МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени М.В.ЛОМОНОСОВА
Биологический факультет
Кафедра биоинженерии
Группа интегративной биологии

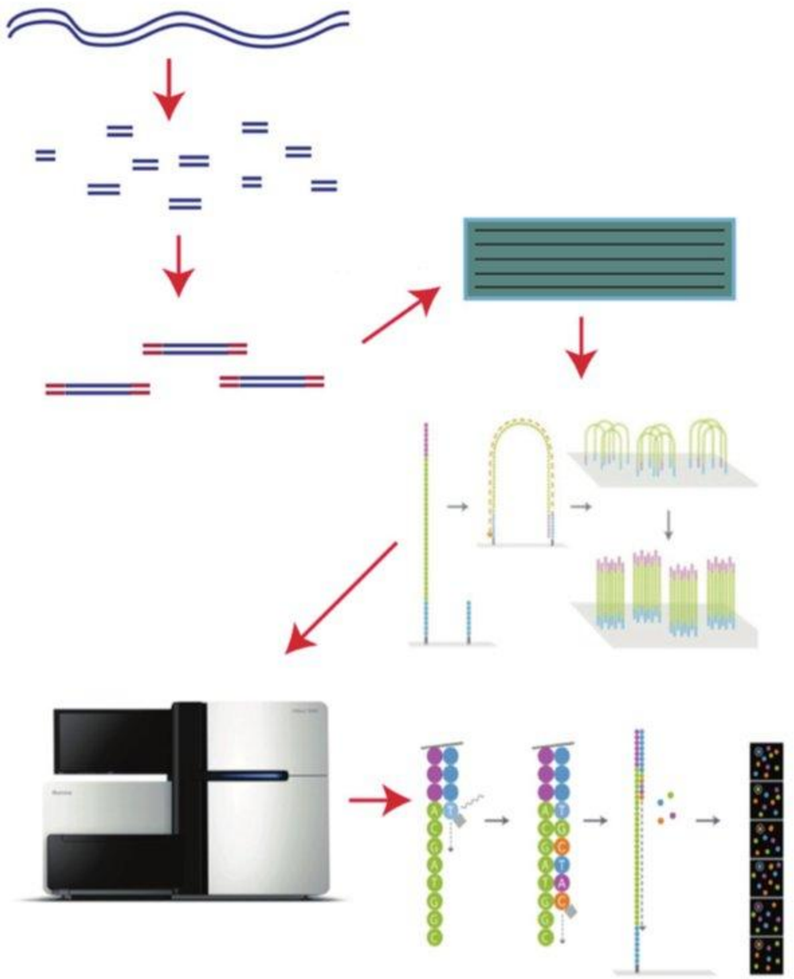
Выпускная квалификационная работа бакалавра

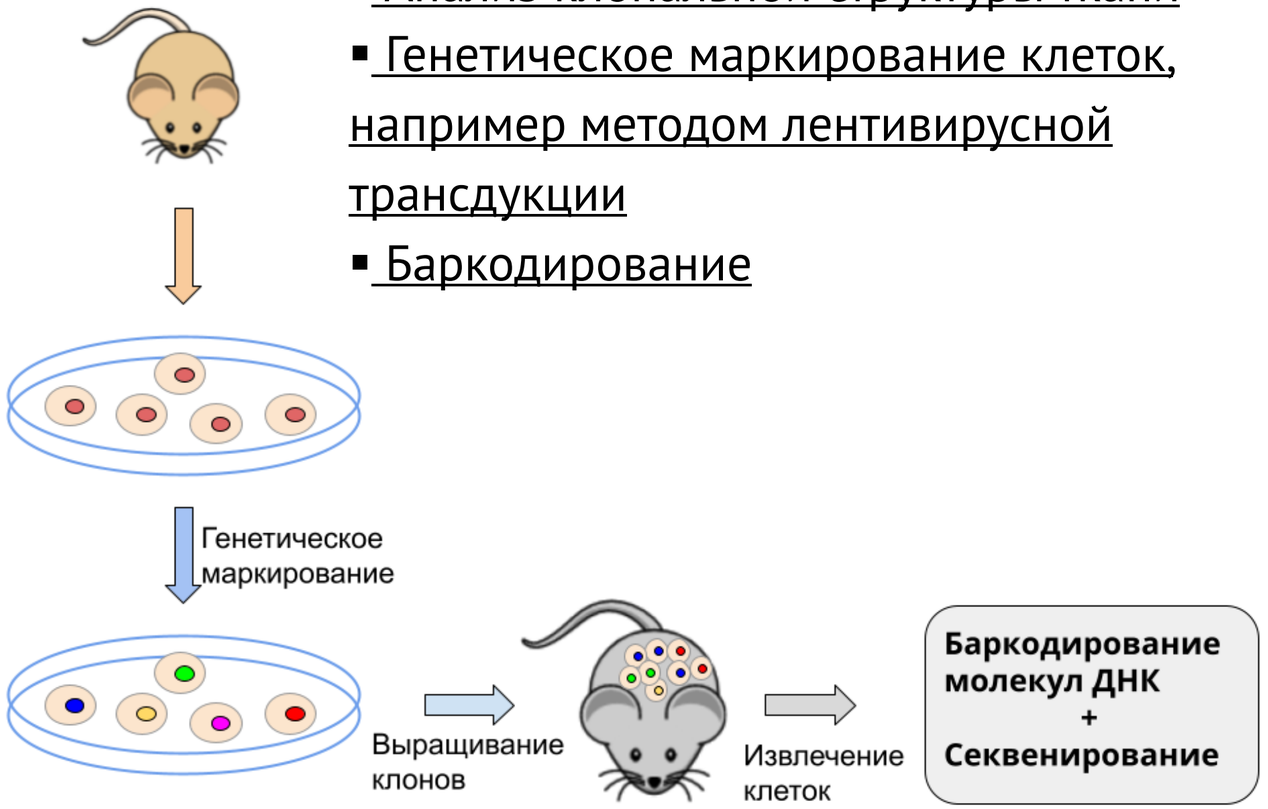
РАЗРАБОТКА МЕТОДОВ АНАЛИЗА ДАННЫХ СЕКВЕНИРОВАНИЯ НОВОГО ПОКОЛЕНИЯ В ЭКСПЕРИМЕНТАХ ПО ЛЕНТИВИРУСНОЙ ТРАНСДУКЦИИ МЕЗЕНХИМАЛЬНЫХ СТВОЛОВЫХ КЛЕТОК

студентка 426 группы **Кожевникова Дарья Дмитриевна,**
Руководитель: д. ф.-м. н. **Шайтан Алексей Константинович**

Москва
2022

Методы секвенирования нового поколения



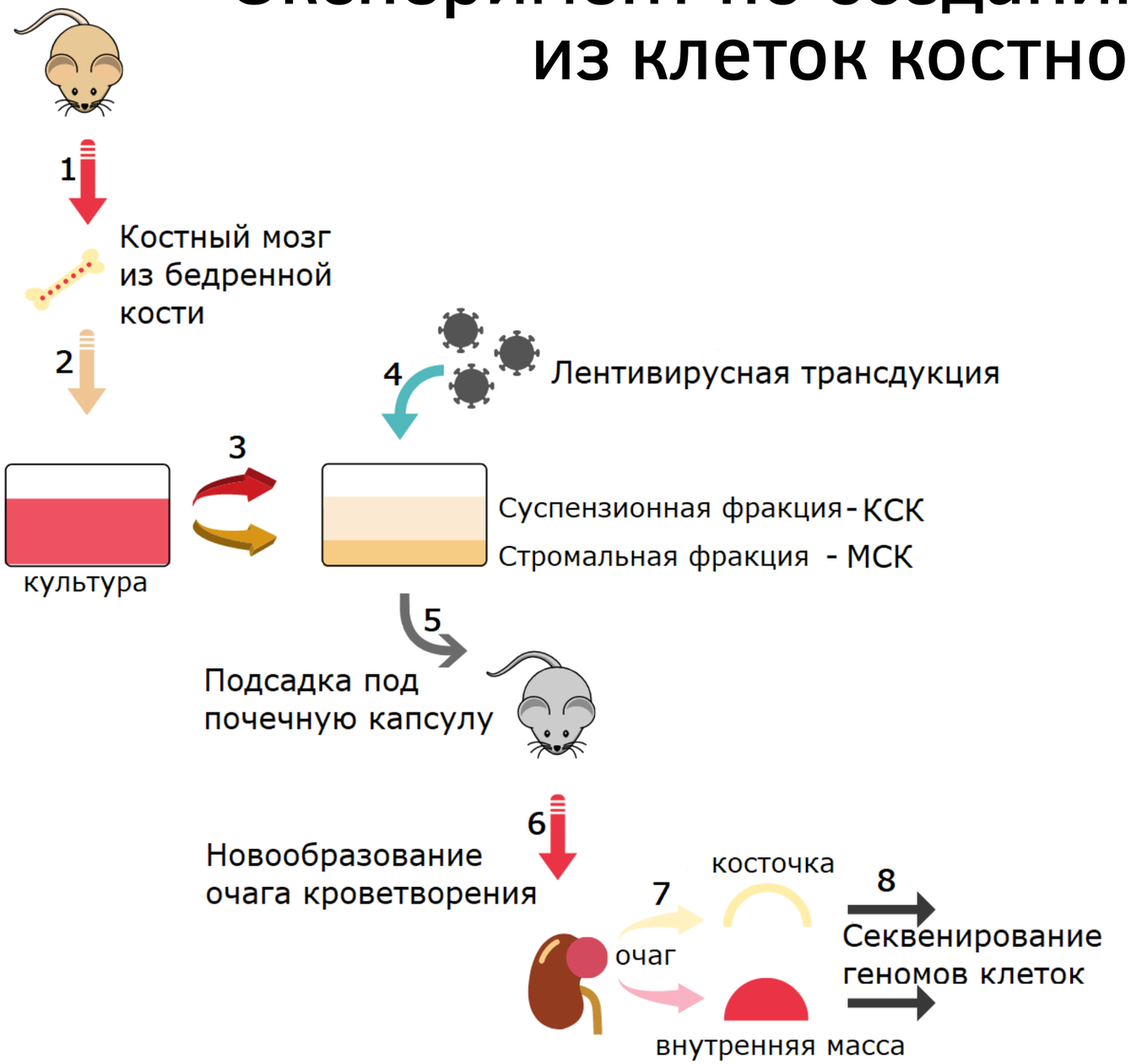


- Анализ клональной структуры ткани
- Генетическое маркирование клеток, например методом лентивирусной трансдукции
- Баркодирование

Платформа **MiSeq** от **Illumina** для коротких прочтений

Dohál, Matúš et al., 2019

Эксперимент по созданию очага кроветворения из клеток костного мозга мыши



2 типа стволовых клеток в костном мозге:

- **МСК** - мезенхимальные стволовые клетки
- КСК - кроветворные стволовые клетки

Эктопический очаг кроветворения:
клетки внутренней массы + косточка

Лентивирусная трансдукция обеспечивает индивидуальное маркирование клеток.
Маркер - сайт интеграции лентивирусного вектора в геном клетки.
Клетки одного клона содержат одинаковый маркер.

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.

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.
- Апробация алгоритма на тестовых данных с повторами.
- Формулирование выводов и предложений по улучшению эксперимента.

Материалы и методы

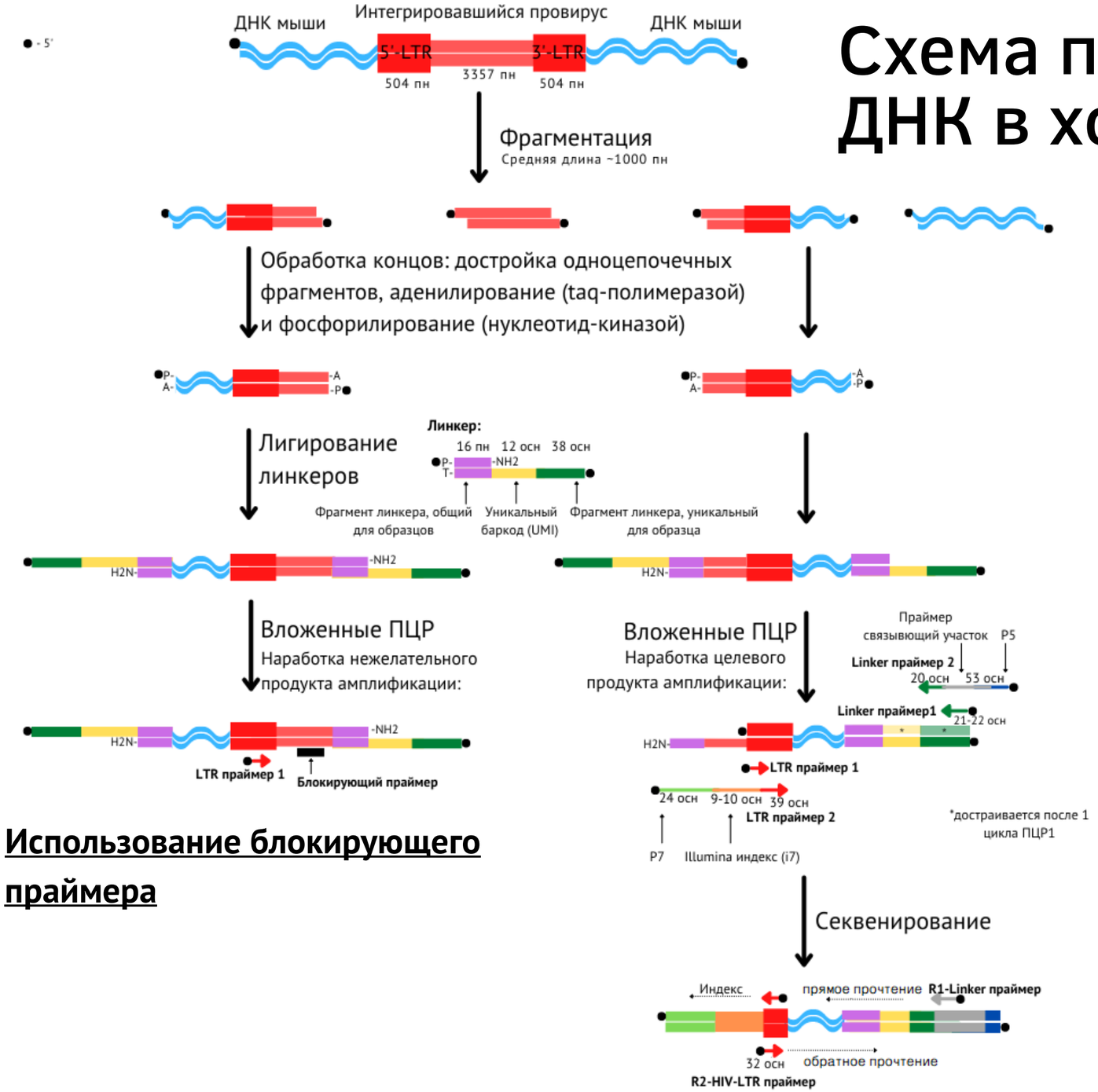
Коллегами из НМИЦ Гематологии были предоставлены:

- Тестовые данные из 15 экспериментов, содержащих по 4 повтора
- Протоколы пробоподготовки

При разработке методов анализа использовались:

- Языки **Python 3 и bash**,
- Система для написания и хранения кода **Jupyter Notebook**,
- Программы и модули для анализа NGS данных:
  - Программа для оценки качества прочтений **FastQC** (Andrews, S. (2010);
  - Программа **BWA** (Li et al., 2009) для картирования прочтений на геном мыши сборки GRCm39 (Church et al., 2011)
  - Программа **Cutadapt** (Martin, 2011) для обрезки прочтений
  - Биоинформатические библиотеки **HTSeq** (Anders et al., 2015), **Bio** (Cock, P.J. et al., 2009), **Levenshtein**.
  - Модуль **Pytexshade** (Integrative Biology Group, 2022) для отображения выравниваний .
  - Библиотеки **Pyvis** (Perrone G. et al., 2020) и **Networkx** (A Hagberg et al., 2008) для анализа и отображения графов.

Схема преобразования молекул ДНК в ходе пробоподготовки



LTR (Long terminal repeat) – длинный концевой повтор – фланкирующий провирусную вставку

Лигирование линкера состоящего из трёх последовательностей:

- **последовательность уникальная для образца**
- **последовательность общая для всех образцов**
- **баркод UMI** (unique molecular identifier) - случайная последовательность из 12 оснований.

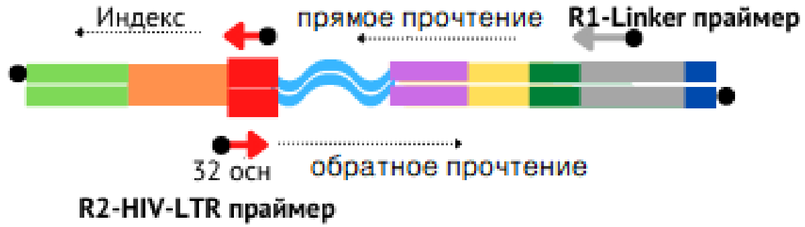
2 этапа ПЦР

Праймеры комплементарны линкеру и LTR

Использование блокирующего праймера

Парноконцевое секвенирование

Предполагаемая структура прочтений



прямое прочтение:

начало: 5' LLLLLLLLLLLLLLLLLLLLNNNNNNNNNNNNCTCCGCTTAAGGGACT ГЕНОМНАЯДНК...

20 осн	12 осн	16 осн	
Фрагмент линкера, общий для образцов	Уникальный баркод (UMI)	Фрагмент линкера, уникальный для образца	ДНК мыши

возможное окончание:

...CTGCTAGA|GATTTTCCACACTGACTAAAAGGGTCTGAGGGATCTCTAXXXXXXXXXXATCTCGTATGCCGTCTTCTGCTTG3'

8 осн	39 осн	10 осн	24 осн
LTR конец (компл цепь)	LTR_праймер_2 LTR конец (компл цепь)	i7 индекс	LTR_праймер_2 P7 конец (компл цепь)

обратное прочтение:

начало: 5' NNNNNNNTCTAGCAGСАЙТИНТЕГРАЦИИ...

7 осн	8 осн	
	LTR конец	ДНК мыши

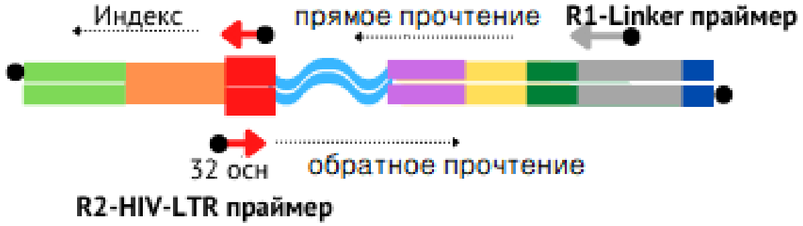
возможное окончание:

...AGTCCCTTAAGCGGAGNNNNNNNNNNNNLLLLLLLLLLLLLLLLLLLLACAATTACCATAGCGTCAGTCCTGGTGTAGATCTCGGTGGTCGCCGTATCATT3'

16 осн	12 осн	20 осн	33 осн	20 осн
Фрагмент линкера, общий для образцов (компл цепь)	Уникальный баркод (UMI) (компл цепь)	Фрагмент линкера, уникальный для образца (компл цепь)	Linker_праймер_2 Праймер связывающий участок (компл цепь)	Linker_праймер_2 P5 конец (компл цепь)

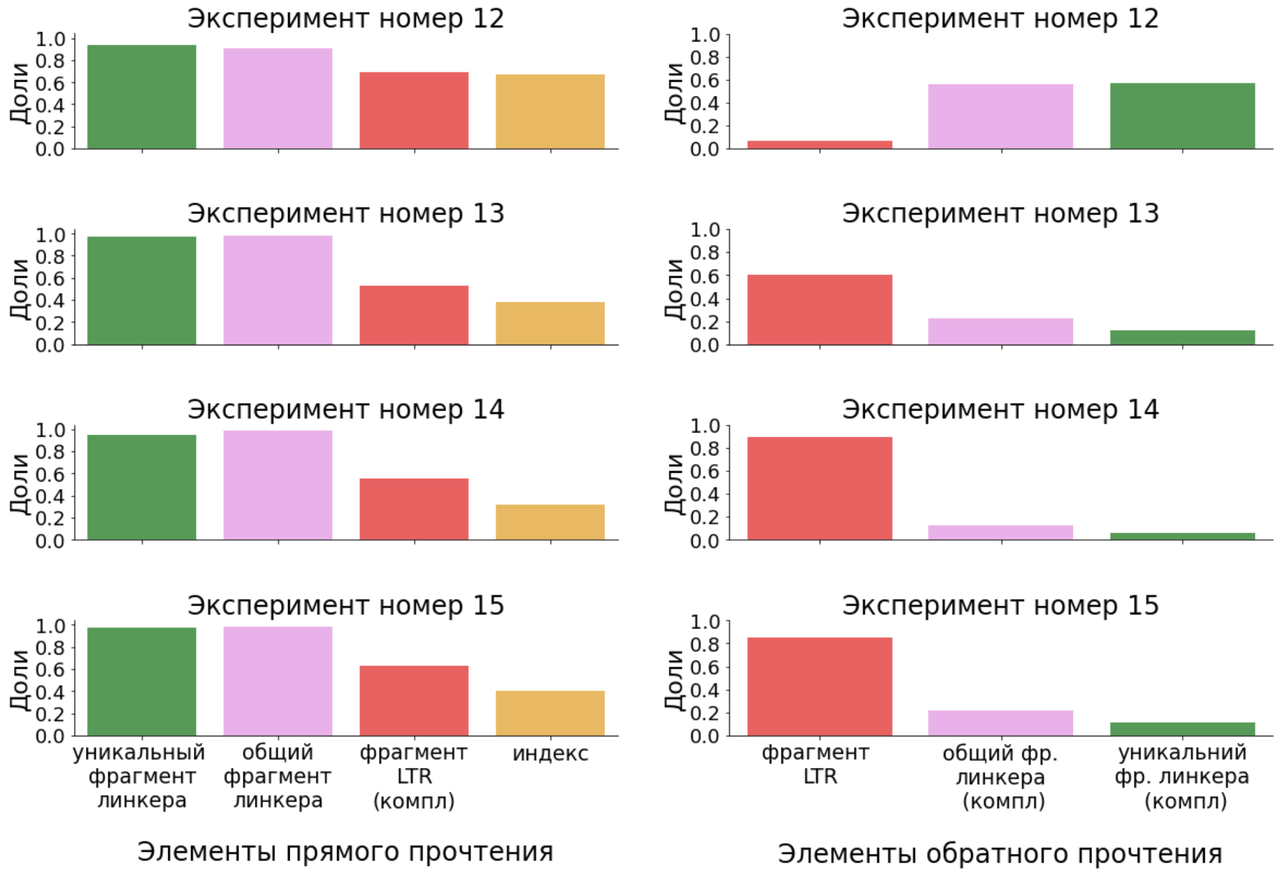
Разработка подходов к оценке качества прочтений

- Оценка качества сырых данных в стандартной программе FastQC
- Разработка метода оценки качества данных на основе соответствия прочтений предполагаемой структуре



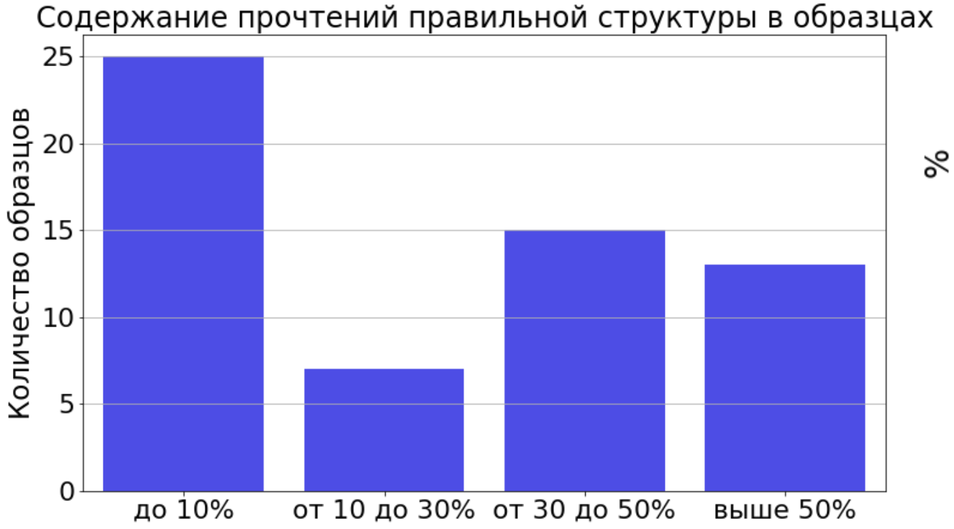
В тестовых данных наблюдаются:

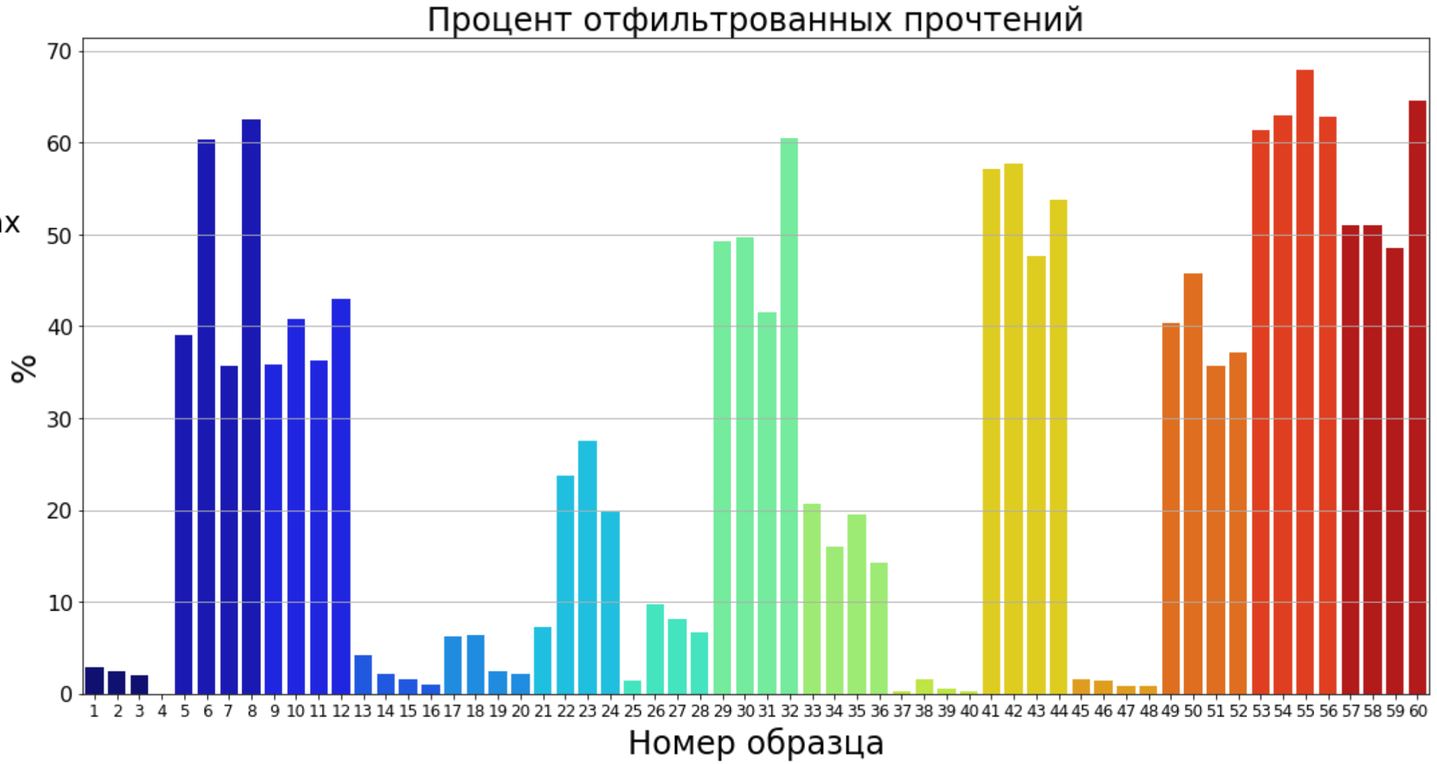
- Кросс-контаминация
- Низкие доли прочтений, содержащих фрагмент LTR



Фильтрация прочтений

- На основе соответствия прочтений ожидаемой структуре
- На основе правильности выравнивания парных прочтений
- Удаление побочных продуктов амплификации





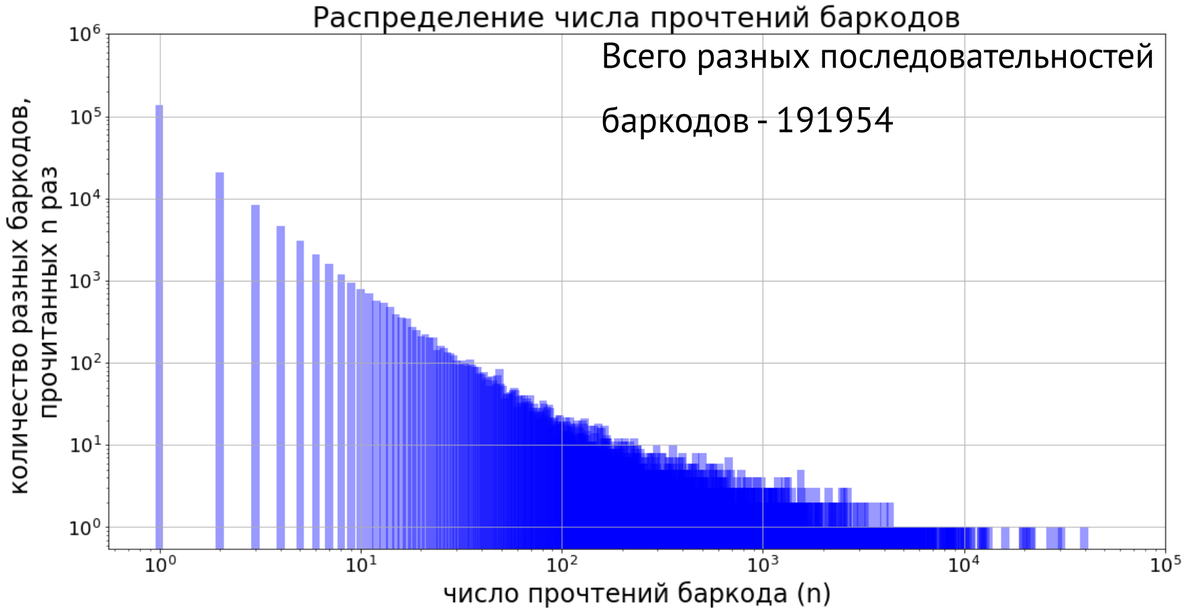
Повторы одного эксперимента покрашены одним цветом

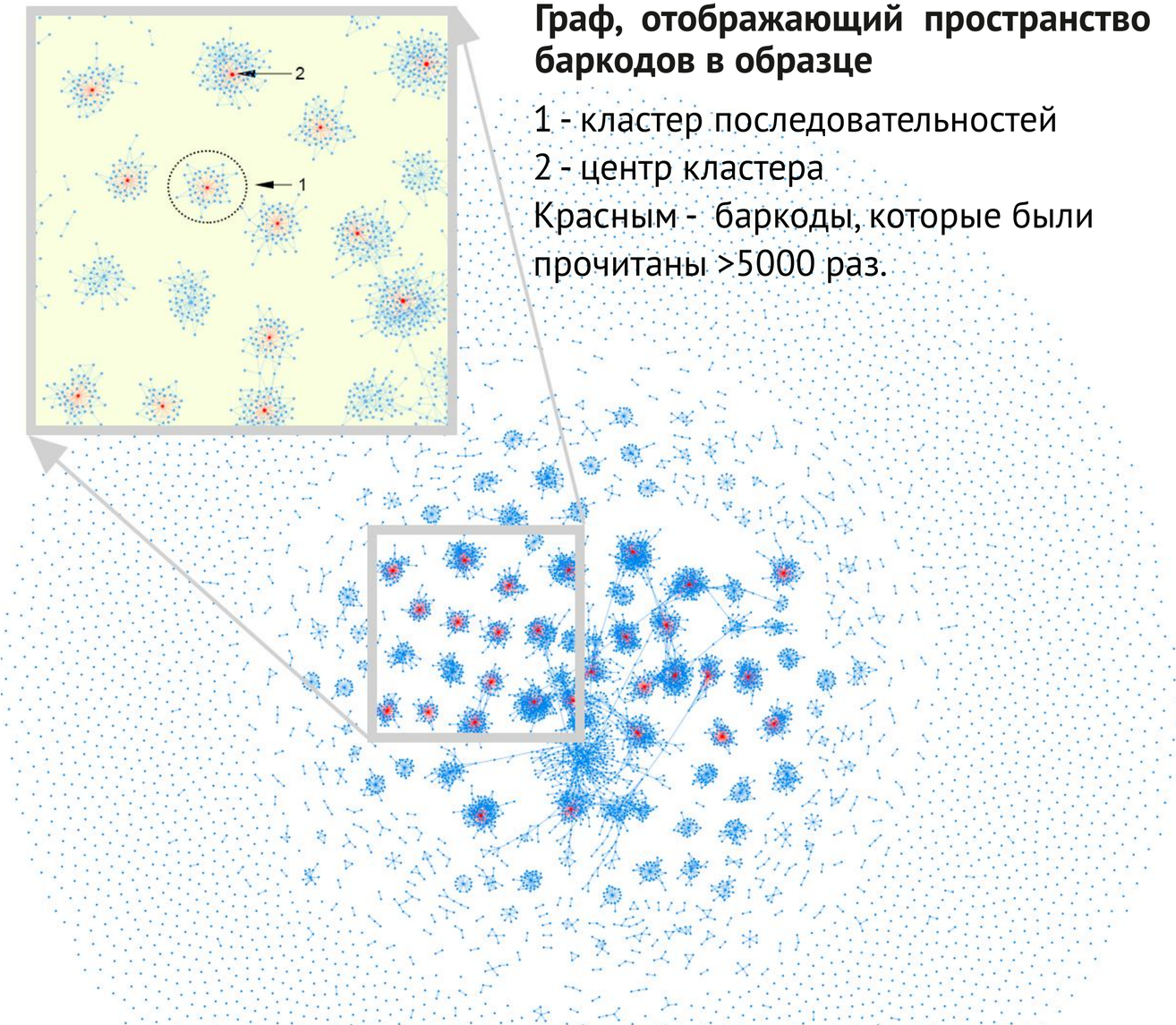
Кластеризация баркодов

Молекулярный баркод маркирует каждую уникальную молекулу на этапе пробоподготовки.

Уникальная последовательность баркода = одна клетка.

Число баркодов = размер клона





Граф, отображающий пространство баркодов в образце

1 - кластер последовательностей

2 - центр кластера

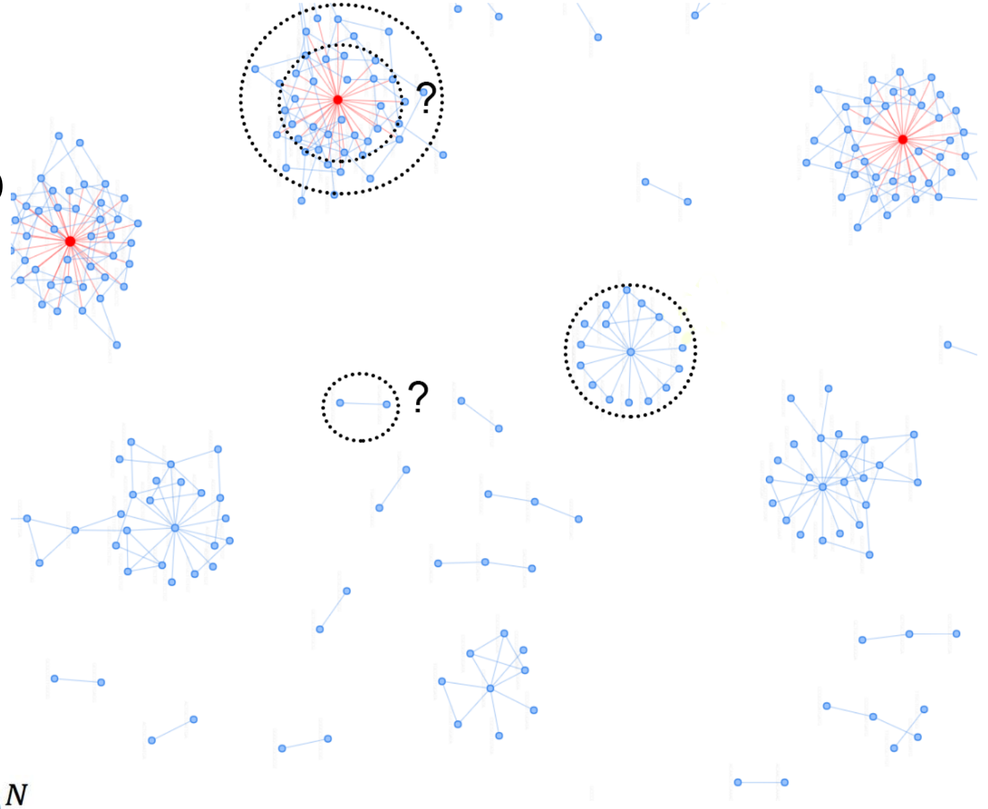
Красным - баркоды, которые были прочитаны >5000 раз.

Кластеризация на основе построения графа:

- вершина - уникальная последовательность баркода,
- ребро - соединяет баркоды с L_dist=1,
- вес вершины - число прочтений последовательности

Алгоритм кластеризации

- Задание направления в графе - ребро направлено в сторону вершины с наибольшим весом (числом прочтений).
- Объединение вершин согласно направлению.
- Здание минимального веса центра кластера



Оценка накопления ошибок при амплификации:

Точности полимеразы error_rate = 0.6*10-4 ошибок/осн/цикл

Вероятность прочтения основания без ошибки: $p = (1 - error\_rate)^N$

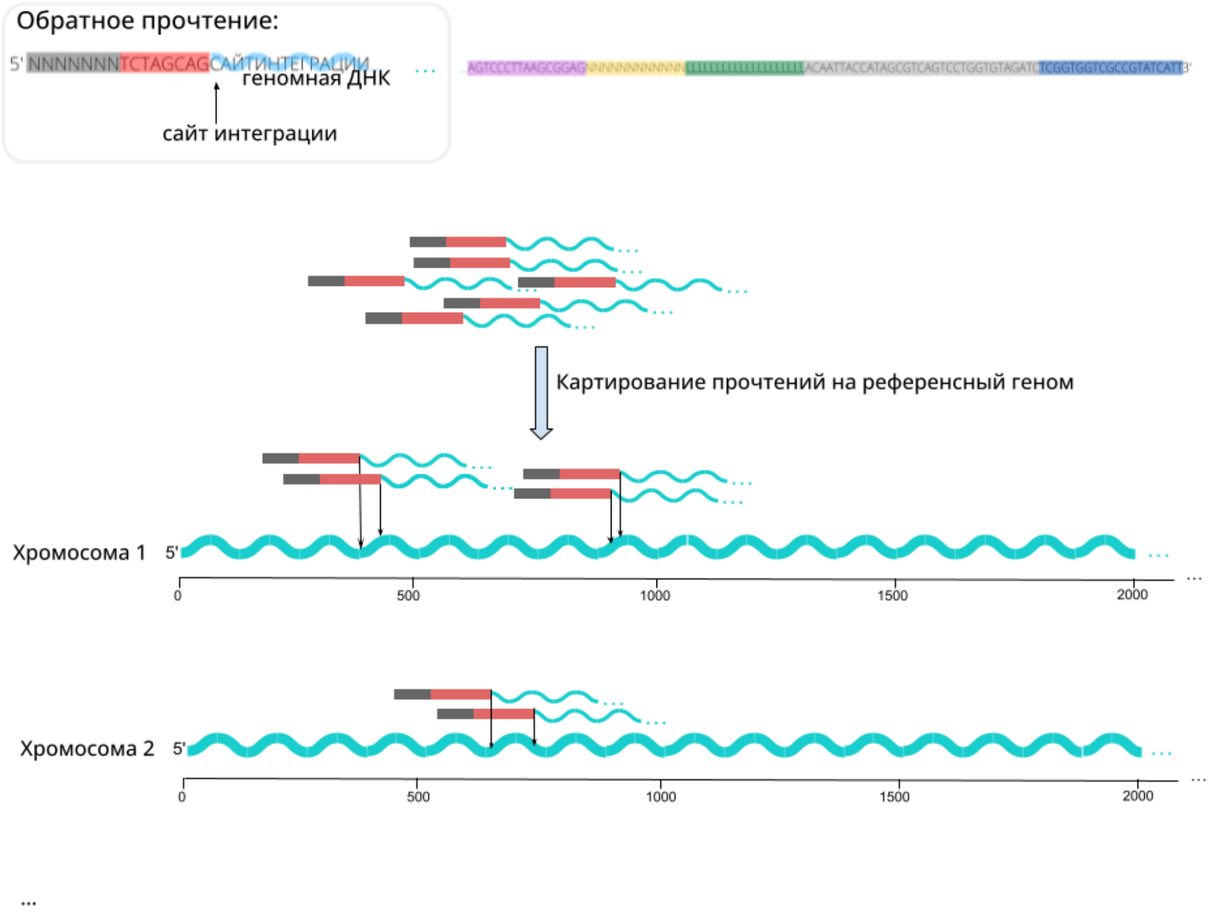
Вероятность появления n ошибок в UMI: $$P(n) = \frac{12!}{\mathrm{n!}\,(12 - \mathrm{n})!} * \mathrm{p}^{12-\mathrm{n}} * (1 - \mathrm{p})^{\mathrm{n}}$$

45 циклов ПЦР на этапе пробоподготовки

Для n=1 и N=45 вероятность P равна 0.0314 - примерно 1 из 30 последовательностей будет содержать хотя бы одну ошибку.

Извлечение сайтов интеграции

- Сайт интеграции – место картирования обратного прочтения на референскный геном
- Анализ равномерности распределения сайтов в технических повторах
- Кластеризация сайтов интеграции - с помощью алгоритма на основе графов
- Удаление шума
- Повторный анализ распределения данных повторах

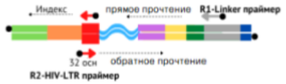


...

Результаты анализа тестовых данных

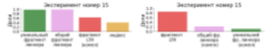
Данные секвенирования

Анализ протоколов пробоподготовки



Оценка качества исходных данных

- Отчеты FastQC
- Диаграммы соответствия структуры прочтений



Картирование прочтений на референсный геном

Фильтрация прочтений

- Отбор прочтений правильной структуры
- Отбор успешно картированные парных прочтений
- Учет наличия побочного продукта амплификации

Построение диаграммы числа прочтений уникальных баркодов

Кластеризация прочтений по баркодам

- Алгоритм на основе графов
- Оценка накопления ошибок в данных

Кластеризованные прочтения

Извлечение сайтов интеграции

- Определение локусов картирования сайтов на геном
- Алгоритм на основе графов

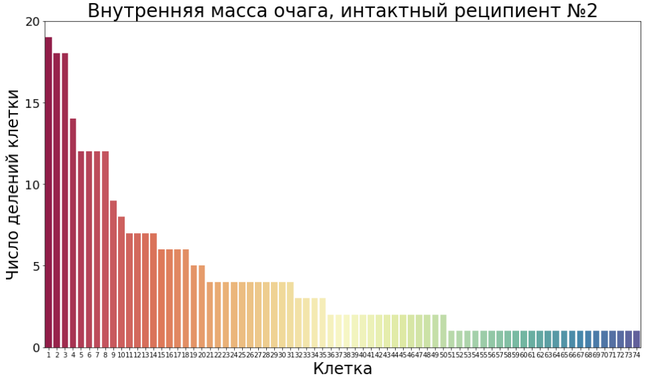
Соотнесение сайтов с баркодами

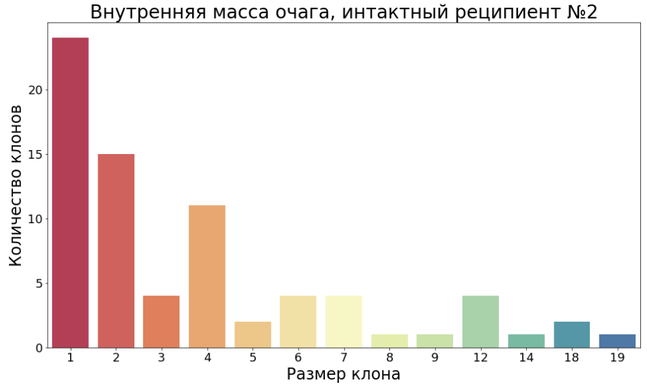
Оценка адекватности результатов

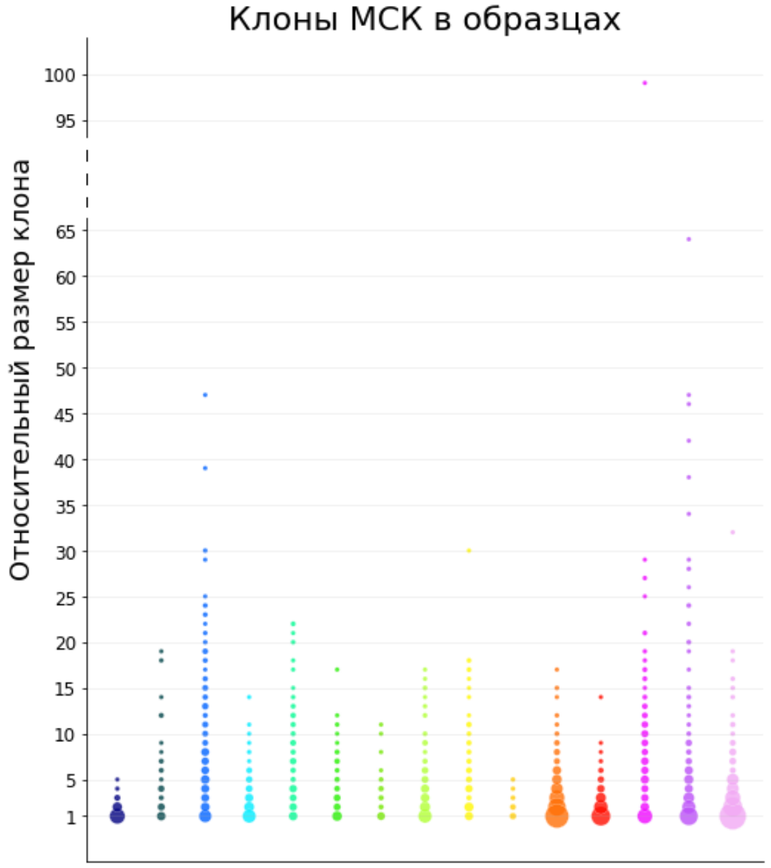
- Оценка равномерности распределения сайтов в повторах

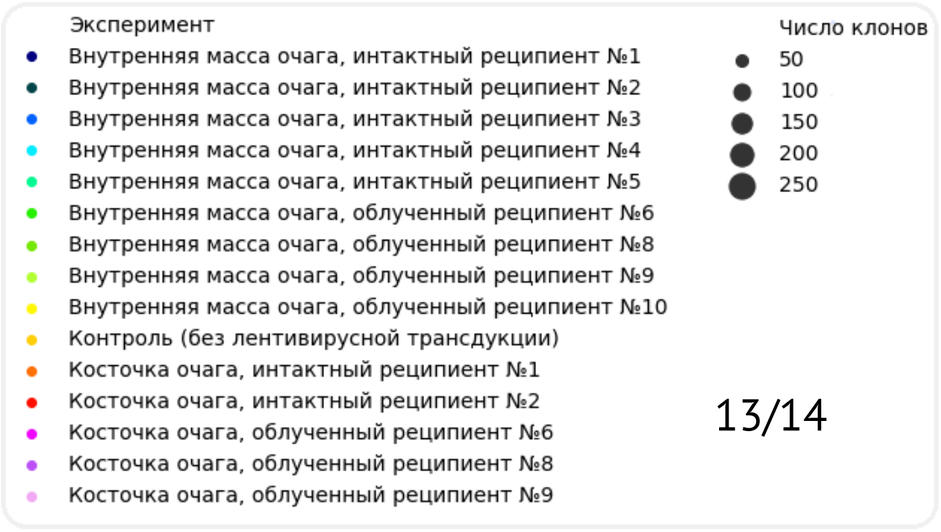
Удаление шума

Результаты - количество клонов и их относительный размер









Выводы

- 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.
- 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.
- Разработанные подходы к анализу данных были апробированы на тестовом наборе экспериментальных данных. Результаты показали низкую (меньше 50%) степень соответствия прочтений предполагаемой структуре в большинстве 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.
- 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.