«Анализ транскриптомных данных»

Лекция #7.

Контроль качества клеток в scRNA-Seq

Серёжа Исаев

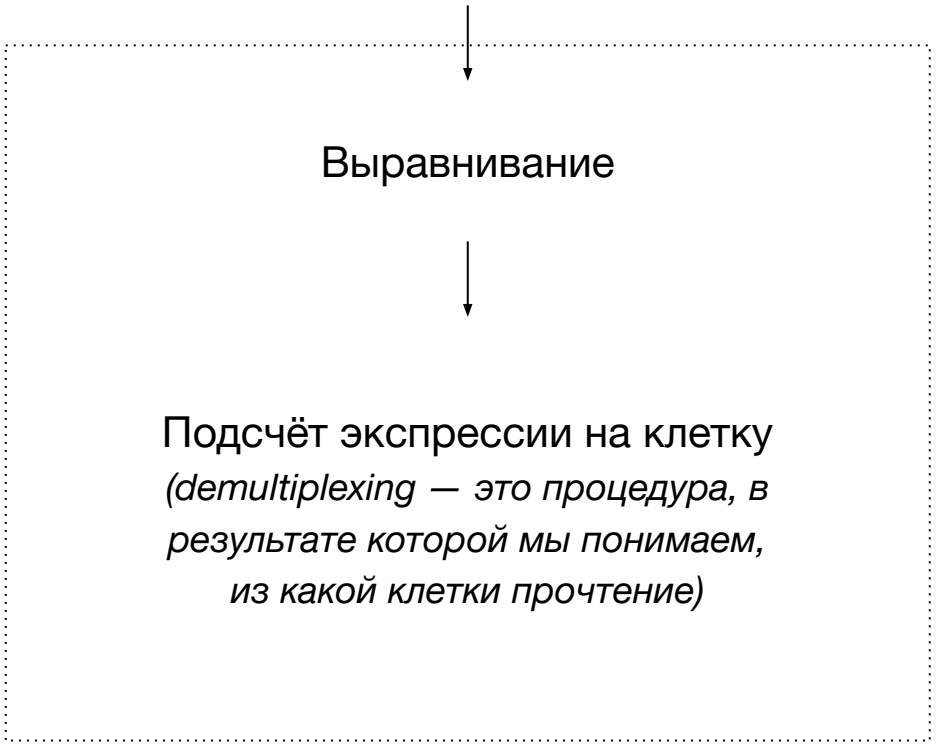
аспирант **ФБМФ МФТИ**
аспирант **MedUni Vienna**

Содержание курса

1. Bulk RNA-Seq:
    a. экспериментальные подходы,
    b. выравнивания и псевдовыравнивания,
    c. анализ дифференциальной экспрессии,
    d. функциональный анализ;

2. Single-cell RNA-Seq:
    a. экспериментальные подходы,
    b. отличия от процессинга bulk RNA-Seq,
    c. методы снижения размерности,
    d. кластера и траектории,
    e. мультимодальные омики одиночных клеток.

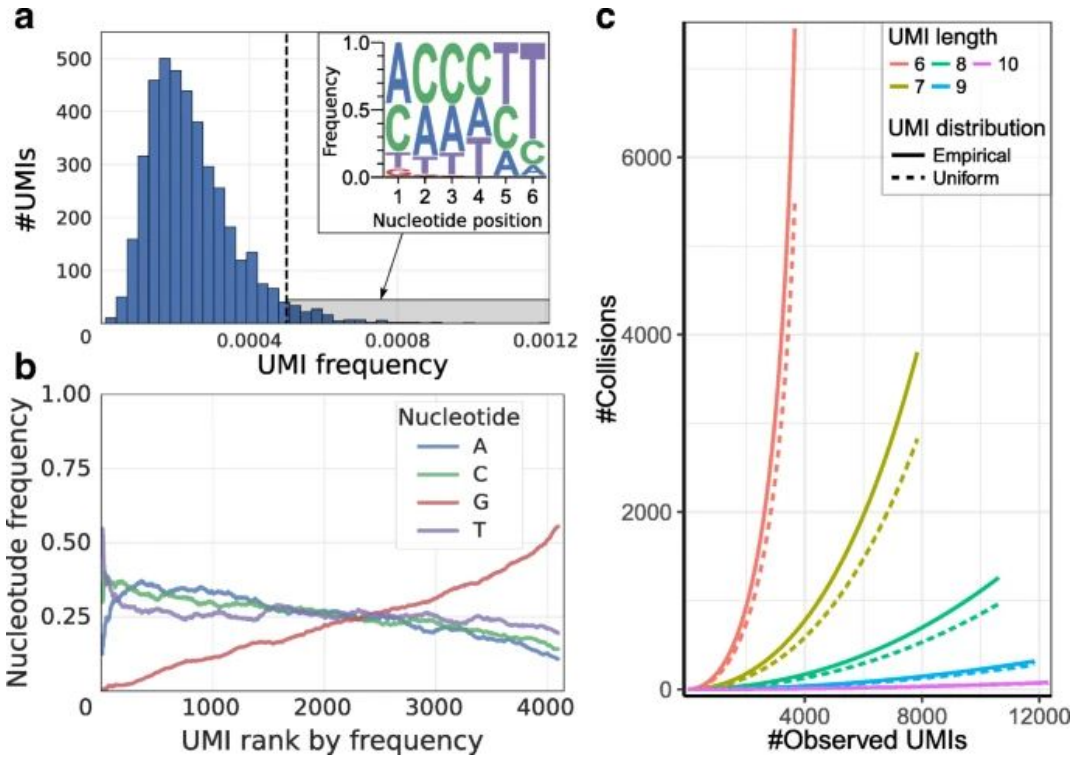
Подсчёт экспрессии



Эти стадии обычно выполняет одна и та же программа автоматически

Проблема демультиплексации: dropEst

Для корректного восстановления последовательности баркода и UMI используются порой нетривиальных подходы типа dropEst



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}
```

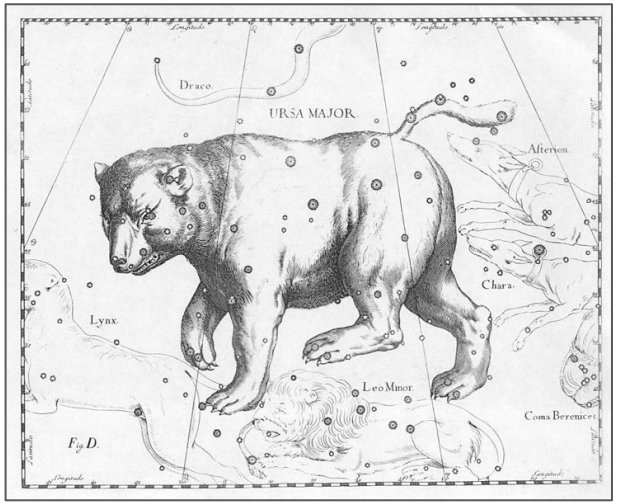
- Подготовленный к работе референсный геном можно найти на сайте Cell Ranger (можно сделать и свой)

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

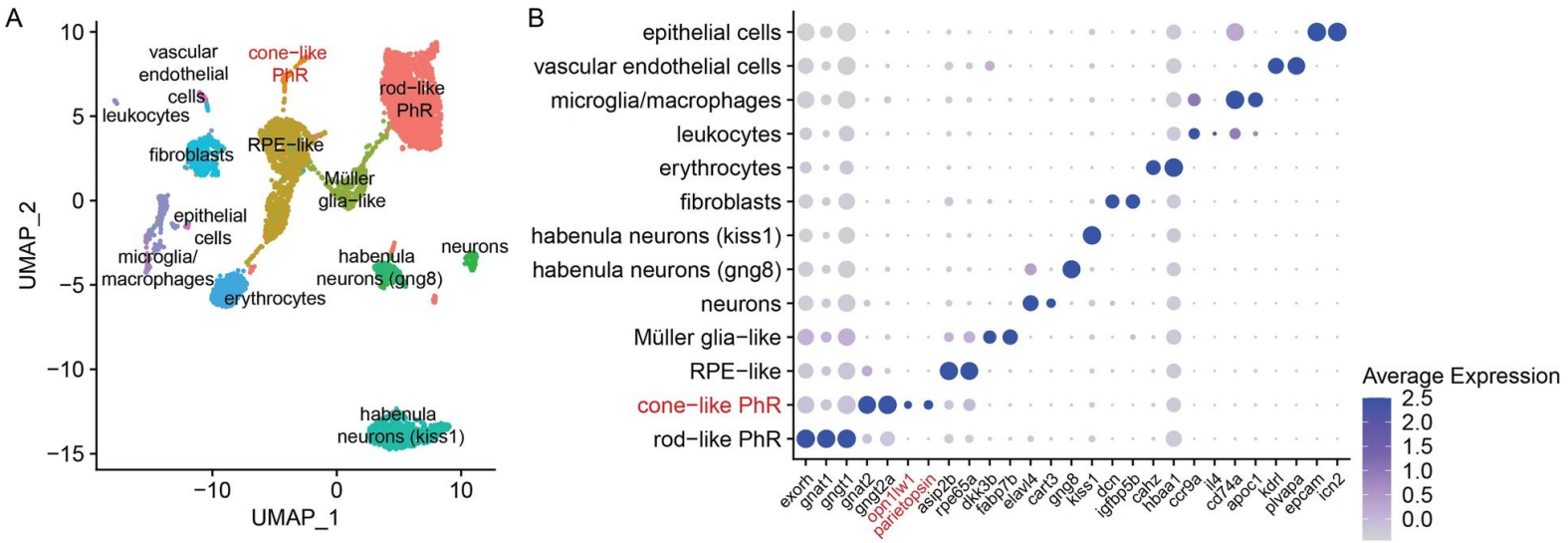
kallisto | bustools

- 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

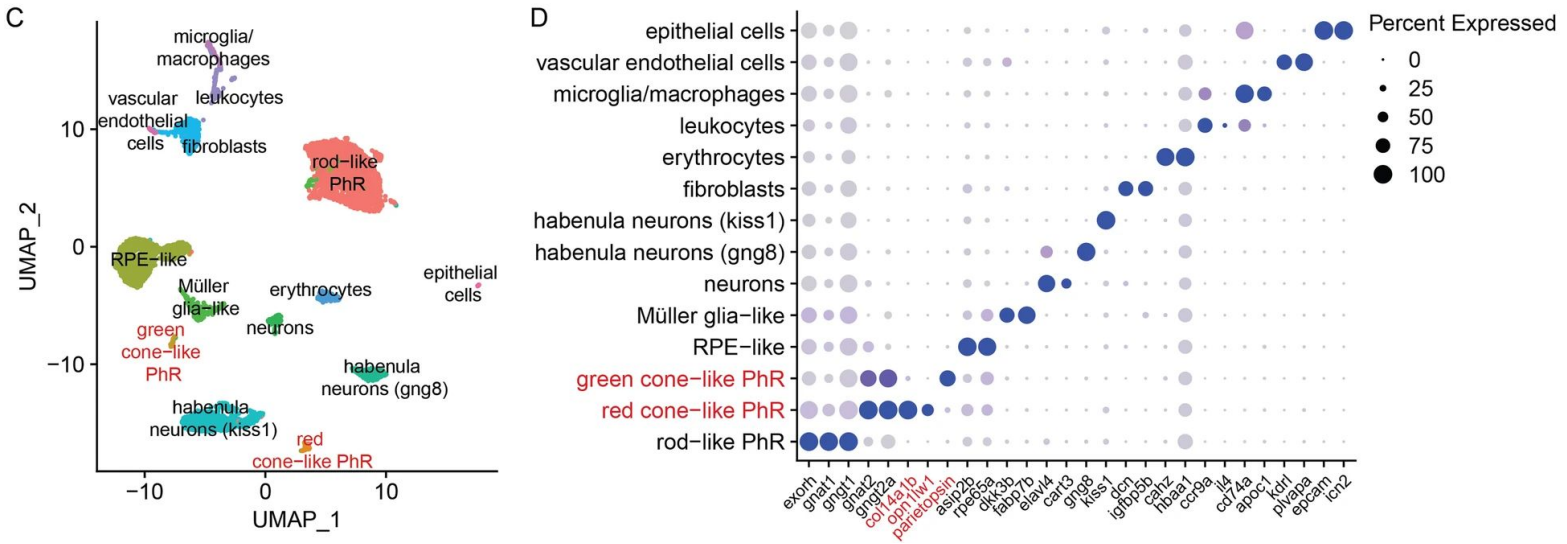
CellRanger



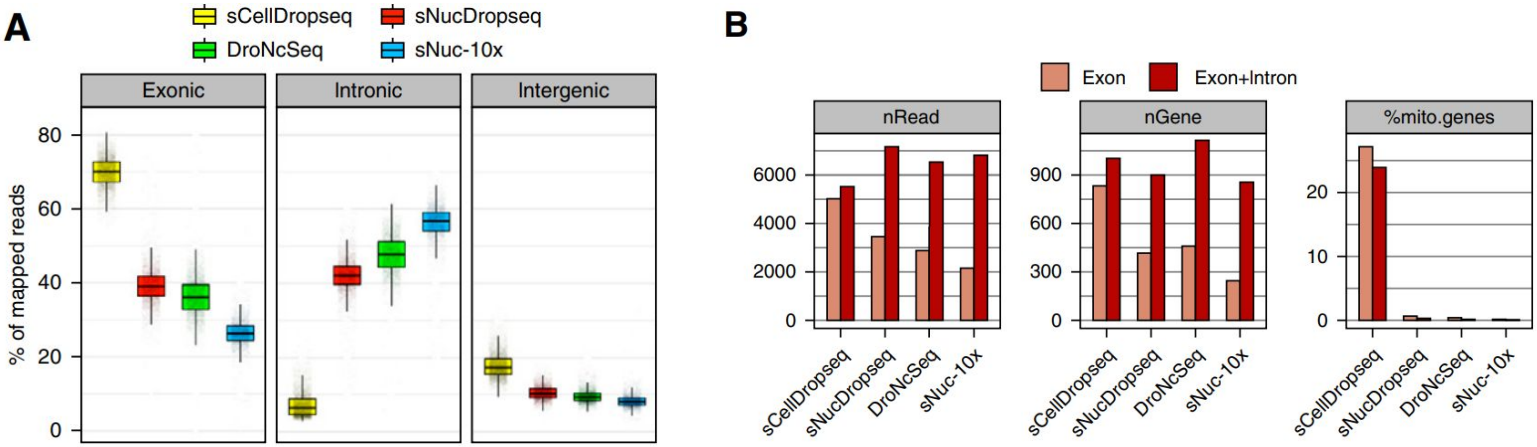
kallisto | bustools и паралоги

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



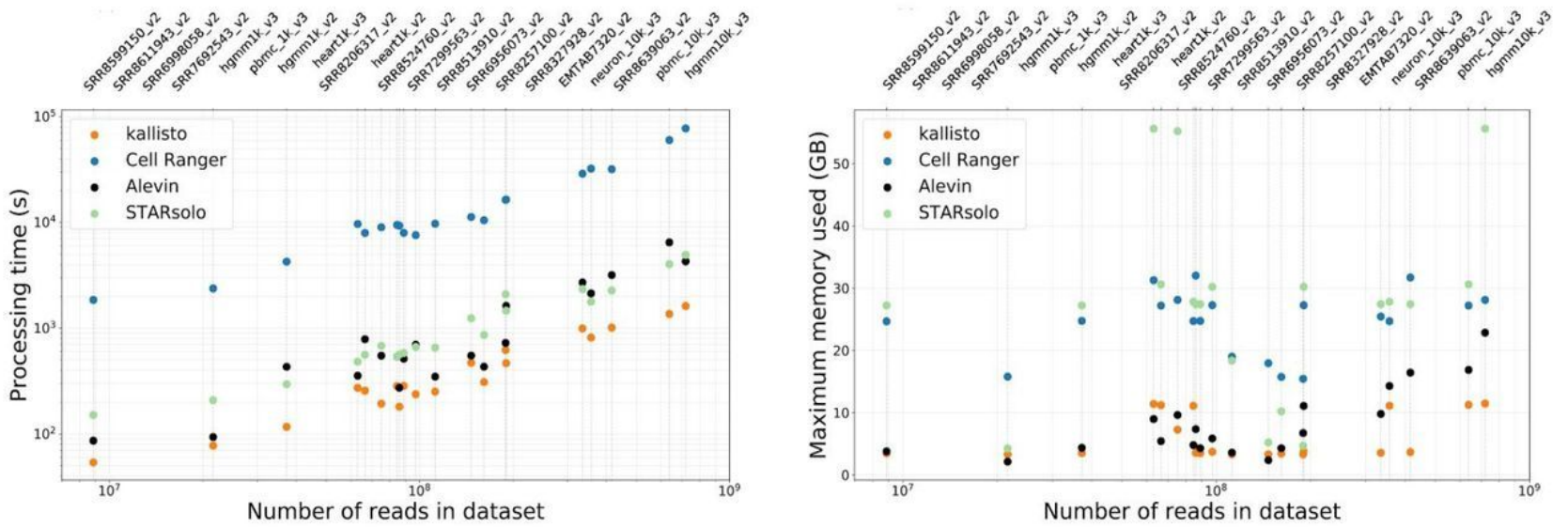
Wu et al. 2018

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

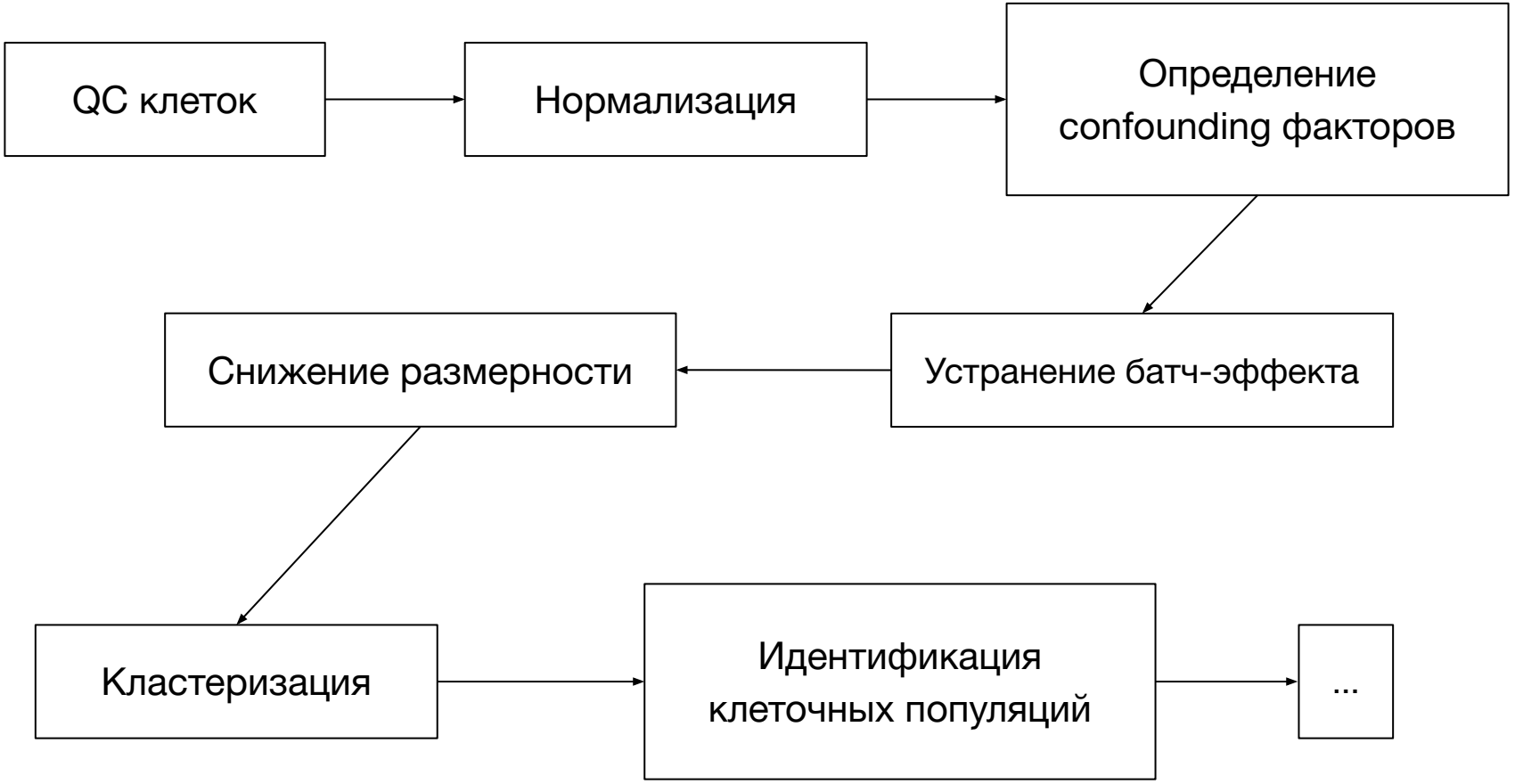
Сравнение пайплайнов

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

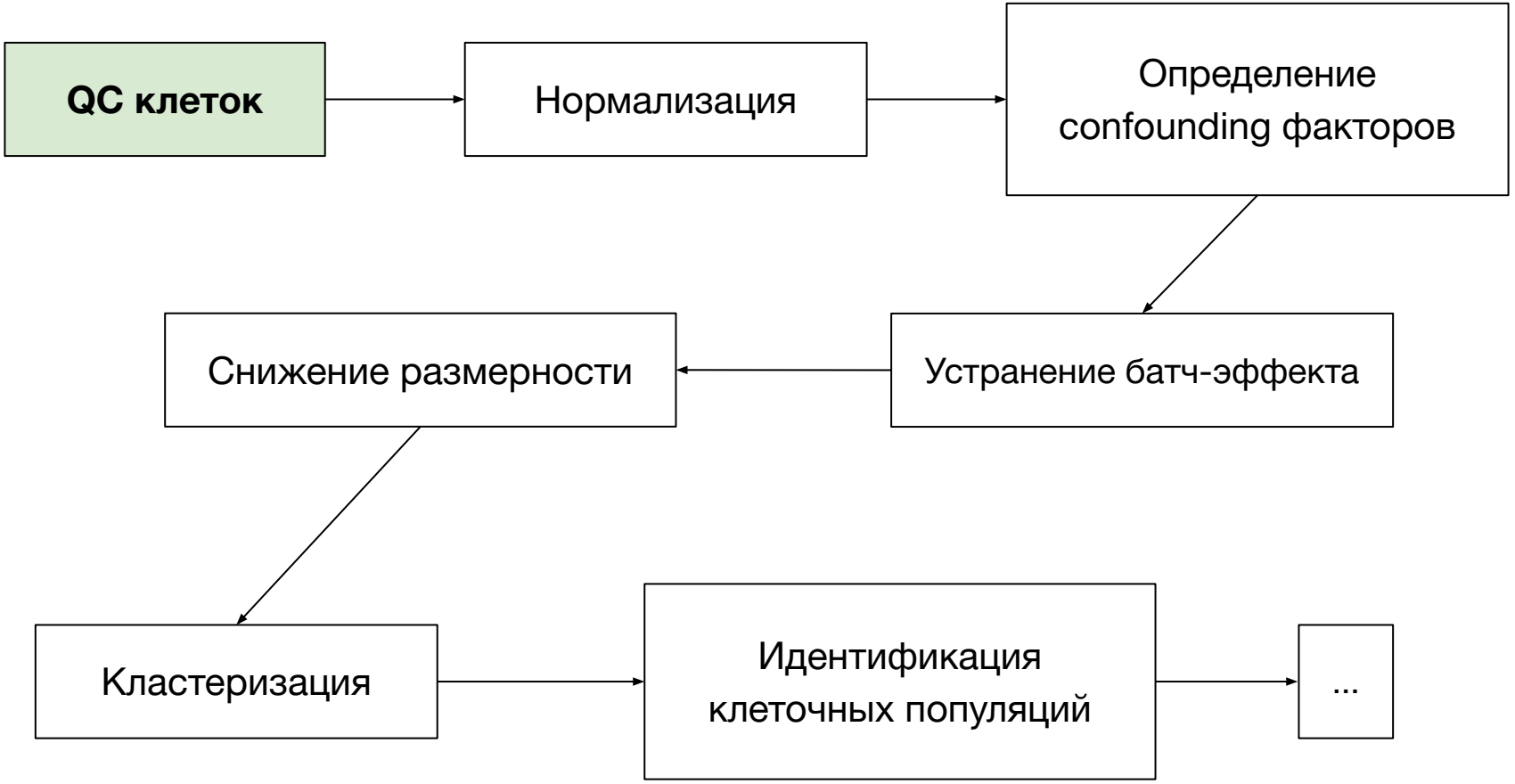
Melsted et al. 2019



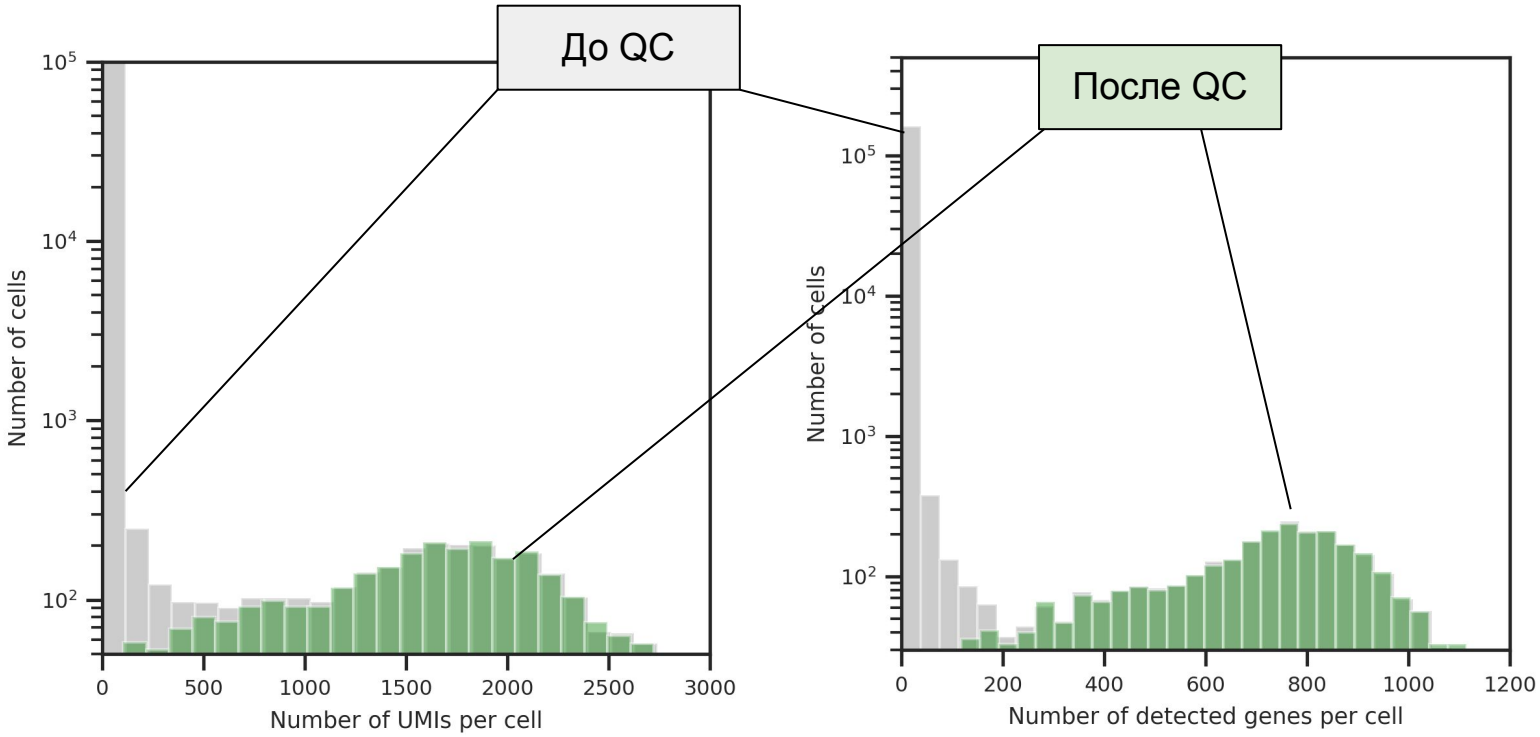
Обработка данных



Обработка данных



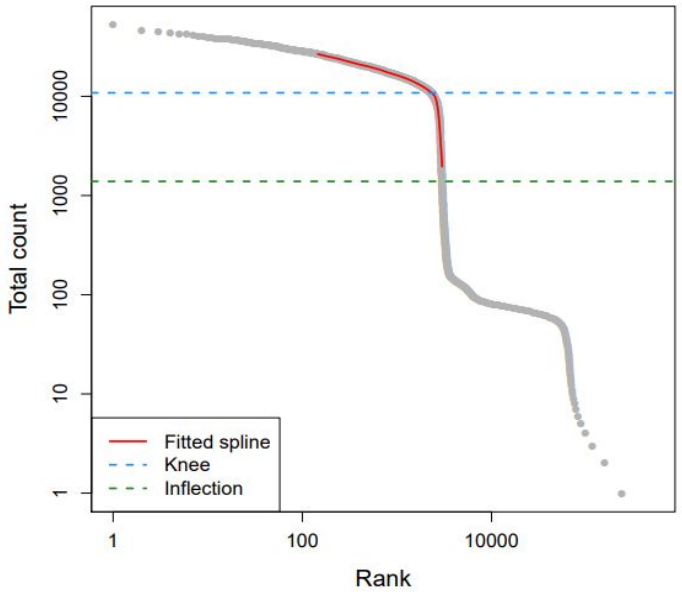
QC клеток

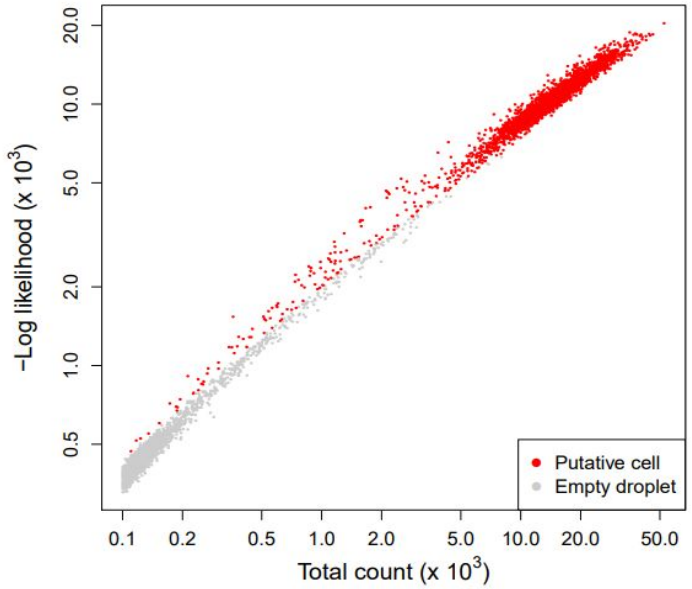

До QC
После QC
Number of cells
Number of UMIs per cell
Number of detected genes per cell
0
500
1000
1500
2000
2500
3000
200
400
600
800
1200
10^2
10^3
10^4
10^5

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

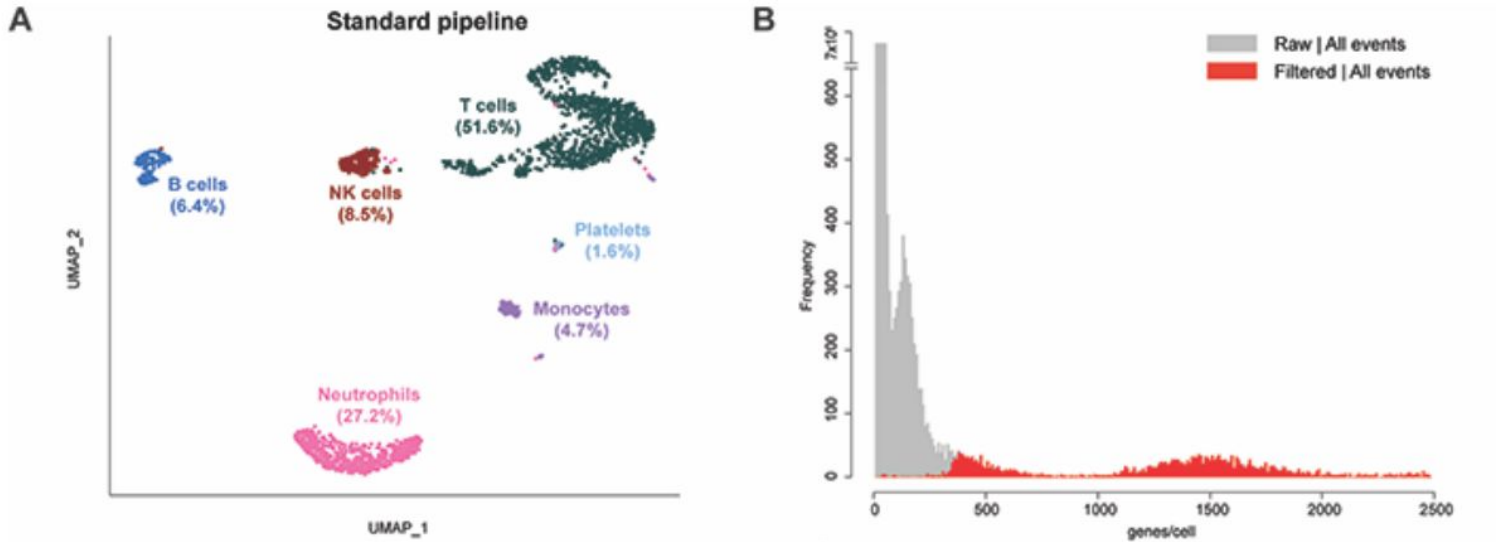




Влияние QC на результат

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

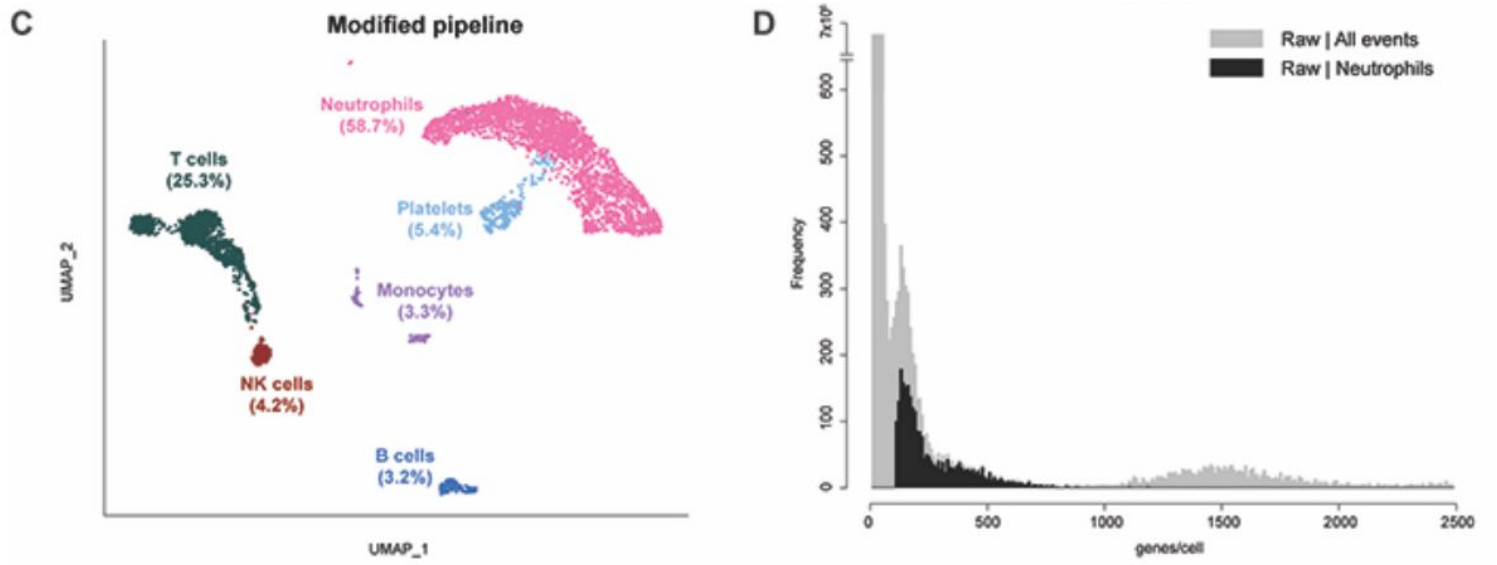
Строгая фильтрация



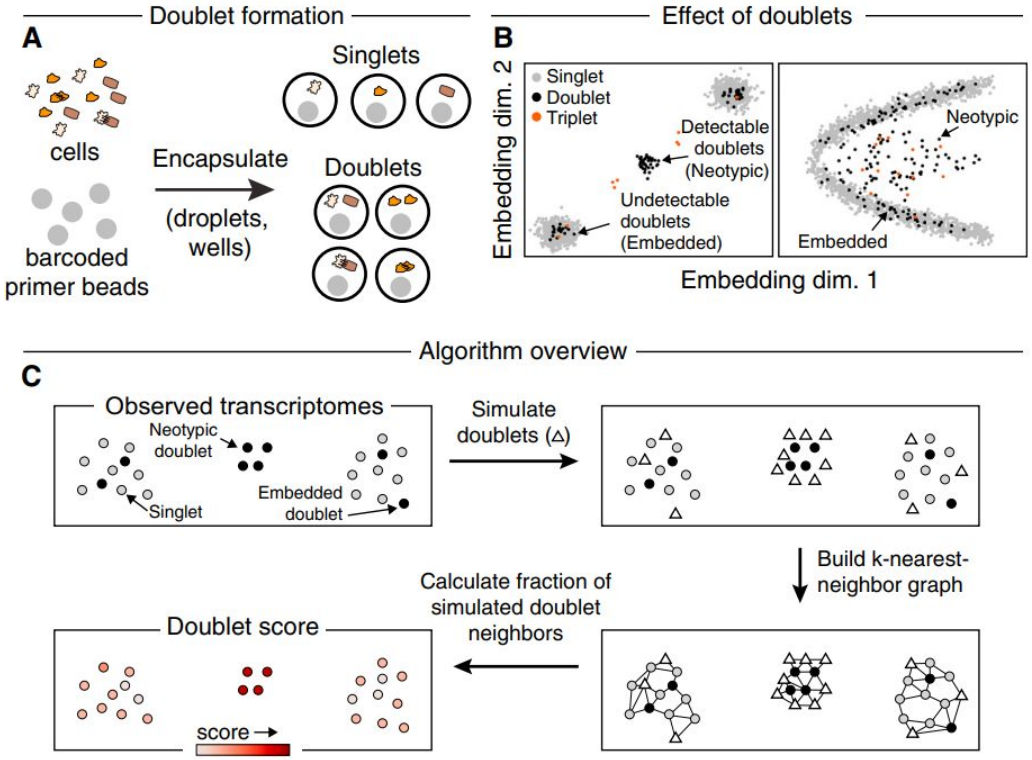
Влияние QC на результат

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**ublet**s)



- 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)

Wolock et al. 2019