МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М.В. ЛОМОНОСОВА

Биологический факультет
Кафедра биоинженерии

РАЗРАБОТКА МЕТОДОВ АНАЛИЗА ДАННЫХ СЕКВЕНИРОВАНИЯ НОВОГО ПОКОЛЕНИЯ В ЭКСПЕРИМЕНТАХ ПО ЛЕНТИВИРУСНОЙ ТРАНСДУКЦИИ МЕЗЕНХИМАЛЬНЫХ СТВОЛОВЫХ КЛЕТОК

Выпускная квалификационная
работа бакалавра

Студентки IV курса
КОЖЕВНИКОВОЙ Дарьи Дмитриевны

Научный руководитель
д. ф.-м. наук, доцент.
Шайтан Алексей Константинович

Москва, 2022

Список сокращений

МСК - мезенхимальные стволовые клетки

ДККМ - длительная культура костного мозга

ДНК – дезоксирибонуклеиновая кислота

LeGO-G2 - Lentiviral gene ontology vector with eGFP inserted

ВИЧ – вирус иммунодефицита человека

LTR - long terminal repeat

UMI – unique molecular identifier

Оглавление

1. Введение

Кроветворение во взрослом организме млекопитающих происходит в костном мозге за счет функционирования двух типов стволовых клеток - стволовых кроветворных клеток (СКК) и мезенхимальных стволовых клеток (МСК). Последние формируют стромальное микроокружение костного мозга, которое регулирует СКК и поддерживает кроветворение. Важные свойства МСК, определяющие возможности их прикладного использования в медицине, это их высокий пролиферативный потенциал, достаточный для серийного переноса кроветворного микроокружения in vivo, и способность дифференцироваться по всем стромальным направлениям - в культуре МСК формируют колониеобразующие единицы фибробластов - полипотентные клетки-предшественницы костной, хрящевой и жировой тканей.

МСК представляют особый интерес в контексте разработки медицинских процедур клеточной терапии, таких как: поддерживающая терапия при заболеваниях костного мозга и трансплантации гемопоэтических клеток; подсадка МСК в зоны поврежденных негемопоэтических тканей для формирования функциональных хрящевых, костных и мышечных структур; подавление иммунных конфликтов при трансплантации и аутоиммунных процессах (Владимирская Е.Б., 2007; Hmadcha et al., 2020).

Несмотря на то что множество исследований и медицинских применений МСК существует на данный момент, все еще отсутствует исчерпывающее представление о структуре популяции МСК, состоящей из множества клеточных клонов с разным пролиферативным потенциалом, и её клеточном вкладе в ткань костного мозга. Для ответа на этот вопрос на сегодняшний день разработаны экспериментальные подходы индивидуального маркирования МСК на уровне генома, что делает возможным наблюдение за их пролиферацией. Одним из способов маркирования является маркирование с помощью лентивирусного вектора. Такие исследования нуждаются в разработке подходов анализа геномных данных секвенирования, получаемых на выходе, для извлечения информации о пролиферативном и

дифференцировочном потенциале исследуемой популяции клеток. Данная работа посвящена разработке методов анализа подобных данных на основе протоколов экспериментов и апробации созданного алгоритма на тестовых данных.

2. Цели и задачи

Целью настоящей работы является разработка и тестирование алгоритмов анализа данных секвенирования нового поколения, полученных в экспериментах по оценке количества и размера клонов мезенхимальных стволовых клеток красного костного мозга мышей на основе их маркирования с помощью лентивирусного вектора.

Тестовые данные были предоставлены коллегами, выполнявшими эксперименты по лентивирусному маркированию клеток костного мозга мышей-доноров и созданию очагов эктопического кроветворения под капсулой почки у облученных и необлученных мышей-реципиентов, пробоподготовку и секвенирование образцов в ФГБУ «НМИЦ Гематологии».

Соответственно цели были поставлены следующие задачи:

- Определение структуры данных на основе анализа протоколов экспериментов.
- Предварительный анализ данных - оценка качества и количества исходных данных.
- Разработка алгоритмов фильтрации и кластеризации данных.
- Валидация алгоритма на основе сравнения результатов анализа данных из технических повторов экспериментов.

Формулирование выводов и предложений по улучшению эксперимента.

3. Обзор литературы

1.1 Методы анализа геномных данных

Методы секвенирования нового поколения, основанные на технологии массового параллельного анализа, широко применяются в геномных исследованиях на сегодняшний день благодаря их высокой производительности, точности, относительно низкой стоимости, простоте выполнения процедуры. Разные технологии секвенирования нового поколения способны к осуществлению анализа коротких или длинных прочтений. Первые подходят для анализа вариантов генов или поиска определенных типов последовательностей в геномах, они более дешевые и дают более точные результаты. Вторые имеют меньшую точность, зато обеспечивают достаточно большую длину прочтения, подходящую для сборки генома de novo и полноразмерного секвенирования изоформ генов (Goodwin et al., 2016). В настоящее время для осуществления анализа коротких прочтений наиболее часто применяются платформы SOLiD/Ion Torrent PGM от Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq от Illumina и GS FLX Titanium/GS Junior от Roche (Liu et al., 2012).

Illumina MiSeq (Ravi et al., 2018) – одна из наиболее популярных платформ для секвенирования во всем мире. Это компактный настольный прибор, который выполняет подготовку кластеров в ячейке, амплификацию и секвенирование геномной ДНК, а так же содержит встроенные функции для анализа данных. Платформа осуществляет как одноконцевые, так и парноконцевые прочтения, длиной до 300 оснований.

Секвенирование на платформах Illumina основано на создании матриц данных. Последовательности для секвенирования иммобилизуют на поверхности проточной ячейки с помощью комплементарного взаимодействия адаптеров на концах этих последовательностей с адаптерами на поверхности ячейки. Далее производится твердофазная амплификация, при которой создаются кластеры размером до 1000 копий последовательностей, расположенных близко друг к другу. Затем осуществляется технология

секвенирования путем синтеза - на каждом цикле в ячейку добавляются флуоресцентно меченые дезоксирибонуклеотиды, происходит присоединение одного нуклеотида за счет 3' терминированности модифицированного основания. Флуорофоры освещают красным светом в случае A и C оснований, зеленым светом в случае G, T, с помощь разных фильтров детектируется флуоресценция и определяется тип присоединенного основания. После чего флуоресцентные метки и 3' терминаторы удаляют и цикл начинается заново. В ходе синтеза в получаемых результатах может накапливаться шум, обусловленный особенностями кинетики ферментов, такими как неполное удаление флуорофоров и 3' терминаторов и остановка синтеза некоторых молекул в кластере. Слишком высокая скорость промывки ячейки реагентами может вызвать припуски в присоединении нуклеотидов к последовательностям в кластере или присоединение нуклеотидов без терминатора, по этим причинам синтез последовательностей в кластере происходит не синхронно. Кроме того накопление ошибок связано с точностью полимеразы, используемой для синтеза. Перечисленные факторы обусловливают ограниченность длины прочтения. Известны также факторы, связанные с буквенной последовательностью анализируемой молекулы. В частности присутствие в ней GCC последовательностей и инвертированных повторов повышает накопление ошибок при чтении (Nakamura et al., 2011). Качество данных также зависит от способа подготовки библиотеки и подбора праймеров (Schirmer et al., 2015)

Для анализа данных секвенирования используются разные методы в зависимости от типа данных (DNA-seq, RNA-seq, Chip-seq, HiC-seq и другие) и длины прочтений. Для данных ДНК секвенирования с короткими прочтениями алгоритм анализа включает предварительную оценку качества прочтений, выравнивание прочтений на референсный геном, извлечение интересующих параметров последовательностей. Для осуществления этих задач используются стандартные программные инструменты, такие как **FastQC** для оценки качества первичных данных, программы **Bowtie2** и **BWA** для выравнивания прочтений на геном.

FastQC – это программа, которая применяется для предварительного анализа качества данных секвенирования (Andrews, S. 2010). Она предоставляет ряд анализов, представляемых в графическом виде, и используется для определения наличия серьезных проблем с качеством прочтений, определяющих их непригодность к анализу. В набор анализов входят графики для описания качества оснований в прочтении, качества прочтения в каждой плитке секвенатора, оценки качества прочтений последовательностей, содержания GC нуклеотидов в прочтениях, распределения прочтений по длине, содержания дупликаций, содержания стандартных адаптерных последовательностей. Параметр качества прочтения основания Phred score выражается через десятичный логарифм вероятности ошибочного прочтения (P): $Q = -10 * log10P$.

Bowtie2 и **BWA** – это программы для выравнивания прочтений на референсный геном. Алгоритм запуска этих программ состоит из двух этапов - индексирования генома – разметки генома на фрагменты для ускорения предстоящего поиска вхождений прочтений в последовательность генома и картирования прочтений на геном. При этом соответствующему алгоритму на вход дается на первом этапе файл с геномом, на втором – файл с построенным геномным индексом и файлы с прочтениями. В программе **BWA** есть три алгоритма для осуществления выравнивания - **BWA-backtrack**, **BWA-SW** и **BWA-MEM**. Первый применяется для прочтений длиной менее 100 оснований, последние два – для прочтений длиной от 70 до 1000 оснований, **BWA-MEM** наиболее новый и обладает высокой точностью.

Новые алгоритмы для анализа данных секвенирования могут быть написаны на языке Python. Существуют готовые библиотеки для анализа данных в биоинформатике, такие как **HTSeq** (Anders et al., 2015) и некоторые модули инструментального набора **Biopython** (Cock, P.J. et al., 2009). **HTSeq** – это библиотека для осуществления высокопроизводительного анализа данных, представленных в виде последовательностей. Ее возможности позволяют производить подсчет прочтений, поиск заданных последовательностей или

свойств в данных, создавать массивы данных. Набор инструментов анализа **Biopython** также содержит в себе модули для анализа последовательностей, в том числе проведения выравниваний последовательностей с различными параметрами.

1.2 Метод формирования очага эктопического кроветворения под капсулой почки мыши

Одним из методов исследования устройства стромальной ткани на клональном уровне является формирование очага эктопического кроветворения под капсулой почки мыши, подробно описанном в источнике (Чертков И.Л., Гуревич О.А., 1984).

В этом методе сначала производится извлечение костного мозга из кости животного, например из бедренной кости мыши, затем выращивание длительной культуры костного мозга (ДККМ) в течение нескольких недель. При посадке ДККМ на среду выращивания стромальные клетки костного мозга выползают из костномозгового цилиндра и распластываются по поверхности культурального флакона, образуя подслой прилипающих клеток. Суспензионная фракция ДККМ не содержит МСК, в ней присутствуют кроветворные клетки. Далее производится подсадка подслоя МСК под почечную капсулу животного. При этом сначала в месте подсадки наблюдаются многочисленные некрозы клеток. На 5-ые сутки образуется монолитный слой пролиферирующих ретикулярных клеток. Часть ретикулярных клеток начинают дифференцироваться в остеобласты, формируя костную ткань. На 9-ые сутки внутрь ткани прорастают кровеносные сосуды, через которые в формирующийся очаг поселяются кроветворные клетки мыши-реципиента. Спустя месяц формируется очаг кроветворения, обособляются внутренний кроветворный слой и костная оболочка.

В образовавшемся очаге стромальные клетки происходят от клеток донора, кроветворные клетки принадлежат реципиенту. В экспериментах с серийными ретрансплантациями эктопических очагов было выяснено, что в месте пересадки очага будет формироваться новая кроветворная территория до

9 пересадки. Число пересадок ограничено пролиферативным потенциалом МСК. Это значение было выяснено в экспериментах по подсадке очага кроветворения облученным мышам, в которых собственное кроветворение было остановлено вследствие гибели пролиферирующих кроветворных клеток и мезенхимальных стволовых клеток. Облученные мыши выживали в случае формирования очага кроветворения и умирали в обратном случае.

Для обеспечения возможности исследования структуры популяции МСК, формирующей эктопический очаг, необходимо дополнить описанный метод этапом предварительного маркирования клеток культуры ДККМ. Эффективное маркирование клонов МСК может быть осуществлено с помощью лентивирусного вектора и библиотеки баркодов, применяемой в пробоподготовке данных секвенирования (Bigildeev et al., 2017). Поэтому в экспериментах, с которыми связана данная работа, через две недели после инициации ДККМ, клетки подслоя были генетически маркированы с помощью трансдукции самоинактивирующимся вектором LeGo на основе ВИЧ-1. На последнем этапе этих экспериментов, схема которых представлена на Рис. 1, из внутренней массы и косточек сформировавшихся очагов кроветворения, получали данные секвенирования, содержащие лентивирусные маркеры.

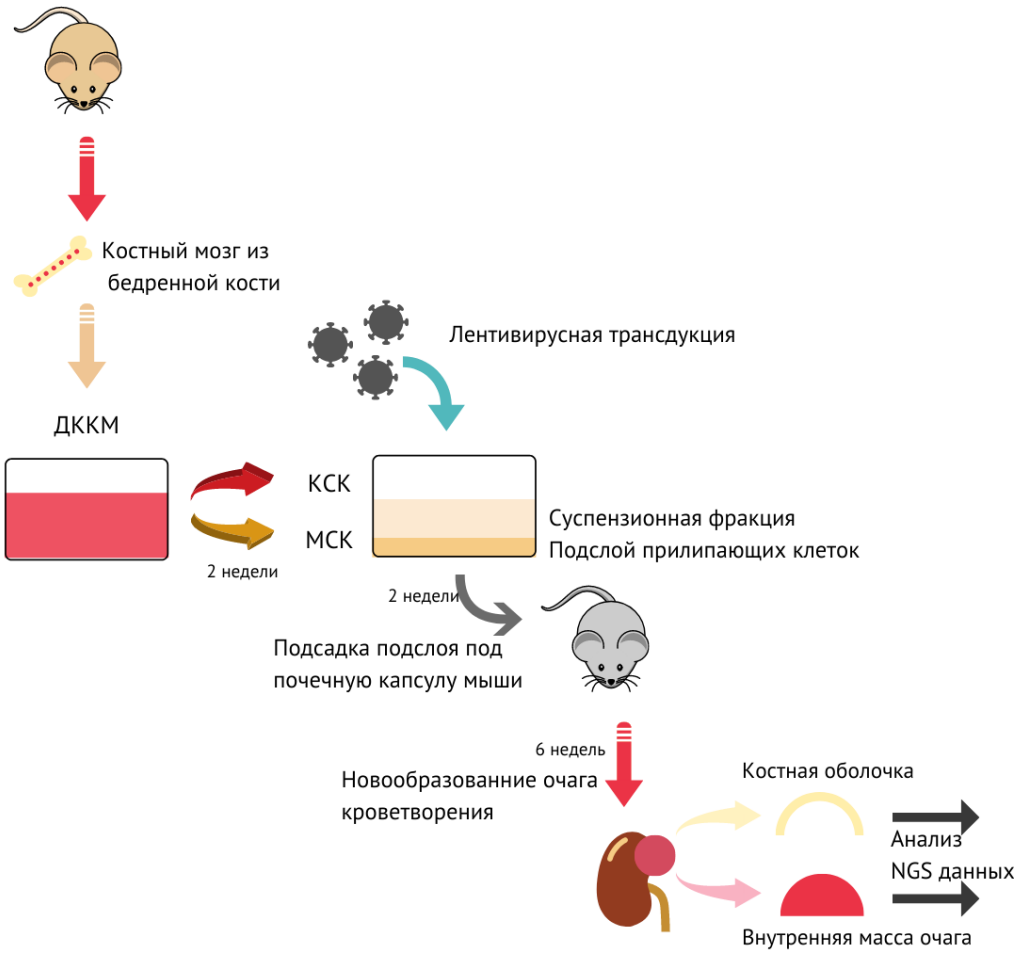


Рис. 1. Схема эксперимента по созданию эктопического очага кроветворения под капсулой почки мыши. Подслой прилипающих клеток содержит МСК, способные дать начало новому очагу кроветворения, поэтому он используется в качестве материала для подсадки мыши-донору. Клетки подслоя предварительно маркируются лентивирусным вектором. В некоторых очагах формируются косточки, из клеток внутренней массы и косточек очагов добывают генетический материал и секвенируют его. (По данным коллег из НМИЦ Гематологии.)

1.3 **Маркирование клеток лентивирусным вектором LeGo**

В экспериментах по созданию эктопических очагов кроветворения (Рис. 1) коллегами из ФГБУ «НМИЦ Гематологии» проводилась лентивирусная трансдукция культуры клеток МСК. Целью трансдукции было индивидуальное маркирование стромальных клеток в новообразованном очаге кроветворения. Индивидуальным наследуемым клеточным маркером служит место интеграции

вставки – фрагмента векторной плазмиды, который интегрируется в геном клеток при вирусной трансдукции. Для маркирования клеток в экспериментах, описанных в предыдущей главе, использовался самоинактивирующийся лентивирусный вектор третьего поколения на основе ВИЧ-1 (Dull et al., 1998; Zufferey et al., 1998). Вектор содержит плазмиду LeGO-G2 (Lentiviral gene ontology vector with eGFP inserted) (Weber et al., 2008) в лентивирусной оболочке.

Изначально плазмида LeGO-G2 была создана для экспрессии кДНК, содержащей ген маркерного белка и, таким образом, маркировки клеток. Для экспрессии маркерного гена зеленого флуоресцентного белка (eGFP) в ней присутствует SFFV промотор. В описанном эксперименте (Рис. 1) маркировка клеток производится за счет случайного выбора места интеграции вставки (провируса) в геном клеток, промотор GFP за ненадобностью был вырезан из плазмиды.

В векторной плазмиде LeGO-G2'(-SFFV) присутствуют следующие последовательности: deltaU3 и R-U5, составляющие 5'LTR (Long terminal repeat – длинный концевой повтор – элемент, фланкирующий провирусную вставку), последовательность которого полностью идентична последовательности 3'SIN LTR; Psi (psi packaging signal) – сайт узнавания, необходимый для упаковки вирусной геномной РНК в нуклеокапсид; RRE (Rev responsible element) - сайт связывания белка Rev (регулятор экспрессии вирусных частиц); cPPT (central polylurine tract) - сайт узнавания ферментов, осуществляющих синтез ДНК провируса, повышает эффективность трансдукции и экспрессию трансгена; WPRE (Woodchuck hepatitus virus post-transcriprional regulatory element) - последовательность, стимулирующая экспрессию трансгена за счет повышения эффективности ядерного экспорта; eGFP - маркер, присутствующий в первоначальной плазмиде. Схема плазмиды показана на рисунке 2. Вставка, интегрирующаяся в геном, показана на Рис. 3.

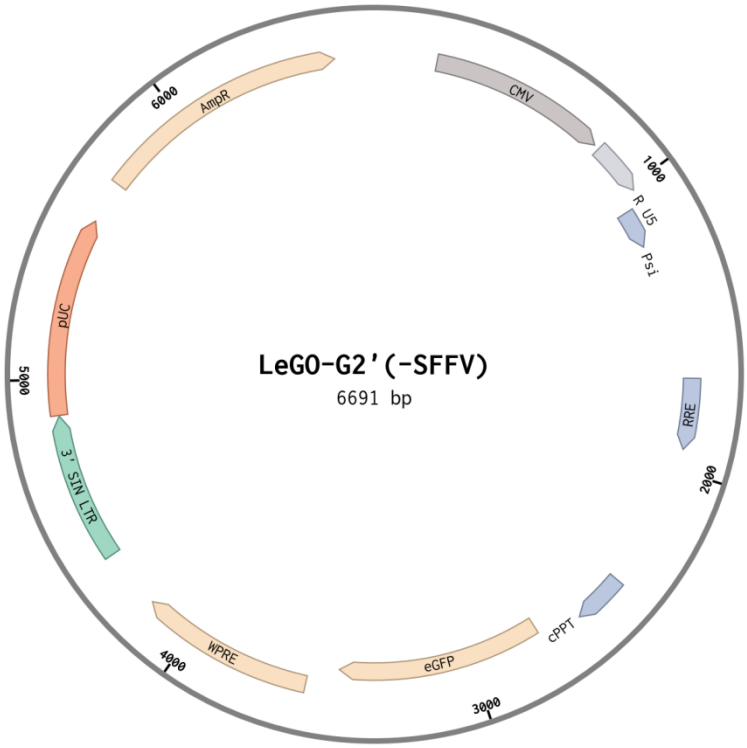


Рис. 2. Векторная плазмида LeGO-G2 с удаленным промотором SFFV, содержится внутри лентивирусного вектора, использованного для маркировки клеток. Состоит из вставки – участка, который интегрируется в геном исследуемых клеток (показан на Рис. 3), и вспомогательных элементов - pUC и AmpR для клонирования вставки в плазмиду и промотора CMV для активации экспрессии генов с плазмиды.

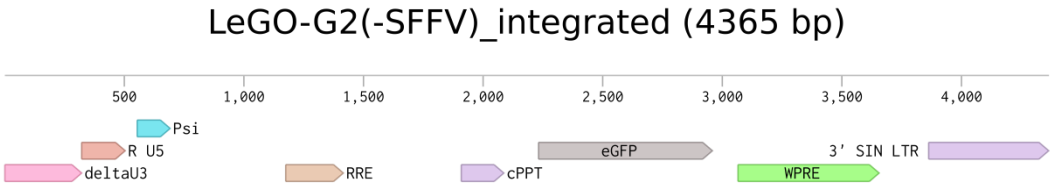


Рис. 3. Вставка - фрагмент векторной плазмиды, интегрирующийся в геном. Содержит фрагменты LTR на обоих концах, регуляторные последовательности Psi, RRE, cPPT, WPRE и ген белка eGFP.

Последовательности LTR (Long terminal repeats) способствуют интеграции в геном хозяина, состоят из регионов U3, R и U5 и находятся с обеих сторон провируса у ретровирусов. Величина вставки между 3' и 5' LTR

может составлять до 8.5 килобаз, но вставки больше 3 килобаз упаковываются менее эффективно.

Компоненты LTR:

- U3 (Unique 3' region) находится на 3' конце вирусной геномной РНК (в провирусе располагается с обоих концов). Содержит последовательности, необходимые для активации транскрипции вирусной геномной РНК. Многие лентивирусные вектора для переноса вставки в геном являются самоинактивирующимися (SIN) векторами. Инактивация осуществляется удалением из векторной плазмиды энхансерной/промоторной последовательности в U3-области 3'-LTR, что обозначается как ΔU3. При обратной транскрипции «дефектная» U3-область копируется на 5'-конец (согласно схеме обратной транскрипции (Cann, 2012)), в результате на обоих концах провируса в области LTR находится дефектная последовательность энхансера/промотора ΔU3. Эта делеция делает невозможной транскрипцию полноразмерного вируса после его включения в геном клетки хозяина.
- R (Repeat region) - находится на 5' и на 3' концах LTR у ретровирусных и лентивирусных векторов. С подпоследовательностью TAR (Trans-activating response element) в регионе R связывется Tat (Trans-activator - активатор транскрипции с LTR промотора в векторных системах 2-го поколения.)
- U5 (Unique 5' region) - находится на 5' конце вирусной геномной РНК (в провирусе располагается с обоих концов)

Интеграция вставки происходит согласно механизму интеграции ВИЧ-1: вирусная интеграза в составе преинтеграционного комплекса катализирует две последовательные Mg-зависимые полинуклеотидил-трансферазные реакции: 3' процессинг кДНК и перенос цепи ДНК в сайт интеграции: вирусная интеграза гидролизует динуклеотид на 3' конце вирусной ДНК, комплементарный инвариантному CA динуклеотиду, образовавшиеся CAOH-3' концы

используются вирусной интеразой для проведения S_N2 реакции в которой сайт интеграции в ДНК хозяина разрезается с образованием липких 5' концов, к которым присоединяются CAOH-3' концы вирусной ДНК (Lesbats et al., 2016; Engelman, 2019). Процессинг может осуществляться димером IN, встраивание - тетрамером, возможны и другие олигомеры. После этого образовавшиеся одноцепочечные участки достраиваются клеточной системой репарации, образуется интегрированный провирус, фланкированный дупликациями сайта интеграции, для ВИЧ-1 длиной чаще всего 5 пн.

Полногеномные исследования вирусной интеграции показали, что во всех видах хозяев ВИЧ-1 чаще всего интегрируется в активно транскрибируемые гены. Это предпочтение объясняют несколько факторов (Engelman et al., 2018; Poletti et al., 2018; Bedwell et al., 2020; Francis et al., 2020):

- Доступность хроматина в этих регионах
- Взаимодействие IN в составе PIC с внутриклеточными белками хозяина: важную роль имеет взаимодействие с внутриклеточным белком LEDGF/p75, в отсутствие которого эффективность интеграции провируса значительно снижается (Engelman et al., 2008), и взаимодействие с белками ядерных пор Nup358, Nup153 и белком CPSF6 (cleavage and polyadenylation specificity factor 6) (Francis et al., 2020, P. 1).
- Путь проникновения вирусного RTC в ядро: “горячие точки” интеграции расположены вблизи ядерных пор, вероятность интеграции в активно транскрибируемый ген - функция от его радиального расстояния от ядерной мембраны. Было показано, что места интеграции ВИЧ-1 расположены вблизи геномных областей, ассоциированных со спеклами, в частности в локусах, ассоциированных с ядерной ламиной (Bedwell et al., 2020; Francis et al., 2020). Причиной к этой направленности является взаимодействие с белками ядерных поровых комплексов и белком CPSF6. Так как топологическая организация хроматина специфична для каждой клетки, горячие точки интеграции варьируют у разных типов клеток.
- Локальное ядерное окружение.

В статье (Serrao et al., 2014) было показано, что HIV-1 предпочитает последовательности (0)RYXRY(4) (R - пурин, Y - пиримидин) в ДНК хозяина. В нескольких исследованиях было показано предпочтение к слабо консервативным палиндромным последовательностям в центре целевой ДНК, кроме того ретровирусы имеют предпочтение к определенным нуклеотидам в сайте интеграции (Wu et al., 2005). Палиндромность сайтов интеграции может быть объяснена симметрией комплекса - каждая молекула интегразы проявляет одинаковые предпочтения к последовательности, поэтому выбираются симметричные последовательности. Несмотря на наличие предпочтений в выборе места интеграции, разброс возможных мест интеграции достаточен для обеспечения маркировки клеток.

Описанные выше факты из литературы были суммированы в схему интеграции, представленную на Рис. 4.

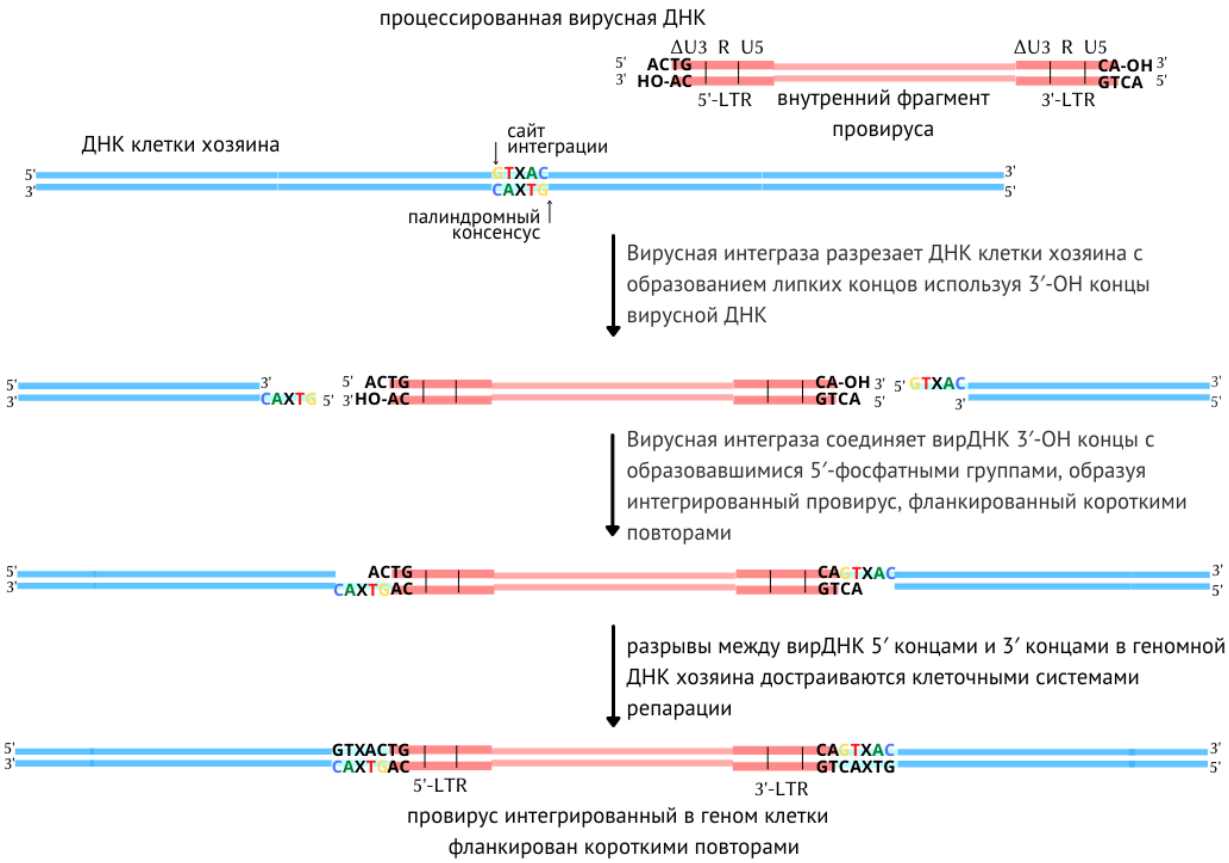


Рис. 4. Схема механизма интеграции провирусного фрагмента ВИЧ-1 в геном клетки хозяина.

1.4 Пробоподготовка и секвенирование образцов

В экспериментах, описанных в главе 1.2, получали образцы ДНК из клеток внутренней массы и косточек очагов эктопического кроветворения. Пробоподготовка образцов для секвенирования была основана на методе, предложенном Sherman E et al. в статье (Sherman et al., 2017). В описанном методе процесс пробоподготовки состоит из следующих этапов: фрагментация ДНК ультразвуком, обработка концов ДНК-фрагментов, лигирование линкеров и два этапа ПЦР. Длина цепей ДНК после фрагментации составляла в среднем 1000 пн, аденилирование 3'-концов производилось с помощью taq-полимеразы, форфорилирование 5'-концов с помощью нуклеотид-киназы, использовался метод вложенных ПЦР реакций (‘nested PCR’). В линкере содержались последовательности: уникальные для биологического образца (одинаковые для технических повторов) (unique linker sequence на Рис. 5); общие для всех образцов (common linker sequence на Рис. 5); уникальные для каждой молекулы ДНК, присутствовавшей в библиотеке на этапе пробоподготовки (Random sequence tag на Рис. 5). Кроме того, использовался блокирующий праймер – олигонуклеотид, комплементарный внутреннему фрагменту провируса и препятствующий наработке нежелательного продукта (Рис. 1).

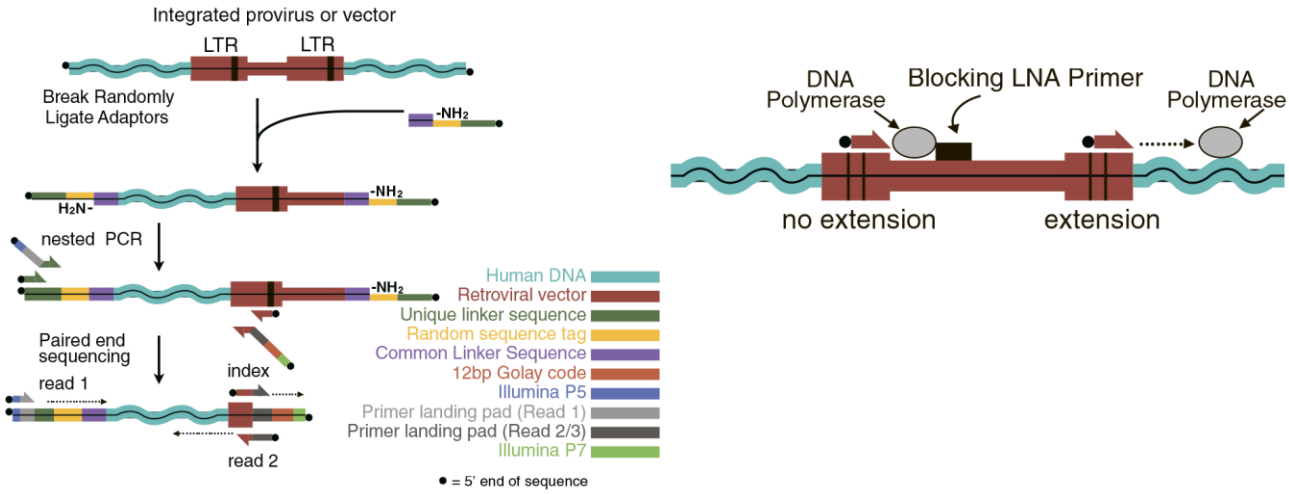


Рис. 5. Схема биохимического метода выделения и секвенирования сайтов интеграции, предложенная в статье (Sherman et al., 2017), и иллюстрация использования блокирующего праймера. Источник: (Sherman et al., 2017).

По известным последовательностям уникальным для образца есть возможность определить образец, к которому принадлежит молекула. При секвенировании на платформе Illumina MiSeq каждая молекула ДНК индексируется, что позволяет определить образец, к которому она принадлежит. MiSeq в автоматическом режиме распределяет прочтения по образцам в соответствии с индексом прочтения. Благодаря двойному индексированию есть возможность уточнить соответствие принадлежности полученных прочтений образцам, в которые они были записаны секвенатором. Благодаря наличию последовательности UMI (unique molecular identifier) можно определить количество разных клеток, содержащих одинаковый сайт интеграции, то есть размер клонов.

2 Материалы и методы

2.1 Тестовые данные

В ходе разработки искомого алгоритма анализа использовались тестовые данные секвенирования геномов клеток эктопических очагов кроветворения, полученные коллегами из ФГБУ «НМИЦ Гематологии» в лаборатории физиологии кроветворения. Ими были проведены эксперименты по созданию очагов эктопического кроветворения под капсулой почки у облученных и необлученных мышей с предварительной обработкой клеток МСК в культуре лентивирусным вектором (Рис. 1). Данные были разделены на 15 образцов, каждый из которых содержал по 4 повтора, количества образцов из разных экспериментальных групп показано в Таблице 1. Образцы были получены из клеток внутренней клеточной массы или косточки очагов облученных или интактных мышей-реципиентов.

Таблица 1. Образцы геномных данных.

Группы образцов	**Кол-во образцов**	**Кол-во тех. повторов**
Очаг в интактной мыши	5	4
Косточка очага в интактной мыши	2	4
Очаг в облученной мыши	4	4
Косточка очага в облученной мыши	3	4
Очаг из незараженной ДККМ (контроль)	1	4
Всего	15	60

Протоколы поведенных экспериментов были предоставлены коллегами и подробно проанализированы в главе 3.1.

2.2 Программное обеспечение

Первичные данные секвенирования представляют собой набор файлов с сырыми данными, полученными с секвенатора Illumina MiSeq, в текстовом формате fastq. Fastq – формат для хранения информации о прочитанных

последовательностях фрагментов ДНК и их побуквенном качестве. Использовалось парноконцевое секвенирование, при котором парные прочтения записываются в отдельные файлы.

Для создания инструмента анализа данных был написан алгоритм на языке Python версии **Python 3.5.6**, код был написан и сохранен в веб-приложении **Jupyter Notebook**. Применялся ряд стандартных программ для анализа данных секвенирования и набор библиотек языка Python для обработки последовательностей.

При извлечении информации о качестве сырых данных использовалась программа **FastQC**.

Выравнивание прочтений, содержащихся в fastq файлах, проводилось с помощью программы **BWA** (Li et al., 2009), использовался алгоритм bwa mem для парных прочтений. Был использован геном мыши сборки GRCm39 (Church et al., 2011). Было проведено сравнение результатов выравнивания сырых и обработанных прочтений, подвергшихся обрезке. Обрезка - удаление адаптерных, линкерных и других дополнительных последовательностей, которые не должны выравниваться на референсный геном, производился с помощью программы **Cutadapt** (Martin, 2011) версии 3.4. Сравнение показало, что выравнивание в **BWA** может быть произведено без тримминга прочтений без потери качества, так как в программе **BWA** есть встроенная процедура отсечения концевых фрагментов прочтений, которые не получилось качественно выровнять (soft clipping).

Для анализа биологических данных были применены инструменты из библиотек **HTSeq** версии 0.12.4 и **Bio** версии 1.79.

Для отображения выравниваний был применен модуль **Pytexshade** (Integrative Biology Group, 2022).

Построение и изображение графов осуществлялось с помощью встроенных методов из библиотек **Pyvis** (Perrone G. et al., 2020) версии 0.1.9 и **Networkx** (A Hagberg et al.,) версии 2.6.2.

3 Результаты и обсуждение

3.1 Анализ протоколов экспериментов и определение структуры данных

Для составления представления о структуре данных, процесс пробоподготовки и секвенирования данных, описанный в главе 1.4, был подробно изучен и сопоставлен с информацией о деталях экспериментов, предоставленных коллегами.

Оригинальная процедура пробоподготовки (описанная в (Sherman et al., 2017)) состоит из этапов фрагментации ДНК, обработки концов ДНК, лигирования линкеров, и двух этапов вложенных ПЦР. Линкер содержит в себе последовательность, специфичную для образца, уникальный баркод – случайную последовательность из 12 пн, обозначаемый как UMI (unique molecular identifier) и 5'-концевую последовательности, общую для всех образцов. Линкер лигируется к каждой молекуле в библиотеке с обеих сторон как показано на Рис. 5. Последовательность UMI важна для анализа, так как маркирует каждую уникальную молекулу ДНК на этапе пробоподготовки перед амплификацией, в дальнейшем анализе данных эти последовательности будут использоваться для определения размера клонов клеток, за счет того, что интересующий фрагмент геномной ДНК, произошедший из каждой клетки помечен своим баркодом. Праймеры для двух этапов вложенных ПЦР попарно комплементарны линкерному фрагменту, общему для образцов, и началу провирусной последовательности LTR. Таким образом при ПЦР будут нарабатываться те молекулы фрагментированной ДНК, в которых присутствует фрагмент LTR. Но поскольку провирус, интегрировавшийся в геном клетки, содержит идентичные последовательности LTR на обоих концах, будут нарабатываться как молекулы, содержащие 3' LTR, так и молекулы, содержащие 5' LTR. С первых будут нарабатываться интересующие нас последовательности, содержащие LTR и фрагмент ДНК генома клетки, в который был интегрирован провирус (то есть сайт интеграции), со второго же будет нарабатываться внутренний фрагмент провируса – последовательность, заключенная в провирусе между двумя концевыми повторами. Второй продукт

не содержит сайта интеграции и не интересен для исследования, его наработка будет загрязнять целевой продукт амплификации. Для подавления синтеза нежелательного продукта используется блокирующий праймер – модифицированный олигонуклеотид, комплементарный внутреннему фрагменту провируса, образующий устойчивый дуплекс с ДНК провируса, и мешающий прохождению полимеразы.

Исходя из описанных выше подробностей, была нарисована более точная и детализированная схема превращений ДНК в ходе пробоподготовки и секвенирования, показанная на Рис. 6.

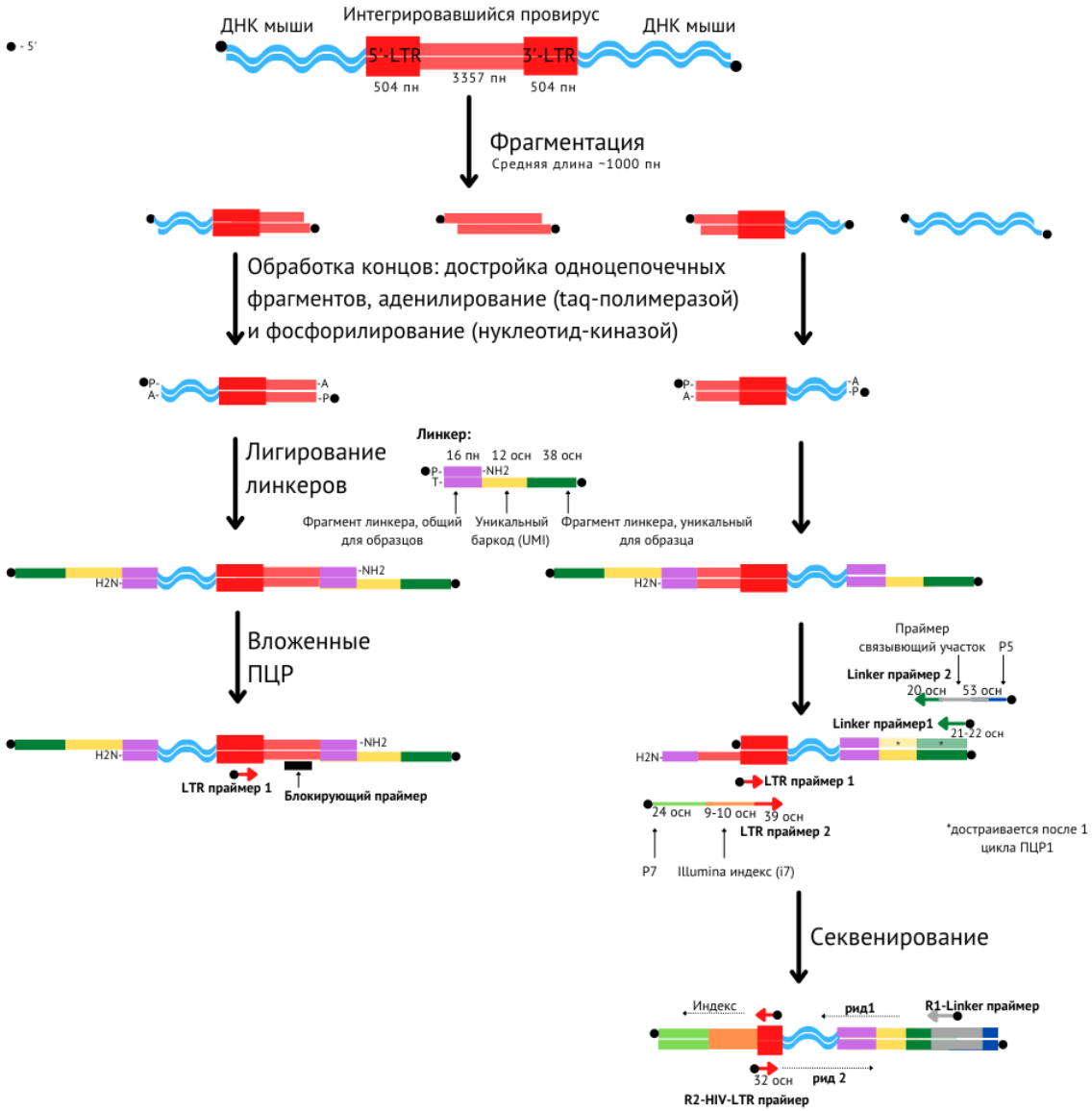


Рис. 6. Детализированная схема превращений ДНК в ходе пробоподготовки и секвенирования геномных данных.

В ходе изучения описанной схемы из оригинальной статьи и сопоставления последовательностей праймеров с последовательностями линкеров и LTR, на Рис. 5 из этой статьи была найдена ошибка – приведенная авторами схема предполагает, что наработка целевого продукта идет с 5' LTR, а побочного – с 3' LTR, тогда как на самом деле получается наоборот.

Для проверки правильности составленной схемы были выполнены выравнивания праймеров для ПЦР на последовательности линкера и LTR, и выравнивания праймеров для секвенирования на последовательности ПЦР праймеров. Для этого был написан код с использованием модуля pairwise2 из библиотеки Bio и инструмента для отображения выравниваний Pytexshade. Были подобраны следующие параметры выравнивания: +1 за совпадение и 0 за несовпадение нуклеотидов в строках, одинаковые штрафы величиной -10 за открытие и продолжение гэпа, отсутствие штрафов за концевые гэпы. Полученные выравнивания для всех праймеров одного из образцов показаны на Рис. 7. Были определены точные позиции расположения праймеров на сиквенсах, которые подтверждают, что использованные праймеры согласуются со схемой Рис. 6.

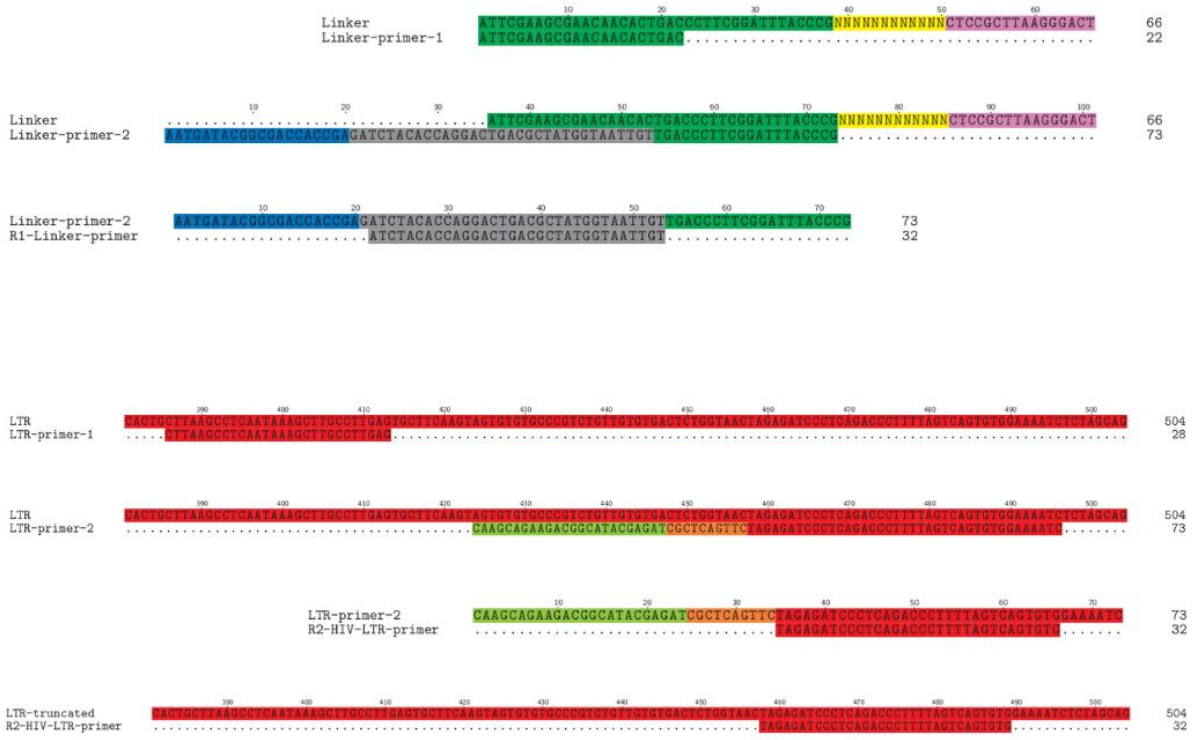


Рис. 7. Выравнивания праймеров для ПЦР (Linker-primer-1, LTR-primer-1 – для первого этапа ПЦР, Linker-primer-2, LTR-primer-2 – для второго этапа ПЦР) на

последовательности линкера и LTR, и праймеров для секвенирвоания (R1-Linker_primer и R2-HIV-LTR-primer – для прямого и обратного прочтения соответственно) на последовательности праймеров второго этапа ПЦР и LTR для R2-HIV-LTR-primer. Все последовательности на выравниваниях расположены в направлении от 5' к 3' концу.

С учетом схемы на Рис. 6, подтвержденной выравниваниями последовательностей праймеров, была определена предполагаемая структура прочтений, представленная на Рис. 8**,** согласно которой будет происходить дальнейший анализ прочтений – извлечение сайтов интеграции из обратного прочтения (рида 2) и уникальных молекулярных баркодов из прямого прочтения (рида 1).

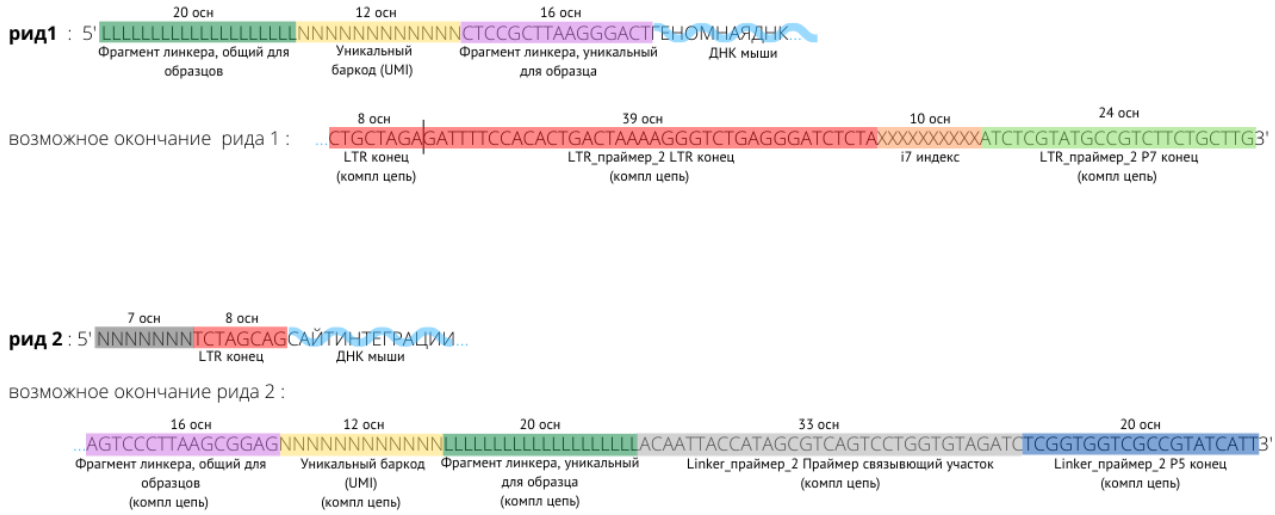


Рис. 8. Предполагаемая структура прочтений. Последовательности, из которых состоит прямое и обратное прочтение (рид 1 и рид 2) выделены цветами, аналогичными цветам этих элементов на схеме Рис. 6.

3.2 Оценка качества и количества прочтений

Для оценки пригодности данных к использованию был выбран стандартный метод оценки качества в программе **FastQC.** Отчет **FastQC** не показал значительных проблем с качеством, которые могли бы помешать целям исследования. Графики побуквенного качества прочтения прямого и обратного прочтений (рида 1 и рида 2), полученные в этом отчете приведены в приложении А. Они показали, что первые 7 оснований рида 2 в среднем

прочитаны с очень низким качеством, в большинстве прочтений на этом месте стоит последовательность NNNNNNN. Поэтому они не рассматривались при дальнейшем анализе. Следующие 8 оснований рида 2 имеют приемлемое качество. Эти нуклеотиды составляют последовательность фрагмента провирусного LTR и имеют большое значение, так как определяют наличие сайта интеграции. Их качества прочтения учитывались в дальнейшем для задания оптимальных параметров фильтрации данных. К концу прочтения качество падает, что может быть обусловлено накоплением ошибок в ходе синтеза и нехваткой реагентов в ячейке секвенатора, это исключает возможность анализа индексов и линкерных фрагментов, расположенных в конце прочтений.

Далее было определено количество прочтений в исходных данных, диаграмма показана на Рис. 9. В 49 образце наблюдалось очень малое количество прочтений, так же сравнительно малое количество наблюдалось в образце 4. Разброс количества данных в повторах оказался неравномерным.

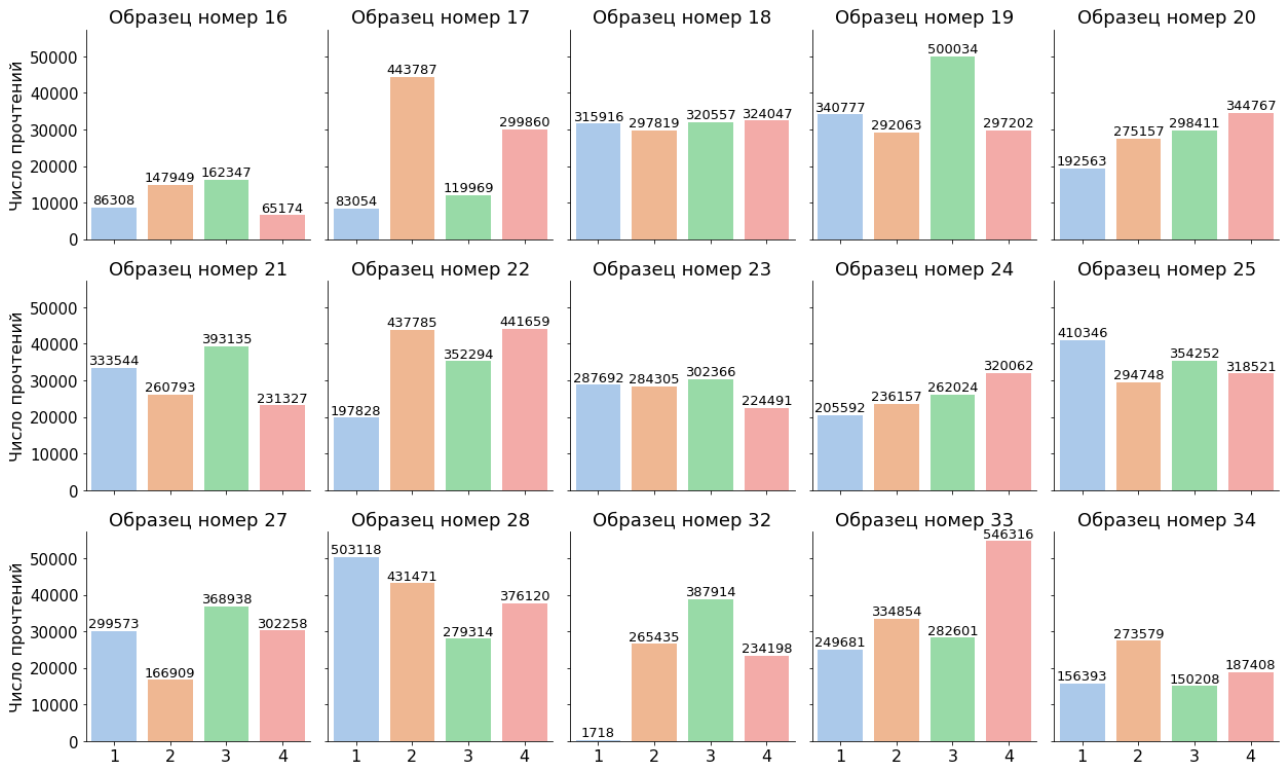


Рис. 9. Диаграмма количества прочтений в исходных fastq файлах. 4 повтора в каждом эксперименте обозначены на диаграмме цветами.

Для дальнейшего анализа качества прочтений был написан алгоритм проверки соответствия прочтений ожидаемой структуре, который определяет доли прочтений, содержащих определенные элементы в своей последовательности. В соответствии со схемой строения прочтений (Рис. 8) в риде 1 было предложено искать линкерные фрагменты в начале прочтения – уникальный для образца и общий для образцов фрагменты, так же проверяли наличие участка, комплементароного LTR, и i7 Illumina индекса в конце прочтения; в риде 2 - фрагмент LTR в начале и комплементарные линкерные фрагменты в конце. Результат представлен на рисунке 10, цветовое обозначение столбца диаграммы соответствует цветовому обозначению последовательностей прочтения на Рис. 8. Из диаграммы видно, что доли прямых прочтений, содержащих правильные последовательности линкерных элементов (зеленые столбики на левой диаграмме), уникальных для этого образца, для разных образцов составляют от 20 до 98%. Как выяснилось, это объясняется тем, что часть прочтений в каждом образце являются химерными. Такие прочтения содержат индекс одного образца (с которым они были прочитаны и определены секвенатором в соответствующий файл), и уникальный линкерный фрагмент, соответствующий другому образцу. Причинами возникновения химерных прочтений предположительно могут быть кросс-контаминации при синтезе линкеров (могут возникать, так как 1-цепочечные ДНК линкеров синтезируются все вместе, а лишь затем разделяются на электрофорезе) и кросс-контаминации в процессе пробоподготовки библиотек (могут возникать, так как после пробоподготовки перед заливкой обработанной библиотеки в ячейку образцы объединяли и между прочтениями могла произойти рекомбинация).

Также из диаграммы видно, что последовательности, расположенные в конце прочтения, находятся у малой доли прочтений, что согласуется с ранее обнаруженным явлением потери качества в конце прочтения.

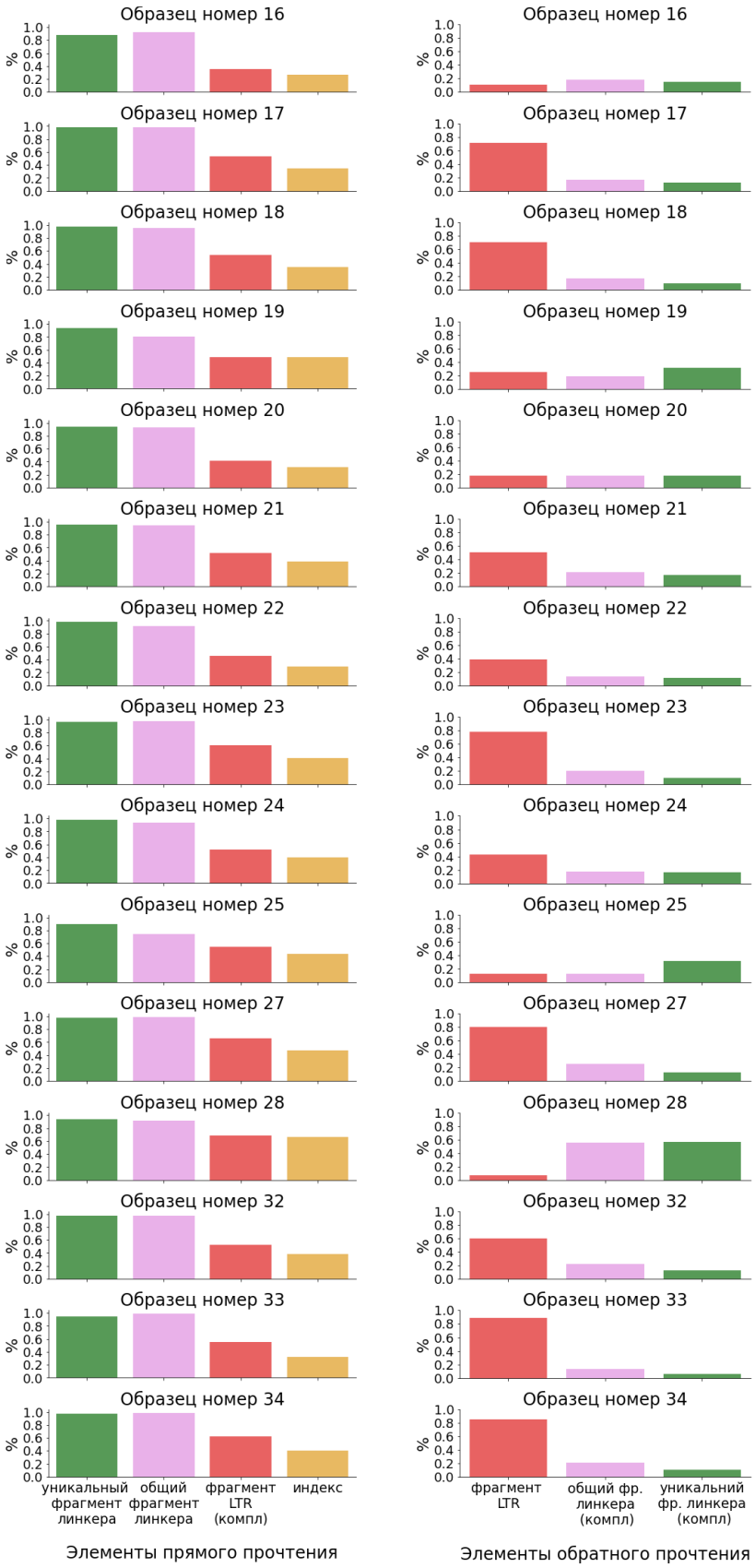


Рис. 10. Диаграмма долей прочтений, содержащих элементы линкера, LTR и индекс. Цветовое обозначение столбцов диаграммы соответствует цветовому обозначению элементов последовательности прочтения на Рис. 8.

Кроме того обнаруживается, что фрагмент LTR (красные на правой диаграмме) присутствует у малого числа прочтений во многих образцах, из чего можно сделать предположение о низкой эффективности лентивирусной трансдукции или же возникновении ошибках в ходе экспериментов, и заключение о необходимости проверки процедуры трансдукции. Таким образом приведенная диаграмма дает представление о степени соответствия прочтений ожидаемой структуре.

Следующим этапом проводилось выравнивание прочтений на референсный геном мыши и количественный анализ результатов. Была использована программа BWA с алгоритмом bwa mem для парных прочтений. На Рис. 11 показан количественный результат выравнивания в виде диаграммы долей корректно выровненных парных прочтений в образцах.

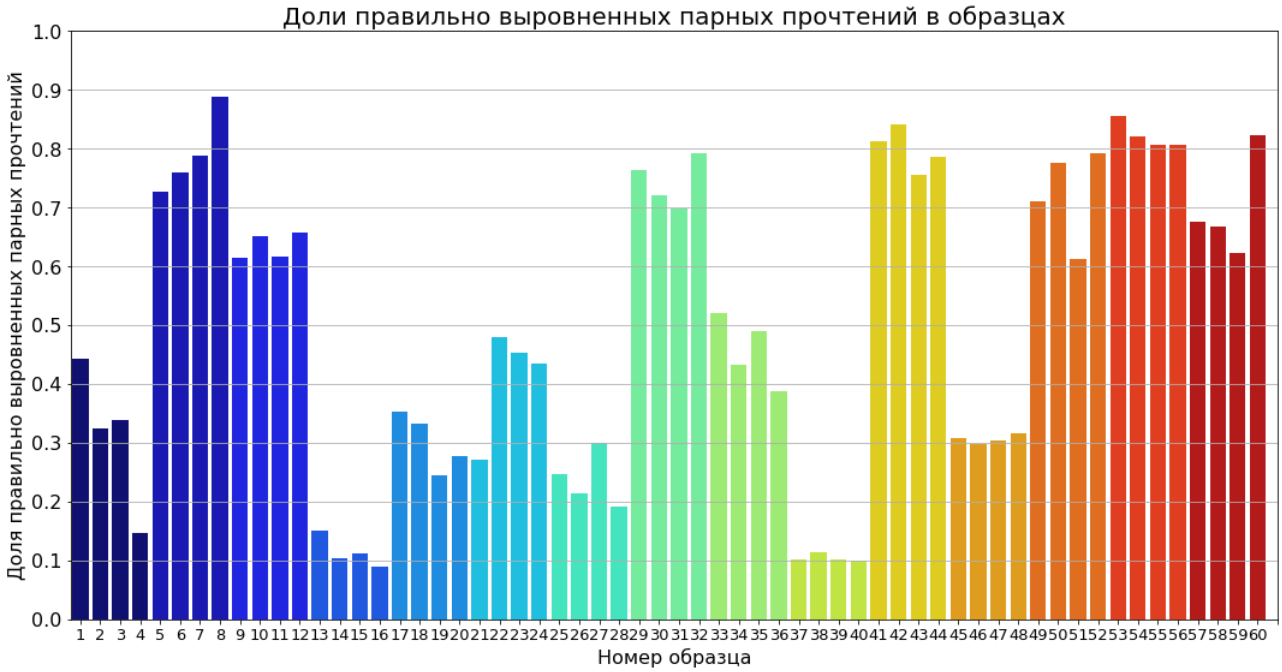


Рис. 11. Диаграмма долей корректно выровненных парных прочтений, повторы одного образца покрашены одинаковым цветом.

Различия в количествах выровненных прочтением среди образцов может объясняться как разной эффективностью трансдукции, так и дополнительными факторами - потерями ДНК в процессе пробоподготовки и долей маркированных клеток среди клеток очага. Так же можно видеть корреляцию с значениями долей обратных прочтений, содержащими LTR (красным на правой

диаграмме Рис. 10), объясняющуюся тем, что после фрагмента LTR располагается сайт интеграции – геномная ДНК мыши. Низкие доли выровненных прочтений и прочтений, содержащих LTR, а так же наличие химерных прочтений означают наличие большого процента загрязнений в данных в виде прочтений, из которых невозможно извлечь интересующую информацию о сайтах интеграции.

3.3 Фильтрация прочтений

Результаты оценки качества данных, приведенные в предыдущей главе, показали присутствие большого процента прочтений, не пригодных для анализа в интересующем контексте. Эти загрязнения необходимо отфильтровать, чтобы оставить только прочтения соответствующие ожидаемой структуре (выясненной в главе 3.1), содержащие сайт интеграции. Сайт интеграции – это место в геноме мыши, в которое был интегрирован провирус, он расположен после фрагмента LTR в обратном прочтении (рид 2 на схеме Рис. 8, сайт интеграции отмечен голубым цветом).

Параметры, по которым логично фильтровать прочтения это: наличие фрагмента провирусного LTR в обратном прочтении; наличие фрагментов линкера - уникального для образца и общего для образцов в прямом прочтении; правильное выравнивание парных прочтений на референсный геном; отсутствие последовательности внутреннего фрагмента провируса в обратном прочтении. Под правильным выравниванием здесь подразумеваются такие выравнивания, в которых: оба прочтения в паре выровнены на разные цепи ДНК и на одну хромосому; длина выравнивания превышает 19 оснований, выравнивания меньшей длины не достоверны, так как могут быть случайными; и при этом расстояние между началом выравнивания обратного прочтения и концом выравнивания прямого прочтения не превышает 3000 пн, цифра выбрана так, чтобы допускать выравнивания длинных последовательностей ДНК (средняя длина ДНК в исходной библиотеке – 1000 пн, разброс – от 60 до 2500, длина прочтения – 251 нуклеотид); кроме того удалялись те пары, у которых выравнивание обратного прочтения на геном начиналось с

последовательности, идентичной фрагменту LTR, так как в этом случае фрагмент произошел не из провируса, а из геномной ДНК, в которой случайно оказалась идентичная ему последовательность.

Поиск линкерных фрагментов в прочтениях выполнялся с точностью до одной ошибки (делеции, инсерции, вставки), так как качество прочтения в этой области достаточно высоко. Для фрагмента LTR в обратном прочтении так же допускалась одна ошибка, но учитывалось его положение в прочтении - начало с 8 буквы прочтения. Были написаны функции для выравнивания на основе алгоритма Pairwise2 bp из библиотеки Bio. Функция, которая допускает одну ошибку в выравнивании, считает Score выравнивания с параметрами: +1 за совпадение нуклеотида в последовательностях, -0.1 за несовпадение, -0.9 за открытие или продолжение гэпа; и сравнивает полученное значение Score с величиной (L-1-0.9), где L – длина выравнивания. Если Score превышает это значение, то выравнивание содержит две или более ошибки, и не отбирается как удовлетворяющее требованиям, в обратном случае – выравнивание отбирается как удовлетворяющее.

Для выравнивания использовался программный пакет BWA (Burrows-Wheeler Aligner). В ходе исследования было выяснено, что BWA одинаково эффективно выравнивает как прочтения с удаленными линкерными участками, так и необработанные прочтения. Для выяснения этого сначала была проведена фильтрация без предварительной обрезки прочтений по алгоритму, указанному выше, затем была проведена обрезка прочтений с помощью программы Cutadapt и последующее выравнивание оставшихся фрагментов геномной ДНК. При сравнении результатов оказалось, что не нашлось такого прочтения, который без предварительной обрезки не был выровнен, а после обрезки был выровнен. Это объясняется тем, что у алгоритма bwa mem, использованного здесь для выравнивания парных прочтений, имеется встроенный soft clipping при выравнивании, то есть он игнорирует края, не поддающиеся выравниванию на референсный геном, и они не мешают анализу остальной части

последовательности. Был сделан вывод о том, что выравнивание без обрезки в данном случае допустимо.

Описанные шаги фильтрации были применены к данным, количественный результат в виде числа прочтений, оставшихся в файлах после фильтрации, показан на Рис.12. Полученная картина хорошо коррелирует с диаграммой, показывающей доли обратных прочтений, в которых содержится фрагмент LTR правильной на правильном месте, показанной на Рис. 13. Что говорит о том, что размер файлов с прочтениями, оставшимися после фильтрации, главным образом лимитировался числом отфильтрованных парных прочтений, содержащих фрагмент LTR в обратном прочтении.

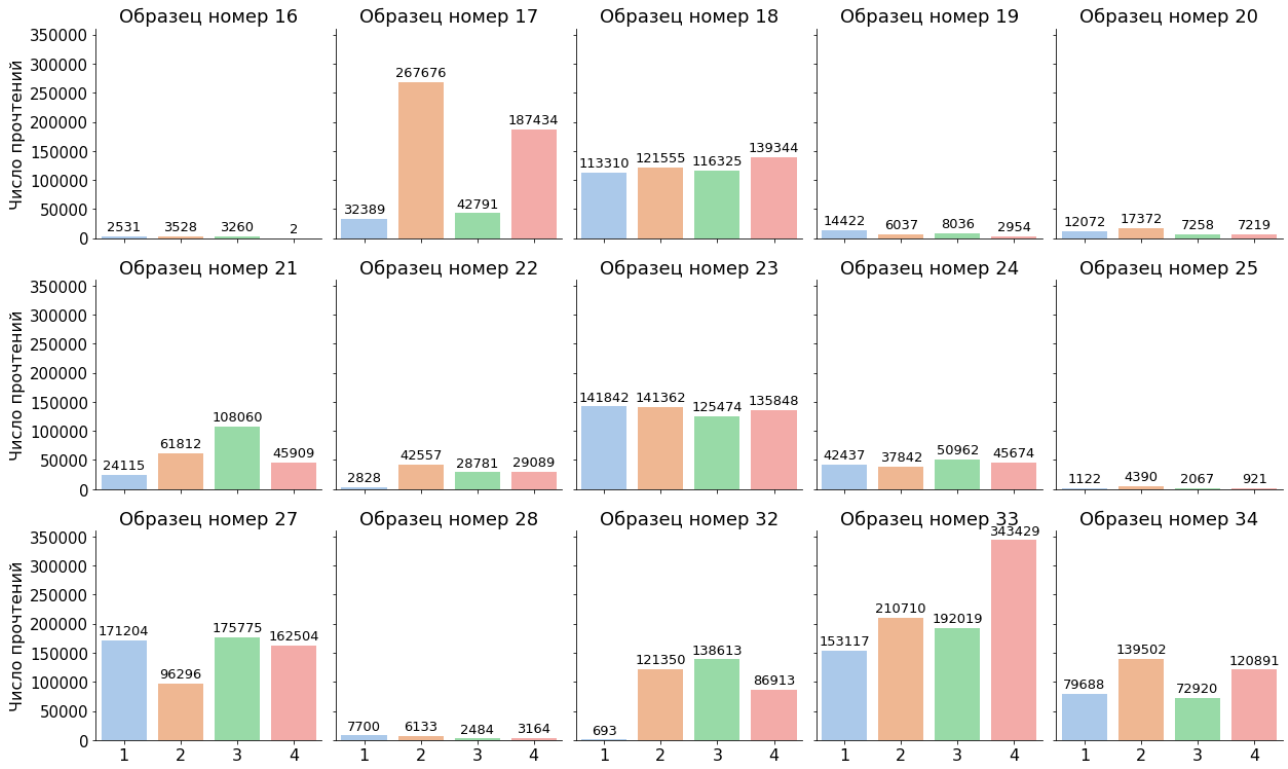


Рис. 12. Диаграмма количества прочтений в файлах, полученных в результате фильтрации.

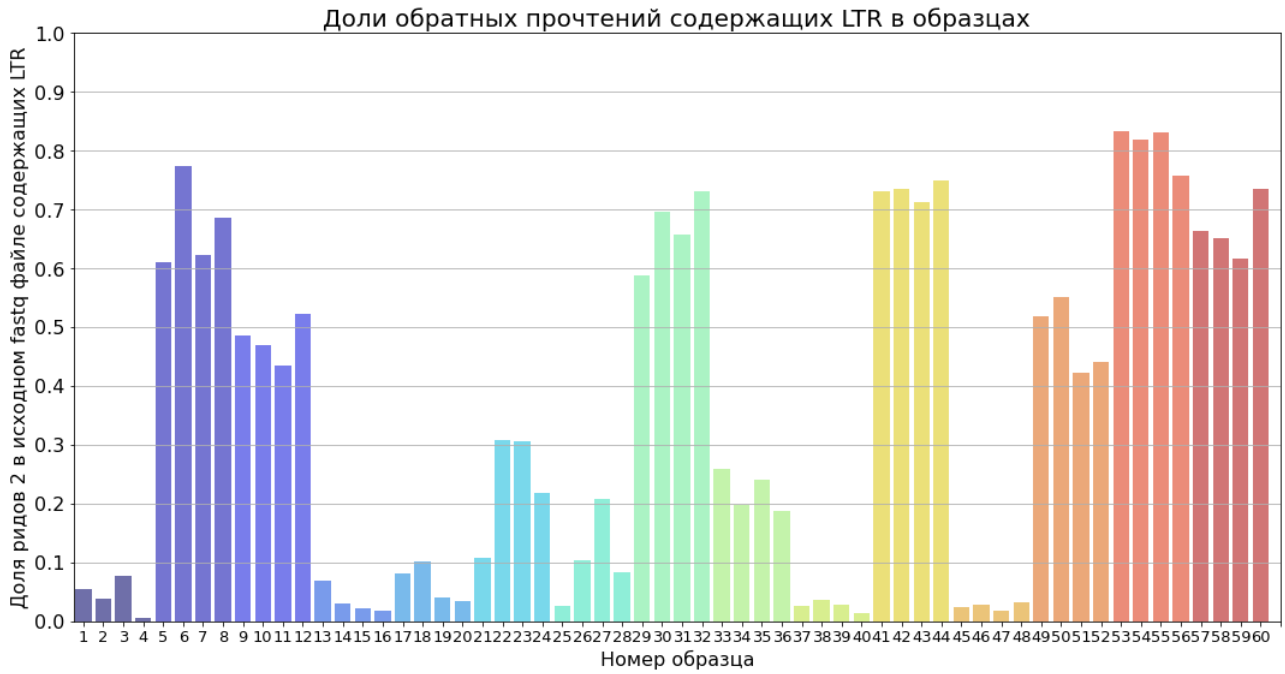


Рис. 13. Диаграмма долей обратных прочтений, содержащих фрагмент LTR.

3.4 Кластеризации данных

Следующим этапом анализа данных после фильтрации необходимо было рассмотреть структуру библиотеки в контексте степени её амплификации. Необходимо выяснить каково среднее число прочтений уникальных молекул ДНК в библиотеке после фрагментации ультразвуком. Число прочтений уникальной молекулы равно числу прочтений уникального молекулярного баркода (UMI - unique molecular identifier), который присутствует в линкере, пришитом к молекуле на этапе пробоподготовки. UMI – это последовательность из случайного набора 12 нуклеотидов, которая, пришиваясь в составе линкера, маркирует каждую уникальную молекулу, поступившую на этап пробоподготовки. В оригинальном исследовании предполагалось, что каждой уникальной последовательности UMI, обнаруженной в данных после анализа, соответствует одна клетка, поступившая в исследование. В дальнейшем анализе эти последовательности будут использоваться для оценки размера клеточных клонов, который соответствует набору UMI, с которыми был прочитан каждый сайт интеграции.

Опираясь на это положение, для количественного описания числа прочтений уникальных молекул на тестовом наборе данных было построено

распределение параметра числа прочтений уникальных последовательностей UMI для всех образцов суммарно, диаграмма показана на Рис. 14. Она показала, что большинство UMI в библиотеке было прочитано от одного до шести раз, так же видно, что присутствуют такие последовательности UMI, которые были прочитаны >1000 раз, что свидетельствует о неравномерной амплификации библиотеки – некоторые последовательности при секвенировании были прочитаны значимо больше раз, чем другие.

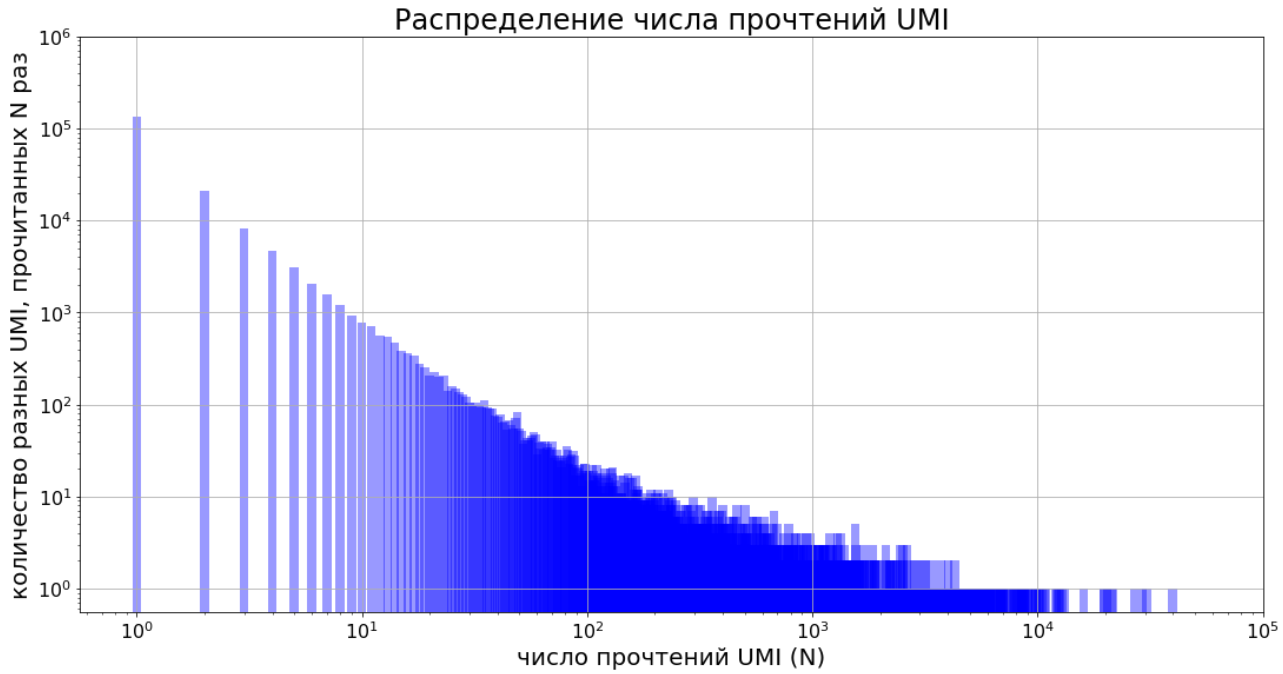


Рис. 14. Диаграмма числа прочтений UMI во всех образцах суммарно.

Тот факт, что присутствуют последовательности, число прочтений которых велико (около тысячи или больше) означает, что при анализе необходимо учесть накопление ошибок в ходе ПЦР на этапах пробоподготовки и секвенирования. Любые последовательности в прочтении, в том числе UMI, могут накапливать мутации при амплификации. В результате этого в прочтениях будут присутствовать наборы схожих, но не полностью идентичных последовательностей, которые могут быть учтены как разные, хотя они происходят от одной последовательности. Для учета этого эффекта было предложено разработать методы кластеризации последовательностей UMI – определение кластеров схожих последовательностей, произошедших от одной оригинальной.

Для отображения кластеров UMI, предположительно присутствующих в данных, были построены графы, отображающие пространство последовательностей UMI для каждого образца. Графы строились по следующему принципу: вершина графа соответствуют уникальной последовательности UMI, ребро соединяет пару UMI, расстояние Левенштейна между которыми равняется единице. Такой граф показан на Рис. 15.

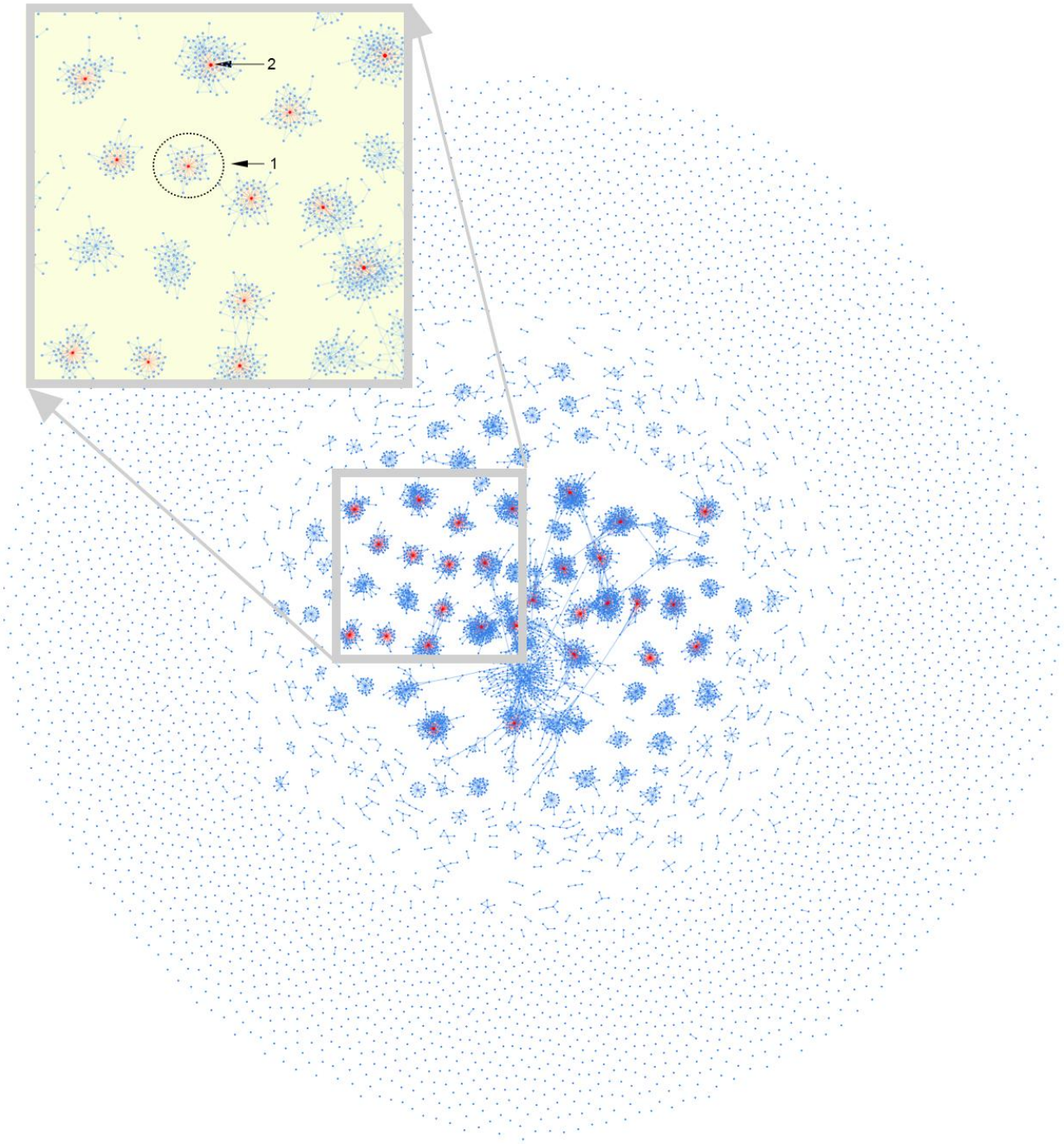


Рис. 15. Граф, отображающий структуру пространства последовательностей UMI в образце номер 17. 1 - кластер последовательностей, 2 – центр кластера - вершина, соответствующая UMI с множественными прочтениями. Красным

цветом покрашены вершины, соответствующие UMI, которые были прочитаны >5000 раз.

На данном графе видны кластеры, центрами которых в большинстве случаев являются UMI с большим числом прочтений. Дальнейшая задача состоит в определении оригинального UMI, то есть центра кластера. Логично, что при кластеризации нужно объединять близкие последовательности в пользу более прочитанной последовательности. Для этого вершинам были приписаны веса, соответствующие числу прочтений последовательности UMI, на их основе в графе было задано направление - ребро направлено в сторону вершины соответствующей последовательности UMI, прочитанной большее количество раз. Таким образом, кластеризация получается автоматически в результате построения такого графа.

Для выполнения дальнейшей процедуры объединения кластеров был написан программный код, работающий с графом по следующему алгоритму: создавался словарь, содержащий названия и веса вершин, степень которых отлична от нуля; он упорядочивался по убыванию величины весов; затем производился проход по всем вершинам из словаря, при этом для каждой из этих вершин C: если эта вершина не была изменена ранее (не была объединена с каким-то другим центром кластера), то для нее определялся набор соседствующих вершин, которые объединяли с ней, их веса прибавлялись к весу C.

Кроме того так как в графах встречаются достаточно запутанные структуры необходимо было выбрать оптимальный критерий отличия исходно различных последовательностей и последовательностей с накопленными ошибками. Таким критерием может служить минимальный вес центра кластера, для определения которого была проведена оценка накопления ошибок в данных на этапе пробоподготовки.

Для оценки накопления ошибок в данных в ходе амплификации на этапе пробоподготовки была построена статистическая модель влияния числа циклов ПЦР на накопление ошибок в последовательностях UMI. По формуле Бернулли

рассчитывались вероятность накопления n ошибок в последовательности UMI в ходе N циклов ПЦР. Вероятность прочтения основания без ошибки зависит от точности полимеразы (error_rate) и равна $p = (1 - error\_rate)^N$. Точность полимеразы, использованной при пробоподготовке (Clontech Advantage 2 PCR Polymerase mix от японской компанией Takara Bio) в три раза выше точности Taq-полимеразы и равна примерно $0.6*10^{-4}$ ошибок/осн/цикл. Формула для расчета вероятности появления n ошибок в UMI:

$$P(n) = \frac{12!}{\mathrm{n!}\,(12 - \mathrm{n})!} * \mathrm{p}^{12-\mathrm{n}} * (1 - \mathrm{p})^{\mathrm{n}}$$

На этапе пробоподготовки по протоколу проводилось 45 циклов ПЦР. Для n=1 и N=45 вероятность P равна 0.0314, что означает, что примерно 1 из 30 последовательностей будет содержать хотя бы одну ошибку.

Из этих соображений был задан минимальный вес центра кластера равный 30. Таким образом, объединение вершин проводилась в пользу вершины с наибольшим весом, при условии, что вес >30. Граф, полученный в результате объединения последовательностей в кластерах, для 17 образца показан в Приложении Б.

3.5 Разработка анализа сравнения данных в технических повторах

Следующая задача анализа - получение сайтов интеграции из прочтений. Сайтом интеграции считалась точка начала выравнивания обратного прочтения на референсный геном. Был написан алгоритм, извлекающий эту информацию из отфильтрованных данных, результат в виде количества сайтов, извлеченных из тестовых данных, указан в таблице 2.

Таблица 2. Количество первичных сайтов интеграции, извлеченных из данных.

Образец	**Число сайтов**	**Источник данных в эксперименте**
16	827	Внутренняя масса очага в интактном реципиенте 1
17	1374	Внутренняя масса очага в интактном реципиенте 2
18	4791	Внутренняя масса очага в интактном реципиенте 3
19	1242	Внутренняя масса очага в интактном реципиенте 4

20	1540	Внутренняя масса очага в интактном реципиенте 5
21	1420	Внутренняя масса очага в облученном реципиенте 6
22	605	Внутренняя масса очага в облученном реципиенте 8
23	3207	Внутренняя масса очага в облученном реципиенте 9
24	1561	Внутренняя масса очага в облученном реципиенте 10
25	287	Контроль (без лентивирусной трансдукции)
27	6338	Косточка очага в интактном реципиенте 1
28	1752	Косточка очага в интактном реципиенте 2
32	3401	Косточка очага в облученном реципиенте 6
33	4871	Косточка очага в облученном реципиенте 8
34	6598	Косточка очага в облученном реципиенте 9

Был написан алгоритм для сопоставления каждого сайта интеграции с набором UMI, которые присутствовали в прочтениях, содержащих сайт. Важным показателем качества анализа является равномерность распределения полученных сайтов по техническим повторам экспериментов. Повторы были сделаны методом разделения библиотеки после 1 этапа ПЦР на 4 образца, при этом данные должны распределяться между повторами равномерно. Результаты были оформлены в виде таблиц, которые содержат значения количества UMI, с которыми были прочитаны сайты в технических повторах, часть этих данных отображена в таблице 3. Ожидалось, что в технических образцах сайты будут прочитаны равномерно. Однако, как видно из таблицы 3 обнаружились сильные перекосы - большинство сайтов были прочитаны с несколькими UMI в одном из технических повторов и с 0 UMI в других повторах. Это означает, что присутствовало распределение исходных молекул ДНК по повторам, что статистически невозможно, большие цифры таких событий должны быть ложными.

Таблица 3. Количества уникальных UMI (N_{UMI}) с которыми прочитаны сайт интеграции в технических повторах. Сайт интеграции записан в виде: №хромосомы:[геномная координата сайта)/цепь на которой сайт расположен

Сайт интеграции	N_{UMI}, повтор 1	N_{UMI}, повтор 2	N_{UMI}, повтор 3	N_{UMI}, повтор 4
13:[14791564,14791565)/+ цепь	1	661	0	3
12:[63600031,63600032)/ −цепь	1	83	0	0
7:[30175405,30175406)/+цепь	1	0	0	0
10:[16400965,16400966)/+цепь	28	0	0	0
11:[87943744,87943745)/ −цепь	1	142	0	0
6:[34342167,34342168)/+цепь	0	461	0	1
…	…	…	…	…

При дальнейшем изучении данных выяснилось, что присутствуют сайты, определенные как разные, но прочитанные с одинаковыми UMI. Причем такие сайты в большинстве очень похожи - геномные позиции различаются в пределах 100 пн. Очевидно это исходно один и тот же сайт, но после пробоподготовки, ПЦР и секвенирования молекулы различаются из-за накопления ошибок в ходе амплификации. Был сделан вывод о том, что необходимо провести кластеризацию сайтов интеграции прежде, чем сопоставлять их с последовательностями UMI. Кластеризация сайтов проводилась аналогично кластеризации UMI – методом построения графа с взвешенными вершинами, отображающего пространство сайтов. Расстояние между вершинами было задано равным разности координат геномных позиций сайтов в том случае если сайты лежат на одной хромосоме, если сайты лежат на разных хромосомах – расстояние задавалось большой цифрой означающей отсутствие связи между сайтами. Родственными считались сайты, лежащие на расстоянии до 100 пн. Изображение такого графа приведено в Приложении В. Количественный результат кластеризации представлен в таблице 4, которая показывает, что равномерность распределения сайтов интеграции в технических повторах повысилась после проведение кластеризации сайтов интеграции. Так же был удален шум в виде сайтов, прочитанных малое количество раз (меньше восьми раз).

Таблица 4. Количество сайтов интеграции после кластеризации.

Образец	**Число сайтов**	**Источник данных в эксперименте**
16	200	Внутренняя масса очага в интактном реципиенте 1
17	99	Внутренняя масса очага в интактном реципиенте 2
18	534	Внутренняя масса очага в интактном реципиенте 3
19	201	Внутренняя масса очага в интактном реципиенте 4
20	187	Внутренняя масса очага в интактном реципиенте 5
21	214	Внутренняя масса очага в облученном реципиенте 6
22	64	Внутренняя масса очага в облученном реципиенте 8
23	387	Внутренняя масса очага в облученном реципиенте 9
24	175	Внутренняя масса очага в облученном реципиенте 10
25	33	Контроль (без лентивирусной трансдукции)
27	730	Косточка очага в интактном реципиенте 1
28	596	Косточка очага в интактном реципиенте 2
32	399	Косточка очага в облученном реципиенте 6
33	573	Косточка очага в облученном реципиенте 8
34	1086	Косточка очага в облученном реципиенте 9

Последним этапом анализа данных, на котором извлекается искомая информация о количестве и размере клонов в исследуемых клетках, является подсчет количества разных UMI, которые присутствовали в прочтениях, содержащих уникальный сайт. Результаты этого подсчета позволяют сформулировать биологически значимые выводы. На тестовых данных было видно, что большинство сайтов прочитано с единичными UMI, что интерпретируется как нахождение единичных клеток содержащих эти сайты в исходных данных. Заметно, что в косточках находится больше сайтов чем во внутренней массе очага, что может быть связанно с значительно большей долей донорских клеток в материале косточки, чем во внутренней массе. Выявленное количество сайтов интеграции в тестовых данных требует интерпретирования и сравнения с оценками, полученными независимым образом.

4 Заключение

Были проанализированы детали эксперимента по созданию эктопических очагов кроветворения из клеток МСК, маркированных лентивирусным вектором, пробоподготовки и секвенирования геномных данных. Разработан алгоритм анализа данных секвенирования геномов маркированных клеток, который включает в себя этапы оценки качества и размера данных, фильтрации данных в соответствии с ожидаемой структурой прочтений, кластеризации сайтов интеграции и последовательностей уникальных баркодов, соотнесения сайтов с баркодами и проверки равномерности распределения полученных данных в технических повторах.

Алгоритм был испытан на тестовом наборе данных из экспериментов по созданию эктопических очагов кроветворения из клеток, маркированных лентивирусом. В процессе оценки качества данных было определено содержание прочтений, пригодных для анализа, оно оказалось малым (порядка нескольких сотен для образцов клеточной массы, около одной тысячи для образцов из косточек) для многих образцов и лимитировалось в основном содержанием фрагмента LTR в обратных прочтениях. В большей части образцов содержание LTR фрагмента в прочтениях не превышало 50%, что может свидетельствовать о низкой эффективности лентивирусной трансдукции в экспериментах, на основе которых был получен тестовый набор данных. Также требуется выяснение наличия факторов, способствующих накоплению загрязнений в данных. Результаты анализа данных показали, что для эффективного анализа необходимо увеличить глубину прочтения библиотеки, так как по проведенной оценке числа прочтений уникальных последовательностей большинство последовательностей прочитано от 1 до 6 раз. Выявленное количество сайтов интеграции в тестовых данных, соответствующее количеству клонов, требует интерпретирования и сравнения с оценками, полученными независимым образом. Созданный алгоритм может быть применен в будущем для анализа количества клонов МСК при наличии более репрезентативного набора данных.

5 Выводы

1. В данной работе была разработана и реализована схема анализа данных секвенирования геномов популяции мезенхимальных стволовых клеток, маркированных лентивирусным вектором, для оценки количества и размера клонов в популяции.
2. Был разработан и апробирован на данных метод, позволяющий при анализе данных секвенирования учитывать мутацации в ДНК-баркодах, основанный на предложенной математической модели накопления мутаций в ДНК-баркодах в ходе ПЦР.
3. Разработанные подходы к анализу данных были апробированы на тестовом наборе экспериментальных данных. Результаты показали низкую (меньше 50%) степень соответствия прочтений предполагаемой структуре в половине образцов, обусловленную отсутствием фрагментов лентивирусных меток в прочтениях и наличием химерных прочтений. Сформулированы рекомендации о разработке методов снижения такого рода загрязнений в данных и проверке эффективности процедуры лентивирусного маркирования независимым образом.
4. Созданный алгоритм может быть использован для обработки экспериментальных данных, которые позволят оценить пролиферативный и дифференцировочный потенциал популяции мезенхимных стволовых клеток в красном костном мозге мышей.

6 Приложения

Приложение А

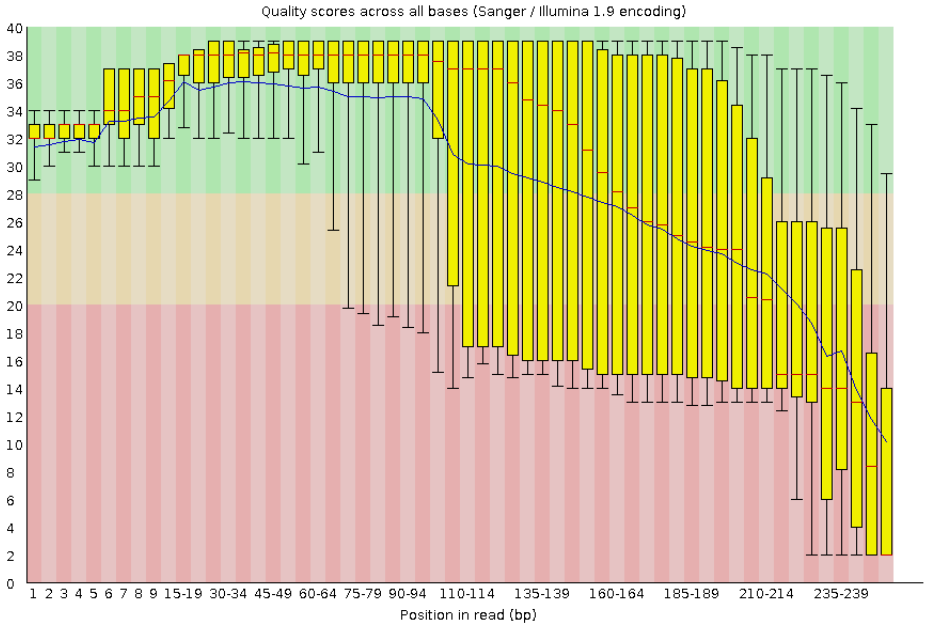


График усредненного значения показателя качества phred score для нуклеотидов в прямом прочтении.

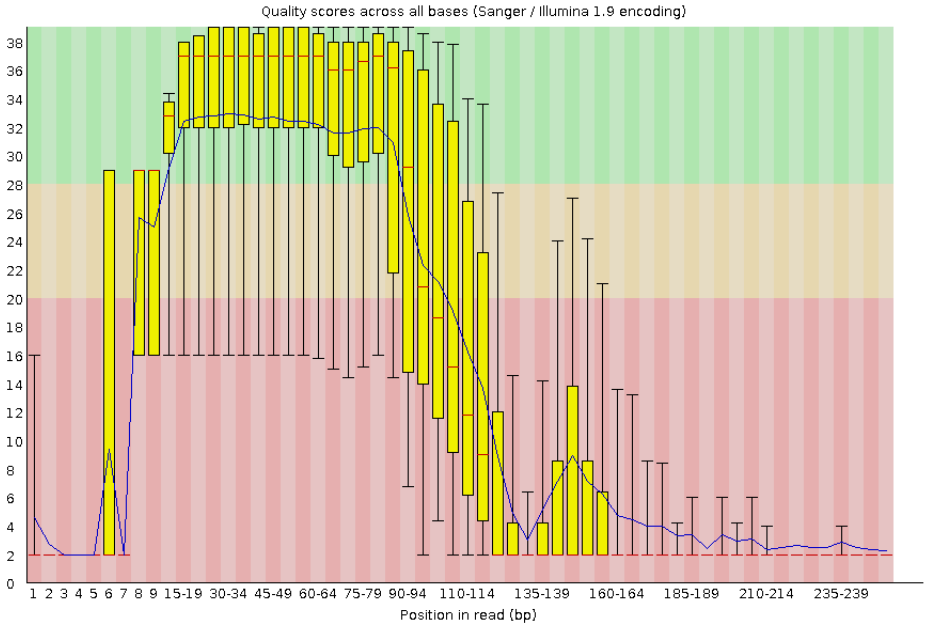
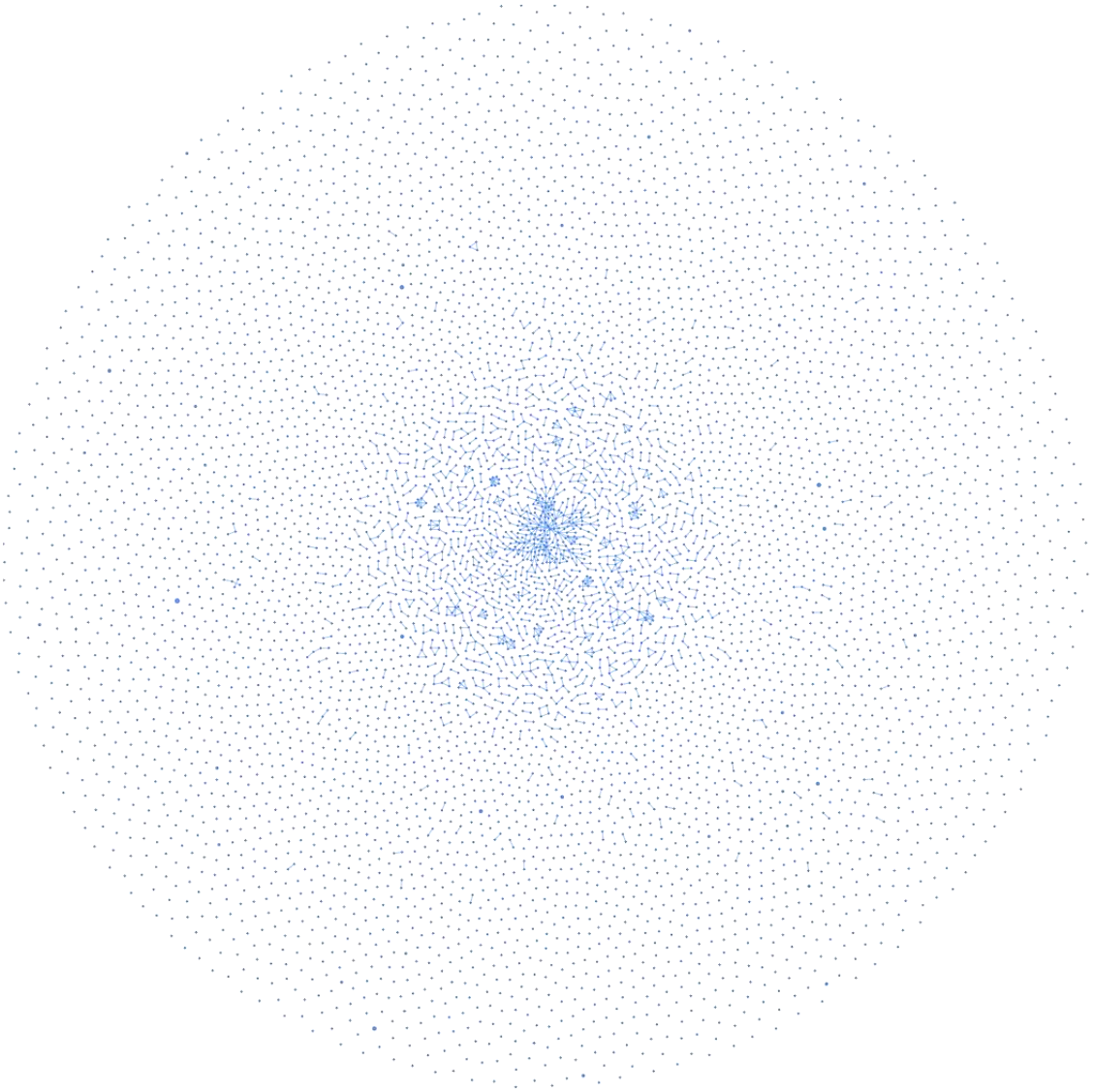


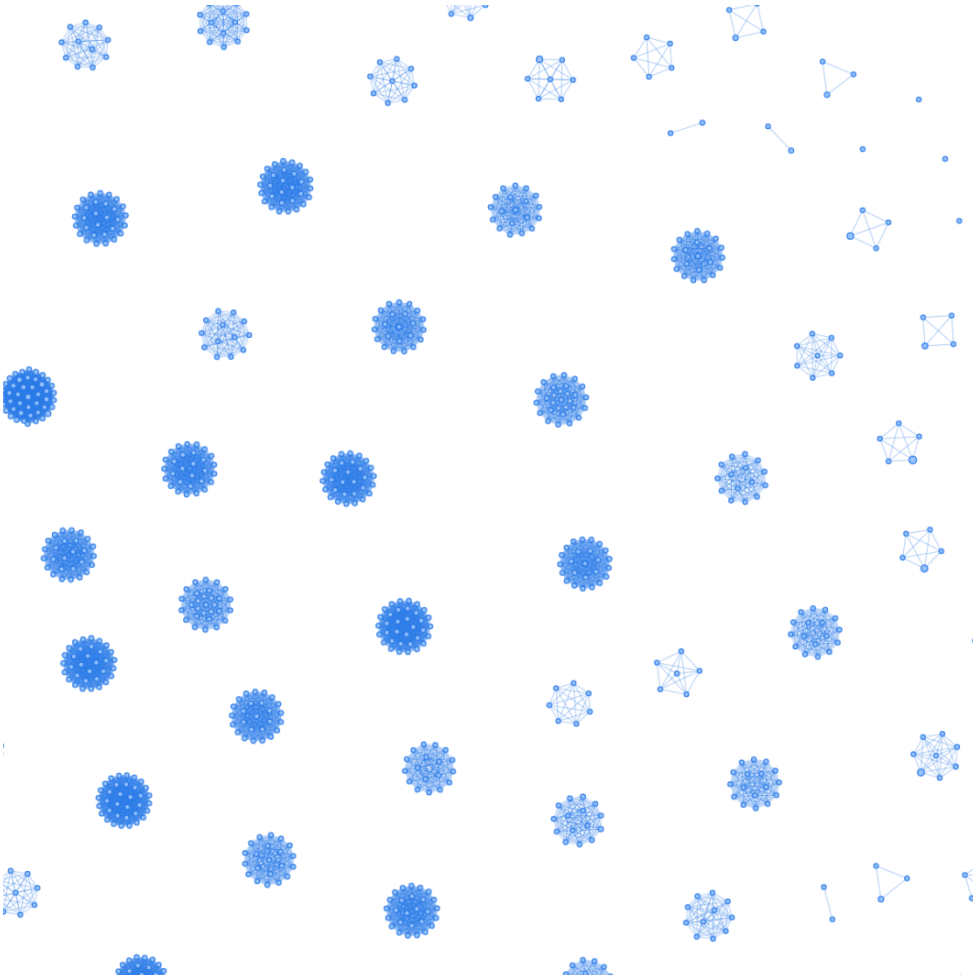
График усредненного значения показателя качества phred score для нуклеотидов в обратном прочтении.

Приложение Б



Результат кластеризации последовательностей UMI для 17 образца.

Приложение В



Фрагмента графа, отображающего полученные кластеры сайтов интеграции.

7 Список литературы

1. Владимирская Е.Б. Мезенхимальные стволовые клетки (мск) в клеточной терапии // Онкогематология. 2007. № 1. P. 4–16.

2. Чертков И.Л., Гуревич О.А. Книга «Стволовая кроветворная клетка и ее микроокружение» — Москва: Медицина, 1984. — 239 С.

3. A Hagberg, D Schult, P Swart. Exploring Network Structure, Dynamics, and Function using NetworkX // in Proceedings of the 7th Python in Science conference (SciPy 2008), G Varoquaux, T Vaught, J Millman (Eds.), pp. 11-15 [Электронный ресурс]. URL: https://conference.scipy.org/proceedings/scipy2008/paper_2/.

4. Anders S., Pyl P.T., Huber W. HTSeq—a Python framework to work with high-throughput sequencing data // Bioinformatics. 2015. V. 31. № 2. P. 166–169.

5. Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data [Электронный ресурс]. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

6. Bedwell G.J., Engelman A.N. Factors that mold the nuclear landscape of HIV-1 integration // Nucleic Acids Res. 2020. V. 49. № 2. P. 621–635.

7. Bigildeev A., Pilunov A., Sats N., Petinati N., Surin V., Drize N. Marking of human multipotent mesenchymal stromal cells by lentiviral barcoded library revealed dynamic polyclonality in their population through passages // Exp. Hematol. 2017. V. 53. P. S111.

8. Cann A.J. Principles of Molecular Virology. , 2012. P. 93.

9. Church D.M., Schneider V.A., et al. Modernizing reference genome assemblies // PLoS Biol. 2011. V. 9. № 7. P. e1001091.

10. Cock, P.J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), pp.1422–1423. [Электронный ресурс]. URL: https://biopython.org/.

11. Dull T., Zufferey R., Kelly M., Mandel R.J., Nguyen M., Trono D., Naldini L. A third-generation lentivirus vector with a conditional packaging system // J. Virol. 1998. V. 72. № 11. P. 8463–8471.

12. Engelman A., Cherepanov P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication // PLoS Pathog. 2008. V. 4. № 3. P. e1000046.

13. Engelman A.N. Multifaceted HIV integrase functionalities and therapeutic strategies for their inhibition // J. Biol. Chem. 2019. V. 294. № 41. P. 15137–15157.

14. Engelman A.N., Singh P.K. Cellular and molecular mechanisms of HIV-1 integration targeting // Cell. Mol. Life Sci. CMLS. 2018. V. 75. № 14. P. 2491–2507.

15. Francis A.C., Marin M., Singh P.K., Achuthan V., Prellberg M.J., Palermino-Rowland K., Lan S., Tedbury P.R., Sarafianos S.G., Engelman A.N., Melikyan G.B. HIV-1 replication complexes accumulate in nuclear speckles and integrate into speckle-associated genomic domains // Nat. Commun. 2020. V. 11. P. 3505.

16. Goodwin S., McPherson J.D., McCombie W.R. Coming of age: ten years of next-generation sequencing technologies // Nat. Rev. Genet. 2016. V. 17. № 6. P. 333–351.

17. Hmadcha A., Martin-Montalvo A., Gauthier B.R., Soria B., Capilla-Gonzalez V. Therapeutic Potential of Mesenchymal Stem Cells for Cancer Therapy // Front. Bioeng. Biotechnol. 2020. V. 8.

18. Integrative Biology Group. A python wrapper for TexShade sequence alignment shader. , 2022.

19. Lesbats P., Engelman A.N., Cherepanov P. Retroviral DNA Integration // Chem. Rev. 2016. V. 116. № 20. P. 12730–12757.

20. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinforma. Oxf. Engl. 2009. V. 25. № 14. P. 1754–1760.

21. Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., Law M. Comparison of next-generation sequencing systems // J. Biomed. Biotechnol. 2012. V. 2012. P. 251364.

22. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads // EMBnet.journal. 2011. V. 17. № 1. P. 10–12.

23. Nakamura K., Oshima T., Morimoto T., Ikeda S., Yoshikawa H., Shiwa Y., Ishikawa S., Linak M.C., Hirai A., Takahashi H., Altaf-Ul-Amin M., Ogasawara N., Kanaya S. Sequence-specific error profile of Illumina sequencers // Nucleic Acids Res. 2011. V. 39. № 13. P. e90.

24. Perrone G., Unpingco J., Lu H. Network visualizations with Pyvis and VisJS // 2020.

25. Poletti V., Mavilio F. Interactions between Retroviruses and the Host Cell Genome // Mol. Ther. Methods Clin. Dev. 2018. V. 8. P. 31–41.

26. Ravi R.K., Walton K., Khosroheidari M. MiSeq: A Next Generation Sequencing Platform for Genomic Analysis // Methods Mol. Biol. Clifton NJ. 2018. V. 1706. P. 223–232.

27. Schirmer M., Ijaz U.Z., D'Amore R., Hall N., Sloan W.T., Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform // Nucleic Acids Res. 2015. V. 43. № 6. P. e37.

28. Serrao E., Krishnan L., Shun M.-C., Li X., Cherepanov P., Engelman A., Maertens G.N. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding // Nucleic Acids Res. 2014. V. 42. № 8. P. 5164–5176.

29. Sherman E., Nobles C., Berry C.C., Six E., Wu Y., Dryga A., Malani N., Male F., Reddy S., Bailey A., Bittinger K., Everett J.K., Caccavelli L., Drake M.J., Bates P., Hacein-Bey-Abina S., Cavazzana M., Bushman F.D. INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes // Mol. Ther. Methods Clin. Dev. 2017. V. 4. P. 39–49.

30. Weber K., Bartsch U., Stocking C., Fehse B. A Multicolor Panel of Novel Lentiviral “Gene Ontology” (LeGO) Vectors for Functional Gene Analysis // Mol. Ther. 2008. V. 16. № 4. P. 698–706.

31. Wu X., Li Y., Crise B., Burgess S.M., Munroe D.J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses // J. Virol. 2005. V. 79. № 8. P. 5211–5214.

32. Zufferey R., Dull T., Mandel R.J., Bukovsky A., Quiroz D., Naldini L., Trono D. Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery // J. Virol. 1998. V. 72. № 12. P. 9873–9880.