На правах рукописи

ЦУКАНОВ АНТОН ВИТАЛЬЕВИЧ

Мультимодельный подход к эффективному картированию сайтов связывания транскрипционных факторов по данным ChIP-seq экспериментов

1.5.8. — математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата биологических наук

Новосибирск - 2023

Работа выполнена в лаборатории эволюционной биоинформатики и теоретической генетики Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск.

Научный руководитель:	**Левицкий Виктор Георгиевич** кандидат биологических наук, старший научный сотрудник лаборатории эволюционной биоинформатики и теоретической генетики ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН», г. Новосибирск
Официальные оппоненты:	**Самсонова Мария Георгиевна** доктор биологических наук, заведующая лабораторией «Цифровые технологии для агробиологии» ФГАОУ ВО Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург **Омелина Евгения Сергеевна** кандидат биологических наук, заведующая лабораторией клеточного деления ФГБУН «Институт молекулярной и клеточной биологии СО РАН», г. Новосибирск
Ведущая организация:	АНО ВО «Университет Сириус», г. Сочи

Защита диссертации состоится « » 2024 г. на утреннем заседании диссертационного совета 24.1.239.01 на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» в конференц-зале Института по адресу: пр. ак. Лаврентьева 10. г. Новосибирск, 630090. +7 (383) 3634906. факс +7 (383) 3331278. e-mail: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института http://www.icgbio.ru

Автореферат разослан «____» _____2024 г.

Ученый секретарь
диссертационного совета
доктор биологических наук Т.М. Хлебодарова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность. Экспрессия генов занимает центральное место в функционировании всех живых систем и имеет сложную систему регуляции, начиная от процесса транскрипции и заканчивая деградацией белка. Одним из ключевых компонентов регуляции экспрессии генов на этапе транскрипции являются транскрипционные факторы (ТФ). ТФ – это белки, которые способны распознавать специфические нуклеотидные последовательности в геномной ДНК, сайты связывания (СС), и связываться с ними (Lambert et al. 2018). Связывание ТФ с ДНК инициирует цепь молекулярных событий, обеспечивающих сборку/регуляцию активности преинициаторного комплекса РНК-полимеразы II за счёт непосредственных или опосредованных контактов с компонентами этого комплекса. Благодаря своей функции ТФ, являются главными компонентами в регуляции транскрипции, а поиск сайтов связывания ТФ (ССТФ), является важной задачей на пути к пониманию процессов регуляции транскрипции (Iwafuchi-Doi 2019; Latchman 2001; Srivastava and Mahony 2020).

Существует множество *in vivo* и *in vitro* экспериментальных методов, таких как ChIP-seq, ChIP-exo, DAP-seq, которые позволяют определять геномные локусы, где ТФ связан с ДНК (Farnham 2009; Furey 2012; Park 2009). Полученные из экспериментов данные секвенирования ДНК после первичной обработки дают только приблизительную информацию о том, где мог находиться ССТФ в виде пиков (локусов генома с картированными прочтениями ДНК) – последовательностей нуклеотидов длиной от 100 п.о. Для разных СС одного ТФ обычно наблюдается некоторая степень вариации, число высококонсервативных позиций в СС одного ТФ может быть очень мало, так что, как правило, даже СС со средней аффинностью могут обладать лишь умеренным сходством между собой. Поэтому, для описания специфичности ССТФ вводится понятие мотива, как общего паттерна нуклеотидного контекста, характерного для предпочтительного формирования комплекса ТФ с ДНК. Длина мотива обычно составляет от 8 до 20 п.о. (Kulakovskiy et al. 2018; O'Malley et al. 2016).Для того, чтобы найти точную форму мотива в наборе пиков используются алгоритмы *de novo* поиска мотивов (Lihu and Holban 2015). Такие алгоритмы могут быть созданы на основе разных математических моделей мотива, но все они предполагают определение и постепенное уточнение мотива на основе его предполагаемого обогащения в пиках по сравнению с некоторой ожидаемой частотой встреч по случайным причинам. Подавляющее большинство широкоиспользуемых реализаций *de novo* поиска мотивов основано на использовании традиционной модели мотива, позиционной весовой матрицы (position weight matrix, PWM) (Berg and von Hippel 1987; Stormo 2000) и наиболее популярные из них это HOMER (Heinz et al. 2010), Streme (Bailey 2021), MEME-ChIP (Machanick and Bailey 2011) и ChIPMunk (Kulakovskiy et al. 2010). Без преувеличения можно сказать, что применение разных реализаций модели PWM входит практически в каждый конвейер обработки полногеномных данных ChIP-seq (Lloyd and Bao 2019). Модель PWM широко применяется для изучения регуляции транскрипции *in silico*. Она используется для поиска ССТФ (Kulakovskiy et al. 2011) в

предполагаемых регуляторных последовательностях, для предсказания *цис*-регуляторных элементов (Nikulova et al. 2012) и для определения возможной регуляторной роли однонуклеотидных полиморфизмов (Boytsov et al. 2022; Macintyre et al. 2010).

Однако, многократно экспериментально показано (Bulyk, Johnson, and Church 2002; Cooper et al. 2023), что модель PWM имеет ограничение, поскольку она предполагает независимость вкладов отдельных позиций в общую оценку аффинности СС по отношению к ТФ. Таким образом, модель PWM не учитывает зависимости между разными позициями сайтов (Benos 2002; Keilwagen and Grau 2015). Помимо этого, существуют и другие особенности связывания ТФ с ДНК, такие как разнообразие структурных типов ССТФ (Jolma et al. 2015; Rogers et al. 2019), нуклеотидный состав флангов ССТФ, возможности разных ТФ действовать в составе гомо- и гетродимеров (Amoutzias et al. 2008), особенности взаимодействия разных ТФ с нуклеосомной ДНК (Zaret and Mango 2016), конформационная структура ДНК ССТФ, всё это не может быть полностью описано в рамках простой модели PWM. Такие ограничения могут снижать способность PWM находить все ССТФ в данных ChIP-seq. В среднем, PWM способна предсказать ССТФ примерно только в половине пиков (Gheorghe et al. 2019; Karimzadeh and Hoffman 2022; Levitsky et al. 2019; Tsukanov, Mironova, and Levitsky 2022; Worsley-Hunt and Wasserman 2014). Такие результаты лишь отчасти объясняются тем, что не всегда ТФ может связываться с ДНК напрямую. Также возможно, что ТФ взаимодействует с ДНК не напрямую, то есть связь с ДНК осуществляется через партнёрский ТФ (ТФ-посредник), и за счёт белок-белковых взаимодействий целевого ТФ и ТФ-посредника появляется некоторая доля пиков с полным отсутствием потенциальных СС целевого ТФ. Помимо этого, отсутствие мотивов целевого ТФ может быть связано с тем, что ССТФ обладают низкой аффинностью, или тем, что такие пики являются ошибками эксперимента (Jain et al. 2015; Teytelman et al. 2013; Worsley-Hunt and Wasserman 2014).

К настоящему времени для *de novo* поиска мотивов разработан и реализован ряд моделей мотивов ССТФ, альтернативных по отношению к традиционной модели PWM, они учитывают разные особенности связывания ТФ с ДНК (Eggeling, Grosse, and Grau 2017; Gheorghe et al. 2019; Keilwagen and Grau 2015; Mathelier and Wasserman 2013; Samee, Bruneau, and Pollard 2019; Siebert and Söding 2016; Yang et al. 2014). Авторы подобных моделей, таких как BaMM (Siebert and Söding 2016), InMoDe (Eggeling et al. 2015) и Slim (Keilwagen and Grau 2015) в своих работах уделяют основное внимание тому, что альтернативные модели могут показывать лучшую точность распознавания ССТФ в сравнении с точностью традиционной модели PWM. Однако, авторы редко уделяют много внимания тому, что их модели могут находить структурные типы ССТФ, отличные от таковых для традиционной модели PWM. Помимо этого, применение только одной модели, не решает проблему наиболее полного распознавания ССТФ в данных ChIP-seq. К сожалению, альтернативные модели для поиска ССТФ не получили широкого применения, несмотря на то что уже более 20 лет

известно о наличии зависимостей частот встреч нуклеотидов в разных позициях ССТФ (Bulyk 2002).

Ранее было показано, что совместное применение моделей SiteGA и PWM, позволяет находить принципиально разные структурные типы ССТФ (Levitsky et al. 2014, 2016), более того, сайты таких разных структурных типов регулировали гены с различными функциями (Levitsky et al. 2016). До сих пор не было массовых и систематических исследований на эту тему. Помимо этого, к настоящему моменту не существует программного комплекса, который позволял бы осуществлять единообразный поиск ССТФ с помощью методологически разных моделей, сопоставлять и объединять результаты поиска мотивов таких разных моделей.

В настоящей работе для массового анализа данных ChIP-seq применялись три модели мотивов PWM, BaMM и SiteGA. Модель BaMM опирается на PWM и расширяет её методологию за счёт того, что добавляет к общей оценке аффинности сайта, равной, согласно модели PWM, сумме вкладов отдельных позиций, вклады от зависимостей близких позиций мотива (Siebert and Söding 2016). Модель SiteGA методологически не связана с моделью PWM и основана на методе дискриминантного анализа, который позволяет выявлять зависимости любых позиций мотива, а точную форму мотива позволяет найти генетический алгоритм, стремящийся найти оптимальный набор локально-позиционированных динуклеотидов с учётом их зависимостей (Levitsky et al. 2007; Tsukanov et al. 2022).

Целью исследования является проведение массового анализа данных ChIP-seq с помощью совместного применения традиционной и альтернативных моделей мотива с целью выявления различных типов нуклеотидного контекста, ответственного за прямые взаимодействия транскрипционных факторов с ДНК

Для того чтобы достичь эту цель, были поставлены следующие **задачи:**

1. Создать программный комплекс для проведения *de novo* поиска мотивов разными моделями, включающий оценку точности моделей, распознавание сайтов в пиках ChIP-seq моделями и объединение результатов их предсказаний.
2. С помощью программного комплекса провести массовый анализ данных ChIP-seq для сотен ТФ для *M. musculus* и *A. thaliana* и оценить, как соотносится точность традиционной и альтернативных моделей в зависимости от типа ДНК-связывающего домена целевого транскрипционного фактора
3. Оценить вклад альтернативных моделей мотива в распознавание сайтов связывания транскрипционных факторов по доле ChIP-seq пиков в зависимости от типа ДНК-связывающего домена целевого транскрипционного фактора.
4. Проверить гипотезу о различных функциях генов, регуляторные районы которых содержат сайты, предсказанные разными моделями мотива.

Научная новизна

Впервые разработан программный комплекс MultiDeNa, который позволяет сочетать методологически разные модели *de novo* поиска мотивов, а именно

традиционную модель PWM, не учитывающую зависимости позиций мотива, и также альтернативные модели, предлагающие разные методологии для выявления зависимостей нуклеотидного контекста мотива. Программный комплекс для каждой модели позволяет выбирать оптимальные параметры для достижения максимальной точности распознавания (например, длину мотива), единообразно оценивать точность распознавания разных моделей, выбирать пороги функций распознавания, осуществлять классификацию ChIP-seq пиков, сравнивая результаты сканирования всех моделей, и выявлять пики, содержащие мотивы только некоторого поднабора моделей, например, всех моделей или только одной модели.

Впервые проведён массовый анализ данных ChIP-seq с помощью мультимодельного подхода для распознавания ССТФ, который позволил показать присутствие значительно большего природного разнообразия ССТФ, связанных с прямыми взаимодействиями ТФ с ДНК, чем это предсказывала модель PWM.

Впервые установлено, что независимые вклады каждой модели в общее распознавание ССТФ существенно зависят от структуры ДНК-связывающего домена ТФ, что подтверждает важность учёта структурного разнообразия ССТФ. Показано, что, используя результаты сочетания разных моделей, можно привязывать сайты, предсказанные разными моделями, к специфическим функциям генов.

Теоретическая и практическая значимость

Разработанный программный комплекс MultiDeNa, позволяет выявлять наиболее полный список СС, с которыми напрямую взаимодействует ТФ, за счёт применения нескольких методологически различных моделей мотивов (PWM, BaMM, SiteGA). Сочетание разных моделей мотива позволяет эффективно выявлять структурное разнообразие СС в зависимости от типа ДНК-связывающего домена ТФ. Программный комплекс MultiDeNa можно использовать в других исследованиях по анализу ChIP-seq экспериментов, с его помощью можно расширить список генов мишеней ТФ, и тем самым прояснить механизмы регуляции транскрипции генов с помощью ТФ.

Положения, выносимые на защиту

1. Разработан программный комплекс MultiDeNa для наиболее полного предсказания в геномах эукариот сайтов связывания транскрипционных факторов (ТФ) на основе данных их массового секвенирования ChIP-seq. Программный комплекс использует методологически разные модели распознавания сайтов – допускающие зависимость между частотами нуклеотидов в разных позициях сайтов (BaMM/SiteGA) и не допускающие её (PWM).
2. Эффективность моделей BaMM/SiteGA в распознавании сайтов связывания ТФ зависит от структуры ДНК-связывающего домена. Наибольший дополнительный вклад эти модели вносят в распознавание сайтов ТФ, содержащих домен типа *Basic helix-loop-helix*, наименьший – *C2H2 zinc finger*.

Вклад автора

Основная часть работы выполнена автором самостоятельно. Автор принимал участие в разработке конвейера программ, проведении вычислительных экспериментов, анализе данных, обсуждении полученных результатов.

Апробация работы

Материалы работы вошли в отчёты по гранту Российского Научного Фонда (№ 21-14-00240. руководитель Левицкий В.Г.)

Результаты диссертации были доложены на научных конференциях: 20-я международная конференция Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020), 6–10 July 2020. Novosibirsk, Russia; VII молодёжная школа-конференция по молекулярной и клеточной биологии Института цитологии РАН, 12–15 октября 2020. Санкт-Петербург, Россия; Системная биология и биоинформатика (SBB-2023), 14-я международная школа молодых ученых, 22–26 мая 2023 г., Новосибирск, Россия.

Структура и объем работы

Работа состоит из введения, обзора литературы, описания материалов и методов, результатов и их обсуждения, заключения, выводов, списка литературы и приложения. Работа изложена на 204 страницах (в том числе 42 страниц в приложении), содержит 46 рисунков и 16 таблиц, включая 3 таблицы из приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Обзор литературы

В первой главе даются представления о том, какую роль играют ТФ в регуляции транскрипции генов. Детально рассматриваются основные домены, составляющие структуру ТФ. Большое внимание уделяется ключевому домену ТФ – ДНК-связывающему домену (ДСД), описывается иерархическая классификация ТФ, основанная на структуре ДСД, включая уровни суперклассов и классов ТФ (Wingender, 2013; Castro-Mondragon et al., 2022). Даётся представление об основных механизмах взаимодействия ТФ с ДНК, представляются традиционные (консенсус и PWM), и альтернативные модели мотива (BaMM, InMoDe, SiteGA и др.). В конце главы даются общие представления об эксперименте ChIP-seq, этапах его биоинформатической обработки, а также обосновывается необходимость сочетания методологически разных моделей поиска мотивов.

Глава 2. Методы

В массовом анализе данных ChIP-seq использовали три модели мотива: PWM, не учитывающую зависимости позиций в мотиве, BaMM / SiteGA, учитывающие зависимости близких / любых позиций в мотиве, соответственно. Из базы данных GTRD (Kolmykov et al. 2021) выделены две коллекции, включающие по 121 / 1556 наборов данных ChIP-seq экспериментов для *A. thaliana* / *M. musculus,* соответственно. В анализ взяли только ChIP-seq эксперименты, для которых (1) известный мотив

целевого ТФ, либо родственного ТФ обогащён, (2) все три модели выявили *de novo* мотив значимо похожий на мотив целевого ТФ. Фильтрацию прошли 68 / 1004 ChIP-seq эксперимента для *A. thaliana* / *M. musculus*, соответственно. Анализ проводили с учётом структуры ДСД ТФ, поэтому данные были разбиты по классам ТФ.

Для оценки точности моделей по каждому набору пиков ChIP-seq строили ROC-кривую с помощью стандартного подхода перекрёстной проверки (ПП). Для проведения ПП в качестве позитивной выборки использовали 1000 лучших по качеству ChIP-seq пиков, а негативная выборка была сгенерирована на основе полного генома. После построения ROC-кривой точность моделей оценивали частичной площадью под кривой (pAUC), ограниченной порогом ошибки перепредсказания 0.001 (FPR = 0.001. False Positive Rate) (рис. 1). С помощью ПП и оценки точности pAUC выбирали оптимальные параметры моделей и сравнивали точности моделей. Далее проводили *de novo* поиск мотивов для каждой модели по установленным параметрам, где в качестве позитивной выборки использовали 1000 лучших по качеству ChIP-seq пиков.

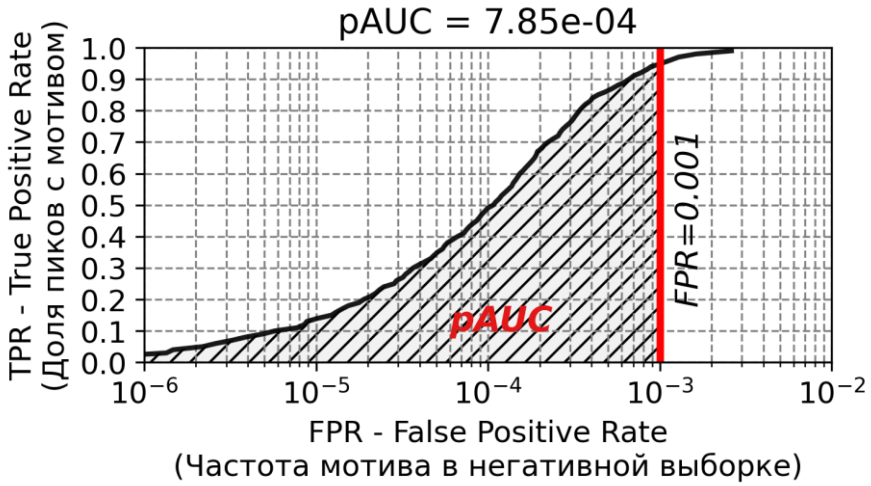

Рис. 1. Пример ROC-кривой, где на оси X - частота мотива в негативной выборке, а на оси Y - доля пиков с мотивом. Частичная площадь под кривой (pAUC) обозначена штриховкой.

Для моделей единообразно установили пороги по частоте встречаемости мотива в промоторах генов всего генома, кодирующих белки (значение ERR, expected recognition rate, ожидаемая частота распознавания). Величина ERR напрямую связана с FPR, поэтому сравнение ROC-кривых проводили на трёх диапазонах: диапазоне жёстких порогов (FPR $\leq 10^{-4}$), диапазоне средних порогов ($10^{-4} < FPR \leq 2.5*10^{-4}$) и диапазоне мягких порогов ($2.5*10^{-4} < FPR \leq 5*10^{-4}$). После выбора порогов, распознавали сайты в пиках ChIP-seq моделями с заданным порогом, далее результаты объединяли и классифицировали пики на основании присутствия сайтов разных моделей.

Для анализа терминов ГО применили пакеты ChIPseeker (Yu, Wang, and He 2015) и clusterProfiler (Yu et al. 2012). В анализ терминов ГО брали только кодирующие белок гены, промоторы которых содержали пики ChIP-seq с предсказанными ССТФ.

Глава 3. Результаты и обсуждения

На рисунке 2 приведены ROC кривые, рассчитанные по данным ChIP-seq для ТФ двух классов, *Basic helix-loop-helix factors (bHLH)* {1.2} и *C2H2 zinc finger factors* {2.3}, которые являются одними из наиболее представительных для *M. musculus* и *A. thaliana*.

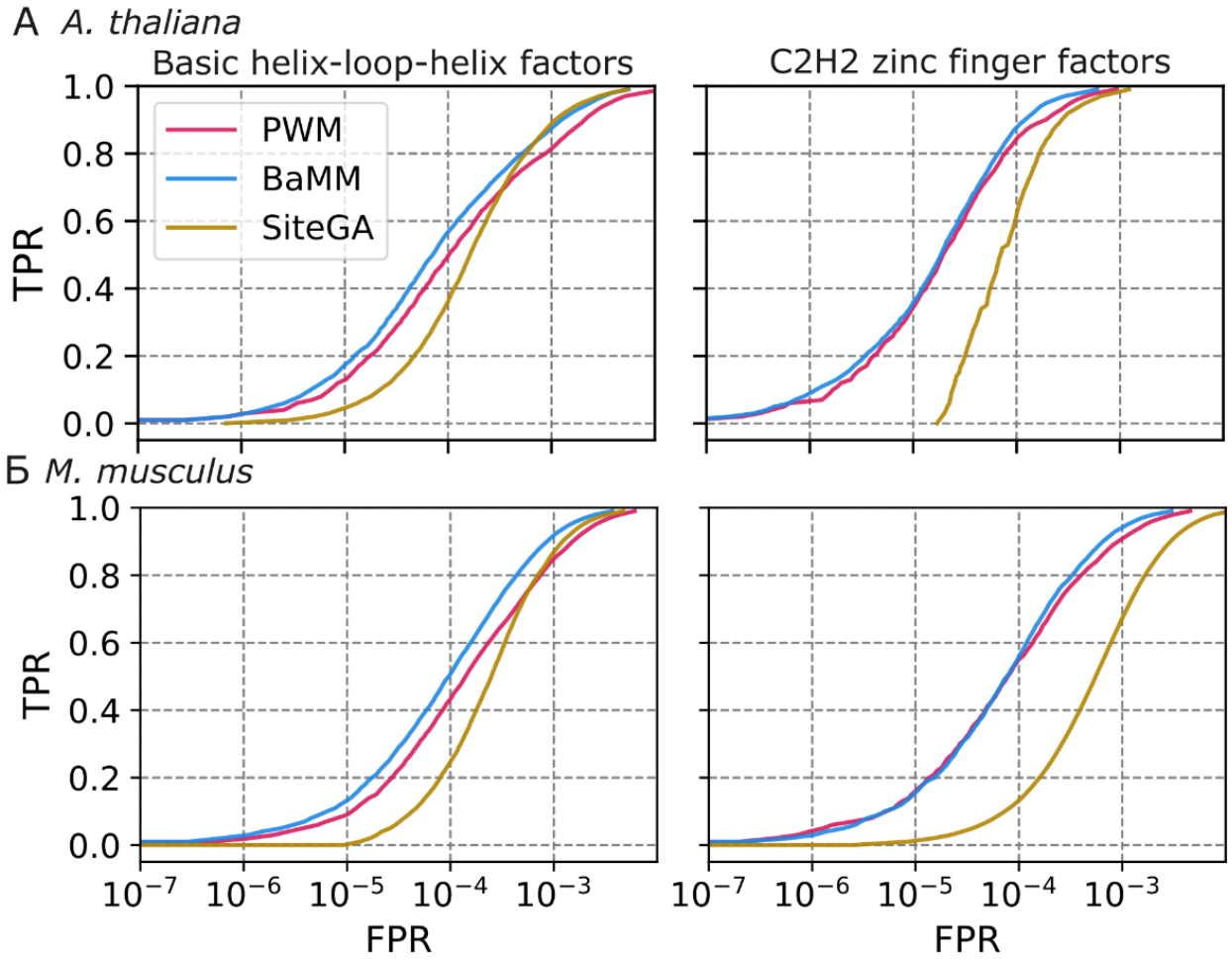

Рис. 2. Характеристика точности распознавания мотивов моделей PWM, BaMM и SiteGA с помощью ROC-кривой для наборов ChIP-seq данных по ТФ классов *Basic helix-loop-helix factors (bHLH)*{1.2} и *C2H2 zinc finger factors* {2.3}. **(А)** *M. musculus*, **(Б)** *A. thaliana.* ROC-кривые получены с применением процедуры ПП. На графиках показаны средние значения FPR по классу (оси X) в зависимости от пороговых значений доли пиков с мотивом (TPR, ось Y).

Полученные ROC-кривые для *M. musculus* и *A. thaliana* имеют схожие закономерности для одинаковых классов ТФ (рис. 2). Для класса *Basic helix-loop-helix factors (bHLH)* {1.2} модель BaMM уже на достаточно жёстких порогах (FPR $\leq 10^{-4}$) заметно превосходит по точности PWM, а для класса *C2H2 zinc finger factors* {2.3} модели PWM и BaMM почти не отличаются при FPR $\leq 10^{-4}$, а при FPR $> 10^{-4}$. отличие очень мало. Модель SiteGA только на мягких порогах и только для класса *Basic helix-loop-helix factors (bHLH)* {1.2} опережает по точности модель PWM.

Далее были построены распределения оценок точности моделей pAUC для наиболее представительных классов ТФ (рис. 3).

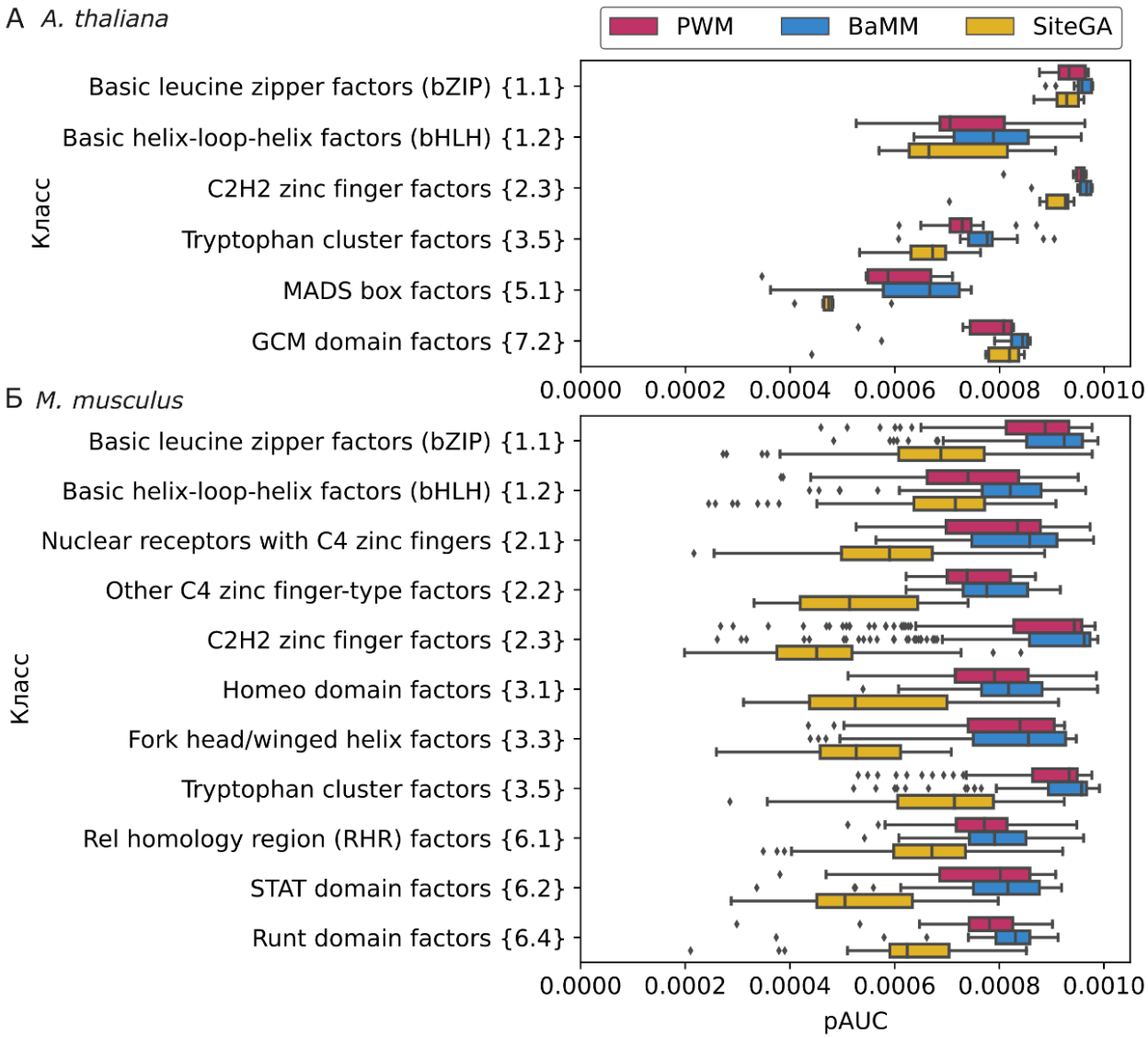

Рис. 3. Диаграмма размаха с распределениями оценки точности pAUC для трёх моделей мотива (PWM, BaMM и SiteGA). Оценки pAUC рассчитаны для разных классов ТФ. На диаграмме представлены распределения квартилей Q_1, Q_2 и Q_3 для pAUC: **(А)** *A. thaliana*, **(Б)** *M. musculus*. Планки погрешностей левее (Q_1) и правее (Q_3) относятся к минимальным/максимальным значениям, если они расположены в пределах 1.5 межквартильных размахов (interquartile range, IQR = Q_3−Q_1) от Q_1 / Q_3, в противном случае они равны {Q_1−1.5*IQR}/{Q_3+ 1.5*IQR}, соответственно. Все значения, которые не попали в пределы планок погрешности отмечены как выбросы.

Распределения оценок точности pAUC, полученные для *M. musculus* и *A. thaliana* имеют общие характерные черты. Как видно из диаграмм (рис. 3), точность моделей согласованно варьирует у ТФ двух видов организмов в зависимости от структуры ДСД. В частности, модели PWM и BaMM достигают высокой точности для классов *Basic leucine zipper factors (bZIP)* {1.1} и *C2H2 zinc finger factors* {2.3}; среди всех классов значения точности модели BaMM выше, чем у других моделей, а у модели SiteGA – ниже, чем у других; PWM везде немного уступает BaMM. Однако, точность модели SiteGA варьирует относительно точности PWM в зависимости от класса ДСД. Максимальное и минимальное отношения точностей моделей SiteGA и PWM

наблюдаются для классов *Basic helix-loop-helix factors (bHLH)* {1.2} и *C2H2 zinc finger factors* {2.3}, соответственно.

По приведённым ROC кривым и распределениям значений оценки точности pAUC (рис. 3. 4) можно заключить, что эффективность моделей зависит от выбора порога: (1) на жёстких и средних порогах эффективны традиционная PWM или альтернативная BaMM, которая является расширением PWM, но (2) при переходе к мягким порогам эффективность альтернативных моделей BaMM и SiteGA заметно повышается относительно модели PWM. Другим важным фактором, который влияет на точность, является класс ДСД, к которому относится ТФ.

Поскольку связывание ТФ с ДНК в значительной степени определяется структурой его ДСД, было предположено, что вклады разных моделей в общее распознавание ССТФ могут зависеть от класса ДСД, к которому относится ТФ. Результаты по распознаванию ССТФ на пороге ERR ≤ 10^{-4} моделями PWM, BaMM и SiteGA для *M. musculus* и *A. thaliana* для классов ТФ представлены на рисунке 4.

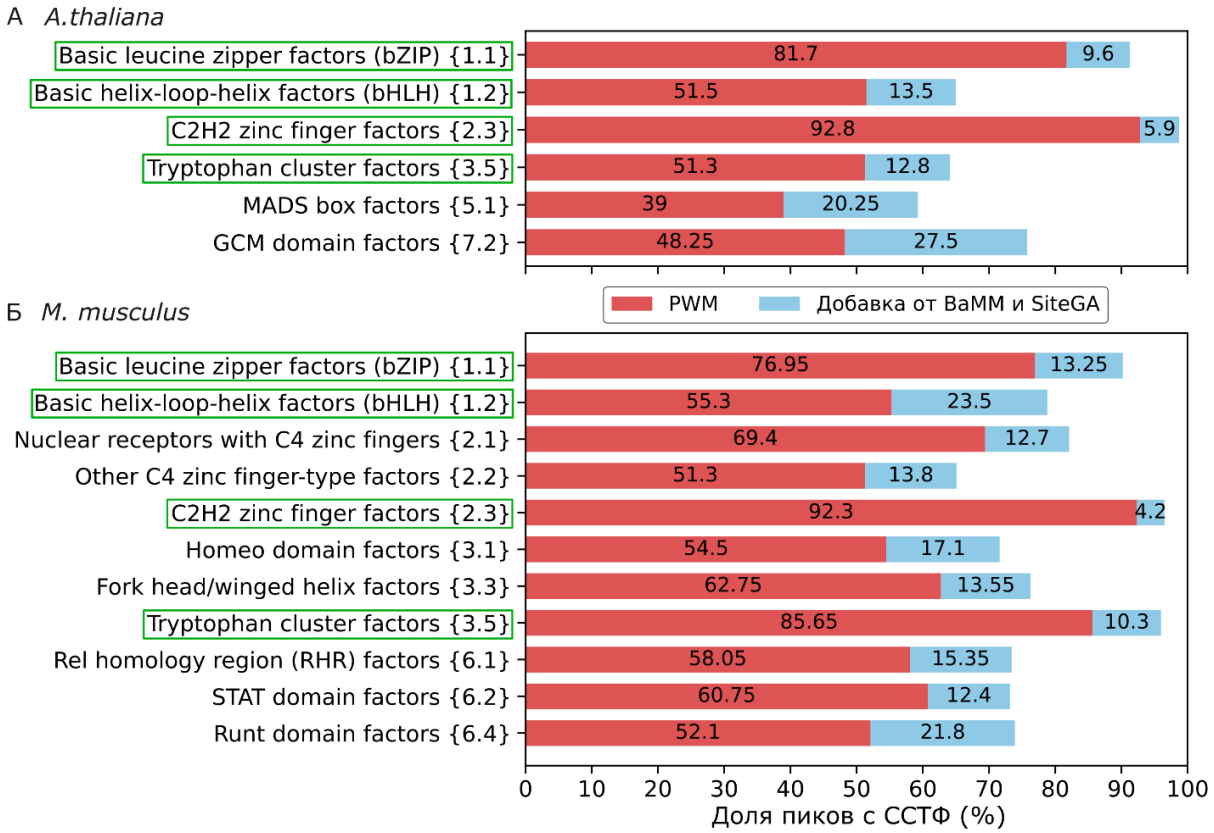

Рис. 4. Распределение фракций доли пиков, содержащих мотивы традиционной модели PWM и добавочной доли с мотивами альтернативных моделей BaMM и/или SiteGA по ТФ разных классов из данных ChIP-seq **(А)** *A. thaliana* и **(Б)** *M. musculus*. Ось X – значения медиан долей пиков с предсказанными мотивами по классу ТФ (в %); ось Y – классы ТФ. ССТФ предсказаны с применением порога ERR ≤ 10^{-4}. Зелёной рамкой обведены общие для двух видов организмов классы ТФ.

Из представленных на рисунке 4 данных можно заключить, что вклад альтернативных моделей зависит от класса ТФ, это справедливо как для *A. thaliana* (рис. 4А), так и для *M. musculus* (рис. 4Б). Прибавка предсказаний альтернативных

моделей BaMM/SiteGA к доле пиков модели PWM варьирует от 4.2% (*M. musculus, C2H2 zinc finger factors* {2.3}) до 27.5% (*A. thaliana, GCM domain factors* {7.2}). Для *A. thaliana* (рис. 4А) особенно выделяются ТФ, относящиеся к классам *MADS box factors* {5.1} и *GCM domain factors* {7.2}, где альтернативные модели увеличивают долю пиков, содержащих ССТФ, по отношению к доле пиков модели PWM, на 20.25% и 27.5%, соответственно. Для *M. musculus* (рис. 4Б) наибольший вклад в распознавание сайтов альтернативные модели продемонстрировали для класса *Basic helix-loop-helix factors (bHLH)* {1.2}, где прибавка доли пиков составила 23.5%, у *A. thaliana* альтернативные модели для этого класса также показали заметный вклад в распознавание сайтов – 13.5%.

На данных для *A. thaliana* альтернативные модели для класса *Basic leucine zipper factors (bZIP)* {1.1} внесли умеренный вклад 9.6%, однако для *M. musculus* он несколько больше – 13.25%. Возможно такое различие, в первую очередь, связано с размером выборок и разнообразием ТФ, так как для *A. thaliana* было 13 ChIP-seq экспериментов для 7 ТФ, а для *M. musculus* - 214 ChIP-seq экспериментов для 20 ТФ.

Наименьший вклад в долю пиков, как для *A. thaliana*, так и для *M. musculus*, альтернативные модели показывают для класса *C2H2 zinc finger factors* {2.3}, где вклад составил 5.9% и 4.2%, соответственно.

В целом, полученные для *M. musculus* и *A. thaliana* результаты хорошо согласуются, альтернативные модели вносят существенный вклад в распознавание сайтов в пиках для разных классов при этом данный вклад зависит от класса ТФ. С одной стороны, у обоих видов организмов альтернативные модели мало расширяют результаты PWM для класса *C2H2 zinc finger factors* {2.3}, для которого, возможно, гипотеза об независимости вкладов нуклеотидных позиций работает наилучшим образом, что может быть связано с протяжёнными ССТФ для данного класса ТФ и минимальным количеством димеров, образуемых ТФ этого класса. Стоит отметить, что для данного класса модель SiteGA имеет наихудшую точность (*M. musculus*, рис. 3А), что так же хорошо согласуется с гипотезой о том, что у данного класса ТФ зависимости в мотиве оказываются очень слабыми. С другой стороны, для класса *Basic helix-loop-helix factors (bHLH)* {1.2} для обоих видов организмов альтернативные модели сделали существенный вклад в распознавание сайтов в пиках. Для этого класса ТФ модель SiteGA имеет более высокую точность по сравнению с другими классами, что может свидетельствовать о наличии существенного вклада зависимостей в общий паттерн нуклеотидного контекста. Это хорошо объясняется тем, что ТФ из класса *Basic helix-loop-helix factors (bHLH)* {1.2} функционируют в составе широкого разнообразия гомо- и гетеродимеров, что существенно влияет на структуру мотивов, а также приводит к различным модификациям ССТФ, которые влияют на конформацию комплекса ТФ-ДНК (de Martin, Sodaei, and Santpere 2021).

С помощью моделей PWM, BaMM и SiteGA, были распознаны сайты в пиках (на среднем пороге ERR $\leq 2.5*10^{-4}$), при этом в анализ взяты все пики для каждого набора пиков ChIP-seq. Далее ССТФ распознанные каждой модели картировали на промоторы генов (±1000 п.о. для *A. thaliana* и ±3000 п.о. для *M. musculus* от сайта старта

транскрипции), в результате для каждой модели были получены списки генов, в промоторах которых есть предсказанные ССТФ. Результаты аннотации для *M. musculus* и *A. thaliana* представлены в виде диаграмм Венна (рис. 5), где изображены доли генов (значения медиан по классам ТФ), содержащих в промоторах генов сайты разных комбинаций моделей. Расчёты произведены отдельно для двух классов ТФ: *Basic helix-loop-helix factors (bHLH)* {1.2} и *C2H2 zinc finger factors* {2.3}.

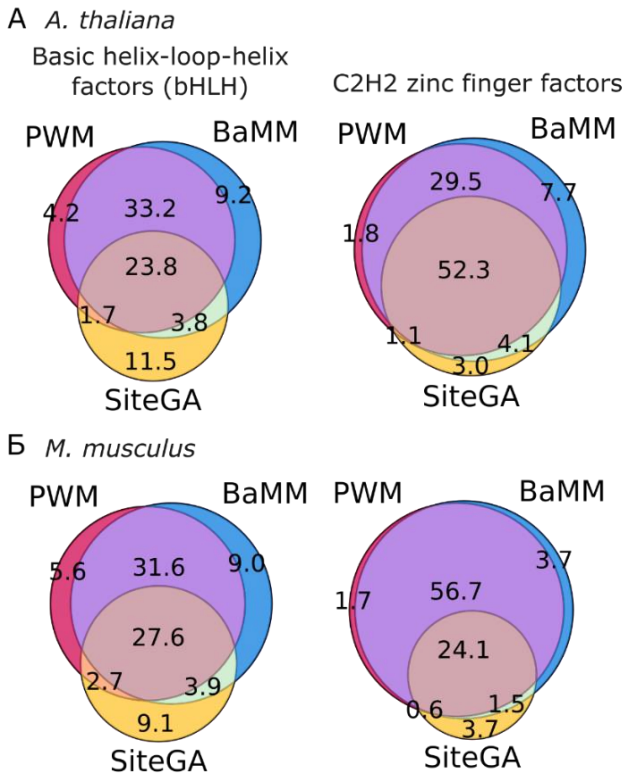

Рис. 5. Диаграммы Венна для классификации промоторов генов, содержащих разные комбинации ССТФ, предсказанных моделями мотива PWM, BaMM и SiteGA. На диаграммах показаны результаты в виде медиан долей генов, содержащих в промоторах предсказанные сайты одной, двух (в разных комбинациях) или трёх моделей для двух классов ТФ: *Basic helix-loop-helix factors (bHLH) {1.2}, C2H2 zinc finger factors {2.3}*. **(А)** *A. thaliana*; **(Б)** *M. musculus*

Полученные результаты показывают, что доли генов содержащие сайты, предсказанные моделями PWM и BaMM, больше соответствующей доли для SiteGA. Для класса *C2H2 zinc finger factors* {2.3} перекрытие долей трёх моделей гораздо больше, а доли генов с сайтами одной из моделей (далее доли «уникальных» генов) гораздо меньше, чем для класса *Basic helix-loop-helix factors (bHLH)* {1.2}. Каждая из моделей имеет от 1.7% до 7.7% «уникальных» генов для класса *C2H2 zinc finger factors* {2.3}, это справедливо как для *A. thaliana* (рис. 5А), так и для *M. musculus* (рис. 5Б). Для класса *Basic helix-loop-helix factors (bHLH)* {1.2} доля «уникальных» генов варьирует от 4.2% до 11.5%. Таким образом, вклад каждой модели меняется в зависимости от класса ТФ, и это утверждение справедливо как для *A. thaliana*, так и для *M. musculus*. Полученные результаты показывают, что заметная часть генов имеет сайты, предсказанные только одной из моделей. Следовательно, можно предположить, что ТФ способны регулировать группы генов, имеющих в промоторах сайты с такой структурой, которая предсказывается только одной моделью мотива. Такие гены могут иметь определенные биологические функции, отличные от функций другой группы генов, где сайты обнаружены другой моделью.

Чтобы проверить эту гипотезу, были получены списки обогащённых терминов ГО для биологических процессов, по результатам картирования сайтов моделями

мотива PWM, BaMM или SiteGA в пиках ChIP-seq, наложенных на промоторы генов. Рассмотрим результаты анализа на примере коллекции данных ChIP-seq *A. thaliana*. Для каждого термина ГО была рассчитана значимость обогащения, скорректированная с учётом множественных сравнений (p_adj) и кратность изменения (англ. fold change, FC), $FC = [(Mot_+GO_+)/Mot_+] / [GO_+/(GO_+ + GO_-)]$ (см. таблицу 1).

Таблица 1. Таблица сопряжённости 2×2 для анализа обогащения терминов ГО.

		Число генов		**Всего**
		С термином ГО	**Без термина ГО**	
Число генов	**С мотивом**	Mot_+GO_+	Mot_+GO_-	Mot_+
	Без мотива	Mot_-GO_+	Mot_-GO_-	Mot_-
Всего		GO_+	GO^-	

Были получены распределения количества обогащенных терминов ГО для каждой из моделей (PWM, BaMM и SiteGA), а также аналогичное распределение для обогащенных терминов ГО по данным хотя бы одной из моделей (рис. 6А) для того, чтобы оценить вклады моделей в расширение списка обогащенных терминов ГО. Для каждой модели было получено распределение количества обогащённых терминов ГО, которые выявляются только одной из моделей (рис. 6Б).

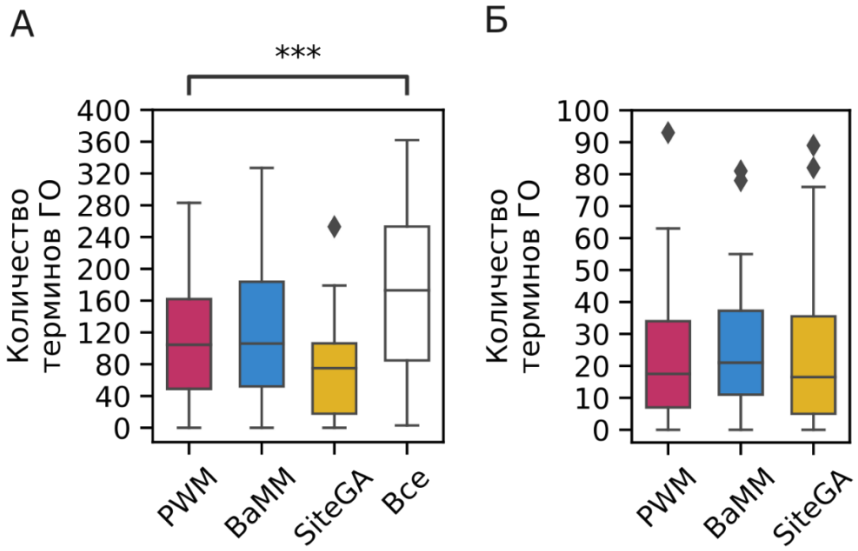

Рис. 6. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации для анализа обогащения терминов ГО на коллекции данных ChIP-seq экспериментов для *A. thaliana*. **(А)** Показаны распределения количества терминов ГО, полученных для каждой из моделей (PWM, BaMM и SiteGA), а также распределение количества терминов ГО, обогащённых хотя бы для одной из моделей (Все). **(Б)** Показаны распределения количества терминов ГО, которые имеют обогащение только для одной модели. Планки погрешностей ниже Q_1 и выше Q_3 относятся к минимальным/максимальным значениям, если они расположены в пределах 1.5 межквартильных размахов (interquartile range, $IQR = Q_3 - Q_1$) от Q_1 / Q_3. в противном случае они равны $\{Q_1 - 1.5*IQR\}/\{Q_3 + 1.5*IQR\}$, соответственно. Все значения, которые не попали в пределы планок погрешности отмечены как выбросы. *** - $p < 0.001$.

По данным, приведенным на рисунке 6. можно заключить, что модель PWM в среднем находит 104 обогащенных термина ГО, а альтернативные модели значимо ($p < 0.05$) увеличивают среднее количество обогащенных терминов ГО до 173. При этом у каждой из трёх моделей есть термины ГО, обогащённые только для неё (рис. 6Б). Следовательно, различная структура мотивов разных моделей связана с отличиями в функциональном статусе генов. Поиск специфических терминов ГО может расширить представление о биологических процессах, которые регулируют ТФ, а также выявить специфические группы генов, которые имеют в промоторах мотивы только одной из моделей.

Для терминов ГО, являющимися общими в парах моделей мотива сравнили кратности изменения FC по всем возможным парам моделей: BaMM/PWM, SiteGA/PWM и SiteGA/BaMM. В каждом наборе данных ChIP-seq посчитали средние значения отношений кратностей изменения по трём парам моделей (FC_{BaMM}/FC_{PWM}, FC_{SiteGA}/FC_{PWM} и FC_{SiteGA}/FC_{BaMM}), после чего было построено распределение этих величин по всей коллекции (рис. 7).

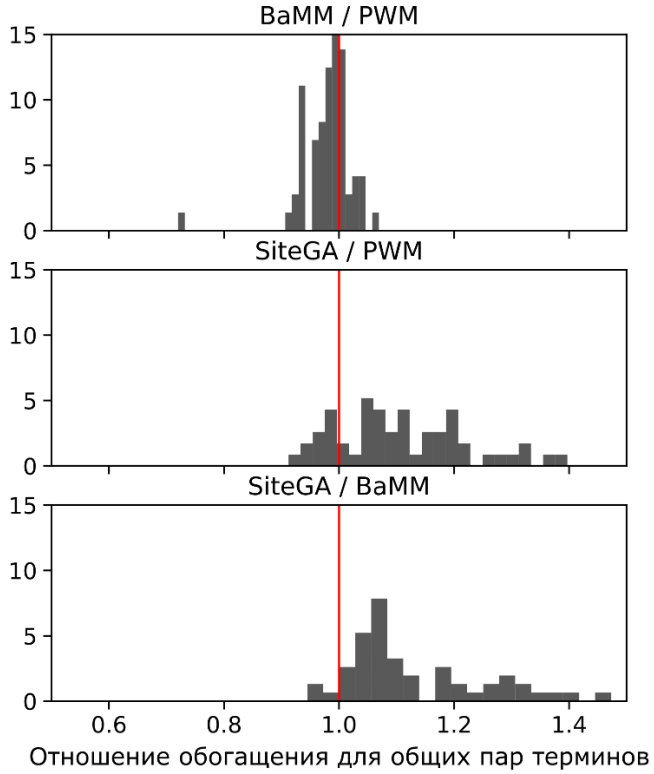

Рис. 7. Гистограммы распределений средних отношений кратностей изменения, рассчитанных для общих терминов ГО для пар моделей (BaMM/PWM, SiteGA/PWM и SiteGA/BaMM) по всей коллекции наборов данных ChIP-seq *A. thaliana.*

Как видно из диаграммы (рис. 7) для пары BaMM/PWM полученное распределение находится в диапазоне от 0.9 до 1.05 с максимумом вблизи единицы (ожидаемое значение), что говорит о схожести величин кратности изменения. Для пар SiteGA/PWM и SiteGA/BaMM распределения существенно сдвинуты вправо относительно единицы. Такой результат свидетельствует о том, что у модели SiteGA кратность изменения для терминов ГО систематически больше таковых для PWM и BaMM (рис. 7). Средние значения отношений кратности изменения FC_{BaMM}/FC_{PWM},

FC_{SiteGA}/FC_{PWM} и FC_{SiteGA}/FC_{BaMM} для полученных распределений равны 0.98. 1.11 и 1.15. соответственно.

Для каждой пары моделей в каждом ChIP-seq эксперименте был применен U-тест Манна-Уитни для сравнения отношений кратностей изменения. После чего для каждой пары моделей был применён метод Фишера, который позволяет получить единое значение p-value (мета p-value) (Heard and Rubin-Delanchy 2018), на основании всех значений p-value, посчитанных для каждого эксперимента. Единое значение p-value позволяет сделать вывод для всей коллекции данных. Результаты расчётов приведены в таблице 2.

Таблица 2. Результаты сравнения кратности изменения по всей коллекции данных ChIP-seq *A. thaliana.* для пар моделей (PMW/BaMM, PWM/SiteGA, BaMM/SiteGA) с использованием U-теста Манна-Уитни.

Пара моделей	**Число экспериментов ChIP-seq**		**Мета p-value****
	Общее	Со значимыми отличиями*	
PMW/BaMM	62	5	P> 0.05
PWM/SiteGA	55	27	$P<3*10^{-37}$
BaMM/SiteGA	55	28	$P<5*10^{-52}$

* - эксперименты для которых U-тест Манна-Уитни показал значение $p < 0.05$

** - мета p-value, посчитанное с помощью метода Фишера (Heard and Rubin-Delanchy 2018), характеризует результат по всей коллекции данных.

Из полученных данных (таблица 2) видно, что модель SiteGA имеет более высокие значения кратности изменения для общих терминов ГО по сравнению с моделями PWM и BaMM на всей коллекции данных. Более высокие значения кратностей обогащения общих терминов ГО у модели SiteGA по сравнению с другими моделями мотивов могут быть связаны с двумя аспектами. Во-первых, модель SiteGA чаще, чем другие модели, предсказывает ССТФ в регуляторных районах генов. Во-вторых, модель SiteGA предсказывает ССТФ в тех генах в которых другие модели не предсказывают ССТФ. Этот результат подчёркивает функциональную важность ССТФ, предсказанных моделью SiteGA в промоторах генов.

Заключение

Был разработан программный комплекс MultiDeNa для анализа данных ChIP-seq, который позволяет совместно применять несколько моделей мотива (PWM, diPWM, BaMM, InMoDe, SiteGA) для распознавания ССТФ в пиках ChIP-seq и проводить классификацию пиков на основании присутствия/отсутствия предсказанных сайтов разными моделями в пиках. Программный комплекс MultiDeNa был апробирован, а затем применён для анализа двух больших коллекций наборов данных экспериментов ChIP-seq *A. thaliana* / *M. musculus*, включающих наборы по 68 / 1003 экспериментам, соответственно.

По сравнению с применением только одной модели PWM, использование нескольких методологически разных моделей позволяет в среднем находить больше

пиков с ССТФ. Вклад альтернативных моделей может существенно отличаться в зависимости от класса ДСД целевого ТФ. Например, сравнение всех классов ТФ по двум видам организма показывает, что медианы по суммарной добавке двух альтернативных моделей к доле пиков, распознанных моделью PWM составляют 13.5% и 13.55% соответственно, у *A. thaliana* и *M. musculus* они выявлены для классов *Fork head/winged helix factors* {3.3} и *Basic helix-loop-helix factors (bHLH)* {1.2}), а максимальное и минимальное значение этой добавки получены для классов *GCM domain factors* {7.2} *A. thaliana* 27.5%, и *C2H2 zinc finger factors* {2.3} *M. musculus* - 4.2%. Следовательно, структурное разнообразие сайтов и вклад зависимостей позиций в информационное содержание их нуклеотидного контекста (оцениваемое моделью мотива как аффинность ССТФ) может зависеть от структуры ДСД. Помимо этого, заметные доли пиков с сайтами только альтернативных моделей мотива BaMM / SiteGA позволяют предполагать, что при заданной ошибке перепредсказания, анализ ChIP-seq данных с привлечением разных моделей мотива определяет значительно больше потенциальных ССТФ, чем может дать одна модель PWM.

Картирование сайтов разных моделей в промоторах генов показало, что часть генов имеют в промоторе сайты только одной из моделей, при этом доли таких генов для альтернативных моделей всегда больше, чем соответствующая доля модели PWM. Определены два класса с наибольшим и наименьшим вкладами альтернативных моделей в распознавания ССТФ, *Basic helix-loop-helix factors (bHLH)* {1.2}, и *C2H2 zinc finger factors* {2.3}. С учётом оценок точности моделей и расчёта долей распознанных пиков и промоторов генов эти классы совпадают для двух видов организмов. Такой результат отражает вклад зависимостей разных позиций в паттерн нуклеотидного контекста, отвечающего за специфичность связывания ТФ класса с геномной ДНК *in vivo*.

Анализ обогащения терминов ГО для коллекций ChIP-seq данных *A. thaliana* и *M. musculus* показал, что альтернативные модели существенно увеличивают количество терминов ГО по сравнению с моделью PWM, что расширяет общий список биологических процессов, с которыми могут быть связаны гены-мишени ТФ. Также для терминов ГО, которые обогащены для всех трёх моделей, именно модель SiteGA имеет значимо большие значения кратности изменения терминов ГО, чем модели PWM и BaMM. Этот результат можно интерпретировать как способность модели SiteGA более надёжно, чем модели PWM и BaMM, выявлять ССТФ в промоторах генов, обладающих специфическими биологическими функциями целевых ТФ.

Выводы

1. Для массового анализа контекстной специфичности мотивов, соответствующих сайтам связывания транскрипционных факторов в геномных последовательностях пиков ChIP-seq экспериментов, впервые разработан программный комплекс MultiDeNa, включающий: (1) модель PWM, предполагающую независимые вклады позиций нуклеотидов сайта в оценку взаимодействия транскрипционного фактора

с ДНК, (2) модель BaMM, учитывающую зависимости между ближними позициями нуклеотидов сайта, и (3) модель SiteGA, учитывающую зависимости частот динуклеотидов между отдельными блоками сайта.

2. На основе программного комплекса MultiDeNa проведен анализ более миллиона геномных последовательностей, выявленных в 1003 ChIP-seq экспериментах для 157 транскрипционных факторов *M. musculus* и 68 ChIP-seq экспериментах для 37 транскрипционных факторов *A. thaliana*. Проведённый анализ показал, что модель BaMM превосходит PWM в точности при распознавании сайтов со средней и низкой консервативностью. Модель SiteGA превосходит PWM в точности при распознавании сайтов с низкой консервативностью для транскрипционных факторов класса *Basic helix-loop-helix factors (bHLH)*.
3. Анализ результатов распознавания сайтов связывания транскрипционных факторов *A. thailana* и *M. musculus*, имеющих ДНК-связывающий домен класса *bHLH*, показал, что модель PWM находит сайты связывания таких факторов только в 52-55% геномных последовательностей пиков ChIP-seq экспериментов. Установлено также, что совместное применение моделей BaMM и SiteGA, дополнительно даёт распознанные сайты связывания транскрипционных факторов класса *bHLH* в 13-23% геномных последовательностей пиков ChIP-seq экспериментов.
4. Показано, что каждая из трёх моделей (PWM, BaMM и SiteGA) выявляет сайты связывания транскрипционных факторов, локализованные в промоторах определенных групп генов, которые достоверно ассоциированы с некоторыми терминами генной онтологии (ГО). Выявлены также термины ГО общие для всех трёх моделей и уникальные для каждой модели. Установлено, что для общих терминов модель SiteGA, по сравнению с моделями PWM/BaMM, имеет значимо более высокую долю генов с предсказанными сайтами в промоторах, например, для коллекции ChIP-seq данных *A. thaliana*: SiteGA против PWM, $p < 4*10^{-33}$, SiteGA против BaMM, $p < 2*10^{-22}$.

Публикации по теме диссертации

1. **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. Frontiers in plant science. 2022; 13. 938545.
2. **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.
3. Жимулёв И.Ф., Ватолина Т.Ю., Левицкий В.Г., Колесникова Т.Д., **Цуканов А.В.** Развитие идеи Н.К. Кольцова о генетической организации междисков политенных хромосом *Drosophila melanogaster*. Онтогенез. 2023; 54(2), 172–175.