ОТЗЫВ

члена диссертационного совета НТУ.1.5.8.01
Орлова Юрия Львовича
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.

Тема диссертационной работы Семёна Константиновича Колмыкова по созданию компьютерных инструментов исследования геномных данных регуляции экспрессии генов актуальна для компьютерной геномики и биоинформатики в целом, соответствует выбранному научному направлению и специальности «Математическая биология, биоинформатика». Целью исследования была разработка компьютерных методов построения карты районов связывания транскрипционных факторов человека на основе массового сравнительного анализа ChIP-seq экспериментов, и контроля качества данных, что является необходимой основой современной компьютерной геномики, задает стандарт обработки данных и представляется актуальным для дальнейших геномных исследований.

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

Научные положения, сформулированные в диссертации С.К. Колмыкова, основаны на большом объёме материала, к результатам грамотно применена соответствующая статистическая обработка. Выводы адекватно сформулированы на основе полученных результатов и их достоверность не вызывает сомнений.

Научная новизна работы определяется тем, что:

В диссертационной работе С.К. Колмыкова предложены и реализованы новые методы оценки качества ChIP-seq экспериментов на основе анализа

согласованности результатов применения четырёх алгоритмов идентификации районов связывания транскрипционных факторов;

Впервые разработан и реализован новый алгоритм на основе применения методов коллективного выбора для последующего отбора наиболее воспроизводимых районов связывания;

Впервые на основе анализа данных полноэкзомного секвенирования были обнаружены ассоциации однонуклеотидных геномных вариантов с нарушениями морфологии сперматозоидов человека.

Теоретическая и практическая значимость работы.

Теоретическая значимость работы обусловлена тем, что предложены новые методы для контроля качества ChIP-seq экспериментов на основе сравнения результатов разных алгоритмов для выявления районов связывания транскрипционных факторов.

Автором диссертационной работы, Семёном Константиновичем Колмыковым, разработан новый алгоритм применения методов коллективного выбора (METARA) для отбора наиболее воспроизводимых районов связывания транскрипционных факторов на основании их ранжирования, что позволило объединить данные из различных ChIP-seq экспериментов.

Практическая значимость работы состоит в создании уникальной коллекции единообразно обработанных экспериментов ChIP-seq и DNase-seq для генома человека. В рамках исследования были впервые идентифицированы геномные вариации, связанные с нарушениями морфологии сперматозоидов человека. Практические результаты работы С.К. Колмыкова были использованы для создания отечественных и международных веб-ресурсов, широко используемых для биомедицинских исследований.

Степень достоверности результатов проведенных исследований.

Результаты исследования С.К. Колмыкова научно обоснованы, представлены в серии рецензируемых научных публикаций. Достоверность

результатов работы подтверждается детальной статистической обработкой, анализом большого массива данных. Созданная база данных GTRD является высоко востребованной для поддержки исследований по биомедицине, что подтверждается высокой цитируемостью.

Публикации основных результатов диссертационной работы.

По теме диссертации С.К. Колмыкова опубликовано 25 научных публикаций, включая 13 публикаций в журналах, индексируемых в международных базах данных Web-of-Science/Scopus, в том числе 8 публикаций в журналах квартиля Q1. Результаты работы были представлены на серии международных научных конференций.

Структура диссертационной работы.

Диссертационная работа С.К. Колмыкова построена по классической схеме, включая Введение, три основных главы – Обзор литературы (Глава 1), Материалы и методы (Глава 2), Результаты и обсуждение (Глава 3), включает необходимые разделы – Заключение, выводы, список литературы, список используемых сокращений. Работа напечатана на 141 странице, содержит 35 рисунков и 5 таблиц. Есть замечания по нумерации рисунков (см. следующий раздел).

Содержание автореферата соответствует содержанию, основным положения и результатам диссертации.

Вопросы по диссертационной работе/ Замечания

Научная часть работы не вызывает сомнений по существу. Однако, очень много стилистических и оформительских замечаний.

Раздел по интерпретации однонуклеотидных геномных вариаций, ассоциированных с нарушениями сперматогенеза, представлен слишком кратко. Стоило бы его расширить как важную практическую часть работы.

Следует избегать сокращений в заголовках, выводах, заключительных фразах, тем более в смеси сокращений на русском и английском. Например «ФАФ METARA»

Аббревиатура РСТФ для районов связывания транскрипционных факторов не является общепринятой. Чаще пишут ССТФ – «сайты» не «районы» в литературе и на русском, и на английском языке.

Замечание по нумерации рисунков - работа содержит 35 рисунков со сложной, вводящей в заблуждение тройной нумерацией.
При этом Рисунок номер 3.1.1 представлен два раза (разные рисунки, один номер), а Рисунок 3.1.2 даже три раза (одинаковый номер).
Конкретнее по дублям с одинаковыми номерами:
Рисунок 3.1.1 – Схема пересечения результатов работы алгоритмов идентификации пиков
Рисунок 3.1.1 – Плотности распределений результатов идентификации РСТФ
Рисунок 3.1.2 – Плотности распределений результатов идентификации РСТФ
Рисунок 3.1.2 – Взаимосвязь различных геномных аннотаций
Рисунок 3.1.2 – ROC-кривые, полученные для ChIP-seq эксперимент
Указаны несуществующие панели рисунков, например Рисунок 3.1.1Ж.
В то же время есть ссылки (см. Рисунок 3.2.1Б) без соответствующей панели.
«(см. Рисунок 9)» - такого номера рисунка нет в работе.
Подпись к рисунку 3.5.2 относится к чему-то другому – см. «Значение над медианой в “ящике с усами” указывает на количество образцов с выбранным генотипом» и далее -
Рисунок 3.5.2 – (А)
Рисунка номер 3.5.1 – нет, но есть следующий номер 3.5.2 и далее.
Панели рисунков (А,Б,В,Г..) подписаны некорректно.
Например, для рисунков 3.2.4-5, 6 и 7 указано в подписи – (АД). Следует хотя бы написать (панели А,Б,В,Г,Д), указать слово «панель» или «фрагмент»

Нужно сразу ставить ссылки на литературу на биологические результаты, упомянутые в тексте,

Например «В последние несколько десятилетий в различных регионах мира наблюдается снижение мужского репродуктивного потенциала..» (стр.6)
Где это показано, в каких регионах мира, где ссылка?
«Большинство известных однонуклеотидных геномных вариантов (SNV) расположено в регуляторных областях генов…» (стр.6) –
Это утверждение тоже нужно подтвердить ссылкой. Можно также сказать, что большинство вариантов как раз аннотировано в белок-кодирующих частях генов. Нужна ссылка на литературу.
«В 2022 году Suryatenggara с соавт. была опубликована статья» - нужна ссылка на эту статью, журнал (добавить ссылку в текст в месте первого упоминания)
Нужны ссылки на литературу или интернет-ссылки (линки) на упоминаемые электронные ресурсы всюду в тексте, например
проект ENCODE, ENCODE Portal, CistromeDB (стр.7).
См также стр.11: «… базах данных: SRA, GEO и ENCODE»
Это конкретные известные ресурсы – указать о чем идет – база данных, портал, указать конкретную ссылку, лучше на научные статьи, можно и то и другое – ссылка на статью и веб-сайт)
Стр.12: «в соответствии с рекомендациями GATK Best Practices» - нужна ссылка на эти рекомендации, что такое GATK?
Стр.13 – опечатки – падежи во фразе про конференции «Международная конференция… Международной конференции…». Большие буквы «И» в названиях конференций не нужны.
Следует пронумеровать формулы в работе, в том числе указывать все параметры, стандартно выделяя параметры курсивом.
Например «В данном контексте λ описывает ожидаемое количество прочтений в рассматриваемом окне поиска пика» - надо отметить, что такое k в формуле (в оной из формул это число событий, в другой число прочтений ДНК).
Цитирование статей в тексте достаточно по фамилиям авторов и году, без инициалов, например (Gaffney D. J. et al., 2012) – инициалы автора избыточны.

Стр. 38 – опечатки в цитировании «(Y et la., 2014) и (Zhou et Troyanskaya, 2015)» - должно быть “et al”?

Сокращения в Списка сокращений не всегда соответствуют терминологии, например - FN - ложно свидетельствующий об отрицательном результате (False Negative).

Можно указать как «число ошибочных ложных предсказаний», но не как «ложно свидетельствующий». Это стандартная терминологии, непонятен новый введенный термин. То же замечание для других сокращений в списке.

Есть замечания и по нумерации рисунков автореферате (например Рисунок 3.4.1) – избыточная нумерация. Формулы в автореферате представлены без детализации, не все параметры указаны. Например, в сложной формуле log*log*... хотя бы поставить скобки, от чего берется аргумент функции.

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, 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., профессор РАН

Орлов
Юрий Львович

Профессор кафедры информационных технологий и обработки медицинских данных
Центра цифровой медицины
Института цифрового биодизайна и моделирования живых систем
Федерального государственного автономного образовательного учреждения высшего образования Первый Московский государственный медицинский университет имени И.М.Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет)

ПОДПИСЬ ЗАВЕРЯЮ
ник отдела
совета
20 г.

Сведения:

Контактные данные:
тел.: +7(495)6091400, e-mail: y.orlov@sechenov.ru
Специальность, по которой официальным оппонентом защищена диссертация:
03.01.09 – «Математическая биология, биоинформатика»

Адрес места работы:
119048, Москва, ул. Трубецкая, д. 8, стр. 2, ФГАОУ ВО Первый МГМУ имени И.М. Сеченова Минздрава России (Сеченовский Университет)
Тел./факс +7(499)2480181; rektorat@sechenov.ru; https://www.sechenov.ru/contacts/