МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

На правах рукописи

Рябых Григорий Кириллович

РНК-хроматиновые взаимодействия: базы данных, интегративный анализ и функциональная аннотация

Специальность 1.5.8. Математическая биология, биоинформатика

Автореферат диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2026

Работа выполнена на факультете биоинженерии и биоинформатики «Московский государственный университет имени М.В. Ломоносова»

Научный руководитель: ***Миронов Андрей Александрович***
доктор биологических наук, профессор

Официальные оппоненты: ***Колпаков Федор Анатольевич***
доктор биологических наук, Научно-технологический университет «Сириус», Научный центр генетики и наук о жизни, направление «Вычислительная биология», научный руководитель

Шайтан Алексей Константинович
доктор биологических наук, член-корреспондент РАН, профессор РАН, Московский государственный университет имени М.В. Ломоносова, биологический факультет, кафедра биоинженерии, профессор

Карягина-Жулина Анна Станиславовна
доктор биологических наук, профессор, Национальный исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи, лаборатория биологически активных наноструктур, главный научный сотрудник

Защита диссертации состоится 18 июня 2026 г. в 16:00 на заседании диссертационного совета МГУ.015.10 Московского государственного университета имени М.В. Ломоносова по адресу: 119234, Москва, Ленинские горы, д. 1, стр. 73, Факультет биоинженерии и биоинформатики, ауд. 221.
E-mail: dissovet@belozersky.msu.ru
С диссертацией можно ознакомиться в отделе диссертаций Научной библиотеки МГУ имени М.В. Ломоносова (Москва, Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/3955
Автореферат разослан «___» мая 2026 г.

Ученый секретарь диссертационного совета,
доктор биологических наук Д.В. Чистяков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность и степень разработанности темы исследования

Значительная часть генома эукариот транскрибируется с образованием кодирующих (мРНК) и некодирующих РНК (нкРНК) (Djebali S. et al. 2012), многие из которых функционируют в ядре и участвуют в регуляции транскрипции, ремоделировании и поддержании пространственной организации хроматина. Классическими примерами таких РНК могут служить MALAT1, NEAT1, XIST, TERC, HOTAIR и другие (Engreitz J. et al. 2016).

Для изучения РНК-хроматиновых взаимодействий разработаны высокопроизводительные методы двух основных классов: «один-против-всех» (OTA), позволяющие картировать контакты конкретной РНК со всеми локусами хроматина, и «все-против-всех» (ATA), выявляющие потенциальные контакты всех РНК со всеми локусами ДНК (Рябых Г.К. и др. 2022). Эти подходы сформировали обширный массив публичных данных, однако их использование ограничено существенной методологической проблемой: результаты разных исследований обрабатываются несопоставимыми вычислительными конвейерами, что затрудняет прямое сравнение, интегративный анализ и функциональную интерпретацию данных.

Несмотря на развитие общедоступных ресурсов для аннотации и анализа некодирующих РНК, специализированная курируемая база данных, содержащая полногеномные РНК-хроматиновые взаимодействия, обработанные по единому стандарту, до настоящего времени отсутствовала. Поэтому область характеризуется противоречием между высокой биологической значимостью нкРНК, быстрым накоплением экспериментальных данных и недостатком унифицированных инструментов для их систематизации, сравнения и интеграции с другими типами полногеномных данных.

Проведенный анализ литературы позволил выделить четыре ключевых пробела: отсутствие стандартизированной обработки данных OTA и ATA; дефицит специализированных курируемых ресурсов; отсутствие полномасштабного сравнительного анализа специфичности, разрешения и воспроизводимости методов; ограниченность интегративных подходов, связывающих карты РНК-хроматиновых контактов с эпигенетическими метками, архитектурой хроматина и профилями экспрессии.

В данной работе предпринята комплексная попытка преодоления этих ограничений путем создания унифицированной аналитической платформы, включающей базу данных RNA-Chrom, стандартизированный протокол обработки, инструменты сравнительного анализа и подходы к функциональной аннотации. Работа направлена на количественную оценку характеристик существующих данных РНК-хроматинового интерактома и создание основы для их более надежного использования в исследованиях функций нкРНК.

Цели и задачи исследования

Цель данной работы – количественно оценить разрешение, полноту выявления, специфичность и воспроизводимость данных РНК-хроматинового интерактома, полученных методами OTA и ATA для человека и мыши, путем систематического сравнительного анализа, а также создать аналитическую инфраструктуру для их стандартизированной обработки, хранения, анализа и интеграции с другими типами полногеномных данных.

Для реализации данной цели были поставлены следующие задачи:

1. Собрать, курировать и систематизировать все общедоступные полногеномные данные о РНК-хроматиновых взаимодействиях, полученные методами OTA и ATA, для человека и мыши.
2. Разработать и реализовать универсальный стандартизированный вычислительный протокол обработки сырых данных всех типов экспериментов РНК-хроматинового интерактома, включая новый инструмент Fastq-dupaway для эффективного удаления ПЦР-дубликатов.
3. Разработать и наполнить аналитическую базу данных RNA-Chrom для хранения, унификации и анализа данных РНК-хроматинового интерактома.
4. Разработать пользовательский веб-интерфейс и его функционал для интерактивного доступа, анализа и визуализации данных.
5. Реализовать и апробировать интеграцию RNA-Chrom с ресурсом HiMoRNA для генерации функциональных гипотез о роли длинных некодирующих РНК в эпигенетической регуляции хроматина.
6. Разработать и применить метрики для оценки специфичности (хроматиновый потенциал) и воспроизводимости (конкордантность) РНК-хроматиновых взаимодействий.
7. Сформулировать практические рекомендации по повышению достоверности анализа данных РНК-хроматинового интерактома.

Объект и предмет исследования

Объектом исследования являются полногеномные данные о физических взаимодействиях молекул РНК с хроматином (РНК-хроматиновый интерактом), полученные экспериментальными методами классов «один-против-всех» (OTA) и «все-против-всех» (ATA) на клеточных линиях человека и мыши.

Предметом исследования являются биоинформатические методы и аналитические подходы для стандартизации, сравнительной оценки и интеграции данных РНК-хроматинового интерактома с целью повышения достоверности их биологической интерпретации.

Научная новизна работы

В диссертационной работе впервые разработан и применен единый стандартизированный вычислительный протокол для обработки сырых данных всех основных методов изучения

РНК-хроматинового интерактома (OTA и ATA), что обеспечило сопоставимость ранее разрозненных наборов данных.

Создана первая специализированная аналитическая база данных RNA-Chrom, курирующая более 5 миллиардов РНК-хроматиновых контактов, полученных в 20 экспериментах ATA и 189 экспериментах OTA для человека и мыши. База данных поддерживается веб-интерфейсом для интерактивного анализа и визуализации.

Разработан новый программный инструмент Fastq-dupaway для ресурсоэффективного удаления ПЦР-дубликатов из данных секвенирования нового поколения (next-generation sequencing, NGS), характеризующийся предсказуемым низким потреблением оперативной памяти, что решает проблему обработки больших наборов данных на инфраструктуре с ограниченными ресурсами.

Предложена и апробирована интегративная схема, связывающая физические РНК-хроматиновые контакты (RNA-Chrom) с данными о корреляциях между экспрессией нкРНК и эпигенетическими метками (HiMoRNA). На примере конкретных нкРНК (PVT1, MIR31HG и др.) показано, что такая интеграция позволяет формулировать интерпретируемые гипотезы о механизмах эпигенетической регуляции экспрессии генов длинными некодирующими РНК.

Впервые проведен полномасштабный сравнительный анализ всего корпуса данных РНК-хроматинового интерактома, позволивший количественно охарактеризовать и сопоставить методы OTA и ATA по ключевым параметрам: разрешение, полнота и специфичность.

Теоретическая и практическая значимость

Результаты работы важны для развития методологии анализа РНК-хроматинового интерактома и интерпретации функций РНК, ассоциированных с хроматином. Теоретическая значимость исследования состоит в разработке единых принципов обработки и количественной оценки данных OTA и ATA, определении их разрешения, полноты, специфичности и воспроизводимости, а также в предложении интегративной схемы перехода от карт физических контактов к функциональным гипотезам о роли нкРНК в регуляции генома.

Практическая значимость работы связана с созданием базы данных RNA-Chrom и веб-интерфейса, позволяющих исследователям работать с унифицированными данными без их самостоятельной повторной обработки. Разработанный инструмент Fastq-dupaway решает задачу ресурсоэффективного удаления ПЦР-дубликатов из NGS-данных и может применяться в других геномных конвейерах. Интеграция RNA-Chrom и HiMoRNA создает рабочий прототип мультиомиксного анализа, а предложенная стратегия отбора достоверных РНК-хроматиновых взаимодействий повышает надежность функциональной интерпретации и отбора генов-кандидатов, регуляторных локусов и проверяемых биологических гипотез.

Методология исследования

Методология исследования носит междисциплинарный характер и построена в соответствии с принятыми стандартами биоинформатики и вычислительной биологии. Она объединяет разработку специализированных программных конвейеров и баз данных (RNA-Chrom, Fastq-dupaway) со статистическим и сравнительным анализом для оценки разрешения, полноты и специфичности данных. Особое внимание уделено предотвращению методологических артефактов: для обеспечения достоверности выводов использованы принципы работы с независимыми репликами, кросс-валидация между разными методами (OTA и ATA) и проверка интегративных гипотез на независимых экспериментальных данных и примерах нкРНК с известными функциями.

Положения, выносимые на защиту

1. Единый стандартизированный протокол обработки и созданная на его основе аналитическая база данных RNA-Chrom обеспечивают сопоставимость полногеномных данных РНК-хроматинового интерактома и создают основу для их сравнительного и интегративного анализа.
2. Программный инструмент Fastq-dupaway обеспечивает ресурсоэффективное удаление ПЦР-дубликатов из данных NGS благодаря предсказуемо низкому потреблению оперативной памяти (~2 ГБ) и высокой скорости работы, что обеспечивает эффективную обработку больших наборов данных NGS.
3. Интеграция RNA-Chrom и HiMoRNA позволяет формулировать и приоритизировать интерпретируемые гипотезы о функциональной роли длинных некодирующих РНК в эпигенетической регуляции хроматина на основе совместного анализа данных о физических контактах РНК с хроматином и корреляциях экспрессии РНК с эпигенетическими метками.
4. Данные OTA характеризуются более высоким разрешением и воспроизводимостью, чем данные ATA, и могут использоваться как референс для валидации взаимодействий, выявленных в полногеномных подходах. Специфический сигнал эффективно выделяется, если отбирать РНК с высоким хроматиновым потенциалом (данные ATA) и воспроизводимые контакты из статистически значимых пиков (данные OTA и ATA).

Степень достоверности данных

Все экспериментальные данные, использовавшиеся в работе, находятся в открытом доступе, и результаты их анализа могут быть воспроизведены. Разработанные аналитическая инфраструктура (база данных RNA-Chrom с веб-интерфейсом) и программа Fastq-dupaway находятся в открытом доступе, а их документация предоставлена для проверки и повторного использования научным сообществом. Результаты, представленные в работе, переносимы между

независимыми экспериментами схожей природы. Обзор литературы и обсуждение подготовлены с использованием актуальной литературы.

Публикации

По теме диссертации опубликовано 6 печатных работ, из них 5 статей в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 1.5.8. Математическая биология, биоинформатика.

Личный вклад автора

В работе (Рябых Г.К. и др. 2022) лично автором проведен детальный литературный анализ методов «один-против-всех» и их применения к биологическим задачам. В работе (Ryabykh G.K. et al 2023) непосредственно автором выполнена разработка базы данных RNA-Chrom, ее функционала и веб-интерфейса. В работе (Ильницкий И.С. и др. 2025) автором выполнена адаптация веб-ресурса RNA-Chrom для интеграции с HiMoRNA, процедура соответствия названий генов из двух баз данных, оценена согласованность результатов HiMoRNA и RNA-Chrom; реализован вариант использования интеграции двух веб-ресурсов на примере днРНК PVT1. В работе (Sigorskikh A.I. et al 2025) под руководством автора диссертации была протестирована новая программа удаления ПЦР-дубликатов на 15 наборах данных секвенирования нового поколения различных типов и размеров. В работе (Ryabykh G.K. et al. 2025) автором выполнен анализ хроматинового потенциала, анализ конкордантности реплик и данных «все-против-всех» в сравнении с «один-против-всех».

Апробация результатов

Результаты работы были представлены на 7 международных и российских научных конференциях:

1. 25-30 сентября 2018, Информационные технологии и системы, Казань, Россия, стендовый доклад, «Анализ данных РНК-ДНК взаимодействий»;
2. 30 сентября - 2 октября 2020, ENCODE 2020: Research Applications and Users Meeting, онлайн, стендовый доклад, «RNA-Chrom: database genome-wide RNA-chromatin interactions»;
3. 10 - 27 ноября 2020, Lomonosov 2020, Москва, Россия, стендовый доклад, «Comparative analysis tools for RNA-chromatin interactome of cells»;
4. 2-6 октября 2022, Информационные технологии и системы, Огниково, Московская область, стендовый доклад, «The RNA-Chrom database opens up new possibilities for the analysis of the RNA-chromatin interactome»;

5. 30 июля - 2 августа 2021, MCCMB-2021 (Moscow Conference on Computational Molecular Biology), Москва, Россия, стендовый доклад, «A new database for genome-wide RNA-chromatin interactome»;
6. 3-6 августа 2023, MCCMB-2023 (Moscow Conference on Computational Molecular Biology 2023), устный доклад, «Сравнительный анализ данных РНК-хроматиновых взаимодействий»;
7. 17-21 сентября 2023, Информационные технологии и системы, Огниково, Московская область, устный доклад, Огниково, Московская область, «Сравнение данных РНК-ДНК контактов из разных экспериментов».

Соответствие диссертации паспорту научной специальности

Представленные в диссертации результаты принадлежат областям исследования «компьютерная системная биология» и «разработка и применение новых вычислительных алгоритмов для анализа экспериментальных данных в биологии и медицине». Диссертация соответствует паспорту специальности 1.5.8. Математическая биология, биоинформатика.

Структура и объем диссертации

Диссертационная работа состоит из титульного листа, оглавления, списка сокращений и условных обозначений, введения, обзора литературы, материалов и методов, результатов, заключения, выводов, списка литературы, списка публикаций по теме диссертации и приложений. Работа изложена на 157 страницах, иллюстрирована 50 рисунками, 19 таблицами и 3 приложениями. Список литературы состоит из 203 источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Материалы и методы исследования

В работе использовались общедоступные полногеномные данные РНК-хроматиновых взаимодействий, включающие 189 наборов OTA и 20 наборов ATA для человека и мыши, загруженные из GEO (*http://www.ncbi.nlm.nih.gov/geo*). Для подготовки и аннотации данных применялись референсные геномы и аннотации генов человека (GRCh38), мыши (GRCm38), домашней свиньи (Sscrofa11.1) и североамериканского красногорлого анолиса (AnoCar2.0v2) из Ensembl, UCSC и GENCODE; аналитические разделы выполнены для человека и мыши, поскольку для этих организмов были доступны как OTA-, так и ATA-данные. Дополнительно использовались 15 наборов NGS-данных разных типов для тестирования Fastq-dupaway (загруженны из SRA, *https://www.ncbi.nlm.nih.gov/sra*), данные корреляций днРНК с эпигенетическими метками из HiMoRNA (Mazurov E. et al. 2022) и данные RNA-seq выбранных клеточных линий человека (MDA-MB-231, MM.1S, H1 ES, K562) и мыши (OPC, ES).

В качестве основных инструментов применялись разработанный стандартизированный

конвейер обработки РНК-хроматинового интерактома, включающий контроль качества (FastQC), дедупликацию (Fastq-dupaway), выравнивание (HISAT2), фильтрацию (SAMtools) и идентификацию значимых взаимодействий (BaRDIC), база данных RNA-Chrom (*https://rnachrom2.bioinf.fbb.msu.ru*) на основе ClickHouse с веб-интерфейсом, программа Fastq-dupaway (*https://github.com/AndrewSigorskih/fastq-dupaway*), а также методы интегративного и сравнительного анализа, включая хроматиновый потенциал и конкордантность взаимодействий.

Результаты и их обсуждение

1. База данных RNA-Chrom

Проведение сравнительного анализа данных полногеномного РНК-хроматинового интерактома – важная научная задача, однако оно сталкивается с отсутствием единых стандартов обработки и консолидированных ресурсов, что не позволяет сравнивать результаты. В качестве решения этой проблемы и для систематизации растущего массива информации была разработана специализированная аналитическая база данных RNA-Chrom (*https://rnachrom2.bioinf.fbb.msu.ru*). База данных аккумулирует 24 набора данных, полученных экспериментальными методами «все-против-всех» («all-to-all», ATA), выявляющими контакты всех транскрибируемых РНК с хроматином, и 189 наборов данных, полученных экспериментальными методами «один-против-всех» («one-to-all», OTA), направленными на изучение контактома конкретной целевой РНК.

Ключевой особенностью RNA-Chrom является применение универсального протокола обработки ко всем наборам данных, начиная с этапа сырых чтений (рис. 1). Этот протокол включает этапы удаления ПЦР-дубликатов, контроля качества прочтений, выравнивания на референсный геном, фильтрации, *de novo* сборки неаннотированных хроматин-ассоциированных РНК (unannotated chromatin-associated RNAs, ucaRNAs) и идентификации значимых взаимодействий, что обеспечивает сопоставимость результатов из разных исследований и выполненных разными методами.

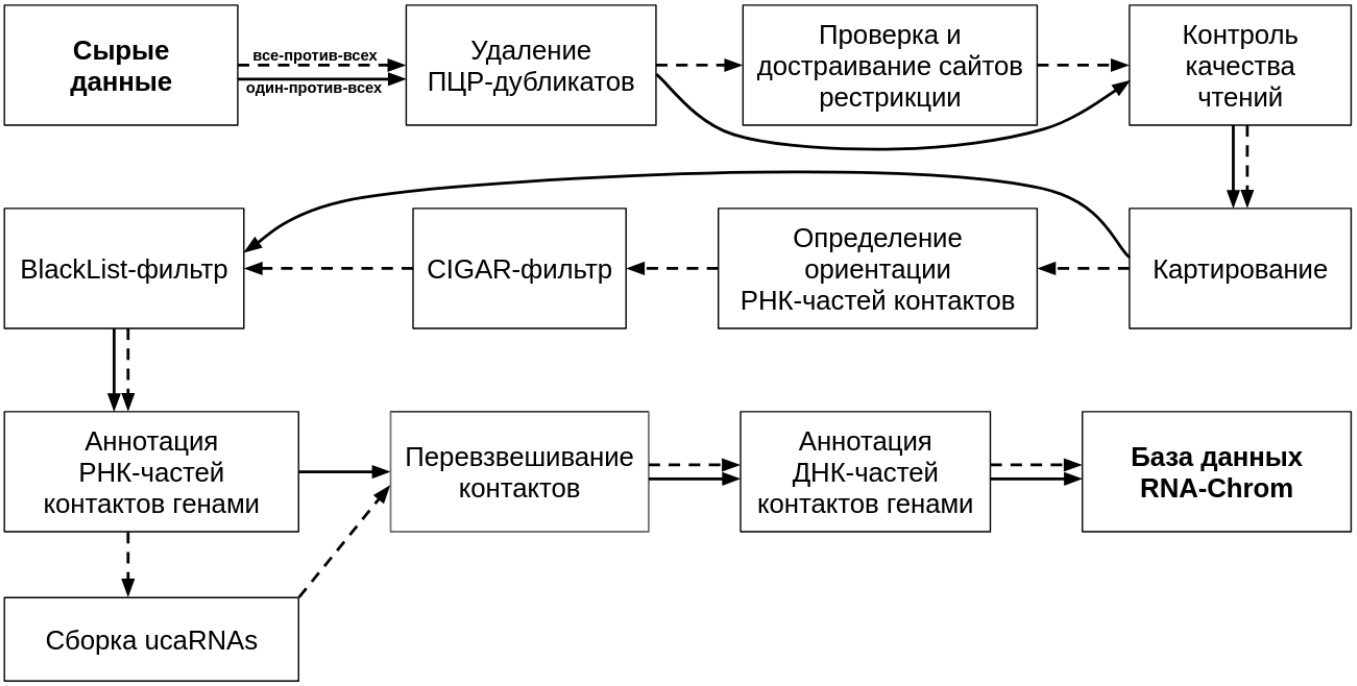

Рисунок 1. Протокол обработки данных РНК-хроматиновых взаимодействий. Пунктирные стрелки соответствуют этапам обработки данных ATA, а сплошные стрелки относятся к данным OTA.

Веб-интерфейс RNA-Chrom поддерживает два основных сценария анализа: «от РНК», когда пользователь получает геномные локусы, контактирующие с выбранной РНК, и «от ДНК», когда для выбранного гена или геномного локуса определяется набор взаимодействующих с ним РНК. Для этих сценариев реализованы таблицы контактирующих РНК и генов, визуализация распределения контактов, экспорт результатов и просмотр карт контактов в UCSC Genome Browser для сопоставления с геномными и эпигенетическими аннотациями. Кроме того, пользователь может экспортировать обработанные данные, карты контактов, таблицы контактирующих РНК/генов и сопутствующие метаданные для последующего анализа.

Для работы с большим объемом данных база данных RNA-Chrom реализована на основе СУБД ClickHouse и содержит более 5 миллиардов аннотированных РНК-хроматиновых контактов. Ресурс интегрирует данные для нескольких модельных организмов: человек (13 экспериментов ATA, 64 OTA), мышь (7 ATA, 125 OTA), домашняя свинья (2 ATA) и североамериканский красногорлый анолис (2 ATA). Сравнительный анализ в диссертации выполнен для человека и мыши, поскольку для этих организмов доступны как ATA-, так и OTA-данные. Аннотация охватывает как известные гены, так и гены неаннотированных хроматин-ассоциированных РНК.

Таким образом, RNA-Chrom представляет собой централизованный и стандартизированный ресурс для анализа РНК-хроматинового интерактома. База данных позволяет исследователям работать с унифицированными данными без необходимости их самостоятельной повторной обработки и создает основу для сравнительного и интегративного анализа функций некодирующих РНК.

2. ПЦР-дедуплексатор Fastq-dupaway

Проблема эффективного удаления ПЦР-дубликатов из данных высокопроизводительного секвенирования (NGS) остается актуальной, особенно при работе с большими наборами данных в условиях ограниченных вычислительных ресурсов. Существующие решения часто имеют узкую специализацию, ограниченную поддержку форматов или непредсказуемое потребление памяти. Для решения этих задач был разработан Fastq-dupaway (*https://github.com/AndrewSigorskih/fastq-dupaway*) – гибкий инструмент, предоставляющий пользователю выбор между скоростью и контролем над ресурсами. Программа поддерживает одно- и парноконцевые данные в форматах FASTQ/FASTA и предоставляет пользователю выбор между быстрым режимом работы и режимами с контролируемым потреблением оперативной памяти.

Fastq-dupaway был протестирован на 15 наборах данных NGS различных типов, включая RNA-seq, ChIP-seq и Hi-C. Сравнительный анализ показал, что режимы на основе внешней сортировки обеспечивают предсказуемое потребление оперативной памяти около 2 ГБ независимо от размера входных данных, что позволяет обрабатывать большие наборы данных на вычислительной инфраструктуре с ограниченными ресурсами. Быстрый режим, основанный на сравнении хэшей последовательностей в оперативной памяти, демонстрирует высокую скорость обработки и может использоваться в случаях, когда ограничение RAM не является критическим.

Сравнение конвейеров обработки данных показало, что применение Fastq-dupaway до выравнивания позволяет снизить вычислительные затраты по сравнению с alignment-based подходами к удалению дубликатов, требующими предварительного картирования, сортировки и последующей обработки BAM-файлов. Таким образом, Fastq-dupaway является универсальным и ресурсоэффективным инструментом предобработки NGS-данных, применимым как в задачах анализа РНК-хроматинового интерактома, так и в других геномных исследованиях.

3. Интеграция баз данных HiMoRNA и RNA-Chrom

Для перехода от карт физических РНК-хроматиновых взаимодействий к функциональной интерпретации была выполнена интеграция баз данных RNA-Chrom и HiMoRNA. База HiMoRNA содержит данные о корреляциях между экспрессией длинных некодирующих РНК (днРНК) и эпигенетическими метками в конкретных геномных локусах, тогда как RNA-Chrom позволяет проверить, контактирует ли соответствующая днРНК с анализируемым локусом. Таким образом, интеграция двух ресурсов позволяет оценивать, могут ли предсказанные ассоциации «днРНК – эпигенетическая метка» быть опосредованы физическим нахождением днРНК в соответствующей области хроматина.

Было установлено взаимно однозначное соответствие между генами в двух базах данных, что позволило для 4124 из 4145 днРНК из HiMoRNA сформировать запросы к RNA-Chrom.

Анализ согласованности показал, что 96,8% днРНК из HiMoRNA (4011 из 4145) имеют хотя бы один подтверждающий контакт в RNA-Chrom. Доля предсказаний HiMoRNA, подтвержденных контактами RNA-Chrom, возрастала при расширении окна вокруг контакта и достигала 53% при ±50 тыс. п.н. Это указывает на то, что функциональная связь днРНК с эпигенетической меткой может быть обусловлена не только точным совпадением координат контакта и пика, но и нахождением РНК в регуляторной окрестности соответствующего локуса.

Статистический анализ с использованием точного теста Фишера выявил 32 пары «днРНК – гистоновая метка», для которых контакты RNA-Chrom значимо чаще подтверждали пики с определенным направлением корреляции. Для 21 днРНК (например, MIR4435-2HG) лучше подтверждались положительно коррелированные пики активирующих меток, включая H3K27ac и H3K4me3, что согласуется с потенциальным участием этих РНК в активации транскрипции. Для 11 днРНК лучше подтверждались отрицательно коррелированные пики, что может указывать на их участие в репрессивной регуляции.

Практическая применимость подхода была продемонстрирована на примерах днРНК MIR31HG и PVT1. Для MIR31HG интеграция подтвердила связь с ранее описанным геном-мишенью GLI2 и позволила выдвинуть гипотезу о более широком участии этой днРНК в регуляции генов сигнального пути Hedgehog. Анализ генов, ассоциированных с подтвержденными контактами MIR31HG и пиками H3K27ac, выявил обогащение генами этого сигнального пути.

Для днРНК PVT1, которая, как известно, рекрутирует репрессорный комплекс PRC2 (EZH2) для подавления гена LATS2, интеграция выявила в HiMoRNA только отрицательно коррелированные с экспрессией PVT1 пики активирующей метки H3K4me3 в области LATS2. Это косвенное подтверждение репрессивной функции. При этом в RNA-Chrom были обнаружены контакты PVT1 в данном регионе, а независимые данные Red-ChIP подтвердили наличие EZH2-опосредованных контактов PVT1 вблизи LATS2.

Таким образом, интеграция RNA-Chrom и HiMoRNA позволяет переходить от корреляционных ассоциаций к проверке физической близости днРНК к потенциальным регуляторным локусам. Такой подход дает возможность формулировать интерпретируемые гипотезы о функциональной роли длинных некодирующих РНК в эпигенетической регуляции генов.

4. Сравнительный анализ данных РНК-хроматинового интерактома: разрешение, полнота и специфичность данных

Проведен комплексный сравнительный анализ данных РНК-хроматинового интерактома, полученных методами «все-против-всех» (ATA) и «один-против-всех» (OTA). Основное внимание уделено трем характеристикам, критически важным для интерпретации полногеномных данных: специфичности сигнала, воспроизводимости между репликами и полноте выявления

РНК-хроматиновых взаимодействий.

Для количественной оценки специфичности связывания РНК с хроматином введен и применен хроматиновый потенциал (chP) – метрика, показывающая, насколько доля контактов данной РНК статистически значимо отличается от ожидаемой при неспецифическом связывании, пропорциональном уровню экспрессии. Как и ожидалось, большинство нкРНК демонстрировало положительные значения chP, что свидетельствует об их повышенном сродстве к хроматину. Однако значительное количество белок-кодирующих РНК (мРНК) также имело chP > 0. Этот факт может объясняться наличием в интронах белок-кодирующих генов неаннотированных функциональных нкРНК или регуляторной ролью некодирующих изоформ самих мРНК. При увеличении порога на chP доля мРНК среди прошедших порог РНК снижается с резким падением почти во всех экспериментах при значениях chP $\geq$ 20. При этом положительный chP и воспроизводимость контактов мРНК в настоящей работе не интерпретируются как прямое доказательство хроматиновой функции зрелых мРНК, поскольку для такого вывода необходим отдельный транскрипт-специфичный анализ экзонных и интронных контактов.

Для оценки разрешения ATA-методов мы анализировали зависимость доли конкордантных контактов между репликами от допустимого геномного расстояния L. Было показано, что медианная доля конкордантных контактов выходит на плато при $L \geq 5000$ п.н., что позволяет рассматривать 5000 п.н. как эмпирическую оценку разрешения ATA-данных. В дальнейшем воспроизводимость контактов мы оценивали на уровне неперекрывающихся геномных бинов размером 5000 п.н. Бин считался конкордантным для данной РНК, если как минимум один контакт в этом бине был обнаружен в обеих репликах, и дискордантным, если в нем есть контакты только в одной реплике. Такой подход позволяет агрегировать данные и количественно оценить воспроизводимость взаимодействий на уровне геномных локусов.

Общая воспроизводимость, а следовательно, и полнота данных для методов Red-C и RADICL-seq оказалась низкой: медианная доля контактов, приходящихся на конкордантные бины размером 5000 п.н., не превышала 2% при анализе всех данных и 5% после фильтрации контактов по пикам BaRDIC. Напротив, воспроизводимость (и, следовательно, полнота) данных GRID-seq оказалась существенно выше, достигая 29% и 82% для всех контактов и контактов из пиков соответственно. Высокую воспроизводимость данных GRID можно объяснить особенностями протокола фиксации. В отличие от методов, использующих только формальдегид (таких как Red-C и RADICL-seq), в протоколе GRID-seq применяется двухэтапная фиксация дисукцинимидилглутаратом (DSG) и формальдегидом. DSG эффективно сшивает белок-белковые взаимодействия до фиксации хроматиновой структуры формальдегидом, что обеспечивает лучшую стабилизацию опосредованных белками РНК-хроматиновых контактов и, как следствие, резкое снижение технического шума. Это подтверждается тем, что при сопоставимом общем числе

контактов медианное количество воспроизводимых контактов в данных GRID-seq на порядок превышало таковое для других методов.

Для всех методов ATA обнаружена устойчивая положительная корреляция между долей конкордантных контактов и двумя параметрами: общим числом детектированных взаимодействий для данной РНК и ее хроматиновым потенциалом (рис. 2). Это позволяет сделать два важных вывода: 1) полнота данных для конкретной РНК сильно зависит от глубины секвенирования, и надежный анализ возможен только для РНК с большим числом контактов (>10 000); 2) воспроизводимость является маркером биологической значимости, а хроматиновый потенциал служит ее предиктором.

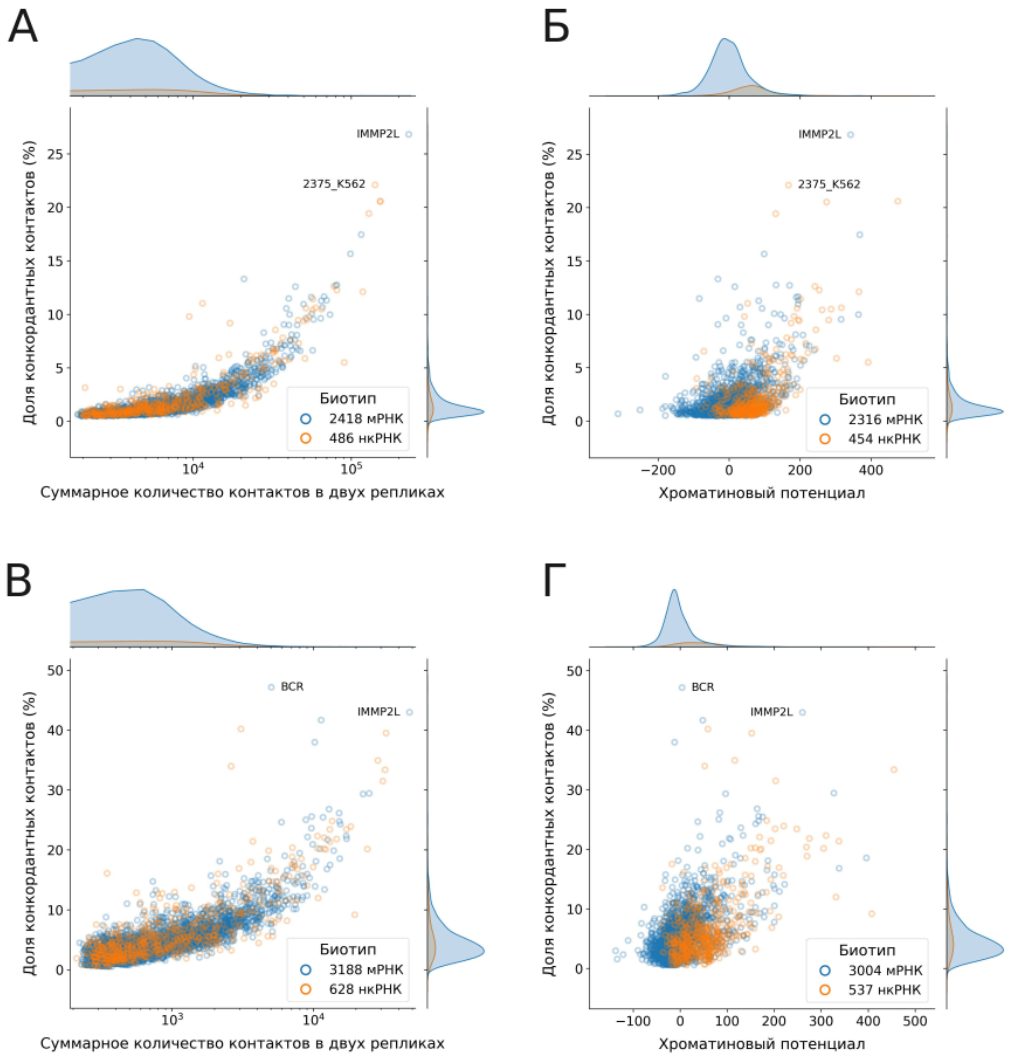

Рисунок 2. Зависимость конкордантности реплик от числа контактов и от хроматинового потенциала. А и Б – Конкордантность рассчитывалась по всем контактам; В и Г – по контактам из пиков BaRDIC. Представлены данные Red-C на клетках K-562, размер бина 5000 п.н. MALAT1 на графике не отображается, так как эта РНК имеет экстремальное значение хроматинового потенциала и доли конкордантных контактов: 991 и 58,2% – на панелях А и Б; 740 и 71,9% – на панелях В и Г.

Данные экспериментов OTA продемонстрировали ожидаемо высокую воспроизводимость между репликами – более 90% контактов были конкордантными даже при строгом размере бина в

1000 п.н. Это указывает на то, что данные OTA обладают разрешением в 1000 п.н. и высокой полнотой, что позволяет рассматривать OTA в качестве «золотого стандарта». Однако критически важным оказался анализ специфичности: при переходе к рассмотрению только тех контактов, которые попадают в статистически значимые пики BaRDIC, уровень конкордантности между репликами падал практически вдвое. Это свидетельствует о том, что значительная часть всех детектированных в OTA взаимодействий (включая воспроизводимые), вероятно, является неспецифическим фоновым сигналом. Таким образом, для повышения специфичности анализа данных OTA также необходима обязательная фильтрация по пикам.

Прямое сопоставление ATA и OTA было ограничено малым числом парных публичных данных для одних и тех же РНК и клеточных линий. Тем не менее на доступных примерах MALAT1 и JPX оно подтвердило принципиальную возможность кросс-валидации взаимодействий между методами и показало, что данные OTA могут использоваться как референс для оценки результатов ATA.

На основании комплексного анализа нами предложена и обоснована двухэтапная стратегия для достоверного анализа данных РНК-хроматинового интерактома, полученного методами OTA и ATA. Применение данной стратегии позволяет существенно повысить специфичность и надежность интерпретации данных полногеномных исследований.

- При анализе данных OTA следует ориентироваться на контакты, прошедшие фильтрацию по пикам (например, с помощью BaRDIC), поскольку они демонстрируют значительно более высокую специфичность. Высокая общая воспроизводимость данных OTA подтверждает их надежность как референса.
- Обратим внимание на то, что хроматиновый потенциал отбирает перспективные РНК, в то время как анализ конкордантности и поиск пиков отбирают значимые контакты РНК с хроматином. Поэтому при анализе данных ATA стратегия должна быть двухуровневой:

1. на первом этапе необходимо отбирать РНК с высоким хроматиновым потенциалом (chP > 20), что позволяет сфокусироваться на молекулах с повышенной вероятностью специфических взаимодействий с хроматином;
2 на втором этапе необходимо отбирать РНК с числом контактов >10 000, а для их анализа использовать исключительно те контакты, которые одновременно и попадают в пики BaRDIC, и воспроизводимы между репликами.

Таким образом, комбинированное использование хроматинового потенциала (для отбора РНК) и конкордантных контактов из пиков (для отбора геномных локусов) позволяет максимально отфильтровать неспецифический шум и выделить наиболее достоверные взаимодействия. Предложенный подход позволяет повысить надежность биоинформатического анализа и интерпретации данных РНК-хроматинового интерактома, что особенно важно для выявления

функционально значимых связей.

ЗАКЛЮЧЕНИЕ

В настоящей диссертационной работе решена комплексная задача по разработке методологических основ, инструментов и аналитической инфраструктуры для систематического изучения РНК-хроматинового интерактома. Для преодоления несопоставимости опубликованных данных была создана база RNA-Chrom и реализован единый вычислительный протокол обработки данных, полученных методами «один-против-всех» и «все-против-всех».

Использование единого протокола обработки позволяет перейти от разрозненного анализа отдельных экспериментов к сопоставимому изучению РНК-хроматиновых взаимодействий в общем контексте разных методов, клеточных систем и геномных локусов. RNA-Chrom обеспечивает возможность анализа как «от РНК», так и «от ДНК», что позволяет не только воспроизводить опубликованные наблюдения, но и выявлять потенциально функциональные связи между РНК, генами и регуляторными областями хроматина.

Для повышения вычислительной эффективности предобработки данных был разработан инструмент Fastq-dupaway, предназначенный для удаления ПЦР-дубликатов из NGS-данных. Его применение позволяет снижать вычислительные затраты при обработке больших наборов данных и делает унифицированный анализ РНК-хроматинового интерактома более доступным для исследователей, работающих с ограниченными вычислительными ресурсами.

Для перехода от карт физических взаимодействий к функциональной интерпретации была выполнена интеграция RNA-Chrom с базой HiMoRNA. Такой подход позволяет сопоставлять данные о контактах РНК с хроматином с корреляциями между экспрессией длинных некодирующих РНК и эпигенетическими метками, что способствует формированию интерпретируемых гипотез о роли днРНК в регуляции генов.

Проведенный сравнительный анализ данных OTA и ATA позволил количественно охарактеризовать их разрешение, полноту и специфичность. Показано, что данные OTA обладают более высоким разрешением и воспроизводимостью, тогда как интерпретация данных ATA требует дополнительного отбора РНК с высоким хроматиновым потенциалом и анализа воспроизводимых контактов, попадающих в статистически значимые пики.

Предложенные решения имеют ограничения применимости: база RNA-Chrom и вычислительный протокол ориентированы прежде всего на общедоступные полногеномные данные OTA и ATA, интеграция с HiMoRNA предназначена для генерации функциональных гипотез, а достоверность анализа зависит от наличия биологических реплик, покрытия данных и качества фильтрации контактов. Тем не менее разработанные инструменты и подходы создают основу для более надежного биоинформатического анализа РНК-хроматинового интерактома и дальнейшей экспериментальной проверки функций некодирующих РНК.

Практическая востребованность RNA-Chrom подтверждается использованием ресурса в независимых исследованиях, включая анализ cis-регуляторных длинных некодирующих РНК и построение интегративных моделей РНК–белок–ДНК взаимодействий. Это показывает, что RNA-Chrom может служить не только хранилищем унифицированных данных, но и опорной платформой для дальнейшей биологической интерпретации РНК-хроматиновых взаимодействий.

Основные результаты и выводы

1. Разработан универсальный стандартизированный протокол обработки данных РНК-хроматинового интерактома, обеспечивающий их сопоставимость, и создан программный инструмент Fastq-dupaway для ресурсоэффективного удаления ПЦР-дубликатов.
2. Создана первая специализированная аналитическая база данных RNA-Chrom, содержащая исчерпывающий массив общедоступных полногеномных данных РНК-хроматинового интерактома человека и мыши, включающий результаты 189 экспериментов OTA и 20 экспериментов ATA (более 5 миллиардов аннотированных контактов РНК с хроматином). Для базы разработан и внедрен пользовательский веб-интерфейс, обеспечивающий сценарии анализа «от РНК» и «от ДНК», а также средства визуализации и фильтрации данных.
3. Осуществлена интеграция RNA-Chrom и HiMoRNA, позволившая сформулировать интерпретируемые гипотезы о потенциальной функциональной роли конкретных днРНК в эпигенетической регуляции генов: в частности, для MIR31HG – в регуляции генов GLI2 и PTCH1 и сигнального пути Hedgehog, а для PVT1 – в регуляции гена LATS2.
4. На едином корпусе данных OTA и ATA количественно охарактеризованы специфичность и воспроизводимость контактов РНК с хроматином. Показано, что данные OTA обладают более высоким разрешением (~1000 п.н.) и воспроизводимостью, тогда как данные ATA характеризуются более низким разрешением (~5000 п.н.), а воспроизводимость сигнала в репликах существенно зависит от протокола фиксации.
5. Выработана практическая стратегия повышения достоверности анализа данных РНК-хроматинового интерактома, основанная на отборе РНК с высоким хроматиновым потенциалом (ATA) и воспроизводимых контактов из статистически значимых пиков (OTA и ATA).

СПИСОК ПУБЛИКАЦИЙ

Статьи в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности и отрасли наук[1].

1. **Рябых Г.К.**, Мыларщиков Д.Е., Кузнецов С.В., Сигорских А.И., Пономарёва Т.Ю., Жарикова А.А., Миронов А.А. РНК-хроматиновый интерактом. Что? Где? Когда? // Молекулярная биология. – 2022. – Т. 56, № 2. – С. 275-295. EDN: COLJSF. Импакт-фактор 0,7 (JIF) (2.25/1.20).
2. **Ryabykh G.K.**, Kuznetsov S.V., Korostelev Y.D., Sigorskikh A.I., Zharikova A.A., Mironov A.A. RNA-Chrom: a manually curated analytical database of RNA–chromatin interactome // Database. – 2023. – vol. 2023. baad025. EDN: YEKQIZ. Импакт-фактор 3,6 (JIF) (1.02/0.40).
3. **Ryabykh G.K.**, Nikolskaya A.I., Garkul L.D., Mironov A.A. Comparative analysis of RNA-chromatin interactome data: resolution, completeness, and specificity // Biochemistry (Moscow). – 2025. – vol. 90, № 11. – pp. 1816-1829. EDN: PORMJW. Импакт-фактор 2,2 (JIF) (1.23/0.50).
4. Sigorskikh A.I., Kompaniets M.A., Ilnitskiy I.S., **Ryabykh G.K.,** Mironov A.A. Fastq-dupaway: a fast and memory-efficient tool for deduplication of single- and paired-end NGS data // Scientific Reports. – 2025. – vol. 15. – 45303. EDN: VBEHEL. Импакт-фактор 3,9 (JIF) (0.88/0.20).
5. Ильницкий И.С., **Рябых Г.К.**, Маракулина Д.А., Миронов А.А., Медведева Ю.А. Интеграция HiMoRNA и RNA-Chrom: подтверждение функциональной роли длинных некодирующих РНК в эпигенетической регуляции генов человека с помощью данных РНК-хроматинового интерактома // Acta Naturae. – 2025. – Т. 17, № 2 (65). – С. 98-109. EDN: PRYTHB. Импакт-фактор 2 (JIF) (1.06/0.20).

Ilnitskiy I.S., **Ryabykh G.K.**, Marakulina D.A., Mironov A.A., Medvedeva Y.A. Integration of HiMoRNA and RNA-Chrom: Validation of the Functional Role of Long Non-coding RNAs in the Epigenetic Regulation of Human Genes Using RNA-Chromatin Interactome Data // Acta Naturae. – 2025. – vol. 17, № 2 (65). – pp. 98-109. EDN: EFZYQO. Импакт-фактор 2 (JIF) (1.06/0.20).

ДРУГИЕ СТАТЬИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **Рябых Г.К.**, Жарикова А.А., Ильницкий И.С., Миронов А.А. РНК-хроматиновые взаимодействия. Анализ данных // Вестник Российского фонда фундаментальных исследований. – 2023. – № 3-4 (119-120). – С. 71-76. EDN: MKHYOB (0.48/0.40).

[1] В скобках приведен объем публикации в условных печатных листах и вклад автора в условных печатных листах