Вавиловский журнал генетики и селекции. 2021;25(1):7-17
DOI 10.18699/VJ21.002

Английский текст https://vavilov.elpub.ru/jour

Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2

А.В. Цуканов[1] ✉, В.Г. Левицкий[1, 2], 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
✉ tsukanov@bionet.nsc.ru

Аннотация. В настоящее время самой распространенной моделью поиска сайтов связывания транскрипционных факторов (ССТФ) в пиках ChIP-seq является позиционная весовая матрица (position weight matrix, PWM). Но эта модель не учитывает взаимосвязи между частотами встреч нуклеотидов в разных позициях ССТФ, поэтому не способна гарантировать определение всех возможных структурных вариантов ССТФ. На сегодняшний день уже предложены альтернативные модели, например BaMM и InMoDe, которые учитывают такие взаимосвязи. Однако применение этих моделей обычно сводилось к сравнению их точности с точностью традиционной модели PWM, тогда как анализ совместной встречаемости и относительного расположения ССТФ разных моделей в пиках не производился. В нашей работе мы предлагаем конвейер программ MultiDeNA, позволяющий сочетать разные модели *de novo* поиска ССТФ для выявления структурной гетерогенности ССТФ в данных ChIP-seq. Разработанный конвейер включает этапы построения моделей на основе заданного набора пиков, оценки точности распознавания моделей с помощью перекрестных тестов, выбора порогов, сканирования пиков ChIP-seq и классификацию пиков по результатам сканирования. С применением конвейера нами проведен анализ 22 экспериментов ChIP-seq для ТФ FOXA2 с помощью четырех моделей: PWM, diPWM, BaMM и InMoDe. Показано, что сочетание моделей позволяет существенно увеличить общее количество распознанных пиков (на 26.3 %) по сравнению с применением только PWM; при этом основной вклад в распознавание внесла модель BaMM. В значительной доле пиков разные модели распознают совпадающие ССТФ; однако для моделей PWM, diPWM, BaMM и InMoDe медианы доли пиков, которые содержали ССТФ только одной модели, составили 1.08, 0.49, 4.15 и 1.73 % соответственно. Таким образом, совокупность ССТФ FOXA2 не описывается полностью только одной моделью, что свидетельствует о наличии структурной гетерогенности в ССТФ у FOXA2.
Ключевые слова: сайты связывания транскрипционных факторов (ССТФ); *de novo* поиск ССТФ; ChIP-seq; гетерогенность ССТФ.

Для цитирования: Цуканов А.В., Левицкий В.Г., Меркулова Т.И. Метод поиска структурной гетерогенности сайтов связывания транскрипционных факторов с использованием альтернативных *de novo* моделей на примере FOXA2. *Вавиловский журнал генетики и селекции*. 2021;25(1):7-17. DOI 10.18699/VJ21.002

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

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

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

Abstract. The most popular model for the search of ChIP-seq data for transcription factor binding sites (TFBS) is the positional weight matrix (PWM). However, this model does not take into account dependencies between nucleotide occurrences in different site positions. Currently, two recently proposed models, BaMM and InMoDe, can do as much. However, application of these models was usually limited only to comparing their recognition accuracies with that of PWMs, while none of the analyses of the co-prediction and relative positioning of hits of different models in peaks has yet been performed. To close this gap, we propose the pipeline called MultiDeNA. This pipeline includes stages of model training, assessing their recognition accuracy, scanning ChIP-seq peaks and their classification based on scan results. We applied our pipeline to 22 ChIP-seq datasets of TF FOXA2 and considered PWM, dinucleotide PWM (diPWM), BaMM and InMoDe models. The combination of these four models allowed a significant increase in the fraction of recognized peaks compared to that for the sole PWM model: the increase was 26.3 %. The BaMM model provided the main contribution to the recognition of sites. Although the major fraction of predicted peaks contained TFBS of different models with coincided positions, the medians of the fraction of peaks containing the predictions of sole models

were 1.08, 0.49, 4.15 and 1.73 % for PWM, diPWM, BaMM and InMoDe, respectively. Thus, FOXA2 BSs were not fully described by only a sole model, which indicates theirs heterogeneity. We assume that the BaMM model is the most successful in describing the structure of the FOXA2 BS in ChIP-seq datasets under study.
Key words: transcription factor binding sites (TFBS); TFBS *de novo* searching; ChIP-seq; heterogeneity of TFBS.

For citation: 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. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):7-17. DOI 10.18699/VJ21.002

Введение

Транскрипционные факторы (ТФ) – белки, способные распознавать определенные участки ДНК (сайты связывания ТФ, ССТФ) (Lambert et al., 2018) и как повышать, так и снижать уровень транскрипции генов (Latchman, 2001). Этап связывания ТФ с ДНК является ключевым для регуляции экспрессии генов, поскольку инициирует цепь молекулярных событий, обеспечивающих сборку/регуляцию активности преинициаторного комплекса РНК-полимеразы II за счет непосредственных или опосредованных контактов с компонентами этого комплекса, а также благодаря привлечению различных модифицирующих хроматин и ремоделирующих белков и, как следствие, локальных изменений структуры хроматина (Iwafuchi-Doi, 2019; Srivastava, Mahony, 2020). Поэтому одна из важнейших задач современной молекулярной биологии – это идентификация всего массива ССТФ в геноме.

В настоящее время для решения этой задачи широко применяется метод, основанный на иммунопреципитации хроматина с использованием антител к исследуемому ТФ с последующим высокопроизводительным секвенированием преципитированной ДНК – ChIP-seq (Farnham, 2009; Park, 2009). Первичная обработка данных экспериментальных методов позволяет выявлять участки ДНК, или пики, для которых ТФ напрямую или через некоторого посредника был связан с ДНК (Furey, 2012). Поскольку длина пиков обычно исчисляется в сотнях пар оснований (п. о.), а протяженность ССТФ не превышает 20–25 п. о. (Levitsky et al., 2007; Kulakovskiy et al., 2018), следующим этапом биоинформатической обработки данных ChIP-seq является поиск ССТФ в полученных пиках. Для этого разработано множество инструментов, большинство из которых основано на использовании позиционных весовых матриц (position weight matrix, PWM) (Stormo, 2000), включая такие популярные, как ChIPMunk (Kulakovskiy, Makeev, 2009) и Homer (Heinz et al., 2010). Без преувеличения можно сказать, что применение разных реализаций модели PWM входит практически в каждый конвейер обработки полногеномных данных (Lloyd, Bao, 2019).

Применение стандартного подхода, основанного на использовании PWM, к обработке данных ChIP-seq показывает, что примерно в половине пиков для большинства ТФ не обнаруживается соответствующих мотивов (Worsley Hunt, Wasserman, 2014; Gheorghe et al., 2019). Традиционно это связывают с главным недостатком PWM – гипотезой независимости частот встреч нуклеотидов в разных позициях ССТФ, которая не всегда подтверждается, что негативно сказывается на точности распознавания (Benos et al., 2002; Keilwagen, Grau, 2015). Поэтому разрабатываются альтернативные методы распознавания ССТФ, где тем или иным способом учитываются зависимости между нуклеотидами в модели сайта (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). С одной стороны, самой простой моделью, которая старается учитывать зависимости между соседними нуклеотидами, является динуклеотидная PWM (dinucleotide position weight matrix, diPWM) (Zhang M., Marr, 1993; Kulakovskiy et al., 2013). С другой стороны, предложены такие модели, как BaMM (Siebert, Söding, 2016) и InMoDe (Eggeling et al., 2017). Они построены с использованием марковских цепей, которые учитывают зависимости позиций с помощью концепции порядка марковской цепи, т. е. участка, длина которого обычно не превышает 5 п. о. и в пределах которого частоты нуклеотидов могут быть зависимыми.

Авторы альтернативных моделей часто доказывают, что их модели могут иметь более высокую точность, чем PWM, однако ни одна из этих моделей сама по себе не решает проблему неполного распознавания ССТФ в пиках ChIP-seq. Мы предполагаем, что частично проблема обусловлена структурной гетерогенностью сайтов связывания для одного и того же ТФ и число распознанных пиков может быть значительно увеличено при одновременном использовании разных моделей. При этом данные ChIP-seq будут содержать как ССТФ, предсказываемые одновременно двумя и более моделями, так и ССТФ, предсказываемые только одной из моделей (Ignatieva et al., 2004; Levitsky et al., 2014, 2016). Ранее при анализе двух независимых экспериментов ChIP-seq для ТФ FOXA2 (Wederell et al., 2008; Wallerman et al., 2009) с помощью альтернативных моделей ChIPMunk (*de novo* PWM (Kulakovskiy, Makeev, 2009)) и SiteGA (по выборке обучения из 53 известных сайтов ТФ подсемейства FOXA (Levitsky et al., 2007)) и экспериментально подобранных порогов моделей (эксперимент EMSA, electrophoretic mobility shift assay – сдвиг в анализе электрофоретической подвижности) удалось обнаружить FOXA2 сайты более чем в 95 % пиков (Levitsky et al., 2014), что согласуется с отсутствием в литературе каких-либо данных о непрямом взаимодействии этого хорошо изученного ТФ с ДНК.

Приведенный пример указывает на перспективность сочетания альтернативных методов поиска ССТФ с матричной моделью для анализа ChIP-seq данных. Однако до сих пор не было систематических исследований на эту тему. Альтернативные модели для поиска ССТФ не получили широкого применения, несмотря на то что уже около 20 лет известно о наличии зависимости частот встреч нуклеотидов в разных позициях ССТФ (Bulyk et al., 2002). В качестве косвенного показателя популярности разных моделей можно привести количество цитирований статей, в которых обсуждаются конкретные программы *de novo*

поиска ССТФ. Так, на конец 2020 г. статьи, посвященные реализации матричной модели в виде программ MEME (Bailey, Elkan, 1994; Machanick, Bailey, 2011), HOMER (Heinz et al., 2010) и ChIPMunk (Kulakovskiy et al., 2010), имеют суммарное количество цитирований более 6000, а статьи, посвященные альтернативным моделям BaMM (Siebert, Söding, 2016; Kiesel et al., 2018), InMoDe (Eggeling et al., 2017) и diChIPMunk (Kulakovskiy et al., 2013), – чуть более 50. При этом конкретные исследования (отдельные эксперименты ChIP-seq) почти всегда анализируются только с использованием стандартной модели PWM. Такое положение можно объяснить следующими причинами: 1) простота применения PWM и доступность в понимании результатов этой модели; 2) недостаточное понимание преимуществ альтернативных моделей, которые, помимо лучшей точности в сравнении с PWM, способны находить ССТФ иной структуры.

В данной работе мы предлагаем конвейер программ, который сочетает четыре модели *de novo* поиска ССТФ, а именно: PWM, реализованную в программе ChIPMunk (Kulakovskiy et al., 2010); diPWM, реализованную в программе diChIPMunk (Kulakovskiy et al., 2013); и две марковские модели – InMoDe (Eggeling et al., 2017) и BaMM (Siebert, Söding, 2016). Конвейер оценивает точность распознавания моделей, выбирает их пороги и проводит классификацию ChIP-seq пиков, сравнивая результаты сканирования всех моделей. Такой подход позволит расширить наши представления о структурном разнообразии ССТФ при прямом связывании ТФ с ДНК, особенно для случаев, когда модель PWM не способна найти ССТФ. Работа конвейера апробирована в ходе анализа данных 22 экспериментов ChIP-seq для ТФ FOXA2.

Материал и методы

Исходные данные. Для анализа использовали набор предобработанных ChIP-seq данных в виде разметки пиков в формате bed из базы данных ReMap http://remap.univ-amu.fr/ (Chèneby et al., 2020). Набор данных включал 22 ChIP-seq эксперимента для ТФ FOXA2 (см. таблицу). Из каждого эксперимента для анализа брали только лучшие 4000 пиков (см. далее раздел «Подготовка первичных данных»).

Список ChIP-seq экспериментов, используемых в работе

№ п/п	GEO/ ENCODE ID	Клеточная линия/ ткань	Обработка	TomTom
1	ENCSR066EBK	Hep-G2	–	+
2	GSE90454	BJ1-hTERT	Mimosine	+
3	GSE90454	A-549	–	+
4	ENCSR000BRE	A-549	–	+
5	GSE92491	BJ1-hTERT	Mimosine	+
6	GSE90454	BJ1-hTERT	–	+
7	ENCSR080XEY	Liver	–	+
8	ENCSR310NYI	Liver	–	+
9	ENCSR000BNI	Hep-G2	–	+
10	GSE90454	BJ1-hTERT	–	+
11	ERP004206	H9	–	+
12	GSE92491	BJ1-hTERT	Mimosine	–
13	GSE90454	KerCT	–	+
14	GSE90454	BJ1-hTERT	Mimosine	–
15	GSE90454	BJ1-hTERT	Mimosine	+
16	GSE90454	BJ1-hTERT	Mimosine	+
17	GSE90454	BJ1-hTERT	GATA4	–
18	ERP008682	Pancreas	CARN1618	+
19	GSE90454	BJ1-hTERT	Mimosine	–
20	GSE92491	BJ1-hTERT	CDT1	+
21	GSE90454	Hep-G2	–	–
22	GSE92491	BJ1-hTERT	FOXA2 and GATA4 coexpression	–

Примечание. GEO/ENCODE – уникальный идентификатор баз данных (GSE*/ENC*); TomTom – результат фильтрации данных с помощью программы TomTom; «+»/«–» – частотная матрица, построенная на основе ССТФ, найденных ChIPMunk (PWM), значимо похожа (*p*-value < 0.001)/не похожа (*p*-value > 0.001) на частотную матрицу ССТФ FOXA2 из HOCOMOCO FOXA2_HUMAN.H11MO.0.A.

Помимо ChIP-seq пиков, на вход в конвейер программ указывали список доступных программ (PWM, diPWM, BaMM, InMoDe) *de novo* поиска ССТФ, включая путь к программам. Также устанавливали версию генома – mm10 или hg38; этот параметр позволяет выбрать список промоторов в формате fasta – 5′-участки кодирующих белок генов (2000 п. о. от сайта старта транскрипции). Общий объем выборки составил 19795 генов для версии генома человека GRCh38.p13 и 19991 ген для версии генома мыши GRCm38.p6. Для извлечения последовательностей нуклеотидов по координатам пиков использовали референсный геном в формате fasta.

Конвейер программ для выявления структурной гетерогенности ССТФ. Нами был разработан конвейер программ MultiDeNA (multiple *de novo* analysis, https://github.com/ubercomrade/MultiDeNA) для поиска ССТФ с помощью нескольких *de novo* моделей в данных ChIP-seq. Данный конвейер программ позволяет получить классификацию пиков ChIP-seq, по результатам которой можно оценить структурное разнообразие ССТФ. В настоящее время конвейер использует модели ChIPMunk (PWM), diChIPMunk (diPWM), BaMM и InMoDe, а также вспомогательные программы bedtools (Quinlan, Hall, 2010) и TomTom (Gupta et al., 2007). Принципиальная схема конвейера программ представлена на рис. 1. Конвейер включает в себя следующие этапы: подготовка данных; построение моделей; оценка точности моделей; выбор порогов, поиск ССТФ в пиках ChIP-seq с фиксированным порогом; классификация ChIP-seq пиков по результатам распознавания ССТФ *de novo* моделями. Каждый этап конвейера программ детально описан ниже.

Подготовка первичных данных включала сортировку пиков по округленному значению $-10 \cdot \log_{10}(p\text{-value})$, которое было ранее вычислено для каждого пика программой MACS (Zhang Y. et al., 2008) и характеризовало

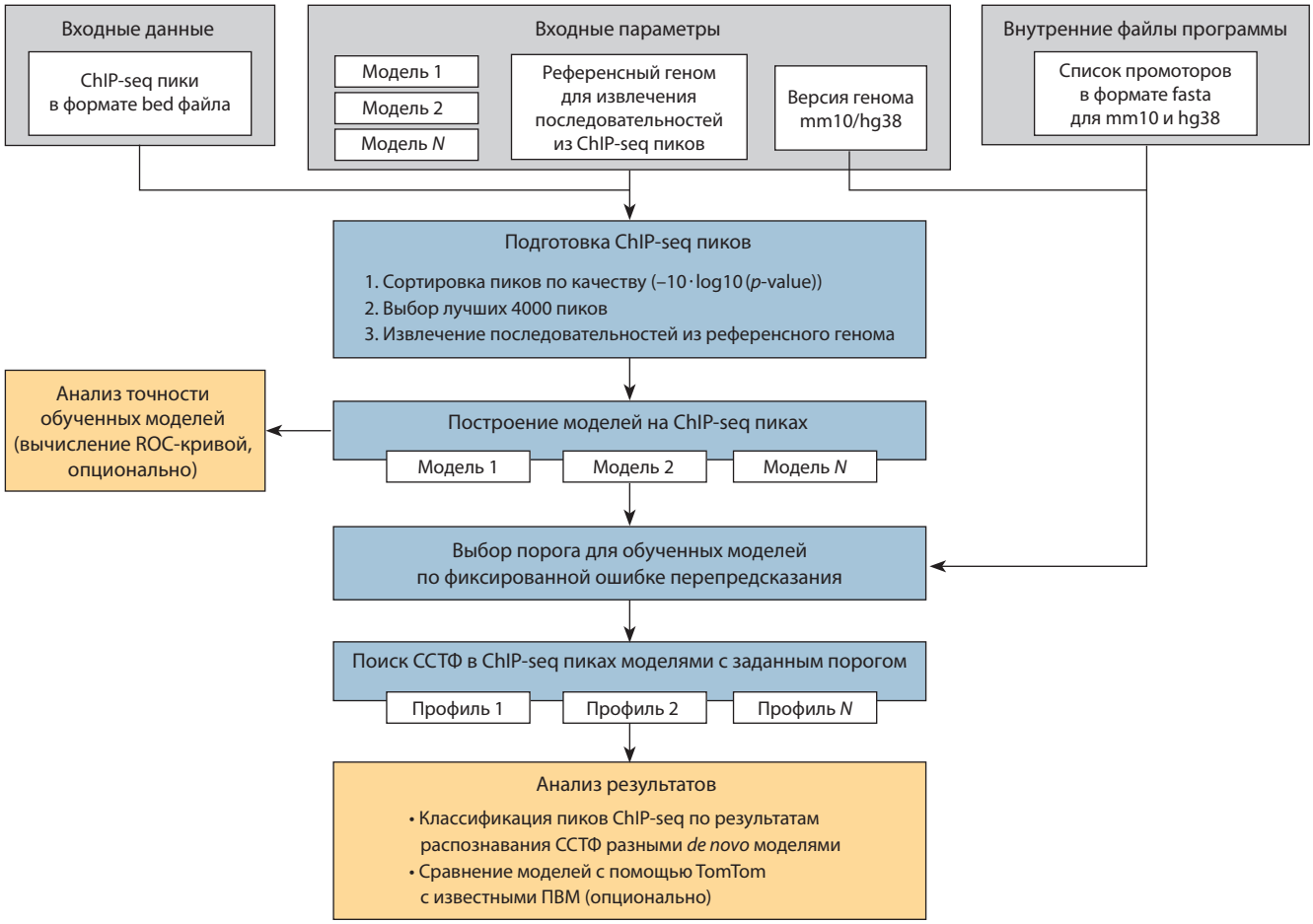


Рис. 1. Принципиальная схема работы конвейера программ.

качество пика. Эту программу конвейер базы ReMap (Chèneby et al., 2020) использовал для обработки сырых данных ChIP-seq. Из каждого набора данных ChIP-seq для анализа мы взяли 4000 лучших по качеству пиков. Далее извлекали нуклеотидные последовательности пиков из генома с помощью bedtools (Quinlan, Hall, 2010).

Построение *de novo* моделей и оценка их точности распознавания ССТФ. Для того чтобы распознавать ССТФ в пиках, необходимо построить *de novo* модели. Построение нетрадиционных моделей ССТФ осуществлялось программами BaMM (Siebert, Söding, 2016) и InMoDe (Eggeling et al., 2017), а модели PWM и diPWM строили соответственно с помощью ChIPMunk и diChIPMunk (Kulakovskiy et al., 2010, 2013).

Чтобы улучшить точность распознавания ССТФ для PWM, подбирали ее оптимальную длину методом перекрестных тестов; эту же длину использовали и при построении других моделей. Метод оценки точности включал следующие этапы: 1) разделение данных на выборку обучения – случайно отобранные 90 % пиков от исходных данных, и контрольную выборку, включавшую оставшиеся 10 % пиков; 2) построение модели на выборке обучения; 3) проверка модели на контрольной выборке для оценки доли верноположительных результатов (ДВР); 4) генерация выборки случайных последовательностей путем случайной перестановки нуклеотидов в последовательностях контрольной выборки; 5) проверка модели на выборке случайных последовательностей для оценки доли ложноположительных результатов (ДЛР); 6) повторение этапов 1–5 несколько раз; 7) вычисление ROC-кривой (receiver operating characteristic) на основе полученных данных. Разные длины модели сравнивали по показателю pAUC (partial area under curve), вычисленному как часть площади под кривой ROC для всех значений ДЛР, меньших 0.001 (McClish, 1989; Siebert, Söding, 2016). Описанный выше способ выбора оптимальной длины PWM на основе наилучшей точности распознавания ССТФ был разработан ранее (Levitsky et al., 2007; Kulakovskiy et al., 2013). Аналогичным методом оценивали точность всех моделей.

После того как модель построена, ее можно применять к последовательности нуклеотидов, равной длине модели. Результатом применения модели является значение функции распознавания. Чем больше это значение, тем выше вероятность того, что оцениваемая последовательность нуклеотидов является функциональным ССТФ.

Выбор порога для моделей на основе фиксированной ошибки перепредсказания. Чтобы корректно сравнивать результаты поиска ССТФ разных моделей, необходимо единообразно установить для всех моделей пороговые значения их функций распознавания. Эти пороги определяли по фиксированной ошибке перепредсказания. Для ее

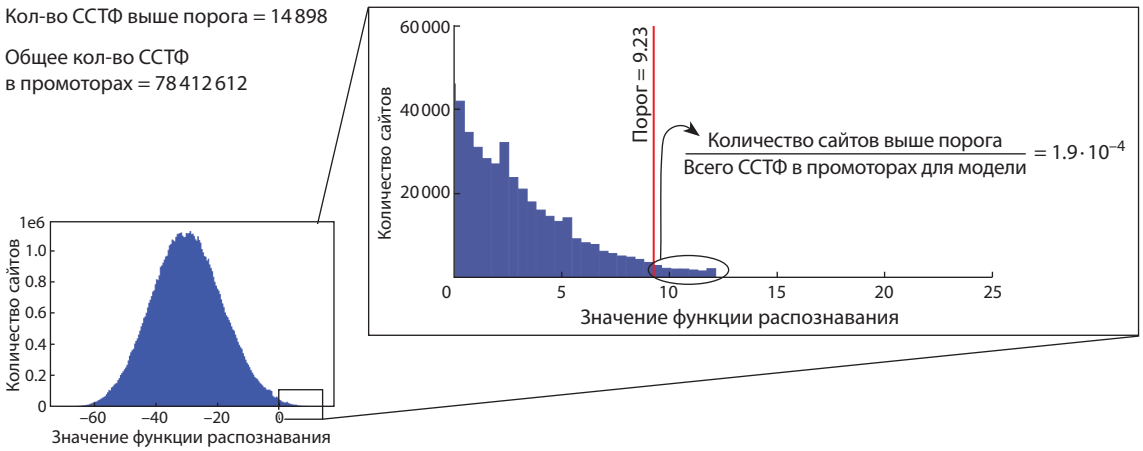


Рис. 2. Выбор порога для модели по фиксированной ошибке перепредсказания с использованием в качестве негативной выборки последовательности промоторов.

вычисления использовали негативную выборку, в которую входили 5′-участки кодирующих белок генов (2000 п. о. от сайта старта транскрипции).

Величину ошибки перепредсказания вычисляли следующим образом. Определяли значение функции распознавания модели для каждого сайта в негативной выборке в каждой позиции и цепи ДНК. Затем оценивали величину ошибки перепредсказания для каждого уникального значения функции распознавания как отношение количества предсказанных ССТФ, для которых значение функции выше этого порога, к общему числу позиций в выборке, доступных для таких ССТФ. При распознавании ССТФ для всех моделей в качестве порога использовали такое значение функции распознавания, при котором ошибка перепредсказания составляла $1.9 \cdot 10^{-4}$. После того как порог выбран для каждой модели, сканировали пики ChIP-seq. Пример выбора порога для PWM длиной 20 п. о. на данных GSE92491 приведен на рис. 2.

Классификация пиков ChIP-seq по результатам распознавания ССТФ разными моделями. После того как для каждой модели был выбран порог, мы искали ССТФ в пиках ChIP-seq. Результаты сканирования записывали в файл bed формата. Далее пики классифицировали на фракции в зависимости от присутствия/отсутствия сайтов, найденных разными моделями (PWM, diPWM, BaMM, InMoDe), как с учетом расположения ССТФ разных моделей в позициях пиков, так и без такого учета (на основе присутствия или отсутствия сайтов в пиках), согласно ранее разработанной методике (Levitsky et al., 2014, 2016). В частности, классификацию пиков с учетом позиций ССТФ разных моделей проводили для каждой пары моделей. Всего было шесть пар моделей: PWM и diPWM, PWM и BaMM, PWM и InMoDe, BaMM и diPWM, BaMM и InMoDe, InMoDe и diPWM. Если в пике присутствовали ССТФ, предсказанные только одной моделью, то данный пик классифицировался как пик соответствующей модели. Если в пике найдены ССТФ, предсказанные двумя разными моделями, то возможны два исхода (рис. 3).

В первом случае, если существует хотя бы одна пара сайтов от разных моделей, которые имеют как минимум одну общую позицию, такой пик классифицируется как

Рис. 3. Пример классификации ChIP-seq двух пиков, в которых обнаружены сайты двух разных моделей: *а* – в пике сайты пересекаются; *б* – пересечения нет.

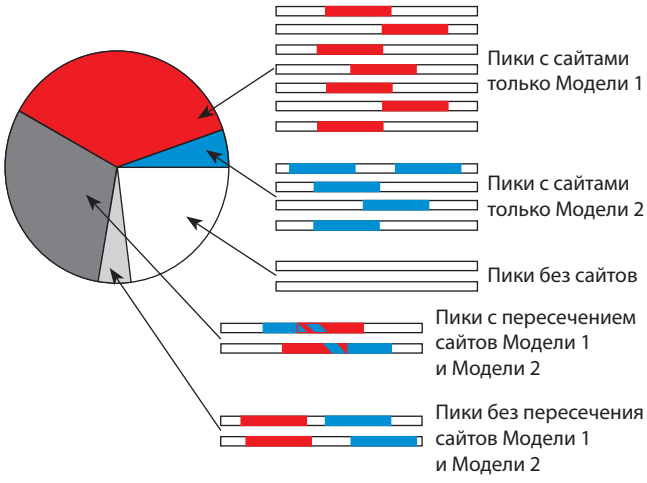


Рис. 4. Классификация ChIP-seq пиков для двух моделей с учетом пересечения ССТФ.

«пересечение сайтов». В другом случае, когда в пике присутствуют ССТФ, найденными разными методами, но их последовательности не пересекаются, пик классифицируется как «нет пересечения». Если в пике нет сайтов, то он классифицируется как «нет сайтов». Представить такую классификацию ChIP-seq пиков для двух моделей можно в виде круговой диаграммы (рис. 4).

Классификацию пиков без учета позиций ССТФ от разных моделей проводили следующим образом. Выделяли группы пиков, где присутствуют только сайты одной из моделей, пики, содержащие сайты всех моделей, а также пики, содержащие сайты комбинации моделей.

Сравнение найденных ССТФ с известными с помощью программы TomTom. Чтобы оценить, соответствуют ли ССТФ, которые находят модели, известным сайтам FOXA2, мы применили программу сравнения мотивов TomTom (Gupta et al., 2007). Эта программа предназначена для оценки значимости схожести частотных матриц. Для каждой PWM модели на основе найденных с ее помощью сайтов строили матрицу частот нуклеотидов. Далее с помощью TomTom оценивали схожесть этой матрицы с частотной матрицей ССТФ FOXA2 из базы данных HOCOMOCO FOXA2_HUMAN.H11MO.0.A (Kulakovskiy et al., 2018). Если при сравнении матриц значение *p*-value было меньше 0.001, то считали, что ChIP-seq обогащен ССТФ FOXA2 (см. таблицу).

Статистический анализ и визуализацию данных выполняли на языке программирования Python 3.8 в среде Jupyter с использованием пакетов numpy, matplotlib, seaborn и statannot. Распределения сравнили с помощью U-критерия Манна–Уитни с поправкой на множественные сравнения Бонферрони.

Результаты и обсуждение

Фильтрация данных на основе сравнения мотивов программой TomTom

Чтобы убедиться, что построенные модели сайтов соответствуют известным сайтам FOXA2 и последующий анализ является корректным, применили фильтр на основе программы оценки сходства мотивов TomTom. Для этого частотные матрицы ССТФ для модели PWM сравнивали с соответствующими матрицами известных ССТФ из базы данных HOCOMOCO. Только в шести из 22 ChIP-seq наборов, согласно TomTom, построенная матричная модель не обладала сходством с известными сайтами FOXA2 (см. таблицу), поэтому в дальнейшем анализе использовали оставшиеся 16 наборов.

Классификация пиков ChIP-seq без учета пересечения ССТФ, найденных разными *de novo* моделями

Основным результатом работы MultiDeNA является классификация пиков, которая позволяет установить, как соотносятся модели между собой, по способности выявлять пики с ССТФ. Всего используются два типа классификации пиков: с учетом пересечения позиций ССТФ разных моделей и без него. Результаты классификации приведем на примере данных GSE90454.FOXA2.KerCT (рис. 5).

Рассмотрим более детально классификацию ChIP-seq пиков по результатам поиска ССТФ четырьмя моделями без учета позиций сайтов. Можно видеть, что все четыре модели совместно распознали 88.35 % пиков (3534 из 4000, сумма всех областей на диаграмме Венна, см. рис. 5, *а*, *б*). Общая для всех методов группа пиков, в которых ССТФ были найдены четырьмя моделями одновременно, составила 34.25 % (1370 из 4000 пиков). Значительный вклад в распознание пиков (34.55 %) вносят нематричные методы (BaMM и InMoDe): 696 + 647 + 39 = = 1382 из 4000, что сопоставимо с фракцией перекрывания всех моделей (1370). При этом самый крупный независимый вклад в распознавание вносит модель BaMM, которая добавляет 17.4 % пиков (696), в отличие от моделей PWM, InMoDe и diPWM, которые добавляют 0.525 % (21), 0.975 % (39) и 0.2 % (8) соответственно.

Чтобы оценить структурное разнообразие ССТФ, мы построили лого для фракций пиков «только PWM», «только diPWM», «только BaMM», «только InMoDe» и «все модели» (см. рис. 5, *в*). Во всех полученных лого можно выделить стандартный консенсус GTAAACA, однако для первых двух нуклеотидов консенсуса у фракций «только PWM», «только diPWM» и «только InMoDe» частота встречаемости GT меньше, чем AT. Можно также отметить, что 5′-концы всех лого разнообразны по информационному и нуклеотидному содержанию.

Классификация пиков ChIP-seq с учетом пересечения ССТФ, найденных разными моделями

Описанная выше классификация пиков без учета позиций ССТФ не учитывает тот факт, что используемые нами модели могут находить сайты в разных позициях одного и того же пика. Чтобы принять во внимание данное обстоятельство, была проведена классификация пиков с учетом позиций ССТФ для каждой пары моделей (PWM–diPWM, PWM–BaMM, PWM–InMoDe, diPWM–BaMM, diPWM–InMoDe, InMoDe–BaMM). Результаты классификации пиков на примере данных GSE90454.FOXA2.KerCT показаны в виде круговых диаграмм (рис. 6).

Все пары сочетаний моделей имеют незначительный класс пиков «нет пересечения», который варьирует от 0.3 до 6.9 %. С другой стороны, для всех случаев характерна большая фракция пиков «только пересечение»: BaMM–InMoDe – 53.6 %, PWM–diPWM – 44.4 %, diPWM–BaMM – 41.0 %, PWM–BaMM – 37.3 %, diPWM–InMoDe – 35.4 %, PWM–InMoDe – 31.6 %; при этом данная фракция больше для методологически близких пар моделей BaMM–InMoDe и PWM–diPWM (см. рис. 6). Класс пиков, где ССТФ находятся только одной из моделей, наиболее выражен для BaMM. В парах PWM–BaMM, diPWM–BaMM и InMoDe–BaMM он преобладает относительно второй модели пары (39.2, 36.4 и 26.8 % соответственно).

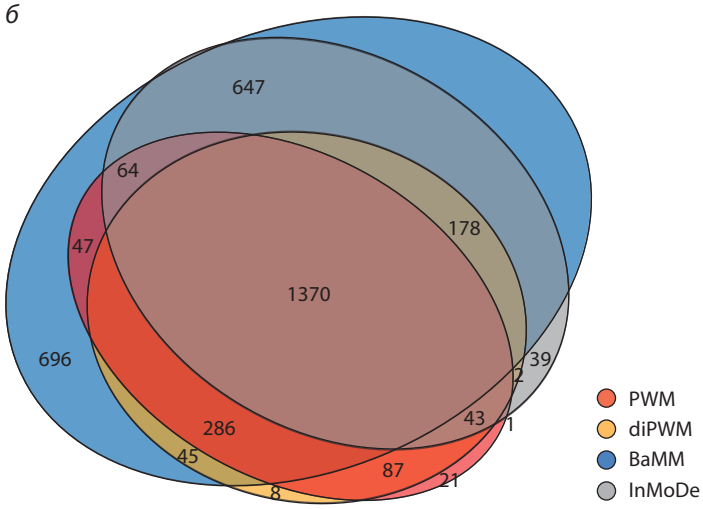
Оценка точности распознавания ССТФ для FOXA2 разными моделями

Чтобы сравнить, насколько точно разные модели способны распознавать ССТФ для FOXA2, по каждому эксперименту для всех четырех моделей рассчитали меру точности распознавания pAUC по кривой ROC, полученной с помощью перекрестного теста (см. выше раздел «Построение *de novo* моделей и оценка их точности распознавания ССТФ») (рис. 7, *а*). Согласно полученным данным, значения медиан pAUC для моделей PWM, diPWM, BaMM и InMoDe равны 8.0E–4, 8.1E–4, 7.3E–4 и 5.6E–4 соответственно. Полученные значения pAUC в парных сравнениях для PWM, diPWM и BaMM значимо не отличаются ($p > 0.05$), однако для InMoDe оно достоверно меньше, чем у остальных моделей ($p < 0.05$).

а

Пики, содержащие ССТФ, распознанные														
всеми моделями	всеми моделями				только двумя моделями						только одной моделью			
	кроме PWM	кроме diPWM	кроме BaMM	кроме InMoDe	PWM diPWM	PWM BaMM	PWM InMoDe	diPWM BaMM	diPWM InMoDe	BaMM InMoDe	PWM	diPWM	BaMM	InMoDe
1370	178	64	43	286	87	47	1	45	2	647	21	8	696	39

б



в

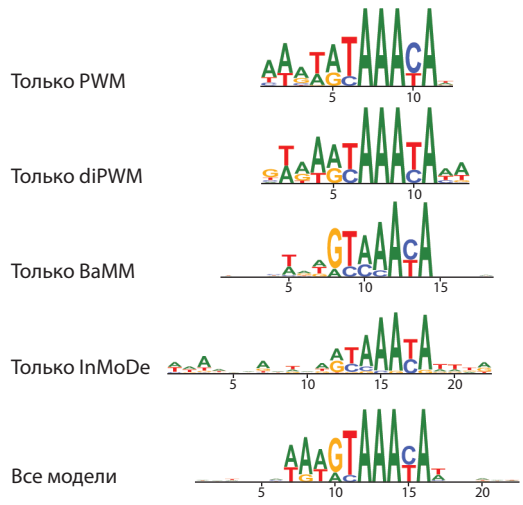


Рис. 5. Классификация пиков по результатам сканирования всеми четырьмя моделями.

а – таблица; *б* – диаграмма Венна; *в* – лого для фракций пиков, содержащих сайты только одной из моделей, и для фракции, где сайты всех моделей пересечены. Проанализирован набор данных GSE90454.FOXA2.KerCT.

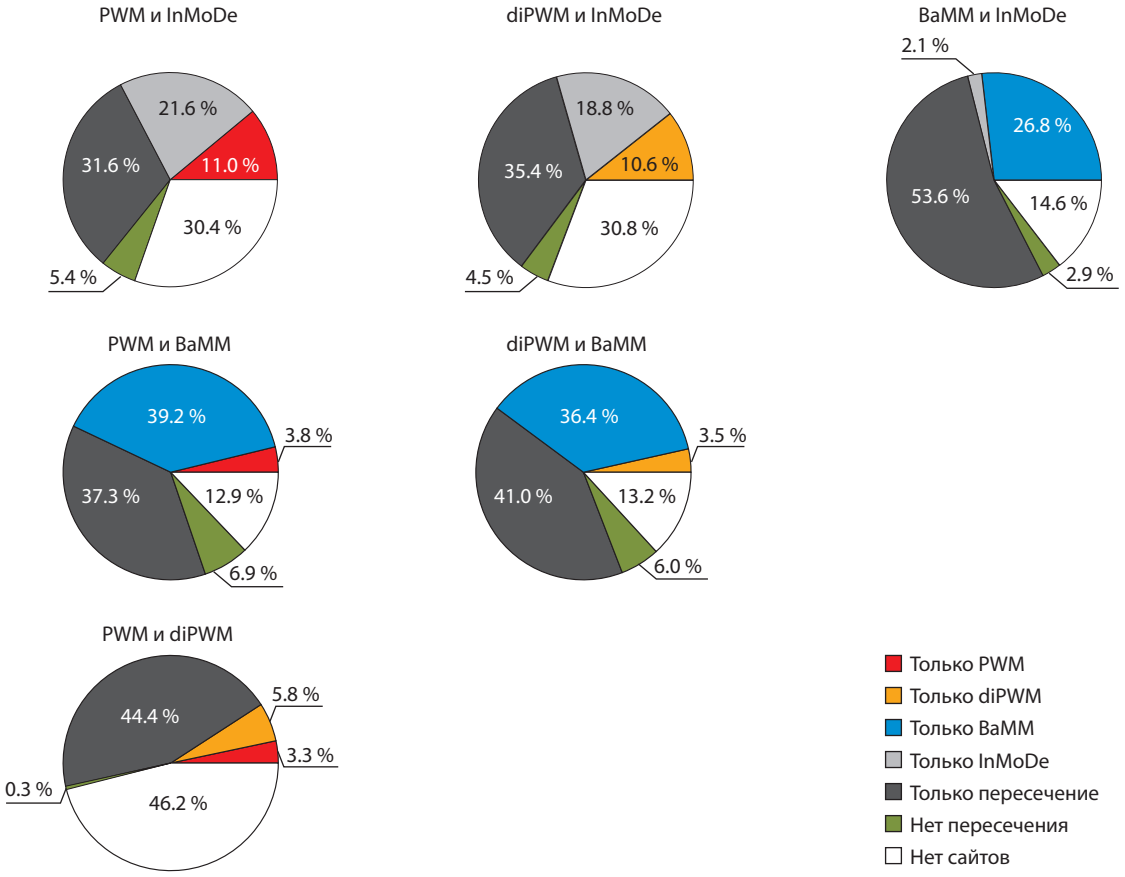


Рис. 6. Классификация ChIP-seq пиков с учетом пересечения ССТФ, распознанных разными моделями на примере данных GSE90454.FOXA2.KerCT.

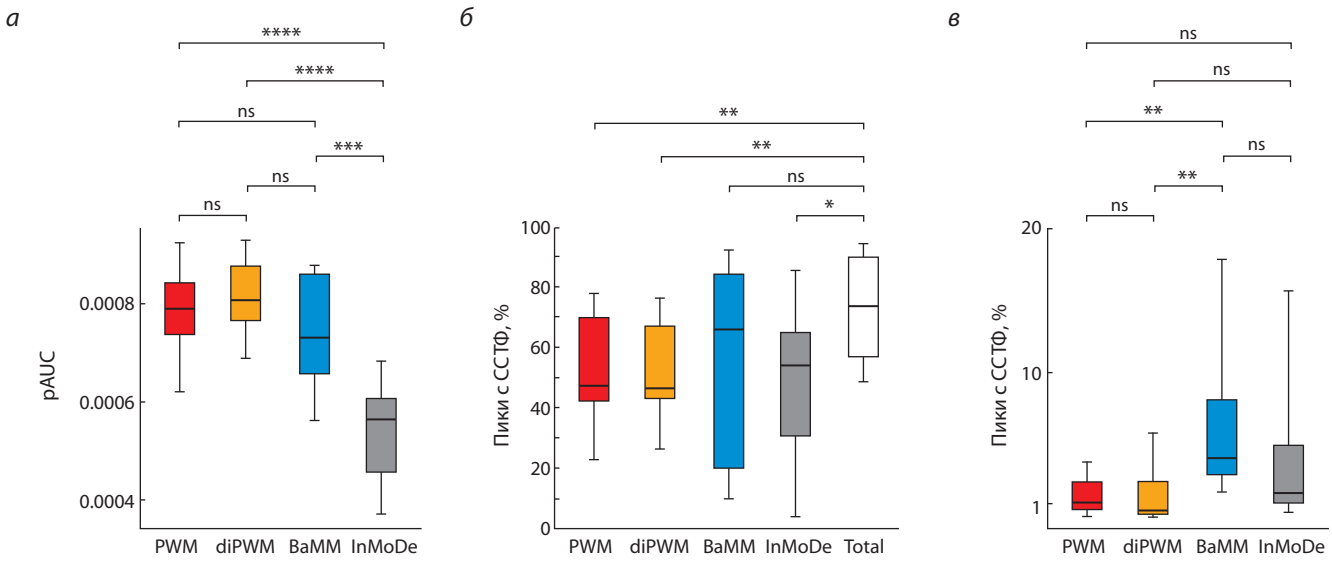


Рис. 7. Диаграммы распределения квартилей для данных: *а* – значения pAUC для всех моделей по всем ChIP-seq экспериментам; *б* – значения доли пиков с ССТФ, распознанных каждой моделью в отдельности (PWM, diPWM, BaMM, InMoDe) и всеми моделями (Total); *в* – значения доли пиков, в которых ССТФ находится только одной из моделей.

ns – $p > 0.05$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$.

Сравнение долей пиков с ССТФ, найденных каждой моделью и всеми моделями. Чтобы исследовать вклады разных моделей в эффективность поиска ССТФ FOXA2 и оценить общий результат использования нескольких моделей для поиска ССТФ, мы определили, в какой доле пиков каждая модель и все модели вместе распознают хотя бы один ССТФ для FOXA2 (см. рис. 7, *б*). Значения медиан доли распознанных пиков составили 47.3, 46.4, 65.8 и 54 % для PWM, diPWM, BaMM и InMoDe соответственно, а медиана доли распознанных пиков при сочетании результатов всех четырех моделей равна 73.6 %. Следовательно, совместно все модели находят на 26.3 % больше пиков, содержащих ССТФ, чем модель PWM, что согласуется с ранее полученным результатом применения двух принципиально разных моделей PWM и SiteGA (Levitsky et al., 2014). При этом доли распознанных пиков для моделей PWM, diPWM и InMoDe значимо отличаются ($p < 0.05$) от результата, полученного сочетанием четырех моделей. Таким образом, подход с сочетанием разных моделей позволяет лучше выявлять пики с ССТФ для FOXA2, чем использование только одной модели. Однако для BaMM доля распознанных пиков статистически не отличается ($p > 0.05$) от результата, полученного сочетанием четырех моделей. Можно предположить, что модель BaMM вносит основной вклад в распознавание пиков FOXA2 и, возможно, лучше описывает структуру сайтов FOXA2. Тем не менее остальные модели добавляют еще 7.8 % пиков к результатам BaMM, что доказывает эффективность совместного использования разных моделей.

Сравнение долей пиков, содержащих ССТФ, распознанные только одной из моделей. Как показано выше, сочетание разных моделей увеличивает количество пиков с ССТФ, соответственно каждая модель должна распознавать ССТФ, которые не распознаются остальными. Чтобы оценить вклады в поиск ССТФ, специфичных только для конкретной модели, были определены доли пиков, содержащих ССТФ только одной из моделей (см. рис. 7, *в*). Как видно из представленных данных, каждая модель (PWM, diPWM, BaMM, InMoDe) способна находить ССТФ, которые не обнаруживаются остальными моделями. Значения медиан по доле пиков, содержащих ССТФ только одной из моделей, для PWM, diPWM, BaMM и InMoDe составили 1.08, 0.49, 4.15 и 1.73 % соответственно. При этом данные по BaMM значимо отличаются ($p < 0.05$) как от PWM, так и от diPWM. Полученный результат подтверждает предположение, что модель BaMM может лучше описывать ССТФ FOXA2. Тем не менее каждая модель вносит вклад в распознавание сайтов. Следовательно, каждая из моделей может выявлять один из структурных вариантов ССТФ, который другие модели не находят.

Перекрестная проверка моделей PWM на данных ChIP-seq, на которых модели не обучались

Чтобы понять, насколько специфика одного ChIP-seq набора, в котором обучалась модель, может повлиять на точность распознавания ССТФ этой же моделью в других ChIP-seq данных, мы провели перекрестную проверку. Оценили точность каждой модели PWM не только внутри того же набора данных, где обучалась модель (для этого случая проводили несколько итераций разделения всей выборки обучения, так что модель обучалась на 90 % пиков, а тестировалась на оставшихся 10 % пиков), но и на остальных 15 наборах данных (контрольных). Для каждого случая рассчитали оценку точности pAUC (см. выше раздел «Построение *de novo* моделей и оценка их точности распознавания ССТФ»), результаты представили в виде тепловой карты (рис. 8). Из тепловой карты видно, что только в трех случаях – ENCSR000BRE.A-549, ENCSR000BNI.Hep-G2 и ERP008682.pancreas – другие

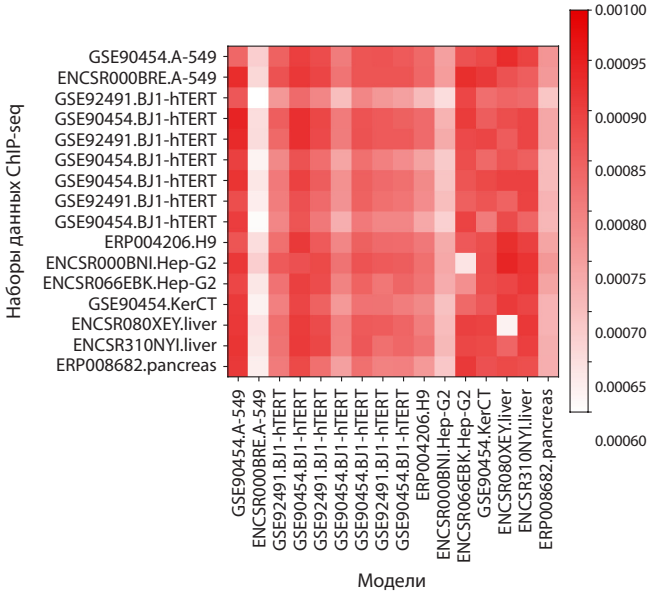


Рис. 8. Тепловая карта сравнения pAUC.

Цвета соответствуют значениям pAUC. Для ячеек, расположенных по диагонали, контрольные и обучающие наборы данных совпадают. В остальных ячейках они различаются. Строки означают модели, столбцы – наборы данных ChIP-seq.

модели имеют очень низкую оценку pAUC, а для случаев GSE90454.A-549, ENCSR066EBK.Hep-G2, GSE90454.KerCT, ENCSR080XEY.liver и ENCSR310NYI.liver все модели имеют высокое значение pAUC.

Обсуждение

На основе полученных данных можно заключить, что совместное использование альтернативных моделей с PWM позволяет расширить количество выявляемых пиков, содержащих ССТФ, относительно PWM.

Такой результат можно объяснить наличием разных структурных типов ССТФ для FOXA2, т. е. их гетерогенностью. Это хорошо согласуется с экспериментальными данными, полученными для ряда других ТФ, включая представителей семейства FOX. Так, было показано, что ТФ HOXB13 и FOXC2 способны связываться с одинаковой аффинностью с совершенно отличными последовательностями CAATAAA/TCGTAAA (Morgunova et al., 2018) и GTAAACA/ACAAATA (Chen et al., 2019) соответственно. Недавно обнаружено, что ТФ FOXN3 может связываться с двумя принципиально различными типами ССТФ, которые имеют разную длину (Rogers et al., 2019). Помимо этого, небольшие изменения в структуре ССТФ зависят от кооперативного взаимодействия между ТФ (Morgunova, Taipale, 2017). Очевидно, что FOXA2 также связывается с разными структурными типами СС.

Чтобы учесть все варианты ССТФ, одной модели PWM для распознавания сайта может быть уже недостаточно. Эта проблема частично решается использованием нескольких PWM (Bi et al., 2011; Mitra et al., 2018) или альтернативных моделей (Mathelier, Wasserman, 2013; Yang et al., 2014; Siebert, Söding, 2016; Eggeling et al., 2017; Gheorghe et al., 2019). Однако ранее альтернативные модели обычно сравнивали с PWM только по точности поиска ССТФ (Siebert, Söding, 2016) либо по количеству распознанных сайтов (Samee et al., 2019). В настоящей работе мы не только сравнили точность и количество распознаваемых пиков, но и оценили, сколько каждая модель привносит своих пиков с ССТФ, совместный вклад моделей в поиск ССТФ, а также как соотносятся между собой пики с ССТФ от разных моделей. Результаты по оценке точности (см. рис. 7, *а*) показали, что на данных по FOXA2 модель InMoDe имеет самую низкую точность относительно других моделей, а модели BaMM, diPWM и PWM сопоставимы между собой. С точки зрения расширения общей доли пиков с ССТФ в рассмотренном наборе данных лучше всего себя показала модель BaMM, поскольку она находит самую крупную долю пиков с ССТФ, которые не выявляются другими моделями. Тем не менее все альтернативные модели (diPWM, BaMM и InMoDe) позволяют расширить набор распознанных ССТФ относительно PWM, а PWM вносит свой независимый вклад в общее количество пиков с распознанными ССТФ.

Заключение

Нами разработан конвейер программ MultiDeNA, который позволяет единообразно обрабатывать данные ChIP-seq с использованием разных моделей поиска ССТФ. В настоящее время с его помощью можно строить модели PWM, diPWM, InMoDe, BaMM. MultiDeNA включает в себя этапы подготовки данных, построения моделей, оценки точности моделей, сканирования пиков, сочетания результатов и их анализа. Разработанным конвейером программ был обработан набор данных из базы ReMap, включающий 22 ChIP-seq эксперимента для ТФ FOXA2. Мы показали, что совместное применение разных моделей позволяет увеличить общее количество распознанных пиков до 73.6 %, относительно модели PWM количество распознанных пиков увеличилось на 26.3 %. Разные модели распознают совпадающие ССТФ в значительной доле пиков, тем самым выявляя наиболее общий структурный тип ССТФ в этих пиках. Также каждая модель находила ССТФ, которые не выявлялись другими моделями. Лучше всего себя показала модель BaMM с 4.15 % пиков, содержащих только ее сайты, против 1.08, 0.49, 1.73 % для PWM, diPWM и InMoDe соответственно. Исходя из результатов можно предположить, что гетерогенность сайтов для FOXA2 не учитывается полностью только одной из моделей. Хуже всего себя в этом плане проявила модель diPWM, которая распознает ССТФ только в 46.4 % пиков. Оптимальной моделью для сайтов FOXA2 оказалась модель BaMM, которая нашла ССТФ в 65.8 % пиков. На основании полученных данных мы предположили, что модель BaMM может лучше описывать ССТФ для FOXA2.

Список литературы / References

Bailey T.L., Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proc. Int. Conf. Intell. Syst. Mol. Biol. 1994;2:28-36. DOI citeulike-article-id:878292. PMID 7584402.

Benos P.V., Bulyk M.L., Stormo G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* 2002;30(20):4442-4451. DOI 10.1093/nar/gkf578.

Bi Y., Kim H., Gupta R., Davuluri R.V. Tree-based position weight matrix approach to model transcription factor binding site profiles. *PLoS One.* 2011;6(9):e24210. DOI 10.1371/journal.pone.0024210.

Bulyk M.L., Johnson P.L.F., Church G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.* 2002;30(5):1255-1261. DOI 10.1093/nar/30.5.1255.

Chen X., Wei H., Li J., Liang X., Dai S., Jiang L., Guo M., Qu L., Chen Z., Chen L., Chen Y. Structural basis for DNA recognition by FOXC2. *Nucleic Acids Res.* 2019;47(7):3752-3764. DOI 10.1093/nar/gkz077.

Chèneby J., Ménétrier Z., Mestdagh M., Rosnet T., Douida A., Rhalloussi W., Bergon A., Lopez F., Ballester B. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.* 2020;48(D1):D180-D188. DOI 10.1093/nar/gkz945.

Eggeling R., Grosse I., Grau J. InMoDe: tools for learning and visualizing intra-motif dependencies of DNA binding sites. *Bioinformatics.* 2017;33(4):580-582. DOI 10.1093/bioinformatics/btw689.

Farnham P.J. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 2009;10(9):605-616. DOI 10.1038/nrg2636.

Furey T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* 2012;13(12):840-852. DOI 10.1038/nrg3306.

Gheorghe M., Sandve G.K., Khan A., Chèneby J., Ballester B., Mathelier A. A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.* 2019;47(4):e21. DOI 10.1093/nar/gky1210.

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

Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell.* 2010;38(4):576-589. DOI 10.1016/j.molcel.2010.05.004.

Ignatieva E.V., Oshchepkov D.Y., Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina T.V., Merkulova T.I. Comparison of the results of search for the SF-1 binding sites in the promoter regions of the steroidogenic genes, using the SiteGA and SITECON methods. In: Proc. Fourth Int. Conf. Bioinform. Genome Regul. Struct. (BGRS). 2004;1:69-72.

Iwafuchi-Doi M. The mechanistic basis for chromatin regulation by pioneer transcription factors. *WIREs Syst. Biol. Med.* 2019;11(1):e1427. DOI 10.1002/wsbm.1427.

Keilwagen J., Grau J. Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.* 2015;43(18):e119. DOI 10.1093/nar/gkv577.

Kiesel A., Roth C., Ge W., Wess M., Meier M., Söding J. The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.* 2018;46(W1):W215-W220. DOI 10.1093/nar/gky431.

Kulakovskiy I.V., Boeva V.A., Favorov A.V., Makeev V.J. Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics.* 2010;26(20):2622-2623. DOI 10.1093/bioinformatics/btq488.

Kulakovskiy I., Levitsky V., Oshchepkov D., Bryzgalov L., Vorontsov I., Makeev V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.* 2013;11(01):1340004. DOI 10.1142/S0219720013400040.

Kulakovskiy I.V., Makeev V.J. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. *Biophysics* (*Oxf.*). 2009;54(6):667-674. DOI 10.1134/S0006350909060013.

Kulakovskiy I.V., Vorontsov I.E., Yevshin I.S., Sharipov R.N., Fedorova A.D., Rumynskiy E.I., Medvedeva Y.A., Magana-Mora A., Bajic V.B., Papatsenko D.A., Kolpakov F.A., Makeev V.J. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. 2018;46(D1):D252-D259. DOI 10.1093/nar/gkx1106.

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.

Latchman D.S. Transcription factors: bound to activate or repress. *Trends Biochem. Sci.* 2001;26(4):211-213. DOI 10.1016/S0968-0004(01)01812-6.

Levitsky V.G., Ignatieva E.V., Ananko E.A., Turnaev I.I., Merkulova T.I., Kolchanov N.A., Hodgman T.C.T. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC Bioinform.* 2007;8(1):1-20. DOI 10.1186/1471-2105-8-481.

Levitsky V.G., Kulakovskiy I.V., Ershov N.I., Oshchepkov D.Y., Makeev V.J., Hodgman T.C., Merkulova T.I. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data. *BMC Genom.* 2014;15(1):80. DOI 10.1186/1471-2164-15-80.

Levitsky V.G., Oshchepkov D.Y., Klimova N.V., Ignatieva E.V., Vasiliev G.V., Merkulov V.M., Merkulova T.I. Hidden heterogeneity of transcription factor binding sites: a case study of SF-1. *Comput. Biol. Chem.* 2016;64:19-32. DOI 10.1016/j.compbiolchem.2016.04.008.

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.

Machanick P., Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics.* 2011;27(12):1696-1697. DOI 10.1093/bioinformatics/btr189.

Mathelier A., Wasserman W.W. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.* 2013;9(9):e1003214. DOI 10.1371/journal.pcbi.1003214.

McClish D.K. Analyzing a portion of the ROC curve. *Med. Decis. Mak.* 1989;9(3):190-195. DOI 10.1177/0272989X8900900307.

Mitra S., Biswas A., Narlikar L. DIVERSITY in binding, regulation, and evolution revealed from high-throughput ChIP. *PLoS Comput. Biol.* 2018;14(4):1-20. 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.

Morgunova E., Yin Y., Das P.K., Jolma A., Zhu F., Popov A., Xu Y., Nilsson L., Taipale J. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife.* 2018;7:1-21. DOI 10.7554/eLife.32963.

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

Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26(6):841-842. DOI 10.1093/bioinformatics/btq033.

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.DOI 10.1016/j.molcel. 2019.01.019.

Samee M.A.H., Bruneau B.G., Pollard K.S. A *de novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Syst.* 2019;8(1):27-42. DOI 10.1016/j.cels.2018.12.001.

Siebert M., Söding J. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.* 2016;44(13):6055-6069. DOI 10.1093/nar/gkw521.

Srivastava D., Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. *Biochim. Biophys. Acta – Gene Regul. Mech.* 2020;1863(6):e194443. DOI 10.1016/j.bbagrm.2019.194443.

Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics.* 2000;16(1):16-23. DOI 10.1093/bioinformatics/16.1.16.

Wallerman O., Motallebipour M., Enroth S., Patra K., Bysani M.S.R., Komorowski J., Wadelius C. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res.* 2009;37(22):7498-7508. DOI 10.1093/nar/gkp823.

Wederell E.D., Bilenky M., Cullum R., Thiessen N., Dagpinar M., Delaney A., Varhol R., Zhao Y., Zeng T., Bernier B., Ingham M., Hirst M., Robertson G., Marra M.A., Jones S., Hoodless P.A. Global analysis of *in vivo* Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.* 2008;36(14):4549-4564. DOI 10.1093/nar/gkn382.

Worsley Hunt R., Wasserman W.W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* 2014;15(7):412. DOI 10.1186/s13059-014-0412-4.

Yang L., Zhou T., Dror I., Mathelier A., Wasserman W.W., Gordân R., Rohs R. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* 2014;42(D1):D148-D155. DOI 10.1093/nar/gkt1087.

Zhang M.O., Marr T.G. A weight array method for splicing signal analysis. *Bioinformatics.* 1993;9(5):499-509. DOI 10.1093/bioinformatics/9.5.499.

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

ORCID ID
A.V. Tsukanov orcid.org/0000-0002-5174-6609
V.G. Levitsky orcid.org/0000-0002-4905-3088

Благодарности. Работа поддержана Российским фондом фундаментальных исследований (№ 18-29-13040) и бюджетным проектом № 0259-2019-0008.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 10.10.2020. После доработки 10.01.2021. Принята к публикации 12.01.2021.