ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК» (ИЦИГ СО РАН)

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

(биологические науки)

ДИССЕРТАЦИЯ

на соискание учёной степени кандидата биологических наук

Научный руководитель:

канд. биол. наук.,

Левицкий Виктор Георгиевич

Новосибирск - 2023

Список сокращений

ГА	генетический алгоритм
ГО	генная онтология
ДСД	ДНК-связывающий домен
ЛПД	локально-позиционированные динуклеотиды
ЛСД	Лиганд-связывающий домен
ММ	макрковская модель
ДРС	домен распознавания сигнала
СС	сайт связывания
ССТФ	сайт связывания транскрипционного фактора
ТАД	трансактивирующий домен
ТФ	транскрипционный фактор
AUC	area under ROC-curve, площадь под ROC-кривой
BaMM	Bayesian Markov Model
CV	cross-validation, перекрёстная проверка
diPWM	динуклетидная позиционная весовая матрица
ERR	expected recognition rate, ожидаемая частота распознавания
FN	false negative, число непредсказанных функциональных объектов
FP	False positive, число предсказанных нефункциональных объектов
FPR	False positive rate, отношение числа предсказанных нефункциональных объектов к общему числу нефункциональных объектов
IQR	Interquartile range, межквартильный диапазон
pAUC	partial AUC, частичная площадь под кривой
PCT	parsimonious context tree, скупое контекстного дерево
PWM	position weight matrix, позиционная весовая матрица

ROC	receiver operating characteristic, рабочая характеристика приёмника, зависимость TPR от FPR
TN	True negative, верно предсказанный нефункциональный объект
TP	True positive, верно предсказанный функциональный объект
TPR	true positive rate, отношение числа предсказанных функциональных объектов к общему числу функциональных объектов

Введение

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 [1]. 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 [2–4].

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 [5–7]. 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. [8, 9].Для того, чтобы найти точную форму мотива в наборе пиков используются алгоритмы *de novo* поиска мотивов [10]. 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) [11, 12] и наиболее популярные из них это HOMER [13], Streme [14], MEME-ChIP [15] и ChIPMunk [16]. Без преувеличения можно сказать, что применение разных реализаций модели PWM входит практически в каждый конвейер обработки полногеномных данных ChIP-seq [17]. Модель PWM широко применяется для изучения регуляции транскрипции *in silico*. Она используется для поиска ССТФ [18] в предполагаемых регуляторных последовательностях, для предсказания *цис*-регуляторных элементов [19] и для определения возможной регуляторной роли однонуклеотидных полиморфизмов [20, 21].

Однако, многократно экспериментально показано [22, 23], 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 [24, 25]. Помимо этого, существуют и другие особенности связывания ТФ с ДНК, такие как разнообразие структурных типов ССТФ [26, 27], нуклеотидный состав флангов ССТФ, возможности разных ТФ действовать в составе гомо- и гетродимеров [28], особенности взаимодействия разных ТФ с нуклеосомной ДНК [29], 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 [30–34]. 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 [30, 35, 36].

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 [25, 31, 37–41]. Авторы подобных моделей, таких как BaMM [39], InMoDe [42] и Slim [25] 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 [43].

Ранее было показано, что совместное применение моделей SiteGA и PWM, позволяет находить принципиально разные структурные типы ССТФ [44, 45], более того, сайты таких разных структурных типов регулировали гены с различными функциями [44]. 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 [39]. 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 [34, 46].

Цель и задачи исследования

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.

Публикации

По материалам диссертации опубликовано 6 работ, из них 3 статьи в рецензируемых научных журналах, входящих в перечень ВАК, 3 тезиса конференций. Получено два авторских свидетельства.

Статьи:

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. Жимулев 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

Авторские свидетельства:

1. Левицкий В.Г., **Цуканов А.В.** Программный комплекс для поиска мотивов в данных полногеномного картирования сайтов связывания транскрипционных факторов ChIP-seq (SiteGA). Свидетельство о

регистрации № 2021616695, зарегистрировано в Реестре программ для ЭВМ 26.04.2021. - https://sites.icgbio.ru/intellectual-property/sitega/

2. Левицкий В.Г., **Цуканов А.В.** Программа для генерации выборки негативных последовательностей ДНК при анализе обогащения мотивов в данных массового секвенирования (SuppressBias). Свидетельство о регистрации №2022612715, зарегистрировано в Реестре программ для ЭВМ 28.02.2022. - https://sites.icgbio.ru/intellectual-property/suppressbias/

Тезисы:

1. **Цуканов А.В.,** Левицкий В.Г. Поиск скрытых сайтов связывания транскрипционных факторов в данных ChIP-seq // VII молодёжная школа-конференция по молекулярной и клеточной биологии Института цитологии РАН, (12–15 октября 2020, Санкт-Петербург, Россия); *Гены и клетки*. 15(3). 154
2. **Tsukanov A.V.**, Levitsky V.G., Merkulova T.I. Analysis of short- and long-range interactions within potential binding sites notably extends the fraction of verified peaks in ChIP-seq data // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2020): The Twelfth International Multiconference (06–10 July 2020, Novosibirsk, Russia); Abstracts. – P. 126-127
3. **Цуканов А.В.**, Левицкий В.Г., Эффективность моделей распознавания сайтов связывания транскрипционных факторов обусловлена структурой ДНКсвязывающих доменов // Системная биология и биоинформатика (SBB-2023): 14-я международная школа молодых ученых (22–26 мая 2023 г., Новосибирск, Россия); Тезисы докладов. – С. 36

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.

1. Литературный обзор

1.1 Функции транскрипционных факторов

Термин транскрипционный фактор (ТФ) используется для белков, которые обладают двумя ключевыми свойствами: (1) способностью связываться со контекст-специфическими сайтами в ДНК и (2) регуляцией уровня транскрипции гена [1]. ТФ достаточно большая группа белков. В среднем доля генов, кодирующих ТФ, составляет от 4 до 10% от всех кодирующих белок генов генома. Например, у *Arabidopsis thaliana* в 2492 генах закодированы ТФ, что составляет 9% от всех кодирующих белок генов [47]. У *Homo sapience* 1639 генов в которых закодированы ТФ, что составляет около 7% от всех кодирующих белок генов [1].

ТФ являются ключевыми элементами в регуляции экспрессии генов на уровне транскрипции. Регуляция транскрипции осуществляется за счёт связывания ТФ с сайтами в ДНК в 5’-проксимальных и 5’-дистальных регуляторных районах генов, таких как промоторы и сайленсеры / энхансеры. События, следующие после связывания ТФ с ДНК, могут приводить как к активации транскрипции: так и к её подавлению [2–4], это может достигаться как за счёт изменения плотности упаковки хроматина [29, 48, 49] и прямого взаимодействия с транскрипционным комплексом и РНК-полимеразой [50], так и через привлечение ко-факторов [51].

Роль ТФ в регуляции значительна, поскольку мутации в ССТФ или самом ТФ могут проводить к изменению аффинности ТФ к данному СС, что в итоге может привести к изменению регуляции транскрипции и, как следствие, к нарушению функций гена и даже различным заболеваниям [1, 52]. Существует множество заболеваний, вызванных ошибками в системе регуляции транскрипции и мутациями в ТФ. ТФ обогащены среди онкогенов [53, 54], треть нарушений развития у человека связаны с потерей функций ТФ [55, 56].

В настоящее время, наиболее изученными являются общие (базальные) ТФ. Они совместно с РНК-полимеразой II (Pol II) формируют преинициаторный комплекс транскрипции, который связывается с ДНК в коровом (базальном)

промоторе [57]. Всего в преинициаторный комплекс входит шесть базальных ТФ: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, каждый из которых выполняет свою роль. TFIID образует комплекс с TATA-связывающим белком (англ. TATA-binding protein, TBP), и, далее, за счёт TBP связывается с TATA-боксом (последовательностью нуклеотидов, которая может быть расположена на расстоянии в 25-35 п.о. от старта транскрипции). Комплекс TFIID-ДНК далее привлекает два следующих фактора TFIIA и TFIIB, где второй играет роль «моста» между комплексом факторов с ДНК и РНК-полимеразой II [50]. Функции остальных факторов представлены в таблице 1. Базальные факторы транскрипции являются конститутивными, без них невозможно начать транскрипцию, и экспрессируются они во всех клетках организма [50].

Таблица 1. Базальные ТФ входящие в состав преинициаторного комплекса

ТФ	Функция
TFIIA	Взаимодействует с TFIID и участвует в связывании с TATA-боксом.
TFIIB	Связующий фактор между промотором и РНК-полимеразой II. Так же участвует распознает BRE элемент в промоторе
TFIID	Включает в себя субъединицу TBP, связывается с TATA-боксом и рядом других элементов проксимальных промоторов.
TFIIE	Активирует TFIIH и стабилизирует открытый промоторный комплекс.
TFIIF	Стабилизирует TFIIB и преинициаторный комплекс
TFIIH	Катализирует раскрытие ДНК и фосфорилирование РНК-полимеразы II

Остальные ТФ являются факультативными, их экспрессия может быть ограничена несколькими типами клеток или стадией развития. Факультативные ТФ связываются с ДНК в цис-регуляторных элементах, участках ДНК, которые содержат большое количество разных ССТФ, на разных расстояниях от старта транскрипции. Они могут связываться недалеко от старта начала транскрипции в районе промотора (не далее 1000-5000 п.о. от старта транскрипции) или в энхансерах (на любых расстояниях от него) [50]. Такие ТФ могут способствовать изменению уровня транскрипции гена разными путями (Рисунок 1).

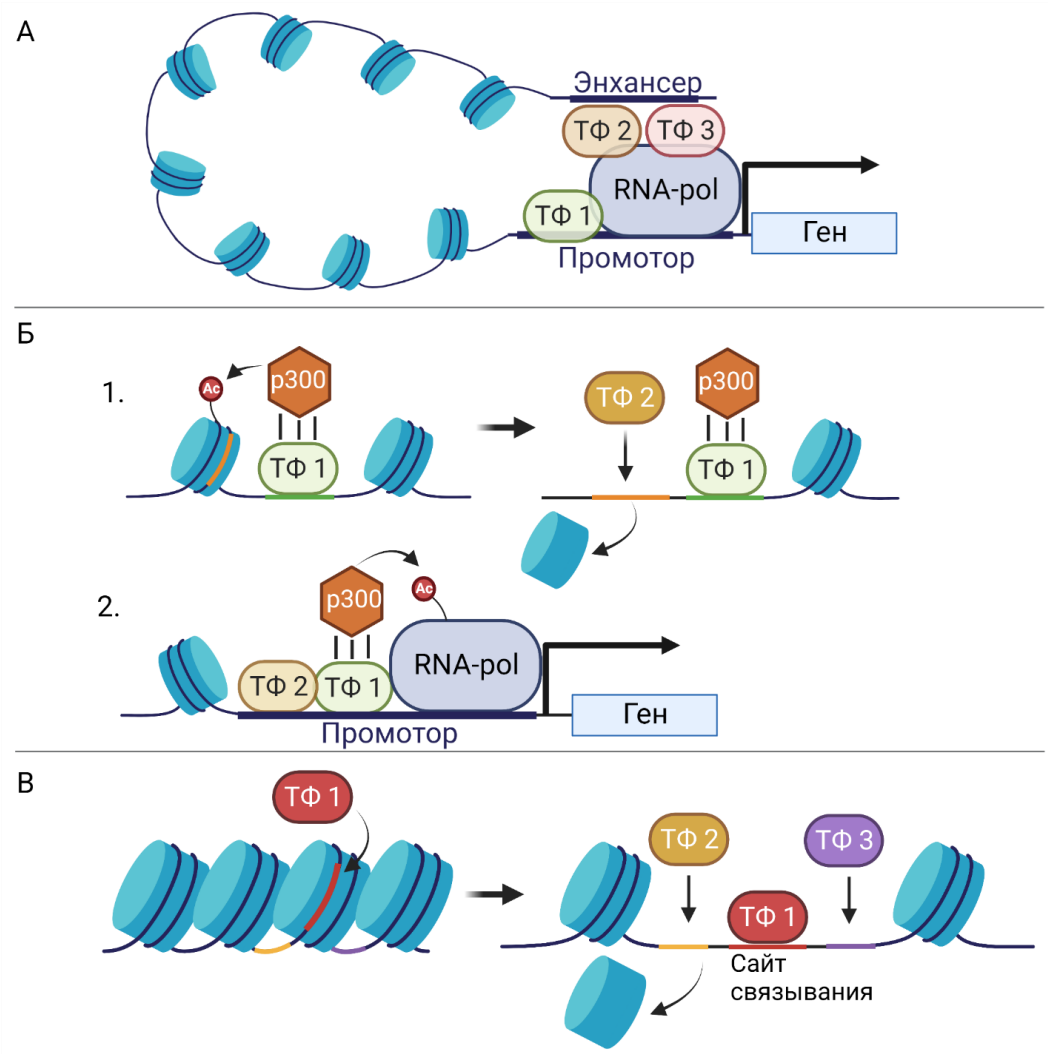

Рисунок 1. Основные функции ТФ. **(А)** ТФ взаимодействуют с преинициаторным комплексом РНК-полимеразы II за счет непосредственных или опосредованных контактов, что приводит к активации транскрипции; **(Б)** ТФ после связывания с ДНК привлекает ацетилтрансферазу p300, далее она ацетилирует гистоны вследствие чего нуклеосома дестабилизируется и освобождает ДНК, открывая доступ для другого ТФ (1). Так же ТФ может привлекать ацетилтрансферазу, которая ацетилирует транскрипционный комплекс, таким образом влияя на его работу (2); **(В)** Пионерские ТФ способны связываться с недоступными для других ТФ СС в составе нуклеосомной ДНК и инициировать ремоделирование хроматина, чтобы облегчить их дальнейшее связывание.

ТФ взаимодействуют с преинициаторным комплексом, они стабилизируют или блокируют его работу (Рисунок 1А) [50, 58]. ТФ могут привлекать ко-факторы как, например, Медиаторный комплекс, который создаёт условия для привлечения РНК-полимеразы II, или ацетилтрансферазу CBP/p300, которая может ацетилировать гистоны. Работа этих кофакторов может приводить к

дестабилизации нуклеосомы и освобождению ДНК от нуклеосом (открытый хроматин), это позволяет открыть доступ для работы других ТФ и РНК-полимеразы II (Рисунок 1Б) [51]. Так же стоит добавить, что другие ко-факторы могут приводить к закрытию хроматина [59, 60]. Помимо этого, ТФ могут изменять структуру хроматина за счёт их прямого воздействия с ДНК в составе нуклеосом [61]. В данном случае открытие хроматина достигается путём вытеснения гистоновых белков. ТФ, которые способны взаимодействовать с ДНК в составе нуклеосомы, а также вследствие этого привлекать другие ТФ к взаимодействию с этой ДНК, принято назвать «пионерскими» (Рисунок 1В, таблица 2) [49, 62, 63].

Связывание пионерских ТФ с нуклеосомной ДНК приводит к открытию хроматина из-за чего ССТФ, которые ранее были недоступны для некоторых ТФ, теперь могут быть ими заняты. Таким образом пионерские ТФ могут оказывать влияние на транскрипцию путём изменения структуры хроматина, открывая возможность для синтеза РНК для ранее не экспрессирующихся генов [64–66]. Обычно пионерские качества ТФ выявлялись при детальном изучении этого ТФ и биологических процессов, в которые ТФ был вовлечён [49, 64]. Одна из первых попыток сделать массовый анализ для выявления пионерских ТФ была сделана в работе [67], где авторы предложили вычислительный метод для определения количественной оценки взаимодействия белков (англ. Protein Interaction Quantitation - PIQ), который позволяет предсказывать взаимодействие ТФ с СС через анализ нескольких профилей эксперимента DNase-seq. Данный метод позволил авторам классифицировать 120 факторов из 733 ТФ, используемых в работе, как пионерских ТФ. Данные результаты значительно расширили список предполагаемых пионерских ТФ [67]. Помимо этого было классифицировано еще две группы ТФ: (1) «поселенцы» (англ. «settlers») – для данной группы ТФ связывание с ДНК напрямую зависит от открытости ДНК, если СС доступен для связывания, то такой ТФ свяжется с ДНК, их количество составило 131 шт.; (2) «мигранты» (англ. «migrants») – для данной группы ТФ

доступность СС не является достаточной причиной для связывания с ДНК, необходимо участие других ТФ и кофакторов, и она является самой крупной по количеству ТФ – 480 шт. [67].

Таблица 2. Основные ТФ, которые проявляют «пионерские» качества [64, 68–71]

Имя ТФ	**Таксон**
FOXA1	Животные
FOXA2	
FOXH1	
FOXD3	
FOXO1	
Ascl1	
C/EBPa	
Ebf1	
Esrrb	
GATA3	
GATA4	
GR/AR	
Klf4	
Neurod1	
Nrf1	
Oct4	
p53	
Pax7	
SPI1	
Sox2	
CLOCK, BMAL1	
MYOD1	
PBX1	
Zelda	
Grainy	
NF-YA	
PR	
ISL1	
LFY	Растения
AP1	
SEP3	
LEC1	

Обобщая всё вышесказанное, ТФ играют огромную роль в регуляции транскрипции, при этом осуществляют её разными способами [49–51, 58, 62, 63], однако ключевым этапом для ТФ остаётся связывание с ДНК, поскольку от этого зависит дальнейшее участие ТФ в регуляции [1]. Этап связывания ТФ с ДНК напрямую определяется структурой ТФ [72], как и его другие функции и особенности ТФ [73], поэтому в следующей главе будут описаны ТФ с точки зрения структуры белка.

1.2 Структура транскрипционных факторов

Как и все белки, ТФ имеют модульную структуру, которая включает в себя несколько доменов, из которых ключевым является ДНК-связывающий домен (ДСД), он осуществляет связывание ТФ с ДНК. Остальные домены являются «эффекторными» и выполняют разные функции, как например трансактивирующий домен (ТАД); домен распознавания сигнала (ДРС) или по-другому лиганд-связывающий домен (ЛСД); домен димеризации (ДД); домен для взаимодействия с эффекторами/ко-факторами и другие [74, 75]. Не все из перечисленных доменов обязательно должны быть у ТФ, за исключением ДСД. На рисунке 2 представлена доменная структура ТФ STAT6 и его комплекса с ДНК.

Поскольку основной функцией ТФ является считывание транскрипционного регуляторного кода ДНК, то основным для ТФ является ДСД, который и выполняет функцию распознавания специфических участков ДНК – ССТФ [76, 77]. Именно структура ДСД определяет последовательность ДНК, с которым будет связываться ТФ [72]. Благодаря данному домену ТФ могут иметь разную аффинность к нуклеотидным последовательностям. Так аффинность к ССТФ может быть в тысячи раз выше по сравнению с аффинностью к случайной нуклеотидной последовательности [78].

В основе классификации ТФ лежит структурная организация ДСД [76, 79]. В ранних работах все ТФ разделили на четыре суперкласса согласно структуре ДСД [76]:

1) факторы, ДСД которых обогащен положительно заряженными аминокислотными остатками (*basic domain*);

2) факторы, ДСД которых содержит координированные атомы цинка (*zinc-coordinated DNA-binding domain*);

3) факторы, ДСД которых содержит ДНК-связывающий мотив типа спираль–поворот–спираль (*helix–turn–helix*);

4) факторы, контактирующие с ДНК по малой бороздке, у которых ДСД представлен в виде β-скэффолда (*betta-scaffold factors with minor grooves contacts*).

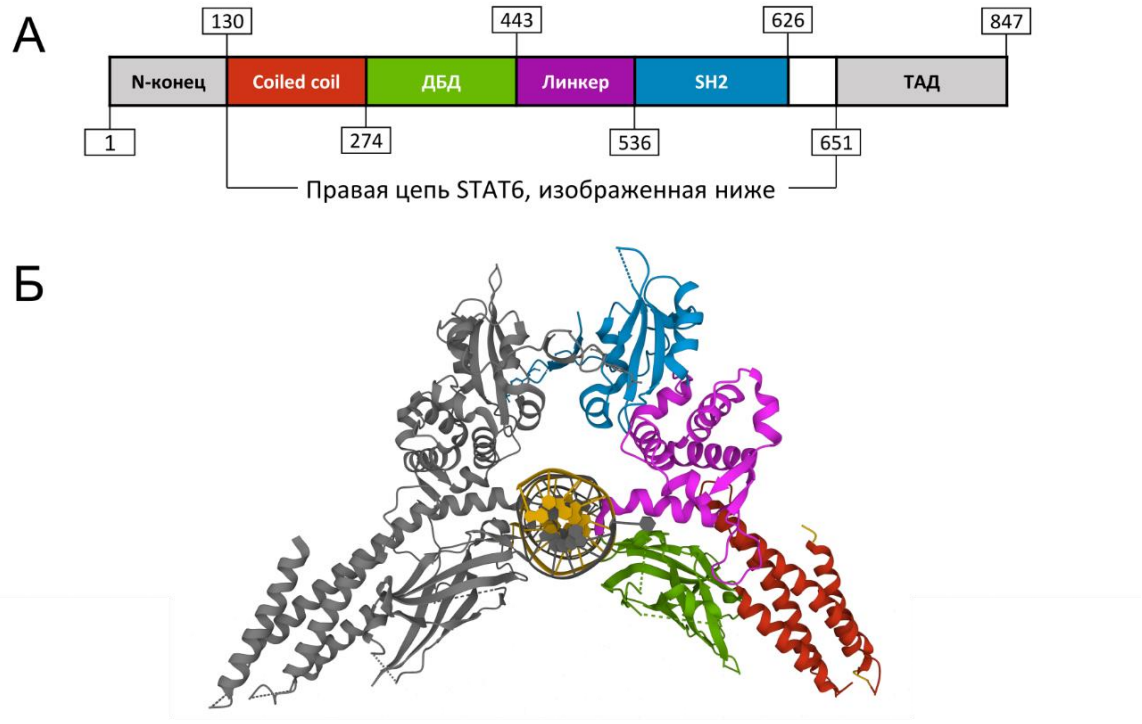

Рисунок 2. Структура ТФ STAT6 и его комплекса с ДНК. **(А)** Доменная структура STAT6, где coiled coil – домен, состоящий из нескольких альфа-спиралей; домен SH2 – позволяет присоединяться к фосфорилированным остаткам тирозина на других белках; линкер – участок, который соединяет между собой другие домены. **(Б)** Трёхмерная структура комплекса STAT6-ДНК, STAT6 является димерным белком и состоит из двух полипептидных цепей [80].

Однако данная классификация не была достаточной, так как с увеличением количества структур известных ТФ, те ТФ, которые не подходили по структуре к первым трем суперклассам, вносились в *четвёртый* суперкласс, из-за чего ТФ *четвертого* суперкласса могли сильно отличаться

друг от друга. В связи с этим возникла необходимость пересмотреть классификацию ТФ. Новую классификацию предложил Эдгар Вингендер, она включала в себя 9 суперклассов ТФ и так же десятый суперкласс (обозначается как нулевой - 0) к которому относят те ТФ, которые не попали в первые девять суперклассов (Рисунок 3) [74].

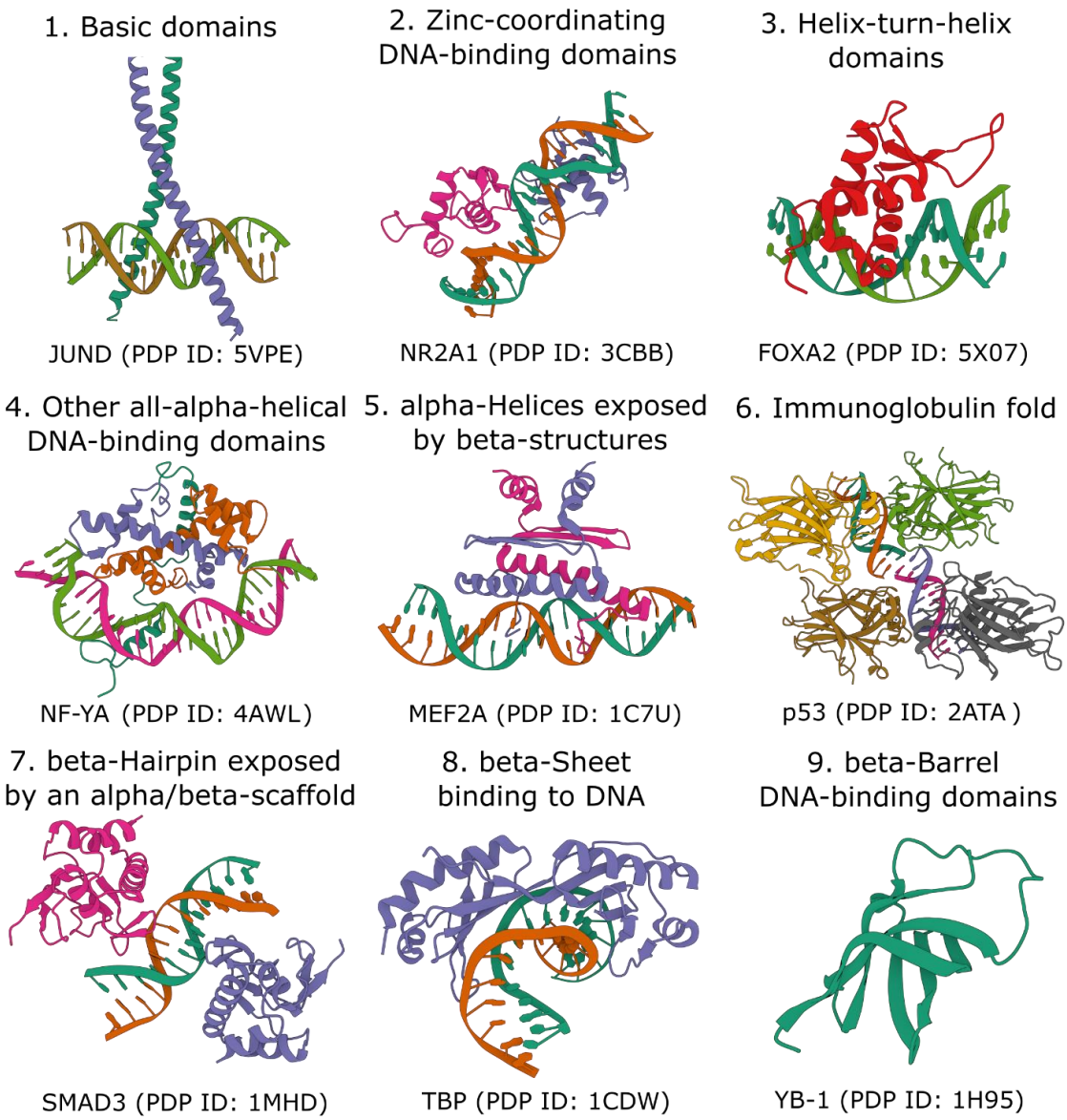

Рисунок 3. Девять суперклассов согласно классификации Вингендер [74, 81–83]. Для каждого суперкласса представлена трёхмерная структура взаимодействия ДСД с ДНК на одном из представителей данного суперкласса (в скобках записаны PDB ID для структуры), исключением является девятый суперкласс, где представлена только структура ДСД. Все структуры были взяты из PDB [84] и нарисованы с помощью molstar [85]. На рисунке суперклассы от 1 до 9 соответствуют ТФ JUND, NR2A1, FOXA2, NF-YA, MEF2A, p53, SMAD3, TBP и YB-1 [86–94]

Данная классификация представлена в базе данных TFClass и доступна по ссылке http://tfclass.bioinf.med.uni-goettingen.de/ [81–83]. Вингендер разделил суперклассы по схожести топологии ДНК-связывающего домена и по способу взаимодействия их с ДНК. Первые три суперкласса остались без изменений, а *четвертый* суперкласс был пересмотрен, из него выделили еще шесть суперклассов, каждый из которых отличается разным соотношением α- и β- структур в ДСД [74]. В базе TFClass применяется классификация ТФ, включающая шесть уровней иерархии, четыре из которых (суперкласс, класс, семейство, подсемейство) являются абстракциями по разным критериям, начиная от уровня суперкласса, определяемого по схожести топологии ДСД. Пятый уровень представляет гены, кодирующие ТФ, а шестой — отдельные генные продукты (Таблица 3).

Таблица 3. Пример классификации ТФ STAT6 согласно TFClass [83].

Уровень	Критерий классификации	Пример
1. Суперкласс	Схожесть топологии ДСД	Immunoglobulin fold {6}
2. Класс	Структурная схожесть ДСД	STAT domain factors {6.2}
3. Семейство	Схожесть аминокислотной последовательности и функции ДСД	STAT {6.2.1}
4. Подсемейство	Особенности аминокислотной последовательности	Нет подсемейства {6.2.1.0}
5. Ген		STAT6 {6.2.1.0.7}

Четвёртый суперкласс (*Other all-alpha-helical DNA-binding domains*), имеет ДСД, полностью состоящий из альфа-спиралей, и не имеет ничего общего с первыми тремя суперклассами по структуре или способу связывания с ДНК. ТФ, входящие в *четвертый* суперкласс, связываются с ДНК через взаимодействие с малыми бороздками двойной спирали ДНК, что приводит к значительному изгибу молекулы ДНК, и является важным условием для активации транскрипции [74].

ТФ из *пятого* суперкласса (*alpha-Helices exposed by beta-structures*) также содержат в структуре ДСД α-спирали, которые контактируют с ДНК, но при этом они экспонируются решеткой из β-цепей. В отличие от ТФ первых четырёх суперклассов, ДНК-связывающие α-спирали данного суперкласса не встраиваются ни в одну из бороздок ДНК, а скорее упаковываются против двойной спирали ДНК [74]. В этот суперкласс входят ТФ классов MADS-box {5.1} и SAND-domain {5.3}.

ДСД *шестого* суперкласса (*Immunoglobulin fold*) имеют укладку полипептидной цепи, похожую на таковую у иммуноглобулина. Она характеризуется β-ядерной структурой, которая обнажает контактную поверхность, состоящую в основном из петель, а также других элементов вторичной структуры, из которых выступают аминокислотные остатки, участвующие в связывании с ДНК [74]. Этот суперкласс включает в себя несколько функционально важных классов, таких как Rel homology region (RHR) factors {6.1} (например, ТФ NF-κB), STAT domain factors {6.2} и p53 domain factors {6.3}.

ТФ *седьмого* суперкласса (*beta-Hairpin exposed by an alpha/beta-scaffold*) имеют ДСД, состоящий из каркаса α- и β-структур, которые обнажают β-шпильку. Именно β-шпилька контактирует с ДНК через большую бороздку. ДСД *восьмого* суперкласса (*beta-Sheet binding to DNA*) связываются с ДНК через одиночные вытянутые нити или β-листы, которые предпочтительно связываются в малой бороздке ДНК. Наиболее известным представителем этой группы является TBP. *Девятый* суперкласс (*beta-Barrel DNA-binding domains*) имеет структуру ДСД, которая включает β-бочонок из переменного количества β-цепей [74]. В последний *нулевой* суперкласс {0} объединены не классифицированные по остальным суперклассам ТФ.

Главной особенностью классификации базы данных TFClass является то, что она иерархичная, поэтому каждый суперкласс делится на классы на основе схожести аминокислотных последовательностей и структуры ДСД [74]. Классы делятся на семейства, которые разделяются не только по

схожести аминокислотных последовательностей ДСД, но и по схожести нуклеотидных последовательностей сайтов, с которыми связываются ТФ [81]. Далее семейства разделяются на подсемейства, но это характерно не для всех семейств [81]. Последние два уровня классификации представляют физические сущности — это ген и продукты этого гена, непосредственно ТФ (белки), поскольку некоторые гены могут иметь несколько продуктов (разные изоформы ТФ) из-за альтернативного сплайсинга [81]. В итоге классификация, без учёта изоформ белка, имеет 5 уровней и для каждого ТФ можно присвоить пятизначный код, например, ТФ STAT6, согласно данной классификации, имеет код — {6.2.1.0.7} (см. Таблица 3).

Помимо ДСД ТФ могут содержать и другие домены, среди которых может быть ТАД, который необходим для взаимодействия факультативных ТФ с базальными ТФ и преинициаторным комплексом. За счёт данного ДСД, ТФ могут связываться с белками коактиваторных или корепрессорных комплексов, а в ряде случаев способны непосредственно взаимодействовать с компонентами базальной транскрипционной машины [77].

Было показано, что значительное количество ТФ имеют ТАД, как, например, E2F1, который непосредственно контактирует с базальными ТФ, включая TBP, TBP-ассоциированные факторы, TFIIA, TFIIB и TFIIH [75, 95, 96]. ТАД также могут взаимодействовать и привлекать компоненты белкового комплекса Медиатора, мультибелкового комплекса, участвующего в активации большого количества генов [97]. ТАД обычно классифицируют по их аминокислотному составу, они могут быть богаты положительно заряженными (кислыми) аминокислотными остатками (как, например, у ТФ E2F1 и p53), остатками глутамина как у ТФ Oct1, Oct2 и Sp1 или остатками пролина как у ТФ AP-2 и CTCF [75].

Функционально ТАД разделяют на две группы, одна отвечает за стимулирование инициации транскрипции, другая стимулирует её элонгацию, на основании различных контактов, которые они устанавливают с базальными факторами транскрипции [98]. Считается, что ТФ с ТАД преимущественно

действуют на уровне инициации транскрипции [99]. Однако контакт между ТАД и преинициаторным комплексом может стимулировать транскрипцию не только за счёт стабилизации преинициаторного комплекса, но также может способствовать увеличению скорости элонгации [75, 100, 101].

Структурные исследования ТАД показали, что большинство ТАД в растворе не структурированы [102]. Например, исследования [103, 104] показали отсутствие структуры в N-концевой области ТФ p53, содержащего ТАД. Дальнейший анализ показал, что специфические мотивы в ТАД сворачиваются в α-спираль при связывании с комплексом инициации транскрипции. Такие исследования предполагают, что субдомены внутри ТАД становятся конформационно ограниченными при взаимодействии с белком-мишенью [105].

Другим важным структурным элементом, который может иметь в своей структуре ТФ, является домен димеризации (ДД) [75]. За счёт взаимодействия ДД двух ТФ могут быть образовываться как гомо- и гетеродимеры (комплексы из двух одинаковых и разных ТФ), а также более крупные структуры (мультимеры), состоящие из трёх и более ТФ, связывающихся с ДНК [106]. Кооперативные взаимодействия между разными ТФ расширяют возможности для распознавания последовательности ДНК, возможно, позволяя связывание сайт-специфического фактора с последовательностью, в меньшей степени соответствующей известному мотиву ТФ [33]. Кроме того, физическая ассоциация ТФ в энхансерах или промоторах не только стабилизирует слабые ДНК-белок взаимодействия одного фактора с ДНК, но также обеспечивает комбинаторную регуляцию, что является важным механизмом, обеспечивающим интеграцию различных сигнальных путей [107].

В дополнение к базальным и сайт-специфическим ТФ существуют другие типы регуляторных белков – ко-факторы (корепрессоры, коактиваторы), некоторые ТФ способны их привлекать к генам-мишеням за счёт наличия специальных доменов белок-белковых взаимодействий. Многие ко-факторы обладают ферментативной активностью, которая помогает в

регуляции генов посредством посттрансляционной модификации гистонов. Белок-белковые взаимодействия между ТФ и ферментами, модифицирующими гистоны, играют важную роль в регуляции транскрипции эукариотических генов. Например, ацетилирование гистонов связано с открытым хроматином и активацией генов, тогда как метилирование гистонов может быть связано как с активацией (например, метилирование лизина 4 гистона H3), так и с репрессией (например, метилирование лизина 9 или лизина 27 гистона H3) [108–110]. Многие ТФ, например, члены семейства E2F-related factors {3.3.2} обладают такими доменами, они могут напрямую взаимодействовать с ацетилтрансферазами гистонов p300/CBP [111–113] через свой C'-концевой ТАД. Белок p300/CBP представляет собой неспецифичную ацетилтрансферазу, поскольку она катализирует ацетилирование лизинов у всех четырёх основных типов гистонов (H2A/H2B/H3/H4), а также ацетилирование более 70 негистоновых белков, включая саму себя. Члены семейства E2F-related factors {3.3.2} также взаимодействуют с репрессивными гистон-модифицирующими комплексами. Например, ТФ E2F6 связывается с репрессорными комплексами GLP и G9a, которые вовлечены в метилирование лизина 9 гистона H3 [114]. Так же высказано предположение, что большое семейство ТФ, содержащих *Krüppel-associated box* (KRAB) домен, в котором насчитывается более 300 различных членов, репрессирует транскрипцию специфических генов посредством взаимодействия с корепрессорным белком KAP1 [59, 60]. В свою очередь, корепрессор KAP1 функционирует как каркас для привлечения изоформ гетерохроматинового белка 1 (HP1), гистоновых деацетилаз и SETDB1, гистоновой метилтрансферазы с SET-доменом, которая метилирует гистон H3 по лизину 9 [59, 60]. Эта модификация связана с закрытым хроматином, и, следовательно, ТФ, несущие KRAB-эффекторные домены, связывают корепрессорный комплекс KAP1 со специфическими геномными сайтами и подавляют экспрессию генов, образуя факультативное гетерохроматиновое окружение [115]. Так как чрезвычайно большое количество ТФ имеют домен

KRAB, то, возможно, он является наиболее распространённым эффекторным доменом, участвующих в репрессии [75].

Некоторые ТФ также имеют ЛСД/ДСР, с которыми специфически связываются различные сигнальные молекулы (например, стероидные и тиреоидные гормоны, витамин D, ретиноевые кислоты, простагландины и др.), меняющие конформацию и активность ТФ [77]. Как правило, ТФ также имеют несколько участков, опознаваемых различными модифицирующими ферментами, осуществляющими фосфорилирование, ацетилирование, сумоилирование, метилирование и убиквитинирование, что включает эти белки в разнообразные сигнальные пути [77]. Для ТФ STAT6 таким доменом является SH2, фосфорилирование тирозина в составе данного домена позволяет ему взаимодействовать с другими белками, которые также имеют фосфорилированный тирозин [116].

1.3 Общие представления о связывании ТФ с ДНК

До настоящего момента не существует полного представления о том каким образом и за счёт чего ТФ связываются с ДНК, но существует множество экспериментальных данных, на основе которых строят разные модели, помогающие разобраться в данной проблеме.

Прежде чем связаться с СС, ТФ должен найти его на ДНК. В нескольких работах [117–119] был предложен ряд механизмов, которые описывают возможные механизмы движения ТФ в процессе поиска СС. Предложенный ряд механизмов получил название «облегченная диффузия». Она включает в себя стадии: 1) трёхмерной диффузии ТФ в нуклеоплазме; 2) одномерного скольжения вдоль молекулы ДНК, которое обеспечено неспецифическим связыванием ТФ с ДНК; 3) переноса ТФ между сегментами ДНК, которые находятся в непосредственной близости; 4) прыжки, в ходе которых ТФ отсоединяется от одного участка ДНК и вновь осуществляет трёхмерную диффузию, а затем присоединяется уже к другому участку ДНК, который

находится на близком расстоянии от первого [120], на рисунке 4 схематично представлены все механизмы движения ТФ.

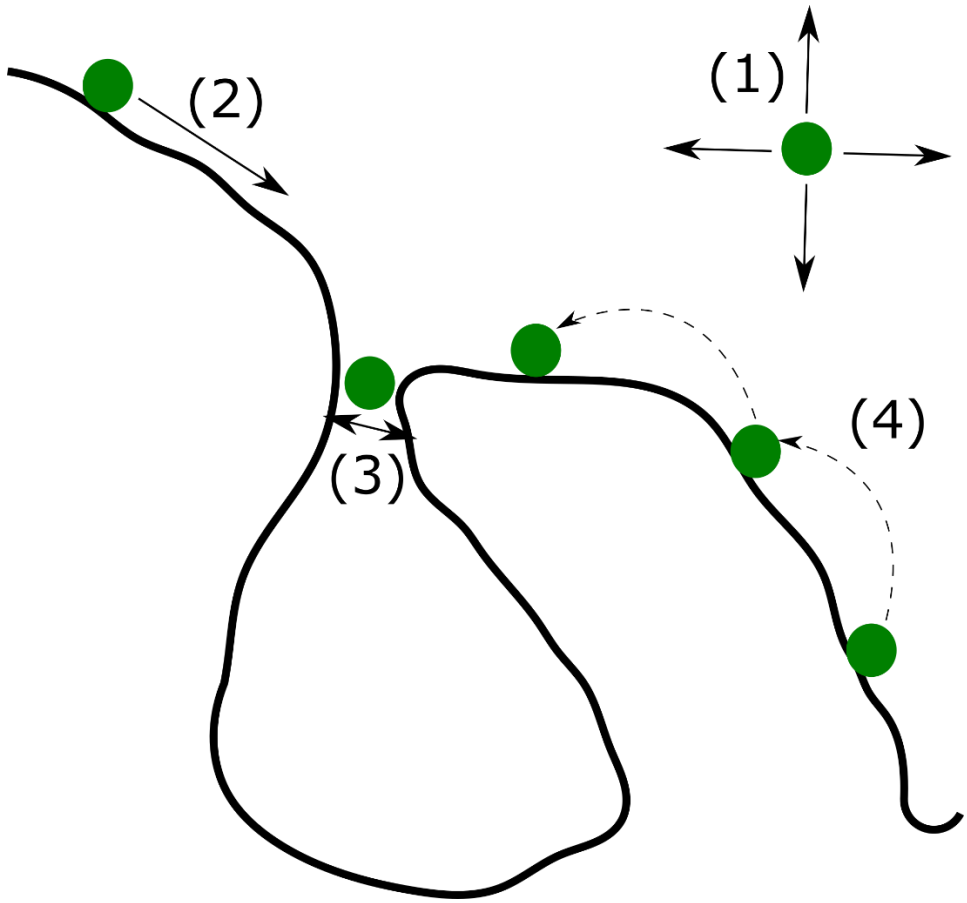

Рисунок 4. Механизм движения ТФ в рамках теории «облегченной диффузии». (1) – трёхмерную диффузия ТФ в нуклеоплазме; (2) – одномерное скольжение вдоль молекулы ДНК; (3) – перенос ТФ между сегментами ДНК; (4) – прыжки, в ходе которых ТФ отсоединяется от одного участка ДНК и вновь присоединяется уже к другому участку ДНК.

В ходе «облегченной диффузии» ТФ находит свой СС, который представляет из себя определенную последовательность нуклеотидов (мотив, ССТФ). ТФ считывают последовательности геномной ДНК тремя основными способами: а именно (1) считывание оснований, (2) непрямое считывание (взаимодействие с фосфатным остовом) и (3) считывание структуры ДНК [121, 122]. В ходе считывания оснований ТФ распознают заданную нуклеотидную последовательность посредством физико-химических взаимодействий между боковыми цепями аминокислот и доступными краями пар оснований ДНК. Данные взаимодействия включают в себя: водородные связи, гидрофобные взаимодействия и солевые мостики [121, 122]. Непрямое

считывание включает главным образом взаимодействие между ТФ и фосфатным остовом ДНК, положение которого зависит от природы основания, но не так сильно, как при считывании оснований. В ходе считывания структуры ДНК ТФ распознают конформационные особенности двойной спирали ДНК, такие как смещения по отношению к осям (shift, slide, rise), углы поворота (twist, tilt и roll), ширина/глубина большой и малой бороздок ДНК и другие, более подробно данный механизм считывания будет рассмотрен далее в разделе 1.4.2.2 [123–127].

Для поиска ССТФ предлагаются разные модели, которые позволяют предсказывать их в ДНК [47, 128]. Обычно такая модель может быть получена как из небольшого набора известных ССТФ (десятки последовательностей) с применением стандартных программ множественного выравнивания [129], так и из экспериментальных данных, таких как ChIP-seq, HS-SELEX, DAP-seq и других, с применением *de novo* подхода поиска мотивов. Алгоритмы *de novo* поиска позволяют выявить мотив через множественное локальное выравнивание для полногеномного набора данных (включающего тысячи последовательностей) согласно заданной модели мотива (консенсус, PWM, и т.д.), которая встречается в нуклеотидных последовательностях (пики ChIP-seq или другие экспериментальные данных) чаще, чем ожидается по случайным причинам [130, 131]. В следующей главе будут описаны основные группы моделей, которые используются для поиска мотивов ССТФ.

1.4 Модели, используемые для описания сайтов связывания транскрипционных факторов

1.4.1 Стандартные модели

1.4.1.1 Консенсус

Самой первой моделью мотива, предложенной для описания ССТФ, был консенсус, который строится на основе выравнивания известных ССТФ. Консенсус, это последовательность, где для каждой позиции ССТФ приписывался определенный совершенный или вырожденный нуклеотид,

который чаще всего встречается в данной позиции сайта. Например, для ТФ STAT6 консенсус выглядит следующим образом – TTCNNNNGAA, где N – это любой нуклеотид [80]. Для представления изменчивости нуклеотидов, описывающих разнообразие сайтов, был предложен расширенный 15-буквенный алфавит IUPAC (Таблица 4), который включает в себя помимо четырёх нуклеотидов любые комбинации двух или трёх из них, или любой из четырёх [132, 133].

Таблица 4. Кодировка IUPAC для расширенного консенсуса

Код	Расшифровка	Объяснение	
A	Аденин		
C	Цитозин		
G	Гуанин		
T	Тимин		
Y	Пиримидин (T или C)	pYrimidine	
R	Пурин (A или G)	puRine	
K	Кето-группа (G или T)	Keto	
M	Амино-группа (A или C)	aMino	
S	Сильные связь (G или C)	Strong	
W	Слабая связь (A или T)	Weak	
B	Все кроме A (G или T или C)	ABcd	Позиция в алфавите
V	Все кроме T (G или C или A)	TUVw	
D	Все кроме C (G или A или T)	CDef	
H	Все кроме G (A или C или T)	GHij	
N	Любой нуклеотид	aNy (любой)	

Однако даже расширенный алфавит для консенсуса очень приблизительно описывает всё разнообразие СС, поэтому для описания ССТФ была предложена другая модель мотива – позиционная весовая матрица (PWM) [134]. В модели консенсуса любой вырожденный нуклеотид предполагает равные частоты составляющих его совершенных нуклеотидов, что является грубым приближением и сильно увеличивает ошибку перепредсказания при распознавании ССТФ по отношению к модели PWM, учитывающей неравномерность частот нуклеотидов в позициях.

1.4.1.2 Позиционная весовая матрица

В настоящее время PWM получила большое распространение, и она используется, в качестве стандартной модели при изучении ССТФ. Как и любая другая модель, PWM строится на основе выравнивания множества ССТФ [135]. Основные этапы, необходимые для получения PWM, представлены на рисунке 5.

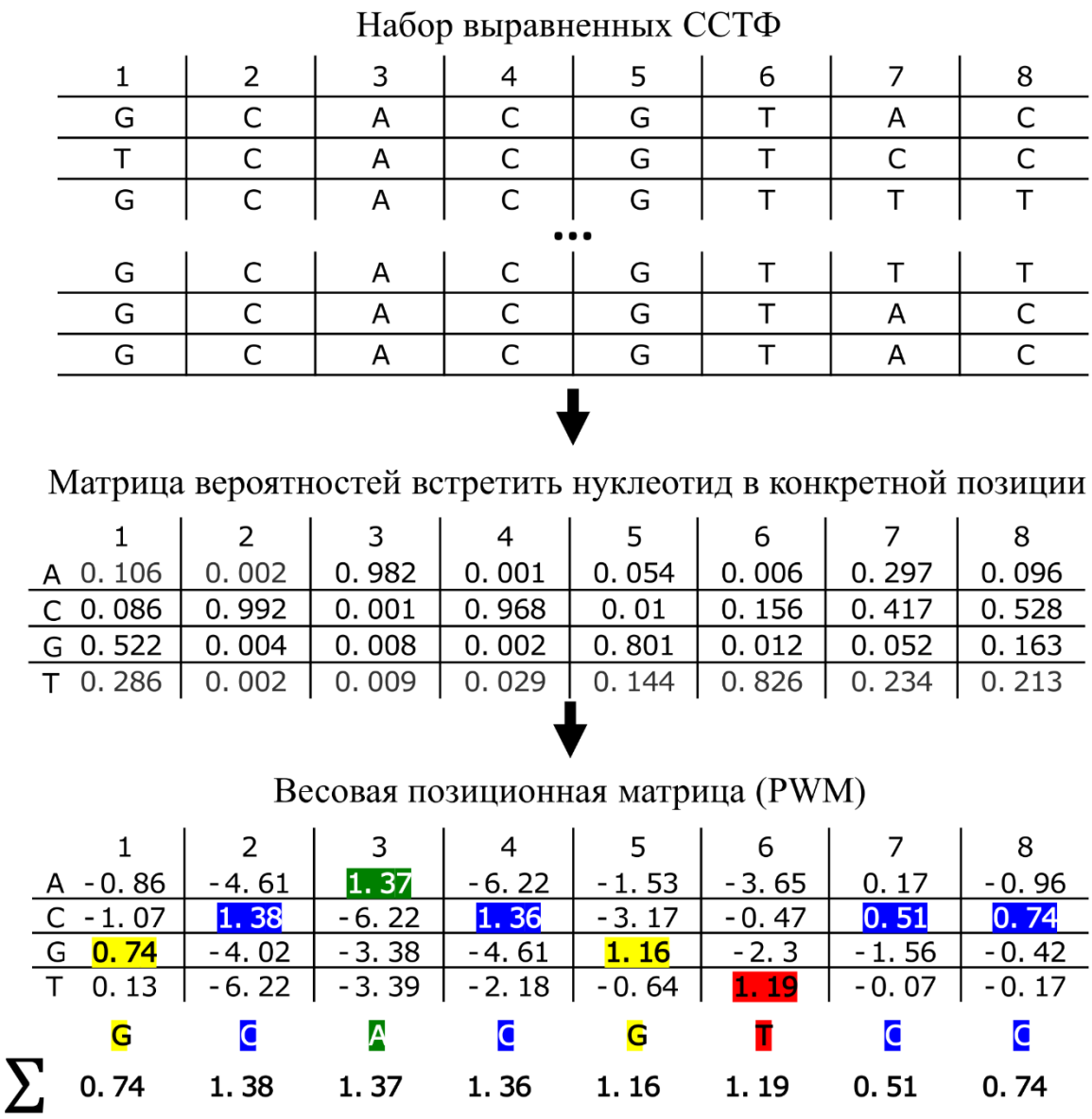

Рисунок 5. Общая схема расчёта PWM из выравненных ССТФ

На первом этапе для каждого нуклеотида в каждой позиции вычисляется вероятность встретить данный нуклеотид по следующей формуле [135]:

$$p(b,i) = \frac{m(b,i)+s(b)}{m+s} \qquad (1)$$

где $p(b, i)$ – вероятность встретить нуклеотид b в позиции i данного ССТФ, $m(b, i)$ – количество нуклеотидов b в данной позиции, m – общее количество нуклеотидов в данной позиции (общее количество выравненных ССТФ), $s(b)$ – в общепринятой английской терминологии обозначается как *pseudocount*, данный параметр особенно важен для малых выборок ССТФ, и в исследованиях его вычисляют или определяют разными способами: s(b) = 0.01, s(b) = 1, s(b) = 2, s(b) = $\sqrt{m}$, s(A)=s(C)=s(T)=s(G)=1/4 [136]. В результате получается матрица частот нуклеотидов (англ. position frequency matrix – PFM). Далее от вероятностей переходят к весам, используя следующую формулу:

$$W_{b,i} = log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

где b ={A,C,G,T}, p(b) – вероятность встретить нуклеотид b в геноме, p(b,i) – вероятность встретить нуклеотид b в позиции i данного ССТФ (1), $W_{b,i}$ – вес нуклеотида b в позиции i.

Используя PWM, можно определить значение функции распознавания заданной последовательности нуклеотидов, используя формулу [135]:

$$S = \sum_{i=1}^{w} W_{l_i,i} \quad (3)$$

где l_i – нуклеотид в позиции i заданной последовательности нуклеотидов, W – длина PWM [135]. Для PWM можно рассчитать порог, значения функции распознавания выше порогового будут соответствовать предсказанным мотивам для данной PWM.

Применяя информационную теорию Шеннона [137] для PWM, можно представить модель PWM в виде лого-диаграммы (Рисунок 6), которое отображает информационное содержание каждого нуклеотида в каждой позиции и даёт визуальное представление о позициях наиболее консервативных нуклеотидов ССТФ [134].

Значения функции распознавания у модели PWM коррелирует со свободной энергией Гиббса, выделяемой при связывании ТФ с ДНК. На

основании этого можно определить, свяжется ли ТФ с заданной нуклеотидной последовательностью или нет [11, 12].

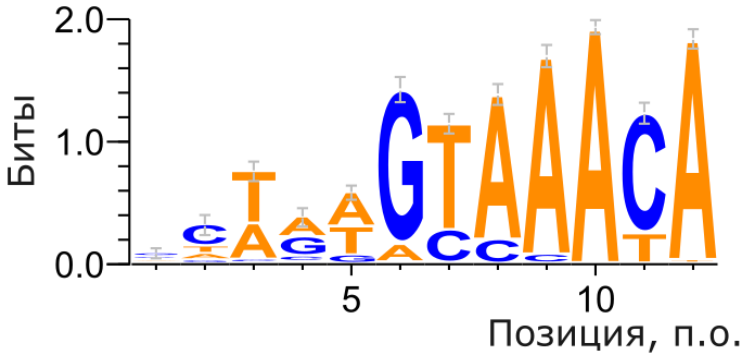

Рисунок 6. Визуальное представление модели PWM для ССТФ FOXA2 в виде лого-диаграммы. Ось абсцисс – позиция в ССТФ, ось ординат – информационное содержание в битах. Изображение получено с помощью WebLogo [2].

Как было показано выше, для того чтобы определить параметры PWM, необходим выравненный набор ССТФ, а результат первичной обработки данных ChIP-seq даёт на выходе только набор геномных последовательностей, который соответствует всем статистически значимым пикам ChIP-seq [138]. Поэтому требуется дополнительная компьютерная обработка данных, включающая использование алгоритмов поиска мотивов *de novo*. Существует множество *de novo* алгоритмов, и их ключевая идея состоит в том, чтобы найти обогащенные мотивы, представляющие общие паттерны часто встречающихся олигонуклеотидов, потенциальных ССТФ, в пиках ChIP-seq. Одни из самых популярных программ поиска мотивов с использованием модели PWM являются HOMER [13], MEME-ChIP [15], STREME [14], ChIPMunk [16].

В настоящее время уже получено огромное количество мотивов разных ТФ, которые представлены в разных базах данных в виде матриц частот нуклеотидов (PFM), из которых уже получают PWM. Наиболее известные базы данных, где представлены PFM: JASPAR (https://jaspar.uio.no/) [139], HOCOMOCO (http://hocomoco11.autosome.ru/) [8], CIS-BP (http://cisbp.ccbr.utoronto.ca/) [140].

После того, как модель мотива построена, с её помощью можно искать ССТФ как в геноме, так и в пиках ChIP-seq. Мотивы, полученные при *de novo* поиске можно сравнить с мотивами известных ТФ с использованием специального программного обеспечения, как например TomTom [141], что является полезным для подтверждения присутствия мотива исследуемого ТФ. Так же можно провести анализ того, какие другие ТФ могут связываться совместно с изучаемым ТФ, на каком расстоянии они находятся друг от друга, возможное перекрытие ССТФ и т.п. Данная информация поможет изучить механизмы действия ТФ [138].

1.4.2 Альтернативные модели ССТФ

Ключевым недостатком модели PWM является то, что она предполагает независимость вкладов отдельных позиций в ССТФ (см. Рисунок 5). Таким образом, она не учитывает зависимости между разными позициями [24, 25], хотя уже давно известно и экспериментально показано, что такие зависимости существуют [23, 43]. Например, если выравнивание ССТФ представляется только тетрануклеотидами GATC или GTAC, то по их выравниванию, построенная PWM будет давать равные высокие значения функции распознавания и для тетрануклеотидов GAAC и GTTC, которые отсутствовали в исходных данных. Существуют различные особенности связывания ТФ с ДНК, которые модель PWM не учитывает, такие как разнообразие структурных типов ССТФ [26, 27], нуклеотидный состав флангов ССТФ [142–145], СС ТФ-гетеродимеров [26], а также конформационные и физико-химические свойства ДНК [124, 125, 127]. Из-за этого PWM не способна находить все ССТФ в данных ChIP-seq; в среднем примерно только в половине пиков данная модель находит сайты [30–32].

Для того чтобы преодолеть этот недостаток, были предложены альтернативные модели мотива ССТФ, которые используют, например, разные вариации Марковских моделей, информацию о конформационных свойствах ДНК, и другие подходы [31, 37–41]. Все это позволяет тем или иным

образом учитывать зависимости между частотами встреч нуклеотидов в ССТФ.

1.4.2.1 Марковские модели

Одним из способов, который позволяет учесть зависимости между частотами встреч нуклеотидов в разных позициях ССТФ, является применение Марковской модели (ММ), где вероятность появления каждого нуклеотида зависит от нескольких предыдущих. Существуют разные варианты реализации таких моделей. Так, в ММ вероятность появления текущего нуклеотида b_j (j – текущая позиция, b – нуклеотид) зависит от k (k – порядок цепи ММ) предыдущих нуклеотидов b_{j-k} ... b_{j-1}. Так же стоит отметить, что ММ бывает гомогенной и негомогенной. Определяется это следующим образом. Для гомогенной ММ справедливо следующее равенство: $\Pr(X_{n+1} = x|\ X_n = y) = \Pr(X_n = x|\ X_{n-1} = y)$, то есть вероятность перехода одного состояния в другое не зависит от n (в нашем случае от позиции нуклеотида). В случае для негомогенной ММ $\Pr(X_{n+1} = x|\ X_n = y) \neq \Pr(X_n = x|\ X_{n-1} = y)$, то есть вероятность перехода одного состояния в другое будет зависеть от позиции. Это означает, что, например, вероятность встретить в мотиве нуклеотид А после нуклеотида С во второй позиции не равна вероятности встретить нуклеотид А после нуклеотида С и в третьей позиции.

Самой простой реализацией ММ является динуклетидная позиционная весовая матрица (diPWM) [146, 147]. Один из вариантов программной реализации модели diPWM для *de novo* поиска ССТФ – это diChIPMunk [148]. diPWM учитывает зависимости соседних нуклеотидов и является негомогенной ММ первого порядка [39].

Для того чтобы учитывать зависимости на расстоянии больше, чем в один нуклеотид, необходимо увеличить порядок ММ, однако с ростом порядка ММ, количество параметров модели растёт экспоненциально [39]. Большое количество параметров модели может приводить к её переобучению, то есть модель учитывает огромное число особенностей выборки обучения,

даже те, которые возникли по случайным причинам, однако данные особенности практически отсутствуют в предполагаемой генеральной совокупности, и в итоге точность модели падает.

В модели *Bayesian Markov Model* (BaMM) применили подход с использованием Байесовской статистики (вероятности), для того чтобы избежать переобучения модели [39]. Центральная идея обучения модели BaMM состоит в том, что для определения вероятности *k*-го порядка $p_i^{(k)} = (b_i|b_{i-k:i-1})$ используется вероятность $(k-1)$ порядка $p_i^{(k-1)} = (b_i|b_{i-k:i-1})$ в качестве априорной информации (Рисунок 7), где b – нуклеотид, k – порядок цепи, i – позиция в ССТФ. Когда все вероятности посчитаны, переходят от вероятностных значений к весовым с помощью *log-odds* преобразования, аналогично, как и для модели PWM, согласно формуле:

$$W_{b,i} = log_2 \frac{p_i^{(k)}(b_i|b_{i-k:i-1})}{p_{bg}^{(k`)}\left(b_j\middle|b_{i-k`:i-1}\right)} (3),$$

где p_{bg} – это Байесовская вероятность k`-го порядка, вычисленная по негативной выборке. Далее по аналогии с моделью мотива PWM рассчитывают значение функции распознавания:

$$S = \sum_{i=1}^{w} W_{b,i} \qquad (4)$$

где l_i – нуклеотид в позиции i заданной последовательности нуклеотидов, W – длина мотива.

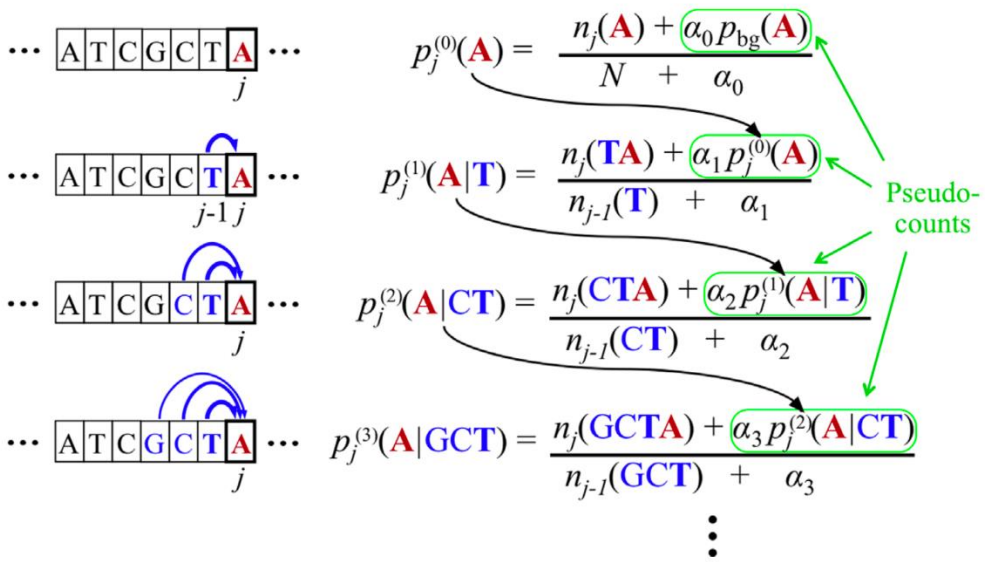

Рисунок 7. Пример расчета параметров модели BaMM [39]

В сравнении с моделью PWM, модель BaMM имеет систематически более высокую точность [34, 39, 149]. Помимо этого, для BaMM, как и для PWM, была показана зависимость между значением функции распознавания и свободной энергией Гиббса при связывании ТФ с СС. Также, показано, что BaMM имеет преимущества над PWM в предсказывании низкоаффинных ССТФ, которые значительно отличаются от консенсуса [39].

Другой реализацией негомогенной ММ является InMoDe [40]. Чтобы избежать проблемы переобучения и избытка параметров модели, авторы InMoDe добавили в модель подход “скупого” контекстного дерева (англ. parsimonious context tree, PCT) [42, 150]. Если в случае негомогенной ММ для каждой позиции и нуклеотида нужно хранить параметры о предыдущих нуклеотидах на k шагов в зависимости от порядка цепи для всех возможных комбинаций, то PCT, представляя все исходы в виде графа, может объединять вершины если значения в них значимо не отличаются, тем самым уменьшая пространство параметров (Рисунок 8) [42, 150]. Модель InMoDe так же, как BaMM, показала прирост точности в распознавании ССТФ в сравнении с PWM [42].

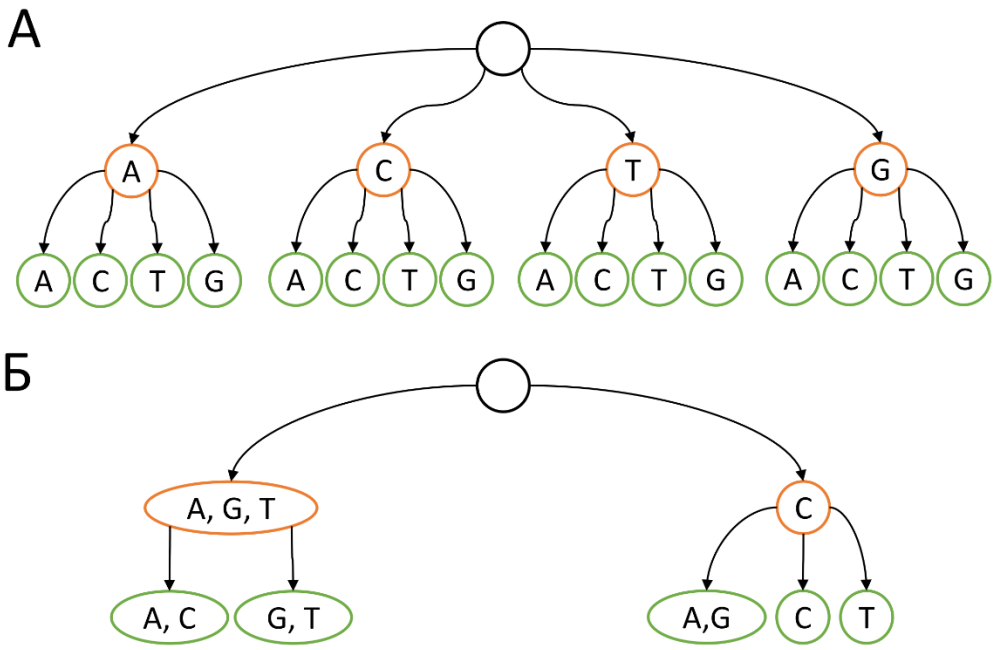

Рисунок 8. Пример контекстного дерева для одной позиции в ССТФ при порядке ММ равной 2 [42, 150]. **(А)** полное контекстное дерева; **(Б)** “скупое” контекстное дерево.

1.4.2.2 Модель, учитывающая структурные особенности ДНК

Связывание ТФ с ДНК является физико-химическим процессом и узнавание ТФ своего сайта основано не только на последовательности нуклеотидов, а также на узнавании структуры двойной спирали ДНК. В связи с этим были предложены модели, которые это учитывают [41, 127, 151–153].

Все такие модели описывают сайт как вектор, включающий разные структурные особенности ДНК. Например, выделяют 12 параметров, которые описывают структурную конформацию ДНК исходя из свойств двух соседних пар оснований [123–127]. Шесть из них отвечают за взаимное расположение пар оснований – shift, slide, rise, helix twist, tilt и roll (Рисунок 9А). Остальные определяют конформацию внутри пары оснований – shear, stretch, stagger, buckle, propeller twist и opening (Рисунок 9Б). Однако существует множество других параметров, например, ширина бороздки ДНК, температура плавления ДНК, изменение энтальпии и другие, которые также можно учитывать при описании физико-химических свойств ДНК [154].

Классической моделью, учитывающей структурные и физические особенности ДНК, является SITEVIDEO, которая использовалась в системе B-DNA VIDEO [127]. Модель SITEVIDEO направлена на выявление в ССТФ таких локальных участков, для которых средние значения конформационных или физико-химических свойств ДНК значимо отличаются от величин, характерных для случайных последовательностей. Однако, в настоящее время данную модель невозможно применять к ChIP-seq данным, так как не существует её адаптации для *de novo* поиска мотивов.

В известных на сегодняшний день *de novo* моделях мотивов, использующих структурный подход, из всего спектра параметров используют только roll, propeller twist, helix twist, а также учитывают ширину малой бороздки (minor groove width, MGW) ДНК [41, 151–153]. Перечисленные параметры вычисляют с помощью симуляции Монте-Карло на разных длинах нуклеотидных последовательностей с помощью программы DNAshapeR [155].

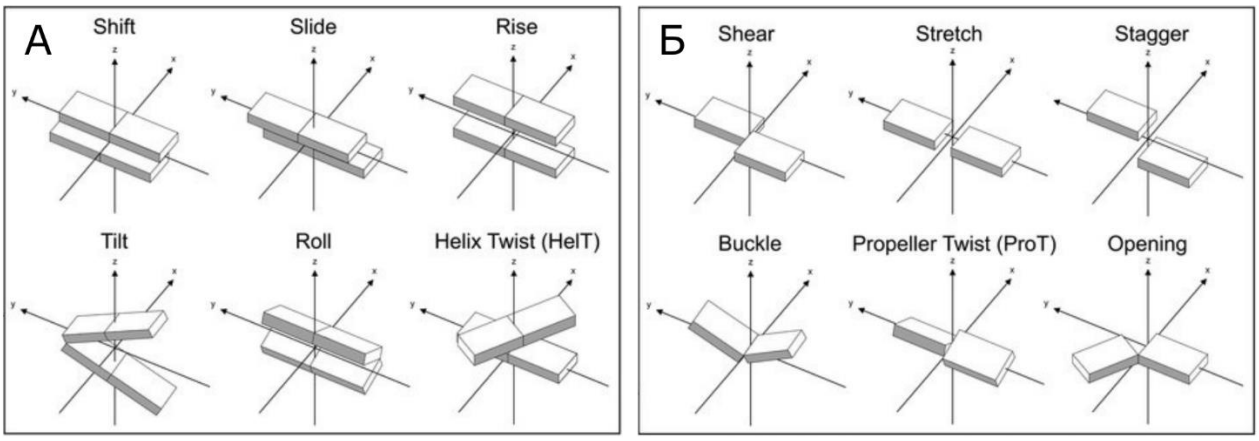

Рисунок 9. Структурные особенности ДНК. **(А)** взаимное расположение пар оснований; **(Б)** конформация пар оснований.

Один из подходов в построении модели с учётом структурных особенностей ДНК основан на применении готового выравнивания ССТФ с помощью известных мотивов в виде PFM [152]. Однако использование готового выравнивания, полученного с помощью традиционной модели PWM, для обучения другой модели, учитывающей структурные особенности ДНК,

имеет существенный недостаток, так как при построении выравнивания ССТФ с помощью модели PWM особенности структуры ДНК не учитывалось, то есть часть информации о структуре ДНК могла быть утеряна. Другой подход отличается лишь тем, что вместо построения выравнивания на основе известного мотива в рамках модели PWM, применяется алгоритм Gibbs sampling [129], с помощью которого строят выравнивание по одному из параметров, описывающих структуру ДНК, и на основании полученного выравнивания строят вектор со средними значениями выбранного параметра, который и используют для поиска мотивов [41]. Для выравненных ССТФ определяют структурные особенности и строят вектор, где каждый элемент для заданной позиции имеет усредненное значение одного из параметров. Далее можно взять любую случайную последовательность, определить для неё структурные параметры, построить вектор, и вычислить расстояние между полученным вектором и вектором модели мотива. Чем меньше будет рассчитанное расстояние, тем вероятнее, что анализируемая последовательность является ССТФ [41, 152].

Использование моделей, которые учитывают структуру ДНК, позволяет более точно предсказывать ССТФ относительно PWM [151, 152]. Применение таких моделей, показало, что помимо улучшения точности, также можно находить новые ССТФ, которые модель PWM не находит. Модели, построенные на учёте структурных особенностей ДНК, как и ММ, используют нуклеотидные зависимости [41].

1.4.2.3 Модели на основе дискриминантного анализа

Дискриминантный анализ (ДА) – это метод статистического анализа, который позволяет разделять наборы данных (классифицировать данные) на основе линейной комбинации признаков. В работах Занга [156] и Соловьёва [157] впервые применили метод ДА к нуклеотидным последовательностям с целью выявления регуляторных последовательностей, в частности, промоторов генов. Обычно при работе с нуклеотидными

последовательностями в качестве признаков используют частоты олигонуклеотидов. Так в работе [156] для ДА признаки рассчитывались через пентануклеотидные частоты, нормированные по отношению к их значениям в соседних районах промоторов (Рисунок 10).

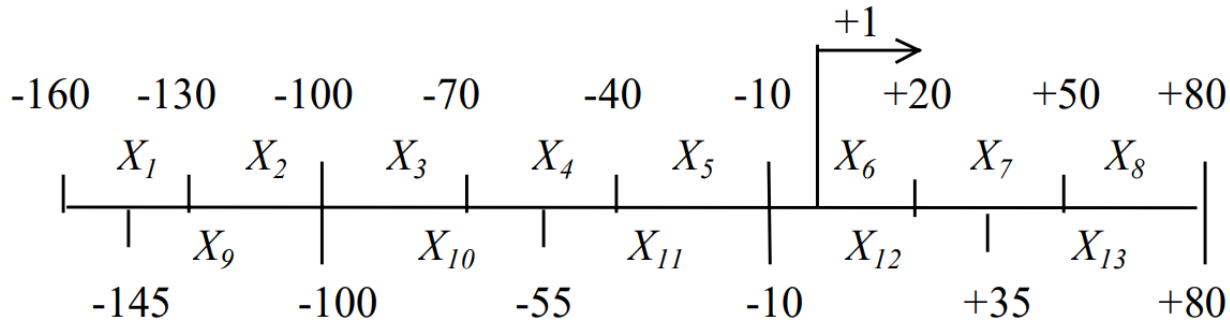

Рисунок 10. Разбиение промоторного района на зоны [156]. Для промоторного района [-160; +80] использованы два набора неперекрывающихся окон – размером 30 и 45 п.о. Стрелкой обозначено положение старта транскрипции.

Величина f_i для *i*-ой переменной вычислялась через частоты f(x) пентануклеотидов для районов номер i-1, i, i+1:

$$f_i = \frac{f(x_i)}{f(x_i)+\frac{f(x_{i-1})+f(x_{i+1})}{2}} \quad (5)$$

Дальнейшем развитием методов ДА для поиска регуляторных элементов была модель SiteGA, которая, в отличие от предыдущих методов, описывает ССТФ, а не промоторы. Модель SiteGA учитывает зависимости между частотами встреч локально-позиционированных динуклеотидов (ЛПД) в ССТФ [46].

Модель SiteGA при обучении осуществляет поиск наиболее оптимального для распознавания ССТФ набора ЛПД. Данная задача решается с помощью генетического алгоритма (ГА), в котором популяция особей, представляет собой наборы ЛПД. Каждый ЛПД особи характеризуется положением $[a, b]$ в пределах всего сайта $[A, B]$, а также типом d_j динуклеотида $(j=1, \ldots, 16)$. Для границ a и b возможной локализации динуклеотида используются позиции его первого нуклеотида, при этом всегда выполняется условие $A \leq a \leq b \leq B-1$. В ходе работы ГА

максимизируется расстояние *D(X)* Махаланобиса, которое является мерой приспособленности особи, рассчитываемое по следующей формуле:

$$D(X) = \sum_{k=1}^{N} \sum_{n=1}^{N} \left(\left[f_n^{(1)}(X) - f_n^{(2)} \right] * S_{n,k}^{-1} * \left[f_k^{(1)}(X) - f_k^{(2)} \right] \right) \quad (6)$$

Здесь X – вектор частот ЛПД, N – общее число ЛПД в текущем наборе, $f_n^{(1)}$ и $f_n^{(2)}$ – средние частоты n-го ЛПД по последовательностям позитивной и негативной выборок, соответственно; S^{-1} – обратная матрица объединённой ковариационной матрицы, которая равна сумме ковариационных матриц для выборок ССТФ и случайных последовательностей по частотам ЛПД. Результатом работы ГА является конкретный набор ЛПД, используемых для построения линейной функции распознавания по частотам ЛПД [46].

Недавно модель SiteGA была адаптирована для работы с ChIP-seq данными, то есть был разработан *de novo* подход с применением модели SiteGA [34]. Реализация модели SiteGA для *de novo* поиска максимизирует произведение расстояния Махаланобиса *D(X)* с консервативностью мотива, оцененной как кратность обогащения частот k-меров заданной длины в позитивной выборке по сравнению с негативной – *E(X)*, итоговая целевая функция максимизации имеет следующий вид:

$$F(X) = D(X) * E(X) \quad (7)$$

Фактор *E(X)* означает среднюю кратность обогащения k-меров определенной длины Z в пределах выравнивания в позитивной выборке по сравнению с негативной выборкой [34].

1.4.3 Оценка точности поиска мотивов разными моделями

Поиск мотивов является задачей классификации, поэтому все стандартные методы статистической оценки, которые используются для задач классификации, также применимы и для поиска мотивов.

После того как модель определила, является ли данная последовательность нуклеотидов сайтом или нет, полученный результат можно отнести к одному из четырёх исходов (Таблица 5). В случае если в

реальности последовательность является сайтом и модель её классифицировала как сайт, то такой исход является *true-positive* (TP) – верно предсказанный сайт. Если в реальности последовательность не является сайтом, и модель классифицирует её как «не-сайт», то такой исход является *true negative* (TN) – верно предсказанный нефункциональный сайт. Если в реальности последовательность не является сайтом, а модель классифицирует её как сайт, то такой исход является *false positive* (FP) – неверно предсказанный сайт. Четвёртый исход это *false negative* (FN) – модель определила последовательность как «не-сайт», хотя в реальности он сайтом является [158, 159].

Таблица 5. Возможные исходы классификации сайт.

		Результат модели	
		сайт	не-сайт
Реальность	Сайт	TP	FP
	не-сайт	FN	TN

Стандартным подходом оценки точности предсказания являются ошибки 1-го и 2-го рода (E_1 и E_2, соответственно). Ошибка 1-го рода (E_1) это недопредсказание – доля сайтов, которые модель не распознала, ошибка 2-го рода (E_2) это перепредсказание – доля последовательностей, которые в реальности не являются сайтами, но модель их распознала как сайт [158].

$$E_1 = \frac{FN}{FN + TN}$$

$$E_2 = \frac{FP}{TP + FP}$$

Визуализировать зависимость этих ошибок друг от друга можно путём построения кривой, которая называется рабочая характеристика приёмника (*receiver operating characteristic*, ROC-*curve*, ROC-кривая) [159], пример этой кривой для трёх моделей на синтетических данных показан на рисунке 11.

Кривая ROC представляет из себя зависимость true positive rate (TPR) – доли верно предсказанных сайтов, к *false positive rate* (FPR) – доле неверно предсказанных сайтов. Сравнивать ROC-кривые можно и визуально, как видно из графика (Рисунок 11), «модель 2» лучше, чем «модель 3», а «модель 1» лучше, чем «модель 2». Однако визуально сравнивать ROC-кривые не всегда эффективно. Поэтому для количественной оценки кривой ошибок используют показатель площадь под ROC-кривой (*area under* ROC-*curve*, AUC). Чем больше значение AUC, тем точнее модель предсказывает сайт. Иногда вместо AUC используют показатель *partial*-AUC (pAUC), посчитанный для значений FPR меньше заданного, что позволяет учитывать лишь критически важный для распознавания диапазон FPR [159]. С помощью ROC-кривой и AUC чаще всего сравнивают точность разных моделей мотивов [25, 39, 42, 152].

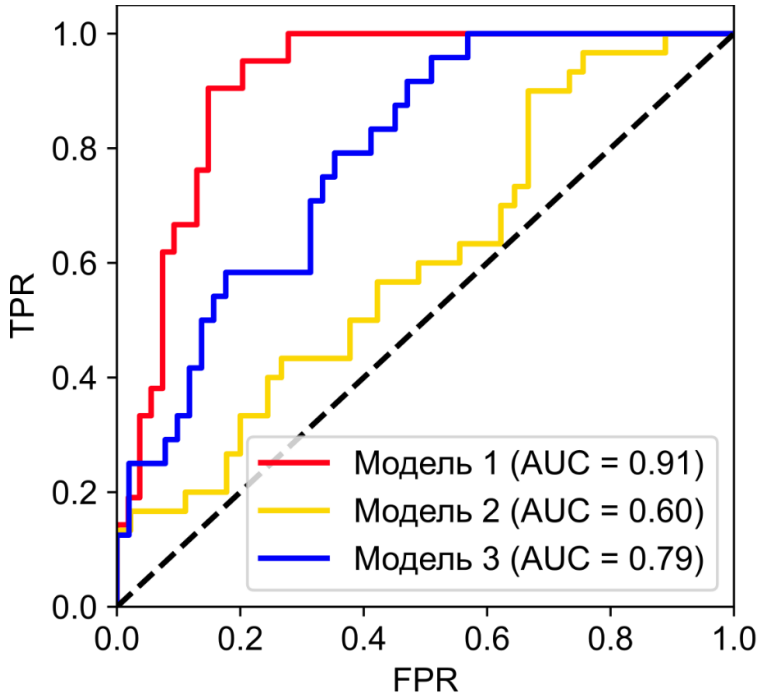

Рисунок 11. Пример ROC-кривой для трёх моделей на синтетических данных. Ось абсцисс – FPR, доля неверно предсказанных сайтов; ось ординат – TPR, доля верно предсказанных сайтов.

Для того, чтобы корректно построить ROC-кривую необходимо иметь (1) позитивную выборку (выборку обучения), на которой модель обучается; (2) контрольную (тестовую) выборку, которая содержит сайты и, используя её, определяются значения TPR; (3) негативную выборку, в которой отсутствует обогащение сайтов и, используя её, определяются значения FPR. Обычно исследователи имеют только одну позитивную выборку в виде исходных данных ChIP-seq, поэтому для оценки точности моделей используют разные методы «перегруппировки» данных позитивной выборки, которые позволяют разделить её на подвыборки обучения и контроля. Наиболее широко применяемым методом «перегруппировки» является перекрёстная проверка (*cross-validation*, CV) [160].

Существует несколько реализаций CV, один из вариантов это *leave-p-out cross-validation* (LpO CV), в котором из общей выборки убирается *p* последовательностей, которые используются для оценки TPR (тестовая выборка), а оставшиеся сайты используют для обучения модели (выборка обучения). В данной реализации CV количество раз, на которые необходимо будет поделить выборку, чтобы учесть все последовательности, определяется, как биноминальный коэффициент, то есть количество сочетаний C_p^n, где n – это общее количество последовательностей в исходной выборке. Частным случаем LpO CV является *leave-one-out cross-validation* (LOOCV), когда *p* равно единице [160].

Другой реализацией является *k-fold* CV, где исходная выборка разделяется заданное количество раз (*k* раз) на выборку обучения и тестовую выборку. Например, если у нас есть исходная выборка с 15 последовательностями, а k=3, то размер тестовой выборки составит $n/k = 15/3 = 5$, а обучающей 10. Далее из исходной выборки последовательно используются первые 5 тестовых последовательностей, а на остальных последовательностях модель обучается, далее используются следующие 5 последовательностей в качестве выборки, а на последнем шаге последние пять последовательностей. Стоит отметить, что перед k-fold CV начальный порядок

последовательностей может быть изменен путем случайных перестановок [160]. Простейшим примером k-fold реализации является разбиение данных на две части [148], что и будет реализовано в представляемой работе.

Еще одной реализацией является Монте Карло CV, данный метод похож на k-fold CV, в нем также выбирается количество итераций разделения выборки. Однако тестовые последовательности каждый раз выбираются случайно из исходного набора, из-за чего одна и так же последовательность может несколько раз участвовать в тестовой выборке [160].

При оценке точностей моделей мотива на данных ChIP-seq обычно в качестве выборок обучения, теста и контроля используют ChIP-seq пики, то есть оценивается наличие или отсутствие мотива в пике [45, 148, 149, 161]. Однако такой подход может порождать некоторые проблемы с расчётом FPR: (1) не корректность сравнения нескольких наборов данных, так как распределения длин пиков (негативной выборки) в разных наборах данных могут отличаться, а вероятность встретить как минимум один мотив в пике по случайным причинам зависит от длины пика; (2) величина FPR как доля пиков несёт меньше смысловой нагрузки, чем вероятность мотива, с точки зрения распознавания сайтов в геноме. Чтобы избежать данных проблем можно величину FPR оценивать по вероятности встретить мотив в негативной выборке [34, 131].

Применяя разные варианты CV можно оценить значения FPR и TPR по исходным данным и модели, а также построить ROC-кривую.

1.5 Структурное разнообразие ССТФ

В действительности для одного ТФ может наблюдаться широкое разнообразие мотивов, с которыми связывается данный ТФ [162]. Данное явление называют гетерогенностью, то есть один и тот же ТФ может иметь разные структурные типы СС [163, 164].

Данную особенность можно выделить для ТФ семейства Forkhead box (FOX) factors {3.3.1} [165], которые могут связываться как с каноничным

консенсусом RYAAAYA (R = G/A, Y = T/C), так и с альтернативным GACGC [165]. В частности, это было экспериментально подтверждено для ТФ FoxN3, который способен связываться с двумя разными консенсусами, при этом сам ТФ никаким внешним модификациям не подвергается [27].

Другим примером может служить ТФ FOXP1 из этого же семейства FOX factors {3.3.1}, он также может связываться с разными мотивами, правда в его случае это достигается за счет альтернативного сплайсинга гена FOXP1. Одна изоформа ТФ FOXP1 связывается с каноничным консенсусом GTAAACA, и отвечает за активацию экспрессии генов необходимых для дифференцировки клеток, другая изоформа FOXP1 связывается с CGATACA и AACAACAA, и отвечает за активацию экспрессии генов связанных с плюрипотентностью клеток [166].

Ещё в одном исследовании ТФ из семейства FOX factors {3.3.1} [167] показали, что FOXC2 помимо классического для семейства ТФ FOX консенсус GTAAACA, так же распознает GTACACA. Для других ТФ из семейства FOX, а именно FOXA2 и FOXM1, было показано, что они также связываются с мотивами GTAAACA и GTACACA. Помимо этого, для FOXC2, FOXA2, FOXM1 показали очень разные аффинности связывания с сайтом ACAAATA: (1) FOXC2 с трудом может связываться с сайтом ACAAATA; (2) FOXA2 все еще может связываться с сайтом ACAAATA, но с гораздо более низкой аффинностью, чем с сайтом GTAAACA; (3) FOXM1 может связываться с сайтом ACAAATA с аффинностью, аналогичной аффинности для сайта GTAAACA. В итоге предположили, что для FOXC2 каноническим мотивом является RYAMACA (R = G/A, Y = T/C, M = A/C) [167].

ТФ HOXB13 и CDX2 из семейства HOX-related factors {3.1.1}, также способны связываться с разными консенсусами: CAATAAA и TCGTAAA, с одинаковой аффинностью. В данном случае гетерогенность ССТФ объясняется влиянием изменений энтропии и энтальпии на изменение свободной энергии Гиббса. При связывании этих ТФ с разными консенсусами свободная энергия Гиббса изменяется одинаково, однако достигается это за

счет разных вкладов энтропии и энтальпии. Так для CAATAAA больший вклад в изменение энергии Гиббса вносит изменение энтальпии, а для TCGTAAA изменение энтропии [168].

Причиной гетерогенности ССТФ, оцениваемой по данным их массового секвенирования на основе картирования *in vivo* (ChIP-seq) может служить димерезация ТФ. Часто ТФ образуют гомо- и гетеродимеры, и в таком случае ССТФ для димера является комбинацией двух «полу-сайтов». Например, для гомодимера MEIS1-MEIS1 консенсус – TGACANNTGTCA (инвертированный повтор гексамера TGACAN без спейсера), а для гетеродимера MEIS1-DLX3 – TGACANNNNAATTG [26]. Некоторые ТФ связываются с ДНК только в виде димеров, например, ТФ, которые относятся к классам *Basic leucine zipper factors* (bZIP) {1.1} и *Basic helix-loop-helix factors* (bHLH) {1.2} [169, 170]. При этом они могут формировать как гомодимеры, так и гетеродимеры с представителями своего класса – bHLH-bHLH, bZIP-bZIP. Сайты связывания таких гетеродимеров не всегда являются суммой полусайтов мономеров, входящих в состав гетеродимера, что порождает структурное разнообразие СС. Одна из причин данного явления – это белок-белковые взаимодействия, из-за которых структура ДСД одного из мономеров может меняться под влиянием другого [169, 170].

Другим важным аспектом наличия гетерогенности ССТФ может быть тот факт, что у эукариот в среднем количество ДСД на один ТФ увеличилось по сравнению с прокариотами (1.44 и 1.04, соответственно). Более того, если рассматривать многоклеточных животных, то у них в среднем приходится 2.75 ДСД на один ТФ [171]. При этом количество отдельных семейств ДСД на один ТФ для эукариот и прокариот сохранилось на прежнем уровне (1.01 и 1.0, соответственно) [171].

Одним из объяснений того, что ТФ имеют СС с существенно отличающейся структурой, является то, что ТФ считывают не последовательность нуклеотидов, а распознают структуру самого сайта, с которым связываются. Ведь в действительности связывание ТФ с ДНК с

физико-химической точки зрения – это взаимодействие между ДСД ТФ и участком ДНК (ССТФ) за счёт водородных связей между аминокислотными остатками белка и нуклеотидами ДНК [79, 172].

Однако, с одной стороны, необходима определённая последовательность нуклеотидов в ССТФ, при которой достигается нужная трёхмерная структура ДНК, необходимая для связывания с ДСД ТФ, при которой достигаются оптимальные условия для формирования водородных связей [173]. Так, наиболее широко применяется метод предсказания структурных особенностей ДНК, основанный на применении симуляции Монте-Карло для нуклеотидных последовательностей разной длины [155]. Данный метод широко используется при разработке моделей поиска ССТФ с использованием структурных особенностей ДНК [38, 41, 174].

С другой стороны, в ряде исследований было показано, что существует зависимость между нуклеотидами в разных позициях ССТФ и при этом такая зависимость не всегда объясняется структурными особенностями ДНК [37, 42, 175, 176]. Из этого следует, что наличие разных типов ССТФ, не всегда объясняется пространственной структурой ДНК. Тем не менее данный аспект является важным для распознавания ТФ своего сайта [38, 41, 174].

На способность ТФ связываться с ДНК может влиять не только СС, но и его окружение. Существует ряд факторов, которые оказывают влияние на связывание ТФ с СС, а именно нуклеотидный состав с флангов СС, структура ДНК флангов СС, наличие корреляции встречаемости нуклеотидов на разных расстояниях друг от друга, наличие вырожденных СС целевого ТФ во флангах [142, 177].

В работе [142] было показано, что разные ТФ предпочитают связываться с сайтом в том случае, когда нуклеотидный состав флангов схож с нуклеотидным составом сайта. Однако этого не всегда достаточно, и в действительности существуют некоторые закономерности по встречаемости нуклеотидов во флангах сайтов [177, 178]. В работе [177] было показано, что ТФ слабо связываются с СС, которые имеют фланги такого же нуклеотидного

состава, как и природные СС, но включающие случайно подобранные нуклеотиды. Однако, если найти зависимости между частотами встреч нуклеотидов, в виде корреляций, и сгенерировать фланги мотива на основе полученных корреляций, то экспрессия гена, который регулируется этим ТФ, с искусственными флангами и СС не будет отличаться от такого же СС с природными флангами [177]. На связывание ТФ с мотивом влияет наличие вырожденных сайтов этого же ТФ во флангах. На примере ТФ MITF было экспериментально показано, что присутствие во флангах вырожденных сайтов способствует связыванию этого ТФ со своим сайтом [142]. Значительный вклад флангов, которые оказывают влияние на связывание ТФ с сайтом, подтверждается тем, что только на основе особенностей флангов, можно построить модели предсказывающие локусы связывания ТФ [142].

Сильная вырожденность регуляторного кода транскрипции имеет глубокий биологический смысл, позволяя различным ССТФ располагаться в пределах ограниченной области с наложением друг на друга, что, с одной стороны, обеспечивает высокую плотность кодирования регуляторной информации, а с другой стороны, создает возможность для формирования очень тонких и специфичных механизмов регуляции транскрипции [77]. Так в недавнем массовом исследовании мотивов ТФ [179] с применением экспериментального метода HT-SELEX (англ. high-throughput systematic evolution of ligands by exponential enrichment) [180] было показано, что около 11% ТФ (19 из 170 ТФ, исследуемых в работе) распознают неканонический мотив, который значительно отличается от канонического мотива, с которым обычно связывается ТФ и который представлен в CIS-BP. В этой же работе авторы подтвердили, что большинство неканонических мотивов возникали как специфические паттерны для ТФ мономеров/димеров или вариаций фланкирующих последовательностей мотива, а также как вариаций канонических мотивов. Возникновение неканонических СС в ДНК происходило абсолютно независимо от канонических СС, и они играют такую же важную биологическую роль, как и канонические СС [179].

1.6 Биоинформатический анализ данных полученных ChIP-seq экспериментом

Для изучения ТФ и их ССТФ *in vivo* используют полногеномный экспериментальный метод ChIP-seq (Chromatin immunoprecipitation sequencing, секвенирование с иммунопреципитацией хроматина). ChIP-seq — метод анализа ДНК-белковых взаимодействий, основанный на иммунопреципитации хроматина (ChIP) и высокоэффективном секвенировании ДНК. Цель этого эксперимента состоит в том, чтобы картировать СС целевого белка с максимальным отношением сигнал-шум во всём геноме [181]. Такие данные накапливается в рамках выполнения таких проектов как ENCODE (https://www.encodeproject.org/) [182], ReMap (http://remap.univ-amu.fr/) [183], Cistrome (http://cistrome.org/db/#/) [184], GTRD (http://gtrd.biouml.org/#!) [185]. Схема проведения эксперимента ChIP-seq представлена на рисунке 12 [181].

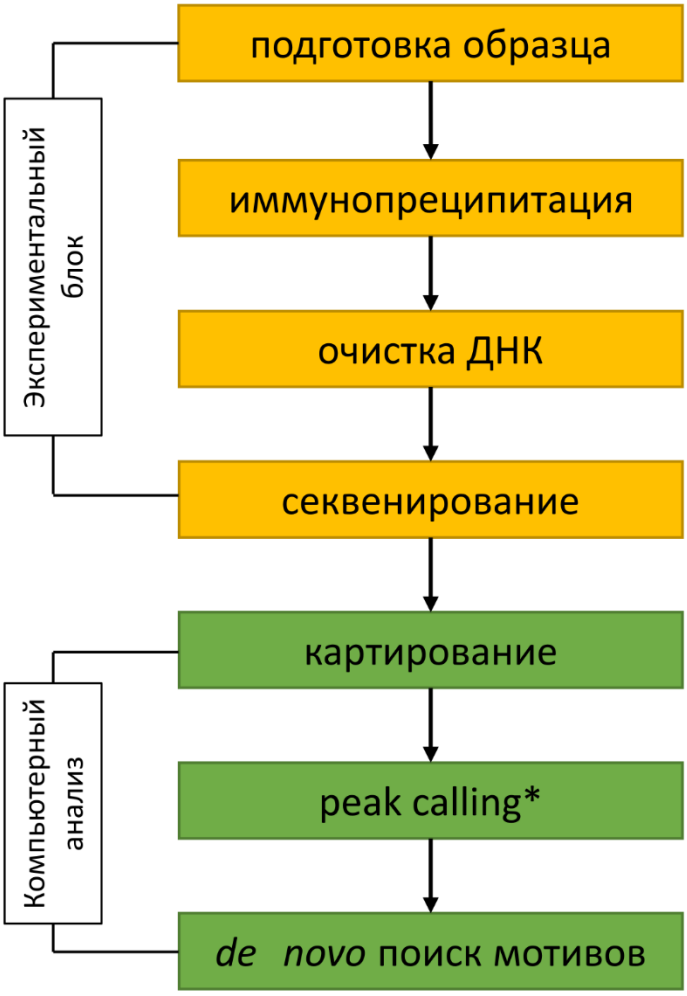

Рисунок 12. Схемы проведения эксперимента ChIP-seq и компьютерной обработки его результатов. **peak calling* –определение локусов, обогащенных прочтениями.

Эксперимент ChIP-seq проводят следующим образом. На первом этапе клетки или ткани обрабатывают химическим агентом, обычно формальдегидом, для ковалентного сшивания белков с ДНК. Затем разрушают клетки при помощи обработки ультразвуком или, в некоторых случаях, ферментативного расщепления, для получения участков ДНК размером порядка 100-300 пар оснований. Далее интересующий белок (ТФ, структурный белок хроматина, гистон, РНК-полимераза и т. п.), связанный со своим участком ДНК, обогащают по отношению к исходному хроматину за счет отчистки антителами, специфичным для данного белка. Обогащенная ДНК очищается и подготавливается для секвенирования последовательностей ДНК, связанных с белком. Результатом работы является файл формата fastq с

прочтениями [181]. После экспериментальной части следует биоинформатический анализ полученных данных, который можно разбить на две части: первичная обработка данных и вторичная обработка данных. В первичную обработку входит проверка качества прочтений, картирование прочтений на референсный геном, получение пиков (peak calling) [186]. Вторичная обработка данных включает, прежде всего, *de novo* поиск мотивов, например, построение позиционной весовой матрицы (PWM), которая является важным инструментом для компьютерных исследований регуляции транскрипции, поиска *цис*-регуляторных элементов и отдельных ССТФ, моделирования генных сетей и аннотации геномной последовательности [8].

1.6.1 Первичная обработка данных ChIP-seq

Перед тем как проводить картирование прочтений на геном, необходимо провести оценку качества полученных прочтений. Обычно оценка качества осуществляется по нескольким параметрам: 1) качество прочтений, 2) G/C состав, 3) присутствие адаптеров (технические последовательности, используемые при секвенировании), 4) обогащение k-мерами и повторами [187]. Для анализа этих параметров существует множество средств, как например FastQC, NGSQC [188], однако данные программы только дают оценку качества данных, но никак не обрабатывают сырые данные. Для этих целей существует ряд программ таких как FASTX-Toolkit, Trimmomatic [189] и dada2 [190], которые позволяют убирать из библиотеки прочтения низкого качества и удалять из прочтений адаптеры [191].

После того как библиотека прочтений прошла проверку на качество, необходимо провести картирование. Для этих целей существует множество программ, как например Bowtie2 [192], BWA [193] или SOAP2 [194]. Результатом данного этапа будет sam или bam файл, который содержит прочтения, выравненные на геномную ДНК.

Для идентификации пиков из геномного профиля ChIP-seq, которые соответствуют местам связывания ТФ, было разработано множество

алгоритмов [195]. Данные ChIP-seq могут содержать три разных типа пиков: 1) короткие области длиной в несколько сотен пар оснований или меньше, 2) более широкие регионы, содержащие до нескольких тысяч пар оснований (п.о.), 3) широкие области, содержащие до нескольких сотен тысяч п.о. Различные типы пиков связаны с разными типами белков, связанными с ДНК. Например, точечная область характерна для многих ТФ. Комбинация точечных и более широких областей характерна для белков, таких как РНК-полимераза II. Широкие области характерны для гистонов и других структурных белков хроматина [195].

Различные инструменты поиска пиков основаны на разных алгоритмах, которые лучше подходят для поиска только конкретного типа пиков. Таким образом, важно выбрать компьютерный инструмент поиска того типа геномных пиков, который изучается в эксперименте, что максимизирует вероятности получения наилучших возможных нуклеотидных последовательностей, соответствующих этим пикам, для поиска нуклеотидных мотивов связывания. К настоящему моменту уже разработано множество программных инструментов для поиска пиков, таких как MACS [196], GEM [197], PICS [198], SISSRS [199], peakROTS [200], mosaics-HMM [201] и BinQuasi [202]. При этом наиболее распространенными программами являются MACS и GEM, которые широко применяются как в отдельных исследованиях [9], так и при разработке баз данных с обработанными ChIP-seq данными [185]. Однако в сравнительном исследовании [203] было показано, что MACS превосходит GEM по чувствительности и точности. MACS применяется во всех основных базах данных, осуществляющих обработку ChIP-seq данных, таких как ReMap [183], CISTROME DB [184], ChIP-Atlas [204] и GTRD [185]. Стоит отметить, что MACS даёт информацию о реальной ширине пика, в отличие от GEM, который даёт информацию только о позиции самой высокой точки пика. Исходя из этого было предположено, что наборы пиков, полученные при массовом секвенировании ССТФ с помощью MACS

[205], более эффективны для последующего биоинформатического анализа, чем полученные с помощью GEM [9].

1.6.2 Вторичная обработка данных ChIP-seq – *de novo* поиск мотивов

Важной задачей в анализе ChIP-seq данных является поиск в наборе ChIP-seq пиков мотивов, которые представляют ССТФ. Для того, чтобы решить данную задачу, применяются специальные алгоритмы *de novo* поиска мотивов. Необходимо отметить, что здесь *de novo* означает, что мы не имеем никакой информации о мотиве, но мы допускаем, что искомый мотив обогащён в пиках [130]. Иногда вместо термина *de novo* используют термин *ab initio*, который имеет такой же смысл. К настоящему моменту разработано множество программ для *de novo* поиска мотивов, в основе которых используется разнообразные алгоритмы для обучения модели (Expectation maximization, Gibbs sampling и др.) [206]. Большинство программ в качестве модели мотива использует PWM несмотря на то, что уже около 20 лет известно о недостатках модели PWM [24, 43], и ряд альтернативных моделей показывал более высокую точность при распознавании мотивов по сравнению с PWM [25, 40, 148, 149]. Реализации *de novo* поиска мотивов, основанные на модели PWM, остаются наиболее широко используемыми. В качестве косвенного показателя популярности программ, которые используют PWM, можно привести количество цитирований статей, в которых обсуждаются конкретные программы *de novo* поиска мотивов. Так, на начало 2023 г. статьи, в которых применялась модель PWM в виде программ STREME [14], MEME/MEME-ChIP [15, 207], HOMER [13] и ChIPMunk [16], имеют суммарное количество цитирований около 8000, а статьи, посвященные альтернативным моделям BaMM [39, 208], InMoDe [40], Slim [25] и diChIPMunk [148] – чуть более 100 цитирований. 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; 3) наличие баз данных с готовыми матричными моделями, таких как HOCOMOCO [8], CIS-BP [140] и JASPAR [209].

В настоящее время разработано множество альтернативных моделей мотивов, однако главной задачей авторов моделей является достижение более высокой точности распознавания, чем у традиционной модели PWM [25, 39, 40, 210], а не сравнение структуры мотива разных моделей. Несмотря на любую, даже самую высокую точность распознавания, ни одна модель не решает проблему полного распознавания ССТФ в пиках ChIP-seq. Данная проблема частично обусловлена структурной гетерогенностью ССТФ для одного и того же ТФ, и число пиков, содержащих мотив, может быть значительно увеличено при одновременном использовании разных моделей [34]. В таком случае ChIP-seq пики будут содержать как мотивы, предсказываемые одновременно двумя и более моделями, так и мотивы, предсказываемые только одной из моделей [34, 44, 45, 211]. Ранее при анализе двух независимых экспериментов ChIP-seq для ТФ FOXA2 [212, 213] с применением модели PWM, реализованной в ChIPMunk, (*de novo*) и альтернативной модели SiteGA (по выборке обучения из 53 известных сайтов ТФ подсемейства FOXA) удалось обнаружить FOXA2 сайты более чем в 95% пиков [45], 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 моделью PWM для анализа ChIP-seq данных. Однако до сих пор не было массовых исследований ChIP-seq данных, где бы применяли для поиска ССТФ сразу несколько методологически разных моделей. Разнообразие структурных типов ССТФ изучено не полностью, до сих пор остаётся не ясно, как СС разных ТФ влияют на уровень транскрипции, и отличаются ли гены биологическими функциями, которые контролируются разными СС одного ТФ.

Не менее важным аспектом, не зависящим от модели мотива, в *de novo* поиске, является негативная выборка относительно которой считают обогащение мотива. В программах используются различные подходы к генерации негативной выборки, так, например, ChIPMunk использует некоторое математическое (статистическое) описание негативной выборки, связанное с частотой нуклеотидов. В MEME негативная выборка генерируется с помощью MM, а в STREME по умолчанию негативная выборка генерируется путём случайной перестановки нуклеотидов выборки обучения, но при этом для обеих программ можно самостоятельно задать негативную выборку. Так же для SiteGA [34] и BaMM [39] негативную выборку можно сгенерировать и задать в качестве параметра. В программе HOMER негативная выборка генерируется путём выбора случайных последовательностей из генома, имеющих схожий нуклеотидный состав с выборкой обучения. От негативной выборки существенно зависит результат, так как значимость обогащения считается относительно неё, и согласно работе [214] лучшим вариантом является выборка, сгенерированная на основе генома и имеющая схожий нуклеотидный состав (G/C состав) с выборкой обучения.

2. Методы

2.1 Используемые данные

Для анализа использовались предобработанные ChIP-seq данные в виде разметки пиков в формате bed из баз данных. Так с использованием базы ReMap (http://remap.univ-amu.fr/) [183] была сформирована небольшая выборка (тестовые данные) по ChIP-seq экспериментам по ТФ FOXA2 для тестирования программного комплекса, а с использованием GTRD (https://gtrd.biouml.org/#!) [185] была сформирована выборка данных для массового анализа ChIP-seq данных для *A. thaliana* и *M. musculus*.

2.2 Конвейер программ для выявления структурной гетерогенности ССТФ.

Разработан программный комплекс MultiDeNA (Multiple *De Novo* Analysis, https://github.com/ubercomrade/MultiDeNA) для совместного применения различных моделей *de novo* поиска мотивов в данных ChIP-seq. Программный комплекс позволяет использовать модели мотивов ChIPMunk (PWM), STREME (PWM), diChIPMunk (diPWM), BaMM, InMoDe и SiteGA, а также вспомогательные программы: bedtools [215], TomTom [141]. Принципиальная схема работы программного комплекса представлена на рисунке 13. Программный комплекс включает в себя следующие этапы: подготовка данных; оценка точности моделей и выбор оптимальных параметров; построение моделей; выбор порогов, поиск мотивов в пиках ChIP-seq с фиксированным порогом; классификация ChIP-seq пиков по результатам распознавания мотивов *de novo* моделями. Каждый этап работы программного комплекса детально описан ниже.

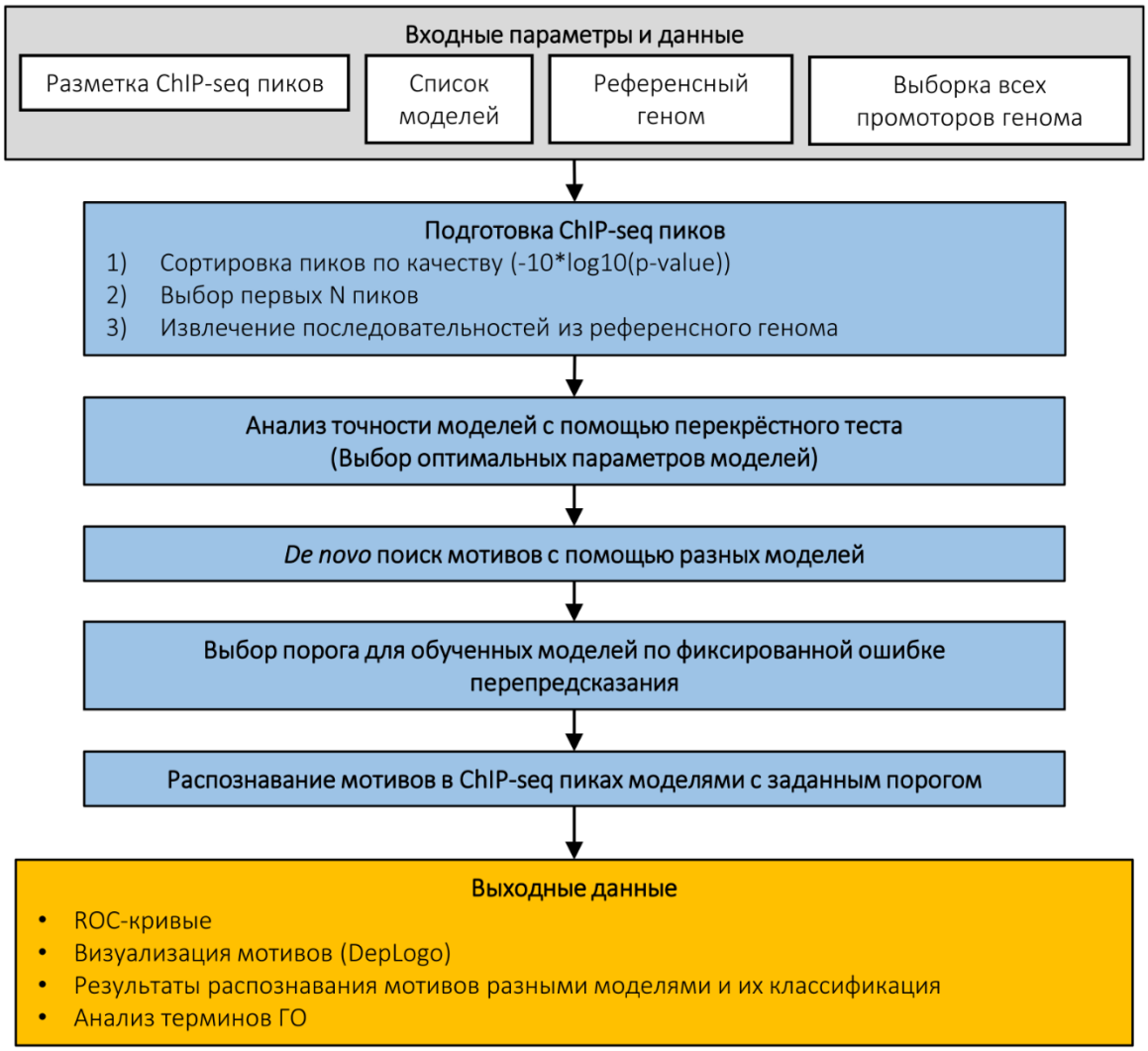

Рисунок 13. Принципиальная схема работы конвейера программ.

2.3 Подготовка данных ChIP-seq

Первым этапом подготовки данных является сортировка пиков по округленному значению $-10*\log_{10}(p\text{-value})$ – характеризует вероятность обогащения пика в иммунопреципитированном образце по сравнению с контролем, которое было ранее вычислено для каждого пика программой MACS2 [196]. Следующий этап это – выбор лучших N пиков для анализа. В данной работе N = 4000 при работе с тестовыми данными по ТФ FOXA2 и N = 1000 при массовом анализе ChIP-seq данных. Последний этап это – извлечение нуклеотидных последовательностей из генома по координатам пиков с помощью bedtools [215].

2.4 Выбор моделей и их параметров

Для того чтобы распознавать сайты в пиках, необходимо построить *de novo* модели мотива. Построение альтернативных моделей мотивов осуществлялось программами BaMM [39], InMoDe [40] и SiteGA [34, 46]; модель diPWM строили с помощью diChIPMunk [148], а модель PWM строили программой ChIPMunk при работе с тестовой выборкой, и STREME [14, 16] при массовом анализе данных ChIP-seq. Использование STREME вместо ChIPMunk при массовом анализе данных ChIP-seq обусловлено тем, что в STREME можно задать сгенерированную негативную выборку.

Чтобы улучшить точность распознавания ССТФ для каждой модели, подбирали оптимальные параметры. Оптимальность параметра подразумевает наивысшую точность распознавания по величине pAUC. Для PWM и diPWM единственный параметр, который можно варьировать это длина мотива, следовательно, для этих моделей подбирали оптимальную длину. Модели, построенные на ММ, помимо длины, в качестве параметра имеют порядок ММ, поэтому для BaMM и InMoDe оптимизировали данный параметр. Для модели SiteGA важным параметром является количество ЛПД. Для того, чтобы оценить точность моделей и подобрать оптимальные параметры, использовали метод оценки точности 2-*fold* CV, который применялся для разных параметров моделей: длина мотива от минимальной W_{min} (по умолчанию 8 п.о.) до максимальной W_{max} (40 п.о. – анализ данных ТФ FOXA2; 20 п.о. – массовый анализ) с шагом dW (2 п.о. при анализе данных ТФ FOXA2 и 4 п.о. при массовом анализе ChIP-seq данных); порядок ММ от 0 до 3 (для BaMM и InMoDe); количество ЛПД от 40 до 100 (для SiteGA). Метод оценки точности включал следующие этапы:

1) генерация негативной выборки последовательностей путем случайной перестановки нуклеотидов в последовательностях контрольной выборки, либо путём извлечения случайных участков из генома, имеющих схожий нуклеотидный состав и длину;

2) разделение данных ChIP-seq на подвыборки обучения и контроля – пики с нечётными и чётными номерами, соответственно (по половине от всех пиков);

3) построение модели на подвыборке обучения с использованием негативной выборки;

4) проверка модели на контрольной подвыборке для оценки TPR;

5) проверка модели на негативной выборке для оценки FPR, как величины встречаемости мотива в негативной выборке;

6) повторение этапов пунктов 1–5, где подвыборка обучения меняется с контрольной;

7) вычисление ROC-кривой на основе полученных данных.

Данную процедуру повторяли для всех комбинаций параметров модели, далее выбирали оптимальные параметры модели, основываясь на максимуме оценки точности pAUC (McClish 1989; Siebert and Söding 2016). pAUC вычислялась как часть площади под кривой ROC для значений FPR (встречаемости мотива в негативной выборке) меньше 0.001. Описанный выше способ выбора оптимальной длины на основе наилучшей точности распознавания ССТФ был разработан ранее [46, 148] и апробирован в массовом анализе в ходе выполнения текущей работы [34].

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

2.5 Выбор порога для моделей на основе фиксированной ошибки перепредсказания

Чтобы корректно сравнивать результаты поиска мотивов с применением разных моделей, необходимо единообразно установить для всех моделей пороговые значения их функций распознавания. Эти пороги определяли для

всех моделей по ожидаемой частоте мотива (англ. expected recognition rate - ERR). Для её вычисления использовали геномную выборку промоторов генов из базы ENSEMBL (https://www.ensembl.org/index.html, https://plants.ensembl.org/index.html), в которую входили 5′-участки кодирующих белок генов (2000 п. о. от сайта старта транскрипции для мыши и человека, 1500 п. о. от сайта старта транскрипции для арабидопсиса), поскольку в этих участках ожидается наличие функциональных ССТФ. Для каждого гена в анализе использован только один сайт старта транскрипции, соответствующий его каноническому транскрипту, указанному в поле ‘gene’, основной согласно использованной аннотации. Для версий генома человека (hg38), мыши (mm10) и арабидопсиса (TAIR10) объём этих выборок составил 19 795, 19 991 и 27 206 генов.

Значение ERR вычисляли следующим образом. Определяли значение функции распознавания модели сайтов во всех промоторах белок-кодирующих генов в каждой позиции и цепи ДНК. Затем значение ERR для каждого уникального значения функции распознавания (порога) вычисляли как отношение количества предсказанных сайтов, для которых значение функции выше этого порога, к общему числу позиций в выборке, доступных для таких сайтов. Таким образом составлена таблица «порог – ERR». При распознавании сайтов для всех моделей выбирали фиксированное значение ERR и вычисляли значение функции распознавания [33]. После выбора порога для каждой модели, сканировали пики ChIP-seq. Пример выбора порога для PWM длиной 20 п.о. с фиксированным значением ERR, равным $1.9 * 10^{-4}$, приведен на рисунке 14.

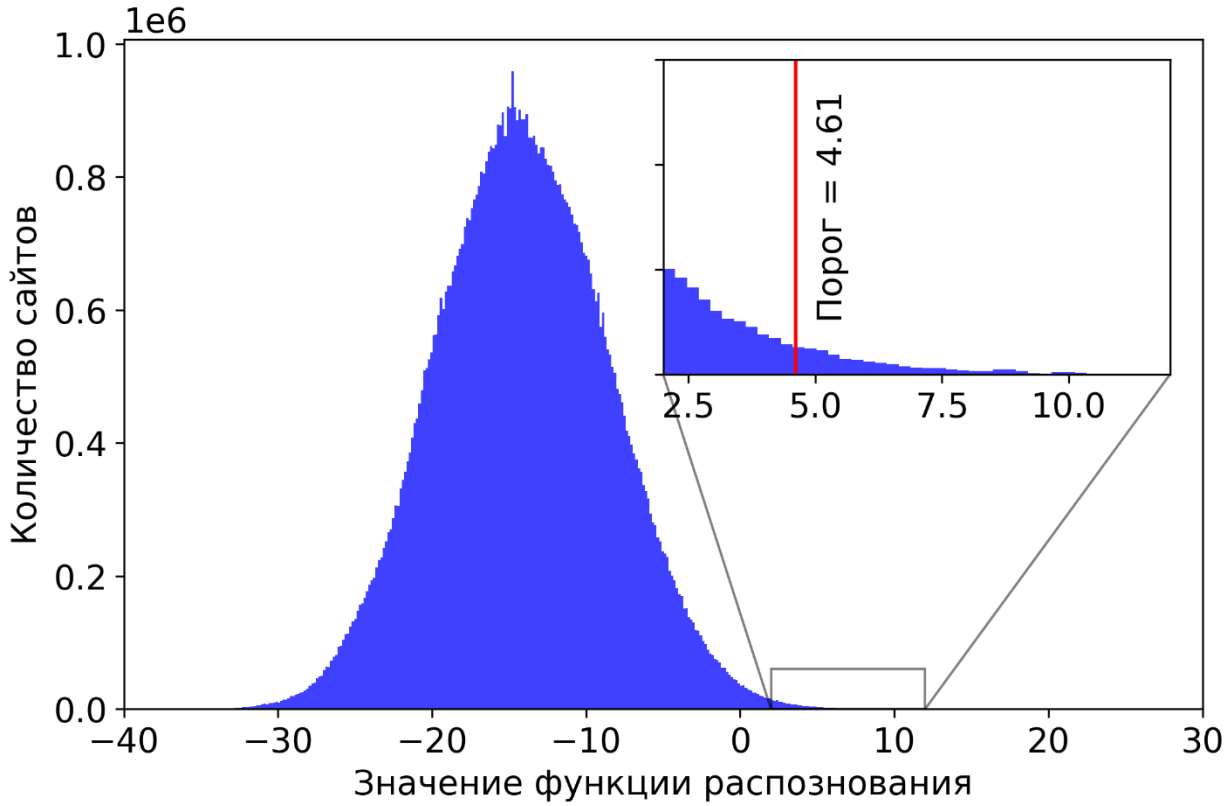

Рисунок 14. Выбор порога для модели по фиксированной ошибке перепредсказания с использованием в качестве негативной выборки последовательности промоторов.

2.6 Классификация пиков ChIP-seq по результатам распознавания сайтов разными моделями мотива

После выбора порога для каждой модели, производилось предсказание сайтов в пиках ChIP-seq. Далее пики классифицировали на фракции в зависимости от присутствия/отсутствия сайтов, найденных разными моделями мотива (PWM, diPWM, BaMM, InMoDe), с учётом расположения сайтов разных моделей в позициях пиков, так и без такого учёта (на основе присутствия или отсутствия сайтов в пиках), согласно ранее разработанной методике [44, 45]. В частности, классификацию пиков с учётом позиций сайтов разных моделей проводили для каждой пары моделей, используемой в анализе данных. Если в пике присутствовали сайты, предсказанные только одной моделью, то данный пик классифицировался как пик соответствующей

модели. Если в пике найдены сайты двух разных моделей, то возможны два исхода (Рисунок 15).

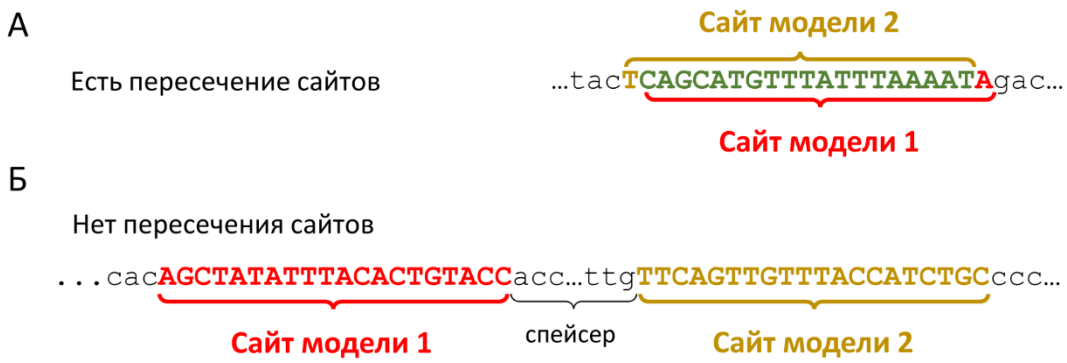

Рисунок 15. Пример классификации ChIP-seq двух пиков, в которых обнаружены сайты двух разных моделей, показывает, что в пике сайты перекрываются (*а*) или нет (*б*).

В первом случае, если существует хотя бы одна пара сайтов разных моделей, которые имеют как минимум одну общую позицию, такой пик классифицируется как «пересечение сайтов». В другом случае, когда в пике присутствуют сайты разных моделей, 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 (Рисунок 16).

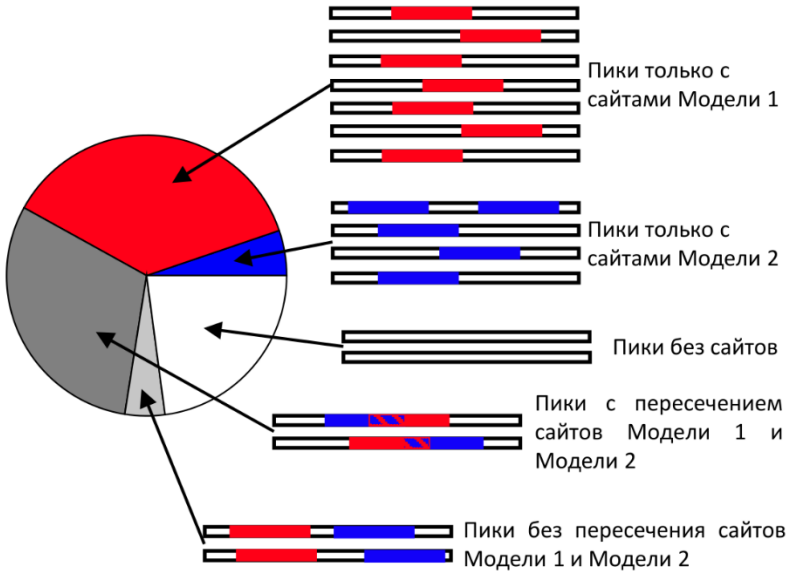

Рисунок 16. Классификация ChIP-seq пиков для двух моделей с учетом пересечения ССТФ.

Классификацию пиков без учета позиций сайтов 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.

2.7 Сравнение найденных мотивов с известными мотивами ТФ с помощью программы TomTom

Чтобы оценить, соответствует ли мотив, которые предсказывает модель, известным мотивам целевого ТФ, была использована программа сравнения мотивов TomTom [141]. Эта программа предназначена для оценки значимости сходства матриц частот нуклеотидов. Для каждой модели на основе найденных с их помощью сайтов строили матрицу частот нуклеотидов. Далее с помощью TomTom оценивали схожесть этой матрицы со всеми известными частотными матрицами целевого ТФ, для которого делался ChIP-seq эксперимент, взятыми из баз данных HOCOMOCO [8], CIS-BP [140], либо JASPAR [139]. Если при сравнении матриц значимость отличия *p*-value была меньше 0.05, то считали, что сходство мотивов значимо, то есть набор пиков ChIP-seq обогащен СС целевого ТФ.

2.8 Аннотация пиков, содержащих ССТФ и анализ терминов ГО

Для анализа терминов ГО применили пакет ChIPseeker [216]. В качестве генов для анализа онтологии брали только те, в промоторах (±1000 п.о. для *A. thaliana* и ±3000 п.о. для *M. musculus* от сайта старта транскрипции) которых находились пики, содержащие ССТФ. Обогащение терминами ГО, а именно терминами «биологический процесс», проводили с использованием пакета clusterProfiler [217]. В результате получали список терминов, для которых вычислялась кратность изменения (англ. fold change, FC) (см. Таблица 6) следующим образом:

$$FC= [(Mot_{+}GO_{+})/Mot_{+}] / [GO_{+}/(GO_{+} + GO_{-})]$$

Также для каждого термина была определена значимость, скорректированная с учётом множественных сравнений (p_adj) методом Беньямини-Хохберга [218].

Таблица 6. Таблица сопряжённости 2 × 2, объясняющая анализ обогащения терминов ГО. Каждая ячейка представляет число генов, имеющих или не имеющих термин GO, и обладающих или не обладающих предсказанными сайтами в промоторах генов, попадающих в пики.

	Число генов с термином ГО	**Число генов без термина ГО**	**Всего генов**
Число генов с мотивом	Mot_+GO_+	Mot_+GO_-	Mot_+
Число генов без мотива	Mot_-GO_+	$Mot\text{-}GO_-$	Mot_-
Всего генов	GO_+	GO^-	

2.9 Сравнение специфики поиска мотивов разными моделями

Разные модели по-разному представляют мотивы. Чтобы оценить, насколько похожи/различаются результаты поиска сайтов для разных пар моделей применили корреляционный тест по значениям функций распознавания моделей, посчитанных для перекрывающихся сайтов. Для этого сопоставляли значения функции распознавания для сайтов, которые предсказаны разными моделями и при этом, как минимум, на 50% пересекают друг друга. Для корректности сравнения результатов разных моделей, использовали таблицу «порог модели – ожидаемая частота мотива ERR». Далее считали значения коэффициент корреляции Пирсона для значений ERR любой заданной пары моделей. Результаты корреляции представляли в виде тепловых карт (Рисунок 17).

На тепловой карте (Рисунок 17) по осям X и Y отложены значения - $\log_{10}$(ERR), которые характеризует сходство между пересекающимися сайтами по значениям функции распознавания (ERR) для разных моделей.

Цветом обозначена доля пиков, для соответствующих сочетаний значений -$\log_{10}$(ERR) по осям X и Y.

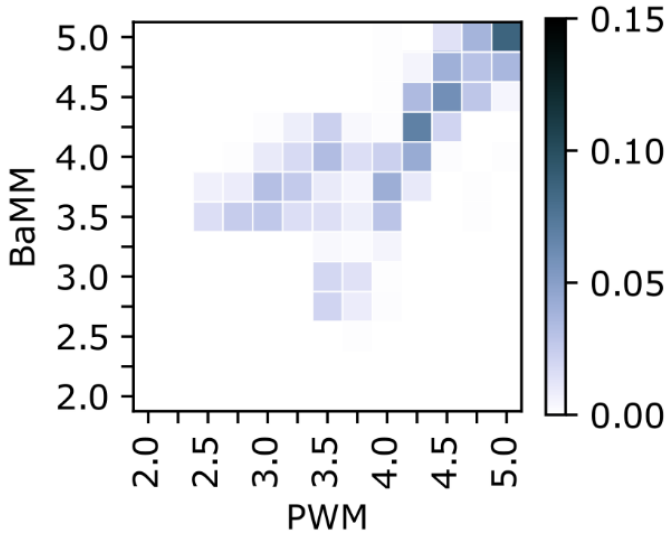

Рисунок 17. Пример тепловой карты для сравнения специфичности распознавания мотивов с моделями PWM и BaMM для ТФ CCA1 (GTRD ID PEAKS042882).

2.10 Статистический анализ и визуализация

Весь статистический анализ осуществляли на языке программирования Python 3.8 в среде Jupyter с использованием пакетов numpy [219], pandas, scipy [220] и statannot. Визуализацию мотивов проводили с помощью инструментов logomaker (пакет Python) [221] (традиционное лого для модели мотива PWM), и DepLogo [222] (альтернативное лого для визуализации зависимостей позиций мотива). Графики рисовали с использованием пакетов языка Python – matplotlib [223] и seaborn [224].

В анализе результатов также использовали иерархическую классификации ТФ по структуре ДСД предложенной Вингендером [81–83]. В качестве источника информации о классах ТФ использовали базу данных JASPAR, где представленная информация о классах ТФ не только для млекопитающих, как в базе TFClass, а также классификация ТФ для других таксонов, включая и растения [225].

3. Результаты и обсуждение

3.1 Анализ данных на примере FOXA2

Для анализа был использован набор предобработанных ChIP-seq данных в виде разметки пиков в формате bed из базы данных ReMap http://remap.univ-amu.fr/ [183]. Набор данных включал 22 ChIP-seq эксперимента для ТФ FOXA2 (Таблица 7).

Таблица 7. Список ChIP-seq экспериментов, используемых в работе

№	GEO/ENCODE ID	Клеточная линия/ткань	Условия	TomTom
1	ENCSR000BNI	Hep-G2	–	+
2	ENCSR000BRE	A-549	–	+
3	ENCSR066EBK	Hep-G2	–	–
4	ENCSR080XEY	Liver	–	+
5	ENCSR310NYI	Liver	–	+
6	ERP004206	H9	–	+
7	ERP008682	Pancreas	CARN1618	+
8	GSM2401464	A-549	–	+
9	GSM2401446	BJ1-hTERT	–	+
10	GSM2977505	BJ1-hTERT	FoxHnf1aCoExp	–
11	GSM2401452	BJ1-hTERT	GATA4 coexpression	–
12	GSM2977503	BJ1-hTERT	Mimosine	–
13	GSM2977504	BJ1-hTERT	MimosineRelease	–
14	GSM2401466	BJ1-hTERT	Mimosine	+
15	GSM2401468	BJ1-hTERT	MimosineRelease	+
16	GSM2977506	BJ1-hTERT	uninduced FOXA2	+
17	GSM2401454	Hep-G2	–	–
18	GSM2401456	KerCT	–	+
19	GSM2430680	BJ1-hTERT	CDT1	+
20	GSM2430677	BJ1-hTERT	GATA4 coexpression	+
21	GSM2430678	BJ1-hTERT	Mimosine	+
22	GSM2430679	BJ1-hTERT	MimosineRelease	+

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), соответственно, на частотную матрицу ССТФ FOXA2 из базы данных HOCOMOCO FOXA2_HUMAN.H11MO.0.A [8].

Из каждого эксперимента для анализа отобрали по 4000 пиков максимального качества. При анализе данных ChIP-seq для FOXA2

использовались модели PWM (ChIPMunk), diPWM (diChIPMunk), BaMM и InMoDe.

3.1.1 Фильтрация данных на основе сравнения мотивов программой TomTom

Чтобы убедиться, что построенные модели мотивов соответствуют известным мотивам FOXA2 и последующий анализ является корректным, был применён фильтр на основе программы оценки сходства мотивов TomTom. Для этого частотные матрицы построенные по результатам распознавания сайтов модели PWM сравнивали с соответствующими частотными матрицами известных ТФ из базы данных HOCOMOCO [8]. Только в 6 из 22 ChIP-seq наборов согласно TomTom построенная модель PWM не обладала сходством с известными сайтами FOXA2 (см. Таблицу 7), поэтому в дальнейшем анализе использовались оставшиеся 16 наборов данных ChIP-seq.

3.1.2 Оценка точности распознавания ССТФ для FOXA2 разными моделями и выбор оптимальных длин

Чтобы выбрать оптимальную длину для каждой модели, на разных длинах и порядках ММ для каждой модели оценили точность поиска ССТФ с помощью перекрёстной проверки. Точность моделей сравнивали по показателю pAUC, посчитанному для диапазона значений FPR < 0.001. На рисунке 18 показано, как в среднем изменяется значение pAUC в зависимости от длины мотива и порядка цепи, если речь идёт о ММ.

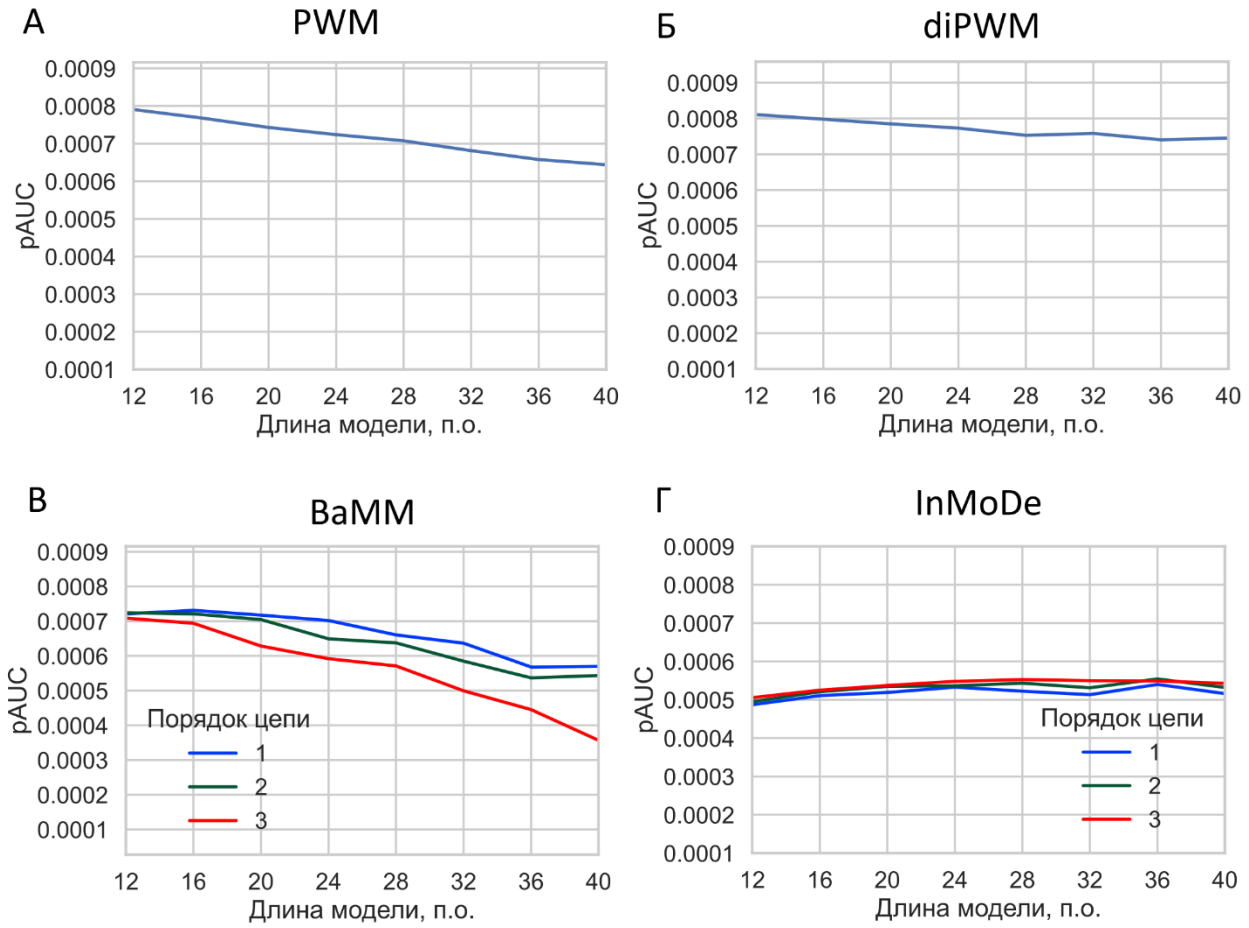

Рисунок 18. Влияние длины мотива и порядка цепи ММ на оценку точности pAUC моделей мотива. PWM **(А)**, diPWM **(Б)**, BaMM **(В)**, InMoDe **(Г)**

Полученные результаты для мотивов FOXA2 демонстрируют, что большинство моделей, за исключением InMoDe, имеют тенденцию терять точность с ростом длины мотива. Возможно, это связано с тем, что ТФ FOXA2 имеет достаточно короткий и консервативный мотив. Для InMoDe наблюдается рост pAUC с ростом длины мотива, который происходит до длины 24 п.о., далее рост точности замедляется. Интересным является и то, что для BaMM изменение порядка цепи приводит к ухудшению точности модели, а для InMoDe порядок оказывает незначительное улучшение.

По максимальному значению pAUC для каждого отдельного ChIP-seq эксперимента и каждой модели выбиралась оптимальная длина с учётом порядка цепи ММ в случае для BaMM и InMoDe (Рисунок 19).

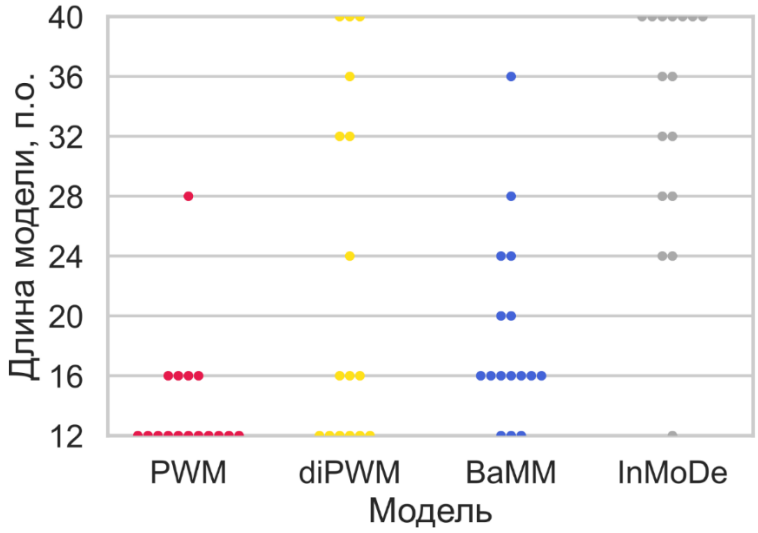

Рисунок 19. Распределения длин для оптимальных моделей мотивов FOXA2.

Как видно из распределений длин мотивов (Рисунок 19), модели мотивов PWM, diPWM преимущественно имеют длину 12 п.о., хотя в отдельных случаях они имеют длину 16 п.о. и больше, из чего можно предположить, что diPWM использует дополнительную информацию, в виде некоторых дополнительных позиций динуклеотидов, вокруг самого консервативного участка мотива. Модель BaMM имеет наибольшую точность при длине 16 п.о., что также выше, чем у PWM. Модель InMoDe в большинстве экспериментов имеет максимальную длину 40 п.о., возможно за счёт устройства данной модели она может учитывать широкий спектр дополнительных зависимостей вокруг консервативного участка мотива, тем самым увеличивая его длину.

Сравнили точность всех моделей относительно друг друга при оптимальных параметрах каждой модели (Рисунок 20). Для каждой из моделей по всем ChIP-seq экспериментам результаты представлены в виде ROC-кривой из медиан значений FPR в каждой точке TPR (Рисунок 20А). Из полученных результатов видно, что точность в последовательности моделей diPWM-PWM-BaMM-InMoDe снижается. Также для каждой модели были рассчитаны распределения значений pAUC (Рисунок 20Б).

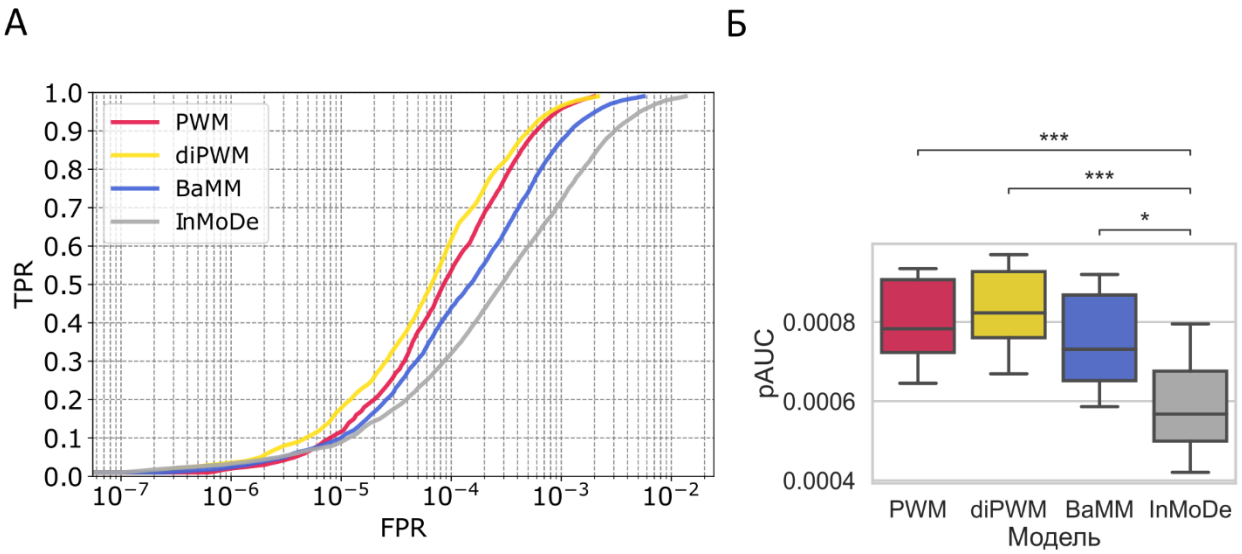

Рисунок 20. Сравнение точности моделей PWM, diPWM, BaMM и InMoDe для мотивов FOXA2: **(А)** ROC-кривые; **(Б)** значения pAUC для всех моделей по всем ChIP-seq экспериментам. * - p < 0.05; ** - p < 0.01; *** p < 0.001.

Согласно полученным результатам, значения медиан pAUC для моделей PWM, diPWM, BaMM и InMoDe равны 7.9E–4, 8.0E–4, 7.3E–4 и 5.6E–4, соответственно. Полученные значения pAUC в парных сравнениях для моделей PWM, diPWM и BaMM значимо не отличаются ($p > 0.05$), однако для модели InMoDe оно достоверно меньше, чем у остальных моделей ($p < 0.05$).

3.1.3 Классификация пиков ChIP-seq без учёта пересечения сайтов, найденных разными *de novo* моделями

Основным результатом работы программного комплекса MultiDeNA является классификация пиков, которая позволяет установить, как соотносятся модели мотивов между собой по способности выявлять пики с мотивами. Всего используются два типа классификации пиков: с учётом пересечения позиций мотивов разных моделей и без него. Частный случай классификации приведён на примере данных GSE90454.FOXA2.KerCT (Рисунок 21).

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
1339	183	33	106	238	53	28	20	38	35	547	15	12	375	309

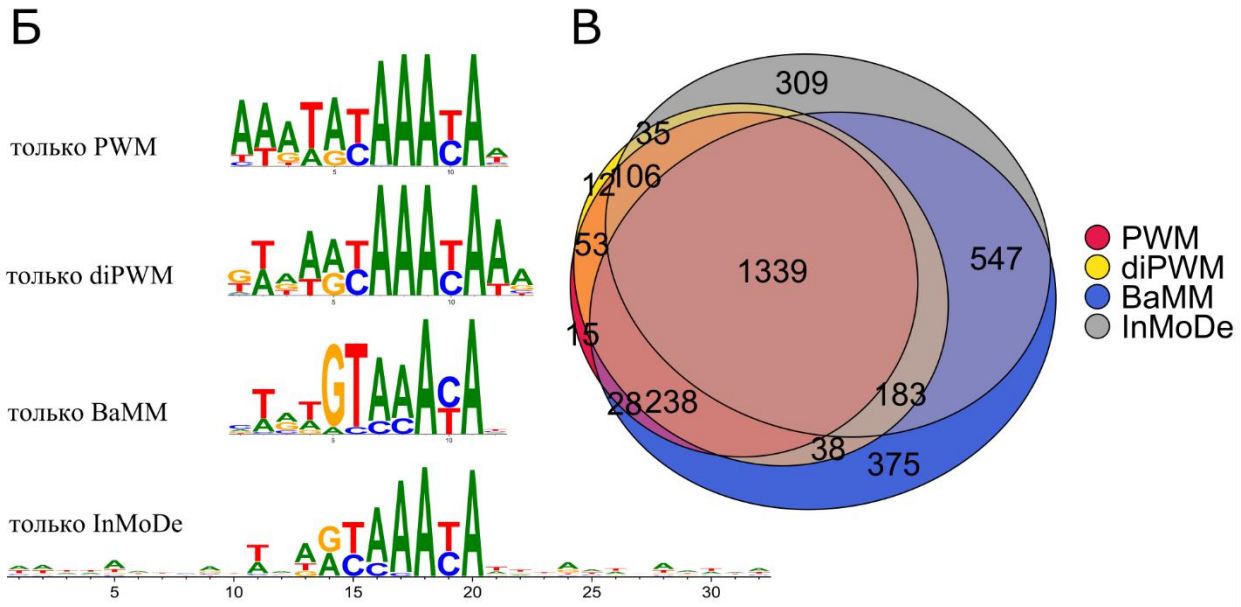

Рисунок 21. Классификация пиков по результатам сканирования всеми четырьмя моделями (PWM, diPWM, BaMM и InMoDe). Проанализирован набор данных GSE90454.FOXA2.KerCT. **(А)** – таблица; **(Б)** – лого-диаграммы для фракций пиков, содержащих сайты только одной из моделей, и для фракций, где сайты всех моделей пересечены; **(В)** – диаграмма Венна, изображающая фракции пиков, содержащих предсказания ССТФ разных комбинаций моделей

Рассмотрим более детально классификацию ChIP-seq пиков по результатам поиска сайтов четырьмя моделями мотива без учета позиций сайтов в пиках. Можно видеть, что суммарно все четыре модели распознали сайты в 83.28 % пиков (3331 из 4000, сумма всех областей на диаграмме Венна, см. Рисунок 21А, Б). Общая для всех моделей группа пиков, в которых сайты были найдены четырьмя моделями одновременно, составила 33.48 % (1339 из 4000 пиков). Самые крупные независимые от других моделей вклады в распознавание вносят модель BaMM, которая добавляет 9.38 % пиков (375), и модель InMoDe которая добавляет 7.73 % пиков (309), в отличие от моделей PWM и diPWM, которые добавляют 0.38 % (15) и 0.3 % (12), соответственно.

При этом модели (BaMM и InMoDe) вносят значительные вклады в распознание пиков в дополнение к результату модели PWM, который составляет 30.78 % (375 + 309 + 547 = 1231 из 4000), что сопоставимо с фракцией перекрывания всех моделей (1339).

Чтобы оценить структурное разнообразие ССТФ, построили лого-диаграммы для фракций пиков «только PWM», «только diPWM», «только BaMM», «только InMoDe» (см. Рисунок 21Б). 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 на всех лого-диаграммах разнообразны по информационному и нуклеотидному содержанию. Отдельного внимания заслуживает лого-диаграмма для набора ССТФ «только InMoDe», где и в 5’- и 3’-концах присутствует небольшое информационное содержание разных нуклеотидов, которое возможно связано с партнёрскими ТФ, которые связываются с ДНК в непосредственной близости от FOXA2.

Чтобы массово исследовать вклады разных моделей в эффективность поиска ССТФ FOXA2 и оценить общий результат использования нескольких моделей, определили, в какой доле пиков каждая модель и все модели вместе распознают хотя бы один сайт FOXA2 по всем наборам ChIP-seq (Рисунок 22).

Значения медиан долей распознанных пиков составили 42.4%, 42.5%, 43.0% и 34.9% для моделей PWM, diPWM, BaMM и InMoDe, соответственно, а медиана доли распознанных пиков при сочетании результатов всех четырех моделей равна 59.8%. Следовательно, совместно все модели находят на 17.4% больше пиков, содержащих мотивы, чем модель PWM, что согласуется с ранее полученным результатом применения двух принципиально разных моделей PWM и SiteGA [45]. При этом доли распознанных пиков для всех моделей относительно друг друга значимо не отличаются ($p < 0.05$). Таким образом, подход с сочетанием разных моделей позволяет лучше выявлять пики с мотивами для FOXA2, чем использование только одной модели.

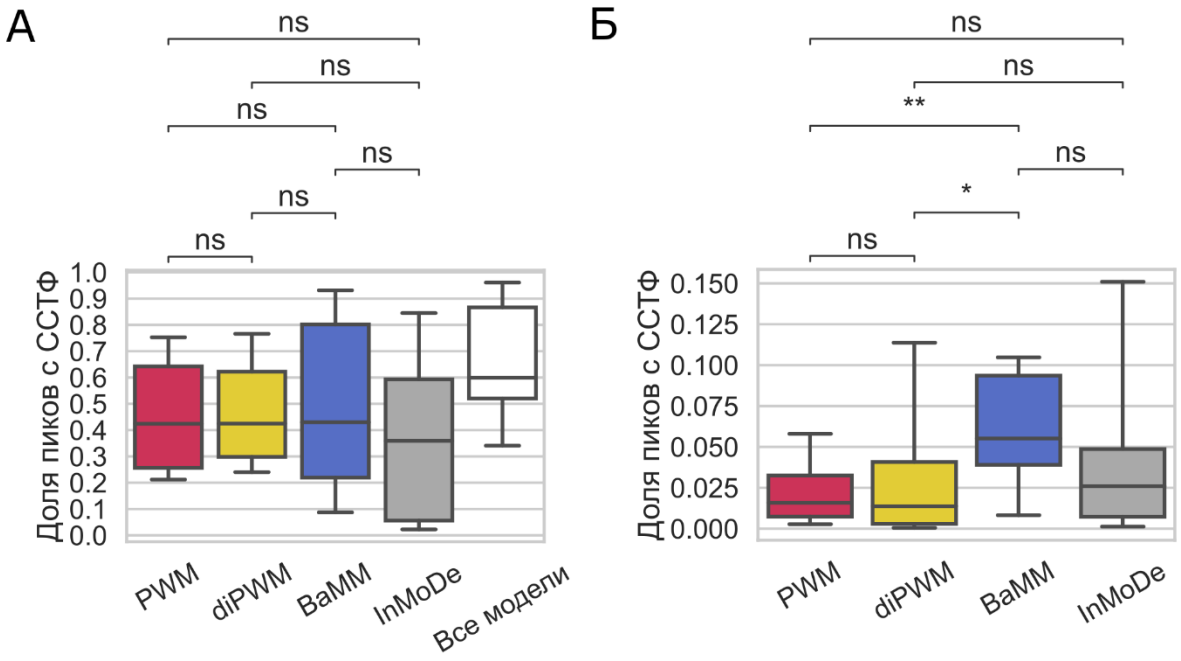

Рисунок 22. Диаграммы распределения квартилей для данных: **(А)** значения доли пиков с сайтами, распознанных каждой моделью мотива в отдельности (PWM, diPWM, BaMM, InMoDe) и всеми моделями (Total); **(Б)** значения доли пиков, в которых находится сайт только одной из моделей. ns – $p > 0.05$; * - $p < 0.05$; ** - $p < 0.01$.

Как показано выше, сочетание разных моделей увеличивает количество пиков с сайтами. Соответственно, каждая модель должна распознавать ССТФ, которые не распознают остальные. Далее пики, содержащие сайты только одной из моделей мотива, будут определены как *уникальные пики*. Полученные доли уникальных пиков для моделей представлены на рисунке 22Б. Как видно, каждая модель (PWM, diPWM, InMoDe, BaMM) способна находить сайты, которые не обнаруживаются остальными моделями. Значения медиан по доле уникальных пиков для PWM, diPWM, BaMM и InMoDe составили 1.6%, 1.4%, 5.5% и 2.6%, соответственно. Как BaMM, так и InMoDe имеют больше уникальных пиков по сравнению с моделью PWM. При этом только для BaMM количество уникальных пиков значимо больше ($p < 0.05$) по сравнению с PWM и diPWM. Полученный результат может свидетельствовать о том, что ММ (BaMM, InMoDe) могут более полно описывать ССТФ FOXA2. Из этого можно предположить, что в ССТФ FOXA2 существуют зависимости разных позиций, и их учёт увеличивает долю

распознанных пиков. 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.

3.1.4. Классификация пиков ChIP-seq с учётом пересечения сайтов, найденных разными моделями

Описанная выше классификация пиков без учёта позиций сайтов 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 мотивов 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 (Рисунок 23).

Все пары сочетаний моделей имеют незначительный размер фракции пиков «без пересечения», варьирующий от 0.3 до 5.8%. С другой стороны, для всех случаев характерен значительный размер фракции пиков «с пересечением»: BaMM-InMoDe – 46.8%, PWM–diPWM – 43.1%, diPWM–BaMM – 41.0%, diPWM–InMoDe – 36.8%, PWM–BaMM – 36.7%, PWM–InMoDe – 31.6%; при этом данная фракция имеет больший размер для методологически близких пар моделей BaMM–InMoDe и PWM–diPWM, и меньший для методологически более удалённых пар моделей PWM–BaMM и PWM–InMoDe (Рисунок 23). Фракция пиков, где мотивы находятся только одной из моделей, выражена для BaMM. В парах PWM–BaMM, diPWM–BaMM и InMoDe–BaMM она преобладает относительно второй модели пары (28.6, 24.6 и 17.0% соответственно).

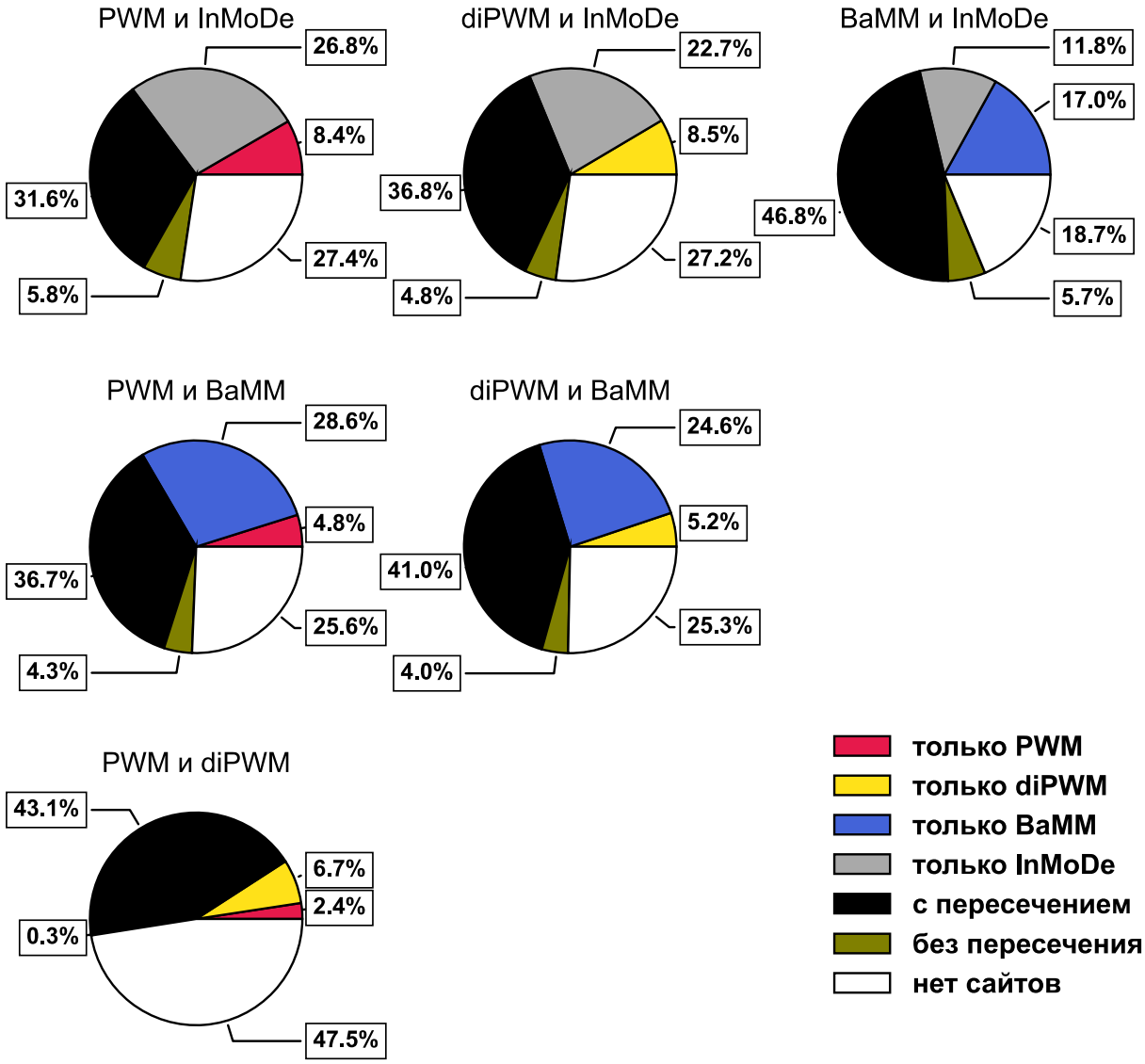

Рисунок 23. Классификация ChIP-seq пиков с учётом пересечения сайтов, распознанных разными моделями мотива на примере набора данных ChIP-seq GSE90454.FOXA2.KerCT.

По всем экспериментам для каждой модели построили распределения долей пиков класса «только одна модель» для всех сочетаний пар моделей, чтобы оценить вклад в поиск мотивов FOXA2 (Рисунок 24).

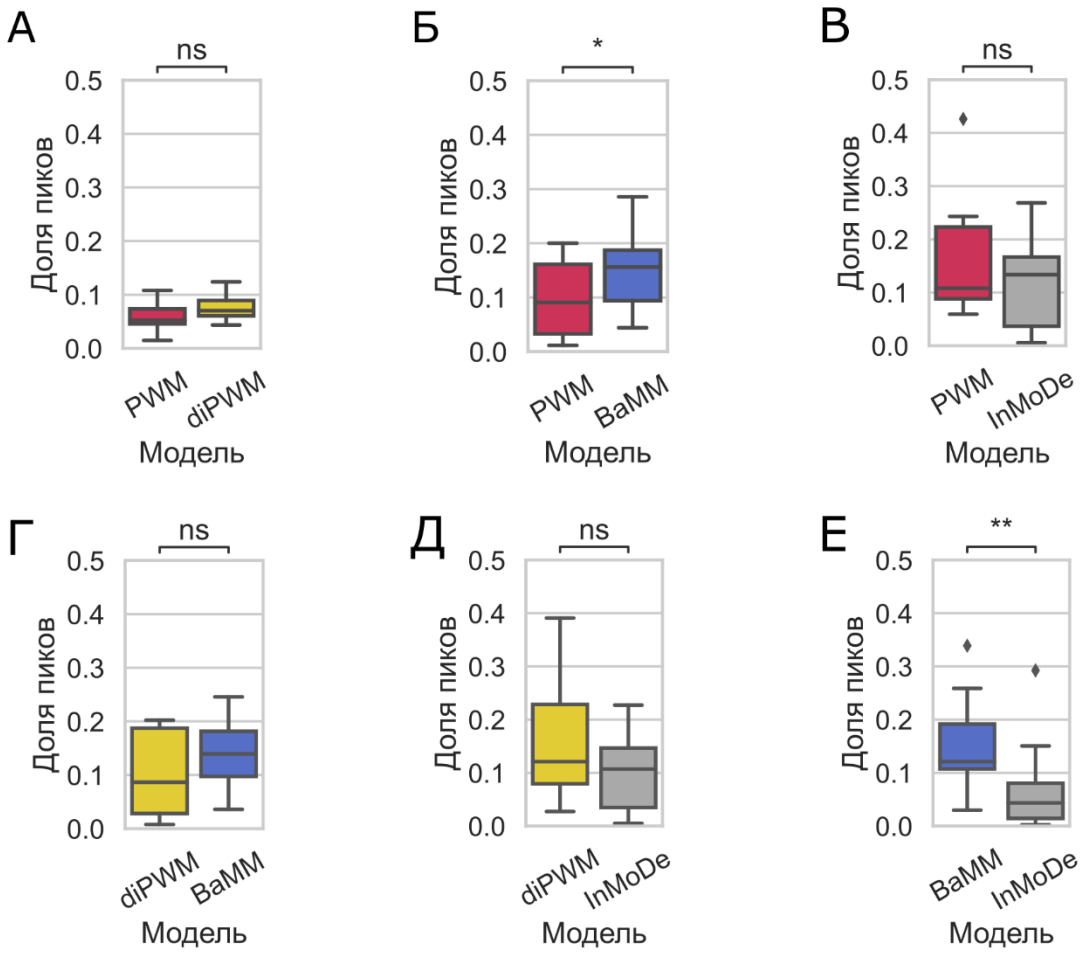

Рисунок 24. Результаты попарного пересечения результатов моделей с последующим сравнением размера фракций пиков «только одна модель». **(А)** PWM-diPWM, **(Б)** PWM-BaMM, **(В)** PWM-InMoDe, **(Г)** diPWM–BaMM, **(Д)** diPWM-InMoDe, **(Е)** InMoDe-BaMM.

Почти во всех парах каждая модель способна находить заметную долю пиков с дополнительными сайтами, которая не находится второй моделью пары. Медиана дополнительного вклада в парах моделей варьируется от 4.3% до 15.6%. В двух отдельных случаях, а именно парах PWM-BaMM и InMoDe-BaMM, модель BaMM находит значимо больше дополнительных сайтов, из чего можно предполагать, что данная модель лучше учитывает особенности структуры ССТФ FOXA2. Тем не менее только использование нескольких моделей, позволяет найти более полный набор сайтов.

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

Чтобы понять, насколько похожи модели мотивов, полученные по разным наборам данных ChIP-seq, мы провели перекрёстную проверку точности распознавания. Оценили точность каждой модели не только для набора данных выборки обучения (каждого отдельного эксперимента ChIP-seq), но и на остальных 15 наборах данных ChIP-seq (контрольные наборы). Для каждого набора рассчитали оценку точности pAUC, результаты представили в виде тепловых карт (Рисунок 25). Из тепловых карт видно, что модели на данных обучения и контроля практически не отличаются. При сравнении медиан pAUC (графики под тепловыми картами соответствующих моделей), посчитанных по наборам данных обучения и контроля для PWM и diPWM, наблюдается небольшое снижение точности моделей на контрольных наборах данных, а для BaMM наоборот, на контрольных наборах данных точность немного выше. Однако стоит отметить, что между значениями pAUC нет статистической разницы. Для InMoDe отличий между значениями pAUC практически нет. Такой результат согласуется с данными, опубликованными в статье [226], где авторы показали, что даже ТФ, имеющие схожие ДСД, т.е. относящиеся к одному семейству, способны с высокой точностью распознавать сайты друг друга. В случае описанного в данном разделе работы (3.1), ТФ был одним и тем же, а отличались только условия экспериментов ChIP-seq. Полученные данные могут свидетельствовать об отсутствии специфики среди СС одного ТФ, работающего в разных условиях клеток, тканей и т.п., возможно большую роль играют другие причины различия СС, такие как партнёрские ТФ [33].

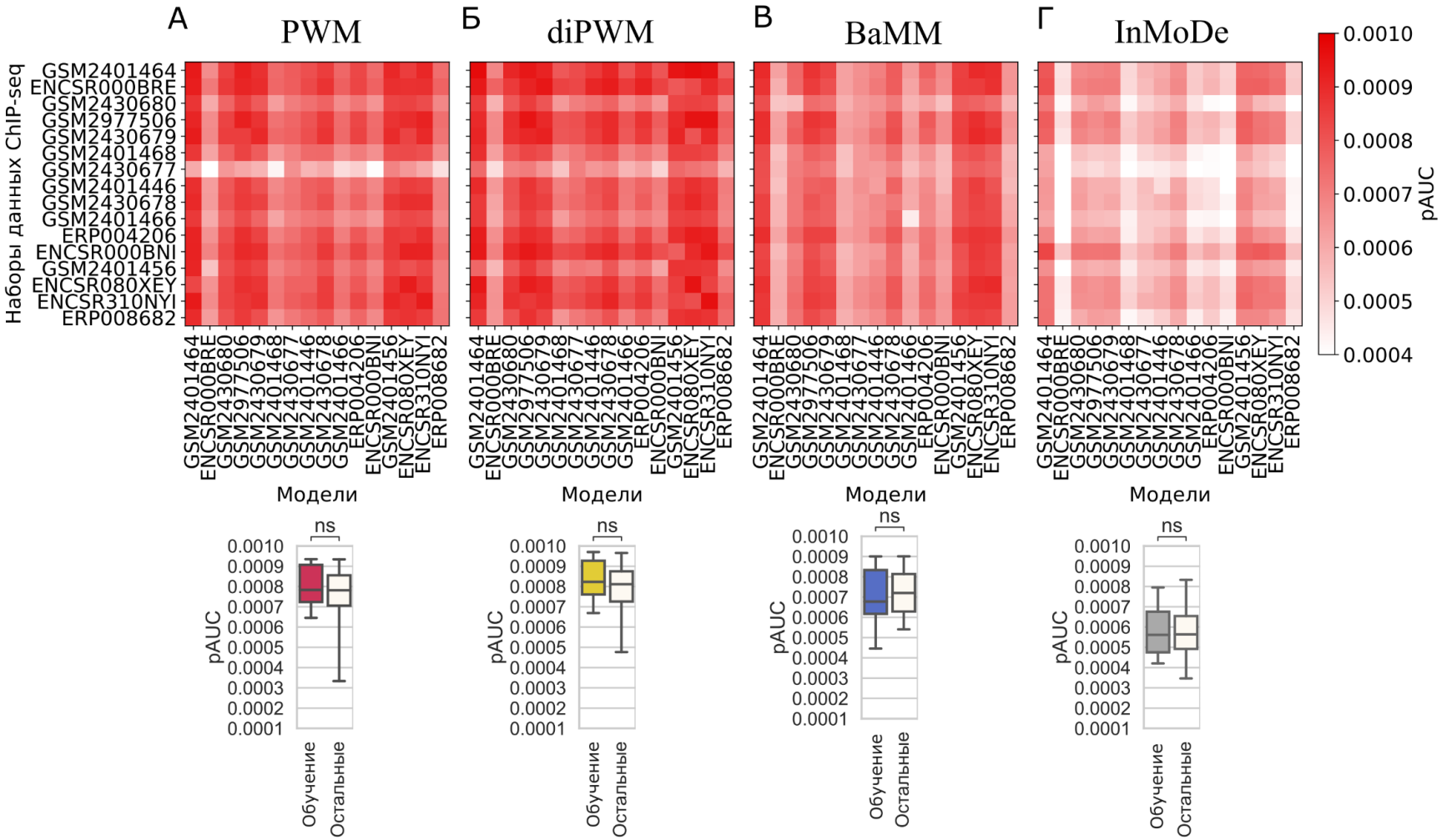

Рисунок 25. Тепловая карта сравнения точности распознавания pAUC. На тепловых картах оттенки цвета обозначают значения pAUC. Для ячеек, расположенных по диагонали, приведены значения pAUC полученные с помощью CV. В остальных ячейках наборы данных для обучения и контроля отличаются. Строки означают модели, а столбцы обозначают наборы данных ChIP-seq. Под тепловой картой представлены диаграммы распределения квартилей pAUC. **(А)** – PWM, **(Б)** – diPWM, **(В)** – BaMM, **(Г)** – InMoDe.

3.2 Массовый анализ данных ChIP-seq для *A. thaliana*

3.2.1 Подготовка данных и выбор оптимальных моделей для анализа

Для массового анализа данных ChIP-seq для *A. thaliana* по разным ТФ была сформирована коллекция из результатов 121 предобработанных экспериментов, наборов данных ChIP-seq в виде разметок пиков в формате bed, взятых из базы данных GTRD (https://gtrd.biouml.org/#!) [185]. Полный список экспериментов ChIP-seq с идентификаторами GTRD, а также с информацией о качестве данных (обогащение целевого мотива с помощью AME [227], сравнениями *de novo* мотивов с известными мотивами целевых ТФ (с помощью TomTom), представлен в Приложении А). Для каждого ChIP-seq эксперимента были взяты пики, длина которых не превышала 3000 п.о., а в анализ отобрали 1000 пиков самого высокого качества. В данном анализе использовались модели PWM, BaMM и SiteGA. Данный выбор моделей обусловлен двумя причинами: (1) при *de novo* поиске мотивов для данных моделей есть возможность задать негативную выборку в качестве параметра. (2) данные модели являются методологически разными.

В работе [214], а также в исследовании сделанным нашим коллективом [228] было показано, что выбор негативной выборки существенно влияет на результат *de novo* поиска мотива. Если использовать в качестве негативной выборки сгенерированные последовательности, полученные путём перемешивания нуклеотидов позитивной выборки, то в значительной доле ChIP-seq экспериментов будут вывялятся неспецифические мотивы такие как поли-A, динуклеотидные повторы и т.п.. Решением данной проблемы является использование негативной выборки, сгенерированной на основе полного генома [228]. Реализации *de novo* поиска мотивов PWM (ChIPMunk), diPWM (diChIPMunk) и InMoDe не могут использовать негативной выборки, сгенерированной из генома, поэтому данные модели не использовались в последующем анализе. Вместо ChIPMunk использовали STREME в качестве *de novo* инструмента для модели мотива PWM. Данной программе можно

задать любую негативную выборку [14]. Недавно разработанная модель SiteGA, так же может использовать любую заданную негативную выборку [34].

Касательно второй причины, BaMM является представителем MM, и замещает diPWM и InMoDe, поскольку diPWM является MM 1-го порядка, а InMoDe обладает худшей точностью среди данных моделей (см. Рисунок 18). SiteGA является моделью, которая учитывает зависимости позиций на любом расстоянии внутри мотива.

3.2.2 Оценка качества исходных данных

Чтобы убедиться в том, что ChIP-seq эксперименты содержат мотивы целевого ТФ и последующий анализ был корректным, данные проверили двумя способами: (1) проверка на обогащение мотива целевого ТФ (или ТФ того же семейства) в пиках ChIP-seq; (2) оценка сходства *de novo* мотивов с мотивом целевого ТФ.

Для того, чтобы проверить, обогащены ли мотивы целевых ТФ в наборах данных ChIP-seq, мы извлекли известные мотивы для всех целевых ТФ из баз данных CIS-BP и JASPAR и далее проверили их обогащение с помощью инструмента AME [227]. Почти для всех ChIP-seq экспериментов в пиках мотивы целевых ТФ были обогащены, $p < 0.05$ в 109 случаях из 121 (см. приложение А, таблица 1).

Проверку сходства *de novo* мотивов с мотивами целевых ТФ делали с помощью инструмента оценки сходства мотивов TomTom. Отдельно для каждой модели PWM, BaMM или SiteGA по результатам применения *de novo* поиска мотивов, в виде выравнивания предсказанных сайтов, строили матрицы частот нуклеотидов, которые использовали в качестве мотивов. Эти мотивы, применяя инструмент TomTom, сравнивали с соответствующими мотивами известных ТФ из баз данных JASPAR и CIS-BP. Так как несколько разных ТФ могут иметь схожие мотивы (и для многих ТФ мотивы малоизучены), особенно это касается ТФ из одного семейства (или даже класса, если семейство не указано), то в качестве результирующего *p-value*

брали наименьшее значение внутри семейства (класса) ТФ к которому относится целевой ТФ. Аннотацию семейств по классификации ДСД для ТФ растений взяли из JASPAR [139], в которую недавно добавили информацию о классах ТФ растений [225]. В результате было обнаружено, что модели PWM, BaMM и SiteGA имели схожие мотивы (построенные по выравниванию матриц частот нуклеотидов) с мотивами целевых ТФ ($p < 0.05$) в 92, 91 и 92 случаях, соответственно (см. Приложение А). Далее, чтобы корректно сравнивать модели и полученные с помощью них результаты, в анализ взяли только те ChIP-seq эксперименты, для которых выполнялись следующие условия: (1) все три модели выявили мотив значимо похожий на мотив целевого ТФ согласно TomTom, (2) мотив целевого ТФ обогащен согласно AME. Данным условиям удовлетворяли 68 ChIP-seq экспериментов, которые и использовались в дальнейшем анализе. Наборы данных ChIP-seq были классифицированы по классам целевых ТФ по ДСД (Таблица 8).

Таблица 8. Информация о классах ТФ, используемых в анализе. Значения в колонках GTRD, AME и TomTom означают количества исходных данных извлечённых из базы данных GTRD до фильтрации, после их фильтрации по обогащению мотивов (до проведения *de novo* поиска мотивов, инструмент AME) и после их фильтрации по сходству мотивов с известными мотивами целевых ТФ (после проведения *de novo* поиска мотивов, инструмент TomTom).

Класс		Число экспериментов			Число ТФ		
		GTRD	AME	TomTom	GTRD	AME	TomTom
Basic leucine zipper factors (bZIP)	{1.1}	12	12	12	6	6	6
Basic helix-loop-helix factors (bHLH)	{1.2}	12	12	12	6	6	6
Other C4 zinc finger-type factors	{2.2}	2	0	0	2	0	0
C2H2 zinc finger factors	{2.3}	11	11	7	4	4	2
CXXC zinc finger factors	{2.6}	1	0	0	1	0	0
Homeo domain factors	{3.1}	11	11	4	6	6	3
Heat shock factors	{3.4}	6	6	0	3	3	0
Tryptophan cluster factors	{3.5}	23	21	13	12	11	9
LEAFY	{3.*}	1	0	0	1	0	0
MADS box factors	{5.1}	13	11	6	6	5	4
GCM domain factors	{7.2}	17	16	10	5	4	4
AP2/EREBP	{7.*}	6	6	2	4	4	2
Ribbon-Helix-Helix factors	{7.*}	2	2	2	1	1	1
B3	{9.*}	1	1	0	1	1	0
Не установлен	{0.*}	3	1	0	2	1	0
Вся коллекция		121	110	68	60	52	37

Примечание. Фигурные скобки представляют числовые обозначения классов согласно иерархической классификации одноимённых ТФ млекопитающих; звёзды в обозначениях классов отмечают классы, не известные у млекопитающих, или отнесённые к суперклассу {0}, содержащему ТФ не классифицированные по известным девяти суперклассам [81–83, 139, 225]

3.2.3 Выбор оптимальных параметров и оценка точности распознавания ССТФ для моделей

Чтобы обеспечить наилучшее качество моделей мотива, для каждой из них выбирали оптимальные параметры. Поиск оптимальных параметров осуществляли с помощью перекрёстной проверки (2-*fol*d CV, см. главу Методы), в качестве негативной выборки использовали случайно выбранные последовательности из генома с нуклеотидным составом и длиной, схожими с таковыми для выборок обучения (см. главу Методы). Предыдущие результаты (см. раздел 3.1) по подбору параметров моделей на данных по ТФ FOXA2 показали, что для большинства моделей (PWM, diPWM, BaMM) рост точности наблюдался только до длины 16 п.о. исключением являлась модель InMoDe, где рост точности продолжался вплоть до 40 п.о., но InMoDe в данном анализе не используется (см. Раздел 3.2.1). Поэтому длину моделей варьировали от 8

до 20 п.о. с шагом в 4 п.о. Для BaMM порядок ММ меняли в диапазоне от 0 до 4 с шагом 1, а количество ЛПД для SiteGA в диапазоне от 40 до 100 с шагом 20. На рисунке 26 представлено распределение полученных оптимальных параметров (длина модели, порядок ММ, количество ЛПД) моделей, при которых модель имела максимальную точность (pAUC).

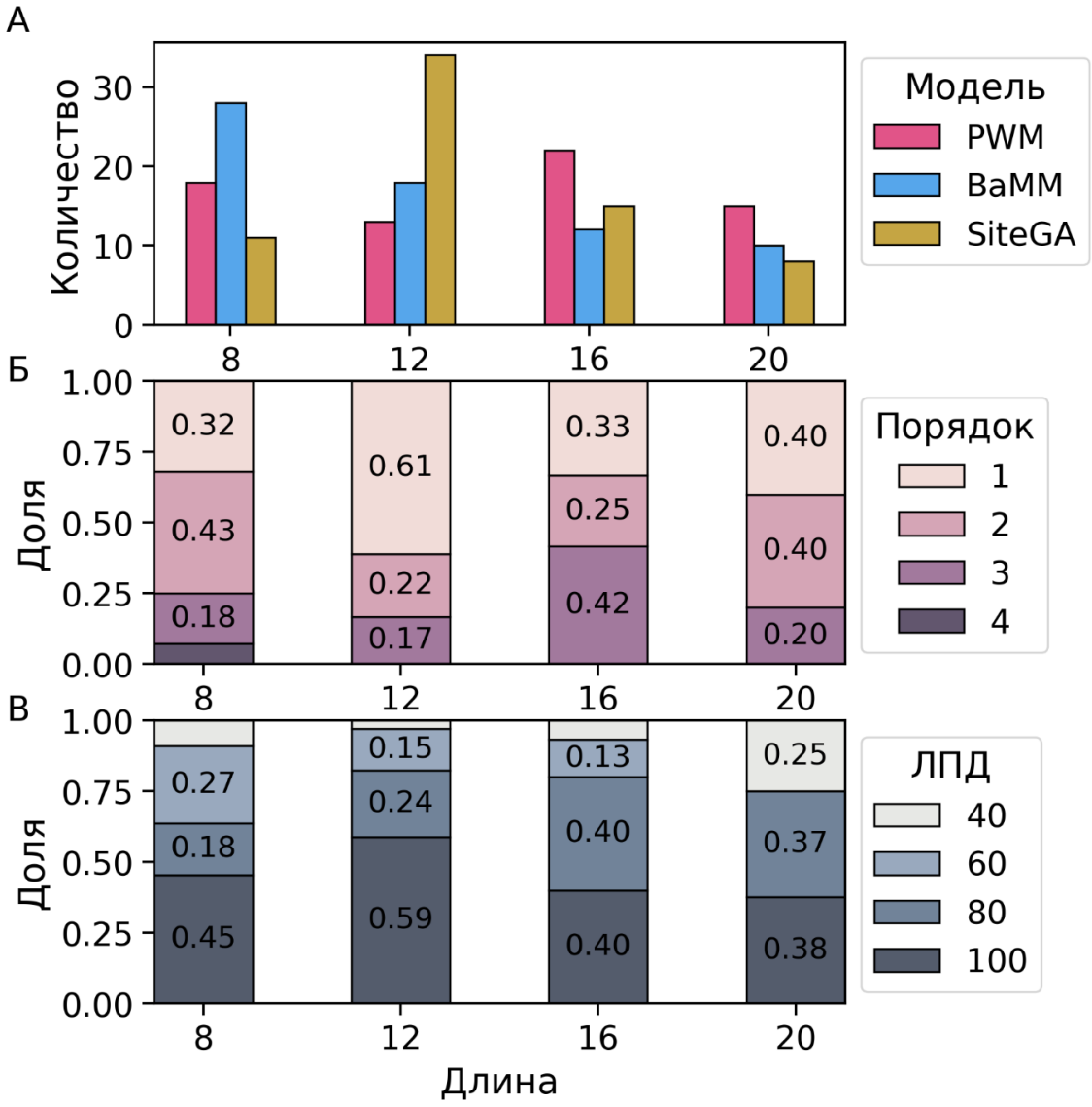

Рисунок 26. Распределение долей наборов данных ChIP-seq (ось Y) в зависимости от параметров моделей (длина, порядок ММ, количество ЛПД) (ось X), подобранных с помощью 2-fold CV. **(А)** Оптимальная длина мотива, модели PWM, BaMM и SiteGA. **(Б)** Порядок цепи ММ, модель BaMM. **(В)** количество ЛПД, модели SiteGA.

Полученные результаты по подбору параметров на основе точности моделей имеют те же тенденции, что и результаты, полученные ранее для ТФ FOXA2 (см. раздел 3.1). Для моделей PWM, BaMM и SiteGA мода длины

мотива составила 16 п.о., 8 п.о. и 12 п.о., при этом для всех моделей длина мотива в 20 п.о. является самой редкой (Рисунок 26А). В изменении порядка ММ для BaMM наблюдаются следующая зависимость (Рисунок 26Б), с ростом длины мотива растёт и значения порядка ММ для BaMM. Если на длинах мотива 8 п.о. и 12 п.о. основная доля мотивов имеют порядок ММ равный 1 и 2, то на длинах 16 п.о. и 20 п.о. увеличивается доля мотивов с порядком ММ, равным 3. Можно предположить, что для модели BaMM увеличение точности при росте длины модели сопряжено с ростом порядка цепи ММ. Для модели SiteGA для всех длин мотивов чаще всего достигается максимальное значение числа ЛПД (100) из допустимого диапазона (40 – 100), то есть данная модель показывает тенденцию роста точности при включении в анализ максимально возможного числа параметров (Рисунок 26В), исключением является длина мотива 16 п.о. и 20 п.о., где доля мотивов с ЛПД 80 равна доле мотивов с ЛПД 100.

Далее для того, чтобы сравнить точности разных моделей, построили ROC-кривые для наиболее крупных, по числу экспериментов, классов ТФ: *Basic leucine zipper factors* (bZIP), *Basic helix-loop-helix factors* (bHLH), *C2H2 zinc finger factors*, *Tryptophan cluster factors* (Рисунок 27). В каждом наборе данных ChIP-seq для каждого значения TPR было посчитано значение FPR, далее, при построении ROC-кривой, все значения FPR при фиксированном TPR усреднялись (Рисунок 27). Предполагается, что шкала значений ERR, которая определяет жёсткий, средний и мягкий пороги распознавания (см. раздел 2.5), очень хорошо соответствует шкале FPR, которая определяет ось X на ROC-кривой, FPR ~ ERR. Следовательно, из полученных данных (Рисунок 27) можно заключить, что для представленных классов ТФ (1) на жёстких порогах (ERR ≤ 1E-4) модель SiteGA имеет точность распознавания хуже, чем у PWM и BaMM; (2) на средних порогах (1E-4 < ERR ≤ 2.5E-4) преимущество PWM и BaMM над SiteGA меньше, при этом BaMM немного лучше чем PWM; (3) на мягких порогах (2.5E-4 < ERR ≤ 5E-4) все три модели показывают схожую точность, при этом для всех классов BaMM лучше чем PWM, а SiteGA

на классах *Basic leucine zipper factors (bZIP), Basic helix-loop-helix factors (bHLH), Tryptophan cluster factors* немного лучше, чем PWM; (4) на сверхмягких порогах (5E-4 < FPR ≤ 1E-3) BaMM и SiteGA лучше, чем PWM, исключением является класс *C2H2 zinc finger factors*, где модель SiteGA немного отстаёт по точности.

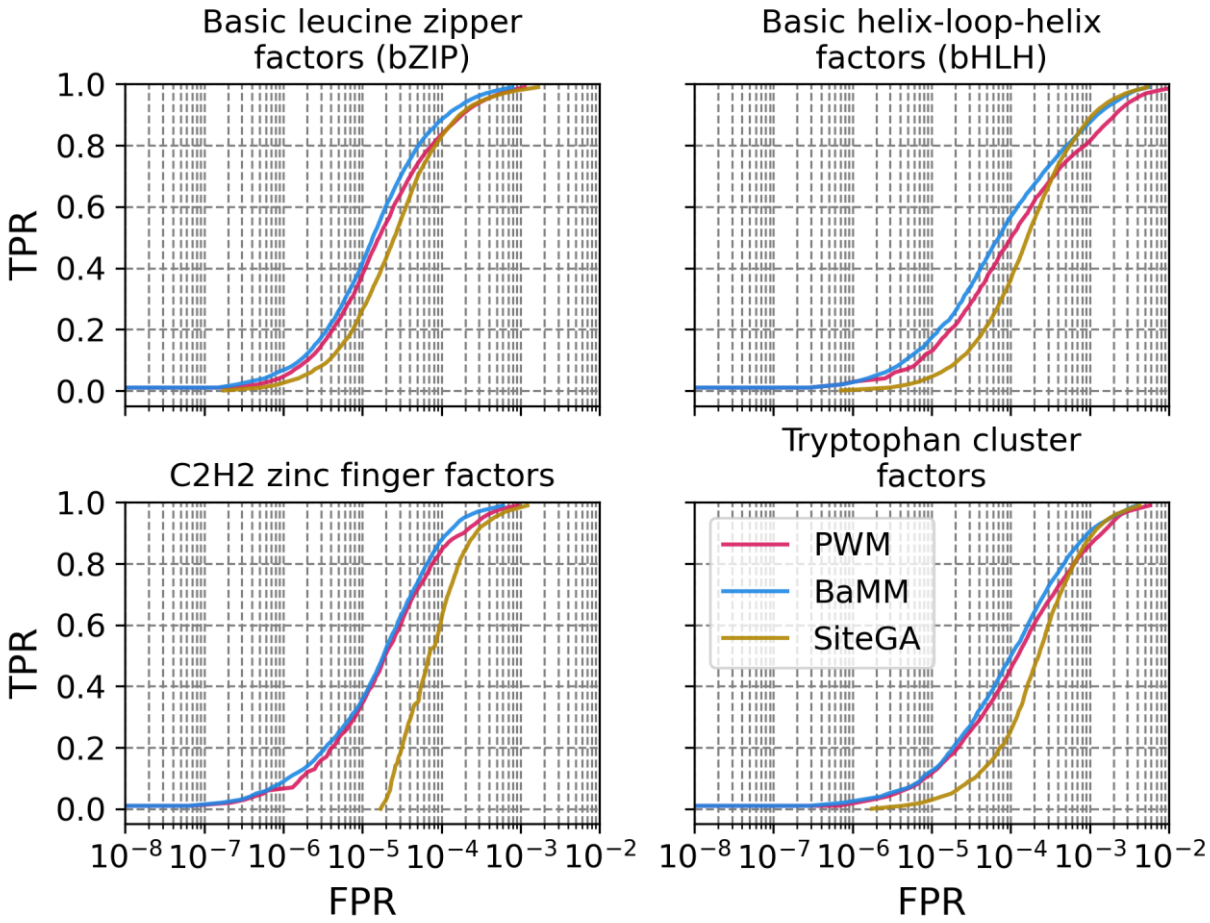

Рисунок 27. Характеристика точности распознавания мотивов моделей PWM, BaMM и SiteGA на данных для *A. thaliana* по четырём классам ТФ: *Basic leucine zipper factors (bZIP), Basic helix-loop-helix factors (bHLH), C2H2 zinc finger factors, Tryptophan cluster factors*. ROC-кривые для моделей, которые были получены с применением 2-fold CV процедуры (см. раздел 2.4). На графиках показаны средние значения FPR (ось X) в зависимости от значений TPR (ось Y).

Далее были рассмотрены распределения точности моделей по показателю pAUC для каждого класса ТФ (Рисунок 28), при этом брали только те классы, для которых количество экспериментов было больше пяти (см. таблица 8).

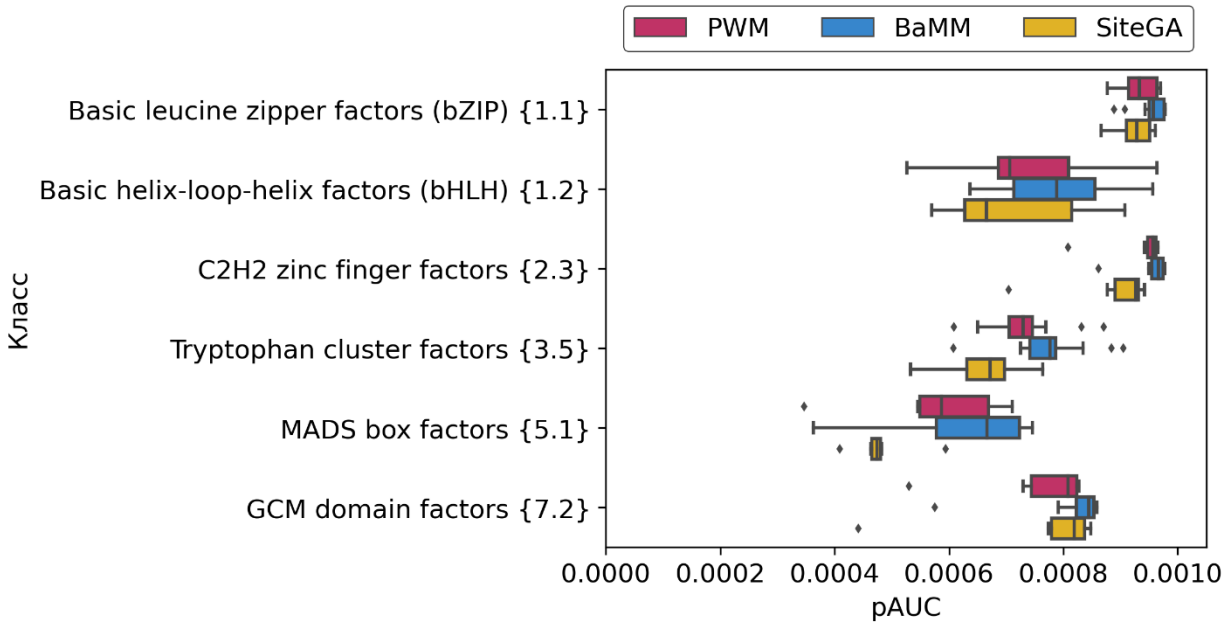

Рисунок 28. Диаграмма размаха с распределениями показателя точности pAUC для трёх моделей (PWM, BaMM и SiteGA) показана отдельно для шести классов ТФ. На диаграмме представлены распределения квартилей Q_1, Q_2 и Q_3 для pAUC. Планки погрешностей ниже Q_1 и выше Q_3 относятся к минимальным/максимальным значениям, если они расположены в пределах 1.5 межквартильных диапазонов (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.

Из приведенных данных (Рисунок 28) видно, что точность моделей может зависеть от структуры ДСД. Все три модели показывают лучшую точность на классах *Basic leucine zipper factors (bZIP)* и *C2H2 zinc finger factors*, а худшую точность на классе *MADS box factors*. Если сравнивать значения медиан, то можно отметить, что во всех случаях модель BaMM точнее моделей PWM и SiteGA. Модель SiteGA имеет наименьшие значения медиан pAUC по сравнению с другими моделями, исключением является *GCM domain factors*, где модель SiteGA показывает более высокую точность по сравнению с PWM. Возможно, в данном классе зависимости между нуклеотидами играют больше роли чем консервативность нуклеотидов в отдельных позициях.

В целом, эффективность моделей зависит от выбора порога: (1) на жёстких или средних порогах эффективно работают модели стандартная PWM или альтернативная BaMM, которая является расширением PWM, но (2) на мягких и сверхмягких порогах лучшую эффективность показывают обе

альтернативных модели BaMM и SiteGA. Другим фактором, который существенно влияет на точность это класс ТФ по структуре ДСД.

Для более детального анализа были рассмотрены три набора данных ChIP-seq для ТФ CCA1, MYC2 и SEP3 (GTRD ID PEAKS042882, PEAKS058394 и PEAKS042820, соответственно). Для всех трёх ТФ, *de novo* модели PWM, BaMM и SiteGA предсказывали сайты значимо похожие (сходство мотивов, $p < 0.05$) на известные мотивы целевых ТФ из базы данных CIS-BP. На рисунке 29 показаны ROC-кривые для трёх наборов данных (панели A-B). На рисунке 29Г сравнивается показатель точности pAUC для моделей. Во всех трёх наборах данных модели PWM и BaMM обладают почти одинаковой точностью, превосходящей точность модели SiteGA. Несмотря на то, что модель SiteGA имеет наименьшую точность среди трёх моделей в диапазоне жёстких (FPR ≤ 1E-4) и средних (1E-4 < FPR ≤ 2.5E-4) порогов, модели PWM и SiteGA имеют схожую точность на мягких (2.5E-4 < FPR ≤ 5E-4) и сверхмягких (5E-4 < FPR ≤ 1E-3) порогах. Данный результат хорошо согласуется с приведёнными выше усреднёнными ROC-кривыми (см. Рисунок 27).

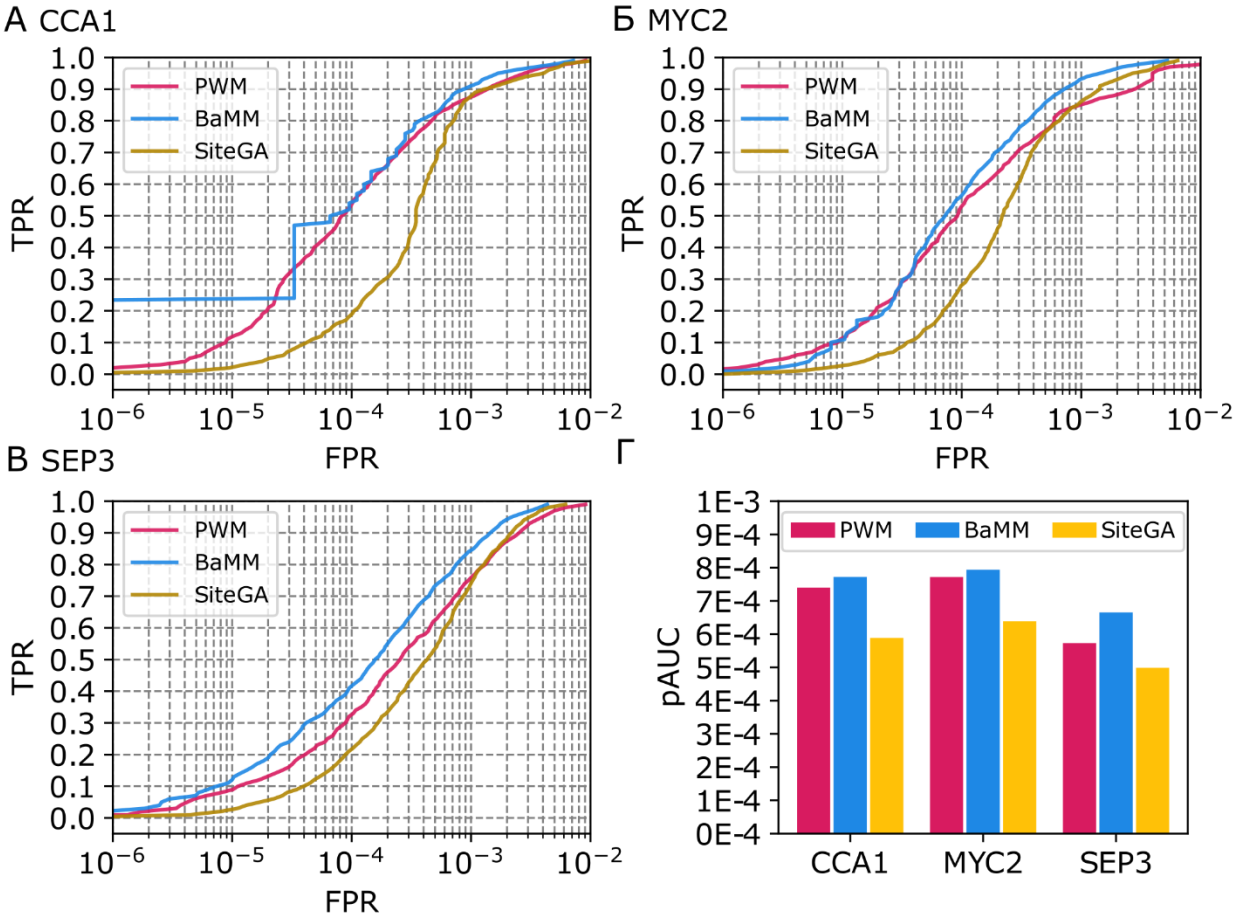

Рисунок 29. Характеристики точности распознавания мотивов моделей PWM, BaMM и SiteGA на трёх примерах ChIP-seq данных. **(А-В)** ROC-кривые для наборов данных ТФ CCA1, MYC2 и SEP3, которые были получены с применением процедуры перекрёстной проверки (см. раздел 2.4). **(Г)** Сравнение средних характеристик точности (pAUC) для трёх моделей и трёх наборов данных.

3.2.4 Сравнение структуры мотивов, распознаваемых разными моделями для одних данных обучения

Представленные выше оценки точности распознавания характеризуют то, насколько точно разные модели способны отличать функциональные сайты от нефункциональных. Однако для того, чтобы понять, могут ли разные модели дополнять друг друга, то есть они распознают сайты с различной структурой, необходимо визуализировать характеристики нуклеотидного контекста, определяющие разные мотивы, например, в виде лого-диаграмм. Однако такой традиционный способ визуализации мотива в виде лого-диаграммы [229] не способен показать зависимости между частотами нуклеотидов в разных позициях мотива. Поэтому для того, чтобы визуально

представить зависимости совместно с традиционной лого-диаграммой была использован инструмент DepLogo [222]. DepLogo берёт на вход выравненные СС и определяет зависимости появления нуклеотидов в различных позициях сайтов через вычисление взаимной информации, после чего изображает эти зависимости в виде треугольной матрицы, где оттенками серого цвета отображается взаимная информация (Рисунок 30). Метод, реализованный в программе DepLogo, последовательно разбивает выравнивание на подмножества в соответствии с информационным содержанием (взаимной информацией) зависимостей между позициями. Результат кластеризации показывает полученные подмножества сайтов, где внутри каждого подмножества совпадения нуклеотидов отмечены цветными прямоугольниками, так что множественные парные зависимости создают общий паттерн для выравнивания СС (Рисунок 30). Под прямоугольниками так же изображается и традиционная лого-диаграмма для выравнивания последовательностей, которая объясняет положение нуклеотидов (Рисунок 30).

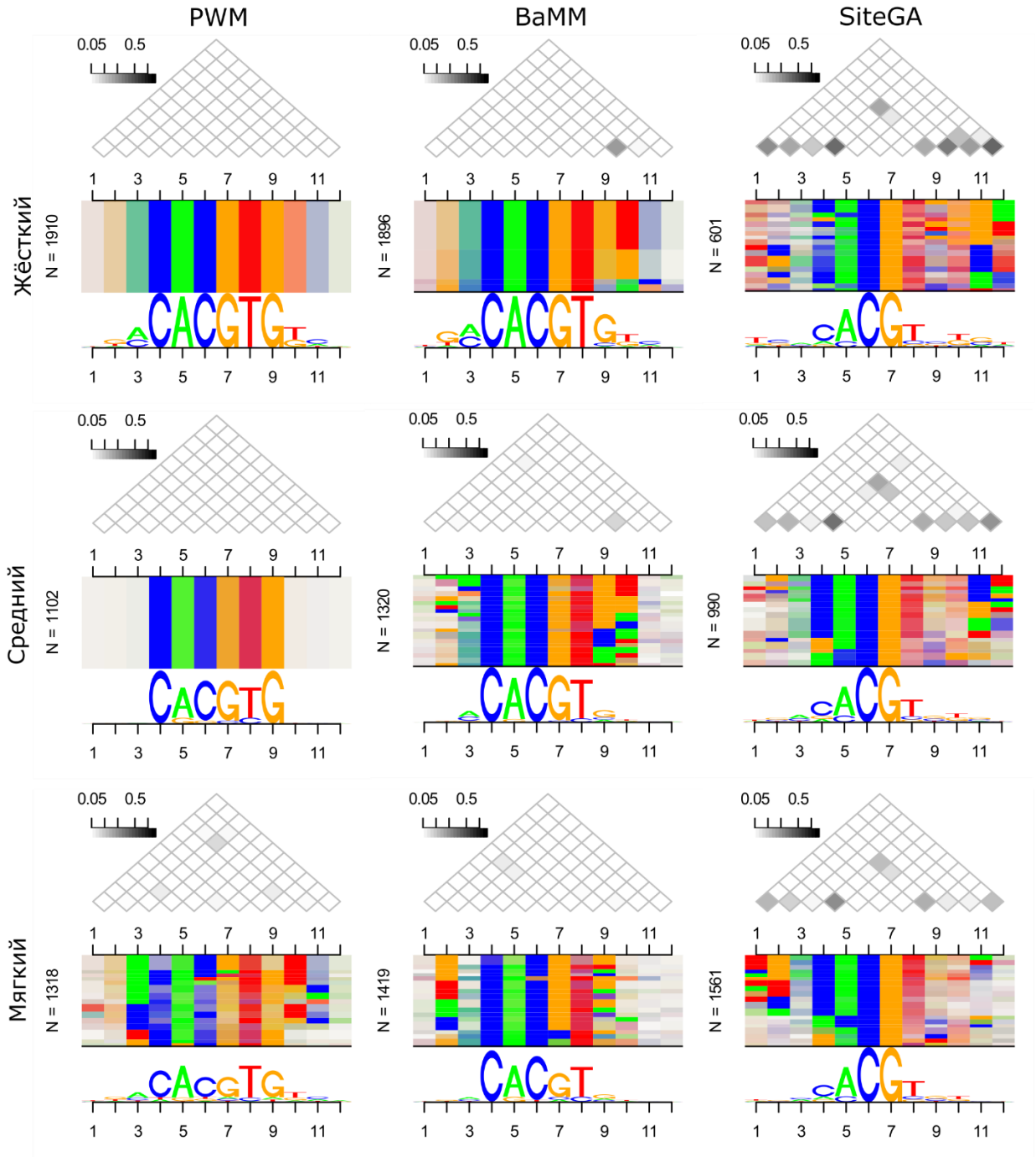

Рисунок 30. Визуализация мотивов разных моделей с помощью инструмента DepLogo. В анализ включены выравнивания предсказанных ССТФ по пикам ChIP-seq для ТФ ABF3 (GTRD ID PEAKS042901, GEO ID GSM2130975, GSM2130977, GSM2130979, трёхдневные проростки). Показаны данные для разных диапазонов порогов: жёстких (ERR ≤ 1E-4), средних (1E-4 < ERR ≤ 2.5E-4) и мягких (2.5E-4 < ERR ≤ 5E-4). В каждом столбце представлены результаты для моделей PWM, BaMM и SiteGA. В каждой из ячеек таблицы 3x3 изображено лого DepLogo [222], которое включает традиционное и альтернативное лого мотива (нижняя и средняя части изображения) и расшифровку альтернативного лого в виде треугольной матрицы (верхняя часть изображения), в которой цветом (оттенки серого) обозначена взаимная информация – мера зависимости нуклеотидов.

Поскольку в предыдущем разделе были показаны особенности поведения ROC-кривых, связанные с величиной порога, необходимо было рассмотреть результаты визуализации DepLogo, полученные на разных порогах распознавания. Чтобы детально визуализировать мотивы с предполагаемой разной аффинностью связывания, для каждой модели мы составили список предсказанных ССТФ во всех пиках и отсортировали их в порядке возрастания ожидаемой частоты распознавания ERR. Затем разделили мотивы на диапазоны ERR с жёстким, средним и мягким порогами распознавания ERR ≤ 1E-4, 1E-4 < ERR ≤ 2.5E-4 и 2.5E-4 < ERR ≤ 5E-4, соответственно. На рисунке 30 представлены традиционные и альтернативные лого мотивов для моделей PWM, BaMM и SiteGA на данных для ТФ ABF3 (GTRD ID PEAKS042901, GEO ID GSM2130975, GSM2130977, GSM2130979, трёхдневные проростки), на трёх диапазонах порога распознавания – жёстком (ERR ≤ 1E-4), среднем (1E-4 < ERR ≤ 2.5E-4) и мягком (2.5E-4 < ERR ≤ 5E-4). Традиционные лого-диаграммы подтверждают, что по крайней мере часть наиболее консервативной последовательности мотива (далее — ядро) выявляется для всех моделей и для всех диапазонов порогов распознавания. Модели PWM, BaMM и, в несколько меньшей степени, модель SiteGA, как правило, сохраняют одни и те же нуклеотиды в ядре на всех диапазонах порога.

Для модели PWM на жёстком пороге консервативность нуклеотидов внутри ядра (большая высота букв в лого-диаграмме) является высокой, однако с уменьшением порога консервативность внутри ядра снижается равномерно. Также стоит отметить, что модель PWM на жёстком и среднем порогах не показывает никаких зависимостей, все треугольные матрицы на DepLogo (Рисунок 30) имеют чистый белый цвет. На мягком пороге DepLogo показывает, что PWM выявляет три слабые зависимости, что видно из треугольной матрицы. Возможно, это случайные зависимости, которые связаны с уменьшением консервативности ядра, поскольку модель PWM не способна учитывать зависимости.

Модель BaMM, как и модель PWM, на жёстком пороге имеет высокую консервативность нуклеотидов ядра, но в отличие от PWM, с уменьшением порога эта консервативность ядра уменьшается неравномерно, поскольку часть нуклеотидов сохраняется, а часть заменяется на другие. Лучше всего данный эффект наблюдается в 9-ой позиции, где на жёстком пороге в данной позиции находится нуклеотид G (вертикальная полоса почти вся оранжевая), а на среднем пороге в данной позиции нуклеотид G стоит в чуть больше чем в половине случаев (вертикальная полоса на половину оранжевая), остальная часть – нуклеотиды C (синий цвет в вертикальной полосе), A (зелёный цвет в полосе) и в незначительной доле случаев T (красный цвет в полосе). На мягком пороге модель BaMM в некоторых позициях имеет более высокую консервативность по сравнению с PWM. На всех порогах модель BaMM выявляет зависимости, однако их интенсивность снижается с уменьшением порога, что так же хорошо видно из треугольной матрицы для 9-ой и 10-ой позиций. Возможно, благодаря тому, что модель BaMM учитывает зависимости, для ядра мотива поддерживается более высокая консервативность даже на мягком пороге.

Модель SiteGA уже на жёстком пороге имеет несколько более вырожденное ядро мотива (caGTt) по сравнению с моделями PWM (CACGTG) и BaMM (CACGTg), но при этом для данной модели выявляется большое количество зависимостей (12 шт. исходя из треугольной матрицы). С уменьшением величины порога консервативность ядра мотива не изменяется даже на самом мягком пороге, основные изменения происходят на флангах мотива. Но с смягчением порога видно, что ослабевают зависимости и их количество уменьшается, если на жёстком пороге их было 12, то на мягком 9. Такой результат можно интерпретировать как способность модели SiteGA сочетать два фактора: зависимости различных позиций в мотиве и сохранение консервативности мотива с точки зрения модели PWM (см. раздел 1.4.2.3, уравнение (5), факторы E(X) и D(X), соответственно). Возможно, для жёстких порогов для модели SiteGA фактор наличия зависимостей вносит больший

вклад, чем фактор сохранения консервативности нуклеотидного состава мотива.

Далее был проведён массовый анализ с применением DepLogo для всего набора ChIP-seq данных, чтобы оценить количество зависимостей, которые выявляют модели PWM, BaMM и SiteGA (Рисунок 31).

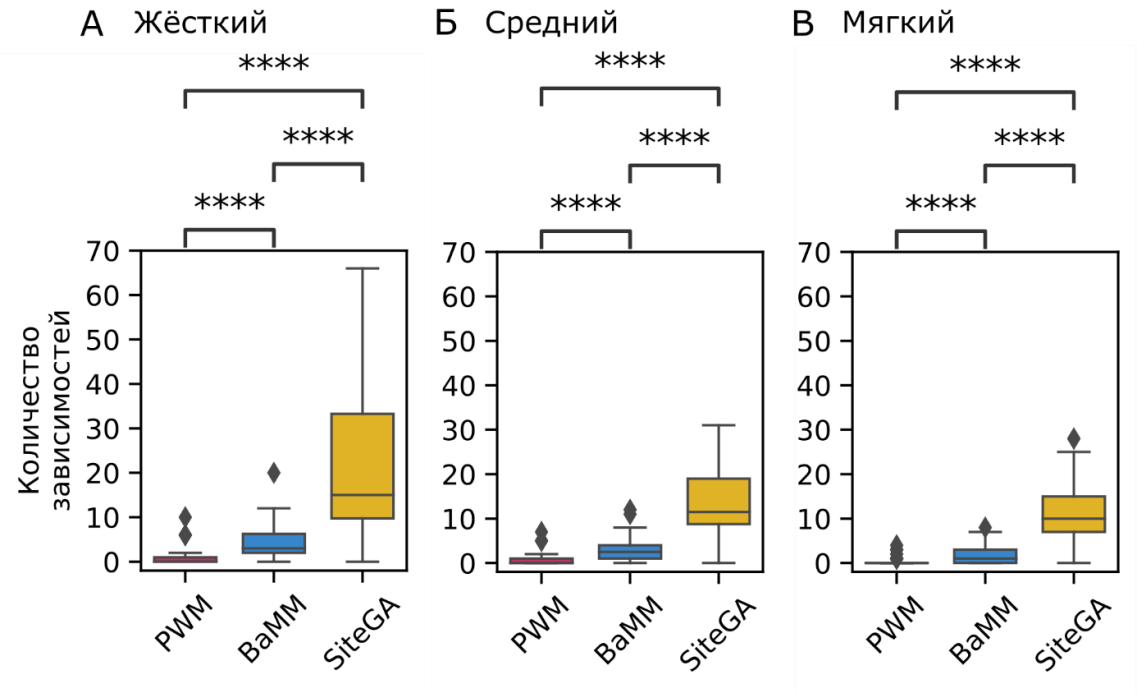

Рисунок 31. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации на выборке данных наборов ChIP-seq для *A. thaliana*. На диаграммах **(А)**, **(Б)**, **(В)** показаны распределения количества зависимостей, полученные на разных порогах: жёстком, среднем и мягком (ERR ≤ 1E-4, ERR ≤ 2.5E-4 и ERR ≤ 5E-4 соответственно) посчитанных с помощью DepLogo с количеством взаимной информации, $p < 0.05$. На диаграммах представлены распределения квартилей Q_1, Q_2 и Q_3 по количеству зависимостей. Планки погрешностей ниже (Q_1) и выше (Q_3) относятся к минимальным/максимальным значениям, если они расположены в пределах полутора межквартильных диапазонов (IQR = Q_3−Q_1) от Q_1 и Q_3, соответственно. В противном случае планки погрешностей установлены в положениях {Q_1−1.5*IQR}/{Q_3+1.5*IQR} для квартилей Q_1/Q_3, соответственно. Все значения, которые не попали в пределы планок погрешности отмечены ромбами как выбросы. **** - $p < 0.0001$.

Из приведённых данных (Рисунок 31) видно, что для PWM медиана количества зависимостей составляет 0 на всех порогах, что и должно наблюдаться для PWM, поскольку она не учитывает зависимости. Однако в части данных PWM имеет ограниченное количество зависимостей между

позициями нуклеотидов, что вероятно связано с незначительным уровнем шума в данных. BaMM находит значимо больше ($p < 0.05$ согласно критерию Манна-Уитни) зависимостей по сравнению с моделью PWM, а модель SiteGA, в свою очередь, имеет значимо больше ($p < 0.05$ согласно критерию Манна-Уитни) таких зависимостей по сравнению с BaMM на всех порогах. Также выявляется следующая закономерность, со смягчением порога для моделей BaMM и SiteGA количество выявляемых зависимостей уменьшается. Так для модели BaMM на мягком, среднем и жёстком порогах медиана количества зависимостей составляет 1, 2.5 и 3, соответственно, а для модели SiteGA 10, 11 и 15, соответственно. Можно предположить, что для альтернативных моделей качество мотива, определяемое через порог, связано с количеством зависимостей, которые учитывает модель.

В заключение можно отметить, что визуализация DepLogo объясняет, почему модель BaMM и особенно модель SiteGA могут не иметь, с точки зрения традиционной лого-диаграммы, позиции консервативных нуклеотидов, но в то же время альтернативные модели могут давать дополнительные предсказания по сравнению с моделью PWM. Более того, дополнительная информация о зависимостях между позициями объясняет, почему модели BaMM и SiteGA последовательно конкурируют с PWM по точности распознавания, особенно в пределах мягкого и сверхмягкого диапазонов порогов ERR ($2.5E\text{-}4 < ERR \leq 5E\text{-}4$ и $5E\text{-}4 < ERR \leq 1E\text{-}3$, соответственно).

3.2.5 Сравнение специфики поиска мотивов разными моделями

Далее было изучено сходство/различие значений функций распознаваний в парах моделей, то есть как разные модели оценивают один и тот же сайт по величине порога распознавания (ERR). Были рассмотрены три возможные комбинации моделей: BaMM/PWM, SiteGA/PWM и SiteGA/BaMM. Для каждой комбинации составили список «совпадающих» сайтов, которые распознаются этими двумя моделями; термин «совпадающие»

сайты подразумевает хотя бы частичное перекрытие сайтов. На рисунке 32 показаны тепловые карты количества совпадающих позиций мотивов, предсказанных разными моделями, с соответствующими оценками долей распознавания для различных попарных комбинаций моделей.

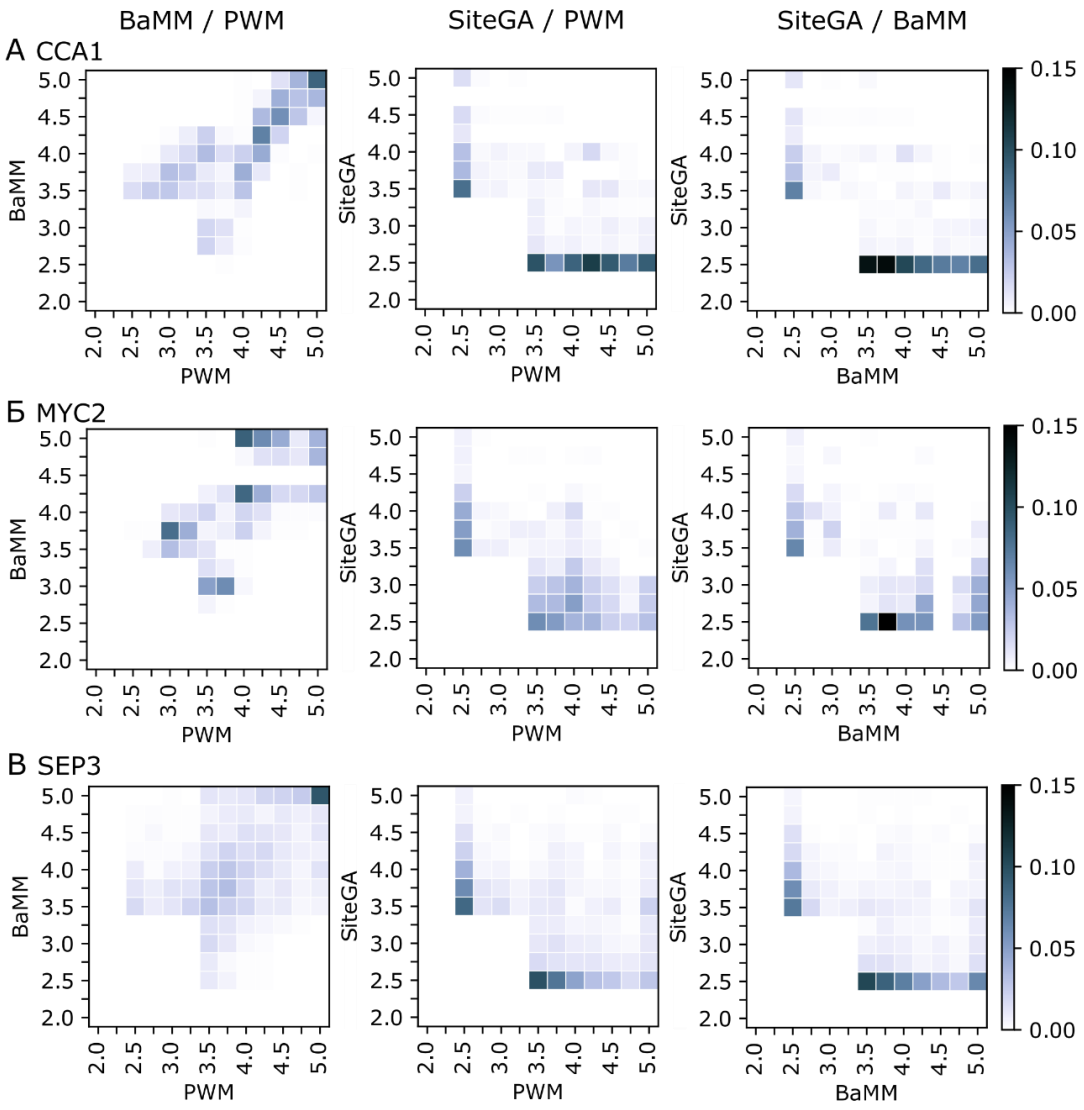

Рисунок 32. ССТФ, обладающие различными оценками аффинности, в разных попарных комбинациях моделей. На панелях **(А)**, **(Б)** и **(В)** показаны тепловые карты, рассчитанные для значений частот распознавания мотивов ERR привязанных к «совпадающим» сайтам, предсказанным моделями для данных CCA1, MYC2 и SEP3 ТФ, соответственно. В строках и столбцах каждой тепловой карты показана оценка частоты встреч мотивов в виде отрицательного десятичного логарифма значения ERR, $-\log_{10}(\mathrm{ERR})$. Цветами разной яркости обозначены доли «совпадающих» сайтов, распознанных моделями с конкретными ERR двух моделей. Максимальное значение доли, равное 1, соответствует полному «совпадению» всех сайтов в паре моделей. В колонках показаны комбинации моделей BaMM/PWM, SiteGA/PWM и SiteGA/BaMM.

Очевидно, что комбинация BaMM/PWM продемонстрировала большее совпадение оценки функции распознавания по сравнению с комбинациями SiteGA/PWM и SiteGA/BaMM. Для пары моделей PWM/BaMM, на примере ТФ CCA1 наблюдается наибольшая схожесть ранжирования значений функций распознавания этих моделей, большая часть «совпадающих» ССТФ лежит на диагональной линии и коэффициент корреляции Пирсона равен 0.86. В случаях для ТФ MYC2 и SEP3, где «совпадающие» ССТФ рассредоточены около диагональной линии, коэффициент корреляции Пирсона равен 0.60 (MYC2) и 0.56 (SEP3), что говорит о различии ранжирования значений функций распознавания моделей PWM и BaMM. Для пар SiteGA/PWM и SiteGA/BaMM во всех случаях не наблюдается корреляции значений функций распознавания в парах моделей, например, если модель SiteGA оценивает сайт как высокоаффинный, то PWM и BaMM как низкоаффинный, и наоборот. Это также подтверждается значениями коэффициента корреляции Пирсона. Для SiteGA/PWM для ТФ CCA1, MYC2 и SEP3 коэффициенты равны -0.63, -0.60 и -0.39, соответственно, а для SiteGA/BaMM для ТФ CCA1, MYC2 и SEP3 коэффициенты равны -0.54, -0.51 и -0.44, соответственно.

Чтобы массово сравнить функции распознавания для «совпадающих» сайтов были посчитаны коэффициенты корреляции для каждого набора ChIP-seq данных и для каждой комбинации моделей. По полученным значениям построили гистограммы распределений коэффициентов корреляций (Рисунок 33).

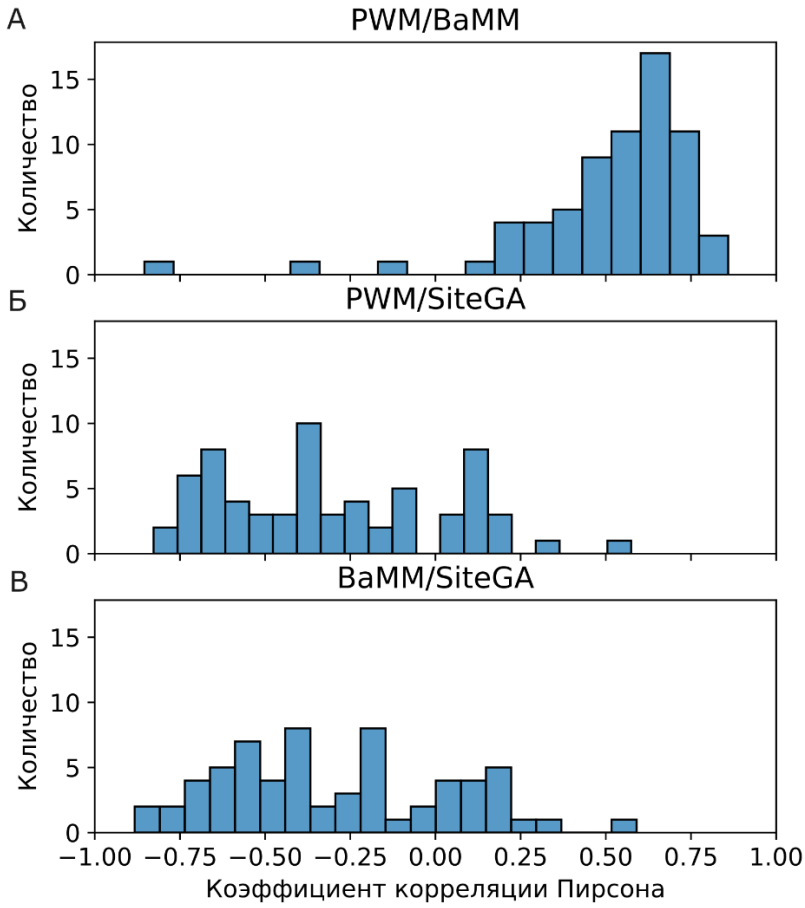

Рисунок 33. Распределение коэффициентов корреляции Пирсона по всей коллекции данных ChIP-seq для *A. thaliana*. Коэффициенты корреляции посчитаны по значениям функций распределения в парах моделей для «совпадающих» мотивов. **(А)** BaMM/PWM; **(Б)** PWM/ SiteGA; **(В)** BaMM/SiteGA

Результаты массового анализа показывают те же закономерности, которые были получены для ТФ CCA1, MYC2 и SEP3. Полученное для моделей PWM и BaMM распределение коэффициентов корреляции имеет практически куполообразную форму, а медиана составляет 0.58. Следовательно для данной пары моделей значения функции распознавания в значительной степени скоррелированы между собой. Тем не менее значение медианы далеко от единицы и, следовательно, данные модели для части ССТФ могут предсказывать разную аффинность. Можно предположить, что эти модели обеспечивают лишь умеренно схожие оценки аффинности ССТФ.

Распределения, полученные для пар PWM/SiteGA и BaMM/SiteGA не имеют выраженной куполообразной формы, для значительной части данных коэффициенты корреляции находятся в диапазоне от -0.85 до 0, однако часть

данных имеют и положительные значения коэффициентов корреляции. Тем не менее, в большинстве случаев для пар PWM/SiteGA и BaMM/SiteGA коэффициенты корреляции являются отрицательными, и медианы равны -0.34 и -0.28, соответственно. Следовательно, можно предположить, что модель SiteGA будет чаще выявлять отличные от других моделей ССТФ, поскольку она иначе оценивает аффинность ССТФ, возможно это связано с тем, что SiteGA в первую очередь оценивает зависимости позиций, а не консервативность нуклеотидов в позициях.

3.2.6 Совместное применение моделей PWM, BaMM и SiteGA для поиска ССТФ

Выше было показано, что мотивы, предсказываемые моделями PWM, BaMM и SiteGA, имеют различную структуру с точки зрения DepLogo, а также показали, что модели по-разному оценивают информационное содержание мотивов. Следовательно, можно ожидать, что модели будут распознавать сайты в разных пиках, и результаты распознавания будут пересекаться лишь частично, а совместное применение моделей увеличит долю пиков с распознанными сайтами. Для того чтобы это проверить, объединили результаты распознавания всех трёх моделей на разных порогах. На рисунке 34 для всех наборов данных ChIP-seq представлены результаты распознавания сайтов на разных порогах: жёстком, среднем и мягком, в виде распределения доли пиков.

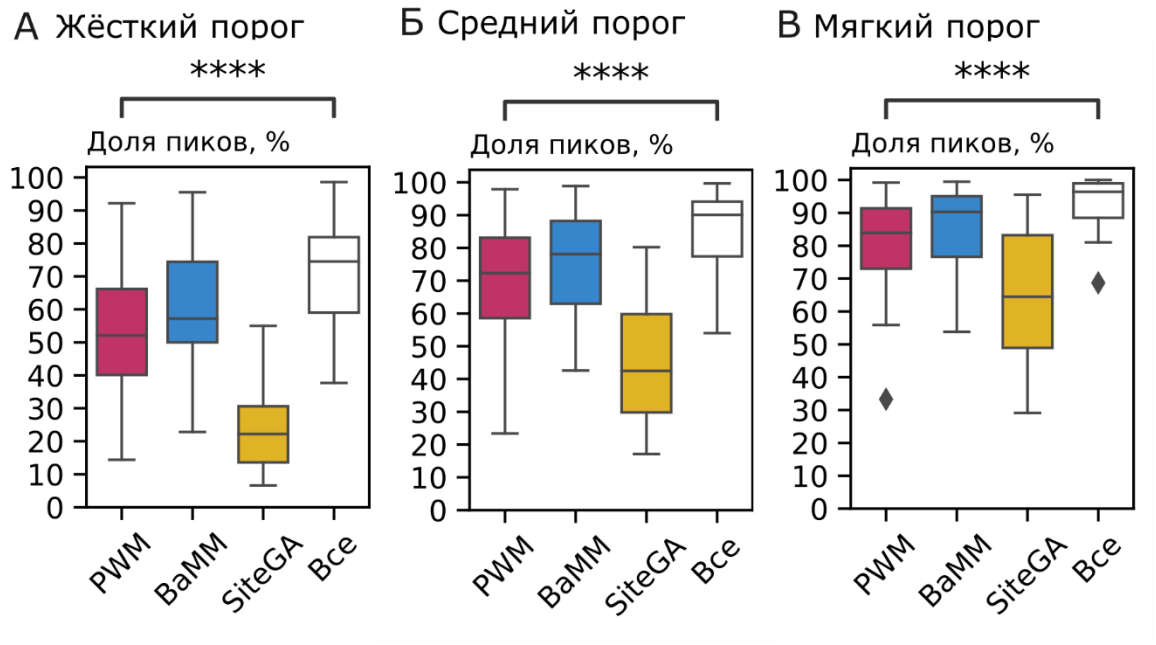

Рисунок 34. Сравнение результатов применения моделей PWM, BaMM и SiteGA по отдельности и объединение результатов моделей (Все) на коллекции данных ChIP-seq экспериментов для *A. thaliana*. На диаграммах **(А)**, **(Б)**, **(В)** показаны распределения долей пиков, полученные на разных порогах: жёстком, среднем и мягком (ERR ≤ 1E-4, ERR ≤ 2.5E-4 и ERR ≤ 5E-4, соответственно). Каждая диаграмма показывает распределение долей пиков, содержащих сайты, предсказанные отдельными моделями мотива (красный, синий и желтый столбцы означают модели PWM, BaMM и SiteGA, соответственно), и доли пиков, содержащие сайты, предсказанных по крайней мере одной из трёх моделей (белые прямоугольники, Все). На диаграммах представлены распределения квартилей Q_1, Q_2 и Q_3 по долям пиков с сайтами. Планки погрешностей ниже (Q_1) и выше (Q_3) относятся к минимальным/максимальным значениям, если они расположены в пределах полутора межквартильных диапазонов (IQR = Q_3−Q_1) от Q_1 и Q_3, соответственно. В противном случае планки погрешностей установлены в положениях {Q_1−1.5*IQR}/{Q_3+1.5*IQR} для квартилей Q_1/Q_3, соответственно. Все значения, которые не попали в пределы планок погрешности отмечены как выбросы. **** - $p < 0.0001$

Из полученных распределений (Рисунок 34) видно, что (а) на всех порогах доли пиков с сайтами моделей мотива PWM и BaMM очень близки, но для BaMM эта доля выше, (б) на жёстком пороге доля пиков с сайтами модели мотива SiteGA более чем в два раза меньше по сравнению с остальными моделями (значения медиан для долей PWM, BaMM и SiteGA составляют: 52.1%, 57.2% и 22.2%, соответственно). При этом по сравнению с моделями PWM/BaMM, модель SiteGA показывает наиболее заметный рост

доли пиков с сайтами при переходе от жёсткого порога к среднему (медианы возрастают до 72.3%, 78.1% и 42.5% для PWM, BaMM и SiteGA, соответственно) и от среднего к мягкому порогу (83.9%, 90.3% и 64.5% для PWM, BaMM и SiteGA, соответственно). Это означает, что (а) большинство сайтов, предсказанных моделью PWM, обладают «высокой консервативностью», в то время как большинство сайтов, предсказанных моделью SiteGA, демонстрируют «низкую консервативность»; (б) сайты, предсказанные моделью BaMM, скорее всего, имеют двойную природу; для жёстких/средних порогов они аналогичны предсказанным сайтам модели PWM; в то время как для мягкого порога BaMM предсказывает сайты с «низкой консервативностью», имеющие зависимости между позициями.

Независимо от выбора порога совместное применение трёх моделей добавляет около 17% к результатам предсказаний модели PWM (22.4%, 17.8% и 12.5% для жёсткого, среднего и мягкого порогов, соответственно, Рисунок 34).

Поскольку связывание ТФ с ДНК определяется ДСД, а структура ДСД у разных ТФ может значительно отличаться, то было предположено, что вклад альтернативных моделей при распознавании ССТФ к модели PWM может зависеть от структуры ДСД. Всего в коллекции ChIP-seq данных есть эксперименты для 68 ТФ, которые относятся к восьми классам по структуре ДСД, но для того, чтобы делать корректные выводы в дальнейший анализ были взяты только те классы, для которых количество экспериментов ChIP-seq было больше 5 (Таблица 8). Результаты по распознаванию ССТФ на жёстком пороге (ERR ≤ 1E-4) моделями мотива PWM, BaMM и SiteGA для классов, прошедших фильтрацию, представлены в таблице 9.

Из полученных результатов (Таблица 9) видно, что для большинства классов ТФ альтернативные модели значительно расширяют результаты, полученные моделью PWM. Особенно выделяются ТФ, относящиеся к классам *MADS box factors* и *GCM domain factors*, где альтернативные модели

увеличивают долю пиков, содержащих ССТФ, по отношению к доле пиков с предсказанными ССТФ моделью PWM, на 20.25% и 27.5%, соответственно.

Таблица 9. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации на жёстком пороге (ERR ≤ 1E-4) на наборе данных ChIP-seq экспериментов для наиболее представительных классов ТФ. В ячейках записаны значения медиан долей пиков с предсказанными ССТФ (в %).

Индекс	Класс	PWM	BaMM	SiteGA	Все*	Все*–PWM
{1.1}	*Basic leucine zipper factors (bZIP)*	81.7	86.9	34.5	91.3	9.6
{1.2}	*Basic helix-loop-helix factors (bHLH)*	51.5	58.6	18.8	65	13.5
{2.3}	*C2H2 zinc finger factors*	92.8	96.5	45.8	98.7	5.9
{3.5}	*Tryptophan cluster factors*	51.3	52.95	20	64.1	12.8
{5.1}	*MADS box factors*	39	48.35	8.3	59.25	20.25
{7.2}	*GCM domain factors*	48.25	55.4	26.5	75.75	27.5

Примечание. *Все – доли пиков, содержащих мотивы, по крайней мере одной из трёх моделей; **Все-PWM – доля пиков с мотивами BaMM или SiteGA, но без мотивов PWM.

У растений ТФ, относящиеся к классу *MADS box factors*, выполняют важные функции в процессе развития разных тканей: корень, цветок и плод [230]. Существует два типа генов *MADS box factors*, называемых типом I и типом II, и у растений эти типы различаются экзон-интронной и доменной структурой белка, скоростью эволюции и функцией развития. Гены типа I далее подразделяются на три группы - M-альфа, M-бета и M-гамма, а гены типа II подразделяются на группы MIKCC и MIKC*. Считается, что разные типы генов *MADS box factors* отвечают за разные биологические функции; в то время как гены типа I могут преимущественно способствовать развитию женского гаметофита, зародыша и семян, а гены группы MIKC* - развитию мужского гаметофита, гены группы MIKCC стали важными для различных аспектов развития спорофита [230]. Другим важным аспектом работы ТФ класса *MADS box factors*, заключается в том, что они формируют гетеродимеры включающие от двух до четырёх ТФ из того же класса [231, 232]. В текущей выборке ТФ в основном представлены гены, связанные с развитием цветка.

Все представители класса *GCM domain factors* в данной работе относятся к семейству WRKY, который есть только у растений [225]. ТФ из данного

семейства выполняют разнообразные биологические функции в отношении устойчивости растений к болезням, реакции на абиотический стресс, дефицит питательных веществ, старения, развития семян и трихом, эмбриогенеза [233]. WRKY могут действовать как активаторы или репрессоры транскрипции и также способны формировать различные комбинации гомо- и гетеродимеров из ТФ, входящих в это же семейство [233].

Для класса *C2H2 zinc finger factors* (Таблица 9), альтернативные модели лишь незначительно расширяют результаты PWM. Доля пиков с сайтами для этого класса ТФ увеличивается на 5.9% за счёт применения альтернативных моделей.

Класс *C2H2 zinc finger factors* является самым крупным по числу ТФ в сравнении с остальными классами для всех эукариотических организмов [234]. ТФ из данного класса связаны с регулированием биологических процессов во время вегетативного роста, репродуктивного развития, а также участвуют в ответе на разные типы стресса: солевой, окислительный, осмотический, холодовой и др. [235, 236]. В частности, ZAT6 связан с устойчивостью к кадмию [237] и солевой устойчивостью [238]. В отличие от других классов, ТФ класса *C2H2 zinc finger factors* специфически связываются с длинными последовательностями ДНК, достигающими нескольких десятков пар оснований. Так же они могут эффективно связываться с ДНК в виде мономеров, в отличие от большинства других ТФ, которые связываются с короткими палиндромными последовательностями в виде гомо- или гетеродимеров. Тем не менее некоторые представители из данного класса также способны формировать как гомодимеры, так и гетеродимеры [234].

Наличие существенного различия вклада альтернативных моделей в распознавание сайтов в пиках для разных классов возможно связано с тем, что часть классов допускает разнообразие структурных вариантов ССТФ (например, за счёт димеризации ТФ или особенностей ДСД). Поэтому альтернативные модели, которые учитывают зависимости позиций, лучше представляют это разнообразие структур ССТФ, чем это делает модель PWM,

и таким образом могут предсказать ССТФ, которые PWM не распознаёт. С другой стороны, для класса *C2H2 zinc finger factors* модель PWM предсказывает сайты практически по всех пиках из-за чего вклад альтернативных моделей незначительный. Поэтому предположение модели PWM о независимости позиций для класса *C2H2 zinc finger factors* более оправдано, чем для других классов ТФ. Это может быть связано с тем, что ТФ класса *C2H2 zinc finger factors* преимущественно связываются как мономеры и имеют очень длинные ССТФ, в сравнении с другими классами ТФ, что даёт возможность иметь консервативные сайты, которые хорошо описывает PWM.

Для того чтобы лучше понять каков вклад каждой модели в распознавание сайтов в пиках, была сделана классификация пиков, которая позволяет установить, в каких пиках содержатся сайты, предсказанные только одной из моделей, двух моделей (всех сочетаний пар), или трёх моделей. На рисунке 35 представлены результаты по классификации пиков для шести классов ТФ (*Basic helix-loop-helix factors (bHLH), Basic leucine zipper factors (bZIP), C2H2 zinc finger factors, GCM domain factors, MADS box factors* и *Tryptophan cluster factors*) в виде диаграмм Венна. На диаграммах Венна (Рисунок 35) показаны медианы долей пиков, с предсказанными ССТФ как отдельными моделями, так и всеми возможными комбинациями двух или трёх моделей.

Из полученных результатов можно отметить, что: (1) большинство пиков содержат сайты моделей мотива PWM и BaMM; (2) для классов *Basic leucine zipper factors (bZIP)* и *C2H2 zinc finger factors* все три модели в большей степени имеют мотивы в одних и тех же пиках (большая доля пиков с мотивами трёх моделей), а доли пиков с сайтами одной из моделей мотива (далее «уникальные» пики) очень малы и варьируют от 0.2% до 1.9%; (3) для остальных классов доля «уникальных» пиков варьирует в диапазоне от 1.5% до 10.4%, при этом вклад каждой модели отличается в зависимости от класса ТФ. Наибольший независимый вклад моделей («уникальная» доля пиков)

наблюдается для класса *MADS box factors*, а наименьший для класса *C2H2 zinc finger factors*.

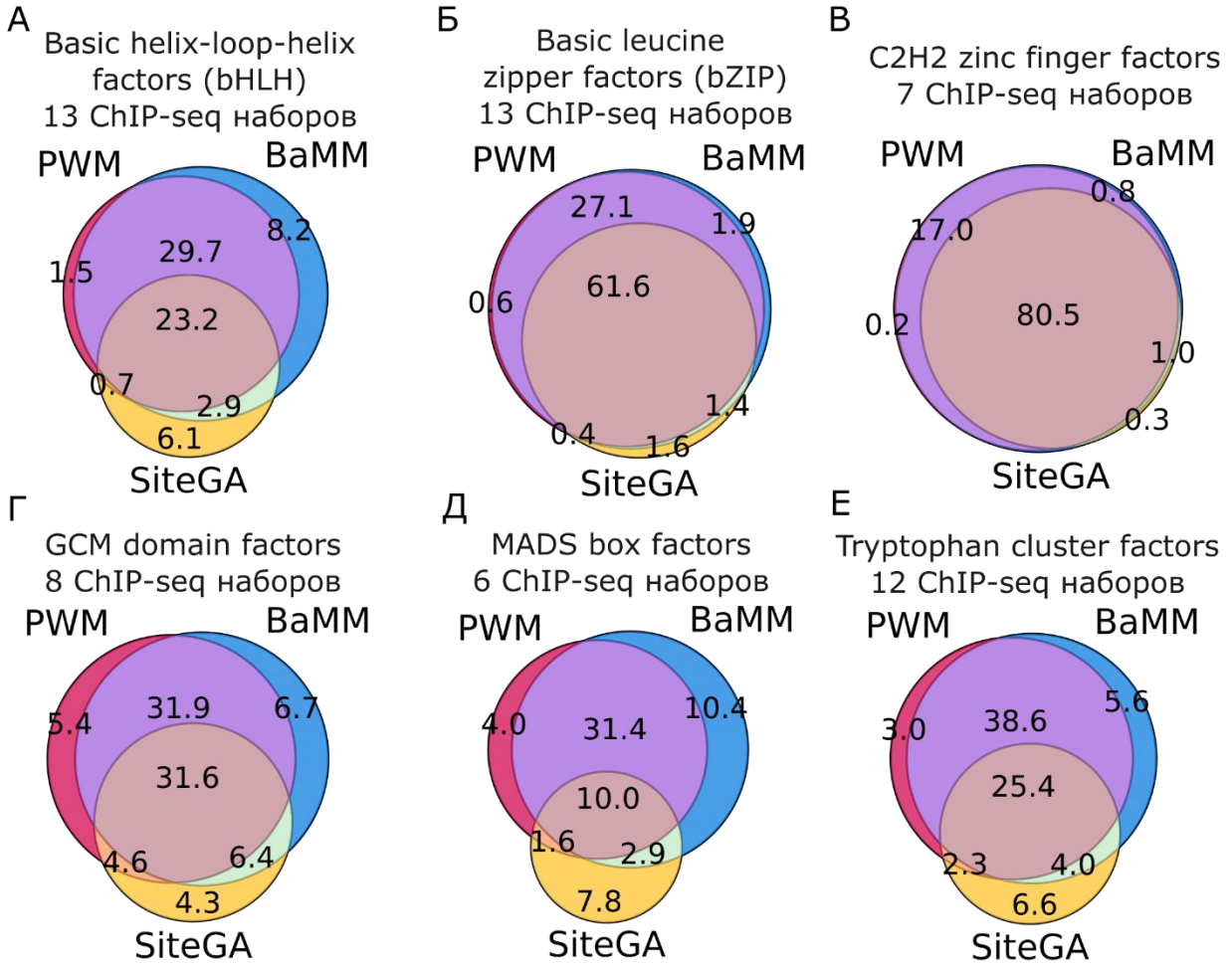

Рисунок 35. Диаграммы Венна для классификации долей пиков, содержащих сайты моделей мотива PWM, BaMM и SiteGA и их сочетаниями. На диаграммах **(А)**, **(Б)**, **(В)**, **(Г)**, **(Д)** и **(Е)** показаны результаты в виде медиан долей пиков для шести классов ТФ: *Basic helix-loop-helix factors (bHLH)*, *Basic leucine zipper factors (bZIP)*, *C2H2 zinc finger factors*, *GCM domain factors*, *MADS box factors* и *Tryptophan cluster factors*. Анализ проводился на среднем пороге (ERR ≤ 2.5E-4) распознавания.

С объединением результатов предсказаний моделей, так же может увеличиваться ошибка перепредсказания. Для того чтобы показать, что объединение результатов моделей лучше по сравнению с моделью PWM, провели следующие расчёты для всей коллекции ChIP-seq *A. thaliana* по классам ТФ *Basic helix-loop-helix factors (bHLH)*, *Basic leucine zipper factors (bZIP) и C2H2 zinc finger factors* (Рисунок 36). Для каждой из трёх моделей

зафиксировали порог по ошибке перепредсказания (FPR = 1E-4, Рисунок 36А; FPR = 2.5E-4, Рисунок 36Б). После объединения результатов распознавания моделей ошибка перепредсказания увеличилась и находится в диапазоне медиан 1.73E-4 – 2.78E-4 для жёсткого порога (Рисунок 36А) и 4.85E-4 – 7.12E-4 для среднего порога (Рисунок 36Б) в зависимости от класса ТФ. Однако при объединении моделей падает ошибка недопредсказания. Если сдвинуть порог модели PWM на соответствующую величину изменения ошибки недопредсказания (1 - TPR), то можно сравнить результат объединения моделей и PWM (Рисунок 36). В результате, объединение трёх моделей имеет меньшую ошибку перепредсказания, чем PWM, если ошибки недопредсказания PWM и объединения моделей равны (Рисунок 36). Стоит отметить, что для классов *Basic helix-loop-helix factors (bHLH)* и *Basic leucine zipper factors (bZIP)* как на жёстком, так и на мягком порогах значение медианы ошибки перепредсказания для объединения моделей ниже, чем для PWM. Для класса *C2H2 zinc finger factors* ошибки перепредсказания объединения моделей и PWM практически равны между собой, что так же подтверждает, что для класса *C2H2 zinc finger factors*, модель PWM хорошо описывает ССТФ и учёт дополнительных зависимостей минимально улучшает результат.

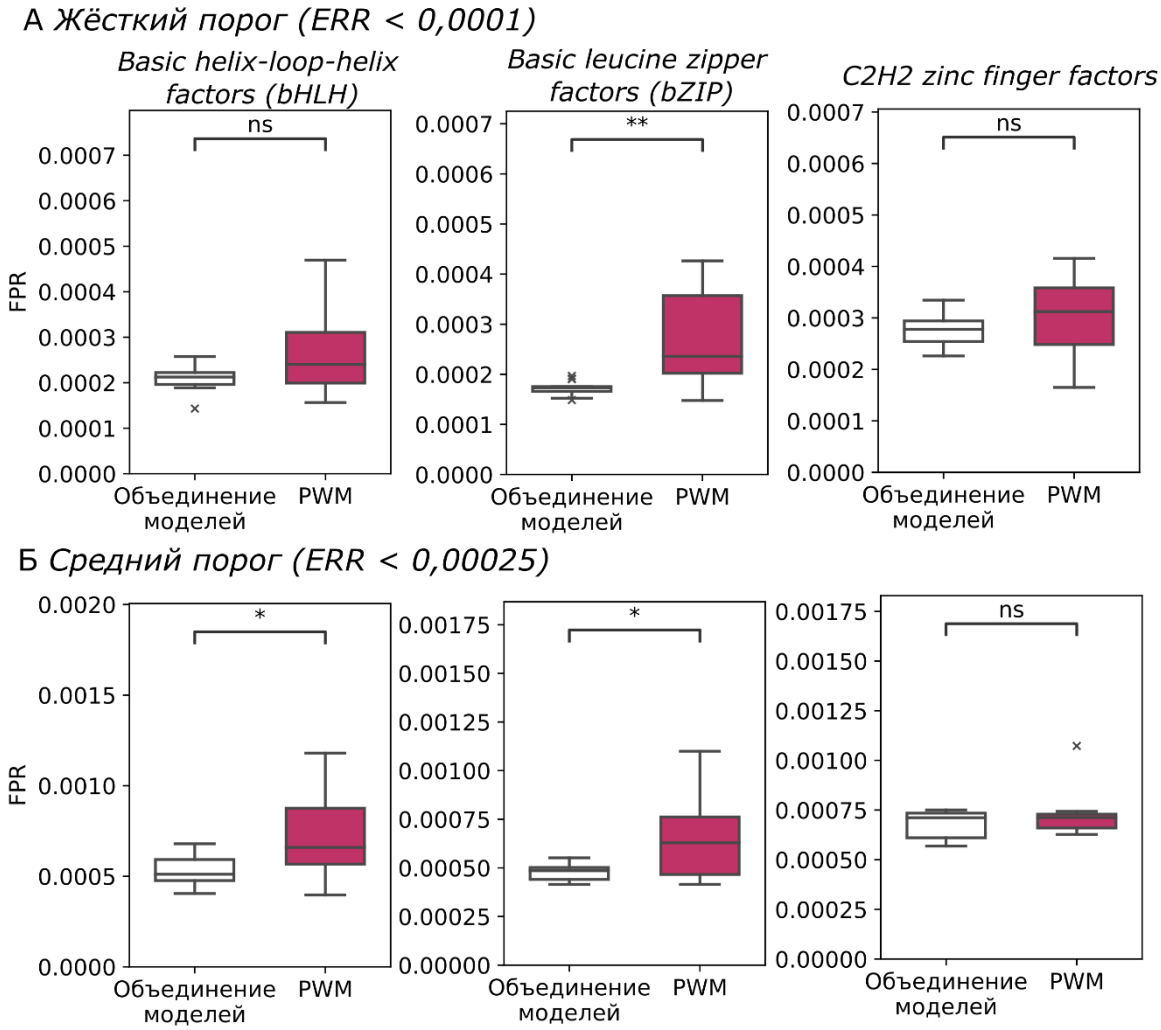

Рисунок 36. Сравнение ошибок перепредсказания для объединения результатов моделей PWM, BaMM и SiteGA и отдельной модели PWM, подсчитанное для одинаковых значений ошибки недопредсказания (1 - TPR) для трёх классов ТФ. На оси Y - FPR, ошибка перепредсказания.

В заключение следует отметить, что совместное применение традиционной и альтернативных моделей мотива увеличило долю пиков, содержащих сайты, по сравнению с применением любой модели по отдельности. Модели BaMM и SiteGA предсказали сайты в части пиков, в которых стандартная модель PWM не находила сайтов. Величины долей пиков, содержащих только сайты одной модели, или определённые сочетания пары моделей, или всех трёх моделей, показывают яркую зависимость от класса ТФ.

3.2.7 Сравнительный анализ списков терминов генной онтологии, полученных путём применения моделей PWM, BaMM и SiteGA

Чтобы выполнить анализ обогащения терминов ГО с помощью всех моделей PWM, BaMM и SiteGA, были распознаны ССТФ в пиках (на среднем пороге ERR ≤ 2.5E-4), при этом если ранее анализ проводили на лучших 1000 пиках, то в данном случае были взяты все пики из каждого набора ChIP-seq данных. Далее сайты разных моделей мотива картировали на 5’-районы генов *A. thaliana* (см. раздел 2.8), в результате для каждой модели были получены списки генов, в промоторах которых есть сайты. Результаты аннотации представлены в виде диаграмм Венна (Рисунок 37), где изображены доли генов (медианные значения по классам ТФ), содержащих в промоторах сайты в разных комбинациях моделей мотива.

Полученные результаты для промоторов (Рисунок 37) имеют те же тенденции, что и ранее представленные результаты по долям пиков (Рисунок 36). Стоит отметить, что существенно выросли доли, где присутствуют предсказания одной модели. Обращает на себя внимание очень существенный рост долей «уникальных» генов для классов *Basic leucine zipper factors (bZIP)* и *C2H2 zinc finger factors*, у которых соответствующие доли, рассчитанные по пикам (Рисунок 37), были очень малы. Тем не менее, класс *C2H2 zinc finger factors* всё ещё отмечается минимальными долями «уникальных» генов (1.8/7.7/3.0 для PWM/BaMM/SiteGA, соответственно). Для этого же класса максимальна доля генов, в которых одновременно все три модели имеют мотивы («пересечение»). Сразу для трёх классов (*GCM domain factors, MADS box factors* и *Tryptophan cluster factors)* суммы долей «уникальных» генов заметно превышают доли генов «пересечения» трёх моделей. С учётом того, что для оставшихся двух классов (*Basic helix-loop-helix factors (bHLH), Basic leucine zipper factors (bZIP))* суммы долей «уникальных» генов приблизительно равны доле «пересечения», мотивы специфичной структуры (в долях «уникальных» генов) составляют заметную долю генов с мотивами всех моделей.

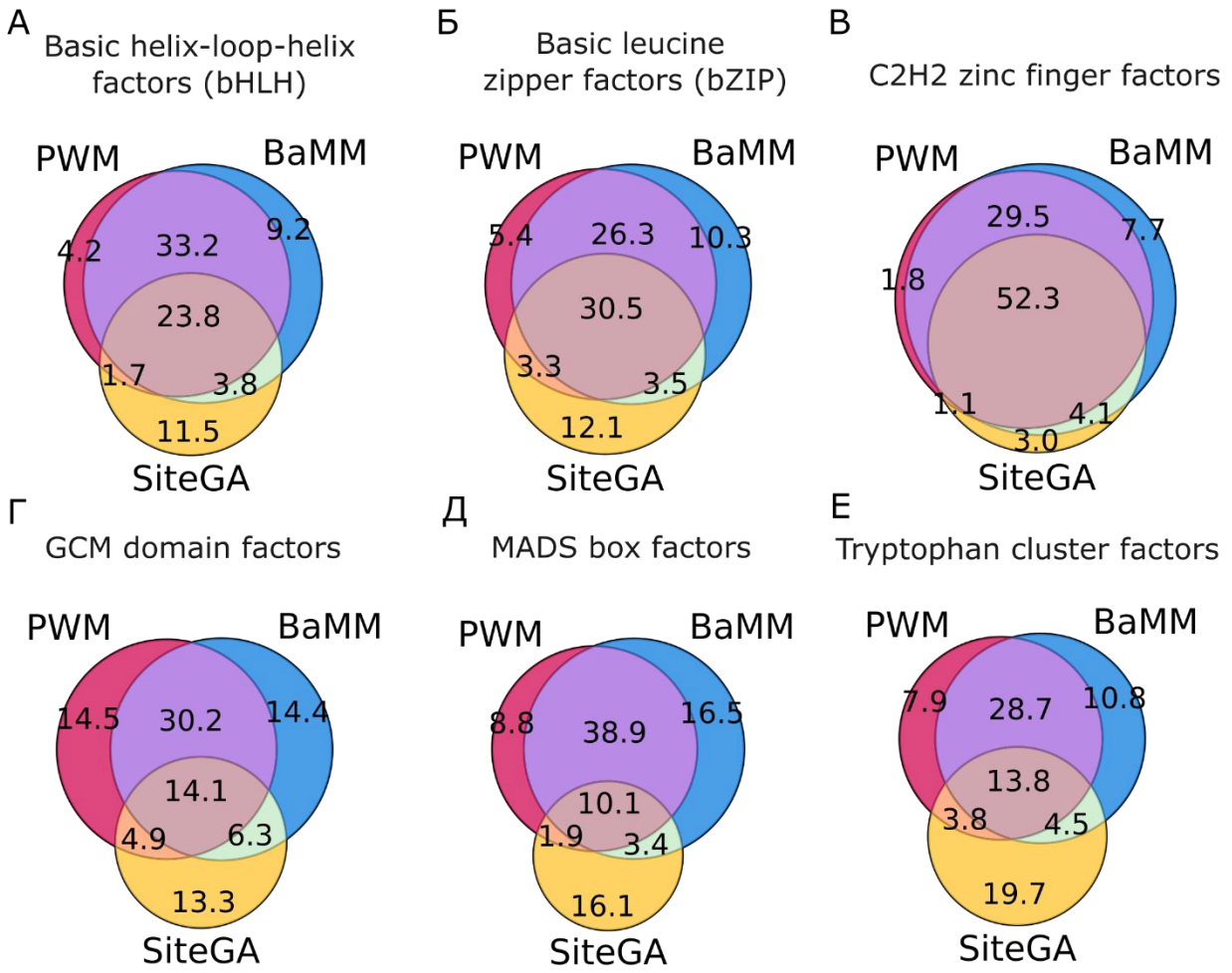

Рисунок 37. Диаграммы Венна для классификации генов, промоторы которых содержат сайты моделей мотива PWM, BaMM и SiteGA. На диаграммах **(А)**, **(Б)**, **(В)**, **(Г)**, **(Д)** и **(Е)** показаны результаты в виде медиан долей генов, содержащих сайты одной, двух (в разных комбинациях моделей) или трёх моделей мотива для шести классов ТФ: *Basic helix-loop-helix factors (bHLH)*, *Basic leucine zipper factors (bZIP)*, *C2H2 zinc finger factors*, *GCM domain factors*, *MADS box factors* и *Tryptophan cluster factors*. Анализ проводился на среднем пороге (ERR ≤ 2.5E-4) распознавания.

Ключевой вывод из результатов состоит в том, что значительная часть генов в промоторах имеет сайты, предсказанные только одной из моделей мотива. Следовательно, можно предположить, что ТФ способен регулировать группу генов, имеющих в промоторах сайты с такой структурой, которая предсказывается только одной моделью. Группа генов, которая регулируется другим структурным вариантом мотива, который распознают альтернативные модели, могут иметь биологические функции, отличные от биологических функций для другой группы генов, где представлен вариант мотива, предсказываемый моделью PWM.

Далее, чтобы это подтвердить, были получены списки обогащённых терминов ГО для биологических процессов, по результатам картирования мотивов моделями PWM, BaMM или SiteGA. Для каждого термина ГО была рассчитана значимость обогащения, скорректированная с учётом множественных сравнений (p_adj) и кратность изменения (англ. fold change, FC) (см. приложение Б и раздел 2). На рисунке 38 показаны значимо обогащенные термины ГО для набора данных ChIP-seq для ТФ CCA1 (GTRD ID PEAKS042882, GEO ID GSM1808452, 14-дневные проростки). Самые значимо обогащенные термины для CCA1, которые выявляют все три модели, это «*circadian rhythm*» и «*rhythmic process*», что хорошо согласуется с функциями данного ТФ (CCA1, Circadian Clock Associated 1), так как его ключевая функция это – регуляция циркадного ритма, при этом всего выявляется 20 общих терминов для трёх моделей (приложение Б). Поскольку независимо от модели мотива выявляется обогащение такими терминами ГО, то можно предположить, что такие термины отражают общие функции генов, связанные с регуляцией транскрипции генов этим ТФ.

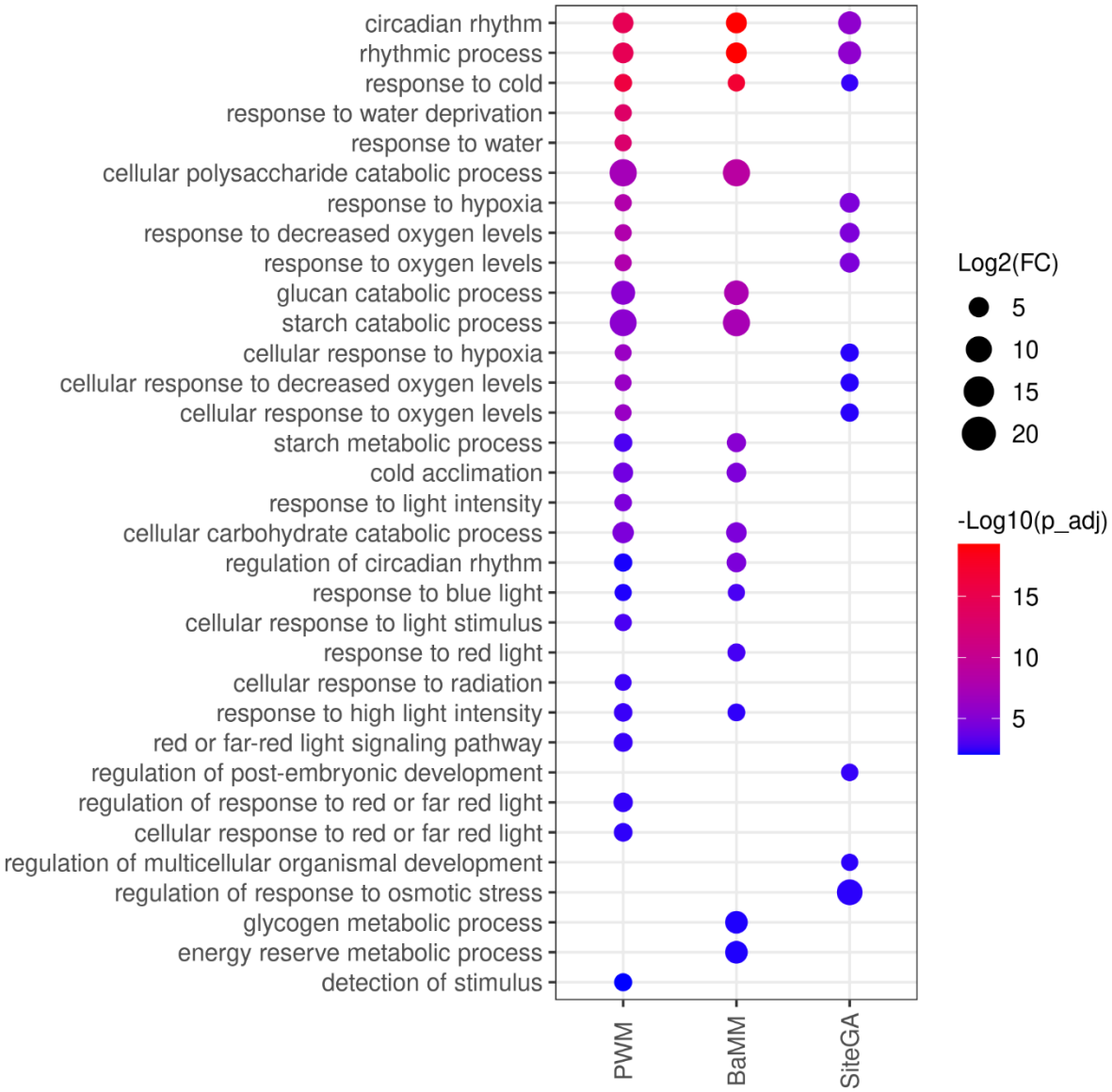

Рисунок 38. Значимо обогащенные термины GO для набора данных ChIP-seq для ТФ CCA1 (GTRD ID PEAKS042882, GEO ID GSM1808452), полученные по результатам картирования мотивов моделей PWM, BaMM и SiteGA в промоторы генов. На оси X обозначены модели мотивов. На оси Y перечислены термины GO. Размер кружка означает отношение долей генов, имеющих термин GO для анализируемого списка и всего генома (кратность изменения, англ. fold change, FC). Цвет характеризует значимость обогащения термина GO (скорректированное значение p-value, p_adj). При построении диаграммы применялись следующие пороги: FC ≥ 3 и p_adj < 0.01.

Из приведённого примера (Рисунок 38, приложение Б) для общих терминов GO было обнаружено, что для подавляющего большинства терминов GO модель SiteGA обладает более высоким значением кратности изменения по сравнению с другими моделями. Например, для термина GO «*circadian*

rhythm» значения кратности изменения для моделей PWM/BaMM/SiteGA составляют 5.3/5.43/7.03. Далее для того, чтобы исследовать это наблюдение, для набора данных ChIP-seq для ТФ CCA1 были построены диаграммы (Рисунок 39), на которых показаны попарные сравнения кратностей изменения для значимых терминов ГО, общих в парах моделей BaMM/PWM, SiteGA/PWM и SiteGA/BaMM.

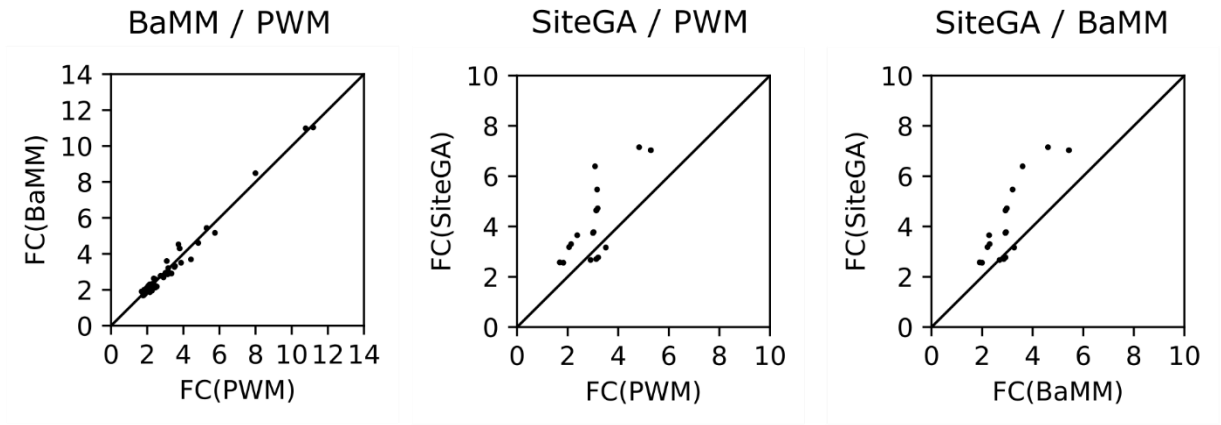

Рисунок 39. Попарные сравнения моделей PWM, BaMM и SiteGA по кратностям изменения. В каждой паре расчёты проведены для выборки общих значимых терминов ГО для трёх моделей мотива (p_adj < 0.05 для каждой из моделей). Анализ проведён для ChIP-seq данных по ТФ CCA1 (GTRD ID PEAKS042882, GEO ID GSM1808452). Оси X и Y на всех диаграммах отображают значение кратности изменения (FC) для соответствующей модели (приложение Б). Диагональ обозначает полное соответствие кратностей изменения в паре моделей.

В паре моделей BaMM/PWM почти все точки очень близки к диагонали (Рисунок 39), и среднее значение отношения кратности изменения модели BaMM к PWM составило 0.98. Для того, чтобы оценить отклонение отношения кратностей изменения в парах моделей от ожидаемого значения 1, применили U-тест Манна-Уитни, который показал, что для пары BaMM/PWM соотношение кратностей изменения не отличается значимо от 1 ($p > 0.05$). В парах моделей SiteGA/PWM и SiteGA/BaMM модель SiteGA имеет более высокие значения кратности изменения, точки лежат выше диагонали (Рисунок 39), а средние значения отношения кратности изменения модели SiteGA к PWM и к BaMM составили 1.36 и 1.35, соответственно. U-тест Манна-Уитни показал, что для обеих пар SiteGA/PWM и SiteGA/BaMM

наблюдается значимое отличия кратности изменения между общими терминами ГО в сторону модели SiteGA ($p < 0.05$).

Для того, чтобы дать оценку наблюдаемому эффекту по всей коллекции данных ChIP-seq, для каждой пары моделей в каждом ChIP-seq эксперименте был применен U-тест Манна-Уитни, чтобы сравнить отношения кратности изменения. После чего для каждой пары моделей был применён метод Фишера, который позволяет получить единое значение p-value (мета p-value) [239], на основании всех p-value, посчитанных для каждого индивидуального эксперимента, что позволяет сделать вывод для всей коллекции данных. Результаты приведены в таблице 10.

Таблица 10. Результаты сравнения кратности изменения по всей коллекции данных ChIP-seq для пар моделей (PMW/BaMM, PWM/SiteGA, BaMM/SiteGA) с использованием U-теста Манна-Уитни

Пара моделей	Кол-во экспериментов ChIP-seq	Кол-во экспериментов ChIP-seq со значимыми отличиями*	Мета p-value**
PMW/BaMM	62	5	> 0.05
PWM/SiteGA	55	27	2.07E-37
BaMM/SiteGA	55	28	4.96E-52

Примечание. * - эксперименты для которых U-тест Манна-Уитни показал значение $p < 0.05$; ** - значения мета p-value, посчитанное с помощью метода Фишера [239], характеризует результат по всей коллекции данных

Из полученных данных (таблица 10) 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 высокая эффективность модели SiteGA по сравнению с другими моделями мотивов может отражать как (1) более точные предсказания ССТФ в регуляторных районах генов, которые также предсказываются моделями PWM/BaMM; так и (2) предсказания SiteGA на генах, в которых отсутствуют предсказания ССТФ других моделей.

Другое важное наблюдение, следующее из данных представленных на рисунке 38 состоит в том, что некоторые термины ГО имеют значимое

обогащение только для одной модели. Например, для ТФ CCA1 количество таких терминов было 20/22/5, для предсказаний моделей PWM/BaMM/SiteGA, соответственно (Приложение Б). Из полученных результатов можно сделать вывод о том, что мотивы разных моделей могут соответствовать специфическим функциям генов. Примечательно, что подобный эффект уже наблюдался ранее, где ССТФ, предсказанные моделью SiteGA для данных ChIP-seq ТФ SF-1 человека, имели видимую тенденцию к терминам ГО, связанным с негативной регуляцией транскрипции и апоптозом, в отличие от ССТФ, предсказанных моделью PWM, связанных с позитивной регуляцией [44].

Далее, для оценки вкладов моделей в расширение списка обогащенных терминов ГО, для каждой из моделей (PWM, BaMM и SiteGA) было получено распределение количества обогащенных терминов ГО (Рисунок 40А), а также было получено распределение количества терминов ГО, которые были обогащены по данным хотя бы одной из моделей (Рисунок 40А). Среди всех терминов для каждой модели было получено распределение количества обогащённых терминов ГО, которые не выявляются другими моделями (Рисунок 40Б).

По данным, приведенным на рисунке 40А, можно заключить, что модель PWM в среднем находит 104 обогащенных ГО термина, а добавка двух альтернативных моделей значимо ($p < 0.001$) увеличивают среднее количество обогащённых терминов ГО до 173. При этом для всех трёх моделей находятся термины ГО, которые обогащены только для одной из моделей (Рисунок 40Б), следовательно, и модель PWM вносит важный вклад в поиск специфических ССТФ, которые, исходя из особенности модели, имеют высокую консервативность. Поиск специфических терминов ГО может позволить расширить представление о биологических процессах, которые регулируют ТФ, а также выявить специфические группы генов, которые имеют в промоторах сайты только одной из моделей.

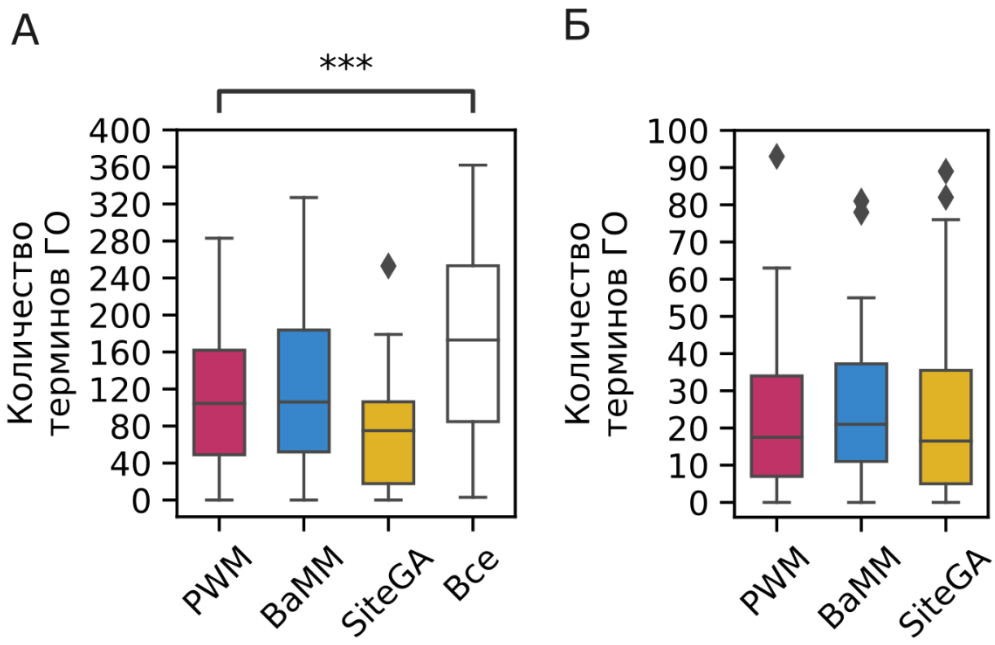

Рисунок 40. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации для анализа обогащения терминов ГО на всей коллекции данных ChIP-seq экспериментов для *A. thaliana*. **(А)** Показаны распределения количества значимых терминов ГО, полученных для каждой из моделей (PWM, BaMM и SiteGA), а также распределение количества значимых терминов ГО, обогащённых хотя бы для одной из трёх моделей (Все). **(Б)** Показаны распределения количества терминов ГО, которые имеют обогащение только для одной модели. На диаграммах представлены распределения квартилей Q_1, Q_2 и Q_3 по количеству ГО терминов. Планки погрешностей ниже (Q_1) и выше (Q_3) относятся к минимальным/максимальным значениям, если они расположены в пределах полутора межквартильных диапазонов (IQR = Q_3–Q_1) от Q_1 и Q_3, соответственно. В противном случае планки погрешностей установлены в положениях {Q_1−1.5*IQR}/{Q_3+1.5*IQR} для квартилей Q_1/Q_3, соответственно. Все значения, которые не попали в пределы планок погрешности отмечены ромбами как выбросы. *** – $p < 0.001$.

3.3 Массовый анализ данных ChIP-seq для *M. musculus*

3.3.1 Подготовка данных и выбор оптимальных моделей для анализа

Для массового анализа данных ChIP-seq была сформирована коллекция из 1556 предобработанных ChIP-seq экспериментов для тканей и органов *M. musculus* в виде разметки пиков в формате bed, взятая из базы данных GTRD (https://gtrd.biouml.org/#!) [185] (приложение А, таблица 2). В данном анализе рассмотрены только эксперименты для тканей и органов, так как общее количество данных ChIP-seq по ТФ млекопитающих в GTRD существенно

больше такового по ТФ растений. В анализ были взяты пики, длина которых не превышала 3000 п.о., а количество пиков для *de novo* поиска мотивов и оценки точности распознавания составило 1000. В данном анализе также использовались три модели мотивов PWM, BaMM и SiteGA.

3.3.2 Оценка качества исходных данных

Данные были отфильтрованы тем же способом, как и для *A. thaliana* (см. раздел 3.2.2). В анализ взяли только те ChIP-seq эксперименты, для которых выполнялись следующие условия: (1) мотив целевого ТФ обогащён согласно AME; (2) все три модели выявили мотив значимо похожий на мотив целевого ТФ согласно TomTom. Данным условиям удовлетворяли 1003 ChIP-seq эксперимента, которые и использовались в дальнейшем анализе. Поскольку часть анализа проводилось с учётом класса ТФ по ДСД [83], данные были разбиты по этим классам (Таблица 11).

Таблица 11. Информация о классах ТФ мыши, используемых в анализе. Значения в колонках GTRD, AME и TomTom означают количества исходных данных извлечённых из базы данных GTRD до фильтрации, после их фильтрации по обогащению мотивов (до проведения *de novo* поиска мотивов, инструмент AME) и после их фильтрации по сходству мотивов с известными мотивами целевых ТФ (после проведения *de novo* поиска мотивов, инструмент TomTom).

Класс		Число экспериментов			Число ТФ		
		GTRD	AME	TomTom	GTRD	AME	TomTom
Basic leucine zipper factors (bZIP)	{1.1}	274	274	214	23	23	20
Basic helix-loop-helix factors (bHLH)	{1.2}	134	130	95	28	27	25
Basic helix-span-helix factors (bHSH)	{1.3}	4	4	4	1	1	1
Nuclear receptors with C4 zinc fingers	{2.1}	187	180	103	24	23	14
C4 zinc finger-type factors	{2.2}	22	22	19	5	5	5
C2H2 zinc finger factors	{2.3}	227	210	178	32	24	19
DM-type intertwined zinc finger factors	{2.5}	2	1	1	2	1	1
CXXC zinc finger factors	{2.6}	19	0	0	4	0	0
Homeo domain factors	{3.1}	45	43	23	24	23	15
Paired box factors	{3.2}	10	10	6	3	3	2
Fork head / winged helix factors	{3.3}	66	66	34	15	15	9
Tryptophan cluster factors	{3.5}	189	189	134	19	19	16
TEA domain factors	{3.6}	6	6	1	2	2	1
High-mobility group (HMG) domain factors	{4.1}	20	19	10	10	9	7
MADS box factors	{5.1}	25	25	4	4	4	3
SAND domain factors	{5.3}	1	0	0	1	0	0
Rel homology region (RHR) factors	{6.1}	114	114	85	9	9	6
STAT domain factors	{6.2}	105	105	56	7	7	7
p53 domain factors	{6.3}	31	31	8	1	1	1
Runt domain factors	{6.4}	35	35	21	3	3	3
T-Box factors	{6.5}	13	13	4	5	5	1
Grainyhead domain factors	{6.7}	3	3	3	1	1	1
SMAD/NF-1 DNA-binding domain factors	{7.1}	9	9	0	2	2	0
TATA-binding proteins	{8.1}	4	4	0	1	1	0
Не установлен	{0.*}	11	0	0	4	0	0
Вся коллекция		1556	1493	1003	230	208	157

Примечание. Фигурные скобки представляют числовые обозначения классов согласно иерархической классификации ТФ TFClass; звёзды в обозначениях классов отмечают классы, отнесённые к суперклассу {0}, содержащему ТФ не классифицированные по известным девяти суперклассам [81–83]

3.3.3 Выбор оптимальных параметров и оценка точности распознавания ССТФ для моделей

Аналогично тому, как это описано выше в разделе 3.2.3, для трёх моделей мотива были выбраны оптимальные параметры для каждого эксперимента в отдельности, далее построены ROC-кривые и посчитаны pAUC.

На рисунке 41 приведены ROC-кривые для классов ТФ: *Basic leucine zipper factors (bZIP)* {1.1}*, Basic helix-loop-helix factors (bHLH)* {1.2}*, C2H2 zinc*

finger factors {2.3}*, Tryptophan cluster factors* {3.5}. ROC-кривые для *M. musculus* (Рисунок 41) имеют схожие закономерности, что ROC-кривые для *A. thaliana* полученные для тех же классов ТФ (раздел 3.2.3, Рисунок 27). Модель BaMM в диапазоне жёстких порогов имеет сопоставимую точность с PWM, но со смягчением порога для модели BaMM наблюдается более высокий рост точности в сравнении с PWM, и на мягких порогах модель BaMM превосходит PWM по точности. Стоит отметить, для классов ТФ *Basic leucine zipper factors (bZIP)* {1.1} и *Basic helix-loop-helix factors (bHLH)* {1.2}, модель BaMM еще на достаточно жёстких порогах обходит по точности PWM, а у класса *C2H2 zinc finger factors* {2.3} ROC-кривые моделей PWM и BaMM имеют идентичную точность, преимущество BaMM немного растёт лишь для мягких порогов. Модель SiteGA только на мягких порогах приближается по точности к PWM/BaMM, и только на классе *Basic helix-loop-helix factors (bHLH)* {1.2} опережает по точности модель PWM.

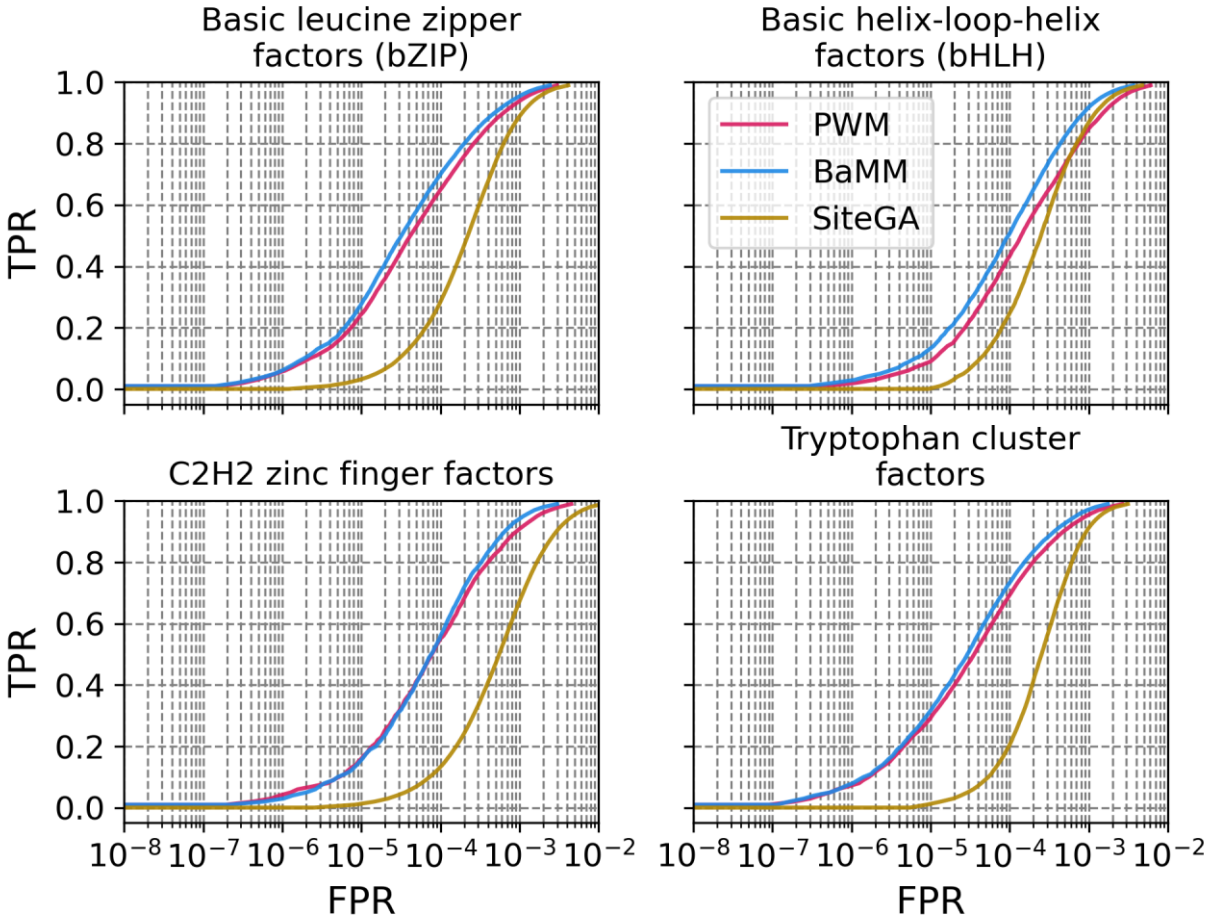

Рисунок 41. Характеристика точности распознавания мотивов моделей PWM, BaMM и SiteGA на данных для *M. musculus* по четырём классам ТФ: *Basic leucine zipper factors (bZIP)* {1.1}, *Basic helix-loop-helix factors (bHLH)* {1.2}, *C2H2 zinc finger factors* {2.3} и *Tryptophan cluster factors* {3.5}. ROC-кривые для моделей, которые были получены с применением 2-fold CV процедуры (см. раздел 2.4). На графиках показаны средние значения FPR (ось X) в зависимости от значений TPR (ось Y).

Далее были построены распределения точности моделей по показателю pAUC для каждого класса ТФ в отдельности (Рисунок 42), при этом брали только те классы, для которых количество экспериментов превышало десять (см. Таблица 11).

Полученные распределения (Рисунок 42) имеют такие же характерные черты и соотношения точностей медиан моделей, что и полученные выше для *A. thaliana* (см. раздел 3.2.3, Рисунок 28). Как видно из диаграммы (Рисунок 42) точность моделей зависит от класса ТФ. В частности, лучшая точность моделей (PWM и BaMM) достигается для классов *Basic leucine zipper factors*

(bZIP) {1.1} и *C2H2 zinc finger factors* {2.3}, а модель SiteGA имеет лучшую точность для класса *Basic helix-loop-helix factors bHLH* {1.2}.

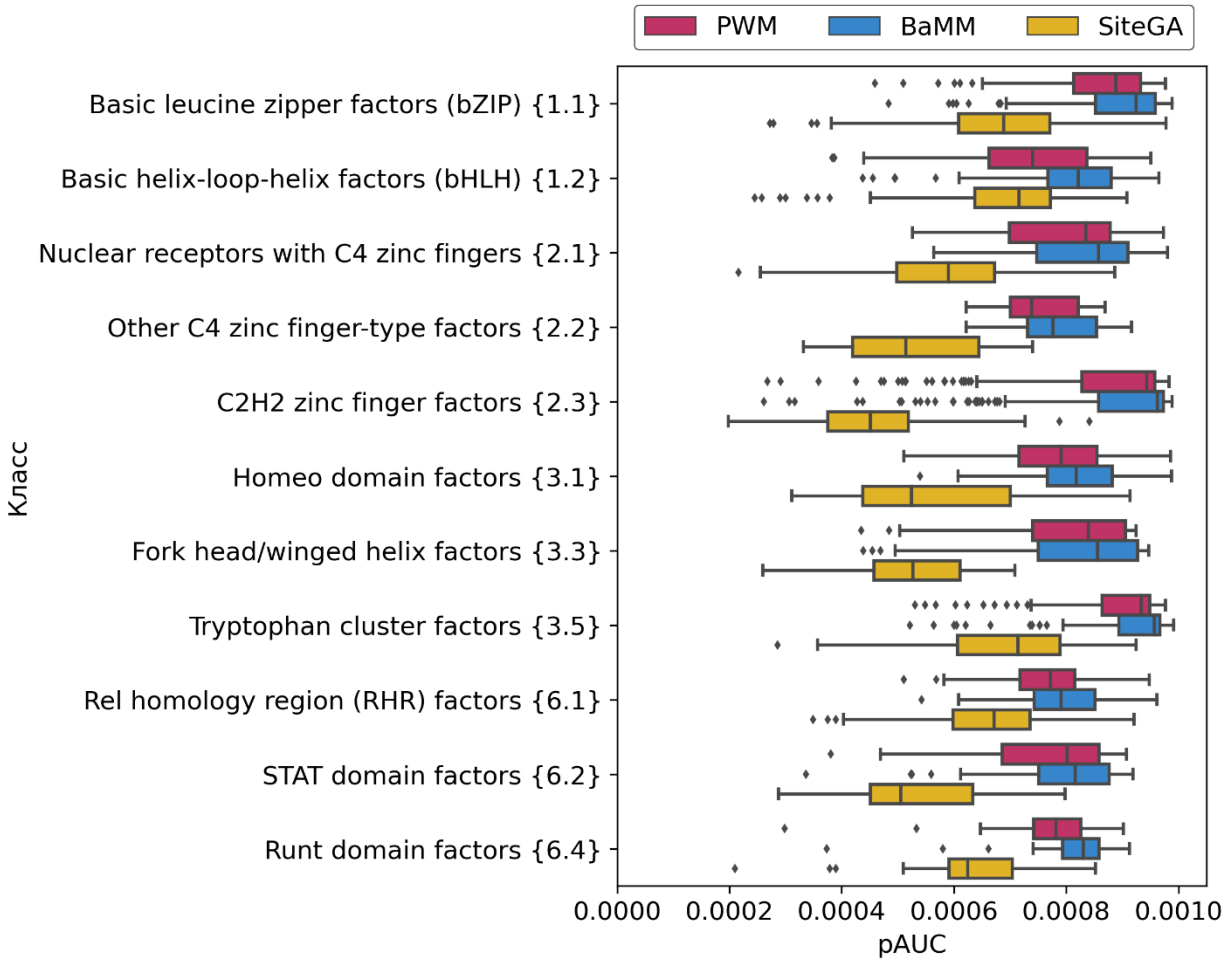

Рисунок 42. Диаграмма размаха с распределениями оценки точности pAUC для трёх моделей (PWM, BaMM и SiteGA), рассчитанные отдельно для ТФ шести классов. На диаграмме представлены распределения квартилей Q_1, Q_2 и Q_3 для pAUC. Планки погрешностей ниже Q_1 и выше Q_3 относятся к минимальным/максимальным значениям, если они расположены в пределах 1.5 межквартильных диапазонов (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.

Значения медиан pAUC для модели BaMM самые высокие по сравнению с остальными моделями, а модель SiteGA имеет наименьшие значения медиан по величине pAUC. Однако можно отметить, что точность модели SiteGA относительно PWM варьируется в зависимости от класса ТФ, и наилучшая точность для модели SiteGA относительно PWM наблюдается на классах *Basic helix-loop-helix factors (bHLH)*{1.2} и *Rel homology region (RHR) factors*{6.1}.

3.3.4 Совместное применение моделей PWM, BaMM и SiteGA для поиска ССТФ

Далее было проведено распознавание сайтов в пиках разными моделями мотивов (порог распознавания ERR ≤ 1E-4) и получены доли пиков, содержащих сайты разных моделей (PWM, BaMM и SiteGA). Результаты распознавания для разных классов ТФ по структуре ДСД представлены в таблице 12, данные приведены только для классов, для которых количество экспериментов было 10 и более.

Таблица 12. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации на жёстком пороге (ERR ≤ 1E-4) для наиболее представительных классов ТФ. В ячейках записаны значения медиан долей пиков с предсказанными ССТФ (в %).

Индекс	Класс	PWM	BaMM	SiteGA	Все*	Все* - PWM
{1.1}	*Basic leucine zipper factors (bZIP)*	76.95	84.5	12.85	90.2	13.25
{1.2}	*Basic helix-loop-helix factors (bHLH)*	55.3	63.6	19.6	78.8	23.5
{2.1}	*Nuclear receptors with C4 zinc fingers*	69.4	75.5	11.1	82.1	12.7
{2.2}	*Other C4 zinc finger-type factors*	51.3	55	7.9	65.1	13.8
{2.3}	*C2H2 zinc finger factors*	92.3	95.3	13.3	96.5	4.2
{3.1}	*Homeo domain factors*	54.5	58.1	11.3	71.6	17.1
{3.3}	*Fork head/winged helix factors*	62.75	67.25	8.1	76.3	13.55
{3.5}	*Tryptophan cluster factors*	85.65	92.8	17.6	95.95	10.3
{6.1}	*Rel homology region (RHR) factors*	58.05	60.9	14.65	73.4	15.35
{6.2}	*STAT domain factors*	60.75	65.2	8.75	73.15	12.4
{6.4}	*Runt domain factors*	52.1	60.9	16.6	73.9	21.8

Примечание. *Все – доли пиков, содержащих мотивы, по крайней мере одной из трёх моделей; **Все-PWM – доля пиков с мотивами BaMM или SiteGA, но без мотивов PWM.

Из представленных в таблице 12 данных можно отметить, что вклад альтернативных моделей зависит от класса ТФ так же, как это было показано ранее для *A. thaliana* (см. раздел 3.2.6). Добавка предсказаний альтернативных моделей BaMM/SiteGA к доле пиков модели PWM, варьируется от 4.2% до 23.5%. Наибольший вклад в распознавание сайтов альтернативные модели продемонстрировали для класса *Basic helix-loop-helix factors (bHLH)* {1.2}, где добавка доли пиков составила 23.5%. На *A. thaliana* наибольший вклад был для классов *MADS box factors* и *GCM domain factors*, однако в коллекции *M. musculus* они не представлены. Тем не менее и у *A. thaliana* альтернативные

модели для класса *Basic helix-loop-helix factors (bHLH)* {1.2} показали достаточно высокий вклад в распознавание сайтов – 13.5%.

Наименьшую добавку доли пиков, как и для *A. thaliana* (см. раздел 3.2.6, таблица 7), альтернативные модели показывают для класса *C2H2 zinc finger factors* {2.3} (таблица 12), где вклад составил 4.2%. На данных для *A. thaliana* альтернативные модели не так эффективно себя показали для класса *Basic leucine zipper factors (bZIP)* {1.1}, где вклад составил 9.6%, однако для *M. musculus* это – 13.25%. Возможно такое различие в первую очередь связано с отличием размеров выборок ChIP-seq 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, следовательно более адекватную оценку для данного класса дают результаты полученные на *M. musculus*.

Полученные для *M. musculus* результаты хорошо согласуются с результатами, описанными выше для *A. thaliana* (см. раздел 3.2.6). Альтернативные модели вносят существенный вклад в распознавание сайтов в пиках для разных классов, при этом данный вклад зависит от класса ТФ. Как и для *A. thaliana* альтернативные модели практически не 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 длинными ССТФ для данного класса ТФ и минимальным количеством димеров ТФ. Стоит отметить, что для данного класса модель SiteGA имеет наихудшую точность (Рисунок 40), что так же хорошо согласуется с гипотезой о том, что у данного класса ТФ зависимости внутри мотива оказываются очень слабыми. С другой стороны, альтернативные модели сделали наибольший вклад в распознавание сайтов в пиках для класса *Basic helix-loop-helix factors (bHLH)* {1.2}, это согласуется и с результатами для *A. thaliana*. Для данного класса ТФ, модель SiteGA имеет наилучшую точность по сравнению с другими классами, что может свидетельствовать о наличии существенного вклада зависимостей в структуру мотива. 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 мотива, а также наличии различных модификаций ТФ, которые влияют на его конформацию и, как следствие, на структуру сайта, с которым связывается ТФ [170].

3.3.5 Сравнительный анализ списков терминов генной онтологии, полученных путём применения моделей PWM, BaMM и SiteGA

Для проведения анализа для поиска обогащённых терминов ГО были распознаны сайты в пиках с помощью всех моделей PWM, BaMM и SiteGA, (на среднем пороге ERR ≤ 2.5E-4). В данном анализе, как и ранее для коллекции *A. thaliana* (см. раздел 3.2.6), в каждом эксперименте использовался полный набор пиков, а не 1000 лучших. Далее распознанные сайты разных моделей картировали на промоторы генов *M. musculus* (см. раздел 2.8), после чего для каждой модели был получен список генов, в промоторах которых эти модели предсказали сайты.

Результаты аннотации представлены в виде диаграмм Венна (Рисунок 43), где изображены доли генов (медианные значения по классам), содержащих в промоторах сайты для разных комбинаций моделей. В предыдущих разделах (см. раздел 3.2.7) уже было показано, что альтернативные модели находят сайты в промоторах, в которых модель PWM не находит сайты ТФ. Результаты на коллекции для *M. musculus* это подтверждают (Рисунок 43). Отдельно стоит отметить, что результаты по долям генов для *M. musculus* (Рисунок 43) и *A. thaliana* (см. раздел 3.2.7, Рисунок 37) хорошо согласуются для классов *Basic helix-loop-helix factors (bHLH)* {1.2}, *Basic leucine zipper factors (bZIP)* {1.1}, *C2H2 zinc finger factors* {2.3}. Для *Basic helix-loop-helix factors (bHLH)* {1.2}, *Basic leucine zipper factors (bZIP)* {1.1} в обеих коллекциях альтернативные модели существенно расширяют долю генов, в промоторах которых есть сайты.

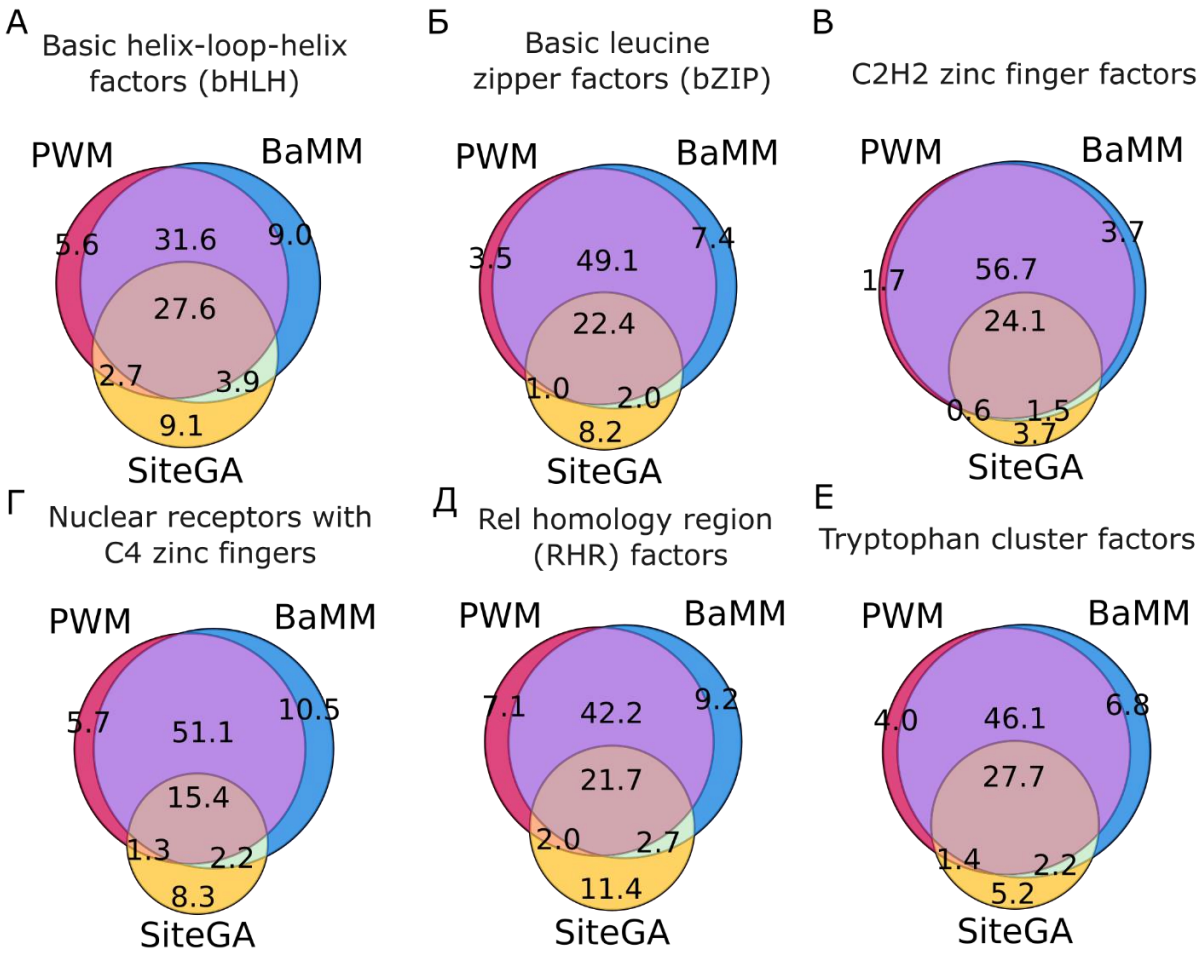

Рисунок 43. Диаграммы Венна для классификации генов, промоторы которых содержат сайты моделей мотива PWM, BaMM и SiteGA. На диаграммах **(А)**, **(Б)**, **(В)**, **(Г)**, **(Д)** и **(Е)** показаны результаты в виде медиан долей промоторов содержащих сайты одной, двух (в разных комбинациях моделей) или трёх моделей мотива для шести классов ТФ: *Basic helix-loop-helix factors (bHLH)* {1.2}, *Basic leucine zipper factors (bZIP)* {1.1}, *C2H2 zinc finger factors* {2.3}, *Nuclear receptors with C4 zinc fingers {2.1}, Rel homology region (RHR) factors* {6.1} и *Tryptophan cluster factors* {3.5}. Анализ проводился на среднем пороге (ERR ≤ 2.5E-4) распознавания. Результаты приведены для коллекции данных ChIP-seq *M. musculus*.

С другой стороны, для класса *C2H2 zinc finger factors* {2.3}, альтернативные модели добавляют меньшие доли генов, чем для других классов. Из данных представленных на рисунке 42, а также данных, которые были представлены выше можно предположить, что одни и те же классы ТФ у организмов разных таксонов имеют схожие механизмы связывания с ДНК, которые отражают модели мотива.

Далее, как и ранее для *A. thaliana*, полученные списки генов *M. musculus* использовали для поиска обогащённых терминов ГО. По всей коллекции

данных для каждого ChIP-seq эксперимента были посчитаны средние значения отношения кратности изменения в парах моделей BaMM и PWM, SiteGA и PWM, SiteGA и BaMM, аналогично тому, как это было описано ранее (см. раздел 3.2.7). После чего было построено распределение этих величин для каждой пары моделей. Результаты анализа приведены на рисунке 44.

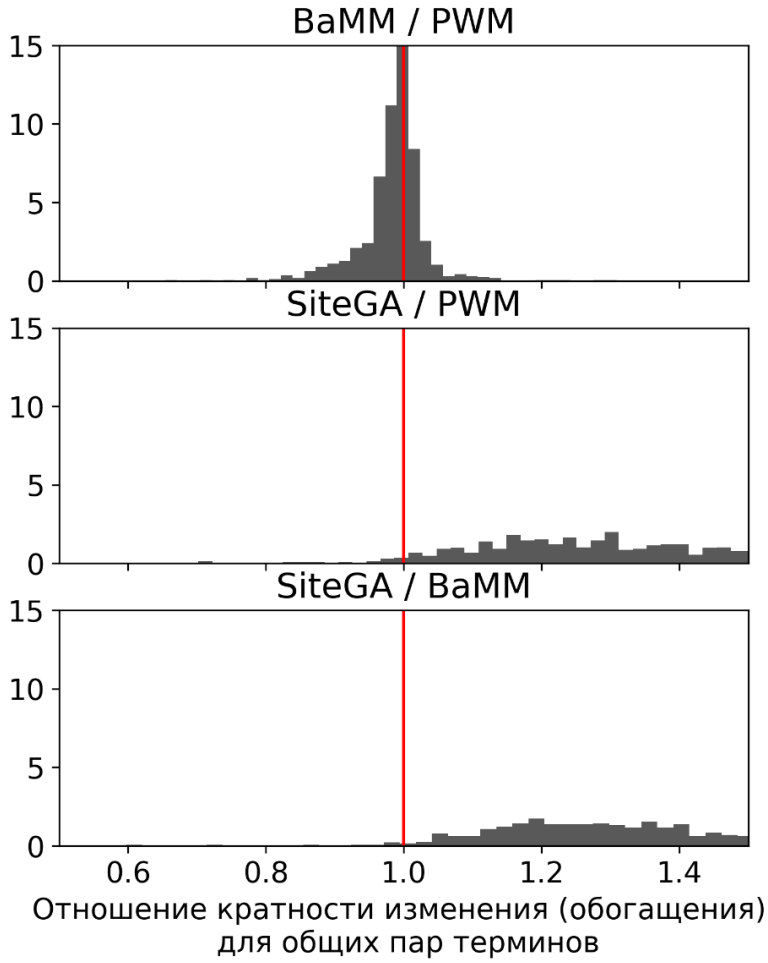

Рисунок 44. Гистограммы распределений отношений средней кратности изменения для общих значимых терминов ГО в парах моделей мотива (BaMM/PWM, SiteGA/PWM и SiteGA/BaMM) по всей коллекции наборов данных ChIP-seq *M. musculus*.

Как видно из диаграммы (Рисунок 44), для пары BaMM/PWM полученное распределение находится в диапазоне от 0.8 до 1.2 с максимумом в области единицы, что говорит о близости величин кратности изменения. Для пар SiteGA/PWM и SiteGA/BaMM распределения существенно сдвинуты вправо относительно единицы, что свидетельствует о том, что у модели SiteGA кратность изменения для терминов ГО существенно больше по

сравнению с кратностями изменения моделей PWM и BaMM (Рисунок 44). Средние значения отношений средней кратности изменения для полученных распределений равны 0.99, 1.53 и 1.57 для пар BaMM/PWM, SiteGA/PWM и SiteGA/BaMM, соответственно.

Для того, чтобы оценить значимость разницы между величинами кратностей изменения ГО внутри каждой пары применили U-тест Манна-Уитни, который применили для каждого ChIP-seq эксперимента по всей коллекции данных (таблица 13). Как и ранее (см. раздел 3.2.7) для каждой пары моделей был применён метод Фишера, для получения единого значение p-value (мета p-value) [239], для того, чтобы дать оценку значимости наблюдаемого эффекта по всей коллекции данных. Результаты приведены в таблице 11.

Таблица 13. Результаты сравнения кратности изменения для пар моделей (PMW/BaMM, PWM/SiteGA, BaMM/SiteGA) с использованием U-теста Манна-Уитни для коллекции данных ChIP-seq для вида организма *M. musculus*.

Пара моделей	Кол-во экспериментов ChIP-seq	Кол-во экспериментов ChIP-seq со значимыми отличиями*	Мета p-value**
BaMM /PMW	942	43	2.63E-24
SiteGA/PWM	810	753	3.04E-260
SiteGA/BaMM	809	776	3.49E-266

Примечание. * - эксперименты для которых U-тест Манна-Уитни показал значение $p < 0.05$; ** - мета p-value, посчитанное с помощью метода Фишера [239], характеризует результат по всей коллекции данных

Из полученных данных (таблица 11) видно, что в парах SiteGA/PWM и SiteGA/BaMM модель SiteGA в подавляющем большинстве экспериментов имеет значимо более высокие значения кратности согласно U-тесту Манна-Уитни и получаемые значения мета p-value существенно меньше 0.05. Полученный для SiteGA результат на *M. musculus* хорошо согласуется с результатом, полученным для *A. thaliana* (см. раздел 3.2.7).

Модель BaMM, в паре BaMM/PMW, имеет значимо большую кратность отношения терминов ГО в сравнении с моделью PWM только в 43

экспериментах из 942, что составляет менее 5% всех экспериментов, однако метод Фишера для объединения p-value даёт значение мета p-value равное 2.63E-24 из-за чего можно было бы сделать вывод, что во всей коллекции данных модель BaMM имеет значимо большую кратность отношения по сравнению с PWM. Однако из данных, приведенных на рисунке 43 можно увидеть, что различие кратностей изменения в паре моделей BaMM/PWM очень мало. Так же стоит отметить, что для SiteGA значение мета p-value на много порядков меньше по сравнению с BaMM. Возможно, значимый результат для модели BaMM связан с тем, что метод Фишера очень чувствителен к отдельным низким значениям p-value, которые могли повлиять на конечное значение мета p-value [239].

Далее для каждой модели оценили общее количество терминов ГО, которые модели (PWM, BaMM и SiteGA) выявляли и построили распределения по результатам, полученным по всей коллекции данных (Рисунок 45А). Помимо этого, были получены распределения количества терминов ГО, которые имели значимое обогащение только у одной из моделей (Рисунок 45Б).

Данные, приведенные для *M. musculus* на рисунке 45, в целом соответствуют аналогичным данным полученным для *A. thaliana* (Рисунок 40). Из диаграммы видно, что модель PWM в среднем находит 975 обогащенных ГО терминов, а альтернативные модели значимо ($p < 0.0001$) расширяют количество обогащенных терминов ГО до 1373. Так же как это было показано на *A. thaliana* (см. раздел 3.2.7) все три модели выявляют термины ГО, которые обогащены только для одной из моделей (Рисунок 45Б), следовательно, все модели вносят важный вклад в поиск специфических сайтов, связанных с разными биологическими функциями. Таким образом, применение нескольких моделей мотива позволяет существенно расширить представление о биологических процессах, связанных с конкретным ТФ, а также выявить специфические группы генов, которые им регулируются.

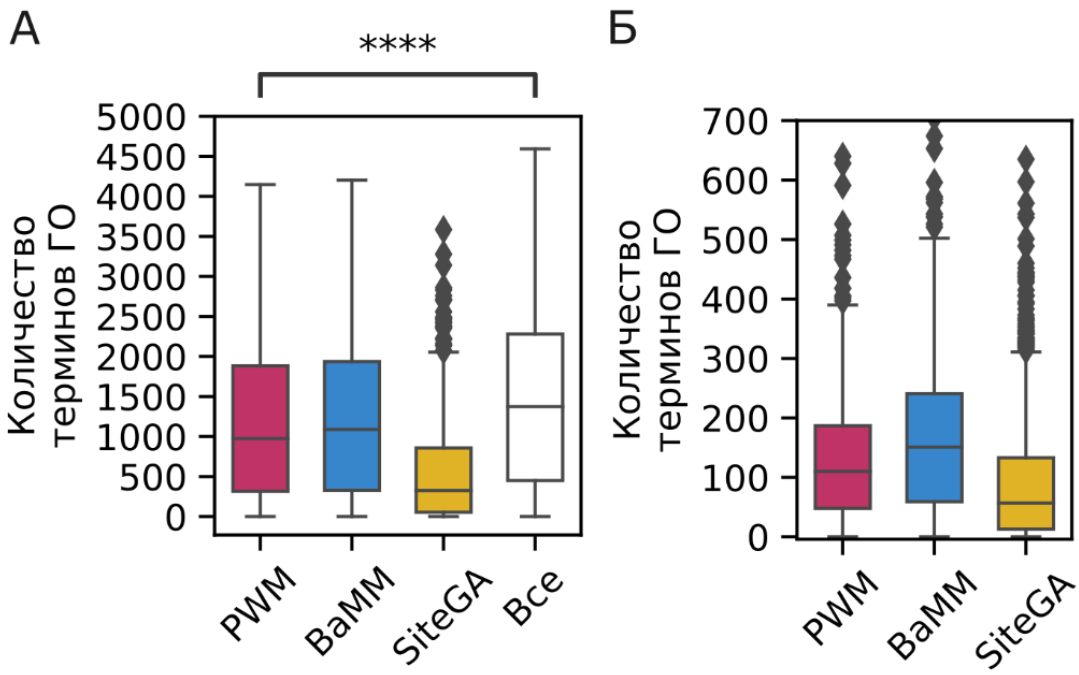

Рисунок 45. Сравнение результатов применения моделей PWM, BaMM и SiteGA и их комбинации для анализа обогащения ГО терминов на всей коллекции данных ChIP-seq экспериментов для *M. musculus*. **(А)** Показаны распределения количества значимых терминов ГО, полученных для каждой из моделей (PWM, BaMM и SiteGA), а также распределение количества значимых терминов ГО, обогащённых хотя бы одной из трёх моделей (Все). **(Б)** Показаны распределения количества значимых терминов ГО, которые имеют обогащение только для одной модели. На диаграммах представлены распределения квартилей Q_1, Q_2 и Q_3 по количеству ГО терминов. Планки погрешностей ниже (Q_1) и выше (Q_3) относятся к минимальным/максимальным значениям, если они расположены в пределах полутора межквартильных диапазонов (IQR = Q_3−Q_1) от Q_1 и Q_3, соответственно. В противном случае планки погрешностей установлены в положениях {Q_1−1.5*IQR}/{Q_3+1.5*IQR} для квартилей Q_1/Q_3, соответственно. Все значения, которые не попали в пределы планок погрешности, отмечены ромбами как выбросы. **** – $p < 0.0001$.

3.3.6 Модель SiteGA распознаёт разные структурные варианты мотива сайтов связывания для транскрипционного фактора JUNB

Полученные выше результаты демонстрируют, что модели способны находить разные структурные варианты сайтов, что выражается как в «уникальной» доли пиков для каждой модели, так и в «уникальной» доле генов. Более того, разная структура сайтов может быть ассоциирована с разными функциями, что продемонстрировали результаты по анализу значимых терминов ГО. При этом среди всех моделей больше всего

выделяется модель SiteGA, которая хоть и распознаёт меньше всего сайтов в пиках, тем не менее, имеет самые большие доли «уникальных» пиков/генов (Рисунки 35, 37, 43). Поэтому с помощью инструмента DepLogo были проанализированы ССТФ, которые распознаёт модель SiteGA (ERR ≤ 5E-4). Для класса *Basic leucine zipper factors (bZIP)*, в частности, для ТФ JUNB было обнаружено, что модель SiteGA за счёт учёта зависимостей может выявлять разные структурных варианты мотива (Рисунок 46А). Детальный разбор результатов применения инструмента DepLogo для ССТФ JUNB (GTRD ID PEAKS040976, GEO GSM2663851, клетки макрофагов, полученных из костного мозга, мышь) показывает, что сайты, предсказанные моделью SiteGA, разделяются на три структурных варианта. Первый и второй варианты имеют в своей структуре спейсер длиной в 1 п.о. между полусайтами и разные правые фланги, а третий вариант имеет спейсер длиной в 2 п.о. (Рисунок 46А). Важно отметить, что для ТФ JUNB модели PWM и BaMM нашли только один структурный вариант мотива со спейсером длины 1 п.о. (Рисунок 46Б). Отсюда следует, что в некоторых случаях модель SiteGA за счёт учёта зависимостей позиций в мотиве, которые не имеют ограничения по расстоянию между нуклеотидами, может обобщать разные структурные варианты сайтов в одном мотиве.

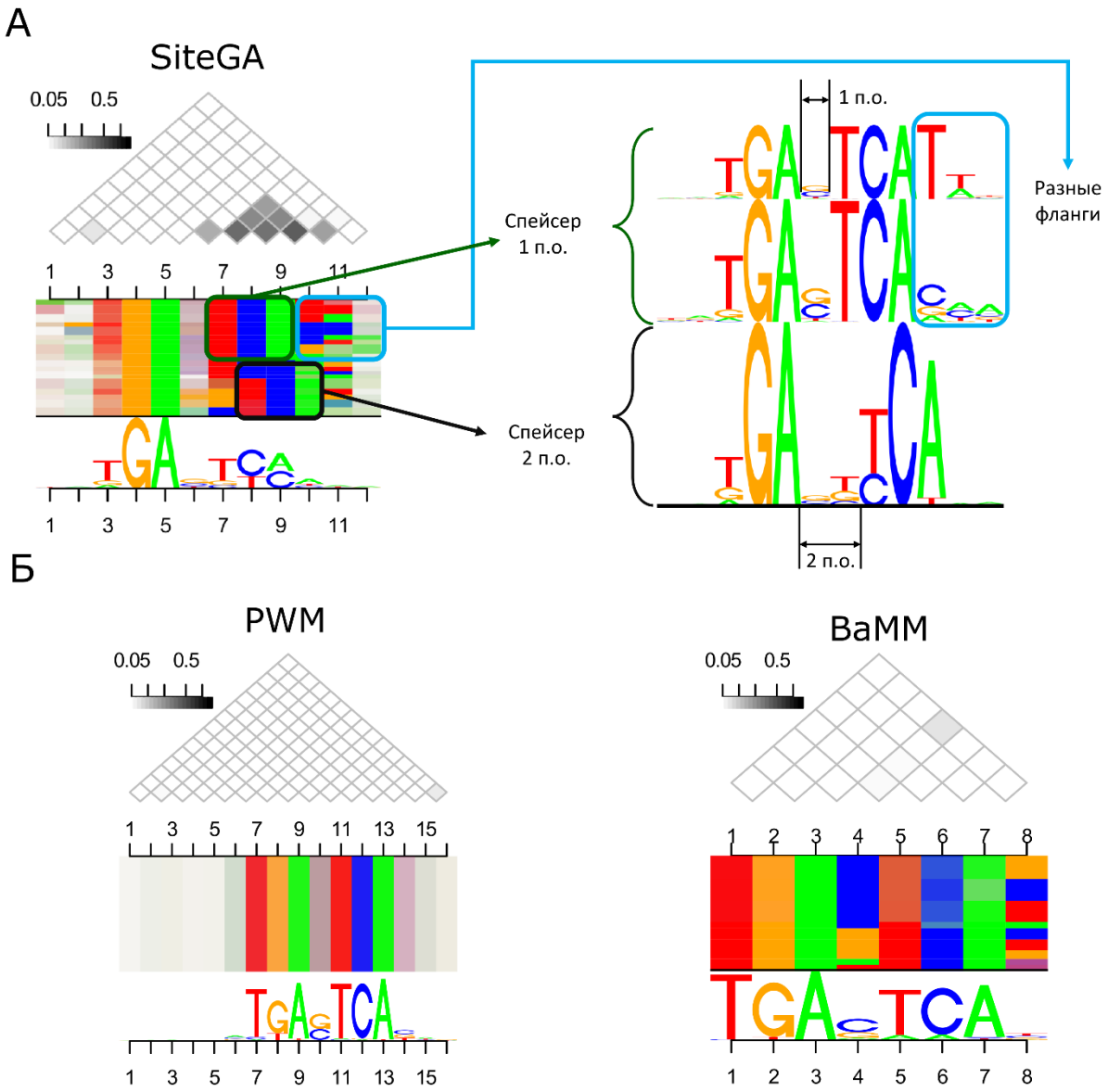

Рисунок 46. Применение инструмента DepLogo [222] для мотивов моделей PWM, BaMM и SiteGA. Для каждой модели в анализ включены выравнивания предсказанных ССТФ по пикам ChIP-seq для ТФ JUNB (GTRD ID PEAKS040976, GEO GSM2663851, клетки макрофагов, полученных из костного мозга, мышь). **(А)** Визуализация DepLogo для модели SiteGA по распознанным сайтам и структурные варианты мотива, которые находит модель SiteGA. **(Б)** Визуализация DepLogo для моделей PWM и BaMM по распознанным сайтом. Расчёты произведены на мягком пороге (ERR ≤ 5E-4).

Заключение

Данная работа посвящена применению методологически разных моделей *de novo* поиска мотивов в данных ChIP-seq. *De novo* поиск мотивов в данных ChIP-seq является важной задачей, так как позволяет лучше понять механизмы регуляции экспрессии генов. В настоящее время большинство исследователей используют инструменты *de novo* поиска мотивов, основанные на применении модели PWM [11]. Однако, экспериментально показано [22], что модель PWM имеет ограничения, поскольку она не учитывает зависимости между разными позициями сайтов [24, 25]. При этом уже разработан ряд альтернативных моделей мотивов, которые учитывают разные особенности связывания ТФ с ДНК [25, 31, 37–41]. Однако, авторы редко уделяют внимания тому, что их модели могут находить разные структурные типы ССТФ, отличные от таковых для традиционной модели PWM. Помимо этого, применение только одной модели, хоть и не PWM, не решает проблему наиболее полного распознавания ССТФ в данных ChIP-seq.

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, $p < 2*10^{-22}$.

Список литературы

1. Lambert, S.A. The Human Transcription Factors / S.A. Lambert, A. Jolma, L.F. Campitelli et al. // Cell. – 2018. – Vol. 172. – № 4. – P. 650-665.

2. Srivastava, D. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns / D. Srivastava, S. Mahony // Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. – 2020. – Vol. 1863. – № 6. – P. 194443.

3. Iwafuchi-Doi, M. The mechanistic basis for chromatin regulation by pioneer transcription factors. Vol. 11 / M. Iwafuchi-Doi. – NLM (Medline), 2019.

4. Latchman, D.S. Transcription factors: Bound to activate or repress. Vol. 26 / D.S. Latchman. – Elsevier Ltd, 2001.

5. Park, P.J. ChIP-seq: Advantages and challenges of a maturing technology. Vol. 10 / P.J. Park. – Nature Publishing Group, 2009.

6. Farnham, P.J. Insights from genomic profiling of transcription factors. Vol. 10 / P.J. Farnham. – Nat Rev Genet, 2009.

7. Furey, T.S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions / T.S. Furey // Nature Reviews Genetics. – 2012. – Vol. 13. – № 12. – P. 840-852.

8. Kulakovskiy, I. V. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. / I. V. Kulakovskiy, I.E. Vorontsov, I.S. Yevshin et al. // Nucleic acids research. – 2018. – Vol. 46. – № D1. – P. D252-D259.

9. O'Malley, R.C. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape / R.C. O'Malley, S.C. Huang, L. Song et al. // Cell. – 2016. – Vol. 165. – № 5. – P. 1280-1292.

10. Lihu, A. A review of ensemble methods for de novo motif discovery in ChIP-Seq data / A. Lihu, tefan Holban // Briefings in Bioinformatics. – 2015. – Vol. 16. – № 6. – P. 964-973.

11. Stormo, G.D. DNA binding sites: Representation and discovery. Vol. 16 / G.D. Stormo. – Oxford University Press, 2000.

12. Berg, O.G. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters / O.G. Berg, P.H. von Hippel // Journal of Molecular Biology. – 1987. – Vol. 193. – № 4. – P. 723-743.

13. Heinz, S. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. / S. Heinz, C. Benner, N. Spann et al. // Molecular cell. – 2010. – Vol. 38. – № 4. – P. 576-89.

14. Bailey, T.L. STREME: accurate and versatile sequence motif discovery / T.L. Bailey // Bioinformatics. – 2021. – Vol. 37. – № 18. – P. 2834-2840.

15. Machanick, P. MEME-ChIP: motif analysis of large DNA datasets / P. Machanick, T.L. Bailey // Bioinformatics. – 2011. – Vol. 27. – № 12. – P. 1696-1697.

16. Kulakovskiy, I. V. Deep and wide digging for binding motifs in ChIP-Seq data / I. V. Kulakovskiy, V.A. Boeva, A. V. Favorov, V.J. Makeev // Bioinformatics. – 2010. – Vol. 26. – № 20. – P. 2622-2623.

17. Lloyd, S.M. Pinpointing the Genomic Localizations of Chromatin-Associated Proteins: The Yesterday, Today, and Tomorrow of ChIP-seq / S.M. Lloyd, X. Bao // Current Protocols in Cell Biology. – 2019. – Vol. 84. – № 1.

18. Kulakovskiy, I. V. A deeper look into transcription regulatory code by preferred pair distance templates for transcription factor binding sites / I. V. Kulakovskiy, A.A. Belostotsky, A.S. Kasianov et al. // Bioinformatics. – 2011. – Vol. 27. – № 19. – P. 2621-2624.

19. Nikulova, A.A. CORECLUST: identification of the conserved CRM grammar together with prediction of gene regulation / A.A. Nikulova, A. V. Favorov, R.A. Sutormin et al. // Nucleic Acids Research. – 2012. – Vol. 40. – № 12. – P. e93-e93.

20. Macintyre, G. is-rSNP: a novel technique for in silico regulatory SNP detection / G. Macintyre, J. Bailey, I. Haviv, A. Kowalczyk // Bioinformatics. – 2010. – Vol. 26. – № 18. – P. i524-i530.

21. Boytsov, A. ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs / A. Boytsov, S. Abramov, A.Z.

Aiusheeva et al. // Nucleic Acids Research. – 2022. – Vol. 50. – № W1. – P. W51-W56.

22. Bulyk, M.L. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Vol. 30 / M.L. Bulyk, P.L.F. Johnson, G.M. Church. – Oxford University Press, 2002.

23. Cooper, B.H. DNA binding specificity of all four Saccharomyces cerevisiae forkhead transcription factors / B.H. Cooper, A.C. Dantas Machado, Y. Gan et al. // Nucleic acids research. – 2023. – Vol. 51. – № 11. – P. 5621-5633.

24. Benos, P. V. Additivity in protein-DNA interactions: how good an approximation is it? / P. V. Benos // Nucleic Acids Research. – 2002. – Vol. 30. – № 20. – P. 4442-4451.

25. Keilwagen, J. Varying levels of complexity in transcription factor binding motifs / J. Keilwagen, J. Grau // Nucleic Acids Research. – 2015. – Vol. 43. – № 18. – P. e119-e119.

26. Jolma, A. DNA-dependent formation of transcription factor pairs alters their binding specificity / A. Jolma, Y. Yin, K.R. Nitta et al. // Nature. – 2015. – Vol. 527. – № 7578. – P. 384-388.

27. Rogers, J.M. Bispecific Forkhead Transcription Factor FoxN3 Recognizes Two Distinct Motifs with Different DNA Shapes / J.M. Rogers, C.T. Waters, T.C.M. Seegar et al. // Molecular Cell. – 2019. – Vol. 74. – № 2. – P. 245-253.e6.

28. Amoutzias, G.D. Choose your partners: dimerization in eukaryotic transcription factors / G.D. Amoutzias, D.L. Robertson, Y. Van de Peer, S.G. Oliver // Trends in biochemical sciences. – 2008. – Vol. 33. – № 5. – P. 220-229.

29. Zaret, K.S. Pioneer Transcription Factors, Chromatin Dynamics, and Cell Fate Control / K.S. Zaret, S.E. Mango // Current opinion in genetics & development. – 2016. – Vol. 37. – P. 76.

30. Worsley-Hunt, R. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets / R. Worsley-Hunt, W.W. Wasserman // Genome Biology. – 2014. – Vol. 15. – № 7. – P. 412.

31. Gheorghe, M. A map of direct TF-DNA interactions in the human genome / M. Gheorghe, G.K. Sandve, A. Khan et al. // Nucleic acids research. – 2019. – Vol. 47. – № 4. – P. e21.

32. Karimzadeh, M. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome / M. Karimzadeh, M.M. Hoffman // Genome biology. – 2022. – Vol. 23. – № 1.

33. Levitsky, V. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package / V. Levitsky, E. Zemlyanskaya, D. Oshchepkov et al. // Nucleic Acids Research. – 2019. – Vol. 47. – № 21. – P. e139-e139.

34. Tsukanov, A. V. Motif models proposing independent and interdependent impacts of nucleotides are related to high and low affinity transcription factor binding sites in Arabidopsis / A. V. Tsukanov, V. V. Mironova, V.G. Levitsky // Frontiers in Plant Science. – 2022. – Vol. 13. – P. 2637.

35. Jain, D. Active promoters give rise to false positive “Phantom Peaks” in ChIP-seq experiments / D. Jain, S. Baldi, A. Zabel et al. // Nucleic Acids Research. – 2015. – Vol. 43. – № 14. – P. 6959-6968.

36. Teytelman, L. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins / L. Teytelman, D.M. Thurtle, J. Rine, A. Van Oudenaarden // Proceedings of the National Academy of Sciences of the United States of America. – 2013. – Vol. 110. – № 46. – P. 18602-18607.

37. Mathelier, A. The Next Generation of Transcription Factor Binding Site Prediction / A. Mathelier, W.W. Wasserman // PLoS Computational Biology. – 2013. – Vol. 9. – № 9.

38. Yang, L. TFBSshape: A motif database for DNA shape features of transcription factor binding sites / L. Yang, T. Zhou, I. Dror et al. // Nucleic Acids Research. – 2014. – Vol. 42. – № D1.

39. Siebert, M. Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences / M. Siebert, J. Söding // Nucleic Acids Research. – 2016. – Vol. 44. – № 13. – P. 6055-6069.

40. Eggeling, R. InMoDe: Tools for learning and visualizing intra-motif dependencies of DNA binding sites / R. Eggeling, I. Grosse, J. Grau // Bioinformatics. – 2017. – Vol. 33. – № 4. – P. 580-582.

41. Samee, M.A.H. A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs / M.A.H. Samee, B.G. Bruneau, K.S. Pollard // Cell Systems. – 2019. – Vol. 8. – № 1. – P. 27-42.e6.

42. Eggeling, R. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data / R. Eggeling, T. Roos, P. Myllymäki, I. Grosse // BMC Bioinformatics. – 2015. – Vol. 16. – № 1.

43. Bulyk, M.L. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors / M.L. Bulyk // Nucleic Acids Research. – 2002. – Vol. 30. – № 5. – P. 1255-1261.

44. Levitsky, V.G. Hidden heterogeneity of transcription factor binding sites: A case study of SF-1 / V.G. Levitsky, D.Y. Oshchepkov, N. V. Klimova et al. // Computational Biology and Chemistry. – 2016. – Vol. 64. – P. 19-32.

45. Levitsky, V.G. Application of experimentally verified transcription factor binding sites models for computational analysis of ChIP-Seq data / V.G. Levitsky, I. V. Kulakovskiy, N.I. Ershov et al. // BMC Genomics. – 2014. – Vol. 15. – № 1. – P. 80.

46. Levitsky, V.G. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions / V.G. Levitsky, E. V. Ignatieva, E.A. Ananko et al. // BMC Bioinformatics. – 2007. – Vol. 8. – № 1. – P. 1-20.

47. Lai, X. Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants / X. Lai, A. Stigliani, G. Vachon et al. // Molecular Plant. – 2019. – Vol. 12. – № 6. – P. 743-763.

48. Klemm, S.L. Chromatin accessibility and the regulatory epigenome / S.L. Klemm, Z. Shipony, W.J. Greenleaf // Nature Reviews Genetics. – 2019. – Vol. 20. – № 4. – P. 207-220.

49. Mayran, A. Pioneer transcription factors shape the epigenetic landscape / A. Mayran, J. Drouin // Journal of Biological Chemistry. – 2018. – Vol. 293. – № 36. – P. 13795-13804.

50. Cramer, P. Organization and regulation of gene transcription / P. Cramer // Nature. – 2019. – Vol. 573. – № 7772. – P. 45-54.

51. Reiter, F. Combinatorial function of transcription factors and cofactors / F. Reiter, S. Wienerroither, A. Stark // Current Opinion in Genetics and Development. – 2017. – Vol. 43. – P. 73-81.

52. Hujoel, M.L.A. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species / M.L.A. Hujoel, S. Gazal, F. Hormozdiari et al. // American Journal of Human Genetics. – 2019. – Vol. 104. – № 4. – P. 611-624.

53. Furney, S.J. Structural and functional properties of genes involved in human cancer / S.J. Furney, D.G. Higgins, C.A. Ouzounis, N. López-Bigas // BMC genomics. – 2006. – Vol. 7.

54. Roey, R. van. Deregulation of Transcription Factor Networks Driving Cell Plasticity and Metastasis in Pancreatic Cancer / R. van Roey, T. Brabletz, M.P. Stemmler, I. Armstark // Frontiers in cell and developmental biology. – 2021. – Vol. 9.

55. Boyadjiev, S. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders / S. Boyadjiev, E. Jabs // Clinical genetics. – 2000. – Vol. 57. – № 4. – P. 253-266.

56. Liu, J. Intrinsic disorder in transcription factors / J. Liu, N.B. Perumal, C.J. Oldfield et al. // Biochemistry. – 2006. – Vol. 45. – № 22. – P. 6873-6888.

57. Tang, H. Protein-protein interactions in eukaryotic transcription initiation: Structure of the preinitiation complex / H. Tang, X. Sun, D. Reinberg, R.H. Ebright // Proceedings of the National Academy of Sciences of the United States of America. – 1996. – Vol. 93. – № 3. – P. 1119-1124.

58. Shlyueva, D. Transcriptional enhancers: From properties to genome-wide predictions / D. Shlyueva, G. Stampfel, A. Stark // Nature Reviews Genetics. – 2014. – Vol. 15. – № 4. – P. 272-286.

59. Urrutia, R. KRAB-containing zinc-finger repressor proteins / R. Urrutia // Genome biology. – 2003. – Vol. 4. – № 10.

60. Ecco, G. KRAB zinc finger proteins / G. Ecco, M. Imbeault, D. Trono // Development (Cambridge, England). – 2017. – Vol. 144. – № 15. – P. 2719-2729.

61. Hübner, M.R. Chromatin organization and transcriptional regulation. Vol. 23 / M.R. Hübner, M.A. Eckersley-Maslin, D.L. Spector. – Elsevier Current Trends, 2013.

62. Deplancke, B. The Genetics of Transcription Factor DNA Binding Variation / B. Deplancke, D. Alpern, V. Gardeux // Cell. – 2016. – Vol. 166. – № 3. – P. 538-554.

63. Zaret, K.S. Pioneer Transcription Factors Initiating Gene Network Changes / K.S. Zaret // Annual review of genetics. – 2020. – Vol. 54. – P. 367.

64. Soufi, A. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming / A. Soufi, M.F.F. Garcia, A. Jaroszewicz et al. // Cell. – 2015. – Vol. 161. – № 3. – P. 555-568.

65. Fernandez Garcia, M. Structural Features of Transcription Factors Associating with Nucleosome Binding / M. Fernandez Garcia, C.D. Moore, K.N. Schulz et al. // Molecular Cell. – 2019. – Vol. 75. – № 5. – P. 921-932.e6.

66. Zeitlinger, J. Seven myths of how transcription factors read the cis-regulatory code / J. Zeitlinger // Current Opinion in Systems Biology. – 2020. – Vol. 23. – P. 22-31.

67. Sherwood, R.I. Discovery of non-directional and directional pioneer transcription factors by modeling DNase profile magnitude and shape / R.I. Sherwood, T. Hashimoto, C.W. O'Donnell et al. // Nature biotechnology. – 2014. – Vol. 32. – № 2. – P. 171.

68. Sierra-Pagan, J.E. The regulatory role of pioneer factors during cardiovascular lineage specification – A mini review / J.E. Sierra-Pagan, D.J. Garry // Frontiers in Cardiovascular Medicine. – 2022. – Vol. 9.

69. Iwafuchi-Doi, M. Pioneer transcription factors in cell reprogramming / M. Iwafuchi-Doi, K.S. Zaret // Genes & Development. – 2014. – Vol. 28. – № 24. – P. 2679-2692.

70. Lai, X. Pioneer Factors in Animals and Plants-Colonizing Chromatin for Gene Regulation / X. Lai, L. Verhage, V. Hugouvieux, C. Zubieta // Molecules (Basel, Switzerland). – 2018. – Vol. 23. – № 8.

71. Sunkel, B.D. Pioneer factors in development and cancer / B.D. Sunkel, B.Z. Stanton // iScience. – 2021. – Vol. 24. – № 10.

72. Baumgarten, N. Improved linking of motifs to their TFs using domain information / N. Baumgarten, F. Schmidt, M.H. Schulz // Bioinformatics. – 2020. – Vol. 36. – № 6. – P. 1655.

73. Ehsani, R. Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function / R. Ehsani, S. Bahrami, F. Drabløs // BMC Bioinformatics. – 2016. – Vol. 17. – № 1.

74. Wingender, E. CRITERIA FOR AN UPDATED CLASSIFICATION OF HUMAN TRANSCRIPTION FACTOR DNA-BINDING DOMAINS / E. Wingender // https://doi.org/10.1142/S0219720013400076. – 2013. – Vol. 11. – № 1.

75. Frietze, S. Transcription Factor Effector Domains / S. Frietze, P.J. Farnham // Sub-cellular biochemistry. – 2011. – Vol. 52. – P. 261.

76. Harrison, S.C. A structural taxonomy of DNA-binding domains The structures of several classes of DNA-binding domains reveal a variety of designs for recognizing a specific site on DNA / S.C. Harrison. – 1991.

77. Merkulova, T.I. [Regulatory transcription codes in eukaryotic genomes]. / T.I. Merkulova, E.A. Ananko, E. V. Ignat'eva, N.A. Kolchanov // Genetika. – 2013. – Vol. 49. – № 1. – P. 37-54.

78. Geertz, M. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform / M. Geertz, D. Shore, S.J. Maerkl // Proceedings of the National Academy of Sciences of the United States of America. – 2012. – Vol. 109. – № 41. – P. 16540-16545.

79. Todeschini, A.L. Transcription factors: Specific DNA binding and specific gene regulation / A.L. Todeschini, A. Georges, R.A. Veitia // Trends in Genetics. – 2014. – Vol. 30. – № 6. – P. 211-219.

80. Li, J. Structural basis for DNA recognition by STAT6 / J. Li, J.P. Rodriguez, F. Niu et al. // Proceedings of the National Academy of Sciences of the United States of America. – 2016. – Vol. 113. – № 46. – P. 13015-13020.

81. Wingender, E. TFClass: an expandable hierarchical classification of human transcription factors / E. Wingender, T. Schoeps, J. Dönitz // Nucleic Acids Research. – 2013. – Vol. 41. – № D1. – P. D165-D170.

82. Wingender, E. TFClass: a classification of human transcription factors and their rodent orthologs / E. Wingender, T. Schoeps, M. Haubrock, J. Dönitz // Nucleic Acids Research. – 2015. – Vol. 43. – № D1. – P. D97-D102.

83. Wingender, E. TFClass: expanding the classification of human transcription factors to their mammalian orthologs / E. Wingender, T. Schoeps, M. Haubrock et al. // Nucleic Acids Research. – 2018. – Vol. 46. – № D1. – P. D343-D347.

84. Burley, S.K. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences / S.K. Burley, C. Bhikadiya, C. Bi et al. // Nucleic acids research. – 2021. – Vol. 49. – № D1. – P. D437-D451.

85. Sehnal, D. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures / D. Sehnal, S. Bittrich, M. Deshpande et al. // Nucleic Acids Research. – 2021. – Vol. 49. – № W1. – P. W431-W437.

86. Yin, Z. Activator Protein-1: redox switch controlling structure and DNA-binding / Z. Yin, M. Machius, E.J. Nestler, G. Rudenko // Nucleic Acids Research. – 2017. – Vol. 45. – № 19. – P. 11425-11436.

87. Lu, P. Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4α: Diabetes gene product / P. Lu, G.B. Rha, M. Melikishvili et al. // Journal of Biological Chemistry. – 2008. – Vol. 283. – № 48. – P. 33685-33697.

88. Li, J. Structure of the Forkhead Domain of FOXA2 Bound to a Complete DNA Consensus Site / J. Li, A.C. Dantas Machado, M. Guo et al. // Biochemistry. – 2017. – Vol. 56. – № 29. – P. 3745-3753.

89. Nardini, M. Sequence-specific transcription factor NF-Y displays histone-like DNA binding and H2B-like ubiquitination / M. Nardini, N. Gnesutta, G. Donati et al. // Cell. – 2013. – Vol. 152. – № 1-2. – P. 132-143.

90. Huang, K. Solution structure of the MEF2A–DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors / K. Huang, J.M. Louis, L. Donaldson et al. // The EMBO Journal. – 2000. – Vol. 19. – № 11. – P. 2615-2628.

91. Kitayner, M. Structural Basis of DNA Recognition by p53 Tetramers / M. Kitayner, H. Rozenberg, N. Kessler et al. // Molecular Cell. – 2006. – Vol. 22. – № 6. – P. 741-753.

92. Shi, Y. Crystal structure of a Smad MH1 domain bound to DNA: Insights on DNA binding in TGF-β signaling / Y. Shi, Y.F. Wang, L. Jayaraman et al. // Cell. – 1998. – Vol. 94. – № 5. – P. 585-594.

93. Nikolov, D.B. Crystal structure of a human TATA box-binding protein/TATA element complex. / D.B. Nikolov, H. Chen, E.D. Halay et al. // Proceedings of the National Academy of Sciences. – 1996. – Vol. 93. – № 10. – P. 4862-4867.

94. Kloks, C.P.A.M. The solution structure and DNA-binding properties of the cold-shock domain of the human Y-box protein YB-1 / C.P.A.M. Kloks, C.A.E.M. Spronk, E. Lasonder et al. // Journal of Molecular Biology. – 2002. – Vol. 316. – № 2. – P. 317-326.

95. Chen, H. What do transcription factors interact with? / H. Chen, B.F. Pugh // Journal of molecular biology. – 2021. – Vol. 433. – № 14. – P. 166883.

96. Horikoshi, M. Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex / M. Horikoshi, T. Hai, Y.S. Lin et al. // Cell. – 1988. – Vol. 54. – № 7. – P. 1033-1042.

97. Poss, Z.C. The Mediator complex and transcription regulation / Z.C. Poss, C.C. Ebmeier, D.J. Taatjes // Critical Reviews in Biochemistry and Molecular Biology. – 2013. – Vol. 48. – № 6. – P. 575.

98. Blau, J. Three functional classes of transcriptional activation domain. / J. Blau, H. Xiao, S. McCracken et al. // Molecular and cellular biology. – 1996. – Vol. 16. – № 5. – P. 2044-55.

99. Choy, B. Eukaryotic activators function during multiple steps of preinitiation complex assembly / B. Choy, M.R. Green // Nature. – 1993. – Vol. 366. – № 6455. – P. 531-536.

100. Fuda, N.J. Defining mechanisms that regulate RNA polymerase II transcription in vivo / N.J. Fuda, M.B. Ardehali, J.T. Lis // Nature. – 2009. – Vol. 461. – № 7261. – P. 186-192.

101. Selth, L.A. Transcript Elongation by RNA Polymerase II / L.A. Selth, S. Sigurdsson, J.Q. Svejstrup // Annual review of biochemistry. – 2010. – Vol. 79. – P. 271-293.

102. Minezaki, Y. Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation / Y. Minezaki, K. Homma, A.R. Kinjo, K. Nishikawa // Journal of molecular biology. – 2006. – Vol. 359. – № 4. – P. 1137-1149.

103. Ayed, A. Latent and active p53 are identical in conformation / A. Ayed, F.A.A. Mulder, G.S. Yi et al. // Nature structural biology. – 2001. – Vol. 8. – № 9. – P. 756-760.

104. Dawson, R. The N-terminal domain of p53 is natively unfolded / R. Dawson, L. Müller, A. Dehner et al. // Journal of Molecular Biology. – 2003. – Vol. 332. – № 5. – P. 1131-1141.

105. Garza, A.S. Role of intrinsically disordered protein regions/domains in transcriptional regulation / A.S. Garza, N. Ahmad, R. Kumar // Life sciences. – 2009. – Vol. 84. – № 7-8. – P. 189-193.

106. Morgunova, E. Structural perspective of cooperative transcription factor binding / E. Morgunova, J. Taipale // Current Opinion in Structural Biology. – 2017. – Vol. 47. – P. 1-8.

107. Spitz, F. Transcription factors: From enhancer binding to developmental control / F. Spitz, E.E.M. Furlong // Nature Reviews Genetics. – 2012. – Vol. 13. – № 9. – P. 613-626.

108. Barrett, R.M. Beyond transcription factors: the role of chromatin modifying enzymes in regulating transcription required for memory / R.M. Barrett, M.A. Wood // Learning & memory (Cold Spring Harbor, N.Y.). – 2008. – Vol. 15. – № 7. – P. 460-467.

109. Stampfel, G. Transcriptional regulators form diverse groups with context-dependent regulatory functions / G. Stampfel, T. Kazmar, O. Frank et al. // Nature. – 2015. – Vol. 528. – № 7580. – P. 147-151.

110. Zabidi, M.A. Regulatory enhancer–core-promoter communication via transcription factors and cofactors / M.A. Zabidi, A. Stark // Trends in genetics : TIG. – 2016. – Vol. 32. – № 12. – P. 801.

111. Trouche, D. The CBP co-activator stimulates E2F1/DP1 activity / D. Trouche, A. Cook, T. Kouzarides // Nucleic acids research. – 1996. – Vol. 24. – № 21. – P. 4139-4145.

112. Morris, L. Regulation of E2F transcription by cyclin E-Cdk2 kinase mediated through p300/CBP co-activators / L. Morris, K.E. Allen, N.B. la Thangue // Nature cell biology. – 2000. – Vol. 2. – № 4. – P. 232-239.

113. Ait-Si-Ali, S. CBP/p300 histone acetyl-transferase activity is important for the G1/S transition / S. Ait-Si-Ali, A. Polesskaya, S. Filleur et al. // Oncogene. – 2000. – Vol. 19. – № 20. – P. 2430-2437.

114. Ogawa, H. A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells / H. Ogawa, K.I. Ishiguro, S. Gaubatz et al. // Science (New York, N.Y.). – 2002. – Vol. 296. – № 5570. – P. 1132-1136.

115. Groner, A.C. KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading / A.C. Groner, S. Meylan, A. Ciuffi et al. // PLoS genetics. – 2010. – Vol. 6. – № 3.

116. Russell, R.B. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains / R.B. Russell, J. Breed, G.J. Barton // FEBS Letters. – 1992. – Vol. 304. – № 1. – P. 15-20.

117. Berg, O.G. Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 1. Models and Theory / O.G. Berg, R.B. Winter, P.H. von Hippel // Biochemistry. – 1981. – Vol. 20. – № 24. – P. 6929-6948.

118. Berg, O.G. How do genome-regulatory proteins locate their DNA target sites? Vol. 7 / O.G. Berg, R.B. Winter, P.H. von Hippel. – Elsevier Current Trends, 1982.

119. Berg, O.G. Diffusion-controlled macromolecular interactions. Vol. 14 / O.G. Berg, P.H. Von Hippel. – 1985. – Mode of access: www.annualreviews.org (date of access: 03.03.2021). – [Electronic resource].

120. Schmidt, H.G. An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding / H.G. Schmidt, S. Sewitz, S.S. Andrews, K. Lipkow // PLoS ONE. – 2014. – Vol. 9. – № 10. – P. 108575.

121. Slattery, M. Absence of a simple code: How transcription factors read the genome / M. Slattery, T. Zhou, L. Yang et al. // Trends in Biochemical Sciences. – 2014. – Vol. 39. – № 9. – P. 381-399.

122. Rohs, R. Origins of specificity in protein-DNA recognition / R. Rohs, X. Jin, S.M. West et al. // Annual review of biochemistry. – 2010. – Vol. 79. – P. 233-269.

123. Hancock, S.P. Control of DNA minor groove width and Fis protein binding by the purine 2-amino group / S.P. Hancock, T. Ghane, D. Cascio et al. // Nucleic Acids Research. – 2013. – Vol. 41. – № 13. – P. 6750-6760.

124. Abe, N. Deconvolving the Recognition of DNA Shape from Sequence / N. Abe, I. Dror, L. Yang et al. // Cell. – 2015. – Vol. 161. – № 2. – P. 307-318.

125. Yang, L. Transcription factor family-specific DNA shape readout revealed by quantitative specificity models / L. Yang, Y. Orenstein, A. Jolma et al. // Molecular Systems Biology. – 2017. – Vol. 13. – № 2. – P. 910.

126. Stella, S. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis / S. Stella, D. Cascio, R.C. Johnson // Genes & Development. – 2010. – Vol. 24. – № 8. – P. 814-826.

127. Ponomarenko, J. V. Conformational and physicochemical DNA features specific for transcription factor binding sites / J. V. Ponomarenko, M.P. Ponomarenko, A.S. Frolov et al. // Bioinformatics. – 1999. – Vol. 15. – № 7-8. – P. 654-668.

128. Hombach, D. A systematic, large-scale comparison of transcription factor binding site models / D. Hombach, J.M. Schwarz, P.N. Robinson et al. // BMC genomics. – 2016. – Vol. 17. – № 1.

129. Lawrence, C.E. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment / C.E. Lawrence, S.F. Altschul, M.S. Boguski et al. // Science (New York, N.Y.). – 1993. – Vol. 262. – № 5131. – P. 208-214.

130. Simcha, D. The limits of de novo DNA motif discovery. / D. Simcha, N.D. Price, D. Geman // Plos one. – 2012. – Vol. 7. – № 11. – P. e47836-e47836.

131. Kulakovskiy, I. V. Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources / I. V. Kulakovskiy, V.J. Makeev // Biophysics. – 2009. – Vol. 54. – № 6. – P. 667-674.

132. Boeva, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. Vol. 7 / V. Boeva. – Frontiers Media S.A., 2016.

133. Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. / A. Cornish-Bowden // Nucleic Acids Research. – 1985. – Vol. 13. – № 9. – P. 3021.

134. Stephens, R.M. Sequence logos: a new way to display consensus sequences / R.M. Stephens. – 1990. – Vol. 18. – № 20. – P. 6097-6100.

135. Wasserman, W.W. Applied bioinformatics for the identification of regulatory elements / W.W. Wasserman, A. Sandelin // Nature Reviews Genetics. – 2004. – Vol. 5. – № 4. – P. 276-287.

136. King, O.D. A non-parametric model for transcription factor binding sites / O.D. King // Nucleic Acids Research. – 2003. – Vol. 31. – № 19. – P. 116e-1116.

137. Shannon, C.E. A Mathematical Theory of Communication / C.E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27. – № 3. – P. 379-423.

138. Bailey, T. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data / T. Bailey, P. Krajewski, I. Ladunga et al. // PLoS Computational Biology. – 2013. – Vol. 9. – № 11. – P. 5-12.

139. Castro-Mondragon, J.A. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles / J.A. Castro-Mondragon, R. Riudavets-Puig, I. Rauluseviciute et al. // Nucleic Acids Research. – 2022. – Vol. 50. – № D1. – P. D165-D173.

140. Weirauch, M.T. Determination and inference of eukaryotic transcription factor sequence specificity. / M.T. Weirauch, A. Yang, M. Albu et al. // Cell. – 2014. – Vol. 158. – № 6. – P. 1431-1443.

141. Gupta, S. Quantifying similarity between motifs. / S. Gupta, J.A. Stamatoyannopoulos, T.L. Bailey, W.S. Noble // Genome biology. – 2007. – Vol. 8. – № 2. – P. R24.

142. Dror, I. A widespread role of the motif environment in transcription factor binding across diverse protein families / I. Dror, T. Golan, C. Levy et al. // Genome Research. – 2015. – Vol. 25. – № 9. – P. 1268-1280.

143. Gordân, R. Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape / R. Gordân, N. Shen, I. Dror et al. // Cell Reports. – 2013. – Vol. 3. – № 4. – P. 1093-1104.

144. Castellanos, M. Eukaryotic transcription factors can track and control their target genes using DNA antennas / M. Castellanos, N. Mothi, V. Muñoz // Nature Communications. – 2020. – Vol. 11. – № 1. – P. 1-13.

145. López-Vidriero, I. DNA features beyond the transcription factor binding site specify target recognition by plant MYC2-related bHLH proteins / I. López-Vidriero, M. Godoy, J. Grau et al. // Plant Communications. – 2021. – Vol. 2. – № 6. – P. 100232.

146. Zhang, M.O. A weight array method for splicing signal analysis / M.O. Zhang, T.G. Marr // Bioinformatics. – 1993. – Vol. 9. – № 5. – P. 499-509.

147. Siddharthan, R. Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix / R. Siddharthan // PLOS ONE. – 2010. – Vol. 5. – № 3. – P. e9722.

148. Kulakovskiy, I. V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. / I. V. Kulakovskiy, V. Levitsky, D. Oshchepkov et al. // Journal of Bioinformatics and Computational Biology. – 2013. – Vol. 11. – № 1. – P. 1340004-1340004.

149. Ge, W. Bayesian Markov models improve the prediction of binding motifs beyond first order. / W. Ge, M. Meier, C. Roth, J. Söding // NAR genomics and bioinformatics. – 2021. – Vol. 3. – № 2. – P. lqab026.

150. Eggeling, R. Inhomogeneous parsimonious Markov models / R. Eggeling, A. Gohr, P.Y. Bourguignon et al. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2013. – Vol. 8188 LNAI. – № PART 1. – P. 321-336.

151. Zhou, T. Quantitative modeling of transcription factor binding specificities using DNA shape / T. Zhou, N. Shen, L. Yang et al. // Proceedings of the National Academy of Sciences of the United States of America. – 2015. – Vol. 112. – № 15. – P. 4654-4659.

152. Mathelier, A. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo / A. Mathelier, B. Xin, T.P. Chiu et al. // Cell Systems. – 2016. – Vol. 3. – № 3. – P. 278-286.e4.

153. Rube, H.T. A unified approach for quantifying and interpreting DNA shape readout by transcription factors / H.T. Rube, C. Rastogi, J.F. Kribelbauer, H.J. Bussemaker // Molecular Systems Biology. – 2018. – Vol. 14. – № 2. – P. 1-16.

154. Friedel, M. DiProDB: A database for dinucleotide properties / M. Friedel, S. Nikolajewa, J. Sühnel, T. Wilhelm // Nucleic Acids Research. – 2009. – Vol. 37. – № SUPPL. 1. – P. D37.

155. Chiu, T.P. DNAshapeR: An R/Bioconductor package for DNA shape prediction and feature encoding / T.P. Chiu, F. Comoglio, T. Zhou et al. // Bioinformatics. – 2016. – Vol. 32. – № 8. – P. 1211-1213.

156. Zhang, M.Q. A discrimination study of human core-promoters. / M.Q. Zhang // Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. – 1998. – P. 240-51.

157. Solovyev, V. The Gene-Finder computer tools for analysis of human and model organisms genome sequences / V. Solovyev, A. Salamov. – 1997.

158. Bajić, V.B. Comparing the success of different prediction software in sequence analysis: a review. / V.B. Bajić // Briefings in bioinformatics. – 2000. – Vol. 1. – № 3. – P. 214-228.

159. Fawcett, T. An introduction to ROC analysis / T. Fawcett // Pattern Recognition Letters. – 2006. – Vol. 27. – № 8. – P. 861-874.

160. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection / R. Kohavi, R. Kohavi. – 1995. – P. 1137--1143.

161. Hartmann, H. P-value-based regulatory motif discovery using positional weight matrices / H. Hartmann, E.W. Guthöhrlein, M. Siebert et al. // Genome Research. – 2013. – Vol. 23. – № 1. – P. 181-194.

162. Badis, G. Diversity and Complexity in DNA Recognition by Transcription Factors / G. Badis, M.F. Berger, A.A. Philippakis et al. // Science. – 2009. – Vol. 324. – № 5935. – P. 1720-1723.

163. Kolchanov, N.A. Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes / N.A. Kolchanov, T.I. Merkulova,

E. V. Ignatieva et al. // Briefings in Bioinformatics. – 2007. – Vol. 8. – № 4. – P. 266-274.

164. Eggeling, R. Disentangling transcription factor binding site complexity / R. Eggeling // Nucleic Acids Research. – 2018. – Vol. 46. – № 20. – P. 1-12.

165. Nakagawa, S. DNA-binding specificity changes in the evolution of forkhead transcription factors / S. Nakagawa, S.S. Gisselbrecht, J.M. Rogers et al. // Proceedings of the National Academy of Sciences of the United States of America. – 2013. – Vol. 110. – № 30. – P. 12349-12354.

166. Gabut, M. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming / M. Gabut, P. Samavarchi-Tehrani, X. Wang et al. // Cell. – 2011. – Vol. 147. – № 1. – P. 132-146.

167. Chen, X. Structural basis for DNA recognition by FOXC2 / X. Chen, H. Wei, J. Li et al. // Nucleic Acids Research. – 2019. – Vol. 47. – № 7. – P. 3752-3764.

168. Morgunova, E. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima / E. Morgunova, Y. Yin, P.K. Das et al. // eLife. – 2018. – Vol. 7. – P. 1-21.

169. Rodríguez-Martínez, J.A. Combinatorial bZIP dimers display complex DNA-binding specificity landscapes / J.A. Rodríguez-Martínez, A.W. Reinke, D. Bhimsaria et al. // eLife. – 2017. – Vol. 6.

170. Martin, X. de. Mechanisms of Binding Specificity among bHLH Transcription Factors / X. de Martin, R. Sodaei, G. Santpere // International Journal of Molecular Sciences. – 2021. – Vol. 22. – № 17.

171. Charoensawan, V. Genomic repertoires of DNA-binding transcription factors across the tree of life / V. Charoensawan, D. Wilson, S.A. Teichmann // Nucleic acids research. – 2010. – Vol. 38. – № 21. – P. 7364-7377.

172. Garvie, C.W. Recognition of Specific DNA Sequences / C.W. Garvie, C. Wolberger // Molecular Cell. – 2001. – Vol. 8. – № 5. – P. 937-946.

173. Zeiske, T. Intrinsic DNA Shape Accounts for Affinity Differences between Hox-Cofactor Binding Sites / T. Zeiske, N. Baburajendran, A. Kaczynska et al. // Cell Reports. – 2018. – Vol. 24. – № 9. – P. 2221-2230.

174. Farrel, A. An efficient algorithm for improving structure-based prediction of transcription factor binding sites / A. Farrel, J. tao Guo // BMC Bioinformatics. – 2017. – Vol. 18. – № 1. – P. 1-11.

175. Tomovic, A. Position dependencies in transcription factor binding sites / A. Tomovic, E.J. Oakeley // Bioinformatics. – 2007. – Vol. 23. – № 8. – P. 933-941.

176. Bulyk, M.L. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Vol. 30 / M.L. Bulyk, P.L.F. Johnson, G.M. Church. – Oxford University Press, 2002.

177. Goldshtein, M. Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry / M. Goldshtein, M. Mellul, G. Deutch et al. // Biophysical Journal. – 2020. – Vol. 118. – № 8. – P. 2015-2026.

178. Sela, I. DNA Sequence correlations shape nonspecific transcription factor-DNA binding affinity / I. Sela, D.B. Lukatsky // Biophysical Journal. – 2011. – Vol. 101. – № 1. – P. 160-166.

179. Chumpitaz-Diaz, L. Systematic identification of non-canonical transcription factor motifs / L. Chumpitaz-Diaz, M.A.H. Samee, K.S. Pollard // BMC Molecular and Cell Biology. – 2021. – Vol. 22. – № 1.

180. Jolma, A. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities / A. Jolma, T. Kivioja, J. Toivonen et al. // Genome research. – 2010. – Vol. 20. – № 6. – P. 861-873.

181. Landt, S.G. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. / S.G. Landt, G.K. Marinov, A. Kundaje et al. // Genome research. – 2012. – Vol. 22. – № 9. – P. 1813-31.

182. Davis, C.A. The Encyclopedia of DNA elements (ENCODE): Data portal update / C.A. Davis, B.C. Hitz, C.A. Sloan et al. // Nucleic Acids Research. – 2018. – Vol. 46. – № D1. – P. D794-D801.

183. Hammal, F. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments / F. Hammal, P. De Langen, A. Bergon et al. // Nucleic acids research. – 2022. – Vol. 50. – № D1. – P. D316-D325.

184. Zheng, R. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis / R. Zheng, C. Wan, S. Mei et al. // Nucleic acids research. – 2019. – Vol. 47. – № D1. – P. D729-D735.

185. Kolmykov, S. GTRD: an integrated view of transcription regulation / S. Kolmykov, I. Yevshin, M. Kulyashov et al. // Nucleic Acids Research. – 2021. – Vol. 49. – № D1. – P. D104-D111.

186. Park, S.-J. A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages / S.-J. Park, J.-H. Kim, B.-H. Yoon, S.-Y. Kim // Genomics & Informatics. – 2017. – Vol. 15. – № 1. – P. 11.

187. Conesa, A. A survey of best practices for RNA-seq data analysis / A. Conesa, P. Madrigal, S. Tarazona et al. // Genome Biology. – 2016. – Vol. 17. – № 1. – P. 1-19.

188. Dai, M. NGSQC: Cross-platform quality analysis pipeline for deep sequencing data / M. Dai, R.C. Thompson, C. Maher et al. // BMC Genomics. – 2010. – Vol. 11. – № SUPPL. 4. – P. 1-9.

189. Bolger, A.M. Trimmomatic: a flexible trimmer for Illumina sequence data / A.M. Bolger, M. Lohse, B. Usadel // Bioinformatics. – 2014. – Vol. 30. – № 15. – P. 2114-2120.

190. Callahan, B.J. DADA2: High-resolution sample inference from Illumina amplicon data / B.J. Callahan, P.J. McMurdie, M.J. Rosen et al. // Nature Methods. – 2016. – Vol. 13. – № 7. – P. 581-583.

191. Abnizova, I. Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing. / I. Abnizova, T. Skelly, F. Naumenko et al. // Journal of bioinformatics and computational biology. – 2010. – Vol. 8. – № 3. – P. 579-91.

192. Langmead, B. Fast gapped-read alignment with Bowtie 2 / B. Langmead, S.L. Salzberg // Nature Methods. – 2012. – Vol. 9. – № 4. – P. 357-359.

193. Li, H. Fast and accurate long-read alignment with Burrows–Wheeler transform / H. Li, R. Durbin // Bioinformatics. – 2010. – Vol. 26. – № 5. – P. 589-595.

194. Li, R. SOAP2: an improved ultrafast tool for short read alignment / R. Li, C. Yu, Y. Li et al. // Bioinformatics. – 2009. – Vol. 25. – № 15. – P. 1966-1967.

195. Tran, N.T.L. A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data / N.T.L. Tran, C.H. Huang // Biology Direct. – 2014. – Vol. 9. – № 1. – P. 1-22.

196. Zhang, Y. Model-based analysis of ChIP-Seq (MACS). / Y. Zhang, T. Liu, C.A. Meyer et al. // Genome biology. – 2008. – Vol. 9. – № 9. – P. R137.

197. Guo, Y. High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints / Y. Guo, S. Mahony, D.K. Gifford // PLoS Computational Biology. – 2012. – Vol. 8. – № 8. – P. e1002638.

198. Zhang, X. PICS: probabilistic inference for ChIP-seq / X. Zhang, G. Robertson, M. Krzywinski et al. // Biometrics. – 2011. – Vol. 67. – № 1. – P. 151-163.

199. Narlikar, L. ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder / L. Narlikar, R. Jothi // Methods in molecular biology (Clifton, N.J.). – 2012. – Vol. 802. – P. 305-322.

200. Kallio, A. Optimizing detection of transcription factor-binding sites in ChIP-seq experiments / A. Kallio, L.L. Elo // Methods in Molecular Biology. – 2013. – Vol. 1038. – № 1. – P. 181-191.

201. Chung, D. MOSAiCS-HMM: A model-based approach for detecting regions of histone modifications from ChIP-seq data / D. Chung, Q. Zhang, S. Keleş // Statistical Analysis of Next Generation Sequencing Data. – 2014. – P. 277-295.

202. Goren, E. BinQuasi: a peak detection method for ChIP-sequencing data with biological replicates / E. Goren, P. Liu, C. Wang et al. // Bioinformatics. – 2018. – Vol. 34. – № 17. – P. 2909-2917.

203. Thomas, R. Features that define the best ChIP-seq peak calling algorithms / R. Thomas, S. Thomas, A.K. Holloway, K.S. Pollard // Briefings in bioinformatics. – 2017. – Vol. 18. – № 3. – P. 441-450.

204. Zou, Z. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data / Z. Zou,

T. Ohta, F. Miura, S. Oki // Nucleic acids research. – 2022. – Vol. 50. – № W1. – P. W175-W182.

205. Lavrekha, V. V. CisCross: A gene list enrichment analysis to predict upstream regulators in Arabidopsis thaliana / V. V. Lavrekha, V.G. Levitsky, A. V. Tsukanov et al. // Frontiers in Plant Science. – 2022. – Vol. 13. – P. 2919.

206. Hashim, F.A. Review of Different Sequence Motif Finding Algorithms. / F.A. Hashim, M.S. Mabrouk, W. Al-Atabany // Avicenna journal of medical biotechnology. – 2019. – Vol. 11. – № 2. – P. 130-148.

207. Bailey, T.L. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. / T.L. Bailey, C. Elkan // Proceedings of International Conference on Intelligent Systems for Molecular Biology. – 1994. – Vol. 2. – P. 28-36.

208. Kiesel, A. The BaMM web server for de-novo motif discovery and regulatory sequence analysis / A. Kiesel, C. Roth, W. Ge et al. // Nucleic Acids Research. – 2018. – Vol. 46. – № W1. – P. W215-W220.

209. Khan, A. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. / A. Khan, O. Fornes, A. Stigliani et al. // Nucleic acids research. – 2018. – Vol. 46. – № D1. – P. D260-D266.

210. DeFord, P. DNA shape complements sequence-based representations of transcription factor binding sites / P. DeFord, J. Taylor // bioRxiv. – 2019. – P. 1-15.

211. Ignatieva, E. V. 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 / E. V. Ignatieva, D.Yu. Oshchepkov, V.G. Levitsky et al. // Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure, Vol 1. – 2004. – Vol. 1. – P. 69-72.

212. Wallerman, O. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing / O. Wallerman, M. Motallebipour, S. Enroth et al. // Nucleic Acids Research. – 2009. – Vol. 37. – № 22. – P. 7498-7508.

213. Wederell, E.D. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing / E.D. Wederell, M. Bilenky, R. Cullum et al. // Nucleic Acids Research. – 2008. – Vol. 36. – № 14. – P. 4549-4564.

214. Worsley-Hunt, R. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment / R. Worsley-Hunt, A. Mathelier, L. del Peso, W.W. Wasserman // BMC Genomics. – 2014. – Vol. 15. – № 1. – P. 472.

215. Quinlan, A.R. BEDTools: A flexible suite of utilities for comparing genomic features / A.R. Quinlan, I.M. Hall // Bioinformatics. – 2010. – Vol. 26. – № 6. – P. 841-842.

216. Yu, G. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization / G. Yu, L.-G. Wang, Q.-Y. He // Bioinformatics. – 2015. – Vol. 31. – № 14. – P. 2382-2383.

217. Yu, G. ClusterProfiler: An R package for comparing biological themes among gene clusters / G. Yu, L.G. Wang, Y. Han, Q.Y. He // OMICS A Journal of Integrative Biology. – 2012. – Vol. 16. – № 5. – P. 284-287.

218. Benjamini, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing / Y. Benjamini, Y. Hochberg // Journal of the Royal Statistical Society: Series B (Methodological). – 1995. – Vol. 57. – № 1. – P. 289-300.

219. Harris, C.R. Array programming with NumPy / C.R. Harris, K.J. Millman, S.J. van der Walt et al. // Nature. – 2020. – Vol. 585. – № 7825. – P. 357-362.

220. Virtanen, P. SciPy 1.0: fundamental algorithms for scientific computing in Python / P. Virtanen, R. Gommers, T.E. Oliphant et al. // Nature Methods 2020 17:3. – 2020. – Vol. 17. – № 3. – P. 261-272.

221. Tareen, A. Logomaker: beautiful sequence logos in Python / A. Tareen, J.B. Kinney // Bioinformatics (Oxford, England). – 2020. – Vol. 36. – № 7. – P. 2272-2274.

222. Grau, J. DepLogo: visualizing sequence dependencies in R / J. Grau, M. Nettling, J. Keilwagen // Bioinformatics. – 2019. – Vol. 35. – № 22. – P. 4812-4814.

223. Hunter, J.D. Matplotlib: A 2D Graphics Environment / J.D. Hunter // Computing in Science & Engineering. – 2007. – Vol. 9. – № 3. – P. 90-95.

224. Waskom, M.L. seaborn: statistical data visualization / M.L. Waskom // Journal of Open Source Software. – 2021. – Vol. 6. – № 60. – P. 3021.

225. Blanc-Mathieu, R. Plant-TFClass: a structural classification for plant transcription factors / R. Blanc-Mathieu, R. Dumas, L. Turchi et al. // bioRxiv. – 2022. – P. 2022.11.22.517060.

226. Ambrosini, G. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study / G. Ambrosini, I. Vorontsov, D. Penzar et al. // Genome Biology. – 2020. – Vol. 21. – № 1. – P. 114.

227. McLeay, R.C. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data / R.C. McLeay, T.L. Bailey // BMC bioinformatics. – 2010. – Vol. 11.

228. Raditsa, V. Massive comparison of the ‘genomic’ and ‘shuffled’ background set generation approaches for efficiency of de novo motif search in A. thaliana ChIP-seq data / V. Raditsa, A. Tsukanov, V. Levitsky // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022): The Thirteenth International Multiconference . – Novosibirsk : Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences, 2022. – P. 90-91.

229. Crooks, G.E. WebLogo: A Sequence Logo Generator / G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner // Genome Research. – 2004. – Vol. 14. – № 6. – P. 1188-1190.

230. Gramzow, L. A hitchhiker’s guide to the MADS world of plants / L. Gramzow, G. Theissen // Genome biology. – 2010. – Vol. 11. – № 6.

231. Goslin, K. Floral Homeotic Factors: A Question of Specificity / K. Goslin, A. Finocchio, F. Wellmer // Plants (Basel, Switzerland). – 2023. – Vol. 12. – № 5.

232. Käppel, S. Cracking the Floral Quartet Code: How Do Multimers of MIKCC-Type MADS-Domain Transcription Factors Recognize Their Target Genes? / S. Käppel, F. Rümpler, G. Theißen // International journal of molecular sciences. – 2023. – Vol. 24. – № 9.

233. Bakshi, M. WRKY transcription factors: Jack of many trades in plants. / M. Bakshi, R. Oelmüller // Plant Signaling & Behavior. – 2014. – Vol. 9. – № 2. – P. e27700-e27700.

234. Fedotova, A.A. C2H2 Zinc Finger Proteins: The Largest but Poorly Explored Family of Higher Eukaryotic Transcription Factors / A.A. Fedotova, A.N. Bonchuk, V.A. Mogila, P.G. Georgiev // Acta Naturae. – 2017. – Vol. 9. – № 2. – P. 47.

235. Han, G. C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants / G. Han, C. Lu, J. Guo et al. // Frontiers in Plant Science. – 2020. – Vol. 11. – P. 115.

236. Liu, Q. Genome-Wide Analysis of C2H2 Zinc-Finger Family Transcription Factors and Their Responses to Abiotic Stresses in Poplar (Populus trichocarpa) / Q. Liu, Z. Wang, X. Xu et al. // PLoS ONE. – 2015. – Vol. 10. – № 8.

237. Chen, J. Zinc-Finger Transcription Factor ZAT6 Positively Regulates Cadmium Tolerance through the Glutathione-Dependent Pathway in Arabidopsis / J. Chen, L. Yang, X. Yan et al. // Plant Physiology. – 2016. – Vol. 171. – № 1. – P. 707-719.

238. Tang, W. Overexpression of Zinc Finger Transcription Factor ZAT6 Enhances Salt Tolerance / W. Tang, C. Luo // Open Life Sciences. – 2018. – Vol. 13. – № 1. – P. 431-445.

239. Heard, N.A. Choosing between methods of combining p -values / N.A. Heard, P. Rubin-Delanchy // Biometrika. – 2018. – Vol. 105. – № 1. – P. 239-246.

Приложение А

Таблица 1. ChIP-seq данные, используемые в анализе ТФ для *A. thaliana*

ID	ТФ	AME (p-value)	TomTom (p-value)		
			PWM	BaMM	SiteGA
PEAKS042938	ABF1	0.00E+00	8.68E-10	3.99E-10	3.15E-07
PEAKS042939	ABF1	0.00E+00	4.18E-10	2.20E-10	1.66E-09
PEAKS042900	ABF3	0.00E+00	3.79E-11	5.76E-10	3.43E-07
PEAKS042901	ABF3	0.00E+00	1.61E-10	4.46E-11	1.10E-07
PEAKS042902	ABF4	0.00E+00	1.88E-10	1.04E-11	6.77E-10
PEAKS042903	ABF4	0.00E+00	2.08E-11	3.16E-11	2.99E-07
PEAKS042822	AGL27	8.94E-154	5.29E-22	3.57E-19	3.01E-02
PEAKS042554	AGL8	1.04E-123	1.79E-01	6.34E-02	1.32E-05
PEAKS042897	AGL8	2.24E-107	9.00E-01	9.08E-01	6.71E-02
PEAKS042553	AGL8	2.78E-216	4.17E-13	5.46E-01	8.88E-04
PEAKS042817	AP1	3.68E-77	9.03E-12	7.59E-11	4.46E-03
PEAKS042819	AP1	2.21E-49	5.75E-13	1.73E-14	1.63E-04
PEAKS042818	AP1	1.01E-71	7.21E-02	1.29E-01	4.60E-01
PEAKS042765	AP2	1.00E-08	9.17E-02	5.33E-02	4.05E-03
PEAKS042831	ARF6	1.09E-97	1.59E-02	3.52E-02	1.35E-01
PEAKS046124	ARR1	1.11E-09	2.17E-01	9.15E-02	2.07E-02
PEAKS046126	ARR1	1.00E+00	1.74E-02	2.05E-02	4.21E-03
PEAKS046127	ARR1	9.12E-103	1.59E-01	5.60E-02	7.42E-02
PEAKS046128	ARR1	1.33E-07	4.14E-02	4.28E-02	1.48E-03
PEAKS046125	ARR1	3.86E-23	3.63E-01	1.14E-01	3.70E-03
PEAKS046123	ARR1	4.12E-20	1.17E-01	1.34E-01	5.21E-02
PEAKS046129	ARR10	1.08E-42	1.46E-01	1.28E-01	5.06E-04
PEAKS046131	ARR12	1.05E-13	7.37E-02	1.30E-01	1.66E-03
PEAKS046130	ARR12	3.50E-05	3.31E-02	1.06E-01	7.48E-03
PEAKS046132	ARR14	6.41E-91	2.75E-03	3.76E-02	4.15E-04
PEAKS042907	AT5G04760	4.95E-132	2.86E-06	4.78E-02	4.23E-02
PEAKS042906	AT5G04760	3.52E-119	1.79E-02	1.36E-02	2.49E-03
PEAKS042912	ATHB-5	4.71E-264	1.38E-06	1.94E-06	3.57E-01
PEAKS042913	ATHB-5	2.14E-259	4.15E-07	7.93E-07	6.00E-02
PEAKS042910	ATHB-6	1.64E-25	1.69E-01	5.59E-03	2.46E-01
PEAKS042911	ATHB-6	8.71E-29	4.52E-03	9.87E-03	1.64E-02
PEAKS042904	ATHB-7	7.67E-212	2.11E-07	1.13E-06	4.73E-02
PEAKS042905	ATHB-7	1.23E-240	2.04E-06	1.01E-07	9.93E-03
PEAKS042945	AZF1	1.58E-16	9.40E-02	1.45E-01	6.31E-03
PEAKS042833	BBM	1.60E-30	1.33E-01	4.96E-02	3.59E-02
PEAKS042834	BBM	1.19E-30	1.74E-01	1.12E-02	3.28E-02
PEAKS042924	BHLH122	1.63E-126	5.20E-10	1.29E-06	6.39E-03
PEAKS042925	BHLH122	5.95E-152	3.02E-08	1.97E-07	3.49E-04
PEAKS042946	BPC1	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS042881	CCA1	4.01E-245	1.16E-06	3.59E-06	1.43E-05
PEAKS042882	CCA1	1.06E-277	4.19E-05	6.50E-06	3.08E-03
PEAKS042920	DREB2A	5.78E-20	1.16E-05	2.34E-06	5.95E-01
PEAKS042921	DREB2A	3.31E-281	4.05E-08	2.40E-08	1.44E-02
PEAKS042826	ERF115	2.41E-128	2.60E-07	2.23E-06	7.44E-03
PEAKS042875	FHY3	1.00E+00	2.86E-08	2.91E-08	2.73E-02
PEAKS042982	GATA21	1.00E+00	2.24E-01	1.12E-01	5.64E-01
PEAKS042983	GATA22	1.00E+00	1.15E-03	3.12E-02	1.92E-02
PEAKS042928	GBF2	0.00E+00	3.05E-12	1.74E-12	2.31E-09

PEAKS042929	GBF2	0.00E+00	1.38E-12	3.38E-12	2.83E-08
PEAKS042937	GBF3	0.00E+00	2.31E-10	1.99E-12	2.06E-09
PEAKS042936	GBF3	0.00E+00	1.45E-10	1.29E-12	7.86E-08
PEAKS042927	HAT22	1.47E-233	7.56E-06	1.83E-05	8.65E-03
PEAKS042926	HAT22	2.34E-213	1.11E-07	2.60E-07	1.42E-01
PEAKS042890	HSFA1A	1.45E-16	8.39E-02	4.92E-02	4.95E-01
PEAKS042892	HSFA1A	2.60E-37	3.39E-06	3.47E-02	7.94E-01
PEAKS042891	HSFA1A	1.02E-16	1.28E-01	5.04E-01	4.72E-01
PEAKS046116	HSFA1B	3.06E-39	8.49E-01	2.56E-01	6.24E-01
PEAKS046117	HSFA1B	1.01E-35	5.43E-01	4.55E-01	3.43E-02
PEAKS042917	HSFA6A	1.62E-221	1.11E-06	3.13E-07	2.66E-01
PEAKS046122	HY5	0.00E+00	1.14E-10	9.78E-11	1.06E-02
PEAKS046121	HY5	0.00E+00	1.25E-10	6.71E-11	1.80E-06
PEAKS050107	IDD4	9.40E-117	4.94E-05	6.65E-05	5.04E-02
PEAKS042821	KAN1	4.35E-53	1.06E-02	2.89E-07	8.33E-03
PEAKS042981	LFY	1.00E+00	6.48E-07	1.53E-06	8.40E-01
PEAKS042832	LHY	5.86E-252	2.17E-07	1.01E-07	7.07E-05
PEAKS056070	MBD9	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS042923	MYB3	2.71E-211	3.23E-08	1.27E-08	2.31E-04
PEAKS042922	MYB3	1.74E-173	6.01E-09	3.32E-08	1.58E-04
PEAKS042915	MYB44	1.57E-138	2.74E-02	1.99E-02	3.77E-04
PEAKS042914	MYB44	6.03E-118	1.62E-01	9.37E-02	6.24E-03
PEAKS058394	MYC2	1.42E-301	1.02E-07	7.53E-07	2.22E-04
PEAKS042932	NAC032	1.94E-117	2.46E-03	3.41E-06	9.40E-03
PEAKS042933	NAC032	1.61E-131	1.39E-03	7.31E-03	5.39E-03
PEAKS058156	PHE1	1.00E+00	8.48E-07	5.09E-07	1.75E-05
PEAKS058157	PHE1	1.00E+00	8.87E-05	4.26E-06	1.74E-02
PEAKS058160	PIE1	1.00E+00	3.04E-01	1.42E-01	5.66E-03
PEAKS042805	PIF1	0.00E+00	1.34E-08	1.26E-08	7.47E-06
PEAKS042804	PIF1	0.00E+00	4.13E-08	1.33E-08	1.34E-05
PEAKS042806	PIF1	0.00E+00	3.00E-08	3.02E-09	5.87E-04
PEAKS042809	PIF4	3.1e-319	1.19E-07	3.15E-09	3.68E-06
PEAKS042873	PIF4	0.00E+00	4.05E-09	2.83E-09	1.75E-02
PEAKS042779	PIF4	2.87E-298	1.12E-07	5.89E-09	3.23E-04
PEAKS042874	PIF5	0.00E+00	1.97E-08	6.13E-09	6.19E-06
PEAKS042778	PIF5	0.00E+00	2.30E-09	1.15E-08	2.29E-06
PEAKS042991	REF6	7.74E-23	2.00E-10	8.39E-12	2.86E-05
PEAKS047354	REF6	6.00E-20	9.42E-11	1.77E-12	3.56E-04
PEAKS047357	REF6	3.32E-16	6.70E-02	6.48E-02	3.01E-01
PEAKS042992	REF6	2.62E-23	3.95E-11	7.77E-12	3.25E-04
PEAKS047355	REF6	1.91E-20	2.89E-10	3.63E-12	3.69E-07
PEAKS042988	REF6	6.25E-20	3.45E-10	9.51E-12	6.68E-06
PEAKS042863	REF6	3.11E-18	1.20E-10	1.94E-12	3.17E-05
PEAKS042771	REV	4.95E-88	8.42E-02	7.57E-02	3.80E-02
PEAKS042845	RGA	1.00E+00	9.88E-01	9.48E-01	8.26E-01
PEAKS042963	RGA	3.56E-18	8.02E-01	7.99E-01	5.37E-01
PEAKS042820	SEP3	6.42E-110	3.48E-16	2.74E-13	3.78E-05
PEAKS042816	SEP3	2.80E-145	1.53E-13	1.15E-15	5.15E-01
PEAKS042815	SEP3	7.98E-143	4.44E-10	5.57E-13	6.01E-05
PEAKS042884	SOC1	2.15E-216	1.47E-12	2.10E-16	3.70E-06
PEAKS042841	SPCH	1.50E-133	3.38E-09	1.35E-07	1.25E-06
PEAKS055376	TCP4	0.00E+00	3.84E-06	4.49E-08	8.61E-06
PEAKS055375	TCP4	6.27E-54	1.10E-03	1.09E-03	8.82E-03
PEAKS042877	TRB1	0.00E+00	2.89E-08	1.72E-08	3.21E-03

PEAKS042876	TRB1	4.46E-253	9.22E-05	3.73E-08	1.33E-04
PEAKS042950	WRKY18	1.08E-219	1.38E-09	1.61E-09	2.30E-05
PEAKS042956	WRKY18	5.03E-169	1.96E-09	6.72E-10	4.96E-07
PEAKS042949	WRKY18	6.58E-154	8.34E-06	2.40E-01	9.51E-04
PEAKS042955	WRKY18	8.35E-114	2.02E-08	2.74E-08	2.45E-02
PEAKS042869	WRKY33	4.73E-121	5.11E-05	1.64E-05	1.09E-01
PEAKS042954	WRKY33	7.04E-212	7.45E-09	1.70E-07	1.21E-05
PEAKS042948	WRKY33	2.88E-252	4.14E-11	4.07E-09	1.69E-02
PEAKS042947	WRKY33	2.64E-110	2.43E-07	1.51E-09	4.15E-02
PEAKS042953	WRKY33	1.03E-23	2.43E-06	1.57E-01	6.05E-06
PEAKS042868	WRKY33	1.20E-03	1.79E-01	4.50E-02	7.76E-02
PEAKS042957	WRKY40	2.58E-16	1.39E-05	6.32E-02	2.19E-03
PEAKS042951	WRKY40	2.21E-15	3.80E-04	1.46E-01	5.43E-02
PEAKS042952	WRKY40	5.13E-247	9.44E-09	7.79E-08	1.62E-04
PEAKS042958	WRKY40	8.11E-168	4.37E-08	1.79E-08	8.83E-07
PEAKS058162	WUS	3.06E-07	2.07E-03	1.96E-03	5.70E-02
PEAKS058163	WUS	2.74E-03	1.64E-01	2.73E-01	4.40E-02
PEAKS042908	ZAT6	4.68E-21	5.90E-06	5.19E-06	4.29E-03
PEAKS042909	ZAT6	1.07E-17	3.75E-06	2.01E-07	2.99E-01

Примечание. ID – уникальный идентификатор базы данных GTRD; AME – результат обогащения частотной матрицы целевого ТФ; TomTom – результат сравнения частотных матриц, полученных с помощью *de novo*, с частотными матрицами целевых ТФ с помощью программы TomTom.

Таблица 2. ChIP-seq данные, используемые в анализе ТФ для *M. musculus*

ID	**ТФ**	**AME (p-value)**	**TomTom (p-value)**		
			PWM	**BaMM**	**SiteGA**
PEAKS035844	AHR	6.47E-17	1.00E+00	1.00E+00	1.00E+00
PEAKS040810	AHR	5.36E-250	1.00E+00	1.00E+00	1.00E+00
PEAKS040811	AHR	5.33E-241	1.00E+00	1.00E+00	1.00E+00
PEAKS040812	AHR	5.98E-253	1.00E+00	1.00E+00	1.00E+00
PEAKS040813	AHR	4.23E-257	1.00E+00	1.00E+00	1.00E+00
PEAKS040814	AHR	1.36E-277	1.00E+00	1.00E+00	1.00E+00
PEAKS040815	AHR	1.08E-217	1.00E+00	1.00E+00	1.00E+00
PEAKS040818	AHR	1.10E-222	1.00E+00	1.00E+00	1.00E+00
PEAKS040612	AIRE	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035586	AR	0.00E+00	3.13E-11	1.92E-11	4.55E-03
PEAKS035588	AR	0.00E+00	1.38E-12	5.80E-13	1.55E-02
PEAKS035590	AR	5.42E-270	1.02E-12	7.09E-13	8.84E-04
PEAKS035592	AR	7.70E-306	6.80E-13	4.31E-13	1.57E-04
PEAKS036281	AR	0.00E+00	5.86E-13	2.68E-13	3.45E-06
PEAKS036285	AR	0.00E+00	1.09E-12	6.08E-13	1.61E-03
PEAKS036287	AR	6.23E-292	1.57E-11	1.00E-11	1.49E-03
PEAKS036289	AR	0.00E+00	3.22E-11	3.24E-11	3.24E-04
PEAKS036541	AR	0.00E+00	1.70E-13	1.25E-13	3.63E-04
PEAKS036542	AR	0.00E+00	4.58E-13	4.23E-12	2.04E-05
PEAKS036543	AR	0.00E+00	3.76E-12	4.21E-12	2.35E-06
PEAKS036544	AR	0.00E+00	2.25E-12	2.98E-12	2.67E-05
PEAKS036545	AR	0.00E+00	4.46E-14	2.23E-14	3.64E-02
PEAKS050693	AR	0.00E+00	1.01E-11	8.58E-12	4.98E-04

PEAKS050694	AR	0.00E+00	3.01E-12	3.48E-12	1.14E-04
PEAKS057419	AR	7.98E-113	4.66E-12	3.26E-12	3.96E-03
PEAKS055351	ARNT2	4.92E-182	5.73E-04	1.46E-03	1.32E-03
PEAKS055352	ARNT2	5.99E-128	9.97E-05	8.57E-05	5.68E-04
PEAKS050661	ARNTL	9.38E-273	5.29E-07	1.82E-07	4.79E-02
PEAKS050663	ARNTL	1.34E-244	6.86E-07	3.94E-07	2.27E-03
PEAKS050664	ARNTL	2.03E-263	1.85E-05	2.69E-07	1.37E-01
PEAKS050674	ARNTL	9.06E-237	6.97E-07	2.97E-07	1.24E-01
PEAKS050676	ARNTL	2.43E-210	5.64E-07	1.02E-05	2.09E-03
PEAKS050678	ARNTL	5.71E-198	7.76E-07	2.71E-07	1.21E-02
PEAKS050680	ARNTL	1.26E-238	3.62E-07	6.50E-06	4.72E-04
PEAKS057547	ARNTL	1.41E-123	1.59E-05	2.97E-07	5.36E-04
PEAKS057548	ARNTL	8.83E-214	9.95E-06	3.56E-06	1.14E-02
PEAKS057549	ARNTL	1.33E-177	6.42E-07	6.28E-06	6.65E-04
PEAKS057550	ARNTL	4.74E-161	8.77E-07	6.59E-06	1.19E-05
PEAKS057551	ARNTL	1.33E-160	1.46E-06	8.00E-06	2.61E-02
PEAKS057553	ARNTL	4.93E-224	8.06E-07	3.42E-07	1.39E-03
PEAKS057554	ARNTL	1.64E-220	5.72E-07	3.77E-07	8.00E-03
PEAKS038157	ASCL1	0.00E+00	1.89E-07	3.63E-09	3.01E-07
PEAKS037311	ATF3	3.84E-276	1.85E-10	7.12E-10	2.31E-03
PEAKS040983	ATF3	5.44E-296	9.41E-07	3.59E-08	4.56E-02
PEAKS040984	ATF3	0.00E+00	1.65E-06	3.25E-07	2.98E-06
PEAKS040985	ATF3	0.00E+00	1.21E-10	3.43E-07	3.33E-03
PEAKS040986	ATF3	0.00E+00	3.08E-10	5.96E-07	2.20E-04
PEAKS040987	ATF3	0.00E+00	9.66E-06	4.14E-08	4.27E-05
PEAKS054850	ATF3	0.00E+00	7.73E-11	1.55E-06	6.23E-04
PEAKS054851	ATF3	5.14E-244	4.78E-07	4.02E-07	8.40E-04
PEAKS054866	ATF3	0.00E+00	4.60E-10	4.77E-07	3.07E-04
PEAKS054867	ATF3	7.17E-245	9.96E-11	1.03E-09	4.26E-02
PEAKS035427	ATOH1	4.17e-317	1.06E-07	7.67E-10	1.16E-06
PEAKS040056	ATOH1	3.71E-76	4.77E-07	3.59E-05	2.70E-01
PEAKS040057	ATOH1	2.70E-72	3.44E-01	5.48E-01	2.46E-02
PEAKS036204	BACH2	0.00E+00	1.09E-07	3.17E-08	9.23E-04
PEAKS040349	BACH2	0.00E+00	6.52E-09	1.88E-08	2.75E-03
PEAKS035635	BATF	7.72E-261	6.34E-06	1.74E-06	4.13E-04
PEAKS036037	BATF	2.40E-230	4.77E-06	4.50E-08	5.31E-03
PEAKS036040	BATF	6.80E-118	1.22E-08	6.37E-06	4.37E-01
PEAKS036041	BATF	4.06E-175	1.88E-06	7.63E-06	1.97E-04
PEAKS036045	BATF	0.00E+00	4.84E-08	1.55E-07	5.20E-05
PEAKS036582	BATF	8.76E-194	6.99E-06	3.79E-06	2.45E-04
PEAKS040193	BATF	5.01E-296	3.15E-07	3.07E-07	2.88E-04
PEAKS040269	BATF	3.77E-258	4.19E-08	6.57E-07	8.94E-03
PEAKS040270	BATF	1.69E-286	3.40E-08	3.95E-06	1.43E-04
PEAKS041042	BATF	1.48E-301	8.63E-09	1.01E-07	2.70E-03
PEAKS041043	BATF	1.18e-311	9.46E-08	9.14E-08	1.79E-03
PEAKS041044	BATF	1.03E-262	1.55E-06	6.50E-07	7.19E-05

PEAKS041045	BATF	7.55E-297	1.16E-06	2.54E-07	2.92E-05
PEAKS041046	BATF	9.72E-48	4.88E-01	9.32E-01	9.82E-01
PEAKS041047	BATF	9.41E-206	7.58E-06	4.34E-06	7.87E-04
PEAKS041048	BATF	7.08E-185	1.10E-08	6.18E-01	9.56E-01
PEAKS041049	BATF	3.73E-247	3.60E-06	1.70E-07	7.90E-02
PEAKS041050	BATF	1.32E-201	8.93E-07	2.21E-06	8.23E-01
PEAKS041051	BATF	3.56E-164	2.79E-06	2.36E-06	8.08E-01
PEAKS041052	BATF	1.72E-276	8.01E-07	1.92E-07	1.10E-04
PEAKS041054	BATF	1.21E-281	6.22E-07	3.39E-08	4.45E-04
PEAKS041055	BATF	9.68E-308	2.47E-07	5.23E-08	1.56E-01
PEAKS037876	BATF3	5.45E-24	1.22E-02	1.06E-02	7.87E-01
PEAKS040668	BCL11B	1.53E-70	1.25E-04	1.12E-05	5.60E-02
PEAKS048666	BCL11B	1.64E-36	2.16E-04	3.97E-04	1.05E-03
PEAKS049421	BCL11B	4.34E-36	5.84E-03	1.52E-04	7.70E-03
PEAKS049422	BCL11B	2.50E-51	7.76E-03	5.66E-04	4.48E-03
PEAKS049423	BCL11B	2.86E-32	8.40E-03	1.39E-06	3.21E-03
PEAKS057190	BCL11B	5.25E-18	1.00E+00	1.00E+00	1.00E+00
PEAKS057194	BCL11B	6.22E-22	1.00E+00	1.00E+00	1.00E+00
PEAKS057195	BCL11B	2.08E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS035263	BCL6	3.32E-102	3.49E-08	2.81E-08	4.25E-03
PEAKS035264	BCL6	1.04E-154	3.28E-09	2.18E-09	1.21E-03
PEAKS038427	BCL6	2.56E-67	5.73E-03	1.11E-02	6.11E-03
PEAKS055017	BCL6	1.57E-157	1.81E-08	2.59E-08	1.19E-02
PEAKS055018	BCL6	5.47E-108	8.49E-08	1.08E-07	2.50E-02
PEAKS055019	BCL6	1.52E-108	2.67E-09	6.08E-10	9.02E-03
PEAKS055020	BCL6	7.58E-158	1.97E-02	1.40E-02	1.49E-02
PEAKS055021	BCL6	1.49E-173	9.14E-04	8.36E-04	5.09E-03
PEAKS055022	BCL6	8.49E-146	1.09E-08	2.85E-08	3.04E-02
PEAKS057221	BCL6	1.88E-205	1.72E-08	1.57E-11	3.97E-03
PEAKS039234	BHLHA15	0.00E+00	1.38E-07	1.69E-07	3.51E-10
PEAKS039235	BHLHA15	0.00E+00	3.28E-07	7.94E-08	1.91E-06
PEAKS055622	BHLHE40	0.00E+00	3.88E-09	1.79E-06	1.72E-04
PEAKS055623	BHLHE40	0.00E+00	1.91E-06	1.22E-08	1.57E-03
PEAKS057576	BHLHE40	3.38E-273	5.40E-08	6.03E-06	2.71E-03
PEAKS040678	BHLHE41	5.54e-311	1.31E-06	4.09E-06	5.12E-05
PEAKS040679	BHLHE41	1.84e-318	1.67E-08	8.47E-06	5.51E-04
PEAKS035381	CDX2	0.00E+00	2.43E-11	8.81E-12	2.14E-04
PEAKS054441	CDX2	2.42E-184	4.07E-07	1.85E-06	1.08E-02
PEAKS055170	CDX2	4.84E-14	1.00E+00	1.00E+00	1.00E+00
PEAKS035677	CEBPA	0.00E+00	2.57E-10	9.24E-10	1.95E-02
PEAKS036125	CEBPA	0.00E+00	7.09E-10	1.31E-09	2.27E-04
PEAKS036126	CEBPA	0.00E+00	1.08E-08	2.83E-08	3.16E-04
PEAKS036127	CEBPA	0.00E+00	3.57E-09	1.17E-08	1.40E-02
PEAKS036128	CEBPA	0.00E+00	2.71E-09	1.45E-09	2.85E-04
PEAKS036129	CEBPA	0.00E+00	3.82E-11	4.09E-10	7.65E-03
PEAKS036130	CEBPA	0.00E+00	2.19E-09	2.77E-09	1.25E-03

PEAKS036131	CEBPA	0.00E+00	3.61E-09	1.09E-09	2.73E-04
PEAKS036132	CEBPA	0.00E+00	5.09E-09	7.53E-09	4.06E-04
PEAKS036241	CEBPA	0.00E+00	2.02E-09	5.07E-10	3.66E-05
PEAKS036243	CEBPA	0.00E+00	7.60E-12	9.11E-11	2.84E-04
PEAKS036316	CEBPA	0.00E+00	2.63E-09	1.62E-10	8.08E-06
PEAKS036317	CEBPA	0.00E+00	1.21E-11	6.89E-09	1.26E-01
PEAKS037787	CEBPA	0.00E+00	1.66E-11	3.16E-11	3.94E-02
PEAKS037788	CEBPA	0.00E+00	1.04E-10	2.02E-11	6.19E-01
PEAKS038345	CEBPA	0.00E+00	7.60E-10	1.41E-09	2.75E-03
PEAKS039023	CEBPA	0.00E+00	1.25E-09	2.05E-10	1.76E-01
PEAKS039031	CEBPA	0.00E+00	3.00E-11	5.64E-10	6.56E-02
PEAKS039045	CEBPA	0.00E+00	5.46E-10	2.01E-10	1.21E-02
PEAKS039785	CEBPA	0.00E+00	1.14E-09	3.94E-09	7.05E-04
PEAKS039786	CEBPA	0.00E+00	9.30E-10	4.18E-11	4.50E-05
PEAKS040017	CEBPA	3e-323	1.90E-07	4.43E-08	2.35E-03
PEAKS040018	CEBPA	0.00E+00	3.15E-09	1.44E-08	3.19E-04
PEAKS040019	CEBPA	0.00E+00	3.78E-12	1.39E-09	3.75E-03
PEAKS041011	CEBPA	0.00E+00	2.12E-09	6.12E-10	1.79E-04
PEAKS048526	CEBPA	0.00E+00	2.12E-10	4.61E-09	9.03E-02
PEAKS054852	CEBPA	0.00E+00	1.82E-09	3.83E-11	6.67E-05
PEAKS054853	CEBPA	0.00E+00	6.83E-13	3.41E-09	8.90E-07
PEAKS057210	CEBPA	0.00E+00	9.77E-12	3.79E-10	1.19E-03
PEAKS057211	CEBPA	0.00E+00	4.86E-10	6.83E-10	9.76E-04
PEAKS059098	CEBPA	0.00E+00	2.31E-08	5.31E-09	9.27E-04
PEAKS059099	CEBPA	0.00E+00	1.32E-09	5.23E-09	8.66E-04
PEAKS059100	CEBPA	0.00E+00	1.01E-08	1.42E-08	7.09E-05
PEAKS059101	CEBPA	0.00E+00	8.29E-09	1.17E-08	7.84E-04
PEAKS035505	CEBPB	0.00E+00	6.23E-09	4.71E-09	1.19E-03
PEAKS035512	CEBPB	1.60E-296	3.91E-09	3.52E-08	2.93E-02
PEAKS035613	CEBPB	0.00E+00	5.54E-11	5.71E-09	1.45E-02
PEAKS035614	CEBPB	0.00E+00	1.68E-11	2.59E-11	1.07E-03
PEAKS035615	CEBPB	3.05E-299	1.81E-10	2.34E-10	1.41E-01
PEAKS035616	CEBPB	0.00E+00	1.30E-09	4.21E-09	1.35E-02
PEAKS036133	CEBPB	0.00E+00	2.46E-08	6.57E-09	5.07E-04
PEAKS036134	CEBPB	1.56E-269	3.94E-07	4.94E-08	1.42E-03
PEAKS036135	CEBPB	0.00E+00	7.65E-10	3.52E-09	1.12E-04
PEAKS036136	CEBPB	0.00E+00	5.15E-09	2.13E-09	2.82E-03
PEAKS036137	CEBPB	0.00E+00	5.30E-10	6.35E-10	3.73E-03
PEAKS036138	CEBPB	1.53E-276	5.34E-08	8.14E-09	1.23E-03
PEAKS036139	CEBPB	0.00E+00	1.58E-09	6.46E-09	5.39E-04
PEAKS036140	CEBPB	0.00E+00	4.26E-08	4.42E-09	1.40E-02
PEAKS036205	CEBPB	0.00E+00	3.93E-09	1.48E-09	1.29E-03
PEAKS036206	CEBPB	0.00E+00	2.61E-10	5.62E-10	6.65E-01
PEAKS036208	CEBPB	0.00E+00	8.02E-10	1.01E-09	4.12E-04
PEAKS036210	CEBPB	0.00E+00	4.18E-10	4.02E-09	2.14E-03
PEAKS037271	CEBPB	9.99e-313	5.98E-10	1.63E-09	4.42E-04

PEAKS037272	CEBPB	0.00E+00	1.99E-11	1.09E-10	1.40E-03
PEAKS037273	CEBPB	0.00E+00	4.93E-09	5.55E-09	2.70E-05
PEAKS038884	CEBPB	0.00E+00	9.04E-11	7.98E-11	1.30E-04
PEAKS038885	CEBPB	0.00E+00	8.66E-10	9.73E-10	7.09E-04
PEAKS039022	CEBPB	0.00E+00	1.10E-09	9.50E-09	4.38E-02
PEAKS039030	CEBPB	0.00E+00	1.66E-11	1.47E-08	8.87E-01
PEAKS039038	CEBPB	0.00E+00	6.86E-12	2.52E-09	4.57E-04
PEAKS039044	CEBPB	0.00E+00	5.26E-12	6.75E-10	1.11E-02
PEAKS039787	CEBPB	0.00E+00	1.04E-10	3.24E-09	3.70E-04
PEAKS039789	CEBPB	0.00E+00	5.97E-11	2.80E-10	1.68E-05
PEAKS039790	CEBPB	0.00E+00	1.87E-05	7.08E-11	3.34E-04
PEAKS040307	CEBPB	0.00E+00	2.59E-11	2.41E-09	2.61E-02
PEAKS040308	CEBPB	0.00E+00	4.83E-10	8.35E-09	4.56E-02
PEAKS040309	CEBPB	0.00E+00	3.33E-10	2.26E-11	3.84E-03
PEAKS040958	CEBPB	0.00E+00	4.83E-11	2.41E-10	7.90E-04
PEAKS040959	CEBPB	0.00E+00	3.17E-10	4.83E-10	4.36E-03
PEAKS040960	CEBPB	0.00E+00	1.77E-09	5.57E-10	8.10E-04
PEAKS040961	CEBPB	0.00E+00	7.23E-09	3.97E-08	5.49E-01
PEAKS040962	CEBPB	0.00E+00	1.20E-09	1.29E-09	6.37E-01
PEAKS049096	CEBPB	0.00E+00	1.45E-09	5.75E-10	3.94E-03
PEAKS049097	CEBPB	0.00E+00	1.05E-09	7.58E-10	9.33E-05
PEAKS049104	CEBPB	0.00E+00	1.34E-09	4.23E-11	1.24E-03
PEAKS049106	CEBPB	0.00E+00	1.62E-12	7.92E-10	6.79E-04
PEAKS049112	CEBPB	0.00E+00	5.63E-09	5.29E-10	7.59E-02
PEAKS049114	CEBPB	0.00E+00	4.32E-09	2.38E-09	9.11E-04
PEAKS049122	CEBPB	0.00E+00	2.52E-11	9.04E-10	1.10E-02
PEAKS049124	CEBPB	0.00E+00	7.55E-11	1.60E-08	1.14E-02
PEAKS049132	CEBPB	0.00E+00	1.72E-12	4.34E-10	7.84E-03
PEAKS049133	CEBPB	0.00E+00	9.36E-11	1.40E-09	1.02E-02
PEAKS049831	CEBPB	0.00E+00	5.51E-09	3.45E-09	4.46E-03
PEAKS057212	CEBPB	0.00E+00	4.09E-09	4.86E-09	2.09E-04
PEAKS057213	CEBPB	0.00E+00	3.01E-10	3.10E-11	5.83E-05
PEAKS040963	CEBPD	0.00E+00	1.44E-11	7.27E-10	2.26E-04
PEAKS040964	CEBPD	0.00E+00	2.58E-09	5.74E-09	9.64E-04
PEAKS040965	CEBPD	0.00E+00	7.74E-10	2.90E-09	5.91E-01
PEAKS040966	CEBPD	0.00E+00	2.86E-09	1.14E-08	3.04E-01
PEAKS040967	CEBPD	0.00E+00	1.13E-09	2.38E-09	2.18E-01
PEAKS039771	CEBPE	3.47E-188	2.54E-07	1.28E-07	5.27E-02
PEAKS039773	CEBPE	7.87E-201	1.79E-07	1.82E-07	3.72E-01
PEAKS050665	CLOCK	1.97E-300	5.69E-07	1.76E-07	3.58E-02
PEAKS050666	CLOCK	8.85E-282	5.18E-07	3.04E-07	1.10E-03
PEAKS050667	CLOCK	1.13E-234	3.68E-07	2.75E-07	1.58E-04
PEAKS050668	CLOCK	1.47E-257	3.92E-07	2.02E-07	1.62E-03
PEAKS050673	CLOCK	7.66E-247	4.53E-07	2.56E-07	1.38E-01
PEAKS050675	CLOCK	1.90E-244	5.25E-07	8.01E-06	1.81E-03
PEAKS050677	CLOCK	2.48E-171	8.33E-07	4.11E-06	3.69E-04

PEAKS050679	CLOCK	1.57E-200	7.53E-07	2.37E-07	2.01E-03
PEAKS040020	CREB1	0.00E+00	2.33E-06	5.28E-09	1.24E-04
PEAKS040021	CREB1	0.00E+00	4.12E-08	5.98E-09	5.51E-05
PEAKS040022	CREB1	0.00E+00	3.82E-07	4.51E-07	2.63E-04
PEAKS040998	CREB1	3.71e-318	1.96E-10	5.58E-09	5.66E-02
PEAKS040999	CREB1	0.00E+00	1.59E-11	1.04E-09	1.91E-03
PEAKS041000	CREB1	0.00E+00	5.80E-12	1.32E-08	2.49E-03
PEAKS041001	CREB1	0.00E+00	1.46E-11	6.58E-09	4.95E-02
PEAKS041002	CREB1	0.00E+00	4.14E-07	4.80E-07	1.04E-02
PEAKS055998	CREB1	0.00E+00	5.49E-07	3.22E-07	1.62E-05
PEAKS055999	CREB1	0.00E+00	1.43E-06	1.37E-08	4.07E-05
PEAKS056002	CREB1	0.00E+00	2.84E-13	1.09E-09	4.37E-02
PEAKS056003	CREB1	0.00E+00	6.04E-08	1.08E-09	1.53E-04
PEAKS035310	CRX	1.54E-252	8.80E-08	7.19E-09	2.40E-04
PEAKS035311	CRX	4.49E-303	2.13E-08	5.72E-09	5.21E-06
PEAKS035258	CTCF	0.00E+00	5.92E-17	5.72E-17	6.73E-03
PEAKS035465	CTCF	0.00E+00	3.09E-21	6.63E-20	1.67E-02
PEAKS035471	CTCF	0.00E+00	5.17E-22	3.25E-22	2.77E-04
PEAKS035472	CTCF	0.00E+00	1.56E-23	2.41E-23	5.13E-03
PEAKS035477	CTCF	0.00E+00	1.98E-20	1.67E-20	9.00E-03
PEAKS035479	CTCF	0.00E+00	7.42E-20	1.03E-18	1.29E-02
PEAKS035480	CTCF	0.00E+00	1.03E-22	8.83E-23	8.57E-03
PEAKS035487	CTCF	0.00E+00	9.02E-22	1.21E-17	1.36E-02
PEAKS035492	CTCF	0.00E+00	3.22E-17	1.51E-17	3.69E-03
PEAKS035498	CTCF	0.00E+00	1.60E-19	3.55E-20	2.72E-02
PEAKS035504	CTCF	0.00E+00	1.05E-16	1.42E-16	1.42E-02
PEAKS035507	CTCF	0.00E+00	5.53E-20	1.36E-23	2.32E-02
PEAKS035519	CTCF	0.00E+00	3.31E-17	3.14E-17	1.84E-03
PEAKS035523	CTCF	0.00E+00	8.59E-17	2.23E-16	4.89E-02
PEAKS035524	CTCF	0.00E+00	1.01E-17	6.92E-22	5.21E-03
PEAKS035526	CTCF	0.00E+00	1.05E-18	1.73E-18	8.17E-03
PEAKS035567	CTCF	0.00E+00	7.43E-21	6.13E-24	1.17E-02
PEAKS035573	CTCF	0.00E+00	3.70E-19	1.42E-16	1.65E-02
PEAKS035725	CTCF	0.00E+00	3.60E-25	1.10E-24	1.25E-02
PEAKS035726	CTCF	0.00E+00	3.46E-19	1.16E-22	4.52E-03
PEAKS035727	CTCF	0.00E+00	6.86E-24	2.34E-23	4.41E-03
PEAKS035728	CTCF	0.00E+00	1.22E-20	1.37E-24	1.85E-02
PEAKS035729	CTCF	0.00E+00	3.63E-17	5.81E-17	7.30E-03
PEAKS035730	CTCF	0.00E+00	3.96E-18	1.37E-17	2.74E-03
PEAKS035731	CTCF	0.00E+00	7.16E-20	1.28E-20	1.34E-02
PEAKS035732	CTCF	0.00E+00	2.78E-21	1.07E-21	1.02E-02
PEAKS035739	CTCF	0.00E+00	7.42E-23	1.90E-22	9.18E-03
PEAKS035740	CTCF	0.00E+00	9.60E-22	3.18E-22	2.60E-02
PEAKS035741	CTCF	0.00E+00	2.01E-20	3.08E-20	1.92E-02
PEAKS035742	CTCF	0.00E+00	2.54E-23	1.40E-23	2.32E-02
PEAKS035743	CTCF	0.00E+00	5.15E-20	1.02E-19	3.72E-02

PEAKS035848	CTCF	0.00E+00	2.93E-17	5.32E-17	5.19E-03
PEAKS035849	CTCF	0.00E+00	1.59E-22	1.27E-25	1.29E-02
PEAKS035850	CTCF	0.00E+00	3.49E-24	4.92E-24	3.49E-03
PEAKS035851	CTCF	0.00E+00	4.61E-24	2.71E-24	3.94E-02
PEAKS036060	CTCF	0.00E+00	1.29E-21	3.48E-22	1.54E-03
PEAKS036061	CTCF	0.00E+00	3.57E-18	1.21E-17	9.64E-03
PEAKS036106	CTCF	0.00E+00	2.89E-24	6.09E-24	4.34E-02
PEAKS036107	CTCF	0.00E+00	2.13E-24	2.80E-24	2.98E-03
PEAKS036174	CTCF	0.00E+00	4.87E-25	4.50E-26	1.46E-02
PEAKS036175	CTCF	0.00E+00	4.77E-23	1.14E-22	7.16E-02
PEAKS036271	CTCF	0.00E+00	1.86E-24	3.91E-24	1.43E-02
PEAKS037162	CTCF	0.00E+00	1.79E-18	1.21E-22	5.19E-02
PEAKS037163	CTCF	0.00E+00	1.38E-17	1.05E-17	2.57E-02
PEAKS037944	CTCF	0.00E+00	5.56E-25	2.19E-25	4.08E-02
PEAKS037945	CTCF	0.00E+00	9.61E-26	2.50E-25	3.00E-03
PEAKS038306	CTCF	0.00E+00	4.67E-18	8.97E-22	5.27E-03
PEAKS039799	CTCF	0.00E+00	3.53E-20	8.51E-19	7.97E-03
PEAKS039800	CTCF	0.00E+00	3.07E-17	4.58E-17	2.55E-03
PEAKS039801	CTCF	0.00E+00	2.33E-24	9.62E-24	1.13E-03
PEAKS039802	CTCF	0.00E+00	4.83E-19	1.82E-22	1.10E-02
PEAKS039850	CTCF	0.00E+00	4.32E-16	2.77E-16	1.66E-02
PEAKS039852	CTCF	0.00E+00	7.65E-19	7.87E-19	7.49E-03
PEAKS039854	CTCF	0.00E+00	1.10E-15	1.27E-20	2.05E-02
PEAKS039856	CTCF	0.00E+00	2.91E-18	1.48E-22	7.69E-03
PEAKS039915	CTCF	0.00E+00	1.91E-24	5.96E-25	6.59E-02
PEAKS040140	CTCF	8.43E-265	6.19E-18	3.81E-16	1.49E-02
PEAKS040325	CTCF	0.00E+00	3.08E-15	1.82E-19	5.80E-03
PEAKS040331	CTCF	0.00E+00	2.20E-23	1.56E-23	5.08E-03
PEAKS040332	CTCF	0.00E+00	1.50E-24	3.34E-24	2.88E-02
PEAKS040337	CTCF	0.00E+00	4.36E-16	2.76E-13	1.82E-02
PEAKS040338	CTCF	0.00E+00	1.26E-18	5.26E-17	2.89E-02
PEAKS040339	CTCF	0.00E+00	7.21E-17	3.11E-15	9.29E-03
PEAKS040340	CTCF	1.15E-14	2.93E-03	4.34E-02	1.07E-02
PEAKS040341	CTCF	0.00E+00	1.28E-21	2.93E-22	8.38E-03
PEAKS040342	CTCF	0.00E+00	5.65E-23	5.02E-23	1.83E-02
PEAKS040412	CTCF	0.00E+00	4.80E-19	4.28E-19	3.83E-03
PEAKS040413	CTCF	0.00E+00	3.93E-26	6.71E-26	1.70E-02
PEAKS040414	CTCF	0.00E+00	8.09E-22	1.63E-21	8.52E-03
PEAKS040666	CTCF	0.00E+00	1.52E-20	9.07E-22	6.97E-03
PEAKS040667	CTCF	0.00E+00	3.80E-24	1.53E-24	3.72E-03
PEAKS040940	CTCF	0.00E+00	7.04E-19	4.10E-19	1.74E-02
PEAKS040941	CTCF	0.00E+00	7.22E-19	5.72E-19	1.17E-02
PEAKS040942	CTCF	0.00E+00	7.39E-21	5.93E-19	1.39E-02
PEAKS040943	CTCF	0.00E+00	1.13E-16	1.10E-16	3.22E-02
PEAKS040944	CTCF	0.00E+00	7.37E-19	4.08E-19	3.49E-03
PEAKS040945	CTCF	0.00E+00	4.64E-20	1.19E-23	2.18E-02

PEAKS040946	CTCF	0.00E+00	1.92E-20	7.42E-17	2.11E-02
PEAKS040947	CTCF	0.00E+00	4.12E-20	4.79E-25	1.44E-02
PEAKS041060	CTCF	0.00E+00	8.33E-23	4.62E-23	7.91E-03
PEAKS041061	CTCF	0.00E+00	2.45E-21	1.51E-21	8.44E-04
PEAKS041062	CTCF	0.00E+00	1.44E-17	6.08E-22	1.90E-02
PEAKS041063	CTCF	0.00E+00	4.66E-16	6.19E-16	2.06E-03
PEAKS041198	CTCF	0.00E+00	1.25E-20	5.34E-21	2.03E-02
PEAKS041243	CTCF	0.00E+00	2.44E-25	5.73E-21	2.13E-03
PEAKS041379	CTCF	0.00E+00	1.76E-19	6.75E-20	4.12E-02
PEAKS041389	CTCF	0.00E+00	1.57E-21	4.75E-26	1.48E-03
PEAKS041458	CTCF	0.00E+00	1.96E-24	1.44E-23	2.22E-02
PEAKS041574	CTCF	0.00E+00	1.07E-21	1.02E-25	7.52E-02
PEAKS041577	CTCF	0.00E+00	1.31E-23	3.34E-23	1.65E-02
PEAKS041630	CTCF	0.00E+00	1.03E-20	1.51E-25	4.17E-02
PEAKS041724	CTCF	0.00E+00	2.00E-21	7.99E-20	2.43E-02
PEAKS041853	CTCF	0.00E+00	3.16E-24	3.31E-24	3.20E-02
PEAKS041854	CTCF	0.00E+00	1.81E-19	1.80E-19	6.76E-02
PEAKS048309	CTCF	0.00E+00	2.79E-19	3.86E-19	1.98E-02
PEAKS048714	CTCF	0.00E+00	1.93E-21	3.28E-17	4.97E-03
PEAKS048715	CTCF	0.00E+00	1.63E-20	6.30E-20	3.25E-03
PEAKS048768	CTCF	0.00E+00	2.07E-22	1.10E-22	2.08E-02
PEAKS048769	CTCF	0.00E+00	1.21E-20	3.57E-17	1.14E-02
PEAKS048774	CTCF	0.00E+00	1.10E-23	9.94E-24	7.46E-03
PEAKS048775	CTCF	0.00E+00	3.80E-19	2.67E-19	5.95E-03
PEAKS049377	CTCF	0.00E+00	1.07E-23	4.01E-24	3.12E-02
PEAKS049619	CTCF	0.00E+00	2.50E-25	1.27E-24	9.05E-03
PEAKS049620	CTCF	0.00E+00	3.54E-18	1.34E-21	4.68E-02
PEAKS049635	CTCF	0.00E+00	1.66E-21	7.67E-21	1.66E-02
PEAKS049636	CTCF	0.00E+00	2.83E-16	2.61E-16	4.16E-02
PEAKS049637	CTCF	0.00E+00	4.42E-20	6.24E-17	4.77E-02
PEAKS049638	CTCF	0.00E+00	3.32E-25	8.71E-25	3.10E-02
PEAKS049816	CTCF	0.00E+00	8.00E-18	5.30E-20	2.27E-02
PEAKS049817	CTCF	0.00E+00	7.15E-22	1.48E-19	6.30E-03
PEAKS049828	CTCF	0.00E+00	3.15E-19	5.85E-24	4.94E-02
PEAKS049829	CTCF	0.00E+00	4.94E-24	5.53E-24	2.14E-02
PEAKS052297	CTCF	0.00E+00	3.04E-17	5.08E-17	3.75E-04
PEAKS052604	CTCF	0.00E+00	1.19E-17	2.19E-21	2.25E-02
PEAKS054477	CTCF	0.00E+00	3.28E-14	4.48E-14	2.11E-02
PEAKS054478	CTCF	0.00E+00	4.67E-21	2.06E-21	1.31E-02
PEAKS054796	CTCF	0.00E+00	2.68E-23	1.29E-23	1.30E-02
PEAKS054797	CTCF	0.00E+00	1.09E-20	2.85E-21	1.55E-02
PEAKS055636	CTCF	0.00E+00	1.54E-24	1.15E-24	1.90E-03
PEAKS055637	CTCF	0.00E+00	2.13E-21	3.28E-22	1.95E-02
PEAKS057534	CTCF	0.00E+00	1.94E-19	2.48E-24	7.27E-02
PEAKS057535	CTCF	0.00E+00	4.05E-20	1.40E-19	1.92E-02
PEAKS057536	CTCF	0.00E+00	9.63E-20	6.82E-14	4.24E-02

PEAKS057537	CTCF	0.00E+00	6.16E-24	3.68E-23	1.44E-02
PEAKS057538	CTCF	0.00E+00	1.99E-20	7.28E-20	1.48E-02
PEAKS057539	CTCF	0.00E+00	7.84E-19	3.67E-19	2.33E-02
PEAKS058095	CTCF	0.00E+00	1.93E-20	7.99E-19	5.24E-02
PEAKS058096	CTCF	0.00E+00	2.92E-15	2.85E-19	7.77E-03
PEAKS059302	CTCF	0.00E+00	1.73E-23	4.66E-24	7.26E-04
PEAKS038307	CTCFL	0.00E+00	7.63E-12	6.47E-12	4.72E-03
PEAKS040452	DLX3	3.43E-79	1.00E+00	1.00E+00	1.00E+00
PEAKS037776	DMRT1	0.00E+00	4.87E-08	2.05E-08	6.35E-05
PEAKS037344	DMRTB1	1.00E+00	1.10E-08	1.94E-08	2.06E-03
PEAKS035852	E2F1	1.87E-23	5.83E-06	2.25E-01	1.81E-01
PEAKS035853	E2F1	1.22E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS035854	E2F1	3.82E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS035855	E2F1	1.73E-09	1.00E+00	1.00E+00	1.00E+00
PEAKS038545	E2F1	4.42E-201	3.35E-01	4.96E-01	4.82E-04
PEAKS035488	E2F4	3.99E-121	1.06E-01	6.54E-02	1.10E-02
PEAKS035856	E2F4	1.53E-97	2.90E-05	2.77E-05	2.39E-02
PEAKS035857	E2F4	2.41E-76	1.06E-01	7.20E-02	2.74E-02
PEAKS035858	E2F4	9.13E-75	2.03E-06	4.14E-05	1.42E-02
PEAKS035859	E2F4	6.91E-85	7.51E-06	3.17E-05	1.00E-03
PEAKS038834	E2F4	1.08E-160	8.21E-02	1.75E-01	1.66E-03
PEAKS035549	EBF1	5e-324	3.61E-07	2.89E-07	1.51E-05
PEAKS035836	EBF1	0.00E+00	2.24E-07	2.63E-07	1.05E-07
PEAKS036423	EBF1	0.00E+00	2.43E-07	1.08E-07	1.15E-02
PEAKS038138	EBF1	0.00E+00	1.76E-07	1.70E-07	9.73E-01
PEAKS038139	EBF1	0.00E+00	4.32E-07	2.42E-07	2.38E-05
PEAKS038140	EBF1	0.00E+00	2.53E-07	1.22E-07	1.41E-03
PEAKS038141	EBF1	0.00E+00	1.74E-06	1.61E-07	1.40E-05
PEAKS038976	EBF1	0.00E+00	1.34E-07	2.66E-08	1.28E-05
PEAKS039240	EBF1	0.00E+00	8.48E-08	6.68E-08	4.36E-06
PEAKS039241	EBF1	0.00E+00	1.98E-07	5.45E-08	1.25E-03
PEAKS040358	EBF1	0.00E+00	1.67E-07	2.85E-08	6.49E-04
PEAKS040359	EBF1	0.00E+00	9.15E-07	4.93E-08	1.40E-03
PEAKS048598	EBF1	0.00E+00	1.46E-07	7.06E-08	2.33E-03
PEAKS048599	EBF1	0.00E+00	2.18E-08	4.87E-08	2.22E-06
PEAKS048600	EBF1	0.00E+00	1.68E-07	1.26E-08	4.93E-07
PEAKS049707	EBF1	1.01E-271	3.42E-07	1.03E-07	6.11E-04
PEAKS049708	EBF1	0.00E+00	6.45E-07	1.18E-07	2.31E-05
PEAKS049709	EBF1	0.00E+00	7.10E-07	1.28E-07	4.71E-05
PEAKS049711	EBF1	0.00E+00	4.37E-07	1.24E-07	3.03E-04
PEAKS039791	EBF2	2.57e-312	1.11E-07	1.42E-07	2.00E-05
PEAKS039792	EBF2	4.39E-262	4.68E-07	1.95E-07	1.69E-03
PEAKS039793	EBF2	2.83E-287	1.83E-07	3.54E-08	2.86E-03
PEAKS039794	EBF2	0.00E+00	2.30E-07	9.04E-08	5.56E-03
PEAKS035917	EGR1	1.24E-235	1.20E-08	1.30E-08	1.40E-01
PEAKS035919	EGR1	3.09E-284	9.87E-10	4.33E-09	5.33E-04

PEAKS036141	EGR1	4.44E-147	3.95E-01	3.24E-01	1.06E-01
PEAKS057617	EGR1	3.35E-273	8.67E-09	4.15E-08	1.47E-02
PEAKS035876	EGR2	1.14E-259	4.33E-06	2.04E-05	4.14E-04
PEAKS035877	EGR2	1.67E-169	1.00E-08	8.00E-05	3.92E-03
PEAKS035878	EGR2	1.65E-235	3.20E-06	1.80E-05	2.90E-04
PEAKS035879	EGR2	1.97E-176	2.75E-05	2.94E-08	1.83E-03
PEAKS036233	EGR2	1.46E-265	3.54E-09	2.10E-05	8.45E-03
PEAKS036234	EGR2	2.06E-216	1.23E-03	4.09E-04	7.40E-02
PEAKS036078	ELF1	0.00E+00	2.49E-10	2.28E-10	7.45E-02
PEAKS036079	ELF1	0.00E+00	1.55E-10	3.91E-11	4.89E-03
PEAKS040401	ELF1	0.00E+00	8.80E-12	8.96E-12	1.61E-03
PEAKS040402	ELF4	0.00E+00	4.06E-11	3.13E-11	8.94E-04
PEAKS040403	ELF4	0.00E+00	1.18E-10	2.31E-13	2.67E-03
PEAKS040404	ELF4	0.00E+00	2.69E-11	5.66E-12	9.67E-03
PEAKS050743	EOMES	4.89E-108	6.99E-01	3.32E-01	1.80E-03
PEAKS053270	ERF	8.62E-108	5.45E-01	4.72E-10	2.00E-02
PEAKS053271	ERF	1.14E-115	1.03E-07	1.01E-06	3.61E-02
PEAKS053272	ERF	9.16E-102	5.23E-01	2.87E-01	1.15E-02
PEAKS053273	ERF	1.10E-86	7.05E-01	5.61E-01	8.79E-03
PEAKS035587	ERG	0.00E+00	3.91E-12	7.57E-12	6.01E-04
PEAKS035589	ERG	0.00E+00	2.70E-10	2.66E-15	8.50E-04
PEAKS035661	ERG	0.00E+00	1.73E-11	1.44E-11	5.67E-05
PEAKS035662	ERG	0.00E+00	7.09E-12	5.28E-12	5.62E-05
PEAKS035603	ESR1	2.26E-61	1.00E+00	1.00E+00	1.00E+00
PEAKS035604	ESR1	0.00E+00	2.16E-13	1.43E-13	9.89E-04
PEAKS035948	ESR1	0.00E+00	7.50E-13	7.17E-13	1.18E-03
PEAKS035949	ESR1	0.00E+00	1.13E-13	2.41E-14	4.13E-04
PEAKS036166	ESR1	0.00E+00	1.35E-12	2.58E-13	7.04E-04
PEAKS036219	ESR1	0.00E+00	5.40E-14	9.18E-16	5.16E-05
PEAKS036220	ESR1	0.00E+00	1.35E-14	3.05E-16	9.92E-05
PEAKS036221	ESR1	2.42E-62	1.00E+00	1.00E+00	1.00E+00
PEAKS036222	ESR1	5.13E-126	1.78E-08	6.60E-09	5.01E-02
PEAKS036566	ESR1	4.90E-142	6.23E-05	9.58E-06	3.86E-06
PEAKS036567	ESR1	5.64E-128	7.26E-04	1.99E-05	1.97E-05
PEAKS037554	ESR1	0.00E+00	2.19E-14	1.23E-14	2.91E-03
PEAKS038233	ESR1	1.24E-285	4.69E-15	1.44E-15	6.10E-03
PEAKS038872	ESR1	7.12E-11	1.00E+00	1.00E+00	1.00E+00
PEAKS040741	ESR1	8.12E-08	1.00E+00	1.00E+00	1.00E+00
PEAKS057223	ESR1	0.00E+00	4.67E-10	1.23E-09	1.35E-04
PEAKS057444	ESR1	0.00E+00	3.36E-13	9.31E-13	7.76E-08
PEAKS057031	ESR2	1.10E-118	3.46E-13	5.61E-13	5.44E-01
PEAKS036172	ESRRA	0.00E+00	2.98E-10	8.29E-08	1.40E-05
PEAKS040010	ESRRA	1.85e-321	3.72E-10	1.61E-09	4.91E-03
PEAKS040112	ESRRA	0.00E+00	2.69E-11	1.91E-12	7.51E-06
PEAKS040113	ESRRA	0.00E+00	2.42E-12	2.19E-07	1.57E-04
PEAKS048563	ESRRG	0.00E+00	7.04E-10	1.05E-09	4.03E-05

PEAKS048564	ESRRG	0.00E+00	6.47E-12	6.04E-12	1.37E-04
PEAKS036080	ETS1	1.00E-75	5.32E-06	1.59E-04	7.88E-02
PEAKS036081	ETS1	4.83E-121	4.20E-08	1.40E-06	2.55E-01
PEAKS036648	ETS1	1.78E-274	2.57E-09	1.94E-09	1.86E-06
PEAKS039242	ETS1	4.40E-292	2.11E-06	6.34E-05	6.11E-02
PEAKS039243	ETS1	2.18E-261	4.44E-05	1.35E-05	8.83E-03
PEAKS040097	ETS1	1.39E-279	7.42E-11	4.01E-10	9.90E-04
PEAKS035862	ETS2	8.19E-72	1.00E+00	1.00E+00	1.00E+00
PEAKS035863	ETS2	9.32E-81	1.00E+00	1.00E+00	1.00E+00
PEAKS041064	ETV6	0.00E+00	1.75E-08	4.40E-10	4.73E-04
PEAKS041065	ETV6	0.00E+00	3.78E-10	4.35E-10	2.82E-04
PEAKS038557	FEV	1.99E-169	5.18E-01	3.68E-01	5.38E-01
PEAKS035515	FLI1	0.00E+00	3.34E-13	1.75E-13	7.79E-07
PEAKS036886	FLI1	0.00E+00	7.53E-09	5.31E-09	1.63E-03
PEAKS040400	FLI1	0.00E+00	3.69E-11	5.23E-11	8.76E-07
PEAKS038481	FOS	0.00E+00	5.96E-07	1.02E-09	1.12E-02
PEAKS038482	FOS	0.00E+00	8.37E-07	1.07E-06	4.85E-07
PEAKS038483	FOS	8.48E-269	4.77E-07	3.58E-07	5.02E-01
PEAKS038484	FOS	0.00E+00	7.15E-07	9.90E-09	4.32E-03
PEAKS039300	FOS	5.61E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS039302	FOS	3.98E-203	9.54E-07	1.90E-10	6.04E-02
PEAKS039303	FOS	0.00E+00	2.36E-10	7.15E-07	1.84E-04
PEAKS039305	FOS	2.47E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS039309	FOS	1.24E-149	3.79E-01	6.37E-01	2.39E-01
PEAKS039310	FOS	0.00E+00	7.14E-11	3.70E-11	3.05E-06
PEAKS039313	FOS	1.67E-178	5.58E-01	2.38E-07	1.11E-05
PEAKS039315	FOS	3.66E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS039317	FOS	1.35E-292	9.80E-07	2.35E-07	2.34E-07
PEAKS039319	FOS	8.05E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS040968	FOS	0.00E+00	2.11E-09	1.88E-10	9.86E-05
PEAKS040969	FOS	0.00E+00	1.40E-06	1.41E-10	2.32E-04
PEAKS040970	FOS	0.00E+00	7.08E-10	5.03E-07	3.69E-07
PEAKS040971	FOS	0.00E+00	1.39E-06	2.63E-10	1.73E-07
PEAKS040972	FOS	0.00E+00	2.86E-06	1.43E-06	3.78E-07
PEAKS049568	FOS	4.16E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS049570	FOS	0.00E+00	3.98E-07	1.59E-09	2.73E-03
PEAKS049571	FOS	0.00E+00	6.33E-07	1.63E-11	6.87E-08
PEAKS049582	FOS	0.00E+00	1.50E-10	1.83E-10	2.16E-04
PEAKS049583	FOS	0.00E+00	3.55E-10	1.31E-06	2.48E-06
PEAKS054856	FOS	0.00E+00	3.58E-07	1.19E-06	1.82E-04
PEAKS054857	FOS	9.09E-130	3.97E-01	3.46E-01	8.80E-01
PEAKS054870	FOS	0.00E+00	2.06E-09	1.67E-06	1.73E-06
PEAKS054871	FOS	2.40E-242	2.94E-11	3.75E-08	3.64E-04
PEAKS039301	FOSB	1.88E-11	1.00E+00	1.00E+00	1.00E+00
PEAKS039304	FOSB	5.93E-273	7.15E-07	5.16E-11	8.74E-02
PEAKS039311	FOSB	2.71E-274	1.22E-10	4.77E-07	9.18E-04

PEAKS039314	FOSB	0.00E+00	4.07E-10	2.20E-10	4.45E-03
PEAKS039316	FOSB	9.78E-113	4.77E-07	7.15E-07	1.91E-01
PEAKS039318	FOSB	0.00E+00	1.85E-08	4.77E-07	1.18E-04
PEAKS039320	FOSB	6.51E-115	2.27E-08	1.19E-06	1.68E-01
PEAKS041066	FOSL2	0.00E+00	4.14E-09	6.64E-10	6.64E-04
PEAKS041067	FOSL2	0.00E+00	1.17E-09	7.19E-07	8.01E-08
PEAKS041068	FOSL2	0.00E+00	3.08E-06	2.35E-07	8.54E-05
PEAKS049574	FOSL2	0.00E+00	1.63E-11	9.39E-07	9.23E-11
PEAKS049575	FOSL2	0.00E+00	5.02E-11	3.04E-11	1.02E-10
PEAKS049576	FOSL2	0.00E+00	7.23E-12	9.54E-07	2.40E-08
PEAKS049577	FOSL2	0.00E+00	6.08E-10	9.09E-07	3.91E-11
PEAKS054858	FOSL2	0.00E+00	5.96E-07	9.54E-07	7.68E-07
PEAKS054859	FOSL2	0.00E+00	5.20E-09	9.54E-07	1.93E-01
PEAKS035679	FOXA1	4.71E-241	1.68E-09	5.47E-10	1.23E-04
PEAKS036274	FOXA1	3.53E-241	1.22E-08	7.72E-11	3.29E-03
PEAKS036286	FOXA1	2.15E-232	3.95E-09	6.27E-10	1.89E-04
PEAKS048589	FOXA1	4.14e-321	1.24E-09	2.78E-10	1.80E-04
PEAKS048590	FOXA1	5e-324	7.75E-09	3.37E-10	5.31E-04
PEAKS048591	FOXA1	5e-324	1.77E-10	2.65E-10	4.91E-03
PEAKS048592	FOXA1	0.00E+00	1.53E-08	4.80E-11	3.59E-03
PEAKS048593	FOXA1	0.00E+00	2.03E-10	7.99E-11	1.84E-02
PEAKS048594	FOXA1	0.00E+00	7.43E-10	4.24E-10	4.91E-05
PEAKS054843	FOXA1	2.07E-250	1.24E-09	2.64E-10	1.41E-03
PEAKS057422	FOXA1	1.77E-282	1.03E-08	1.43E-10	1.21E-04
PEAKS057424	FOXA1	7.86E-142	7.83E-10	6.32E-10	3.54E-04
PEAKS057426	FOXA1	1.40E-248	4.80E-08	3.62E-10	2.11E-02
PEAKS035249	FOXA2	7.62e-310	1.17E-09	1.53E-10	1.64E-03
PEAKS035599	FOXA2	7.81E-175	2.98E-08	1.55E-09	8.94E-05
PEAKS035678	FOXA2	9.20E-289	3.06E-08	1.04E-08	6.57E-04
PEAKS037143	FOXA2	5.51E-228	2.25E-09	8.99E-11	1.10E-02
PEAKS038234	FOXA2	3.05E-235	2.37E-08	1.88E-11	4.62E-03
PEAKS039577	FOXA2	5.00E-192	4.32E-08	6.58E-09	2.68E-01
PEAKS038882	FOXA3	1.22E-202	8.94E-10	3.96E-10	2.03E-03
PEAKS038883	FOXA3	1.38e-314	4.17E-10	1.20E-10	1.32E-03
PEAKS038924	FOXF1	1.22e-315	4.41E-07	4.37E-06	1.60E-03
PEAKS049810	FOXG1	1.06E-77	5.67E-06	7.04E-06	1.25E-02
PEAKS049811	FOXG1	2.61E-71	1.95E-01	6.24E-08	8.43E-02
PEAKS055621	FOXH1	5.06E-129	1.00E+00	1.00E+00	1.00E+00
PEAKS037388	FOXL2	3.99E-205	5.67E-10	3.54E-10	7.75E-03
PEAKS037553	FOXL2	2.46E-224	2.26E-09	1.43E-09	2.37E-04
PEAKS038632	FOXN1	5.83E-47	1.00E+00	1.00E+00	1.00E+00
PEAKS036075	FOXO1	1.96E-169	2.40E-11	6.68E-11	1.62E-02
PEAKS036257	FOXO1	1.34E-220	4.49E-08	2.62E-08	1.49E-02
PEAKS036258	FOXO1	2.89E-103	1.37E-06	2.15E-06	6.53E-03
PEAKS036930	FOXO1	3.56E-258	4.69E-11	2.49E-10	5.09E-03
PEAKS037345	FOXO1	2.67E-248	5.17E-10	1.03E-07	1.77E-04

PEAKS050687	FOXO1	4.86E-186	3.01E-06	4.44E-06	7.36E-04
PEAKS050688	FOXO1	6.91E-113	1.23E-01	7.89E-05	3.23E-02
PEAKS050758	FOXO1	1.22E-125	3.14E-01	1.60E-01	6.99E-02
PEAKS055001	FOXP1	2.75E-48	2.05E-01	7.36E-02	2.85E-01
PEAKS036084	FOXP3	1.17E-43	5.61E-02	4.98E-02	1.65E-03
PEAKS039656	FOXP3	1.54E-24	1.00E+00	1.00E+00	1.00E+00
PEAKS039658	FOXP3	2.29E-23	7.61E-02	6.35E-02	6.04E-02
PEAKS039659	FOXP3	1.80E-14	2.77E-02	7.57E-02	3.16E-01
PEAKS039660	FOXP3	2.61E-13	4.19E-03	9.44E-02	1.44E-03
PEAKS040450	FOXP3	4.80E-52	6.38E-02	6.44E-02	2.83E-01
PEAKS040451	FOXP3	1.58E-77	6.33E-02	9.09E-02	1.13E-02
PEAKS055683	FOXP3	1.10E-28	6.03E-02	4.60E-02	1.11E-01
PEAKS055684	FOXP3	7.55E-20	1.00E+00	1.00E+00	1.00E+00
PEAKS057187	FOXP3	1.14E-46	9.48E-02	7.21E-02	6.09E-02
PEAKS057188	FOXP3	4.96E-42	1.01E-01	1.96E-02	6.69E-02
PEAKS057192	FOXP3	1.52E-42	4.96E-02	7.75E-02	1.42E-01
PEAKS057193	FOXP3	1.51E-53	6.25E-02	6.89E-02	2.99E-01
PEAKS059306	FOXP3	8.26E-20	7.76E-02	8.69E-02	1.05E-02
PEAKS039651	GABPA	0.00E+00	4.10E-13	2.00E-13	2.50E-07
PEAKS040405	GABPA	0.00E+00	1.83E-10	2.56E-11	2.75E-04
PEAKS040406	GABPA	0.00E+00	3.36E-13	3.81E-15	4.07E-05
PEAKS040407	GABPA	0.00E+00	2.58E-09	1.14E-15	3.86E-04
PEAKS035331	GATA1	0.00E+00	1.26E-09	4.69E-09	1.06E-03
PEAKS035481	GATA1	0.00E+00	1.48E-09	5.67E-08	1.12E-03
PEAKS035518	GATA1	1.67E-190	4.52E-10	5.74E-04	3.79E-03
PEAKS036144	GATA1	2.81E-308	3.40E-08	1.72E-08	2.25E-03
PEAKS035260	GATA2	4.75E-201	1.62E-09	5.25E-10	8.78E-03
PEAKS048667	GATA3	9.08E-136	6.27E-09	8.09E-04	1.53E-03
PEAKS036330	GATA4	1.46E-251	7.40E-11	2.03E-09	5.07E-05
PEAKS039539	GATA4	4.33E-129	1.80E-08	8.01E-08	4.49E-02
PEAKS039540	GATA4	3.54E-142	4.41E-09	1.17E-07	4.12E-03
PEAKS039541	GATA4	4.81E-292	1.01E-06	5.66E-07	1.73E-01
PEAKS039542	GATA4	0.00E+00	2.68E-07	1.76E-07	1.30E-01
PEAKS041833	GATA4	1.45E-292	2.38E-07	4.86E-11	1.12E-04
PEAKS057460	GATA4	6.63E-222	1.26E-06	2.78E-04	1.99E-02
PEAKS057461	GATA4	3.08E-189	4.02E-07	6.07E-03	1.74E-04
PEAKS057462	GATA4	4.70E-214	1.24E-06	1.40E-02	7.50E-04
PEAKS057463	GATA4	2.60E-229	4.02E-11	2.57E-10	1.13E-02
PEAKS057464	GATA4	1.21E-207	1.42E-08	1.58E-08	1.57E-04
PEAKS057465	GATA4	1.45E-222	4.77E-06	2.83E-02	8.30E-03
PEAKS057474	GATA4	2.34E-246	6.38E-01	7.43E-01	4.24E-02
PEAKS037140	GATA6	2.36E-303	2.34E-09	2.26E-07	1.38E-02
PEAKS039700	GATA6	0.00E+00	7.53E-12	6.19E-10	1.71E-06
PEAKS039701	GATA6	0.00E+00	1.63E-11	1.55E-10	1.22E-03
PEAKS036387	GFI1	1.14E-142	4.42E-13	1.55E-12	6.87E-03
PEAKS038227	GFI1	5.06E-157	9.25E-08	1.61E-08	7.54E-03

PEAKS037839	GRHL2	1.86E-115	5.88E-07	2.40E-06	1.58E-04
PEAKS049676	GRHL2	8.86E-229	1.38E-06	2.36E-06	1.68E-03
PEAKS049680	GRHL2	7.17E-178	5.80E-08	3.08E-07	2.67E-02
PEAKS052556	GTF2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS052551	HAND2	4.17E-29	3.64E-03	2.06E-03	1.19E-02
PEAKS041069	HIF1A	2.58E-214	1.20E-05	6.89E-06	2.34E-02
PEAKS041070	HIF1A	2.76E-225	9.47E-06	4.51E-05	1.70E-03
PEAKS050761	HNF1A	1.19E-12	1.00E+00	1.00E+00	1.00E+00
PEAKS048565	HNF1B	0.00E+00	9.31E-08	5.77E-12	6.58E-02
PEAKS035790	HNF4A	1.47E-256	4.72E-09	1.23E-07	5.80E-04
PEAKS035791	HNF4A	8.54E-288	9.91E-11	2.89E-11	4.09E-05
PEAKS036278	HNF4A	6.21E-258	1.01E-06	1.90E-06	3.47E-06
PEAKS037160	HNF4A	2.78E-288	5.15E-09	1.01E-08	1.04E-02
PEAKS037161	HNF4A	0.00E+00	3.61E-11	8.98E-11	4.97E-03
PEAKS038886	HNF4A	4.07e-318	1.41E-10	5.27E-10	5.70E-05
PEAKS038887	HNF4A	0.00E+00	1.11E-11	3.10E-11	2.54E-03
PEAKS040474	HNF4A	5.79E-262	2.80E-09	1.35E-09	2.65E-04
PEAKS040475	HNF4A	0.00E+00	1.93E-09	4.09E-09	2.41E-04
PEAKS040511	HNF4A	3.98E-218	8.49E-07	2.50E-10	2.31E-03
PEAKS049330	HNF4A	4.57E-278	3.13E-11	5.84E-11	1.50E-04
PEAKS049331	HNF4A	4.59E-277	5.06E-09	1.69E-10	1.10E-02
PEAKS050010	HNF4A	0.00E+00	9.54E-09	3.66E-11	1.14E-04
PEAKS050011	HNF4A	0.00E+00	9.23E-09	6.02E-11	3.29E-04
PEAKS050012	HNF4A	0.00E+00	7.78E-09	4.87E-11	1.60E-04
PEAKS050013	HNF4A	0.00E+00	1.08E-08	5.21E-11	1.35E-04
PEAKS059094	HNF4A	2.08E-264	2.90E-06	6.25E-08	7.20E-04
PEAKS059095	HNF4A	1.03E-264	1.40E-10	9.00E-11	6.87E-02
PEAKS059096	HNF4A	5.74E-241	6.92E-08	1.04E-07	7.57E-02
PEAKS059097	HNF4A	1.81E-247	2.55E-10	4.36E-09	2.37E-03
PEAKS040476	HNF4G	2.95E-248	3.41E-08	7.77E-07	1.08E-03
PEAKS035431	HOXA9	4.10E-69	7.56E-04	6.89E-06	4.75E-01
PEAKS042586	HOXA9	2.61E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS049616	HOXD11	7.49E-113	1.07E-05	2.79E-06	4.38E-03
PEAKS035768	IKZF1	1.45E-35	6.82E-05	1.18E-04	3.87E-03
PEAKS035778	IKZF1	1.61E-96	1.46E-03	2.37E-05	2.34E-02
PEAKS035779	IKZF1	7.84E-85	8.39E-09	1.68E-08	2.43E-04
PEAKS035975	IKZF1	4.43E-25	1.40E-02	1.32E-02	3.14E-04
PEAKS035976	IKZF1	1.50E-18	3.40E-04	1.23E-03	4.00E-04
PEAKS036142	IKZF1	1.14E-67	2.44E-04	2.19E-02	4.22E-04
PEAKS036426	IKZF1	1.30E-66	3.04E-04	3.20E-07	1.70E-02
PEAKS037407	IKZF1	7.50E-64	1.28E-05	4.83E-06	3.96E-02
PEAKS037408	IKZF1	7.26E-79	7.76E-05	1.00E-04	1.24E-02
PEAKS037409	IKZF1	3.14E-43	1.00E+00	1.00E+00	1.00E+00
PEAKS037410	IKZF1	2.36E-15	1.69E-04	2.26E-06	1.48E-01
PEAKS037888	IKZF1	5.71E-65	8.31E-07	9.92E-07	1.66E-04
PEAKS039244	IKZF1	5.65E-89	1.99E-04	2.76E-05	2.95E-04

PEAKS038501	IKZF2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038502	IKZF2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035767	IKZF3	3.09E-92	1.60E-10	2.12E-11	1.62E-05
PEAKS039237	IKZF3	1.83E-89	1.67E-04	2.94E-06	1.00E-03
PEAKS039575	INSM1	4.53E-12	1.00E+00	1.00E+00	1.00E+00
PEAKS035868	IRF1	2.61E-25	5.37E-02	3.40E-02	4.95E-04
PEAKS035869	IRF1	0.00E+00	7.37E-07	4.41E-07	2.86E-05
PEAKS035870	IRF1	2.64E-37	1.00E+00	1.00E+00	1.00E+00
PEAKS035871	IRF1	1.05E-141	4.28E-05	6.10E-07	1.04E-03
PEAKS037845	IRF1	0.00E+00	3.57E-07	3.67E-07	1.26E-04
PEAKS038918	IRF1	0.00E+00	2.95E-09	2.39E-09	4.92E-05
PEAKS038919	IRF1	2.61E-308	2.88E-09	6.31E-09	4.10E-04
PEAKS038920	IRF1	0.00E+00	8.39E-08	1.37E-08	6.07E-06
PEAKS040302	IRF1	0.00E+00	1.49E-06	2.51E-07	5.79E-05
PEAKS040303	IRF1	0.00E+00	1.20E-06	1.67E-07	4.53E-05
PEAKS040779	IRF1	0.00E+00	1.32E-10	7.24E-11	3.10E-02
PEAKS042179	IRF1	0.00E+00	1.79E-07	2.49E-07	3.06E-07
PEAKS042180	IRF1	0.00E+00	2.85E-07	1.57E-06	4.07E-04
PEAKS042181	IRF1	0.00E+00	1.19E-06	2.77E-06	7.69E-05
PEAKS042182	IRF1	0.00E+00	5.23E-07	3.18E-07	2.21E-05
PEAKS055531	IRF2	6.82E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS037919	IRF3	1.21E-252	2.10E-06	8.11E-07	3.40E-03
PEAKS037920	IRF3	0.00E+00	1.81E-05	1.06E-07	1.62E-07
PEAKS037924	IRF3	4.38E-248	2.53E-06	1.03E-06	1.59E-03
PEAKS048604	IRF3	1.02E-71	1.43E-04	2.49E-04	1.53E-02
PEAKS048605	IRF3	5.49E-145	9.66E-04	2.40E-04	8.70E-04
PEAKS048606	IRF3	1.25E-92	1.15E-02	8.65E-04	9.97E-04
PEAKS035634	IRF4	1.01E-19	6.93E-01	7.44E-02	2.27E-01
PEAKS035637	IRF4	5.64E-37	1.17E-01	9.46E-02	1.40E-01
PEAKS036029	IRF4	1.17E-33	4.43E-01	2.11E-01	1.08E-02
PEAKS036030	IRF4	4.12E-30	1.84E-01	4.90E-02	3.66E-02
PEAKS036043	IRF4	1.11E-29	1.00E+00	1.00E+00	1.00E+00
PEAKS036044	IRF4	5.09E-25	1.00E+00	1.00E+00	1.00E+00
PEAKS036046	IRF4	6.63E-18	1.00E+00	1.00E+00	1.00E+00
PEAKS036335	IRF4	5.89E-15	3.60E-01	2.52E-01	2.04E-01
PEAKS036425	IRF4	4.25E-183	1.03E-06	1.17E-06	6.22E-04
PEAKS036583	IRF4	4.11E-29	2.07E-01	7.06E-02	4.89E-02
PEAKS039245	IRF4	7.35E-175	3.08E-06	3.85E-06	3.01E-03
PEAKS039246	IRF4	2.82E-173	2.35E-07	2.07E-07	3.57E-03
PEAKS040194	IRF4	2.40E-22	1.07E-01	4.05E-02	1.73E-01
PEAKS040195	IRF4	8.21E-94	2.77E-12	1.95E-11	1.87E-04
PEAKS040196	IRF4	7.52E-83	3.59E-08	5.35E-08	5.24E-02
PEAKS040197	IRF4	1.44E-35	3.92E-01	1.58E-01	3.24E-01
PEAKS040198	IRF4	5.43E-30	1.04E-01	4.13E-02	4.75E-02
PEAKS040199	IRF4	2.97E-51	2.92E-02	1.67E-01	4.51E-02
PEAKS040200	IRF4	1.06E-28	2.01E-01	1.88E-01	1.19E-01

PEAKS040267	IRF4	9.39E-19	3.72E-01	3.65E-01	2.94E-02
PEAKS040268	IRF4	7.13E-15	5.48E-01	2.36E-01	5.13E-01
PEAKS040781	IRF4	2.34E-14	5.68E-01	2.78E-01	2.11E-01
PEAKS040782	IRF4	4.95E-13	1.00E+00	1.00E+00	1.00E+00
PEAKS041006	IRF4	1.14E-46	1.00E+00	1.00E+00	1.00E+00
PEAKS041071	IRF4	2.11E-29	4.07E-02	4.03E-02	3.95E-01
PEAKS041072	IRF4	1.34E-97	1.56E-11	2.29E-10	2.25E-02
PEAKS041073	IRF4	8.44E-31	2.43E-01	6.60E-02	3.98E-03
PEAKS041074	IRF4	1.04E-24	2.60E-02	6.94E-02	1.14E-01
PEAKS041075	IRF4	1.41E-29	6.75E-02	6.08E-02	1.41E-01
PEAKS041076	IRF4	5.35E-88	1.00E+00	1.00E+00	1.00E+00
PEAKS041077	IRF4	1.24E-33	5.57E-01	1.94E-01	3.28E-02
PEAKS041078	IRF4	4.79E-45	5.22E-01	5.40E-01	6.36E-01
PEAKS041079	IRF4	9.86E-20	1.00E+00	1.00E+00	1.00E+00
PEAKS041080	IRF4	1.08E-76	1.00E+00	1.00E+00	1.00E+00
PEAKS041081	IRF4	6.25E-24	5.71E-01	1.04E-01	2.48E-01
PEAKS041083	IRF4	4.44E-22	6.91E-01	1.67E-01	2.71E-02
PEAKS041084	IRF4	2.28E-26	6.13E-01	1.41E-01	2.72E-01
PEAKS041085	IRF4	1.87E-32	5.69E-01	7.05E-02	3.33E-01
PEAKS041086	IRF4	3.62E-33	3.15E-01	7.45E-02	2.98E-01
PEAKS041087	IRF4	3.25E-23	3.17E-01	1.08E-01	1.98E-01
PEAKS041088	IRF4	1.06E-75	1.00E+00	1.00E+00	1.00E+00
PEAKS046788	IRF4	7.82E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS046791	IRF4	1.81E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS048607	IRF5	2.52E-86	1.33E-01	4.49E-06	3.30E-04
PEAKS048608	IRF5	9.11E-104	2.37E-07	2.56E-05	4.61E-03
PEAKS048609	IRF5	1.05E-60	1.00E+00	1.00E+00	1.00E+00
PEAKS037877	IRF8	2.61E-173	4.08E-07	5.62E-07	5.42E-03
PEAKS037879	IRF8	3.31E-230	9.13E-08	1.85E-07	1.86E-03
PEAKS037941	IRF8	6.16E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS038228	IRF8	1.13E-122	4.99E-07	2.72E-06	8.68E-03
PEAKS038915	IRF8	0.00E+00	1.97E-07	1.76E-07	1.06E-03
PEAKS038916	IRF8	0.00E+00	1.80E-07	1.59E-07	8.61E-04
PEAKS038917	IRF8	0.00E+00	6.66E-07	1.48E-07	1.61E-05
PEAKS048610	IRF8	1.61E-198	1.24E-07	2.72E-07	1.98E-03
PEAKS048611	IRF8	9.40E-204	5.72E-06	5.07E-07	2.06E-03
PEAKS048612	IRF8	2.64E-214	5.34E-07	3.14E-07	8.13E-04
PEAKS048613	IRF8	2.13E-201	6.42E-06	3.52E-08	3.88E-03
PEAKS048614	IRF8	9.69E-213	5.12E-06	2.82E-08	1.53E-03
PEAKS048615	IRF8	9.75E-185	3.51E-06	5.65E-08	1.97E-03
PEAKS055166	IRF8	2.18E-32	9.71E-09	3.25E-07	2.19E-02
PEAKS048308	ISX	3.50E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS048310	ISX	3.31E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS036034	JUN	1.14E-130	3.22E-06	1.71E-06	1.89E-04
PEAKS036048	JUN	1.76E-111	1.19E-05	3.38E-06	5.03E-01
PEAKS036578	JUN	8.37E-207	9.46E-10	7.77E-05	4.58E-03

PEAKS037022	JUN	0.00E+00	6.15E-07	4.70E-07	4.63E-07
PEAKS038512	JUN	0.00E+00	3.28E-07	4.53E-06	1.74E-03
PEAKS048577	JUN	3.10E-08	1.00E+00	1.00E+00	1.00E+00
PEAKS048580	JUN	1.05E-02	1.00E+00	1.00E+00	1.00E+00
PEAKS049100	JUN	4.89E-96	5.98E-01	4.27E-01	8.90E-01
PEAKS049102	JUN	0.00E+00	7.57E-06	2.98E-06	1.81E-02
PEAKS049109	JUN	0.00E+00	2.45E-05	1.18E-06	5.70E-05
PEAKS049111	JUN	0.00E+00	2.38E-07	2.98E-06	4.82E-05
PEAKS049116	JUN	0.00E+00	1.14E-05	1.36E-07	1.27E-02
PEAKS049118	JUN	0.00E+00	1.66E-05	5.73E-09	2.07E-05
PEAKS049128	JUN	0.00E+00	1.10E-05	1.21E-08	2.16E-05
PEAKS049129	JUN	0.00E+00	3.70E-06	1.51E-08	1.36E-04
PEAKS049134	JUN	0.00E+00	2.90E-07	3.97E-08	6.53E-07
PEAKS049139	JUN	0.00E+00	9.66E-06	2.50E-06	8.69E-06
PEAKS049832	JUN	0.00E+00	1.75E-05	1.71E-05	6.63E-05
PEAKS054854	JUN	0.00E+00	2.53E-07	3.58E-06	3.90E-05
PEAKS054855	JUN	0.00E+00	2.82E-07	1.57E-08	1.54E-02
PEAKS054868	JUN	0.00E+00	9.61E-10	1.99E-09	4.47E-04
PEAKS054869	JUN	0.00E+00	1.54E-05	8.62E-07	3.27E-06
PEAKS036047	JUNB	0.00E+00	7.83E-08	3.15E-09	3.33E-06
PEAKS039312	JUNB	0.00E+00	2.50E-06	1.28E-07	5.59E-07
PEAKS040265	JUNB	0.00E+00	5.90E-06	2.15E-06	2.80E-07
PEAKS040266	JUNB	0.00E+00	4.36E-06	4.91E-08	1.18E-04
PEAKS040304	JUNB	0.00E+00	7.57E-10	3.71E-10	1.79E-06
PEAKS040305	JUNB	3.38E-180	2.98E-07	7.53E-01	1.48E-01
PEAKS040306	JUNB	1.31E-271	2.59E-06	4.02E-06	2.22E-04
PEAKS040888	JUNB	1.78E-292	2.01E-08	3.55E-06	1.58E-03
PEAKS040889	JUNB	0.00E+00	2.81E-08	1.21E-07	1.15E-04
PEAKS040973	JUNB	0.00E+00	1.58E-05	1.23E-07	8.52E-07
PEAKS040974	JUNB	0.00E+00	6.62E-06	3.81E-06	1.36E-06
PEAKS040975	JUNB	0.00E+00	5.07E-09	1.12E-08	3.95E-07
PEAKS040976	JUNB	0.00E+00	9.76E-09	7.91E-06	6.22E-06
PEAKS040977	JUNB	0.00E+00	3.06E-05	1.89E-06	5.15E-02
PEAKS048616	JUNB	7.08E-199	4.53E-06	2.62E-06	1.34E-02
PEAKS048617	JUNB	8.25E-225	1.19E-06	2.03E-06	4.82E-05
PEAKS048618	JUNB	2.16E-109	3.59E-02	3.10E-02	1.04E-01
PEAKS054860	JUNB	0.00E+00	1.50E-05	3.70E-06	1.61E-03
PEAKS054872	JUNB	0.00E+00	1.83E-08	1.67E-06	8.35E-03
PEAKS036049	JUND	0.00E+00	1.81E-05	1.43E-08	4.19E-07
PEAKS040890	JUND	0.00E+00	8.34E-06	7.53E-09	1.20E-04
PEAKS040891	JUND	0.00E+00	1.54E-05	3.77E-09	2.93E-06
PEAKS040978	JUND	0.00E+00	1.20E-08	6.13E-09	1.51E-01
PEAKS040979	JUND	0.00E+00	2.50E-08	1.75E-08	6.53E-02
PEAKS040980	JUND	0.00E+00	3.36E-07	3.91E-09	6.81E-02
PEAKS040981	JUND	0.00E+00	2.65E-07	7.89E-09	2.94E-07
PEAKS040982	JUND	0.00E+00	7.06E-06	2.29E-07	5.07E-02

PEAKS049578	JUND	0.00E+00	8.52E-09	8.20E-09	1.08E-07
PEAKS049579	JUND	0.00E+00	1.70E-09	4.02E-09	7.23E-07
PEAKS049580	JUND	0.00E+00	1.70E-09	3.72E-09	2.09E-09
PEAKS049581	JUND	0.00E+00	3.10E-09	3.71E-09	2.00E-07
PEAKS049584	JUND	0.00E+00	1.30E-07	1.51E-08	7.71E-02
PEAKS049585	JUND	0.00E+00	3.49E-08	3.88E-08	2.57E-04
PEAKS049586	JUND	0.00E+00	3.16E-08	4.51E-09	5.87E-04
PEAKS049587	JUND	0.00E+00	2.56E-09	1.74E-08	1.65E-04
PEAKS054862	JUND	0.00E+00	2.43E-07	3.92E-06	4.47E-06
PEAKS054863	JUND	0.00E+00	6.44E-09	2.23E-07	1.44E-06
PEAKS054874	JUND	0.00E+00	5.79E-10	3.34E-06	1.88E-04
PEAKS054875	JUND	0.00E+00	2.41E-07	2.59E-08	1.01E-03
PEAKS036273	KLF1	1.82E-63	9.29E-02	7.70E-02	3.02E-02
PEAKS038538	KLF15	3.19E-36	1.00E+00	1.00E+00	1.00E+00
PEAKS038344	KLF4	2.58E-121	6.36E-08	8.19E-08	5.44E-03
PEAKS039877	KLF4	4.10E-39	1.00E+00	1.00E+00	1.00E+00
PEAKS036343	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036344	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036345	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036346	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036347	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036348	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036376	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036377	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039730	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039803	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039804	KMT2B	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS042178	LEF1	4.87E-58	8.55E-01	8.01E-01	8.21E-01
PEAKS054995	LEF1	3.25E-298	1.77E-14	1.30E-12	4.93E-03
PEAKS039247	LHX2	2.18E-218	1.31E-06	3.27E-06	4.29E-02
PEAKS049705	LHX2	6.04E-302	9.06E-07	6.66E-07	6.57E-05
PEAKS049706	LHX2	2.42e-311	1.02E-05	2.27E-05	1.48E-04
PEAKS039229	LHX6	3.23E-268	4.23E-04	1.77E-04	1.47E-05
PEAKS036502	MAF	3.47E-191	2.09E-09	1.52E-09	9.24E-03
PEAKS036503	MAF	5.54E-200	3.55E-11	7.83E-10	4.09E-04
PEAKS041057	MAF	6.83E-154	4.30E-07	1.46E-08	1.80E-04
PEAKS041058	MAF	3.70E-150	1.12E-10	1.59E-11	4.92E-05
PEAKS038652	MAFB	3.84E-255	3.55E-06	2.03E-09	6.52E-04
PEAKS038653	MAFB	8.42E-269	6.85E-06	1.04E-08	1.22E-04
PEAKS035880	MAFF	0.00E+00	3.73E-10	1.64E-10	2.20E-04
PEAKS035881	MAFF	0.00E+00	2.37E-13	3.44E-11	1.02E-04
PEAKS035882	MAFF	0.00E+00	1.42E-12	3.65E-13	5.96E-03
PEAKS035883	MAFF	0.00E+00	4.37E-13	1.98E-13	1.83E-02
PEAKS038870	MAFG	0.00E+00	2.26E-10	4.22E-11	4.11E-04
PEAKS038871	MAFG	0.00E+00	2.50E-12	2.05E-13	1.46E-04
PEAKS040348	MAFK	0.00E+00	4.27E-10	5.04E-11	4.12E-05

PEAKS035467	MAX	6.49E-192	6.20E-06	1.23E-09	1.86E-02
PEAKS049931	MBD1	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS052831	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS052832	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS053153	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS053154	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS055122	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS055123	MECP2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS037872	MEF2A	0.00E+00	5.83E-12	8.89E-13	1.15E-02
PEAKS037873	MEF2A	0.00E+00	2.62E-12	2.94E-13	1.98E-01
PEAKS037870	MEF2C	6.39E-228	1.85E-11	6.08E-15	4.73E-01
PEAKS037871	MEF2C	7.05E-232	8.59E-13	9.79E-13	9.89E-01
PEAKS055083	MEF2C	5e-324	5.57E-09	3.78E-11	1.86E-02
PEAKS057466	MEF2C	8.40E-224	7.90E-12	5.75E-08	1.33E-01
PEAKS057467	MEF2C	1.11E-219	3.24E-13	5.45E-15	9.81E-01
PEAKS057468	MEF2C	3.85E-190	2.45E-13	1.05E-12	7.89E-01
PEAKS057469	MEF2C	5.82E-205	2.07E-12	7.40E-12	9.63E-01
PEAKS057475	MEF2C	5.22E-171	6.78E-10	1.99E-10	5.77E-01
PEAKS037447	MEF2D	0.00E+00	1.22E-09	1.98E-08	7.64E-03
PEAKS037448	MEF2D	0.00E+00	6.11E-10	1.96E-08	7.21E-02
PEAKS058108	MEF2D	0.00E+00	2.86E-14	7.78E-11	1.09E-03
PEAKS058109	MEF2D	0.00E+00	1.19E-10	1.05E-08	6.17E-01
PEAKS035335	MYB	1.59E-80	2.15E-05	1.19E-05	3.08E-03
PEAKS035584	MYBL1	1.52E-93	6.16E-04	2.16E-04	1.35E-04
PEAKS040028	MYBL1	9.48E-103	1.82E-04	1.85E-04	1.77E-02
PEAKS035964	MYC	7.58E-84	9.68E-06	5.77E-06	2.32E-02
PEAKS036488	MYC	2.62E-305	2.40E-09	2.11E-09	3.00E-04
PEAKS036666	MYC	2.35E-181	4.47E-07	6.54E-10	7.18E-03
PEAKS039097	MYC	4.74E-95	8.14E-06	9.92E-06	4.45E-04
PEAKS040310	MYC	5.52E-140	9.54E-06	9.47E-07	3.66E-03
PEAKS040311	MYC	8.83E-212	7.63E-06	1.99E-09	4.17E-05
PEAKS040312	MYC	7.73E-212	8.13E-06	7.37E-06	6.26E-03
PEAKS040313	MYC	2.62E-190	4.08E-07	2.87E-05	9.12E-03
PEAKS040314	MYC	2.33E-211	7.63E-06	6.86E-02	2.02E-04
PEAKS041116	MYC	1.72E-106	2.71E-03	1.12E-02	2.62E-03
PEAKS042251	MYC	9e-323	2.65E-09	1.25E-09	6.29E-04
PEAKS042252	MYC	7.63E-276	1.86E-07	3.96E-09	4.28E-06
PEAKS042253	MYC	1.35E-264	9.54E-06	1.88E-05	9.38E-04
PEAKS049409	MYC	1.90E-280	7.63E-06	3.56E-09	3.68E-02
PEAKS049504	MYC	6.71E-188	8.11E-02	3.01E-01	3.56E-02
PEAKS049505	MYC	5.50E-29	2.81E-01	4.97E-01	7.83E-02
PEAKS049508	MYC	6.07E-87	4.20E-02	3.31E-02	9.19E-03
PEAKS049509	MYC	7.50E-192	7.49E-03	2.14E-01	8.00E-02
PEAKS057309	MYC	6.49E-147	4.19E-01	8.77E-02	5.57E-03
PEAKS035495	MYOD1	0.00E+00	4.69E-08	3.63E-08	1.13E-08
PEAKS035690	MYOD1	7.78E-117	1.91E-06	1.92E-06	1.63E-04

PEAKS035691	MYOD1	0.00E+00	8.94E-10	7.67E-09	1.95E-08
PEAKS036373	MYOD1	0.00E+00	4.06E-11	1.45E-07	2.21E-06
PEAKS036714	MYOD1	0.00E+00	1.56E-09	4.46E-09	4.08E-06
PEAKS036729	MYOD1	0.00E+00	4.85E-09	7.88E-09	2.05E-04
PEAKS037024	MYOD1	0.00E+00	2.03E-09	8.34E-10	3.96E-06
PEAKS055156	MYOD1	0.00E+00	1.06E-09	2.33E-09	1.19E-06
PEAKS055157	MYOD1	0.00E+00	3.05E-06	5.05E-08	1.40E-04
PEAKS055159	MYOD1	0.00E+00	3.78E-08	2.66E-08	1.61E-06
PEAKS055161	MYOD1	0.00E+00	6.21E-10	6.77E-10	4.97E-03
PEAKS059241	MYOD1	0.00E+00	3.23E-09	3.97E-09	3.15E-04
PEAKS059242	MYOD1	0.00E+00	9.66E-10	1.26E-09	1.31E-05
PEAKS035509	MYOG	0.00E+00	1.91E-06	3.68E-07	2.20E-07
PEAKS036746	MYOG	0.00E+00	1.91E-06	2.60E-09	4.80E-09
PEAKS036759	MYOG	0.00E+00	4.92E-12	6.01E-10	7.56E-08
PEAKS037025	MYOG	0.00E+00	5.68E-13	1.55E-08	7.19E-09
PEAKS038399	NANOG	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038400	NANOG	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036598	NCOA2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039543	NCOA2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039544	NCOA2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039545	NCOA2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039576	NEUROD1	1.80E-294	7.63E-11	9.86E-13	4.00E-03
PEAKS037963	NEUROD2	1.58E-307	1.51E-07	1.47E-10	1.15E-01
PEAKS037964	NEUROD2	0.00E+00	6.78E-12	2.28E-11	3.69E-05
PEAKS037965	NEUROD2	2.27E-298	3.42E-09	8.27E-09	5.28E-06
PEAKS057246	NFATC1	5.57E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS037285	NFATC2	3.82E-168	4.27E-07	5.12E-01	6.42E-03
PEAKS036124	NFE2	4.76E-235	2.87E-10	8.47E-15	1.97E-02
PEAKS036889	NFE2	0.00E+00	2.28E-10	1.20E-10	5.45E-07
PEAKS036890	NFE2	0.00E+00	2.37E-08	2.92E-10	5.42E-04
PEAKS038494	NFE2L2	7.42E-236	3.40E-10	5.89E-12	1.81E-04
PEAKS038495	NFE2L2	3.19E-299	8.14E-13	1.34E-12	9.90E-02
PEAKS038626	NFE2L2	3.42E-284	1.14E-07	1.08E-07	9.67E-03
PEAKS038628	NFE2L2	0.00E+00	8.12E-20	7.52E-19	9.57E-04
PEAKS048772	NFE2L2	1.49E-250	9.35E-10	6.74E-06	7.54E-03
PEAKS048773	NFE2L2	1.38E-264	1.13E-13	1.93E-06	2.90E-01
PEAKS050707	NFE2L2	1.14E-151	1.31E-06	1.45E-09	1.20E-02
PEAKS050708	NFE2L2	0.00E+00	1.33E-10	4.59E-19	1.30E-02
PEAKS039228	NKX2-1	8.69E-104	1.05E-09	2.42E-09	2.18E-02
PEAKS059308	NKX2-2	2.49E-34	4.61E-03	1.61E-03	1.94E-01
PEAKS059309	NKX2-2	2.59E-60	4.93E-06	2.77E-06	1.38E-02
PEAKS036920	NKX2-5	4.99E-106	8.80E-02	1.34E-01	1.31E-01
PEAKS036086	NKX6-1	7.64E-123	4.28E-07	9.27E-05	1.75E-04
PEAKS035194	NR1D1	3.54E-102	1.57E-06	1.95E-07	1.24E-03
PEAKS035783	NR1D1	1.24E-45	2.15E-01	7.52E-02	4.69E-03
PEAKS037977	NR1D1	2.35E-04	1.00E+00	1.00E+00	1.00E+00

PEAKS037978	NR1D1	8.59E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS037987	NR1D1	1.44E-72	1.11E-07	1.00E-08	1.88E-01
PEAKS037991	NR1D1	6.61E-85	9.34E-02	8.90E-02	2.51E-02
PEAKS038537	NR1D1	1.58E-166	4.15E-07	2.01E-07	4.85E-02
PEAKS038982	NR1D1	7.66E-113	2.78E-08	3.83E-09	1.64E-02
PEAKS057427	NR1D1	5.59E-182	1.52E-10	9.15E-09	1.01E-03
PEAKS057428	NR1D1	2.94E-179	3.59E-08	1.06E-09	1.43E-01
PEAKS057429	NR1D1	1.23E-200	5.35E-08	7.95E-08	7.62E-04
PEAKS057430	NR1D1	2.95E-204	6.26E-09	1.12E-07	7.73E-04
PEAKS057908	NR1D1	5.96E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS035784	NR1D2	7.24E-50	3.82E-02	2.36E-08	1.35E-01
PEAKS042375	NR1H2	2.92E-129	9.81E-10	1.16E-08	1.16E-03
PEAKS039922	NR1H3	1.13E-119	1.00E+00	1.00E+00	1.00E+00
PEAKS039923	NR1H3	1.24E-164	1.00E+00	1.00E+00	1.00E+00
PEAKS039924	NR1H3	1.99E-87	1.00E+00	1.00E+00	1.00E+00
PEAKS039925	NR1H3	2.30E-112	1.00E+00	1.00E+00	1.00E+00
PEAKS039926	NR1H3	5.47E-59	1.00E+00	1.00E+00	1.00E+00
PEAKS042374	NR1H3	6.29e-319	1.00E+00	1.00E+00	1.00E+00
PEAKS059123	NR1H3	1.87E-123	1.00E+00	1.00E+00	1.00E+00
PEAKS059124	NR1H3	5.44E-47	1.00E+00	1.00E+00	1.00E+00
PEAKS059125	NR1H3	2.93E-63	1.00E+00	1.00E+00	1.00E+00
PEAKS059126	NR1H3	6.25E-53	1.00E+00	1.00E+00	1.00E+00
PEAKS036235	NR2F2	4.40E-104	3.76E-01	3.76E-01	1.42E-03
PEAKS036207	NR3C1	5.86E-24	1.00E+00	1.00E+00	1.00E+00
PEAKS036209	NR3C1	7.15E-301	2.77E-10	9.05E-11	1.63E-03
PEAKS036211	NR3C1	0.00E+00	3.54E-11	1.06E-11	6.35E-04
PEAKS036212	NR3C1	0.00E+00	1.55E-10	1.90E-11	5.12E-04
PEAKS036214	NR3C1	6.34E-301	4.99E-10	3.59E-10	1.30E-03
PEAKS036215	NR3C1	1.54E-43	1.00E+00	1.00E+00	1.00E+00
PEAKS036216	NR3C1	2.70E-270	3.97E-11	3.49E-11	1.67E-03
PEAKS037265	NR3C1	1.43E-29	1.00E+00	1.00E+00	1.00E+00
PEAKS037266	NR3C1	1.37E-35	1.00E+00	1.00E+00	1.00E+00
PEAKS037267	NR3C1	6.18E-26	1.73E-01	1.41E-04	8.53E-03
PEAKS037268	NR3C1	1.30E-168	1.10E-10	1.38E-10	3.79E-03
PEAKS037269	NR3C1	2.86E-144	5.29E-10	2.54E-10	1.26E-01
PEAKS037270	NR3C1	0.00E+00	1.58E-11	2.56E-11	2.74E-03
PEAKS037705	NR3C1	2.48E-210	5.61E-10	7.29E-10	2.00E-01
PEAKS037706	NR3C1	0.00E+00	9.75E-11	8.98E-12	8.32E-03
PEAKS040015	NR3C1	1.44E-176	5.28E-09	2.67E-10	9.41E-04
PEAKS040016	NR3C1	1.12E-127	3.19E-11	1.46E-09	3.94E-03
PEAKS048675	NR3C1	4.42E-225	7.19E-08	4.61E-08	3.04E-02
PEAKS048677	NR3C1	7.89E-180	3.65E-09	1.08E-09	2.72E-02
PEAKS050685	NR3C1	2.03E-288	4.15E-11	6.85E-11	1.89E-04
PEAKS050686	NR3C1	2.92E-209	1.31E-10	3.33E-11	9.20E-02
PEAKS038164	NR5A1	0.00E+00	4.26E-11	2.46E-10	8.48E-01
PEAKS035220	NR5A2	1.52E-286	2.95E-12	2.68E-13	6.64E-02

PEAKS042291	OLIG2	2.04E-102	1.91E-06	9.47E-08	1.49E-02
PEAKS049624	OSR1	1.32E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS036637	OTX2	1.21E-228	2.95E-09	1.79E-06	4.17E-02
PEAKS036638	OTX2	4.72E-251	3.93E-06	2.38E-06	9.64E-02
PEAKS052804	OTX2	2.65E-139	1.51E-07	1.55E-08	9.08E-05
PEAKS052806	OTX2	4.82E-192	4.41E-06	4.06E-08	2.81E-01
PEAKS036574	OVOL2	0.00E+00	8.90E-11	1.65E-10	1.17E-04
PEAKS035556	PAX5	1.22E-56	3.20E-05	3.75E-05	7.38E-01
PEAKS035557	PAX5	4.57E-55	1.00E+00	1.00E+00	1.00E+00
PEAKS039248	PAX5	4.53E-100	9.49E-06	5.89E-06	5.77E-03
PEAKS039249	PAX5	5.23E-96	7.56E-06	5.78E-07	2.51E-02
PEAKS048635	PAX5	4.53E-89	2.31E-07	2.57E-07	3.40E-02
PEAKS048636	PAX5	8.61E-160	4.19E-05	4.45E-05	9.03E-03
PEAKS049203	PAX5	8.86E-129	2.78E-07	1.27E-05	2.18E-02
PEAKS037886	PAX6	2.16E-110	1.51E-08	1.49E-09	2.85E-01
PEAKS037887	PAX6	4.06E-121	9.00E-08	1.97E-09	1.82E-01
PEAKS035694	PAX7	0.00E+00	2.83E-09	4.45E-10	2.35E-08
PEAKS038950	PBX1	9.20E-170	2.57E-03	7.99E-04	6.74E-03
PEAKS038980	PBX1	5.80E-173	2.11E-03	7.14E-09	4.58E-02
PEAKS049655	PDX1	8.72E-30	8.09E-03	7.75E-03	2.39E-04
PEAKS035795	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035796	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036145	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038873	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS050639	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS050640	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS057626	PGR	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035398	POU2F1	1.04E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS037836	POU2F1	4.96E-10	1.00E+00	1.00E+00	1.00E+00
PEAKS037837	POU2F1	7.07E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS049292	POU2F3	0.00E+00	5.52E-08	4.91E-08	2.99E-02
PEAKS036635	POU5F1	0.00E+00	1.45E-08	4.54E-09	2.03E-02
PEAKS036636	POU5F1	0.00E+00	3.66E-07	1.19E-07	2.04E-03
PEAKS039671	PPARA	3.72E-247	1.00E+00	1.00E+00	1.00E+00
PEAKS039673	PPARA	0.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039675	PPARA	1.95E-289	1.00E+00	1.00E+00	1.00E+00
PEAKS039677	PPARA	3.33E-266	1.00E+00	1.00E+00	1.00E+00
PEAKS055023	PPARA	7.80E-87	1.00E+00	1.00E+00	1.00E+00
PEAKS055024	PPARA	3.93E-218	1.00E+00	1.00E+00	1.00E+00
PEAKS055025	PPARA	3.25E-120	1.00E+00	1.00E+00	1.00E+00
PEAKS055026	PPARA	1.62E-89	1.00E+00	1.00E+00	1.00E+00
PEAKS055027	PPARD	9.90E-150	6.06E-09	9.92E-11	4.24E-03
PEAKS055028	PPARD	1.09E-140	1.22E-08	5.58E-10	1.86E-03
PEAKS036102	PPARG	5.95E-203	2.51E-07	1.79E-07	7.26E-02
PEAKS036103	PPARG	7.92E-240	1.38E-08	3.66E-10	3.00E-01
PEAKS036104	PPARG	8.80E-241	2.09E-07	1.43E-07	3.80E-02

PEAKS037701	PPARG	7.94E-194	1.74E-08	4.57E-10	3.45E-02
PEAKS037702	PPARG	9.58E-197	3.86E-06	1.46E-06	6.03E-02
PEAKS037703	PPARG	8.31E-213	2.63E-07	3.69E-10	1.74E-01
PEAKS039024	PPARG	4.70E-208	3.65E-09	3.95E-10	1.42E-01
PEAKS039032	PPARG	2.70E-193	4.61E-07	3.92E-10	4.83E-02
PEAKS039046	PPARG	2.02E-291	1.14E-07	6.99E-09	2.07E-02
PEAKS039797	PPARG	5.10E-189	1.11E-06	1.01E-09	8.58E-02
PEAKS039798	PPARG	2.79E-229	4.34E-07	3.70E-10	4.43E-02
PEAKS040603	PPARG	6.43E-245	9.27E-08	7.11E-09	9.01E-02
PEAKS040615	PPARG	1.32E-126	5.02E-08	5.91E-09	1.26E-02
PEAKS042490	PPARG	5.26E-43	1.60E-05	1.98E-01	9.88E-02
PEAKS042491	PPARG	2.43E-207	1.17E-08	3.45E-09	3.18E-01
PEAKS054849	PPARG	5.55E-174	6.94E-09	1.82E-09	4.45E-01
PEAKS037844	PRDM1	5.19E-160	8.34E-04	6.29E-04	4.80E-04
PEAKS038654	PRDM1	1.17E-10	1.00E+00	1.00E+00	1.00E+00
PEAKS049648	PRDM1	7.33E-199	1.56E-04	1.59E-04	5.25E-05
PEAKS037668	PRDM16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS052630	PRDM9	6.80E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS053234	PRDM9	1.22E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS053237	PRDM9	1.36E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS053238	PRDM9	6.25E-14	1.00E+00	1.00E+00	1.00E+00
PEAKS053239	PRDM9	1.19E-02	1.00E+00	1.00E+00	1.00E+00
PEAKS035596	PTF1A	8.95E-191	4.20E-05	2.86E-07	8.57E-06
PEAKS039232	PTF1A	6.51E-194	3.13E-06	2.03E-04	1.43E-04
PEAKS039233	PTF1A	2.55E-144	2.54E-05	7.74E-06	3.52E-06
PEAKS049408	PTF1A	5.56E-167	5.57E-06	2.71E-07	1.29E-05
PEAKS048625	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048626	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048627	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048628	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048629	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048630	PURB	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036433	RARB	6.49E-51	1.00E+00	1.00E+00	1.00E+00
PEAKS037973	RARB	6.40E-46	1.00E+00	1.00E+00	1.00E+00
PEAKS059131	RBPJ	4.89E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS059133	RBPJ	7.90E-13	1.00E+00	1.00E+00	1.00E+00
PEAKS059134	RBPJ	2.59E-12	1.00E+00	1.00E+00	1.00E+00
PEAKS035598	RBPJL	2.61E-149	1.73E-05	1.05E-06	3.58E-05
PEAKS035911	REL	2.32E-175	5.39E-07	1.14E-07	1.46E-03
PEAKS035358	RELA	4.57E-202	2.58E-09	1.89E-08	1.25E-03
PEAKS035606	RELA	1.95E-08	1.00E+00	1.00E+00	1.00E+00
PEAKS035607	RELA	6.96E-254	1.41E-08	2.24E-09	1.97E-04
PEAKS035608	RELA	1.35E-221	1.22E-07	1.26E-09	1.23E-04
PEAKS035892	RELA	7.42E-127	6.16E-01	4.24E-01	6.51E-02
PEAKS035893	RELA	8.04E-167	1.44E-07	3.05E-07	1.68E-03
PEAKS035894	RELA	9.88E-136	1.24E-08	4.85E-09	3.82E-01

PEAKS035895	RELA	1.85E-152	8.32E-09	8.74E-08	1.01E-04
PEAKS036247	RELA	3.35E-178	4.41E-09	5.13E-09	6.21E-04
PEAKS036248	RELA	4.73E-176	2.74E-07	9.69E-10	1.14E-03
PEAKS036311	RELA	3.77E-218	9.64E-08	2.36E-08	1.39E-03
PEAKS036312	RELA	3.76E-211	5.57E-09	9.71E-10	4.26E-05
PEAKS036315	RELA	1.16E-254	1.01E-08	7.12E-10	1.28E-03
PEAKS036320	RELA	8.29E-173	1.79E-09	1.65E-08	8.47E-04
PEAKS036321	RELA	3.61E-254	2.09E-08	1.04E-08	4.93E-03
PEAKS036324	RELA	2.51E-200	5.14E-09	9.22E-09	1.99E-04
PEAKS036325	RELA	8.71E-151	4.20E-09	6.01E-09	2.66E-04
PEAKS037537	RELA	6.58E-187	6.91E-09	5.16E-08	3.67E-03
PEAKS037538	RELA	4.53E-168	2.34E-08	5.85E-08	7.43E-01
PEAKS037927	RELA	0.00E+00	1.99E-08	2.70E-08	1.33E-04
PEAKS037928	RELA	0.00E+00	3.95E-09	1.99E-08	6.53E-04
PEAKS037929	RELA	3.49E-252	2.04E-09	8.62E-09	6.82E-04
PEAKS037930	RELA	3.00E-235	2.33E-09	4.54E-09	9.64E-05
PEAKS038485	RELA	7.97E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS038486	RELA	1.60E-43	1.00E+00	1.00E+00	1.00E+00
PEAKS038489	RELA	1.42E-144	1.88E-04	2.76E-09	1.22E-03
PEAKS038490	RELA	4.96E-94	3.37E-02	3.45E-02	4.64E-04
PEAKS038491	RELA	1.18E-139	4.24E-03	1.09E-02	5.74E-03
PEAKS039927	RELA	8.65E-97	8.90E-03	9.37E-03	6.18E-03
PEAKS039928	RELA	4.38E-199	5.63E-09	6.40E-08	5.99E-04
PEAKS039929	RELA	1.48E-121	2.84E-03	2.92E-03	3.16E-04
PEAKS039930	RELA	2.50E-133	2.84E-09	4.73E-09	1.09E-02
PEAKS039931	RELA	5.11E-81	9.74E-03	7.57E-03	1.71E-02
PEAKS039932	RELA	2.25E-190	2.01E-09	3.42E-08	1.50E-03
PEAKS039933	RELA	5.02E-140	3.14E-09	7.03E-08	9.11E-04
PEAKS039934	RELA	1.16E-131	5.70E-09	1.15E-09	9.90E-05
PEAKS040670	RELA	1.38E-188	2.48E-08	2.84E-08	1.15E-04
PEAKS040671	RELA	4.79E-149	1.35E-08	1.21E-08	2.06E-03
PEAKS040989	RELA	1.72E-261	1.91E-08	3.46E-07	2.49E-04
PEAKS040990	RELA	9.15E-250	1.21E-08	5.38E-08	2.19E-04
PEAKS040991	RELA	1.85E-238	1.03E-08	9.53E-09	4.39E-05
PEAKS040992	RELA	7.06E-259	9.76E-08	2.45E-08	7.13E-04
PEAKS040994	RELA	3.63E-255	7.65E-08	3.58E-08	9.75E-04
PEAKS040995	RELA	3.73E-227	1.94E-07	1.04E-08	6.02E-06
PEAKS040996	RELA	2.13E-193	1.45E-08	2.12E-07	1.61E-04
PEAKS040997	RELA	6.88E-217	1.30E-08	1.23E-08	9.70E-04
PEAKS048680	RELA	1.62E-132	1.32E-06	6.18E-10	3.64E-04
PEAKS048681	RELA	2.03E-21	1.00E+00	1.00E+00	1.00E+00
PEAKS049101	RELA	5.56E-229	1.24E-09	3.41E-09	1.69E-04
PEAKS049103	RELA	2.25E-239	1.90E-05	4.74E-08	5.54E-02
PEAKS049108	RELA	1.29E-196	7.63E-08	3.16E-08	1.39E-03
PEAKS049110	RELA	1.28E-211	5.17E-08	4.53E-08	1.48E-04
PEAKS049117	RELA	4.23E-268	2.17E-08	5.17E-09	2.88E-04

PEAKS049119	RELA	7.03E-225	3.21E-09	1.41E-08	9.17E-04
PEAKS049130	RELA	6.78E-243	6.21E-09	2.04E-08	7.36E-05
PEAKS049131	RELA	6.94E-42	1.00E+00	1.00E+00	1.00E+00
PEAKS049135	RELA	2.99E-43	1.00E+00	1.00E+00	1.00E+00
PEAKS049138	RELA	7.54E-210	4.06E-08	1.09E-08	1.42E-04
PEAKS049792	RELA	1.20E-63	1.00E+00	1.00E+00	1.00E+00
PEAKS049793	RELA	2.71E-202	2.16E-08	2.37E-08	4.66E-04
PEAKS049794	RELA	8.23E-173	6.14E-08	1.07E-08	4.00E-04
PEAKS049795	RELA	2.02E-191	1.39E-08	3.52E-08	9.43E-04
PEAKS049796	RELA	8.57E-171	1.36E-08	6.65E-09	1.06E-03
PEAKS049797	RELA	6.54E-162	1.02E-09	6.90E-09	9.08E-05
PEAKS049799	RELA	5.40E-201	7.67E-09	2.96E-09	3.01E-04
PEAKS049800	RELA	1.49E-144	7.45E-09	1.91E-08	9.58E-01
PEAKS049801	RELA	2.39E-171	1.26E-08	1.80E-10	6.56E-05
PEAKS049802	RELA	2.65E-178	1.87E-08	8.61E-10	7.39E-04
PEAKS049830	RELA	0.00E+00	5.20E-09	2.09E-09	1.11E-03
PEAKS049983	RELA	7.85E-84	1.40E-03	4.09E-04	9.35E-02
PEAKS049984	RELA	5.22E-77	1.35E-01	4.09E-02	7.71E-01
PEAKS049986	RELA	5.33E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS049987	RELA	2.10E-78	1.00E+00	1.00E+00	1.00E+00
PEAKS049988	RELA	2.09E-77	6.39E-09	5.27E-07	4.19E-01
PEAKS049990	RELA	4.68E-11	1.00E+00	1.00E+00	1.00E+00
PEAKS050657	RELA	2.28E-112	2.10E-02	1.98E-02	2.44E-01
PEAKS050658	RELA	2.16E-101	1.69E-02	1.61E-02	2.78E-03
PEAKS050659	RELA	4.07E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS050660	RELA	7.73E-03	2.37E-02	2.02E-02	8.62E-01
PEAKS050671	RELA	1.33E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS050755	RELA	1.06E-217	6.37E-09	1.45E-08	1.72E-05
PEAKS059247	RELA	0.00E+00	2.62E-09	5.63E-09	3.55E-04
PEAKS035889	RELB	1.81E-183	3.69E-08	1.31E-07	3.11E-04
PEAKS035891	RELB	2.08E-88	2.87E-08	2.15E-08	2.48E-01
PEAKS036719	REST	0.00E+00	1.50E-17	1.05E-14	4.72E-03
PEAKS049047	REST	0.00E+00	3.42E-18	5.03E-17	4.45E-02
PEAKS049048	REST	0.00E+00	1.11E-15	6.66E-15	7.69E-02
PEAKS049049	REST	0.00E+00	8.12E-20	1.13E-17	1.04E-01
PEAKS049050	REST	0.00E+00	3.51E-16	3.51E-15	1.12E-01
PEAKS038844	RFX1	0.00E+00	7.25E-07	3.03E-12	2.14E-03
PEAKS038845	RFX1	0.00E+00	7.04E-07	2.22E-12	2.10E-01
PEAKS038015	RFX2	0.00E+00	1.28E-06	1.42E-12	2.30E-01
PEAKS038016	RFX2	0.00E+00	8.83E-07	4.14E-12	1.48E-01
PEAKS039259	RORC	5.05E-119	1.03E-09	1.49E-10	4.28E-01
PEAKS039261	RORC	1.98E-38	1.00E+00	1.00E+00	1.00E+00
PEAKS041089	RORC	2.46E-162	1.48E-09	1.44E-06	1.94E-01
PEAKS041091	RORC	0.00E+00	1.81E-10	1.51E-10	3.15E-02
PEAKS041092	RORC	0.00E+00	2.32E-11	2.66E-11	1.88E-02
PEAKS035774	RUNX1	2.66E-162	8.05E-06	2.16E-10	1.87E-02

PEAKS035896	RUNX1	3.43E-248	5.85E-01	4.80E-11	2.06E-01
PEAKS035897	RUNX1	3.67E-78	2.46E-06	5.67E-06	5.26E-03
PEAKS035898	RUNX1	2.31E-115	2.80E-10	1.38E-06	4.01E-02
PEAKS036191	RUNX1	6.86E-244	7.81E-13	1.14E-12	6.11E-06
PEAKS036887	RUNX1	1.29E-30	9.43E-05	2.17E-03	1.05E-01
PEAKS037023	RUNX1	7.22E-205	2.86E-08	3.85E-09	7.36E-07
PEAKS040480	RUNX1	8.44E-11	1.00E+00	1.00E+00	1.00E+00
PEAKS040481	RUNX1	5.75E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS040913	RUNX1	1.35E-192	9.73E-11	3.44E-11	2.43E-02
PEAKS048531	RUNX1	4.06E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS048532	RUNX1	2.18E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS048533	RUNX1	1.70E-115	1.56E-05	3.96E-05	1.69E-01
PEAKS048534	RUNX1	1.89E-63	4.13E-01	1.03E-01	7.14E-05
PEAKS049059	RUNX1	3.07E-116	3.85E-01	1.33E-05	3.51E-01
PEAKS049060	RUNX1	6.32E-54	7.75E-02	4.09E-02	1.72E-02
PEAKS049061	RUNX1	4.38E-98	2.22E-01	1.99E-01	5.50E-04
PEAKS049062	RUNX1	3.43E-51	5.51E-02	2.01E-01	8.09E-03
PEAKS049090	RUNX1	1.43E-280	1.39E-11	8.52E-13	4.92E-04
PEAKS049091	RUNX1	0.00E+00	1.41E-06	5.76E-13	1.95E-07
PEAKS036576	RUNX2	5.24E-150	4.96E-09	1.98E-11	2.45E-05
PEAKS039021	RUNX2	0.00E+00	3.12E-11	3.60E-14	1.03E-01
PEAKS039029	RUNX2	0.00E+00	3.26E-12	1.80E-13	1.94E-05
PEAKS039037	RUNX2	2.18E-267	3.33E-06	6.03E-12	1.50E-05
PEAKS039043	RUNX2	0.00E+00	4.21E-09	2.29E-09	2.33E-06
PEAKS039851	RUNX2	1.07E-284	3.21E-11	1.97E-11	1.30E-03
PEAKS039853	RUNX2	4.24E-295	3.99E-14	2.70E-14	1.91E-02
PEAKS039855	RUNX2	4.15E-164	3.89E-12	3.90E-07	1.74E-04
PEAKS039857	RUNX2	2.03E-200	2.22E-06	9.94E-12	4.88E-04
PEAKS036336	RUNX3	4.71E-241	1.79E-06	1.02E-06	2.64E-04
PEAKS036338	RUNX3	6.18e-315	2.70E-12	4.93E-11	3.88E-06
PEAKS036340	RUNX3	5.25E-278	7.19E-14	2.51E-08	4.46E-03
PEAKS036341	RUNX3	3.22E-242	7.37E-11	4.79E-11	1.52E-04
PEAKS036342	RUNX3	3.77E-217	7.92E-07	2.83E-01	1.09E-03
PEAKS036349	RUNX3	1.11E-133	1.04E-06	2.04E-05	1.66E-04
PEAKS036255	RXRA	5.41E-180	5.11E-09	1.93E-07	2.93E-04
PEAKS036431	RXRA	6.60E-179	4.14E-08	1.59E-07	1.72E-04
PEAKS036434	RXRA	2.74E-174	2.67E-06	7.75E-07	8.37E-05
PEAKS036449	RXRA	7.78E-89	6.21E-06	3.18E-07	3.62E-04
PEAKS036450	RXRA	3.31E-93	4.02E-07	5.97E-09	7.95E-02
PEAKS036451	RXRA	1.31E-89	1.99E-04	9.40E-08	9.05E-05
PEAKS036452	RXRA	5.60E-84	1.77E-06	1.56E-07	2.34E-02
PEAKS036453	RXRA	1.59E-172	2.67E-10	5.77E-09	9.80E-05
PEAKS036454	RXRA	1.49E-158	4.98E-04	7.27E-09	4.73E-05
PEAKS036455	RXRA	1.39E-182	2.20E-09	1.29E-07	1.78E-04
PEAKS036456	RXRA	1.72E-169	3.23E-09	9.63E-08	1.67E-03
PEAKS039019	RXRA	7.37E-111	1.00E-01	8.15E-02	6.50E-02

PEAKS039020	RXRA	1.85E-108	3.97E-05	1.01E-04	7.85E-03
PEAKS039027	RXRA	2.50E-86	2.51E-02	4.85E-02	7.76E-01
PEAKS039028	RXRA	5.48E-94	6.65E-02	2.69E-02	3.40E-02
PEAKS039035	RXRA	1.58E-99	1.50E-01	1.02E-07	2.24E-01
PEAKS039036	RXRA	4.85E-99	1.74E-02	8.32E-02	3.75E-01
PEAKS039041	RXRA	0.00E+00	8.76E-10	5.94E-10	2.79E-02
PEAKS039042	RXRA	0.00E+00	6.70E-10	6.93E-10	2.63E-02
PEAKS039670	RXRA	6.46E-138	6.84E-08	4.63E-07	1.74E-04
PEAKS039672	RXRA	5.31E-132	3.60E-09	6.81E-09	7.45E-05
PEAKS039674	RXRA	4.99E-171	2.70E-09	5.95E-10	1.33E-03
PEAKS039676	RXRA	2.54E-167	6.25E-08	1.29E-08	4.65E-05
PEAKS042495	RXRA	5.91E-116	1.88E-01	2.06E-01	1.50E-03
PEAKS057208	RXRA	1.80E-217	4.30E-08	5.59E-10	1.46E-06
PEAKS057209	RXRA	7.06E-209	1.24E-08	8.51E-08	4.30E-06
PEAKS038516	SALL4	1.00E+00	7.43E-02	9.01E-02	3.97E-03
PEAKS038517	SALL4	1.00E+00	1.29E-01	1.15E-01	2.08E-02
PEAKS038518	SALL4	1.00E+00	3.24E-01	3.99E-01	2.79E-03
PEAKS037860	SATB1	3.99E-75	1.00E+00	1.00E+00	1.00E+00
PEAKS036054	SIX2	9.89E-189	3.98E-12	2.48E-11	3.60E-02
PEAKS036055	SIX2	1.94E-162	4.16E-08	7.94E-08	3.44E-02
PEAKS049615	SIX2	2.87E-151	7.01E-09	2.75E-09	6.00E-02
PEAKS049621	SIX2	1.25E-215	1.43E-12	5.68E-12	7.19E-02
PEAKS049622	SIX2	2.67E-190	6.52E-10	7.85E-11	4.13E-01
PEAKS049623	SIX2	4.94E-208	2.50E-09	3.58E-12	2.82E-01
PEAKS038488	SMAD3	3.67E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS038493	SMAD3	6.98E-60	9.28E-01	6.68E-01	1.44E-03
PEAKS037135	SMAD4	2.73E-54	7.44E-06	4.57E-01	5.58E-01
PEAKS042204	SMAD4	4.17E-49	3.32E-01	4.89E-01	8.90E-03
PEAKS055750	SMAD4	1.12E-32	1.00E+00	1.00E+00	1.00E+00
PEAKS059127	SMAD4	3.09E-25	9.04E-01	8.79E-01	8.06E-01
PEAKS059128	SMAD4	7.77E-16	9.57E-01	8.90E-01	5.31E-02
PEAKS059129	SMAD4	1.15E-51	9.71E-01	9.51E-01	4.31E-01
PEAKS059130	SMAD4	8.44E-45	9.51E-01	9.20E-01	9.75E-01
PEAKS052779	SMYD3	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS052780	SMYD3	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038220	SOX10	1.54E-119	1.11E-08	1.14E-06	4.70E-01
PEAKS057174	SOX11	0.00E+00	4.34E-10	9.40E-12	3.40E-04
PEAKS039191	SOX2	1.61E-143	9.84E-10	6.45E-11	4.48E-03
PEAKS048661	SOX2	5.61E-63	6.89E-01	7.20E-01	1.12E-04
PEAKS048662	SOX2	2.19E-78	4.19E-01	5.90E-01	6.93E-03
PEAKS048663	SOX2	5.58E-26	6.75E-01	6.63E-01	6.48E-03
PEAKS049524	SOX2	1.34E-108	5.14E-10	9.20E-11	4.07E-04
PEAKS055224	SOX2	3.01E-144	2.06E-09	5.03E-08	2.63E-01
PEAKS048643	SOX30	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036535	SOX4	4.16E-283	2.98E-09	8.48E-09	4.82E-03
PEAKS057173	SOX4	0.00E+00	2.55E-13	2.89E-12	8.65E-04

PEAKS035367	SOX6	1.70E-105	3.77E-01	2.91E-01	4.53E-03
PEAKS038125	SOX9	2.63E-06	7.04E-01	6.57E-01	2.83E-01
PEAKS038504	SOX9	5.15E-54	3.14E-08	3.78E-09	3.59E-04
PEAKS038700	SP7	1.00E+00	2.81E-01	2.66E-01	1.36E-01
PEAKS039079	SP7	1.00E+00	1.23E-01	1.97E-02	6.61E-02
PEAKS039254	SP9	8.88E-08	6.82E-03	4.10E-03	4.36E-03
PEAKS050021	SPEN	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS050033	SPEN	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036237	SPI1	0.00E+00	2.34E-13	2.49E-14	1.11E-07
PEAKS036239	SPI1	0.00E+00	1.19E-14	1.68E-14	5.06E-06
PEAKS036251	SPI1	0.00E+00	1.06E-12	1.54E-17	4.87E-06
PEAKS036252	SPI1	0.00E+00	1.47E-13	1.20E-16	2.73E-04
PEAKS036305	SPI1	0.00E+00	5.03E-12	6.33E-12	1.09E-06
PEAKS036306	SPI1	0.00E+00	4.27E-12	4.44E-12	4.55E-06
PEAKS036307	SPI1	0.00E+00	1.86E-12	5.54E-12	6.00E-07
PEAKS036309	SPI1	0.00E+00	8.20E-13	2.56E-12	5.52E-04
PEAKS036318	SPI1	0.00E+00	2.60E-12	3.07E-12	7.14E-06
PEAKS036319	SPI1	0.00E+00	4.26E-11	1.98E-11	1.11E-06
PEAKS036322	SPI1	0.00E+00	3.05E-11	1.89E-12	2.98E-06
PEAKS036323	SPI1	0.00E+00	1.17E-11	1.61E-11	2.49E-05
PEAKS036424	SPI1	0.00E+00	1.63E-14	2.16E-15	1.74E-06
PEAKS036891	SPI1	0.00E+00	1.69E-11	1.81E-11	3.39E-07
PEAKS037583	SPI1	0.00E+00	1.21E-11	3.72E-15	3.41E-05
PEAKS037584	SPI1	0.00E+00	5.94E-11	3.12E-11	6.90E-07
PEAKS037585	SPI1	0.00E+00	3.32E-13	3.51E-11	3.00E-06
PEAKS037586	SPI1	0.00E+00	2.23E-13	1.34E-11	8.43E-06
PEAKS037587	SPI1	0.00E+00	1.33E-11	1.20E-11	2.53E-05
PEAKS037588	SPI1	0.00E+00	3.06E-14	2.55E-14	1.69E-04
PEAKS037590	SPI1	0.00E+00	9.60E-13	1.77E-13	1.10E-02
PEAKS037591	SPI1	0.00E+00	1.17E-12	1.39E-12	1.44E-04
PEAKS037592	SPI1	0.00E+00	1.11E-14	5.37E-13	2.13E-05
PEAKS037593	SPI1	0.00E+00	2.08E-12	1.50E-12	8.78E-06
PEAKS037594	SPI1	0.00E+00	1.43E-12	2.42E-13	1.50E-07
PEAKS038347	SPI1	0.00E+00	4.30E-11	1.00E-11	8.21E-04
PEAKS038348	SPI1	0.00E+00	7.49E-12	1.13E-11	2.80E-06
PEAKS038479	SPI1	0.00E+00	1.11E-10	2.29E-10	7.86E-05
PEAKS038480	SPI1	0.00E+00	3.65E-10	2.66E-09	1.23E-04
PEAKS038649	SPI1	0.00E+00	5.30E-11	3.79E-14	1.79E-05
PEAKS038650	SPI1	0.00E+00	1.39E-11	1.93E-14	3.94E-02
PEAKS038651	SPI1	0.00E+00	8.99E-12	9.81E-13	4.99E-04
PEAKS038921	SPI1	0.00E+00	2.99E-12	1.26E-11	1.36E-05
PEAKS038922	SPI1	0.00E+00	3.70E-12	1.62E-12	5.52E-04
PEAKS038923	SPI1	0.00E+00	1.01E-11	1.59E-15	3.22E-05
PEAKS040350	SPI1	0.00E+00	2.40E-09	1.23E-10	9.86E-03
PEAKS040351	SPI1	0.00E+00	3.06E-07	4.19E-08	1.71E-03
PEAKS040408	SPI1	0.00E+00	3.56E-12	1.17E-12	6.52E-06

PEAKS040409	SPI1	0.00E+00	3.82E-12	2.04E-14	2.43E-05
PEAKS040453	SPI1	0.00E+00	8.10E-12	3.19E-11	1.54E-08
PEAKS040478	SPI1	0.00E+00	2.41E-10	1.21E-14	8.40E-02
PEAKS040479	SPI1	0.00E+00	3.03E-13	2.12E-09	3.27E-04
PEAKS042478	SPI1	0.00E+00	2.61E-11	6.46E-14	1.81E-04
PEAKS042479	SPI1	0.00E+00	2.79E-11	8.80E-15	3.28E-05
PEAKS048427	SPI1	0.00E+00	1.12E-10	7.71E-11	6.01E-03
PEAKS048428	SPI1	0.00E+00	7.93E-12	2.43E-11	1.36E-05
PEAKS048682	SPI1	0.00E+00	2.14E-12	1.97E-11	1.32E-03
PEAKS048683	SPI1	0.00E+00	5.83E-12	1.08E-12	1.53E-04
PEAKS048684	SPI1	0.00E+00	1.72E-12	5.95E-12	5.31E-05
PEAKS049092	SPI1	0.00E+00	2.28E-12	1.91E-13	1.48E-06
PEAKS049093	SPI1	0.00E+00	1.01E-11	3.21E-11	1.40E-07
PEAKS049098	SPI1	0.00E+00	2.37E-13	9.35E-12	5.48E-07
PEAKS049099	SPI1	0.00E+00	8.72E-13	7.42E-13	8.02E-07
PEAKS049105	SPI1	0.00E+00	7.62E-12	6.68E-12	1.07E-04
PEAKS049107	SPI1	0.00E+00	3.18E-11	5.59E-13	4.97E-06
PEAKS049113	SPI1	0.00E+00	8.52E-13	2.34E-12	3.82E-04
PEAKS049115	SPI1	0.00E+00	9.78E-12	4.73E-12	1.36E-04
PEAKS049120	SPI1	0.00E+00	4.88E-12	3.87E-13	1.41E-05
PEAKS049121	SPI1	0.00E+00	2.38E-12	2.73E-13	6.52E-07
PEAKS049123	SPI1	0.00E+00	1.95E-13	3.70E-15	1.29E-03
PEAKS049125	SPI1	0.00E+00	9.34E-12	9.08E-12	1.48E-04
PEAKS049136	SPI1	0.00E+00	1.47E-12	2.25E-13	2.29E-06
PEAKS049137	SPI1	0.00E+00	1.47E-12	7.38E-12	5.68E-03
PEAKS049363	SPI1	0.00E+00	9.03E-13	2.16E-14	7.16E-07
PEAKS049364	SPI1	0.00E+00	2.51E-12	3.13E-13	2.53E-03
PEAKS049728	SPI1	0.00E+00	1.03E-13	6.08E-10	1.38E-04
PEAKS049729	SPI1	0.00E+00	1.77E-13	1.95E-14	3.77E-03
PEAKS054864	SPI1	0.00E+00	4.91E-11	1.75E-11	1.52E-03
PEAKS054865	SPI1	0.00E+00	1.99E-12	1.23E-13	1.01E-05
PEAKS055167	SPI1	0.00E+00	2.46E-13	1.06E-12	2.18E-04
PEAKS058091	SPI1	1.90E-90	3.37E-08	8.83E-07	1.40E-04
PEAKS058093	SPI1	2.20E-190	9.11E-09	2.58E-09	1.27E-04
PEAKS036519	SREBF1	6.52E-22	1.00E+00	1.00E+00	1.00E+00
PEAKS036520	SREBF1	9.06E-20	1.00E+00	1.00E+00	1.00E+00
PEAKS036521	SREBF1	2.36E-93	1.00E+00	1.00E+00	1.00E+00
PEAKS036522	SREBF1	5.67E-42	1.00E+00	1.00E+00	1.00E+00
PEAKS036523	SREBF1	7.36E-113	6.08E-08	7.94E-08	2.13E-05
PEAKS036524	SREBF1	1.43E-42	1.00E+00	1.00E+00	1.00E+00
PEAKS039920	SREBF1	7.57E-21	1.14E-05	3.45E-04	2.08E-01
PEAKS039921	SREBF1	7.92E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS040237	SREBF1	1.22E-162	5.95E-08	1.97E-08	1.51E-02
PEAKS036443	SRF	4.96E-11	1.00E+00	1.00E+00	1.00E+00
PEAKS036742	SRF	1.55E-219	1.00E+00	1.00E+00	1.00E+00
PEAKS047267	SRF	3.34E-139	1.00E+00	1.00E+00	1.00E+00

PEAKS047269	SRF	3.39E-28	1.00E+00	1.00E+00	1.00E+00
PEAKS047270	SRF	5.44E-17	1.00E+00	1.00E+00	1.00E+00
PEAKS047271	SRF	1.14E-97	1.00E+00	1.00E+00	1.00E+00
PEAKS047272	SRF	2.95E-43	1.00E+00	1.00E+00	1.00E+00
PEAKS047277	SRF	5.94E-117	1.00E+00	1.00E+00	1.00E+00
PEAKS047278	SRF	1.75E-36	1.00E+00	1.00E+00	1.00E+00
PEAKS048586	SRF	0.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS048587	SRF	8.88E-178	1.00E+00	1.00E+00	1.00E+00
PEAKS035277	STAT1	5.05E-123	9.91E-08	2.64E-07	5.91E-03
PEAKS035278	STAT1	3.00E-44	8.06E-01	9.39E-01	1.56E-03
PEAKS035279	STAT1	7.15E-184	7.79E-07	1.98E-06	5.43E-05
PEAKS035280	STAT1	2.81E-115	7.38E-01	1.26E-07	5.83E-04
PEAKS035281	STAT1	4.19E-140	7.41E-01	7.52E-01	5.83E-04
PEAKS035347	STAT1	1.19E-43	1.81E-02	5.50E-02	1.96E-04
PEAKS035565	STAT1	2.12E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS037797	STAT1	3.94E-83	1.00E+00	1.00E+00	1.00E+00
PEAKS037799	STAT1	9.35E-24	1.00E+00	1.00E+00	1.00E+00
PEAKS037801	STAT1	1.02E-29	1.00E+00	1.00E+00	1.00E+00
PEAKS037806	STAT1	2.47E-94	6.43E-04	4.88E-03	3.96E-02
PEAKS037807	STAT1	5.75E-21	7.84E-01	4.63E-01	1.10E-01
PEAKS037808	STAT1	7.28E-45	1.00E+00	1.00E+00	1.00E+00
PEAKS040294	STAT1	8.86E-278	1.03E-09	1.95E-01	1.54E-01
PEAKS040295	STAT1	8.23E-299	6.17E-09	1.40E-08	1.81E-02
PEAKS040298	STAT1	2.12E-110	1.05E-03	4.36E-04	2.11E-02
PEAKS040299	STAT1	0.00E+00	7.34E-13	2.09E-08	2.97E-04
PEAKS040300	STAT1	1.12E-300	1.90E-10	4.55E-11	6.87E-02
PEAKS040783	STAT1	2.30E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS040784	STAT1	1.77E-36	1.00E+00	1.00E+00	1.00E+00
PEAKS040785	STAT1	2.21E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS040789	STAT1	3.20E-08	1.00E+00	1.00E+00	1.00E+00
PEAKS040790	STAT1	4.25E-09	1.00E+00	1.00E+00	1.00E+00
PEAKS040791	STAT1	7.08E-04	1.00E+00	1.00E+00	1.00E+00
PEAKS042433	STAT1	1.30E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS048631	STAT1	5.61E-204	4.39E-07	7.54E-08	6.87E-05
PEAKS048632	STAT1	0.00E+00	1.23E-07	4.77E-07	1.43E-05
PEAKS048633	STAT1	6.92E-166	4.84E-04	1.62E-05	2.03E-04
PEAKS040786	STAT2	2.17E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS040787	STAT2	3.56E-20	1.00E+00	1.00E+00	1.00E+00
PEAKS040788	STAT2	1.01E-14	1.00E+00	1.00E+00	1.00E+00
PEAKS040792	STAT2	2.10E-13	1.00E+00	1.00E+00	1.00E+00
PEAKS040793	STAT2	4.10E-14	1.00E+00	1.00E+00	1.00E+00
PEAKS040794	STAT2	1.88E-13	1.00E+00	1.00E+00	1.00E+00
PEAKS057110	STAT2	0.00E+00	3.05E-06	1.17E-05	5.20E-03
PEAKS035900	STAT3	2.06E-49	1.20E-03	4.49E-02	1.79E-01
PEAKS035901	STAT3	5.87E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS035902	STAT3	1.08E-02	1.00E+00	1.00E+00	1.00E+00

PEAKS036035	STAT3	2.59E-12	1.00E+00	1.00E+00	1.00E+00
PEAKS036036	STAT3	1.86E-197	3.91E-10	2.60E-12	2.08E-03
PEAKS036050	STAT3	2.13E-61	2.26E-03	4.61E-03	4.00E-03
PEAKS037798	STAT3	4.08E-26	1.00E+00	1.00E+00	1.00E+00
PEAKS037800	STAT3	2.48E-145	3.99E-08	1.50E-09	3.06E-01
PEAKS037802	STAT3	1.27E-158	1.38E-08	3.00E-12	3.79E-04
PEAKS037803	STAT3	1.29E-19	1.00E+00	1.00E+00	1.00E+00
PEAKS037804	STAT3	2.51E-144	2.13E-09	2.36E-10	2.40E-04
PEAKS037805	STAT3	1.64E-156	9.26E-01	4.65E-12	8.89E-03
PEAKS040256	STAT3	4.57E-155	5.70E-12	4.60E-11	9.48E-04
PEAKS040258	STAT3	3.46E-147	1.30E-12	2.26E-11	2.77E-04
PEAKS040957	STAT3	2.07E-257	3.04E-15	9.31E-12	1.45E-01
PEAKS041094	STAT3	1.67E-142	2.05E-10	8.81E-11	1.50E-05
PEAKS041095	STAT3	3.12E-135	3.82E-09	2.56E-12	2.63E-02
PEAKS041096	STAT3	4.54E-159	5.47E-10	1.09E-11	8.72E-04
PEAKS041097	STAT3	8.74E-144	1.35E-09	3.73E-11	3.60E-01
PEAKS041099	STAT3	7.96E-156	4.99E-10	8.82E-11	1.09E-01
PEAKS041100	STAT3	1.15E-150	2.74E-11	2.13E-10	3.75E-03
PEAKS041101	STAT3	1.19E-139	8.00E-08	4.61E-10	3.08E-03
PEAKS041102	STAT3	3.82E-154	9.57E-01	1.13E-10	3.54E-03
PEAKS042228	STAT3	8.86E-199	1.47E-10	6.51E-14	3.31E-03
PEAKS049835	STAT3	9.73E-278	1.11E-09	6.88E-13	5.00E-02
PEAKS049836	STAT3	9.63E-242	6.08E-11	2.92E-11	6.23E-01
PEAKS049839	STAT3	6.28E-271	1.59E-11	5.66E-12	5.16E-04
PEAKS035301	STAT4	8.74E-224	3.24E-14	8.03E-17	1.06E-05
PEAKS042427	STAT4	1.57E-140	4.20E-11	9.48E-17	4.15E-02
PEAKS042428	STAT4	1.89E-151	9.66E-13	4.41E-16	1.21E-04
PEAKS042429	STAT4	2.08E-95	6.74E-01	3.34E-08	1.64E-02
PEAKS042431	STAT4	2.01E-03	1.00E+00	1.00E+00	1.00E+00
PEAKS035252	STAT5A	3.78E-09	1.00E+00	1.00E+00	1.00E+00
PEAKS035288	STAT5A	4.88E-76	6.69E-01	5.62E-01	1.20E-01
PEAKS035296	STAT5A	0.00E+00	3.84E-11	1.44E-11	4.95E-02
PEAKS035298	STAT5A	0.00E+00	6.11E-11	1.31E-10	6.39E-03
PEAKS035639	STAT5A	0.00E+00	8.63E-11	2.29E-10	1.33E-05
PEAKS035954	STAT5A	2.14E-300	9.77E-11	6.86E-11	1.42E-04
PEAKS035958	STAT5A	0.00E+00	2.26E-10	4.11E-10	6.70E-06
PEAKS036303	STAT5A	0.00E+00	9.51E-10	8.00E-12	7.49E-05
PEAKS036304	STAT5A	0.00E+00	5.48E-09	6.00E-10	4.05E-07
PEAKS036398	STAT5A	0.00E+00	9.66E-10	2.88E-09	6.93E-06
PEAKS036399	STAT5A	0.00E+00	3.10E-08	9.81E-09	2.55E-06
PEAKS038238	STAT5A	0.00E+00	1.67E-08	5.08E-09	2.08E-06
PEAKS038239	STAT5A	0.00E+00	2.98E-08	1.55E-09	3.44E-08
PEAKS038878	STAT5A	0.00E+00	1.34E-09	8.21E-10	9.63E-05
PEAKS038881	STAT5A	1.44E-168	2.33E-10	4.22E-10	4.62E-03
PEAKS039935	STAT5A	1.85E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS039936	STAT5A	0.00E+00	9.31E-10	4.39E-11	6.73E-01

PEAKS042584	STAT5A	8.32E-15	1.00E+00	1.00E+00	1.00E+00
PEAKS035299	STAT5B	0.00E+00	6.64E-10	3.87E-10	6.14E-07
PEAKS035640	STAT5B	0.00E+00	1.52E-08	6.04E-11	3.62E-08
PEAKS035642	STAT5B	0.00E+00	3.83E-09	3.24E-09	1.75E-10
PEAKS035955	STAT5B	7.85E-300	2.32E-10	3.22E-10	1.57E-05
PEAKS035959	STAT5B	0.00E+00	9.24E-11	2.62E-10	2.29E-04
PEAKS042198	STAT5B	0.00E+00	8.74E-09	4.52E-10	1.37E-01
PEAKS042200	STAT5B	0.00E+00	4.02E-09	4.38E-09	3.71E-02
PEAKS042230	STAT5B	6.47E-301	3.33E-09	9.02E-10	2.39E-03
PEAKS042231	STAT5B	4.82e-320	6.68E-11	3.51E-10	9.26E-03
PEAKS042232	STAT5B	0.00E+00	6.33E-10	5.44E-10	2.64E-05
PEAKS042233	STAT5B	1.95E-275	9.15E-09	1.27E-09	1.87E-03
PEAKS042234	STAT5B	0.00E+00	1.49E-09	5.14E-10	6.01E-05
PEAKS035211	STAT6	6.48E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS035212	STAT6	9.24E-281	1.79E-11	9.98E-11	7.19E-01
PEAKS040296	STAT6	0.00E+00	2.04E-12	5.74E-10	1.80E-04
PEAKS040297	STAT6	0.00E+00	3.32E-10	9.31E-11	3.80E-05
PEAKS042487	STAT6	8.91E-294	1.24E-02	1.21E-13	6.71E-04
PEAKS042488	STAT6	2.08E-291	1.19E-13	5.28E-10	2.10E-04
PEAKS042489	STAT6	2.39E-264	2.30E-13	8.54E-03	1.71E-05
PEAKS057572	STAT6	0.00E+00	2.40E-12	1.27E-09	1.95E-04
PEAKS035259	TAL1	5.20E-277	1.00E+00	1.00E+00	1.00E+00
PEAKS035330	TAL1	0.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035499	TAL1	0.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS035513	TAL1	1.41E-178	1.00E+00	1.00E+00	1.00E+00
PEAKS036143	TAL1	0.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS037158	TBP	2.11E-113	1.00E+00	1.00E+00	1.00E+00
PEAKS037159	TBP	2.38E-96	1.00E+00	1.00E+00	1.00E+00
PEAKS050777	TBP	5.86E-45	1.00E+00	1.00E+00	1.00E+00
PEAKS050779	TBP	2.11E-39	1.00E+00	1.00E+00	1.00E+00
PEAKS049348	TBR1	8.71E-68	8.31E-01	7.84E-01	5.12E-01
PEAKS049349	TBR1	2.91E-66	7.98E-01	8.16E-01	2.38E-01
PEAKS036073	TBX21	2.30E-105	1.99E-01	2.96E-01	3.30E-03
PEAKS040795	TBX21	4.17E-54	2.27E-01	4.93E-01	5.04E-02
PEAKS040796	TBX21	1.56E-90	9.97E-08	1.30E-10	8.43E-02
PEAKS059303	TBX21	1.09E-40	1.70E-01	2.91E-01	7.12E-01
PEAKS035802	TBX3	5.34E-09	2.97E-04	5.20E-02	2.40E-01
PEAKS057470	TBX5	3.52E-264	5.19E-07	2.11E-07	1.25E-01
PEAKS057471	TBX5	1.73E-172	2.26E-07	2.99E-07	3.30E-07
PEAKS057472	TBX5	2.62E-118	5.38E-06	1.24E-06	3.67E-04
PEAKS057473	TBX5	3.93E-153	8.13E-07	2.19E-06	1.85E-03
PEAKS057476	TBX5	1.96E-118	9.07E-07	2.33E-07	4.13E-05
PEAKS035469	TCF12	0.00E+00	1.34E-06	2.53E-06	1.24E-06
PEAKS036745	TCF3	0.00E+00	3.05E-08	1.82E-08	4.25E-03
PEAKS038888	TCF3	1.86E-170	2.44E-04	1.25E-04	2.78E-06
PEAKS038889	TCF3	7.80E-177	2.56E-07	9.59E-08	5.76E-08

PEAKS038890	TCF3	4.19E-147	5.11E-05	2.53E-06	3.81E-04
PEAKS038891	TCF3	1.40E-168	1.50E-05	1.16E-05	1.33E-06
PEAKS038892	TCF3	1.15E-181	1.23E-07	4.94E-05	2.63E-01
PEAKS038893	TCF3	1.03E-178	2.49E-06	1.26E-07	5.63E-08
PEAKS038894	TCF3	3.31E-164	3.20E-03	9.51E-01	3.54E-04
PEAKS038895	TCF3	1.72E-173	2.35E-05	3.84E-06	1.71E-05
PEAKS038896	TCF3	2.97E-89	4.62E-01	3.48E-01	5.13E-01
PEAKS038897	TCF3	4.61E-141	5.51E-05	3.49E-06	5.04E-06
PEAKS039238	TCF3	1.68E-280	8.72E-05	9.17E-06	8.90E-05
PEAKS039239	TCF3	8.20E-281	1.91E-06	1.07E-05	3.33E-03
PEAKS049613	TCF3	1.24E-280	4.39E-05	1.54E-04	3.31E-04
PEAKS049614	TCF3	4.48E-184	2.79E-05	1.60E-05	2.50E-07
PEAKS036253	TCF7	9.27E-142	9.12E-14	2.15E-13	3.56E-04
PEAKS036254	TCF7	7.83E-150	2.94E-11	4.98E-13	1.91E-05
PEAKS035805	TCF7L2	2.80E-217	1.40E-14	1.08E-14	2.69E-03
PEAKS037791	TCF7L2	6.64e-310	6.04E-13	5.12E-13	7.12E-01
PEAKS039250	TEAD1	5.31E-306	2.26E-08	1.41E-10	1.23E-04
PEAKS039251	TEAD1	0.00E+00	5.60E-11	1.11E-10	3.52E-01
PEAKS049506	TEAD4	0.00E+00	3.98E-11	4.01E-11	5.04E-01
PEAKS049507	TEAD4	8.50E-283	1.72E-09	3.02E-10	5.80E-01
PEAKS049510	TEAD4	0.00E+00	4.93E-11	8.21E-11	4.30E-01
PEAKS049511	TEAD4	0.00E+00	1.37E-08	7.66E-11	2.70E-01
PEAKS055535	TET2	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036276	TFAP2A	3.31E-230	2.56E-07	4.39E-08	2.70E-03
PEAKS036280	TFAP2A	2.90E-133	2.11E-10	1.54E-08	4.88E-03
PEAKS036290	TFAP2A	2.83E-276	1.46E-06	6.15E-09	9.87E-05
PEAKS036291	TFAP2A	2.17E-243	1.53E-08	1.66E-09	8.66E-04
PEAKS036665	TFAP4	7.90E-183	7.93E-06	5.90E-06	3.39E-03
PEAKS055119	THRA	6.08E-260	1.41E-07	9.50E-06	7.07E-02
PEAKS038329	TP53	9.01E-172	5.18E-13	1.65E-13	5.24E-01
PEAKS038330	TP53	0.00E+00	6.45E-14	4.79E-13	2.28E-03
PEAKS039947	TP53	4.92E-56	1.00E+00	1.00E+00	1.00E+00
PEAKS039948	TP53	4.14E-05	1.00E+00	1.00E+00	1.00E+00
PEAKS039949	TP53	1.83E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS039950	TP53	3.39E-07	1.00E+00	1.00E+00	1.00E+00
PEAKS039951	TP53	0.00E+00	1.26E-08	2.57E-12	2.56E-01
PEAKS039952	TP53	0.00E+00	2.66E-13	1.82E-13	3.85E-03
PEAKS039953	TP53	1.97E-236	4.98E-12	1.01E-09	8.32E-01
PEAKS039954	TP53	1.93E-52	1.00E+00	1.00E+00	1.00E+00
PEAKS039955	TP53	0.00E+00	7.91E-14	1.72E-08	6.77E-01
PEAKS039956	TP53	0.00E+00	4.57E-14	2.51E-14	5.94E-01
PEAKS039958	TP53	0.00E+00	1.10E-13	2.03E-08	4.01E-01
PEAKS039959	TP53	0.00E+00	4.47E-09	5.74E-10	1.62E-01
PEAKS039960	TP53	0.00E+00	1.63E-08	5.66E-08	4.50E-01
PEAKS039961	TP53	0.00E+00	2.05E-08	1.94E-09	5.04E-04
PEAKS039962	TP53	0.00E+00	5.41E-08	5.14E-09	8.88E-01

PEAKS039963	TP53	0.00E+00	1.19E-12	1.40E-08	9.85E-01
PEAKS039964	TP53	0.00E+00	7.87E-08	7.61E-14	1.19E-01
PEAKS039965	TP53	0.00E+00	1.36E-13	1.61E-08	9.01E-01
PEAKS039966	TP53	0.00E+00	1.27E-13	1.06E-04	8.88E-01
PEAKS039967	TP53	0.00E+00	5.35E-12	2.81E-12	9.53E-01
PEAKS039968	TP53	0.00E+00	2.04E-14	3.32E-14	5.51E-01
PEAKS039969	TP53	0.00E+00	2.45E-13	3.97E-14	1.19E-02
PEAKS039970	TP53	0.00E+00	8.38E-12	2.68E-12	3.42E-05
PEAKS039971	TP53	0.00E+00	2.82E-13	1.74E-13	7.47E-01
PEAKS039972	TP53	0.00E+00	7.38E-13	3.69E-13	3.42E-03
PEAKS042188	TP53	2.01E-184	3.49E-01	3.61E-01	6.54E-01
PEAKS042189	TP53	6.08e-320	1.33E-08	1.14E-09	1.40E-03
PEAKS042190	TP53	0.00E+00	4.82E-13	2.41E-09	9.36E-05
PEAKS042191	TP53	0.00E+00	1.08E-09	1.54E-12	3.43E-01
PEAKS055154	TWIST2	3.70E-31	3.07E-03	1.03E-02	1.22E-02
PEAKS055155	TWIST2	1.26E-05	4.98E-02	4.59E-01	1.33E-03
PEAKS055158	TWIST2	1.70E-253	7.53E-09	4.12E-08	9.03E-07
PEAKS055160	TWIST2	7.60E-233	4.87E-06	2.22E-07	7.40E-07
PEAKS035508	USF1	0.00E+00	2.38E-12	5.34E-12	6.83E-11
PEAKS049126	USF2	0.00E+00	1.47E-10	1.34E-11	5.93E-04
PEAKS049127	USF2	0.00E+00	1.35E-09	2.10E-09	2.30E-03
PEAKS038136	VDR	1.88E-227	8.31E-11	9.39E-10	5.36E-01
PEAKS038137	VDR	1.79E-271	4.07E-09	1.86E-08	1.81E-01
PEAKS039017	VDR	9.97E-18	1.00E+00	1.00E+00	1.00E+00
PEAKS039018	VDR	4.38E-204	2.59E-11	1.28E-11	8.44E-01
PEAKS039026	VDR	1.20E-176	3.97E-10	3.60E-10	8.17E-01
PEAKS039033	VDR	8.07E-06	1.00E+00	1.00E+00	1.00E+00
PEAKS039034	VDR	5.02E-173	6.42E-07	1.01E-10	1.20E-01
PEAKS039040	VDR	6.50E-128	1.64E-11	6.51E-12	7.95E-01
PEAKS050751	VDR	8.22E-22	1.00E+00	1.00E+00	1.00E+00
PEAKS050752	VDR	1.71E-177	1.02E-02	1.01E-11	3.25E-03
PEAKS039190	VSX2	0.00E+00	1.70E-04	2.33E-05	7.49E-03
PEAKS038833	WIZ	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS037674	WT1	1.55E-217	5.85E-04	1.03E-05	6.97E-05
PEAKS049625	WT1	1.24E-222	9.96E-12	5.96E-12	5.32E-05
PEAKS040238	XBP1	4.26E-83	1.00E+00	1.00E+00	1.00E+00
PEAKS040240	XBP1	3.04E-140	1.00E+00	1.00E+00	1.00E+00
PEAKS040241	XBP1	1.64E-138	1.00E+00	1.00E+00	1.00E+00
PEAKS040242	XBP1	1.25E-106	1.00E+00	1.00E+00	1.00E+00
PEAKS040244	XBP1	2.53E-154	1.00E+00	1.00E+00	1.00E+00
PEAKS040245	XBP1	9.83E-95	1.00E+00	1.00E+00	1.00E+00
PEAKS036085	YY1	0.00E+00	9.39E-10	6.74E-10	4.77E-03
PEAKS038524	YY1	0.00E+00	1.61E-08	4.74E-12	2.82E-03
PEAKS038525	YY1	0.00E+00	6.86E-11	5.69E-11	1.83E-02
PEAKS038526	YY1	0.00E+00	9.29E-09	2.99E-09	3.43E-04
PEAKS038527	YY1	0.00E+00	3.21E-11	3.18E-11	2.48E-02

PEAKS038513	ZBTB16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038514	ZBTB16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS038515	ZBTB16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039142	ZBTB16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS039143	ZBTB16	1.00E+00	1.00E+00	1.00E+00	1.00E+00
PEAKS036489	ZBTB17	1.00E+00	1.14E-04	2.77E-04	5.96E-02
PEAKS036490	ZBTB17	1.00E+00	6.43E-05	1.26E-05	6.76E-02
PEAKS059294	ZBTB7B	5.73E-51	7.47E-03	1.26E-04	1.18E-03
PEAKS039167	ZFAT	1.00E+00	1.08E-01	1.52E-02	5.38E-03
PEAKS037378	ZIC1	7.46E-58	2.55E-03	1.01E-01	2.85E-04
PEAKS037379	ZIC1	3.45E-111	1.92E-04	3.85E-02	7.56E-03
PEAKS042519	ZNF652	7.37E-28	1.15E-01	3.39E-03	1.41E-02
PEAKS042520	ZNF652	1.00E-28	1.75E-02	1.78E-02	6.64E-03
PEAKS042521	ZNF652	5.09E-50	4.35E-02	1.85E-02	3.38E-03
PEAKS042522	ZNF652	1.69E-34	1.78E-02	6.95E-03	1.27E-02

Примечание. ID – уникальный идентификатор базы данных GTRD; AME – результат обогащения частотной матрицы целевого ТФ; TomTom – результат сравнения частотных матриц, полученных с помощью *de novo*, с частотными матрицами целевых ТФ с помощью программы TomTom.

Приложение Б

Таблица. Значимо обогащенные термины ГО для списков генов, соответствующих пикам (Peaks) CCA1, и ССТФ, предсказанных в пределах пиков с помощью моделей PWM, BaMM и SiteGA.

Модель	ID	Описание	Обогащение	p_adj	Общий термин
PWM	GO:0009409	response to cold	3.51	8.57E-17	*
PWM	GO:0007623	circadian rhythm	5.29	2.43E-15	*
PWM	GO:0048511	rhythmic process	5.29	2.43E-15	*
PWM	GO:0009414	response to water deprivation	3.21	5.21E-14	*
PWM	GO:0009415	response to water	3.13	1.41E-13	*
PWM	GO:0001101	response to acid chemical	2.91	4.37E-12	*
PWM	GO:0001666	response to hypoxia	3.19	6.89E-09	*
PWM	GO:0036293	response to decreased oxygen levels	3.14	9.79E-09	*
PWM	GO:0070482	response to oxygen levels	3.13	9.81E-09	*
PWM	GO:0044247	cellular polysaccharide catabolic process	11.19	7.89E-08	
PWM	GO:0071456	cellular response to hypoxia	3.03	4.89E-07	*
PWM	GO:0036294	cellular response to decreased oxygen levels	3.00	5.18E-07	*
PWM	GO:0071453	cellular response to oxygen levels	3.00	5.18E-07	*
PWM	GO:0005983	starch catabolic process	10.77	3.62E-06	
PWM	GO:0009251	glucan catabolic process	7.99	4.68E-06	
PWM	GO:0009642	response to light intensity	3.34	2.17E-05	
PWM	GO:0044275	cellular carbohydrate catabolic process	5.75	2.20E-05	
PWM	GO:0009651	response to salt stress	2.15	2.53E-05	
PWM	GO:0006979	response to oxidative stress	2.17	4.27E-05	
PWM	GO:0009631	cold acclimation	4.83	6.78E-05	*
PWM	GO:0009408	response to heat	2.53	1.19E-04	
PWM	GO:0048580	regulation of post-embryonic development	2.14	7.90E-04	*
PWM	GO:0051253	negative regulation of RNA metabolic process	2.46	8.84E-04	
PWM	GO:0005982	starch metabolic process	3.80	1.14E-03	
PWM	GO:0071482	cellular response to light stimulus	3.23	1.31E-03	
PWM	GO:2000026	regulation of multicellular organismal development	2.06	1.58E-03	*
PWM	GO:0009639	response to red or far red light	2.38	1.91E-03	*
PWM	GO:0045934	negative regulation of nucleobase-containing compound metabolic process	2.27	2.28E-03	
PWM	GO:0071478	cellular response to radiation	3.03	2.53E-03	
PWM	GO:0071214	cellular response to abiotic stimulus	2.47	2.53E-03	
PWM	GO:0104004	cellular response to environmental stimulus	2.47	2.53E-03	

PWM	GO:0009644	response to high light intensity	3.87	2.81E-03	
PWM	GO:0010017	red or far-red light signaling pathway	4.14	2.92E-03	
PWM	GO:2000030	regulation of response to red or far red light	4.42	3.45E-03	
PWM	GO:0071489	cellular response to red or far red light	4.00	3.92E-03	
PWM	GO:0044262	cellular carbohydrate metabolic process	1.87	6.03E-03	
PWM	GO:0009637	response to blue light	3.16	6.03E-03	*
PWM	GO:0051239	regulation of multicellular organismal process	1.83	6.12E-03	*
PWM	GO:0042752	regulation of circadian rhythm	3.73	6.77E-03	
PWM	GO:0006091	generation of precursor metabolites and energy	1.88	8.66E-03	
PWM	GO:0051606	detection of stimulus	3.61	8.73E-03	
PWM	GO:0045892	negative regulation of transcription, DNA-templated	2.23	9.12E-03	
PWM	GO:1902679	negative regulation of RNA biosynthetic process	2.21	9.99E-03	
PWM	GO:1903507	negative regulation of nucleic acid-templated transcription	2.21	9.99E-03	
PWM	GO:0009645	response to low light intensity stimulus	6.10	1.21E-02	
PWM	GO:0015979	photosynthesis	2.17	1.25E-02	
PWM	GO:0019725	cellular homeostasis	2.08	1.34E-02	
PWM	GO:0048584	positive regulation of response to stimulus	2.05	1.67E-02	
PWM	GO:0031324	negative regulation of cellular metabolic process	1.78	1.72E-02	
PWM	GO:0010286	heat acclimation	3.45	2.03E-02	
PWM	GO:0010119	regulation of stomatal movement	2.74	2.04E-02	
PWM	GO:0051172	negative regulation of nitrogen compound metabolic process	1.80	2.49E-02	
PWM	GO:0009890	negative regulation of biosynthetic process	1.85	2.63E-02	
PWM	GO:0010200	response to chitin	4.45	2.63E-02	
PWM	GO:0010600	regulation of auxin biosynthetic process	6.36	2.63E-02	
PWM	GO:0055065	metal ion homeostasis	2.15	2.65E-02	
PWM	GO:0010224	response to UV-B	3.07	2.65E-02	
PWM	GO:2000113	negative regulation of cellular macromolecule biosynthetic process	1.88	2.65E-02	
PWM	GO:0009909	regulation of flower development	2.29	2.78E-02	
PWM	GO:0010118	stomatal movement	2.29	2.78E-02	
PWM	GO:0010558	negative regulation of macromolecule biosynthetic process	1.87	2.78E-02	
PWM	GO:0009611	response to wounding	2.01	2.80E-02	
PWM	GO:0031327	negative regulation of cellular biosynthetic process	1.82	3.43E-02	

PWM	GO:0009787	regulation of abscisic acid-activated signaling pathway	2.44	3.43E-02	
PWM	GO:1901419	regulation of response to alcohol	2.44	3.43E-02	
PWM	GO:1905957	regulation of cellular response to alcohol	2.44	3.43E-02	
PWM	GO:0050801	ion homeostasis	1.87	3.67E-02	
PWM	GO:0010114	response to red light	3.08	3.67E-02	*
PWM	GO:0006006	glucose metabolic process	3.62	3.76E-02	
PWM	GO:0006073	cellular glucan metabolic process	2.02	3.80E-02	
PWM	GO:0009646	response to absence of light	3.54	4.24E-02	
PWM	GO:0051241	negative regulation of multicellular organismal process	2.37	4.32E-02	
PWM	GO:0098771	inorganic ion homeostasis	1.89	4.45E-02	
PWM	GO:0014070	response to organic cyclic compound	1.81	4.57E-02	
PWM	GO:0009743	response to carbohydrate	2.20	4.57E-02	
PWM	GO:0015980	energy derivation by oxidation of organic compounds	2.20	4.57E-02	
PWM	GO:0044042	glucan metabolic process	1.96	4.84E-02	
PWM	GO:0048831	regulation of shoot system development	2.13	4.88E-02	
PWM	GO:0048366	leaf development	1.69	4.97E-02	*
BaMM	GO:0007623	circadian rhythm	5.43	5.41E-20	*
BaMM	GO:0048511	rhythmic process	5.43	5.41E-20	*
BaMM	GO:0009409	response to cold	3.27	9.98E-18	*
BaMM	GO:0009414	response to water deprivation	2.93	1.89E-13	*
BaMM	GO:0009415	response to water	2.86	5.50E-13	*
BaMM	GO:0001101	response to acid chemical	2.69	6.65E-12	*
BaMM	GO:0044247	cellular polysaccharide catabolic process	11.04	1.42E-09	
BaMM	GO:0001666	response to hypoxia	2.98	3.76E-09	*
BaMM	GO:0036293	response to decreased oxygen levels	2.93	5.69E-09	*
BaMM	GO:0070482	response to oxygen levels	2.92	5.84E-09	*
BaMM	GO:0009251	glucan catabolic process	8.49	1.85E-08	
BaMM	GO:0005983	starch catabolic process	10.99	3.89E-08	
BaMM	GO:0071456	cellular response to hypoxia	2.95	3.89E-08	*
BaMM	GO:0036294	cellular response to decreased oxygen levels	2.92	4.39E-08	*
BaMM	GO:0071453	cellular response to oxygen levels	2.92	4.39E-08	*
BaMM	GO:0005982	starch metabolic process	4.30	5.39E-06	
BaMM	GO:0048580	regulation of post-embryonic development	2.30	5.39E-06	*
BaMM	GO:2000026	regulation of multicellular organismal development	2.21	1.67E-05	*
BaMM	GO:0009631	cold acclimation	4.60	1.94E-05	*
BaMM	GO:0044275	cellular carbohydrate catabolic process	5.17	2.23E-05	
BaMM	GO:0042752	regulation of circadian rhythm	4.53	2.27E-05	
BaMM	GO:0051239	regulation of multicellular organismal process	2.00	6.88E-05	*

BaMM	GO:0006979	response to oxidative stress	2.00	1.11E-04	
BaMM	GO:0009642	response to light intensity	2.91	1.11E-04	
BaMM	GO:0009651	response to salt stress	1.86	7.68E-04	
BaMM	GO:0009639	response to red or far red light	2.28	1.19E-03	*
BaMM	GO:0009637	response to blue light	3.21	1.26E-03	*
BaMM	GO:0048366	leaf development	1.90	1.45E-03	*
BaMM	GO:0009408	response to heat	2.18	1.45E-03	
BaMM	GO:0010114	response to red light	3.60	1.59E-03	*
BaMM	GO:0006091	generation of precursor metabolites and energy	1.91	1.64E-03	
BaMM	GO:0044262	cellular carbohydrate metabolic process	1.83	2.93E-03	
BaMM	GO:0009644	response to high light intensity	3.50	3.79E-03	
BaMM	GO:0051241	negative regulation of multicellular organismal process	2.63	3.79E-03	
BaMM	GO:0071396	cellular response to lipid	1.75	5.83E-03	
BaMM	GO:0051253	negative regulation of RNA metabolic process	2.12	5.83E-03	
BaMM	GO:0010119	regulation of stomatal movement	2.78	5.93E-03	
BaMM	GO:0006073	cellular glucan metabolic process	2.14	5.93E-03	
BaMM	GO:0009787	regulation of abscisic acid-activated signaling pathway	2.58	5.93E-03	
BaMM	GO:1901419	regulation of response to alcohol	2.58	5.93E-03	
BaMM	GO:1905957	regulation of cellular response to alcohol	2.58	5.93E-03	
BaMM	GO:0005977	glycogen metabolic process	6.79	5.93E-03	
BaMM	GO:0006112	energy reserve metabolic process	6.79	5.93E-03	
BaMM	GO:0015979	photosynthesis	2.11	6.72E-03	
BaMM	GO:0010118	stomatal movement	2.33	7.86E-03	
BaMM	GO:0044042	glucan metabolic process	2.09	8.05E-03	
BaMM	GO:0009739	response to gibberellin	2.55	9.55E-03	
BaMM	GO:2000030	regulation of response to red or far red light	3.69	1.15E-02	
BaMM	GO:0071215	cellular response to abscisic acid stimulus	2.00	1.16E-02	
BaMM	GO:0097306	cellular response to alcohol	2.00	1.16E-02	
BaMM	GO:0009311	oligosaccharide metabolic process	2.86	1.25E-02	
BaMM	GO:0045934	negative regulation of nucleobase-containing compound metabolic process	1.96	1.27E-02	
BaMM	GO:0009269	response to desiccation	4.95	1.27E-02	
BaMM	GO:0009312	oligosaccharide biosynthetic process	3.92	1.28E-02	
BaMM	GO:0015980	energy derivation by oxidation of organic compounds	2.27	1.31E-02	
BaMM	GO:0046351	disaccharide biosynthetic process	4.24	1.46E-02	
BaMM	GO:0045892	negative regulation of transcription, DNA-templated	2.02	1.82E-02	
BaMM	GO:0009611	response to wounding	1.96	1.82E-02	
BaMM	GO:1902679	negative regulation of RNA biosynthetic process	2.00	2.00E-02	

BaMM	GO:1903507	negative regulation of nucleic acid-templated transcription	2.00	2.00E-02	
BaMM	GO:0031324	negative regulation of cellular metabolic process	1.68	2.05E-02	
BaMM	GO:0000272	polysaccharide catabolic process	2.26	2.35E-02	
BaMM	GO:0005984	disaccharide metabolic process	2.87	2.60E-02	
BaMM	GO:2000113	negative regulation of cellular macromolecule biosynthetic process	1.79	2.72E-02	
BaMM	GO:0009743	response to carbohydrate	2.16	2.82E-02	
BaMM	GO:0048581	negative regulation of post-embryonic development	2.40	2.82E-02	
BaMM	GO:0010558	negative regulation of macromolecule biosynthetic process	1.78	2.83E-02	
BaMM	GO:0009890	negative regulation of biosynthetic process	1.74	2.83E-02	
BaMM	GO:0009785	blue light signaling pathway	4.85	2.83E-02	
BaMM	GO:0030522	intracellular receptor signaling pathway	4.85	2.83E-02	
BaMM	GO:0051017	actin filament bundle assembly	4.85	2.83E-02	
BaMM	GO:0061572	actin filament bundle organization	4.85	2.83E-02	
BaMM	GO:0009909	regulation of flower development	2.12	3.18E-02	
BaMM	GO:0042542	response to hydrogen peroxide	2.75	3.23E-02	
BaMM	GO:0009646	response to absence of light	3.32	3.30E-02	
BaMM	GO:0016052	carbohydrate catabolic process	1.89	3.45E-02	
BaMM	GO:0031327	negative regulation of cellular biosynthetic process	1.72	3.67E-02	
BaMM	GO:0055065	metal ion homeostasis	1.96	3.90E-02	
BaMM	GO:0009694	jasmonic acid metabolic process	2.98	4.06E-02	
BaMM	GO:2000377	regulation of reactive oxygen species metabolic process	2.98	4.06E-02	
BaMM	GO:0010117	photoprotection	6.79	4.57E-02	
SiteGA	GO:0007623	circadian rhythm	7.03	2.33E-06	*
SiteGA	GO:0048511	rhythmic process	7.03	2.33E-06	*
SiteGA	GO:0001666	response to hypoxia	4.72	1.90E-05	*
SiteGA	GO:0036293	response to decreased oxygen levels	4.65	1.90E-05	*
SiteGA	GO:0070482	response to oxygen levels	4.64	1.90E-05	*
SiteGA	GO:0048580	regulation of post-embryonic development	3.30	3.12E-03	*
SiteGA	GO:0009409	response to cold	3.16	3.12E-03	*
SiteGA	GO:2000026	regulation of multicellular organismal development	3.18	4.38E-03	*
SiteGA	GO:0047484	regulation of response to osmotic stress	9.59	4.51E-03	
SiteGA	GO:0071456	cellular response to hypoxia	3.78	4.87E-03	*
SiteGA	GO:0036294	cellular response to decreased oxygen levels	3.75	4.87E-03	*
SiteGA	GO:0071453	cellular response to oxygen levels	3.75	4.87E-03	*
SiteGA	GO:0009639	response to red or far red light	3.65	1.11E-02	*

SiteGA	GO:0009631	cold acclimation	7.15	1.55E-02	*
SiteGA	GO:0090351	seedling development	3.70	1.63E-02	
SiteGA	GO:0009414	response to water deprivation	2.77	1.72E-02	*
SiteGA	GO:0001101	response to acid chemical	2.67	1.72E-02	*
SiteGA	GO:0009845	seed germination	3.86	1.81E-02	
SiteGA	GO:0009637	response to blue light	5.47	1.81E-02	*
SiteGA	GO:0009415	response to water	2.70	1.83E-02	*
SiteGA	GO:0009765	photosynthesis, light harvesting	8.18	1.83E-02	
SiteGA	GO:0010114	response to red light	6.39	1.83E-02	*
SiteGA	GO:0051239	regulation of multicellular organismal process	2.56	2.05E-02	*
SiteGA	GO:0051094	positive regulation of developmental process	4.11	3.75E-02	
SiteGA	GO:0048366	leaf development	2.57	3.88E-02	*