру фиксировали на штативе сверху, объективом вертикально вниз (Приложение 3). Изображения сохранялись в формате jpg. Пример изображения, полученного в результате выполнения протокола, приведен на рис. 1.

При съемке в чашке Петри находилось около 100–160 зерен, принадлежащих одному образцу. Для каждого образца было сделано 9 изображений его реплик, полученных путем случайного перемешивания зерен в чашке Петри.

Разметка данных. С целью разработки алгоритма сегментации для 212 изображений 59 образцов, отобранных случайным образом, была выполнена ручная разметка зерен и границ чашки Петри с помощью программы LabelMe (https://github.com/wkentaro/labelme). Пример фрагмента размеченного изображения показан в Приложении 4. Кроме того, каждому изображению была присвоена метка в зависимости от типа пигментации соответствующего образца на основе экспериментально полученных данных.

Предсказание пигментации зерен с помощью методов машинного обучения. Общая схема предсказания типа пигментации включала сегментацию изображения для выделения на нем области, занятой зернами, и предсказание наличия пигментов определенного типа с помощью трех методов: 1) методом случайного леса с использованием цветовых дескрипторов изображения; 2) сверточной нейронной сетью архитектуры ResNet-18; 3) сверточной нейронной сетью архитектуры EfficientNetB0.

Схема разбиения данных для валидации и тестирования. Для методов машинного обучения изображения были разделены на три подвыборки: обучающую (60 % данных: 423 изображения, 47 образцов) – для обучения модели; валидационную (20 % данных: 144 изображения, 16 образцов) – для выбора лучшей модели в процессе обучения; тестовую (20 % данных: 144 изображения, 16 образцов) – для оценки точности выбранной модели. Для финальной оценки точности была использована отложенная выборка из 29 образцов, включавшая 261 изображение. Данные о разбиении образцов представлены в Приложении 5.

Оценка точности классификации образцов. Выходные данные обученных моделей классификации для каждого изображения были представлены двумя бинарными числами, каждое из которых характеризовало наличие или отсутствие антоцианов и меланина. Чтобы оценить точность метода на тестовой выборке, для каждого изображения сравнивались предсказанный набор таких чисел и истинный набор, так что если тестовая выборка содержала M изображений, проводилось $2M$ бинарных сравнений и на их основе подсчитывались следующие метрики: истинно положительные предсказания класса (TP, true positive), истинно отрицательные предсказания класса (TN, true negative), общее количество представителей положительного (P) и отрицательного (N) классов. С использованием этих величин вычислялось значение метрики «точность» (ACC, accuracy) согласно формуле

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}.$$

Модель для идентификации области зерен на изображении. Чтобы на изображении выделить зерна в чашках Петри, была реализована модель нейронной сети, сегментирующая исходные изображения. Ее выходными данными являлись бинарные маски. Для этого использовалась сеть U-Net с кодером ResNet-18. Архитектура U-Net разрабатывалась специально для сегментации биомедицинских изображений (Ronneberger et al., 2015). Модель основана на использовании преобразования свертки (convolution) и состоит из двух частей: кодер и декодер (рис. 2). Полноразмерное изображение на входе сети преобразуется кодером в результате нескольких шагов, включающих два последовательных слоя свертки размером 3×3, после которых идет преобразование ReLU (слои ‘conv 3×3, ReLU’ на рис. 2) и пулинг с функцией максимума 2×2 с шагом 2 (слои ‘max pool 2×2’). Кодер осуществляет понижающую дискретизацию изображения. Декодер, наоборот, выполняет повышающую дискретизацию изображения, используя серию операций, обратных пулингу, расширяющих карту признаков. Затем следует свертка 2×2, которая уменьшает количество каналов признаков (слои ‘up-conv 2×2’). Далее идет конкатенация с соответствующим образом обрезанной по краям картой признаков из сжимающего пути и две свертки 3×3 (слои ‘copy and crop’ на рис. 2), после каждой из которой применяется операция ReLU.

Сегментация позволяла выделить на изображении область чашки Петри с зернами, которая использовалась для вычисления их цветовых дескрипторов. Для каждого изображения было извлечено 2380 числовых параметров, характеризующих цвет пикселей зерен. Это средние значения интенсивности каналов для четырех цветовых пространств (RGB, HSV, Lab, YCrCb), значения гистограмм распределений интенсивностей цветовых компонент и т. п. Детальное описание полученных характеристик приведено в Приложении 6.

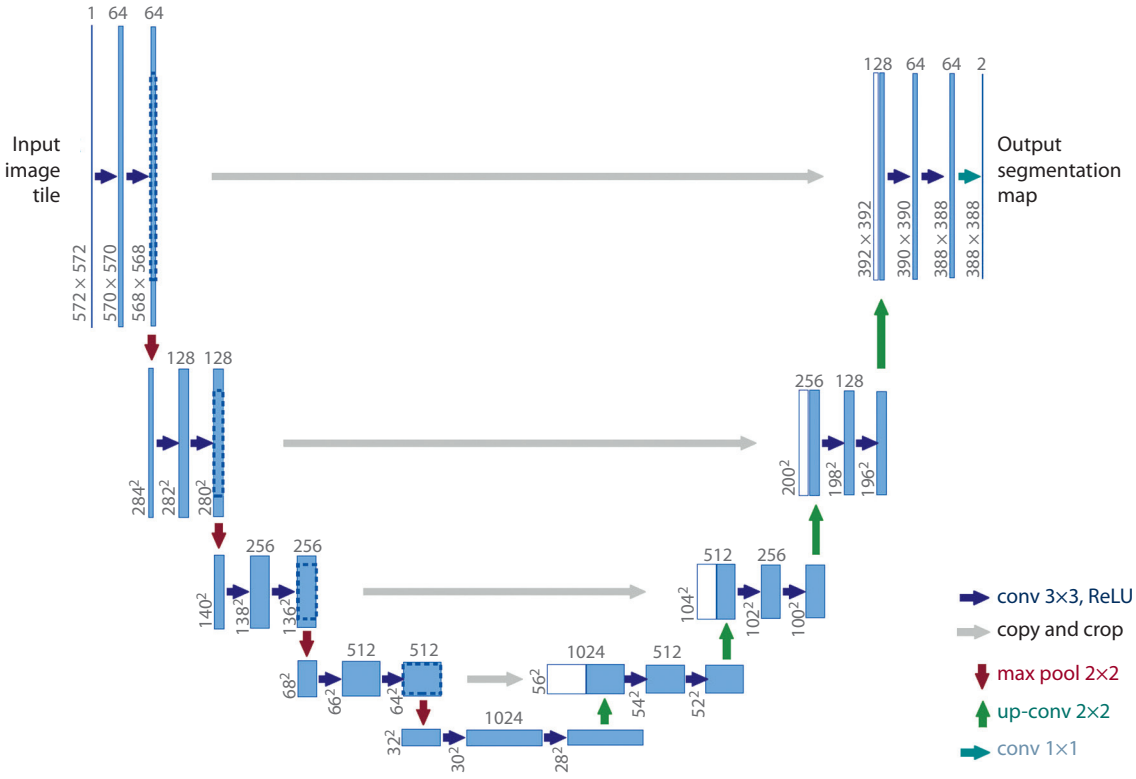


Рис. 2. Архитектура сети U-Net, из (Ronneberger et al., 2015).

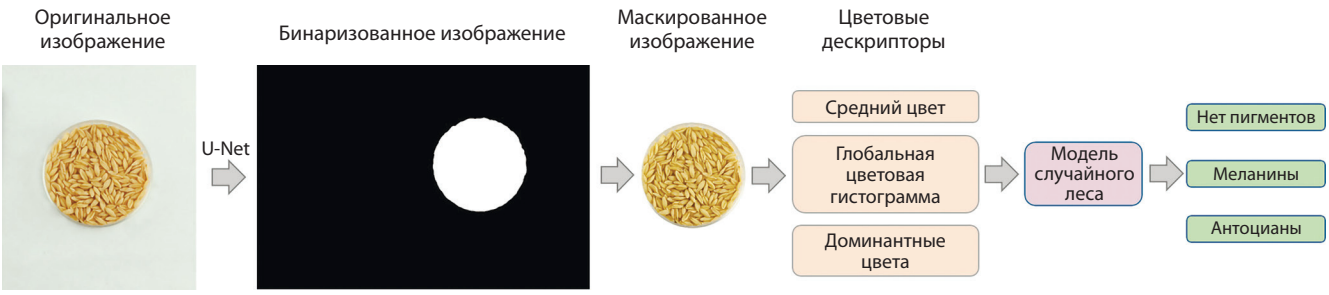


Рис. 3. Схема модели классификации зерен ячменя на основе изображений и алгоритма случайного леса с использованием цветовых дескрипторов.

Фильтрация данных. Для фильтрации малозначимых признаков были удалены следующие из них: значения которых одинаковы для всех изображений, значения которых не превышают 0.01 на более 20 % изображений. Дополнительно для уменьшения избыточности были удалены признаки, которые имели значение коэффициента корреляции Спирмена более 0.97 с другими признаками. В результате фильтрации осталось 345 дескрипторов из 2380.

Анализ данных. Для того чтобы оценить распределение образцов в пространстве изучаемых признаков, использовались метод главных компонент (Jolliffe, 2002) и нелинейный алгоритм снижения размерности t-SNE (van der Maaten, Hinton, 2008). Эти методы позволяют визуализировать многомерные данные путем отображения объектов в многомерном пространстве в пространство меньшей размерности (обычно двух- или трехмерное).

Модель классификации пигментного состава зерен на основе цветовых дескрипторов методом случайного леса

Рассматривалась классификация изображений зерен на четыре класса: без пигментации, пигментация антоцианами и/или меланинами, так что изображение зерен с пигментацией могло быть классифицировано как «содержащие антоцианы и меланины одновременно». Первая модель классификации была построена на основе алгоритма «случайный лес» (Random Forest), реализованного в пакете Scikit-learn (Pedregosa et al., 2011). На вход алгоритма подавались значения 345 дескрипторов цвета, описанных выше. Схема обработки данных для этой модели показана на рис. 3. Дополнительно при помощи метода главных компонент количество признаков было уменьшено до 13, которые объясняют 81.2 % дисперсии данных и дают мак-

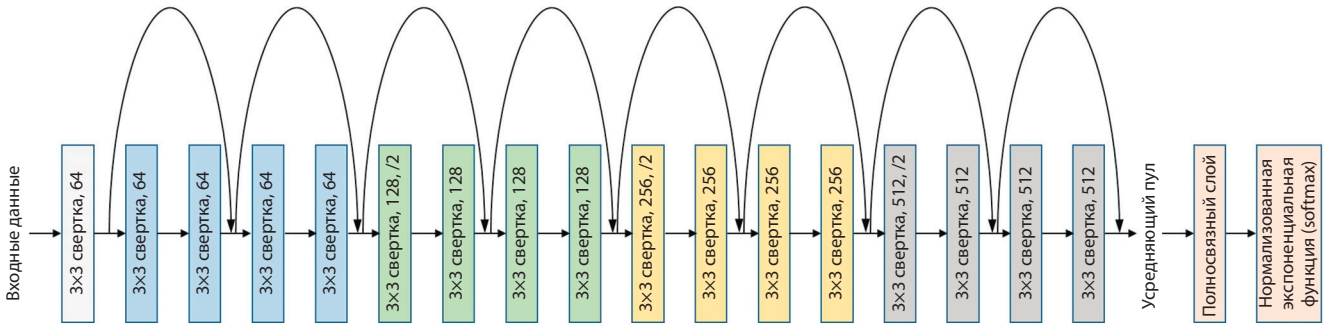


Рис. 4. Схема архитектуры сети ResNet-18.
Разноцветными прямоугольниками показаны слои сети различной структуры.

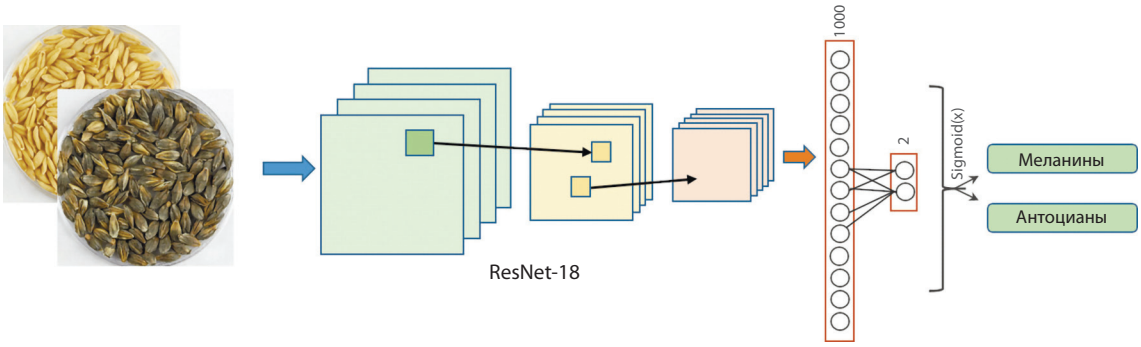


Рис. 5. Схема модели ResNet-18 классификации изображений зерен ячменя на основе сверточной нейронной сети.

симальную точность на тестовой выборке. Модель классификации обозначена нами как RF13.

Модели классификации пигментного состава зерен с использованием метода глубокого машинного обучения

Модель классификации на основе сети архитектуры ResNet-18. В дополнение к методу классификации изображений с помощью модели случайного леса, для предсказания типа образца были реализованы три модели, основанные на методах глубокого машинного обучения. Эти методы в настоящее время широко используются для анализа изображений растений и показали свою высокую точность.

Одна из моделей – нейронная сеть архитектуры ResNet-18, описанная в статье (He et al., 2016). ResNet – это семейство сверточных нейронных сетей (convolutional neural networks, CNN) сходной архитектуры, отличающихся количеством слоев (18, 34, 50, 101 и 152). В нашей работе использовалась модель глубиной 18 слоев, как наиболее простая и быстрая. Она состоит из последовательно идущих 17 слоев, включающих преобразование свертки, соединенных альтернативным путем для сигнала, и одного полносвязного слоя (рис. 4). Каждые четыре слоя происходит операция субдискретизации, при которой длина и ширина слоя становятся меньше в два раза, а число каналов двукратно возрастает. На рис. 4 это слои, обозначенные как «3×3 свертка, N», где N – число каналов.

На вход данной сети подавались прямоугольные изображения, в которые были вписаны области чашек Петри (рис. 5). Выходной слой включал два числа в интервалах от 0 до 1, предсказывающих присутствие (1) меланина или антоциана. В случае если значение числа было больше 0.5, считалось, что соответствующий пигмент присутствует в оболочке зерна. Такой метод позволял классифицировать изображения по наличию в зернах двух пигментов как по отдельности, так и совместно, а также идентифицировать их отсутствие в случае, если оба числа были меньше 0.5. Эта модель классификации была обозначена нами в работе как ResNet-18.

Модель на основе сегментации с головой для классификации. Для классификации зерен по наличию пигментов можно использовать параметры нейронной сети, которые были получены при сегментации изображений с помощью модели на основе U-Net. Это позволяет улучшить точность предсказания для алгоритмов подобного рода и решать две задачи одновременно (сегментация и классификация). Для решения этой задачи в уже имеющуюся модель на основе сегментации с архитектурой U-Net был добавлен дополнительный выходной классифицирующий слой («голова классификации») (рис. 6). На выходе классифицирующего слоя, как и в модели ResNet-18, имеются два числа, которые позволяют определить наличие в зернах антоцианов и/или меланинов (см. рис. 6). Для этой сети была использована топология кодера архитектуры EfficientNetB0, подробно представленная в статье (Tan, Le, 2019). Такая топология сети позволяла не только сегментировать изображение, выделяя на нем область зерен в чашке Петри, но и одновременно производить классификацию всего изображения по наличию или отсутствию двух пигментов. Данная модель классификации обозначена в работе как U-Net+ClassHead.

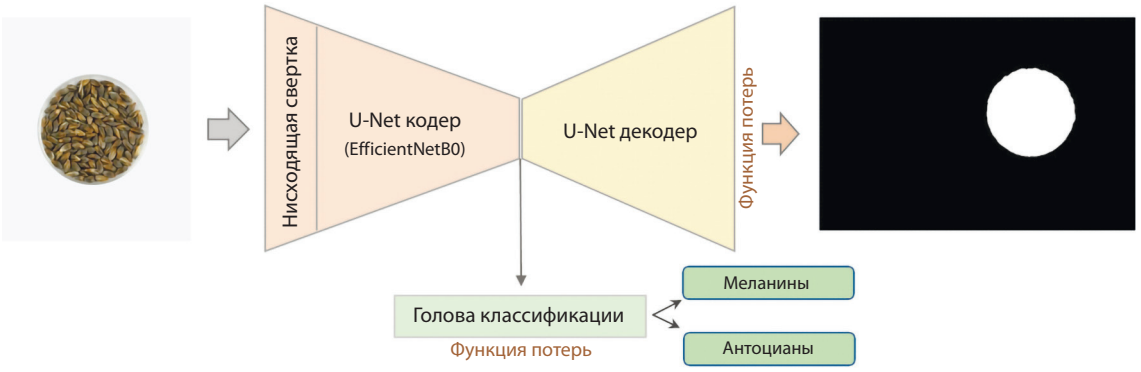


Рис. 6. Схема модели U-Net+ClassHead на основе U-Net сегментации с головой для одновременной сегментации и классификации изображений зерен ячменя.

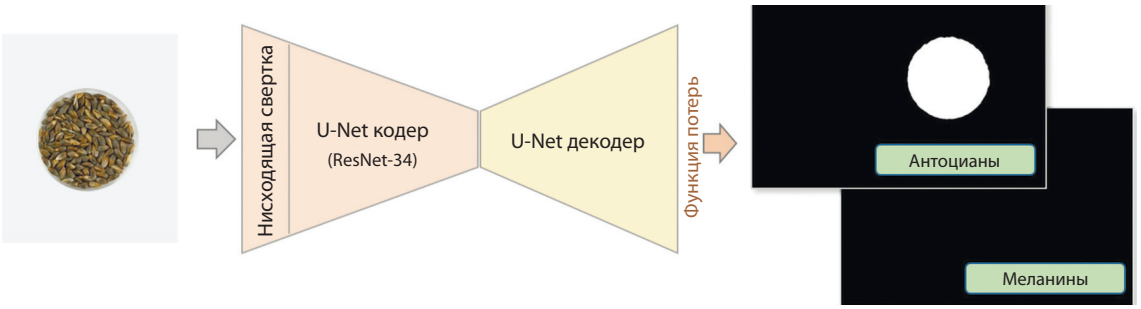


Рис. 7. Схема модели U-Net+ClassSegment для классификации на основе 2-канальной сегментации.

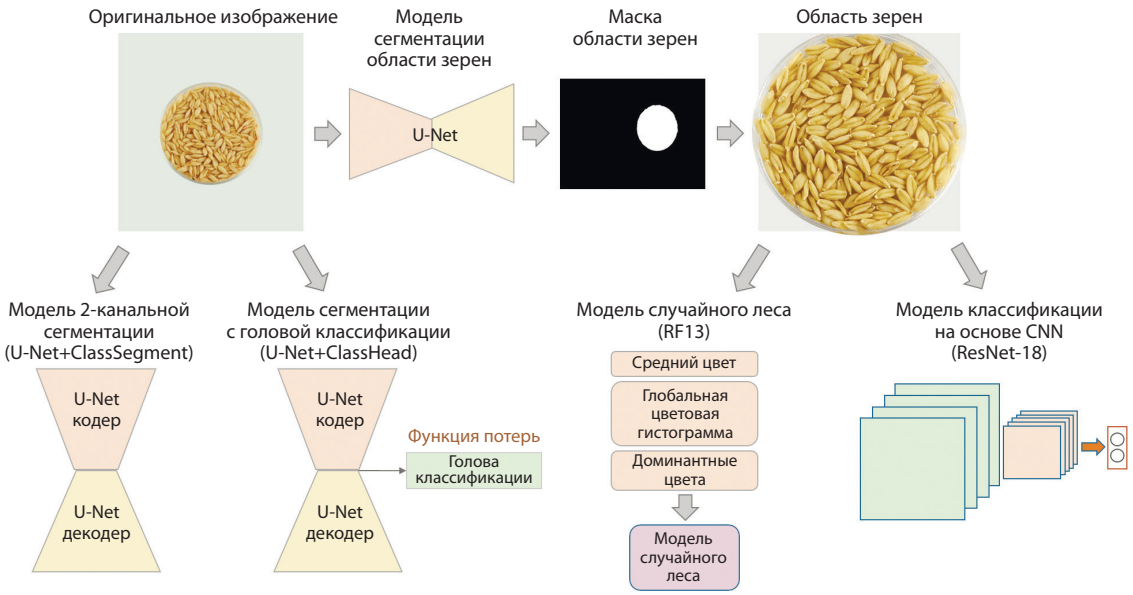


Рис. 8. Общая схема анализа изображений предложенными в работе моделями.

Модель 2-канальной сегментации. Для классификации изображений можно использовать сеть U-Net, видоизмененную таким образом, что она будет сегментировать каждый пиксель по наличию определенной пигментации. На выходе эта сеть выдает двухканальную маску, в которой каждый канал сегментирует область изображения, если образец содержит определенный пигмент (рис. 7). Эта модель, U-Net+ClassSegment, была основана на архитектуре U-Net с кодером ResNet-34. Для определения класса всего изображения мы считали, что если хоть один пиксель после сегментации был классифицирован как содержащий пигмент, то весь образец содержит данный пигмент.

Остальные технические параметры обучения моделей, такие как число эпох обучения, размер батча, использованная функция потерь и параметры оптимизатора, приведены в Приложении 7.

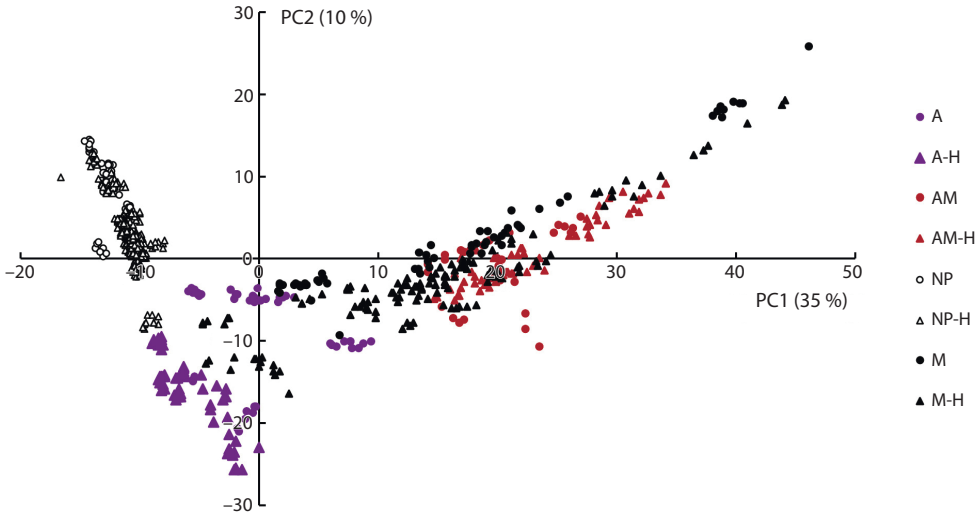


Рис. 9. Диаграмма рассеяния образцов в пространстве первых двух компонент, полученных на основе PCA для цветовых характеристик зерен.

Ось *X* – компонента PC1, ось *Y* – компонента PC2. В скобках для компонент указаны доли дисперсии. Здесь и на рис. 10 обозначения образцов по наличию антоцианов и пленчатости/голозерности: A, AM, M, NP – антоцианы, антоцианы и меланины, меланины и отсутствие пигментов соответственно; H – пленчатые зерна.

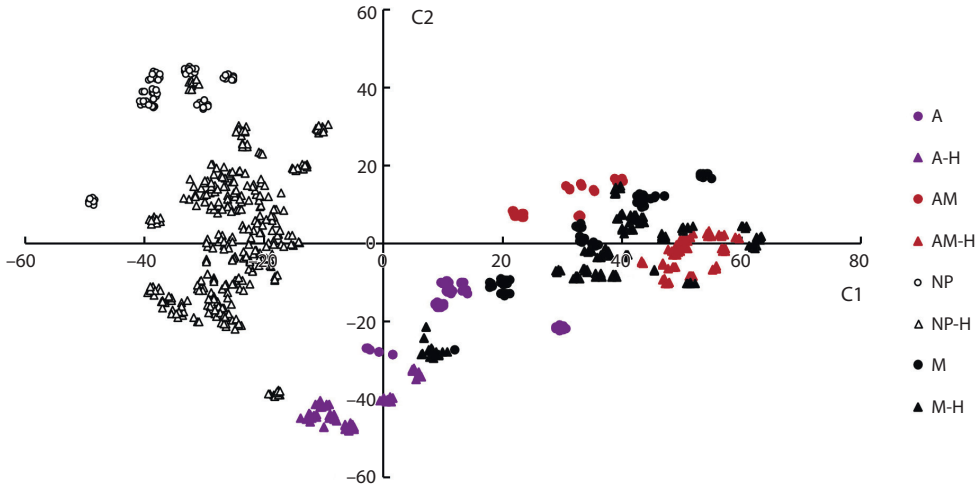


Рис. 10. Диаграмма рассеяния образцов в пространстве первых двух компонент, полученных в результате алгоритма t-SNE для цветовых характеристик зерен.

Ось *X* – компонента C1, ось *Y* – компонента C2.

Таким образом, в работе рассматривались две модели классификации на основе U-Net сегментации на исходном изображении (U-Net+ClassHead и U-Net+ClassSegment) и две модели классификации, для которых отдельно выделялась область зерен на исходных изображениях с помощью модели U-Net сегментации (RF13 и ResNet-18). Общая схема анализа изображений предложенными в работе моделями сегментации и классификации приведена на рис. 8.

Результаты

Цветовые характеристики зерен

Методы PCA и t-SNE были применены для отображения образцов зерен в пространстве обобщенных характеристик размерности 2 с использованием 345 информативных характеристик (см. Материалы и методы). Отфильтрованные признаки были подвергнуты нормализации (центрирование по математическому ожиданию и приведение к единичной дисперсии). При этом все изображения анализировались независимо, т. е. каждая точка на диаграммах PCA (рис. 9) и t-SNE (рис. 10) соответствует одному изображению.

Пигментированные (залитые значки) и непигментированные образцы (незалитые значки) хорошо разделяются на обеих диаграммах (см. рис. 9 и 10). При этом на диаграмме t-SNE (см. рис. 10) различия более выражены. Также образцы с антоцианами (фиолетовые значки) на диаграмме отделены от образцов, содержащих оба пигмента (красные). Области, занимаемые этими образцами

Таблица 1. Оценка точности классификации (ACC) образцов ячменя по содержанию антоцианов и меланинов в оболочке зерен для четырех моделей на валидационной, тестовой и отложенной выборках

Модель классификации	Валидационная выборка	Тестовая выборка	Отложенная выборка
RF13	0.896	0.903	0.652
ResNet-18	0.938	0.934	0.817
U-Net+ClassHead	0.906	0.962	0.821
U-Net+ClassSegment	0.917	0.903	0.819

Таблица 2. Параметры оценки точности классификации образцов ячменя по содержанию антоцианов и меланинов в оболочке зерен для модели U-Net+ClassHead на тестовой и отложенной выборках

Параметр	Тестовая выборка		Отложенная выборка	
	Меланин	Антоцианы	Меланин	Антоцианы
F-мера	1.0	0.937	0.983	0.488
Чувствительность	1.0	0.881	1.0	0.389
Положительное предсказуемое значение	1.0	1.0	0.966	0.656

на диаграммах, не пересекаются. В то же время области, занимаемые образцами с антоцианами (фиолетовые значки) и меланинами (черные), имеют пересечение. Образцы, содержащие как антоцианы, так и меланины, и образцы, содержащие только меланины, тесно пересекаются (справа по оси X и ближе к 0 по оси Y).

Отметим также влияние пленчатости зерен на характеристики их цвета, которое заметно на двух указанных графиках. Прежде всего, пленчатость/голозерность не влияет на разделение областей для разных классов зерен по пигментации за исключением пар, содержащих антоцианы либо меланин: пленчатые и голозерные образцы с одним типом окраски находятся ближе друг к другу, чем образцы с другим типом окраски. Особенно явно это проявляется для зерен без пигментации (незалитые значки). Для зерен с меланином одна из групп пленчатых образцов имеет цветовые характеристики, весьма схожие с зернами, содержащими антоцианы (на графиках эта группа расположена внутри области, занимаемой образцами с антоцианами, и отстоит далеко от других зерен с окраской меланином). Внутри одного пигментного класса голозерные и пленчатые образцы занимают разные области и хорошо разделяются (см. рис. 10, образцы без пигментации, образцы с антоцианами и образцы с антоцианами и меланинами). Эти результаты показывают, что в большинстве случаев пленчатость не изменяет тип окраски зерен, но вносит существенный вклад в изменение цветовых характеристик оболочки.

Классификация зерен по содержанию пигментов

В результате обучения моделей классификации образцов зерен по содержанию пигментов были получены оценки точности на валидационной, тестовой и отложенной выборках (табл. 1).

Лучший результат классификации на отложенной выборке демонстрирует модель сегментации с «головой классификации» (U-Net+ClassHead). Данные о параметрах оценок точности данной модели приведены в табл. 2.

Матрица ошибок предсказания класса образцов зерен (Приложение 8) позволяет определить, что большинство неверных классификаций модели приходится на предсказание содержания антоцианов в пленчатых образцах, что согласуется с графиками PCA и t-SNE (см. рис. 9 и 10), на которых области для пленчатых образцов, содержащих меланины и антоцианы, значительно перекрываются с областями образцов, содержащих только меланины. Причем образцов, для которых сеть предсказала отсутствие пигментов, но содержащих антоцианы, значительно больше образцов, для которых предсказано наличие антоцианов, тогда как они отсутствовали. Ошибки наблюдаются и для голозерных образцов. В первую очередь это образцы, содержащие антоцианы, для которых сеть предсказала их отсутствие. В небольшом числе образцов, содержащих меланины, эти пигменты не были идентифицированы с помощью нейронной сети. В то же время образцы, содержащие антоцианы, были классифицированы сетью как содержащие меланины.

Согласно непараметрическому тесту Манна–Уитни, точность определения антоцианов различается (p-value = = 0.004) для голозерных и пленчатых образцов. На определение меланинов пленчатость/голозерность не оказывает влияния.

Несколько меньшую точность показал метод U-Net+ ClassSegment. Можно сделать вывод о лучшей обобщающей способности моделей, которые одновременно решают несколько разных задач (multi task learning). Обе модели, основанные на этом подходе, существенно выигрывают как у метода случайного леса, использующего цветовые дескрипторы (наиболее низкая точность), так и у классификации при помощи модели ResNet-18. Надо отметить, что результаты точности на отложенной выборке существенно ниже, чем на тестовой.

Обсуждение

Методы анализа цифровых RGB изображений для изучения физиологических свойств зерен широко применяются для злаков (Neuman et al., 1989; Huang et al., 2015; Sabanci et al., 2017; Kozłowski et al., 2019; Комышев и др., 2020; Zykin et al., 2020). Они используются в том числе для классификации зерен как по пигментному составу, так и по сортам.

В нашей работе мы анализировали методы классификации зерен по цветовым характеристикам на классы по наличию двух типов пигментов. Мы показали, что методы глубокого машинного обучения позволяют получить более высокую точность классификации зерен, чем использование цветовых дескрипторов. Аналогичные выводы были сделаны при классификации зерен ячменя по видам (Kozłowski et al., 2019). Наши результаты также демонстрируют, что применение многозадачного подхода (multi task learning) дает более высокую точность классификации.

На отложенной выборке точность классификации оказалась существенно меньшей по сравнению с тестовой выборкой. Предположительно, одной из причин этого могло быть то, что баланс меток различных классов в обучающей, валидационной и тестовой выборках был одинаковым и не совпадал с соотношением в отложенной выборке. В частности, изображений с зернами без пигментов в отложенной выборке было в 1.5 раза меньше, чем в обучающей, а для классификации такие зерна представляют собой наиболее простой случай. Также на основании извлеченных цветовых дескрипторов был обучен бинарный классификатор, который отличал зерна из отложенной выборки от других зерен с точностью ACC = 1. Это означает, что между данными сериями изображений есть существенные различия, которые могут быть объяснены тем, что в отложенной выборке были выбраны зерна из других коллекций или вариации в условиях съемки этих изображений могли оказать существенное влияние. Этим можно объяснить небольшое уменьшение точности при качестве классификации у модели случайного леса.

Наш анализ также показал, что пленчатость или голозерность у ячменя – это признак, который влияет на его цветовые характеристики. Этот признак может оказывать влияние и на классификацию зерен по наличию пигментов.

Заключение

Предложенные методы на основе анализа цифровых изображений с помощью алгоритмов компьютерного зрения и машинного обучения показали приемлемую классифицирующую способность в задаче определения содержания меланина и антоцианов в зернах ячменя. Результаты демонстрируют, что применение алгоритма Random Forest с использованием цветовых дескрипторов уступает в итоговой точности подходам на основе сверточных нейронных сетей. Метод случайного леса, кроме того, оказывается чувствительным к малым вариациям протокола или условий съемки, теряя обобщающую способность по сравнению со сверточными нейронными сетями. Возможные пути улучшения модели на основе данного алгоритма – тщательный подбор признаков и предварительная нормализация изображений, подаваемых на вход. Классическая архитектура классификационной модели уступает по точности модели 2-канальной сегментации целого изображения. Сегментация с «головой классификации» дала наилучшие результаты (ACC = 0.821) и является предпочтительной в задаче определения содержания пигментов ячменя.

Список литературы / References

Комышев Е.Г., Генаев М.А., Афонников Д.А. Анализ цветовых и текстурных характеристик зерен злаков на цифровых изображениях. *Вавиловский журнал генетики и селекции*. 2020;24(4):340-347. DOI 10.18699/VJ20.626
[Komyshev E.G., Genaev M.A., Afonnikov D.A. Analysis of color and texture characteristics of cereals on digital images. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2020;24(4):340-347. DOI 10.18699/VJ20.626]

Шоева О.Ю., Стрыгина К.В., Хлесткина Е.К. Гены, контролирующие синтез флавоноидных и меланиновых пигментов ячменя. *Вавиловский журнал генетики и селекции*. 2018;22(3):333-342. DOI 10.18699/VJ18.369
[Shoeva O.Yu., Strygina K.V., Khlestkina E.K. Genes determining the synthesis of flavonoid and melanin pigments in barley. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2018;22(3):333-342. DOI 10.18699/VJ18.369 (in Russian)]

Abdel-Aal E.S.M., Hucl P. A rapid method for quantifying total anthocyanins in blue aleurone and purple pericarp wheats. *Cereal Chem*. 1999;76(3):350-354. DOI 10.1094/CCHEM.1999.76.3.350

Afonnikov D.A., Genaev M.A., Doroshkov A.V., Komyshev E.G., Pshenichnikova T.A. Methods of high-throughput plant phenotyping for large-scale breeding and genetic experiments. *Russ. J. Genet*. 2016;52(7):688-701. DOI 10.1134/S1022795416070024

Afonnikov D.A., Komyshev E.G., Efimov V.M., Genaev M.A., Koval V.S., Gierke P.U., Börner A. Relationship between the characteristics of bread wheat grains, storage time and germination. *Plants*. 2022;11(1):35. DOI 10.3390/plants11010035

Arif M.A.R., Komyshev E.G., Genaev M.A., Koval V.S., Shmakov N.A., Börner A., Afonnikov D.A. QTL analysis for bread wheat seed size, shape and color characteristics estimated by digital image processing. *Plants*. 2022;11(16):2105. DOI 10.3390/plants11162105

Ceccarelli S., Grando S., Van Leur J.A.G. Genetic diversity in barley landraces from Syria and Jordan. *Euphytica*. 1987;36(2):389-405. DOI 10.1007/BF00041482

Choo T.M., Vigier B., Ho K.M., Ceccarelli S., Grando S., Franckowiak J.D. Comparison of black, purple, and yellow barleys. *Genet. Resour. Crop Evol*. 2005;52(2):121-126. DOI 10.1007/s10722-003-3086-4

Downie A.B., Zhang D., Dirk L.M.A., Thacker R.R., Pfeiffer J.A., Drake J.L., Levy A.A., Butterfield D.A., Buxton J.W., Snyder J.C. Communication between the maternal testa and the embryo and/or endosperm affect testa attributes in tomato. *Plant Physiol*. 2003;133(1):145-160. DOI 10.1104/pp.103.022632

Ferdinando M.D., Brunetti C., Fini A., Tattini M. Flavonoids as antioxidants in plants under abiotic stresses. In: Ahmad P., Prasad M. (Eds.) Abiotic Stress Responses in Plants. New York: Springer, 2012;159-179. DOI 10.1007/978-1-4614-0634-1_9

Flintham J., Adlam R., Bassoi M., Holdsworth M., Gale M. Mapping genes for resistance to sprouting damage in wheat. *Euphytica*. 2002;126:39-45. DOI 10.1023/A:1019632008244

Glagoleva A.Y., Shoeva O.Y., Khlestkina E.K. Melanin pigment in plants: current knowledge and future perspectives. *Front. Plant Sci*. 2020;11:770. DOI 10.3389/fpls.2020.00770

He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016. IEEE, 2016;770-778. DOI 10.1109/CVPR.2016.90

Huang M., Wang Q.G., Zhu Q.B., Qin J.W., Huang G. Review of seed quality and safety tests using optical sensing technologies. *Seed Sci. Technol*. 2015;43(3):337-366. DOI 10.15258/sst.2015.43.3.16

Jana B.K., Mukherjee S.K. Notes on the distribution of phytomelanin layer in higher plants – a short communication. *J. Pharm. Biol*. 2014;4(3):131-132

Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics. New York: Springer, 2002. DOI 10.1007/b98835

Khojastehnazhand M., Roostaei M. Classification of seven Iranian wheat varieties using texture features. *Expert Syst. Appl*. 2022;199: 117014. DOI 10.1016/j.eswa.2022.117014

Kolhar S., Jagtap J. Plant trait estimation and classification studies in plant phenotyping using machine vision. A review. *Inf. Process. Agric*. 2023;10(1):114-135. DOI 10.1016/j.inpa.2021.02.006

Kozłowski M., Górecki P., Szczypiński P.M. Varietal classification of barley by convolutional neural networks. *Biosyst. Eng*. 2019;184: 155-165. DOI 10.1016/j.biosystemseng.2019.06.012

Li Z., Guo R., Li M., Chen Y., Li G. A review of computer vision technologies for plant phenotyping. *Comput. Electron. Agric*. 2020;176: 105672. DOI 10.1016/j.compag.2020.105672

Neuman M.R., Sapirstein H.D., Shwedyk E., Bushuk W. Wheat grain colour analysis by digital image processing II. Wheat class discrimination. *J. Cereal Sci*. 1989;10(3):183-188. DOI 10.1016/S0733-5210(89)80047-5

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. 2011;12:2825-2830

Ronneberger O., Fischer P., Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N., Hornegger J., Wells W., Frangi A. (Eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Vol. 9351. Cham: Springer, 2015;234-241. DOI 10.1007/978-3-319-24574-4_28

Sabanci K., Kayabasi A., Toktas A. Computer vision-based method for classification of wheat grains using artificial neural network. *J. Sci. Food Agric*. 2017;97(8):2588-2593. DOI 10.1002/jsfa.8080

Sharma R., Kumar M., Alam M.S. Image processing techniques to estimate weight and morphological parameters for selected wheat refractions. *Sci. Rep*. 2021;11(1):20953. DOI 10.1038/s41598-021-00081-4

Souza F.H., Marcos-Filho J. The seed coat as a modulator of seed-environment relationships in Fabaceae. *Braz. J. Bot*. 2001;24(4):365-375. DOI 10.1590/S0100-84042001000400002

Tan M., Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, 9–15 June 2019. ICML, 2019;6105-6114

van der Maaten L., Hinton G. Visualizing data using t-SNE. *J. Mach. Learn. Res*. 2008;9(11):2579-2605.

Wang Y.H., Su W.H. Convolutional neural networks in computer vision for grain crop phenotyping: a review. *Agronomy*. 2022;12(11):2659. DOI 10.3390/agronomy12112659

Zykin P.A., Andreeva E.A., Tsvetkova N.V., Voylokov A.V. Anatomical and image analysis of grain coloration in rye. *Preprints.* 2020; 2020110530. DOI 10.20944/preprints202011.0530.v1

ORCID ID

D.A. Afonnikov orcid.org/0000-0001-9738-1409

Благодарности. Разработка протокола фенотипирования, алгоритма классификации и тестирования проводилась при финансовой поддержке Российского научного фонда (проект № 22-74-00122, https://rscf.ru/project/22-74-00122/). Для анализа данных использовались вычислительные ресурсы ЦКП «Биоинформатика» при поддержке бюджетного проекта FWNR-2022-0020.
Авторы выражают благодарность Е.А. Заварзину и А.И. Ивлевой за участие в обучении моделей нейронных сетей.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 30.06.2023. После доработки 27.09.2023. Принята к публикации 28.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):869-877
DOI 10.18699/VJGB-23-100

Перевод на английский язык https://vavilov.elpub.ru/jour

Математическое моделирование динамики кворум-эффекта в накопительной культуре люминесцентных бактерий *Photobacterium phosphoreum* 1889

С.И. Барцев[1, 2] ✉, А.Б. Сарангова[2]

[1] Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
[2] Сибирский федеральный университет, Красноярск, Россия
✉ bartsev@yandex.ru

Аннотация. В начале статьи обсуждается уровень необходимой феноменологичности сложных моделей. При работе со сложными системами, к которым, безусловно, относятся живые организмы и экологические системы, с необходимостью приходится использовать феноменологическое описание. Приведена иллюстрация феноменологического подхода, который ухватывает наиболее существенные даже не закономерности, а общие принципы или паттерны взаимодействий, причем конкретные значения параметров не могут быть вычислены из первых принципов, а определяются эмпирически. Также эмпирически и прагматически выбирается соответствующая интерпретация. Однако для моделирования более широкого круга ситуаций возникает необходимость понижать уровень феноменологии, переходить на более детальное описание системы, вводя взаимодействие между выделенными элементами системы. Формулируются требования к модели системы, совмещающей экологический, метаболический и генетический уровни описания клеточной культуры. Разработана математическая модель динамики кворум-эффекта в процессе роста накопительной культуры люминесцентных бактерий при разных концентрациях питательного субстрата. Модель содержит четыре блока, описывающие экологический, энергетический, кворумный и люминесцентный аспекты развития культуры. Модель продемонстрировала хорошее соответствие экспериментальным данным, полученным в ходе выполнения работы. При анализе модели отмечены три странности в поведении культуры, которые, предположительно, могут изменить представление о некоторых процессах, имеющих место при развитии культуры люминесцентных бактерий. Полученные результаты позволяют предположить наличие некоторой дополнительной системы контроля люминесцентной реакции через пути синтеза ФМН·H_2 или алифатического альдегида. В этом случае обобщенное описание вклада энергетического метаболизма в люминесценцию только через АТФ является слишком сильным упрощением. В результате анализа результатов сопоставления модельной динамики с экспериментом возникло расхождение между измеряемой в эксперименте концентрацией субстрата (пептона) и его эффективным действием на рост популяции бактерий. Это расхождение, по-видимому, указывает на то, что пептон не является ведущим субстратом и рост лимитируют биогены, содержащиеся в дрожжевом экстракте, концентрация которого в этих экспериментах не изменялась. Отмеченные расхождения между ожиданиями и результатами обработки экспериментальных данных вместе с предположениями о причинах этих расхождений задают направление дальнейших экспериментальных и теоретических исследований механизмов кворум-эффекта в культуре люминесцентных бактерий.
Ключевые слова: кворум-эффект; математическая модель; люминесцентные бактерии.

Для цитирования: Барцев С.И., Сарангова А.Б. Математическое моделирование динамики кворум-эффекта в накопительной культуре люминесцентных бактерий *Photobacterium phosphoreum* 1889. *Вавиловский журнал генетики и селекции*. 2023;27(7):869-877. DOI 10.18699/VJGB-23-100

Mathematical modeling of quorum sensing dynamics in batch culture of luminescent bacterium *Photobacterium phosphoreum* 1889

S.I. Bartsev[1, 2] ✉, A.B. Sarangova[2]

[1] Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
[2] Siberian Federal University, Krasnoyarsk, Russia
✉ bartsev@yandex.ru

Abstract. At the beginning of the paper, the level of necessary phenomenology of complex models is discussed. When working with complex systems, which of course include living organisms and ecological systems, it is necessary to use a phenomenological description. An illustration of the phenomenological approach is given, which captures the most significant general principles or patterns of interactions; the specific values of the parameters cannot be calculated

from the first principles, but are determined empirically. An appropriate interpretation is also chosen empirically and pragmatically. However, in order to simulate a wider range of situations, it becomes necessary to lower the level of phenomenology, switch to a more detailed description of the system, introducing interaction between selected elements of the system. The requirements for a system model combining ecological, metabolic and genetic levels of cell culture description are formulated. A mathematical model of quorum sensing dynamics during the growth of batch culture of luminescent bacteria at different concentrations of the nutrient substrate has been developed. The model contains four blocks describing ecological, energy, quorum and luminescent aspects of bacterial culture growth. The model demonstrated good agreement with the experimental data obtained. When analyzing the model, three oddities in the behavior of the culture were noted, which presumably can change the idea of some processes taking place during the development of a culture of luminescent bacteria. The results obtained suggest the presence of some additional control system for the luminescent reaction via the synthesis pathways of FMN·H_2 or aliphatic aldehyde. In this case, the generalized description of the contribution of energy metabolism to luminescence only through ATP is too strong a simplification. As a result of comparing the model dynamics with the experiment, a discrepancy arose between the concentration of the substrate (peptone) measured in the experiment and its effective influence on the bacterial population growth. This discrepancy seems to indicate peptone is not the leading substrate, and growth is limited by nutrients contained in the yeast extract, the concentration of which did not change in these experiments. The discrepancies noted between the expectations and the results of experimental data processing, together with the assumptions about the causes of these discrepancies, set the direction for further experimental and theoretical studies of quorum sensing mechanisms in a culture of luminescent bacteria

Key words: quorum sensing; mathematical model; luminescent bacteria.

For citation: Bartsev S.I., Sarangova A.B. Mathematical modeling of quorum sensing dynamics in batch culture of luminescent bacterium *Photobacterium phosphoreum* 1889. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):869-877. DOI 10.18699/VJGB-23-100

Введение

При работе со сложными системами, к которым, безусловно, относятся живые организмы и экологические системы, с необходимостью приходится использовать феноменологическое описание. Одним из примеров феноменологического описания популяции является уравнение Ферхюльста. Несмотря на то что формально уравнения на численность можно использовать лишь вблизи порога выживания популяции (Горбань и др., 1982), это уравнение достаточно хорошо описывает динамику различных процессов: накопительной культуры микроорганизмов, развития эпидемии в фиксированных условиях, развития популяции после инвазии и динамики продаж в условиях ограниченной емкости рынка. По-видимому, это связано с тем, что на завершающей стадии развития процесса, когда величина переменной приближается к значению предельной емкости среды, удельная скорость роста приближается к 0, что, по сути, соответствует приближению к порогу выживания.

Уравнение Ферхюльста можно записать несколькими вариантами, которым соответствуют разные интерпретации. Рассмотрим для примера два из них:

$$\dot{N} = \mu_0(N_{\max} - N)N, \quad (1a)$$

$$\dot{N} = \mu_0 N - \alpha N^2. \quad (1b)$$

В первом варианте $N_{\max}$ называют емкостью среды, понимаемой как предельная численность популяции, которая может существовать в данных условиях, а произведение $\mu_0 N_{\max}$ представляет собой удельную скорость роста популяции при близкой к нулю ее численности. Емкость среды феноменологически включает всевозможные факторы, ограничивающие рост популяции: ограниченность субстрата, ингибирование метаболитами, фиксированную поверхность светосбора для растений. Этот вариант хорошо соответствует интерпретации роста растительной или микробной популяции.

Во втором варианте μ_0 – удельная скорость роста, α – коэффициент, описывающий внутривидовую конкуренцию, которая может реализоваться разными механизмами: конкуренцией за пищу и/или вытеснением с охотничьей территории и прямыми столкновениями особей. Такая интерпретация представляется более подходящей для животного мира.

Эти примеры приведены для иллюстрации феноменологического подхода, который ухватывает наиболее существенные даже не закономерности, а общие принципы или паттерны взаимодействий, причем конкретные значения параметров не могут быть вычислены из первых принципов, а определяются эмпирически. Эмпирически и прагматически выбирается также соответствующая интерпретация.

Однако для моделирования более широкого круга ситуаций возникает необходимость понижать уровень феноменологии, переходить на более детальное описание системы, вводя взаимодействие между выделенными элементами системы. Так, встречаются случаи, когда уравнение Ферхюльста описывает динамику накопительной культуры недостаточно точно. В этом случае приходится, например, учитывать динамику субстрата и вводить, допустим, субстратное ингибирование роста культуры. При этом мы все равно остаемся на очень высоком уровне феноменологии, продолжая описывать зависимость роста культуры формулой Моно и ее различными модификациями и усложнениями, сводя весь метаболизм клетки или многоклеточного организма к одной ключевой ферментативной реакции.

Потребность в понижении уровня феноменологии возникает, когда исследователь встречается с явлениями, которые не укладываются в имеющуюся модель. При этом часто приходится переходить на уровень генетической и/или метаболической регуляции клеточных процессов. Один из примеров, требующих понижения феноменоло-

гичности используемых моделей, – кворум-эффект (КЭ) (Miller, Bassler, 2001). Примечательно и символично, что кворум-эффект (quorum sensing), представляющий собой проявление событий молекулярного уровня на уровне популяции, был открыт на люминесцентных бактериях, чье свечение служит естественным индикатором текущего состояния клеточного метаболизма (Nealson et al., 1970).

Кворум-эффект заключается в том, что экспрессия некоторых генов запускается при достижении определенной пороговой плотности популяции. На уровне бактерий этот эффект основан на синтезе и выделении во внешнюю среду сигнальных молекул (аутоиндукторов), концентрация которых изменяется в зависимости от количества окружающих клеток, и при превышении некоторой пороговой концентрации запускается экспрессия определенных генов. Поскольку кворум-эффект встречается у широкого круга организмов (например, у насекомых (Anstey et al., 2009) и рыб (Makris et al., 2009)), то его изучение само по себе представляется важным. Кроме того, выявление закономерностей проявления кворум-эффекта и его прогноз важны для микробиологического синтеза продуктов, запускаемого этим эффектом. Примером такого продукта является бактериальная люцифераза, препараты которой используются для лабораторных и токсикологических экспресс-биотестов. При этом люминесцентные бактерии являются удобным инструментом исследования кворум-эффекта, поскольку люминесценция представляет собой естественную функцию клеток, что дает возможность изучать процесс на нативных клетках без внедрения специальных флуоресцентных красителей и без стимулирующего флюоресценцию излучения. Эволюционный смысл кворум-эффекта у люминесцентных бактерий находит объяснение в рамках гипотезы, что механизм отбора связан с рассеиванием и размножением бактерий (Nealson, Hastings, 1979). Являясь морскими энтеробактериями, светящиеся бактерии, растущие на субстрате (поверхность мертвых организмов или фекальные шарики), при достаточной плотности культуры могут производить свечение, способное привлечь организмы к их поглощению, обеспечивая тем самым круговорот бактерий по кишечным трактам морских животных.

Целью настоящей работы была разработка математической модели и ее программной реализации для анализа экспериментальных данных по кворум-эффекту в накопительной культуре люминесцентных бактерий. Для конкретизации требований к модели сформулируем своеобразное техническое задание (ТЗ). Во-первых, модель должна описывать динамику роста бактерий в накопительной культуре, во-вторых, она должна описывать динамику свечения бактериальной культуры, которая регулируется кворум-эффектом, т. е. событиями на молекулярном уровне, и, в-третьих, модель должна быть максимально простой. Сложная модель содержит большое количество параметров, значения которых неизвестны, т. е. мы следуем парадигме, что чем меньше подгоночных параметров в модели, описывающей сложные процессы, тем больше она отображает сущность моделируемых процессов.

В третьем пункте нашего условного ТЗ упоминается сложность модели, и поскольку этот пункт взыскует к простоте создаваемой модели, то требуется хотя бы краткое обсуждение этого термина. К сожалению, универсального определения сложности нет, об этом говорит огромное (более сорока) количество существующих определений сложности (Edmonds, 1999). Особенности применения этого термина к описанию живых эволюционирующих систем позволяют сузить набор возможных определений (Барцев, Барцева, 2010). В случае математических моделей, построенных в виде систем обыкновенных дифференциальных уравнений (ОДУ), часто используемых для описания химической (биохимической) кинетики и динамики экологических систем, естественным (или по крайней мере широко используемым) показателем сложности является количество дифференциальных уравнений в системе. По-видимому, не зря методы, позволяющие понизить размерность системы ОДУ, например, за счет выделения подсистемы быстрых движений и применения теоремы Тихонова (Романовский и др., 1984), называются методами упрощения систем кинетических уравнений.

Правда остается вопрос о сложности самих уравнений, вернее, их правых частей. Очевидно, что функции, включающие большее число (если так можно выразиться) нелинейностей, например слагаемых с большими степенями в дробно-рациональной функции, могут обеспечить более разнообразное поведение. Количественный подход к оценке сложности систем ОДУ с учетом степени нелинейности правых частей может быть основан на теореме Корзухина (Жаботинский, 1974), которая утверждает, что для системы с нелинейными правыми частями можно построить систему уравнений химической кинетики (содержащую члены, которые описывают реакции не выше второго порядка) такую, что поведение некоторых из переменных новой системы будет совпадать с поведением переменных исходной системы. Количество уравнений второй, развернутой системы могло бы служить мерой сложности модели с учетом степени нелинейности используемых правых частей. Поскольку наша задача состоит не в получении точной оценки сложности разрабатываемой модели, а лишь в построении максимально простой модели, обеспечивающей адекватное описание реальной системы, то мы будем минимизировать количество дифференциальных уравнений модели и одновременно использовать минимальные степени переменных в правых частях уравнений.

Методы и материалы

Экспериментальная часть. Объектом исследования являются светящиеся бактерии *Photobacterium phosphoreum* 1889 из коллекции Института биофизики СО РАН. Рост бактерий оценивали путем измерения оптической плотности при 660 нм на спектрофотометре Agilent Cary 60. Для измерения биолюминесценции реакционной смеси использовали люминометр (Promega GloMax 20/20 Luminometer, США). Выращивание бактерий было проведено на жидкой среде для морских бактерий (г/л): NaCl – 28.5, KCl – 0.5, $CaCl_2$ – 0.5, $MgCl_2$ – 4.5, дрожжевой экстракт – 1, пептон – 10; pH 7.6.

Математическая модель. Биолюминесцентная система бактерий к настоящему времени очень хорошо изучена (Brodl et al., 2018). Известны ферменты, экспрессируемые

совместно при срабатывании КЭ, достаточно хорошо изучены пути синтеза субстратов люминесцентной реакции. Для нас, чтобы не погружаться в детали кинетики полиферментной системы, важно следующее: непосредственными субстратами люминесцентной реакции являются: восстановленный флавинмононуклеотид (ФМН·H_2), длинноцепочечный алифатический альдегид – тетрадеканаль и молекулярный кислород. Флавин восстанавливается ферментом НАДН:ФМН-оксидоредуктазой, а альдегид синтезируется с помощью ферментного комплекса редуктазы жирной кислоты с потреблением АТФ. Тем самым люминесцентная реакция непосредственно связана с энергетическим метаболизмом клетки, и ее свечение зависит не только от количества люциферазы в клетке, но и от состояния ее энергетического метаболизма.

Следовательно, уже на уровне описания роста культуры в модель нужно закладывать оценку состояния ее энергетического метаболизма. Подробно свойства полиферментной системы энергетического метаболизма исследовались в почти забытой (если судить по статистике цитирования из ResearchGate) работе Е.Е. Селькова, составляющей часть коллективной монографии (Иваницкий и др., 1978). Одним из важнейших свойств энергетического метаболизма является поддержание в достаточно широких пределах постоянства внутриклеточной концентрации АТФ, чтобы обеспечить развязку (относительную независимость) внутриклеточных потребителей энергии. В упомянутой работе рассматривался случай постоянства скорости поступления субстрата при варьировании нагрузки (активности обобщенной АТФ-азы). В этой модели нужно учитывать и изменение скорости поступления субстрата (в нашем случае будем рассматривать его концентрацию в среде), и изменения АТФ-азной активности, связанные с разными фазами роста культуры. В нашей модели вся модель Селькова, в соответствии с ТЗ3, не будет воспроизведена, но некоторые его идеи будут использованы.

При написании модели, которая, с одной стороны, описывает переменные, характеризующие бактериальную культуру (концентрацию субстрата и плотность биомассы в колбе), а с другой – должна описывать среднюю внутриклеточную концентрацию АТФ, необходимо согласовать скорости процессов. Если обозначить объем колбы за V_c, а суммарный объем бактериальных клеток за V_b, то между скоростями процессов, выраженными в концентрациях за единицу времени, – v_c и v_b соответственно, вследствие закона сохранения должно выполняться следующее соотношение: $v_c \cdot V_c = v_b \cdot V_b$, где правая и левая части равенства описывают скорость изменения массы реагента. Отсюда следует, что скорости внутриклеточных процессов должны превышать (концентрационные) скорости тех же процессов в V_c/V_b раз, и мы будем иметь систему с различными характерными временами изменения переменных. Обозначим отношение V_b/V_c как малый параметр ε_0.

С учетом вышесказанного «экологическая часть» модели может быть записана в следующем виде:

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = [f_G(S,a) - M_N(a)]N, \\ \varepsilon_0 \dot{a} = f_E(S,a) \cdot \dfrac{N}{\varepsilon_1 + N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2 + a}, \end{cases} \quad (2)$$

где S – концентрация питательного субстрата; N – биомасса бактерий; a – усредненная внутриклеточная концентрация АТФ в клетках бактериальной культуры.

При этом функция $f_G(S,a) = \dfrac{V_G S}{K_G + S} \cdot \dfrac{a}{K_{aG} + a}$ описывает АТФ-зависимый синтез биомассы, функция $f_E(S,a) = \dfrac{V_E S}{K_E + S} \cdot \dfrac{a}{K_{aE} + a^2}$ – производство АТФ, выражение $\dfrac{k_d a}{\varepsilon_2 + a}$ описывает активность обобщенной АТФ-азы, а функция $M_N(a) = \dfrac{m}{1 + A_N a}$ – интенсивность отмирания бактерий, зависящую от внутриклеточной концентрации АТФ.

Как можно видеть, в данной модели обобщенные активности анаболитных и катаболитных путей описываются отдельными функциями, поэтому вводить специально так называемый экономический коэффициент нет необходимости, более того, соотношение скоростей синтеза биомассы и окисления органики может меняться в ходе развития культуры. Вид функции $f_E(S,a)$, точнее ее часть, описывающая зависимость активности синтеза АТФ от ее концентрации, выбран в соответствии с моделью Селькова (Иваницкий и др., 1978). Последнее слагаемое в уравнении, описывающем концентрацию АТФ, представляет вклад обобщенной АТФ-азы, т. е. совокупность всех базовых процессов в клетке. При малых значениях коэффициента ε_2 активность АТФ-азы будет слабо меняться в широком диапазоне концентраций АТФ, и только при низких значениях будет наблюдаться падение АТФ-азной активности, что представляется естественным.

Наличие малого параметра в третьем уравнении делает концентрацию АТФ быстрой переменной и позволяет исследовать свойства этого уравнения отдельно от остальных переменных, полагая остальные (экологические) переменные константами (Романовский и др., 1984). Мы не будем делать полный анализ устойчивости этого уравнения вследствие его громоздкости, нам достаточно проверить возможность существования устойчивого квазистационарного состояния этого уравнения данной динамической системы и оценить зависимость его устойчивости от значений экологических переменных.

Из рис. 1 можно видеть, что в зависимости от набора параметров уравнение системы может иметь: *А*) одно устойчивое нулевое стационарное состояние либо три стационарных состояния в зависимости от концентрации субстрата *S*; *Б*) одно неустойчивое нулевое и одно устойчивое стационарные состояния при любых значениях концентрации *S*. Поскольку на данном этапе мы не озабочены точным соответствием параметров модели энергетической системы клетки реальным данным, то, следуя подходу Селькова и заявленному ТЗ, выберем вариант, с одной стороны, обеспечивающий клетке стабильное удовлетворение ее энергетических потребностей, а с другой – делающий это наиболее простым способом. Этому требованию удовлетворяет набор параметров, порождающий зависимости, представленные на рис. 1, *Б*.

При определенных значениях параметров существует интервал изменения концентрации АТФ, в котором скорость синтеза АТФ положительна, что приводит к росту ее концентрации до тех пор, пока концентрация не попадет в область с отрицательным значением скорости, что и

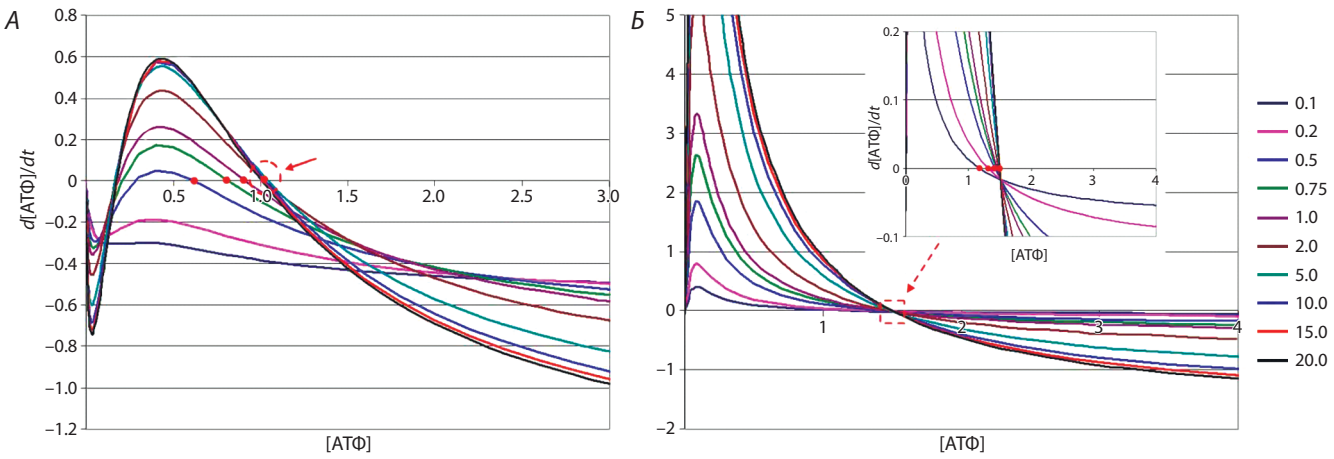


Рис. 1. Зависимость скорости изменения концентрации АТФ от ее концентрации при различных концентрациях субстрата (показаны справа), при различных значениях параметра.

Случай *А*: в системе при $S > S_{min}$ могут существовать три стационарных состояния, из которых одно неустойчивое, а одно соответствует нулевой концентрации АТФ. Штриховой овал выделяет группу малоразличимых стационарных состояний при различных значениях *S*. Случай *Б*: в системе существуют одно устойчивое и одно неустойчивое нулевое стационарные состояния. Красными кружками показаны устойчивые стационарные состояния при различных концентрациях субстрата. Соответствующие наборы параметров для случая *А*: $V_g = 1.22$, $K_g = 1.94$, $K_a = 0.01$, $V_e = 2$, $K_e = 1$, $K_{ae} = 0.2$, $k_d = 0.5$, $\varepsilon_2 = 0.05$; для случая *Б*: $V_g = 2.18$, $K_g = 4$, $K_a = 0.004$, $V_e = 3.299$, $K_e = 4$, $K_{ae} = 0.008$, $k_d = 0.026$, $\varepsilon_2 = 0.85$.

обеспечивает существование устойчивого стационарного состояния (см. рис. 1).

Обеспечив, условно говоря, жизнедеятельность клетки, можно перейти к построению модели кворум-эффекта. Рассмотрим модель КЭ (Williams et al., 2008), которая впоследствии была использована в ряде работ других авторов (Melke et al., 2010; Djezzar et al., 2019). Согласно этой модели, аутоиндуктор *AHL* (*A*) и рецептор *LuxR* (*R*) образуют димеризованный комплекс, который регулирует выработку как *R*, так и *A*. Кроме того, существует ненулевой, базовый, независимый от концентрации индуктора синтез *LuxR*. Модель имеет следующий вид:

$$\begin{cases} \dot{R} = C_R + \dfrac{V_R D}{K_R + D} - k_3 R - k_1 R A + k_2 C, \\ \dot{C} = k_1 R A - k_2 C - 2k_4 C^2 + 2k_5 D, \\ \dot{D} = k_4 C^2 - k_5 D. \end{cases} \quad (3)$$

В этой системе первое уравнение описывает скорость изменения концентрации *LuxR*, которая положительно зависит от суммы базовой (C_R) и автоиндуцированной скоростей синтеза, причем последняя пропорциональна вероятности инициации транскрипции, контролируемой связыванием комплекса $(LuxR\text{-}A)_2$ (*D*) с соответствующим сайтом связывания в регуляторной последовательности оперона. Второе и третье уравнения описывают образование комплекса *LuxR-A* (*C*) с последующим образованием димерного комплекса $(LuxR\text{-}A)_2$ (*D*).

Следуя (Williams et al., 2008) и ТЗЗ, будем предполагать существование квазистационарного состояния для переменных *C* и *D*. Тогда уравнение, описывающее поведение *LuxR* при концентрации аутоиндуктора, рассматриваемой как внешний параметр, имеет вид:

$$\dot{R} = C_R + \frac{V_R \gamma R^2 A^2}{K_R + \gamma R^2 A^2} - k_3 R, \quad (4)$$

где $\gamma = \dfrac{k_4 k_1^2}{k_5 k_2^2}$.

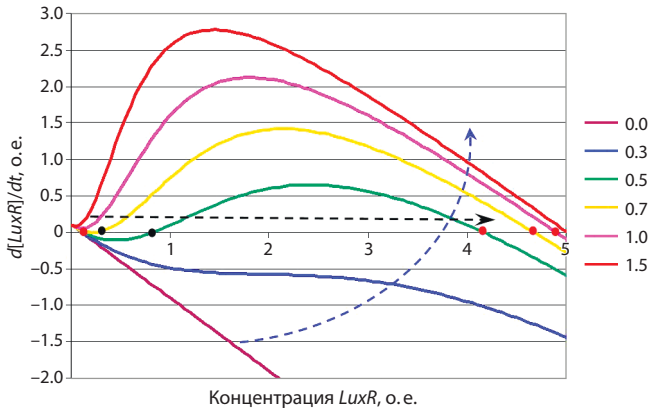


Рис. 2. Зависимость скорости изменения концентрации *LuxR* от его концентрации при различных концентрациях аутоиндуктора, показанных справа.

Красными кружками показаны устойчивые стационарные состояния при различных концентрациях аутоиндуктора, черными – неустойчивые состояния. Изогнутая штриховая стрелка указывает направление изменения концентрации аутоиндуктора, прямая – направление переключения в новое состояние.

Для анализа свойств этого уравнения можно применить прием, использованный для третьего уравнения системы (2), т.е. рассмотреть его в координатах (*R*, *dR/dt*) при разных концентрациях аутоиндуктора, что является несложным делом (рис. 2). Из рисунка видно, что при нулевой и низких концентрациях субстрата может существовать только одно устойчивое стационарное состояние, соответствующее низкой концентрации *LuxR*. По мере повышения концентрации аутоиндуктора появляются еще два стационарных состояния – устойчивое и неустойчивое, однако система произвольно не может перейти в состояние с высоким уровнем экспрессии *LuxR*. При дальнейшем росте концентрации аутоиндуктора левое колено кривой выходит из отрицательной полуплоскости,

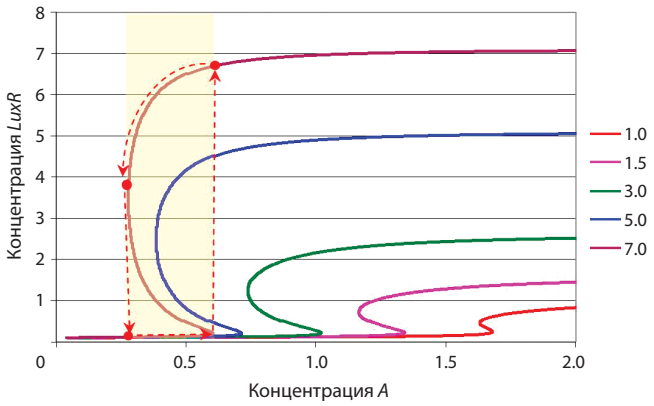


Рис. 3. Стационарные кривые, показывающие зависимость стационарной концентрации *LuxR* от концентрации аутоиндуктора при $\alpha = 0.1$, $\sigma = 1$ и разных значениях параметра β (справа).

что приводит к исчезновению неустойчивого и устойчивого состояний, и система быстро переходит в состояние с высокой концентрацией *LuxR*.

Более наглядно процесс переключения можно показать, если предположить, что реализуется квазистационарное состояние системы, описываемой уравнением (4). В этом случае можно либо применить построение графиков неявно заданных функций в системах компьютерной алгебры типа Maxima, либо, приравняв правую часть к 0, получить выражение для явной функции

$$A = \frac{1}{R}\sqrt{\frac{\sigma(R-\alpha)}{(\alpha+\beta)-R}}, \qquad (5)$$

где $\sigma = K_R\gamma$; $\alpha = C_R/k_3$; $\beta = V_R/k_3$. При этом должно выполняться условие $\alpha < R < \alpha+\beta$.

Для наглядности можно протабулировать (5) как обычную функцию в Excel, а потом перевернуть координаты – сделать (A, R) (рис. 3). Наглядно видно, что при превышении некоторой пороговой концентрации $A(\beta)$ происходит резкий переход в состояние высокого уровня экспрессии *LuxR*, причем в системе наблюдается гистерезис, который в природных условиях может наблюдаться при угнетении роста бактерий и постепенном разрушении аутоиндуктора.

После обкатки модели КЭ и ориентировочной оценки значений параметров, которые необходимы для реализации КЭ, вернемся к построению модели. Несколько модифицируем рассмотренную выше известную модель, чтобы обеспечить ее концептуальное единство, а именно сделаем интенсивный синтез *LuxR* энергозависимым. При этом фоновый синтез аутоиндуктора и *LuxR* оставим условно энергонезависимым, считая, что расходы на их синтез входят в активность обобщенной АТФ-азы (2):

$$\begin{cases} \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R\gamma R^2A^2}{K_R+\gamma R^2A^2}\cdot\dfrac{a}{\varepsilon_3+a} - k_3R. \end{cases} \qquad (6)$$

Забегая вперед, можно сказать, что использование более сложного уравнения, предполагающего, что одновременно с синтезом *LuxR* интенсифицируется синтез аутоиндуктора, как это сделано в модели (Melke et al., 2010), оказалось не обязательным, чтобы описать экспериментальные данные. Кроме того, для простоты предполагается, что концентрация аутоиндуктора в среде и в клетке совпадает, что позволяет обойтись без выделения малого параметра. В итоге наша модель, объединяющая экологические и внутриклеточные молекулярные процессы, выглядит следующим образом:

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = [f_G(S,a) - M_N(a)]N, \\ \varepsilon_0\dot{a} = f_E(S,a)\cdot\dfrac{N}{\varepsilon_1+N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2+a}, \\ \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R\gamma R^2A^2}{K_R+\gamma R^2A^2}\cdot\dfrac{a}{\varepsilon_3+a} - k_3R. \end{cases} \qquad (7)$$

Займемся конструированием завершающего люминесцентного блока модели. Во-первых, предположим, что синтез люциферазы идет параллельно с синтезом *LuxR* и тоже является энергозависимым. Кроме того, учтем энергонезависимый процесс инактивации люциферазы. Однако в данном эксперименте мы регистрируем не количество люциферазы в культуре, а интенсивность люминесценции. Как сказано выше, для обеспечения свечения от клетки должны поступать НАДН и АТФ. Учитывать эти потоки раздельно возможно, но вряд ли имеет смысл, так как от наличия НАДН зависит активность цитохромной цепи, производящей АТФ. Поскольку эти процессы тесно связаны и наличие АТФ означает наличие НАДН, будем в модели (следуя ТЗЗ) рассматривать зависимость свечения только от АТФ. В результате получаем общую модель рассматриваемой системы, где наблюдаемый показатель – свет, описывается функцией *Light*(t):

$$\begin{cases} \dot{S} = -[f_G(S,a) + f_E(S,a)]N, \\ \dot{N} = f_G(S,a)N, \\ \varepsilon_0\dot{a} = f_E(S,a)\cdot\dfrac{N}{\varepsilon_1+N} - f_G(S,a) - \dfrac{k_d a}{\varepsilon_2+a}, \\ \dot{A} = C_A - k_0 A, \\ \dot{R} = C_R + \dfrac{V_R\gamma R^2A^2}{K_R+\gamma R^2A^2}\cdot\dfrac{a}{\varepsilon_3+a} - k_3R, \\ \dot{L} = \dfrac{V_L R}{K_L+R}\cdot\dfrac{a}{\varepsilon_3+a} - k_{dL}L, \end{cases} \qquad (8)$$

$$Light(t) = L(t)\cdot\frac{a(t)}{\varepsilon_4+a(t)}.$$

Отличие экологической части этой модели от (2) заключается в том, что поскольку в эксперименте рассматривается накопительная культура от инокуляции до логарифмической фазы роста включительно, без рассмотрения стационарной фазы и фазы отмирания, то смертность бактерий можно в данном эксперименте не учитывать.

Математическая модель была реализована в среде открытого ПО SciLab 6.1. Для определения параметров математической модели по экспериментальным данным использовался метод Нелдера–Мида, код которого присутствует среди сопровождающих примеров программного обеспечения.

Результаты

Проведенные пробные эксперименты на рекомендованной по прописям богатой (10 г/л пептона) и бедной (0.1 г/л пептона) средах показали наличие кворум-эффекта в обоих случаях. Кривые динамики биомассы и свечения приведены на рис. 4.

Уже рассматривая полученные кривые, без всякой модели, можно видеть (см. рис. 4, *а*), что до начала КЭ (в течение 6 часов культивирования) наблюдается постепенное снижение интенсивности люминесценции, осуществляемой люциферазой, принесенной с инокулятом. После достижения максимума свечения (~11 часов) происходит резкий спад интенсивности свечения. Почти очевидно, что такой спад не может быть связан с инактивацией люциферазы, что потребовало бы предположить существование специальной системы, разрушающей люциферазу сразу после синтеза, да еще в условиях энергетического голода. По-видимому, именно падение концентраций НАДН и АТФ на заключительной стадии логарифмической фазы роста культуры и обусловило это падение свечения. Тогда как медленное падение интенсивности свечения, проходившее в условиях избытка субстрата и интенсивной работы энергетического метаболизма, демонстрирует процесс инактивации люциферазы, точнее всего комплекса ферментов, обслуживающих свечение бактерий.

В то же время динамика свечения культуры в условиях бедной среды (см. рис. 4, *б*) ставит интересные вопросы. Видно, что за 7 часов культивирования биомасса бактерий достигла примерно трети от биомассы, достигнутой бактериями за это же время в богатой среде. При этом темпы роста культуры были хоть и не очень большими, но примерно постоянными на всем рассмотренном периоде, чего нельзя сказать о другом эксперименте. Очевидное ускорение роста культуры в богатой среде после 4-часового роста может указывать на субстратное угнетение при данных концентрациях субстрата.

Интересно, что при этом на бедной среде КЭ начался на 2 часа раньше, чем на богатой. Не исключено, что так влияет субстратное ингибирование, но этот вопрос требует дальнейшего исследования и большего количества экспериментального материала. На текущем этапе исследований наша задача состоит в разработке адекватной модели, удовлетворяющей заявленному в начале статьи ТЗ, и предварительной проверке адекватности этой модели на имеющихся экспериментальных данных.

Результаты вычислительного моделирования приведены на рис. 5 и 6. Подстройка параметров модели проходила в два этапа. Сначала подстраивалась экологическая часть, описывающая динамику биомассы бактериальной культуры, концентрации субстрата и средней по культуре внутриклеточной концентрации АТФ. Результаты приведены на трех верхних графиках представленных рисунков. Надо отметить достаточно хорошее соответствие между модельной кривой, описывающей динамику биомассы, и экспериментальными точками. Как показали дополнительные расчеты, ни модель Ферхюльста, ни введение в модель фактора субстратного ингибирования не дают улучшенного описания. Возможны два варианта: либо наблюдаемое расхождение имеет статистическую природу, либо в системе действует механизм, ускоряющий рост

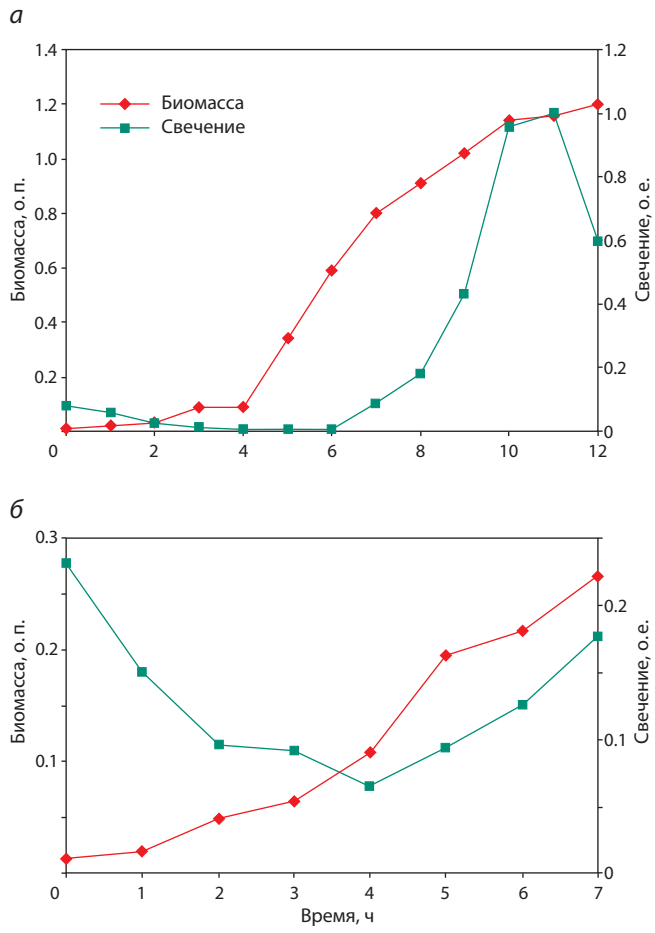


Рис. 4. Динамика роста биомассы и люминесценции культуры *Photobacterium phosphoreum* 1889 на стандартной (*а*) и бедной (*б*) среде.

после достижения некоторого порога. Для дальнейшего анализа потребуются дополнительные эксперименты, которые планируются.

Следует отметить ожидаемое поведение концентрации АТФ, которая, как видно из рис. 1, должна претерпевать незначительные изменения при варьировании концентрации субстрата в определенном интервале и достаточно резко меняться при выходе из этого интервала.

На втором этапе подстраивалась часть модели, описывающая КЭ и люминесценцию, причем в качестве опорных данных использовались уже данные по динамике люминесценции. При этом «эколого-энергетические» параметры модели не менялись.

Рисунок 5 демонстрирует модельную динамику аутоиндуктора и *LuxR*, а также динамику количества люциферазы, которая повторяет динамику экспрессии *LuxR*. Важно отметить, что модель хорошо описывает медленное падение свечения на начальном этапе развития культуры и быстрое ее уменьшение на завершающей стадии, отличающееся по темпам от уменьшения количества люциферазы, что отражает энергетическое состояние клеток.

В случае моделирования поведения культуры на бедной среде (см. рис. 6) отметим следующий момент. Модель, содержащая большое количество подстраиваемых параметров, способна описывать разнообразные варианты ди-

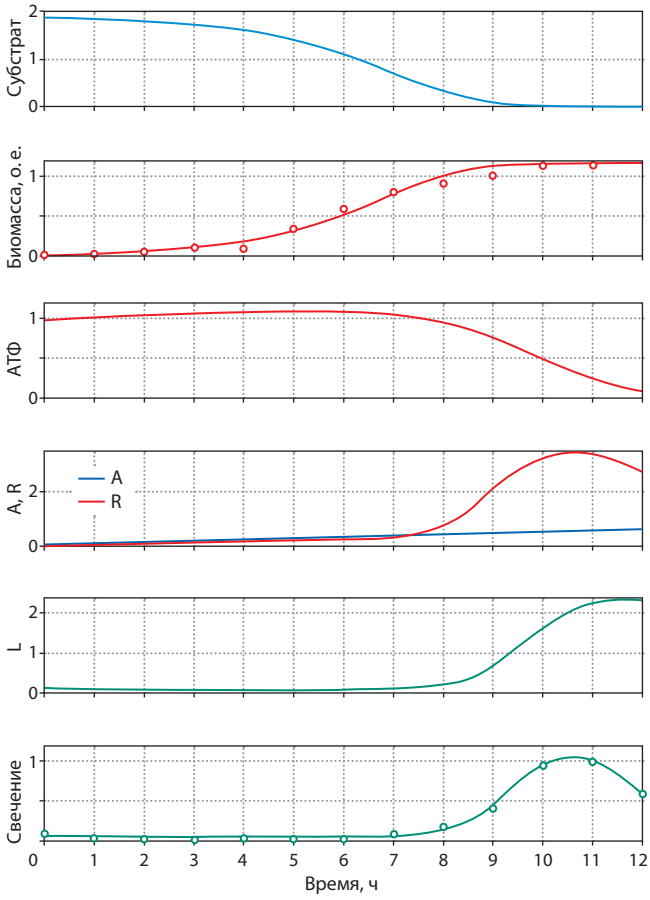


Рис. 5. Модельная и реальная динамика переменных в культуре люминесцентных бактерий.

Кружки – экспериментальные данные.

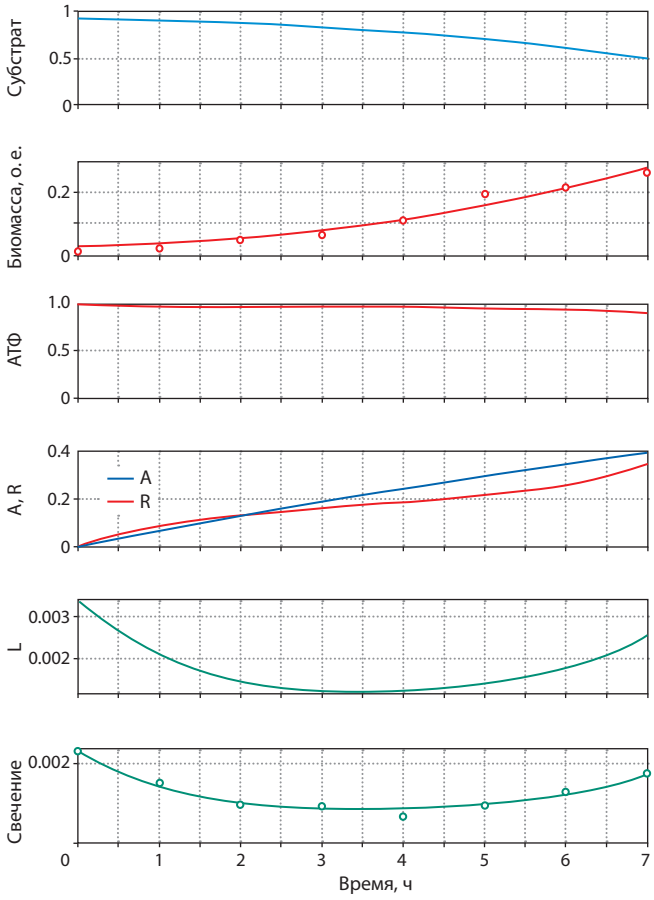


Рис. 6. Модельная и реальная динамика переменных в культуре люминесцентных бактерий на бедной среде.

Кружки – экспериментальные данные.

намики, и вопрос состоит в том, насколько эти параметры соответствуют биологическим представлениям об исследуемой системе. Модельные кривые хорошо соответствуют экспериментальным данным (см. рис. 6). Сопоставим в разделе «Обсуждение» изменения в константах, которые произвела система подстройки параметров при описании роста культуры на бедной среде. При этом общие для двух случаев значения параметров модели следующие: $V_g = 2.18$, $K_g = 3.99$, $K_a = 0.0033$, $V_e = 3.30$, $K_e = 4.02$, $K_{ae} = 0.008$, $a_0 = 1.40$, $k_d = 0.0315$, $k_0 = 0.082$, $V_R = 1.50$, $C_A = 0.14$, $C_R = 0.011$, $k_3 = 0.057$, $\gamma = 0.331$, $K_R = 0.06$, $K_L = 0.17$, $\varepsilon_0 = 0.01$, $\varepsilon_1 = 0.001$, $\varepsilon_2 = 1.54$, $\varepsilon_3 = 0.39$, $\varepsilon_4 = 3.34$.

Обсуждение

Чтобы получить хорошее описание динамики культуры в обоих экспериментах, потребовалось изменить не очень большое количество параметров (см. таблицу). Отметим, что изменение S_0 является ожидаемым, другое дело, что почти двукратное уменьшение S_0 в модели плохо согласуется со стократным уменьшением концентрации пептона в среде. Это расхождение можно предварительно объяснить тем, что, по-видимому, пептон не является ведущим субстратом и рост лимитируют биогены, содержащиеся в дрожжевом экстракте, концентрация которого в этих экспериментах не изменялась.

Сопоставление параметров модели
для двух видов питательных сред

Среда	S_0	k_{dL}	V_L
Богатая	1.84	0.75	9.02
Бедная	1.06	0.41	0.15

В отношении изменений двух других параметров возникают вопросы. Столь существенное (60 раз!) падение константы V_L можно объяснить только наличием некоторой дополнительной системы контроля люминесцентной реакции через пути синтеза ФМН · H_2 или алифатического альдегида. В этом случае обобщенное описание вклада энергетического метаболизма только через АТФ является слишком сильным упрощением.

Почти двукратное уменьшение константы k_{dL} при росте на бедной среде тоже трудно объяснить. Гипотезы по этому поводу строить преждевременно, к вопросу можно вернуться после получения дополнительных экспериментальных данных.

Отмеченные расхождения между ожиданиями и результатами обработки экспериментальных данных вместе с

предположениями о причинах этих расхождений задают направление дальнейших экспериментальных и теоретических исследований механизмов кворум-эффекта в культуре люминесцентных бактерий.

Заключение

Результаты сопоставления модели, построенной в рамках представленной логики, и экспериментальных данных показывают, что предложенная модель в целом удовлетворяет условному техническому заданию, которое было сформулировано во введении. Действительно, модель вполне удовлетворительно описывает динамику биомассы бактерий в накопительной культуре и хорошо описывает динамику свечения бактериальной культуры, которая регулируется кворум-эффектом.

А вот третьему требованию ТЗ о максимальной простоте модели трудно дать окончательную оценку. С одной стороны, не исключено, что данная модель может быть упрощена, чтобы описывать поведение культур бактерий в условиях, приближенных к условиям рассмотренных экспериментов. С другой стороны, работа с моделью (подбор параметров) сформировала ощущение, что данная модель недостаточно робастна (недостаточно груба) в отношении вариации параметров. Это проявлялось, в частности, в том, что метод Нелдера–Мида, как и любой метод локального поиска, достаточно часто находит ближайший минимум целевой функции, который соответствует значениям параметров, имеющих отдаленное отношение к биологическому смыслу, например устремление константы Моно к 0. Не исключено, что модель, в которой смысловые блоки (экологический, энергетический, кворумный, люминесцентный) более артикулированы, более автономны в русле идей Е.Е. Селькова, будет по своей устойчивости к внешним и внутренним возмущениям больше похожа на живое существо.

Список литературы / References

Барцев С.И., Барцева О.Д. Эвристические нейросетевые модели в биофизике: приложение к проблеме структурно-функционального соответствия. Красноярск: Сиб. федер. ун-т, 2010
[Bartsev S.I., Bartseva O.D. Heuristic Neural Network Models in Biophysics: Application to the problem of structure–function mapping. Krasnoyarsk: Siberian Federal University Publ., 2010 (in Russian)]

Горбань А.Н., Охонин В.А., Садовский М.Г., Хлебопрос Р.Г. Простейшее уравнение математической экологии. Препр. ИЛиД СО АН СССР, 1982
[Gorban A.N., Okhonin V.A., Sadovskiy M.G., Khlebopros R.G. The simplest equation of mathematical ecology. Preprint of the Sukachev Forest and Timber Institute, Siberian Branch of the USSR Academy of Sciences, 1982 (in Russian)]

Жаботинский А.М. Концентрационные колебания. М.: Наука, 1974
[Jabotinsky A.M. Concentration Oscillations. Moscow: Nauka Publ., 1974 (in Russian)]

Иваницкий Г.Р., Кринский В.И., Сельков Е.Е. Математическая биофизика клетки. М.: Наука, 1978
[Ivanitsky G.R., Krinsky V.I., Selkov E.E. Mathematical Biophysics of the Cell. Moscow: Nauka Publ., 1978 (in Russian)]

Романовский Ю.М., Степанова Н.В., Чернавский Д.С. Математическая биофизика. М.: Наука, 1984
[Romanovsky Yu.M., Stepanova N.V., Chernavsky D.S. Mathematical Biophysics. Moscow: Nauka Publ., 1984 (in Russian)]

Anstey M.L., Rogers S.M., Ott S.R., Burrows M., Simpson S.J. Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. *Science*. 2009;323(5914):627-630. DOI 10.1126/science.1165939

Brodl E., Winkler A., Macheroux P. Molecular mechanisms of bacterial bioluminescence. *Comput. Struct. Biotechnol. J.* 2018;16:551-564. DOI 10.1016/j.csbj.2018.11.003

Djezzar N., Pérez I.F., Djedi N., Duthen Y. A computational multiagent model of bioluminescent bacteria for the emergence of self-sustainable and self-maintaining artificial wireless networks. *Informatica.* 2019;43(3):395-408. DOI 10.31449/inf.v43i3.2381

Edmonds B. Syntactic Measures of Complexity. Doctoral Thesis. Manchester, UK: Univ. of Manchester, 1999.

Makris N.C., Ratilal P., Jagannathan S., Gong Z., Andrews M., Bertsatos I., Godø O.R., Nero R.W., Jech J.M. Critical population density triggers rapid formation of vast oceanic fish shoals. *Science*. 2009;323(5922):1734-1737. DOI 10.1126/science.1169441

Melke P., Sahlin P., Levchenko A., Jönsson H. A cell-based model for quorum sensing in heterogeneous bacterial colonies. *PLoS Comput. Biol*. 2010;6(6):e1000819. DOI 10.1371/journal.pcbi.1000819

Miller M.B., Bassler B.L. Quorum sensing in bacteria. *Annu. Rev. Microbiol.* 2001;55(1):165-199. DOI 10.1146/annurev.micro.55.1.165

Nealson K.H., Hastings J.W. Bacterial bioluminescence: its control and ecological significance. *Microbiol. Rev.* 1979;43(4):496-518. DOI 10.1128/mr.43.4.496-518.1979

Nealson K.H., Platt T., Hastings J.W. Cellular control of the synthesis and activity of the bacterial luminescent system. *J. Bacteriol.* 1970; 104(1):313-322. DOI 10.1128/jb.104.1.313-322.1970

Williams J.W., Cui X., Levchenko A., Stevens A.M. Robust and sensitive control of a quorum-sensing circuit by two interlocked feedback loops. *Mol. Syst. Biol.* 2008;4:234. DOI 10.1038/msb.2008.70

ORCID ID
S.I. Bartsev orcid.org/0000-0003-0140-4894

Благодарности. Исследование выполнено в рамках государственного задания Министерства науки и высшего образования РФ, проект № 0287-2021-0018.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 18.07.2023. После доработки 16.09.2023. Принята к публикации 16.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):878-883
DOI 10.18699/VJGB-23-101

Перевод на английский язык https://vavilov.elpub.ru/jour

Математическая модель системы жизнеобеспечения на основе водорослей, замкнутая по кислороду и углекислому газу

Д.А. Семёнов ✉, А.Г. Дегерменджи

Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
✉ semenov@ibp.ru

Аннотация. Целью исследования было сравнить методы количественного анализа, применявшиеся на ранних этапах создания прототипов замкнутых систем, с современными подходами анализа данных. В качестве примера рассмотрена математическая модель устойчивого сосуществования двух микроводорослей в смешанной проточной культуре, предложенная Болсуновским и Дегерменджи в 1982 г. Модель построена на основе детального теоретического описания взаимодействия видов и субстрата (в данном случае освещенности). Возможность управления соотношением видов позволяет регулировать ассимиляционный коэффициент (AQ), т.е. отношение поглощенного углекислого газа к выделенному кислороду. Задача управления ассимиляционным коэффициентом системы жизнеобеспечения до сих пор актуальна, микроводоросли рассматриваются как перспективные генераторы кислорода и в современных работах. При этом акцент в них сделан на эмпирических методах моделирования, в частности на анализе больших данных; также работы не выходят за пределы задачи управления монокультурой микроводорослей. В настоящем исследовании мы обращаем внимание на три результата, по нашему мнению, удачно дополняющих современные методы. Во-первых, модель позволяет использовать результаты экспериментов с монокультурами, во-вторых, предсказывает преобразование данных к виду, удобному для дальнейшего анализа, в том числе для вычисления AQ. В-третьих, модель позволяет гарантировать устойчивость полученного приближения и в дальнейшем искать решение как малую поправку эмпирическими методами.
Ключевые слова: система жизнеобеспечения (СЖО); математическая модель; смешанная культура двух водорослей.

Для цитирования: Семёнов Д.А., Дегерменджи А.Г. Математическая модель системы жизнеобеспечения на основе водорослей, замкнутая по кислороду и углекислому газу. *Вавиловский журнал генетики и селекции*. 2023;27(7):878-883. DOI 10.18699/VJGB-23-101

Alga-based mathematical model of a life support system closed in oxygen and carbon dioxide

D.A. Semyonov ✉, A.G. Degermendzhi

Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
✉ semenov@ibp.ru

Abstract. The purpose of the study was to compare quantitative analysis methods used in the early stages of closed-loop system prototyping with modern data analysis approaches. As an example, a mathematical model of the stable coexistence of two microalgae in a mixed flow culture, proposed by Bolsunovsky and Degermendzhi in 1982, is considered. The model is built on the basis of a detailed theoretical description of the interaction between species and substrate (in this case, illumination). The ability to control the species ratio allows you to adjust the assimilation quotient (AQ), that is, the ratio of carbon dioxide absorbed to oxygen released. The problem of controlling the assimilation coefficient of a life support system is still relevant; in modern works, microalgae are considered as promising oxygen generators. At the same time, modern works place emphasis on empirical modeling methods, in particular, on the analysis of big data, and the work does not go beyond the task of managing a monoculture of microalgae. In our work, we pay attention to three results that, in our opinion, successfully complement modern methods. Firstly, the model allows the use of results from experiments with monocultures. Secondly, the model predicts the transformation of data into a form convenient for further analysis, including for calculating AQ. Thirdly, the model allows us to guarantee the stability of the resulting approximation and further refine the solution by small corrections using empirical methods.
Key words: life support system (LSS); mathematical model; mixed culture of two algae.

For citation: Semyonov D.A., Degermendzhi A.G. Alga-based mathematical model of a life support system closed in oxygen and carbon dioxide. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):878-883. DOI 10.18699/VJGB-23-101

Введение

Сейчас сложные системы преимущественно рассматриваются как «черный ящик», генерирующий большой объем данных. Развитию соответствующих методов анализа способствовало значительное повышение доступности методов регистрации данных и снижение стоимости вычислительных мощностей. При проектировании замкнутых систем жизнеобеспечения данные продолжают быть малочисленными и дорогими. Теоретические подходы, основанные на детальном описании компонентов сложных систем, могут предсказать удобные подходы к предварительной обработке данных. Математические модели, стремящиеся описать сложную систему минимально сложным образом, преобразуют массив экспериментальных данных к удобному не только для анализа, но и для восприятия человеком-оператором виду. Кроме того, математические модели помогают решать актуальные до сих пор задачи. Эти положения мы иллюстрируем на примере управления ассимиляционным коэффициентом (AQ) смешанной культуры двух водорослей.

Может ли нас научить чему-нибудь ранний опыт создания прототипов замкнутых систем жизнеобеспечения? История конструирования замкнутых систем жизнеобеспечения (ЗСЖО) насчитывает уже более полувека. В связи с возрождением интереса к проектированию баз на Луне и Марсе в последнее десятилетие актуальность этого направления заметно возросла (Keller et al., 2021; Liu et al., 2021). Поскольку на начальных этапах создавались и подробно изучались некоторые прототипы, в дальнейшем отклоненные по разным причинам, то возникает желание изучить опыт этих работ для возможного применения в современных проектах. В современных публикациях настойчиво предлагаются универсальные шаги по проектированию отдельных модулей системы жизнеобеспечения (СЖО) и ее апробации (Heinicke, Verseux, 2023; Metelli et al., 2023). Могут ли прежние подходы оказаться полезны для новых проектов? Возникает также соблазн сравнить использовавшиеся тогда методы с распространенными сейчас, в частности с методами анализа больших данных.

Подобный, на предварительном этапе мысленный эксперимент удобно провести для системы, имеющей достаточно детальное теоретическое описание. В нашем случае это система совместного культивирования двух водорослей (*Chlorella vulgaris* и *Spirulina platensis*), используемая в качестве генератора кислорода для систем жизнеобеспечения. Идея применения водорослей для создания систем жизнеобеспечения до сих пор актуальна (Häder, 2020; Fahrion et al., 2021; Matula et al., 2021; Keller et al., 2023). В частности, *Chlorella vulgaris* и *Spirulina platensis* рассматриваются как перспективные виды для этой задачи (Helisch et al., 2020; Cycil et al., 2021; Matula, Nabity, 2021; Matula et al., 2021). Мы не можем уверенно утверждать, что все авторы названных работ искренне убеждены в будущей роли микроводорослей в СЖО. По нашему мнению, более перспективны для решения задачи обеспечения человека кислородом и пищей высшие растения. Однако мы, как, возможно, и многие из перечисленных авторов, считаем микроводоросли удачным учебным пособием. Благодаря ряду преимуществ культивирование микроводорослей является хорошим модельным объектом. Например, в литературе можно встретить работы, посвященные управлению монокультурами микроводорослей (Hu et al., 2008, 2012, 2014), где демонстрируется эффективность различных методов управления. То есть теоретическая работа по управлению культивированием микроводорослей в серии трех статей носит методический характер. Мы видим возможность дополнить эту серию статей, обратившись к анализу модели сорокалетней давности. В рамках работ по созданию замкнутых систем жизнеобеспечения в 1982 г. была предложена модель управления смешанной проточной культурой двух водорослей (Болсуновский, Дегерменджи, 1982).

Использование водорослей в качестве единственных автотрофов в системе жизнеобеспечения позволяет применить удобное упрощение для рассуждений о стехиометрии восстановления кислорода и связывания углекислого газа в водорослевом культиваторе. В первом приближении можно считать, что весь углекислый газ выделяется человеческим организмом в реакциях окисления углеводов и жиров. Это предположение основано на том, что использование человеческим организмом аминокислот в качестве значимого источника энергии возможно при несбалансированном рационе, чрезмерных физических нагрузках или при некоторых хронических заболеваниях. Исключив три эти возможности, будем считать, что аминокислоты вносят незначительный вклад в дыхание. Углеводы и жиры – основные источники энергии человеческого организма и основные продукты биосинтеза водорослей.

Следующим удобным упрощением может быть игнорирование синтеза аминокислот водорослями. К сожалению, состав биомассы обеих водорослей указывает на то, что белки присутствуют в больших количествах. Тем не менее можно допустить первое приближение, за которым должна последовать корректировка модели при необходимости замыкания азотного обмена. То есть сколько человек окислил углеводов и жиров, столько углеводов и жиров должны синтезировать водоросли для связывания избыточного углекислого газа и регенерации использованного человеком кислорода. Использование высших растений не позволило бы прибегнуть к такому простому первому приближению, так как кроме углеводов, жиров и белков в состав высших растений в заметных количествах входит лигнин, значительно отличающийся по стехиометрии как от углеводов, так и от жиров.

В зависимости от рациона и уровня физических нагрузок организм человека может использовать разные субстраты для получения энергии. При достаточном доступе кислорода основным источником энергии является окисление жирных кислот в митохондриях. При нехватке кислорода организм человека предпочитает углеводы в качестве основного источника энергии. Таким образом, соотношение выделяемого человеком углекислого газа и поглощаемого кислорода может варьировать почти от 0.7 (окисление жиров) до 1.0 (окисление углеводов). Для человека возможно даже кратковременное превышение респираторного индекса 1.0 в результате интенсивных физических нагрузок (ацидоз с потерей бикарбонатов) и даже длительное превышение при условии углеводного питания и увеличения массы тела при накоплении жиров. В отличие от человека, водоросли в среднем за цикл своей

жизни придерживаются относительного постоянства состава. Поскольку в анализируемой проточной культуре не наблюдалось синхронизации и колебаний численности, то можно для каждого из двух видов водорослей использовать усредненные значения ассимиляционных индексов.

Ассимиляционные индексы отражают стехиометрическую пропорцию, в которой связываемые молекулы углекислого газа относятся к вырабатываемым молекулам кислорода. Так как мы договорились в первом приближении описывать весь метаболизм балансом жиров и углеводов, то оставим за рамками этой статьи исследование возможности сместить ассимиляционный индекс водорослей вариациями азотного питания (Белянин и др., 1980). Примем ситуацию с азотным питанием стабильной и предположим, что ассимиляционный индекс системы из двух водорослей может меняться в пределах, указанных в литературе. Метаболическое постоянство автотрофов и метаболическая пластичность человека должны как-то согласовываться в рамках работы ЗСЖО. Диапазон возможного суммарного ассимиляционного индекса двух водорослей ограничивает рацион и метаболическую активность человека, поселенного в ЗСЖО. Важным допущением будем считать, что мы можем придерживаться в среднем указанного диапазона, рационально управляя диетой и физической активностью человека. Тогда, например, в зависимости от долговременного повышения уровня физических нагрузок респираторный коэффициент человека может сместиться, что потребует смещения ассимиляционного коэффициента СЖО. Конструкция СЖО должна предполагать возможность подстраиваться под потребности метаболизма человека. В анализируемой модели нас будет интересовать возможность управления составом смешанной культуры водорослей и управления суммарным ассимиляционным коэффициентом.

Материалы и методы

Оценка ассимиляционных индексов смешанной культуры двух водорослей. Для того чтобы более детально представлять себе процессы газообмена в исследуемой системе, воспользуемся брутто-формулами биомассы хлореллы ($C_{6.0}H_{9.7}O_{2.635}N_{0.937}$) и спирулины ($C_{6.0}H_{10.84}O_{2.06}N_{0.87}$) (Белянин и др., 1980). Так как система на первом этапе считается не замкнутой по азоту, то можно упростить формулы, считая, что основной формой усвоения азота водорослями является мочевина или ионы аммония, а также убрав из формул кислород в виде воды. Находим остаток в виде ($C_{6.0}H_{1.6}$) для хлореллы и ($C_{6.0}H_{4.11}$) для спирулины. Откуда следует, что синтез биомассы хлореллы и спирулины позволяет на один поглощенный литр углекислого газа выделить 1.13 и 1.3425 л кислорода соответственно. Что отвечает ассимиляционным коэффициентам AQ = 0.885 для хлореллы и AQ = 0.745 для спирулины.

Ассимиляционный индекс смешанной культуры может быть легко получен из массовых соотношений водорослей в культуре:

$$AQ = X \cdot 0.885 + (1 - X) \cdot 0.745,$$

где X – доля спирулины в культуре. Так, для первоначально полученной устойчивой смешанной культуры $X = 0.6$ и $AQ = 0.6 \cdot 0.885 + 0.4 \cdot 0.745 = 0.829$. Управление составом смешанной культуры дает возможность варьировать значение AQ в диапазоне от 0.745 (монокультура спирулины) до 0.885 (монокультура хлореллы).

Математическая модель. Для того чтобы прогнозировать стационарное состояние популяций водорослей в проточном культиваторе, необходима математическая модель, обобщающая информацию о влиянии управляющих факторов на систему из двух видов. Именно такая модель проточного культиватора с двумя водорослями была построена в работе (Болсуновский, Дегерменджи, 1982). Модель описывает сосуществование двух видов, конкурирующих за лимитирующий субстрат. Лимитирующим субстратом в данном случае является световой поток. В модели есть область параметров освещенности, в которой устойчиво сосуществуют два вида; кроме этого, существуют области доминирования для каждого вида, когда конкурирующий вид вытесняется. Разумеется, есть и диапазон параметров, не позволяющий воспроизводиться ни одному из видов, – они просто вымываются из культиватора при заданном протоке и недостаточной освещенности. Проток вещества в культиваторе стабилизировался через регистрацию поглощения хлорофилла на длине волны 680 нм, т. е. система поддерживала постоянную оптическую плотность среды. Управлять системой можно регулируя скорость протока (т. е. оптическую плотность среды в культиваторе) и интенсивность освещенности. Модель не учитывает фотоингибирование роста спирулины при высокой интенсивности освещенности, а также эффекты метаболического ингибирования при высокой плотности популяций. Математическая часть модели получена в результате количественного описания экспериментов (Белянин, Болсуновский, 1980) при помощи дифференциальных уравнений с применением последующей процедуры линеаризации (Болсуновский, Дегерменджи, 1982).

Модель представляет собой систему из двух дифференциальных уравнений, каждое из которых отражает динамику численности одной водоросли. Уравнения имеют вид:

$$X'_1 = (\mu_1 - D_п)X_1;\ \mu_1 = a_1 E/(b_1 + E),$$
$$X'_2 = (\mu_2 - D_п)X_1;\ \mu_2 = a_2 E/(b_2 + E),$$
$$E = E_0(1 - \gamma_1 X_1 - \gamma_1 X_2),$$
$$D_п = \mu_1 X_1 + \mu_2 X_2.$$

Здесь E – средняя освещенность, учитывающая поглощение света культурами водорослей, получена после разложения в ряд Тейлора и отбрасывания нелинейных слагаемых, учитывая низкую оптическую плотность смешанной культуры; $D_п$ – скорость протока, в дальнейшем анализе заменяемая оптической плотностью культуры, как экспериментально измерявшейся величины. Уравнения отражают конкуренцию за свет как за субстрат. Этот субстрат, как известно из экспериментальных данных по монокультурам, усваивается согласно уравнению Михаэлиса–Ментен. Кривые удельного роста в монокультурах демонстрируют, что спирулина эффективнее усваивает свет при низкой освещенности, а хлорелла – при высокой освещенности (рис. 1).

В областях параметров, характерных для устойчивой совместной культуры двух водорослей (низкая плотность

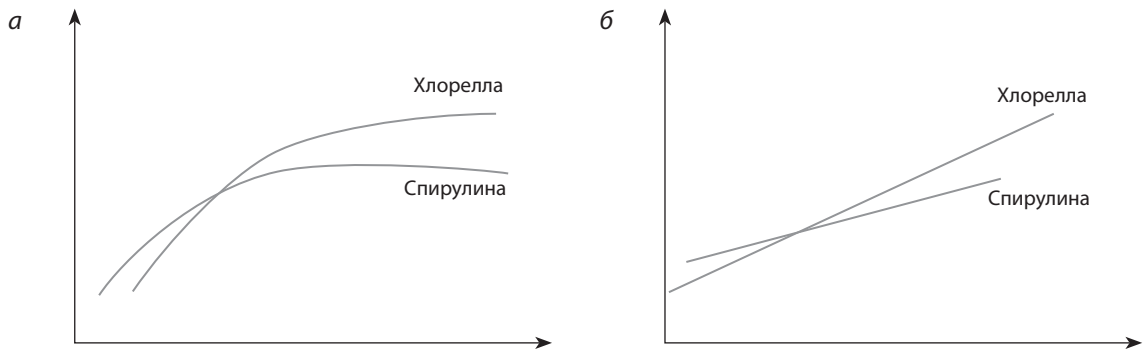


Рис. 1. Удельная скорость роста в зависимости от освещенности монокультур хлореллы и спирулины (*а*). Представление тех же данных в обратных координатах (*б*) демонстрирует хорошее согласование с уравнением Михаэлиса–Ментен.

популяции и малый световой поток), модель должна давать наименьшее расхождение с экспериментальными данными. Для изменения соотношения видов в культиваторе в этих условиях достаточно небольшого по величине изменения режима освещенности или соответственного изменения протока. Долговременное увеличение и уменьшение потребности в кислороде в ЗСЖО может быть компенсировано соответствующим масштабированием культиватора.

Результаты

Первое впечатление – культура двух практически не взаимодействующих видов при конкуренции за единственный общий субстрат должна приводить к устойчивому состоянию, когда доминирует один вид, а другой вытесняется. Оказывается, понять, почему сосуществование возникает, можно при внимательном анализе взаимодействия видов с субстратом в монокультуре. Хлорелла не только лучше себя чувствует при высокой освещенности, но и создает некоторое преимущество для спирулины в смешанной культуре по сравнению с монокультурой. То есть спирулина в присутствии хлореллы может существовать в области более высокой освещенности. Хлорелла «затеняет» спирулину, создавая ей более комфортные условия. Более детальный анализ биологии этих видов позволил выявить приспособления к высокой и низкой освещенности, а также адаптацию к разным диапазонам спектра (Болсуновский, Дегерменджи, 1982). Но даже без учета этой приспособленности к разным частям спектра и на материале экспериментов с монокультурами удается получить нетривиальную динамику в модели смешанной культуры. Математическая модель помогает перейти от качественного объяснения к количественным предсказаниям.

Модель позволяет получить область устойчивого сосуществования двух видов в непрерывной культуре. Графически область представлена на плоскости в координатах освещенность (E_0) и оптическая плотность культуры на длине волны 680 нм (C), отражающая скорость протока в культиваторе (рис. 2).

Необходимо обратить внимание, что экстраполяция результатов модели в область высокой освещенности и высокой плотности культуры нежелательна, так как в этой области экспериментально показано действие факторов, не учтенных при моделировании (Белянин и др., 1980).

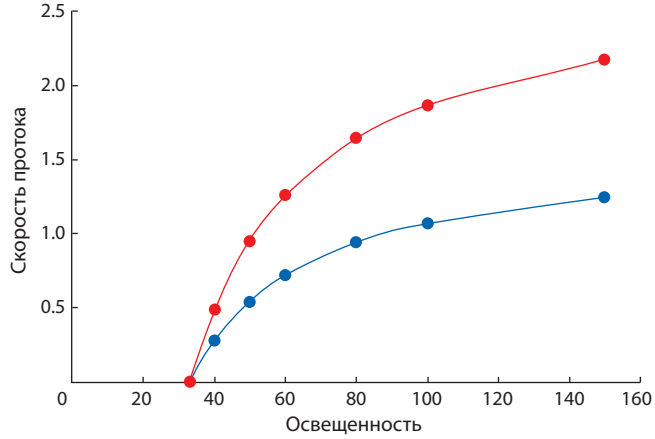


Рис. 2. Область существования устойчивой культуры двух водорослей ограничена двумя кривыми на плоскости Освещенность/Скорость протока.

На верхней границе (красная кривая) смешанная кривая превращается в монокультуру хлореллы, на нижней (синяя кривая) – в монокультуру спирулины.

Модель позволяет вычислить стационарные концентрации компонентов, т. е. плотности численности отдельных видов:

$$\alpha_1 X_1 = 2K_1K_2(E/E_0 - 1 + C/2K_2)/(K_1 - K_2),$$
$$\alpha_2 X_2 = 2K_1K_2(-E/E_0 + 1 - C/2K_1)/(K_1 - K_2).$$

Для задачи управления газовым составом в ЗСЖО важно определить, где выполняется соотношение $X_1/X_2 = \text{const}$. Математическая модель была призвана качественно объяснить наблюдающееся явление, а именно устойчивое сосуществование двух видов. Ожидать точного предсказания положений равновесия во всей области существования системы не приходится, но модель может дать хорошее первое приближение для решения этой задачи на практике.

Такой приближенный алгоритм поиска равновесного состояния системы послужит «ручкой грубой настройки». Более точный подбор параметров может быть осуществлен экспериментальным путем.

Для того чтобы понять, как модель может использоваться для анализа экспериментальных данных, вообразим, что данные есть, а теоретических представлений о том,

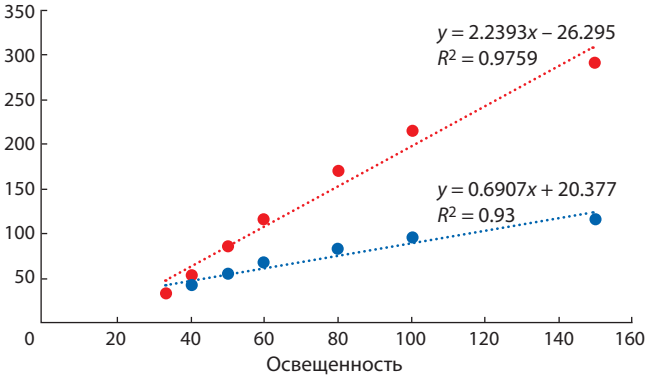


Рис. 3. Результат эмпирического подбора преобразования, «спрямляющего» графики данных в новых координатах.

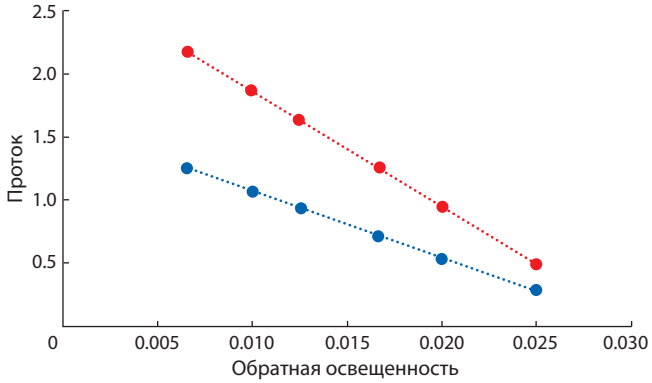


Рис. 4. Как следует из модели, для поиска положений равновесия в смешанной культуре удобно представить данные в координатах Обратная освещенность/Проток.

как функционирует система, нет. Прагматичным будет подход, состоящий в поиске преобразования кривых, ограничивающих область существования, в прямые в новых координатах. Тогда все прямые на этой плоскости, проходящие через точку пересечения и лежащие в области существования смешанной культуры, можно принять за $X_1/X_2 = \text{const}$. Например, для данного типа кривых аппроксимацией может быть преобразование вида

$$C(E) = K \cdot \ln(E) - \text{const},$$

где K и const подбирались бы методом наименьших квадратов.

Результаты обратного преобразования графиков $E = \exp(C/K + \text{const})$ показаны на рис. 3. Можно отметить, что после преобразования точки хорошо аппроксимируются прямой линией.

Все возможные устойчивые положения равновесия системы, допускающие сосуществование двух видов, преобразуются в пучок прямых, проходящих через одну точку. Для каждой подобной прямой можно принять AQ = const. Так как AQ получен простой суперпозицией ассимиляционных индексов двух водорослей, то естественно предположить, что на плоскости, где данные о монокультурах представлены прямыми линиями, данные о смешанной культуре тоже будут представлены прямыми линиями.

Следует обратить внимание на два наглядных факта: 1) выбранная аппроксимация чувствительна к тому, в какой области набраны экспериментальные данные; 2) аппроксимация дает систематическую ошибку, занижая результаты при средней освещенности и завышая в области низкой и высокой освещенности.

Теперь сравним этот подход с тем, который вытекает из знания точного решения модели. Точную аппроксимацию модельного решения даст преобразование к координатам $(1/E_0; C)$, тогда точные решения преобразуются в прямые (рис. 4). Все точки, подчиняющиеся соотношениям $X_1/X_2 = \text{const}$, также будут лежать на прямых, проходящих через общую точку пересечения. Именно такую аппроксимацию можно рекомендовать для дальнейшего применения при обработке экспериментальных данных в качестве первого приближения.

Представим ситуацию, когда у нас есть экспериментальные данные, полученные в современных условиях. Допустим, в культиваторе установилось стационарное состояние. В эксперименте можно контролировать скорость протока и освещенность. Можно при помощи газоанализа получить значение AQ для стационарного случая, а затем вычислить соотношение видов в культуре. Можно представить и непосредственное измерение соотношения видов. Используя современные методы, например проточную цитометрию, можно в автоматическом режиме получать данные об установившемся соотношении X_1/X_2. Все эти данные можно использовать для восстановления параметров калибровочных графиков вида $X_1/X_2 = \text{const}$. То есть теория помогает выбрать процедуру предобработки данных для дальнейшего анализа, например, методами математической статистики, или искусственными нейронными сетями, или даже в виде графических построений. Более того, теория получена преимущественно исходя из данных об удельной скорости роста водорослей в монокультурах. Опираясь на данные о монокультурах, эмпирические методы просто не могут предсказать соотношения в смешанной культуре, поэтому для эмпирических методов, к которым относятся все современные методы анализа больших данных, понадобятся не только большие, но еще и достаточно труднодоступные данные.

Как теперь определить положение прямой с заданным соотношением X_1/X_2? Нижний график – это оптическая плотность монокультуры спирулины, верхний – оптическая плотность монокультуры хлореллы. Для того чтобы на заданном уровне освещенности найти положение с заданным соотношением X_1/X_2, надо поделить вертикальный отрезок, соединяющий нижнюю и верхнюю прямую, в соотношении X_1/X_2. Устойчивость решения математической модели гарантирует, что последующее экспериментальное уточнение положения равновесия будет небольшим. Эмпирические методы в настоящее время не дают представления об устойчивости полученных с их помощью прогнозов.

Заключение

При создании сложных биотехнологических систем достаточно простые и наглядные математические модели могут быть хорошим дополнением современных методов анализа данных. В случае труднодоступности эксперимен-

тальных данных единственным способом получить прогноз поведения системы становится создание адекватной математической модели. Кроме того, в случае замкнутых систем жизнеобеспечения немаловажна возможность понимания устройства системы со стороны человека-оператора, как правило, обитателя этой системы. Чем более простые и наглядные механизмы будут заложены в конструкцию системы жизнеобеспечения, тем выше будет ее надежность.

Список литературы / References

Белянин В.Н., Болсуновский А.Я. Регулирование видового состава двукомпонентного сообщества водорослей в эксперименте. В: Параметрическое управление биосинтезом микроводорослей. Новосибирск: Наука, 1980;72-80
[Belyanin V.N., Bolsunovskiy A.Ya. Regulation of species range in a two-component algae community in an experiment. In: Parametric Control of Microalgal Biosynthesis. Novosibirsk: Nauka Publ., 1980;72-80 (in Russian)]

Белянин В.Н., Сидько Ф.Я., Тринкеншу А.П. Энергетика фотосинтезирующей культуры растений. Новосибирск: Наука, 1980
[Belyanin V.N., Sydko F.Ya., Trinkenschu A.P. Energetics of Photosynthesizing Plant Culture. Novosibirsk: Nauka Publ., 1980 (in Russian)]

Болсуновский А.Я., Дегерменджи А.Г. Изучение фотосинтетического механизма сосуществования видов в смешанной проточной культуре «хлорелла-спирулина». В: Вопросы управления биосинтезом низших растений. Новосибирск: Наука, 1982;99-116
[Bolsunovskiy A.Ya., Degermendzhi A.G. Study of the photosynthetic mechanism of coexistence of species in a mixed continuous-flow chlorella-spirulina culture. In: Issues of Controlling the Biosynthesis in Lower Plants. Novosibirsk: Nauka Publ., 1982;99-116 (in Russian)]

Cycil L.M., Hausrath E.M., Ming D.W., Adcock C.T., Raymond J., Remias D., Ruemmele W.P. Investigating the growth of algae under low atmospheric pressures for potential food and oxygen production on Mars. *Front. Microbiol*. 2021;12:733244. DOI 10.3389/fmicb.2021.733244

Fahrion J., Mastroleo F., Dussap C.-G., Leys N. Use of photobioreactors in regenerative life support systems for human space exploration. *Front. Microbiol*. 2021;12:699525. DOI 10.3389/fmicb.2021.699525

Häder D. On the way to Mars-flagellated algae in bioregenerative life support systems under microgravity conditions. *Front. Plant Sci*. 2020;10:1621. DOI 10.3389/fpls.2019.01621

Heinicke C., Verseux C. The MaMBA facility as a testbed for bioregenerative life support systems. *Life Sci. Space Res.* (*Amst*.). 2023;36:86-89. DOI 10.1016/j.lssr.2022.08.009

Helisch H., Keppler J., Detrell G., Belz S., Ewald R., Fasoulas S., Heyer A.G. High density long-term cultivation of *Chlorella vulgaris* SAG 211-12 in a novel microgravity-capable membrane raceway photobioreactor for future bioregenerative life support in SPACE. *Life Sci. Space Res.* (*Amst.*). 2020;24:91-107. DOI 10.1016/j.lssr.2019.08.001

Hu D., Liu H., Yang C., Hu E. The design and optimization for light-algae bioreactor controller based on Artificial Neural Network-Model Predictive Control. *Acta Astronaut*. 2008;63(7-10):1067-1075. DOI 10.1016/j.actaastro.2008.02.008

Hu D., Li M., Zhou R., Sun Y. Design and optimization of photo bioreactor for O_2 regulation and control by system dynamics and computer simulation. *Bioresour. Technol*. 2012;104:608-615. DOI 10.1016/j.biortech.2011.11.049

Hu D., Li L., Li Y., Li M., Zhang H., Zhao M. Gas equilibrium regulation by closed-loop photo bioreactor built on system dynamics, fuzzy inference system and computer simulation. *Comput. Electron. Agric*. 2014;103:114-121. DOI 10.1016/j.compag.2014.02.002

Keller R.J., Porter W., Goli K., Rosenthal R., Butler N., Jones J.A. Biologically-based and physiochemical life support and in situ resource utilization for exploration of the Solar System – reviewing the current state and defining future development needs. *Life*. 2021;11(8):844. DOI 10.3390/life11080844

Keller R., Goli K., Porter W., Alrabaa A., Jones J.A. Cyanobacteria and algal-based biological life support system (BLSS) and planetary surface atmospheric revitalizing bioreactor brief concept review. *Life*. 2023;13(3):816. DOI 10.3390/life13030816

Liu H., Yao Z., Fu Y., Feng J. Review of research into bioregenerative life support system(s) which can support humans living in space. *Life Sci. Space Res.* (*Amst.*). 2021;31:113-120. DOI 10.1016/j.lssr.2021.09.003

Matula E.E., Nabity J.A. Effects of stepwise changes in dissolved carbon dioxide concentrations on metabolic activity in *Chlorella* for spaceflight applications. *Life Sci. Space Res.* (*Amst.*). 2021;29:73-84. DOI 10.1016/j.lssr.2021.03.005

Matula E.E., Nabity J.A., McKnight D.M. Supporting simultaneous air revitalization and thermal control in a crewed habitat with temperate *Chlorella vulgaris* and eurythermic antarctic chlorophyta. *Front. Microbiol*. 2021;12:709746. DOI 10.3389/fmicb.2021.709746

Metelli G., Lampazzi E., Pagliarello R., Garegnani M., Nardi L., Calvitti M., Gugliermetti L., Restivo Alessi R., Benvenuto E., Desiderio A. Design of a modular controlled unit for the study of bioprocesses: eowards solutions for Bioregenerative Life Support Systems in space. *Life Sci. Space Res.* (*Amst.*). 2023;36:8-17. DOI 10.1016/j.lssr.2022.10.006

ORCID ID
D.A. Semyonov orcid.org/0000-0002-4993-6358
A.G. Degermendzhi orcid.org/0000-0001-8649-5419

Благодарности. Исследование выполнено за счет гранта Российского научного фонда № 23-44-00059, https://rscf.ru/project/23-44-00059/

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 20.07.2023. После доработки 18.09.2023. Принята к публикации 19.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):884-889
DOI 10.18699/VJGB-23-102

A phenomenological model of non-genomic variability of luminescent bacterial cells

S.I. Bartsev[1, 2]

[1] Institute of Biophysics of the Siberian Branch of the Russian Academy of Sciences, Federal Research Center "Krasnoyarsk Science Center SB RAS", Krasnoyarsk, Russia
[2] Siberian Federal University, Krasnoyarsk, Russia
bartsev@yandex.ru

Abstract. The light emitted by a luminescent bacterium serves as a unique native channel of information regarding the intracellular processes within the individual cell. In the presence of highly sensitive equipment, it is possible to obtain the distribution of bacterial culture cells by the intensity of light emission, which correlates with the amount of luciferase in the cells. When growing on rich media, the luminescence intensity of individual cells of brightly luminous strains of the luminescent bacteria *Photobacterium leiognathi* and *Ph. phosporeum* reaches 10^4–10^5 quanta/s. The signal of such intensity can be registered using sensitive photometric equipment. All experiments were carried out with bacterial clones (genetically homogeneous populations). A typical dynamics of luminous bacterial cells distributions with respect to intensity of light emission at various stages of batch culture growth in a liquid medium was obtained. To describe experimental distributions, a phenomenological model that links the light of a bacterial cell with the history of events at the molecular level was constructed. The proposed phenomenological model with a minimum number of fitting parameters (1.5) provides a satisfactory description of the complex process of formation of cell distributions by luminescence intensity at different stages of bacterial culture growth. This may be an indication that the structure of the model describes some essential processes of the real system. Since in the process of division all cells go through the stage of release of all regulatory molecules from the DNA molecule, the resulting distributions can be attributed not only to luciferase, but also to other proteins of constitutive (and not only) synthesis.
Key words: non-genomic variability; phenomenological model; luminescent bacteria.

For citation: Bartsev S.I. A phenomenological model of non-genomic variability of luminescent bacterial cells. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):884-889. DOI 10.18699/VJGB-23-102

Феноменологическая модель негеномной изменчивости люминесцентных бактериальных клеток

С.И. Барцев[1, 2]

[1] Институт биофизики Сибирского отделения Российской академии наук, Федеральный исследовательский центр «Красноярский научный центр СО РАН», Красноярск, Россия
[2] Сибирский федеральный университет, Красноярск, Россия
bartsev@yandex.ru

Аннотация. Свет, испускаемый люминесцентными бактериями, может служить уникальным природным каналом передачи информации о процессах внутри отдельной клетки. При наличии высокочувствительного оборудования можно получить распределение клеток бактериальной культуры по интенсивности свечения, которая коррелирует с количеством люциферазы в клетках. При выращивании на богатых питательных средах интенсивность свечения отдельных клеток ярко светящихся штаммов люминесцентных бактерий *Photobacterium leiognathi* и *Ph. phosporeum* достигает 10^4–10^5 квантов/с. Сигнал такой интенсивности может быть зарегистрирован с помощью чувствительного фотометрического оборудования. Все эксперименты проводились с бактериальными клонами – генетически однородными популяциями. Получена типичная динамика распределения светящихся бактериальных клеток по интенсивности свечения на различных стадиях периодического выращивания культуры в жидкой среде. Для описания экспериментальных распределений была построена феноменологическая модель, которая связывает излучение бактериальной клетки с историей событий на молекулярном уровне. Предложенная феноменологическая модель с минимальным числом подстроечных параметров (1.5) обеспечивает удовлетворительное описание сложного процесса формирования распределения клеток по интенсивности свечения на разных стадиях роста бактериальной культуры. Это может свидетельствовать о том, что структура модели описывает некоторые существенные процессы реальной системы. Поскольку в процессе деления все клетки проходят стадию отсоединения всех регуляторных молекул от молекулы ДНК, результирующие распределения можно отнести не только к люциферазе, но и к другим белкам конститутивного (и не только) синтеза.
Ключевые слова: негеномная изменчивость; феноменологическая модель; люминесцентные бактерии.

Introduction

The heterogeneity of isogenic bacterial populations, or, in other words, non-genomic variability of cells, is increasingly attracting the attention of researchers. This is partly due to the development of methods for tracking individual cell parameters, down to the dynamics of protein synthesis during the cell cycle (Taheri-Araghi et al., 2015; Andryukov et al., 2021). On the other hand, understanding the mechanisms or causes of phenotypic differences of cells from an isogenic population is important both for the formation of fundamental concepts of intracellular processes organization and for increasing the efficiency of solving practical problems in medicine and biotechnology.

The cell cycle is a potentially significant source of non-genomic variability. During the cell cycle, the protein abundance in the cell undergoes two-fold changes. In the case of an asynchronous population, these changes can contribute significantly to phenotypic variability. However, another possible source of heterogeneity is related to the cell cycle. It has been shown quite a long time ago (Shkolnik, 1989) that the widely used allometric dependences (when different variables N_i are related by relations of the form $N_i = \alpha_i N_1^{\beta_i}$), when describing growth curves, lead to a contradiction with observations. So in the case of an allometric growth model, a cell dies after a small number of generations due to the fact that certain substances abundance approaches zero. Then a phenomenological trigger model combining allometric growth with switches was proposed. According to the model, the passage of a cell through various phases of the cell cycle is accompanied by sharp changes in the allometric ratios of growth variables. There are certain combinations of parameters that can be conditionally associated with multidimensional switching surfaces – the boundaries of cellular phases – from cell birth to division. When passing the next boundary, the rates of change in cellular variables switch. This model was further developed (Zinovyev et al., 2022) and demonstrated strong agreement with experimental data.

According to this model, switching should occur in a certain sequence and in a fairly uniform manner, but for a non-synchronous culture such switching can make a significant contribution to the variability of phenotypic traits. However, it should be noted that this model was compared with data on the dynamics of variable eukaryotic cells and it is possible that in bacterial cells the limitations of allometric growth are overcome in another way.

Thus, experimental observations of protein synthesis inside bacterial cells (Kiviet et al., 2014) show that the activation of particular protein synthesis occurs without pronounced patterns. Another paper on the topic (Walker et al., 2016) notes that the contribution of the bacterial cell cycle to expression noise consists of two parts: a deterministic fluctuation synchronous with the cell cycle and a stochastic component caused by variable timing of gene replication. It was shown earlier (Taniguchi et al., 2010) that proteins with strong expression have a coefficient of variation of ~30 %, which indicates an "external" factor not associated with fluctuations in the abundance of a small number of molecules.

Fluorescence microscopy is primarily used to monitor protein synthesis at the single-cell scale, which is essential for studying non-genomic variation. However, it is noted that with the current level of device sensitivity stimulating light has a negative effect on the physiological state of cells (Taheri-Araghi et al., 2015).

A unique alternative to fluorescence microscopy is the use of luminescence of luminescent bacteria (Deryabin, 2009) as a channel of information about the state of intracellular processes (Berzhanskaya et al., 1975; Bartsev, Gitelzon, 1985). The uniqueness of luminescence lies in the fact that the cell emits light while in its native state, which significantly reduces the probability of artifacts. Moreover, since the intensity of cell luminescence depends both on the abundance of luciferase and on the presence of substrates for the luciferase reaction, the luminescence of a bacterium is a kind of multiplexer – information from different input channels can be transmitted through one output channel – about the expression of the luciferase operon, on the one hand, and the state of the cell's energy metabolism, on the other.

The goal of the work is to assess the degree of variability of individual bacterial cells regarding luminescence intensity at different stages of development of batch culture of bacteria, and to test the simplest possible approach to the mathematical description of this variability.

Experiment description

When growing on rich media, the luminescence intensity of individual cells of brightly luminous strains of luminescent bacteria *Photobacterium leiognathi* and *Ph. phosporeum* reaches 10^4–10^5 quanta/s. Such signal can be registered using sensitive photometric equipment. The strains used did not demonstrate the typical quorum effect (Brodl et al., 2018) and an increase in their luminescence was observed from the beginning of culture growth.

Without delving into the details of the experimental setup, which operates in the photon counting mode, and the routine for measuring the distribution of bacterial cells according to luminescence intensity (Bartsev, Shenderov, 1985), let us proceed to the description of the results. It should be noted that all experiments were carried out with bacterial clones (genetically homogeneous populations).

During the registration of distributions, the bacteria were in a medium containing only glucose as an energy substrate, i. e. bacterial growth was stopped and the luciferase abundance during the measurement can be considered unchanged. At least, control experiments showed that over a typical period of time the luminescence intensity of individual bacterial cells did not undergo noticeable changes.

A typical view of luminous bacteria distribution at various stages of batch culture growth in a liquid medium is shown in Figure 1.

An immediate question arises regarding the potential mechanism behind the observed variation in the phenotypic trait. The simplest explanation for the observed variability can be suggested immediately – the intensity of the emission is determined by the variability of the bacterial cell volumes. However, direct measurements of cell volume variation in *B. subtilis* and *E. coli* showed that the coefficient of variation (CV) of cell volume is ~23 % (van Heerden et al., 2017), while the average CV of bacterial luminescence intensity

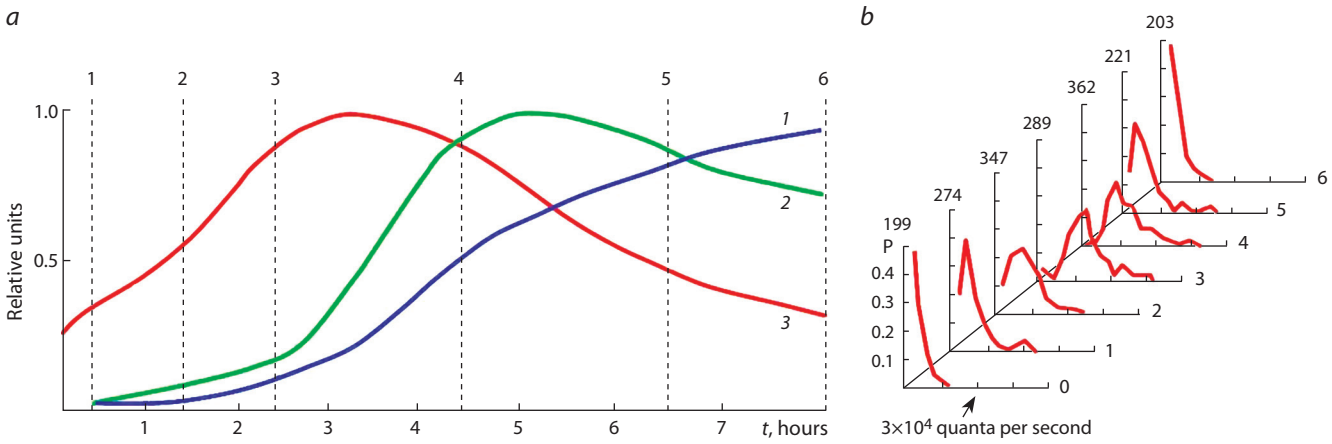


Fig. 1. Dynamics of luminescent bacteria culture parameters (*a*) and cell distributions by luminescence intensity (*b*).

Curves of culture parameters are given in relative units: *1* – optical density; *2* – culture luminescence intensity; *3* – the average intensity of a single cell. The dashed lines indicate sampling times, and their numbers correspond to the numbers of distributions.

is ~50 % and can exceed 70 %. Therefore, there is an additional factor that provides a significant variability in cell luminescence.

On possible causes of non-genomic variability

Under normal growth conditions, the luminescence intensity of a bacterial cell is determined by the abundance of luciferase, the enzyme responsible for catalyzing the luminescent reaction, as well as a set of enzymes that supply the necessary substrates for this reaction (Brodl et al., 2018). Proteins involved in bacterial bioluminescence, notably, LuxCDABEG, are encoded by the lux operon and are highly conserved among different bacterial strains. The *luxA* and *luxB* genes encode a heterodimeric luciferase; the *luxCs*, *luxDs*, and *luxE* gene products are components of the fatty acid reductase complex; and *luxG* encodes flavin reductase.

It is natural to assume that in the presence of an energy substrate, as was the case in the experiments performed, the intensity of bacterial luminescence is determined primarily by the expression of the luciferase operon. Other factors, such as the contribution of uneven distribution of protein, mRNA and ribosomes during division, variability in the amount of mRNA due to the small number of molecules, the transition of genes from active to passive state due to reversible binding of a transcription factor, conformation of the DNA molecule that prevents binding RNA polymerases show less variability (Paulsson, 2004; Schwabe, Bruggeman, 2014; Kuwahara et al., 2015; van Heerden et al., 2017; Dessalles et al., 2020) than observed in the experiment. In addition, the resulting cell distributions by protein amount give a distribution close to normal, while asymmetric distributions were observed in the experiment. In addition to this, these distributions demonstrated characteristic dynamics during the development of the enrichment culture, and an adequate model for the formation of distributions of luminescent bacteria by luminescence intensity should, at least qualitatively, reproduce the experimental dynamics.

With a large number of molecules, which is the case for luciferase, fluctuations in its amount between daughter cells are determined by fluctuations in the uneven volumes of daughter cells, which cannot explain the observed CV value. At the same time, it was shown (Taniguchi et al., 2010) that proteins with strong expression have a coefficient of variation of ~30 %, which indicates an "external" factor not associated with fluctuations in a small number of molecules.

Mathematical model derivation

Without delving into the details of the processes of transcription and translation, let us consider a possible phenomenological stochastic mechanism for generating significant variability in the amount of luciferase in cells. The amount of luciferase in a cell of age $\tau - z(\tau)$ is the sum of the amount of luciferase received by the cell after division (x) and the amount of luciferase accumulated by age $\tau - y(\tau)$:

$$z(\tau) = x + y(\tau). \quad (1)$$

Immediately after division, when $\tau = 0$, the cell contains only the luciferase produced in the previous cell cycle. Let $f(x)$ be the distribution of cells of a narrow age interval according to the amount of luciferase obtained during division, which does not change throughout the entire cell cycle. The form of this distribution is not known and must be obtained by solving the model equation.

Type of cells distribution from a narrow age interval according to the amount of luciferase synthesized and accumulated by age $\tau - P(y, \tau)$ can be obtained from the following considerations. For the sake of simplicity, let's assume that luciferase synthesis begins immediately after cell division, closely associated with the release of DNA from all transcription factors (in our case, the luciferase gene repressor), proceeds at a constant rate, and stops after binding the repressor to the operator.

Let's assume that τ' is the moment when the repressor binds to the operator. Then the amount of luciferase synthesized by time τ is described by the following expression:

$$y(\tau) = \alpha \int_0^{\tau} \theta(\tau' - \eta) d\eta, \quad (2)$$

where α is the rate of enzyme synthesis; θ is the Heaviside step function.

Since $y(\tau)$ is also a function of the random variable τ', distribution $P(y, \tau)$ is described by the following expression:

$$P(y,\tau)=\int_0^{\tau} g(\tau')\delta(y-\alpha\tau')d\tau'+\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')d\tau', \quad (3)$$

where $g(\tau')$ is the distribution describing the proportion of the cell population in which the binding of the repressor to the operator occurred in the interval $[\tau', \tau'+d\tau']$; $\delta(x)$ is the Dirac delta function.

This integral is split into two integrals with integration limits $[0, \tau)$ and $[\tau, \infty)$, and the cells in which the binding of the repressor to the operator occurred by the age τ ($\tau'<\tau$) fall into the first integral, the rest ($\tau'\geq\tau$) fall into in the second. Let's do some calculations:

$$P(y,\tau)=\int_0^{\tau} g(\tau')\,\delta\Big(y-\alpha\int_0^{\tau'}d\eta\Big)\,d\tau'+\int_{\tau}^{\infty} g(\tau')\,\delta\Big(y-\alpha\int_0^{\tau}d\eta\Big)\,d\tau',$$

$$P(y,\tau)=\int_0^{\infty} g(\tau')\,\delta\Big[y-\alpha\int_0^{\tau}\theta(\tau'-\eta)\,d\eta\Big]\,d\tau',$$

$$P(y,\tau)=\frac{1}{\alpha}g\Big(\frac{y}{\alpha}\Big)\theta(\alpha\tau-y)+\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'.$$

Since the total amount of luciferase in a cell ($z(\tau)$) is the sum of independent random variables x and y, then the distribution of cells in a narrow time interval of age τ by the total amount of luciferase has the following form:

$$L(z,\tau)=\int_0^{\infty}\int_0^{\infty} f(x)P(y,\tau)\,\delta(z-x-y)\,dx\,dy,$$

$$L(z,\tau)=\int_0^{\infty} f(z-y)\,P(y,\tau)\,dy,$$

$$L(z,\tau)=\int_0^{\infty} f(z-y)\frac{1}{\alpha}g\Big(\frac{y}{\alpha}\Big)\theta\Big(\tau-\frac{y}{\alpha}\Big)dy+$$

$$+\int_0^{\infty} f(z-y)\,\delta(y-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'\,dy.$$

By changing the variables $\tau' = y/\alpha$ we get:

$$L(z,\tau)=\int_0^{\tau} f(z-\alpha\tau')\,g(\tau')\,d\tau'+f(z-\alpha\tau)\int_{\tau}^{\infty} g(\tau')\,d\tau'. \quad (4)$$

As a result, an expression for the distribution of cells by the amount of luciferase for a narrow age range of age τ was obtained. In order to obtain the equations for the distribution function $f(x)$ and the expression for $\Phi(z)$ – the distribution function of the cell population by the amount of luciferase, it is necessary to know the age structure of the population.

The form of cells distribution by age $\Psi(\tau)$ is obtained from the equation (Romanovsky et al., 1984):

$$\frac{\partial n}{\partial t}+\frac{\partial n}{\partial \tau}=-\omega(\tau)n,$$

where $n(t, \tau)d\tau$ is the number of cells of age in the interval $[\tau, \tau+d\tau]$ at the moment t; $\omega(\tau)$ is the rate of cell loss from a given age interval due to division.

Let us consider the case of a stationary age distribution of bacteria, i.e. $n(t, \tau)/N(t)$, is fixed, but the total number of cells $N(t)$ increases. In the case of a stationary distribution, the specific growth rate of cells number in a given age interval is equal to the specific population growth rate:

$$\frac{\partial n(t,\tau)}{\partial t}=\mu n(t,\tau). \quad (5)$$

Dividing this equation by $N(t)$ we get the equation for frequencies:

$$\frac{\partial \Psi}{\partial \tau}=-[\omega(\tau)+\mu]\Psi, \quad \Psi(\tau)=\frac{n(t,\tau)}{N(t)}.$$

For simplicity, we set the division rate as a step function (Romanovsky et al., 1984, p. 88):

$$\omega(\tau)=C\Theta(\tau-\tau_1)=\begin{cases}0, & \tau<\tau_1\\ C, & \tau\geq\tau_1\end{cases},$$

then the distribution density of dividing cells looks like:

$$\Omega(\tau)=\begin{cases}0 & ,\ \tau<\tau_1\\ Ce^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases},$$

where C is the intensity of cell division events. And as a result:

$$\Psi(\tau)=\begin{cases}\Psi_0 e^{-\mu\tau} & ,\ \tau<\tau_1\\ \Psi_0 e^{-\mu\tau}e^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases}.$$

It remains to determine the form of the function $g(\tau)$. Assumptions about the constant amount of the repressor in the cell and the irreversibility of its binding to the operator allow us to represent the distribution of cells over the time that elapsed from replication (division) to the moment of binding the repressor to the operator in the form of an exponential distribution:

$$g(\tau)=Ae^{-A\tau},$$

where A is the intensity of events.

As a result of all substitutions, we obtain a model for the distribution of luciferase over the cells of the bacterial culture:

$$f\Big(\frac{z}{2}\Big)=2\int_0^{\infty}\Omega(\tau)\,d\tau\left[\int_0^{\tau} f(z-\alpha\tau')\,Ae^{-A\tau'}d\tau'+2f(z-\alpha\tau)e^{-A\tau}\right],$$

$$\Phi(z)=\int_0^{\infty}\Psi(\tau)\,d\tau\left[\int_0^{\tau} f(z-\alpha\tau')\,Ae^{-A\tau'}d\tau'+f(z-\alpha\tau')e^{-A\tau}\right],$$

where

$$\Omega(\tau)=\begin{cases}0 & ,\ \tau<\tau_1,\\ Ce^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases}\quad \Psi(\tau)=\begin{cases}\Psi_0 e^{-\mu\tau} & ,\ \tau<\tau_1\\ \Psi_0 e^{-\mu\tau}e^{-C(\tau-\tau_1)} & ,\ \tau\geq\tau_1\end{cases},$$

and where $f(z)$ is the density of distribution of cells from a narrow age interval according to the amount of luciferase obtained during division; $\Phi(z)$ is the density of cell distribution according to the intracellular amount of luciferase; $\Psi(\tau)$ is distribution density of culture cells by age; $\Omega(\tau)$ is distribution density of dividing cells; A is the intensity of binding the repressor to the operator; α is the rate of luciferase synthesis; C is the intensity of cell division events; τ_1 is the minimum age of the beginning of cell division τ.

Computer simulation

If the resulting equations cannot be solved analytically, then successive approximations are used. But first the values of the model parameters need to be chosen. Note that if the intensity of the repressor binding the activator (parameter A) is equal to zero, then constitutive protein synthesis throughout the entire cell cycle takes place. It is natural to compare this synthesis with the growth of cell volume.

That is, the parameters C and τ_1 can be determined from other independent distributions (van Heerden et al., 2017), assuming that the coefficients of variation of distributions by volume in luminescent bacteria and other gram-negative bacte-

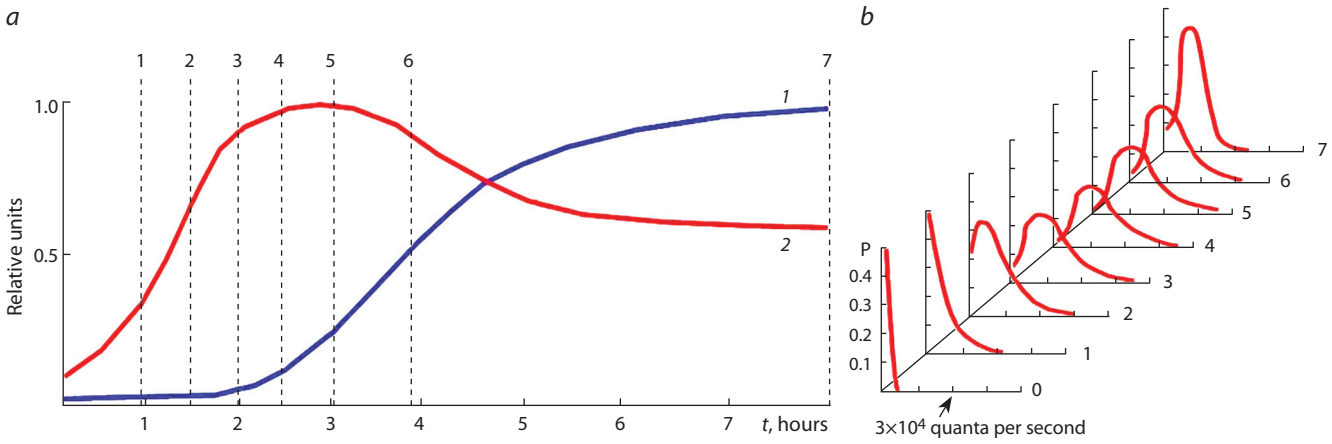


Fig. 2. Model dynamics of luminescent bacteria culture parameters (*a*) and cell distributions by luminescence intensity (*b*).

Curves of culture parameters are given in relative units: *1* – biomass; *2* – the average intensity of a single cell emission. The dashed lines indicate the moments of "sampling", and the numbers correspond to the numbers of the distributions.

ria are close. The coefficient of variation of the model distribution is close to the value of 24 % at $C = 4$ and $\tau_1 = 3/4\ \tau_0$, where τ_0 is the average generation time in the population. These values were used for further simulation. When modeling the dynamics of light intensity distributions during population growth, at the next iteration step the value of the specific growth rate μ was substituted from population growth simulation describing the growth of a real culture.

Thus, as a result, there are only two adjustable parameters, or rather, one and a half – the parameter α (the rate of luciferase synthesis) is, in fact, a scale factor. It shows the relative value of the luminescence intensity, mediated in the experiment by the quantum efficiency of the luciferase itself, the geometry of the recording system that determines the amount of light from a bacterium that hits the photocathode of the photomultiplier, the quantum yield of the photocathode, and the fraction of single-electron pulses cut off by the discriminator at the PMT output.

So to describe the dynamics of distributions obtained in the experiment, the model has one adjustable parameter, A, the intensity of repressor-operator binding events. The results of calculations for the most suitable value for describing real distributions, which is $A = 2$, are shown in Figure 2.

When comparing Figures 2 and 1, one can see a quite satisfactory correspondence between them. It is worth noting that this correspondence was obtained with one fitting parameter, which apparently indicates that the proposed model describes something significant in the simulated real system.

It should be noted that luciferase inactivation was not taken into account when deriving the model, which was done to simplify the model; however, it is a common practice (Schwabe, Bruggeman, 2014, p. 306). Palliative inactivation of luciferase can be introduced externally – simply by shifting the distribution points to 0 in proportion to their distance from the origin. In this case, the visual representation of the model would be closer to the experimental data.

However, one property of the model is of interest, which manifested itself in the shift of distributions to 0 at the last stages of population development. By distribution No. 4, the model has almost reached a stationary state and should have remained in it. But since the model takes into account the increase in the duration of the generation time due to the slowdown in culture growth, the established balance between the rate of luciferase synthesis and its distribution between two daughter cells is disturbed.

Since the rate of synthesis of a particular protein is related to the state of basic metabolism, a slowdown in the cell growth rate and accordingly an increase in the generation time leads to a decrease in the rate of luciferase synthesis (decreasing α coefficient). But the intensity of repressor-operator binding events (a physical, energy-independent process) remains the same. However, on the time scale of the cell itself (the unit of measurement is generation time), the rate of luciferase synthesis remained the same, while the intensity of switching events of the luciferase operon increased. Therefore, according to the model, there is a close relationship between the rate of cell growth and the content of luciferase in it, and the higher the rate, the more luciferase is synthesized per cell cycle and vice versa.

The proposed model based on switching off the operon some time after the birth corresponds to the results on the dependence of fluorescent protein expression on cell age (van Heerden et al., 2017, Fig. 4, B, C). It should be noted that the imposition of the age distribution on the expression level curve (Fig. 4, C) was not done entirely correctly by the authors – they have expression even at negative ages (beyond the left border of the age distribution). When bringing the expression level to the age distribution, it would be even more clearly visible, as can be judged by the saturation of the blue area in Fig. 4, B, that the expression level is maximum immediately after the birth of the cell and then decreases with age, which corresponds to the proposed model.

Conclusion

In conclusion, it can be noted that the proposed phenomenological model with a minimum number of adjustable parameters (1.5) satisfactorily describes a rather complex process that takes place during the growth of a bacterial culture. This may be an indication that the structure of the model describes some essential processes of the real system. Since

in the process of division all cells go through the stage of release of all regulatory molecules from the DNA molecule, the resulting distributions can be realized not only in relation to luciferase, but also to other proteins of constitutive (and not only) synthesis.

References

Andryukov B.G., Timchenko N.F., Lyapun I.N., Bynina M.P., Matosova E.V. Heterogeneity in isogenic bacteria populations and modern technologies of cell phenotyping. *J. Microbiol. Epidemiol. Immunobiol.* 2021;98(1):73-83. DOI 10.36233/0372-9311-33 (in Russian)

Bartsev S.I., Gitelson J.I. On the temporary organization of bacterial luminescence. *Studia Biophisica.* 1985;105(3):149-156 (in Russian)

Bartsev S.I., Shenderov A.N. Dynamics of distributions of luminescent bacteria according to the intensity of luminescence in periodic culture. Krasnoyarsk: Preprint Institute of Physics SB AS USSR, 1985 (in Russian)

Berzhanskaya L.Yu., Gitelson J.I., Fish A.M., Chumakova R.I. On the pulsed nature of bacterial bioluminescence. *Doklady Akademii Nauk SSSR*. 1975;222(5):1220-1222 (in Russian)

Brodl E., Winkler A., Macheroux P. Molecular mechanisms of bacterial bioluminescence. *Comput. Struct. Biotechnol. J.* 2018;16:551-564. DOI 10.1016/j.csbj.2018.11.003

Deryabin D.G. Bacterial Bioluminescence: Fundamental and Applied Aspects. Moscow: Nauka Publ., 2009 (in Russian)

Dessalles R., Fromion V., Robert P. Models of protein production along the cell cycle: an investigation of possible sources of noise. *PLoS One.* 2020;15(1):e0226016. DOI 10.1371/journal.pone.0226016

Kiviet D.J., Nghe P., Walker N., Boulineau S., Sunderlikova V., Tans S.J. Stochasticity of metabolism and growth at the single-cell level. *Nature*. 2014;514(7522):376-379. DOI 10.1038/nature 13582

Kuwahara H., Arold S.T., Gao X. Beyond initiation-limited translational bursting: the effects of burst size distributions on the stability of gene expression. *Integr. Biol.* 2015;7(12):1622-1632. DOI 10.1039/c5ib00107b

Paulsson J. Summing up the noise in gene net works. *Nature*. 2004; 427(6973):415-418. DOI 10.1038/nature02257

Romanovsky Yu.M., Stepanova N.V., Chernavsky D.S. Mathematical Biophysics. Moscow: Nauka Publ., 1984 (in Russian)

Schwabe A., Bruggeman F.J. Contributions of cell growth and biochemical reactions to nongenetic variability of cells. *Biophys. J.* 2014;107(2):301-313. DOI 10.1016/j.bpj.2014.05.004

Shkolnik E.M. Dynamic models of the cell cycle. In: Bykov V.I. (Ed.) Dynamics of Chemical and Biological Systems. Novosibirsk: Nauka Publ., 1989;230-260 (in Russian)

Taheri-Araghi S., Brown S.D., Sauls J.T., McIntosh D.B., Jun S. Single-cell physiology. *Annu. Rev. Biophys.* 2015;44:123-142. DOI 10.1146/annurev-biophys-060414-034236

Taniguchi Y., Choi P.J., Li G.-W., Chen H., Babu M., Hearn J., Emili A., Xie X.S. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science.* 2010;329(5991): 533-538. DOI 10.1126/science.1188308

van Heerden J.H., Kempe H., Doerr A., Maarleveld T., Nordholt N., Bruggeman F.J. Statistics and simulation of growth of single bacterial cells: illustrations with *B. subtilis* and *E. coli*. *Sci. Rep.* 2017; 7(1):16094. DOI 10.1038/s41598-017-15895-4

Walker N., Nghe P., Tans S.J. Generation and filtering of gene expression noise by the bacterial cell cycle. *BMC Biol*. 2016;14:11. DOI 10.1186/s12915-016-0231-z

Zinovyev A., Sadovsky M., Calzone L., Fouché A., Groeneveld C.S., Chervov A., Barillot E., Gorban A.N. Modeling progression of single cell populations through the cell cycle as a sequence of switches. *Front. Mol. Biosci*. 2022;8:793912. DOI 10.3389/fmolb. 2021.793912

ORCID ID

S.I. Bartsev orcid.org/0000-0003-0140-4894

Acknowledgements. The study was funded by State Assignment of the Ministry of Science and Higher Education of the Russian Federation (project No. 0287-2021-0018).

I am grateful to L.Yu. Berzhanskaya for involving me in this work and to V.A. Okhonin and A.N. Shenderov for useful comments and advice in carrying out this work.

Conflict of interest. The author declares no conflict of interest.

Received July 18, 2023. Revised September 16, 2023. Accepted September 18, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):890-897
DOI 10.18699/VJGB-23-103

DyCeModel: a tool for 1D simulation for distribution of plant hormones controlling tissue patterning

D.S. Azarova[1], N.A. Omelyanchuk[1], V.V. Mironova[2], E.V. Zemlyanskaya[1, 3], V.V. Lavrekha[1, 3]

[1] Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud University, Nijmegen, the Netherlands
[3] Novosibirsk State University, Novosibirsk, Russia
vvl@bionet.nsc.ru

Abstract. To study the mechanisms of growth and development, it is necessary to analyze the dynamics of the tissue patterning regulators in time and space and to take into account their effect on the cellular dynamics within a tissue. Plant hormones are the main regulators of the cell dynamics in plant tissues; they form gradients and maxima and control molecular processes in a concentration-dependent manner. Here, we present DyCeModel, a software tool implemented in MATLAB for one-dimensional simulation of tissue with a dynamic cellular ensemble, where changes in hormone (or other active substance) concentration in the cells are described by ordinary differential equations (ODEs). We applied DyCeModel to simulate cell dynamics in plant meristems with different cellular structures and demonstrated that DyCeModel helps to identify the relationships between hormone concentration and cellular behaviors. The tool visualizes the simulation progress and presents a video obtained during the calculation. Importantly, the tool is capable of automatically adjusting the parameters by fitting the distribution of the substance concentrations predicted in the model to experimental data taken from the microscopic images. Noteworthy, DyCeModel makes it possible to build models for distinct types of plant meristems with the same ODEs, recruiting specific input characteristics for each meristem. We demonstrate the tool's efficiency by simulation of the effect of auxin and cytokinin distributions on tissue patterning in two types of *Arabidopsis thaliana* stem cell niches: the root and shoot apical meristems. The resulting models represent a promising framework for further study of the role of hormone-controlled gene regulatory networks in cell dynamics.
Key words: computer modeling; developmental trajectory; input data; genetic algorithm; phytohormones.

For citation: Azarova D.S., Omelyanchuk N.A., Mironova V.V., Zemlyanskaya E.V., Lavrekha V.V. DyCeModel: a tool for 1D simulation for distribution of plant hormones controlling tissue patterning. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):890-897. DOI 10.18699/VJGB-23-103

DyCeModel: программное средство для одномерного моделирования распределения гормонов растений, контролирующих образование структуры ткани

Д.С. Азарова[1], Н.А. Омельянчук[1], В.В. Миронова[2], Е.В. Землянская[1, 3], В.В. Лавреха[1, 3]

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

Аннотация. Для изучения механизмов роста и развития необходимо анализировать динамику распределения регуляторов по ткани во времени и пространстве и учитывать их влияние на клеточную динамику внутри ткани. Растительные гормоны являются основными регуляторами динамики клеток в тканях растений; они образуют градиенты и максимумы и контролируют молекулярные процессы в зависимости от концентрации. Мы представляем DyCeModel, программный инструмент, реализованный в среде MATLAB для одномерного моделирования ткани с динамическим клеточным ансамблем, где изменения концентрации гормона (или другого активного вещества) в клетках описываются обыкновенными дифференциальными уравнениями. Мы применили DyCeModel для моделирования динамики клеток в меристемах растений с различной клеточной структурой и продемонстрировали, что DyCeModel помогает выявить взаимосвязь между концентрацией гормонов и поведением клеток. Инструмент визуализирует ход моделирования и предоставляет видео, полученное в ходе расчета. Важно отметить, что инструмент способен автоматически подбирать параметры, подгоняя распределение концентраций веществ, предсказанное в модели, к экспериментальным данным, полученным по изображениям с микроскопа. Примечательно, что DyCeModel позволяет строить модели для различных типов меристем растений на основе одних и тех же обыкновенных дифференциальных уравнений, используя для каждой меристемы специфические входные характеристики. Эффективность инструмента продемонстриро-

вана путем моделирования влияния распределения ауксина и цитокинина на формирование паттерна ткани в двух типах ниш стволовых клеток *Arabidopsis thaliana*: апикальных меристемах корня и побега. Полученные модели представляют собой перспективный фреймворк для дальнейшего изучения роли контролируемых гормонами генных регуляторных сетей в динамике клеток.
Ключевые слова: компьютерное моделирование; траектория развития; входные данные; генетический алгоритм; фитогормоны.

Introduction

Understanding the control of cell division and differentiation in stem cell niches is among the major issues in plant developmental biology (Hayashi et al., 2023). Although many components of the molecular regulatory networks, which underlie these processes, have been identified, complex interactions and numerous players hinder detailed study on the mechanisms of their functioning. For example, it is still largely unknown how the formation of plant hormone concentration gradients results in particular alterations in the cellular dynamics of developing tissues and organs (Rutten et al., 2022). Dissection of these issues requires application of computer modeling to predict the output in cellular dynamics and to determine whether various developmental pathways exist under certain conditions (Fisher et al., 2023).

Nowadays, developmental biology has recruited experts in mathematical modeling and computer sciences to create appropriate tools. Numerical simulations were successfully used to study the influence of phytohormone concentration distribution on the functioning of plant stem cell niches in 1D and 2D models describing cell divisions, growth, and differentiation under control of signaling molecules (Kitano et al., 2005; Nikolaev et al., 2006; Mironova et al., 2010; Muraro et al., 2013; Band et al., 2014; De Rybel et al., 2014; Lavrekha et al., 2014; Dubreuil et al., 2018; Savina et al., 2020; Hartmann et al., 2021). At the same time, these models stay within the limits of a certain meristem, and are not applicable to a wider range of plant stem cell niches. A general description of the basic set of processes related to the redistribution of hormone gradients and cellular response to this may serve as a basis for the investigation of the common and specific features of various plant meristems.

To solve this kind of problem, professional tools have started to be developed, helping researchers to create extensible computer models, which enable applying the same mathematical model equations to various plant systems (Hay Mele et al., 2015; Schölzel et al., 2021). For example, Cell Designer is a tool for simulating biochemical networks (Kitano et al., 2005) without reference to the tissue topology. A similar tool, PySB, has ample opportunity to create, extend and combine models based on genetic networks with high complexity (Lopez et al., 2013). This Python-based software is highly flexible because it provides the possibility of direct manipulation of equations. BioNetGen allows to create models both using a graphic editor and describing models manually inside the program code that simplifies reconstruction of molecular networks (Harris et al., 2016). BioNetGen has a convenient graphical representation for the solution of equations. SBMLToolbox provides the possibility to create, validate and calculate models with ODEs using SBML in MATLAB and Octave (Keating et al., 2006). DBSolve features abundances of certain molecules in a system, displaying it dynamically as a bar graph (Gizzatkulov et al., 2010). MGSmodeller is a Java application, which enables hierarchical data presentation and editing, and implements dynamic calculation tools in reconstructing molecular genetic networks and solving inverse problems (Kazantsev et al., 2008). The COPASI software is able to describe models of biological processes, such as metabolic networks, cellular signaling pathways, regulatory networks, infectious diseases and many others, simulate and analyze these models, create analysis reports and import/export models (reviewed in Bergmann et al., 2017). In COPASI, models are defined as chemical reactions between molecules. The model analyzer includes steady-state analysis, stoichiometric analysis, time history modeling using deterministic and stochastic modeling algorithms, metabolic control analysis, optimization and parameter estimation. VCell is a computing system for modeling physicochemical and electrophysiological processes in living cells (Loew, Schaff, 2001; Moraru et al., 2008). The tool allows the user to enter a description of cell physiology, biochemical reactions, and automatically or manually input mathematical equations. The resulting simulations are displayed on dynamic spatial regions of various shapes, including irregular 3D geometries derived from experimental images. VCell can also implement rule-based models, which allows the representation of species as structured objects consisting of molecules and uses reaction rules to define molecular interactions. SpringSaLaD is a software platform based on spatial stochastic modeling of biochemical systems (Michalski, Loew, 2016). SpringSaLaD models molecules as a group of connected spherical regions with excluded volume. This allows establishing a connection between molecular dynamics modeling and processes at the cellular level. SpringSaLaD is a standalone tool that supports model building, simulation, visualization, and data analysis through a graphical user interface.

The tools listed above develop models for metabolic and signal transduction pathways, and gene regulation networks. Such tools do not implement embedding of the generated mathematical models into cell ensembles to study the influence of regulatory networks on cell divisions, growth and differentiation (Kitano et al., 2005; Keating et al., 2006; Kazantsev et al., 2008; Gizzatkulov et al., 2010; Lopez et al., 2013; Harris et al., 2016).

On the other hand, there are programs that along with simulation of gene networks also consider the influence of regulatory circuits on cell growth or divisions. CompuCell3D is a tool for constructing dynamic multicellular 2D and 3D models to simulate cells that lack a cell wall (Swat et al., 2012). It is based on the lattice-based Glazier–Graner–Hogeweg (GGH) Monte Carlo multi-cell modeling, which employs an energetic approach to model growth, intercellular communication and maintenance of cell shape. Molecular processes, namely, the production and diffusion of substances, are described via ODE

solvers. The VirtualLeaf program simulates the relationship between gene expression and the biophysics of plant cell growth (Merks et al., 2011). The model is a set of cells and cell walls, through which chemical substances can move, affecting gene expression and properties of the cell wall. Cellzilla is a 2D tissue modeling platform using Cellerator, a tool describing biochemical interactions via simplified notation as reactions and converting them automatically to the corresponding differential equations by an inner computer algebra system (Shapiro et al., 2013). In Cellzilla, cells are represented by a polygonal grid of well-mixed compartments. Cell components can interact through Cellerator reactions, which describe diffusion and transport. Dynamic simulation consists of cell growth and division. Despite these advantages, modern software tools for modeling usually use manual setting of parameters, and do not support automatic parameter fitting, which may be critical for some models.

A recent trend is further improvement of computer tools, which can be used by biologists for in-depth study of developmental processes at the multicellular level. One of the current challenges is the creation of software that constructs numerical models along various plant organs utilizing uniformly described processes and provides automatic parameters setting. Here we present a tool creating one-dimensional computer models that provide embedding of signaling molecules into a dynamically developing cellular ensemble, where, based on the same set of processes, it is possible to model cellular dynamics in various plant tissues. To build realistic computer models, it is necessary to apply experimental data. The tool we have developed takes experimental data into account already at the first stage of parameter fitting, which brings the constructed models as close to reality as possible.

Materials and methods

DyCeModel overview. DyCeModel allows creating a dynamic one-dimensional cell lattice, embedding it into a mathematical model in ODE, and performing numerical analysis. It contains five script files (.m files) executed in the MATLAB software environment (Fig. 1). The substance_eq.m block incorporates an ODE system for description of synthesis, degradation, passive and active transport for the substances of interest. By default, DyCeModel provides examples of functions, which describe these processes for two substances according to Michaelis–Menten kinetics and Generalized Hill function method (Likhoshvai, Ratushny, 2007), Fick's law of diffusion and the mass action law (for describing active transport). Alternatively, users can build their own functions instead of the default ones. The parameters_fitting.m block describes the realization of a genetic algorithm to assess the similarity of the modeled substance distribution to the experimental data. The model_parameters.m block contains all model parameter default values for the ODE system and describes the model configuration of substance influxes. The grow_eq.m block describes the cell growth function, tool_1d_model.m ensures the simulation procedure. Importantly, there are two different strategies of applying DyCeModel: using the parameters_fitting.m block or omitting it. In the latter case, the user should define all parameters in the model_parameters.m file.

The input data. A pre-processed experimentally obtained microscopic image, which visualizes the distribution profile of the substance concentration within the modeled tissue, is an input for the parameters_fitting.m block. DyCeModel accepts TIFF, GIF, JPEG, PNG formats and some other graphic file formats supported by MATLAB, and it is capable of process-

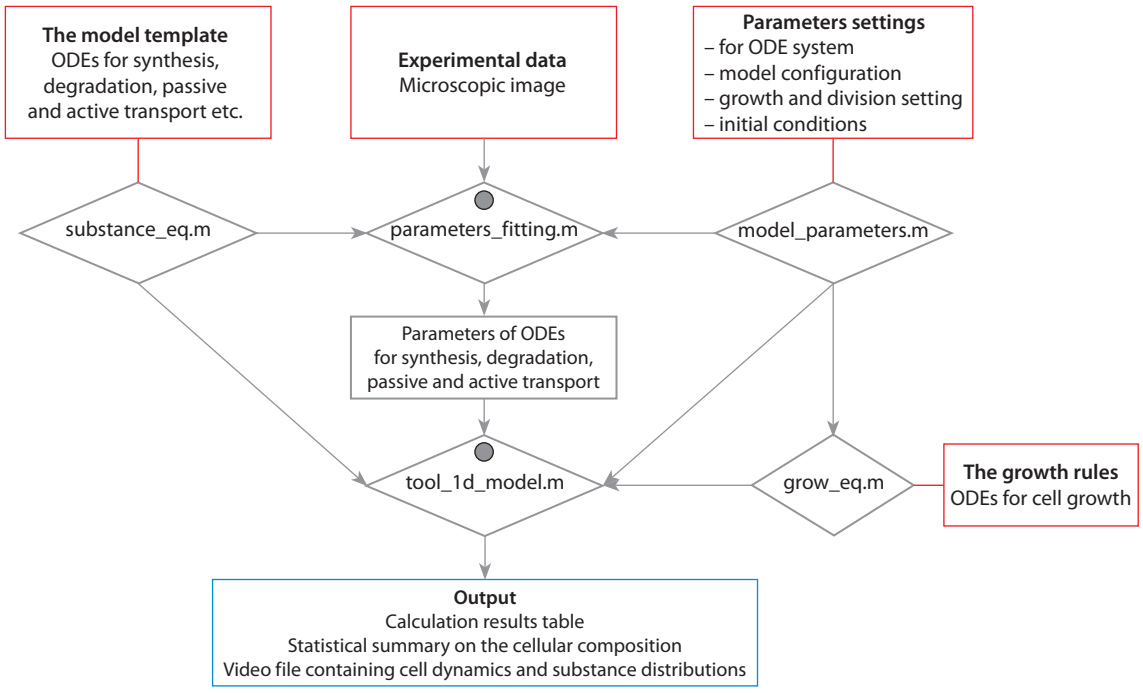


Fig. 1. DyCeModel pipeline for creating mathematical models.

The input data are marked in red. The output data are depicted in blue. Gray circles indicate the presence of visualization modules. Five script files are given in rhombuses.

ing the signal localized in the cytosol or in the nucleus. The image must be well focused. The aforementioned image preprocessing consists in excision of a rectangular area containing the modeled axis along the tissue, which should be parallel to the long side of the rectangle. This area should not contain microscope artifacts. To obtain noise-free measurements, the user can decrease the size of the rectangular area (the minimum size of the uploaded rectangular image is 1 pixel in width and 90 pixels in length). There are no strict requirements for image resolution.

The ODE system of the mathematical model is an input, which is written in the substance_eq.m file block (see Fig. 1). The default example equations can be changed according to the user's request. The model configuration is defined by the initial number of cells and position of the substance influxes in model_parameters.m file. For the model simulation, initial concentrations of all substances, as well as growth and division settings should be defined. The user also sets the number of calculation steps in order to define the time of investigation. If automatic parameter fitting is going to be omitted, the user can optionally set the parameters for the ODE system in the model_parameters.m file. All default example model parameters are consistent with the default ODE system and calculation procedure.

The parameter fitting. First, the parameters_fitting.m script quantifies the distribution of the substance concentration along the selected axis from the microscopic image. These data will be used as target distribution, which the algorithm should reproduce as precisely as possible according to the model equations and configuration. At this stage, the concentration distribution can be manually corrected if it is distorted in the microscopic image. Next, the genetic algorithm is used to find a set of model parameters, which allow reproducing the input experimental data on the distribution of the substance concentration the most accurately (Fig. 2) (Dubitzky et al., 2013).

Initially, the parameters_fitting.m script generates individuals: the sets of model parameters assigned to random values. Each individual is characterized with the fitness function value that scores the similarity of the modeled distribution of the substance concentration to the experimental data. The root-mean-square deviation (RMSD) metric is used as a fitness function. A lower fitness value corresponds to a better quality of the solution. The genetic algorithm is implemented in the following three steps.

Step 1 is "mutation", which changes a randomly selected parameter in each parameter set by the value of λ (which is also randomly selected in the interval from 0 to 1). For each individual, we calculate the model with a new parameter set. "Mutation" is fixed if it brings the solution closer to the target distribution. Step 2 is "crossover", the exchange of the parameter values between two individuals. In the first new set of parameters, a few (the number is defined randomly at each step of the algorithm) are picked from individual 1, and the rest are taken from individual 2. The second new set of parameters incorporates the values for the corresponding few parameters from individual 2, and the values for the rest of the parameters are taken from individual 1. The model is calculated with two new sets of parameters after the "crossover", and the recombination event is fixed if it brings the solution closer to the target. Step 3 supports biologically reasonable limitations on parameter values, which the user can set up manually in the "Biological limits" block of the parameters_fitting.m script (see Fig. 2). The restrictions may apply, for example, to the parity of the passive transport of different substances, the parity of active and passive transport of the same substance, the parity of the substance inflow and synthesis, etc. Taking into account reasonable biological restrictions, the algorithm "rewards" the realistic parameter values during selection, which both favors identification of the local optimum corresponding to the real processes, and speeds up the algorithm.

The fitting ends when the difference between the substance distribution calculated with the adjusted parameters and target substance distribution from the microscopic image becomes less than the threshold. The selected parameters set (the "Par" variable) is saved in a file. After executing the parameters_fitting.m script, it is recommended to inspect the selected parameters, since not all biological limitations could be taken into account during the selection. The user can view the "Par" variable and, if there are obvious inconsistencies in parameter matching, restart the parameter fitting.

Calculation of the mathematical model. When the ODE system and cell growth rules are defined, the user can load the mandatory parameters of the model with the model_parameters() function, including the initial number of cells, the initial concentrations of substances, the initial cell sizes, the maximum number of cells to be monitored, cell division parameters and cell growth settings according to the function described in the grow_eq.m file. Then the user uploads the set of parameters for the model ODE system, which are either obtained during the parameter fitting procedure or defined manually in the model_parameters.m script. After that, the model can be calculated (Fig. 3). We proceed under the assumption that cell dynamic events such as division or differentiation are discrete processes. Therefore, the calculation of ODEs is periodically interrupted to check if the conditions for cell division and differentiation specified in the tool_1d_model.m and model_parameters.m files are met. Optionally, the user decides which substances will regulate the ability to divide and the probability of cell division. All calculation results obtained during the simulation of the model are recorded in a video file, which represents the redistribution of the substance concentrations on a one-dimensional dynamic cellular ensemble.

Images used in the study. To model root apical meristem, we used publicly available images for 9-day-old *Arabidopsis thaliana* seedlings expressing *DR5::GFP* auxin sensor (Ottenschläger et al., 2003) or *TCSn::GFP* cytokinin sensor (Zürcher et al., 2013), which were obtained using a confocal fluorescence microscope (FV-1200, Olympus) (Sakamoto et al., 2019). To model shoot apical meristem, we took publicly available images for 7-day-old *A. thaliana* seedlings expressing *TCSn::GFP* cytokinin sensor obtained by a confocal microscope (Leica) (Zürcher et al., 2016). As a visualization of auxin distribution in the shoot apical meristem, we used images of auxin immunolocalization in the inflorescences of 22-day-old *A. thaliana* seedlings taken by a confocal microscope (LSM, FluoView1000, Olympus) (Banasiak et al., 2019).

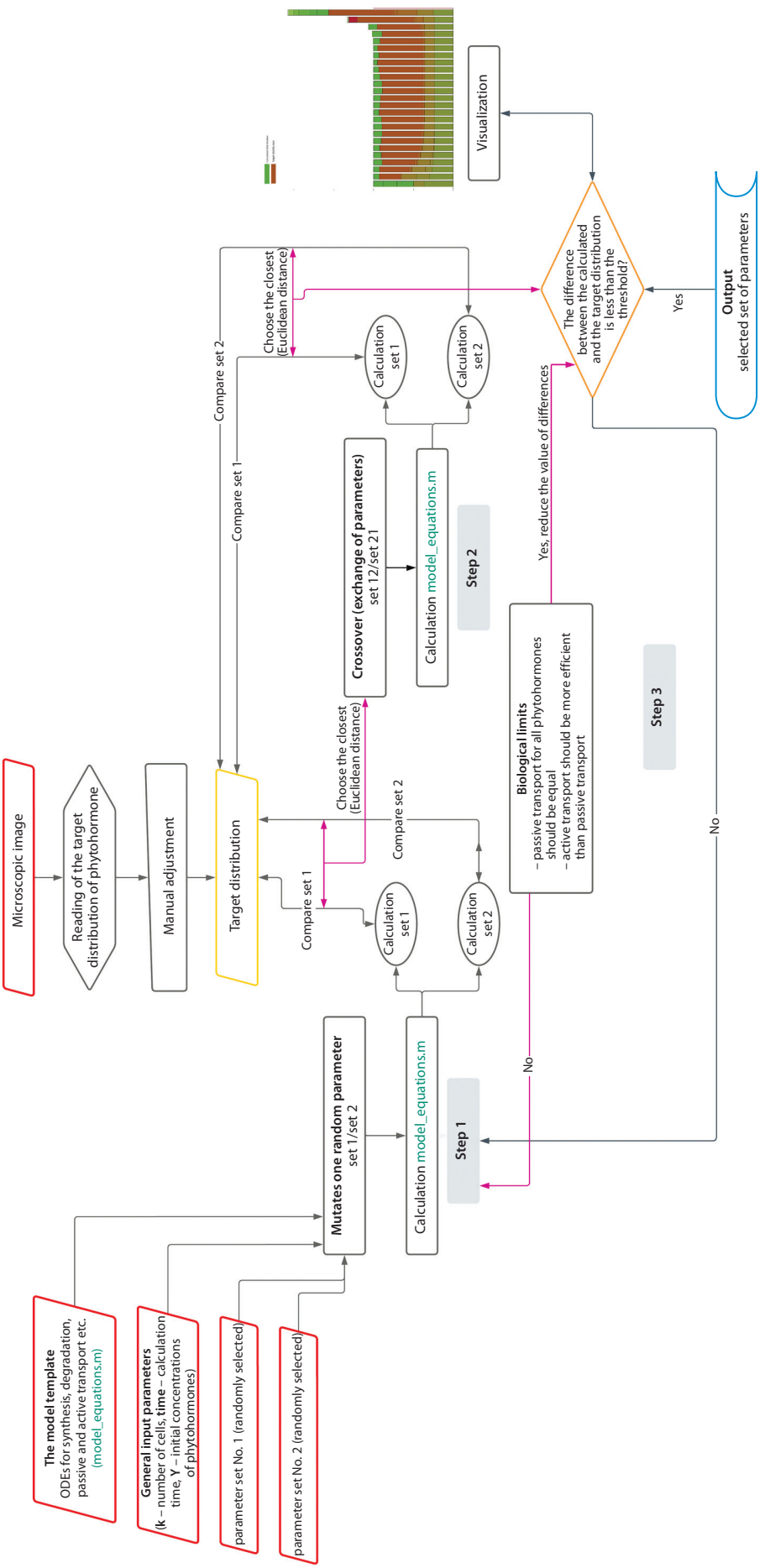


Fig. 2. The scheme of semi-automatic parameter fitting by the genetic algorithm method.
User-downloaded inputs are marked by red. Output data is depicted in blue. Data comparison blocks are marked by yellow. Black arrows indicate the processes executable in the model, pink arrows connect the parts within the comparison blocks. Orange indicates the exit block from the cyclic selection algorithm.

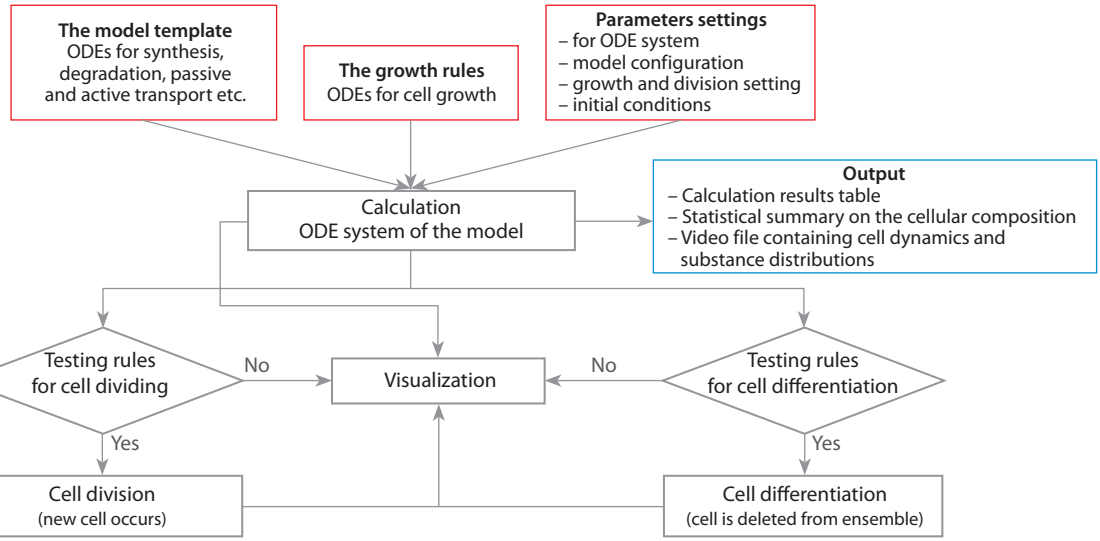


Fig. 3. The framework for calculation of the ODE system on a dynamic cell ensemble.

Input data are marked by red. Output data are depicted in blue.

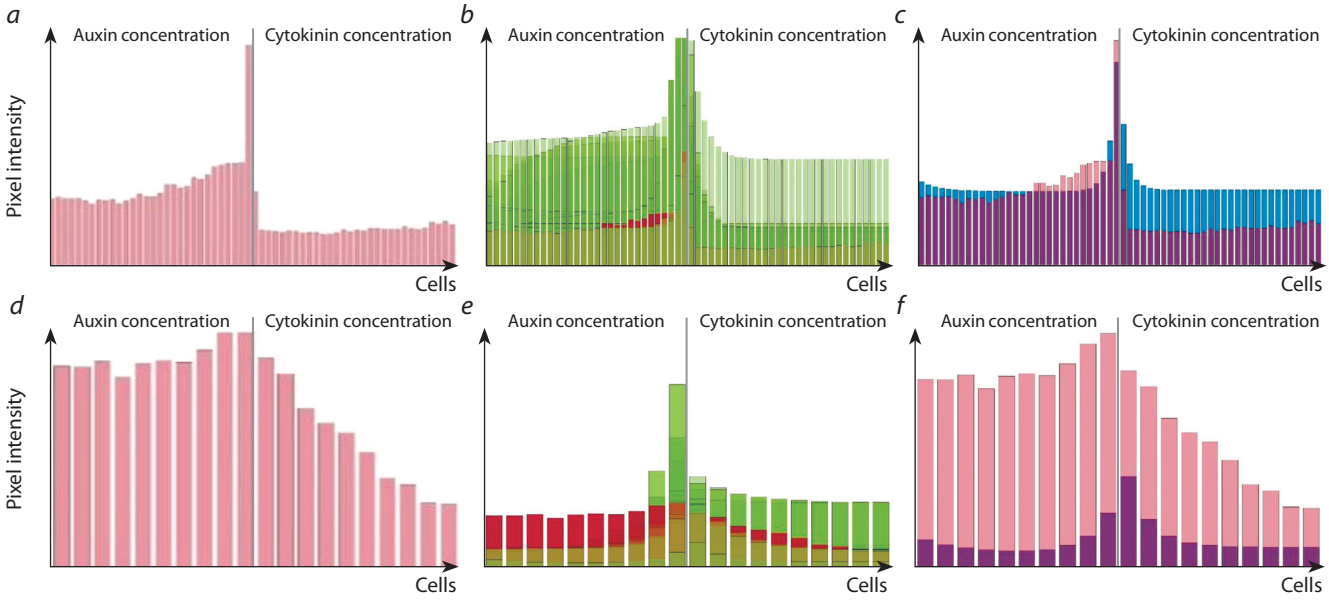


Fig. 4. DyCeModel solutions on auxin and cytokinin distribution within the root (*a–c*) and shoot (*d–f*) apical meristems along the central axis. *a*, *d*, Obtained signal intensity of hormone distributions along the allocated area; *b*, *e*, visualization of the parameter fitting process; *c*, *f*, the result of automatic parameter fitting.

Pink or burgundy indicates the signal intensity distribution obtained from the experimental data. The distributions of phytohormones during each step of parameter selection are indicated in green. Blue marks the distribution of the substances when calculating the model with the automatically selected set of parameters.

Results and discussion

A one-dimensional model of *Arabidopsis thaliana* root apical meristem built with DyCeModel

To demonstrate the performance of DyCeModel, we used it to create a 1D model of *A. thaliana* root apical meristem. Plant hormones auxin and cytokinin play major roles in regulation of maintenance of its structure (Yamoune et al., 2021). We built an ODE system based on mathematical models of auxin and cytokinin distribution published earlier (Mironova et al., 2010; Lavrekha et al., 2014). To set the parameters for the model ODE system, we used automatic parameter fitting. To define the target distribution of the hormone concentrations in the root tip, we used publicly available microscopic images described in the "Materials and methods" section. We used the following parameter value limitations during parameter fitting: approximately equal diffusion parameter values for auxin and cytokinin, prevalence of active auxin transport over the passive transport, prevalence of auxin flow into the meristem over its biosynthesis, which is typical for the root apical meristem (Overvoorde et al., 2010). Figure 4, *a–c* demonstrates auxin and cytokinin distributions in the root apical meristem gener-

ated using the DyCeModel tool. The equation describing the dependence of cell growth on auxin was built on principles similar to the Hartmann model (Hartmann et al., 2021), where the growth rate is directly proportional to auxin concentration in the cell and inversely proportional to the cell size. Cell division can occur if the cell attains minimum size required for division and possesses a certain ratio of auxin and cytokinin concentrations. The probability of cell division is 0.1, the values were obtained by analyzing images and 24-hour videos with marked division events in the meristem of *A. thaliana* (Marhava et al., 2019). Cell differentiation occurs if the cell size approaches the "maximum cell size" parameter value.

Then we built a functional model of the root apical meristem and obtained a stationary solution. Analysis of the steady-state solution of the root apical meristem model showed that it is consistent with experimental data (García-Gómez et al., 2017; Hu et al., 2021). The distribution of auxin had the shape of an inverted dome and reached a maximum in cells representing the quiescent center. The concentration of cytokinin decreased nonlinearly towards the stem cell niche and reached a minimum in the initial cells, which corresponds to experimental data. In the constructed model obtained with DyCeModel, the correct location and size of the zone of high proliferative activity were specified and remained stable for a long period of calculation, corresponding to those in the root meristem *in vivo*. Similar zones of proliferative activity were also formed in two other models of the root apical meristem (Mironova et al., 2010; Lavrekha et al., 2014).

DyCeModel enables modeling distinct types of plant meristems based on the same ODEs

We speculated that recruiting specific input characteristics for distinct meristems could enable the modeling of distinct types of plant meristems with DyCeModel based on the same ODEs. Therefore, we applied DyCeModel to build a model of *A. thaliana* shoot apical meristem using the same mathematical model equations and rules as for root apical meristem. The images used for automatic parameter fitting are described in the "Materials and methods" section. We used the following parameter value limitations during parameter fitting: approximately equal diffusion parameter values for auxin and cytokinin and a low level of auxin synthesis. In the model of the shoot apical meristem, we obtained a hormone distribution profile that qualitatively corresponds to experimental data (Heisler, Jönsson, 2006). Auxin and cytokinin concentrations decreased nonlinearly with distance from the stem cells. One-dimensional simulations of the shoot apical meristem of *A. thaliana* established a dynamic balance between dividing and differentiated cells. In this way, zones of proliferative activity were identified, and the number of cells within this zone was maintained at a certain level throughout the entire model calculation. At the same time, the identified parameters of passive transport, degradation, and growth were the same for the model of shoot meristem and root meristem, and the parameters determining cell division remained similar.

Conclusion

The DyCeModel tool constructs mathematical models of hormone distribution based on the processes of their synthesis, degradation, diffusion and active transport in a dynamically developing cellular ensemble. Such models are necessary to consider the influence of hormone distribution on cell growth and division. The developed DyCeModel tool is quite flexible, it provides embedding, addition, mixing of already existing mathematical models. Adding each model to the scripts switches on machine selection of unknown parameters, which speeds up the work with the model and makes it more stable. In addition, DyCeModel makes a statistical summary on the cellular composition that can be used for predictions about the influence of hormones on proliferative cell activity.

Using DyCeModel, we built a functional model of the root apical meristem, which was consistent with the experimental data. Next, we applied DyCeModel to build a model of the shoot apical meristem using the same mathematical model equations as for the root apical meristem model and demonstrated that the parameters of passive transport, degradation, growth, even the parameters determining cell division remain similar between root and shoot models. The resulting one-dimensional models can be further used as a framework to study the role of hormone-controlled gene networks in cell dynamics in two types of meristem.

References

Banasiak A., Biedroń M., Dolzblasz A., Berezowski M.A. Ontogenetic changes in auxin biosynthesis and distribution determine the organogenic activity of the shoot apical meristem in *pin1* mutants. *Int. J. Mol. Sci.* 2019;20(1):180. DOI 10.3390/ijms20010180

Band L.R., Wells D.M., Fozard J.A., Ghetiu T., French A.P., Pound M.P., Wilson M.H., Yu L., Li W., Hijazi H.I., Oh J., Pearce S.P., Perez-Amador M.A., Yun J., Kramer E., Alonso J.M., Godin C., Vernoux T., Hodgman T.C., Pridmore T.P., Swarup R., King J.R., Bennett M.J. Systems analysis of auxin transport in the *Arabidopsis* root apex. *Plant Cell*. 2014;26(3):862-875. DOI 10.1105/tpc.113.119495

Bergmann F.T., Hoops S., Klahn B., Kummer U., Mendes P., Pahle J., Sahle S. COPASI and its applications in biotechnology. *J. Biotechnol.* 2017;261:215-220. DOI 10.1016/j.jbiotec.2017.06.1200

De Rybel B., Adibi M., Breda A.S., Wendrich J.R., Smit M.E., Novák O., Yamaguchi N., Yoshida S., van Isterdael G., Palovaara J., Nijsse B., Boekschoten M.V., Hooiveld G., Beeckman T., Wagner D., Ljung K., Fleck C., Weijers D. Integration of growth and patterning during vascular tissue formation in *Arabidopsis*. *Science.* 2014;345(6197):1255215. DOI 10.1126/science.1255215

Dubitzky W., Wolkenhauer O., Cho K.-H., Yokota H. (Eds.) Encyclopedia of Systems Biology. New York: Springer, 2013. DOI 10.1007/978-1-4419-9863-7

Dubreuil C., Jin X., Grönlund A., Fischer U. A local auxin gradient regulates root cap self-renewal and size homeostasis. *Curr. Biol*. 2018;28(16):2581-2587.e3. DOI 10.1016/j.cub.2018.05.090

Fischer S.C., Bassel G.W., Kollmannsberger P. Tissues as networks of cells: towards generative rules of complex organ development. *J. R. Soc. Interface*. 2023;20(204):20230115. DOI 10.1098/rsif.2023.0115

García-Gómez M.L., Azpeitia E., Álvarez-Buylla E.R. A dynamic genetic-hormonal regulatory network model explains multiple cellular behaviors of the root apical meristem of *Arabidopsis thaliana*. *PLoS Comput. Biol*. 2017;13(4):e1005488. DOI 10.1371/journal.pcbi.1005488

Gizzatkulov N.M., Goryanin I.I., Metelkin E.A., Mogilevskaya E.A., Peskov K.V., Demin O.V. DBSolve Optimum: a software package for kinetic modeling which allows dynamic visualization of simulation results. *BMC Syst. Biol*. 2010;4(1):109. DOI 10.1186/1752-0509-4-109

Harris L.A., Hogg J.S., Tapia J.-J., Sekar J.A.P., Gupta S., Korsunsky I., Arora A., Barua D., Sheehan R.P., Faeder J.R. BioNetGen 2.2: ad-

vances in rule-based modeling. *Bioinformatics*. 2016;32(21):3366-3368. DOI 10.1093/bioinformatics/btw469

Hartmann F.P., Rathgeber C.B.K., Badel É., Fournier M., Moulia B. Modelling the spatial crosstalk between two biochemical signals explains wood formation dynamics and tree-ring structure. *J. Exp. Bot*. 2021;72(5):1727-1737. DOI 10.1093/jxb/eraa558

Hay Mele B., Giannino F., Vincenot C.E., Mazzoleni S., Cartení F. Cell-based models in plant developmental biology: insights into hybrid approaches. *Front. Environ. Sci*. 2015;3:73. DOI 10.3389/fenvs.2015.00073

Hayashi M., Mähönen A.P., Sakakibara H., Torii K.U., Umeda M. Plant Stem Cells: the source of plant vitality and persistent growth. *Plant Cell Physiol*. 2023;64(3):271-273. DOI 10.1093/pcp/pcad009

Heisler M.G., Jönsson H. Modeling auxin transport and plant development. *J. Plant Growth Regul*. 2006;25:302-312. DOI 10.1007/s00344-006-0066-x

Hu Y., Omary M., Hu Y., Doron O., Hoermayer L., Chen Q., Megides O., Chekli O., Ding Z., Friml J., Zhao Y., Tsarfaty I., Shani E. Cell kinetics of auxin transport and activity in *Arabidopsis* root growth and skewing. *Nat. Commun*. 2021;12(1):1657. DOI 10.1038/s41467-021-21802-3

Kazantsev F.V., Akberdin I.R., Bezmaternykhi K.D., Lashin S.A., Podkolodnaya N.N., Likhoshvai V.A. MGSmodeller – a computer system for reconstruction, calculation and analysis mathematical models of molecular genetic system. In: Abstracts of the VI International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2008), Novosibirsk, June 22–28. Novosibirsk: ICG, 2008;113

Keating S.M., Bornstein B.J., Finney A., Hucka M. SBMLToolbox: an SBML toolbox for MATLAB users. *Bioinformatics*. 2006;22(10):1275-1277. DOI 10.1093/bioinformatics/btl111

Kitano H., Funahashi A., Matsuoka Y., Oda K. Using process diagrams for the graphical representation of biological networks. *Nat. Biotechnol*. 2005;23(8):961-966. DOI 10.1038/nbt1111

Lavrekha V.V., Omelyanchuk N.A., Mironova V.V. Mathematical model of phytohormone regulation of root meristematic zone formation. *Vavilov J. Genet. Breed*. 2014;18(4/2):963-972 (in Russian)

Likhoshvai V., Ratushny A. Generalized hill function method for modeling molecular processes. *J. Bioinform. Comput. Biol*. 2007;5(2B):521-531. DOI 10.1142/s0219720007002837

Loew L.M., Schaff J.C. The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol.* 2001;19(10):401-406. DOI 10.1016/S0167-7799(01)01740-1

Lopez C.F., Muhlich J.L., Bachman J.A., Sorger P.K. Programming biological models in Python using PySB. *Mol. Syst. Biol*. 2013;9(1):646. DOI 10.1038/msb.2013.1

Marhava P., Hoermayer L., Yoshida S., Marhavý P., Benková E., Friml J. Re-activation of stem cell pathways for pattern restoration in plant wound healing. *Cell*. 2019;177(4):957-969.e13. DOI 10.1016/j.cell.2019.04.015

Merks R.M.H., Guravage M., Inzé D., Beemster G.T.S. VirtualLeaf: an open-source framework for cell-based modeling of plant tissue growth and development. *Plant Physiol*. 2011;155(2):656-666. DOI 10.1104/pp.110.167619

Michalski P.J., Loew L.M. SpringSaLaD: a spatial, particle-based biochemical simulation platform with excluded volume. *Biophys. J.* 2016;110(3):523-529. DOI 10.1016/j.bpj.2015.12.026

Mironova V.V., Omelyanchuk N.A., Yosiphon G., Fadeev S.I., Kolchanov N.A., Mjolsness E., Likhoshvai V.A. A plausible mechanism for auxin patterning along the developing root. *BMC Syst. Biol*. 2010;4(1):98. DOI 10.1186/1752-0509-4-98

Moraru I.I., Schaff J.C., Slepchenko B.M., Blinov M.L., Morgan F., Lakshminarayana A., Gao F., Li Y., Loew L.M. Virtual Cell modelling and simulation software environment. *IET Syst. Biol.* 2008;2(5):352-362. DOI 10.1049/iet-syb:20080102

Muraro D., Byrne H., King J., Bennett M. The role of auxin and cytokinin signalling in specifying the root architecture of *Arabidopsis thaliana*. *J. Theor. Biol*. 2013;317:71-86. DOI 10.1016/j.jtbi.2012.08.032

Nikolaev S.V., Kolchanov N.A., Fadeev S.I., Kogai V.V., Mjolsness E. Investigation of a one-dimensional model of the regulation of the size of the renewal zone in biological tissue, taking into account cell division. *Computational Technologies*. 2006;11(2):67-81. (in Russian)

Ottenschläger I., Wolff P., Wolverton C., Bhalerao R.P., Sandberg G., Ishikawa H., Evans M., Palme K. Gravity-regulated differential auxin transport from columella to lateral root cap cells. *Proc. Natl. Acad. Sci. USA*. 2003;100(5):2987-2991. DOI 10.1073/pnas.0437936100

Overvoorde P., Fukaki H., Beeckman T. Auxin control of root development. *Cold Spring Harb. Perspect. Biol.* 2010;2(6):a001537. DOI 10.1101/cshperspect.a001537

Rutten J., van den Berg T., Tusscher K.T. Modeling auxin signaling in roots: auxin computations. *Cold Spring Harb. Perspect. Biol*. 2022;14(2):a040089. DOI 10.1101/cshperspect.a040089

Sakamoto T., Sotta N., Suzuki T., Fujiwara T., Matsunaga S. The 26S proteasome is required for the maintenance of root apical meristem by modulating auxin and cytokinin responses under high-boron stress. *Front. Plant Sci*. 2019;10:590. DOI 10.3389/fpls.2019.00590

Savina M.S., Pasternak T., Omelyanchuk N.A., Novikova D.D., Palme K., Mironova V.V., Lavrekha V.V. Cell dynamics in WOX5-overexpressing root tips: the impact of local auxin biosynthesis. *Front. Plant Sci*. 2020;11:560169. DOI 10.3389/fpls.2020.560169

Schölzel C., Blesius V., Ernst G., Goesmann A., Dominik A. Countering reproducibility issues in mathematical models with software engineering techniques: a case study using a one-dimensional mathematical model of the atrioventricular node. *PLoS One*. 2021;16(7):e0254749. DOI 10.1371/journal.pone.0254749

Shapiro B.E., Meyerowitz E.M., Mjolsness E. Using Cellzilla for plant growth simulations at the cellular level. *Front. Plant Sci*. 2013;4:408. DOI 10.3389/fpls.2013.00408

Swat M.H., Thomas G.L., Belmonte J.M., Shirinifard A., Hmeljak D., Glazier J.A. Multi-scale modeling of tissues using CompuCell3D. *Methods Cell Biol.* 2012;110:325-366. DOI 10.1016/B978-0-12-388403-9.00013-8

Yamoune A., Cuyacot A.R., Zdarska M., Hejatko J. Hormonal orchestration of root apical meristem formation and maintenance in Arabidopsis. *J. Exp. Bot*. 2021;72(19):6768-6788. DOI 10.1093/jxb/erab360

Zürcher E., Tavor-Deslex D., Lituiev D., Enkerli K., Tarr P.T., Müller B. A robust and sensitive synthetic sensor to monitor the transcriptional output of the cytokinin signaling network in planta. *Plant Physiol*. 2013;161(3):1066-1075. DOI 10.1104/pp.112.211763

Zürcher E., Liu J., di Donato M., Geisler M., Müller B. Plant development regulated by cytokinin sinks. *Science*. 2016;353(6303):1027-1030. DOI 10.1126/science.aaf7254

ORCID ID

D.S. Azarova orcid.org/0009-0006-2030-6842
V.V. Mironova orcid.org/0000-0003-3438-0147
E.V. Zemlyanskaya orcid.org/0009-0005-7316-7690
V.V. Lavrekha orcid.org/0000-0001-8813-8941

Acknowledgements. The model development was supported by the budget project FWNR-2022-0020. All computational experiments were supported by the Russian Science Foundation, grant No. 20-14-00140.

Conflict of interest. The authors declare no conflict of interest.

Received August 16, 2023. Revised October 2, 2023. Accepted October 5, 2023.

Вавиловский журнал генетики и селекции. 2023;27(7):898-905
DOI 10.18699/VJGB-23-104

Перевод на английский язык https://vavilov.elpub.ru/jour

Лабораторные информационные системы для управления исследовательскими работами в биологии

А.М. Мухин[1, 2, 3] ✉, Ф.В. Казанцев[1, 2, 3], С.А. Лашин[1, 2, 3]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Курчатовский геномный центр ИЦиГ СО РАН, Новосибирск, Россия
[3] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ mukhin@bionet.nsc.ru

Аннотация. Современная исследовательская работа в биологии нередко требует усилий одной или нескольких групп исследователей. Часто это группы специалистов из смежных областей, которые генерируют и обмениваются данными разных форматов и размеров. Без применения современных подходов автоматизации работы и версионирования данных (когда данные от разных сотрудников сохраняются в разные моменты времени) коллективная работа быстро переходит в неуправляемый хаос. В настоящем обзоре приведен ряд информационных систем, предназначенных для решения озвученных задач. Их применение для организации научной деятельности позволяет управлять потоком действий и данных, добиваясь работы всех участников с актуальной информацией, и решением вопроса воспроизводимости как экспериментальных, так и вычислительных результатов. Описаны методики по организации потоков данных в рамках работы коллектива, принципы по организации метаданных и онтологий. Рассмотрены информационные системы Trello, Git, Redmine, SEEK, OpenBIS и Galaxy. Описана их функциональность и сфера использования. Выбирая те или иные инструменты, важно понимать цель внедрения, определить набор задач, которые они должны решать, и исходя из этого формулировать требования и отслеживать применение рекомендаций на местах. Задачи по созданию структуры онтологий, метаданных, схем хранения данных и программных систем являются ключевыми для коллектива, который решился на проведение работ по автоматизации оборота данных. Не всегда возможно внедрить такие системы целиком, но все же следует стремиться к этому через поэтапное внедрение принципов по организации данных и задач с освоением отдельных программных инструментов. Следует отметить, что системы Trello, Git и Redmine проще в использовании, настройке и поддержке для малых исследовательских групп. В то же время SEEK, OpenBIS и Galaxy более специфичные, их применение целесообразно в случае, если возможностей простых систем уже недостаточно.
Ключевые слова: управление; LIMS; ELN; FAIR; системы контроля версий; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

Для цитирования: Мухин А.М., Казанцев Ф.В., Лашин С.А. Лабораторные информационные системы для управления исследовательскими работами в биологии. *Вавиловский журнал генетики и селекции*. 2023;27(7):898-905. DOI 10.18699/VJGB-23-104

Laboratory information systems for research management in biology

A.M. Mukhin[1, 2, 3] ✉, F.V. Kazantsev[1, 2, 3], S.A. Lashin[1, 2, 3]

[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
[3] Novosibirsk State University, Novosibirsk, Russia
✉ mukhin@bionet.nsc.ru

Abstract. Modern investigations in biology often require the efforts of one or more groups of researchers. Often these are groups of specialists from various scientific fields who generate and share data of different formats and sizes. Without modern approaches to work automation and data versioning (where data from different collaborators are stored at different points in time), teamwork quickly devolves into unmanageable confusion. In this review, we present a number of information systems designed to solve these problems. Their application to the organization of scientific activity helps to manage the flow of actions and data, allowing all participants to work with relevant information and solving the issue of reproducibility of both experimental and computational results. The article describes methods for organizing data flows within a team, principles for organizing metadata and ontologies. The information systems Trello, Git, Redmine, SEEK, OpenBIS and Galaxy are considered. Their functionality and scope of use are described. Before using any tools, it is important to understand the purpose of implementation, to define the set of tasks they should solve, and, based on this, to formulate requirements and finally to monitor the application of recommendations in the field. The tasks of creating a framework of ontologies, metadata, data

warehousing schemas and software systems are key for a team that has decided to undertake work to automate data circulation. It is not always possible to implement such systems in their entirety, but one should still strive to do so through a step-by-step introduction of principles for organizing data and tasks with the mastery of individual software tools. It is worth noting that Trello, Git, and Redmine are easier to use, customize, and support for small research groups. At the same time, SEEK, OpenBIS, and Galaxy are more specific and their use is advisable if the capabilities of simple systems are no longer sufficient.
Key words: management; LIMS; ELN; FAIR; version control systems; Trello; GitHub; Redmine; SEEK; OpenBIS; Galaxy.

For citation: Mukhin A.M., Kazantsev F.V., Lashin S.A. Laboratory information systems for research management in biology. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):898-905. DOI 10.18699/VJGB-23-104

Введение

Современная исследовательская работа в биологии нередко требует усилий одной или нескольких групп исследователей. Часто это группы специалистов из смежных областей, которые генерируют и обмениваются данными разных форматов и размеров. Для автоматизации и компьютерной поддержки этой работы используют различные инструменты каталогизации, протоколирования хода протекания экспериментов и фиксации результатов: бумажные блокноты и лабораторные журналы, программы ведения электронных таблиц, составление отчетов в разных текстовых редакторах. Без применения современных подходов автоматизации работы и версионирования данных в коллективе быстро наступает «неуправляемый хаос». Критическим местом организации взаимодействия в коллективе является сложность процедуры передачи знаний от одного члена команды другому, так как такие знания не формализованы и часто содержат пометки, понятные только автору. Все это приводит к задержкам в проведении следующих этапов исследования или в оформлении публикаций. Иногда сотрудники забывают записывать новые факты и заметки либо вообще не ведут никакого учета промежуточных этапов работы. Это приводит к безвозвратным потерям знаний и тратам ресурсов на повторные эксперименты и наблюдения.

При сборе первичных данных исследователи могут также допускать ошибки в обработке значений или приписывании их к той или иной категории. Например, транскриптомные данные могут быть ошибочно приписаны к организму, отличному от того, откуда они получены; данные могут быть записаны не в унифицированном виде, с использованием значений разных типов (целое число, число с плавающей точкой, строка, дата и т. п.). При работе с Excel может произойти ошибочное преобразование строк в числа с плавающей точкой, что критично для интерпретации результатов исследования (Zeeberg et al., 2004), поэтому неявных преобразований данных необходимо избегать. В статье (Roche et al., 2015) были проанализированы биоресурсные коллекции (БРК) в области Экологии и Эволюции. Выяснилось, что 56 % этих БРК были неполными, т. е. в табличных данных были пустые значения, а 64 % собраны таким образом, что повторно использовать хранящиеся данные невозможно ввиду ошибок записи значений.

Поэтому перед каждым коллективом стоит задача по грамотной формализации процессов управления данными и обмена знаниями между сотрудниками. Далее мы рассмотрим конкретные методологии организации данных и реализующие их информационные системы и программные инструменты, которые используются научными организациями для распределения задач и автоматизации потока рабочих данных.

Методологии организации данных и процессов

Для решения задачи организации потоков научных данных существует несколько путей, но все они требуют от коллектива исследователей создания систем договоренностей по управлению, обработке и передаче научной информации. Системы автоматизации с предоставлением управляемого доступа помогают в сохранении знаний, регламентов и других «сущностей» лабораторной работы, не требуют постоянных согласований. В самом начале этих работ встают следующие вопросы: 1) использование существующих стандартов оформления данных, разработанных профессиональным сообществом; 2) формализация или создание единого «рабочего языка» внутри коллектива; 3) развертывание, внедрение и сопровождение информационной системы и настройка прав доступа для групп пользователей.

Переход на существующие стандарты и форматы представления данных или создание собственных форматов с исчерпывающей документацией, достаточной для однозначной интерпретации значений, позволяет преодолеть проблему передачи знаний между сотрудниками внутри коллектива и вне его. Сопроводительная документация будет использована для автоматизации работы с информационной системой, например для построения модулей генерации сводных диаграмм и отчетов. Формальные схемы описания результатов научной деятельности в последнее время полезны для быстрого поиска информации и интерпретации этих файлов не только машинами, но и людьми. В качестве примеров могут служить математические модели в форматах SBML (Hucka et al., 2019), SBGN (Novère et al., 2009), поддерживаемые сообществом CO.MBINE (Schreiber et al., 2015). Отметим также подход MIRIAM для описания целостных биохимических систем (Novère et al., 2005) и формат MIAME (Brazma et al., 2001) для описания результатов секвенирования на микрочипах или РНК-последовательностей.

Когда определены стандарты представления данных, наступает этап формализации или создания единого рабочего языка и протоколов обмена внутри коллектива для упорядочения передачи знаний предметной области. Если оставить подход к оформлению данных «как удобно/как раньше», то вопрос с неоднозначными или пропущенными

знаниями в базе не будет решен, что в дальнейшем приведет к дополнительным затратам ресурсов на исправление данных на более поздних стадиях работы. В решении задачи формализации и создании единого рабочего языка могут помочь инструменты с онтологиями (Guizzardi, 2020). Онтологии являются более широким классом систем организации знаний по описанию результатов в сравнении с вышеупомянутыми формальными схемами. В системах онтологий можно устанавливать «понятия» и «отношения» между понятиями, а не строго следовать за готовой схемой, предложенной кем-то ранее. Онтологии создаются с целью описания смысловой информации и однозначной интерпретации системы понятий и процессов внутри коллектива и за его пределами. Коллективы используют как простые методы описания онтологий, такие как язык логики первого порядка, так и более сложные древовидные структуры, например OntoUML (Guizzardi et al., 2018) или схемы RDF (Gutierrez et al., 2007). Для составления онтологических связей предметной области также набирает популярность математическая теория категорий (Kuś, Skowron, 2019), призванная соединять различные области математики и предметные области друг с другом. Был реализован графический язык «онтологических журналов» (англ. Ologs, по сути описания предметной области в виде графов, где в узлах описаны объекты с определенными свойствами, а в ребрах – функции по преобразованию из одного объекта в другой) с использованием основ данной теории (Spivak, Kent, 2012). В настоящее время инструментарий и язык теории категорий не используются широко в научных публикациях и системах, однако есть работы по реализации этого языка в нейробиологии (Brown, Porter, 2003) и по математическому описанию развивающейся модели памяти (Ehresmann, Vanbremeersch, 2007).

Одним из путей формализации этапов работы лаборатории является создание метаданных – информации, описывающей сами данные (Roche et al., 2015). Формат их описания довольно свободный. Метаданные могут быть описаны/представлены в виде структурированного файла (XML или JSON) или таблиц баз данных как реляционной (Postgrespro.ru), так и документно-ориентированной структуры (MongoDB.com). Описанием может быть любая информация, например: что означают колонки в таблицах, какие используются единицы измерений, из какого организма были получены материалы, каким образом получались эти результаты. Метаданные могут дополнять системы онтологий и формальные схемы представления научных результатов для быстрого поиска нужной информации и однозначного интерпретирования результатов.

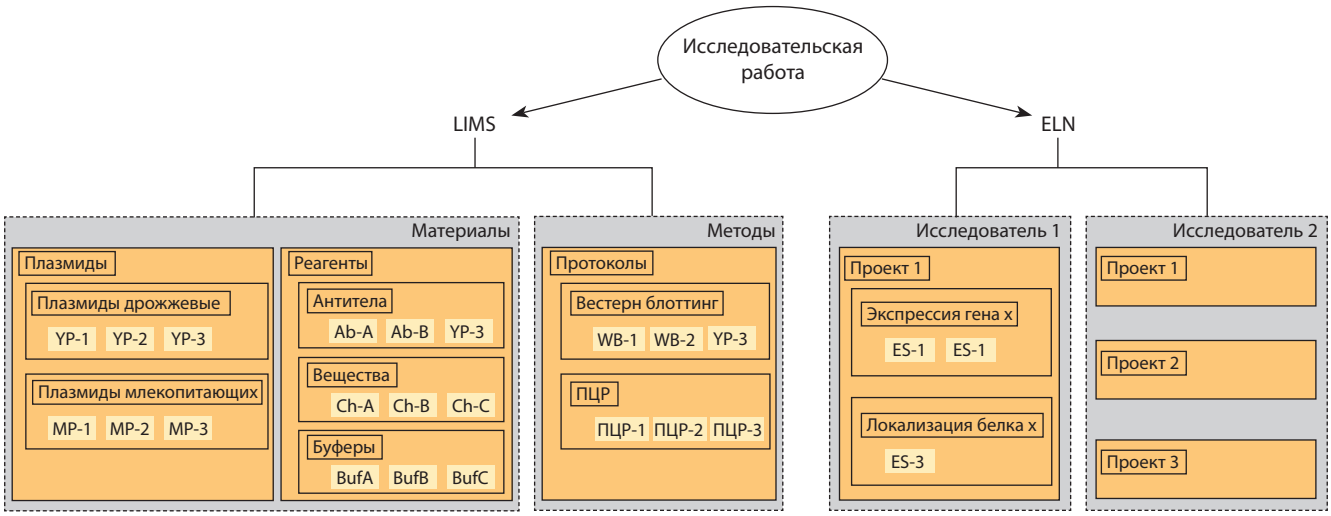
Сообщество исследователей FAIR (Wilkinson et al., 2016) предложило свой набор принципов описания данных и метаданных в задачах хранения и передачи информации как между коллективами исследователей, так и между различными программами анализа данных. Ими были сформулированы следующие четыре принципа, которыми должна обладать лабораторная информационная система:

1. Определенность (Findable) – (мета)данные уникальные и однозначно определяемые. Система должна обладать базовым механизмом чтения подробного описания и возможностью искать эти данные по ключевым полям.
2. Доступность (Accessible) – данные доступны для чтения как людям, так и компьютерам для дальнейшей работы. Достигается с помощью стандартных форматов и протоколов.
3. Интерпретируемость (Interoperable) – (мета)данные описаны в машиночитаемом виде, в удобном формате и аннотированы с помощью онтологии.
4. Повторная используемость (Reusable) – (мета)данные достаточно хорошо описаны, чтобы передавать эти данные другим людям и системам для дальнейшего анализа. Этот пункт является логичным следствием выполнения вышеупомянутых пунктов.

Далее рассмотрим программные инструменты для решения задач управления данными и автоматизации исследовательских работ.

Программные инструменты

Две концепции – LIMS и ELN (Barillari et al., 2016), которые реализовываются в программных комплексах для задач контроля выполнения исследовательских работ, приведены на рисунке.



Описание структуры данных, которые хранятся в LIMS и ELN системах.

LIMS (Laboratory Information Management System) – система управления лабораторной информацией. В ее задачи входят управление и контроль за лабораторными материалами и методами. С помощью этой системы исследователи могут осуществлять документооборот с администрацией и компаниями, создавать расписание использования инструментов, учет реактивов, объектов исследований и др.

ELN (Electronic Laboratory Notebook) – электронный лабораторный журнал. В задачи таких систем входит управление проектами, экспериментами, пользователями, исследовательскими группами, а также протоколирование (журналирование) и контроль проведения экспериментов. По сути, эти системы заменяют функции бумажных блокнотов для ведения и передачи заметок по ходу экспериментов.

Trello

Trello (https://trello.com/) является условно-бесплатным веб-сервисом по организации рабочего процесса и коммуникации. В этой системе пользователи настраивают виртуальную доску, на которой располагаются «карточки» с «заданиями». Сама доска разделена на участки, между которыми перемещаются эти карточки, демонстрируя движение по этапам работ. Чаще всего участки доски помечают статусами выполнения работ, например: «задачи в очереди», «в работе», «ждут отклика», «задача выполнена». Возможно самостоятельно создавать участки/разделы по своему сценарию, наиболее отражающему рабочий процесс коллектива. Таким образом, сотрудники и руководители могут: 1) наблюдать в режиме реального времени за прогрессом работ; 2) изменять статусы задач, добавлять к задачам комментарии; 3) связывать друг с другом задачи; 4) реагировать на ранних этапах в случаях зависших работ.

К недостаткам Trello можно отнести невозможность модифицировать функционал системы собственными модулями и ограниченную функциональность в бесплатной версии. Аналогами можно считать решения Яндекс.Трекер (https://cloud.yandex.ru/services/tracker), GutHub Projects (https://docs.github.com/en/issues/planning-and-tracking-with-projects/learning-about-projects/quickstart-for-projects) и Kanboard (https://kanboard.org/). Предложенные инструменты ориентированы на реализацию требований ELN, однако пользователи могут адаптировать их под решение задач LIMS. Они направлены на управление процессами работы коллектива. Для организации хранения и перемещения самих данных надо использовать другие инструменты.

GitHub

При совместной работе коллектива над кодами программ, документами и отчетами стоит важная задача по контролю за изменениями. Почтовые клиенты и пересылка по сети от человека к человеку плохо справляются с этой задачей, так как самим пользователям нужно контролировать актуальность версий этих документов. Также не решается задача версионирования данных и текста ввиду отсутствия системы централизации хранения файлов и фиксации их изменений. Именно эти задачи можно решить с помощью программы Git (Chacon, Straub, 2014).

Программа Git создает в локальной папке файлы репозитория, позволяющие перемещаться между изменениями в файлах. Как правило, данную систему используют программисты для одновременной работы над проектом, сравнивая и объединяя изменения кода от разных разработчиков. Проекты с открытым кодом обычно хранятся публично в серверах проекта GitHub (https://github.com). Некоторые исследовательские группы используют систему контроля версий Git для подготовки статей и диссертаций. К примеру, с их помощью писалась математическая книга по гомотопической теории типов (The Univalent Foundations Program, 2013). Над книгой работало около 20 человек, и сервис облачного хранения Dropbox не справлялся с задачей синхронизацией текста. В результате команда выпустила книгу объемом 600 страниц менее чем за полгода (https://math.andrej.com/2013/06/20/the-hott-book/).

Сам GitHub нельзя установить на локальном компьютере, однако есть аналогичные решения с возможностью установки в локальное хранилище, например GitLab (https://gitlab.com), Gogs (https://gogs.io), Gitea (https://gitea.com), GitWeb (https://git-scm.com/docs/gitweb). В рамках этих систем возможно решать задачи ELN и задачи LIMS, но пользователям придется подробно разобраться с Git.

Redmine

Redmine (https://redmine.org/) используется в качестве системы контроля проектов и распределения задач. Чаще всего главный управляющий проекта (менеджер, заведующий лабораторией и т. д.) создает набор задач и назначает ответственных исполнителей. Исполнители по мере выполнения меняют статус готовности задачи. Система автоматически отслеживает состояние задач проекта и строит сводные диаграммы, на которых видно расхождение по срокам между планом и фактическим исполнением. Также в основные функции данной системы входят:

- создание и ограничение ролей – администратор может создать несколько дополнительных ролей пользователей и установить для них правила работы в системе (чтение и/или запись «задачи», вики-страниц и т. д.);
- гибкая система по контролю ошибок – функция широко используется в сфере разработки ПО, когда тестировщики или пользователи добавляют «задачу» вида «ошибка» в систему для оповещения разработчиков;
- календарь и диаграмма Ганта. Служат для отслеживания сроков исполнения задач;
- добавление новостей по проекту с оповещением участников;
- добавление документов и файлов в систему;
- оповещение пользователей по электронной почте или RSS-ленте;
- оформление знаний для каждого проекта в формате Википедии – электронная энциклопедия/справочник в виде интернет-страниц;
- система форумов для каждого проекта – возможность публично обсудить в одном месте решение задач. Воз-

можность быстро пробежать глазами цепочки сообщений по теме;
- учет времени работы над задачами и проектом в целом;
- создание пользовательских форм и полей для дополнительного описания «задач», «проектов», «пользователей» и других сущностей в рамках данной системы.

Для данной системы существует функция по ее разворачиванию в локальной информационной среде (вплоть до персонального компьютера). Также пользователи могут реализовывать новую функциональность через реализацию подмодулей (плагинов). К недостаткам Redmine можно отнести отсутствие доски задач по типу Trello, которая понятна и проста в использовании, а также ограниченность функциональности стандартной версии. Поэтому для полноценной работы приходится устанавливать сторонние подмодули.

На основе программного комплекса Redmine построены рабочие процессы многих коллективов в секторе информационных технологий. В 2019 г. была начата реализация проекта ENVRI-FAIR (Petzold et al., 2019) по объединению ресурсов и данных между кластером Европейской инфраструктуры экологических исследований (ENVRI) и вычислительным облаком Европейской «Открытой науки» (EOSC) с использованием Redmine (эта информация была получена из технической документации данного проекта). На основе Redmine возможно реализовать решение задач и ELN, и LIMS.

Система SEEK

Система SEEK (Wolstencroft et al., 2015) предназначена для управления, распространения и изучения математических моделей и ассоциированных данных системной биологии. SEEK организовывает информацию исследовательского проекта, включающего экспериментальные данные и результаты биоинформатической обработки в рамках структуры из трех сущностей: Исследования, Стадии, Образцы (ISA) (Rocca-Serra et al., 2010). «Исследование» раскрывает суть конкретного проекта (кто выполняет работу, какой институт, время проведения исследования). «Стадия» описывает конкретный этап исследования (экскреция ДНК или белка из ткани исследуемого организма, картирование РНК-прочтений на референсный геном и т. д.). «Образец» – единица результата выполненной работы. Также в системе можно устанавливать ассоциативную связь между образцами.

Достоинством этой системы является связывание данных между собой в рамках вышеописанной структуры с описанием коллектива исследователей, а также переформатирование метаданных в граф знаний RDF (Gutierrez et al., 2007) с помощью сервера Virtuoso (Software, 2022). Метаданные описываются в основном в табличной форме (сокр. ISA-Tab), также есть возможность использования JSON схемы. Для ручной аннотации данных разработчики SEEK предлагают программное обеспечение FightField. Поиск данных по графу RDF с помощью языка запросов SPARQL является гибким в использовании в сравнении с SQL, в котором, помимо написания правил отбора данных, от пользователя требуется вручную расписать список таблиц и способ их объединения. Проблема SQL также в том, что пользователь вынужден оптимизировать свои запросы для быстрого выполнения поиска.

Основным направлением SEEK является хранение и передача математических моделей биологических процессов. Ресурс также позволяет работать с SBML моделями и открывать их в JWS Online (Olivier, Snoep, 2004) и в COPASI (Hoops et al., 2006). Эта система в основном реализовывает требования ELN по биоинформатическим проектам, а LIMS не реализован в ней.

Система OpenBIS

В рамках работы лаборатории перед исследователями стоит задача по созданию протоколов экспериментов, следованию этим протоколам с фиксацией результатов, фиксации событий и т. д. Необходимо выстраивать результаты серии данных в рамках одного проекта, например, связывание экспериментов с различными организмами, их фенотипами, генотипами, средой развития и другими данными. OpenBIS (Bauch et al., 2011) предоставляет функционал по хранению и выстраиванию метаданных под подробное описание экспериментов, их результатов, параметров и т. д. Система OpenBIS состоит из трех модулей: сервер приложения, сервер данных и база метаданных.

- Сервер приложения является точкой доступа для пользователей. Модуль реализовывает доступ к программному комплексу через графический пользовательский интерфейс, а также по HTTP протоколу (для OpenBIS предоставлены библиотеки на языках программирования Python, Java и Matlab для взаимодействия по сети). Для добавления новых функций (например, хранение данных по масс-спектрометрии) OpenBIS предоставляет систему модулей, каждый из которых должен быть реализован на языке программирования Jython. Этот модуль разделяет полномочия среди пользователей (чтение данных, чтение/запись данных).
- Сервер данных выполняет работу по организации хранения первичных данных на дисковых накопителях.
- База метаданных представляет собой систему управления базой данных (СУБД) PostgreSQL. Этот модуль связывает данные в проектах, хранит метаданные, указывает на данные из сервера данных, обеспечивает задачи поиска в данных.
- Возможность ссылок к данным на внешних ресурсах (модуль BigDataLink). Метаданные сохраняются в базе метаданных, при этом исходная информация не хранится на сервере данных, а остается на сторонних ресурсах. Эта функция используется в случае работы с файлами большого размера.
- Расширение функционала с помощью библиотек на Java, Python, JavaScript, Matlab для взаимодействия с системой OpenBIS (получение/загрузка данных, поиск метаданных). Эти библиотеки используют аппаратный интерфейс REST API сервиса OpenBIS; таким образом, можно реализовать модули для взаимодействия с системой на других языках программирования. Может использоваться для реализации автоматизированных вычислений с привлечением хранимых данных из системы OpenBIS.

- Структура хранения данных является иерархической и организована следующим образом: область (space), проект (project), эксперимент/коллекция (experiment/collection), Объект/Образец (Object/Sample), данные (Data Set).
- Для связи объектов и данных друг с другом существует метод по установлению связей «предок–потомок», т. е. система может создавать граф объектов и данных.
- Импорт/экспорт данных в табличном виде.
- Реализация дополнительного функционала самой системы с помощью системы модулей.
- Система выполняет аудит каждого изменения в своих базах данных.
- Семантическое аннотирование данных – описание результатов в удобном и интерпретируемом формате. Для описания семантики используется RDF схема (Gutierrez et al., 2007).
- Интеграция с системой SEEK.

Система OpenBIS хорошо себя зарекомендовала для первичного хранения биологической информации, полученной в ходе экспериментов. В работе (Friedrich et al., 2015) была реализована система по добавлению и учету экспериментальных данных по различным тканям организмов при применении разных препаратов. На первом уровне системы хранения описывается объект исследования (например, определенная мышь в лаборатории, которой ввели конкретный препарат). На втором уровне – определенная биологическая ткань, которую извлекли из объекта. На третьем уровне – последовательности (нуклеотидные или белковые), полученные из исследуемой ткани объекта. Система основывается на требованиях LIMS и ELN, является образцовой их реализацией.

Galaxy

Выше были описаны в основном системы для контроля лабораторных данных, однако для биоинформатических лабораторий задачи стоят точно такие же: контроль за потоком данных, воспроизводимость вычислений, доступ к данным и их сохранение в сервере. Для решения подобных задач была реализована система Galaxy (Galaxy Community, 2022). Galaxy состоит из следующих модулей: 1) сервер с программным и графическим интерфейсом; 2) рабочие процессы, которые и запускают аналитические конвейеры по запросу пользователей. Пользователи могут запускать самостоятельным образом установленные в сервере программы и там же хранить свои данные (последовательности, аннотации, список белков и т. д.).

Доступна реализация вычислительных конвейеров в виде графа, где в вершинах обозначены программы с настроенными параметрами, а связаны они между собой ребрами, которые обозначают направление движения данных от выхода одной программы ко входу другой. Также

Сравнение программных решений

Название системы	Основная сфера работы	Уровни иерархии	Использование метаданных	LIMS	ELN	Используемые средства разработки	Развертывания
Trello	Организация задач в виде заметок на доске (Kanban стиль)	Проект Стадии Задача		–/+	+	Невозможно установить в локальной среде	Не нуждается в развертывании, для локального развертывания требуются другие инструменты (например, Kanboard)
Git	Версионирование текстовых файлов	Свободное	Файлы изменений и дерево «коммитов»	–/+	+/–	Само приложение git, GitHub нельзя установить локально	Для эффективной работы требуются навыки работы с git. Также стоит решить, использовать сторонний сервис (например, GitHub) или разворачивать локальный сервер (GitLab, Gitea)
Redmine	Организация работы по проектам (используется в IT)	Проекты Задачи	Сервер PostgreSQL Можно добавлять пользовательские поля для описания	+/–	+	Ruby, PostgreSQL	Требуется развертывание системы и базы данных
OpenBIS	Управление лабораторией (LIMS) и проектами (ELN)	Проекты Эксперименты Образцы Набор данных	Сервер PostgreSQL Пользовательские поля	+	+	Java, PostgreSQL	Требуется развертывание системы и базы данных
SEEK	Управление данными и моделями системной биологии (ELN)	ISA стандарт: Исследование Стадия Образец	Схемы RDF Пользовательские поля на уровне «Образец»	–	+	Ruby, MySQL, Virtuoso	Требуется развертывание системы и базы данных
Galaxy	Воспроизводимость вычислительных экспериментов/протоколов	Отсутствует, есть связь между данными	База метаданных PostgreSQL	–	+	Python, PostgreSQL	Требуется развертывание системы и базы данных, также требуется настройка кластера

эти процессы могут запускать программы на удаленном сервере или кластере, а обмен файлами выполнять через общую файловую систему. Воспроизводимость вычислительных программ достигается с помощью системы окружений Conda (Yan Y., Yan J., 2018), когда для каждой программы создается свое независимое окружение (набор библиотек, программ и модулей на Python\R строго определенных версий). Может также использоваться система легковесной виртуализации Docker (Rad et al., 2017), в рамках которой программа запускается в «виртуальной» и «легковесной» операционной системе семейства Linux. Galaxy является FAIR-подобной системой (Hiltemann et al., 2023). По сути, Galaxy реализовывает ELN систему требований, но в области биоинформатических конвейеров, т. е. не является полноценной ELN. LIMS не реализован в полной мере, есть лишь многопользовательский вход и ограничение на хранение результатов вычислений.

Заключение

В настоящей работе было рассмотрено ограниченное множество информационных решений в сфере организации проектной деятельности лабораторий, работающих в области биологии. Краткие характеристики систем описаны в таблице. Такие решения, как OpenBIS, SEEK и Galaxy, были созданы специально для сопровождения научных работ, тогда как Trello и Redmine являются системами управления проектами более общих категорий, хотя и могут использоваться в работе научных групп. Программный комплекс Git может быть рассмотрен крупными коллективами как инструмент для обмена и версионирования программного кода, данных, текстов статей, монографий и других научных текстов. Следует отметить, что Git не предназначен для хранения бинарных файлов (в частности, файлов в формате DOCX, PDF и др.), так как учитывает лишь изменения текстовых файлов. Более подходящие форматы для такого использования Git – это Markdown и LaTeX.

Перед внедрением тех или иных инструментов важно понимать цели их внедрения. Исходя из целей сформулировать требования, определить набор задач, которые должна решать система, а также отслеживать применение рекомендаций конкретными исполнителями. Учитывая сложность перечисленных процессов, можно рекомендовать начинать с внедрения открытых форматов и стандартов по представлению и передаче биологических данных, предложенных и развиваемых научным сообществом. Использование систем документооборота общего назначения в лаборатории позволит получить опыт эксплуатации, что, в свою очередь, поможет определить форматы данных, протоколы работы и программные продукты, необходимые для работы лаборатории, и исходя из этого принимать решение о масштабировании автоматизации работы с данными, включая создание структур онтологий, метаданных, схем хранения, сценариев работы программных систем.

Список литературы / References

Barillari C., Ottoz D.S.M., Fuentes-Serna J.M., Ramakrishnan C., Rinn B., Rudolf F. openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics.* 2016;32(4):638-640. DOI 10.1093/bioinformatics/btv606

Bauch A., Adamczyk I., Buczek P., Elmer F.J., Enimanev K., Glyzewski P., Kohler M., Pylak T., Quandt A., Ramakrishnan C., Beisel C., Malmström L., Aebersold R., Rinn B. openBIS: a flexible framework for managing and analyzing complex data in biology research. *BMC Bioinformatics*. 2011;12:468. DOI 10.1186/1471-2105-12-468

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C.A., Causton H.C., Gaasterland T., Glenisson P., Holstege F.C., Kim I.F., Markowitz V., Matese J.C., Parkinson H., Robinson A., Sarkans U., Schulze-Kremer S., Stewart J., Taylor R., Vilo J., Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. 2001;29(4):365-371. DOI 10.1038/ng1201-365.

Brown R., Porter T. Category Theory and Higher Dimensional Algebra: potential descriptive tools in neuroscience. *arXiv*. 2003. DOI 10.48550/arXiv.math/0306223

Chacon S., Straub B. Pro Git. Kaliforniya: Apress Berkli, 2014. DOI 10.1007/978-1-4842-0076-6

Ehresmann A., Vanbremeersch J. Memory Evolutive Systems: Hierarchy, Emergence, Cognition. Elsevier Science, 2007.

Friedrich A., Kenar E., Kohlbacher O., Nahnsen S. Intuitive web-based experimental design for high-throughput biomedical data. *BioMed Res. Int.* 2015;2015:958302. DOI 10.1155/2015/958302

Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res*. 2022;50(W1):W345-W351. DOI 10.1093/nar/gkac247

Guizzardi G. Ontology, ontologies and the "I" of FAIR. *Data Intell*. 2020;2(1-2):181-191. DOI 10.1162/dint_a_00040

Guizzardi G., Fonseca C.M., Benevides A.B., Almeida J.P.A., Porello D., Sales T.P. Endurant Types in Ontology-Driven Conceptual Modeling: Towards OntoUML 2.0. In: Conceptual Modeling – 37th International Conference, Xi'an, China, October 22–25, 2018. Proceedings. Berlin: Springer, 2018;136-150. DOI 10.1007/978-3-030-00847-5_12

Gutierrez C., Hurtado C.A., Vaisman A. Introducing time into RDF. *IEEE Trans. Knowl. Data Eng*. 2007;19(2):207-218. DOI 10.1109/TKDE.2007.34

Hiltemann S., Rasche H., Gladman S., Hotz H.-R., Larivière D., Blankenberg D., Jagtap P.D., Wollmann T., Bretaudeau A., Goué N., Griffin T.J., Royaux C., Bras Y.L., Mehta S., Syme A., Coppens F., Droesbeke B., Soranzo N., Bacon W., Psomopoulos F., Gallardo-Alba C., Davis J., Föll M.C., Fahrner M., Doyle M.A., Serrano-Solano B., Fouilloux A.C., van Heusden P., Maier W., Clements D., Heyl F., Network G.T., Grüning B., Batut B. Galaxy Training: a powerful framework for teaching! *PLoS Comput. Biol*. 2023;19(1):e1010752. DOI 10.1371/journal.pcbi.1010752

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

Hucka M., Bergmann F.T., Chaouiya C., Dräger A., Hoops S., Keating S.M., König M., Le Novère N., Myers C.J., Olivier B.G., Sahle S., Schaff J.C., Sheriff R., Smith L.P., Waltemath D., Wilkinson D.J., Zhang F. The Systems Biology Markup Language (SBML): language specification for Level 3 Version 2 Core Release 2. *J. Integr. Bioinform*. 2019;16(2):20190021. DOI 10.1515/jib-2019-0021

Kuś M., Skowron B. (Eds.) Category Theory in Physics, Mathematics, and Philosophy, Springer Proceedings in Physics. Cham: Springer, 2019. DOI 10.1007/978-3-030-30896-4

MongoDB: The Developer Data Platform [WWW Document], n.d. MongoDB. URL https://www.mongodb.com (accessed 9.19.23)

Novère N.L., Finney A., Hucka M., Bhalla U.S., Campagne F., Collado-Vides J., Crampin E.J., Halstead M., Klipp E., Mendes P., Nielsen P., Sauro H., Shapiro B., Snoep J.L., Spence H.D., Wanner B.L. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol*. 2005;23(12):1509-1515. DOI 10.1038/nbt1156

Novère N.L., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., Wegner K., Aladjem M.I., Wimalaratne S.M., Bergman F.T., Gauges R., Ghazal P., Kawaji H., Li L., Matsuoka Y., Villéger A., Boyd S.E., Calzone L., Courtot M., Dogrusoz U., Freeman T.C., Funahashi A., Ghosh S., Jouraku A., Kim S., Kolpakov F., Luna A., Sahle S., Schmidt E., Watterson S., Wu G., Goryanin I., Kell D.B., 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

Olivier B.G., Snoep J.L. Web-based kinetic modelling using JWS Online. *Bioinformatics*. 2004;20(13):2143-2144. DOI 10.1093/bioinformatics/bth200

Petzold A., Asmi A., Vermeulen A., Pappalardo G., Bailo D., Schaap D., Glaves H.M., Bundke U., Zhao Z. ENVRI-FAIR-interoperable environmental FAIR data and services for society, innovation and research. In: 15th International Conference on eScience (eScience), San Diego, CA, USA, 2019. IEEE, 2019;277-280. DOI 10.1109/eScience.2019.00038

PostgreSQL: the world's most advanced open source database [WWW Document], n.d. URL https://www.postgresql.org/

Rad B.B., Bhatti H.J., Ahmadi M. An introduction to Docker and analysis of its performance. *Int. J. Comput. Sci. Netw. Secur.* 2017;17(3):228-235

Rocca-Serra P., Brandizi M., Maguire E., Sklyar N., Taylor C., Begley K., Field D., Harris S., Hide W., Hofmann O., Neumann S., Sterk P., Tong W., Sansone S.-A. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*. 2010;26(18):2354-2356. DOI 10.1093/bioinformatics/btq415

Roche D.G., Kruuk L.E.B., Lanfear R., Binning S.A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol*. 2015;13(11):e1002295. DOI 10.1371/journal.pbio.1002295

Schreiber F., Bader G.D., Golebiewski M., Hucka M., Kormeier B., Novère N.L., Myers C., Nickerson D., Sommer B., Waltemath D., Weise S. Specifications of standards in systems and synthetic biology. *J. Integr. Bioinform*. 2015;12(2):1-3. DOI 10.1515/jib-2015-258

Software OpenLink. Virtuoso Open-Source Edition: Building. 2022. URL https://github.com/openlink/virtuoso-opensource

Spivak D.I., Kent R.E. Ologs: a categorical framework for knowledge representation. *PLoS One*. 2012;7(1):e24274. DOI 10.1371/journal.pone.0024274

The Univalent Foundations Program. Homotopy Type Theory: Univalent Foundations of Mathematics. Princeton, NJ: Institute for Advanced Study, 2013

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.W., da Silva Santos L.B., Bourne P.E., … van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J., Mons B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*. 2016;3:160018. DOI 10.1038/sdata.2016.18

Wolstencroft K., Owen S., Krebs O., Nguyen Q., Stanford N.J., Golebiewski M., Weidemann A., Bittkowski M., An L., Shockley D., Snoep J.L., Mueller W., Goble C. SEEK: a systems biology data and model management platform. *BMC Syst. Biol*. 2015;9:33. DOI 10.1186/s12918-015-0174-y

Yan Y., Yan J. Hands-On Data Science with Anaconda: Utilize the right mix of tools to create high-performance data science applications. Packt Publishing Ltd., 2018

Zeeberg B.R., Riss J., Kane D.W., Bussey K.J., Uchio E., Linehan W.M., Barrett J.C., Weinstein J.N. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*. 2004;5:80. DOI 10.1186/1471-2105-5-80

ORCID ID
A.M. Mukhin orcid.org/0000-0002-1102-0934
F.V. Kazantsev orcid.org/0000-0002-5711-7539
S.A. Lashin orcid.org/0000-0003-3138-381X

Благодарности. Работа выполнена при поддержке Курчатовского геномного центра Института цитологии и генетики СО РАН (№ 075-15-2019-1662).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 13.07.2023. После доработки 28.09.2023. Принята к публикации 29.09.2023.

Вавиловский журнал генетики и селекции. 2023;27(7):906-915
DOI 10.18699/VJGB-23-105

Перевод на английский язык https://vavilov.elpub.ru/jour

Анализ транскрипционной активности модельных piggyBac-трансгенов, стабильно интегрированных в разные локусы генома культивируемых клеток CHO при отсутствии селекционного давления

Л.А. Яринич, А.А. Огиенко, А.В. Пиндюрин, Е.С. Омелина ✉

Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ omelina@mcb.nsc.ru

Аннотация. Культивируемые клетки яичника китайского хомячка (CHO) наиболее часто используются для синтеза рекомбинантных белков в биофармацевтическом производстве. При получении стабильных клеточных линий-продуцентов локус интеграции трансгена в геном оказывает большое влияние на уровень его экспрессии (явление, известное как эффект положения гена). Соответственно, поиск локусов генома, обеспечивающих высокий уровень продукции белков, является актуальной практической задачей. В данной работе мы использовали метод TRIP для исследования влияния локального окружения хроматина на активность трансгенов, встроенных в разные локусы генома культивируемых клеток CHO. С этой целью репортерные конструкции, кодирующие белок eGFP под контролем четырех разных промоторов, были стабильно встроены в геном клеток CHO при помощи транспозона piggyBac. При этом каждый отдельный трансген содержал уникальную метку – ДНК-штрихкод. Полученная трансгенная поликлональная популяция клеток была культивирована в течение месяца без какой-либо селекции. Далее при помощи присутствующих в конструкциях штрихкодов и высокопроизводительного секвенирования были определены сайты локализации трансгенов в геноме, измерена их представленность в популяции, а также транскрипционная активность. Всего удалось полностью охарактеризовать около 640 трансгенов, более-менее равномерно распределенных по всем хромосомам клеток CHO. Более половины трансгенов оказались полностью молчащими. Наиболее активные трансгены выявлены в окрестностях геномных сайтов инициации транскрипции – в промоторных и 5′-некодирующих районах генов. Наибольшей активностью обладали трансгены, несущие полноразмерный промотор гена *EF-1α* китайского хомячка. Трансгены с укороченным вариантом этого же промотора, а также трансгены с промотором мышиного гена *PGK* (m*PGK*) были соответственно в среднем в 10 и 19 раз менее активны. В целом в результате данной работы выявлены сочетания локусов генома культивируемых клеток CHO и промоторных элементов, которые обеспечивают разные уровни транскрипционной активности модельной репортерной конструкции.
Ключевые слова: TRIP; штрихкод; эффект положения гена; трансген; хроматин; транскрипция.

Для цитирования: Яринич Л.А., Огиенко А.А., Пиндюрин А.В., Омелина Е.С. Анализ транскрипционной активности модельных piggyBac-трансгенов, стабильно интегрированных в разные локусы генома культивируемых клеток CHO при отсутствии селекционного давления. *Вавиловский журнал генетики и селекции*. 2023;27(7):906-915. DOI 10.18699/VJGB-23-105

Analysis of the transcriptional activity of model piggyBac transgenes stably integrated into different loci of the genome of CHO cells in the absence of selection pressure

L.A. Yarinich, A.A. Ogienko, A.V. Pindyurin, E.S. Omelina ✉

Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ omelina@mcb.nsc.ru

Abstract. CHO cells are most commonly used for the synthesis of recombinant proteins in biopharmaceutical production. When stable producer cell lines are obtained, the locus of transgene integration into the genome has a great influence on the level of its expression. Therefore, the identification of genomic loci ensuring a high level of protein production is very important. Here, we used the TRIP assay to study the influence of the local chromatin environment on the activity of transgenes in CHO cells. For this purpose, reporter constructs encoding eGFP under the control of four promoters were stably integrated into the genome of CHO cells using the piggyBac transposon. Each individual transgene contained a unique tag, a DNA barcode, and the resulting polyclonal cell population was cultured for almost a month without any selection. Next, using the high-throughput sequencing, genomic localizations of barcodes, as well as their abundances in the population and transcriptional activities were identified. In total,

~640 transgenes more or less evenly distributed across all chromosomes of CHO cells were characterized. More than half of the transgenes were completely silent. The most active transgenes were identified to be inserted in gene promoters and 5′ UTRs. Transgenes carrying Chinese hamster full-length promoter of the *EF-1α* gene showed the highest activity. Transgenes with a truncated version of the same promoter and with the mouse *PGK* gene promoter were on average 10 and 19 times less active, respectively. In total, combinations of genomic loci of CHO cells and transgene promoters that together provide different levels of transcriptional activity of the model reporter construct were described.
Key words: TRIP; barcode; chromatin position effect; transgene; chromatin; transcription.

For citation: Yarinich L.A., Ogienko A.A., Pindyurin A.V., Omelina E.S. Analysis of the transcriptional activity of model piggyBac transgenes stably integrated into different loci of the genome of CHO cells in the absence of selection pressure. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):906-915. DOI 10.18699/VJGB-23-105

Введение

Метод TRIP (thousands of reporters integrated in parallel) позволяет выполнять масштабные исследования влияния локального окружения хроматина на активность трансгенов. Метод основан на использовании ДНК-штрихкодов (далее – просто штрихкодов) и исходно был апробирован на культивируемых эмбриональных стволовых клетках мыши с помощью транспозона piggyBac в качестве средства доставки трансгенов в геном (Akhtar et al., 2013). Штрихкод – это короткая последовательность ДНК (длиной 16–20 п. н.), являющаяся уникальной для каждой копии трансгена, используемой в исследовании. Крайне важно, что штрихкод располагается в пределах транскрибируемой части трансгенов, что обеспечивает его присутствие не только в ДНК, но также в составе молекул мРНК, синтезированных с трансгенов. Соответственно, штрихкод можно применять и для количественного измерения уровня транскрипционной активности трансгенов.

Транспозон piggyBac позволяет эффективно модифицировать различные клеточные линии и организмы (Wilson et al., 2007), в том числе протяженными конструкциями (Ding et al., 2005). Кроме того, транспозон piggyBac характеризуется относительно равномерным профилем встраивания в клеточный геном (Huang et al., 2010). В экспериментах TRIP система для трансгенеза исследуемых клеток состоит из двух плазмидных конструкций: конструкции для экспрессии транспозазы piggyBac, катализирующей встраивание трансгена в случайное место генома, и собственно трансгена – целевой конструкции (состоящей из промотора, репортерного гена, штрихкода и сигнала полиаденилирования), расположенной между обращенными повторами транспозона piggyBac (Akhtar et al., 2014; Lebedev et al., 2019). Котрансфекция клеток такими плазмидными конструкциями позволяет получить поликлональную популяцию трансгенных клеток, в которой каждая индивидуальная встройка трансгена в геноме маркирована уникальной последовательностью штрихкода. После размножения трансфицированных клеток из них выделяют геномную ДНК и тотальную РНК. С помощью образца геномной ДНК идентифицируют сайты локализации трансгенов в геноме и определяют представленность каждого штрихкода в популяции клеток. На основе образца тотальной РНК определяют представленность каждого штрихкода в общей массе транскриптов, синтезированных с трансгенов. Наконец, соотношение представленности каждого штрихкода в молекулах мРНК и его представленности в геноме трансгенных клеток позволяет количественно измерить уровень транскрипционной активности каждого отдельного трансгена (Akhtar et al., 2014).

В данной работе с помощью метода TRIP исследовано влияние локального окружения хроматина на транскрипционную активность трансгенов в клетках яичника китайского хомячка CHO. Культивируемая клеточная линия CHO наиболее часто используется для наработки разнообразных белков (Xu et al., 2023). Несмотря на доступность ряда других культивируемых клеток млекопитающих, таких как клетки почки детеныша хомячка, клетки мышиной миеломы NS0, эмбриональные клетки почки человека (HEK293) и эмбриональные клетки сетчатки глаза человека PerC6, более 70 % всех рекомбинантных терапевтических белков производится в клетках яичника китайского хомячка (Kim et al., 2012; Ritacco et al., 2018; Gupta et al., 2021). Популярность клеток CHO можно объяснить следующими причинами. Во-первых, применение клеток CHO для производства рекомбинантных белков является безопасным, так как клетки CHO невосприимчивы к заражению вирусами человека (Lalonde, Durocher, 2017). Во-вторых, клетки CHO обладают способностью к эффективной посттрансляционной модификации и продуцируют рекомбинантные белки с совместимыми для человека гликоформами (Stach et al., 2019). Наконец, клетки линии CHO обладают высокой скоростью роста и относительно легко адаптируются к росту в суспензии, что является предпочтительной характеристикой для крупномасштабного культивирования в биореакторах (Ritacco et al., 2018; Dahodwala, Lee, 2019). В настоящее время биореакторы объемом более 10 тыс. литров используются для суспензионных культур рекомбинантных клеток CHO, продуцирующих терапевтические антитела (Kim et al., 2012).

Локализация в геноме оказывает большое влияние на уровень экспрессии рекомбинантного гена (явление, известное как эффект положения) (Gierman et al., 2007; Babenko et al., 2010; Ruf et al., 2011; Chen M. et al., 2013; Elgin, Reuter, 2013). Интеграция в неактивный гетерохроматин приводит к незначительной экспрессии трансгена или ее полному отсутствию, тогда как интеграция в активный эухроматин часто делает возможной экспрессию трансгена. Однако просто интеграции в эухроматин может быть недостаточно для обеспечения длительной экспрессии рекомбинантного гена. Экспрессия трансгена в клетках

млекопитающих во многих случаях быстро инактивируется (замалчивается), в частности, вероятно, из-за влияния соседнего конденсированного хроматина.

Интеграция трансгенов в транскрипционно-активные области генома является одной из стратегий, позволяющих избежать их инактивации. Данная работа направлена на анализ транскрипционной активности piggyBac-трансгенов, интегрированных в разные локусы генома культивируемых клеток CHO при отсутствии селекционного давления.

Материалы и методы

Приготовление конструкции pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242. Плазмида pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242 была приготовлена на основе ранее описанной универсальной конструкции (Lebedev et al., 2019). Встройку амплифицировали с использованием праймеров mPGK-EcoRI-F и eGFP-XbaI-R (табл. 1), используя в качестве матрицы плазмиду mPGK-Puro-IRES-eGFP-sNRP-pA (Akhtar et al., 2013). В 50 мкл реакционной смеси добавляли 1 нг матрицы плазмиды, 2.5 е. а. Phusion-полимеразы (ThermoFisher Scientific), по 1 мкл 10 мкМ праймеров и дНТФ до 0.2 мМ. Условия ПЦР: 98 °С в течение 30 с, 35 циклов: 98 °С 10 с, 62 °С 10 с, 72 °С 1 мин, инкубация 10 мин при 72 °С.

Получение генно-инженерных конструкций с различными промоторами генов китайского хомячка. Плазмиду pPB-mPGK-Puro-IRES-eGFP-PI.11-TR.242 гидролизовали по сайтам рестрикции EcoRI, BglII, AgeI. Для получения встроек амплифицировали последовательности промоторов гена *PGK* китайского хомячка и длинного и короткого варианта промотора гена *EF-1α* с использованием праймеров hamPgk1-EcoRI-F и hamPgk1-BglII-R, CHEF-1-v1-EcoRI-F и CHEF-1-v1-BclI-R, CHEF-

Таблица 1. Список праймеров, использованных в работе

Название праймера	Последовательность (5′→3′)
mPGK-EcoRI-F	aaagaattctcgacaattctaccgggtagg
eGFP-XbaI-R	aaatctagaccctccggattacttg
hamPgk1-EcoRI-F	aaagaattcaggtccctggggattcca
hamPgk1-BglII-R	aaaagatctcggtaggatcaagaggctcag
CHEF-1-v1-EcoRI-F	aaagaattccacgttgtgcatagaaacagatgc
CHEF-1-v1-BclI-R	aaatgatcatggttttcacaacaccttaaaaaaaaagttcg
CHEF-1-v2-EcoRI-F	aaagaattcaagcttctgtggatagaaaatgattag
CHEF-1-v2-BclI-R	aaatgatcactgcgttctgacggcaaac
Plasmid-1	ccgcttaattaatccagcttttgttc
pPB-eGFP-PI-6-R	ctcgagctctcgatctctagacc
pPB-eGFP-PI-11-R	ctcactagctcgatctctagacc
pPB-eGFP-PI-16-R	ctcttgtactcgatctctagacc
pPB-eGFP-PI-28-R	ctcctcggctcgatctctagacc
PB-Barcode-PI-6-Gibson-F	gtctagagatcgagagctcgaggN_{18}gagttgtggccggcccttgtg
PB-Barcode-PI-11-Gibson-F	gtctagagatcgagctagtgaggN_{18}gagttgtggccggcccttgtg
PB-Barcode-PI-16-Gibson-F	gtctagagatcgagtacaagaggN_{18}gagttgtggccggcccttgtg
PB-Barcode-PI-28-Gibson-F	gtctagagatcgagccgaggagggN_{18}gagttgtggccggcccttgtg
PB-Gibson-R1	aacaaaagctggattaattaagcggccgcatacgcgtatactagattaaccc
Libr-cDNA-for	gtctcgtgggctcggagatgtgtataagagacaggtcctgctggagttcgtgac
Libr-cDNA-A16-rev	tcgtcggcagcgtcagatgtgtataagagacagcctatggtcgccagggttttcccagtcacaagg
Libr-cDNA-A23-rev	tcgtcggcagcgtcagatgtgtataagagacagtaattgcgcgccagggttttcccagtcacaagg
Libr-P5-for	aatgatacggcgaccaccgagatctacactcgtcggcagcgtc
Libr-P7-rev	caagcagaagacggcatacgagatgtctcgtgggctcgg
PB-outer-F-2	ttttacgcatgattatctttaacgtacgtc
cDNA-ampl-R	cgccagggttttcccagtcacaag
PB-cDNA-fwd-A7	tcgtcggcagcgtcagatgtgtataagagacagagcgagctgtcacaagggccggccacaa
InvPCR-F-Nextera2	gtctcgtgggctcggagatgtgtataagagacaggtacgtcacaatatgattatctttctag

Примечание. N_{18} – случайная 18-буквенная последовательность ДНК-штрихкода.

1-v2-EcoRI-F и CHEF-1-v2-BclI-R (см. табл. 1) соответственно. В 50 мкл реакционной смеси добавляли 50 нг матрицы геномной ДНК, выделенной из культивируемых клеток CHO, 2.5 е. а. Phusion-полимеразы (ThermoFisher Scientific), по 1 мкл 10 мкМ праймеров и дНТФ до 0.2 мМ. Условия ПЦР: 98 °С 30 с, 35 циклов: 98 °С 10 с, 62 °С 10 с, 72 °С 1 мин, инкубация 10 мин при 72 °С.

Штрихкодированные плазмидные библиотеки были приготовлены в соответствии с ранее описанным протоколом (Lebedev et al., 2019) с помощью метода бесшовного клонирования по Гибсону. Для этого с помощью ПЦР были получены векторы и встройки, содержащие 18-буквенный ДНК-штрихкод и промоторный индекс. Для амплификации векторов использовали праймеры Plasmid-1 и pPB-eGFP-PI-6-R / pPB-eGFP-PI-11-R / pPB-eGFP-PI-16-R / pPB-eGFP-PI-28-R (см. табл. 1) соответственно для конструкций с промотором гена *PGK* китайского хомячка / промотором гена m*PGK* / коротким вариантом промотора гена *EF-1α* / длинным вариантом промотора гена *EF-1α*. Для амплификации штрихкодированных встроек для получения конструкций с промотором гена *PGK* китайского хомячка / промотором гена m*PGK* / коротким вариантом промотора гена *EF-1α* / длинным вариантом промотора гена *EF-1α* использовали праймеры PB-Gibson-R1 и PB-Barcode-PI-6-Gibson-F / PB-Barcode-PI-11-Gibson-F / PB-Barcode-PI-16-Gibson-F / PB-Barcode-PI-28-Gibson-F (см. табл. 1) соответственно. В 50 мкл реакционной смеси добавляли 1 нг матрицы, 2.5 е. а. Phusion-полимеразы (ThermoFisher Scientific), по 1 мкл 10 мкМ праймеров и дНТФ до 0.2 мМ. Условия ПЦР: 98 °С 30 с, 35 циклов: 98 °С 10 с, 62 °С 10 с, 72 °С 1 мин, инкубация 10 мин при 72 °С. После очистки 200 нг «вектора» и 135 нг «встроек» смешивали с 10 мкл 2× NEBuilder HiFi DNA Assembly Master в суммарном объеме 20 мкл. Лигирование ДНК и трансформацию бактерий выполняли в соответствии с ранее описанным протоколом (Lebedev et al., 2019). Штрихкодированные плазмидные библиотеки выделяли с помощью набора Mega Plasmid Kit (Qiagen).

Получение поликлональной трансгенной популяции клеток CHO. За 24 ч до трансфекции культивируемые клетки CHO-S (далее – клетки CHO; любезно предоставлены лабораторией иммуногенетики ИМКБ СО РАН) рассаживали в 12-луночный культуральный планшет в концентрации $1.5 \cdot 10^5$ клеток на лунку в 1 мл среды IMDM с добавлением 10 % сыворотки крупного рогатого скота. Проводили котрансфекцию клеток смесью плазмидных штрихкодированных библиотек (3 мкг) и плазмидой pRP[Exp]-mCherry-CAG>hyPBase (VectorBuilder #VB160216-10057; любезно предоставлена проф. В.В. Верхушей, Медицинский колледж им. Альберта Эйнштейна, Бронкс, Нью-Йорк, США) (0.3 мкг) с использованием реагента X-tremeGENE HP DNA transfection reagent (Roche). Трансгенную популяцию клеток культивировали в течение месяца при отсутствии селекционного давления.

Выделение геномной ДНК. Геномную ДНК выделяли из $5 \cdot 10^7$ клеток полученной поликлональной трансгенной популяции с помощью набора PureLink® Genomic DNA Kit (Invitrogen) в соответствии с рекомендациями производителя.

Выделение тотальной РНК, обратная транскрипция. Тотальную РНК выделяли из $5 \cdot 10^7$ клеток полученной поликлональной трансгенной популяции с помощью реагента RNAzol RT (Molecular Research Center) в соответствии с рекомендациями производителя. Выделенную РНК инкубировали с 20 е. а. эндонуклеазы DpnI (New England Biolabs) и 3 е. а. ДНКазы I (ThermoFisher Scientific) 30 мин при 37 °С. Для очистки РНК использовали набор CleanRNA Standard («Евроген»). 2 мкг тотальной РНК смешивали с 1 мкл 50 мМ олиго(dT) праймера в суммарном объеме 13.5 мкл, смесь инкубировали в течение 5 мин при 65 °С. Последующую реакцию обратной транскрипции проводили в объеме 20 мкл со следующими компонентами: 13.5 мкл матрицы РНК с отожженными праймерами, 4 мкл 5× буфера RT (ThermoFisher Scientific), 1 мкл 10 мМ дНТФ, 1 мкл RNaseOUT (ThermoFisher Scientific), 100 е. а. обратной транскриптазы RevertAid (ThermoFisher Scientific). Смесь инкубировали 60 мин при 42 °С, инактивировали фермент в течение 10 мин при 70 °С.

Приготовление образцов нормирования и экспрессии. Проводили два раунда ПЦР. Для первого раунда амплификации использовали: 600 нг матрицы геномной ДНК (для образцов нормирования) или 3 мкл кДНК (для образцов экспрессии), 0.5 мкл 10 мкМ праймеров Libr-cDNA-for и Libr-cDNA-A16-rev / Libr-cDNA-A23-rev (см. табл. 1) соответственно для образцов нормирования/экспрессии), 2 мкл 2.5 мМ дНТФ, 1.25 е. а. Phusion HS II ДНК-полимеразы (ThermoFisher Scientific), 5 мкл 5× буфера Phusion HF (ThermoFisher Scientific) и бидистиллированную воду до конечного объема 25 мкл. Условия первого раунда ПЦР: 98 °С 1 мин, 15 циклов: 98 °С 30 с, 70 °С 30 с, 72 °С 30 с, инкубация 5 мин при 72 °С. Второй раунд амплификации проводили в объеме 25 мкл со следующими компонентами: 0.5 мкл ПЦР-продукта первого раунда, по 0.25 мкл 10 мкМ праймеров Libr-P5-for и Libr-P7-rev (см. табл. 1), 2 мкл 2.5 мМ дНТФ, 1.25 е. а. Phusion HotStart II ДНК-полимеразы (ThermoFisher Scientific), 5 мкл 5× буфера Phusion HighFidelity (ThermoFisher Scientific). Условия второго раунда ПЦР: 98 °С 1 мин, 23 цикла: 98 °С 30 с, 61 °С 30 с, 72 °С 30 с, инкубация 5 мин при 72 °С.

Приготовление образцов картирования. 2 мкг геномной ДНК инкубировали с 10 е. а. эндонуклеазы рестрикции DpnII (New England Biolabs) при 37 °С в течение 16 ч, затем очищали с помощью набора реагентов GeneJET PCR Purification Kit (ThermoFisher Scientific). 600 нг фрагментированной геномной ДНК смешивали с 4 мкл 100 мМ АТФ, 2.5 е. а. T4 ДНК-лигазы («Евроген») в суммарном объеме 400 мкл. Лигазную смесь инкубировали 2 ч при комнатной температуре и 16 ч при 4 °С, лигазу инактивировали при 65 °С в течение 10 мин. К реакции лигирования добавляли 100 мкл бидистиллированной воды и 500 мкл раствора фенол : хлороформ (в соотношении 1:1), перемешивали, центрифугировали при комнатной температуре 5 мин 10000 g, переносили верхнюю фазу в новую пробирку. К полученному раствору добавляли 1/10 объема 3М NaOAc (pH 5.5), 2.5 объема 96 % этилового спирта, инкубировали 2 ч при –70 °С, центрифугировали 30 мин при 4 °С, 14000 об/мин. Удаляли супернатант, осадок промывали 750 мкл охлажденного 70 % этилового спирта, центрифугировали 10 мин при 4 °С, 14000 об/мин.

Удаляли супернатант, осадок высушивали 15 мин при 37 °C, ДНК растворяли в 30 мкл бидистиллированной воды.

Для приготовления образцов картирования проводили три раунда ПЦР. Для первого раунда амплификации использовали: 5 мкл очищенной лигазной смеси, по 0.5 мкл 10 мкМ праймеров PB-outer-F-2 и cDNA-ampl-R (см. табл. 1), 2 мкл 2.5 мМ дНТФ, 1.25 е. а. Phusion HS II ДНК-полимеразы (ThermoFisher Scientific), 5 мкл 5× буфера Phusion HighFidelity (ThermoFisher Scientific) и бидистиллированную воду до конечного объема 25 мкл. Второй и третий раунды амплификации проводили в том же составе с использованием 10 мкМ праймеров PB-cDNA-fwd-A7 и InvPCR-F-Nextera2 (см. табл. 1) для второго раунда, Libr-P5-for и Libr-P7-rev (см. табл. 1) для третьего раунда, в качестве матрицы использовали 1 мкл смеси ПЦР первого и второго раундов соответственно. Условия первого раунда ПЦР: 98 °C 1 мин, 12 циклов: 98 °C 30 с, 65 °C 30 с, 72 °C 2 мин, инкубация 5 мин при 72 °C. Условия второго раунда ПЦР: 98 °C 1 мин, 12 циклов: 98 °C 30 с, 62 °C 30 с, 72 °C 2 мин, инкубация 5 мин при 72 °C. Условия третьего раунда ПЦР: 98 °C 1 мин, 16 циклов: 98 °C 30 с, 61 °C 30 с, 72 °C 2 мин, инкубация 5 мин при 72 °C. Образец картирования (5 мкг) обрабатывали 10 е. а. эндонуклеазы рестрикции NotI для удаления побочных продуктов в суммарном объеме 100 мкл при 37 °C в течение 2 ч.

Секвенирование и анализ данных. Секвенирование образцов проводили на платформе Genolab 2 × 75 п. о. (https://genomed.ru/). Fastq-файлы были демультиплексированы с помощью инструмента sabre (https://github.com/najoshi/sabre). В результате получено 4.5 млн, 1.6 млн и около 1 млн прочтений для образцов картирования, нормирования и экспрессии соответственно. Анализ качества прочтений fastq файлов для каждого образца выполняли с помощью специального инструмента FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Далее с помощью инструмента TASK (The TRIP Analysis Software Kit, https://trip.nki.nl/) были установлены последовательности достоверно выявляемых штрихкодов, а также их нормированные уровни экспрессии и локализация в версиях генома китайского хомячка CriGri-PICRH-1.0 (GCA_003668045.2) и Cgr1.0 (GCA_000448345.1). Версия генома CriGri-PICRH-1.0 характеризуется наличием очень протяженных последовательностей, соответствующих всем ожидаемым хромосомам клеток CHO, и поэтому именно эта версия генома была использована в работе как основная, тогда как версия генома Cgr1.0 была ранее использована для картирования типов хроматина в клетках CHO-K1 (Feichtinger et al., 2016). Для определения наиболее «надежных» (далее – отфильтрованных) трансгенов дополнительно использовали следующие параметры инструмента TASK: norm > = 5, reads_r > = 10, freq1_r > 0.60. Данные по типам хроматина были взяты для временной точки Tp0, соответствующей 4 ч культивирования клеток (https://cho-epigenome.boku.ac.at/JB/). Позиционная весовая матрица для геномных последовательностей, перекрывающих сайты инсерции трансгенов, была построена с помощью специального приложения pLogo (https://plogo.uconn.edu/) (O'Shea et al., 2013).

Результаты и обсуждение

Чтобы изучить активность нескольких промоторных элементов в разных локальных окружениях хроматина в культивируемых клетках CHO, на основе транспозона piggyBac были сконструированы штрихкодированные модельные трансгены, несущие ген устойчивости к пуромицину (далее – $Puro^R$) и ген улучшенного зеленого флуоресцентного белка eGFP под контролем четырех следующих промоторов: 1) промотора гена *PGK* мыши (m*PGK*), использованного ранее для аналогичного исследования на культивируемых мышиных эмбриональных стволовых клетках (Akhtar et al., 2013); 2) промотора гена *PGK* китайского хомячка, гомологичного промотору m*PGK*; 3) полноразмерного (long) и 4) усеченного (short) вариантов промотора гена *EF-1α* китайского хомячка (Running Deer, Allison, 2004; Orlova et al., 2014; Wang et al., 2017) (рис. 1, *А*). При этом в конструкциях с каждым отдельным промотором непосредственно перед 18-буквенным штрихкодом присутствовал также специфический 5-буквенный мотив (промоторный индекс), позволяющий одновременно использовать все четыре штрихкодированных модельных трансгена в одном эксперименте (Gisler et al., 2019) (см. рис. 1, *А*). Полученные штрихкодированные плазмидные библиотеки с промоторами long *EF-1α*, short *EF-1α*, m*PGK* и *PGK* были смешаны в молярных пропорциях 7:7:7:1. Меньшая доля конструкции с промотором *PGK* объясняется ее использованием в данном эксперименте в качестве контроля. Мы также использовали эту конструкцию для получения стабильных трансгенных популяций клеток CHO при селекции пуромицином (результаты исследования будут сообщены отдельно), и нам представлялось полезным иметь в будущем техническую возможность для корректного сравнения данных для таких разных трансгенных популяций.

Культивируемые клетки CHO (сублинии CHO-S) были котрансфицированы вышеописанной смесью модельных трансгенов, а также плазмидой, кодирующей транспозазу piggyBac. Спустя 72 ч после трансфекции экспрессия белка eGFP наблюдалась примерно в 40 % клеток (см. рис. 1, *Б*). После этого трансфицированные клетки культивировали в отсутствие какой-либо селекции еще 25 дней с целью размножить трансгенные клетки и параллельно избавиться от молекул плазмидной ДНК, которые могут загрязнить интересующие нас данные. Действительно, в результате в популяции наблюдались множественные клоны трансгенных клеток (см. рис. 1, *Б*).

Из полученной поликлональной популяции клеток были выделены геномная ДНК и тотальная РНК, на основе которых определены геномные локализации и нормированные уровни экспрессии штрихкодированных трансгенов. Всего в трансгенной популяции выявлен 641 уникально штрихкодированный и картированный в геноме трансген. Эти трансгены присутствовали в более-менее ожидаемых количествах на всех хромосомах клеток CHO (см. рис. 1, *В*). Анализ геномных последовательностей, перекрывающих сайты инсерции трансгенов, выявил их АТ-обогащенность, а также присутствие центрального мотива ttaa (см. рис. 1, *Г*) – черты, характерные для транспозона piggyBac (Fraser et al., 1996; Li et al., 2013; Chen Q. et al., 2020).

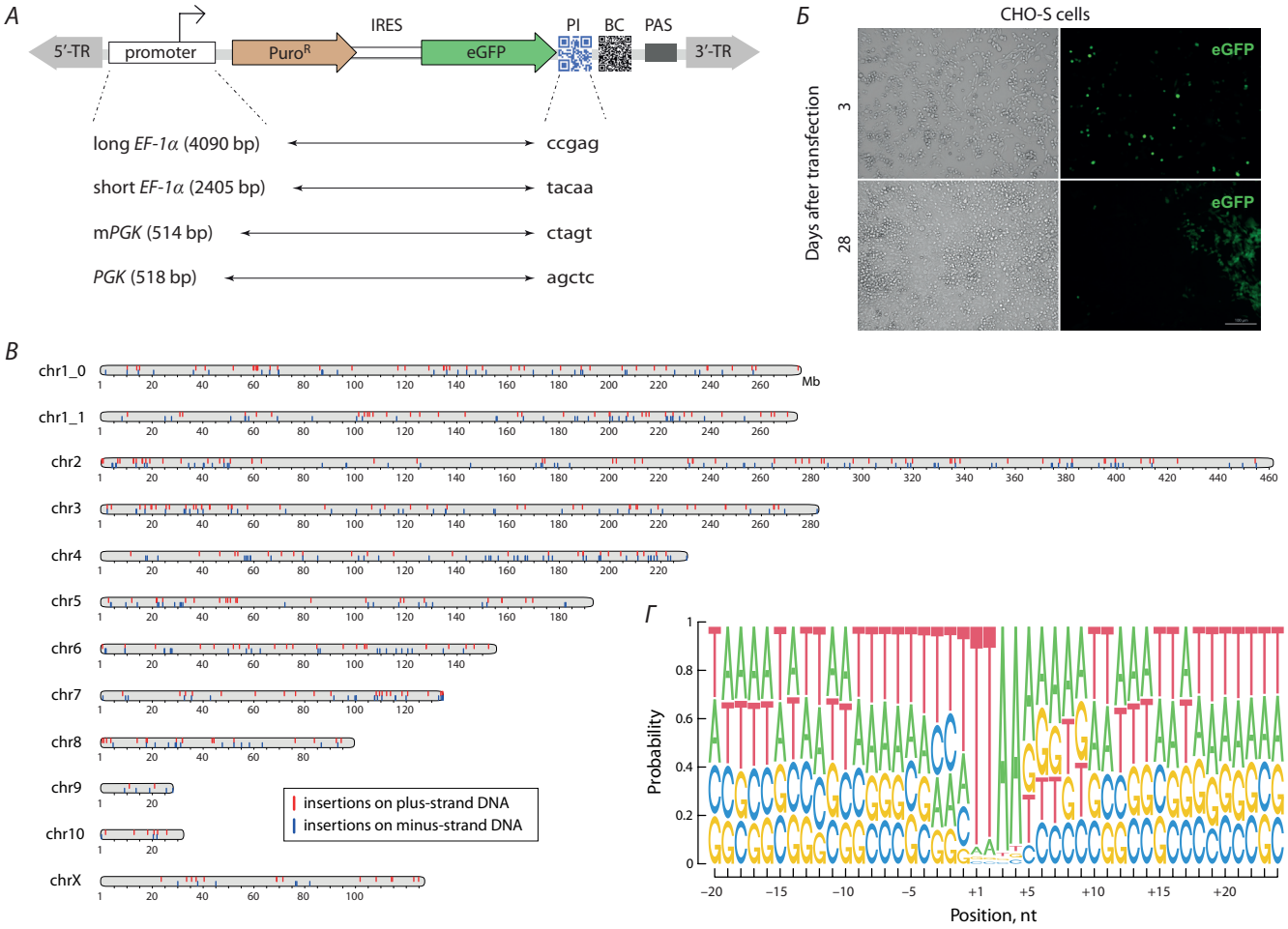


Рис. 1. Стабильная интеграция модельных piggyBac-трансгенов в геном культивируемых клеток CHO.

A – схема штрихкодированных репортерных конструкций, использованных в работе. 5′-TR и 3′-TR – обращенные концы транспозона piggyBac; IRES – участок внутренней посадки рибосомы; PI – промоторный индекс; BC – штрихкод; PAS – сигнал полиаденилирования. *Б* – клетки CHO спустя 3 дня и 28 дней после трансфекции. *B* – распределение всех уникально картированных трансгенов по хромосомам китайского хомячка. Красными и синими черточками обозначены встройки трансгенов по плюс- и минус-цепи ДНК соответственно. *Г* – анализ мотивов генома, по которым произошла интеграция всех уникально картированных трансгенов. Позиции от +1 до +4 соответствуют последовательности, которая при встройке транспозона piggyBac дуплицируется и фланкирует интегрированный трансген.

Среди выявленных трансгенов 38.8 % оказались с промоторным индексом tacaa (соответствующим промотору short *EF-1α*), 24.3 % – с промоторным индексом ccgag (соответствующим промотору long *EF-1α*), 32.2 % – с промоторным индексом ctagt (соответствующим промотору m*PGK*), 4.7 % – с промоторным индексом agctc (соответствующим промотору *PGK* китайского хомячка) (рис. 2, *A*).

Анализ активности репортерных конструкций, находящихся под контролем четырех разных промоторов, выявил наличие большого числа молчащих (т. е. транскрипционно неактивных) трансгенов с каждым промотором (табл. 2), что, наиболее вероятно, связано с отсутствием селекции антибиотиком при получении поликлональной популяции трансгенных клеток CHO.

Сравнение активностей промоторов среди отфильтрованных экспрессирующихся трансгенов (144 шт.) показало, что основная часть высокоактивных репортерных конструкций находится под контролем полноразмерного варианта промотора гена *EF-1α* (см. рис. 2, *Б*, табл. 2). В частности, среди 10 % наиболее активных отфильтрованных трансгенов доли промоторов распределяются следующим образом: long *EF-1α* – 70 %, m*PGK* – 20 %, short *EF-1α* – 10 %, *PGK* – 0 %. Надо отметить, что из-за малого количества исследованных трансгенов с промотором *PGK* результаты по его активности носят очень предварительный характер.

Две трети всех трансгенов встроились в геном клеток CHO внутри генов (которые были определены как –1000 п. н. от дистального сайта инициации транскрипции до сайта терминации транскрипции), причем преимущественно в интроны (42.3 %), промоторы (8.5 %) и 5′-некодирующие области (9.4 %) (см. рис. 2, *B*). Необходимо отметить, что промоторы были определены как районы от –1000 до +100 п. н. относительно сайтов инициации транскрипции. Схожие паттерны интеграции трансгенов на основе транспозона piggyBac были описаны ранее для культивируемых клеток других видов (Ding et al., 2005; Wilson et al., 2007; Galvan et al., 2009; Li et al., 2013). Анализ 10 % наиболее активных отфильтрованных

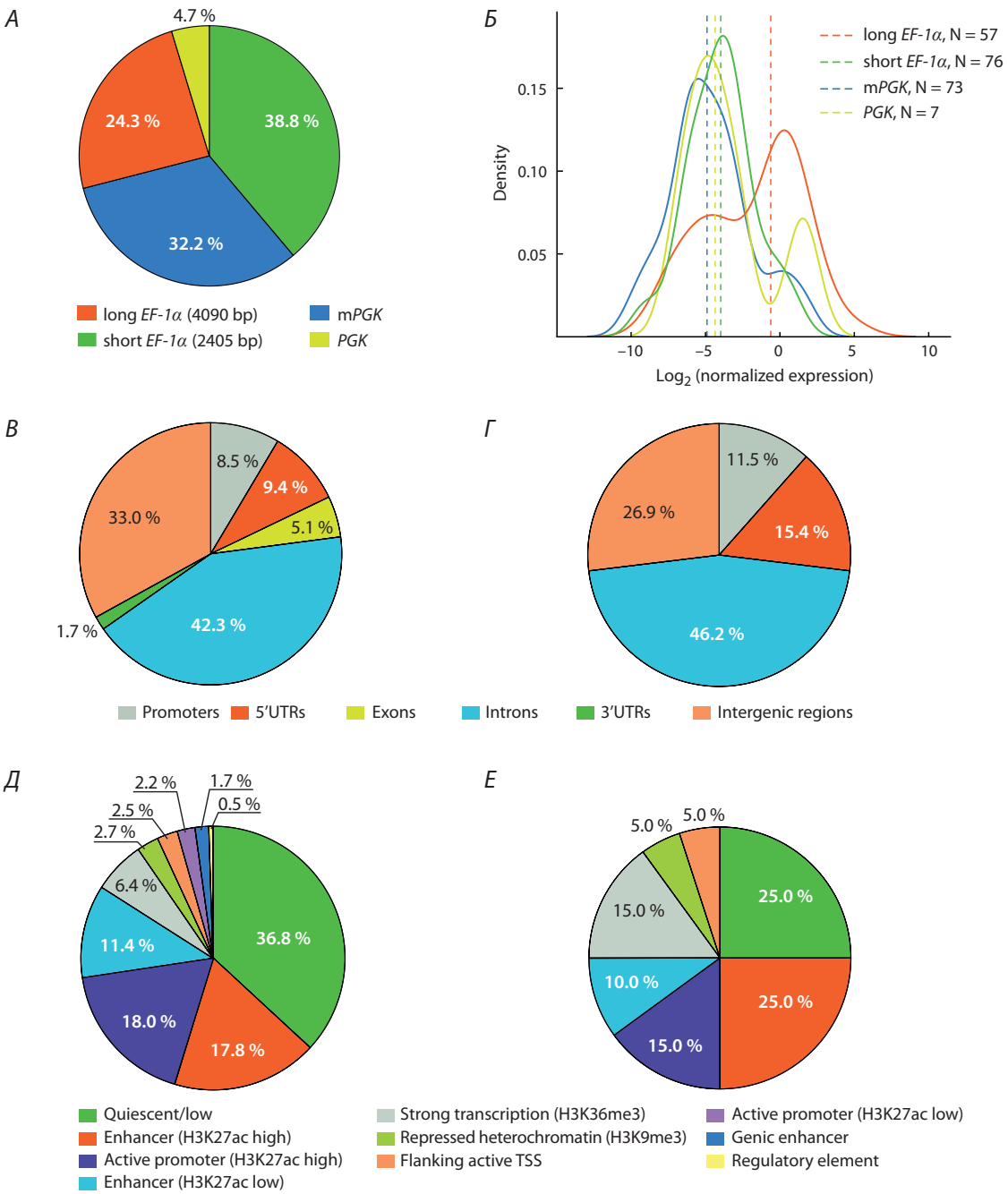


Рис. 2. Характеристика исследованных трансгенов.

А – распределение всех выявленных трансгенов по исследуемым промоторам long *EF-1α*, short *EF-1α*, m*PGK* и *PGK*. *Б* – сравнение активностей промоторов на основе данных для отфильтрованных 144 экспрессирующихся трансгенов (см. Материалы и методы). Штриховыми вертикальными линиями показаны медианные значения нормированной экспрессии для каждого промотора. *В*, *Г* – распределение всех выявленных трансгенов (*В*) и 10 % наиболее активных отфильтрованных трансгенов (*Г*) по элементам генов (промоторам, 5′- и 3′-некодирующим районам, экзонам, интронам), а также межгенным промежуткам. *Д*, *Е* – распределение всех трансгенов (*Д*) и 10 % наиболее активных отфильтрованных трансгенов (*Е*) по типам хроматина, определенным ранее для клеток CHO-K1 (Feichtinger et al., 2016).

репортерных конструкций (21 шт.) выявил увеличение доли трансгенов как раз именно в 5′-некодирующих областях генов, промоторах и интронах (в 1.6, 1.4 и 1.1 раза соответственно) (см. рис. 2, *Г*). Интересно, что трансгены чаще локализовались ближе к началу, чем к концу генов. Значения медианных расстояний от позиции локализации трансгена в геноме до ближайших сайтов инициации и терминации транскрипции оказались равны соответственно 11.4 и 20.2 т. п. н. для полного набора исследованных репортерных генов (641 шт.). При этом для 10 % наиболее активных отфильтрованных трансгенов (21 шт.) такие значения были равны 6.6 и 17.8 т. п. н. соответственно.

Для изучения влияния локального окружения хроматина на активность репортерных генов были использованы ранее опубликованные данные по распределению 11 типов хроматина в геноме клеток CHO-K1 (Feichtinger

Таблица 2. Сравнительная активность исследованных промоторов

Промотор	Количество трансгенов	Доля трансгенов, %		Значение медианной активности промотора, отн. ед.
		молчащих	активных	
long *EF-1α*	156	42.31	57.69	18.93
short *EF-1α*	249	60.24	39.76	1.84
m*PGK*	206	53.88	46.12	1
PGK	30	66.67	33.33	1.42
Всего	641	54.13	45.87	

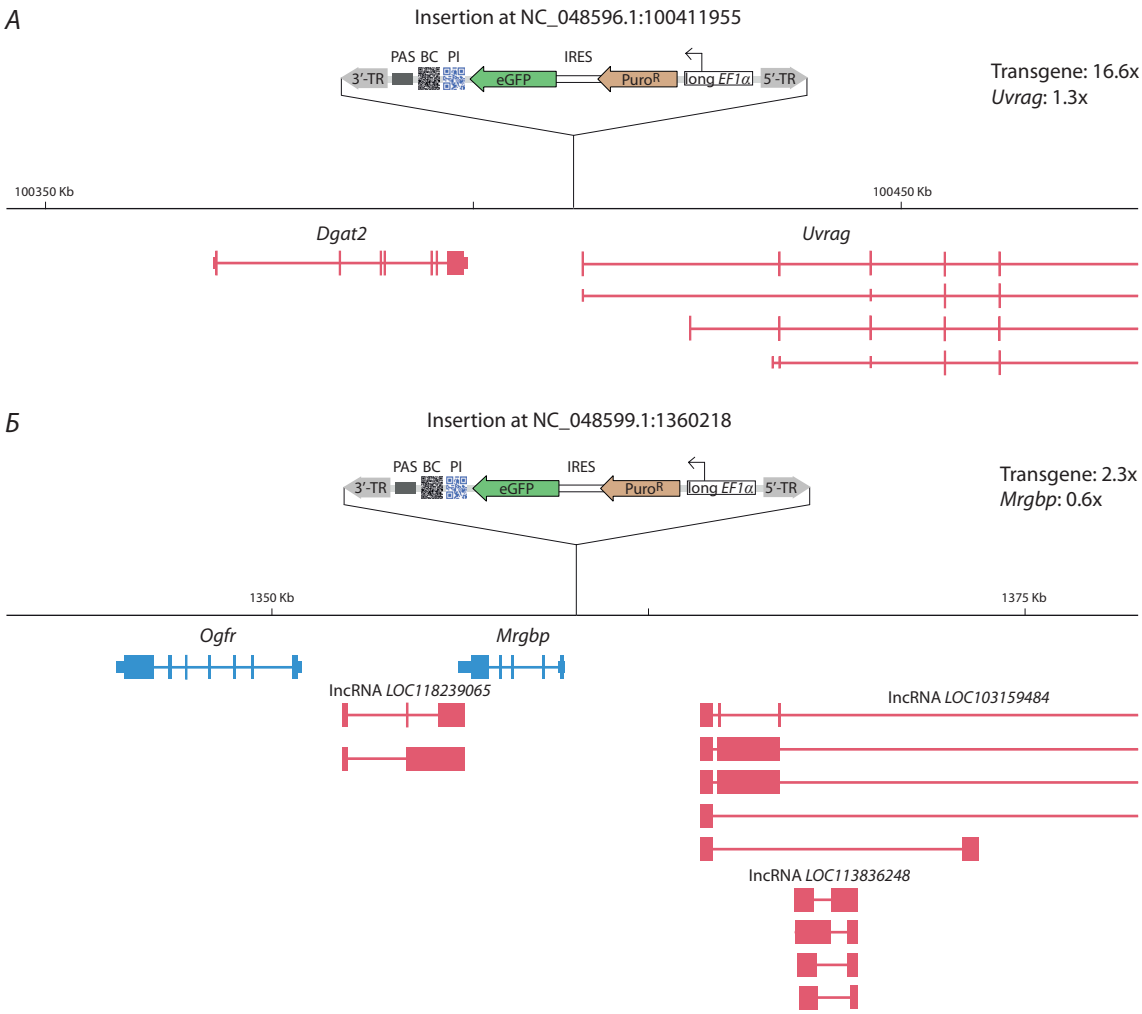


Рис. 3. Примеры геномной локализации трансгенов из числа наиболее активных (*А*) и со средней транскрипционной активностью (*Б*).

Схемы встроек трансгенов не масштабированы относительно геномной ДНК. Активности трансгенов и ближайших к ним генов указаны относительно среднего уровня экспрессии всех изученных штрихкодированных трансгенов и всех эндогенных генов соответственно.

et al., 2016). Изначально эти данные были определены для версии генома китайского хомячка, отличной от использованной для всех описанных выше анализов (подробности см. в разделе Материалы и методы), поэтому только для 595 из 641 трансгена удалось определить типы хроматина, перекрывающие позиции их локализации в геноме. При этом лишь 39.5 % трансгенов оказались расположены в неактивных типах хроматина “Quiescent/low”, “Repressed heterochromatin (H3K9me3)” и “Polycomb repressed regions (H3K27me3)”, которые суммарно покрывают более 88 % генома клеток китайского хомячка (Feichtinger et al., 2016). Остальные 60.5 % трансгенов были выявлены в различных активных типах хроматина (см. рис. 2, *Д*).

Наиболее активные трансгены чаще обнаруживались в районах генома, ассоциированных с активными типами

хроматина “Enhancer (H3K27ac high)”, “Strong transcription (H3K36me3)” и “Flanking active TSS”, а также с неактивным типом хроматина “Repressed heterochromatin (H3K9me3)” (см. рис. 2, *E*). Последнее достаточно неожиданное наблюдение, возможно, связано с тем, что типы хроматина были определены для другой (отличной от использованной в данной работе) сублинии клеток CHO.

Поскольку, как отмечено выше, две трети всех трансгенов локализовались внутри генов (см. рис. 2, *B*), стоит отметить, что встройка трансгена даже в какой-то важный ген, вероятно, лишь незначительно сказывается на жизнеспособности клеток. Это связано как с тем, что далеко не каждая встройка чужеродной последовательности в пределах гена существенно нарушает его функцию, так и с наличием второй нативной копии этого гена в геноме. Вместе это обеспечивает выживание таких трансгенных клеток в поликлональной популяции. Шансы повредить сразу оба аллеля гена в использованном нами экспериментальном подходе ничтожно малы. Для этого сразу два трансгена должны встроиться в геном одной и той же клетки, причем в разные аллели одного и того же гена. Таким образом, выявленные в данной работе геномные позиции активных трансгенов вполне могут претендовать на рассмотрение в качестве перспективных сайтов для направленной интеграции целевых биотехнологических трансгенов, даже если они располагаются внутри активных генов (рис. 3).

Заключение

В полученной при отсутствии селекционного давления поликлональной популяции культивируемых клеток CHO более половины модельных трансгенов, стабильно интегрированных в геном, оказались транскрипционно неактивными. По сравнению с полным набором трансгенов, наиболее активные трансгены локализовались в 1.6 и 1.4 раза чаще в районах промоторов и 5′-некодирующих областей генов соответственно. Также наиболее активные трансгены локализовались в 2.3 и 1.4 раза чаще в транскрипционно активных типах хроматина “Strong transcription (H3K36me3)” и “Enhancer (H3K27ac high)” соответственно. Трансгены, содержащие полноразмерный промотор гена *EF-1α* китайского хомячка, оказались в среднем наиболее активными. При этом медианная активность короткого варианта промотора гена *EF-1α* была в 10 раз ниже медианной активности полноразмерного промотора этого гена (см. табл. 2). Это можно объяснить наличием важных сайтов связывания транскрипционных факторов в полноразмерной версии промотора гена *EF-1α*. Геномные сайты локализации наиболее активных встроек модельных трансгенов могут представлять интерес для дальнейших экспериментов как перспективные позиции для направленной интеграции целевых биотехнологических конструкций.

Список литературы

Akhtar W., de Jong J., Pindyurin A.V., Pagie L., Meuleman W., de Ridder J., Berns A., Wessels L.F.A., van Lohuizen M., van Steensel B. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*. 2013;154(4):914-927. DOI 10.1016/j.cell.2013.07.018

Akhtar W., Pindyurin A.V., de Jong J., Pagie L., ten Hoeve J., Berns A., Wessels L.F.A., van Steensel B., van Lohuizen M. Using TRIP for genome-wide position effect analysis in cultured cells. *Nat. Protoc*. 2014;9(6):1255-1281. DOI 10.1038/nprot.2014.072

Babenko V.N., Makunin I.V., Brusentsova I.V., Belyaeva E.S., Maksimov D.A., Belyakin S.N., Maroy P., Vasil′eva L.A., Zhimulev I.F. Paucity and preferential suppression of transgenes in late replication domains of the *D. melanogaster* genome. *BMC Genomics*. 2010;11:318. DOI 10.1186/1471-2164-11-318

Chen M., Licon K., Otsuka R., Pillus L., Ideker T. Decoupling epigenetic and genetic effects through systematic analysis of gene position. *Cell Rep.* 2013;3(1):128-137. DOI 10.1016/j.celrep.2012.12.003

Chen Q., Luo W., Veach R.A., Hickman A.B., Wilson M.H., Dyda F. Structural basis of seamless excision and specific targeting by *piggyBac* transposase. *Nat. Commun*. 2020;11(1):3446. DOI 10.1038/s41467-020-17128-1

Dahodwala H., Lee K.H. The fickle CHO: a review of the causes, implications, and potential alleviation of the CHO cell line instability problem. *Curr. Opin. Biotechnol*. 2019;60:128-137. DOI 10.1016/j.copbio.2019.01.011

Ding S., Wu X., Li G., Han M., Zhuang Y., Xu T. Efficient transposition of the *piggyBac* (*PB*) transposon in mammalian cells and mice. *Cell*. 2005;122(3):473-483. DOI 10.1016/j.cell.2005.07.013

Elgin S.C.R., Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in *Drosophila*. *Cold Spring Harb. Perspect. Biol*. 2013;5(8):a017780. DOI 10.1101/cshperspect.a017780

Feichtinger J., Hernández I., Fischer C., Hanscho M., Auer N., Hackl M., Jadhav V., Baumann M., Krempl P.M., Schmidl C., Farlik M., Schuster M., Merkel A., Sommer A., Heath S., Rico D., Bock C., Thallinger G.G., Borth N. Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time. *Biotechnol. Bioeng*. 2016;113(10):2241-2253. DOI 10.1002/bit.25990

Fraser M.J., Ciszczon T., Elick T., Bauser C. Precise excision of TTAA-specific lepidopteran transposons *piggyBac* (IFP2) and *tagalong* (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol. Biol*. 1996;5(2):141-151. DOI 10.1111/j.1365-2583.1996.tb00048.x

Galvan D.L., Nakazawa Y., Kaja A., Kettlun C., Cooper L.J.N., Rooney C.M., Wilson M.H. Genome-wide mapping of *PiggyBac* transposon integrations in primary human T cells. *J. Immunother*. 2009;32(8):837-844. DOI 10.1097/CJI.0b013e3181b2914c

Gierman H.J., Indemans M.H.G., Koster J., Goetze S., Seppen J., Geerts D., van Driel R., Versteeg R. Domain-wide regulation of gene expression in the human genome. *Genome Res*. 2007;17(9):1286-1295. DOI 10.1101/gr.6276007

Gisler S., Gonçalves J.P., Akhtar W., de Jong J., Pindyurin A.V., Wessels L.F.A., van Lohuizen M. Multiplexed Cas9 targeting reveals genomic location effects and gRNA-based staggered breaks influencing mutation efficiency. *Nat. Commun*. 2019;10(1):1598. DOI 10.1038/s41467-019-09551-w

Gupta K., Modi D., Jain R., Dandekar P. A stable CHO K1 cell line for producing recombinant monoclonal antibody against TNF-α. *Mol. Biotechnol*. 2021;63(9):828-839. DOI 10.1007/s12033-021-00329-4

Huang X., Guo H., Tammana S., Jung Y.-C., Mellgren E., Bassi P., Cao Q., Tu Z.J., Kim Y.C., Ekker S.C., Wu X., Wang S.M., Zhou X. Gene transfer efficiency and genome-wide integration profiling of *Sleeping Beauty*, *Tol2*, and *piggyBac* transposons in human primary T cells. *Mol. Ther*. 2010;18(10):1803-1813. DOI 10.1038/mt.2010.141

Kim J.Y., Kim Y.-G., Lee G.M. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Appl. Microbiol. Biotechnol*. 2012;93(3):917-930. DOI 10.1007/s00253-011-3758-5

Lalonde M.-E., Durocher Y. Therapeutic glycoprotein production in mammalian cells. *J. Biotechnol*. 2017;251:128-140. DOI 10.1016/j.jbiotec.2017.04.028

Lebedev M.O., Yarinich L.A., Ivankin A.V., Pindyurin A.V. Generation of barcoded plasmid libraries for massively parallel analysis of chromatin position effects. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2019;23(2):203-211. DOI 10.18699/VJ19.483

Li M.A., Pettitt S.J., Eckert S., Ning Z., Rice S., Cadiñanos J., Yusa K., Conte N., Bradley A. The *piggyBac* transposon displays local and distant reintegration preferences and can cause mutations at non-canonical integration sites. *Mol. Cell. Biol.* 2013;33(7):1317-1330. DOI 10.1128/MCB.00670-12

Orlova N.A., Kovnir S.V., Hodak J.A., Vorobiev I.I., Gabibov A.G., Skryabin K.G. Improved elongation factor-1 alpha-based vectors for stable high-level expression of heterologous proteins in Chinese hamster ovary cells. *BMC Biotechnol.* 2014;14:56. DOI 10.1186/1472-6750-14-56

O'Shea J.P., Chou M.F., Quader S.A., Ryan J.K., Church G.M., Schwartz D. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods*. 2013;10(12):1211-1212. DOI 10.1038/nmeth.2646

Ritacco F.V., Wu Y., Khetan A. Cell culture media for recombinant protein expression in Chinese hamster ovary (CHO) cells: history, key components, and optimization strategies. *Biotechnol. Prog*. 2018;34(6):1407-1426. DOI 10.1002/btpr.2706

Ruf S., Symmons O., Uslu V.V., Dolle D., Hot C., Ettwiller L., Spitz F. Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nat. Genet.* 2011;43(4):379-386. DOI 10.1038/ng.790

Running Deer J., Allison D.S. High-level expression of proteins in mammalian cells using transcription regulatory sequences from the Chinese hamster EF-1α gene. *Biotechnol. Prog*. 2004;20(3):880-889. DOI 10.1021/bp034383r

Stach C.S., McCann M.G., O'Brien C.M., Le T.S., Somia N., Chen X., Lee K., Fu H.Y., Daoutidis P., Zhao L., Hu W.S., Smanski M. Model-driven engineering of N-linked glycosylation in Chinese hamster ovary cells. *ACS Synth. Biol*. 2019;8(11):2524-2535. DOI 10.1021/acssynbio.9b00215

Wang X., Xu Z., Tian Z., Zhang X., Xu D., Li Q., Zhang J., Wang T. The EF-1α promoter maintains high-level transgene expression from episomal vectors in transfected CHO-K1 cells. *J. Cell. Mol. Med*. 2017;21(11):3044-3054. DOI 10.1111/jcmm.13216

Wilson M.H., Coates C.J., George A.L., Jr. *PiggyBac* transposon-mediated gene transfer in human cells. *Mol. Ther*. 2007;15(1):139-145. DOI 10.1038/sj.mt.6300028

Xu W.-J., Lin Y., Mi C.-L., Pang J.-Y., Wang T.-Y. Progress in fed-batch culture for recombinant protein production in CHO cells. *Appl. Microbiol. Biotechnol*. 2023;107(4):1063-1075. DOI 10.1007/s00253-022-12342-x

ORCID ID

L.A. Yarinich orcid.org/0000-0003-0469-0371
A.A. Ogienko orcid.org/0000-0002-0896-1899
A.V. Pindyurin orcid.org/0000-0001-6959-0641
E.S. Omelina orcid.org/0000-0002-2189-5101

Благодарности. Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (Соглашение № 075-15-2021-1086, контракт RF----193021X0015, 15.ИП.21.0015).

Авторы благодарны А.В. Таранину (Институт молекулярной и клеточной биологии СО РАН), В.В. Верхуше (Медицинский колледж им. Альберта Эйнштейна, Бронкс, Нью-Йорк, США), В.С. Фишману (Институт цитологии и генетики СО РАН) за предоставление культивируемых клеток линии CHO-S, плазмиды pRP[Exp]-mCherry-CAG>hyPBase и помощь в извлечении данных по распределению типов хроматина в геноме китайского хомячка.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 13.09.2023. После доработки 21.09.2023. Принята к публикации 27.09.2023.

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» – 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 на платформе Web of Science, 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 ИЦиГ СО РАН, 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/

✉ 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.
Фотография на обложке О.В. Андреенкова.