МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА

На правах рукописи

Жарикова Анастасия Александровна

Биоинформатический анализ РНК-хроматиновых взаимодействий

03.01.09 - математическая биология, биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Москва - 2022

Работа выполнена на факультете биоинженерии и биоинформатики Московского государственного университета им. М. В. Ломоносова.

Научный руководитель	**Миронов Андрей Александрович**, *доктор биологических наук, профессор*
Официальные оппоненты	**Орлов Юрий Львович**, *доктор биологических наук, профессор РАН, профессор кафедры информационных и интернет-технологий (КИИТ) Института цифровой медицины ФГАОУ ВО Первый МГМУ имени И.М. Сеченова Минздрава России (Сеченовский Университет)*
	Кулаковский Иван Владимирович, *доктор биологических наук, ведущий научный сотрудник Института белка РАН, Пущино*
	Медведева Юлия Анатольевна, *кандидат биологических наук, руководитель группы регуляторной транскриптомики и эпигеномики, старший научный сотрудник Федерального государственного учреждения «Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук»*

Защита диссертации состоится 29 июня 2022 года в 15 часов 00 минут на заседании диссертационного совета МГУ.03.04 Московского государственного университета им. М.В. Ломоносова по адресу: 119234, Москва, Ленинские горы, дом 1, стр. 73, Факультет биоинженерии и биоинформатики, ауд. 221.

E-mail: dissovet@belozersky.msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27) и на сайте ИАС “ИСТИНА”: https://istina.msu.ru/dissertations/460193362/

Автореферат разослан “ ” мая 2022 года.

Ученый секретарь диссертационного совета,
кандидат химических наук И.В. Шаповалова

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) (Djebali et al., 2012). Генные аннотации постоянно расширяются в основном за счет включения в них не только отдельных представителей нкРНК, но и целых новых классов нкРНК. Так, в 2006 году сразу несколькими группами были представлены короткие некодирующие РНК, взаимодействующие с белками класса PIWI (piРНК) (Zharikova et al., 2016), 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) (Holdt et al., 2018), консорциумом FANTOM было показано существование транскрипционного потенциала некоторых энхансерных областей (Andersson et al., 2014).

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 (Engreitz et al., 2016). 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 (Quinn et al., 2016). Упомянутые выше piРНК также работают в ядре, реализуя в том числе ко-транскрипционное подавление транспозонов (Ozata et al., 2019).

Несмотря на всю важность функций, в которых принимают участие длинные и короткие нкРНК, механизмы действий изучены лишь для некоторых из них. 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. Пристального внимания и изучения заслуживает каждый представитель этой группы.

Методы, применяемые для изучения РНК, ассоциированных с хроматином, существуют давно и постоянно развиваются. Еще в 50-х годах 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 выявить локусы ДНК, с которыми взаимодействуют РНК (Ryabykh et al., 2022). 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.

Степень разработанности темы исследования

За последние шесть лет появилось сразу несколько методов для изучения РНК-ДНК интерактома (рис. 1); такие подходы получили название “все-против-всех” (Sridhar et al., 2017; Li et al., 2017; Bell et al., 2018; Yan et al., 2019; Bonetti et al., 2020; Gavrilov et al., 2020; Calandrelli et al., 2020; Li et al., 2021).

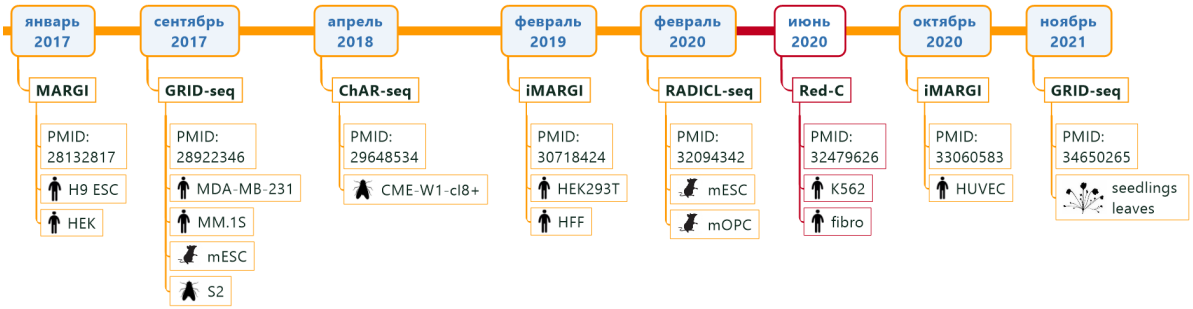

Рисунок 1. Временная шкала публикации работ, посвященных исследования РНК-ДНК интерактома.

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, что затрудняет их совместный анализ и сравнение результатов. 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 из Института биологии гена РАН принимала участие в обработке данных одного из методов типа “все-против-всех” - Red-C. Наиболее близкими к Red-C с точки зрения экспериментальной процедуры являются протоколы GRID-seq и RADICL-seq. 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.

Цели 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

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)*.

Апробация результатов

Полученные результаты были представлены на заседании ученого совета факультета биоинженерии и биоинформатики МГУ м. 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; ИТиС - 2017 в Уфе, Россия; FEBS Congress - 2018 в Праге, Чехия.

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.

Результаты и обсуждение

В данной работе мы предлагаем ознакомиться с многоступенчатым биоинформатическим подходом к анализу данных полногеномных РНК-ДНК взаимодействий. Разработанный подход позволяет начинать анализ с первичных чтений, до какой бы то ни было обработки, производить ряд фильтраций, учитывающих особенности экспериментальной подготовки, формировать РНК-ДНК контакты, идентифицировать хроматин-ассоциированные РНК 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.

Данные

Наиболее полный анализ представлен для протокола 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 (Li et al., 2017) и RADICL-seq (Bonetti et al., 2020), как наиболее близкие к Red-C с экспериментальной точки зрения.

Разнообразие и объем анализируемых данных по РНК-ДНК интерактому можно увидеть в таблице 1.

Таблица 1.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

Исходное количество чтений во всех экспериментах достаточно высокое (более 120 млн чтений для клеточного типа), а наблюдаемый разброс можно объяснить количеством представленных реплик. В процессе обработки контактов все реплики до момента аннотации РНК-частей генами были обработаны независимо. Перед этапом

аннотации реплики были объединены с целью увеличения количества данных и покрытия и далее исследовались совместно.

Протокол Red-C

Разработка биоинформатического протокола для обработки любых данных невозможна без детального изучения особенностей экспериментальной методики. Кратко опишем основные этапы экспериментальной методики Red-C (рис. 2). Клетки фиксируют формальдегидом (1%), для фрагментации хроматина была использована рестриктаза NlaIII. В результате получаются комплексы РНК-белок-ДНК, к которым лигируют особым образом сконструированный полярный линкер так, что к одному его концу специфически пришивается РНК, а к другому - ДНК. Также линкер содержит биотиновую метку для последующей экстракции полученной химерной молекулы и сайт для рестриктазы MmeI с помощью которой ограничивают длину ДНК-части (режет на ~20 нуклеотидов в сторону). Далее на матрице РНК достраивают вторую цепь, получая полностью двухцепочечный фрагмент, к которому пришивают необходимые адаптеры и секвенируют.

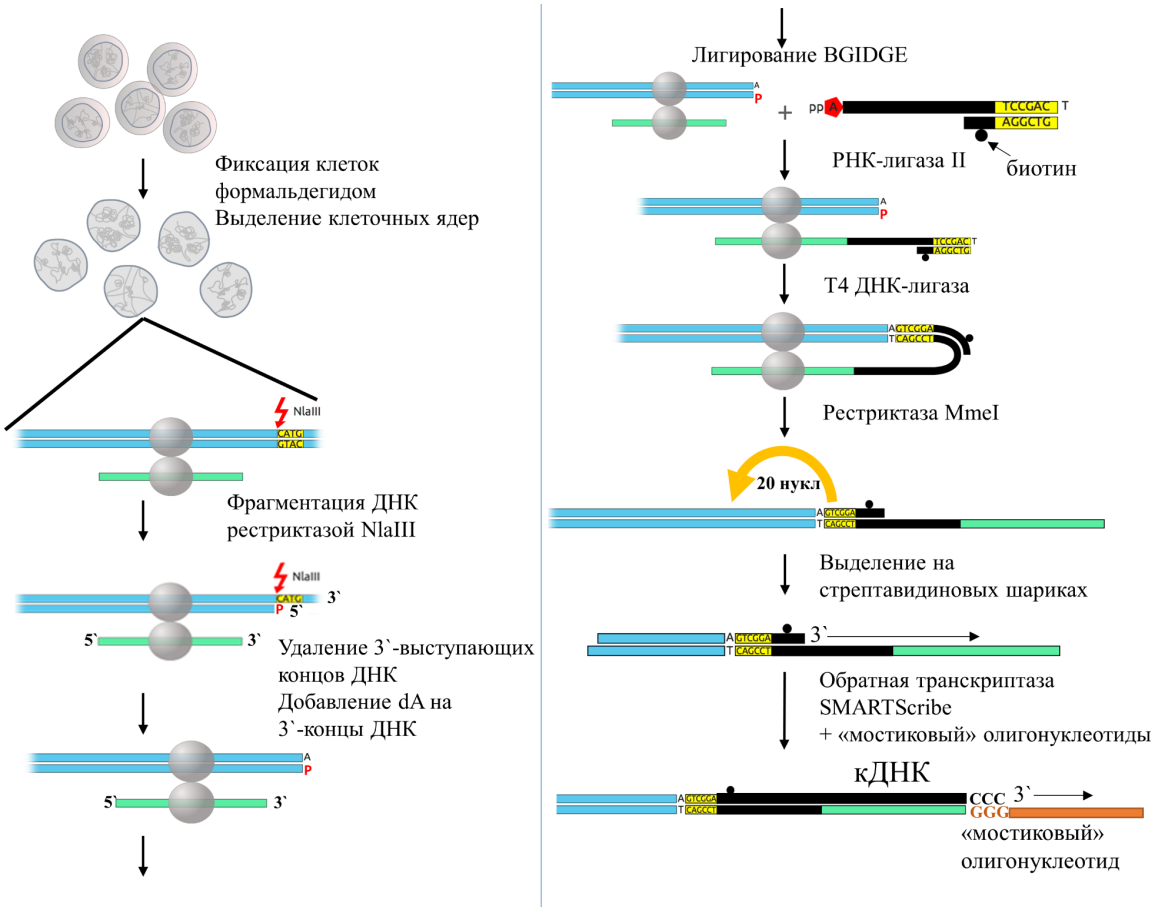

Рисунок 2. Схема экспериментального протокола Red-C.

Структура чтений библиотеки Red-C

По результатам секвенирования библиотек из протокола Red-C были получены парноконцевые чтения, 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 3.

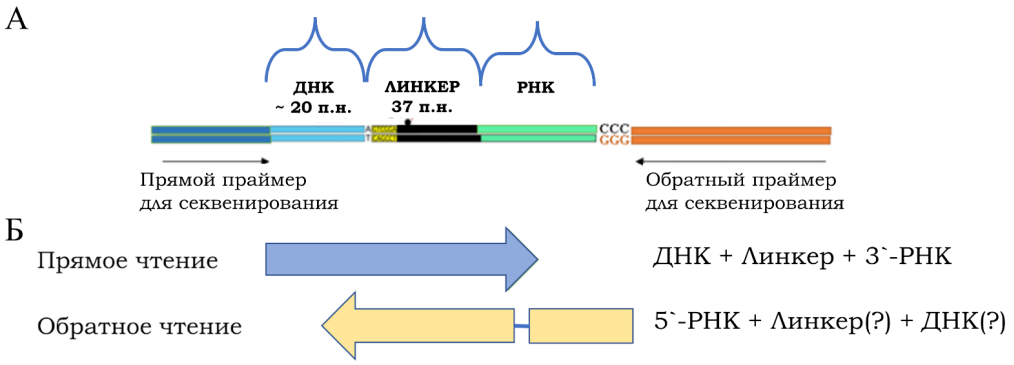

Рисунок 3. (А). Схема химерной конструкции для секвенирования, полученной в ходе эксперимента Red-C. (Б). Ожидаемая структура прямого и обратного прочтений.

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 хроматин-ассоциированные РНК любой длины. Таким образом, если удалось зафиксировать длинный фрагмент РНК (больше длины чтений), 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 исследования корректности работы протокола были поставлены контрольные эксперименты, в которых варьировали присутствие необходимых ферментов для лигирования ДНК, а также добавляли или не добавляли обработку РНКазой А (рис. 4).

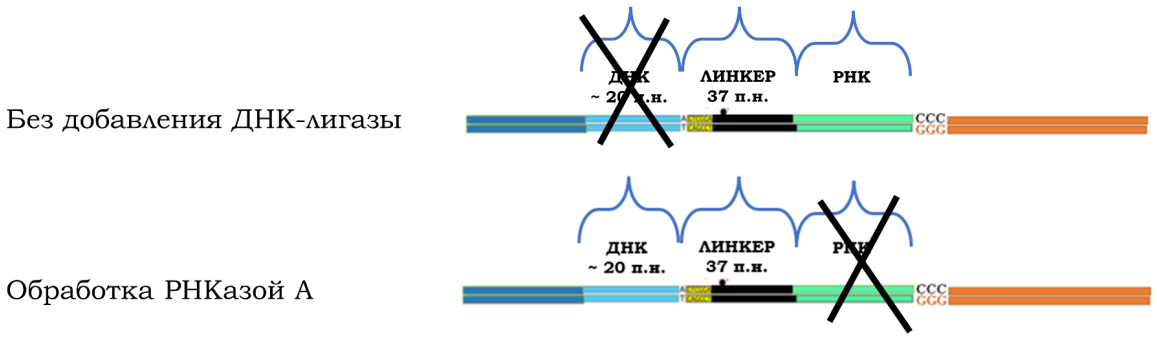

Рисунок 4. Ожидаемая структура химерной конструкции для контрольных библиотек (Red-C).

С помощью электрофоретического анализа удалось показать, что четкий фрагмент необходимого размера был обнаружен только в случае добавления всех ферментов, предусмотренных протоколом. Было проведено секвенирование контрольных библиотек и оказалось, что чтения с ожидаемой структурой составляли подавляющее большинство. 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.

Анализ данных РНК-ДНК интерактома

Сборка контактов РНК-ДНК

Всю процедуру биоинформатической обработки можно разделить на три многоступенчатых этапа (рис. 5):

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.

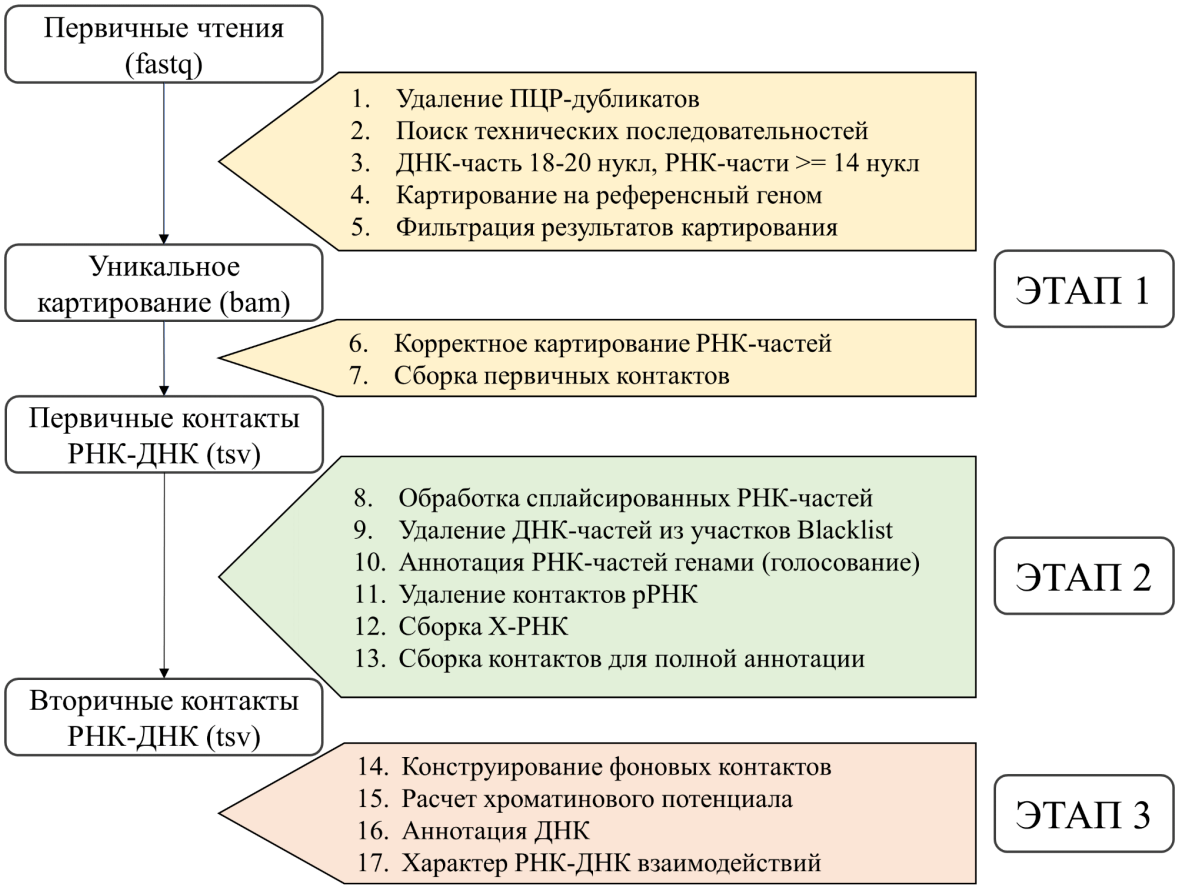

Рисунок 5. Схема анализа данных РНК-ДНК интерактома для протокола Red-C.

На примере эксперимента Red-C (клеточная линия K562) кратко разберем основные шаги протокола.

Этап №1. Первичная обработка результатов картирования. 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). Следующий шаг является одним из самых важных - поиск и удаление технических последовательностей. Основным критерием отбора пар чтений для дальнейшего анализа было наличие полной последовательность линкера в прямый чтениях. На данном этапе удалено еще ~ 10% данных. Для картирования на референсный геном отбирали ДНК-части контактов длиной 18-20 нуклеотидов (определено действием MmeI) и РНК-части контактов не менее 14 нуклеотидов в длину для более точного картирования в дальнейшем. На этом

шаге было потеряно почти 83 млн чтений (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. По причине плохого качества одной из частей отфильтровали еще ~ 12% чтений. Полученные части контактов (3`РНК, 5`РНК, ДНК) из эксперимента Red-C были независимо картированы на канонические хромосомы генома человека версий hg19 и hg38. 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 потерь на этом этапе было множественное картирование хотя бы одной из трех частей контакта.

Следующие фильтры суммарно отбраковали 6.7 млн контактов (~ 2%). Мы ожидаем, что 3`-РНК и 5`-РНК части контакта 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.

Данные из протоколов GRID-seq и RADICL-seq были проведены через Этап №1 с модификациями, учитывающими особенность экспериментов, сотрудниками нашей лаборатории.

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.

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. 2).

Таблица 2. 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

Рассмотрим некоторые характеристики полученных чтений. 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 (рис. 6). Также только для протокола Red-C удалось зафиксировать небольшое количество РНК-частей, картированных на референсный геном с разрывами.

Для всех протоколов более половины контактов детектированы между РНК и ДНК частями, пришедшими в разных хромосом, однако для Red-C процент таких контактов один из самых высоких (~80%).

Рассмотрим 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. 7). Для всех протоколов наибольшую плотность контактов можно наблюдать рядом с РНК-частью, с увеличением расстояния

от РНК-части, плотность контактов снижается (рис. 7), т.е. скорее всего мы детектируем процесс транскрипции. 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. 7Г). Это же наблюдение отмечают авторы статьи, что подтверждает возможность применения представленного протокола к другим данным типа “все-против-всех” без драматических искажений результатов.

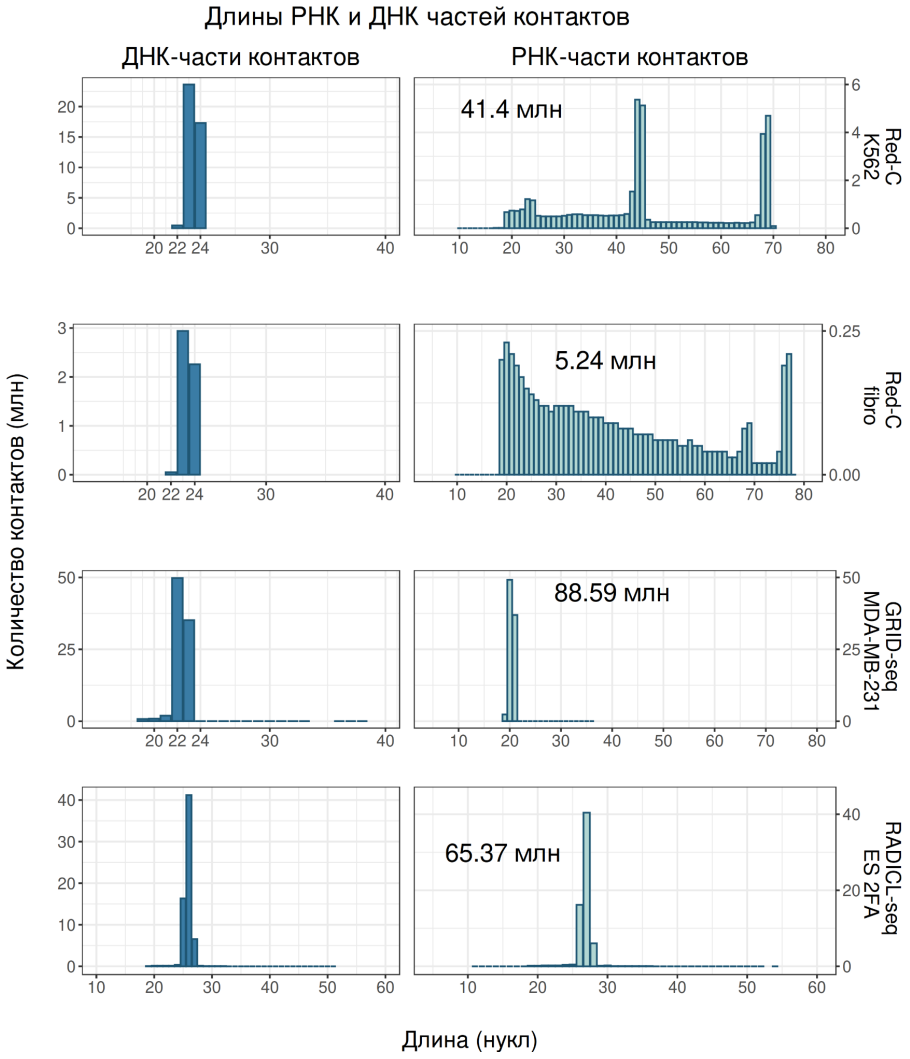

Рисунок 6. Распределение длин РНК и ДНК-частей контактов. Для протоколов GRID-seq и RADICL-seq приведено по одному примеру, остальные библиотеки выглядят аналогично.

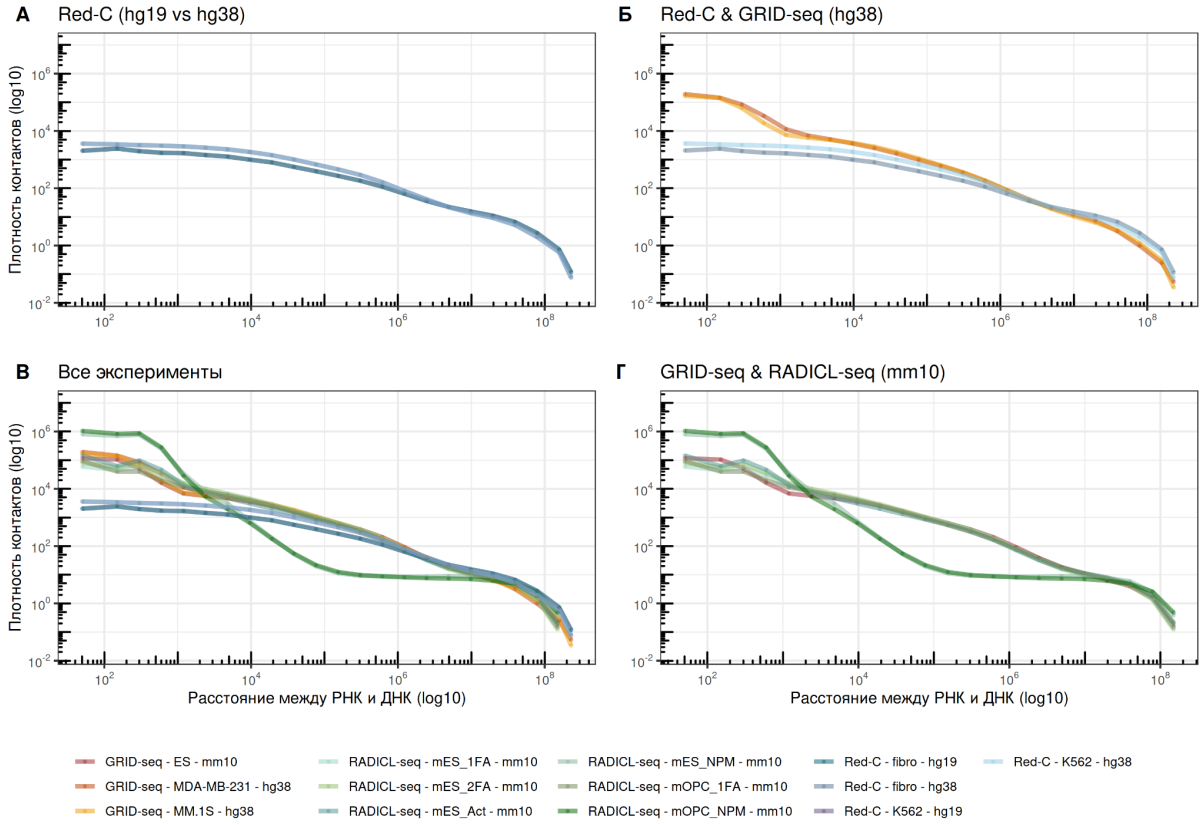

Рисунок 7. Зависимость плотности контактов РНК от расстояния между РНК-частью и местом контактов для внутрихромосомных контактов. (А) Протокол Red-C, клеточные линии K562 и фибробласты, картирование на референсные геномы версий hg19 и hg38. (Б) Протоколы Red-C и GRID-seq, картирование на референсный геном версии hg38. (В) Все эксперименты. (Г) Протоколы GRID-seq и RADICL-seq, картирование на референсный геном версии mm10.

Аннотация РНК-частей генами

Чтобы установить, какие именно РНК были детектированы как контактирующие с хроматином, было необходимо аннотировать РНК-части с помощью одной из существующих разметок генов.

Для генома человека за основу была выбрана разметка генов по версии GENCODE (release 35, GRCh38 (hg38); release 27, GRCh37 (hg19)), содержащая основные типы РНК. Было решено дополнить разметку GENCODE некоторыми треками малых РНК из геномного браузера и добавить разметку очень длинных некодирующих РНК (vlinc). Перед аннотацией данные о всех репликах внутри одной клеточной линии были объединены для увеличения покрытия. Для разрешения случаев попадания РНК-части контакта на пересечение двух или более генов была разработана процедура голосования, при которой предпочтение отдавалось гену с наибольшей

плотностью покрытия. По результатам аннотации оказалось, что большая часть РНК-фрагментов попала в генную разметку, для протокола Red-C процент аннотированных РНК-частей оказался максимальным, превысив 90% (рис. 8).

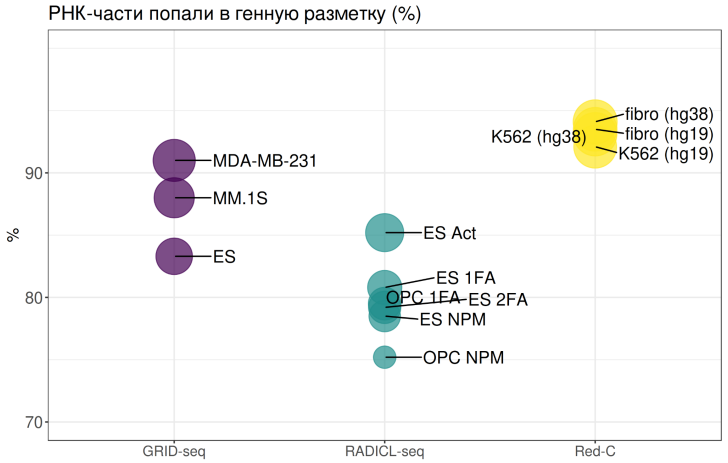

Рисунок 8. Процент РНК-частей контактов, попавших в генную разметку.

Более детальное исследование типов РНК показало, что для всех протоколов ~80% РНК пришли из белок-кодирующих генов, а в эксперименте RADICL-seq детектировано наибольшее количество рРНК (~20% от всех контактов). Для протоколов Red-C и GRID-seq было обнаружено менее 5% рРНК. Контакты, где РНК-часть контактов попала в рРНК, были удалены.

Сборка X-РНК

РНК-части контактов, которые не попали в генную разметку, были обработаны отдельно с целью поиска гипотетически новых РНК, которые могут находиться в плотной связке с хроматином. Неаннотированные РНК-части контактов были 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”. Полученные X-РНК добавлены к генной разметке, а попавшие в них контакты возвращены в общую выборку.

Полученные X-РНК были разделены по типам в зависимости от значений двух характеристик (рис. 9): близости к аннотированным генам и взаимной ориентации.

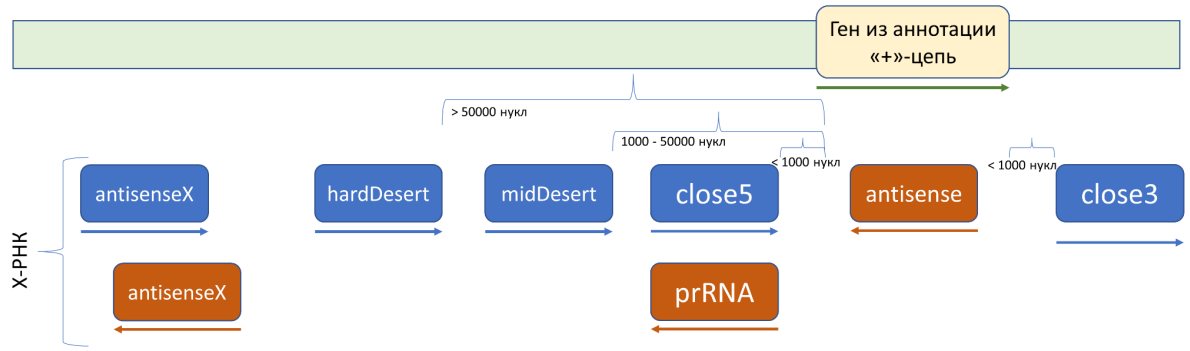

Рисунок 9. Схематичная иллюстрация правил, на основании которых X-РНК были разделены на классы.

Распределение X-РНК по классам можно увидеть на рисунке 10.

Распределение X-РНК по типам (%)

	GRID-seq hg38 MDA-MB-231	GRID-seq hg38 MM.1S	GRID-seq mm10 ES	RADICL-seq mm10 mES_1FA	RADICL-seq mm10 mES_2FA	RADICL-seq mm10 mES_Act	RADICL-seq mm10 mES_NPM	RADICL-seq mm10 mOPC_1FA	RADICL-seq mm10 mOPC_NPM	Red-C hg19 fibro	Red-C hg19 K562	Red-C hg38 fibro	Red-C hg38 K562
midDesert_prRNA_antisenseX	0.02	0.03	0.12	0.07	0.17		0.03	0.07	0.08		0.02		
midDesert_prRNA	0.62	0.5	0.93	0.45	0.38	1.85	0.3	0.39	0.23	0.56	0.36		0.92
midDesert_antisenseX_antisense	0.67	0.68	1.72	1.49	1.68	0.81	0.81	2.11	2.27		0.08		0.17
midDesert_antisenseX	0.57	0.47	1.19	1.3	1.76	0.35	0.84	2.06	2.39		0.11		0.14
midDesert_antisense	10.56	10.59	11.23	11.25	12.36	19.61	12.13	12.94	13.09	6.15	2.81	15.19	7.05
midDesert	8.77	7.41	12.33	8.17	9.78	13.49	11.12	6.69	6.14	4.47	2.49	9.49	8.12
hardDesert_prRNA_antisenseX	0.02		0.01				0.03	0.01					0.03
hardDesert_prRNA	0.14	0.08	0.19	0.13	0.07		0.07	0.12	0.05		0.02	0.63	0.06
hardDesert_antisenseX_antisense	0.05	0.03	0.16	0.15	0.14	0.12	0.03	0.38	0.4				0.03
hardDesert_antisenseX	0.26	0.24	0.22	0.16	0.17		0.17	1.1	1.22				0.08
hardDesert_antisense	3.01	2.65	2.58	2.12	2.27	3.69	2.08	2.76	2.93		0.02	1.9	2.49
hardDesert	3.46	2.97	5.17	2.14	2.26	1.15	2.32	2.83	2.65		0.02	6.96	6.97
close5_prRNA_antisenseX	0.03	0.03	0.06	0.03	0.03	0.12	0.03	0.05	0.04		0.04		
close5_prRNA	0.31	0.47	0.47	0.26	0.2	0.81	0.24	0.31	0.2	3.35	1.67	1.27	0.36
close5_close3_prRNA_antisenseX									0.02		0.04		
close5_close3_prRNA			0.02	0.01	0.01		0.03	0.01	0.01		1.14		
close5_close3_antisenseX_antisense		0.03	0.02	0.06	0.01			0.02	0.06		0.17		
close5_close3_antisenseX			0.06		0.01			0.03	0.06		0.13		
close5_close3_antisense	0.07	0.13	0.16	0.13	0.12	0.23	0.13	0.09	0.12	4.47	7.84		0.17
close5_close3	0.15	0.18	0.3	0.29	0.21	0.23	0.34	0.14	0.19	0.56	3.02		0.36
close5_antisenseX_antisense	0.6	0.55	0.84	0.76	0.92	0.35	0.77	1.87	2.4		0.61		0.11
close5_antisenseX	0.5	0.37	1.01	1.11	1.29	0.58	0.44	1.95	1.96		0.23		0.22
close5_antisense	4.9	6.25	4.96	5.86	5.21	7.04	5.24	6.55	6.9	26.82	18.72	9.49	5.21
close5	5.61	6.94	6.55	7.18	6.51	12	6.49	5.11	4.87	13.41	9.36	8.23	7.22
close3_prRNA_antisenseX	0.02	0.03	0.05	0.06	0.07		0.07	0.11	0.08		0.04		
close3_prRNA	0.84	0.71	0.5	0.51	0.63	0.46	0.5	0.6	0.54	1.12	2.72	0.63	0.78
close3_close5_prRNA_antisenseX											0.04		
close3_close5_prRNA	0.02	0.03	0.02	0.03	0.03		0.1	0.02	0.02	1.12	1.37		0.06
close3_close5_antisenseX_antisense	0.02	0.03	0.02	0.01	0.04			0.03	0.04		0.13		
close3_close5_antisenseX	0.02		0.03	0.01	0.01			0.04	0.11		0.02		0.03
close3_close5_antisense	0.05	0.13	0.18	0.22	0.17		0.37	0.14	0.18	2.79	9.53		0.11
close3_close5	0.14	0.11	0.33	0.31	0.22	0.23	0.4	0.18	0.23		2.85		0.2
close3_antisenseX_antisense	3.4	2.26	3.85	4.94	5.06	0.35	2.72	6.83	7.19		0.87		1.54
close3_antisenseX	1.43	1.1	2.76	2.2	2.1	0.58	1.14	2.63	2.9		0.27		0.7
close3_antisense	25.58	27.65	18.71	24.96	22.7	17.42	26.85	22.2	21.75	27.93	23.74	31.01	29.5
close3	28.18	27.36	23.24	23.57	23.35	18.45	24.19	19.63	18.66	7.26	9.53	15.19	27.37

Рисунок 10. Распределение X-РНК по классам (в процентах для каждого эксперимента).

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.

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 (Imada L. et al., 2020) показало, что 104 из 112 (93%) X-РНК пересекаются с РНК, аннотированными в FANTOM, которые были заявлены как новые РНК. Для протокола Red-C (hg19) были оставлены только X-РНК, находящиеся в отдалении от известных генов. Для клеточной линии K562 осталось 1867 таких X-РНК.

По окончании Этапа №2 мы получили финальную выборку контактов. Для протокола Red-C (K562) представлена диаграмма истощения, где проиллюстрированы основные этапы фильтрации (рис. 11).

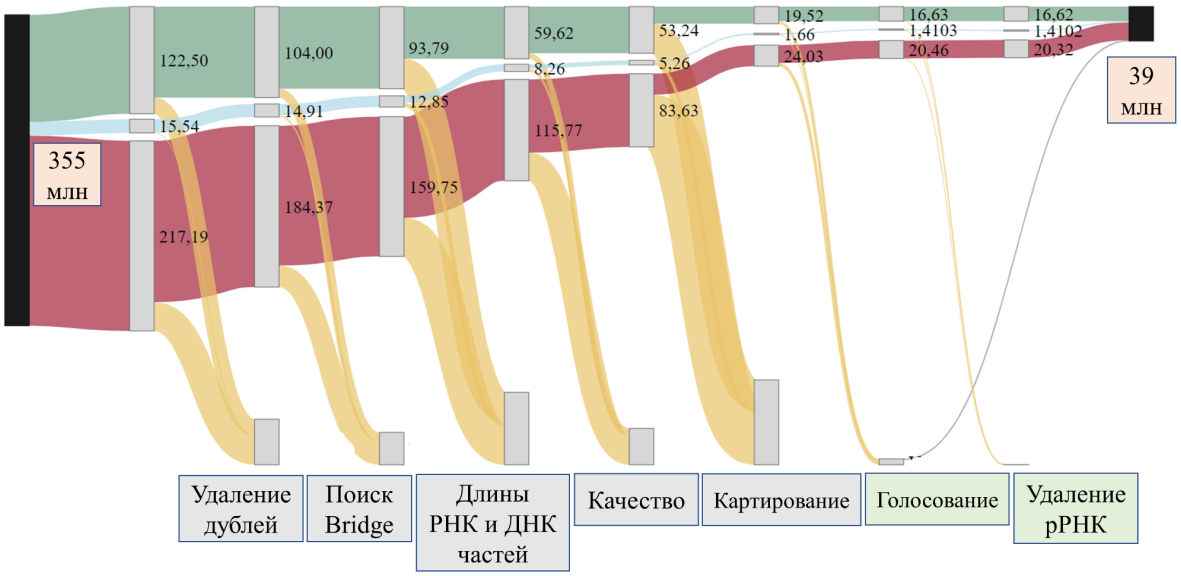

Рисунок 11. Диаграмма истощения для протокола Red-C (K562; hg38). Лентами зеленого, голубого и малинового цветов обозначены отдельные реплики; лентой желтого цвета обозначены контакты, не прошедшие очередной фильтр. Названия фильтров указаны в нижней части рисунка в боксах серого (относятся к Этапу №1) 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. 12), 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). 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.

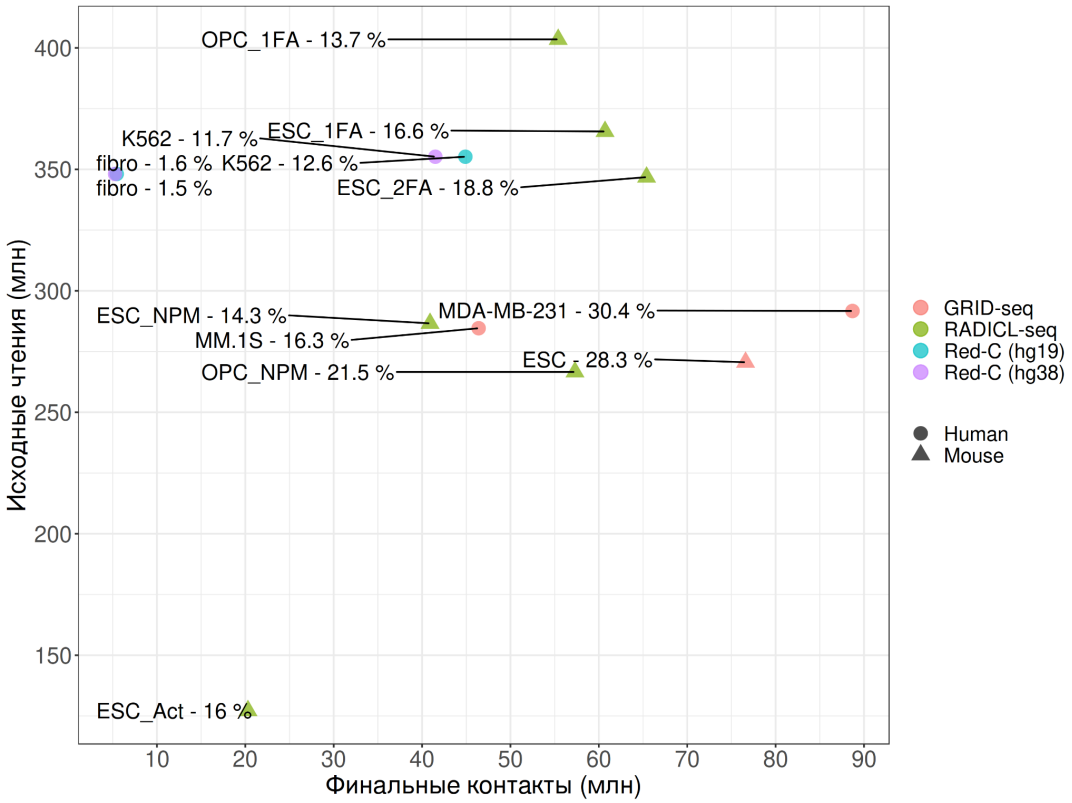

Рисунок 12. Соотношение количества исходных чтений к количеству контактов, прошедших Этап №1 и №2. Протоколы Red-C, GRID-seq, RADICL-seq. Реплики объединены.

Конструирование фона

Согласно предложенному в протоколе GRID-seq подходу построения эндогенного фона фоновые контакты были определены как контакты белок-кодирующих генов с “не материнскими” хромосомами. Для экспериментов протокола Red-C в фон попало более половины всех контактов.

Для каждого участка референсного генома длиной в 500 нуклеотидов было рассчитано суммарное количество фоновых контактов, полученный сигнал сглажен с помощью программы Stereogene. Затем каждый индивидуальный РНК-ДНК контакт был нормирован на значение фона согласно координате ДНК-части контакта. Дополнительно полученные веса были перенормированы так, чтобы сумма контактов до нормировки на фон была равна сумме контактов после нормировки. Полученные нормированные на фон веса контактов были использованы далее при изучении характера взаимодействия РНК с хроматином.

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. 13). Была 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 оценки статистической значимости в качестве фонового распределения chP были использованы значения chP для мРНК.

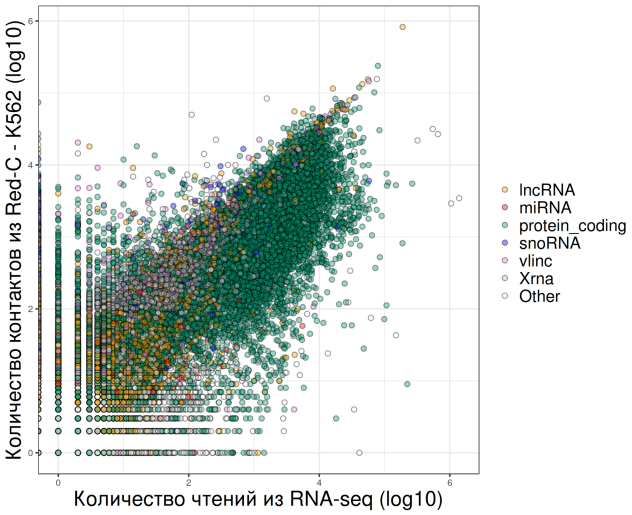

Рисунок 13. Зависимость количества контактов РНК с хроматином от уровня их экспрессии.

На рисунке 14 можно увидеть распределение по классам РНК с высоким хроматиновым потенциалом для эксперимента Red-C (клеточная линия K562; hg38).

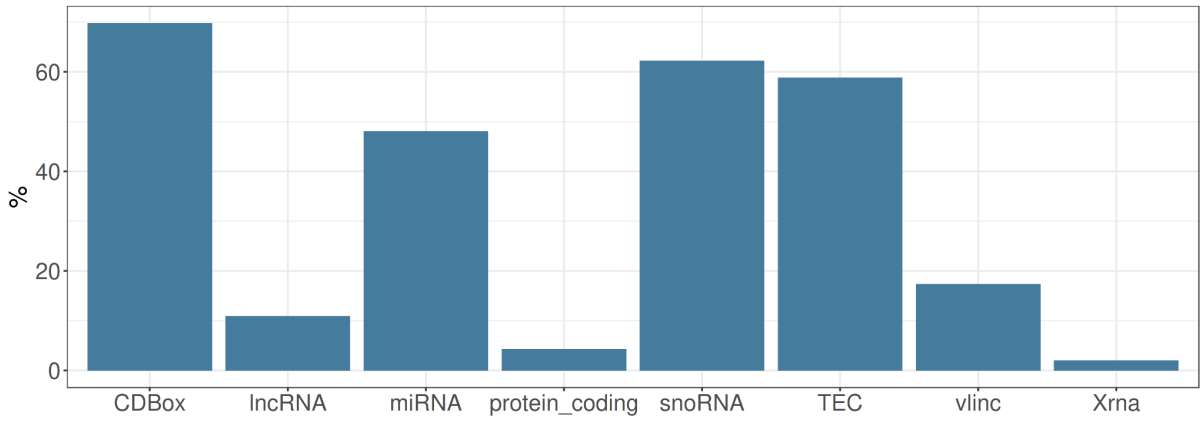

Рисунок 14. Распределение РНК с значимо высоким хроматиновым потенциалам (количество контактов > 100; поправленный p-value < 0.1) по классам. Для каждого класса представлен процент РНК с высоким chP к общему количеству РНК в классе. Протокол Red-C, клеточная линия K562, версия референсного генома - hg38.

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 Х-РНК обладают высоким 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.

К сожалению, для протоколов кроме Red-C авторы не предоставили данных о секвенировании РНК. Для расчета хроматинового потенциала для других протоколов 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. В результате тенденция в представленности классов РНК среди хаРНК с высоким значением chP сохраняется.

Исследование характера взаимодействия РНК с хроматином

Для изучения того, с какими именно участками хроматина контактируют найденные РНК, ДНК-части также могут быть аннотированы какой-либо разметкой, позволяющей охарактеризовать свойства хроматина. Для каждой РНК были рассчитаны непересекающиеся интервалы, разбивающие геном на участки разноудаленные от гена (рис. 15): 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).

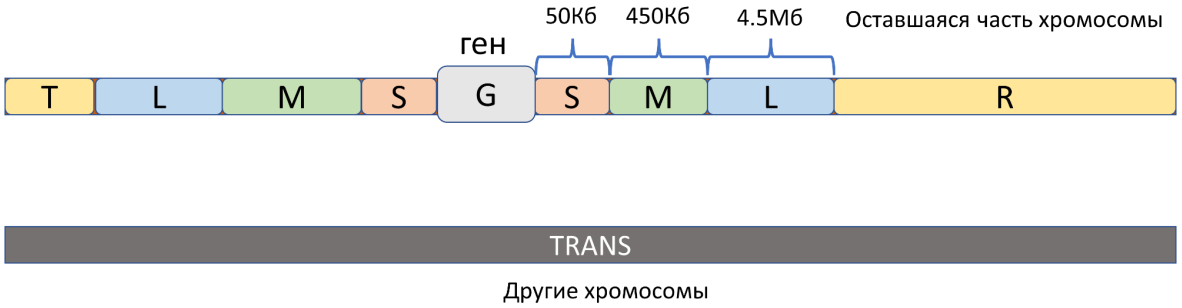

Рисунок 15. Схема конструкции непересекающихся интервалов вокруг контактирующей с хроматином РНК в зависимости от их удаленности от соответствующего гена.

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 (состояния хроматина определены по разметке из работы Ernst et al, 2014) 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. 16).

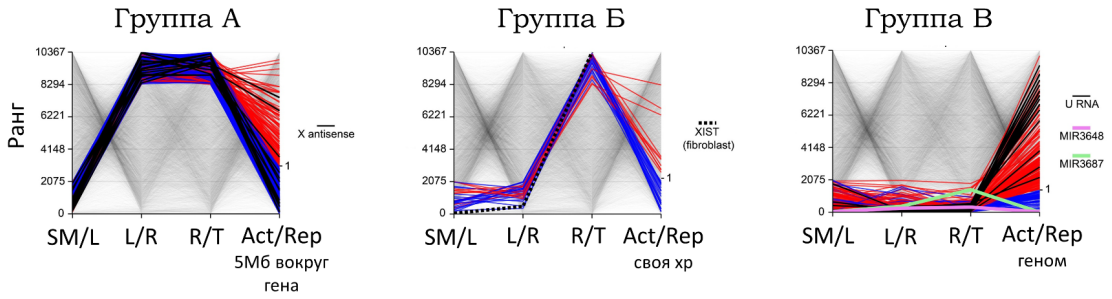

Рисунок 16. Группы РНК, выделенные по характеру взаимодействия с хроматином: группа А: контактирующие преимущественно недалеко от своего гена, 313 РНК, отношение Act/Rep рассчитано для интервала в 5Мб вокруг своего гена; группа Б: предпочитающие взаимодействовать с “родной” хромосомой аналогично XIST, 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).

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 Х-РНК. В группу В входят РНК, контактирующие полногеномно, включая MALAT1 и малые РНК разных классов. 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. 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.

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. 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).

Сосредоточившись на белок-кодирующих генах было показано, что мРНК чаще контактируют с областями, следующими за стартом транскрипции (по ходу транскрипции), чем с предшествующими ему, что можно объяснить тем, что РНК тянется за РНК-полимеразой в процессе транскрипции гена (рис. 17).

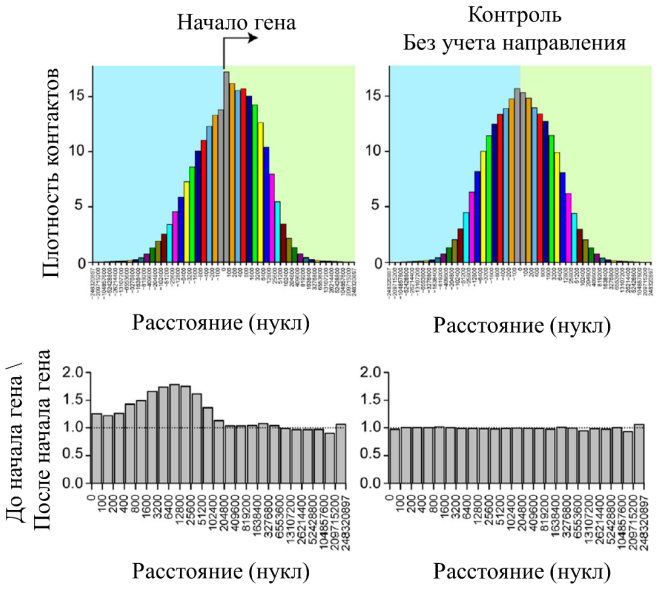

Рисунок 17. Частота контактов фрагментов мРНК с интервалами по ходу (слева) или в обратную сторону (справа) относительно направления транскрипции. Пары столбцов одного цвета представляют результаты для равноотстоящих от начала гена интервалов. Ниже показаны соотношения частот между равноотстоящими от начала гена интервалами.

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. Было выявлено 3572 неизвестных ранее хроматин-ассоциированных РНК (Х-РНК) (Red-C; K562). 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.

Научные статьи по теме диссертации, опубликованные в журналах SCOPUS, WOS, RSCI[1]

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

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

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).

Другие научные работы, опубликованные по теме диссертации:

1. Рябых Г., Миронов А., **Жарикова А.**, Сигорских А., Коростелев Ю. Инструменты сравнительного анализа РНК-ДНК интерактома клеток. Сборник трудов 44-й междисциплинарной школы-конференции ИППИ РАН. Москва, 2020, 342-343.
2. Gavrilov, A., **Zharikova, A.**, Galitsyna, A., Razin, S., and Mironov, A. A whole-genome map of chromatin-bound RNAs. (2018). FEBS open bio 8, Suppl 1, 137–137.
3. **Жарикова А.**, Галицына А., Гаврилов А., Логачева М., Разин С., Миронов А. Исследование свойств РНК-ДНК контактов в хроматине. Сборник трудов 42-й междисциплинарной школы-конференции ИППИ РАН "Информационные технологии и системы 2018", 656-658.
4. **Жарикова А.**, Галицына А., Гаврилов А., Миронов А., Разин С. Взаимодействие РНК и ДНК в хроматине. Сборник трудов 41-й междисциплинарной школы-конференции ИППИ РАН "Информационные технологии и системы 2017", 516-517.