Вавиловский журнал генетики и селекции
Vavilov Journal of Genetics and Breeding. 2024;28(8):822-833

doi 10.18699/vjgb-24-90

Программный комплекс MetArea для анализа взаимоисключающей встречаемости в парах мотивов сайтов связывания транскрипционных факторов по данным ChIP-seq

В.Г. Левицкий [1, 2], А.В. Цуканов [1], Т.И. Меркулова [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

levitsky@bionet.nsc.ru

Аннотация. Технология ChIP-seq, основанная на иммунопреципитации хроматина (ChIP), позволяет картировать набор геномных локусов (пиков), содержащих сайты связывания (СС) для исследуемого (целевого) транскрипционного фактора (ТФ). ТФ может распознавать несколько структурно различных мотивов СС. Мультибелковый комплекс, картируемый в эксперименте ChIP-seq, включает целевой и другие «партнерские» ТФ, связанные белок-белковыми взаимодействиями. Не все из этих ТФ связываются с ДНК напрямую. Поэтому и целевой, и партнерские ТФ распознают обогащенные мотивы СС в пиках. Для поиска обогащенных мотивов по данным ChIP-seq применяется подход *de novo* поиска. Для пары обогащенных мотивов СС ТФ в наборе пиков может быть обнаружена совместная или взаимоисключающая встречаемость: совместная отражает более частое нахождение двух мотивов СС ТФ в одних пиках, а взаимоисключающая – в разных пиках. Мы предлагаем программный комплекс (ПК) MetArea для выявления пар мотивов СС ТФ со взаимоисключающей встречаемостью по данным ChIP-seq. ПК MetArea предназначен для предсказания структурного разнообразия мотивов СС одного ТФ и функциональной связи мотивов СС разных ТФ. Функциональная связь мотивов двух разных ТФ предполагает, что они взаимозаменяемы в составе мультибелкового комплекса, который использует СС этих ТФ для прямого связывания с ДНК в различных пиках. ПК MetArea рассчитывает оценки точности распознавания pAUPRC (частичная площадь под кривой Precision–Recall) для каждого из двух входных одиночных мотивов, определяет их «объединенный» мотив и оценивает точность для него. Целью анализа является поиск пар одиночных мотивов A и B, для которых точность объединенного мотива A&B выше точностей обоих одиночных мотивов.

Ключевые слова: *de novo* поиск мотивов; кривая PR; площадь под кривой; структурные варианты мотивов сайтов связывания транскрипционных факторов; кооперативное действие транскрипционных факторов.

Для цитирования: Левицкий В.Г., Цуканов А.В., Меркулова Т.И. Программный комплекс MetArea для анализа взаимоисключающей встречаемости пар мотивов сайтов связывания транскрипционных факторов по данным ChIP-seq. *Вавиловский журнал генетики и селекции*. 2024;28(8):822-833. doi 10.18699/vjgb-24-90

Финансирование. Работа поддержана государственным бюджетным проектом № FWNR-2022-0020 Института цитологии и генетики СО РАН.

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

MetArea: a software package for analysis of the mutually exclusive occurrence in pairs of motifs of transcription factor binding sites based on ChIP-seq data

V.G. Levitsky [1, 2], A.V. Tsukanov [1], T.I. Merkulova [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

levitsky@bionet.nsc.ru

Abstract. ChIP-seq technology, which is based on chromatin immunoprecipitation (ChIP), allows mapping a set of genomic loci (peaks) containing binding sites (BS) for the investigated (target) transcription factor (TF). A TF may recognize several structurally different BS motifs. The multiprotein complex mapped in a ChIP-seq experiment includes target and other "partner" TFs linked by protein-protein interactions. Not all these TFs bind to DNA directly. Therefore, both target and partner TFs recognize enriched BS motifs in peaks. A *de novo* search approach is used to search for enriched TF BS motifs in ChIP-seq data. For a pair of enriched BS motifs of TFs, the co-occurrence or mutually exclusive occurrence can be detected from a set of peaks: the co-occurrence reflects a more frequent occurrence of two motifs in the same peaks, while the mutually exclusive means their more frequent detection in different peaks. We propose

the MetArea software package to identify pairs of TF BS motifs with the mutually exclusive occurrence in ChIP-seq data. MetArea was designed to predict the structural diversity of BS motifs of the same TFs, and the functional relation of BS motifs of different TFs. The functional relation of the motifs of the two distinct TFs presumes that they are interchangeable as part of a multiprotein complex that uses the BS of these TFs to bind directly to DNA in different peaks. MetArea calculates the estimates of recognition performance pAUPRC (partial area under the Precision–Recall curve) for each of the two input single motifs, identifies the "joint" motif, and computes the performance for it too. The goal of the analysis is to find pairs of single motifs A and B for which the accuracy of the joint A&B motif is higher than those of both single motifs.

Key words: *de novo* motif search; PR curve; area under curve; structural variants of transcription factor binding site motifs; cooperative action of transcription factors.

For citation: Levitsky V.G., Tsukanov A.V., Merkulova T.I. MetArea: a software package for analysis of the mutually exclusive occurrence in pairs of motifs of transcription factor binding sites based on ChIP-seq data. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2024;28(8):822-833. doi 10.18699/vjgb-24-90

Введение

Транскрипционные факторы (ТФ) – это белки, обладающие способностью специфического связывания ДНК и регулирующие таким образом транскрипцию генов. К ТФ относится около 1600 белков человека (Lambert et al., 2018). Сайты связывания (СС) ТФ в геномной ДНК эукариот представляют собой короткие участки длиной обычно от 6 до 20 пар нуклеотидов (п. н.) (Vorontsov et al., 2024). Как правило, ТФ способны связываться не с одной последовательностью ДНК, а со многими сходными. Мотив СС ТФ в ДНК – это общее представление доступного разнообразия таких сходных последовательностей (D'haeseleer, 2006). Установить четко закономерности, определяющие аффинность нуклеотидных последовательностей геномной ДНК к ТФ, очень трудно: даже умеренно консервативными в мотивах СС ТФ, т. е. неизменными в большинстве природных СС, является всего лишь несколько позиций нуклеотидов, число их обычно гораздо меньше половины длины мотива. Разнообразие мотивов СС ТФ *in vivo* пока слабо изучено из-за огромного разнообразия механизмов связывания ТФ с ДНК, включающего, наряду с прямым взаимодействием ТФ с ДНК, механизмы связывания при помощи или через посредника в составе мультибелкового комплекса с другими ТФ, использование пространственной структуры ДНК в составе нуклеосомы и т. п. (Morgunova, Taipale, 2017; Levitsky et al., 2020; Zeitlinger, 2020).

Самой популярной моделью мотивов СС ТФ является традиционная позиционная весовая матрица (ПВМ) (Wasserman, Sandelin, 2004; Tognon et al., 2023). Она оценивает аффинность сайта как сумму вкладов (весов) всех его позиций, где вес каждой позиции задан типом ее нуклеотида. Альтернативные модели мотива способны дополнить предсказания модели ПВМ (Levitsky et al., 2007; Siebert, Söding, 2016; Tsukanov et al., 2022), т. е. предсказывать СС ТФ в таких локусах генома, где модель ПВМ этого не делает. Общим отличием всех альтернативных моделей мотива от традиционной модели ПВМ является участие в оценке аффинности сайтов вкладов зависимостей частот нуклеотидов в разных позициях мотива.

Способность каждого ТФ взаимодействовать с ДНК обеспечивается его ДНК-связывающим доменом (ДСД). Структура ДСД ТФ определяет варианты мотивов его СС (Wingender, 2013; Lambert et al., 2018; Nagy G., Nagy L., 2020). Иерархическая классификация ТФ по структуре ДСД в базе данных TFClass (Wingender, 2013; Wingender et al., 2013, 2015, 2018) определяет классы ТФ по структуре ДСД. Например, в базе данных Hocomoco (Vorontsov et al., 2024) аннотированы мотивы СС 949 разных ТФ человека. Эти ТФ относятся к 34 классам, однако на 10 классов, имеющих не менее 10 ТФ, приходится 858 ТФ (более 90 % от всех 949 ТФ), а три крупнейших класса – C2H2 zinc finger factors {2.3}, Homeo domain factors {3.1} и Basic helix-loop-helix factors (bHLH) {1.2} – насчитывают 373, 184 и 76 ТФ соответственно. Выравнивание последовательностей ДСД ТФ определяет семейства и подсемейства ТФ ниже классов по иерархии.

Транскрипционные факторы эукариот взаимодействуют с ДНК *in vivo* в составе мультибелковых комплексов, включающих несколько ТФ. ТФ в составе таких комплексов можно назвать «партнерскими», так как между ними есть белок-белковые взаимодействия. Общее (кооперативное) действие нескольких ТФ на регуляторный район гена способно менять локальное окружение хроматина и регулировать транскрипцию гена (Morgunova, Taipale, 2017; Zeitlinger, 2020; Georgakopoulos-Soares et al., 2023). Для ТФ многих классов характерна возможность связываться с совершенно структурно разными СС (Rogers et al., 2019; Vorontsov et al., 2024). Например, ТФ класса Nuclear receptors with C4 zinc fingers {2.1} (Ядерные рецепторы с цинковыми пальцами С4) могут связываться и как мономеры, и как димеры. В случае димера СС включает два полусайта, спейсер между которыми и ориентация цепей ДНК могут варьировать. ТФ класса Basic leucine zipper factors (bZIP) {1.1} (Лейциновая застежка) связываются только как димеры, так что два полусайта всегда расположены в одной цепи ДНК, а спейсер почти не изменяется (Nagy G., Nagy L., 2020). Здесь и далее в фигурных скобках индексы обозначены согласно данным базы TFClass (Wingender et al., 2013, 2015, 2018). Есть несколько типов ДСД ТФ эукариот, которые могут функционировать как димеры, включающие пары близкородственных ТФ (Amoutzias et al., 2008). ТФ, похожие по структуре ДСД, часто распознают схожие мотивы СС ТФ (Lambert et al., 2018; Ambrosini et al., 2020). Единственное явное исключение из этого правила – мотивы СС ТФ класса C2H2 цинковых пальцев (C2H2 zinc finger factors {2.3}).

Идентификация СС ТФ в геномах значительно продвинулась с появлением методов высокопроизводительного массового секвенирования, в частности экспериментальной технологии ChIP-seq. Эта технология дает для целевого ТФ набор районов генома (пиков) длиной обычно

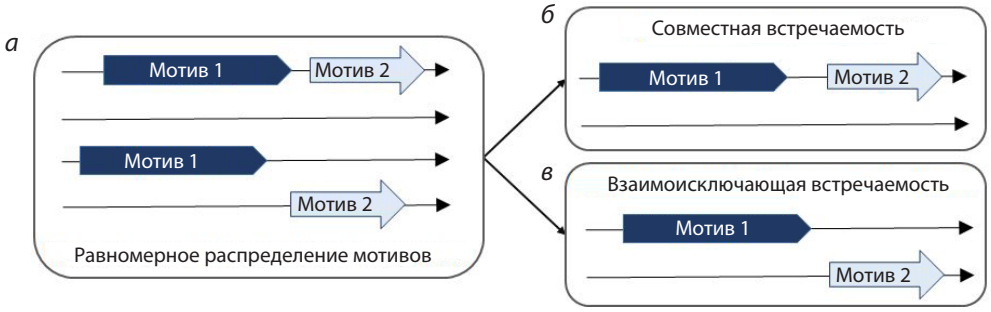

Рис. 1. Схема отличий терминов совместной и взаимоисключающей встречаемости мотивов СС ТФ.

Пусть частота встреч каждого из двух мотивов в пике 50 %. *а* – два мотива появляются в пиках независимо друг от друга, есть четыре равновероятных варианта картирования мотивов в пиках; *б* – совместная встречаемость: в пике есть только оба мотива или нет ни одного из двух; *в* – взаимоисключающая встречаемость: в пике есть только один мотив из двух. Стрелки от панели *а* к панелям *б* и *в* означают, что четыре варианта панели *а* точно разделяются на две группы по два варианта на панелях *б* и *в*.

несколько сотен пар нуклеотидов, где связывание мультибелкового комплекса многих ТФ, который включает и целевой ТФ, было экспериментально картировано. Поэтому два типа пиков отвечают за прямое и непрямое связывание целевого ТФ с геномной ДНК. Прямое связывание означает, что целевой ТФ взаимодействует с ДНК напрямую, а непрямое – что у целевого ТФ есть только белок-белковые взаимодействия с одним или несколькими партнерскими ТФ, которые, в свою очередь, взаимодействуют с ДНК напрямую. Из наличия прямого/непрямого связывания следует, что в пиках обогащены мотивы СС целевого/партнерских ТФ и что мотивы целевого ТФ есть только в части пиков. Для отражения роста содержания мотивов СС ТФ в геномных локусах, полученных по данным массового секвенирования ChIP-seq, применяется термин «обогащение» – повышенное содержание мотива СС ТФ по сравнению с ожидаемым по случайным причинам. Выборку последовательностей ДНК, по которой оценивается ожидаемое содержание мотива, называют негативной. Нами показано, что для пиков ChIP-seq более эффективно в негативную выборку брать любые участки генома, подходящие пикам по G/C-составу, чем использовать синтетические последовательности, полученные из пиков путем перемешивания нуклеотидов (Raditsa et al., 2024).

Когда по заданному набору данных массового секвенирования ChIP-seq определены обогащенные мотивы СС, на механизмы действия ТФ могут указать статистические закономерности встреч мотивов в парах. Ранее были определены понятия синергии и антагонизма мотивов в составе композиционного элемента (КЭ) – устойчивого сочетания для пары мотивов (Kel et al., 1995). Синергия означает, что результат действия пары ТФ значительно превосходит результат действия каждого из них в отдельности. При антагонизме ТФ, наоборот, мешают друг другу. Например, один из двух ТФ – активатор, а другой – репрессор, так что один из них вытесняет другого. К сожалению, понятия «синергия» и «антагонизм» относятся к устойчивой паре двух встречаемых в ДНК мотивов, и эти два случая невозможно различить по частотам совместных встреч пары мотивов.

С начала эпохи массового секвенирования СС ТФ прошло более 15 лет (Jonhson et al., 2007). Сегодня роль биоинформатического анализа полногеномных данных для понимания механизмов работы ТФ трудно переоценить. В случае данных ChIP-seq бионформатический анализ имеет дело не с отдельными локусами генома, а с набором сотен или даже тысяч таких локусов, в которых в целом наблюдается как прямое, так и непрямое связывание целевого ТФ. При переходе от отдельного рассмотрения частот двух мотивов СС ТФ в наборе пиков ChIP-seq к закономерностям связи между этими двумя мотивами целесообразно учитывать две возможности:

- два мотива чаще, чем ожидается по случайным причинам, встречаются вместе в одних пиках и реже по отдельности в разных пиках;
- два мотива встречаются чаще в разных пиках и реже в одних пиках.

Поэтому мы предлагаем для пары мотивов СС ТФ термины совместной и взаимоисключающей встречаемости (рис. 1).

Совместная встречаемость в паре мотивов отражает наличие КЭ, т. е. пары близко расположенных мотивов СС ТФ в ДНК, между которыми есть небольшой спейсер или которые перекрываются (Kel et al., 1995; Levitsky et al., 2019). Взаимоисключающая встречаемость в паре может означать, что либо эта пара представляет собой два структурных варианта СС одного ТФ (он связывается в разных пиках по-разному), либо это СС разных ТФ. Полагая, что два мотива СС соответствуют двум разным ТФ в составе одного мультибелкового комплекса, можно предположить замену одного ТФ, напрямую взаимодействующего с ДНК, на другой ТФ. Поэтому тренд расхождения мотивов СС двух ТФ по разным пикам может указывать на функциональную связь этих мотивов, в простейшем случае представляющую указанную выше замену. При совместной встречаемости, в случае как синергии, так и антагонизма, два ТФ связываются с ДНК вблизи друг от друга (по крайней мере некоторое время они могут быть в контакте даже при антагонизме); скорее всего, они входят в состав одного мультибелкового комплекса. При взаимоисключающей встречаемости, наоборот, мотивы

СС и соответствующие ТФ находятся в разных участках ДНК (разных пиках). Следовательно, мы предполагаем, что два мотива представляют альтернативные следы одной общей молекулярной функции ТФ:

- один ТФ распознаёт два различных по структуре мотива СС, или
- в составе мультибелкового комплекса связывание с ДНК происходит за счет разных ТФ и мотивов их СС.

Обе эти возможности иллюстрирует рис. 2.

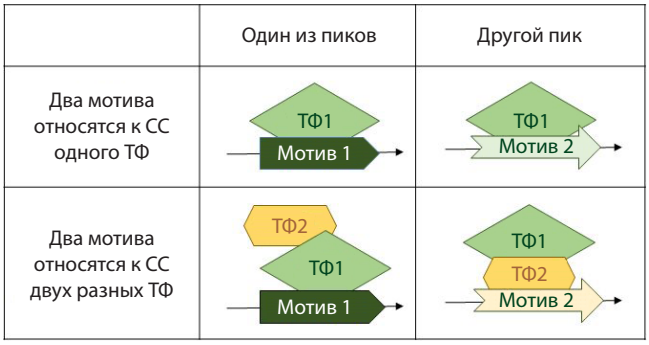

Рис. 2. Предполагаемое происхождение взаимоисключающей встречаемости двух мотивов СС ТФ в наборе пиков ChIP-seq.

Две колонки представляют два разных пика. Взаимоисключающая встречаемость в паре мотивов может означать, что либо пара мотивов объясняется двумя структурно разными мотивами одного ТФ (этот ТФ связывается с двумя мотивами в разных пиках), либо пара мотивов соответствует СС разных ТФ. В этом случае мы предполагаем, что в составе некоторого мультибелкового комплекса происходит замена взаимодействующего напрямую с ДНК одного ТФ на другой (ТФ1 на ТФ2).

Площадь под кривой ROC (AUC ROC, Area Under Curve) – это традиционная количественная оценка точности бинарного классификатора. Термин ROC (Receiver Operating Characteristic curve) означает кривая «Рабочая характеристика приемника». Для мотива СС ТФ кривая ROC определяется как зависимость доли распознанных последовательностей позитивной выборки (TPR, True Positive Rate) от доли распознанных последовательностей негативной выборки (FPR, False Positive Rate). Однако для моделей распознавания мотивов СС ТФ в данных ChIP-seq эффективно измерять FPR не как долю последовательностей негативной выборки, а как ожидаемую частоту мотива в ней. Это позволяет гораздо точнее оценить предсказания модели мотива на жестких и даже средних по жесткости порогах распознавания (Tsukanov et al., 2022). Для модели распознавания мотивов СС ТФ точность распознавания может рассчитываться как частичная площадь под кривой ROC (pAUC ROC) (Tsukanov et al., 2022). Значение pAUC ROC равно части площади под кривой, ограниченной максимально допустимой ожидаемой частотой мотива. Площадь под кривой ROC интегрирует доли пиков, обладающих предсказанными СС ТФ (доля верно предсказанных пиков, ось Y) в широком диапазоне порогов распознавания, исчисляемых как частота мотива в негативной выборке (ось X).

В этой работе мы предлагаем подход MetArea, который рассматривает два отдельных «одиночных» мотива, а также некий «объединенный» мотив, означающий появление любого их двух одиночных. Для распознавания объединенного мотива в последовательности ДНК достаточно распознать в ней хотя бы один из двух одиночных мотивов при заданном пороге ожидаемой частоты мотива. Точный расчет частоты такого объединенного мотива даже по одной последовательности ДНК представляет собой некоторые трудности из-за огромного разнообразия возможных перекрываний одиночных мотивов. Поэтому для оценки точности модели мотива мы разработали и применили меру точности «Частичная площадь под кривой PR (Precision–Recall)», для расчета которой нужно отслеживать только число распознанных последовательностей позитивной и негативных выборок.

Кривая PR – это зависимость меры Precision (отношения числа предсказанных последовательностей в позитивной выборке к числу предсказанных последовательностей в позитивной и негативной выборках) от меры Recall (отношения числа предсказанных последовательностей позитивной выборки к объему выборки). Кривая PR – это альтернатива более популярной кривой ROC (Davis, Goadrich, 2006; Keilwagen et al., 2019). Достоинством меры «Площадь под кривой PR» по сравнению с мерой «Площадь под кривой ROC» является соотношение вкладов мягких и жестких порогов распознавания, соответствующих предсказанным сайтам с низкой и высокой аффинностью. По сравнению с кривой ROC, кривая PR обеспечивает бóльшие вклады от сайтов с высокой аффинностью, чем от сайтов с низкой. Кривая ROC поступает наоборот. Согласно кривой PR, вклад от сайтов с некоторой низкой аффинностью может даже стремиться к нулю, если такие сайты не содержат специфического нуклеотидного контекста. Это связано с равными вероятностями распознавания объектов в позитивной и негативной выборках (Saito, Rehmsmeier, 2015).

Мы разработали программный комплекс (ПК) MetArea для выявления пар мотивов СС ТФ со взаимоисключающей встречаемостью. ПК MetArea производит расчеты оценок точности pAUPRC (частичная площадь под кривой PR) для каждого из двух входных одиночных мотивов, а также для их сочетания – «объединенного мотива». Это позволяет выявить взаимоисключающую встречаемость двух входных мотивов.

Материалы и методы

В анализе использовались данные ChIP-seq из базы данных GTRD (Kolmykov et al., 2021). По каждому эксперименту ChIP-seq в анализ взят набор из 1000 пиков наилучшего качества, согласно предобработке инструментом MACS2 (Zhang et al., 2008). В настоящей работе в анализ взяты обогащенные мотивы, полученные из результатов *de novo* поиска мотивов, и мотивы СС ТФ мыши *Mus musculus* из базы данных Hocomoco (https://hocomoco12.autosome.org/) (Vorontsov et al., 2024). *De novo* поиск мотивов традиционной ПВМ и альтернативной SiteGA моделей мотивов СС ТФ проведен с помощью STREME https://meme-suite.org/meme/tools/streme (Bailey, 2021) и https://github.com/parthian-sterlet/sitega (Tsukanov et al., 2022). Значимость сходства обогащенных мотивов из результатов *de novo* поиска мотивов STREME с мотивами известных ТФ из баз данных Hocomoco, Cis-BP (Weirauch et al., 2014) и JASPAR (Rauluseviciute et al., 2024) оценивалась

с помощью инструмента TomTom (https://meme-suite.org/meme/tools/tomtom) (Gupta et al., 2007). Для анализа ПК MetArea допускает также мотивы из баз данных Hocomoco и JASPAR, отбираемые согласно принятой ранее методике (ПК MCOT) (Levitsky et al., 2019). Самый лучший хит модели мотива должен иметь ожидаемую частоту в выборке всех промоторов генов генома, кодирующих белки, не менее 2E–5. Лучший хит дает предсказанный сайт с максимально возможным значением функции распознавания модели мотива.

Итого в состав ПК MetArea включены 1420/1142 мотива для 942/713 ТФ человека/мыши из базы данных Hocomoco и 556/151 мотив для 555/148 ТФ растений/насекомых из базы данных JASPAR. ПК MetArea доступен по адресу https://github.com/parthian-sterlet/metarea. Детальное описание алгоритма ПК MetArea см. далее в разделе результатов. В ПК MetArea для оценки сходства анализируемых мотивов модели ПВМ (матриц частот нуклеотидов) реализован подход из ПК MCOT (Levitsky et al., 2019).

Результаты

Общее описание ПК MetArea

ПК MetArea позволяет анализировать как пары мотивов традиционной модели ПВМ, так и пары мотивов традиционной ПВМ и альтернативной SiteGA моделей (Levitsky et al., 2007; Tsukanov et al., 2022). Общая схема работы конвейера ПК MetArea представлена на рис. 3.

Входными данными и параметрами ПК MetArea являются:

- Два мотива: сочетание двух мотивов модели ПВМ, заданными двумя матрицами частот нуклеотидов (МЧН), или сочетание мотива модели ПВМ, заданной МЧН, и модели мотива SiteGA, заданной своей весовой матрицей, см. https://github.com/parthian-sterlet/sitega (Tsukanov et al., 2022).
- Позитивная выборка в формате FASTA (набор пиков ChIP-seq, NF последовательностей, Number of Foreground sequences).
- Негативная выборка в формате FASTA (NB последовательностей, Number of Background sequences). Ее рекомендуется приготовить предварительно по позитивной выборке и полному геному с помощью ПК AntiNoise (Raditsa et al., 2024), https://github.com/parthian-sterlet/antinoise. Для каждой последовательности позитивной выборки по ее длине и G/C-составу в полном геноме случайным образом находится несколько последовательностей негативной выборки. Далее в анализе NF/NB = 5.
- Выборка промоторов всех генов генома, необходимая для определения порогов распознавания на основе расчета таблиц ‘Threshold vs. ERR’ («Порог функции распознавания vs. Частота мотива в выборке всех промоторов генома») для каждого из входных мотивов.
- Порог ERR_{MAX} максимальной ожидаемой частоты для каждого входного мотива (Expected Recognition Rate, ERR).
- Таблицы ‘Threshold vs. ERR’ для каждого входного мотива.

Максимальная частота мотива 0.01 означает, что специфичность СС соответствует одному сайту на сто нуклеотидных позиций. Рекомендуемый интервал порога ожидаемой частоты мотива ERR_{MAX} – от 0.001 до 0.01. Далее в расчетах принято значение параметра $ERR_{MAX} = 0.002$. Ранее мы использовали таблицы ‘Threshold vs. ERR’ для унификации порогов распознавания разных мотивов (Levitsky et al., 2019; Tsukanov et al., 2021, 2022). Каждый мотив вместе со своей таблицей ‘Threshold vs. ERR’ подается в файле бинарного формата, генерируемого компонентами ПК MetArea для расчета ожидаемых частот мотива для моделей мотива ПВМ и SiteGA.

Выходными данными ПК MetArea являются:

- Текстовый файл с кривыми PR для каждого из входных мотивов, а также их объединенного мотива.
- Текстовый файл со значениями оценок точности распознавания pAUPRC для каждого из входных мотивов, а также их объединенного мотива, значением отношения площадей под кривыми (см. ниже), оценкой сходства мотивов (только для пар мотивов модели ПВМ).

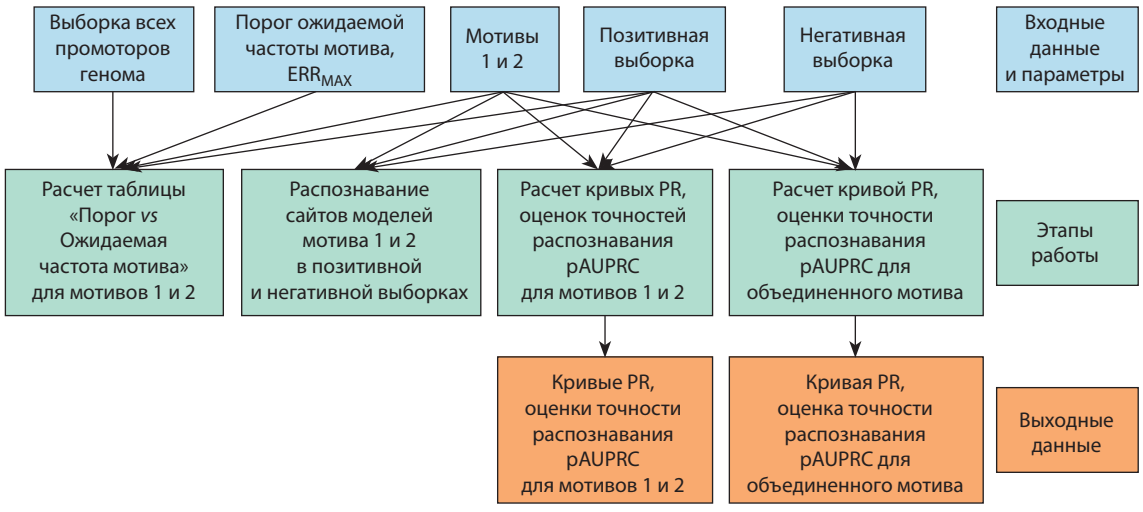

Рис. 3. Общая схема работы ПК MetArea.

Определение порогов распознавания для разных мотивов

Пороги функций распознавания каждого из двух входных мотивов согласно предварительно рассчитанным таблицам 'Threshold vs. ERR' переводятся в общую шкалу ожидаемой частоты мотива, ERR (Levitsky et al., 2019; Tsukanov et al., 2021, 2022). Это необходимо для построения PR кривой объединенного мотива. Ожидаемая частота мотива ERR для входных мотивов рассчитывается до порога ERR_{MAX}, так что все ожидаемые частоты удовлетворяют условию $ERR < ERR_{MAX}$.

Ожидаемую частоту мотива в выборке промоторов вычисляли следующим образом. Определяли значения функции распознавания мотива для каждого предсказанного сайта в выборке в каждой позиции и цепи ДНК. Затем для каждого порога функции распознавания ожидаемую частоту мотива вычисляли как отношение числа предсказанных СС, для которых значения этой функции равны порогу распознавания или выше него, к общему числу позиций, доступных для таких СС в выборке с учетом цепей ДНК.

Статистические метрики и кривая PR

Кривую PR, согласно (Davis, Goadrich, 2006), для модели мотива СС ТФ можно определить так: по оси X – отношение числа последовательностей позитивной выборки (пиков) с предсказанными сайтами к числу всех пиков (TPR, True Positive Rate, Recall, REC):

$$REC = \frac{TP}{TP+FN}. \quad (1)$$

Здесь TP/FN (True Positives/False Negatives) – число верно/неверно предсказанных последовательностей позитивной выборки (TP + FN = NF).

По оси Y кривой PR – отношение числа предсказанных последовательностей позитивной выборки к числу всех предсказанных последовательностей позитивной и негативной выборок (Precision, PREC), согласно (Davis, Goadrich, 2006):

$$PREC = \frac{TP}{TP+FP}. \quad (2)$$

Здесь FP (False Positives) – число предсказанных последовательностей негативной выборки. С учетом отличия объемов позитивной (NF) и негативной (NB) выборок мы поправили расчет величины Precision следующим образом:

$$PREC = \frac{TPR}{TPR+FPR} = \frac{TP/NF}{TP/NF+FP/NB} = \frac{TP}{TP+(NF/NB)\times FP}. \quad (3)$$

Здесь TPR и FPR – доли предсказанных последовательностей в позитивной и негативной выборках. Коэффициент NF/NB учитывает различие объемов негативной (NB) и позитивной (NF) выборок. Ожидаемое по случайным причинам число предсказанных последовательностей позитивной (TP) и негативной (FP) выборок пропорционально объемам выборок, NF и NB, соответственно. Целью замены формулы (2) на формулу (3) введением коэффициента NF/NB является унификация вида кривой PR для разных отношений объемов позитивной и негативной выборок.

Частичная площадь под кривой PR и отношение площадей под кривыми

Алгоритм MetArea использует таблицы «Порог функции распознавания vs. Частота мотива в выборке всех промоторов генома», описанные выше. Затем выполняется распознавание двух входных одиночных мотивов в позитивной и негативной выборках. Далее рассчитывается мера pAUPRC для одиночных мотивов, а также для объединенного мотива. Расчет частичной площади под кривой PR (pAUPRC) ограничен условиями, налагаемыми на меры Recall (ось X) и Precision (ось Y), т.е. площадь является частичной как по оси X, так и по оси Y (рис. 4).

Условием частичности площади под кривой PR по оси X является участие в расчетах меры pAUPRC части всего диапазона меры Recall от 0 до 1. Это условие означает, что в определении меры площади участвуют не все пики с предсказанными сайтами, а только те пики, наилучшие хиты в которых имеют ожидаемую частоту меньше порога, $ERR < ERR_{MAX}$ (см. рис. 4). Порог частоты $ERR_{MAX} = 0.002$ выбран более мягким, чем ранее использованный для анализа мотивов целевых ТФ порог $ERR_{MAX} = 0.001$

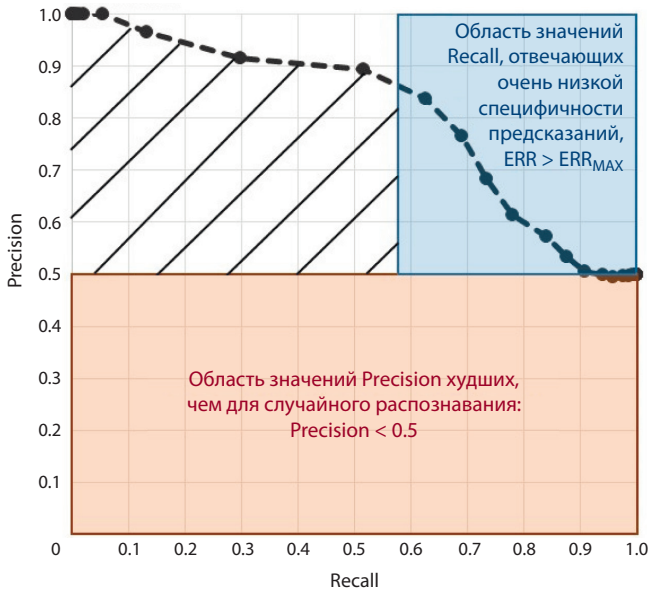

Рис. 4. Схема расчета частичной площади под кривой PR.

Ось X – мера Recall, вероятность предсказания последовательности позитивной выборки, Recall = TPR = TP/NF, формула (1). Ось Y – мера Precision, отношение вероятности предсказания последовательности позитивной выборки к сумме вероятностей предсказания последовательностей позитивной и негативной выборок, Precision = TPR/(TPR+FPR), формула (3). Розовая область отмечает значения Precision < 0.5, соответствующие предсказаниям худшим, чем предсказания случайной модели, равновероятно распознающей последовательности позитивной и негативной выборок. Precision > 0.5/Precision < 0.5 – области отбора в сторону позитивной/негативной выборок. Голубым показана область предсказанных последовательностей позитивной выборки с очень низкой специфичностью, соответствующих ожидаемой частоте мотива, большей порога, $ERR > ERR_{MAX}$. Для генерации данных примера негативной выборки взято нормальное распределение со средним и стандартным отклонением $(\mu_N, \sigma_N) = (5, 2.5)$, а позитивная выборка – это смесь 50 % на 50 % нормальных распределений $(\mu_{P1}, \sigma_{P1}) = (10, 1)$ и $(\mu_{P2}, \sigma_{P2}) = (5.5, 4)$. Эти распределения моделируют сайты, проходящие и не проходящие порог ERR_{MAX} ожидаемой частоты мотива. Штриховка обозначает область, по которой определяется метрика частичной площади под кривой pAUPRC.

(Tsukanov et al., 2022), поскольку ранее проводился анализ мотивов целевых ТФ ChIP-seq экспериментов, а ПК MetArea анализирует как мотивы СС целевых ТФ, так и менее консервативные мотивы СС партнерских ТФ.

Условием частичности площади под кривой PR по оси *Y* является вычитание из каждого значения меры Precision ее ожидаемого значения $PREC_{EXP}$ (см. рис. 4) (Saito, Rehmsmeier, 2015). Для модели, равновероятно распознающей данные позитивной и негативной выборки, PR кривая представляет собой горизонтальную линию:

$$PREC_{EXP} = \frac{NF}{NF+NB} = 0.5. \quad (4)$$

Значение $PREC_{EXP}$ постоянно и равно 0.5, поскольку ранее значение FP было нормализовано, так что объемы выборок в этой формуле уже можно считать равными. В итоге частичная площадь под кривой PR в ПК MetArea вычисляется как следующая сумма:

$$pAUPRC = \frac{2}{NF} \times \sum_{i=1}^{NI} \left[\left\{ \frac{PREC(i)+PREC(i-1)}{2} - PREC_{EXP} \right\} \times \right.$$
$$\left. \times \left\{ REC(i) - REC(i-1) \right\} \right]. \quad (5)$$

Здесь NI – самый мягкий порог, определяемый, как описано выше, по ожидаемым частотам и входному параметру ERR_{MAX}. Коэффициент 2/NF в формуле (5) необходим для нормировки значения pAUPRC на максимальное значение 1, так как при максимальных значениях меры Precision, равных 1, максимальное значение первого множителя под суммой, $\{(PREC(i)+PREC(i-1))/2 - PREC_{EXP}\}$, равно 0.5, а максимальное значение суммы всех вторых множителей, $\{REC(i)-REC(i-1)\}$, равно NF – объему позитивной выборки.

Критерий предсказания функциональной связи мотивов отражает повышение оценки точности объединенного мотива по сравнению с оценками точности одиночных мотивов и является количественной оценкой взаимоисключающей встречаемости в парах мотивов. Для пары мотивов A и B критерий требует более высокого значения оценки точности pAUPRC(A&B) объединенного мотива A&B по сравнению со значениями оценок точностей обоих одиночных мотивов, pAUPRC(A) и pAUPRC(B). Рассчитанное таким образом Отношение Площадей Под Кривыми (ОППК, Ratio of Areas Under Curves, RAUC) должно быть больше единицы:

$$RAUC(A, B) = \frac{pAUPRC(A\&B)}{Max\{pAUPRC(A), pAUPRC(B)\}} > 1. \quad (6)$$

Варианты применения ПК MetArea

Входными данными ПК MetArea могут быть мотивы СС ТФ, для которых предполагается обогащение в позитивной выборке по сравнению с негативной, например, такие мотивы – это результат *de novo* поиска мотивов (Bailey, 2021). Отдельные варианты применения ПК реализуют массовый анализ коллекций мотивов СС ТФ из баз данных Hocomoco и JASPAR. Анализ множества пар мотивов позволяет найти пары, для которых обнаруживается наибольший рост оценки точности распознавания pAUPRC при объединении мотивов. ПК MetArea допускает несколько вариантов применения, реализованных в виде отдельных программ. В следующих вариантах применения рассматривается модель мотива ПВМ:

- два заданных мотива;
- несколько заданных мотивов, для K мотивов проверяются все возможные $\{K \times (K-1)/2\}$ пары;
- заданный мотив против всех (M) мотивов СС известных ТФ из базы данных. Для заданного мотива проверяются все его M пар с мотивами из коллекции Hocomoco (человек, мышь) или JASPAR (растения, насекомые);
- все мотивы СС известных ТФ из базы данных. Из всех M мотивов известных ТФ из коллекции Hocomoco или JASPAR отбираются K мотивов с наивысшими оценками точности pAUPRC и проверяются все возможные пары этих мотивов, их всего $\{K \times (K-1)/2\}$.

Вариант применения для мотивов моделей ПВМ и SiteGA:

- мотив ПВМ и мотив SiteGA.

Далее приведены примеры результатов анализа данных ChIP-seq для разных вариантов применения ПК MetArea.

Анализ нескольких заданных мотивов модели ПВМ

Рассмотрим набор данных ChIP-seq для ТФ BHLHA15 (Hess et al., 2016) (GTRD PEAKS039234, GEO GSE86289) для поджелудочной железы мыши. *De novo* поиск мотивов с помощью инструмента STREME (Bailey, 2021) показал, что среди пяти мотивов с самым высоким обогащением четыре имеют значимое сходство (p-value < 0.001) (Gupta et al., 2007) с известными мотивами СС ТФ BHLHA15 из Hocomoco (BHA15.H12CORE.0.P.B, мотивы 1 и 5; BHA15.H12CORE.1.SM.B, мотивы 2 и 4). Мотивы 1/5 и 2/4 соответствуют консенсусу E-бокс CAnnTG со спейсерами GC и TA, поэтому они обозначены BHLHA15_GC_1/BHLHA15_GC_2 и BHLHA15_TA_1/BHLHA15_TA_2 соответственно. Мотив 3 имеет значимое сходство (p-value < <0.001) с мотивом СС ТФ CTCF (CTCF.H12CORE.0.P.B) (рис. 5, *а*).

Анализ значений оценок точности распознавания pAUPRC для одиночных мотивов и их попарных объединенных мотивов (см. рис. 5, *б*) проводится на основе соответствующих значений ОППК для пар мотивов (см. рис. 5, *в*), оценка сходства пар мотивов нужна для контроля значимо похожих мотивов (см. рис. 5, *г*). Высокие ОППК показаны для пар мотивов BHLHA15_GC_1/BHLHA15_TA2 и BHLHA15_GC_1/BHLHA15_TA1, кривые PR по ним показаны на рис. 5, *д*, *е*. Мотив CTCF имеет высокие ОППК с мотивами BHLHA15_GC1 и BHLHA15_TA2 (см. рис. 5, *в*). В паре мотивов BHLHA15_TA2 и CTCF обнаружено максимальное значение ОППК = 1.48 (см. рис. 5, *в*). Полученные результаты согласуются со способностью ТФ BHLHA15 связываться с ДНК только в составе димера ТФ класса bHLH (Amoutzias et al., 2008). Тренд к расхождению разных по структуре СС ТФ BHLHA15 в разные пики может означать, что в состав димера могут входить разные ТФ класса bHLH (включая и ТФ BHLHA15) и что на связывание димера влияют другие партнерские ТФ, формирующие с ТФ BHLHA15 мультибелковые комплексы, так что ДСД ТФ BHLHA15 принимает разную конформацию. Например, таким партнерским ТФ может быть CTCF, мотив СС

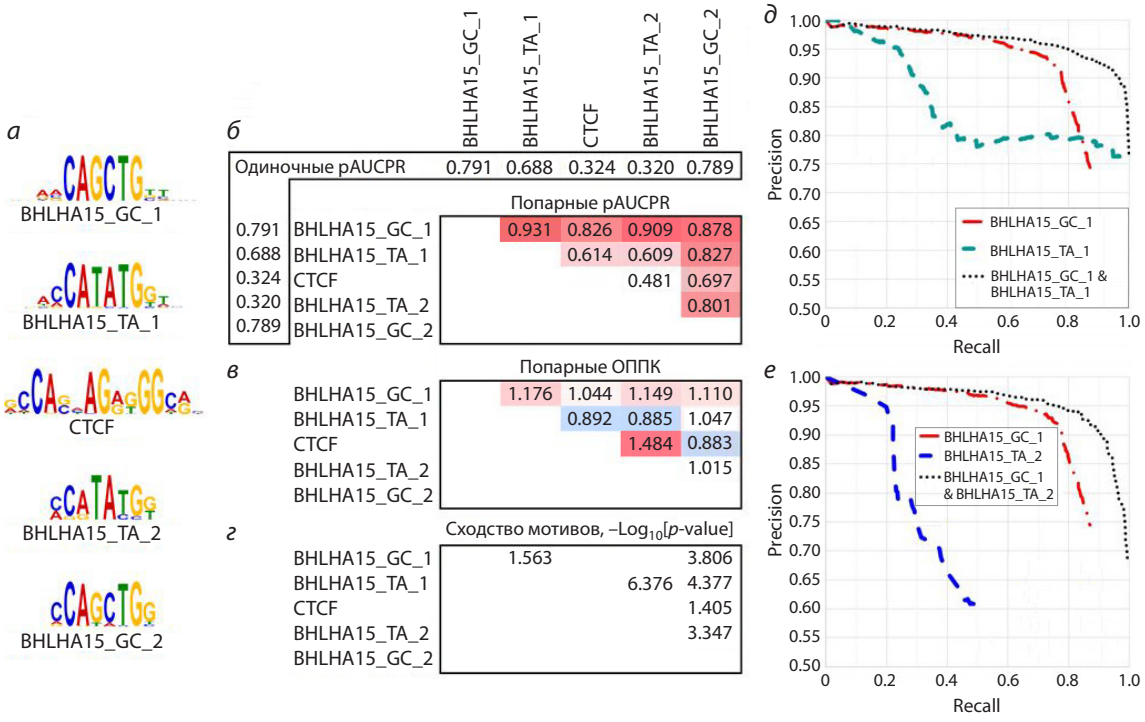

Рис. 5. Анализ пяти самых обогащенных мотивов из результатов *de novo* поиска мотивов (STREME) (Bailey, 2021) для набора данных ChIP-seq по ТФ BHLHA15 мыши (Hess et al., 2016, GTRD PEAKS039234, GEO GSM2299654/GSM2299655).

а – лого пяти мотивов, сортировка по значимости обогащения, полученной инструментом STREME; обозначения мотивов СС ТФ BHLHA15 согласно динуклеотиду в его спейсере в общем консенсусе CAnnTG; *б* – таблица попарных величин оценок точности pAUPRC объединенных мотивов, построенных по попарным сочетаниям мотивов; в заголовках указаны значения pAUPRC для одиночных мотивов, оттенки красного цвета отмечают максимальные значения pAUPRC объединенного мотива; *в* – таблица ОППК в парах мотивов, оттенки красного и синего цвета отмечают значения больше и меньше единицы; *г* – таблица значимостей сходства мотивов, $-\mathrm{Log}_{10}[p\text{-value}]$; *д*, *е* – кривые PR для одиночных мотивов и их попарных объединенных мотивов BHLHA15_GC_1/BHLHA15_TA1 и BHLHA15_GC_1/BHLHA15_TA2.

которого также обогащен (см. рис. 5, *а*). Согласно экспериментальным данным, (1) несколько ТФ из класса bHLH имеют белок-белковые взаимодействия с ТФ CTCF (база BIOGRID https://thebiogrid.org/); (2) анализ партнерских ТФ по геномной колокализации (Hu et al., 2020) подтверждает, что несколько ТФ класса bHLH колокализованы с ТФ CTCF в одних геномных локусах *in vivo*.

Анализ всех мотивов СС известных ТФ из базы данных

Рассмотрим набор данных ChIP-seq по ТФ AR (Androgene Receptor, рецептор андрогена) для простаты мыши (Chen et al., 2013) (GTRD PEAKS035588, GEO GSM1145307). На рис. 6 для этого набора данных ChIP-seq показана матрица попарных значений ОППК для 15 самых обогащенных мотивов СС ТФ согласно мере pAUPRC из всех 1142 мотивов СС ТФ мыши из базы данных Hocomoco. Семь из 15 мотивов принадлежат СС ТФ AR и его гомологам из подсемейства GR-like (NR3C) {2.1.1.1} семейства Steroid hormone receptors {2.1.1} класса Nuclear receptors with C4 zinc fingers {2.1}. Это семейство определяет целевой ТФ AR и вероятные мотивы его СС. Остальные восемь мотивов принадлежат СС ТФ из подсемейств FOXA {3.3.1.1}, FOXJ {3.3.1.10}, FOXM {3.3.1.13} и FOXP {3.3.1.16} одного семейства FOX {3.3.1} класса Fork head/winged helix factors {3.3}. ТФ этого семейства – предполагаемые партнерские ТФ для ТФ AR, например ТФ Foxa1, – известен для этой же ткани простаты (Yang, Yu, 2015).

Значения ОППК больше 1 получены почти для всех пар мотивов GR-like/FOX. Например, значение ОППК = = 1.03 в паре ANDR.H12CORE.0.P.B (ранг pAUPRC 1) и FOXA2.H12CORE.0.PSM.A (ранг 5) соответствует максимальному значению 0.853 pAUPRC по парам мотивов GR-like/FOX. Значения ОППК для пар мотивов GR-like/GR-like превышают 1 лишь для некоторых пар мотивов. Мотив ANDR.H12CORE.2.P.B (ранг 7), отличный по консенсусу от всех остальных мотивов GR-like (AAACA вместо GNACA, см. рис. 6, колонка Лого), имеет высокие значения ОППК. Также это единственный мотив, для которого ОППК больше 1 в парах со всеми остальными мотивами GR-like и FOX. В частности, среди пар мотивов GR-like/GR-like максимальное значение pAUPRC 0.876 при ОППК = 1.06 достигается в паре мотивов ANDR.H12CORE.0.P.B (ранг 1) и ANDR.H12CORE.2.P.B (ранг 7). Высокие значения ОППК в парах мотивов GR-like/GR-like обнаружены и для мотива MCR.H12CORE.1.SM.B, однако он имеет самый низкий ранг pAUPRC – 15. MCR.H12CORE.1.SM.B является мотивом связывания мономера, а не димера. Среди пар мотивов FOX/FOX значений ОППК больше 1 почти нет.

В целом высокие значения ОППК многих пар мотивов GR-like/GR-like позволяют предположить, что ТФ AR связывается в разных пиках с помощью различных структурных вариантов мотивов GR-like. Аналогичное предположение можно сделать и о связывании димера, вклю-

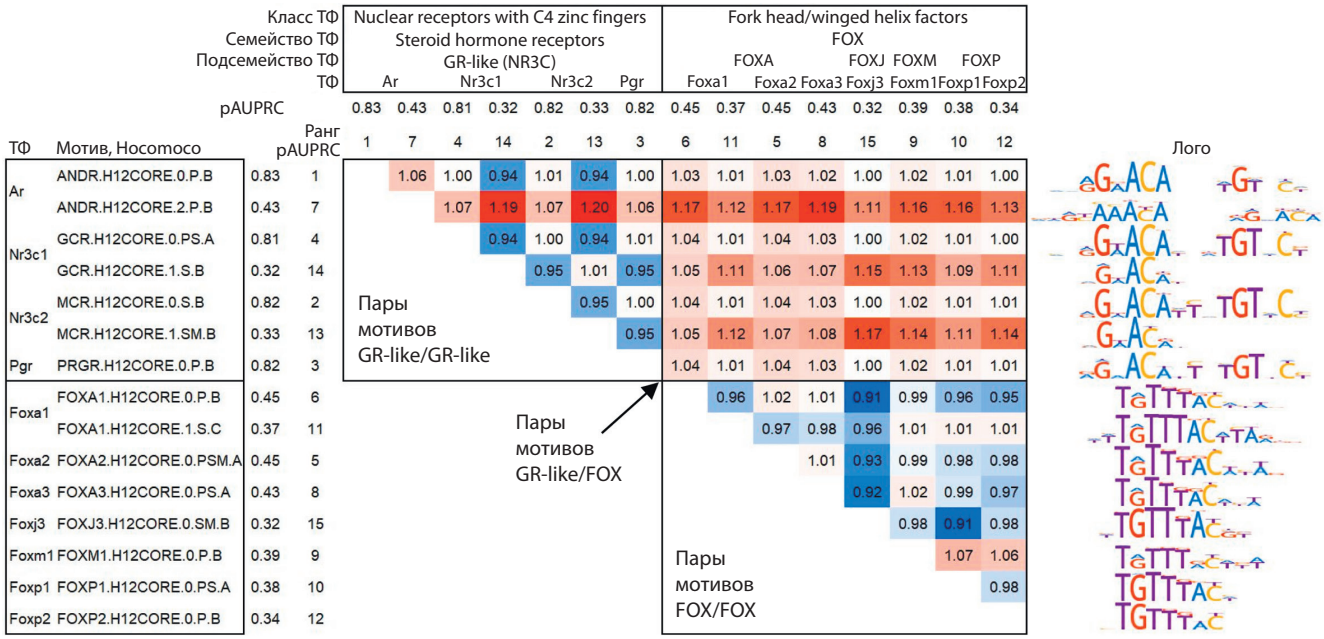

Рис. 6. Результаты анализа мотивов СС известных ТФ из базы данных Hocomoco для набора данных ChIP-seq по ТФ AR в простате мыши (Chen et al., 2013).

В анализ включены 15 самых обогащенных мотивов по оценкам точности pAUPRC. Их значения и ранги, а также имена ТФ из базы данных Hocomoco приведены в заголовках рядов и колонок. В заголовках рядов указаны идентификаторы мотивов из Hocomoco, а в заголовках колонок – имена классов, семейств и подсемейств ТФ. В таблице оттенками красного/синего цвета обозначены изменения ОППК в большую/меньшую сторону от нейтрального значения 1. В самой правой колонке указаны лого мотивов из базы данных Hocomoco. Черными рамками отмечены мотивы GR-like и FOX в заголовках рядов и колонок, а также пары мотивов СС ТФ GR-like/GR-like, GR-like/FOX и FOX/FOX в таблице.

чающего ТФ AR и ТФ семейства FOX на основе высоких значений ОППК для пар мотивов GR-like/FOX. Полученные результаты для ChIP-seq данных по ТФ AR означают, что связывание с ДНК ТФ AR происходит в составе димеров AR/AR и AR/Foxa1 (если в условиях эксперимента с мотивами FOX связывается именно ТФ Foxa1) и что оба ТФ допускают большое разнообразие разных структурных типов СС, так что разные пары мотивов расходятся по разным пикам.

Анализ пары мотивов моделей ПВМ и SiteGA

Рассмотрим набор данных ChIP-seq для ТФ E2F4 для дендритных клеток первичного врожденного иммунитета, полученных из костного мозга мыши, стимулированных патогенным компонентом липополисахаридом в течение 120 мин (Garber et al., 2012) (GTRD PEAKS035857, GEO GSM881061). Кривые PR для мотивов ПВМ, SiteGA и их объединенного мотива ПВМ & SiteGA, рассчитанные с помощью ПК MetArea, представлены на рис. 7. Значения pAUPRC для мотивов ПВМ, SiteGA и объединенного мотива ПВМ & SiteGA равны 0.457, 0.358 и 0.47 соответственно, значение ОППК объединенного мотива равно 1.028.

Модели мотивов ПВМ и SiteGA основаны на совершенно разных методологических принципах (Levitsky et al., 2007). Модель ПВМ представляет сайты с высокой аффинностью, определяемые самыми консервативными позициями и наиболее частыми нуклеотидами в них. Модель SiteGA подразумевает сайты, содержащие зависимости разных позиций, которые, по-видимому, возникают из-за кооперативных действий как минимум двух ТФ при связывании с ДНК (Morgunova, Taipale, 2017; Levitsky et al., 2020). Сайты модели SiteGA заметно менее консервативны, чем сайты модели ПВМ, модель SiteGA способна лучше, чем модель ПВМ, предсказывать сайты с низкой аффинностью (Tsukanov et al., 2022). Объединение моделей ПВМ и SiteGA позволяет улучшить распознавание сайтов с низкой аффинностью, что отражает бóльшая

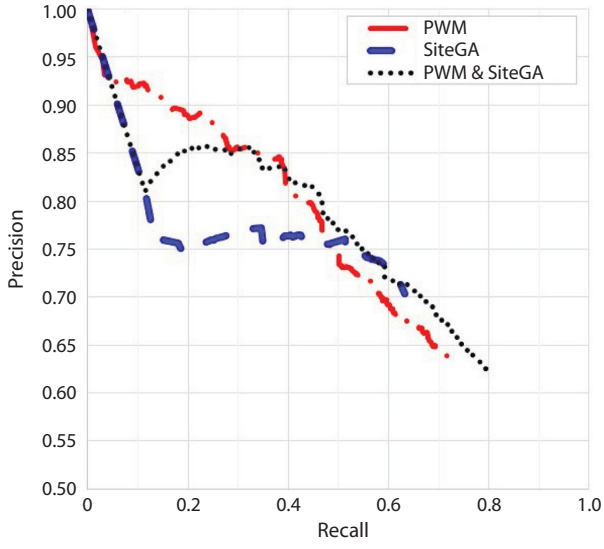

Рис. 7. Результаты анализа пары мотивов моделей ПВМ и SiteGA с помощью ПК MetArea.

Красным, синим и черным цветом показаны кривые PR для мотивов ПВМ, SiteGA и объединенного мотива ПВМ & SiteGA. В анализе использован набор данных ChIP-seq для ТФ E2F4 (GTRD PEAKS035857, GEO GSM881061).

протяженность кривой PR объединенного мотива ПВМ & SiteGA по оси *X* (Recall) по сравнению с каждым из одиночных мотивов ПВМ и SiteGA. Хотя объединенный мотив и имеет меньшие значения величины Precision (см. рис. 7, ось *Y*), чем модель ПВМ, более широкий интервал величин Recall (ось *X*) определяет рост меры pAUPRC объединенного мотива. Одиночные мотивы до порога ожидаемой частоты мотива $ERR_{MAX} = 0.002$ распознают 73.2 % (ПВМ) и 63.3 % (SiteGA) пиков, объединенный мотив распознает 79.9 %.

Гипотеза о том, что модели ПВМ и SiteGA представляют собой различные структурные варианты СС ТФ E2F4, подтверждается инструментом TomTom сравнения мотивов (p-value < 0.05) (Gupta et al., 2007): для модели ПВМ – по ее матрице частот нуклеотидов, а для модели SiteGA, как и ранее (Tsukanov et al., 2022), – по матрице частот нуклеотидов, построенной по предсказанным сайтам. На способность ТФ E2F4 связываться с разными структурными типами СС указывает также то, что в эксперименте M. Garber с коллегами (2012) в одних условиях определены геномные локусы связывания 25 ТФ и показано, что локусы ТФ E2F4 значимо перекрываются с локусами пяти ТФ: EGR2, EGR1, IRF2, ETS2 и E2F1. Следовательно, можно предполагать, что ТФ E2F4 входит в одни мультибелковые комплексы с этими ТФ. Поэтому в разных локусах ТФ E2F4 вынужден в большей или меньшей степени изменять свои СС, чтобы адаптироваться к СС партнерских ТФ.

Обсуждение

В нашей работе предложен новый подход MetArea для выявления взаимоисключающей встречаемости в парах мотивов СС ТФ на основе анализа наборов данных ChIP-seq. Если два мотива являются структурно различными мотивами СС одного ТФ в разных пиках одного набора, то взаимоисключающая встречаемость обусловлена тем, что в пиках этот ТФ предпочитает либо один, либо другой структурный тип СС, но реже наблюдаются два СС разной структуры в одном пике. Если мотивы СС относятся к двум разным ТФ, то взаимоисключающую встречаемость можно объяснить тем, что в составе мультибелкового комплекса, включающего оба ТФ, в разных пиках один или другой ТФ связывается с ДНК напрямую, но реже наблюдаются также СС обоих ТФ в одном пике.

В ходе разработки ПК MetArea мы отказались от применения метрики частичной площади под кривой ROC (pAUC ROC) (Levitsky, Tsukanov, 2024) и использовали метрику площади по кривой PR (Davis, Goadrich, 2006) для определения метрики частичной площади под кривой PR. Ранее высказано предположение (Davis, Goadrich, 2006), что применение метрики площади под кривой AUC ROC может быть некорректным, если реальные пороги распознавания классификатора должны быть достаточно жесткими, а преимущество одной модели относительно другой набирается в интервале более мягких порогов распознавания (на правом хвосте кривой ROC). Для корректного сравнения двух мотивов в таком случае вместо меры площади под кривой AUC ROC мы ранее использовали меру «Частичная площадь под кривой ROC, pAUC», которая вместо полноразмерного диапазона ошибок перепредсказания (False Positive Rate, доля в распознавании объектов негативной выборки, ось *X* кривой ROC) от 0 до 1 использует только некую ее левую часть, отбрасывая диапазон слишком больших ошибок перепредсказания. Этот подход нами реализован для сравнения точности распознавания мотивов СС ТФ моделей ПВМ, BaMM и SiteGA (Tsukanov et al., 2022), где пороги распознавания при вычислении оценки точности pAUC ROC были ограничены условием ERR < 0.001.

К сожалению, такой способ не подходит для расчета оценки точности объединенного мотива, необходимого при реализации подхода MetArea. Причина этого в определении частоты объединенного мотива (т.е. числа его хитов), которое возможно при условии неперекрывания хитов одиночных мотивов, а в случае их перекрывания частота объединенного мотива должна некоторым образом понижаться. Альтернативным способом избавления от завышенной оценки точности, даваемой мерой AUC ROC, является переход от кривой ROC к кривой PR и вычисление площади под PR (Davis, Goadrich, 2006; Keilwagen et al., 2019).

Ранее было предложено несколько подходов, нацеленных на выявление встречаемости разных мотивов СС ТФ или разных наборов мотивов в различных долях пиков одного набора пиков ChIP-seq. Инструмент DIVERSITY (Mitra et al., 2018) разбивает пики ChIP-seq одного набора на несколько неперекрывающихся групп, так что каждая группа описывается своим обогащенным мотивом из результатов *de novo* поиска. Позднее авторы допустили, что каждая группа представляется не одним мотивом, а комбинацией нескольких мотивов. Инструмент cisDIVERSITY (Biswas, Narlikar, 2021) для набора пиков проводит *de novo* поиск обогащенных мотивов с помощью модели ПВМ, а затем разделяет найденные мотивы по нескольким неперекрывающимся группам пиков так, что все группы составляют весь набор пиков. Каждый из мотивов имеет разные частоты по группам; например, в некоторых группах частота мотива выше, чем в других группах, а в других группах мотива может не быть. Задачи инструментов DIVERSITY/cisDIVERSITY и MetArea сходны в том, что происходит разделение разных мотивов по разным долям пиков. Однако инструменты DIVERSITY/cisDIVERSITY: (1) выявляют все разнообразие мотивов и разделяют все пики на группы, с целью найти для разных групп разные мотивы или их комбинации; (2) ограничены традиционной моделью мотива ПВМ. ПК MetArea: (1) рассматривает только пары мотивов, с целью за счет максимизации меры точности pAUPRC для объединенного мотива найти пары, наилучшим образом дополняющие друг друга; (2) рассматривает как традиционную ПВМ, так и альтернативные модели мотива СС ТФ.

Заключение

Нами разработан ПК MetArea, который по заданному набору пиков ChIP-seq рассчитывает меру точности «Частичная площадь под кривой PR» (pAUPRC) для двух входных одиночных мотивов СС ТФ, определяет по ним объединенный мотив и также рассчитывает меру pAUPRC для него. Создание объединенного мотива по двум одиночным мотивам и расчет для него оценки точно-

сти pAUPRC позволяют сравнить в единой шкале два одиночных мотива и их общее действие. Превышение оценки точности объединенного мотива над оценками точности обоих одиночных мотивов означает их взаимоисключающую встречаемость. Результаты анализа с помощью ПК MetArea позволяют предсказывать функциональную связь двух мотивов, а значит, и соответствующих им ТФ. В частности, ПК MetArea может предложить существенные аргументы за или против гипотезы о том, что два мотива являются структурными вариантами СС одного ТФ. Аналогично предлагается поддержка или опровержение гипотезы о том, что мотивы СС представляют два ТФ, вместе вовлеченных в регуляцию транскрипции генов в составе одного мультибелкового комплекса. В итоге для заданного набора ChIP-seq данных ПК MetArea предсказывает (1) структурное разнообразие СС отдельных ТФ и (2) пары мотивов СС разных ТФ, действующих для регуляции транскрипции генов в составе одних мультибелковых комплексов из многих ТФ.

Список литературы / References

Ambrosini G., Vorontsov I., Penzar D., Groux R., Forne O., Nikolaeva D.D., Ballester B., Grau J., Grosse I., Makeev V., Kulakovskiy I., Buche P. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biol*. 2020;21:114. 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;37:2834-2840. doi 10.1093/bioinformatics/btab203

Biswas A., Narlikar L. A universal framework for detecting cis-regulatory diversity in DNA regions. *Genome Res*. 2021;31(9):1646-1662. doi 10.1101/gr.274563.120

Chen Y., Chi P., Rockowitz S., Iaquinta P.J., Shamu T., Shukla S., Gao D., Sirota I., Carver B.S., Wongvipat J., Scher H.I., Zheng D., Sawyers C.L. ETS factors reprogram the androgen receptor cistrome and prime prostate tumorigenesis in response to PTEN loss. *Nat. Med*. 2013;19(8):1023-1029. doi 10.1038/nm.3216

Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. New York: Assoc. for Computing Machinery, 2006;233-240. doi 10.1145/1143844.1143874

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

Garber M., Yosef N., Goren A., Raychowdhury R., Thielke A., Guttman M., Robinson J., Minie B., Chevrier N., Itzhaki Z., Blecher-Gonen R., Bornstein C., Amann-Zalcenstein D., Weiner A., Friedrich D., Meldrim J., Ram O., Cheng C., Gnirke A., Fisher S., Friedman N., Wong B., Bernstein B.E., Nusbaum C., Hacohen N., Regev A., Amit I. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell*. 2012;47(5):810-822. doi 10.1016/j.molcel.2012.07.030

Georgakopoulos-Soares I., Deng C., Agarwal V., Chan C.S.Y., Zhao J., Inoue F., Ahituv N. Transcription factor binding site orientation and order are major drivers of gene regulatory activity. *Nat. Commun*. 2023;14:2333. doi 10.1038/s41467-023-37960-5

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

Hess D.A., Strelau K.M., Karki A., Jiang M., Azevedo-Pouly A.C., Lee A.H., Deering T.G., Hoang C.Q., MacDonald R.J., Konieczny S.F. MIST1 links secretion and stress as both target and regulator of the unfolded protein response. *Mol. Cell. Biol*. 2016;36(23):2931-2944. doi 10.1128/MCB.00366-16

Hu G., Dong X., Gong S., Song Y., Hutchins A.P., Yao H. Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions. *Nucleic Acids Res*. 2020;48(17):9606-9620. doi 10.1093/nar/gkaa705

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

Keilwagen J., Posch S., Grau J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol*. 2019;20(1):9. doi 10.1186/s13059-018-1614-y

Kel O.V., Romaschenko A.G., Kel A.E., Wingender E., Kolchanov N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res*. 1995;23(20):4097-4103. doi 10.1093/nar/23.20.4097

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

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinformatics*. 2007;8(1):481. doi 10.1186/1471-2105-8-481

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:e139. doi 10.1093/nar/gkz800

Levitsky V., Oshchepkov D., Zemlyanskaya E., Merkulova T. Asymmetric conservation within pairs of co-occurred motifs mediates weak direct binding of transcription factors in ChIP-Seq data. *Int. J. Mol. Sci*. 2020;21(17):E6023. doi 10.3390/ijms21176023

Levitsky V.G., Tsukanov A.V. MetArea tool for predicting structural variability and cooperative binding of transcription factors in ChIP-seq data. In: 14th International Conference on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2024). 2024;136-138. doi 10.18699/bgrs2024-1.2-17

Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol*. 2018;14(4):e1006090. doi 10.1371/journal.pcbi.1006090

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

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

Rauluseviciute I., Riudavets-Puig R., Blanc-Mathieu R., Castro-Mondragon J.A., Ferenc K., Kumar V., Lemma R.B., Lucas J., Chèneby J., Baranasic D., Khan A., Fornes O., Gundersen S., Johansen M., Hovig E., 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

Rogers J.M., Waters C.T., Seegar T.C.M., Jarrett S.M., Hallworth A.N., Blacklow S.C., Bulyk M.L. Bispecific forkhead transcription factor

FoxN3 recognizes two distinct motifs with different DNA shapes. *Mol. Cell*. 2019;74(2):245-253.e6. doi 10.1016/j.molcel.2019.01.019

Saito T., Rehmsmeier M. The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi 10.1371/journal.pone.0118432

Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res*. 2016;44:6055-6069. doi 10.1093/nar/gkw521

Tognon M., Giugno R., Pinello L. A survey on algorithms to characterize transcription factor binding sites. *Brief. Bioinform*. 2023;24(3):bbad156. doi 10.1093/bib/bbad156

Tsukanov A.V., Levitsky V.G., Merkulova T.I. Application of alternative *de novo* motif recognition models for analysis of structural heterogeneity of transcription factor binding sites: a case study of FOXA2 binding sites. *Vavilov J. Genet. Breed*. 2021;25(1):7-17. doi 10.18699/VJ21.002

Tsukanov A.V., Mironova V.V., Levitsky V.G. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis. *Front. Plant Sci*. 2022;13:938545. doi 10.3389/fpls.2022.938545

Vorontsov I.E., Eliseeva I.A., Zinkevich A., Nikonov M., Abramov S., Boytsov A., Kamenets V., Kasianova A., Kolmykov S., Yevshin I.S., Favorov A., 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., Lambert S.A., Mann I., Cook K., Zheng H., Goity A., van Bakel H., Lozano J.C., Galli M., Lewsey M.G., Huang E., Mukherjee T., Chen X., Reece-Hoyes J.S., Govindarajan S., Shaulsky G., Walhout A.J.M., 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. 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

Yang Y.A., Yu J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes Dis*. 2015;2(2):144-151. doi 10.1016/j.gendis.2015.01.003

Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code. *Curr. Opin. Syst. Biol*. 2020;23:22-31. doi 10.1016/j.coisb.2020.08.002

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li W., Liu X.S. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9:R137. doi 10.1186/gb-2008-9-9-r137

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 19.10.2024. После доработки 20.11.2024. Принята к публикации 21.11.2024.