МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА,
ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ

На правах рукописи

Жарикова Анастасия Александровна

Биоинформатический анализ РНК-хроматиновых взаимодействий

03.01.09 - математическая биология, биоинформатика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата биологических наук

Научный руководитель:
д.б.н., профессор Миронов Андрей Александрович

Москва – 2022

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ

мРНК - матричные РНК

нкРНК - некодирующие РНК

хаРНК - хроматин-ассоциированные РНК

мяРНК - малые ядерные РНК

мякРНК - малые ядрышковые РНК

ТАДы - топологически ассоциированные домены

ПЦР - полимеразно-цепная реакция

Кб - килобаза (1000 нуклеотидов)

Мб - мегабаза (1 млн нуклеотидов)

DSG - disuccinimidyl glutarate

RPKM - reads per kilobase of transcript per million mapped reads

FDR - false discovery rate

TCGA - the cancer genome atlas

NPM - non-protein mediated

RAP - RNA antisense purification

CHART-seq - capture hybridization analysis of RNA targets

ChIRP-seq - chromatin Isolation by RNA purification

MARGI - mapping of RNA-genome interactions

Red-C - RNA ends on DNA capture

RADICL-seq - RNA and DNA interacting complexes ligated and sequenced

GRID-seq - global RNA interactions with DNA

ChIP-seq - chromatin immunoprecipitation

eCLIP - enhanced crosslinking and immunoprecipitation

RIC-seq - RNA in situ conformation sequencing

MARIO - mapping RNA interactome in vivo

GRO-seq - global run-on sequencing

ChAR-seq - chromatin-associated RNA sequencing

ATAC-seq - assay for Transposase-Accessible Chromatin

ВВЕДЕНИЕ

Актуальность темы исследования

С приходом технологий высокопроизводительного секвенирования в лабораторную практику удалось установить, что внушительная часть генома эукариот способна к транскрипции с образованием большого количества РНК, включая белок-кодирующие (мРНК), а также разнообразные длинные и короткие некодирующие РНК (нкРНК) [1–4]. Генные аннотации постоянно расширяются в основном за счет включения в них не только отдельных представителей нкРНК, но и целых новых классов нкРНК [5,6]. Так, в 2006 году сразу несколькими группами были представлены короткие некодирующие РНК, взаимодействующие с белками класса PIWI (piРНК) [7,8], которые до сих пор являются самым многочисленным классом РНК, согласно актуальным аннотациям. В научных кругах ведутся оживленные споры относительно существования и функций кольцевых РНК (circРНК) [9], консорциумом FANTOM было показано наличие транскрипционного потенциала некоторых энхансерных областей [10].

Молекулы РНК могут выполнять свои функции не только в цитоплазме клетки, но и в ядре, где они принимают активное участие в таких важных для жизнедеятельности клетки процессах, как регуляция транскрипции, ремоделирование и поддержание структуры хроматина, формирование ядерных телец [3,4,11,12]. Хрестоматийным примером нкРНК, работающей в непосредственной связке с хроматином, может служить длинная нкРНК XIST, которая участвует в инактивации Х-хромосомы у самок млекопитающих. MALAT1 и NEAT1, тоже представители класса длинных нкРНК, участвуют в формировании ядерных спеклов и параспеклов, соответственно [13]. Упомянутые выше piРНК также работают в ядре, реализуя в том числе ко-транскрипционное подавление транспозонов [14].

Несмотря на всю важность функций, в которых принимают участие длинные и короткие нкРНК, механизмы действий изучены лишь для некоторых из них [15]. Если малые нкРНК объединяют в классы, исходя из их общих механизмов действий

и основных функций, то группа длинных нкРНК содержит совершенно разнородные РНК, выполняющие множество разных функций, а их количество сопоставимо с количеством белок-кодирующих РНК [16–18]. Пристального внимания и изучения заслуживает каждый представитель этой группы.

Методы, применяемые для изучения РНК, ассоциированных с хроматином, существуют давно и постоянно развиваются. Еще в середине прошлого века с помощью биохимических методов был установлен сам факт существования фракции хроматин-ассоциированных РНК (хаРНК), а сегодня с помощью современных лабораторных протоколов и высокопроизводительного секвенирования можно получать карты взаимодействия РНК с хроматином в достаточно хорошем разрешении. Существует целый спектр методик, позволяющих полногеномно выявить локусы ДНК, с которыми взаимодействуют РНК [19]. Однако, до 2017 года такие методы позволяли в рамках одного эксперимента изучать только одну или небольшое количество заранее известных РНК. Подобные подходы называют “один-против-всех”.

Появление полногеномных протоколов, с помощью которых можно было бы сразу для всех потенциальных хаРНК установить их локусы взаимодействия с хроматином, радикальным образом продвинуло бы вперед исследования в области некодирующих РНК.

В данной работе представлен биоинформатический подход, позволяющий анализировать результаты оригинального экспериментального протокола по определению РНК-ДНК интерактома - Red-C.

Степень разработанности темы исследования

За последние шесть лет появилось сразу несколько методов для изучения РНК-ДНК интерактома; такие подходы получили название “все-против-всех” [20–27].

Представленные методики идеологически похожи между собой и базируются на лигировании расположенных близко в пространстве макромолекул, что

порождает химерную РНК-ДНК конструкцию, последовательность которой расшифровывается при помощи высокопроизводительного секвенирования с последующей биоинформатической обработкой. Все манипуляции проводят после фиксации клеток, чаще с помощью формальдегида. Ключевым фигурантом в процессе подготовки объекта для секвенирования является особым образом сконструированный полярный линкер. Структура линкера позволяет с одной его стороны лигировать фрагмент РНК, а с другой - фрагмент ДНК так, что в процессе обработки можно точно установить чтения, пришедшие с соответствующей нуклеиновой кислоты. Отличия методов заключаются в основном в деталях структуры линкера, способах фиксации клеток и их количестве, применяемых рестриктазах, длинах секвенируемых фрагментов, подходах к анализу результатов секвенирования. Предложенные методы практически не пересекаются с точки зрения выбора объекта исследования, каждый протокол реализован на уникальной клеточной линии. Результаты обработаны с помощью биоинформатических протоколов, созданных специально для конкретного метода. Протоколы обработки данных отличаются в том числе выбором программ для картирования чтений на референсный геном, версиями референсных геномов, а также источниками генной аннотации. Все это затрудняет совместный анализ результатов этих методов и их сравнение. Во всех протоколах представлены контрольные эксперименты, позволяющие убедиться в корректности пробоподготовки.

Авторы опубликованных методов наблюдают высокий уровень шума в данных, большое количество детектируемых мРНК, а также наибольшую плотность контактов РНК рядом со своим геном. Тем не менее везде отмечают согласованность в поведении выбранных контрольных РНК между полногеномным подходом и исследованием единичной РНК по данным “один-против-всех”.

Наша группа в коллаборации с лабораторией С.В. Разина принимала участие в обработке данных одного из методов “все-против-всех” - Red-C [25]. Наиболее близкими к Red-C с точки зрения экспериментальной процедуры являются

протоколы GRID-seq [21] и RADICL-seq [24]. Авторы GRID-seq отметили корреляцию количества контактов РНК с уровнем экспрессии по данным GRO-seq, предложен способ выделения специфических пиков контактов РНК с ДНК. Протокол GRID-seq был реализован на клеточных линиях человека, мыши и мухи, выделены РНК, которые предпочитают связываться с разными локусами на хроматине, что может говорить о специфичности этих РНК в клеточных регуляторных путях. Протокол RADICL-seq реализован на клеточных линиях мыши (эмбриональные стволовые клетки и клетки-предшественники олигодендроцитов). Авторы отмечают, что РНК, локализованные внутри топологически ассоциированных доменов (ТАД), предпочитали контактировать с ДНК из этих же ТАДов. Также были выделены РНК с тканеспецифичным относительно исследуемых клеточных линий профилем взаимодействия с хроматином.

Цель и задачи работы

Цель настоящей работы заключается в биоинформатическом анализе данных полногеномного РНК-ДНК интерактома на примере экспериментального протокола Red-C.

Были поставлены следующие **задачи**:

1. Анализ первичных результатов секвенирования данных протокола Red-C.
2. Сборка и фильтрация РНК-ДНК контактов.
3. Разработка нормировок и метрик, позволяющих выявить хроматин-ассоциированные РНК.
4. Аннотация РНК-частей контактов генной разметкой с разрешением ситуаций неоднозначной аннотации.
5. Сборка новых (неаннотированных ранее) хроматин-ассоциированных РНК.
6. Изучение характера взаимодействия выявленных хроматин-ассоциированных РНК с ДНК.
7. Распространение разработанного подхода на другие данные из экспериментов по изучению РНК-ДНК интерактома.

Объект и предмет исследования

Объектом исследования являются РНК, которые выполняют в ядре регуляторные функции, взаимодействуя с хроматином. Предметом исследования являются данные секвенирования, полученные в результате выполнения экспериментальных полногеномных протоколов по изучению хроматин-ассоциированных РНК. Это новый тип данных, позволяющий в рамках одного эксперимента установить для всех потенциальных хроматин-ассоциированных РНК локусы их взаимодействия с хроматином.

Научная новизна

В работе представлен анализ данных из оригинальной работы по изучению РНК-ДНК интерактома с помощью метода Red-C, опубликованный впервые. Данные подобного типа появились в 2017 году, представлены всего лишь в нескольких публикациях и предоставляют возможность изучать РНК, ассоциированные с хроматином, не имея никаких априорных знаний об этих РНК. Предложенный алгоритм анализа разработан специально для протокола Red-C, однако может быть с легкостью применен и к результатам, полученным с помощью других схожих протоколов. Для аннотации РНК-частей генами была разработана процедура голосования, учитывающая случаи неоднозначной аннотации. На основании дополнительной информации об уровне экспрессии разработана и рассчитана метрика хроматинового потенциала. Предложен подход к изучению характера взаимодействия РНК с хроматином. Несмотря на то, что авторы аналогичных работ отмечали, что наблюдают фрагменты РНК, которые детектированы как контактирующие с ДНК, но не попадали в генную разметку, анализ таких РНК-частей произведен не был. В данной работе таким неаннотированным РНК-частям уделено особое внимание, в результате чего удалось собрать гипотетически новые хроматин-ассоциированные РНК.

Практическая значимость

Разработанный биоинформатический подход к анализу полногеномных данных РНК-хроматиновых взаимодействий, позволяет единообразно обрабатывать любые данные из методов типа “все-против-всех”, вне зависимости от исходного протокола. Для анализа можно использовать необходимые референсные геномы любой версии сборки, любые генные аннотации. Первичный анализ данных, состоящий из технических этапов получения последовательностей РНК и ДНК-частей контактов по данным секвенирования, их картирование на референсный геном и сборка контактов, порождает огромное количество материала. Этапы последующего анализа позволяют выявить потенциальные хроматин-ассоциированные РНК, установить характер их взаимодействия с хроматином, рассчитать хроматиновый потенциал. Таким образом можно отобрать небольшое количество РНК-кандидатов с заданными характеристиками и известной последовательностью, включая ранее не аннотированные РНК, для последующей экспериментальной проверки.

Предложенный протокол был использован при создании базы данных RNAChrom, посвященной анализу хроматин-ассоциированных РНК (https://rnachrom2.bioinf.fbb.msu.ru/). Существующие на сегодняшний день данные из экспериментов по изучению РНК-ДНК интерактома были обработаны единым образом и доступны для анализа средствами базы данных и для загрузки.

Методология и методы исследования

Работа была выполнена с использованием разнообразных программ и пакетов, а также программных сценариев, написанных самостоятельно.

Для манипуляции с геномными интервалами были использованы программа bedtools и пакет для R GenomicRanges. В качестве источника базовой генной разметки для человека и мыши был выбран проект GENCODE, аннотация дополнена разметкой малых РНК и очень длинных некодирующих РНК. Для работы с табличными данными, а также для визуализации результатов были в

основном использованы возможности Tidyverse (коллекция пакетов для R). Исследование корреляции полногеномных разметок, а также процедура сглаживания полногеномных сигналов были осуществлены средствами программы Stereogene.

Для анализа данных RNA-seq применялся общепринятый подход. Секвенированные прочтения были картированы на референсный геном с помощью программы HISAT2, учитывающей возможность сплайсинга. Из находящихся в открытом доступе результатов проектов RNA Atlas и ENCODE были получены данные об уровне экспрессии (RNA-seq) для нескольких клеточных линий человека (K562, дермальные фибробласты, MDA-MB-231), а также для мышиных эмбриональных стволовых клеток.

Проведенный анализ реализован на языках программирования R с использованием вспомогательных сценариев на bash.

Положения, выносимые на защиту

1. Предложенный биоинформатический подход для анализа данных РНК-ДНК интерактома, полученных из экспериментов, основанных на лигировании расположенных близко в пространстве макромолекул, позволяет производить нормировку, учитывающую фоновые взаимодействия, разрешать ситуации неоднозначной аннотации в генной разметке и может быть применен к любым данным такого типа.
2. С помощью предложенной метрики хроматинового потенциала для протокола по изучению РНК-ДНК интерактома Red-C (клеточная линия K562) было выявлено 1823 хроматин-ассоциированных РНК, которые взаимодействуют с хроматином чаще, чем это ожидается, исходя из уровня их экспрессии.
3. Выявлены неизвестные ранее хроматин-ассоциированные РНК, произведена их классификация.
4. Хроматин-ассоциированные РНК можно классифицировать в зависимости от удаленности места контакта РНК от своего гена и характера взаимодействия с состояниями хроматина.

Личный вклад автора

Личный вклад автора заключается в разработке многоступенчатого биоинформатического подхода для обработки полногеномных данных РНК-ДНК взаимодействий (протокол Red-C), включая исследование контрольных экспериментов, сбор необходимых метрик по каждому этапу анализа, конструирование трека фоновых контактов и расчет хроматинового потенциала. Этапы первичной подготовки данных, включающие удаление технических последовательностей, картирование РНК и ДНК-частей контактов на референсный геном, сборку первичных контактов, разработаны при активном участии автора. Технически первичная подготовка данных реализована и имплементирована для данных Red-C Александрой Галицыной (https://github.com/agalitsyna/RedClib), для экспериментов GRID-seq и RADICL-seq RedClib модифицирован и применен Юрием Коростелевым и Андреем Сигорских.

Также автором были обработаны все дополнительные данные, необходимые для анализа (результаты секвенирования РНК от исходных чтений, разметка состояний хроматина и пр.).

В личный вклад автора входила биологическая интерпретация и визуализации полученных результатов, представление результатов на научных конференциях, участие в подготовке публикаций в рецензируемых научных журналах.

Степень достоверности данных

Данные, представленные в работе, получены с использованием современных программ и пакетов. Результаты воспроизводимы. Обзор литературы и обсуждение подготовлены с использованием актуальной литературы.

Публикации по теме диссертации[1]

По материалам диссертации опубликовано 4 статьи в рецензируемых научных журналах, в том числе в *Nucleic Acids Research*, *Methods in molecular biology (Clifton, N.J.)* и *Молекулярная биология* (2 статьи).

1. **A. A. Zharikova** and A. A. Mironov. pirnas: Biology and bioinformatics. Молекулярная биология, 50(1):80–88, 2016 [IF = 1.678] (0,5 / 0,45)
2. Potashnikova, D. M., Golyshev, S. A., Penin, A. A., Logacheva, M. D., Klepikova, A. V., **Zharikova, A. A.**, Mironov, A. A., Sheval, E. V., & Vorobjev, I. A. (2018). FACS Isolation of Viable Cells in Different Cell Cycle Stages from Asynchronous Culture for RNA Sequencing. Methods in molecular biology (Clifton, N.J.), 1745, 315–335. [IF = 1.7] (1,3 / 0,2)
3. 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. (2020). Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. Nucleic acids research, 48(12), 6699–6714. [IF = 16.971] (1 / 0,3)
4. Ryabykh, G. K., Mylarshchikov, D. E., Kuznetsov, S. V., Sigorskikh, A. I., Ponomareva, T. Y., **Zharikova, A. A.**, & Mironov, A. A. (2022). Молекулярная биология, 56(2), 275–295 [1.678] (1,3 / 0,25)

Апробация результатов

Полученные результаты были представлены на заседании ученого совета факультета биоинженерии и биоинформатики МГУ им. М.В. Ломоносова 15 ноября 2021 года и обсуждены на конференциях: MCCMB - 2021 в Москве, Россия; “Ломоносов - 2020” в Москве, Россия; ИТиС - 2018 в Казани, Россия; FEBS Congress - 2018 в Праге, Чехия; ИТиС - 2017 в Уфе, Россия.

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

Структура диссертации

Работа состоит из введения, обзора литературы, описания материалов и методов, результатов и их обсуждения, заключения, выводов, списка публикаций и списка цитируемой литературы, содержащего 104 ссылки. Работа изложена на 128 страницах текста, содержит 9 таблиц и 58 рисунков.

ОБЗОР ЛИТЕРАТУРЫ

В настоящее время исследования, касающиеся изучения нуклеиновых кислот, редко обходятся без секвенирования. Арсенал молекулярно-биологических протоколов огромен и непрестанно пополняется. Также стремительно развиваются и биоинформатические алгоритмы, подходы и протоколы, разрабатываемые специально для учета особенностей тех или иных экспериментальных данных.

Нуклеиновые кислоты выполняют свои функции в клетке, вступая в многочисленные взаимодействия в том числе друг с другом и с белками. Существуют различные подходы, позволяющие установить такие взаимодействия. Методы, основанные на иммунопреципитации, позволяют для индивидуальных белков определить места их связывания с хроматином (ChIP-seq) [28] или обнаружить РНК, с которыми эти белки могут взаимодействовать (eCLIP) [29]. С помощью таких протоколов как RIC-seq [30] и MARIO [31] можно получить информацию о РНК-РНК интерактоме.

В 2002 году впервые был опубликован метод 3C, который положил начало группе протоколов, позволяющих изучать пространственную организацию хроматина в клетках [32]. Сегодня эксперименты по изучению организации хроматина в ядре в тандеме с соответствующими биоинформатическими протоколами обработки данных позволяют наблюдать жизнь хроматина в гораздо более подробном разрешении, различая даже особенности его пространственной организации в единичных клетках [33]. Так, примерно за 20 лет подходы к изучению структуры хроматина шагнули далеко вперед.

Внутри ядра царит не только хроматин. Ядерные хромосомы транскрибируют гораздо больше РНК, чем необходимо только для синтеза белка [34]. Большинство этих РНК в принципе не способны к трансляции и относятся к огромной гетерогенной группе некодирующих РНК (нкРНК). Многие некодирующие РНК выполняют свои регуляторные функции в ядре, вступая во взаимодействие с хроматином, действуя *in cis*, т.е. в непосредственной близости от своего гена или на далеких расстояниях *in trans* [2,35,36].

Еще в 60-х годах прошлого века с помощью биохимических методов было установлено, что с хроматином (в том числе непосредственно с гистонами) связана внушительная фракция РНК [37–40]. Однако, что это за РНК, сколько их, к какому классу они принадлежат и какие функции выполняют было неизвестно. Чуть позже, в 80-х годах, было показано, что большое количество РНК связано с ядерным матриксом. Более того, ингибирование транскрипции с помощью, например, актиномицина Д вызывает крупномасштабные изменения в морфологии ядра, включая агрегацию хроматиновых белков [41]. Далее в основном с помощью молекулярно-биохимических методов изучали механизмы функционирования таких некодирующих РНК как XIST [42], HOTAIR [43] и др.

Сегодня мы уже знаем достаточно много примеров РНК, которые действительно выполняют свои функции в ядре, принимая на себя регуляторные, архитектурные и прочие функции [44–47]. Таким образом исследование структуры хроматина, образования ядерных телец и механизмов управления экспрессией неотрывно связано с изучением хроматин-ассоциированных РНК.

Полногеномные способы обнаружения РНК-ДНК взаимодействий

Для того, чтобы понять, как именно РНК выполняет свои функции в связке с хроматином, необходимо прежде всего установить, с какими локусами ДНК она взаимодействует. Существует более десятка подходов, позволяющих решить поставленную задачу, используя современные технологии высокопроизводительного секвенирования. Эти методы можно условно разделить на две группы, различающиеся производительностью и разрешением.

Методы “один-против-всех”

К наиболее широко используемым методам группы “один-против-всех” относятся такие подходы как RAP, CHART-seq, ChIRP-seq, ChOP-seq. В рамках одного эксперимента они позволяют исследовать одну конкретную РНК на предмет ее взаимодействия с хроматином. С точки зрения эксперимента представленные протоколы похожи между собой. С помощью одного из фиксирующих агентов (формальдегида, ультрафиолета и др.) сшивают макромолекулярные комплексы, после чего хроматин фрагментируют ультразвуком или ферментативно. Далее из полученной смеси необходимо выделить только такие комплексы, которые содержат исследуемую хроматин-ассоциированную РНК или ее фрагмент. Для этого используют заранее синтезированные биотинилированные олигонуклеотиды, комплементарные к целевой РНК. Геномную ДНК элюируют в присутствии РНКазы H, освобождая ее от РНК, а обработка протеиназой К помогает избавиться от белков. Полученные фрагменты ДНК являются именно теми локусами, с которыми взаимодействует исследуемая хаРНК. Последовательность этих локусов определяют с помощью секвенирования.

Основная проблема подходов “один-против-всех” заключается в подборе комплементарных олигонуклеотидов таким образом, чтобы избежать пространственных затруднений. РНК имеет вторичную структуру, взаимодействует с ДНК и белками, в результате чего участки РНК, доступные для связывания, ограничены. В методе ChIRP-seq [48] используют короткие ДНК-зонды (~25 нуклеотидов), которые специфически и без пересечений выстилают бОльшую часть целевой РНК. Авторы протокола CHART-seq [47] выбирают такие олигонуклеотиды, которые наилучшим образом связываются с хаРНК по итогам исследования на предмет чувствительности комплекса олигонуклеотид-хаРНК к РНКазе H, определяя количество связавшейся хаРНК с помощью кПЦР для каждого олигонуклеотида. В методе RAP [49] используют более длинные олигонуклеотиды (~120 нукл), которые покрывают всю хаРНК с перекрытием.

Как любой полногеномный эксперимент методы “один-против-всех” могут детектировать неспецифические взаимодействия. Для повышения специфичности на уровне эксперимента авторы предлагают разные подходы. В методе ChIRP-seq все подобранные олигонуклеотиды условно разделяют на четные и нечетные, полностью проводят все экспериментальные процедуры независимо для двух наборов зондов, а затем сравнивают полученные результаты, отбирая для работы только сигналы, подтвержденные дважды. В CHART-seq с помощью дополнительного эксперимента с неспецифическими зондами получают сигнал потенциального шума, который учитывают в дальнейшей обработке.

Несомненным плюсом вышеописанных подходов является высокое разрешение, которое достигается подбором специфических олигонуклеотидов к исследуемой РНК, особенностями протоколов, повышающими специфичность, и дополнительными контрольными экспериментами.

С точки зрения биоинформатической обработки результатов секвенирования протоколы группы “один-против-всех” также довольно схожи. Анализ состоит из стандартных шагов, включающих исследование качества полученных чтений, картирование на референсный геном, поиск и удаление ПЦР-дубликатов. Основным результатом анализа является определение локусов ДНК, которые значимо обогащены контактами хаРНК. Для поиска этих участков в основном используют стандартные программы поиска пиков (MACS2 [50], HOMER [51] и др.), используя информацию о фоновых взаимодействиях. Стоит отметить, что исходно данные программы были разработаны для анализа данных метода ChIP-seq (ДНК-белковые взаимодействия).

Примеры хроматин-ассоциированных РНК

Протоколы группы “один-против-всех” появились более 10 лет назад, с их помощью изучены несколько десятков разных хаРНК в клеточных линиях дрозофилы, мыши и человека [19,52]. Рассмотрим несколько примеров хаРНК, в изучении которых были применены методы “один-против-всех”.

У видов, где определение пола происходит с помощью половых хромосом, часто встает проблема разного уровня экспрессии генов, связанных с полом у самцов и у самок. С помощью эпигенетических механизмов можно скомпенсировать экспрессию таких генов. В разрешении этой проблемы на примере *D. melanogaster* и млекопитающих непосредственное участие принимают хаРНК [53,54]. У самок млекопитающих одна из копий Х-хромосомы инактивирована, находится в виде гетерохроматизированной структуры, именуемой тельце Барра, и теряет практически всю транскрипционную активность. Сразу стоит отметить, что некоторые гены на подавленной Х-хромосоме все же остаются активными. Например, длинная нкРНК Firre, ген которой локализован на Х-хромосоме и избегает инактивации, является хаРНК, контактирует со своей хромосомой в радиусе 5Мб, а также с небольшим количеством локусов на других хромосомах [55]. Вероятно, Firre работает как фактор, способствующий сближению некоторых геномных локусов. Вернемся к процессу инактивации Х-хромосомы у самок млекопитающих. Инициирует и играет ключевую роль в процессе подавления Х-хромосомы длинная нкРНК XIST [56], транскрипты которой распространяются вдоль хромосомы, подлежащей инактивации, привлекая многие белковые факторы, в том числе факторы ремоделирования хроматина [56–58]. В результате инактивированная копия половой хромосомы оказывается прижата к ядерной ламине, разрушается структура ТАДов, характерных для активной копии [58,59]. С помощью методов “один-против-всех” (RAP и CHART-seq) был исследован механизм действия XIST [49,60]. Было изучено несколько временных точек на клеточных линиях, соответствующих периодам до начала инактивации Х-хромосомы (мышиные эмбриональные стволовые клетки), после окончания инактивации (фибробласты), ряд промежуточных состояний. Показано, что при инициации процесса инактивации Х-хромосомы XIST устремляется сначала к активно экспрессирующимся локусам, а затем распространяется на другие участки, причем процесс модерируется конформацией самой хромосомы. При поддержании Х-хромосомы в неактивном

состоянии XIST взаимодействует с ДНК без какого-либо предпочтения по уровню экспрессии.

Дозовая компенсация Х-хромосомы у *D.melanogaster* реализована по обратному принципу. Две нкРНК roX1 и roX2 входят в состав рибонуклеопротеинового комплекса MSL (male-specific lethal), который способствует увеличению транскрипции на мужской Х-хромосоме [61]. Полногеномные карты РНК-ДНК взаимодействий для этих нкРНК показали схожий профиль их контактов с хроматином, а также корреляцию с сайтами связывания белка MSL3, который также входит в состав комплекса MSL [48,62–64].

Еще один метод класса “один-против-всех” - CHIRT-seq - помог разобраться в локализации популяции теломерных нкРНК - TERRA. Было показано, что эти хаРНК взаимодействуют с хроматином не только в теломерных областях, где расположены кодирующие их гены, но и *in trans*, принимая участие в том числе и в процессе инактивации Х-хромосомы [65–67].

хаРНК могут не только подавлять, но и активировать. нкРНК HOTTIP активирует некоторые гомеобоксные гены HoxA, находящиеся в ее близком окружении, привлекая комплекс WDR5, который в свою очередь способствует возникновению активирующей гистоновой метки H3K4me3 [68]. Экспрессию гомеобоксных генов регулируют и другие хаРНК. нкРНК HOTAIR и Haunt действуют похожим образом, привлекая PRC2 (ингибиторный комплекс) и некоторые другие белковые комплексы, которые в свою очередь так изменяют эпигенетические метки, что экспрессия таргетных генов снижается [43,69,70].

В клеточном ядре существуют разнообразные тельца - субкомпартменты, не окруженные мембранами, но морфологически различимые. хаРНК играют важную роль в обеспечении жизнедеятельности многих таких телец, являясь для некоторых из них каркасной основой.

Длинная нкРНК MALAT1 локализована в ядерных спеклах - тельцах, аккумулирующих факторы сплайсинга и процессинга РНК. MALAT1 может

взаимодействовать с этими белковыми факторами и принимать участие в регуляции экспрессии генов и альтернативного сплайсинга [71,72]. С помощью метода RAP показано, что MALAT1 взаимодействует с хроматином в активно экспрессирующихся участках, в основном контактируя с телами генов, избегая одноэкзонные гены и гены гистонов [73].

Другая длинная нкРНК - NEAT1 - ведет себя практически аналогично MALAT1, взаимодействуя полногеномно с активным хроматином, однако предпочитая сайты инициации и терминации транскрипции [74]. NEAT1 является в полной мере архитектурной РНК, т.к. на ней организуются чувствительные к изменению метаболической активности клетки ядерные тельца - параспеклы [75–77].

Методы “все-против-всех”

Основное ограничение описанных выше протоколов заключается в том, что в рамках одного эксперимента можно детектировать локусы взаимодействия с хроматином только одной и, что самое главное, заранее известной РНК. Разработка и введение в лабораторную практику методов, позволяющих определять полный РНК-ДНК интерактом, значительно бы продвинули исследования в области изучения функций длинных и коротких некодирующих РНК. Начиная с 2017 года были предложены сразу шесть методов, закрывающих этот пробел (рис. 1, табл. 1), которые получили название “все-против-всех”.

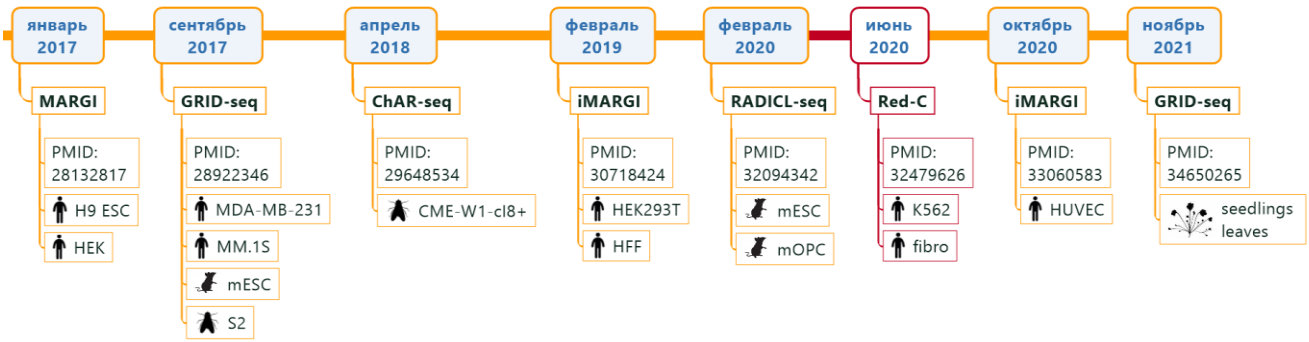

Рисунок 1. Временная шкала публикации работ, посвященных исследованиям РНК-ДНК интерактома с указанием организма и клеточной линии.

Таблица 1. Описание клеточных типов и линий, используемых в публикациях с методами “все-против-всех”.

Организм	Клетки	Описание
MARGI		
Человек	H9 ESC	эмбриональные стволовые клетки
Человек	HEK	эмбриональные почки
GRID-seq		
Человек	MDA-MB-231	клетки рака груди
Человек	MM.1S	множественная миелома
Мышь	mESC	эмбриональные стволовые клетки
Муха	S2	клетки Шнайдера[1]
Арабидопсис	seedlings leaves	саженцы и листья
ChAR		
Муха	CME-W1-cI8+	имагинальный диск; крыло
iMARGI		
Человек	HEK293T	эмбриональные почки
Человек	HFF	фибробласты крайней плоти
Человек	HUVEC	эндотелиальные клетки пупочной вены
RADICL-seq		
Мышь	mESC	эмбриональные стволовые клетки
Мышь	mOPC	клетки-предшественники олигодендроцитов[2]
Red-C		
Человек	K562	клеточная линия хронического миелолейкоза
Человек	fibro	нормальные фибробласты кожи

[1] первичная культура поздних эмбриональных стадий

[2] (женские 46XX) любезно предоставленные доктором М. Лагарковой (ФНКЦ ФХМ, Россия)

Все методы идеологически используют один и тот же основной подход - сначала клетки фиксируют *in vivo*, фрагментируют нуклеиновые кислоты с помощью ферментов, а затем проводят лигирование близко расположенных в пространстве молекул РНК и локусов ДНК при участии специфически сконструированного линкера (или бриджа), несущего биотиновую метку, необходимую для последующего выделения целевых конструктов. Линкер последовательно лигируют сначала к РНК, а затем к ДНК. Во всех публикациях большое внимание уделяют контрольным экспериментам, позволяющим убедиться, что линкер в результате манипуляций в нужный момент специфически

присоединяет РНК и ДНК, не допуская ДНК-ДНК и РНК-РНК сшивок. Дополнительно в некоторых протоколах проводят эксперименты по оценке неспецифических взаимодействий, добавляя к выбранным для исследования клеткам клетки другого организма (например, дрозофилы), после чего подсчитывают долю обнаруженных межвидовых контактов. Специфичность лигирований обеспечивают конструкция линкера, а также заранее особым образом модифицированные концы РНК и ДНК молекул, находящихся в комплексе. В результате получают большое количество химерных молекул (в общем виде: РНК-линкер-ДНК), нуклеотидную последовательность которых расшифровывают с помощью применения технологий высокопроизводительного секвенирования.

Биоинформатическая обработка результатов секвенирования также состоит из нескольких ключевых этапов. Сначала необходимо найти только такие прочтения, которые содержат последовательность линкера, т.к. только они могут быть сиквенсами целевых химерных молекул. Далее выделяют РНК и ДНК части контактов нужной длины и картируют их на последовательность референсного генома.

Стоит отметить, что во всех публикациях после картирования работают только с уникально картированными чтениями. Известно, что геномы высших эукариот содержат огромное количество повторяющихся последовательностей. Используя только уникальное картирование, невозможно изучить поведение хаРНК, пришедших из повторов, или хаРНК, контактирующих с повторяющимися областями последовательности генома, что является существенным ограничением в представленных результатах методов “все-против-всех”. Разработка подходов, позволяющих корректно обрабатывать результаты множественного картирования, является перспективным направлением в изучении хаРНК.

Далее РНК-части аннотируют, используя какую-либо из существующих разметок генов нужного организма. Локусы ДНК также могут быть проаннотированы разнообразными разметками: генами, промоторами и энхансерами, сайтами связывания транскрипционных факторов, эпигенетическими

метками, транскрипционно активными доменами и пр. На основании полученных аннотированных контактов можно делать выводы о предпочтении взаимодействия конкретной РНК или группы РНК с специфическими участками хроматина, исследовать характер взаимодействия РНК с хроматином. Для подтверждения корректности метода сравнивают профили контактов известных хаРНК, полученные из эксперимента “все-против-всех” с аналогичными данными из экспериментов “один-против-всех”.

Далее разберем экспериментальные особенности каждого опубликованного метода “все-против-всех”, подходы в анализе результатов секвенирования, основные выводы и наблюдения.

MARGI

Метод MARGI (mapping RNA-genome interactions) [20] опубликован в феврале 2017 года и представляет собой первую технологию, с помощью которой можно выявить все РНК-хроматиновые взаимодействия в интактных клетках в рамках одного эксперимента (рис. 2).

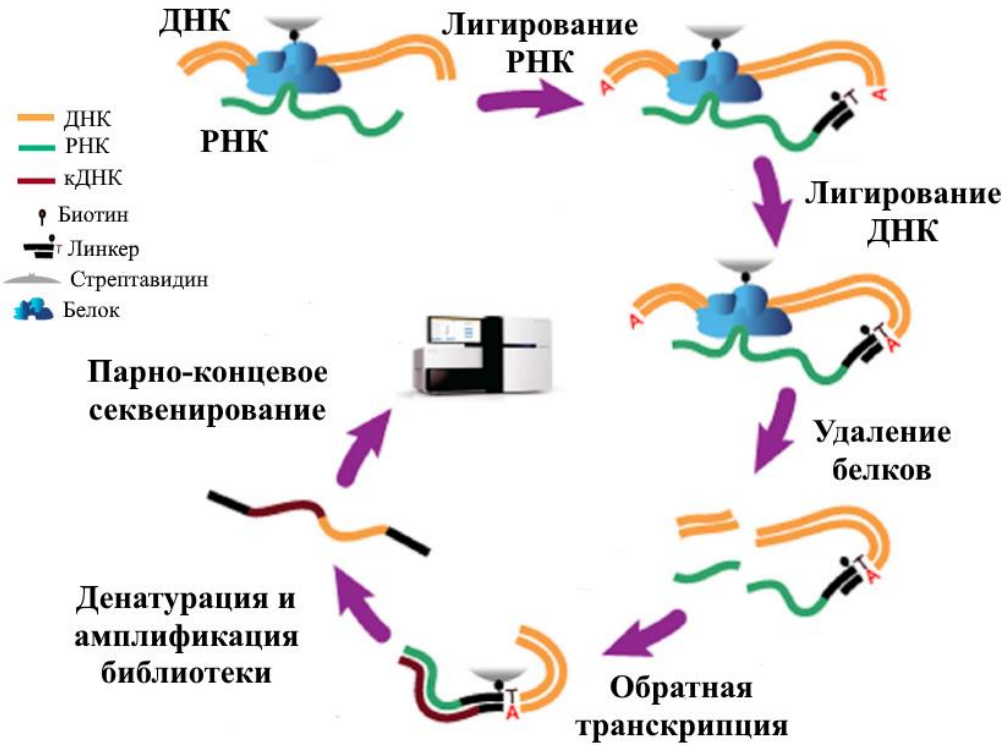

Рисунок 2. Схема эксперимента MARGI. Адаптировано из [20].

Авторам удалось разработать два подхода: pxMARGI (proximity MARGI) - дает возможность установить все взаимодействия, нельзя отличить случайные неспецифические контакты РНК с хроматином от функциональных; diMARGI (direct MARGI) - разработан с целью увидеть белок- или РНК-опосредованные контакты с хроматином. Авторы отмечают, что протокол pxMARGI способен детектировать контакты РНК с гетерохроматином, тогда как через призму diMARGI можно говорить в основном о функциональных контактах с открытым хроматином.

Далее описаны основные шаги пробоподготовки, необходимые для понимания работы метода.

Согласно протоколу клетки сначала фиксируют, обрабатывая 1% формальдегидом (в случае pxMARGI) или дисукцинимидил глутаратом (DSG) и 3% формальдегидом (в случае diMARGI). Формальдегид обратимо сшивает нуклеиновые кислоты и белки на расстоянии ~2 ангстрема [78], а DSG необратимо сшивает белки на расстоянии ~7 ангстрем [79]. Фиксация позволяет прочно сшить присутствующие в клетке макромолекулярные комплексы. Клеточные мембраны разрушают, выделяют ядра, а затем и содержимое ядер. Хроматин фрагментировали ферментативно с помощью рестриктазы HaeIII в эксперименте pxMARGI и с помощью ультразвука в эксперименте diMARGI, отбирая для работы только растворимую фракцию. Дополнительно в протоколе diMARGI были удалены никак не связанные с хроматином или случайно прилипшие РНК до стадий лигирования. Белки, в том числе опосредующие РНК-ДНК взаимодействия, биотинилируют. Данный шаг позволит далее зафиксировать комплексы с белками на гранулах со стрептавидином, что значительно снизит случаи детекции ложных неспецифических контактов и упростит работу с материалом. После процедуры биотинилирования белков молекулы ДНК подвергают ферментации с помощью рестриктазы HaeIII, формирующей “тупые” концы, комплексы фиксируют на гранулах со стрептавидином.

Далее необходимо создать химерную конструкцию для последующего секвенирования. Ключом к реализации этой идеи является лигирование к взаимодействующим нуклеиновым кислотам особым образом сконструированного линкера (рис. 3): основную длину линкера составляет центральная двухцепочечная ДНК, несущая сайт узнавания рестриктазы BamHI и биотиновую метку, а также последовательности, комплементарные праймерам для последующего секвенирования, по бокам выступает одноцепочечная ДНК разной длины. Линкер лигируют сначала к 3`-ОН РНК-части комплекса с помощью мутированной Т4 РНК-лигазы 2, которая не нуждается в АТФ, что предотвращает РНК-РНК сшивки и другие нежелательные реакции. Выступающий одноцепочечный 5`-конец линкера несет 5`App, а первые два нуклеотида рандомные, что обеспечивает эффективное лигирование любой последовательности 3`-конца фрагмента РНК. Используемая лигаза (Т4 РНК-лигаза 2) специфически соединяет 3`-ОН одноцепочечной РНК с 5`App одноцепочечной РНК или ДНК. Далее с другой стороны линкера присоединяют фрагмент ДНК. Оставшийся свободным конец линкера двухцепочечный за исключением выступающего 3`-Т. Перед лигированием фрагменты ДНК из РНК-ДНК комплекса дополнительно обрабатывают, довешивая 3`-А. Таким образом возможно осуществить лигирование линкера и ДНК за “липкие” концы.

В результате получают химерные молекулы РНК:линкер:ДНК, которые выделяют за биотиновую метку и проводят реакцию обратной транскрипции, достраивая недостающую цепь в области РНК.

После денатурации получают одноцепочечный фрагмент ДНК, который закольцовывают с помощью специальной лигазы circligase, что является отличительной особенностью технологии MARGI. Рестриктазой BamHI получившиеся кольца разрезают (сайт узнавания для рестриктазы находится примерно в середине последовательности линкера (рис. 3)), получая целевую химерную конструкцию для парно-концевого секвенирования. С помощью описанных манипуляций стало возможно точно позиционировать и секвенировать

РНК-части и фрагменты ДНК с разных праймеров для секвенирования, получая Read1 (РНК-часть) и Read2 (ДНК-часть). Секвенирование проводили на приборе компании Illumina в парно-концевом режиме по 100 циклов с каждого конца. В результате получали достаточно длинные чтения как для РНК-части, так и для локуса ДНК, что является несомненным преимуществом метода MARGI.

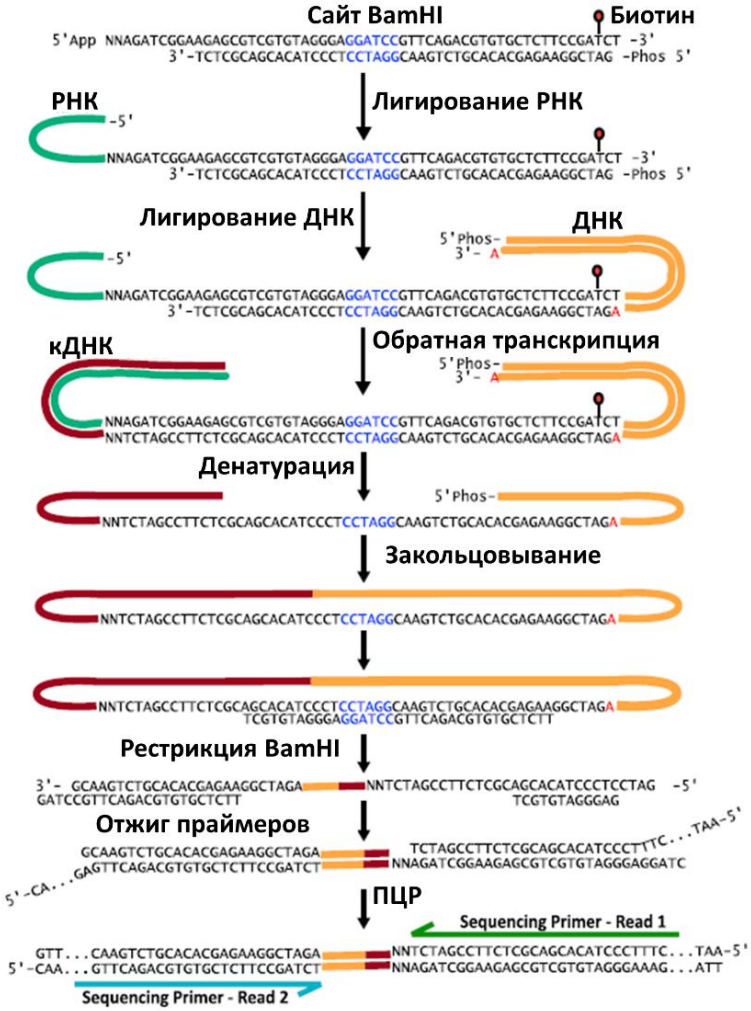

Рисунок 3. Схема эксперимента MARGI. Этапы манипуляции с линкером. Последовательность линкера представлена в явном виде. Адаптировано из [20].

Подготовку библиотек для секвенирования предваряет огромное количество манипуляций, которые должны проходить строго задуманным образом, чтобы полученные в результате секвенирования данные имели смысл и были пригодны для дальнейшего анализа. Необходимо убедиться, что ожидаемый тип нуклеиновой

кислоты присоединяется к намеченному концу ликнера, а также корректно срабатывают все этапы процесса закольцовывания. В противном случае было бы невозможно с уверенностью различить фрагменты РНК и ДНК. Для осуществления такого контроля был предложен следующий эксперимент (контроль на полярность линкера). Подготовили две одинаковые библиотеки с разным способом фрагментации ДНК: ультразвуком (режет в практически случайных местах) и рестриктазой HaeIII (оставляет “CC” на 5`-конце). В случае корректной работы метода ожидали увидеть обогащение “CC” только в начале ДНК-чтений (Read2) для библиотеки с обработкой HaeIII, тогда как все другие варианты чтений должны были начинаться с случайных нуклеотидов, что и было показано.

Для контроля специфичности и корректности лигирования было реализовано два эксперимента. Подготовлено три библиотеки: весь протокол выполнен полностью (1), отсутствует этап лигирования линкера к РНК (2), отсутствует этап лигирования линкера к ДНК (3). В результате библиотеки 2 и 3 не дали итогового продукта. Также был проведен эксперимент для оценки уровня неспецифического лигирования - к культуре исследуемых клеток добавили клетки дрозофилы, провели полностью все шаги экспериментального протокола и подсчитали процент межвидовых РНК-ДНК контактов. Оказалось, что таких контактов регистрируется всего 2,18%, что говорит о низком уровне случайного лигирования.

В протоколе MARGI были использованы две клеточные линии человека (HEK293T и H9 ESC) и ~ 400 млн клеток на образец.

Биоинформатическая обработка результатов секвенирования состояла из нескольких этапов: удаление дублированных чтений с помощью программы FastUniq [80], удаление технических последовательностей, картирование чтений (независимо РНК и ДНК части) на референсный геном человека (версия hg38) с помощью программы STAR, при картировании РНК-частей допускали возможность картирования с разрывом. В качестве источника генной разметки использовали аннотацию Ensemble (release 84).

После биоинформатической обработки результатов получившеся контакты были разделены на три группы: проксимальные или близкие (РНК и ДНК фрагменты одного контакта картированы на одну и ту же хромосому не далее 2 Кб друг от друга), дистальные или дальние (РНК и ДНК фрагменты одного контакта картированы на одну и ту же хромосому на расстоянии более 2 Кб друг от друга) и межхромосомные (РНК и ДНК фрагменты одного контакта картированы на разные хромосомы). В обеих клеточных линиях в случае эксперимента pxMARGI ~ 80% контактов оказались близкими, менее 4% попали в группу дальних контактов и 15-20% межхромосомных контактов. В эксперименте diMARGI подавляющее большинство контактов (~ 96%) оказались близкими, количество дистальных контактов сократилось до ~ 1%, остальные контакты попали в группу межхромосомных (менее 5%). Далее работали только с далекими и межхромосомными контактами.

На основании пуассоновской статистической модели отбирали такие хаРНК, которые были покрыты РНК-частями больше, чем ожидалось по случайным причинам, предполагая, что количество РНК-частей пропорционально длине гена, далее отбирали значимые находки с FDR < 0.0001 . Для протокола pxMARGI удалось идентифицировать 2864 (HEK) и 1933 (ESC) нк РНК, подавляющее большинство которых в обеих клеточных линиях были классифицированы как псевдогены, антисенс РНК или длинные межгенные некодирующие РНК. В случае diMARGI обнаружено 747 (HEK) и 467 (ESC) некодирующих ассоциированных с хроматином РНК, в обеих клеточных линиях находки были обогащены малыми ядрышковыми РНК.

Сравнение результатов протоколов pxMARGI и diMARGI, реализованных в двух разных клеточных линиях, позволило установить для ряда хаРНК характер взаимодействия с хроматином (случайный или функциональный) и изучить его специфичность относительно клеточной линии. Наиболее интересным наблюдением оказалось поведение длинной некодирующей РНК XIST, которая была идентифицирована, как высоко контактирующая РНК в pxMARGI (обе

клеточные линии) и только в HEK293T в случае diMARGI. Таким образом в эмбриональных стволовых клетках XIST предположительно взаимодействует исключительно с конденсированным хроматином (протокол diMARGI сильно хуже детектирует такие контакты), что согласуется с ее функциями, тогда как в клетках HEK293T XIST может прикрепляться к более доступным участкам хроматина в том числе через белковые мостики. Интенсивность контактов с хроматином двух ассоциированных с XIST РНК - TSIX и FTX - согласуется с таковой для XIST по всем протоколам и клеточным линиям.

Для того, чтобы изучить участки на хроматине, с которыми контактируют хаРНК, авторы воспользовались программой MACS2, которая позволяет выделить области, обогащенные выровненными прочтениями (пики), применив MACS2 к ДНК-частям контактов. Программа разработана специально для обработки данных ChIP-seq, валидность применения MACS2 к данным РНК-ДНК контактов авторами статьи не обсуждается. В результате для протокола pxMARGI было идентифицировано более чем в 10 раз больше пиков в обеих клеточных линиях, чем в diMARGI, также пики piMARGI были шире. При сравнении находок между клеточными линиями вне зависимости от протокола в HEK293T пиков было больше, чем в стволовых клетках, где хроматин в бОльшей степени гетерохроматизирован. Все это говорит в пользу утверждения, что pxMARGI (но не diMARGI) позволяет детектировать контакты РНК с конденсированным хроматином.

Также показано, что в обеих клеточных линиях для обоих протоколов пики контактов хаРНК обогащены в промоторных областях. В 20Кб окружении стартов транскрипции контакты хаРНК положительно коррелируют с данными ChIP-seq на метки H3K4me3 и H3K27ac и отрицательно коррелируют с меткой H3K9me3. При разбиении всего генома на участки длиной 1000 нуклеотидов контакты хаРНК из diMARGI сохраняют вышеописанную корреляцию с гистоновыми метками, а контакты хаРНК из протокола pxMARGI отрицательно коррелируют с меткой H3K9me3. Метка H3K9me3 отсутствует в принципе в близком окружении от всех

пиков эксперимента MARGI. Авторы объясняют это тем, что в случае закрытого хроматина РНК не связываются с участками, несущими метку H3K9me3.

GRID-seq

Следующим в группе “все-против-всех” был опубликован протокол GRID-seq (Global RNA Interactions with DNA by deep sequencing) [21]. В данном методе авторы предлагают более глубокий анализ, а также значительное разнообразие данных. Протокол был реализован на клеточных линиях сразу трех организмов: человек (MDA-MB-231 и MM.1S), мышь (mESC) и дрозофила (S2).

Подготовка образцов включала практически все описанные в протоколе MARGI шаги.

Клетки фиксировали DSG и 3% формальдегидом, выделяли ядра и фрагментировали ДНК с помощью частощепящей рестриктазы AluI. Для формирования химерных молекул, содержащих фрагменты гипотетически контактирующих РНК и ДНК, также использовали полярно сконструированный линкер, преаденилированный с 5`-конца, к которому *in situ* лигировалась одноцепочечная РНК из зафиксированного комплекса. Далее с помощью обратной транскриптазы на матрице РНК синтезируется кДНК, после чего удаляют линкеры без РНК, затем к свободной стороне линкера лигируют фрагмент ДНК из комплекса. Получившийся конструкт выделяют на стрептавидиновых шариках за счет расположенной на линкере биотиновой метке.

Важным отличием технологии GRID-seq от MARGI является заключительная стадия формирования химерной молекулы перед секвенированием. В GRIG-seq отсутствует этап образования кольца, после всех манипуляций образуется конструкт РНК:линкер:ДНК. На концах линкера с обоих сторон находится сайт узнавания рестриктазы MmeI, которая режет на ~20 нуклеотидов в сторону. Таким образом библиотеки для секвенирования представляют собой линкер, к которому с двух концов пришиты по ~20 нуклеотидов, которые соответствуют фрагменту РНК и локусу ДНК, с которым эта РНК гипотетически взаимодействует. Сразу стоит отметить, что в GRID-seq чтения в 5 раз короче, чем в протоколе MARGI, что

можно отнести к недостаткам метода GRID-seq. Итоговый целевой фрагмент имеет длину ~85 нуклеотидов, для его секвенирования достаточно реализовать одно-концевой протокол. Был проведен эксперимент, доказывающий корректность лигирования ожидаемой нуклеиновой кислоты к запланированному концу линкера, аналогичный протоколу MARGI, но с применением рестриктазы AluI.

Авторы отмечают, что РНК-части сохраняли информацию об ориентации цепи, а фрагменты ДНК теряли эту информацию. Все эксперименты были сделаны в двух репликах, в каждом случае реплики хорошо коррелировали между собой по количеству чтений на каждую РНК.

Анализ распределения картирования РНК и ДНК-частей контактов на генную архитектуру показал, что фрагменты РНК в подавляющем большинстве приходят из аннотированных участков (в основном из экзонов генов), а локусы ДНК - из промоторных и межгенных областей. Оказалось, что интенсивность контакта с хроматином для индивидуальной РНК лучше коррелируют с уровнем экспрессии зарождающихся транскриптов (определены с помощью эксперимента GRO-seq), чем с уровнем экспрессии, измеренным с помощью эксперимента RNA-seq. Авторы отмечают, что, по-видимому, протокол GRID-seq детектирует в основном контакты зарождающихся транскриптов с хроматином.

Для еще одной проверки корректности работы метода сравнили профили контактов одной из самых высоко контактирующих в GRID-seq длинных некодирующих РНК MALAT1 в двух экспериментах: GRID-seq (“все-против-всех”) и RAP (“один-против-всех”) на одинаковой линии клеток (мышиные эмбриональные стволовые клетки). Было показано, что профили контактов прекрасно совпадают вдоль всего генома. Однако, согласно эксперименту GRID-seq, РНК MALAT1 предпочитала взаимодействовать с сайтами старта транскрипции, в то время как по результатам нескольких экспериментов “один-против-всех” было показано, что MALAT1 взаимодействует с телами генов. Скорее всего это отличие появляется по техническим причинам из-за особенностей метода GRID-seq. Тем не менее можно утверждать, что GRID-seq детектирует

биологически осмысленные взаимодействия, т.к. в рамках одной ткани для MALAT1 GRID-seq и RAP-DNA определяют практически одинаковый набор таргетных генов (пересечение ~ 90%). Причем, если сравнивать GRID-seq и CHART (“один-против-всех”) для MALAT1, но для разных клеточных линий, то пересечение таргетных генов оказывается минимальным (~ 10%).

Аналогично MARGI авторами GRID-seq был поставлен межвидовой контроль на специфические взаимодействия (к клеткам человека линии MDA-MB-231 добавили клетки мухи (*D. melanogaster*) S2) и обнаружено ~ 8,5 % межвидовых контактов. Далее такие контакты использовали для построения трека экзогенных фоновых неспецифических РНК-ДНК взаимодействий. Дополнительно авторы предложили способ конструирования эндогенного трека неспецифических взаимодействий: все контакты мРНК с “не материнскими” хромосомами. Было показано, что области, куда попадают экзогенные и эндогенные фоновые неспецифические взаимодействия, хорошо коррелируют между собой внутри одной ткани. Также профили фоновых контактов были довольно схожи между разными тканями одного организма (показано на примере человека). Таким образом, для выявления особо “липких” участков хроматина не требуется осуществлять эксперимент по обнаружению межвидовых контактов, что значительно облегчает экспериментальную работу. Авторы предложили подход, позволяющий для каждой РНК определять значимые локусы взаимодействия с хроматином (пики), подсчитывая отношения сигнала конкретной РНК к сигналу эндогенного фона в заранее определенных интервалах длиной 1Кб. Пик считался значимым, если для интервала в 10Кб было обнаружено как минимум 3 интервала в 1Кб, где уровень сигнала конкретной РНК превышал уровень фона не менее, чем в 2 раза.

Используя полученные знания, было показано, что большинство дальних контактов и примерно половина близких контактов попадает на участки неспецифического связывания. Также неспецифические транс-контакты приходятся в основном на открытый хроматин, коррелируя с треком РНК-

полимеразы II и эпигенетическими метками открытого хроматина (H3K4me1 и H3K27ac).

Процедура, используемая для поиска значимых контактов, включала в себя довольно жесткие этапы фильтрации и в результате можно было видеть только мажорные контакты. Авторы отмечают, что удалось обнаружить менее 1000 хаРНК (преобладают белок-кодирующие (~ 88%)), из которых считанные единицы контактируют *in trans* (MALAT1, NEAT1, U2), контакты всех остальных РНК обнаруживаются в основном в непосредственной близости от своего же гена (~1Мб вокруг старта транскрипции для млекопитающих и ~200Кб для *D.melanogaster*). Это наблюдение воспроизводимо между организмами, тканями и репликами. Тем не менее было обнаружено большое количество хаРНК, профиль контактов которых различался между тканями. Используя дополнительные разметки генома, удалось показать, что контакты хаРНК преобладают в области активных энхансеров, а некоторые хаРНК могут взаимодействовать с активными промоторами и энхансерами на далеких расстояниях ткане-специфичным образом.

На примере длинной некодирующей РНК roX2 (эксперимент на мухе (*D.melanogaster*)) была показана высокая корреляция между тремя типами данных: полногеномные контакты roX2 (GRID-seq), полногеномные контакты roX2 (CHiRP и CHART (методы группы “один-против-всех”)), а также пики ДНК-связывающего белка MSL, который также связывается и с roX2 (ChIP-seq). При сравнении профиля контактов для одной и той же РНК, полученных разными методами, авторы отмечают, что в экспериментах “один-против-всех” заведомо бОльшая глубина покрытия, а ширина локализуемых пиков сильно меньше, чем в эксперименте “все-против-всех” (4,5 Кб для ChIRP и 83 Кб для GRID-seq), т.е. удается более точно позиционировать места контактов. Тем не менее с помощью GRID-seq удалось подтвердить > 90% мест контактов для roX2 пусть и с худшим разрешением.

Интересные наблюдения показали результаты сравнения данных РНК-ДНК взаимодействий с Hi-C картами, отражающими 3D организацию хроматина.

Оказалось, что профили РНК-ДНК и ДНК-ДНК контактов хорошо коррелируют для РНК, принимающих участие в образовании РНК-ДНК интерактома, причем примерно половина контактов РНК приходилась на соседний транскрипционно активный доман, что может свидетельствовать в пользу того, что хаРНК способствуют формированию хромосомных территорий.

В данной работе была предпринята попытка разобраться, как взаимодействуют друг с другом энхансеры и промоторы активных генов. Используя в качестве нулевой модели транс контакты мРНК, удалось установить ~11000 значимых промотор-энхасерных взаимодействий и чуть более 8 000 промотор-промоторных взаимодействий в клетках MM.1S. При анализе полученной информации было показано, что обычные энхансеры в среднем контактировали на более дальние расстояния, чем супер-энхансеры. Оказалось, что каждый промотор может регулировать до четырех других генов. Были установлены промотор-промоторные взаимодействия между хромосомами, чего раньше показано не было. Показано, что хаРНК может взаимодействовать с многими энхансерами, но только с 1-2 супер энхансерами, но энхансер вне зависимости от типа взаимодействует только с 1-2 РНК. Таким образом, экспрессия каждого гена может находиться под контролем многих энхансеров, но каждый конкретный энхансер контролирует очень ограниченный набор генов.

Авторы также отмечают, что в результате обработки было найдено некоторое количество новых неаннотированных хаРНК, но их изучение не входило в опубликованную работу.

ChAR-seq

Метод ChAR-seq (Chromatin-associated RNA sequencing) [22] реализован на клеточной линии мухи *D.melanogaster*, полученной из крылового диска самца с нормальным кариотипом, для этой клеточной линии хорошо охарактеризованы эпигеном и транскриптом.

Для проведения эксперимента отбирают 100-150 млн клеток, которые фиксируют 1% формальдегидом и пермеабилизуют. РНК частично фрагментируют

и удаляют растворимую фракцию, избавляясь таким образом от свободной РНК. Для формирования РНК-ДНК комплексов используют полярный линкер, который лигируют сначала к РНК, которую с помощью обратной транскрипции переводят в кДНК, а затем к ДНК, обработанную рестриктазой DpnII. Последовательность линкера не имеет аналогов в геномах дрожжей, мухи, мыши и человека. После лигирования линкера к нуклеиновым кислотам все обрабатывают ультразвуком для получения фрагментов нужной длины (~200 нуклеотидов). Химерные молекулы выделяют за биотин, пришитый к линкеру. Секвенирование проводили в одноконцевом режиме, чтения длиной 152 нукл.

Как и в предыдущих методах разделение РНК и ДНК частей было возможно благодаря полярности линкера. Выделенные фрагменты РНК были картированы на транскриптом, а участки ДНК - на геном, отбирали только уникально картированные чтения, удаляя регионы повторов и BlackList [81] (участки, содержащие в основном простые повторы; разработан для протокола ChiP-seq). В результате было получено 22.2 млн контактов для 16800 транскриптов. Показано, что добавление РНКазы А и РНКазы Н до лигирования РНК к линкеру сильно снижает количество химерных молекул с линкером, т.е. формирование финальных молекул РНК специфично. В данной работе был сделан интересный контрольный эксперимент для проверки поведения 3D структуры хроматина при проведении протокола. Авторы сымитировали эксперимент Hi-C, обработав ДНК DpnII, после чего биотинилировали получившиеся хвосты, сшили и секвенировали. Таким образом были получены ДНК-ДНК взаимодействия, которые хорошо коррелировали с результатами настоящего Hi-C в тех же клетках, но сильно отличались от РНК-ДНК интерактома. Таким образом показано, что ТАДы сохраняются, а РНК-ДНК контакты не являются ДНК-ДНК взаимодействиями. Полученные результаты по количеству контактов воспроизводимы между репликами (в единицах CPKM). После аннотации и анализа распределения РНК-частей по цепям, авторы указали, что были обнаружены неаннотированные РНК, но их анализ не входил в работу.

По характеру взаимодействия с хроматином РНК были разделены на три группы: контактируют только рядом с местом своей транскрипции (в основном это мРНК); контакты распределены в основном *in trans* практически по всему геному (малые ядерные РНК и ряд нк РНК); нкРНК roX1 и roX2, как часть комплекса компенсации дозы, которые контактируют преимущественно с хромосомой X.

При попытке выявить достоверно контактирующие с хроматином РНК оказалось, что ~88% всех РНК имеют очень маленькое число контактов, зато оставшиеся 12% дают 88% всех контактов; слабо контактирующие РНК были удалены из анализа. Дополнительно были привлечены данные по экспрессии РНК в аналогичных клетках, показана корреляция между нормированными уровнями экспрессии (FPKM) и контактов с хроматином (CPKM). Результаты воспроизводимы между репликами.

Для двух хорошо изученных ассоциированных с хроматином РНК roX1 и roX2 удалось сравнить профили контактов между экспериментами ChAR-seq и ChIRP-seq (“один-против-всех”), показав их высокую корреляцию между собой, что говорит о хорошей специфичности метода даже для индивидуальных РНК. Разрешение метода авторы оценивают ~200 нукл, что соответствует длине фрагментов рестрикции после обработки ДНК рестриктазой DpnII.

Большое количество контактов (~23%) принадлежат классу малых ядерных РНК (мяРНК). Эти малые РНК можно подразделить на два класса: входящие в состав большой сплайсосомы и малой сплайсосомы. Оказалось, что контакты мяРНК, которые входят в состав большой сплайсосомы, кластеризуются по ДНК-частям. Некоторые сплайсосомальные РНК контактировали с телами генов, границами ТАДов, находившимися в открытом хроматине по данным ATAC-seq.

iMARGI

Данный метод представлен авторами протокола MARGI, описанного выше. Отличие iMARGI [23] заключается в том, что этапы фрагментирования хроматина и лигирования были проведены *in situ*, что позволило значительно снизить числе

клеток, необходимых для проведения эксперимента: с 400 млн до 5 млн. Протокол реализован на двух клеточных линиях человека (HEK293T и HFF).

Если в предыдущих подходах авторы отмечали, что в результате было детектировано больше близких контактов, то в случае iMARGI более половины контактов РНК с хроматином были классифицированы как далекие, а для индивидуальных РНК показана отрицательная корреляция частоты контактов с расстоянием от своего гена. Сравнение результатов MARGI и iMARGI показали, что профили контактов iMARGI лучше коррелировали с вариацией pxMARGI, который не различает РНК или белок-опосредованные взаимодействия от неспецифических.

Далее авторы воспользовались данными о гибридных транкриптах из опухолевых клеток, представленных в базе данных TCGA (The Cancer Genome Atlas). Оказалось, что 5 из 10 наиболее часто взаимодействующих пар генов в iMARGI были определены как гибридные транскрипты. Дополнительно была проанализирована экспрессия 96 новых опухолевых образцов легких, в результате чего было обнаружено 42 гибридных транскрипта. Большая часть из них (37 из 42) совпадала с РНК-ДНК контактами нормальных клеток, а при образовании гибридного транскрипта EML4-ALK не происходит геномной перестройки между соответствующими генами. На основании этих результатов авторы предположили, что образование гибридных транскриптов может происходить в результате того, что РНК оказалась в близком окружении от другого гена, и в этот момент может произойти транс-сплайсинг или геномная перестройка.

RADICL-seq

Метод RADICL-seq [24] представлен в 2020 году и реализован на двух клеточных линиях мыши: эмбриональных стволовых клетках (ESC) и клетках-предшественниках олигодендроцитов (OPC). Для клеточной линии ESC эксперимент проведен при различных концентрациях фиксирующего агента (1% и 2% формальдегида), а также при обработке ингибитором РНК-полимеразы II (актиномицином Д). Для обеих клеточных линий реализованы дополнительные

эксперименты, помогающие определить только РНК-ДНК взаимодействия, не опосредованные белками. Для этого перед стадиями лигирования образцы были обработаны протеиназой К в денатурирующих условиях; такие эксперименты обозначены как NPM (non-protein mediated). Показано, что обработка разной концентрацией формальдегида дают схожие результаты. В NPM образцах авторы отмечают резкое снижение транс-контактов. В отличие от предыдущих протоколов, где фрагментацию хроматина осуществляли ферментативно, в RADICL-seq хроматин фрагментировали с помощью неспецифической ДНКазы I. Дополнительно перед стадиями лигирования выделенные ядра обрабатывали РНКазой H для удаления РНК из РНК-ДНК дуплексов. Также отличается рестриктаза, которую используют после лигирования (авторы используют EcoP15I). Картирование как РНК, так и ДНК-частей было реализовано с помощью программы для картирования BWA [82], не допускающего сплайсинг. При аннотации фрагмент контакта рассматривали как точку (использовали центр участка), пересекая непосредственно с элементами генной разметки, учитывая ориентацию РНК-частей контактов. Предложена процедура выделения специфических взаимодействий, при которой геном разбивали на интервалы (или бины) в 25Кб, для каждой РНК подсчитывали число контактов, попавших в каждый бин. Затем для каждого бина проводили правосторонний биномиальный тест, где число испытаний было определено, как общее количество контактов конкретной РНК, а вероятность успеха определяли как величину, обратную количеству бинов, с которыми контактировала исследуемая РНК. Локусами с значимыми контактами считали такие бины, p-value которых было менее 0.05 после коррекции на множественное тестирование.

Показана корреляция уровня контактов РНК с уровнем их экспрессии, причем корреляция была больше при сравнении с экспрессионным профилем ядерной фракции, чем цитоплазматической. В основном контакты были детектированы с эухромтаином. Большое количество специфических контактов приходится на мРНК. При сравнении результатов по двум исследуемым в

протоколе клеточным линиям обнаружены РНК, контактирующие специфическим образом. Некоторые из этих РНК являются соответствующими клеточными маркерами. Отмечено наличие полимеразного следа, т.е. плотность контактов РНК убывала при увеличении расстояния от своего гена.

Были привлечены дополнительные данные по исследованию R-петель (по данным DRIP-seq) и ТАДов (по данным Hi-C). Показано, что РНК, пришедшие из ТАДов, предпочитали контактировать с локусами ДНК, принадлежащими этим же ТАДам, а РНК, гены которых закодированы вне границ ТАДов, преимущественно контактировали с ДНК вне ТАДов. Значимые контакты были обогащены в пиках DRIP-seq. Для двух нкРНК MALAT1 и MEG3 были предсказаны ДНК:ДНК:РНК-триплексы, расположенные недалеко от их ДНК-контактов.

Несмотря на то, что при анализе, как и во всех аналогичных протоколах, были использованы только уникально картированные чтения, удалось захватить некоторые РНК, пришедшие из повторов. Для эмбриональных стволовых клеток были обнаружены малые ядерные РНК и РНК, пришедшие с повторов класса SINE. Показано, что мяРНК контактируют в основном *in trans*, а SINE-РНК - на расстоянии 10Кб-1Мб от своего гена. В клеточной линии OPC обнаружены РНК, пришедшие с повторов классов LINE и LTR, взаимодействующие с хроматином на значительном расстоянии от своего гена (более 100Кб).

Проведено сравнение с протоколом GRID-seq, который был реализован в том числе и на эмбриональных стволовых клетках мыши. В RADICL-seq отмечают большее количество специфических дальних контактов, а также лучшую корреляцию профиля контактов для некоторых РНК с данными “один-против-всех” (показано для нкРНК MALAT1 и RN7SK).

Red-C

Протокол Red-C опубликован в 2020 году [25] и имеет ряд отличий, как в экспериментальной процедуре, так и в биоинформатическом подходе к обработке результатов. В эксперименте были использованы две человеческие клеточные линии: K562 и нормальные фибробласты кожи. Для каждой ткани протокол

реализован в трех репликах, одна из которых отличается малым количеством контактов. Показано, что реплики хорошо коррелируют между собой по количеству контактов для индивидуальных РНК.

Для реализации протокола необходимо ~2.5 млн клеток, в качестве фиксирующего агента был применен 1% формальдегид. Используемый в эксперименте линкер представляет собой конструкцию, которая с одной стороны является двухцепочечной ДНК с сайтом MmeI, а с другой - одноцепочечной ДНК. Предполагается, что к двухцепочечному концу линкера будет лигирован фрагмент ДНК, а рестриктаза MmeI служит для определения длины этого фрагмента, разрезая последовательность ДНК на 20 нуклеотидов левее своего сайта. Соответственно, к одноцепочечному концу линкера будет лигирован фрагмент РНК, длина которого ничем не ограничена. После лигирования к линкеру нуклеиновых кислот при синтезе кДНК используется особая ревертаза, из технологии SMART-seq [83], что является отличительной особенностью экспериментальной части метода Red-C. Данная ревертаза сначала в качестве матрицы использует одноцепочечный фрагмент линкера, затем проходит через границу линкер:РНК и далее достраивает кДНК по матрице РНК. Дойдя до 5`-конца фрагмента РНК ревертаза нематрично достраивает 2-3 цитозина, которые могут комплементарно связаться с последовательность GGG, которая находится на 3`-конце адаптеров для секвенирования. Далее ревертаза продолжает работу, достраивая кДНК на матрице адаптера (рис. 4).

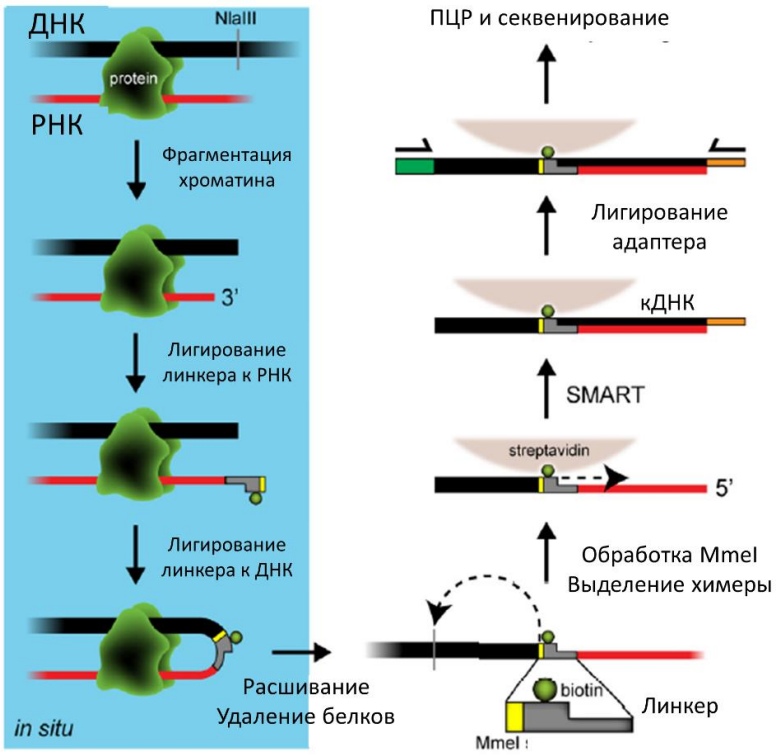

Рисунок 4. Схема эксперимента Red-C. Адаптировано из [25].

Секвенирование проводилось в парно-концевом режиме, причем в случае Red-C химерная конструкция ДНК:линкер:РНК устроена таким образом, что прямое чтение содержит последовательность фрагмента ДНК, затем линкер и далее фрагмент РНК (3`-часть РНК). Обратное же прочтение содержит последовательность РНК, секвенированную с 5`-конца. Если фрагмент РНК длиннее чтения, то никаких технических последовательностей обратное прочтение содержать не будет. В противном случае обратное прочтение будет содержать кроме последовательности РНК еще и последовательность линкера и возможно захватит фрагмент ДНК, лигированный с другого конца линкера. Таким образом в протоколе Red-C репортируют сразу две части РНК.

Картирование РНК и ДНК частей на референсный геном осуществлено с помощью программы HISAT2 [84], допуская сплайсинг в случае РНК. Для аннотации РНК-частей генами была предложена процедура голосования, при которой в случае попадания РНК-части на пересечение генов по одной цепи отдавали предпочтение гену с наибольшей плотностью контактов. Также авторы

провели анализ РНК-частей, которые по итогам первичной аннотации не попали в известную разметку, собрав с помощью кластеризации ~2000 гипотетических новых хаРНК, названных X-РНК. В качестве модели эндогенного фона авторы используют транс-контакты мРНК, аналогично работе GRID-seq, однако при расчете используют бины меньшего размера (500 нуклеотидов вместо 1Кб в GRID-seq). Процедура выявления специфичных контактов относительно эндогенного фона аналогична таковой в GRID-seq. В качестве подтверждения корректности метода было показано, что значимые контакты нкРНК XIST ожидаемо присутствуют только на X-хромосоме, MALAT1 контактирует полногеномно и является одной из наиболее часто контактирующей РНК, контакты мРНК сосредоточены в основном вблизи от своего гена, хотя есть и дальние взаимодействия, включая контакты с другими хромосомами.

Авторы отмечают, что более 70% контактов приходятся на мРНК, тем не менее были детектированы длинные и короткие нкРНК разных классов. Показана корреляция уровня экспрессии по данным собственных RNA-seq с уровнем контактов для индивидуальных РНК. Выделены группы РНК, которые контактируют с хроматином по-разному в зависимости от расстояния от своего гена. Некоторые индивидуальные РНК и классы РНК предпочитали взаимодействовать с хроматином, находящимся в определенном состоянии. Например, малые ядерные РНК чаще взаимодействовали с активным хроматином, а некоторые представители классов очень длинных нкРНК (vlinc) и X-РНК взаимодействовали с репрессированным хроматином на расстоянии 5Мб от своего гена. Выявлены две микроРНК, предпочитающие контактировать с 18 хромосомой, причем исключительно с подавленным хроматином. В ходе изучения частот контактов интронв и экзонов белок-кодирующих генов был подтвержден ко-транскрипционный сплайсинг.

Предложена метрика хроматинового потенциала, с помощью которой можно выявить РНК, контактирующие с хроматином чаще, чем это ожидается из уровня их экспрессии. Сконцентрировавшись только на РНК с высоким хроматиновым

потенциалом показана ассоциация профиля контактов некоторых РНК с пиками из эксперимента по РНК иммунопреципитации (fRIP-seq) [85]. Такие РНК оказались ассоциированы с белками комплекса Polycomb, HDAC, DNMT1. Значительная часть хаРНК взаимодействовала с ADAR.

Другие подходы к изучению РНК-ДНК взаимодействий

В 2021 году был опубликован метод RD-SPRITE (RNA and DNA Split-Pool Recognition of Interactions by Tag Extension) [86], с помощью которого можно в том числе исследовать полногеномные РНК-ДНК взаимодействия. Предложенный в работе экспериментальный протокол, а также характер полученных данных в корне отличается от методов “все-против-всех”. Ключевая идея метода заключается в том, что после фиксации клеток *in situ*, фрагментации хроматина и специфической обработки концов фрагментов нуклеиновых кислот полученные макромолекулярные комплексы разделяют и итеративно баркодируют (технология split-and-pool), смешивая и разделяя заново перед каждым шагом баркодирования. После этого комплексы расшивают, последовательности нуклеиновых кислот секвенируют. Чтения, несущие одинаковую последовательность набора баркодов, кластеризуют, предполагая, что они одновременно входят в состав одного и того же комплекса. Зафиксированные макромолекулярные комплексы могут содержать фрагменты только РНК или только ДНК, а также РНК и ДНК сразу, причем для одного комплекса может быть зафиксировано сразу несколько фрагментов любых нуклеиновых кислот. Таким образом можно исследовать взаимодействие РНК и ДНК в пространстве. Методы “все-против-всех”, так же как и метод Hi-C, основаны на лигировании только близко расположенных в пространстве макромолекул и позволяют установить исключительно факт взаимодействия РНК с ДНК. В протоколе RD-SPRITE нет таких ограничений, можно изучать композицию достаточно крупных комплексов, фиксируя фрагменты РНК и ДНК, которые входят в состав этого комплекса одновременно. Авторам удалось выделить, например, сплайсосомальный кластер, куда входят такие РНК, как MALAT1 и мяРНК, тельца гистоновых локусов, зафиксировать инактивированную Х-

хромосому. Авторы отмечают ряд технических ограничений метода, которые заключаются в невозможности исследовать композиции комплексов в динамике, а также отмечают, что способ фиксации клеток позволяет в основном детектировать только взаимодействия, опосредованные белками.

Все вышеописанные методы фиксировали в основном РНК-ДНК контакты, опосредованные белками, что следует из способов фиксации. Однако, существуют способы детекции прямых контактов РНК и ДНК.

С помощью метода DRIP-seq можно идентифицировать R-петли [87]. R-петля представляет собой РНК-ДНК дуплекс, который формируется, например, во время транскрипции, когда новосинтезированная РНК комплементарно связывается с локусом ДНК, вытесняя некомплементарную ДНК, что приводит к образованию петли. Показано, что R-петли участвуют в процессе регуляции экспрессии генов, рекомбинации и репарации ДНК [88].

Также РНК могут непосредственно взаимодействовать с двухцепочечной ДНК, формируя хугстиновские пары с образованием ДНК:ДНК:РНК триплексов. Так, вышеописанная длинная нкРНК HOTAIR взаимодействует с ДНК именно таким способом, формируя ДНК:ДНК:РНК триплексы в специфических локусах, а затем привлекает модифицирующие хроматин комплексы, реализуя подавление транскрипции [89]. Аналогичный механизм показан для нкРНК Fendrr, MEG3, PARTICLE и некоторых других [89].

МАТЕРИАЛЫ И МЕТОДЫ

Данные полногеномного РНК-ДНК интерактома

Red-C

Основным набором данных в представленной работе являются результаты секвенирования эксперимента Red-C [25] (GEO: GSE136141). Были получены парно-концевые сиквенсы следующих библиотек (указано количество первичных чтений):

1. Клеточная линия K562
    1.1. SRR10010326 - повтор №1 (повторное секвенирование SRR10010328) - 122.5 млн пар чтений (Illumina HiSeq 2500; 100+100 нукл)
    1.2. SRR10010327 - контроль, без обработки ДНК-лигазой - 12.2 млн пар чтений (Illumina HiSeq 2500; 125+125 нукл)
    1.3. SRR10010328 - повтор №1 - 15.5 млн пар чтений (Illumina MiSeq; 80+80 нукл)
    1.4. SRR10010329 - контроль, обработка РНКазой - 10.6 млн пар чтений (Illumina MiSeq; 80+80 нукл)
    1.5. SRR10010330 - повтор №2 - 217 млн пар чтений (Illumina HiSeq 2500; 125+125 нукл)

2. Клеточная линия нормальных женских человеческих фибробластов кожи (fibro)
    2.1. SRR10010323 - повтор №1 (повторное секвенирование SRR10010324) - 320 млн пар чтений (Illumina HiSeq 2500; 133+133 нукл)
    2.2. SRR10010324 - повтор №1 - 9.3 млн пар чтений (Illumina HiSeq 2500; 125+125 нукл)
    2.3. SRR10010325 - повтор №2 - 18.6 млн пар чтений (Illumina HiSeq 2500; 125+125 нукл)

GRID-seq и RADICL-seq

В нашей группе были обработаны контакты РНК с хроматином из двух ранее опубликованных работ: GRID-seq [21] и RADICL-seq [24] (указано количество первичных чтений):

1. GRID-seq - клеточная линия MDA-MB-231
    1.1. GSM2188866 - повтор №1 - 146.9 млн чтений (Illumina HiSeq 2500; 100 нукл)
    1.2. GSM2188867 - повтор №2 - 144.8 млн чтений (Illumina HiSeq 2500; 100 нукл)

2. GRID-seq - клеточная линия MM.1S
    2.1. GSM2188868 - повтор №1 - 135 млн чтений (Illumina HiSeq 2500; 100 нукл)
    2.2. GSM2188869 - повтор №2 - 149.6 млн чтений (Illumina HiSeq 2500; 100 нукл)
3. GRID-seq - клеточная линия mESC
    3.1. GSM2396700 - повтор №1 - 157 млн чтений (Illumina HiSeq 2500; 100 нукл)
    3.2. GSM2396701 - повтор №2 - 113.6 млн чтений (Illumina HiSeq 2500; 100 нукл)
4. RADICL-seq - клеточная линия mES
    4.1. SRR9201799 - повтор №1, обработка 1% формальдегидом - 141.4 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.2. SRR9201801 - повтор №2, обработка 1% формальдегидом - 120.5 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.3. SRR9201803 - повтор №3, обработка 1% формальдегидом - 103.8 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.4. SRR9201805 - повтор №1, обработка 2% формальдегидом - 141.8 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.5. SRR9201807 - повтор №2, обработка 2% формальдегидом - 107.3 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.6. SRR9201809 - повтор №3, обработка 2% формальдегидом - 97.6 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.7. SRR9201811 - повтор №1, обработка актиномицином - 80.8 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.8. SRR9201813 - повтор №2, обработка актиномицином - 46.4 млн чтений (Illumina HiSeq 2500; 150 нукл)
    4.9. SRR9201815 - повтор №1, обработка NPM - 90.6 млн чтений (Illumina HiSeq 2500; 150 нукл)

4.10. SRR9201817 - повтор №2, обработка NPM - 94.8 млн чтений (Illumina HiSeq 2500; 150 нукл)

4.11. SRR9201819 - повтор №3, обработка NPM - 101.2 млн чтений (Illumina HiSeq 2500; 150 нукл)

5. RADICL-seq - клеточная линия mOPC

5.1. SRR9201821 - повтор №1, обработка 1% формальдегидом - 109.6 млн чтений (Illumina HiSeq 2500; 150 нукл)

5.2. SRR9201823 - повтор №2, обработка 1% формальдегидом - 177.1 млн чтений (Illumina HiSeq 2500; 150 нукл)

5.3. SRR9201825 - повтор №3, обработка 1% формальдегидом - 116.7 млн чтений (Illumina HiSeq 2500; 150 нукл)

5.4. SRR9201827 - повтор №1, обработка NPM - 183.4 млн чтений (Illumina HiSeq 2500; 150 нукл)

5.5. SRR9201829 - повтор №2, обработка NPM - 83.2 млн чтений (Illumina HiSeq 2500; 150 нукл)

Данные секвенирования РНК

Red-C

Из клеточной линии K562, использованной в эксперименте по определению РНК-ДНК интерактома, коллегами из лаборатории С.В. Разина была выделена тотальная РНК (с деплецией рибосомальной РНК), в результате секвенирования получены одноконцевые чтения с точной информацией о цепи [25] (GEO: GSE136141):

1. Клеточная линия K562:

1.1. SRR10010331 - повтор №1, 25.5 млн чтений (Illumina NextSeq 500; 75 нукл)

1.2. SRR10010332 - повтор №2, 22.6 млн чтений (Illumina NextSeq 500; 75 нукл)

RNA Atlas

Авторы проекта RNA Atlas [90] предоставляют доступ к данным секвенирования РНК для многих клеточных линий человека, сделанных по одному протоколу: тотальная РНК (с деплецией рибосомальной РНК), в результате секвенирования получены парноконцевые чтения с точной информацией о цепи.

1. Клеточная линия K562:
    1.1. SRR10266766 - повтор №1, 15.3 млн чтений (Illumina HiSeq 4000; 75 нукл)
    1.2. SRR10266767 - повтор №2, 16 млн чтений (Illumina HiSeq 4000; 75 нукл)
2. Клеточная линия MDA-MB-231
    2.1. SRR10261661 - повтор №1, 16 млн чтений (Illumina HiSeq 4000; 75 нукл)
    2.2. SRR10261662 - повтор №1, 14 млн чтений (Illumina HiSeq 4000; 75 нукл)
3. Клеточная линия дермальных фибробластов:
    3.1. SRR10264465 - повтор №1, 9.4 млн чтений (Illumina HiSeq 4000; 75 нукл)
    3.2. SRR10264466 - повтор №2, 9.9 млн чтений (Illumina HiSeq 4000; 75 нукл)

ENCODE

Данные о секвенировании РНК мышиных эмбриональных стволовых клеток были получены из базы ENCODE [91], где представлены парноконцевые чтения с точной информацией о цепи, полученные с помощью необходимого протокола: тотальная РНК с деплецией рибосомальной РНК.

1. Клеточная линия мышиных эмбриональных стволовых клеток (E14):
    1.1. SRR5048190 - повтор №1, 118 млн чтений (Illumina HiSeq 2000; 100 нукл)

1.2. SRR5048191 - повтор №2, 164 млн чтений (Illumina HiSeq 2000; 100 нукл)

Геномы

1. Последовательность генома человека версии GRCh37 (hg19) получен из Assembly (NCBI), только канонические хромосомы (chr1 - chr22; chrX; chrY).
2. Последовательность генома человека версии GRCh38.p13 (hg38) получен из Assembly (NCBI), только канонические хромосомы (chr1 - chr22; chrX; chrY).
3. Последовательность генома мыши версии GRCm38.p6 (mm10) получен из Assembly (NCBI), только канонические хромосомы (chr1 - chr19; chrX; chrY).

Разметка генов

Для геномов человека и мыши за основу была выбрана разметка генов по версии GENCODE [92]:

- release 35 для версии GRCh38
- release 27 для версии GRCh37
- release M25 для версии GRCm38

Дополнительно для человека была добавлена разметка генов очень длинных РНК (vlinc) из работы [93] в количестве 2762 штук. Изначально разметка опубликована в координатах референсного генома версии hg19, с помощью LiftOver [94] координаты генов были переведены в координаты генома человека версии hg38. Также для человека (только для версии hg19) добавлена разметка piРНК из базы данных piRNABank [95]. РНК классов piРНК и vlinc не представлены в генной аннотации, предоставляемой проектом GENCODE.

Из геномного браузера были получены разметки генов малых РНК (мяРНК, мякРНК, микроРНК, тРНК, а также РНК из repeatMasker), которые представлены в малом количестве в разметках ENCODE.

Полную информацию о количестве генов и их источниках для всех референсных геномах можно увидеть в таблице 2.

Таблица 2. Сводная информация о количестве и источниках генных разметок, используемых в работе.

Тип РНК	Источник	hg19	hg38	mm10
3prime_overlapping_ncRNA	GENCODE	31	0	3
antisense_RNA	GENCODE	5536	0	2991
bidirectional_promoter_lncRNA	GENCODE	18	0	198
CDBox	UCSC - sno/miRNA	0	269	0
HAcaBox	UCSC - sno/miRNA	0	112	0
IG_C_gene	GENCODE	14	14	13
IG_C_pseudogene	GENCODE	9	9	1
IG_D_gene	GENCODE	37	37	19
IG_D_pseudogene	GENCODE	0	0	3
IG_J_gene	GENCODE	18	18	14
IG_J_pseudogene	GENCODE	3	3	0
IG_LV_gene	GENCODE	0	0	4
IG_pseudogene	GENCODE	1	1	2
IG_V_gene	GENCODE	142	144	218
IG_V_pseudogene	GENCODE	191	188	158
intron	UCSC - tRNA Genes	34	0	0
lincRNA	GENCODE	7551	0	5629
lncRNA	GENCODE	0	16899	0
macro_lncRNA	GENCODE	1	0	2
miRNA	GENCODE | UCSC - sno/miRNA	3043 | 0	1881 | 1918	2202 | 0
misc_RNA	GENCODE	2032	2212	562
non_coding	GENCODE	3	0	0
piRNA	piRNABank	667836	0	0
polymorphic_pseudogene	GENCODE	62	49	89
processed_pseudogene	GENCODE	10161	10169	10002
processed_transcript	GENCODE	541	0	779
protein_coding	GENCODE	20210	19941	21846
pseudogene	GENCODE	565	18	61
ribozyme	GENCODE	0	8	22
RNA	UCSC - RepeatMasker	717	666	691
rRNA	GENCODE | UCSC - RepeatMasker	526 | 1707	47 | 1751	354 | 1563
rRNA_pseudogene	GENCODE	0	497	0
scaRNA	GENCODE | UCSC - sno/miRNA	0 | 0	49 | 21	51 | 0
scRNA	GENCODE | UCSC - RepeatMasker	1 | 1288	1 | 1334	1 | 8320
sense_intronic	GENCODE	909	0	328
sense_overlapping	GENCODE	189	0	29
snoRNA	GENCODE	1458	943	1507
snRNA	GENCODE | UCSC - RepeatMasker	1913 | 4259	1901 | 4285	1383 | 3004
sRNA	GENCODE	0	5	2
srpRNA	UCSC - RepeatMasker	1452	1595	437
TEC	GENCODE	1019	1058	3238
TR_C_gene	GENCODE	5	6	8
TR_D_gene	GENCODE	4	4	4
TR_J_gene	GENCODE	73	79	70
TR_J_pseudogene	GENCODE	5	4	10
TR_V_gene	GENCODE	96	106	144
TR_V_pseudogene	GENCODE	30	33	34
transcribed_processed_pseudogene	GENCODE	451	500	300
transcribed_unitary_pseudogene	GENCODE	110	138	26
transcribed_unprocessed_pseudogene	GENCODE	795	941	271
translated_processed_pseudogene	GENCODE	2	2	0
translated_unprocessed_pseudogene	GENCODE	0	1	2
tRNA	UCSC - tRNA Genes | UCSC - RepeatMasker	570 | 1759	629 | 1777	434 | 4755
unitary_pseudogene	GENCODE	95	97	61
unprocessed_pseudogene	GENCODE	2521	2615	2723
vaultRNA	GENCODE	1	1	0
vlinc	vlinc	2762	2762	0

Разметка состояний хроматина

Типы состояний хроматина были определены согласно предложенной аннотации из работы *Ernst et al.* [96] для клеточной линии K562. Используя различные комбинации эпигенетических меток, авторы выделяют 15 состояний, которые представляют собой набор неперекрывающихся интервалов, покрывающих геном человека практически по всей длине. Координаты данной разметки предоставлены для версии генома hg19. Выделяют: активные промоторы (1), слабые промоторы (2), неактивные промоторы, (3), сильные энхансеры (4 и 5), слабые энхансеры (6 и 7), CTCF-зависимые инсуляторы (8), переходные стадии транскрипции (9), элонгация транскрипции (10), слабая транскрипция (11), подавленные участки (Polycomb) (12), гетерохроматин (13) и области, богатые повторами и CNV (14 и 15). Предложенные состояния хроматина были сгруппированы в более крупные: активный хроматин (Act) (1+2+4+5+6+7+9+10+11) и репрессированный хроматин (Rep) (3+12+13).

Полногеномные разметки

Координаты BlackList были получены для человека (версии hg19 и hg38) и для мыши (версии mm10) из ENCODE (Accession: ENCSR636HFF).

Разметка локусов ДНК, гиперчувствительных к обработке ДНКазой I, получена из ENCODE для клеточной линии K562, версия референсного генома hg19, файл в формате bigwig (GEO: GSM816655; ENCFF352SET).

Данные о временных профилях репликации получены из работы *Hansen et al.* [97] для клеточной линии K562, версия референсного генома hg19. Предоставлено 6 файлов в формате bigwig, отражающих паттерны репликации в зависимости от стадии клеточного цикла, выделяя G1 (соответствует ранней репликации), S1, S2, S3, S4, S5, G2 (соответствует поздней репликации).

Программы и пакеты

Для анализа качества чтений была использована программа fastQC (версия 0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), на вход которой

были поданы чтения в формате fastq. Были проанализированы чтения, полученные из экспериментов по определению полногеномного РНК-ДНК интерактома (для протоколов Red-C, GRID-seq и RADICL-seq), а также все использованные в работе чтения из экспериментов по исследованию экспрессии (RNA-seq).

Программа MultiQC (версия 1.9) [98] была использована для визуализации результатов по исследованию качества чтений с помощью fastQC, позволяя объединять информацию по многим образцам сразу.

Картирование чтений, полученных из экспериментов по определению уровня экспрессии (RNA-seq), на референсный геном было осуществлено с помощью программы HISAT2 (версия 2.0.5), которая позволяет учитывать сплайсинг. Для поиска дифференциально экспрессирующихся генов использован пакет для R DESeq2, предполагающий отрицательное биномиальное распределение количества прочтений на ген. Анализ дифференциальной экспрессии был реализован на данных из работы *Potashnikova et al.* [99] по изучению клеточного цикла (клеточная линия K562).

Для работы с результатами выравнивания прочтений на референсный геном (в форматах sam и bam) были использованы возможности программы samtools (версия 0.1.18). Данная программа позволяет сортировать, индексировать, фильтровать результаты картирования, а также получать информацию о количестве картированных прочтений, что было необходимо при сборе соответствующих метрик.

Для подсчета корреляций полногеномных разметок и профилей контактов была использована программа Stereogene [100]. На вход программе необходимо подать соответствующие разметки в формате bedgraph, после чего в окнах заданного размера подсчитываются коэффициенты корреляции. Статистическая значимость корреляций оценивается с помощью теста на основе перестановок. Для клеточной линии K562 (версия референсного генома hg19) были исследованы корреляции:

- фоновых контактов из протокола Red-C с разметкой областей генома, чувствительных к обработке ДНКазой I (размер окна 1 Мб);
- профиля контактов микроРНК MIR3687 из протокола Red-C с временными профилями репликации (размер окна 20 Мб).

Для сглаживания трека фоновых контактов также была использована программа Stereogene, модуль Smoother (размер окна 1 Мб).

Аппроксимация распределения хроматинового потенциала для мРНК была осуществлена с помощью пакета для R fitdistrplus (версия 1.1-5).

Для работы с геномными интервалами использована программа bedtools (версия 2.29.2) [101], а также пакет для языка R GenomicRanges (версия 1.46.1).

Для визуализации результатов были использованы следующие пакеты для языка R: ggplot2 (версия 3.3.5), cowplot (версия 1.1.1), karyoploteR (версия 1.20.3), VennDiagram (версия 1.6.20).

Работа с табличными данными была осуществлена с помощью вспомогательных сценариев на bash, написанных самостоятельно, а также с помощью пакетов для языка R: tidyverse (версия 1.3.1), data.table (версия 1.14.0).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

На сегодняшний день представлено не так много экспериментальных подходов для получения данных об РНК-ДНК интерактоме (протоколы типа “все-против-всех”), а специально разработанных программ и алгоритмов для полного анализа такого типа данных не существует совсем.

В разработке одного из таких протоколов - Red-C - поучаствовала наша группа под руководством Андрея Александровича Миронова [25]. Работа над протоколом Red-C была осуществлена в коллаборации с лабораторией Сергея Владимировича Разина из Института биологии гена РАН, сотрудники которого полностью осуществляли экспериментальную часть проекта.

В данной работе мы предлагаем ознакомиться с многоступенчатым биоинформатическим подходом к анализу данных полногеномных РНК-ДНК

взаимодействий. Данный подход позволяет начинать анализ с “сырых” или первичных чтений, до какой бы то ни было обработки, производить ряд фильтраций, учитывающих особенности экспериментальной подготовки, формировать РНК-ДНК контакты, идентифицировать хроматин-ассоциированные РНК (хаРНК) согласно любой выбранной генной разметке, изучать характер взаимодействия РНК с хроматином, производить различные нормировки. Также мы предлагаем подход, позволяющий идентифицировать гипотетические новые РНК, ассоциированные с хроматином, но не представленные в существующей разметке генов.

Основные результаты и наблюдения показаны для эксперимента Red-C, т.к. мы имели доступ к абсолютно всем исходным данным, полученным сразу после секвенирования, а представленный протокол был разработан специально для метода Red-C. В качестве дополнительных источников для тестирования возможности применения предлагаемого биоинформатического протокола к другим данным схожего типа и сравнения некоторых этапов анализа были выбраны эксперименты GRID-seq [21] и RADICL-seq [24], как наиболее близкие к Red-C с экспериментальной точки зрения.

Количество чтений в экспериментах “все-против-всех”

Разнообразие и объем анализируемых данных по РНК-ДНК интерактому можно увидеть в таблице 3.

Исходное количество чтений во всех экспериментах достаточно высокое (более 120 млн чтений для клеточного типа), а наблюдаемый разброс можно объяснить количеством представленных реплик. В таблице 4 можно увидеть аналогичные данные о количестве первичных чтений для каждой реплики каждого протокола индивидуально.

Таблица 3. Количество первичных чтений в экспериментах “все-против-всех”. Все реплики объединены по тканям в рамках одного протокола и организма. Количество представленных реплик указано в столбце “Реплики”. Клетки: K562 - клеточная линия хронического миелолейкоза; fibro - нормальные женские фибробласты кожи; MDA-MB-231 - клетки рака груди; MM-1S - множественная миелома; mESC - эмбриональные стволовые клетки мыши с обработками 1% формальдегидом (1FA), 2% формальдегидом (2FA), протеиназой К в денатурирующих условиях (NPM), актиномицином Д (Act); mOPC - клетки-предшественники олигодендроцитов мыши с обработками 1% формальдегидом (1FA), протеиназой К в денатурирующих условиях (NPM).

Протокол	Организм	Клетки	Реплики	Чтения (млн)
Red-C	Человек	K562	3	355.2
Red-C	Человек	fibro	3	348.1
GRID-seq	Человек	MDA-MB-231	2	291.7
GRID-seq	Человек	MM-1S	2	284.6
GRID-seq	Мышь	mESC	2	270.6
RADICL-seq	Мышь	mESC_1FA	3	365.6
RADICL-seq	Мышь	mESC_2FA	3	346.8
RADICL-seq	Мышь	mESC_NPM	3	286.6
RADICL-seq	Мышь	mESC_Act	2	127.2
RADICL-seq	Мышь	mOPC_1FA	3	403.5
RADICL-seq	Мышь	mOPC_NPM	2	266.6

Внутри одной клеточной линии количество чтений по репликам может сильно различаться (табл. 4), что особенно ярко видно для протокола Red-C. В процессе обработки контактов все реплики до момента аннотации РНК-частей генами были обработаны независимо. Мы обращали внимание на то, как ведут себя реплики внутри одного протокола и одной клеточной линии, удается ли наблюдать для них одинаковые результаты и схожие тенденции. При аннотации РНК-частей генами реплики были объединены с целью увеличения количества данных и покрытия и далее были исследованы совместно.

Таблица 4. Количество первичных чтений в экспериментах “все-против-всех” отдельно для каждой реплики. Количество чтений (в млн) указано в последнем столбце таблицы “Чтения”.

Протокол	Организм	Клетки	Реплики (ID)	Чтения (млн)
Red-C	Человек	K562	SRR10010326	122.5
Red-C	Человек	K562	SRR10010328	15.5
Red-C	Человек	K562	SRR10010330	217.0
Red-C	Человек	fibro	SRR10010323	320.0
Red-C	Человек	fibro	SRR10010324	9.3
Red-C	Человек	fibro	SRR10010325	18.6
GRID-seq	Человек	MDA-MB-231	GSM2188866	146.9
GRID-seq	Человек	MDA-MB-231	GSM2188867	144.8
GRID-seq	Человек	MM-1S	GSM2188868	135.0
GRID-seq	Человек	MM-1S	GSM2188869	149.6
GRID-seq	Мышь	mESC	GSM2396700	157.0
GRID-seq	Мышь	mESC	GSM2396701	113.6
RADICL-seq	Мышь	mESC_1FA	SRR9201799	141.4
RADICL-seq	Мышь	mESC_1FA	SRR9201801	120.5
RADICL-seq	Мышь	mESC_1FA	SRR9201803	103.8
RADICL-seq	Мышь	mESC_2FA	SRR9201805	141.8
RADICL-seq	Мышь	mESC_2FA	SRR9201807	107.3
RADICL-seq	Мышь	mESC_2FA	SRR9201809	97.6
RADICL-seq	Мышь	mESC_NPM	SRR9201815	90.6
RADICL-seq	Мышь	mESC_NPM	SRR9201817	94.8
RADICL-seq	Мышь	mESC_NPM	SRR9201819	101.2
RADICL-seq	Мышь	mESC_Act	SRR9201811	80.8
RADICL-seq	Мышь	mESC_Act	SRR9201813	46.4
RADICL-seq	Мышь	mOPC_1FA	SRR9201821	109.6
RADICL-seq	Мышь	mOPC_1FA	SRR9201823	177.1
RADICL-seq	Мышь	mOPC_1FA	SRR9201825	116.7
RADICL-seq	Мышь	mOPC_NPM	SRR9201827	183.4
RADICL-seq	Мышь	mOPC_NPM	SRR9201829	83.2

Анализ качества результатов секвенирования

Приступая к анализу данных высокопроизводительного секвенирования, прежде всего необходимо убедиться, что полученные чтения хорошего качества и пригодны для дальнейшей работы. В работе были использованы первичные чтения из трех экспериментов “все-против-всех”, а также чтения из работ по секвенированию РНК.

Качество полученных чтений для данных РНК-ДНК интерактома было проанализировано дважды. Первый раз было проверено качество первичных чтений, т.е. чтений, полученных сразу после секвенирования, или чтений, предоставленных авторами (для протоколов GRID-seq и RADICL-seq). На основании первичной оценки принималось решение о целесообразности дальнейшей работы с данными и необходимости дополнительной фильтрации. Повторно качество чтений было проанализировано после дополнительных манипуляций (удаление дублированных чтений, технических последовательностей, а также нуклеотидов и чтений плохого качества), перед картированием.

Качество всех чтений реплик основной библиотеки протокола Red-C (рис. 5 А и Б), включая контрольные библиотеки и данные секвенирования, были признаны удовлетворительными и пригодными для дальнейшей работы. Чтения в экспериментах GRID-seq (рис. 5В) и RADICL-seq (рис. 5Г) короткие, что следует из протоколов, а также практически во всех первичных библиотеках к концу чтений качество сильно падало (особенно в случае GRID-seq), тем не менее все данные были взяты в работу, ситуация с качеством была исправлена на соответствующем этапе фильтрации.

Качество чтений из внешних экспериментов по секвенированию РНК для клеточных линий человека (K562, фибробласты, MDA-MB-231 из проекта RNA Atlas [90]) и мыши (эмбриональные стволовые клетки из проекта ENCODE) были признаны удовлетворительными и взяты в работу без дополнительной фильтрации по качеству.

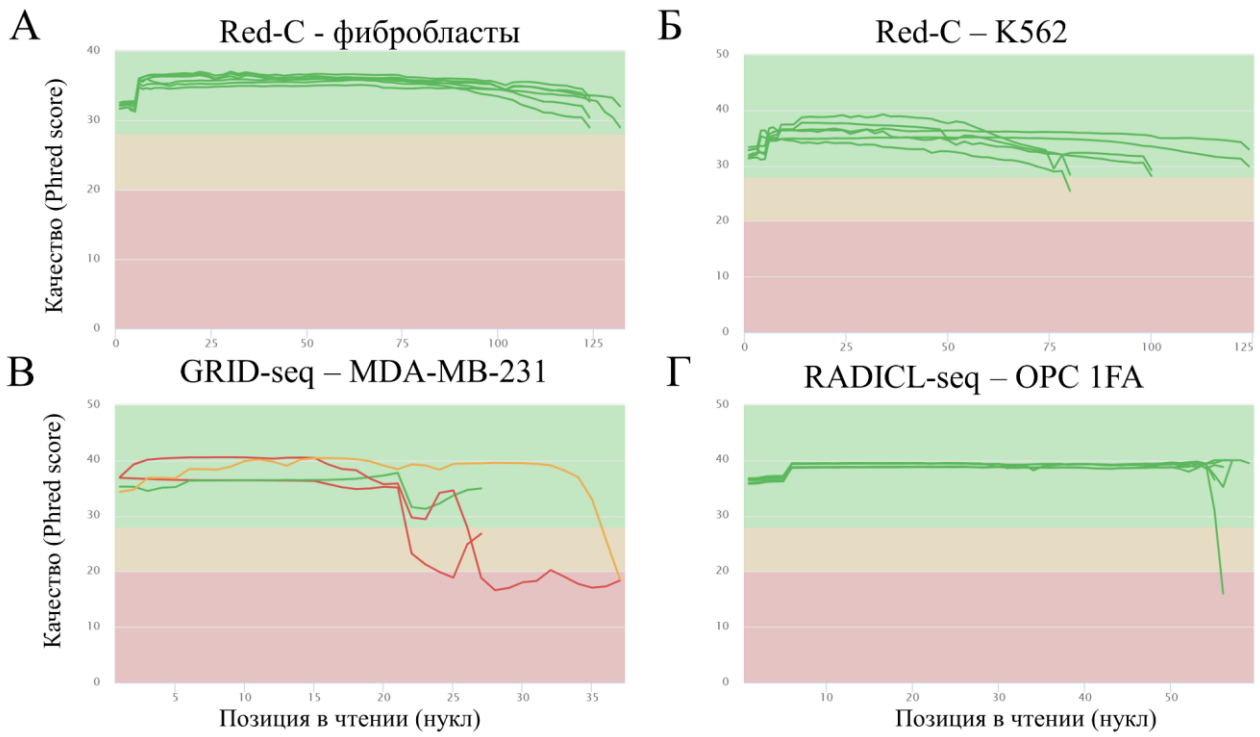

Рисунок 5. Распределение медианы качества нуклеотидов исходных прямых и обратных прочтений (Phred quality score). Зеленая зона: Phred quality score > 28; точность определения нуклеотида ~99.8%. Желтая зона: Phred quality score 20-28; точность определения нуклеотида 99-99.8%. Красная зона: Phred quality score < 20; точность определения нуклеотида менее 99%. Представлены данные для основных библиотек из протоколов: (А) Red-C (клеточная линия нормальных фибробластов кожи); (Б) - Red-C (клеточная линия K562); (В) - GRID-seq (клеточная линия MDA-MB-231); (Г) - RADICL-seq (клеточная линия OPC, фиксация 1% формальдегидом).

Структура чтений библиотеки Red-C

По результатам секвенирования были получены парно-концевые чтения (80-125 нукл в зависимости от модели секвенатора, см. раздел “Материалы и методы”), содержащие не только РНК и ДНК части гипотетически контактирующих молекул нуклеиновых кислот, но и технические последовательности. Схематичное изображение конструкции, подготовленной для секвенирования, можно увидеть на рисунке 6.

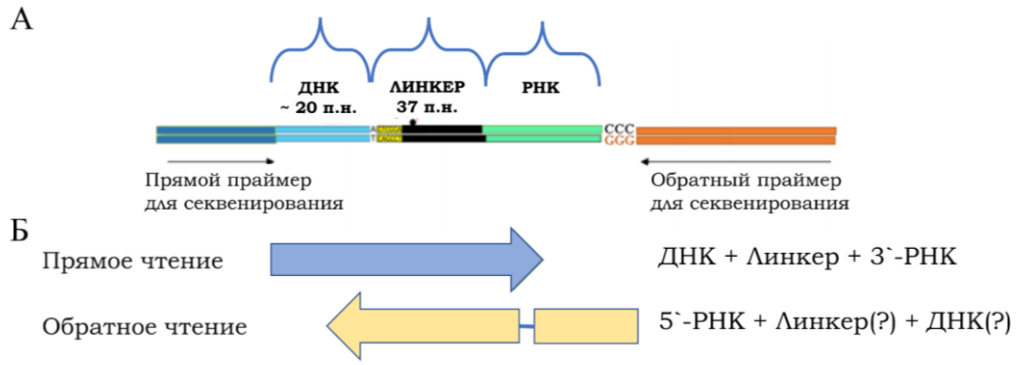

Рисунок 6. (А). Схема химерной конструкции для секвенирования, полученной в ходе эксперимента Red-C. (Б). Ожидаемая структура прямого и обратного прочтений.

Из последовательности прямого чтения можно получить ДНК-часть контакта и 3`-РНК-часть, которая примыкает непосредственно к линкеру. Обратное прочтение содержит последовательность 5`-РНК-части контакта. Протокол Red-C позволяет исследовать хаРНК любой длины. Таким образом, если удалось зафиксировать длинный фрагмент РНК (больше длины чтения), то обратное прочтение будет содержать только последовательность РНК-части, если же РНК оказалась короткой или в какой-то степени деградировала в ходе эксперимента, в последовательности обратного чтения мы увидим фрагмент РНК, затем последовательность линкера (или его часть) и даже возможно фрагмент ДНК.

Исследование контрольных экспериментов Red-C

Для подтверждения корректности работы протокола Red-C были поставлены контрольные эксперименты, в которых варьировали присутствие необходимых ферментов для лигирования ДНК, а также добавляли или не добавляли обработку РНКазой А при разном количестве циклов ПЦР (рис. 7). Эксперименты были выполнены сотрудниками лаборатории С.В. Разина.

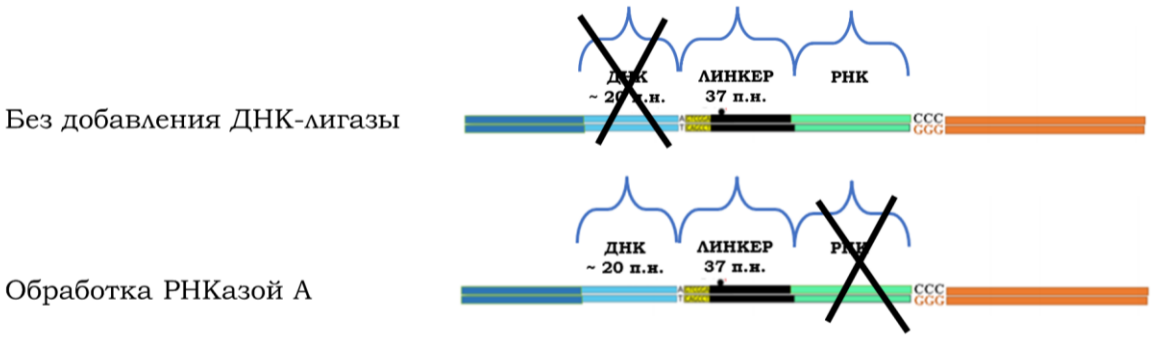

Рисунок 7. Ожидаемая структура химерной конструкции для контрольных библиотек (протокол Red-C).

Как и ожидалось, четкий фрагмент необходимого размера был обнаружен только в случае добавления всех ферментов, предусмотренных протоколом (лучший результат на 15 циклах ПЦР при электрофоретическом анализе) (рис. 8).

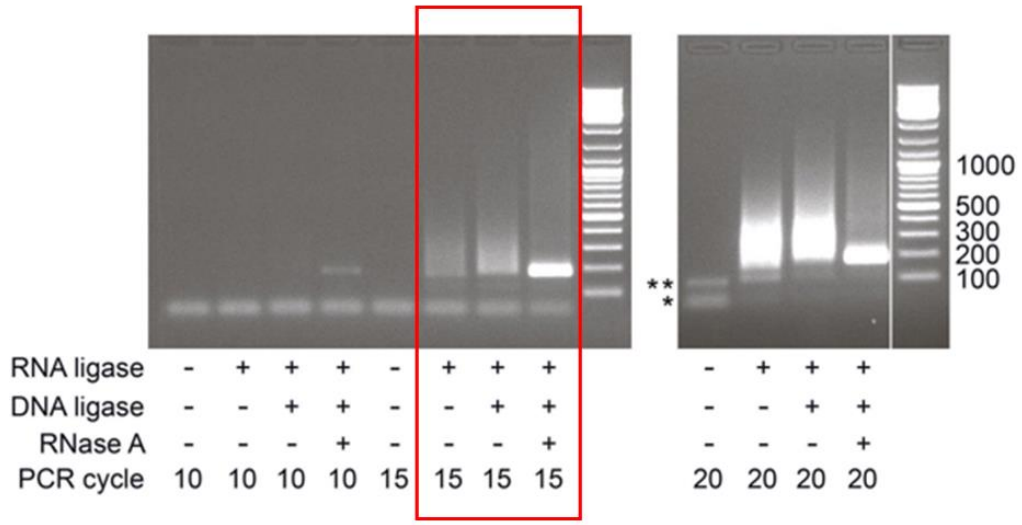

Рисунок 8. Электрофорез продуктов ПЦР основного и контрольных опытов с клетками K562 (протокол Red-C). Ожидаемые размеры продуктов ПЦР: химера без частей ДНК и РНК, 163 п.н. (только технические последовательности); химера с частью ДНК без части РНК, ~181 п.н. (технические последовательности и ДНК-часть 18-20 нуклеотидов); химера с частью РНК без части ДНК, > 163 п.н. (технические последовательности и РНК-часть любой длины); химера с частями ДНК и РНК > 181 п.н. (технические последовательности, ДНК-часть 18-20 нуклеотидов, РНК-часть любой длины). Фотография электрофореза предоставлена Алексеем Гавриловым.

Библиотеки, отмеченные красной рамкой, были отправлены на секвенирование. На рисунке 9 можно увидеть структуру целевой химерной конструкции, где технические последовательности представлены в явном виде.

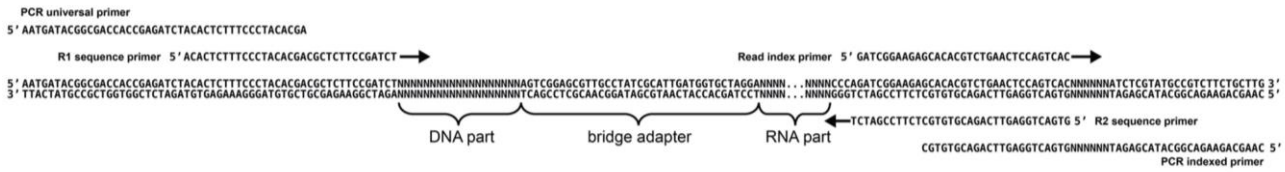

Рисунок 9. Структура химерной конструкции для протокола Red-C. Технические последовательности представлены в явном виде. Адаптировано из [25].

Посмотрим на последовательности прямых чтений трех библиотек: основной и двух контролей.

На сиквенсах прямых чтений из основной бибилиотеки (рис. 10) видно, что вокруг линкера с нужных сторон расположены последовательности ДНК (~ 20 нукл) и РНК (между линкером и праймером с CCC). Можно заметить, что по длине РНК часть может сильно варьировать, что будет учтено при дальнейшей фильтрации.

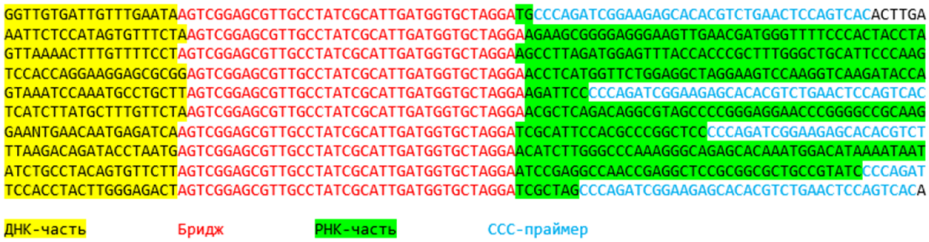

Рисунок 10. Прямые чтения из сиквенса основной библиотеки эксперимента Red-C, выбранные случайным образом, клеточная линия K562. Исследуемые последовательности отмечены цветом.

На рисунке 11 представлен сиквенс контрольной библиотеки без обработки ДНК лигазой. В данном случае с одной (левой) стороны линкера (AGT...) не должно было присоединиться ничего, а с другой - мы ожидаем увидеть РНК-часть (аналогично рис. 10), что и было показано.

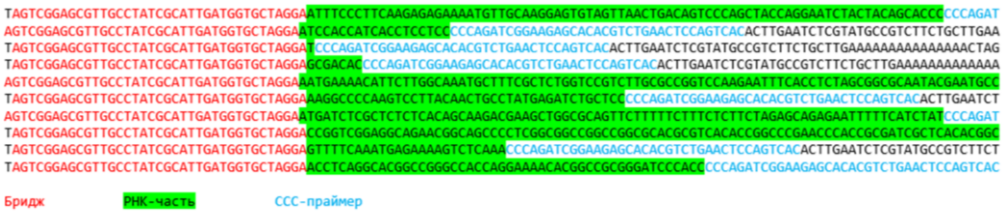

Рисунок 11. Прямые чтения из сиквенса контрольной библиотеки без обработки ДНК лигазой эксперимента Red-C, выбранные случайным образом, клеточная линия K562. Исследуемые последовательности отмечены цветом.

В результатах сиквенса контрольной библиотеки с обработкой РНКазой А показано присоединений слева от линкера ДНК-части ожидаемой длины ~ 20 нукл, а с другой стороны линкера сразу идет последовательность праймера, который начинается с CCC (рис. 12).

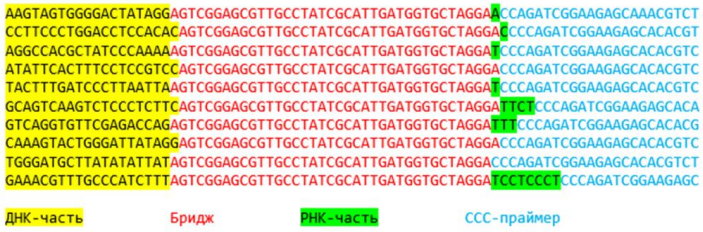

Рисунок 12. Прямые чтения из сиквенса контрольной библиотеки с обработкой РНКазой А эксперимента Red-C, выбранные случайным образом, клеточная линия K562. Исследуемые последовательности отмечены цветом.

Таким образом было показано, что метод работает корректно, полученная химерная конструкция содержит последовательности соответствующих нуклеиновых кислот ожидаемых длин, лигированных с запланированными в эксперименте концами линкера.

Биоинформатический протокол анализа РНК-ДНК интерактома

Для каждого протокола процесс сборки финальных РНК-ДНК контактов состоял из множества шагов, учитывающих особенности эксперимента. Стоит отметить, что авторы публикаций не всегда предоставляют доступ к полным данным.

Всю процедуру биоинформатической обработки можно разделить на три многоступенчатых этапа (рис. 13):

1. От первичных чтений до первичных контактов.
2. Фильтрация и аннотация полученных контактов, сборка РНК, не представленных в генной разметке.
3. Конструирование фона, расчет хроматинового потенциала, исследование характера взаимодействия РНК с хроматином.

Дополнительно по каждому пункту этапов были собраны метрики для оценки корректности работы процедуры и для анализа данных.

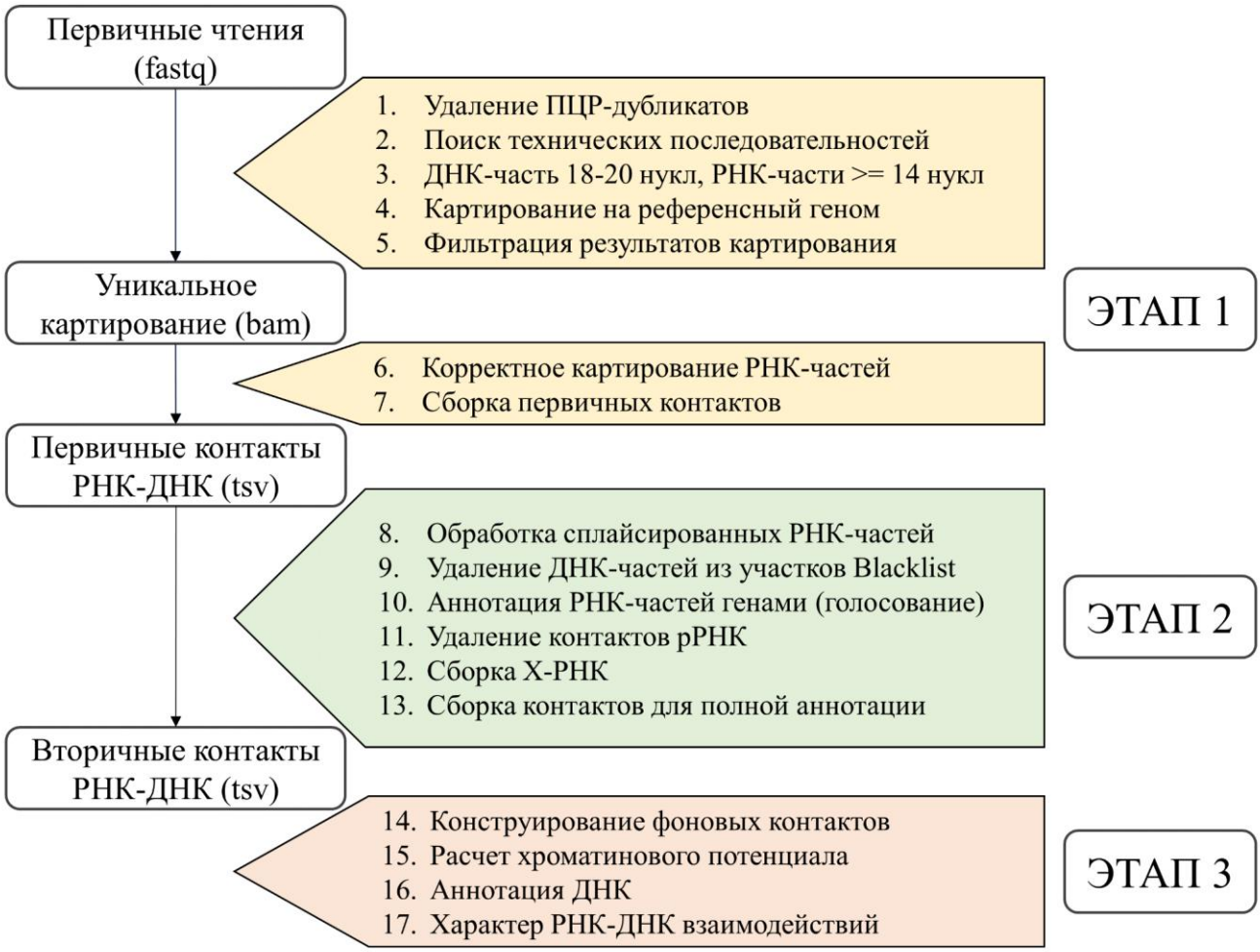

Рисунок 13. Схема анализа данных РНК-ДНК интерактома для протокола Red-C.

Преимущественно на примере эксперимента Red-C (клеточная линия K562) разберем каждый этап отдельно.

Первичная подготовка данных

Этап №1 реализован и имплементирован для данных Red-C в виде RedClib Александрой Галицыной, для экспериментов GRID-seq и RADICL-seq RedClib модифицирован и применен Юрием Коростелевым и Андреем Сигорских.

Удаление ПЦР-дубликатов

Прежде всего из исходных данных, до какой бы то ни было обработки, были удалены такие чтения, которые полностью дублировали последовательности друг друга (одновременно и по прямому, и по обратному прочтению). Такие “стопки” чтений могли образоваться в результате ПЦР во время пробоподготовки. Из каждой “стопки” была оставлена одна пара прочтений. На этом этапе было удалено ~52 млн чтений (14.6%; здесь и далее процент был рассчитан от исходного количества контактов). В случае протокола GRID-seq данный этап был пропущен, т.к. предоставленные авторами данные уже прошли процедуру удаления гипотетических дубликатов согласно процедуре, разработанной в эксперименте GRID-seq.

Поиск технических последовательностей

Следующий шаг является одним из самых важных для протокола Red-C - поиск и удаление технических последовательностей. В случае корректного присоединения линкера к нуклеиновым кислотам мы бы ожидали в прямых прочтениях обнаружить полную последовательность линкера (допускалась одна ошибка кроме сайта GA, граничащего с РНК-частью), а последовательности обратных прочтений должны были бы начинаться с CCC\GGG (первые 3 нуклеотида после последовательности R2 праймера для секвенирования) (рис. 9). Были отобраны только пары чтений, отвечающие этим двум условиям.

Далее было необходимо подготовить РНК и ДНК части отобранных пар чтений к картированию, удалив все технические последовательности. Из прямого чтения была взята часть левее линкера (ДНК-часть), а также часть правее линкера (3`-РНК-часть). Дополнительно в прямом чтении искали и удаляли фрагменты

последовательности CCC\GGG + праймер R2, т.к. РНК-часть могла оказаться слишком короткой и закончиться раньше, чем максимальная длина чтения. В обратном прочтении искали последовательность линкера и выделяли нуклеотиды, начинающиеся сразу правее линкера (5`-РНК-часть).

Для других протоколов этот шаг был также пропущен, т.к. авторы предоставили доступ к чтениям, откуда уже были удалены все технические последовательности, присущие конкретному протоколу.

Фрагменты, соответствующие ДНК-части контакта, согласно протоколу на одном из первых этапов экспериментальной процедуры были обработаны рестриктазой NlaIII. В дальнейшую работу были взяты только такие контакты, где ДНК-часть примыкала к CATG, что соответствует сайту рестрикции NlaIII. Для увеличения процента уникального картирования ДНК-части были дополнены нуклеотидами CATG. Чуть меньше 37 млн (~10.5%) чтений не удовлетворяли всем вышеописанным условиям.

Для эксперимента GRID-seq была проведена аналогичная процедура, однако в протоколе была использована рестриктаза AluI, что было учтено. Для RADICL-seq этот этап был пропущен, т.к. хроматин обрабатывали ДНКазой I.

Фильтрация РНК и ДНК фрагментов контактов по длине

Для картирования на референсный геном были отобраны ДНК-части длиной 18-20 нуклеотидов (до достраивания сайтов рестрикции) и РНК-части не менее 14 нуклеотидов в длину для более точного картирования в дальнейшем. На этом шаге было потеряно достаточно много контактов, почти 83 млн (более 23%), в основном из-за того, что одна из РНК частей была меньше требуемого размера. Видимо, в ходе экспериментального протокола молекулы РНК могли деградировать.

Независимо от поиска и удаления технических последовательностей был проведен анализ качества нуклеотидов. Финальные РНК и ДНК части, пошедшие на картирование, содержали в основном нуклеотиды хорошего качества. По причине плохого качества одной из частей отфильтровали еще 41,5 млн чтений

(~12%). Стоит отметить, что в случае GRID-seq по причине плохого качества было потеряно ~20% данных.

Картирование на референсный геном

Все три части контактов эксперимента Red-C (3`-РНК, 5`-РНК, ДНК) были независимо картированы на канонические хромосомы генома человека версий hg19 и hg38 для того, чтобы в дальнейшем использовать различные готовые разметки для аннотации ДНК и РНК частей вне зависимости от того, для какой версии генома они предоставлены.

Для протоколов GRID-seq и RADICL-seq реплики человеческих клеточных линий были картированы на референсный геном человека версии hg38, а реплики мышиных клеточных линий - на референсный геном mm10.

Фильтрация результатов картирования

Далее были отобраны только такие контакты, для которых и ДНК и обе РНК части были картированы на геном уникально и не более чем с двумя ошибками. На этом этапе мы потеряли самое большое количество данных - 97 миллионов (~27%). Стоит отметить, что на референсный геном было картировано ~99% чтений, а чтений с большим количеством ошибок было менее 1%. Таким образом, основной причиной потери данных на этом этапе было множественное картирование хотя бы одной из трех частей контакта.

Результаты картирования чтения из экспериментов GRID-seq и RADICL-seq были таким же образом отфильтрованы. Для всех клеточных линий на этом этапе также был отмечен самый большой процент потерянных данных.

Исследование корректности картирования РНК-частей контактов

Шаг 6 Этапа №1 характерен только для протокола Red-C, а нижеописанные фильтры суммарно отбраковали 6.7 млн контактов (~ 2%).

Мы ожидаем, что 3`-РНК и 5`-РНК части должны быть сиквенсами фрагмента одной молекулы РНК, прочитанной с разных концов, т.е. должны быть

картированы на одну хромосому, в взаимно обратной ориентации друг относительно друга. Дополнительно был введен критерий на удаленность картирования РНК частей: не далее чем 10 Кб друг от друга. Контакты, не прошедшие эти фильтры, были удалены.

Эксперимент Red-C единственный, где для предполагаемой ассоциированной с хроматином РНК репортируют сразу два фрагмента. Можно использовать эту информацию, чтобы оценить, насколько длинные молекулы РНК могут быть детектированы в эксперименте.

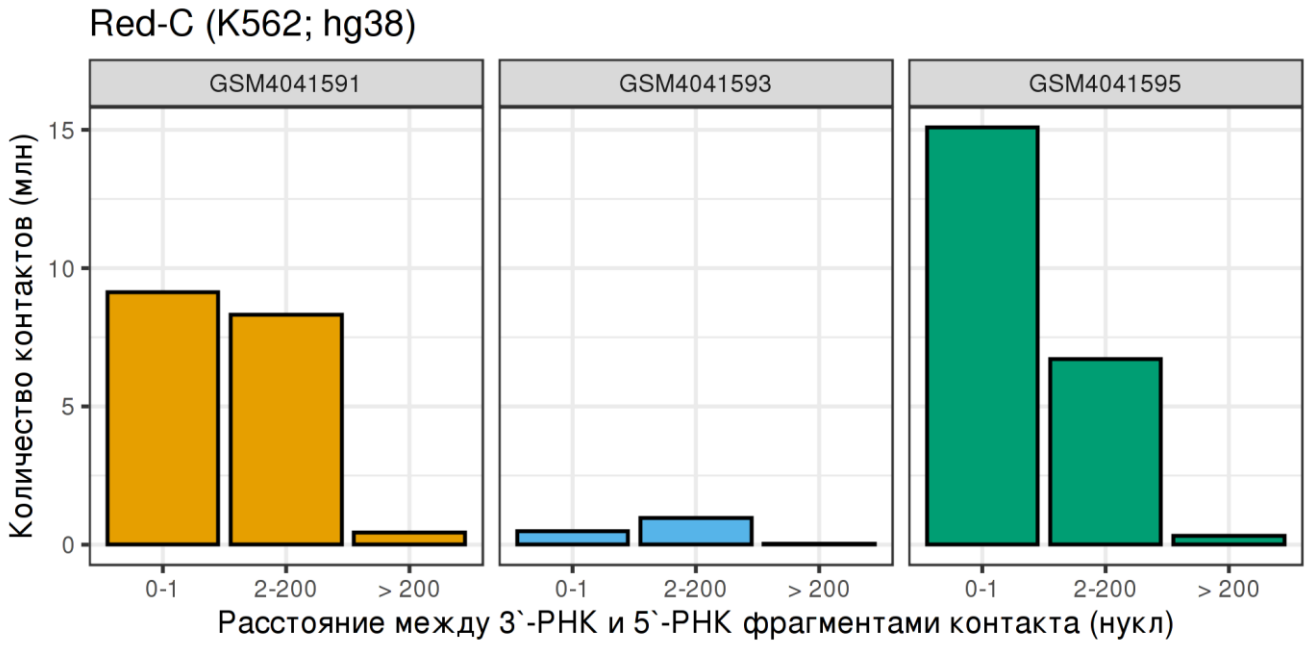

Рисунок 14. Распределение расстояний между 3`- и 5`-РНК частями из эксперимента Red-C (клеточная линия K562), картированных на референсный геном человека версии hg38. Реплики указаны отдельно.

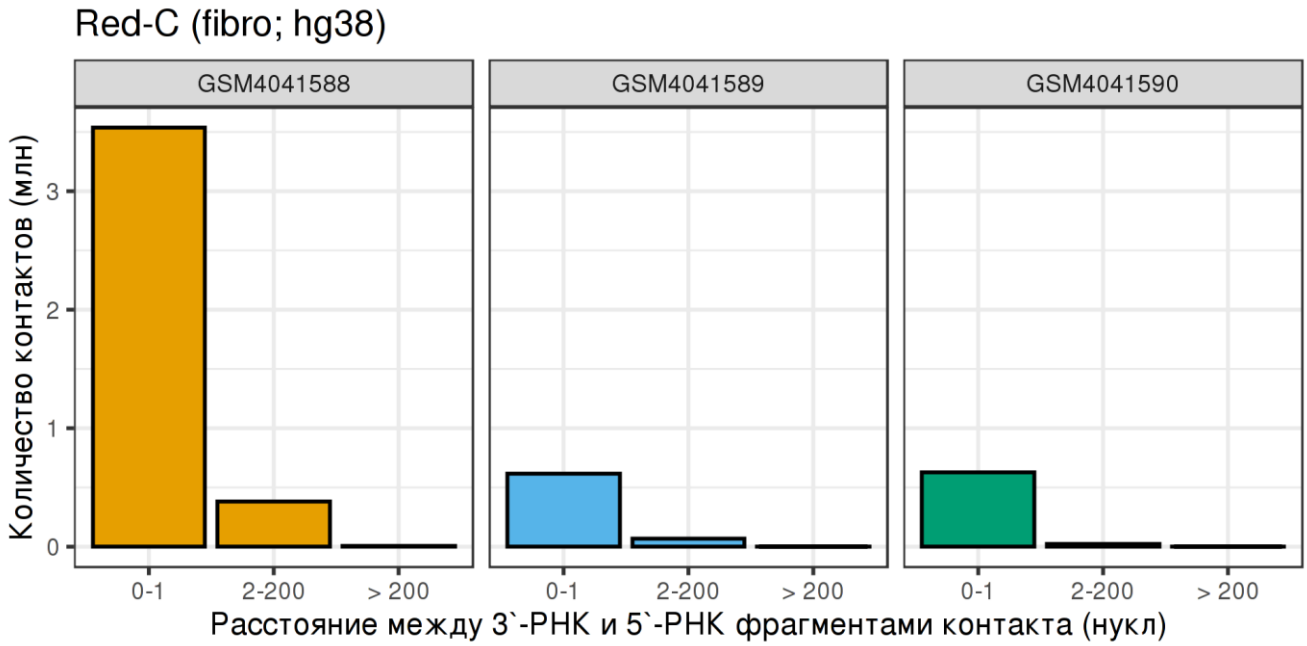

Рисунок 15. Распределение расстояний между 3`- и 5`-РНК частями из эксперимента Red-C (клеточная линия нормальных фибробластов), картированных на референсный геном человека версии hg38. Реплики указаны отдельно.

На рисунках 14 и 15 можно увидеть распределение расстояния между 3`- и 5`- фрагментами молекулы РНК, которую можно детектировать в РНК-ДНК контактах эксперимента Red-C. Во всех трех репликах, как в случае с K562, так и для фибробластов, видно, что в основном детектируемые фрагменты РНК довольно короткие, порядка 200 нуклеотидов, т.е. обе части РНК не сильно отличаются друг от друга, прямые и обратные прочтения практически полностью перекрывают друг друга. Для единообразия с другими экспериментами, где есть возможность изучать только один фрагмент РНК, далее мы работали только с 3`РНК-частью, примыкающей непосредственно к линкеру. Видимо, в ходе формирования химерной молекулы РНК-линкер-ДНК молекулы РНК подвергаются довольно сильным воздействиям и в значительной степени деградируют. Также эту гипотезу подтверждает факт потери большого количества контактов на этапе фильтрации слишком коротких РНК и ДНК фрагментов.

В случае картирования частей контактов на референсный геном человека версии hg19 получены аналогичные результаты (данные не приведены).

Сборка первичных РНК-ДНК контактов

Финальный шаг Этапа №1 предполагает получение файла, содержащего первичные РНК-ДНК контакты, где указаны координаты для РНК и ДНК частей относительно соответствующего референсного генома. Сборка первичных контактов заключалась в сопоставлении РНК и ДНК-частей, пришедших из одного чтения (по идентификатору чтения) и удовлетворяющих всем вышеописанным условиям.

На этом заканчивается Этап №1. Далее для унифицированной обработки файлов с контактами из разных протоколов важно, чтобы входные данные имели одинаковую структуру.

Для каждой хромосомы в отдельности (по РНК-части контактов) необходим файл, содержащий информацию о картировании РНК и ДНК части каждого контакта, независимо для каждой реплики при наличии таковых (табл. 5).

Таблица 5. Описание столбцов файла, содержащего информацию о каждом РНК-ДНК контакте. Такие файлы были одинаковым образом подготовлены для всех реплик каждого экспериментального протокола и служили входными данными для Этапа №2 биоинформатического протокола.

№	Название	Значение	Описание
1	read_ID	D00795:32:CA2K6ANXX:1:1104:8070:1930	Имя чтения
2	rna_chr	chr1	Координаты РНК-части: хромосома
3	rna_start	205149406	Координаты РНК-части: начало чтения
4	rna_end	205149482	Координаты РНК-части: конец чтения
5	rna_strand	+	Координаты РНК-части: цепь
6	rna_cigar	76M	Поле CIGAR для РНК-части
7	dna_chr	chr9	Координаты ДНК-части: хромосома
8	dna_start	26042113	Координаты ДНК-части: начало чтения
9	dna_end	26042137	Координаты ДНК-части: конец чтения
10	dna_strand	-	Координаты ДНК-части: цепь
11	dna_cigar	24M	Поле CIGAR для ДНК-части

Исследование первичных РНК-ДНК контактов

Начиная с Этапа №2 данные РНК-ДНК интерактома, полученные с помощью всех экспериментов (Red-C, GRID-seq, RADICL-seq) для всех клеточных типов и тканей, приведены к единому виду. Последующий анализ осуществлен над всеми данными абсолютно одинаковым образом вне зависимости от экспериментальной стратегии.

Обработка сплайсированных РНК-частей контактов

При картировании РНК-частей контактов на референсный геном была использована программа для картирования HISAT2, допускающая сплайсинг. Способ картирования и длина РНК-частей позволяла детектировать чтения, картированные с разрывом, будем называть такие чтения чтениями со сплайсингом.

Для поиска и обработки чтений со сплайсингом была проанализирована информация о картировании, представленная в поле CIGAR для РНК-части. Чтения могут быть картированы тремя способами:

- с полным совпадением с референсным геномом по всей длине чтения (CIGAR вида “25M”, где M - match) - такие чтения проходили далее без изменений;
- содержат один пропущенный интервал (CIGAR вида “30M65N10M”, где M - match, N - skipped region) - для таких чтений оставляли наиболее протяженный участок, картированный без разрывов (рис. 16);
- более сложные варианты картирования (чтения с сложным сплайсингом): несколько пропущенных интервалов (CIGAR вида: “8M1113N56M79N8M”), картирование с вставками или делециями - такие чтения были удалены.

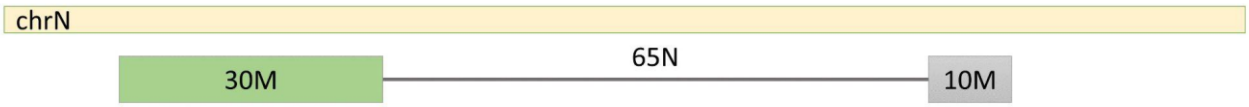

Рисунок 16. Схематичное изображение сплайсированного фрагмента РНК с одним пропущенным интервалом. Согласно процедуре были использованы координаты самого длинного непрерывно картированного фрагмента чтения (отмечено зеленым).

Удаленных чтений по причине детекции сложного сплайсинга в РНК-части оказалось очень мало (табл. 6). В GRID-seq сплайсированных контактов не оказалось совсем как и в целом детектированных случаев спласинга, для RADICL-seq сложный сплайсинг также не детектирован.

Таблица 6. Процент контактов с сплайсированными РНК-частями, включая случаи сложного сплайсинга.

Образец	Детектировано контактов с сплайсингом в РНК-части (%)	Детектировано контактов с сложным сплайсингом в РНК-части (%)
Red-C - hg19		
K562	1.6900	0.26
fibro	2.7000	0.40
Red-C - hg38		
K562	2.0000	0.26
fibro	3.6000	0.45
GRID-seq - hg38		
MDA-MB-231	0.0000	0.00
MM.1S	0.0000	0.00
GRID-seq - mm10		
ES	0.0000	0.00
RADICL-seq - mm10		
ES 1FA	0.0220	0.00
ES 2FA	0.0100	0.00
ES Act	0.0010	0.00
ES NPM	0.0550	0.00
OPC 1FA	0.0140	0.00
OPC NPM	0.0017	0.00

Метрики

На данном этапе мы располагаем полным набором РНК-ДНК контактов, прошедших все стадии технической обработки и полностью готовых для дальнейшего анализа. Для каждого контакта выделены корректно и уникально картированные на референсный геном РНК и ДНК части нужной длины. Все реплики подлежат независимой обработке. Обратим внимание на некоторые более подробные и разнообразные характеристики полученных контактов, которые позволят лучше понять, с чем именно предстоит работать.

Длины детектируемых РНК и ДНК-частей протоколов “все-против-всех” определяются в процессе экспериментальных процедур, исходя из особенности работы ферментов, используемых при обработке хроматина и химерных РНК:линкер:ДНК последовательностей. Также в ходе биоинформатической обработки длины частей могут претерпевать изменения при удалении нуклеотидов низкого качества. Шаг 3 Этапа №1 не пропускает РНК и ДНК части короче 14 нуклеотидов.

На рисунках 17-19 представлены распределения длин РНК и ДНК-частей контактов в исследуемых протоколах “все-против-всех”.

Как видно из рисунка 17 на примере Red-C (K562) в подавляющем большинстве случаев длина ДНК-частей контактов составляет 23-24 нуклеотида, что следует из протокола с учетом достроенных сайтов рестрикции. Длины РНК-частей варьируют в зависимости от модели секвенатора и протокола секвенирования. Реплики ведут себя согласованно. Результаты, полученные для разных версий референсного генома человека, не противоречат друг другу, хотя и имеют незначительные отличия в количестве контактов (данные не приведены).

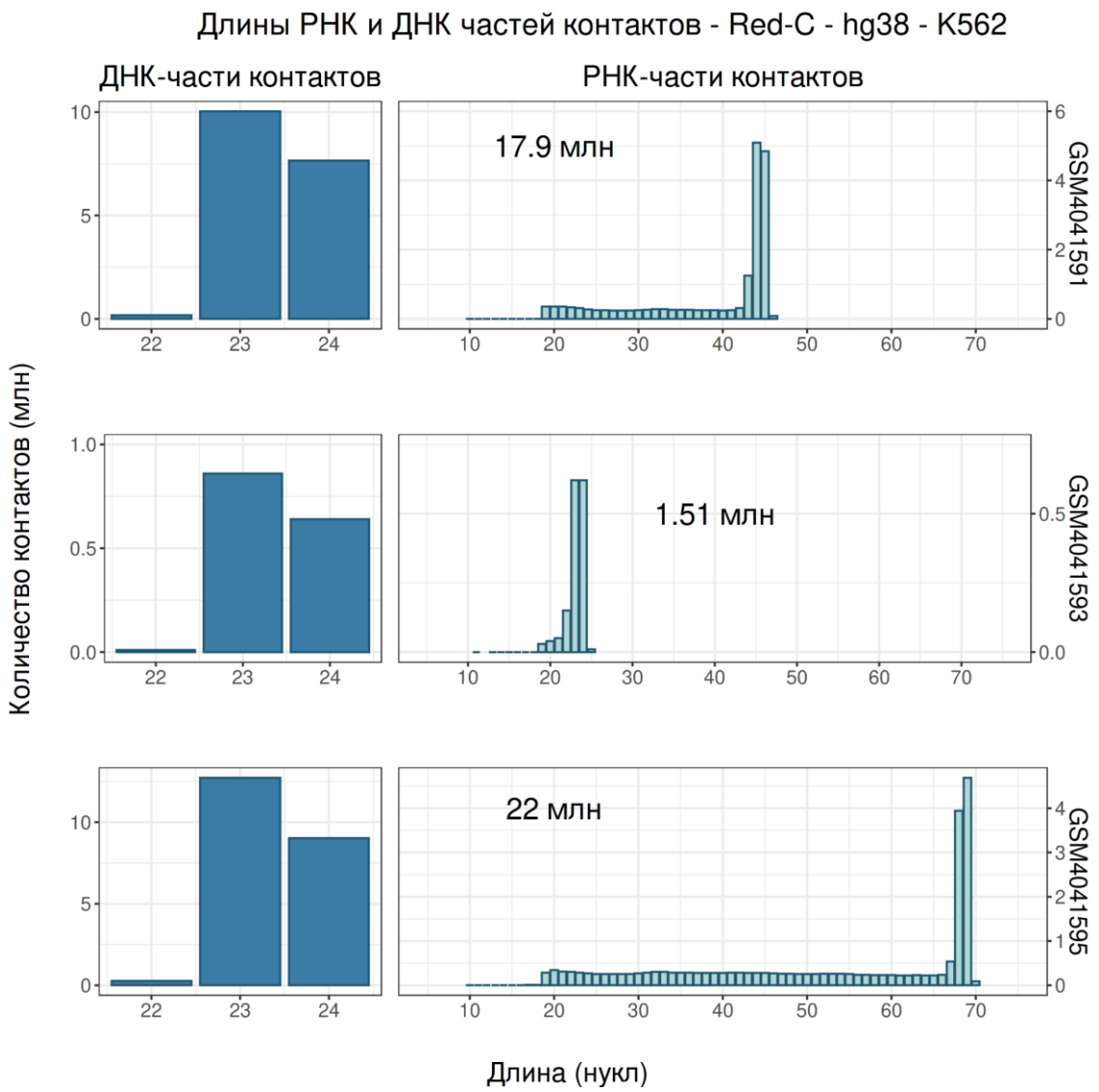

Рисунок 17. Распределение длин РНК и ДНК-частей контактов из протокола Red-C для клеточной линии K562 (картирование на референсный геном версии hg38). Все реплики представлены отдельно.

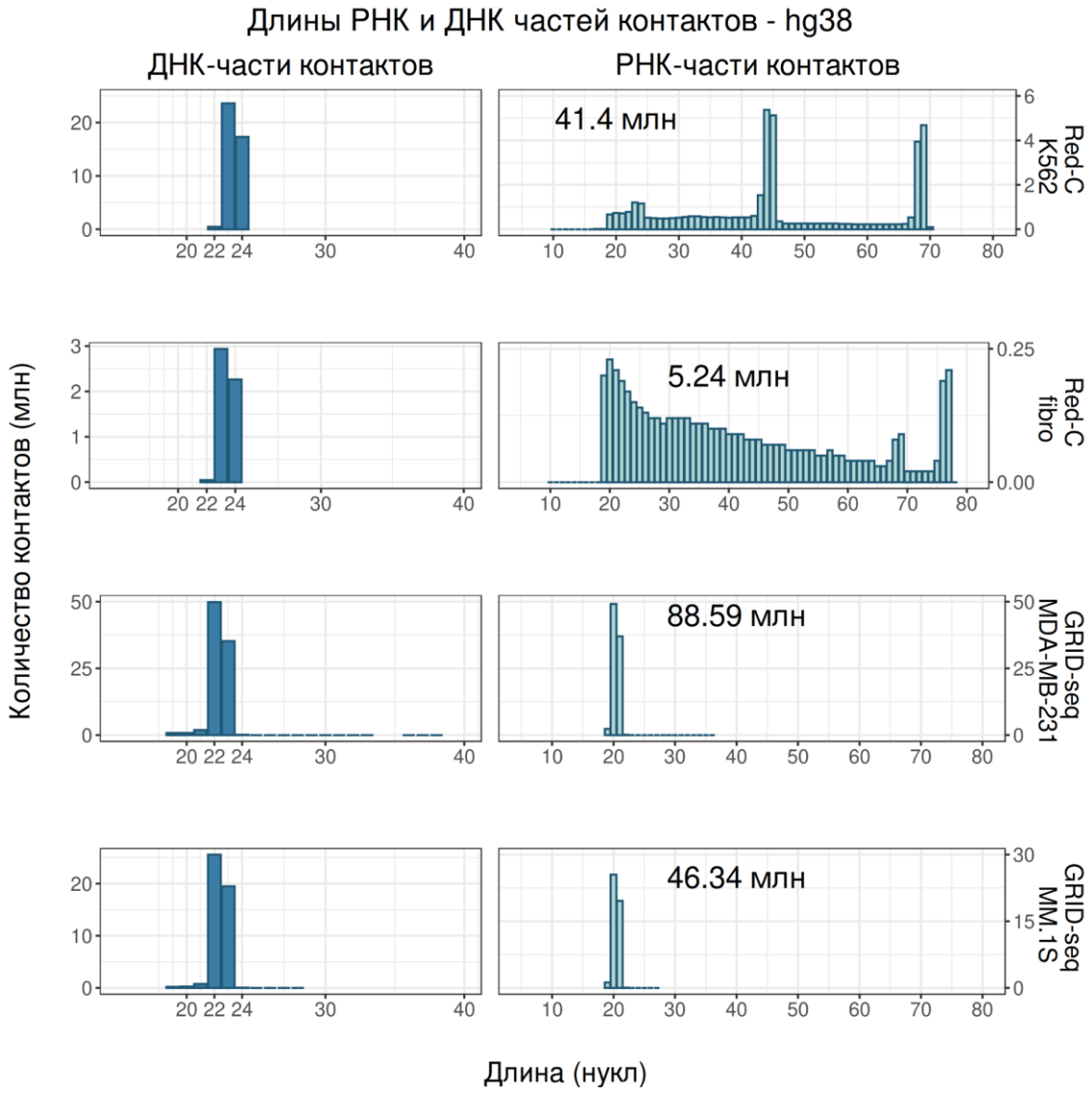

Рисунок 18. Распределение длин РНК и ДНК-частей контактов для протоколов GRID-seq и Red-C (картирование на референсный геном версии hg38), реплики объединены.

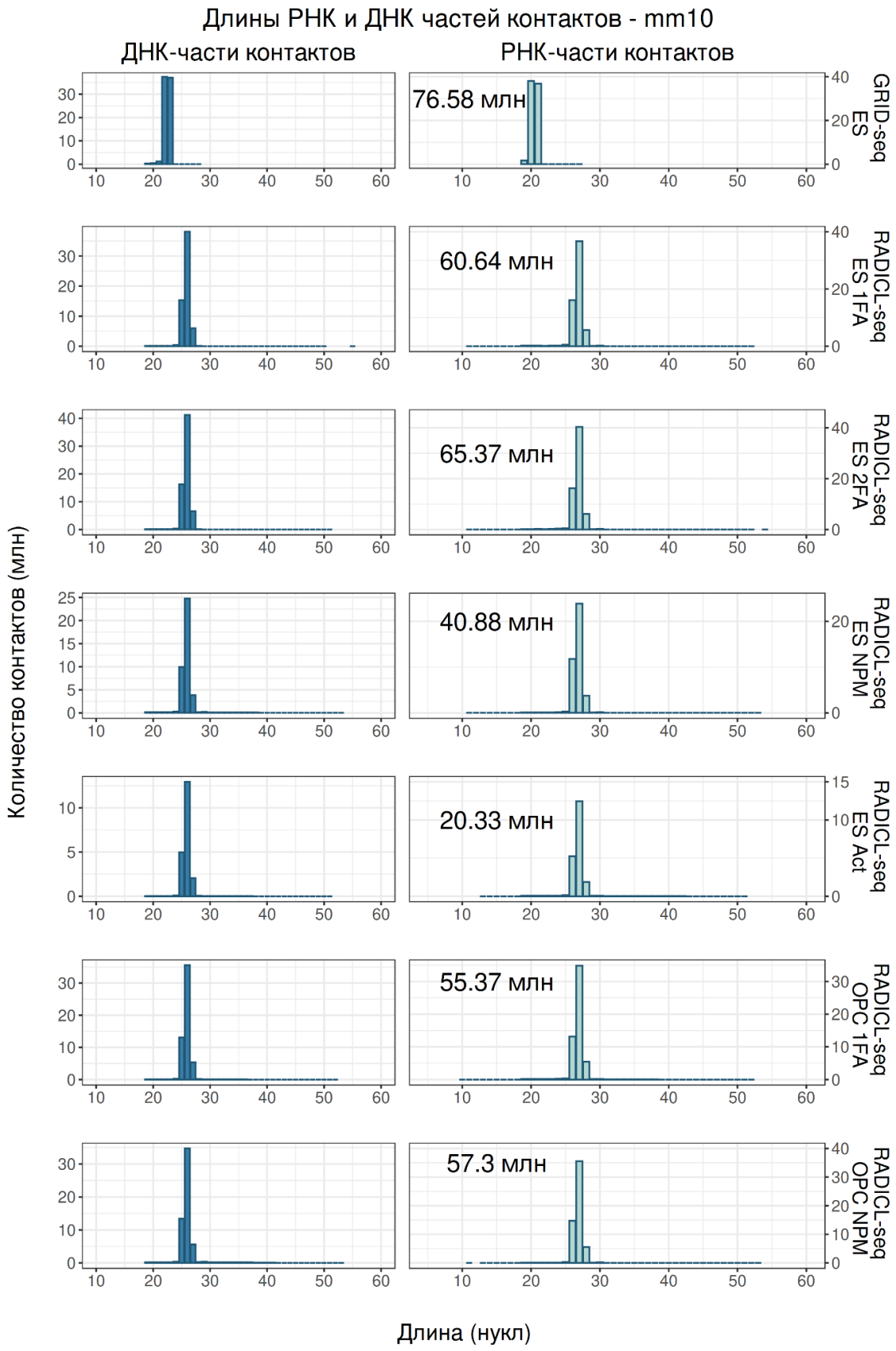

Рисунок 19. Распределение длин РНК и ДНК-частей контактов для протоколов GRID-seq и RADICL-seq (картирование на референсный геном версии mm10), реплики объединены.

Как видно из рисунков 17-19 длины РНК-частей контактов в протоколе Red-C достигают 70-80 нуклеотидов, во всех остальных экспериментах РНК-части более чем в два раза короче и не превышают 30 нуклеотидов. Длины ДНК-частей контактов во всех протоколах сопоставимы и находятся в диапазоне 23-28 нуклеотидов.

Каждый контакт представляет собой два интервала с координатами на референсном геноме. Один из них соответствует локусу, порождающему РНК-часть, т.е. ген, с которого была синтезирована РНК, а другой - участок ДНК, с которым эта РНК взаимодействует. Для каждого такого контакта можно определить расстояние между местом синтеза РНК и местом контактирования. Контакты можно разделить на внутрихромосомные, где РНК контактирует с хромосомой, на которой закодирован ее ген, и на межхромосомные, где РНК контактирует с другими хромосомами (“не материнскими”). Внутрихромосомные расстояния между РНК и ДНК-частями поделим на несколько интервалов: менее 1000 нуклеотидов (< 1Кб) от своего гена, от 1Кб до 10Кб, от 10Кб до 100 Кб, от 100 Кб до 1 Мб (1 мегабаза = 1 млн нуклеотидов) и более 1Мб. Посмотрим, как распределены по этим интервалам детектируемые контакты (в процентах) (рис. 20).

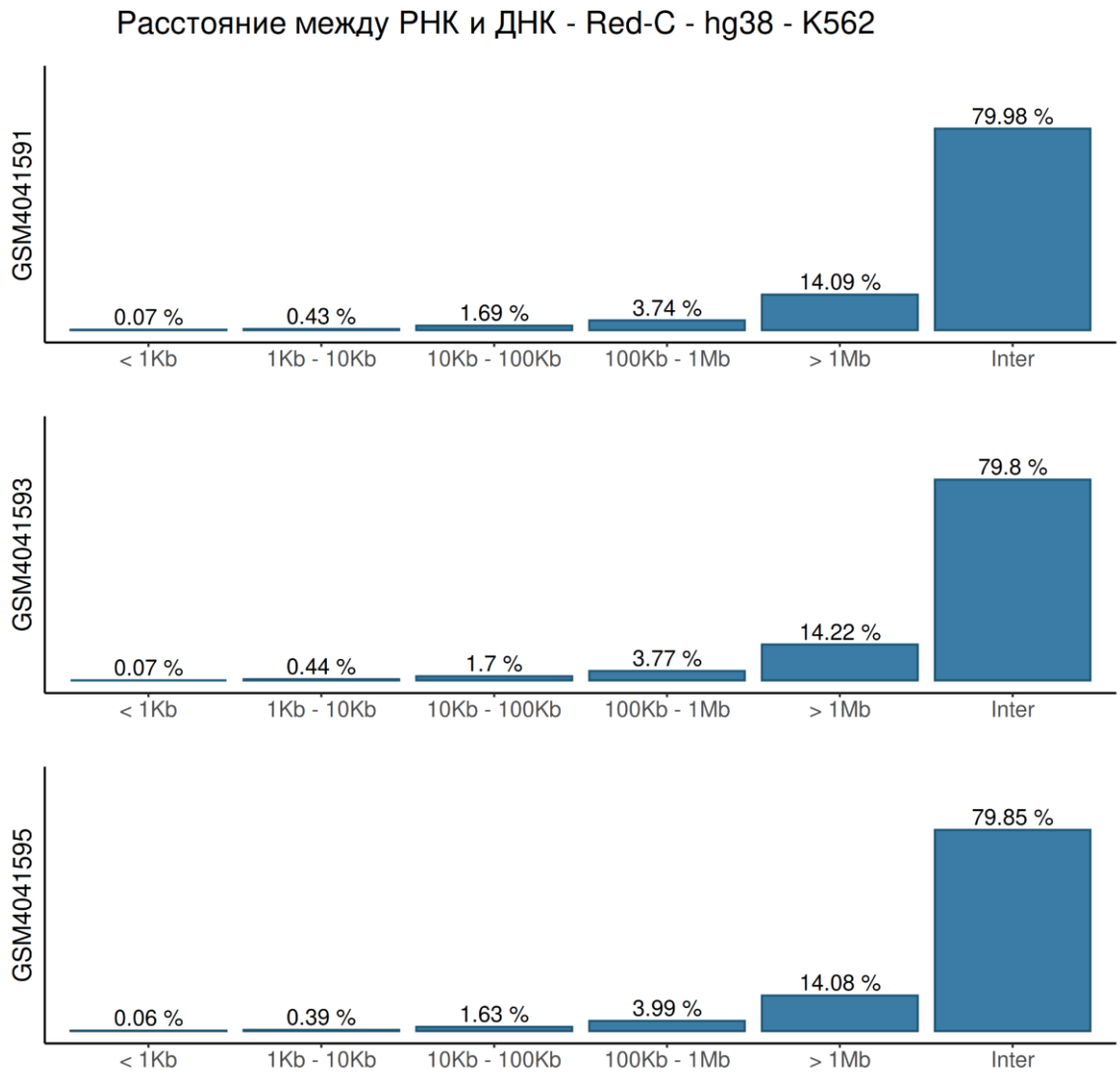

Рисунок 20. Распределение количества РНК-ДНК контактов в зависимости от расстояния между РНК и ДНК-частями (в процентах). Протокол Red-C, клеточная линия K562, картирование на референсный геном версии hg38. Все реплики представлены отдельно.

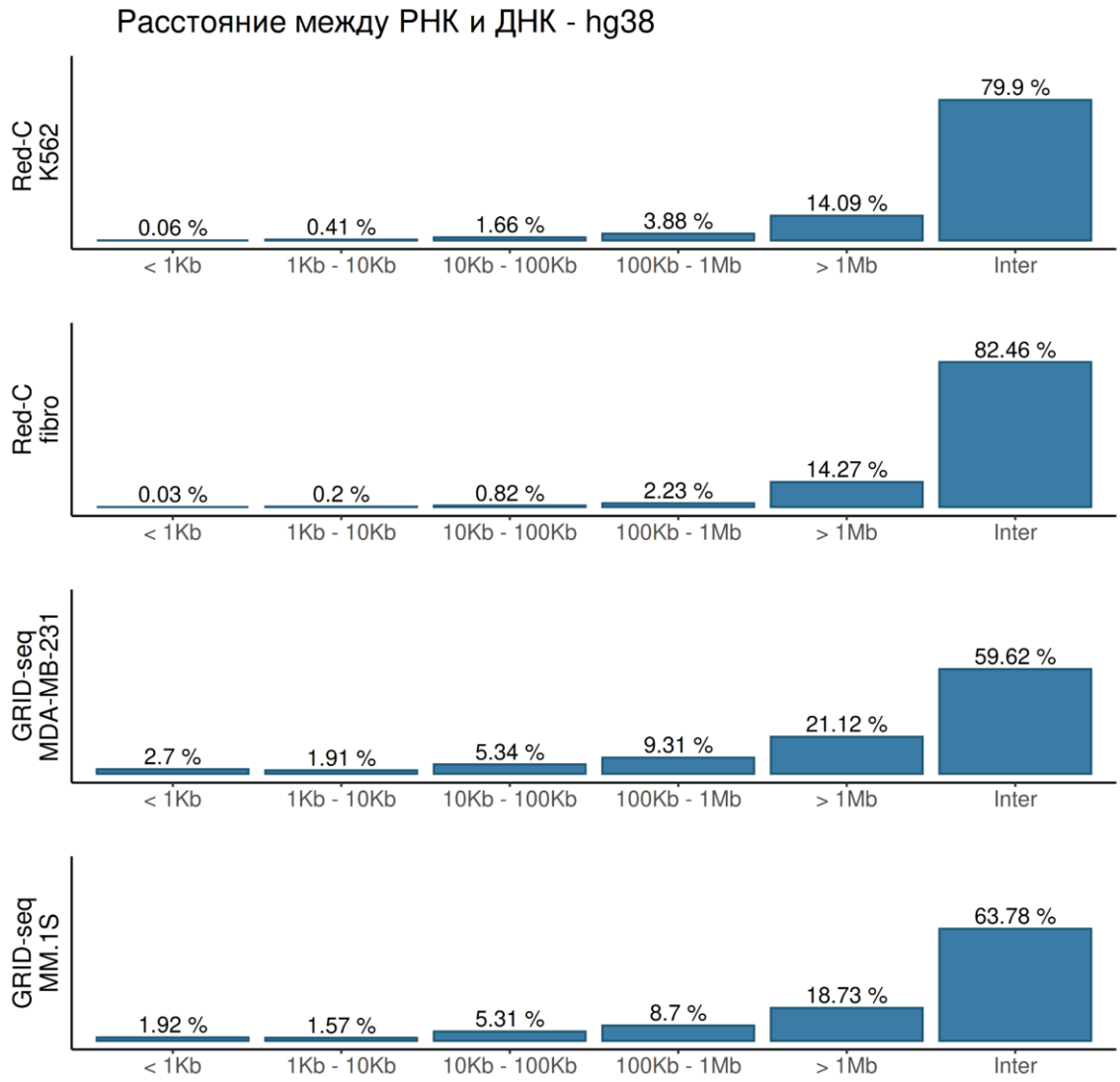

Рисунок 21. Распределение количества РНК-ДНК контактов в зависимости от расстояния между РНК и ДНК-частями (в процентах). Протоколы Red-C и GRID-seq, картирование на референсный геном версии hg38. Все реплики объединены.

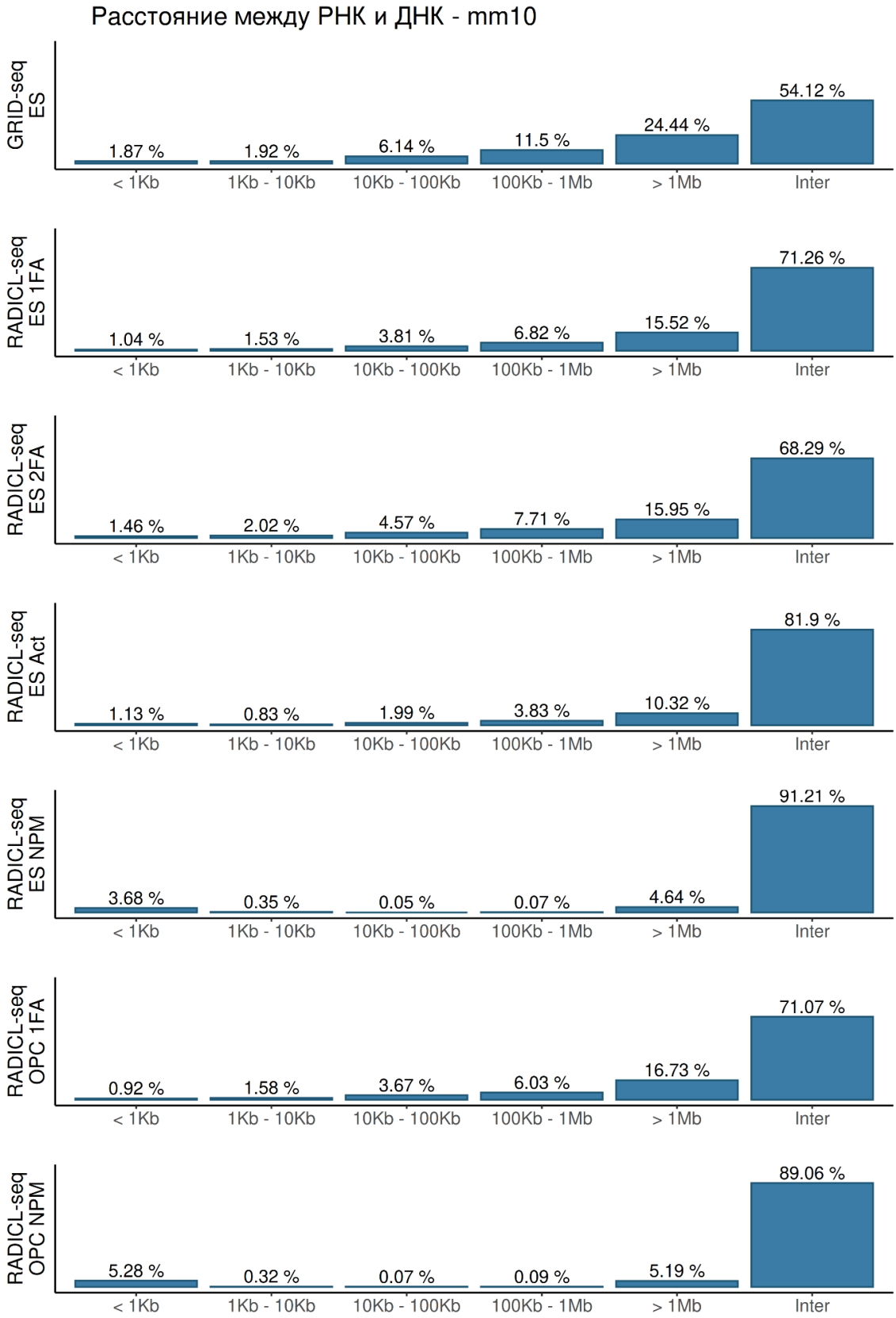

Рисунок 22. Распределение количества РНК-ДНК контактов в зависимости от расстояния между РНК и ДНК-частями (в процентах). Протоколы GRID-seq и RADICL-seq, картирование на референсный геном версии mm10. Все реплики объединены.

Для всех протоколов более половины контактов определены как межхромосомные, однако наибольший процент наблюдается для эксперимента Red-C, а также для протокола RADICL-seq в случае обработок NPM и актиномицином Д (Act) (рис. 20-22).

Рассмотрим еще более подробно только внутрихромосомные контакты. Для каждого контакта хромосому, с которой пришла РНК-часть контакта, разобьем на непересекающиеся участки переменной длины так, что вокруг РНК-части, выделим короткие локусы в 100 нуклеотидов, далее с увеличением расстояния от РНК-части увеличим и длины локусов в 2 раза, пока не покроем хромосому целиком. Теперь для каждого такого интервала можно рассчитать количество и плотность попавших в них контактов, не различая локусы, находящиеся на одинаковом расстоянии в 5`- и 3`-областях относительно РНК-части (рис. 23).

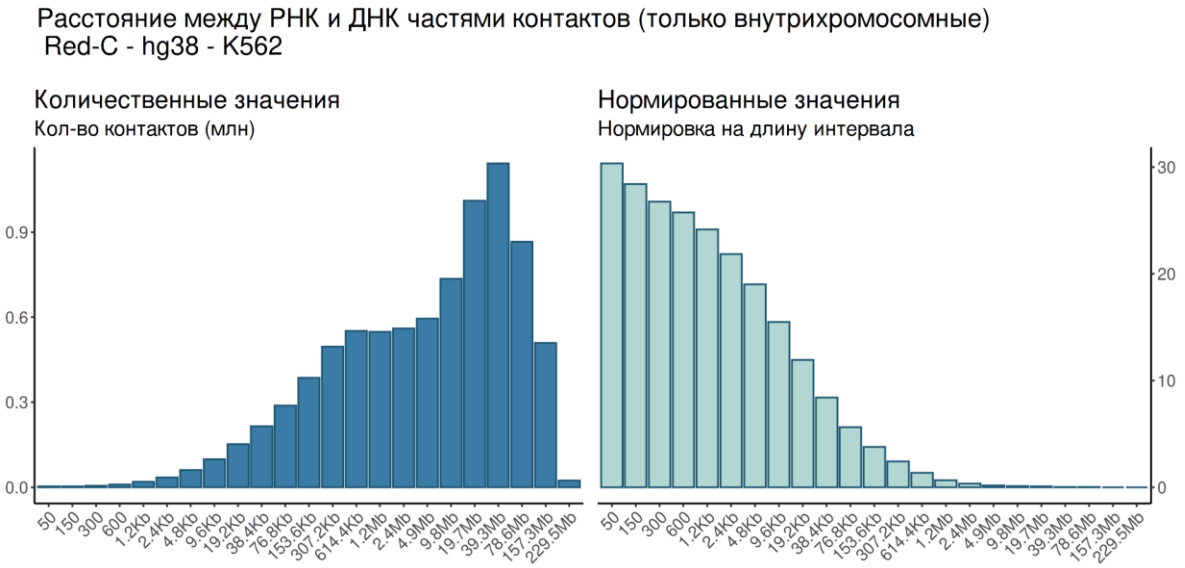

Рисунок 23. Распределение расстояния между РНК и ДНК частями внутрихромосомных контактов. Слева: абсолютные значения количества контактов; справа: плотность контактов (для каждого интервала количество контактов нормирована на длину интервала). Протокол Red-C, клеточная линия K562, картирование на референсный геном K562.

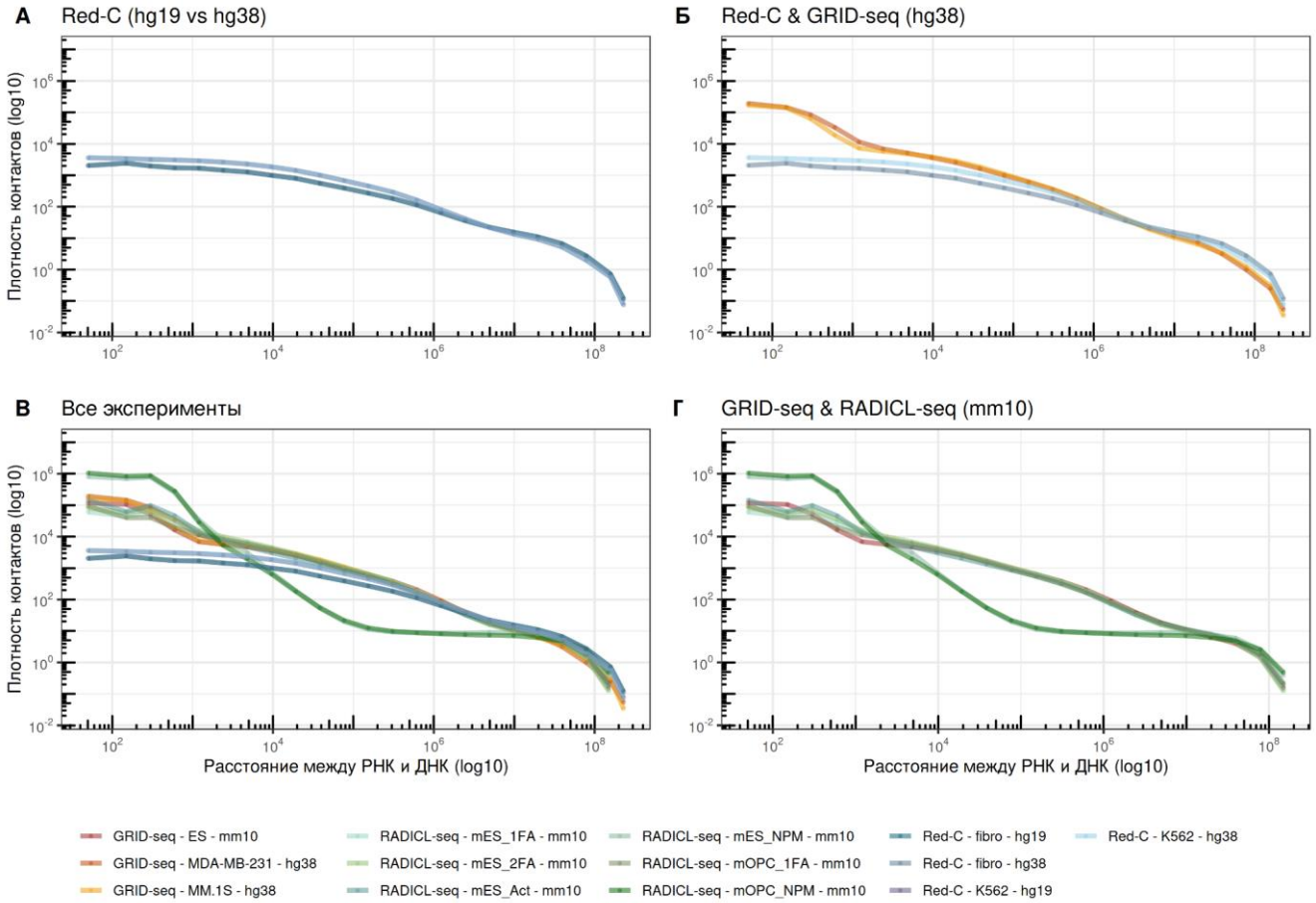

Рисунок 24. Зависимость плотности контактов РНК от расстояния между РНК-частью и местом контактов для внутрихромосомных контактов. (А) Протокол Red-C, клеточные линии K562 и фибробласты, картирование на референсные геномы версий hg19 и hg38. (Б) Протоколы Red-C и GRID-seq, картирование на референсный геном версии hg38. (В) Все эксперименты. (Г) Протоколы GRID-seq и RADICL-seq, картирование на референсный геном версии mm10.

Для всех протоколов наибольшую плотность контактов можно наблюдать рядом с РНК-частью, с увеличением расстояния от РНК-части, плотность контактов снижается (рис. 23, 24). Для экспериментов RADICL-seq можно увидеть резкий спад плотности контактов РНК с хроматином на расстоянии ~100Кб от РНК-части, причем только для образцов NPM (рис. 24Г). Это же наблюдение отмечают авторы статьи, что подтверждает возможность применения представленного протокола к другим данным типа “все-против-всех” без драматических искажений результатов.

Для каждой хромосомы посчитаем число контактов отдельно по РНК-части и ДНК-части.

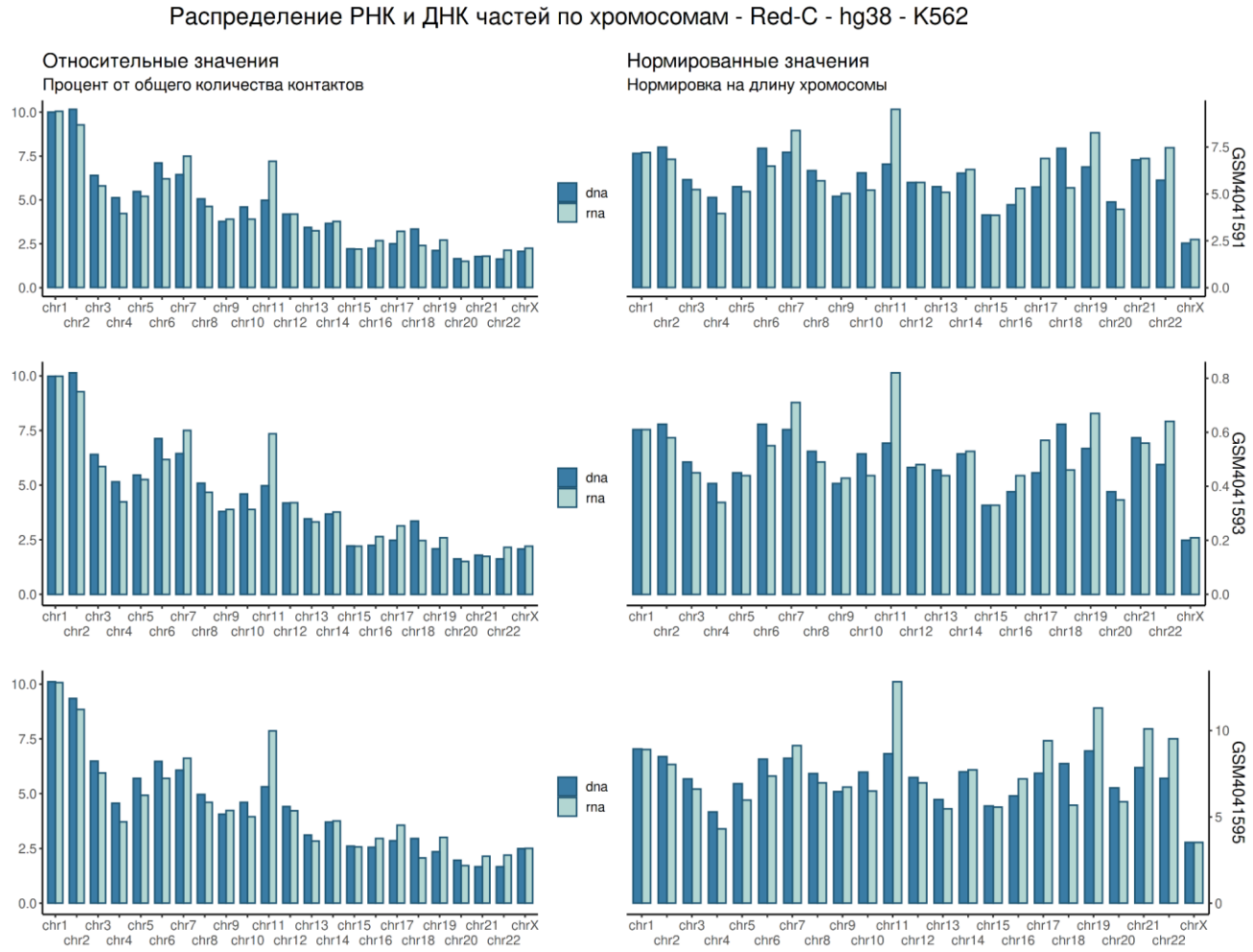

Рисунок 25. Распределение контактов по хромосомам отдельно для РНК-частей и для ДНК-частей. Протокол Red-C, клеточная линия K562, референсный геном версии hg38. Относительные значения - процент контактов для каждой хромосомы. Нормированные значения - количество контактов нормировано на длину хромосомы в нуклеотидах. Все реплики указаны отдельно.

Можно отметить, что есть хромосомы, которые порождают больше РНК-частей, чем содержат локусов ДНК, с которыми контактируют хаРНК (например, хромосома 11). Обратная ситуация также имеет место быть, хотя и менее выражена (рис. 25). Реплики для Red-C ведут себя аналогичным образом.

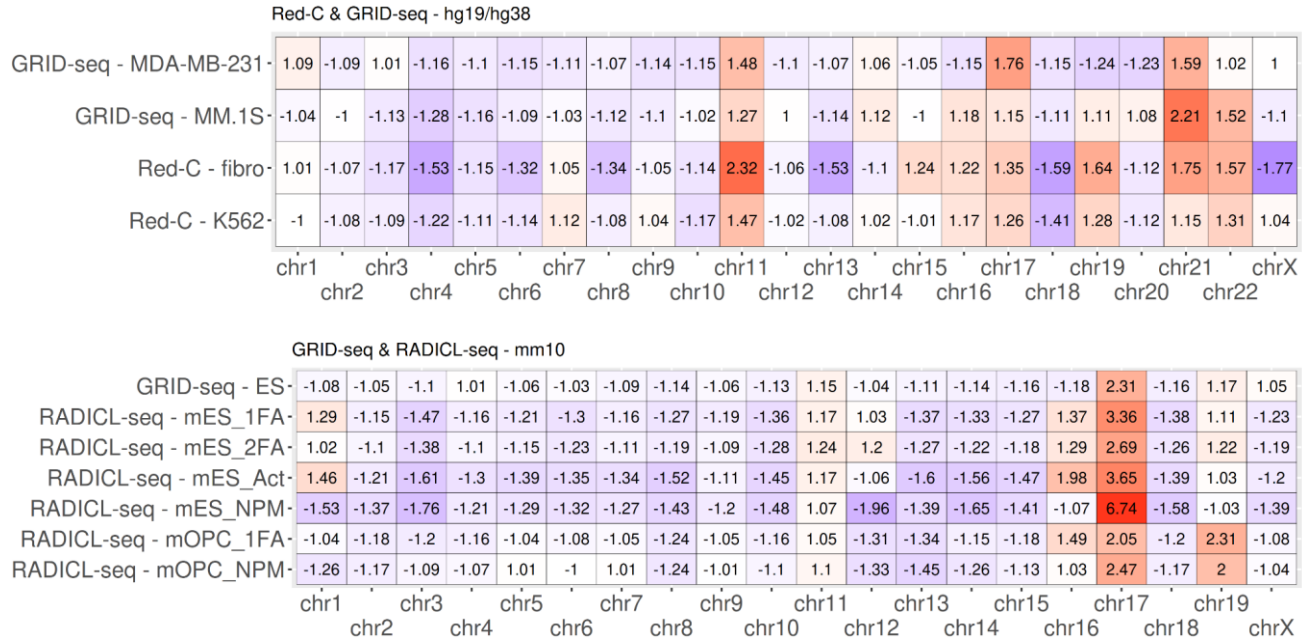

Рисунок 26. Отношение количества РНК-частей к количеству ДНК-частей, пришедших с каждой хромосомы. Сверху: протоколы Red-C и GRID-seq, картирование на референсный геном версии hg38; снизу: протоколы GRID-seq и RADICL-seq, картирование на референсный геном версии mm10. Все реплики объединены. Отрицательные значения покрашены градиентом синего - хромосома больше контактирует, чем порождает контактирующие РНК. Положительные значения покрашены градиентом красного - хромосома больше порождает хроматин ассоциированные РНК, чем контактирует с другими РНК.

Из рисунка 26 видно, что в целом вне зависимости от протокола хромосомы демонстрируют одинаковую тенденцию, хотя есть и исключения. Для протоколов, реализованных на человеческих клеточных линиях, можно выделить хромосомы 11, 17, 21 и 22 с которых приходит много РНК, взаимодействующих с хроматином, тогда как с хромосомами 18 и 4 наоборот много контактируют. Для экспериментов, реализованных на мышиных клеточных линиях, бросается в глаза хромосома 17, которая порождает большое количество РНК, взаимодействующих с хроматином.

Удаление ДНК-частей контактов из BlackList

Авторы протокола RADICL-seq предлагают исключать из анализа контакты, которые ДНК-частью попадают в участки из ENCODE BlackList, что и было сделано. Данная процедура может вызывать сомнения, т.к. авторы создавали Blacklist на основании данных экспериментов ChIP-seq, чтобы в итоге удалить

артефактные сигналы, попадающие в основном в области генома, содержащие простые повторы. Тем не менее, как видно из таблицы 7, контакты, попадающие в Blacklist своими ДНК-частями составляют практически для всех экспериментов менее одного процента.

Таблица 7. Количество контактов, попавших ДНК-частью в разметку BlackList для соответствующего референсного генома (в процентах).

Образец	Попало в blacklist (%)
Red-C - hg19	
K562	0.21
fibro	0.26
Red-C - hg38	
K562	0.12
fibro	0.16
GRID-seq - hg38	
MDA-MB-231	0.41
MM.1S	0.29
GRID-seq - mm10	
ES	0.24
RADICL-seq - mm10	
ES 1FA	0.50
ES 2FA	0.44
ES Act	0.33
ES NPM	1.47
OPC 1FA	0.36
OPC NPM	0.88

Аннотация РНК-частей контактов генами

Чтобы установить, какие именно РНК были детектированы как контактирующие с ДНК, было необходимо аннотировать РНК-часть с помощью подготовленной нами разметки генов (см. раздел “Материалы и методы”).

Координаты генов человека из используемой нами сборной разметки суммарно покрывают ~60% генома (для мыши ~ 47%). Внутри разметки координаты генов часто пересекаются между собой по одноименной цепи (иллюстративный пример на рис. 26; сводная статистика на табл. 8), все перекрытия подсчитаны для “плюс” и “минус”-цепей независимо.

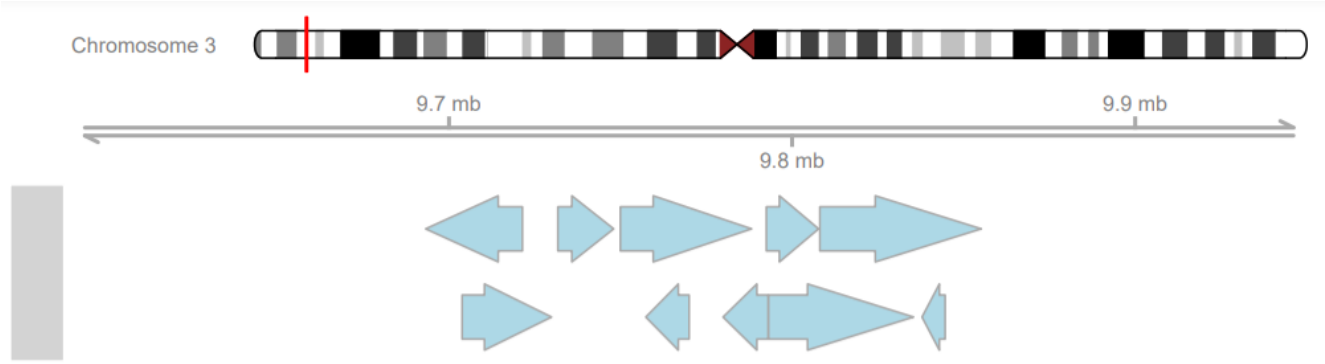

Рисунок 27. Фрагмент хромосомы 3 генома человека (hg38). Синими стрелками показаны гены, закодированные на представленном фрагменте, направление стрелок соответствует ориентации гена.

Таблица 8. Количество нуклеотидов, покрытых генами. Данные представлены для референсных геномов человека (hg19, hg38) и мыши (mm10). Указан процент от суммарной длины генной разметки с неоднозначной аннотацией по одной цепи.

Версия генома	Гены покрывают "плюс"-цепь, млрд нукл	Гены покрывают "минус"-цепь, млрд нукл	Разметка с перекрытием, %[1]
hg19	1.04	0.98	10.52
hg38	0.96	0.92	9.57
mm10	0.65	0.62	2.40

[1] подсчитано независимо по цепям

Для каждого контакта была реализована процедура первичной аннотации РНК-части, в ходе которой были найдены все гены, пересекающие по координатам РНК-часть контакта вне зависимости от цепи.

Протоколы “все-против-всех” в большинстве своем сохраняют информацию о том, на какой цепи был закодирован ген хаРНК. Однако, исходя из особенностей эксперимента, цепь РНК-части, полученная в результате картирования на референсный геном, либо соответствует цепи гена, либо должна быть заменена на противоположную. Решение о сохранении или изменении цепи РНК-части принималось на основании результатов дополнительной обработки информации, полученной после первичной аннотации. Для каждого эксперимента мы рассчитывали отношение количеств случаев, когда цепь РНК-части и

пересекающего ее гена не совпадают (Diff), к таковым, когда они совпадают (Same). Если отношение Diff/Same оказывалось больше 1, то цепь РНК-части следовало заменить на противоположную, в противном случае цепь РНК-части соответствовала цепи гена из аннотации. Как видно из рисунка 28 в случае протокола Red-C все реплики ведут себя схожим образом, а результаты свидетельствуют о том, что цепь РНК-части должна быть изменена (Diff/Same > 1 для всех хромосом). Среди всех рассматриваемых протоколов “все-против-всех” цепь РНК-части сохранена только для эксперимента RADICL-seq (рис. 29).

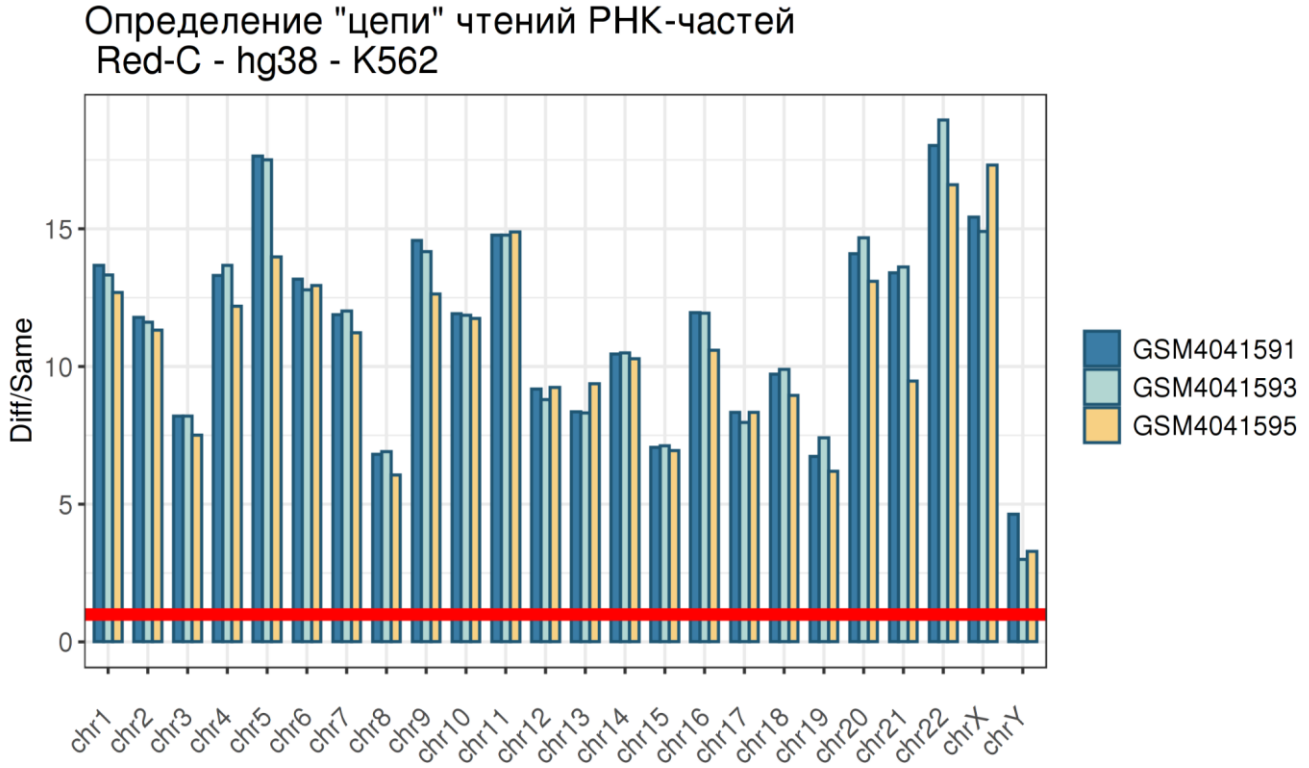

Рисунок 28. Отношения количества случаев, когда после первичной аннотации цепи РНК-части контакта и гена не совпадали (Diff) к случаям, когда цепи РНК-части и гена совпадали (Same). Красная линия соответствует значению Diff/Same = 1. Протокол Red-C, клеточная линия K562, реплики и хромосомы представлены отдельно.

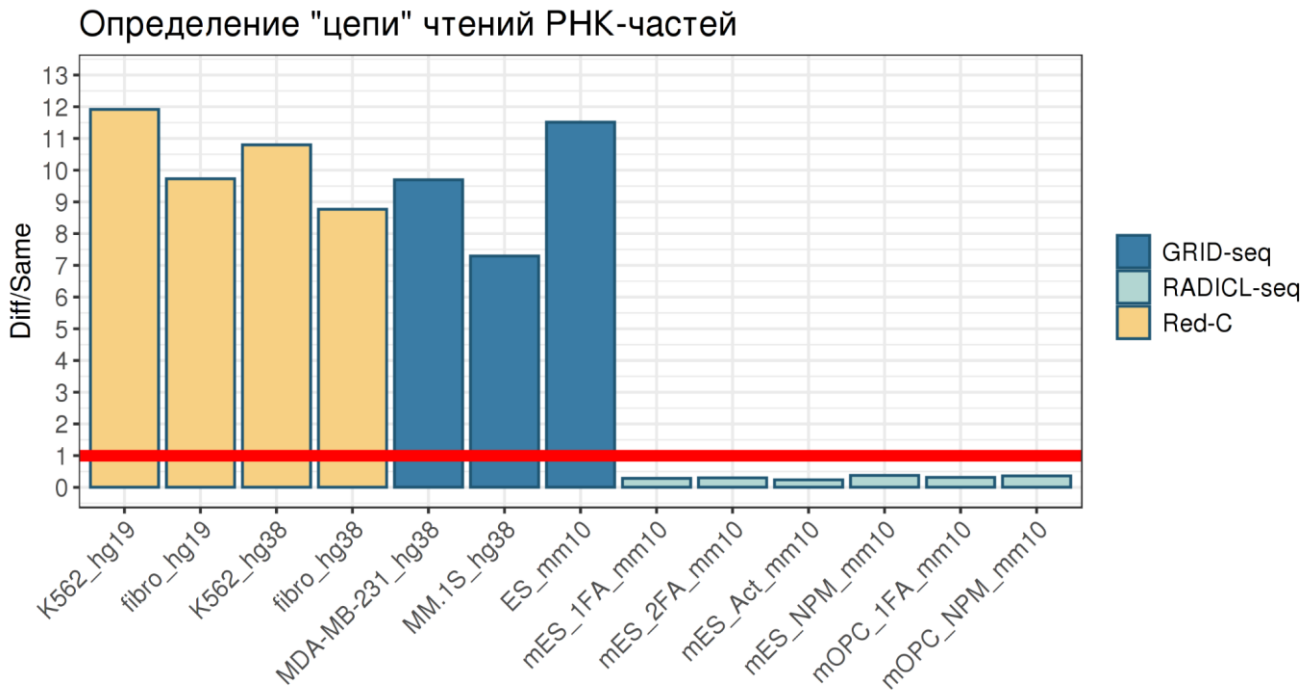

Рисунок 29. Отношения количества случаев, когда после первичной аннотации цепи РНК-части контакта и гена не совпадали (Diff) к случаям, когда цепи РНК-части и гена совпадали (Same). Красная линия соответствует значению Diff/Same = 1. Реплики объединены.

Это важных шаг протокола, пренебрежение которым не позволит корректно проаннотировать РНК-части контактов выбранной генной разметкой с учетом цепи. В результате для каждой РНК-части были оставлены только такие пересекающие ее гены, которые несут нужную цепь, согласно вышеописанной процедуре.

После первичной аннотации и процедуры определения корректной цепи РНК-часть контакта все равно могла быть аннотирована сразу несколькими генами по причине пересечения их координат, как это обсуждалось выше. Для разрешения случаев попадания РНК-части контакта на пересечение двух и более генов была разработана процедура голосования. Перед голосованием данные о всех репликах внутри одной клеточной линии для каждого протокола были объединены для увеличения покрытия. Для каждого гена независимо была рассчитана плотность покрытия, как количество РНК-частей, попавших в ген, деленное на полную длину гена. Затем были найдены РНК-части, приписанные сразу нескольким генам. В таких случаях для РНК-части оставляли только один ген, для которого плотность

покрытия оказалась наибольшей. Такой подход позволяет разрешать случаи, когда в длинном гене, например мРНК, закодированы малые нкРНК, которые и дают основной сигнал (рис. 30).

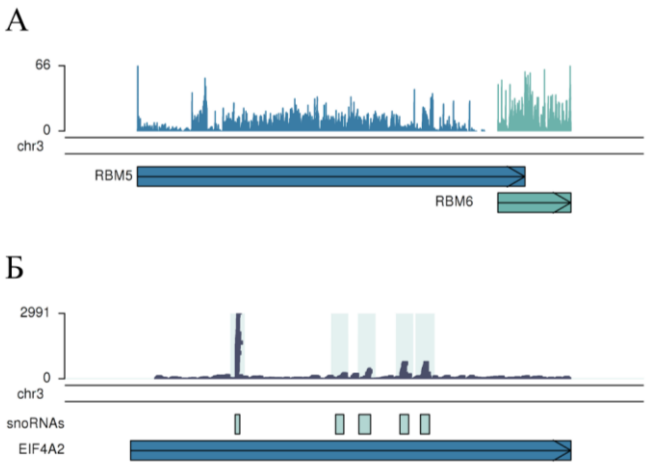

Рисунок 30. Примеры реализации процедуры голосования при разрешении неоднозначной аннотации. Гены обозначены разноцветными прямоугольниками. Чтения, аннотированные одним из генов покрашены соответствующим цветом (А) или выделены цветной рамкой (Б).

В ситуации с конфликтной аннотацией в случае протоколов Red-C и GRID-seq попала пятая часть всех контактов, для протокола RADICL-seq доля контактов, принимавших участие в голосовании достигала 40% (рис. 31).

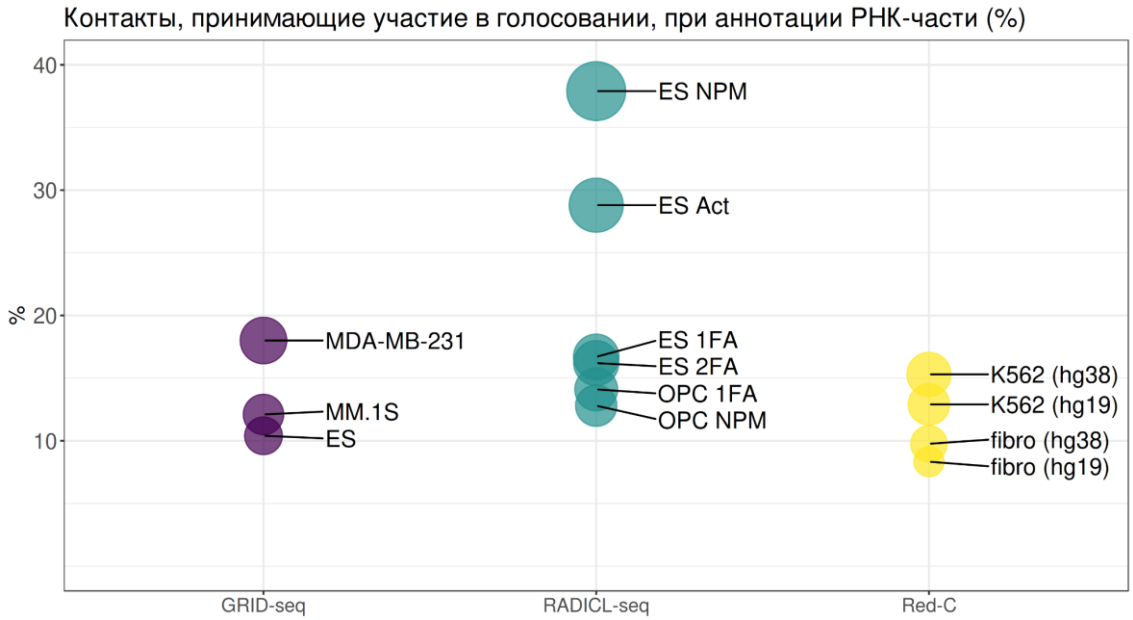

Рисунок 31. Процент РНК-частей, участвовавших в голосовании.

По итогам голосования в генную разметку попало более 90% контактов по РНК-части для Red-C, для остальных протоколов 75-90% (рис. 32).

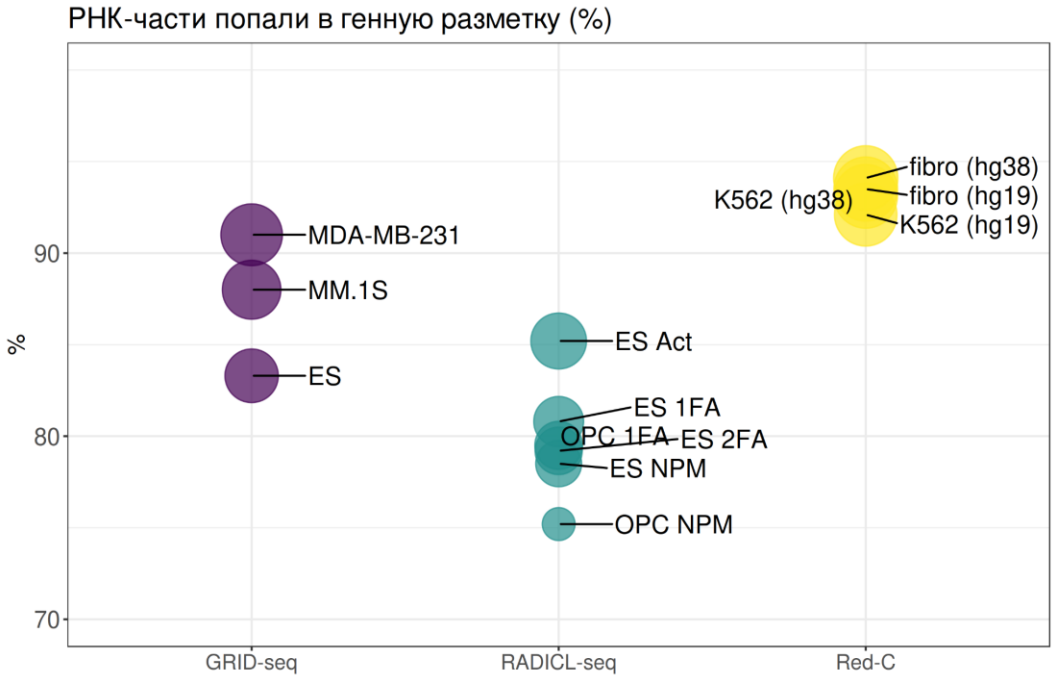

Рисунок 32. Процент контактов, где РНК-часть попала в границы существующей генной аннотации для соответствующей версии референсного генома; после процедуры голосования.

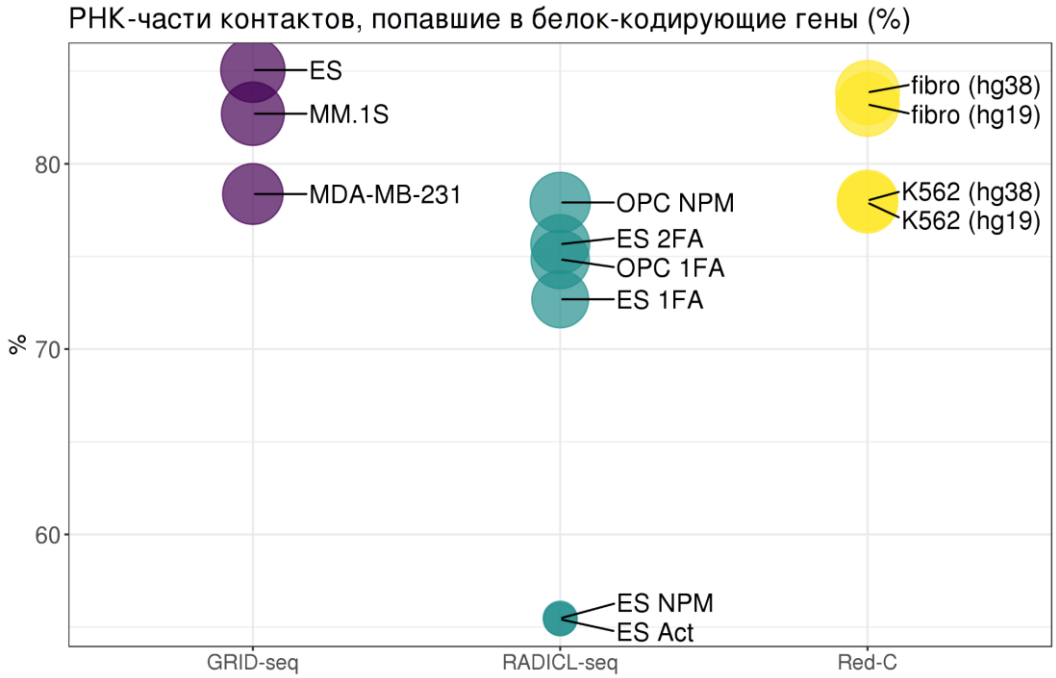

Рисунок 33. Процент контактов, где РНК-часть попала в границы мРНК.

Во всех случаях большинство РНК-частей оказалось в границах белок-кодирующх генов (рис. 33). Такое же наблюдение фиксируют авторы всех методов “все-против-всех”, рассуждая о большом количестве контактов мРНК.

Интересно отметить, что нкРНК NEAT1 и MALAT1 входят в топ 10 наиболее контактирующих РНК в протоколе Red-C для обеих клеточных линий. Суммарно для них детектировано более 1 млн контактов (K562). Гены этих нкРНК закодированы на хромосоме 11, которая была замечена как порождающая большое количество РНК-частей (рис. 26). В случае протокола GRID-seq, MALAT1 также содержит большое количество контактов, однако наиболее контактирующими являются более десятка малых ядерных РНК, гены которых расположены на 17 хромосоме (рис. 26).

Удаление контактов рибосомальных РНК

Дополнительно были удалены контакты, РНК-часть которых пришла с генов аннотированных как рибосомальные РНК. Для протоколов Red-C и GRID-seq рРНК породили ~5% контактов, а в случае RADICL-seq доля контактов рРНК достигала 30% (рис. 34). Стоит отметить, что в случае экспериментов, реализованных на мышиных клеточных линиях, особенно для метода RADICL-seq было замечено большое скопление высоко контактирующих рРНК, пришедших с хромосомы 17. Этот факт может объяснить наблюдение о том, что в случае мышиных клеточных линий хромосома 17 была отмечена, как наиболее богатая на РНК, контактирующие с хроматином (рис. 29).

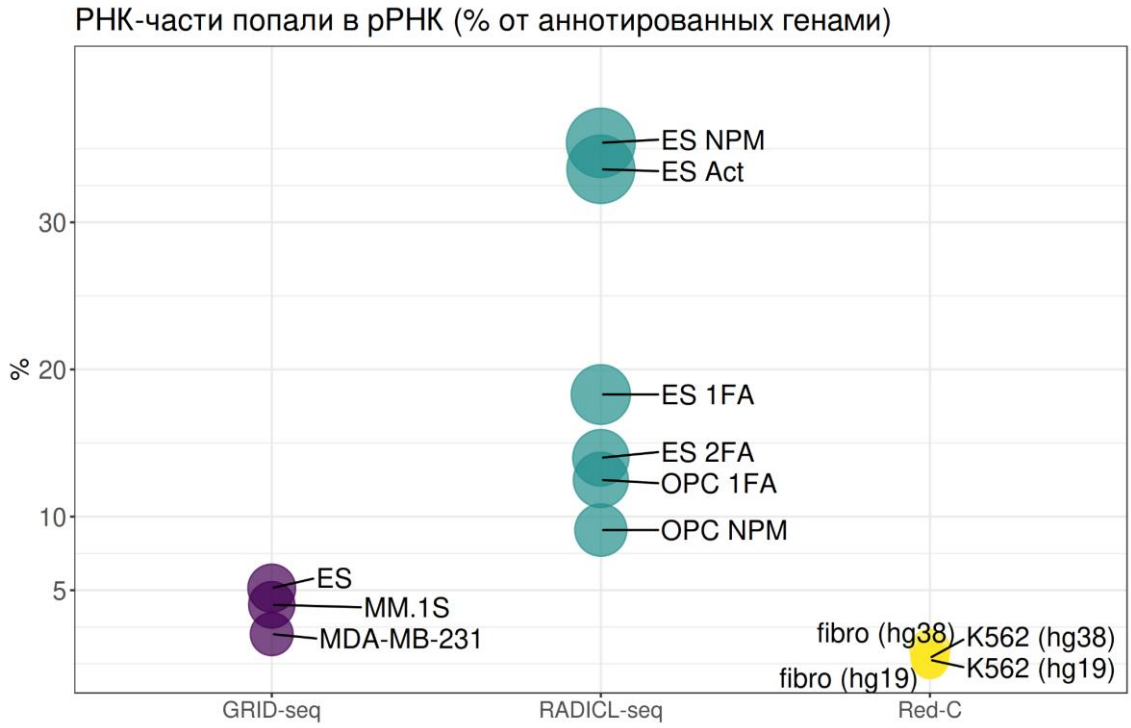

Рисунок 34. Процент контактов, где РНК-часть попала в границы генов, аннотированных как рРНК.

Сборка новых РНК, не представленных в генной разметке

РНК-части, которые не попали в генную разметку, были обработаны отдельно с целью поиска возможных новых РНК, которые не попадают в разметку генов, т.к. могут находиться в плотной связке с хроматином. Неаннотированные РНК-части были выделены и кластеризованы по координатам так, чтобы расстояние между РНК-частями не превышало 100 нуклеотидов, а в итоговый кластер попало не менее 100 РНК-частей. Для каждого эксперимента было собрано несколько тысяч таких кластеров, которым было дано общее название “X-РНК” (рис. 35). Полученные X-РНК добавлены к генной разметке, а попавшие в них контакты возвращены в общую выборку.

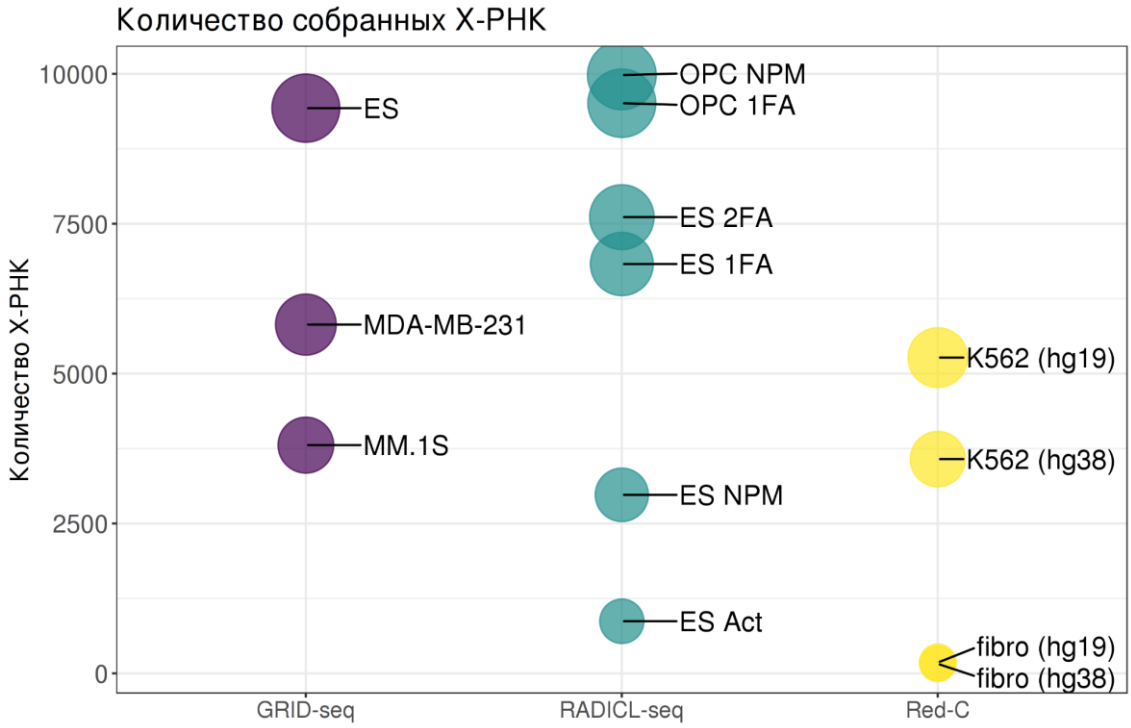

Рисунок 35. Количество собранных X-РНК.

Для протокола Red-C на рисунке 36 можно увидеть соотношение длин собранных X-РНК и количества их контактов.

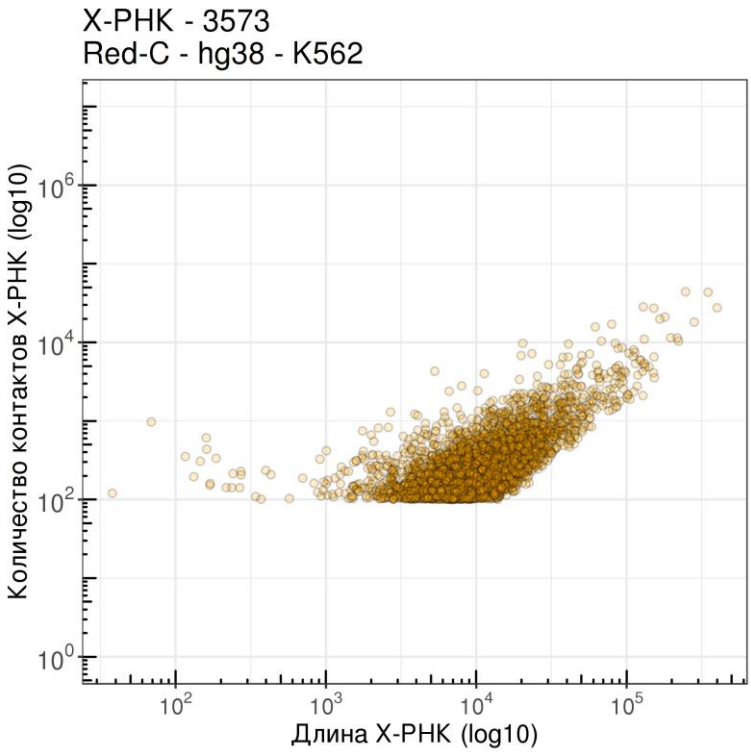

Рисунок 36. Соотношение количества контактов X-РНК и их длины. Протокол Red-C, клеточная линия K562, референсный геном версии hg38.

Полученные X-РНК были разделены по типам в зависимости от значений двух характеристик (рис. 37): близости к аннотированным генам и взаимной ориентации. Выделены следующие классы X-РНК: antisense - на противоположной цепи закодирован известный ген; close3 и close5 - находятся не далее 1000 нукл от 3`- и 5`-концов известных генов, соответственно; prRNA - находятся не далее 1000 нуклеотидов от 5`-конца известного гена, но на другой цепи; midDesert - в окрестностях 1Кб-50Кб на одноименной цепи не аннотировано ни одного известного гена; hardDesert - в окрестностях более 50Кб не аннотировано ни одного известного гена; antisenseX - на противоположной цепи тоже оказалась X-РНК.

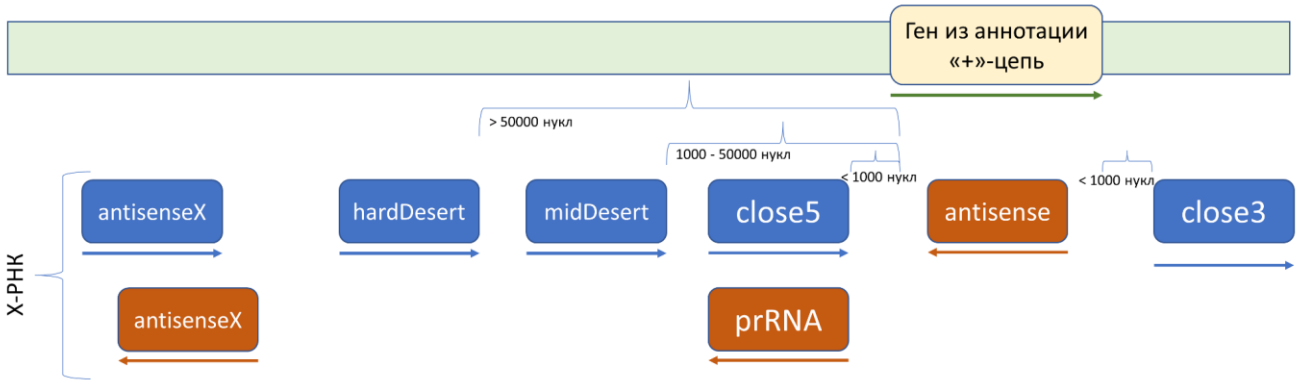

Рисунок 37. Схематичная иллюстрация правил, на основании которых X-РНК были разделены на классы.

Распределение X-РНК по классам можно увидеть на рисунке 38.

Наибольшее количество X-РНК для всех протоколов принадлежит классам, характеризующим их расположение близко к существующим генам по одноименной цепи (содержат “close3” и “close5”). Это явление можно объяснить тем, что позиционирование 3`- и 5`-концов генов в выбранной нами генной разметке может быть не очень точным и мы видим некоторое количество чтений, несколько выходящих за рамки существующей аннотации.

Довольно сильное различие в представленности классов между двумя версиями сборки генома человека (hg19 vs hg38, протокол Red-C) следует из того, что в разметку для версии hg19 были включены piРНК в количестве ~ 670000 генов, согласно аннотации piRNABank. Можно увидеть, что для версии hg19 сильно упал

процент X-РНК, находящихся на удалении от существующих генов. Видимо, в случае hg38 многие X-РНК из генной пустыни на самом деле являются piРНК. Этот вопрос требует дополнительного исследования и не был решен в данной работе.

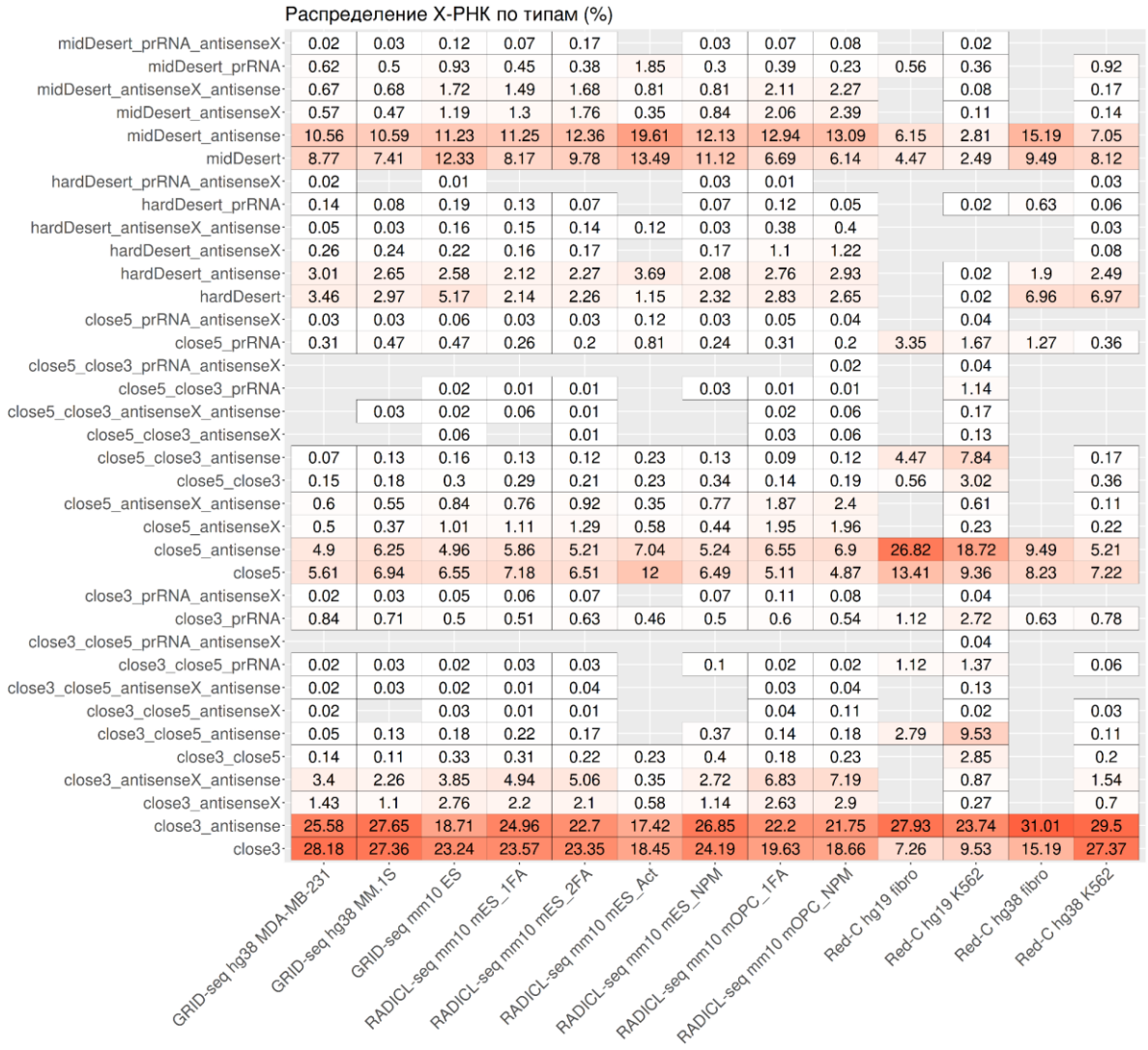

Рисунок 38. Распределение X-РНК по классам (в процентах для каждого эксперимента).

Наиболее интересные X-РНК относятся к тем, которые находятся в отдалении от известных генов (midDesert и hardDesert), а также аннотированы как антисенс РНК. Рассмотрим только такие X-РНК, отобрав дополнительно те из них, которые имеют более 1000 контактов. Для протокола Red-C (hg38) нашлось 112 таких X-РНК. Около 20 из них также обнаружены в человеческих клеточных линиях из протокола GRID-seq (hg38). Сравнение X-РНК с разметкой FANTOM

[102] показало, что 104 из 112 (93%) Х-РНК пересекаются с РНК, аннотированными в FANTOM, которые были заявлены как новые РНК.

Для протокола Red-C (hg19) были оставлены только Х-РНК, находящиеся в отдалении от известных генов. Для клеточной линии К562 осталось 1867 таких Х-РНК.

нкРНК выполняют множество функций не только внутри ядра и далеко не все их них являются ассоциированными с хроматином. Показано, например, что нкРНК принимают участие в регуляции клеточного цикла [103]. Экспрессия собранных для протокола Red-C (клеточная линия К562; версия референсного генома hg19) Х-РНК была изучена в данных по исследованию клеточного цикла в линии К562 [99]. Работа была сделана в коллаборации с лабораторией Евгения Валерьевича Шеваля, мы осуществляли биоинформатический анализ результатов секвенирования. В данной работе клетки были разделены по стадиям клеточного цикла и выделены популяции клеток, находящихся в стадиях G1, ранней стадии S, средней стадии S, а также смесь клеток в стадиях G2 и митозе. Мы располагали исходными чтениями, которые были картированы на референсный геном человека (hg19) с помощью программы HISAT2. Для собранных Х-РНК был исследован уровень экспрессии в каждой популяции клеток с определенным клеточным циклом. В каждой популяции были обнаружены несколько сотен Х-РНК (G1 - 158; ранняя S - 218; средняя S - 306; G2 и митоз - 203), имеющих ненулевую экспрессию. Однако, ни одна из обнаруженных Х-РНК не была охарактеризована как дифференциально экспрессирующаяся между какими бы то ни было парами популяций клеток. Видимо, если эти РНК и существуют, то никакого влияния на процессы, связанные с регуляцией клеточного цикла, обнаруженные Х-РНК не оказывают.

Сборка РНК-ДНК контактов до полной аннотации

Прежде чем двигаться дальше зафиксируем количества контактов, прошедшие Этап №1 и Этап №2, для экспериментального протокола Red-C, а также

оценим, сколько данных было потеряно от исходного количества секвенированных чтений. Для Red-C приведены результаты, разделенные по репликам, для двух клеточных линий, а также для двух версий сборки референсного генома человека (табл. 9).

Видно, что в процентном отношении реплики K562 сохранили примерно одинаковое количество данных, несмотря на сильную разницу в количестве чтений. Фибробласты, к сожалению, лишились более 90% всех данных, потеряв на этапе отбора уникального картирования больше чтений, чем это произошло в случае клеточной линии K562.

Таблица 9. Количество и процент контактов, прошедших Этапы №1 и №2 биоинформатического анализа. Протокол Red-C, клеточные линии K562 и фибробласты. Реплики представлены отдельно.

Образец	Всего чтений (млн)	Контакты, после Этапов №1 и №2 (млн)	Контакты, после Этапов №1 и №2 (%)
Red-C (hg19) - K562			
SRR10010326	122.5	18.50	15.10
SRR10010328	15.5	1.60	10.30
SRR10010330	217.0	24.80	11.40
Red-C (hg19) - фибробласты			
SRR10010323	320.0	4.00	1.25
SRR10010324	9.3	0.70	7.50
SRR10010325	18.6	0.68	3.60
Red-C (hg38) - K562			
SRR10010326	122.5	17.90	14.60
SRR10010328	15.5	1.50	9.70
SRR10010330	217.0	22.10	10.20
Red-C (hg38) - фибробласты			
SRR10010323	320.0	3.90	1.20
SRR10010324	9.3	0.69	7.40
SRR10010325	18.6	0.65	3.50

По окончании Этапа №2 мы получили финальную выборку контактов. Для протокола Red-C (K562) представлена диаграмма истощения, где проиллюстрированы основные этапы фильтрации (рис. 39). Несмотря на большую разницу в количестве исходных чтений между репликами протокола Red-C (клеточная линия K562), все реплики ведут себя одинаково и демонстрируют

схожую тенденцию к потере контактов в ходе биоинформатической обработки данных.

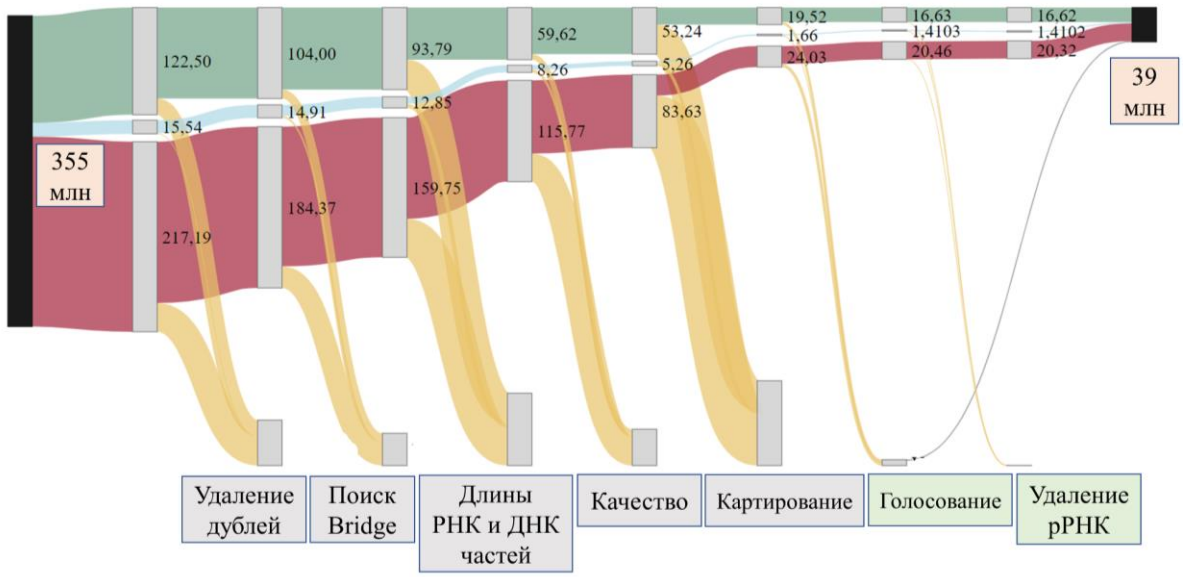

Рисунок 39. Диаграмма истощения для протокола Red-C (K562; hg38). Лентами зеленого, голубого и малинового цветов обозначены отдельные реплики; лентами желтого цвета обозначены контакты, не прошедшие очередной фильтр. Названия фильтров указаны в нижней части рисунка в боксах серого (относятся к Этапу №1) и зеленого (относятся к Этапу №2) цветов. Суммарные по всем репликам значения исходных чтений (355 млн) и финальных контактов (39 млн) указаны в боксах розового цвета.

По результатам Этапа №1 и Этапа №2 оказалось, что для большинства экспериментов сохранилось всего лишь 11-19% данных (рис. 40), что составляет более 40 миллионов контактов в каждом случае. Исключением являются фибробласты из эксперимента Red-C, которые сохранили менее 2% данных от исходного количества. Такая большая потеря данных приводит к снижению покрытия интересующих нас локусов ДНК и фрагментов РНК, хотя исходно была заложена большая глубина покрытия (более 350 миллионов чтений для протокола Red-C). Таким образом мы можем рассчитывать только на изучение наиболее часто контактирующих с хроматином РНК, а поиск и изучение более локально действующих РНК, представленных в меньшем количестве, требует увеличения глубины секвенирования.

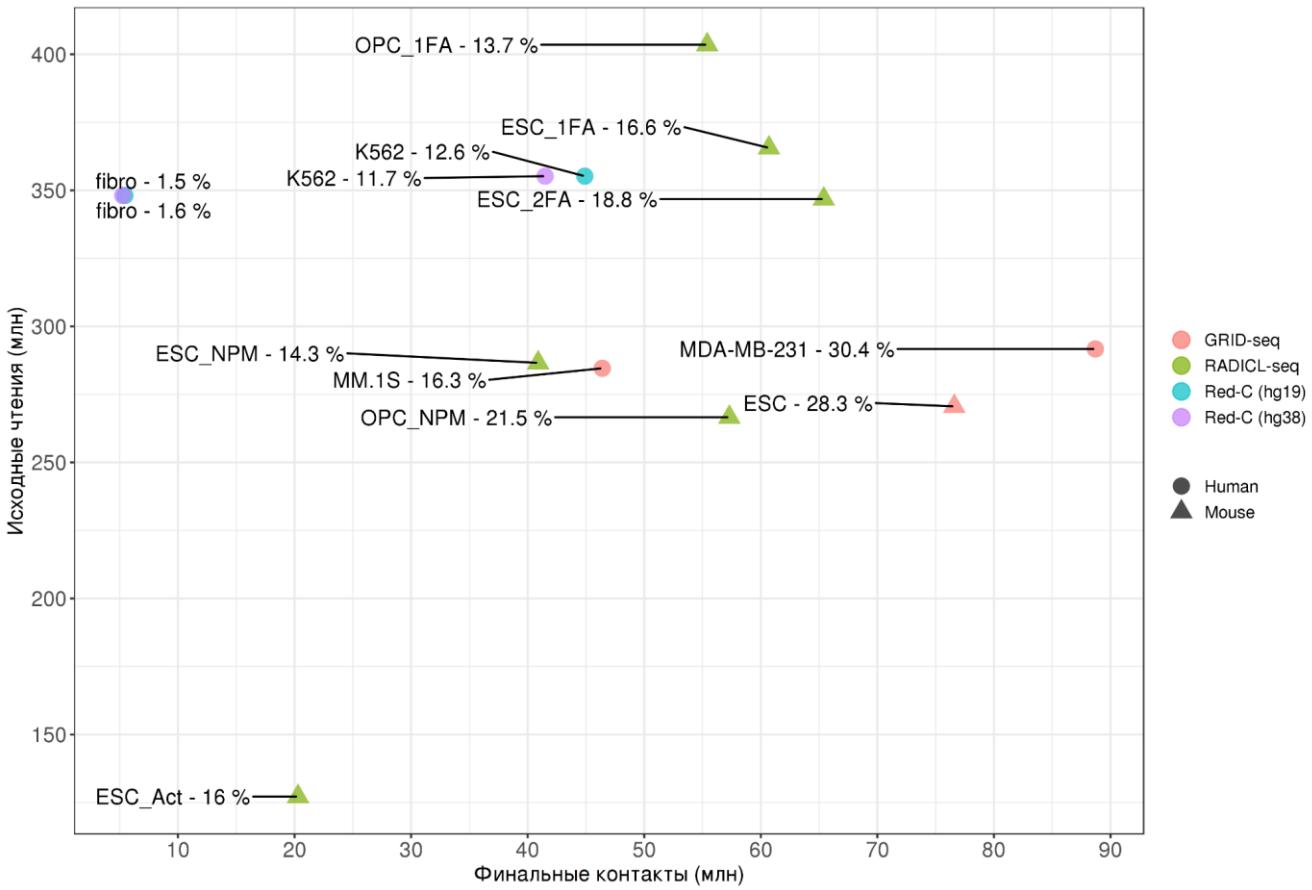

Рисунок 40. Соотношение количества исходных чтений к количеству контактов, прошедших Этап №1 и №2. Протоколы Red-C, GRID-seq, RADICL-seq. Реплики объединены.

Также стоит отметить, что для всех протоколов реплики одной клеточной линии ведут себя схожим образом в плане истощения, процент потерянных контактов аналогичен поведению данных метода Red-C. Для протокола Red-C результаты, полученные для разных версий референсного генома практические не отличаются.

Исследование вторичных РНК-ДНК контактов

Этап №3 включает в себя разработку нормировки, привлечение внешних данных, расчет хроматинового потенциала, а также изучение характера взаимодействий с хроматином хаРНК.

Конструирование фона

Транскрипты мРНК порождают большое количество РНК-ДНК контактов, в том числе и *in trans*. Согласно предложенному в протоколе GRID-seq подходу построения эндогенного фона фоновые контакты были определены как контакты белок-кодирующих генов с “не материнскими” хромосомами (рис. 41). Все такие контакты, а именно ДНК-части этих контактов, были собраны отдельно для дальнейшей обработки.

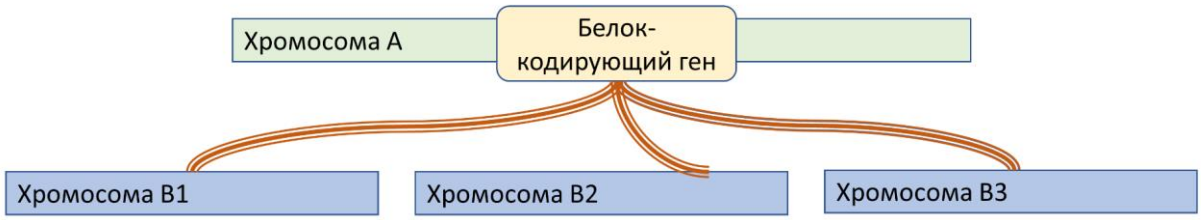

Рисунок 41. Схема, иллюстрирующая правила определения фоновых контактов: контакты мРНК с хромосомами, на которых не закодирован их ген.

Для каждого белок-кодирующего гена было рассчитано число контактов для их транскрипта. Как видно из рисунка, существуют гены, число контактов которых очень сильно отличаются в большую сторону от основной массы мРНК. Из выборки контактов, составляющих фон, были удалены 50 наиболее контактирующих мРНК (отмечены оранжевым на рис. 42), которые могли бы повлиять на поведение трека фоновых контактов. Также удалены 1000 наименее контактирующих мРНК (отмечены синим на рис. 42).

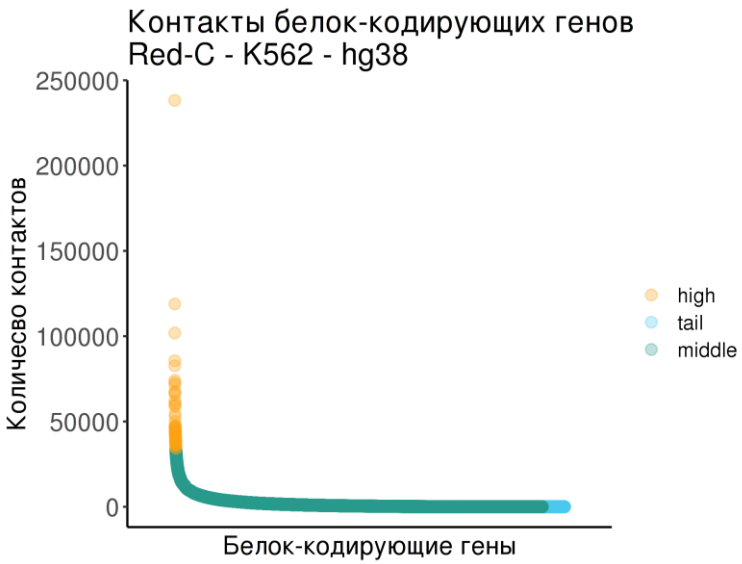

Рисунок 42. Количества контактов белок-кодирующих генов. Гены упорядочены по уменьшению числа контактов (слева направо). high - 50 наиболее контактирующих белок-кодирующих гена; tail - 1000 наименее контактирующих белок-кодирующих гена; middle - оставшаяся часть белок-кодирующих генов, контакты которых составляют фон.

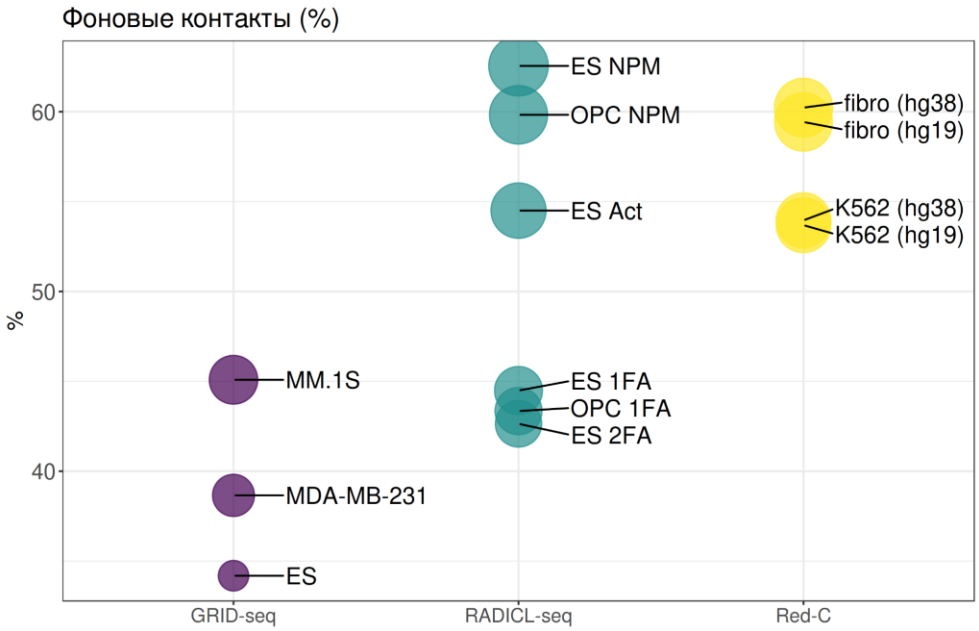

Рисунок 43. Процент контактов, определенных как фоновые.

Оказалось, что для протокола Red-C и для половины экспериментов из протокола RADICL-seq более половины контактов были определены как фоновые (рис. 43).

Далее полученный трек - ДНК-части фоновых контактов - сгладили с помощью программы Stereogene (опция Smoother), в результате чего были

получены значения сглаженного фона вдоль всего референсного генома в интервалах длиной 500 нуклеотидов. Каждому индивидуальному контакту был присвоен вес до нормировки, равный единице, а также значение фона по координате ДНК-части контакта.

Нормировка на фон для каждого контакта i была осуществлена согласно следующей формуле:

N1i = $(1 \,/\, LibsizeRD) \,/\, ((BackValue \,/\, LibsizeBack) \,+\, \mu)$,

где LibsizeRD - размер библиотеки контактов, т.е. все контакты для данного образца (реплики объединены), BackValue - значение фона для индивидуального контакта, согласно координате ДНК-части, LibsizeBack - количество контактов, определенных как фоновые, μ - псевдокаунт << 1. Дополнительно полученные веса были перенормированы так, чтобы сумма контактов до нормировки на фон была равна сумме контактов после нормировки.

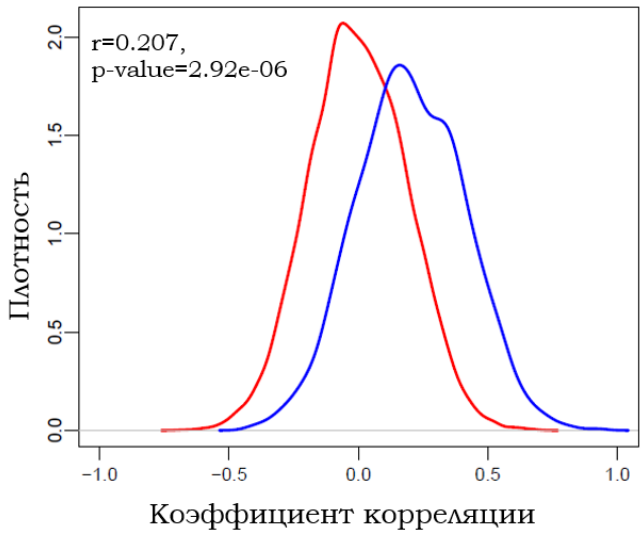

Рисунок 44. Сглаженные распределения коэффициентов корреляций для геномных окон из выдачи StereoGene (полногеномные разметки фоновых контактов (ДНК-части) и локусов генома, гиперчувствительных к обработке ДНКазой I). Приведено наблюдаемое (синий) и фоновое (красный) распределения коэффициентов корреляций. Смещение наблюдаемого распределения от фонового означает наличие корреляции.

Далее мы работали с нормированными на фон весами контактов.

В нашей лаборатории показано, что трек фоновых контактов положительно коррелирует с открытым хроматином (рис. 44). Открытый хроматин определен согласно разметке, полученной в результате исследования областей генома, гиперчувствительных к обработке ДНКазой I (см. раздел “Материалы и методы”).

Расчет хроматинового потенциала

Для протокола Red-C мы располагали данными секвенирования РНК (тотальная РНК, цепь-специфичная библиотека, одноконцевые чтения), полученного в лаборатории С.В. Разина, для той же клеточной линии K562, что была использована в эксперименте Red-C. Было показано, что для индивидуальных РНК количество РНК-ДНК контактов коррелирует с уровнем экспрессии этой РНК (коэффициент корреляции Пирсона = 0.817, p-value << 0.001) (рис. 45).

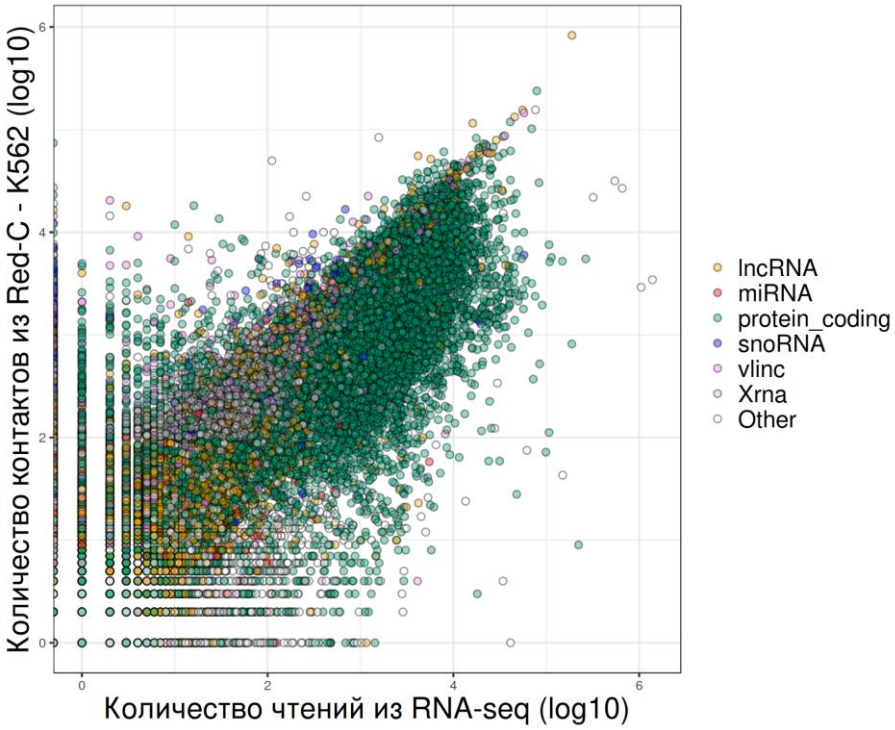

Рисунок 45. Зависимость количества контактов РНК с хроматином от уровня их экспрессии. Протокол Red-C, клеточная линия K562, референсный геном версии hg38.

Процедура обработки данных секвенирования РНК была максимально приближена к таковой для данных “все-против-всех”, а именно были использованы одинаковые программы для картирования, те же пороги при фильтрации

результатов картирования, а также применена процедура голосования при определения уровня экспрессии для генов.

Была предложена и разработана метрика хроматинового потенциала (chP), основанная на сравнении уровня экспрессии РНК с количеством ее контактов с хроматином. На основании этой метрики отобраны РНК, которые контактируют с хроматином значимо чаще, чем это ожидается, исходя из уровня экспрессии этих РНК.

Для каждой РНК i хроматиновый потенциал рассчитан по следующей формуле:

$$chPi = log(\frac{RDcnt + \mu}{LibsizeRD}) / (\frac{RScnt + \mu}{LibsizeRS}),$$

где RDcnt - суммарное количество контактов для каждой РНК i, LibsizeRD - размер библиотеки контактов, т.е. все контакты для данного образца (реплики объединены), RScnt - суммарное количество чтений для РНК i, пришедших из эксперимента по секвенированию РНК, LibsizeRS - размер библиотеки из эксперимента по секвенированию РНК, μ - псевдокаунт, равный 1.

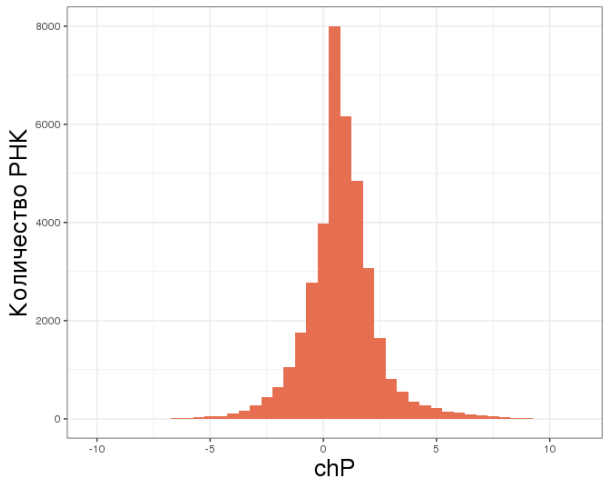

Рисунок 46. Распределение значений хроматинового потенциала для всех РНК из эксперимента Red-C, клеточная линия K562.

Значение хроматинового потенциала для всех РНК из протокола Red-C (клеточная линия K562; hg38) распределено, как показано на рисунке 46.

Данное распределение включает chP и для мРНК, на рисунке 47 представлено распределение хроматинового потенциала только для мРНК.

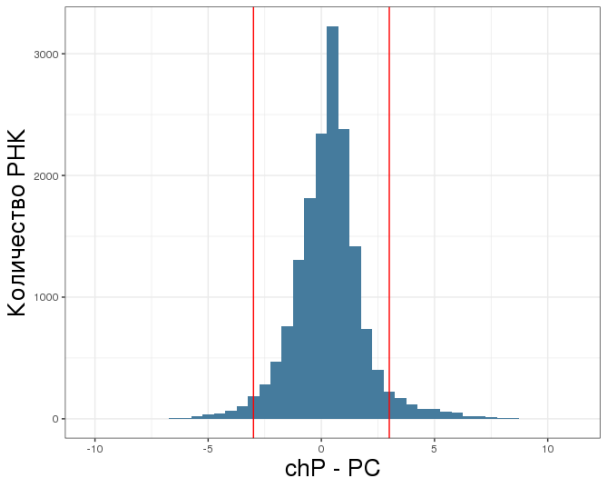

Рисунок 47. Распределение значений хроматинового потенциала для мРНК из эксперимента Red-C, клеточная линия K562. Красными линиями отмечены значения chP, равные -3 и 3.

Для оценки статистической значимости метрики хроматинового потенциала в качестве фоновой модели возьмем значения chP для мРНК, удалив правый и левый хвосты распределения с порогом 3 и -3, соответственно. Полученное фоновое распределение было аппроксимировано нормальным распределением.

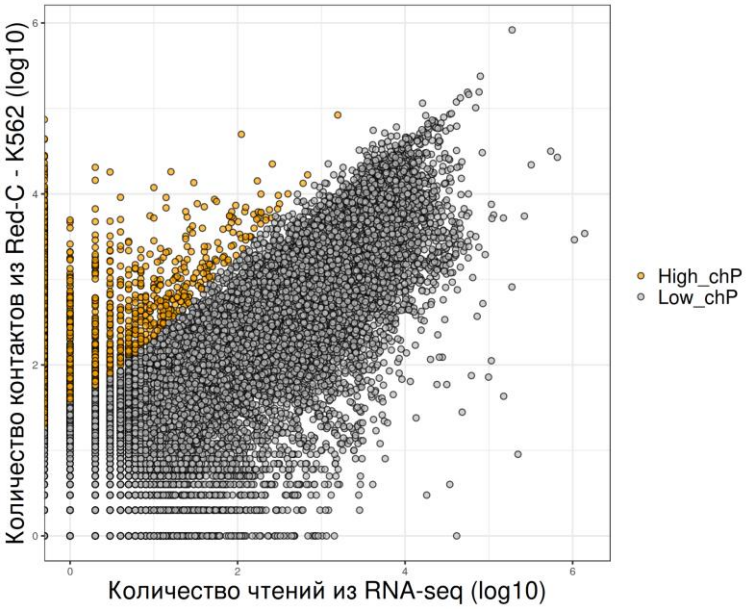

Рисунок 48. Зависимость количества контактов РНК с хроматином от уровня их экспрессии. High-chP - РНК с высоким хроматиновым потенциалом; Low_chP - РНК с низким хроматиновым потенциалом.

Далее мы можем оценить p-value любого значения хроматинового потенциала (с учетом поправки на множественное тестирование) и выделить РНК с значимо высоким значением chP (поправленное значение p-value < 0.1) (рис. 48).

На рисунке 49 можно увидеть распределение по классам РНК с высоким хроматиновым потенциалом для эксперимента Red-C (клеточная линия K562; hg38).

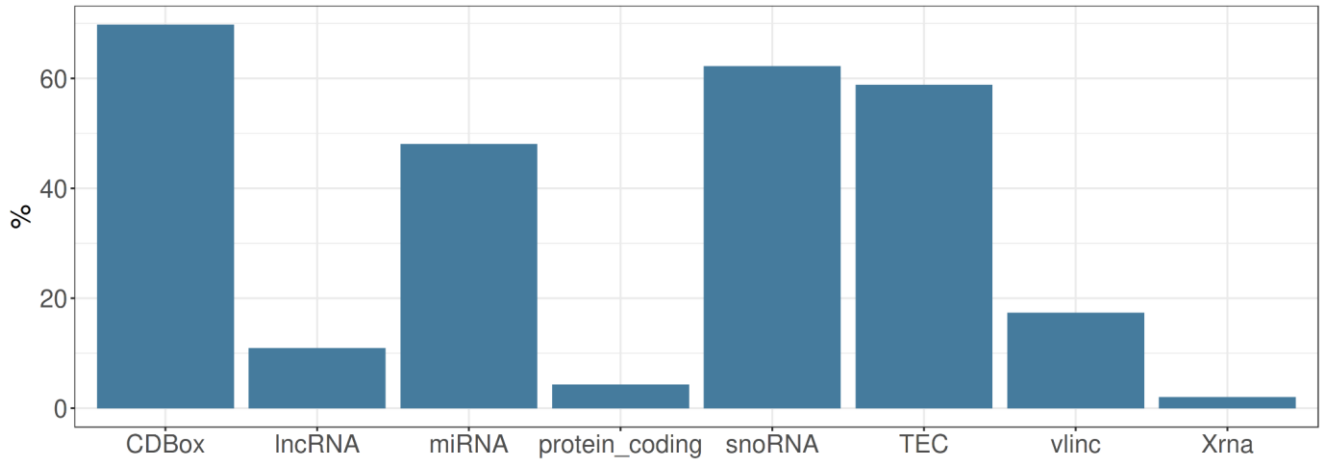

Рисунок 49. Распределение РНК с высоким хроматиновым потенциалам (количество контактов > 100) по классам. Для каждого класса представлен процент РНК с высоким chP к общему количеству РНК в классе. Протокол Red-C, клеточная линия K562, версия референсного генома - hg38.

Как было показано ранее для клеточной линии K562 из протокола Red-C на белок-кодирующие гены приходится более 70% всех контактов. Однако, только ~4% мРНК обладают высоким chP. Более половины представителей классов малых РНК, а также ~10-20% lncRNA и vlinc, некоторые X-РНК обладают высоким chP. Стоит отметить РНК из класса TEC, которые также представлены в данном анализе. Этот класс РНК описан в GENCODE как неизвестные гипотетические РНК разной длины, для которых не показан сплайсинг, но есть потенциал к полиаденилированию. Такие РНК согласно GENCODE подлежат экспериментальной проверке на предмет возможности к трансляции, хотя для некоторых длинных нкРНК показано полиаденилирование [104]. Аналогичный анализ данных протокола Red-C (клеточная линия K562) для версии сборки hg19

дает похожие результаты. Для hg19 были аннотированы piРНК, ~52% которых обладают высоким значением chP.

РНК с высоким значением хроматинового потенциала могут действительно оказаться функциональными хаРНК. Для них характерна довольно низкая экспрессия по данным RNA-seq, но при этом достаточно высокое количество контактов с ДНК. Таким образом они могут все время находиться в плотной связке с хроматином, выполняя свои функции. В качестве примера можно привести нкРНК XIST, обладающую значимо высоким значением chP.

С другой стороны, не все хаРНК обладают значимо высоким хроматиновым потенциалом. Например, длинные нкРНК MALAT1 и NEAT1 имеют несколько десятков тысяч контактов и положительное значение chP, однако уровень их экспрессии насколько высок, что перевес в сторону РНК-ДНК контактов оказывается незначимым.

Также стоит обратить внимание на тот факт, что для расчета хроматинового потенциала были использованы данные RNA-seq, полученные по протоколу выделения тотальной РНК с деплецией рРНК. Обычно в ходе пробоподготовки таких библиотек есть стадия отбора полученных фрагментов РНК по размеру, которая отсекается все короткие (~ меньше 200-300 нуклеотидов) РНК. Таким образом в данных RNA-seq экспрессия малых РНК будет сильно занижена, но многие из них обладают большим количеством РНК-ДНК контактов. В результате значение chP для таких РНК скорее всего будет значимо высоким.

Некоторые малые РНК (например, piРНК) закодированы в областях генома, заведомо богатых повторами. Используя для анализа только уникально картированные на референсный геном чтения мы не учитываем в полной мере ни потенциал взаимодействия этих РНК с хроматином, ни уровень их экспрессии.

Для исследования малых РНК стоит использовать информацию об экспрессии, предоставленную специально для малых РНК, и рассчитывать хроматиновый потенциал более аккуратно.

К сожалению, для протоколов кроме Red-C авторы не предоставили данных о секвенировании РНК. Для расчета хроматинового потенциала было решено воспользоваться общедоступными данными по секвенированию РНК требуемых клеточных линий, сделанных с помощью аналогичного для Red-C подхода - тотальная РНК с деплецией рРНК и с сохранением информации о цепи (см. раздел “Материалы и методы”).

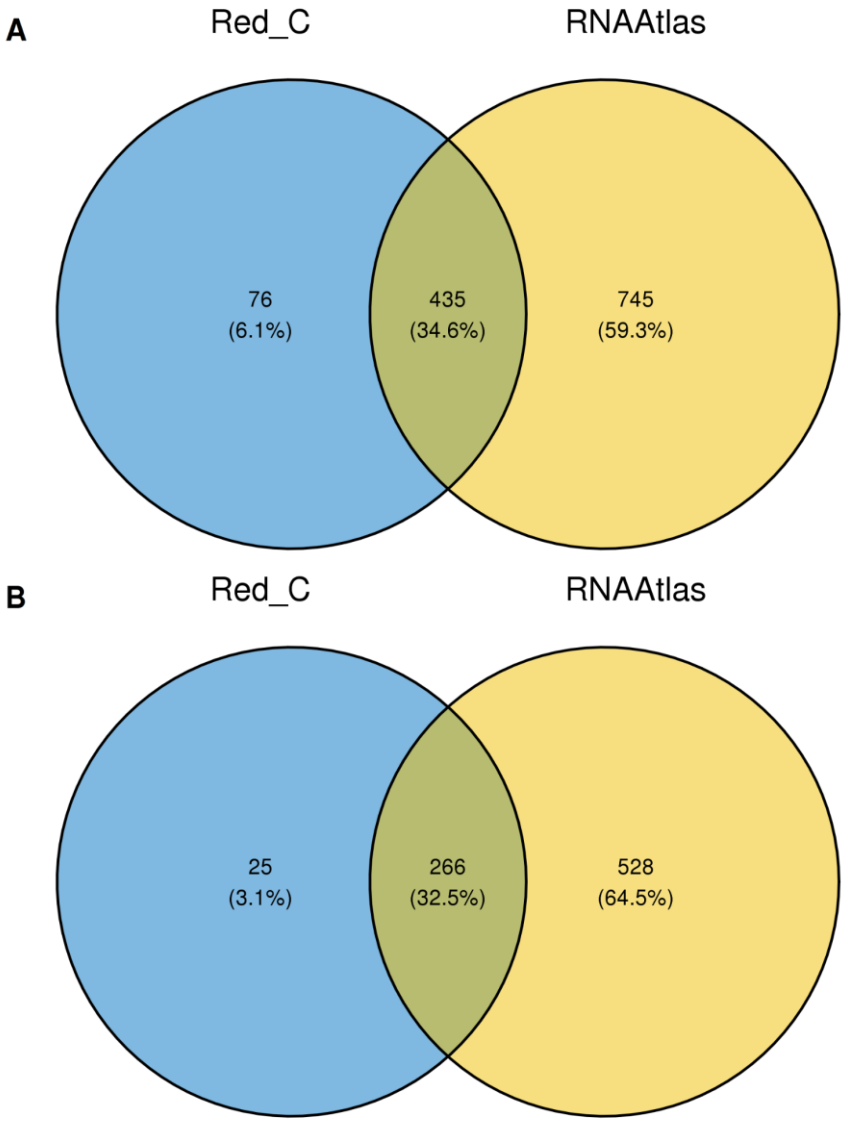

Рисунок 50. Пересечение наборов РНК, обладающих высоким хроматиновым потенциалом (chP) для протокола Red-C, клеточная линия K562 (референсный геном версии hg38). Круг синего цвета - chP рассчитан с использованием данных об экспрессии РНК из оригинальной статьи протокола Red-C; круг желтого цвета - chP рассчитан с использованием данных об экспрессии РНК из проекта RNAAtlas. (А) - РНК всех типов, (Б) - не мРНК.

На примере Red-C (клеточная линия K562) показано, что уровни экспрессии генов из данных по секвенированию РНК от авторов Red-C и из общедоступных данных из RNA Atlas не противоречат друг другу (коэффициент корреляции Пирсона = 0.92, p-value << 0.00001). Однако, если сравнить значения хроматинового потенциала, то видно, что по данным RNA Atlas было определено гораздо больше РНК с высоким хроматиновым потенциалом, чем при использовании результатов секвенирования РНК от авторов Red-C (рис. 50). Стоит отметить, что практически все не белок-кодирующие хаРНК с высоким chP, рассчитанным по RNA-seq из Red-C были детектированы как таковые и в данных, где chP был рассчитан по RNA-seq из RNA Atlas (рис. 50).

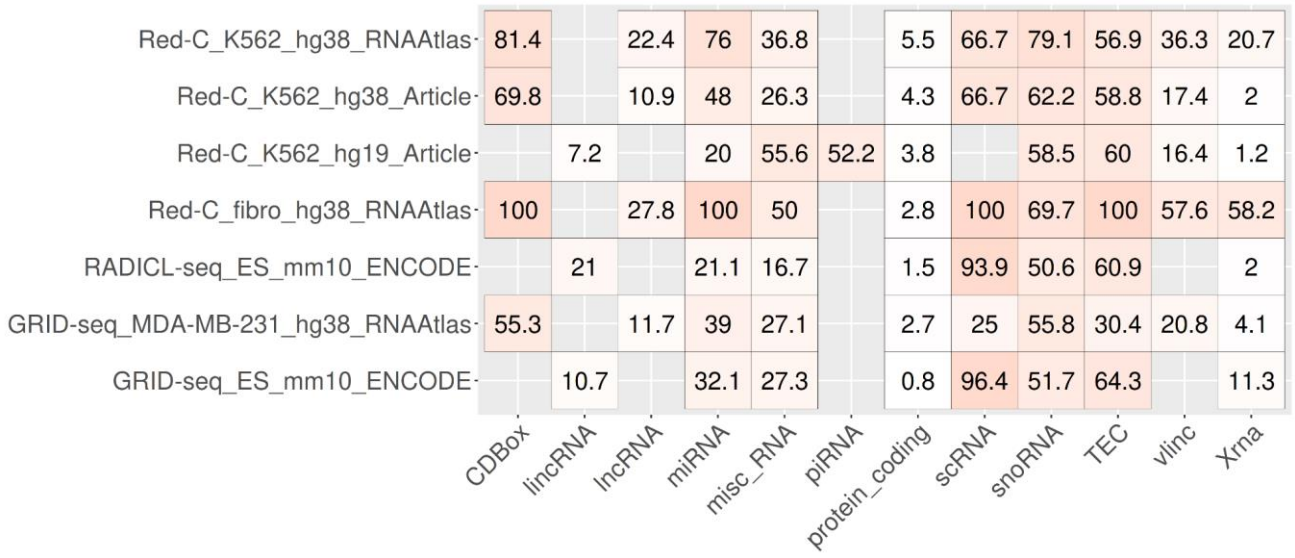

Рисунок 51. Представленность РНК с высоким хроматиновым потенциалом по типам и по экспериментам. Цифры в клетках тепловой карты карты отражают проценты РНК данного класса с высоким chP от общего количества РНК данного класса, детектированных в соответствующем протоколе “все-против-всех”.

Для протоколов Red-C, GRID-seq и RADICL-seq была также рассчитана метрика хроматинового потенциала. Данные представлены только для тех клеточных линий, для которых удалось найти информацию об уровне экспрессии (см. раздел “Материалы и методы”) (рис. 51). В случае мышиных эмбриональных

стволовых клеток для протокола RADICL-seq была выбрана обработка 1% формальдегидом.

Можно заметить, что высоким хроматиновым потенциалом обладают многие классы нкРНК разной длины, в то время как для всех протоколов, клеточных линий и референсных геномов мРНК с высоким chP не превышает 6%. Интересно, что для протокола Red-C количество X-РНК с высоким хроматиновым потенциалом, рассчитанным на основании разных источников об уровне экспрессии генов, сильно различается. В случае использования результатов RNA-seq, сделанных ровно для той же версии клеточной линии, на которой реализован сам протокол Red-C, X-РНК с высоким chP всего 8 штук. В то время как использование публичных данных дает уже 252 X-РНК с высоким chP. Использование сторонних данных, особенно результатов такого чувствительного метода как RNA-seq, может вносить существенные искажения в результаты исследования.

Рисунок 52. Распределение количества X-РНК с высоким хроматиновым потенциалом по классам (протокол Red-C, клеточная линия K562). Хроматиновый потенциал рассчитан с применением RNA-seq от авторов Red-C (Article) и сторонних данных (RNAAtlas).

Полученные наблюдения могут быть объяснены не столько биологическими законами, сколько batch-эффектом, происходящим вследствие недокументированных отличий при ведении клеточных линий (количество делений после разморозки, точные условия содержания клеточной линии и многое другое).

Если все же детально посмотреть на типы Х-РНК с высоким chP, рассчитанным с использованием внешних данных, то можно увидеть, что помимо Х-РНК, примыкающих к границам гена, детектировано достаточно много Х-РНК из генных пустынь (рис. 52). Возможно, это специфические РНК, экспрессирующиеся в клеточной линии К562, содержащейся в конкретных условиях.

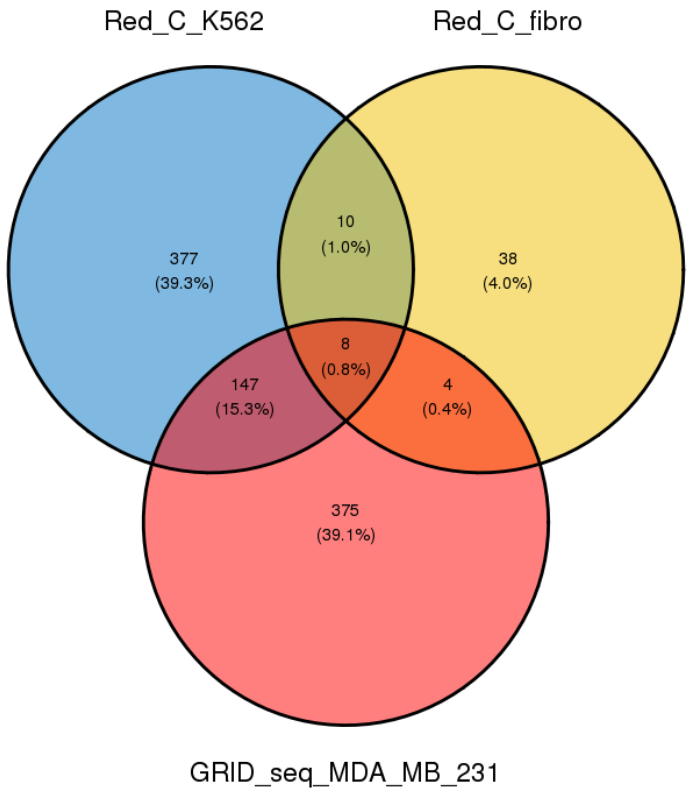

Рисунок 53. РНК с высоким хроматиновым потенциалом, не принадлежащие классам мРНК и Х-РНК. Показано пересечение списка хаРНК с высоким chP, обнаруженным в протоколах Red-C (клеточные линии К562 и нормальные фибробласты) и GRID-seq (клеточная линия MDA-MB-231). В качестве источника данных об уровне экспрессии во всех случая использованы данные из RNA Atlas для соответствующей клеточной линии; версия референсного генома hg38.

Было изучено, насколько сильно пересекаются списки РНК с высоким chP для разных протоколов без учета Х-РНК, т.к. они были собраны индивидуально для каждой клеточной линии, и без мРНК. Для протоколов, реализованных на клеточных линиях человека метрику chP удалось подсчитать для Red-C (клеточные линии К562 и фибробласты) и для GRID-seq (клеточная линия MDA-MB-231), в качестве источника данных об уровне экспрессии во всех случая использованы данные из RNA Atlas для соответствующей клеточной линии (рис. 53).

Для фибробластов из протокола Red-C в принципе обнаружено очень мало РНК с высоким хроматиновым потенциалом, что может быть объяснено фактом потери более 90% данных в ходе обработки. Для протоколов Red-C (К562) и GRID-seq обнаружено ~150 РНК с высоким chP для обеих клеточных линий. Таким образом на основании метрики chP можно отбирать хаРНК, которые в зависимости от клеточной линии и протокола ведут себя по-разному, а также конститутивные хаРНК. Хотя стоит принимать во внимание, что в данном случае мы сравниваем между собой не только клеточные линии, но и протоколы.

Аналогичное исследование можно сделать для эмбриональных стволовых клеток мыши в сравнении двух протоколов: GRID-seq и RADICL-seq (Рис. 51). Хроматиновый потенциал был рассчитан на основании одного и того же RNA-seq, полученного из проекта ENCODE, версия референсного генома mm10. В данном случае 126 нкРНК имеют высокий chP в обоих протоколах, 371 нкРНК обладает высоким chP только в случае протокола GRID-seq, 99 нкРНК детектируются в RADICL-seq, как РНК с высоким chP.

Аннотация ДНК-частей контактов

Изучаемые контакты содержат РНК-часть и ДНК-часть. Для того, чтобы понять, какие именно РНК были зафиксированы в эксперименте, как контактирующие с хроматином, РНК-части, как показано выше, были аннотированы генами с применением разработанной процедуры голосования.

Для изучения того, с какими именно участками хроматина контактируют найденные РНК, ДНК-части также могут быть аннотированы какой-либо

разметкой, позволяющей охарактеризовать свойства хроматина. Задача несколько осложняется тем, что не всегда желаемые разметки существуют для нужной клеточной линии.

В данной работе была выбрана разметка состояний хроматина согласно предложенной аннотации из работы *Ernst et al.* [96] для клеточной линии K562. Авторы выделяют 15 состояний, основываясь на комбинации различных эпигенетических меток. Полногеномные профили контактов (по ДНК-части) индивидуальных РНК из протокола Red-C достаточно разреженные даже в случае многих высоко контактирующих РНК. Исследование каждого из 15 состояний хроматина отдельно не представляется возможным, т.к. на отдельные состояния в случае индивидуальных РНК может приходиться крайне мало контактов или же сигнал отсутствует вовсе. Предложенные состояния хроматина были сгруппированы в две крупных группы непересекающихся интервалов, соответствующие активному и репрессированному хроматину (см. раздел “Материалы и методы”). ДНК-части РНК-ДНК контактов для протокола Red-C короткие (22-24 нуклеотида) (рис. 17). При аннотации ДНК-часть каждого контакта рассматривалась как точка, соответствующая меньшей координате ДНК-части. Таким образом мы избегали проблемы конфликтных ситуаций в случае попадания ДНК-части на границу интервалов, соответствующих разным состояниям хроматина.

Изучение характера РНК-ДНК взаимодействий

Как было показано выше, РНК может контактировать на разном расстоянии от своего гена, может взаимодействовать не только с хромосомой, на которой закодирован ее ген, но и с другими хромосомами. На основании удаленности контактов РНК от своего гена можно изучить характер взаимодействия конкретной РНК с хроматином.

Для каждой РНК были рассчитаны непересекающиеся интервалы, разбивающие геном на участки разноудаленные от гена (рис. 54): непосредственно область самого гена (gene - G); 0-50 Кб вокруг гена (short - S); 50-500 Кб вокруг

гена (medium - M); 0.5-5 Мб вокруг гена (long - L); более 5 Мб на “материнской хромосоме” (remote - T); все другие хромосомы (trans - T). Интервалы short и medium были объединены в интервал SM (0.5 Мб вокруг гена).

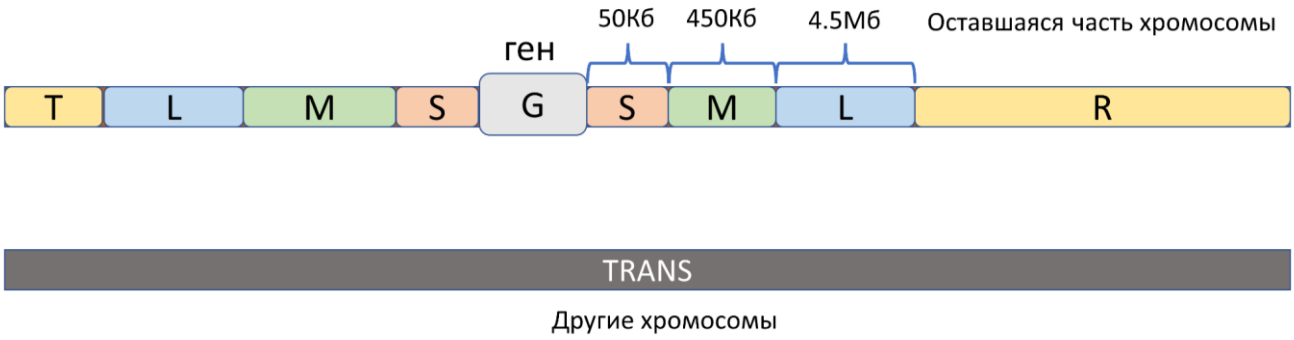

Рисунок 54. Схема конструкции непересекающихся интервалов вокруг контактирующей с хроматином РНК в зависимости от их удаленности от соответствующего гена.

Были отобраны такие РНК, для которых зарегистрирован хотя бы один контакт с хроматином в соответствующих интервалах (SM, L, R, T), а общее количество контактов превышало бы 500. Получилось 10367 РНК для протокола Red-C (клеточная линия K562, референсный геном сборки hg19). Для каждой РНК в интервалах SM, L, R, T была рассчитана плотность контактов, а также отношения SM/L, L/R, R/T. Полученные отношения были независимо друг от друга z-трансформированы и поделены на 5 квантилей.

Дополнительно для каждой РНК были рассчитаны плотности контактов в участках активного (Act) и подавленного (Rep) хроматина (см. раздел “Материалы и методы”) в близком окружении от своего гена (~ 5Мб вокруг гена), внутрихромосомные и полногеномные. Для оценки того, с каким именно хроматином предпочитает взаимодействовать РНК, было рассчитано отношение плотностей контактов локусов активного хроматина к локусам подавленного хроматина (Act/Rep).

Далее были выделены 3 группы РНК, схожие по своему характеру взаимодействия с хроматином (рис. 55).

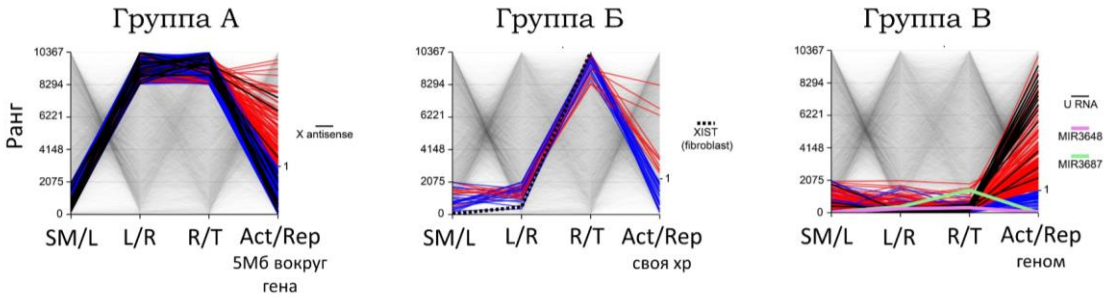

Рисунок 55. Группы РНК, выделенные по характеру взаимодействия с хроматином: группа А: контактирующие преимущественно недалеко от своего гена, 313 РНК, отношение Act/Rep рассчитано для интервала в 5Мб вокруг своего гена; группа Б: предпочитающие взаимодействовать с “родной” хромосомой аналогично XIST, 30 РНК, нкРНК XIST (фибробласты) выделена пунктирной линией, отношение Act/Rep рассчитано для своей хромосомы; группа В: контактирующие полногеномно, 224 РНК, отношение Act/Rep рассчитано полногеномно. Данные представлены на примере протокола Red-C, клеточная линия K562, референсный геном сборки hg19. Красными линиями выделены РНК, взаимодействующие преимущественно с активным хроматином (Act/Rep > 1), синими линиями - с неактивным (Act/Rep < 1).

Композицию классов РНК, составляющих вышеописанные три группы можно увидеть на рисунке 56. Как и было отмечено ранее, среди всех РНК преобладают белок-кодирующие РНК. В группах А и Б практически полностью отсутствуют малые РНК, которые сосредоточены в основном в группе В. Иначе ведут себя РНК из классов vlinc и Х-РНК, которые находятся в основном в группе А, но также представлены в группе Б. В группе А можно обратить внимание на 16 Х-РНК (подгруппа antisense), среди которых для 13ти Х-РНК отношение Act/Rep < 1 в близком окружении от своего гена, т.е. они взаимодействуют с неактивным хроматином в радиусе 5Мб от своего гена. В группе Б детектировано довольно мало РНК, всего 30 штук, однако они довольно интересны с точки зрения характера взаимодействия с хроматином. Выделенные РНК взаимодействуют преимущественно с своей хромосомой, подобно XIST, причем большинство из них при этом контактирует с неактивным хроматином. Для 10 из 12 мРНК группы Б отношение Act/Rep > 1, т.е. с неактивным хроматином контактируют в основном

нкРНК, включая одну X-РНК. В группу В входят РНК, контактирующие полногеномно, включая MALAT1 и малые РНК разных классов.

Такой подход к исследованию характера взаимодействия РНК с хроматином может способствовать идентификации потенциальных регуляторных РНК, которые действуют *in cis*, или ведут себя аналогично какой-то известной хаРНК (например, XIST).

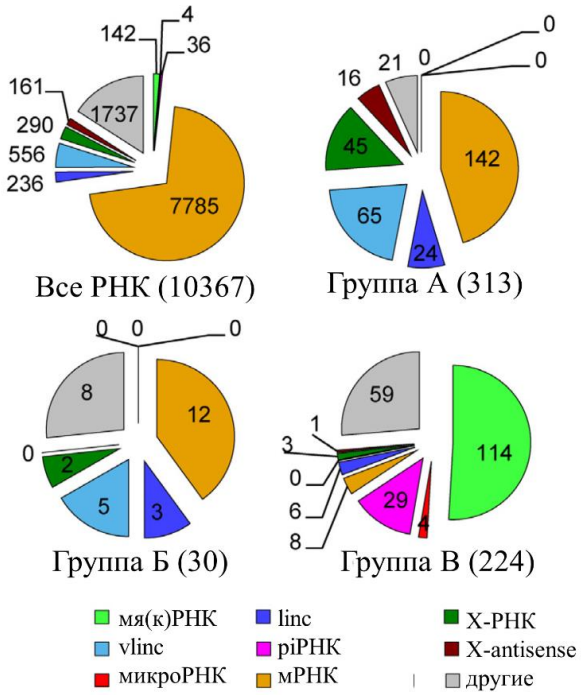

Рисунок 56. Количество РНК по типам в вышеописанных трех группах, а в всех выборке РНК, для которых зарегистрирован хотя бы один контакт с хроматином в соответствующих интервалах (SM, L, R, T), а общее количество контактов превышало бы 500 (Все РНК).

В группе В были обнаружены две микроРНК - MIR3648 и MIR3687, гены которых расположены в 5`-области пре-рРНК. Эти микроРНК взаимодействуют полногеномно в основном с неактивным хроматином. Коллегами было показано, что эти две микроРНК предпочитают взаимодействовать с участками ДНК, где не аннотированны гены, включая хромосому 18, обедненную генами. Стоит отметить, что ранее было показано, что хромосома 18 скорее является местом посадки хаРНК, чем их источником. Для MIR3687, для которой было установлено более 11000 контактов с хроматином, показана ассоциация профиля РНК-ДНК контактов с

регионами поздней репликации по данным Repli-seq (смещение наблюдаемого распределения от фонового вправо для S3, S4, G2) (рис. 57).

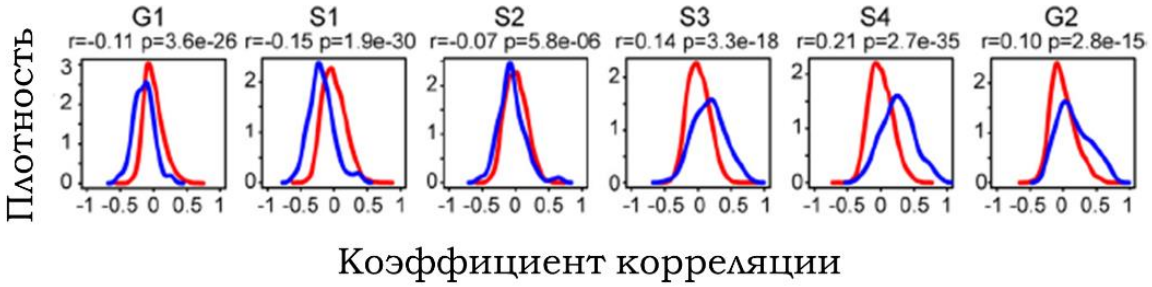

Рисунок 57. Сглаженные распределения коэффициентов корреляций для геномных окон из выдачи StereoGene (профиль контактов MIR3687 (протокол Red-C, клеточная линия K562) и временные профили репликации по данным Repli-seq [97], упорядоченные от ранней к поздней: G1 (соответствует ранней репликации), S1, S2, S3, S4, S5, G2 (соответствует поздней репликации)). Приведено наблюдаемое (синий) и фоновое (красный) распределения коэффициентов корреляций. Смещение наблюдаемого распределения от фонового означает наличие корреляции.

Сосредоточившись на белок-кодирующих генах было показано, что мРНК чаще контактируют с областями, следующими за стартом транскрипции (по ходу транскрипции), чем с предшествующими ему. Для каждой мРНК хромосома, на которой закодирован ее ген, была разделена на непересекающиеся интервалы переменной длины в 3`- и 5`-области от начала гена с учетом направления транскрипции конкретного гена. Самые близкие к началу гена интервалы составляли в длину 100 нуклеотидов, каждый последующий интервал был в 2 раза длиннее предыдущего. Для каждого такого интервала было рассчитано количество контактов индивидуальной мРНК. Затем данные по всем мРНК были объединены, а для каждого интервала рассчитана плотность контактов (суммарное количество количество контактов для интервала, деленное на длину интервала). В качестве контроля были выполнены аналогичные расчеты, однако направление транскрипции генов было проигнорировано. Видно, что наибольшая плотность контактов зарегистрирована в самых близких от старта транскрипции интервалах, что видимо соответствует взаимодействию зарождающейся РНК с ДНК

посредством транскрипционного комплекса. По мере удаления от старта транскрипции плотность контактов уменьшается (рис. 58, верхняя панель). В случае учета направления транскрипции распределение плотностей асимметрично относительно начала гена (рис. 58, наверху слева). Частота взаимодействия мРНК с 3`-областями в ~1,5 выше, чем с 5`-областями относительно начала гена, эффект наблюдается в пределах 100 Кб от начала гена (рис. 58, внизу слева). Показанное наблюдение можно объяснить тем, что РНК следует за РНК-полимеразой в процессе транскрипции гена.

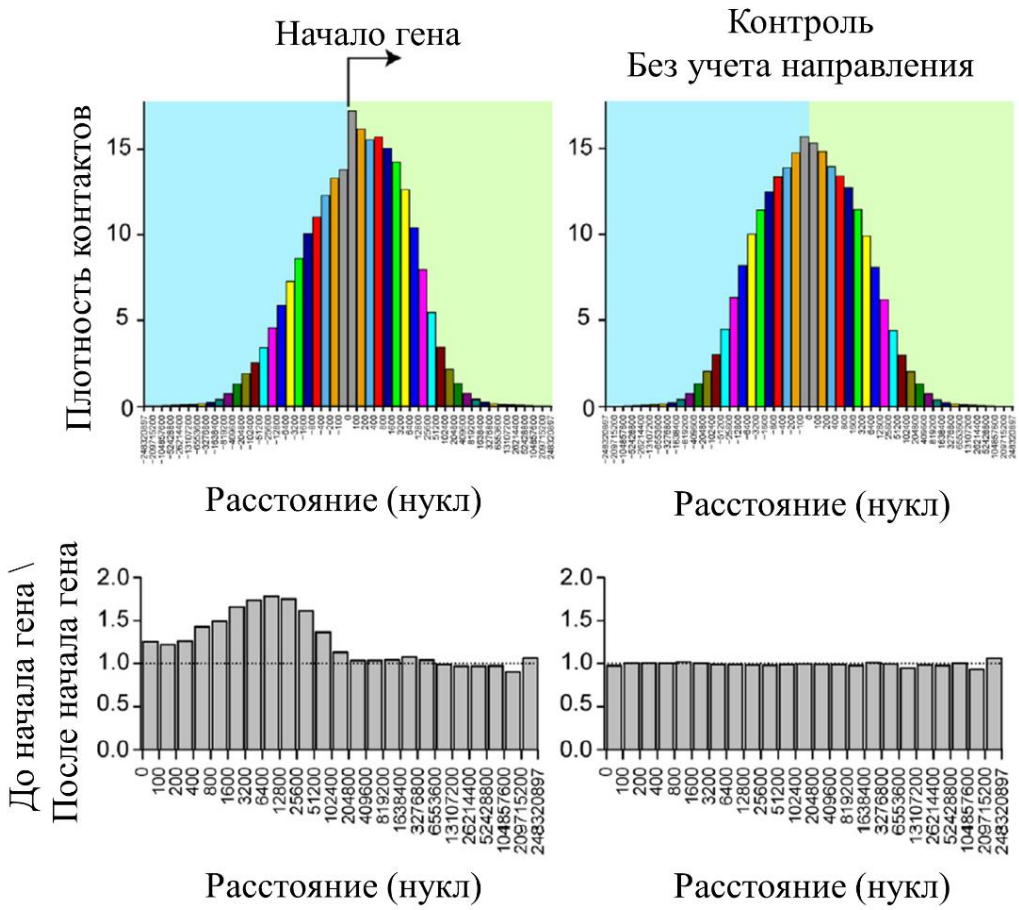

Рисунок 58. Частота контактов фрагментов мРНК с интервалами по ходу (слева) или в обратную сторону (справа) относительно направления транскрипции. Левая панель: для каждого белок-кодирущего гена было учтено направление транскрипции; правая панель (контроль): направление транскрипции белок-кодирующих генов было проигноировано. Пары столбцов одного цвета представляют результаты для равноотстоящих от начала гена интервалов. Нижняя панель: показаны соотношения плотностей контактов в 3`- и 5`-интервалах, равноотстоящих от начала гена.

ЗАКЛЮЧЕНИЕ

Любые методы полногеномного анализа дают с одной стороны большое количество информации, с другой - несут в себе много неспецифики и шума. Опубликованные протоколы по исследованию РНК-ДНК интерактома, включая метод Red-C, предоставляют для изучения новый тип данных. Главной ценностью методов “все-против-всех” является возможность изучения хроматин-ассоциированных РНК, не обладая при этом никакой априорной информацией о последовательности, типе, длине этих РНК. Проблемой данных из методов “все-против-всех” помимо неспецифических взаимодействй и детекции полимеразного следа является небольшое количество клеточных типов, на которых реализованы эти протоколы.

В данной работе предпринята попытка разработать такой подход к анализу данных методов “все-против-всех”, который мог бы быть применен не только к результатам протокола Red-C, но и для всех аналогичных протоколов.

Механизмы действия индивидуальных хроматин-ассоциированных РНК могут быть очень разными. В данной работе показано, что можно выделять группы РНК, основанные на характере взаимодействия РНК с хроматином в зависимости от удаленности локусов контактов от своего гена, по предпочтению к взаимодействию с тем или иным состоянием хроматина, а также учитывать уровень экспрессии хроматин-ассоциированных РНК. Таким образом можно выделить хаРНК, предпочитающие взаимодействовать вблизи от своего гена, полногеномно или же обладать специфическим паттерном взаимодействия с хроматином, как, например, выделенные XIST-подобные хаРНК.

Разметка генов не является единой, устоявшейся и точной даже для человека. Существует несколько крупных консорциумов, которые довольно часто публикуют очередные версии разметки генов, в которых появляются новые гены, а уже аннотированные могут поменять класс или уточнить свои границы. Исследование РНК, не представленных в актуальной разметке генов, вполне может привести к открытию действительно новых функциональных РНК. В данной

работе представлены новые ранее неаннотированные РНК, включая и антисенс РНК (Х-РНК). Возможно, что Х-РНК, в том числе подтвержденные консорциумом FANTOME, могут положить начало исследованию интересных хаРНК.

На сегодняшний день для многих клеточных линий опубликовано огромное количество данных разного генеза: изучение пространственной организации хроматина, экспрессионные профили, информация о сайтах связывания транскрипционных факторов и эпигенетических метках и многие другие. Несомненно эти данные хотелось бы интегрировать и комбинировать между собой, т.к. они позволяют смотреть на одну клеточную линию с разных сторон. Тем не менее, такие данные являются специфичными для конкретной клеточной линии и, как показано в работе на примере расчета хроматинового потенциала, могут вносить batch-эффект, если эксперименты сделаны на клетках одного типа, но в разных лабораториях.

Такие методы как Hi-C, RNA-seq и многие другие изначально были разработаны для исследования группы клеток, но со временем были реализованы их вариации применительно к единичным клеткам. К сожалению, пока что протоколы изучения РНК-ДНК интерактома в единичных клетках не представлены. Несомненно, появление таких протоколов - важный и логичный шаг в развитии области.

ВЫВОДЫ

1. Разработанный биоинформатический подход для анализа данных РНК-ДНК интерактома, полученных из экспериментов, основанных на лигировании расположенных близко в пространстве макромолекул, может быть применен к любым данным по изучению РНК-хроматиновых взаимодействий из протоколов “все-против-всех”.
2. Разработанная процедура голосования для аннотации РНК-частей контактов генами позволяет однозначно аннотировать ~20% РНК-частей, попавших в ситуацию конфликта генной аннотации.

3. На основании разработанной метрики хроматинового потенциала выявлено 1823 хроматин-ассоциированные РНК (Red-C; K562), которые взаимодействуют с хроматином значимо чаще, чем это ожидается, исходя из их уровня экспрессии.
4. Были выявлены неизвестные ранее хроматин-ассоциированных РНК (X-РНК). Произведена классификация обнаруженных РНК.
5. Исследован характер взаимодействия РНК с хроматином в зависимости от удаленности места контакта РНК от своего гена. Произведена соответствующая классификация хроматин-ассоциированных РНК.
6. Для мРНК показано наличие полимеразного следа - контакты мРНК предпочтительно располагаются в 3`-области от старта транскрипции.
7. Проведен сравнительный анализ основных этапов биоинформатического анализа для других полногеномных протоколов по изучению РНК-хроматиновых взаимодействий.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

2. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **81**, 145–166 (2012).

3. Fu, X.-D. Non-coding RNA: a new frontier in regulatory biology. *Natl. Sci. Rev.* **1**, 190–204 (2014).

4. Vance, K. W. & Ponting, C. P. Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet. TIG* **30**, 348–355 (2014).

5. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).

6. Chen, J. *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* **17**, 19 (2016).

7. Zharikova, A. A. & Mironov, A. A. [piRNAs: Biology and Bioinformatics]. *Mol. Biol. (Mosk.)* **50**, 80–88 (2016).

8. Yamanaka, S., Siomi, M. C. & Siomi, H. piRNA clusters and open chromatin structure. *Mob. DNA* **5**, 22 (2014).

9. Holdt, L. M., Kohlmaier, A. & Teupser, D. Molecular roles and function of circular RNAs in eukaryotic cells. *Cell. Mol. Life Sci. CMLS* **75**, 1071–1098 (2018).

10. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

11. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).

12. Sun, X. *et al.* Chromatin-enriched RNAs mark active and repressive cis-regulation: An analysis of nuclear RNA-seq. *PLoS Comput. Biol.* **16**, e1007119 (2020).

13. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47–62 (2016).

14. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.* **20**, 89–108 (2019).

15. Zhang, G. *et al.* Comprehensive analysis of long noncoding RNA (lncRNA)-chromatin interactions reveals lncRNA functions dependent on binding diverse regulatory elements. *J. Biol. Chem.* **294**, 15613–15622 (2019).

16. Maracaja-Coutinho, V. *et al.* Noncoding RNAs Databases: Current Status and Trends. *Methods Mol. Biol. Clifton NJ* **1912**, 251–285 (2019).

17. Ayupe, A. C. *et al.* Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome. *RNA Biol.* **12**, 877–892 (2015).

18. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).

19. Ryabykh, G. K. *et al.* [RNA-Chromatin Interactome: What? Where? When?]. *Mol. Biol. (Mosk.)* **56**, 275–295 (2022).

20. Sridhar, B. *et al.* Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Curr. Biol.* **27**, 602–609 (2017).

21. Li, X. *et al.* GRID-seq reveals the global RNA-chromatin interactome. *Nat. Biotechnol.* **35**, 940–950 (2017).

22. Bell, J. C. *et al.* Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *eLife* **7**, e27024 (2018).

23. Yan, Z. *et al.* Genome-wide colocalization of RNA-DNA interactions and fusion RNA pairs. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3328–3337 (2019).

24. Bonetti, A. *et al.* RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. *Nat. Commun.* **11**, 1018 (2020).

25. Gavrilov, A. A. *et al.* Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res.* **48**, 6699–6714 (2020).

26. Calandrelli, R. *et al.* Stress-induced RNA-chromatin interactions promote endothelial dysfunction. *Nat. Commun.* **11**, 5211 (2020).

27. Li, L. *et al.* Global profiling of RNA-chromatin interactions reveals co-regulatory gene expression networks in Arabidopsis. *Nat. Plants* **7**, 1364–1378 (2021).

28. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).

29. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).

30. Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA-RNA spatial interactions. *Nature* **582**, 432–437 (2020).

31. Nguyen, T. C. *et al.* Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. *Nat. Commun.* **7**, 12023 (2016).

32. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

33. Galitsyna, A. A. & Gelfand, M. S. Single-cell Hi-C data analysis: safety in numbers. *Brief. Bioinform.* **22**, bbab316 (2021).

34. Kugel, J. F. & Goodrich, J. A. Non-coding RNAs: key regulators of mammalian transcription. *Trends Biochem. Sci.* **37**, 144–151 (2012).

35. Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nat. Rev. Mol. Cell Biol.* **17**, 756–770 (2016).

36. Jalali, S., Singh, A., Maiti, S. & Scaria, V. Genome-wide computational analysis of potential long noncoding RNA mediated DNA:DNA:RNA triplexes in the human genome. *J. Transl. Med.* **15**, 186 (2017).

37. Huang, R. C. & Bonner, J. Histone-bound RNA, a component of native nucleohistone. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 960–967 (1965).

38. Bonner, J. & Widholm, J. Molecular complementarity between nuclear DNA and organ-specific chromosomal RNA. *Proc. Natl. Acad. Sci. U. S. A.* **57**, 1379–1385 (1967).

39. Beiderbeck, R. & Richter, G. Characterization of rapidly labelled RNA associated with DNA in Chlorella. *Arch. Mikrobiol.* **67**, 256–272 (1969).

40. Bynum, J. W. & Volkin, E. Chromatin-associated RNA: differential extraction and characterization. *Biochim. Biophys. Acta* **607**, 304–318 (1980).

41. Nickerson, J. A., Krochmalnic, G., Wan, K. M. & Penman, S. Chromatin architecture and nuclear RNA. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 177–181 (1989).

42. Borsani, G. *et al.* Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325–329 (1991).

43. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).

44. Papanicolaou, N. & Bonetti, A. The New Frontier of Functional Genomics: From Chromatin Architecture and Noncoding RNAs to Therapeutic Targets. *SLAS Discov. Adv. Life Sci. R D* **25**, 568–580 (2020).

45. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat. Rev. Mol. Cell Biol.* **19**, 143–157 (2018).

46. Werner, M. S. *et al.* Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. *Nat. Struct. Mol. Biol.* **24**, 596–603 (2017).

47. Simon, M. D. *et al.* The genomic binding sites of a noncoding RNA. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20497–20502 (2011).

48. 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. *Mol. Cell* **44**, 667–678 (2011).

49. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973 (2013).

50. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

51. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

52. 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.* **27**, 44 (2020).

53. Disteche, C. M. Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560 (2012).

54. Brockdorff, N. *et al.* Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351**, 329–331 (1991).

55. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* **21**, 198–206 (2014).

56. Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575–579 (2016).

57. Hasegawa, Y. *et al.* The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev. Cell* **19**, 469–476 (2010).

58. Minajigi, A. *et al.* Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, (2015).
59. Chen, C.-K. *et al.* Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science* **354**, 468–472 (2016).
60. Simon, M. D. *et al.* High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. *Nature* **504**, 465–469 (2013).
61. Gelbart, M. E. & Kuroda, M. I. Drosophila dosage compensation: a complex voyage to the X chromosome. *Dev. Camb. Engl.* **136**, 1399–1410 (2009).
62. Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115–118 (2011).
63. Meller, V. H. & Rattner, B. P. The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *EMBO J.* **21**, 1084–1091 (2002).
64. Quinn, J. J. *et al.* Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.* **32**, 933–940 (2014).
65. Chu, H.-P. *et al.* TERRA RNA Antagonizes ATRX and Protects Telomeres. *Cell* **170**, 86-101.e16 (2017).
66. Marión, R. M. *et al.* TERRA regulate the transcriptional landscape of pluripotent cells through TRF1-dependent recruitment of PRC2. *eLife* **8**, e44656 (2019).
67. Chu, H.-P. *et al.* PAR-TERRA directs homologous sex chromosome pairing. *Nat. Struct. Mol. Biol.* **24**, 620–631 (2017).
68. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–124 (2011).
69. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
70. Micsinai, M. *et al.* Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res.* **40**, e70 (2012).

71. Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39 (2007).
72. Zhang, X., Hamblin, M. H. & Yin, K.-J. The long noncoding RNA Malat1: Its physiological and pathophysiological functions. *RNA Biol.* **14**, 1705–1714 (2017).
73. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
74. West, J. A. *et al.* The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell* **55**, 791–802 (2014).
75. Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell* **33**, 717–726 (2009).
76. Imamura, K. *et al.* Long noncoding RNA NEAT1-dependent SFPQ relocation from promoter region to paraspeckle mediates IL8 expression upon immune stimuli. *Mol. Cell* **53**, 393–406 (2014).
77. Hirose, T. *et al.* NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell* **25**, 169–183 (2014).
78. Murgatroyd, C., Hoffmann, A. & Spengler, D. In vivo ChIP for the analysis of microdissected tissue samples. *Methods Mol. Biol. Clifton NJ* **809**, 135–148 (2012).
79. Tian, B., Yang, J. & Brasier, A. R. Two-step cross-linking for analysis of protein-chromatin interactions. *Methods Mol. Biol. Clifton NJ* **809**, 105–120 (2012).
80. Xu, H. *et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads. *PloS One* **7**, e52249 (2012).
81. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

83. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* **30**, 892–897 (2001).

84. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

85. G Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol.* **17**, 28 (2016).

86. Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).

87. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* **45**, 814–825 (2012).

88. Groh, M. & Gromak, N. Out of balance: R-loops in human disease. *PLoS Genet.* **10**, e1004630 (2014).

89. Li, Y., Syed, J. & Sugiyama, H. RNA-DNA Triplex Formation by Long Noncoding RNAs. *Cell Chem. Biol.* **23**, 1325–1333 (2016).

90. Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nat. Biotechnol.* **39**, 1453–1465 (2021).

91. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

92. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

93. St Laurent, G. *et al.* VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol.* **14**, R73 (2013).

94. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-598 (2006).

95. Sai Lakshmi, S. & Agrawal, S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* **36**, D173-177 (2008).

96. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

97. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 139–144 (2010).

98. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

99. Potashnikova, D. M. *et al.* FACS Isolation of Viable Cells in Different Cell Cycle Stages from Asynchronous Culture for RNA Sequencing. *Methods Mol. Biol. Clifton NJ* **1745**, 315–335 (2018).

100. Stavrovskaya, E. D. *et al.* StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data. *Bioinforma. Oxf. Engl.* **33**, 3158–3165 (2017).

101. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

102. Imada, E. L. *et al.* Recounting the FANTOM CAGE-Associated Transcriptome. *Genome Res.* **30**, 1073–1081 (2020).

103. Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H. & Ohhata, T. Cell cycle regulation by long non-coding RNAs. *Cell. Mol. Life Sci. CMLS* **70**, 4785–4794 (2013).

104. Rashid, F., Shah, A. & Shan, G. Long Non-coding RNAs in the Cytoplasm. *Genomics Proteomics Bioinformatics* **14**, 73–80 (2016).