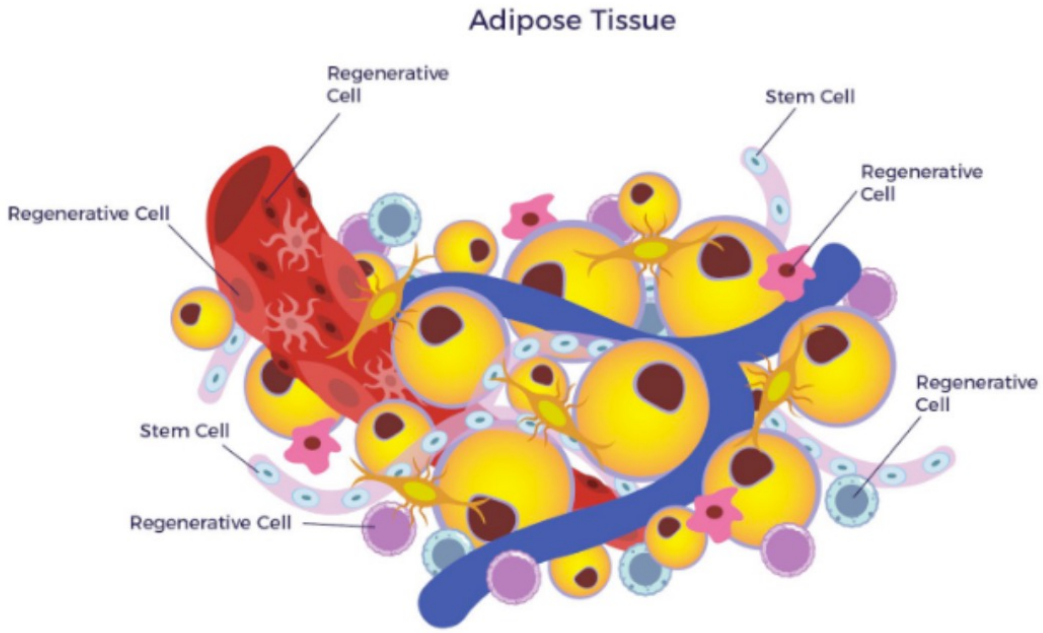
Single Cell RNA-seq: Characteristics, Preprocessing and QC

Hubert Rehrauer

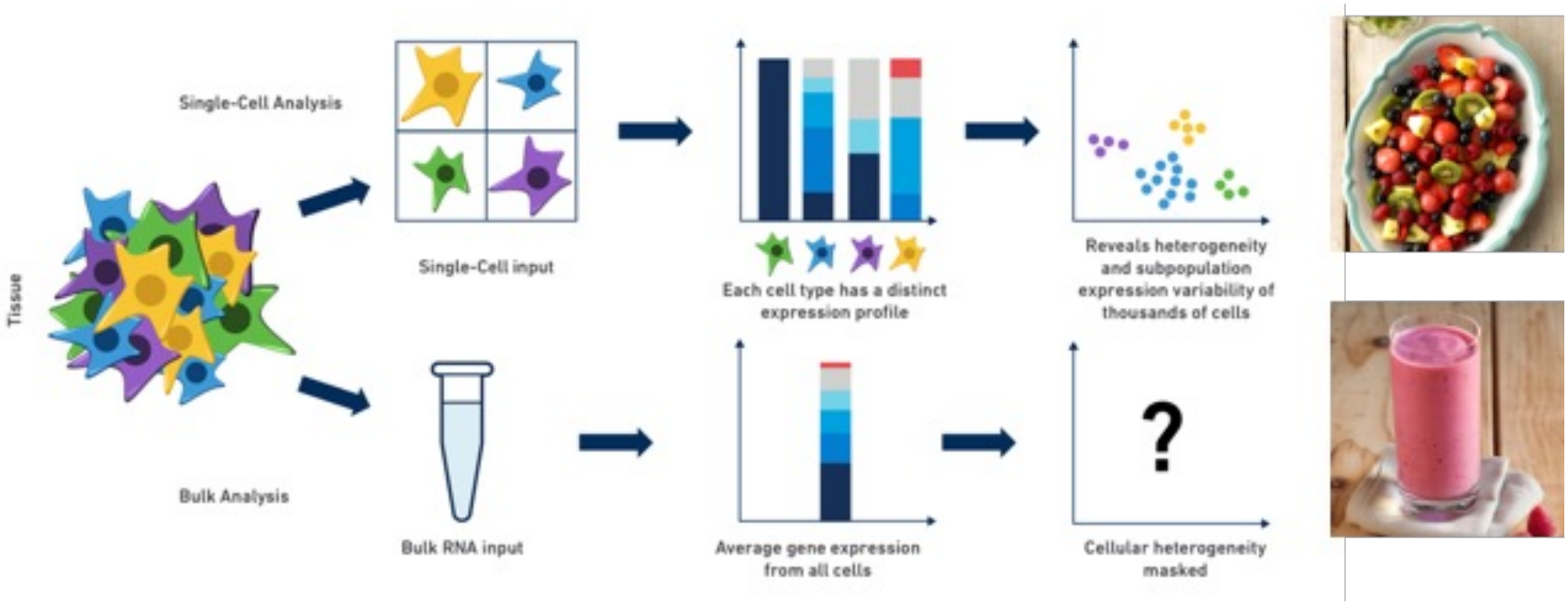
University of Zurich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

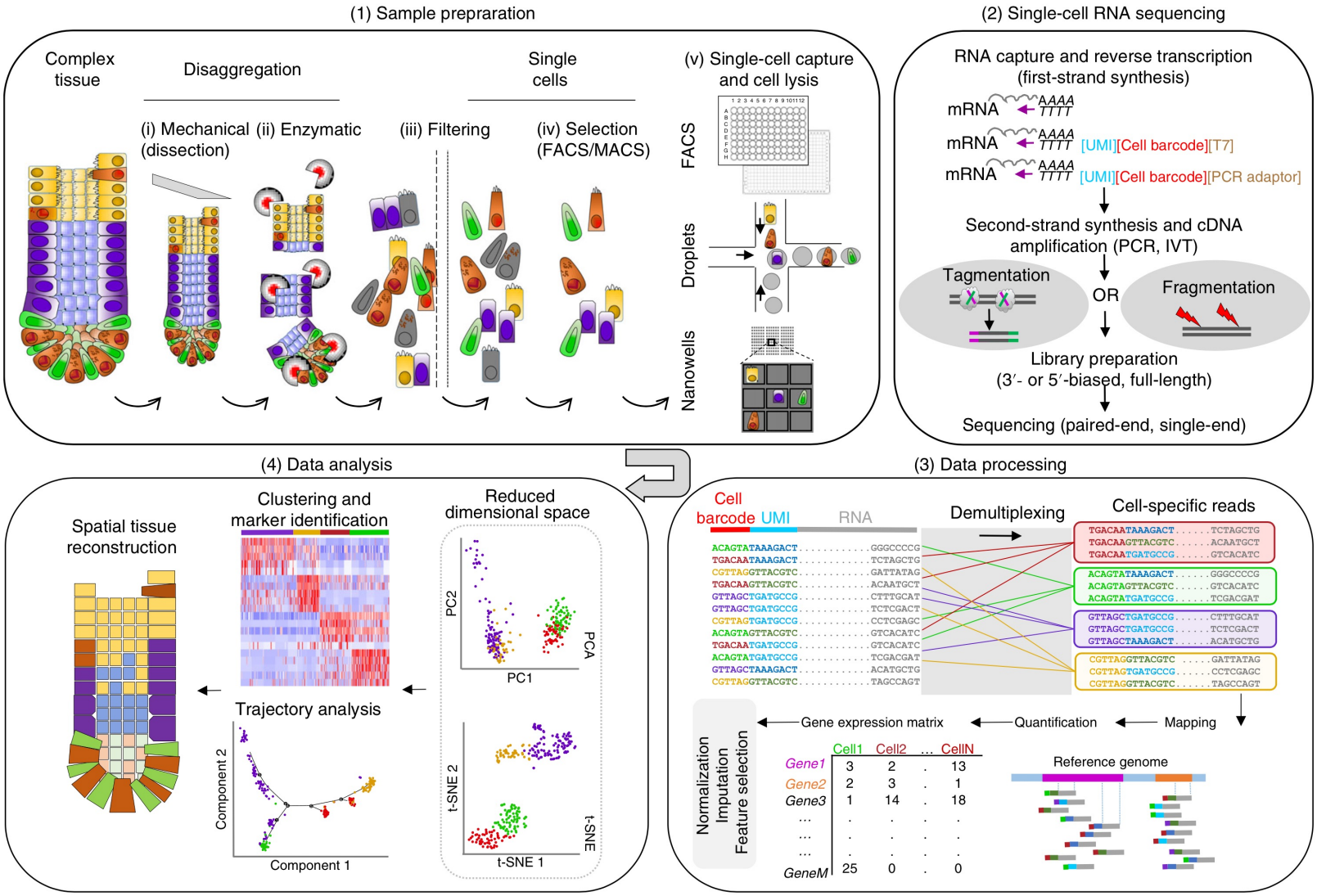
Tissues are heterogeneous



levelanatomyandphysiology.com


Single-Cell Analysis
Tissue
Single-Cell input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA input
Average gene expression from all cells
?
Cellular heterogeneity masked

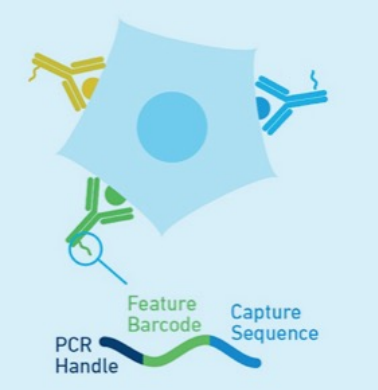
scRNA-seq

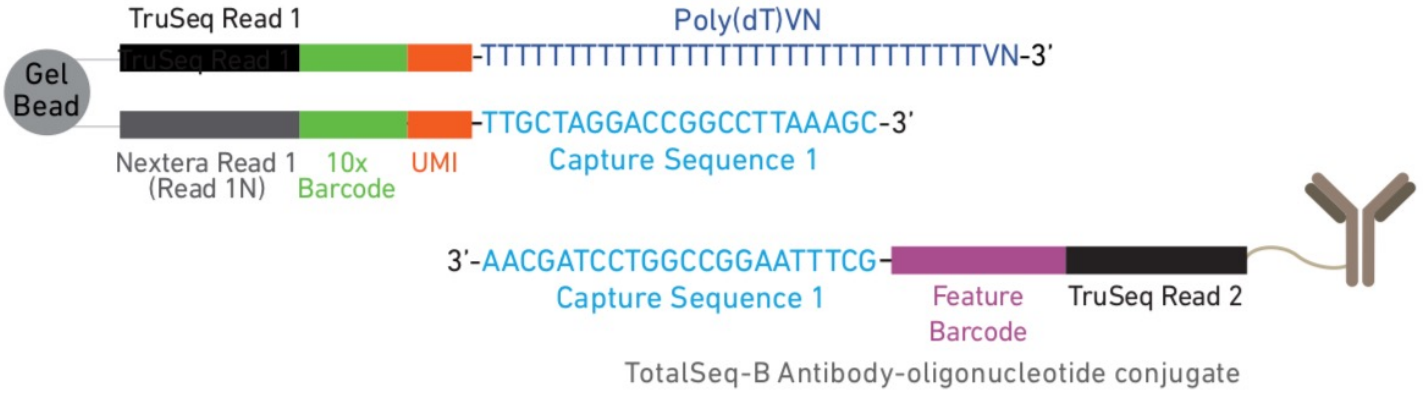


Lafzi et al. https://www.nature.com/articles/s41596-018-0073-y

Cell hashing (feature barcoding)

CITE-seq





- Can be used
  - to combine cells from different pools in one experiment
  - detect presence of surface markers of cells
  - detect doublets

Molecule numbers in cells

Table 1. RNA content of a typical human cell

Total RNA per cell	~10–30 pg
Proportion of total RNA in nucleus	~14%
DNA:RNA in nucleus	~2:1
mRNA molecules	$2 \times 10^5 - 1 \times 10^6$

Table 2. RNA distribution in a typical mammalian cell

RNA species	Relative amount
rRNA (28S, 18S, 5S)	80–85%
tRNAs, snRNAs, low MW species	15–20%
mRNAs	1–5%

Table 3. mRNA classification based on abundance

Abundance class	Copies/cell	Number of different messages/cell	Abundance of each message
Low	5–15	11,000	<0.004%
Intermediate	200–400	500	<0.1%
High	12,000	<10	3%

https://www.gene-quantification.de/

From the cell to the RNA library

- ~ 30% of the cells are lost in the fluidics …
- Major loss point: Reverse transcription (RT)
  - Estimate: 5 – 20% of the mRNA make it to the cDNA
- Full length cDNA amplification:
  - 11 – 13 cycles
- Fragmentation & PCR amplification:
  - Depending on the amount of generated cDNA: 6 – 15 cycles

→Many duplicates; UMIs are essential to improve accuracy

→Many dropouts for low abundance mRNA

Data characteristics

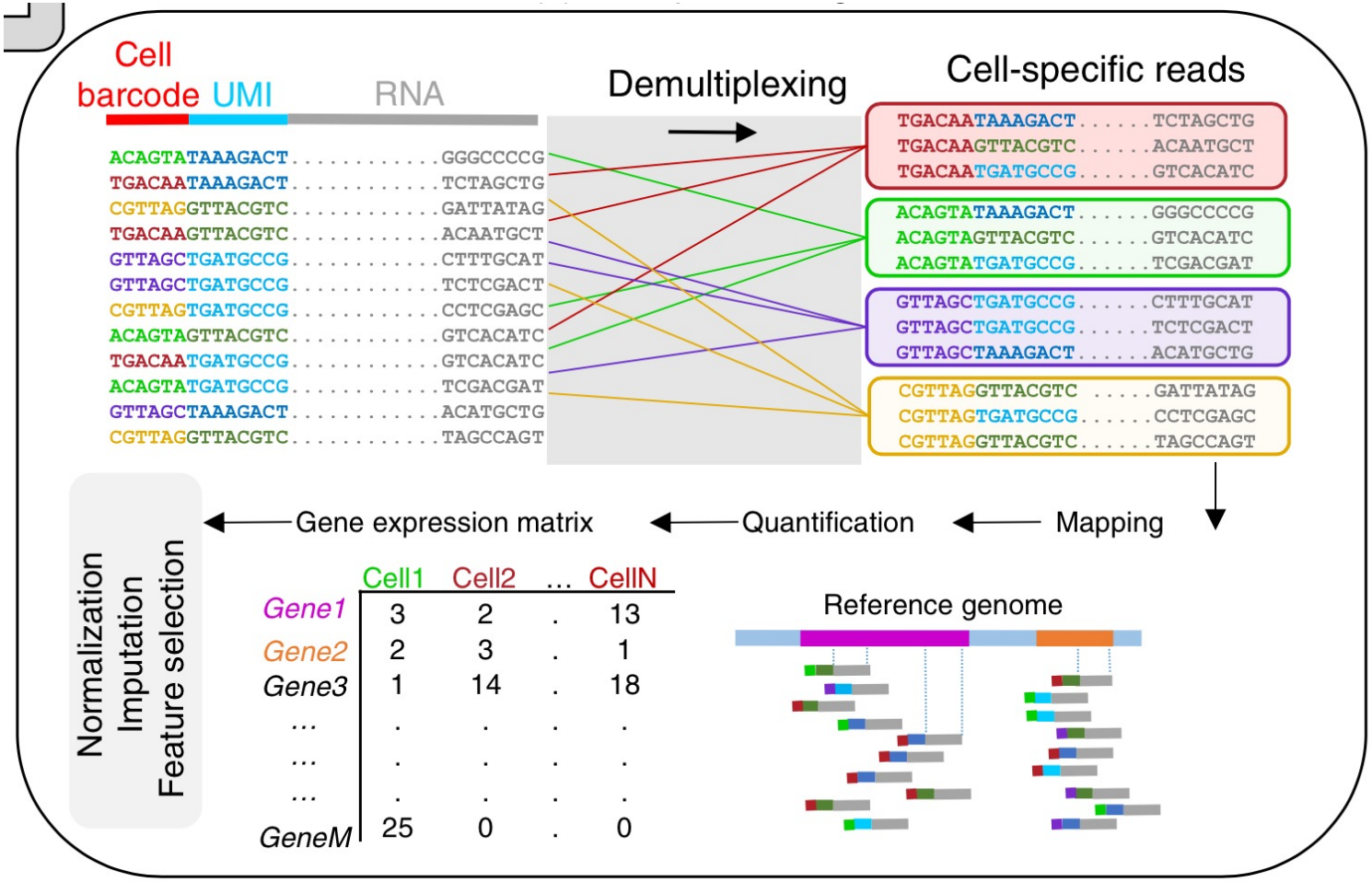
- bulk RNA-seq:
  - 10k to 10Mio cells → billions of mRNA molecules → 100s of billions of fragments after fragmentation and amplification → sequencing samples ~20 Mio reads out of a much larger pool of fragments
  - 10-20k different genes detected in a sample
- single cell RNA-seq
  - 1 cell → 200k mRNA molecules
  - 70-90% of mRNA molecules are typically lost during library prep; precise numbers are unknown
  - sample prep **overamplifies deliberately**, i.e. sequencing is expected to capture multiple reads that originate from the same mRNA fragments; without strong amplification the detection sensitivity of current technologies is too low
  - protocols include Unique Molecular Identifiers (UMIs) that allow to identify and deduplicate reads
  - 200 – 8000 different genes detected in a cell
  - transcription happens in **transcriptional bursts** → considerable cell-to-cell variability

UMIs and 3'-tagging

- The use of UMIs implies that only one end of the transcript can be captured! Expression quantification is only at the **gene level** not at the isoform level
- bulk RNA-seq:
  - typically, coverage of entire transcript body
- single cell RNA-seq
  - transcript end tagging and UMI deduplication

  or:
  - whole transcript coverage but increased variability caused by clonal reads

Generating the single cell count matrix

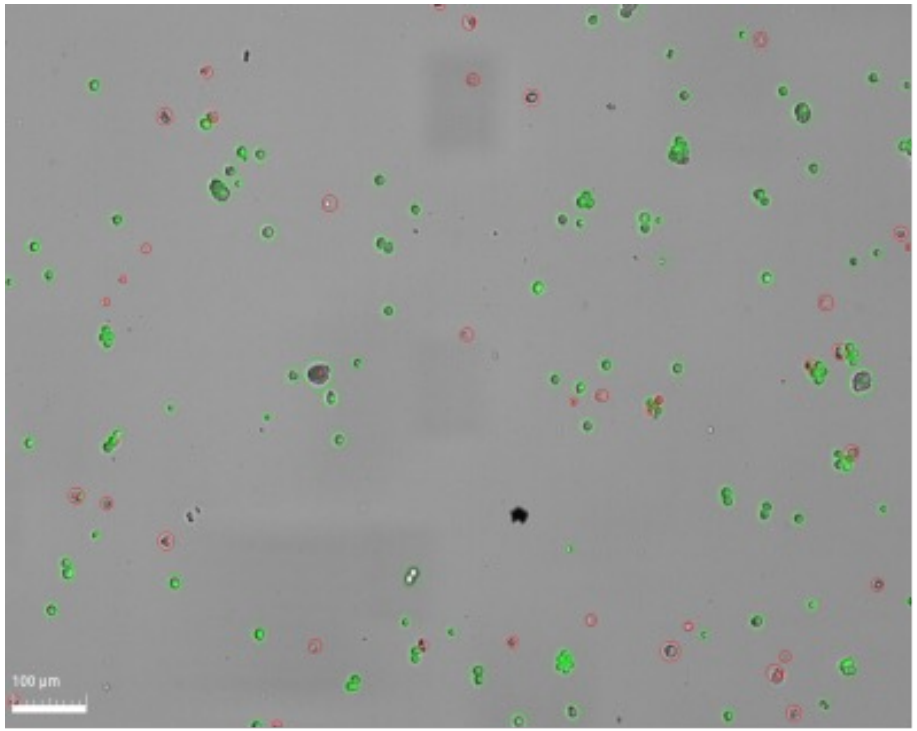


- Expression matrix cells vs genes
- Measurement does not capture the identity of the cells
- **Identity of cells needs to be derived from gene expression**
- Up to 95% of the expression matrix may be zero

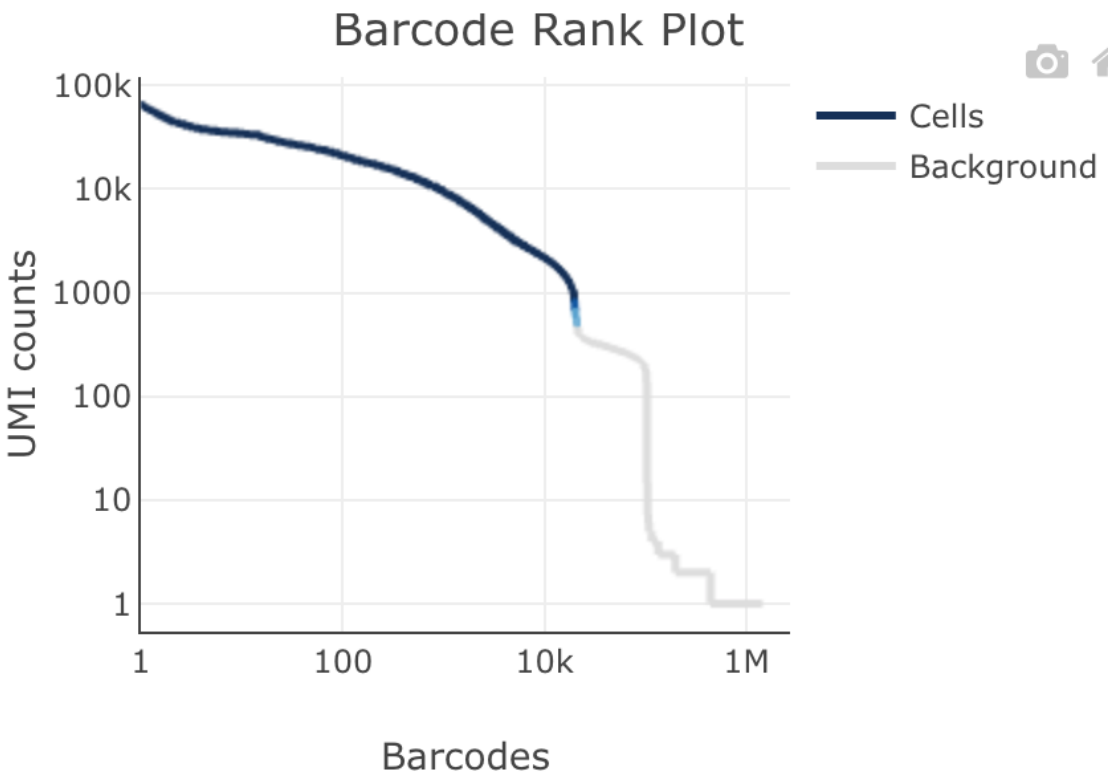
Cell viability

- Microscopy of cell suspension
  - Green: live
  - Red: dead
  - Unstained: debris
- Cell viability is the key parameter that drives the quality (ideally above 90%)

Low quality example with ~80% viability



EmptyDrops algorithm



- In a typical single-cell 10X experiment >1Mio barcodes are detected
- These are error-corrected barcodes; error-correction is done by matching barcode reads against *known-good* barcodes from 10X
- Vast majority do not represent cells but are empty drops capturing few ambient mRNA molecules

Empty drops

- The expression matrix contains also columns/barcodes that correspond to empty drops
  - Reads in empty drops correspond to free-floating RNA in the suspension of single cells
  - Empty drops are expected to have few reads and the reads in empty drops all come from the **same distribution of ambient RNA**
- Implemented in Bioconductur package *DropletUtils* and 10X Genomics CellRanger software

Method | Open Access | Published: 22 March 2019

EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data

Aaron T. L. Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree & John C. Marioni

Genome Biology **20**, Article number: 63 (2019) | Cite this article

EmptyDrops algorithm

Steps:

1. Select y% of the barcodes with lowest total UMI counts (those that are for sure background)
   → assume they represent empty cells and build ambient RNA model
   → call barcodes that deviate from ambient model as valid cells
2. Additionally, call all cells above the *knee* point as valid cells

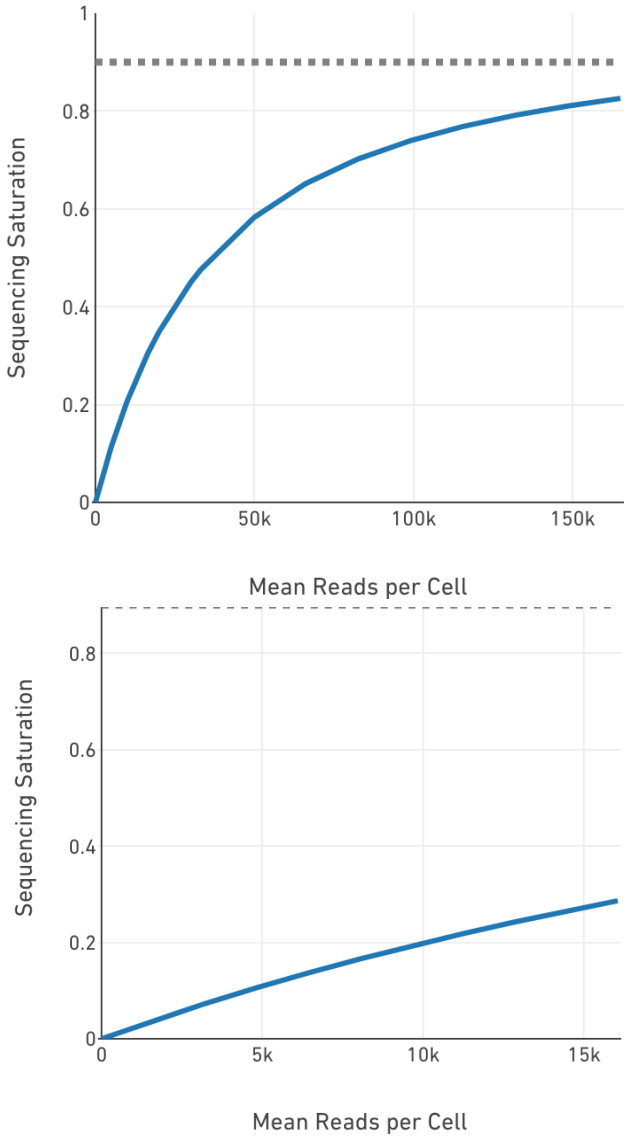
In essence barcodes are called *cells* if they

- have either sufficient reads (above knee-point)
- if they are significantly different from background distribution

Cellranger implements a similar approach, Cellranger is often too optimistic and includes too many barcodes as cells

Sequencing Depth

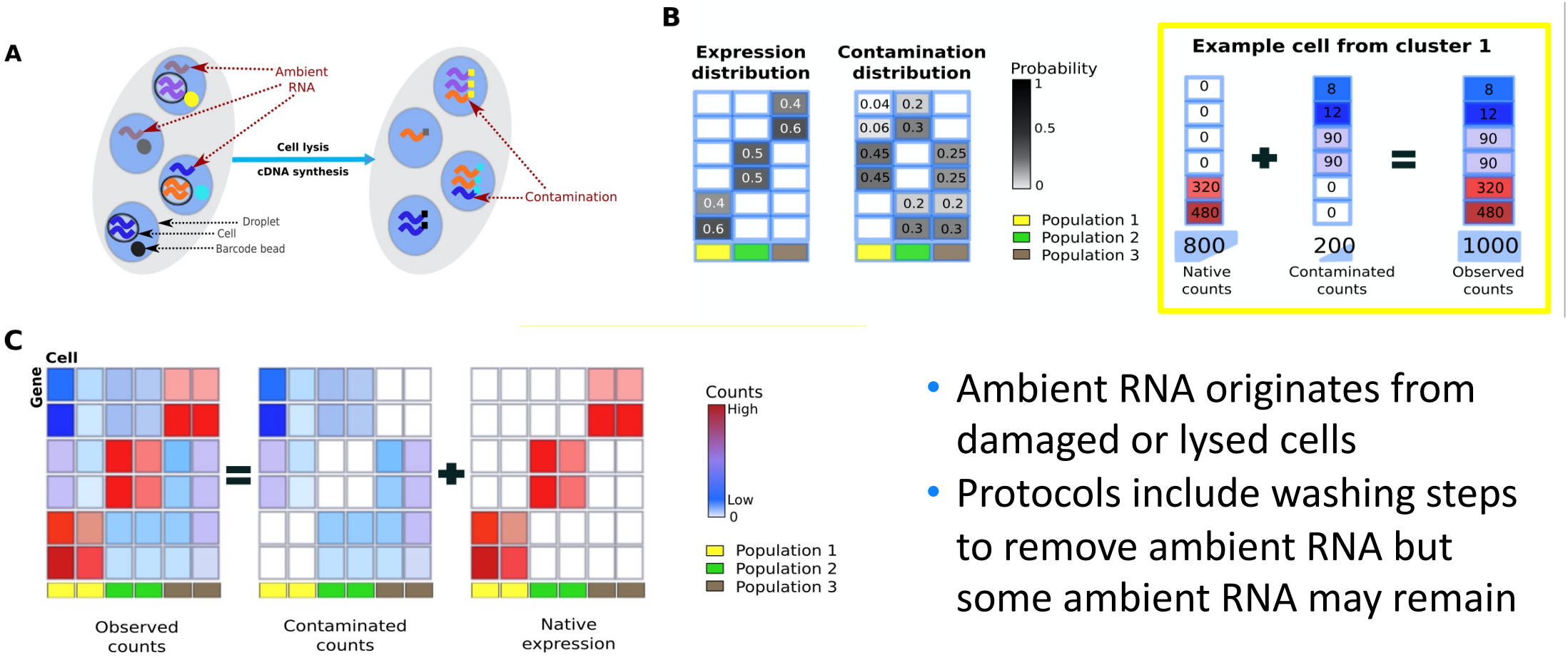
- Based on resampling one can estimate how many genes would be additionally detected if more reads were sequenced
- Typically, 50’000 reads per cell are targeted, since the yield in terms of detected cells may vary the actual number of reads per cell can be very different



Ambient RNA

- EmptyDrops removes barcodes that do contain only ambient RNA
- Ambient RNA still may affect expression of true cells by an additive contribution
- This may be relevant for cells that have low transcriptional activity
- Ambient RNA contributions can be estimated (and removed)
  - SoupX
  - DecontX
  - CellBender

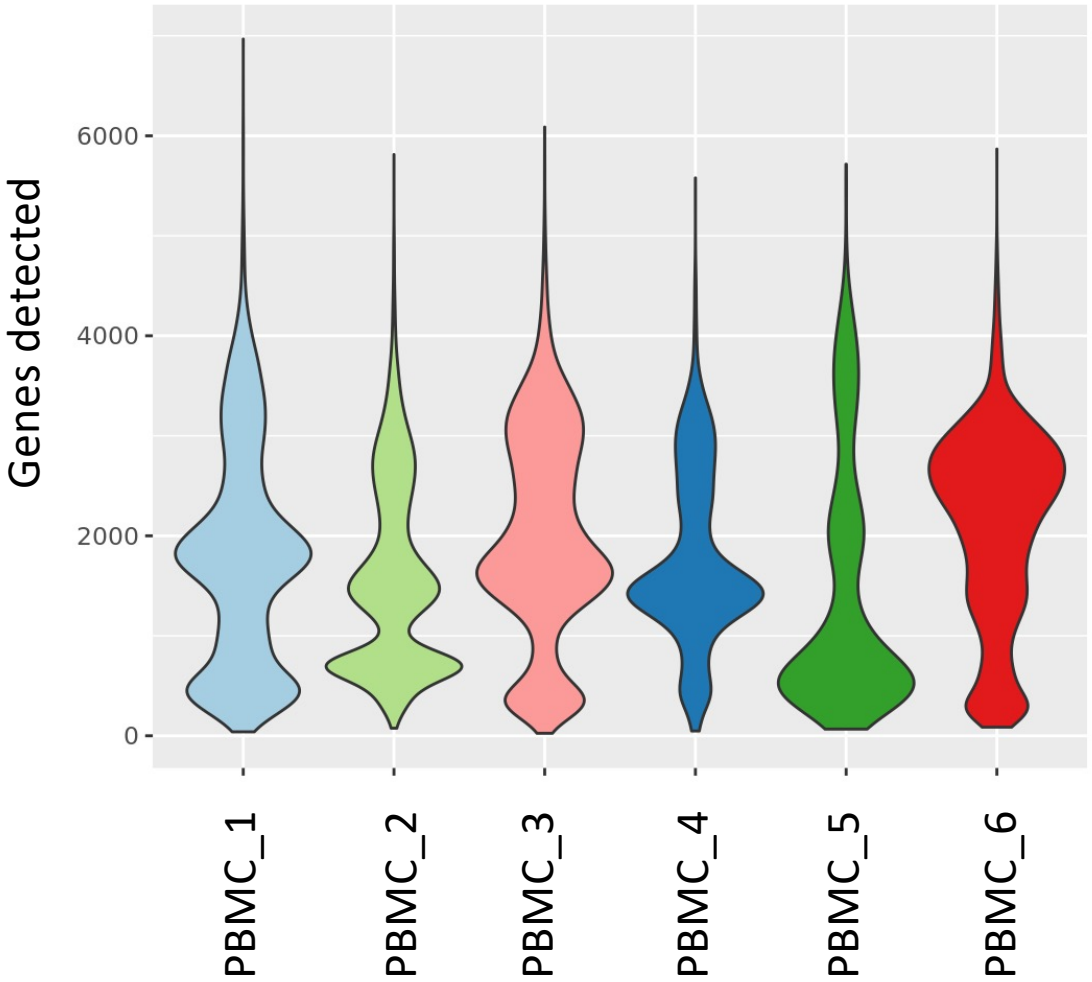
Ambient RNA



- Ambient RNA originates from damaged or lysed cells
- Protocols include washing steps to remove ambient RNA but some ambient RNA may remain

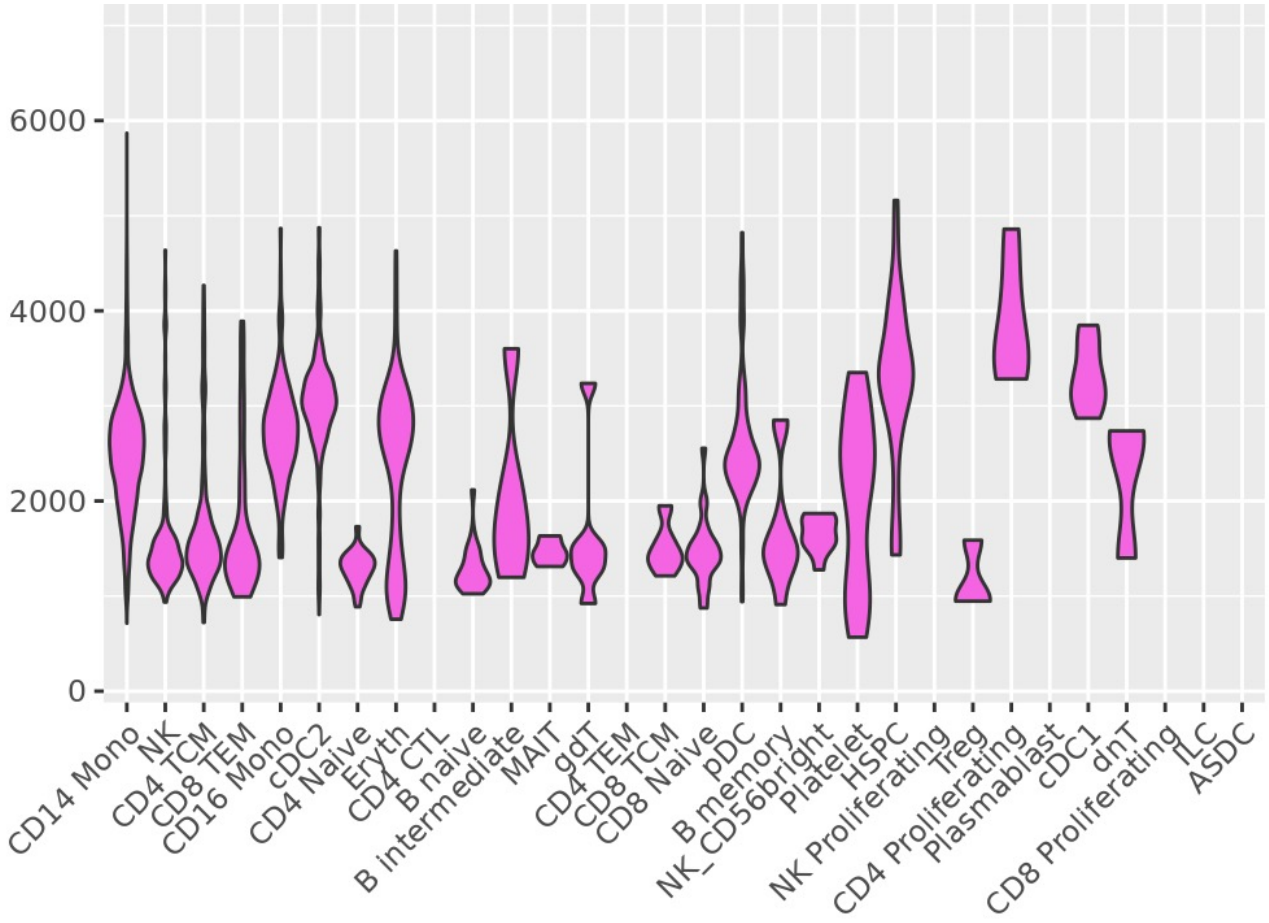
Yang et al., DecontX, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1950-

Filtering example: Peripheral Blood Mononuclear Cells (PBMCs)



- distribution
  - is multi-modal
  - differs between samples

Filtering example: PBMCs



- PBMCs consist of different cell types with varying overall transcriptional activity
- Risk: cell types with “few” genes expressed may be filtered out, e.g. neutrophils

Quality control: Stressed or dying cells

Look at specific gene groups (no hard thresholds available):

- **Mitochondrial genes:** Apoptosis, membrane leakage, damaged cells
- Other gene groups
  - Ribosomal protein genes (RPS*/RPL*)
  - Stress response genes (FOS, FOSB, …)
  - Apoptosis markers
  - Cell cycle markers: “M-phase only”
  - MALAT1: long-noncoding RNA that can have poly-A

Quality control: Low quality cell

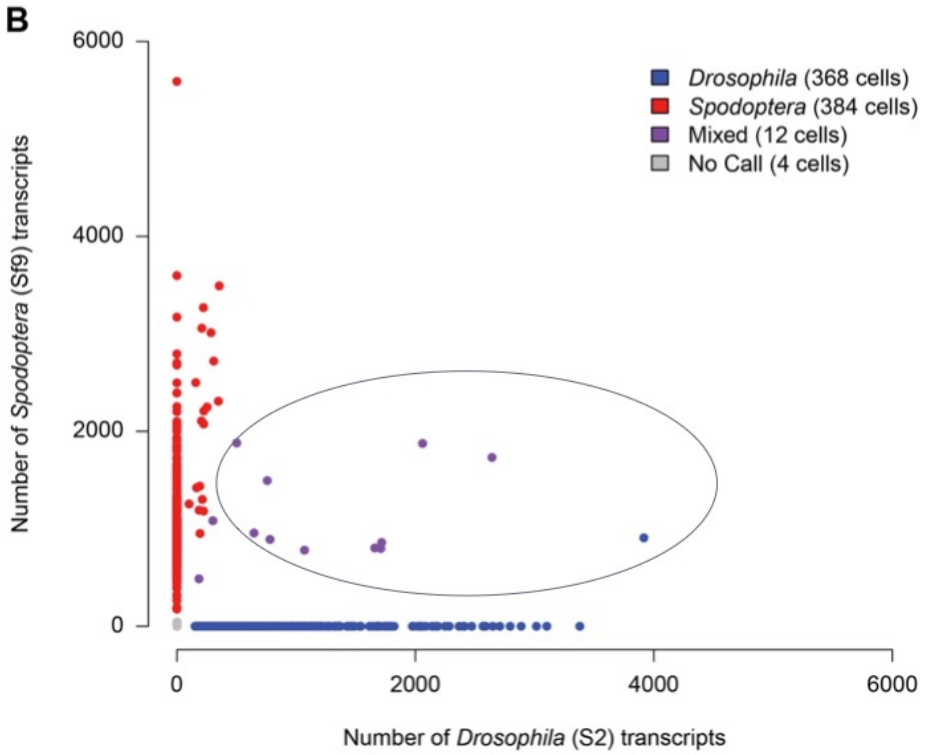
- Remove barcodes with
  - high fraction of mitochondrial genes (in apoptotic or partially lysed cells)
  - zero mitochondrial genes (cells that have lost the cytoplasm?)

- Potential confounder: number of mitochondria differs between cells
  - red blood cells: no mitochondria
  - liver cells: up to 2000 mitochondria

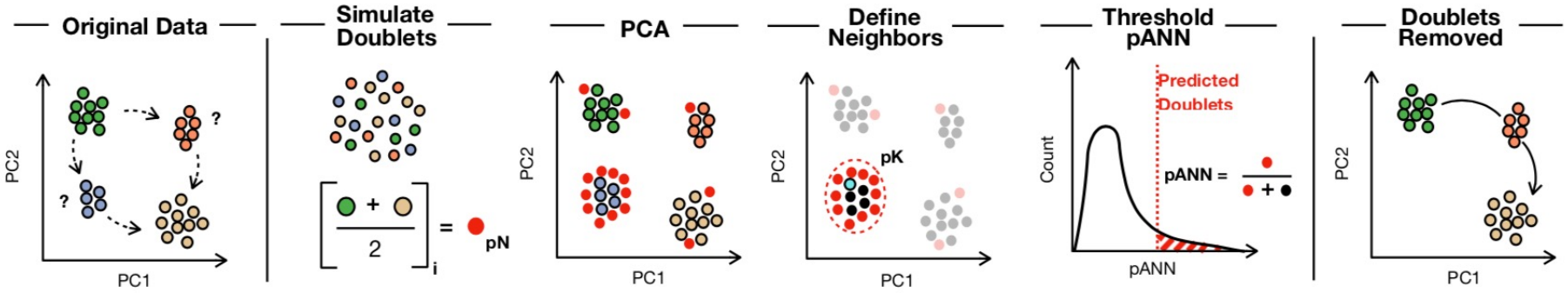
Quality control: Doublets

- Barcode collisions: not enough barcodes
- Technical doublets: two cells in the same droplet; tech specs of 10X Gennomics: +1% per 1000 cells
- Biological doublets: two cells sticking tightly together and form a unit; need to do single-nucleus RNA-seq

- Test datasets for doublets consist of cells from two species



Doublet Detection



- see also:
  https://github.com/plger/scDblFinder

https://github.com/chris-mcginnis-ucsf/DoubletFinder

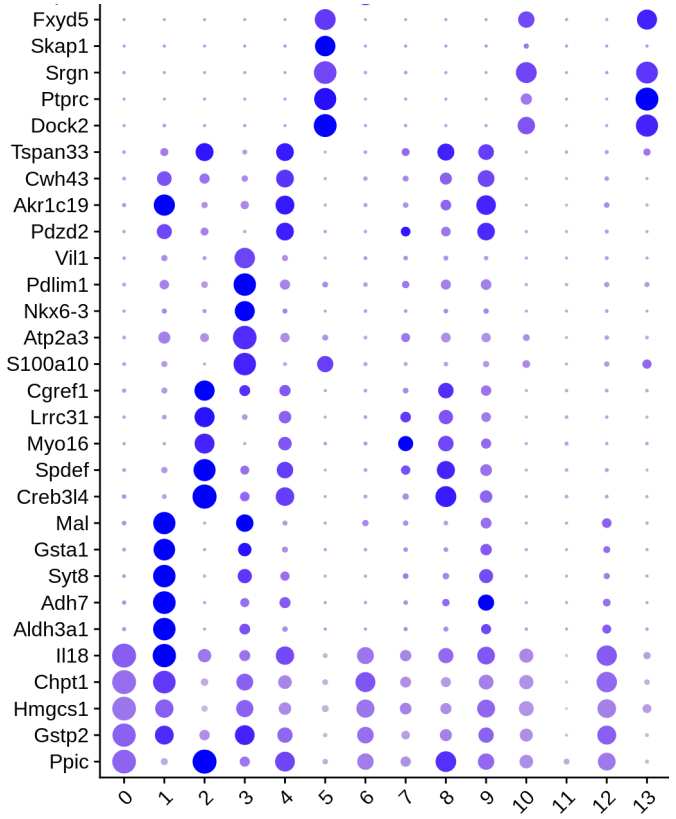
Quality control thresholds

- Quality metrics, like the number of reads or genes detected per cell, do depend on factors like
  - cell type
  - sequencing depth
  - overall quality
  - …
- As a consequence thresholds for these metrics need to be adapted to individual samples and cell types

Cluster Marker genes are not specific

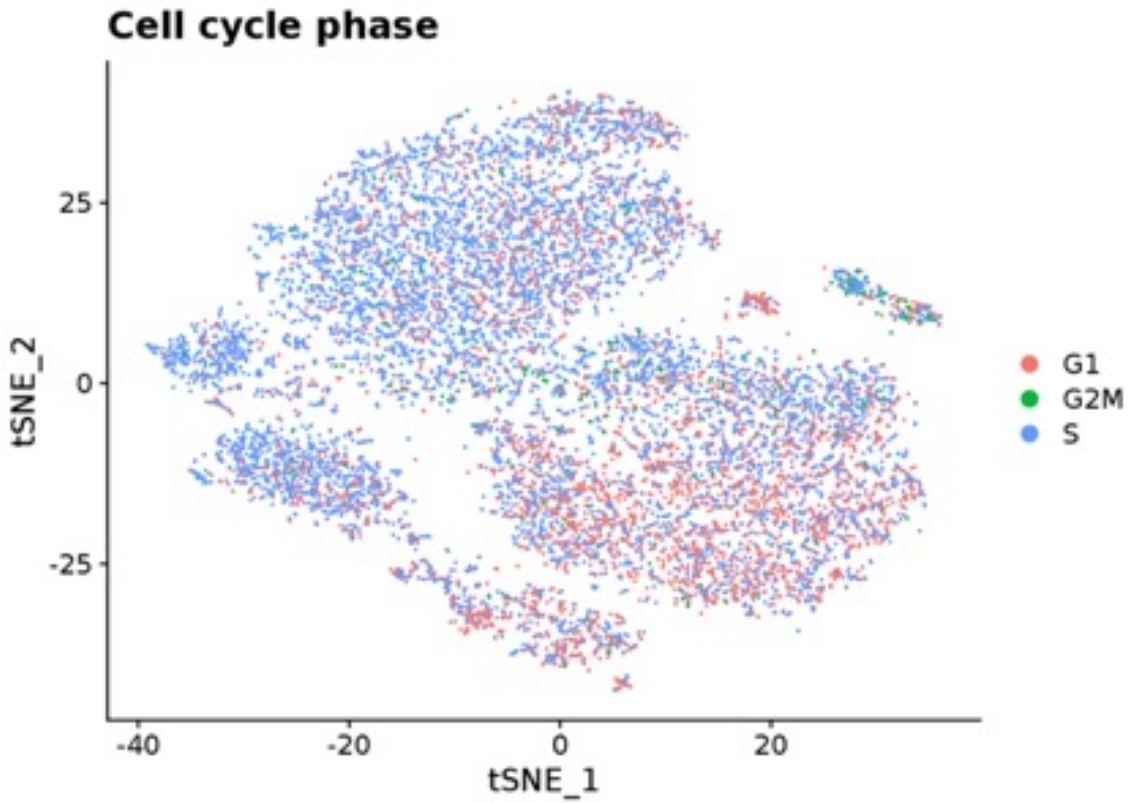
- If we compute cell clusters (== cell types) and the marker genes for these clusters (genes higher expressed in one cluster relative to all other clusters, then the top marker genes should be **cluster-specific** (only expressed in one cluster)
- Reasons for unspecific markers:
  - Too many clusters
  - Ambient RNA

Example of cluster marker genes, where cluster zero represents low count, low quality cell fragments or ambient RNA



Cell cycle: Cause of variation in gene expression

- Cell cycle can be a quality indicator too. Cells in M-phase might indicate dying or damaged cells
- Depending on the research goal, cell-cycle associated variation
  - can be used to discriminate cells
  - is unwanted variation and cell-cycle should be considered as latent variable

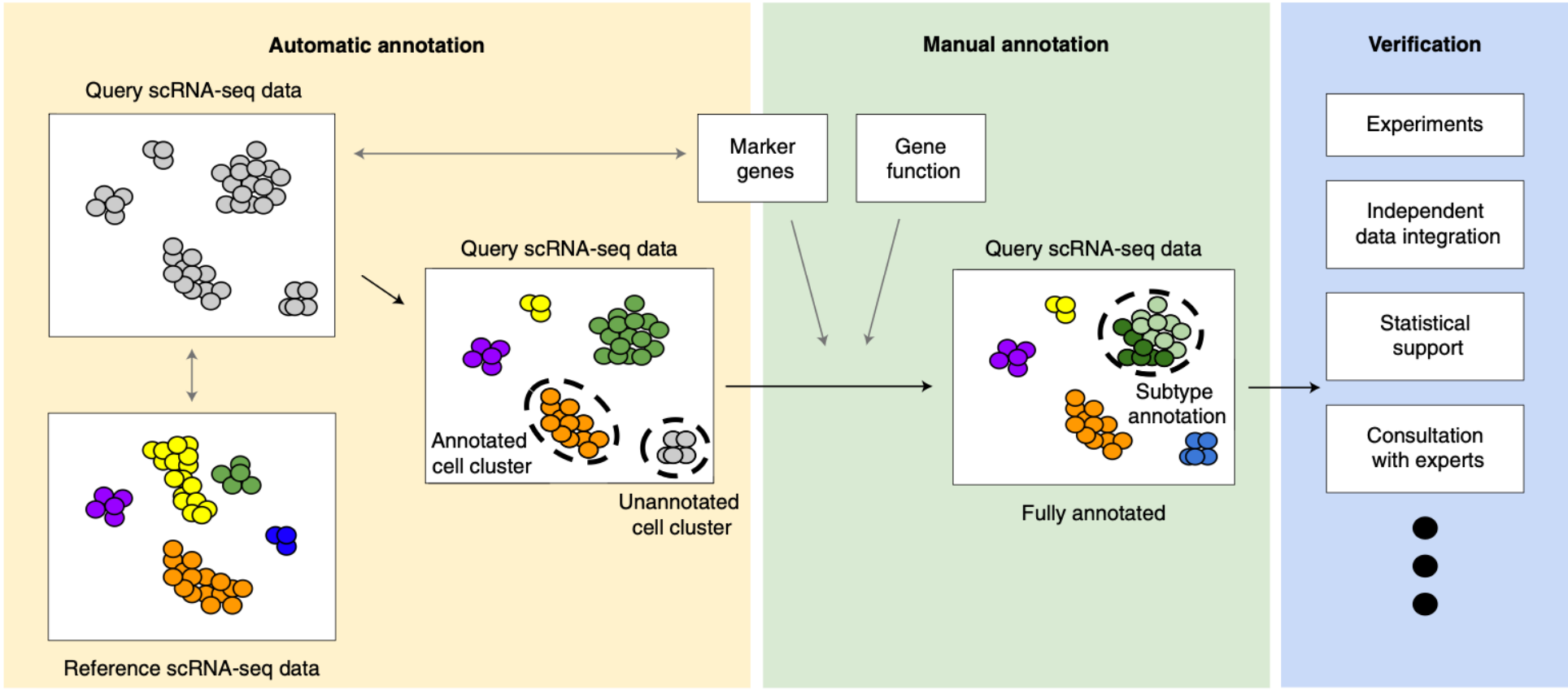


General Quality Measures

- General quality measures
  - Cell viability (%)
  - Knee plot
  - Reads in cells (%)
  - Sequencing depth & sequencing saturation
- Indicate low quality, but are often not actionable

Per Cell Quality Measures

- # UMIs per cell
  - Low: cell fragment; high: doublet
- # genes per cell:
  - Low: cell fragment; high: doublet
- Percent mitochondrial reads:
  - Apoptosis, membrane leakage, damaged cells
- Percent ribosomal protein genes (RPL/RPS)
- Cell-cycle M-phase genes
- Ambient contamination score



Clarke et al. 2021

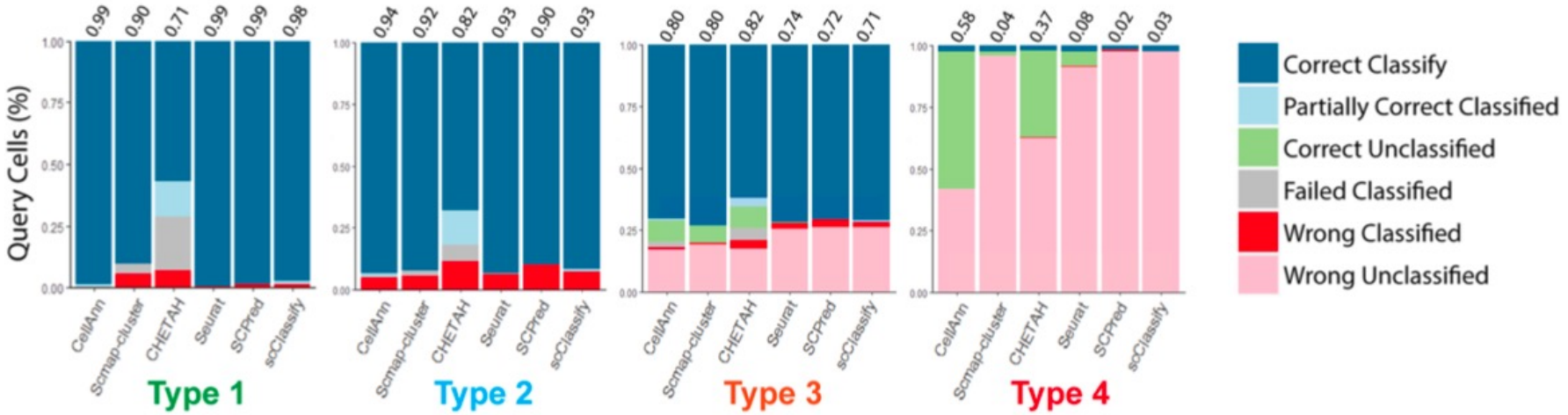
Cell type: annotation approaches and tools

	marker genes	reference expression profiles
per cell	• **AUCell** • scROSHI	• **SingleR** • scArches • **Azimuth** • scmap-cell
per cluster	• **enrichR** • **scType**	• **SingleR** • FR-Match • CellAnn • scmap-cluster

Per-cell vs per-cluster labeling

- generally, both approaches agree for the major cell-types
- discrepancies occur when trying to identify sub-types
- major discrepancies might be explained by
  - cluster resolution inappropriate
  - reference data / marker gene set inappropriate
  - subtype of cells
- discrepancies suggest manual resolution

Reference-based Annotation



- generally good performance
- situation where query and reference do not overlap are of a concern

Lyu, CellAnn paper, 2023

Reference Based Celltype Annotation

- **CellAnn:** compute **correlations** of cluster averages, analyze correlations to define thresholds, compute rank-test to define sub-types
- **Seurat:** integrate data in UMAP space and assign nearby cell type
- **scmap-cell:** approximate **nearest neighbors**
- **SingleR:** identify variable genes in the reference set, **correlate** individual cells with the reference using the variable genes

Single Cell Reference Data

- "Cell-by-gene": https://cellxgene.cziscience.com
- https://atlas.brain-map.org/
- https://www.humancellatlas.org/
- https://www.flycellatlas.org/
- https://azimuth.hubmapconsortium.org/references/

- Single Cell Expression Atlas: https://www.ebi.ac.uk/gxa/sc/home
  - individual studies
- BROAD:
  - https://singlecell.broadinstitute.org/single_cell
- ...

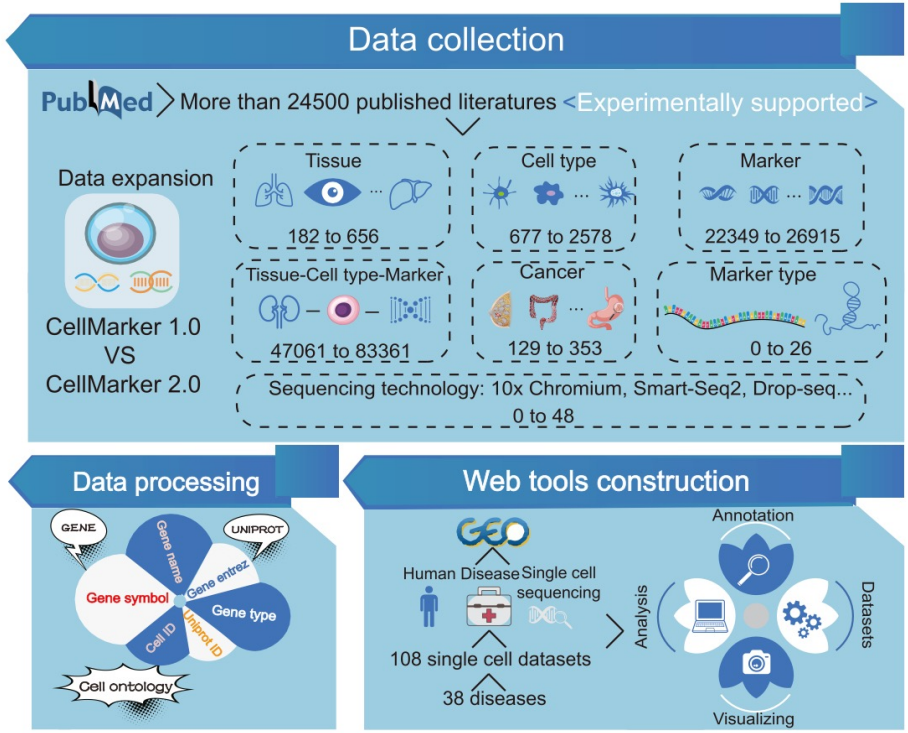
Generally: The more similar the reference is to your system, the better

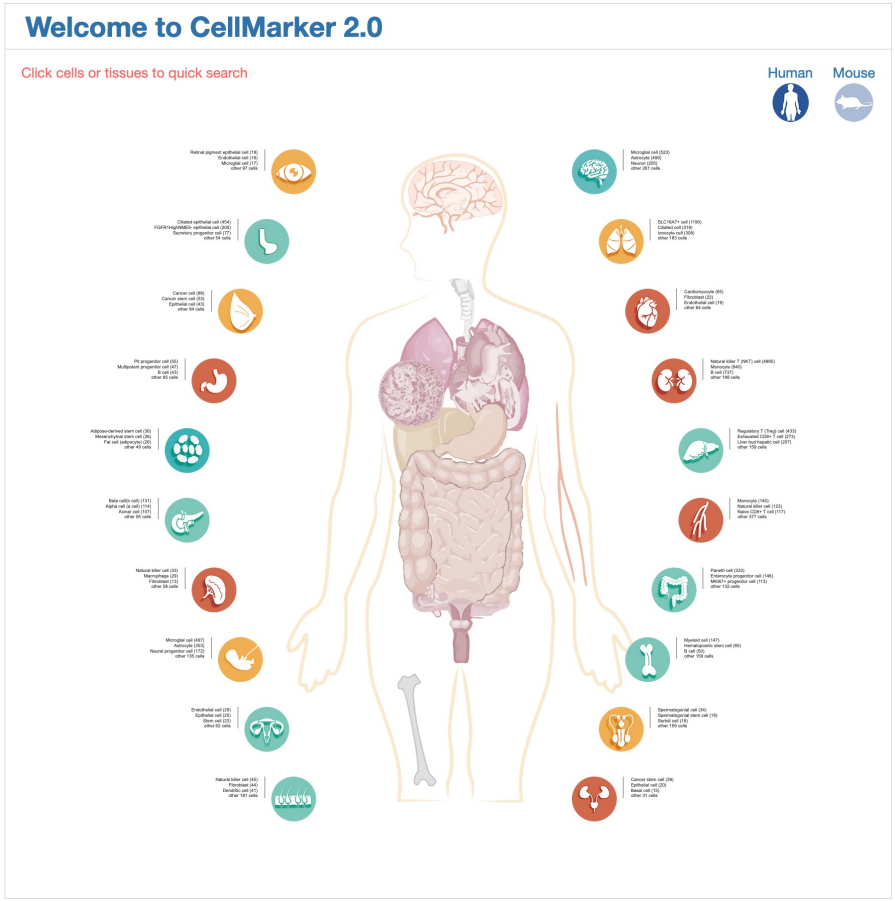
Marker Gene Based Celltype Annotation

Tools:

- **AUCell:** Area und the curve to estimate marker enrichment
- **enrichR:** compute cluster-markers, subsequently compute fuzzy enrichment among cell-type markers
- **scROSHI:** rank test to identify if cell-type markers are higher expressed than others

CellMarker 2.0

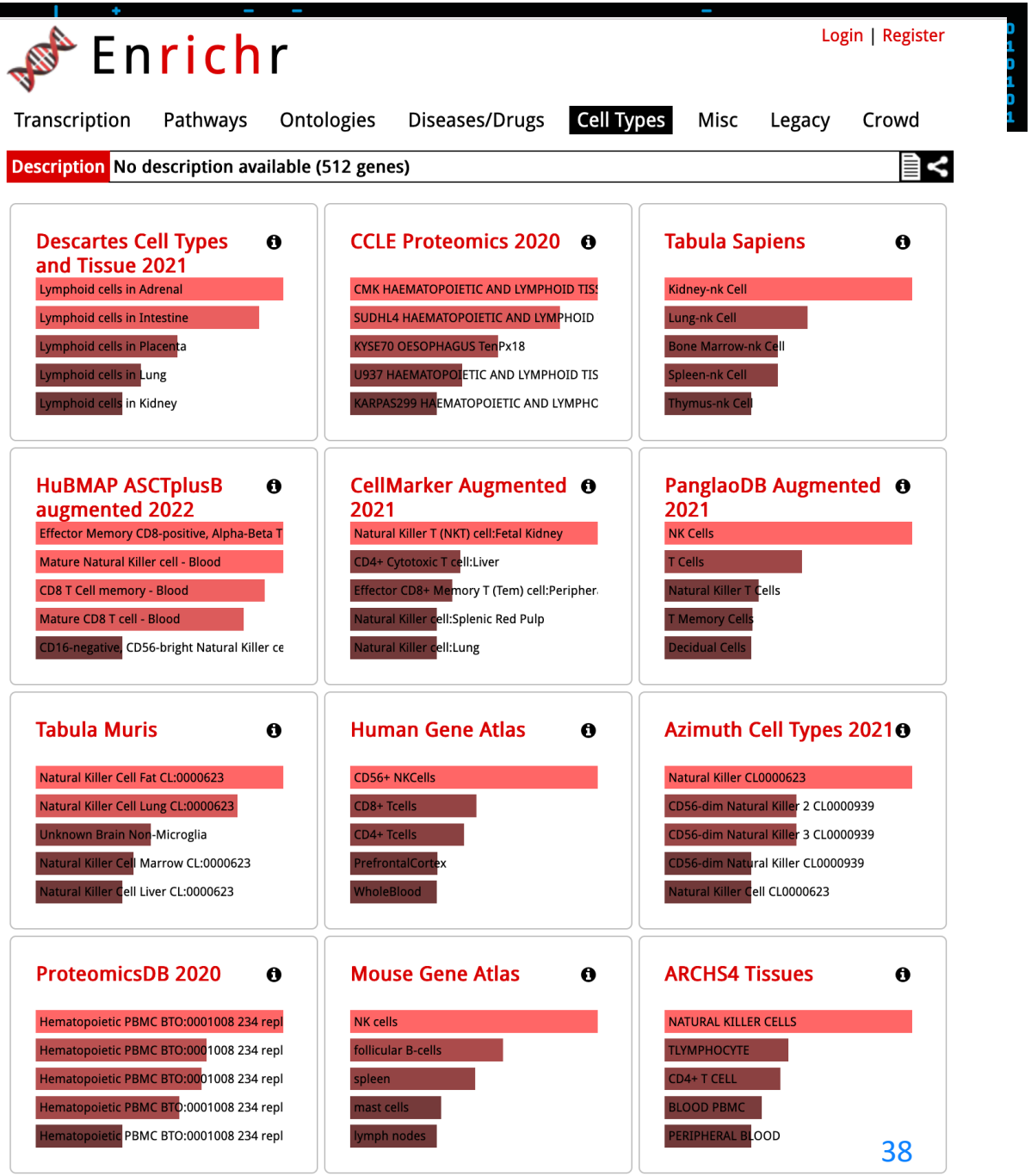




EnrichR

- database of databases
- diverse source of gene lists
- hosts also gene sets specific to cell types
- Gene sets are defined for human/mouse

- EnrichR databases for model organisms exist but do not include cell type specific gene sets

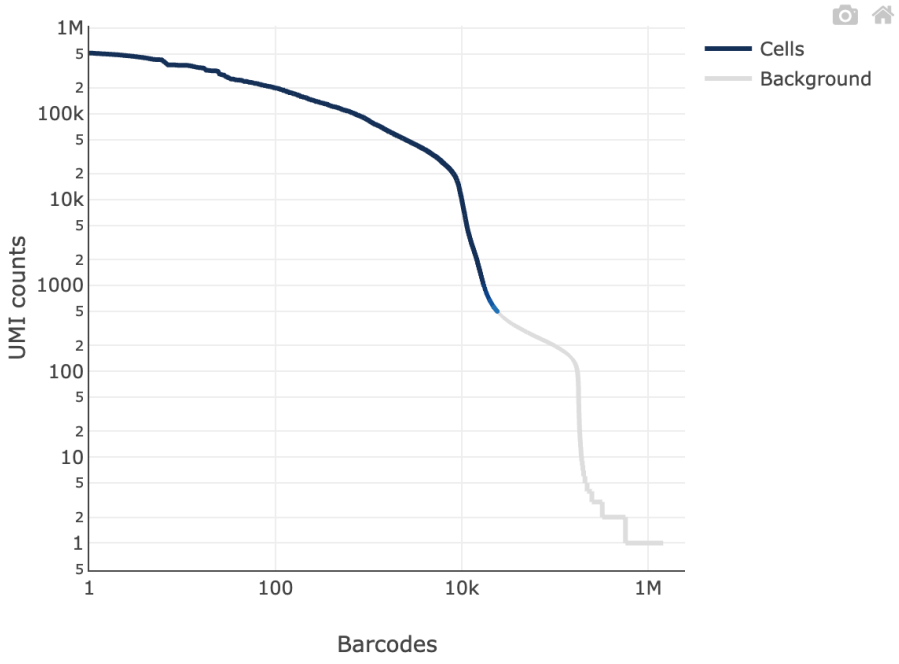


Summary

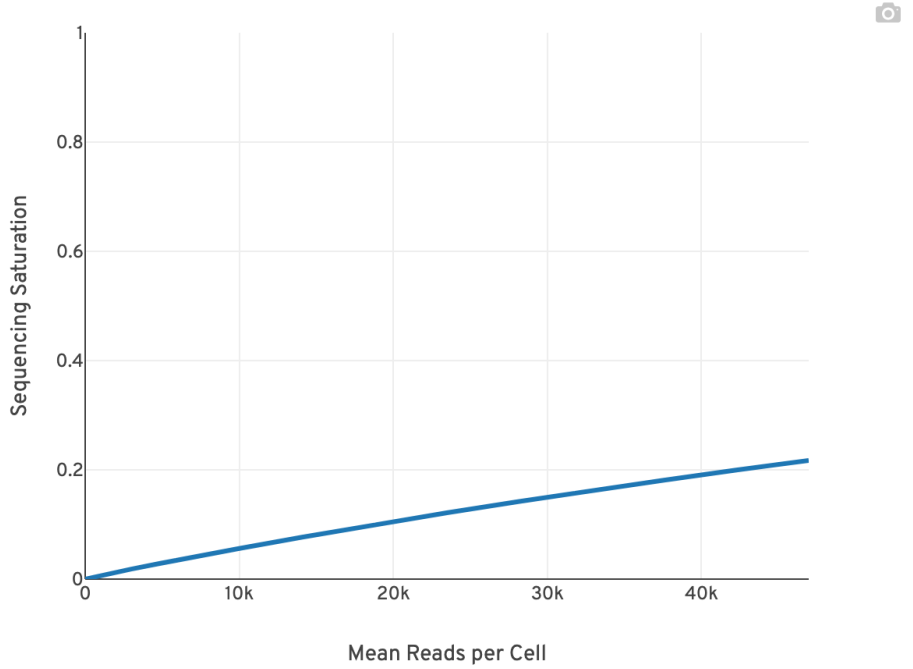
- Quality control is an essential step in single cell processing
- Cell-type annotation typically requires manual curation
- Major cell-type are well automatically detected, subtypes however are often misclustered and misassigned in automatic workflows

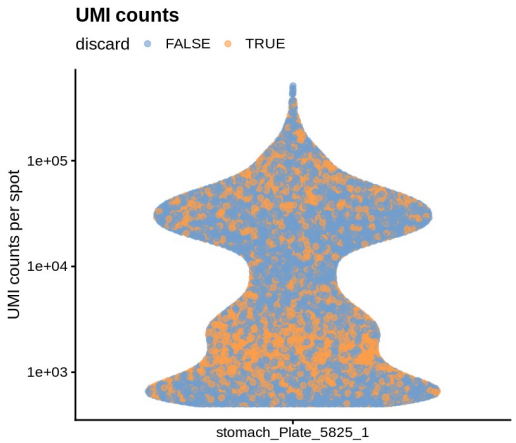
Low Quality Example

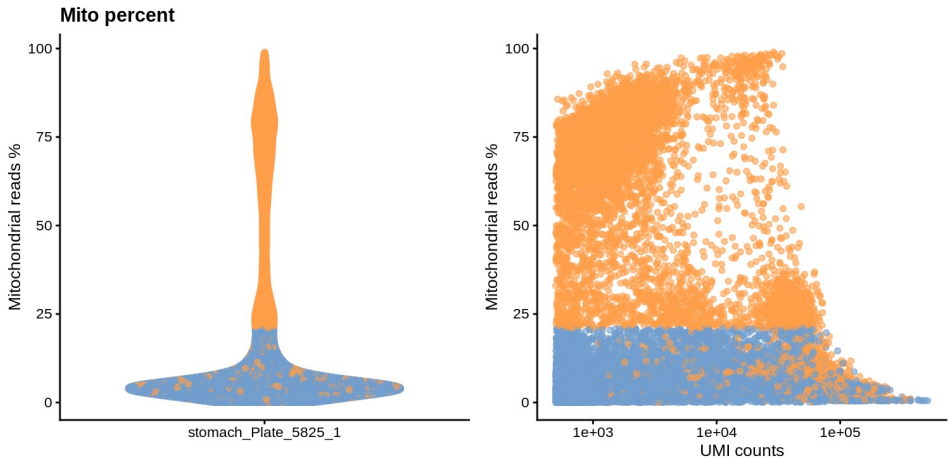
GEX Barcode Rank Plot

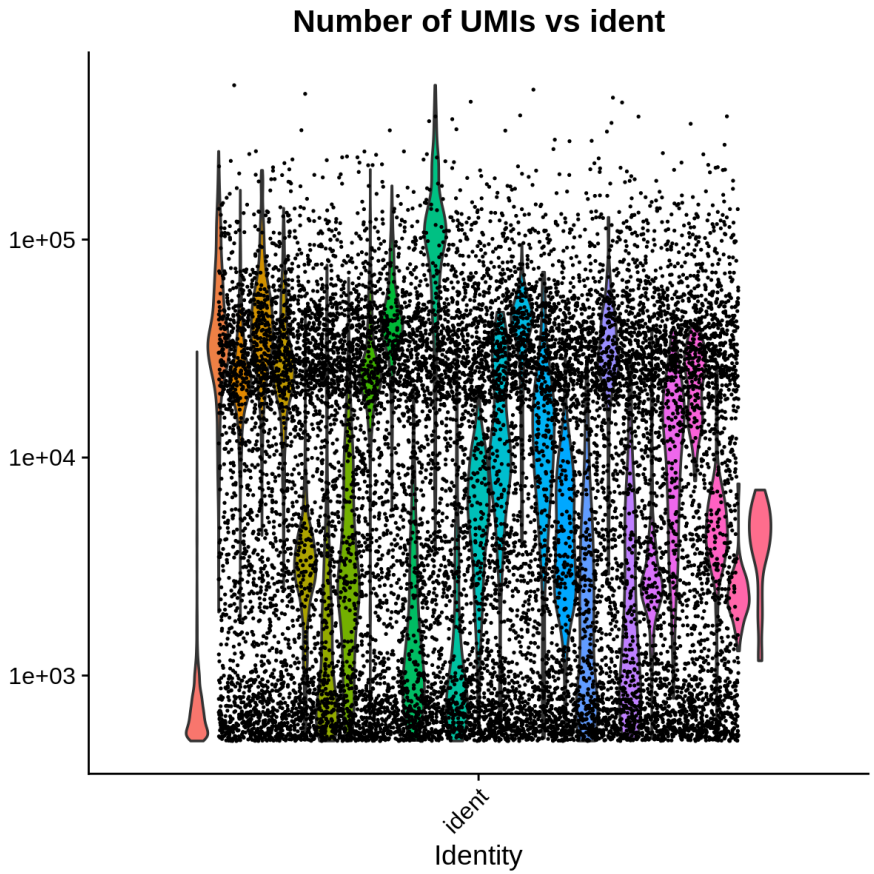


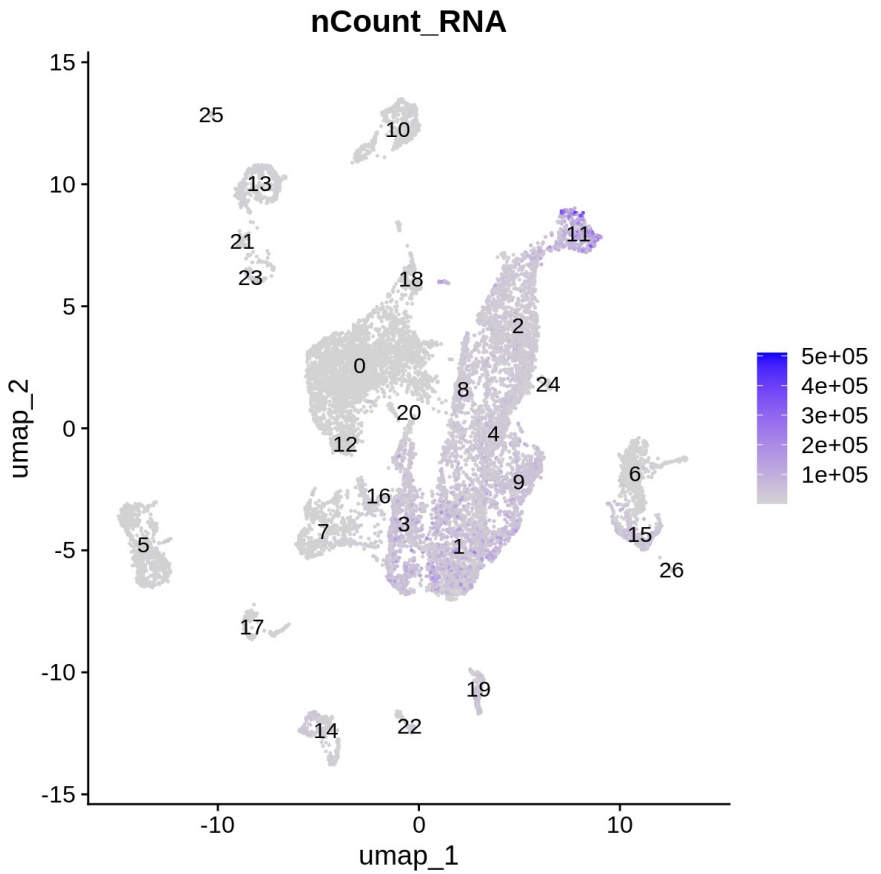
Sequencing Saturation

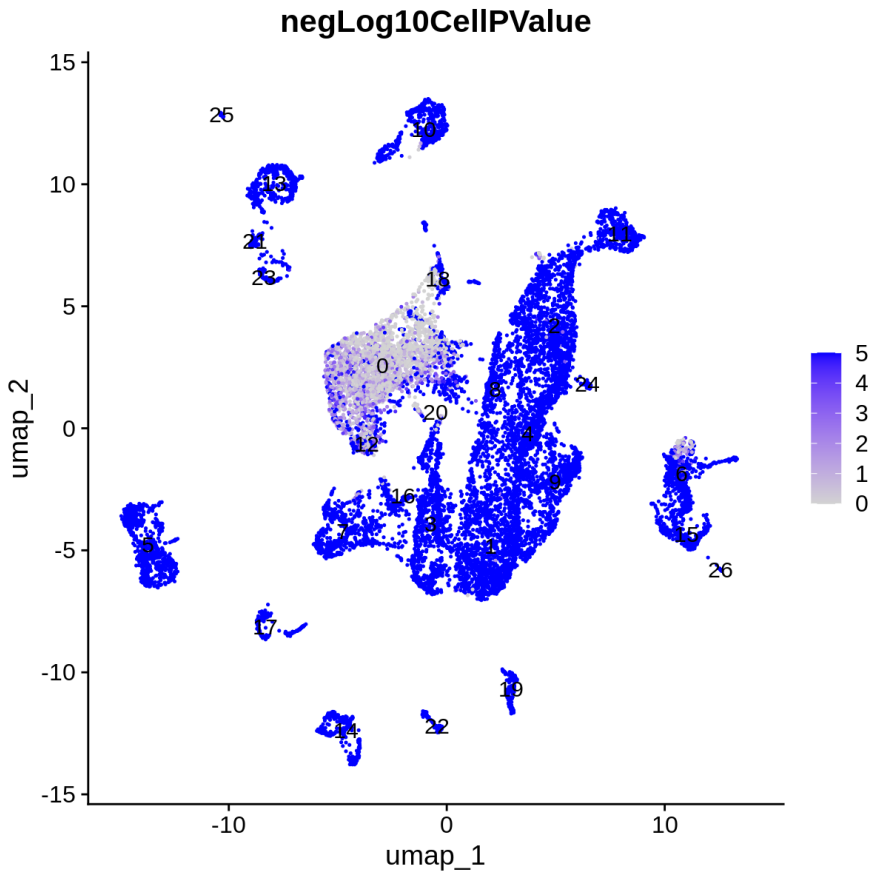



UMI counts
discard FALSE TRUE
UMI counts per spot
1e+05
1e+04
1e+03
stomach_Plate_5825_1


Mito percent
Mitochondrial reads %
100
75
50
25
0
stomach_Plate_5825_1
1e+03
1e+04
1e+05
UMI counts


Number of UMIs vs ident
1e+05
1e+04
1e+03
ident
Identity


nCount_RNA
umap_2
umap_1
15
10
5
0
-5
-10
-15
-10
0
10
25
10
13
11
21
23
18
2
0
8
24
20
4
12
9
6
16
3
7
1
5
15
26
17
19
14
22
5e+05
4e+05
3e+05
2e+05
1e+05


negLog10CellPValue
umap_2
umap_1
15
10
5
0
-5
-10
-15
-10
0
10
5
4
3
2
1
0