ОТЗЫВ

члена диссертационного совета НТУ.1.5.8.01
Попова Даниила Викторовича
на диссертацию **Колмыкова Семёна Константиновича**
«Разработка методов контроля качества и построения карты геномных районов связывания транскрипционных факторов на основе сравнительного анализа ChIP-seq экспериментов»,
представленной на соискание ученой степени кандидата биологических наук по специальности 1.5.8. Математическая биология, биоинформатика

Актуальность темы. Регуляция экспрессии генов является фундаментальным биологическим процессом, поэтому исследование молекулярных механизмов, регулирующих этот процесс – это одна из ключевых задач в молекулярной биологии. Несмотря на многолетнюю историю изучения этой проблемы, конкретные механизмы регуляции экспрессии генов изучены явно недостаточно. Частично, это связано с высокой сложностью исследуемой системы, а именно с тем, что *i*) экспрессия любого гена регулируется многими транскрипционными факторами, связывающимися с относительно небольшим промоторным регионом, *ii*) большим количеством транскрипционных факторов (более 1 500) и *iii*) их генов-мишеней (около 20 000). Идентификация районов связывания транскрипционных факторов необходима для понимания механизмов регуляции экспрессии генов, в том числе для определения конкретных мотивов связывания транскрипционных факторов. С другой стороны, важно отметить, что современные полногеномные методы оценки районов связывания транскрипционных факторов имеют высокий уровень шума. Поэтому для получения надежных данных требуется усреднение результатов разные экспериментов, зачастую имеющих разное качество, а именно их обработка по единому алгоритму анализа с контролем и удалением ненадежных данных. Решению этой проблемы посвящена диссертационная работа С.К. Колмыкова. Актуальность темы работы не взывает сомнения.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации. В работе сформулированы три положения, выносимых на защиту.

На первом этапе работы автор выбрал алгоритм единообразной аннотации данных всех доступных ChIP-seq и DNase-seq экспериментов, выполненных на человеческих тканях/клетках, и внес их результаты в базу данных GTRD (около 1 500 для каждого из таких экспериментов). Это позволило значительно расширить базу данных и провести единообразный анализ данных более чем 11 000 ChIP-seq экспериментов. Затем было проведено сопоставление результатов полногеномного поиска регионов связывания транскрипционных факторов (РСТФ) с помощью четырех наиболее популярных алгоритмов идентификации пиков и выявлены РСТФ, идентифицированные по результатам одного или нескольких алгоритмов. Это позволило автору охарактеризовать такие РСТФ, а именно изучить их пересечение с регионами открытого хроматина, воспроизводимость и среднюю консервативность, а также наличие в них известных мотивов связывания с транскрипционными факторами. Ожидаемо более надежные характеристики были найдены для РСТФ, идентифицированных с помощью нескольких алгоритмов. Для фильтрации данных низкого качества автором был предложен анализ доли ложноположительных РСТФ (FPCM). Затем, был предложен и реализован анализ доли ложно-невыявленных РСТФ (FNCM); это позволило обнаружить, что MACS2 – алгоритм идентификации РСТФ для данных ChIP-seq, превосходит остальные алгоритмы и в отличии от них слабо зависит от доступности инпут-контроля. Совместное использование оценок FPCM и FNCM позволило автору предложить комплексный подход для оценки качества данных ChIP-seq для выявления наиболее достоверных наборов РСТФ.

На следующем этапе работы для выявления наиболее значимых РСТФ для заданного транскрипционного фактора автор предложил и использовал мета-анализ данных всех ChIP-seq экспериментов из базы данных GTRD

(METARA). Для этой задачи были использованы подходы с агрегацией рангов и метод коллективного выбора. Способность предложенного мета-анализа ранжировать наборы РСТФ по степени их правдоподобности была подтверждена с помощью исследования распределения пиков внутри областей с открытым хроматином и доли пиков, находящихся рядом с известными мотивами транскрипционных факторов. Помимо этого, было показано, что в ряде случаев наиболее правдоподобные РСТФ имеют низкие оценки по этим метрикам, что объясняются специфическими функциями транскрипционных факторов, связывающихся в этих регионах (ремоделирование хроматина). Используя алгоритм METARA, были построены карты геномных районов связывания 1391 человеческого транскрипционного фактора и корегулятора.

На финальном этапе работы было исследована роль различных однонуклеотидных геномных вариаций в нарушении сперматогенеза. Для этого были использованы собственные результаты морфологического анализа сперматозоидов людей с разной выраженностью нарушений и данные полноэкзомного секвенирования (выявившие 135 геномных вариаций), а также данные по геномным вариациям и экспрессии генов из базы данных GTX и Human Protein Atlas. Из рассматриваемых геномных вариаций были отобраны только четыре, которые входят в наиболее правдоподобные РСТФ, оцененные по алгоритму METARA, и имеют аллельный дисбаланс (база данных ADASTRA).

Полученные результаты корректно обосновывают все положения, выносимые на защиту; выводы, представленные в работе, корректно обобщают полученные результаты.

Научная новизна работы определяется тем, что автором был предложен и обоснован новый алгоритм оценки качества всех доступных человеческих ChIP-seq экспериментов, основанный на оценке согласованности результатов четырех наиболее популярных алгоритмов идентификации районов связывания транскрипционных факторов. С

помощью методов коллективного выбора был предложен и реализован новый алгоритм для отбора наиболее правдоподобного РСТФ. Это позволило автору построить полную карту геномных РСТФ с учетом известных мотивов связывания транскрипционных факторов и районов открытого хроматина. Помимо этого, были выявлены ассоциации нарушений морфологии сперматозоидов человека с однонуклеотидными геномными вариантами, характерными для российской популяции; для четырех из которых был выявлен аллельный дисбаланс.

Теоретическая и практическая значимость работы. Теоретическая значимость работы связана с тем, что предложены новый алгоритм анализа и контроля данных ChIP-seq экспериментов. Практическая значимость заключается в том, что все предложенные и успешно апробированные алгоритмы анализа были реализованы в виде программ, интегрированных в платформу BioUML и систему управления распределенными вычислениями e-grid, созданных с непосредственным участием автора диссертационной работы. Основной результат диссертационной работы – полная карта геномных РСТФ с учетом известных мотивов связывания транскрипционных факторов и районов открытого хроматина – представлен в открытой базе данных GTRD, которая является наиболее полной отечественной и одной из наиболее полных и популярных в мире базой данных по регуляторным элементам генома. Высокая практическая значимость полученных результатов, подтверждается тем, что результаты диссертационной работы, представленные в базе данных GTRD, были успешно использованы для создания различных российских и международных инструментов/баз данных для исследования регуляции генома (HOCOMOCO, ADASTRA, ANANASTRA, BaMM motif и miSigDB).

Степень достоверности результатов проведенных исследований. Результаты диссертационной работы были широко представлены и успешно

апробированы на ряде ведущих отечественных и зарубежных конференций, а также в девяти публикациях в ведущих профильных международных журналах. Достоверность полученных данных также подтверждается высокой цитируемостью ключевых публикаций, подготовленных по результатам диссертационной работы.

Публикации основных результатов диссертационной работы. По теме диссертации опубликовано 9 статей в ведущих международных рецензируемых журналах, индексируемых в базах данных Scopus и WoS, среди которых – 8 входят в категорию Q1. В трех из опубликованных статей С.К. Колмыков является первым автором, что подчеркивает его ключевой вклад в получение, обработку и представление данных, полученных в рамках диссертационной работы.

Структура диссертационной работы. Диссертационная работа построена по классическому типу и изложена на 141-й странице. Во Введении четко сформулированы цель и задачи исследования. Обзор литературы написан ясно и дает представление о современных алгоритмах идентификации РСТФ и контроля качества данных ChIP-seq экспериментов, а также о методах коллективного выбора. Отдельный раздел посвящен описанию нарушений в морфологии сперматозоидов.

В разделе Материалы и методы подробно описаны подходы и алгоритмы анализа данных, но недостаточно подробно описаны экспериментальные методы, результаты которых использовались для идентификации и анализа однонуклеотидных геномных вариантов (см.ниже).

В разделе Результаты и обсуждение подробно описаны и иллюстрированы полученные данные. В разделе Заключение проведено обобщение основных результатов работы. Ключевые результаты корректно представлены в Выводах, которые соответствуют задачам, поставленным в диссертационной работе.

Содержание автореферата соответствует содержанию, основным положения и результатам диссертации.

Вопросы по диссертационной работе / Замечания

1. Работа имеет ряд опечаток (особенно в разделе 3.5, посвященном идентификации и анализу однонуклеотидных геномных вариантов) и неточностей в обозначении номеров рисунков. Часть используемых в тексте аббревиатур не представлена в списке сокращений, что несколько затрудняет восприятие результатов работы.

2. При описании степени разработанности темы отсутствует какая-либо информация о литературных данных по ассоциации однонуклеотидных геномных вариантов с нарушениями сперматогенеза. Более того, при обсуждении собственных данных нет описания того, насколько полученные результаты соответствуют литературе и что они добавляют в этой области.

3. При описании методов, автор указал, что «процесс отбора участников (эксперимента) подробно описан в работе Osadchuk и соавторов». Отсутствие в диссертации этой важной информации затрудняет интерпретацию данных по ассоциациям с нарушением структуры сперматозоидов, представленных в работе. Помимо этого, хотелось бы видеть хотя бы краткое описание методов подготовки проб и протокола анализа для полноэкзомного секвенирования, что также важно для корректной интерпретации полученных данных.

4. Хотелось бы видеть данные по средним размерам (и разбросам) для идентифицированных РСТФ и распределению плотности РСТФ относительно стартов транскрипции: имеется ли смещение распределения плотностей относительно стартов транскрипции?

5. Медианная воспроизводимость РСТФ даже в группе F4 не превышает 0,5 (Рисунок 3.1.2 Б). Чем объясняется такое относительно невысокое значение?

6. На рисунке 3.4.3 представлены примеры разных динамик значений ФАФ и AUC для двух разных транскрипционных факторов (разная направленность изменений, наличие осцилляций в доле мета-кластеров в открытом хроматине для фактора JUN). Хотелось бы получить комментарии о причинах этих различий и особенностях динамики.

7. Ложноположительные результаты ChIP-seq экспериментов, связанные с наличием белок-белковых взаимодействий – это одна из ключевых проблем этого метода. Автор обсуждает эту проблему при описании высокой вариабельности в доле МСТФ и РСТФ в зависимости от транскрипционного фактора. Возможно ли использование каких-либо специальных подходов (при проведении ChIP-seq экспериментов или обработке их результатов) для снижения влияния белок-белок взаимодействий на результаты идентификации РСТФ. Планируется ли как-то отражать в базе данных GTRD информацию о результатах с высоким потенциальным влиянием белок-белковых взаимодействий?

8. На рисунке 3.5.3 представлена частота различных аллелей в пробах со сперматозоидами с морфологическими нарушениями. В отличие от предыдущих рисунков представлено только два генотипа. С чем связано отсутствие данных по одному из гомозиготных вариантов?

Заданные вопросы носят дискуссионный характер; отмеченные недостатки не снижают высокого качества исследования и не влияют на главные теоретические и практические результаты диссертации, описанные выше. Результаты оригинальны, обладают научной новизной и практически значимы.

Заключение. Диссертационная работа Колмыкова Семёна Константиновича является законченной научно-квалификационной работой, выполненной автором на высоком научном уровне. Диссертация соответствует пп. 2, 5, 11 и 12 паспорта научной специальности 1.5.8. Математическая биология, биоинформатика.

Диссертационная работа Колмыкова Семёна Константиновича «Разработка методов контроля качества и построения карты геномных районов связывания транскрипционных факторов на основе сравнительного анализа ChIP-seq экспериментов» отвечает требованиям пп.2.1–2.6 Положения о присуждении ученых степеней Автономной некоммерческой образовательной организацией высшего образования «Научно-технологический университет «Сириус» утвержденного приказом от 25 декабря 2023 г. № 350/1-ОД-У, предъявляемым к диссертациям на соискание ученой степени кандидата наук, а ее автор, Колмыков С.К., заслуживает присуждения ученой степени кандидата биологических наук по специальности 1.5.8. Математическая биология, биоинформатика.

Член диссертационного совета
НТУ.1.5.8.01
Ведущий научный сотрудник-
заведующий
лабораторией физиологии
мышечной деятельности
Федерального государственного
бюджетного учреждения науки
Государственного научного центра
Российской Федерации
Института медико-биологических
проблем
Российской академии наук,
профессор РАН, доктор
биологических наук
по специальности 03.03.01 –
«Физиология»»

Попов
Даниил Викторович

Сведения:

Адрес организации:
Адрес: 123007, Российская Федерация, г. Москва, Хорошевское шоссе, д. 76А
Федеральное государственное бюджетное учреждение науки
Государственный научный центр Российской Федерации
Институт медико-биологических проблем
Российской академии наук
Контактный телефон: +7 499 195 6566
e-mail: danil-popov@imbp.ru

Подпись д.б.н. Попова Д.В. удостоверяю:

Ученый секретарь
Федерального государственного бюджетного учреждения науки
Государственного научного центра Российской Федерации
Института медико-биологических проблем
Российской академии наук, доктор биологических наук

Левинских М.А.
16 октября 2024 г.