МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Слушатель Чувакова Любовь Николаевна

Москва, 2023 год.

# Содержание

Введение .......................................................................................................... 4

1. Аналитическая часть .................................................................................. 6

1.1. Постановка задачи ................................................................................... 6

1.2. Описание используемых методов ......................................................... 13

1.2.1 Scanpy…………….................................................................................. 14

1.2.2 Scrublet .................................. 15

1.2.3 PCA ............................................ 15

1.2.4 Метод k-ближайших соседей .............................................................. 16

1.2.5 tSNE .................................................................................. 17

1.2.6 Leiden кластеризация.............................................................................. 18

1.2.7 UMAP ........................................................................... 19

1.2.8 DBSCAN ...........................................................................

1.2.9 RANDOM FOREST ...........................................................................

1.2.10 Нейронная сеть ..................................................................................... 20

2. Практическая часть ................................................................................... 28

2.1. Предобработка данных ..................................................... 28

2.1.1 Удаление пустых клеток ............... 28

2.1.2 Удаление клеток с митохондриальными генами ............................ 29

2.1.3 Удаление дублетов ............. 30

2.1.4 Нормализация и логарифмирование данных

2.2 Анализ данных ................................................................................................. 31

2.2.1 Снижение размерности – Метод главных компонент ............................... 34

2.2.2 Метод k-ближайших соседей .............................................................. 16

2.2.3 tSNE .................................................................................. 17

2.2.4 Leiden кластеризация.............................................................................. 18

2.2.5 DBSCAN ...........................................................................

2.2.6 RANDOM FOREST ...........................................................................

2.2.7 Нейронная сеть ..................................................................................... 20

2.3. Разработка приложения .......................................................................... 46

2.4. Создание удаленного репозитория ........................................................ 46

Заключение ..................................................................................................... 47

Библиографический список .......................................................................... 49

# Введение

Тема данной работы - Кластеризация клеток периферической крови от здорового донора после single-cell RNA sequencing.

Single-cell RNA sequencing (scRNA-seq) – относительно новый метод секвенирования РНК, который активно внедряется в научную практику. Этот метод позволяет получать данные от отдельных клеток различной этиологии, оценить, сколько и каких уже известных клеток было в образце, пути дифференциации клеток, а также возможные функции клеток в их микроокружении. Помимо этого, scRNA-seq позволяет выявить, выделить и проанализировать новые популяции клеток. Чтобы выделить популяции клеток, необходимо их кластеризовать, что уже успешно реализуется с помощью методов машинного обучения.

В данной работе использован датасет, полученный в результате scRNA-seq образца периферической крови, и после его картирования на референсный геном. Исходный датасет подвергнут препроцессингу, в виде удаления пустых значений (пустых капель), капель с погибшими клетками (с высоким содержанием митохондриальных генов), дуплетов, нормализации. Проведен кластерный анализ разными методами машинного обучения, для определения количества кластеров популяций клеток крови, с возможностью дальнейшего более глубокого анализа какого-либо кластера.

1. **Аналитическая часть.**

Полный цикл scRNA-Seq, от пробоподготовки до конечной кластеризации обычно включает следующие этапы: выделение отдельной клетки, экстракция и амплификация нуклеиновых кислот, подготовка библиотеки секвенирования, секвенирование и анализ биоинформатических данных (рис.1)

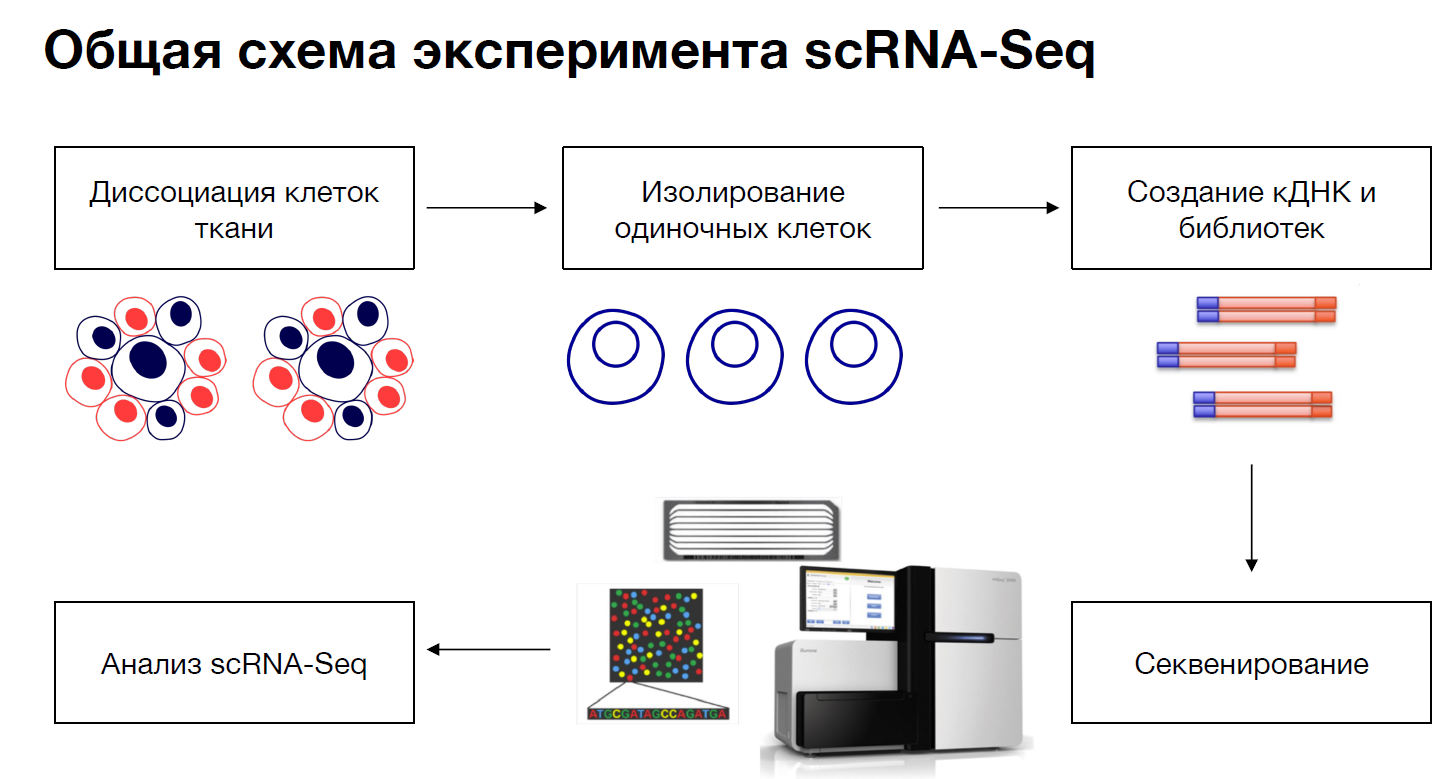
****

Рис 1. Общая схема эксперимента

Первым шагом метода является разделение и инкапсуляция каждой клетки и подготовка библиотеки. Клетки инкапсулируются в гелевые шарики в эмульсии (GEM) (рис. 2).

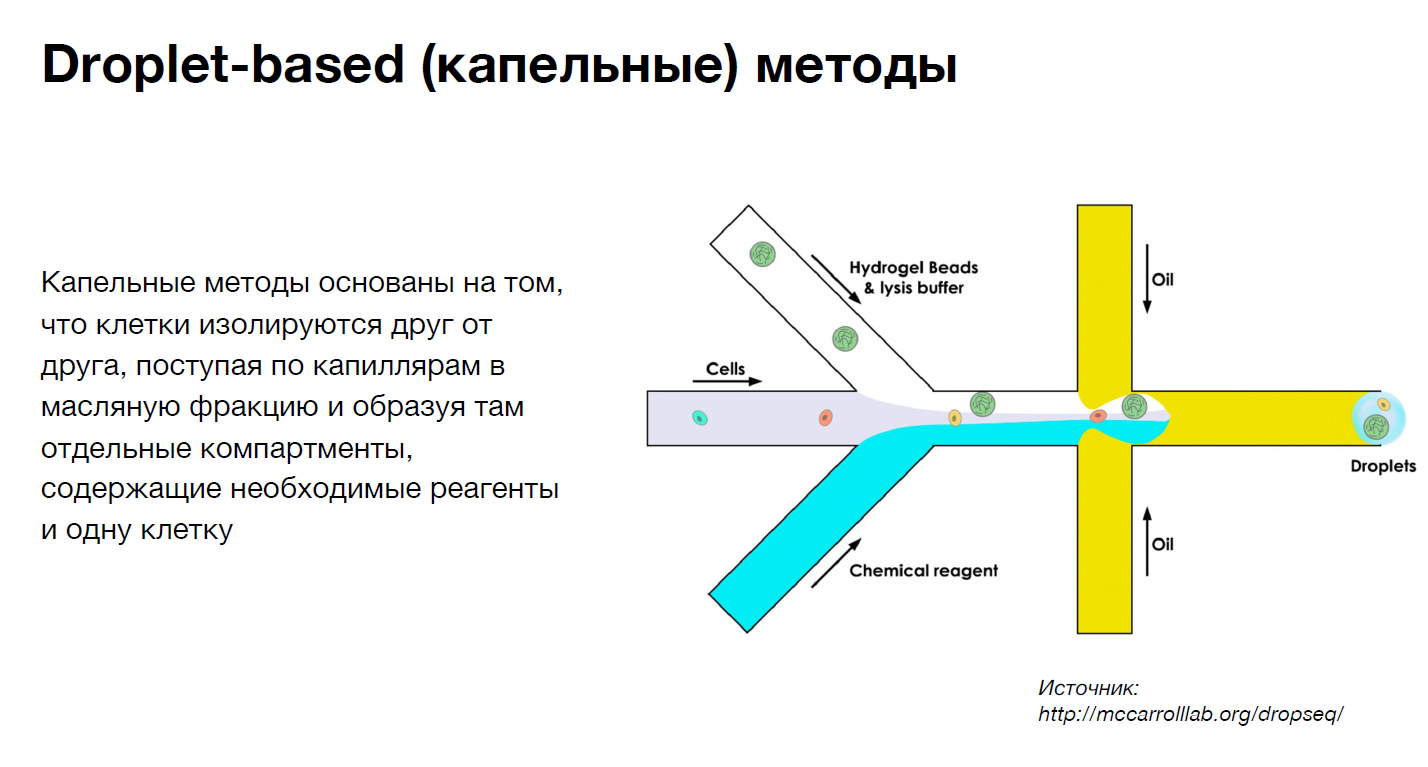
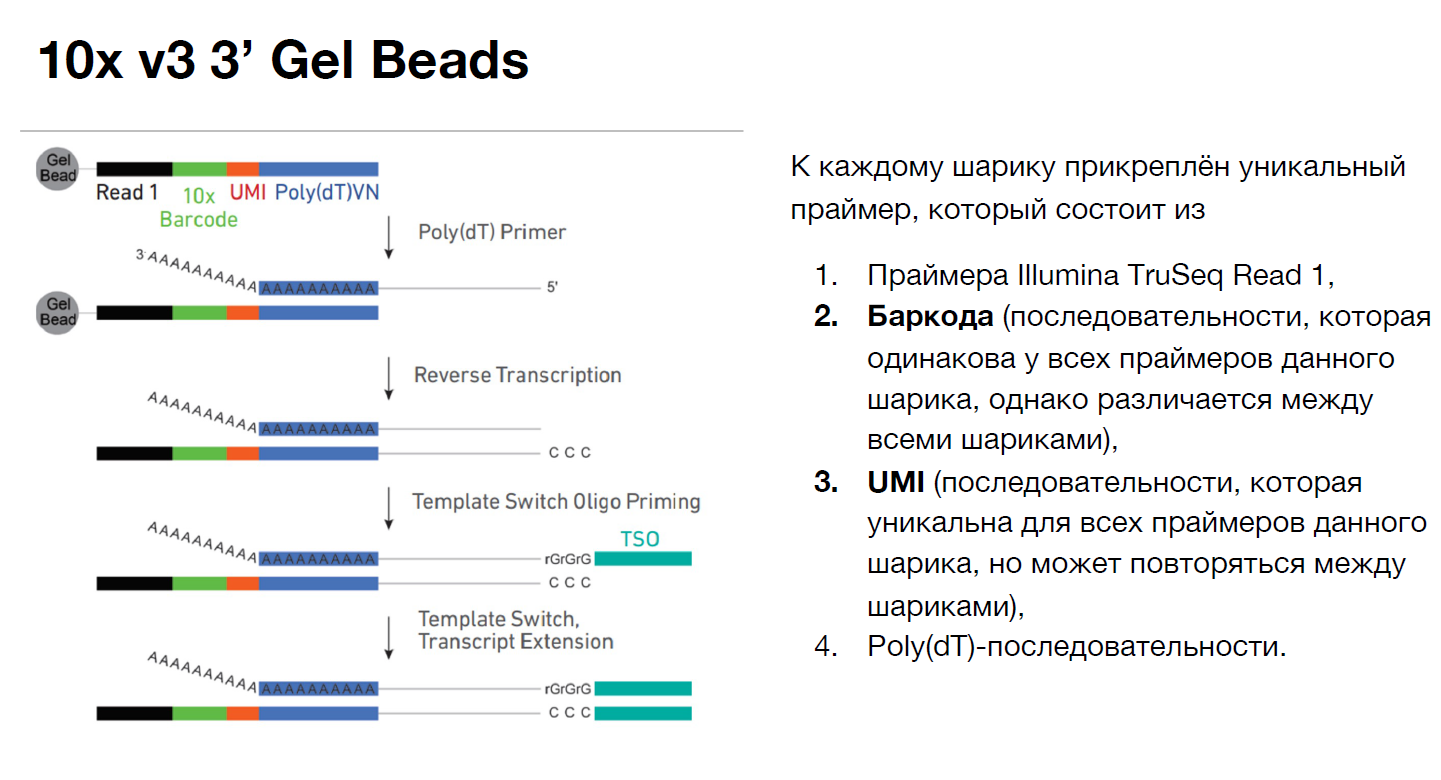
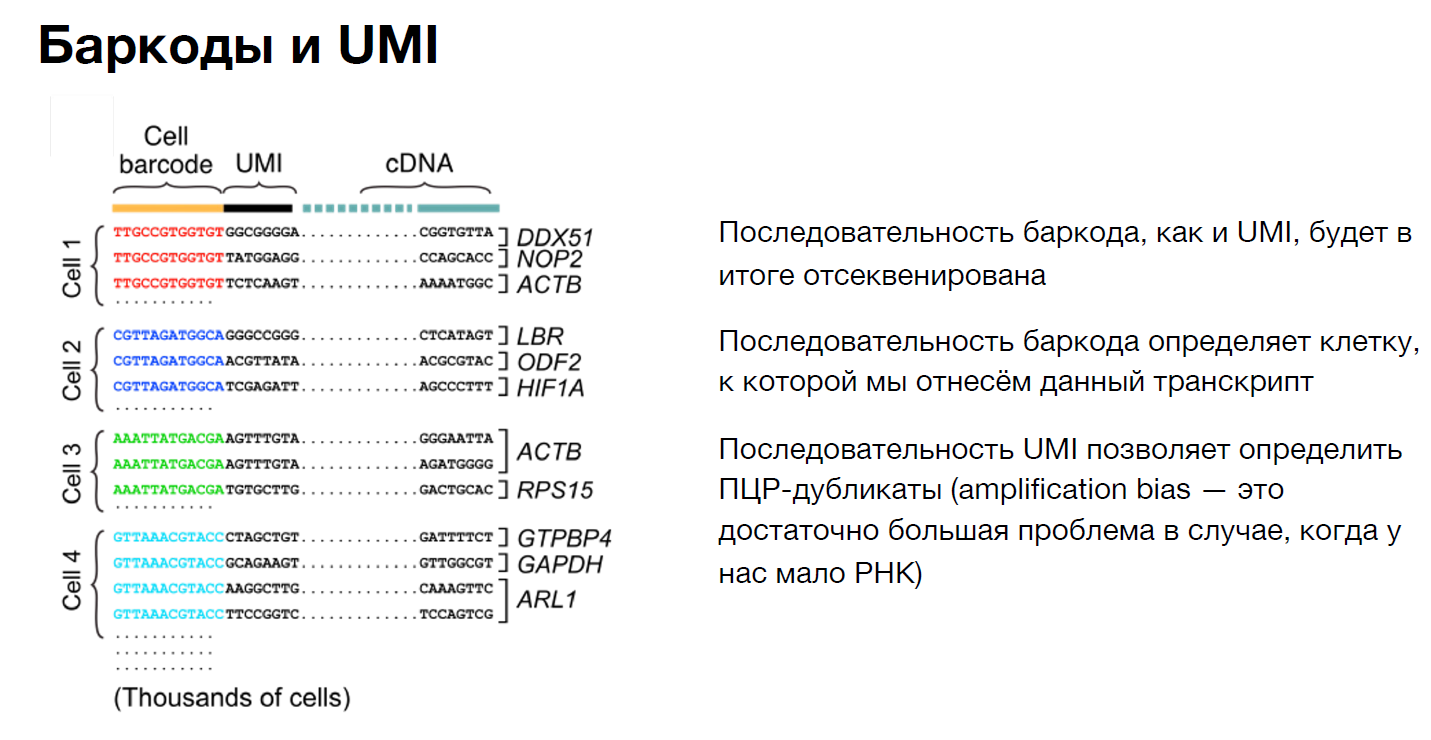
****

Рис. 2. Изоляция диссоциированных клеток в дроплеты

Каждый функциональный GEM содержит одну ячейку, одну гелевую бусину и реагенты RT. На гелевой бусине связываются олигнонуклеотиды, состоящие из 4 отдельных частей: праймер для ПЦР (необходим для секвенирования); 10X олигонуклеотиды со штрих-кодом; последовательность уникального молекулярного идентификатора (UMI); Последовательность PolydT (которая позволяет захватывать полиаденилированные молекулы мРНК) (рис. 3).

****

****

Внутри каждой реакционной везикулы GEM одна клетка лизируется и подвергается обратной транскрипции. кДНК из одной и той же клетки идентифицируются благодаря общему штрих-коду 10X. Кроме того, количество UMI выражает уровень экспрессии генов, и его анализ позволяет обнаруживать высоковариабельные гены. Эти данные часто используются либо для классификации клеточных фенотипов, либо для идентификации новых субпопуляций.

Завершающим этапом платформы является секвенирование. Созданные библиотеки можно напрямую использовать для секвенирования всего транскриптома одной клетки или рабочих процессов целевого секвенирования.

В дальнейшем осуществляется непосредственно биоинформатический анализ данных. Из секвенатора выгружаются данные в формате .fastq. Выравнивание и подсчет экспрессии генов на клетку выполняется программой. Золотым стандартом является программа Cell Ranger. Ее достоинством является широкий спектр функций, начиная от автоматичского определения координат координаты баркода / UMI в прочтениях, работает с разными модификациями scRNA-Seq, с помощью флагов можно провести препроцессинг данных. Однако, он очень требовательный к ресурсам (1 Тб дискового пространства, 128 Гб RAM, 16 ядер) и очень долго работает (один образец может рассчитываться 12 часов).

В простейшем случае аутпут содержит 4 файла:

1. raw\_feature\_bc\_matrix.tar.gz — матрица со всеми “клетками” из датасета (в дальнейшем и будет анализироваться)

a. barcodes.tsv.gz — названия клеток (баркоды)

b. features.tsv.gz — названия и id генов

c. matrix.mtx.gz — непосредственно матрица экспрессии в sparce-виде

2. filtered\_feature\_bc\_matrix.tar.gz — то же, что и пункт 1, только с уже

отфильтрованными клетками (Cell Ranger фильтрует очень неплохо)

a. barcodes.tsv.gz

b. features.tsv.gz

c. matrix.mtx.gz

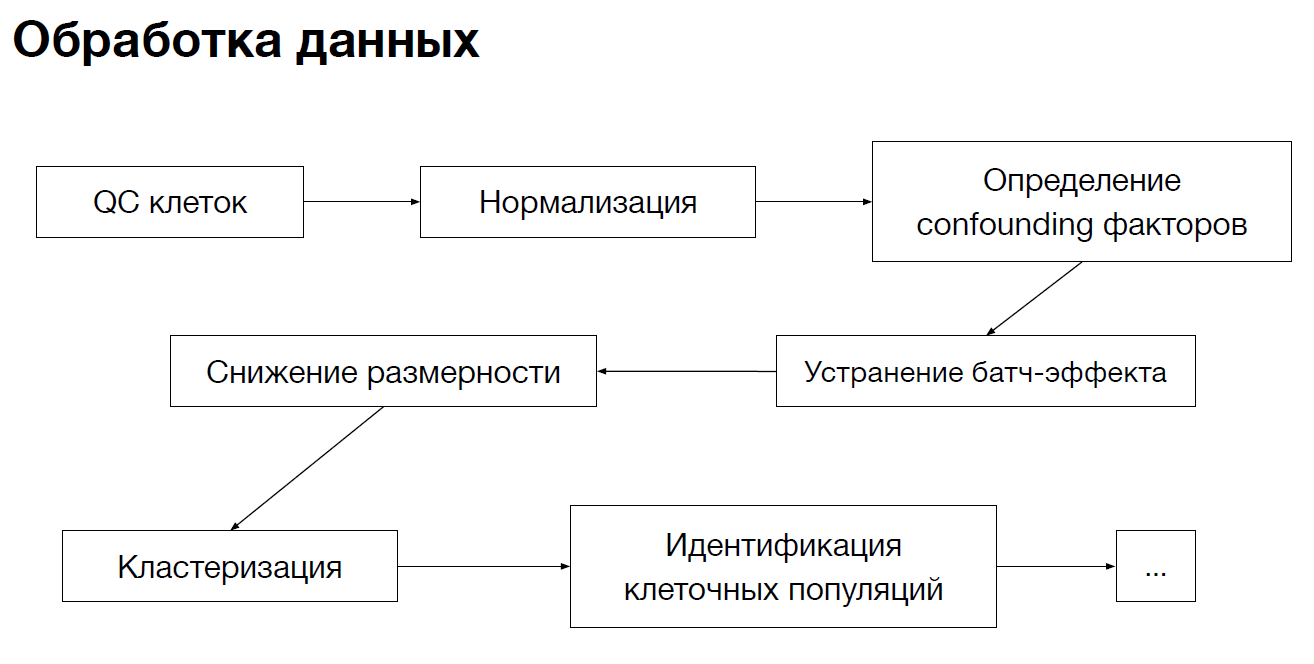
3. metrics\_summary.csv — таблица с основными метриками

4. web\_summary.html — графический веб-отчёт о качестве выравнивания и т. п.

* 1. Постановка задачи

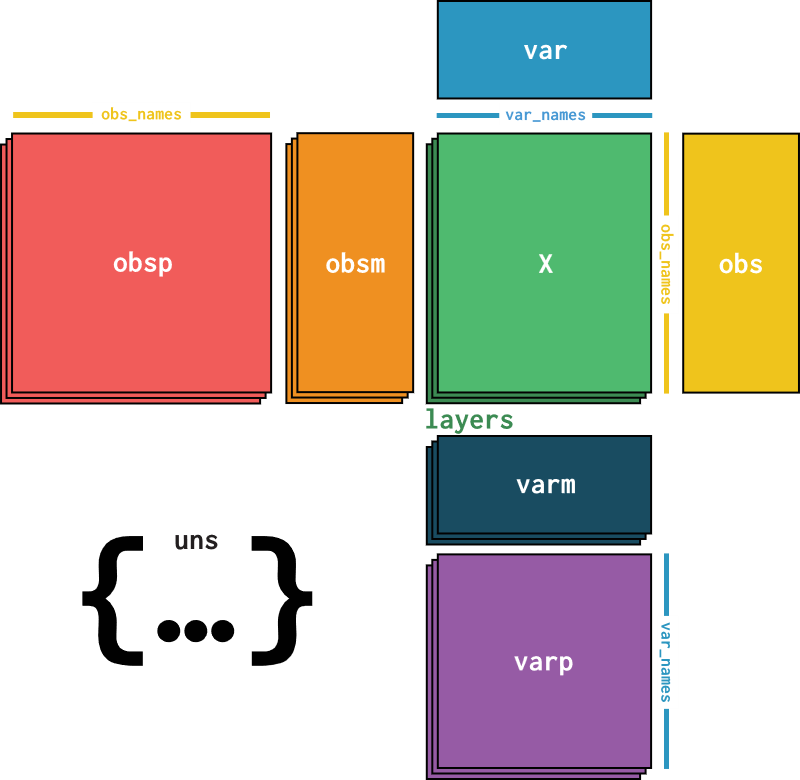
В папке outs содержатся файлы после отработки CellRanger. Датасет raw\_feature\_bc\_matrix.tar.gz находится на google disk, так как Githab не позволяет подгружать большие датасеты с весом, более 25 мб непосредственно в репозиторий и более 100 мб через командную строку.

Для того, чтобы кластеризовать данные после scRNA-Seq и в дальнейшем выбрать и проанализировать интересующие кластеры клеток периферической крови, необходимо сначала провести препроцессинг данных, который включает в себя QC клеток (удаление пустых капель (аналог удаления строк с NA), капель с погибшими клетками (где большое количество митохондриальных генов), дуплетов, нормализацию данных, возможные дополнительные этапы и затем непосредственно кластеризация клеточных популяций (рис.4).

****

# Данные представлены в формате AnnData – Annotated data – это так называемый контейнерный формат данных, где данные взаимосвязаны между собой, но представлены в нескольких таблицах (рис.5).

***Class*anndata.AnnData(*X=None*, *obs=None*, *var=None*, *uns=None*, *obsm=None*, *varm=None*, *layers=None*, *raw=None*, *dtype=None*, *shape=None*, *filename=None*, *filemode=None*, *asview=False*, *\**, *obsp=None*, *varp=None*, *oidx=None*, *vidx=None*)**



[AnnData](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.html#anndata.AnnData) хранит в себе матрицу [X](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.X.html#anndata.AnnData.X) вместе с аннотациями исследований [obs](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.obs.html#anndata.AnnData.obs) ([obsm](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.obsm.html#anndata.AnnData.obsm), [obsp](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.obsp.html#anndata.AnnData.obsp)), вариативными данными [var](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.var.html#anndata.AnnData.var) ([varm](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.varm.html#anndata.AnnData.varm), [varp](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.varp.html#anndata.AnnData.varp)), и неструктурированными аннотациями [uns](https://anndata.readthedocs.io/en/latest/generated/anndata.AnnData.uns.html#anndata.AnnData.uns) (https://anndata.readthedocs.io/en/latest/#). Данный формат возможно перевести в data frame, однако часть данных будет потеряна, что не всегда может положительно сказаться на кластерном анализе.

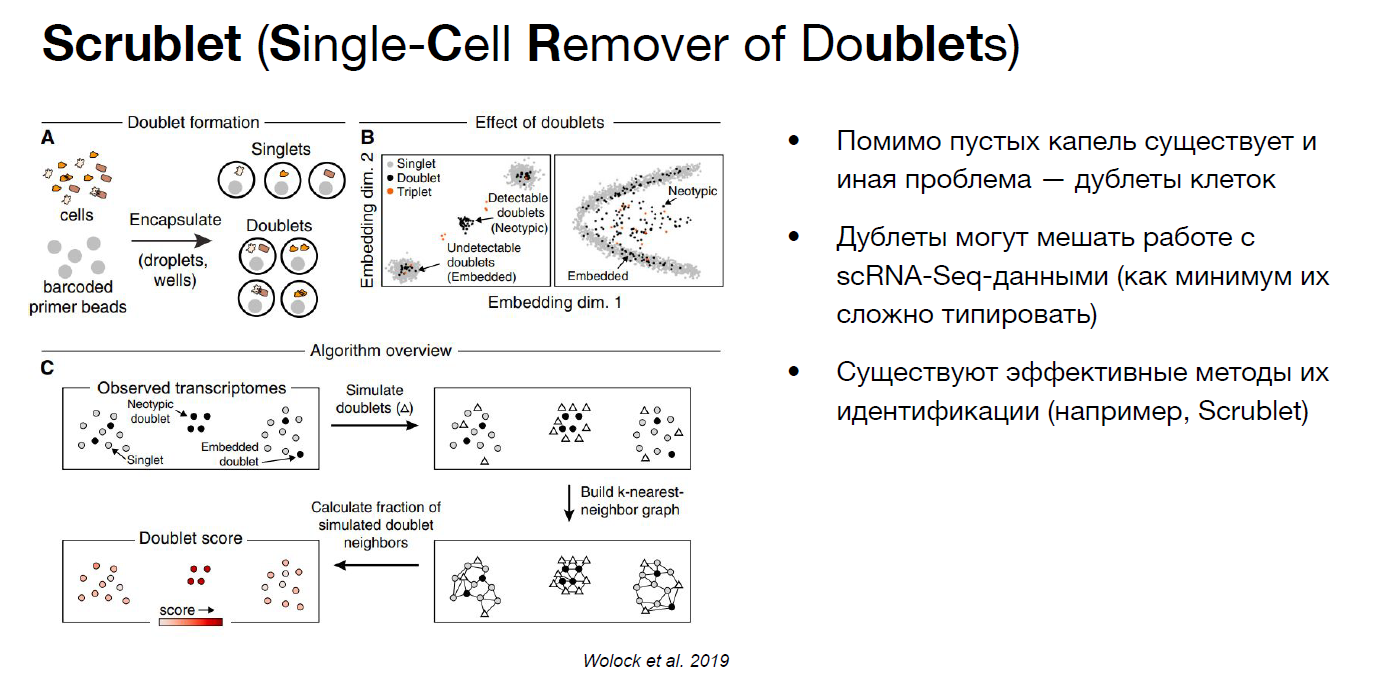
Также дается характеристика датасета – число входных и выходных переменных (факторов), объем выборки, характеристика выборки с точки зрения ее особенностей (выбросы, пропуски и т.д.).

* 1. Описание используемых методов
     1. Scanpy – Single-Cell Analysis in Python

Scanpy — это масштабируемый инструментарий для анализа данных об экспрессии генов в отдельных клетках, созданный совместно с anndata. Он включает в себя предварительную обработку, визуализацию, кластеризацию, вывод траектории и тестирование дифференциальных выражений. Реализация на основе Python эффективно работает с наборами данных, содержащими более миллиона ячеек (<https://scanpy-tutorials.readthedocs.io/en/latest/index.html>**)**

1.2.2. Scrublet

Scrublet используется для удаления капель, где оказалось 2 и более клеток. С такой ячейки приходит смешанная информация, которую затем анализировать не представляется возможным, поэтому данной утилитой строки с дублетами удаляются. Подход достаточно простой, сначала на строках генерируются фантомные дублеты и с ними сравниваются другие строки, и где оказывается похожее распределение по баркодам и UMI, те строки помечаются как дублеты и в дальнейшем из датасета удаляются.

****

1.2.3. PCA

Метод главных компонент (PCA) - один из основных способов уменьшить [размерность](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%B0) данных, потеряв наименьшее количество [информации](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D0%B8%D1%8F), а также снизить шум в данных, который также может помешать адекватной кластеризации данных.

1.2.4 Метод k-ближайших соседей

Еще один метод классификации, который адаптирован для регрессии – метод k-ближайших соседей (k Nearest Neighbors). Суть метода проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься.

Для реализации метода необходима метрика расстояния между объектами. Используется, например, эвклидово расстояние для количественных признаков или расстояние Хэмминга для категориальных.

1.2.5 tSNE

tSNE - это алгоритм [машинного обучения](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5) для [визуализации](https://ru.wikipedia.org/wiki/%D0%92%D0%B8%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85), является техникой [нелинейного снижения размерности](https://ru.wikipedia.org/w/index.php?title=%D0%9D%D0%B5%D0%BB%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D0%B5_%D1%81%D0%BD%D0%B8%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5_%D1%80%D0%B0%D0%B7%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D1%81%D1%82%D0%B8&action=edit&redlink=1), хорошо подходящей для вложения данных высокой размерности для визуализации в пространство низкой размерности (двух- или трехмерное). В частности, метод моделирует каждый объект высокой размерности двух- или трёхмерной точкой таким образом, что похожие объекты моделируются близко расположенными точками, а непохожие точки моделируются с большой вероятностью точками, далеко друг от друга отстоящими.

1.2.6 Leiden

Алгоритм Лейдена использует идею ускорения локального перемещения узлов и идею перемещения узлов к случайным соседям (https://www.nature.com/articles/s41598-019-41695-z)

1.2.7. UMAP

Uniform Manifold Approximation and Projection (UMAP) — алгоритм [машинного обучения](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), выполняющий [нелинейное снижение размерности](https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction). похожий на [t-SNE](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%BE%D1%85%D0%B0%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%B2%D0%BB%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81%D0%BE%D1%81%D0%B5%D0%B4%D0%B5%D0%B9_%D1%81_t-%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5%D0%BC), но с более сильным математическим обоснованием.

При снижении размерности UMAP сначала выполняет построение [взвешенного графа](https://ru.wikipedia.org/wiki/%D0%93%D0%BB%D0%BE%D1%81%D1%81%D0%B0%D1%80%D0%B8%D0%B9_%D1%82%D0%B5%D0%BE%D1%80%D0%B8%D0%B8_%D0%B3%D1%80%D0%B0%D1%84%D0%BE%D0%B2#%D0%B2%D0%B7%D0%B2%D0%B5%D1%88%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B3%D1%80%D0%B0%D1%84), соединяя ребрами только те объекты, которые являются ближайшими соседями. Множество из ребер графа — это [нечёткое множество](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D1%87%D1%91%D1%82%D0%BA%D0%BE%D0%B5_%D0%BC%D0%BD%D0%BE%D0%B6%D0%B5%D1%81%D1%82%D0%B2%D0%BE) с [функцией принадлежности](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%D0%BF%D1%80%D0%B8%D0%BD%D0%B0%D0%B4%D0%BB%D0%B5%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D0%B8), она определяется как вероятность существования ребра между двумя вершинами. Затем алгоритм создает граф в низкоразмерном пространстве и приближает его к исходному.

1.2.8 DBSCAN

Основанная на плотности пространственная кластеризация для приложений с шумами ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) Density-based spatial clustering of applications with noise, DBSCAN) — это алгоритм [кластеризации данных](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), основанной на плотности — если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

1.2.9 Случайный лес (RandomForest)

Случайный лес (RandomForest) - представитель ансамблевых методов.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

Преимущества случайного леса:

* высокая точность предсказания;
* редко переобучается;
* практически не чувствителен к выбросам в данных;
* одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков;
* высокая параллелизуемость и масштабируемость.

Из недостатков можно отметить, что его построение занимает больше времени. Так же теряется интерпретируемость.

1.2.10 Нейронная сеть

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Структура нейронной сети пришла в мир программирования из биологии. Вычислительная единица нейронной сети — нейрон или персептрон.

У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа.

Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения.

Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется:

* входной слой — его размер соответствует входным параметрам;
* скрытые слои — их количество и размерность определяем специалист;
* выходной слой — его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением.

Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась.

Для обновления весов в модели используются различные оптимизаторы.

Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения.

Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении.

2. Практическая часть

2.1. Предобработка данных ..................................................... 28

2.1.1 Удаление пустых клеток ............... 28

2.1.2 Удаление погибших клеток с митохондриальными генами ............................ 29

2.1.3 Удаление дублетов ............. 30

2.1.4 Нормализация и логарифмирование данных

2.2 Анализ данных ................................................................................................. 31

2.2.1 Снижение размерности – Метод главных компонент ............................... 34

2.2.2 Метод k-ближайших соседей .............................................................. 16

2.2.3 tSNE .................................................................................. 17

2.2.4 Leiden кластеризация.............................................................................. 18

2.2.5 DBSCAN ...........................................................................

2.2.6 RANDOM FOREST ...........................................................................

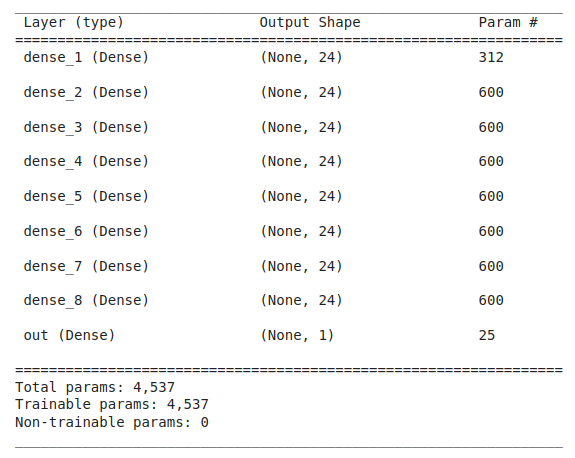
2.2.7 Нейронная сеть ..................................................................................... 20

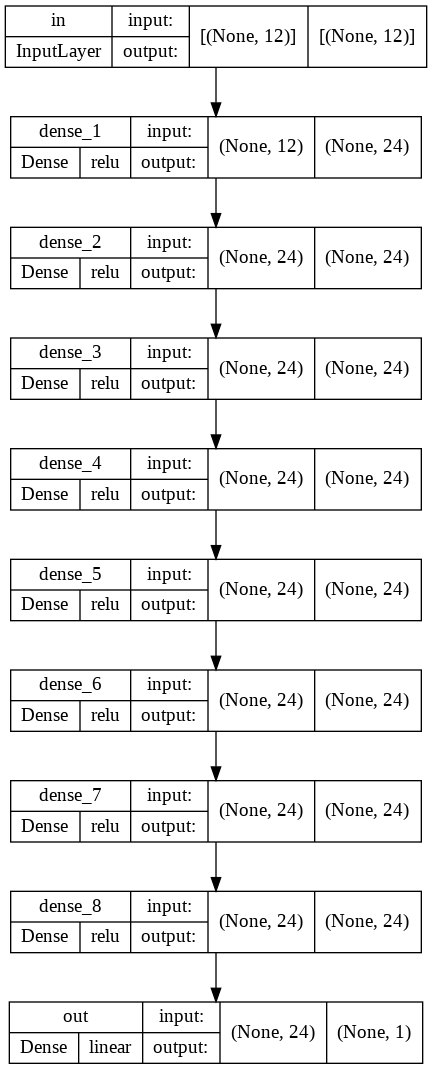
## 2.4.2 Нейросеть из библиотеки tensorflow

Строю нейронную сеть с помощью класса keras.Sequential со следующими параметрами:

* входной слой для 12 признаков;
* выходной слой для 1 признака;
* скрытых слоев: 8;
* нейронов на каждом скрытом слое: 24;
* активационная функция скрытых слоев: relu;
* оптимизатор: Adam;
* loss-функция: MeanAbsolutePercentageError.

Архитектура нейросети приведена на рисунках 28 и 29.

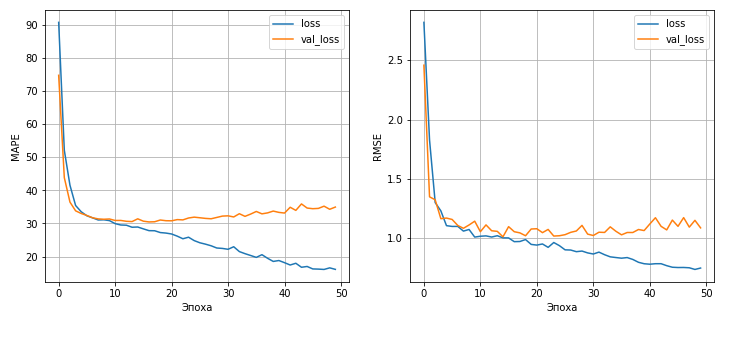
Рисунок 28 — Архитектура нейросети в виде summary

Рисунок 29 — Архитектура нейросети в виде графа

Запускаю обучение нейросети со следующими параметрами:

* пропорция разбиения данных на тестовые и валидационные: 30%;
* количество эпох: 50.
* раннюю остановку не использую.

График обучения приведен на рисунке 30, ошибка — в таблице 2.

Рисунок 30 — График обучения нейросети

Видно, что примерно до 8 эпохи обучение шло хорошо, а потом сеть начала переобучаться. Значение loss на тестовых выборках продолжило уменьшаться, а на валидационной начало расти.

Одним из способов борьбы с переобучением может быть ранняя остановка обучения, если val\_loss начинает расти. Для этого в tensorflow используются callbacks. Попробую взять нейросеть с той же архитектурой и запустить обучение с ранней остановкой. График обучения приведен на рисунке 31, а ошибка — в таблице 2. Очевидно, что решение проблемы переобучения повышает точность модели на новых данных.

Еще одним методом борьбы с переобучением является добавление Dropout-слоев. Построим модель аналогичной архитектуры, только после каждого скрытого слоя добавим слой Dropout с параметром 0.05. Такой слои слои выключат 5% случайных нейронов на каждом слое.

График обучения приведен на рисунке 32, а ошибка — в таблице 2. Видно, что Dropout-слои справились с переобучением.

Использование ранней остановки сокращает время на обучение модели, а использование Dropout увеличивает. Но уменьшается риск, что мы остановились слишком рано.

Визуализация результатов показывает, что нейросеть из библиотеки tensorflow старалась подстроиться к данным. Выглядят результаты «похоже», но метрики разочаровывают. Лучшая обобщающая спрособность и меньшие значения ошибок на тестовом множестве оказались у нейросети, обученной с ранней остановкой. Но и она предсказывает гораздо хуже базовой модели.

## 2.6. Разработка приложения

Разработка приложения, которое будет проводить весь необходимый анализ, представлена в виде файла app.py, однако, так и не была реализована на платформе. Причин несколько: 1) при запуске CellRanger возможно с помощью флагов прописать все необходимые этапы препроцессинга и получить уже готовые к кластеризации данные. 2) уже вышло приложение Azimuth, которое на вход принимает файл после картирования (в т.ч. после отработки CellRanger) в формате .h5ad, самостоятельно осуществляет препроцессинг, картирование и визуализацию данных (особенно касается часто секвенируемых фрагментов органов или тканей). 3) Если же транскриптом мало изучен, тогда есть смысл все этапы проводить вручную.

## 2.7. Создание удаленного репозитория

Для данного исследования был создан удаленный репозиторий на GitHub, который находится по адресу <https://github.com/e-ginger/bmstu-ds-course>. На него были загружены результаты работы: исследовательский notebook, код приложения.

# Заключение

В ходе выполнения данной работы мы прошли практически весь Dataflow pipeline, рассмотрели большую часть операций и задач, которые приходится выполнять специалисту по работе с данными.

Этот поток операций и задач включает:

* изучение теоретических методов анализа данных и машинного обучения;
* изучение основ предметной области, в которой решается задача;
* извлечение и транформацию данных. Здесь нам был предоставлен готовый набор данных, поэтому через трудности работы с разными источниками и парсингом данных мы еще не соприкоснулись;
* проведение разведочного анализа данных статистическими методами;
* DataMining — извлечение признаков из датасета и их анализ;
* разделение имеющихся, в нашем случае размеченных, данных на обучающую, валидационную, тестовую выборки;
* выполнение предобработки (препроцессинга) данных для обеспечения корректной работы моделей;
* построение аналитического решения. Это включает выбор алгоритма решения и модели, сравнение различных моделей, подбор гиперпараметров модели;
* визуализация модели и оценка качества аналитического решения;
* сохранение моделей;
* разработка и тестирование приложения для поддержки принятия решений специалистом предметной области, которое использовало бы найденную модель;
* внедрение решения и приложения в эксплуатацию. Этот блок задач мы тоже пока не затронули.

В этой работе мы имели дело не с учебными наборами данных, которые дают хорошо изученные решения, а с реальной производственной задачей. И к сожалению, не смогли поставленную задачу решить — не получили моделей, которые бы описывали закономерности предметной области. Я проделала максимум исследований, которые в моей компетенции как начинающего дата-сайентиста, применила большую часть знаний, полученных в ходе прохождения курса.

Возможные причины неудачи:

* нечеткая постановка задачи, отсутствие дополнительной информации о зависимости признаков с точки зрения физики процесса. Незначимые признаки являются для модели шумом, и мешают найти зависимость целевых от значимых входных признаков;
* исследование предварительно обработанных данных. Возможно, на "сырых", не предобработанных данных можно было бы получить более качественные модели, воспользовавшись другими методами очистки и подготовки;
* мой недостаток знаний и опыта. Нейросети являются самым современным подходам к решению такого рода задач. Они способны находить скрытые и нелинейные зависимости в данных. Но выбор оптимальной архитектуры нейросети является неочевидной задачей.

Дальнейшие возможные пути решения этой задачи могли бы быть:

* углубиться в изучение нейросетей, попробовать различные архитектуры, параметры обучения и т.д.;
* провести отбор признаков разными методами. Испробовать методы уменьшения размерности, например метод главных компонент;
* после уменьшения размерности градиентный бустинг может улучшить свои результаты. Так же есть большой простор для подбора гиперпараметров для этого метода;

проконсультироваться у экспертов в предметной области. Возможно, они могли бы поделиться знаниями, необходимыми для решения задачи.

**Секвени́рование РНК** ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *RNA sequencing, RNA-seq*) — метод определения [первичной структуры](https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D0%B2%D0%B8%D1%87%D0%BD%D0%B0%D1%8F_%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%B0) [молекул](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%BB%D0%B5%D0%BA%D1%83%D0%BB%D0%B0) [РНК](https://ru.wikipedia.org/wiki/%D0%A0%D0%9D%D0%9A), представляющий собой высокочувствительный и точный инструмент для изучения [транскриптома](https://ru.wikipedia.org/wiki/%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%BC" \o "Транскриптом). Под этим может подразумеваться как [секвенирование](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) [мРНК](https://ru.wikipedia.org/wiki/%D0%9C%D0%A0%D0%9D%D0%9A), так и определение последовательности [некодирующих РНК](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%BA%D0%BE%D0%B4%D0%B8%D1%80%D1%83%D1%8E%D1%89%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A" \o "Некодирующие РНК). Современное полногеномное секвенирование основано на прямом секвенировании фрагментов [кДНК](https://ru.wikipedia.org/wiki/%D0%9A%D0%94%D0%9D%D0%9A" \o "КДНК)[[1]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-1).

В отличие от другого широкомасштабного метода анализа транскриптома — [экспрессионных микрочипов](https://ru.wikipedia.org/wiki/%D0%94%D0%9D%D0%9A-%D0%BC%D0%B8%D0%BA%D1%80%D0%BE%D1%87%D0%B8%D0%BF), РНК-секвенирование позволяет получать данные об [аллель](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D0%BB%D0%B5%D0%BB%D1%8C)-специфичной [экспрессии генов](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2), [сплайсинговых](https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%BB%D0%B0%D0%B9%D1%81%D0%B8%D0%BD%D0%B3) вариантах транскриптов, пост- и ко-трансляционном [редактировании РНК](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B4%D0%B0%D0%BA%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A), [однонуклеотидных полиморфизмах](https://ru.wikipedia.org/wiki/%D0%9E%D0%B4%D0%BD%D0%BE%D0%BD%D1%83%D0%BA%D0%BB%D0%B5%D0%BE%D1%82%D0%B8%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BF%D0%BE%D0%BB%D0%B8%D0%BC%D0%BE%D1%80%D1%84%D0%B8%D0%B7%D0%BC), а также [химерных генах](https://ru.wikipedia.org/wiki/%D0%A5%D0%B8%D0%BC%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B3%D0%B5%D0%BD). Кроме того, РНК-секвенирование позволяет получить абсолютную количественную информацию о представленности различных транскриптов в пробе, в отличие от относительных количественных данных микрочипов[[2]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-2)[[3]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-3).

Совершенствование технологий секвенирования РНК наряду с развитием секвенирования РНК одиночных клеток ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *single-cell RNA-seq*) позволяет более детально изучать [этиологию](https://ru.wikipedia.org/wiki/%D0%AD%D1%82%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F) и [патогенез](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D1%82%D0%BE%D0%B3%D0%B5%D0%BD%D0%B5%D0%B7) различных заболеваний[[4]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-4)[[5]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-5).



## Содержание

* [1История](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%98%D1%81%D1%82%D0%BE%D1%80%D0%B8%D1%8F)
* [2Методы](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%D1%8B)
  + [2.1Основные принципы секвенирования РНК](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9E%D1%81%D0%BD%D0%BE%D0%B2%D0%BD%D1%8B%D0%B5_%D0%BF%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF%D1%8B_%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_%D0%A0%D0%9D%D0%9A)
    - [2.1.1Создание библиотеки поли(А)-транскриптов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%A1%D0%BE%D0%B7%D0%B4%D0%B0%D0%BD%D0%B8%D0%B5_%D0%B1%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B8_%D0%BF%D0%BE%D0%BB%D0%B8(%D0%90)-%D1%82%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%B2)
    - [2.1.2Удаление рибосомной РНК](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%A3%D0%B4%D0%B0%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%80%D0%B8%D0%B1%D0%BE%D1%81%D0%BE%D0%BC%D0%BD%D0%BE%D0%B9_%D0%A0%D0%9D%D0%9A)
    - [2.1.3Фрагментация](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%A4%D1%80%D0%B0%D0%B3%D0%BC%D0%B5%D0%BD%D1%82%D0%B0%D1%86%D0%B8%D1%8F)
    - [2.1.4Адаптеры и направление цепей](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%90%D0%B4%D0%B0%D0%BF%D1%82%D0%B5%D1%80%D1%8B_%D0%B8_%D0%BD%D0%B0%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D1%86%D0%B5%D0%BF%D0%B5%D0%B9)
    - [2.1.5Амплификация и молекулярная маркировка](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%90%D0%BC%D0%BF%D0%BB%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F_%D0%B8_%D0%BC%D0%BE%D0%BB%D0%B5%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%BC%D0%B0%D1%80%D0%BA%D0%B8%D1%80%D0%BE%D0%B2%D0%BA%D0%B0)
  + [2.2Секвенирование РНК для особых целей](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A_%D0%B4%D0%BB%D1%8F_%D0%BE%D1%81%D0%BE%D0%B1%D1%8B%D1%85_%D1%86%D0%B5%D0%BB%D0%B5%D0%B9)
    - [2.2.1Измерение профиля экспрессии генов с помощью методов, основанных на использовании тэгов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%98%D0%B7%D0%BC%D0%B5%D1%80%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BF%D1%80%D0%BE%D1%84%D0%B8%D0%BB%D1%8F_%D1%8D%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%B8_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2_%D1%81_%D0%BF%D0%BE%D0%BC%D0%BE%D1%89%D1%8C%D1%8E_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D0%BE%D0%B2,_%D0%BE%D1%81%D0%BD%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%D0%BD%D0%B0_%D0%B8%D1%81%D0%BF%D0%BE%D0%BB%D1%8C%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B8_%D1%82%D1%8D%D0%B3%D0%BE%D0%B2)
    - [2.2.2Прямое секвенирование РНК](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9F%D1%80%D1%8F%D0%BC%D0%BE%D0%B5_%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A)
* [3Проблемы](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9F%D1%80%D0%BE%D0%B1%D0%BB%D0%B5%D0%BC%D1%8B)
* [4Применение](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9F%D1%80%D0%B8%D0%BC%D0%B5%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5)
  + [4.1Определение профиля экспрессии генов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9E%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BF%D1%80%D0%BE%D1%84%D0%B8%D0%BB%D1%8F_%D1%8D%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%B8_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2)
  + [4.2Определение мест альтернативного сплайсинга и выявление однонуклеотидных полиморфизмов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9E%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BC%D0%B5%D1%81%D1%82_%D0%B0%D0%BB%D1%8C%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B3%D0%BE_%D1%81%D0%BF%D0%BB%D0%B0%D0%B9%D1%81%D0%B8%D0%BD%D0%B3%D0%B0_%D0%B8_%D0%B2%D1%8B%D1%8F%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BE%D0%B4%D0%BD%D0%BE%D0%BD%D1%83%D0%BA%D0%BB%D0%B5%D0%BE%D1%82%D0%B8%D0%B4%D0%BD%D1%8B%D1%85_%D0%BF%D0%BE%D0%BB%D0%B8%D0%BC%D0%BE%D1%80%D1%84%D0%B8%D0%B7%D0%BC%D0%BE%D0%B2)
  + [4.3Изучение редактирования РНК](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%98%D0%B7%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%80%D0%B5%D0%B4%D0%B0%D0%BA%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_%D0%A0%D0%9D%D0%9A)
  + [4.4РНК-секвенирование раковых транскриптомов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%A0%D0%9D%D0%9A-%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D1%80%D0%B0%D0%BA%D0%BE%D0%B2%D1%8B%D1%85_%D1%82%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%82%D0%BE%D0%BC%D0%BE%D0%B2)
  + [4.5Детекция гибридных генов](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%94%D0%B5%D1%82%D0%B5%D0%BA%D1%86%D0%B8%D1%8F_%D0%B3%D0%B8%D0%B1%D1%80%D0%B8%D0%B4%D0%BD%D1%8B%D1%85_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2)
  + [4.6ENCODE и modENCODE](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#ENCODE_%D0%B8_modENCODE)
* [5Примечания](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9F%D1%80%D0%B8%D0%BC%D0%B5%D1%87%D0%B0%D0%BD%D0%B8%D1%8F)
* [6Литература](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#%D0%9B%D0%B8%D1%82%D0%B5%D1%80%D0%B0%D1%82%D1%83%D1%80%D0%B0)

## История[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=1) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=1)]

Технологическая платформа для быстрого широкомасштабного секвенирования была создана в 2005 году фирмами [454 Life Sciences](https://ru.wikipedia.org/wiki/%D0%9F%D0%B8%D1%80%D0%BE%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)[[6]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-6) и [Illumina](https://ru.wikipedia.org/wiki/Illumina/Solexa_method" \o "Illumina/Solexa method) (ранее Solexa)[[7]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-7), и сначала использовалась для [секвенирования геномов](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D0%BE%D0%BB%D0%BD%D0%BE%D0%B3%D0%B5%D0%BD%D0%BE%D0%BC%D0%BD%D0%BE%D0%B5_%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Whole_genome_sequencing). Первые работы по секвенированию транскриптомов появились в 2008 году. В числе первых были секвенированы транскриптом [дрожжей](https://ru.wikipedia.org/wiki/%D0%94%D1%80%D0%BE%D0%B6%D0%B6%D0%B8)[[8]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-8), [арабидопсиса](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B7%D1%83%D1%85%D0%BE%D0%B2%D0%B8%D0%B4%D0%BA%D0%B0_%D0%A2%D0%B0%D0%BB%D1%8F" \o "Резуховидка Таля)[[9]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-9) и [мыши](https://ru.wikipedia.org/wiki/%D0%9C%D1%8B%D1%88%D0%B8)[[10]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-10).

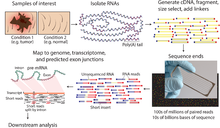
В настоящее время РНК-секвенирование осуществляется в основном с использованием трех инструментальных платформ широкомасштабного секвенирования: Illumina, 454 Life Sciences и [SOLiD](https://ru.wikipedia.org/wiki/SOLiD" \o "SOLiD)[[11]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-11).

В 2019 году удалось секвенировать [РНК](https://ru.wikipedia.org/wiki/%D0%A0%D0%9D%D0%9A) из кожи, хрящей, печени и скелетных мышц щенка Тумата (волка или собаки) возрастом 14300 лет[[12]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-12).

## Методы[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=2) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=2)]

### Основные принципы секвенирования РНК[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=3) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=3)]

Большинство экспериментов по секвенированию РНК проводятся на оборудовании, которое предназначено для секвенирования молекул ДНК. В связи с этим необходимым шагом для секвенирования РНК является создание библиотеки [кДНК](https://ru.wikipedia.org/wiki/%D0%9A%D0%94%D0%9D%D0%9A" \o "КДНК), полученной из исследуемой тотальной РНК. Каждая кДНК из такой библиотеки представляет собой фрагмент ДНК разного размера, фланкированный по обоим краям специальными [адаптерами](https://ru.wikipedia.org/w/index.php?title=%D0%90%D0%B4%D0%B0%D0%BF%D1%82%D0%B5%D1%80_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Adapter_(genetics)). Наличие адаптеров необходимо для последующей [амплификации](https://ru.wikipedia.org/wiki/%D0%90%D0%BC%D0%BF%D0%BB%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D1%8F_(%D0%BC%D0%BE%D0%BB%D0%B5%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)) образцов и секвенирования. Методы создания [библиотек кДНК](https://ru.wikipedia.org/w/index.php?title=%D0%91%D0%B8%D0%B1%D0%BB%D0%B8%D0%BE%D1%82%D0%B5%D0%BA%D0%B0_%D0%BA%D0%94%D0%9D%D0%9A&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/cDNA_library) варьируются в зависимости от конечной цели исследования и типа изучаемой РНК (РНК может различаться в размере, [последовательности](https://ru.wikipedia.org/wiki/%D0%9D%D1%83%D0%BA%D0%BB%D0%B5%D0%BE%D1%82%D0%B8%D0%B4%D0%BD%D0%B0%D1%8F_%D0%BF%D0%BE%D1%81%D0%BB%D0%B5%D0%B4%D0%BE%D0%B2%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%8C), структурных особенностях а также в концентрации). Перед созданием библиотеки кДНК, подходящей для конкретного эксперимента, необходимо ответить на следующие вопросы: 1) какие именно молекулы РНК представляют интерес; 2) как получить кДНК желаемого размера; 3) каким способом лучше присоединить адаптерные последовательности к краям кДНК для амплификации и секвенирования[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).

[](https://commons.wikimedia.org/wiki/File:Journal.pcbi.1004393.g002.png?uselang=ru)

Общая схема секвенирования РНК.

#### Создание библиотеки поли(А)-транскриптов**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=4)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=4)**]**

Секвенирование [полиаденилированной](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%B8%D0%B0%D0%B4%D0%B5%D0%BD%D0%B8%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5" \o "Полиаденилирование) РНК находит широкое применение в секвенировании РНК. У [эукариот](https://ru.wikipedia.org/wiki/%D0%AD%D1%83%D0%BA%D0%B0%D1%80%D0%B8%D0%BE%D1%82%D1%8B) большая часть белок-кодирующих РНК ([мРНК](https://ru.wikipedia.org/wiki/%D0%9C%D0%A0%D0%9D%D0%9A)) и [длинных некодирующих РНК](https://ru.wikipedia.org/wiki/%D0%94%D0%BB%D0%B8%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BD%D0%B5%D0%BA%D0%BE%D0%B4%D0%B8%D1%80%D1%83%D1%8E%D1%89%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A) (РНК длиной более 200 [пар оснований](https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B0%D1%80%D0%B5%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BE%D1%81%D0%BD%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F) (п. о.)) содержат поли-(А)-хвосты. Наличие поли-(А)-хвоста делает технически простым обогащение препарата суммарной РНК поли-(А)-содержащими РНК (1—5 % от всей суммарной клеточной РНК). Отбор поли-А содержащих РНК можно производить с помощью [магнитных](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D0%B3%D0%BD%D0%B8%D1%82) или [целлюлозных](https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BB%D0%BB%D1%8E%D0%BB%D0%BE%D0%B7%D0%B0) бусин, покрытых [праймерами](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B0%D0%B9%D0%BC%D0%B5%D1%80" \o "Праймер), содержащими олиго-dT-участки[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13). Веб-сайт «The Protocol Online»[[14]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-14) предоставляет список нескольких протоколов, относящихся к выделению мРНК.

#### Удаление рибосомной РНК**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=5)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=5)**]**

Неполиаденилированные РНК, такие как мРНК [прокариот](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%BA%D0%B0%D1%80%D0%B8%D0%BE%D1%82%D1%8B), фрагменты мРНК, полученные из препаратов, зафиксированных [формалином](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D0%B8%D0%BD), и транскрипты без поли-(А)-хвостов у эукариот, зачастую являются объектами исследований. Самая большая трудность в секвенировании таких РНК заключается в необходимости очистить суммарную РНК от [рибосомной РНК](https://ru.wikipedia.org/wiki/%D0%A0%D0%B8%D0%B1%D0%BE%D1%81%D0%BE%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D0%A0%D0%9D%D0%9A" \o "Рибосомальная РНК) (рРНК), которая превалирует в образце (например, в активно делящихся клетках млекопитающих количество рРНК от суммарной РНК может доходить до 80 %[[15]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-15))[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13). Существует несколько способов элиминации рРНК:

1. Первый подход основан на специфичных к последовательностям пробах, которые могут быть [гибридизованы](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D0%B1%D1%80%D0%B8%D0%B4%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_%D0%94%D0%9D%D0%9A" \o "Гибридизация ДНК) с рРНК. Нежелательные рРНК или их кДНК гибридизуют с [биотинилированной](https://ru.wikipedia.org/wiki/%D0%91%D0%B8%D0%BE%D1%82%D0%B8%D0%BD" \o "Биотин) [ДНК](https://ru.wikipedia.org/wiki/%D0%94%D0%9D%D0%9A) или же с пробами, содержащими «закрытые» [нуклеиновые кислоты](https://ru.wikipedia.org/wiki/%D0%9D%D1%83%D0%BA%D0%BB%D0%B5%D0%B8%D0%BD%D0%BE%D0%B2%D1%8B%D0%B5_%D0%BA%D0%B8%D1%81%D0%BB%D0%BE%D1%82%D1%8B) ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *locked nucleic acid, LNA*), и затем проводят очистку на [стрептавидиновых](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D1%80%D0%B5%D0%BF%D1%82%D0%B0%D0%B2%D0%B8%D0%B4%D0%B8%D0%BD" \o "Стрептавидин) бусинах. В другом методе (методе направленной деградации ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *probe-directed degradation, PDD*[[16]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-16)) рРНК помечаются [антисмысловыми](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D1%82%D0%B8%D1%81%D0%BC%D1%8B%D1%81%D0%BB%D0%BE%D0%B2%D1%8B%D0%B5_%D0%A0%D0%9D%D0%9A" \o "Антисмысловые РНК) олиго-ДНК-праймерами и обрабатываются [РНКазой Н](https://ru.wikipedia.org/wiki/%D0%A0%D0%9D%D0%9A%D0%B0%D0%B7%D0%B0_H" \o "РНКаза H). В третьем методе из всех кДНК, полученных и с рРНК и с других РНК, делают кольцевые молекулы, а затем гибридизуют с пробами, содержащими рРНК. Гибридизованные последовательности расщепляются при последующей обработке дуплекс-специфической нуклеазой ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *duplex-specific nuclease, DSN*), которая обладает специфичностью к двуцепочечной ДНК. Последний метод имеет ограничения ввиду необходимости большого количества РНК[[17]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-17).
2. Другой подход избавления от рРНК основан на использовании специфических праймеров NSR ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *not-so-random (NSR) primers*), которые связываются только с интересующими молекулами РНК во время [обратной транскрипции](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%80%D0%B0%D1%82%D0%BD%D0%B0%D1%8F_%D1%82%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%86%D0%B8%D1%8F) при получении кДНК. Данный метод, запущенный на рынок под названием Ovation компанией NuGEN, использует [гексамерные](https://ru.wikipedia.org/wiki/%D0%9E%D0%BB%D0%B8%D0%B3%D0%BE%D0%BC%D0%B5%D1%80" \o "Олигомер) или гептамерные праймеры, последовательности которых отсутствуют в рРНК. Одним из самых ярких преимуществ данного метода является хорошая работа праймеров NSR в отношении частично деградированной РНК, а также с количественно малыми образцами. Очень часто данный подход используют при изучении транскриптомов прокариот, так как создание библиотеки поли-(А)-содержащих РНК в этом случае невозможно ввиду отсутствия полиаденилирования РНК у прокариот[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).
3. К третьей группе можно отнести методы, которые используют некоторые особенности рРНК для её последующего удаления. Так, первый метод, известный как СоТ-гибридизация, основан на тепловой денатурации, отжиге и селективной деградации с помощью дуплекс-специфической нуклеазы. Двуцепочечные кДНК, полученные с РНК, превалирующей в образце, будут избирательно подвергаться деградации за счет более быстрой [кинетики](https://ru.wikipedia.org/wiki/%D0%A5%D0%B8%D0%BC%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D0%B8%D0%BD%D0%B5%D1%82%D0%B8%D0%BA%D0%B0) отжига по сравнению с другой РНК, которой в образце намного меньше. Второй метод основан на использовании [фермента](https://ru.wikipedia.org/wiki/%D0%A4%D0%B5%D1%80%D0%BC%D0%B5%D0%BD%D1%82) *TEX*[[18]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-18) ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *terminator 5'-phosphate exonuclease*), который узнает молекулы РНК, имеющие на 5'-конце [фосфат](https://ru.wikipedia.org/wiki/%D0%A4%D0%BE%D1%81%D1%84%D0%B0%D1%82), как у рРНК и [тРНК](https://ru.wikipedia.org/wiki/%D0%A2%D0%A0%D0%9D%D0%9A)[[19]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-19).

#### Фрагментация**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=6)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=6)**]**

После процедуры создания библиотеки поли-(А)-транскриптов либо процедуры удаления рРНК образцы РНК подвергаются фрагментации (обычно перед проведением обратной транскрипции все образцы РНК делаются одинакового размера). Отчасти это обусловлено ограниченными возможностями секвенирующих платформ. Так например, Illumina позволяет секвенировать образцы размером до 1500 п. о. В качестве альтернативы можно не фрагментировать РНК, а сначала сделать из неё кДНК, а затем уже полученную кДНК подвергнуть фрагментации[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).

#### Адаптеры и направление цепей**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=7)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=7)**]**

В стандартных протоколах по созданию библиотек для секвенирования РНК перед амплификацией и секвенированием к кДНК желаемого размера [лигируются](https://ru.wikipedia.org/w/index.php?title=%D0%A5%D0%B8%D0%BC%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BB%D0%B8%D0%B3%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5&action=edit&redlink=1" \o "Химическое лигирование (страница отсутствует)) ДНК-адаптеры. Несмотря на простоту, в данном подходе теряется информация о том, какая из цепей ДНК соответствует смысловой цепи РНК. Особенно это критично в исследованиях для поиска и идентификации антисмысловых и новых видов РНК. В связи с этим разработаны несколько методов, которые позволяют выявить направление цепи молекул РНК в соответствующей библиотеке кДНК[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A" \l "cite_note-:1-13).

1. Первый подход подразумевает присоединение разных адаптеров непосредственно к 5'-концу и к 3'-концу РНК . Изначально этот метод был создан для секвенирования [микроРНК](https://ru.wikipedia.org/wiki/%D0%9C%D0%B8%D0%BA%D1%80%D0%BE%D0%A0%D0%9D%D0%9A" \o "МикроРНК). Сначала у фрагментированной РНК убирается фосфатная группа с 3'-конца, а на 5'-конец, наоборот, навешивается. Данная процедура сопровождается последовательным лигированием 5'-аденилированного 3' адаптера с помощью T4 РНК лигазы II и присоединением 5'-адаптера с помощью T4 РНК лигазы I. Различие в адаптерах на разных концах РНК сохраняет информацию о её направлении (имеется в виду, что после проведения процедуры обратной транскрипции информация о том, какая их цепей кДНК соответствует исходной последовательности РНК, сохранится)[[20]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-20).
2. Второй подход основан на включении [дУТФ](https://ru.wikipedia.org/wiki/%D0%94%D0%B5%D0%B7%D0%BE%D0%BA%D1%81%D0%B8%D1%83%D1%80%D0%B8%D0%B4%D0%B8%D0%BD" \o "Дезоксиуридин) во вторую цепь кДНК. Помеченная цепь может быть деградирована непосредственно перед амплификацией с помощью [урацил-ДНК-гликозилазы](https://ru.wikipedia.org/w/index.php?title=%D0%A3%D1%80%D0%B0%D1%86%D0%B8%D0%BB-%D0%94%D0%9D%D0%9A-%D0%B3%D0%BB%D0%B8%D0%BA%D0%BE%D0%B7%D0%B8%D0%BB%D0%B0%D0%B7%D0%B0&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Uracil-DNA_glycosylase) — фермента, который выщепляет [урацил](https://ru.wikipedia.org/wiki/%D0%A3%D1%80%D0%B0%D1%86%D0%B8%D0%BB) из ДНК, содержащей дУТФ. Считается, что этот метод наиболее эффективный из всех[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).
3. Третий подход включает в себя несколько методов. В одном из них производят замену матрицы после отжига на неё случайного гексамерного праймера, содержащего тэг (короткую, до 20 п.о., уникальную последовательность)[[21]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-21). В другом методе ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *breath adapter directional sequencing, BrAD-seq*) в момент временного разделения цепей ДНК вводят последовательность с тэгом[[22]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-22).

#### Амплификация и молекулярная маркировка**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=8)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=8)**]**

Перед секвенированием кДНК её необходимо амплифицировать с помощью [ПЦР](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%B8%D0%BC%D0%B5%D1%80%D0%B0%D0%B7%D0%BD%D0%B0%D1%8F_%D1%86%D0%B5%D0%BF%D0%BD%D0%B0%D1%8F_%D1%80%D0%B5%D0%B0%D0%BA%D1%86%D0%B8%D1%8F). Непосредственно перед проведением ПЦР можно ввести молекулярные маркеры. Эта процедура особенно актуальна, если РНК в образце изначально немного, как, например, в случае секвенирования РНК одной [клетки](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B5%D1%82%D0%BA%D0%B0_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F))[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).

### Секвенирование РНК для особых целей[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=9) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=9)]

#### Измерение профиля экспрессии генов с помощью методов, основанных на использовании тэгов**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=10)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=10)**]**

Секвенирование DGE (от [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *digital gene expression*), или Tag-seq — это метод глубокого секвенирования, полученный из [SAGE](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D1%80%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D1%8D%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%B8_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Serial_analysis_of_gene_expression) (от [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *Serial Analysis of Gene Expression*). Как и в SAGE, метод включает в себя присоединение мРНК за поли-А хвост к бусинам, покрытым олиго-dT-праймерами; синтез первой и второй цепи кДНК на бусинах; расщепление двуцепочечной кДНК часто щепящей [эндонуклеазой рестрикции](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D1%81%D1%82%D1%80%D0%B8%D0%BA%D1%82%D0%B0%D0%B7%D0%B0). Оставшийся 3'-конец, который присоединен к бусинам, лигируется со своим адаптером, находящимся на 5'-конце. В адаптере есть сайт узнавания для специфической эндонуклеазы рестрикции TE (от [англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *tagging enzyme*). TE расщепляет кДНК, в ходе чего образуется короткий тэг длинной 21 п. о., который затем лигируется со следующим адаптером, находящимся на 3'-конце. кДНК амплифицируется с помощью ПЦР и секвенируется. Так как секвенируется только короткий тэг из целого транскрипта, секвенирование DGE является более экономичным вариантом в сравнении со стандартным секвенированием РНК. Секвенирование DGE сохраняет информацию о том, какая из цепей кДНК соответствует исходной РНК. Также этот метод находит широкое применение в случае, если полноразмерный геном или транскриптом организма недоступен для полноразмерного выравнивания с [прочтениями](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D1%80%D0%BE%D1%87%D1%82%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Read_(biology)), полученными в ходе секвенирования[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13)[[23]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-23).

Секвенирование 3'-концов включает в себя целый ряд методов, большинство из которых было специально разработано для поиска альтернативного сплайсинга и сайтов полиаденилирования у эукариот[[13]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-:1-13).

#### Прямое секвенирование РНК**[**[**править**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=11)**|**[**править код**](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=11)**]**

Так как обратная транскрипция РНК с помощью [обратной транскриптазы](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%80%D0%B0%D1%82%D0%BD%D0%B0%D1%8F_%D1%82%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%82%D0%B0%D0%B7%D0%B0) дает большое число ошибок и артефактов, которые могут препятствовать корректному качественному и количественному анализу транскриптов[[24]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-24), компанией [Helicos](https://ru.wikipedia.org/wiki/Helicos_Biosciences" \o "Helicos Biosciences) была начата разработка технологии мономолекулярного прямого секвенирования РНК ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *Direct RNA Sequencing, DRSTM*). Этот метод предполагает секвенирование РНК массово-параллельным образом, без получения кДНК, лигирования, амплификации и других процедур, которые могут изменить образец[[25]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-25).

## Проблемы[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=12) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=12)]

Основная проблема технологии RNA-seq заключается в том, что исходно неизвестно, какому транскрипту соответствует прочитанный фрагмент. Особенно сложно решить данную проблему в случае исследования транскриптома высших эукариот с частым альтернативным сплайсингом и присутствием в геноме большого числа [паралогов](https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D1%80%D0%B0%D0%BB%D0%BE%D0%B3" \o "Паралог). Существует два подхода для восстановления транскриптов по прочитанным фрагментам: [картирование](https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D1%80%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%BA%D0%BE%D1%80%D0%BE%D1%82%D0%BA%D0%B8%D1%85_%D0%BF%D1%80%D0%BE%D1%87%D1%82%D0%B5%D0%BD%D0%B8%D0%B9) на [геном](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%BE%D0%BC) отдельных прочитанных фрагментов[[26]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-26) или восстановление структуры транскрипта *[de novo](https://ru.wikipedia.org/w/index.php?title=De_novo&action=edit&redlink=1" \o "De novo (страница отсутствует))*[[en]](https://en.wikipedia.org/wiki/de_novo) с последующим картированием полноразмерного транскрипта на геном[[27]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-27).

## Применение[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=13) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=13)]

### Определение профиля экспрессии генов[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=14) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=14)]

Метод секвенирования РНК становится основным методом определения того, какие гены и на каком уровне экспрессируются в клетке. С помощью РНК секвенирования можно определять различия в [экспрессии генов](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2) на различных стадиях [развития](https://ru.wikipedia.org/wiki/%D0%9E%D0%BD%D1%82%D0%BE%D0%B3%D0%B5%D0%BD%D0%B5%D0%B7) организма[[28]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-28) или в разных [тканях](https://ru.wikipedia.org/wiki/%D0%A2%D0%BA%D0%B0%D0%BD%D1%8C_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F))[[29]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-29). Например, разработан метод локализации *[in situ](https://ru.wikipedia.org/wiki/In_situ" \o "In situ)* последовательностей РНК-транскриптов с помощью [флуоресцентного](https://ru.wikipedia.org/wiki/%D0%A4%D0%BB%D1%83%D0%BE%D1%80%D0%B5%D1%81%D1%86%D0%B5%D0%BD%D1%86%D0%B8%D1%8F) секвенирования ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) [*Fluorescent in situ Sequencing, FISSEQ*](https://ru.wikipedia.org/w/index.php?title=Fluorescent_in_situ_Sequencing&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/Fluorescent_in_situ_sequencing)), который позволяет изучать [фенотип](https://ru.wikipedia.org/wiki/%D0%A4%D0%B5%D0%BD%D0%BE%D1%82%D0%B8%D0%BF) клеток и регуляцию активности генов непосредственно в биологическом образце (на срезах тканей)[[30][30]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-%D0%B0%D0%B2%D1%82%D0%BE%D1%81%D1%81%D1%8B%D0%BB%D0%BA%D0%B01-30). Также можно определить, [транскрипция](https://ru.wikipedia.org/wiki/%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D0%BA%D1%80%D0%B8%D0%BF%D1%86%D0%B8%D1%8F_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)) каких генов изменяется при развитии болезней и [рака](https://ru.wikipedia.org/wiki/%D0%97%D0%BB%D0%BE%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BE%D0%BF%D1%83%D1%85%D0%BE%D0%BB%D1%8C)[[31]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-31). В связи с удешевлением [методов секвенирования нового поколения](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4%D1%8B_%D1%81%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_%D0%BD%D0%BE%D0%B2%D0%BE%D0%B3%D0%BE_%D0%BF%D0%BE%D0%BA%D0%BE%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F) появилась возможность определять экспрессию генов у любого человека для диагностики заболеваний. Наряду с секвенированием РНК для измерения профиля экспрессии генов также широко используется [кэп-анализ экспрессии генов](https://ru.wikipedia.org/wiki/%D0%9A%D1%8D%D0%BF-%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D1%8D%D0%BA%D1%81%D0%BF%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%B8_%D0%B3%D0%B5%D0%BD%D0%BE%D0%B2)[[32]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-32).

### Определение мест альтернативного сплайсинга и выявление однонуклеотидных полиморфизмов[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=15) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=15)]

Секвенирование РНК — наиболее удобный способ определения мест [альтернативного сплайсинга](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D1%8C%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B9_%D1%81%D0%BF%D0%BB%D0%B0%D0%B9%D1%81%D0%B8%D0%BD%D0%B3), а также количественного соотношения различных альтернативных форм транскрипта[[33]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-33)[[34]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-34). Другие методы не позволяют картировать места альтернативного сплайсинга на всем протяжении генома. Также как и определение экспрессии генов, определение соотношения альтернативных форм транскриптов можно проводить на различных стадиях развития организма или в разных тканях.

РНК-секвенирование позволяет различить транскрипты с отличием в одном нуклеотиде, поэтому может быть использовано как для выявления экспрессируемых однонуклеотидных полиморфизмов в генах, так и для изучения процесса редактирования РНК[[35]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-35)[[36]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-%D0%B0%D0%B2%D1%82%D0%BE%D1%81%D1%81%D1%8B%D0%BB%D0%BA%D0%B02-36).

### Изучение редактирования РНК[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=16) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=16)]

*См. также:*[*Редактирование РНК*](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B4%D0%B0%D0%BA%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A)

Редактирование РНК — процесс пост- или ко- транскрипционной модификации [рибонуклеотидов](https://ru.wikipedia.org/wiki/%D0%A0%D0%B8%D0%B1%D0%BE%D0%BD%D1%83%D0%BA%D0%BB%D0%B5%D0%BE%D1%82%D0%B8%D0%B4" \o "Рибонуклеотид) в молекуле РНК. В большинстве случаев редактирование РНК приводит к замене [аденозина](https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D0%B5%D0%BD%D0%BE%D0%B7%D0%B8%D0%BD) [инозином](https://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%BE%D0%B7%D0%B8%D0%BD)[[36]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-%D0%B0%D0%B2%D1%82%D0%BE%D1%81%D1%81%D1%8B%D0%BB%D0%BA%D0%B02-36); [катализаторами](https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D1%82%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%82%D0%BE%D1%80) указанных изменений являются [белки](https://ru.wikipedia.org/wiki/%D0%91%D0%B5%D0%BB%D0%BE%D0%BA) [семейства](https://ru.wikipedia.org/wiki/%D0%91%D0%B5%D0%BB%D0%BA%D0%BE%D0%B2%D0%BE%D0%B5_%D1%81%D0%B5%D0%BC%D0%B5%D0%B9%D1%81%D1%82%D0%B2%D0%BE) [ADAR](https://ru.wikipedia.org/w/index.php?title=ADAR&action=edit&redlink=1)[[en]](https://en.wikipedia.org/wiki/ADAR). В дальнейшем инозин распознаётся клеточной машинерией (например, рибосомой) как [гуанозин](https://ru.wikipedia.org/wiki/%D0%93%D1%83%D0%B0%D0%BD%D0%BE%D0%B7%D0%B8%D0%BD" \o "Гуанозин), что приводит к возникновению различий между закодированной в геноме информацией и её интерпретацией[[37]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-37).

Основным методом выявления внесённых изменений является сравнение последовательности нуклеотидов геномной ДНК и соответствующих участков РНК[[38]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-38).

Важным прогностическим признаком обнаружения сайтов редактирования РНК является наличие [эволюционно](https://ru.wikipedia.org/wiki/%D0%AD%D0%B2%D0%BE%D0%BB%D1%8E%D1%86%D0%B8%D1%8F_(%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F)) [консервативных](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BD%D1%81%D0%B5%D1%80%D0%B2%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D1%8B%D0%B5_%D0%BF%D0%BE%D1%81%D0%BB%D0%B5%D0%B4%D0%BE%D0%B2%D0%B0%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D0%B8) нуклеотидных последовательностей в окружении места редактирования[[39]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-39).

Вследствие значительного прогресса в развитии методов массового параллельного секвенирования стало технически возможным проводить секвенирование полного транскриптома исследуемого организма с целью выявления связанных с редактированием РНК событий. Однако в силу [генетического разнообразия](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%B5%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D1%80%D0%B0%D0%B7%D0%BD%D0%BE%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%B8%D0%B5) наличие различий в определённой позиции между последовательностью РНК и [референсным геномом](https://ru.wikipedia.org/w/index.php?title=%D0%A0%D0%B5%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%81%D0%BD%D1%8B%D0%B9_%D0%B3%D0%B5%D0%BD%D0%BE%D0%BC&action=edit&redlink=1" \o "Референсный геном (страница отсутствует))[[en]](https://en.wikipedia.org/wiki/Reference_genome) не означает присутствия сайта редактирования в этой позиции, так как идентификация сайтов редактирования РНК подразумевает секвенирование как геномной ДНК, так и кДНК, выделенных из одного и того же организма. Также необходимо принимать во внимание то, что уровни редактирования РНК различаются в разных тканях организма[[40]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-40).

Для упрощения процедуры идентификации сайтов редактирования РНК предпринимаются попытки разработать программные пакеты, использующие только транскриптомные данные и не требующие секвенирования геномной ДНК. Возможным решением может послужить программное обеспечение GIREMI[[41]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-41) ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *Genome-independent Identification of RNA Editing by Mutual Information*), которое способно детектировать сайты редактирования РНК, используя исключительно последовательности транскриптов[[42]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-42).

### РНК-секвенирование раковых транскриптомов[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=17) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=17)]

РНК-секвенирование широко используется в настоящее время для исследование особенностей транскриптома [раковых клеток](https://ru.wikipedia.org/wiki/%D0%97%D0%BB%D0%BE%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BA%D0%BB%D0%B5%D1%82%D0%BA%D0%B8), в том числе появление химерных транскриптов[[43]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-%D0%B0%D0%B2%D1%82%D0%BE%D1%81%D1%81%D1%8B%D0%BB%D0%BA%D0%B03-43) и продуктов альтернативного сплайсинга, специфичных для [раковых](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D0%BA_(%D0%B7%D0%B0%D0%B1%D0%BE%D0%BB%D0%B5%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)) клеток[[44]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-44).

### Детекция гибридных генов[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=18) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=18)]

Гибридизация генов происходит из-за различных структурных модификаций в геноме и может быть связана с раком[[45]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-45). Возможность анализировать весь транскриптом образца с помощью секвенирования РНК делает этот метод привлекательным для поиска подобных частых преобразований при раковой трансформации клеток[[43]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-%D0%B0%D0%B2%D1%82%D0%BE%D1%81%D1%81%D1%8B%D0%BB%D0%BA%D0%B03-43).

### ENCODE и modENCODE[[править](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&veaction=edit&section=19) | [править код](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A&action=edit&section=19)]

*Основная статья:*[***ENCODE***](https://ru.wikipedia.org/wiki/ENCODE)

Секвенирование РНК является одним из основных методов исследований, проводимых в рамках проектов ENCODE и modENCODE, направленных на создание базы данных элементов [генома человека](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%BE%D0%BC_%D1%87%D0%B5%D0%BB%D0%BE%D0%B2%D0%B5%D0%BA%D0%B0)[[46]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-46) и основных модельных объектов [молекулярной биологии](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%BB%D0%B5%D0%BA%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D0%B0%D1%8F_%D0%B1%D0%B8%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F)[[47]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-47)[[48]](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D0%BA%D0%B2%D0%B5%D0%BD%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A0%D0%9D%D0%9A#cite_note-48).

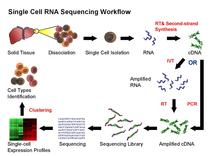
## Transcriptome sequencing (scRNA-seq)[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=11)]

*Main article:*[*Single-cell transcriptomics*](https://en.wikipedia.org/wiki/Single-cell_transcriptomics)

Standard methods such as [microarrays](https://en.wikipedia.org/wiki/Microarray) and bulk [RNA-seq](https://en.wikipedia.org/wiki/RNA-seq) analyze the RNA expression from large populations of cells. These measurements may obscure critical differences between individual cells in mixed-cell populations.[[42]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Shapiro-42)[[43]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Kolodziejczyk-43)

Single-cell RNA sequencing (scRNA-seq) provides the [expression profiles](https://en.wikipedia.org/wiki/Gene_expression_profiling) of individual cells and is considered the [gold standard](https://en.wikipedia.org/wiki/Gold_standard_(test)) for defining cell states and phenotypes as of 2020.[[44]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-44) Although it is impossible to obtain complete information on every RNA expressed by each cell, due to the small amount of material available, gene expression patterns can be identified through gene [clustering analyses](https://en.wikipedia.org/wiki/Cluster_analysis).[[45]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-45) This can uncover rare cell types within a cell population that may never have been seen before. For example, one group of scientists performing scRNA-seq on neuroblastoma tumor tissue identified a rare pan-neuroblastoma cancer cell, which may be attractive for novel therapy approaches.[[46]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-46)

### Methods[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=12" \o "Edit section: Methods)]

[](https://en.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf)

Single-cell RNA sequencing workflow

Current scRNA-seq protocols involve isolating single cells and their RNA, and then following the same steps as bulk RNA-seq: [reverse transcription](https://en.wikipedia.org/wiki/Reverse_transcription) (RT), amplification, library generation and sequencing. Early methods separated individual cells into separate wells; more recent methods encapsulate individual cells in droplets in a microfluidic device, where the reverse transcription reaction takes place, converting RNAs to cDNAs. Each droplet carries a DNA "barcode" that uniquely labels the cDNAs derived from a single cell. Once reverse transcription is complete, the cDNAs from many cells can be mixed together for sequencing, because transcripts from a particular cell are identified by the unique barcode.[[47]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-47)[[48]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-48)

Challenges for scRNA-Seq include preserving the initial relative abundance of mRNA in a cell and identifying rare transcripts.[[49]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Hebenstreit-49) The reverse transcription step is critical as the efficiency of the RT reaction determines how much of the cell's RNA population will be eventually analyzed by the sequencer. The [processivity](https://en.wikipedia.org/wiki/Processivity) of reverse transcriptases and the priming strategies used may affect full-length cDNA production and the generation of libraries biased toward 3’ or 5' end of genes.

In the amplification step, either PCR or in vitro transcription (IVT) is currently used to amplify cDNA. One of the advantages of PCR-based methods is the ability to generate full-length cDNA. However, different PCR efficiency on particular sequences (for instance, GC content and snapback structure) may also be exponentially amplified, producing libraries with uneven coverage. On the other hand, while libraries generated by IVT can avoid PCR-induced sequence bias, specific sequences may be transcribed inefficiently, thus causing sequence drop-out or generating incomplete sequences.[[1]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Eberwine-1)[[42]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Shapiro-42) Several scRNA-seq protocols have been published: Tang et al.,[[50]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-50) STRT,[[51]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-51) SMART-seq,[[52]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-52) SORT-seq,[[53]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-53) CEL-seq,[[54]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-54) RAGE-seq,[[55]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-55) Quartz-seq.[[56]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-56) , and C1-CAGE.[[57]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-57) These protocols differ in terms of strategies for reverse transcription, cDNA synthesis and amplification, and the possibility to accommodate sequence-specific barcodes (i.e., [UMIs](https://en.wikipedia.org/wiki/Unique_molecular_identifier)) or the ability to process pooled samples.[[58]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-58)

In 2017, two approaches were introduced to simultaneously measure single-cell mRNA and protein expression through oligonucleotide-labeled antibodies known as REAP-seq,[[59]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-59) and CITE-seq.[[60]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-60) Collecting cellular contents following electrophysiological recording using patch-clamp has also allowed development of the [Patch-Seq](https://en.wikipedia.org/wiki/Patch-sequencing) method, which is steadily gaining ground in neuroscience.[[61]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-61)

#### Example of a droplet based platform - 10X method**[**[**edit**](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=13)**]**

This platform of single cell RNA sequencing allows to analyze transcriptomes on a cell-by-cell basis by the use of microfluidic partitioning to capture single cells and prepare [next-generation sequencing](https://en.wikipedia.org/wiki/Next_generation_sequencing) (NGS) [cDNA libraries](https://en.wikipedia.org/wiki/CDNA_libraries).[[62]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-62) The droplets based platform enables massively parallel sequencing of mRNA in a large numbers of individual cells by capturing single cell in oil droplet.[[63]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-63)

Overall, in a first stage individual cells are captured separately and lysed, then [reverse transcription](https://en.wikipedia.org/wiki/Reverse_transcription) (RT) of [mRNA](https://en.wikipedia.org/wiki/Messenger_RNA) is performed and [cDNA](https://en.wikipedia.org/wiki/Complementary_DNA) library is obtained. To select mRNA, the RT is performed with a single-stranded sequence of deoxythymine (oligo dT) [primer](https://en.wikipedia.org/wiki/Primer_(molecular_biology)) which bind specifically the poly(A) tail of mRNA molecules. Subsequently, the amplified cDNA library is used for sequencing.[[64]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-64)

So, the first step of the method is the single cell encapsulation and library preparation. Cells are encapsulated into Gel Beads-in-emulsion (GEMs) thanks to an automate. To form these vesicle, the automate uses a [microfluidic chip](https://en.wikipedia.org/wiki/Microfluidic_chip) and combines all components with oil. Each functional GEM contains a single cell, a single Gel Bead, and RT reagents. On the Gel Bead, olignonucleotides composed by 4 distincts parts are bind: [PCR primer](https://en.wikipedia.org/wiki/PCR_Primer) (essential for the sequencing) ; 10X barcoded oligonucleotides ; Unique Molecular Identifier (UMI) sequence ; PolydT sequence (that enables capture of [poly-adenylated](https://en.wikipedia.org/wiki/Poly-adenylated_tail) mRNA molecules).[[65]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-65) Within each GEM reaction vesicle, a single cell is lysed and undergo reverse transcription. cDNA from the same cell are identified thanks to a common 10X barcode. In addition, the number of UMIs express the gene expression level and its analyse allows to detect highly variable genes. Those data are often used for either cellular phenotype classification or new subpopulation identification.[[66]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-66)

The final step of the platform is the sequencing. Libraries generated can be directly used for single cell whole transcriptome sequencing or target sequencing workflows. The sequencing is performed by using the [Illumina dye sequencing](https://en.wikipedia.org/wiki/Illumina_dye_sequencing) method. This sequencing method is based on sequencing by synthesis (SBS) principle and the use of reversible dye-terminator that enables the identification of each single nucleotid. In order to read the transcript sequences on one end, and the barcode and UMI on the other end, paired-end sequencing readers are required.[[67]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-67)

The droplet-based platform allows the detection of rare cell types thanks to its high throughput. In fact, 500 to 10,000 cells are captured per sample from a single cell suspension. The protocol is performed easily and allows a high cell recovery rate of up to 65%. The global workflow of the droplet-based platform takes 8 hours and so is faster than the Microwell-based method (BD Rhapsody), which takes 10 hours. However, it presents some limitations as the need of fresh samples and the final detection of only 10% mRNA.

The major difference between the droplet-based method and the microwell-based method is the technique used for partitioning cells.[[68]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-68)

### Limitations[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=14" \o "Edit section: Limitations)]

Most RNA-seq methods depend on [poly(A) tail](https://en.wikipedia.org/wiki/Polyadenylation) capture to enrich mRNA and deplete abundant and uninformative rRNA. Thus, they are often restricted to sequencing polyadenylated mRNA molecules. However, recent studies are now starting to appreciate the importance of non-poly(A) RNA, such as long-noncoding RNA and microRNAs in gene expression regulation. Small-seq is a single-cell method that captures small RNAs (<300 nucleotides) such as microRNAs, fragments of tRNAs and small nucleolar RNAs in mammalian cells.[[69]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-69) This method uses a combination of “oligonucleotide masks” (that inhibit the capture of highly abundant 5.8S rRNA molecules) and size selection to exclude large RNA species such as other highly abundant rRNA molecules. To target larger non-poly(A) RNAs, such as long non-coding mRNA, histone mRNA, circular RNA, and enhancer RNA, size selection is not applicable for depleting the highly abundant ribosomal RNA molecules (18S and 28s rRNA).[[70]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Hayashi_619-70) Single-cell RamDA-Seq is a method that achieves this by performing reverse transcription with random priming (random displacement amplification) in the presence of “not so random” (NSR) primers specifically designed to avoid priming on rRNA molecule.[[71]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-71) While this method successfully captures full-length total RNA transcripts for sequencing and detected a variety of non-poly(A) RNAs with high sensitivity, it has some limitations. The NSR primers were carefully designed according to rRNA sequences in the specific organism (mouse), and designing new primer sets for other species would take considerable effort. Recently, a CRISPR-based method named scDASH (single-cell depletion of abundant sequences by hybridization) demonstrated another approach to depleting rRNA sequences from single-cell total RNA-seq libraries.[[72]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-72)

Bacteria and other prokaryotes are currently not amenable to single-cell RNA-seq due to the lack of polyadenylated mRNA. Thus, the development of single-cell RNA-seq methods that do not depend on poly(A) tail capture will also be instrumental in enabling single-cell resolution microbiome studies. Bulk bacterial studies typically apply general rRNA depletion to overcome the lack of polyadenylated mRNA on bacteria, but at the single-cell level, the total RNA found in one cell is too small.[[70]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Hayashi_619-70) Lack of polyadenylated mRNA and scarcity of total RNA found in single bacteria cells are two important barriers limiting the deployment of scRNA-seq in bacteria.

### Applications[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=15" \o "Edit section: Applications)]

scRNA-Seq is becoming widely used across biological disciplines including [Developmental biology](https://en.wikipedia.org/wiki/Developmental_biology),[[73]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-73) [Neurology](https://en.wikipedia.org/wiki/Neurology),[[74]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-74) [Oncology](https://en.wikipedia.org/wiki/Oncology),[[75]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-75)[[76]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-76)[[77]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-77) [Immunology](https://en.wikipedia.org/wiki/Immunology),[[78]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-78)[[79]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-79) Cardiovascular research[[80]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-80)[[81]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-81) and [Infectious disease](https://en.wikipedia.org/wiki/Infectious_disease_(medical_specialty)).[[82]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-82)[[83]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-83)

Using [machine learning](https://en.wikipedia.org/wiki/Machine_learning) methods, data from bulk RNA-Seq has been used to increase the signal/noise ratio in scRNA-Seq. Specifically, scientists have used gene expression profiles from [pan-cancer](https://en.wikipedia.org/wiki/Pan-Cancer_Analysis) datasets in order to build [coexpression networks](https://en.wikipedia.org/wiki/Gene_co-expression_network" \o "Gene co-expression network), and then have applied these on single cell gene expression profiles, obtaining a more robust method to detect the presence of mutations in individual cells using transcript levels.[[84]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-MercatelliRay2019-84)

Some scRNA-seq methods have also been applied to single cell microorganisms. SMART-seq2 has been used to analyze single cell eukaryotic microbes, but since it relies on poly(A) tail capture, it has not been applied in prokaryotic cells.[[85]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-85) Microfluidic approaches such as Drop-seq and the Fluidigm IFC-C1 devices have been used to sequence single malaria parasites or single yeast cells.[[86]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-86)[[87]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-87) The single-cell yeast study sought to characterize the heterogeneous stress tolerance in isogenic yeast cells before and after the yeast are exposed to salt stress. Single-cell analysis of the several transcription factors by scRNA-seq revealed heterogeneity across the population. These results suggest that regulation varies among members of a population to increase the chances of survival for a fraction of the population.

The first single-cell transcriptome analysis in a prokaryotic species was accomplished using the terminator exonuclease enzyme to selectively degrade rRNA and [rolling circle amplification](https://en.wikipedia.org/wiki/Rolling_circle_replication) (RCA) of mRNA.[[88]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-88) In this method, the ends of single-stranded DNA were ligated together to form a circle, and the resulting loop was then used as a template for linear RNA amplification. The final product library was then analyzed by microarray, with low bias and good coverage. However, RCA has not been tested with RNA-seq, which typically employs next-generation sequencing. Single-cell RNA-seq for bacteria would be highly useful for studying microbiomes. It would address issues encountered in conventional bulk metatranscriptomics approaches, such as failing to capture species present in low abundance, and failing to resolve heterogeneity among cell populations.

scRNA-Seq has provided considerable insight into the development of embryos and organisms, including the worm [Caenorhabditis elegans](https://en.wikipedia.org/wiki/Caenorhabditis_elegans),[[89]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-89) and the regenerative planarian [Schmidtea mediterranea](https://en.wikipedia.org/wiki/Schmidtea_mediterranea" \o "Schmidtea mediterranea)[[90]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-90)[[91]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-91) and axolotl [Ambystoma mexicanum](https://en.wikipedia.org/wiki/Ambystoma_mexicanum).[[92]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-92)[[93]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-93) The first vertebrate animals to be mapped in this way were [Zebrafish](https://en.wikipedia.org/wiki/Zebrafish)[[94]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-94)[[95]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-95)[[96]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-96) and [*Xenopus laevis*](https://en.wikipedia.org/wiki/Xenopus_laevis).[[97]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-97) In each case multiple stages of the embryo were studied, allowing the entire process of development to be mapped on a cell-by-cell basis. [Science](https://en.wikipedia.org/wiki/Science_Magazine) recognized these advances as the 2018 [Breakthrough of the Year](https://en.wikipedia.org/wiki/Breakthrough_of_the_Year).[[98]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-98)

A molecular cell atlas of mice testes was established to define BDE47-induced prepubertal testicular toxicity using the ScRNA-seq approach, providing novel insight into our understanding of the underlying mechanisms and pathways involved in BDE47-associated testicular injury at a single-cell resolution.[[99]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-99)

## Considerations[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=16" \o "Edit section: Considerations)]

### Isolation of single cells[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=17)]

There are several ways to isolate individual cells prior to whole genome amplification and sequencing. [Fluorescence-activated cell sorting](https://en.wikipedia.org/wiki/Fluorescence-activated_cell_sorting) (FACS) is a widely used approach. Individual cells can also be collected by micromanipulation, for example by serial dilution or by using a patch pipette or nanotube to harvest a single cell.[[15]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Genome-wide_detection_of_single-nuc-15)[[100]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-100) The advantages of micromanipulation are ease and low cost, but they are laborious and susceptible to misidentification of cell types under microscope. [Laser-capture microdissection](https://en.wikipedia.org/wiki/Laser-capture_microdissection) (LCM) can also be used for collecting single cells. Although LCM preserves the knowledge of the spatial location of a sampled cell within a tissue, it is hard to capture a whole single cell without also collecting the materials from neighboring cells.[[42]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Shapiro-42)[[101]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-101)[[102]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-102) [High-throughput methods](https://en.wikipedia.org/wiki/High-throughput_screening) for single cell isolation also include [microfluidics](https://en.wikipedia.org/wiki/Microfluidics). Both FACS and microfluidics are accurate, automatic and capable of isolating unbiased samples. However, both methods require detaching cells from their microenvironments first, thereby causing perturbation to the transcriptional profiles in RNA expression analysis.[[103]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-103)[[104]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-104)

### Number of cells to be analyzed[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=18)]

#### scRNA-Seq**[**[**edit**](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=19)**]**

Generally speaking, for a typical bulk cell [RNA-sequencing](https://en.wikipedia.org/wiki/RNA-Seq) (RNA-seq) experiment, ten million reads are generated and a gene with higher than the threshold of 50 reads per kb per million reads (RPKM) is considered expressed. For a gene that is 1kb long, this corresponds to 500 reads and a minimum [coefficient of variation](https://en.wikipedia.org/wiki/Coefficient_of_variation) (CV) of 4% under the assumption of the [Poisson distribution](https://en.wikipedia.org/wiki/Poisson_distribution). For a typical mammalian cell containing 200,000 mRNA, sequencing data from at least 50 single cells need to be pooled in order to achieve this minimum CV value. However, due to the efficiency of reverse transcription and other noise introduced in the experiments, more cells are required for accurate expression analyses and cell-type identification.[[42]](https://en.wikipedia.org/wiki/Single-cell_sequencing#cite_note-Shapiro-42)

## See also[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_sequencing&action=edit&section=20" \o "Edit section: See also)]

* [Single-cell analysis](https://en.wikipedia.org/wiki/Single-cell_analysis)
* [Single-cell transcriptomics](https://en.wikipedia.org/wiki/Single-cell_transcriptomics)
* [Single cell epigenomics](https://en.wikipedia.org/wiki/Single_cell_epigenomics)
* [DNA sequencing](https://en.wikipedia.org/wiki/DNA_sequencing)
* [Whole genome sequencing](https://en.wikipedia.org/wiki/Whole_genome_sequencing)

# Single-cell transcriptomics

1 language

* [Article](https://en.wikipedia.org/wiki/Single-cell_transcriptomics)
* [Talk](https://en.wikipedia.org/wiki/Talk:Single-cell_transcriptomics)
* [Read](https://en.wikipedia.org/wiki/Single-cell_transcriptomics)
* [Edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit)
* [View history](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=history)

Tools

From Wikipedia, the free encyclopedia

**Single-cell transcriptomics** examines the [gene](https://en.wikipedia.org/wiki/Gene) expression level of individual [cells](https://en.wikipedia.org/wiki/Cell_(biology)) in a given population by simultaneously measuring the RNA concentration (conventionally only [messenger RNA](https://en.wikipedia.org/wiki/Messenger_RNA) (mRNA)) of hundreds to thousands of genes.[[1]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-Kanter_2015_53-1) Single-cell transcriptomics makes it possible to unravel [heterogeneous](https://en.wikipedia.org/wiki/Heterogenous) cell populations, reconstruct cellular developmental pathways, and model transcriptional dynamics — all previously masked in bulk RNA sequencing.[[2]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-2)

## Background[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=1)]

The development of high-throughput [RNA sequencing](https://en.wikipedia.org/wiki/RNA_sequencing) (RNA-seq) and [microarrays](https://en.wikipedia.org/wiki/Microarrays) has made gene expression analysis a routine. RNA analysis was previously limited to tracing individual [transcripts](https://en.wikipedia.org/wiki/Transcription_(biology)) by [Northern blots](https://en.wikipedia.org/wiki/Northern_blot) or [quantitative PCR](https://en.wikipedia.org/wiki/Quantitative_PCR). Higher throughput and speed allow researchers to frequently characterize the expression profiles of populations of thousands of cells. The data from bulk [assays](https://en.wikipedia.org/wiki/Assays) has led to identifying genes differentially expressed in distinct cell populations, and [biomarker](https://en.wikipedia.org/wiki/Biomarker) discovery.[[3]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-3)

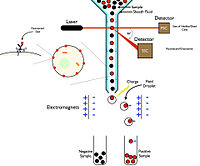
These studies are limited as they provide measurements for whole tissues and, as a result, show an average expression profile for all the constituent cells. This has a couple of drawbacks. Firstly, different cell types within the same tissue can have distinct roles in multicellular organisms. They often form subpopulations with unique transcriptional profiles. Correlations in the gene expression of the subpopulations can often be missed due to the lack of subpopulation identification.[[1]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-Kanter_2015_53-1) Secondly, bulk assays fail to recognize whether a change in the expression profile is due to a change in regulation or composition — for example if one cell type arises to dominate the population. Lastly, when your goal is to study cellular progression through [differentiation](https://en.wikipedia.org/wiki/Cellular_differentiation), average expression profiles can only order cells by time rather than by developmental stage. Consequently, they cannot show trends in gene expression levels specific to certain stages.[[4]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-4)

Recent advances in biotechnology allow the measurement of gene expression in hundreds to thousands of individual cells simultaneously. While these breakthroughs in [transcriptomics technologies](https://en.wikipedia.org/wiki/Transcriptomics_technologies) have enabled the generation of single-cell transcriptomic data, they also presented new computational and analytical challenges. Bioinformaticians can use techniques from bulk RNA-seq for single-cell data. Still, many new computational approaches have had to be designed for this data type to facilitate a complete and detailed study of single-cell expression profiles.[[5]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-5)

## Experimental steps[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=2)]

There is so far no standardized technique to generate single-cell data: all methods must include cell isolation from the population, [lysate](https://en.wikipedia.org/wiki/Lysate) formation, amplification through [reverse transcription](https://en.wikipedia.org/wiki/Reverse_transcription) and quantification of expression levels. Common techniques for measuring expression are quantitative PCR or RNA-seq.[[6]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-6)

### Isolating single cells[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=3)]

[](https://en.wikipedia.org/wiki/File:Fluorescence_Assisted_Cell_Sorting_(FACS)_B2.jpg)

Fluorescence Assisted Cell Sorting workflow (FACS)

There are several methods available to isolate and amplify cells for single-cell analysis. Low throughput techniques are able to isolate hundreds of cells, are slow, and enable selection. These methods include:

* [Micropipetting](https://en.wikipedia.org/w/index.php?title=Micropipetting&action=edit&redlink=1)
* [Cytoplasmic aspiration](http://www.single-cell-analysis.com/tag/cytoplasmic-aspiration/)
* [Laser capture microdissection](https://en.wikipedia.org/wiki/Laser_capture_microdissection).

High-throughput methods are able to quickly isolate hundreds to tens of thousands of cells.[[7]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-7) Common techniques include:

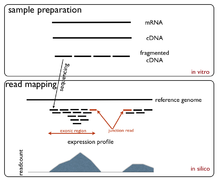
* [Fluorescence activated cell sorting](https://en.wikipedia.org/wiki/Flow_cytometry#Fluorescence-activated_cell_sorting_(FACS)) (FACS)
* [Microfluidic](https://en.wikipedia.org/wiki/Microfluidic) devices

Combining FACS with scRNA-seq has produced optimized protocols such as SORT-seq.[[8]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-8) A list of studies that utilized SORT-seq can be found here.[[9]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-9) Moreover, combining microfluidic devices with scRNA-seq has been optimized in 10x Genomics protocols.[[10]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-10)

### Quantitative PCR (qPCR)[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=4)]

To measure the level of expression of each transcript qPCR can be applied. Gene specific [primers](https://en.wikipedia.org/wiki/Primer_(molecular_biology)) are used to amplify the corresponding gene as with regular [PCR](https://en.wikipedia.org/wiki/Polymerase_chain_reaction) and as a result data is usually only obtained for sample sizes of less than 100 genes. The inclusion of [housekeeping genes](https://en.wikipedia.org/wiki/Housekeeping_genes), whose expression should be constant under the conditions, is used for normalisation. The most commonly used house keeping genes include [GAPDH](https://en.wikipedia.org/wiki/GAPDH) and α-[actin](https://en.wikipedia.org/wiki/Actin), although the reliability of normalisation through this process is questionable as there is evidence that the level of expression can vary significantly.[[11]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-11) [Fluorescent dyes](https://en.wikipedia.org/wiki/Fluorescent_dyes) are used as [reporter molecules](https://en.wikipedia.org/wiki/Reporter_gene) to detect the PCR product and monitor the progress of the amplification - the increase in fluorescence intensity is proportional to the [amplicon](https://en.wikipedia.org/wiki/Amplicon) concentration. A plot of fluorescence vs. cycle number is made and a threshold fluorescence level is used to find cycle number at which the plot reaches this value. The cycle number at this point is known as the threshold cycle (Ct) and is measured for each gene.[[12]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-12)

### Single-cell RNA-seq[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=5)]

[](https://en.wikipedia.org/wiki/File:RNA_Seq_Experiment.png)

RNA Seq Experiment

The [Single-cell](https://en.wikipedia.org/wiki/Single_cell_sequencing) [RNA-seq](https://en.wikipedia.org/wiki/RNA-Seq) technique converts a population of RNAs to a library of [cDNA](https://en.wikipedia.org/wiki/CDNA) fragments. These fragments are sequenced by high-throughput [next generation sequencing](https://en.wikipedia.org/wiki/Next_generation_sequencing) techniques and the reads are mapped back to the reference genome, providing a count of the number of reads associated with each gene.[[13]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-13)

Normalisation of RNA-seq data accounts for cell to cell variation in the efficiencies of the cDNA library formation and sequencing. One method relies on the use of [extrinsic](https://en.wikipedia.org/wiki/Extrinsic) RNA spike-ins (RNA sequences of known sequence and quantity) that are added in equal quantities to each cell [lysate](https://en.wikipedia.org/wiki/Lysate) and used to normalise read count by the number of reads mapped to spike-in [mRNA](https://en.wikipedia.org/wiki/Messenger_RNA).[[14]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-14)

Another control uses [unique molecular identifiers](https://en.wikipedia.org/wiki/Unique_molecular_identifiers) (UMIs)-short DNA sequences (6–10nt) that are added to each cDNA before amplification and act as a bar code for each cDNA molecule. Normalisation is achieved by using the count number of unique UMIs associated with each gene to account for differences in amplification efficiency.[[15]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-Quantitative_single-cell_RNA-seq_wi-15)

A combination of both spike-ins, UMIs and other approaches have been combined for more accurate normalisation.

### Considerations[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=6)]

A problem associated with single-cell data occurs in the form of zero inflated gene expression distributions, known as technical dropouts, that are common due to low mRNA concentrations of less-expressed genes that are not captured in the reverse transcription process. The percentage of mRNA molecules in the cell lysate that are detected is often only 10-20%.[[16]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-16)

When using RNA spike-ins for normalisation the assumption is made that the amplification and sequencing efficiencies for the [endogenous](https://en.wikipedia.org/wiki/Endogenous) and spike-in RNA are the same. Evidence suggests that this is not the case given fundamental differences in size and features, such as the lack of a [polyadenylated](https://en.wikipedia.org/wiki/Polyadenylation) tail in spike-ins and therefore shorter length.[[17]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-17) Additionally, normalisation using UMIs assumes the cDNA library is sequenced to saturation, which is not always the case.[[15]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-Quantitative_single-cell_RNA-seq_wi-15)

## Data analysis[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=7)]

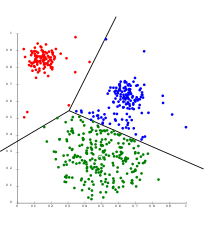
Insights based on single-cell data analysis assume that the input is a matrix of normalised gene expression counts, generated by the approaches outlined above, and can provide opportunities that are not obtainable by bulk.

Three main insights provided:[[18]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-Computational_and_analytical_challe-18)

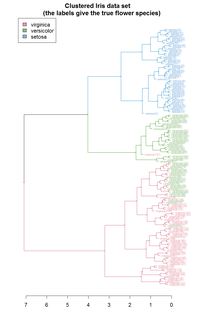
1. Identification and characterization of cell types and their spatial organisation in time
2. Inference of [gene regulatory networks](https://en.wikipedia.org/wiki/Gene_regulatory_networks) and their strength across individual cells
3. Classification of the [stochastic](https://en.wikipedia.org/wiki/Stochastic) component of transcription

The techniques outlined have been designed to help visualise and explore patterns in the data in order to facilitate the revelation of these three features.

### Clustering[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=8)]

[](https://en.wikipedia.org/wiki/File:KMeans-Gaussian-data.svg)

K-Means-Gaussian-data

[](https://en.wikipedia.org/wiki/File:Iris_dendrogram.png)

Iris dendrogram produced using a Hierarchical clustering algorithm

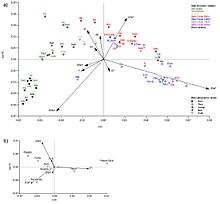
[Clustering](https://en.wikipedia.org/wiki/Subgroup_analysis) allows for the formation of subgroups in the cell population. Cells can be clustered by their transcriptomic profile in order to analyse the sub-population structure and identify rare cell types or cell subtypes. Alternatively, genes can be clustered by their expression states in order to identify covarying genes. A combination of both clustering approaches, known as [biclustering](https://en.wikipedia.org/wiki/Biclustering" \o "Biclustering), has been used to simultaneously cluster by genes and cells to find genes that behave similarly within cell clusters.[[19]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-19)

Clustering methods applied can be [K-means clustering](https://en.wikipedia.org/wiki/K-means_clustering), forming disjoint groups or [Hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering), forming nested partitions.

#### Biclustering**[**[**edit**](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=9)**]**

Biclustering provides several advantages by improving the resolution of clustering. Genes that are only informative to a subset of cells and are hence only expressed there can be identified through biclustering. Moreover, similarly behaving genes that differentiate one cell cluster from another can be identified using this method.[[20]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-20)

### Dimensionality reduction[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=10)]

[](https://en.wikipedia.org/wiki/File:PCA_of_Guinean_and_other_African_populations_Y_chromosome_haplogroup_frequencies.jpg)

PCA example of Guinean and other African populations Y chromosome haplogroup frequencies

[Dimensionality reduction](https://en.wikipedia.org/wiki/Dimensionality_reduction) algorithms such as [Principal component analysis](https://en.wikipedia.org/wiki/Principal_component_analysis) (PCA) and [t-SNE](https://en.wikipedia.org/wiki/T-SNE) can be used to simplify data for visualisation and pattern detection by transforming cells from a high to a lower [dimensional space](https://en.wikipedia.org/wiki/Dimensional_space). The result of this method produces graphs with each cell as a point in a 2-D or 3-D space. Dimensionality reduction is frequently used before clustering as cells in high dimensions can wrongly appear to be close due to distance metrics behaving non-intuitively.[[21]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-21)

#### Principal component analysis**[**[**edit**](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=11)**]**

The most frequently used technique is PCA, which identifies the directions of largest [variance](https://en.wikipedia.org/wiki/Variance) [principal components](https://en.wikipedia.org/wiki/Principal_components) and transforms the data so that the first principal component has the largest possible variance, and successive principle components in turn each have the highest variance possible while remaining orthogonal to the preceding components. The contribution each gene makes to each component is used to infer which genes are contributing the most to variance in the population and are involved in differentiating different subpopulations.[[22]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-22)

### Differential expression[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=12)]

Detecting differences in gene expression level between two populations is used both single-cell and bulk transcriptomic data. Specialised methods have been designed for single-cell data that considers single cell features such as technical dropouts and shape of the distribution e.g. [Bimodal](https://en.wikipedia.org/wiki/Bimodal) vs. [unimodal](https://en.wikipedia.org/wiki/Unimodal).[[23]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-23)

#### Gene ontology enrichment**[**[**edit**](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=13)**]**

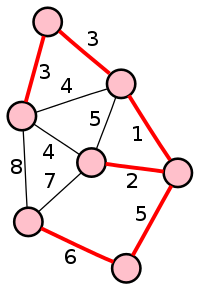
[Gene ontology](https://en.wikipedia.org/wiki/Gene_ontology) terms describe gene functions and the relationships between those functions into three classes:

1. Molecular function
2. Cellular component
3. Biological process

[Gene Ontology (GO) term enrichment](https://en.wikipedia.org/w/index.php?title=Gene_Ontology_(GO)_term_enrichment&action=edit&redlink=1) is a technique used to identify which GO terms are over-represented or under-represented in a given set of genes. In single-cell analysis input list of genes of interest can be selected based on differentially expressed genes or groups of genes generated from biclustering. The number of genes annotated to a GO term in the input list is normalised against the number of genes annotated to a GO term in the background set of all genes in genome to determine statistical significance.[[24]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-hag1-24)

### Pseudotemporal ordering[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=14)]

*Main article:*[*Trajectory inference*](https://en.wikipedia.org/wiki/Trajectory_inference)

[](https://en.wikipedia.org/wiki/File:7n_graph_with_minimal_spanning_tree.svg)

Graph with minimal spanning tree

Pseudo-temporal ordering (or trajectory inference) is a technique that aims to infer gene expression dynamics from snapshot single-cell data. The method tries to order the cells in such a way that similar cells are closely positioned to each other. This trajectory of cells can be linear, but can also bifurcate or follow more complex graph structures. The trajectory, therefore, enables the inference of gene expression dynamics and the ordering of cells by their progression through differentiation or response to external stimuli. The method relies on the assumptions that the cells follow the same path through the process of interest and that their transcriptional state correlates to their progression. The algorithm can be applied to both mixed populations and temporal samples.

More than 50 methods for pseudo-temporal ordering have been developed, and each has its own requirements for prior information (such as starting cells or time course data), detectable topologies, and methodology.[[25]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-25) An example algorithm is the Monocle algorithm[[26]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-26) that carries out dimensionality reduction of the data, builds a [minimal spanning tree](https://en.wikipedia.org/wiki/Minimal_spanning_tree) using the transformed data, orders cells in pseudo-time by following the longest connected path of the tree and consequently labels cells by type. Another example is the diffusion pseudotime (DPT) algorithm,[[24]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-hag1-24) which uses a diffusion map and diffusion process. Another class of methods such as MARGARET [[27]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-27) employ graph partitioning for capturing complex trajectory topologies such as disconnected and multifurcating trajectories.

### Network inference[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=15)]

Gene regulatory network inference is a technique that aims to construct a network, shown as a graph, in which the [nodes](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) represent the genes and edges indicate co-regulatory interactions. The method relies on the assumption that a strong statistical relationship between the expression of genes is an indication of a potential functional relationship.[[28]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-28) The most commonly used method to measure the strength of a statistical relationship is [correlation](https://en.wikipedia.org/wiki/Correlation). However, correlation fails to identify [non-linear](https://en.wikipedia.org/wiki/Non-linear) relationships and [mutual information](https://en.wikipedia.org/wiki/Mutual_information) is used as an alternative. Gene clusters linked in a network signify genes that undergo coordinated changes in expression.[[29]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-29)

### Integration[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=16)]

The presence or strength of technical effects and the types of cells observed often differ in single-cell transcriptomics datasets generated using different experimental protocols and under different conditions. This difference results in strong [batch effects](https://en.wikipedia.org/wiki/Batch_effects) that may bias the findings of statistical methods applied across batches, particularly in the presence of [confounding](https://en.wikipedia.org/wiki/Confounding).[[30]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-30) As a result of the aforementioned properties of single-cell transcriptomic data, batch correction methods developed for bulk sequencing data were observed to perform poorly. Consequently, researchers developed statistical methods to correct for batch effects that are robust to the properties of single-cell transcriptomic data to integrate data from different sources or experimental batches. Laleh Haghverdi performed foundational work in formulating the use of mutual nearest neighbors between each batch to define batch correction vectors.[[31]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-31) With these vectors, you can merge datasets that each include at least one shared cell type. An orthogonal approach involves the projection of each dataset onto a shared low-dimensional space using [canonical correlation analysis](https://en.wikipedia.org/wiki/Canonical_correlation_analysis).[[32]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-32) Mutual nearest neighbors and canonical correlation analysis have also been combined to define integration "anchors" comprising reference cells in one dataset, to which query cells in another dataset are normalized.[[33]](https://en.wikipedia.org/wiki/Single-cell_transcriptomics#cite_note-33)

## See also[[edit](https://en.wikipedia.org/w/index.php?title=Single-cell_transcriptomics&action=edit&section=17)]

* [RNA-Seq](https://en.wikipedia.org/wiki/RNA-Seq)
* [Single-cell analysis](https://en.wikipedia.org/wiki/Single-cell_analysis)
* [Single-cell sequencing](https://en.wikipedia.org/wiki/Single_cell_sequencing)
* [Transcriptome](https://en.wikipedia.org/wiki/Transcriptome)
* [Transcriptomics](https://en.wikipedia.org/wiki/Transcriptomics)