Федеральное государственное бюджетное научное учреждение
«Федеральный исследовательский центр
Институт цитологии и генетики
Сибирского отделения
Российской академии наук»
630090, г.Новосибирск, пр. Академика Лаврентьева, д.10

ОТЗЫВ

члена диссертационного совета НТУ.1.5.8.01
Афонникова Дмитрия Аркадьевича
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.

В процессе развития организма гены экспрессируются специфическим образом, формируя различные типы клеток. Это требует сложной регуляции экспрессии генов, которая в значительной степени происходит во время транскрипции генов. Она осуществляется в результате серии биофизических событий, управляемых огромным количеством молекул, образующих более крупные сети и проходящих через множество временных и функциональных этапов, которые варьируются от специфических взаимодействий ДНК-белок до набора и сборки нуклеопротеиновых комплексов. Одну из ключевых ролей в этих процессах играют транскрипционные факторы (ТФ) – белки, которые связываются с ДНК в промоторе гена и контролируют скорость транскрипции. Информация о регуляторных участках ДНК, районах связывания этих транскрипционных факторов (РСТФ) позволяет оценивать регуляторный потенциал генов, определять какими ТФ происходит регуляция их экспрессии.

Современные экспериментальные технологии на основе технологий иммунопреципитации хроматина и высокопроизводительного секвенирования (ChIP-seq) позволяют идентифицировать РСТФ в масштабе всего генома, в определенных тканях, типах клеток и при разном на них воздействии внешних условий. Эти методы продемонстрировали свою высокую эффективность, однако обладают и рядом недостатков, таких как высокий уровень шума, неоднозначность результатов при обработке данных различными алгоритмами.

В этой связи актуальной задачей является разработка методов оценки доли ошибок в определении РСТФ для заданного ChIP-seq эксперимента на основании сравнения результатов нескольких алгоритмов идентификации РСТФ, определения наиболее достоверных из них на основе мета-анализа большого количества экспериментов, а также изучение влияния на функции РСТФ однонуклеотидных вариаций.

Работа Колмыкова С.К. была посвящена решению этих актуальных задач.

Степень обоснованности научных положений, выводов и рекомендаций, сформулированных в диссертации.

Научные положения, выносимые на защиту, являются обоснованными. Для их получения обработан большой объем данных по экспериментам ChiP-seq, представленном в базе данных (БД) GTRD. Они базируются на тщательной проверке статистических гипотез, лежащих в их основе. Программные решения базируются на использовании платформы BioUML, которая зарекомендовала себя как надежная и широко используемая система.

Научная новизна работы определяется тем, что:

Впервые предложены и реализованы новые методы оценки качества ChIP-seq экспериментов (FPCM и FNCM), на большом количестве систематизированных данных из БД GTRD исследована связь между функциональными характеристиками РСТФ, взятых из аннотации генома и надежностью предсказания их локализации на основе экспериментов ChIP-seq. Этот анализ проведен в рамках одной методологии для нескольких метрик надёжности идентификации РСТФ компьютерными методами. Показано, что функциональная нагрузка РСТФ положительно связана как с согласованностью предсказания разными методами, так и с оценкой FPCM.

Впервые разработан метод мета-анализа данных ChIP-seq METARA и с его помощью 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 на основе разработанных методов проведено исследование связи между однонуклеотидными вариациями и нарушением морфологии сперматозоидов человека. 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 и практическая значимость работы.

В работе предложен комплекс методов, позволяющих оценивать качество ChIP-seq экспериментов на основе сравнения результатов разных алгоритмов для выявления РСТФ, что позволило общее оценить общее количество таких районов и долю ложно идентифицированных РСТФ.

Разработан новый алгоритм применения методов коллективного выбора, METARA, для последующего отбора наиболее воспроизводимых районов связывания транскрипционных факторов.

На основе комплекса разработанных методов впервые идентифицированы однонуклеотидные вариации, ассоциированные с различными нарушениями морфологии сперматозоидов, характерные для популяции, проживающей на территории Российской Федерации.

Степень достоверности результатов проведенных исследований.

Результаты работы являются статистически достоверными. Автор провел все необходимые статистические тесты для оценки достоверности

результатов и выявления достоверных факторов, влияющих на полученные оценки.

Публикации основных результатов диссертационной работы.

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 квартиля Q1.

Структура диссертационной работы.

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.

В обзоре литературы приводятся основные сведения о молекулярных механизмах регуляции транскрипции, экспериментальных методах определения РСТФ, приведены базы данных и Интернет-порталы, в которых представлен большой массив данных, полученных в результате экспериментов ChIP-seq. Описаны современные алгоритмы и программы по идентификации РСТФ на основе обработки данных ChIP-seq, описаны меры качества при идентификации РСТФ, характерные ошибки методов. Также приводится информация о влиянии однонуклеотидных геномных вариантов на регуляцию транскрипции и раздел по описанию биологии сперматозоидов. Большая часть обзора посвящена описанию математических методов для мета-анализа, в частности, методов коллективного выбора. В целом, обзор литературы формирует основу для целей и задач, поставленных в работе, подтверждает ее актуальность.

В главе «Материалы и методы» описана процедура отбора данных для анализа из БД GTRD. Предложен метод оценки качества ChIP-seq экспериментов. Описано получение материала для анализа связи

полиморфизмов ДНК и качества семенной жидкости у добровольцев мужского пола из шести городов России и Беларуси, описан метод идентификации однонуклеотидных полиморфизмов по данным полноэкзомного секвенирования.

В главе «Результаты и обсуждение» представлены результаты работы. В разделе 3.1 исследована зависимость между согласованностью предсказания локализации РСТФ разными программами и качеством исходных данных ChIP-seq, консервативностью последовательностей РСЕФ, состоянием хроматина, значениями AUC, локализации соседних мотивов РСТФ. Показано, что эти характеристики связаны с согласованностью предсказания РСТФ разными методами: чем больше методов предсказывают РСТФ, тем чаще характеристики, отражающие его функциональную важность коррелируют.

В разделе 3.2 проведена всесторонняя оценка, как разработанная автором метрика False Positive Control Metric (FPCM) связана с характеристиками РСТФ, характеризующими его функциональность (аналогичные характеристикам предыдущего раздела). Этот вопрос был исследован для сайтов, которые идентифицировались лишь одним компьютерным методом (наименее надежные). Автору удалось показать статистически достоверную связь между FPCM и параметров AUC, открытостью хроматина, воспроизводимостью эксперимента и консервативностью последовательности. Тем самым показана эффективность этой метрики для оценки качества идентификации РСТФ.

Предложена метрика оценки доли ложно отрицательных предсказаний локализации РСТФ (FNCM). Показана ее статистическая связь с характеристиками качества экспериментов, показаны различия этого параметра для разных алгоритмов распознавания РСТФ.

В разделе 3.4 приведено описание метода мета-анализа данных ChIP-seq (METARA). На основе его использования проведена аннотация генома человека и определены карты геномных РСТФ 1391 ТФ и ко-факторов

человека. Проведен анализ характеристик РСТФ для ChIP-seq экспериментов разных транскрипционных факторов (локализация в участках открытого хроматина и наличие соответствующих мотивов).

В разделе 3.5 автор применил комплекс разработанных методов для исследования проблемы связи между SNV и нарушением в функции и структуре сперматозоидов. Для решения этой задачи использовались как данные полноэкзомных экспериментов, так и из различных баз данных по экспрессии генов и регуляции человека. В результате выявлен ряд полиморфизмов, которые могут оказывать влияние на процессы сперматогенеза, проведена интерпретация возможных молекулярных механизмов, вовлеченных в этот процесс.

Содержание автореферата соответствует содержанию, основным положения и результатам диссертации.

Вопросы по диссертационной работе/ Замечания

1. Мера FPCM предложенная автором для оценки доли ложноположительных предсказаний РСТФ основана на предположении, что «неизвестное число подлинных РСТФ является случайной величиной с распределением Пуассона», однако какого-либо обоснования этого не приведено. На чем оно основано и можно ли каким-то образом проверить такое предположение?
2. На рисунке 3.4.4 приведен несколько неожиданный результат: в области РСТФ для белка TBP наблюдается наименьшая среди всех факторов доля пиков с мотивом связывания ТФ. Почему такое происходит, ведь мотив TBP, TATA-бокс, известен как классический и наиболее изученный ранее мотив, а связывание его с TBP это сигнал к старту транскрипции для большинства генов? Чем такой результат можно объяснить?
3. К замечаниям следует отнести небрежность в оформлении текста диссертации: (1) Встречается три разных рисунка с одним номером 3.1.2, два рисунка с номером 3.1.1; (2) на некоторых рисунках

присутствуют надписи на английском; (3) не для всех сокращений приведена расшифровка в специальном списке, например ФАФ; (4) в некоторых местах показатель степени значения p-value не приведен в верхнем регистре; (5) на рис. 3.1.2 стр. 74 пропущена панель E, на рис. 3.3.2 не подписаны и не описаны панели; неудачные фразы – кальки с английского (орфаны, корзина и пр.), опечатки.

4. Полагаю, что в формулировке цели работы напрасно отсутствует фраза о практическом применении комплекса разработанных методов для исследования связи между однонуклеотидными вариациями и нарушением структуры сперматозоидов у мужчин, проживающих на территории Российской Федерации.

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, 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
д.б.н., доцент
в.н.с. «ФИЦ Институт цитологии и генетики СО РАН»

Афонников Дмитрий Аркадьевич

Сведения:

Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук»

Адрес организации: 630090, Новосибирск, пр-т Академика Лаврентьева, д.10

Телефон: +7 (383) 363-49-63

e-mail: ada@bionet.nsc.ru

Подпись
удостоверяю зав.канцелярией
ИЦиГ СО РАН
«16» 10 2024 г.