

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ

ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ МНОЖЕСТВЕННО-КАРТИРОВАННЫХ ЧТЕНИЙ ПРИ
ИЗУЧЕНИИ ДНК-РНК ИНТЕРАКТОМА

FUNCTIONAL ANALYSIS OF MULTI-MAPPED READS IN DNA-RNA INTERACTOME
STUDIES

Курсовая работа студента 3-го курса
Косимова Мухаммадфирдавса Назаровича

Научные руководители:
доцент ФББ МГУ, к.б.н.
А.А.Жарикова

проф ФББ МГУ., к.ф.-м.н.,
д.б.н. А.А.Миронов

Москва – 2024 г.

1. ОГЛАВЛЕНИЕ

1. ОГЛАВЛЕНИЕ.....	2
2. СПИСОК СОКРАЩЕНИЙ.....	3
3. ВВЕДЕНИЕ.....	4
4. ЦЕЛИ И ЗАДАЧИ.....	5
5. ОБЗОР ЛИТЕРАТУРЫ.....	6
5.1 Повторы в геноме.....	6
5.1.1. Не транспозонные элементы.....	6
5.1.2. Транспозоны.....	7
5.1.2.1. Ретротранспозоны.....	8
5.1.2.2. ДНК-транспозоны.....	10
5.2. T2T.....	12
5.3. Repbase Update.....	14
5.4. Программы для анализа повторяющихся элементов генома.....	15
5.4.1. Repeatmasker.....	15
5.4.2. mHi-C.....	16
5.4.3. TEtranscripts.....	17
5.5. ДНК-РНК интерактом.....	18
6. МАТЕРИАЛЫ И МЕТОДЫ.....	21
6.1. Данные РНК-ДНК интерактома.....	21
6.2. Картирование.....	21
6.3. Обработка результатов картирования.....	21
6.4. Аннотация.....	21
6.5. Голосование.....	21
7. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ.....	23
7.1. Анализ аннотаций.....	23
7.2. Картируемость.....	24
7.3. Уникальные РНК-части - любая ДНК-часть.....	27
7.3.1 Уникальная РНК - любая ДНК, картированная на один класс повторов.....	28
7.3.2 Уникальная РНК - множественная ДНК, картированная на один класс повторов.....	29
7.3.3 Анализ обогащения.....	32
7.4. Множественные РНК - любая ДНК.....	33
7.4.1 Множественные РНК, картированные на один биотив - любые ДНК, картированные на один класс повторов.....	33
7.4.2 2-9 раз картированная РНК - любая ДНК, попавшая на повтор.....	34
7.4.3 10+ раз картированная РНК - любая ДНК, попавшая на повторы.....	36
8. ВЫВОДЫ.....	38
9. СПИСОК ЛИТЕРАТУРЫ.....	39

2. СПИСОК СОКРАЩЕНИЙ

NGS — next generation sequencing, высокопроизводительное секвенирование

bp — base pair, пара оснований

kbp — kilobase pairs, тысяча пар оснований

Mbp — megabase pairs, миллион пар оснований

УКД — уникально картированная ДНК

МКР — множественно картированная РНК

МКД — множественно картированная ДНК

УКР — уникально картированная РНК

МКП — множественно картированное прочтение

УКП — уникально картированное прочтение

lncRNA — long noncoding RNA, длинная некодирующая РНК

НТЭ — не транспозонные элементы

STR — short tandem repeats, короткие tandemные повторы

rRNA — ribosomal RNA, рибосомальная РНК

snRNA — small nuclear RNA, малая ядерная РНК

snoRNA — small nucleolar RNA, малая ядрышковая РНК

rDNA — ribosomal DNA, рибосомальная ДНК

кДНК - ДНК, полученная в результате обратной транскрипции

LINE — long interspersed nuclear elements, длинные диспергированные ядерные элементы

SINE — short interspersed nuclear elements, короткие диспергированные ядерные элементы

MITE — miniature inverted-repeat transposable element, миниатюрный транспозируемый элемент с инвертированными повторами

TIR — terminal inverted repeats, терминальные инвертированные повторы

tRNA — transfer RNA, транспортная РНК

LTR — long terminal repeats, длинные концевые повторы

TRPT — target-primed reverse transcription, таргет-специфичная обратная транскрипция

RC — повторы, транспозирующиеся по механизму катящегося кольца

GRCh38 — Genome Reference Consortium

T2T — Telomere-to-Telomere

RU — Repbase Update

ТАД — топологически ассоциированные домены

EM — Expectation-Maximization

3. ВВЕДЕНИЕ

В геномах эукариот, в частности человека, представлено большое разнообразие повторяющихся элементов, которые могут составлять более половины общей длины всего генома. В геноме человека повторяющиеся элементы занимают около 50% суммарной длины. До недавних пор повторяющиеся элементы генома считались “мусорными” в связи с их затруднительным анализом и, как результат, обедненной информации об их функциональной значимости. С развитием технологий геномного и транскриптомного секвенирования и алгоритмов анализа данных, все больше становится известно о широком спектре клеточных процессов, в которых задействованы повторяющиеся элементы в геноме [1].

Присутствие повторов усложняет анализ данных высокопроизводительного секвенирования (NGS), поскольку последовательности прочтений из этих областей могут быть короче самого повтора и, следовательно, могут быть картированы на несколько мест в геноме. Большинство существующих алгоритмов для анализа данных NGS не способны эффективно обрабатывать множественные картировки, что влечет за собой потерю существенной части информации и затрудняет биологическую интерпретацию результатов.

Существующие референсные сборки геномов эукариот практически не включают локусы, содержащие такие типы повторов, как, например, центромерные и теломерные, так как они состоят из повторяющихся массивов сателлитных повторов. Проблема определения последовательности повторяющихся элементов была частично решена с разработкой революционных протоколов секвенирования третьего поколения длина прочтений которых может достигать 15 тысяч пар оснований (15 kbp) на платформе PacBio и 4 миллионов пар оснований (4 Mbp) на Oxford Nanopore Sequencing (ONT) [2]. Например, геномная сборка человека T2T (telomere to telomere) была получена с использованием этих протоколов и включает в себя полный набор tandemных и центромерных повторов, а также полностью разрешенные последовательности акроцентрических хромосом [3]. Тем не менее, многие распространенные в практике протоколы секвенирования дают короткие чтения. Например, протокол GRID-seq [4,5] для анализа полногеномного интерактома между ДНК и РНК генерирует химерные прочтения длиной ~85 пар оснований (bp). Стандартные биоинформационные программные конвейеры, используемые для анализа ДНК-РНК интерактомов подразумевают использование только уникально картированных прочтений (УКП), что приводит к потере трех типов контактов: уникально-картированная ДНК-часть: множественно-картированная РНК-часть (УКД:МКР), множественно-картированная ДНК-часть: уникально-картированная РНК-часть (МКД:УКР), МКР:МКД. В различных экспериментах потеря данных, связанная с игнорированием множественно-картированных

прочтений (МКП), может составлять больше половины всех контактов. В основном теряются контакты, приходящиеся на повторяющиеся участки генома, что способствует обеднению информации об участии повторов в ДНК-РНК контактах.

4. ЦЕЛИ И ЗАДАЧИ

Цель:

Анализ полногеномного ДНК-РНК интерактома человека из данных GRID-seq с учетом МКП

Задачи:

- разработка протокола, позволяющего учитывать МКД при анализе контактов между УКР и МКД
- выявление тенденций во взаимодействиях разных биотипов РНК с повторяющимися участками генома в зависимости от картируемости попадающих на них прочтений

5. ОБЗОР ЛИТЕРАТУРЫ

5.1 Повторы в геноме

В геномах эукариот содержится большое количество повторяющихся элементов, которые до недавних пор считались “мусорной ДНК”. Однако все больше становится известно о различных процессах в которых вовлечены повторы в геноме [6]. Сейчас известно, что повторяющиеся элементы генома могут транскрибироваться: например длинная некодирующая РНК (lncRNA) TERRA, транскрибуемая из теломерных областей, регулирует активность теломеразы и поддерживает целостность теломер [7]. Транскрипты, считываемые с повторяющихся участков, могут транслироваться в составе 5' UTR, 3'UTR или инtronов по механизму Repeat-associated Non-ATG translation (RAN), вызывая тем самым ряд неврологических заболеваний [8]. Таким образом, с развитием технологий анализа геномных, транскриптомных и интерактомных данных, повторяющиеся элементы генома уже не считаются бесполезными.

Повторяющиеся элементы генома можно классифицировать на основании:

- частоты

повторяемости повторов: часто и умеренно повторяющиеся;

- организации структуры
- повторов : tandemные и диспергированные повторы; обильности : транспозонные элементы и не транспозонные элементы (НТЭ)

5.1.1. Не транспозонные элементы

Не транспозонные повторяющиеся элементы в геноме (НТЭ) - последовательности, не меняющие свое положение и не увеличивающие свою представленность в геноме. НТЭ в основном представлены tandemными повторами - короткими последовательностями, которые повторяются непосредственно друг за другом, образуя массивы tandemных повторов.

Тандемные повторы в зависимости от длины повторяющегося участка делятся на микросателлиты, также называемые простыми повторами или Short Tandem Repeats (STR), длина которых до 10 оснований, минисателлиты, длина которых до 100 оснований, и макросателлиты, длина которых может достигать нескольких тысяч пар оснований. Тандемные повторы составляют большую часть центромерных и теломерных частей хромосом и составляют около 8 % от суммарной длины генома человека [9].

Тандемные повторы встречаются как в конститтивном гетерохроматине, так и в кодирующих областях эухроматина, где изменения в последовательности или в количестве

повторяющихся элементов может приводить к изменению последовательности белка, его сворачиванию и как результат - изменению фенотипа [10]. Так, например, изменения количества единиц повторов в tandemных повторах белков клеточной стенки *Saccharomyces cerevisiae* и *Candida glabrata* приводят к модулированным свойства клеточной адгезии [11]. На данный момент известно более 60 заболеваний, вызываемых мутациями в tandemных повторах [9]. Например, увеличение количества остатков глицина в белке хантингтин, вызываемая экспансией tandemных повторов, способствует развитию заболевания Хантингтона [12]. Было показано, что почти все короткие tandemные повторы, ассоциированные с заболеваниями, находятся на границах топологически ассоциированных доменов [13].

Около 10% генома человека составляет такой класс повторов, как альфа-сателлитная ДНК, преимущественно расположенный в центромерных участках. Этот класс повторов транскрибируется в lncRNA, которые составляют значительную часть транскриптома [14].

К НТЭ также относятся гены некодирующих РНК: рибосомальная РНК (rRNA), малая ядерная РНК (snRNA), малая ядрышковая РНК (snoRNA) и др.

Рибосомальная ДНК (rDNA) представлена tandemными повторами, образующими массивы двух видов: 5S rDNA, расположенные одним массивом на хромосоме 1 и 45S rDNA массивы, расположенные на коротких плечах акроцентрических хромосом, образующих ядрышковые организаторы (хромосомы 13, 14, 15, 21, 22). Каждый повтор массива 45S rDNA состоит из 13 kb “рибосомальной части”, содержащей в себе будущие индивидуальные 18S, 5.8S, 28S rRNA, и 30 kb межгенного спейсера. Количество копий генов rRNA в геноме человека варьируется от 200 до 600 [15].

5.1.2. Транспозоны

Транспозоны - мобильные участки генома, способные независимо реплицироваться в клетке. Их размеры варьируются от 100 bp до 10 kb. Классически делятся на 2 класса: класс I - ретротранспозоны и класс II - ДНК-транспозоны.

Ретротранспозоны встраиваются в геном через РНК-интермедиат, который обратной транскриптазой превращается в комплементарную ДНК (кДНК) перед встройкой. Ретротранспозоны часто называют “cut-and-paste” элементами, так как обычно встраиваемый фрагмент не подвергается изменениям. ДНК-транспозоны же называют “cut-and-paste”, так как они вырезаются и встраиваются в новый геномный локус.

Транспозоны также делятся на автономные и неавтономные в зависимости от того способны ли они сами осуществлять свою транспозицию. Автономные элементы ее осуществляют путем транскрипции и трансляции ферментативной машинерии, необходимой для

транспозиции. Неавтономные элементы для передвижения используют ферменты автономных транспозонов [16]. Некоторые из неавтономных элементов образуются из автономных потерей транскрибуируемый машинерии : так например, не автономные элементы Miniature Inverted-repeat Transposable Elements (MITE'ы) из класса ДНК-транспозонов сохранили только терминальные инвертированные повторы (TIR's), т.е. сайты связывания транспозазы не описанного предкового автономного семейства ДНК-транспозонов [17]. Другие же *de novo* образуются из некодирующих генов, как например Short interspersed nuclear elements (SINE'ы) образуются из транспортной РНК (tRNA). SINE'ы также являются примером неавтономных транспозонов, так как для их транспозиции нужны ферменты, получаемые после транскрипции и трансляции Long interspersed nuclear elements (LINE'ы) [16].

5.1.2.1. Ретротранспозоны

Ретротранспозоны делятся на 3 подкласса: повторы с длинными концевыми повторами (LTR), занимающие ~8 % генома человека, таргет-инициируемые не-LTR повторы, занимающие примерно треть генома человека и TR повторы, транспозия которых осуществляется тирозин-рекомбиназой.

Таргет-инициируемые не-LTR повторы представлены тремя основными видами повторов: Alu, L1, SVA. В геноме человека более 500000 копий повторов L1, а длина каждого около 6 kb, суммарно они занимают около 17% генома человека. L1 повторы состоят из 3' UTR, 5'UTR с промотором РНК полимеразы II и двух открытых рамок считывания. Первая рамка ORF1 встречается у некоторых представителей не-LTR повторов и кодирует РНК-связывающий белок, который участвует в распознавании и перемещении транскрипта в ядро. ORF2 кодирует фермент с эндонуклеазной и РНК-зависимой ДНК полимеразной активностью. Этот комплекс ферментов обеспечивает таргет-специфичную обратную транскрипцию (TRPT) (см. рис. 1), что делает L1 единственным автономным видом повторов в геноме человека. Однако из всех копий L1, менее стабильно способны к транспозиции, остальные же не активны из-за внутренних перестроек и мутаций [19]. Часто этап обратной транскрипции прерывается до достижения 5'-конца, что приводит к потере промотора во встраиваемом повторе [18].

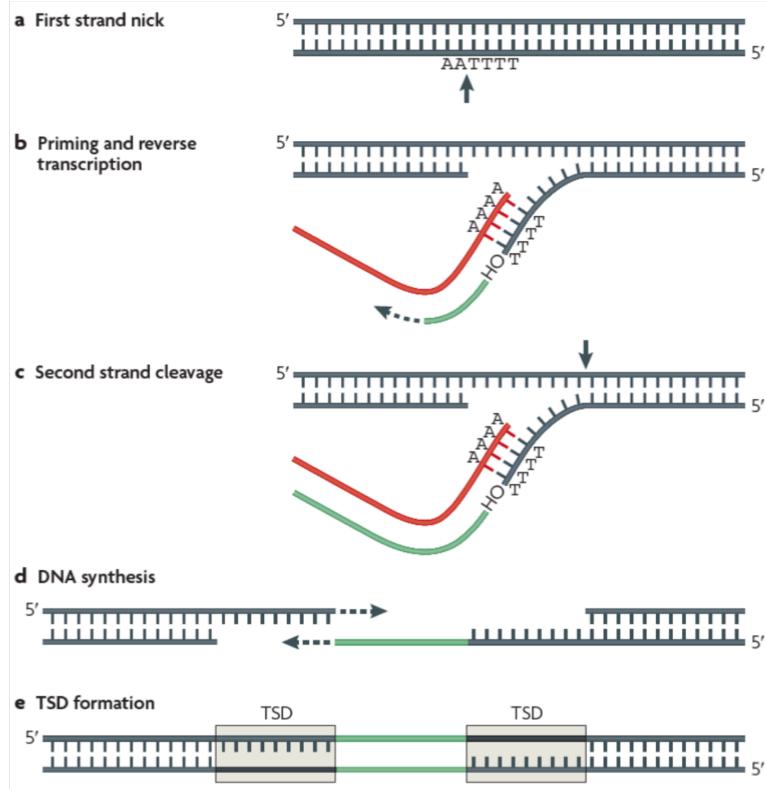


Рис. 1. Схема интеграции L1 элемента в геном механизмом таргет-специфичной обратной транскрипции. а) Считываемая с ORF2 эндонуклеаза вносит одноцепочечный разрыв в таргетную последовательность, оставляя 3'-поли-Т конец. б) Свободная гидроксильная группа на 3'-конце затем используется для проведения обратной транскрипции с встраиваемого L1 транскрипта (красный). в) Затем разрезается вторая цепь и д) свободный 3'-гидроксил используется для достройки цепи. Точный механизм перехода из в) в д) не известен. Аналогично происходит встройка Alu повторов и SVA [18]

В геноме человека количество Alu повторов значительно превышает миллион копий, каждая из которых длиной ~300 bp, что суммарно занимает ~11% всего генома человека. На 5'-конце находится промотор РНК-полимеразы III, но сигнала терминации нет, поэтому транскрипция продолжается до тех пор, пока не встретится терминатор. С получившегося транскрипта никаких ферментов не считывается, поэтому Alu повторы неавтономные. Для своей транспозиции они пользуются ферментами L1 повторов.

SVA повторы содержат ~3000 копий и длина каждой ~2 kb. Каждый повтор состоит из повторяющегося гексамерного участка, Alu-подобного региона, участка с варьирующим числом tandemных повторов, HERV-K10-подобного региона и сайта полиаденилирования. Предположительно SVA-повторы транскрибируются РНК-полимеразой II. Внутри самого повтора промотора нет, поэтому приходится рассчитывать на наличие промотора во

фланкирующем повтор участке. SVA также для своей транспозиции рассчитывают на машинерию L1 повторов [18].

Оставшиеся неактивные ретротранспозоны занимают ~ 6 % генома человека. Известно, что некоторые из них раньше были активны: L2 были автономными ретротранспозонами, а MIR-повторы пользовались их машинерией для своей транспозиции [20].

LTR повторы устроены сложнее и больше похожи на ретровирусы. Примером элементов LTR являются ERV1, известные как "эндогенные вирусы". Они содержат гены *gag* и *pol*, которые транскрибируются единой полицистронной РНК. Продуктами гена *pol* являются протеаза (PR), обратная транскриптаза (RT), рибонуклеаза Н и интеграза (IN). Сначала LTR элемент транскрибируется, затем для обратной транскрипции инкапсулируется в цитоплазме в белок Gag. Для инициации обратной транскрипции в качестве праймера используется tRNA, выполняющая роль затравки. Полученная кДНК встраивается в геном по механизму "cut-and-paste" с помощью интегразы.

Наименее изученным подклассом ретротранспозонов являются YR-ретротранспозоны. Они содержат длинные концевые повторы, чем похожи на подкласс LTR повторов, однако вместо интегразы в них закодирована тирозин-рекомбиназа. YR-ретротранспозоны подразделяются на несколько суперсемейств по последовательности терминальных повторов, но функции терминальных повторов и уникальной для подкласса тирозин-рекомбиназы все еще не известны [16].

Также кратко стоит упомянуть Penelope повторы, которые впервые были найдены как мутагенные агенты в *Drosophila virilis*. Они обладают двумя особенностями: наличие терминальных повторов, подобных LTR, и GIY-YIG мотивы в эндонуклеазе, которые не схожи ни с какими другими эндонуклеазами других ретротранспозонов. По результатам филогенетического анализа выяснили, что повторы Penelope далеки и от LTR-повторов и от не-LTR-повторов, а наиболее близки они к теломеразе, что свидетельствует о вероятно раннем ответвлении этих элементов в эволюции эукариот [21].

5.1.2.2. ДНК-транспозоны

Суммарно ДНК-транспозоны занимают ~3% генома человека [18]. На данный момент известно 4 подкласса ДНК-транспозонов: "cut-and-paste" элементы, транспозиция которых осуществляется DDE-транспозазой (названная так из-за наличия DDE аминокислот в активном центре), криптоны, хелитроны и маверики. Самыми простыми подклассами являются транспозоны DDE и криптоны, они обычно содержат одну рамку считывания, которая кодирует рекомбиназу, фланкированную короткими терминальными инвертированными повторами (TIRs) [16]. Транспозоны DDE являются наиболее

разнообразными и широко распространенными среди всех видов транспозонов, включая по меньшей мере 17 крупных суперсемейств, определенных по филогенетическим отношениям транспозаз [22–25]. Криптоны же редки у эукариот. Механизм DDE транспозации отличается между суперсемействами, но у всех эукариот с изученным механизмом DDE этот процесс инициируется транспозаз-катализированной нуклеофильной атакой молекулой воды на TIR'ы, вызывая тем самым разрывы в цепи. Хотя сам процесс перемещения такого транспозона не приводит к увеличению количества копий, данные подклассы могут размножаться в процессе репликации, перемещаясь от реплицированных участков на не реплицированные. К увеличению числа копий также приводит процесс гомологичной рекомбинации двухцепочечных разрывов, которые образуются при вырезании ДНК-транспозона, что приводит к увеличению числа копий.

Хелитроны широко распространены во многих эукариотических организмах, включая такие, как *Drosophila melanogaster*, *Caenorhabditis elegans*, и *Arabidopsis thaliana*. Хелитроны по большей части представлены не автономными элементами и не обладают стандартными характеристиками ДНК-транспозонов, как например TIR'ы. Поэтому до 2000-х годов хелитроны оставались малоизученными [26]. Также хелитроны отличаются от других ДНК-транспозонов тем, что их изначальная последовательность остается неизмененной, поэтому их механизм транспозации называют “peel-and-paste” (механизм “катящегося кольца” или RC), при котором сначала происходит отщепление смысловой цепи, за которым может следовать синтез антисмысловой цепи. По мере того, как ДНК сворачивается в круг к концу хелитрона, образуется ковалентно связанный круговой двухцепочечный ДНК-интермедиат (RC интермедиат) [26]. Затем RC вставляется в геном, увеличивая тем самым количество копий этого повтора. Однако не все хелитроны увеличивают количество своих копий. Известно, что хелитроны генома кукурузы способны непосредственно вырезаться из генома, что указывает на то, что еще предстоит работа по выяснению механизмов транспозации хелитронов [27].

Маверики (или полинтоны) — еще один малоизученный подкласс ДНК-транспозонов, особенностью которых является исключительно большая длина (15-20 Кб). Эти повторы содержат содержат до двадцати белок-кодирующих генов, фланкированных длинными (400-700-п.н.) TIR'ами [16]. Эти элементы широко распространены у эукариот, но обычно присутствуют в низком количестве копий (десятки на геном), за несколькими исключениями, как, например, у протиста *T. vaginalis*, где они занимают треть генома [28]. Было выяснено, что маверики имеют много сходств с разными группами ДНК-содержащих вирусов [29].

5.2. T2T

Референсный геном человека от Genome Reference Consortium (GRCh38) является наиболее актуальной версией генома человека на сегодняшний день. Для него есть множество доступной информации в UCSC: аннотация генов, проекты по анализу экспрессии генов, регуляция генов, аннотация повторов и др. [30]. Не смотря на высокое качество сборки от GRCh38, в ней содержится сотни гэпов и ошибок, которые возникли из-за сложностей при сборке повторяющихся участков генома. Суммарно 151 Mbp остаются не разрешенными в GRCh38, большая часть которых представлена центромерными и теломерными участками хромосом, недавними сегментными дупликациями и массивами rDNA [3]. С развитием протоколов секвенирования третьего поколения: PacBio HiFi и Oxford Nanopore, а также новых алгоритмов сборки консорциуму Telomere-to-Telomere (T2T) удалось разрешить все гэпы и получить полную версию референсного генома (T2T-CHM13). Таким образом удалось разрешить 8% генома человека. Сравнительная характеристика сборок представлена в таблице 1.

	GRCH38	T2T-CHM13	Разница ($\pm\%$)
Собрано оснований (Gbp)	2.92	3.05	+4.5
Неразмещенные основания (Mbp)	11.42	0	-100.0
Гэповые основания (Mbp)	120.31	0	-100.0
Число контигов	949	24	-97.5
NG50 контигов (Mbp)	56.41	154.26	+173.5

Таблица 1. Сравнительная статистика собранных последовательностей в GRCH38 и T2T-CHM13 версиях генома человека. Собрано оснований - все собранные не-N основания. Неразмещенные основания - те, чье положение не определено в хромосоме. NG50 контигов - это наибольшее значение, при котором контиги по крайней мере этого размера составляют более половины размера генома размером 3,05 Gbp [3].

Появление T2T-CHM13 дало первое полное представление о последовательности центромер, теломер, массивов tandemных повторов, сегментных дупликаций (segdups) и p-плеч акроцентрических хромосом в геноме человека. Консорциум T2T также позволил получить представление об организации и функции сегментных дупликаций, центромер,

эпигенетических особенностях повторов и генома, а также о генетической вариативности человека с помощью подходов сравнительной геномики и популяционной генетики [3,31–35]. В результате сборки и аннотации не распознанных ранее областей было обнаружено 3604 новых гена, большинство из которых являются активными паралогами других генов. Эти гены в основном локализуются в перицентромерных областях и коротких плечах акроцентрических хромосом.

Разрешение последовательностей коротких плеч акроцентрических хромосом позволило установить, что последовательности коротких дистальных плеч, располагающихся от теломер до массивов рДНК, схожи во всех акроцентрических хромосомах и содержат симметрично расположенные инвертированные сегментные дупликации и акроцентрические повторы. Проксимальные короткие плечи, расположенные между центромерами и массивами рДНК, довольно разнообразны: они содержат сегментные дупликации, составные массивы транспозонов и массивы сателлитов. Также было выяснено, что размеры массивов рДНК варьируются от 0.7 Mbp до 3,6 Mbp и располагаются в виде массивов, направленных от центромер к теломерам. Однако, несмотря на вариабельность размеров, все сателлитные массивы внутри акроцентрических хромосом имеют высокое сходство (более 90%).

Консорциум T2T помимо аннотации генов предоставил обновленную аннотацию повторов, полученную Repeatmasker'ом (о нем в следующих разделах), показав обогащение T2T-CHM13 последовательностями различных классов повторов в сравнении с GRCH38 (см. таблицу 2).

	GRCH38	T2T-CHM13	Разница ($\pm\%$)
Количество повторов (%)	51.89	53.94	
Количество повторов (Mbp)	1,516.37	1,647.81	+8.7
LINE	626.33	631.64	+0.8
SINE	386.48	390.27	+1.0
LTR	267.52	269.91	+0.9
Satellite	76.51	150.42	+96.6

DNA (ДНК-транспозоны)	108.53	109.35	+0.8
Simple repeat	36.5	77.69	+112.9
Low complexity	6.16	6.44	+4.6
Retroposon	4.51	4.65	+3.3
rRNA	0.21	1.71	+730.4

Таблица 2. Сравнительная статистика повторяющихся последовательностей в GRCH38 и T2T-CHM13 версиях генома человека. Информация о содержании различных классов повторов получена Repeatmasker'ом [3].

5.3. Repbase Update

В 1992 году появилась курируемая база данных Repbase Update (RU), которая изначально содержала информацию о повторяющихся элементах в геноме человека. Со временем RU стала обширной базой данных, содержащая прототипные последовательности повторов из геномов разных эукариот [36]. Repbase Update активно используется при анализе геномов эукариот, в частности исследований эволюции транспозонов и их влияние на геном [37–39]. В RU хранятся преимущественно транспозонные элементы, среди не транспозонных же элементов в базе данных хранятся мультикопийные гены (rRNA, tRNA, snRNA) и некоторые интегрированные вирусные последовательности. Однако RU обеднена информацией об этих повторах в сравнении со специальными базами данных для этих повторов: SILVA [40], GtRNADB [41], 5SRNADB [42].

На данный момент транспозонные элементы в RU делятся на 3 группы: ДНК транспозоны, LTR ретротранспозоны (включая ретровирусы) и не-LTR ретротранспозоны (включая SINE). Далее они подразделяются на 65 суперсемейств или клад (см. таблицу 3). Причем термин “суперсемейство” обычно относится к ДНК-транспозонам в то время как “клада” относится к LTR ретротранспозонам и не-LTR ретротранспозонам.

Группа	Суперсемейство/клада
ДНК транспозон	Academa, Cryptona (CryptonAa, CryptonFa, CryptonIa, CryptonSa, CryptonVa), Dadaa, EnSpm/CACTA, Ginger1a, Ginger2a, Harbinger, hAT, Helitron, IS3EUa, ISL2EU, Kolobok, Mariner/Tc1, Merlin, MuDR, Novosib, P, piggyBac, Polinton, Solaa (Sola1a,

	Sola2a, Sola3a), Transib, Zatora, Zisuptona
LTR ретротранспозон	BEL, Copia, DIRS, Gypsy, ERV1, ERV2, ERV3, ERV4a, Lentivirusa
не-LTR ретротранспозон	Ambala, CR1, CRE, Crack, Daphne, Hero, I, Ingi, Jockey, Kiria, L1, L2, L2A, L2B, Loa, NeSL, Nimb, Outcast, Penelope, Proto1, Proto2, R1, R2, R4, RandI/Dualen, Rex1, RTE, RTETP, RTEX, Tad1, Tx1, Vingia SINE (SINE1/7SL, SINE2/tRNA, SINE3/5S, SINE4a, SINEUa)

Таблица 3. Классификации транспозонов в RU [36].

Записи в RU регулярно обновляются. Обновление происходит в различных формах: замена исходной образцовой последовательности на консенсус, уточнение или расширение последовательности, добавление последовательностей белков, удаление чужеродных (вставленных или прилегающих) последовательностей, переклассификация, переименование записи или ее удаление. Удаленные старые версии записей можно найти либо в приложении к ежемесячному выпуску RU, либо в архивированных выпусках RU.

5.4. Программы для анализа повторяющихся элементов генома

5.4.1. Repeatmasker

Программа Repeatmasker разработана Арианом Смитом и Робертом Хаблей в 1998. Она используется для идентификации, аннотации и маскирования повторяющихся элементов в контиге или геноме и более того является самой используемой из программ для детекции транспозонных элементов. Программа позволяет идентифицировать некоторые классы tandemных повторов, транспозонных элементов и некодирующих РНК.

У программы есть версия для локального скачивания и WEBRepeatMasker, работающий на сервере института системной биологии (ISB | Seattle, WA). WEBRepeatMasker обладает ограничениями на размеры подаваемых файлов: контиг должен быть не больше 100 kbp. Для поиска используется Basic Local Alignment Search Tool (BLAST). На вход программе подается библиотека повторов и геном/контиг для поиска. Консенсусные последовательности повторяющихся элементов генома можно найти в Repbase Update (RU). Repeatmasker выравнивает каждую консенсусную последовательность с базой данных [43]. Выдает программа три файла: map_file, содержащий информацию о каждом найденном повторе, .masked файл с подаваемым контигом или геномом, где все нуклеотиды найденных повторов

заменены на N, и .tbl с описательной статистикой о представленности каждого класса повторов [36].

Одно из основных применений Repeatmasker'а заключается в маскировании повторяющихся последовательностей перед последующим поиском BLASTX или BLASTN. Если повторяющиеся последовательности не замаскированы, то при поиске последовательности, содержащей повторяющийся участок, будет наблюдаться большое количество незначимых находок по этим повторяющимся участкам. Маскировка этих последовательностей решает проблему. Помимо маскировки, Repeatmasker выдает аннотацию найденных повторов, что упрощает в дальнейшем анализ определенных классов повторов [43]. Для поиска повторов низкой сложности Repeatmasker пользуется Tandem repeats Finder [37].

5.4.2. mHi-C

Стандартные протоколы анализа данных ДНК-ДНК интерактома, полученные протоколом Hi-C, не используют прочтения, картированные в несколько локусов. Таким образом теряется значительная часть информации, в частности о контактах различных классов повторов. Для разрешения этой проблемы был придуман алгоритм mHi-C, использующий байесовский вывод вероятностных моделей и способный распределить часть множественно-картированных прочтений по геномным локусам.

Алгоритм состоит из двух основных частей. Первый осуществляет препроцессинг данных: фильтрацию контактов, а также бинирование хромосом - разделение хромосом на бины определенного размера для рассмотрения контактов между геномными бинами вместо отдельных локусов. Второй этап же применяет модель к данным для вычисления вероятности контакта между двумя бинами.

Алгоритм был применен к данным Hi-C нескольких клеточных линий человека, мыши, а также *P. falciparum*. На этих данных было показано, что учитывание множественно-картированных прочтений вносит большой вклад в построение карт геномных контактов, разрешая в них пробелы, а также увеличивает глубину покрытия секвенирования. Было установлено, что карты контактов, созданные с помощью mHi-C, довольно точно воспроизводят контакты, обнаруживаемые в различных репликах. Используя mHi-C, также удалось выявить новые значимые контакты и взаимодействия между промоторами и энхансерами, а также более точно определить границы топологически ассоциированных доменов (ТАД) [44].

5.4.3. TEtranscripts

Большинство программ для анализа данных секвенирования РНК не учитывают МКП. Это приводит к неверным оценкам экспрессии генов с повторяющихся участков генома. TEtranscripts способен при оценке уровня экспрессии учитывать МКП. Общая схема работы представлена на рисунке 2.

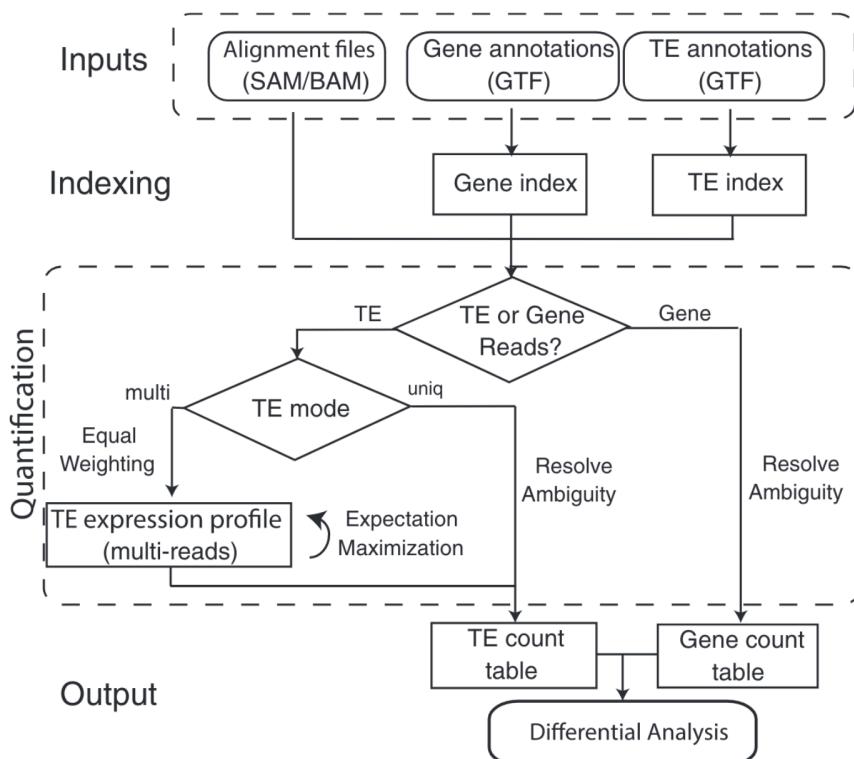


Рис. 2. Схема работы TEtranscripts. TEtranscripts может работать в двух различных режимах: uniq (учитывает только УКП, попавшие на транспозоны) и multi (учитывает УКП и МКП, попавшие на транспозоны). В режиме multi используется итерационный алгоритм для оптимального распределения МКП [45].

Программе подаются картированные чтения в формате .sam или .bam и две аннотации формате .gtf : генов и транспозонов. Для быстрого поиска всех генов и повторов в данных картированиях TEtranscripts индексирует подаваемые аннотации, создавая хэш-таблицу с ключами - идентификаторами последовательностей, значениями - геномными интервалами. Следующим этапом является распределение чтений по элементам аннотации. С УКП все просто: их относят в элемент на который они попали, с МКП же все сложнее. TEtranscripts, пользуясь номенклатурой транспозонов из базы данных RU, группирует транспозоны по схожести последовательностей между собой. Комбинированной оценкой экспрессии для всех элементов мы получаем более надежные и воспроизводимые результаты, чем при анализе отдельных случаев транскрибуемых транспозонов.

TEtranscripts может учитывать только УКП (uniq mode), а может учитывать и МКП (multi mode). В multi mode МКП назначаются веса, равные $1/(количество\ картировок)$, таким образом множественно картированные прочтения суммарно вносит такой же вклад в оценку экспрессии, как и уникально-картированные прочтения. Это важно для поддержания размера библиотеки для каждого образца, который рассчитывается на основе количества выровненных прочтений. Это в свою очередь важно, так как размер библиотеки используется для нормализации при сравнении между несколькими библиотеками.

Затем используется алгоритм Expectation-Maximization (EM), который чередует вычисление дробного распределения каждого множественно-картированного прочтения среди его транскриптов (Е-шаг) и оценку экспрессии всех транскриптов (М-шаг), пока оцененные экспрессии не сойдутся. Полученные оценки экспрессии транскриптов от МКП объединяются с данными для УКП чтобы получить финальные оценки экспрессии для транскриптов. Затем осуществляется анализ дифференциальной экспрессии аналогично пакету DESeq2 [46].

TEtranscripts сравнили с другими аналогичными программами: HTSeq-count [47], Cufflinks [48] и RepEnrich [49] на синтетических и реальных данных. На синтетических данных было показано что TEtranscripts восстановил 88.84% оценок экспрессией, опередив все остальные программы, а на реальных данных результаты TEtranscripts повторяют данные qPCR для нескольких транскриптов *Drosophila melanogaster*.

5.5. ДНК-РНК интерактом

Существует множество некодирующих РНК, берущих на себя регуляторные функции. Было показано, что хроматин-ассоциированные транскрипты участвуют в регуляции экспрессии генов, поддержании структуры хроматина, клеточном делении и в других процессах [50–52]. Известными примерами ДНК-РНК взаимодействий являются: инактивация X-хромосомы у самок млекопитающих, осуществляемая lncRNA Xist; альфа-сателлитные РНК в перицентромерных районах могут стабилизировать связывание гистоновых метилтрансфераз в гетерохроматине; РНК TERRA поддерживает целостность теломер; snRNA участвуют в сборке сплайкосомы и в комплексе с белками сближают пространственно участки РНК, обеспечивая тем самым сплайсинг.

Взаимодействия ДНК и РНК бывают разной природой: опосредованные белком, с образованием РНК-ДНК триплексов, при которых между ДНК и РНК образуются хугстиновские связи, а также трехцепочные структуры, состоящие из ДНК:РНК гибрида, стабилизированного водородными связями и смещенной цепи ДНК (R-петли) [53].

С развитием высокопроизводительного секвенирования появляется все больше методов для изучения ДНК-РНК интерактома. На данный момент существует множество протоколов получения данных о контактах РНК с определенными геномными локусами. Эти протоколы можно разделить на “один против всех”: получение информации о контактах определенной РНК с геномными локусами, и “все-против-всех”: получения информации о контактах всех РНК с геномной ДНК. К протоколам всех-против-всех относятся: GRID-seq [4,5], ChAR-seq [54], RADICL-seq [55], iMARGI [56,57], MARGI [58] и Red-C [59]. Рассмотрим протокол получения данных ДНК-РНК интерактома на примере GRID-seq (см. рис. 3).

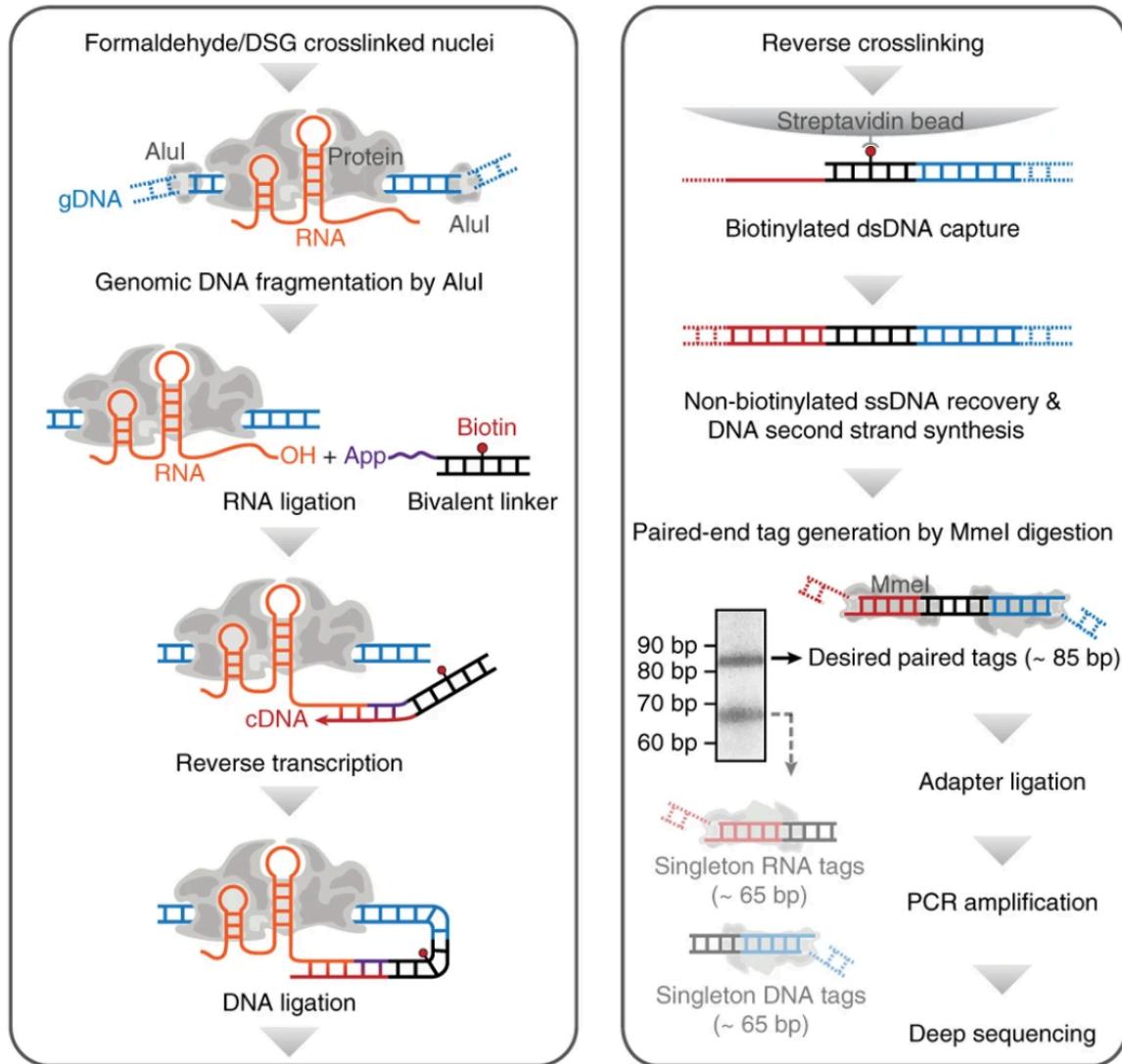


Рис. 3. Схематичное представление протокола GRID-seq. Слева показаны этапы, проводящиеся *in situ*. Справа показаны этапы, проводящиеся *in vitro* [4].

Для получения данных о ДНК-РНК интерактоме в первую очередь проводится фиксация клеток фиксирующим агентом: сукцинимидил глутарат и формальдегид. Выделяется клеточное ядро и затем хроматин подвергается обработке эндонуклеазами рестрикции. В

качестве будущего бриджа используется биотинилированный линкер, содержащий одноцепочечный РНК участок для лигирования к контактирующей РНК и двухцепочечный участок для лигирования к контактирующей ДНК. Для начала проводят лигирование РНК-частей и используя ДНК-праймер синтезируют кДНК обратной транскриптазой. Затем лигируют ДНК-часть линкера с порезанным рестриктазой фрагментом генома, после чего проводится аффинная очистка на стрепатвидиновых шариках. Получаемые двухцепочечные участки обрабатываются рестриктазой MmeI, которая разрезает последовательность на расстоянии 20 нуклеотидов в обе стороны от сайты узнавания. Получаются химерные чтения вида РНК-часть - бридж - ДНК-часть, которые секвенируются и подвергаются дальнейшему биоинформационическому анализу.

Ранее упомянутые протоколы “все-против-всех” во многом схожи с GRID-seq, но есть и ряд отличий. Например, GRID-seq применяется к выделенным ядрам, а MARGI применяется к лизатам клеток. Также в MARGI и iMARGI присутствует этап закольцовывания будущего химерного прочтения и линеаризации его по сайту рестрикции в линкере, в результате чего получаются остатки линкера по краям химерного прочтения, а контактирующие участки внутри. RADICL-seq в отличии от GRID-seq использует для фрагментации хроматина ДНКазу I, что делает получаемые фрагменты близкими по длине, и перед лигированием проба обрабатывается РНКазой H, что способствует понижению количества неспецифических контактов РНК с участком на расстоянии 1 kb от ее гена.

6. МАТЕРИАЛЫ И МЕТОДЫ

6.1. Данные РНК-ДНК интерактома

Для изучения ДНК-РНК интерактомы были использованы, данные полученные протоколом GRID-seq [4] для клеточной линии В-лимфобласт MM.1S. В разделе 4.6. описан протокол GRID-seq. Предобработанные авторами статьи данные можно скачать в NCBI (GEO: GSM2188868) [60].

6.2. Картрирование

ДНК и РНК-части по отдельности картировали на референсный T2T геном человека. Картрирование проводилось с помощью программы HISAT2 (2.2.1) с параметрами --no-softclip -k 20 --no-spliced-alignment для ДНК-частей и --no-softclip -k 20 --no-spliced-alignment --known-splicesite-infile \${ss} для РНК-частей [61]. Картрирование проводилось без возможности сплайсинга для РНК-частей, так как прочтения имеют довольно короткую длину (до 28 bp) из-за чего явление сплайсинга практически не детектируется.

6.3. Обработка результатов картрирования

Для работы с .sam и .bam файлами использовался SAMtools (1.3.1) [62]. Полученная информация о картрировках прочтений была переведена с использованием BEDTools bamtobed в .bed формат [63].

6.4. Аннотация

Затем ДНК и РНК-прочтения были аннотированы аннотацией генов по версии RefSeq (T2T-CHM13v2.0) [3] и аннотацией повторяющихся элементов Repeatmasker [64].

К аннотациям и полученным файлам также применялись BamTools [65] и BEDOPS [66] для приведения файлов в удобные для анализа форматы. Полученный набор файлов использовался для анализа контактов скриптами, написанными самостоятельно на языках Python и R

6.5. Голосование

Уникально картрированные чтения проходят через стадию голосования. Голосование - скрипт Лаборатории Биоинформатики ФББ МГУ для обработки случаев, когда РНК-прочтение попадает на пересечение генов. Скриптом подсчитываются количество прочтений, попадающих на каждый ген и нормализуется на длину соответствующего гена. Все

прочтения, попадающие на пересечения, назначаются тому гену, который содержит большую плотность попадаемых прочтений.

Для УКП, прошедших этап голосования, и их соответствующих ДНК-частей было подсчитано количество контактов между каждой парой биотипов (от РНК-части) и классов повторов (от ДНК-части) и проведен анализ обогащения точным тестом Фишера. Шаблон четырехпольной таблицы представлен в таблице 4. Шаблон заполняется значениями для каждой пары биотипов и классов повторов. К полученным p-value была применена поправка Бонферрони.

	Определенный класс повторов	Остальные классы повторов
Определенный биотипы		
Остальные биотипы		

Таблица 4. Шаблон четырехпольной таблицы. В ячейках располагается количество контактов между соответствующими биотипами и классами повторов.

Весь биоинформатический пайплайн представлен на рисунке 4.

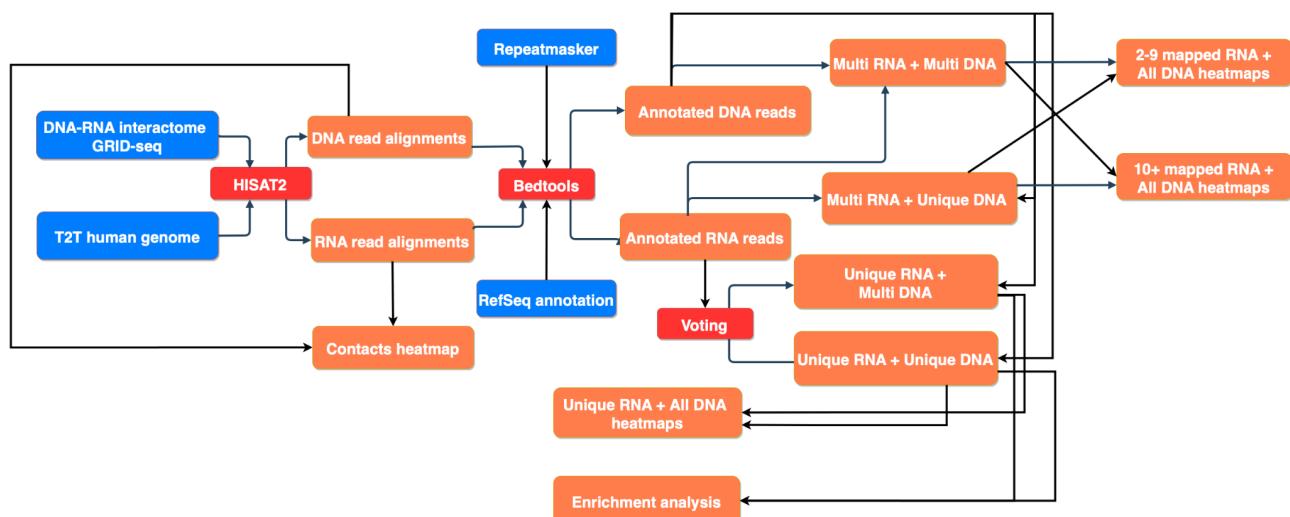


Рис. 4. Биоинформационический пайплайн

Используемые библиотеки для анализа данных в Python: numpy, pandas, pysam. Для визуализации использовались библиотеки matplotlib, seaborn.

Использованные пакеты в R: dplyr, readr, pheatmap, stringr, data.table, pheatmap.

7. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

7.1. Анализ аннотаций

В дальнейшем анализе будут использованы генная аннотация RefSeq и аннотация Repeatmasker. Предварительно эти аннотации были изучены на представленность разных биотипов и классов повторов и занимаемую ими длину в геноме человека (см. рис. 5).

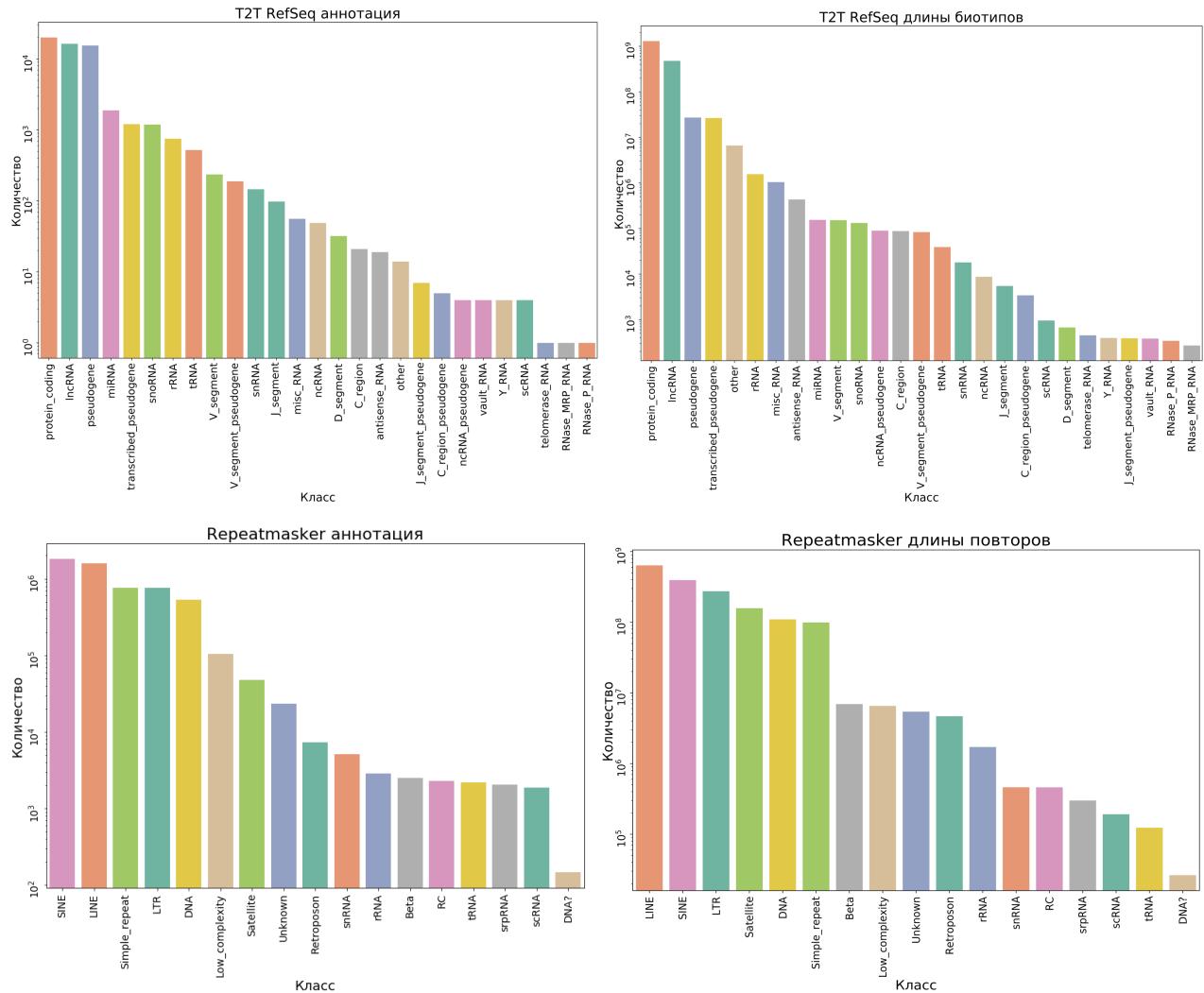


Рис. 5. Представленность биотипов и классов повторов в генной аннотации RefSeq и аннотации повторов Repeatmasker, соответственно. Верхняя пара графиков соответствует анализу генной аннотации, нижня - Repeatmasker. Слева показано количество элементов, относящихся к биотипам и классам повторов. Справа их суммарная длина.

Наиболее представленными повторами являются SINE'ы, LINE'ы и LTR. Они же и занимают большую часть генома среди повторяющихся элементов. Среди биотипов же наиболее представлены mRNA, lncRNA и псевдогены.

7.2. Картируемость

Данные ДНК-РНК-интерактома были картированы на референсный геном T2T с параметрами, позволяющими прочтениям картироваться до 20 раз включительно. Учитывая, что части ДНК/РНК могут не картироваться, мы можем наблюдать девять вариантов картирования РНК-ДНК пар при первичном анализе контактов (см. рис. 6).

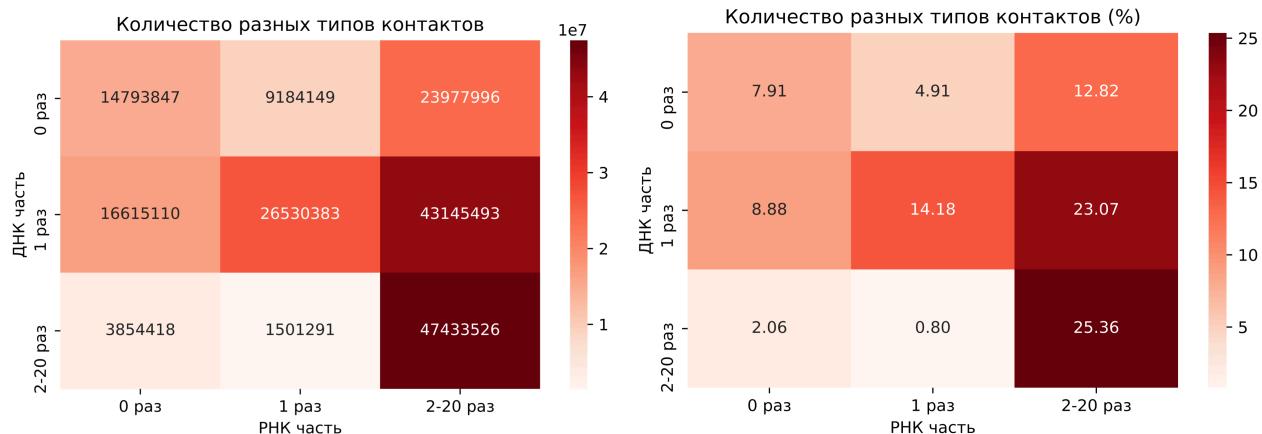


Рис. 6. Теплокарты количества каждого варианта контакта (слева) и процентного содержания определенного типа контакта от всех (справа). Классов три : 0 раз - не картированные прочтения, 1 раз - УКП , 2-20 - МКП.

Среди этих вариантов превалирующим является контакт между множественной картированной ДНК (МКД) и множественно картированной РНК (МКР). Учитывая длину ридов, выдаваемых протоколом GRID-seq (до 28 bp) [4], множественные картировки становятся более вероятными.

Ранее упоминалось, что было введено ограничение на максимальное количество картировок равное 20. Таким образом мы можем наблюдать на контакты в контексте количества картировок ДНК/РНК частей соответствующих контактов (см. рис. 7). Стоит отметить: 20 картировок включают в себя прочтения, которые картировались ровно 20 раз, и чтения, которые картировались бы более 20 раз, но в силу поставленного ограничения принимают значение 20.

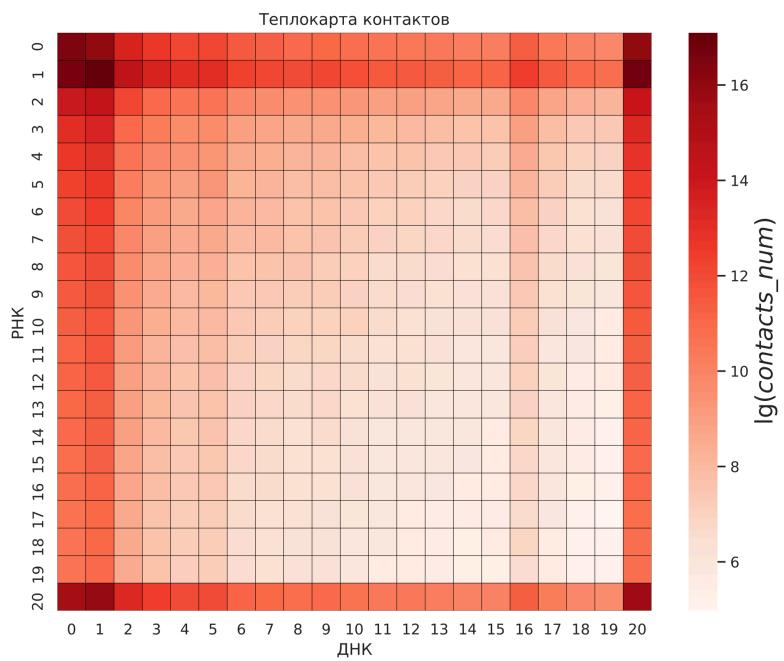


Рис. 7. Теплокарта количества ДНК-РНК контактов для разных картировок. Количество картировок логарифмировано. 0-20 - это количество картировок, соответствующих ДНК/РНК частей.

Больше всего контактов наблюдается там, где части ДНК/РНК картируются уникально или же сразу 20 раз. В остальных случаях с увеличением числа картировок количество контактов уменьшается.

Далее было решено отдельно рассмотреть контакты между частями, картированными 2-5 раз. Для каждого множественно картированного чтения рассматривалось, на какие биотипы оно попадает, с целью идентификации ситуаций, когда одно прочтение картируется на разные биотипы (см. рис. 8). Ожидаемо значительная часть картировок в случае ДНК-частей и большая часть картировок в случае РНК-частей приходится на пары между псевдогенами и другими биотипами. Также основными донорами небольшого количества множественных картировок являются mRNA и lncRNA

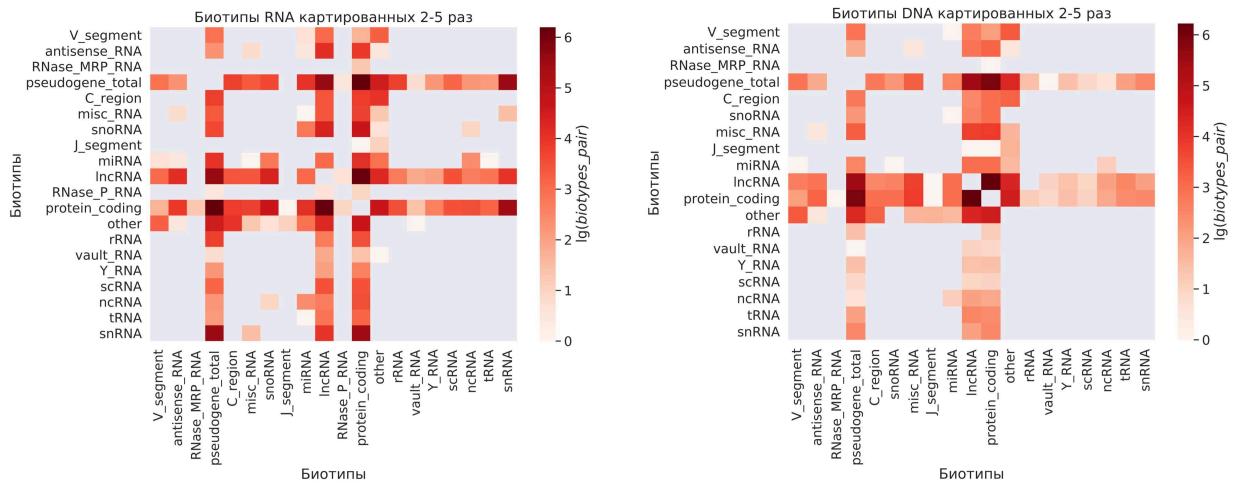


Рис. 8. Теплокарта пар биотипов, на которые 2-5 раз картируются РНК прочтения (слева) и ДНК прочтения (справа). Значения логарифмированы.

Контакты можно подразделить на цис-контакты: РНК и ДНК части приходятся на одну хромосому, и транс-контакты: РНК и ДНК части приходятся на разные хромосомы (см. рис. 9). Здесь мы впервые сталкиваемся с проблемой, возникающей в связи с множественной картируемостью. Множественно картированное прочтение может образовать сразу оба типа контактов. В таком случае мы будем говорить не об уникальных парах ридов, как было ранее, а о всех парах картировок. В дальнейшем, если постановка задачи будет подразумевать возможность прочтения относиться к нескольким классам, то мы будем оперировать именно парами картировок.

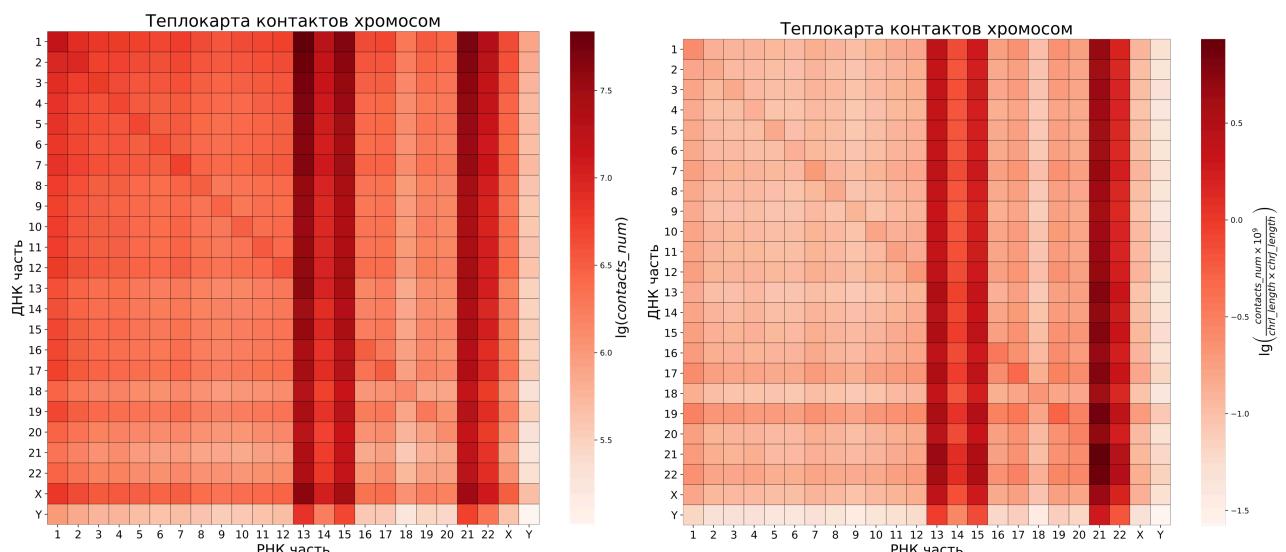


Рис. 9. Теплокарта количества пар картировок ДНК-РНК частей по хромосомам. Количество пар картировок было логарифмировано (слева) и нормализовано на длины взаимодействующих хромосом и также логарифмировано (справа)

Заметное количество контактов приходится на РНК-прочтения из 13, 14, 15, 21, 22 хромосом, что соответствует ядрышковым организаторам. Также заметна диагональная линия, указывающая на то что контакты предпочтительнее происходят между ДНК/РНК частями с одной хромосомы.

Заметно обогащены контактами 19 и 17 хромосомы, которые обладают самой большой плотностью генов [67,68]. Обеднена же контактами 18 хромосома с самой низкой плотностью генов [69].

Также было рассмотрено сколько генов в аннотации содержат только УКР, только МКР и смешанные прочтения (см. рис. 10).

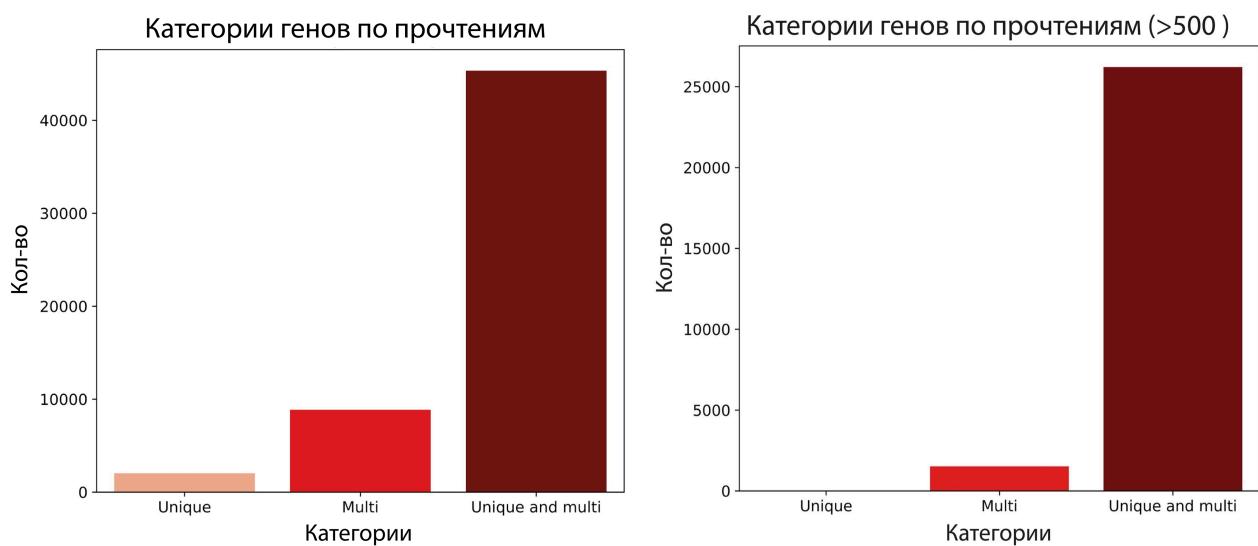


Рис. 10. Количество генов на которые картировались только УКР, только МКР и смешанные прочтения. Слева показаны все гены, а справа гены, содержащие более 500 контактов с ДНК локусами.

На большую часть генов попадают как и УКР, так и МКР. При этом высоко контактирующие гены (более 500 контактов) на которые попадают только УКР вовсе не наблюдается.

В анализах ДНК-РНК интерактома особый интерес представляют именно высоко контактирующие гены и наблюдаемые результаты указывают на то, что игнорирование МКП приводит кискажению профиля контактируемости некодирующих РНК с хроматином и соответственно биологической интерпретации

7.3. Уникальные РНК-части - любая ДНК-часть

В этом разделе мы работаем с уникально картированными РНК-прочтениями, прошедшиими этап голосования, необходимый для разрешения тех случаев, когда РНК часть попадает на пересечение генов.

Были рассмотрены контакты биотипов УКР и соответствующих классов повторов (см. рис. 11).

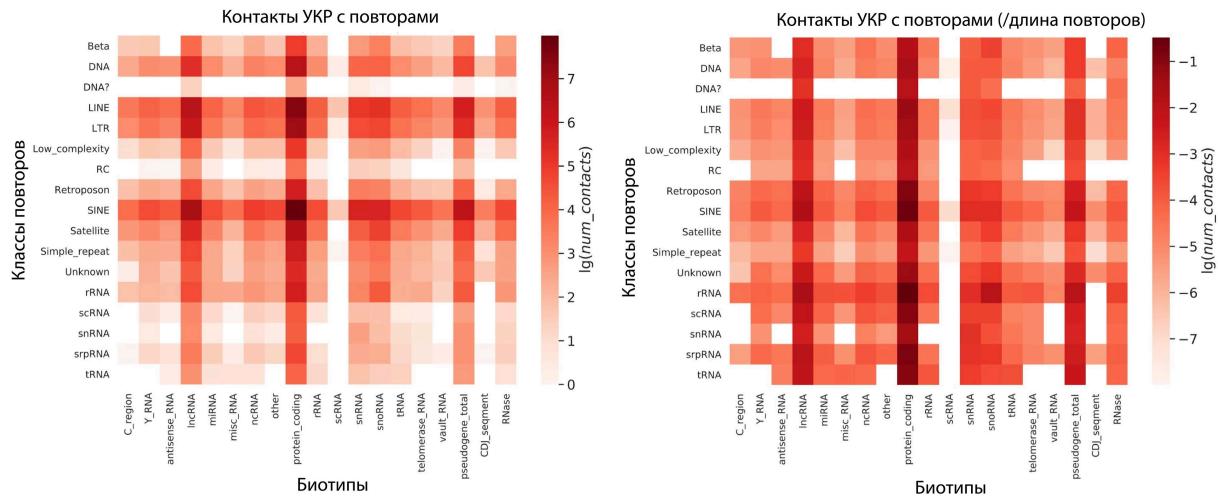


Рис. 11. Теплокарты контактов УКР и любых ДНК (слева) и нормализованных контактов на суммарную длину повторов (справа).

Активно в представленных контактах участвуют mRNA, lncRNA и, по-видимому, транскрибуемая часть псевдогенов. Без нормализации на длину наблюдается много контактов от трёх самых распространённых классов повторов: LINE, SINE, LTR. После нормализации на суммарные длины повторов их контакты перестают выделяться на фоне других. ДНК-транспозоны (DNA) и предполагаемые ДНК-транспозоны (DNA?) ведут себя по-разному в этих контактах, да и в целом в дальнейших анализах эта тенденция будет прослеживаться

7.3.1 Уникальная РНК - любая ДНК, картированная на один класс повторов

На специфичность контактов между биотипами РНК и классами повторов могут указывать те случаи, когда ДНК-часть во всех своих картировках попадает на определенный класс повторов (см. рис. 12). Также при этих условиях можно рассмотреть уникальные пары идентификаторов вместо всех пар картировок.

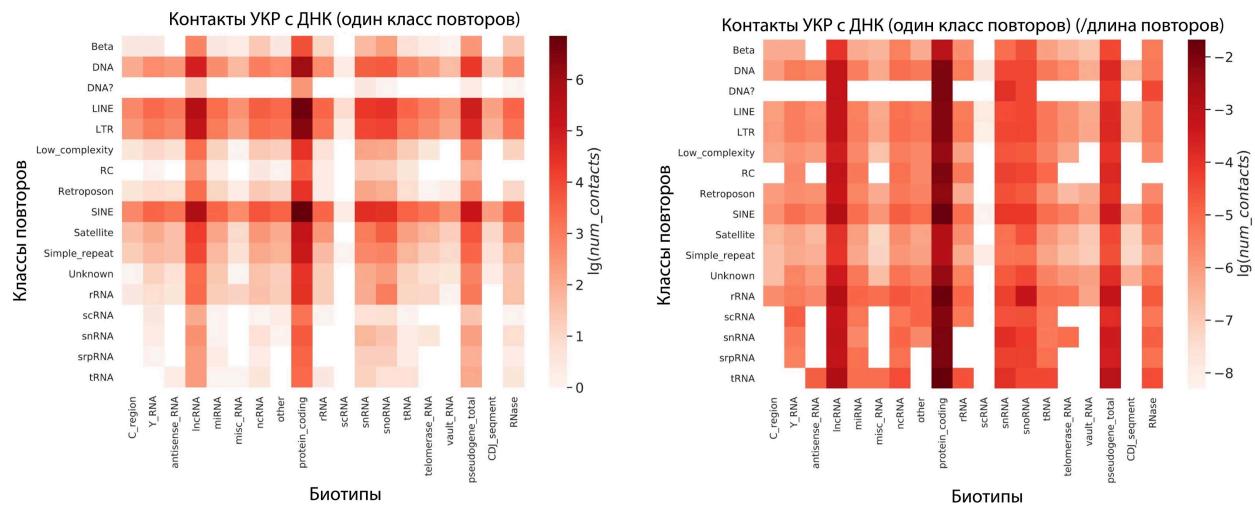


Рис. 12. Теплокарты контактов уникально картированных РНК и любых ДНК, попавших во всех своих картировках на один класс повторов (слева) и нормализованных контактов на суммарную длину повторов (справа).

По сравнению с рассмотренными ранее всеми контактами (в дальнейшем больше/меньше при сравнении теплокарт будет подразумевать увеличенную/пониженнную представленность контакта между определенным классом повторов и биотипом транскриптов относительно остальных контактов в этих теплокартах), здесь наблюдается меньше шума: пропали контакты между повторами srpRNA с биотипами C region, antisense RNA, miscRNA, telomerase RNA, vault RNA. Аналогично scRNA и snRNA потеряли наблюдаемые ранее контакты с упомянутыми биотипами транскриптов.

7.3.2 Уникальная РНК - множественная ДНК, картированная на один класс повторов
Самой обедненной комбинацией контактов являются те, где РНК-часть уникально картирована, а ДНК-часть множественно. В разделе выше рассматриваемая специфичность также зависит от уникально картированных ДНК-частей, так как они, очевидно, во всех своих картировках попадают на один класс повторов. Поэтому теперь будут рассматриваться только множественно картированные ДНК-прочтения.

Для начала рассмотрим зависимость количества контактов от расстояния между ДНК/РНК-частями и "распластанность" этих контактов по длине всей хромосомы на примере 3-й хромосомы (см. рис. 13). Отметим, что одной РНК-части могут соответствовать разные расстояния контактов в связи с множественностью соответствующей ей ДНК-части. Все эти расстояния учитываются.

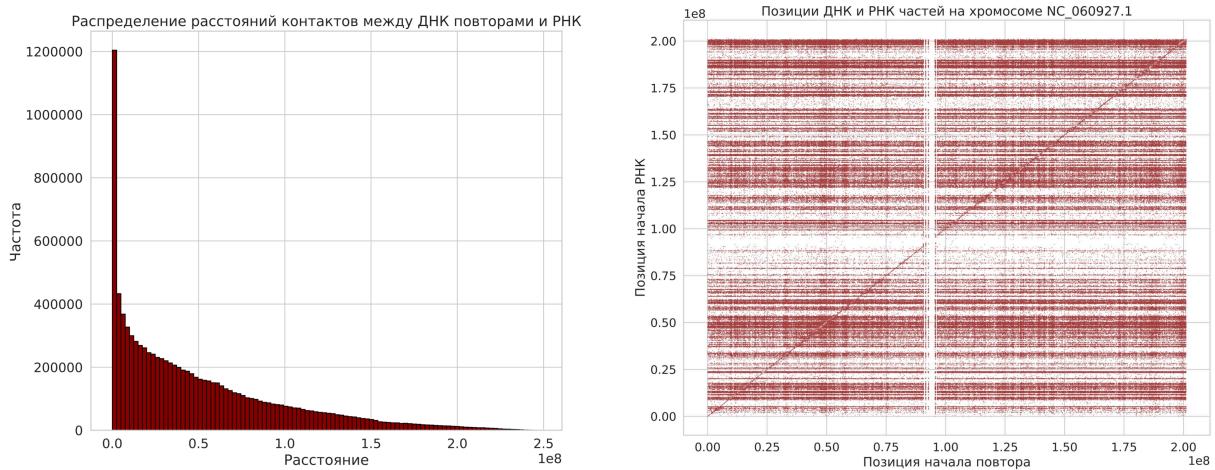


Рис. 13. Зависимость количества контактов от расстояния между ДНК/ РНК частями, образующими контакт (слева) и карта контактов для 3 хромосомы (справа).

Наблюдается явление скейлинга: уменьшение количества контактов с увеличением расстояния между контактирующими ДНК и РНК-частями в связи с пространственным удалением контактирующих частей. На карте контактов это же явление наблюдается в виде диагональной линии. Также на карте контактов наблюдается обединённый контактами «крест», указывающий на затруднительность детекции контактов в центромерных участках. В основном в таких взаимодействиях участвуют LINE’ы, SINE’ы и LTR’ы, что может быть объяснено тем, что это самые распространенные классы повторов в геноме человека (см. рис. 14).

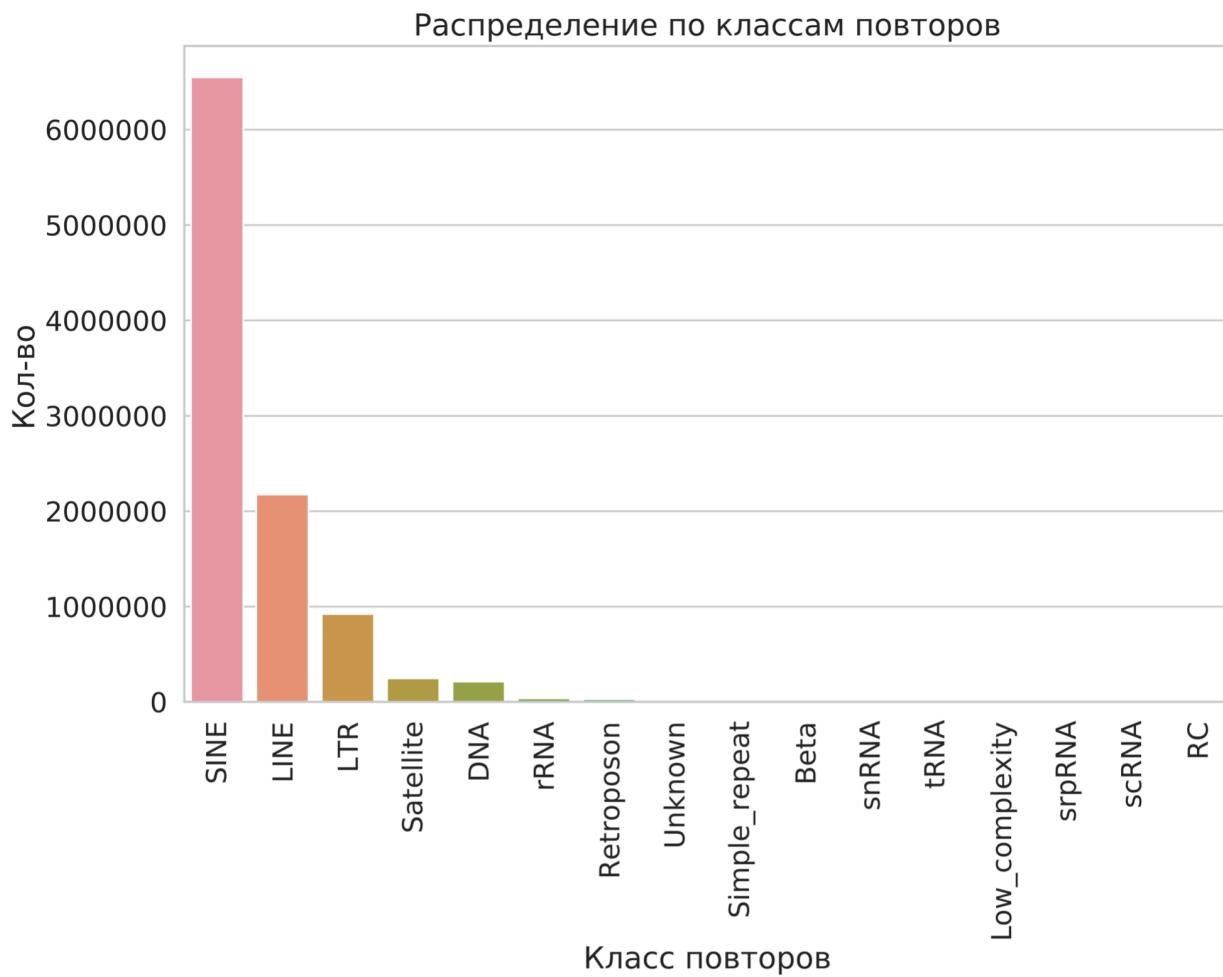


Рис. 14. Количество повторов, участвующих во взаимодействии УКР и МКД, попавшей везде на один класс повторов

Теперь проанализируем контакты в контексте взаимодействий биотипов РНК с классами повторов (см рис. 15). В сравнении с рассматриваемыми ранее контактами, где учитывались УКД, мы наблюдаем массивную потерю контактов ДНК-частей из повторов srpRNA, snRNA, scRNA, RC (хелиитроны), Low complexity repeats. Также мы полностью теряем контакты приходящиеся на класс DNA?.

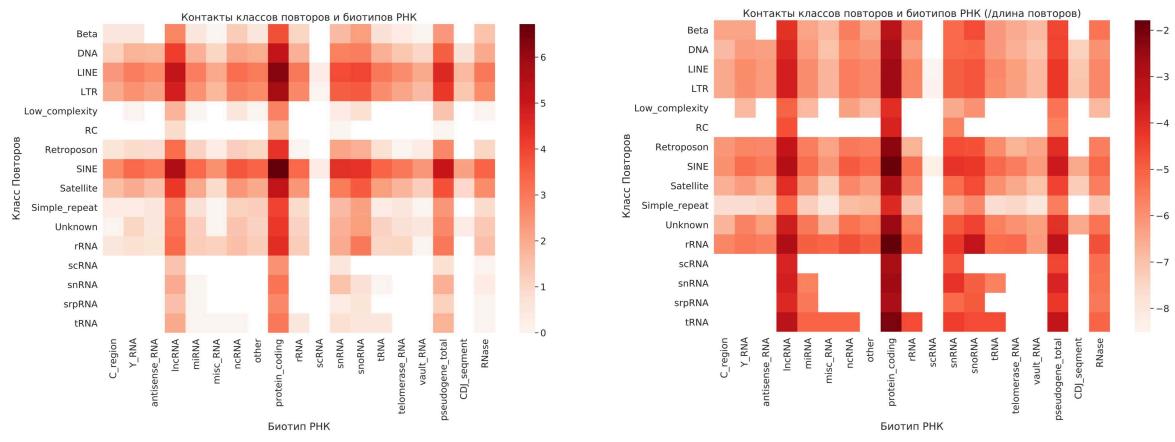


Рис. 15. Теплокарта контактов между УКР (разделение по биотипам) и МКД, попавшие на один класс повторов (разделение по повторам) (слева) и нормализованных на суммарную длину классов повторов (справа). Значения логарифмированы

7.3.3 Анализ обогащения

Проголосованные РНК-части классифицировали по попадающим биотипам, а соответствующие им ДНК-части по классам повторов. При этом в случае ДНК-частей не рассматриваются те прочтения, которые попадают на пересечения в аннотации повторов. Провели анализ обогащения для выявления предпочтений во взаимодействиях биотипов РНК с классами повторов (см. рис. 16)

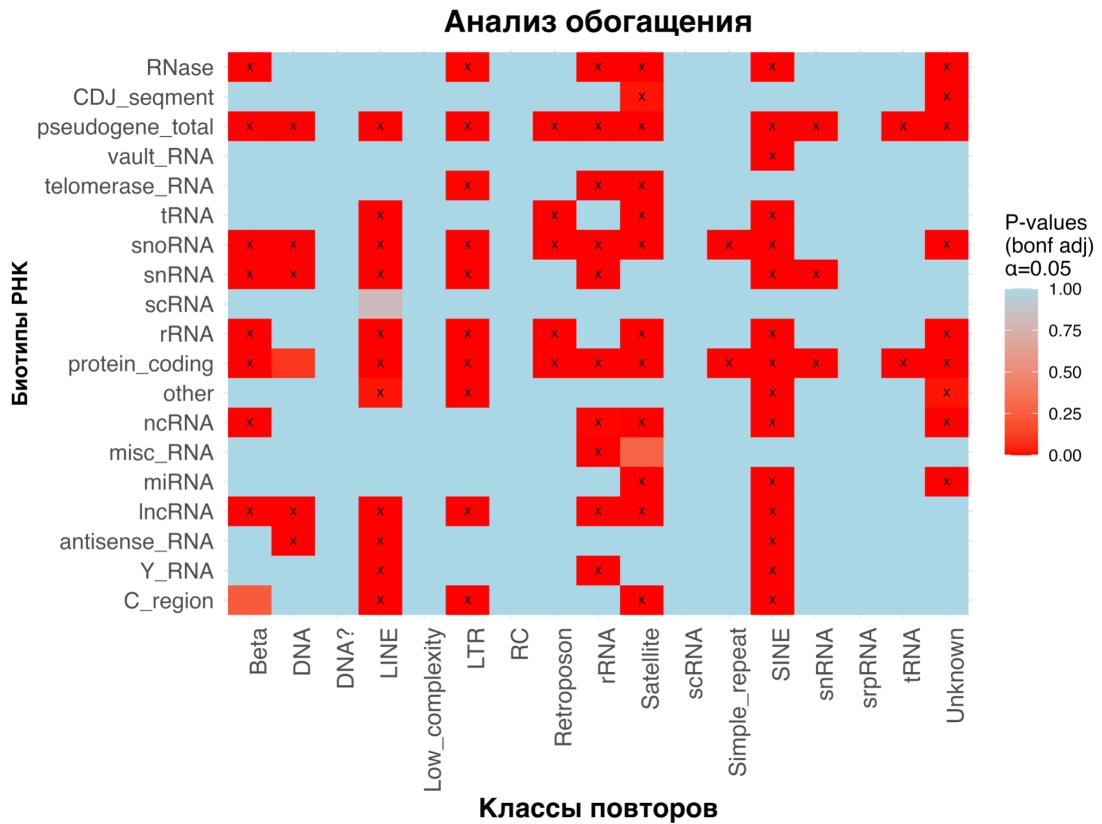


Рис. 16 Анализ обогащения контактов между биотипами РНК и классами повторов на основании УКР, прошедших голосование и соответствующих им ДНК-частей. Анализ проведен тестом Фишера с поправкой Бонферрони на множественное тестирование. Статистически значимые контакты обозначены “x”.

Мы наблюдаем обогащенные контакты от mRNA, lncRNA, транскрибуемых псевдогенов. Предполагается, что взаимодействие антисенс РНК с LINE-элементами может быть обусловлено тем, что некоторые антисенс РНК частично совпадают с последовательностями повторов LINE-1, придавая этим антисенс РНК пространственную близость к данным повторам [70]. Аналогично есть lncRNA SINEUPS, содержащие в себе повторы класса SINE B2 [71]. Есть и случаи, когда Alu повторы (SINE'ы) фланкируют единицы кластеров микроРНК [72]. Это случаи, которые указывают на то, что часть наблюдаемых значимых контактов могут быть объяснены пространственной близостью контактирующих биотипов и повторов.

При этом довольно многие из наблюдаемых значимых контактов на данный момент объяснения в литературе не содержит. Так например: vault RNA значимо взаимодействует только с SINE'ами, объединенные в один класс C,D,G segments - только с сателлитными повторами и т.д. Наблюдаемые тенденции могут лежать в основу дальнейшего анализа контактов между этими определенными биотипами и классами повторов.

7.4. Множественные РНК - любая ДНК

7.4.1 Множественные РНК, картированные на один биотип - любые ДНК, картированные на один класс повторов

Рассмотрим случаи, когда ДНК части попадают на один класс повторов и взаимодействуют с множественно картированной РНК, попавшей везде на один биотип (см. рис. 17). Ожидаемо большую часть таких РНК составляют tRNA и лишь затем mRNA, snoRNA и lncRNA.

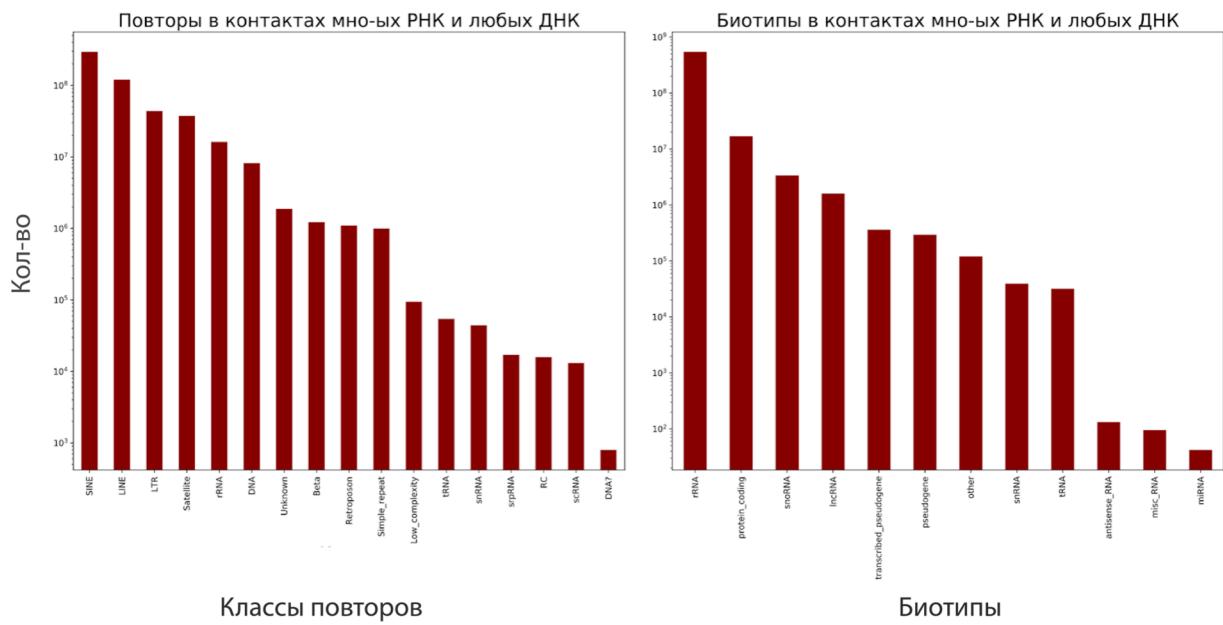


Рис. 17. Представленность классов повторов и биотипов РНК в контактах между МКР, попавшей везде на один биотип и любой ДНК, попавшей на один класс повторов.

При рассмотрении контактов между этими классами повторов и биотипами транскриптов, мы все также наблюдаем превалирующие контакты rRNA со всеми классами повторов. Также в контактах активно участвуют mRNA, lncRNA, псевдогены и snoRNA (см. рис. 18).

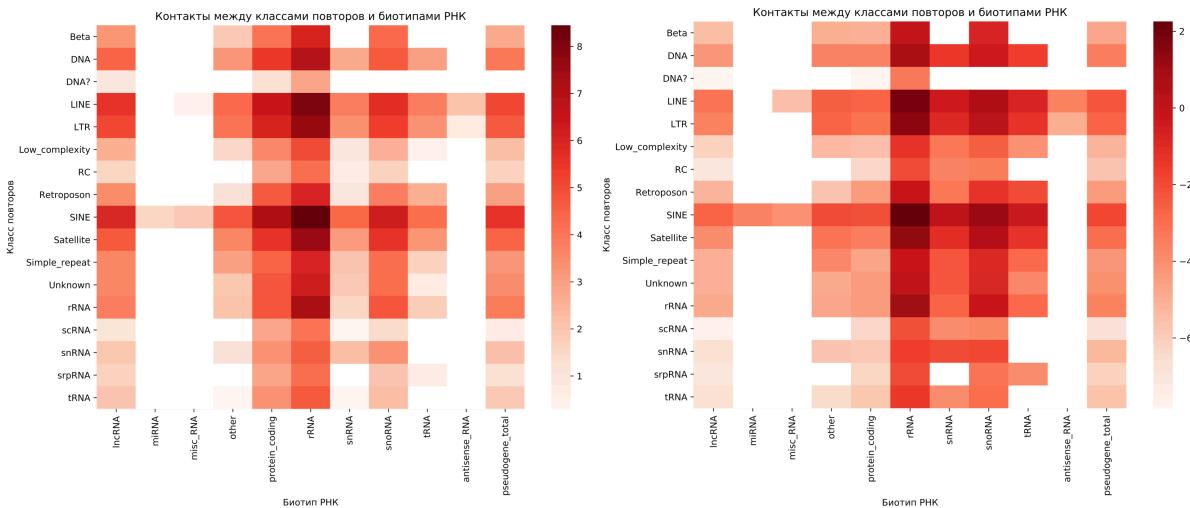


Рис. 18. Представленность классов повторов и биотипов РНК в контактах между МКР, попавшей везде на один биотип и любой ДНК, попавшей на один класс повторов.

7.4.2 2-9 раз картированная РНК - любая ДНК, попавшая на повтор

Если рассматривать все множественные картировки вместе, то тенденции во взаимодействии прочтений, которые картируются много раз (10-20) будут затмевать те

прочтения, что картируются малое число раз (2-9), как мы видели ранее на примере представленности rRNA во взаимодействиях между МКР и любой ДНК. Поэтому было решено отдельно рассмотреть контакты между РНК прочтениями, картированными 2-9 раз и любой ДНК (см. рис. 19)

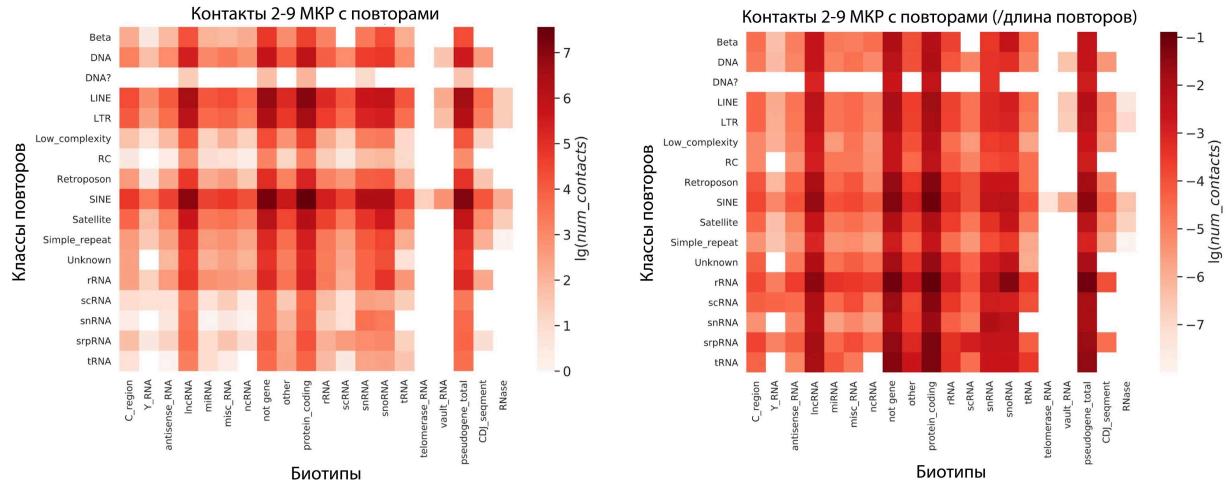


Рис. 19. Теплокарты контактов 2-9 раз картированных РНК, классифицированных по биотипам и любых ДНК, попавших на повторы (слева) и нормализованных контактов на суммарную длину повторов (справа). Значения логарифмированы

Наблюдаемые контакты в целом схожи с ранее проведенным анализом для УКР, однако здесь мы наблюдаем больше контактов от scRNA и меньше контактов от telomerase RNA и vault RNA (больше/меньше относительно контактов между другими классами повторов и биотипами РНК в пределах одной темплокарты). Так как в этом разделе мы работаем со всеми МКР, то у нас могут возникать картировки РНК-частей на локусы не аннотированные как гена (столбец not gene).

Для выявления специфичных контактов были рассмотрены случаи, когда РНК часть во всех картировках попадает на определенный биотип, а ДНК-часть на определенный класс повторов (см. рис. 20).

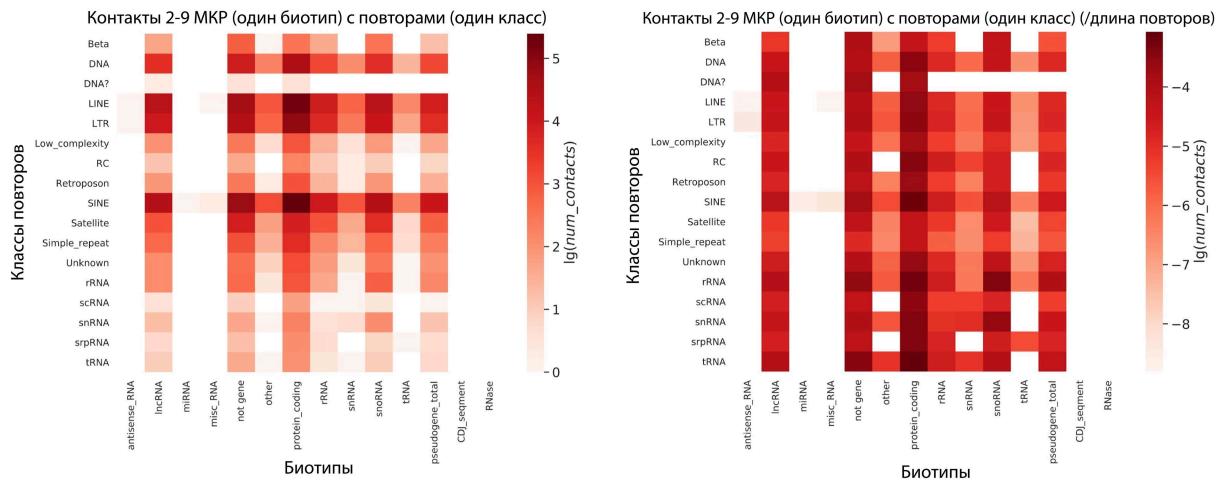


Рис. 20. Теплокарты контактов 2-9 раз картированных РНК, которые во всех своих картировках попали на один и тот же биотип, и любых ДНК, попавших во всех своих картировках на определенный класс повторов (слева) и нормализованных контактов на суммарную длину повторов (справа).

В сравнении с аналогичным графиком для УКР, были потеряны контакты для ряда биотипов: C region, Y RNA, telomerase RNA, vault RNA, ncRNA, scRNA. Интересно, что сохранился класс РНК-частей, не попавших на аннотированные гены.

7.4.3 10+ раз картированная РНК - любая ДНК, попавшая на повторы

Аналогичный анализ был проведен для РНК-частей, картированных 10-20+ раз (см рис 21-22). По сравнению с 1-9 раз картированными РНК, наблюдается значительно увеличивающееся количество контактов, РНК-части которых соответствуют rRNA.

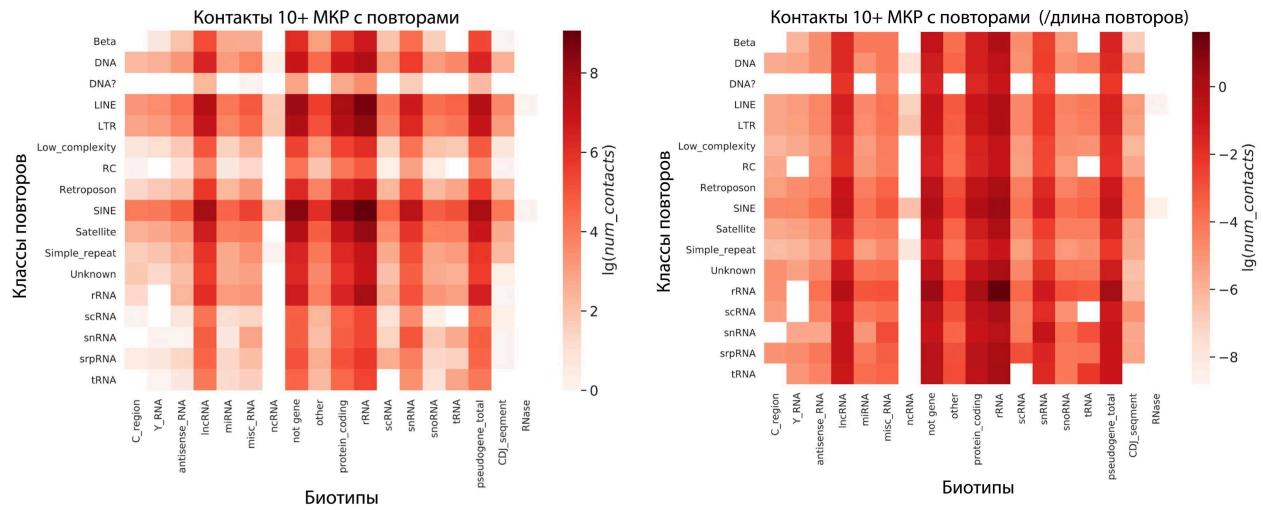


Рис. 21. Теплокарты контактов 10+ раз картированных РНК, классифицированных по биотипам и любых ДНК, попавших на повторы (слева) и нормализованных контактов на суммарную длину повторов (справа). Значения логарифмированы

По сравнению с аналогичным анализом для РНК-частей, картированных 2-9 раз, мы замечаем уменьшение контактов от ncRNA и snoRNA. Ожидаемо, при анализе только специфических взаимодействий, большая часть контактов исходит от rRNA, даже с нетранскрибуируемыми классами повторов.

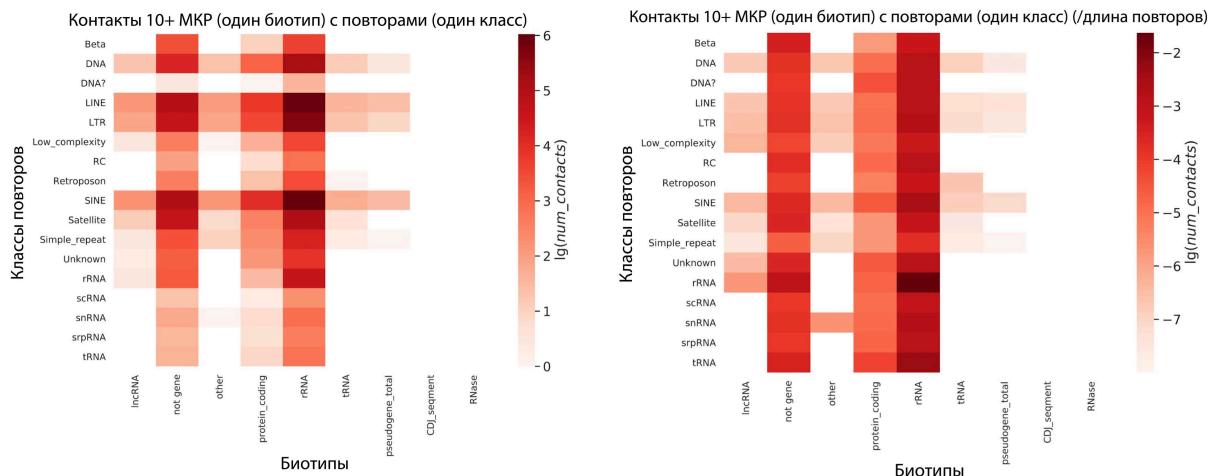


Рис. 22. Теплокарты контактов 10+ раз картированных РНК, которые во всех своих картировках попали на один и тот же биотип, и любых ДНК, попавших во всех своих картировках на определенный класс повторов (слева) и нормализованных контактов на суммарную длину повторов (справа).

8. ВЫВОДЫ

Был проведен анализ данных, полученных в результате эксперимента для установления полногеномного РНК-ДНК интерактома GRID-seq.

- В ходе анализа разработан протокол, позволяющий анализировать контакты между УКР и ДНК-частями, картированными на геном любое количество раз. Предложенный подход позволил дополнить информацию о хроматин-ассоциированных РНК, закодированных в геноме в нескольких копиях.
- Показано, что с геномными локусами, несущими повторяющиеся элементы разных классов, взаимодействуют хроматин-ассоциированные РНК, а также удалось выявить тенденции взаимодействия уникальных и множественно картируемых прочтений, приходящихся на различные гены и типы повторов.

9. СПИСОК ЛИТЕРАТУРЫ

1. Liao X. et al. Repetitive DNA sequence detection and its role in the human genome // Commun. Biol. 2023. Vol. 6, № 1. P. 954.
2. Athanasopoulou K. et al. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics // Life. 2021. Vol. 12, № 1. P. 30.
3. Nurk S. et al. The complete sequence of a human genome // Science. 2022. Vol. 376, № 6588. P. 44–53.
4. Li X. et al. GRID-seq reveals the global RNA–chromatin interactome // Nat. Biotechnol. 2017. Vol. 35, № 10. P. 940–950.
5. Li L. et al. Global profiling of RNA–chromatin interactions reveals co-regulatory gene expression networks in Arabidopsis // Nat. Plants. 2021. Vol. 7, № 10. P. 1364–1378.
6. Biscotti M.A., Olmo E., Heslop-Harrison J.S.P. Repetitive DNA in eukaryotic genomes // Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. 2015. Vol. 23, № 3. P. 415–420.
7. Chebly A. et al. Telomeric Repeat-Containing RNA (TERRA): A Review of the Literature and First Assessment in Cutaneous T-Cell Lymphomas // Genes. 2022. Vol. 13, № 3. P. 539.
8. Zu T., Pattamatta A., Ranum L.P.W. Repeat-Associated Non-ATG Translation in Neurological Diseases // Cold Spring Harb. Perspect. Biol. 2018. Vol. 10, № 12. P. a033019.
9. Hannan A.J. Expanding horizons of tandem repeats in biology and medicine: Why ‘genomic dark matter’ matters // Emerg. Top. Life Sci. / ed. Yuen R. 2023. Vol. 7, № 3. P. 239–247.
10. Schaper E., Gascuel O., Anisimova M. Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes // Mol. Biol. Evol. 2014. Vol. 31, № 5. P. 1132–1148.
11. Saguez C. et al. Functional variability in adhesion and flocculation of yeast megasatellite genes // Genetics / ed. Louise Glass N. 2022. Vol. 221, № 1. P. iyac042.
12. Podvin S. et al. Mutant Huntingtin Protein Interaction Map Implicates Dysregulation of Multiple Cellular Pathways in Neurodegeneration of Huntington’s Disease // J. Huntingt. Dis. 2022. Vol. 11, № 3. P. 243–267.
13. Sun J.H. et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries // Cell. 2018. Vol. 175, № 1. P. 224–238.e15.
14. McNulty S.M., Sullivan B.A. Alpha satellite DNA biology: finding function in the recesses of the genome // Chromosome Res. 2018. Vol. 26, № 3. P. 115–138.
15. Smirnov E. et al. Variability of Human rDNA // Cells. 2021. Vol. 10, № 2. P. 196.
16. Wells J.N., Feschotte C. A Field Guide to Eukaryotic Transposable Elements // Annu. Rev. Genet. 2020. Vol. 54, № 1. P. 539–561.
17. Feschotte C., Zhang X., Wessler S.R. Miniature Inverted-Repeat Transposable Elements and Their Relationship to Established DNA Transposons // Mobile DNA II. 1st ed. / ed. Craig N.L. et al. Wiley, 2007. P. 1145–1158.
18. Cordaux R., Batzer M.A. The impact of retrotransposons on human genome evolution // Nat. Rev. Genet. 2009. Vol. 10, № 10. P. 691–703.
19. Brouha B. et al. Hot L1s account for the bulk of retrotransposition in the human population // Proc. Natl. Acad. Sci. 2003. Vol. 100, № 9. P. 5280–5285.
20. International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome // Nature. 2001. Vol. 409, № 6822. P. 860–921.
21. Evgen’ev M.B., Arkhipova I.R. *Penelope*-like elements – a new class of retroelements: distribution, function and possible evolutionary significance // Cytogenet. Genome Res. 2005. Vol. 110, № 1–4. P. 510–521.
22. Arkhipova I.R. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories // Mob. DNA. 2017. Vol. 8, № 1. P. 19.
23. Yuan Y.-W., Wessler S.R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies // Proc. Natl. Acad. Sci. 2011. Vol. 108, № 19. P. 7884–7889.

24. Bao W. et al. New Superfamilies of Eukaryotic DNA Transposons and Their Internal Divisions // Mol. Biol. Evol. 2009. Vol. 26, № 5. P. 983–993.
25. Feschotte C., Pritham E.J. DNA Transposons and the Evolution of Eukaryotic Genomes // Annu. Rev. Genet. 2007. Vol. 41, № 1. P. 331–368.
26. Grabundzija I., Hickman A.B., Dyda F. Helraiser intermediates provide insight into the mechanism of eukaryotic replicative transposition // Nat. Commun. 2018. Vol. 9, № 1. P. 1278.
27. Li Y., Dooner H.K. Excision of *Helitron* Transposons in Maize // Genetics. 2009. Vol. 182, № 1. P. 399–402.
28. Pritham E.J., Putliwala T., Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses // Gene. 2007. Vol. 390, № 1–2. P. 3–17.
29. Kapitonov V.V., Jurka J. Self-synthesizing DNA transposons in eukaryotes // Proc. Natl. Acad. Sci. 2006. Vol. 103, № 12. P. 4540–4545.
30. Navarro Gonzalez J. et al. The UCSC Genome Browser database: 2021 update // Nucleic Acids Res. 2021. Vol. 49, № D1. P. D1046–D1057.
31. Vollger M.R. et al. Segmental duplications and their variation in a complete human genome // Science. 2022. Vol. 376, № 6588. P. eabj6965.
32. Hoyt S.J. et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements // Science. 2022. Vol. 376, № 6588. P. eabk3112.
33. Gershman A. et al. Epigenetic patterns in a complete human genome // Science. 2022. Vol. 376, № 6588. P. eabj5089.
34. Altemose N. et al. Complete genomic and epigenetic maps of human centromeres // Science. 2022. Vol. 376, № 6588. P. eabl4178.
35. Aganezov S. et al. A complete reference genome improves analysis of human genetic variation // Science. 2022. Vol. 376, № 6588. P. eabl3533.
36. Jurka J. et al. Repbase Update, a database of eukaryotic repetitive elements // Cytogenet. Genome Res. 2005. Vol. 110, № 1–4. P. 462–467.
37. Huda A., Mariño-Ramírez L., Jordan I.K. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation // Mob. DNA. 2010. Vol. 1, № 1. P. 2.
38. Suh A. et al. Multiple Lineages of Ancient CR1 Retroposons Shaped the Early Genome Evolution of Amniotes // Genome Biol. Evol. 2015. Vol. 7, № 1. P. 205–217.
39. Han M.-J. et al. Spy: A New Group of Eukaryotic DNA Transposons without Target Site Duplications // Genome Biol. Evol. 2014. Vol. 6, № 7. P. 1748–1757.
40. Quast C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools // Nucleic Acids Res. 2012. Vol. 41, № D1. P. D590–D596.
41. Chan P.P., Lowe T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence // Nucleic Acids Res. 2009. Vol. 37, № Database. P. D93–D97.
42. Szymanski M. et al. 5SRNAdb: an information resource for 5S ribosomal RNAs // Nucleic Acids Res. 2016. Vol. 44, № D1. P. D180–D183.
43. Tempel S. Using and Understanding RepeatMasker // Mobile Genetic Elements / ed. Bigot Y. Totowa, NJ: Humana Press, 2012. Vol. 859. P. 29–51.
44. Zheng Y., Ay F., Keles S. Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies // eLife. 2019. Vol. 8. P. e38070.
45. Jin Y. et al. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets // Bioinformatics. 2015. Vol. 31, № 22. P. 3593–3599.
46. Anders S., Huber W. Differential expression analysis for sequence count data // Genome Biol. 2010. Vol. 11, № 10. P. R106.
47. Anders S., Pyl P.T., Huber W. HTSeq—a Python framework to work with high-throughput sequencing data // Bioinformatics. 2015. Vol. 31, № 2. P. 166–169.
48. Trapnell C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation // Nat. Biotechnol. 2010. Vol. 28, № 5. P. 511–515.

49. Criscione S.W. et al. Transcriptional landscape of repetitive elements in normal and cancer human cells // *BMC Genomics*. 2014. Vol. 15, № 1. P. 583.
50. Cech T.R., Steitz J.A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones // *Cell*. 2014. Vol. 157, № 1. P. 77–94.
51. Nozawa R.-S., Gilbert N. RNA: Nuclear Glue for Folding the Genome // *Trends Cell Biol.* 2019. Vol. 29, № 3. P. 201–211.
52. Kopp F., Mendell J.T. Functional Classification and Experimental Dissection of Long Noncoding RNAs // *Cell*. 2018. Vol. 172, № 3. P. 393–407.
53. Hegazy Y.A., Fernando C.M., Tran E.J. The balancing act of R-loop biology: The good, the bad, and the ugly // *J. Biol. Chem.* 2020. Vol. 295, № 4. P. 905–913.
54. Jukam D. et al. Chromatin-Associated RNA Sequencing (ChAR-seq) // *Curr. Protoc. Mol. Biol.* 2019. Vol. 126, № 1. P. e87.
55. Bonetti A. et al. RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions // *Nat. Commun.* 2020. Vol. 11, № 1. P. 1018.
56. Wu W. et al. Mapping RNA–chromatin interactions by sequencing with iMARGI // *Nat. Protoc.* 2019. Vol. 14, № 11. P. 3243–3272.
57. Yan Z. et al. Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs // *Proc. Natl. Acad. Sci.* 2019. Vol. 116, № 8. P. 3328–3337.
58. Sridhar B. et al. Systematic Mapping of RNA-Chromatin Interactions In Vivo // *Curr. Biol.* 2017. Vol. 27, № 4. P. 602–609.
59. 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.* 2020. Vol. 48, № 12. P. 6699–6714.
60. BioSample for BioProject (Select 324602) - BioSample - NCBI [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/biosample?Db=biosample&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_biosample&LinkReadableName=BioSample&ordinalpos=1&IdsFromResult=324602 (accessed: 26.04.2024).
61. Kim D. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype // *Nat. Biotechnol.* 2019. Vol. 37, № 8. P. 907–915.
62. Li H. et al. The Sequence Alignment/Map format and SAMtools // *Bioinformatics*. 2009. Vol. 25, № 16. P. 2078–2079.
63. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features // *Bioinformatics*. 2010. Vol. 26, № 6. P. 841–842.
64. GitHub - marbl/CHM13: The complete sequence of a human genome [Electronic resource]. URL: <https://github.com/marbl/CHM13> (accessed: 26.04.2024).
65. Barnett D.W. et al. BamTools: a C++ API and toolkit for analyzing and managing BAM files // *Bioinformatics*. 2011. Vol. 27, № 12. P. 1691–1692.
66. Neph S. et al. BEDOPS: high-performance genomic feature operations // *Bioinformatics*. 2012. Vol. 28, № 14. P. 1919–1920.
67. Grimwood J. et al. The DNA sequence and biology of human chromosome 19 // *Nature*. 2004. Vol. 428, № 6982. P. 529–535.
68. Zody M.C. et al. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage // *Nature*. 2006. Vol. 440, № 7087. P. 1045–1049.
69. Nusbaum C. et al. DNA sequence and analysis of human chromosome 18 // *Nature*. 2005. Vol. 437, № 7058. P. 551–555.
70. Cruickshanks H.A. et al. Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene TFPI-2 in cancer // *Nucleic Acids Res.* 2013. Vol. 41, № 14. P. 6857–6869.
71. Takahashi H. et al. Identification of functional features of synthetic SINEUPs, antisense lncRNAs that specifically enhance protein translation // *PLOS ONE* / ed. Preiss T. 2018. Vol. 13, № 2. P. e0183229.
72. Wickramage I. et al. SINE RNA of the imprinted miRNA clusters mediates constitutive type

III interferon expression and antiviral protection in hemochorionic placentas // Cell Host Microbe. 2023. Vol. 31, № 7. P. 1185-1199.e10.