HSE University
Faculty of Computer science

«Анализ данных NGS»

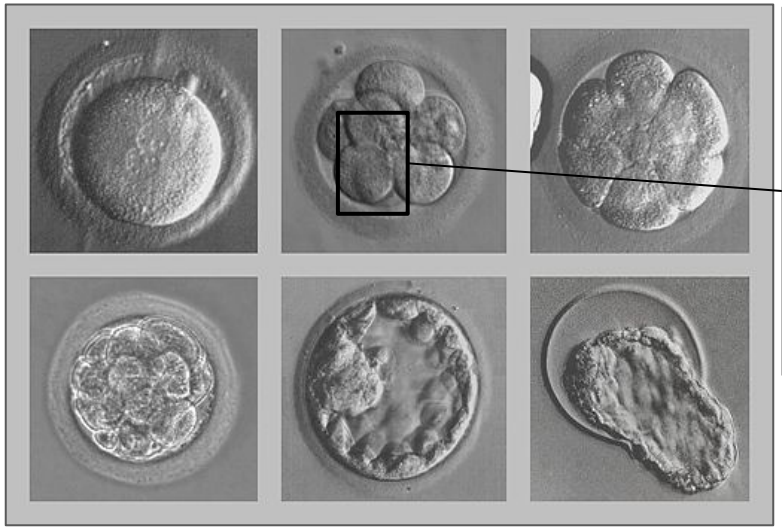
Лекция #5

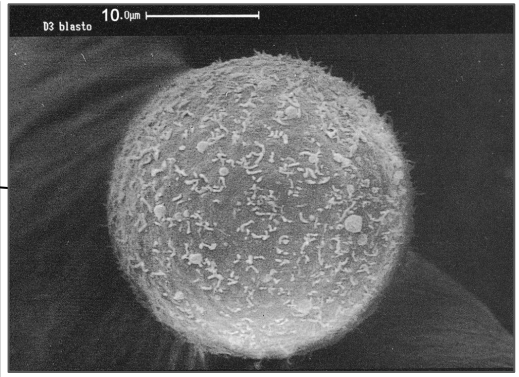
scRNA-Seq

Серёжа Исаев

аспирант **MedUni Vienna**

Tang et al., 2009 — первая работа

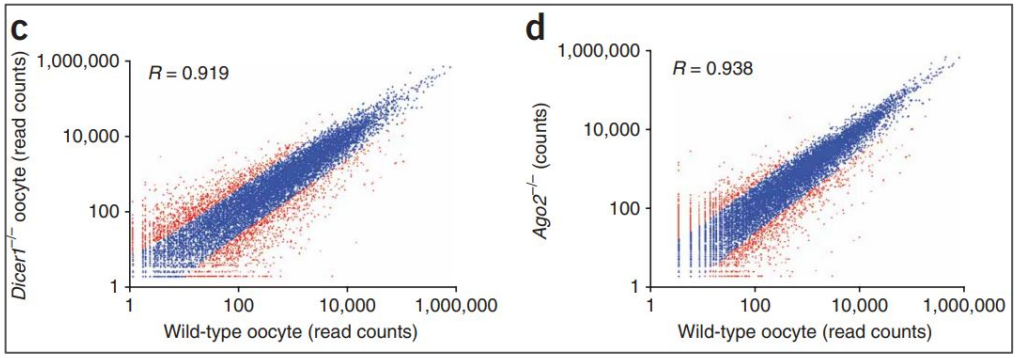




Бластомер. Источник:
https://www.ehd.org/

Оплодотворённая яйцеклетка, восьмиклеточная стадия, стадия адгезии, морула, бластоцист, зона вылупления.
Источник: http://nobelprize.org/

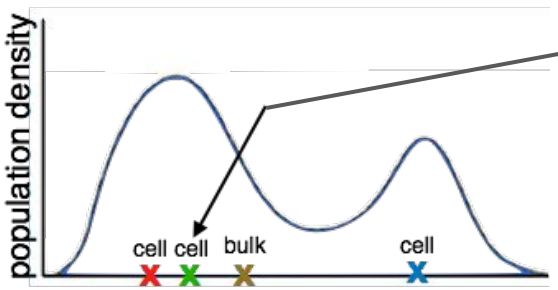
Сравнение профилей экспрессии нормального бластомера и бластомера с нокаутами из Tang et al. 2009



	Bulk RNA-Seq	scRNA-Seq
Начало	2008	2009
Экспрессия	средний уровень экспрессии	распределение уровней экспрессии
Количество транскриптов	~15-20 000 на образец	~200-10 000 на клетку
% Транскриптома	80-95%	10-50%

Bulk RNA-Seq

	Exp
Gene1	100
Gene2	5.5
...	
Gene N	0.5



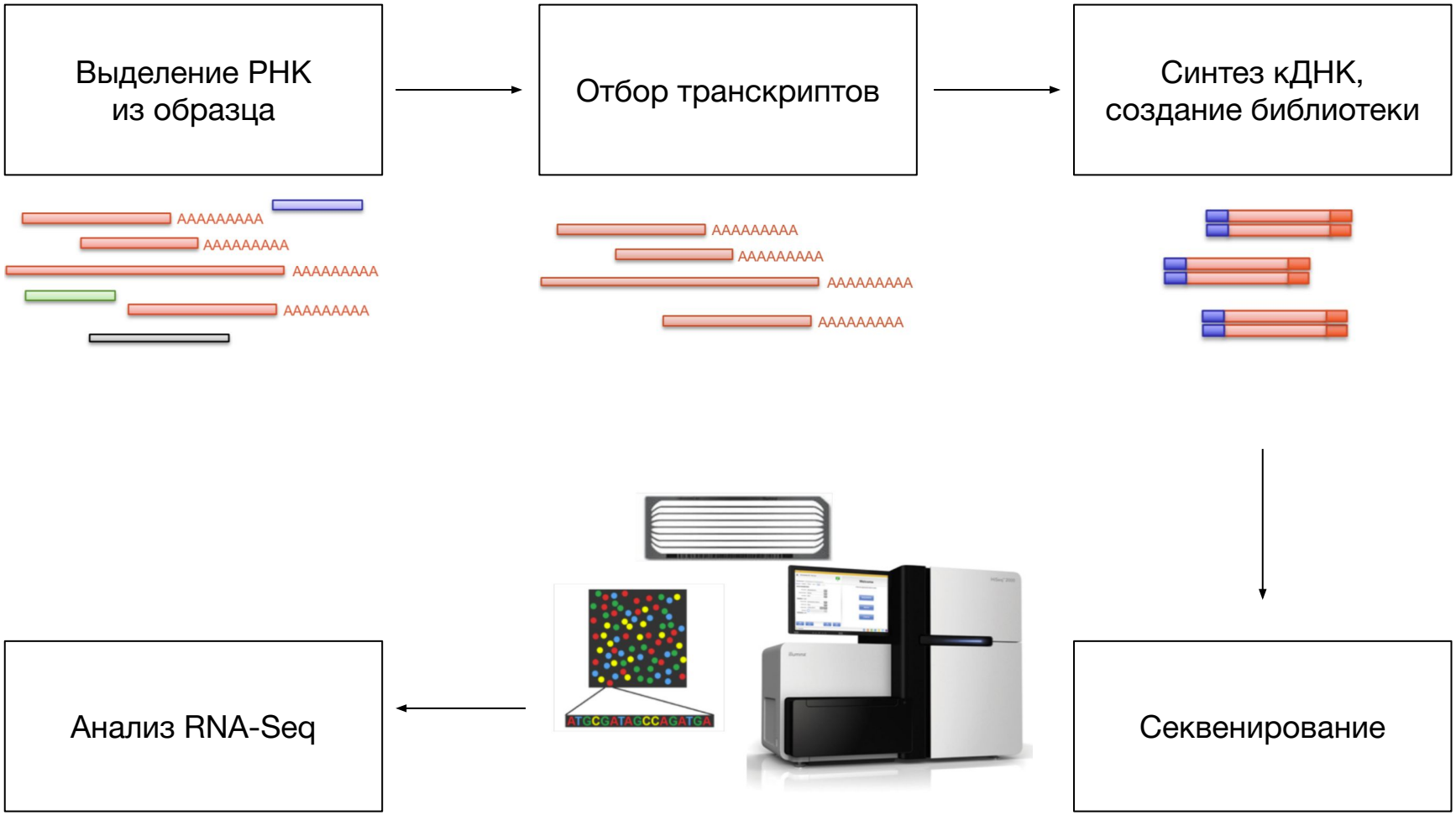
scRNA-Seq

	Cell1	Cell2	...	Cell K
Gene1	3	0		2
Gene2	0	2		1
...				
Gene N	0	13		2

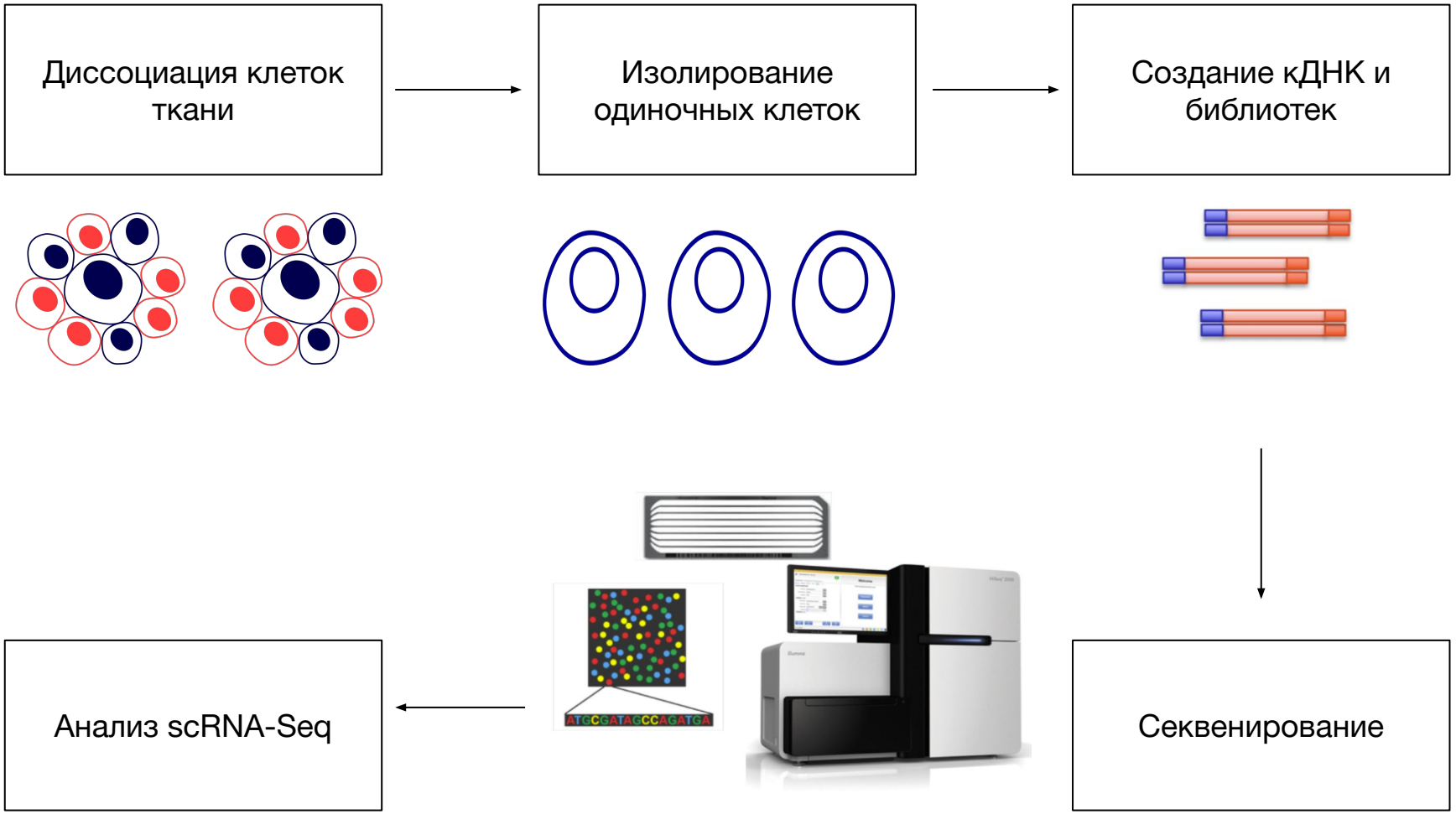
Планирование эксперимента

Сначала задайте вопросы — **какое именно явление я хочу изучить?** и **какая у меня гипотеза?** — и только потом ставьте эксперимент. Это касается не только scRNA-Seq, но и вообще любых исследований.

Общая схема эксперимента RNA-Seq



Общая схема эксперимента scRNA-Seq

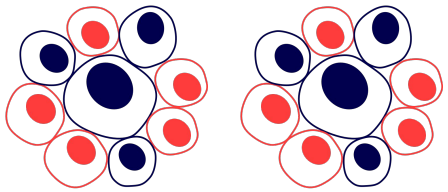


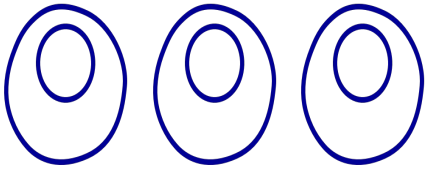
Общая схема эксперимента scRNA-Seq

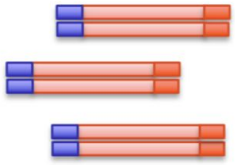
Диссоциация клеток ткани

Изолирование одиночных клеток

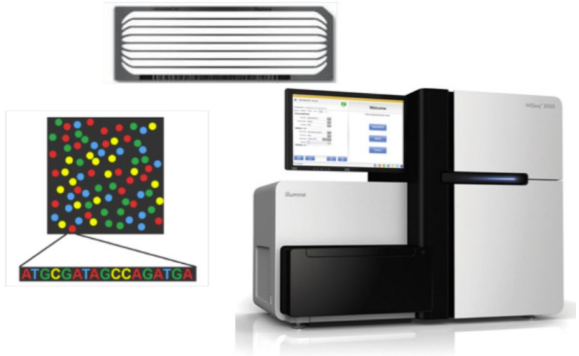
Создание кДНК и библиотек







Секвенирование



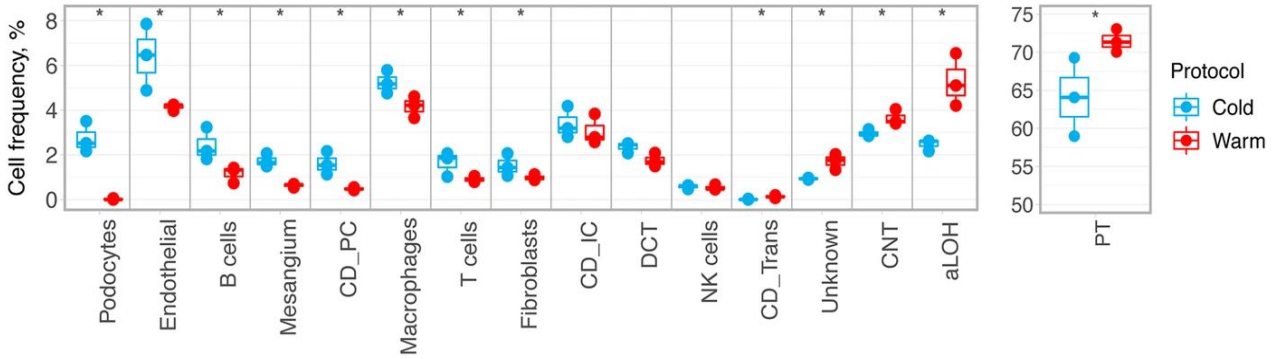
Анализ scRNA-Seq

Диссоциация клеток

В основном клетки в изучаемых тканях находятся в “сцепленном” состоянии (они соединены при помощи молекул адгезии и т. п.). Для того, чтобы их “расцепить”, необходимо провести диссоциацию ткани:

- диссоциация при нагревании (напр., Multi-tissue dissociation kit 2) — может вызвать активацию транскрипции генов теплового шока,
- диссоциация на холоду (с использованием протеазы *Bacillus Licheniformis*).

Сравнение методов диссоциации ткани из Denisenko et al., 2020. Некоторые клеточные типы не детектируются при диссоциации “горячим” методом



Создание клеточного атласа всего организма

Диссоциация ткани — это один из самых важных шагов при пробоподготовке scRNA-Seq. Во время этой процедуры клетки могут

1. умереть (и тогда мы увидим смещённый клеточный состав),
2. изменить свой экспрессионный профиль (и тогда мы увидим тот же клеточный состав, но не в нативном состоянии),
3. диссоциировать неполностью (и тогда мы увидим большое количество дублетов).

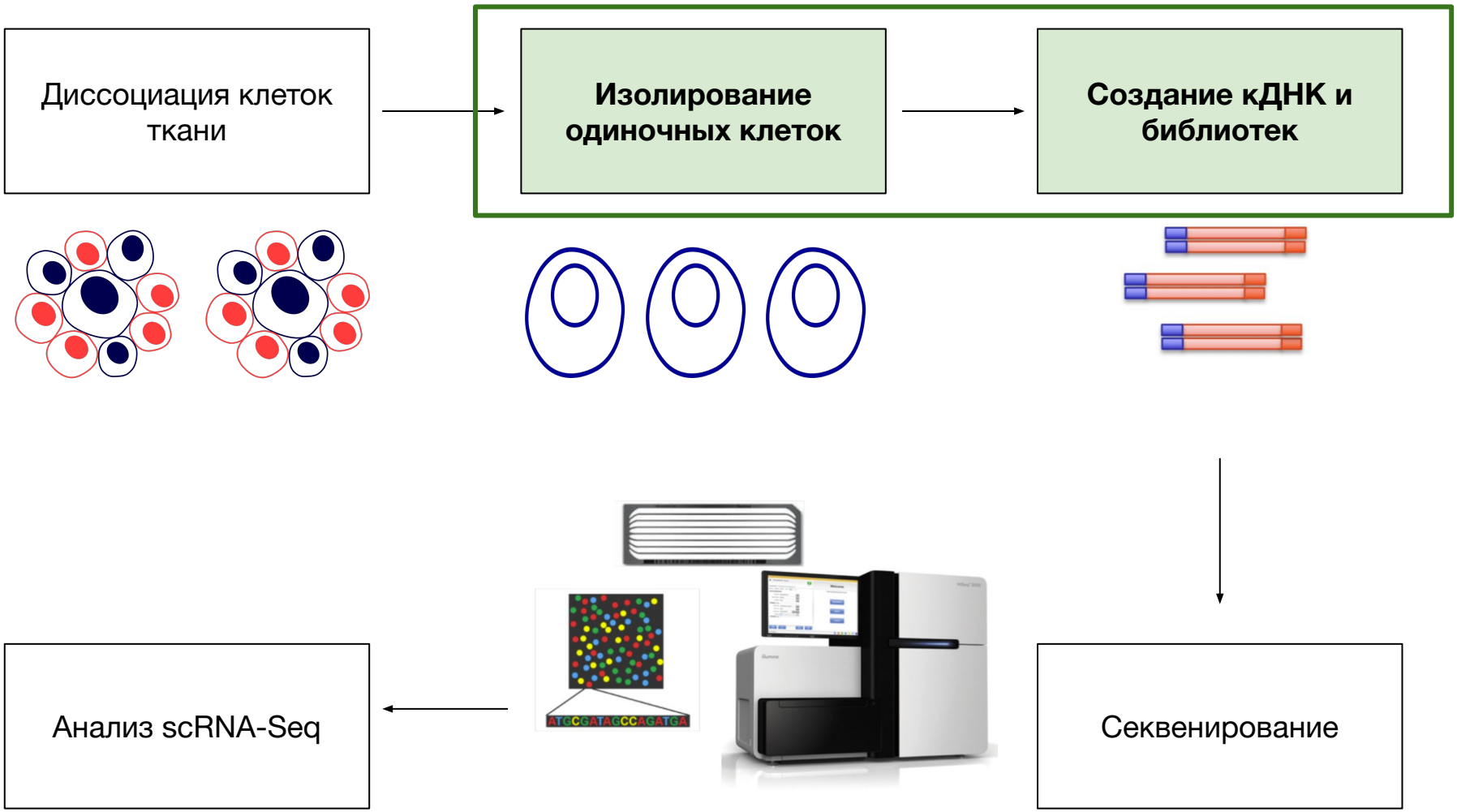
Для каждой ткани используется собственный протокол диссоциации. И это является очень большой проблемой для создания пан-тканевого клеточного атласа.

scRNA-Seq vs. snRNA-Seq

- Профилирование ядер из крупных клеток (> 40 мкм), которые не проходят через микрофлюидику
- Позволяет профилировать отдельные ядра, выделенные из замороженных тканей, отделяя получение ткани от немедленной обработки образца
- snRNA-Seq может также обрабатывать образцы, которые не могут быть успешно диссоциированы, даже если они свежие, из-за хрупкости клеток

Ядра имеют меньшее количество мРНК по сравнению с клетками, и их сложнее обогатить для конкретных интересующих типов клеток.

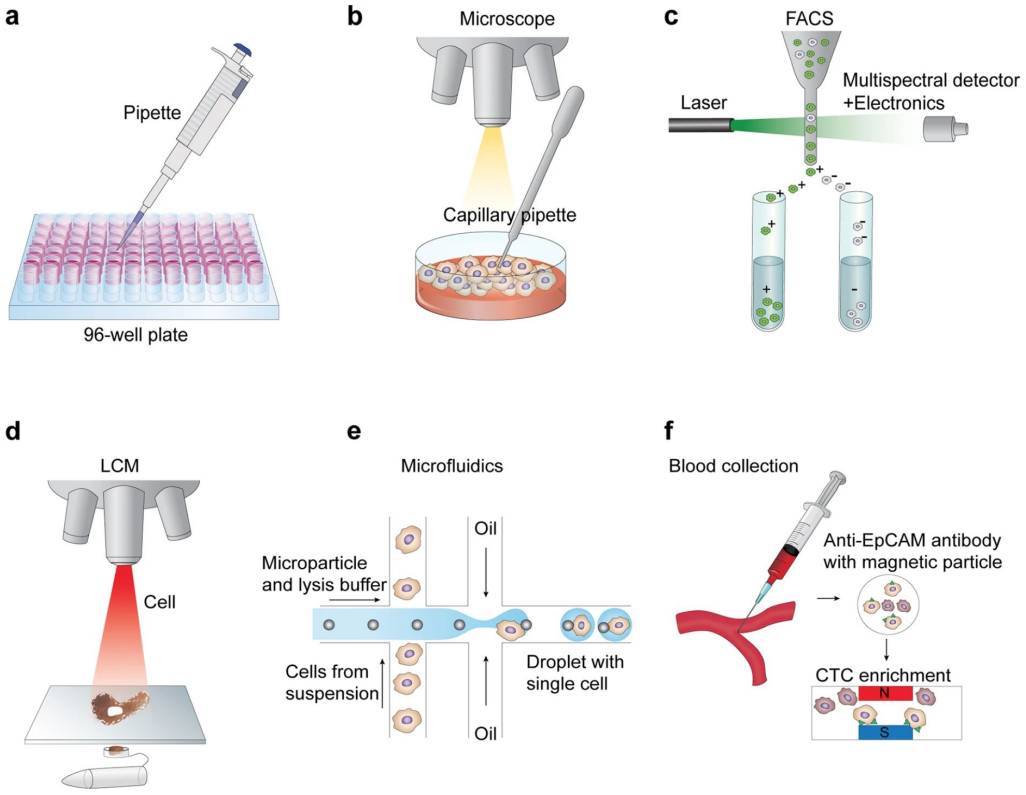
Общая схема эксперимента scRNA-Seq



Изолирование одиночных клеток

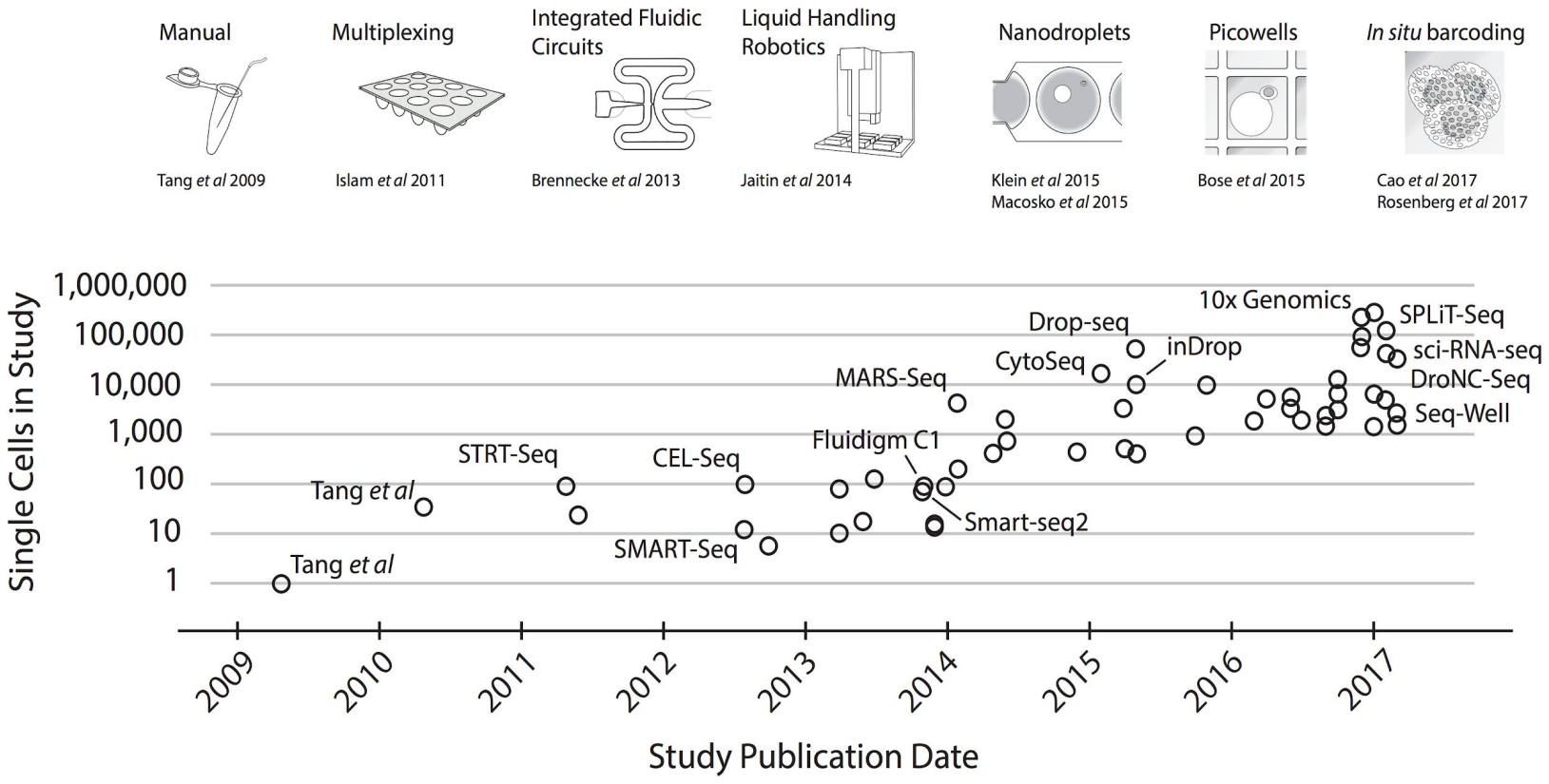
Существует множество различных способов изолировать одиночные клетки

Как правило, стадия изолирования одиночных клеток очень тесно связана с дальнейшими стадиями подготовки библиотек, поэтому рассмотрим их вместе



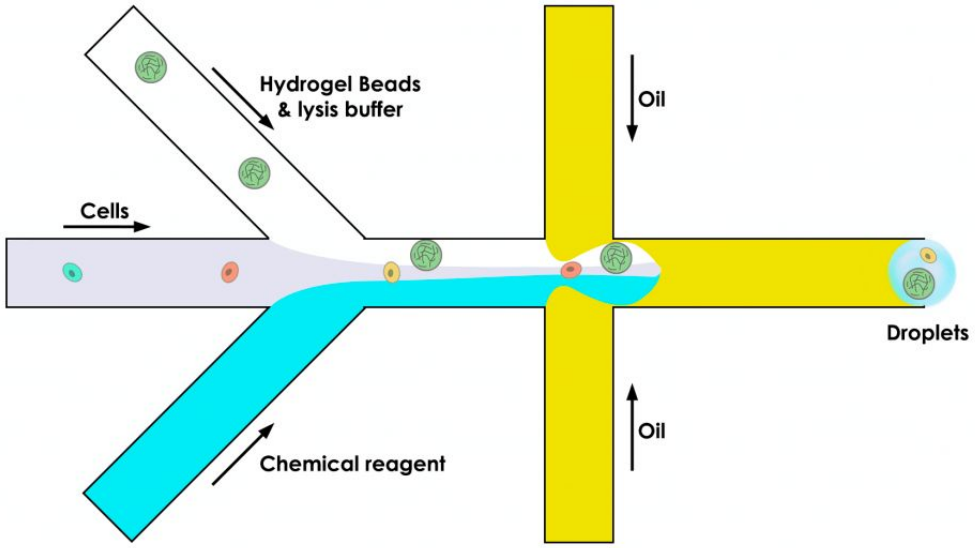
Hwang et al. 2018

История развития методов scRNA-Seq



Droplet-based (капельные) методы

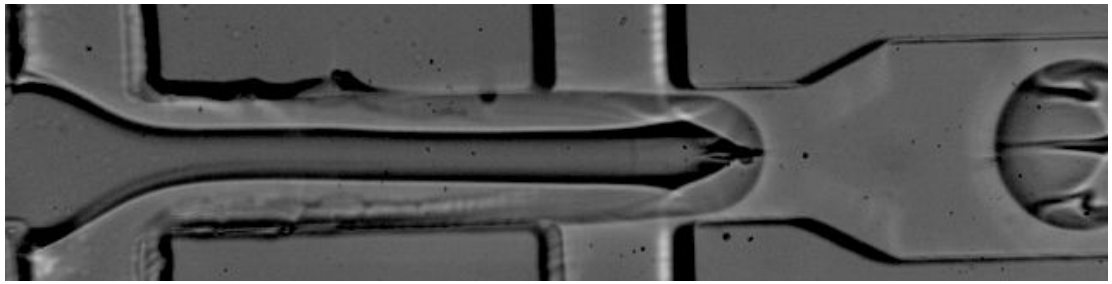
Капельные методы основаны на том, что клетки изолируются друг от друга, поступая по капиллярам в масляную фракцию и образуя там отдельные компартменты, содержащие необходимые реагенты и одну клетку



Источник: http://mccarrolllab.org/dropseq/

Droplet-based (капельные) методы

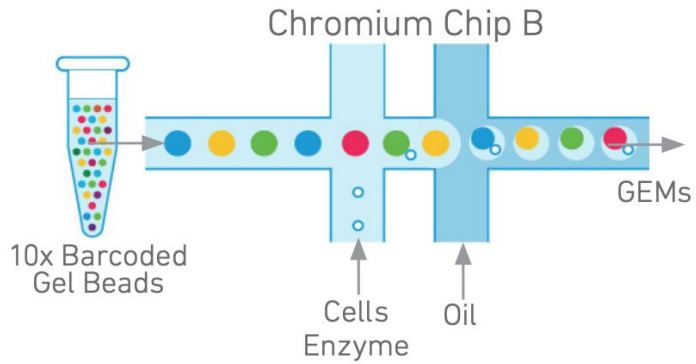
Капельные методы основаны на том, что клетки изолируются друг от друга, поступая по капиллярам в масляную фракцию и образуя там отдельные компартменты, содержащие необходимые реагенты и одну клетку



Источник: https://www.elveflow.com/microfluidic-reviews/droplet-digital-microfluidics/drop-seq/#_ftn4

10x Chromium

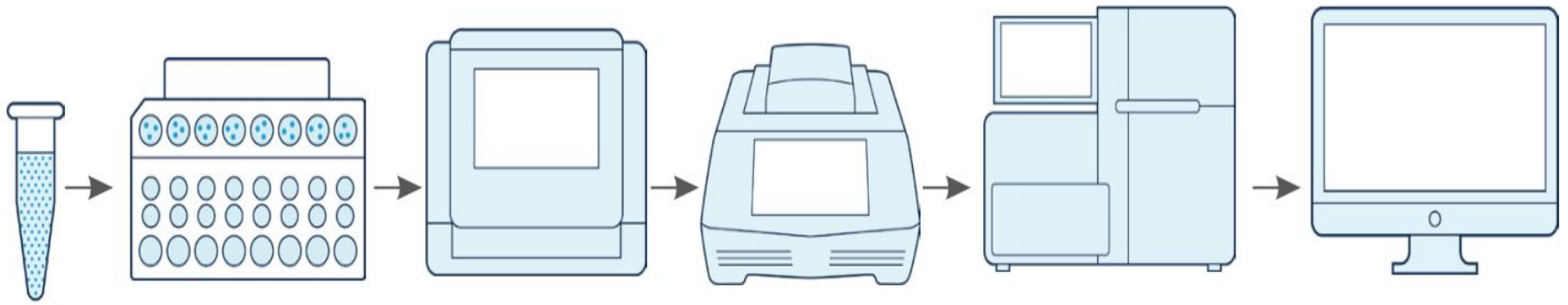
Контроллер 10x Chromium является сейчас одной из самых популярных платформ для создания библиотек scRNA-Seq



Источник: 10xgenomics.com

Процесс пробоподготовки

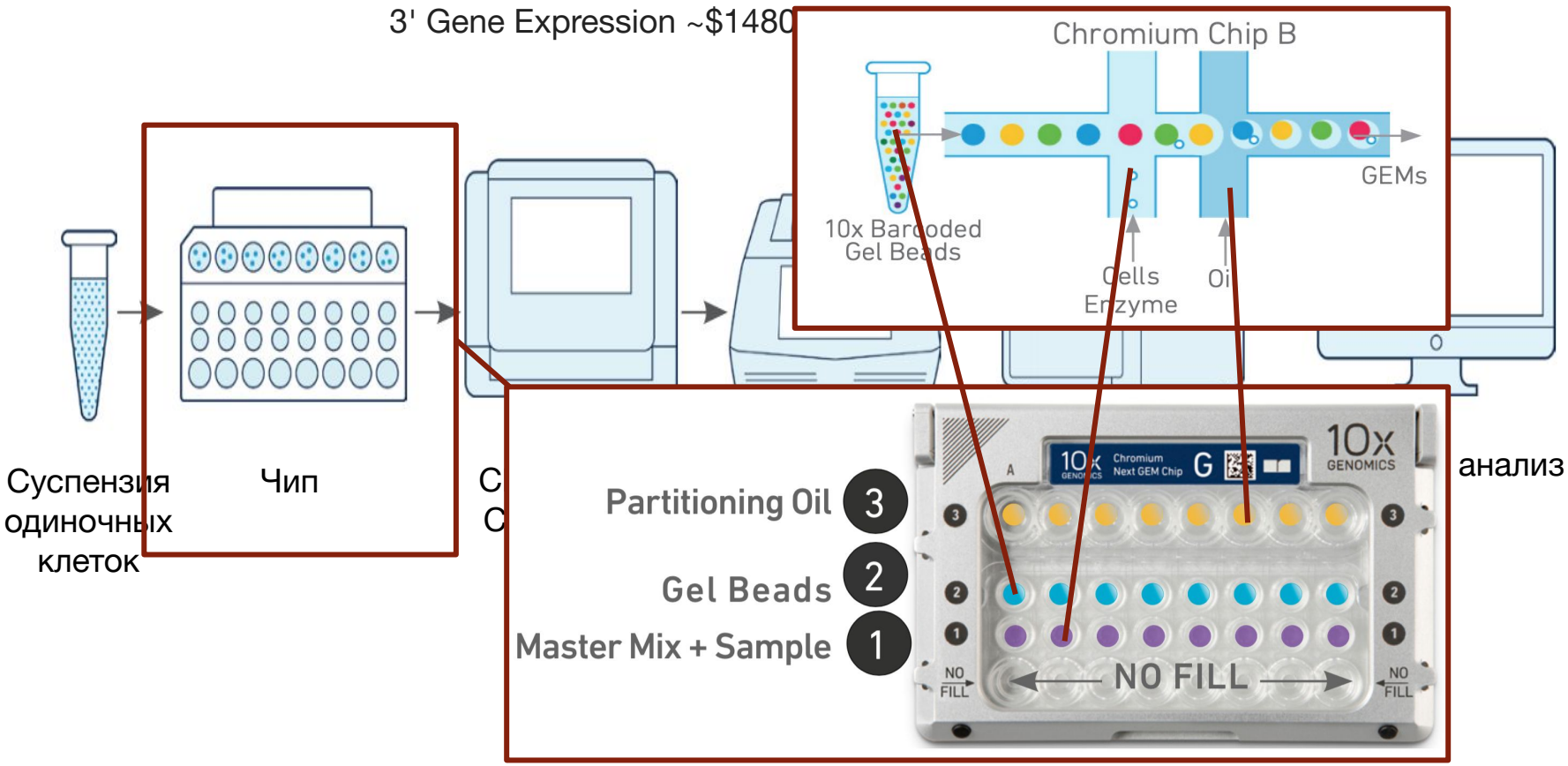
3' Gene Expression ~$1480/образец (WES ~150$/образец)



Суспензия одиночных клеток | Чип | Chromium Controller | Post GEM-RT Cleanup & cDNA Amplification | Секвенирование | Расчет и анализ

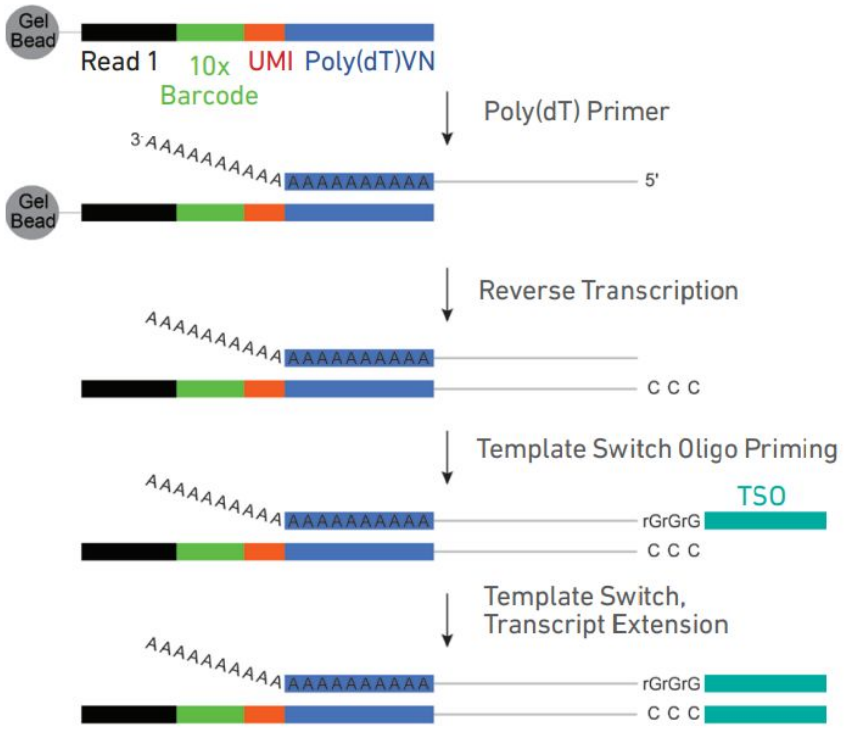
Источник: 10xgenomics.com

Процесс пробоподготовки



Источник: 10xgenomics.com

10x v3 3’ Gel Beads

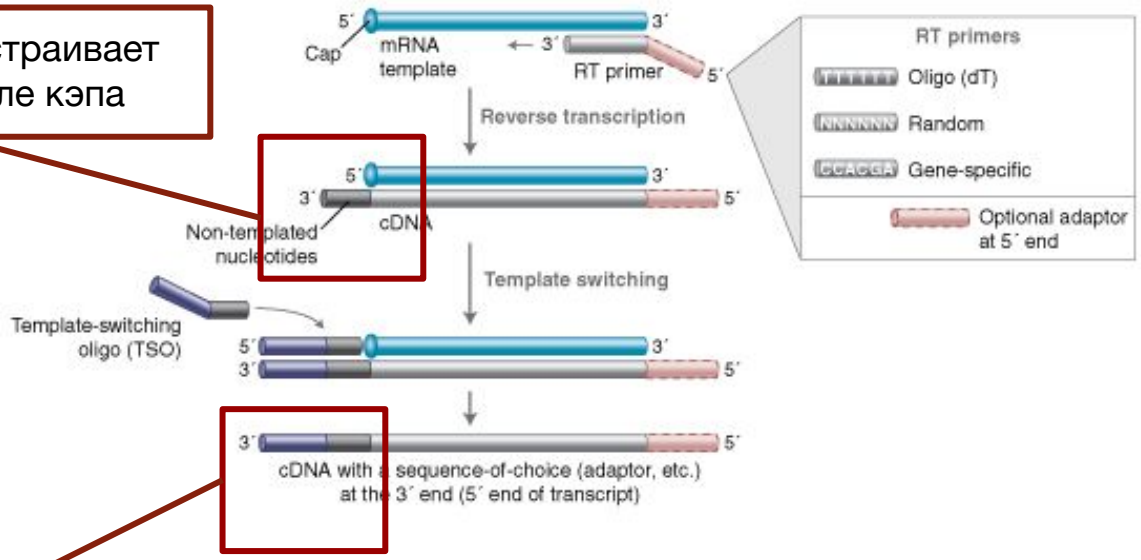


К каждому шарику прикреплён уникальный праймер, который состоит из

1. Праймера Illumina TruSeq Read 1,
2. **Баркода** (последовательности, которая одинакова у всех праймеров данного шарика, однако различается между всеми шариками),
3. **UMI** (последовательности, которая уникальна для всех праймеров данного шарика, но может повторяться между шариками),
4. Poly(dT)-последовательности.

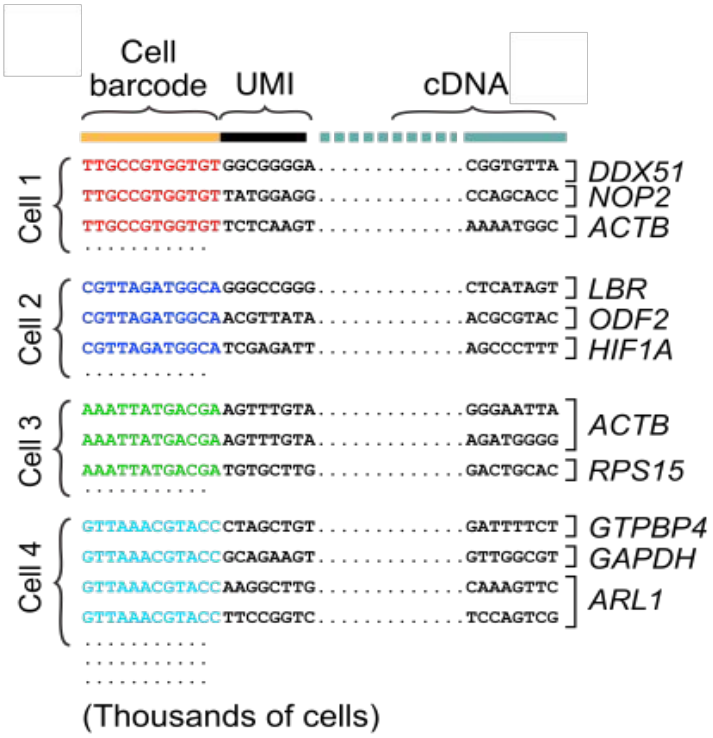
Template Switch Oligo (TSO)

Полимераза достраивает CCC строго после кэпа



После гибридизации TSO и CCC происходит смена матрицы синтеза

Баркоды и UMI



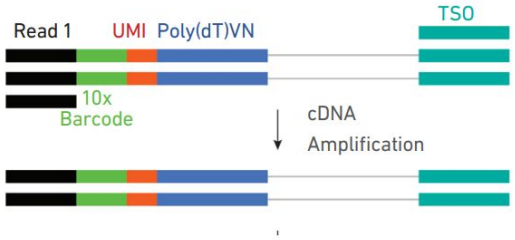
Последовательность баркода, как и UMI, будет в итоге отсеквенирована

Последовательность баркода определяет клетку, к которой мы отнесём данный транскрипт

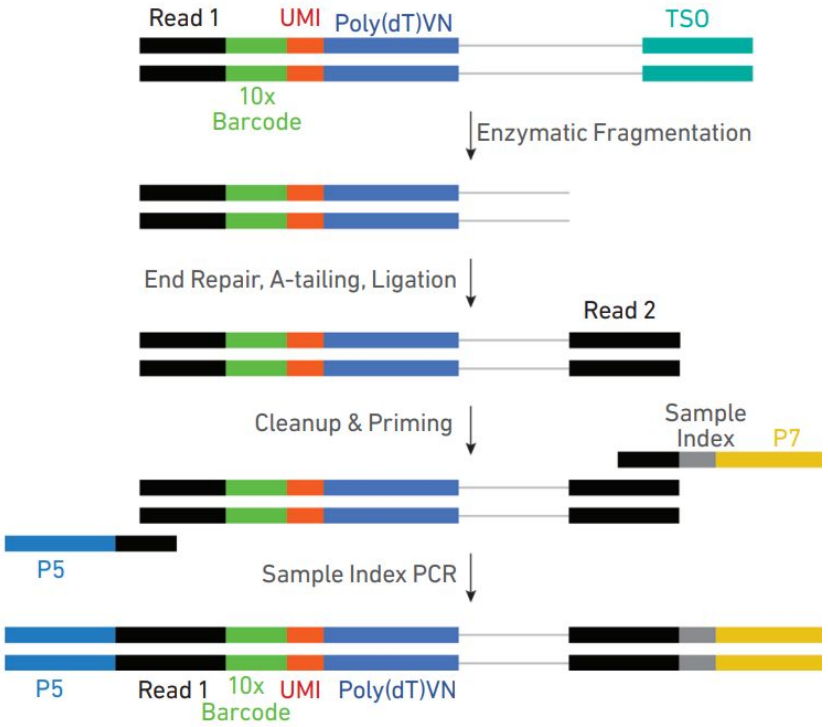
Последовательность UMI позволяет определить ПЦР-дубликаты (amplification bias — это достаточно большая проблема в случае, когда у нас мало РНК)

Подготовка библиотек для секвенирования

ПЦР



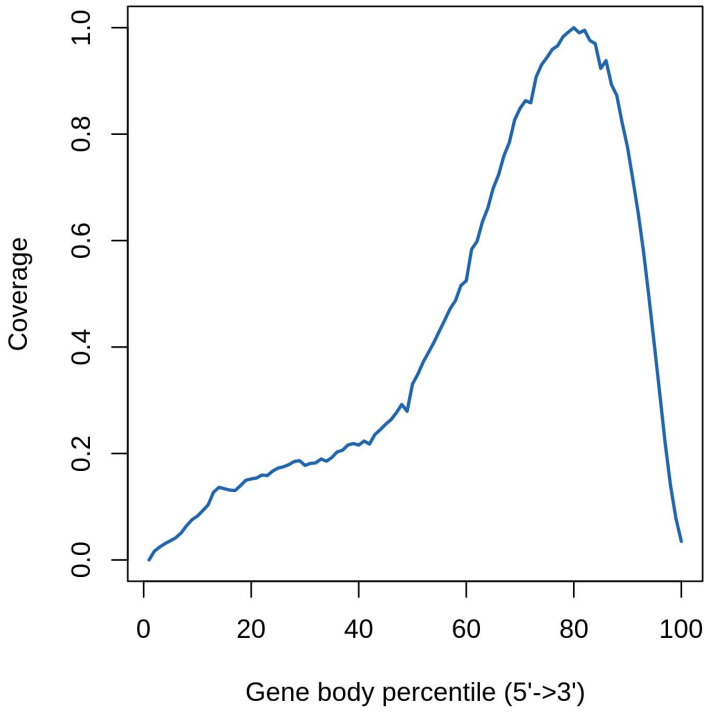
Подготовка библиотеки



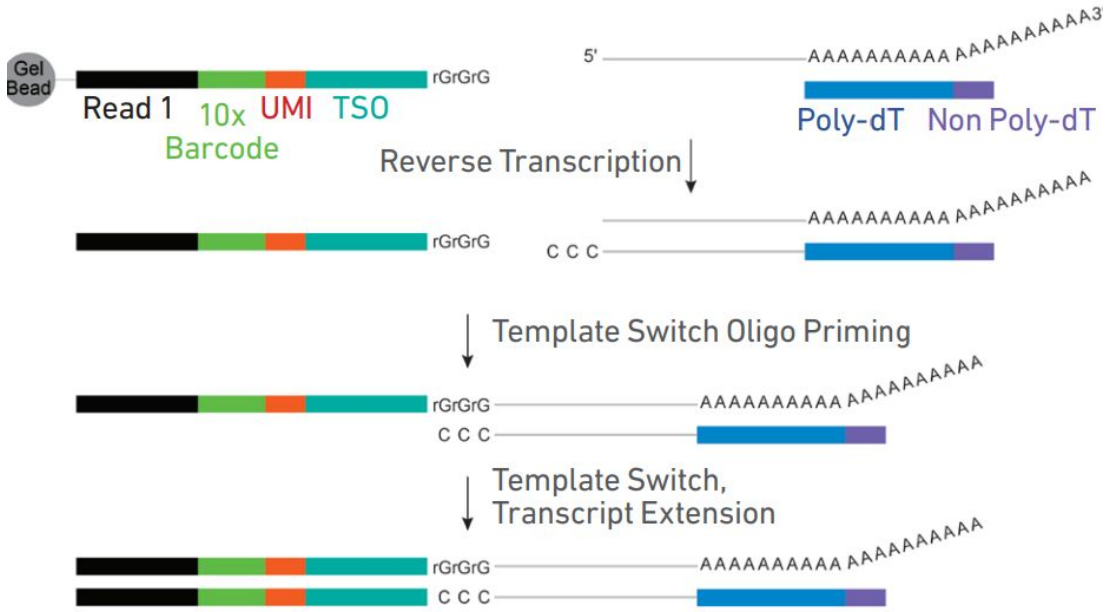
Покрытие транскрипта ридами (3’-протокол)

Несмотря на TSO, у нас присутствует стадия фрагментации кДНК. Более того, после этой стадии пришивается только один из концов, праймер для секвенирования другого конца всегда зафиксирован на 3’-конце последовательности. Как итог у нас возникает неравномерность в покрытии транскрипта ридами.

Источник:
https://liulab-dfci.github.io/MAESTRO/example/RNA_infrastructure_10x/RNA_infrastructure_10x.html



10x v3 5’ Gel Beads



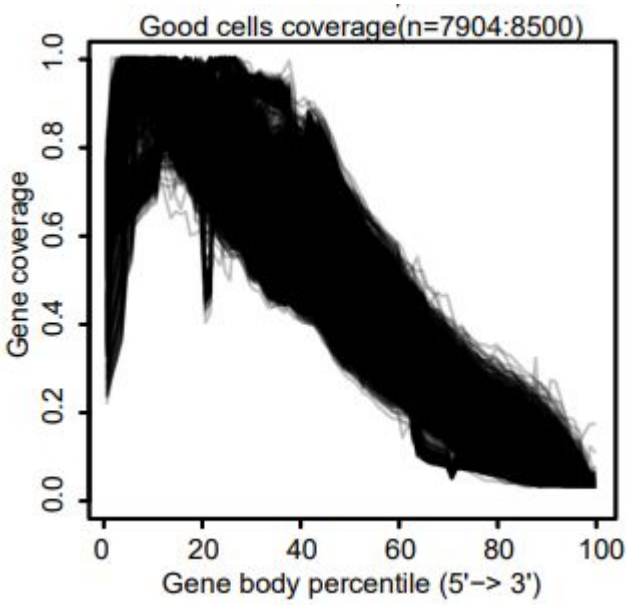
Бывает так (когда? зачем?), что нам катастрофически необходимо хорошее покрытие на 5’-конце транскриптов

Элегантное решение данной проблемы — это праймер не из баркода-UMI-oligo(dT), а из баркода-UMI-TSO

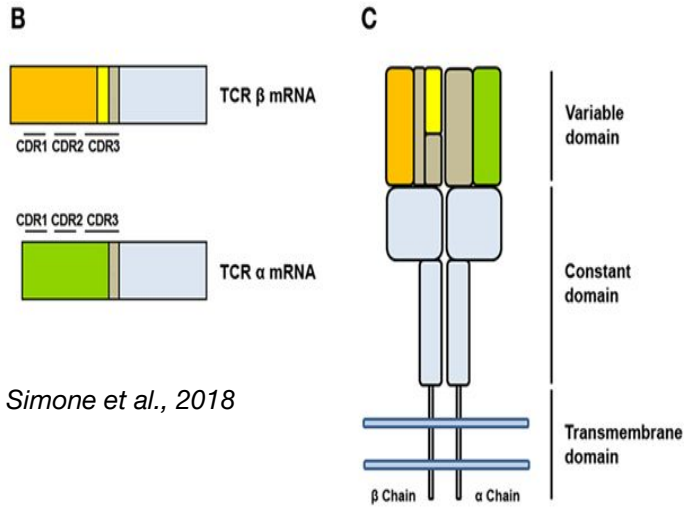
Покрытие транскрипта ридами (5’-протокол)

В случае с 5’-протоколом также наблюдается неравномерность покрытия транскрипта прочтениями, однако уже в сторону 5’-конца транскрипта

Обогащение 5’-концов транскрипта необходимо тогда, когда нам нужно дополнительно обогатить библиотеку V(D)J-фрагментами для определения Т- или В-клеточных рецепторов иммунных клеток



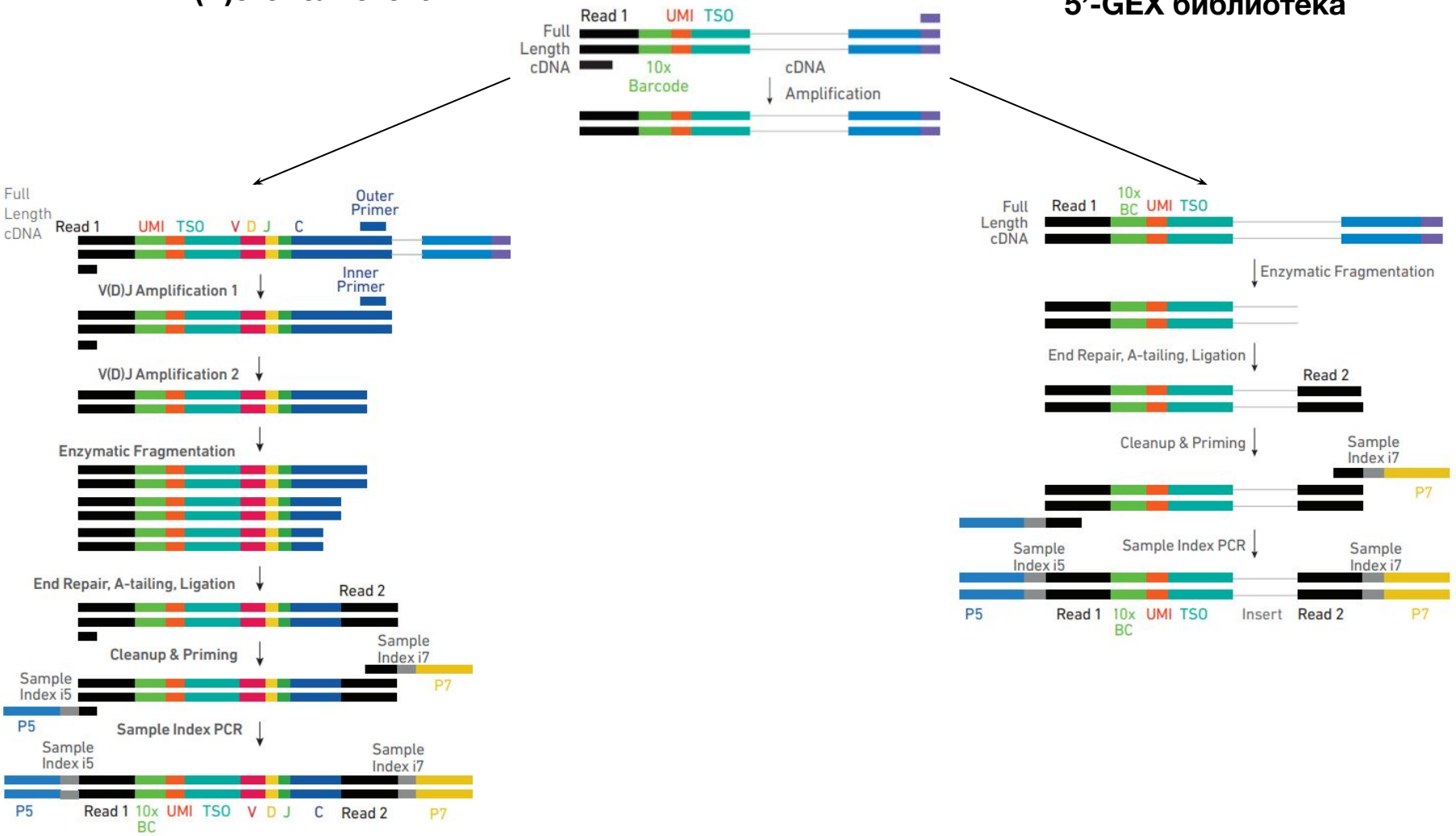
Abugessaisa et al., 2018



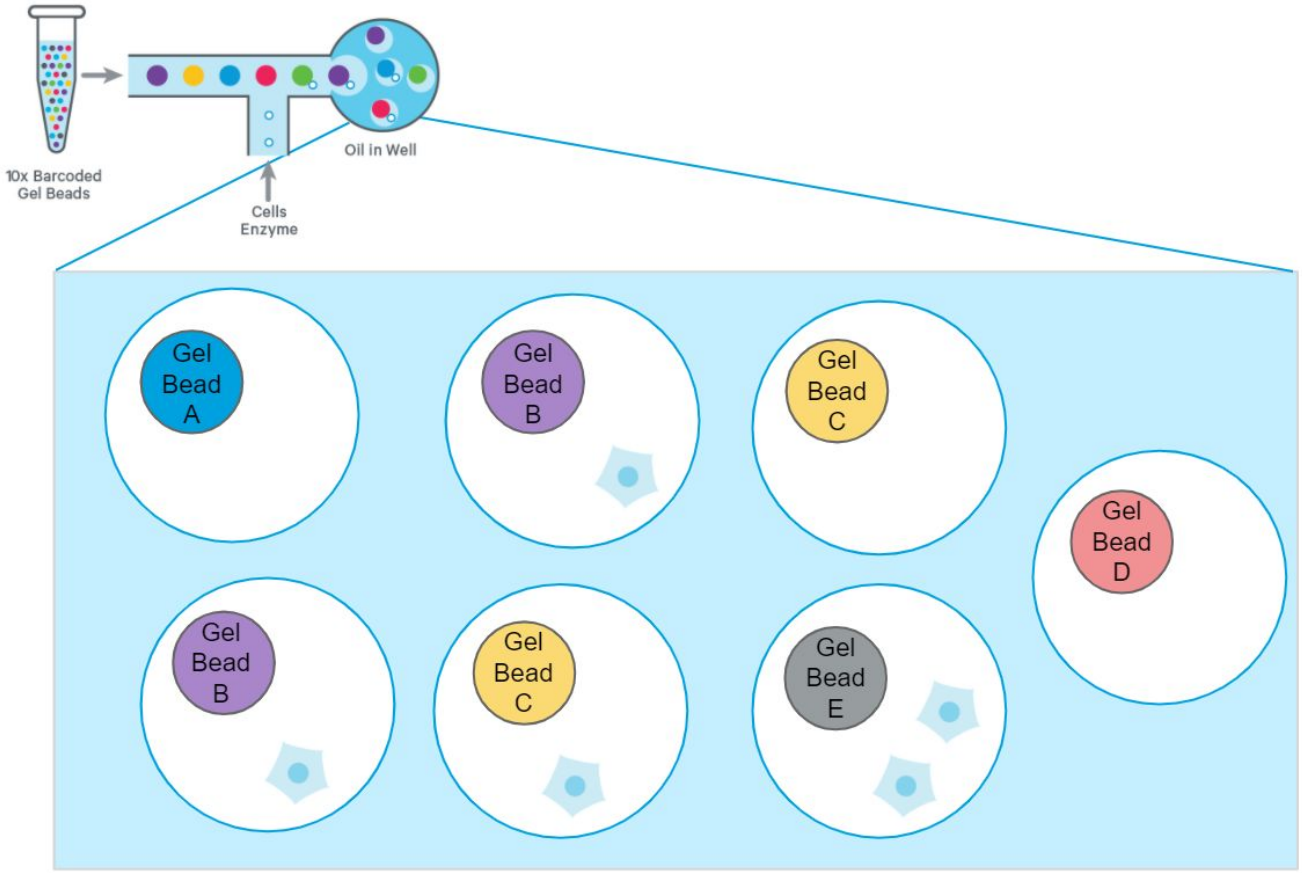
Simone et al., 2018

V(D)J-библиотека

5’-GEX библиотека

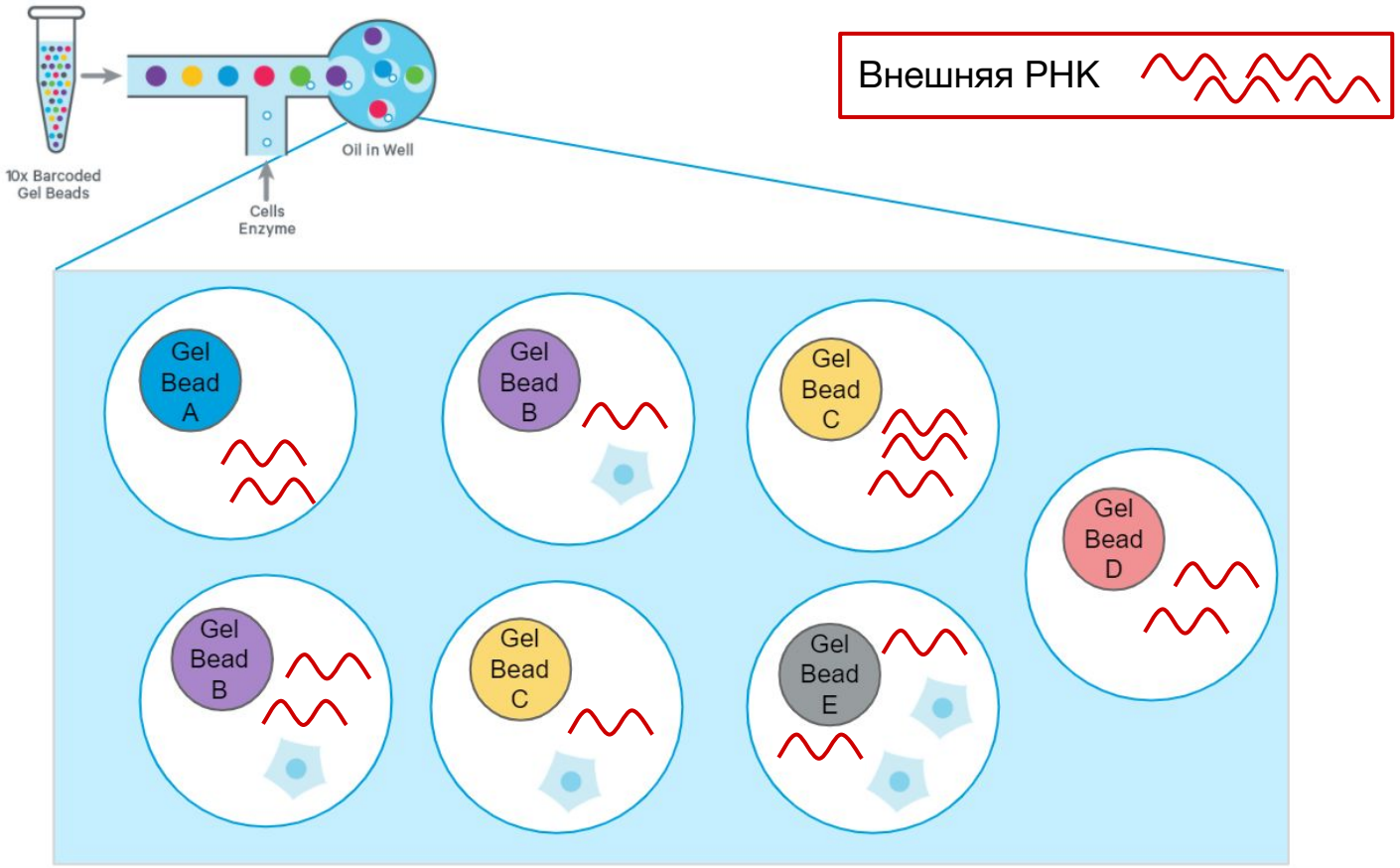


Пустые капли и дублеты



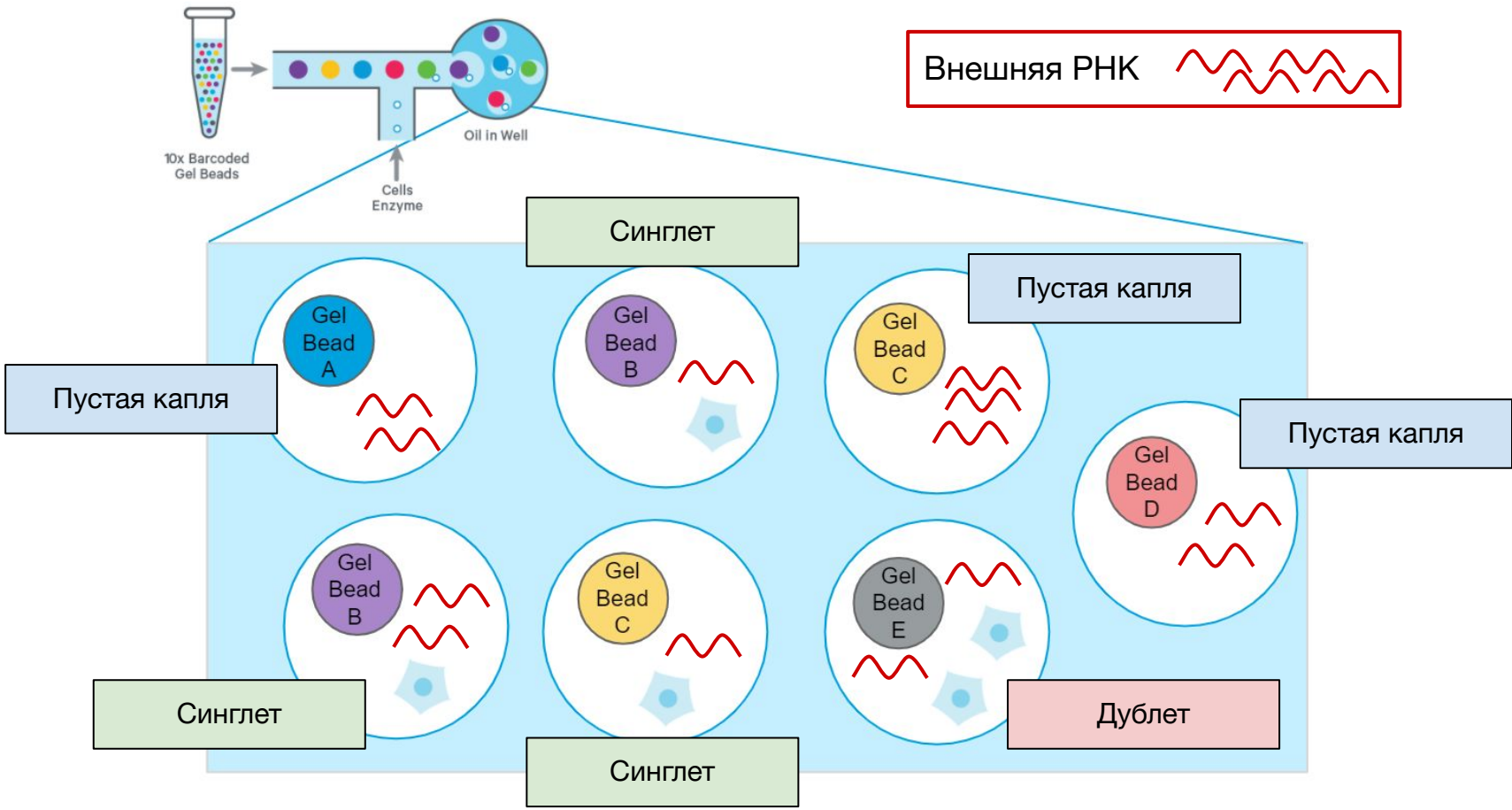
Источник: 10xgenomics.com

Пустые капли и дублеты



Источник: 10xgenomics.com

Пустые капли и дублеты



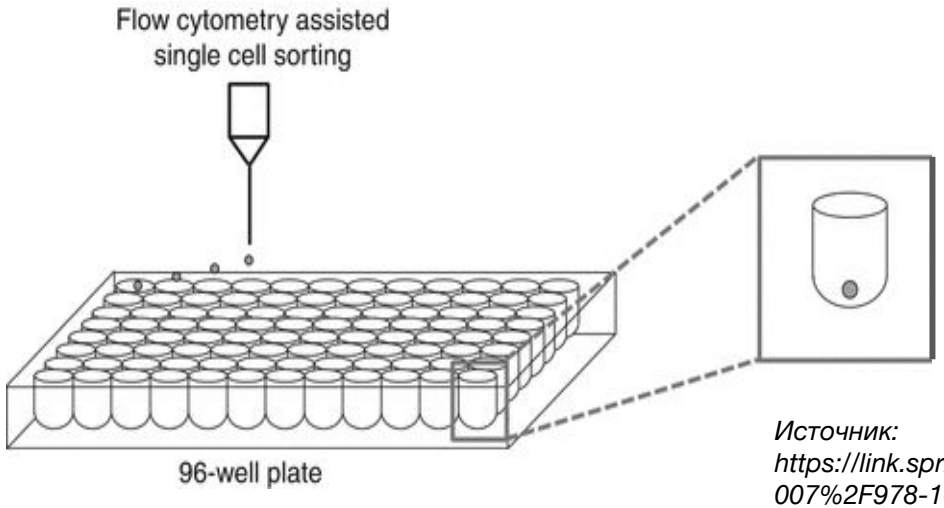
Источник: 10xgenomics.com

Smart-seq

Разделение клеток на лунки при помощи сортера

Существует ряд методов, которые требуют изолирования одиночных клеток при помощи клеточного сортера (например, методы класса Smart-seq)

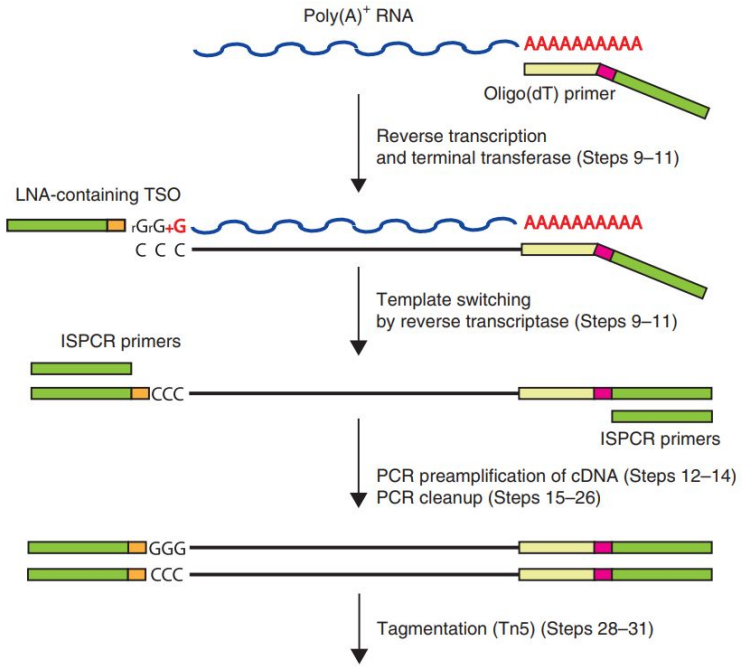
Главная проблема таких методов в ограниченном числе клеток, которые можно отсеквенировать за раз

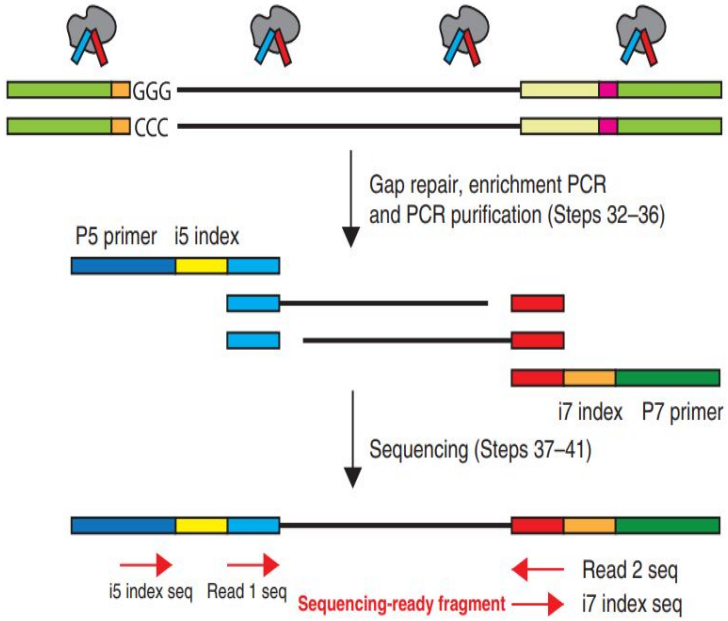


Источник:
https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_625

Smart-seq2

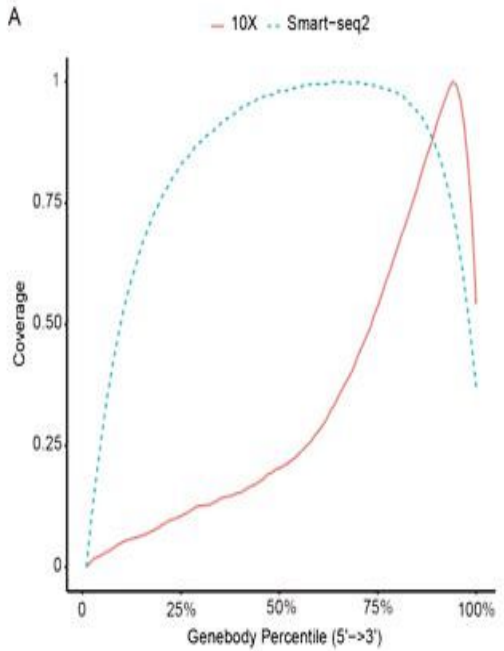
Контроль за ПЦР-дубликатами не ведётся (**отсутствуют UMI**), а идентичность клеток определяется образец-специфичными адаптерами для секвенирования (отдельный на каждую лунку)

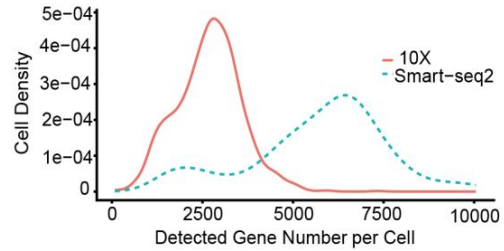




Picelli et al., 2014

Smart-seq2 vs. 10x Chromium GEX



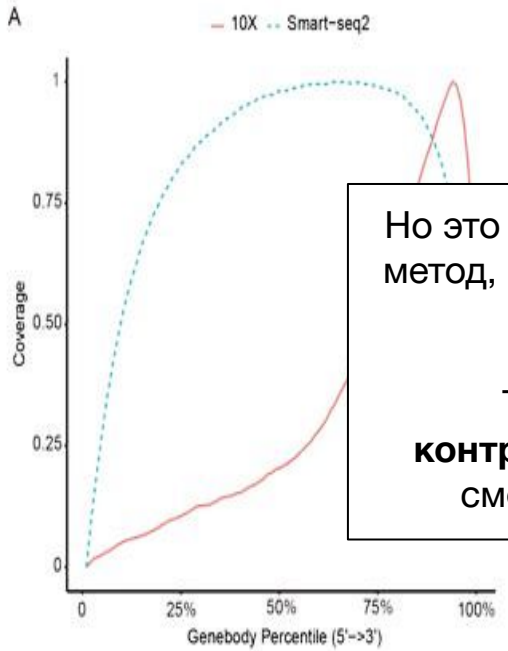


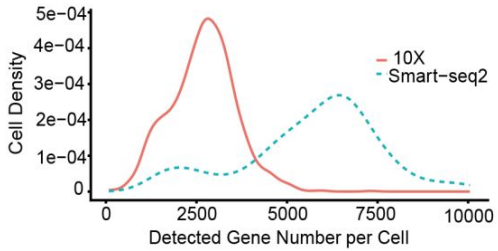
Smart-seq2 позволяет достичь более равномерного покрытия, чем 10x Chromium

Smart-seq2 охватывает экспрессию большего числа генов, чем 10x Chromium

Источник: https://daydaynews.cc/en/science/single-cell-transcriptome-smart-seq-2-or-10x-g-c.html

Smart-seq2 vs. 10x Chromium GEX





Но это гораздо более **дорогой** и **трудоёмкий** метод, поэтому чаще вместо него используют 10x Chromium

Также Smart-seq2 **не позволяет контролировать ПЦР-дубликаты** и часто смещает анализируемые экспрессии

ее
Chromium

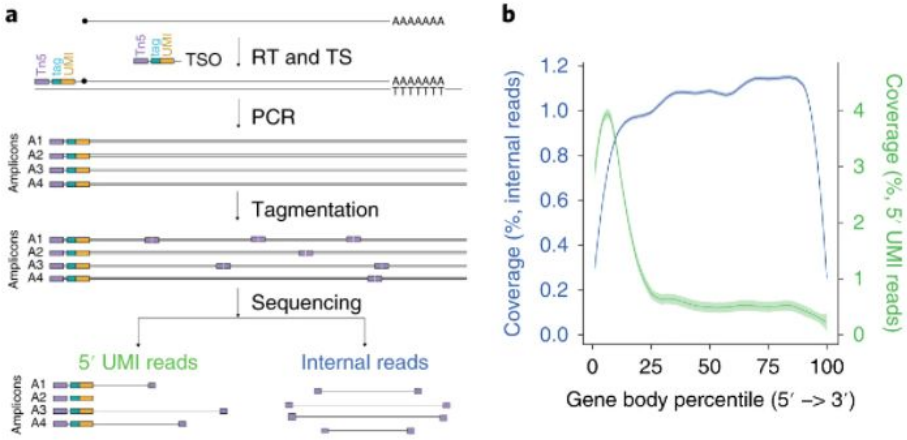
Smart-seq2 охватывает экспрессию большего числа генов, чем 10x Chromium

Источник: https://daydaynews.cc/en/science/single-cell-transcriptome-smart-seq-2-or-10x-g-c.html

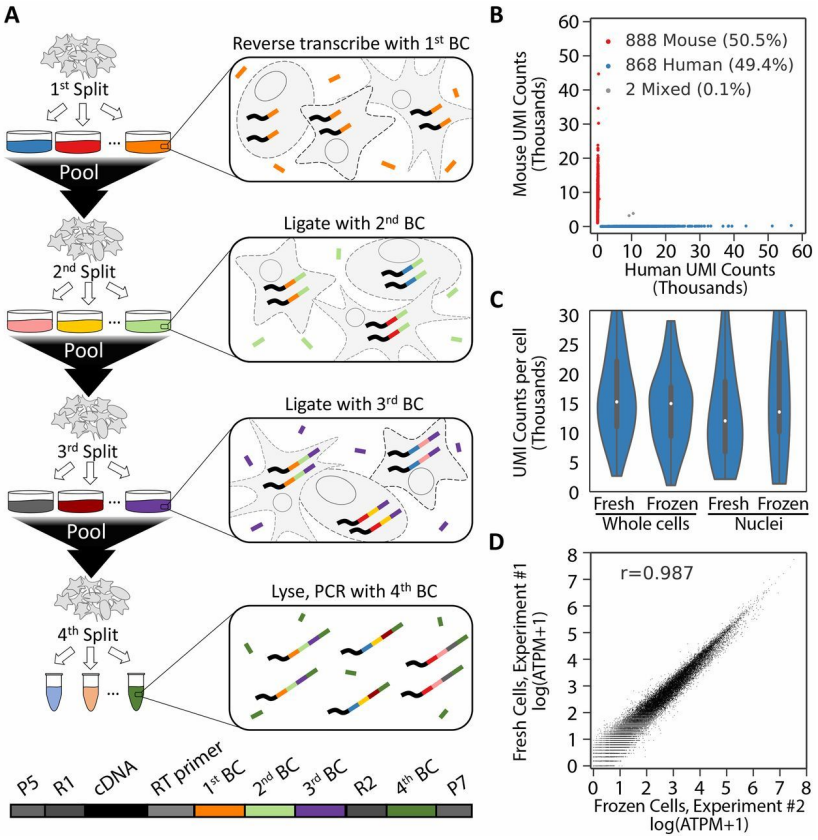
Smart-seq3

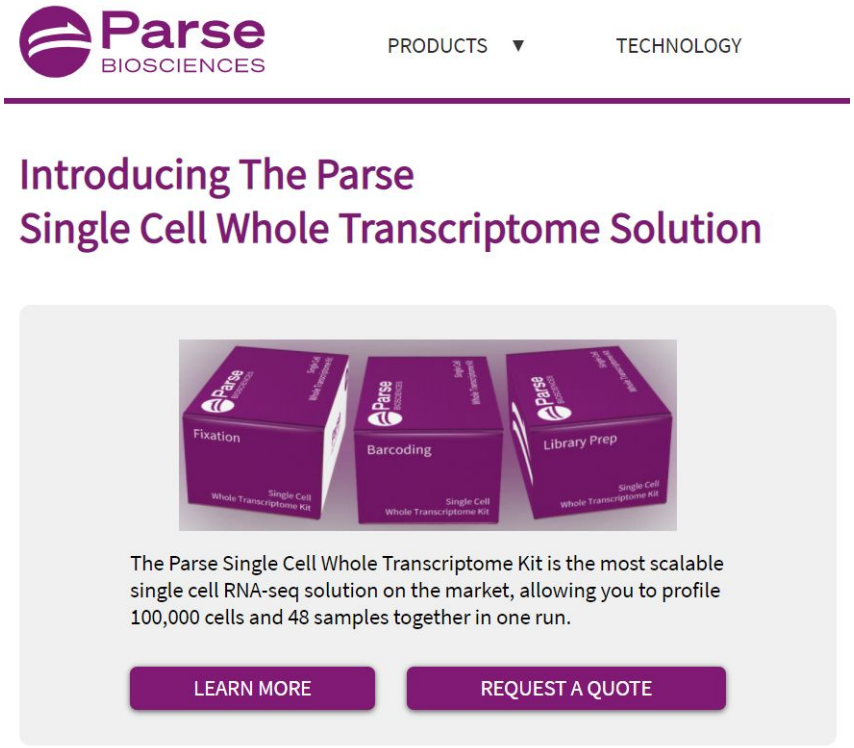
Добавлены UMI, перед которыми содержится специальный тег, который позволяет распознать последовательность UMI

Как итог мы имеем как контроль за ПЦР-дубликатами, так и полное покрытие “внутренними” ридами, что важно, например, в случае, когда нам хочется изучить мутации

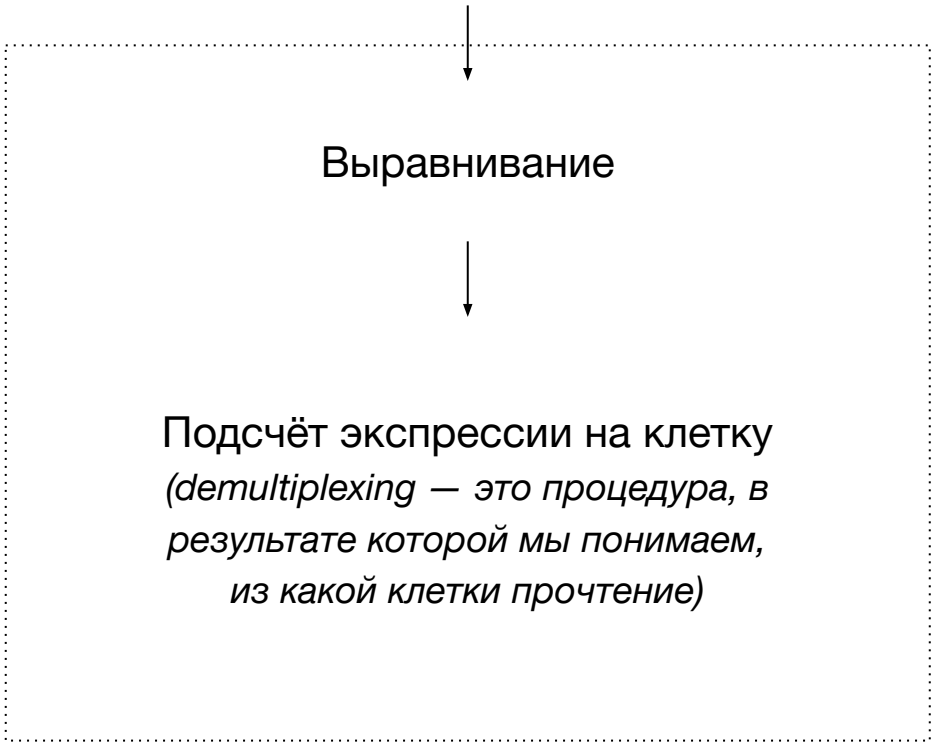


Parse Biosciences Evercode (ex SPLiT-Seq)





Подсчёт экспрессии



Эти стадии обычно выполняет одна и та же программа автоматически

Cell Ranger

- Подходит только для библиотек, полученных при помощи 10x Chromium
- Автоматически определяет версию химии 10x ⇒ не нужно прописывать координаты баркода / UMI в прочтениях (это сильно облегчает работу)
- Основан на STAR, а потому **очень** требовательный к ресурсам (1 Тб дискового пространства, 128 Гб RAM, 16 ядер)
- **Очень долго** работает (один образец может рассчитываться 12 часов)
- Умеет работать с данными CITE-Seq и большим количеством иных модификаций scRNA-Seq-эксперимента
- Может вернуть **.bam-файл с картированием**, если попросить его это сделать

Cell Ranger

- **Очень** просто запускается:

```
cellranger count \
  --id={id запуска} \
  --transcriptome={путь до директории с референсным геномом}
  --fastqs={директория с прямыми прочтениями},{директория с обратными прочтениями} \
  --sample={название образца} \
  --localcores={число ядер}
```

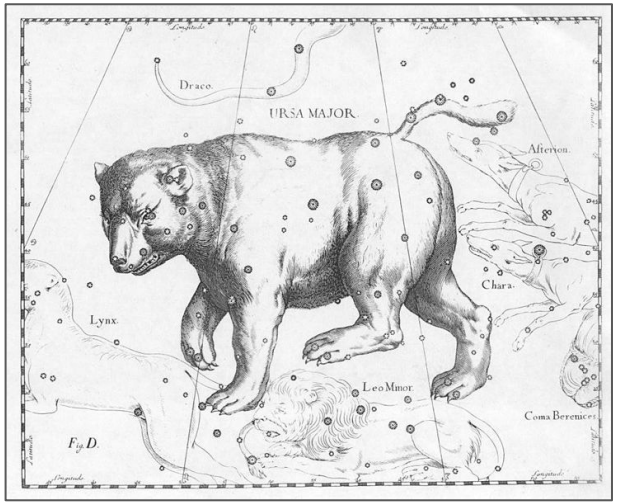
- Подготовленный к работе референсный геном можно найти на сайте Cell Ranger (можно сделать и свой)

Cell Ranger

- В простейшем случае аутпут содержит 4 файла:

1. `raw_feature_bc_matrix.tar.gz` — матрица со всеми “клетками” из датасета
   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.

kallisto | bustools



- Подходит для большого числа различных библиотек (в основном 10x Chromium, но не только). BUS расшифровывается как barcode | UMI | sequence, поэтому подойдут практически любые UMI-based методы
- kallisto | bustools основан на псевдовыравниваниях с использованием kallisto, поэтому он **не требовательный к железу**
- Работает, как правило, в **несколько раз быстрее**, чем Cell Ranger

- Необходимо впрямую прописывать координаты баркода, UMI и последовательности на прочтениях. Чуть менее user-friendly, чем Cell Ranger
- Умеет работать с **CITE-Seq** и некоторыми другими протоколами
- **Не возвращает выравнивание!**

kallisto | bustools

- Запускается очень просто:

```
kb count \
  -i {файл с индексом} \
  -g {файл с соответствием транскриптов генам} \
  -x {версия химии 10x или описание координатов баркода и UMI} \
  {прямые прочтения} {обратные прочтения}
```

- Индекс (он же референс) можно сделать самостоятельно или загрузить с сайта kallisto | bustools уже созданный
- **Не делает автоматическую фильтрацию клеток**! Выводит относительно мало статистики

kallisto | bustools

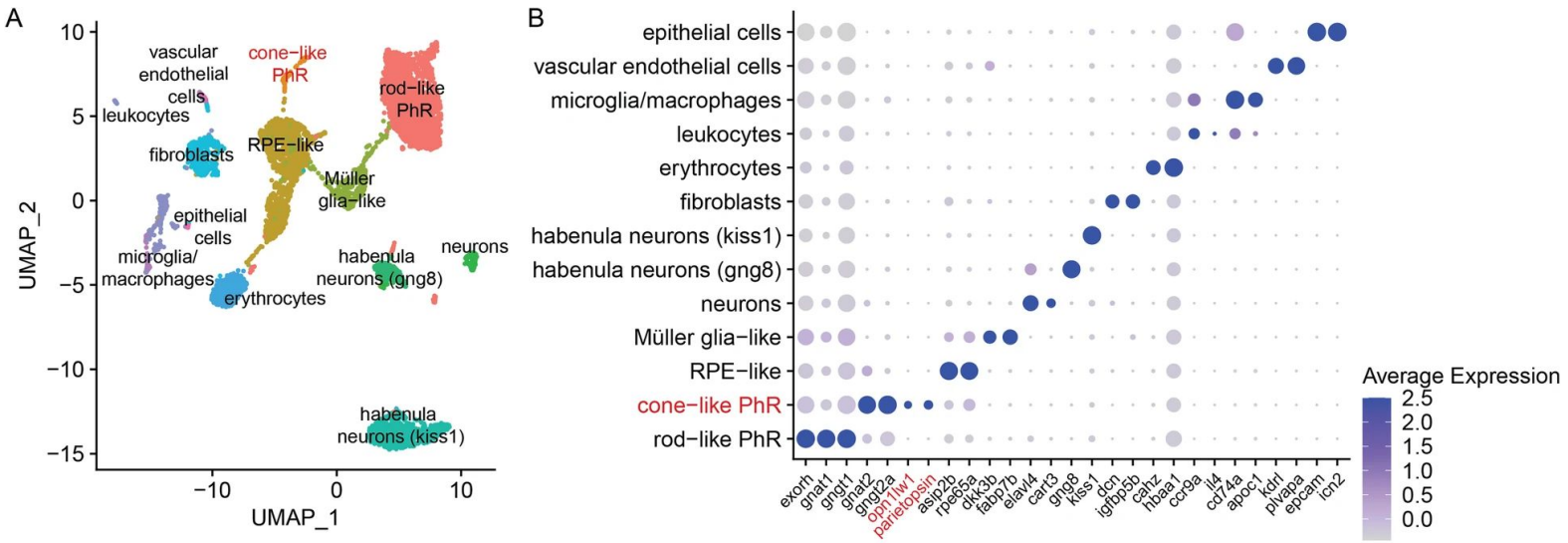
- В простейшем случае аутпут содержит 1 файл и 1 папку:

1. `counts_unfiltered` — матрица со всеми “клетками” из датасета
   a. `cells_x_genes.barcodes.txt`— названия клеток (баркоды)
   b. `cells_x_genes.genes.txt`— названия и id генов
   c. `cells_x_genes.mtx`— непосредственно матрица экспрессии в sparce-виде
2. `inspect.json` — .json-файл с краткой статистикой по QC клеток

kallisto | bustools и паралоги

Из-за того, что прочтения, которые были откартированы неоднозначно, просто отбрасываются при процессинге при помощи STAR (= CellRanger), то часто возникает проблема различить типы клеток, отличающиеся по экспрессии паралогичных генов

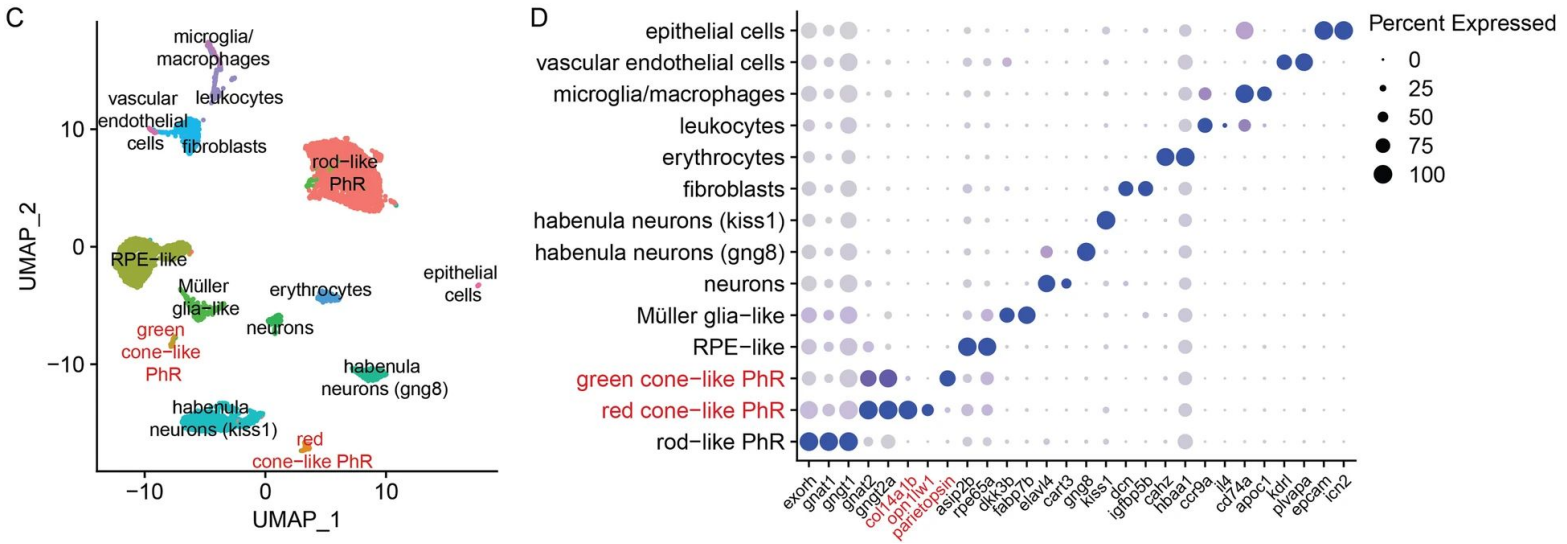
CellRanger



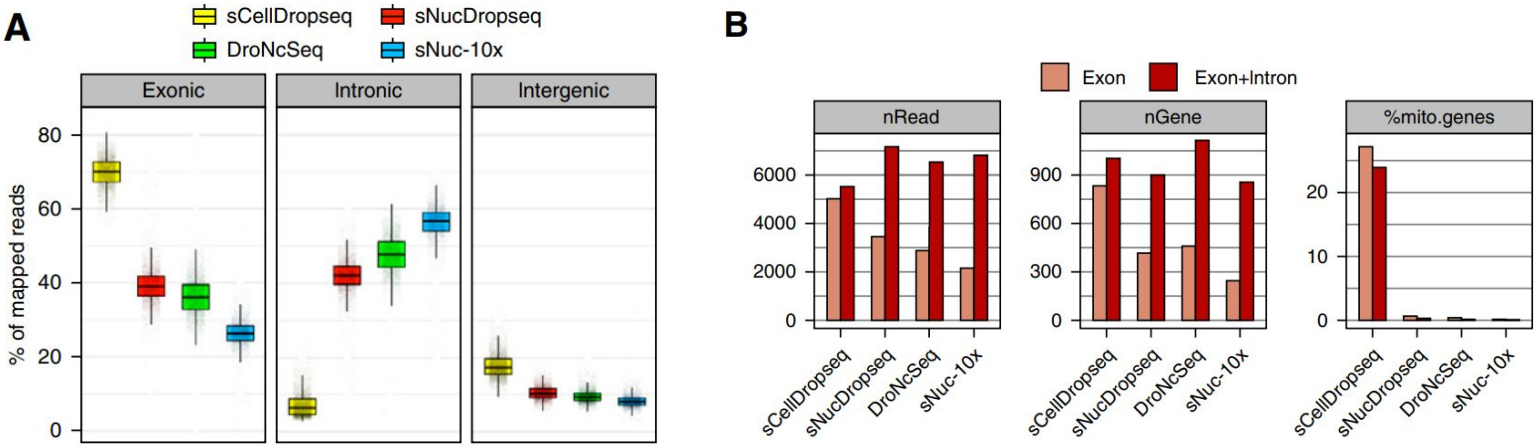
kallisto | bustools и паралоги

Из-за того, что прочтения, которые были откартированы неоднозначно, просто отбрасываются при процессинге при помощи STAR (= CellRanger), то часто возникает проблема различить типы клеток, отличающиеся по экспрессии паралогичных генов

kallisto | bustools



Картирование snRNA-Seq



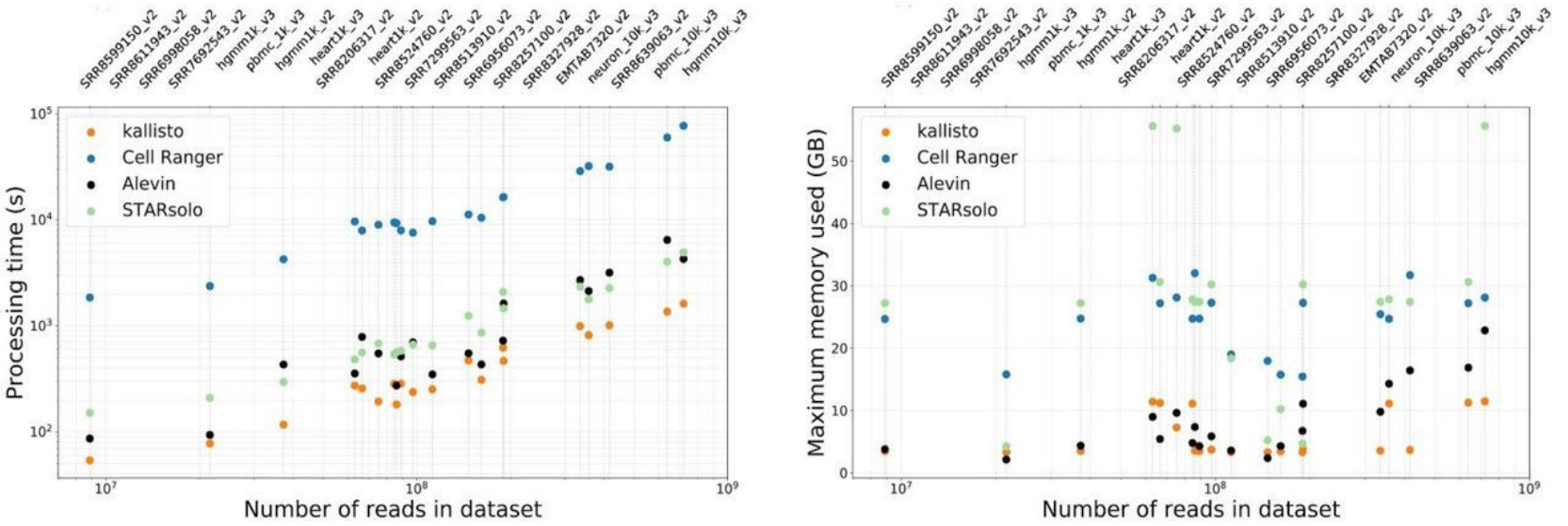
Wu et al. 2018

- В snRNA-Seq большая часть прочтений ложится в интронные регионы, это необходимо учитывать при выравнивании

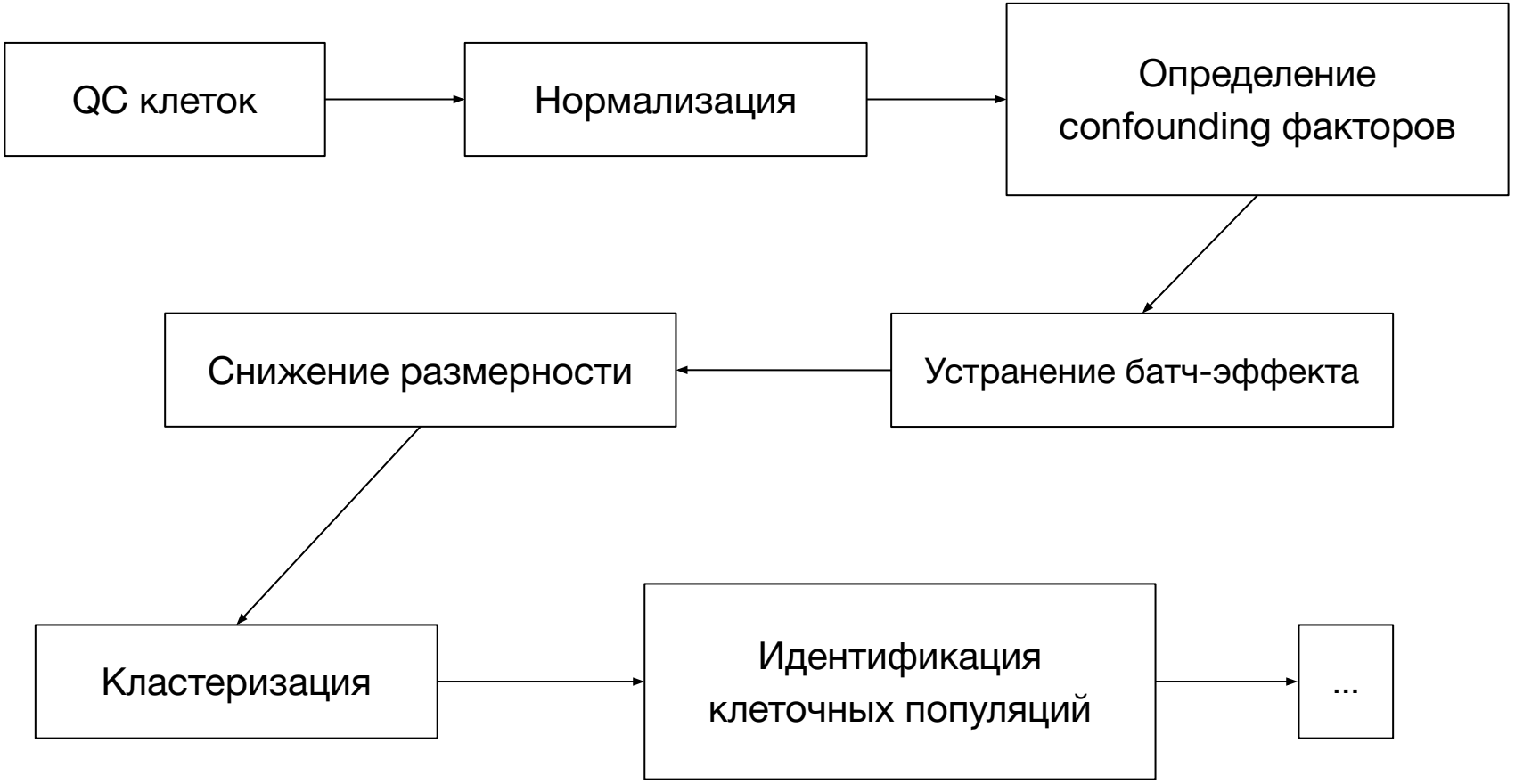
Сравнение пайплайнов

- Cell Ranger — это самый затратный и медленный пайплайн, однако именно он является сейчас «золотым стандартом» препроцессинга данных scRNA-Seq

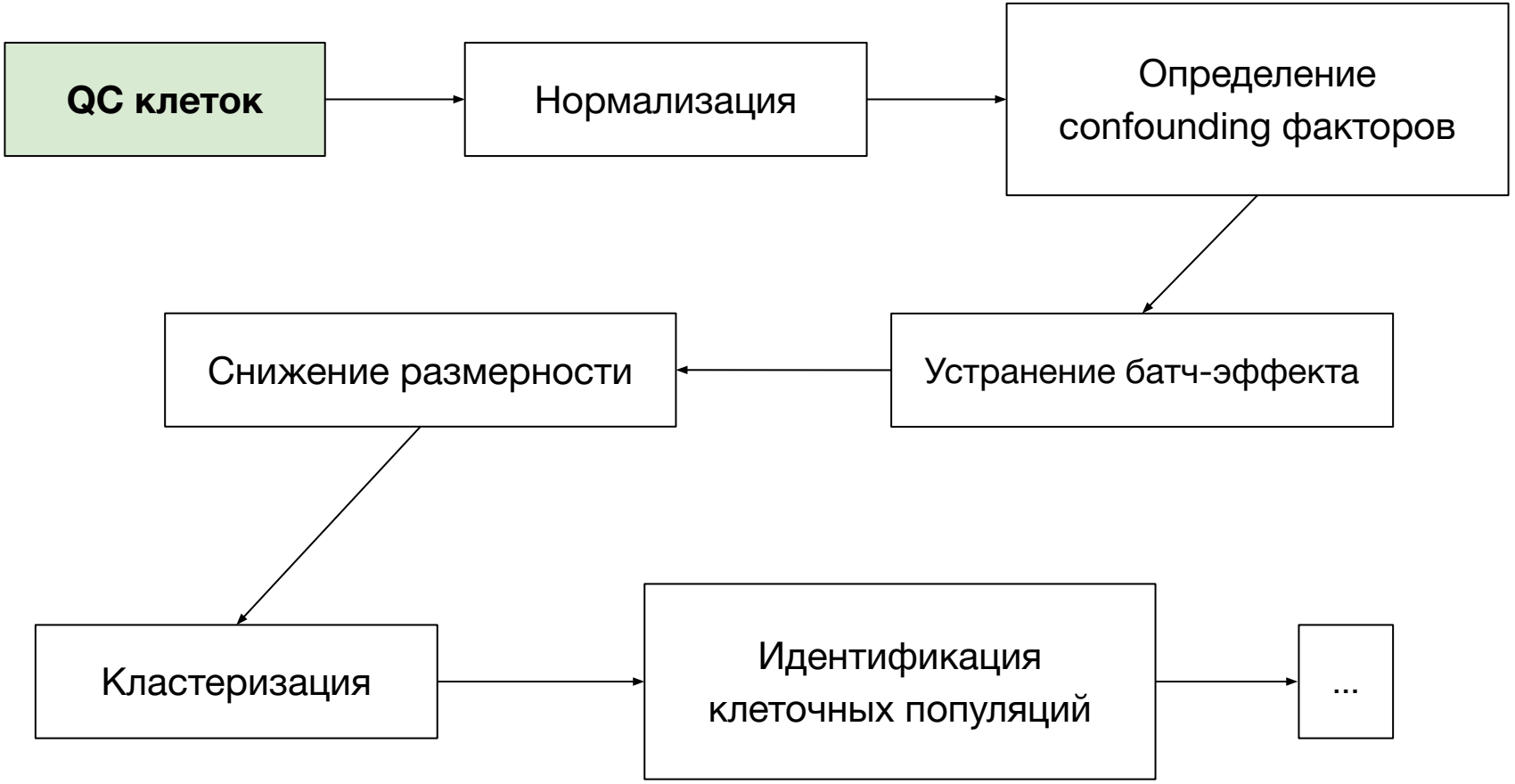
Melsted et al. 2019



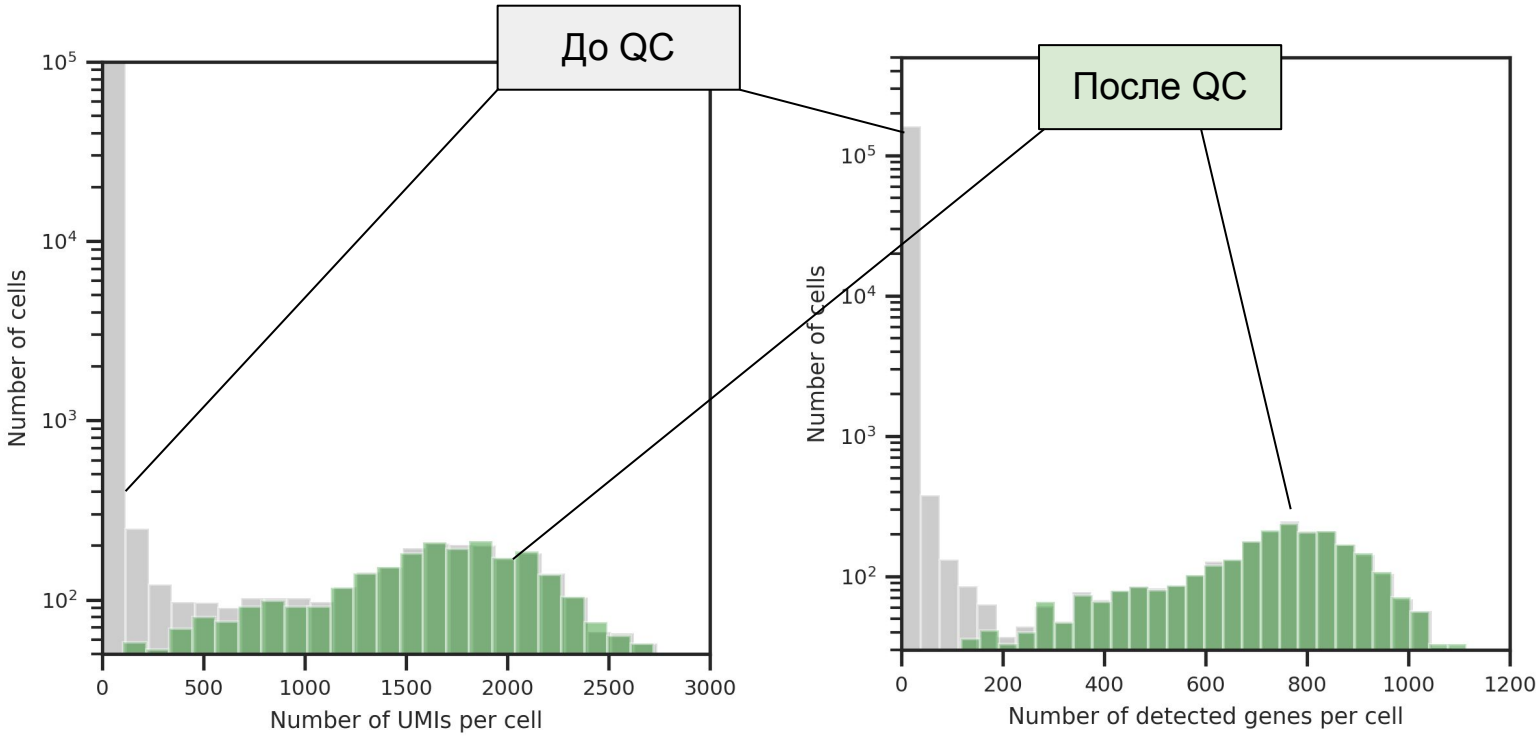
Обработка данных



Обработка данных



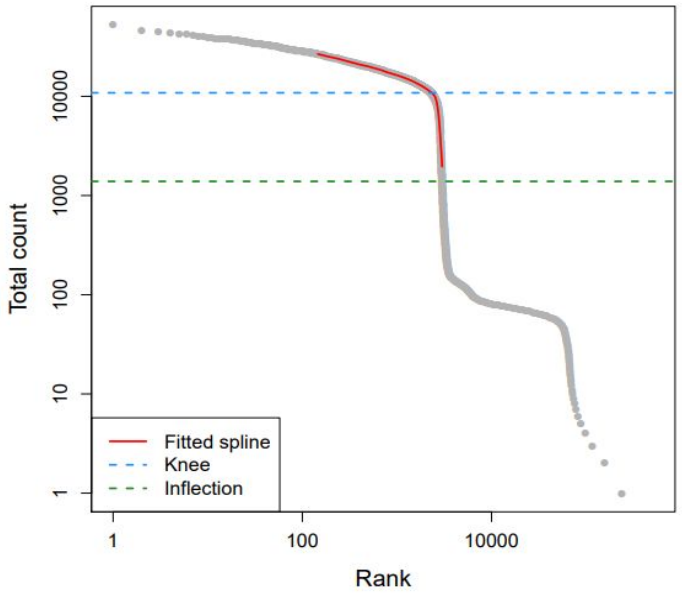
QC клеток

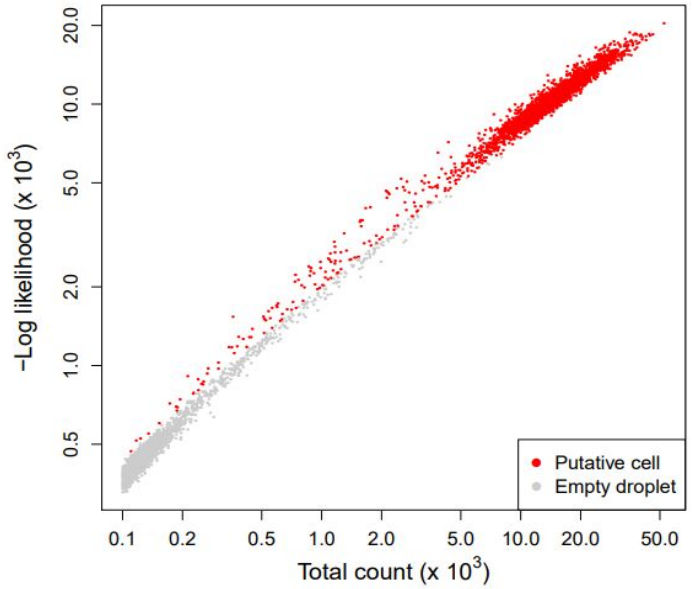

До QC
После QC
Number of cells
10^5
10^4
10^3
10^2
0
500
1000
1500
2000
2500
3000
Number of UMIs per cell
Number of cells
10^5
10^4
10^3
10^2
0
200
400
600
800
1000
1200
Number of detected genes per cell

QC клеток

- Для идентификации пустых капель (без клеток) можно использовать пакет DropletUtils с его функцией emptyDrops (есть только на R)
- Всегда необходимо смотреть на распределение числа UMI / генов / митохондриальной экспрессии на клетку
- Клетки с высокой митохондриальной экспрессией мы считаем плохими (их тоже имеет смысл выфильтровывать)

Lun et al. 2019

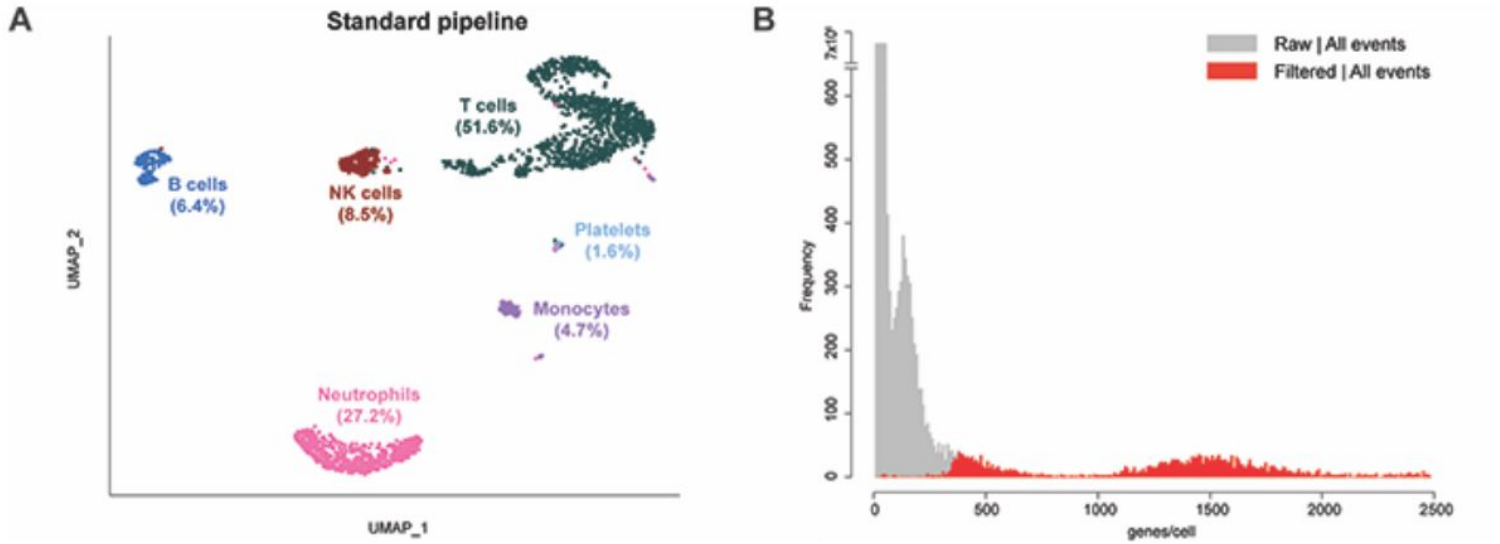




Влияние QC на результат

Различные типы клеток могут иметь разное количество UMI на клетку из-за биологической разницы (например, в случае с нейтрофилами это явнее всего — почему?)

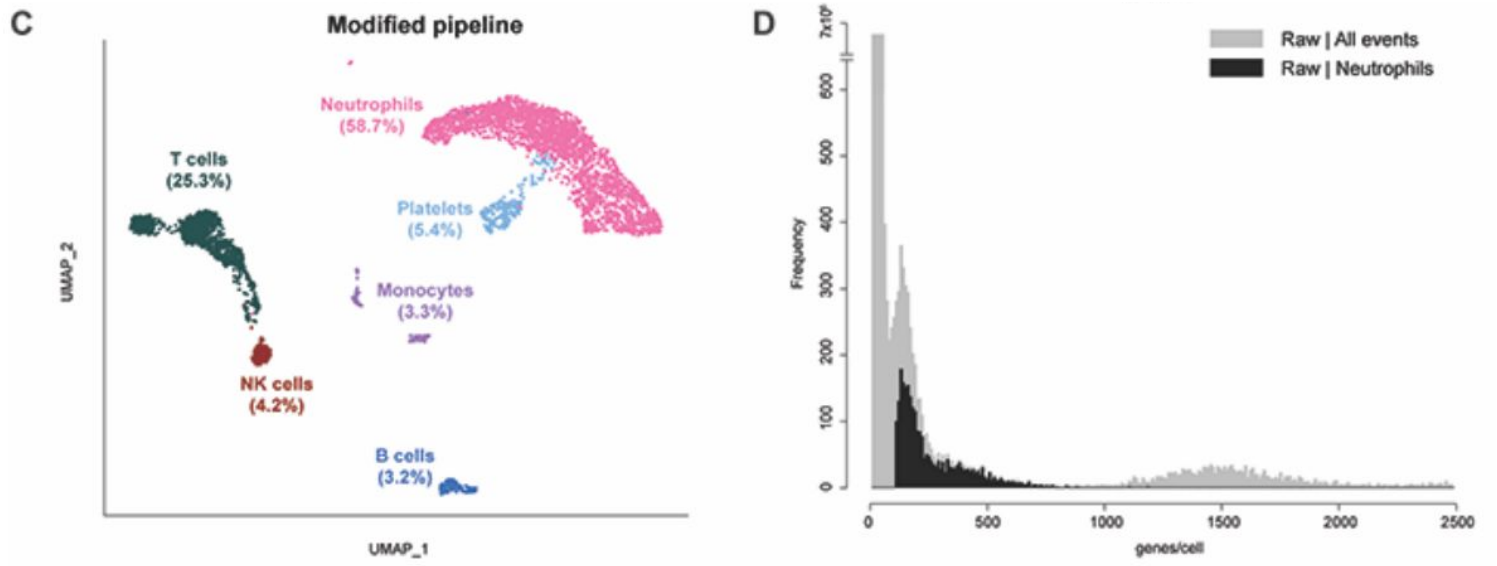
Строгая фильтрация



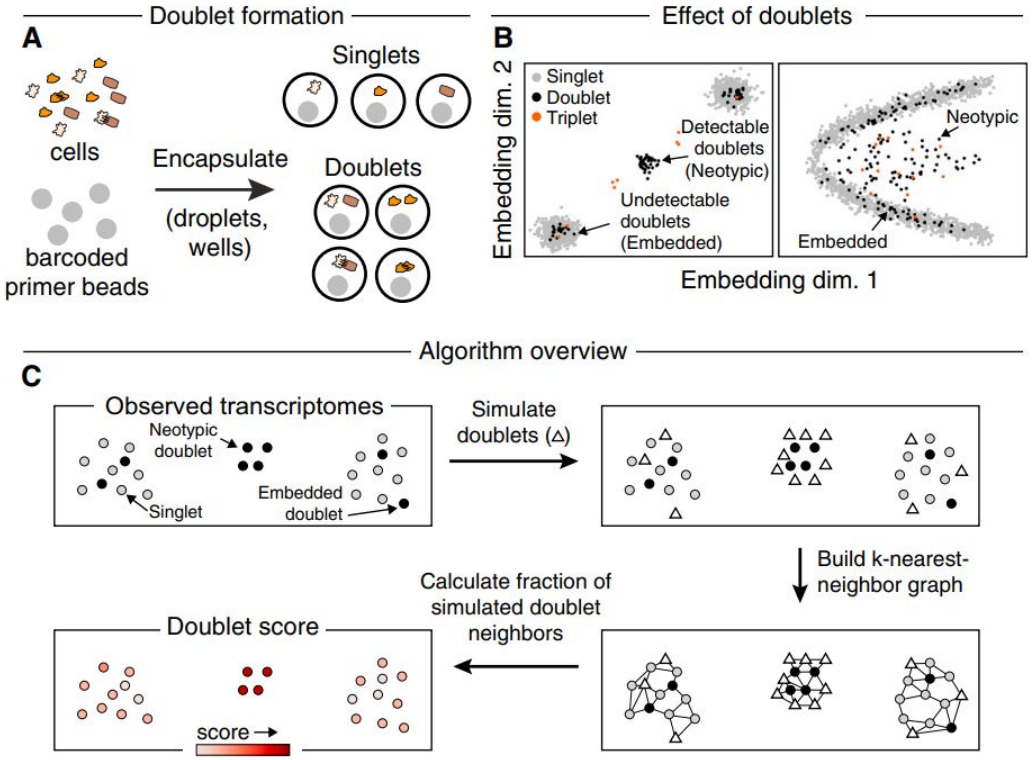
Влияние QC на результат

Различные типы клеток могут иметь разное количество UMI на клетку из-за биологической разницы (например, в случае с нейтрофилами это явнее всего — почему?)

Нестрогая фильтрация



Scrublet (**S**ingle-**C**ell **R**emover of Do**ublets**)



Wolock et al. 2019

- Помимо пустых капель существует и иная проблема — дублеты клеток
- Дублеты могут мешать работе с scRNA-Seq-данными (как минимум их сложно типировать)
- Существуют эффективные методы их идентификации (например, Scrublet)