МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ

На правах рукописи

Рябых Григорий Кириллович

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

Специальность 1.5.8. Математическая биология, биоинформатика

Диссертация на соискание ученой степени

кандидата биологических наук

Научный руководитель:

доктор биологических наук, профессор

Миронов Андрей Александрович

Москва – 2026

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность и степень разработанности темы исследования

Известно, что значительная часть генома эукариот транскрибируется с образованием большого количества разнообразных РНК, включая мРНК и такие некодирующие РНК (нкРНК), как микроРНК, малые ядерные и малые ядрышковые РНК, энхансерные РНК, длинные нкРНК (днРНК) и другие [1]. Кодирующие и некодирующие транскрипты могут выполнять свои функции не только в цитоплазме, но и в ядре клетки, где активно участвуют в процессах регуляции транскрипции, а также в ремоделировании и поддержании пространственной структуры хроматина [2]. Классическими примерами таких РНК могут служить MALAT1, NEAT1, XIST, TERC, HOTAIR и другие [2].

Механизмы взаимодействия нкРНК с хроматином, модификаторами хроматина или с другими белками изучают с помощью многочисленных разработанных за последнее время экспериментальных подходов (см. обзор [3]). Среди них можно выделить два класса методов: первый определяет взаимодействия конкретной РНК со всеми локусами хроматина (далее «один-против-всех» или OTA) и позволяет установить сайты связывания одной конкретной РНК во всем геноме – карту контактов (RAP [4], CHART-seq [5], ChIRP-seq [6], dChIRP-seq [7], ChOP-seq [8] и CHIRT-seq [9]), в то время как методы второго класса определяют взаимодействия всех РНК со всеми локусами ДНК (далее «все-против-всех» или ATA) и позволяют получить данные обо всех потенциальных РНК-хроматиновых взаимодействиях в клетке (MARGI [10], GRID-seq [11], ChAR-seq [12], iMARGI [13,14], RADICL-seq [15], Red-C [16]).

Некодирующие РНК являются ключевыми регуляторами фундаментальных клеточных процессов. Нарушение их функций ассоциировано с развитием широкого спектра заболеваний, таких как онкологические [17–19], нейродегенеративные [20,21] и аутоиммунные патологии [22,23], что определяет их значимость как для фундаментальной науки, так и для биомедицины. Методы OTA и ATA позволяют картировать физические взаимодействия нкРНК с хроматином [24], однако стремительное накопление данных выявило серьезную методологическую проблему: результаты, полученные в разных лабораториях, обрабатываются с помощью несопоставимых вычислительных конвейеров, что делает их прямое сопоставление и интегративный анализ практически невозможным.

При этом методы OTA часто служат важнейшим инструментом заключительной экспериментальной валидации в исследованиях функциональной роли нкРНК. Например, поиск нкРНК, специфичных для клеток рака толстой кишки, с помощью дифференциального анализа транскриптома позволил отобрать потенциальную хроматин-ассоциированную РНК (хаРНК), функционально связанную с этим типом рака – lincDUSP [25]. А идентификация сайтов связывания lincDUSP с хроматином с помощью ChIRP-seq продемонстрировала связь днРНК с

генами важных сигнальных путей при раке толстой кишки. Таким образом, данные ОТА формируют критическую доказательную базу для многих гипотез.

Фундаментальные основы понимания роли нкРНК в регуляции хроматина заложены в классических работах, посвященных таким молекулам, как XIST [4], HOTAIR [6], U1 [26], NEAT1 и MALAT1 [27]. Для этих и многих других нкРНК с помощью генетических, биохимических и цитологических методов были установлены конкретные биологические функции, связанные с дозовой компенсацией, эпигенетическим сайленсингом, формированием ядерных доменов и контролем альтернативного сплайсинга.

На уровне экспериментальных методов за последнее десятилетие наблюдается значительный прогресс. Разработаны и широко применяются высокопроизводительные технологии для полногеномного картирования РНК-хроматиновых взаимодействий, что подтверждается сотнями исследований, а их результаты в значительном объеме депонированы в публичных репозиториях, что создает обширную эмпирическую базу для анализа.

Параллельно развивается инфраструктура для работы с данными о нкРНК. Существует ряд общедоступных баз данных, таких как NONCODE [28], LNCipedia [29] и RNAInter [30], которые аккумулируют информацию об аннотации, экспрессии и взаимодействиях нкРНК с различными молекулами. Однако эти ресурсы носят общий характер и, как правило, либо не содержат полногеномных данных РНК-хроматиновых взаимодействий, либо включают лишь их незначительную часть. Специализированный ресурс LnChrom [31] был закрыт, что усугубляет проблему отсутствия курируемой базы данных для систематизации знаний в этой области.

Несмотря на очевидные успехи, проведенный анализ литературных данных позволяет выявить существенные методологические пробелы, определяющие степень разработанности темы:

1. Отсутствие стандартизации в обработке данных. Существующие вычислительные конвейеры для анализа экспериментов ОТА и АТА являются узкоспециализированными, что приводит к использованию несопоставимых алгоритмов фильтрации, нормализации и идентификации значимых взаимодействий. Результаты, полученные разными группами, зачастую невозможно непосредственно сравнивать или интегрировать.
2. Дефицит специализированных курируемых ресурсов. В настоящее время отсутствует база данных, которая целенаправленно аккумулировала бы все доступные полногеномные данные РНК-хроматинового интерактома, обработанные по единому стандарту. В общих базах данных этот тип информации представлен фрагментарно.
3. Отсутствие систематического сравнительного анализа. Не проводилось полномасштабного исследования, которое бы количественно оценивало и сопоставляло ключевые характеристики методов ОТА и АТА – такие как специфичность, разрешение, воспроизводимость – на единой совокупности данных.

4. Ограниченность интегративных подходов. Стратегии совместного анализа карт РНК-хроматиновых контактов с другими типами полногеномных данных (эпигенетические метки, архитектура хроматина, профили экспрессии) носят эпизодический характер и не систематизированы в рамках удобных аналитических платформ.

Таким образом, область характеризуется парадоксальной ситуацией: при понимании биологической значимости нкРНК и обилии экспериментальных данных РНК-хроматинового интерактома отсутствуют унифицированные методы обработки и анализа этих данных, стандартизированные ресурсы для их хранения и инструменты для интегративной оценки, что затрудняет формирование целостной картины регуляторных сетей, опосредованных нкРНК.

В данной диссертационной работе предпринята комплексная попытка преодоления этих ограничений путем создания унифицированной аналитической платформы, включающей базу данных RNA-Chrom, стандартизированные протоколы обработки и инструменты для сравнительного анализа. Мы первые провели полномасштабное исследование характеристик и согласованности всего корпуса существующих данных РНК-хроматинового интерактома. Результаты этой работы не только дают количественную оценку точности, полноты и специфичности современных методов, но и закладывают основу для их более эффективного использования в интегративных исследованиях функциональной роли 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, полученные соискателем, представляют существенный научный интерес и важны для развития представлений в области изучения РНК-хроматинового интерактома. Сделанные выводы могут послужить основой для разработки целого комплекса дальнейших молекулярно-биологических экспериментов и биоинформатических исследований, направленных на изучение функций РНК, ассоциированных с хроматином.

Теоретическая значимость исследования обусловлена следующим:

1. Работа вносит вклад в развитие методологии анализа полногеномных данных РНК-хроматинового интерактома, устанавливая количественные критерии оценки их качества.
2. Полученные сравнительные характеристики методов OTA и ATA формируют более четкое понимание их областей применения, ограничений и взаимодополняемости, что важно для планирования будущих экспериментов.
3. Предложенная интегративная схема демонстрирует путь от данных о физических контактах к выдвижению гипотез о функциональных механизмах, обогащая теорию системной регуляции генома нкРНК.

Практическая значимость исследования обусловлена следующим:

1. Созданная база данных RNA-Chrom и сопровождающий ее веб-интерфейс (*https://rnachrom2.bioinf.fbb.msu.ru*) представляют собой готовый к использованию аналитический ресурс для научного сообщества, позволяющий исследователям работать с унифицированными данными без необходимости их самостоятельной сложной обработки.
2. Разработанный программный инструмент Fastq-dupaway (*https://github.com/AndrewSigorskih/fastq-dupaway*) решает конкретную вычислительную проблему обработки больших данных NGS и может быть внедрен в конвейеры других геномных исследований.
3. Установленная интеграция веб-сервисов RNA-Chrom и HiMoRNA создает прецедент и рабочий прототип для мультиомиксного анализа, который можно расширять путем подключения к RNA-Chrom дополнительных веб-сервисов с полногеномными данными (Hi-C, ChIP-seq и другие).
4. Результаты сравнительного анализа данных РНК-хроматинового интерактома и предложенная стратегия достоверного анализа интерактома позволяют повысить надежность биоинформатического анализа и интерпретации данных РНК-хроматинового интерактома, что особенно важно для выявления функционально значимых связей и интеграции с другими омиксными данными.

5. В отличие от традиционного сценария, при котором биологическая интерпретация данных РНК-хроматиновых взаимодействий требует ручного сбора разрозненных экспериментов, их повторной обработки и последующего сопоставления с геномными и эпигенетическими аннотациями, RNA-Chrom предоставляет единое пространство для такого анализа. Это позволяет быстрее переходить от наблюдаемого контакта РНК с хроматином к формированию списка генов-кандидатов, регуляторных локусов и проверяемых функциональных 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.

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».

Публикации по теме исследования

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, – pp. 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).

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

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 (2025). 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 публикации по теме исследования

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).

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.

Глава 1. ОБЗОР ЛИТЕРАТУРЫ[2]

С каждым годом появляется все больше доказательств того, что некодирующие РНК (нкРНК) у животных и растений участвуют в широком спектре биологических процессов, таких как регуляция дифференцировки клеток и экспрессии генов, ремоделирование хроматина, поддержание структуры хроматина, сплайсинг, процессинг РНК, образование биомолекулярных конденсатов и другие. Нарушение регуляторных путей, опосредованных нкРНК, ассоциировано с развитием различных заболеваний, что подчеркивает важность понимания механизмов их действия [32]. Значительная часть функций нкРНК реализуется в ядре клетки, что обусловливает необходимость детального изучения РНК-хроматинового интерактома.

Молекулы РНК взаимодействуют со множеством белков, хроматином и другими РНК. Экспериментальные методы, позволяющие идентифицировать локусы ДНК, с которыми контактируют нкРНК, можно разделить на две группы: «один-против-всех» («one-to-all» или OTA) и «все-против-всех» («all-to-all» или ATA). Первая группа методов (RAP [4], CHART-seq [5], ChIRP-seq [6], dChIRP-seq [7], ChOP-seq [8], CHIRT-seq [9]) определяет контакты заранее известной РНК с хроматином, а вторая группа методов (MARGI [10], GRID-seq [11], ChAR-seq [12,33], iMARGI [13], RADICL-seq [15], Red-C [16]) направлена на определение всех возможных контактов РНК-ДНК в клетке.

1.1. Эксперименты OTA: взаимодействия конкретных РНК с хроматином

Группа методов OTA (RAP, CHART-seq, ChIRP-seq, dChIRP-seq, ChOP-seq, CHIRT-seq) во многом однородна. Взаимодействия РНК-ДНК фиксируются комбинацией или единственным сшивающим агентом, например, формальдегидом, дисукцинимидилглутаратом, глутаральдегидом или ультрафиолетовым излучением. Хроматин фрагментируется ДНКазой и/или ультразвуком. Сшитые комплексы (РНК-геномная ДНК) захватываются биотинилированными ДНК-олигонуклеотидами, антисмысловыми по отношению к целевой РНК, и отбираются с помощью стрептавидиновых шариков. Далее геномную ДНК обычно элюируют в присутствии РНКазы H и очищают от белков с помощью протеазы K. В конце секвенируют фрагменты геномной ДНК, ранее сшитые с целевой РНК, и анализируют полученные данные (рис. 1).

[2] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **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, – pp. baad025. EDN: YEKQIZ. Импакт-фактор 3,6 (JIF) (1.02/0.40). 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 (2025). EDN: VBEHEL. Импакт-фактор 3,9 (JIF) (0.88/0.20). **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). **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).

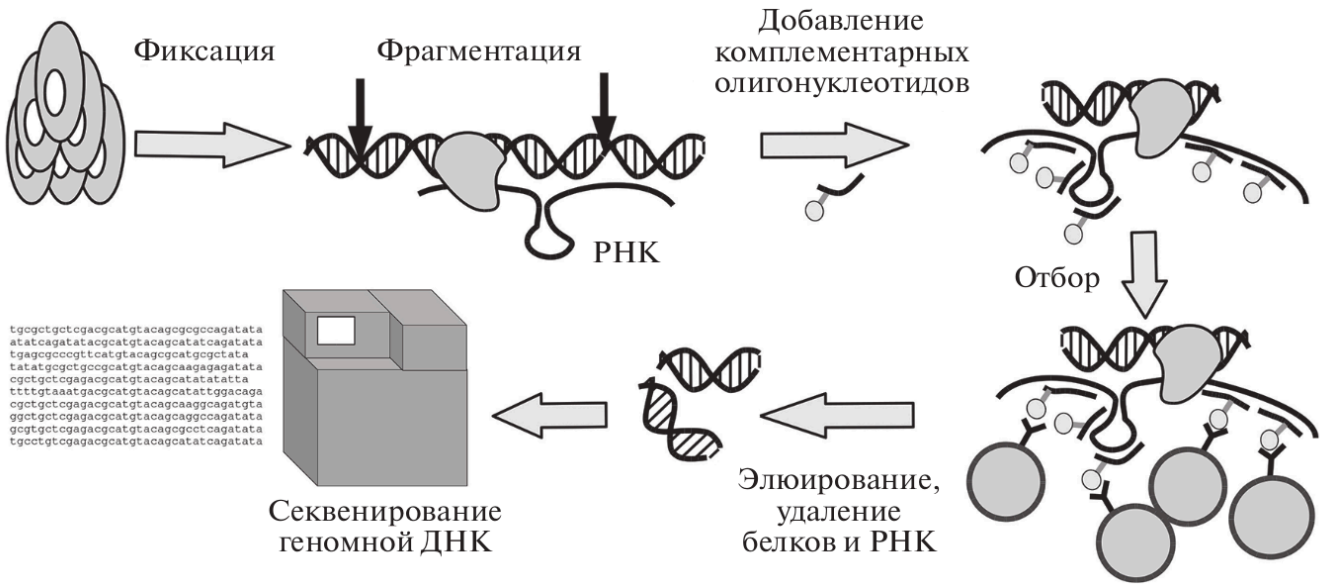


Рисунок 1. Общая схема протоколов ОТА.

Доступные участки целевой РНК, как правило, заранее неизвестны, поэтому подбор антисмысловых зондов является одной из главных проблем методов ОТА. Например, для решения этой проблемы в методе CHART-seq предлагается использовать малое количество коротких зондов, которые в комплексе с РНК обладают высокой чувствительностью к РНКазе H [5]. Этот отбор позволяет выявить наиболее доступные для гибридизации зондов области РНК. Напротив, в методе ChIRP-seq предлагается использовать выстилающие наборы зондов, покрывающие РНК без пересечений (в случае HOTAIR зонды покрыли около 50% ее последовательности) [6]. Исключением стала многокопийная РНК TERRA, содержащая повторяющиеся субтеломерные последовательности, что делало затруднительным использование выстилающих зондов. Метод CHIRT-seq, объединивший и оптимизировавший методы ChIRP-seq и CHART-seq, решил данную проблему, использовав для специфического захвата TERRA один олигонуклеотид, комплементарный теломерному повтору [9]. Отметим, что все методы, борясь с неспецифическими взаимодействиями зондов, отбраковывают зонды, которые оказались комплементарными другим РНК или участкам геномной ДНК.

Если геномный локус, с которым контактирует целевая РНК, известен заранее, то поиск ее контактов можно ограничить этим локусом, амплифицировав ДНК только этого локуса (метод ChOP-qPCR), как в случае управляющей импринтингом нкРНК Kcnq1ot1 [34]. Аналогичным образом с помощью CHART-qPCR проанализированы РНК STEEL [35], а с помощью ChIRP-qPCR – FMR1 [36].

Домены РНК, ответственные за взаимодействие с хроматином, можно выявлять с использованием делеционного анализа (в ряде работ из XIST удаляли повторы «B» [37], «E» [38] и «A» [4,39], а из ANRIL – экзон 8 [40]) или домен-специфичного метода dChIRP-seq, как в случае roX1 [7].

Популярность экспериментальных протоколов менялась со временем. В настоящее время в большинстве работ используют протокол ChIRP-seq (рис. 2).

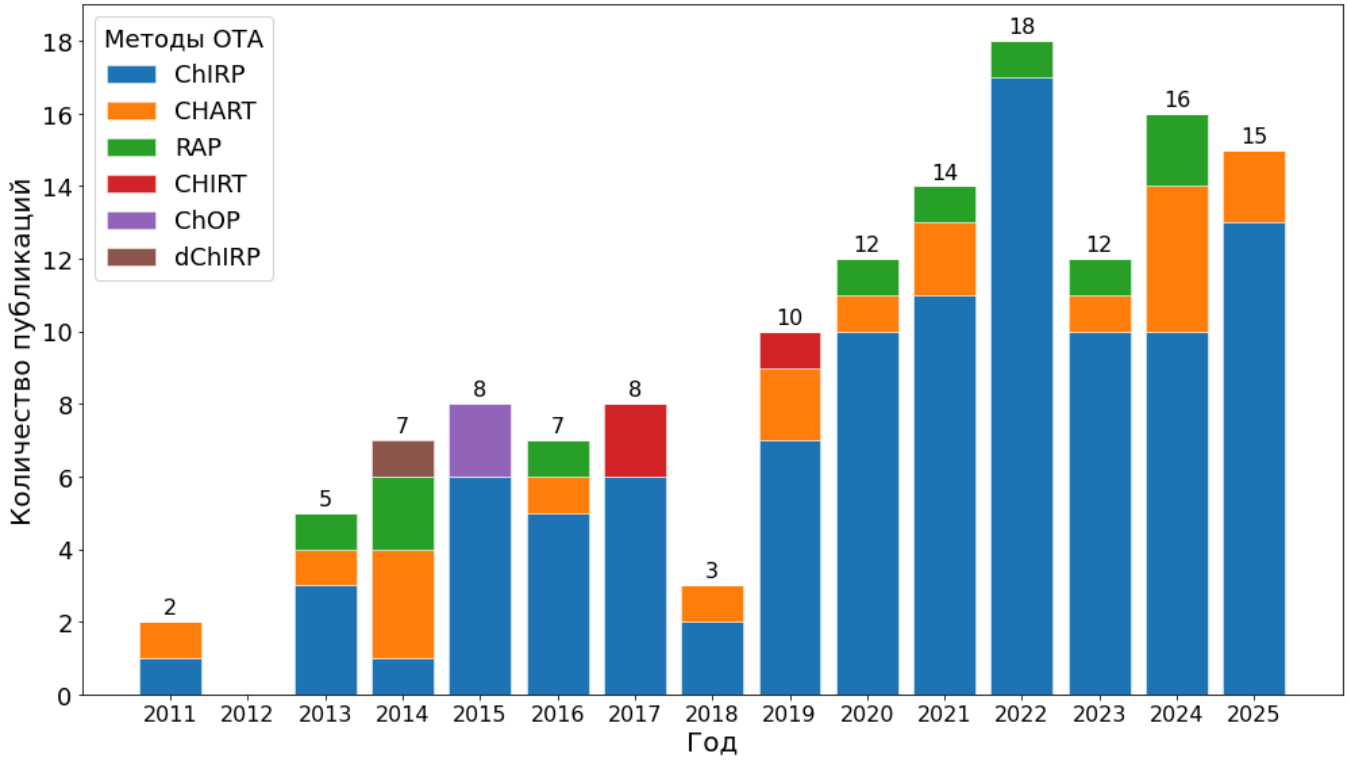


Рисунок 2. Использование методов ОТА в разные годы.

Основными задачами обработки данных ОТА являются: во-первых, выделение участков генома, обогащенных контактами данной РНК с хроматином (далее «пики»), во-вторых, фильтрация специфичных пиков от артефактных.

Типичный протокол обработки данных ОТА состоит из следующих шагов: удаление ПЦР-дубликатов и картирование прочтений на геном; выделение участков со статистически значимым превышением покрытия над покрытием фона (определение пиков) и/или нормализация на фон; поиск и фильтрация пиков.

В связи с особенностью решаемой биологической задачи в некоторых работах выполняли аллель-специфичное картирование [37,41]. Чаще всего при не аллель-специфичном выравнивании оставляли уникально картированные на геном прочтения.

Для учета артефактных сигналов используют сравнение с фоновыми данными («input»), а также фильтрацию с помощью данных, полученных в контрольных экспериментах: геномную ДНК элюируют без добавления РНКазы H [9] или образец перед выделением обрабатывают РНКазой [42]; используют зонды, не имеющие отношения к целевой РНК, например, комплементарные зондам основного эксперимента [8] или антисмысловые зонды к РНК, которые не транскрибируются в данной клеточной культуре, например, к РНК из другого организма – мРНК гена LacZ [43]. В качестве контроля используют также клеточные линии, в которых

изучаемая РНК либо не экспрессировалась [44], либо ее ген был «нокаутирован» [45].

В большинстве работ пики определяли с использованием программы MACS2 [46], первоначально разработанной для поиска пиков в данных ChIP-seq, однако такие программы как SICER [47], HOMER [48] и SPP [49] также использовали. В зависимости от целевой РНК количество обнаруженных пиков варьировало от сотен до десятков тысяч. Любопытно, что при использовании в качестве фона контрольных данных в некоторых работах [5,50] были получены пики, отличные от пиков, полученных при нормировании на «input», в то время как в [51] показана слабая зависимость между полученными пиками и примененной фоновой моделью.

Для фильтрации случайного шума также применяют биологические реплики с четными и нечетными зондами. Основной протокол обработки таких данных состоит из генерации консенсусного трека для четных и нечетных зондов и фильтрации пиков по порогам обогащения, покрытия и корреляции между четными и нечетными данными (подробнее этот подход описан в [6]).

Другой подход к использованию контрольных данных для исключения неспецифических пиков состоит в том, чтобы отфильтровать те пики, которые попадают в области, обогащенные пиками из контрольных экспериментов (со смысловыми зондами, зондами против LacZ, с предобработкой РНКазой) [52,53].

Определение пиков проводят не всегда – в некоторых случаях треки покрытия сглаживают с помощью SPP, после чего нормируют на фон и/или треки, полученные из других экспериментов или с другого аллеля [37,54].

Отметим, что РНК может взаимодействовать с геномной ДНК несколькими способами. Напрямую – формируя такие структуры, как РНК-ДНК-триплексы (при этом образуются хугстиновские связи между РНК и ДНК) и R-петли (уотсон-криковское спаривание РНК-ДНК), или через белкового посредника [3].

Поиск прямого взаимодействия некодирующих РНК с ДНК можно проводить с использованием нескольких подходов как независимо, так и в комбинации. Простейший подход ищет потенциальные комплементарные взаимодействия с ДНК в местах, ассоциированных с пиками РНК-хроматиновых контактов. Однако применение, например, алгоритма EMBOSS Water [55] не выявило значительной комплементарности между транскриптами Paupar и Dali и их соответствующими контактами [50,56]. С помощью программы Triplexator [57] показано, что РНК Dali не образует триплексные структуры РНК-ДНК [50]. Таким образом, РНК Paupar и Dali, вероятно, не связываются с ДНК напрямую. С другой стороны, с помощью программы Triplexator обнаружены потенциальные триплексы у MEG3 и HOTAIR [8], а «Triplex Domain Finder» [58] – у EPR [59] и ANRIL [40]. Здесь следует отметить, что программы поиска потенциальных триплексов имеют невысокое качество предсказания [60].

Для обнаружения потенциального белкового партнера проводят *de novo* анализ ДНК-мотивов по данным связывания изучаемой РНК с хроматином методами MEME, DREME, MEME-ChIP, MEME SUITE [61] и HOMER [48]. Далее обогащенные мотивы можно сравнивать с известными мотивами ДНК-связывающих белков, например, с помощью TOMTOM [62]. В нкРНК LncHSC-2 [43] обнаружен мотив связывания изоформы гемопоэтического фактора транскрипции E2A. Показано пересечение пиков связывания в данных ChIP-seq E2A и ChIRP-seq LncHSC-2. Сделанные наблюдения, а также результаты нокдауна гена E2A, показали, что LncHSC-2 отвечает за связывание E2A на некоторых сайтах-мишенях.

Поиск ДНК-мотивов может дать неожиданный результат. Например, в нкРНК MEG3 [8], HOTAIR, roX2 [6], TERRA [9], ANRIL [40], linc-ASEN [63] и Tug1 [64] идентифицированы GA-богатые ДНК-мотивы, которые, как известно, могут образовывать триплексные структуры [65–67]. Это указывает на потенциально важную роль GA-богатых мотивов в полногеномном таргетинге длинных нкРНК. И если взглянуть на MEG3, то выстраивается цельная картина: за нацеливание на хроматин отвечает одна часть РНК, а за привлечение EZH2 (часть комплекса PRC2), с помощью которой происходит модификация хроматина, отвечает совсем другая часть РНК MEG3 [8].

Полученные пики можно связать с регуляторными областями генов и другими геномными областями с помощью программ GREAT [68], CEAS [69] (lincDUSP [25], TERRA [9,51]). Гены, ассоциированные с контактами, можно охарактеризовать по их принадлежности к тому или иному процессу, выполнив анализ онтологий генов (GO-анализ, программы Panther [70], DAVID [71], QuickGO [72], Enrichr [73]). В дальнейшем анализе данные ОТА можно сравнивать с другими экспериментальными данными (например, с GRID-seq, ChIP-seq, RNA-seq, Hi-C и др.) и геномными аннотациями (SINE, LINE, CpG-острова, цвета хроматина, энхансеры и т.д.). Например, сравнение с данными RNA-seq после нокдауна или сверхэкспрессии изучаемой РНК (ANRIL [40], lnc-NR2F1 [42], Bloodlinc [52], EPR [59], LED [74]) позволяет определить первичные мишени среди генов, чья экспрессия значительно изменялась при манипулировании изучаемой РНК. Другой вариант последующего анализа: определение колокализации сайтов связывания нкРНК с сайтами связывания белков или белковых комплексов, способствующих активации транскрипции (например, MSL, Mediator, PGC-1α, факторы транскрипции bHLH, INO80 и др.), метками активного хроматина (H3K4me3, H3K4me1, H3K27ac, H3K9ac, РНК-полимераза II и др.), а также ассоциация сайтов связывания нкРНК с репрессирующими гистоновыми метками. Такие стратегии позволили показать, что нкРНК roX [5,6], Dali [50], Hotchon [44], Evx1as [75], Tug1 [64], lnc-NR2F1 [42], HAND2-AS1 [53], ERV-9 [76], Charme [77] и LED [74] являются активирующими. С другой стороны, обогащение репрессирующими гистоновыми метками, колокализация с сайтами связывания PRC1/PRC2 (или их субъединицами), BAF, P-TEFb и другими белковыми

партнерами свидетельствуют о том, что XIST [4,37–39,41,54,78], HOTAIR [6], MEG3 [8], Linc-ASEN [63], 7SK [79] относятся к репрессирующим РНК. Однако это вовсе не значит, что РНК может быть только активирующей или только репрессирующей. Например, в зависимости от образования комплекса с TrxG или с PRC2 нкРНК SRA будет способствовать установлению либо активирующих (H3K4me3), либо репрессирующих (H3K27me3) гистоновых меток в соответствующем геномном локусе и в зависимости от этого будет выполнять либо активирующую, либо репрессирующую функцию [80]. Paupar [56], ANRIL [40], TERRA [9,81] также ведут себя двойственно.

Данные ОТА часто используются для поиска различий между состояниями: между ортологами roX1 и roX2 нескольких видов *Drosophila* [82], до и после теплового шока (нкРНК B2 мыши) [83], нокдауна белкового партнера РНК HOTAIR, EZH2 [6], между разными временными точками дифференциации клеток (XIST [4], Evx1as [75], PAR-TERRA [51], ERV-9 [76]), до и после дополнительной обработки клеток ингибитором элонгации транскрипции флавопиридолом (MALAT1, U1 [26]) или ингибитором бромодомена JQ1 (7SK [79]).

В отличие от белок-кодирующих генов, гены длинных нкРНК часто менее консервативны на уровне последовательности [84,85], однако другие характеристики нкРНК: синтеничные отношения с соседними генами, сходство коротких фрагментов последовательностей и вторичная структура – часто остаются консервативными (roX1/2 [82]). Важно отметить, что из этого правила есть исключения, например, нкРНК Dali [50], Paupar [56], Firre [86], Bloodlinc [52], DINO [87], Charme [77], HAND2-AS1 [53], lnc-NR2F1 [42], EPR [59], Eprn [88], 7SK [79], MALAT1, NEAT1 [27] являются консервативными, в то время как нкРНК DACOR1 [89] – неконсервативная.

1.1.1. Биологические примеры

Дозовая компенсация Х-хромосомы у *Drosophila*, roX1/2. Рибонуклеопротеин MSL (male-specific lethal), в состав которого входят нкРНК roX2 и roX1, обеспечивает эквивалентную экспрессию генов, расположенных на Х-хромосоме, между мужскими (XY) и женскими (XX) особями *Drosophila* за счет увеличения транскрипции с единственной мужской Х-хромосомы примерно в 2 раза [90]. Полученные с помощью методов CHART-seq и ChIRP-seq в 2011 году полногеномные карты контактов roX2 с хроматином подтвердили, что roX2 в основном локализуется на Х-хромосоме и связывается в тех же участках хроматина, как и белок MSL3 (субъединица MSL-комплекса) [5,6]. Анализ ДНК-мотивов roX2 в данных ChIRP-seq выявил мотив, почти идентичный мотиву MSL [6]. Плотность контактов roX2 увеличивается с 5′- до 3′-конца каждого гена на Х-хромосоме, что соответствует представлению о том, что комплекс roX-MSL усиливает элонгацию, а не инициацию транскрипции [91].

Модификация метода ChIRP-seq (dChIRP-seq – «domain-specific chromatin isolation by RNA

purification» [7]) позволила уточнить, какие именно домены РНК roX1 участвуют в MSL- и CLAMP-опосредованном взаимодействии с хроматином. Карты связывания другой нкРНК, roX1 [92], показали схожие сайты связывания с хроматином, что подтверждает ранее известные данные о функциональной избыточности и взаимозаменяемости roX1 и roX2. Сравнение dChIRP-seq roX1 и ChIRP-seq roX2 с данными Hi-C позволили предложить модель конформации X-хромосомы, в которой локус roX2 и некоторые CES-сайты (специфические сайты на X-хромосоме, с которыми взаимодействуют обе РНК и MSL-белки) кластеризуются в территорию дозовой компенсации, в то время как roX1-локус находится вне этой территории [7].

Поиск ортологов roX1 и roX2 выявил 19 и 28 новых ортологов соответственно, у 35 видов рода Drosophila, эволюционно разошедшихся около 40 миллионов лет назад [82]. Сравнительный анализ данных ChIRP-seq по взаимодействию ортологов roX1 и roX2 с хроматином у *D. melanogaster, D. willistoni, D. virilis и D. busckii* показал, что сайты связывания roX1/2 ассоциированы с X-хромосомой и в большинстве своем эволюционно динамичны. Обнаружено, что новые сайты могут возникать из интронных полипиримидиновых трактов. Более детальный анализ показал, что для сайтов связывания roX1/2 важна близость, а не точное расположение относительно генов. В случае слияния аутосомы с X-хромосомой и образования нео-X-хромосомы (как у *D. willistoni*) новые X-сцепленные гены также могут подвергаться дозовой компенсации.

Дозовая компенсация X-хромосомы у млекопитающих, XIST. Инактивация X-хромосомы («X-chromosome inactivation», или XCI) – это механизм регулирования экспрессии X-сцепленных генов у женских особей млекопитающих, позволяющий им компенсировать удвоенное количество половых хромосом [93–95]. В процессе XCI неактивная X-хромосома (Xi) подвергается реконфигурации в уникальную структуру – тельце Барра. Центральный игрок этого процесса – длинная нкРНК XIST, которая распространяется вдоль Xi, привлекая факторы сайленсинга и индуцируя глобальное подавление топологически ассоциированных доменов (ТАД) [37,96–98]. Она может выполнять функцию каркаса в сборке репрессивных комплексов, таких как Polycomb (PRC1/PRC2), которые моноубиквитинируют гистон H2A по лизину 119 (H2AK119ub) и триметилируют гистон H3 по лизину 27 (H3K27me3) [98–101], или формировать репрессивный компартмент, например, взаимодействуя непосредственно с рецептором ламина B (LBR) и привлекая Xi в ядерную ламину [39].

В 2013 году были опубликованы две работы, в которых методы ОТА (RAP [4] и CHART-seq [41]) впервые применили для изучения распространения РНК XIST по X-хромосоме. Экспериментальные данные получены для нескольких временных точек: перед XCI (эмбриональные стволовые клетки мыши), после завершения XCI (терминально дифференцированные фибробласты) и три/две промежуточные стадии соответственно. Как и ожидалось, чем ближе клеточная линия приближалась к терминально дифференциированной

стадии, тем больше становилось XIST-обогащенных сегментов на X-хромосоме, тем больше была плотность контактов на этих сегментах по сравнению с плотностью на аутосомах. Проанализировав корреляции между профилями распространения XIST вдоль Xi (по стадиям развития) и геномными и хроматиновыми аннотациями, пришли к выводу, что во время *de novo* XCI XIST сначала распространяется по богатым генами регионам («ранние» домены). Любопытно, что XIST с большей вероятностью нацелена на гены в областях активного хроматина, а затем распространяется в области, более бедные генами («поздние» домены), причем XIST и PRC2 совместно мигрируют в новые регионы внутри «ранних» доменов Xi. Показано, что именно конформация хромосомы, а не аффинность «ранних» доменов играет важную роль в определении ранних сайтов локализации XIST на Xi при *de novo* XCI [4]. Чтобы выяснить, как распространяется XIST по Xi после завершения XCI, проанализировали распространение XIST по Xi в клеточной линии MEF (эмбриональные фибробласты мыши) через 1, 3 и 8 ч после удаления XIST с помощью заблокированных нуклеиновых кислот LNA-C1 и LNA-4978. Оказалось, что во время поддержания Xi в соматических клетках XIST распространяется как на «ранние», так и на «поздние» домены одновременно [41]. Таким образом, РНК XIST распространяется по X-хромосоме по двум разным механизмам в зависимости от стадии: de novo XCI или поддержания XCI.

РНК XIST имеет доменную структуру из консервативных повторяющихся мотивов (повторы A–F) [102], роль которых в распространении XIST известна не до конца. Для выяснения роли этих доменов применили делеционный анализ. После удаления повтора E паттерны распространения данной РНК по Xi на 3-й день дифференцировки эмбриональных стволовых клеток (embryonic stem cells, ESC) сильно коррелировали с паттернами XIST дикого типа, однако на 7-й день микроскопическими методами (FISH и 3D-SIM) выявили значительную дисперсию XIST-ΔE внутри ядра по сравнению с диким типом. Это показывает, что повтор E не влияет на начальное распространение XIST по Xi, но необходим на более поздних стадиях XCI [38].

С помощью методов ChIP-seq и CHART-seq показано [37], что удаление повтора B в клеточной линии MEF (процесс XCI завершен), необходимого XIST для взаимодействия с белком HNRNPK и привлечения через него PRC1 и PRC2, вызывает истощение покрытия XIST и почти полную потерю меток H2AK119ub и H3K27me3 на Xi. Причем дефект распространения XIST связан с потерей PRC1 и PRC2, что подтверждают данные независимых двойных нокаутов по компонентам PRC1 и PRC2. Также показано, что повтор A, с которым взаимодействует PRC2 [100], необходим XIST на начальных этапах XCI для подавления экспрессии генов в «ранних» доменах и изменения структуры хромосомы [4]. Так, с повтором A напрямую взаимодействует LBR [98,103] при de novo XCI, что способствует привлечению покрытой XIST ДНК в ядерную ламину и распространению XIST на другие участки X-хромосомы, которые физически стали ближе к локусу

транскрипции XIST [39].

В дальнейшем была опубликована модель пошагового сворачивания Xi. При *de novo* XCI XIST сначала распространяется по богатым генами «ранним» доменам (которые сильно коррелируют с A-компартментами укладки хроматина), а далее по «поздним» доменам (B-компартменты). В этот момент A/B-компартменты XIST-зависимо ремоделируются в S1/S2-компартменты (специфичные только для Xi). Далее архитектурный белок SMCHD1 ослабляет (но не удаляет полностью) структуру ТАДов Xi, объединяет S1/S2 и формирует два мегадомена, разделенных Dxz4-локусом. Удаление SMCHD1 в клетках с завершенной XCI вызывает повторное появление S1/S2-компартментов и обогащение XIST в S1 (но не в S2), что приводит к аберрантному обогащению H3K27me3 на Xi, сегментарной эрозии гетерохроматина и неудачному подавлению экспрессии генов. Более того, истощение XIST, HNRNPK или PRC1 препятствует S1/S2-компартментализации хроматина [54,78].

TERRA. На теломерах активно синтезируется гетерогенная популяция длинных нкРНК, получивших название TERRA. Известно, что TERRA является неотъемлемой частью архитектуры теломер. Однако цитологические исследования показали, что только около половины обнаруживаемых транскриптов TERRA локализованы в теломерах. С помощью CHIRT-seq удалось показать, что TERRA связывает хроматиновые мишени по всему геному, в том числе на теломерах [9,81].

С помощью CHIRT-seq на женских клетках ESC, MEF и двух промежуточных стадиях обнаружена субпопуляция TERRA, транскрибируемая из псевдоаутосомных регионов (pseudoautosomal regions, PAR) X-хромосом – PAR-TERRA [51]. Оказалось, что РНК PAR-TERRA необходима для гомологичного спаривания X-хромосом (X–X) и собственно для процесса XCI, который управляется Xic (X-inactivation center) через нкРНК XITE, TSIX и XIST. Во время раннего развития счетный механизм определяет количество X-хромосом и инициирует XCI в клетках с двумя или более X-хромосомами. XIST инициирует сайленсинг всей случайно выбранной хромосомы Xi, тогда как TSIX противодействует XCI на оставшейся активной Xa. Переход от двухаллельной к моноаллельной экспрессии гена TSIX фиксирует выбор аллеля. Переход TSIX в моноаллельное состояние происходит во время гомологичного спаривания между двумя аллелями Xic. Показано, что PAR-TERRA взаимодействует с Xic в течение всего периода спаривания Xic-Xic, а взаимодействия Xic-Xic, PAR-PAR и Xic-PAR регулируются PAR-TERRA.

Уточнение молекулярно-биологического механизма действия NEAT1 и MALAT1. Известно, что высококонсервативная нкРНК MALAT1 локализуется в ядерных спеклах [104], взаимодействует с серин/аргинин-богатыми белками, участвующими в сплайсинге, и может регулировать экспрессию генов и альтернативный сплайсинг [105]. Для более детального изучения молекулярной функции MALAT1 в 2014 году с помощью методов RAP–RNA и RAP–DNA

получены полногеномные карты контактов MALAT1 с другими РНК и с хроматином соответственно [26]. Локализация MALAT1 на хроматине оказалась сильно связанной с активно транскрибируемыми генами, причем наибольшее обогащение MALAT1 наблюдалось примерно на 500 нуклеотидов ниже аннотированного сигнала полиаденилирования. Эксперимент RAP–DNA при подавленной элонгации транскрипции показал, что обогащение MALAT1 в ранее активных генах значительно снижено, т.е. локализация MALAT1 на хроматине зависит от транскрипции его ДНК-мишеней. Кроме того, локализация MALAT1 на хроматине зависит от экзон-интронной структуры генов: мультиэкзонные и альтернативно сплайсируемые гены наиболее обогащены контактами MALAT1. Одноэкзонные гены негистоновых белков менее обогащены, а гены гистонов редко контактируют с MALAT1.

Длинная нкРНК NEAT1 входит в состав ядерных параспеклов [106] и участвует в регуляции транскрипции [107,108]. Она, как и MALAT1, взаимодействует с сотнями активных генов, причем со многими эти РНК взаимодействуют совместно [27], однако их паттерны связывания различаются: NEAT1 связывает сайты инициации и терминации транскрипции генов, а MALAT1 – тела генов и сайты терминации транскрипции. Дополнительные эксперименты CHART-seq с подавлением элонгации транскрипции показали, что NEAT1 более заметно реагирует на изменение уровня транскрипции его ДНК-мишеней, чем MALAT1 – ингибирование элонгации транскрипции ведет к отчетливому смещению локализации NEAT1 в направлении сайтов инициации транскрипции [27].

U1. В 2010 году заметили, что сплайсосомная РНК U1 связана с другим процессом – защитой транскриптома клетки. U1 предотвращает преждевременное расщепление и полиаденилирование (premature cleavage and polyadenylation, PCPA) зарождающихся транскриптов [109]. В 2014 году с помощью методов RAP–RNA и RAP–DNA получены полногеномные карты контактов U1 с другими РНК и с хроматином соответственно [26]. Оказалось, что локализация U1 на хроматине ассоциирована с транскрипцией ДНК-мишеней, при этом 5′- и 3′-концы активных генов обогащены контактами U1 больше, чем их тела. После подавления элонгации транскрипции локализация U1 изменилась: контакты U1 на 3′-концах генов в значительной степени исчезли, но на 5′-концах генов обогащение U1 сохранялось. Предполагается, что U1 локализуется на хроматине двумя способами: 1) U1 связывается с зарождающимися РНК, защищая транскриптом от PCPA, что отражается в обогащении 3′-концов генов контактами U1; 2) U1 связывается с 5′-концами генов по механизму, который не зависит от элонгации транскрипции, и поэтому может участвовать в регуляции инициации транскрипции [26].

Ингибирование транскрипции и истощение SNRNP70 (компонента малого ядерного рибонуклеопротеина U1) на 90% приводит к снижению связывания MALAT1 с активными генами, что подтверждает гипотезу о том, что связанный с активной РНК-полимеразой II малый ядерный

рибонуклеопротеин U1 прямо влияет на локализацию некоторых длинных нкРНК на хроматине [110].

HOTAIR. Длинная межгенная нкРНК HOTAIR транскрибируется из HOXC-локуса, может связывать комплекс PRC2 и влиять на степень его обогащения на генах-мишенях по всему геному [111–113]. Чтобы понять, как HOTAIR направляет PRC2 к генам-мишеням, в 2011 году с помощью ChIRP-seq получены карты контактов HOTAIR с хроматином на клетках рака молочной железы [6]. Обнаружены 832 сайта связывания HOTAIR с хроматином, которые располагались в основном в энхансерах и интронах, а также ассоциировались с сайтами связывания субъединиц PRC2 (EZH2 и SUZ12). Далее провели ChIRP-seq HOTAIR на клетках MDA-231, истощенных по EZH2-субъединице, которая непосредственно связывается с HOTAIR. Оказалось, что паттерн локализации HOTAIR на хроматине в значительной степени сохранился, что указывает на способность HOTAIR связываться с хроматином без PRC2. В совокупности эти результаты подтверждают, что HOTAIR может активно привлекать комплексы, модифицирующие хроматин.

lnc-NR2F1. Ген длинной нкРНК lnc-NR2F1 (или NR2F1-AS1) располагается в геноме рядом с белок-кодирующим геном NR2F1, который кодирует фактор транскрипции, участвующий в нейрогенезе. Последовательность lnc-NR2F1 консервативна у мыши и человека, что не характерно для длинных нкРНК. Данные ChIRP-seq для lnc-NR2F1 мыши анализировали по стандартной схеме [42]: GREAT-анализ выявил ассоциации пиков с конкретными генами, а GO-анализ отобранных генов показал обогащение этого набора генами, вовлеченными в нейронные процессы. Сравнение карт контактов lnc-NR2F1 с общедоступными данными ChIP-seq показало, что пики lnc-NR2F1 колокализуются с энхансерными хроматиновыми метками H3K27ac и H3K4me3. Согласно анализу ДНК-мотивов сайтов связывания lnc-NR2F1, подобные мотивы найдены у некоторых факторов транскрипции, связанных с нейрогенезом (Ngn2 и Ascl1).

Так как lnc-NR2F1 человека имеет несколько изоформ, предположили, что разные домены lnc-NR2F1 могут иметь различную локализацию на хроматине. С помощью ChIRP-seq картированы контакты двух изоформ lnc-NR2F1: короткой и длинной. Число сайтов связывания длинной изоформы оказалось примерно в 10 раз больше, чем короткой: 4404 против 415, причем эти множества сайтов слабо перекрывались. Это, а также GO-анализ генов-мишеней, согласуется с гипотезой о том, что длинная изоформа участвует в нейрогенезе [42].

Консервативность lnc-NR2F1, сходство множеств генов человека и мыши, ассоциированных с контактами этой РНК, а также соответствующих мотивов ДНК указывают на возможные консервативные функции lnc-NR2F1 у мыши и человека [42].

Firre. Длинная межгенная нкРНК Firre локализуется на X-хромосоме и взаимодействует с ядерным матричным фактором hnRNPU через повторяющиеся последовательности. Заметное обогащение локуса Firre сайтами связывания CTCF и меткой H3K4me3, а также сниженное число

сайтов связывания XIST, LaminB1 и метки H3K27me3 в этом локусе свидетельствуют о том, что этот ген избегает инактивации Х-хромосомы [86]. Используя метод RAP, определили сайты связывания Firre с хроматином и обнаружили значительно обогащенные пики на некоторых хромосомах, однако преобладал сигнал в области ~5 миллионов п.н. вокруг сайта транскрипции. Дополнительные FISH-эксперименты подтвердили фокусную локализацию Firre в ее собственном локусе. По-видимому, Firre может служить локальным организующим фактором для топологического сближения транс-сайтов и геномного локуса.

SRA. Колокализация активной гистоновой метки H3K4me3 и репрессирующей метки H3K27me3 определяет бивалентные сайты, которые маркируют гены, готовые к индукции экспрессии, в том числе гены, участвующие в дифференцировке плюрипотентных стволовых клеток. За метилирование H3K4 и H3K27 ответственны белки группы Trithorax (TrxG) и PRC2 соответственно. Длинная нкРНК SRA (Steroid Receptor RNA Activator) может связываться как с TrxG, так и с PRC2. Используя метод ChIRP-seq, в геноме плюрипотентных стволовых клеток человека [80] идентифицировали 7899 сайтов связывания SRA, из которых 1570 несли метку H3K4me3, 735 – H3K27me3 и 894 сайта имели сигнатуру бивалентного домена (8% от всех картированных бивалентных доменов). Отсюда следует, что в зависимости от сайта SRA может доставлять одну или обе эти модификации. Пересекая сайты связывания SRA с сайтами p68 (SRA-ассоциированная РНК-хеликаза), CTCF, NANOG, предположили, что CTCF способствует установлению бивалентных состояний в местах, где также присутствует SRA: NANOG привлекает на хроматин SRA и связанные с ним комплексы TrxG и PRC2, а p68 облегчает SRA-опосредованное метилирование H3K4.

Выполнив ChIRP-seq на клеточной линии K562 эритробластов человека [114], обнаружили 2790 сайтов связывания SRA по всему геному, большинство из которых располагались на расстоянии до 50 тысяч п.н. от сайта инициации транскрипции. Анализ транскриптома показывает, что SRA способствует экспрессии эритроидассоциированных генов и при этом действует как репрессор генов, ассоциированных с лейкоцитами. Кроме того, подавление SRA с помощью малых шпилечных РНК снижает экспрессию эритроидспецифических маркеров TFRC и GYPA и подавляет экспрессию глобиновых генов. Таким образом показано, что нкРНК SRA способствует транскрипции эритроидных генов.

DACOR1. Для эпигенома опухолевых клеток характерна глобальная потеря метилирования ДНК, которое вносит вклад в геномную нестабильность и аберрантную экспрессию генов. Длинная нкРНК DACOR1 взаимодействует с поддерживающей ДНК-метилтрансферазой DNMT1 в клеточной линии HCT116 (рак толстой кишки), экспрессируется на высоком уровне в нормальных клетках здоровой толстой кишки и на сниженном – в клонах раковых клеток толстой кишки. С помощью ChIRP-seq найдено 338 сайтов связывания DACOR1 с хроматином: 161 сайт

располагается около 150 аннотированных генов и 177 в межгенных областях. При сравнении сайтов связывания DACOR1, расположенных рядом с аннотированными генами, с дифференциально метилированными областями (differentially methylated regions, DMR), определенными в когорте опухолей толстой кишки и в соответствующих нормальных тканях, выявили перекрывание 31 сайта с DMR. Эти результаты показывают, что, взаимодействуя как с DNMT1, так и с хроматином, DACOR1 потенциально привлекает и/или собирает DNMT1-комплекс в определенных геномных локусах для регуляции эпигенетических модификаций и, следовательно, экспрессии конкретных генов и путей [89].

p53-связанные РНК: LED, DINO, SLEAR. Белок p53 связывается с энхансерами, чтобы регулировать ключевые гены-мишени. Картирование 6270 p53-регулируемых энхансеров показало, что большинство из них не содержит сайтов связывания p53. Поскольку длинные нкРНК являются важными регуляторами динамики хроматина, предположили, что p53-индуцированные длинные нкРНК вносят вклад в активацию энхансеров с помощью p53 [74]. Из индуцированных p53 нкРНК были выбраны три – RP3-510D11.2, loc643401 и LED, но только подавление LED ослабляло функцию p53. Выполнив ChIRP-seq, получили 1698 сайтов связывания LED с хроматином, которые затем сопоставили с аннотацией состояний хроматина [115]. Сайты связывания LED присутствуют во всех состояниях хроматина, однако значительное обогащение наблюдается в сильных энхансерах. Дальнейшее сопоставление LED-ассоциированных энхансеров с данными GRO-seq, гистоновыми метками, а также с изменением уровней гистоновых меток после нокдауна LED позволило сделать вывод, что LED регулирует энхансеры, влияя на установку активной энхансерной гистоновой метки H3K9ac. Показано также, что длинные нкРНК DINO [87] и SLEAR [116] ассоциированы с p53 и непосредственно участвуют в связывании p53 с некоторыми из его генов-мишеней.

РНК, регулирующие гены HOX: Haunt, HOTTIP. Идентифицировав длинные нкРНК, участвующие в плюрипотентности и дифференцировке эмбриональных стволовых клеток (ESC), на примере РНК Haunt постарались определить функции ESC-специфичных нкРНК [45]. Ген Haunt расположен на ~40 тысяч п.н. выше кластера HOXA. Помимо нокаута, нокдауна и сверхэкспрессии Haunt с последующим анализом дифференциальной экспрессии генов и других экспериментов выполнен ChIRP-seq для РНК Haunt. В недифференцированных ESC контакты Haunt непрерывно покрывали геномную область размером 1 миллион п.н. с сигналами, сосредоточенными между ближайшими к гену Haunt HOXA-кластером и геном Skap2. Сигналы ChIRP-seq резко снижались ниже локуса Haunt в непосредственной близости от гена Hoxa1 и уменьшались дальше от 5'- к 3'-концу кластера HOXA, что сопровождалось появлением длинного участка репрессивных меток H3K27me3, покрывающих всю область HOXA (105 тысяч п.н.). Не обнаружено никаких значительных пиков Haunt за пределами ее гена и близлежащих локусов, что подтверждает

цис-функцию Haunt на хроматине. После добавления ретиноевой кислоты (морфоген, регулирующий активацию генов HOX *in vivo* и *in vitro*) РНК Haunt стала чаще контактировать с 5'-стороной области HOXA с пиком в локусах Hoxa5 и Hoxa6, показывая 6–7-кратное увеличение плотности контактов Для сравнения, уровень ассоциации Haunt с хроматином не изменился в дистальном Hoxa13 и увеличился только в 2 раза в локусах Haunt и Skap2. Таким образом, вместо неспецифического связывания хроматина из-за повышенной экспрессии после обработки ESC ретиноевой кислотой, усиленная ассоциация Haunt с локусами HOXA может быть специфически связана с функцией РНК Haunt. Сильное и прямое взаимодействие Haunt с 5'-стороной кластера HOXA указывает на то, что Haunt прямо участвует в модуляции индукции HOXA.

Данные ChIRP-seq позволили показать, что РНК HOTTIP является цис-регулятором поздних генов HOXA (HOXA9–HOXA13). Однако анализ 3767 сайтов связывания показал, что помимо регуляции генов HOXA, HOTTIP регулирует гемопоэтический ландшафт хроматина и транскрипционную программу путем взаимодействия со специфическими для гемопоэза факторами транскрипции и эпигенетическими регуляторами [117].

1.2. Эксперименты ATA: взаимодействия всех РНК с хроматином

С 2017 опубликовано шесть основополагающих работ, в которых предложено несколько экспериментальных протоколов и биоинформатических подходов, позволяющих определять полный РНК-хроматиновый интерактом: MARGI [10], GRID-seq [11], ChAR-seq [12], iMARGI [13,14], RADICL-seq [15], Red-C [16]. Все методы идеологически используют один и тот же основной подход – сначала клетки фиксируют, фрагментируют ДНК с помощью эндонуклеаз или ультразвука, а затем проводят лигирование пространственно сближенных молекул РНК и локусов ДНК при участии специфически сконструированного линкера, несущего биотиновую метку, необходимую для последующего выделения целевых конструкций. При этом сначала лигируют линкер к РНК, а потом к хроматиновой ДНК (хрДНК). Далее проводят обратную транскрипцию и получают химеры кДНК–линкер–хрДНК (рис. 3). В первой работе [10] все процедуры проводили в ядерном лизате, что требовало много клеток (порядка 400 миллионов). В других методах процедуры проводили in situ, поэтому клеток требовалось на два порядка меньше.

В ряде работ (рис. 3, верхняя панель) линкер содержит сайт узнавания MmeI (GRID [11], Red-C [16]) или EcoP15I (RADICL [15]). Эти рестриктазы замечательны тем, что точка расщепления удалена на 20 или 27 нуклеотидов от сайта узнавания. В методах GRID и RADICL линкер несет два сайта рестрикции, направленные как в сторону кДНК, так и в сторону хрДНК, в то время как в Red-C линкер содержит только один сайт, направленный в сторону хрДНК (рис. 3, верхняя панель). Фрагменты, полученные после обработки рестриктазами, выделяют на стрептавидине и секвенируют либо с одного конца (GRID и RADICL), либо с двух (Red-C). В

методах MARGI и iMARGI [10,13,14] конструкцию кДНК–линкер–хрДНК закольцовывают, после чего кольцо линеаризуют рестриктазой BamHI, сайт которой присутствует в линкере, так, что один конец содержит фрагмент линкера и кДНК, а второй – хрДНК и другой фрагмент линкера (рис. 3, нижняя панель). Далее конструкцию отбирают на стрептавидиновых шариках и секвенируют с двух концов.

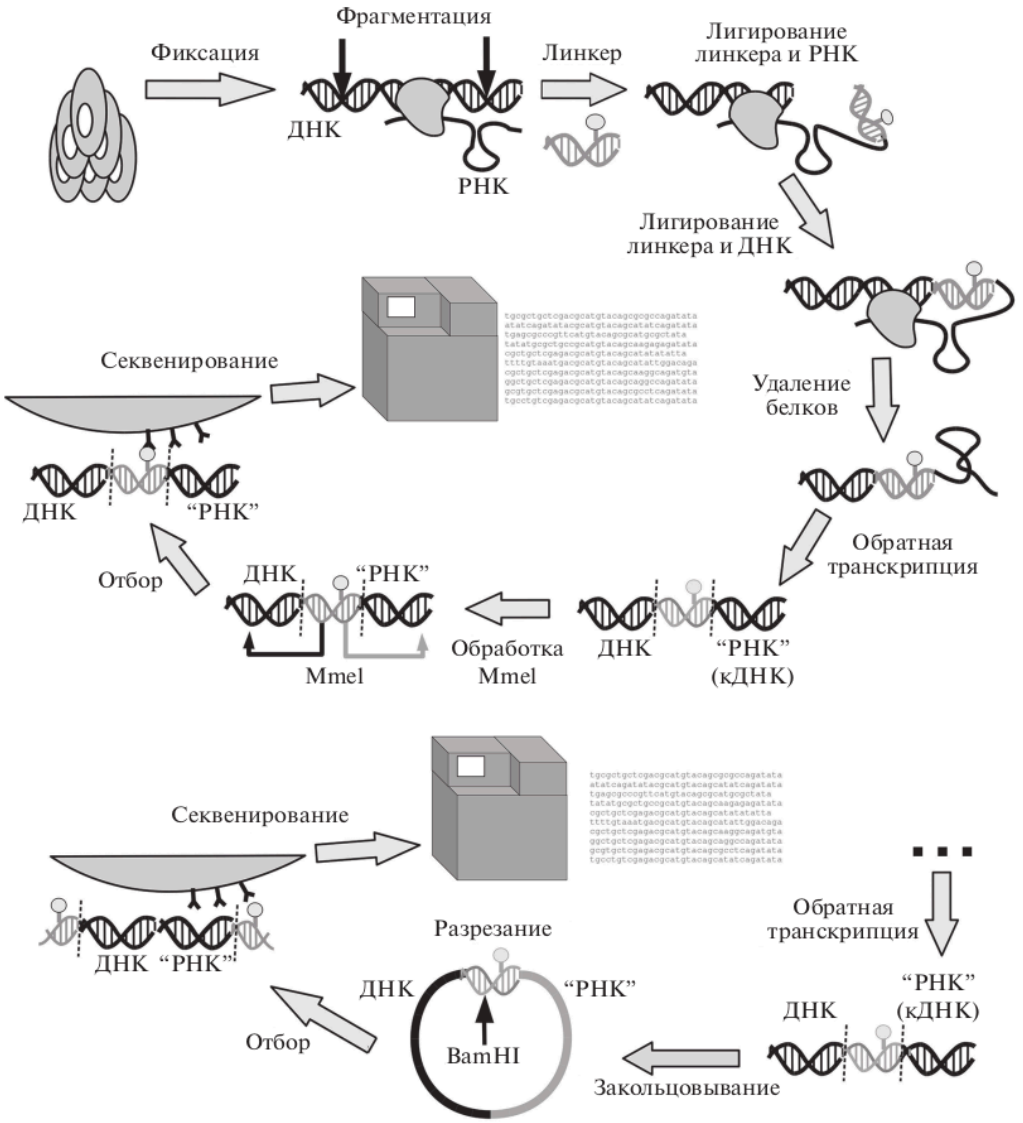


Рисунок 3. Общая схема протоколов ATA. Верхняя панель: GRID, RED-C, RADICL (в RADICL вместо MmeI использована рестриктаза EcoP15I). Нижняя панель: MARGI, iMARGI (показаны этапы после обратной транскрипции, этапы до обратной транскрипции совпадают с GRID, RED-C, RADICL).

В работах использовали разнообразные контроли. В частности, оценивали фоновые взаимодействия, добавляя к клеткам, выбранным для исследования, клетки другого организма (например, дрозофилы), после чего определяли долю обнаруженных межвидовых контактов (MARGI, GRID [10,11]). Эти контрольные эксперименты показали, что лишь небольшая доля (около 2%) контактов являются межвидовыми. В методах ATA корректность протокола проверяют

по соответствию полногеномных данных результатам некоторых экспериментов ОТА.

Отметим, что в используемых методах фиксации предпочтительно сшивают нуклеиновые кислоты с белками, а нуклеиновые кислоты сшиваются между собой менее эффективно. Есть РНК-ДНК-контакты, обусловленные просто процессом транскрипции, которые должны породить большое количество тривиальных контактов (полимеразный след). Поэтому в ряде работ применяют дополнительные процедуры подготовки библиотек. В частности, в методе MARGI [10] используют два варианта фиксации – жесткую с 1%-ным формальдегидом (подход pxMARGI), и фиксацию смесью формальдегида и дисукцинимидилглутарата (подход diMARGI), где ожидается меньший уровень полимеразного следа. Для уменьшения шума в работе [15] (RADICL) использовали два дополнительных варианта подготовки биологического материала. В первом для подавления полимеразного следа клетки инкубировали с ингибитором инициации транскрипции актиномицином D. Во втором хроматин после сшивки сразу же расшивали и обрабатывали протеиназой K в денатурирующих условиях, чтобы оставить только прямые РНК-ДНК-взаимодействия.

В ходе биоинформатического анализа секвенированные фрагменты проходят проверку качества, удаляются ПЦР-дубликаты, затем из последовательностей одноконцевых или парноконцевых чтений удаляют технические последовательности, определяя фрагменты, соответствующие РНК- и ДНК-частям, которые картируют на последовательность соответствующего генома или транскриптома. Таким образом получают координаты фрагмента РНК и локуса ДНК, которые, согласно результатам протокола и секвенирования, определены как контактирующие друг с другом. Все методы анализа подразумевают исключительно уникальное картирование и РНК-, и ДНК-частей, что приводит к исключению из рассмотрения хаРНК, пришедших из повторов, и контактов хаРНК с повторами. Отметим, что использование рестриктазы EcoP15I с более далекой точкой расщепления (27 вместо 20 у MmeI) позволило в эксперименте RADICL-seq получить больше уникально картированных фрагментов [15].

В работах АТА показана клеточная специфичность профилей контактов РНК, особенно РНК-маркеров клеточных линий. Значительная часть РНК-ДНК-контактов приходится на активный хроматин. В частности, уровень покрытия ДНК контактами положительно коррелирует с гистоновыми метками транскрипционно активного хроматина (H3K27ac и H3K4me3) [10]. Большая часть контактов приходится на эухроматин (при сравнении с DHS-seq и ATAC-seq [15]), ДНК-части обогащены в промоторах и межгенных участках [11]. Отметим, что это обогащение может быть артефактом экспериментов, поскольку открытый хроматин более доступен для нуклеаз и поэтому он дает больше контактов. Рассмотрение контактов различных РНК с хроматином [16] показало, что некоторые РНК предпочитают гетерохроматин, Kcnq1ot1 взаимодействует с хроматином, подавленным Polycomb, и с сайтами связывания CTCF, а энхансерные РНК –

преимущественно с промоторами активно транскрибируемых генов и другими энхансерами. Малые ядерные РНК предпочитали активный хроматин, а vlincRNA (очень длинные нкРНК [118]) – гетерохроматин на расстоянии более 10 миллионов п.н. от своего гена. Показано также, что некоторые микроРНК чаще контактируют с гетерохроматином и B-компартментами, что указывает на организацию гетерохроматина вокруг ядрышка.

Показано, что РНК с высоким уровнем контактов с хроматином ассоциированы с белками (по данным fRIP для K562), в частности, с белками комплекса Polycomb, HDAC и DNMT1 [16]. Значительная часть хаРНК взаимодействовала также с ADAR.

Изучение связи РНК-хроматиновых контактов с архитектурой хроматина [11,15] показало высокую глобальную согласованность между РНК-ДНК-взаимодействиями и ДНК–ДНК-взаимодействиями соответствующих генов в пределах ±1 миллиона п.н. в эмбриональных стволовых клетках мыши и ±200 тысяч п.н. в клетках S2 *Drosophila* [11]. Большая часть РНК из генов внутри ТАДов контактировала с ДНК преимущественно внутри своих же ТАДов. Отмечено, что границы ТАДов обогащены ДНК-частями во всех линиях и во всех экспериментальных условиях [15]. Там же показано, что ассоциированные с хроматином РНК предпочитают взаимодействовать с тем же типом хроматинового компартмента, из которого они происходят. Отмечено, что до 30% близких взаимодействий можно объяснить пространственной укладкой хроматина. Протоколы полногеномного изучения архитектуры хроматина и РНК-ДНК-интерактома используют одни и те же подходы: фиксацию клеток и лигирование близко расположенных друг к другу молекул. В результате биоинформатической обработки данных в протоколах Hi-C остается ~75% от исходных чтений, в то время как в случае РНК-ДНК-взаимодействий пригодной для анализа остается только четверть данных, что говорит о несовершенстве экспериментальных подходов при массовом изучении хаРНК.

Согласно [13], 5 из 10 наиболее часто взаимодействующих пар генов определены в iMARGI как гибридные транскрипты в TCGA (The Cancer Genome Atlas). Помимо этого, проанализированы 96 новых опухолевых образцов, в которых нашли 42 гибридных транскрипта, 37 из которых совпадали с РНК-ДНК-контактами в нормальных клетках. Исходя из этих результатов, предложена новая модель образования гибридных транскриптов: два транскрипта сближаются в пространстве (либо из-за сближения их генов, либо из-за того, что РНК оказалась в пространственной близости другого гена), а затем происходит транс-сплайсинг или геномная перестройка.

Применение протокола iMARGI позволило показать, что при искусственно вызванном стрессе, имитирующем сахарный диабет, дисфункция клеток эндотелия обусловлена нарушением экспрессии генов, опосредованным хаРНК [14]. Данные iMARGI указывают на индуцированное стрессом возникновение ряда контактов РНК–хроматин, в том числе, с участием ингибитора активатора плазминогена SERPINE1 и нкРНК LINC00607. Эти гены известны как

коэкспрессируемые в дисфункциональных клетках эндотелия, включая клетки, полученные от больных диабетом. Нокдаун LINC00607 привел к подавлению SERPINE1 и других генов, вносящих вклад в эндотелиальную дисфункцию, подтверждая тем самым предположение о том, что взаимодействия РНК с хроматином вносят вклад в регуляцию транскрипции во время дисфункции эндотелия.

Для завершения обзора современных методов изучения взаимодействий РНК с хроматином необходимо отметить подход RD-SPRITE (RNA & DNA split-pool recognition of interactions by tag extension) [119], который принципиально отличается как по структуре генерируемых данных, так и по стратегии их биоинформатической обработки от разобранных выше методов АТА. В отличие от методов на основе проксимальной лигации, RD-SPRITE основан на комбинаторном баркодировании индивидуальных нуклеиновых комплексов, что позволяет детектировать одновременное взаимодействие нескольких молекул РНК и ДНК. В результате анализа выявляются не попарные взаимодействия, а пространственные хабы – устойчивые мультимолекулярные ансамбли, в которых координированно собираются РНК, ДНК и белки для выполнения конкретной ядерной функции. Это делает RD-SPRITE уникальным инструментом для исследования роли конкретных РНК в формировании трехмерных компартментов, участвующих в важных ядерных функциях, включая процессинг РНК, формирование гетерохроматина и регуляцию экспрессии генов.

1.3. Обсуждение методов ОТА и АТА

Данные экспериментов ОТА (как и АТА) отражают физическую близость РНК и конкретных хроматиновых локусов, что делает их центральными для понимания механизма действия нкРНК. Однако, чтобы определить функциональную роль РНК в соответствующем ДНК-локусе, необходимы дополнительные полногеномные данные, например, о структуре хроматина, экспрессии генов или о локализации ДНК-связывающих и хроматин-модифицирующих белков. В частности, более детально партнеров изучаемой РНК выявляют, используя модификации основных методов ОТА, такие как RAP-RNA [26], CHART-MS [27], CHART-RNA [120], ChIRP-RNA [121] и другие. Помимо методов ОТА, АТА и их производных существует множество других подходов, позволяющих исследовать РНК-ДНК-взаимодействия не только прямо (поиск R-петель [122], триплексных структур [123]), но и косвенно (определение предположительно хроматин-ассоциированных РНК путем комбинации методов изучения РНК-белковых взаимодействий и ChIP-seq) [124].

Обе группы методов активно используются в исследованиях, однако, как правило, изолированно друг от друга, что не позволяет выработать единые стандарты для повышения достоверности и значимости выводов при работе с данными РНК-хроматинового интерактома. До

сих пор отсутствует систематическое сравнение данных ATA и OTA по ключевым характеристикам, таким как точность, полнота и специфичность.

Несмотря на быстрое развитие этих технологий, получаемые данные характеризуются рядом существенных методологических проблем и систематических смещений. Во-первых, плотность контактов РНК зависит от расстояния между геном-источником РНК и целевыми локусами ДНК, расположенными на той же хромосоме [11,13,15,16,27,33]. Это смещение мы будем называть «РНК-ДНК-скейлингом» (или «РД-скейлингом») по аналогии со «скейлингом» в данных ДНК–ДНК-интерактома (метод Hi-C) [125]. Во-вторых, значительное влияние оказывает доступность хроматина, которую будем называть «фоном». Для оценки фона используются либо данные «input» в экспериментах OTA, либо контакты белок-кодирующих РНК (мРНК) – в экспериментах ATA [11], поскольку предполагается, что мРНК белок-кодирующих генов в основной своей массе не должны специфически контактировать с хроматином. Кроме того, по построению эти эксперименты имеют ограниченную точность определения контактов. В экспериментах ATA сшивка РНК с хроматином может происходить на некотором расстоянии от реального контакта (рис. 4А). В то же время в экспериментах OTA точность определения позиции контакта зависит только от размеров фрагментов ДНК (рис. 4Б). Особую проблему представляет наличие неспецифических взаимодействий. Значительная часть наблюдаемых контактов может объясняться электростатическим притяжением между отрицательно заряженной РНК и положительно заряженными гистоновыми хвостами, а также предпочтительным сшиванием формальдегидом аминогрупп [126], представленных на лизинах и аргининах гистонов. Хотя аффинность таких неспецифических взаимодействий относительно низка, их совокупный вклад оказывается существенным из-за огромного количества потенциальных сайтов связывания (рис. 4В). С другой стороны, технические ограничения существующих экспериментальных методов приводят к потере части истинных контактов. Совокупность этих факторов ставит под сомнение специфичность выявляемых взаимодействий и поднимает фундаментальные вопросы о точности, полноте и специфичности данных РНК-хроматинового интерактома.

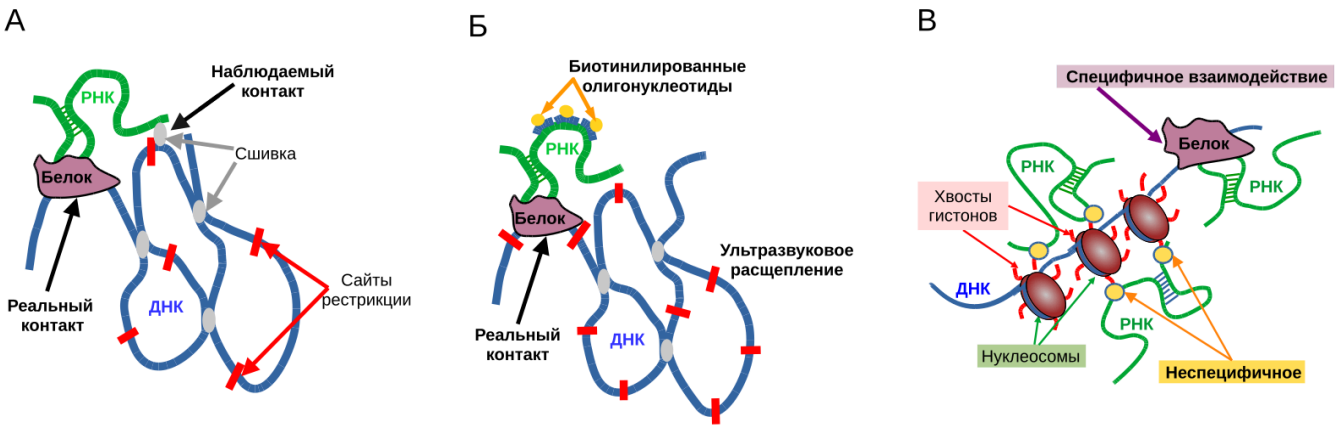


Рисунок 4. Точность определения позиции реального контакта различается в протоколах ATA и OTA. А – Источник смещения позиции в ATA данных – структура хроматина. Б – OTA – смещение наблюдаемой позиции контакта от реальной определяется только размером фрагментов ДНК. В – Возможный источник неспецифических взаимодействий.

1.4. Базы данных нкРНК

Системный анализ функциональных механизмов некодирующих РНК в значительной степени опирается на специализированные базы данных (БД), аккумулирующие экспериментальные данные и их аннотации. Эти ресурсы можно условно разделить на две крупные группы, различающиеся по широте охвата и целевой направленности [127,128].

К первой группе относятся интегративные базы данных, целью которых является консолидация разнородных данных о нкРНК: от базовой аннотации последовательностей и экспрессии до предсказанных функций. Например, NONCODE V6 [28] объединяет аннотации для более чем 644 тысяч транскриптов длинных нкРНК (днРНК) человека, мыши и 23 видов растений, предоставляет профили тканеспецифичной экспрессии и оценку консервативности на уровне транскриптов. Подобные ресурсы, включая БД RNAInter v4 [30], нацеленную на РНК-интерактом (взаимодействия РНК с белками, ДНК и другими РНК), выполняют критически важную функцию: они агрегируют разноплановые данные, создавая основу для первичной функциональной интерпретации и последующей интеграции с другими омиксными данными, что способствует формированию новых гипотез.

Ко второй группе принадлежат специализированные базы данных, фокусирующиеся на конкретном биологическом контексте (например, сердечно-сосудистая система [129]), патологическом процессе или определенном типе данных. Их преимущество заключается не в широте охвата, а в глубине и качестве курирования информации в заданной области. Примерами служат ресурсы, курирующие списки днРНК, играющих причинную роль в онкогенезе (Cancer LncRNA Census 2 [130]) и других заболеваниях (LncRNADisease v3 [131]), а также базы, аккумулирующие специфические типы данных, как HiMoRNA, которая напрямую связывает профили экспрессии днРНК с силой гистоновых модификаций в конкретных локусах [132].

Подобные ресурсы незаменимы для валидации гипотез и получения высококачественных, тематически сфокусированных данных. Однако их многообразие приводит к фрагментации знаний, вынуждая исследователя обращаться к множеству источников для формирования целостной картины.

Особый интерес представляют базы данных, посвященные РНК-интерактому, то есть совокупности взаимодействий нкРНК с другими молекулами. До недавнего времени для такого важного и активно изучаемого типа взаимодействий, как контакты РНК с хроматином, не существовало узкоспециализированного ресурса, который агрегировал бы результаты полногеномных методов ATA (например, MARGI, GRID-seq, ChAR-seq) и OTA (например, ChIRP-seq, CHART-seq, RAP) и предоставлял удобные инструменты для их анализа. Публикация нашей базы данных RNA-Chrom в 2023 году [133] позволила заполнить этот пробел. До ее появления основными источниками подобной информации оставались ресурсы LnChrom [31] (в настоящее время не поддерживается), RNAInter v4 [30] и NPInter v5 [134] (в настоящее время недоступна).

Элементарной записью в RNAInter v4 является экспериментально подтвержденный контакт между молекулами (РНК-ДНК, РНК-белок, РНК-РНК), для каждого из которых рассчитан показатель достоверности, основанный на качестве эксперимента и его воспроизводимости. RNAInter v4 выполняет роль мета-агрегатора, интегрируя данные из множества источников. Несмотря на свою комплексность, RNAInter v4 содержит относительно небольшое количество полногеномных экспериментов по РНК-хроматиновым взаимодействиям (в частности, из методов ATA представлен только MARGI) [30].

База данных NPInter v5, актуальная на момент публикации в 2023 году, заявила о включении нескольких полногеномных РНК-ДНК интерактомов, полученных методами iMARGI, ChAR-seq, GRID-seq, ChIRP-seq и CHART-seq [134]. Однако с тех пор объем публикаций новых данных ATA и OTA продолжил расти, а обновление РНК-хроматинового раздела в NPInter v5 не производилось.

Таким образом, исследователь, изучающий РНК-хроматиновые взаимодействия, до появления RNA-Chrom сталкивался с необходимостью:

1. Обращаться к интегративным платформам типа RNAInter v4, где соответствующие данные могли быть представлены неполно.
2. Работать непосредственно с оригинальными исследованиями и архивами «сырых» экспериментальных данных, например, «Gene Expression Omnibus» (GEO, *http://www.ncbi.nlm.nih.gov/geo*), ENCODE (*https://www.encodeproject.org*), что требует специальных биоинформатических навыков.

Существование этого пробела подчеркивало актуальную потребность в создании

специализированного ресурса, который не только каталогизировал бы полногеномные РНК-ДНК контакты, но и предоставлял инструменты для их сравнительного анализа и визуализации в геномном контексте (например, интеграция с браузерами генома вроде UCSC Genome Browser [135]).

1.5 Обработка данных

Обработка данных, полученных методами секвенирования нового поколения, в том числе при картировании РНК-хроматиновых взаимодействий, осуществляется с помощью специализированных вычислительных конвейеров. В настоящее время как минимум для каждого основного экспериментального метода существуют свои публично доступные конвейеры, например, iMARGI-docker (*https://github.com/Zhong-Lab-UCSD/iMARGI-Docker*), Flypipe (*https://github.com/straightlab/flypipe*), ChARtools (*https://github.com/straightlab/chartools*), GridTools (*https://github.com/GridTools/gridtools*), RADICL-seq (*https://github.com/fagostini/RADICL_analysis*), RedClib (*https://github.com/agalitsyna/RedClib*), ChRD-PET (*https://github.com/fengchuiguo1994/ChRDPETpipeline*), ChIRP-seq (*https://github.com/bdo311/chirpseq-analysis*), представляющие собой последовательность взаимосвязанных этапов. Цель этих конвейеров – трансформировать первичные данные секвенирования в биологически интерпретируемую информацию. Типичный конвейер обработки данных как OTA, так и ATA включает следующие ключевые шаги:

1. Контроль качества: оценка технических параметров прочтений (длина, качество оснований, доля адаптеров) с помощью инструментов, таких как FastQC или MultiQC [136].
2. Предобработка: удаление адаптерных последовательностей и обрезка низкокачественных концов прочтений, например, с помощью Trimmomatic [137].
3. Удаление ПЦР-дубликатов: идентификация и удаление артефактных копий молекул, возникших в процессе амплификации. Проблематике и инструментам этого этапа посвящен следующий раздел обзора.
4. Картирование: выравнивание прочтений на референсный геном с помощью специализированных программ, таких как HISAT2 [138], STAR [139], BWA [140] или Bowtie2 [141].
5. Идентификация сигнала: статистическое выделение геномных локусов, достоверно обогащенных сигналом взаимодействий (пиков), по сравнению с фоновым распределением, например, с помощью MACS2 [46] или BaRDIC [142].

Однако узкая специализация существующих конвейеров под конкретные экспериментальные протоколы создает серьезную методологическую проблему. Разработка нового метода зачастую влечет за собой создание уникального конвейера обработки. В результате данные,

полученные в разных исследованиях, обрабатываются неединообразно, что делает их прямое сопоставление и совместный интеграционный анализ крайне затруднительным. Эта неоднородность проявляется даже на уровне отдельных этапов: например, в конвейерах для данных ATA ПЦР-дубликаты часто удаляют до этапа картирования, тогда как для данных OTA применяют методы, работающие с уже картированными чтениями, или не удаляют их вовсе. Ситуацию усугубляет и эволюция самих конвейеров: так, для метода ChAR-seq первоначальный конвейер Flypipe был позднее заменен на ChARtools.

Таким образом, отсутствие стандартизированного подхода к обработке приводит к неоднозначности в получении итоговых данных и к дефициту унифицированных репозиториев готовых контактов. Это подчеркивает важность создания единого протокола обработки и курируемой базы данных, таких как RNA-Chrom. В контексте стандартизации и крупномасштабной обработки данных выбор оптимального инструмента для ключевого этапа удаления ПЦР-дубликатов приобретает особую значимость, поскольку от его эффективности и воспроизводимости напрямую зависит надежность всех последующих выводов.

1.5.1 ПЦР-дубликаты

В настоящее время методы секвенирования нового поколения (next-generation sequencing, NGS) широко используются в различных биологических приложениях [143]. Процесс секвенирования включает несколько ключевых этапов: подготовку образца, фрагментацию нуклеиновых кислот, амплификацию и секвенирование. На каждом этапе возникают систематические ошибки, которые могут в различной степени влиять на качество данных и их интерпретацию [144]. В частности, во время амплификации (также известной как полимеразная цепная реакция, ПЦР) образуются копии исходных молекул ДНК (ПЦР-дубликаты), которые могут составлять значительную часть данных секвенирования (в некоторых случаях – десятки процентов от общего объема) [145]. Степень искажения данных секвенирования в результате ПЦР-дубликатов остается предметом постоянных дискуссий [145–148]. Суть этой дискуссии заключается в том, является ли дедупликация необходимым шагом для устранения технических артефактов, искажающих количественный анализ, или же она может непреднамеренно удалять биологически значимый сигнал в определенных экспериментальных условиях, например, в условиях изначально низкой изменчивости.

Помимо решений на уровне подготовки образцов, таких как использование уникальных молекулярных идентификаторов (unique molecular identifiers, UMI) для маркировки молекул-шаблонов перед амплификацией, был разработан широкий спектр биоинформатических инструментов для идентификации и удаления ПЦР-дубликатов из данных секвенирования. Вычислительный анализ данных на основе UMI выполняется с помощью специализированных

инструментов, таких как описанные в [149] и BBTools Clumpify [150], и их анализ выходит за рамки данной работы. Все инструменты дедупликации можно разделить на две категории: основанные на выравнивании («alignment-based») и «*de novo*-based» [151]. Ключевым компонентом alignment-based методов является выравнивание прочтений по референсному геному или транскриптому, что делает эти методы зависимыми от доступности и качества референсного генома/транскриптома. К числу наиболее известных инструментов в этой категории относятся Picard MarkDuplicates [152] и SAMtools markdup [153]. В отличие от этого, de novo-based методы дедупликации идентифицируют и группируют идентичные или очень похожие прочтения без выравнивания их на референсный геном. Многочисленные инструменты реализуют этот подход, включая FastUniq [154], Seqkit rmdup [155], BBTools Clumpify [150], CD-HIT-DUP [156], Fastx Toolkit Collapser [157] и другие [158–165].

Традиционно сопоставление прочтений с референсным геномом является наиболее трудоемким этапом предварительной обработки данных NGS, что побуждает ученых применять подходы к раннему удалению ПЦР-дубликатов для оптимизации вычислительных процессов. Однако переход к *de novo*-based методам дедупликации создает новую проблему: эти методы обычно требуют загрузки всего набора данных в оперативную память (RAM), что осуществимо для небольших наборов данных, таких как RNA-seq, но становится критическим ограничением для современных наборов данных, содержащих сотни миллионов прочтений. Требования к оперативной памяти могут достигать десятков или даже сотен гигабайт (ГБ) (см. раздел 3.2.2 «Сравнение Fastq-dupaway с *de novo*-based инструментами ПЦР-дедупликации»), что увеличивает вычислительные затраты и делает обработку недоступной для исследователей, не имеющих доступа к высокопроизводительной вычислительной инфраструктуре. Кроме того, некоторые инструменты изменяют идентификаторы прочтений (Fastx Toolkit Collapser) или некорректно обрабатывают парные прочтения (FastUniq).

В настоящее время существует множество вычислительных подходов биоинформатической обработки РНК-хроматиновых данных, которые либо не удаляют ПЦР-дубликаты, либо используют разные методы, которые хорошо подходят лишь для конкретного типа данных. Отсутствие универсальной программы удаления ПЦР-дубликатов затрудняет создание стандартизированного подхода обработки данных РНК-хроматинового интерактома и их сравнительного анализа.

Глава 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1.База данных RNA-Chrom[3]

2.1.1 Данные полногеномного РНК-ДНК интерактома

Поскольку первые статьи с методами ATA появились только в 2017 году, найти соответствующие им данные было несложно. Совсем иначе обстояли дела с данными OTA. Сначала мы искали их в «Gene Expression Omnibus» (http://www.ncbi.nlm.nih.gov/geo) по ключевым словам «RAP-seq», «CHART-seq», «ChIRP-seq», «dChIRP-seq», «ChOP-seq», «CHIRT-seq» и учитывали только наборы данных человека и мыши. Затем мы прошлись по статьям, которые ссылались на основные методы OTA (RAP [4], CHART-seq [5], ChIRP-seq [6], dChIRP-seq [7], ChOP-seq [8] и CHIRT-seq [9]). Удивительно, но мы нашли большое количество публикаций, в которых использовались методы OTA, но не было общедоступных данных. Только один автор ответил на наш запрос и предоставил нам RAP-данные для РНК Firre [86]. Всего было найдено 75 статей, в которых были данные в открытом доступе.

2.1.2. Универсальный протокол обработки данных в базе данных RNA-Chrom

2.1.2.1. Удаление ПЦР-дубликатов

К сожалению, ПЦР-дубликаты нам пришлось удалять с помощью двух программ, так как FastUniq [154] не умеет работать с одноконцевыми чтениями. Таким образом возможные ПЦР-дубликаты в данных ATA и в данных OTA с парноконцевыми чтениями были удалены с помощью FastUniq, в то время как SeqKit rmdup [155] использовался для обработки данных OTA с одноконцевыми чтениями. Для данных iMARGI [13,14] мы следовали рекомендациям авторов оригинальной статьи, чтобы выполнить этот шаг после этапа «Проверка и достраивание сайтов рестрикции». Важно отметить, что все реплики были обработаны независимо, начиная с шага «Удаление ПЦР-дубликатов» и до «BlackList-фильтр» включительно.

2.1.2.2. Проверка и достраивание сайтов рестрикции

Поскольку в методах GRID-seq [11], Red-C [16] и iMARGI [13,14] для фрагментации геномной ДНК использовались эндонуклеазы (в RADICL-seq [15] применялась ДНКаза I), было важно отфильтровать чтения, соответствующие ДНК-частям контактов, у которых 3’-конец не

[3] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **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, – pp. baad025. EDN: YEKQIZ. Импакт-фактор 3,6 (JIF) (1.02/0.40).

заканчивался половиной сайта рестрикции (при ориентации исходной химеры контакта: 5'-ДНК-линкер-РНК-3'). Далее к 3’-концам ДНК-частей добавлялась вторая половина сайта рестрикции, что несколько увеличивало длину чтений и повышало эффективность уникального картирования. Данная процедура проводилась в строгом соответствии с рекомендациями оригинальных статей.

2.1.2.3. Контроль качества чтений

Мы использовали TRIMMOMATIC (v0.39) [137] (параметры: «window size» = 5, «quality threshold» = 26, «minlen» = 14) для обнаружения позиции низкого качества в каждом чтении. Несколько датасетов низкого качества с более чем 50% отброшенных данных после триммирования были повторно триммированы с более низким порогом качества («quality threshold» = 22) и дополнительным параметром «LEADING» = 22.

Качество чтений оценивалось с помощью программы FastQC (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc*) перед обработкой и после этапа триммирования.

2.1.2.4. Картирование

Данные ОТА были картированы на геном с помощью программы HISAT2 (версия 2.1.0) [138] (параметры для одноконцевых чтений: -k 100 --no-spliced-alignment --no-softclip; для парноконцевых чтений: -k 100 --no-spliced-alignment --no-softclip --no-discordant --no-mixed). ДНК-части и РНК-части контактов (данные АТА) были независимо сопоставлены с геномом той же программой (параметры для ДНК-частей: -k 100 --no-spliced-alignment --no-softclip, для РНК-частей: -k 100 --no-softclip --dta-cufflinks --known-splicesite-infile). Аннотация сайтов сплайсинга для соответствующих геномов была получена с помощью скрипта «hisat2_extract_splice_sites.py» [138]. Чтения в SAM-файлах были отфильтрованы на предмет уникального картирования с максимум 2 несовпадениями относительно эталонного генома.

2.1.2.5. Определение ориентации РНК-частей контактов

В ходе обработки данных мы заметили, что РНК-части контактов в ряде экспериментов могли представлять собой не те части последовательностей генов, с которых транскрибировались соответствующие РНК, а обратные комплементы этих последовательностей. Другими словами, в некоторых экспериментах могла быть секвенирована «прямая» цепь считываемой части кДНК, а в других – «обратная».

Для того чтобы определить, верна ли эта гипотеза, был проведен эксперимент, основанный на следующем предположении: в любой жизнеспособной клеточной линии гены рибосомных

белков должны быть высоко экспрессированы, и, вероятно, значительная часть имеющихся у нас данных – это именно матричные РНК этих белков, контактирующие с хроматином на пути к ядерным порам. Для каждого набора данных мы выбрали РНК-части контактов, которые были выровнены в координатах генов рибосомных белков на обеих цепях, а затем вычислили доли чтений, выровненных на генной цепи и цепи, комплементарной гену (рис. 5А).

Если с генной цепью было сопоставлено больше чтений, чем с ее комплементарной, то во время секвенирования «правильная» цепь кДНК считывалась в соответствии с последовательностью РНК, контактирующей с хроматином, и наоборот. Части РНК из экспериментов, которые имели «неправильные» последовательности цепи кДНК, необходимо было перевернуть перед дальнейшим анализом, хотя это явно не было указано ни в одной из исходных статей.

Примечательно, что при рассмотрении человеческих данных были прочитаны следующие цепи: «правильная» цепь кДНК (в случае эксперимента Red-C [16]); обратная цепь (для экспериментов GRID-seq [11] и iMARGI [13,14]). Тогда как в случае MARGI [10], похоже, что в основном была прочитана случайная цепь, и ориентация РНК-частей контактов была утеряна (рис. 5А). Можно заметить, что для некоторых наборов данных MARGI (SRR5278097, SRR5278097, SRR5278100, SRR5278102) цепи были определены однозначно. Однако из-за низкого покрытия генов рибосомных белков в этих наборах данных (см. приложение А, табл. А.1) и потери ориентаций РНК-частей в других наборах данных MARGI мы решили исключить эксперимент MARGI из дальнейшего анализа.

В случае мышиных данных мы видим, что мышиный GRID-seq ведет себя как человеческий GRID-seq, а RADICL-seq [15] ведет себя как Red-C (рис. 5Б).

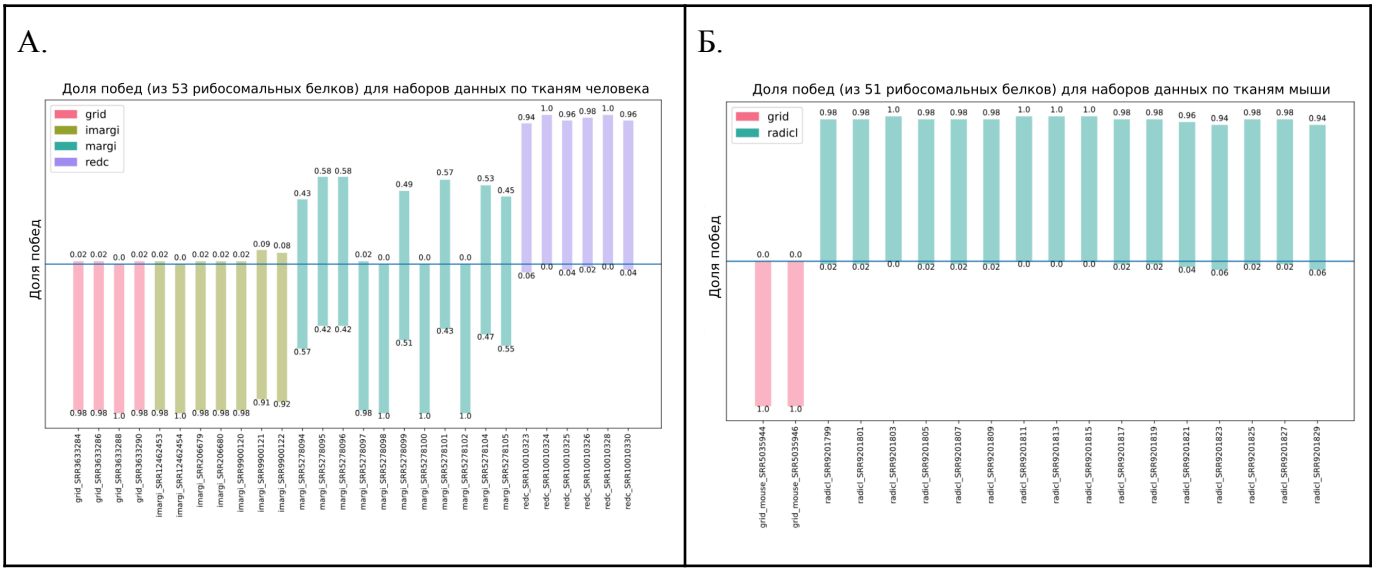


Рисунок 5. Ориентация РНК-частей чтений контактов на подвыборке из генов, кодирующих рибосомальные белки. А. Человеческие наборы данных, отобраны 53 гена: RPL22, RPL11, RPL5, RPL31, RPL37A, RPL32, RPL15, RPL14, RPL29, RPL24, RPL22L1, RPL39L, RPL35A, RPL9, RPL34, RPL37, RPL26L1, RPL10A, RPL7L1, RPL7, RPL30, RPL8, RPL35, RPL12, RPL7A, RPLP2, RPL27A, RPL41, RPL6, RPLP0, RPL21, RPL10L, RPL36AL, RPL4, RPLP1, RPL3L, RPL13, RPL26, RPL23A, RPL23, RPL19, RPL27, RPL38, RPL28, RPL3, RPL36A, RPL13A, RPL39, RPL10 , RPL17, RPL36, RPL18A, RPL18. Б. Мышиные наборы данных, отобран 51 ген: Rpl7, Rpl31, Rpl37a, Rpl7a, Rpl12, Rpl35, Rpl22l1, Rpl34, Rpl11, Rpl22, Rpl9, Rpl5, Rplp0, Rpl6, Rpl21, Rpl32, Rpl28, Rpl13a, Rpl18, Rpl27a, Rplp2, Rpl18a, Rpl13, Rplp1, Rpl4, Rpl29, Rpl14, Rpl41, Rpl26, Rpl23a, Rpl23, Rpl19, Rpl27, Rpl38, Rpl10l, Rpl15, Rpl37, Rpl30, Rpl8, Rpl3, Rpl39l, Rpl35a, Rpl24, Rpl3l, Rpl10a, Rpl7l1, Rpl36, Rpl17, Rpl39, Rpl10, Rpl36a.

2.1.2.6. CIGAR-фильтр

Чтения, соответствующие РНК-частям контактов, могут быть картированы тремя способами:

1) с полным совпадением с референсным геномом по всей длине чтения (CIGAR вида «25M», где M – совпадение). Такие чтения оставлялись без изменений;
2) содержащие один пропущенный интервал (CIGAR вида «30M65N10M», где M – совпадение, N – пропущенный регион). Для таких чтений оставлялся самый длинный участок, картированный без разрывов;
3) более сложные варианты картирования (чтения со сложным сплайсингом): множественные пропущенные интервалы (CIGAR вида: «8M1113N56M79N8M»), картирование со вставками или делециями. Все такие чтения были удалены.

2.1.2.7. BlackList-фильтр

ДНК-части контактов, которые попали в регионы из ENCODE BlackList для GRCh38 и GRCm38 (идентификатор: ENCSR636HFF), были удалены. Для части данных ОТА, а именно для «input»-библиотек (данные с фоновыми или неспецифическими контактами), BlackList-фильтр не

применялся, чтобы избежать краевых эффектов на этапе «Перевзвешивание контактов».

2.1.2.8. Аннотация РНК-частей контактов генами

Для любого РНК-ДНК контакта важно знать, какому гену принадлежит РНК-часть (рис. 6). Для этого мы собрали общую аннотацию генов (см. приложение А, табл. А.2 и А.3), балансируя между ее большим размером и низкой представленностью определенных биотипов РНК. Кластеры неаннотированных РНК-частей контактов были названы ucaRNAs и тоже добавлены в общую аннотацию генов. Если названия генов повторялись в общей аннотации генов, то им присваивался порядковый номер, чтобы все названия генов в базе данных были уникальными. Например, ген «WASIR1» был обнаружен дважды в GENCODE-аннотации, поэтому мы присвоили копиям порядковый номер: «WASIR1_1» и «WASIR1_2» соответственно. Поскольку данные OTA не имеют никакой информации о РНК-частях контактов, мы присвоили каждой ДНК-части из эксперимента координаты соответствующего исходного гена РНК. В случае данных ATA, если РНК-часть контакта пересекает ген по крайней мере на 1 нуклеотид, эта РНК-часть приписывалась этому гену. Если РНК-часть контакта пересекает более одного гена на одной и той же цепи, эта РНК-часть была назначена гену, показывающему наибольшее покрытие РНК-частями, которое определялось как общее количество РНК-частей, сопоставленных с геном, разделенное на длину гена («процедура голосования»).

На данном и последующих этапах реплики уже объединены для увеличения объема данных и покрытия. В базу данных были добавлены только те контакты, у которых РНК-части пересекали гены из общей аннотации. Остальные контакты были названы «синглтонами» и не использовались. Пропустив данные через все предыдущие шаги протокола, мы получили итоговое количество контактов для каждого эксперимента.

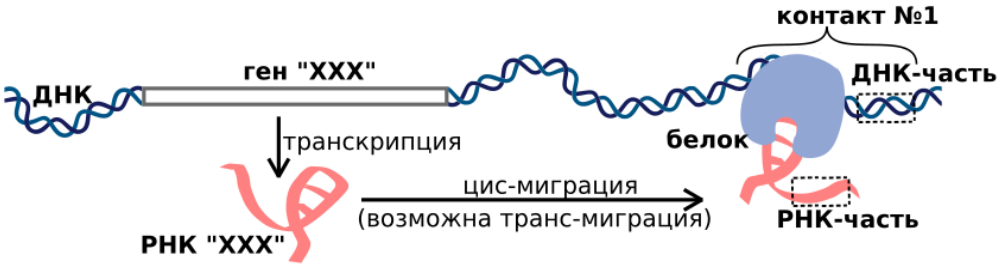


Рисунок 6. РНК «XXX» взаимодействует с локусом ДНК и образует контакт № 1. В случае методов OTA мы видим только ДНК-части контактов, тогда как в случае методов ATA мы видим как ДНК-части, так и РНК-части контактов. Цис- и транс-миграция – это миграция РНК в окрестности родительской хромосомы и за ее пределами соответственно.

2.1.2.9. Сборка ucaRNAs

Значительное количество РНК-частей не было аннотировано ни одной из использованных нами аннотаций генов. Некоторые из этих чтений могут принадлежать неизвестным

некодирующим РНК. В связи с этим, транскрипты, не соответствующие ни одному известному гену из базы данных GENCODE [166] (версия аннотации 35 для человека и 25 для мыши), были собраны с помощью StringTie [167], а затем отфильтрованы по нескольким критериям: длине, расстоянию до ближайшего известного гена на той же цепи, консервативности на разных таксономических уровнях и высокому покрытию. Мы назвали группу транскриптов, которые прошли все фильтры, неаннотированными ассоциированными с хроматином РНК (unannotated chromatin-associated RNAs, ucaRNAs). Каждой ucaRNA был присвоен уникальный идентификатор на основе ее местоположения в геноме. Например, X_1_13_a_hg38 – это ucaRNA, расположенная на хромосоме 1 генома человека (версия hg38) в 13-м бине (каждая хромосома была разделена на бины по 10 000 п.н.). А буква «a» указывает на то, что исходный ген этой ucaRNA является первым в бине относительно начала соответствующей хромосомы.

2.1.2.10. Перевзвешивание контактов

Для каждого эксперимента была проведена следующая процедура. Каждому контакту было присвоено единичное значение («n-reads (raw)» или $n_{raw} = 1$), поскольку один контакт соответствует 1 паре РНК- и ДНК-частей. После этого мы разделили геном на бины по 500 п.н. и определили трек с фоновым сигналом в зависимости от типа данных:

1) **Эксперименты ATA.** Для каждого эксперимента мы удаляли 50 наиболее контактирующих и 1000 наименее контактирующих мРНК и, согласно подходу, предложенному в протоколе GRID-seq [11], суммировали число транс-контактов (не с родительской хромосомой) мРНК в каждом бине.
2) **Эксперименты OTA.** Для каждого бина мы суммировали количество «input»-контактов, которые серединой чтения попадали в соответствующий бин. Если для конкретного эксперимента не было «input»-библиотеки, то фон делался постоянным, то есть мы назначали ровно один контакт каждому бину.

Полученный трек мы сгладили с помощью программы Smoother, встроенной в StereoGene (v.2.20) [168], (параметры: bin = 500, wSize = 1000000, flankSize = 10000, kernelSigma = 3000, kernelType = NORMAL) и использовали сглаженные значения в соответствующих бинах в качестве фонового сигнала – «n-reads (background)» или «n_{bg}».

Затем для каждого контакта мы нормализовали n_{raw} на значение фонового сигнала (n_{bg}) в бине, который соответствовал геномной координате ДНК-части контакта. Для работы с ДНК-частями, картированными в области с нулевым значением фонового сигнала, мы добавили псевдосчетчик к n_{bg}. Таким образом, мы получили нормализованное значение («n-reads (normalized)» или n_{norm}). Эта нормализация гарантирует, что сумма нормализованных значений равна количеству чтений в эксперименте:

$$n_{norm} = \frac{n_{raw}}{n_{bg} + 0.5} \cdot (\sum n_{raw} \; / \; \sum \frac{n_{raw}}{n_{bg} + 0.5})$$

Для экспериментов OTA были посчитаны пики (области генома, обогащенные контактами РНК с хроматином) с помощью программы MACS2 [46] со следующими параметрами: -Q 0.05 -FORMAT BED (если чтения одноконцевые) или -Q 0.05 -FORMAT BEDPE (если чтения парноконцевые). Контакты, ДНК-части которых пересекали пики не менее чем на 1 п.н., будут использоваться при дальнейшем построении аналитических графиков с заранее определенными «n-reads (raw)» и «n-reads (normalized)». Применив этот фильтр, мы получили «n-reads (raw & in peaks)» и «n-reads (norm. & in peaks)».

Таким образом, в нашей базе данных существует четыре типа нормализации. В качестве последнего шага перед загрузкой данных в RNA-Chrom мы проаннотировали ДНК-части контактов генами и окологенными областями. Чтобы получить сопоставимую характеристику контактируемости для РНК в экспериментах ATA, мы ввели метрику «CPKM» (Contacts Per Kilobase of RNA length per Million filtered contacts in the experiment) – количество контактов, деленное на килобазу длины соответствующей РНК и на миллион отфильтрованных контактов в эксперименте. Важно отметить, что для данных OTA метрика «CPKM» не имеет смысла.

2.1.3 Веб-сервис RNA-Chrom

Из множества платформ/JavaScript-библиотек для одностраничной веб-разработки мы выбрали наиболее популярные: «Node.js» (*https://nodejs.org/en*) (асинхронная событийно-управляемая среда выполнения JavaScript), «React.js» (*https://reactjs.org*) и «Redux» (*https://github.com/reduxjs/redux*). Библиотека «Material-UI V4» (*https://mui.com*) была взята в качестве основы для элементов веб-интерфейса, а «Plotly JavaScript Open Source Graphing Library» (*https://plot.ly/javascript*) была использована для создания интерактивных графиков. В качестве веб-микрофреймворка для разработки бэкенда мы использовали веб-микрофреймворк «Quart» (*https://pgjones.gitlab.io/quart*).

2.2. ПЦР-дедуплексатор[4]

2.2.1. Среда, в которой проводилось тестирование программ

Все тесты проводились на высокопроизводительном вычислительном сервере, оснащенном

[4] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: 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 (2025). EDN: VBEHEL. Импакт-фактор 3,9 (JIF) (0.88/0.20).

процессором Intel® Xeon® Gold 6226 CPU @ 2.70GHz и 1.5 ТБ RAM, работающем в кластере GPFS. GPFS поддерживается корпоративными полками с жесткими дисками SAS объемом 14 ТБ со скоростью вращения 7200 оборотов в минуту, сконфигурированными в RAID 6 для повышения производительности и подключенными через InfiniBand к серверам «dss». Тесты последовательного ввода-вывода с использованием команды *dd* в файловой системе GPFS показали скорость записи 1,3 ГБ/с и скорость чтения 1,1 ГБ/с.

2.2.2. Наборы данных для сравнительного анализа программ

Все наборы данных, использованные в этом сравнительном исследовании, находятся в открытом доступе в архиве последовательностей NCBI (*https://www.ncbi.nlm.nih.gov/sra*). Идентификационные номера наборов данных и экспериментальные сведения приведены в табл. 3.

2.2.3. Запуск программ

Все программы дедупликации ПЦР-дублей запускались на одном ядре CPU (с поддержкой одного потока), чтобы обеспечить равноправное сравнение между BBTools Clumpify (поддерживающей обработку данных на нескольких ядрах одновременно) и другими инструментами, не обладающими такой возможностью. Каждая программа запускалась пять раз на каждом наборе данных для обеспечения надежности измерений и учета потенциальной вариабельности производительности.

Показатели производительности, включая пользовательское время (User time), системное время (System time), затраченное время (Elapsed time) и использование оперативной памяти, измерялись с помощью команды Unix */usr/bin/time*. Время CPU рассчитывалось как сумма пользовательского и системного времени. Полные параметры запуска программ и их версии приведены в дополнительных материалах нашей статьи [169].

2.3. Интеграция баз данных HiMoRNA и RNA-Chrom[5]

Базы данных используют разные источники аннотации генов. Поэтому, во-первых, необходимо было установить соответствие между генами из разных версий GENCODE-аннотаций, так как для длинных некодирующих РНК HiMoRNA использует «gencode basic annotation v31», а RNA-Chrom – «gencode basic annotation v35». Мы использовали три показателя сходства: (1) одинаковые имена генов (совпадение «gene_name», рис. 7A); (2) одинаковые идентификаторы

[5] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: 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).

генов (совпадение «gene_id», рис. 7А); и (3) индекс Жаккара (отношение длины перекрытия генов к длине их объединения) больше 0.99 (Индекс Жаккара > 0.99, рис. 7А). К сожалению, имена/идентификаторы и координаты генов в разных источниках аннотаций не всегда совпадают. Чтобы решить эту проблему, мы пересекли 4145 генов днРНК из HiMoRNA с 60619 генами из RNA-Chrom по геномным координатам с использованием bedtools (команда intersect), в результате чего получилось 6778 пар генов (из-за того, что гены из HiMoRNA несколько раз пересеклись с генами из RNA-Chrom, пар получилось больше 4145). Два гена из HiMoRNA не пересеклись ни с одним из генов из RNA-Chrom (ENSG00000267034.1, ENSG00000280076.1). Далее для каждой пары генов был рассчитан индекс Жаккара. Используя описанные выше три показателя сходства, мы разбили 6778 пар генов на шесть групп (рис. 7А). Взяв в качестве главной метрики сходства «Индекс Жаккара > 0.99», мы выделили четыре группы (рис. 7А, группы 2, 4, 5 и 6), между которыми гены не пересекались (рис. 7Б). Таким образом мы определили 4100 однозначных соответствий между генами. Из оставшихся 43 генов из HiMoRNA нам удалось сопоставить 24 гена с генами из RNA-Chrom по метрике совпадение «gene_name». Всего мы получили 4124 гена днРНК, общих как для HiMoRNA, так и для RNA-Chrom. Таблица соответствия генов днРНК («lncRNA correspondence table») доступна для скачивания на веб-ресурсе HiMoRNA.

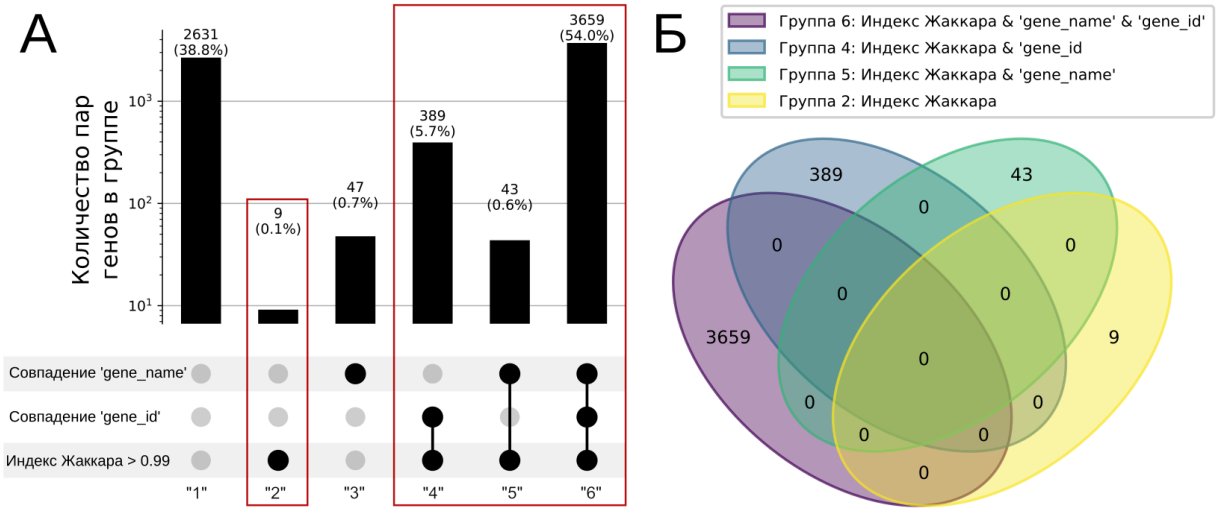


Рисунок 7. Пересечение 4145 генов из HiMoRNA с 60619 генами из RNA-Chrom. А – разделение пар генов на шесть групп в зависимости от показателей сходства, которым они удовлетворяют. Красными прямоугольниками выделены те группы, в которых достигнуто однозначное соответствие между генами. Б – диаграмма Венна между группами генов 2, 4, 5 и 6 (суммарное количество пар генов в четырех группах равно 4100).

Во-вторых, к веб-интерфейсу RNA-Chrom была добавлена новая опция – обработка параметров (локус, название РНК, внутренний для RNA-Chrom идентификатор РНК, организм) из специального вида URL-ссылки (например, *https://rnachrom2.bioinf.fbb.msu.ru/basic_graphical_summary_dna_filter?locus=chrX:*

23456-24253566&name=XIST&rnaID=227896&organism=Homo+sapiens) и предоставление информации о контактах запрашиваемой в URL-ссылке днРНК с хроматином в разных типах экспериментов на новой странице в браузере.

В-третьих, на стороне HiMoRNA были сделаны следующие улучшения:

1) Для правильной генерации URL-ссылки была добавлена таблица соответствия генов днРНК между RNA-Chrom и HiMoRNA.
2) На веб-страницу «Страница результатов поиска» была добавлена кнопка «Перейти в RNA-Chrom БД» («Go to RNA-Chrom DB») с выпадающим списком (рис. 25), который позволяет сгенерировать три типа URL-ссылки для перехода на страницу RNA-Chrom:
    a) с контактами данной днРНК в определенном геномном локусе, расширенном на 1 / 5 / 10 / 25 / 50 / 100 тысяч п.н.;
    b) со всеми контактами данной днРНК;
    c) со всеми РНК, которые имеют контакты в определенном геномном локусе.

2.3.1. Односторонний точный тест Фишера

В большинстве триад («днРНК–пик эпигенетической модификации–ассоциированный с пиком ген») обнаруживаются пики гистоновых модификаций как с отрицательной, так и с положительной корреляцией экспрессии днРНК и уровня сигнала пика (далее «-» и «+» пики соответственно). Обнаружение «+» пика соответствует предположению, что днРНК участвует в установке модификации гистона, тогда как обнаружение «-» пика соответствует предположению, что днРНК участвует в удалении модификации гистона.

Чтобы оценить, насколько хорошо предсказания согласуются с опубликованными на данный момент результатами экспериментальных исследований, были отобраны днРНК и соответствующие им расширенные на +/- 25 тысяч п.н. «-» и «+» – гистоновые пики (положительно и отрицательно скоррелированные), у которых доля соответствующих пиков, поддерживаемых контактами хотя бы по одной из гистоновых меток, больше 0.4. Далее мы посчитали правосторонний и левосторонний тест Фишера отдельно для каждой днРНК и гистоновой метки (например, таблица сопряженности для пары «PVT1–H3K27ac» – табл. 1).

Таблица 1. Таблица сопряженности для расчета правостороннего и левостороннего теста Фишера, например, для пары днРНК PVT1 – пики гистоновой метки H3K27ac»

	Общее количество «-» пиков H3K27ac	Общее количество «+» пиков H3K27ac
Суммарное количество «+» и «-» пиков, поддерживаемых контактами PVT1	Количество поддерживаемых контактами PVT1 «-» пиков H3K27ac	Количество поддерживаемых контактами PVT1 «+» пиков H3K27ac
Суммарное количество «+» и «-» пиков, неподдерживаемых контактами PVT1	Количество неподдерживаемых контактами PVT1 «-» пиков H3K27ac	Количество неподдерживаемых контактами PVT1 «+» пиков H3K27ac

2.3.2. Данные Red-ChIP

При рассмотрении днРНК PVT1 в качестве примера интеграции HiMoRNA и RNA-Chrom мы дополнительно валидировали ее контакты с хроматином с помощью внешних данных Red-ChIP [170], которые доступны в Gene Expression Omnibus под номером GSE174474, образцы GSM5315228 и GSM5315229 (клеточная линия hES). Метод Red-ChIP фиксирует контакты РНК с хроматином, опосредованные белком EZH2, компонентом комплекса PRC2, который устанавливает, в том числе гистоновую модификацию H3K27me3. Первичную обработку этих данных проводили в соответствии с протоколом, использованным в базе данных RNA-Chrom. Далее мы определили участки генома, обогащенные контактами днРНК PVT1 с хроматином, с помощью программы BaRDIC (--qval_type all; --qval_threshold 1) [142]. Таким образом, мы обнаружили 3242 потенциально функциональные геномные области, в которых связывание PVT1 опосредовано белком EZH2.

2.4. Сравнительный анализ[6]

2.4.1. Данные

Данные РНК-хроматинового интерактома человека и мыши были взяты из базы данных RNA-Chrom [133]. Из всех данных ATA мы оставили только те, для которых были найдены данные

[6] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **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).

секвенирования РНК (RNA-seq) той же клеточной линии. При наличии более двух реплик в данных ATA мы отбирали две более полные реплики. Данные RNA-seq были взяты из базы данных GEO и обработаны аналогично процедуре обработки данных ATA, описанной в RNA-Chrom. Список данных представлен в приложении B в табл. B.1 и B.2. В анализ были включены только те РНК, которые демонстрировали больше 1000 контактов с хроматином в каждой реплике, что обеспечивает достаточную статистическую мощность для выявления «пиков» (участков генома, обогащенных контактами данной РНК с хроматином) программой BaRDIC [142]. Рибосомальные РНК были исключены из анализа. Например, при применении данного фильтра в экспериментах «RADICL, ES (NPM)» и «RADICL, ES (ActD)» остается меньше 1000 РНК и меньше 50% контактов от изначального размера отобранных реплик (см. приложение B, рис. B.1). Учитывая значительную перепредставленность ближних контактов (РД-скейлинг), в дальнейших анализах в данной статье не включались взаимодействия, находящиеся ближе 1 Мб от генов, кодирующих соответствующие РНК.

2.4.2. Использование BaRDIC, выбор порога

Как и большинство полногеномных данных, данные РНК-хроматинового интерактома характеризуются высоким уровнем неспецифического сигнала («шума»). Для выявления значимых взаимодействий применяются специализированные алгоритмы обнаружения пиков, направленные на идентификацию статистически значимых кластеров взаимодействий в конкретных геномных локусах.

Для выделения пиков в данных в настоящей работе мы используем разработанный нами ранее алгоритм BaRDIC [142], который учитывает РД-скейлинг и степень открытости хроматина.

Этот алгоритм для каждого локуса использует вероятностную оценку принадлежности контактов в локусе хроматина к пику или шуму. Далее, применяется поправка на множественное тестирование методом Бенджамини–Хохберга (FDR, частота ложных обнаружений), которая контролирует долю ложноположительных результатов на основе фонового распределения. Однако в нашем случае значительное перекрытие распределений сигнала и шума приводит к потере значительной доли истинных взаимодействий при использовании строгого порога на FDR. Чтобы избежать данную проблему, мы использовали гибкий критерий отбора – для каждой РНК выбирали топ-10% пиков с наименьшим FDR. Поскольку размеры пиков могли достигать десятков тысяч п.н. из-за разреженности данных, все сравнения проводились на уровне отдельных контактов, пересекающих эти пики.

Для анализа данных ATA BaRDIC был запущен с параметрами по умолчанию. При обработке экспериментов OTA, обладающих лучшим покрытием контактами, были выставлены следующие параметры: --trans_min 400 п.н.; --cis_start 100 п.н.; --trans_step 50 п.н. Фон

рассчитывали по input-данным с конвертацией в BedGraph, размер окна составлял 1000 п.н.

2.4.3. Хроматиновый потенциал

Практически во всех работах, посвященных экспериментам ATA, отмечено, что количество контактов РНК с хроматином линейно зависит от уровня экспрессии соответствующей РНК [11,12,15,16,33]. Нормировка на уровень экспрессии позволяет выделить РНК, демонстрирующие повышенную склонность к взаимодействию с хроматином – те молекулы, чья частота контактов существенно превышает ожидаемую при данном уровне экспрессии.

Для оценки склонности РНК контактировать с хроматином мы вводим понятие «хроматиновый потенциал». Пусть N_c – суммарное количество контактов отобранных РНК в эксперименте ATA с учетом фильтра на РД-скейлинг; N_e – полное число уникально картированных и проаннотированных генами чтений в эксперименте RNA-seq; n_c^i – число контактов с учетом фильтра на РД-скейлинг конкретной i-й РНК в эксперименте ATA; n_e^i – число чтений конкретной i-й РНК в эксперименте RNA-seq. Для сравнения этих наблюдений применим Z-тест пропорций. Для каждой i-й РНК вычислим Z-статистику (Z_i):

$$Z_i = \frac{p_c^i - p_e^i}{\sqrt{p_i(1-p_i)\,(1/N_c + 1/N_e)}};\;\; p_c^i = \frac{n_c^i}{N_c};\; p_e^i = \frac{n_e^i}{N_e};\; p_i = \frac{n_c^i + n_e^i}{N_c + N_e}$$

Статистика Z имеет стандартное нормальное распределение, и поэтому по этой статистике можно оценить *p-value* и FDR Бенджамини–Хохберга. Значение Z-статистики будем называть хроматиновым потенциалом. Хроматиновый потенциал является более адекватной мерой, чем простое отношение числа контактов к уровню экспрессии, поскольку он учитывает статистическую значимость отклонения. Отношение числа контактов к уровню экспрессии сильно смещено в сторону РНК с низким покрытием в данных RNA-seq, для которых знаменатель дроби (уровень экспрессии) оценивается с большой ошибкой, что приводит к большому разбросу значений (до шести порядков; см. приложение В, рис. В.2).

Однако надо иметь в виду следующие обстоятельства. Во-первых, для такого анализа нужно цепь-ориентированное тотальное секвенирование РНК с деплецией рибосомальной РНК. Во-вторых, этот анализ применим только для длинных РНК, поскольку стандартные данные RNA-seq не позволяют адекватно оценить уровень экспрессии РНК короче 100 нуклеотидов [171].

2.4.4. Воспроизводимость (конкордантность) контактов в репликах

Для оценки случайности совпадений можно использовать простейшую модель. Разобьем геном на непересекающиеся фрагменты (бины) фиксированного размера (*bin* п.н.) и предположим, что РНК контактирует с геномной ДНК равномерно, тогда в одной реплике эксперимента вероятность попадания хотя бы одного контакта в бин оценим как: $p_{bin}^e(i) = n_i^e/N_{bin}$, где i – номер

РНК, *e* – номер реплики, n_i^e – число бинов с контактами i-й РНК, N_{bin} – полное число бинов, на которое был разбит соответствующий геном. Здесь мы пренебрегаем смещениями в данных, в частности доступностью хроматина, и считаем, что размер бина достаточно мал. Чтобы избежать влияния РД-скейлинга, мы выбираем бины, отстоящие от гена i-й РНК дальше, чем на 1 Мб. Тогда вероятность того, что в один бин попадут контакты i-й РНК из двух реплик (*a* и *b*) будет равна $p_{bin}(i) = p_{bin}{}^a(i) \cdot p_{bin}{}^b(i)$. Можно сделать грубую оценку вероятности наблюдения *k* совпадающих бинов с помощью распределения Бернулли:

$$P^i_{obs} = C^k_{N_{bin}} p_{bin}(i)^k \cdot (1 - p_{bin}(i))^{N_{bin}-k}$$

Это позволяет сделать вероятностную оценку соответствия реплик или экспериментов. Определим $\lambda(i) = (n_i^a n_i^b / N_{bin})$. Для $\lambda \geq 10$ можно применить нормальное приближение и оценить вероятность такого события при условии, что реплики имеют независимые контакты:

$$P^i_{obs} \simeq \mathcal{N}\left(\lambda, \sqrt{\lambda}\right); \;\; P(X \geq k) \approx 1 - \Phi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right)$$

Для $\lambda < 10$ используем приближение Пуассона:

$$P(X \geq k) \approx 1 - F_{Poisson}(k-1, \lambda)$$

Глава 3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1. База данных RNA-Chrom[7]

Сравнительный анализ данных РНК–хроматинового интерактома представляет большой научный интерес. Поскольку не существует стандартного протокола обработки этих данных и существующие базы данных содержат очень небольшое количество полногеномных данных РНК–хроматинового интерактома, сложно провести сравнительный анализ соответствующих данных, доступных в открытом доступе.

Для решения этой задачи мы разработали специализированную аналитическую базу данных (*https://rnachrom2.bioinf.fbb.msu.ru*), которая содержит 213 наборов данных РНК-хроматинового интерактома типа ATA и OTA. Мы стандартизировали протокол обработки данных и реализовали его, начиная с сырых чтений (рис. 8). База данных RNA-Chrom [133] позволяет пользователю не только скачать данные, обработанные по единому протоколу, и подробные метаданные, но и выполнить различный анализ и сравнение данных в режиме реального времени (см. раздел «3.1.4.

[7] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **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, – pp. baad025. EDN: YEKQIZ. Импакт-фактор 3,6 (JIF) (1.02/0.40).

Функционал базы данных RNA-Chrom»):

- он может выбрать РНК и посмотреть, с какими геномными локусами/генами она контактирует в разных экспериментах (анализ «от РНК»);
- он может выбрать геномный локус и получить список контактирующих с ним РНК, после чего детально изучить характер контактируемости любой из этих РНК с выбранным локусом (анализ «от ДНК»).

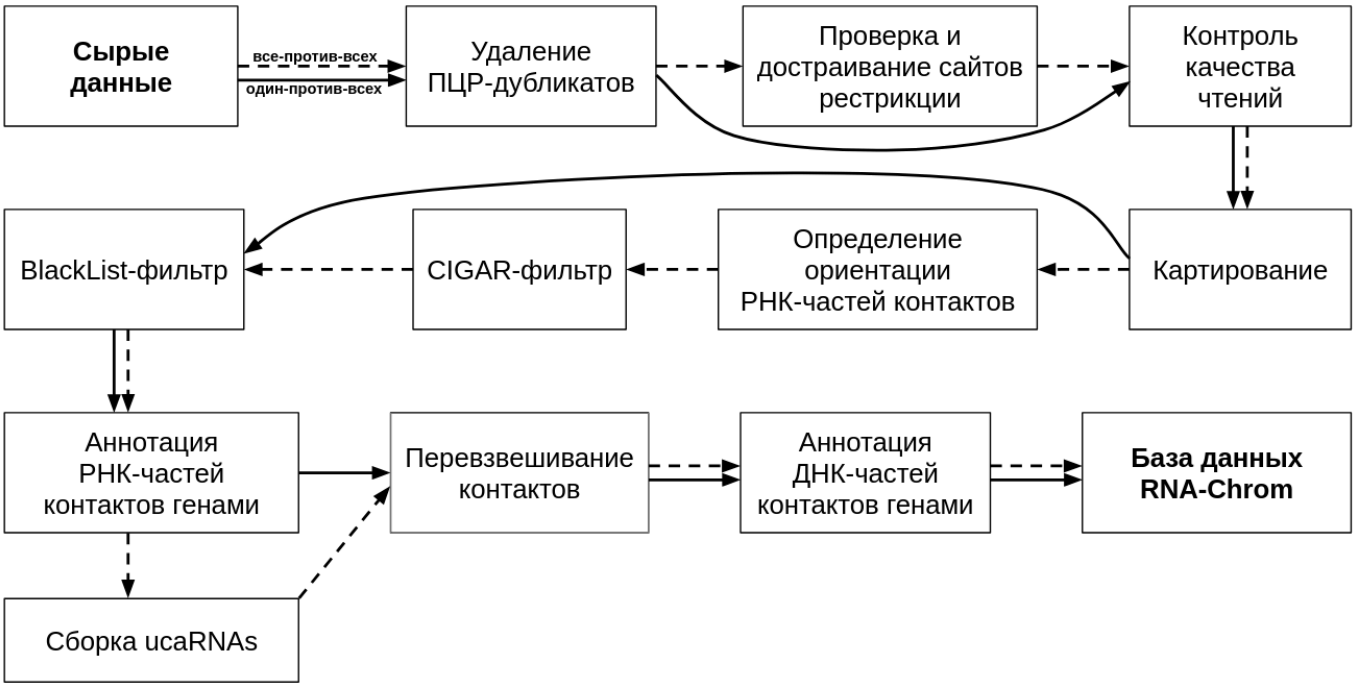


Рисунок 8. Протокол обработки данных РНК-хроматиновых взаимодействий. Пунктирные стрелки соответствуют этапам обработки данных ATA, а сплошные стрелки относятся к данным OTA.

Также пользователь может просмотреть карты контактов в UCSC Genome Browser [135], чтобы изучать их детальнее и сравнивать с другими данными, например, данными метилирования ДНК и гистоновых модификаций. Мы считаем, что база данных RNA-Chrom позволит исследователям выйти на более систематический уровень работы с РНК-хроматиновым интерактомом, что поможет расширить понимание биологической роли некодирующих РНК в самых разных процессах.

Для человека были собраны 13 экспериментов ATA и 64 эксперимента OTA для 33 РНК, для мыши – 7 экспериментов ATA и 125 экспериментов OTA для 35 РНК. Дополнительно в базу данных RNA-Chrom были включены по два эксперимента ATA для *Sus scrofa* и *Anolis carolinensis*. В аналитических разделах диссертации использовались только данные человека и мыши, поскольку для свиньи и анолиса число доступных экспериментов было невелико и отсутствовали данные OTA. Поскольку отрицательные контроли были доступны не для всех экспериментов, они не были включены в универсальный протокол обработки данных и, следовательно, в базу данных. Подводя итог, база данных RNA-Chrom содержит более 5 миллиардов РНК-хроматиновых

контактов и 232870 человеческих и 88914 мышиных генов. Общая генная аннотация включает в себя общедоступные генные аннотации (77743 гена для человека и 74581 ген для мыши), а также кластеры неаннотированных РНК-частей (155127 ucaRNAs для человека и 14333 ucaRNAs для мыши).

3.1.1. Предобработка данных РНК-хроматинового интерактома

Существует множество подходов к обработке данных РНК-хроматинового интерактома, но каждый из них специализирован под данные, полученные определенным экспериментальным методом. Например, авторы базы данных LnChrom использовали протокол из статьи ChIRP [6], в то время как мы будем основываться на протоколе, примененном в эксперименте Red-C [16]. В нашем протоколе сырые данные последовательно проходят этапы, указанные на рис. 8 (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.1.2. Универсальный протокол обработки данных в базе данных RNA-Chrom»). Важно отметить, что все реплики были обработаны независимо, начиная с шага «Удаление ПЦР-дубликатов» и до «BlackList-фильтр» включительно.

Согласно сводной статистике, для данных ATA по сравнению с данными OTA наибольшее количество чтений отфильтровывается на этапе «Картирование» (рис. 9). Это связано с тем, что для данных ATA мы требуем, чтобы РНК- и ДНК-части каждого контакта были правильно картированы, в противном случае они будут удалены. Что касается данных OTA, среди них есть несколько наборов данных, которые не заслуживают доверия. Например, данные GSM3073889 и GSM3073888 (человеческая lincDUSP РНК) имеют меньше 4000 сырых чтений и ноль пиков MACS2. Однако, мы решили не исключать их из базы данных.

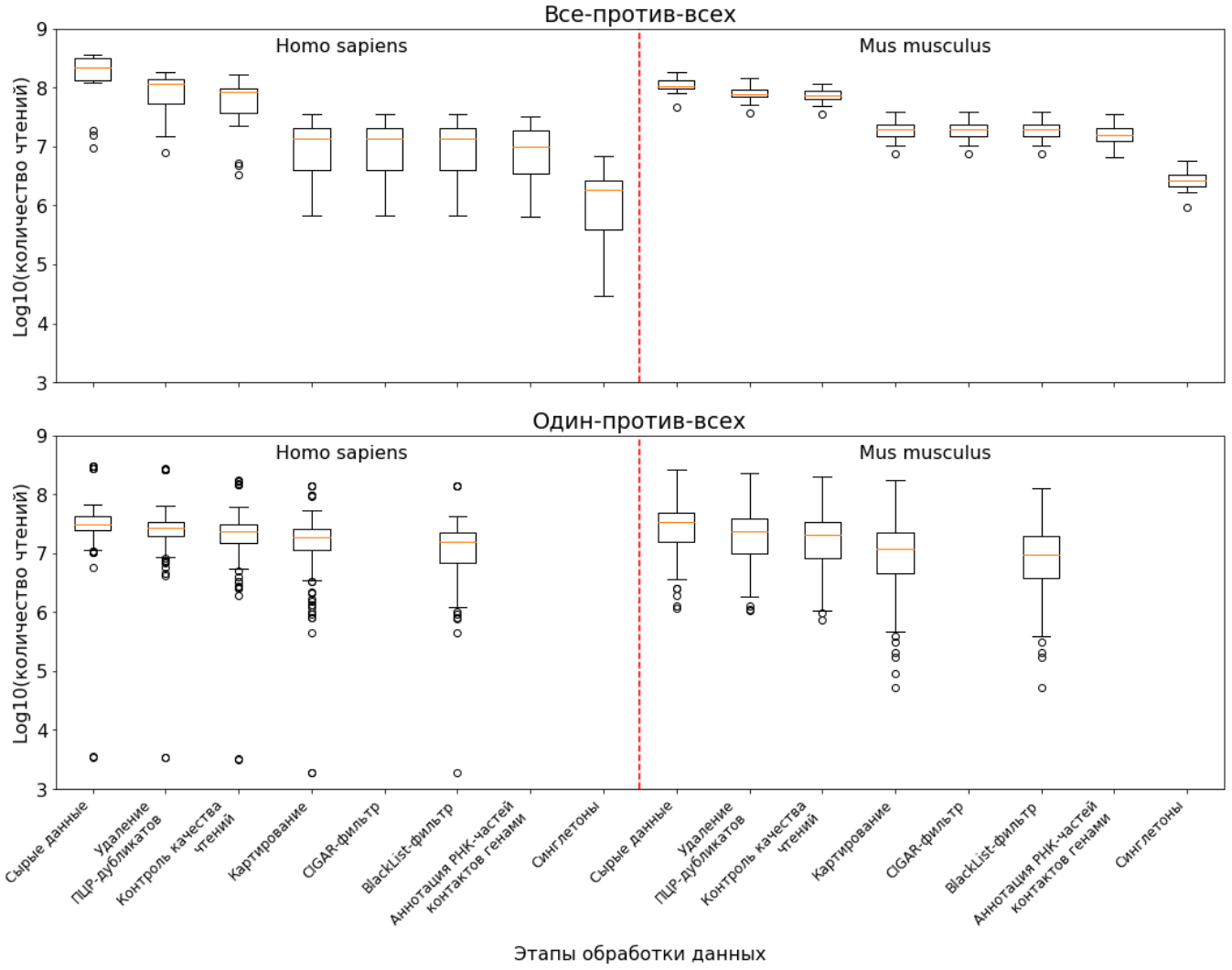


Рисунок 9. Распределение количества чтений в наборах данных, оставшихся после соответствующего шага обработки и всех предыдущих. Верхняя панель: диаграммы типа «ящик с усами», построенные на основе данных ATA («все-против-всех»). Нижняя панель: диаграммы типа «ящик с усами», построенные на основе данных OTA («один-против-всех»).

3.1.2. Разработка БД и структура данных

База данных RNA-Chrom должна не только хранить большое количество данных РНК-хроматинового интерактома, но и позволять анализировать все эти данные в режиме реального времени. В связи с этим мы решили реализовать ее с помощью системы управления базами данных ClickHouse (*https://clickhouse.com*).

База данных содержит 6 таблиц типа «MergeTree» (рис. 10):

1) таблица «genes» с человеческими и мышиным генами;
2) таблица «contacts» с человеческими и мышиным РНК-ДНК контактами;
3) таблица «rna_dna_experiments_3» с метаданными 213 экспериментов;
4) таблица «processing_metadata_2» с метаданными о каждой стадии обработки данных по всем экспериментам;

5) таблица «temporary_type_chr» с человеческими и мышиными каноническими хромосомами;
6) сгруппированная по генам и экспериментам таблица РНК-ДНК контактов – «experiment_genes_14».

Данные в этих таблицах отсортированы по первичному ключу (или ключам), по этим колонкам автоматически будет создан индекс, позволяющий быстрее находить необходимые данные .

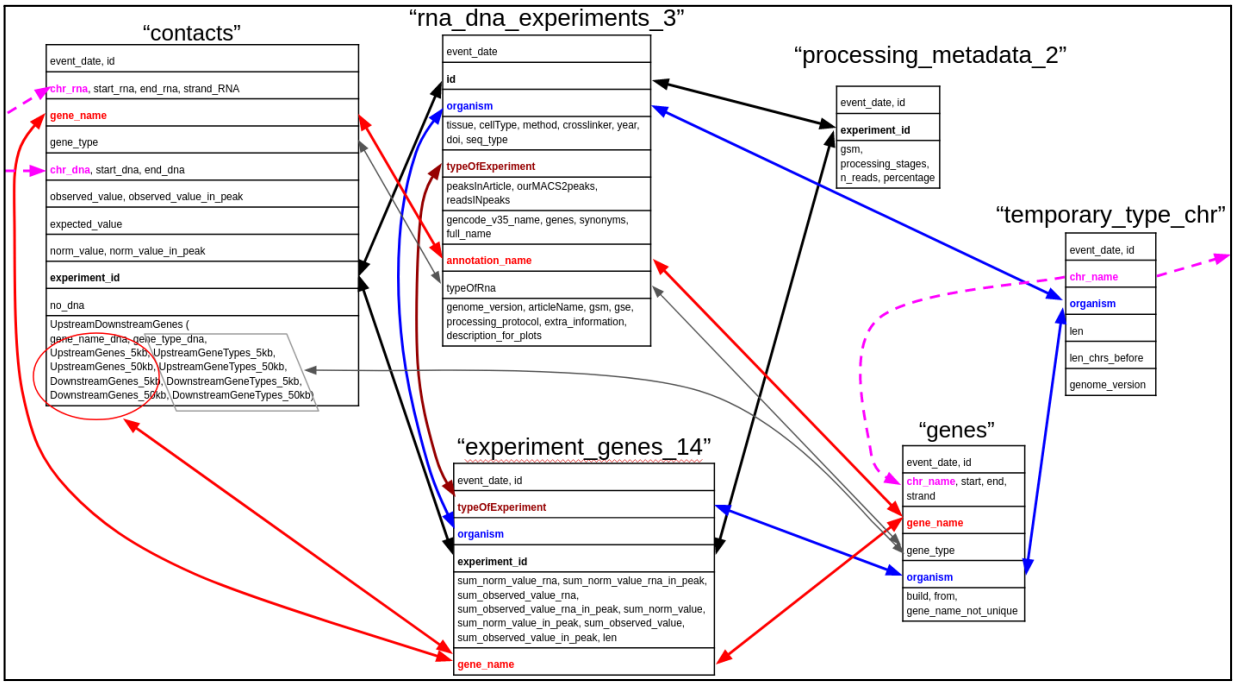


Рисунок 10. Структура базы данных RNA-Chrom.

3.1.3. Разработка веб-сервиса RNA-Chrom

Наш веб-сервис состоит из трех частей (рис. 11, см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.1.3 Веб-сервис RNA-Chrom»): базы данных, «бэкенда» (связывает фронтенд и БД) и «фронтенда» (отвечает за пользовательский интерфейс). В результате веб-сервис работает следующим образом:

1) пользователь открывает веб-страницу в браузере;
2) нажимает на какую-нибудь кнопку;
3) GET\POST-запрос приходит на бэкенд;
4) с бэкенда отправляется http-запрос в базу данных;
5) из БД на бэкенд приходят соответствующие данные;
6) эти данные предобрабатываются на бэкенде и в json-формате отправляются пользователю;
7) пользователь у себя в браузере видит эти данные в таблицах или графиках.

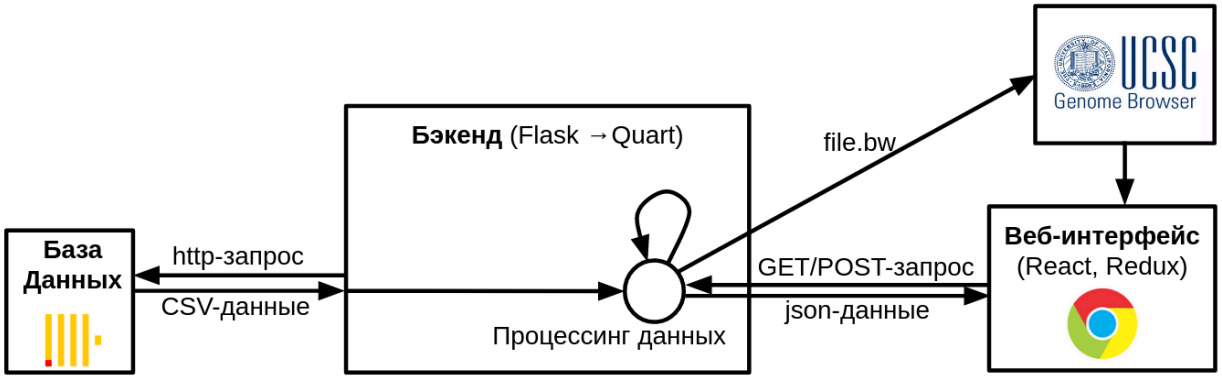


Рисунок 11. Схема работы веб-сервиса RNA-Chrom.

3.1.4. Функционал базы данных RNA-Chrom

С помощью RNA-Chrom пользователь может выполнять два типа анализа данных РНК-хроматинового интерактома, которые могут переходить друг в друга, в режиме реального времени. Первый тип анализа мы назвали «от РНК», поскольку первым шагом является выбор интересующей РНК. Этот анализ позволяет пользователю ответить на вопрос «С какими геномными локусами контактирует выбранная РНК?». В то время как второй тип анализа, «от ДНК», начинается с выбора интересующего геномного локуса, и пользователь получит ответ на вопрос «Какие РНК контактируют с выбранным целевым локусом?». Эти типы анализа включают следующее:

1) таблица с РНК, которые контактируют со всем геномом (анализ «от РНК») или с выбранным геном/локусом (анализ «от ДНК»), с соответствующими характеристиками их контактируемости (рис. 12, 13В, 14В);
2) таблица генов, с которыми выбранная РНК контактирует напрямую или в 50’000-нуклеотидной окрестности (рис. 12, 13Д, 14Д);
3) три типа аналитических графиков (рис. 12):
    a) распределение плотности контактов на целевом локусе или по всему геному (рис. 13Г, 14Г);
    b) изменение плотности контактов в зависимости от расстояния между геном, кодирующим РНК, и целевыми геномными локусами («РД-скейлинг») (рис. 13Г);
    c) распределение РНК-частей контактов по телу гена, кодирующего соответствующую РНК (рис. 13Е, 14Е);
4) возможность просмотра карт контактов в UCSC Genome Browser (рис. 12, 13Ж, 14Ж).

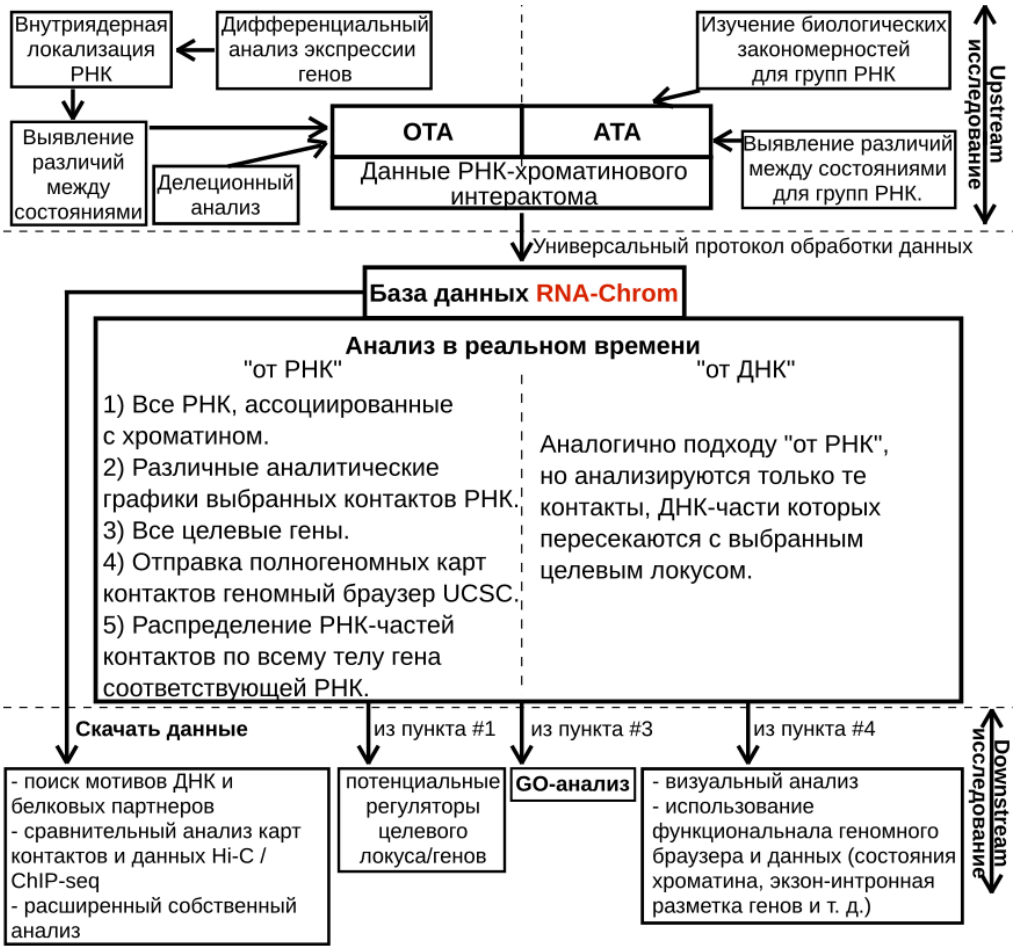


Рисунок 12. Функционал базы данных RNA-Chrom и возможный последующий анализ.

Кроме того, база данных RNA-Chrom позволяет пользователю скачивать для собственного исследования или последующего анализа:

1) обработанные единым протоколом данные о взаимодействиях РНК с хроматином (рис. 19, подробнее в разделе «4.1.5. Дополнительные веб-страницы»);
2) карты контактов отдельных РНК (рис. 13Г, 14Г: «Contacts Summary»);
3) таблицу генов, с которыми выбранная РНК контактирует напрямую или в 50’000-нуклеотидной окрестности (рис. 13Д, 14Д: «All target genes»);
4) таблицу с РНК, которые контактируют со всем геномом или с выбранным геном/локусом (рис. 13В, 14В: «Complete table of RNAs»);
5) дополнительные метаданные.

3.1.4.1. Анализ «от РНК»

Используя веб-интерфейс, пользователь может проанализировать контакты любой РНК из нашей общей аннотации. Для проведения анализа «от РНК» необходимо выполнить следующие шаги (рис. 13):

1) Пользователь должен выбрать анализ «от РНК» на стартовой странице (рис. 13А). Страница

«RNA-Chrom: where does the selected RNA contact with chromatin?» откроется в новой вкладке.

2) Следующий шаг зависит от того, знает ли пользователь имя интересующей его РНК или он хочет просмотреть таблицу со всеми РНК и с соответствующими метриками контактируемости («Complete table of RNAs»), чтобы выбрать из нее РНК для дальнейшего анализа.
    a) Например, если пользователь интересуется NR2F1-AS1 РНК (*Homo sapiens*), он должен ввести имя РНК в поле «Select RNA», выбрать нужную РНК в выпадающем списке и затем нажать кнопку «GRAPHICAL SUMMARY» (рис. 13Б). Страница «GRAPHICAL SUMMARY» откроется в новой вкладке, которая будет содержать список экспериментов, содержащих контакты соответствующей РНК, аналитические интерактивные графики и различные дополнительные опции для дальнейшего анализа.
    b) Пользователь может нажать кнопку «Просмотр всех РНК» («BROWSE ALL RNAS»). Работая с таблицей «Complete table of RNAs» можно использовать различные фильтры: искать по имени РНК или подслову («Search by RNA name»), выбрать из выпадающего списка РНК по имени («Select RNA names»), выбрать биотип РНК («Select RNA types»), ввести геномные координаты («Genomic loci») и т. д. (рис. 13В). Например, пользователь может заполнить фильтры «Select RNA types», «Minimum gene length» и «Organism» значениями «Xrna», «1000» и «Mus musculus» соответственно, а затем нажать кнопку «Применить фильтры» («APPLY FILTERS»). Чтобы перейти на страницу с аналитическими графиками («Graphical Summary»), пользователь должен нажать на название интересующей его РНК (например, X_17_3984_a_mm10, так как эта РНК имеет самый большой «CPKM»).
3) Страница «Graphical Summary» (рис. 13Г) состоит из одной таблицы («Contacts Summary») и трех аналитических графиков («Contacts Distribution», «Comparative Heatmap» и «Distance Distribution»).
4) В таблице «Contacts Summary» пользователь может выбрать, например, «Exp.ID: 14» (RADICL, mES R08), а затем нажать на одну из четырех кнопок, например, «Все целевые гены» («ALL TARGET GENES»). Страница «Все целевые гены» («All target genes») откроется в новой вкладке.
5) На странице «All target genes» (рис. 13Д) пользователь может продолжить анализ в режиме «от ДНК» (будет описан дальше) и посмотреть, например, какие РНК взаимодействуют с тем или иным целевым геном, или использовать фильтры для получения списка генов, которые можно скачать для последующего самостоятельного анализа (например, «Gene

Ontology»).

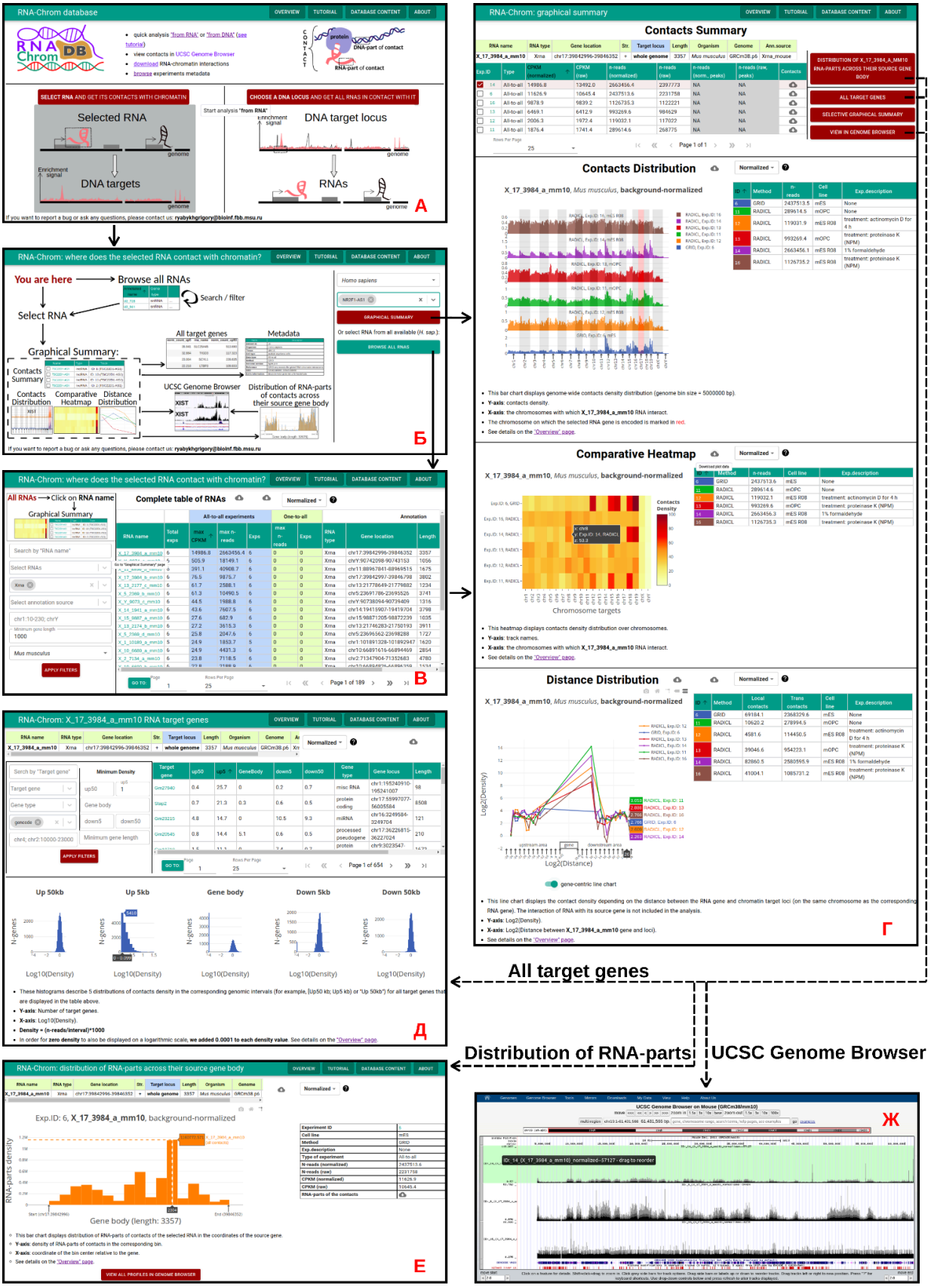


Рисунок 13. Анализ «от РНК» для X_17_3984_a_mm10 РНК (*Mus musculus*). Стрелки отражают направление исследовательского процесса. А. Пользователь выбирает анализ «от РНК». Б. Затем пользователь либо вводит интересующее название РНК в поле «Select RNA», либо нажимает

кнопку «BROWSE ALL RNAS». В. «Complete table of RNAs». Здесь пользователь выбирает РНК для анализа. Г. Страница «Graphical Summary» состоит из «Contacts Summary» и трех аналитических графиков. Выбрав одну или несколько карт контактов X_17_3984_a_mm10 РНК в таблице «Contacts Summary», пользователь продолжает их анализ, нажимая одну из четырех кнопок, расположенных справа от таблицы. Д. Страница «All target genes» отображает связь контактов с генами и фланкирующими их областями. Применив несколько фильтров, пользователь скачивает список целевых генов. С этого момента пользователь может переключиться на анализ «от ДНК». Для этого пользователь нажимает на интересующий целевой ген. Е. Распределение РНК-частей X_17_3984_a_mm10 РНК по телу кодирующего ее гена (может отражать экзон-интронную структуру, множественные изоформы транскрибированного гена и т. д.) Пользователь может отправить распределения для всех экспериментов в UCSC Genome Browser для более подробного изучения или скачать их. Ж. Пользователь отправляет карты контактов X_17_3984_a_mm10 РНК (ДНК-части) в UCSC Genome Browser, если он хочет просмотреть их в более высоком разрешении или визуально сопоставить их с геномными аннотациями (наборами генов, эпигенетическими метками и т. д.) или данными (ChIP-seq, Hi-C и т. д.).

3.1.4.2. Анализ «от ДНК»

Этот тип анализа позволяет пользователю найти все РНК, которые контактируют с выбранным геном или локусом. Для проведения анализа «от ДНК» необходимо выполнить следующие шаги (рис. 14):

1) Пользователь должен выбрать анализ «от ДНК» на стартовой странице (рис. 14А). Страница «RNA-Chrom: what RNAs contact with the selected target locus?» откроется в новой вкладке.
2) Пользователь должен нажать на кнопку «Выбрать ДНК-локус» («CHOOSE A DNA LOCUS»), расположенную в правой части страницы (рис. 14Б).
3) Он может выбрать организм «*Mus musculus*», ввести, например, точные или приблизительные координаты HoxA-кластера (chr6:52’015’389-52’270’886) и нажать кнопку «Применить» («APPLY»). Другой способ – выбрать локус по названию гена (рис. 14Б).
4) После выбора локуса появляется список РНК, которые контактируют с выбранным локусом. Для работы с этим списком можно использовать различные фильтры, а также сортировать список различными способами (рис. 14В).
    a) Например, пользователь может заполнить фильтры «Select RNA types» и «Select annotation source» значениями «lincRNA» и «gencode» соответственно и нажать кнопку «Применить фильтры» («APPLY FILTERS»).
    b) Он может выбрать нормализацию «Norm. & in peaks (one-to-all)» и отсортировать таблицу по столбцу «max n-reads», который относится к данным ОТА.
    c) Чтобы перейти на страницу «Graphical Summary», необходимо нажать на название интересующей РНК (например, Halr1, поскольку эта РНК находится в верхней части таблицы и, как известно, участвует в модуляции индукции HoxA [34]).

5) Страница «Graphical Summary» (рис. 14Г) состоит из таблицы «Contacts Summary» и аналитического графика «Contacts Distribution».
6) В таблице «Contacts Summary» можно выбрать, например, «Exp.ID: 102» (ChIRP, 46C ES, treatment: LIF withdrawal & 2uM RA – Day 1).
7) Пользователь может выбрать нормализацию «Norm. & in peaks (one-to-all)» и нажать кнопку «VIEW IN GENOME BROWSER». UCSC Genome Browser откроется в новой вкладке (рис. 14Ж).
8) Если нажать кнопку «Все контакты» («ALL CONTACTS») (рис. 14Г), то в новой вкладке откроется страница «Graphical Summary», соответствующая анализу «от РНК». Здесь пользователь может продолжить анализ в режиме «от РНК».
9) Пользователь может нажать кнопку «Все целевые гены» («ALL TARGET GENES»), выбрать нормализацию «Norm. & in peaks (one-to-all)» и заполнить фильтр «Search by target gene» значением «hox» (рис. 14Д). Затем он должен нажать кнопку «Применить фильтры» («APPLY FILTERS»). Как и ожидалось, пользователь увидит множество «Ноха»-генов. Со страницы «Все целевые гены» («All target genes») можно продолжить анализ в режиме «от ДНК» и посмотреть, какие РНК взаимодействуют с этим целевым геном или скачать список генов для последующего собственного анализа.

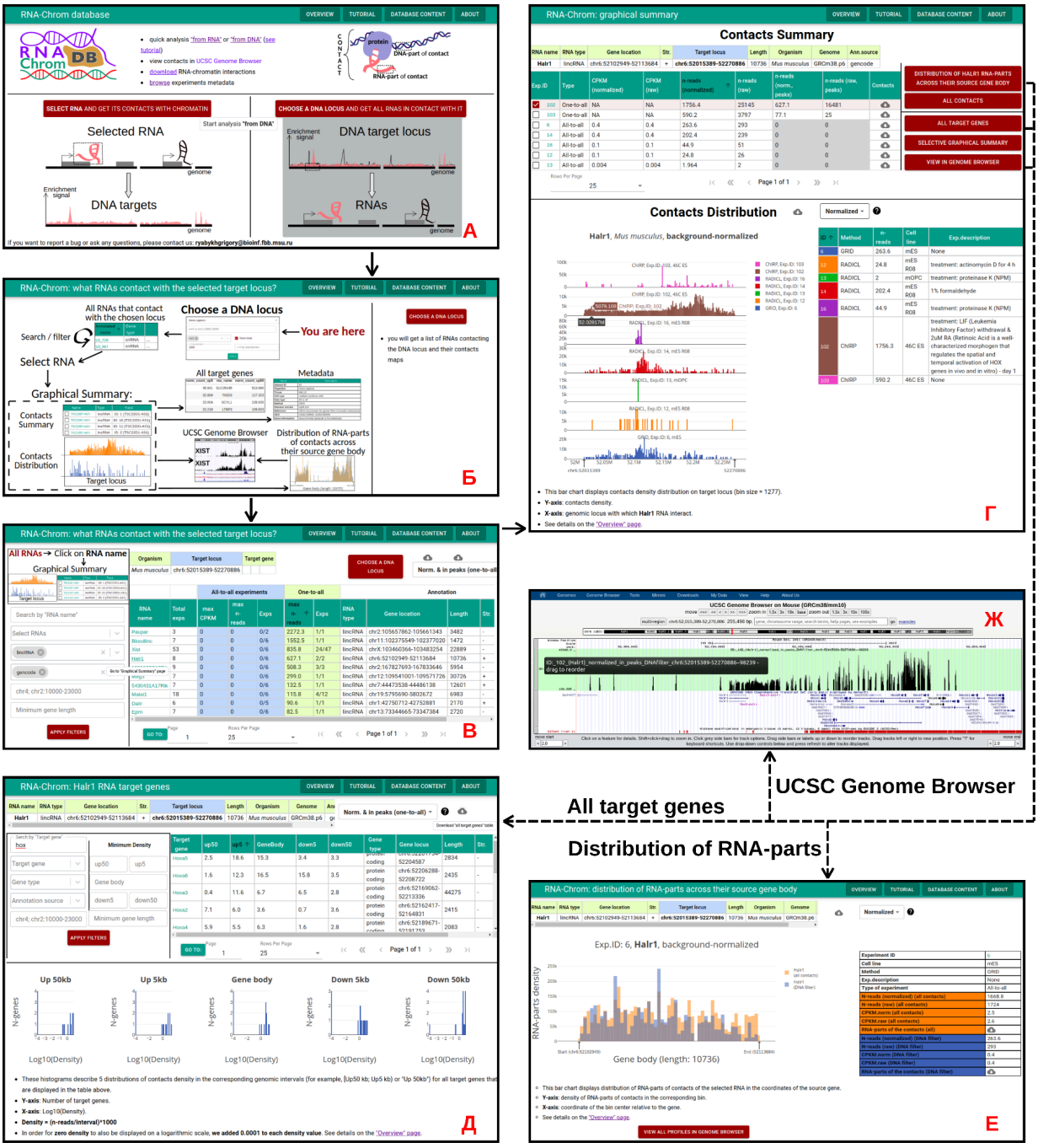


Рисунок 14. Анализ «от ДНК» для HoxA-кластера (chr6:52 015 389-52 270 886, *M. musculus*). Стрелки отражают направление исследовательского процесса. А. Пользователь выбирает анализ «от ДНК» и Б. нажимает кнопку «CHOOSE A DNA LOCUS», выбирает организм, вводит примерные координаты HoxA-кластера и нажимает кнопку «APPLY». В. Все РНК, которые контактируют с выбранным локусом, представлены в таблице. При необходимости пользователь применяет фильтры. Чтобы перейти на страницу «Graphical Summary», пользователь нажимает на интересующее название РНК (например, Halr1). Г. Страница «Graphic Summary» разделена на два блока: «Contacts Summary» и «Contacts Distribution». Пять кнопок, расположенных справа от таблицы «Contacts Summary», представляют пять вариантов дальнейшей работы с контактами Halr1. Ж. Чтобы получить более подробную информацию о локусе, пользователь отправляет карты контактов в UCSC Genome Browser. Д. Страница «All target genes» отображает связь контактов с генами, расположенными в выбранном локусе, а также фланкирующие их регионы. Список генов можно фильтровать разными способами. С этого момента пользователь продолжает анализ

способом «от ДНК». E. Гистограмма «Distribution of Halr1 RNA-parts across their source gene body» строится как для всех контактов, так и для контактов с целевым локусом. Эти распределения можно скачать или отправить в UCSC Genome Browser для более детального изучения.

С биологической точки зрения RNA-Chrom переводит анализ РНК-хроматиновых взаимодействий из режима описания отдельных контактов в режим приоритизации биологически осмысленных гипотез. До создания такого ресурса исследователь, как правило, был вынужден отдельно собирать опубликованные эксперименты, использовать несовместимые результаты их авторской обработки (или самому пытаться обработать сырые данные) и вручную сопоставлять найденные контакты с генами, эпигенетическими метками и другими геномными аннотациями. В RNA-Chrom эти действия сведены к единому сценарию работы с унифицированными данными: можно быстро определить, с какими локусами контактирует интересующая РНК, какие гены попадают в область ее контактов, какие еще РНК взаимодействуют с тем же локусом, и сопоставить эти данные с независимыми эпигенетическими и пространственными аннотациями в UCSC Genome Browser. Тем самым сервис позволяет быстрее переходить от списка контактов к формулированию проверяемых гипотез о возможной регуляторной функции конкретных РНК.

3.1.5. Дополнительные веб-страницы

RNA-Chrom, помимо стартовой веб-страницы (рис. 15) и страниц с аналитическими графиками и таблицами, включает в себя дополнительные страницы, содержащие подробную информацию о данных и метаданных, протоколе обработки, функционале, примерах «от РНК» / «от ДНК» анализах и авторах базы данных. Мы их создали для того, чтобы у пользователя было как можно больше информации о данном ресурсе, которая поможет легко и эффективно использовать его в исследовательских задачах.

На стартовую веб-страницу (рис. 15) пользователь попадает сразу, когда проходит по ссылке *https://rnachrom2.bioinf.fbb.msu.ru*. В самом верху веб-сайта можно заметить зеленоватого цвета полоску с 4 кнопками («OVERVIEW», «TUTORIAL», «DATABASE CONTENT», «ABOUT»), которые ведут на соответствующие страницы. Ниже перечислены возможности RNA-Chrom, а внизу изображены две инфографики, отражающие «от РНК» и «от ДНК» анализ РНК-хроматинового интерактома.

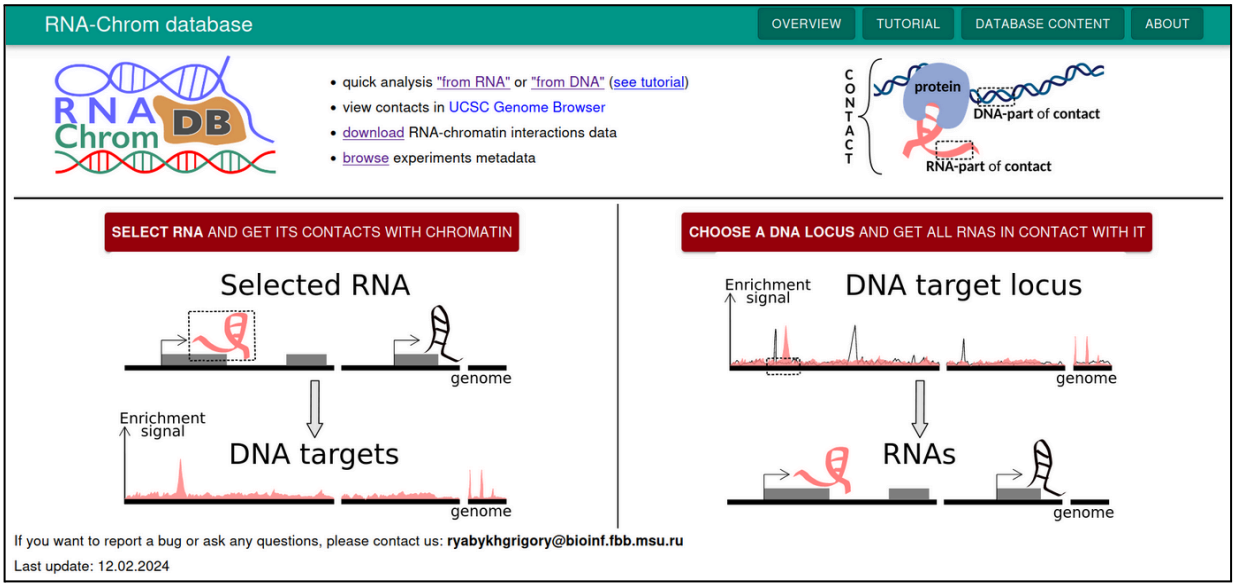
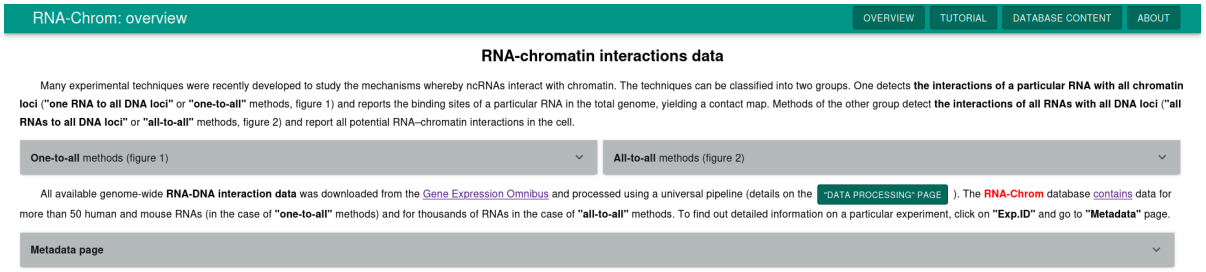


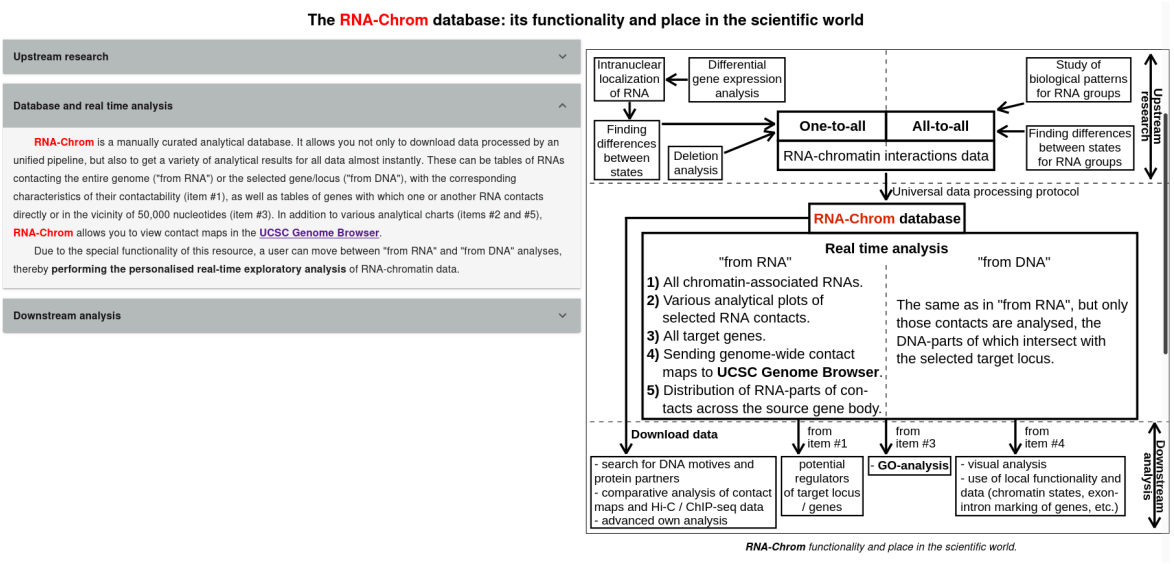
Рисунок 15. Стартовая страница RNA-Chrom БД.

Страница «Overview» состоит из трех частей: «RNA-Chromatin interactions data», «The RNA-Chrom database: its functionality and place in the scientific world», «Available types of data analysis». Первая часть (рис. 16А) описывает данные, которые представлены в данной БД, ссылается на страницу «Data processing» и содержит подробное описание страницы «Metadata». Вторая часть (рис. 16Б) – место RNA-Chrom в цепочке «Постановка биологической задачи» - «выбор объекта исследования» - «проведение эксперимента и получение данных» - «обработка данных» - «анализ данных» и ее функционал. Третья (рис. 16В) – какие функции доступны пользователю, какие формулы использовались для предварительной обработки данных для графиков и таблиц, и какую информацию пользователь может извлечь из этих графиков и таблиц.

А.

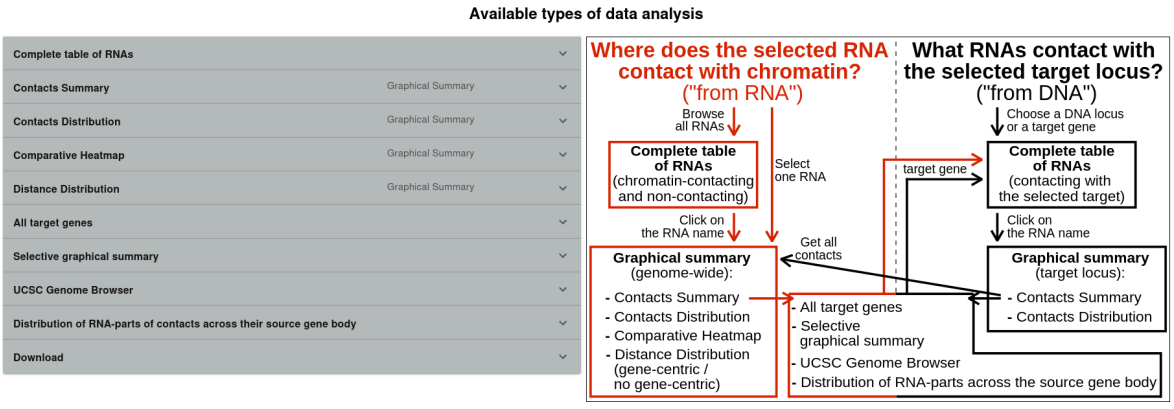


Б.



RNA-Chrom functionality and place in the scientific world.

В.

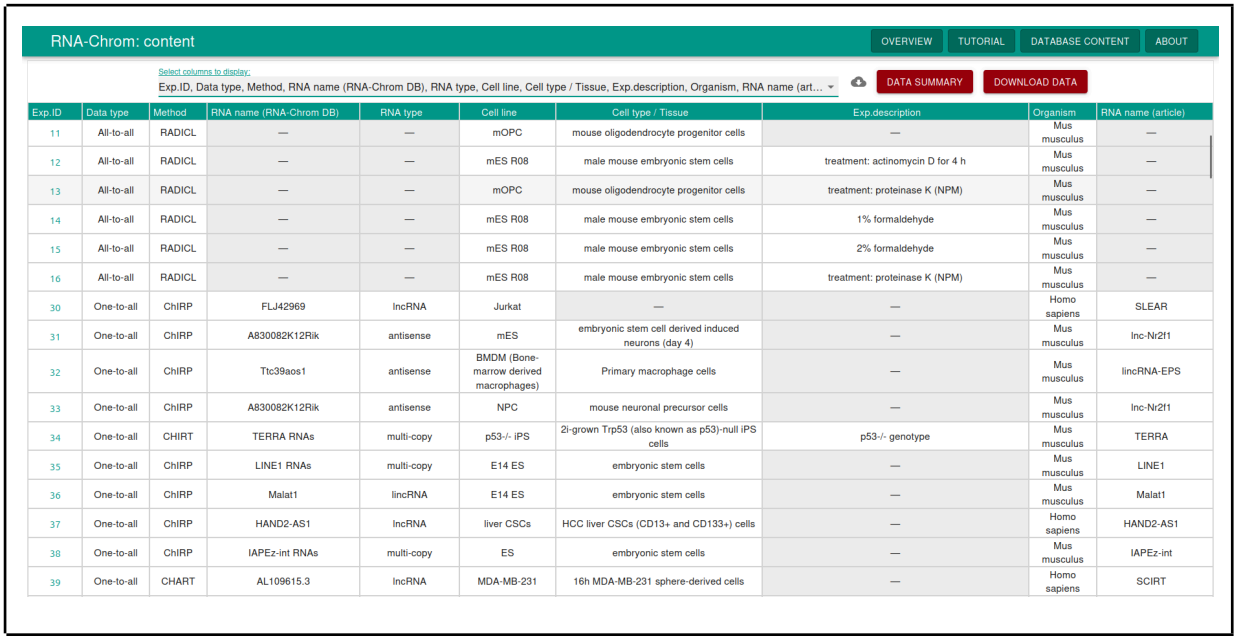


Full workflow of data analysis.

Рисунок 16. Страница «Overview». А. Первый раздел; Б. второй раздел; В. третий раздел.

Страница «Data processing» подробно описывает каждый этап обработки данных РНК-хроматинового интерактома. Страница «Tutorial» содержит примеры простого и продвинутого анализа «от РНК» и «от ДНК». Страница «About» содержит информацию об авторах этого проекта, благодарности коллегам и список статей нашей группы по теме РНК-хроматинового интерактома.

Страница «Content» (рис. 17) представляет собой таблицу с полной метаинформацией по всем экспериментам из базы данных RNA-Chrom. Здесь пользователь может нажать кнопку «DOWNLOAD DATA», чтобы скачать данные по каждому эксперименту (контакты со всеми нормализациями, синглтоны, пики и т. д.). Чтобы узнать подробную информацию по конкретному эксперименту, пользователь должен нажать на соответствующий «Exp.ID» и перейти на страницу «Metadata».



Exp.ID	Data type	Method	RNA name (RNA-Chrom DB)	RNA type	Cell line	Cell type / Tissue	Exp.description	Organism	RNA name (article)
11	All-to-all	RADICL	—	—	mOPC	mouse oligodendrocyte progenitor cells	—	Mus musculus	—
12	All-to-all	RADICL	—	—	mES R08	male mouse embryonic stem cells	treatment: actinomycin D for 4 h	Mus musculus	—
13	All-to-all	RADICL	—	—	mOPC	mouse oligodendrocyte progenitor cells	treatment: proteinase K (NPM)	Mus musculus	—
14	All-to-all	RADICL	—	—	mES R08	male mouse embryonic stem cells	1% formaldehyde	Mus musculus	—
15	All-to-all	RADICL	—	—	mES R08	male mouse embryonic stem cells	2% formaldehyde	Mus musculus	—
16	All-to-all	RADICL	—	—	mES R08	male mouse embryonic stem cells	treatment: proteinase K (NPM)	Mus musculus	—
30	One-to-all	ChIRP	FLJ42969	lncRNA	Jurkat	—	—	Homo sapiens	SLEAR
31	One-to-all	ChIRP	A830082K12Rik	antisense	mES	embryonic stem cell derived induced neurons (day 4)	—	Mus musculus	lnc-Nr2f1
32	One-to-all	ChIRP	Ttc39aos1	antisense	BMDM (Bone-marrow derived macrophages)	Primary macrophage cells	—	Mus musculus	lincRNA-EPS
33	One-to-all	ChIRP	A830082K12Rik	antisense	NPC	mouse neuronal precursor cells	—	Mus musculus	lnc-Nr2f1
34	One-to-all	CHIRT	TERRA RNAs	multi-copy	p53-/- iPS	2i-grown Trp53 (also known as p53)-null iPS cells	p53-/- genotype	Mus musculus	TERRA
35	One-to-all	ChIRP	LINE1 RNAs	multi-copy	E14 ES	embryonic stem cells	—	Mus musculus	LINE1
36	One-to-all	ChIRP	Malat1	lincRNA	E14 ES	embryonic stem cells	—	Mus musculus	Malat1
37	One-to-all	ChIRP	HAND2-AS1	lncRNA	liver CSCs	HCC liver CSCs (CD13+ and CD133+) cells	—	Homo sapiens	HAND2-AS1
38	One-to-all	ChIRP	IAPEz-int RNAs	multi-copy	ES	embryonic stem cells	—	Mus musculus	IAPEz-int
39	One-to-all	CHART	AL109615.3	lncRNA	MDA-MB-231	16h MDA-MB-231 sphere-derived cells	—	Homo sapiens	SCIRT

Рисунок 17. Страница «Content».

Страница «Metadata» состоит из трех частей: «All metadata information on a particular experiment & RNAs distribution», «Analytical plots», «Summary statistics of the data processing protocol». Первая часть (рис. 18А) описывает все метаданные по конкретному эксперименту и представленность каждого биотипа РНК в соответствующих данных. Вторая часть (рис. 18Б) – графики распределения контактов по биотипам РНК и по отдельным РНК. Третья часть (рис. 18В) – статистику по каждому этапу обработки данных. Пользователь может открыть страницу «Metadata», нажав на идентификатор эксперимента («Exp.ID») везде, где он появляется, например, на страницах «Graphical Summary», «All target genes» и других.

Рисунок 18. Страница «Metadata». А. Первый раздел; Б. второй раздел; В. третий раздел.

Заархивированные данные РНК-хроматинового интерактома по каждому эксперименту можно скачать со специальной веб-страницы (рис. 19). На нее можно попасть либо со стартовой страницы (рис. 15), перейдя по соответствующей гиперссылке, либо со страницы «Content» (рис. 17), нажав на кнопку «DOWNLOAD DATA». Доступная для скачивания таблица «Содержание

архивных данных» («Contents of archived data») отражает содержание каждого архива. В свою очередь в папках «All-to-all-data» и «One-to-all-data» хранятся непосредственно архивы с данными OTA и ATA соответственно.

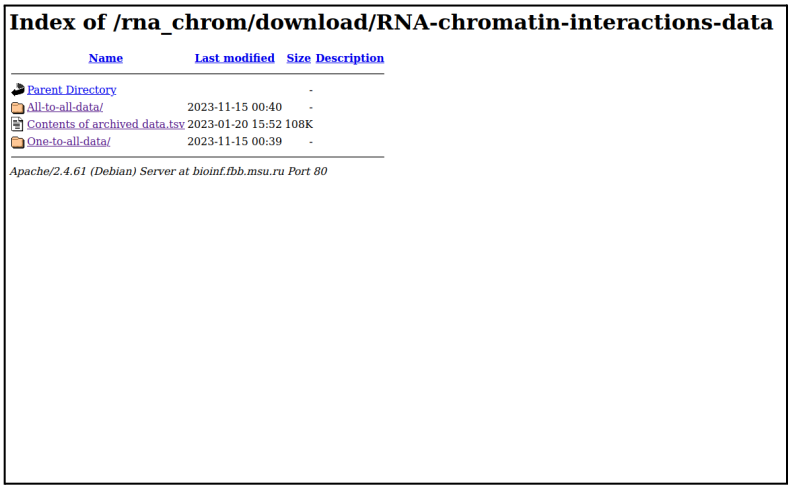


Рисунок 19. Веб-страница с заархивированными данными РНК-хроматинового интерактома.

3.2. ПЦР-дедуплексатор Fastq-dupaway[8]

3.2.1. Разработка и реализация программы

Программа для удаления ПЦР-дубликатов Fastq-dupaway предоставляет множество вариантов для гибкой настройки рабочего процесса. Программа работает с одноконцевыми или парноконцевыми входными данными, поддерживает форматы файлов FASTQ и FASTA и работает как с обычными, так и с gz-сжатыми файлами. Пользователь может выбрать один из двух алгоритмов дедупликации: (i) режим «sequence-based», который обеспечивает контроль над верхним пределом использования оперативной памяти (RAM) и логикой сравнения последовательностей, но требует примерно в 2 раза больше дискового пространства, чем входной файл; и (ii) режим «fast», разработанный для скорости и удаляющий только точные дубликаты, но не позволяющий устанавливать ограничение на RAM (см. приложение Б, рис. Б.1).

3.2.1.1. Режим «sequence-based»

Этот режим работает путем прямого сравнения последовательностей. Во время выполнения использование памяти контролируется таким образом, чтобы оставаться на уровне или ниже заданного пользователем порогового значения.

Сначала данные из входного файла (или файлов) сортируются по их последовательности.

[8] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: 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 (2025). EDN: VBEHEL. Импакт-фактор 3,9 (JIF) (0.88/0.20).

Программа использует вариант алгоритма «внешней сортировки», аналогичного сортировке слиянием, со временно́й сложностью *O(N*logN)*. В этом режиме программа создает временные файлы, общий размер которых примерно в 2 раза превышает размер входного файла. Установленное ограничение на объем оперативной памяти не влияет на общее использование дискового пространства, но напрямую влияет на производительность во время выполнения. На втором этапе программа считывает отсортированные входные данные и удаляет повторяющиеся записи за один проход. Объем данных, считываемых и записываемых во время операций ввода-вывода (I/O) при выполнении sequence-based алгоритма, в 3 раза превышает объем входных данных. Логика определения «дубликата» поддерживает три варианта:

- «tight» режим (включен по умолчанию). Удаляются только полные дубликаты последовательностей. Последовательности разной длины автоматически считаются недублированными.
- «loose» режим. Этот режим дает результаты, идентичные результатам FastUniq [154], причем оба инструмента чувствительны к порядку входных файлов. Однако Fastq-dupaway в «loose» режиме достигает этого за счет более эффективной с точки зрения использования оперативной памяти реализации потоковой обработки. При сравнении последовательностей последовательности разной длины считаются дубликатами, если более короткая последовательность точно соответствует префиксу более длинной последовательности.
- «tail-hamming» режим основан на предположении, что несоответствия, вызванные ошибками секвенирования, с большей вероятностью наблюдаются на концах прочтений, а не в их начале [172]. В этом режиме два прочтения считаются дубликатами, если они расположены рядом в отсортированном файле, и расстояние Хэмминга между ними не превышает заданный пользователем порог.

В случае парных прочтений логика обнаружения дубликатов та же, что и выше, с одним дополнением: для удаления пары Б как «левое», так и «правое» прочтения пары Б должны быть дубликатами соответствующих прочтений пары А.

3.2.1.2. Режим «fast»

Этот режим работает путем сравнения хешей последовательностей. Во время выполнения последовательности упаковываются в массивы 64-битных целых чисел. В этом режиме использование памяти не может быть ограничено, и он обнаруживает и удаляет только точные дубликаты. Однако он быстрее, чем sequence-based режим и не требует больших дисковых ресурсов.

3.2.1.3. Рекомендации по выбору режима

Мы рекомендуем использовать «fast» режим для относительно небольших наборов данных, когда требуются быстрые результаты или при работе в системе с высокой нагрузкой I/O. «Sequence-based» режим больше подходит для больших наборов данных в системах с ограниченными вычислительными ресурсами, при условии, что производительность I/O не является ограничивающим фактором. «Tight» режим предлагает сбалансированный базовый уровень для дедупликации, «loose» режим представляет собой эффективную реализацию алгоритма FastUniq, а режим «tail-hamming» можно рассматривать как несколько более агрессивный вариант режима «tight».

3.2.2. Сравнение Fastq-dupaway с *de novo*-based инструментами ПЦР-дедупликации

Предварительное тестирование *de novo*-based методов дедупликации показало, что несколько программ имеют существенные ограничения (табл. 2). Например, Fastx Toolkit Collapser принимает входные данные в формате FASTQ, но выдает файл FASTA с измененными идентификаторами прочтений. Seqkit rmdup ограничен одноконцевыми прочтениями, в то время как FastUniq обрабатывает только парные прочтения, и его выходные данные зависят от порядка входных файлов. В отличие от него, Fastq-dupaway демонстрирует гибкость и универсальность, эффективно обрабатывая все эти сценарии. Следует отметить, что как FastUniq, так и «loose» режим Fastq-dupaway дают результаты, которые сильно зависят от порядка входных файлов. Такое поведение может быть связано с присущей неоднозначностью в определении идентичности дубликатов при реализации логики FastUniq, которая основана на сравнении префиксов последовательностей. Подробное описание алгоритма сравнения с примерами приведено в руководстве на GitHub (*https://github.com/AndrewSigorskih/fastq-dupaway*). Различные инструменты обладают некоторыми дополнительными функциями, не представленными в табл. 2. В частности, BBTools Clumpify может идентифицировать оптические дубликаты и выполнять исправление ошибок.

Таблица 2. Сравнение характеристик *de novo*-based инструментов ПЦР-дедупликации. Колонки «А» – На результат влияет порядок входных файлов; «Б» – Существует параметр «N-mismatch»; «В» – Есть ли какие-либо настраиваемые параметры?; «Г» – Может работать с файлами, сжатыми в формате gz; «Д» – Обрабатывает неоднозначные нуклеотиды (N); «Е» – Основной язык программирования.

Программа удаления ПЦР-дубликатов	**Тип входных данных**	**Изменяются ли идентификаторы прочтения?**	**А**	**Б**	**В**	**Г**	**Д**	**Е**
Fastq-dupaway «tight»	Любые	Нет	Нет	Нет	Да	Да	Нет	C++
Fastq-dupaway «loose»	Любые	Нет	Да	Нет	Да	Да	Нет	C++
Fastq-dupaway «tail-hamming»	Любые	Нет	Нет	Да	Да	Да	Нет	C++
Fastq-dupaway «fast»	Любые	Нет	Нет	Нет	Да	Да	Нет	C++
FastUniq	Парные	Нет	Да	Нет	Нет	Нет	Нет	C
BBTools Clumpify	Любые	Нет	Нет	Да	Да	Да	Да	Java
CD-HIT-DUP	Любые	Нет	Нет	Да	Да	Нет	Нет	C++
Fastx Toolkit Collapser	Одноконцевые	Да (вывод в формате FASTA)	–	Нет	Нет	Нет	Нет	C/C++
Seqkit rmdup	Одноконцевые	Нет	–	Нет	Да	Нет	Нет	GO

Мы сравнили Fastq-dupaway (в различных режимах) с пятью широко используемыми *de novo*-based инструментами дедупликации – FastUniq, BBTools Clumpify, CD-HIT-DUP, Fastx Toolkit Collapser и Seqkit rmdup – на 15 наборах данных различных типов и размеров (табл. 3). Для обеспечения объективности сравнения все инструменты запускались на одном ядре центрального процессора (CPU), который поддерживает один поток.

Таблица 3. Характеристики наборов данных NGS, использованных для сравнительной оценки инструментов удаления ПЦР-дубликатов. Данные доступны по адресу *https://www.ncbi.nlm.nih.gov/sra*

Тип протокола	**Датасет**	**Тип прочтений**	**Количество прочтений**	**Размер датасета (ГБ)**	**Организм**
RADICL-seq	SRR9201799, SRR9201800	парноконцевые	141,369,583	36	*Mus musculus*
GRID-seq	SRR3633290, SRR3633291	парноконцевые	149,560,162	42	*Mus musculus*
ChIRP-seq	SRR1425229	одноконцевые	297,394,120	53	*Homo sapiens*
CHART-seq	SRR10044362	одноконцевые	60,857,691	10.8	*Homo sapiens*
Whole Genome Sequencing	SRR2014554	парноконцевые	24,627,585	12.5	*Escherichia coli RR1*
Whole Genome Sequencing	SRR19505554	парноконцевые	175,735,042	104	*Homo sapiens*
Whole Genome Sequencing	SRR19505555	парноконцевые	201,266,001	112.4	*Homo sapiens*
ChIP-seq (H3K27me3)	SRR8902551	парноконцевые	31,868,491	12.3	*Mus musculus*
ChIP-seq (H3K4me1)	SRR504934	одноконцевые	31,269,559	5.3	*Homo sapiens*
ChIP-seq (CTCF)	SRR10950502	парноконцевые	51,668,100	16.5	*Mus musculus*
Hi-C	SRR9675763	парноконцевые	305,604,544	106.6	*Homo sapiens*
Hi-C	SRR8902547	парноконцевые	150,153,852	47.6	*Mus musculus*

Hi-C	SRR1658643	парноконцевые	1,094,811,672	538	*Homo sapiens*
Exome-seq	SRR24907572	парноконцевые	91,397,861	62.4	*Homo sapiens*
Exome-seq	SRR13232316	парноконцевые	127,197,855	91.4	*Homo sapiens*

Fastq-dupaway демонстрирует различные характеристики производительности по разным метрикам. Что касается затраченного времени выполнения (elapsed time), то «fast» режим оказался самым быстрым среди всех протестированных инструментов, в то время как «tight», «loose» и «tail-hamming» режимы были примерно в 1,5 раза медленнее, чем Seqkit rmdup и FastUniq (рис. 20А), что отражает накладные расходы на I/O при операциях с диском. Напротив, все режимы Fastq-dupaway демонстрируют наилучшую производительность по времени работы CPU (рис. 20Б). Характеристики использования оперативной памяти дополнительно подчеркивают полезность инструмента. Режимы «tight», «loose» и «tail-hamming» поддерживают стабильно низкий объем RAM, составляющий приблизительно 2 ГБ, что делает их подходящими для обработки больших наборов данных на стандартных персональных компьютерах. «Fast» режим использует меньше RAM, чем большинство инструментов, уступая по производительности только Seqkit rmdup, который ограничен одноконцевыми чтениями (рис. 20В). Такие программы, как Clumpify и CD-HIT-DUP, требуют значительного объема RAM, что делает их подходящими в основном для вычислительных кластеров с большими объемами RAM (в среднем, в 2–3 раза превышающими размер обрабатываемых данных) (рис. 20В). Это ограничение стало особенно очевидным при обработке большого набора данных Hi-C (SRR1658643, 538 ГБ, 1 094 811 672 парных чтения), где CD-HIT-DUP и FastUniq потребовали приблизительно 1 ТБ RAM. BBtools Clumpify завис на этом наборе данных при работе только с одним ядром CPU, но стабильно работает с четырьмя ядрами. В том же тесте Fastq-dupaway в «fast» режиме использует в 6 раз меньше RAM, чем FastUniq и CD-HIT-DUP (рис. 20В), что подтверждает его преимущества при обработке больших наборов данных на оборудовании с ограниченными ресурсами.

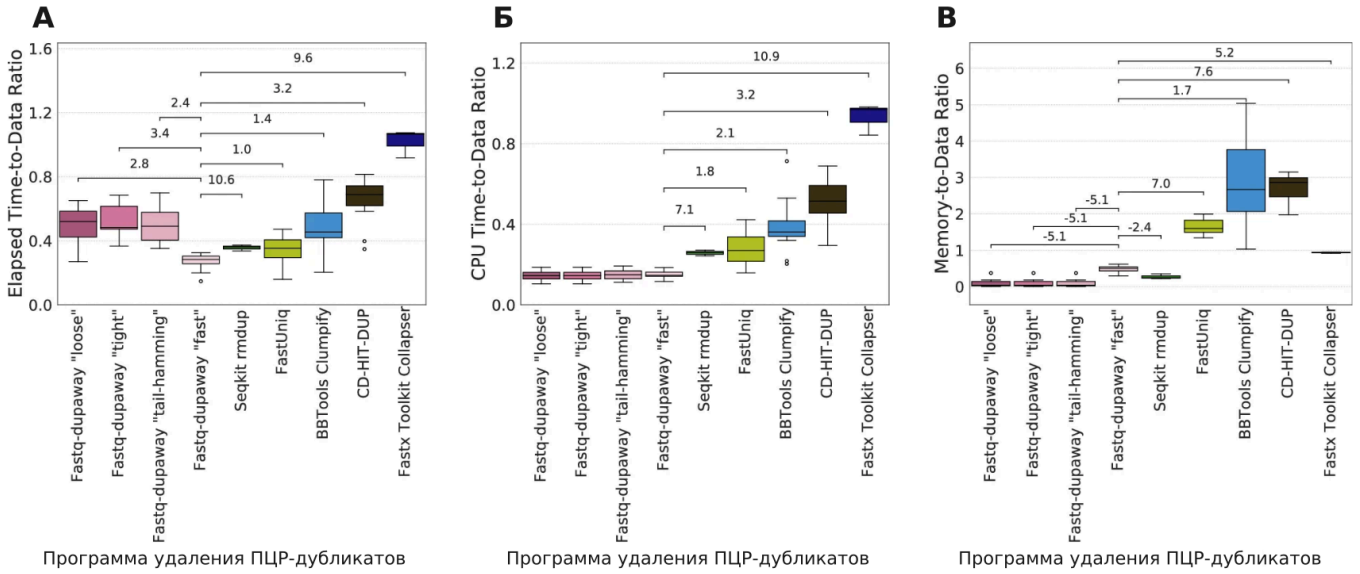


Рисунок 20. Сравнение производительности инструментов дедупликации по трем показателям. (А) Соотношение затраченного времени к объему данных (Elapsed Time-to-Data Ratio, мин/ГБ), (Б) Соотношение времени CPU к объему данных (CPU Time-to-Data Ratio, мин/ГБ) и (В) Соотношение RAM к объему данных (Memory-to-Data Ratio, безразмерная величина). Все показатели рассчитывались путем деления абсолютного значения на соответствующий размер набора данных (ГБ). Более низкие значения указывают на лучшую производительность по всем показателям. Статистический анализ сравнивал каждый инструмент с Fastq-dupaway в «fast» режиме (8 сравнений на каждый показатель) с использованием t-критерия Уэлча с поправкой Бенджамини-Хохберга на ложноположительные результаты. Горизонтальные соединители выделяют статистически значимые различия (FDR < 0,05), а значения строго стандартизированной разницы средних (SSMD) с поправкой на ковариацию указаны над каждым соединителем.

Как показано на рис. 20Б (соотношение времени CPU к объему данных), «fast» режим Fastq-dupaway демонстрирует несколько более низкую производительность по сравнению с режимами «tight», «loose» и «tail-hamming». Однако на рис. 20А (соотношение затраченного времени к объему данных) эта зависимость перевернута, и «fast» режим оказывается самым быстрым среди всех. Это кажущееся несоответствие объясняется принципиальной разницей в том, как эти режимы используют системные ресурсы. Режимы «tight», «loose» и «tail-hamming» интенсивно используют дисковое пространство, применяя внешний алгоритм сортировки, который генерирует значительные количество операций I/O. Хотя такой подход минимизирует потребление RAM, связанная с этим задержка дискового I/O учитывается в измерении затраченного времени, а не в чистом времени работы CPU. Напротив, «fast» режим работает преимущественно с RAM, жертвуя большим объемом оперативной памяти ради значительно более высокой производительности. Следовательно, его общее время выполнения (затраченное время) меньше, поскольку он избегает узкого места доступа к диску, несмотря на то, что требует больше циклов CPU для процессов хеширования и дедупликации в памяти.

Стоит отметить, что в то время как большинство инструментов, включая все режимы

Fastq-dupaway, работают в однопоточном режиме, BBTools Clumpify предлагает возможности многопоточности. Наше сравнение его однопоточного и многопоточного выполнения (см. приложение Б, табл. Б.1) показало, что многопоточность может обеспечить Clumpify увеличение скорости обработки до трех раз.

Для основного тестирования производительности все программы запускались пять раз на каждом наборе данных. По мере увеличения размера обрабатываемых данных вариативность времени работы программ возрастает, причем этот эффект более выражен для затраченного времени по сравнению со временем CPU (см. приложение Б, рис. Б.2 и Б.3). Это расхождение возникает потому, что затраченное время включает не только вычислительное время (время CPU), но и накладные расходы, такие как операции I/O, задержки освобождения памяти и другие системные задержки. Вариативность в «sequence-based» режимах fastq-dupaway может зависеть от нагрузки I/O сервера. В отличие от этого, вариативность других инструментов, менее зависимых от I/O, может быть обусловлена различными факторами, детальная оценка и минимизация которых выходят за рамки типичных научных приложений. BBTools Clumpify продемонстрировал наибольшую нестабильность производительности, особенно в использовании RAM (см. приложение Б, рис. Б.4).

В случае если между прочтениями не допускаются несовпадения, различные методы демонстрируют схожую эффективность в удалении ПЦР-дубликатов (рис. 21). Однако при учете потенциальных различий в несколько нуклеотидов (например, два), наблюдается существенная вариативность результатов в разных программах, при этом доля дополнительно удаленных прочтений варьируется до трех раз (рис. 22). Эта изменчивость обусловлена фундаментальной проблемой транзитивности при идентификации ПЦР-дубликатов с допустимыми несовпадениями и отсутствием единого стандартизированного подхода к их идентификации. Например, при использовании расстояния Хэмминга прочтения «А» и «Б» могут отличаться на один нуклеотид, как и прочтения «Б» и «В», в то время как прочтения «А» и «В» могут отличаться на два нуклеотида. В зависимости от алгоритма, реализованного в конкретной программе, может быть удалено либо прочтение «Б», либо оба прочтения «А» и «В» (рис. 23). Это нарушение транзитивности означает, что при допущении несовпадений проблема становится вырожденной, и единого истинного решения не существует. Следовательно, в данном контексте невозможно определить или оценить стандартные показатели ошибок первого и второго рода. Окончательный набор удаленных прочтений полностью определяется конкретным алгоритмом, используемым каждым инструментом. Таким образом, как «Программа 1», так и «Программа 2» на рис. 23 предоставляют логически обоснованные, но разные решения одной и той же некорректно поставленной задачи.

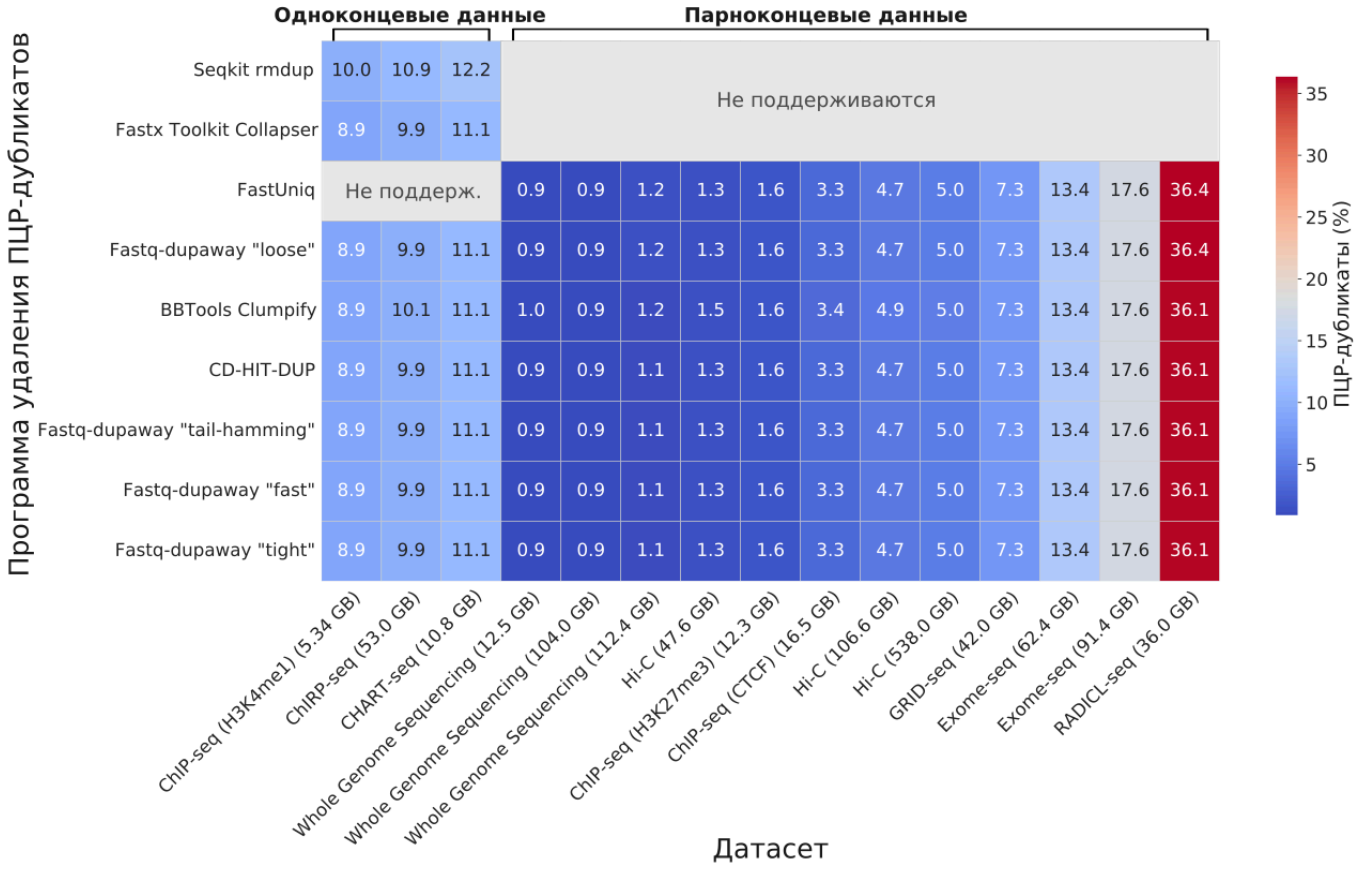


Рисунок 21. Процент ПЦР-дубликатов, идентифицированных каждым инструментом для каждого набора данных. Для ПЦР-дубликатов не допускалось ни одного несовпадения. Для набора данных «Hi-C (538,0 ГБ)» BBTools Clumpify выполнялся в многопоточном режиме. «Not Supported»: инструмент не поддерживает соответствующий тип данных.

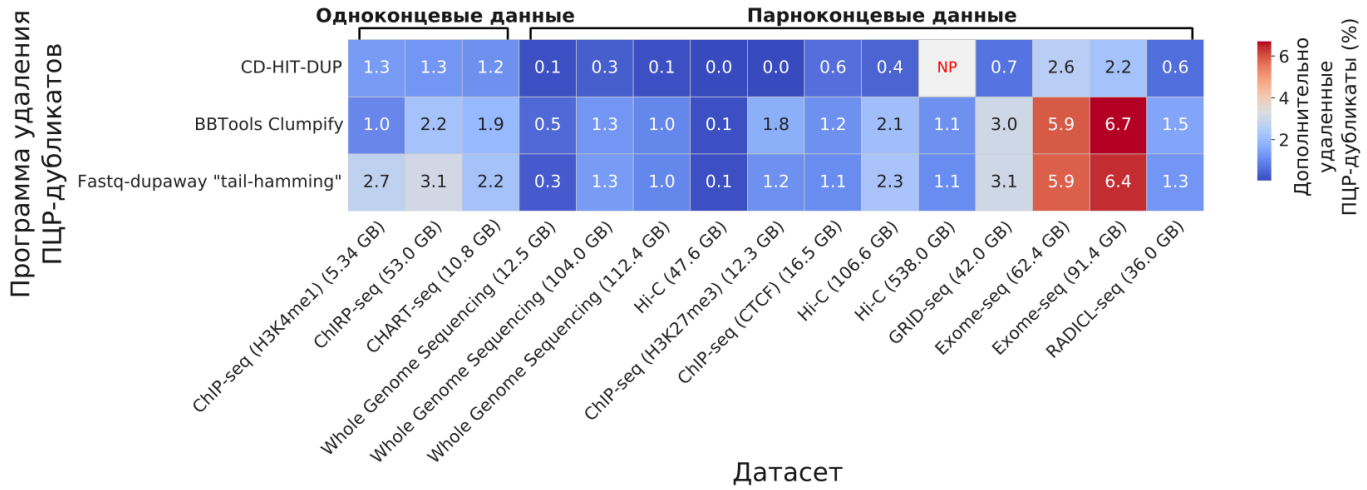


Рисунок 22. Процент дополнительных ПЦР-дубликатов, удаленных каждым инструментом для каждого набора данных при допущении двух несовпадений. Для набора данных «Hi-C (538,0 ГБ)» BBTools Clumpify выполнялся в многопоточном режиме. «NP»: не обработано из-за ошибки.

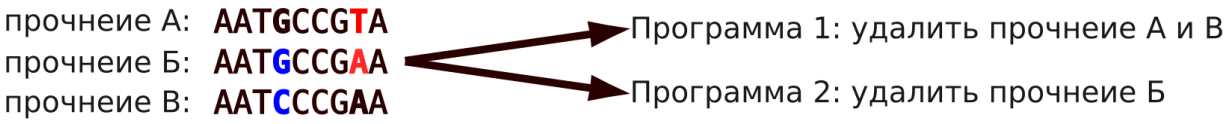


Рисунок 23. Проблема транзитивности при идентификации ПЦР-дубликатов с несовпадениями.

Анализ производительности с допустимым уровнем несоответствия в два нуклеотида показывает, что Fastq-dupaway в режиме «tail-hamming» обеспечивает более высокую скорость обработки данных по сравнению с BBTools Clumpify и CD-HIT-DUP, при этом стабильно используя ровно 2 ГБ оперативной памяти независимо от размера данных (см. приложение Б, табл. Б2-Б4). В то же время, при обработке большого набора данных Hi-C (SRR1658643, 538 ГБ) как BBTools Clumpify, так и CD-HIT-DUP дали сбой, по-видимому, из-за ограничений памяти. Этот результат подчеркивает надежность алгоритма Fastq-dupaway при обработке больших наборов данных при сохранении предсказуемого потребления памяти.

3.2.3. Сравнение Fastq-dupaway с alignment-based инструментами ПЦР-дедупликации

Прямое сравнение точности alignment-based и *de novo*-based методов удаления ПЦР-дубликатов представляет собой сложную задачу, поскольку эти подходы опираются на принципиально разные принципы идентификации дубликатов. Методы, основанные на выравнивании последовательностей, используют геномные координаты для различения истинных ПЦР-дубликатов от отдельных молекул, имеющих идентичные последовательности, но происходящих из разных геномных локусов, особенно в повторяющихся регионах. В отличие от них, *de novo*-based методы, которые полагаются исключительно на идентичность последовательностей, не могут сделать это различие и, следовательно, могут ошибочно классифицировать недублированные прочтения из повторяющихся регионов как дубликаты. Однако это теоретическое преимущество методов, основанных на выравнивании, часто нивелируется на практике. В стандартных анализах коротких прочтений (таких как RNA-seq или ChIP-seq) многократно картированные прочтения обычно отфильтровываются, и для удаления дубликатов рассматриваются только уникально картированные прочтения [173]. Для специализированных анализов, направленных на повторяющиеся регионы, используются специальные инструменты [174]. Кроме того, эффективность инструментов, основанных на выравнивании, по своей сути зависит от выбора программы картирования, ее параметров и качества сборки референсного генома, что вносит дополнительные переменные, усложняющие прямые сравнения. В совокупности эти соображения показывают, что проведение контролируемого и справедливого сравнения эффективности удаления дубликатов является особенно сложной задачей, и этот вопрос заслуживает отдельного углубленного исследования (см.,

например, [154]).

В данной работе мы сравнили время CPU и используемый объем оперативной памяти конвейеров обработки данных NGS, интегрирующих либо *de novo*-based метод дедупликации (Fastq-dupaway «tight», fastp [175], HISAT2 [138]), либо alignment-based метод (fastp, HISAT2, Samtools sort, Picard MarkDuplicates). Результаты тестирования на наборах данных различного размера показывают, что alignment-based конвейер требует значительно больше времени CPU и RAM для завершения, чем *de novo*-based конвейер (рис. 24 и приложение Б, рис. Б.5). Основной вклад в это замедление вносит этап сортировки картированных прочтений (Samtools sort), который является обязательным для alignment-based подхода, а также относительно низкая скорость алгоритма Picard MarkDuplicates. Другим фактором, хотя и менее значительным, является сокращение времени картирования при дедупликации прочтений на этапе FASTQ-файла (*de novo*-based подход) перед выравниванием, в отличие от картирования всего исходного набора данных. Как и ожидалось, чем выше исходная доля ПЦР-дубликатов, тем более выраженной становится эта разница (см. приложение Б, рис. Б.6).

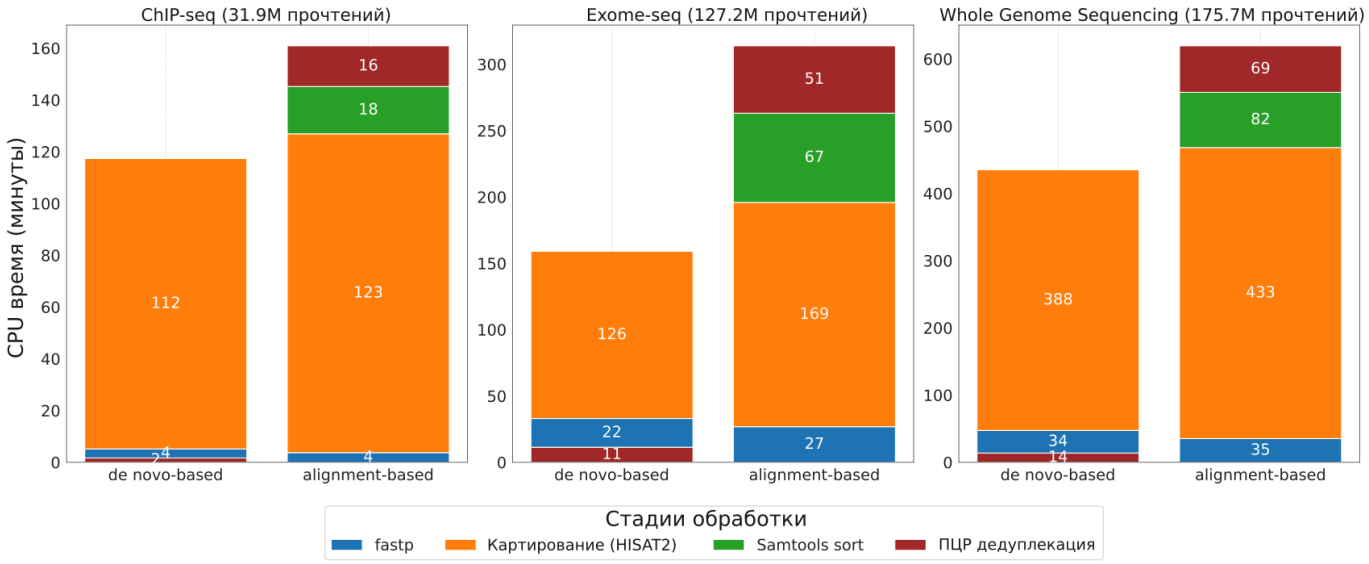


Рисунок 24. CPU время конвейеров, использующих alignment-based и *de novo*-based подходы к дедупликации. ПЦР-дубликаты были идентифицированы с нулевым количеством несовпадений. Каждое значение соответствует медиане пяти запусков соответствующего инструмента. Порядок этапов обработки снизу вверх отражает последовательный порядок выполнения программ в соответствующих конвейерах.

В заключение следует отметить, что alignment-based дедупликация является вычислительно затратной и требует значительных временных затрат, что ограничивает ее практическое применение для анализа больших наборов данных NGS в условиях ограниченных вычислительных ресурсов.

3.3. Интеграция баз данных HiMoRNA и RNA-Chrom[9]

3.3.1. Интеграция баз данных

Поскольку HiMoRNA содержит миллионы эпигенетических пиков, целесообразно отобрать для последующего анализа наиболее надежные из них. Чтобы облегчить эту задачу, мы провели интеграцию пиков HiMoRNA с данными РНК-хроматинового интерактома из RNA-Chrom. Для этого установили взаимно однозначное соответствие между генами двух баз данных и модифицировали веб-интерфейсы (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.3. Интеграция баз данных HiMoRNA и RNA-Chrom»). При таком подходе HiMoRNA может генерировать специальный URL-запрос к 4124 из 4145 днРНК из RNA-Chrom, что, в частности, позволяет ответить на вопрос о том, с какими другими локусами хроматина контактирует исследуемая РНК. Этот подход позволяет значительно расширить представление о функции конкретной РНК.

Общая схема интеграции представлена на рис. 25. Чтобы воспользоваться интеграцией, для начала необходимо найти целевую днРНК в базе данных HiMoRNA. Находясь на главной странице HiMoRNA, пользователь может скачать саму базу данных, добавленные в рамках интеграции «Таблицу генов» («Gene table») и «Таблицу соответствия длинных некодирующих РНК» («lncRNA correspondence table»), для поиска в них интересующих генов/днРНК по геномным координатам. Мы предоставили эту опцию, поскольку идентификатор Ensembl или названия днРНК и ассоциированные с гистоновыми модификациями гены, которые пользователь хочет использовать, могут не совпадать с приведенными в HiMoRNA. На странице поиска пользователю необходимо настроить фильтры под свою задачу, указывая интересующие днРНК, модификации гистонов, геномные координаты и гены, ассоциированные с выбранной модификацией гистонов.

Попав на страницу с результатами поиска, пользователь может более подробно изучить найденные по запросу предсказания, в частности, перейдя в базу данных RNA-Chrom. Для этого следует выбрать интересующую триаду «днРНК–пик эпигенетической модификации–ассоциированный с пиком ген» в интерактивной таблице результатов, а затем нажать на кнопку «Перейти в RNA-Chrom БД» («Go to RNA-Chrom DB»). В выпадающем списке необходимо нажать соответствующую кнопку для перехода на страницу: 1) с контактами данной днРНК в области конкретного пика (надо выбрать, на сколько расширить координаты пика при поиске контактов); 2) со всеми контактами данной днРНК; 3) со всеми днРНК, которые имеют

[9] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: 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).

контакты в данном локусе. Далее пользователь будет перенаправлен на веб-страницу базы данных RNA-Chrom с графической сводкой днРНК-хроматинового интерактома, которая позволяет уточнить, опосредована ли функциональная связь «днРНК–эпигенетическая модификация» из HiMoRNA физическим нахождением днРНК у соответствующего геномного локуса, а также какие еще днРНК потенциально могут участвовать в регуляции данного локуса. Для визуального анализа контакты всех интересующих экспериментов можно загрузить в UCSC Genome Browser (нажать на «VIEW IN GENOME BROWSER»). Выбрав один эксперимент по РНК-хроматиновому интерактому, пользователь может получить список генов, которые располагаются на интересующем участке генома, со статистикой контактируемости днРНК с ними (нажать на «ALL TARGET GENES»). Этот список генов пользователь может скачать и далее работать с ним, например, выполнив GO-анализ. Примеры использования интеграции баз данных HiMoRNA и RNA-Chrom приведены и подробно рассмотрены в разделе «4.3.3 Варианты использования».

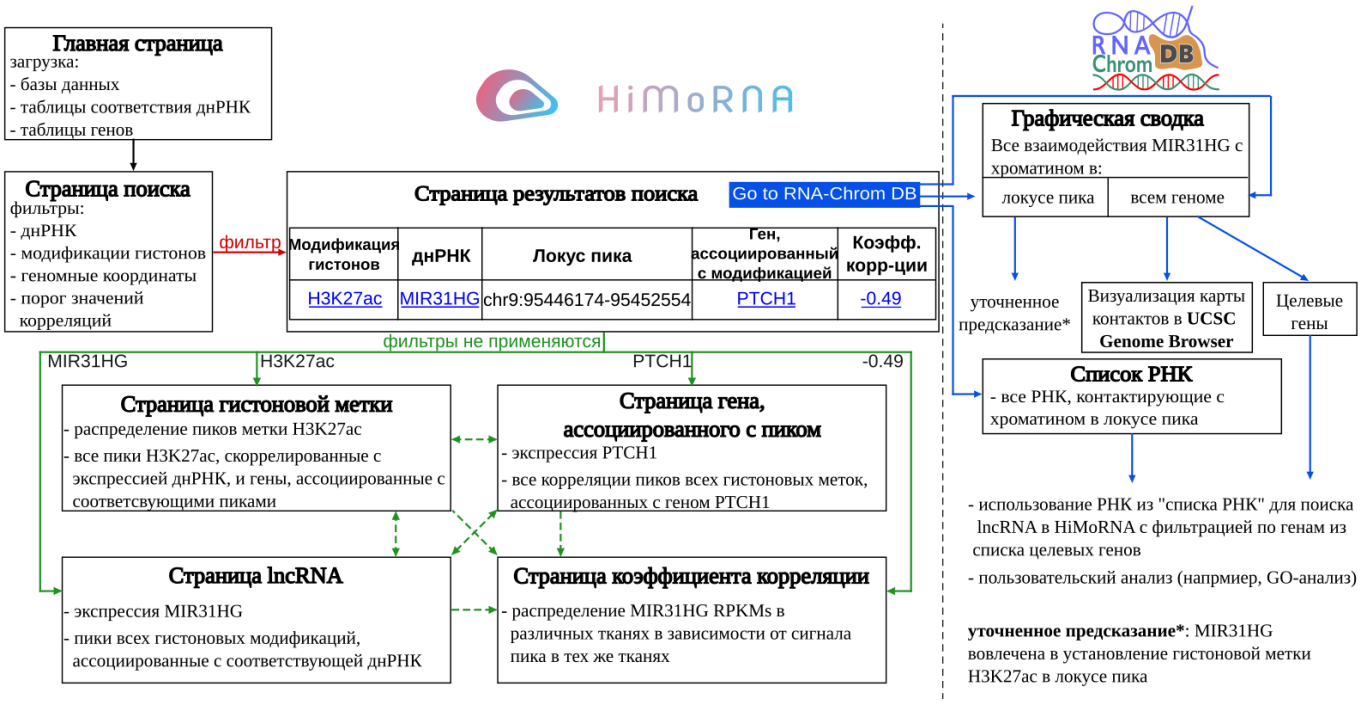


Рисунок 25. Сценарий использования баз данных HiMoRNA и RNA-Chrom после интеграции. Прямоугольники представляют веб-страницы, стрелки – переходы между ними.

3.3.2. Согласованность результатов HiMoRNA и RNA-Chrom

Для оценки полноты интеграции проанализировали частоту подтверждения скоррелированных с экспрессией днРНК гистоновых пиков HiMoRNA данными о контактах соответствующей днРНК с хроматином. Из 4145 днРНК, присутствующих в HiMoRNA, 4011 (96.8%) имеют хотя бы один контакт в базе данных RNA-Chrom, 29 РНК не согласуются между базами данных и еще 105 (2.5%) не имеют контактов в RNA-Chrom. Среди интересующих нас 4011 днРНК только 35.5% имеют хотя бы один пик, который поддерживается контактами

соответствующей днРНК. Однако, принимая во внимание, что по построению экспериментальных протоколов реальное взаимодействие днРНК с хроматином может происходить на удалении от экспериментально фиксируемого контакта, мы предлагаем расширять координаты контакта для более точной оценки соответствия предсказанных пиков HiMoRNA и информации из RNA-Chrom. При расширении контактов на +/- 1, +/- 5, +/- 10, +/- 25 и +/- 50 тысяч п.н. процент РНК, у которых пики из HiMoRNA подтверждаются хотя бы одним контактом, увеличивается до 38.5, 42.7, 45.7, 50.1 и 53% соответственно. В частности, для днРНК MALAT1, HOXC-AS2, NEAT1, NR2F1-AS1, PVT1, MEG3 и ряда других доля расширенных на +/- 25 тысяч п.н. пиков из HiMoRNA, подтвержденных в RNA-Chrom, приближается к 1 (рис. 26). Однако чаще встречаются днРНК, у которых доля расширенных на +/- 25 тысяч п.н. и подтвержденных контактами пиков сильно меньше 1 (JPX, AP005263.1, MIR31HG) или приближается к 0 (MAPKAPK5-AS1). Это, по-видимому, связано с тем, что базы данных HiMoRNA и RNA-Chrom содержат неполную информацию о днРНК из-за строгой фильтрации предсказаний и несовершенства экспериментальных данных РНК–хроматиновых взаимодействий соответственно. Например, в RNA-Chrom половина рассматриваемых в данной статье днРНК имеют меньше 200 контактов (рис. 26Б), так как для большинства днРНК имеются лишь данные «все-против-всех», полученные экспериментальными методами, которые недостаточно полно определяют контакты низко экспрессирующихся РНК.

Следует подчеркнуть, что величина расширения координат контактов влияет не только на численную долю подтверждаемых ассоциаций, но и на биологическую интерпретацию соответствий. Малые окна расширения (например, ±1–5 тыс. п.н.) задают более строгий критерий и позволяют говорить о близком позиционном совпадении контакта с эпигенетическим пиком, однако при этом, вероятно, недооценивают число реальных соответствий из-за ограниченного разрешения методов РНК-хроматинового интерактома. Напротив, при расширении окна до ±25–50 тыс. п.н. возрастает чувствительность, но такие соответствия следует интерпретировать уже не как точное совпадение координат, а как принадлежность контакта к той же регуляторной окрестности. В настоящей работе для детального разбора отдельных примеров использовано расширение на ±25 тыс. п.н. как компромиссное значение, поскольку оно существенно увеличивает долю подтверждаемых ассоциаций по сравнению с отсутствием расширения, тогда как дальнейшее увеличение окна дает лишь умеренный дополнительный прирост.

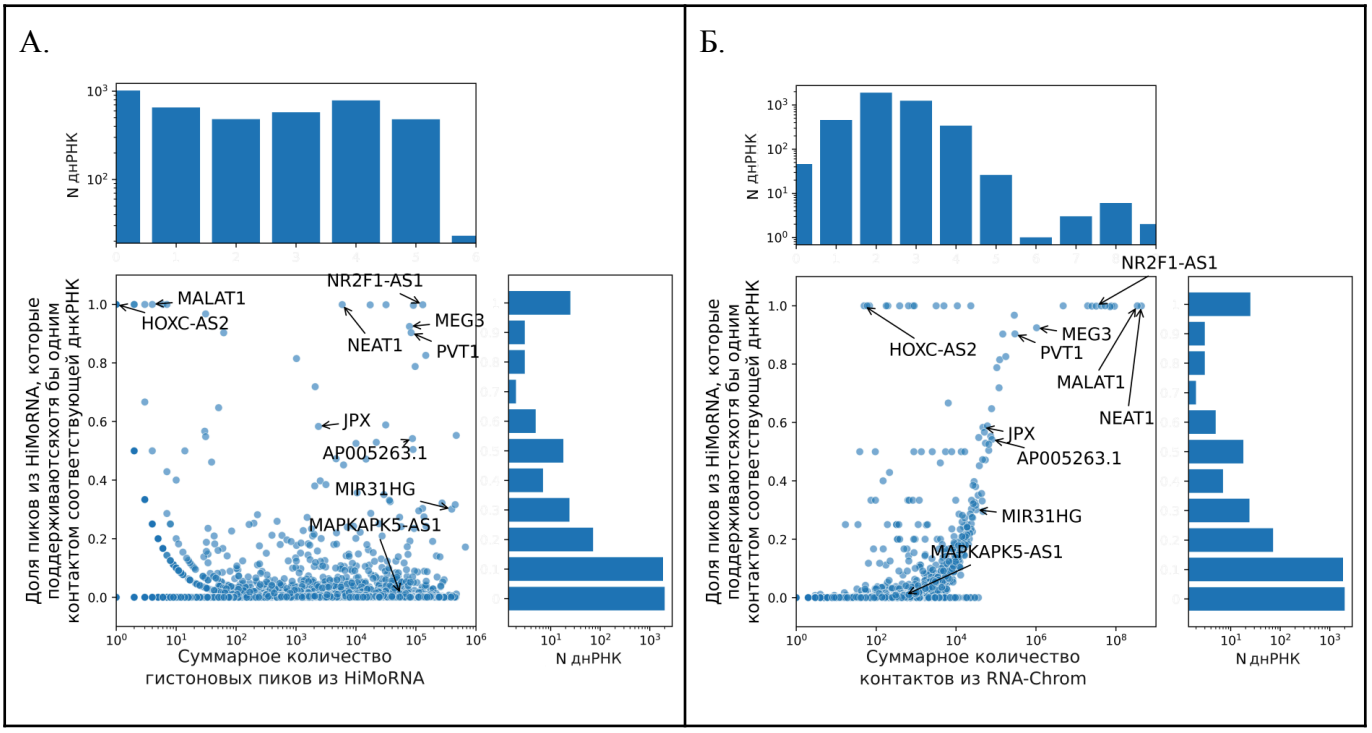


Рисунок 26. Доли пиков HiMoRNA, которые подтверждаются хотя бы одним контактом соответствующей днРНК из RNA-Chrom, относительно (А) суммарного количества пиков HiMoRNA для соответствующей днРНК и (Б) суммарного количества контактов для соответствующей днРНК из RNA-Chrom. Геномные координаты контактов расширены на +/- 25 тысяч п.н.

В базе данных HiMoRNA встречаются триады с отрицательной или положительной корреляцией экспрессии днРНК и сигнала эпигенетического пика («-» и «+» пики соответственно). Для того, чтобы оценить, насколько хорошо предсказания на основе интегрированных баз данных согласуются с опубликованными на данный момент экспериментальными исследованиями, мы отобрали 30 днРНК и соответствующие им гистоновые пики, для которых статистически значимо (односторонний точный тест Фишера, *p-value* < 0.001) преобладают «+» или «-» пики хотя бы одной из гистоновых меток, подтвержденные контактами из RNA-Chrom с расширением на +/- 25 тысяч п.н. (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.3.1. Односторонний точный тест Фишера», рис. 27). Отфильтровав результаты по *p-value* < 0.001, мы получили следующие пары «днРНК-гистоновая метка»:

1) 21 днРНК, у которых «+» пики соответствующих гистоновых меток лучше поддерживаются контактами из RNA-Chrom нежели «-» пики (правосторонний точный тест Фишера, *p-value* < 0.001).
2) 11 днРНК, у которых «-» пики соответствующих гистоновых меток лучше поддерживаются контактами из RNA-Chrom, нежели «+» пики (левосторонний точный тест Фишера, *p-value* < 0.001).

Ранее было показано потенциальное участие значительной части выявленных днРНК в

эпигенетической регуляции посредством гистоновых модификаций. Разберем случаи, когда «+» пики статистически значимо лучше поддерживаются контактами из RNA-Chrom, нежели «-» пики. Например, MIR4435-2HG участвует в установлении активаторной метки H3K27ac в энхансерном регионе локуса RPTOR [176]. Наши данные показывают, что у MIR4435-2HG, помимо H3K27ac, вероятно, существуют и другие мишени эпигенетической регуляции посредством установления таких меток, как H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2 (рис. 27A). Аналогично на основании данных для MIR31HG [177], SNHG1, PVT1 [178–180] и днРНК lnc-Nr2f1 мыши (предположительно имеющей функциональную консервативность с днРНК NR2F1-AS1 человека) [42] мы определили согласующиеся с этими данными гистоновые модификации: NR2F1-AS1 – H3K27ac, MIR31HG – H3K4me3, SNHG1 – H3K27me3, PVT1 – H3K27me3. Помимо этого, мы выявили функциональную связь этих днРНК с другими эпигенетическими метками: NR2F1-AS1 – H3K4me2, MIR31HG – H3K79me2, SNHG1 – H3K4me3, PVT1 – H3K4me1, H3K4me3, H3K9me3 (рис. 27A).

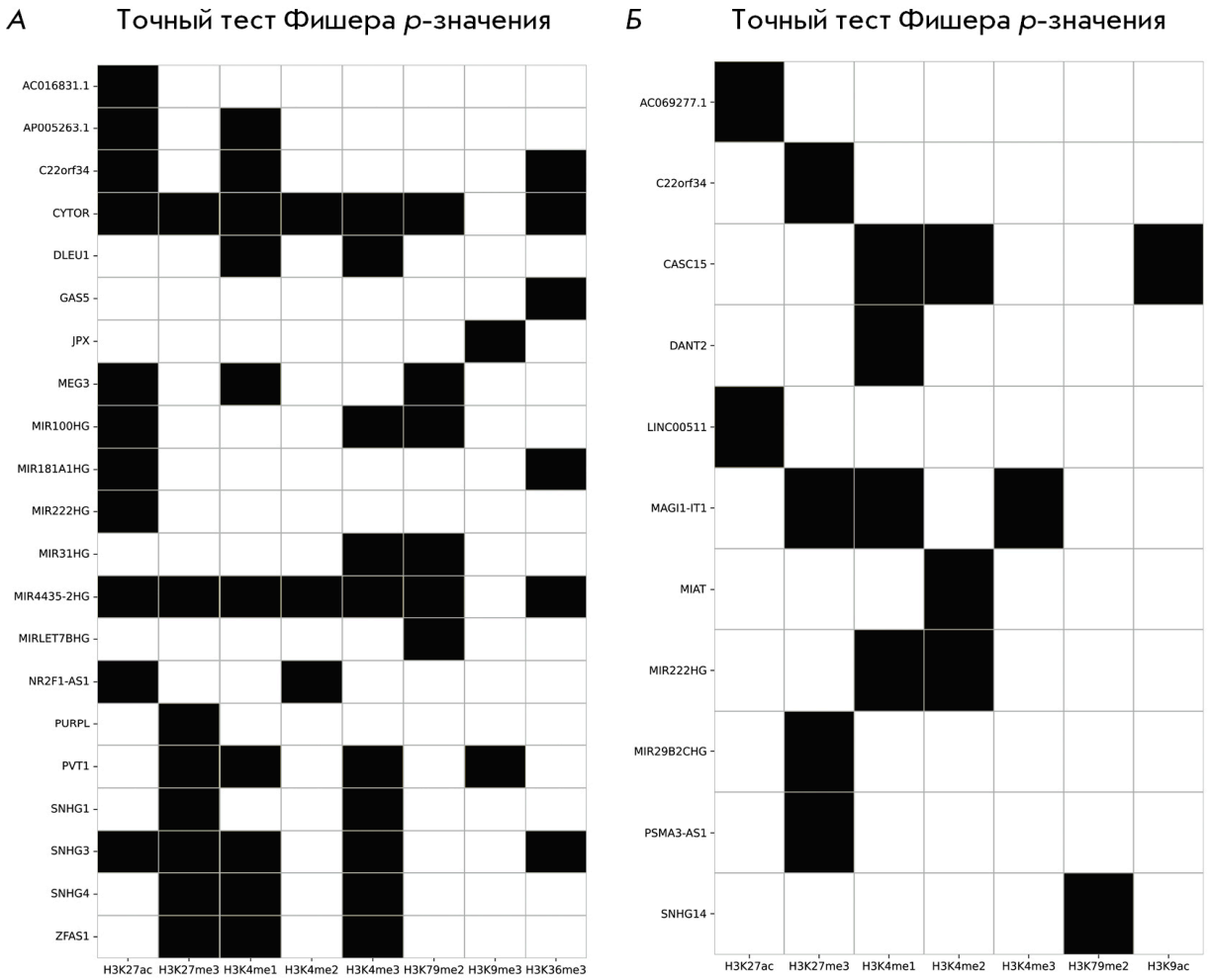


Рисунок 27. Тепловая карта с результатами точного теста Фишера («Fisher exact test») для пар «днРНК – расширенные на +/- 25 тысяч п.н. пики гистоновой метки». Черный цвет означает, что доля «-» / «+» гистоновых пиков, поддерживаемых контактами соответствующей днРНК, больше 0.4 и *p-value* точного теста Фишера меньше 10^{-3}, иначе – белый цвет. А – правосторонний тест Фишера («right-tailed»): «+» пики соответствующих гистоновых меток лучше поддерживаются контактами из RNA-Chrom, нежели «-» пики. Б – левосторонний тест Фишера («left-tailed»): «-» пики соответствующих гистоновых меток лучше поддерживаются контактами из RNA-Chrom, нежели «+» пики.

У ряда днРНК (ZFAS1, SNHG4, SNHG1, SNHG3, PVT1, MIR4435-HG, CYTOR) выявлено большое количество подтвержденных контактами «+» пиков H3K27me3 и H3K4me3, которые статистически значимо больше поддерживаются контактами из RNA-Chrom, нежели «-» пики (рис. 27А), отвечающие за противоположные состояния хроматина. По аналогии с существующими днРНК, устанавливающими обе этих гистоновых метки в зависимости от ассоциации с различными эффекторными белками (как, например, нкРНК SRA [80], ANRIL [40]), можно предположить, что для них тоже характерны более сложные механизмы регуляции активности хроматина.

Случаи, когда «-» пики статистически значимо больше поддержаны контактами из RNA-Chrom нежели «+» пики, вероятно, можно объяснить тем, что соответствующие днРНК

регулируют удаление гистоновых меток посредством привлечения деметилаз и деацетилаз к соответствующим геномным локусам (рис. 27Б). Мы не можем оценить качество нашего предсказания для этих днРНК, поскольку экспериментальные статьи, показывающие такого рода регуляцию этих днРНК, не найдены. Мы предполагаем, что полученные в данном разделе пары «днРНК-гистоновая метка» (рис. 27), являются потенциальными объектами дальнейших исследований.

3.3.3. Варианты использования

Основная цель интеграции HiMoRNA и RNA-Chrom – уточнить функциональную связь внутри триад «днРНК–пик эпигенетической модификации–ассоциированный с пиком ген» с помощью данных о локализации соответствующей днРНК в геномной области вблизи пиков конкретной модификации гистонов. Далее мы приведем примеры пользовательского исследования нескольких днРНК, механизм действия которых известен.

3.3.3.1. днРНК MIR31HG

Длинная некодирующая РНК MIR31HG является известным регулятором гистоновых меток H3K1me1, H3K4me3 и H3K27ac. Ранее сообщалось о снижении уровней H3K4me1 и H3K27ac в энхансерной области гена GLI2 и H3K4me3 и H3K27ac в промоторной области гена FABP4 после нокдауна MIR31HG [177,181]. Это наблюдение можно проверить, используя нашу интеграцию HiMoRNA и RNA-Chrom. Для этого мы создали запрос в HiMoRNA: днРНК MIR31HG, метки гистонов H3K4me1 и H3K27ac, координаты двух выбранных генов указаны с увеличенной на 10 тысяч п.н. промоторной областью в поле геномных координат (рис. 28А). В результате веб-ресурс HiMoRNA сгенерировал таблицу с пиками H3K27ac и H3K4me1, которые коррелируют с экспрессией MIR31HG в различных тканях (рис. 28Б). Затем мы выбрали триаду с пиком H3K27ac и перешли на страницу RNA-Chrom с экспериментально обнаруженными контактами MIR31HG с хроматином в области выбранного пика (при нажатии на «Go to RNA-Chrom DB», рис. 28В). Выбрав РНК-хроматиновый эксперимент в верхней таблице и нажав на «All target genes» (рис. 29А), получили таблицу, в которой, в частности, отражено взаимодействие MIR31HG с геном GLI2 (рис. 29Б).

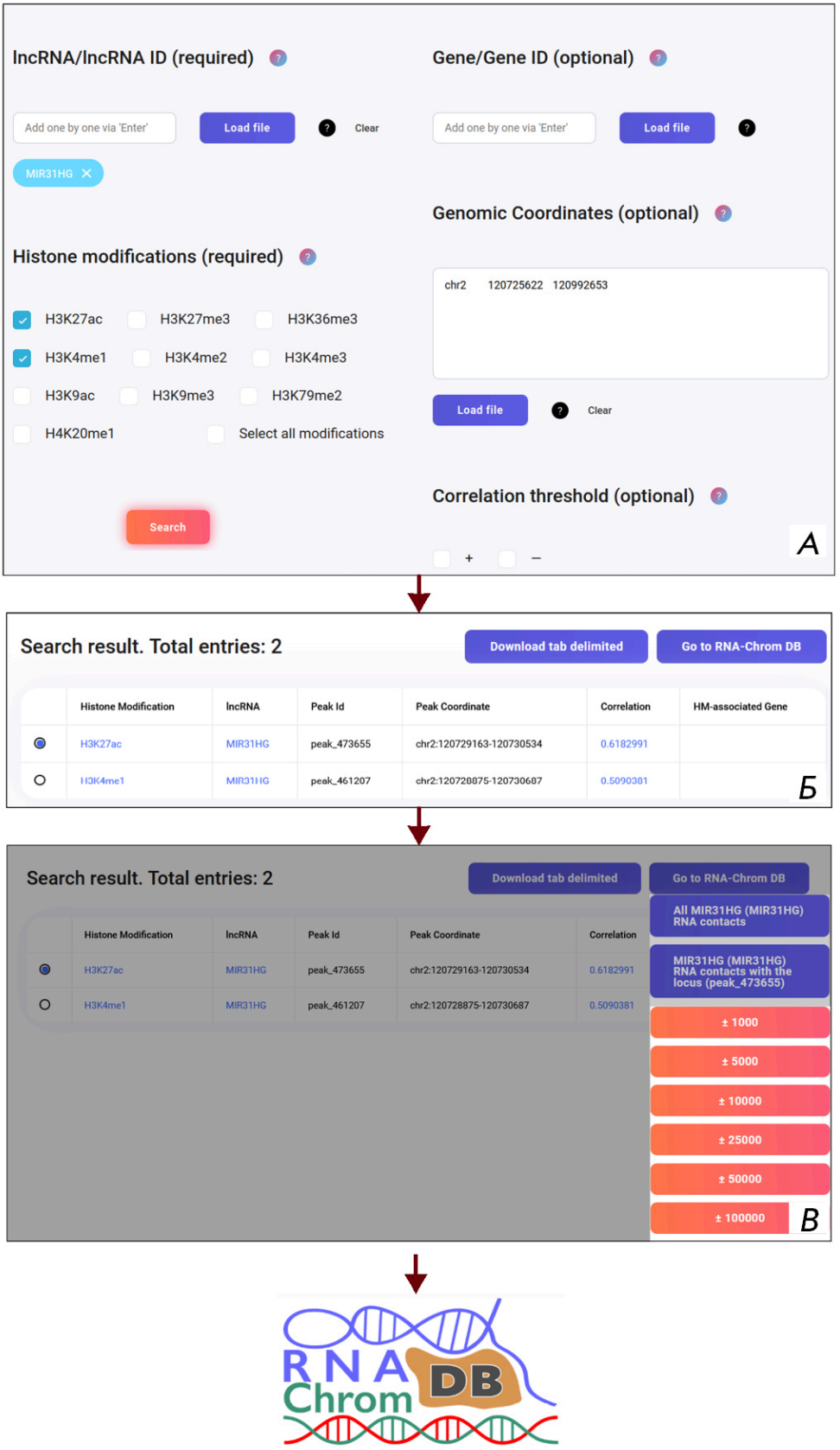


Рисунок 28. Вариант использования интеграции баз данных HiMoRNA и RNA-Chrom на примере днРНК MIR31HG. А – создание запроса в HiMoRNA на MIR31HG, гистоновые модификации H3K4me1 и H3K27ac, гены GLI2 и FABP4. Б – таблица с результатами поиска. В – переход в RNA-Chrom.

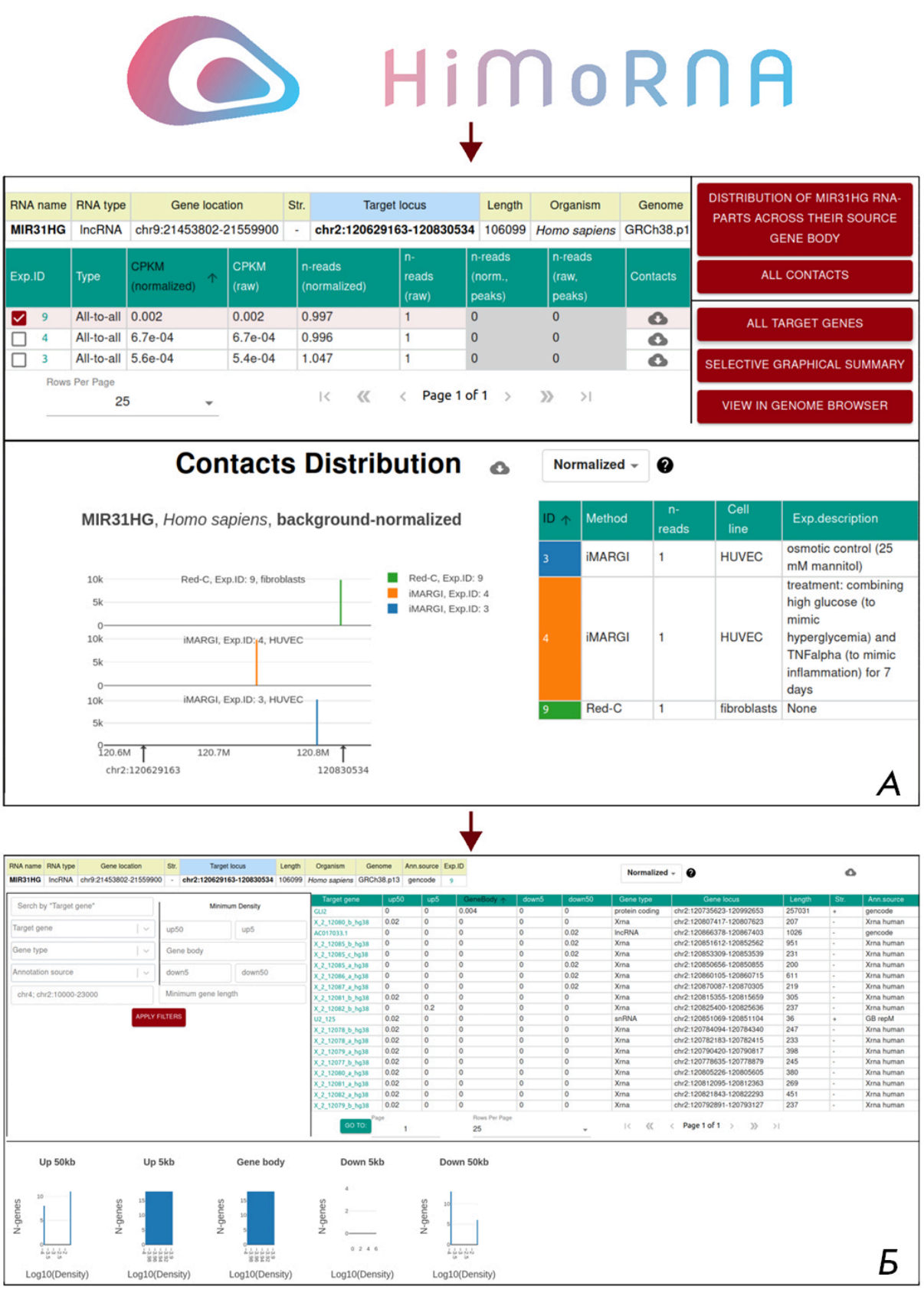


Рисунок 29. Вариант использования интеграции баз данных HiMoRNA и RNA-Chrom на примере днРНК MIR31HG. А – страница RNA-Chrom с контактами MIR31HG с хроматином в области расширенного пика HiMoRNA. Б – таблица со всеми генами из области расширенного пика, с которыми взаимодействует или нет MIR31HG (эксперимент Exp.ID: 9).

Чтобы выяснить, может ли интеграция HiMoRNA и RNA-Chrom дать новую биологическую информацию о функциональной роли днРНК, мы предположили, что MIR31HG может регулировать не только GLI2, но и другие гены, принадлежащие регуляторному пути «Sonic hedgehog» (KEGG:04340). С этой целью идентифицировали соответствующие гены с помощью базы данных KEGG Pathway [182], после чего был сформирован новый запрос в HiMoRNA: днРНК MIR31HG, гистоновые метки H3K4me1 и H3K27ac, названия 56 генов из пути «Hedgehog signaling pathway». В результате мы получили таблицу из 162 триад, которые можно проверить с помощью ресурса RNA-Chrom. Например, в локусе пика H3K27ac_963553 (chr9:95446174-95452554) MIR31HG контактирует с геном PTCH1, кодирующим рецептор «Sonic hedgehog». Чтобы оценить, насколько существенно список генов, ассоциированных со скоррелированными с экспрессией MIR31HG пиками H3K27ac и H3K4me1, обогащен генами из регуляторного пути «Hedgehog signaling pathway», был проведен анализ KEGG обогащений при помощи веб-ресурса «g:Profiler» [183]. В качестве запроса подавались отобранные для MIR31HG и H3K27ac/H3K4me1 гены, а в качестве бэкграунда использовались все остальные ассоциированные с пиками HiMoRNA гены. По результатам нашего анализа, гены, принадлежащие пути «Hedgehog signaling pathway», обогащены пиками H3K27ac (*p-value* = 2.090 × 10-2), но не H3K4me1. Это наблюдение позволяет предположить участие MIR31HG в регуляции пути «Hedgehog signaling pathway» через установление гистоновой модификации H3K27ac в соответствующих геномных локусах.

3.3.3.2. днРНК PVT1

днРНК PVT1 ингибирует экспрессию гена LATS2 в клетках немелкоклеточного рака легкого путем рекрутирования EZH2 (субъединица комплекса PRC2) на соответствующий промотор [184]. Мы выполнили поиск триад в HiMoRNA: днРНК PVT1, все гистоновые модификации, ген LATS2. В результате получили пики только для активирующей метки H3K4me3, которые отрицательно коррелировали с экспрессией PVT1, что косвенно согласуется с опубликованными данными [184], поскольку PVT1 привлекает EZH2 и участвует в установлении репрессивной метки H3K27me3. В RNA-Chrom мы наблюдали контакты вокруг одного из пиков H3K4me3 (peak_169403, chr13:21045571-21046978) в двух экспериментах (клеточные линии K562 и MDA-MB-231). Визуализация контактов PVT1 в Genome Browser [135] подтверждает наличие этого пика в промоторной области гена LATS2 (рис. 30). Дополнительное подтверждение регуляции LATS2 с помощью днРНК PVT1 было получено на основе данных Red-ChIP (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.3.2. Данные Red-ChIP»): обнаружен пик EZH2-опосредованных контактов PVT1 (chr13:21168000-21224000, *q-value* = 0.09) в 106.4 тысяч п.н. от 5’-конца гена LATS2 (рис. 30).

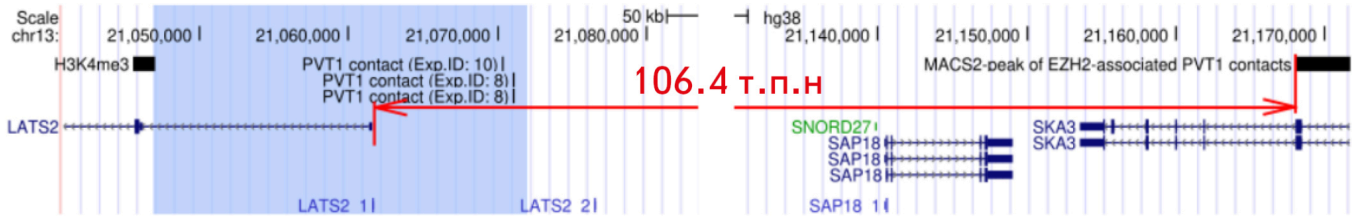


Рисунок 30. Представление в UCSC Genome Browser в области гена LATS2 и его промоторной окрестности пика H3K4me3, скоррелированного с экспрессией днРНК PVT1, контакты днРНК PVT1 из двух экспериментов (RNA-Chrom Exp.ID: 8, 10) и пик EZH2-опосредованных контактов PVT1. Синяя область отражает расширение координат пика H3K4me3 на 25 тысяч п.н., в пределах которых были отобраны контакты из RNA-Chrom.

Отсутствие в HiMoRNA триад «днРНК PVT1–пик H3K27me3–ген LATS2» с положительной корреляцией, по-видимому, связано с чрезмерно строгой фильтрацией пиков метки H3K27me3 при создании базы данных. Приведенные выше примеры из раздела «Варианты использования» подтверждают, что интеграцию можно успешно применять для генерации гипотез о роли днРНК в эпигенетической регуляции конкретных генов для дальнейшей экспериментальной проверки.

3.4. Сравнительный анализ данных РНК-хроматинового интерактома: разрешение, полнота и специфичность данных[10]

3.4.1. Хроматиновый потенциал

Во всех полногеномных исследованиях РНК-хроматиновых взаимодействий (эксперименты ATA) наблюдается значительное преобладание контактов мРНК. Это обусловлено тем, что мРНК имеют, как правило, более высокий уровень экспрессии по сравнению с нкРНК. Хроматиновый потенциал (chP) (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.4.3. Хроматиновый потенциал») позволяет ответить на вопрос, является ли доля контактов данной РНК статистически значимо отличной от той, которую можно было бы ожидать, если бы все РНК контактировали с хроматином неспецифически и пропорционально своему уровню экспрессии. Если предположить, что большинство контактов мРНК с хроматином неспецифичны, то можно ожидать, что некодирующие РНК будут демонстрировать более высокое сродство к хроматину. Как и ожидалось, большинство нкРНК демонстрировали хроматиновый потенциал больше нуля (рис. 31А и приложение В, рис. В.3), однако большое количество мРНК также имело положительный хроматиновый потенциал. При увеличении порога на chP доля мРНК среди прошедших порог РНК снижается (рис. 31Б и приложение В, табл. В.3) с резким падением почти во всех экспериментах

[10] При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором лично или в соавторстве, в которых, согласно Положению о присуждении ученых степеней в МГУ, отражены основные результаты, положения и выводы исследования: **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).

при значениях chP ≥ 20.

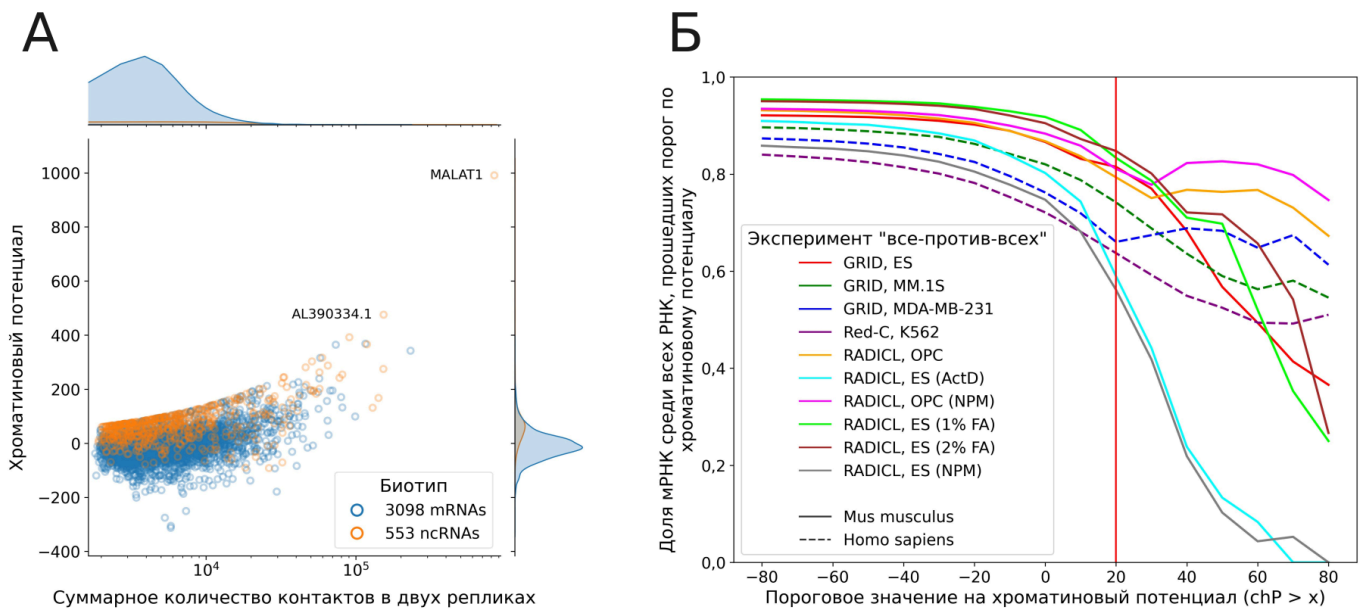


Рисунок 31. Характеристики хроматинового потенциала РНК. А. Зависимость хроматинового потенциала от числа контактов РНК в эксперименте Red-C K562. Голубой цвет – белок-кодирующие РНК, оранжевый – некодирующие РНК. Б. Доля мРНК в зависимости от порога на хроматиновый потенциал (chP > x) для разных экспериментов ATA. Доля нкРНК соответствует 1 минус доля мРНК. ActD – обработка актиномицином D; NPM – обработка протеиназой K; 1% FA – обработка сшивающим агентом формальдегидом в концентрации 1%; 2% FA – обработка сшивающим агентом формальдегидом в концентрации 2%.

Тот факт, что даже при больших порогах на хроматиновый потенциал остается достаточно много белок-кодирующих РНК, может быть связан с несколькими обстоятельствами. Например, некоторые белок-кодирующие гены содержат в своих интронных областях функциональные нкРНК [185,186], среди которых можно ожидать значительное количество неаннотированных нкРНК. Вероятно, положительный хроматиновый потенциал некоторых мРНК связан именно с этими нкРНК. С другой стороны, некодирующие изоформы мРНК могут сами играть определенную роль в регуляции хроматина [187]. Вместе с тем, положительный хроматиновый потенциал у белок-кодирующей РНК сам по себе не следует интерпретировать как доказательство прямой функции зрелой мРНК в регуляции хроматина. В рамках использованного здесь ген-уровневого анализа невозможно надежно отделить сигналы, связанные с экзонными участками зрелой мРНК, от сигналов, происходящих из интронных областей того же гена, в том числе от неаннотированных нкРНК или некодирующих изоформ. Поэтому такие случаи следует рассматривать прежде всего как приоритетные кандидаты для дальнейшего транскрипт-специфичного анализа, а не как окончательно установленные функциональные взаимодействия.

3.4.2. Сравнение реплик в данных ATA

Для оценки консистентности взаимодействий РНК с хроматином между репликами мы оценили долю воспроизводимых контактов. Поскольку точная координата контакта в методах ATA может быть смещена из-за особенностей протокола, мы ввели параметр геномного расстояния (L), в пределах которого контакты, принадлежащие одной РНК, но обнаруженные в разных репликах, считались конкордантными. Чтобы определить порог L, адекватно отражающий разрешение метода, мы для каждой РНК вычисляли долю ее контактов, для которых в другой реплике был обнаружен хотя бы один контакт той же РНК в пределах заданного расстояния L. Анализ зависимости этой доли от L для данных «GRID, ES, Mus musculus» показал, что медианная доля конкордантных контактов перестает существенно возрастать при $L \geq 5000$ п.н. (см. приложение В, рис. В.4), выходя на плато. Это указывает, что 5000 п.н. являются эмпирической оценкой точности позиционирования контакта в методах ATA. На основе этого результата для последующего анализа мы разбили геном на непересекающиеся фрагменты (бины) фиксированного размера (*bin* п.н.). Основной анализ проводили с размером бина 5000 п.н., что соответствует эмпирически оцененной точности позиционирования. Для проверки устойчивости результатов и моделирования сценария «высокого разрешения» был также использован размер бина, равный 1000 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.

Для оценки консистентности реплик в экспериментах ATA мы проанализировали набор РНК, имеющих более 1000 контактов с хроматином в каждой реплике. При отборе РНК фильтр на исключение регионов РД-скейлинга (в пределах 1 Мб от гена-источника РНК) не применялся, он применялся уже при отборе контактов. Важно отметить, что мы оценивали не просто наличие хотя бы одного конкордантного бина, а статистическую значимость общего уровня конкордантности для каждой РНК в целом. Для этого мы подсчитывали общее количество конкордантных и дискордантных бинов для каждой РНК и применяли статистический критерий для проверки гипотезы о неслучайности наблюдаемого уровня совпадений (см. «Глава 2. МАТЕРИАЛЫ И МЕТОДЫ», раздел «2.4.4. Воспроизводимость (конкордантность) контактов в репликах»). РНК считалась конкордантной, если рассчитанный для нее FDR был меньше 0,05. Как видно в приложении В на рис. В.5, наличие единичных конкордантных бинов не гарантирует прохождения этого строгого порога значимости.

В табл. 4 и в приложении В на рис. В.6 показано количество РНК, которое имеет конкордантные бины между репликами с FDR меньше 0,05. Анализ проводился в четырех условиях, позволяющих оценить влияние двух факторов: размер геномного бина (1000 п.н. против

5000 п.н.) и применение фильтрации контактов (все контакты по сравнению с контактами, попавшими в пики BaRDIC). При анализе результатов прежде всего выделяются эксперименты GRID, для которых ни одно из этих условий не оказало влияния: количество конкордантных РНК оставалось неизменным и почти всегда равно исходному количеству отобранных РНК (обсуждение см. ниже).

Таблица 4. Количество РНК, имеющие конкордантные бины в репликах (FDR < 0,05). Отбирались только РНК с числом контактов > 1000 в каждой реплике. При отборе контактов применялся фильтр на расстояние от гена-источника РНК равный 1 Мб.

Эксперимент	**Исходное количество мРНК (нкРНК)**	**Количество конкордантных мРНК (нкРНК), все контакты**		**Количество конкордантных мРНК (нкРНК), контакты из пиков**	
		Бин 1000 п.н.	**Бин 5000 п.н.**	**Бин 1000 п.н.**	**Бин 5000 п.н.**
Red-C, K562, *H. sapiens*	3230 (636)	1571 (341)	2418 (486)	2779 (556)	3188 (628)
GRID, MM.1S, *H. sapiens*	3771 (413)	3771 (413)	3771 (413)	3771 (413)	3771 (413)
GRID, MDA_MB_231, *H. sapiens*	4844 (653)	4844 (653)	4844 (653)	4844 (653)	4844 (653)
GRID, ES, *M. musculus*	4706 (436)	4706 (429)	4706 (427)	4706 (432)	4706 (435)
RADICL (2% FA), ES, *M. musculus*	2758 (162)	1829 (87)	2552 (124)	2226 (131)	2704 (158)
RADICL, OPC, *M. musculus*	2580 (197)	1954 (136)	2484 (175)	2203 (161)	2555 (191)
RADICL (ActD), ES, *M. musculus*	657 (87)	345 (42)	576 (76)	512 (74)	646 (86)
RADICL (NPM), OPC, *M. musculus*	3734 (275)	504 (40)	1464 (103)	2128 (136)	2839 (200)
RADICL (1% FA), ES, *M. musculus*	2079 (117)	1533 (66)	1986 (80)	1811 (102)	2056 (115)
RADICL (NPM), ES, *M. musculus*	643 (149)	643 (149)	643 (149)	643 (148)	643 (148)

Для других данных ATA, как и ожидалось, при использовании всех контактов количество статистически значимо конкордантных РНК было существенно ниже для строгого условия (размер бина – 1000 п.н.) по сравнению с условием, соответствующим разрешению метода (размер бина –

5000 п.н.). Это подтверждает, что более крупный бин лучше агрегирует технические вариации и точнее отражает воспроизводимость взаимодействий. Наиболее важным наблюдением стало то, что предварительный отбор контактов, принадлежащих пикам BaRDIC, значительно повышал консистентность реплик. Эта фильтрация либо увеличивала количество конкордантных РНК, либо позволяла достичь сопоставимого уровня конкордантности даже при использовании строгого размера бина в 1000 п.н. по сравнению с анализом всех контактов с размером бина 5000 п.н. Таким образом, выделение пиков взаимодействия РНК с хроматином с помощью BaRDIC эффективно фильтрует случайные взаимодействия и выделяет наиболее достоверные, воспроизводимые контакты РНК с хроматином, существенно повышая консистентность между репликами.

После выделения статистически значимо конкордантных РНК мы оценили полноту данных АТА, рассчитав для них медианную долю контактов, приходящихся на конкордантные бины размером 5000 п.н., что соответствует оцененному разрешению методов АТА. Данная метрика отражает долю воспроизводимых между репликами взаимодействий от общего числа детектированных контактов (см. приложение В, табл. В.4 и В.5).

Были выявлены кардинальные различия между методами. Полнота данных для Red-C и RADICL-seq не превышала 2% при анализе всех контактов и 5% – для контактов, отфильтрованных по пикам BaRDIC. Напротив, полнота данных GRID-seq оказалась существенно выше, достигая 29% и 82% для всех контактов и контактов из пиков соответственно.

Высокую воспроизводимость данных GRID, вероятно, 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 – это сшивающий агент с длинным спейсером (7,7 Å), который эффективно сшивает белок-белковые взаимодействия, стабилизируя белковые комплексы до фиксации хроматиновой структуры формальдегидом [188]. Это позволяет более эффективно «запечатывать» опосредованные белками РНК-хроматиновые взаимодействия, которые составляют основную долю специфических контактов.

Данное предположение подтверждается на количественном уровне. Несмотря на сопоставимость медиан общего количества контактов конкордантных РНК во всех данных АТА (см. приложение В, рис. В.7А), медианное количество воспроизводимых (конкордантных) контактов в данных GRID-seq на порядок превышало соответствующие показатели для данных Red-C и RADICL-seq (см. приложение В, рис. В.7Б). Это свидетельствует о том, что протокол с DSG не просто увеличивает объем данных, а кардинально повышает удельный вес специфического воспроизводимого сигнала в общем массиве. Таким образом, протокол с дополнительной обработкой DSG обеспечивает более полный и стабильный захват мультибелковых комплексов, что приводит к значительному снижению технического шума и повышению воспроизводимости

между репликами. В то же время фиксация одним формальдегидом может недостаточно стабилизировать крупные надмолекулярные комплексы, что, в свою очередь, повышает долю случайных нестабильных взаимодействий и снижает общую конкордантность.

Несмотря на кардинальные различия в абсолютном уровне конкордантности между методами, для всех экспериментов ATA мы обнаружили общую закономерность: воспроизводимость контактов положительно коррелирует с общим числом взаимодействий РНК и с ее хроматиновым потенциалом (рис. 32 и приложение B, рис. B.8). Обнаруженная зависимость позволяет сделать два важных вывода о природе данных РНК-хроматинового интерактома.

1. Полнота данных является функцией глубины секвенирования для конкретной РНК. Низкая воспроизводимость РНК с малым числом контактов (<10 000) указывает на то, что для таких молекул данные являются существенно неполными и содержат высокий уровень шума. Достаточная полнота достигается лишь при большом количестве взаимодействий, что свидетельствует о необходимости глубокого секвенирования для надежного выявления интерактома отдельных РНК.
2. Воспроизводимость является маркером биологической значимости. Положительная корреляция между хроматиновым потенциалом и долей конкордантных контактов свидетельствует о том, что чем специфичнее взаимодействие (выше хроматиновый потенциал), тем оно стабильнее и лучше воспроизводится между репликами. Это укрепляет позицию хроматинового потенциала не только как меры специфичности, но и как предиктора достоверности и воспроизводимости взаимодействий.

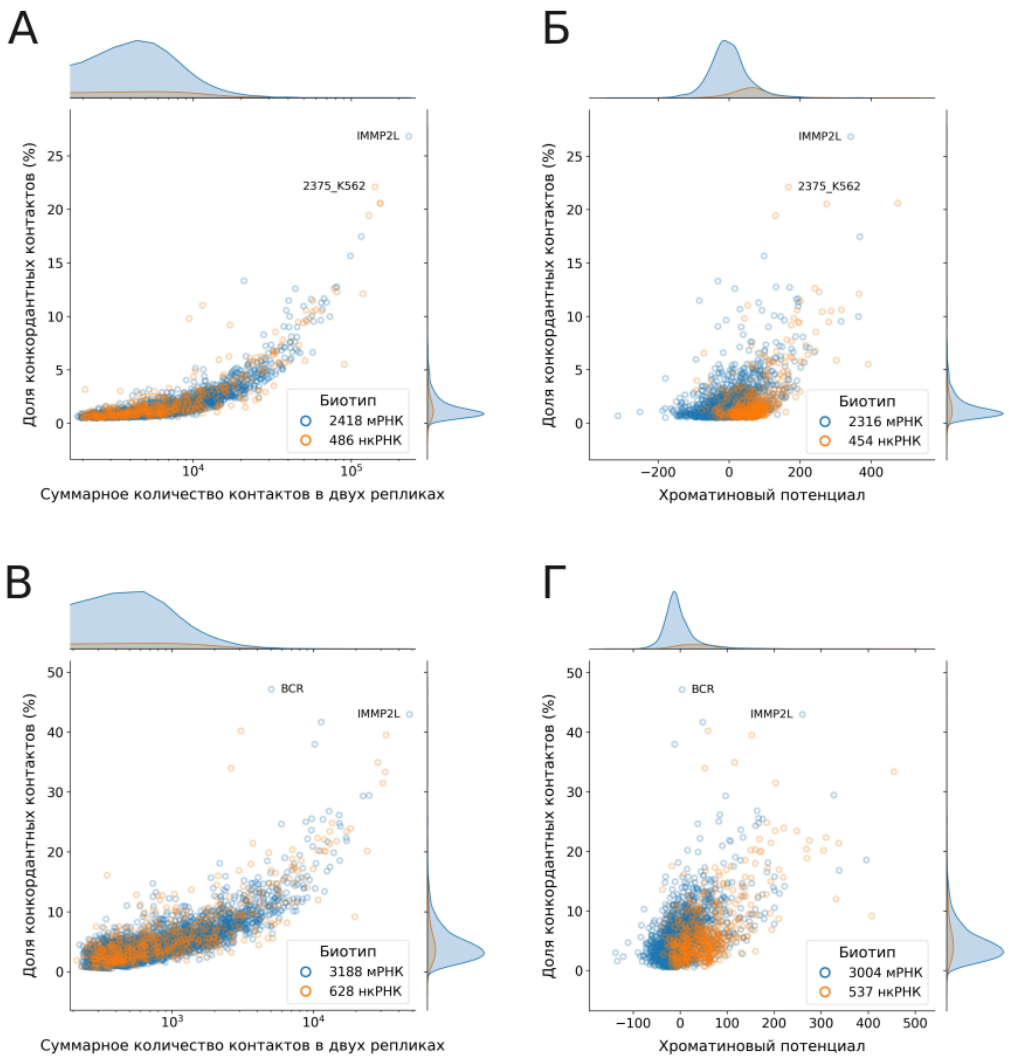


Рисунок 32. Зависимость конкордантности реплик от числа контактов и от хроматинового потенциала. А и Б – Конкордантность рассчитывалась по всем контактам; В и Г – по контактам из пиков BaRDIC. Представлены данные Red-C на клетках K-562, размер бина 5000 п.н. MALAT1 на графике не отображается, так как 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.

Также следует отметить, что медианные доли конкордантности контактов между мРНК и нкРНК практически не отличались (см. приложение В, табл. В.4 и В.5, колонки 1 и 4), что свидетельствует об отсутствии зависимости воспроизводимости от биотипа РНК. Для интерпретации неожиданно высокого уровня конкордантности контактов мРНК, сопоставимого с таковым для нкРНК, были предложены две не исключающие друг друга гипотезы.

1. Существование неспецифических, но статистически воспроизводимых взаимодействий, при которых электростатические или иные слабые силы могут приводить к массовому, но стабильному связыванию РНК с хроматином.
2. Наличие у части мРНК неизвестных специфических функций, связанных с непосредственным взаимодействием с хроматином (например, опосредованных

некодирующими изоформами или неаннотированными интронными нкРНК).

Таким образом, наблюдаемая воспроизводимость контактов части мРНК допускает как минимум две интерпретации: существование устойчивого воспроизводимого фонового сигнала и наличие подмножества действительно функциональных случаев. Более надежное разграничение этих вариантов требует отдельного анализа распределения контактов по экзонным и интронным областям конкретных транскриптов, а также учета локализации контактов относительно статистически значимых пиков. Такой анализ выходит за рамки настоящей работы, поэтому сочетание положительного хроматинового потенциала и высокой конкордантности для мРНК в дальнейшем рассматривается нами прежде всего как критерий отбора кандидатов для углубленного исследования, а не как прямое доказательство хроматиновой функции зрелых мРНК.

3.4.3. Сравнение реплик в данных OTA

Для оценки воспроизводимости экспериментов с индивидуальными РНК мы проанализировали консистентность реплик в соответствующих наборах данных (табл. 5 и приложение В, рис. В.9). Во-первых, был подтвержден ожидаемо высокий уровень воспроизводимости: на полном наборе данных OTA доля конкордантных контактов между репликами превышала 90% уже при размере бина 1000 п.н. Это указывает на то, что данные OTA обладают разрешением в 1000 п.н. и высокой полнотой. Во-вторых, был выявлен критически важный аспект, касающийся специфичности сигнала. При переходе к анализу только тех контактов, которые попадают в пики, выявленные программой BaRDIC (которая отфильтровывает редкие одиночные контакты в пользу статистически значимых кластеров), уровень конкордантности снижался практически вдвое. Это резкое падение позволяет сделать вывод о том, что значительная доля (более половины) всех детектированных контактов, в том числе конкордантных, в экспериментах OTA, вероятно, является неспецифической.

Таблица 5. Доля конкордантных контактов в репликах OTA (%). d0, d3 и d7 – 0, 3 и 7 дней клеточной дифференцировки соответственно; *p-value* < 0,05.

РНК	**Эксперимент**	**Бин = 1000 п.н.**		**Бин = 5000 п.н.**	
		Все контакты, %	**Контакты из пиков, %**	**Все контакты, %**	**Контакты из пиков, %**
JPX	CHART, ES d0 (GSM4278791, GSM4278795)	99,5	53,3	100,0	78,2
JPX	CHART, ES d3 (GSM4278799, GSM4278803)	99,3	36,8	100,0	70,6
JPX	CHART, ES d7 (GSM4278807, GSM4278811)	99,2	44,5	100,0	75,0
MALAT1	ChIRP, ES, genotype: Ythdc1-cKO (conditional); treatment: DMSO, (GSM4669091, GSM4669092)	79,9	26,9	99,6	50,6
MALAT1	ChIRP, ES, genotype: Mettl3-WT, (GSM4875651, GSM4875652)	92,4	40,9	99,9	66,3

Сравнительный анализ воспроизводимости между методами ATA и OTA позволяет сделать следующие выводы:

1. данные ATA (кроме GRID) характеризуются низкой воспроизводимостью между репликами (медианная доля конкордантных контактов <5%), что указывает на их существенную неполноту;
2. данные OTA, напротив, демонстрируют высокую воспроизводимость (>90%), что подтверждает их полноту и позволяет рассматривать их в качестве надежного референса («золотого стандарта») для валидации взаимодействий, выявленных в полногеномных подходах (данные ATA).

3.4.4. Сравнение экспериментов ATA и OTA

Высокая воспроизводимость данных OTA, продемонстрированная в предыдущем разделе, позволяет использовать их в качестве референса для оценки степени согласованности данных

полногеномных подходов (ATA) с этим референсом. Проведение такого сравнительного анализа сопряжено со значительными ограничениями, так как требует наличия данных обоих типов для одних и тех же РНК в идентичных клеточных линиях и со схожими условиями культивирования, а также достаточного количества контактов в данных ATA для обеспечения статистической мощности. Публично доступные данные OTA, соответствующие условиям экспериментов ATA, были найдены лишь для двух РНК – MALAT1 и JPX.

Для нкРНК MALAT1 и JPX мы провели сравнение, используя бины, размером 5000 п.н. В качестве меры согласованности мы рассчитывали долю контактов из данных ATA, которые попадали в бины, обогащенные контактами из пиков BaRDIC соответствующего эксперимента OTA. Анализ проводился как для всех контактов ATA, так и для подмножества, отфильтрованного по пикам BaRDIC. Множества контактов из реплик ATA были объединены для увеличения мощности данных. Результаты для нкРНК MALAT1 представлены в табл. 6 и в приложении В на рис. В.10, для нкРНК JPX – в табл. 7.

Таблица 6. Процент согласованных контактов РНК MALAT1 с хроматином в данных ATA при сравнении с контактами РНК MALAT1 из экспериментов OTA (контакты из пиков BaRDIC) в эмбриональных стволовых клетках мыши. В скобках представлен результат для контактов ATA из пиков BaRDIC. Размер бина – 5000 п.н.; *p-value* < 0,05.

Эксперимент	**Количество контактов в данных ATA**	**RAP**		**ChIRP**		
		pSM33 ES, DMSO 1 hour, %	**V6.5 ES, %**	**ES, Ythdc1-cKO; DMSO, %**	**ES, Mettl3-WT, %**	**E14 ES, %**
GRID, ES, *M. musculus*	522 741 (109 371)	38,0 (46,5)	55,8 (61,6)	50,6 (51,0)	58,8 (58,3)	53,8 (57,0)
RADICL (1% FA), ES, *M. musculus*	636 802 (138 422)	42,9 (56,2)	61,5 (74,1)	50,4 (49,6)	58,9 (57,2)	55,1 (58,6)
RADICL (2% FA), ES, *M. musculus*	484 878 (99 985)	41,5 (51,0)	59,7 (69,0)	50,7 (49,6)	59,2 (58,2)	54,9 (56,8)

Таблица 7. Процент согласованных контактов РНК JPX с хроматином в данных ATA (все контакты) при сравнении с контактами РНК JPX из экспериментов OTA (контакты из пиков BaRDIC) в эмбриональных стволовых клетках мыши. В скобках представлен *p-value* конкордантности. Размер бина – 5000 п.н., d0, d3 и d7 – 0, 3 и 7 дней клеточной дифференцировки соответственно.

Эксперимент	**Количество контактов Jpx в данных ATA**	**CHART, ES**		
		d0	**d3**	**d7**
GRID, ES, *M. musculus*	459	57,1 (0,05)	61,9 (0,22)	63,6 (0,002)
RADICL (1% FA), ES, *M. musculus*	341	57,2 (0,09)	62,8 (0,15)	61,0 (0,03)
RADICL (2% FA), ES, *M. musculus*	332	54,5 (0,24)	56,6 (0,65)	63,9 (0,005)

Для нкРНК MALAT1 с чрезвычайно высоким уровнем взаимодействий в данных ATA была выявлена значительная доля совпадений (~50%) с данными OTA, что указывает на хорошую согласованность методов. При этом примерно половина контактов MALAT1, детектированных только методом ATA, не подтверждается независимым методом OTA, что позволяет оценить долю неспецифического сигнала в данных ATA для данной РНК в ~50%. Важно отметить, что в данном случае мы не наблюдаем значительного преимущества метода GRID-seq, которое было столь явным при анализе консистентности реплик ATA. Это, вероятно, связано с тем, что для MALAT1 общее количество контактов и их воспроизводимость настолько высоки во всех экспериментах ATA, что эффект от более специфичного протокола фиксации нивелируется на фоне доминирующего сигнала.

Ситуация для нкРНК JPX, характеризующейся низким уровнем контактов в данных ATA, фундаментально отличается. Совпадение с данными OTA составило ~60%, что позволяет грубо оценить долю неспецифических контактов в ~40%. Низкое абсолютное число контактов делает эту оценку менее надежной. Как и ожидалось, основываясь на известной ассоциации JPX с XIST [189], большинство ее контактов локализовано на X-хромосоме. Низкое количество контактов не позволило применить фильтрацию по пикам BaRDIC, которая, вероятно, повысила бы специфичность анализа.

Проведенный анализ демонстрирует принципиальную возможность кросс-валидации, но также подчеркивает его ограничения. К сожалению, для РНК со средним уровнем взаимодействий – наиболее интересной для оценки специфичности методов ATA – проведение сравнительного

анализа с ОТА оказалось невозможным из-за отсутствия парных данных в согласованных биологических условиях. Таким образом, данные ОТА служат надежным референсом в первую очередь для высоко контактирующих РНК, в то время как оценка специфичности АТА для остального интерактома требует разработки альтернативных подходов.

3.4.5. Сравнение экспериментов ОТА

Для оценки согласованности данных ОТА мы провели сравнительный анализ карт РНК-хроматиновых взаимодействий для различных нкРНК в клетках человека (рис. 33) и мыши (см. приложение В, рис. В.11). В качестве меры сходства мы использовали отношение конкордантных контактов к общему количеству выявленных взаимодействий в сравниваемых экспериментах ОТА (меру Жаккара).

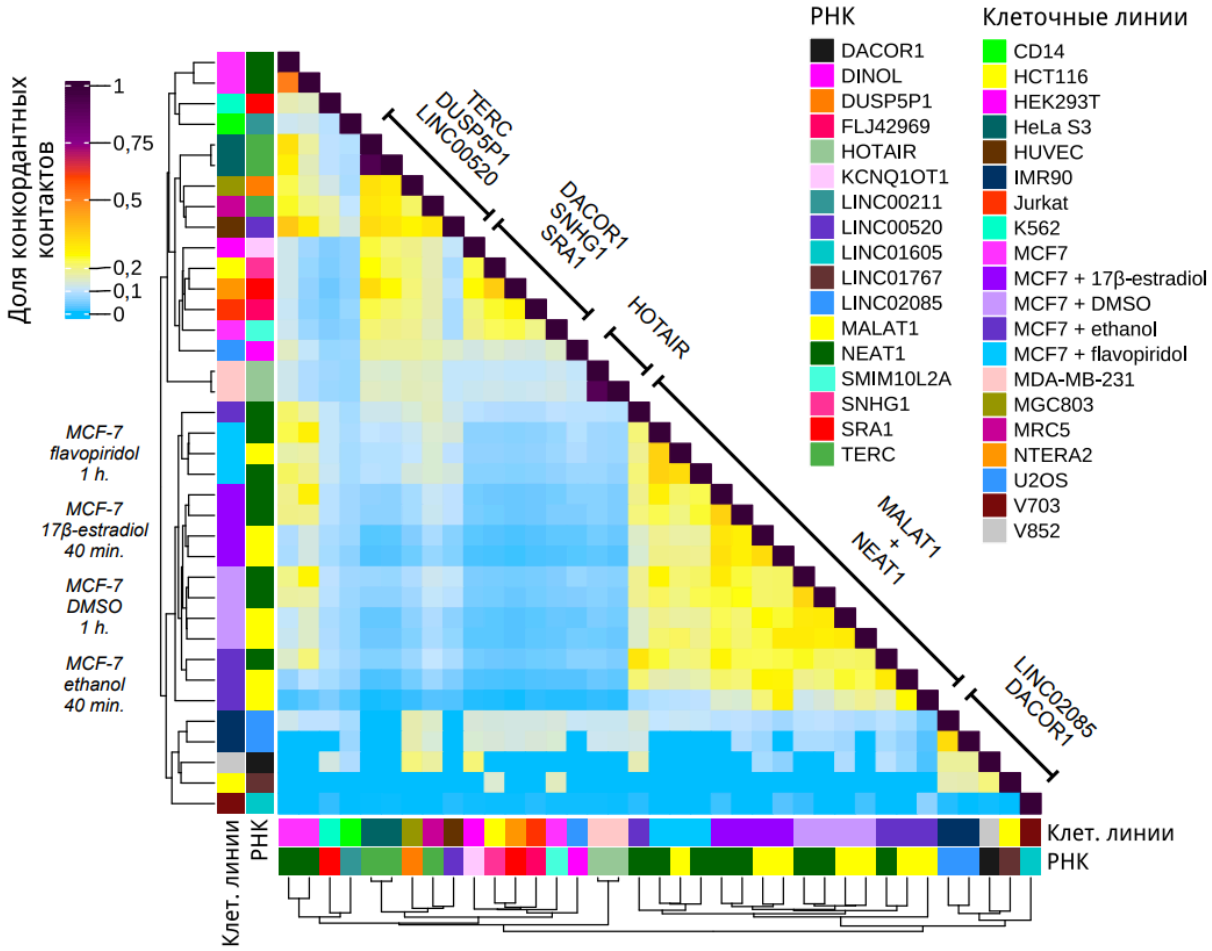


Рисунок 33. Тепловая карта, отражающая долю согласованных контактов (из пиков BaRDIC, FDR < 0,05) из экспериментов группы «один-против-всех» для клеточных линий человека. Незначимые обогащения (*p-value* > 0,05) обнулены. Кластеризация проведена по клеточным типам и РНК, используемым в эксперименте. Размер бина – 1000 п.н.

Анализ выявил кластеры высокой функциональной согласованности, а также перекрытия, вероятно, связанные с общими принципами организации хроматина. На тепловой карте для

человека наблюдаются выраженные кластеры, соответствующие определенным РНК, таким как MALAT1, NEAT1 и HOTAIR. Наиболее ярким примером ожидаемого сходства оказались профили MALAT1 и NEAT1. Высокая конкордантность их хроматиновых контактов хорошо согласуется с их известной колокализацией в ядре: NEAT1 является структурной основой параспеклов, а MALAT1 – ключевым компонентом ядерных спеклов [106,190]. Обе РНК ассоциированы с активными генами и участвуют в регуляции сплайсинга [27], что объясняет сходство их хроматинового ландшафта.

Были также обнаружены перекрытия, например, между контактами LINC02085 и DACOR1. LINC02085 участвует в NF-kB-зависимой регуляции [191], а DACOR1 – в поддержании паттернов метилирования ДНК [89], что может отражать их совместную вовлеченность в эпигенетический контроль.

В то же время анализ выявил кластеры сходства, не имеющие очевидного функционального объяснения. Например, профиль теломеразной РНК TERC показал значительную конкордантность с такими РНК, как SRA1, SNHG1 и KCNQ1OT1, прямые функциональные связи с которыми неизвестны. Этот результат указывает на существование фонового сигнала. Если принять, что большинство детектируемых взаимодействий РНК с хроматином опосредовано белками, то низкую специфичность этих контактов можно объяснить не особенностями экспериментальных методов, а относительно низкой специфичностью РНК-связывающих доменов белков [192,193], приводящей к общим паттернам ассоциации для функционально несвязанных РНК.

В отличие от данных для человека, данные ОТА для мыши в основном посвящены исследованию XIST. Наблюдаемая высокая конкордантность профилей XIST с такими РНК, как ее известный активатор JPX [189], служит дополнительным внутренним контролем качества данных и подтверждает специфичность метода для функционально связанных пар (см. приложение В, рис. В.11).

3.4.6. Стратегия анализа данных РНК-хроматинового интерактома

На основании полученных результатов для повышения достоверности и значимости выводов при работе с данными РНК-хроматинового интерактома мы рекомендуем следующий подход.

- При анализе данных ОТА следует ориентироваться на контакты, прошедшие фильтрацию по пикам (например, с помощью BaRDIC), поскольку они демонстрируют значительно более высокую специфичность. Высокая общая воспроизводимость данных ОТА 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.

Следует отметить ограничения применимости предложенной стратегии. Она разработана на материале общедоступных полногеномных данных РНК-хроматинового интерактома человека и мыши, обработанных по единому протоколу, и потому наиболее надежна именно в этих условиях. Для данных ATA ее применение оправдано при наличии биологических реплик, цепь-ориентированных данных тотального RNA-seq с деплецией рибосомальной РНК, достаточной глубины секвенирования и достаточного числа контактов, позволяющего оценивать хроматиновый потенциал, конкордантность и статистически значимые пики; в рамках настоящей работы это соответствовало отбору РНК с chP > 20 и числом контактов > 10 000. Для низкопокрытых библиотек, низкоэкспрессирующихся РНК и наборов данных без реплик предлагаемая схема может быть недостаточно чувствительной. Для данных OTA наиболее достоверными следует считать контакты, прошедшие фильтрацию по пикам, тогда как внепиковые контакты характеризуются более низкой специфичностью и требуют более осторожной интерпретации.

ЗАКЛЮЧЕНИЕ

В настоящей диссертационной работе решена комплексная задача по разработке методологических основ, инструментов и аналитической инфраструктуры для систематического изучения РНК-хроматинового интерактома.

Была разработана и реализована специализированная аналитическая база данных RNA-Chrom, представляющая собой первый курируемый ресурс, содержащий все доступные полногеномные данные о взаимодействиях РНК с хроматином, полученные методами OTA и ATA для человека, мыши, домашней свиньи и североамериканского красногорлого анолиса. Для обеспечения сопоставимости разрозненных данных был создан и применен единый стандартизированный вычислительный конвейер их обработки. Ресурс снабжен веб-интерфейсом, предоставляющим функционал для двух типов интерактивного анализа («от РНК» и «от ДНК»), что позволяет исследователям эффективно работать с унифицированными данными без необходимости сложной самостоятельной обработки.

Использование единого протокола обработки меняет интерпретацию опубликованных экспериментов по меньшей мере в трех отношениях. Во-первых, результаты, полученные разными методами и в разных работах, становятся непосредственно сопоставимыми, что позволяет отделять устойчиво воспроизводимые особенности контактома от артефактов частных вычислительных конвейеров. Во-вторых, интерпретация смещается от простого перечисления контактирующих участков к оценке их вероятной функциональной значимости за счет анализа статистически значимых пиков, генов в контактирующей области и сопоставления с независимыми эпигенетическими данными. В-третьих, благодаря режиму анализа «от ДНК» опубликованный эксперимент с одной РНК можно рассматривать не изолированно, а в контексте всех РНК, взаимодействующих с тем же локусом, что позволяет точнее оценивать специфичность предполагаемой регуляторной связи.

В биологическом плане это означает переход от разрозненного анализа отдельных экспериментов к унифицированному поиску потенциально функциональных РНК-хроматиновых связей. RNA-Chrom позволяет не только воспроизводить опубликованные наблюдения, но и переинтерпретировать их в общем контексте других экспериментов, геномных локусов и контактирующих РНК. Благодаря этому сервис ускоряет отбор генов-мишеней и регуляторных локусов-кандидатов для последующей экспериментальной проверки.

Для решения критической проблемы вычислительной эффективности на этапе предобработки данных был разработан новый инструмент Fastq-dupaway. Программа характеризуется предсказуемо низким потреблением оперативной памяти (~2 ГБ) и высокой скоростью работы, что позволяет осуществлять дедупликацию больших наборов данных на инфраструктуре с ограниченными ресурсами. Сравнительный анализ подтвердил его

преимущества по соотношению производительности и потребления ресурсов перед существующими аналогами.

Для перехода от карт физических взаимодействий к функциональным гипотезам была реализована интеграция баз данных RNA-Chrom и HiMoRNA. Этот подход позволяет сопоставлять информацию о физической локализации нкРНК в геномных локусах с данными о корреляциях их экспрессии с эпигенетическими метками. На примере lncRNA PVT1 и MEG3 показано, что такая интеграция способствует генерации интерпретируемых гипотез о механизмах эпигенетической регуляции генов, выполняемой длинными нкРНК.

Проведен первый полномасштабный сравнительный анализ характеристик данных, полученных методами OTA и ATA. Введена и апробирована метрика хроматинового потенциала, позволяющая выявлять РНК со статистически значимо повышенным сродством к хроматину. Установлены ключевые различия между методами: данные ATA характеризуются существенно более низким разрешением (~5000 п.н. против ~1000 п.н. у OTA), меньшей воспроизводимостью на уровне отдельных контактов и сильным влиянием протокола фиксации на качество сигнала. На основе полученных результатов предложен двухуровневый подход к анализу данных ATA, включающий фильтрацию РНК по высокому хроматиновому потенциалу с последующим отбором воспроизводимых контактов, принадлежащих статистически значимым пикам.

Следует отметить, что предложенные в работе решения имеют определенные ограничения применимости. База RNA-Chrom и унифицированный вычислительный протокол ориентированы на общедоступные полногеномные данные OTA и ATA и требуют дополнительной адаптации при переносе на иные типы экспериментов. Интеграция RNA-Chrom и HiMoRNA предназначена прежде всего для генерации функциональных гипотез и ограничена полнотой обеих баз данных, особенно для низкоэкспрессирующихся днРНК с малым числом контактов. Выбор режима Fastq-dupaway определяется доступными вычислительными ресурсами: режим «fast» обеспечивает максимальную скорость, но не позволяет ограничивать использование оперативной памяти, тогда как «sequence-based» режимы предпочтительны для больших наборов данных при отсутствии узкого места по операциям ввода-вывода. Предложенная стратегия достоверного анализа наиболее надежна при наличии биологических реплик, после фильтрации по пикам и для данных ATA достаточного покрытия.

Практическая востребованность RNA-Chrom подтверждается тем, что ресурс уже использовался в независимых исследованиях для решения задач более высокого уровня, включая анализ cis-регуляторных длинных некодирующих РНК [203] и построение интегративных моделей РНК–белок–ДНК взаимодействий [193]. 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 СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

РНК – рибонуклеиновая кислота

мРНК – белок-кодирующие РНК

нкРНК – некодирующие РНК

днРНК – длинные некодирующие РНК

ДНК – дезоксирибонуклеиновая кислота

ОТА («один-против-всех» или «one-to-all») – экспериментальные методы, позволяющие определить контакты ранее известной РНК с хроматином

АТА («все-против-всех» или «all-to-all») – экспериментальные методы, направленные на определение всех возможных контактов РНК-ДНК в клетке

хаРНК – хроматин-ассоциированная РНК

ПЦР – полимеразная цепная реакция

NGS – next-generation sequencing, секвенирование нового поколения

ГБ - гигабайт

п.н. – пара нуклеотидов

пик – участок генома, обогащенный контактами данной РНК с хроматином

input – данные с фоновыми или неспецифическими контактами

ТАД – топологически ассоциированный домен

хрДНК – хроматиновая ДНК

кДНК – комплементарная ДНК

БД – база данных

RAM – оперативная память

ucaRNAs – unannotated chromatin-associated RNAs, неаннотированные хроматин-ассоциированные РНК

ТБ – терабайт

CPU – центральный процессор

мин – минута

с – секунда

FDR – частота ложных обнаружений

I/O – input-output, ввод-вывод

GEO – Gene Expression Omnibus, архивами «сырых» экспериментальных данных

РД-скейлинг – зависимость плотности контактов РНК от расстояния между геном-источником РНК и целевыми локусами ДНК, расположенными на той же хромосоме

chP – хроматиновый потенциал

DSG – дисукцинимидилглутарат

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Djebali S., Davis C.A., Merkel A., Dobin A., Lassmann T., Mortazavi A., Tanzer A., Lagarde J., Lin W., Schlesinger F., Xue C., Marinov G.K., Khatun J., Williams B.A., Zaleski C., Rozowsky J., Röder M., Kokocinski F., Abdelhamid R.F., Alioto T., Antoshechkin I., Baer M.T., Bar N.S., Batut P., Bell K., Bell I., Chakrabortty S., Chen X., Chrast J., Curado J., Derrien T., Drenkow J., Dumais E., Dumais J., Duttagupta R., Falconnet E., Fastuca M., Fejes-Toth K., Ferreira P., Foissac S., Fullwood M.J., Gao H., Gonzalez D., Gordon A., Gunawardena H., Howald C., Jha S., Johnson R., Kapranov P., King B., Kingswood C., Luo O.J., Park E., Persaud K., Preall J.B., Ribeca P., Risk B., Robyr D., Sammeth M., Schaffer L., See L.-H., Shahab A., Skancke J., Suzuki A.M., Takahashi H., Tilgner H., Trout D., Walters N., Wang H., Wrobel J., Yu Y., Ruan X., Hayashizaki Y., Harrow J., Gerstein M., Hubbard T., Reymond A., Antonarakis S.E., Hannon G., Giddings M.C., Ruan Y., Wold B., Carninci P., Guigó R., Gingeras T.R. *Landscape of transcription in human cells* // Nature 2012.– Vol. 489 № 7414.– P. 101–108.
2. Engreitz J., Ollikainen N., Guttman M. *Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression* // 2016.– Vol.17.– P. 756–770.
3. Mishra K., Kanduri C., Kanduri C. *Understanding Long Noncoding RNA and Chromatin Interactions: What We Know So Far* // Non-Coding RNA 2019.– Vol. 5 № 4.– P. 1–28.
4. Engreitz J.M., Pandya-Jones A., McDonel P., Shishkin A., Sirokman K., Surka C., Kadri S., Xing J., Goren A., Lander E.S., Plath K., Guttman M. *The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome* // Science 2013.– Vol. 341 № 6147.– P. 1–8.
5. Simon M.D., Wang C.I., Kharchenko P.V., West J.A., Chapman B.A., Alekseyenko A.A., Borowsky M.L., Kuroda M.I., Kingston R.E. *The genomic binding sites of a noncoding RNA* // PNAS 2011.– Vol. 108 № 51.– P. 20497–20502.
6. Chu C., Qu K., Zhong F.L., Artandi S.E., Chang H.Y. *Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions* // Molecular Cell 2011.– Vol. 44 № 4.– P. 667–678.
7. Quinn J.J., Ilik I.A., Qu K., Georgiev P., Chu C., Akhtar A., Chang H.Y. *Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification* // Nature Biotechnology 2014.– Vol. 32 № 9.– P. 933–940.
8. Mondal T., Subhash S., Vaid R., Enroth S., Uday S., Reinius B., Mitra S., Mohammed A., James A.R., Hoberg E., Moustakas A., Gyllensten U., Jones S.J.M., Gustafsson C.M., Sims A.H., Westerlund F., Gorab E., Kanduri C. *MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures* // Nature Communications 2015.– Vol. 6 № 7743.– P. 1–17.
9. Chu H.-P., Cifuentes-Rojas C., Kesner B., Aeby E., Lee H.-g., Wei C., Oh H.J., Boukhali M., Haas W., Lee J.T. *TERRA RNA Antagonizes ATRX and Protects Telomeres* // Cell 2017.– Vol. 170 № 1.– P. 86-101.e16.
10. Sridhar B., Rivas-Astroza M., Nguyen T.C., Chen W., Yan Z., Cao X., Hebert L., Zhong S. *Systematic Mapping of RNA-Chromatin Interactions In Vivo* // Current Biology 2017.– Vol. 27 № 4.– P. 602–609.
11. Li X., Zhou B., Chen L., Gou L.-T., Li H., Fu X.-D. *GRID-seq reveals the global RNA–chromatin interactome* // Nature Biotechnology 2017.– Vol. 35 № 10.– P. 940–950.
12. Bell J.C., Jukam D., Teran N.A., Risca V.I., Smith O.K., Johnson W.L., Skotheim J.M.,

Greenleaf W.J., Straight A.F. *Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts* // eLife 2018.– Vol.7.– P. 1–28.
13. Yan Z., Huang N., Wu W., Chen W., Jiang Y., Chen J., Huang X., Wen X., Xu J., Jin Q., Zhang K., Chen Z., Chien S., Zhong S. *Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs* // PNAS 2019.– Vol. 116 № 8.– P. 3328–3337.
14. Calandrelli R., Xu L., Luo Y., Wu W., Fan X., Nguyen T., Chen C.-J., Sriram K., Tang X., Burns A.B., Natarajan R., Chen Z.B., Zhong S. *Stress-induced RNA–chromatin interactions promote endothelial dysfunction* // Nature Communications 2020.– Vol. 11 № 5211.– P. 1–13.
15. Bonetti A., Agostini F., Suzuki A.M., Hashimoto K., Pascarella G., Gimenez J., Roos L., Nash A.J., Ghilotti M., Cameron C.J.F., Valentine M., Medvedeva Y.A., Noguchi S., Agirre E., Kashi K., Samudyata, Luginbühl J., Cazzoli R., Agrawal S., Luscombe N.M., Blanchette M., Kasukawa T., Hoon M.D., Arner E., Lenhard B., Plessy C., Castelo-Branco G., Orlando V., Carninci P. *RADICL-seq identifies general and cell type–specific principles of genome-wide RNA-chromatin interactions* // Nature Communications 2020.– Vol. 11 № 1018.– P. 1–14.
16. Gavrilov A.A., Zharikova A.A., Galitsyna A.A., Luzhin A.V., Rubanova N.M., Golov A.K., Petrova N.V., Logacheva M.D., Kantidze O.L., Ulianov S.V., Magnitov M.D., Mironov A.A., Razin S.V. *Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics* // Nucleic Acids Research 2020.– Vol. 48 № 12.– P. 6699–6714.
17. Tuna M., Mills G.B., Amos C.I. *The role of long non‐coding RNAs in lung cancer metastasis: Molecular mechanisms, pathogenesis and clinical implications* // Clinical and Translational Medicine 2025.– Vol. 15 № e70429.– P. 1–28.
18. Xu J., Xu J., Liu X., Jiang J. *The role of lncRNA-mediated ceRNA regulatory networks in pancreatic cancer* // Cell Death Discovery 2022.– Vol. 8 № 287.– P. 287.
19. Nylund P., Garrido-Zabala B., Kalushkova A., Wiklund H.J. *The complex nature of lncRNA-mediated chromatin dynamics in multiple myeloma* // Frontiers in Oncology 2023.– Vol. 13 № 1303677.– P. 1–9.
20. Yang S., Yang H., Luo Y., Deng X., Zhou Y., Hu B. *Long non-coding RNAs in neurodegenerative diseases* // Neurochemistry International 2021.– Vol. 148 № 105096.
21. Esmaeili A., Yazdanpanah N., Rezaei N. *LncRNAs Orchestrating Neuroinflammation: A Comprehensive Review* // Cellular and Molecular Neurobiology 2025.– Vol. 45 № 21.– P. 1–27.
22. Xu W., Wu Q., Huang A. *Emerging Role of LncRNAs in Autoimmune Lupus* // Inflammation 2022.– Vol. 45 № 3.– P. 937–948.
23. Zhang Y., Wang X., Zhang C., Yi H. *The dysregulation of lncRNAs by epigenetic factors in human pathologies* // Drug Discovery Today 2023.– Vol. 28 № 9.
24. 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.– Vol. 56 № 2.– P. 275–295.
25. Forrest M.E., Saiakhova A., Beard L., Buchner D.A., Scacheri P.C., LaFramboise T., Markowitz S., Khalil A.M. *Colon Cancer-Upregulated Long Non-Coding RNA lincDUSP Regulates Cell Cycle Genes and Potentiates Resistance to Apoptosis* // Scientific Reports 2018.– Vol. 8 № 7324.– P. 1–12.
26. Engreitz J.M., Sirokman K., McDonel P., Shishkin A.A., Surka C., Russell P., Grossman S.R., Chow A.Y., Guttman M., Lander E.S. *RNA-RNA Interactions Enable Specific Targeting of*

Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites // Cell 2014.– Vol. 159 № 1.– P. 188–199.

27. West J.A., Davis C.P., Sunwoo H., Simon M.D., Sadreyev R.I., Wang P.I., Tolstorukov M.Y., Kingston R.E. *The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites* // Molecular Cell 2014.– Vol. 55 № 5.– P. 791–802.
28. Zhao L., Wang J., Li Y., Song T., Wu Y., Fang S., Bu D., Li H., Sun L., Pei D., Zheng Y., Huang J., Xu M., Chen R., Zhao Y., He S. *NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants* // Nucleic Acids Research 2021.– Vol.49.– P. D165–D171.
29. Volders P.-J., Anckaert J., Verheggen K., Nuytens J., Martens L., Mestdagh P., Vandesompele J. *LNCipedia 5: towards a reference set of human long non-coding RNAs* // Nucleic Acids Research 2019.– Vol.47.– P. D135–D139.
30. Kang J., Tang Q., He J., Li L., Yang N., Yu S., Wang M., Zhang Y., Lin J., Cui T., Hu Y., Tan P., Cheng J., Zheng H., Wang D., Su X., Chen W., Huang Y. *RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility* // Nucleic Acids Research 2022.– Vol.50.– P. D326–D332.
31. Yu F., Zhang G., Shi A., Hu J., Li F., Zhang X., Zhang Y., Huang J., Xiao Y., Li X., Cheng S. *LnChrom: a resource of experimentally validated lncRNA–chromatin interactions in human and mouse* // Database 2018.– Vol.2018.– P. 1–7.
32. Mattick J.S., Amaral P.P., Carninci P., Carpenter S., Chang H.Y., Chen L.-L., Chen R., Dean C., Dinger M.E., Fitzgerald K.A., Gingeras T.R., Guttman M., Hirose T., Huarte M., Johnson R., Kanduri C., Kapranov P., Lawrence J.B., Lee J.T., Mendell J.T., Mercer T.R., Moore K.J., Nakagawa S., Rinn J.L., Spector D.L., Ulitsky I., Wan Y., Wilusz J.E., Wu M. *Long non-coding RNAs: definitions, functions, challenges and recommendations* // Nature Reviews Molecular Cell Biology 2023.– Vol. 24 № 6.– P. 430–447.
33. Limouse C., Smith O.K., Jukam D., Fryer K.A., Greenleaf W.J., Straight A.F. *Global mapping of RNA-chromatin contacts reveals a proximity-dominated connectivity model for ncRNA-gene interactions* // Nature Communications 2023.– Vol. 14 № 6073.– P. 1–21.
34. Pandey R.R., Mondal T., Mohammad F., Enroth S., Redrup L., Komorowski J., Nagano T., Mancini-Dinardo D., Kanduri C. *Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation* // Mol Cell 2008.– Vol. 32 № 2.– P. 232–246.
35. Man H.S.J., Sukumar A.N., Lam G.C., Turgeon P.J., Yan M.S., Ku K.H., Dubinsky M.K., Ho J.J.D., Wang J.J., Das S., Mitchell N., Oettgen P., Sefton M.V., Marsden P.A. *Angiogenic patterning by STEEL, an endothelial-enriched long noncoding RNA* // PNAS 2018.– Vol. 115 № 10.– P. 2401–2406.
36. Colak D., Zaninovic N., Cohen M.S., Rosenwaks Z., Yang W.-Y., Gerhardt J., Disney M.D., Jaffrey S.R. *Promoter-Bound Trinucleotide Repeat mRNA Drives Epigenetic Silencing in Fragile X Syndrome* // Science 2014.– Vol. 343 № 6174.– P. 1002–1005.
37. Colognori D., Sunwoo H., Kriz A.J., Wang C.-Y., Lee J.T. *Xist Deletional Analysis Reveals an Interdependency between Xist RNA and Polycomb Complexes for Spreading along the Inactive X* // Molecular Cell 2019.– Vol. 74 № 1.– P. 101-117.e10.
38. Pandya-Jones A., Markaki Y., Serizay J., Chitiashvili T., Leon W.R.M., Damianov A., Chronis C., Papp B., Chen C.-K., McKee R., Wang X.-J., Chau A., Sabri S., Leonhardt H., Zheng S., Guttman M., Black D.L., Plath K. *A protein assembly mediates Xist localization and gene silencing* // Nature 2020.– Vol. 587 № 7832.– P. 145–151.

39. Chen C.-K., Blanco M., Jackson C., Aznauryan E., Ollikainen N., Surka C., Chow A., Cerase A., McDonel P., Guttman M. *Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing* // Science 2016.– Vol. 354 № 6311.– P. 468–472.
40. Alfeghaly C., Sanchez A., Rouget R., Thuillier Q., Igel-Bourguignon V., Marchand V., Branlant C., Motorin Y., Behm-Ansmant I., Maenner S. *Implication of repeat insertion domains in the trans-activity of the long non-coding RNA ANRIL* // Nucleic Acids Research 2021.– Vol. 49 № 9.– P. 4954–4970.
41. Simon M.D., Pinter S.F., Fang R., Sarma K., Rutenberg-Schoenberg M., Bowman S.K., Kesner B.A., Maier V.K., Kingston R.E., Lee J.T. *High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation* // 2013.– Vol. 504 № 7480.– P. 465–469.
42. Ang C.E., Ma Q., Wapinski O.L., Fan S., Flynn R.A., Lee Q.Y., Coe B., Onoguchi M., Olmos V.H., Do B.T., Dukes-Rimsky L., Xu J., Tanabe K., Wang L., Elling U., Penninger J.M., Zhao Y., Qu K., Eichler E.E., Srivastava A., Wernig M., Chang H.Y. *The novel lncRNA lnc-NR2F1 is pro-neurogenic and mutated in human neurodevelopmental disorders* // eLife 2019.– Vol.8.– P. 1–29.
43. Luo M., Jeong M., Sun D., Park H.J., Rodriguez B.A.T., Xia Z., Yang L., Zhang X., Sheng K., Darlington G.J., Li W., Goodell M.A. *Long Non-Coding RNAs Control Hematopoietic Stem Cell Function* // Cell Stem Cell 2015.– Vol. 16 № 4.– P. 426–438.
44. Carlson H.L., Quinn J.J., Yang Y.W., Thornburg C.K., Chang H.Y., Stadler H.S. *LncRNA-HIT Functions as an Epigenetic Regulator of Chondrogenesis through Its Recruitment of p100/CBP Complexes* // PLOS Genetics 2015.– Vol. 11 № 12.– P. 1–30.
45. Yin Y., Yan P., Lu J., Song G., Zhu Y., Li Z., Zhao Y., Shen B., Huang X., Zhu H., Orkin S.H., Shen X. *Opposing Roles for the lncRNA Haunt and Its Genomic Locus in Regulating HOXA Gene Activation during Embryonic Stem Cell Differentiation* // Cell Stem Cell 2015.– Vol. 16 № 5.– P. 504–516.
46. Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W., Liu X.S. *Model-based Analysis of ChIP-Seq (MACS)* // Genome Biology 2008.– Vol. 9 № 9.– P. 1–9.
47. Xu S., Grullon S., Ge K., Peng W. *Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) to Map Regions of Histone Methylation Patterns in Embryonic Stem Cells* // Methods Mol Biol 2014.– Vol.1150.– P. 97–111.
48. Heinz S., Benner C., Spann N., Bertolino E., Lin Y.C., Laslo P., Cheng J.X., Murre C., Singh H., Glass C.K. *Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities* // Molecular Cell 2010.– Vol. 38 № 4.– P. 576–589.
49. Kharchenko P.V., Tolstorukov M.Y., Park P.J. *Design and analysis of ChIP-seq experiments for DNA-binding proteins* // Nature Biotechnology 2008.– Vol. 26 № 12.– P. 1351–1359.
50. Chalei V., Sansom S.N., Kong L., Lee S., Montiel J.F., Vance K.W., Ponting C.P. *The long non-coding RNA Dali is an epigenetic regulator of neural differentiation* // eLife 2014.– Vol.3.– P. 1–24.
51. Chu H.-P., Froberg J.E., Kesner B., Oh H.J., Ji F., Sadreyev R., Pinter S.F., Lee J.T. *PAR-TERRA directs homologous sex chromosome pairing* // Nature Structural & Molecular Biology 2017.– Vol. 24 № 8.– P. 620–631.
52. Alvarez-Dominguez J.R., Knoll M., Gromatzky A.A., Lodish H.F. *The Super-Enhancer-Derived alncRNA-EC7/ Bloodlinc Potentiates Red Blood Cell Development*

in trans // Cell Rep. 2017.– Vol. 19 № 12.– P. 2503–2514.
53. Wang Y., Zhu P., Luo J., Wang J., Liu Z., Wu W., Du Y., Ye B., Wang D., He L., Ren W., Wang J., Sun X., Chen R., Tian Y., Fan Z. *LncRNA HAND2‐AS1 promotes liver cancer stem cell self‐renewal via BMP signaling* // The EMBO Journal 2019.– Vol. 38 № 17.– P. 1–17.
54. Wang C.-Y., Jégu T., Chu H.-P., Oh H.J., Lee J.T. *SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X* // Cell 2018.– Vol. 174 № 2.– P. 406-421.e25.
55. Rice P., Longden L., Bleasby A. *EMBOSS: the European Molecular Biology Open Software Suite* // Trends Genet. 2000.– Vol. 16 № 6.– P. 276–277.
56. Vance K.W., Sansom S.N., Lee S., Chalei V., Kong L., Cooper S.E., Oliver P.L., Ponting C.P. *The long non-coding RNA Paupar regulates the expression of both local and distal genes* // The EMBO Journal 2014.– Vol. 33 № 4.– P. 296–311.
57. Buske F.A., Bauer D.C., Mattick J.S., Bailey T.L. *Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data* // Genome Research 2012.– Vol. 22 № 7.– P. 1372–1381.
58. Kuo C.-C., Hänzelmann S., Cetin N.S., Frank S., Zajzon B., Derks J.-P., Akhade V.S., Ahuja G., Kanduri C., Grummt I., Kurian L., Costa I.G. *Detection of RNA–DNA binding sites in long noncoding RNAs* // Nucleic Acids Research 2019.– Vol. 47 № 6.– P. 1–12.
59. Zapparoli E., Briata P., Rossi M., Brondolo L., Bucci G., Gherzi R. *Comprehensive multi-omics analysis uncovers a group of TGF-β-regulated genes among lncRNA EPR direct transcriptional targets* // Nucleic Acids Research 2020.– Vol. 48 № 16.– P. 9053–9066.
60. Matveishina E., Antonov I., Medvedeva Y.A. *Practical Guidance in Genome-Wide RNA:DNA Triple Helix Prediction* // International Journal of Molecular Sciences 2020.– Vol. 21 № 3.– P. 1–12.
61. Bailey T.L., Johnson J., Grant C.E., Noble W.S. *The MEME Suite* // Nucleic Acids Research 2015.– Vol.43.– P. W39–W49.
62. Gupta S., Stamatoyannopoulos J.A., Bailey T.L., Noble W.S. *Quantifying similarity between motifs* // Genome Biology 2007.– Vol. 8 № 2.– P. 1–9.
63. Lee H.C., Kang D., Han N., Lee Y., Hwang H.J., Lee S.-B., You J.S., Min B.S., Park H.J., Ko Y.-G., Gorospe M., Lee J.-S. *A novel long noncoding RNA Linc-ASEN represses cellular senescence through multileveled reduction of p21 expression* // Cell Death & Differentiation 2020.– Vol. 27 № 6.– P. 1844–1861.
64. Long J., Badal S.S., Ye Z., Wang Y., Ayanga B.A., Galvan D.L., Green N.H., Chang B.H., Overbeek P.A., Danesh F.R. *Long noncoding RNA Tug1 regulates mitochondrial bioenergetics in diabetic nephropathy* // Journal of Clinical Investigation 2016.– Vol. 126 № 11.– P. 4205–4218.
65. Martianov I., Ramadass A., Barros A.S., Chow N., Akoulitchev A. *Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript* // Nature 2007.– Vol. 445 № 7128.– P. 666–670.
66. Khomyakova E.B. *Parallel intramolecular DNA triple helix with G and T bases in the third strand stabilized by Zn2+ ions* // Nucleic Acids Research 2000.– Vol. 28 № 18.– P. 3511–3516.
67. Besch R., Giovannangeli C., Kammerbauer C., Degitz K. *Specific inhibition of ICAM-1 expression mediated by gene targeting with Triplex-forming oligonucleotides* // J. Biol. Chem. 2002.– Vol. 277 № 36.– P. 32473–32479.
68. McLean C.Y., Bristor D., Hiller M., Clarke S.L., Schaar B.T., Lowe C.B., Wenger A.M.,

Bejerano G. *GREAT improves functional interpretation of cis-regulatory regions* // Nature Biotechnology 2010.– Vol. 28 № 5.– P. 495–501.
69. Shin H., Liu T., Manrai A.K., Liu X.S. *CEAS: cis-regulatory element annotation system* // Bioinformatics 2009.– Vol. 25 № 19.– P. 2605–2606.
70. Mi H., Muruganujan A., Ebert D., Huang X., Thomas P.D. *PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools* // Nucleic Acids Research 2019.– Vol.47.– P. D419–D426.
71. Huang D.W., Sherman B.T., Lempicki R.A. *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources* // Nat. Protoc. 2009.– Vol. 4 № 1.– P. 44–57.
72. Binns D., Dimmer E., Huntley R., Barrell D., O'Donovan C., Apweiler R. *QuickGO: a web-based tool for Gene Ontology searching* // Bioinformatics 2009.– Vol. 25 № 22.– P. 3045–3046.
73. Kuleshov M.V., Jones M.R., Rouillard A.D., Fernandez N.F., Duan Q., Wang Z., Koplev S., Jenkins S.L., Jagodnik K.M., Lachmann A., McDermott M.G., Monteiro C.D., Gundersen G.W., Ma'ayan A. *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update* // Nucleic Acids Research 2016.– Vol.44.– P. W90–W97.
74. Léveillé N., Melo C.A., Rooijers K., Díaz-Lagares A., Melo S.A., Korkmaz G., Lopes R., Moqadam F.A., Maia A.R., Wijchers P.J., Geeven G., Boer M.L.D., Kalluri R., Laat W.D., Esteller M., Agami R. *Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA* // Nature Communications 2015.– Vol. 6 № 6520.– P. 1–12.
75. Luo S., Lu J.Y., Liu L., Yin Y., Chen C., Han X., Wu B., Xu R., Liu W., Yan P., Shao W., Lu Z., Li H., Na J., Tang F., Wang J., Zhang Y.E., Shen X. *Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells* // Cell Stem Cell 2016.– Vol. 18 № 5.– P. 637–652.
76. Hu T., Pi W., Zhu X., Yu M., Ha H., Shi H., Choi J.-H., Tuan D. *Long non-coding RNAs transcribed by ERV-9 LTR retrotransposon act in cis to modulate long-range LTR enhancer function* // Nucleic Acids Research 2017.– Vol. 45 № 8.– P. 4479–4492.
77. Ballarino M., Cipriano A., Tita R., Santini T., Desideri F., Morlando M., Colantoni A., Carrieri C., Nicoletti C., Musarò A., O' Carroll D., Bozzoni I. *Deficiency in the nuclear long noncoding RNA Charme causes myogenic defects and heart remodeling in mice* // The EMBO Journal 2018.– Vol. 37 № 18.– P. 1–16.
78. Wang C.-Y., Colognori D., Sunwoo H., Wang D., Lee J.T. *PRC1 collaborates with SMCHD1 to fold the X-chromosome and spread Xist RNA between chromosome compartments* // Nature Communications 2019.– Vol. 10 № 1.– P. 1–18.
79. Flynn R.A., Do B.T., Rubin A.J., Calo E., Lee B., Kuchelmeister H., Rale M., Chu C., Kool E.T., Wysocka J., Khavari P.A., Chang H.Y. *7SK-BAF axis controls pervasive transcription at enhancers* // Nature Structural & Molecular Biology 2016.– Vol. 23 № 3.– P. 231–238.
80. Wongtrakoongate P., Riddick G., Fucharoen S., Felsenfeld G. *Association of the Long Non-coding RNA Steroid Receptor RNA Activator (SRA) with TrxG and PRC2 Complexes* // PLOS Genetics 2015.– Vol. 11 № 10.– P. 1–20.
81. Marión R.M., Montero J.J., Silanes I.L.D., Graña-Castro O., Martínez P., Schoeftner S., Palacios-Fábrega J.A., Blasco M.A. *TERRA regulate the transcriptional landscape of pluripotent cells through TRF1-dependent recruitment of PRC2* // eLife 2019.– Vol.8.– P. 1–32.
82. Quinn J.J., Zhang Q.C., Georgiev P., Ilik I.A., Akhtar A., Chang H.Y. *Rapid evolutionary*

turnover underlies conserved lncRNA–genome interactions // Genes & Development 2016.– Vol. 30 № 2.– P. 191–207.
83. Zovoilis A., Cifuentes-Rojas C., Chu H.-P., Hernandez A.J., Lee J.T. *Destabilization of B2 RNA by EZH2 Activates the Stress Response* // Cell 2016.– Vol. 167 № 7.– P. 1788–1802.
84. Chodroff R.A., Goodstadt L., Sirey T.M., Oliver P.L., Davies K.E., Green E.D., Molnár Z., Ponting C.P. *Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes* // Genome Biology 2010.– Vol. 11 № 7.– P. 1–16.
85. Ulitsky I., Shkumatava A., Jan C.H., Sive H., Bartel D.P. *Conserved Function of lincRNAs in Vertebrate Embryonic Development despite Rapid Sequence Evolution* // Cell 2011.– Vol. 147 № 7.– P. 1537–1550.
86. Hacisuleyman E., Goff L.A., Trapnell C., Williams A., Henao-Mejia J., Sun L., McClanahan P., Hendrickson D.G., Sauvageau M., Kelley D.R., Morse M., Engreitz J., Lander E.S., Guttman M., Lodish H.F., Flavell R., Raj A., Rinn J.L. *Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre* // Nature Structural & Molecular Biology 2014.– Vol. 21 № 2.– P. 198–206.
87. Schmitt A.M., Garcia J.T., Hung T., Flynn R.A., Shen Y., Qu K., Payumo A.Y., Peres-da-Silva A., Broz D.K., Baum R., Guo S., Chen J.K., Attardi L.D., Chang H.Y. *An inducible long noncoding RNA amplifies DNA damage signaling* // Nature Genetics 2016.– Vol. 48 № 11.– P. 1370–1376.
88. Li M.A., Amaral P.P., Cheung P., Bergmann J.H., Kinoshita M., Kalkan T., Ralser M., Robson S., Meyenn F.V., Paramor M., Yang F., Chen C., Nichols J., Spector D.L., Kouzarides T., He L., Smith A. *A lncRNA fine tunes the dynamics of a cell state transition involving Lin28, let-7 and de novo DNA methylation* // eLife 2017.– Vol.6.– P. 1–24.
89. Merry C.R., Forrest M.E., Sabers J.N., Beard L., Gao X.-H., Hatzoglou M., Jackson M.W., Wang Z., Markowitz S.D., Khalil A.M. *DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer* // Human Molecular Genetics 2015.– Vol. 24 № 21.– P. 6240–6253.
90. Gelbart M.E., Kuroda M.I. *Drosophila dosage compensation: a complex voyage to the X chromosome* // Development 2009.– Vol. 136 № 9.– P. 1399–1410.
91. Larschan E., Bishop E.P., Kharchenko P.V., Core L.J., Lis J.T., Park P.J., Kuroda M.I. *X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila* // Nature 2011.– Vol. 471 № 7336.– P. 115–118.
92. Meller V.H., Rattner B.P. *The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex* // EMBO J. 2002.– Vol. 21 № 5.– P. 1084–1091.
93. Disteche C.M. *Dosage compensation of the sex chromosomes* // Annu. Rev. Genet. 2012.– Vol.46.– P. 537–560.
94. Lee J.T. *Epigenetic regulation by long noncoding RNAs* // Science 2012.– Vol. 338 № 6113.– P. 1436–1439.
95. Wutz A. *Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation* // Nat. Rev. Genet. 2011.– Vol. 12 № 8.– P. 542–553.
96. Giorgetti L., Lajoie B.R., Carter A.C., Attia M., Zhan Y., Xu J., Chen C.J., Kaplan N., Chang H.Y., Heard E., Dekker J. *Structural organization of the inactive X chromosome in the mouse* // Nature 2016.– Vol. 535 № 7613.– P. 575–579.
97. Splinter E., Wit E.D., Nora E.P., Klous P., Van De Werken H.J.G., Zhu Y., Kaaij L.J.T., Van IJcken W., Gribnau J., Heard E., De Laat W. *The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA* // Genes & Development

2011.– Vol. 25 № 13.– P. 1371–1383.
98. Minajigi A., Froberg J.E., Wei C., Sunwoo H., Kesner B., Colognori D., Lessing D., Payer B., Boukhali M., Haas W., Lee J.T. *A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation* // Science 2015.– Vol. 349 № 6245.– P. 1–12.
99. Schoeftner S., Sengupta A.K., Kubicek S., Mechtler K., Spahn L., Koseki H., Jenuwein T., Wutz A. *Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing* // The EMBO Journal 2006.– Vol. 25 № 13.– P. 3110–3122.
100. Zhao J., Sun B.K., Erwin J.A., Song J.-J., Lee J.T. *Polycomb Proteins Targeted by a Short Repeat RNA to the Mouse X Chromosome* // Science 2008.– Vol. 322 № 5902.– P. 750–756.
101. Chu C., Zhang Q.C., Rocha S.T., Flynn R.A., Bharadwaj M., Calabrese J.M., Magnuson T., Heard E., Chang H.Y. *Systematic Discovery of Xist RNA Binding Proteins* // Cell 2015.– Vol. 161 № 2.– P. 404–416.
102. Brockdorff N. *Local Tandem Repeat Expansion in Xist RNA as a Model for the Functionalisation of ncRNA* // Non-coding RNA 2018. – Vol. 4 № 4.– P. 1–11.
103. McHugh C.A., Chen C.-K., Chow A., Surka C.F., Tran C., McDonel P., Pandya-Jones A., Blanco M., Burghard C., Moradian A., Sweredoski M.J., Shishkin A.A., Su J., Lander E.S., Hess S., Plath K., Guttman M. *The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3* // Nature 2015.– Vol. 521 № 7551.– P. 232–236.
104. Hutchinson J.N., Ensminger A.W., Clemson C.M., Lynch C.R., Lawrence J.B., Chess A. *A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains* // BMC Genomics 2007.– Vol.8.– P. 1–16.
105. Zhang X., Hamblin M.H., Yin K.-J. *The long noncoding RNA Malat1: Its physiological and pathophysiological functions* // RNA Biol. 2017.– Vol. 14 № 12.– P. 1705–1714.
106. Clemson C.M., Hutchinson J.N., Sara S.A., Ensminger A.W., Fox A.H., Chess A., Lawrence J.B. *An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles* // Molecular Cell 2009.– Vol. 33 № 6.– P. 717–726.
107. Imamura K., Imamachi N., Akizuki G., Kumakura M., Kawaguchi A., Nagata K., Kato A., Kawaguchi Y., Sato H., Yoneda M., Kai C., Yada T., Suzuki Y., Yamada T., Ozawa T., Kaneki K., Inoue T., Kobayashi M., Kodama T., Wada Y., Sekimizu K., Akimitsu N. *Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli* // Molecular Cell 2014.– Vol. 53 № 3.– P. 393–406.
108. Hirose T., Virnicchi G., Tanigawa A., Naganuma T., Li R., Kimura H., Yokoi T., Nakagawa S., Bénard M., Fox A.H., Pierron G. *NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies* // Molecular Biology of the Cell 2014.– Vol.25.– P. 169–183.
109. Kaida D., Berg M.G., Younis I., Kasim M., Singh L.N., Wan L., Dreyfuss G. *U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation* // Nature 2010.– Vol. 468 № 7324.– P. 664–668.
110. Yin Y., Lu J.Y., Zhang X., Shao W., Xu Y., Li P., Hong Y., Cui L., Shan G., Tian B., Zhang Q.C., Shen X. *U1 snRNP regulates chromatin retention of noncoding RNAs* // Nature 2020.– Vol. 580 № 7801.– P. 147–150.
111. Gupta R.A., Shah N., Wang K.C., Kim J., Horlings H.M., Wong D.J., Tsai M.-C., Hung T., Argani P., Rinn J.L., Wang Y., Brzoska P., Kong B., Li R., West R.B., Van De Vijver M.J., Sukumar S., Chang H.Y. *Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis* // Nature 2010.– Vol. 464 № 7291.– P. 1071–1076.

112. Rinn J.L., Kertesz M., Wang J.K., Squazzo S.L., Xu X., Brugmann S.A., Goodnough L.H., Helms J.A., Farnham P.J., Segal E., Chang H.Y. *Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs* // Cell 2007.– Vol. 129 № 7.– P. 1311–1323.
113. Tsai M.-C., Manor O., Wan Y., Mosammaparast N., Wang J.K., Lan F., Shi Y., Segal E., Chang H.Y. *Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes* // Science 2010.– Vol. 329 № 5992.– P. 689–693.
114. Sawaengdee W., Cui K., Zhao K., Hongeng S., Fucharoen S., Wongtrakoongate P. *Genome-Wide Transcriptional Regulation of the Long Non-coding RNA Steroid Receptor RNA Activator in Human Erythroblasts* // Frontiers in Genetics 2020.– Vol.11.– P. 1–15.
115. Ernst J., Kheradpour P., Mikkelsen T.S., Shoresh N., Ward L.D., Epstein C.B., Zhang X., Wang L., Issner R., Coyne M., Ku M., Durham T., Kellis M., Bernstein B.E. *Mapping and analysis of chromatin state dynamics in nine human cell types* // Nature 2011.– Vol. 473 № 7345.– P. 43–49.
116. Fan Z., Chen X., Liu L., Zhu C., Xu J., Yin X., Sheng Y., Zhu Z., Wen L., Zuo X., Zheng X., Zhang Y., Xu J., Huang H., Zhou F., Sun L., Luo J., Zhang D., Chen X., Cui Y., Hao Y., Cui Y., Zhang X., Chen R. *Association of the Polymorphism rs13259960 in SLEAR With Predisposition to Systemic Lupus Erythematosus* // Arthritis Rheumatol. 2020.– Vol. 72 № 6.– P. 985–996.
117. Luo H., Zhu G., Xu J., Lai Q., Yan B., Guo Y., Fung T.K., Zeisig B.B., Cui Y., Zha J., Cogle C., Wang F., Xu B., Yang F.-C., Li W., So C.W.E., Qiu Y., Xu M., Huang S. *HOTTIP lncRNA Promotes Hematopoietic Stem Cell Self-Renewal Leading to AML-like Disease in Mice* // Cancer Cell 2019.– Vol. 36 № 6.– P. 645–659.
118. Laurent G.S., Shtokalo D., Dong B., Tackett M.R., Fan X., Lazorthes S., Nicolas E., Sang N., Triche T.J., McCaffrey T.A., Xiao W., Kapranov P. *VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer* // Genome Biology 2013.– Vol. 14 № 7.– P. 1–20.
119. Quinodoz S.A., Jachowicz J.W., Bhat P., Ollikainen N., Banerjee A.K., Goronzy I.N., Blanco M.R., Chovanec P., Chow A., Markaki Y., Thai J., Plath K., Guttman M. *RNA promotes the formation of spatial compartments in the nucleus* // Cell 2021.– Vol. 184 № 23.– P. 5775–5790.
120. Wang Y., Hu S.-B., Wang M.-R., Yao R.-W., Wu D., Yang L., Chen L.-L. *Genome-wide screening of NEAT1 regulators reveals cross-regulation between paraspeckles and mitochondria* // Nat. Cell Biol. 2018.– Vol. 20 № 10.– P. 1145–1158.
121. Binder S., Hösler N., Riedel D., Zipfel I., Buschmann T., Kämpf C., Reiche K., Burger R., Gramatzki M., Hackermüller J., Stadler P.F., Horn F. *STAT3-induced long noncoding RNAs in multiple myeloma cells display different properties in cancer* // Scientific Reports 2017.– Vol. 7 № 7976.– P. 1–13.
122. Guh C.-Y., Hsieh Y.-H., Chu H.-P. *Functions and properties of nuclear lncRNAs–from systematically mapping the interactomes of lncRNAs* // J. Biomed. Sci. 2020.– Vol. 27 № 44.
123. Cetin N.S., Kuo C.-C., Ribarska T., Li R., Costa I.G., Grummt I. *Isolation and genome-wide characterization of cellular DNA:RNA triplex structures* // Nucleic Acids Research 2019.– Vol. 47 № 5.– P. 2306–2321.
124. Rom A., Melamed L., Gil N., Goldrich M.J., Kadir R., Golan M., Biton I., Perry R.B.-T., Ulitsky I. *Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability* // Nature Communications 2019.– Vol. 10 № 5092.– P. 1–15.

125. Lieberman-Aiden E., Van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., Dekker J. *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome* // Science 2009.– Vol. 326 № 5950.– P. 289–293.
126. Hoffman E.A., Frey B.L., Smith L.M., Auble D.T. *Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes* // Journal of Biological Chemistry 2015.– Vol. 290 № 44.– P. 26404–26411.
127. Pinkney H.R., Wright B.M., Diermeier S.D. *The lncRNA Toolkit: Databases and In Silico Tools for lncRNA Analysis* // Non-Coding RNA 2020.– Vol. 6 № 4.– P. 1–25.
128. Chaudhary U., Banerjee S. *Decoding the Non-coding: Tools and Databases Unveiling the Hidden World of "Junk" RNAs for Innovative Therapeutic Exploration* // ACS Pharmacology & Translational Science 2024.– Vol. 7 № 7.– P. 1901–1915.
129. Balamurali D., Stoll M. *Non-Coding RNA Databases in Cardiovascular Research* // Noncoding RNA 2020.– Vol. 6 № 3.– P. 1–13.
130. Vancura A., Lanzós A., Bosch-Guiteras N., Esteban M.T., Gutierrez A.H., Haefliger S., Johnson R. *Cancer LncRNA Census 2 (CLC2): an enhanced resource reveals clinical features of cancer lncRNAs* // NAR Cancer 2021.– Vol. 3 № 2.– P. 1–15.
131. Lin X., Lu Y., Zhang C., Cui Q., Tang Y.-D., Ji X., Cui C. *LncRNADisease v3.0: an updated database of long non-coding RNA-associated diseases* // Nucleic Acids Research 2023.– Vol. 52 № D1.– P. D1365–D1369.
132. Mazurov E., Sizykh A., Medvedeva Y.A. *HiMoRNA: A Comprehensive Database of Human lncRNAs Involved in Genome-Wide Epigenetic Regulation* // Non-Coding RNA 2022.– Vol. 8 № 1.– P. 1–7.
133. 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.– Vol.2023.– P. 1–10.
134. Zheng Y., Luo H., Teng X., Hao X., Yan X., Tang Y., Zhang W., Wang Y., Zhang P., Li Y., Zhao Y., Chen R., He S. *NPInter v5.0: ncRNA interaction database in a new era* // Nucleic Acids Research 2023.– Vol. 51 № D1.– P. D232–D239.
135. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. *The Human Genome Browser at UCSC* // Genome Res. 2002.– Vol. 12 № 6.– P. 996–1006.
136. Ewels P., Magnusson M., Lundin S., Käller M. *MultiQC: summarize analysis results for multiple tools and samples in a single report* // Bioinformatics 2016.– Vol. 32 № 19.– P. 3047–3048.
137. Bolger A.M., Lohse M., Usadel B. *Trimmomatic: a flexible trimmer for Illumina sequence data* // Bioinformatics 2014.– Vol. 30 № 15.– P. 2114–2120.
138. Kim D., Paggi J.M., Park C., Bennett C., Salzberg S.L. *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype* // Nature Biotechnology 2019.– Vol. 37 № 8.– P. 907–915.
139. Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. *STAR: ultrafast universal RNA-seq aligner* // Bioinformatics 2012.– Vol. 29 № 1.– P. 15–21.
140. Li H., Durbin R. *Fast and accurate short read alignment with Burrows–Wheeler transform* // Bioinformatics 2009.– Vol. 25 № 14.– P. 1754–1760.
141. Langmead B., Salzberg S.L. *Fast gapped-read alignment with Bowtie 2* // Nature Methods

2012.– Vol. 9 № 4.– P. 357–359.
142. Mylarshchikov D.E., Nikolskaya A.I., Bogomaz O.D., Zharikova A.A., Mironov A.A. *BaRDIC: robust peak calling for RNA–DNA interaction data* // NAR Genomics and Bioinformatics 2024.– Vol. 6 № 2.– P. 1–10.
143. Mandlik J.S., Patil A.S., Singh S. *Next-Generation Sequencing (NGS): Platforms and Applications* // Journal of Pharmacy and Bioallied Sciences 2024.– Vol.16.– P. S41–S45.
144. Dijk E.L, Jaszczyszyn Y., Thermes C. *Library preparation methods for next-generation sequencing: tone down the bias* // Experimental Cell Research 2014.– Vol. 322 № 1.– P. 12–20.
145. Rochette N.C., Rivera‐Colón A.G., Walsh J., Sanger T.J., Campbell‐Staton S.C., Catchen J.M. *On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity* // Molecular Ecology Resources 2023.– Vol. 23 № 6.– P. 1299–1318.
146. Dozmorov M.G., Adrianto I., Giles C.B., Glass E., Glenn S.B., Montgomery C., Sivils K.L., Olson L.E., Iwayama T., Freeman W.M., Lessard C.J., Wren J.D. *Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data* // BMC Bioinformatics 2015.– Vol.16.– P. 1–11.
147. Parekh S., Ziegenhain C., Vieth B., Enard W., Hellmann I. *The impact of amplification on differential expression analyses by RNA-seq* // Scientific Reports 2016.– Vol. 6 № 25533.– P. 1–11.
148. Bansal V. *A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments* // BMC Bioinformatics 2017.– Vol.18.– P. 114–175.
149. Deng J., Zhang J., Tian S., DiCarlo J., Xu H., Rulli S.J., Shaffer J.M., Gupta V., Karakoyun T. *UMI-nea: a fast, robust tool for reference-free UMI deduplication and accurate quantification* // Bioinformatics 2025.– Vol. 41 № 9.– P. 1–4.
150. The bbtools toolkit. Available from https://archive.jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide.
151. Tian S., Peng S., Kalmbach M., Gaonkar K.S., Bhagwate A., Ding W., Eckel-Passow J., Yan H., Slager S.L. *Identification of factors associated with duplicate rate in ChIP-seq data* // PLOS ONE 2019.– Vol. 14 № 4.– P. 1–22.
152. The picard toolkit. Available from https://broadinstitute.github.io/picard.
153. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. *The Sequence Alignment/Map format and SAMtools* // Bioinformatics 2009.– Vol. 25 № 16.– P. 2078–2079.
154. Xu H., Luo X., Qian J., Pang X., Song J., Qian G., Chen J., Chen S. *FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads* // PLoS ONE 2012.– Vol. 7 № 12.– P. 1–6.
155. Shen W., Le S., Li Y., Hu F. *SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation* // PLOS ONE 2016.– Vol. 11 № 10.– P. 1–10.
156. Li W., Fu L., Niu B., Wu S., Wooley J. *Ultrafast clustering algorithms for metagenomic sequence analysis* // Briefings in Bioinformatics 2012.– Vol. 13 № 6.– P. 656–668.
157. The fastx-toolkit. Available from http://hannonlab.cshl.edu/fastx_toolkit.
158. Gaia A.S.C., Gomes De Sá P.H.C., De Oliveira M.S., De Oliveira Veras A.A. *NGSReadsTreatment – A Cuckoo Filter-based Tool for Removing Duplicate Reads in NGS Data* // Scientific Reports 2019.– Vol. 9 № 11681.
159. Hu J., Luo S., Tian M., Ye A.Y. *TrieDedup: a fast trie-based deduplication algorithm to handle ambiguous bases in high-throughput sequencing* // BMC Bioinformatics 2024.– Vol.

25 № 154.– P. 1–13.
160. Heiden J.A.V., Yaari G., Uduman M., Stern J.N.H., O'Connor K.C., Hafler D.A., Vigneault F., Kleinstein S.H. *pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires* // Bioinformatics 2014.– Vol. 30 № 13.– P. 1930–1932.
161. González-Domínguez J., Schmidt B. *ParDRe: faster parallel duplicated reads removal tool for sequencing studies* // Bioinformatics 2016.– Vol. 32 № 10.– P. 1562–1564.
162. Expósito R.R., Veiga J., González-Domínguez J., Touriño J. *MarDRe: efficient MapReduce-based removal of duplicate DNA reads in the cloud* // Bioinformatics 2017.– Vol. 33 № 17.– P. 2762–2764.
163. Manconi A., Moscatelli M., Armano G., Gnocchi M., Orro A., Milanesi L. *Removing duplicate reads using graphics processing units* // BMC Bioinformatics 2016.– Vol. 17 № S12.– P. 59–71.
164. Burriesci M.S., Lehnert E.M., Pringle J.R. *Fulcrum: condensing redundant reads from high-throughput sequencing studies* // Bioinformatics 2012.– Vol. 28 № 10.– P. 1324–1327.
165. Liu Y., Zhang X., Zou Q., Zeng X. *Minirmd: accurate and fast duplicate removal tool for short reads via multiple minimizers* // Bioinformatics 2021.– Vol. 37 № 11.– P. 1604–1606.
166. Frankish A., Diekhans M., Jungreis I., Lagarde J., Loveland J.E., Mudge J.M., Sisu C., Wright J.C., Armstrong J., Barnes I., Berry A., Bignell A., Boix C., Sala S.C., Cunningham F., Di Domenico T., Donaldson S., Fiddes I.T., Girón C.G., Gonzalez J.M., Grego T., Hardy M., Hourlier T., Howe K.L., Hunt T., Izuogu O.G., Johnson R., Martin F.J., Martínez L., Mohanan S., Muir P., Navarro F.C.P., Parker A., Pei B., Pozo F., Riera F.C., Ruffier M., Schmitt B.M., Stapleton E., Suner M.-M., Sycheva I., Uszczynska-Ratajczak B., Wolf M.Y., Xu J., Yang Y.T., Yates A., Zerbino D., Zhang Y., Choudhary J.S., Gerstein M., Guigó R., Hubbard T.J.P., Kellis M., Paten B., Tress M.L., Flicek P. *GENCODE 2021* // Nucleic Acids Research 2021.– Vol. 49 № D1.– P. D916–D923.
167. Pertea M., Pertea G.M., Antonescu C.M., Chang T.-C., Mendell J.T., Salzberg S.L. *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads* // Nature Biotechnology 2015.– Vol. 33 № 3.– P. 290–295.
168. Stavrovskaya E.D., Niranjan T., Fertig E.J., Wheelan S.J., Favorov A.V., Mironov A.A. *StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data* // Bioinformatics 2017.– Vol. 33 № 20.– P. 3158–3165.
169. 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.
170. Gavrilov A.A., Sultanov R.I., Magnitov M.D., Galitsyna A.A., Dashinimaev E.B., Aiden E.L., Razin S.V. *RedChIP identifies noncoding RNAs associated with genomic sites occupied by Polycomb and CTCF proteins* // PNAS 2022.– Vol. 119 № 1.– P. 1–3.
171. Alberti A., Belser C., Engelen S., Bertrand L., Orvain C., Brinas L., Cruaud C., Giraut L., Da Silva C., Firmo C., Aury J.-M., Wincker P. *Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data* // BMC Genomics 2014.– Vol. 15 № 1.– P. 1–13.
172. Stoler N., Nekrutenko A. *Sequencing error profiles of Illumina sequencing instruments* // NAR Genomics and Bioinformatics 2021.– Vol. 3 № 1.– P. 1–9.
173. Da Paz M.A., Warger S., Taher L. *Disregarding multimappers leads to biases in the functional assessment of NGS data* // BMC Genomics 2024.– Vol.25.– P. 1–9.
174. Morrissey A., Shi J., James D.Q., Mahony S. *Accurate allocation of multimapped reads*

enables regulatory element analysis at repeats // Genome Research 2024.– Vol. 34 № 6.– P. 937–951.
175. Chen S., Zhou Y., Chen Y., Gu J. *fastp: an ultra-fast all-in-one FASTQ preprocessor* // Bioinformatics 2018.– Vol. 34 № 17.– P. i884–i890.
176. Hartana C.A., Rassadkina Y., Gao C., Martin-Gayo E., Walker B.D., Lichterfeld M., Yu X.G. *Long noncoding RNA MIR4435-2HG enhances metabolic function of myeloid dendritic cells from HIV-1 elite controllers* // Journal of Clinical Investigation 2021.– Vol. 131 № 9.– P. 1–17.
177. Chen W., Wang F., Yu X., Qi J., Dong H., Cui B., Zhang Q., Wu Y., An J., Ni N., Liu C., Han Y., Zhang S., Schmitt C.A., Deng J., Yu Y., Du J. *LncRNA MIR31HG fosters stemness malignant features of non-small cell lung cancer via H3K4me1- and H3K27Ace-mediated GLI2 expression* // Oncogene 2024.– Vol. 43 № 18.– P. 1328–1340.
178. Li B., Li A., You Z., Xu J., Zhu S. *Epigenetic silencing of CDKN1A and CDKN2B by SNHG1 promotes the cell cycle, migration and epithelial-mesenchymal transition progression of hepatocellular carcinoma* // Cell Death & Disease 2020.– Vol. 11 № 10.– P. 1–15.
179. Li Z., Guo X., Wu S. *Epigenetic silencing of KLF2 by long non-coding RNA SNHG1 inhibits periodontal ligament stem cell osteogenesis differentiation* // Stem Cell Research & Therapy 2020.– Vol. 11 № 435.– P. 1–11.
180. Nylund P., Garrido-Zabala B., Atienza Párraga A., Vasquez L., Pyl P.T., Harinck G.M., Ma A., Jin J., Öberg F., Kalushkova A., Wiklund H.J. *PVT1 interacts with polycomb repressive complex 2 to suppress genomic regions with pro-apoptotic and tumour suppressor functions in multiple myeloma* // Haematologica 2023.– Vol. 109 № 2.– P. 567–577.
181. Huang Y., Jin C., Zheng Y., Li X., Zhang S., Zhang Y., Jia L., Li W. *Knockdown of lncRNA MIR31HG inhibits adipocyte differentiation of human adipose-derived stem cells via histone modification of FABP4* // Scientific Reports 2017.– Vol. 7 № 8080.– P. 1–13.
182. Kanehisa M., Sato Y., Kawashima M., Furumichi M., Tanabe M. *KEGG as a reference resource for gene and protein annotation* // Nucleic Acids Research 2016.– Vol.44.– P. D457–D462.
183. Kolberg L., Raudvere U., Kuzmin I., Adler P., Vilo J., Peterson H. *g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)* // Nucleic Acids Research 2023.– Vol. 51 № W1.– P. W207–W212.
184. Wan L., Sun M., Liu G.-J., Wei C.-C., Zhang E.-B., Kong R., Xu T.-P., Huang M.-D., Wang Z.-X. *Long Noncoding RNA PVT1 Promotes Non–Small Cell Lung Cancer Cell Proliferation through Epigenetically Regulating LATS2 Expression* // Molecular Cancer Therapeutics 2016.– Vol. 15 № 5.– P. 1082–1094.
185. Lin S.-L., Miller J.D., Ying S.-Y. *Intronic MicroRNA (miRNA)* // BioMed Research International 2006.– Vol.2006.– P. 1–13.
186. Bergeron D., Faucher-Giguère L., Emmerichs A.-K., Choquet K., Song K.S., Deschamps-Francoeur G., Fafard-Couture É., Rivera A., Couture S., Churchman L.S., Heyd F., Elela S.A., Scott M.S. *Intronic small nucleolar RNAs regulate host gene splicing through base pairing with their adjacent intronic sequences* // Genome Biology 2023.– Vol.24.– P. 1–25.
187. Nam J.-W., Choi S.-W., You B.-H. *Incredible RNA: Dual Functions of Coding and Noncoding* // Molecules and Cells 2016.– Vol. 39 № 5.– P. 367–374.
188. Machyna M., Simon M.D. *Catching RNAs on chromatin using hybridization capture methods* // Briefings in Functional Genomics 2018.– Vol. 17 № 2.– P. 96–103.

189. Oh H.J., Aguilar R., Kesner B., Lee H.-G., Kriz A.J., Chu H.-P., Lee J.T. *Jpx RNA regulates CTCF anchor site selection and formation of chromosome loops* // Cell 2021.– Vol. 184 № 25.– P. 6157–6173.
190. Tripathi V., Ellis J.D., Shen Z., Song D.Y., Pan Q., Watt A.T., Freier S.M., Bennett C.F., Sharma A., Bubulya P.A., Blencowe B.J., Prasanth S.G., Prasanth K.V. *The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation* // Molecular Cell 2010.– Vol. 39 № 6.– P. 925–938.
191. Cai D., Han J.-D.J. *Aging-associated lncRNAs are evolutionarily conserved and participate in NFκB signaling* // Nat Aging. 2021.– Vol. 1 № 5.– P. 438–453.
192. Stitzinger S.H., Sohrabi-Jahromi S., Söding J. *Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains* // NAR Genomics and Bioinformatics 2023.– Vol. 5 № 2.– P. 1–10.
193. Khlebnikov D.A., Nikolskaya A.I., Zharikova A.A., Mironov A.A. *Comprehensive analysis of RNA–chromatin, RNA–, and DNA–protein interactions* // NAR Genomics and Bioinformatics 2025.– Vol. 7 № 1.– P. 1–15.
194. Liu W., Ma Q., Wong K., Li W., Ohgi K., Zhang J., Aggarwal A.K., Rosenfeld M.G. *Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release* // Cell 2013.– Vol. 155 № 7.– P. 1581–1595.
195. Studniarek C., Tellier M., Martin P.G.P., Murphy S., Kiss T., Egloff S. *The 7SK/P-TEFb snRNP controls ultraviolet radiation-induced transcriptional reprogramming* // Cell Rep. 2021.– Vol. 35 № 2.
196. Chakraborty D., Paszkowski-Rogacz M., Berger N., Ding L., Mircetic J., Fu J., Iesmantavicius V., Choudhary C., Anastassiadis K., Stewart A.F., Buchholz F. *lncRNA Panct1 Maintains Mouse Embryonic Stem Cell Identity by Regulating TOBF1 Recruitment to Oct-Sox Sequences in Early G1* // Cell Rep. 2017.– Vol. 21 № 11.– P. 3012–3021.
197. Dong A., Preusch C.B., So W.-K., Lin K., Luan S., Yi R., Wong J.W., Wu Z., Cheung T.H. *A long noncoding RNA, LncMyoD, modulates chromatin accessibility to regulate muscle stem cell myogenic lineage progression* // PNAS 2020.– Vol. 117 № 51.– P. 32464–32475.
198. Zhu G., Luo H., Feng Y., Guryanova O.A., Xu J., Chen S., Lai Q., Sharma A., Xu B., Zhao Z., Feng R., Ni H., Claxton D., Guo Y., Mesa R.A., Qiu Y., Yang F.-C., Li W., Nimer S.D., Huang S., Xu M. *HOXBLINC long non-coding RNA activation promotes leukemogenesis in NPM1-mutant acute myeloid leukemia* // Nature Communications 2021.– Vol. 12 № 1956.– P. 1–17.
199. Powell W.T., Coulson R.L., Crary F.K., Wong S.S., Ach R.A., Tsang P., Yamada N.A., Yasui D.H., LaSalle J.M. *A Prader–Willi locus lncRNA cloud modulates diurnal genes and energy expenditure* // Human Molecular Genetics 2013.– Vol. 22 № 21.– P. 4318–4328.
200. Lu J.Y., Shao W., Chang L., Yin Y., Li T., Zhang H., Hong Y., Percharde M., Guo L., Wu Z., Liu L., Liu W., Yan P., Ramalho-Santos M., Sun Y., Shen X. *Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation* // Cell Reports 2020.– Vol. 30 № 10.– P. 3296–3311.
201. Liu J., Gao M., He J., Wu K., Lin S., Jin L., Chen Y., Liu H., Shi J., Wang X., Chang L., Lin Y., Zhao Y.-L., Zhang X., Zhang M., Luo G.-Z., Wu G., Pei D., Wang J., Bao X., Chen J. *The RNA m6A reader YTHDC1 silences retrotransposons and guards ES cell identity* // Nature 2021.– Vol.591.– P. 322–326.
202. Xu W., Li J., He C., Wen J., Ma H., Rong B., Diao J., Wang L., Wang J., Wu F., Tan L., Shi Y.G., Shi Y., Shen H. *METTL3 regulates heterochromatin in mouse embryonic stem cells* //

Nature 2021.– Vol.591.– P. 317–321.
203. Dhaka B., Zimmerli M., Hanhart D., Moser M.B., Guillen-Ramirez H., Mishra S., Esposito R., Polidori T., Widmer M., García-Pérez R., Kruithof-de Julio M., Pervouchine D., Melé M., Chouvardas P., Johnson R. *Functional identification of cis-regulatory long noncoding RNAs at controlled false discovery rates* // Nucleic Acids Research 2024.– Vol.52.– P. 2821–2835.

ПРИЛОЖЕНИЕ А

Данные для определения ориентации РНК-частей контактов и аннотации генов в базе данных RNA-Chrom

Таблица А.1. Данные, которые участвуют в определении ориентации РНК-частей контактов. *Homo sapiens*.

Датасет	Количество чтений, участвующих в анализе	Медианное покрытие генов, кодирующих рибосомальные белки
imargi_SRR206679	23682	317
imargi_SRR206680	73941	953
margi_SRR5278094	6671	74
margi_SRR5278095	1246	15
margi_SRR5278096	1580	17
margi_SRR5278097	3166	26
margi_SRR5278098	3691	37
margi_SRR5278099	3100	38
margi_SRR5278100	62782	632
margi_SRR5278101	3656	45
margi_SRR5278102	1531	17
margi_SRR5278104	3127	39
margi_SRR5278105	721	8
grid_SRR3633284	71923	908
grid_SRR3633286	71580	891
grid_SRR3633288	64394	687.0
grid_SRR3633290	68753	765.0
redc_SRR10010323	2484	28.0
redc_SRR10010324	205427	1797.0
redc_SRR10010325	87569	1053.0
redc_SRR10010326	21091	180.0
redc_SRR10010328	314345	3454.0

redc_SRR10010330	2857	36.0
imargi_SRR12462453	13820	222.0
imargi_SRR12462454	10470	136.0
imargi_SRR9900120	44049	705.0
imargi_SRR9900121	23886	345.0
imargi_SRR9900122	27872	372.0

Таблица А.2. Аннотации генов человека (только канонические хромосомы).

Аннотации	**Источник**	**Количество генов**	**Описание**
gencode	GENCODE v35 [166]	60619	Оставили только те записи, у которых в третьем столбце было указано «gene».
vlinc	статья [118]	2762	Координаты генов, полученные на геноме hg19, были перенесены на hg38 геном с помощью инструмента liftOver (UCSC Genome Browser, стандартные параметры).
GB_snomirna	UCSC Genome Browser (таблица wgRna)	2320	NaN
GB_trna	UCSC Genome Browser (таблица tRNAs)	629	NaN
GB_repM	UCSC Genome Browser (таблица rmsk)	11408	Фильтрация по столбцу «repClass», выбранные значения: snRNA, rRNA, scRNA, tRNA, RNA, srpRNA.
from_article	статьи [42,63,89]	3	«lnc-NR2F1_short», «Linc-ASEN» и «DACOR1» не были обнаружены ни в одной из приведенных выше аннотаций генов, но для этих РНК имеются данные ОТА.
Xrna_human	собраны программой StringTie [167]	155127	Использованы данные из статей: GRID-seq [11], Red-C [16] и iMARGI [13,14] (см. «МАТЕРИАЛЫ И МЕТОДЫ», раздел «Универсальный протокол обработки данных в базе данных RNA-Chrom», раздел «Сборка Х-РНК»).
RNA-Chrom DB	статьи [76,79,194,195]	2	В некоторых экспериментах изучались многокопийные РНК, истинные родительские гены которых мы не можем определить. В таких случаях мы ввели групповое название для соответствующих многокопийных РНК: «ERV-9 RNAs» и «7SK RNAs».

Таблица А.3. Аннотации генов мыши (только канонические хромосомы).

Аннотации	**Источник**	**Количество генов**	**Описание**
gencode	GENCODE M25 [166]	55364	Оставили только те записи, у которых в третьем столбце было указано «gene».
GB_trna	UCSC Genome Browser (таблица tRNAs)	434	NaN
GB_repM	UCSC Genome Browser (таблица rmsk)	18770	Фильтрация по столбцу «repClass», выбранные значения: snRNA, rRNA, scRNA, tRNA, RNA, srpRNA.
from_article	статьи [43,196–198]	4	«Panct1», «LncHSC-2», «LncMyoD», «HOXBLINC» не были обнаружены ни в одной из приведенных выше аннотаций генов, но для этих РНК имеются данные ОТА.
Xrna_mouse	собраны программой StringTie [167]	14333	Использованы данные из статей: GRID-seq [11] и RADICL-seq [15] (см. «МАТЕРИАЛЫ И МЕТОДЫ», раздел «Универсальный протокол обработки данных в базе данных RNA-Chrom», раздел «Сборка Х-РНК»).
RNA-Chrom DB	статьи [9,26,51,79,81,83,88,199–202]	9	В некоторых экспериментах изучались многокопийные РНК, истинные родительские гены которых мы не можем определить. В таких случаях мы

			ввели групповое название для соответствующих многокопийных РНК: «U1 RNAs", «TERRA RNAs», «PAR-TERRA RNAs», «LINE1 RNAs», «IAP RNAs», «B2 RNAs», «7SK RNAs», «116HG RNAs», «IAPEz-int RNAs».

ПРИЛОЖЕНИЕ Б

Архитектура и сравнительный анализ производительности программы Fastq-dupaway

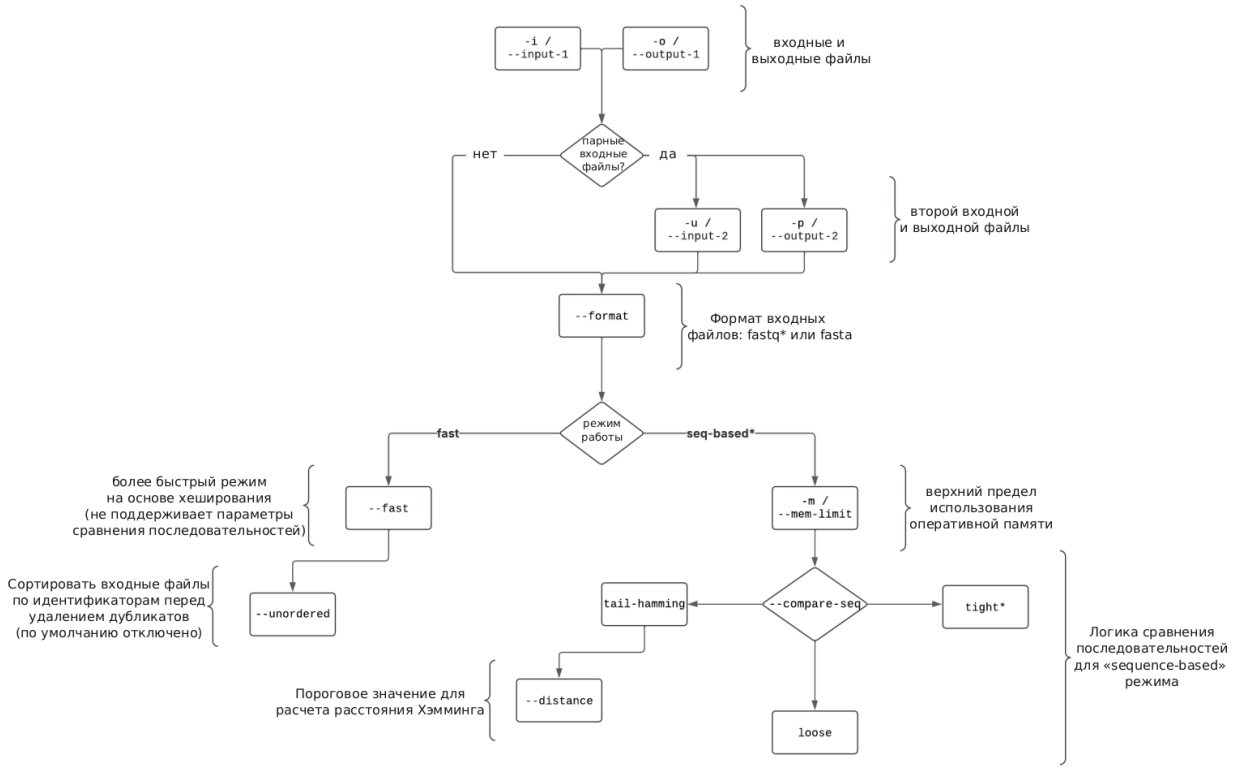


Рисунок Б.1. Блок-схема со всеми доступными опциями программы Fastq-dupaway. Опции, включенные по умолчанию, отмечены знаком «*».

Таблица Б.1. Показатели производительности BBTools Clumpify в многопоточном режиме (4 потока) по сравнению с однопоточным выполнением. ПЦР-дубликаты были идентифицированы с нулевым количеством несовпадений. Коэффициенты производительности рассчитывались как медиана однопоточного режима, деленная на медиану многопоточного режима.

Тип протокола (ГБ)	**Отношение затраченного времени**	**Отношение времени CPU**
CHART-seq (10,8)	2,4	0,7
ChIP-seq (CTCF) (16,5)	2,1	0,7
ChIP-seq (H3K27me3) (12,3)	2,4	0,8
ChIP-seq (H3K4me1) (5,34)	1,8	0,7
ChIRP-seq (53,0)	3	1
Exome-seq (62,4)	1,6	0,6
Exome-seq (91,4)	1,4	0,6
GRID-seq (42,0)	3	0,7
Hi-C (47,6)	2,5	0,8
Hi-C (106,6)	3	1
RADICL-seq (36,0)	2,8	0,8
Whole Genome Sequencing (12,5)	2,5	0,9
Whole Genome Sequencing (104,0)	2,5	0,9
Whole Genome Sequencing (112,4)	2,6	0,8
Hi-C (538,0)	None	None

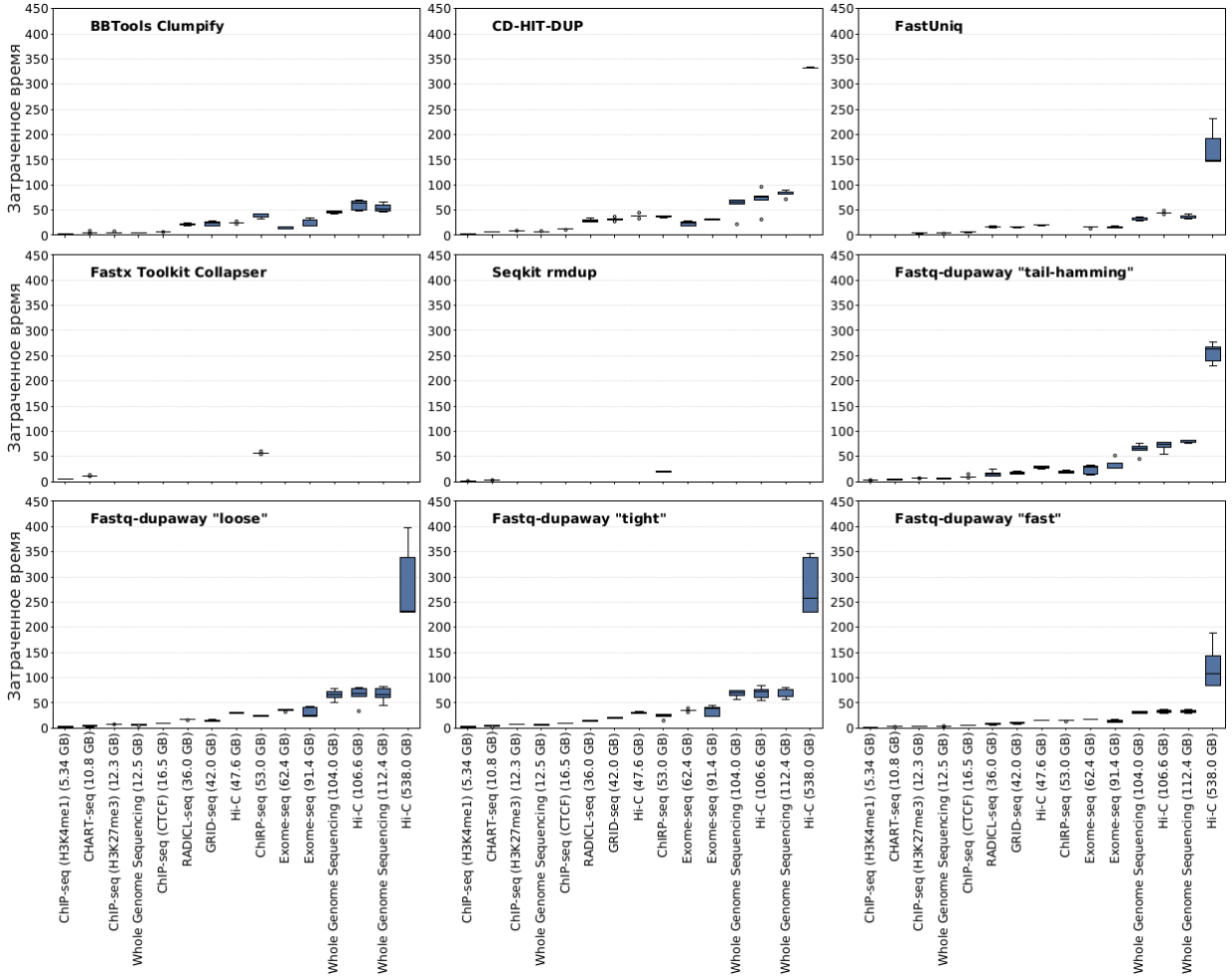


Рисунок Б.2. Распределение затраченного времени для инструментов, запущенных 5 раз на каждом наборе данных. ПЦР-дубликаты были выявлены с нулевым количеством несовпадений. Для BBTools Clumpify отсутствует «ящик с усами» для данных «Hi-C (538 ГБ)», поскольку программа завершилась с ошибкой. Для программ FastUniq, Fastx Toolkit Collapser и Seqkit rmdup отсутствуют «ящики с усами» для наборов данных, не поддерживаемых этими инструментами.

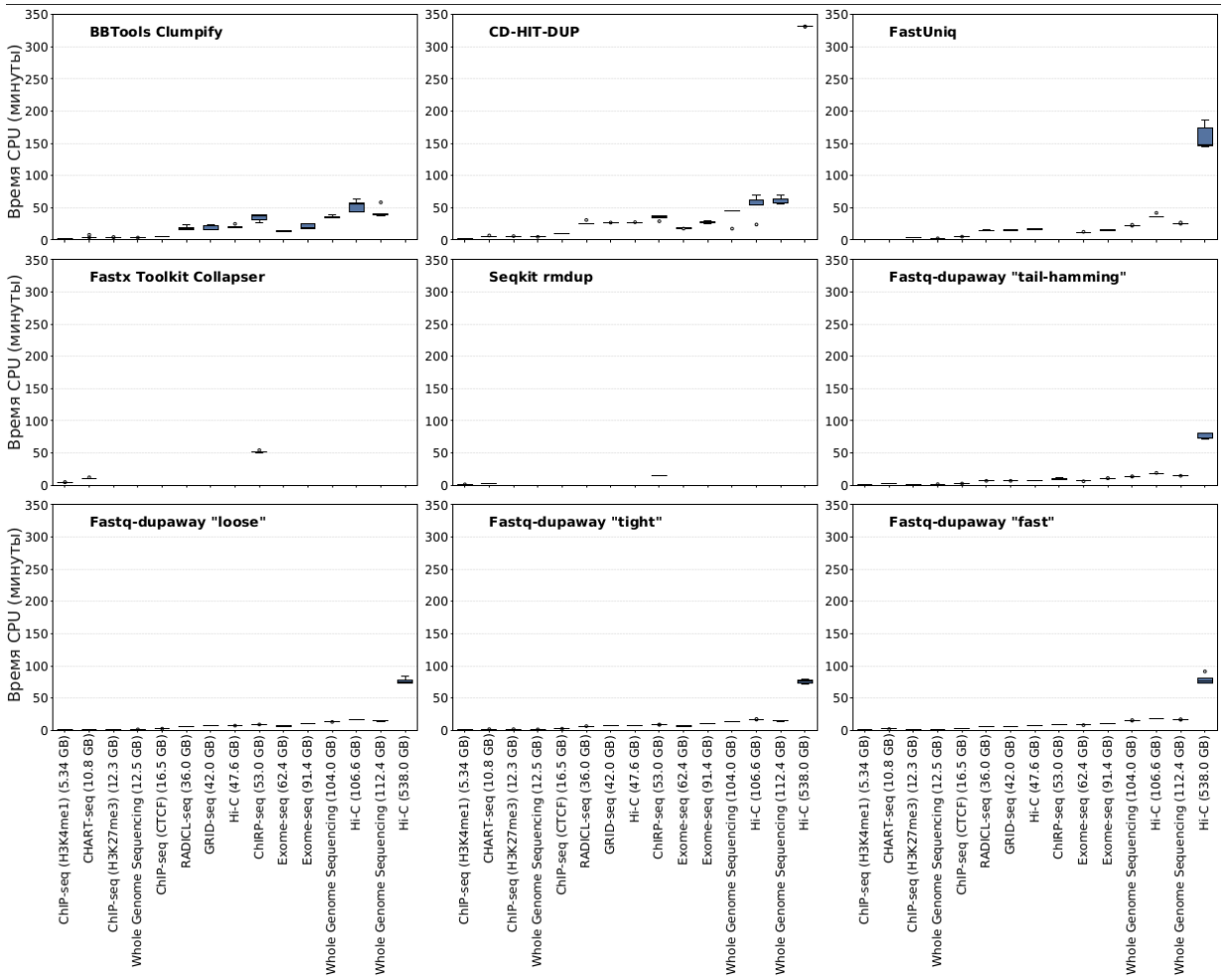


Рисунок Б.3. Распределение времени CPU для инструментов, запущенных 5 раз на каждом наборе данных. ПЦР-дубликаты были выявлены с нулевым количеством несовпадений. Для BBTools Clumpify отсутствует «ящик с усами» для данных «Hi-C (538 ГБ)», поскольку программа завершилась с ошибкой. Для программ FastUniq, Fastx Toolkit Collapser и Seqkit rmdup отсутствуют «ящики с усами» для наборов данных, не поддерживаемых этими инструментами.

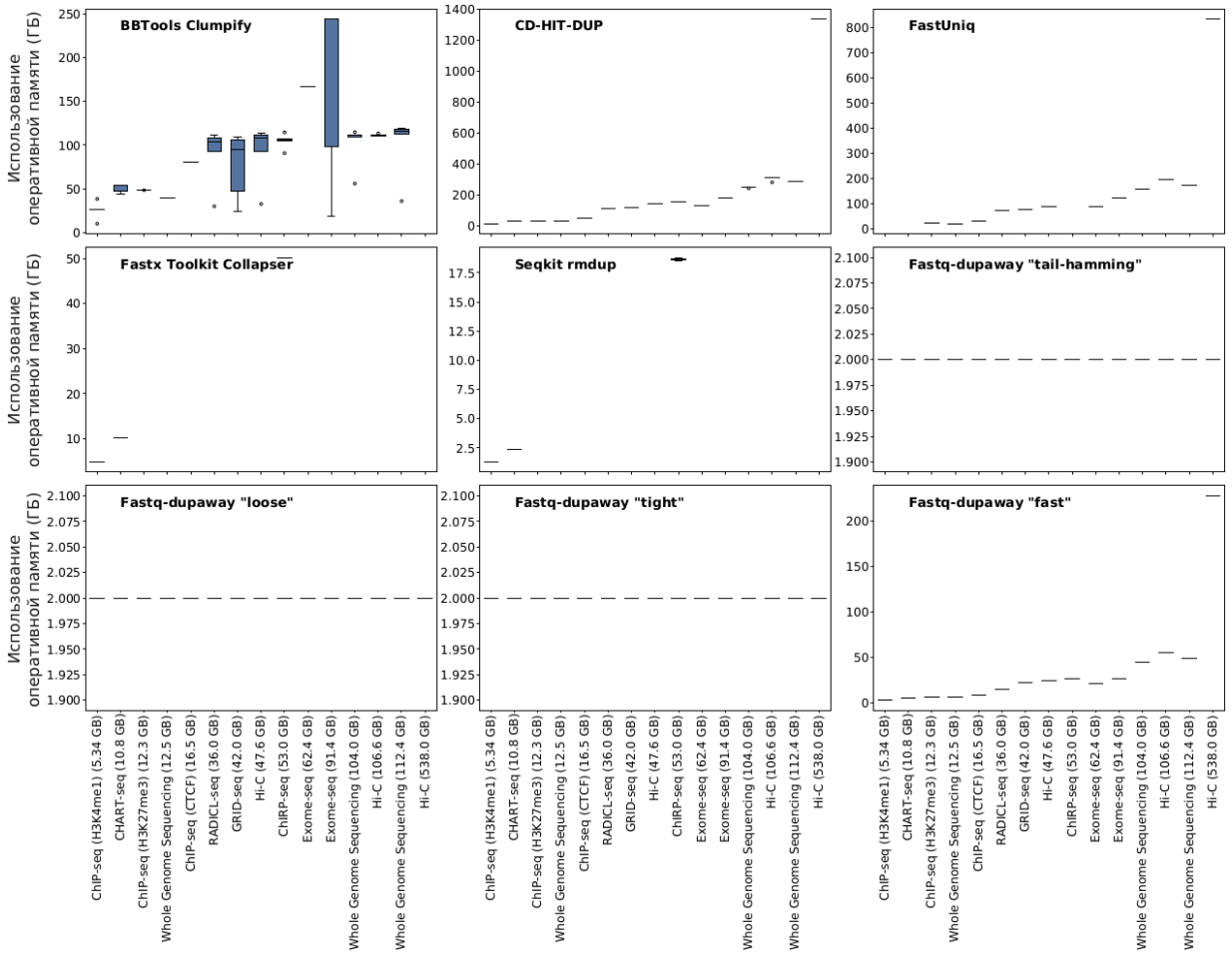


Рисунок Б.4. Распределение использованной оперативной памяти для инструментов, запущенных 5 раз на каждом наборе данных. ПЦР-дубликаты были выявлены с нулевым количеством несовпадений. Для BBTools Clumpify отсутствует «ящик с усами» для данных «Hi-C (538 ГБ)», поскольку программа завершилась с ошибкой. Для программ FastUniq, Fastx Toolkit Collapser и Seqkit rmdup отсутствуют «ящики с усами» для наборов данных, не поддерживаемых этими инструментами.

Таблица Б.2. Время CPU (в минутах) для различных методов. Были идентифицированы ПЦР-дубликаты с допустимым количеством несовпадений до двух. Каждое значение соответствует медиане пяти запусков соответствующего инструмента.

Датасет	BBTools Clumpify	CD-HIT-DUP	Fastq-dupaway "tail-hamming"	Тип протокола	Размер датасета (ГБ)
SRR10044362	8.2	7.5	2.2	CHART-seq	10.8
SRR10950502	5.7	15.3	2.4	ChIP-seq (CTCF)	16.48
SRR8902551	4.5	10.7	1.6	ChIP-seq (H3K27me3)	12.34
SRR504934	2.6	3.0	1.1	ChIP-seq (H3K4me1)	5.34
SRR1425229	32.4	39.1	8.5	ChIRP-seq	53.0
SRR24907572	13.8	31.2	6.9	Exome-seq	62.4
SRR13232316	19.0	42.5	10.6	Exome-seq	91.4
SRR3633290, SRR3633291	23.1	29.9	6.8	GRID-seq	42.0
SRR8902547	25.8	47.2	7.2	Hi-C	47.6
SRR9675763	52.6	109.0	16.8	Hi-C	106.6
SRR1658643	Error	Error	80.3	Hi-C	538.0
SRR9201799, SRR9201800	19.5	40.8	6.5	RADICL-seq	36.0
SRR2014554	4.0	8.4	1.5	Whole Genome Sequencing	12.5
SRR19505554	42.2	80.2	13.3	Whole Genome Sequencing	104.0
SRR19505555	51.3	94.6	14.4	Whole Genome Sequencing	112.4

Таблица Б.3. Затраченное время (в минутах) для различных методов. Были идентифицированы ПЦР-дубликаты с допустимым количеством несовпадений до двух. Каждое значение соответствует медиане пяти запусков соответствующего инструмента.

Датасет	BBTools Clumpify	CD-HIT-DUP	Fastq-dupaway "tail-hamming"	Тип протокола	Размер датасета (ГБ)
SRR10044362	8.7	8.5	3.9	CHART-seq	10.8
SRR10950502	7.0	19.1	8.8	ChIP-seq (CTCF)	16.48
SRR8902551	5.1	11.9	5.3	ChIP-seq (H3K27me3)	12.34
SRR504934	3.2	3.6	2.3	ChIP-seq (H3K4me1)	5.34
SRR1425229	34.8	39.5	12.5	ChIRP-seq	53.0
SRR24907572	14.5	32.7	26.9	Exome-seq	62.4
SRR13232316	19.1	42.7	25.9	Exome-seq	91.4
SRR3633290, SRR3633291	25.6	38.5	19.2	GRID-seq	42.0
SRR8902547	29.2	61.3	28.4	Hi-C	47.6
SRR9675763	59.6	130.3	45.2	Hi-C	106.6
SRR1658643	Error	Error	240.2	Hi-C	538.0
SRR9201799, SRR9201800	22.5	43.6	13.4	RADICL-seq	36.0
SRR2014554	4.5	9.9	5.1	Whole Genome Sequencing	12.5
SRR19505554	50.5	98.2	59.0	Whole Genome Sequencing	104.0
SRR19505555	58.5	113.2	45.2	Whole Genome Sequencing	112.4

Таблица Б.4. Использованный объем оперативной памяти (ГБ) для различных методов. Были идентифицированы ПЦР-дубликаты с допустимым количеством несовпадений до двух. Каждое значение соответствует медиане пяти запусков соответствующего инструмента.

Датасет	BBTools Clumpify	CD-HIT-DUP	Fastq-dupaway "tail-hamming"	Тип протокола	Размер датасета (ГБ)
SRR10044362	54.4	40.3	2.0	CHART-seq	10.8
SRR10950502	80.3	71.8	2.0	ChIP-seq (CTCF)	16.48
SRR8902551	49.0	52.1	2.0	ChIP-seq (H3K27me3)	12.34
SRR504934	26.9	16.2	2.0	ChIP-seq (H3K4me1)	5.34
SRR1425229	105.9	187.9	2.0	ChIRP-seq	53.0
SRR24907572	166.7	184.1	2.0	Exome-seq	62.4
SRR13232316	243.6	243.1	2.0	Exome-seq	91.4
SRR3633290, SRR3633291	111.4	132.6	2.0	GRID-seq	42.0
SRR8902547	116.1	206.5	2.0	Hi-C	47.6
SRR9675763	112.4	479.9	2.0	Hi-C	106.6
SRR1658643	Error	Error	2.0	Hi-C	538.0
SRR9201799, SRR9201800	101.5	127.3	2.0	RADICL-seq	36.0
SRR2014554	39.3	41.2	2.0	Whole Genome Sequencing	12.5
SRR19505554	117.4	373.9	2.0	Whole Genome Sequencing	104.0
SRR19505555	118.6	432.0	2.0	Whole Genome Sequencing	112.4

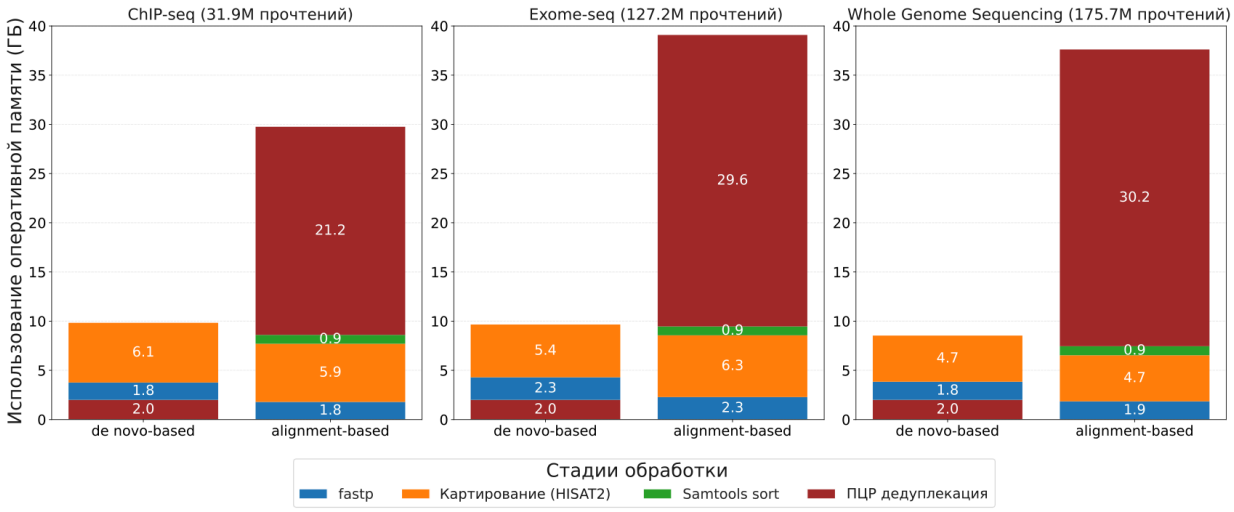


Рисунок Б.5. Использование оперативной памяти (ГБ) конвейерами обработки данных, использующими *de novo*-based и alignment-based подходы к дедупликации. ПЦР-дубликаты были идентифицированы с нулевым количеством несовпадений. Каждое значение соответствует медиане пяти запусков соответствующего инструмента. Порядок этапов обработки снизу вверх отражает последовательный порядок выполнения программ в соответствующих конвейерах.

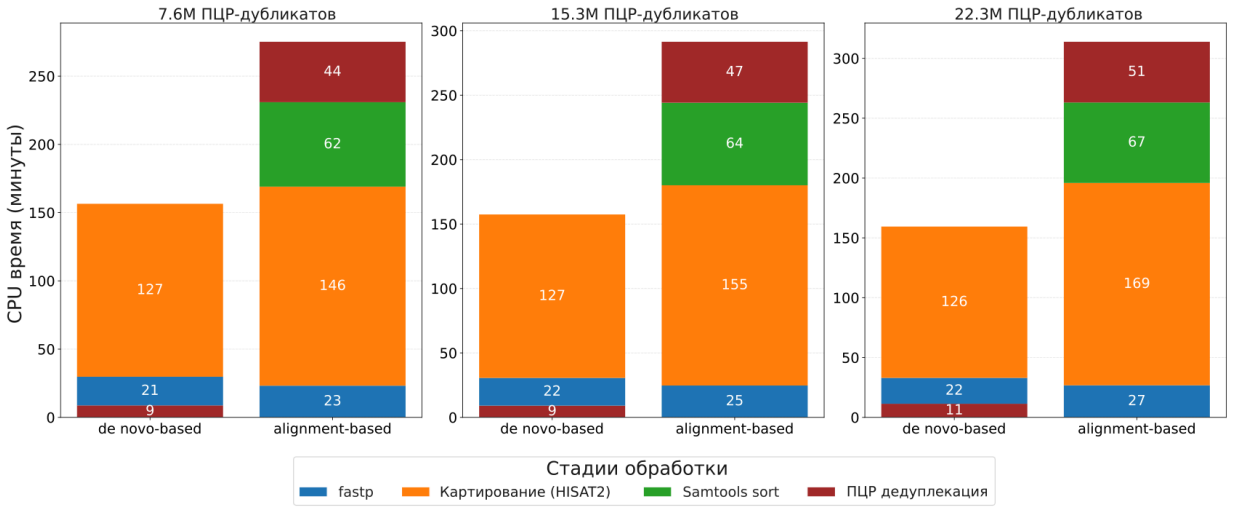


Рисунок Б.6. CPU время конвейеров, использующих de novo-based и alignment-based подходы к дедупликации. ПЦР-дубликаты были идентифицированы с нулевым количеством несовпадений. Данные представлены для трех наборов данных с различным количеством ПЦР-дубликатов, полученных из SRR13232316: (слева) ~6,8% дубликатов, (в центре) ~12,7% дубликатов, (справа) исходная библиотека SRR13232316 с ~17,5% дубликатов. Во всех этих наборах данных 104,9 миллиона прочтений являются общими и не идентифицируются как дубликаты ПЦР в соответствии с режимом «tight» Fastq-dupaway.

ПРИЛОЖЕНИЕ В

Материалы, дополняющие главу по сравнительному анализу РНК-хроматинового интерактома

Таблица В.1. Данные ATA. ActD – обработка актиномицином D, NPM – обработка протеиназой К, 1% FA – обработка сшивающим агентом формальдегидом в концентрации 1%, 2% FA – обработка сшивающим агентом формальдегидом в концентрации 2%.

Данные ATA (экспериментальный метод, клеточная линия, организм)	**Номера реплик**	**Количество контактов**
GRID, MM.1S, *Homo sapiens*	GSM2188868, GSM2188869	37 817 797
Red-C, K562, *H. sapiens*	GSM4041591, GSM4041595	41 777 653
RADICL, OPC, *Mus musculus*	GSM3852782-GSM3852783, GSM3852784-GSM3852785	27 898 478
RADICL, ES (ActD), *M. musculus*	GSM3852772-GSM3852773, GSM3852774-GSM3852775	13 586 144
RADICL, OPC (NPM), *M. musculus*	GSM3852788-GSM3852789, GSM3852790-GSM3852791	42 648 037
RADICL, ES (1% FA), *M. musculus*	GSM3852760-GSM3852761, GSM3852762-GSM3852763	28 609 319
RADICL, ES (2% FA), *M. musculus*	GSM3852766-GSM3852767, GSM3852768-GSM3852769	28 849 211
RADICL, ES (NPM), *M. musculus*	GSM3852776-GSM3852777, GSM3852778-GSM3852779	10 444 053
GRID, ES, *M. musculus*	GSM2396700, GSM2396701	59 958 702
GRID, MDA_MB_231, *H. sapiens*	GSM2188866, GSM2188867	63 227 196

Таблица В.2. Данные RNA-seq с деплецией рибосомальной РНК.

Клеточная линия, организм	**Номера реплик**
MDA-MB-231, *H. sapiens*	GSM7143069-GSM7143071
MM.1S, *H. sapiens*	GSM5788444 (SRR17510863, SRR17510864), GSM5788446 (SRR17510859, SRR17510860)
H1 ES, *H. sapiens*	GSM3630264, GSM3630265
K562, *H. sapiens*	GSM4744788, GSM4744789, GSM4744790
OPC, *M. musculus*	GSM3034716, GSM3034717, GSM3034718
ES, *M. musculus*	GSM4775002, GSM4775004

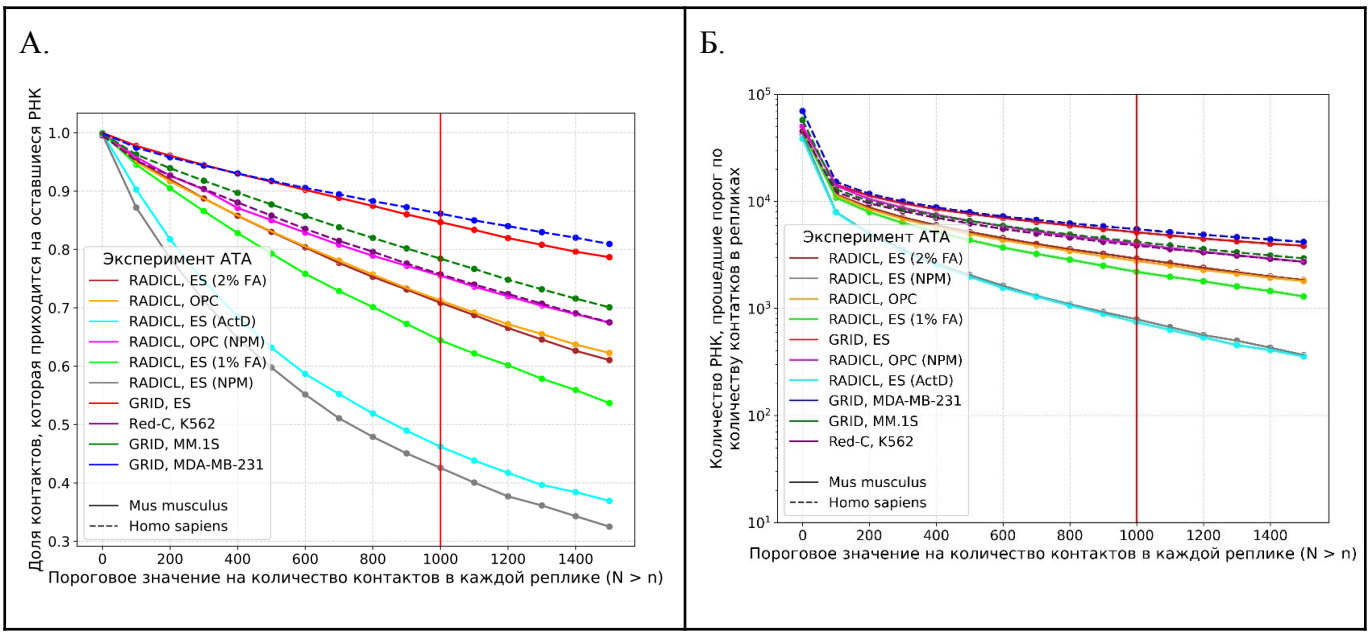


Рисунок В.1. Доля контактов (А) и количество РНК (Б) в зависимости от порога на количество контактов в каждой реплике для разных экспериментов. ActD – обработка актиномицином D; NPM – обработка протеиназой К; 1% FA – обработка сшивающим агентом формальдегидом в концентрации 1%; 2% FA – обработка сшивающим агентом формальдегидом в концентрации 2%.

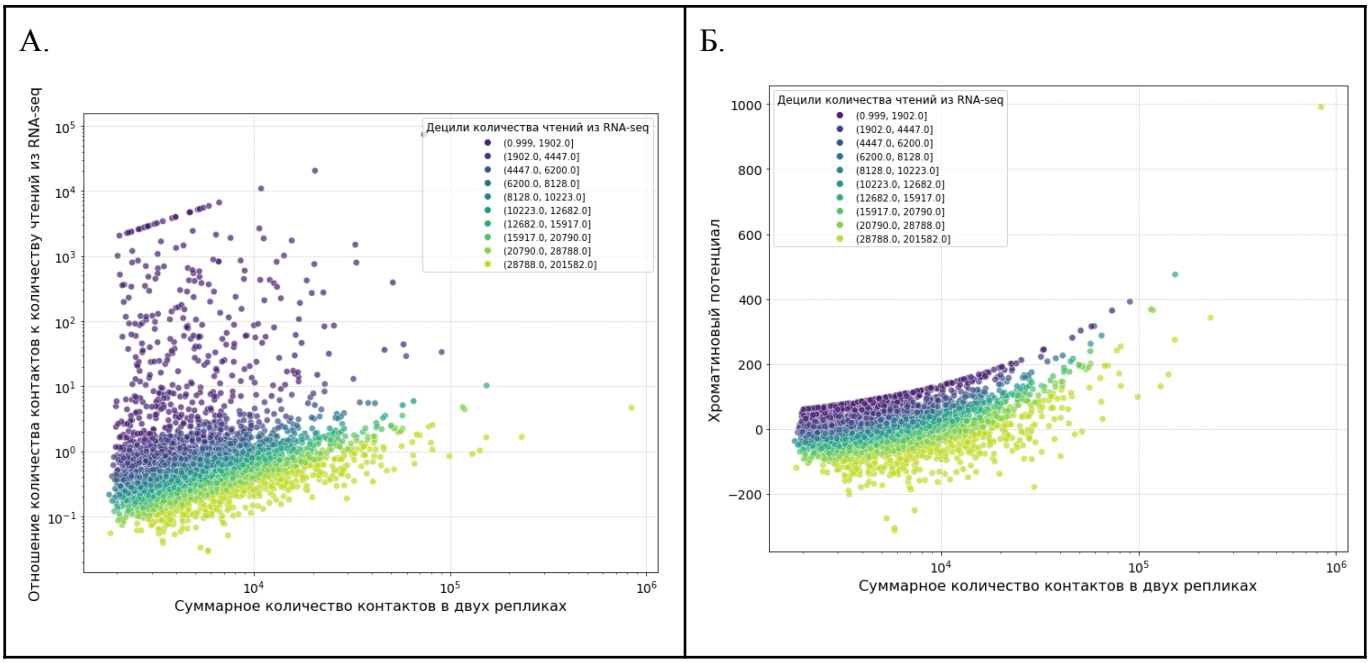
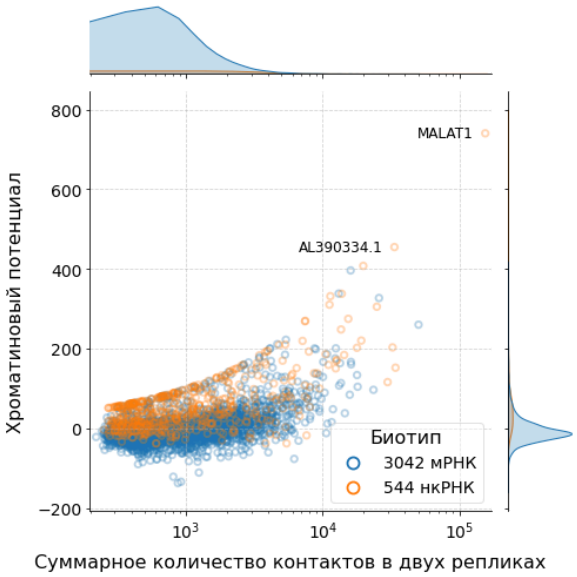
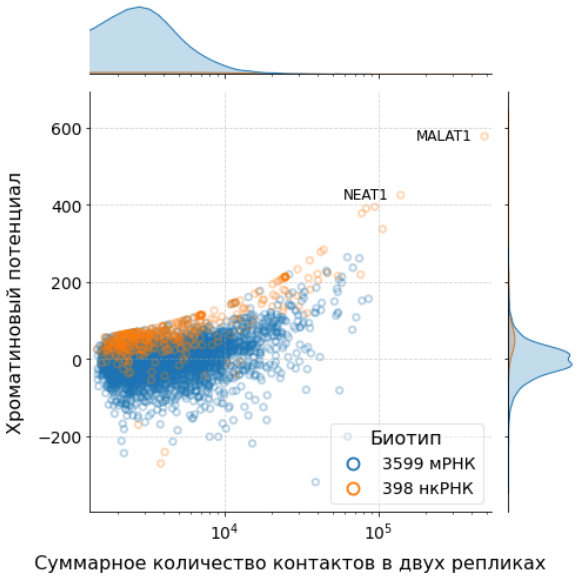


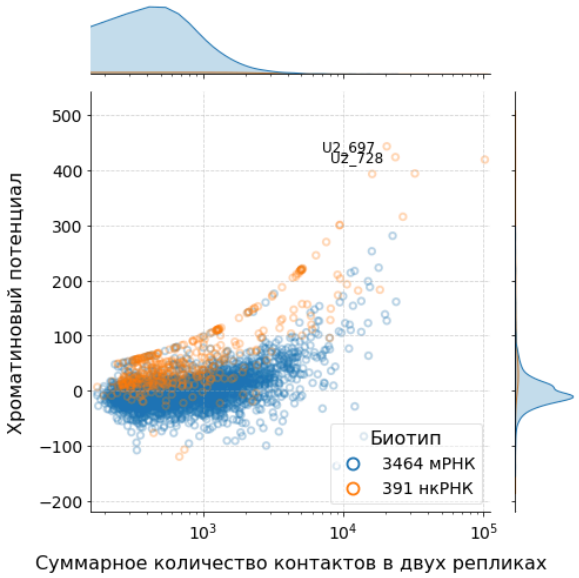
Рисунок В.2. Распределение (А) отношения числа контактов к уровню экспрессии (Б) хроматинового потенциала в зависимости от количества контактов. Представлены данные Red-C на клетках K562.



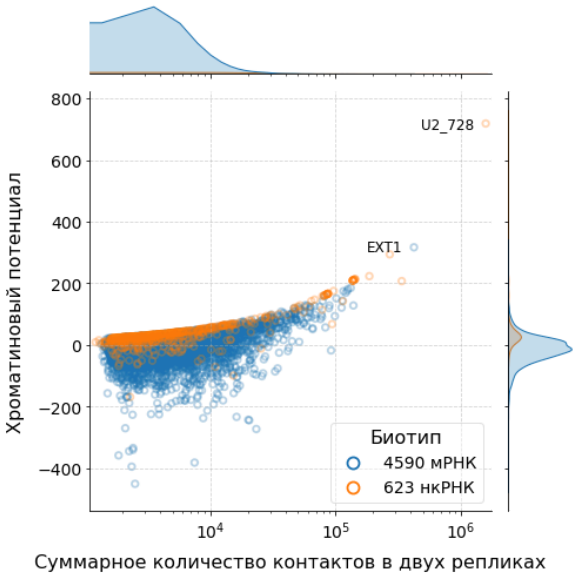
А. Red-C, K562, H. sapiens, контакты в пиках



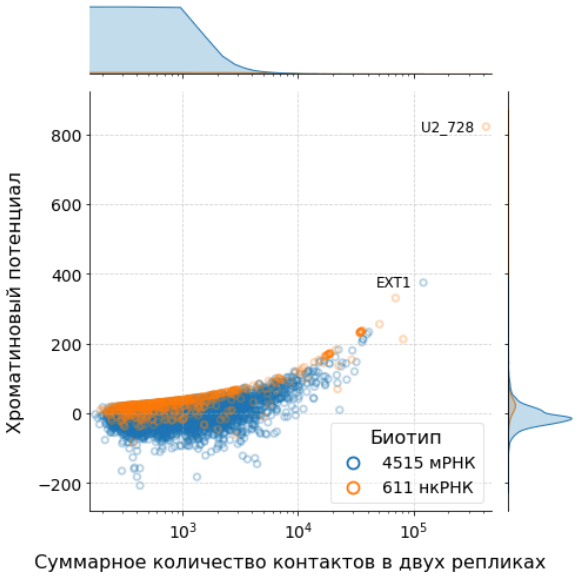
Б. GRID, MM.1S, H. sapiens, все контакты



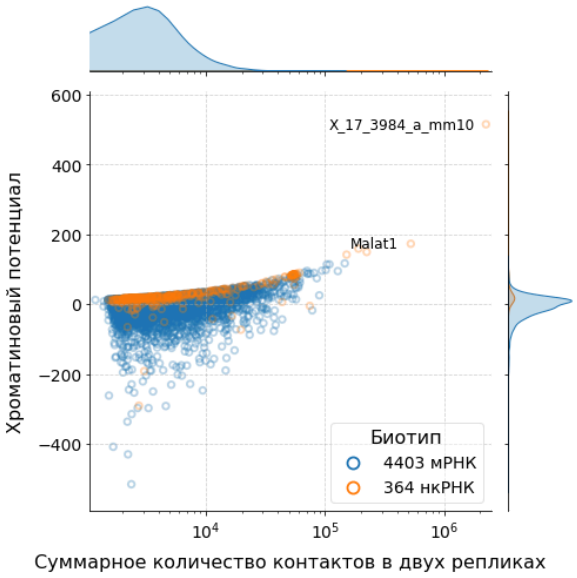
В. GRID, MM.1S, H. sapiens, контакты в пиках



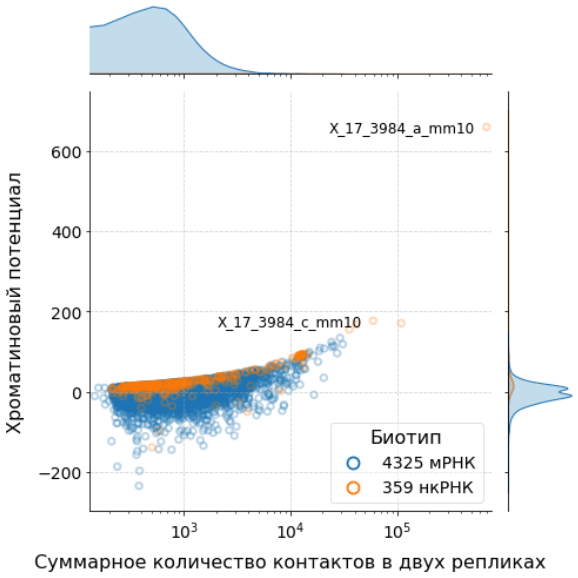
Г. GRID, MDA_MB_231, H. sapiens, все контакты



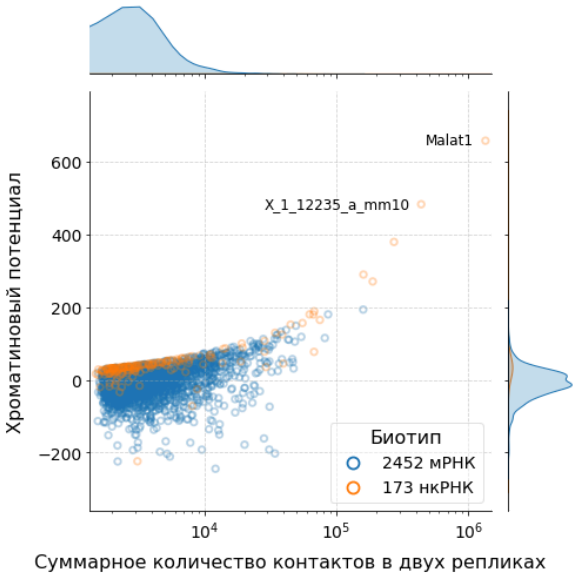
Д. GRID, MDA_MB_231, H. sapiens, контакты в пиках



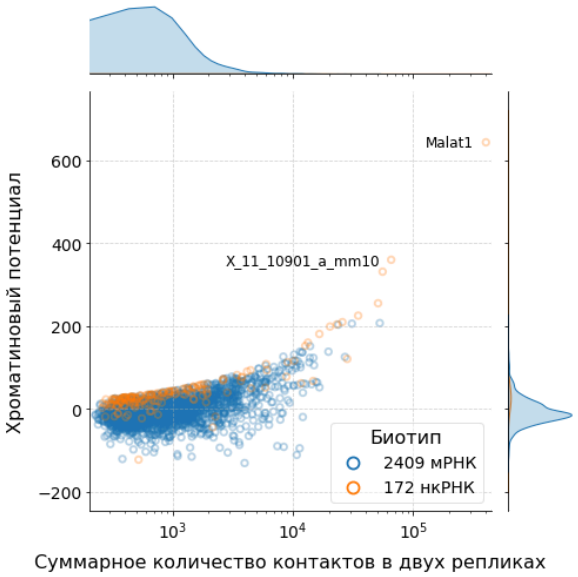
Е. GRID, ES, M. musculus, все контакты



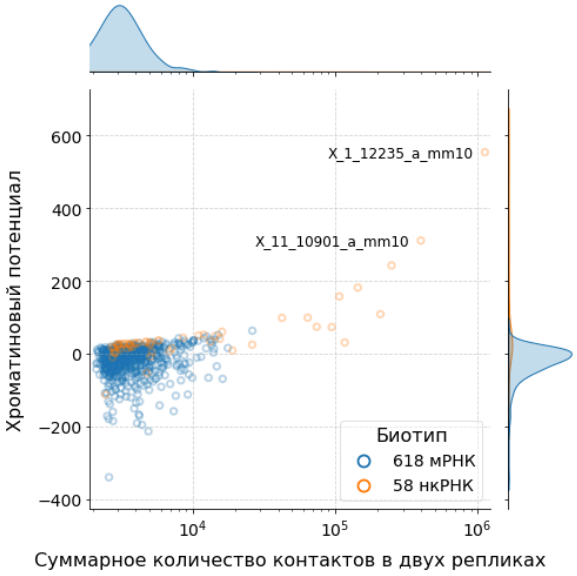
Ж. GRID, ES, M. musculus, контакты в пиках



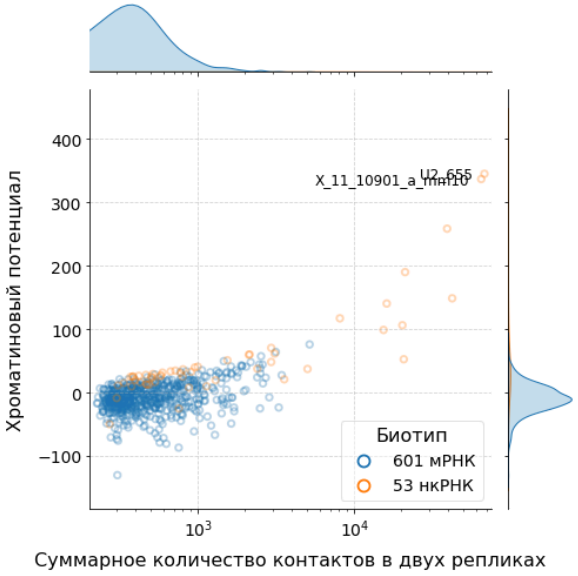
З. RADICL, OPC, M. musculus, все контакты



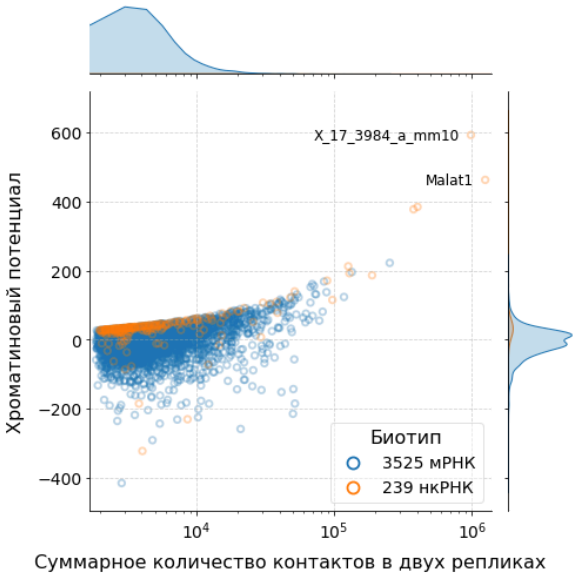
И. RADICL, OPC, M. musculus, контакты в пиках



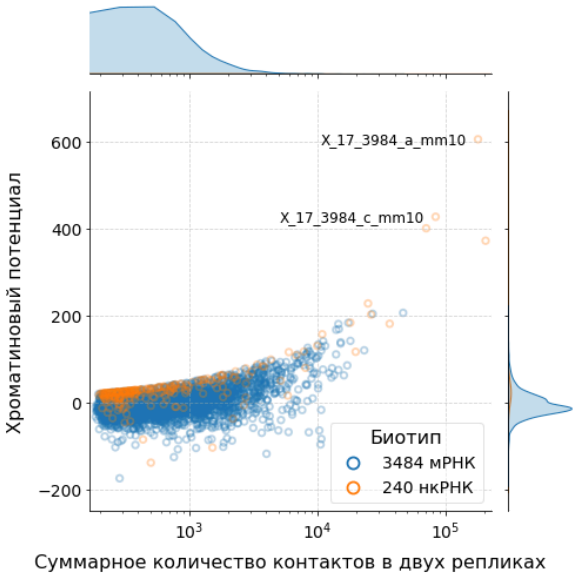
К. RADICL (ActD), ES, M. musculus, все контакты



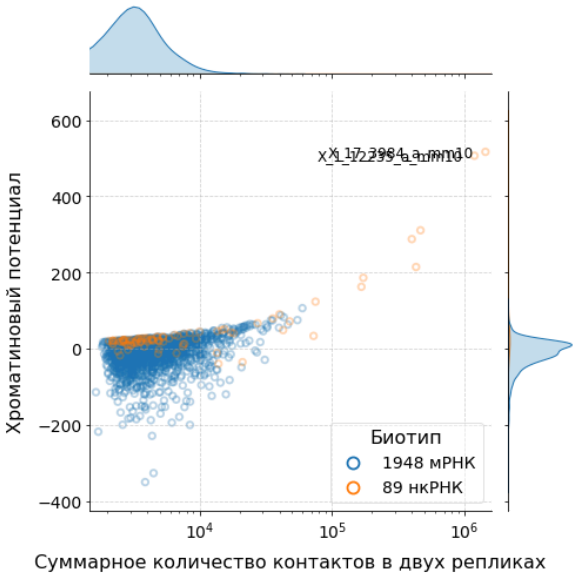
Л. RADICL (ActD), ES, M. musculus, контакты в пиках



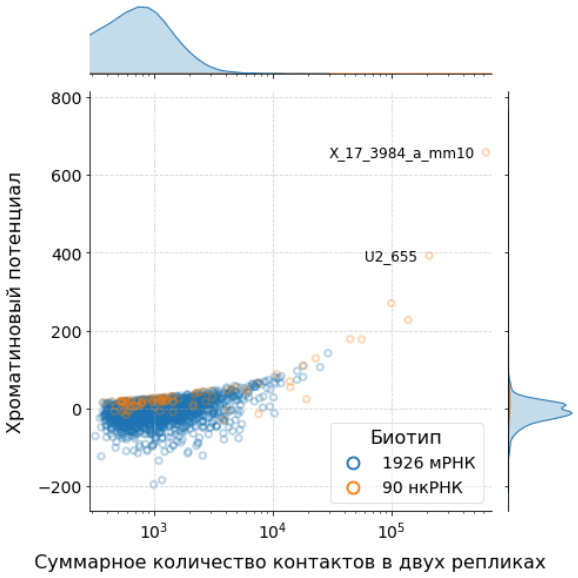
M. RADICL (NPM), OPC, M. musculus, все контакты



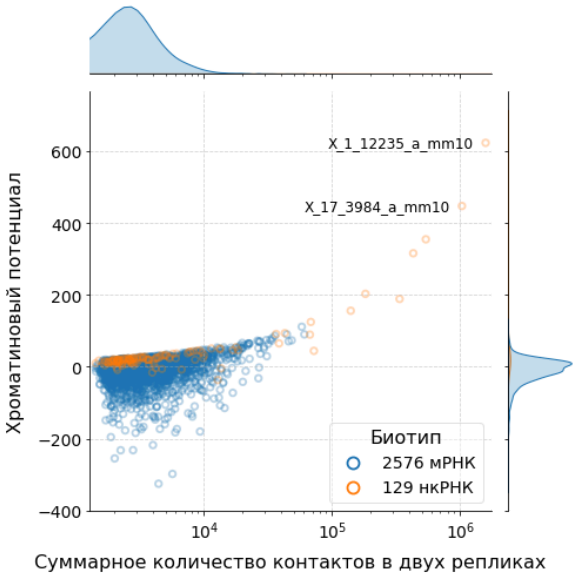
H. RADICL (NPM), OPC, M. musculus, контакты в пиках



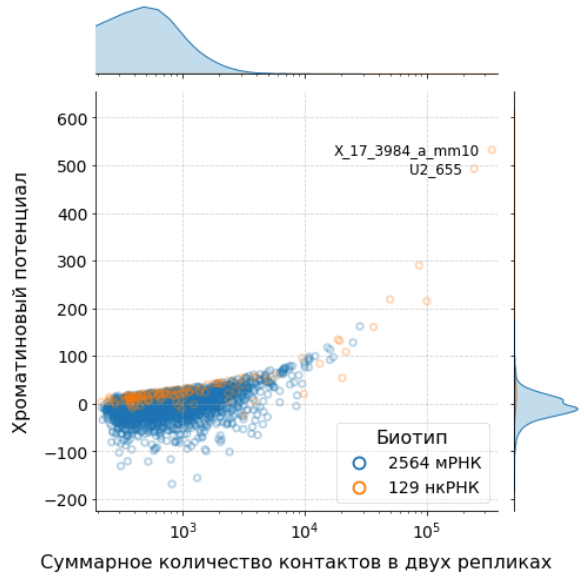
O. RADICL (1% FA), ES, M. musculus, все контакты



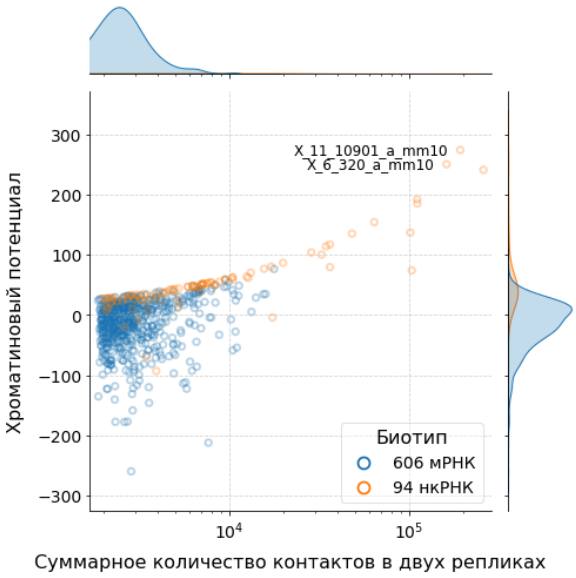
П. RADICL (1% FA), ES, M. musculus, контакты в пиках



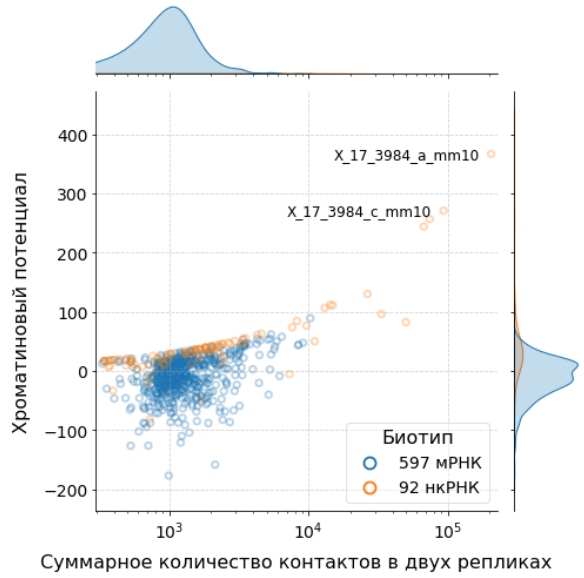
P. RADICL (2% FA), ES, M. musculus, все контакты



C. RADICL (2% FA), ES, M. musculus, контакты в пиках



T. RADICL (NPM), ES, M. musculus, все контакты



У. RADICL (NPM), ES, M. musculus, контакты в пиках

Рисунок В.3. Зависимость хроматинового потенциала от числа контактов для всех контактов и для контактов из пиков BaRDIC.

Таблица В.3. Число РНК, оставшихся при применении фильтра на хроматиновый потенциал (chP).

Эксперимент	chP > 0		chP > 20		chP > 50	
	мРНК	нкРНК	мРНК	нкРНК	мРНК	нкРНК
Red-C, K562, *H. sapiens*	1256	486	769	438	362	328
GRID, MM.1S, *H. sapiens*	1705	374	1002	349	338	235
GRID, MDA_MB_231, *H. sapiens*	1806	564	890	458	250	116
GRID, ES, *M. muculus*	2201	340	691	157	63	48
RADICL, OPC, *M. musculus*	1056	161	564	147	129	40
RADICL, ES (ActD), *M. musculus*	219	54	56	39	2	13
RADICL, OPC (NPM), *M. musculus*	1658	219	861	201	195	41
RADICL, ES (1% FA), *M. musculus*	891	80	285	57	30	13
RADICL, ES (2% FA), *M. musculus*	1150	121	361	65	38	15
RADICL, ES (NPM), *M. musculus*	263	89	105	82	4	35

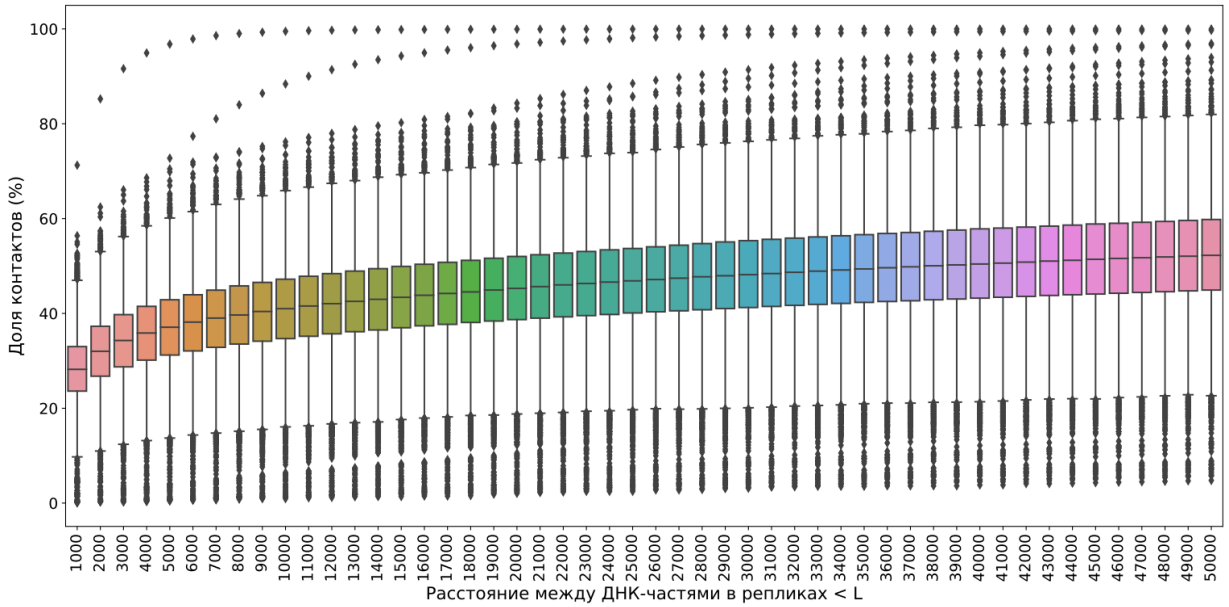


Рисунок В.4. Доли контактов отобранных РНК (больше 1000 контактов с хроматином в каждой реплике) в данных «GRID, ES, Mus musculus» в зависимости от порога на геномное расстояние *L*, в пределах которого контакты, принадлежащие одной РНК, обнаружены в разных репликах.

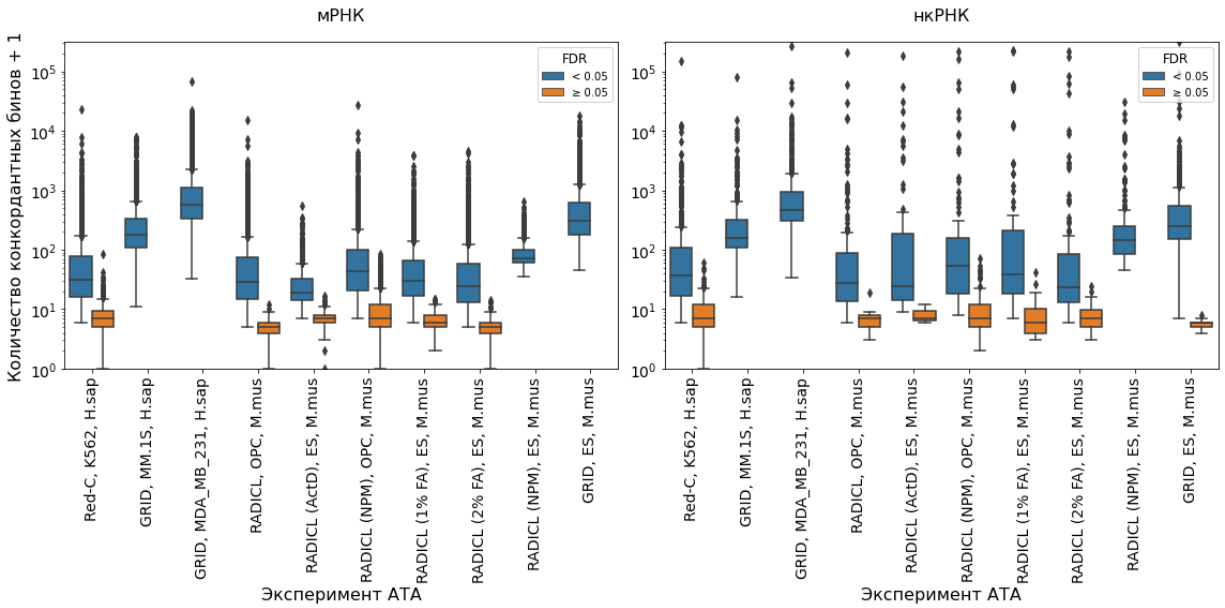


Рисунок B.5. Количество конкордантных бинов у отобранных РНК (больше 1000 контактов с хроматином в каждой реплике) в данных ATA. Размер бина – 5000 п.н., использовались все контакты, находящиеся дальше 1 Мб от генов, кодирующих соответствующие РНК. M.mus – *Mus musculus*; H.sap – *Homo sapiens.*

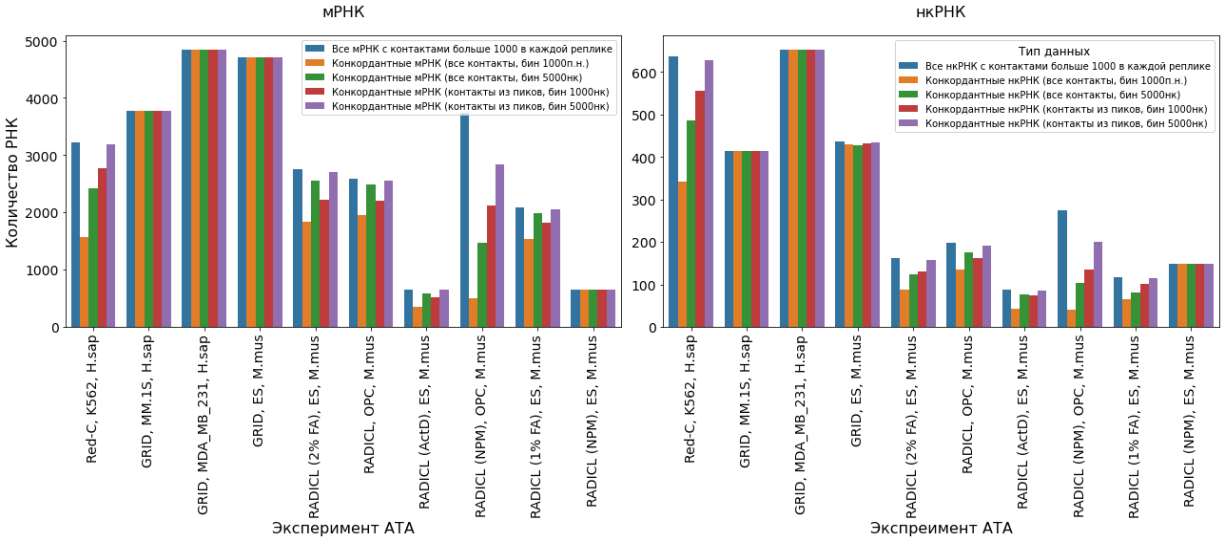


Рисунок B.6. Количество РНК, имеющие конкордантные бины в репликах (FDR < 0,05). В анализе участвовали РНК с числом контактов >1000 в каждой реплике. При отборе контактов применялся фильтр на расстояние от гена-источника РНК равный 1 Мб. M.mus – *Mus musculus*; H.sap – *Homo sapiens.*

Таблица В.4. Медианный процент конкордантных контактов мРНК и нкРНК в различных экспериментах ATA. Размер бина – 5000 п.н.; фильтр на РД-скейлинг – 1 Мб; ActD – обработка актиномицином D; NPM – обработка протеиназой К; 1% FA – обработка сшивающим агентом формальдегидом в концентрации 1%; 2% FA – обработка сшивающим агентом формальдегидом в концентрации 2%.

Эксперимент	**мРНК все**	**мРНК (chP > 20)**	**мРНК (chP > 50)**	**нкРНК все**	**нкРНК (chP > 20)**	**нкРНК (chP > 50)**
	1	2	3	4	5	6
GRID, ES, *M. musculus*	15,7	22,2	34,6	14,2	14,9	9,8
GRID, MM.1S, *H. sapiens*	10,4	11,9	14,3	10,1	10,3	10,6
GRID, MDA_MB_231, *H. sapiens*	28,5	32,4	38,6	28,6	28,9	31,6
Red-C, K562, *H. sapiens*	1,1	1,8	2,5	1,3	1,4	1,6
RADICL, OPC, *M. musculus*	1,8	4,0	8,6	1,7	1,8	6,7
RADICL (ActD), ES, *M. musculus*	1,1	3,1	0,0	1,4	2,5	11,4
RADICL (NPM), OPC, *M. musculus*	1,0	1,5	2,8	1,0	1,0	2,0
RADICL (1% FA), ES, *M. musculus*	1,2	3,4	11,3	1,1	2,9	7,0
RADICL (2% FA), ES, *M. musculus*	1,6	4,3	13,5	1,6	2,4	9,0
RADICL (NPM), ES, *M. musculus*	0,5	1,1	3,5	0,6	1,1	6,8

Таблица В.5. Медианный процент конкордантных контактов мРНК и нкРНК в различных экспериментах ATA. Размер бина – 5000 п.н.; фильтр на РД-скейлинг – 1 Мб. Контакты из пиков. ActD – обработка актиномицином D; NPM – обработка протеиназой К; 1% FA – обработка сшивающим агентом формальдегидом в концентрации 1%; 2% FA – обработка сшивающим агентом формальдегидом в концентрации 2%.

Эксперимент	**мРНК все**	**мРНК (chP > 20)**	**мРНК (chP > 50)**	**нкРНК все**	**нкРНК (chP > 20)**	**нкРНК (chP > 50)**
	1	2	3	4	5	6
GRID, ES, *M. musculus*	58,6	57,3	65,1	56,7	45,1	20,5
GRID, MM.1S, *H. sapiens*	46,9	44,5	44,6	44,6	43,1	39,7
GRID, MDA_MB_231, *H. sapiens*	81,6	73,8	69,8	81,1	77,4	64,7
Red-C, K562, *H. sapiens*	4	7,7	9,2	4,6	5,7	6
RADICL, OPC, *M. musculus*	6	13,8	21,1	5,2	5,7	12,4
RADICL (ActD), ES, *M. musculus*	4,2	10,2	10,3	5	7,6	19,6
RADICL (NPM), OPC, *M. musculus*	3,1	5,5	8,2	3,1	3,4	6,5
RADICL (1% FA), ES, *M. musculus*	3,8	9	22,4	3,2	5,5	10,2
RADICL (2% FA), ES, *M. musculus*	5,6	14,6	29,5	4	4,9	12,5
RADICL (NPM), ES, *M. musculus*	2,5	3,5	7,4	2,6	3,5	10,8

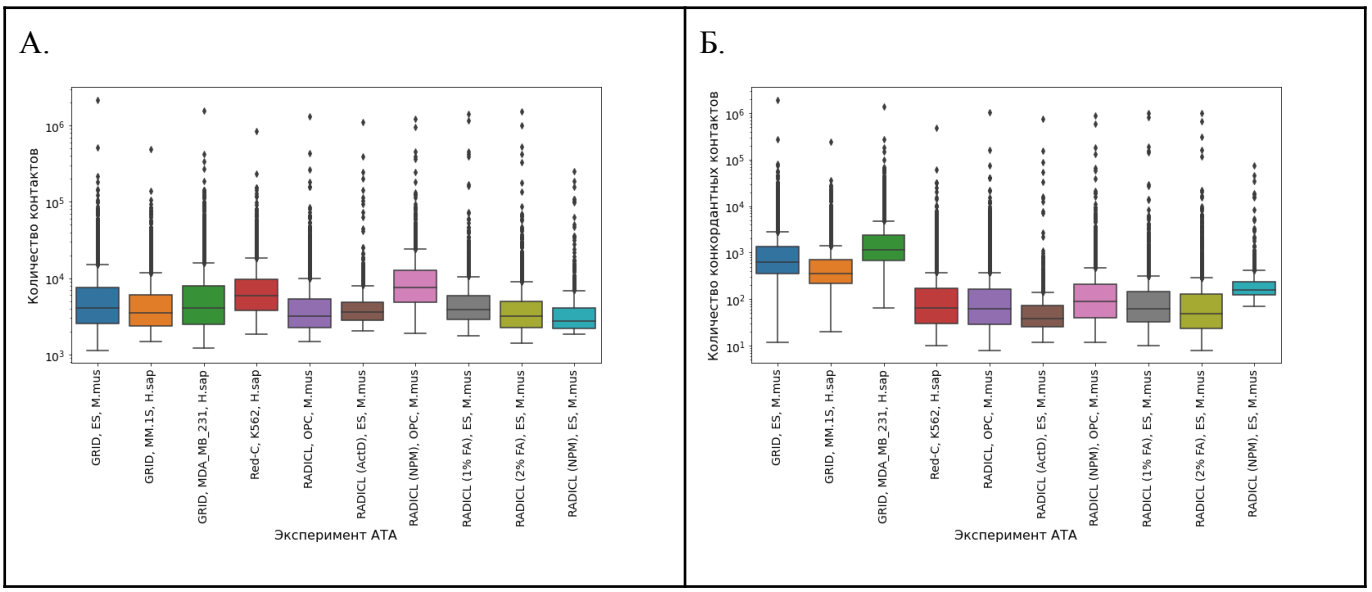
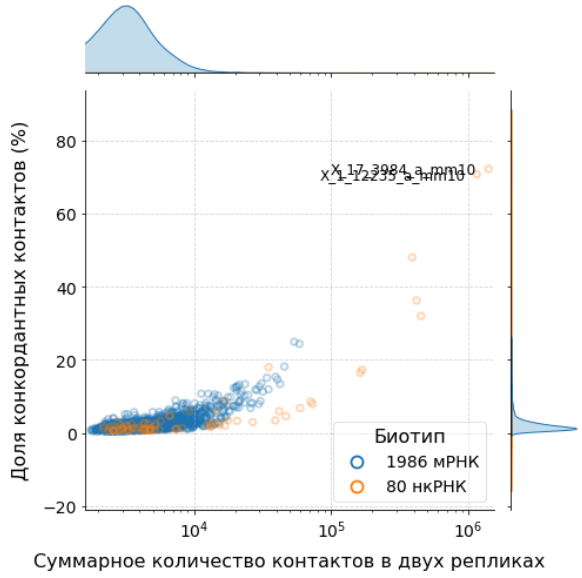
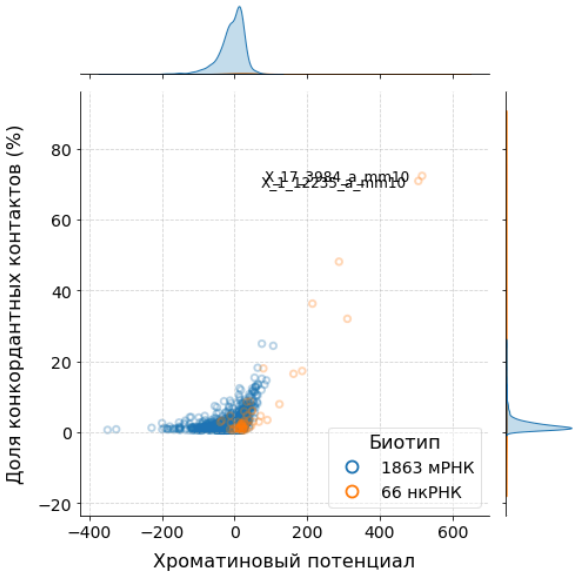


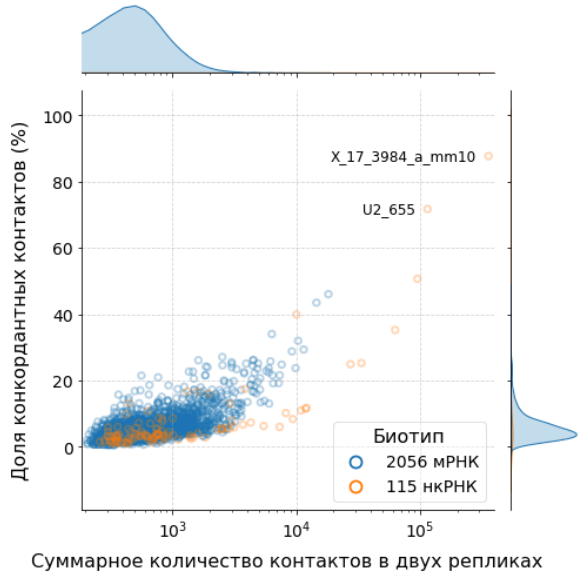
Рисунок В.7. А. Количество контактов у конкордантных РНК в данных ATA. Б. Количество воспроизводимых контактов у конкордантных РНК в данных ATA. Размер бина – 5000 п.н., использовались все контакты, находящиеся дальше 1 Мб от генов, кодирующих соответствующие РНК. M.mus – *Mus musculus*; H.sap – *Homo sapiens.*



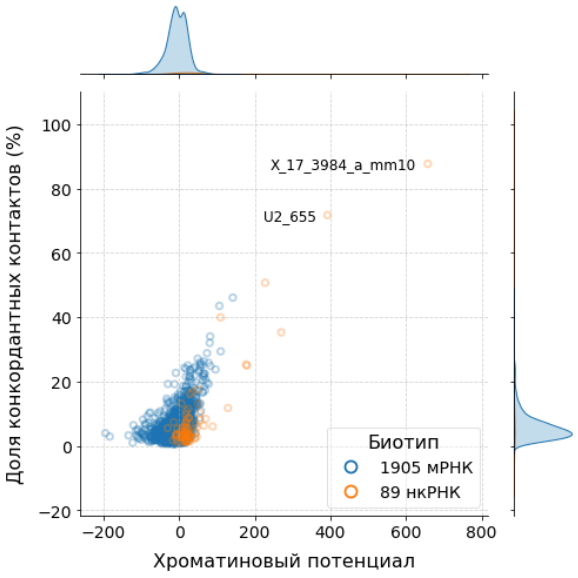
A1. RADICL (1% FA), ES, M. musculus, все контакты



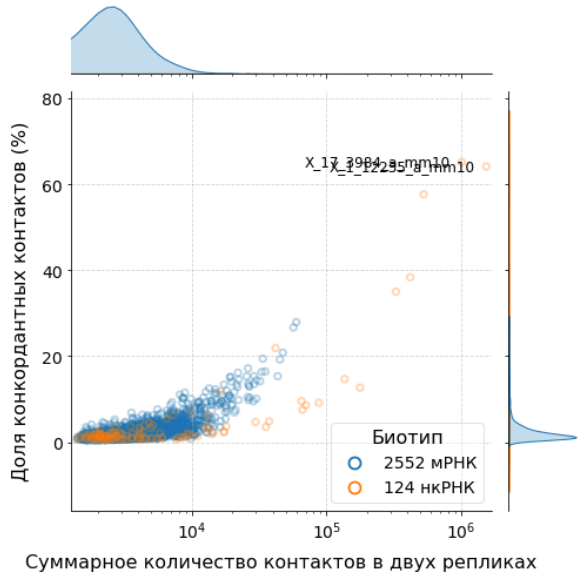
A2. RADICL (1% FA), ES, M. musculus, все контакты



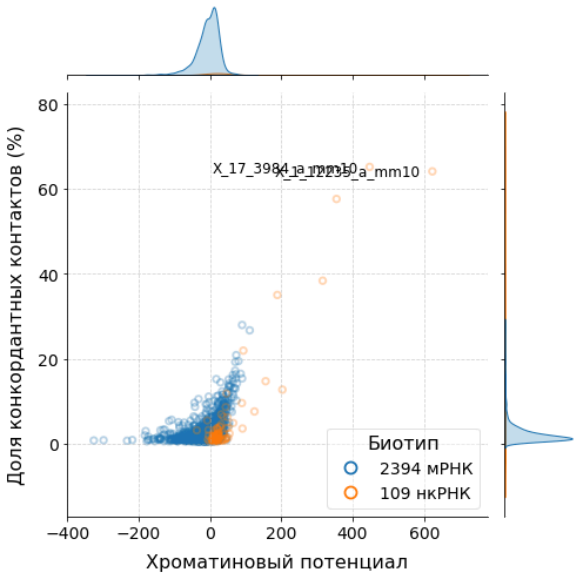
A3. RADICL (1% FA), ES, M. musculus, контакты в пиках



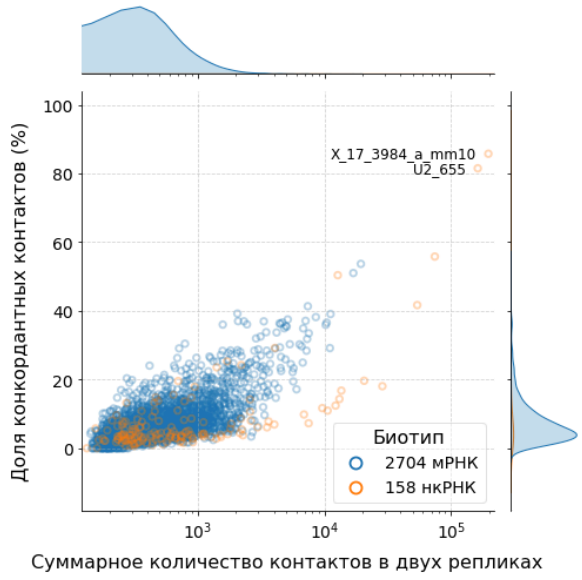
A4. RADICL (1% FA), ES, M. musculus, контакты в пиках



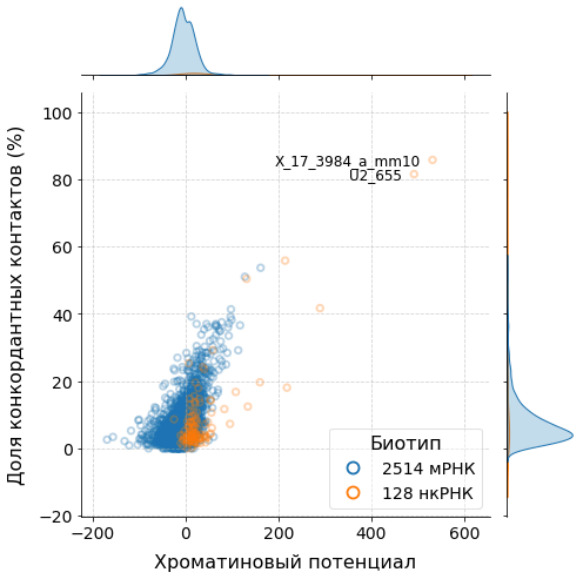
Б1. RADICL (2% FA), ES, M. musculus, все контакты



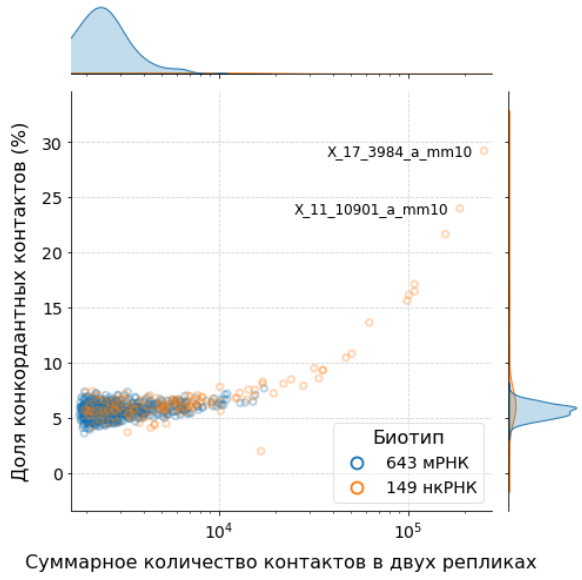
Б2. RADICL (2% FA), ES, M. musculus, все контакты



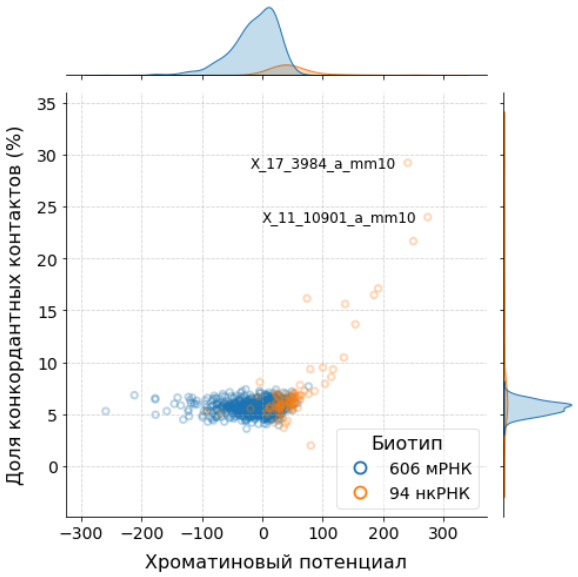
Б3. RADICL (2% FA), ES, M. musculus, контакты в пиках



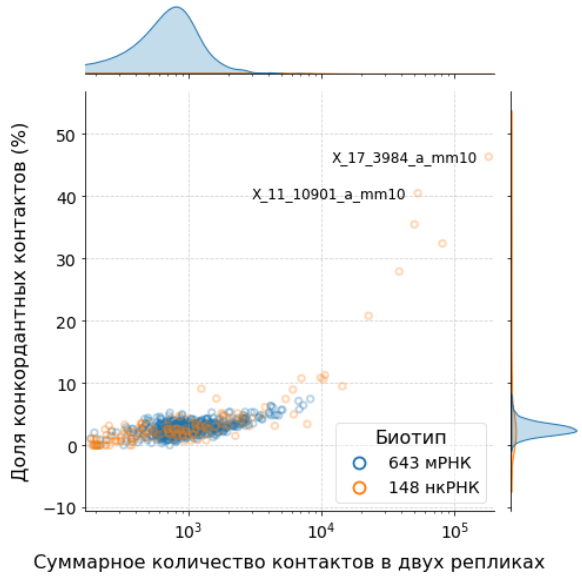
Б4. RADICL (2% FA), ES, M. musculus, контакты в пиках



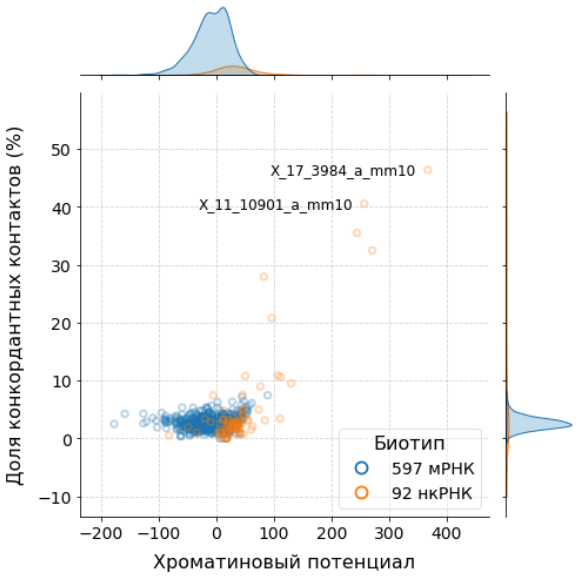
B1. RADICL (NPM), ES, M. musculus, все контакты



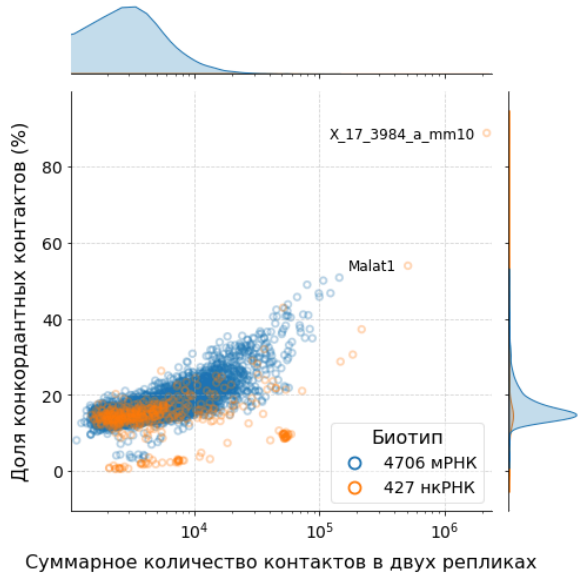
B2. RADICL (NPM), ES, M. musculus, все контакты



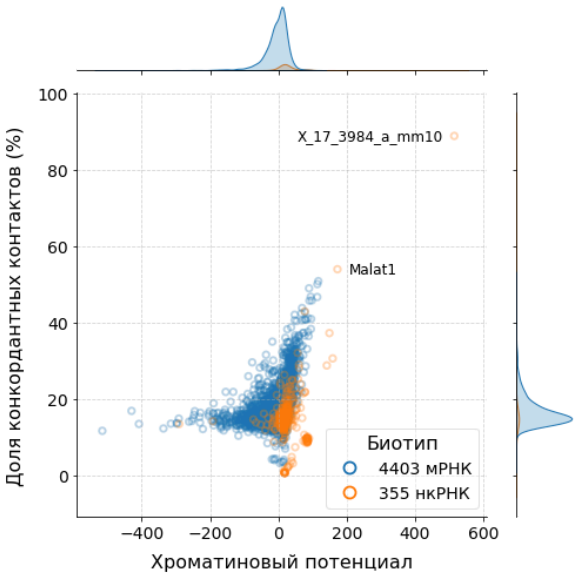
B3. RADICL (NPM), ES, M. musculus, контакты в пиках



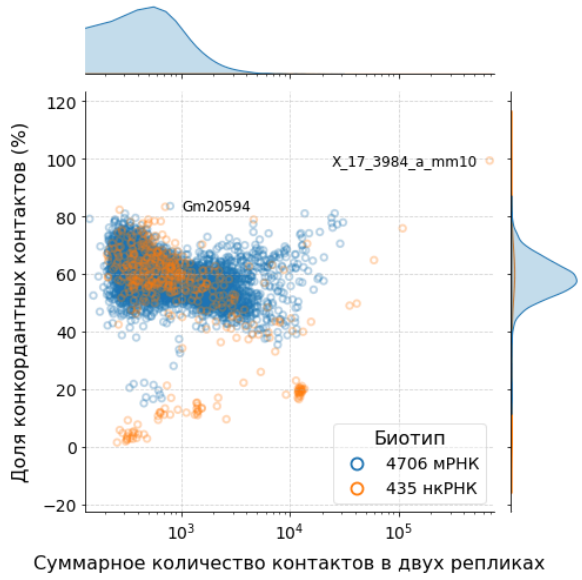
B4. RADICL (NPM), ES, M. musculus, контакты в пиках



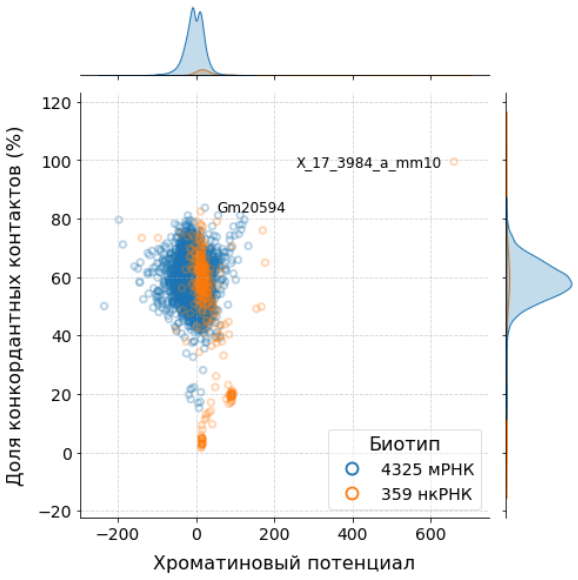
Г1. GRID, ES, M. musculus, все контакты



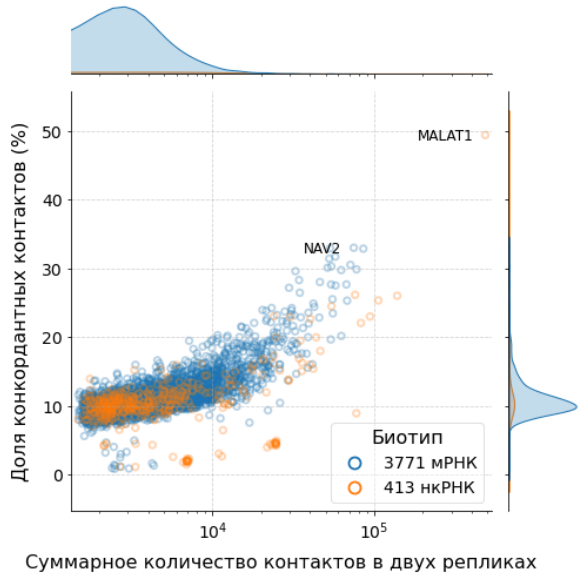
Г2. GRID, ES, M. musculus, все контакты



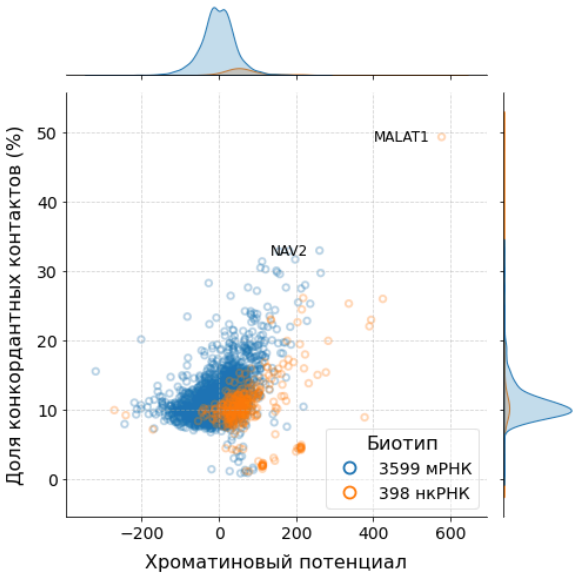
Г3. GRID, ES, M. musculus, контакты в пиках



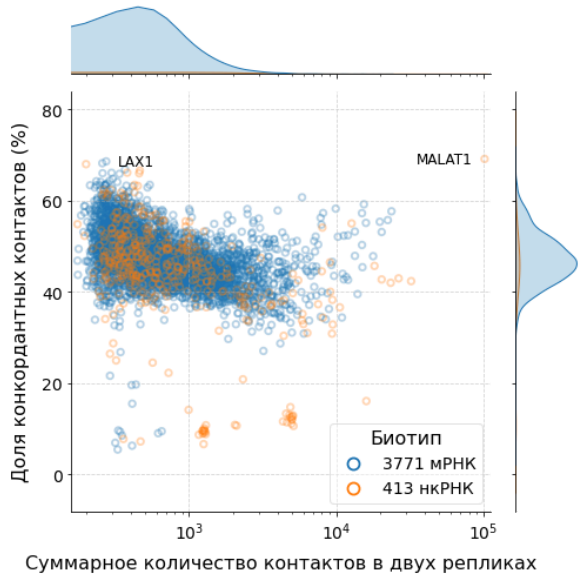
Г4. GRID, ES, M. musculus, контакты в пиках



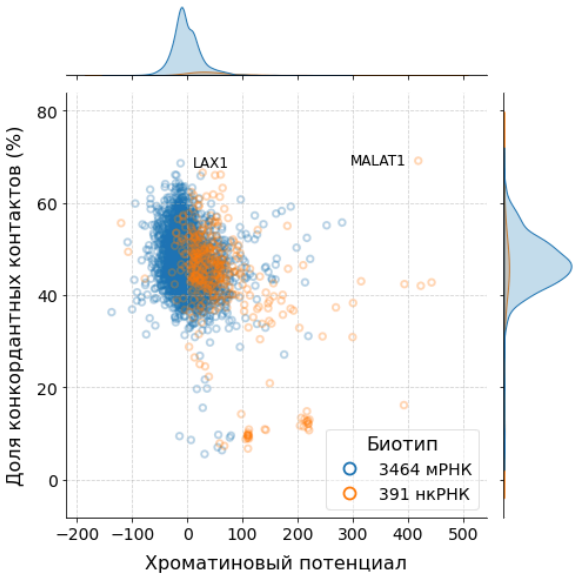
Д1. GRID, MM.1S, H. sapiens, все контакты



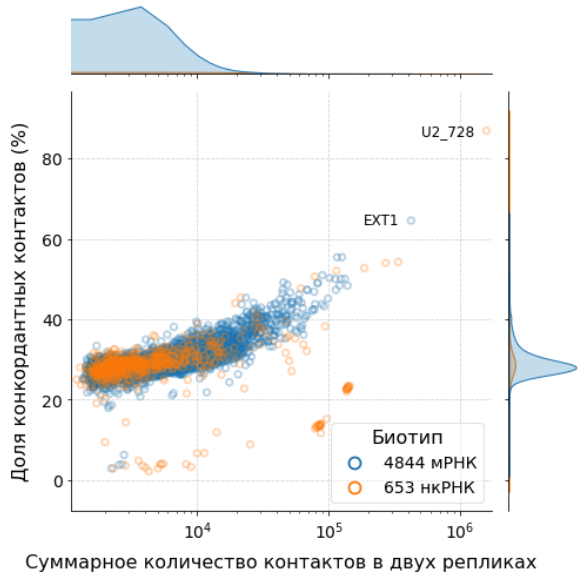
Д2. GRID, MM.1S, H. sapiens, все контакты



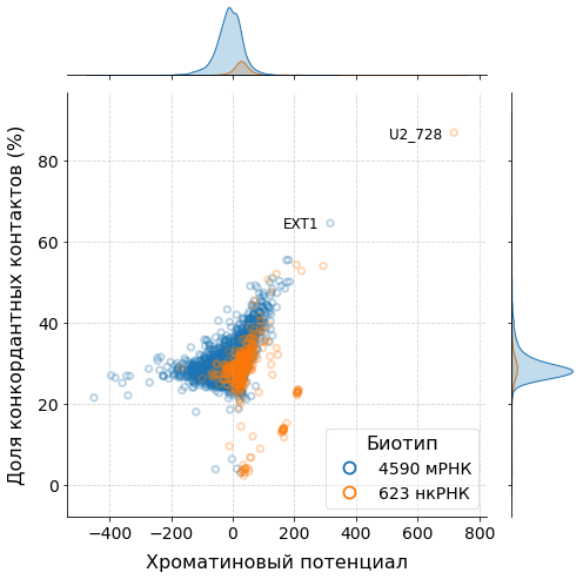
Д3. GRID, MM.1S, H. sapiens, контакты в пиках



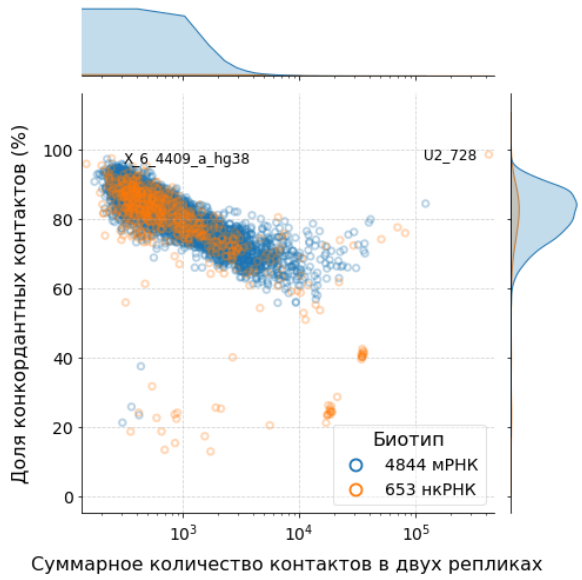
Д4. GRID, MM.1S, H. sapiens, контакты в пиках



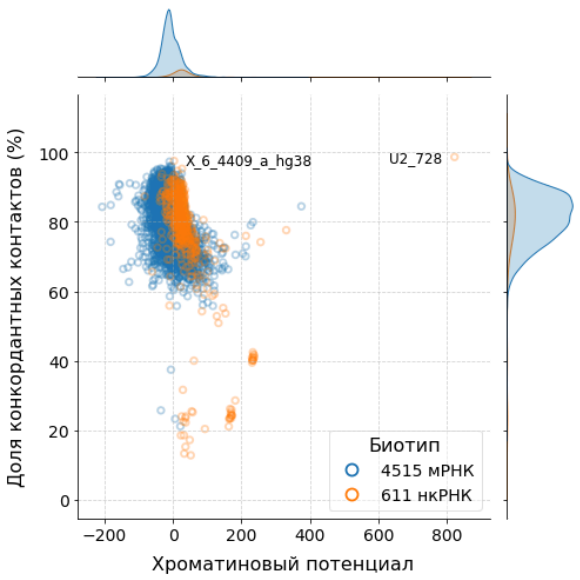
E1. GRID, MDA_MB_231, H. sapiens, все контакты



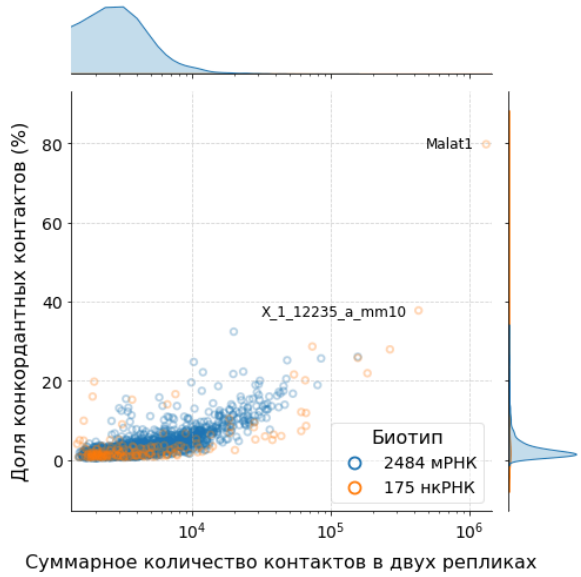
E2. GRID, MDA_MB_231, H. sapiens, все контакты



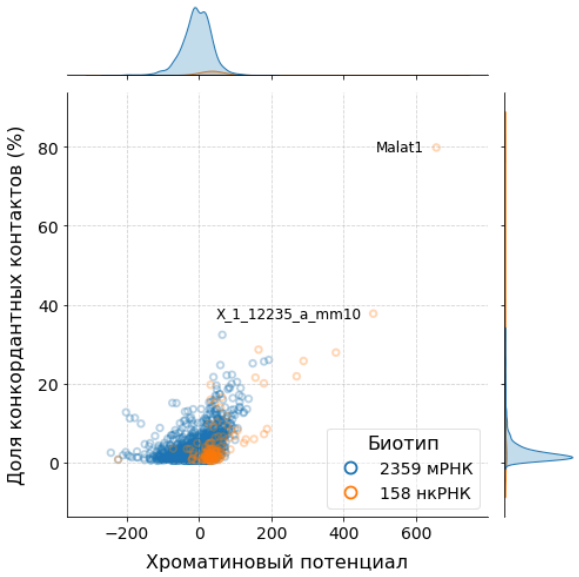
E3. GRID, MDA_MB_231, H. sapiens, контакты в пиках



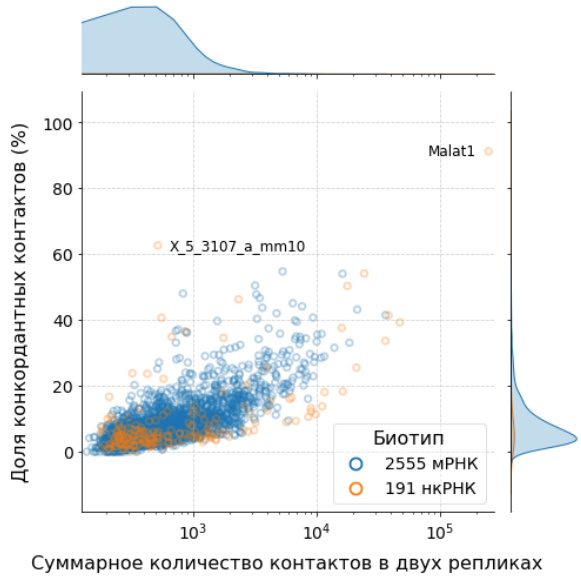
E4. GRID, MDA_MB_231, H. sapiens, контакты в пиках



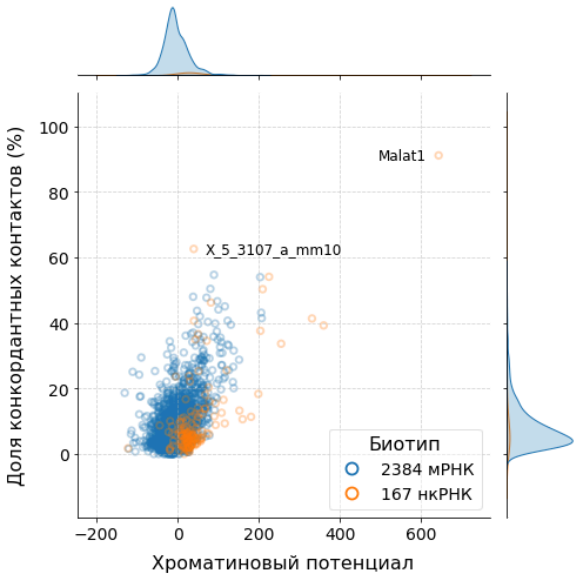
Ж1. RADICL, OPC, M. musculus, все контакты



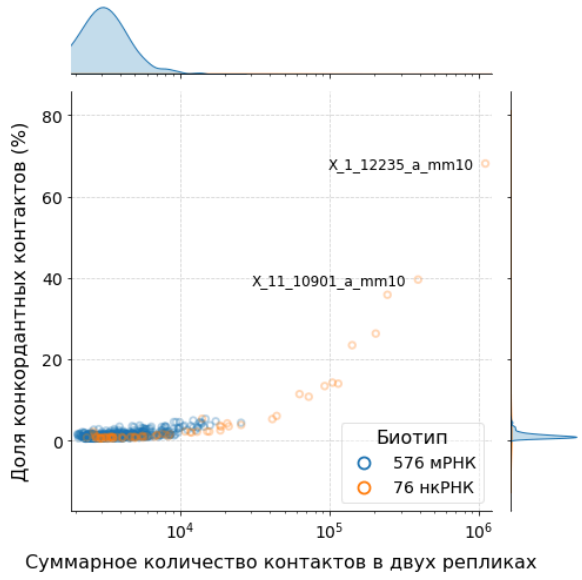
Ж2. RADICL, OPC, M. musculus, все контакты



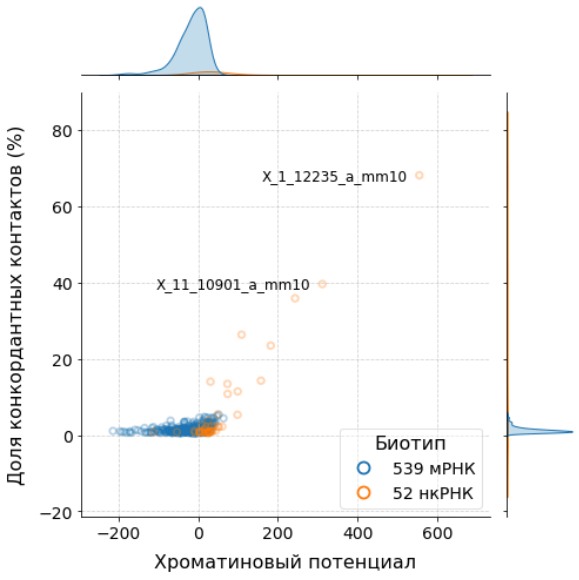
Ж3. RADICL, OPC, M. musculus, контакты в пиках



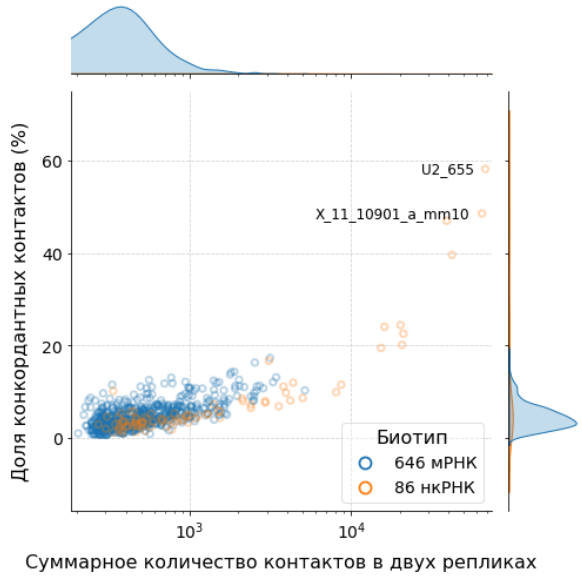
Ж4. RADICL, OPC, M. musculus, контакты в пиках



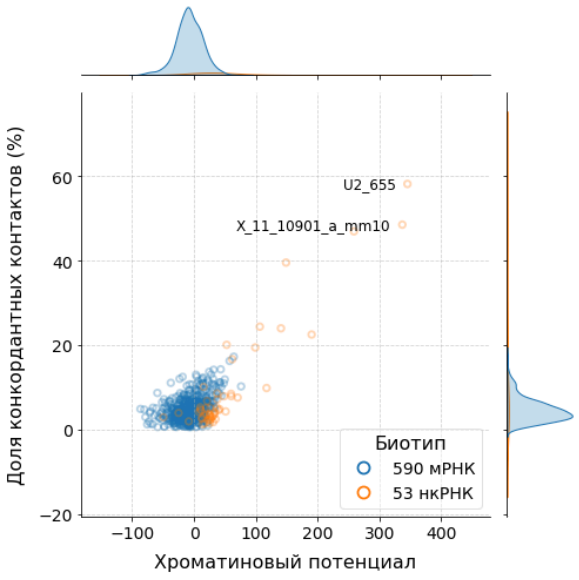
31. RADICL (ActD), ES, M. musculus, все контакты



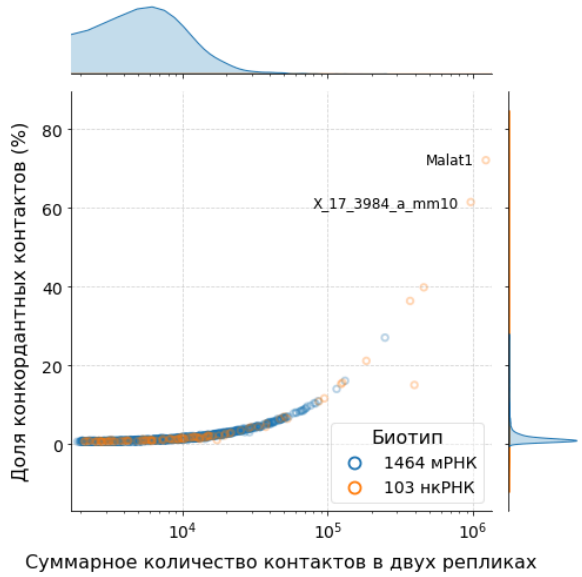
32. RADICL (ActD), ES, M. musculus, все контакты



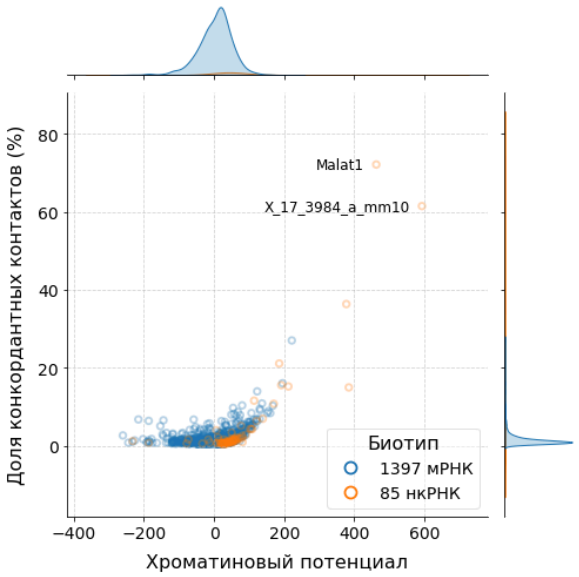
33. RADICL (ActD), ES, M. musculus, контакты в пиках



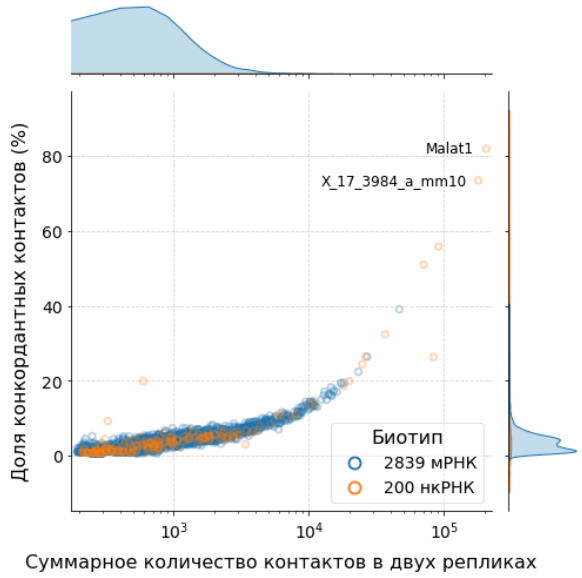
34. RADICL (ActD), ES, M. musculus, контакты в пиках



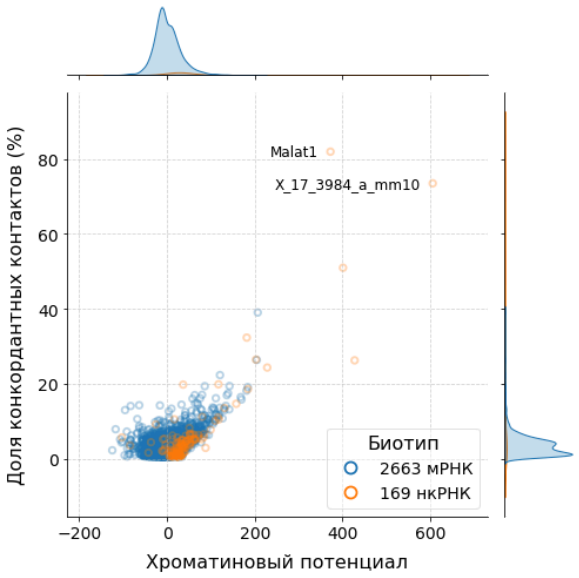
И1. RADICL (NPM), OPC, M. musculus, все контакты



И2. RADICL (NPM), OPC, M. musculus, все контакты



И3. RADICL (NPM), OPC, M. musculus, контакты в пиках



И4. RADICL (NPM), OPC, M. musculus, контакты в пиках

Рисунок В.8. Зависимость доли конкордантных контактов от полного числа контактов, от контактов из пиков BaRDIC и от хроматинового потенциала для разных экспериментов ATA. M.mus – *Mus musculus*; H.sap – *Homo sapiens.*

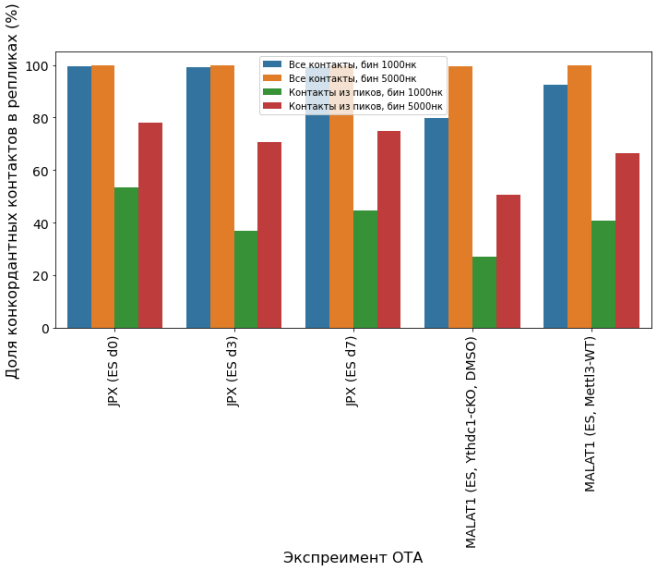


Рисунок В.9. Уровень конкордантности реплик экспериментов с индивидуальными РНК (ОТА).

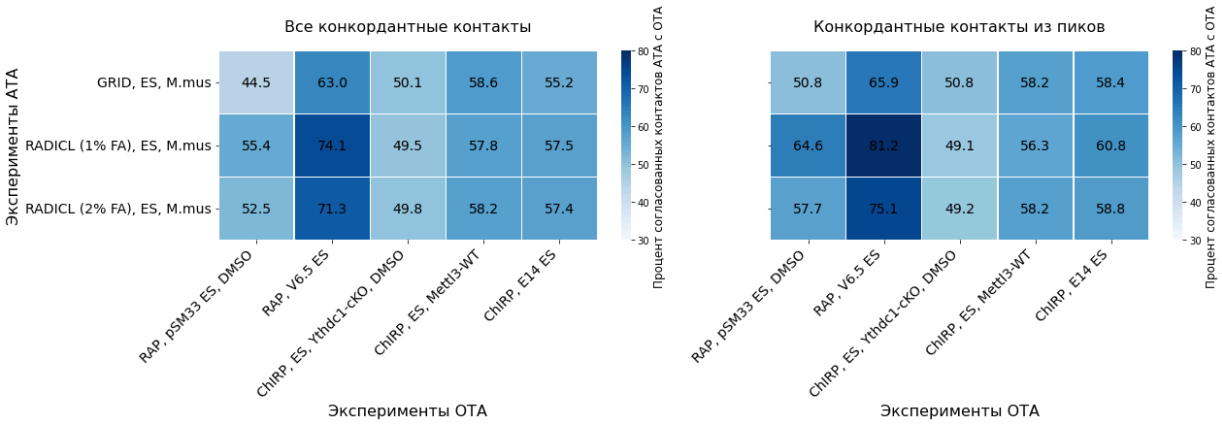


Рисунок В.10. Согласованность конкордантных контактов (слева) и конкордантных контактов из пиков BaRDIC нкРНК MALAT1 из данных АТА с контактами из пиков BaRDIC нкРНК MALAT1 из экспериментов ОТА в эмбриональных стволовых клетках мыши. Размер бина – 5000 п.н., *p-value* < 0,05. M.mus – *Mus musculus*.

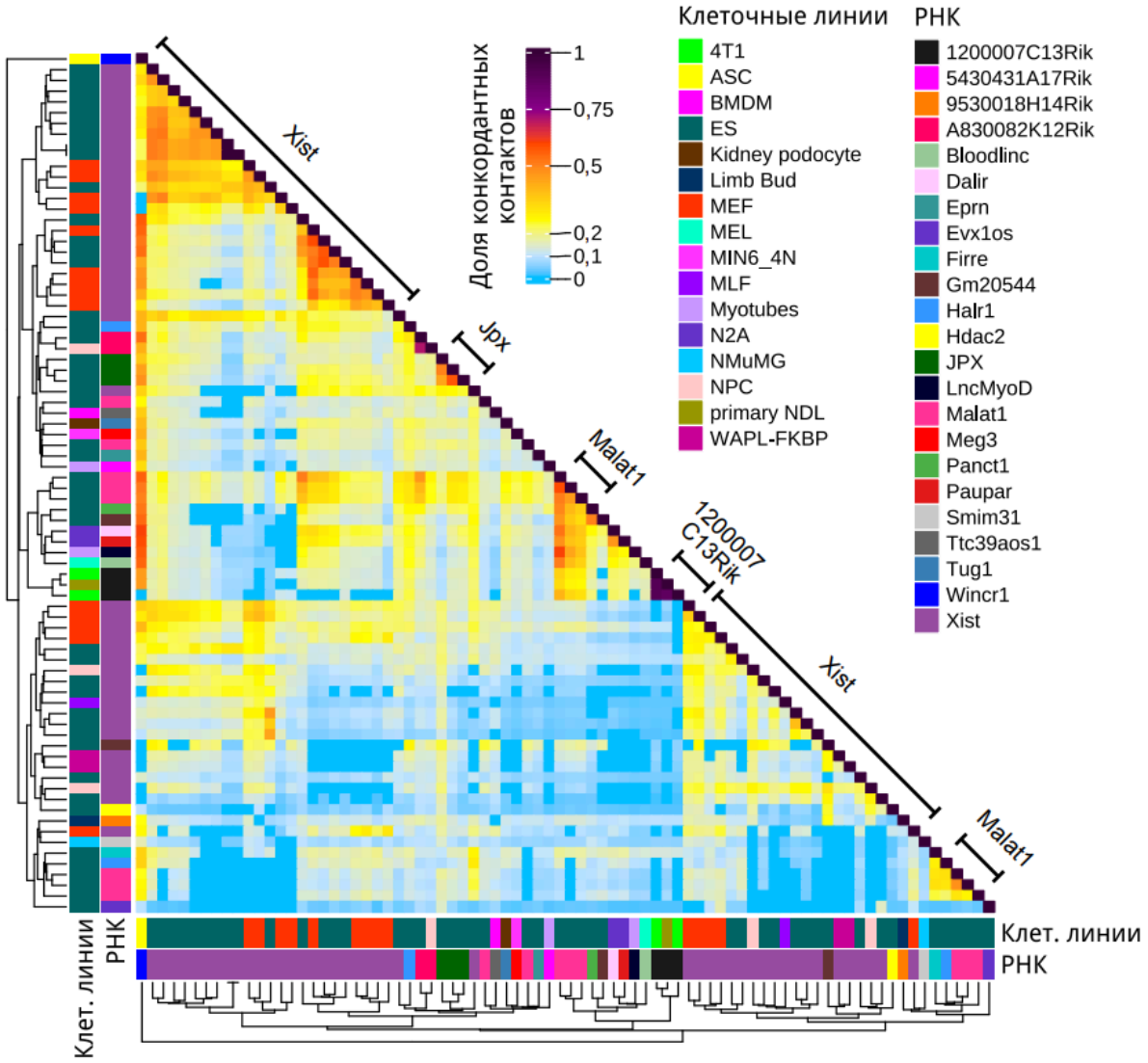


Рисунок В.11. Тепловая карта, отражающая долю согласованных контактов (из пиков BaRDIC, FDR < 0,05) из экспериментов группы «один-против-всех» для клеточных линий мыши. Незначимые обогащения (*p-value* > 0,05) обнулены. Кластеризация проведена по клеточным типам и РНК, используемым в эксперименте. Размер бина составляет 1000 п.н.