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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

федеральная территория «Сириус»

2024

Работа выполнена в автономной некоммерческой образовательной организации высшего образования «Научно-технологический университет «Сириус», федеральная территория «Сириус».

Научный руководитель:

Колпаков Федор Анатольевич

доктор биологических наук

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

Кондрахин Юрий Васильевич

кандидат биологических наук

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт информационных и вычислительных технологий Сибирского отделения Российской академии наук»

Защита состоится 28 октября 2024 г. в 15.00 на заседании диссертационного совета НТУ.1.5.8.01 на базе АНОО ВО «Университет «Сириус» по адресу 354340, Краснодарский край, федеральная территория «Сириус», Олимпийский пр., д.1.

С диссертацией можно ознакомиться в библиотеке и на сайте АНОО ВО «Университет «Сириус»:
https://siriusuniversity.ru/sveden/science/obyavleniya-o-zashchitakh/8548/

Автореферат разослан “___“ _____ 2024 г.

Ученый секретарь диссертационного совета,
кандидат биологических наук Акбердин И.Р.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

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.

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.

СОДЕРЖАНИЕ РАБОТЫ

Во введении раскрыта актуальность рассматриваемой научной проблемы по теме диссертационной работы, приведена степень разработанности выбранного направления, сформулированы цель и задачи исследования, научные положения, выносимые на защиту, изложена научная новизна, теоретическая и практическая значимость работы, логическая структура диссертации, представлен личный вклад автора.

В первой главе рассмотрен экспериментальный подход к идентификации РСТФ на основе иммунопреципитации хроматина с последующим высокопроизводительным секвенированием ДНК (ChIP-seq). На примере алгоритма MACS2 рассмотрен процесс идентификации РСТФ в данных ChIP-seq. Отдельное внимание уделено методам оценки качества исходных данных, а также рассмотрены подходы к мета-анализу имеющихся ChIP-seq данных. В частности, в обзоре литературе рассмотрены различные группы методов коллективного выбора (Rank Aggregation). Описаны подходы выявления однонуклеотидных геномных вариаций, имеющих функциональное значение в контексте регуляции транскрипции. Дано

краткое описание структуры сперматозоидов человека, а также наиболее встречающихся аномалий морфологии сперматозоидов.

Вторая глава посвящена материалам и методам, используемым в рамках данной диссертационный работы.

На первом этапе исследования производился сбор информации 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), UBERON (Mungall et al., 2012), Cellosaurus (Bairoch, 2018), BRENDA (Gremse et al., 2010), Experimental factor ontology (Malone et al., 2010), Plant ontology (Cooper et al., 2013) (Kulyashov M. et al, 2020).

В рамках диссертационной работы был расширен функционал программы GEOminer для поддержки аннотации DNase-seq и ATAC-seq данных, а также интегрированы упомянутые выше 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., 2021) и системы управления распределенными вычислениями e-grid - собственная распределенная вычислительная платформа для параллельной обработки данных на нескольких вычислительных узлах (Kolmykov et al., 2021).

В рамках диссертационного исследования основной фокус направлен на ChIP-seq и DNase-seq данные для человека. В актуальной версии базы данных GTRD (GTRD v 21.12) (Kolmykov et al., 2021) содержится 35719 ChIP-seq экспериментов для *Homo sapiens*, охватывающих различные клеточные типы и ткани, а также широкий спектр экспериментальных условий, для 1391 ТФ и кофактора. Также в работе использовались 1701 DNase-seq эксперимент.

Для анализа качества NGS данных использовался стандартный подход. Помимо расчета базовых статистик выравнивания (samtools flagstat) производится оценка сложности библиотеки (NRF, PBC1 и PBC2) и оценка отношения сигнал-шум при помощи метода кросс-корреляции (NSC и RSC) (Landt et al., 2012). По завершении стадии идентификации РСТФ производится вычисление доли прочтений, попавших в границы полученных РСТФ (FRiP).

Анализ достоверности районов связывания транскрипционных факторов был проведен на основании дополнительных данных:

- Поиск мотивов связывания ТФ (МСТФ) из БД HOCOMOCO в РСТФ
- Оценки консервативности РСТФД на основании данных из БД UCSC (phastCons и phyloP);
- Сопоставления с районами открытого хроматина (DOX);
- Оценки воспроизводимости РСТФ в других ChIP-seq экспериментах.

Для поиска однонуклеотидных геномных вариантов (SNV) в данной работе анализировались данные 367 образцов полноэкзомного секвенирования, полученные в секторе репродуктивных технологий человека ИЦИГ СО РАН под руководством Осадчук Л. В. Полноэкзомное секвенирование проводилось на базе Сектора геномных исследований ИЦиГ СО РАН. Идентификация и фильтрация SNV осуществлялась в соответствии с рекомендациями GATK Best Practices при помощи HaplotypeCaller и GenotypeGVCFs из Genome Analysis Toolkit (GATK) v4.1.4.1 (Poplin et al., 2017).

Полученные наборы геномных вариаций были проаннотированы с помощью Annovar (dbSNP146) (Wang et al., 2010). Для определения приоритетности вариаций, ассоциированных с нарушенным сперматогенезом, их эффекты 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).

Для интерпретации SNV, ассоциированных с низким качеством семенной жидкости, в контексте регуляции транскрипции были использованы базы данных: GTRD, ADASTRA и GTEx.

Третья глава посвящена описанию и обсуждению основных результатов диссертационной работы. Данная глава состоит из 5 разделов.

Первый раздел посвящен исследованию воспроизводимости РСТФ в рамках одного ChIP-seq эксперимента каждым из 4 алгоритмов идентификации РСТФ: GEM, MACS2, PICS и SISSRs. На основании сопоставления результатов работы упомянутых выше алгоритмов полученные РСТФ разделялись на 4 группы: F1, F2, F3 и F4 (см. рисунок 3.1.1). Было показано, что доля группы F4 в среднем составляет менее 10% РСТФ в ChIP-seq экспериментах. Поскольку успешность идентификации РСТФ напрямую связана с качеством ChIP-seq эксперимента, была исследована взаимосвязь воспроизводимости РСТФ с качеством ChIP-seq данных, полученных на основе рекомендаций проекта ENCODE.

В исследовании была показана статистически значимая зависимость между качеством ChIP-seq данных и воспроизводимостью РСТФ разными алгоритмами. В частности, в экспериментах с высоким качеством существенно снижается (на ~20 процентных пунктов) доля F1 РСТФ. Однако даже для высококачественных ChIP-seq данных средняя доля F4 РСТФ составляла всего ~15%.

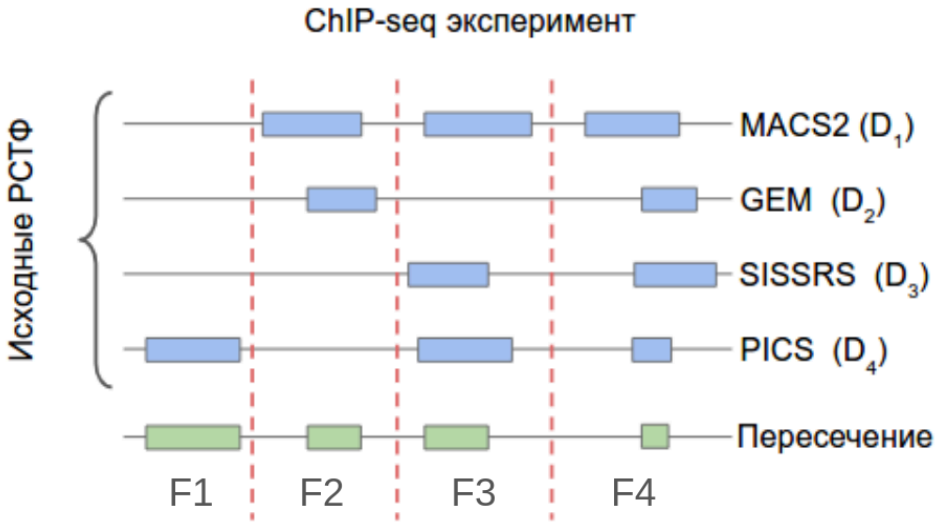

Рисунок 3.1.1 Схема пересечения результатов работы алгоритмов идентификации пиков и получения обобщённого набора РСТФ с разбиением РСТФ на подгруппы по уровню воспроизводимости среди использованных алгоритмов. D_i - множество РСТФ выявленных заданным методом в эксперименте, f_i - число получившихся районов связывания, которые были составлены ровно из i пиков.

Далее была исследована взаимосвязь воспроизводимости РСТФ внутри эксперимента с различными геномными аннотациями: районами открытого хроматина, генетической консервативностью района, содержанием мотивов связывания ТФ (МСТФ); Также была оценена воспроизводимость РСТФ относительно других ChIP-seq экспериментов для выбранного ТФ, доступных в БД GTRD.

Было показано, что чем большим количеством методов (рис. 3.1.2) идентифицируется РСТФ, тем:

1) чаще он встречается в районах открытого хроматина (рис. 3.1.2.А);
2) чаще воспроизводится в других экспериментах для заданного ТФ (рис. 3.1.2.Б);
3) является более консервативным (рис. 3.1.2.В);
4) в большем количестве содержит мотивы, представленные позиционной весовой матрицей, связывания соответствующего ТФ (рис. 3.1.2.Г).

Во втором разделе описывается разработанный подход оценки доли ложно-выявленных РСТФ, полученных на основании обработки ChIP-seq экспериментов при помощи 4 методов идентификации РСТФ.

Оценка доли ложноположительных (FP) РСТФ при помощи FPCM (False Positive Control Metric) основывается на предположении о том, что их большая часть находится в группе F1, т. е. подтверждается только одним методом. Таким образом оценка FPCM может быть представлена в виде:

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

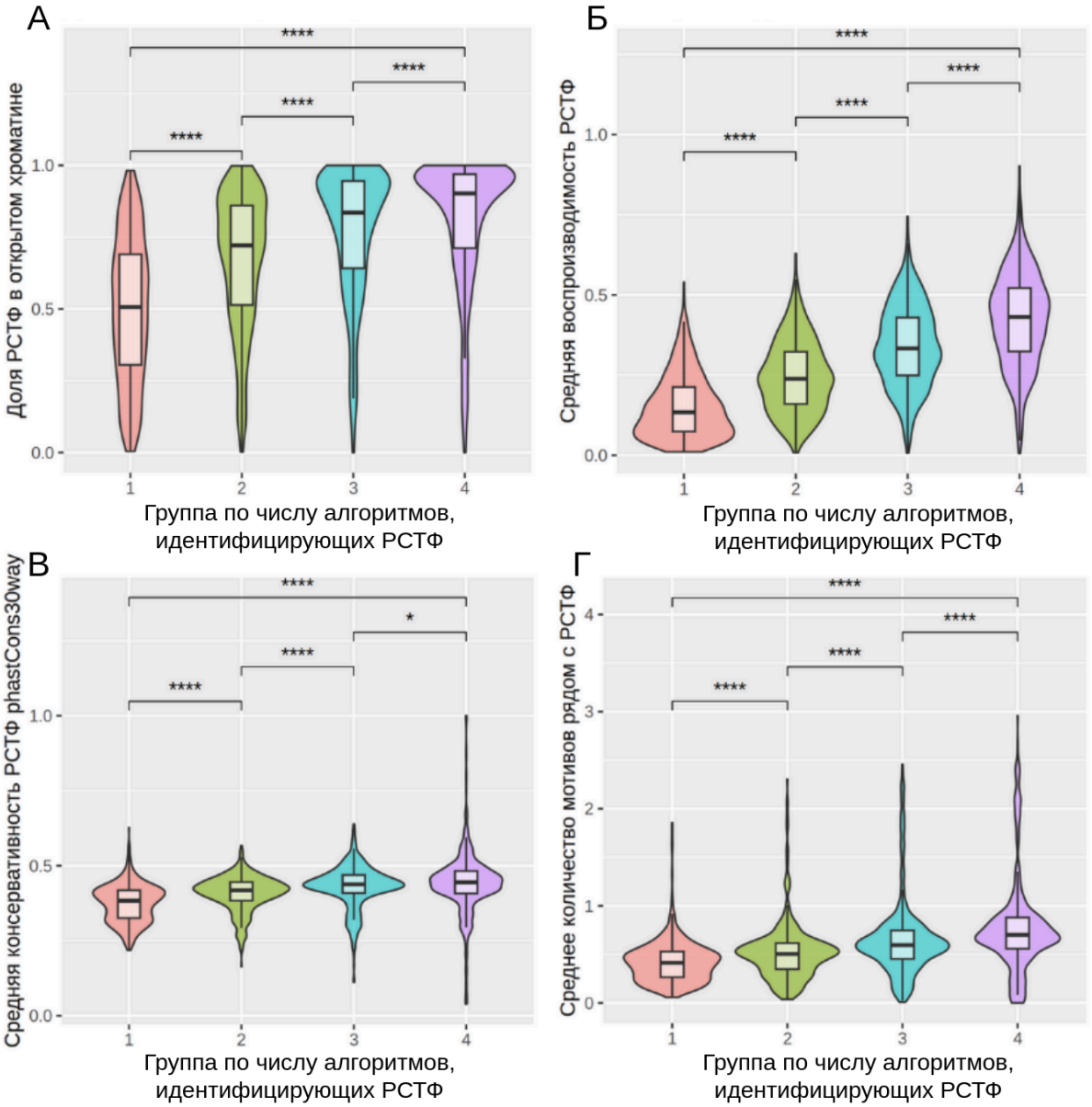

Рисунок 3.1.2 Взаимосвязь различных геномных аннотаций с принадлежностью к одной из 4 групп РСТФ, обусловленных степенью пресечения результатов работы 4 алгоритмов идентификации РСТФ (А) Распределения доли РСТФ в открытом хроматине; (Б) Распределения воспроизводимости пиков в других экспериментах для заданного ТФ; (В) Распределения значений эволюционной консервативности пиков методом PhastCons30Way; (Г) Распределения среднего числа мотивов связывания ТФ (p-value < 10^{-4}) в окрестности РСТФ. Для определения уровня достоверности различий между группами использовался непараметрический критерий Уилкоксона (* - pvalue < 10^{-2}; **** - pvalue < 10^{-5}).

Предполагается, что неизвестное число подлинных РСТФ является случайной величиной с распределением Пуассона. То есть 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 районов связывания. Тогда:

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

Разработанный алгоритм был реализован на языке программирования Java в виде программного модуля для платформы BioUML (Kolpakov et al., 2019).

Была исследована возможность использования значений FPCM для принятия решения по удалению F1 РСТФ из дальнейшего анализа. Для этого 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. Например, на рисунке 3.2.1. демонстрируется, что на основании метрики FPCM даже среди качественных ChIP-seq экспериментов можно выделить эксперименты, для которых характерна в среднем более низкая воспроизводимость F1 РСТФ.

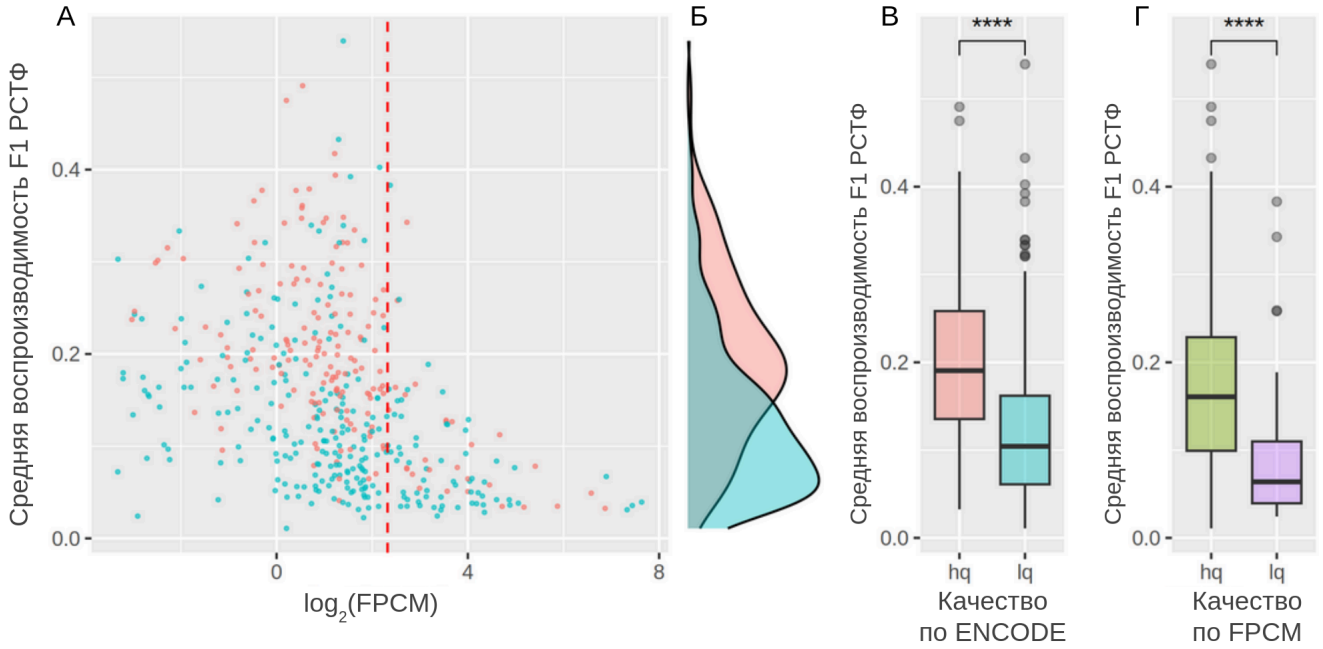

Рисунок 3.2.1. (А) - Взаимоотношение значений log2(FPCM) и воспроизводимости РСТФ среди всех экспериментов (количество экспериментов с РСТФ / количество экспериментов) в подгруппе F1. Синим цветом обозначены эксперименты, которые по критериям ENCODE относятся к данным с низким качеством (lq); Красным цветом - качественные ChIP-seq эксперименты (hq); Красный пунктир - условный порог FPCM = 5. **(Б)** - Плотности распределений значений AUC для F1 и F4 в хороших и плохих данных (красный и синий цвета соответственно). **(В)** - бокс-плот, описывающий распределения с графика **Б**. **(Г)** - бокс-плот, описывающий распределение значений AUC в F1 в экспериментах с FPCM < 5 (зелёный цвет; умеренное количество FP; hq) и для экспериментов с FPCM > 5 (фиолетовый цвет; высокое содержание FP; lq). Для определения уровня достоверности различий между группами на рисунках: В и Г, использовался непараметрический критерий Уилкоксона (**** - p-value < 10^{-5}).

Далее был проведен анализ взаимосвязи между значениями FPCM и изменением эффективности идентификации МСТФ в полном наборе РСТФ в ответ на удаление F1 РСТФ (см. Рисунок 3.2.2). На первом этапе был проведён анализ эффективности идентификации МСТФ на основании использования 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 3.2.2. демонстрируется резкое ухудшение эффективности предсказания МСТФ по PWM в F1 РСТФ при достижении определенных значений FPCM.

В рамках данного раздела было показано, что FPCM позволяет идентифицировать поднаборы экспериментах, которые демонстрируют:

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

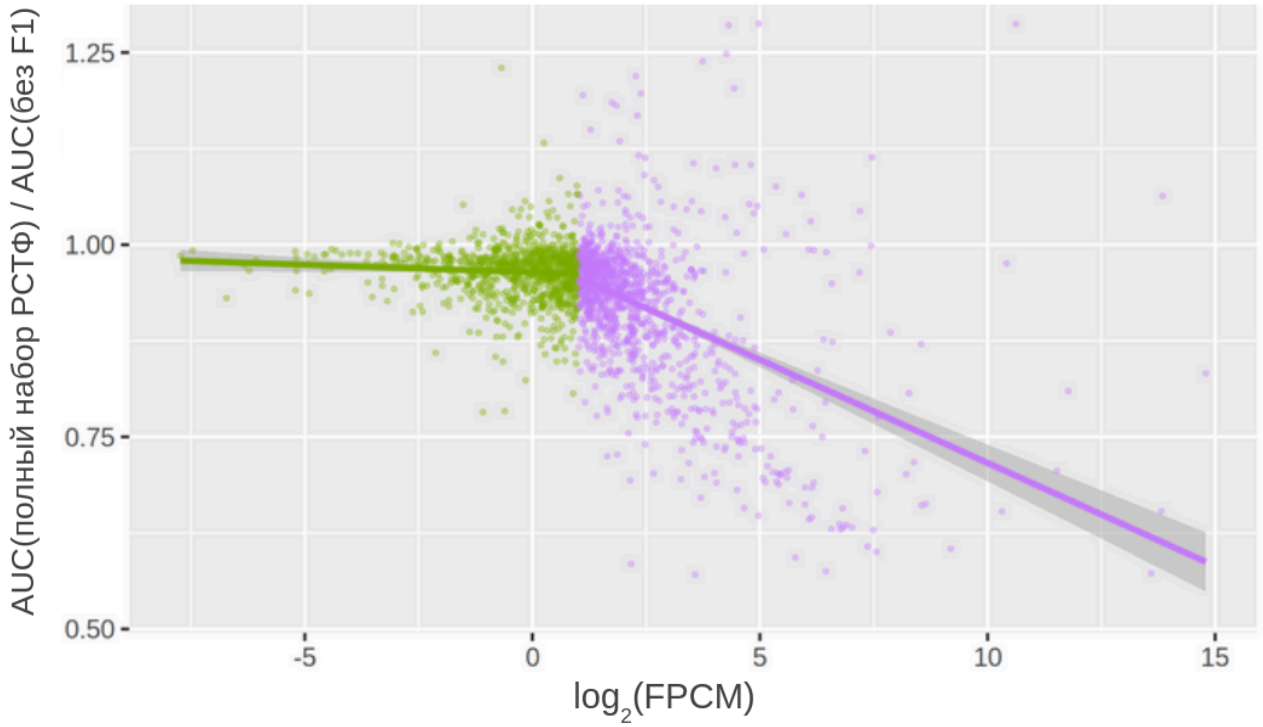

Рисунок 3.2.2. Взаимоотношение оценки FPCM и изменения значений AUC для предсказания мотивов связывания ТФ до и после удаления F1 подгруппы из РСТФ, идентифицированных только одним методом идентификации пиков (MACS2). На рисунке отображены только ChIP-seq эксперименты, которые прошли фильтрацию оценками качества, использующимися в проекте ENCODE. Зелёным и фиолетовым цветом обозначены эксперименты, располагающиеся ниже и выше относительно выбранного условного порогового значения FPCM=5.

В третьем разделе описывается разработанный подход оценки доли ложно неидентифицированных РСТФ, полученных на основании обработки ChIP-seq экспериментов при помощи 4 методов идентификации пиков.

Метод FNCM (False Negative Control Metric) - определяется как отношение количества РСТФ, выявленных выбранным методом, и ожидаемым количеством 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 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, 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

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

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

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 методы: 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 (Kolpakov et al., 2019) на языке Java был реализован алгоритм расчета значений FNCM, а также алгоритм оценки истинного размера набора РСТФ на основании пересечения предоставленных пользователем наборов РСТФ.

Четвертый раздел посвящен разработанному подходу последовательного применения методов коллективного выбора для мета-анализа ChIP-seq данных.

Для выявления наиболее воспроизводимых РСТФ на основе мета-анализа результатов всех ChIP-seq 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 методов коллективного выбора (МКВ):

1-й этап – метод Борда применяется для ранжирования РСТФ, выявленных каждым методом идентификации РСТФ (MACS, SISSRs, GEM и PICS), на основании различных характеристик качества, присваиваемых соответствующим методом (например, p-value, “fold enrichment” и др.);

2-й этап – полученные и упорядоченные на основании полученных весов списки РСТФ также подаются на вход методу Борда;

3-й этап – полученный список РСТФ группируется с другими подобными списками, полученными из других экспериментов для рассматриваемого ТФ, и затем обрабатываются используемым методом коллективного выбора.

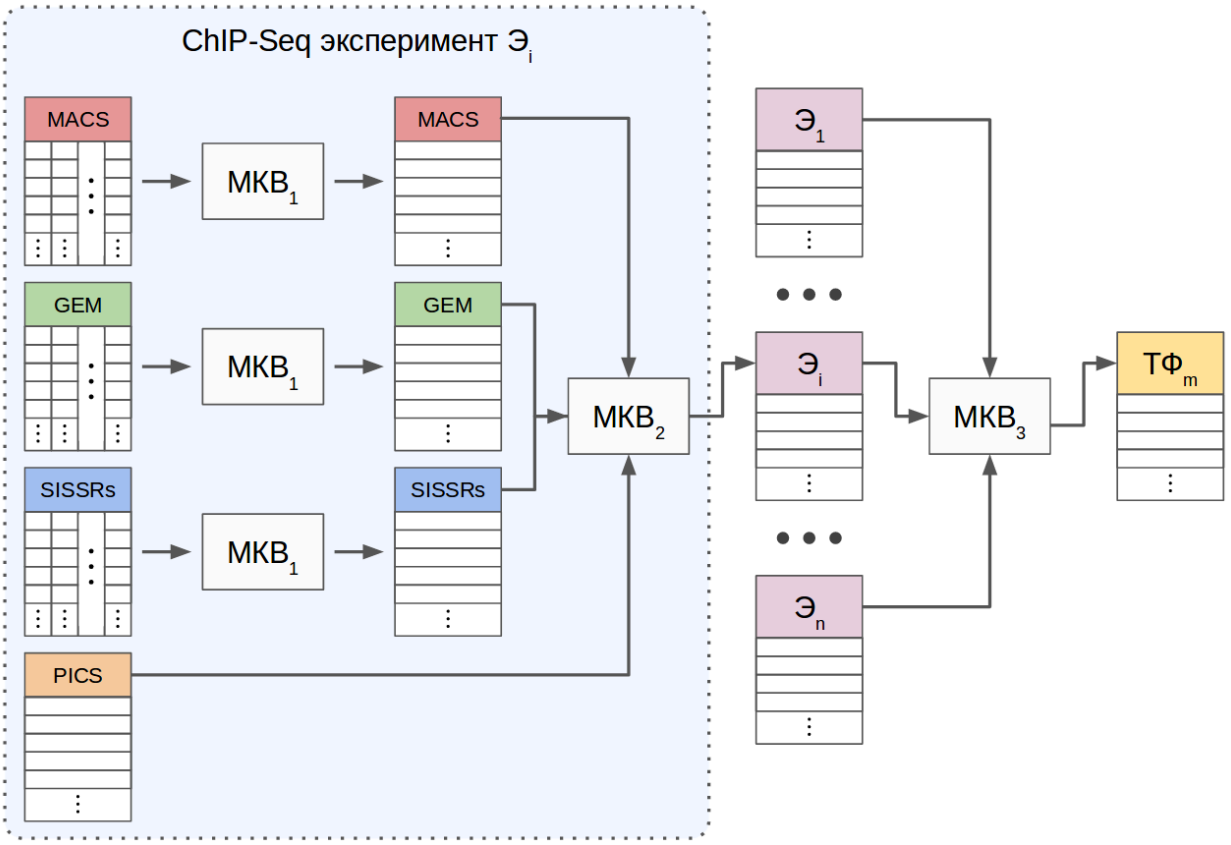

Рисунок 3.4.1. Схема многостадийного применения методов коллективного выбора, METARA. МКВ$_i$ - i-ая стадия применения агрегирующей функции (метода коллективного выбора).

В результате описанного выше многостадийного алгоритма METARA для каждого ТФ был построен набор мета-кластеров, ранжированный на основании значений финальной агрегирующей функции (ФАФ). При помощи предложенного алгоритма были построены карты геномных районов связывания 1391 ТФ и

кофактора 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 3.4.2 приведен пример такого анализа для ТФ NRF1. Для проанализированных 119 ТФ показано, чем выше ФАФ (меньше номер группы мета-кластеров), тем чаще мета-кластеры:

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

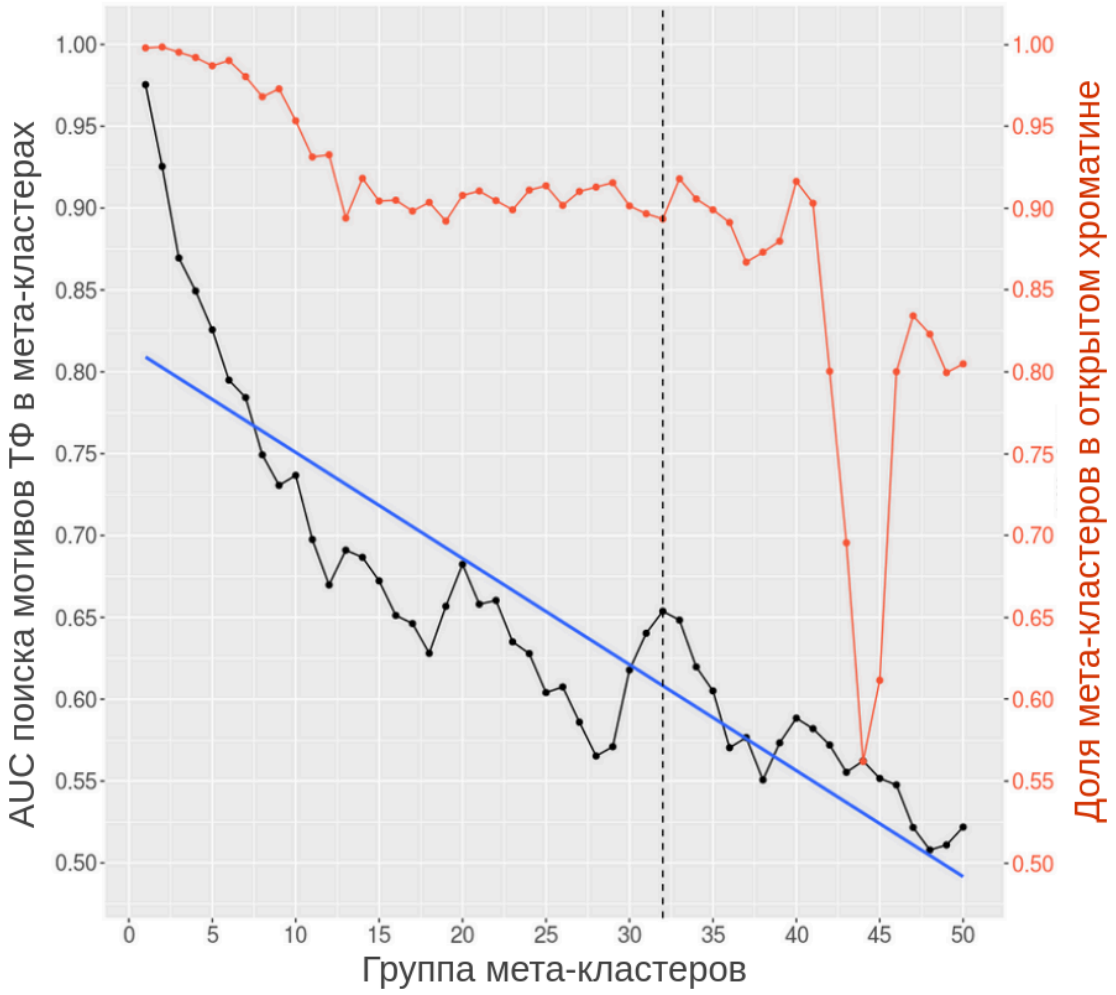

Рисунок 3.4.2. Взаимоотношение значений ФАФ и значений AUC для поиска мотивов связывания ТФ (чёрная линия), а также взаимоотношение значений ФАФ и доли мета-кластеров в районах открытого хроматина (красная линия).

Пятый раздел посвящен интерпретации с точки зрения регуляции транскрипции однонуклеотидных геномных вариантов, ассоциированных с нарушениями морфологии сперматозоидов.

На основании полноэкзомного анализа ассоциации было идентифицировано 135 SNV достоверно (FDR<0.05) ассоциированных с морфологическими нарушениями сперматозоидов. Найденные геномные варианты располагаются в границах 63 генов. Два однонуклеотидных варианта являются синонимичными заменами в генах, кодирующих транскрипционные факторы: AKNA и ZNF704. ТФ AKNA участвует в регуляции организации микротрубочек.

Также 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.

Таблица 3.5.1. SNV, влияющие на эффективность связывания ТФ с ДНК.

dbSNP ID	Ген: Эффект	Связанный с SNV признак	ASB	eQTL
rs138595914	*PCNT* : 5 Prime UTR *C21orf58* : 2KB Upstream	Аномалии средней части	⬇ AR	⬆ *SPATC1L* (в семенниках) ⬇ :*C21orf58*, *DIP2A*, *FTCD*, *MCM3AP*, *PCNT*, *PRMT2*, *SPATC1L*, *YBEY*
rs2304961	*ABR* : Intron Variant (16 интрон)	Аномалии средней части	⬇ CTCF	—
rs2270420	*POMT2* : Синонимичная замена (1 экзон) *GSTZ1* : 2KB Upstream	Двойная головка	⬇ SRBP2	⬇ *GSTZ1*
rs71486131	*MARVELD1* : Синонимичная замена (1 экзон)	Маленькая головка	⬆ CTCF	⬆ *MARVELD1* ⬆ *ZFYVE27*

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 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. **Semyon Kolmykov**, Yuriy Kondrakhin, Ivan Yevshin, Ruslan Sharipov, Mikhail Kulyashov, Fedor Kolpakov, 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.

Подписано в печать 25.09.2024
Формат 60 х 90 1/16. Печ. л. 1,5
Тираж 100 экз. Заказ №1

Отпечатано в типографии “Cheese Photo”
354340, Краснодарский край, федеральная территория «Сириус», Хуторская, 10