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

doi 10.18699/vjgb-25-99

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Введение

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

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

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

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

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

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

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

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

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

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

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

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

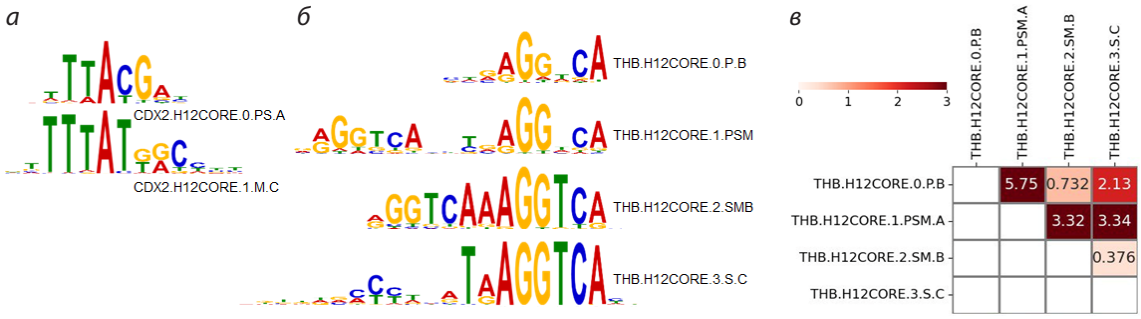

Рис. 1. Сходство разных мотивов сайтов связывания, представляющих отдельные ТФ.

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

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

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

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

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

Материалы и методы

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

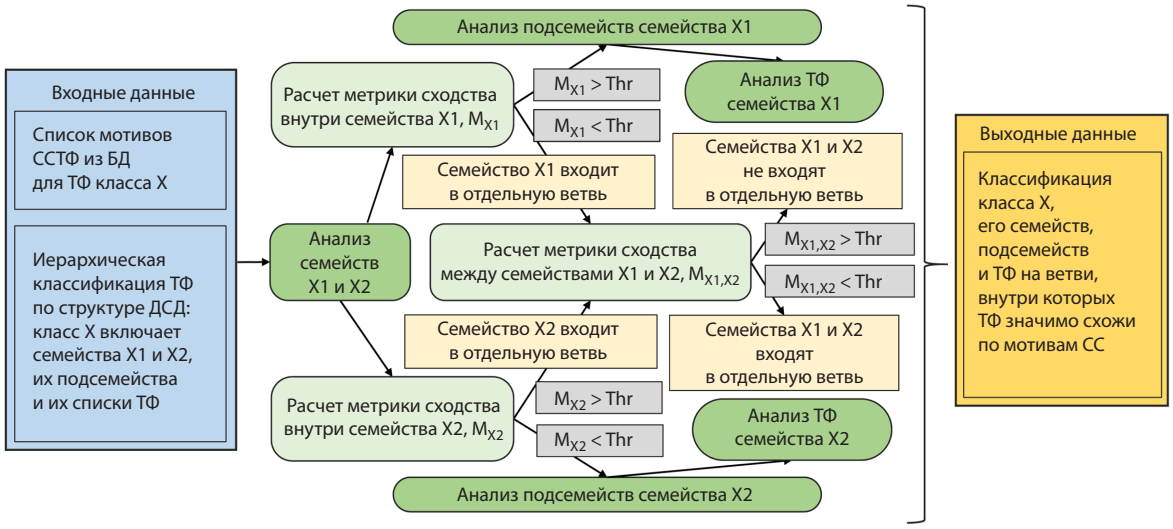

Рис. 2. Схема анализа для определения ветвей схожих мотивов ССТФ. На схеме детально представлен этап анализа одного класса X из двух семейств X1 и X2. Голубой цвет показывает входные данные, темно-зеленый – этапы анализа, светло-зеленый – расчеты метрик сходства, серый – проверки условий сходства мотивов,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

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

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

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

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

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

Ветвь может включать одну или несколько сестринских единиц классификации:

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

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

Результаты

Сходство транскрипционных факторов в сестринских подсемействах одних семейств

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

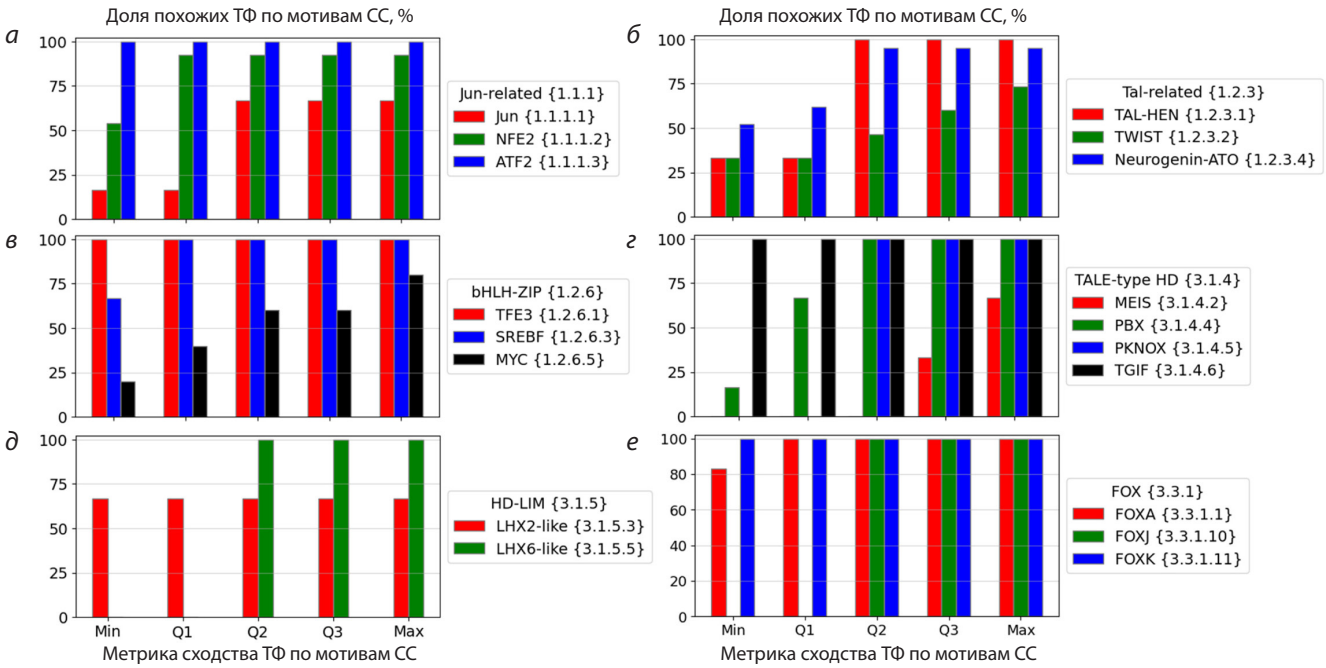

Рис. 3. Доля значимо похожих ТФ по мотивам сайтов связывания для подсемейств разных семейств при использовании пяти метрик сходства: Min, Q1, Q2, Q3 и Max.

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

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

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

Анализ сходства транскрипционных факторов человека

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

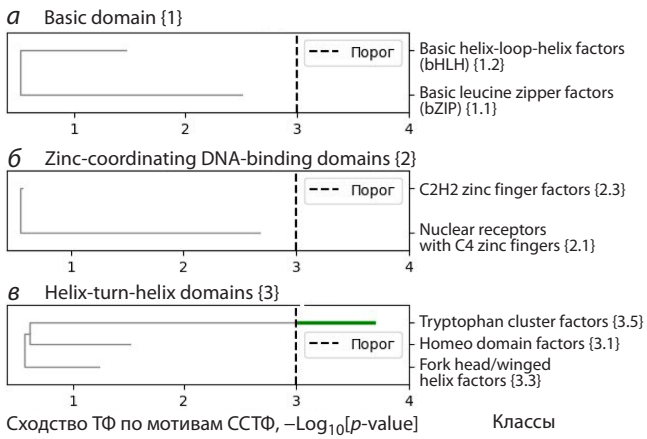

Рис. 4. Сходство ТФ по мотивам сайтов связывания в крупнейших классах трех крупнейших суперклассов человека.

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

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

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

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

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

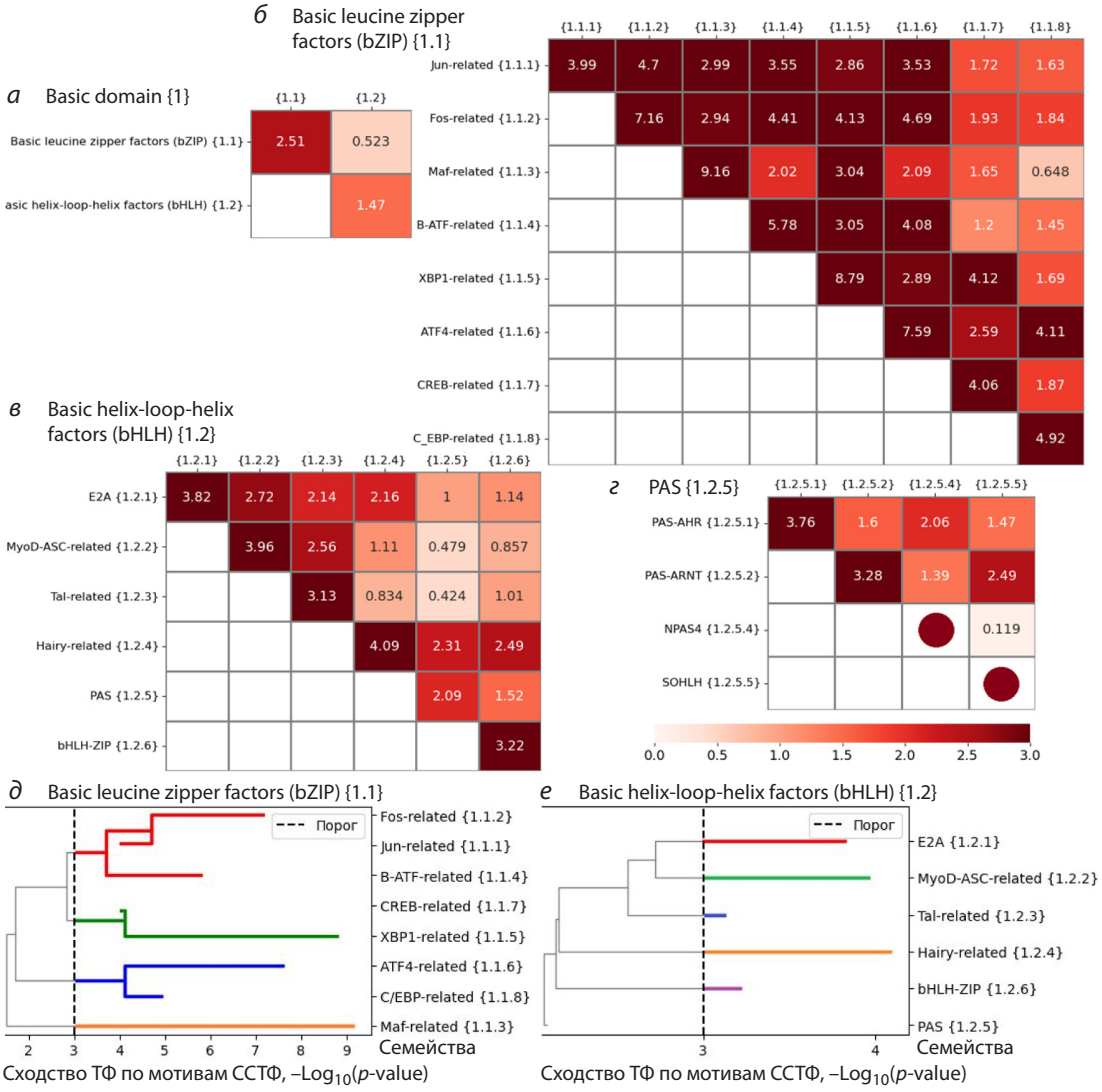

Рис. 5. Сходство ТФ по мотивам сайтов связывания для суперкласса Basic domain {1}.

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

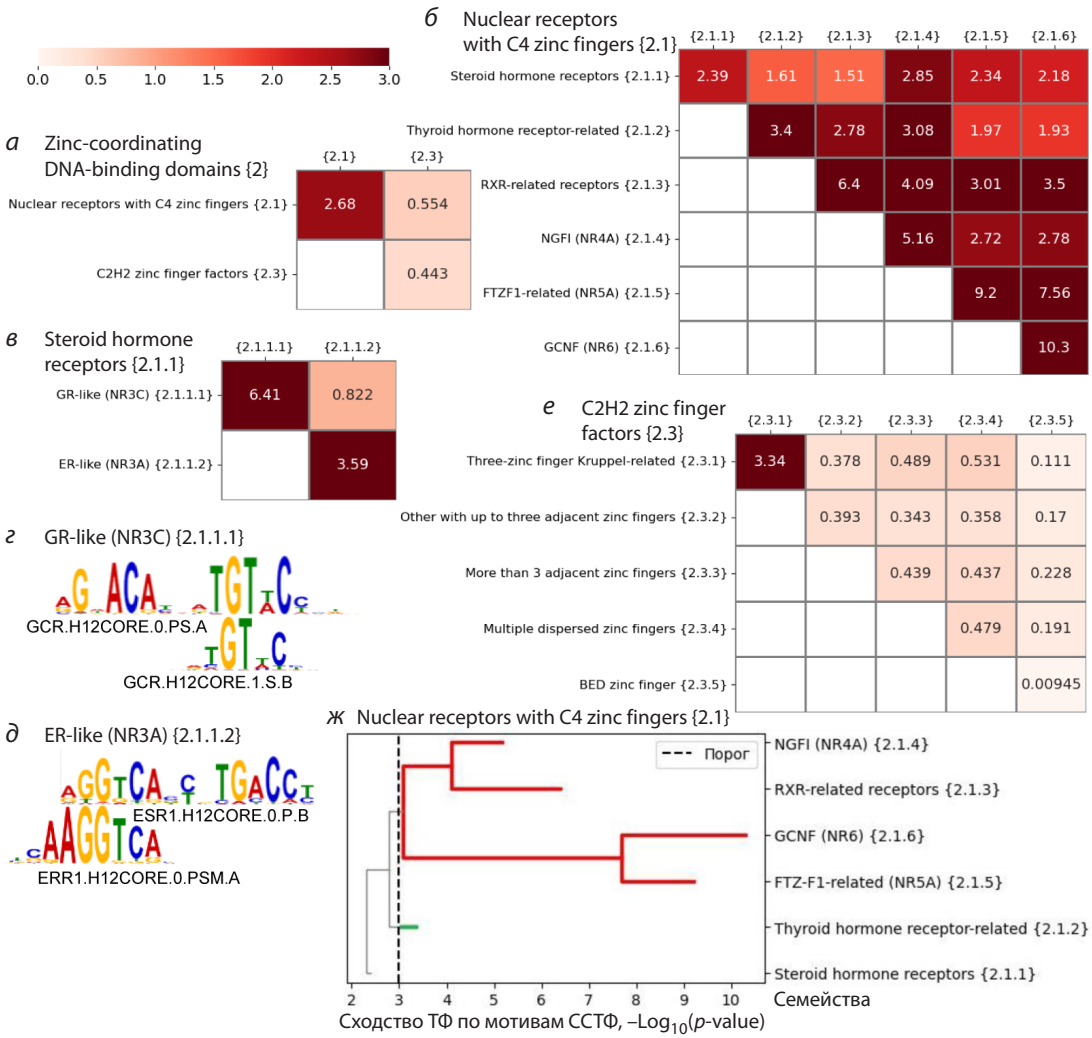

Рис. 6. Сходство ТФ по мотивам сайтов связывания для суперкласса Zinc-coordinating DNA-binding domains {2}.

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

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

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

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

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

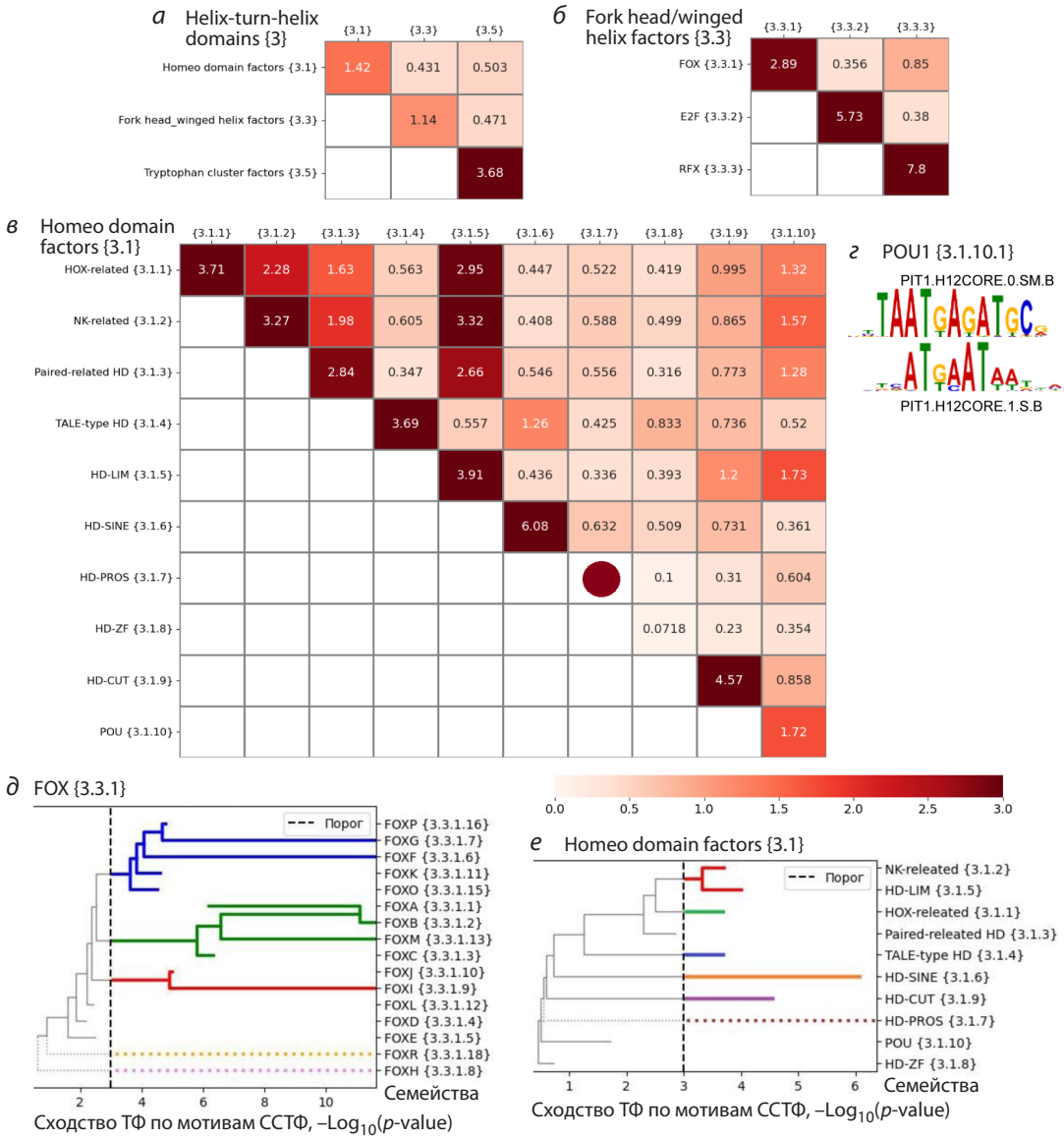

Рис. 7. Сходство ТФ по мотивам сайтов связывания для суперкласса Helix-turn-helix domains {3}.

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

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

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

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

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

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

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

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

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

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

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

Анализ сходства транскрипционных факторов дрозофилы

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

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

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

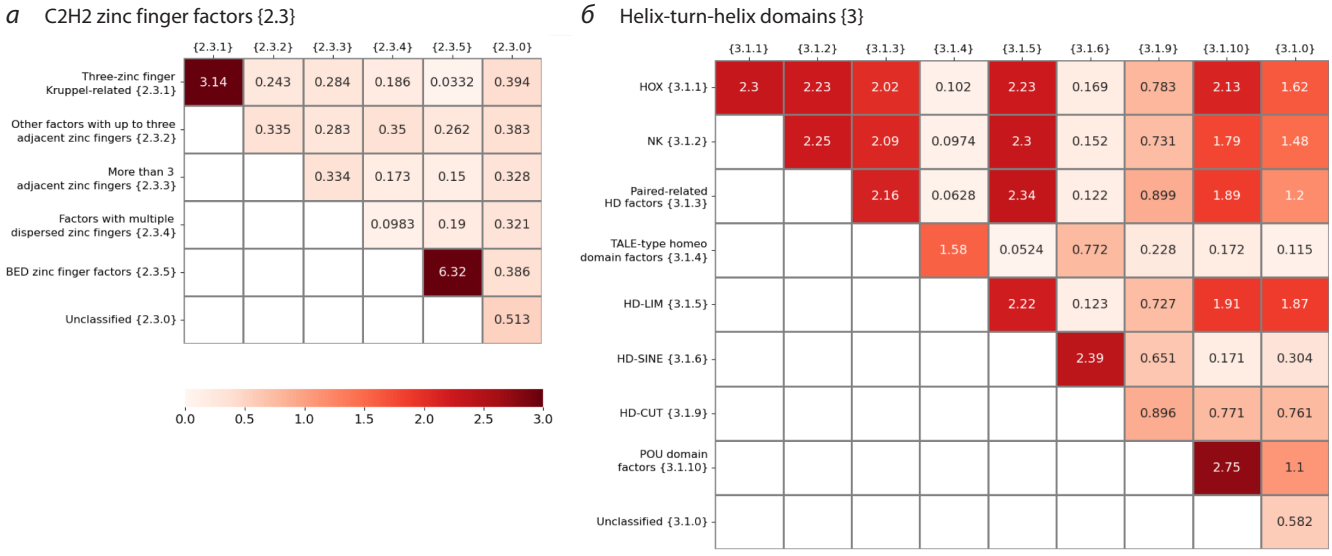

Рис. 8. Сходство ТФ дрозофилы из двух крупных классов по мотивам сайтов связывания.

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

Обсуждение

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

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

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

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

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

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

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

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

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

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

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

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

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

Заключение

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

Список литературы / References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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