Single cell RNA-seq analysis

BBS230B - July 15th, 2023

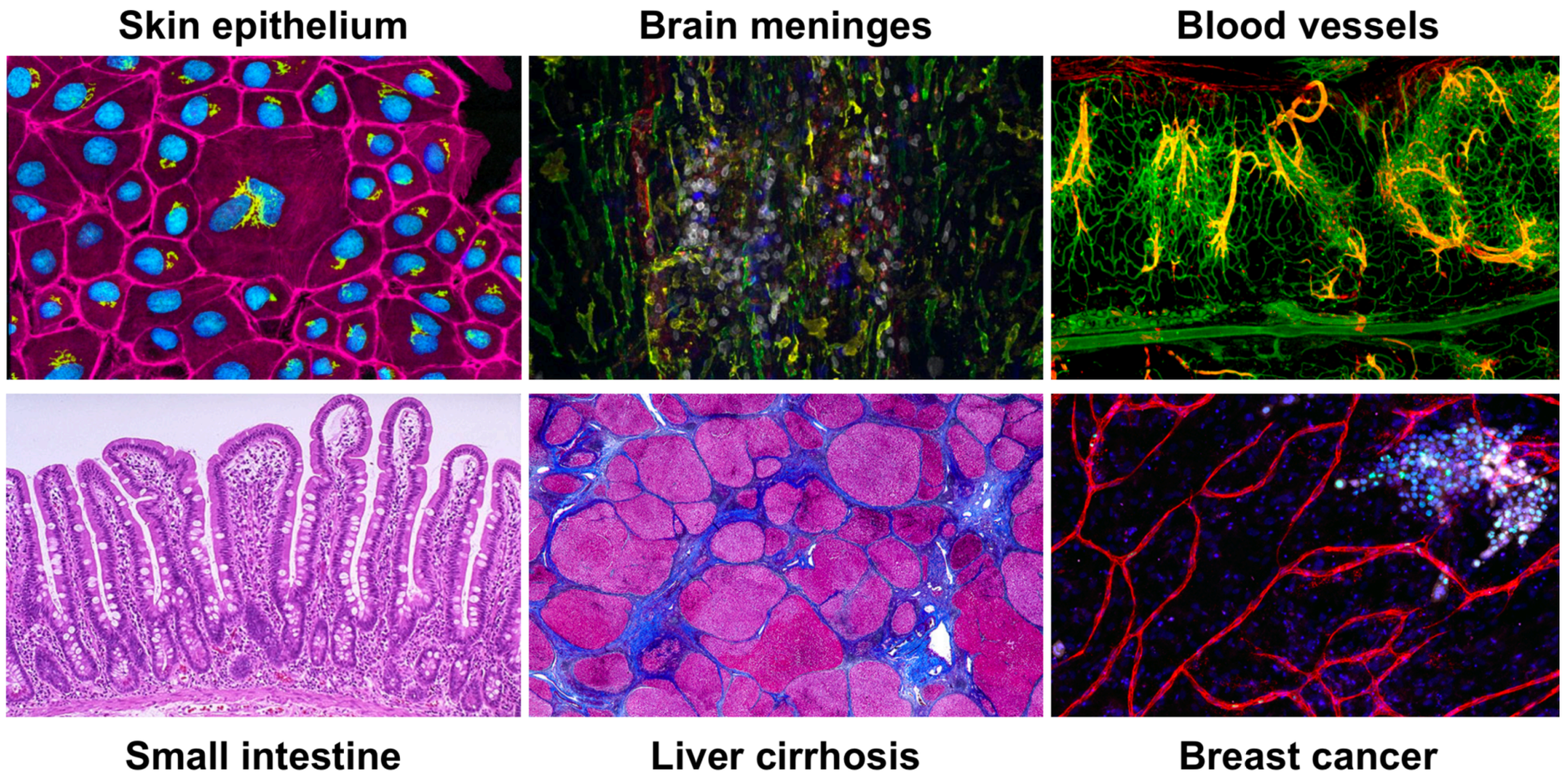
Radhika S. Khetani, PhD

rkhetani@hsph.harvard.edu

https://bioinformatics.sph.harvard.edu/

Why single-cell RNA-seq?

Single-cell RNA-seq (scRNA-seq) allows us to evaluate the transcriptome at the level of individual cells. This offers a glimpse into the incredible diversity of cell types, states, and interactions.



http://www.cell.com/pictureshow/skin | https://library.med.utah.edu/WebPath/webpath.html

Why single-cell RNA-seq?

- To explore which cell types are present in a tissue
- To identify unknown/rare cell types or states
- To elucidate the changes in gene expression during differentiation processes or across time or states
- To identify genes that are differentially expressed in particular cell types between conditions (e.g. treatment or disease)
- To explore changes in expression among a cell type while incorporating spatial, regulatory, and/or protein information

Single-cell RNA-seq platforms

In wells
(FACS-assisted)

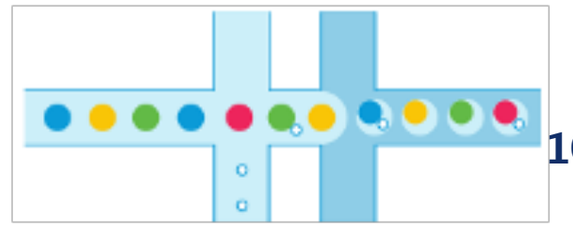
CEL-Seq
MARS-Seq
SMART-Seq
SCRB-Seq
Quartz-Seq

In nano/microwells

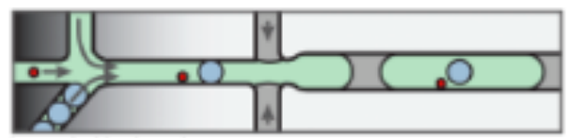
Seq-Well

BD Rhapsody

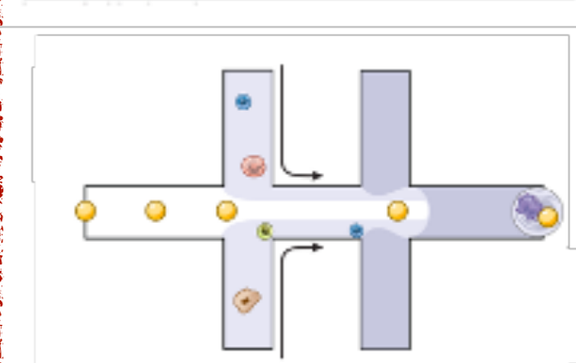
In droplets



Chromium
10x Genomics

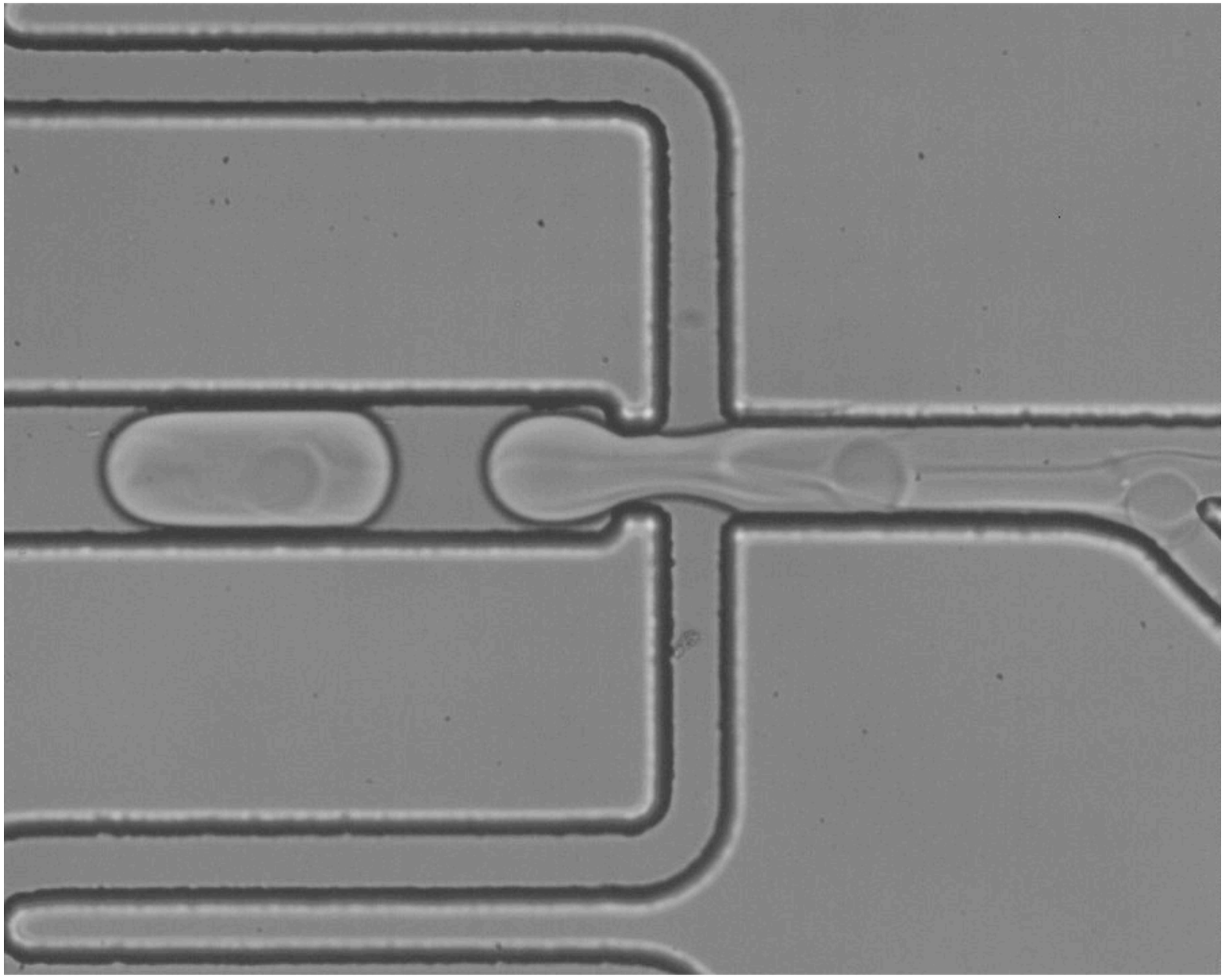


InDrops



DropSeq

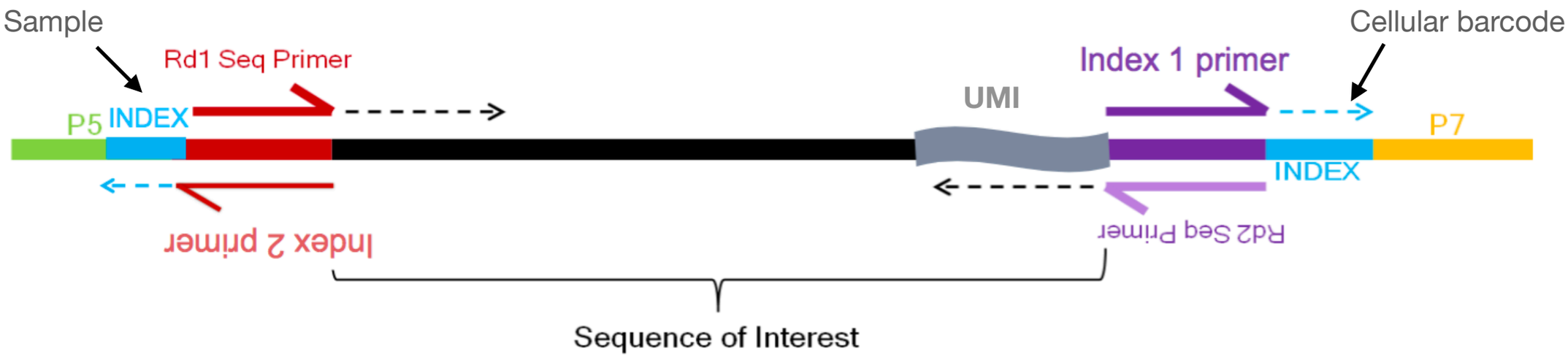
Slide adapted from “Introduction to Single Cell RNA-sequencing: a practical guideline”, Mandovi Chatterjee, Ph. D.



Video generated by
Single Cell Core @ HMS

Slide taken from “Introduction to Single Cell RNA-sequencing: a practical guideline”, Mandovi Chatterjee, Ph. D.

Components of a scRNA-sequencing read



- **Sample index:** determines which sample the read originated from (red bottom arrow)
  - Added during library preparation - needs to be documented

- **Cellular barcode:** determines which cell the read originated from (purple top arrow)
  - Each library preparation method has a stock of cellular barcodes used during the library preparation

- **Unique molecular identifier (UMI):** determines which transcript molecule the read originated from
  - The UMI will be used to collapse PCR duplicates (purple bottom arrow)

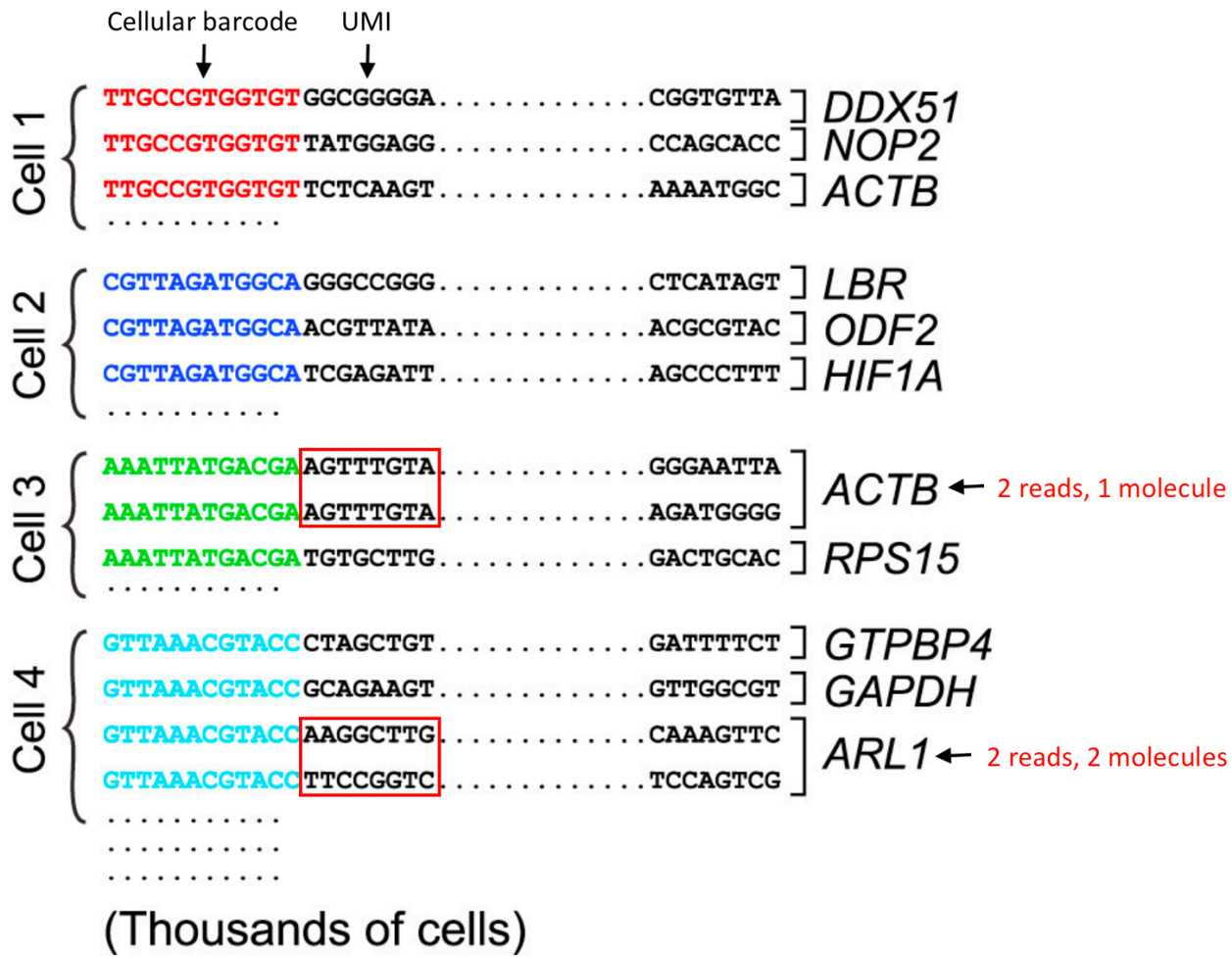
- **Sequencing read1:** the Read1 sequence (red top arrow)
- **Sequencing read2:** the Read2 sequence (purple bottom arrow)

Image credit: *Single Cell Sequencing Core at HMS*

Understanding UMIs

- Reads with **different UMIs** mapping to the same transcript were derived from **different molecules** and are biological duplicates - each read should be counted.

- Reads with the **same UMI** originated from the **same molecule** and are technical duplicates - the UMIs should be collapsed to be counted as a single read.



Bulk v.s. Single-cell RNA-seq

- Bulk RNA-seq provides an overview of average differences in gene expression. For certain scenarios this has proven to be sufficient (i.e. biomarkers for cancer).

- Bulk RNA-seq is useful if you are not expecting or not concerned about cellular heterogeneity

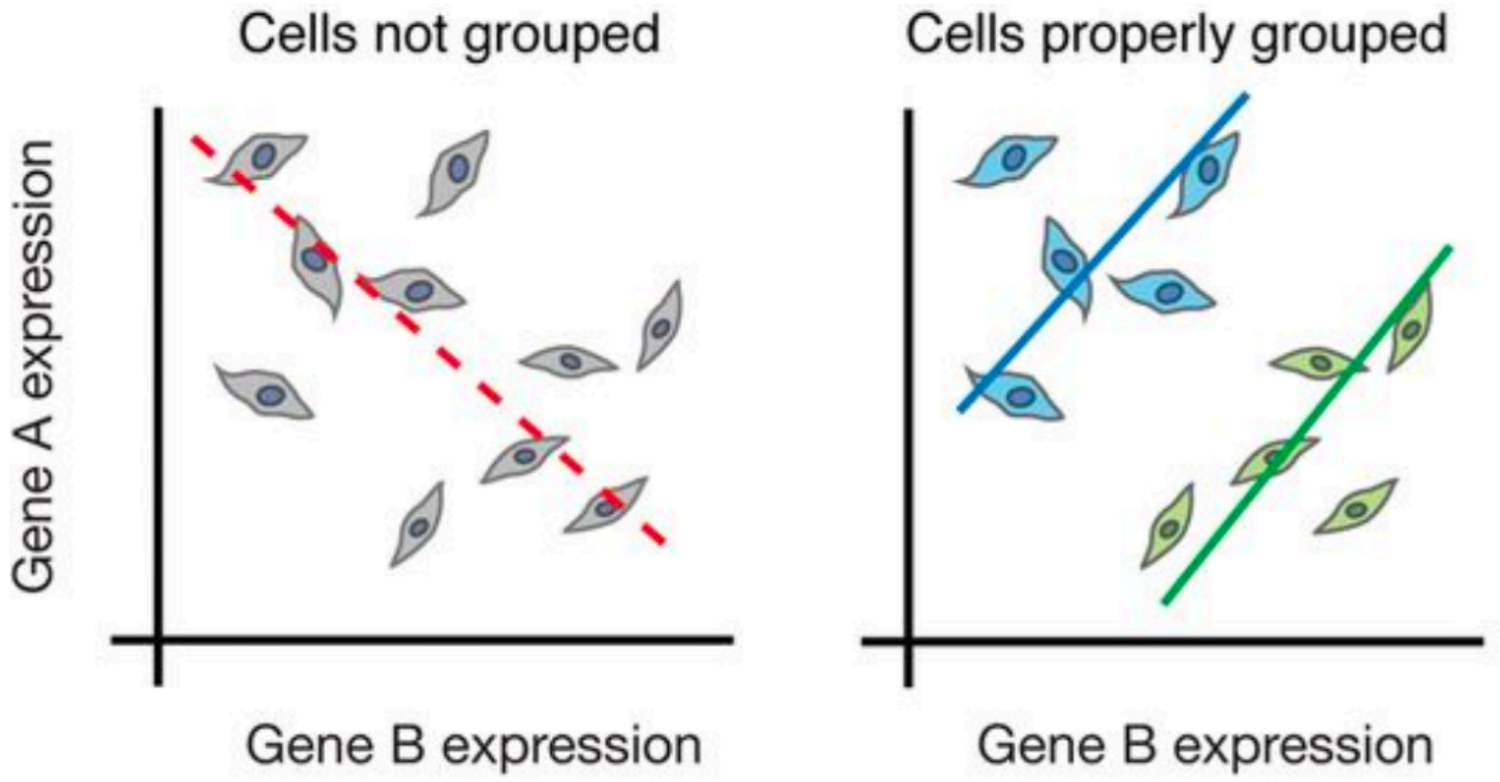


Image credit: *Trapnell, C. Defining cell types and states with single-cell genomics, Genome Research 2015 (doi: https://dx.doi.org/10.1101/gr.190595.115)*

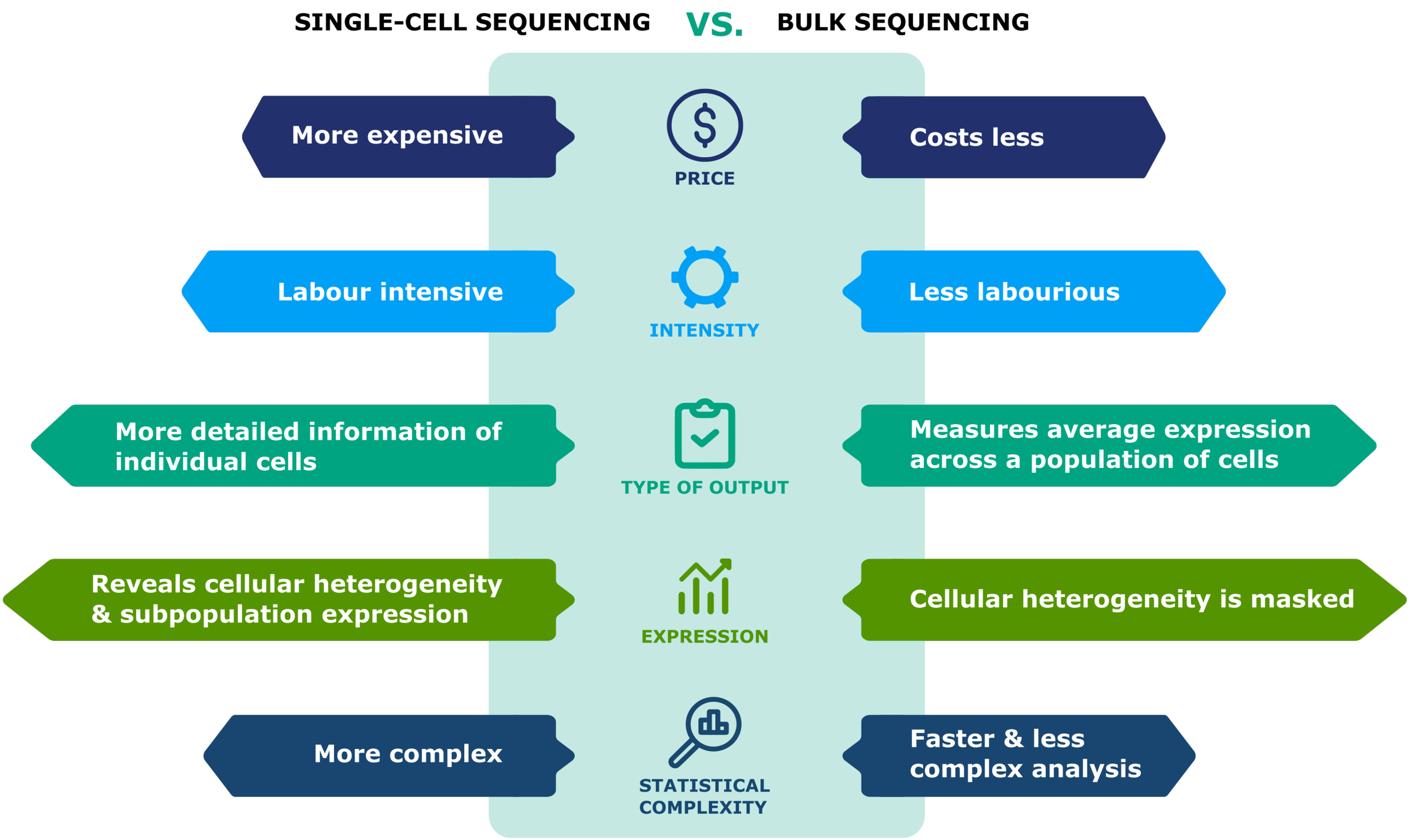


Image source: https://lizard.bio/blog/single-cell-vs-bulk/

Challenges with scRNA-seq data

- Large volume of data
- Low depth of sequencing per cell (zero-inflation)
  - Often detecting only 10-50% of the transcriptome per cell
- Biological variability across cells/samples can obscure the cell type identities
- Technical variability across cells/samples

Increased complexity and richer datasets, means more room for misinterpretations and deriving wrong conclusions!

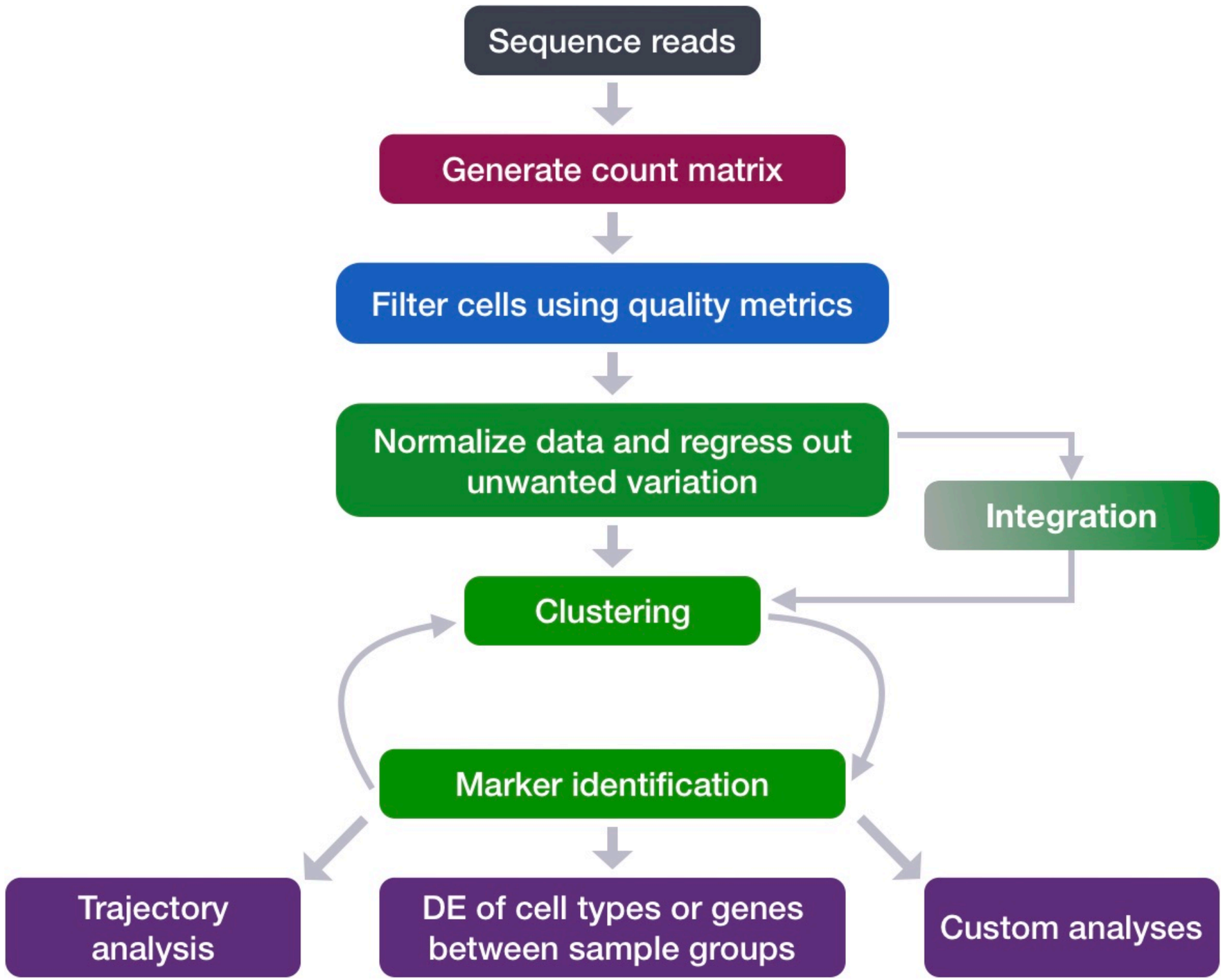
Recommendations

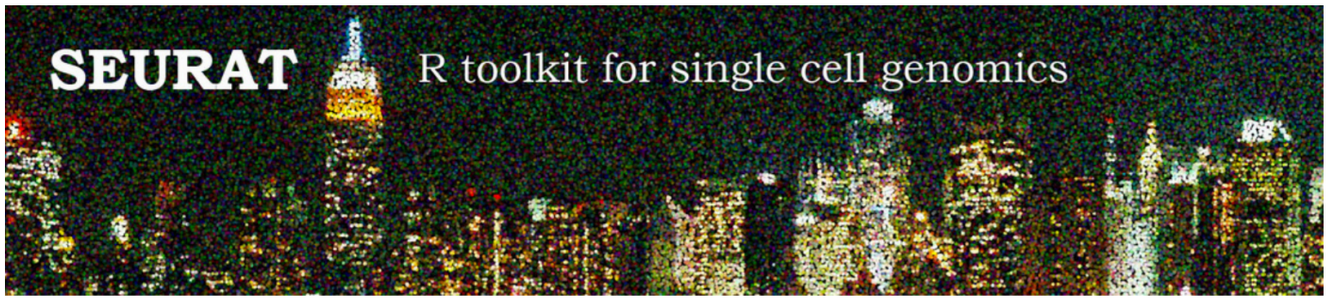
- Do not perform single-cell RNA-seq unless it is necessary for the experimental question of interest.
  - Could you answer the question using bulk sequencing, which is simpler and less costly? Perhaps FACS sorting the samples could allow for bulk analysis?

- Understand the details of the experimental question you wish to address.
  - Library preparation method and analysis workflow can vary based on the specific experiment and tissue

- Avoid technical sources of variability, if possible:
  - Discuss experimental design with experts prior to the initiation of the experiment
  - Isolate RNA from samples at same time
  - Prepare libraries at same time or alternate sample groups to avoid batch confounding

Single-cell RNA-seq analysis workflow





Official release of Seurat 4.0

We are excited to release Seurat v4.0! This update brings the following new features and functionality:

- **Integrative multimodal analysis.** The ability to make simultaneous measurements of multiple data types from the same cell, known as multimodal analysis, represents a new and exciting frontier for single-cell genomics. In Seurat v4, we introduce weighted nearest neighbor (WNN) analysis, an unsupervised strategy to learn the information content of each modality in each cell, and to define cellular state based on a weighted combination of both modalities. In our new paper, we generate a CITE-seq dataset featuring paired measurements of the transcriptome and 228 surface proteins, and leverage WNN to define a multimodal reference of human PBMC. You can use WNN to analyze multimodal data from a variety of technologies, including CITE-seq, ASAP-seq, 10X Genomics ATAC + RNA, and SHARE-seq.
  - Paper: Integrated analysis of multimodal single-cell data
  - Vignette: Multimodal clustering of a human bone marrow CITE-seq dataset
  - Portal: Click here
  - Dataset: Download here
- **Rapid mapping of query datasets to references.** We introduce Azimuth, a workflow to leverage high-quality reference datasets to rapidly map new scRNA-seq datasets (queries). For example, you can map any scRNA-seq dataset of human PBMC onto our reference, automating the process of visualization, clustering annotation, and differential expression. Azimuth can be run within Seurat, or using a standalone web application that requires no installation or programming experience.
  - Vignette: Mapping scRNA-seq queries onto reference datasets
  - Web app: Automated mapping, visualization, and annotation of scRNA-seq datasets from human PBMC

Links

Download from CRAN at
https://cloud.r-project.org/package=Seurat

Browse source code at
https://github.com/satijalab/seurat/

Report a bug at
https://github.com/satijalab/seurat/issues

License

Full license

MIT + file LICENSE

Community

Code of conduct

Citation

Citing Seurat

Developers

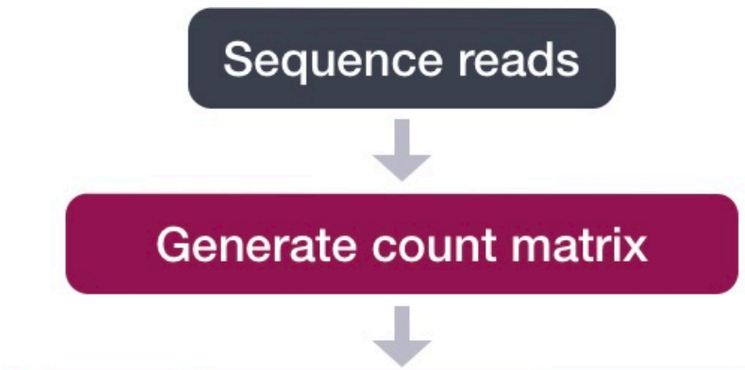
Paul Hoffman
Author, maintainer

Satija Lab and Collaborators
Funder

All authors...

https://satijalab.org/seurat/

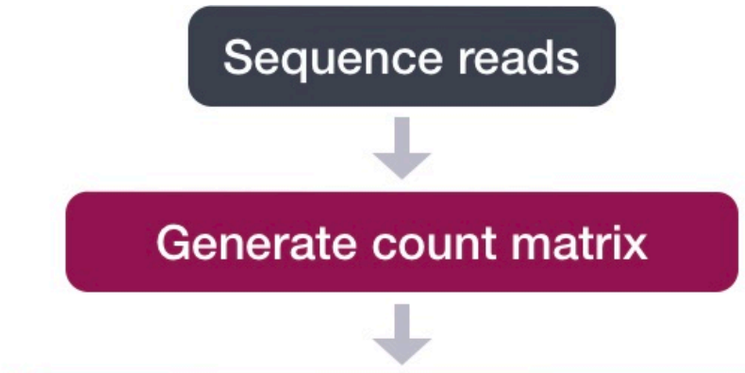
From FASTQ to counts



Tools for this part of the workflow include Alevin, UMI-tools, and Cell Ranger (10X data). While each tool will do things slightly differently the steps below are common to all:

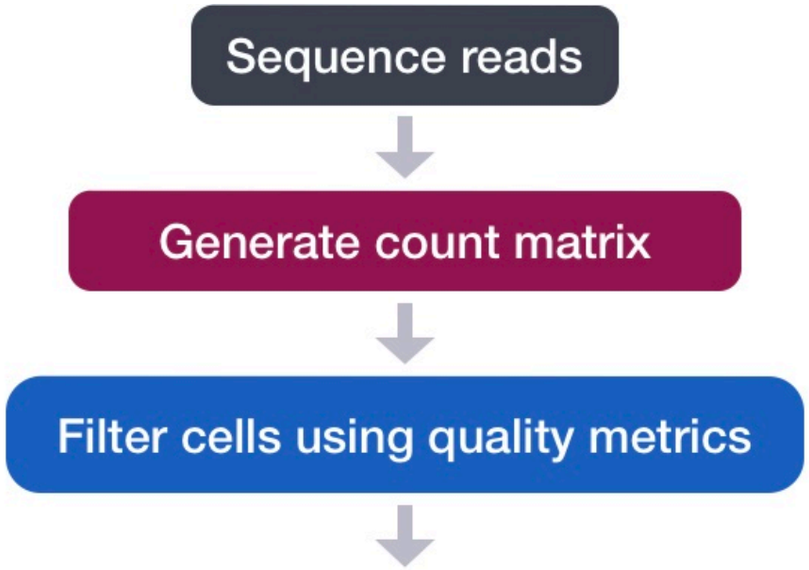
1. Formatting reads and filtering noisy cellular barcodes
2. Demultiplexing the samples
3. Mapping/pseudo-mapping to transcriptome
4. Collapsing UMIs and quantification of reads

From FASTQ to counts



	Cell1	Cell2	…	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
…	.	.	.	.
…	.	.	.	.
…	.	.	.	.
GeneM	25	0	.	0

Quality control of count matrix



Quality control of count matrix

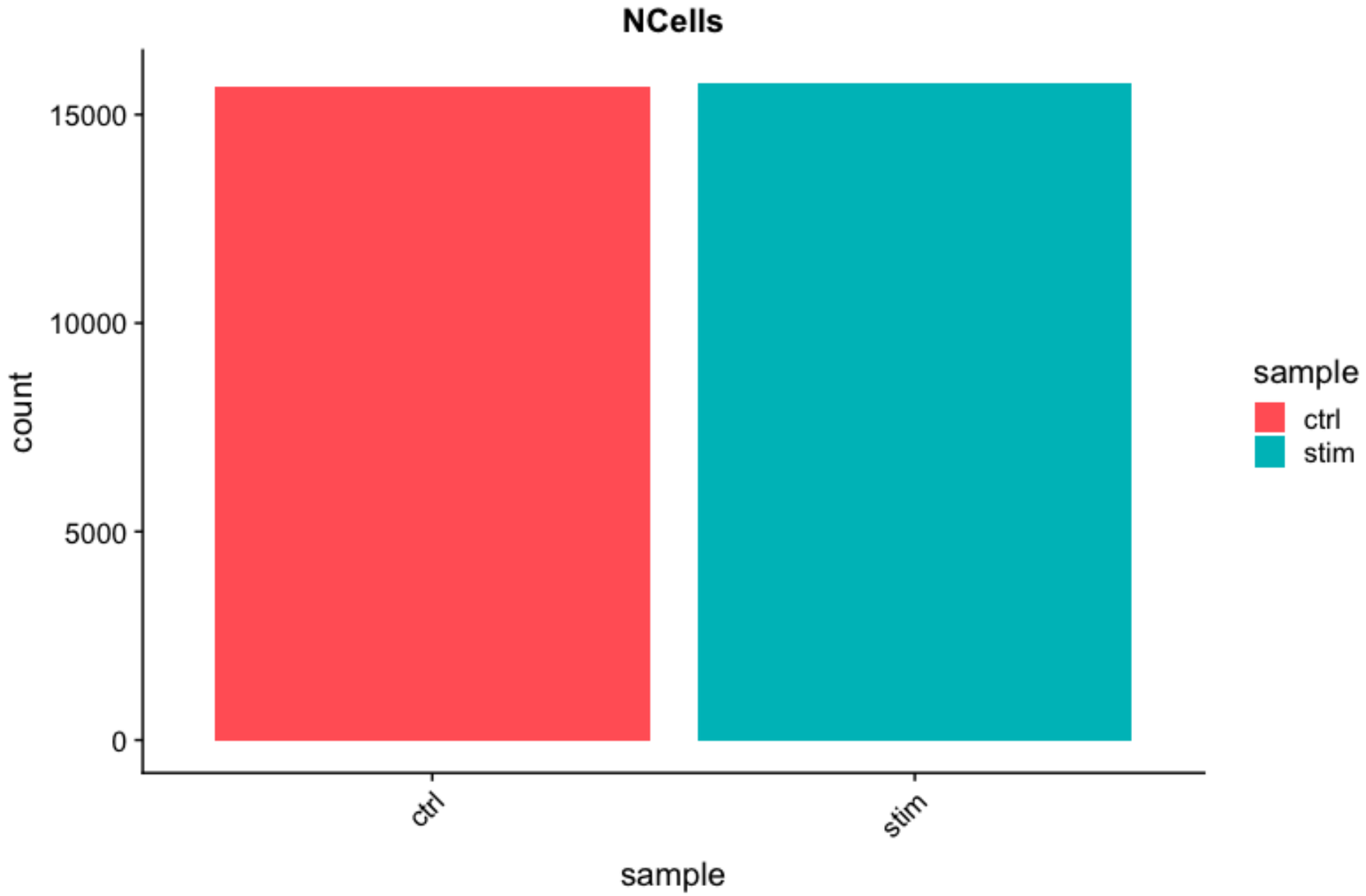
Goals:

- To **filter the data to only include true cells that are of high quality**, so that when we cluster our cells it is easier to identify distinct cell type populations
- To **identify any failed samples** and either try to salvage the data or remove from analysis, in addition to, trying to understand why the sample failed

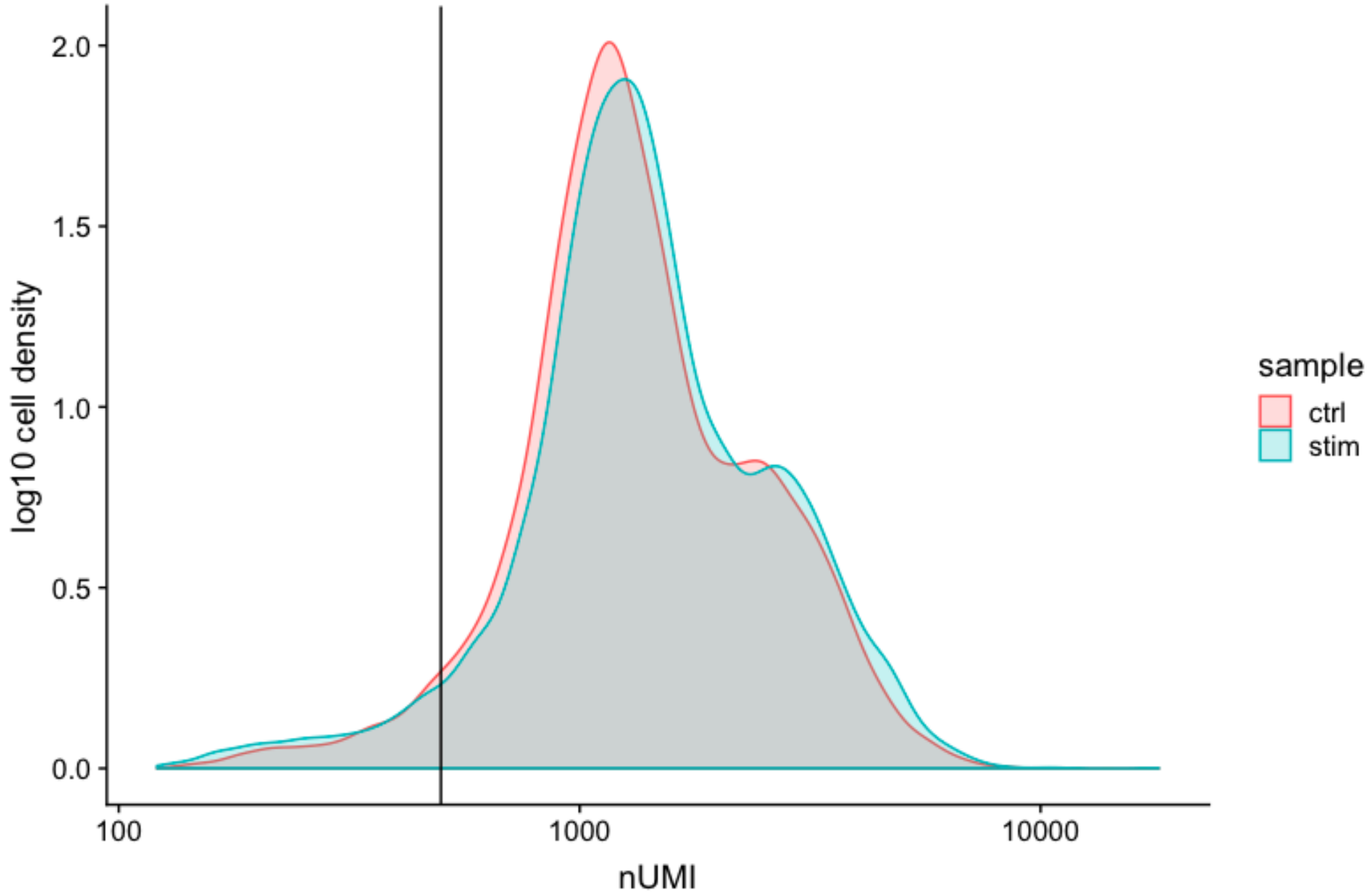
Challenges:

- Delineating cells that are poor quality from less complex cells
- Choosing appropriate thresholds for filtering, so as to keep high quality cells without removing biologically relevant cell types

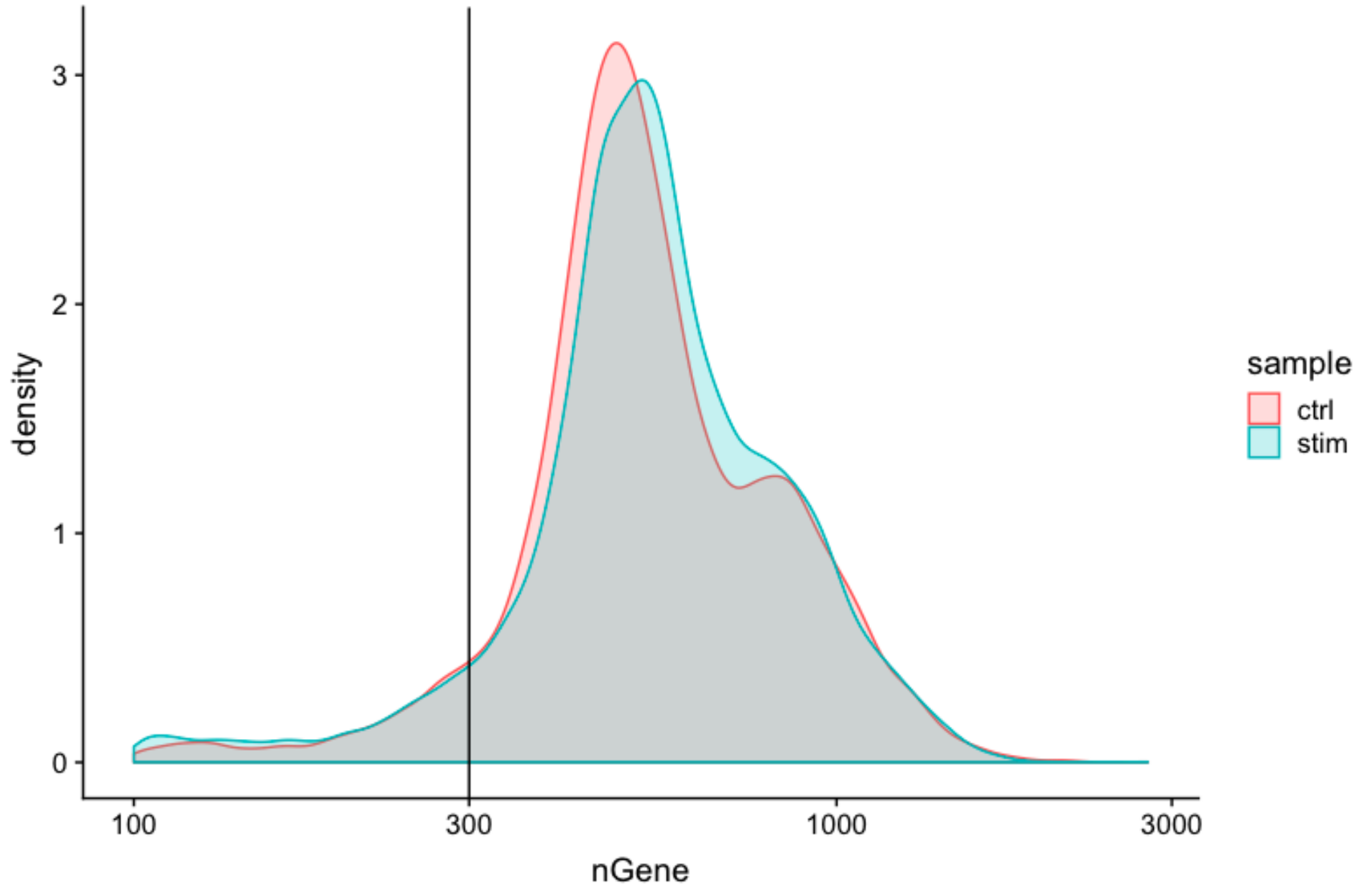
Quality metrics:
Cell counts per sample



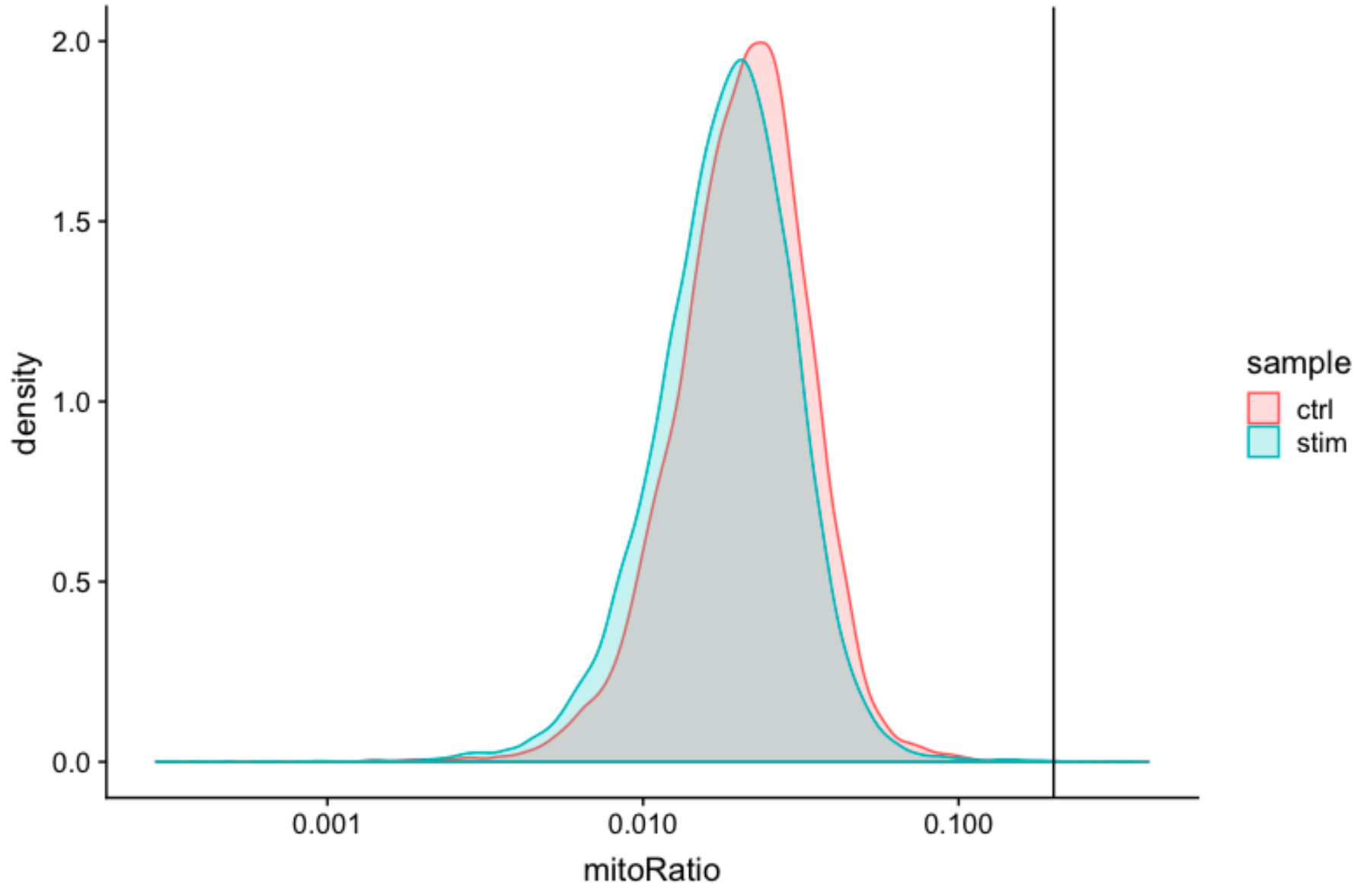
Quality metrics:
UMI counts per cell



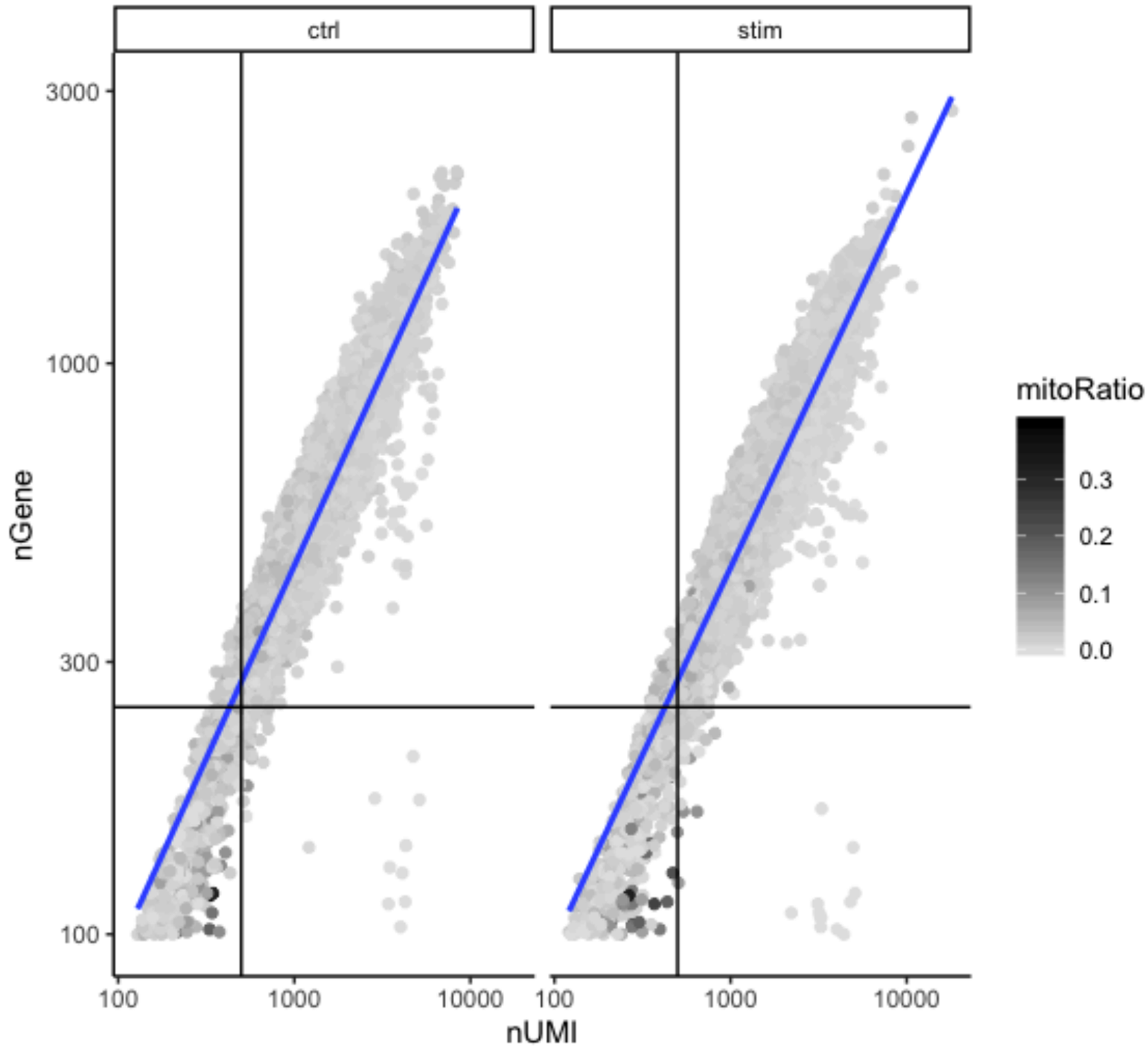
Quality metrics:
Genes detected per cell



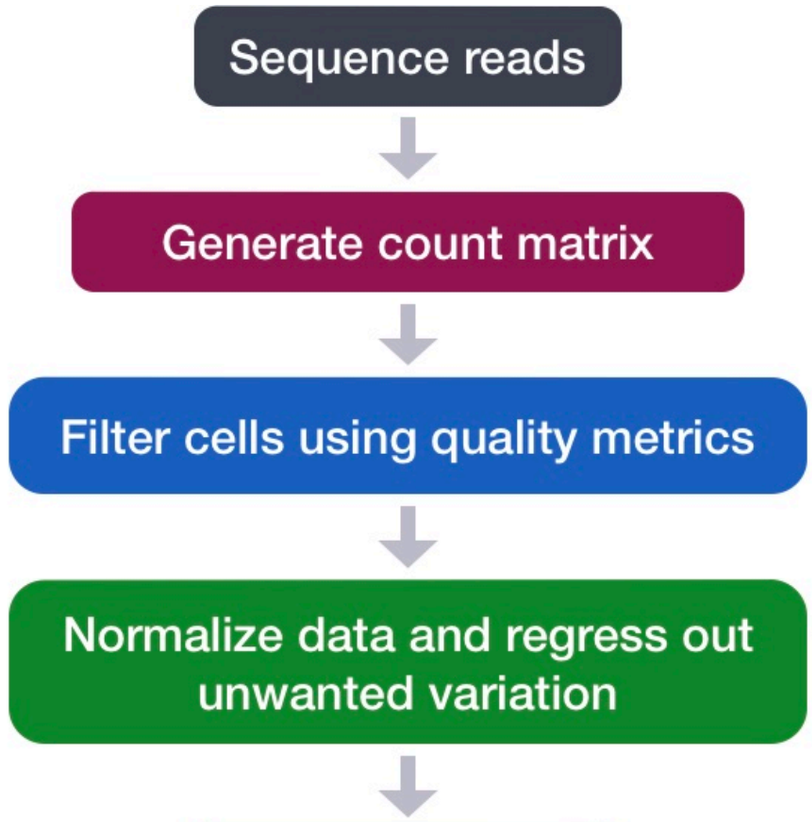
Quality metrics:
Mitochondrial counts ratio



Quality metrics:
Joint filtering effects



Normalizing and removing unwanted variation



Normalizing and removing unwanted variation

Goals:

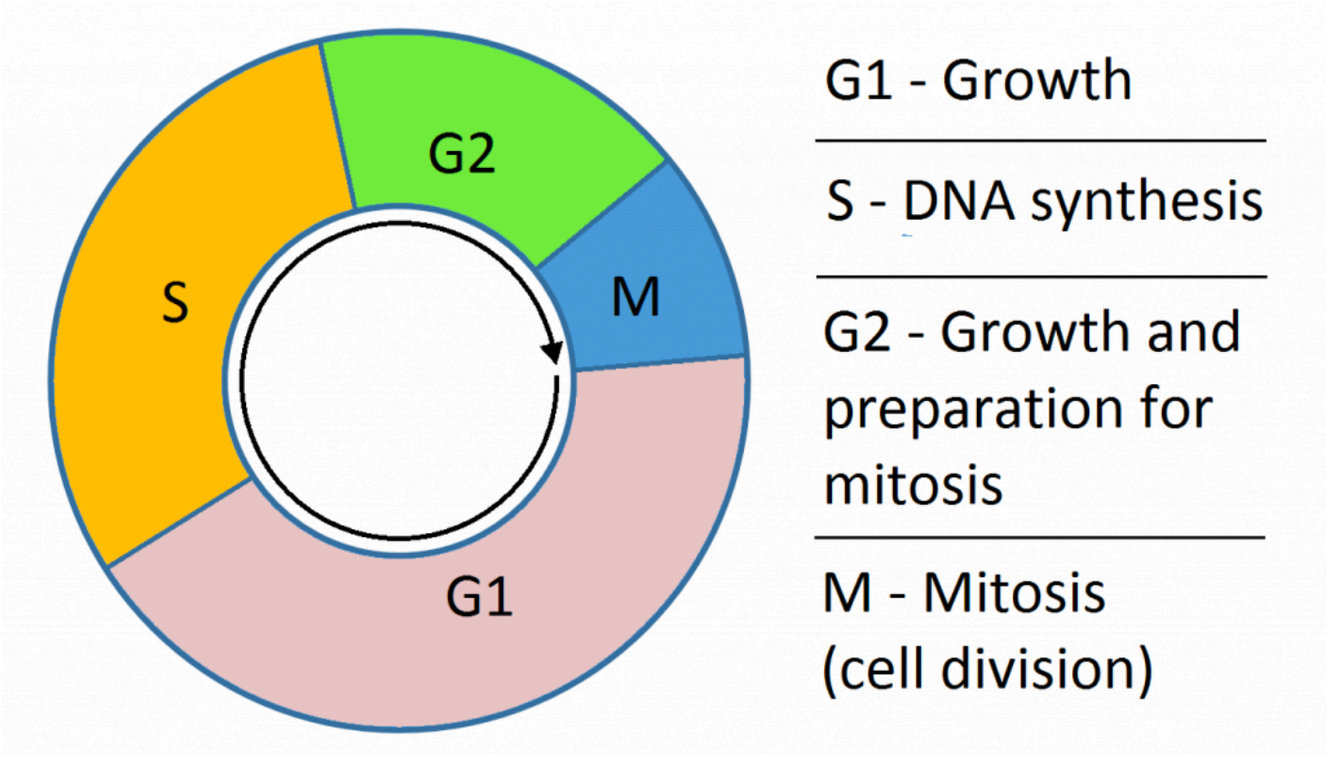
- To identify sources of **unwanted variation** in the data
- To accurately **normalize and scale the gene expression values** to account for differences in sequencing depth and overdispersed count values.

Challenges:

- Determining whether or not there really is an effect from unwanted sources to be concerned about.
- Correcting for covariates while being mindful not to remove true biological signal.

Exploring sources of unwanted variation:

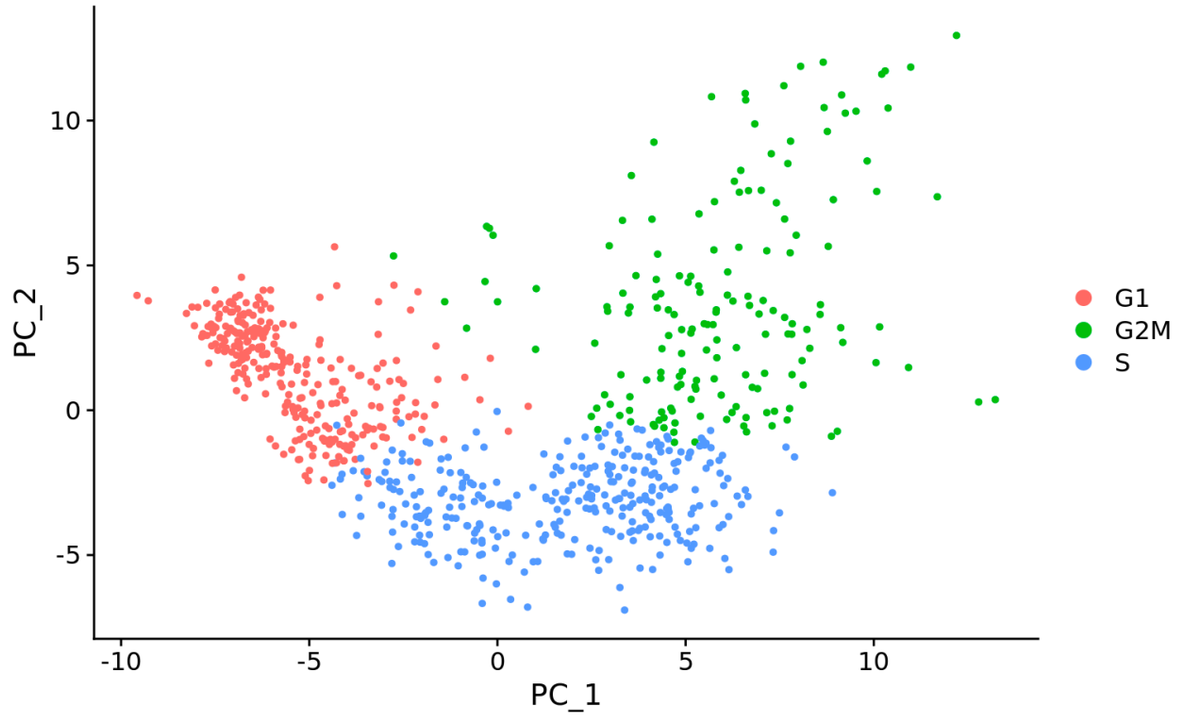
Cell cycle effects



- Cell cycle heterogeneity can drive the changes in expression, and subsequently the way in which cells cluster together.
- For each cell, compute cell cycle phase scores based on the expression of canonical markers.

Exploring sources of unwanted variation:

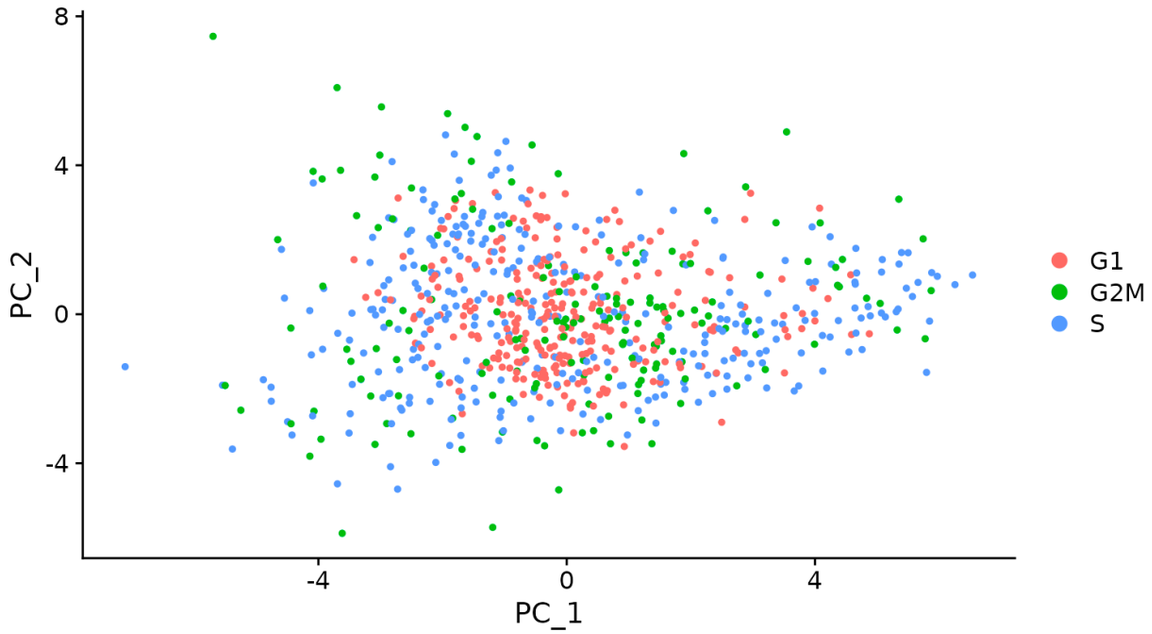
Cell cycle effects



- Scores are used to classify cells into one of three phases
- Plot PCA and color cells by phase
- Do we see separation of cells by cell cycle phase?

Image source: https://satijalab.org/seurat/articles/cell_cycle_vignette.html

Cell cycle effects regressed out

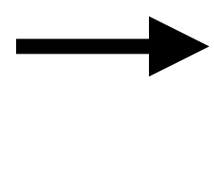


Now, cell-cycle heterogeneity does not contribute to PCA or downstream analysis.

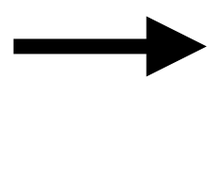
Image source: https://satijalab.org/seurat/articles/cell_cycle_vignette.html

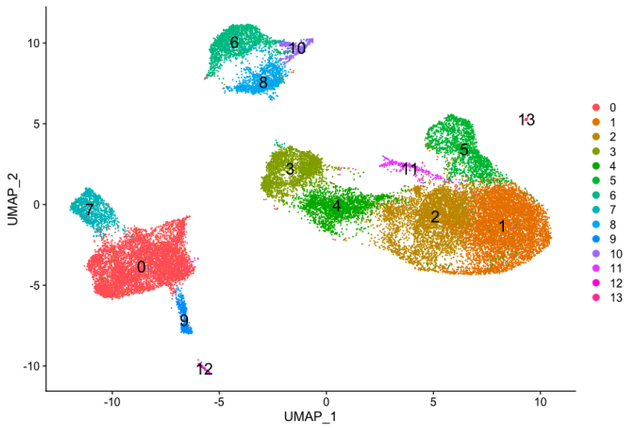
Normalization

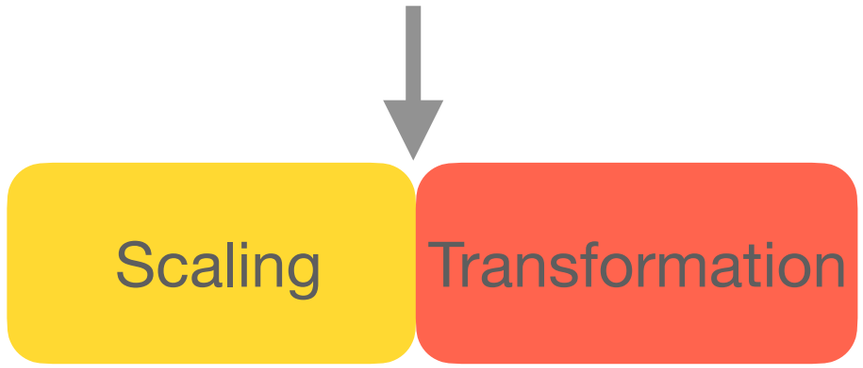
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



Normalization







Adapted from “Normalization methods for single-cell RNA-seq data”, F. Wagner

SCTransform

- A novel statistical approach for the modeling, normalization, and variance stabilization of UMI count data for scRNA-seq

  - Uses a **regularized negative binomial model** to remove the variation due to sequencing depth (total nUMIs per cell)

  - Can add **additional covariates** to include in the model (to regress out effects)

  - The output (residuals) is the normalized expression levels for each transcript tested

Hafemeister C and Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression, bioRxiv 2019 (https://doi.org/10.1101/576827)

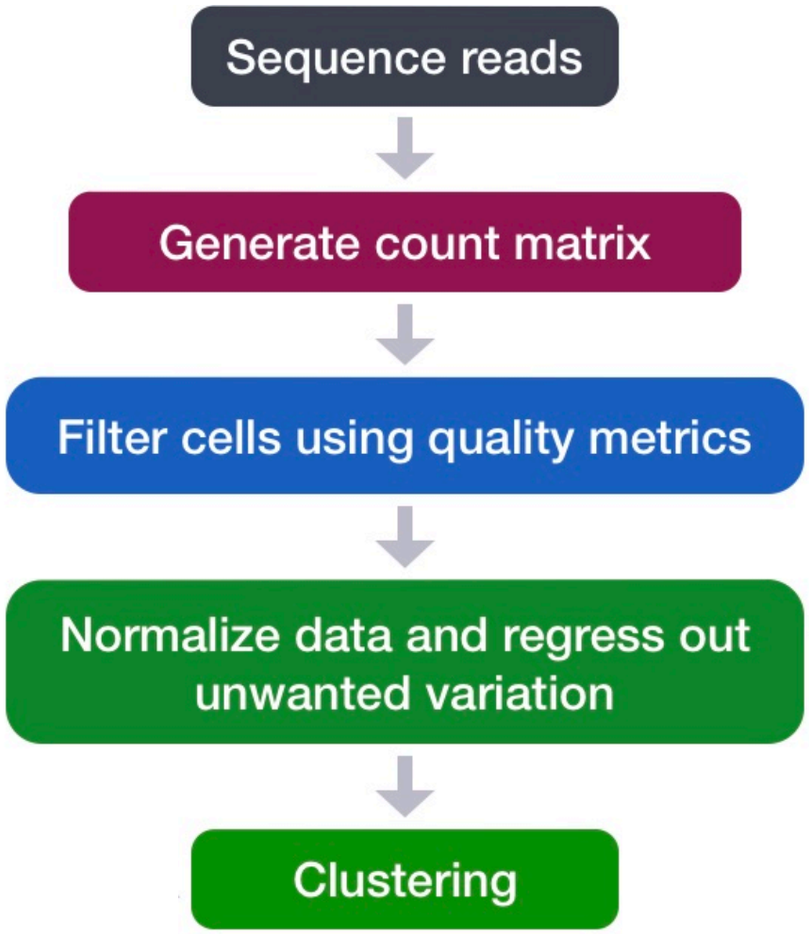
Recommendations

- Always explore the data to see if there are any effects from potential sources of unwanted variation, e.g. mitochondrial counts, cell cycle, etc.

- Do not correct if there is no effect observed.

- Have a good expectation of cell types to be present, and whether cells might be differentiating.

Clustering


Sequence reads
Generate count matrix
Filter cells using quality metrics
Normalize data and regress out unwanted variation
Clustering

Clustering

Goals:

- To **generate cell type-specific clusters** and use known cell type marker genes to determine the identities of the clusters.

- To **d**etermine **whether clusters represent true cell types** or cluster due to biological or technical variation

Challenges:

- **Identifying poor quality clusters** that may be due to uninteresting biological or technical variation

- Identifying the cell types of each cluster

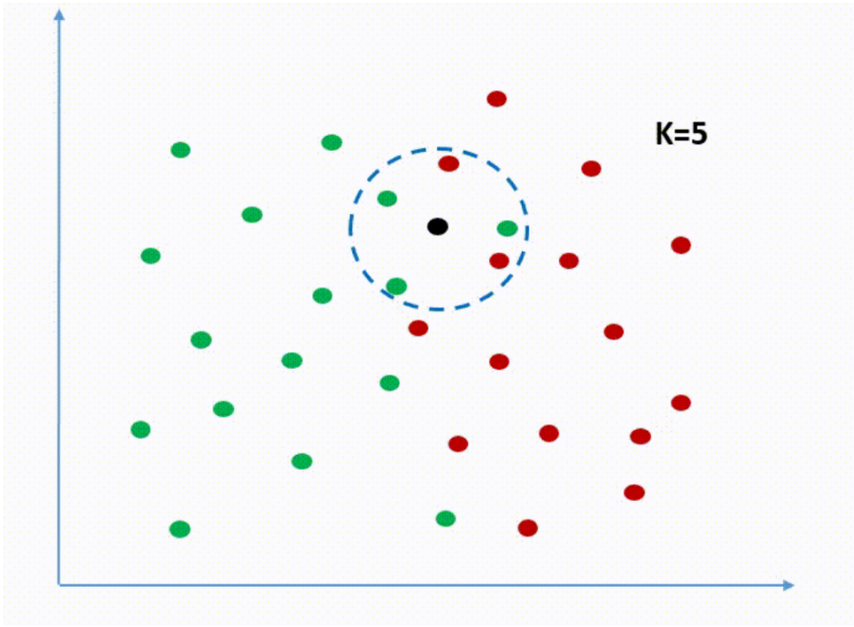
- **Maintaining patience** as this can be a highly iterative process

Clustering of cells

1. First, construct a K-nearest neighbor (KNN) graph based on the distance in PCA space.

    - Edges drawn between cells with similar features expression patterns.

    - Refine the edge weights between any two cells based on shared overlap in their local neighborhoods.

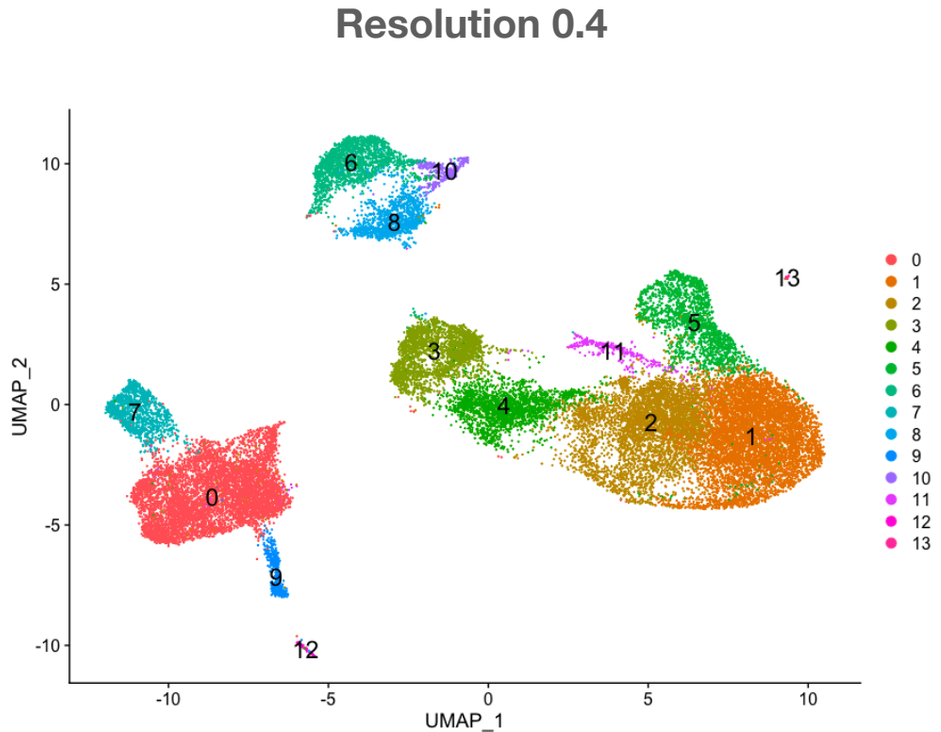


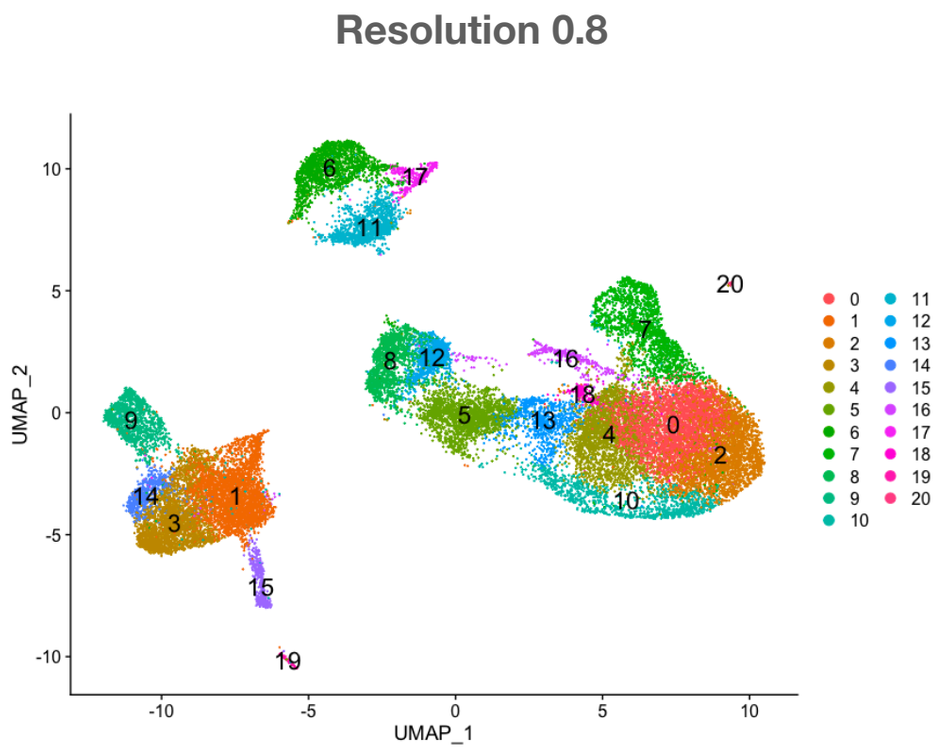
Clustering of cells

2. Iteratively group cells together

    - A *resolution* parameter can be specified by the user, to set the granularity of downstream clustering

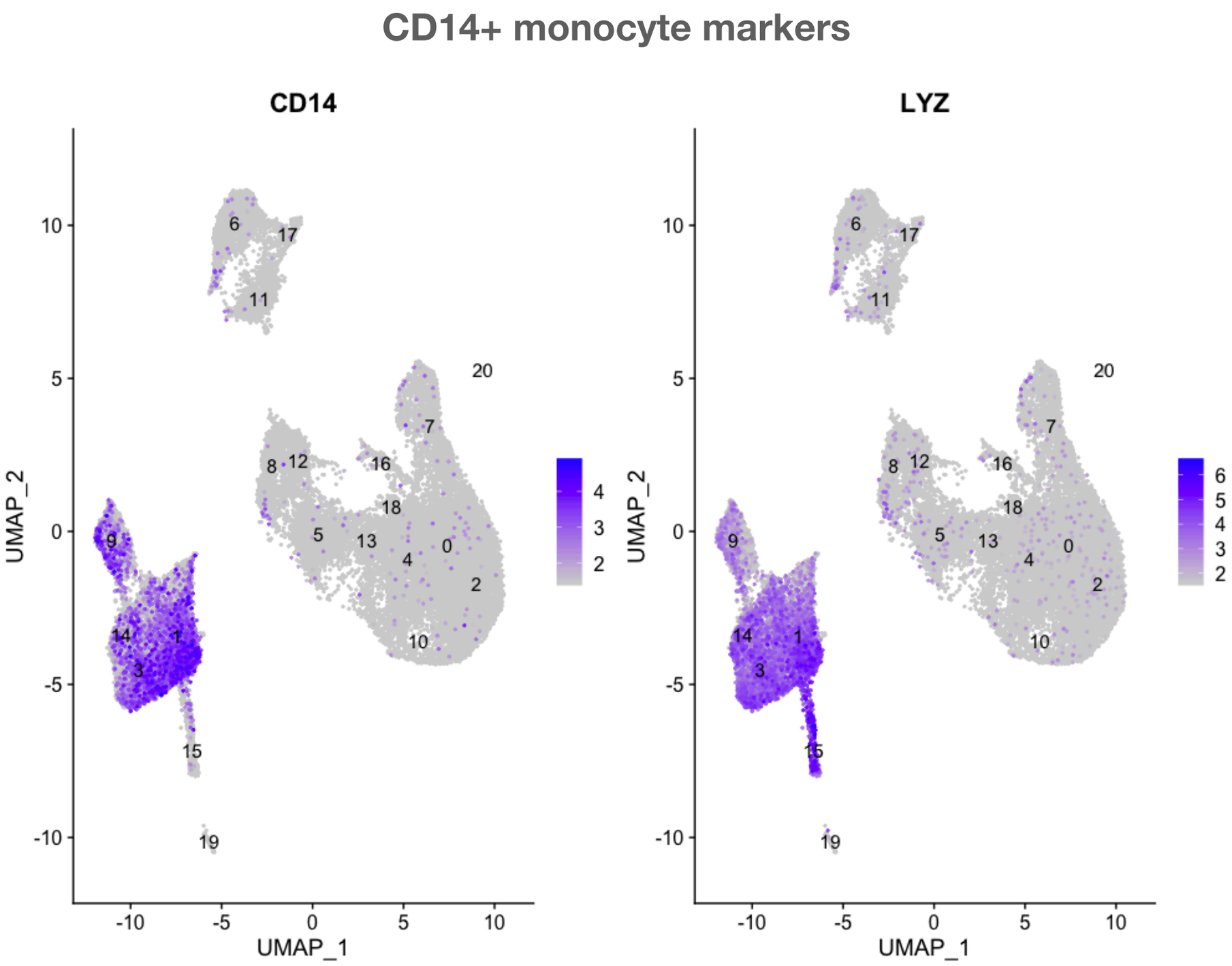
    - Increasing resolution will increase total number of clusters





Cluster quality control

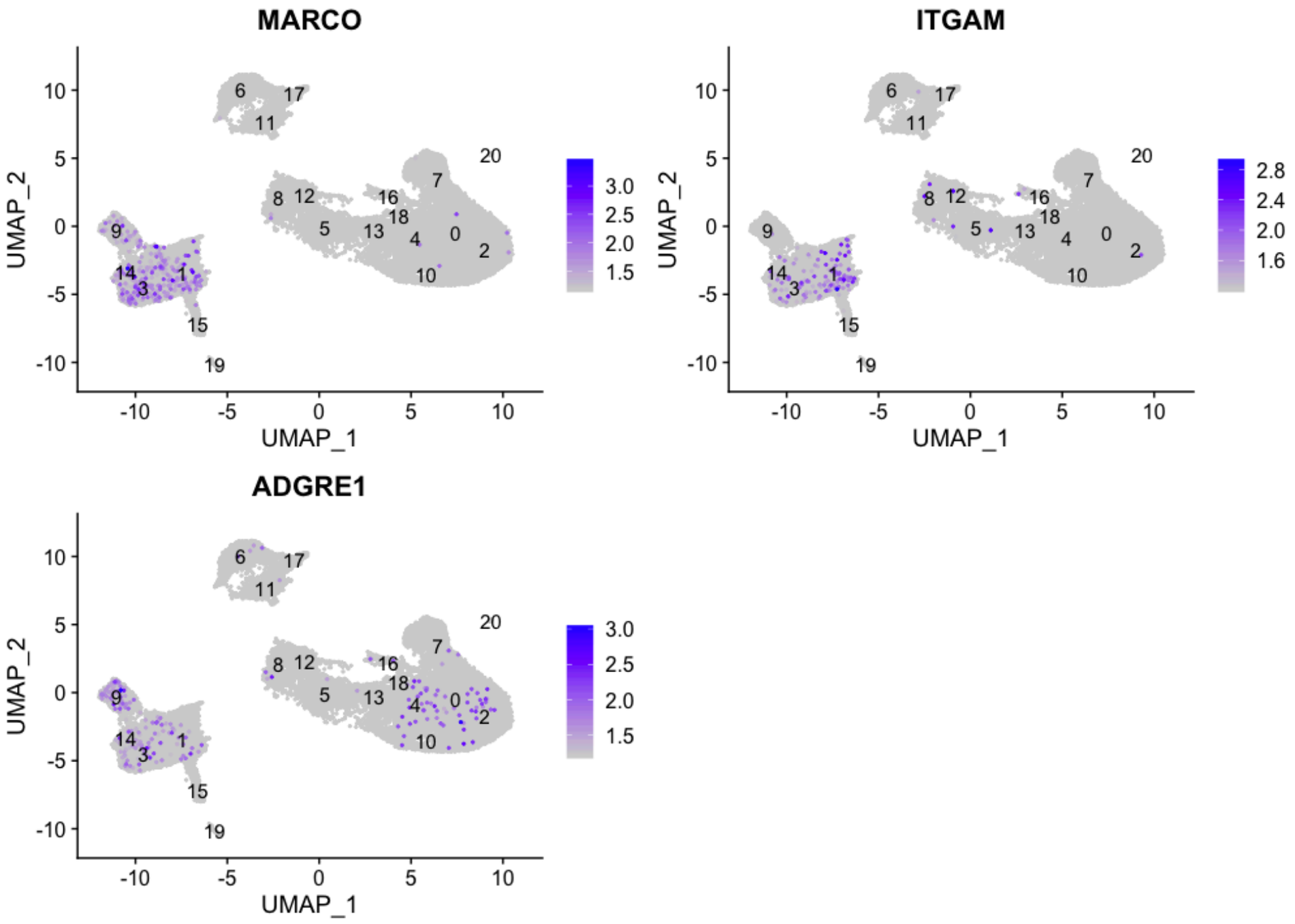
Exploring known cell type markers



Cluster quality control

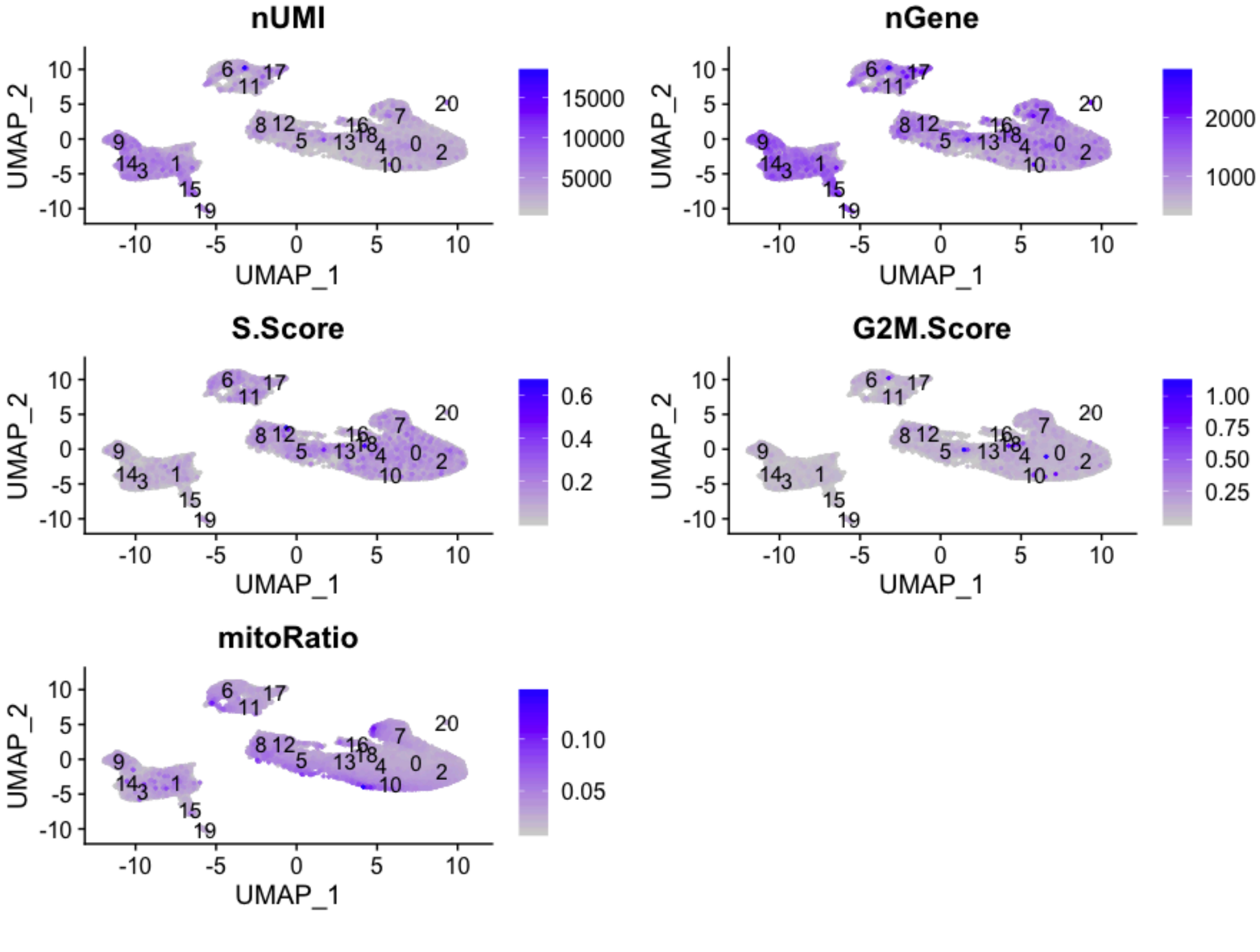
Exploring known cell type markers

Macrophages?



Cluster quality control

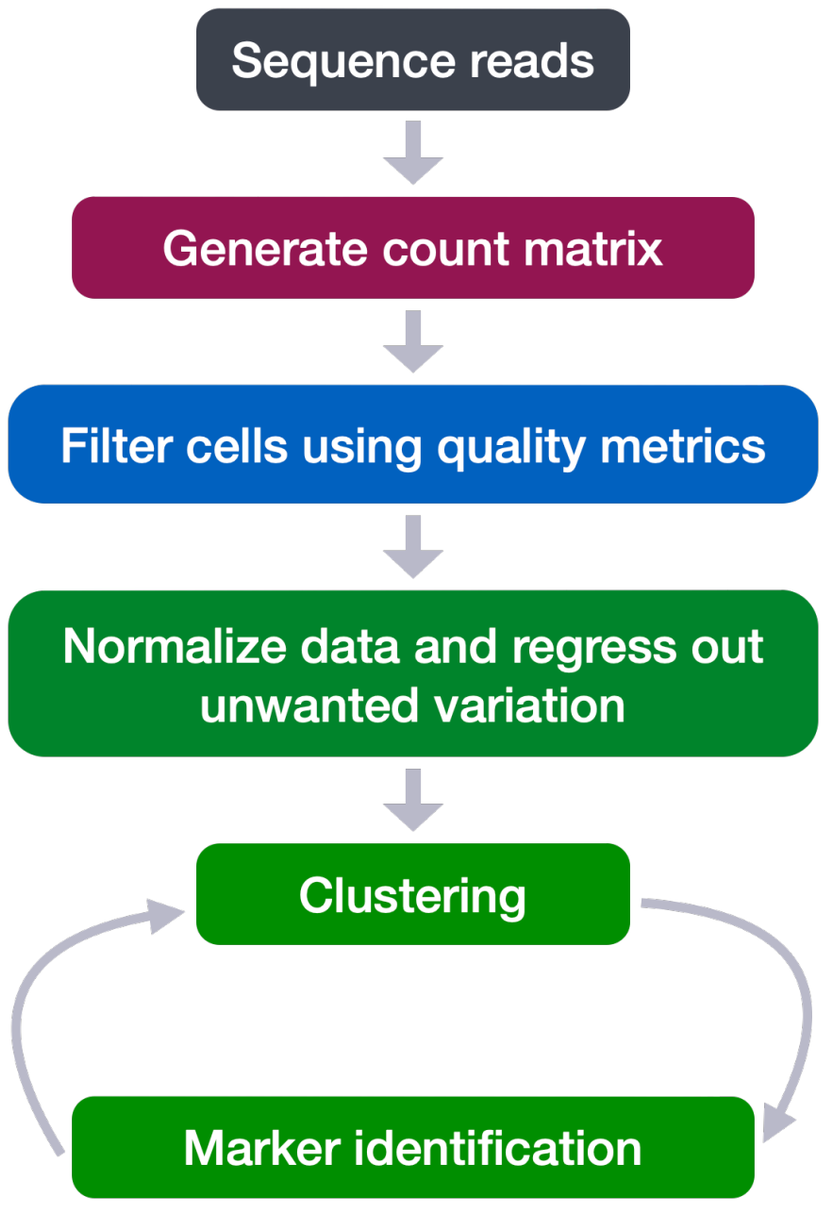
Segregation of clusters by various sources of uninteresting variation



Recommendations

- ❖ Have a good idea of the expectations for your dataset:
  - What cell types are you expecting?
  - Are cells differentiating?
  - Do you expect cell types of low complexity or high mitochondrial content?
- ❖ Identify any junk clusters for removal (i.e. low nUMIs/nGenes)
- ❖ If not detecting cell types as separate clusters:
  - Try changing the cluster resolution
  - Alter the number of PCs used for clustering
  - Subset the data to keep clusters of interest and re-cluster

Marker Identification



Goals:

- To evaluate expression differences between clusters to define gene markers
- To **assign cell types** to clusters or acquire higher confidence in cell type identities already determined
- To determine whether there's a need to **re-cluster based on cell type markers**, perhaps clusters need to be merged or split

Challenges:

- Over-interpretation of the results
- Combining different types of marker identification

Methods for marker identification

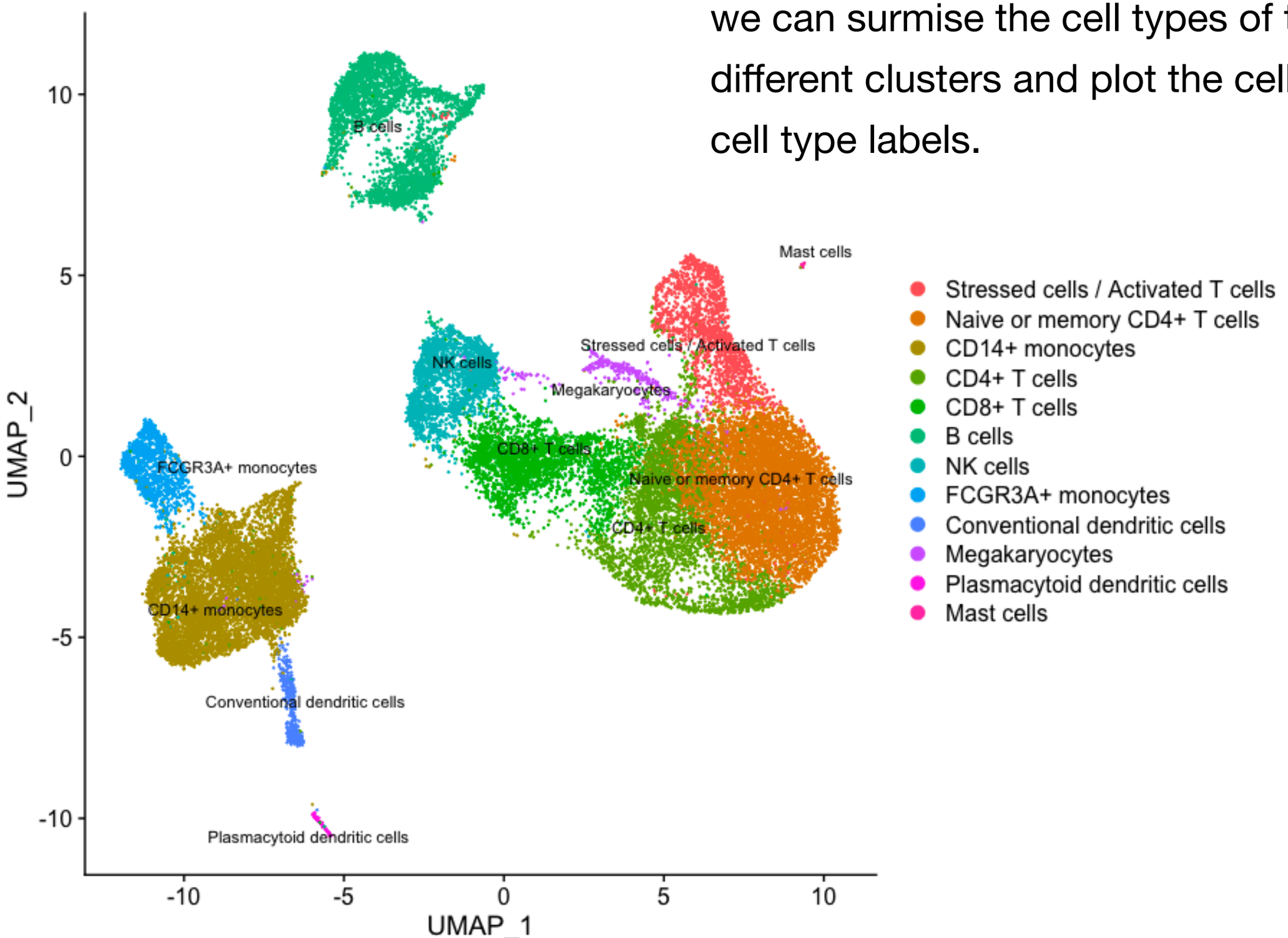
1. **Identification of all markers for each cluster:** this analysis compares each cluster against all others and outputs the genes that are differentially expressed/present.
    - Useful for identifying unknown clusters and improving confidence in hypothesized cell types.

2. **Marker identification between specific clusters:** this analysis explores differentially expressed genes between specific clusters.
    - Useful for determining differences in gene expression between clusters that appear to be representing the same celltype (i.e with markers that are similar) from the above analyses.

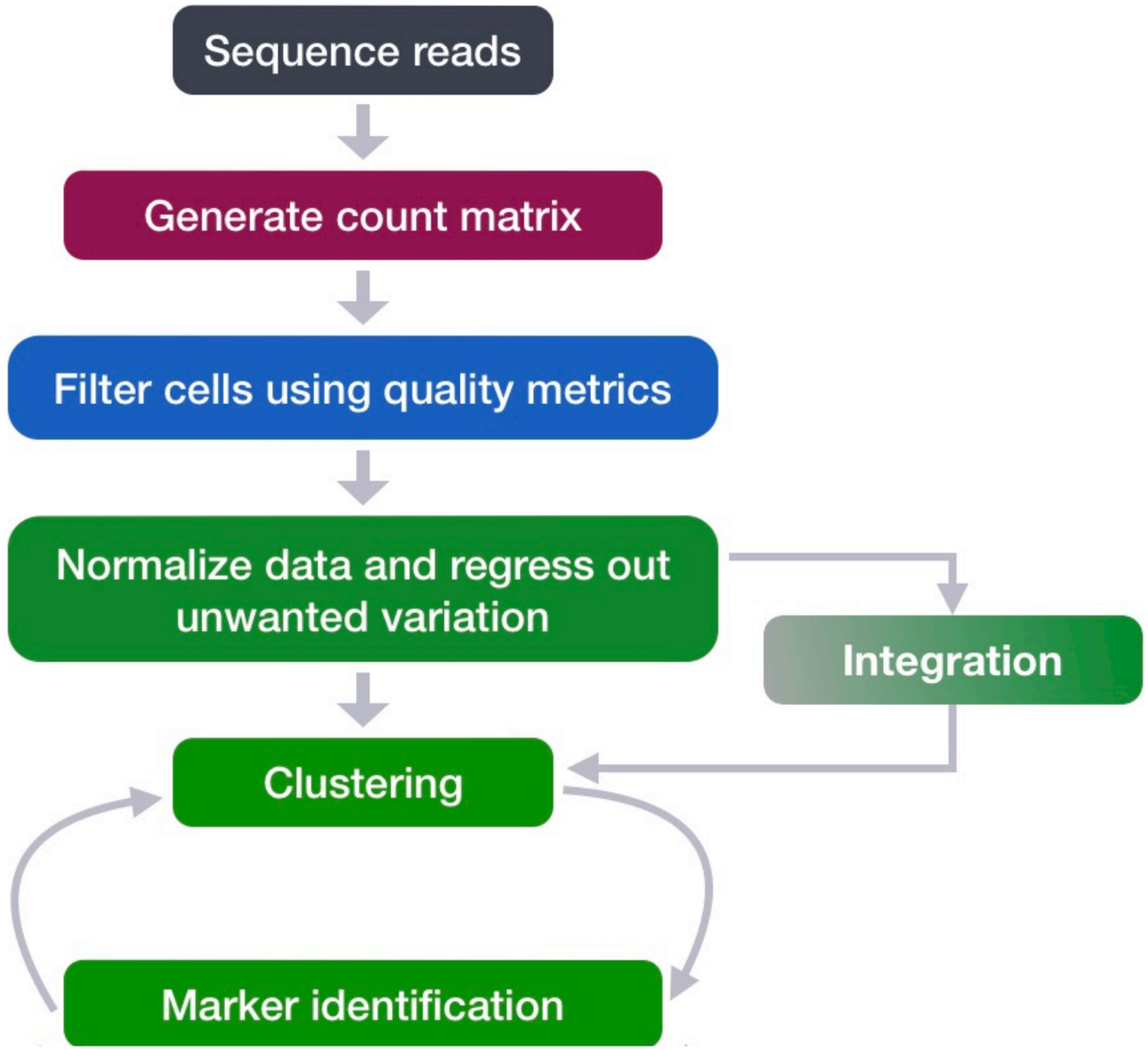
3. **Identification of conserved markers for each cluster:**
    - Useful with more than one condition to identify cell type markers that are conserved across conditions.

The final product

Taking all of this information together, we can surmise the cell types of the different clusters and plot the cells with cell type labels.



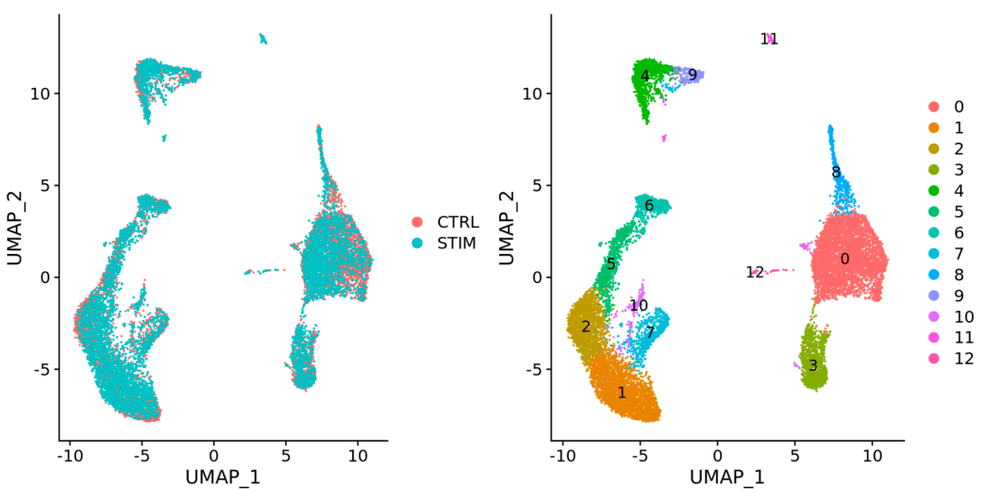
Working with multiple samples



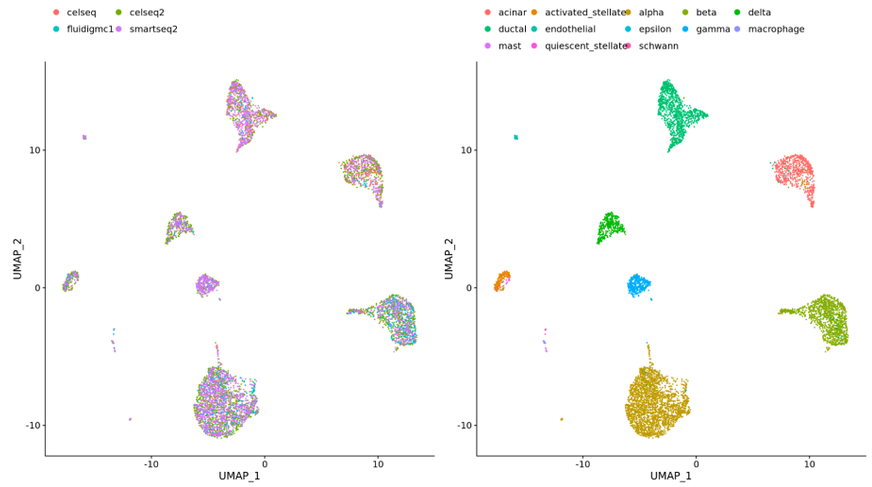
Integration

Using the shared highly variable genes from each group, to “integrate” or “harmonize” the groups by aligning samples based on the “common set of biological features”.

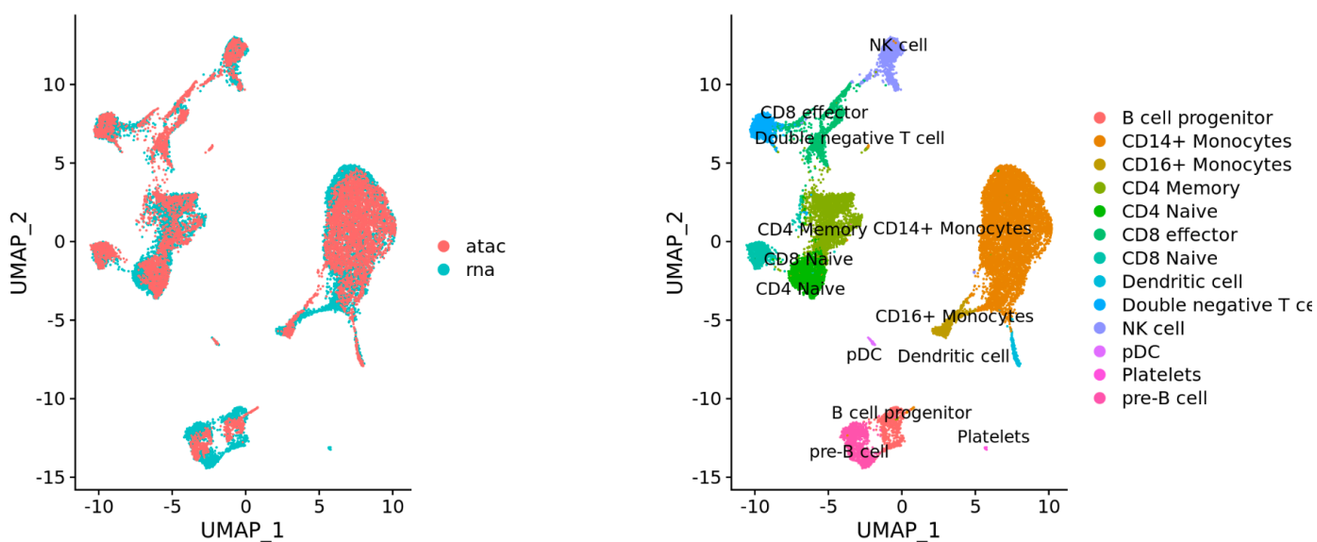
Different conditions



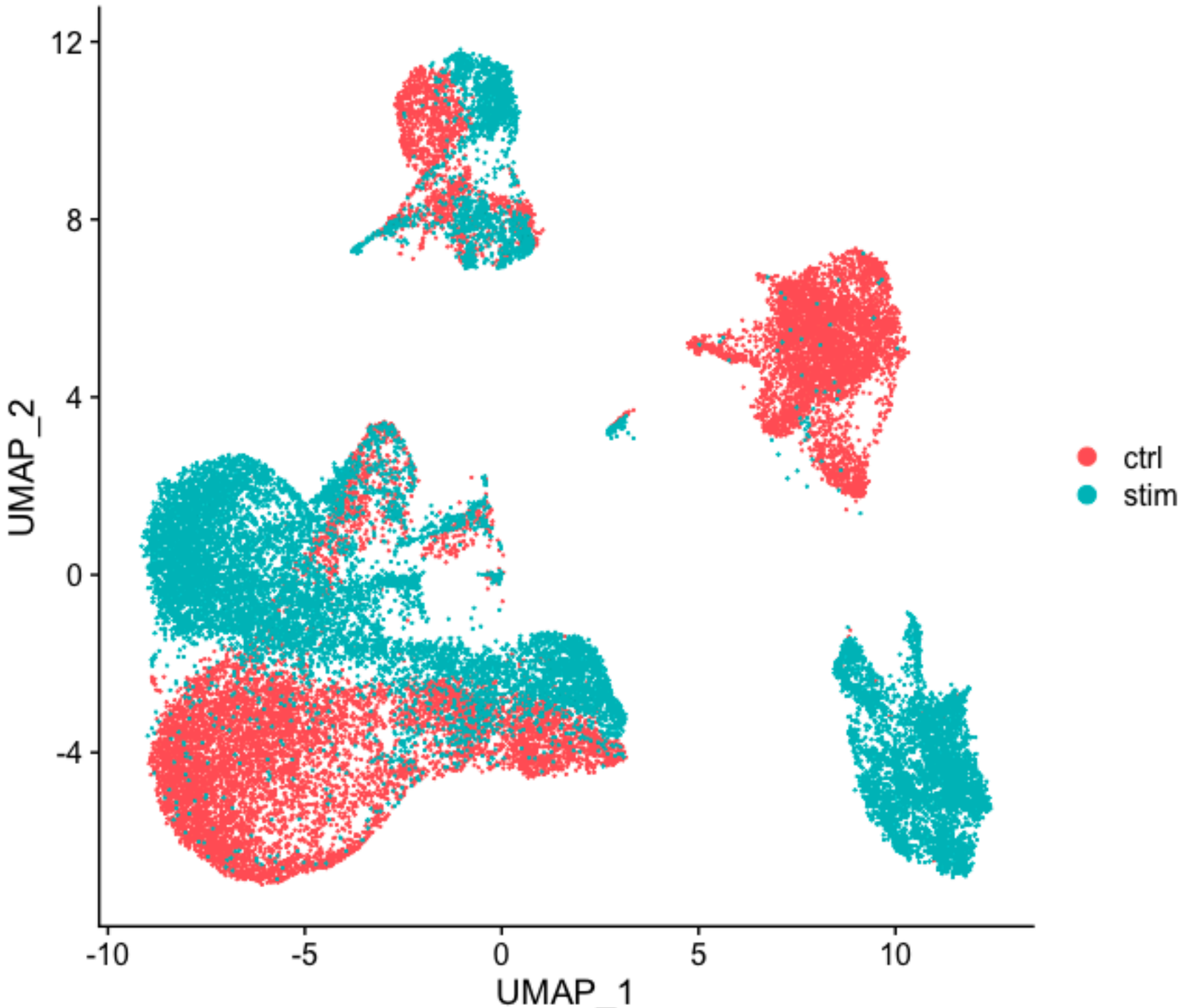
Different datasets



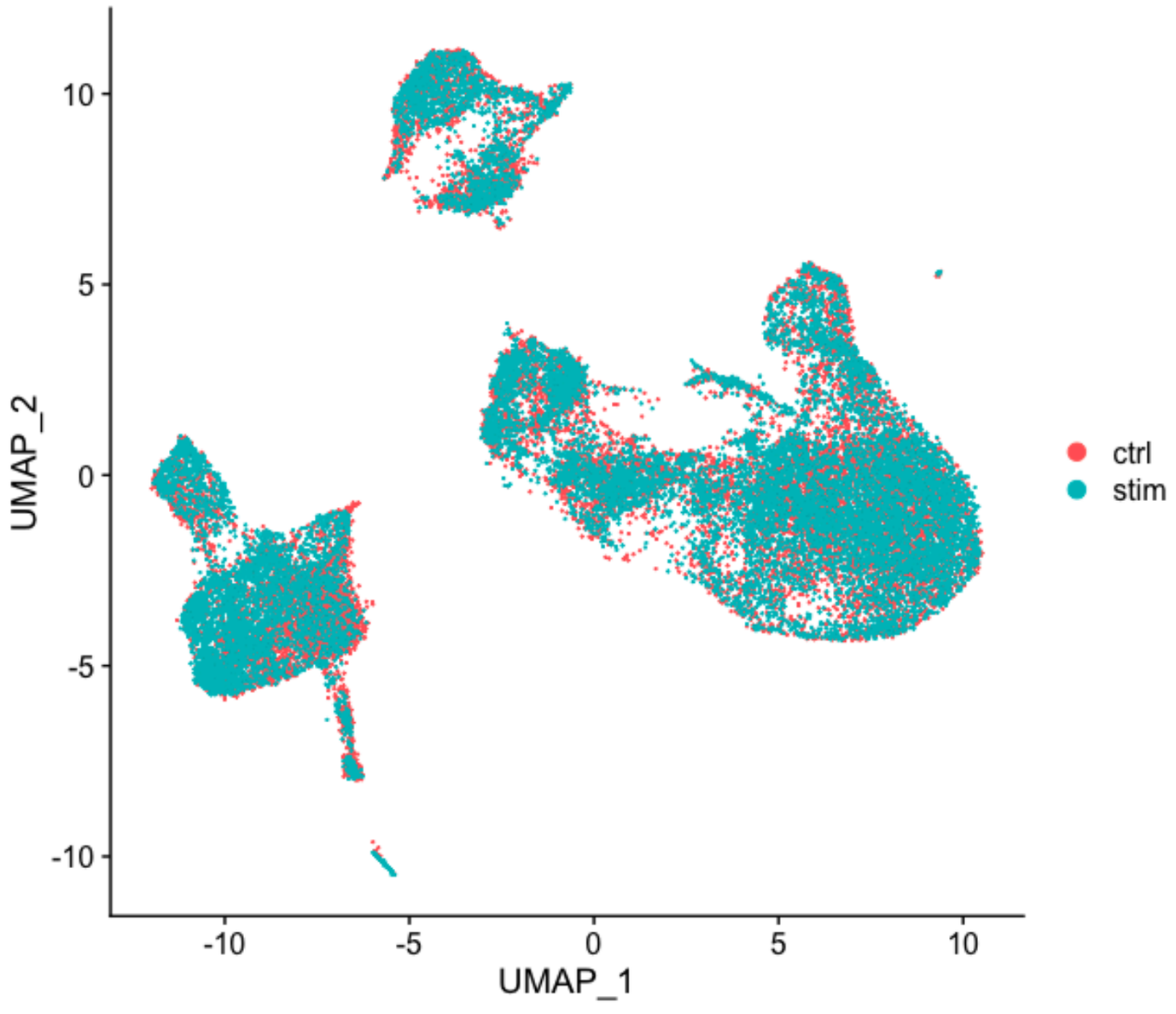
Different modalities

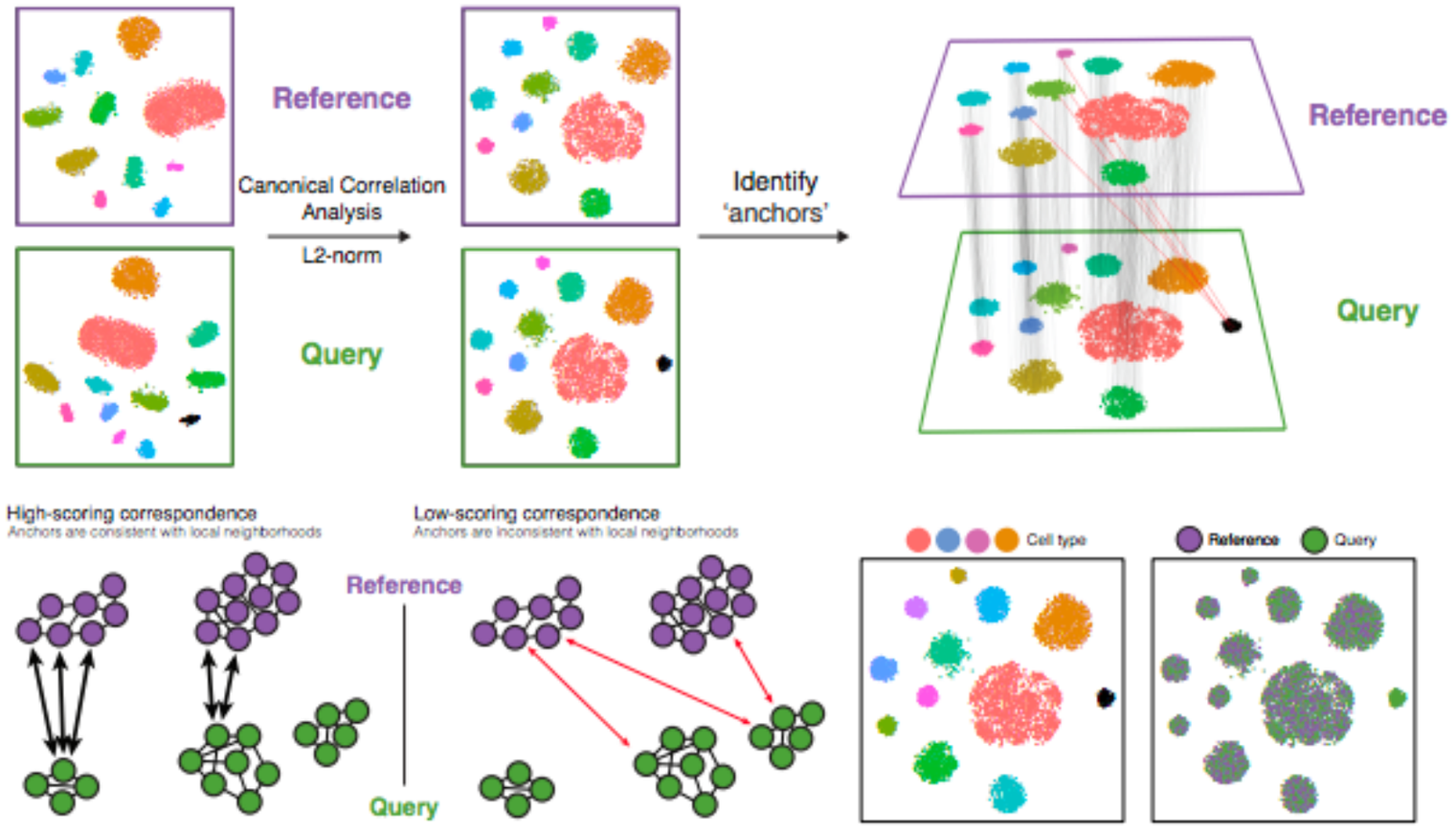


To integrate or not to integrate?



Clusters will now be a representation of cells from both conditions allowing for more interpretable results downstream.





1. CCA (a form of PCA) identifies shared sources of variation between the conditions/groups.

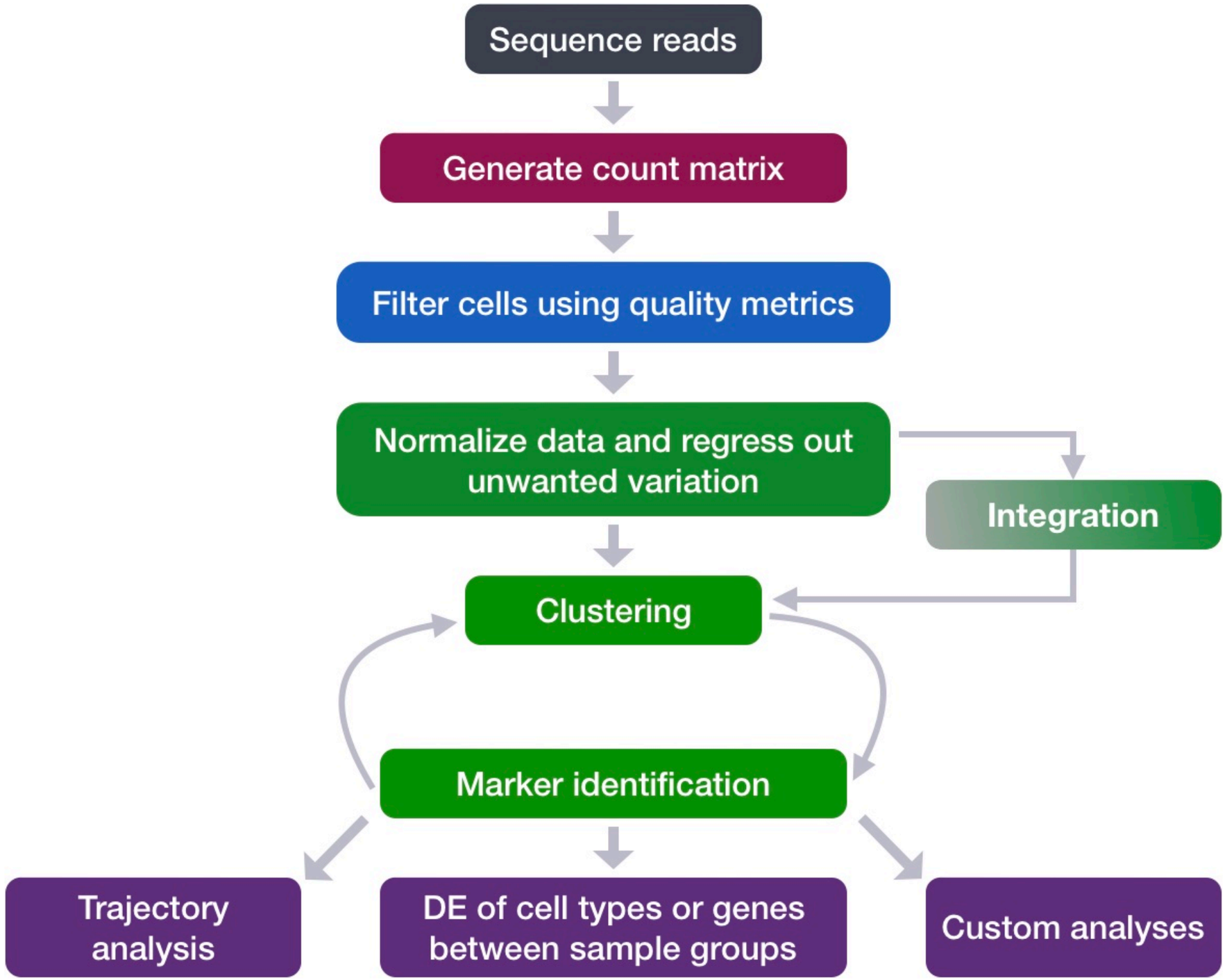
2. Identify anchors or mutual nearest neighbors (MNNs) across conditions. If two cells are 'best buddies' in both directions they will be marked as anchors.

3. Filter anchors to remove incorrect ones.

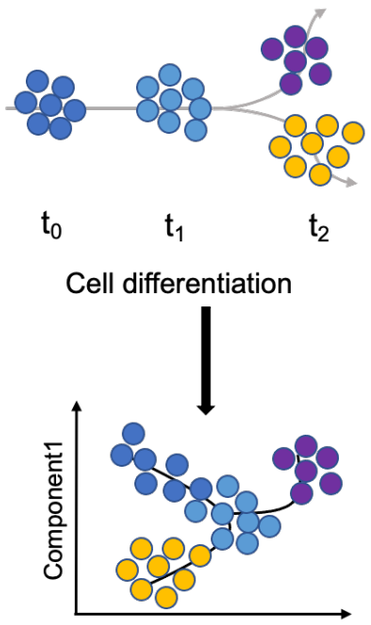
4. Integrate. Use anchors and corresponding scores to transform the cell expression values.

Stuart T and Butler A, et al. Comprehensive integration of single cell data, bioRxiv 2018 (https://doi.org/10.1101/460147)

What's next?


Sequence reads
Generate count matrix
Filter cells using quality metrics
Normalize data and regress out unwanted variation
Integration
Clustering
Marker identification
Trajectory analysis
DE of cell types or genes between sample groups
Custom analyses

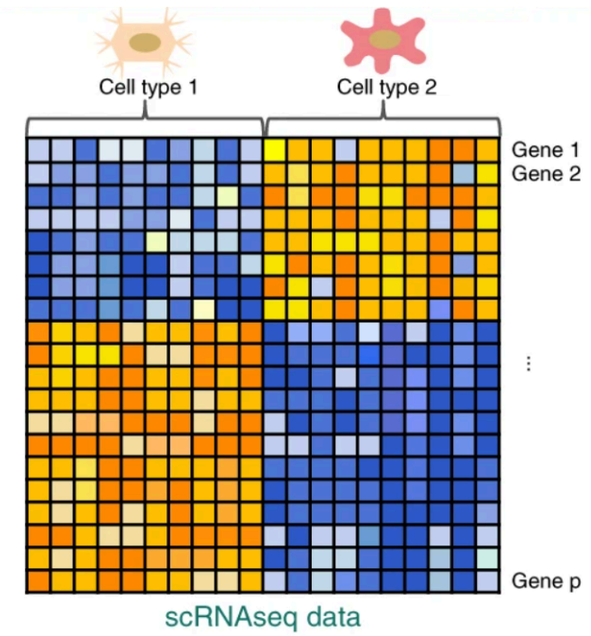
Trajectory analysis



Trajectory analysis, or lineage tracing, could be performed if trying to determine the progression between cell types or cell states. For example, we could explore:

- Differentiation processes
- Expression changes over time
- Cell state changes in expression

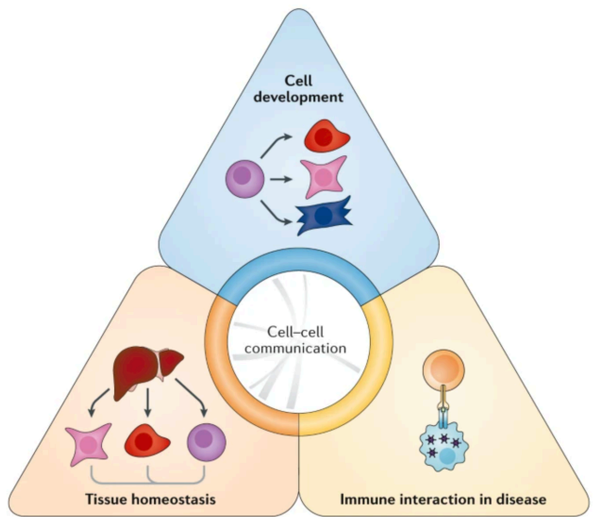
DE of cell types or genes between sample groups



For individual clusters, perform differential expression analysis between conditions/groups

- Biological replicates are **necessary** to proceed with this analysis
- Pseudobulk analysis is a popular approach

Custom analyses



- Deciphering cell-cell interactions (Ligand-receptor analyses)
- Sub-clustering to identify cell subtypes
- Experiments to validate specific results
- …

Helpful resources

- Seurat vignettes
- Seurat cheatsheet
- Satija Lab: Single Cell Genomics Day
- “Principal Component Analysis (PCA) clearly explained”, a video from Josh Starmer
- Additional information about cell cycle scoring
- CellMarker resource
- HBC Introduction to single-cell RNA-seq analysis workshop materials

Acknowledgments

Thanks to many people who contributed to the workshop materials which this lecture is based off of:

- All HCBC team members, especially Shannan Ho Sui, Meeta Mistry and Mary Piper

Shannan Ho Sui
Director

Meeta Mistry

Mary Piper

- HMS Single Cell Core, Mandovi Chatterjee and Arpita Kulkarni

HARVARD MEDICAL SCHOOL

HBC
Harvard Chan Bioinformatics Core