

## ОТЗЫВ

на автореферат диссертационной работы 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

Диссертационная работа Колмыкова С.К. посвящена очень актуальной проблеме разработки и унификации конвейеров обработки результатов массового секвенирования сайтов связывания (СС) факторов транскрипции (ТФ) в полном геноме, и, в частности, в применении к экспериментальным данным технологии ChIP-seq. В работе на большом объеме экспериментальных данных ChIP-seq проведены сравнение разных методов первичной обработки сырых данных, применение и разработка метрик сравнения качества и, что не менее важно, приводится пример применения единообразно обработанных экспериментов для массового анализа с целью поиска однонуклеотидных геномных вариантов, ассоциированных с нарушениями морфологии сперматозоидов.

В целом, соискатель заслуживает присуждения искомой степени по совокупности актуальности работы, достигнутым результатам, включая и набор публикаций. Есть достаточно обширный список замечаний по изложению работы в автореферате, подозрительно много ошибок в логике изложения. Все замечания, кроме одного носят технический характер. Не является техническим только замечание о термине воспроизводимость результатов картирования ChIP-seq и других данных. Не ясно, какой вклад здесь вносит разнообразие конкретных биологических условий экспериментов, а какой алгоритмы обработки. В тексте нет чёткого разъяснения, в какой степени наблюдаемая изменчивость данных определяется самим объектом, а в какой – биоинформатической обработкой данных.

### Список замечаний

стр. 3. Очень резкая смена темы повествования сразу на первой странице текста, в самом первом разделе, Актуальность темы исследования: ... 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... Важное достоинство GTRD состоит в том, что она ориентирована на получение биологических результатов на основе массового анализа собранных в базе данных, именно это должно быть написано при переходе от общего введения (NGS, ChIP-seq, GTRD) к вопросам массового анализа данных, в частности, от мета-анализа ChIP-seq к теме анализа однонуклеотидных геномных вариантов (SNV) и исследования мужского репродуктивного потенциала, нужно было бы осветить именно как удобство и адаптацию GTRD для получения конкретных биологических результатов.

стр. 4 ...Также до конца нерешённым остается вопрос об интеграции имеющихся данных ... Здесь ошибки и в логике, и в русском языке. Более точно писать - Также не решённым до конца.

Я бы склонился ближе к раздельному написанию, это следует из того, что вопрос как раз решался, но до конца не был решён, поэтому его нельзя считать нерешённым.

Перед списком задач не написано, что это список задач.

Вводится термин однонуклеотидных геномных вариантов (SNV), а затем на стр. 5 – ..однонуклеотидные геномные вариации... Нужно быть строже в терминологии

стр. 6 ...НОСОМОСО (<https://hocomoco11.autosome.ru/>)... Чуть менее года назад вышла версия 12 <https://hocomoco12.autosome.ru/> с гораздо большим числом мотивов сайтов связывания транскрипционных факторов - в 2-3 раза, чем было в версии 11, в версии не использованы данные GTRD? Насколько мне известно переход к версии 12 существенно улучшил как количество, так и качество данных, между двумя выпусками прошло 6 лет. Среди авторов версии Носомосо 12 соискателя найти можно (NAR 2024), а вот версии Носомосо 11 его нет (NAR 2018), см. <https://doi.org/10.1093/nar/gkx1106> и <https://doi.org/10.1093/nar/gkad1077> Возникает вопрос о том, насколько велик вклад автора в версии Носомосо 11 и 12, лучше избегать таких вопросов, так как не все знают, что Носомосо и GTRD плотно интегрированы, а аспирантура автора была закончена в 2020 г., и до этого он несколько лет уже работал с базой данных GTRD, в частности по конвейеру обработки данных для базы данных Носомосо.

стр. 7 ...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... Необходимо указать версию GTRD, она регулярно обновляется.

стр. 8 опечатка – в обзоре литературе

стр. 10 ...Поиск мотивов связывания ТФ (МСТФ) из БД НОСОМОСО в РСТФ...

не указан инструмент

стр. 11. Автореферат заслуживает своей нумерации рисунков. Рисунок 3.1.1 и т.д. - он первый в автореферате

стр. 13, 14 PWM - не определён термин, AUC - не определён термин

рис. 18 ...Анализ воспроизводимости РСТФ разными алгоритмами идентификации РСТФ в рамках одного ChIP-seq эксперимента показал, что в среднем полностью воспроизводится ~10% от общего числа РСТФ... Термин воспроизводимость тут спорен, так как неконсервативность РСТФ в разных экспериментах ChIP-seq не означает, что где-то есть ошибка, я подозреваю, что данные GTRD как раз и могут помочь лучше понимать специфичность работы ТФ в разных типах клеток и на разных стадиях развития. Стремление получить идеальные мета-кластеры приведёт, к отмечено в автореферате, к генам домашнего хозяйства и убьёт все эффекты тканеспецифичности. Не понятно, речь в заключении идёт о наложении результатов peak callers разных, или о наложении ChIP-seq экспериментов разных. В целом, данные агрегации данных отражают частоты экспериментов ChIP-seq, проведённых в разных тканях, стадиях. Некоторые ТФ узкоспецифичны по тканям (AR), другие не очень (Foxa1). Это поле для артефактных выводов из анализа данных.

стр. 21 ...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... Это в какой степени отражает, что очень много данных ChIP-seq для AR по ткани простаты, например ?

Однако, несмотря на приведённые замечания к автореферату, на основании его содержания можно заключить, что работа Колмыкова С.К. соответствует требованиям пп. 2.1–2.6 Положения о присуждении учёных степеней Автономной некоммерческой образовательной организацией высшего образования «Научно-технологический университет «Сириус» утверждённого приказом от 25 декабря 2023 г. № 350/1-ОД-У, предъявляемым к диссертациям на соискание учёной степени кандидата наук, а её автор, Колмыков С.К., заслуживает присуждения учёной степени кандидата биологических наук по специальности 1.5.8. Математическая биология, биоинформатика.

к.б.н., с.н.с. лаборатории эволюционной биоинформатики и теоретической генетики  
Федерального исследовательского центра Институт цитологии и генетики Сибирского  
отделения Российской академии наук

22 октября 2024 года

Левицкий Виктор Георгиевич

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

630090, Новосибирск, ул. Академика Лаврентьева, 10

Тел.: +7(383)363-49-80

E-mail: [icg-adm@bionet.nsc.ru](mailto:icg-adm@bionet.nsc.ru)

Сайт: <http://www.bionet.nsc.ru/>

Подпись Левицкого В.Г. заверяю  
ученый секретарь ИИГ СО РАН



к.б.н.

Г. В. Орлова