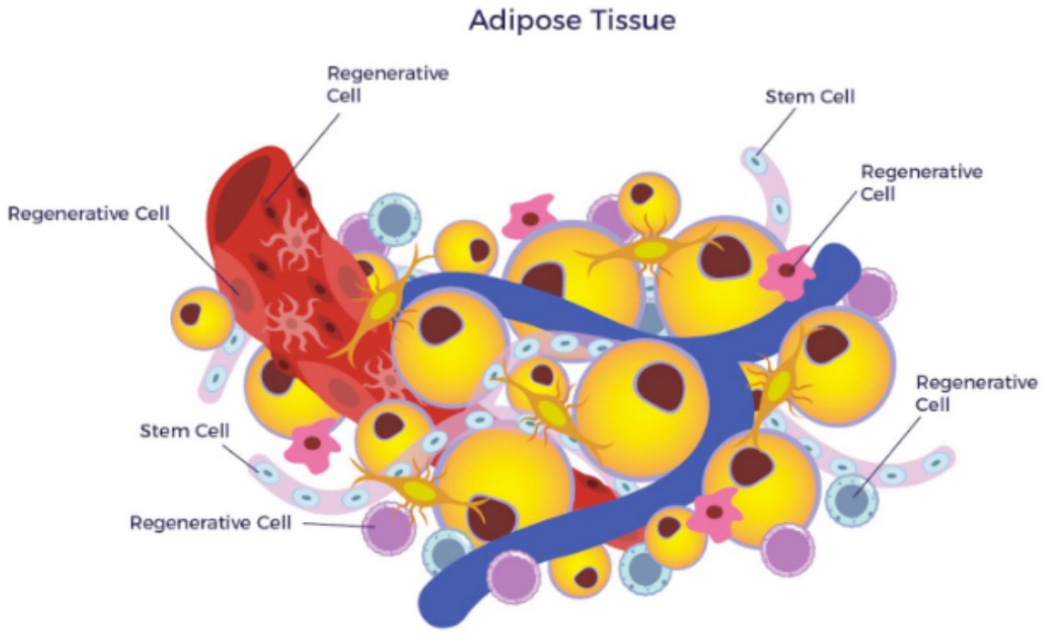
Single Cell RNA-seq: Characteristics, Preprocessing and QC

Hubert Rehrauer

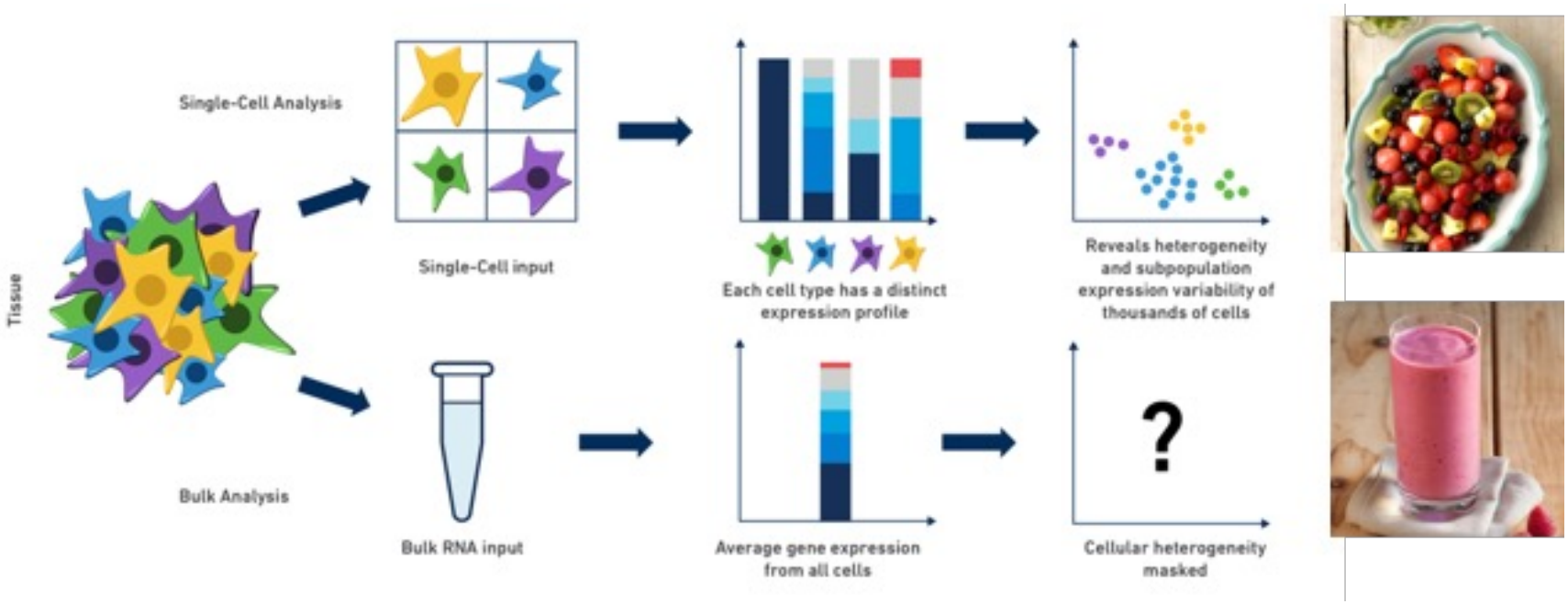
University of Zurich[UZH]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

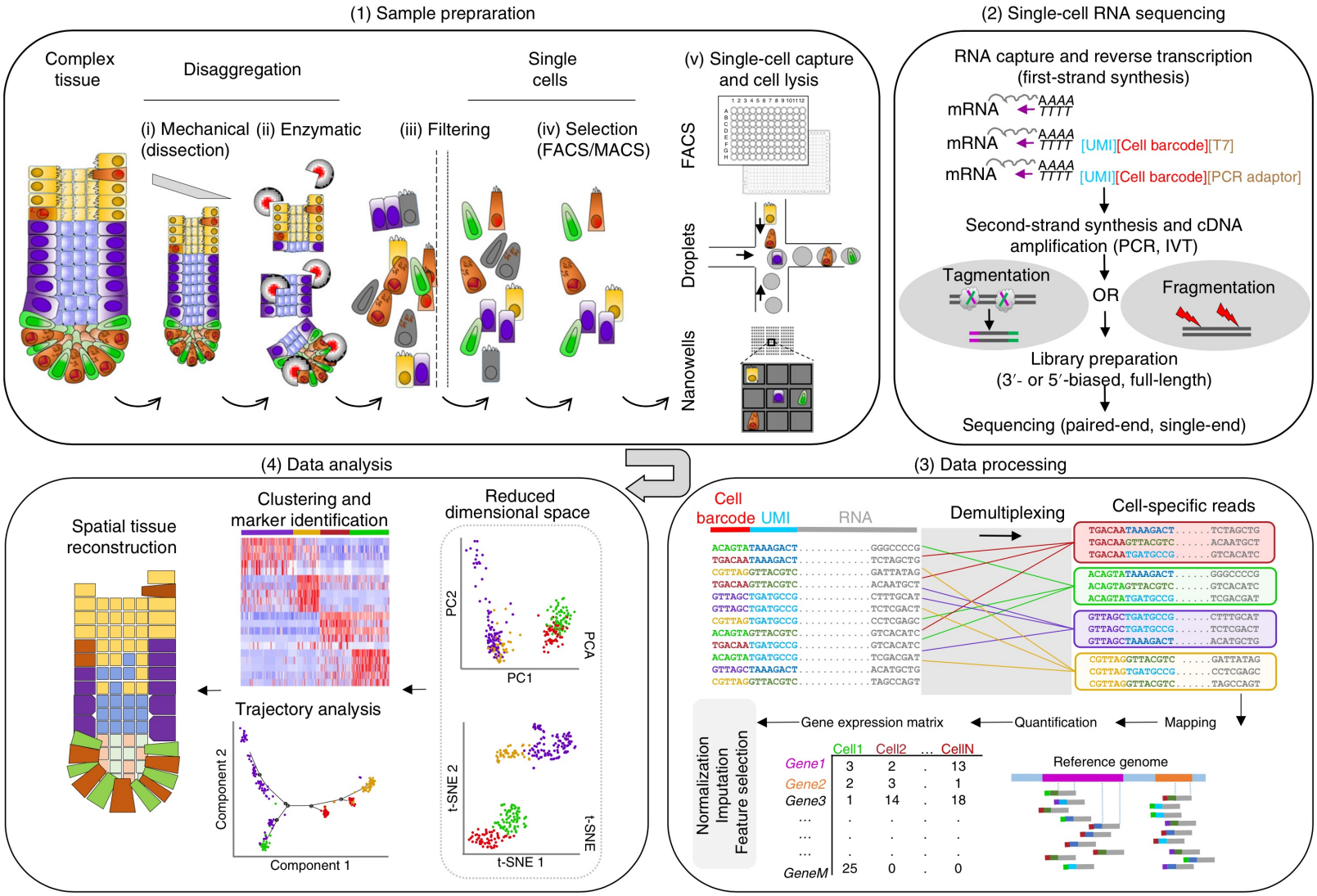
Tissues are heterogeneous



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

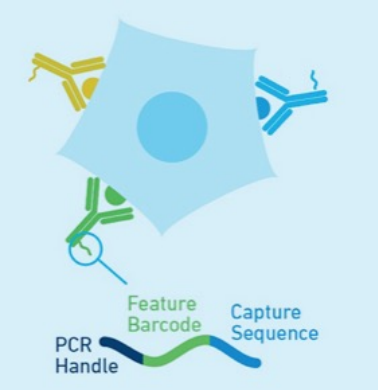
scRNA-seq

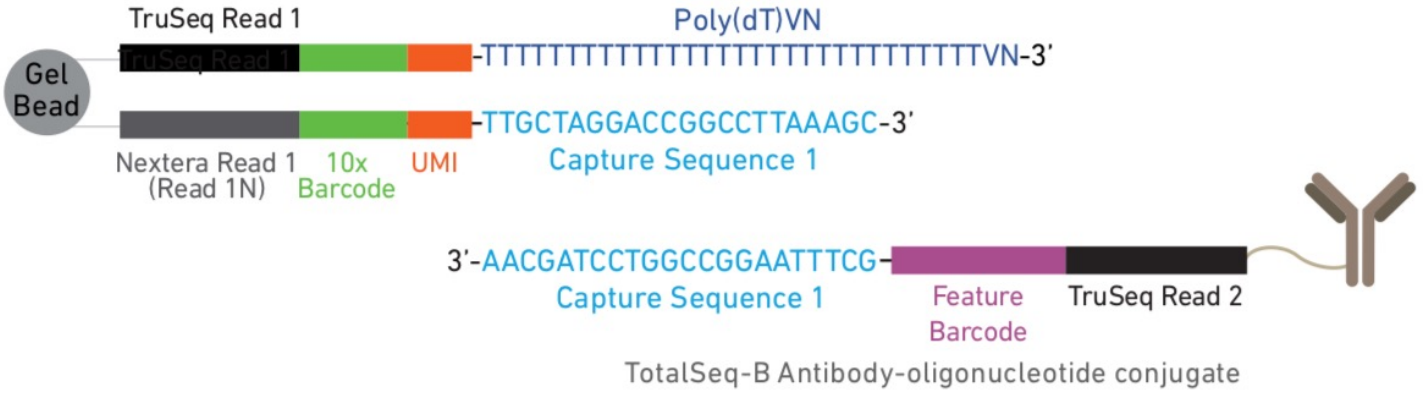


Lafzi et al. https://www.nature.com/articles/s41596-018-0073-y

Cell hashing (feature barcoding)

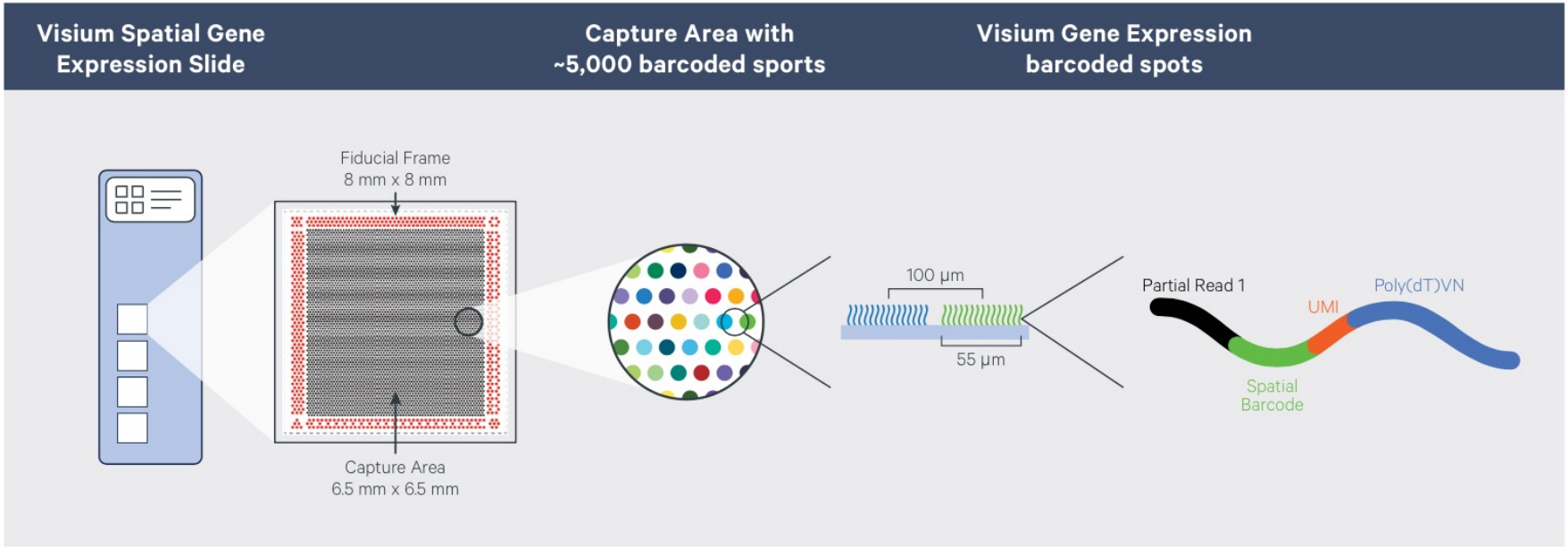
CITE-seq





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

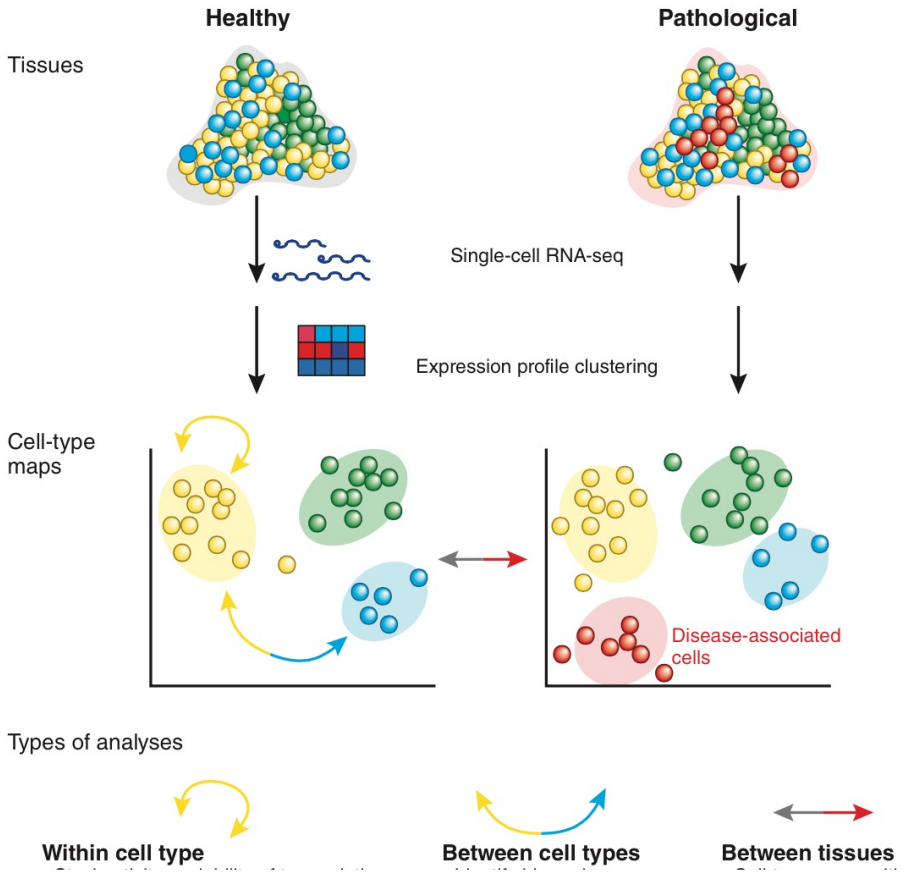
scRNA-seq with location information (spatial transcriptomics)



- can be combined with imaging the tissue slice
- each spot is characterized by expression profile + image

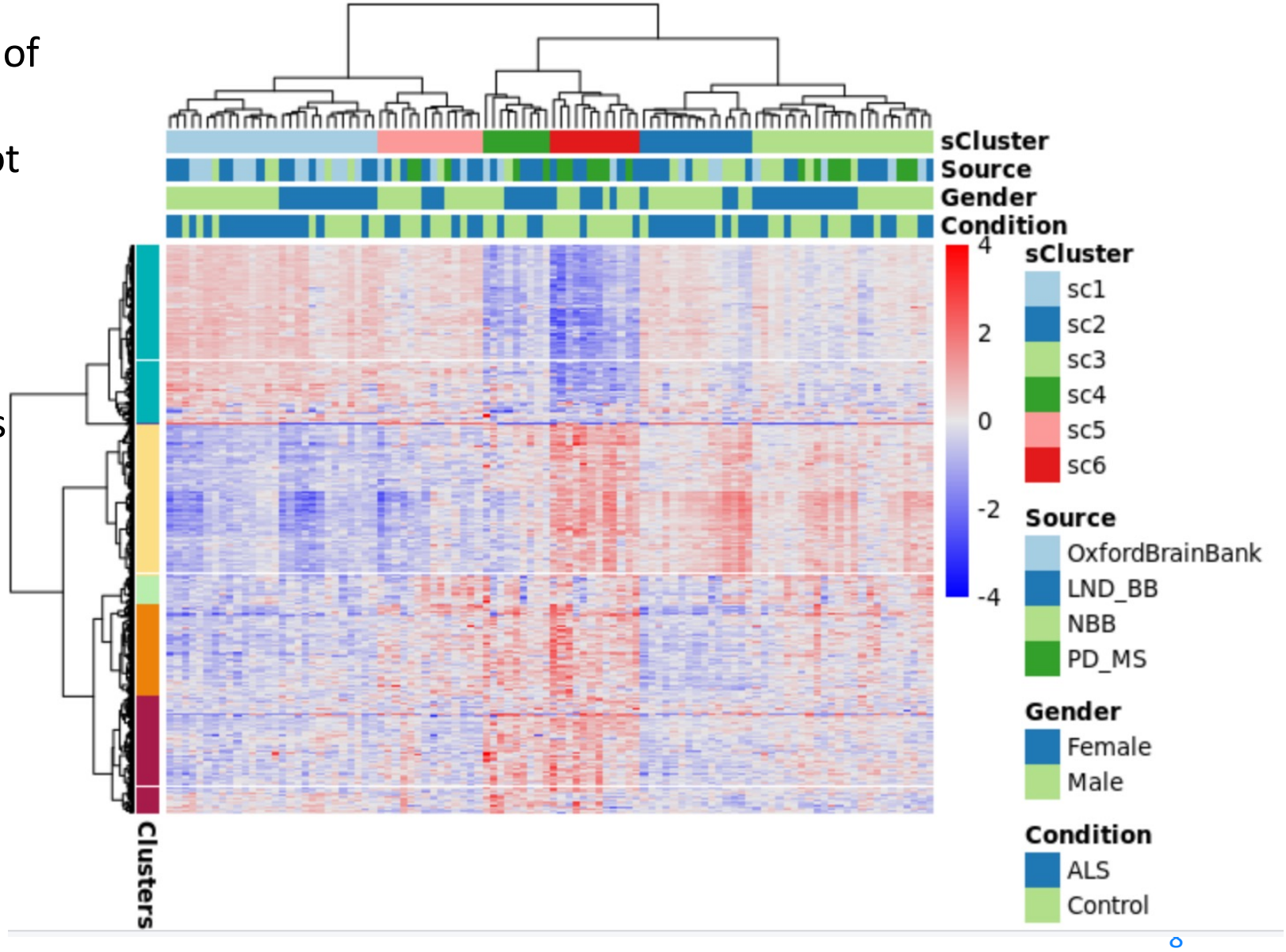
scRNA-seq vs bulk RNA-seq

- bulk RNA-seq
  - lyse entire tissue → expression vector is average of all cells
  - sort cells by surface marker → expression profile is average of all cells exhibiting that surface marker
- cell sorting can affect cell viability and can thus affect mRNA measurements

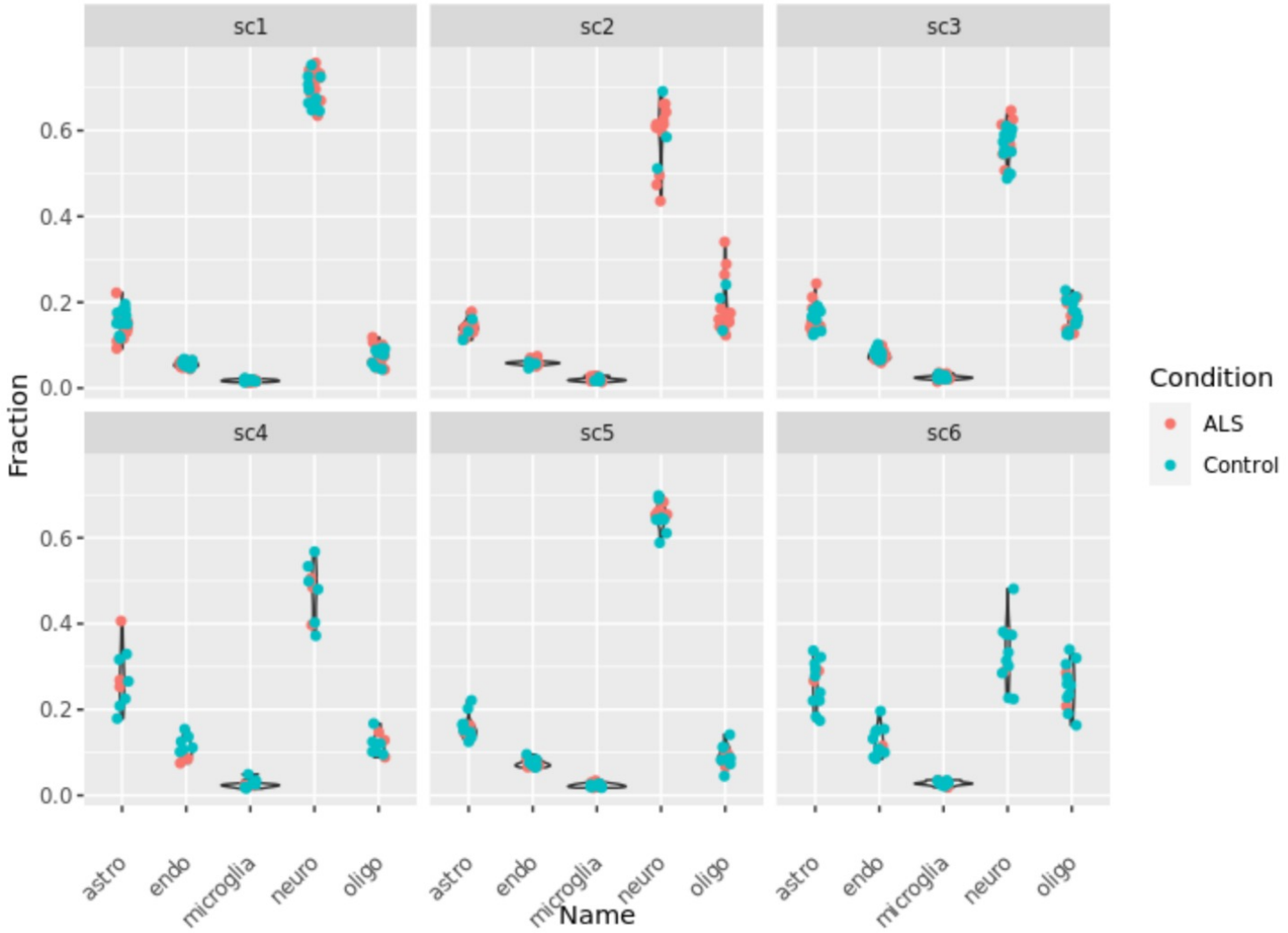


Example: bulk RNA-seq of brain tissue

- Sample clusters do not correlate with experimental factors
- Gene clusters can be associated to
  - cell type markers
  - gender

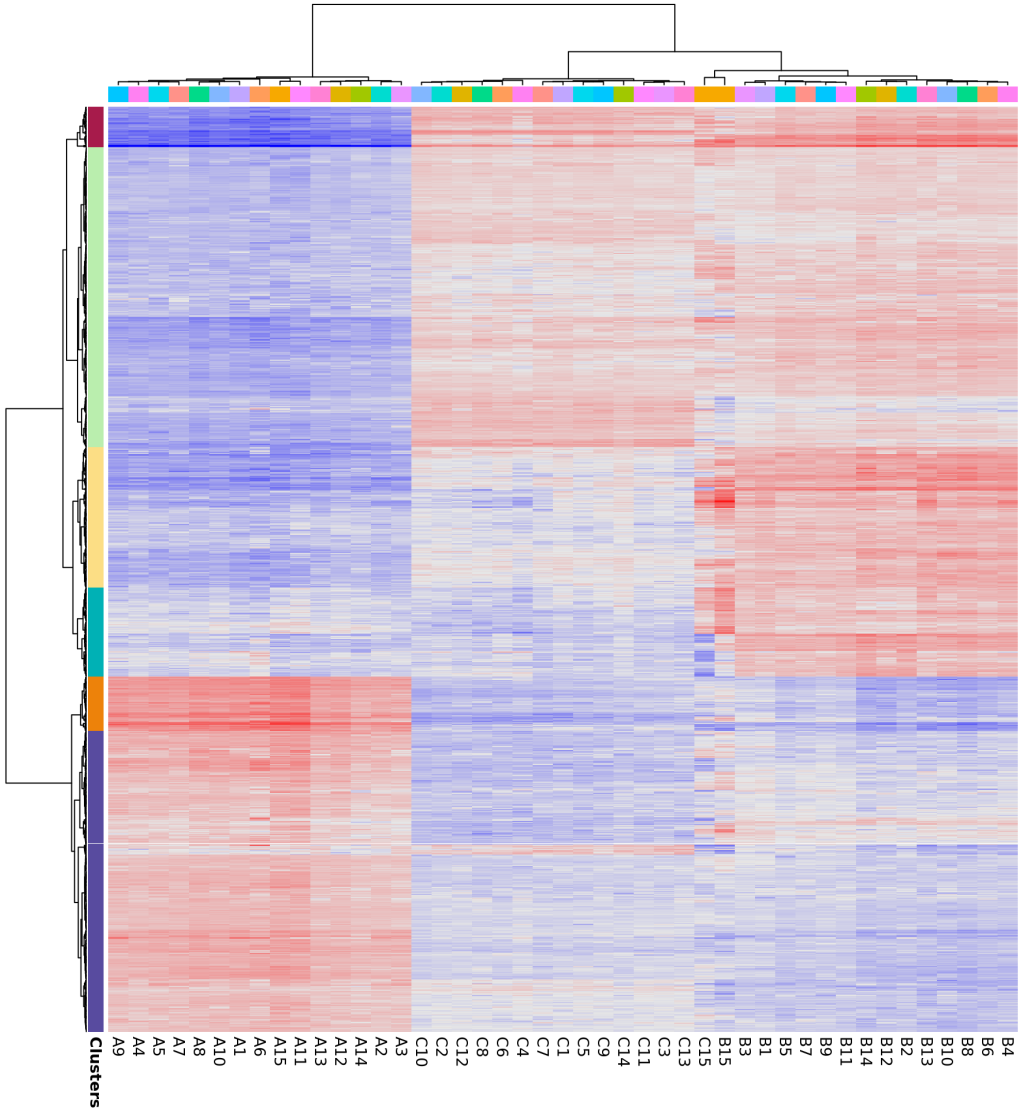


Cell-type deconvolution



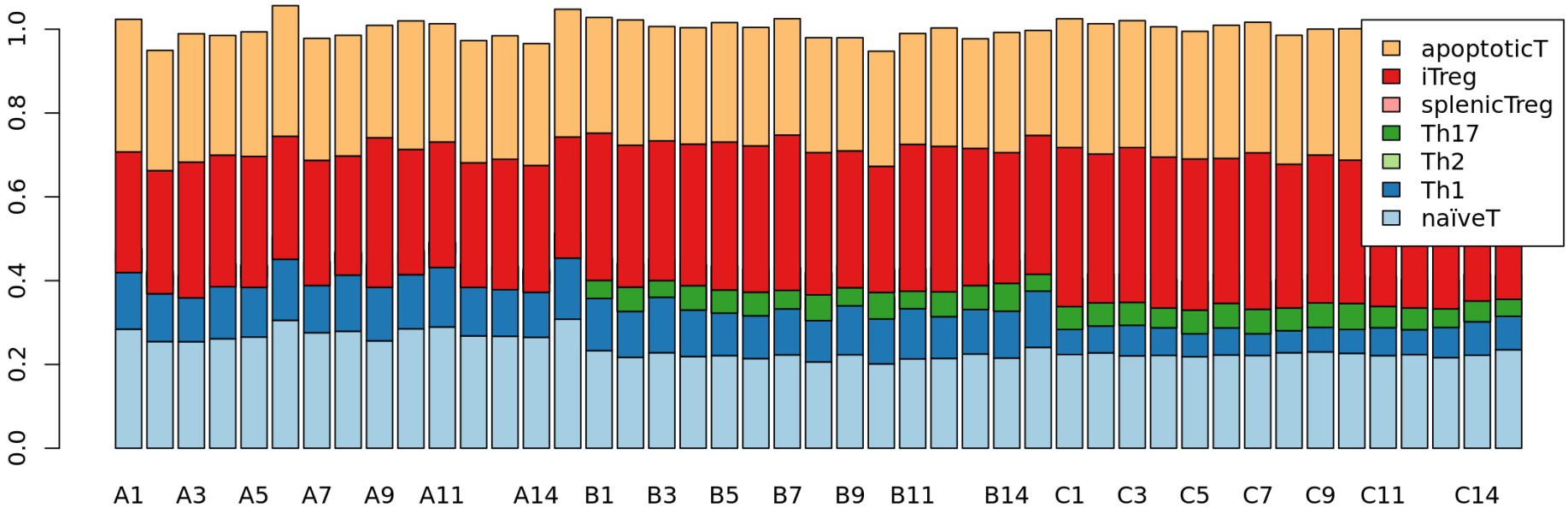
Batch effect: Treated T-cells

- bulk RNA-seq of T-cells treated with different drugs
- Samples cluster according to batch



Batch effect: Treated T-cells

- Batch A samples do not have the helper T cells Th17



Data characteristics

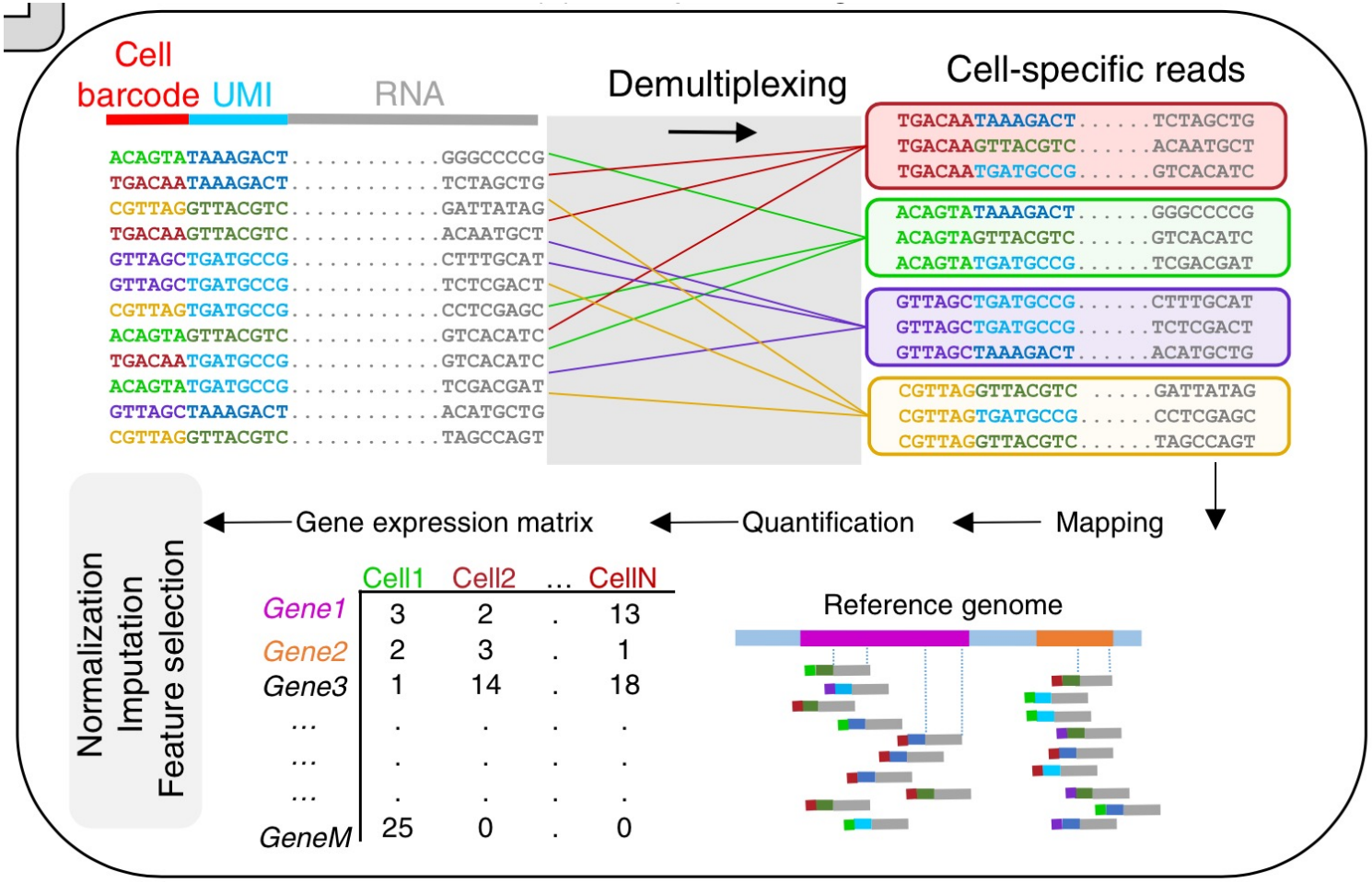
- bulk RNA-seq:
  - 10k to 10Mio cells → billions of mRNA molecules → 100s of billions of fragments after fragmentation and amplification → sequencing samples ~20 Mio reads out of a much larger pool of fragments
  - 10-20k different genes detected in a sample
- single cell RNA-seq
  - 1 cell → 200k mRNA molecules
  - 70-90% of mRNA molecules are typically lost during library prep; precise numbers are unknown
  - library 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

Filtering of cells/barcodes

Reasons for low quality cells

- Empty drops (ambient RNA)
- Sequencing depth
- Low quality bead
- Non-viable or apoptotic cell
- Nucleus only
- Doublet
- …

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

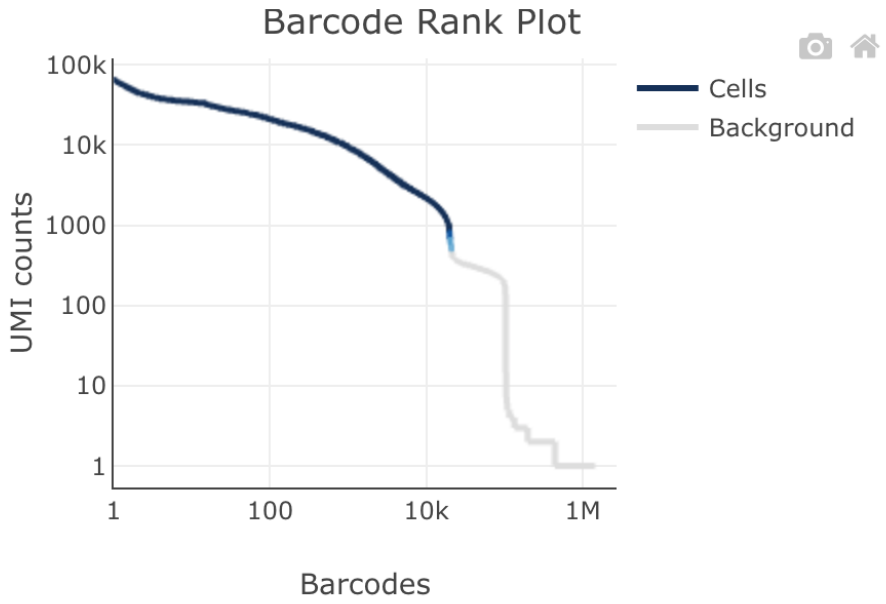
Method | Open Access | Published: 22 March 2019

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



- 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

EmptyDrops algorithm

The algorithm has two key steps:

1. Select x% of the barcodes with the highest total UMI counts → primary mode of cells with high RNA content cells.
2. Select y% of the barcodes with lowest total UMI counts
   → assume they represent empty cells and build ambient RNA model
   → call barcodes that disagree with ambient model as valid cells

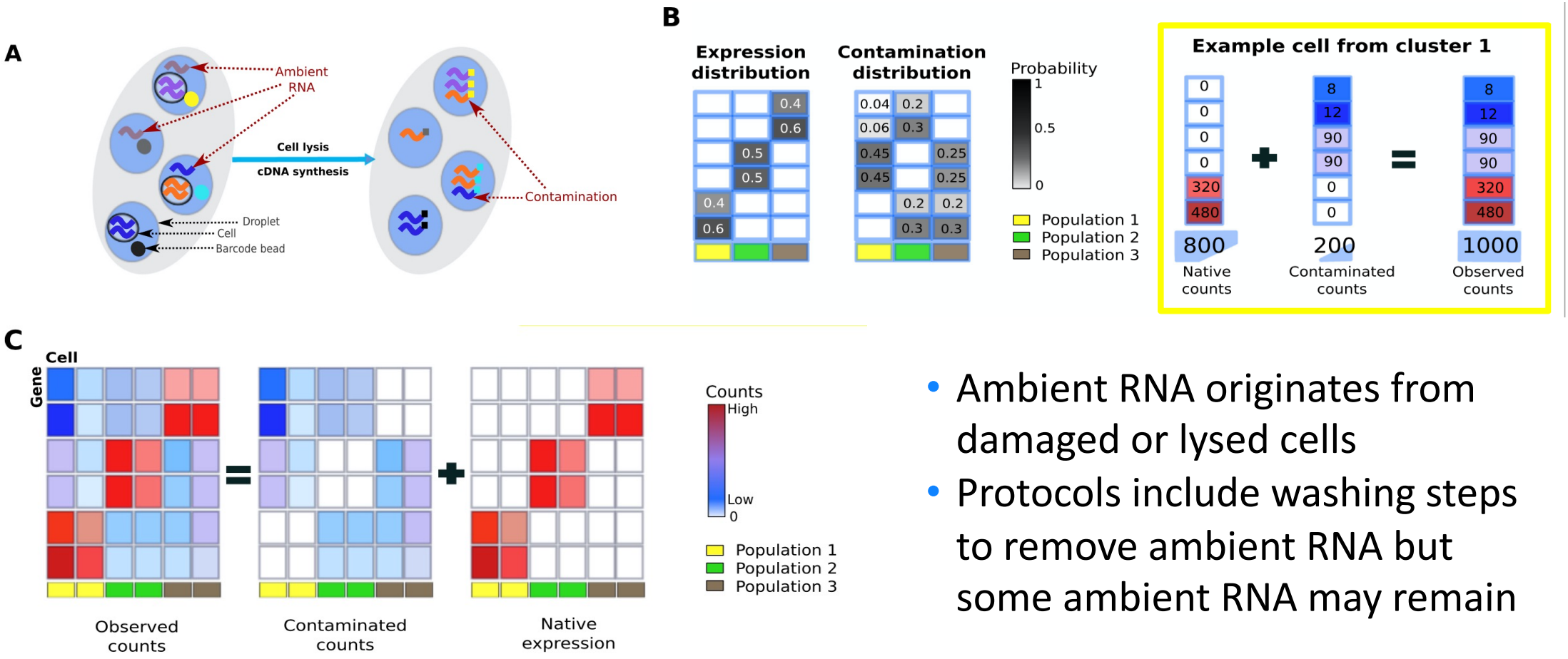
Barcodes are called cells if they

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

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 may be removed using Bayesian approach.

Ambient RNA

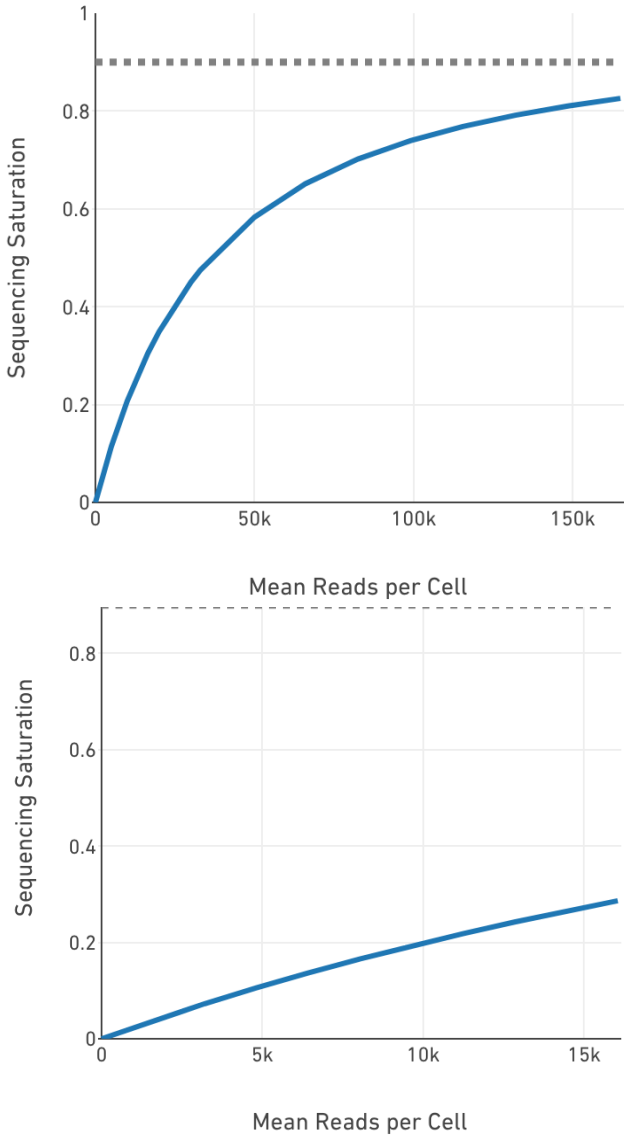


- Ambient RNA originates from damaged or lysed cells
- Protocols include washing steps to remove ambient RNA but some ambient RNA may remain

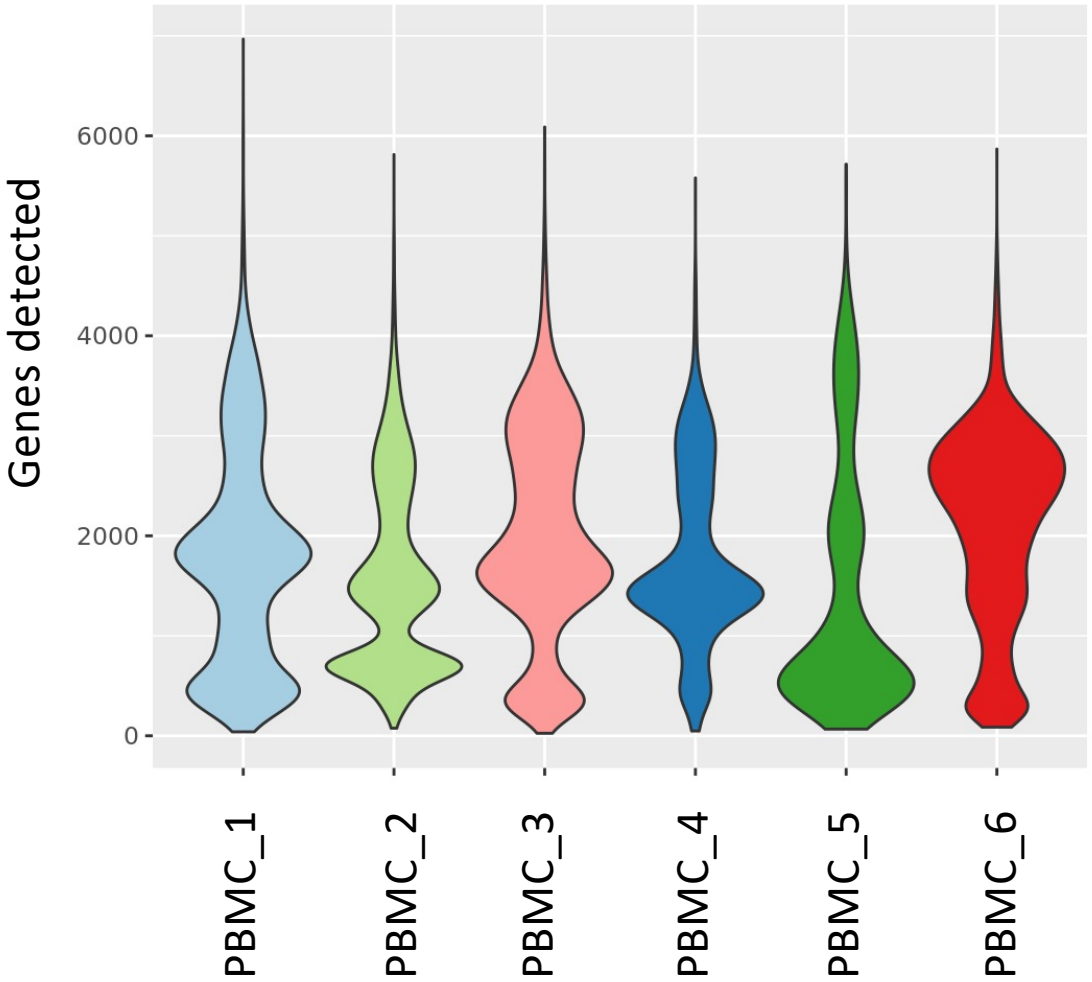
Yang et al., DecontX, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1950-

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

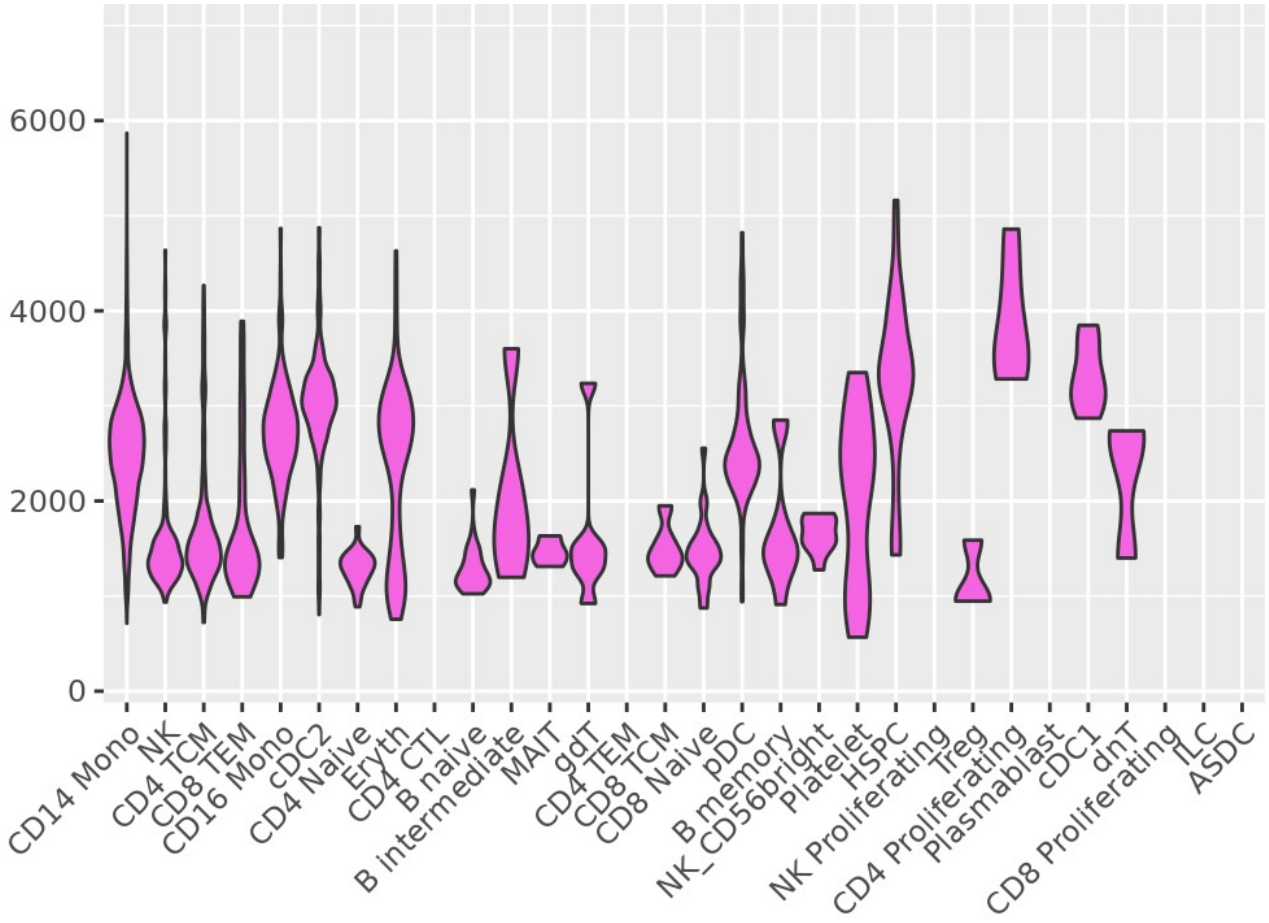


Filtering example: Peripheral Blood Mononuclear Cells (PBMCs)



- 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 detected may be filtered ab

Quality control: Low quality bead

- Remove barcodes with few reads sequenced
- Remove barcodes with few genes detected

- Risk: might remove cells with low transcriptional activity

Quality control: Low quality cell

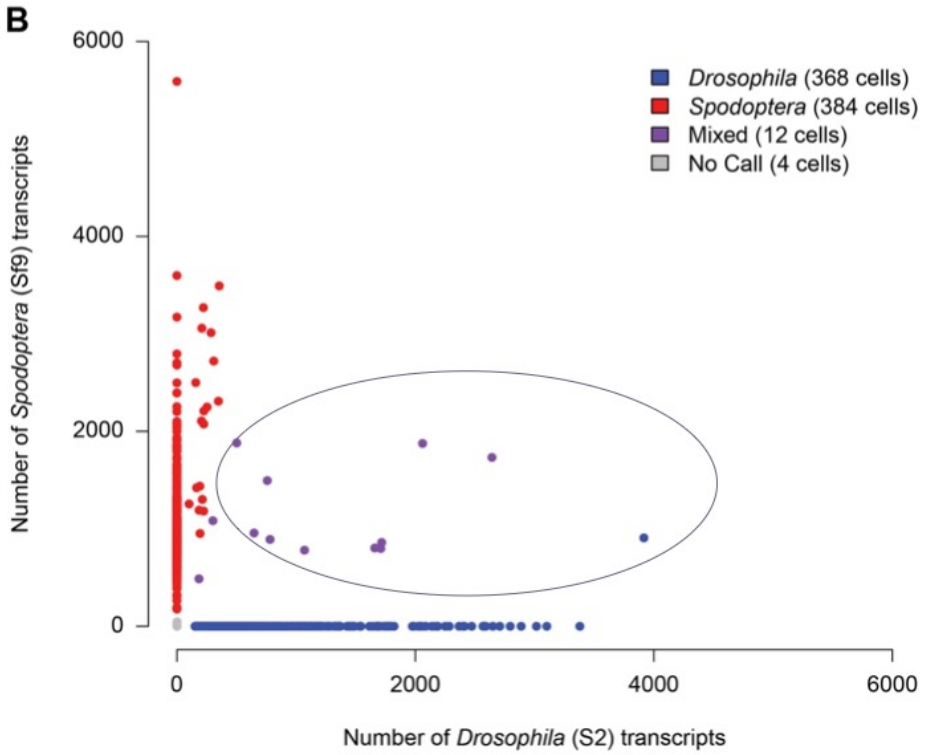
- Remove barcodes with
  - high fraction of mitochondrial genes (in apoptotic or partially lysed cells)
  - zero mitochondrial genes (cells that have lost the cytoplasm?)

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

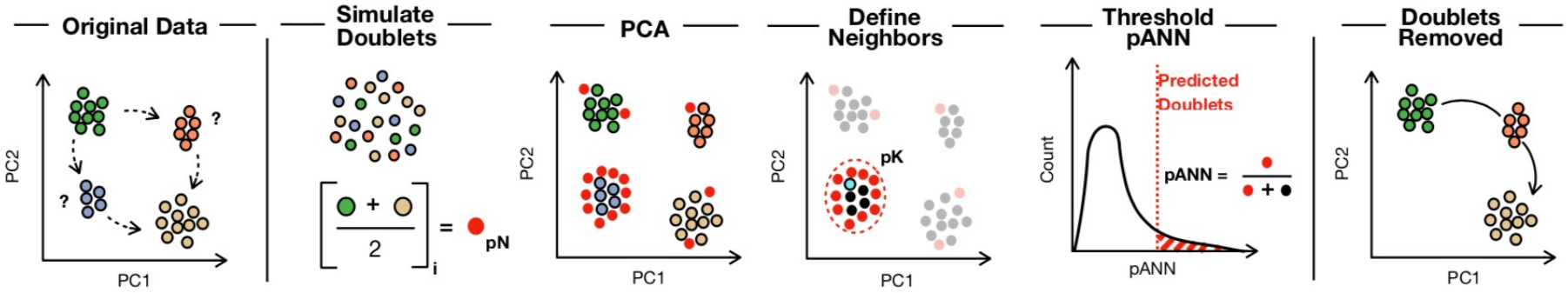
Quality control: Doublets

- Barcode collisions: not enough barcodes
- Technical doublets: two cells in the same droplet (10X specs: +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

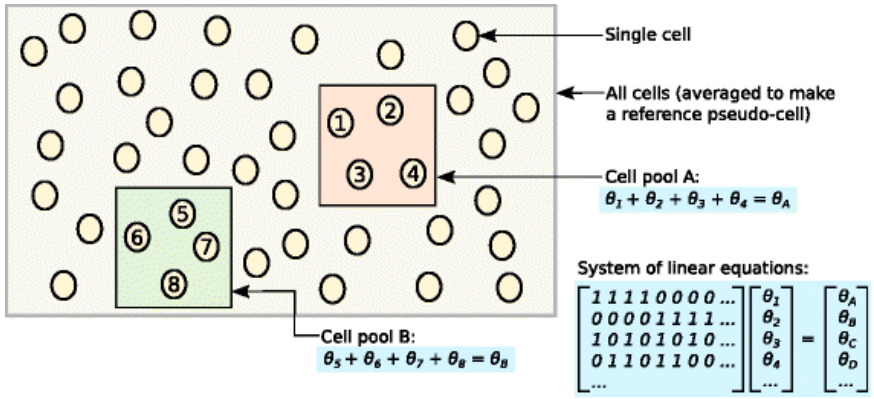
- Quality metrics, like the number of reads or genes detected per cell, do depend on factors like
  - cell type
  - sequencing depth
  - overall quality
  - …
- As a consequence thresholds for these metrics need to be adapted to individual samples and cell types

Normalization

- Simple straightforward normalization:
  - scale by total number of counts per cell
  - transform the result with log(x+1)

- Expression is at Log-scales
- Avoids that zero counts become negative infinite at log scale

scran: Normalization using pools of cells



- Define a pool of cells
- Sum expression values across all cells in the pool
- Normalize the cell pool against an average reference, using the summed expression values
- Repeat this for many different pools of cells to construct a linear system
- Deconvolute the pool-based size factors to their cell-based counterparts

Lun, Genome Biology, 2016, https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7

Normalization: Performance comparison papers

nature methods

Analysis | Published: 26 February 2018

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson & Mark D Robinson

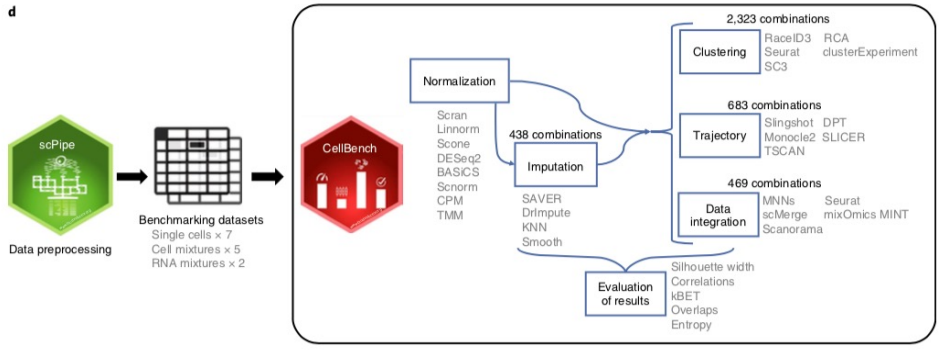
Nature Methods **15**, 255–261(2018) | Cite this article

10k Accesses | **61** Citations | **158** Altmetric | Metrics

nature methods ANALYSIS
https://doi.org/10.1038/s41592-019-0425-8

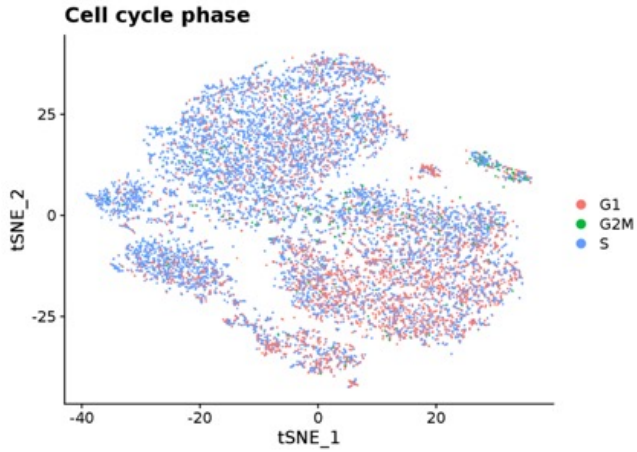
Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments

Luyi Tian[1,2*], Xueyi Dong[1,3], Saskia Freytag[1,4], Kim-Anh Lê Cao[5], Shian Su[1], Abolfazl JalalAbadi[5], Daniela Amann-Zalcenstein[1,2], Tom S. Weber[1,2], Azadeh Seidi[6], Jafar S. Jabbari[6], Shalin H. Naik[1,2] and Matthew E. Ritchie[1,2*]



Normalization: Cell cycle causes unwanted variation

- Requires normalization considering cell cycle as latent variable



Dimensionality Reduction

- Removes redundancy in data
- Makes subsequent analyses more efficient and robust
  - clustering
  - classification
  - cell characterization
- Many features (high dimensionality) make classification more erroneous

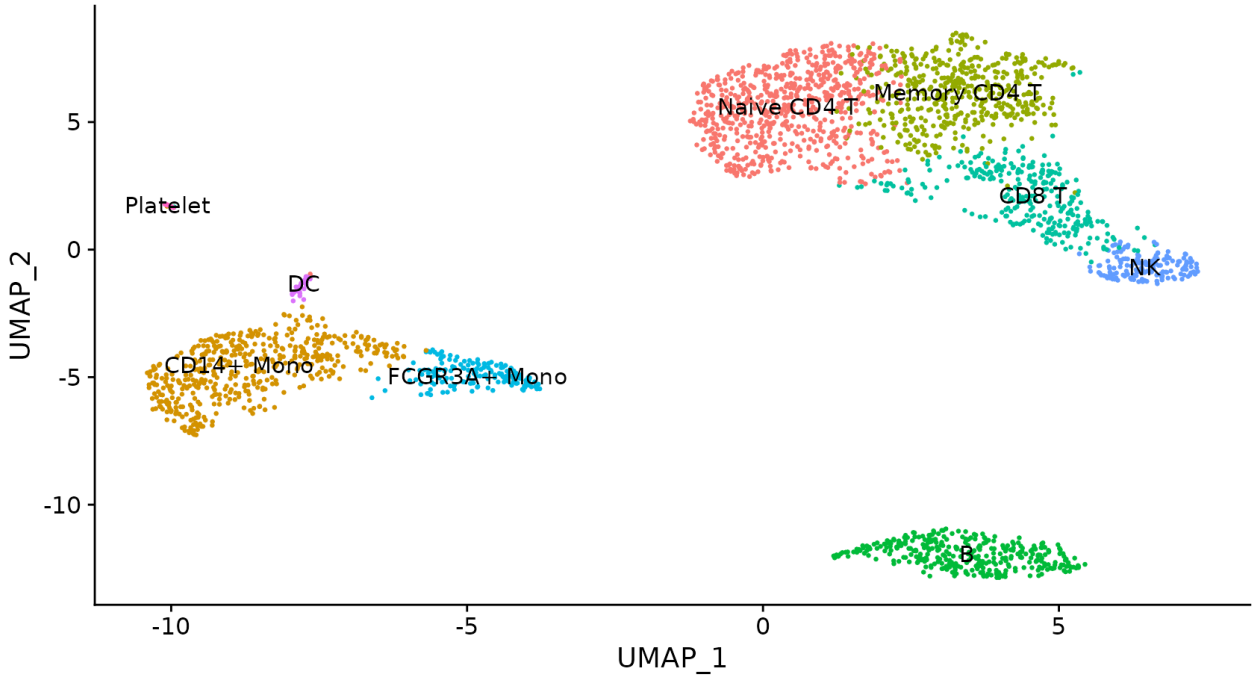
Dimensionality Reduction Methods

- Matrix Factorization
  - PCA
  - MDS – multi-dimensional scaling
  - ….
- Graph-based
  - t-SNE
  - UMAP

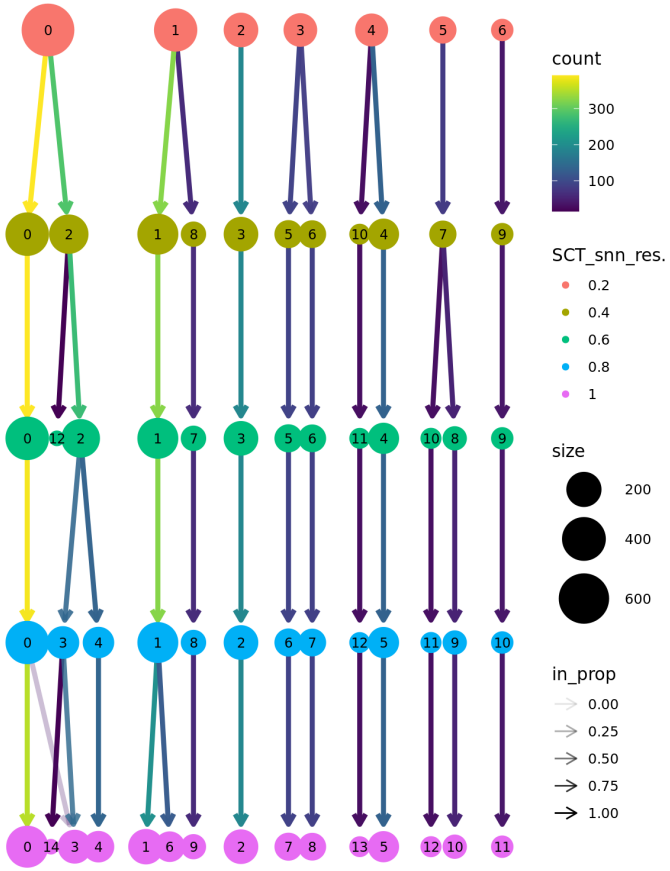
- Autoenconder (based on matrix factorization)

Two-dimensional representation

- cells cluster according to **cell types**
- **cell type** is the major cause for variation in sc data



Cluster resolution should match cell types of interest



- resolution of clustering controls cell type granularity
- Often there is no perfect clustering
- If in doubt, overclustering is preferred
  - it is easy to handle a cell type split in two clusters
  - it is cumbersome to have a cluster that contains two cell types

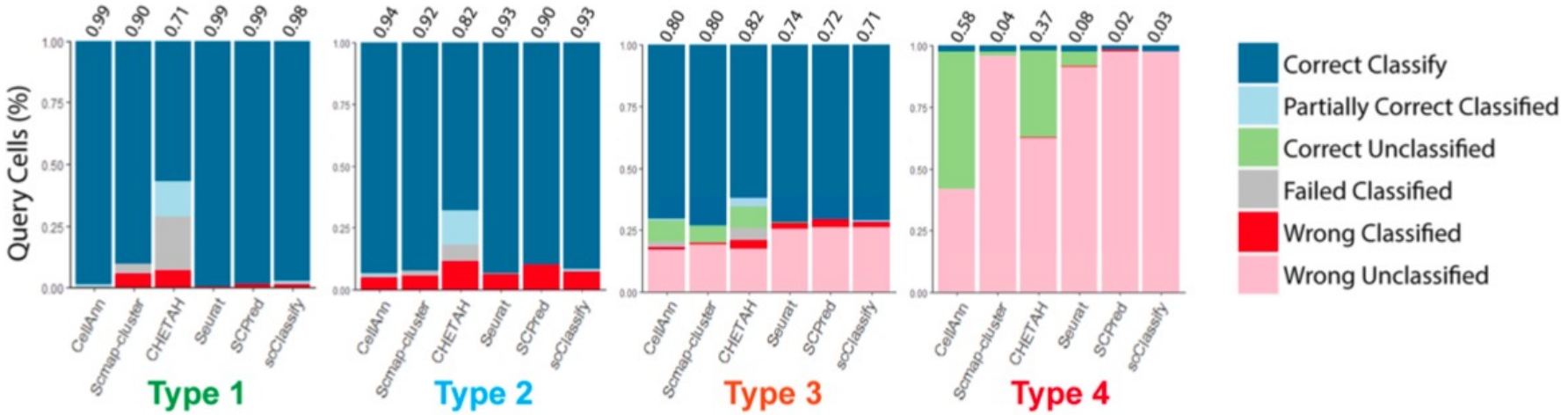
Cell type annotation tools

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

Per-cell vs per-cluster labeling

- generally they should agree for the major cell-types
- discrepancies occur at sub-types
- often only a small percentage of cells (<10%) has single-cell labels that diverge from the cluster label
- 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

Algorithms:

- **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

- https://atlas.brain-map.org/
- https://www.humancellatlas.org/
- https://tabula-sapiens-portal.ds.czbiohub.org/
- https://www.flycellatlas.org/
- https://azimuth.hubmapconsortium.org/references/

- Sigle Cell Expression Atlas: https://www.ebi.ac.uk/gxa/sc/home
  - individual studies
- Chan Zuckerberg: https://cellxgene.cziscience.com/collections
- BROAD:
  - https://singlecell.broadinstitute.org/single_cell
- ...

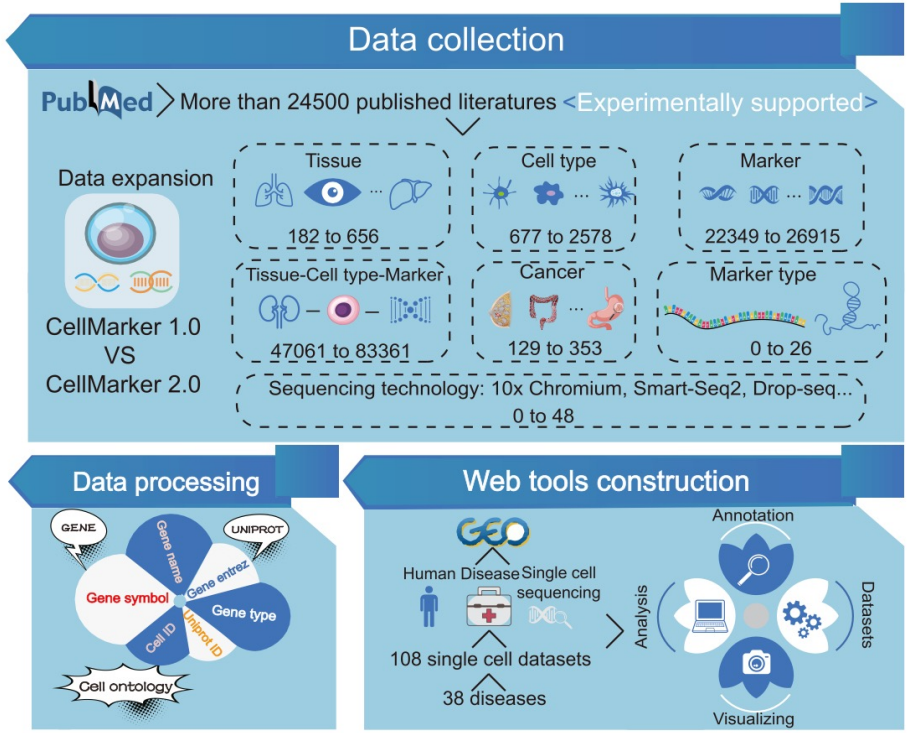
Generally: The more similar to your system, the better

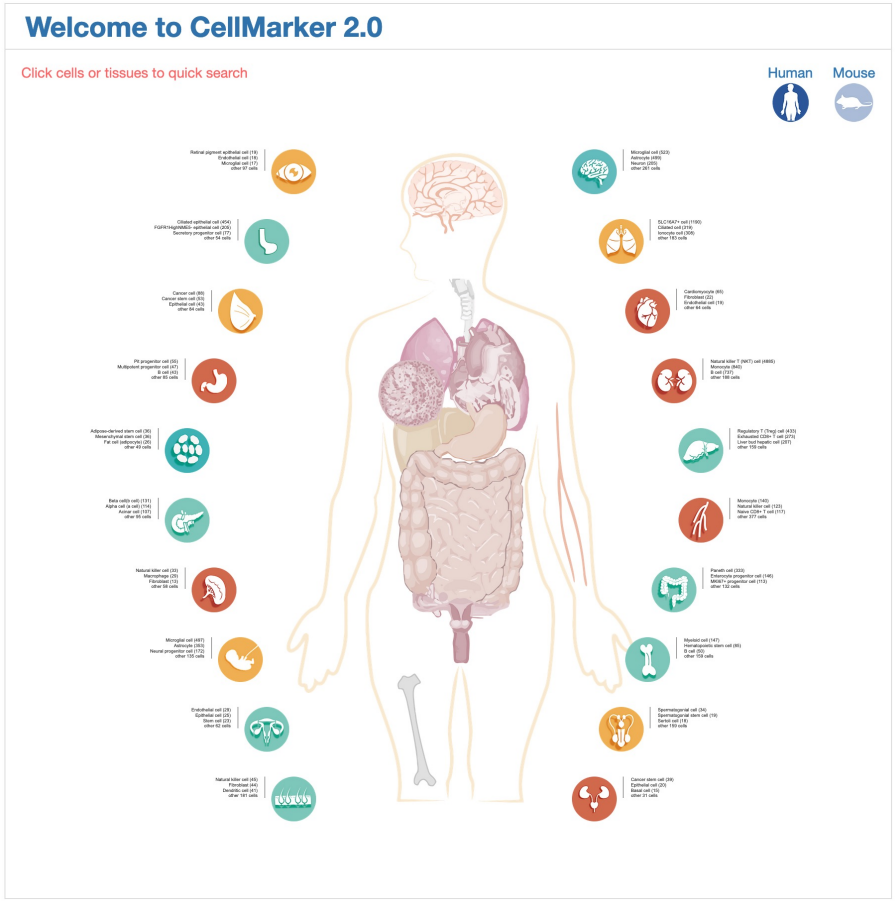
Celltype Marker-Based Annoation

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

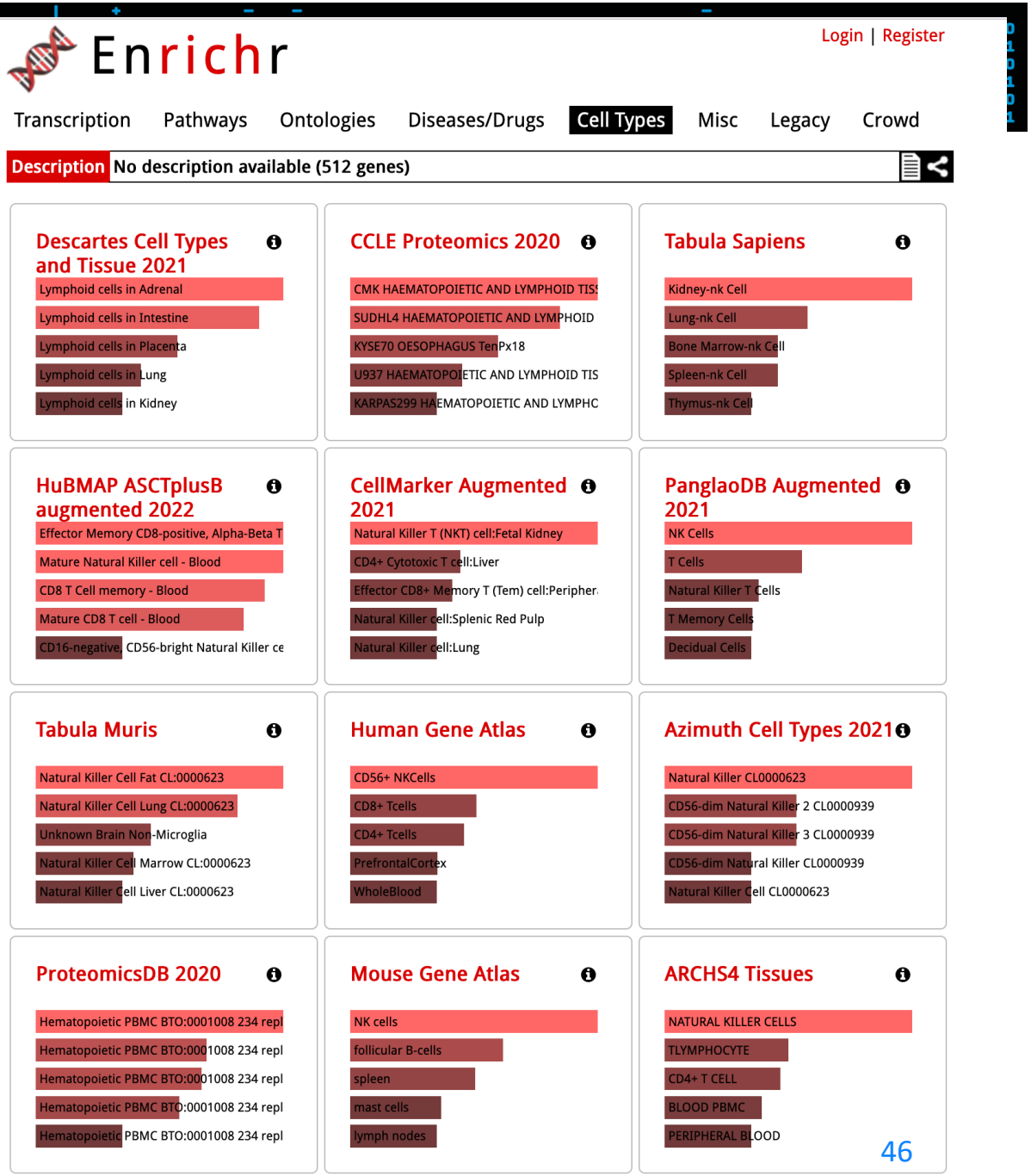




EnrichR

- 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



Azimuth

- Next to reference scRNA-seq data, Azimuth hosts also corresponding cell type markers

▾ celltype.l2 (default)

Label	Expanded Label	OBO Ontology ID	Markers
B intermediate	Intermediate B cell	mature B cell	MS4A1, TNFRSF13B, IGHM, IGHD, AIM2, CD79A, LINC01857, RALGPS2, BANK1, CD79B
B memory	Memory B cell	memory B cell	MS4A1, COCH, AIM2, BANK1, SSPN, CD79A, TEX9, RALGPS2, TNFRSF13C, LINC01781
B naive	Naive B cell	naive B cell	IGHM, IGHD, CD79A, IL4R, MS4A1, CXCR4, BTG1, TCL1A, CD79B, YBX3

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