Автономная некоммерческая образовательная организация высшего образования
«Научно-технологический университет «Сириус»

На правах рукописи

Колмыков Семён Константинович

Разработка методов контроля качества и построения карты геномных районов связывания транскрипционных факторов на основе сравнительного анализа ChIP-seq экспериментов

1.5.8. Математическая биология, биоинформатика

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
доктор биологических наук
Колпаков Федор Анатольевич

Федеральная территория «Сириус»
2024

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ

ADASTRA - База данных по аллель-специфичным сайтам связывания транскрипционных факторов человека (Allelic Dosage-corrected Allele-Specific human TRAnscription factor binding sites)

ASB - аллель-специфичное связывание ТФ (Allele-specific binding)

AUC - Площадь под кривой (Area Under the Curve)

ChIP-seq - Иммунопреципитация хроматина с последующим глубоким секвенированием (Chromatin ImmunoPrecipitation followed by massively parallel/deep Sequencing)

DNase-seq - Определение чувствительности к эндонуклеазе ДНКаза I на основе секвенирования нового поколения (Deoxyribonuclease I Sequencing)

ENCODE - Энциклопедия элементов ДНК (Encyclopedia of DNA Elements). Международный проект по идентификации функциональных регуляторных элементов.

eQTL - Локус количественного признака экспрессии; геномный локус, от генотипа которого зависит уровень экспрессии гена-мишени (Expression quantitative trait loci)

FN - ложно свидетельствующий об отрицательном результате (False Negative)

FNCM - Метод оценки доли ложно неидентифицированных РСТФ (False Negative Control Metric)

FP - ложно свидетельствующий о положительном результате (False Positive)

FPCM - Метод оценки доли ложно идентифицированных РСТФ (False Positive Control Metric)

FRiP - Доля прочтений в ChIP-seq пиках (Fraction of Reads in Peaks)

GTRD - База данных по регуляции транскрипции (Gene Transcription Regulation Database)

HOCOMOCO - Коллекция мотивов для сайтов связывания транскрипционных факторов человека и мыши (Homo sapiens Comprehensive Model Collection)

NGS - технологии секвенирования нового поколения (Next Generation Sequencing), или технологии массового параллельного секвенирования

NRF - Доля неизбыточных прочтений (Non-Redundant Fraction)

PBC1 - Коэффициент ограничения ПЦР 1 (PCR Bottlenecking Coefficient 1)

PBC2 - Коэффициент ограничения ПЦР 2 (PCR Bottlenecking Coefficient 2)

PWM - Позиционно-весовая матрица (Position Weight Matrix)

RBP - РНК-связывающие белки

ROC - Операционная характеристика приемника (Receiver Operating Characteristic)

SNV - Однонуклеотидный геномный вариант (Single Nucleotide Variant)

UMI - Уникальный молекулярный идентификатор (Unique Molecular Identifier)

БД - база данных

МКВ - метод коллективного выбора

ММАЖ - Множественные морфологические аномалии жгутиков сперматозоидов

ВВЕДЕНИЕ

Актуальность темы исследования

Регуляция транскрипции осуществляется на разных уровнях при помощи разных механизмов (структура хроматина, метилирование ДНК, модификации гистонов и другие), однако именно транскрипционные факторы (ТФ) и их сайты связывания являются основными компонентами регуляции транскрипции.

Основным методом массового экспериментального определения районов связывания транскрипционных факторов (РСТФ) является метод ChIP-seq. В рамках данного метода из клетки выделяют ДНК и фрагментируют на небольшие нуклеотидные последовательности, затем проводят иммунопреципитацию, используя антитела к соответствующему ТФ. В результате с антителами связываются комплексы, состоящие из исследуемого ТФ и фрагмента ДНК. Для анализа нуклеотидной последовательности данных фрагментов ДНК используются методы массового параллельного секвенирования (NGS). Затем, проанализированные фрагменты ДНК картируются на референсный геном и при помощи различных алгоритмов определяются районы с большим количеством таких картированных фрагментов – РСТФ.

Для метода ChIP-seq характерен высокий уровень шума, что привело к созданию различных алгоритмов (MACS2, GEM, SISSR, PICS и другие, для обзора см. Jeon et al., 2020, Thomas et al., 2017), которые дают существенно разные результаты при обработке результатов одного и того же эксперимента. На данный момент не существует "золотого стандарта" для валидации правильности определения РСТФ. Для косвенной оценки качества построенного набора РСТФ можно использовать частоту наличия в них известных мотивов для заданного ТФ и степень пересечения РСТФ с районами открытого хроматина, которые могут быть определены при помощи методов: DNase-seq и ATAC-seq. Таким образом,

актуальной является задача разработки методов оценки доли ложно идентифицированных и ложно неидентифицированных РСТФ для заданного ChIP-seq эксперимента на основании сравнения результатов нескольких алгоритмов идентификации РСТФ.

База данных GTRD - Gene Transcription Regulation Database (Kolmykov et al., 2021) является крупнейшей в мире базой данных по регуляции транскрипции. В ней хранятся однообразно аннотированные и обработанные результаты десятков тысяч экспериментов по регуляции транскрипции, большинство из которых составляют ChIP-seq, DNase-seq и ATAC-seq эксперименты. Важной особенностью базы данных GTRD является использование онтологий клеточных типов и экспериментальных условий, что позволяет выделить группы экспериментов, проведенных в одинаковых условиях. Поэтому актуальной является задача разработки алгоритма определения наиболее достоверных РСТФ на основе мета-анализа сходных ChIP-seq экспериментов для заданного ТФ.

В последние несколько десятилетий в различных регионах мира наблюдается снижение мужского репродуктивного потенциала, что выражается в уменьшении концентрации сперматозоидов в эякуляте, доли подвижных и морфологически нормальных сперматозоидов, в увеличении доли мужского фактора в бесплодных парах и росте врожденных аномалий мужской репродуктивной системы, приводящих к бесплодию. Качество семенной жидкости является важным компонентом репродуктивного мужского здоровья. Современные молекулярно-генетические подходы, в первую очередь, секвенирование нового поколения (NGS), значительно расширяют возможности исследования генома: выявления значимых ассоциаций между фенотипическими и молекулярно-генетическими маркерами и идентификации новых генов, вовлеченных в контроль мужской фертильности. Большинство известных однонуклеотидных геномных вариантов (SNV) расположено в регуляторных областях генов и могут влиять на эффективность связывания существующих ТФ.

Один из актуальных подходов для идентификации пар SNV-ТФ является анализ аллель-специфичного связывания по данным ChIP-seq экспериментов. Такая информация представлена в базе данных ADASTRA - Allelic Dosage-corrected Allele-Specific human TRAnscription factor binding sites (Abramov et al., 2021), которая построена на основе информации из базы данных GTRD. Таким образом, приобретает актуальность интерпретации SNV, ассоциированных с нарушениями сперматогенеза, с точки зрения регуляции транскрипции.

Степень разработанности темы

Существует набор широко апробированных методов для оценки качества ChIP-seq экспериментов, предложенных в рамках проекта ENCODE. Однако основная часть разработанных характеристик качества направлена на контроль ложно предсказанных районов связывания транскрипционных факторов (РСТФ). В 2022 году Suryatenggara с соавт. была опубликована статья, посвященная пересечению результатов работы различных алгоритмов идентификации РСТФ в ChIP-seq экспериментах для выявления наиболее достоверных РСТФ.

Также до конца нерешённым остается вопрос об интеграции имеющихся данных для получения более достоверных результатов картирования районов связывания транскрипционных факторов на геном. Для решения данной задачи крупные базы данных ChIP-seq экспериментов: ENCODE Portal, CistromeDB и ReMap работают в направлении улучшения интерфейсов доступа к хранящимся данным, предоставляя тем самым пользователям возможность одновременно анализировать и сопоставлять разные типы экспериментов. Также, в рамках баз данных ENCODE Portal и ReMap осуществляется мета-анализ хранящихся в рассматриваемых базах данных позиционных методов NGS.

Цель и задачи диссертационного исследования

Целью данной работы является разработка методов контроля качества и построения карты наиболее воспроизводимых геномных районов связывания транскрипционных факторов человека на основе массового сравнительного анализа ChIP-seq экспериментов.

Для достижения этой цели были поставлены и решены следующие задачи:

1. Внести в базу данных GTRD описания хранящихся в открытом доступе ChIP-seq и DNase-seq экспериментов для человека. Реализовать конвейер для стандартизации обработки данных DNase-seq.
2. Разработать методы оценки качества ChIP-seq данных на основе анализа согласованности результатов применения четырёх алгоритмов идентификации районов связывания транскрипционных факторов: MACS2, GEM, SISSRs и PICS.
3. Разработать метод для приоритезации воспроизводимых районов связывания транскрипционных факторов. Используя предложенный метод, построить карту геномных районов связывания транскрипционных факторов человека. Сравнить расположение таких районов и мотивов связывания соответствующих транскрипционных факторов, а также районов открытого хроматина.
4. Идентифицировать однонуклеотидные геномные варианты, ассоциированные с нарушениями морфологии сперматозоидов, используя данные полноэкзомного секвенирования, и проанализировать их возможное влияние на регуляцию транскрипции на основе построенной карты районов связывания транскрипционных факторов.

Научная новизна

В диссертационной работе предложены и реализованы новые методы оценки качества ChIP-seq экспериментов (FPCM и FNCM) на основе анализа

согласованности результатов применения четырёх алгоритмов идентификации районов связывания транскрипционных факторов: MACS2, GEM, SISSRs и PICS.

Разработан и реализован новый алгоритм на основе применения методов коллективного выбора, METARA, для последующего отбора наиболее воспроизводимых районов связывания ТФ на основании значений финальной агрегирующей функции. Используя предложенный метод, построена наиболее полная карта геномных районов связывания транскрипционных факторов человека. Проведен массовый анализ расположения наиболее воспроизводимых районов связывания транскрипционных факторов относительно мотивов связывания соответствующих транскрипционных факторов, а также районов открытого хроматина.

Впервые, при анализе данных полноэкзомного секвенирования были обнаружены ассоциации однонуклеотидных геномных вариантов с различными нарушениями морфологии сперматозоидов человека. Найденные 135 геномных вариантов были рассмотрены с точки зрения влияния на регуляцию транскрипции. Были выявлены как однонуклеотидные варианты, располагающихся в генах, кодирующих факторы транскрипции, так и геномные варианты, приводящие к изменению эффективности связывания транскрипционных факторов, участвующих в регуляции сперматогенеза, с ДНК.

Теоретическая значимость диссертационного исследования

Предложены новые методы для контроля качества ChIP-seq экспериментов на основе сравнения результатов разных алгоритмов для выявления РСТФ, что позволило общее оценить как общее количество таких районов, так и долю ложно идентифицированных РСТФ.

Разработан новый алгоритм применения методов коллективного выбора, METARA, для последующего отбора наиболее воспроизводимых районов связывания транскрипционных факторов на основании их ранжирования, что позволило объединить данные из различных ChIP-seq экспериментов в базе данных GTRD.

В рамках диссертационного исследования были впервые идентифицированы однонуклеотидные геномные вариации, ассоциированные с различными нарушениями морфологии сперматозоидов, характерные для популяции, проживающей на территории Российской Федерации.

Практическая значимость диссертационного исследования

Была создана уникальная коллекция единообразно обработанных ChIP-seq и DNase-seq экспериментов для человека. Построенные наиболее полные карты геномных районов связывания ТФ и районов открытого хроматина могут быть использованы для решения широкого спектра задач в области регуляторной геномики человека. Результаты данной работы использованы при создании отечественной базы данных GTRD. База данных GTRD является высоко востребованной для поддержки исследований по биомедицине, что подтверждается высокой цитируемостью (две публикации, в которых принял участие автор, в специализированных выпусках Nucleic Acids Research 2019 и 2021 года набрали в совокупности более 300 цитирований по версии Semantic Scholar (https://www.semanticscholar.org/), включая цитирования в журналах Nature и Science). Интеграция в базу данных GTRD онтологий тканей и клеточных типов, полученных с помощью ресурсов: BRENDA, UBERON, Cell Ontology и Cellosaurus сделала возможным автоматизированное сопоставление данных из GTRD с другими базами данных.

Результаты работы были использованы для создания отечественных и международных веб-ресурсов: HOCOMOCO (https://hocomoco11.autosome.ru/), ADASTRA (https://adastra.autosome.ru/), ANANASTRA (https://ananastra.autosome.ru/), BaMM motif (https://bammmotif.soedinglab.org/), mSigDB (https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#C3), широко используемых для биомедицинских исследований.

Методология и методы исследования

В рамках данной работы в базу данных GTRD было добавлено описание ChIP-seq и DNase-seq экспериментов для человека, доступных в крупнейших базах данных: SRA, GEO и ENCODE. Для систематизации экспериментов по тканям и клеточным типам были использованы онтологии: BRENDA, UBERON, Cell Ontology и Cellosaurus. Методологической основой для оценки качества данных секвенирования следующего поколения (NGS) являются рекомендации международного исследовательского консорциума ENCODE.

Для валидации разработанных в рамках данной работы методов анализа качества ChIP-seq экспериментов и построения карты геномных районов связывания транскрипционных факторов был использован комплексный подход оценки достоверности полученных районов. С одной стороны, данный подход основывается на анализе воспроизводимости районов связывания в других ChIP-seq и DNase-seq экспериментах. С другой стороны, используются вычислительные методы оценки эволюционной консервативности рассматриваемых регионов из базы данных UCSC и идентификации мотивов связывания транскрипционных факторов на основе позиционно-весовых матриц из базы данных HOCOMOCO v11.

Идентификация однонуклеотидных геномных вариантов в данных полноэкзомного секвенирования выполнялась в соответствии с рекомендациями GATK Best Practices. Для интерпретации геномных вариантов, ассоциированных с различными нарушениями морфологии сперматозоидов, в контексте регуляции транскрипции были использованы базы данных: GTRD, ADASTRA и GTEx.

Положения, выносимые на защиту

1. Для районов связывания транскрипционных факторов, выявляемых только одним из алгоритмов (MACS2, GEM, SISSRs или PICS) при высоких значениях разработанной оценки доли ложно идентифицированных районов (FPCM) характерны: сниженная воспроизводимость в других ChIP-seq экспериментах, сниженная эволюционная консервативность, более низкие вероятности расположения в районах открытого хроматина и наличия мотивов связывания транскрипционных факторов.
2. Новый алгоритм METARA, разработанный на основе применения методов коллективного выбора, позволяет приоритезировать воспроизводимые районы связывания транскрипционных факторов с ДНК: чем выше вес, присвоенный алгоритмом, тем более вероятнее выявленный район располагается в районе открытого хроматина и тем чаще он содержит мотивы связывания транскрипционных факторов, предсказанные позиционной весовой матрицей.
3. Показано, что четыре однонуклеотидных геномных варианта: rs138595914, rs2304961, rs2270420, rs71486131 ассоциированы с нарушениями морфологии сперматозоидов. Выявленные однонуклеотидные варианты располагаются в наиболее воспроизводимых районах связывания транскрипционных факторов, участвующих в регуляции сперматогенеза: AR, CTCF и SRBP2, и влияют на эффективности их связывания с ДНК.

Степень достоверности и апробация результатов

Результаты работы были представлены и обсуждены на следующих российских и международных конференциях: Международная конференция по биоинформатике, структуре и регуляции генома (BGRS\SB'2018, BGRS\SB'2020, BGRS\SB'2022, BGRS\SB'2024, г. Новосибирск, Россия), Международный конгресс “Биотехнология: Состояние И Перспективы Развития“ (25-27 февраля 2019 г., Москва, Россия), XXIV съезд физиологического общества им. И.П. Павлова (11–15 сентября 2023 г., Санкт-Петербург, Россия), Международной конференции “Распределенные Информационно-вычислительные Ресурсы. Цифровые Двойники И Большие Данные.” (DICR-2019, 3-6 декабря 2019 г., Новосибирск, Россия), Международной московской конференции по вычислительной молекулярной биологии (MCCMB’2023, г. Москва, Россия).

Публикации

Материалы диссертационной работы отражены в 25 научных публикациях, включая: 13 публикаций в журналах, индексируемых в международных базах данных Web of Science/Scopus, из которых 8 публикаций Q1.

Личный вклад автора

База данных GTRD - результат работы большого количества аннотаторов и биоинформатиков. В ходе диссертационной работы автором лично проаннотировано 1701 DNase-seq и 1347 ChIP-seq экспериментов для человека. Доработана программа для полуавтоматической аннотации NGS данных,

GEOminer. Реализован конвейер по анализу данных DNase-seq. Результаты представлены в публикациях (Yevshin et al., 2018; Kolmykov et al., 2020; Kolpakov et al., 2019; Kolmykov et al., 2021a; Kolpakov et al., 2021).

В работах (Kulyashov et al., 2020a; Kulyashov et al., 2020б) совместно с Куляшовым М. А. была проведена интеграция в БД GTRD различных онтологий клеточных типов и экспериментальных условий.

В методологической работе (Kolmykov et al., 2019) автором была выполнена разработка, реализация и валидация новых методов анализа качества ChIP-seq экспериментов на основе оценки доли ложноположительных (FPCM) и ложноотрицательных (FNCM) пиков в ChIP-seq данных.

Разработан и валидирован алгоритм многостадийного применения методов коллективного выбора (METARA) для мета-анализа ChIP-seq экспериментов. Результаты представлены в публикациях (Kolmykov et al., 2020; Kolmykov et al., 2021a).

В работах, посвященных базам данных: HOCOMOCO и ADASTRA (Abramov et al., 2021; Boytsov et al., 2022; Vorontsov et al., 2024), автор участвовал в подготовке и экспертной оценке информации из базы данных GTRD.

Автором работы были идентифицированы однонуклеотидные геномные варианты в данных полноэкзомного секвенирования и проведён анализ их ассоциации с нарушениями морфологии сперматозоидов. Реализованный сценарий идентификации однонуклеотидных вариаций представлен в публикации (Kolmykov et al., 2021б). При помощи результатов применения алгоритма METARA и данных из БД ADASTRA было исследовано влияние выявленных геномных вариаций на эффективность связывания транскрипционных факторов в наиболее воспроизводимых районах связывания транскрипционных факторов.

Структура и объем диссертации

Диссертационная работа состоит из введения, обзора литературы, пяти разделов с описанием результатов работы, заключения, выводов, списка публикаций по теме диссертации, списка литературы (159 источников). Работа изложена на 141 странице, содержит 35 рисунков и 5 таблиц.

Благодарности

Автор глубоко признателен научным руководителям: к.б.н. Кондрахину Ю.В. и д.б.н. Колпакову Ф.А.; коллегам и соавторам: Осадчуку А.В., Кулаковскому И.В., Акбердину И.Р., Куляшову М.А., Пономаренко М.П., Евшину И.С., Шарипову Р.Н., Жатченко С.А., Пинтусу С.С., Левицкому В.Г., Вишнивецкой А.П. – за ценные дискуссии и поддержку, оказанную на всех этапах выполнения работы.

Кроме того, автор выражает благодарность сотрудникам Сектора репродуктивных технологий человека ИЦИГ СО РАН под руководством д.б.н. Осадчук Л.В, сотрудникам Сектора геномных исследований ИЦиГ СО РАН и лично к.б.н. Васильеву Г.В. – за подготовку образцов, проведение и предоставление результатов полноэкзомного секвенирования.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 Регуляция транскрипции

Транскрипция - это фундаментальный процесс, лежащий в основе ответа клеток живых организмов на различные внутренние и внешние сигналы. Он включает в себя сложные взаимодействия между молекулами ДНК, регуляторными белками и структурой хроматина (см. Рисунок 1.1.1). Изучение механизмов транскрипционной регуляции имеет решающее значение для расшифровки функций генов, понимания процессов развития организма и механизмов возникновения различных заболеваний (Bolt et al., 2020; Kuznetsova et al., 2020; Dupont et al., 2022; Grosveld et al., 2021).

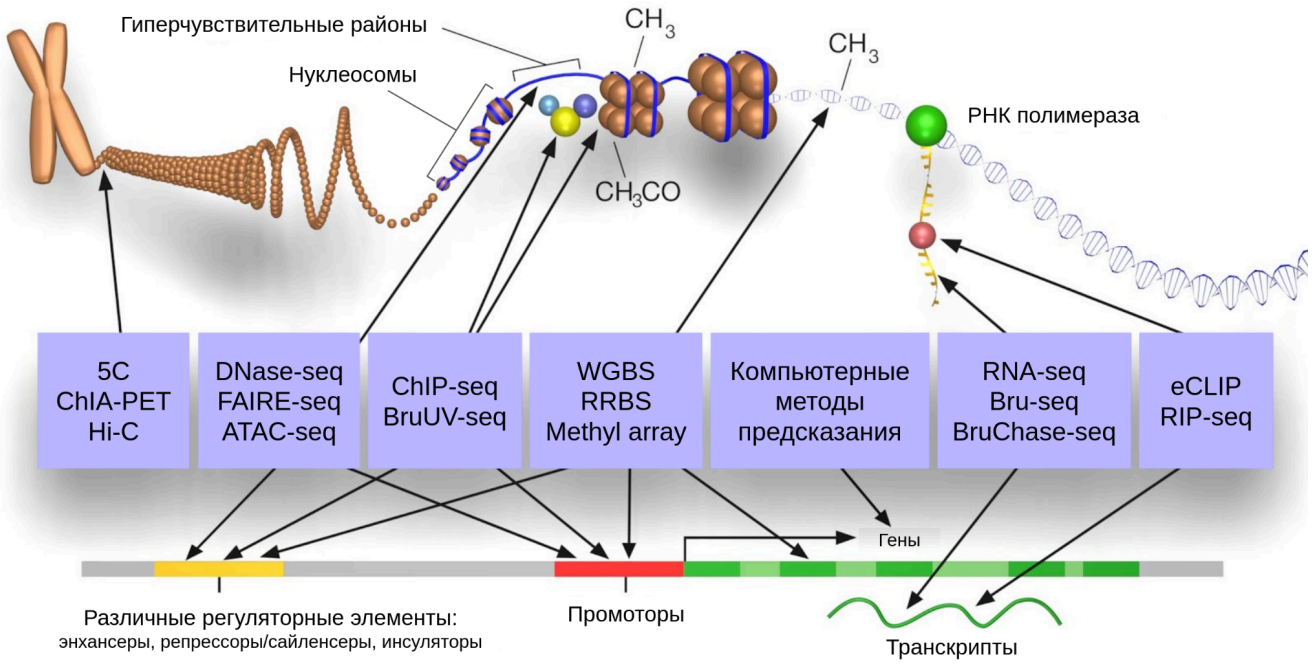


Рисунок 1.1.1 - Обобщенное представление основных компонентов регуляции транскрипции и экспериментальных NGS методов их изучения (Moore et al., 2020)

Транскрипционные факторы (ТФ) играют ключевую роль в регуляции экспрессии генов на уровне транскрипции. Регуляция осуществляется путем

связывания ТФ с сайтами в ДНК, расположенными в регуляторных областях генов, таких как промоторы, сайленсеры и энхансеры. После связывания ТФ с ДНК могут происходить события, приводящие как к активации транскрипции, так и к её подавлению. Это может осуществляться путем изменения плотности упаковки хроматина, прямого взаимодействия транскрипционного комплекса и РНК-полимеразой, а также через привлечение кофакторов (Lambert et al., 2018).

Изначально для изучения экспрессии генов и транскрипционной регуляции использовались такие традиционные методы, как секвенирование по Сэнгеру, микрочипы и иммунопреципитация хроматина (ChIP). Однако эти методы имели ограничения по пропускной способности, разрешению и способности обеспечить комплексное представление о геноме (Johnson et al., 2007). Появление NGS в середине 2000-х годов произвело революцию в изучении транскрипционной регуляции, позволив с высоким разрешением изучить эти процессы в масштабах всего генома. Этот прогресс позволил разработать различные методы на основе NGS, предназначенные для изучения различных аспектов транскрипционной регуляции, приведенные в таблице 1.1.1.

Таблица 1.1.1 – Сводная таблица методов NGS по исследованию регуляции транскрипции

Метод	**Описание**	**Основные приложения**	**Комментарий**	**Необходимое количество клеток**	**Глубина секвенирования (прочтений)**
Методы определение конформации хромосом					
5C (Dostie et al., 2006)	Chromosome Conformation Capture Carbon Copy; картирует взаимодействия между геномными локусами	Изучение 3D-организации генома	Высокое разрешение картирования взаимодействий	10^6 - 10^7	$5x10^7$ - 10^8
ChIA-PET (Fullwood et al., 2009)	анализ взаимодействий хроматина с помощью секвенирования парных концов (Chromatin Interaction Analysis by Paired-End-Tag sequencing), выявляет опосредованные белками взаимодействия хроматина	Картирование дальних взаимодействий хроматина	Прямое выявление опосредованных белками взаимодействий	10^7 - 10^8	10^8 - $2x10^8$
Hi-C (Lieberman-Aiden et al., 2009)	Конформации хромосом высокого порядка (High conformation Capture), выявляет пространственную организацию геномов	Картирование взаимодействий на уровне всего генома	Всеобъемлющее представление архитектуры хроматина, однако сложный анализ данных	10^6 - 10^7	$5x10^8$ - 10^9
Методы определения районов открытого хроматина					
DNase-seq (Boyle et al., 2008)	Идентификация гиперчувствительных к DNase I участков ДНК, которые соответствуют открытым и доступным для ТФ регионам хроматина.	Выявление активных регуляторных элементов; картирование открытого хроматина	Высокая чувствительность и специфичность	10^6	$5x10^7$ - 10^8

ATAC-seq (Buenrostro et al., 2013)	Анализ доступности хроматина с использованием транспозазы Tn5, которая встраивает адаптеры секвенирования в открытые участки хроматина	Выявление активных регуляторных элементов; картирование открытого хроматина	Низкие требования к входному материалу, быстрая постановка	$5x10^5$	$5x10^7$ - 10^8
FAIRE-seq (Giresi et al., 2007)	Секвенирование изолированных регуляторных элементов, полученных при помощи фиксации белков с помощью формальдегида; выявляет участки, свободные от нуклеосом	Выявление активных регуляторных элементов; картирование открытого хроматина	Простая процедура, не требует антител. Однако низкое разрешение по сравнению с DNase-seq	10^6 - 10^7	$2x10^7$ - $5x10^7$
Методы исследования взаимодействия белков с ДНК/РНК					
ChIP-seq (Johnson et al., 2007)	Иммунопреципитация хроматина с последующим секвенированием; картирует взаимодействия белков с ДНК	Выявление геномных районов связывания белков	Высокая специфичность и чувствительность, однако требует высококачественных антител	10^6 - 10^7	$2x10^7$ - $5x10^7$
BruUV-seq (Paulsen et al., 2014)	Использует ультрафиолетовый свет для введения блокирующих транскрипцию повреждений ДНК для создания ковалентных сшивок между РНК и белками. Исследование недавно синтезированных РНК; дает информацию о динамике транскрипции и позволяет идентифицировать гены, активно транскрибируемые в момент инкубации	Изучение транскрипционной активности, стартов транскрипции и РНК-белковых взаимодействий		10^6	$1x10^7$ - $2x10^7$

eCLIP (Van Nostrand et al., 2016)	Позволяет идентифицировать сайты связывания РНК-связывающих белков (RBP) с транскриптами благодаря этапу UV кросс-линкинга	Изучение взаимодействий РНК с белками	Высокая чувствительность, прямое выявление сайтов связывания	10^6	$2x10^7$ - $5x10^7$
RIP-seq (Keene et al., 2006)	Позволяет идентифицировать РНК, связанные с конкретными RBP благодаря этапу иммунопреципитации со специфичными антителами.	Изучение взаимодействий РНК с белками	Выявляет связанные с белками молекулы РНК	10^6	$2x10^7$ - $5x10^7$
Методы исследования паттернов метилирования ДНК					
WGBS (Lister et al., 2009)	Предоставляет карту метилирования ДНК по всему геному, что позволяет исследовать метилирование в промоторах, экзонах, интронах и межгенных областях.	Профилирование метилированных районов на уровне целого генома	Высокое разрешение, полногеномный анализ районов метилирования	10^5 - 10^6	$5x10^8$ - $1x10^9$
RRBS (Meissner et al., 2005)	Фокусируется на областях генома, богатых CpG-островками, что позволяет получить подробную информацию о метилировании в этих областях с меньшими затратами и меньшим объемом данных по сравнению с WGBS.	Таргетное картирование метилированных районов	Низкие требования к входному материалу; Ограничен CpG-обогащенными областями	10^4 - 10^5	$3x10^7$ - $5x10^7$
Methyl array (Bibikova et al., 2006)	Метод анализа районов метилирования ДНК на основе микрочипов в заранее выбранных районах генома	Профилирование метилирования таргетных районов	Низкое разрешение по сравнению с методами на основе NGS секвенирования	10^4 - 10^5	Не применимо
Методы исследования активности транскрипции					
RNA-seq (Wang et	Секвенирование кДНК для анализа транскриптома: экспрессия генов,	Количественная оценка экспрессии	Высокая чувствительность	10^5 - 10^6	$2x10^7$ - $5x10^7$

al., 2009)	идентификация сплайсинговых вариантов транскриптов, SNV, поиск химерных генов и т.д.	генов			
CAGE-seq (Takahashi et al., 2012)	Cap Analysis Gene Expression Sequencing, определяет старты транскрипции путем секвенирования 5'-концов мРНК	Количественная оценка экспрессии генов; Анализ сайтов начала транскрипции, промоторных и энхансерных районов	Высокая точность определения сайтов начала транскрипции	10^5 - 10^6	10^7 - $2x10^7$
Bru-seq (Paulsen et al., 2013)	Дает информацию о динамике транскрипции и позволяет идентифицировать гены, активно транскрибируемые в момент инкубации	Изучение транскрипционной активности		10^6	10^7 - $2x10^7$
BruChase-seq (Rabani et al., 2011)	Позволяет анализировать, как быстро различные РНК транскрипты деградируют после их синтеза	Изучение стабильности и скорости распада РНК		10^6	10^7 - $2x10^7$

Описанные выше методы высокопроизводительных исследований генерируют огромные массивы информации по ключевым факторам регуляции транскрипции. Исходные данные таких экспериментов представлены в специальных хранилищах данных, основными из которых являются SRA (https://www.ncbi.nlm.nih.gov/sra) и GEO (https://www.ncbi.nlm.nih.gov/geo/). Эти данные были получены как отдельными лабораториями, так и большими международными консорциумами:

- ENCODE (Luo et al., 2020) - энциклопедия (регуляторных) ДНК элементов;
- FANTOM5 (Abugessaisa et al., 2021) - функциональная аннотация и уровень экспрессии функциональных ДНК участков;
- Roadmap Epigenomics Project (Zhao et al., 2020) - данные по эпигеномике и метилированию ДНК;
- GTEx - Genotype-Tissue Expression – коллекция вариаций генной экспрессии среди индивидуумов и в 44 различных тканях тела человека, а также паттернов тканеспецифичности для выявления генетических основ болезней человека;

Однако эти данные недостаточно интегрированы друг с другом, что существенно затрудняет их совместное использование как для понимания механизмов регуляции транскрипции, так и для решения практических задач - например, предсказание возможных эффектов одиночных нуклеотидных замен (SNV) в регуляторных районах генов. В частности, интеграцию большого объема схожих данных затрудняют различия как в качестве обрабатываемых данных, так и различия в пайплайнах обработки экспериментальных данных.

1.2 Алгоритмы идентификации районов связывания транскрипционных факторов в данных ChIP-seq экспериментов

Метод иммунопреципитации хроматина с последующим секвенированием (ChIP-seq) впервые был описан в 2007 году (Johnson et al., 2007) и к настоящему времени является стандартом для *in vivo* картирования районов связывания белка с ДНК, а результаты применения данного подхода сыграли важную роль в исследовании различных эпигенетических механизмов регуляции. На рисунке 1.2.1 представлена схема проведения ChIP-seq эксперимента. На первом этапе происходит сшивка связанных белков с хроматином, фрагментация хроматина, захват фрагментов ДНК, связанных с одним исследуемым белком, с помощью специфичного для него антитела и секвенирование концов захваченных фрагментов с помощью секвенирования нового поколения (NGS). На последующих этапах выполняется биоинформатический анализ полученных прочтений. В частности, прочтения картируются на референсный геном, образуя скопления в местах взаимодействия белка с ДНК. Такие скопления, пики, являются потенциальными районами связывания ТФ (РСТФ).

Для поиска РСТФ по данным ChIP-seq опубликовано более 30 алгоритмов (пикколлеров) (Thomas et al., 2017). В настоящее время уже проведены различные сравнительные анализы таких алгоритмов. В частности, один из первых сравнительных анализов был опубликован еще в 2009 году (Laajala et al., 2009). Как правило, эти сравнения обычно производились на небольшом количестве наборов экспериментов, а также с использованием различающихся критериев сравнения. В результате этого различные сравнительные анализы приводили нередко к противоречивым результатам. Например, в трех работах (Harmanci et al., 2014; Koohy et al., 2014; Micsinai et al., 2012) были получены противоречивые

результаты сравнения алгоритмов: MACS, SICER (Zang C. et al., 2009) и F-Seq (Boyle A. P. et al., 2008). Такое состояние дел однозначно указывает на крайнюю необходимость в создании более совершенных как метрик, так и критериев сравнения, а также создания единого набора надежных ChIP-seq данных, который бы использовался в дальнейших сравнительных анализах.

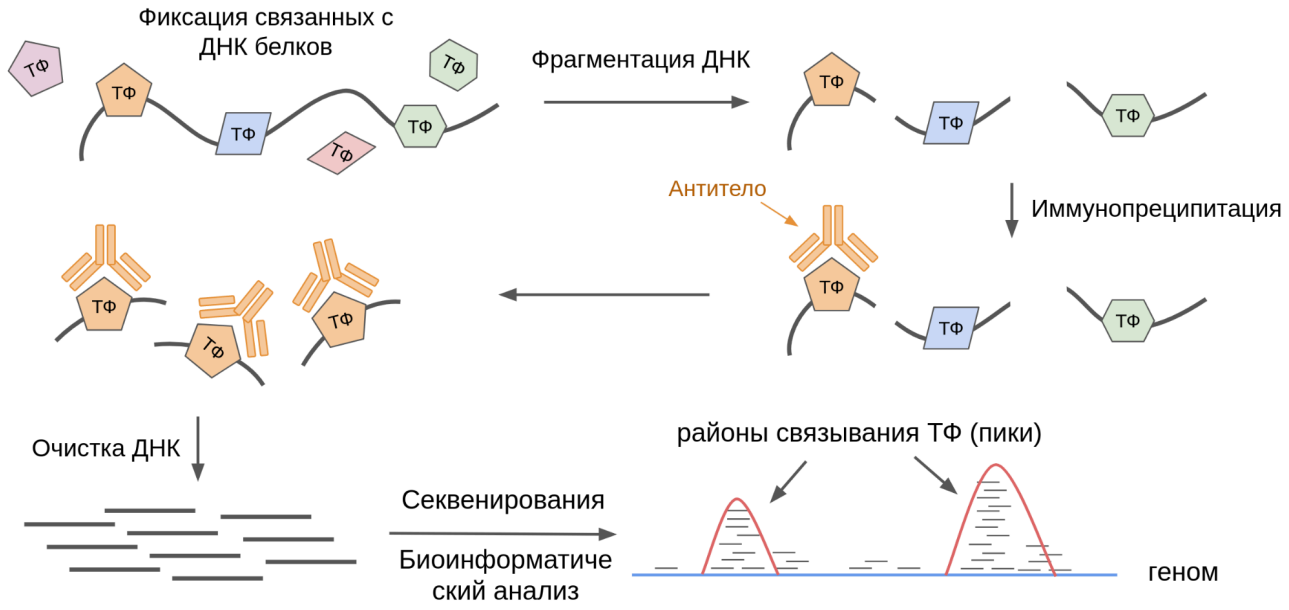


Рисунок 1.2.1 - Схема эксперимента по иммунопреципитации хроматина с последующим высокопроизводительным секвенированием ДНК (ChIP-seq)

Рассмотрим процесс идентификации пиков на примере алгоритма MACS2. На первом этапе работы алгоритма MACS2 происходит удаление из набора дуплицированных прочтений, т. е. прочтений, у которых совпадают координаты выравнивания на референсный геном. По умолчанию алгоритм оставляет только один экземпляр подобных прочтений. Однако пользователь может либо самостоятельно указать максимальное количество прочтений-дубликатов, либо воспользоваться специальной опцией (auto), которая автоматически рассчитает максимально допустимое число прочтений-дубликатов на основании биномиального распределения (p-value < 10^{-5}) в каждом конкретном случае.

Подобное внимание к обработке дуплицированных прочтений объясняется существенным их влиянием на результаты идентификации пиков (Tian et al., 2019). В частности, дубликаты прочтений являются следствием чрезмерной амплификации изначально недостаточно большой библиотеки последовательностей ДНК (Klepikova A. V. et al., 2017). ПЦР-амплификация является основным источником дуплицированных прочтений, так называемых "ПЦР-дубликатов" (Tian S. et al., 2019). Для решения данной проблемы для анализа различных типов NGS экспериментов общепринятым является включение этапа удаления ПЦР-дубликатов из набора выровненных прочтений. Наиболее часто встречается использование утилиты MarkDuplicates из программного пакета Picard (https://broadinstitute.github.io/picard/).

Однако совпадающие по локализации прочтения могут быть обусловлены естественными причинами. Поскольку при проведении ChIP-seq эксперимента секвенируется относительно небольшая часть генома, то с увеличением глубины секвенирования, растёт вероятность получения естественных дубликатов (Parekh S. et al., 2016). Таким образом, удаление дубликатов может привести к неверной оценке наблюдаемых профилей обогащения. В некоторых случаях используют удаление ПЦР-дубликатов только при совпадении координат парно-концевых прочтений. В качестве решения этой проблемы со стороны постановки NGS эксперимента могут быть использованы UMI-баркоды, наборы уникальных последовательностей нуклеотидов установленной длины, которые случайным образом лигируются к адаптерным последовательностям, и позволяют, в том числе, идентифицировать ПЦР-дубликаты (Tsagiopoulou M. et al., 2021).

В рамках проекта ENCODE был составлен черный список (ENCODE blacklist) районов в геномах *H. sapiens*, *M. musculus*, *C. elegans* и *D. melanogaster* (Amemiya et al., 2019). Описанные районы демонстрируют аномальный, неструктурированный или высокий уровень сигнала в NGS данных вне зависимости от клеточной линии и типа эксперимента. При стандартном анализе

ChIP-seq данных рекомендуется не учитывать пики, идентифицированные в данных областях генома (Cancer Genome Atlas Research Network, 2011).

При рассмотрении профилей выравнивания прочтений в окрестности сайта связывания ТФ наблюдается бимодальность распределения выровненных прочтений (Wilbanks et Facciotti, 2010). При этом каждая мода располагается на отдельной цепи ДНК (см. Рисунок 1.2.2), а расстояние между ними соответствует средней длине секвенируемого фрагмента ДНК. Подобная асимметрия обогащения прочтений вокруг сайта связывания влечёт за собой размывание границ в процессе его идентификации. Для устранения данного эффекта на следующем этапе алгоритма MACS2 происходит смещение каждого прочтения на половину длины фрагмента в направлении от 5’-конца к 3’-концу и расширение всех прочтений относительно их центров до размеров секвенируемого фрагмента.

Для оценки длины фрагмента в описываемом алгоритме существует два подхода. Первый подход применим только в случае парно-концевых данных и заключается в вычислении среднего расстояния между 5’-концами парных прочтений. Второй подход основан на вычислении среднего расстояния между двумя модами для районов с высоким уровнем обогащения. Данный набор состоит 1000 районов случайно выбранных из районов, для которых отношение количества прочтений в заданном районе существенно превышает количество прочтений в случае, если бы все прочтения были распределены по геному равномерно. При этом размер заданного района поиска, равный по длине двум (bandwidth), и пороговая величина обогащения (mfold) может задаваться пользователем отдельно.

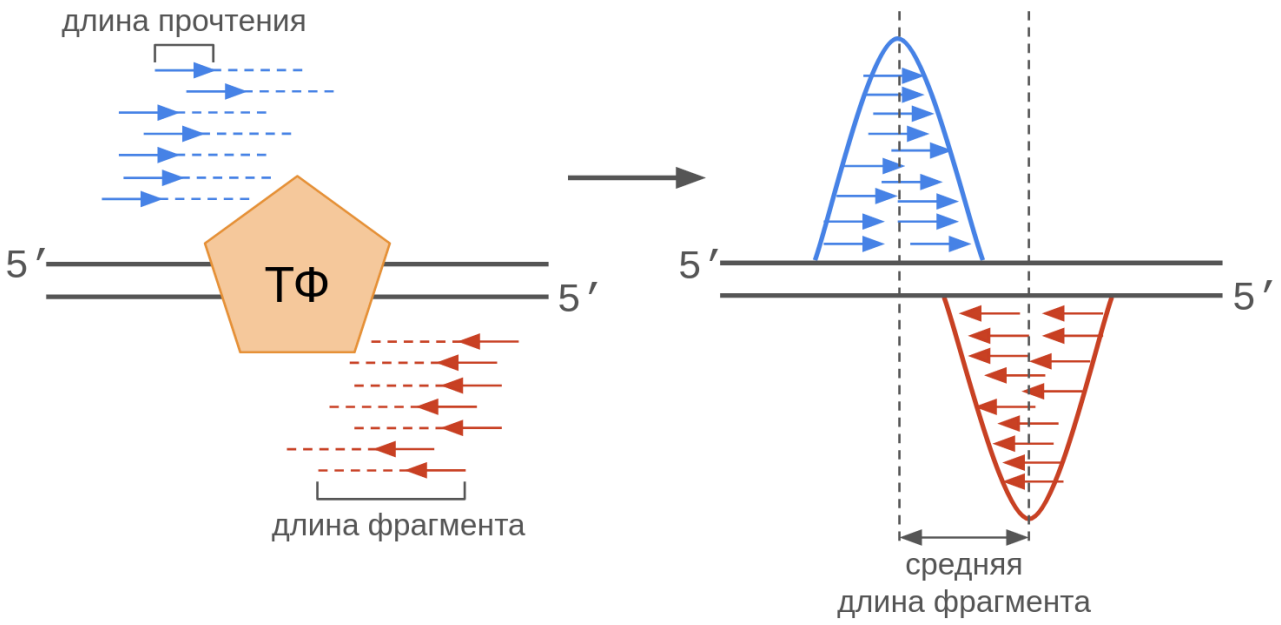


Рисунок 1.2.2 – Распределение прочтений в ChIP-seq эксперименте по цепями ДНК

На следующем этапе происходит поиск потенциальных пиков, для этого MACS2 сканирует профиль выравнивания прочтения рамкой размером 2d, где d - размер фрагмента, полученный на предыдущем этапе. Для описания распределения прочтений по геному MACS2 использует распределение Пуассона.

$$P(Y = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

В данном контексте λ описывает ожидаемое количество прочтений в рассматриваемом окне поиска пика.

$$\lambda = \frac{\text{число событий } (k)}{\text{число измерений } (n)} = \frac{\text{длина прочтения * количество прочтений}}{\text{эффективная длина генома}}$$

Для оценки соответствия количества прочтений распределению Пуассона в заданном окне поиска в алгоритме MACS2 используется динамический параметр lambda (λlocal). Данный параметр оценивается несколько раз, варьируя размер окна поиска: 1000 п.н., 5000 п.н., 10000 п.н. и для района, равного по длине эффективной длине генома (ЭДГ). В качестве искомого λlocal выбирается λ с наибольшим значением, т. е. λlocal = max(λ1000, λ5000, λ10000, λЭДГ).

Следует отметить, что использование ЭДГ вместо полной длины генома связано с тем, что в геноме присутствуют районы с низкой уникальностью, что приводит к тому что, множество прочтений может быть с равным успехом картировано на несколько позиций в геноме, мульти-выравнивание (multimapping). Таким образом, для оценки эффективной длины генома необходимо оценить суммарную длину всех уникальных последовательностей длиной k (k-mer), где k - длина прочтения. Использование эффективной длины генома помогает скорректировать потерю сигнала в районах с низким покрытием.

Однако, повышенное число прочтений не обязательно являются следствием связывания белка с молекулой ДНК в данном районе. Многие исследования указывают на то, что распределение прочтений по геному в ChIP-seq данных не равномерно и зависит от многих факторов. Для нивелирования описываемых эффектов в повсеместную практику вошло использование контрольных экспериментов (Landt et al., 2012).

Существует два основных типа контролей для ChIP-seq экспериментов. В первом случае в качестве контроля используется ДНК выделенная в тех же условиях из той же ткани или клеточного типа, но без стадии иммунопреципитации (input DNA; ДНК-контроль или инпут-контроль). Данный тип контроля позволяет скорректировать эффекты, вызванные неравномерностью фрагментации ДНК при пробоподготовке, опосредованные большей подверженностью районов открытого хроматина к сонификации, по сравнению с остальными районами генома. Однако данный тип контроля не корректирует ошибки, вызванные неспецифическими взаимодействиями антител на этапе иммунопреципитации (Xu J. et al., 2021; Kharchenko et al., 2008).

Для борьбы с данным эффектом в качестве контроля проводится ещё один ChIP-seq эксперимент с использованием антител к белкам, которые не способны связаться с исследуемым белком, например, к внеядерным антигенам. В частности, наибольшую популярность имеют антитела к иммуноглобулинам

класса G (IgG-контроль или “mock IP” контроль). Некоторые работы указывают, что данный тип контроля также содержит в себе информацию о неравномерности сонификации библиотеки (Kidder et al. 2011; Landt et al. 2012). Таким образом, использование IgG-контролей является предпочтительным при проведении ChIP-seq исследований, поскольку описывает шум, появляющийся на разных этапах проведения ChIP-seq эксперимента. В 2014 году Marinov с соавторами при проведении массового анализа качества ChIP-seq данных показали, что IgG-контроли позволяют обнаружить события неспецифического связывания (Marinov et al., 2014).). Несмотря на это, IgG-контроли проигрывают в своей популярности ДНК-контролям. Например, в базе данных ENCODE из 3205 контролей только 48 помечены как IgG-контроли. Более того сравнение IgG-контролей и ДНК-контролей, проведённое Xu с соавторами, показало, что как в IgG-контролях, так и в ДНК-контролях обогащение прочтениями ассоциировано с уровнем доступности рассматриваемого района генома (Xu et al., 2021). Таким образом, популярность ДНК-контролей, как более простого в подготовке контроля, более чем оправдана.

Некоторые алгоритмы идентификации пиков поддерживают использование контрольных экспериментов. В частности, алгоритм MACS2, сперва приводит контрольный и исследуемый эксперимент к равному количеству прочтений. При этом, большая по размеру библиотека линейно приводится к размеру меньшей библиотеки. Затем, профиль выравнивания прочтений из контрольного эксперимента используется для расчета значения λlocal. Таким образом, при оценке наличия обогащения в рассматриваемом районе исследуемый эксперимент сравнивается с контрольным.

Поскольку идентификация каждого пика - независимое событие, в результате среди тысяч пиков может быть идентифицировано существенное количество ложноположительных пиков. Для того, чтобы внести поправку на множественное сравнение на финальном этапе алгоритма MACS2 используется

поправка Бенджамини-Хохберга (Benjamini-Hochberg correction) (Benjamini et Hochberg, 1995).

1.3 Анализ качества ChIP-seq экспериментов

В рамках проекта ENCODE (ENCODE Project Consortium et al., 2012) разработан и достаточно широко апробирован целый набор методов как для оценки качества выравнивания прочтений (например, NRF, PBC1, PBC2 (https://www.encodeproject.org/data-standards/terms/#library)), так и для оценивания качества РСТФ (например, FRiP, IDR (Landt et al., 2012)).

Сложность библиотеки связана со многими факторами, такими как качество антител, количества ДНК, особенности фрагментации или чрезмерная амплификация библиотеки с помощью ПЦР (Bailey et al., 2013). Сложность исследуемой библиотеки напрямую зависит от уникальности последовательностей, входящих в её состав. Для оценки сложности библиотеки используются 2 характеристики: Non-Redundant Fraction (NRF) и PCR Bottlenecking Coefficients (PBC).

NRF - Доля неизбыточных прочтений (Non-Redundant Fraction) - отношение количества прочтений, для которых число дубликатов не превышает заданный порог к общему числу прочтений. Пороговое значение зависит от ожидаемого числа дуплицированных прочтений. Например, для организмов с относительно небольшими геномами (например, дрожжей и бактерий) данное пороговое значение может быть больше 1. В экспериментах, где большее число прочтений имеет отношение к сигналу (например, при исследовании сайтов посадки РНК-полимеразы II), также может наблюдаться повышенное число естественных дубликатов прочтений. Также обогащение прочтениями протяжённых участков генома с низкой сложностью снижает значение NRF (Nakato et al., 2017).

Поскольку NRF зависит от глубины секвенирования, то для сравнения между собой нескольких экспериментов по данному показателю, необходимо принимать во внимание размеры исследуемых библиотек. Согласно рекомендациям консорциума ENCODE (https://www.encodeproject.org/data-standards/terms/) библиотека, состоящая хотя бы наполовину из уникальных прочтений, является приемлемой по данному показателю качества. Идеальной по данному показателю библиотека начинает считаться при NRF > 0.8 при минимальном размере библиотеке равном 10^7 прочтений.

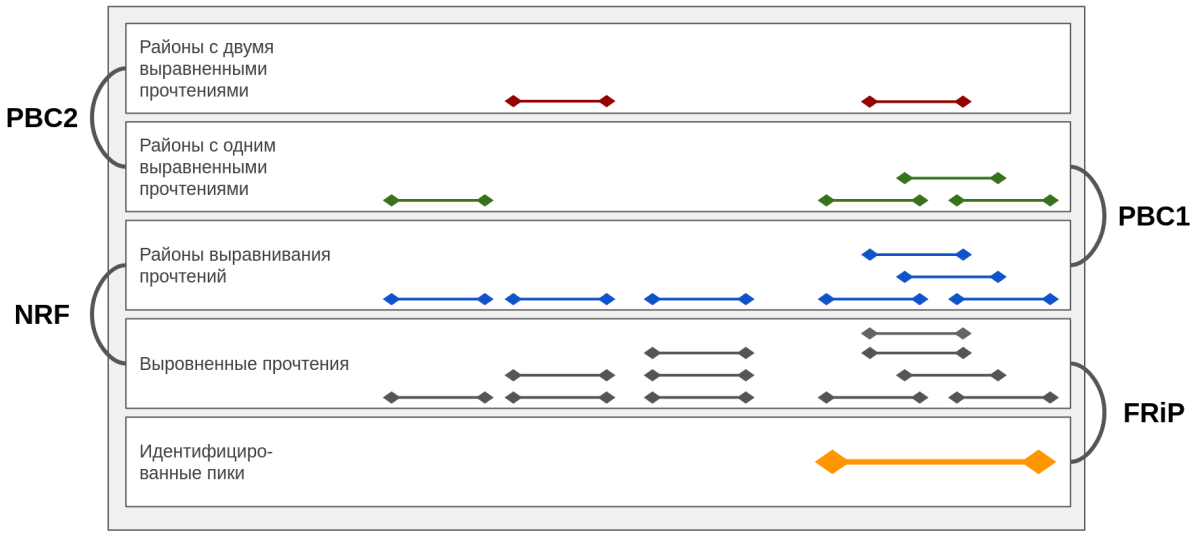
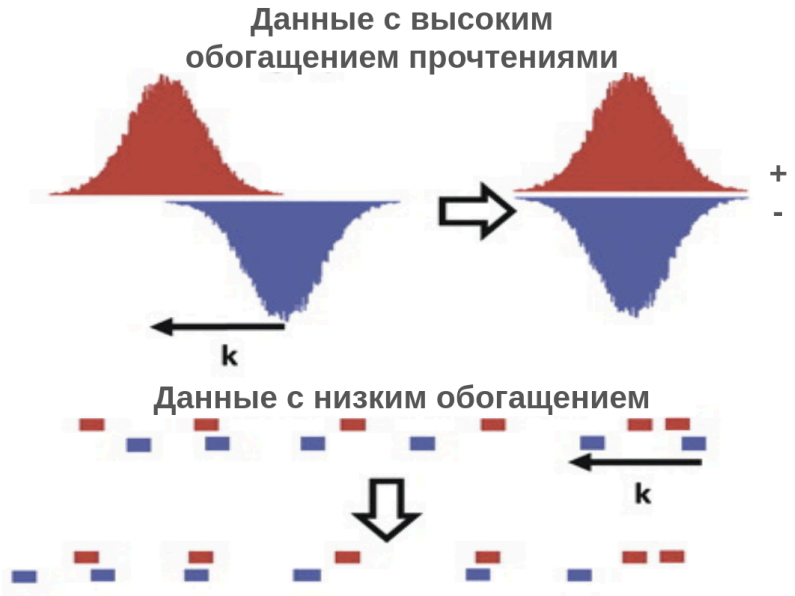


Рисунок 1.3.1 – Визуализация подмножеств, сформированных на основании выровненных прочтений, необходимых для расчета характеристик качества: FRiP, NRF, PBC1, PBC2

PBC - коэффициент ограничения ПЦР (PCR Bottlenecking Coefficients) также позволяет оценить сложность исследуемой библиотеки. Встречается два варианта расчёта данной характеристики. Первый вариант (PBC1) является отношением количества позиций на геноме, на которые выровнялось ровно одно прочтение, к общему числу уникальных районов выравнивания. Второй вариант представляет собой отношение количества районов, на которые выровнялось только одно

прочтение, к районам, на которые картировалось ровно два прочтения (https://www.encodeproject.org/data-standards/terms/#library). PBC представляет собой более консервативную, по сравнению с NRF, оценку сложности библиотеки, поскольку вместо общего числа прочтений, используется количество позиций, на которые были выровнены прочтения (Qin et al., 2016). Согласно рекомендациям консорциума ENCODE негативный эффект ПЦР-амплификации в исследуемой библиотеке наблюдается при снижении значений PBC1 и PBC2 ниже 0.9 и 10, соответственно. Однако приемлемые значения PBC1 и PBC2 начинаются со значений: 0.8 и 3.0 соответственно (Landt et al., 2012).

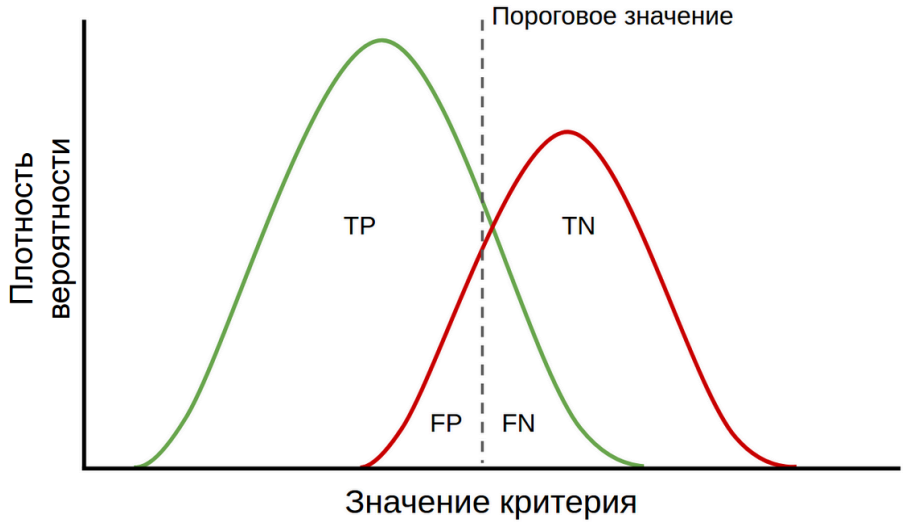


k - средняя длина фрагмента

Рисунок 1.3.2 – Визуализация смещения обогащения выравненными прочтениями относительно прямой и обратной цепей ДНК

FRiP - доля картированных чтений, попадающих в регионы, называемые пиками (FRiP), т.е. количество полезных прочтений в значительно обогащенных прочтениями районах, деленное на количество всех картированных прочтений. В целом, значения FRiP положительно коррелируют с количеством потенциальных районов связывания (Landt et al., 2012).

Кроме того, каждый из методов идентификации пиков, как правило, также присваивает каждому идентифицированному пику различные характеристики качества, например, такие как p-value, Q-value, fold enrichment, number of tags. Фиксируя те или иные пороговые значения для этих характеристик, можно снижать/повышать количество ложно предсказанных районов (FP), но при этом будет автоматически повышаться/снижаться количество непредсказанных районов (FN) (см. Рисунок 1.3.3).



FP - ложноположительные классификации, FN - ложноотрицательные классификации, TN - истинно отрицательные классификации, TP - истинно положительные классификации.
Рисунок 1.3.3 – Распределение значений исследуемого критерия и определение порогового значения

При наличии нескольких биологических реплик для выявления подмножества наиболее “качественных” РСТФ (без возрастания количества ложно непредсказанных РСТФ) можно дополнительно рассмотреть для каждого фиксированного РСТФ такую характеристику, как количество его пересечений с другими РСТФ для данного ТФ, полученными из других экспериментов (Kulakovskiy, 2018). Полезность такой дополнительной характеристики уже продемонстрирована на примере применения комбинированного теста Фишера (Fisher’s combined probability test) к 7 биологическим репликам, полученным на клеточной линии K562 (Wang et al., 2018). Важно отметить, что несмотря на

наличие целого списка уже существующих характеристик качества, их систематический и исчерпывающий сравнительный анализ в литературе на сегодняшний день отсутствует. Тем самым, задача создания наиболее эффективной(ых) характеристики(к) и построения наиболее точного метода оценивания качества фиксированного множества РСТФ до сих пор является актуальной. Наконец, необходимо подчеркнуть, что основная часть всех указанных выше характеристик качества направлена на контроль ложно предсказанных РСТФ. Тем самым, незаслуженно мало уделяется внимания контролированию ложно непредсказанных РСТФ.

1.4 Мета-анализ ChIP-seq экспериментов

Увеличивающиеся с каждым годом общие объемы ChIP-seq экспериментов, хранящиеся в специализированных базах данных: ChIP-Atlas (Oki et al., 2018), CistromeDB (Zheng et al., 2019), GTRD (Yevshin et al., 2019), ReMap (Chèneby et al., 2020) и прочих, поднимают вопросы о унификации анализа качества, обработки и обобщения имеющихся данных. Данные задачи могут решаться как в рамках анализа пиков идентифицированных в рамках одного эксперимента с несколькими репликами, так и в рамках массового анализа большого количества ChIP-seq экспериментов для конкретного ТФ.

Для первого случая, консорциумом ENCODE было предложено использование метода IDR (Li Q. et al., 2011). Основной задачей метода IDR является оценка воспроизводимости наборов РСТФ, полученных в экспериментах с двумя репликами, и фильтрация наиболее достоверных на основании этой информации. Данный метод основывается на том, что при сравнении двух упорядоченных списков РСТФ ранги истинных РСТФ будут обладать высокой корреляцией, а шум будет демонстрировать её отсутствие. Таким образом, найдя точку (IDR threshold), в которой наблюдается значительное увеличение

согласованности между рангами и репродуцибельностью можно отделить сигнал от шума. Данный подход также имеет свои ограничения.

В частности, данный метод чувствителен к ситуациям, когда одна из реплик обладает более низким качеством (обладает меньшим отношением сигнала к шуму), а также чувствителен к наличию пиков с одинаковыми рангами в исследуемых наборах пиков (Yang Y. et al., 2014). Также, данный метод оценивает взаимоотношение только между парой наборов пиков, т. е. при повышении числа наборов необходимо либо выбирать для дальнейшего анализа только два набора, либо разрабатывать дополнительные стратегии попарного применения данного подхода с последующим обобщением результатов.

В качестве альтернативы данному подходу, Yajie Yang с соавторами в 2014 (Yang Y. et al., 2014) было рассмотрено использование правила большинства (если > 50% включают в себя рассматриваемый РСТФ, то данный РСТФ считается истинным) для обобщения результатов нескольких реплик ChIP-seq экспериментов, если количество реплик > 2. Таким образом, если истинный РСТФ был потерян в одном наборе данных, данный РСТФ может быть восстановлен за счёт привлечения других реплик. Было показано, что использование данного подхода позволяет получать более надежные наборы РСТФ, по сравнению с наборами пиков, получаемых на основании метода IDR, примененного для всех возможных пар реплик. Данный подход также может быть использован для нахождения наиболее достоверного набора пиков на основании набора различных экспериментов для рассматриваемого ТФ. Однако, использование правила большинства для анализа большого количества экспериментов может привести к повышению количества ложноотрицательных РСТФ, в особенности, по отношению к менее представленным клеточным типам и экспериментальным условиям.

Использование экспериментов с большим количеством биологических реплик, а также дополнительных экспериментов для рассматриваемого ТФ, может

повысить надёжность идентификации РСТФ. Например, в рамках базы данных ReMap (Chèneby J. et al., 2020) был проведен массовый анализ ChIP-seq экспериментов с целью выявления районов скопления РСТФ. Для каждого ТФ было проведено пересечение всех доступных в базе данных ChIP-seq экспериментов, таким образом сформировав кластеры РСТФ. Затем на основе средних значений координат РСТФ определяются границы данных кластеров (non-redundant peaks). Таким образом каждому кластеру присваивалось значение, соответствующее количеству входящих в кластер РСТФ из разных ChIP-seq экспериментов. Подобный анализ по обобщению ChIP-seq экспериментов и построению кластеров РСТФ (мета-кластеров) также проводится в рамках базы данных GTRD (Kolmykov et al., 2020). Однако данный подход не дает достаточной информации для вычисления порогового значения для последующего выделения набора наиболее достоверных РСТФ.

1.5 Влияние однонуклеотидных геномных вариантов на регуляцию транскрипции

Выявление и интерпретация однонуклеотидных геномных вариантов (SNV) является важной задачей при анализе индивидуальной генетической информации, поскольку они могут играть важную роль в формировании различных заболеваний и сложных фенотипических признаков (Hirschhorn J. N. et al., 2020). В обзоре Merkulov с соавторам, посвященному роли SNV в транскрипционной регуляции, отмечается особая важность цис-регуляторного эффекта некодирующих SNV, как одного из основных факторов фенотипической вариабельности сложных признаков (Merkulov et al., 2018).

Рост числа секвенированных персональных геномов человека позволил установить ассоциации между SNV и вариативностью уровней экспрессии как проксимально, так и дистально лежащих генов (cis-eQTL, trans-eQTL). На данный момент крупнейшим проектом, направленным на изучение тканеспецифичной экспрессии генов и нахождение ассоциаций между генетическими вариантами и изменением экспрессии генов человека, является GTEx (Genotype-Tissue Expression) (GTEx Consortium et al., 2017). В работе (Gaffney D. J. et al., 2012) было показано, что приблизительно 40% используемых в исследовании eQTL SNV располагаются в области открытого хроматина, в частности, в РСТФ. Также была произведена попытка использования информации о доступности хроматина, модификации гистонов, локализации РСТФ, полученных из ChIP-seq данных, для оценки функциональной значимости SNV (Gaffney D. J. et al., 2012).

Однако оценка функциональной значимости SNV все еще является сложной задачей, особенно по отношению к генетическим вариантам, расположенным в регуляторных районах генов, что привело к развитию различных вычислительных подходов оценки способности таких SNV влиять как на эффективность связывания ТФ, так и, в общем, приводить к изменениям уровней экспрессии близлежащих генов. Например, одним из методов оценки потенциала SNV к изменению уровней экспрессии близлежащих генов является метод FIRE (Functional Inference of Regulators of Expression), основанный на алгоритме случайного леса (random forest). На кросс-валидации при решении задачи дискриминации cis-eQTL SNV и нефункциональных SNV показатель точности (AUC) на обучающей выборке был равен 0,807 (Ioannidis et al., 2017). В 2017 году была предпринята попытка интеграции данных eQTL и GWAS с данными о доступности хроматина, в частности, применение метода Wellington для de novo предсказания ССТФ. Было показано, что большинство некодирующих SNV в тканях мозга локализуются в функциональных районах генов, но вне предсказанных ССТФ, что может свидетельствовать о недостаточной

чувствительности выбранного метода предсказания. Также подобные результаты могут указывать на сложность предсказания ССТФ по данным о доступности хроматина в столь динамичной системе. Используемый в данной работе подход также не способен дать полной картины относительно ТФ, способных к ремоделированию структуры хроматина (Handel et al., 2017).

SNV, расположенные в ССТФ, могут изменять аффинность ТФ к ДНК, что приводит либо к повышению, либо к снижению транскрипционной активности генов-мишеней. Активное применение различных высокопроизводительных методов исследований (ChIP-seq, DNase-seq, WGS, WES и др.) привело к накоплению больших объемов данных, позволяющих не только выявлять SNV, лежащих в основе формирования аллель-специфичного связывания ТФ (на основании данных ChIP-seq и WGS), но и создавать более сложные модели предсказания влияния некодирующих SNV на интенсивность связывания ТФ. Следует отметить, что изменение интенсивности связывания ТФ может быть обусловлено не только наличием замен нуклеотидов, но и различиями в эпигеноме и нарушениями связывания кооперативных ТФ (Kasowski et al., 2010).

В настоящее время существует несколько вычислительных методов оценки влияния некодирующих SNV на интенсивность связывания ТФ с функциональным сайтом, однако, практически все они обладают относительно невысокой точностью предсказания. Например, были разработаны подходы, основанные на определении разницы между интенсивностями связывания ТФ, базирующихся на последовательности ДНК, для двух аллелей (is-rSNP (Macintyre et al., 2010), BayesPI-BAR (Wang, Batmanov, 2015) и deltaSVM (Lee et al., 2015)). В свою очередь, методы, основанные на известных некодирующих полиморфизмах, ассоциированных с заболеваниями человека, такие как GWAVA (Ritchie et al., 2014), FunSeq2 (Y et la., 2014) и DeepSEA (Zhou et Troyanskaya, 2015), существенно ограничены небольшим объемом данных для обучения. Метод CADD (Kircher et al., 2014), основанный на эволюционной консервативности,

показывает низкую способность к выявлению некодирующих SNV. Одними из наиболее предпочтительными методами для детерминация функциональных и нефункциональных SNV являются три родственные классификационные модели (DeFine-regression, DeFine-classification и DeFine-combined), основанные на предсказаниях интенсивности связывания, которые, в свою очередь, предсказываются с помощью алгоритма DeFine (Deep learning based Functional impact of non-coding variants evaluator).

Описание алгоритма DeFine, использующего нейронные сети, а также сравнительный анализ трех классификационных моделей (DeFine-regression, DeFine-classification и DeFine-combined) с CADD, GWAVA и DeepSEA приведено в (Wang, 2018). Сравнительный анализ проводился независимо друг от друга на трех типах наборов данных: HGMD (Human Gene Mutation Database), GWAS и eQTL. С одной стороны, сравнительный анализ однозначно свидетельствует о том что наиболее точной классификационной моделью оказалась модель DeFine-combined, у которой показатели точности - значения AUC (Area Under Curve) - оказались наибольшими на всех трех используемых наборах данных HGMD, GWAS и eQTL: 0.847, 0.652, 0.619. С другой стороны, такие значения AUC однозначно свидетельствуют об умеренной точности (AUC = 0.847 на множестве HGMD) и весьма посредственной (AUC = 0.652 и 0.619 на множествах GWAS и eQTL). Другими словами, задача создания существенно более точных моделей предсказания является достаточно актуальной.

Также в 2021 году Абрамовым с соавторами был представлен алгоритм идентификации аллель-специфичного связывания ТФ (Allele-specific binding; ASB) на основании массового анализа ChIP-seq данных из БД GTRD, а также БД с результатами применения данного алгоритма, ADASTRA. Суть данного алгоритма заключается в поиске диплоидных SNV в ChIP-seq пиках и последующий анализ распределения количества прочтений между двумя аллелями (см. Рисунок 1.5.1) (Abramov et al., 2021).

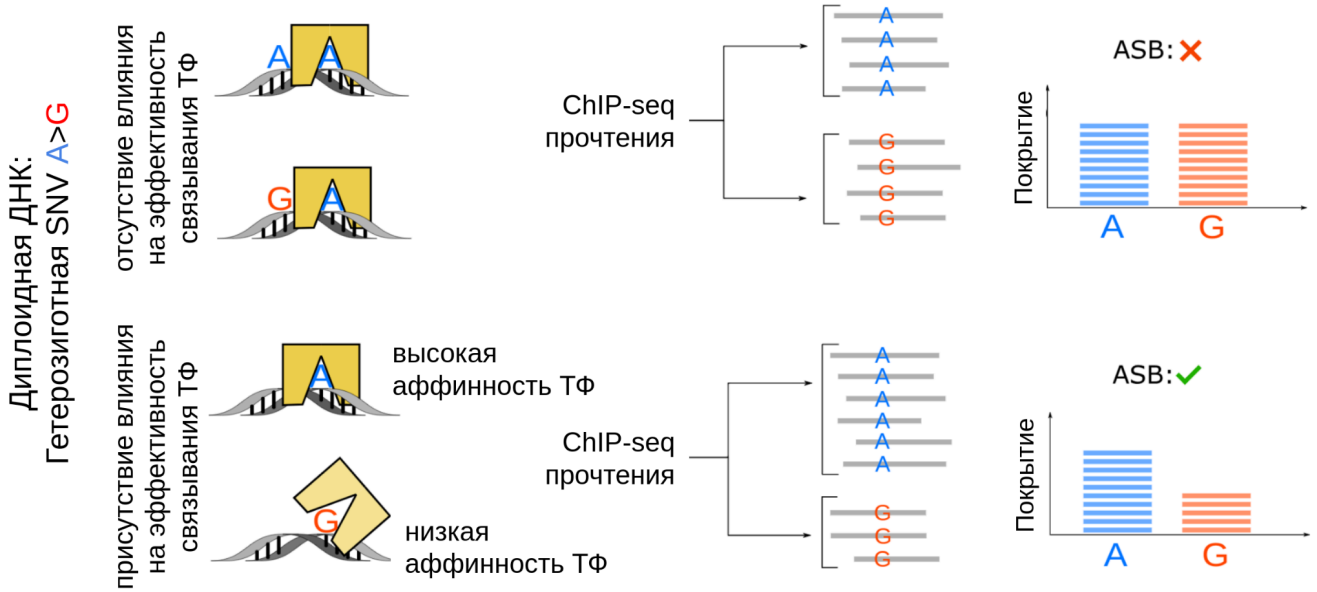


Рисунок 1.5.1 – Обобщенная схема идентификации событий аллель-специфичного связывания ТФ (ASB) на основании анализа ChIP-seq данных (Abramov et al., 2021).

1.6 Морфология сперматозоидов

Сперматозоиды являются одними из самых маленьких клеток в организме человека, диаметром от 2,5 до 4 мкм. На рисунке 1.6.1 изображена схема строения зрелого сперматозоида (Белякова с соавт., 2006). Сложность строения сперматозоида обусловлена его узкой специализацией. Зрелый сперматозоид состоит из головки, шейки, промежуточного (среднего) отдела и хвоста.

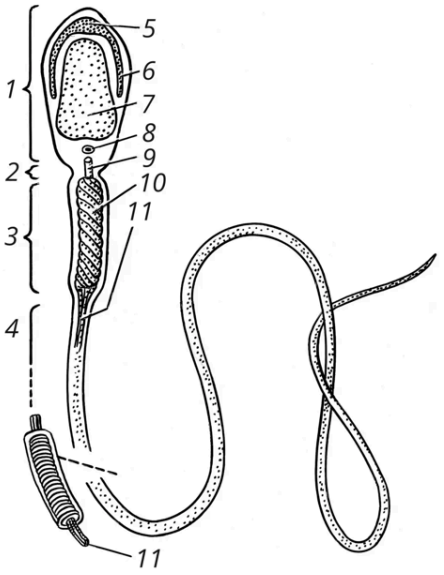


Рисунок 1.6.1 — Схема строения нормального сперматозоида. 1 – головка; 2 – шейка; 3 – промежуточный (средний) отдел; 4 – хвост; 5 – акросома; 6 – головной чехлик; 7 – ядро; 8 – проксимальная центриоль; 9 – дистальная центриоль; 10 – спираль из митохондрий; 11 – осевая нить

Головка сперматозоида человека содержит плотно упакованный генетический материал, состоящий из гаплоидного набора хромосом. В передней её части располагается мембранный пузырек, акросома. Акросома содержит в себе гидролитические ферменты, играющие главную роль в растворении оболочки ооцита при оплодотворении. В шейке сперматозоида располагаются две центриоли, лежащих под прямым углом друг к другу; от проксимальной пары центриолей берут начало микротрубочки, участвующие в образовании хвоста и формирующие осевую нить жгутика. В средней части вокруг осевой нити располагается спиралевидное скопление митохондрий, называемое митохондрионом. Хвост сперматозоида представляет собой длинный жгутик, который составляет около 90% от общей длины сперматозоида, состоящий из пары центральных микротрубочек, окруженной девятью парами периферических микротрубочек (Грин с соавт., 2004).

Морфологические аномалии мужских половых клеток могут препятствовать успешному проникновению клеток через цервикальный канал. В работе Клещёва с соавторами описываются различные аномалии морфологии сперматозоидов. На рисунке 1.6.2 представлены некоторые примеры таких аномалий (Kleshchev et al., 2023).

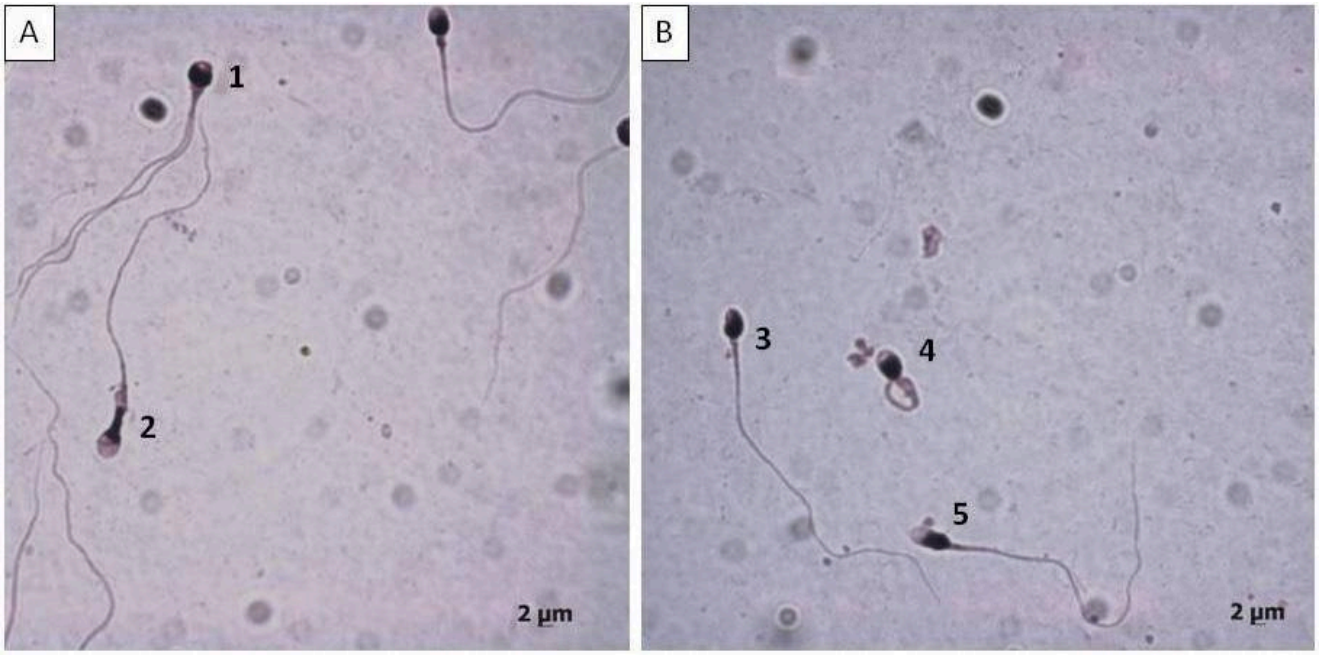


Рисунок 1.6.2 — Некоторые аномалии морфологии сперматозоидов. (А) круглая головка с аномально маленькой акросомой и двойным хвостом; 2 – грушевидная головка. (Б) 3 – аморфная головка с аномальной (маленькой) акросомой; 4 – аномально маленькая акросома и скрученный хвост; 5 – удлиненная головка.

Среди аномалий морфологии сперматозоидов часто встречаются изменения формы головки, такие как: круглая, тонкая, овальная, грушевидная или сдвоенная формы. Также возможны изменения размеров акросомы или присутствие в ней посторонних вакуолей. Аномалии шейки проявляются в виде: перекручивания, утолщения, истончения, или нарушения однородности консистенции, и могут приводить к не прочному соединению головки с хвостом, часто выражающемуся в присутствии в эякуляте хвостов без головок. Аномалии развития хвоста сперматозоида могут быть представлены в виде его отсутствия или чрезмерной

изогнутости, а также увеличения или уменьшения его длины (Kleshchev et al., 2023).

Всемирная организация здравоохранения в 2010 году утвердила схему оценки морфологических показателей сперматозоидов, предложенной Kruger T. F. в 1987 году (Kruger et al., 1987). Данный подход заключается детальном исследовании морфологических признаков сперматозоидов. Следует отметить, что в эякуляте всегда присутствует большое количество сперматозоидов, демонстрирующих различные отклонения морфологии; нормой считается присутствие в эякуляте более 4% нормальных сперматозоидов, способных к оплодотворению. Характеристики нормального сперматозоида по Крюгеру:

- Головка овальная, длиной 4–5 мкм, шириной 2,5–4 мкм.
- Головка соединяется с шейкой под углом в 90 градусов.
- Акросома занимает от 40% до 60% площади головки.
- Шейка ровная, тонкая, длиной 7–8 мкм, максимальная толщина — 1 мкм.
- Цитоплазматическая капля на шейке должна составлять менее трети размера головки.
- Хвост должен быть один, извитый, длиной ~45 мкм (90% длины всего сперматозоида).

Исследование механизмов, лежащих в основе нарушений морфологии сперматозоидов, затрудняется сложностью процесса созревания данных клеток (Green et al., 2018). Данный процесс, называемый сперматогенезом, протекает в извитых канальцах семенников и включает в себя этапы: пролиферации сперматогониев, дифференцировки в сперматоциты, мейотического деления сперматоцитов, созревания круглых сперматид и спермиогенез (созревания узкоспециализированных сперматозоидов) (см. Рисунок 1.6.3).

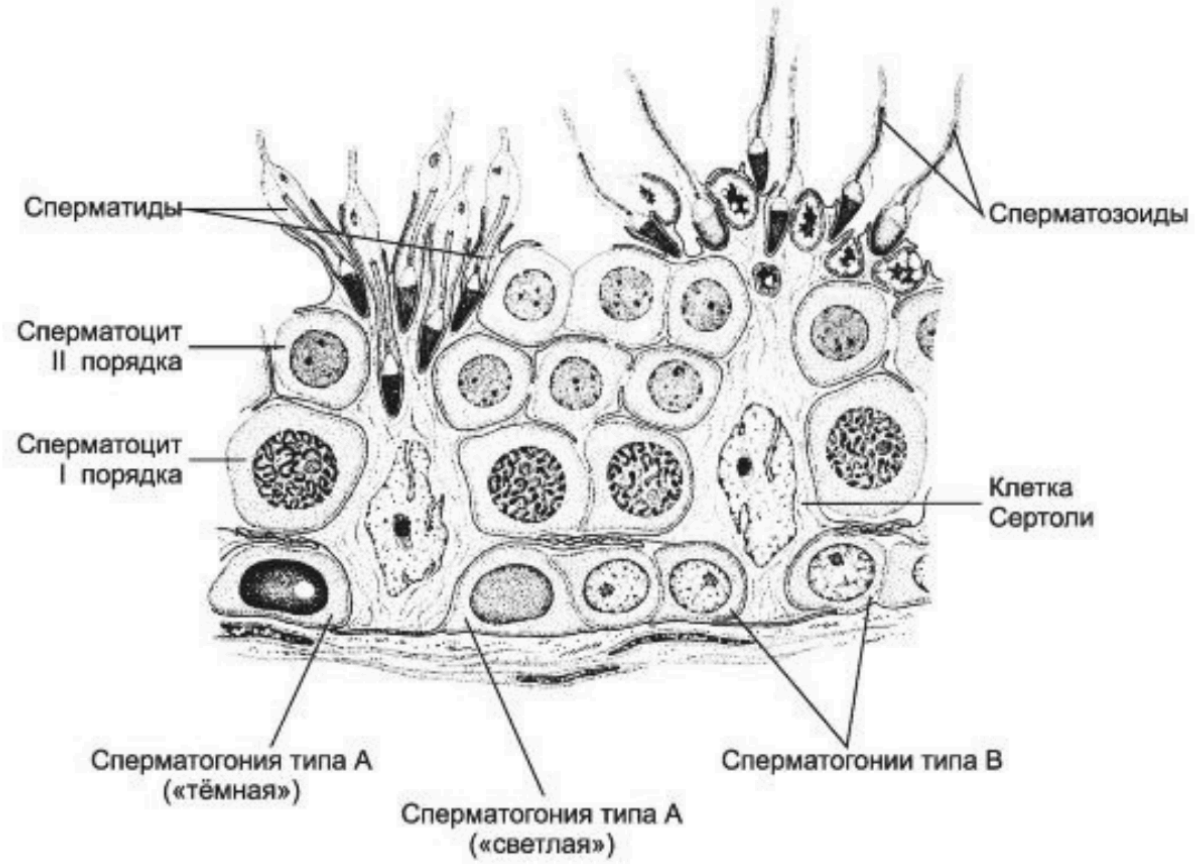


Рисунок 1.6.3 – Сперматогенез в извитых семенных канальцах

ТФ являются неотъемлемой частью регуляции различных этапов сперматогенеза, в частности, в обеспечении нормальной морфологии сперматозоидов (Green et al., 2018). Нарушения в функционировании или изменения уровней их экспрессии могут приводить к дефектам формы головки, формирования жгутиков и средней части сперматозоидов, являясь причиной формирования мужского бесплодия (Du et al., 2021; Cannarella et al., 2020; Blanco et Cocquet, 2019; Silva et al., 2015; Vernet et al., 2016).

Следует отметить, что в настоящий момент влияние событий аллель-специфичного связывания ТФ в контексте нарушения процессов сперматогенеза, в частности, их влияния на морфологию сперматозоидов, да этого не изучалось.

1.7 Определение чувствительности к ДНКазе I (DNase-seq)

В настоящее время наиболее популярным экспериментальным методом полногеномного картирования РСТФ является ChIP-seq. Однако данный метод позволяет провести полногеномное картирование районов связывания в одном эксперименте только одного ТФ и требует использования высококачественных специфичных для исследуемого белка антител. В свою очередь, методы, позволяющие картировать районы открытого хроматина (РОХ) (такие как DNAse-seq, ATAC-seq и FAIRE-seq) дают возможность осуществить полногеномное картирование предполагаемых районов взаимодействия всех ТФ с ДНК в одном эксперименте.

Одним из методов выявления районов открытого хроматина (РОХ) в геноме является секвенирование гиперчувствительных сайтов к ферменту ДНКаза I (DNase-seq). Эти регионы часто связаны с регуляторными элементами, такими как энхансеры, промоторы и другие сайты связывания транскрипционных факторов. DNase-seq позволяет реконструировать регуляторный ландшафт генома путем картирования районов, где хроматин более доступен для взаимодействия с ферментом ДНКаза I (Song, Crawford, 2010) (см. Рисунок 1.7.1).

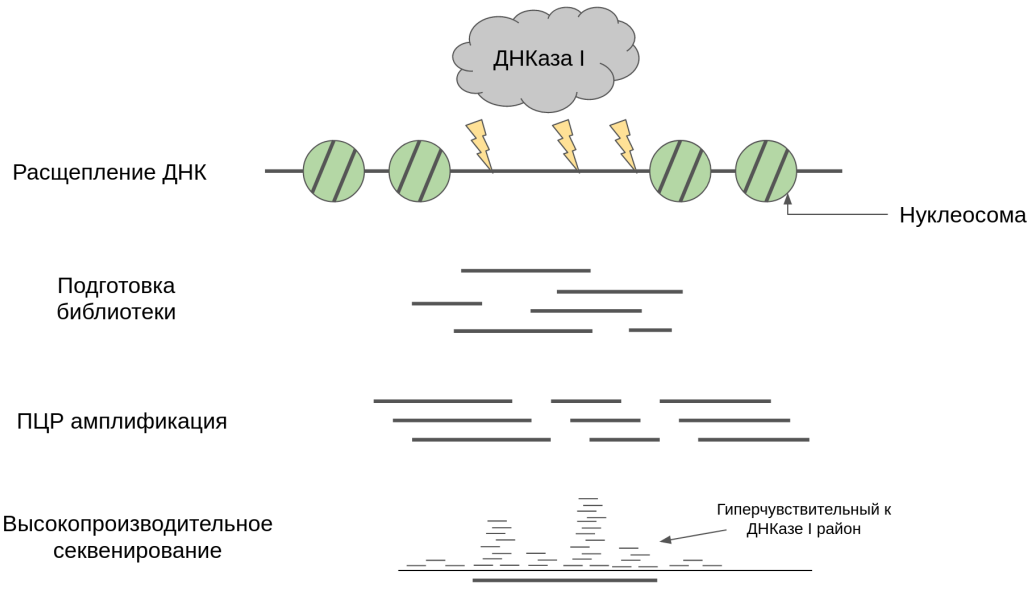


Рисунок 1.7.1 - Упрощенная схема DNase-seq эксперимента

Поскольку размер свободных от хроматина участков ДНК может достигать нескольких сотен пар оснований, выделение в них предполагаемых сайтов связывания ТФ является нетривиальной задачей, что привело к развитию вычислительных методов предсказания сайтов связывания ТФ на основании доступности хроматина.

С 2010 года было разработано множество методов предсказания сайтов связывания ТФ по данным доступности хроматина (например, CENTIPEDE (Pique-Regi R. et al. 2011), MILLIPEDE, PIQ (Sherwood R. I. et al., 2014), Wellington (Piper J. et al., 2013), BinDNase (Kähärä J. et al., 2015), DeFCoM (Quach B. et al., 2017), Romulus (Jankowski et al., 2016). В подавляющем большинстве данные методы работают только с данными DNAse-seq, демонстрируя более низкое качество предсказания функциональных районов при использование данных ATAC-seq (Wang et al., 2018). Все многообразие данных методов можно разделить на две группы: мотив-центрические и *de novo* методы. Часть методов предсказания сайтов связывания ТФ *de novo* базируется на распознавании небольших участков профилей DNase-seq данных, или футпринтов, обладающих пониженной чувствительностью к ДНКазе I и располагающихся в регионах с высокой активностью данного фермента (Neph et al., 2012, Wellington и DNase2TF (Sung M. H. et al., 2014)). Другая группа *de novo* методов основывается на применении скрытых марковских моделей (Boyle et al., 2011 и HINT (Gusmao E. G. et al., 2014)). Мотив-центрические методы основываются на анализе профиля доступности хроматина вокруг потенциального сайта связывания ТФ, найденного на основании известного мотива. Большинство мотив-центрических методов являются представителями кластерного анализа с использованием смеси распределений (FLR (footprint likelihood ratio; Yardimci et al., 2014)), байесовской смеси распределений (CENTIPEDE) и комбинации гауссовского процесса и распространения ожидания (PIQ). Помимо этого, построены различные

классификационные модели, такие как логистическая регрессия (BinDNase) и нелинейные классификационные модели (например, DeFCoM).

Таким образом, на данный момент существует больше 10 методов для предсказания функциональных сайтов по данным о доступности хроматина. В 2016 году были произведены две попытки систематического сравнения этих методов (Quach, Furey, 2016; Gusmao et al., 2016). Результаты двух сравнений оказались достаточно противоречивы, особенно при сравнении мотив-центрических методов. Это говорит о том, что на тот момент еще не выработаны универсальные критерии (меры сравнения) методов подобного типа и отсутствовали универсальные выборки как для их обучения, так и для их тестирования.

1.8 Методы коллективного выбора

В связи с постоянно растущим количеством информации связанной с результатами различных типов полногеномных экспериментов, важной задачей является её интеграция и последующий анализ в контексте решения той или иной задачи. Однако одной из главных проблем является разнородность имеющейся информации. Например, некоторые экспериментальные данные не могут рассматриваться вместе по многим причинам: различия в использованных методах и дизайне экспериментов, или различия в качестве исходных данных и их последующем анализе. Одним из важнейших инструментов мета-анализа способным помочь в решении данной задачи являются методы коллективного выбора. Суть данных методов заключается в том, чтобы привести имеющиеся значения признаков в каждом из экспериментов к ранговым переменным, а затем на основании полученных упорядоченных списков обобщить имеющуюся информацию. Такой подход позволяет значительно снизить влияние шума на

конечный результат (Li et al., 2017; Lin, 2010).

В конце двадцатого века в связи с появление интернета и развитием поисковых систем началось активное развитие новых методов коллективного выбора. Появившиеся в то время методы основывались на использовании значений, присужденных элементу из упорядочиваемого множества каждым из ранжировщиков для составления искомой агрегирующей функции. Однако поскольку в большинстве случаев была доступна только информация о положении элемента в наборе упорядоченных списков, стали появляться методы, основывающиеся только на этой информации. Несмотря на меньшую производительность, данная группа методов стала приобретать наибольшую популярность (Randa, Straccia, 2003). В частности, были предложены методы, использующие различные модификации метода Борда. Впоследствии были разработаны более производительные методы коллективного выбора основанные на использовании Марковских цепей (Dwork C. et al., 2001), нечёткой логики (Beg et al., 2004), генетического алгоритма (Pihur et al., 2009), теории графов и др. В данной группе методов наибольшей производительностью обладают методы основанные на марковских цепях. В частности, набор методов, предложенных Dwork с соавторами (Dwork et al., 2001), а также их модификации, активно использовались для решения задач в различных областях, в том числе и биоинформатике (Li et al., 2017).

Одной из главных проблем методов коллективного выбора можно назвать наличие различающихся по качеству источников упорядоченных списков. Например, при мета-анализе биологических данных, можно наблюдать различия в качестве получения как экспериментальных данных, так и их последующего анализа, что может существенно повлиять на результат применения методов коллективного выбора. Существует несколько способов решения этой проблемы. В работе Ling и Ding 2009 (Ling, Ding, 2009), описывается приписывание весового коэффициента каждому источнику упорядоченных списков, однако не было

предложено прозрачной процедуры присвоения весовых коэффициентов. Также для решения этой проблемы возможно использование методов, основанных на обучении с учителем. Ограничением в данном случае является наличие адекватной обучающей выборки достаточного объема, в большинстве случаев отсутствующей в контексте решения биологических задач. Альтернативой данным подходам являются методы, основанные на байесовской статистике (Deng et al., 2014; Badgeley et al., 2015; Li et al., 2018).

Другой проблемой, связанной с использованием методов коллективного выбора, является ориентированность большинства методов на работу с полными списками, т. е. упорядоченными списками, содержащими в себе информацию о всех элементах исследуемого множества. Тем не менее, на практике исследователи чаще встречаются либо с неполными списками, представляющими лишь подмножество исследуемого множества элементов, что может быть следствием качества проб-подготовки, используемой методологии и последующего анализа эксперимента; либо с частично отсортированными списками, когда можно упорядочить только часть элементов, а положение остальных элементов задать относительно уже упорядоченных (Li et al., 2018). В 2017 году Li с соавторами провели сравнительный анализ методов коллективного выбора и продемонстрировали влияние различных типов списков на производительность методов (Li et al., 2017).

На сегодняшний день существует большое количество методов коллективного выбора. В обзоре Li с соавторами 2017 года предлагается классификация данных методов, которая будет использована далее при детальном рассмотрении описанных выше методов (рис. 1.8.1).

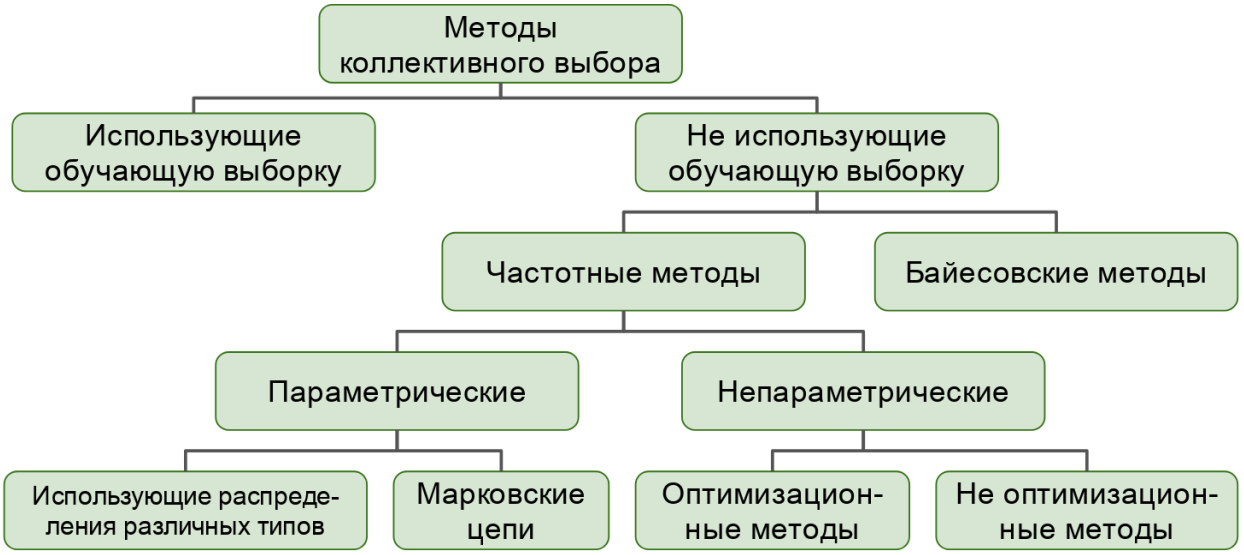


Рисунок 1.8.1 – Классификация методов коллективного выбора по Li с соавторами (Li et al., 2017)

1.8.1 Непараметрические методы

1.8.1.1 Неоптимизационные методы

Основной отличительной чертой не оптимизационных методов можно назвать их прозрачность и интуитивность. К данной группе методов относится группа методов, использующая основные статистические показатели (арифметическое и геометрическое среднее, медиана, L2-norm) для присуждения конечного ранга каждому элементу. К данной группе методов относится метод Борда (Lin, 2010). Пусть $\{\tau_l(i)\}_{1\le l\le L}$ - ранг элемента i в каждом из L упорядоченных списков. Тогда агрегирующая функция f(i) может принимать вид:

$$f(i) = median\{|\tau_1(i)|,\ldots, |\tau_l(i)|,\ldots, |\tau_L(i)|\},$$ в случае медианы;

$$f(i) = \left(\prod_{l=1}^{L} |\tau_l(i)|\right)^{1/L},$$ в случае среднего геометрического;

$$f(i) = \sum_{l=1}^{L} \frac{|\tau_l(i)|^p}{L},$$ в случае p-norm.

Различные модификации метода Борда показывают хорошие результаты при работе с близкими по качеству полными списками, но демонстрируют малую устойчивость к влиянию списков с низким качеством (Lin, 2010).

Также к не оптимизационным методам можно отнести некоторые методы отбора признаков (Haury et al., 2011). Например, суть метода Stability Selection (Meinshausen, Bühlmann, 2010) заключается в присвоении более высокого ранга элементу обладающему большим числом рангов превышающих установленный порог в исходных списках. Данный подход обладает большей устойчивостью к наличию списков с низким качеством.

1.8.1.2 Оптимизационные методы

Суть оптимизационных методов состоит в минимизации значения величины, характеризующей несогласованность между набором исходных упорядоченных списков и конечным ранжированием. В качестве оптимизируемой величины зачастую используют либо расстояние Кендалла (Kendall's tau distance) (Kendall M. G., 1938), либо расстояние Спирмена (Spearman's footrule distance) (Spearman C., 1961). В отличии от методов Борда при использовании этих непараметрических мер расхождения оцениваются парные отношения внутри набора исходных списков.

Пусть $\tau_1,...,\tau_L$ - набор исходных упорядоченных списков, тогда расстояние Кендалла для пары элементов в двух списках можно выразить как:

$$d_K(u,v) = I[(R_{\tau_1}(u) - R_{\tau_1}(v))(R_{\tau_2}(u) - R_{\tau_2}(v)) < 0],\ i,j = 1,...,|T|$$

Где *I()* - характеристическая функция, принимающая значение 1, если неравенство выполняется, и 0 - в обратном случае, а $R_{\tau_l}(n)$- ранг элемента n в списке l. Таким образом расстояние Кендалла между двумя списками будет равно:

$$K(\tau_1, \tau_2) = \sum_{i,j} d_K(i, j)$$

В случае с неполными списками при сопоставлении положения пары элементов в двух списках, в случае отсутствия элемента в одном из списков, в качестве ранга элемента в данном списке будет взято значение k + 1, где k количество элементов в данном списке. Расстояние между парой элементов u и v в двух списках можно выразить следующим образом:

$$d_k(u, v) = \begin{cases} I\{[R_{\tau_1}(u) - R_{\tau_1}(v)] \\ \quad [R_{\tau_2}(u) - R_{\tau_2}(v)] < 0\} & \text{если } (u, v) \in B^c \\ p & \text{в противном случае,} \end{cases}$$

Где p - параметр, принимающий значение от 0 до 1. Благодаря варьированию параметра p можно настраивать чувствительность системы к потере информации о ранге элемента в списке (Lin, 2010).

Одним из способов оптимизации значения выбранной меры расхождения является использование методов стохастической оптимизации. В частности, в работе Lin и Ding (Ling, Ding, 2009) описывают алгоритм явного упорядочивания (Order Explicit Algorithm; OEA), основанный на методе перекрестной энтропии, относящийся к группе методов Монте-Карло (cross-entropy Monte Carlo; CEMC) (Rubinstein and Kroese, 2004). В своей работе исследователи продемонстрировали эффективность данного метода при работе с неполными списками на примере отбора мРНК-мишеней для микроРНК-155 на основе нескольких списков потенциальных мишеней данной микроРНК, полученных в результате применения различных программных пакетов.

1.8.2 Параметрические методы

1.8.2.1 Марковские цепи

Dwork с соавторами в 2001 году одними из первых описали применение марковских цепей в контексте решения задач коллективного выбора (Dwork et al., 2001). Суть метода метода заключается в построении матрицы переходных вероятностей $P = \{P(u \rightarrow v)\}_{u,v \in U}$, где $P(u \rightarrow v)$ - вероятность перехода ранга элемента из u в v, таким образом, чтобы в стационарном распределении марковской цепи присваивалась большая вероятность элементу с более высоким рангом. Методы, основанные на использовании марковских цепей, также, в отличие от методов Борда, опираются на парные отношения внутри набора исходных списков. В обзоре Lin 2010 года приводится несколько вариантов построения матрицы переходных вероятностей, которые можно считать вариациями алгоритмов, описанных Dwork с соавторами и DeConde с соавторами (DeConde et al., 2006):

MC1:

Для каждой пары элементов u и v ($u \neq v$) из исследуемого множества S:

$$P(u \rightarrow v) = \begin{cases} 1/|S| & \text{если } R_{\tau_l}(u) > R_{\tau_l}(v) \text{ хотя бы для одного исходного списка} \\ 0 & \text{в противном случае,} \end{cases}$$

В случае, когда u = v, вероятность перехода рассчитывается по формуле:

$$P(u \rightarrow u) = 1 - \sum_{u \neq v} P(u \rightarrow v)$$

Таким образом, при построении матрицы переходных вероятностей по

данному алгоритму, для присуждения ненулевой вероятности перехода цепи из элемента u в v необходимо, чтобы элемент u был ранжирован выше элемента v хотя бы в одном из исходных списков.

MC2:

Для каждой пары элементов u и v (u ≠ v) из исследуемого множества S:

$$P(u \to v) = \begin{cases} 1/|S| & \text{если } R_{\tau_l}(u) > R_{\tau_l}(v) \text{ для большинства исходных списков} \\ 0 & \text{в противном случае,} \end{cases}$$

Вероятность перехода для случаев u = v рассчитывается по формуле:

$$P(u \to u) = 1 - \sum_{u \neq v} P(u \to v)$$

Построение матрицы переходных вероятностей данным способом позволяет лучше работать со списками сильно отличающимися по качеству данных, на которых они были основаны, поскольку для присвоения ненулевой вероятности требуется, чтобы элемент обладал большим или равным рангом в половине использующихся списков.

MC3:

Для каждой пары элементов u и v (u ≠ v) из исследуемого множества S:

$$P(u \to v) = \sum_{l=1}^{L} I(R_{\tau_l}(u) > R_{\tau_l}(v))/(L|S|)$$

Вероятность перехода для случаев u = v рассчитывается по формуле:

$$P(u \to u) = 1 - \sum_{u \neq v} P(u \to v)$$

Данный способ построение матрицы переходных вероятностей позволяет учитывать количество исходных списков, в которых элемент u обладает большим или равным рангом, что может быть полезно при интеграции данных, полученных при использовании различных экспериментальных платформ.

Применение данного подхода было описано DeConde с соавторами в 2006

году. В частности, было описано успешное применение алгоритма построение матрицы переходных вероятностей MC3 для последующего анализа разнородных данных микрочип-экспериментов, направленных на выявление дифференциально экспрессирующихся генов при раке простаты.

1.8.2.2 Методы, использующие различные типы распределений

Stuart с соавторами (Stuart et al, 2003) предложили использовать значение коэффициента корреляции Пирсона для идентификации и ранжирования пар генов, коэкспрессирующихся у различных организмов. Значение p-value рассчитывалось на основании распределения порядковых статистик. Позднее, Aerts с соавторами (Aerts et al, 2006) была предложена модификация данного метода. Однако использование данного метода для набора неполных списков является затруднительным.

Другим представителем данной группы методов является описанный Kolde с соавторами метод robust rank aggregation (RRA) (Kolde et al, 2012). В отличие от описанных выше методов коллективного выбора, направленных на минимизацию несогласованности окончательного ранжирования со всеми исходными списками, данный метод основывается на утверждении, что каждый список является истинным только относительно ограниченного набора элементов, так как описывает тот или иной аспект фиксируемого биологического процесса. Для того, чтобы выделить это подмножество элементов, вводится нулевая модель, которая описывает случайное распределение рангов элементов, т. е. предполагается что все исходные списки неинформативны, а элементы в них ранжированы случайным образом. Алгоритм принимает во внимание положение элемента в наборе исходных списков, сравнивает с нулевой моделью и присуждает каждому элементу значение p-value. Пусть $r=\{r_1,...,r_n\}$ - упорядоченное множество из n

элементов, где $r_1 \leq \cdots \leq r_n$, а $\beta_{k,n}(r)$ - вероятность получить $r'_k \leq r_{k,}$, где r' - случайно упорядоченное множество элементов сгенерированное нулевой моделью. Тогда вероятность того количество наблюдаемых неравенств $r'_k \leq r_{k,}$, что будет меньше или равна n - k, может быть выражена через биномиальное распределение:

$$\beta_{k,n}(x) = \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l},$$

где $x=P(r'_k \leq r_{k,})$. Также значение $\beta_{k,n}(x)$ может быть выражено через бета-распределение, поскольку r'_k - является порядковой статистикой для n значений равномерно распределенных в интервале [0; 1]. Поскольку количество истинных (информативных) рангов неизвестно, в данной работе финальное значение, используемое для искомого ранжирования, определяется как минимальное значение p-value:

$$p(r) = min_{k=1,\ldots,n} \beta_{k,n}(r)$$

Искомое ранжирование представляет собой упорядочивание по значению p-value после применения поправки на множественное сравнение Бонферрони.

1.8.3 Байесовские методы

Байесовские методы основаны на расчёте апостериорной вероятности для определения положения элемента в упорядоченном списке. Одной из основных задач методов, основанных на Байесовской статистике, является решение проблемы варьирования качества исходных списков. Например, метод BARD (Bayesian aggregation of rank data) для разделения списков с низким качеством каждому списку присваивает параметр, характеризующий качество данного списка, а все элементы в каждом списке разбивает на два подмножества: шум и сигнал (Deng et al., 2014).

Пусть I_i - характеристическая функция элемента i из исследуемого множества элементов U, которая принимает значение равное 1, если элемент относится к подмножеству элементов, отражающих реальное ранжирование (U_R), и равное 0, если элемент можно отнести к подмножеству элементов с ложным ранжированием, т. е. к шуму (U_B). Также предполагается, что каждый список τ_k можно описать при помощи трёх упорядоченных множеств: τ_k^0, $\tau_k^{1|0}$ и τ_k^1, где τ_k^0 и τ_k^1 отражают ранжирование элементов в подмножествах U_B и U_R соответственно (т. е. $\tau_k^0 = \tau_{k\in U_B}$; $\tau_k^1 = \tau_{k\in U_R}$), а $\tau_k^{1|0}$ представляет относительное ранжирование элементов из подмножества U_R в U_B ($\tau_k^{1|0}(i) = \tau_{k|\{i\}\cup U_B}(i)$). Особенностью метода BARD является использование степенного закона для описания функциональной зависимости ранжирования $k \in U_R$ в U_B, имеющего вид степенной функции:

$$P(\tau_k^{1|0}(i) = t) \propto t^{-\gamma_k},$$

Где γ_k ($\gamma_k > 0$) - степенной коэффициент, демонстрирующий насколько хорошо элементы списка τ_k могут быть разделены на два подмножества. В данном случае коэффициент γ_k используется для оценки качества исходных списков. Однако в большинстве случаев γ_k, как и I, неизвестно, поэтому последующие действия будут направлены на определение значений данных величин. Для нахождения искомых параметров используется формула апостериорной вероятности следующего вида:

$$P(I, \gamma \mid \tau_1,\ldots,\tau_m) \propto P(\tau_1,\ldots,\tau_m \mid I, \gamma)\pi(I, \gamma),$$

где $\pi(I, \gamma)$ - априорная вероятность.

Для оценки качества списков используется апостериорное среднее:

$$\bar{\gamma}_k := \int \gamma_k P(\gamma_k \mid \tau_1,\ldots,\tau_m) d\gamma_k$$

Для ранжирования элементов в списке предлагается использовать частную вероятность вида:

$$p_i := P(I_i = 1 \mid \tau_1,\ldots,\tau_m)$$

Саму же систему можно описать в следующем виде:

$$P(\tau_1,\ldots,\tau_m \mid I,\ \gamma) = \prod_{k=1}^{m} P(\tau_k \mid I,\ \gamma_k) = \prod_{k=1}^{m} P(\tau_k^0, \tau_k^{1|0}, \tau_k^1 \mid I,\ \gamma_k)$$

$$= \prod_{k=1}^{m} P(\tau_k^0 \mid I) \times P(\tau_k^{1|0} \mid I,\ \gamma_k) \times P(\tau_k^1 \mid \tau_k^{1|0};\ I),$$

где $P(\tau_k^0 \mid I)$ и $P(\tau_k^1 \mid \tau_k^{1|0};\ I)$ - равномерные распределения, а

$$P(\tau_k^{1|0} \mid I,\ \gamma_k) = \prod_{k=1}^{m} P(\tau_k^{1|0}\ (i) \mid I,\ \gamma_k)\ \text{, где}$$

$$P(\tau_k^{1|0}\ (i) = t \mid I,\ \gamma_k) \propto t^{-\gamma_k}.$$

1.8.4 Методы использующие обучение с учителем

Для решения проблемы связанной с вариативностью качества исходных списков могут быть использованы методы, основанные на обучении с учителем. В 2007 году Liu с соавторами предложили использовать дополнительную стадию присвоения весовых коэффициентов исходным списков для повышения производительности методов Борда (Supervised Borda Fuse) и методов, построенных на основе Марковских цепей, описанных выше (Liu et al., 2007). Другим методом коллективного выбора, построенным на основе обучения с учителем, относится метод RankBoost (Freund et al. 2003).

Однако, поскольку в большинстве случаев исследователи не располагают обучающей выборкой достаточного размера, данный тип методов обладает меньшей популярностью по сравнению с методами не использующие обучение с учителем.

1.8.5 Сравнение производительности методов коллективного выбора

В 2017 году было проведено сравнение методов коллективного выбора для неполных списков в контексте биологических данных (Li et al., 2017). В сравнении участвовали 13 методов коллективного выбора: методы Борда (среднее арифметическое и геометрическое, медиана, L2-norm), методы CEMC (CEMC.k, оптимизирующий расстояние Кендалла, и CEMC.s, оптимизирующий расстояние Спирмена), методы, основанные на использование марковских цепей (MC1, MC2 и MC3), Байесовские методы (BARD и BIRRA) и методы, использующие распределения различного типа (RRA и Stuart et al.). Вне зависимости от модели и сценария одним из худших методов всегда оказывался метод BIRRA (Badgeley et al., 2014). rGEO, MC2 и MC3 показывали хороший уровень производительности при работе с неполными списками, а tGEO показывал хороший уровень производительности при работе с неполными списками с неполным ранжированием (top-k списки). RRA демонстрировал низкую или среднюю производительность, в отличие от метода Stuart et al., в некоторых случаях показывающих высокий уровень производительности как с неполными, так и с top-k списками. Методы, построенные на стохастической оптимизации, демонстрировали средний уровень производительности во всех рассмотренных сценариях.

1.9 Заключение по обзору литературы

Как следует из обзора литературы, для исследования процессов регуляции транскрипции существует множество высокопроизводительных экспериментальных подходов, основанных на методах NGS. На сегодняшний день в открытом доступе хранятся десятки тысяч различных типов NGS экспериментов. В рамках развития крупных БД ведется активная работа по унификации анализа качества, обработки и обобщения имеющихся данных. В частности, большой интерес представляет проведение массового сравнительного анализа ChIP-seq экспериментов, направленных на идентификацию геномных районов связывания ТФ, как основных компонентов в регуляции транскрипции. Существует множество метрик, описывающих качество ChIP-seq данных. Большой вклад был сделан международным исследовательским консорциумом ENCODE. Однако взаимосвязь качества ChIP-seq данных и согласованности результатов их обработки различными алгоритмами идентификации РСТФ требует дальнейшего изучения.

Значительное количество исследований указывают на ключевую роль ТФ в регуляции различных этапов сперматогенеза, в частности, в обеспечении нормальной морфологии сперматозоидов. Однако остается не изученным вопрос о влиянии событий аллель-специфичного связывания ТФ в контексте данных процессов.

Таким образом, обзор литературы подтверждает актуальность целей и задач, сформулированных в разделе “Введение”.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1. Единообразная аннотация и анализ NGS данных

На первом этапе производится сбор информации (мета-данные) о доступных экспериментальных данных по регуляции транскрипции. В зависимости от источника данных, применялись разные подходы: хорошо структурированная информация из проекта ENCODE собиралась автоматически (программно), в то время как для аннотации данных из GEO была создана специальная программа GEOminer (Yevshin et al., 2019). Данная программа принимает на вход метаданные, описывающие серию экспериментов (GSE), в формате MINiML (MIAME Notation in Markup Language) и предоставляет полученную информацию аннотатору посредством графического интерфейса. Помимо этого, GEOminer имеет набор полей, заполняемых аннотатором на основании имеющейся об эксперименте информации, обеспечивающих единообразность аннотации экспериментов.

Для понимания регуляции транскрипции необходима интеграция различных типов NGS данных по клеточным типам (линиям) и экспериментальным условиям. Для этого была проведена работа по привязке единого словаря клеточных типов БД GTRD к существующими онтологиями клеточных линий и типов Cell Ontology (Diehl et al., 2016), UBER-anatomy ontology (Mungall et al., 2012), Cellosaurus (Bairoch, 2018), Brenda tissue ontology (Gremse et al., 2010), Experimental factor ontology (Malone et al., 2010), Plant ontology (Cooper et al., 2013) (Kulyashov M. et al, 2020). использование этого словаря и онтологии экспериментальных условий при аннотации всех экспериментальных данных в БД GTRD.

В рамках диссертационной работы был расширен функционал программы GEOminer для поддержки аннотации DNase-seq данных, а также интегрированы упомянутые выши онтологии клеточных типов. Для учёта дополнительных параметров постановки экспериментов: условия обработки, стадию развития организма, генотип и др., в GEOminer были введены дополнительные ключи, которые формализуются в виде набора "ключ-значение" и привязываются к соответствующим экспериментам в GTRD.

Далее все данные единым образом обрабатывались и проходили контроль качества, используя сценарии обработки данных для платформы BioUML (Kolpakov et al., 2019) и системы управления распределенными вычислениями e-grid - собственная распределенная вычислительная платформа для параллельной обработки данных на нескольких вычислительных узлах (Kolmykov et al., 2021).

В настоящее время в базе данных GTRD накапливаются данные для 10 видов организмов: *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae, Schizosaccharomyces pombe* и *Arabidopsis thaliana*. В рамках диссертационного исследования акцент сделан на ChIP-seq и DNase-seq данные для человека. В актуальной версии базы данных GTRD (GTRD v21.12) (Kolmykov et al., 2019) содержится 35719 ChIP-seq экспериментов для *Homo sapiens*, охватывающие различные клеточные типы и ткани, а также широкий спектр экспериментальных условий, для 1391 ТФ и кофакторов. Также в работе использовалось 1701 DNase-seq экспериментов.

2.2. Обработка ChIP-seq и DNase-seq экспериментов

Первые стадии обработки используемых типов полногеномных экспериментов общие: оценка качества исходных данных при помощи

программного пакета FastQC, удаление адаптерных последовательностей с использованием программы trimmomatic (Bolger et al., 2014) и выравнивание набора прочтений на референсный геном (GRCh38) при помощи программы bowtie2 (Langmead, Salzberg, 2012) (см. Рисунок 2.2.1).

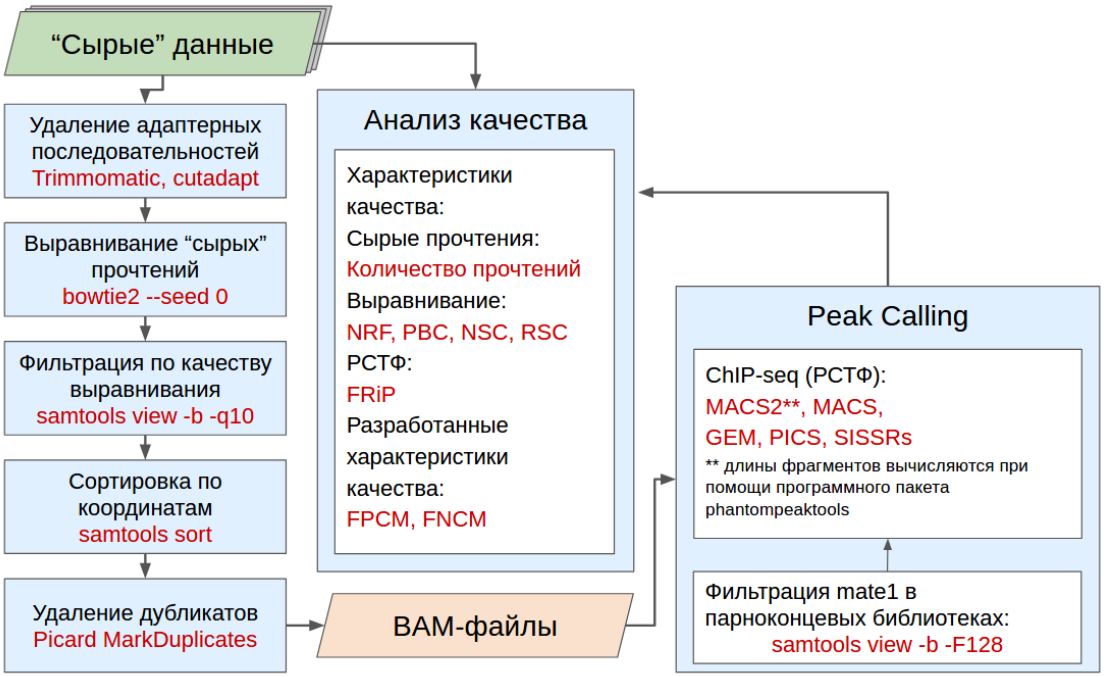


Рисунок 2.2.1 – Схема обработки ChIP-seq экспериментов.

На следующем этапе производится оценка качества ChIP-seq эксперимента: оценка сложности библиотеки (NRF, PBC1 и PBC2) и оценка отношения сигнал-шум при помощи кросс-корреляции (NSC и RSC) (Landt et al., 2012). Для ChIP-Seq экспериментов, библиотека которых представлена парноконцевыми прочтениями, проводится поиск и удаление ПЦР-дубликатов при помощи программного пакета Picard (https://broadinstitute.github.io/picard).

Последующий этап поиска РСТФ проводится при помощи следующего набора методов идентификации пиков: MACS2 (Zhang Y. et al., 2008), GEM (Guo Y. et al., 2012), SISSRs (Narlikar L. et al., 2012) и PICS (Zhang X. et al., 2011). Процесс обработки данных ChIP-Seq экспериментов завершается расчетом доли

прочтений, попавших в полученные РСТФ (FRiP), а также значений FPCM и FNCM (см. далее).

Для выявления районов высокой аффинности к ферменту ДНКаза I были использован алгоритм идентификации пиков MACS2 (Zhang et al., 2008). Вследствие различия в подготовки библиотеки, для single-cut DNase-seq экспериментов MACS2 запускался с параметрами “–nomodel –shift -100 –extsize 200”, для остальных случаев использовались параметры по умолчанию. Завершающий этап обработки DNase-seq экспериментов - идентификация в полученных районах высокой аффинности к ферменту ДНКаза I футпринтов при помощи программы Wellington (Piper et al., 2013).

Для обработки рассматриваемых типов данных были разработаны сценарии для:

1. системы управления распределенными вычислениями eGrid. Данная система была разработана для контроля за распределенной обработкой NGS данных, входящих в состав GTRD). Описываемые выше сценарии реализованы в виде программ, написанных на языке Java, и необходимых для обработки NGS данных программных пакетов.
2. платформы BioUML (Kolpakov et al., 2019) - web-платформы для анализа биомедицинских данных (https://ict.biouml.org). BioUML включает в себя широкий спектр возможностей, включая доступ к базам данных с экспериментальными данными, инструменты для формализованного описания структуры и функционирования биологических систем, а также инструменты для их визуализации, моделирования, подбора параметров и анализа. В данном случае разработанные сценарии реализованы в виде сценариев (workflow) и, в отличие от eGrid, доступны для всех пользователей данной платформы. Каждый пользователь имеет возможность изменять параметры используемых анализов в собственной копии используемого сценария. Также, благодаря наглядному представлению,

workflow и имеющемуся функционалу редактора сценариев, пользователь имеет возможность как изменять структуру, так и строить на их основе собственные сценарии обработки данных.

2.3 Оценка качества ChIP-seq данных

В рамках данной работы на различные этапы конвейера по обработке ChIP-seq данных были добавлены стадии оценки качества полученных данных. Таким образом, для содержащихся в базе данных GTRD v.20.06 21988 ChIP-seq экспериментов (наборов картированных РСТФ) для человека был рассчитан набор различных характеристик качества.

Доля уникальных позиций выравнивания (Non-Redundant Fraction; NRF) отражает сложность библиотеки и рассчитывается как отношение количества районов выравнивания прочтений к количеству выровненных прочтений (см. Рисунок 1.3.1).

Характеристики PCR Bottlenecking Coefficient 1 и 2 (PBC1 и PBC2) также характеризуют сложность библиотеки. Данные характеристика отражают ассимтерию в распределении количества совпадающих выравниваний прочтений в сторону уникально выравниваненных прочтений. PBC1 рассчитывается как отношение количества районов с одним выровненным прочтением к количеству районов выравнивания прочтений. PBC2 - отношение количества районов с одним выровненным прочтением к количеству районов с двумя выровненными прочтениями.

Для разбиения ChIP-seq экспериментов на эксперименты с высоким и низким качеством были использованы рекомендации проекта ENCODE, представленные в таблице 2.3.1.

Таблица 2.3.1 – Референсные значения характеристик качества ChIP-seq экспериментов. Построенные на основании рекомендаций проекта ENCODE

NRF	**Сложность библиотеки**	**PBC1**	**PBC2**	**Уровень ограничения ПЦР**
< 0.5	Сомнительная	< 0.7	< 1	Сильный
0.5 ≤ NRF < 0.8	Приемлемо	0.7 ≤ PBC1 ≤ 0.9	1 ≤ PBC2 ≤ 3	Умеренный
0.8 ≤ NRF < 0.9	Соответствует нормам	0.8 ≤ PBC1 < 0.9	3 ≤ PBC2 < 10	Легкий
> 0.9	Идеально	≥ 0.9	≥ 10	Отсутствует

2.4 Оценка эволюционной консервативности районов связывания транскрипционных факторов

Для оценки эволюционной консервативности РСТФ были использованы результаты работы двух алгоритмов: PhastCons (Siepel et al., 2005) и phyloP (Pollard et al., 2010), доступные в БД UCSC (Kent et al., 2002). В частности, были использованы результаты применения данных алгоритмов на 30 видах: PhastCons30way и phyloP30way. Геномная последовательность была разбита на набор наполовину перекрывающихся между собой участков длиной 200 п.н. (корзины). Для каждой корзины было подсчитано среднее значение консервативность позиций по PhastCons30way и phyloP30way.

2.5 Исследуемая популяция славян

В исследовании принимали участие добровольцы мужского пола из шести городов России и Беларуси: Архангельск, Новосибирск, Кемерово, Улан-Удэ,

Якутск и Минск. Процесс отбора участников подробно описан в работе Osadchuk и соавторов (Osadchuk et al., 2020).

Участникам проводили забор крови для выделения и секвенирования генетического материала, а также сбор семенной жидкости для последующего анализа её характеристик. Формирование популяционных выборок в шести городах, многофакторное фенотипирование показателей сперматогенеза и анкетирование добровольцев осуществлялось сотрудниками Отдела молекулярной генетики человека Сектора прикладных репродуктивных технологий человека ИЦиГ СО РАН. Полноэкзомное секвенирование проводилось на базе Сектора геномных исследований ИЦиГ СО РАН. Исследование поддержано грантами РНФ № 19–15-00075 и № 19-15-00075-П.

Для проведения полноэкзомного секвенирования были выбраны 367 представителей славянской этнической группы. В выборку вошли 174 участников с нормоспермией и 193 участника с различными формами патоспермии: 145 с олигоспермией, 26 с азоспермией, 188 с астеноспермией и 105 с тератоспермией. Некоторые из участников имели сочетанные нарушения качества семенной жидкости.

2.6 Идентификация и анализ однонуклеотидных геномных вариантов

При помощи алгоритма BWA-MEM (Li, 2013) прочтения были выровнены на референсный геном (GRCh38. p13; GCA_000001405.15). Для удаления ПЦР-дубликатов использовался инструмент Picard MarkDuplicates. Средняя глубина покрытия секвенирования для полноэкзомных данных составила 47.5, при этом покрытие >20X для 82% целевых участков.

Идентификация и фильтрация однонуклеотидных вариаций выполнялись в соответствии с рекомендациями GATK Best Practices. Идентификацию SNP, а

также инсерций и делеций (INDELs) проводили с помощью HaplotypeCaller и GenotypeGVCFs из Genome Analysis Toolkit (GATK) v4.1.4.1 (Poplin et al., 2017).

Для последующей фильтрации полученных на первом этапе наборов мутаций был использован алгоритм Variant Quality Score Recalibration (VQSR). В качестве обучающих данных использовались HapMap v3.3, dbSNP 146 (Sherry et al., 2001) и данные 1000 Genomes (1000 Genomes Project Consortium, 2015). На втором этапе фильтрации изученных наборов SNPs и INDELs рассматриваемые данные были преобразованы в формат PLINK BED (Chang et al., 2015). Отфильтровывались слабо представленные (<2% от числа образцов) полиморфизмы, а также отклоняющиеся от равновесия Харди-Вайнберга с порогом 10^{-6} в контроле. Затем были удалены SNV с частотой минорных аллелей <10%. Также был проанализирован уровень гетерозиготности образцов, участвовавших в исследовании. Согласно полученным результатам, ни один из образцов не отклонялся от среднего уровня гетерозиготности на ±3 SD.

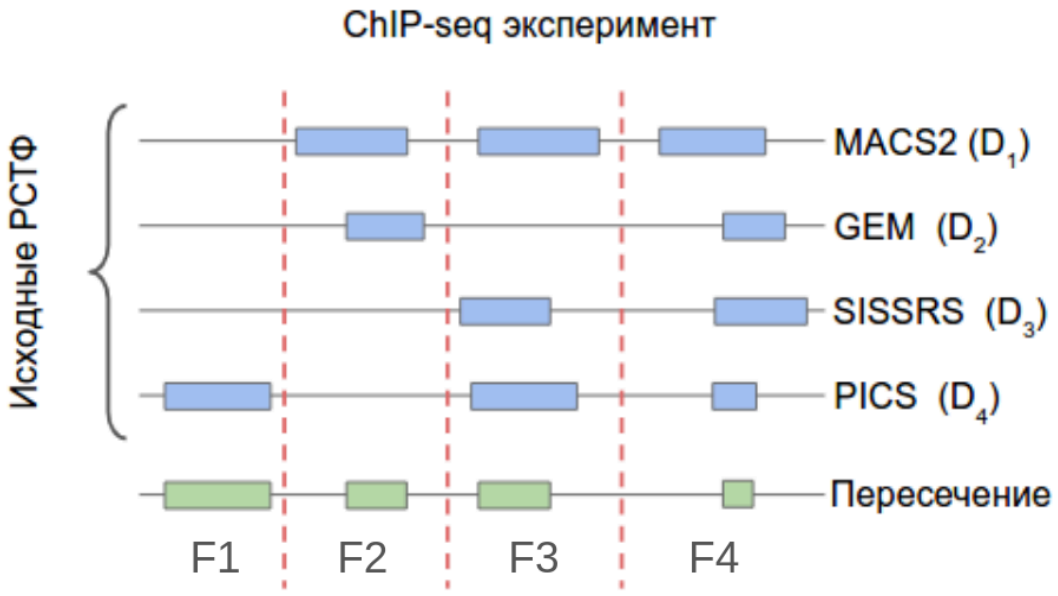
Полученные наборы SNV были аннотированы с помощью Annovar (dbSNP146) (Wang et al., 2010). Для определения приоритетности вариаций, ассоциированных с нарушенным сперматогенезом, их потенциальные эффекты на конечный продукт определяли с помощью программы Ensembl Variant Effect Predictor (McLaren et al., 2016). Кроме того, на основе геномной аннотации были выявлены SNV, расположенные в границах генов, экспрессирующихся в тканях мужской репродуктивной системы. Данные об экспрессии генов были получены из базы данных Human Protein Atlas (Uhlén et al., 2015).

При помощи eQTL из БД GTEx была получена информация об ассоциированных с SNV изменениях уровней экспрессии генов в различных тканях и органах мужской репродуктивной системы. Для оценки влияния SNV на эффективность связывания ТФ с соответствующим РСТФ использовались данные по аллельный дисбаланс связывания ТФ из БД ADASTRA v.5.1.3.

ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1 Взаимосвязь воспроизводимости РСТФ различными алгоритмами идентификации пиков с правдоподобностью

В рамках данной работы была исследована воспроизводимость результатов работы 4 алгоритмов идентификации пиков: GEM, MACS2, PICS и SISSRs (см. Рисунок 3.1.1) на одном и том же наборе из 11836 ChIP-seq экспериментов. На основании сопоставления результатов работы упомянутых выше алгоритмов полученные РСТФ разделялись по уровню воспроизводимости на 4 группы: F1, F2, F3 и F4; Группа F1 соответствовала пикам, идентифицированным только одним из методов, а F4 - РСТФ, которые встречались в результатах применения всех исследуемых алгоритмов (см. Рисунок 3.1.1).



Di - множество РСТФ выявленных заданным методом в эксперименте, Fi - число получившихся районов связывания, которые были составлены ровно из i пиков.

Рисунок 3.1.1 – Схема пересечения результатов работы алгоритмов идентификации пиков и получения обобщённого набора РСТФ с разбиением РСТФ на подгруппы по уровню воспроизводимости среди использованных алгоритмов.

Проведенный сравнительный анализ показал, что, хотя все 4 метода выявили общее подмножество пиков, F4, были выявлены и уникальные пики, обнаруженные каждым алгоритмом, F1. Было показано, что наиболее представленной группой является группа F1 (в среднем ~65% от общего числа РСТФ), а доля группы F4 в среднем составляла менее 10% от общего числа РСТФ в ChIP-seq экспериментах.

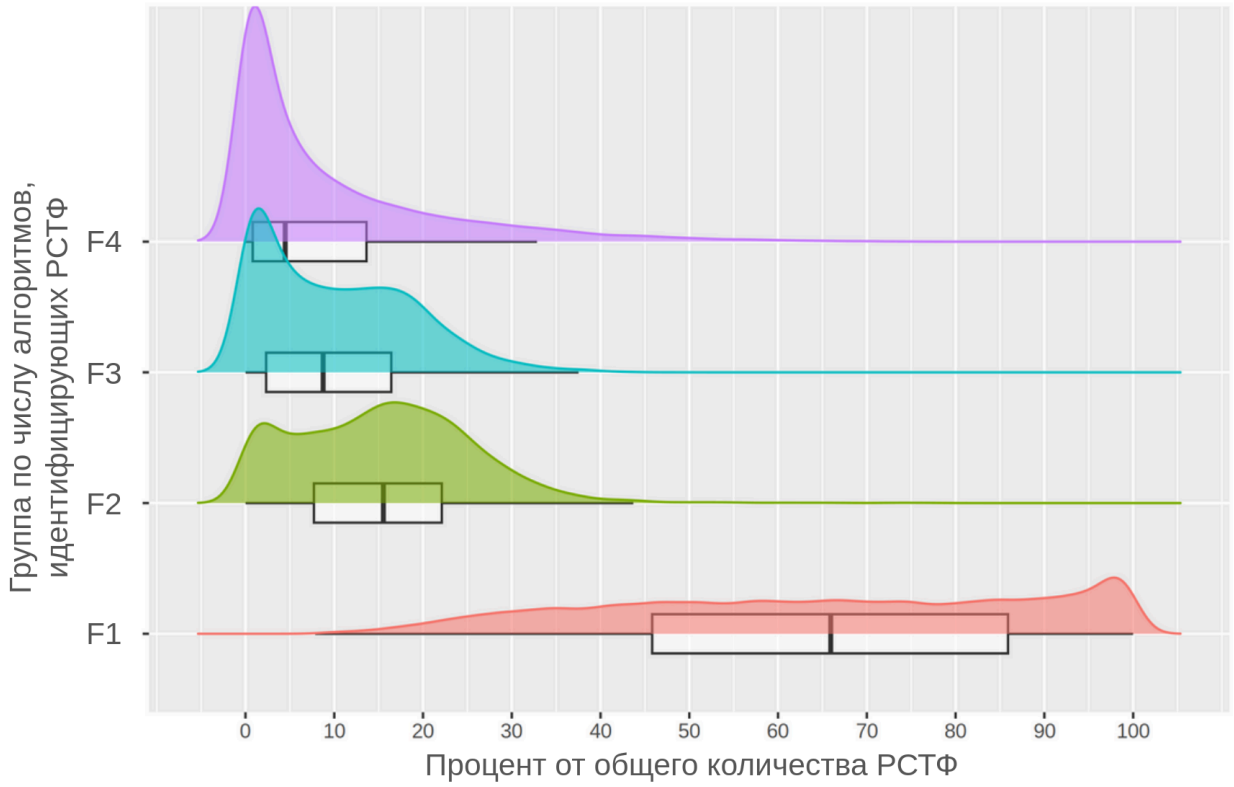


Рисунок 3.1.1 – Плотности распределений результатов идентификации РСТФ при помощи 4 алгоритмов на основании их пересечения между собой. F1, F2, F3, F4 - доли групп, включающие в себя РСТФ, полученные при пересечении 1, 2, 3 и 4 алгоритмов соответственно.

Полученные результаты в целом согласуются с результатами сопоставления наборов пиков, полученных при помощи разных алгоритмов (Osmanbeyoglu et al., 2012; Eder et Grebien, 2022). Например, в работе Osmanbeyoglu с соавторами были проанализированы результаты идентификации потенциальных РСТФ SRC-1 в ChIP-seq данных при помощи 3 алгоритмов: MACS, BayesPeak (Spyrou et al., 2009) и T-PIC (Hower et al., 2011). Было показано, что на долю РСТФ,

идентифицированными всеми 3 методами приходилось ~12% (4811 из 38711). Однако доля данной подгруппы пиков увеличивалась по мере ослабления порога значимости в используемых алгоритмах, что указывает на вариабельность уровня воспроизводимости РСТФ в зависимости от строгости выбранных параметров.

Поскольку успешность идентификации РСТФ напрямую связана с качеством ChIP-seq эксперимента (Eder et Grebien, 2022; Bailey et al., 2013; Landt et al., 2012; Suryatenggara et al., 2022), была исследована взаимосвязь воспроизводимости РСТФ с качеством ChIP-seq данных, полученных на основе рекомендаций проекта ENCODE. Для оценки наличия различий в представленности каждой группы РСТФ между экспериментами с высоким и низким качеством был использован критерий суммы рангов Уилкоксона (U-тест). Для каждой из 4 групп РСТФ наблюдались статистически значимые различия (p-value < 0.05) представленности группы РСТФ в зависимости от качества ChIP-seq эксперимента. Следует отметить, что даже в данных с высоким качеством средняя доля F1 РСТФ в экспериментах равнялась ~45%, а полностью воспроизводятся в среднем только ~15% РСТФ. В данном контексте интересно, чем отличаются между собой РСТФ в зависимости от степени их воспроизводимости на уровне одного ChIP-seq эксперимента.

Во множестве работ была также показана ассоциация РСТФ с различными геномными аннотациями: районами открытого хроматина (doi: 10.1186/s13059-018-1614-y; John et al., 2011; Lamparter et al., 2017), генетической консервативностью района (Tuğrul et al., 2017; Ballester et al., 2014; Håndstad et al., 2011), а также содержанием мотивов связывания ТФ (Keilwagen et al., 2019; Ambrosini et al., 2020). Таким образом, если пик содержит в себе мотив связывания ТФ, лежит в области открытого хроматина, а также располагается в районе с повышенной генетической консервативностью, с большей вероятностью окажется истинным РСТФ. В дальнейшем будем называть такие пики более правдоподобными.

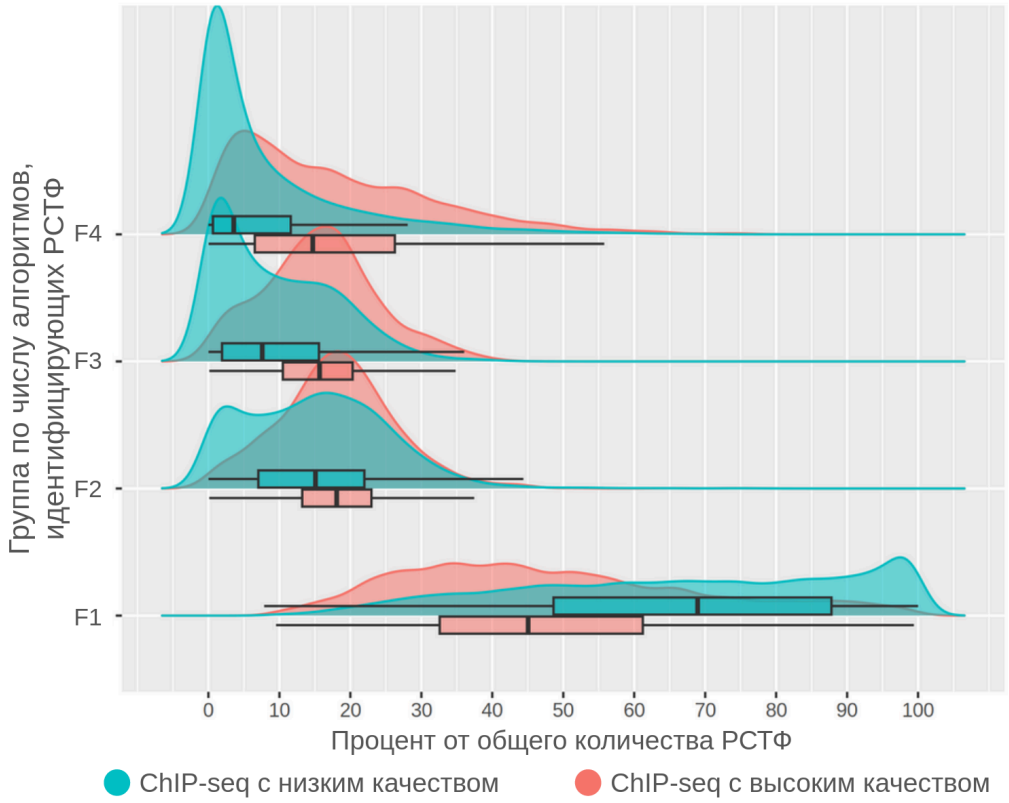
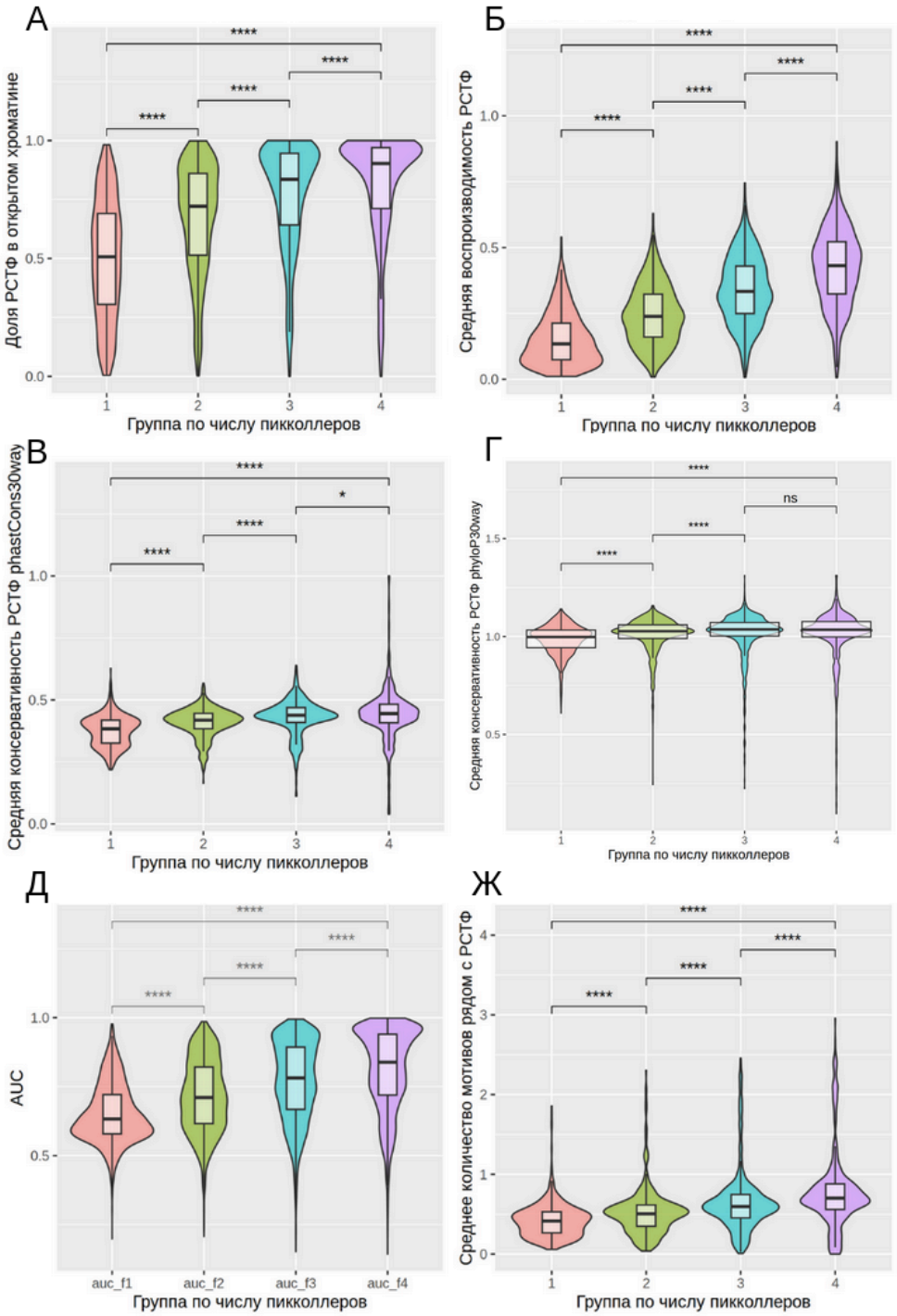


Рисунок 3.1.2 – Плотности распределений результатов идентификации РСТФ при помощи 4 алгоритмов на основании их пересечения между собой в зависимости от качества исходных данных.

Далее была исследована взаимосвязь воспроизводимости РСТФ внутри ChIP-seq эксперимента с различными геномными аннотациями: районами открытого хроматина, генетической консервативностью района, содержанием мотивов связывания ТФ (МСТФ). В первую очередь были привлечены данные о районах открытого хроматина, полученные на основании DNase-seq экспериментов. С целью исключить влияние различий, обусловленных условиями проведения ChIP-seq и DNase-seq экспериментов, для сопоставления наборов пиков были использованы только DNase-seq, совпадающие по типу ткани и условиям проведения эксперимента с ChIP-seq экспериментами. Таким образом, для проведения данного сопоставления были отобраны 1073 ChIP-seq экспериментов, охватывающие 13 ТФ в 195 тканях/клеточных типах. Следует отметить, что большая часть отобранных ChIP-seq экспериментов описывали

условия, которые в исходных дизайнах исследований выполняли роль контрольных условий.

Результаты сопоставления пиков, разбитых в соответствии с уровнем воспроизводимости 4 алгоритмами на 4 подгруппы: F1, F2, F3 и F4, с районами открытого хроматина представлены на рисунке 3.2.1А. Для оценки достоверности различий между распределениями доли пиков в районах открытого хроматина в зависимости от воспроизводимости пиков был применён U-тест. Наблюдались статистически значимые отличия между всеми исследуемыми группами РСТФ. Из полученных данных следует, что F4 РСТФ с большей вероятностью будут располагаться в РОХ, по сравнению и РСТФ из других групп. Несмотря на то, что тенденция располагаться в районах открытого хроматина разнится между ТФ (Kolmykov et al., 2023; Li et al., 2019), сниженная доля F1 РСТФ в РОХ может свидетельствовать о наличии в данной группе РСТФ большего числа ложно идентифицированных РСТФ, по сравнению с другими группами РСТФ, поскольку данный анализ проводился на множестве ТФ и рассматривались средние значения встречаемости РСТФ в РОХ.



(А) Распределения доли РСТФ в открытом хроматине; (Б) Распределения воспроизводимости пиков в других экспериментах для заданного ТФ; (В) Распределения значений эволюционной консервативности пиков методом PhastCons30Way; (Г) Распределения значений эволюционной консервативности пиков методом phyloP30way; (Д) Распределения значений AUC для идентификации мотивов связывания ТФ в РСТФ; (Ж) Распределения среднего числа мотивов связывания ТФ (p-value < 10-4) в окрестности РСТФ. Для определения уровня достоверности различий между группами использовался непараметрический критерий Уилкоксона (ns - p-value > 5*10-2; * - 10-2 < p-value < 5*10-2; **** - p-value < 10-5).

Рисунок 3.1.2 – Взаимосвязь различных геномных аннотаций с принадлежностью к одной из 4 групп пиков, обусловленных степенью пресечения результатов работы 4 алгоритмов идентификации пиков: F1, F2, F3 и F4.

На следующем этапе было проведено сопоставление РСТФ с идентифицированными на геноме мотивами связывания ТФ (МСТФ). Несмотря на то, что между РСТФ и МСТФ наблюдается лишь частичное перекрытие (Håndstad et al., 2011), среди наиболее обогащённых прочтениями пиков в ChIP-seq экспериментах зачастую обнаруживаются мотивы связывания ТФ (Czipa et al., 2020). В целом, точность идентификации сайтов связывания ТФ в наборе РСТФ при помощи позиционных весовых матриц (PWM) зависит, по меньшей мере, от следующих четырех факторов:

1. качество матрицы;
2. качество метода оценки точности распознавания (scoring method);
3. качество исходных данных;
4. неизвестная доля непрямого связывания ТФ, когда рассматриваемый ТФ связан с фрагментом ДНК опосредованно, благодаря белок-белковым взаимодействия с другим ТФ, который, в свою очередь, напрямую связывается с ДНК.

Чтобы исследовать взаимосвязь между успешностью идентификации мотива связывания ТФ (МСТФ) и принадлежности РСТФ к определённой подгруппе, были построены ROC-кривые (receiver operating characteristic) и вычислены значения AUC (area under ROC curve, площадь под ROC-кривой). Для решения данной задачи были рассмотрены модели, основанные на использовании позиционных весовых матриц, а именно, HOCOMOCO v.11 (Kulakovskiy et al., 2018). На рисунке 3.1.2 представлен пример результата построения ROC-кривых для разных подгрупп РСТФ для эксперимента GTRD:EXP069682 (GEO: GSM3962467). Результаты массового анализа эффективности идентификации мотивов связывания ТФ на основании значений AUC отображены на рисунке 3.1.1Д. Для оценки достоверности различий распределений значений AUC между группами РСТФ был использован U-тест.

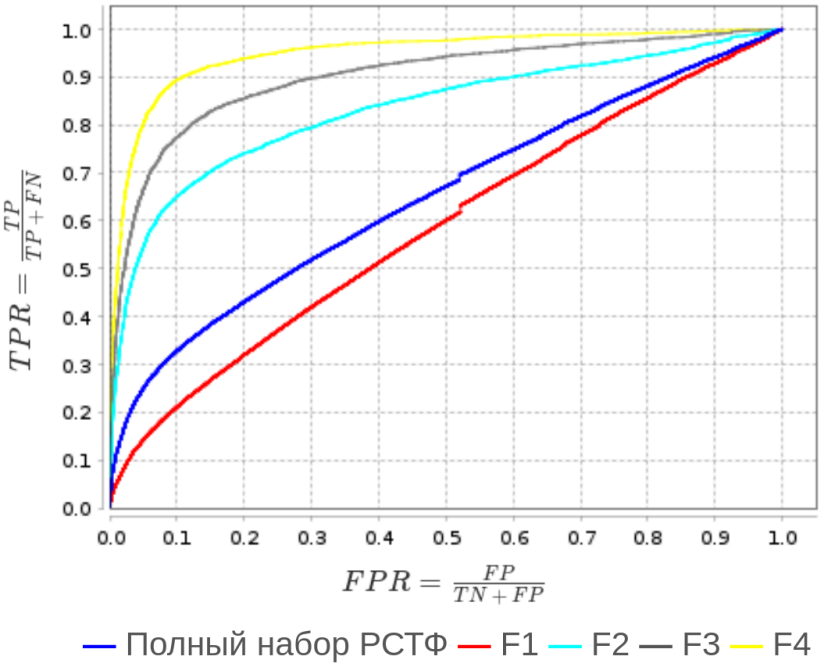


Рисунок 3.1.2 – ROC-кривые, полученные для ChIP-seq эксперимента GTRD: EXP069682 (GEO: GSM3962467) при идентификации мотива связывания FOXA1_HUMAN.H11MO.0.A из базы данных HOCOMOCO.

Также было рассмотрено обогащение окрестностей РСТФ соответствующими мотивами ССТФ (см. Рисунок 3.1.1Ж). Для этого последовательность генома была разбита на набор полуперекрывающихся корзин размером 200 п. н. Полученные интервалы были пересечены с результатами картирования МСТФ (p-value < 0.001) и каждой корзине было присвоено число, соответствующее количеству найденных в ней МСТФ. При проведении обобщения данных по каждому эксперименту из нескольких перекрывающихся с пиком корзин, предпочтение отдавалось корзине с наибольшим количеством мотивов ССТФ.

Результаты проведенного анализа свидетельствуют о наличии статистически значимых различий между всеми группами РСТФ в контексте эффективности предсказания МСТФ. Группа F4 РСТФ демонстрировала наибольшие значения AUC, а F1 - наименьшие. Полученные данные согласуются со средним количеством МСТФ, лежащих в окрестностях РСТФ, где также были найдены статистически значимые различия между группами.

Ещё одним индикатором правдоподобности РСТФ может выступать его встречаемость в других ChIP-seq экспериментах, особенно, идентичных по условиям проведения. Для оценки воспроизводимости РСТФ в схожих ChIP-seq экспериментах были отобраны ТФ, для которых было доступно более **10** экспериментов, и была подсчитана доля экспериментов, в которых они были также идентифицированы (см. Рисунок 3.2.1Б).

Таким образом, из полученных данных следует, что F1 группа РСТФ включает в себя большую, относительно других групп РСТФ, долю ложно идентифицированных РСТФ. С другой стороны, группа F4 РСТФ представлена наиболее правдоподобными РСТФ на основании исследованных геномных аннотаций. К подобному выводу пришли и Suryatenggara с соавторами (Suryatenggara et al., 2022). В данном исследовании при помощи 4 алгоритмов идентификации пиков: MACS2, GEM, HOMER и Genrich были проанализированы 64 ChIP-seq эксперимента, направленные на исследование 20 ТФ и различных модификаций гистонов. Авторами было продемонстрировано, что F4 РСТФ обладает большим, по сравнению с другими группами РСТФ, процентным соотношением РСТФ с МСТФ. Также был проведён анализ взаимосвязи воспроизводимости РСТФ на уровне одного ChIP-seq эксперимента с воспроизводимостью РСТФ в схожих по условиям проведения ChIP-seq экспериментах. Данный анализ проводился на 5 ТФ: CTCF, JUND, MYC, REST и YY1; Для каждого ТФ было использовано по 2 набора ChIP-seq данных, полученных в разных лабораториях. Несмотря на значительно отличающиеся количества идентифицированных РСТФ в разных лабораториях, F4 РСТФ, полученные из данных одной лаборатории почти полностью воспроизводился в наборе данных из второй лаборатории для ТФ: MYC, REST и YY1. Для CTCF и JUND согласованность F4 РСТФ была значительно меньше, однако всё ещё описывала большую часть воспроизводящихся РСТФ.

Заключение к главе 3.1

Проведённый анализ воспроизводимости РСТФ разными алгоритмами идентификации пиков в рамках одного ChIP-seq эксперимента показал, что в среднем полностью воспроизводится ~10% от общего числа РСТФ; Наиболее представленной группой является группа F1 (в среднем ~65% от общего числа РСТФ). Такая вариативность подчеркивает важность использования нескольких алгоритмов идентификации пиков и сопоставления их результатов для получения набора наиболее правдоподобных РСТФ.

Степень воспроизводимости РСТФ разными алгоритмами напрямую связана с достоверностью рассматриваемых РСТФ. Наблюдаются статистически значимые различия между подгруппами РСТФ в контексте расположения РСТФ в областях открытого хроматина, более консервативных районах, а также в районах, демонстрирующих более выраженное обогащение мотивами связывания ТФ.

Было показано, что чем большим количеством методов идентифицируется РСТФ, тем:

1. чаще он встречается в районах открытого хроматина;
2. чаще воспроизводится в других экспериментах для заданного ТФ;
3. является более консервативным;
4. в большем количестве содержит мотивы, представленные позиционной весовой матрицей, связывания соответствующего ТФ.

Из полученных данных следует рекомендация: при необходимости выделения наиболее достоверных РСТФ, полученных в ChIP-seq эксперименте, необходимо рассматривать РСТФ, входящих в состав подгруппы F4.

Однако следует отметить, что существуют эксперименты, демонстрирующие значения характеристик, ассоциированных с достоверностью

РСТФ, на уровне РСТФ из подгруппы F4. Также подгруппа F1 является самой многочисленной. Таким образом, существуют случаи, когда РСТФ подгруппы F1 представлены правдоподобными РСТФ. Разработка метода оценки для выделения таких случаев является одной из задач данной кандидатской диссертации.

3.2 Оценка доли ложноположительных РСТФ. FPCM

Для оценки доли ложноположительных (FP) РСТФ в рамках данной работы была разработана оценка FPCM (False Positive Control Metric). FPCM основывается на предположении о том, что большая часть FP РСТФ находится в группе F1, т. е. подтверждается только одним методом. Таким образом метрика FPCM может быть представлена в виде:

$FPCM = \frac{f_1}{f_1^e}$, где f_1^e - ожидаемое количество истинных пиков в F1

Предполагается, что неизвестное число подлинных РСТФ является случайной величиной с распределением Пуассона. Т. е. для оценки ожидаемого числа F1 РСТФ необходимо решить систему из 3 уравнений, полученных из функции вероятности распределения Пуассона:

$$p_1 = \lambda e^{-\lambda},\ p_2 = \lambda^2 \frac{e^{-\lambda}}{2},\ p_3 = \lambda^3 \frac{e^{-\lambda}}{6},$$

где λ - неизвестный параметр распределения Пуассона, а p_i - вероятность случайно выбранного объединенного РСТФ быть составленным из i сайтов связывания. Таким образом, FPCM можно выразить следующим образом:

$$f_1^e = 2\frac{f_2^2}{3f_3} \text{ и } FPCM = \frac{f_1}{f_1^e} = \frac{3f_1f_3}{2f_2^2}$$

В целом, можно предположить, что каждый набор F1 может состоять из истинно положительных F1 РСТФ (True Positive Orphans; TPOs) и

ложноположительных орфанов (False Positive Orphans; FPOs), т. е. Количество F1 РСТФ может быть выражено как:

$$f_1 = (\text{количество } TPOs) + (\text{количество } FPOs).$$

Поскольку мы можем оценить количество истинно положительных орфанов, мы также можем оценить долю ложноположительных орфанов (p_{false}):

$$p_{false} = \frac{f_1 - f_1^e}{f_1} = 1 - \frac{1}{FPCM}\ .$$

Таким образом, мы можем выразить характеристику FPCM как 1 / (1 - p_{false}). Если f_1 и f_1^e близки друг к другу по значению, то предположение о Пуассоновском распределении соблюдается. В таком случае значение p_{false} должно быть близким к 0, а значение FPCM должно быть близко к 1. Однако если распределение Пуассона серьезно нарушается, p_{false} принимает большие значения, а значение FPCM значительно превышает 1.

Разработанный алгоритм был реализован на языке программирования Java в виде анализа для платформы BioUML (см. Рисунок 3.2.1).

Использование ChIP-seq контролей снижает долю ложно идентифицированных РСТФ, в связи с чем является необходимым при планировании ChIP-seq экспериментов (Landt et al., 2012). Таким образом, отдельный интерес представляет изучение влияния наличия контролей в ChIP-seq экспериментах на значения FPCM. В рамках данной работы из полной выборки ChIP-seq экспериментов для человека были отобраны 3402 эксперимента, для которых был доступен инпут-контроль. Для исследования зависимости значений FPCM от наличия инпут-контроля выбранные эксперименты были обработаны повторно, но без учета инпут-контроля. Затем были посчитаны значения FPCM для двух наборов РСТФ для каждого эксперимента. На рисунке 3.2.2 представлены плотности распределений значений FPCM в зависимости от присутствия инпут-контроля. На основании U-критерия наблюдается

статистически значимые различия между значениями FPCM в рассматриваемых группах.

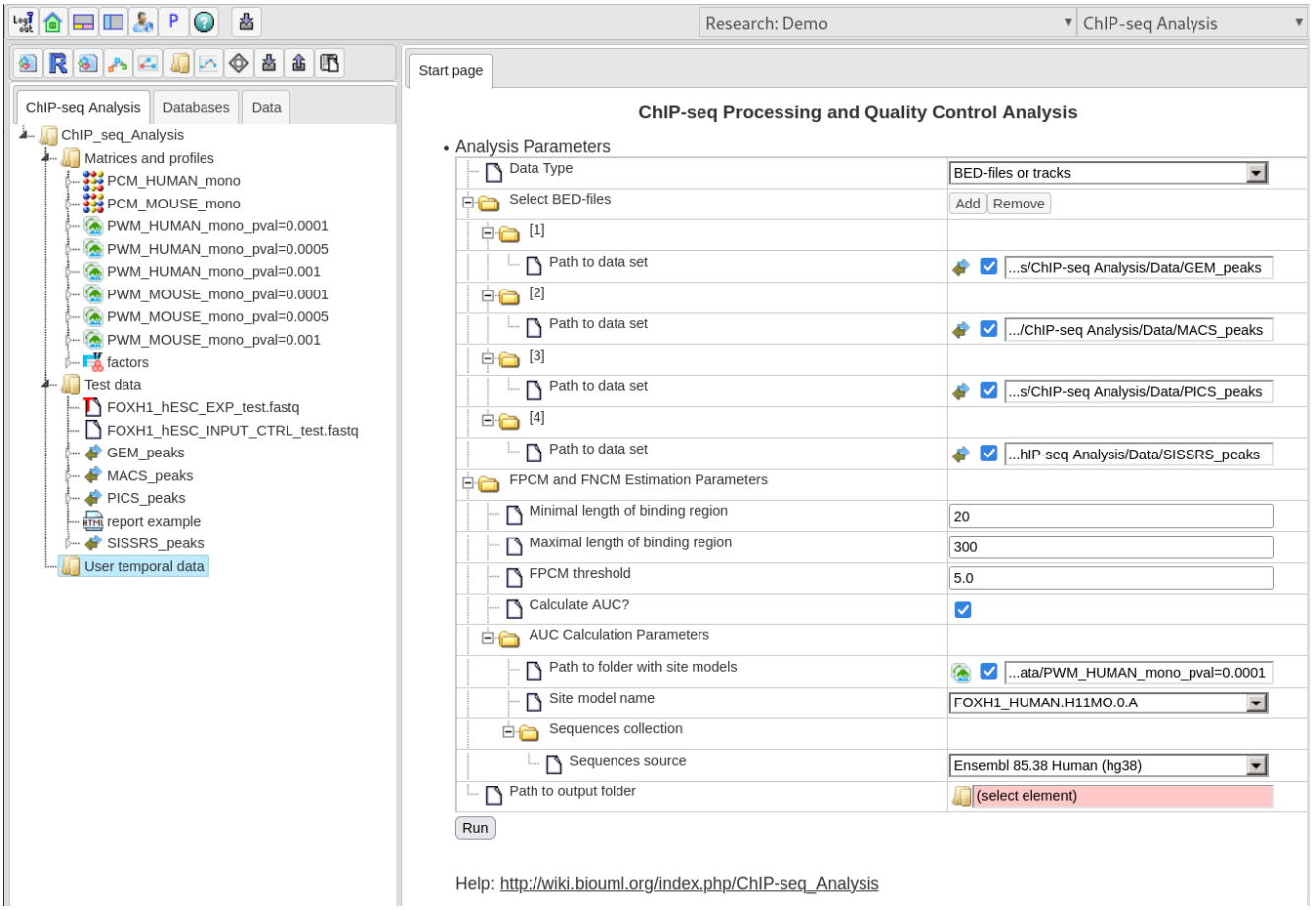


Рисунок 3.2.1 – Интерфейс программы на платформе BioUML, реализующей алгоритм расчёта характеристик FPCM и FNCM, а также оценивает истинное количество РСТФ на основании пересечения предоставленных пользователем наборов РСТФ.

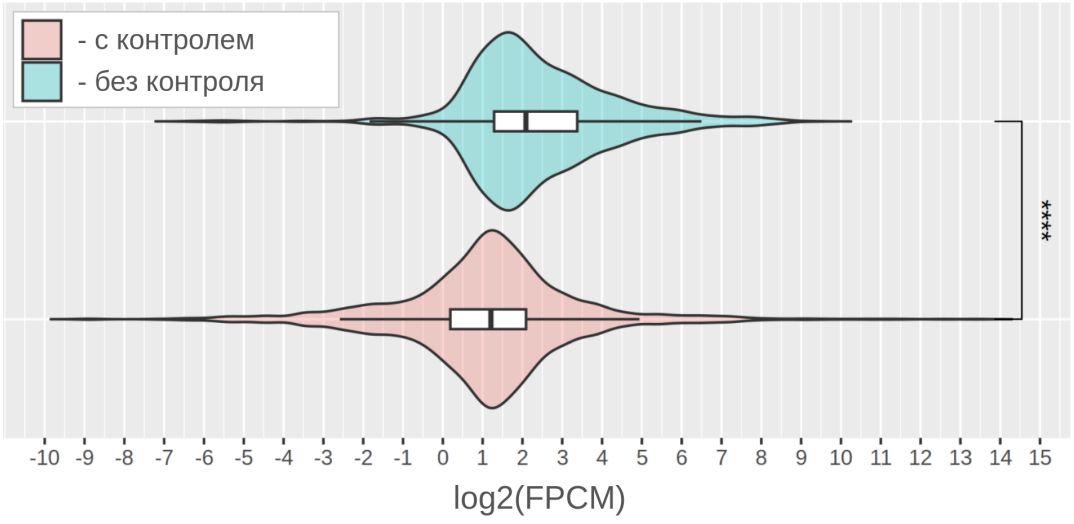


Рисунок 3.2.2 – Плотности распределений значений FPCM в наборах РСТФ зависимости от присутствия инпут-контроля в процессе обработки данных

Далее была исследована зависимость значений FPCM от результатов пересечения F1 РСТФ с другими типами данных: открытым хроматином, расположением предсказанных МСТФ в РСТФ, консервативностью РСТФ и воспроизводимостью РСТФ в других ChIP-seq экспериментах (см. Рисунки 3.2.3-3.2.6). Для решения данной задачи было отобрано 453 ChIP-seq эксперимента которые отвечали следующим требованиям:

1. Для заданного ТФ доступны данные DNase-seq экспериментов, проведенные в схожих экспериментальных условиях;
2. Для заданного ТФ доступна PWM в HOCOMOCO;
3. Для данного ТФ доступно минимум 10 ChIP-seq экспериментов.

Сперва была исследована взаимосвязь между значениями FPCM и эффективностью обнаружения РСТФ с МСТФ (см. Рисунок 3.2.3). Значения AUC для обнаружения РСТФ с МСТФ были рассмотрены для F1 и F4 РСТФ независимо друг от друга. На рисунках 3.2.3 А и Д демонстрируется выраженное снижение значений AUC для F1 РСТФ после достижения определенных значений FPCM, тогда как для F4 РСТФ снижения значений AUC для F4 РСТФ не наблюдается. Чтобы проанализировать взаимосвязь значений AUC с качеством ChIP-seq экспериментов, данные были разделены на основании значений характеристик качества согласно рекомендациям проекта ENCODE на две группы: данные с низким и высоким качеством (см. Рисунки 3.2.3 В и Ж). Согласно U-критерию Манна-Уитни наблюдаются достоверные различия значений AUC между данными с высоким и низким качеством; При этом, в группе F1 РСТФ различия между медианами более выраженные по сравнению с F4 РСТФ. Далее на основании значений FPCM данные были также разделены на две группы: эксперименты с умеренной и повышенной долей ложно положительных F1 РСТФ; В качестве порогового значения было выбрано FPCM=5. Следует отметить, что разбиение на две группы на основании выбранного порогового значения FPCM демонстрирует более выраженные различия в медианах между полученными

группами в F1 РСТФ, по сравнению с разбиением на основании качества данных согласно рекомендациям ENCODE. Также, на основании значений FPCM даже среди качественных ChIP-seq экспериментов можно выделить эксперименты, для которых характерна в среднем более низкие значения AUC в F1 РСТФ.

На следующем этапе была исследована взаимосвязь между значениями FPCM и долей F1 и F4 РСТФ в РОХ (см. Рисунок 3.2.4). На рисунке 3.2.4А демонстрируется выраженное снижение доли F1 РСТФ в РОХ после достижения определенных значений FPCM, тогда как для F4 РСТФ подобной картины не наблюдается (см. Рисунок 3.2.4Д). Также наблюдаются достоверные различия доли F1 и F4 РСТФ в РОХ между данными с высоким и низким качеством (см. Рисунки 3.2.4 В и Ж); При этом, в группе F1 РСТФ различия между медианами более выраженные по сравнению с F4 РСТФ. В группах, полученных на основании значений FPCM, наблюдаются достоверные различия в медианах между группами с высоким и умеренным содержанием ложно положительных РСТФ среди F1. Следует отметить, что по сравнению с разбиением на основании качества данных согласно рекомендациям ENCODE, между группами, полученных при помощи FPCM, не было найдено достоверных различий в доли F4 РСТФ в РОХ. Как и в случае с идентификацией МСТФ в РСТФ, на основании значений FPCM даже среди качественных ChIP-seq экспериментов можно выделить эксперименты, для которых характерна в среднем более низкое содержание F1 РСТФ в РОХ.

Далее была исследована взаимосвязь между значениями FPCM и воспроизводимостью F1 и F4 РСТФ в других ChIP-seq экспериментах для рассматриваемого ТФ (см. Рисунок 3.2.5). Воспроизводимость РСТФ определялась как отношение числа экспериментов, в которых встречается РСТФ, к общему числу экспериментов. В целом, наблюдается картина, схожая с описанным ранее взаимоотношением F1 и F4 РСТФ с РОХ: только для F1 РСТФ наблюдается выраженное снижение воспроизводимости в других экспериментах

после достижения определенных значений FPCM (см. Рисунок 3.2.5 А и Д). Также наблюдаются достоверные различия доли F1 и F4 РСТФ между данными с высоким и низким качеством (см. Рисунки 3.2.5 В и Ж), а также статистически значимые различия между наборами экспериментов, разделенными по FPCM. При этом, в группе F1 РСТФ различия между медианами более выраженные по сравнению с F4 РСТФ. Как и ранее, в числе экспериментов со сниженной воспроизводимостью F1 РСТФ и высокими значениями FPCM попали ChIP-seq эксперименты с высоким качеством по ENCODE.

На следующем этапе была исследована взаимосвязь между значениями FPCM и эволюционной консервативностью районов, в которых располагаются F1 и F4 РСТФ (см. Рисунок 3.2.6). Для этого геномная последовательность была разбита на набор наполовину перекрывающихся между собой участков длиной 200 п.н., корзины. Для каждой корзины было подсчитано среднее значение консервативность позиций нуклеотидов по PhastCons30way. Затем полученные районы были сопоставлены с F1 и F4 РСТФ. При пересечении РСТФ с более чем 1 корзиной, присваивалось наибольшее из значений оценки консервативности участка. Как и в ранее рассмотренных аннотациях наблюдается снижение консервативности района с РСТФ после прохождение определенного порога FPCM. Также наблюдаются статистически значимые различия в консервативности районов с F1 РСТФ как в наборах экспериментов, которые были разделены по качеству, так и в наборах, разделенных на основании значений FPCM. Среди качественных ChIP-seq экспериментов также выделяются эксперименты со сниженной консервативностью районов с F1 РСТФ и повышенным значением FPCM.

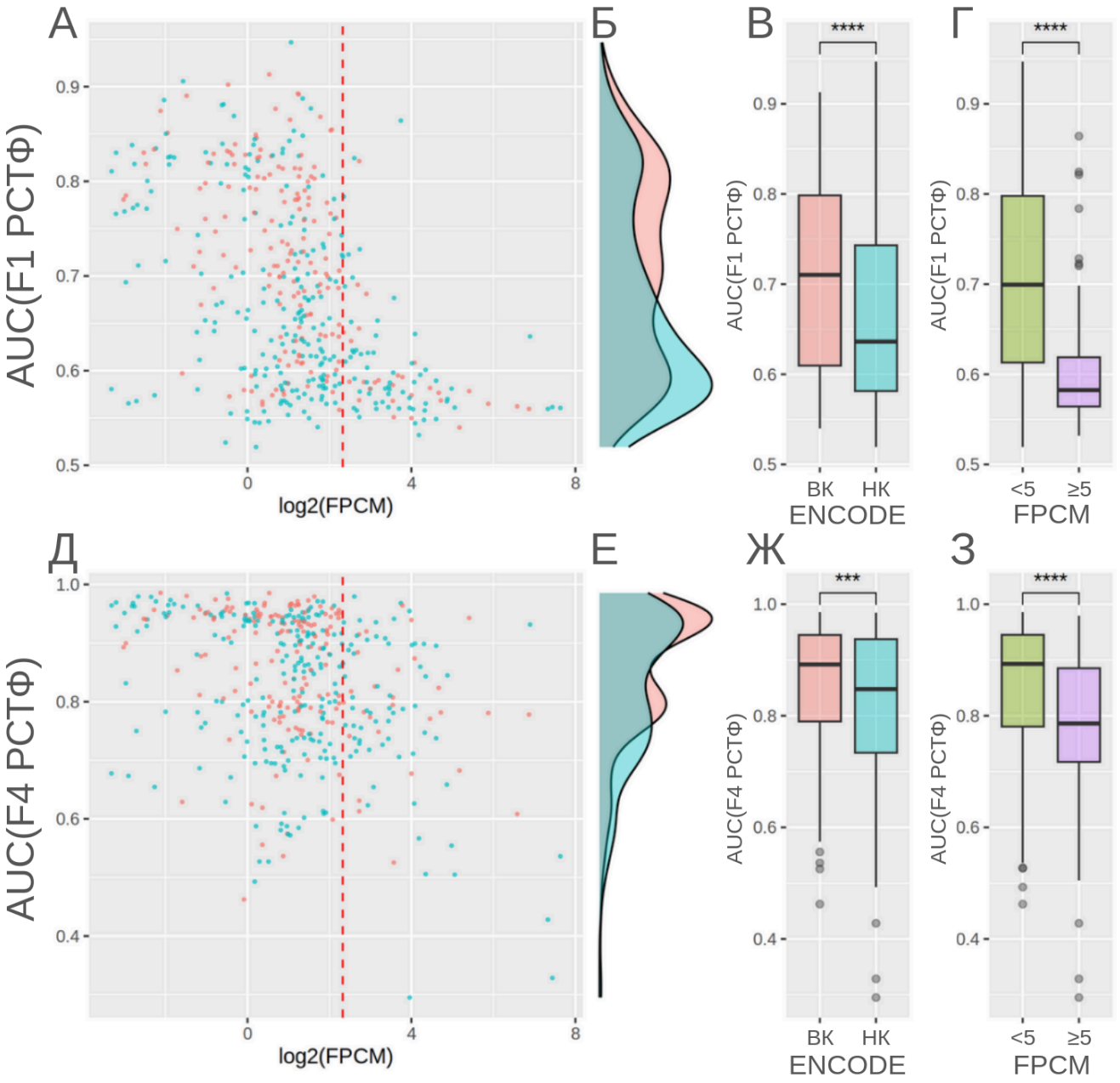


Рисунок 3.2.3 – (АД) - Взаимоотношение значений log2(FPCM) и эффективностью идентификации мотивов ССТФ в подгруппе пиков F1 (верхний ряд) и в группе F4 (нижний ряд). Синим цветом обозначены эксперименты, которые по критериям ENCODE относятся к данным с низким качеством; Красным цветом - качественные ChIP-seq эксперименты по критериям ENCODE; Красный пунктир - пороговое значение FPCM = 5. (БЕ) - Плотности распределений значений AUC для F1 и F4 для данных с высоким (красный цвет) и низким (синий цвет) качеством. (ВЖ) - диаграммы размаха, описывающие распределения с рисунка (БЕ). (ГЗ) - диаграммы размаха, описывающие распределение значений AUC в F1 и F4 в экспериментах с FPCM < 5 (зелёный цвет; умеренное количество FP) и для экспериментов с FPCM > 5 (фиолетовый цвет; высокое содержание FP).

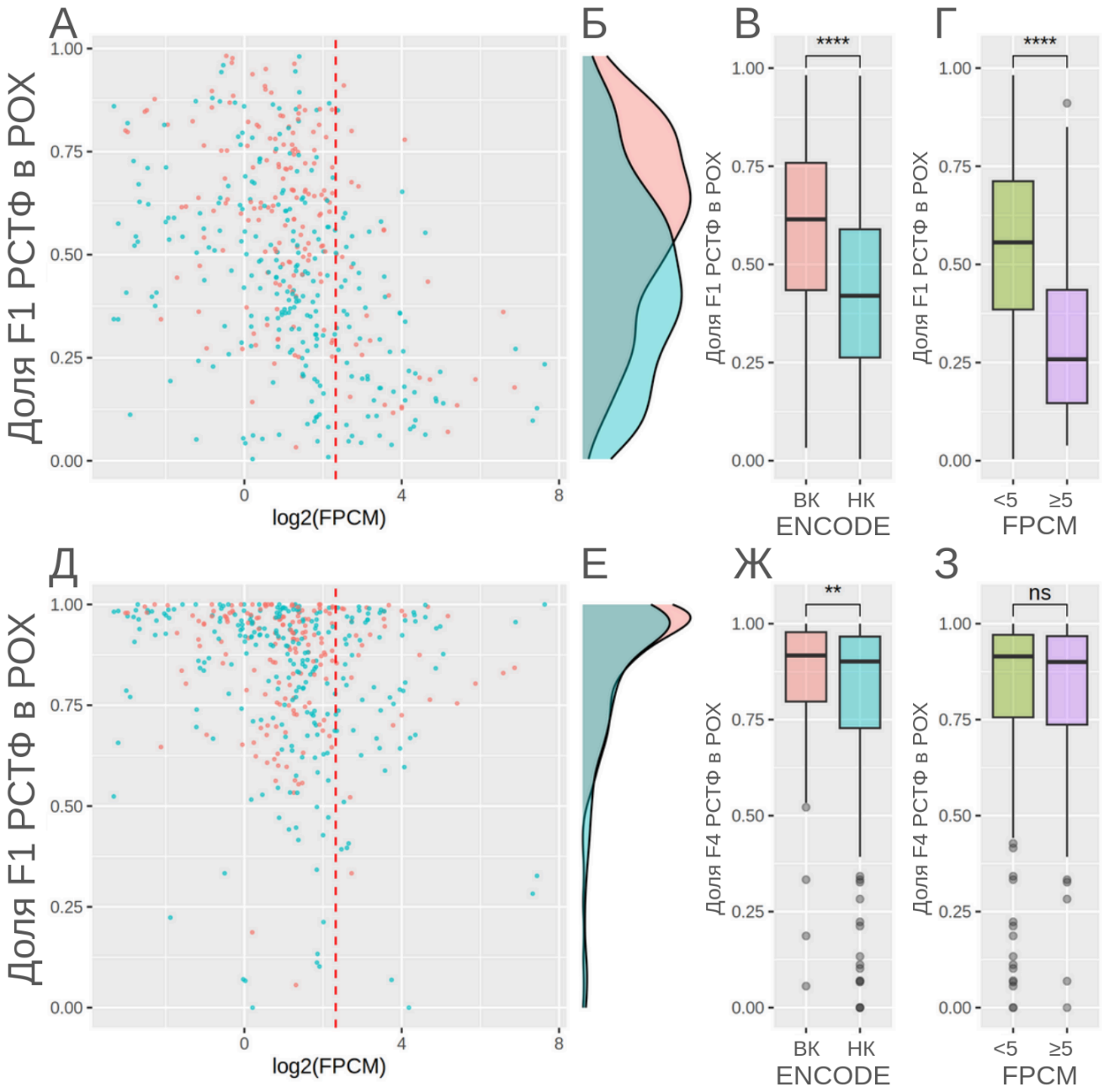


Рисунок 3.2.4 – (АД) - Взаимоотношение значений log2(FPCM) и долей РСТФ в открытом хроматине в подгруппе F1 (один пикколлер; верхний ряд) и в подгруппе F4 (нижний ряд; РСТФ подтверждается всеми пикколлерами). Синим цветом обозначены эксперименты, которые по критериям ENCODE относятся к данным с низким качеством; Красным цветом - качественные ChIP-seq эксперименты по критериям ENCODE; Красный пунктир - условный порог FPCM = 5. (БЕ) - Плотности распределений значений AUC для F1 и F4 в хороших и плохих данных (красный и синий цвета соответственно). (ВЖ) - Ящики с усами (жирная линия в ящике - медиана, а границы прямоугольника - первый и третий квартиль), описывающие распределения с рисунка (БЕ). (ГЗ) - ящики с усами , описывающие распределение значений AUC в F1 и F4 в экспериментах с FPCM < 5 (зелёный цвет; умеренное количество FP) и для экспериментов с FPCM > 5 (фиолетовый цвет; высокое содержание FP).

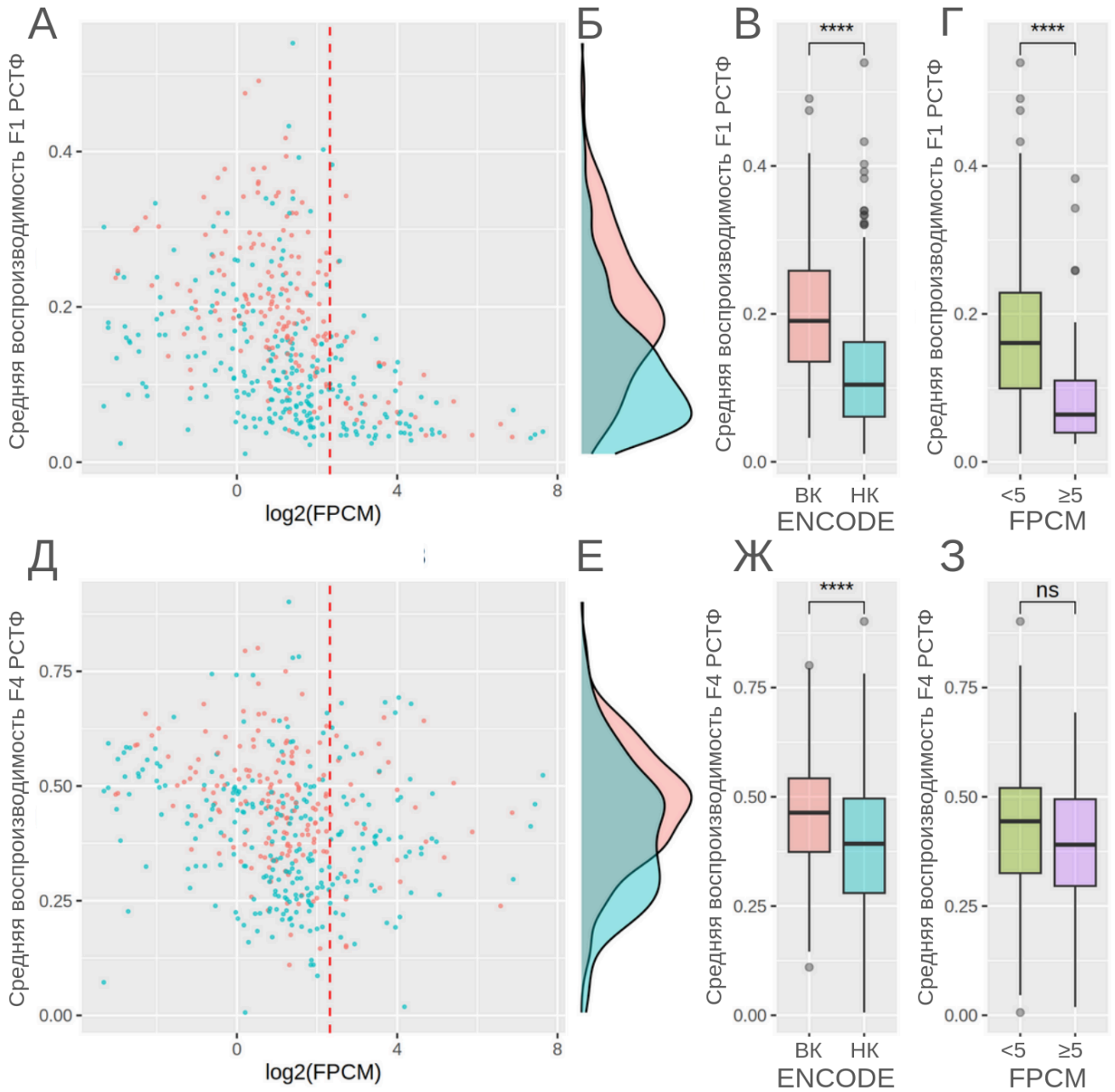


Рисунок 3.2.5 – (АД) - Взаимоотношение значений log2(FPCM) и воспроизводимостью РСТФ среди всех экспериментов (количество экспериментов с РСТФ / количество экспериментов) в подгруппе f1 (один пикколлер; верхний ряд) и в подгруппе f4 (нижний ряд; все пикколлеры). Синим цветом обозначены эксперименты, которые по критериям ENCODE относятся к данным с низким качеством; Красным цветом - качественные ChIP-seq эксперименты; Красный пунктир - условный порог FPCM = 5. (БЕ) - Плотности распределений значений AUC для f1 и f4 в хороших и плохих данных (красный и синий цвета соответственно). (ВЖ) - Ящики с усами (жирная линия в ящике - медиана, а границы прямоугольника - первый и третий квартиль), описывающие распределения с рисунка (БЕ). (ГЗ) - ящики с усами , описывающие распределение значений AUC в F1 и F4 в экспериментах с FPCM < 5 (зелёный цвет; умеренное количество FP) и для экспериментов с FPCM > 5 (фиолетовый цвет; высокое содержание FP)

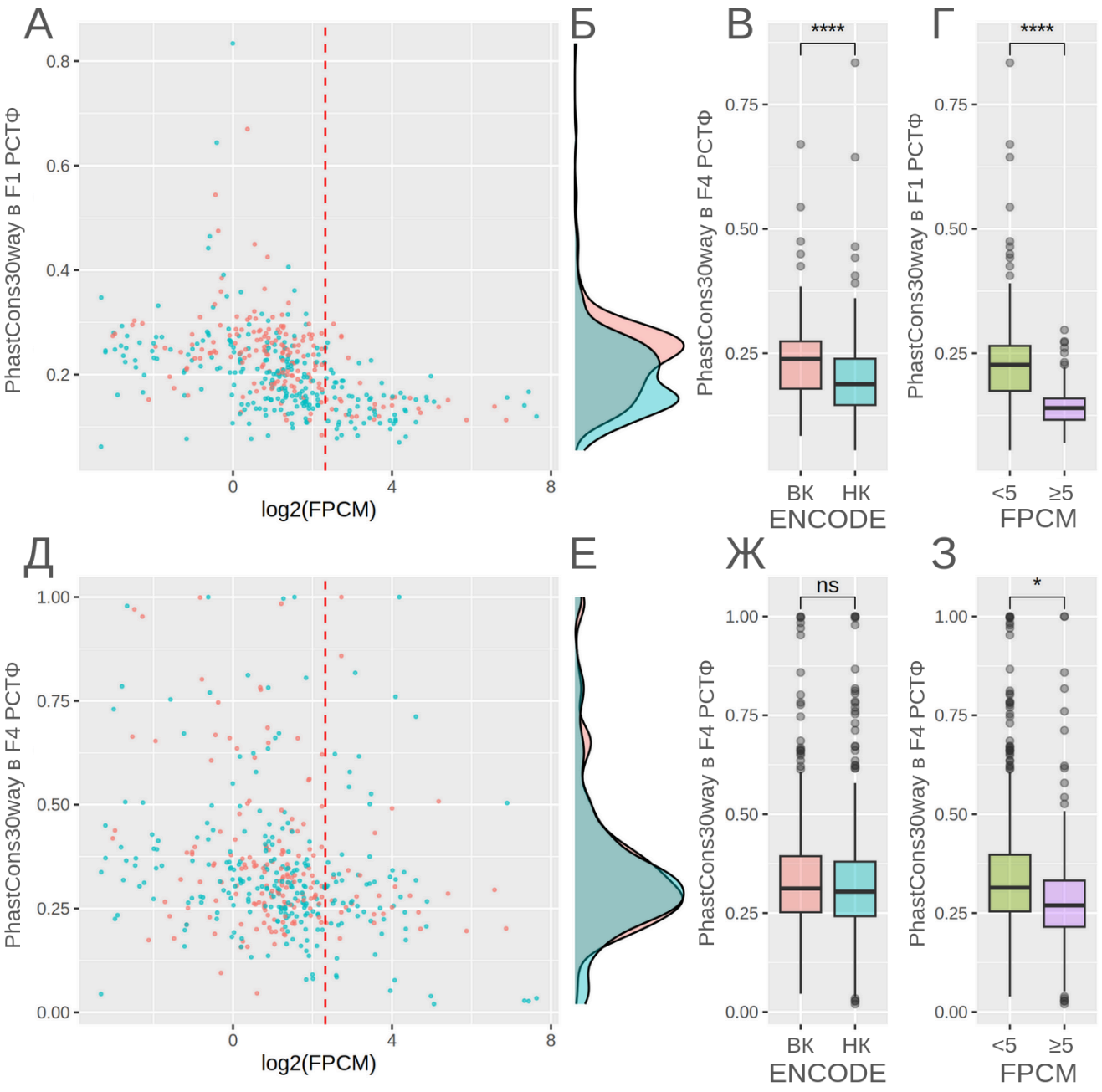


Рисунок 3.2.6 – (АД) - Взаимоотношение значений log2(FPCM) и эволюционной консервативностью района по PhastCons30way (количество экспериментов с РСТФ / количество экспериментов) в подгруппе F1 (один пикколлер; верхний ряд) и в подгруппе F4 (нижний ряд; все пикколлеры). Синим цветом обозначены эксперименты, которые по критериям ENCODE относятся к данным с низким качеством; Красным цветом - качественные ChIP-seq эксперименты; Красный пунктир - условный порог FPCM = 5. (БЕ) - Плотности распределений значений AUC для f1 и f4 в хороших и плохих данных (красный и синий цвета соответственно). (ВЖ) - Ящики с усами (жирная линия в ящике - медиана, а границы прямоугольника - первый и третий квартиль), описывающие распределения с рисунка (БЕ). (ГЗ) - ящики с усами , описывающие распределение значений AUC в F1 и F4 в экспериментах с FPCM < 5 (зелёный цвет; умеренное количество FP) и для экспериментов с FPCM > 5 (фиолетовый цвет; высокое содержание FP)

Далее была исследована возможность использования значений FPCM для принятия решения по удалению F1 РСТФ из дальнейшего анализа. Для этого был проведен анализ взаимосвязи между значениями FPCM и изменением эффективности идентификации МСТФ в полном наборе РСТФ в ответ на удаление F1 РСТФ (см. Рисунок 3.2.7). На первом этапе был проведён анализ эффективности идентификации МСТФ на основании использовании PWM из БД HOCOMOCO для 5855 ChIP-seq экспериментов, относящихся к 17 наиболее представленным в БД GTRD ТФ. Затем, анализ был повторен на наборах РСТФ, из которых были исключены РСТФ из группы F1. После этого было подсчитано отношение значений AUC до удаления группы F1 к значениям после удаления F1 РСТФ. На рисунке 3.2.6 демонстрируется резкое ухудшение эффективности предсказания МСТФ в полных наборах РСТФ при достижении определенных значений FPCM. Следует обратить внимание, что данная картина не меняется при рассмотрении только ChIP-seq экспериментов, которые определены как качественные согласно рекомендациям проекта ENCODE. Таким образом, на основании высоких значений FPCM можно рекомендовать удалять из последующего анализа F1 РСТФ, тогда как при умеренных значениях FPCM удаление F1 РСТФ не приведёт к существенному увеличению доли РСТФ с МСТФ.

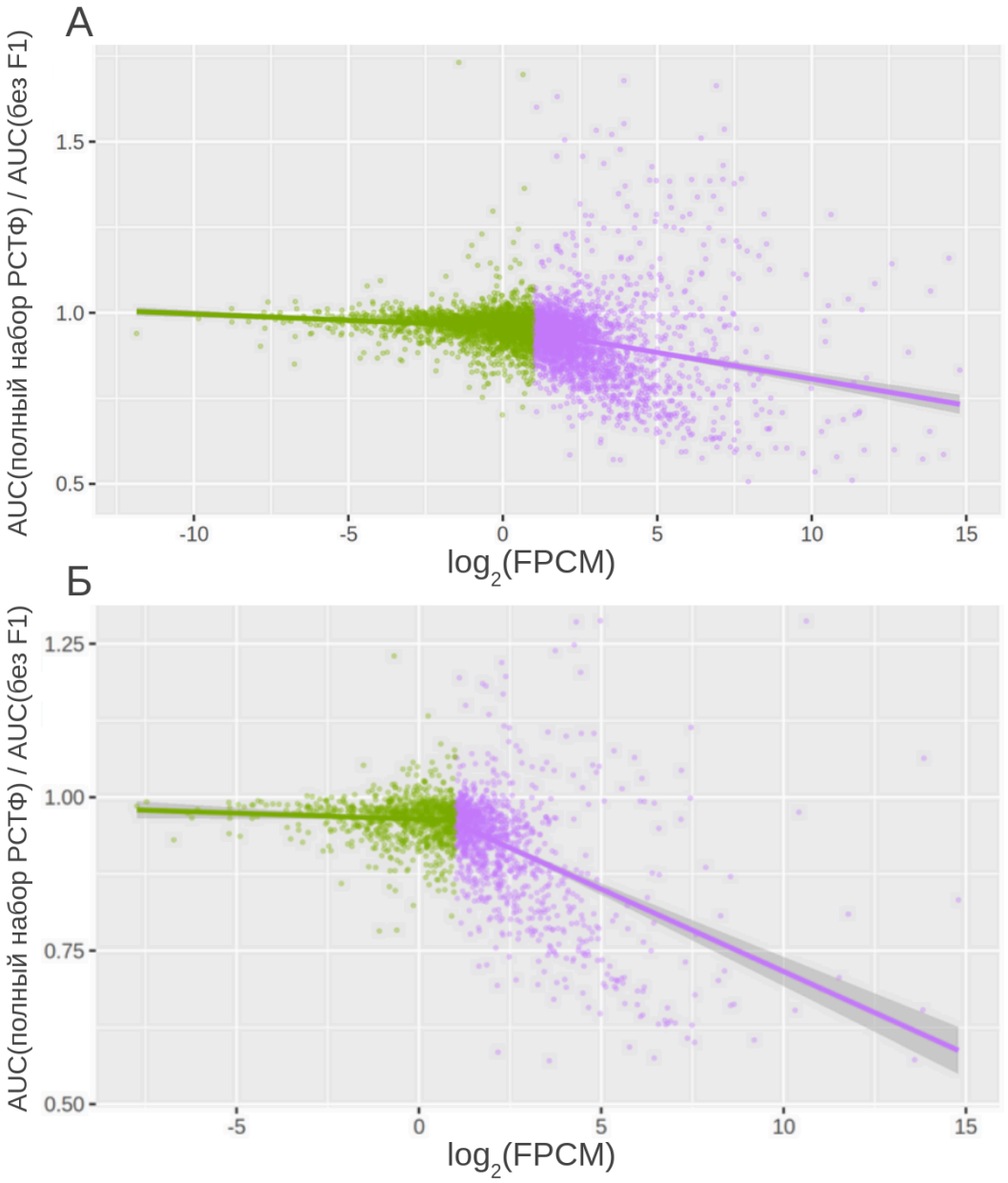


Рисунок 3.2.7 – (А-Б) Взаимоотношение оценки FPCM и изменения значений AUC для предсказания мотивов связывания ТФ до и после удаления F1 подгруппы из РСТФ, идентифицированных только одним методом идентификации пиков (MACS2). Зелёным и фиолетовым цветом обозначены эксперименты, располагающиеся ниже и выше относительно выбранного условного порогового значения FPCM=5. На рисунке (Б) отображены только ChIP-seq эксперименты, которые прошли фильтрацию по качеству согласно рекомендациям проекта ENCODE.

Заключение к главе 3.2

Была предложена и валидирована новая метрика, для оценки доли ложно положительных РСТФ в ChIP-seq эксперименте на основе анализа пересечения

результатов работы 4 алгоритмов идентификации пиков — FPCM. Для платформы BioUML на языке Java был реализован алгоритм расчета значений FPCM.

В рамках данной главы было показано, что даже для ChIP-seq экспериментов с высоким качеством FPCM позволяет идентифицировать поднаборы экспериментов, которые демонстрируют:

- сниженное количество МСТФ, представленные позиционной весовой матрицей, связывания соответствующего ТФ, в F1 РСТФ;
- сниженное количество F1 РСТФ в РОХ;
- более низкую воспроизводимость F1 РСТФ в других ChIP-seq экспериментах для выбранного ТФ;
- более низкую эволюционную консервативность районов с F1 РСТФ.

Также на основании пересечения F1 РСТФ с другими типами данных было продемонстрировано, что на основании характеристики FPCM даже среди качественных, согласно рекомендациям проекта ENCODE, ChIP-seq экспериментов можно выделить эксперименты, для которых характерно снижение доли правдоподобных РСТФ в F1 РСТФ на основании сопоставления с дополнительными аннотациями. Было продемонстрировано, что повышенные значения FPCM могут выступать рекомендацией к удалению из дальнейшего анализа группы F1 РСТФ.

3.3 Оценка доли ложно-невыявленных РСТФ. FNCM

Для оценки доли ложно-невыявленных РСТФ в рамках данной работы была разработана оценка FNCM (False Negative Control Metric). FNCM определяется как отношение количества РСТФ, выявленных конкретным методом, к ожидаемому количеству подлинных РСТФ:

$$FNCM\left(D_i\right) = \frac{\left|D_i\right|}{N^{gen}}$$, где $|D_i|$ - количество РСТФ в наборе D_i,

где N^{gen} - оценка общего количества подлинных РСТФ заданного типа.

N^{gen} оценивается как среднее значение четырех различных оценок (E_C, E_{LB}, E_Z и E_{ML}), используемых для оценки размера популяций, для N^{gen}, т.е.

$$FNCM\left(D_i\right) = \frac{\left|D_i\right|}{N_1^e}\text{, где } N_1^e = \frac{\left(E_C+E_{LB}+E_Z+E_{ML}\right)}{4},$$

где E_C - оценка Чао (Chao's estimate) (Chao, 1987), E_{LB} - оценка Ланумтинга-Бонинга (Lanumteang-Bohning's estimate) (Lanumteang et Bohning, 2011), E_Z - оценка Зельтермана (Zelterman's estimate) (Zelterman, 1988) и E_{ML} - оценка, основанная на функции максимального правдоподобия (maximum likelihood estimate) (McCrea, Morgan, 2014), которые имеют следующий вид:

$$E_C = n + \frac{f_1^2}{2f_2}, E_{LB} = n + \frac{3f_1^3 f_3}{4f_2^3}, E_Z = \frac{n}{1-exp(\frac{-2f_2}{f_1})}, E_{ML} = \frac{n}{1-exp(-\lambda^*)},$$

где λ^* рассчитывается путем максимизации логарифмической функции правдоподобия $L(\lambda)$ для положительного распределения Пуассона.

$$L(\lambda) = constant + \log log\,(\lambda)\sum_{i=1}^{k}\left(i * f_i\right) - n\log log\left(e^{\lambda} - 1\right).$$

Также был предложен альтернативный метод оценки неизвестного числа подлинных РСТФ для N^{gen}; Данный подход рассматривает все k(k-1)/2 различных пар (D_i, D_j) при $i \neq j$ и рассчитали для каждой пары (D_i, D_j) оценку Чапмана (Chapman's estimate) (Chapman, 1951) $E_{i,j}$ по формуле

$$E_{i,j} = \frac{\left(\left|D_i\right|+1\right)\left(\left|D_j\right|+1\right)}{\left|D_i \cap D_j\right|+1} - 1.$$

Затем осуществляется проверка на наличие выбросов в полученном наборе оценок Чапмана ($E_{Chap} = \{E_{i,j}\}$) и последующее их удаление. Произвольный элемент X в выборке классифицируется как выброс, если имеет место следующее неравенство:

$$\left|\left(X - X_0\right)\right| > 3\sigma,$$

где X_0 и σ - среднее значение и стандартное отклонение, когда элемент X временно удален из выборки E_{Chap}. Наконец, N^{gen} оценивается как среднее значение выборки E_{Chap}, а FNCM(D_i) выражается как

$$FNCM\left(D_i\right) = \frac{\left|D_i\right|}{N_2^e},$$

где $N_2^e =$ среднее значение выборки E_{Chap}.

Значение FNCM варьируется в диапазоне [0,0; 1,0]. Чем ближе значение FNCM к 1, тем ниже ошибка недопредсказания, в то время как значения ближе к 0 указывают на то, что большое количество подлинных РСТФ в рассматриваемом наборе было упущено.

Стоит отметить, что предложенные метрики FNCM и FPCM можно использовать для сравнения РСТФ между разными ChIP-seq экспериментами для одного и того же ТФ в сходных экспериментальных условиях.

Для платформы BioUML на языке Java был реализован алгоритм расчета метрики FNCM, а также алгоритм оценки истинного размера набора РСТФ на основании пересечения предоставленных пользователем наборов РСТФ (Рисунок 3.2.1).

С целью выявления взаимоотношений внутри характеристик качества, а также FPCM и FNCM, для всех возможных комбинаций пар характеристик был посчитан коэффициент парной корреляции Пирсона (см. Рисунок 9). Для большинства случаев не наблюдалось сильной корреляции между характеристиками. На общем фоне выделяются группы характеристик FNCM и FRiP, демонстрирующие относительно высокие уровни корреляции значений для всех используемых методов идентификации РСТФ (0.36-0.68 и 0.46-0.9), и умеренно низкий уровень корреляции между значениями между данными группами характеристик (-0.15-0.51). Стоит отметить, что умеренно низкий уровень отрицательной корреляции наблюдался между значениями коэффициента относительной кросс-корреляции (RSC) и характеристиками, оценивающими

сложность библиотеки (NRF и PBC1), что может быть объяснено снижением данных показателей для экспериментов с большим размером библиотеки прочтений. Подобного поведения также придерживались значения FRiP и FNCM для PICS и SISSRs; Интересно, что MACS2 и GEM демонстрировали обратную картину, что может говорить о наличии связи между сложностью библиотеки ChIP-seq эксперимента и эффективностью работы данных методов идентификации РСТФ.

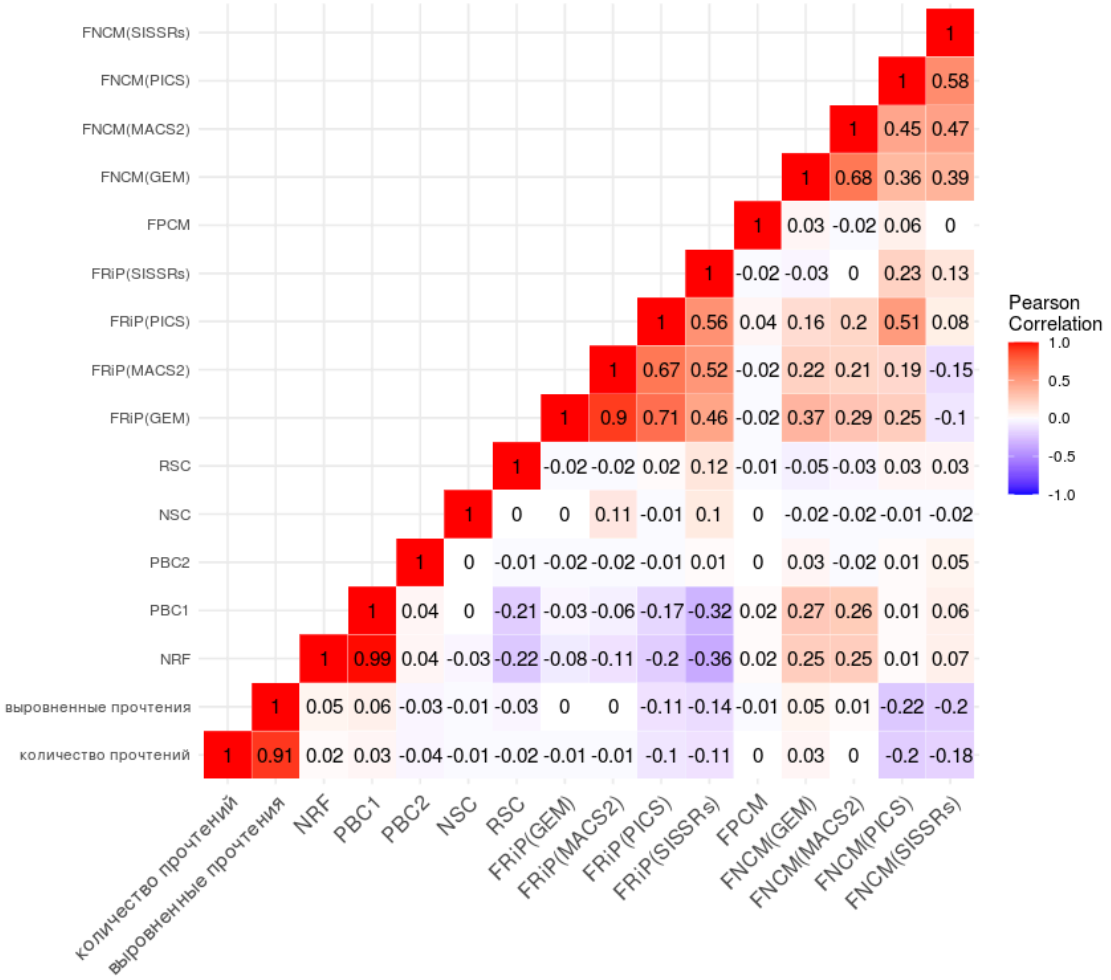


Рисунок 3.3.1 – Значения парного коэффициента корреляции Пирсона для характеристик качества ChIP-seq экспериментов.

Для более глубокого изучения связи между предложенными характеристиками (FPCM и FNCM) и характеристиками качества был проведен регрессионный анализ. Для этой цели были применены три модели

множественной регрессии, а именно: метод простых наименьших квадратов (ordinary least squares; OLS), алгоритм случайного леса (random forest; RF) и метод опорных векторов (support vector machine; SVM) для 11836 наборов ChIP-seq данных для человека в GTRD. Сила взаимосвязей между характеристиками качества и FNCM/FPCM была измерена путем вычисления корреляции Пирсона между наблюдаемыми и предсказанными значениями характеристик. Чтобы избежать переобучения регрессионных моделей, весь набор данных был разделён на два равных подмножества: обучающую и тестовую выборки. Данная процедура по разделению исследуемого набора данных с последующей оценкой эффективности предсказательных моделей была проведена 10 раз. Максимальные средние значения коэффициента корреляции (0,657 и 0,611) были достигнуты с помощью использования алгоритма случайного леса (см. Таблицу 3.3.1). В первом случае регрессионная модель описывала соотношение между FNCM(PICS) и показателями качества описанными ENCODE, тогда как вторая регрессионная модель описывала отношение между FNCM(GEM) и характеристиками, полученными от данного метода идентификации пиков. В целом, умеренные значения корреляций указывают на то, что между предложенными характеристиками качества и существующими метриками качества нет сильных взаимосвязей, в частности. Другими словами, нет комбинаций известных признаков, которые могли бы заменить FNCM или FPCM.

Таблица 3.3.1 – Взаимосвязь между FPCM и FNCM и другими характеристиками качества.

Тип характеристики качества	**Характеристика качества**	**Регрессионная модель**	**Коэффициент корреляции Пирсона между предсказанными и наблюдаемыми значениями характеристик качества**
Метрики качества, предложенные в рамках проекта ENCODE	FNCM (GEM)	OLS RF SVM	0.472 0.611 0.545
	FNCM (MACS)	OLS RF SVM	0.336 0.413 0.327
	FNCM (PICS)	OLS RF SVM	0.415 0.657 0.475
	FNCM (SISSRs)	OLS RF SVM	0.259 0.451 0.295
	FPCM	OLS RF SVM	0.044 0.104 0.064

Стоит отметить поведение FPCM и FNCM, когда большая часть показателей качества ENCODE указывают на низкое качество ChIP-seq эксперимента. С одной стороны, на основании таких характеристик, как NRF, PBC1, NSC и RSC, можно рекомендовать исключить эти данные из дальнейшего анализа. С другой стороны, почти во всех случаях FNCM указывает на высокий уровень ложноотрицательных РСТФ. Другими словами, используемым методам идентификации пиков не удалось идентифицировать множество подлинных РСТФ. Однако FPCM указывает на низкую долю ложноположительных РСТФ, на основании чего рекомендуется не производить модификаций наборов РСТФ для последующего анализа. Таким образом, использование FPCM для фильтрации ChIP-seq экспериментов возможно только при участии других характеристик качества, поскольку данная характеристика оценивает только взаимоотношение внутри нескольких наборов РСТФ.

На основании значений FNCM было проведено сравнение используемых алгоритмов идентификации пиков. MACS2 показал лучшие результаты в 69,5% экспериментов, в то время как GEM, SISSRs и PICS превзошли своих конкурентов в 8,4%, 13,9% и 8,2% экспериментов, соответственно. Данный вывод о превосходстве MACS2 над другими методами идентификации пиков подтверждается в проведенном в 2017 году сравнением производительности данной группы методов (Thomas et al. 2017).

Использование ChIP-seq контролей снижает число идентифицированных пиков в ChIP-seq экспериментах, поскольку помогает отсеять ложноположительные пики (Liang et Keleş, 2012). В данном контексте были исследованы значения FNCM для 4 алгоритмов идентификации пиков в зависимости от присутствия инпут-контроля при обработке данных. Для этого были отобраны 3402 ChIP-seq эксперимента, для которых в БД GTRD присутствовали инпут-контроли. Результаты расчёта FNCM в присутствии/отсутствии инпут-контролей представлены в таблице 3.3.2. На основании U-критерия Манна-Уитни наблюдается статистически достоверные различия в значениях FNCM в наборах РСТФ, полученных в присутствии/отсутствии инпут-контролей для алгоритмов: GEM, PICS и SISSRs;

Таблица 3.3.2 – Средние значения FNCM для 11836 ChIP-seq экспериментов для человека из GTRD.

Метрики качества	**Все эксперименты**	**с контролем**	**без контроля**	**U-test**	**p-value**
FNCM(GEM)	0,509	0,516	0,484	3,997	$6,4 * 10^{-5}$
FNCM(MACS2)	0,651	0,645	0,672	0,864	0,389
FNCM(PICS)	0,36	0,292	0,62	28,461	$< 10^{-14}$
FNCM(SISSRs)	0,454	0,398	0,668	24,753	$< 10^{-14}$

Также следует обратить внимание на характер распределения плотностей вероятности значений FNCM для каждого из методов идентификации пиков (см.

Рисунок 3.3.2). Различия в плотностях распределения вероятности значения FNCM также указывают на снижение доли ложноотрицательных РСТФ в случаях использования методов PICS или SISSRs для идентификации РСТФ в экспериментах с контролем, а для MACS2 и GEM отличия менее выражены.

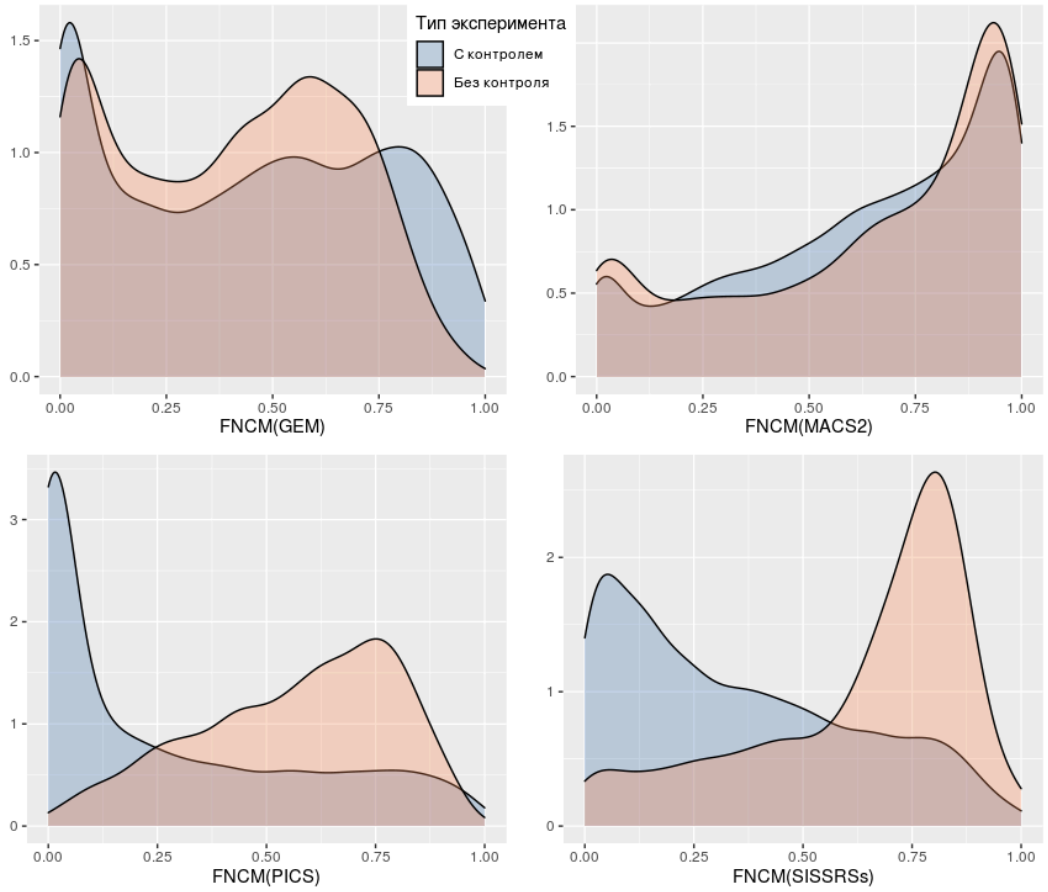


Рисунок 3.3.2 – Наблюдаемые распределения плотности вероятности значений FNCM для различных методов идентификации пиков.

Заключение по главе 3.3

Была разработана и валидирована новая метрика оценки доли ложно недопредсказанных РСТФ в ChIP-seq эксперименте на основании пересечения нескольких алгоритмов идентификации пиков — FNCM. Для платформы BioUML на языке Java был реализован алгоритм расчета значений FNCM.

Сравнительный анализ алгоритмов идентификации пиков на основании значений FNCM показал превосходство MACS2 над остальными алгоритмами вне зависимости от доступности инпут-контроля при обработке ChIP-seq данных.

В данной главе было показана возможность совместного использования разработанных оценок, FPCM и FNCM, в сочетании с другими оценками качества данных, что позволяет комплексно подходить к оценке качества данных и выявлять наборы наиболее достоверных РСТФ.

3.4 METARA

Для выявления наиболее достоверных РСТФ на основе мета-анализа результатов ChIP-seq анализа всех экспериментов из БД GTRD для заданного ТФ был разработан новый метод METARA – METa Analysis of ChIP-seq datasets through the Rank Aggregation (рис. 3.4.1, Kolmykov et al., 2019, 2020). Выявленные при его помощи РСТФ, встречающиеся в нескольких экспериментах одновременно, называются мета-кластерами. Данный метод представляет собой трехэтапное применение метода коллективного выбора (МКВ):

1-й этап – $МКВ_1$ применяется для ранжирования РСТФ, найденных каждым методом (MACS, SISSRs, GEM и PICS), на основании различных характеристик качества, присваиваемых соответствующим методом. Например, MACS2 присваивает такие характеристики, как «Fold Enrichement», «FDR» (уровень ложного обнаружения), «количество выровненных прочтений» и «-lg(p-value)»;

2-й этап – полученные и упорядоченные на основании применённой функции ранжирования списки эксперимента $Э_i$ для некоторого $ТФ_m$ подаются на вход выбранному МКВ;

3-й этап – полученный на предыдущем этапе список РСТФ группируется с другими подобными списками, полученными из всех экспериментов по рассматриваемому ТФ, и затем обрабатываются используемым МКВ.

В результате описанного выше многостадийного алгоритма METARA для каждого ТФ был построен набор мета-кластеров, ранжированный на основании значений финальной агрегирующей функции (ФАФ).

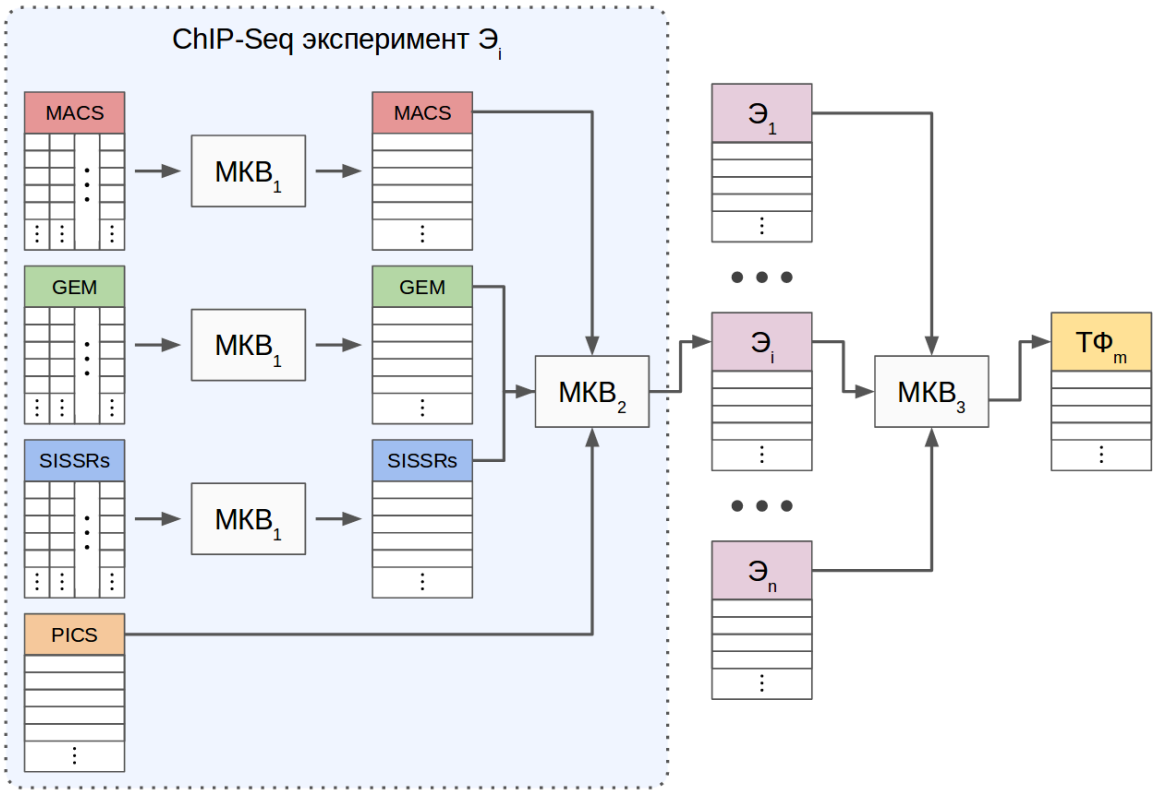


Рисунок 3.4.1 – Схема многостадийного применения методов коллективного выбора. МКВi - i-ая стадия применения агрегирующей функции (метода коллективного выбора), основанной на вычислении среднего арифметического значения.

Описанный алгоритм был также реализован на языке Java в виде анализа для платформы BioUML (Рисунок 3.4.2). В качестве агрегирующей функций доступны: арифметическое и геометрическое средние, медиана, L1-norm, L2-norm, а также методы, основанные на использовании Марковских цепей. Для построения матрицы переходных вероятностей использовались алгоритмы MC1,

MC2 и MC3, которые были подробно описаны в соответствующей главе в разделе, посвященному обзору литературы.

Для получения более корректного результата при использовании метода Борда для набора неполных списков, использовалась следующая формула вычисления ранга $R_{\tau_l}(i)$ для отсутствующего в рассматриваемом списке τ_l элемента i:

$$R_{\tau_l}(i) = \frac{1}{n} \sum_{m=k+1}^{n} m = \frac{1}{2n}(n - k)(n + k + 1),$$

где n - общее количество элементов, k - ранг последнего упорядоченного элемента в рассматриваемом списке τ_l.

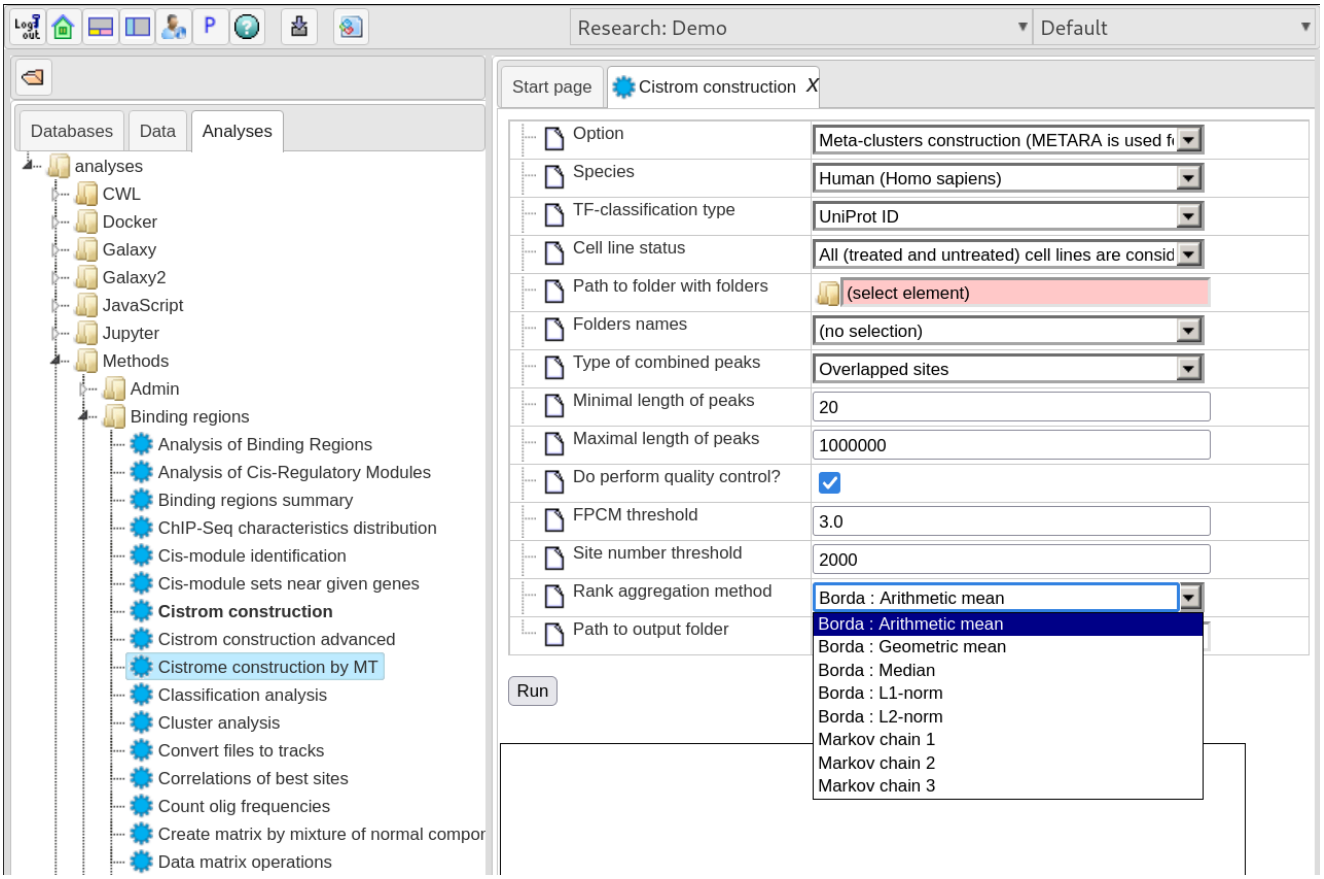


Рисунок 3.4.2 – Интерфейс программы на платформе BioUML, реализующий применение различных агрегирующий функций, для генерации на основании предоставленных пользователем наборов РСТФ обобщённого набора РСТФ.

При помощи предложенного алгоритма были построены карты геномных районов связывания 1391 ТФ и кофакторов человека. Полученные районы вошли в состав БД GTRD и доступны по ссылке: http://gtrd.biouml.org:8888/egrid/bigBeds/hg38/ChIP-seq.

Был проведен анализ правдоподобности мета-кластеров в зависимости от значений ФАФ. Для этого полученные мета-кластеры для заданного ТФ были разбиты на 50 равных по размеру подгрупп на основании значений ФАФ. Из каждой подгруппы было случайным образом взято по 5000 мета-кластеров. Для каждой подгруппы было посчитано значение AUC, характеризующее эффективность поиска мотивов связывания в рассматриваемой подгруппе, а также подсчитана доля мета-кластеров, располагающихся в районах открытого хроматина. На рисунке 3.4.3 приведен пример такого анализа для ТФ NRF1.

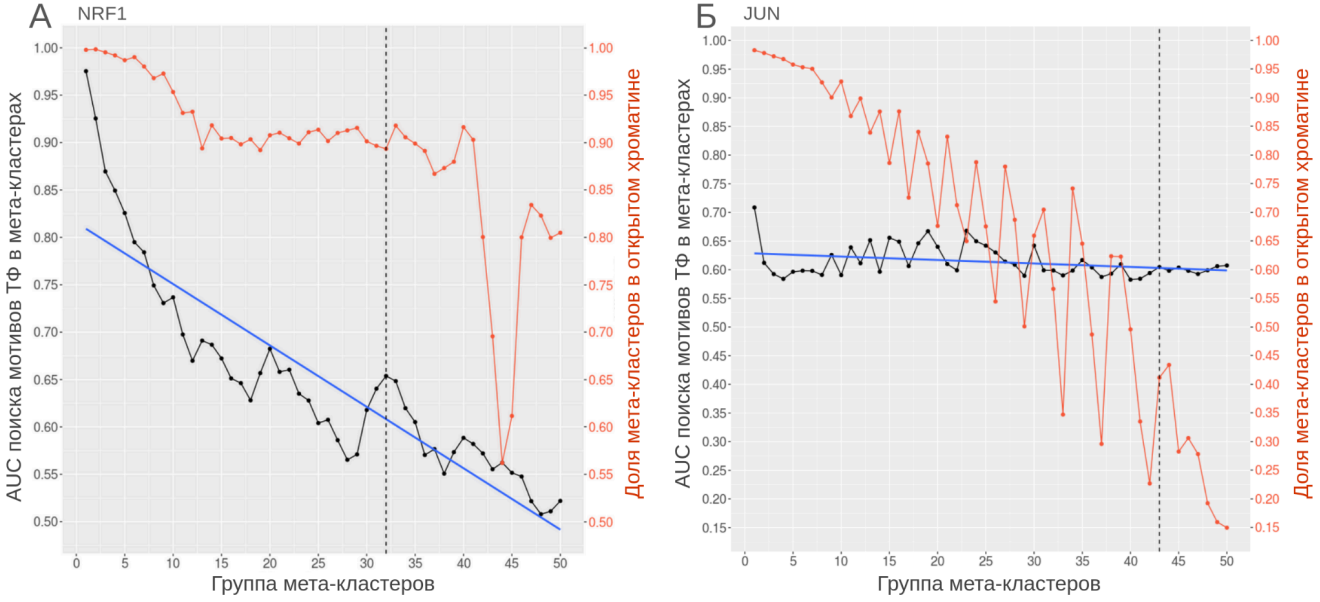


Рисунок 3.4.3 – Взаимоотношение значений ФАФ и значений AUC для поиска мотивов связывания ТФ NRF1 и JUN (чёрная линия), а также взаимоотношение значений ФАФ и доли мета-кластеров NRF1 и JUN в районах открытого хроматина (красная линия).

Для описания зависимости RA-score и значений AUC была использована линейная регрессия. Для проанализированных 119 ТФ показано, чем выше ФАФ (меньше номер группы мета-кластеров), тем чаще мета-кластеры:

1. встречаются в районах открытого хроматина (для 108 из 119 ТФ);
2. содержит мотивы, представленные позиционной весовой матрицей, связывания соответствующего ТФ (для 101 из 119 ТФ).

Полученные результаты свидетельствуют о том, что существуют ТФ, для которых, как минимум, характерна относительно сниженная тенденция содержать МСТФ в РСТФ, а также РСТФ в РОХ. Для более детального исследования тенденции РСТФ различных ТФ включать в себя МСТФ из базы данных GTRD были отобраны ChIP-seq эксперименты, для которых имеются PWM-матрицы мотивов связывания ТФ в базе данных HOCOMOCO. Затем из данного набора экспериментов были отобраны эксперименты, удовлетворяющие характеристикам качества, рекомендованным проектом ENCODE. Таким образом был составлен набор из 3426 ChIP-seq экспериментов. Затем, к полученному набору экспериментов был применён алгоритм METARA. При объединении результатов работы 4 алгоритмов идентификации пиков на основании значения FPCM (FPCM > 5) принималось решение об исключении из дальней обработки подгруппы пиков F1. Таким образом, были получены наборы кластеров РСТФ для 362 ТФ. Для дальнейшего анализа для каждого набора РСТФ был выбрано пороговое значение ФАФ. Для выбора порогового значения ФАФ были взяты значения ФАФ, ограничивающие топ 200000 РСТФ в наборах кластеров РСТФ. Затем данный порог был применён на уровне отдельных экспериментов, чтобы отобрать наиболее правдоподобные РСТФ. Полученные из каждого эксперимента наборы РСТФ были пересечены с позициями МСТФ, предсказанными при помощи PWM с порогом p-value < 0.0001.

На рисунке 3.4.4 представлены распределения доли наиболее правдоподобных РСТФ с МСТФ для ТФ, для которых в анализе присутствовало 10 или более ChIP-seq экспериментов. Из полученных данных видно, что наблюдается высокая вариабельность в доли МСТФ в РСТФ в зависимости от ТФ. Так, например, для ТФ: CTCFL, USF2, CTCF, EGR1 и USF1 характерна высокая

(>75%) доля РСТФ с МСТФ, тогда как для ТФ: TBP, PBX2, MYB, IRF4, SMAD3, STAT3, TCF12, GATA3, RBPJ, ATF3, RUNX1 и OTX2 характерна низкая (<25%) доля РСТФ с МСТФ. Вероятно, для ТФ с меньшей долей РСТФ в МСТФ свойственно проявление меньшей специфичности относительно последвательности их РСТФ; либо для данных ТФ более характерно опосредованное связывание с ДНК посредством белок-белковым взаимодействиям. Полученные результаты согласуются с исследованием Ambrosini с соавторами (Ambrosini et al., 2020). В данной работе исследовали эффективность PWM из разных БД: JASPAR (Castro-Mondragon et al., 2022), HOCOMOCO, и CIS-BP (Weirauch et al., 2014) в предсказании РСТФ, полученных на основе различных экспериментальных данных: ChIP-seq, HT-SELEX и PBM, для различных ТФ. Было также показано, что эффективность предсказания РСТФ по МСТФ сильно разнится в зависимости от ТФ, а также для некоторых ТФ наблюдается достижения наибольшей эффективности предсказания РСТФ при использовании PWM для МСТФ, относящихся к ТФ из других семейств ТФ. Авторы статьи связывают это преобладанием кооперативного связывания для таких ТФ.

Для анализа тенденции РСТФ различных ТФ располагаться в РОХ 3426 ChIP-seq эксперимента были сопоставлены с набором данных DNase-seq экспериментов из БД GTRD по их экспериментальным условиям. Таким образом были отобраны совпадающие по условиям проведения эксперименты. Затем полученные ранее наиболее правдоподобные, на основании значений ФАФ, РСТФ из ChIP-seq экспериментов были сопоставлены с районами открытого хроматина и для каждого ChIP-seq эксперимента посчитана доля рассматриваемых РСТФ в РОХ. На рисунке 3.4.5 представлены распределения доли РСТФ в РОХ для ТФ, для которых в анализе присутствовало 10 или более ChIP-seq экспериментов.

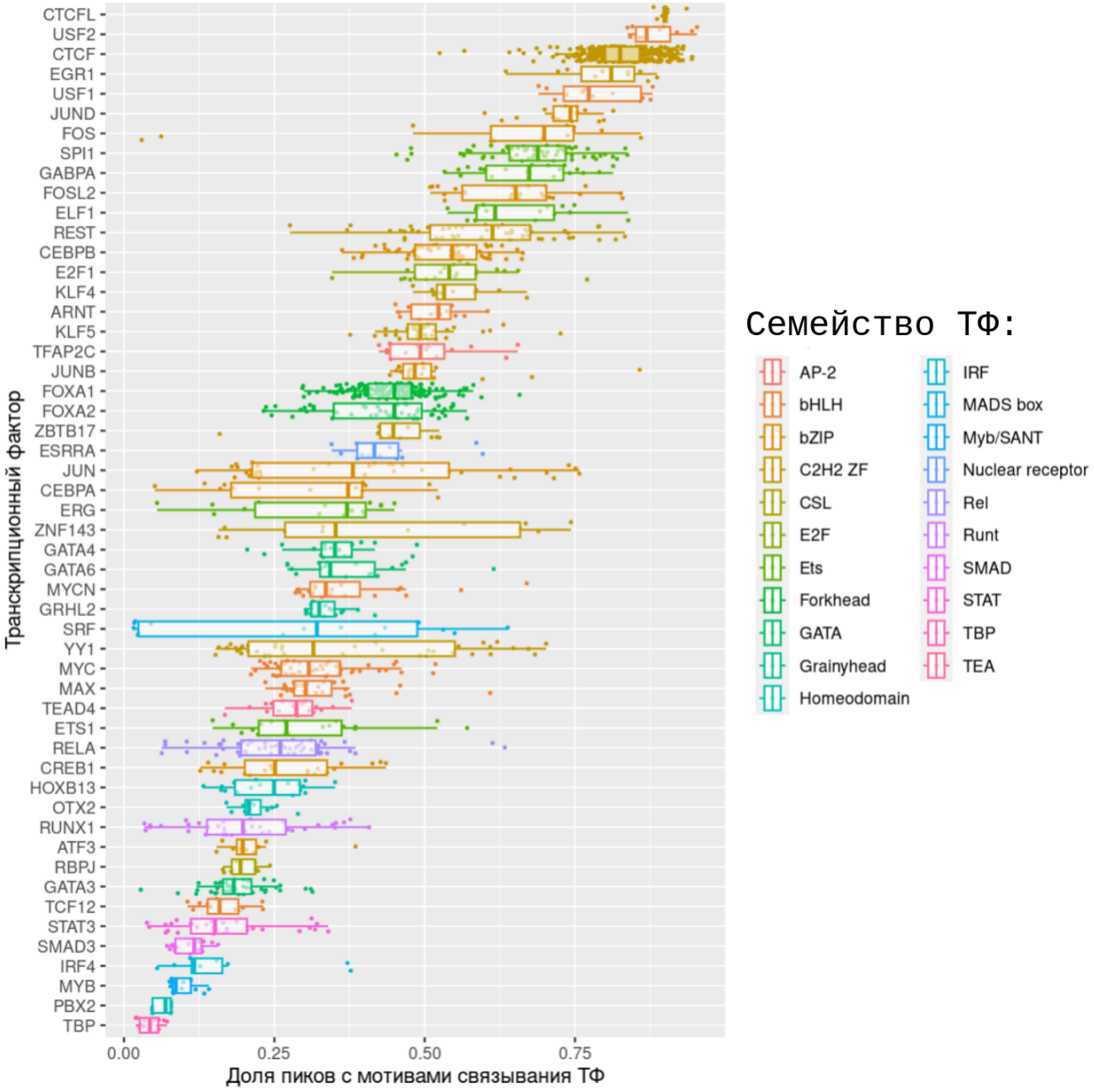


Рисунок 3.4.4 – Распределения доли пиков, в окрестностях которых (+- 30 п.н.) были найдены мотивы связывания ТФ (p-value < 0.0001). Число экспериментов для каждого ТФ > 10

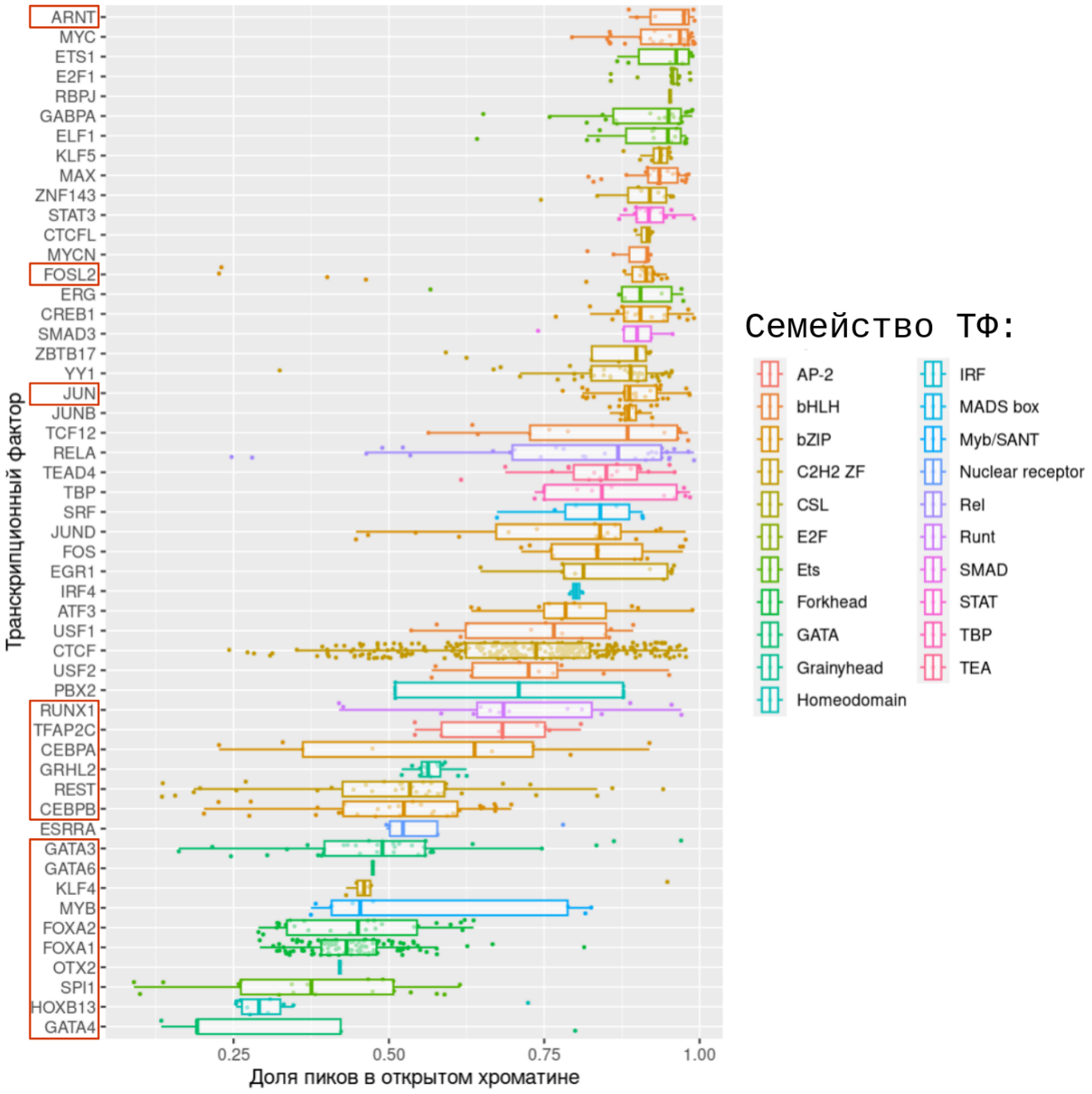


Рисунок 3.4.5 – Распределения доли пиков внутри открытого хроматина. Число экспериментов для каждого ТФ > 10. Красным отмечены ТФ, ассоциированные с процессами ремоделинга хроматина по Lemma et al. (Lemma et al., 2022).

Примечательно, что такие ТФ, как GATA4, HOXB13, SPI1, OTX2, FOXA1 и FOXA2, демонстрируют сниженную долю РСТФ в РОХ. Чтобы глубже понять это наблюдение, была проанализирована связь этих белков с процессами ремоделинга хроматина. Пионерные и фланкирующие ТФ, связанные с доступностью хроматина, были определены на основе исследования Lemma et al (Lemma et al.,

2022). Большинство ТФ, связанных с ремоделированием хроматина, демонстрируют более низкий процент РСТФ в РОХ.

Заключение по главе 3.4

Был предложен и на языке Java реализован алгоритм многостадийного применения методов коллективного выбора для выявления наиболее достоверных РСТФ на основе мета-анализа результатов ChIP-seq анализа всех экспериментов из БД GTRD для заданного ТФ, METARA. Разработанный метод поддерживает использование различных агрегирующих функций: арифметическое и геометрическое средние, медиана, L1-norm, L2-norm, а также методы, основанные на использовании Марковских цепей.

При помощи анализа групп мета-кластеров в зависимости от значений ФАФ было продемонстрировано, что чем выше ФАФ, тем чаще мета-кластеры:

1. Встречаются в районах открытого хроматина (в 91% случаев);
2. Содержит мотивы, представленные позиционной весовой матрицей, связывания соответствующего ТФ (в 85% случаев).

Таким образом было показана способность алгоритма METARA ранжировать наборы РСТФ по степени их правдоподобности.

При помощи значений ФАФ для каждого из 3426 ChIP-seq экспериментов, прошедших рекомендуемые пороги качества ENCODE, были отобраны наиболее правдоподобные РСТФ. На основании поиска МСТФ в РСТФ было продемонстрировано наличие ТФ, для которых наиболее правдободобные РСТФ имеют сниженную тенденцию содержать в себе МСТФ (менее 25% по медиане). Например, TBP, PBX2, MYB, IRF4, SMAD3, STAT3, TCF12, GATA3 и др. Также были идентифицированы ТФ, для которых менее выражена тенденция РСТФ располагаться в РОХ. Например, GATA4, HOXB13, SPI1, OTX2, FOXA1 и FOXA2. Было показано, что сниженная доля РСТФ в РОХ свойственна ТФ, ассоциированным с ремоделированием хроматина.

При помощи предложенного алгоритма были построены карты геномных районов связывания 1391 ТФ и кофакторов человека. Полученные районы вошли в состав БД GTRD и доступны по ссылке: http://gtrd.biouml.org:8888/egrid/bigBeds/hg38/ChIP-seq.

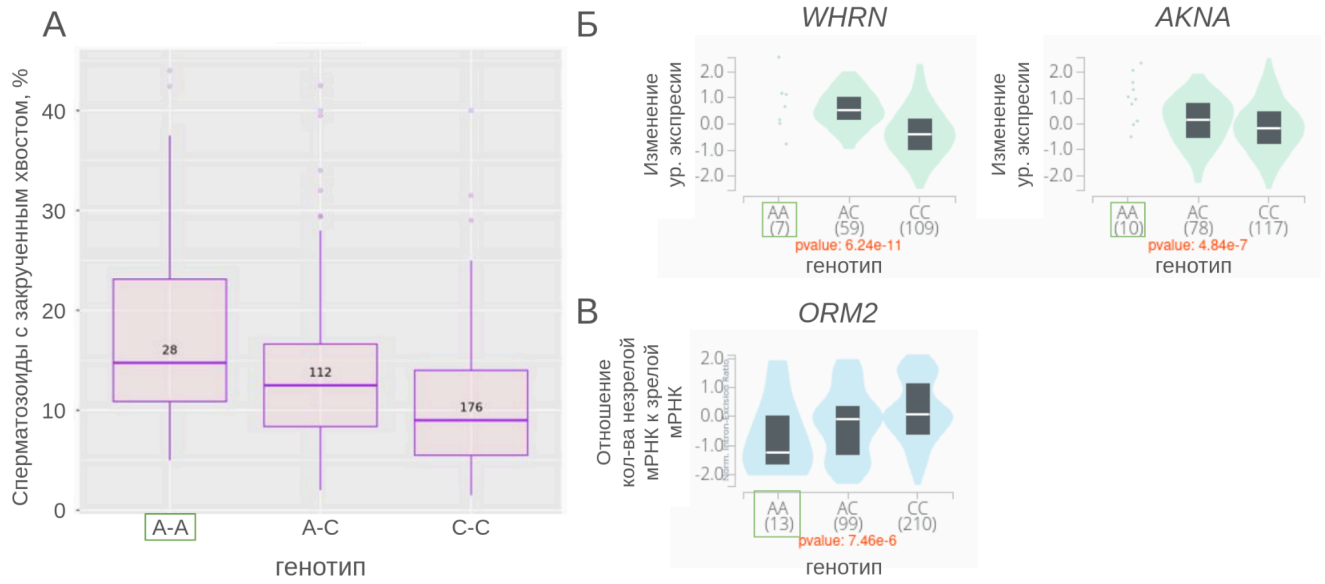
3.5 Интерпретация однонуклеотидных геномных вариаций, ассоциированных с нарушениями сперматогенеза, с точки зрения регуляции транскрипции

На основании полноэкзомного анализа ассоциации было идентифицировано 135 SNP достоверно (FDR<0.05) ассоциированных с морфологическими нарушениями сперматозоидов. Найденные геномные варианты располагаются в 63 генах, 2 из которых являются ТФ: AKNA и ZNF704. Данные SNV расположены в интронах и не влекут за собой изменений в аминокислотных последовательностях ТФ (Таблица 3.5.1). Однако SNV, расположенные в нетранслируемых областях генов могут быть причиной изменений на транскрипционном и трансляционном уровнях. Геномные вариации, расположенные в интронах, могут влиять на эффективность сплайсинга и других процессов процессинга РНК (Robert et Pelletier, 2018). В частности, интронные SNV могут приводить к пропуску экзонов или включению псевдоэкзонов. Также наличие интронных SNV может привести к изменению структуры пре-мРНК, что может пагубно отразиться на ее стабильности (Lin et al., 2019). Более того, SNV могут располагаться в сайтах связывания ТФ и быть причиной изменения аффинности связывания ТФ с ДНК (Ecker et al., 2017).

Сперва обратим внимание на геномные варианты, расположенные в границах генов, кодирующих ТФ. SNV rs2787348 (chr9:114359560), расположенный в первом интроне гена *AKNA*, достоверно ассоциирован с уменьшением процента сперматозоидов с закрученным хвостом в семенной

жидкости (см. Рисунок 3.5.1). Согласно dbSNP частота данного аллеля в сибирской популяции составляет 0.07. ТФ AKNA участвует в регуляции организации микротрубочек, которая имеет решающее значение для правильного формирования клеточных структур и клеточной мобильности (Ramírez-González et al., 2021). Таким образом, как изменения в структуре AKNA, так и изменения уровня экспрессии гена может привести к нарушению морфологии сперматозоидов, таким как обретение закрученных хвостов. Следует обратить внимание, что рассматриваемый SNV ассоциирован положительный эффект, т. е. гомозигота по rs2787348 ассоциирована с меньшим процентом сперматозоидов с закрученными хвостами. На основе анализа данных eQTL (Expression Quantitative Trait Loci) из БД GTEx также демонстрируется ассоциация данного SNV со снижением уровней экспрессии генов: *AKNA* и *WHRN* (см. Рисунок 3.5.2). Однако данный эффект не был показан в тканях органов мужской репродуктивной системы. Согласно Human Protein Atlas *WHRN* демонстрирует повышенный уровень экспрессии в ранних и поздних сперматидах. Таким образом, требуется экспериментальное подтверждение зависимости уровней экспрессии генов *AKNA* и *WHRN* с увеличением доли сперматозоидов с закрученными хвостами. Также, согласно данным sQTL (Splicing Quantitative Trait Loci) из БД GTEx SNV rs2787348 ассоциирован с повышением в семенниках отношения пре-мРНК к зрелой мРНК (Intron-Excision Ratio; IER) в гене *ORM2* (см. Рисунок 3.5.2), что говорит о снижении эффективности сплайсинга. ORM2 относится к белкам острой фазы и является одним из маркеров воспаления. Данный белок выполняет транспортную функцию, связываясь с различными гидрофобными лигандами (Nishi et al., 2011), также снижение уровня экспрессии данного гена тесно связано с иммуносупрессией в опухолях печени (Zhu et al., 2020). Согласно Human Protein Atlas *ORM2* активно экспрессируется в поздних сперматидах. Поскольку воспалительные процессы негативным образом влияют на сперматогенез (Hasan et al., 2022), повышенный уровень экспрессии *ORM2* может создавать

благоприятные условия на поздних этапах созревания сперматозоидов, подавляя активность иммунной системы.



Значение над медианой в “ящике с усами” указывает на количество образцов с выбранным генотипом. Зеленым прямоугольник отмечена гомозигота по референсному аллелю.
Рисунок 3.5.2 – (А) Распределение процента сперматозоидов с закрученным хвостом в зависимости от присутствия SNV rs2787348 в генотипе; (Б) – Взаимосвязь изменения уровней экспрессии генов: WHRN и AKNA в головном мозге в зависимости от присутствия SNV rs2787348 в генотипе; (В) – отношения пре-мРНК к зрелой мРНК гена ORM2 в семенниках в зависимости от присутствия SNV rs2787348 в генотипе.

Полученные SNV были проанализированы с точки зрения влияния на специфичность связывания ТФ. Для этого были отобраны SNV, расположенные в границах РСТФ. Однако в открытом доступе очень мало данных по РСТФ из данных, полученных на органах мужской репродуктивной системы. Поэтому из рассматриваемых SNV для дальнейшей интерпретации были отобраны только те, что входят в топ 5% РСТФ после применения алгоритма METARA ко всем имеющимся для данного ТФ ChIP-seq данным в БД GTRD. Затем, на основании данных из БД ADASTRA были отобраны SNV, демонстрирующие аллельный дисбаланс связывания соответствующих ТФ (FDR<0.05). Таким образом было отобрано 4 из 135 SNV (см. Рисунок 3.5.3).

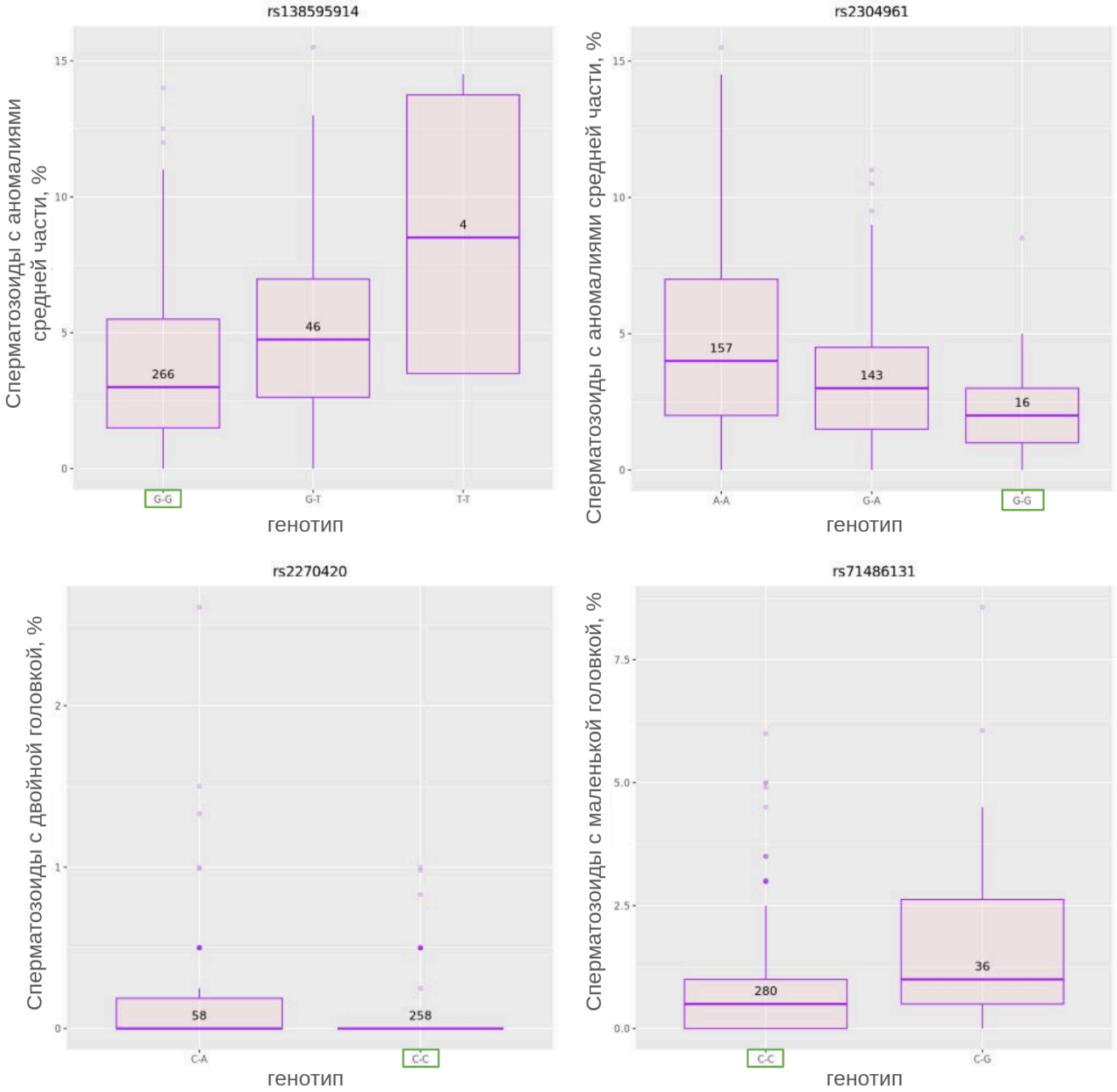


Рисунок 3.5.3 – Распределение процента сперматозоидов с различными морфологическими нарушениями сперматозоидов в зависимости от присутствия SNV в генотипе. Значение над медианой в “ящике с усами” указывает на количество образцов с выбранным генотипом. Зеленым прямоугольник отмечена гомозигота по референсному аллелю.

Также при помощи БД GTEx была получена информация об ассоциированных с SNV изменениях уровней экспрессии генов в различных тканях и органах мужской репродуктивной системы (Таблица 3.5.1). Для валидации наличия экспрессии рассматриваемых ТФ в органах мужской

репродуктивной системы, а также присутствии в данных тканях самих белков, была использована БД Human Protein Atlas.

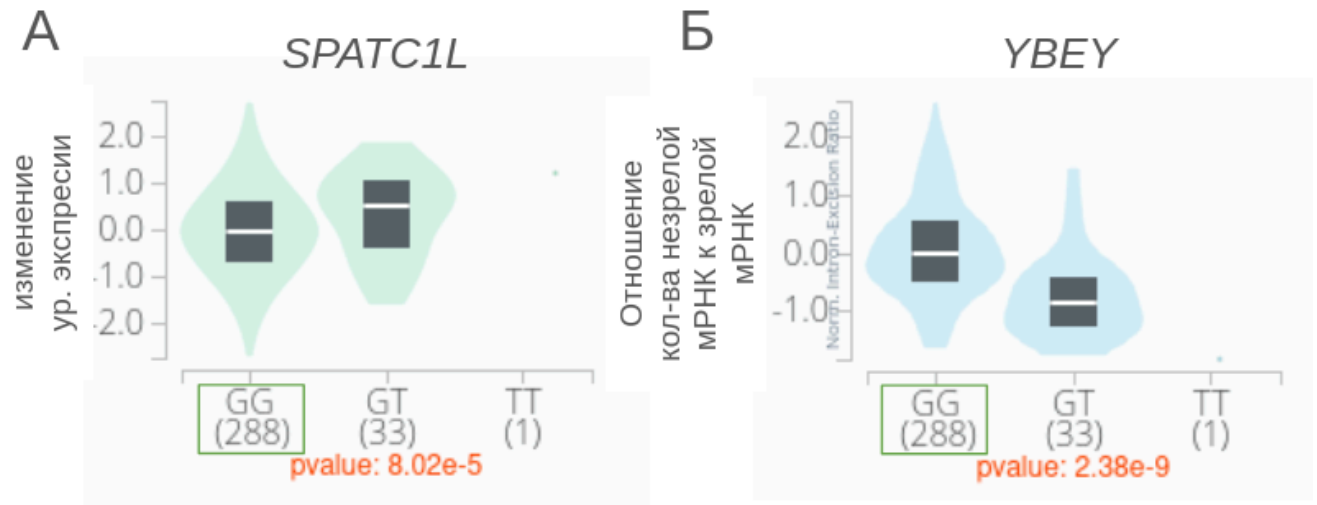
Таблица 3.5.1 – SNV, расположенные в РСТФ, для которых показан аллельный дисбаланс связывания ТФ по БД ADASTRA (FDR<0.05).

dbSNP ID	Ген	Ген: Эффект	Связанный с SNP признак	Влияние на эффективность связывания ТФ	eQTL
rs138595914	C21orf58 PCNT	PCNT : 5 Prime UTR Variant C21orf58 : 2KB Upstream Variant	Аномалии средней части	⬇AR	⬆SPATC1L в семенниках ⬇:C21orf58, DIP2A, FTCD, MCM3AP, PCNT, PRMT2, SPATC1L, YBEY
rs2304961	ABR	ABR : Intron Variant (16-ый интрон)	Аномалии средней части	⬇CTCF	—
rs2270420	GSTZ1 POMT2	POMT2 : Synonymous Variant (1-й экзон) GSTZ1 : 2KB Upstream Variant	Двойная головка	⬇SRBP2	⬇GSTZ1
rs71486131	MARVELD1	MARVELD1 : Synonymous Variant (1-й экзон)	Маленькая головка	⬆CTCF	⬆MARVELD1 ⬆ZFYVE27

rs138595914 (chr21:46324213) статистически значимо ассоциирован с увеличением процента сперматозоидов с аномалиями в средней части. В БД dbSNP частота данного аллеля в сибирской популяции составляет 0.5. Данный SNV располагается в границах РСТФ андрогенового рецептора (AR) и снижает эффективность связывания данного ТФ с ДНК. Следует обратить внимание, что AR может быть как активатором, так и репрессором. В частности, супрессорная активность может быть опосредована способностью AR рекрутировать гистоновые деацетилазы (Hodgson et al., 2008) и ДНК-метилтрансферазы (Ylitalo et al., 2021), что приводит к деацетилированию гистонов и метилированию CpG островков, соответственно. Также супрессорная активность AR может

обусловлена взаимодействием с белками-корепрессорами: NCoR и SMRT (Hodgson et al., 2008).

Согласно eQTL из БД GTEx также наблюдается ассоциация данного SNV с повышением уровня экспрессии гена *SPATC1L* в семенниках (см. Рисунок 3.5.4), а также снижением уровней экспрессии генов: *C21orf58, DIP2A, FTCD, MCM3AP, PCNT, PRMT2, SPATC1L, YBEY* в других тканях. Следует обратить внимание, что в 31 ткани *SPATC1L* снижение уровня экспрессии гена ассоциированы с данным SNV, тогда как только в семенниках наблюдается позитивная корреляция с присутствием данного аллеля.



Зеленым прямоугольник отмечена гомозигота по референсному аллелю.
Рисунок 3.5.4 – (А) – Взаимосвязь изменения уровней экспрессии гена SPATC1L в семенниках в зависимости от присутствия SNV rs138595914 в генотипе; (Б) – отношения пре-мРНК к зрелой мРНК гена *YBEY* в семенниках в зависимости от присутствия SNV rs138595914 в генотипе.

Ген *SPATC1L* (Spermatogenesis and centriole associated 1 like) специфично экспрессируется в ранних и поздних сперматидах. Более того белок SPATC1L локализован в центросомах поздних сперматид и зрелых сперматозоидов. В работе Li с соавторами (Li et al., 2022) была показана ассоциация мутаций в данном гене с ацефалией сперматозоидов. Вероятно, повышенная экспрессия данного гена также приводит к нарушению морфологии сперматозоидов, в частности, к аномалиям средней части.

Согласно данным sQTL наблюдает повышение доли зрелой мРНК гена *YBEY*. *YBEY* кодирует белок, участвующий в процессе созревания рибосом и функционировании митохондрий (Summer et al., 2020). Нарушения в функции митохондрий могут негативным образом сказаться на структуре средней части сперматозоида или функциональные проблемы, связанные с подвижностью и выживаемостью сперматозоидов.

rs2304961 (chr17:1058660) статистически значимо ассоциирован с увеличением процента сперматозоидов с аномалиями в средней части. Согласно dbSNP частота данного аллеля в сибирской популяции составляет 0.21. Данный SNV располагается в 16 интроне гена *ABR*. Продукт гена *ABR* . Ген *ABR* демонстрирует повышенный уровень экспрессии в поздних сперматидах (159.3 nTPM), помимо этого данный ген также высоко экспрессируется в астроцитах (162.9 nTPM) и клетках микроглии (139.6 nTPM). Данный SNV располагается в границах РСТФ CTCF и снижает эффективность связывания данного ТФ с ДНК. Однако в БД GTEx нет данных об ассоциации rs2304961 с изменением уровней экспрессии генов.

rs2270420 (chr14:77320520) статистически значимо ассоциирован с увеличением доли сперматозоидов с двойной головкой. В БД dbSNP частота данного аллеля в сибирской популяции составляет 0.4. Данный SNV располагается в границах РСТФ SRBP2 (Sterol regulatory element-binding protein 2; SREBF2) и снижает эффективность связывания данного ТФ с ДНК. Согласно eQTL из БД GTEx данный аллель ассоциирован с незначительным снижением уровня экспрессии гена *GSTZ1* в легких. Согласно Human Protein Atlas *GSTZ1* специфично экспрессируется в клетках гепатоцитов (416.3 nTPM) и в поздних спертматидах (400.4 nTPM).

Также рассматриваемый SNV, rs2270420, расположен в первом интроне гена *POMT2*, который демонстрирует повышенную экспрессию в семенниках. Продукт гена POMT2 . Несмотря на то, что по данным GTEx не наблюдается ассоциации с

изменением уровня экспрессии данного гена, данный SNV может влиять как на стабильность мРНК, так и на эффективность сплайсинга.

rs71486131 (chr10:97713972) представляет собой синонимичную замену в 1 интроне гена *MARVELD1* и статистически значимо ассоциирован с увеличением доли сперматозоидов с маленькой головкой. В БД dbSNP частота данного аллеля в сибирской популяции составляет 0.5. Согласно eQTL из БД GTEx также наблюдается ассоциация данного SNV с повышением уровней экспрессии генов: *MARVELD1* и *ZFYVE27*; Однако данных по изменениям уровней экспрессии генов в органах мужской репродуктивной системы в GTEx обнаружено не было.

Заключение к главе 3.5

На основании полноэкзомного анализа ассоциации было идентифицировано 135 SNV достоверно (FDR<0.05) ассоциированных с морфологическими нарушениями сперматозоидов. Найденные геномные варианты располагаются в 63 генах, 2 из которых являются синонимичными заменами в ТФ: AKNA и ZNF704. SNV rs2787348 (chr9:114359560), Располагается в первом интроне гена AKNA и достоверно ассоциирован с уменьшением процента сперматозоидов с закрученным хвостом в семенной жидкости. ТФ AKNA участвует в регуляции организации микротрубочек, которая имеет решающее значение для правильного формирования клеточных структур и клеточной мобильности.

Помимо этого были идентифицированы 4 SNV: rs138595914, rs2304961, rs2270420, rs71486131, которые, с одной стороны, располагаются в наиболее правдоподобных на основании значений ФАФ РСТФ, с другой стороны, демонстрируют аллельный дисбаланс связывания соответствующих ТФ (FDR<0.05). В частности, для SNV rs138595914 (chr21:46324213), ассоциированного с увеличением процента сперматозоидов с аномалиями в

средней части, было показано снижение специфичности связывания AR с ДНК, сопровождающееся увеличением уровня экспрессии гена *SPATC1L* в семенниках.

ЗАКЛЮЧЕНИЕ

Анализ воспроизводимости РСТФ разными алгоритмами идентификации РСТФ в рамках одного ChIP-seq эксперимента показал, что в среднем полностью воспроизводится ~10% от общего числа РСТФ, а наиболее представленной группой является группа F1 (в среднем ~65% от общего числа РСТФ).

Было показано, что степень воспроизводимости РСТФ разными алгоритмами напрямую связана с правдоподобностью рассматриваемых РСТФ. Наблюдаются статистически значимые различия между подгруппами РСТФ в контексте расположения РСТФ в областях открытого хроматина, более консервативных районах, а также в районах, демонстрирующих более выраженное обогащение мотивами связывания ТФ. Такая вариативность подчеркивает важность сопоставления результатов работы различных алгоритмов для получения набора наиболее правдоподобных РСТФ.

В рамках данной работы было показано, что существуют ChIP-seq эксперименты, в которых F1 РСТФ содержат больший процент правдоподобных РСТФ, по сравнению с другими экспериментами. Был предложен и валидирована новый метод, для оценки доли ложно идентифицированных РСТФ в ChIP-seq эксперименте на основе анализа пересечения результатов работы 4 алгоритмов идентификации РСТФ — FPCM. Также была разработана оценка доли ложно неидентифицированных РСТФ в ChIP-seq эксперименте на основании пересечения нескольких алгоритмов идентификации РСТФ — FNCM. Для платформы BioUML на языке Java был реализован алгоритм расчета значений FPCM и FNCM.

В рамках данной работы было показано, что даже для ChIP-seq экспериментов с высоким качеством FPCM позволяет идентифицировать поднаборы экспериментов, которые демонстрируют:

- сниженное количество МСТФ, представленные позиционной весовой матрицей, связывания соответствующего ТФ, в F1 РСТФ;
  сниженное количество F1 РСТФ в РОХ;
- более низкую воспроизводимость F1 РСТФ в других ChIP-seq экспериментах для выбранного ТФ;
- более низкую эволюционную консервативность районов с F1 РСТФ.

Также на основании пересечения F1 РСТФ с другими типами данных было продемонстрировано, что на основании характеристики FPCM даже среди качественных ChIP-seq экспериментов можно выделить эксперименты, для которых характерно снижение доли правдоподобных РСТФ в F1 РСТФ. Было продемонстрировано, что повышенные значения FPCM могут выступать рекомендацией к удалению из дальнейшего анализа группы F1 РСТФ.

Таким образом, совместное использование разработанных методов, FPCM и FNCM, в сочетании с другими оценками качества данных позволяет комплексно подходить к оценке качества данных и выявлять наборы наиболее достоверных РСТФ.

Для выявления наиболее воспроизводимых РСТФ на основе мета-анализа результатов ChIP-seq данных всех экспериментов из БД GTRD для заданного ТФ был предложен и реализован алгоритм многостадийного применения методов коллективного выбора – METARA. Разработанный метод поддерживает использование различных агрегирующих функций: арифметическое и геометрическое средние, медиана, L1-norm, L2-norm, а также методы, основанные на использовании Марковских цепей.

На основании значений ФАФ для каждого из 3426 ChIP-seq экспериментов, прошедших рекомендуемые пороги качества ENCODE, были отобраны наиболее правдоподобные РСТФ. Были идентифицированы ТФ, для которых менее выражена тенденция РСТФ располагаться в РОХ. Например, GATA4, HOXB13, SPI1, OTX2, FOXA1 и FOXA2. Было показано, что сниженная доля РСТФ в РОХ

свойственна ТФ, потнциально ассоциированна с цчастем ТФ в ремоделинге хроматина.

На основании полноэкзомного анализа ассоциаций было идентифицировано 135 SNV достоверно (FDR<0.05) ассоциированных с морфологическими нарушениями сперматозоидов. Найденные однонуклеотидные геномные варианты располагаются в 63 генах, 2 из которых кодируют ТФ: AKNA и ZNF704.

Также были идентифицированы 4 SNV: rs138595914, rs2304961, rs2270420, rs71486131, которые, с одной стороны, располагаются в наиболее правдодобных на основании значений ФАФ РСТФ, с другой стороны, демонстрируют аллельный дисбаланс связывания соответствующих ТФ (FDR<0.05). В частности, для SNV rs138595914 (chr21:46324213), ассоциированного с увеличением процента сперматозоидов с аномалиями в средней части, было показано снижение специфичности связывания AR с ДНК, сопровождающееся увеличением уровня экспрессии гена SPATC1L в семенниках.

Рекомендации и перспективы дальнейшей разработки темы

Планируется дополнительная валидация и исследование применимости разработанного метода, FNCM, для оценки количества РСТФ на основании сравнения большого набора схожих ChIP-seq экспериментов. Также необходимы дальнейшие исследования для реализации алгоритма определения оптимального порога для FPCM для принятия решения об удалении F1 РСТФ в ChIP-seq экспериментах.

Одним из перспективных направлений исследования является создание алгоритма определения порога для выявления наиболее воспроизводимых РСТФ на основании значений ФАФ METARA, который бы учитывал вариабельность представленных условий проведения ChIP-seq экспериментов.

В рамках подзадачи исследования нарушений сперматогенеза планируется дополнительно исследовать два этноса, проживающие на территории Российской Федерации: буряты и якуты. Данный анализ позволит сформировать более полную картину о популяционной специфичности ассоциации однонуклеотидных вариантов со сниженным репродуктивным потенциалом.

ВЫВОДЫ

1. В базу данных GTRD внесена информация о 1347 ChIP-seq экспериментах, что позволило создать уникальную коллекцию из 15982 единообразно обработанных ChIP-seq экспериментов для человека, описывающих районы связывания 1391 транскрипционного фактора и кофактора. В базу данных GTRD внесено описание 1701 DNase-seq эксперимента и реализован конвейер их обработки, что позволило сформировать коллекцию районов открытого хроматина для 444 различных тканей и клеточных типов человека.
2. Разработаны и реализованы в виде программных модулей для биоинформатической платформы BioUML два новых метода оценки качества ChIP-seq данных на основе анализа согласованности результатов четырёх алгоритмов идентификации районов связывания транскрипционных факторов (MACS2, GEM, SISSRs и PICS):
    - метод оценки доли ложно идентифицированных районов связывания транскрипционных факторов (FPCM) - оценивает отклонение от распределения Пуассона доли районов, идентифицированных только одним из четырёх алгоритмов в ChIP-seq эксперименте;
    - метод оценки доли ложно неидентифицированных районов связывания транскрипционных факторов (FNCM) - является адаптацией экологических подходов по оценке размеров популяции, примененной для расчета верхней границы количества таких районов в ChIP-seq эксперименте и оценки вклада в это значение результатов применения каждого из алгоритмов идентификации районов связывания транскрипционных факторов.
3. Разработан и реализован в виде программного модуля для биоинформатической платформы BioUML новый алгоритм, METARA, для

приоритезации наиболее воспроизводимых районов связывания транскрипционных факторов путем вычисления значения финальной агрегирующей функции. При помощи предложенного алгоритма были построены карты геномных районов связывания 1391 транскрипционного фактора и кофактора человека для базы данных GTRD. Для 119 транскрипционных факторов человека, наиболее полно представленных в базе данных GTRD, показана корреляция между значениями финальной агрегирующей функцией и:

- расположением районов связывания транскрипционных факторов в районах открытого хроматина (91% случаев);
- наличием в районах связывания транскрипционных факторов мотивов связывания транскрипционных факторов (85% случаев).

4. На основе анализа данных полноэкзомного секвенирования впервые идентифицированы ассоциации 135 однонуклеотидных геномных вариантов с различными нарушениями морфологии сперматозоидов человека. Два однонуклеотидных варианта являются синонимичными заменами в генах, кодирующих транскрипционные факторы: AKNA и ZNF704. Выявлено четыре однонуклеотидных варианта: rs138595914, rs2304961, rs2270420, rs71486131, которые расположены в наиболее воспроизводимых геномных районах связывания трёх транскрипционных факторов: AR, CTCF и SRBP2, участвующих в регуляции сперматогенеза, и влияют на эффективность их связывания с ДНК.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, входящих в наукометрические базы Web of Science и Scopus:

1. **Kolmykov S.**, Yevshin I.S., Kulyashov M., Sharipov R.N., Kondrakhin Y., Makeev V.J., Kulakovskiy I.V., Kel A., Kolpakov F. GTRD: an integrated view of transcription regulation //Nucleic Acids Research – 2021. – Т. 49 – № D1. – С. D104-D111 (Q1).
2. **Kolmykov S.K.**, Kondrakhin Y.V., Yevshin I.S., Sharipov R.N., Ryabova A.S., Kolpakov F.A. Population size estimation for quality control of ChIP-Seq datasets //PloS ONE – 2019. – Т. 14. – №. 8. – С. e0221760 (Q1).
3. **Kolmykov S.**, Vasiliev G., Osadchuk L., Kleschev M., Osadchuk A. Whole-Exome Sequencing Analysis of Human Semen Quality in Russian Multiethnic Population //Frontiers in Genetics – 2021 – Т. 12. – С. 662846 (Q2).
4. Vorontsov IE, Eliseeva IA, Zinkevich A, Nikonov M, Abramov S, Boytsov A, Kamenets V, Kasianova A, **Kolmykov S**, Yevshin IS, Favorov A, Medvedeva YA, Jolma A, Kolpakov F, Makeev VJ, Kulakovskiy IV. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors //Nucleic Acids Research – 2024 – Т. 52. – №. D1. – С. D154-D163 (Q1).
5. Kolpakov F., Akberdin I.R., Kiselev I.N., **Kolmykov S.K.**, Kondrakhin Y., Kulyashov M., Kutumova E.O., Pintus S.S., Ryabova A., Sharipov R.N., Yevshin I.S., Zhatchenko S., Kel A. BioUML – towards a universal research platform //Nucleic Acids Research – 2022. – Т. 50. – №. W1. – С. W124-W131 (Q1).
6. Kolpakov F., Akberdin I., Kashapov T., Kiselev I., **Kolmykov S.**, Kondrakhin Y., Kutumova E., Mandrik N., Pintus S., Ryabova A., Sharipov R.N., Yevshin I., Kel A. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data //Nucleic Acids Research – 2019. – Т. 47. – №. W1. – С.

W225-W233 (Q1).

7. Abramov S., Boytsov A., Bykova D., Penzar D.D., Yevshin I., **Kolmykov S.K.**, Fridman M.V., Favorov A.V., Vorontsov I.E., Baulin E., Kolpakov F.A., Makeev V.J., Kulakovskiy I.V. Landscape of allele-specific transcription factor binding in the human genome //Nature Communications – 2021. – Т. 12. – №. 1. – С. 2751 (Q1).
8. Boytsov A, Abramov S, Aiusheeva AZ, Kasianova AM, Baulin E, Kuznetsov IA, Aulchenko YS, **Kolmykov S**, Yevshin I, Kolpakov F, Vorontsov IE, Makeev VJ, Kulakovskiy IV. ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs //Nucleic Acids Research – 2022 – Т. 50. – №. W1. – С. W51-W56 (Q1).
9. Yevshin I., Sharipov R., **Kolmykov S.**, Kondrakhin Y., Kolpakov F. GTRD: a database on gene transcription regulation-2019 update //Nucleic Acids Research – 2018. – Т. 47. – №. D1. – С. D100-D105 (Q1).

Другие публикации из списка Scopus:

10. **Kolmykov S. K.**, Kondrakhin Y. V., Sharipov R. N., Yevshin I. S., Ryabova A. S., & Kolpakov F. A. (2020, July). Meta-analysis of ChIP-seq datasets through the rank aggregation approach //2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB). – IEEE, 2020. – С. 180-184.
11. Kulyashov M. A., **Kolmykov S. K.**, Yevshin I. S., & Kolpakov F. A. (2020, July). Advanced data curation in GTRD database: hierarchical dictionaries of cell types and experimental factors //2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB). – IEEE, 2020. – С. 23-27.
12. **Kolmykov S.K.**, Evshin I.S., Kolpakov F.A. Analysis of NGS Data on the Transcriptional Regulation //CEUR Workshop Proceedings. – 2019. – С. 19-22.
13. Kulyashov M.A., **Kolmykov S.K.,** Evshin I.S., Kolpakov F.A. Description, Characteristic And Algorithm For Creation Of A Dictionary Of Cell Types And

Tissues In The Gtrd Database //CEUR Workshop Proceedings. – 2020. – Т. 2569. – С. 13-18.

Публикации в других изданиях, в сборниках трудов конференций:

14. **Kolmykov S**, Kulyashov M, Sokolova T, Prasolov D, Kolpakov F. Exploring the interplay between reproducibility of open chromatin regions and transcription factor functional activity //Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2024). – 2024. – С. 127-129.
15. **Kolmykov S,** Kondrakhin Y, Kolpakov F. Relationship Between Transcription Factor Binding Regions and Open Chromatin Regions in Human Based on GTRD Data //Proceedings of 11th Moscow Conference on Computational Molecular Biology MCCMB'23 (August 3-6, 2023). – 2023. – С. 1-4.
16. Осадчук А.В., Васильев Г.В., **Колмыков С.К.**, Иванов М.К., Прасолова М.А., Клещев М.А., Осадчук Л.В., Евразийский тренд фенотипической и генетической изменчивости мужского репродуктивного потенциала в популяциях Российской Федерации и Республики Беларусь //Сборник тезисов XXIV съезда физиологического общества им. ИП Павлова. – 2023. – С. 329-329.
17. **Kolmykov S**, Kondrakhin Y, Sharipov R, Yevshin I, Ryabova A, Kolpakov F. Transcription factor binding sites: data integration, stable identifiers and incremental builds //Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2022). – 2022. – С. 77-77.
18. Sharipov R, Kondrakhin Y, **Kolmykov S**, Yevshin I, Ryabova A, Kolpakov F. Heterogeneity of transcription factor binding sites within ChIP-Seq datasets //Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2022). – 2022. – С. 94-94.
19. **Колмыков С.К.**, Евшин И.С., Колпаков Ф.А., Анализ NGS Данных По Регуляции Транскрипции //Распределенные Информационно-Вычислительные

Ресурсы. Цифровые Двойники И Большие Данные. (DICR-2019). Труды XVII Международной конференции. – 2019. – С 107-112.

20. Куляшов М.А., **Колмыков С.К.**, Евшин И.С., Колпаков Ф.А., Описание, Характеристика И Алгоритм Создания Словаря Клеточных Типов И Тканей В Базе Данных GTRD //Распределенные Информационно-Вычислительные Ресурсы. Цифровые Двойники И Большие Данные. (DICR-2019). Труды XVII Международной конференции. – 2019. – С 119-125.
21. **Kolmykov S.K.**, Kleshev M.A., Vasiliev G.V., Osadchuk A.V., Ponomarenko M.P., Osadchuk L.V., Whole-Exome Sequencing Association Studies On Impaired Spermatogenesis In Different Ethnic Groups In Russia //Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2020). – 2020. – С. 462-463.
22. Sharipov R.N., Yevshin I.S., Kondrakhin Yu.V., Ryabova A.S., **Kolmykov S.K.**, Kolpakov F.A., Peak Caller Comparison Through Quality Control Of Chip-Seq Datasets //Bioinformatics of genome regulation and structure/systems biology (BGRS/SB-2020). – 2020. – С. 105-106.
23. **Kolmykov S.K.**, Kondrakhin Yu.V., Yevshin I.S., Sharipov R.N., Kulyashov M.A., Kolpakov F.A, Human cistrome - genome-wide map of human transcription factor binding sites derived from GTRD database //Moscow Conference on Computational Molecular Biology (MCCMB). – 2019.
24. Yevshin I.S., Sharipov R.N., **Kolmykov S.K**., Kondrakhin Yu.V., Kolpakov F.A, Gtrd: a database on gene transcription regulation //Биотехнология: состояние и перспективы развития. – 2019. – С. 389-390.
25. **Kolmykov S.**, Kondrakhin Y., Kolpakov F. New method for estimation of number of transcription factor binding sites using results of processing of ChIP-seq data by different peak callers //Systems Biology and Bioinformatics (SBB-2018). – 2018. – С. 52-52.

СПИСОК ЛИТЕРАТУРЫ

1. Abugessaisa I. et al. FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs //Nucleic acids research. – 2021. – Т. 49. – №. D1. – С. D892-D898.
2. Aerts S. et al. Gene prioritization through genomic data fusion //Nature biotechnology. – 2006. – Т. 24. – №. 5. – С. 537-544.
3. Ambrosini G. et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study //Genome biology. – 2020. – Т. 21. – С. 1-18.
4. Amemiya H. M., Kundaje A., Boyle A. P. The ENCODE blacklist: identification of problematic regions of the genome //Scientific reports. – 2019. – Т. 9. – №. 1. – С. 9354.
5. Badgeley M. A., Sealfon S. C., Chikina M. D. Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation //Bioinformatics. – 2015. – Т. 31. – №. 2. – С. 209-215.
6. Bailey T. et al. Practical guidelines for the comprehensive analysis of ChIP-seq data //PLoS computational biology. – 2013. – Т. 9. – №. 11. – С. e1003326.
7. Ballester B. et al. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways //Elife. – 2014. – Т. 3. – С. e02626.
8. Beg M. M. S., Ahmad N. Fuzzy logic and rank aggregation for the world wide web //Fuzzy Logic and the Internet. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2004. – С. 27-46.
9. Benjamini Y., Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing //Journal of the Royal statistical society: series B (Methodological). – 1995. – Т. 57. – №. 1. – С. 289-300.
10. Bibikova M. et al. High-throughput DNA methylation profiling using universal

bead arrays //Genome research. – 2006. – T. 16. – №. 3. – C. 383-393.

11. Blanco M., Cocquet J. Genetic factors affecting sperm chromatin structure //Genetic damage in human spermatozoa. – 2019. – C. 1-28.
12. Bolt C. C., Duboule D. The regulatory landscapes of developmental genes //Development. – 2020. – T. 147. – №. 3. – C. dev171736.
13. Boyle A. P. et al. High-resolution mapping and characterization of open chromatin across the genome //Cell. – 2008. – T. 132. – №. 2. – C. 311-322.
14. Boyle A. P. et al. F-Seq: a feature density estimator for high-throughput sequence tags //Bioinformatics. – 2008. – T. 24. – №. 21. – C. 2537-2538.
15. Boyle A. P. et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells //Genome research. – 2011. – T. 21. – №. 3. – C. 456-464.
16. Buenrostro J. D. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position //Nature methods. – 2013. – T. 10. – №. 12. – C. 1213-1218.
17. Cannarella R. et al. Molecular biology of spermatogenesis: novel targets of apparently idiopathic male infertility //International journal of molecular sciences. – 2020. – T. 21. – №. 5. – C. 1728.
18. Castro-Mondragon J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles //Nucleic acids research. – 2022. – T. 50. – №. D1. – C. D165-D173.
19. Chao A. Estimating the population size for capture-recapture data with unequal catchability //Biometrics. – 1987. – C. 783-791.
20. Chapman D. G. Some properties of the hypergeometric distribution with applications to zoological censuses //Univ. Calif. Stat. – 1951. – T. 1. – C. 60-131.
21. Chèneby J. et al. ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments //Nucleic acids research. – 2020. – T. 48. – №. D1. – C. D180-D188.

22. Czipa E. et al. ChIPSummitDB: a ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them //Database. – 2020. – T. 2020. – C. baz141.

23. DeConde R. P. et al. Combining results of microarray experiments: a rank aggregation approach //Statistical applications in genetics and molecular biology. – 2006. – T. 5. – №. 1.

24. Deng K. et al. Bayesian aggregation of order-based rank data //Journal of the American Statistical Association. – 2014. – T. 109. – №. 507. – C. 1023-1039.

25. Dostie J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements //Genome research. – 2006. – T. 16. – №. 10. – C. 1299-1309.

26. Du L. et al. Novel gene regulation in normal and abnormal spermatogenesis //Cells. – 2021. – T. 10. – №. 3. – C. 666.

27. Dupont S., Wickström S. A. Mechanical regulation of chromatin and transcription //Nature Reviews Genetics. – 2022. – T. 23. – №. 10. – C. 624-643.

28. Dwork C. et al. Rank aggregation methods for the web //Proceedings of the 10th international conference on World Wide Web. – 2001. – C. 613-622.

29. Ecker S. et al. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types //Genome biology. – 2017. – T. 18. – C. 1-17.

30. Eder T., Grebien F. Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection //Genome Biology. – 2022. – T. 23. – №. 1. – C. 119.

31. Evenson D. P. Evaluation of sperm chromatin structure and DNA strand breaks is an important part of clinical male fertility assessment //Translational andrology and urology. – 2017. – T. 6. – №. Suppl 4. – C. S495.

32. Freund Y. et al. An efficient boosting algorithm for combining preferences //Journal of machine learning research. – 2003. – T. 4. – №. Nov. – C. 933-969.

33. Freund Y., Schapire R. E. Large margin classification using the perceptron algorithm //Proceedings of the eleventh annual conference on Computational learning theory. – 1998. – С. 209-217.

34. Fu Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer //Genome biology. – 2014. – Т. 15. – С. 1-15.

35. Fullwood M. J. et al. An oestrogen-receptor-α-bound human chromatin interactome //Nature. – 2009. – Т. 462. – №. 7269. – С. 58-64.

36. Gaffney D. J. et al. Dissecting the regulatory architecture of gene expression QTLs //Genome biology. – 2012. – Т. 13. – С. 1-15.

37. Giresi P. G. et al. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin //Genome research. – 2007. – Т. 17. – №. 6. – С. 877-885.

38. Green C. D. et al. A comprehensive roadmap of murine spermatogenesis defined by single-cell RNA-seq //Developmental cell. – 2018. – Т. 46. – №. 5. – С. 651-667. e10.

39. Grosveld F., van Staalduinen J., Stadhouders R. Transcriptional regulation by (super) enhancers: from discovery to mechanisms //Annual review of genomics and human genetics. – 2021. – Т. 22. – №. 1. – С. 127-146.

40. Guo Y., Mahony S., Gifford D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. – 2012.

41. Gusmao E. G. et al. Analysis of computational footprinting methods for DNase sequencing experiments //Nature methods. – 2016. – Т. 13. – №. 4. – С. 303-309.

42. Gusmao E. G. et al. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications //Bioinformatics. – 2014. – Т. 30. – №. 22. – С. 3143-3151.

43. Handel A. E. et al. Most brain disease-associated and eQTL haplotypes are not located within transcription factor DNase-seq footprints in brain //Human

Molecular Genetics. – 2017. – T. 26. – №. 1. – C. 79-89.

44. Håndstad T. et al. A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites //PloS one. – 2011. – T. 6. – №. 4. – C. e18430.

45. Harmanci A., Rozowsky J., Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework //Genome biology. – 2014. – T. 15. – C. 1-15.

46. Hasan H. et al. Mechanism of inflammatory associated impairment of sperm function, spermatogenesis and steroidogenesis //Frontiers in Endocrinology. – 2022. – T. 13. – C. 897029.

47. Haury A. C., Gestraud P., Vert J. P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures //PloS one. – 2011. – T. 6. – №. 12. – C. e28210.

48. Hirschhorn J. N. et al. A comprehensive review of genetic association studies //Genetics in medicine. – 2002. – T. 4. – №. 2. – C. 45-61.

49. Hodgson M. C. et al. Structural basis for nuclear receptor corepressor recruitment by antagonist-liganded androgen receptor //Molecular cancer therapeutics. – 2008. – T. 7. – №. 10. – C. 3187-3194.

50. Hower V., Evans S. N., Pachter L. Shape-based peak identification for ChIP-Seq //BMC bioinformatics. – 2011. – T. 12. – C. 1-9.

51. Ioannidis N. M. et al. FIRE: functional inference of genetic variants that regulate gene expression //Bioinformatics. – 2017. – T. 33. – №. 24. – C. 3895-3901.

52. Jankowski A., Tiuryn J., Prabhakar S. Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data //Bioinformatics. – 2016. – T. 32. – №. 16. – C. 2419-2426.

53. John S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns //Nature genetics. – 2011. – T. 43. – №. 3. – C. 264-268.

54. Johnson D. S. et al. Genome-wide mapping of in vivo protein-DNA interactions

//Science. – 2007. – T. 316. – №. 5830. – C. 1497-1502.

55. Kähärä J., Lähdesmäki H. BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data //Bioinformatics. – 2015. – T. 31. – №. 17. – C. 2852-2859.

56. Kasowski M. et al. Variation in transcription factor binding among humans //science. – 2010. – T. 328. – №. 5975. – C. 232-235.

57. Keene J. D., Komisarow J. M., Friedersdorf M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts //Nature protocols. – 2006. – T. 1. – №. 1. – C. 302-307.

58. Keilwagen J., Posch S., Grau J. Accurate prediction of cell type-specific transcription factor binding //Genome biology. – 2019. – T. 20. – C. 1-17.

59. Kendall M. G. A new measure of rank correlation //Biometrika. – 1938. – T. 30. – №. 1-2. – C. 81-93.

60. Kent W. J. et al. The human genome browser at UCSC //Genome research. – 2002. – T. 12. – №. 6. – C. 996-1006.

61. Kharchenko P. V., Tolstorukov M. Y., Park P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins //Nature biotechnology. – 2008. – T. 26. – №. 12. – C. 1351-1359.

62. Kircher M. et al. A general framework for estimating the relative pathogenicity of human genetic variants //Nature genetics. – 2014. – T. 46. – №. 3. – C. 310-315.

63. Klepikova A. V. et al. Effect of method of deduplication on estimation of differential gene expression using RNA-seq //PeerJ. – 2017. – T. 5. – C. e3091.

64. Kleshchev M., Osadchuk L., Osadchuk A. Age-related changes in sperm morphology and analysis of multiple sperm defects //Frontiers in Bioscience-Scholar. – 2023. – T. 15. – №. 3. – C. 12.

65. Kolde R. et al. Robust rank aggregation for gene list integration and

meta-analysis //Bioinformatics. – 2012. – T. 28. – №. 4. – C. 573-580.

66. Kolpakov F. et al. BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data //Nucleic acids research. – 2019. – T. 47. – №. W1. – C. W225-W233.

67. Koohy H. et al. A comparison of peak callers used for DNase-Seq data //PloS one. – 2014. – T. 9. – №. 5. – C. e96303.

68. Kruger T. F. et al. New method of evaluating sperm morphology with predictive value for human in vitro fertilization //Urology. – 1987. – T. 30. – №. 3. – C. 248-251.

69. Kulakovskiy I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis //Nucleic acids research. – 2018. – T. 46. – №. D1. – C. D252-D259.

70. Kulyashov M. A. et al. Description, characteristic and algorithm for creation of a dictionary of cell types and tissues in the GTRD database //CEUR Workshop Proceedings. – 2020. – T. 2569. – C. 13-18.

71. Kuznetsova T. et al. Transcriptional and epigenetic regulation of macrophages in atherosclerosis //Nature Reviews Cardiology. – 2020. – T. 17. – №. 4. – C. 216-228.

72. Laajala T. D. et al. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments //BMC genomics. – 2009. – T. 10. – C. 1-15.

73. Lambert S. A. et al. The human transcription factors //Cell. – 2018. – T. 172. – №. 4. – C. 650-665.

74. Lamparter D. et al. Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility //PLoS computational biology. – 2017. – T. 13. – №. 1. – C. e1005311.

75. Landt S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia //Genome research. – 2012. – Т. 22. – №. 9. – С. 1813-1831.

76. Langmead B., Salzberg S. L. Fast gapped-read alignment with Bowtie 2 //Nature methods. – 2012. – Т. 9. – №. 4. – С. 357-359.

77. Lanumteang K., Böhning D. An extension of Chao's estimator of population size based on the first three capture frequency counts //Computational Statistics & Data Analysis. – 2011. – Т. 55. – №. 7. – С. 2302-2311.

78. Lee D. et al. A method to predict the impact of regulatory variants from DNA sequence //Nature genetics. – 2015. – Т. 47. – №. 8. – С. 955-961.

79. Lemma R. B. et al. Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers //Epigenetics & chromatin. – 2022. – Т. 15. – №. 1. – С. 13.

80. Liang K., Keleş S. Normalization of ChIP-seq data with control //BMC bioinformatics. – 2012. – Т. 13. – С. 1-10.

81. Lieberman-Aiden E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome //science. – 2009. – Т. 326. – №. 5950. – С. 289-293.

82. Lin H. et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants //Genome biology. – 2019. – Т. 20. – С. 1-16.

83. Lin S., Ding J. Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies //Biometrics. – 2009. – Т. 65. – №. 1. – С. 9-18.

84. Lin S. Rank aggregation methods //Wiley Interdisciplinary Reviews: Computational Statistics. – 2010. – Т. 2. – №. 5. – С. 555-570.

85. Li Q. et al. Measuring reproducibility of high-throughput experiments. – 2011.

86. Lister R. et al. Human DNA methylomes at base resolution show widespread

epigenomic differences //nature. – 2009. – T. 462. – №. 7271. – C. 315-322.

87. Liu Y. T. et al. Supervised rank aggregation //Proceedings of the 16th international conference on World Wide Web. – 2007. – C. 481-490.

88. Li X. et al. A Bayesian latent variable approach to aggregation of partial and top‐ranked lists in genomic studies //Statistics in medicine. – 2018. – T. 37. – №. 28. – C. 4266-4278.

89. Li X., Wang X., Xiao G. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications //Briefings in bioinformatics. – 2019. – T. 20. – №. 1. – C. 178-189.

90. Li Y. Z. et al. Biallelic mutations in spermatogenesis and centriole-associated 1 like (SPATC1L) cause acephalic spermatozoa syndrome and male infertility //Asian journal of andrology. – 2022. – T. 24. – №. 1. – C. 67-72.

91. Li Z. et al. Identification of transcription factor binding sites using ATAC-seq //Genome biology. – 2019. – T. 20. – C. 1-21.

92. Luo Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal //Nucleic acids research. – 2020. – T. 48. – №. D1. – C. D882-D889.

93. Macintyre G. et al. is-rSNP: a novel technique for in silico regulatory SNP detection //Bioinformatics. – 2010. – T. 26. – №. 18. – C. i524-i530.

94. Marinov G. K. et al. Large-scale quality analysis of published ChIP-seq data //G3: Genes, Genomes, Genetics. – 2014. – T. 4. – №. 2. – C. 209-223.

95. McCrea R. S., Morgan B. J. T. Analysis of capture-recapture data. – CRC Press, 2014.

96. Meinshausen N., Bühlmann P. Stability selection //Journal of the Royal Statistical Society Series B: Statistical Methodology. – 2010. – T. 72. – №. 4. – C. 417-473.

97. Meissner A. et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis //Nucleic acids research. – 2005. – T.

33. – №. 18. – С. 5868-5877.

98. Merkulov V. M., Leberfarb E. Y., Merkulova T. I. Regulatory SNPs and their widespread effects on the transcriptome //Journal of biosciences. – 2018. – Т. 43. – С. 1069-1075.

99. Moore J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes //Nature. – 2020. – Т. 583. – №. 7818. – С. 699-710.

100. Micsinai M. et al. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments //Nucleic acids research. – 2012. – Т. 40. – №. 9. – С. e70-e70.

101. Nakato R., Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation //Briefings in bioinformatics. – 2017. – Т. 18. – №. 2. – С. 279-290.

102. Narlikar L., Jothi R. ChIP-Seq data analysis: identification of Protein–DNA binding sites with SISSRs peak-finder //Next Generation Microarray Bioinformatics: Methods and Protocols. – 2012. – С. 305-322.

103. Nishi K. et al. Structural insights into differences in drug-binding selectivity between two forms of human α1-acid glycoprotein genetic variants, the A and F1* S forms //Journal of Biological Chemistry. – 2011. – Т. 286. – №. 16. – С. 14427-14434.

104. Oki S. et al. ChIP‐Atlas: a data‐mining suite powered by full integration of public ChIP‐seq data //EMBO reports. – 2018. – Т. 19. – №. 12. – С. e46255.

105. Osadchuk L. et al. Study of semen quality, reproductive hormone levels, and lipid levels in men from Arkhangelsk, a city in North of European Russia //American Journal of Men's Health. – 2020. – Т. 14. – №. 4. – С. 1557988320939714.

106. Osmanbeyoglu H. U. et al. Improving ChIP-seq peak-calling for functional co-regulator binding by integrating multiple sources of biological information //BMC genomics. – BioMed Central, 2012. – Т. 13. – С. 1-11.

107. Parekh S. et al. The impact of amplification on differential expression analyses by RNA-seq //Scientific reports. – 2016. – T. 6. – №. 1. – C. 25533.

108. Paulsen M. T. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response //Proceedings of the National Academy of Sciences. – 2013. – T. 110. – №. 6. – C. 2240-2245.

109. Paulsen M. T. et al. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA //Methods. – 2014. – T. 67. – №. 1. – C. 45-54.

110. Pihur V., Datta S., Datta S. RankAggreg, an R package for weighted rank aggregation //BMC bioinformatics. – 2009. – T. 10. – C. 1-10.

111. Piper J. et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data //Nucleic acids research. – 2013. – T. 41. – №. 21. – C. e201-e201.

112. Pique-Regi R. et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data //Genome research. – 2011. – T. 21. – №. 3. – C. 447-455.

113. Pollard K. S. et al. Detection of nonneutral substitution rates on mammalian phylogenies //Genome research. – 2010. – T. 20. – №. 1. – C. 110-121.

114. Qin Q. et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline //BMC bioinformatics. – 2016. – T. 17. – C. 1-13.

115. Quach B., Furey T. S. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter //Bioinformatics. – 2017. – T. 33. – №. 7. – C. 956-963.

116. Rabani M. et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells //Nature biotechnology. – 2011. – T. 29. – №. 5. – C. 436-442.

117. Ramírez-González A. et al. Functional role of AKNA: A scoping review //Biomolecules. – 2021. – T. 11. – №. 11. – C. 1709.

118. Ritchie G. R. S. et al. Functional annotation of noncoding sequence variants //Nature methods. – 2014. – T. 11. – №. 3. – C. 294-296.

119. Robert F., Pelletier J. Exploring the impact of single-nucleotide polymorphisms on translation //Frontiers in genetics. – 2018. – T. 9. – C. 507.

120. Rubinstein R. Y., Kroese D. P. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning. – New York : Springer, 2004. – T. 133.

121. Silva J. V. et al. Profiling signaling proteins in human spermatozoa: biomarker identification for sperm quality evaluation //Fertility and Sterility. – 2015. – T. 104. – №. 4. – C. 845-856. e8.

122. Sherwood R. I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape //Nature biotechnology. – 2014. – T. 32. – №. 2. – C. 171-178.

123. Siepel A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes //Genome research. – 2005. – T. 15. – №. 8. – C. 1034-1050.

124. Song L., Crawford G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells //Cold Spring Harbor Protocols. – 2010. – T. 2010. – №. 2. – C. pdb. prot5384.

125. Spearman C. The proof and measurement of association between two things. – 1961.

126. Spyrou C. et al. BayesPeak: Bayesian analysis of ChIP-seq data //BMC bioinformatics. – 2009. – T. 10. – C. 1-17.

127. Stuart J. M. et al. A gene-coexpression network for global discovery of conserved genetic modules //science. – 2003. – T. 302. – №. 5643. – C. 249-255.

128. Summer S. et al. YBEY is an essential biogenesis factor for mitochondrial ribosomes //Nucleic Acids Research. – 2020. – T. 48. – №. 17. – C. 9762-9786.

129. Sung M. H. et al. DNase footprint signatures are dictated by factor dynamics and DNA sequence //Molecular cell. – 2014. – T. 56. – №. 2. – C. 275-285.

130. Suryatenggara J. Integrated Analysis Pipeline For Unbiased Chip-Seq Analysis. – 2022.

131. Takahashi H. et al. 5′ end–centered expression profiling using cap-analysis gene expression and next-generation sequencing //Nature protocols. – 2012. – T. 7. – №. 3. – C. 542-561.

132. Thomas R. et al. Features that define the best ChIP-seq peak calling algorithms //Briefings in bioinformatics. – 2017. – T. 18. – №. 3. – C. 441-450.

133. Tian S. et al. Identification of factors associated with duplicate rate in ChIP-seq data //PloS one. – 2019. – T. 14. – №. 4. – C. e0214723.

134. Tsagiopoulou M. et al. UMIc: a preprocessing method for UMI deduplication and reads correction //Frontiers in Genetics. – 2021. – T. 12. – C. 660366.

135. Tuğrul M. et al. Dynamics of transcription factor binding site evolution //PLoS genetics. – 2015. – T. 11. – №. 11. – C. e1005639.

136. Van Nostrand E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP) //Nature methods. – 2016. – T. 13. – №. 6. – C. 508-514.

137. Vernet N. et al. Mouse Y-encoded transcription factor Zfy2 is essential for sperm head remodelling and sperm tail development //PLoS One. – 2016. – T. 11. – №. 1. – C. e0145398.

138. Wang J., Batmanov K. BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations //Nucleic acids research. – 2015. – T. 43. – №. 21. – C. E147-e147.

139. Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data //Nucleic acids research. – 2010. – T. 38. – №. 16. – C. e164-e164.

140. Wang M. et al. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants //Nucleic acids research. – 2018. – T. 46. – №.

11. – C. e69-e69.

141. Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics //Nature reviews genetics. – 2009. – T. 10. – №. 1. – C. 57-63.

142. Weirauch M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity //Cell. – 2014. – T. 158. – №. 6. – C. 1431-1443.

143. Wilbanks E. G., Facciotti M. T. Evaluation of algorithm performance in ChIP-seq peak detection //PloS one. – 2010. – T. 5. – №. 7. – C. e11471.

144. Xu J. et al. To mock or not: a comprehensive comparison of mock IP and DNA input for ChIP-seq //Nucleic acids research. – 2021. – T. 49. – №. 3. – C. e17-e17.

145. Yang Q. et al. Sperm telomere length is positively associated with the quality of early embryonic development //Human reproduction. – 2015. – T. 30. – №. 8. – C. 1876-1881.

146. Yang Y. et al. Leveraging biological replicates to improve analysis in ChIP-seq experiments //Computational and structural biotechnology journal. – 2014. – T. 9. – №. 13. – C. e201401002.

147. Yardımcı G. G. et al. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection //Nucleic acids research. – 2014. – T. 42. – №. 19. – C. 11865-11878.

148. Yevshin I. et al. GTRD: a database on gene transcription regulation—2019 update //Nucleic acids research. – 2019. – T. 47. – №. D1. – C. D100-D105.

149. Ylitalo E. B. et al. A novel DNA methylation signature is associated with androgen receptor activity and patient prognosis in bone metastatic prostate cancer //Clinical Epigenetics. – 2021. – T. 13. – №. 1. – C. 133.

150. Zang C. et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data //Bioinformatics. – 2009. – T. 25. – №. 15. – C. 1952-1958.

151. Zelterman D. Robust estimation in truncated discrete distributions with

application to capture-recapture experiments //Journal of statistical planning and inference. – 1988. – Т. 18. – №. 2. – С. 225-237.

152. Zhang X. et al. PICS: probabilistic inference for ChIP-seq //Biometrics. – 2011. – Т. 67. – №. 1. – С. 151-163.

153. Zhang Y. et al. Model-based analysis of ChIP-Seq (MACS) //Genome biology. – 2008. – Т. 9. – С. 1-9.

154. Zhao L. et al. Integrative analysis of reference epigenomes in 20 rice varieties //Nature communications. – 2020. – Т. 11. – №. 1. – С. 2658.

155. Zheng R. et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis //Nucleic acids research. – 2019. – Т. 47. – №. D1. – С. D729-D735.

156. Zhou J., Troyanskaya O. G. Predicting effects of noncoding variants with deep learning–based sequence model //Nature methods. – 2015. – Т. 12. – №. 10. – С. 931-934.

157. Zhu H. Z. et al. Downregulation of orosomucoid 2 acts as a prognostic factor associated with cancer-promoting pathways in liver cancer //World journal of gastroenterology. – 2020. – Т. 26. – №. 8. – С. 804.

158. Н. Грин, У. Стаут, Д. Тейлор Биология. В 3 томах. Т. 1 / Н. Грин, У. Стаут, Д. Тейлор — 3-е изд. — Москва: Мир, 2004 — 514 с.

159. Г. А. Белякова, Е. Л. Богатырёва, Т. А. Вершинина и др., Биология. Современная иллюстрированная энциклопедия / Г. А. Белякова, Е. Л. Богатырёва, Т. А. Вершинина и др., — Москва: Росмэн-Пресс, 2006 — 304 с.