Single-cell RNA-seq data analysis workshop

Audience	Computational skills required	Duration
Biologists	Introduction to R	3-session online workshop (~7.5 hours of trainer-led time)

Description

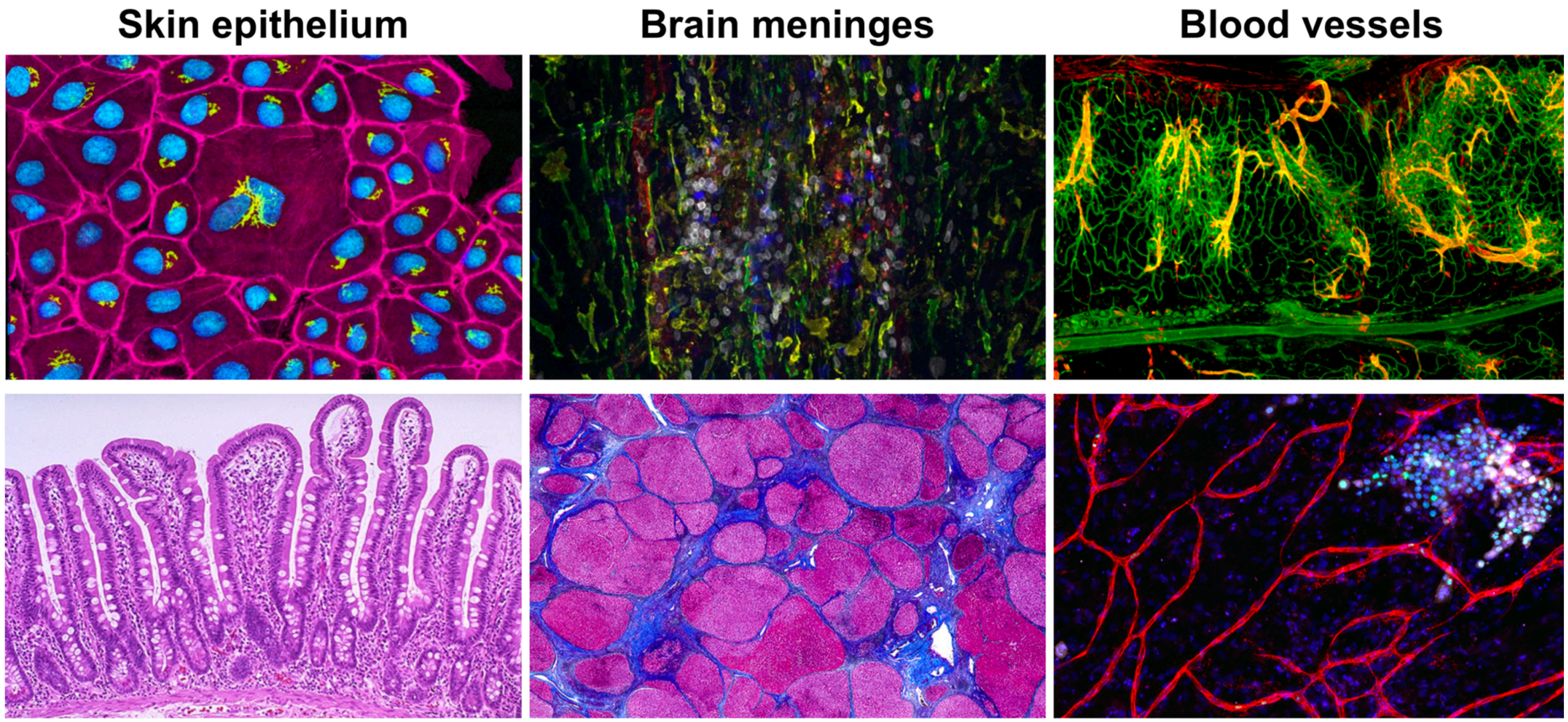
This repository has teaching materials for a hands-on **Introduction to single-cell RNA-seq analysis** workshop. This workshop will instruct participants on how to design a single-cell RNA-seq experiment, and how to efficiently manage and analyze the data starting from count matrices. This

Mary Piper

https://hbctraining.github.io/scRNA-seq_online/

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)

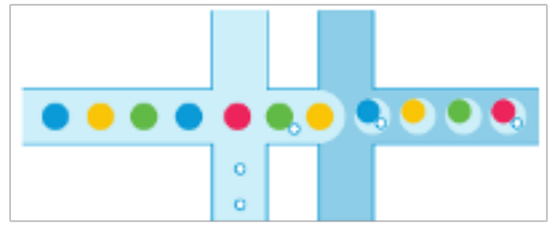
CEL-Seq
MARS-Seq
SMART-Seq
SCRB-Seq
Quartz-Seq

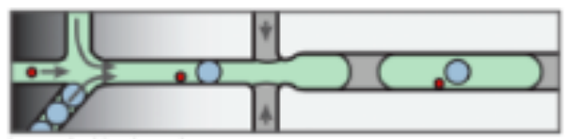
In nano/microwells

Seq-Well

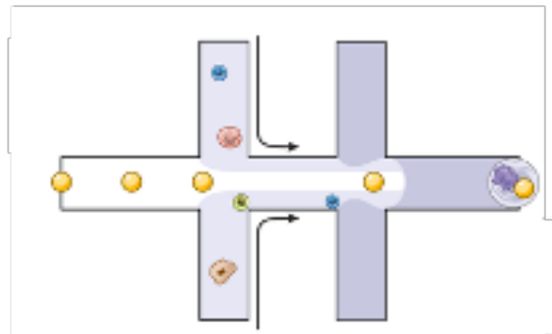
BD Rhapsody

In droplets





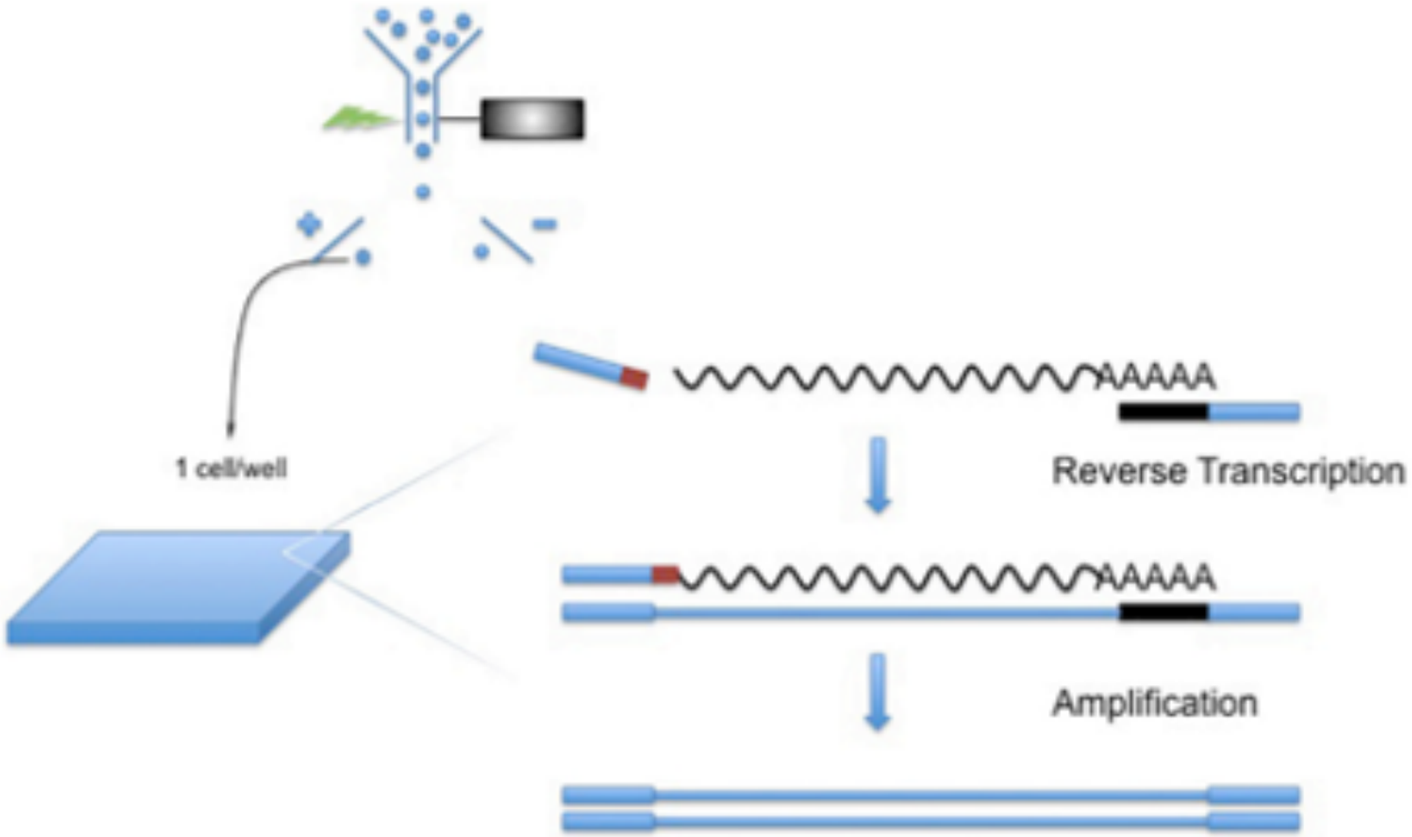
InDrops



DropSeq

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

SMART-seq (v3): Full length transcript



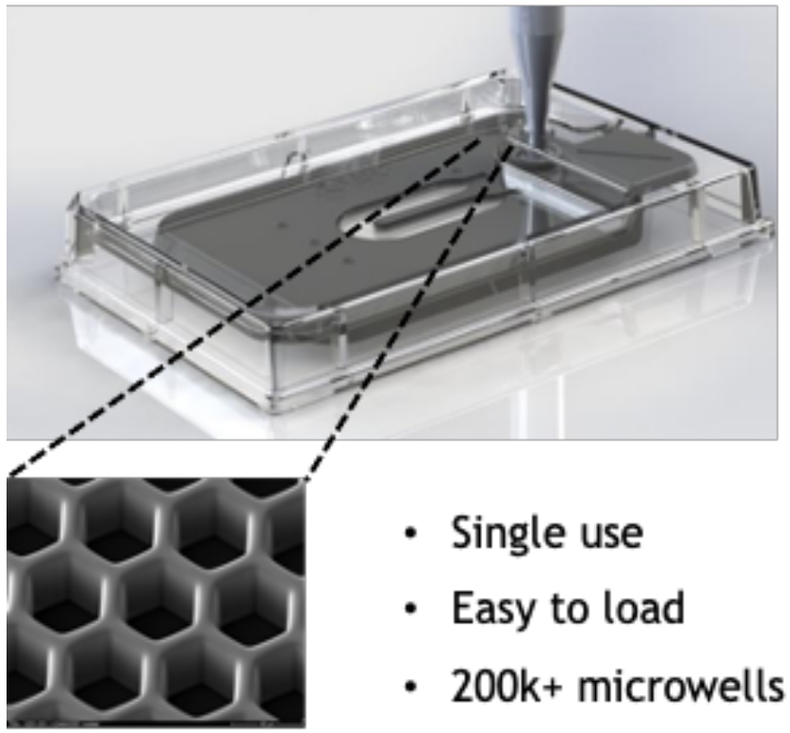
- Sorted cells of interest are picked and placed into single well.
- Only single cell method that gives full transcript information.
- Currently best option for low cell number samples (100s to 1000s)

H Lim et al, Profiling Individual Human Embryonic Stem Cells by Quantitative RT-PCR. J. Vis. Exp. (87), e51408, 2014 (doi:10.3791/51408)
M Hagemann-Jensen et al, Single-cell RNA counting at allele- and isoform-resolution using Smart-seq3 bioRxiv 2019 (doi: https://doi.org/10.1101/817924)

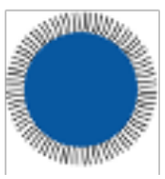
Slide adapted from “Introduction to Single Cell RNA-sequencing: a practical guideline”, Mandovi Chatterjee, Ph. D.

Microwell-based platform: BD Rhapsody

Cartridge



- Single use
- Easy to load
- 200k+ microwells



- Magnetic bead with immobilized oligos

Scanner

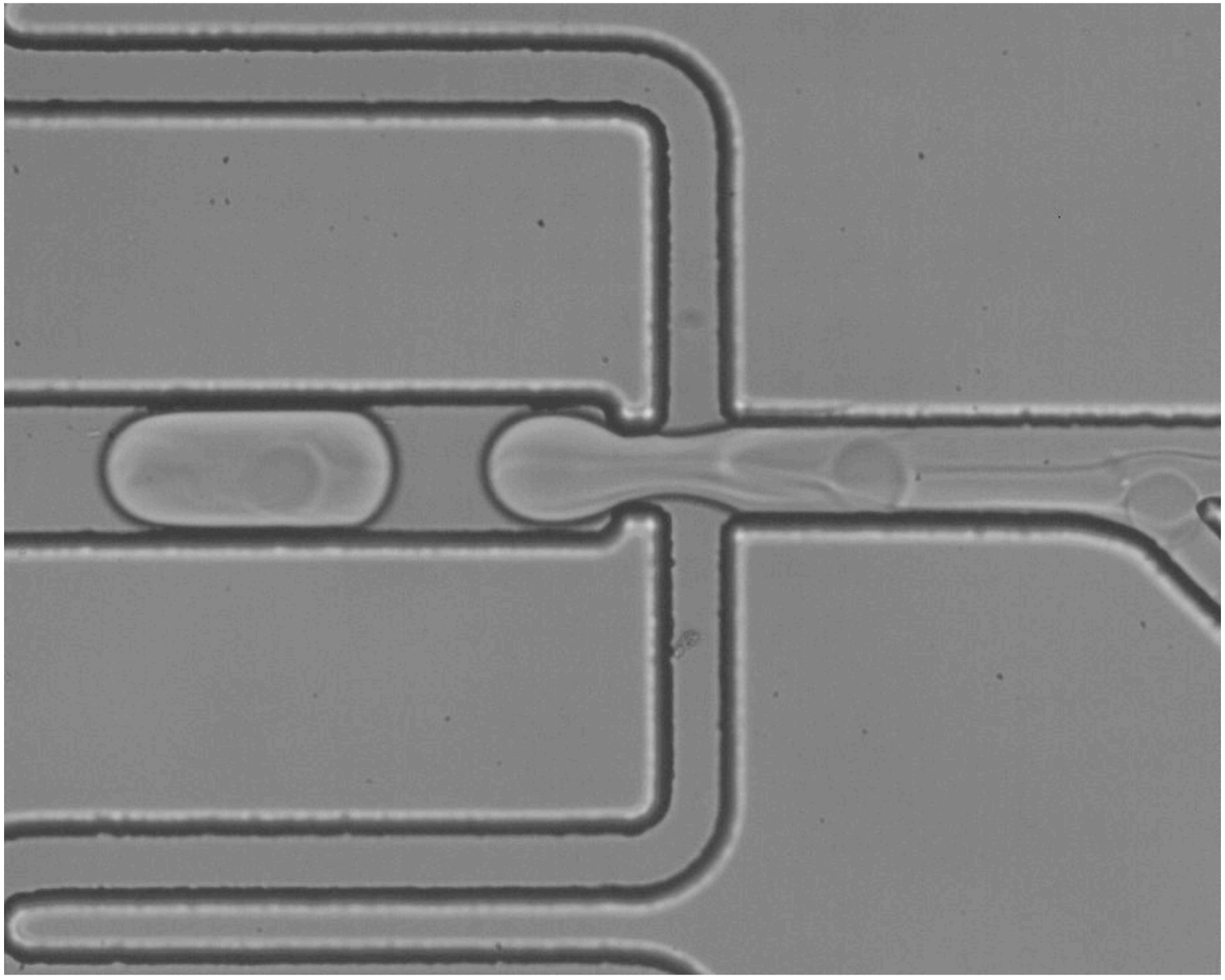
- Helps with the workflow
- Real-time cell count and viability
- True doublet rate
- Comprehensive statistics report

Rhapsody Express

- Manipulation of Microwell cartridge
- Portable/Benchtop
- Can be purchased individually

https://www.bdbiosciences.com/en-us/products/instruments/single-cell-multiomics-systems/rhapsody

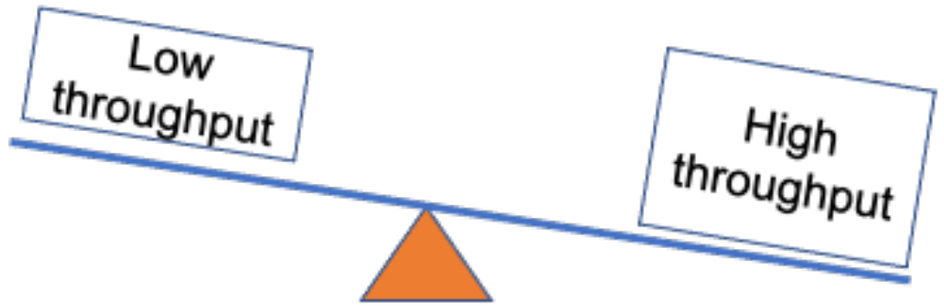
Slide adapted from “Introduction to Single Cell RNA-sequencing: a practical guideline”, Mandovi Chatterjee, Ph. D.



Video generated by
Single Cell Core @ HMS

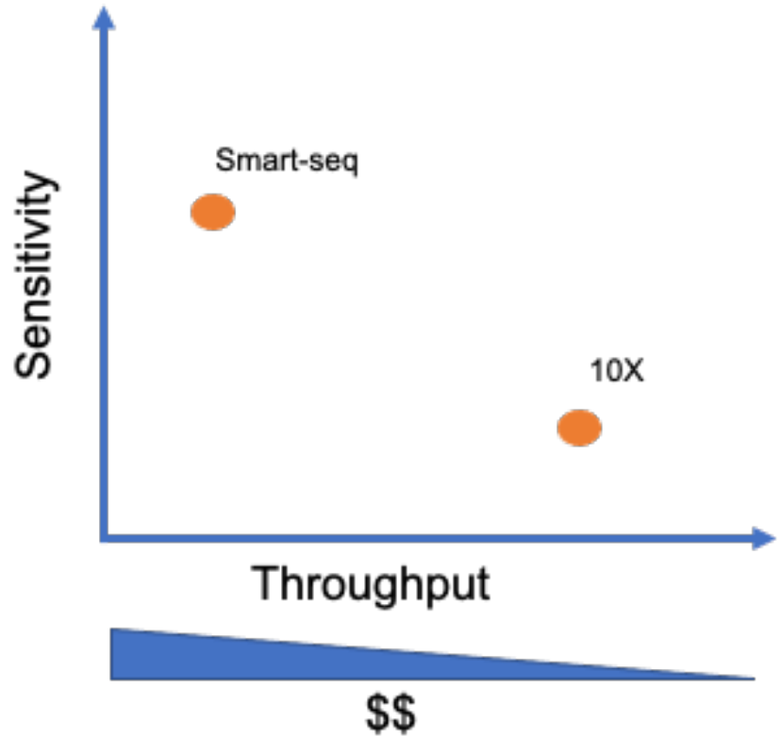
Slide taken from “Introduction to Single Cell RNA-sequencing: a practical guideline”, Mandovi Chatterjee, Ph. D.

Which platform should I use?



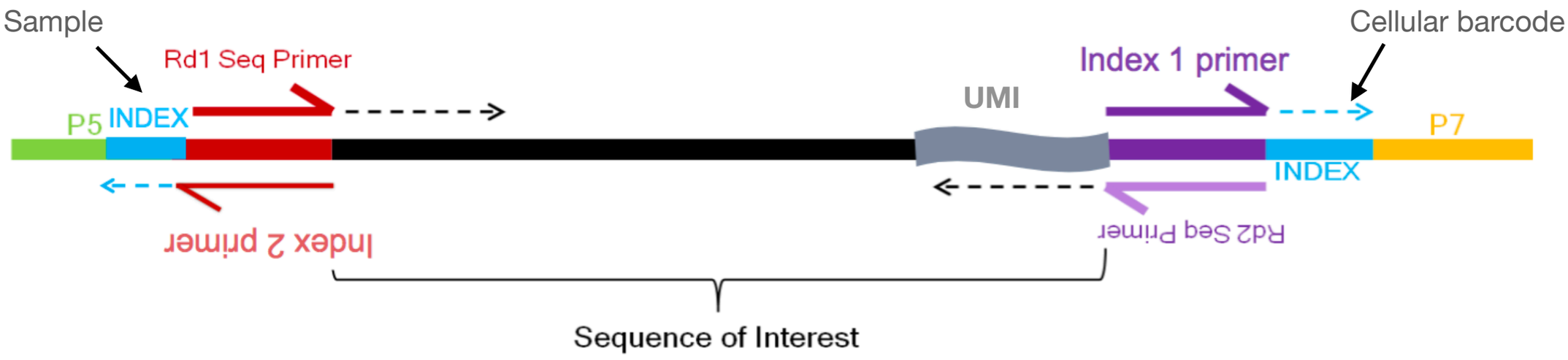
- Limited starting material, not enough for the high-throughput platforms.
- Interested in studying full length transcript, e.g. splice forms

- Need to profile a lot of cells- to detect a very rare sub-population in the tissue.
- Looking for overall heterogeneity in the tissue



Slide adapted 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)

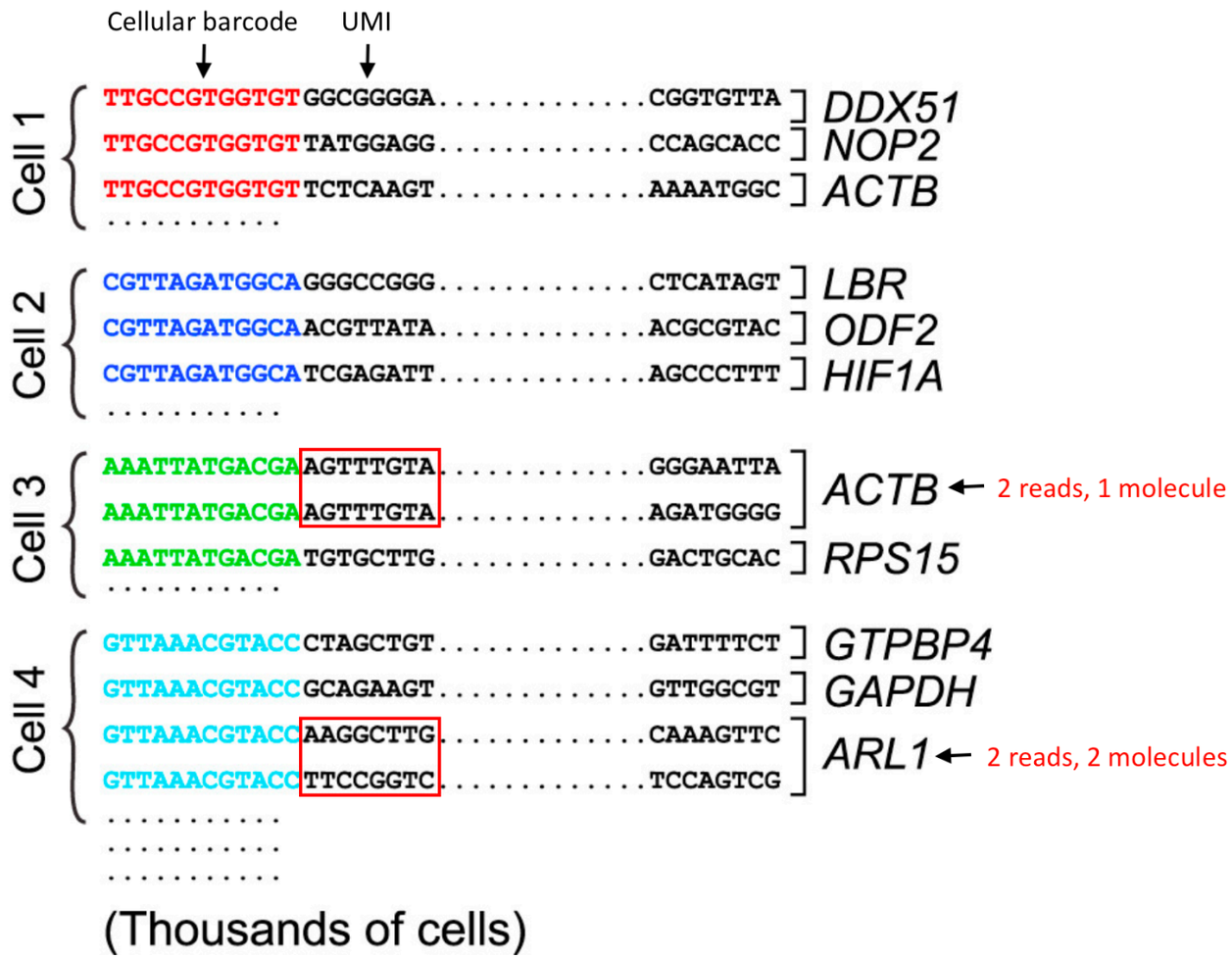
- **Sequencing read1:** the Read1 sequence (red top arrow)
- **Sequencing read2:** the Read2 sequence (purple bottom arrow)

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

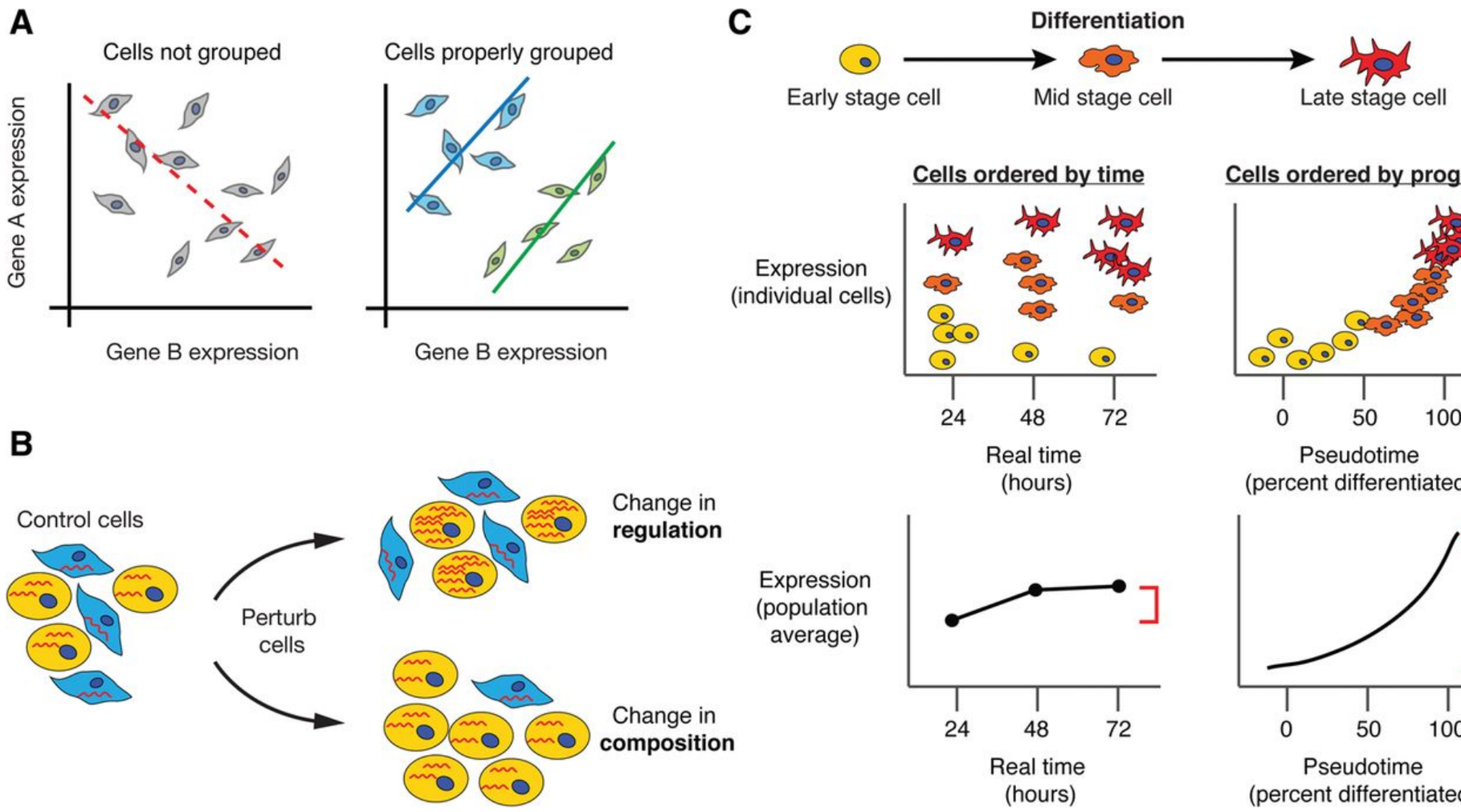


Image credit: *Trapnell, C. Defining cell types and states with single-cell genomics, Genome Research 2015 (doi: https://dx.doi.org/10.1101/gr.190595.115)*

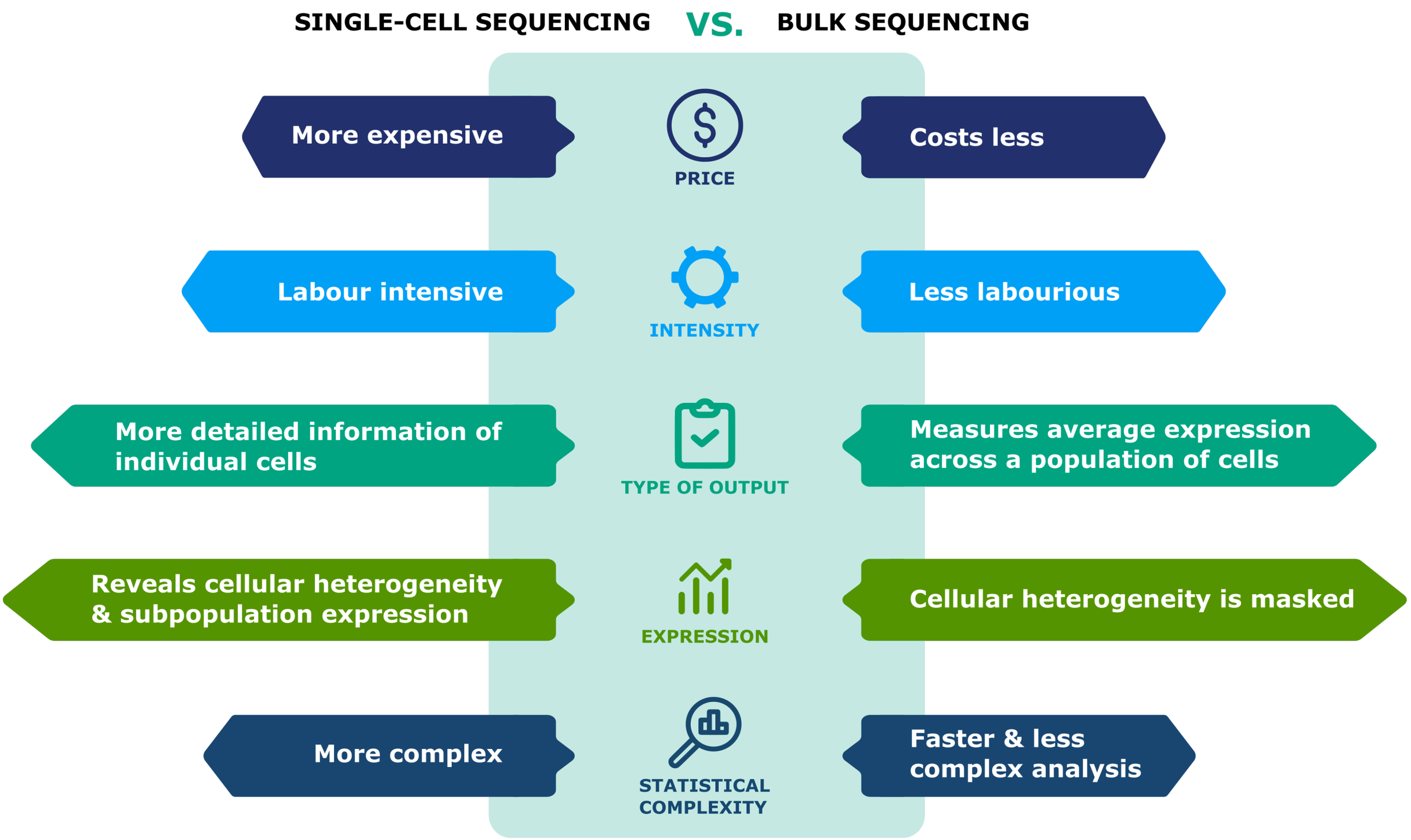


Image source: https://lizard.bio/blog/single-cell-vs-bulk/

Challenges with scRNA-seq data

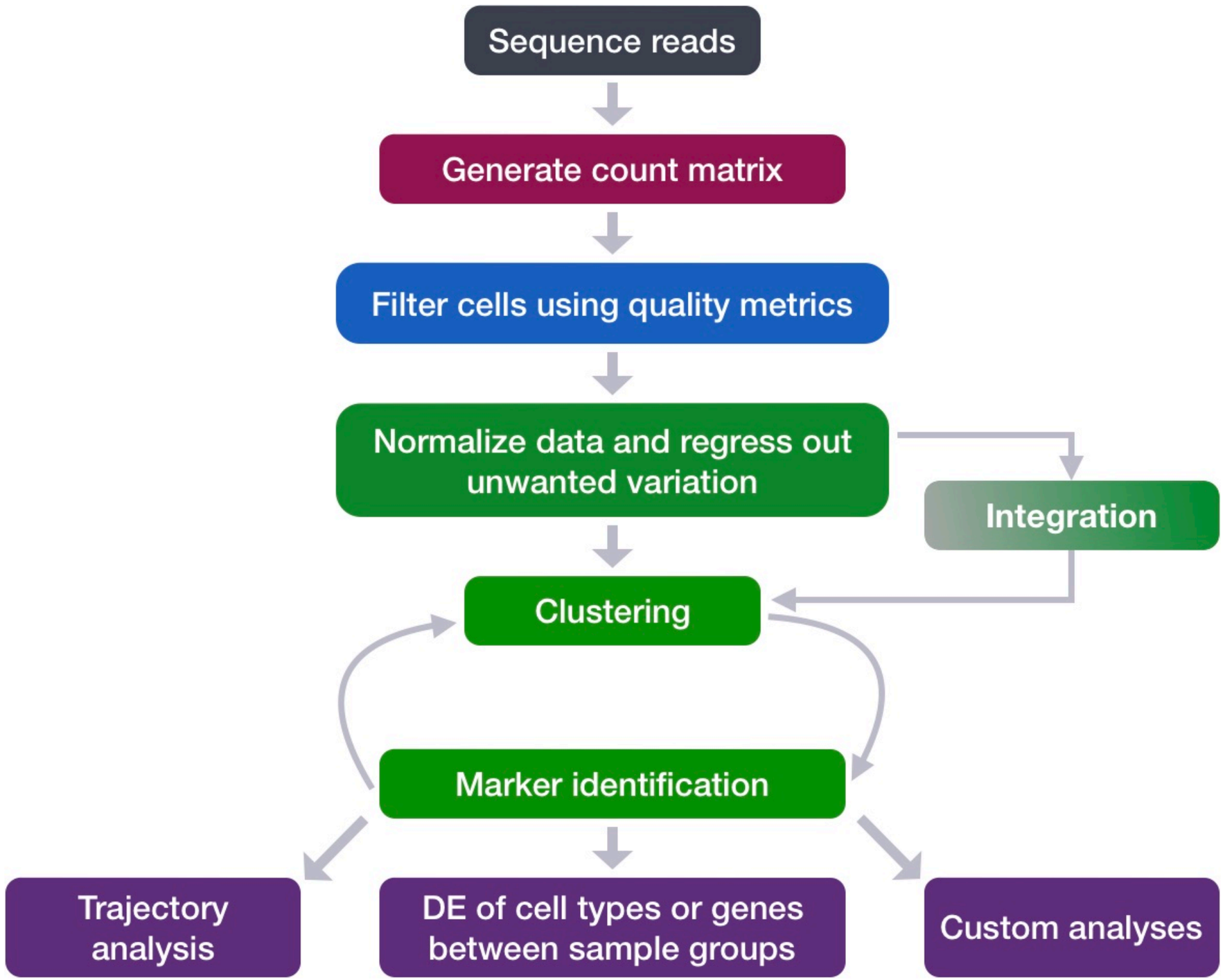
- Large volume of data
- Low depth of sequencing per cell (zero-inflation)
  - Often detecting only 10-50% of the transcriptome per cell
- Biological variability across cells/samples can obscure the cell type identities
- Technical variability across cells/samples

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

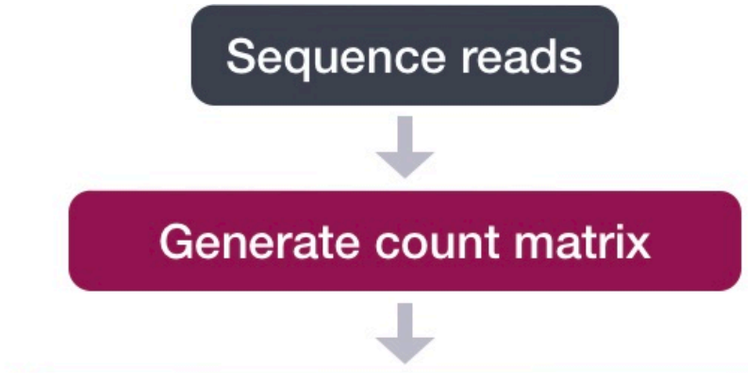
Recommendations

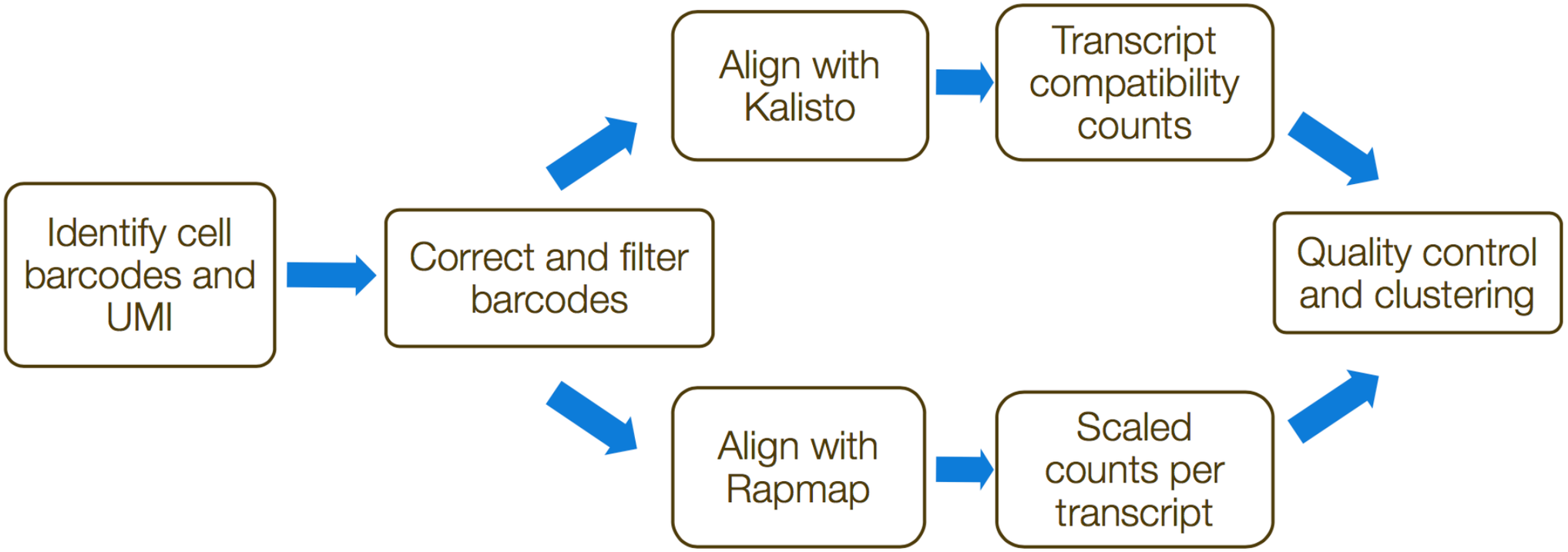
- Do not perform single-cell RNA-seq unless it is necessary for the experimental question of interest.
- Understand the details of the experimental question you wish to address.
- Avoid technical sources of variability, if possible:

Single-cell RNA-seq analysis workflow



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:

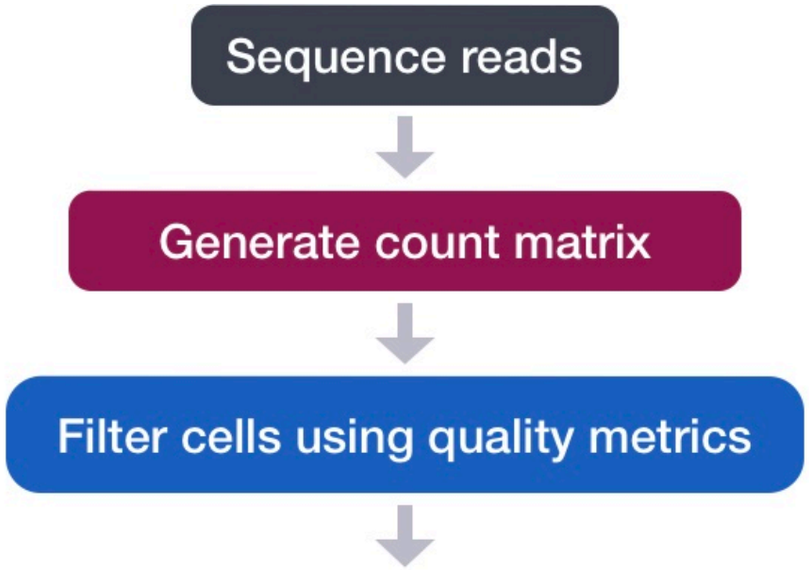
1. Formatting reads and filtering noisy cellular barcodes
2. Demultiplexing the samples
3. Mapping/pseudo-mapping to transcriptome
4. Collapsing UMIs and quantification of reads

Count matrix

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

Image credit: *extracted from Lafzi et al. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies, Nature Protocols 2018 (https://doi.org/10.1038/s41596-018-0073-y)*

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

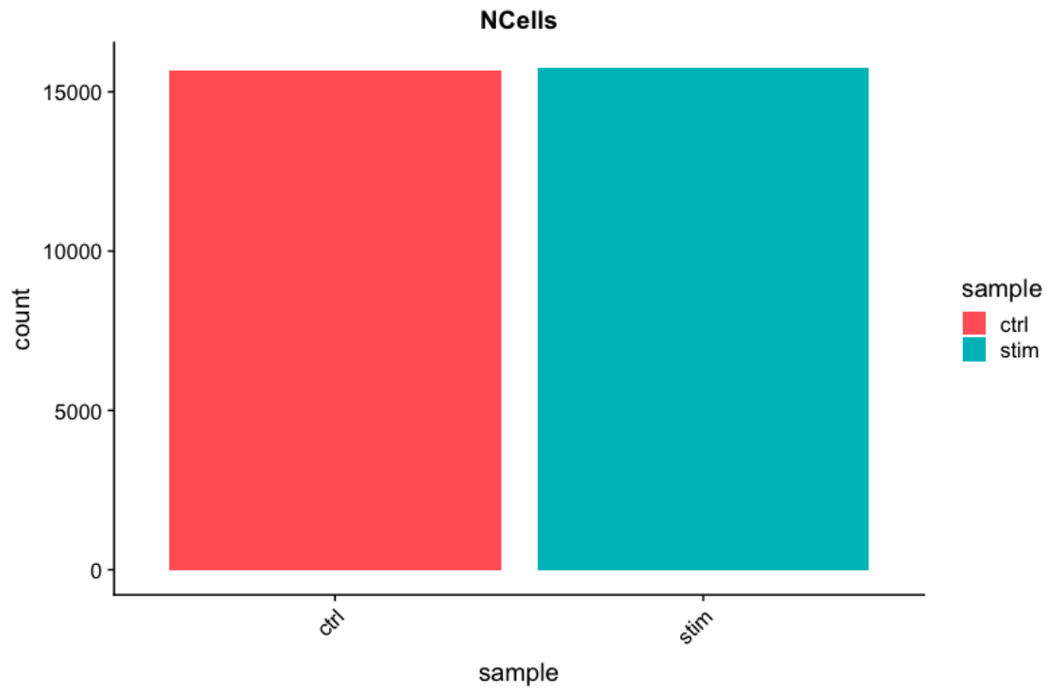
- To **identify any failed samples** and either try to salvage the data or remove from analysis, in addition to, trying to understand why the sample failed

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

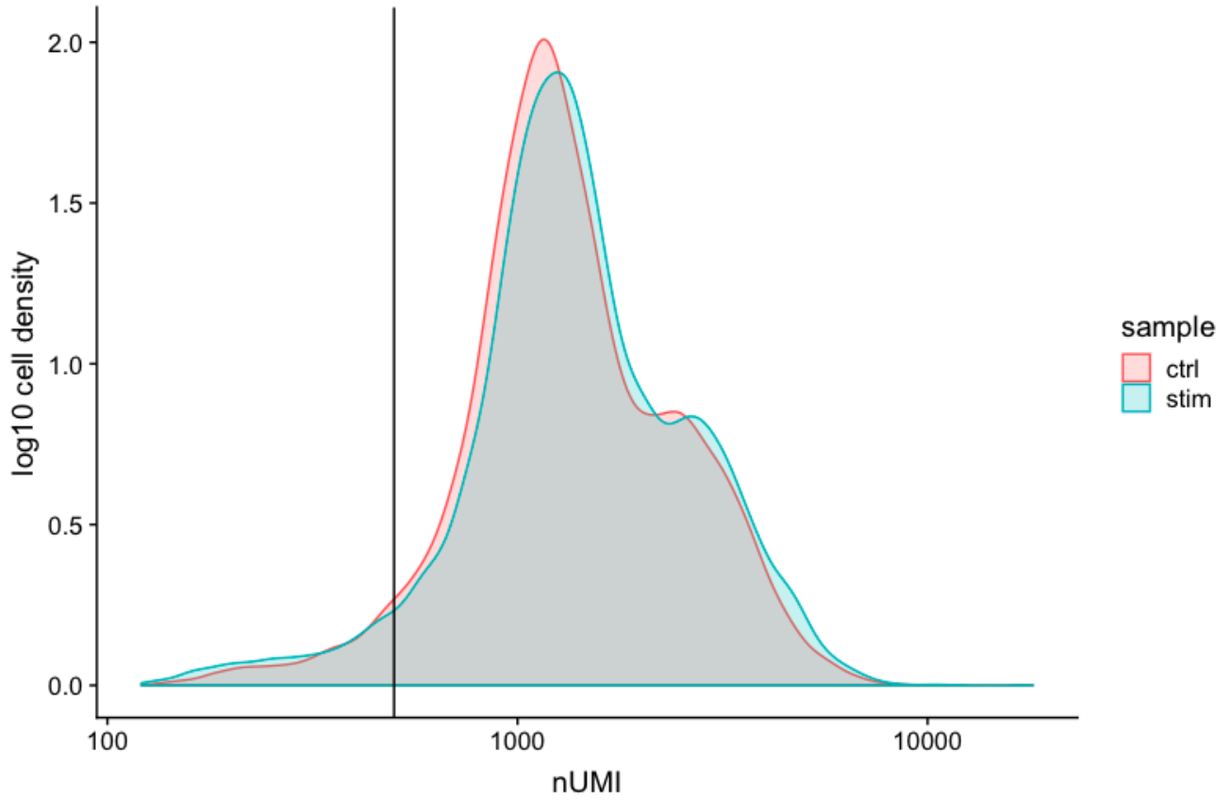
Quality metrics:
Cell counts per sample



In an ideal world, you would expect the number of unique cellular barcodes to correspond to the number of cells you loaded. *However, this is not the case.*

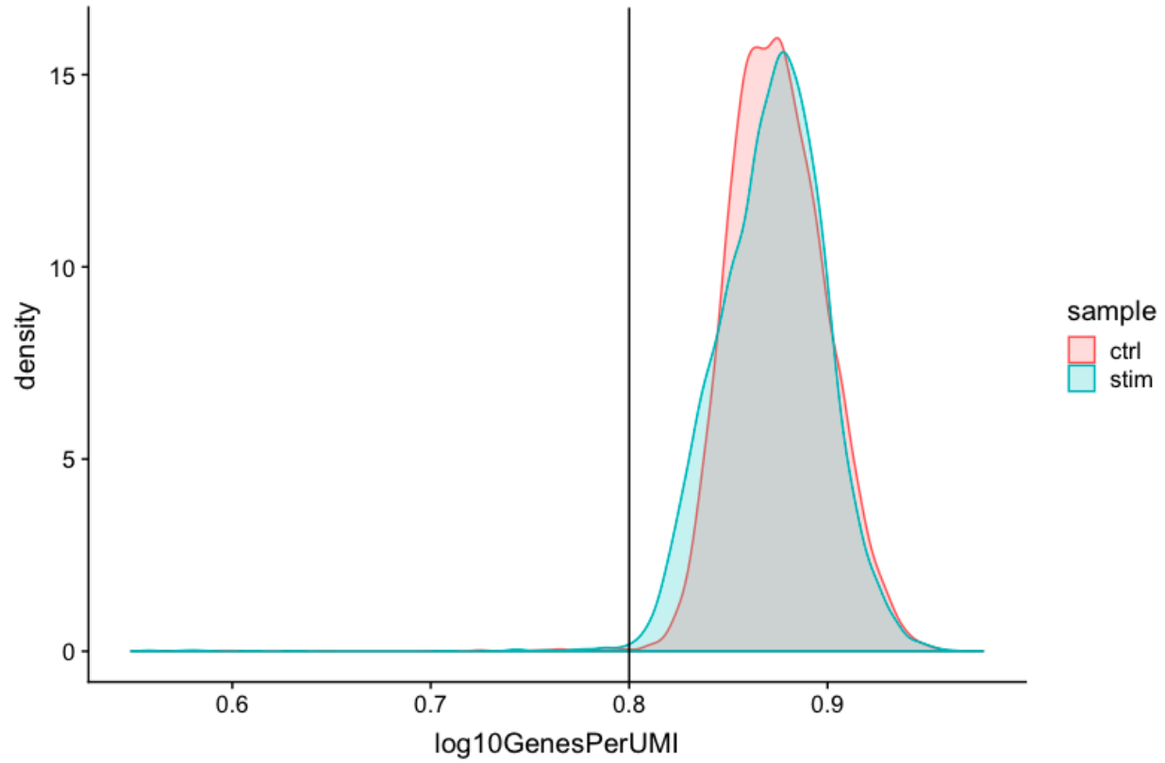
- Numbers of cells will appear to be higher if:
  - Barcoded beads in the droplet with no actual cell present
  - More than one barcode in the droplet
  - Dead or dying cells encapsulated into the droplet
- Number of cells (post-filtering) are expected to be lower due to capture rates.
  - inDrops cell capture efficiency is 70-80%
  - 10X cell capture efficiency is between 50-60%
  - The capture efficiency can be even lower if cell concentration used for library preparation was not accurate (i.e. use a hemocytometer to count cells)

Quality metrics:
UMI counts per cell



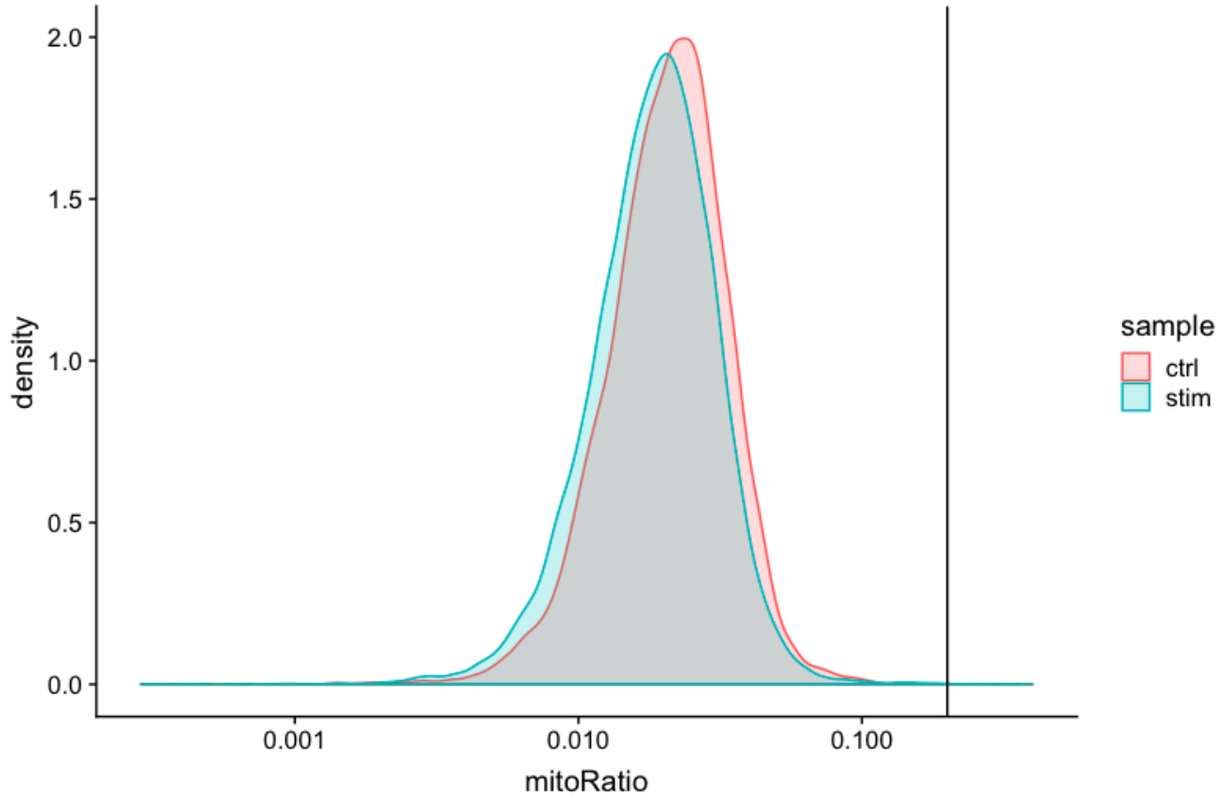
- 500 counts per cell is the low end of what is acceptable
- Counts between 500-100 is usable, but probably should have sequenced deeper

Quality metrics: Complexity



- Novelty score is computed to evaluate each cell in terms of how complex the RNA species are.
    - High values (> 0.8): indicates that for a given number of transcripts identified, there is an equally high number of genes detected. There is a reasonable amount of complexity in the transcriptome profile
    - Low values (< 0.7): indicates that for a given number of transcripts there is a low number of corresponding genes detected. A small set of genes were sequenced over and over again.

Quality metrics:
Mitochondrial counts ratio

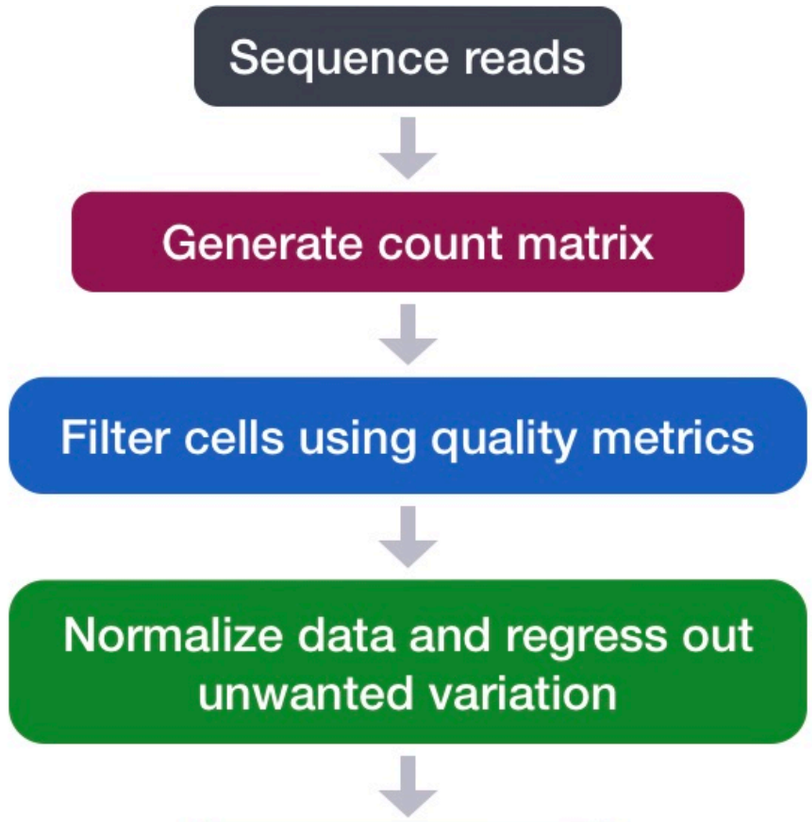


- High fractions of mitochondrial reads can indicate presence of dead or dying cells
- Typically, cells with higher than 20% of reads mapping to mitochondrial genes are filtered out

Recommendations

- Cell-level filtering
  - Be permissive as possible with filtering thresholds
  - Evaluate joint effects, as considering any of these QC metrics in isolation can lead to misinterpretation of cellular signals
  - Have a good idea of your expectations for the cell types to be present prior to performing the QC.
- Gene-level filtering
  - There will be many genes with zero counts. These can dramatically reduce the average expression for a cell.
  - If a gene is only expressed in a handful of cells, it is not all that meaningful. Removing these genes will effectively remove genes with zero counts in all cells, too.

Normalizing and removing unwanted variation



Goals:

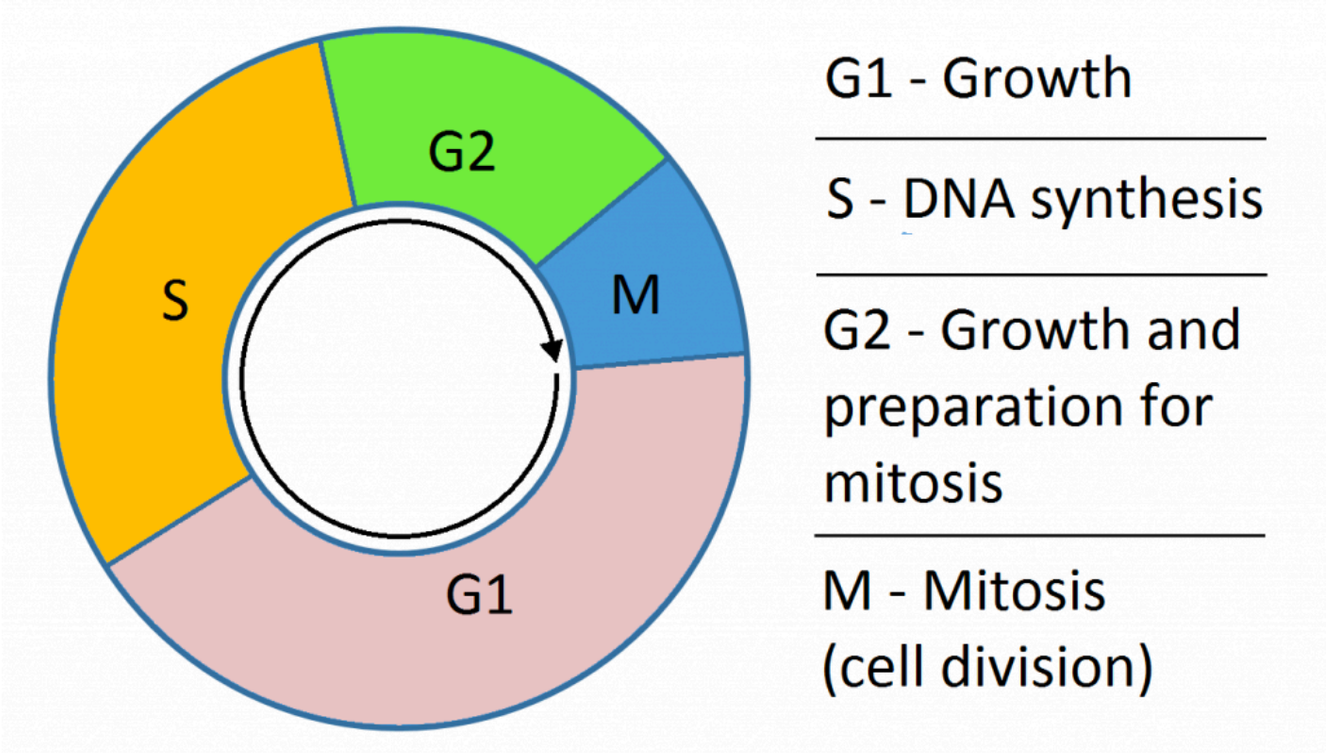
- To identify sources of **unwanted variation** in the data
- To accurately **normalize and scale the gene expression values** to account for differences in sequencing depth and overdispersed count values.
- To identify **the most variant genes** most likely to be indicative of the different cell types present.

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

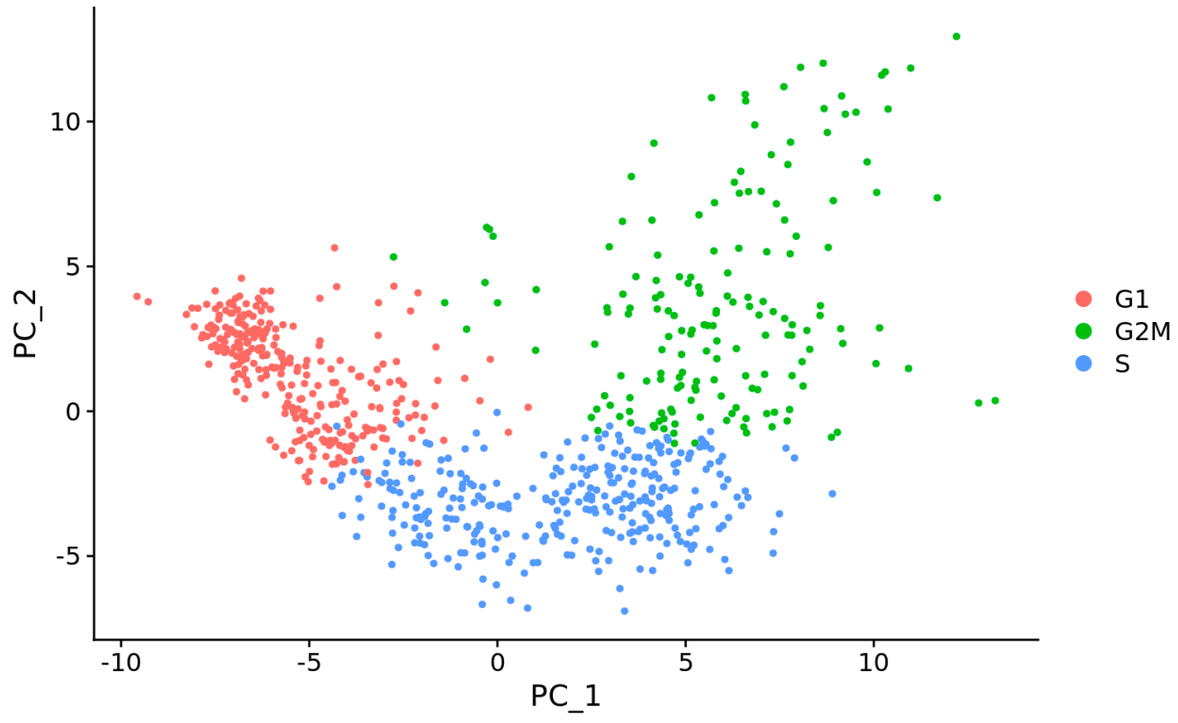


- Cell cycle heterogeneity can drive the changes in expression, and subsequently the way in which cells cluster together.

- For each cell, compute cell cycle phase scores based on the expression of canonical markers.

Exploring sources of unwanted variation:

Cell cycle effects

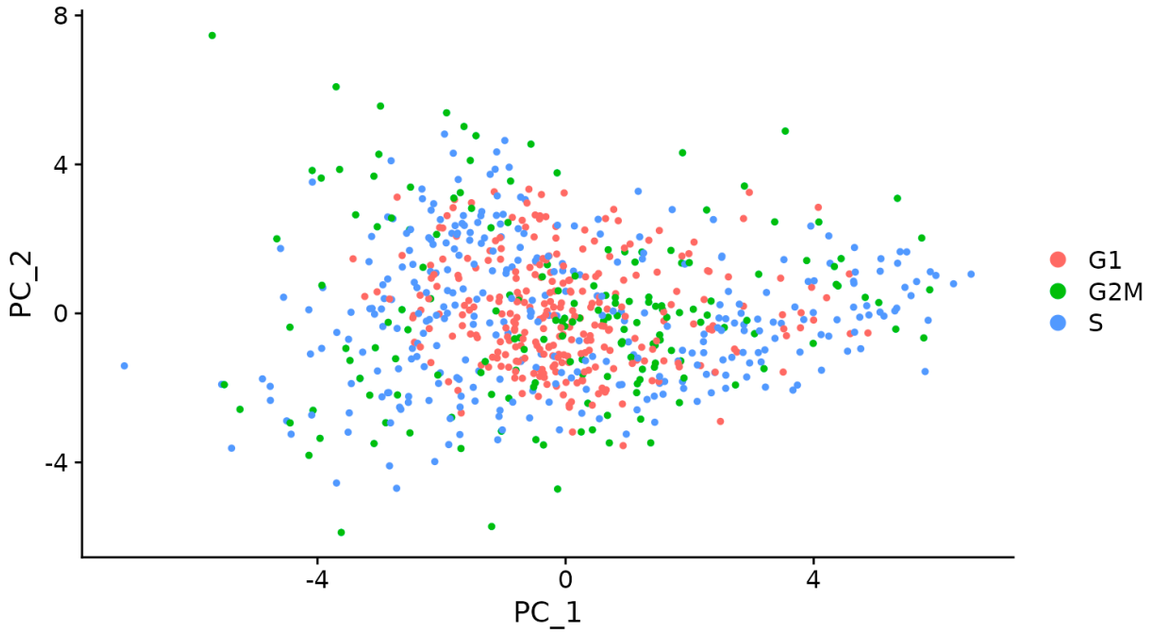


- Scores are used to classify cells into one of three phases
- Plot PCA and color cells by phase
- Do we see separation of cells by cell cycle phase?

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

Exploring sources of unwanted variation:

Cell cycle effects regressed out

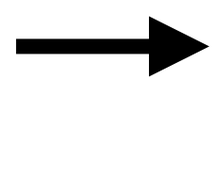


Now, cell-cycle heterogeneity does not contribute to PCA or downstream analysis.

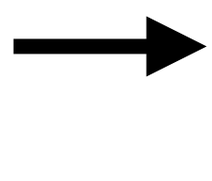
Image source: https://satijalab.org/seurat/articles/cell_cycle_vignette.html

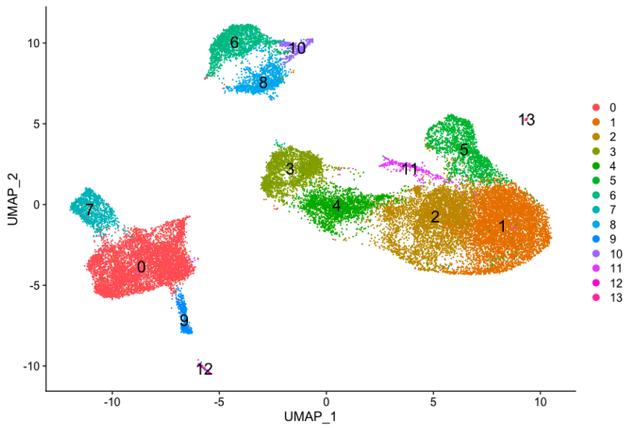
Normalization

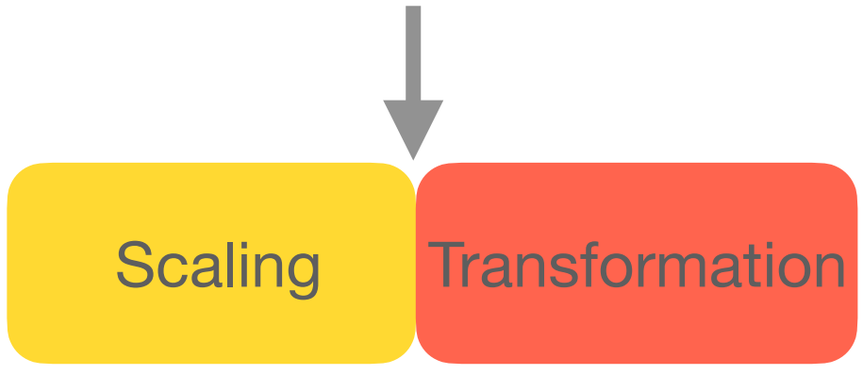
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



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)

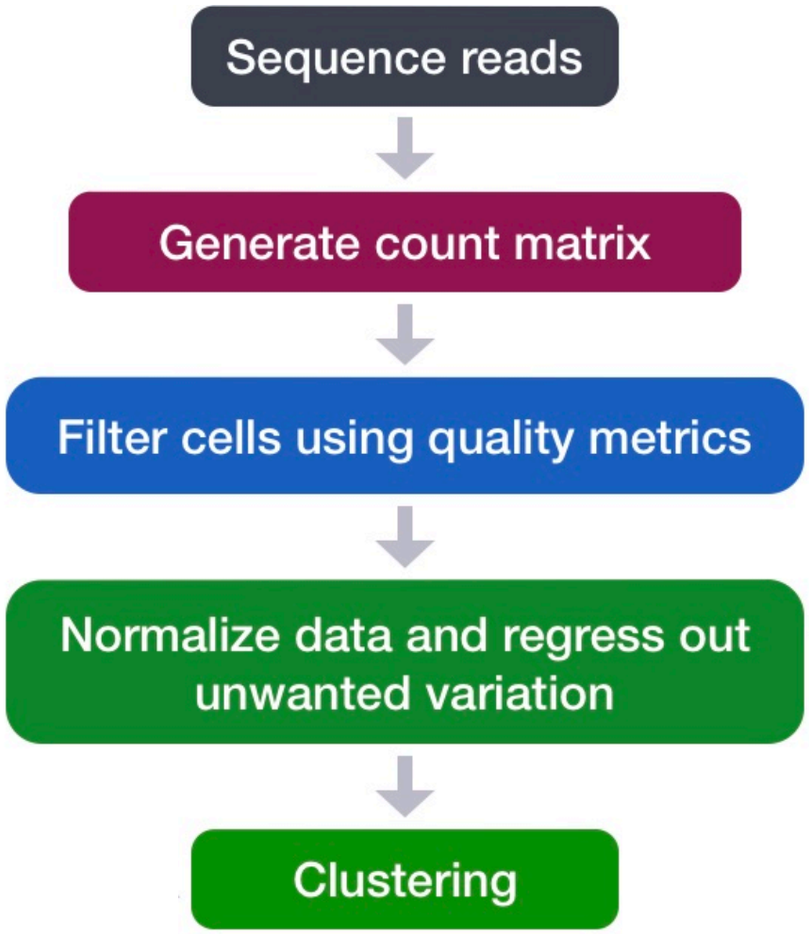
  - The output (residuals) is the normalized expression levels for each transcript tested

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.
- 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

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

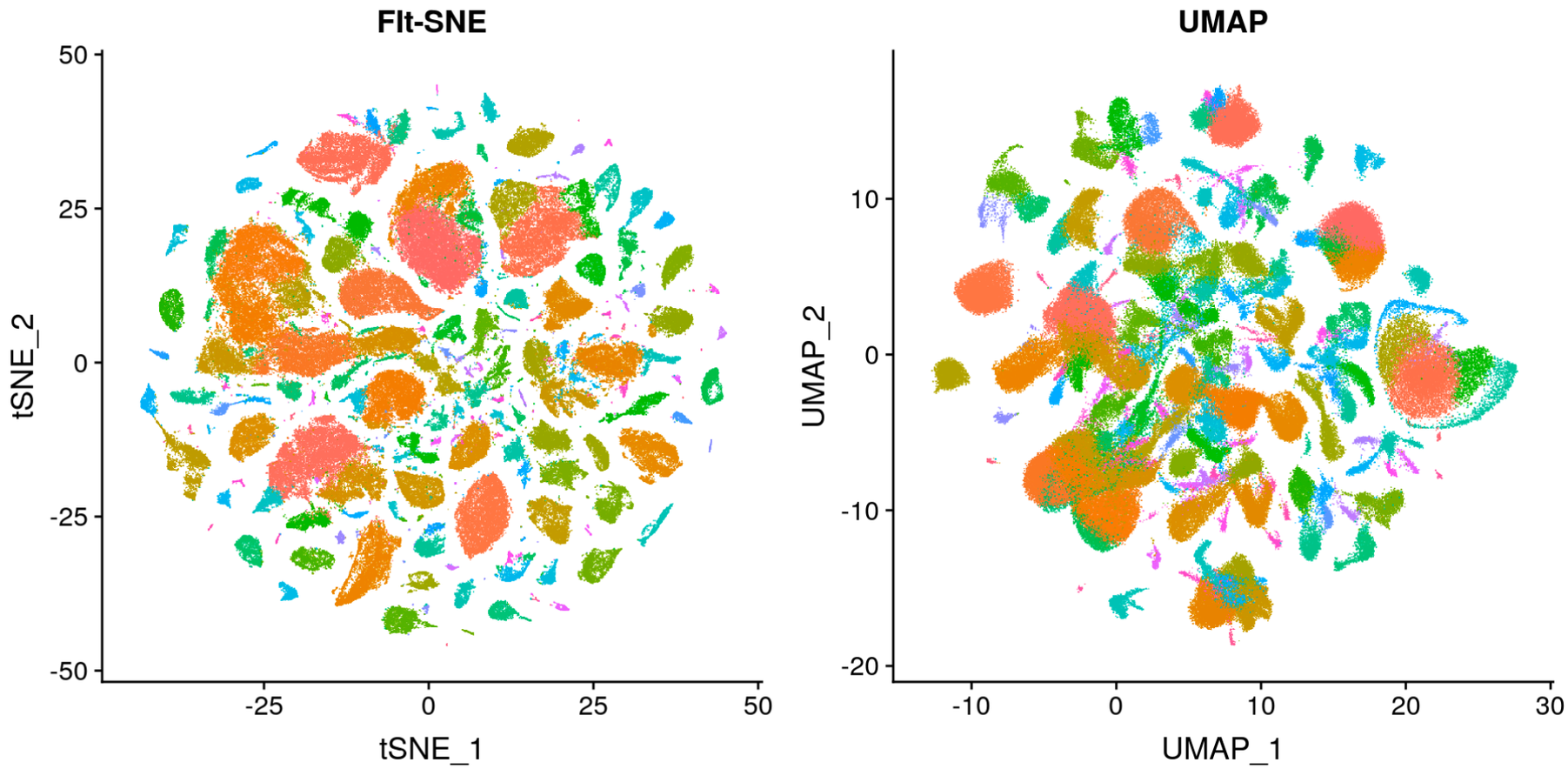
Challenges:

- **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

Visualizing the data

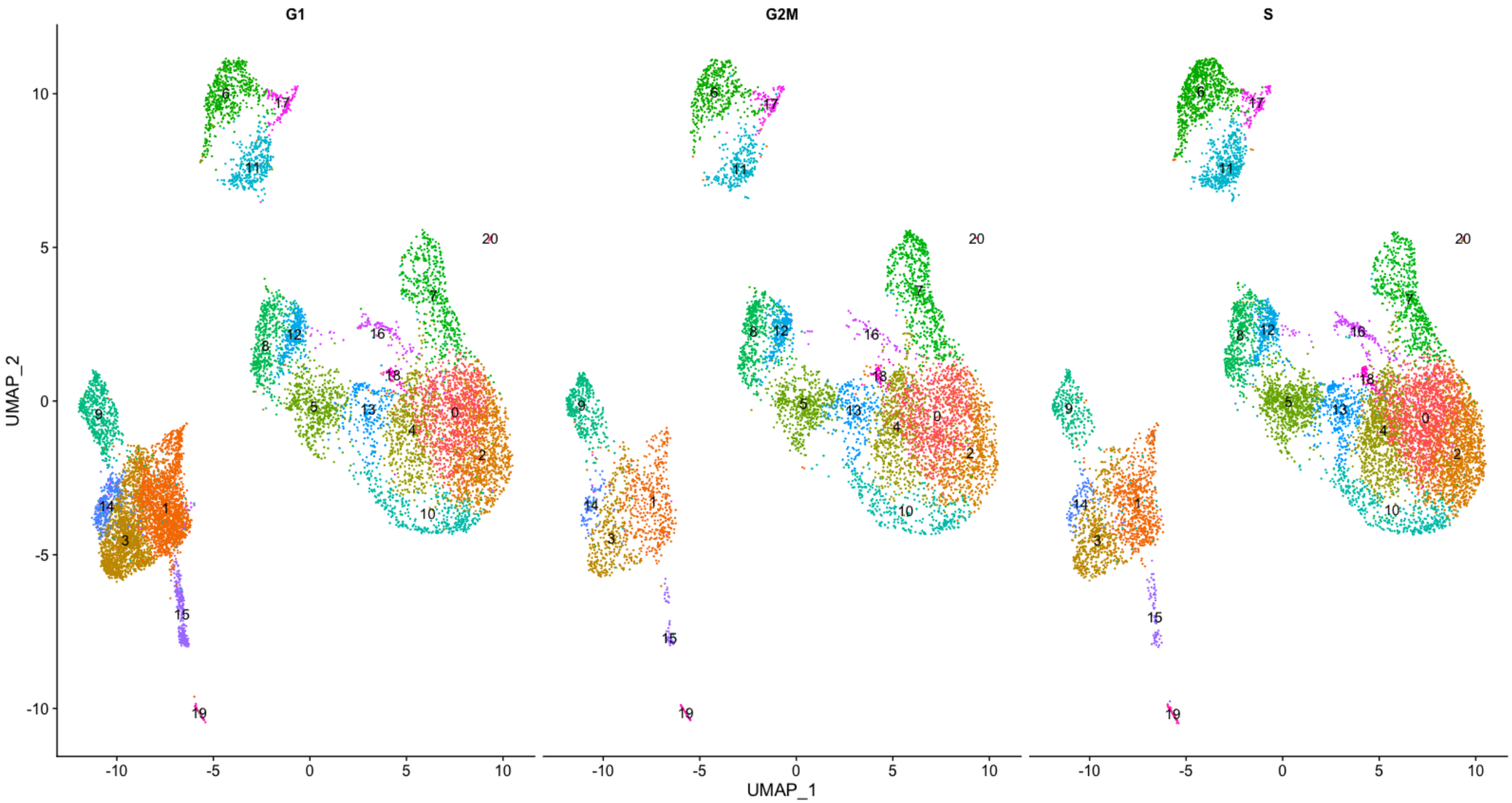


- Take the top principal components to arrange cells in multi-dimensional space.
- **UMAP and t-SNE are dimension reduction techniques** designed to preserve local structure (group neighboring points together).
- Neither approach guarantees that the relatedness between clusters of cells are totally preserved, but UMAP has been shown to have increased preservation of global structure.

Image source: https://satijalab.org/seurat/archive/v3.0/mca.html

Cluster quality control

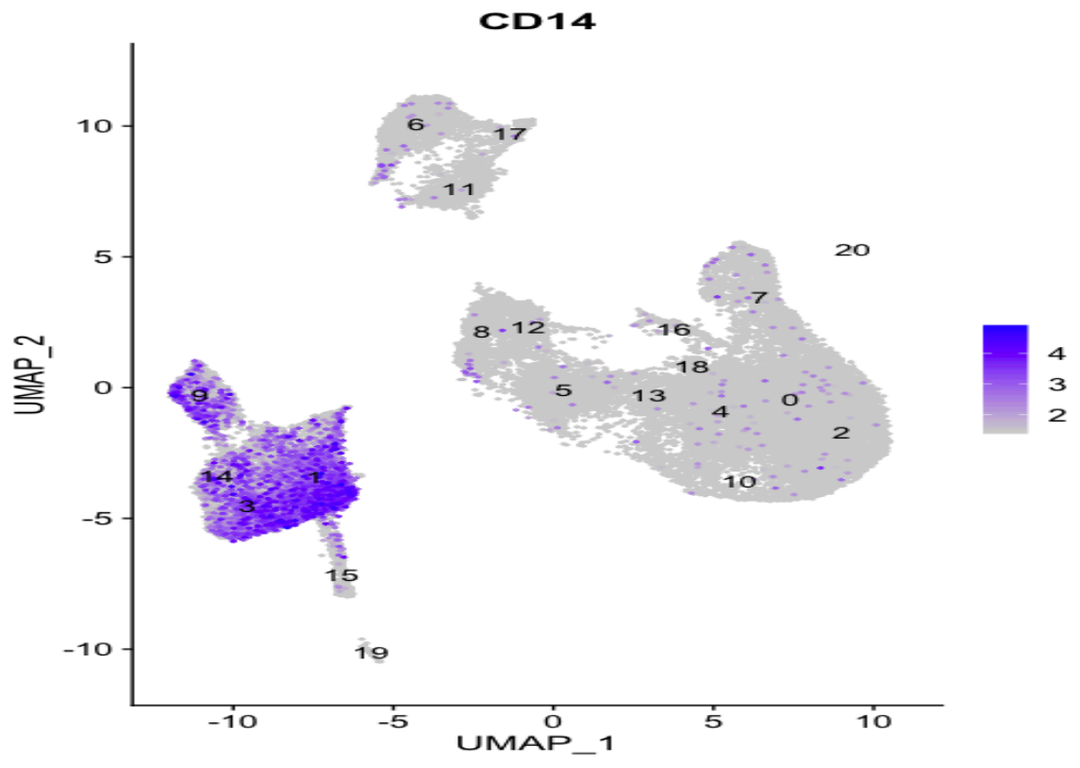
Segregation of clusters by cell cycle phase

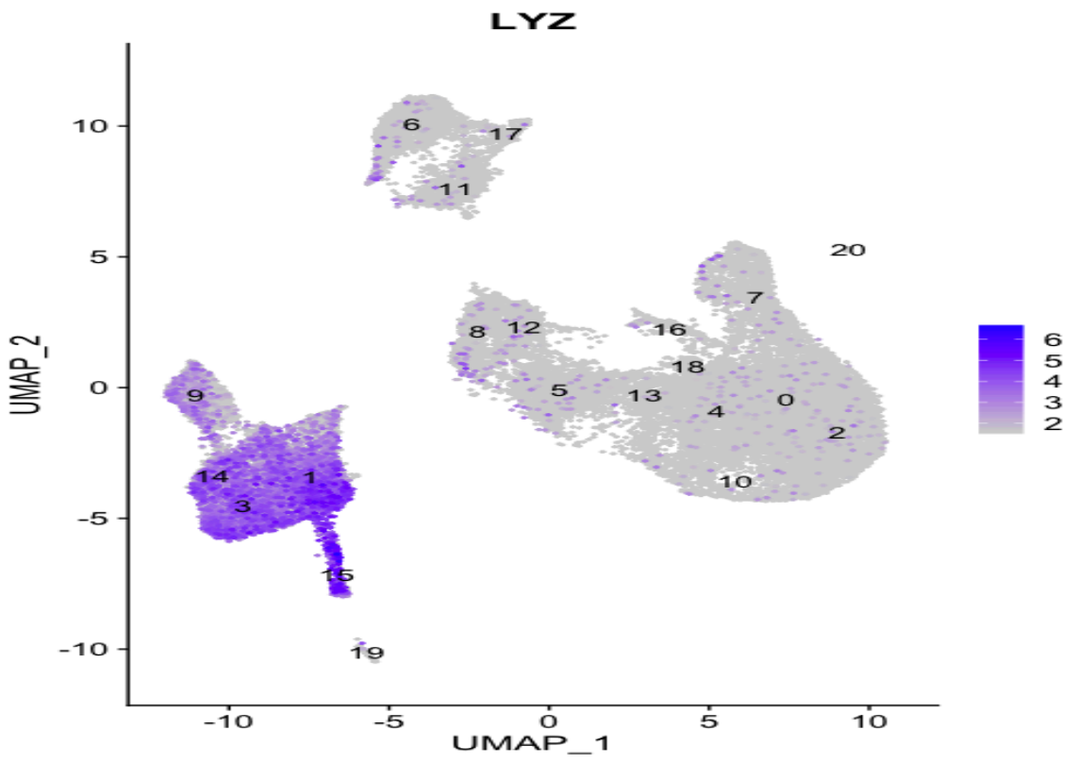


Cluster quality control

Exploring known cell type markers

CD14+ monocyte markers

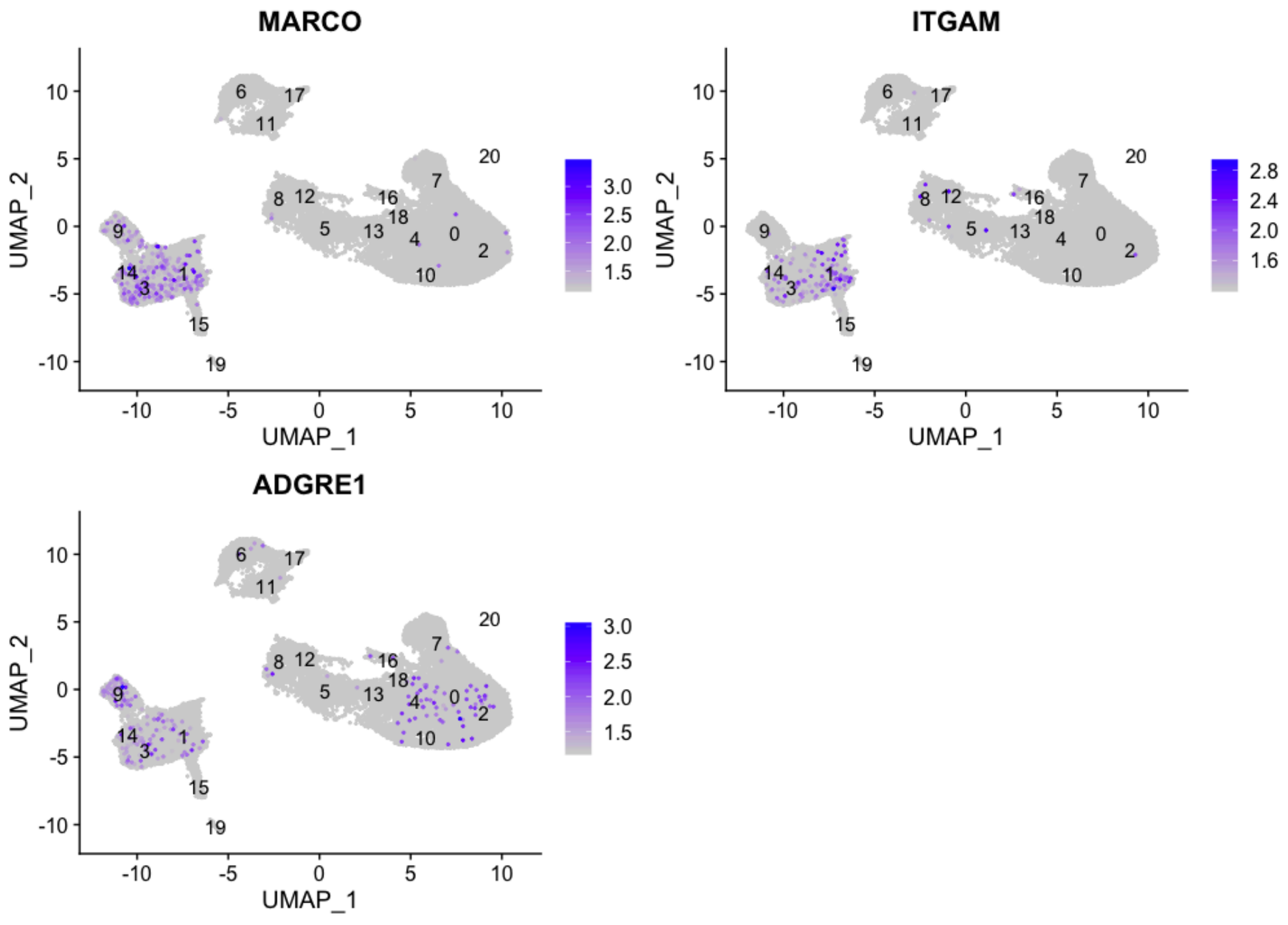




Cluster quality control

Exploring known cell type markers

Macrophages?



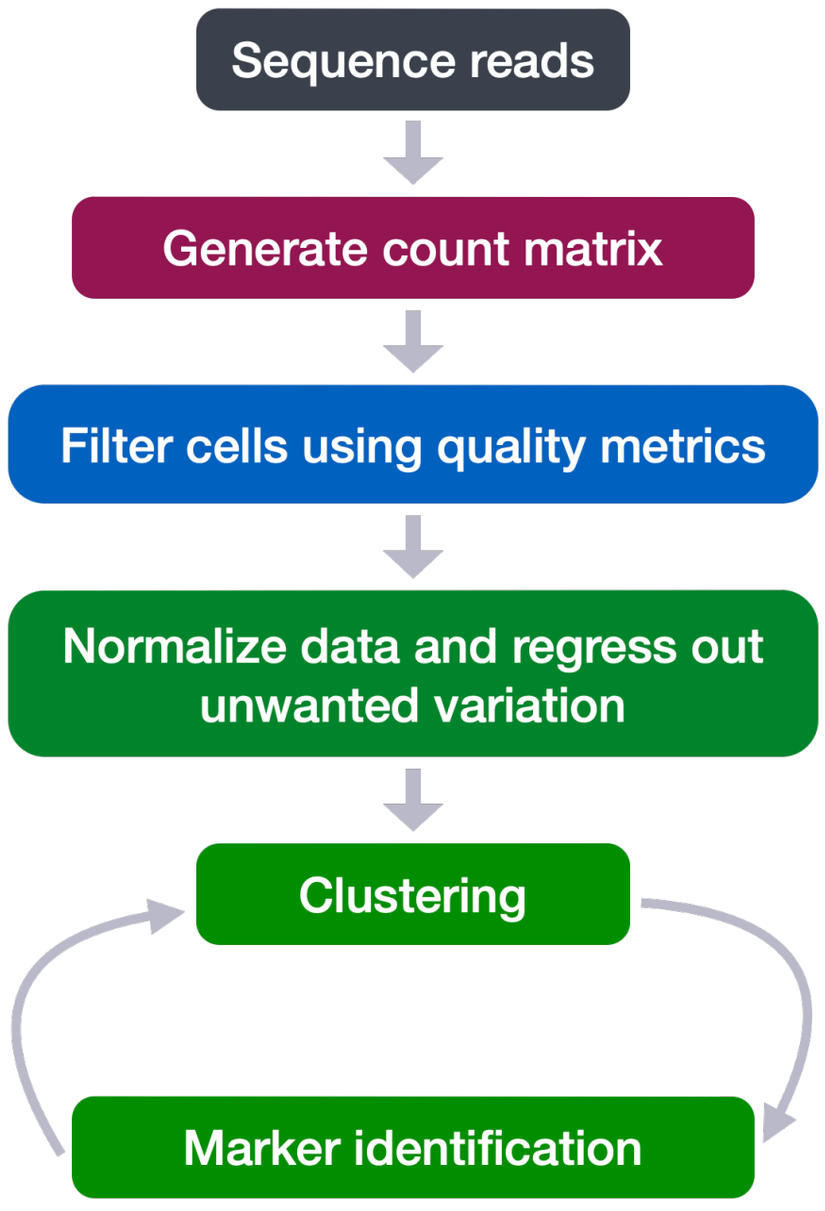
Recommendations

- 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



Methods for marker identification

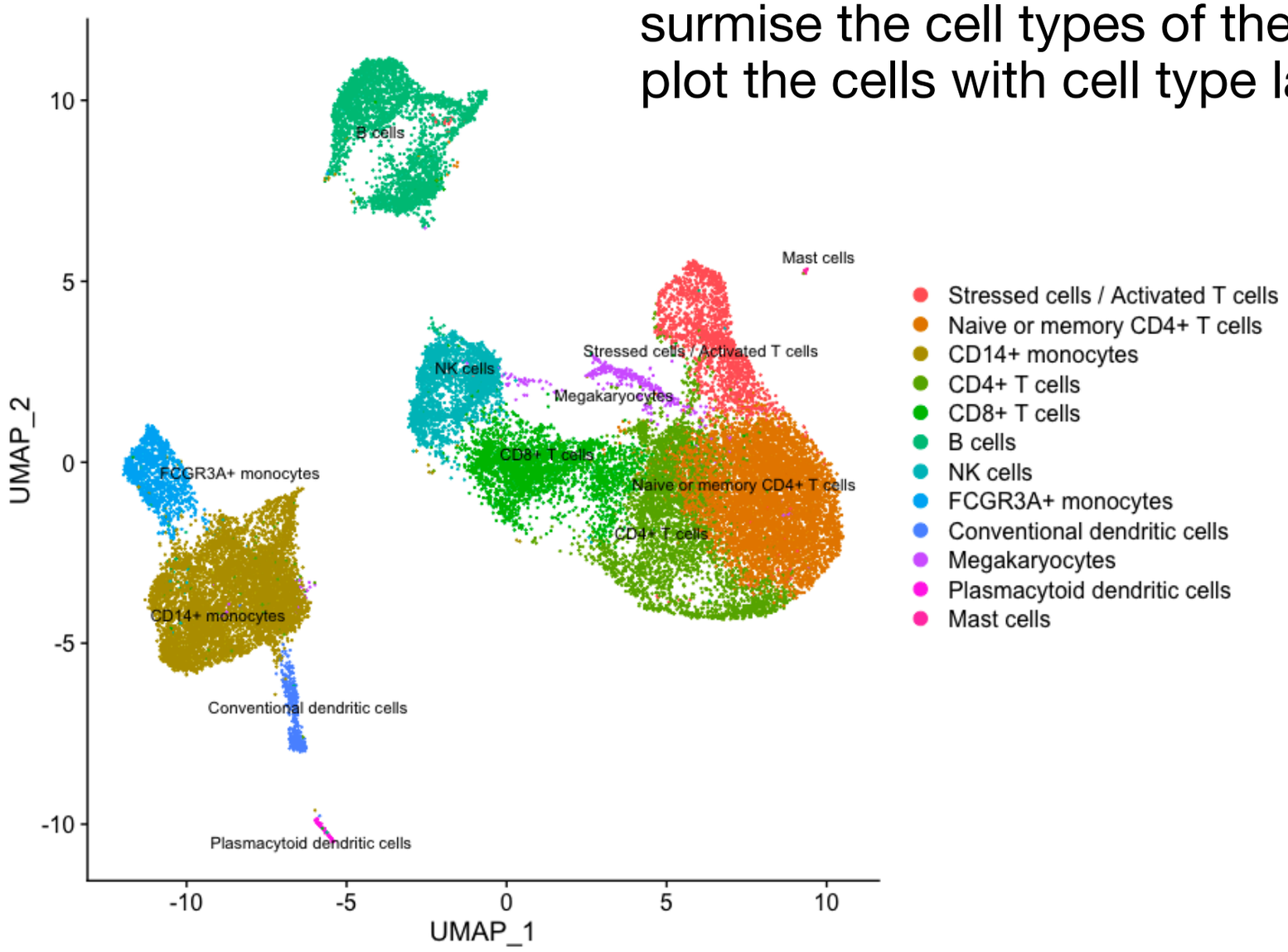
1. **Identification of all markers for each cluster:** this analysis compares each cluster against all others and outputs the genes that are differentially expressed/present.
    - Useful for identifying unknown clusters and improving confidence in hypothesized cell types.

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.



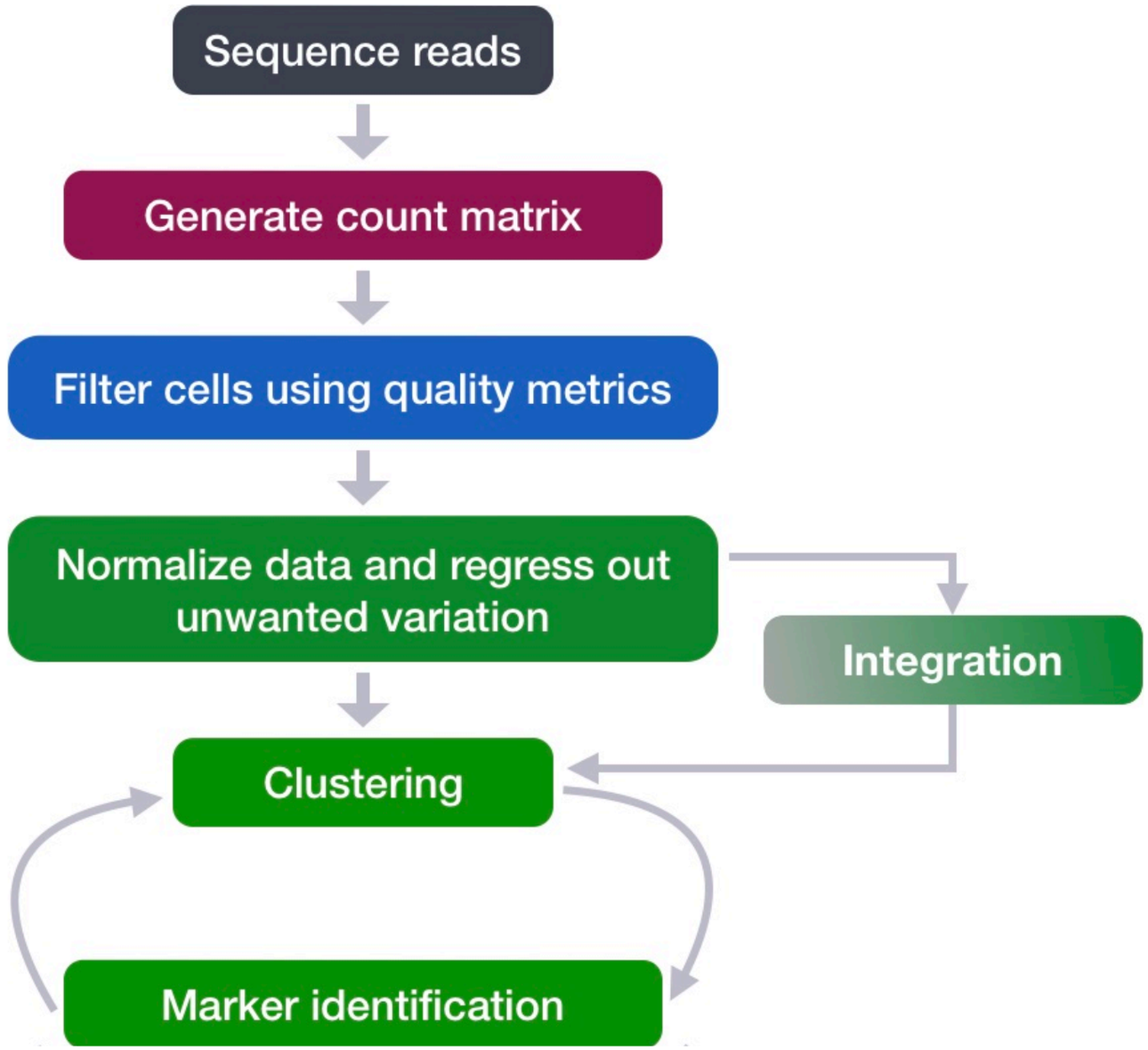
Recommendations

- Inflated p-values can lead to over-interpretation of results (essentially each cell is used as a replicate). Top markers are most trustworthy.

- Test out different levels of stringency to compare and contrast marker gene lists.

- Think of the results as hypotheses that need verification. Follow up with bench validation.

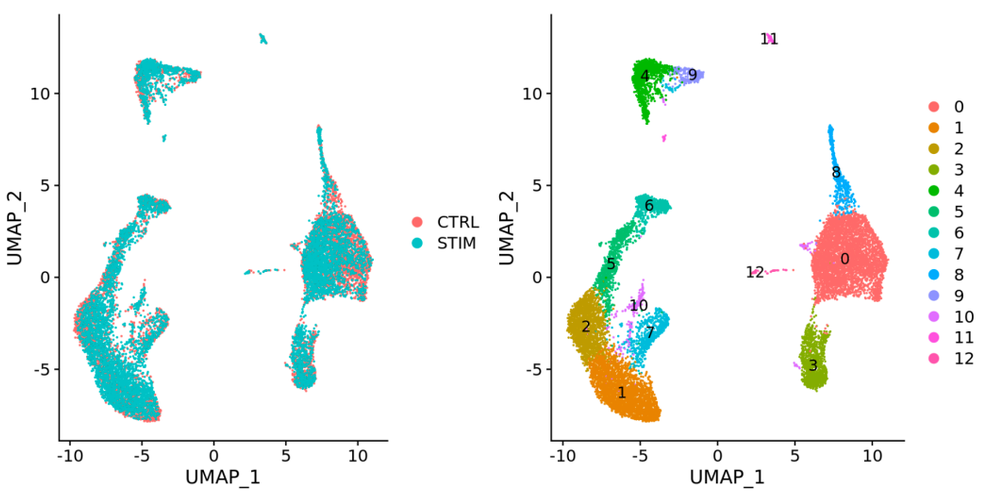
Working with multiple samples



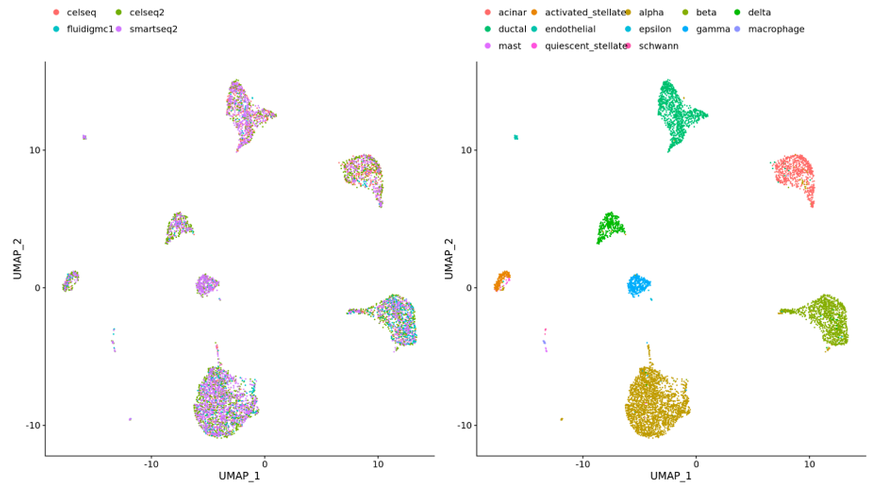
Integration

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”.

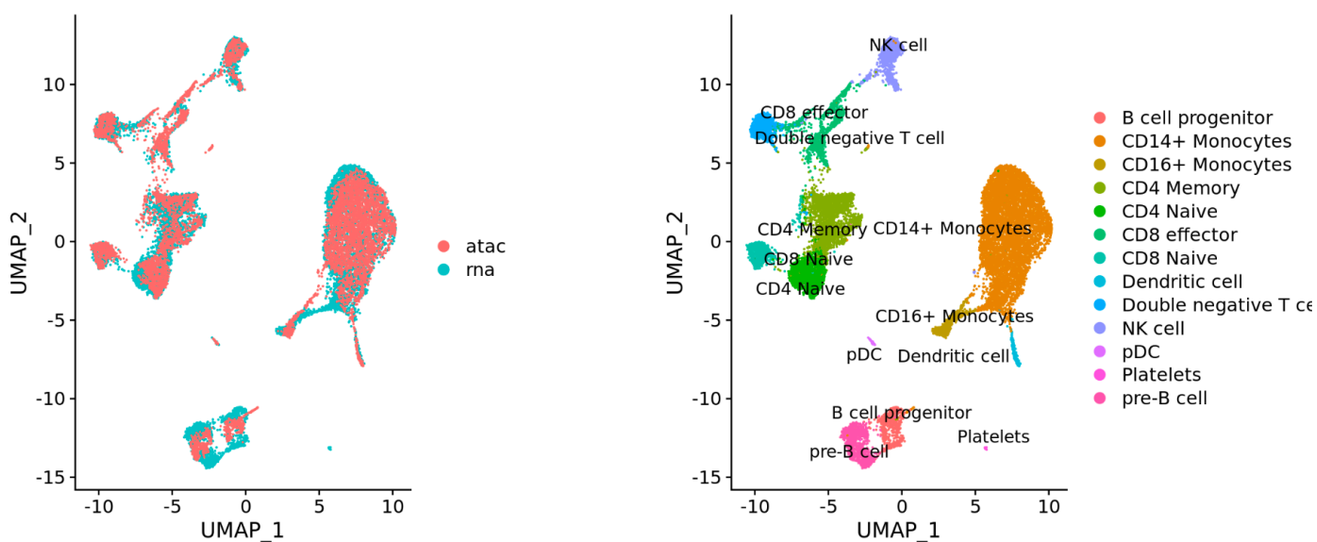
Different conditions



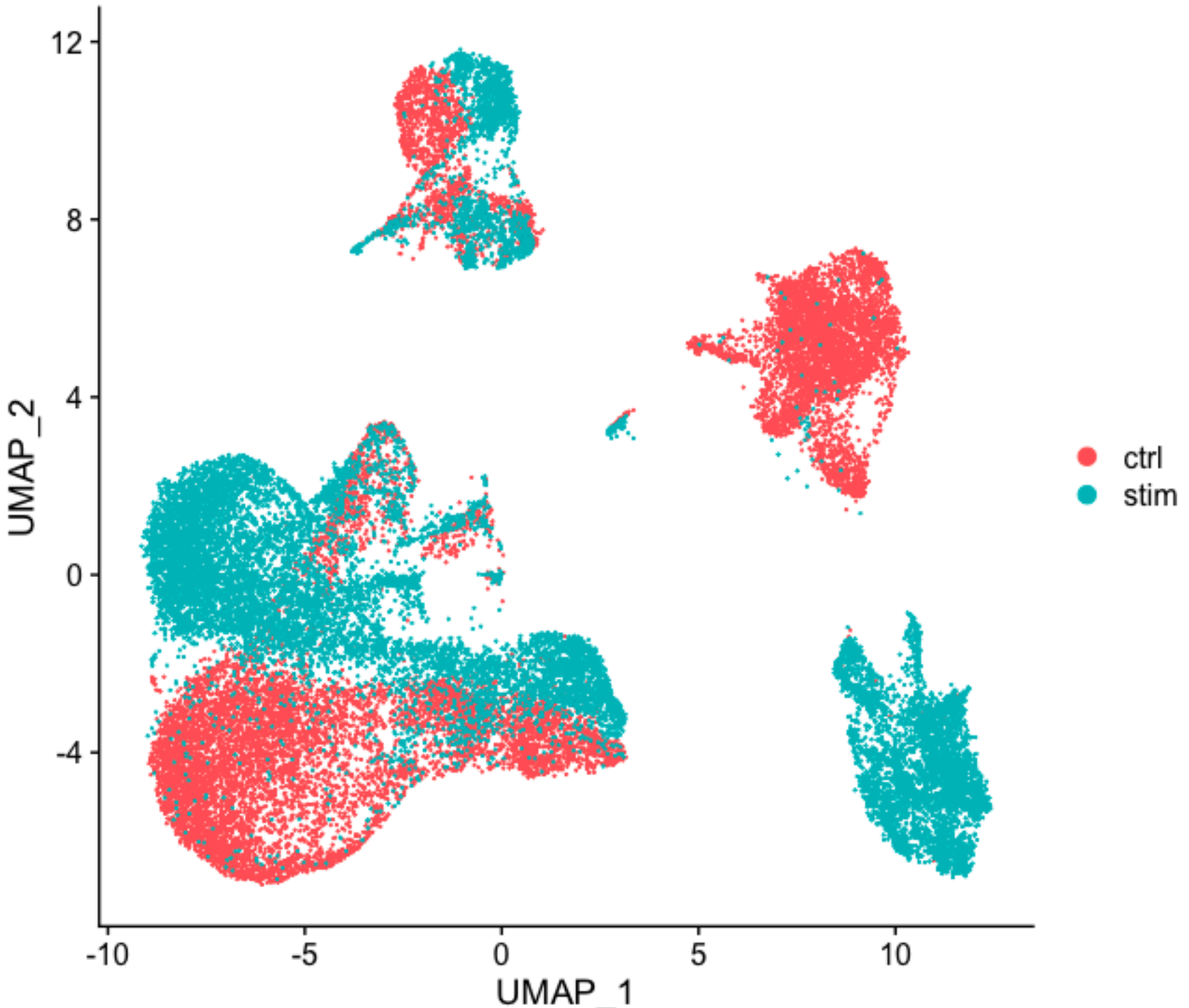
Different datasets



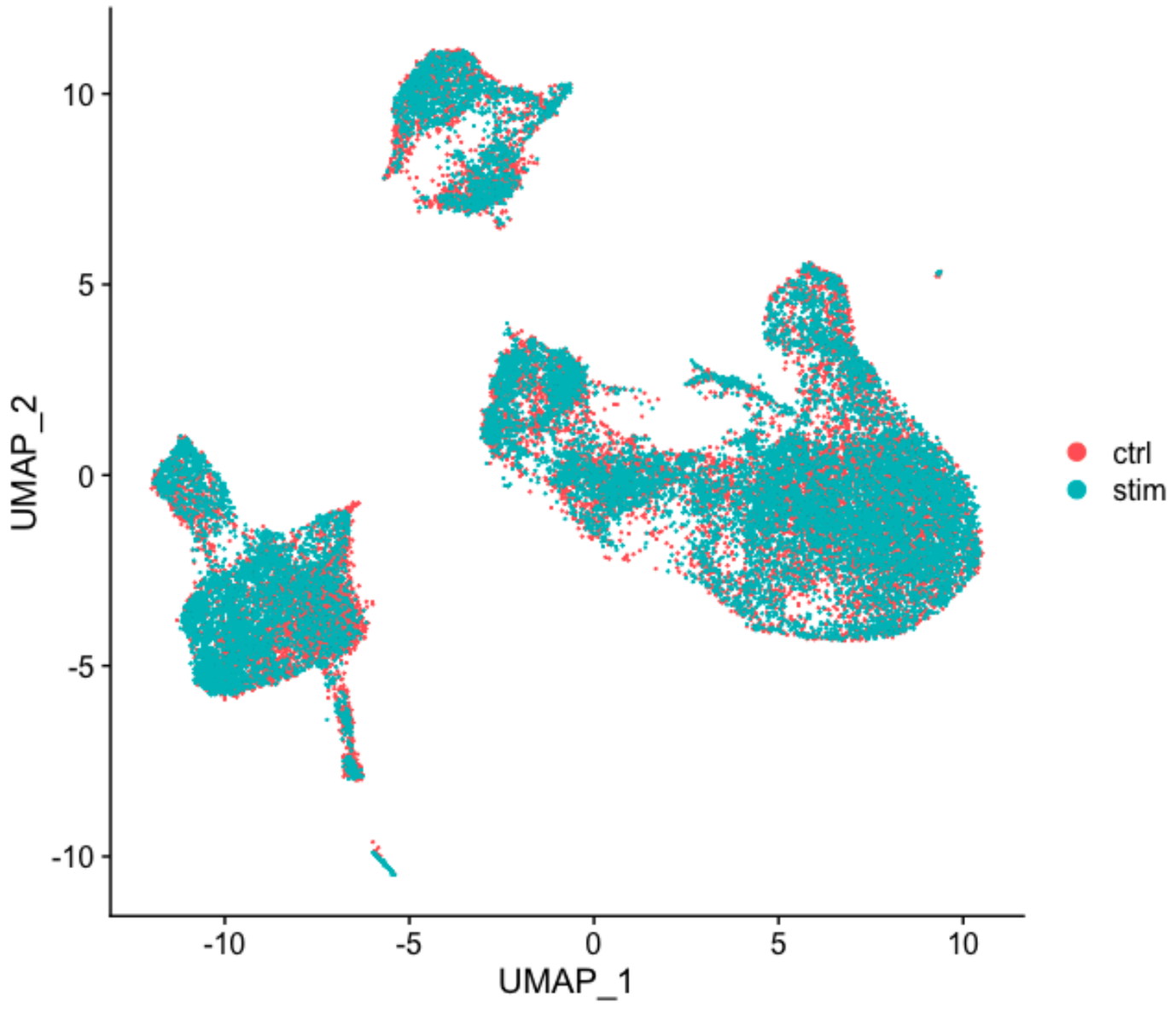
Different modalities



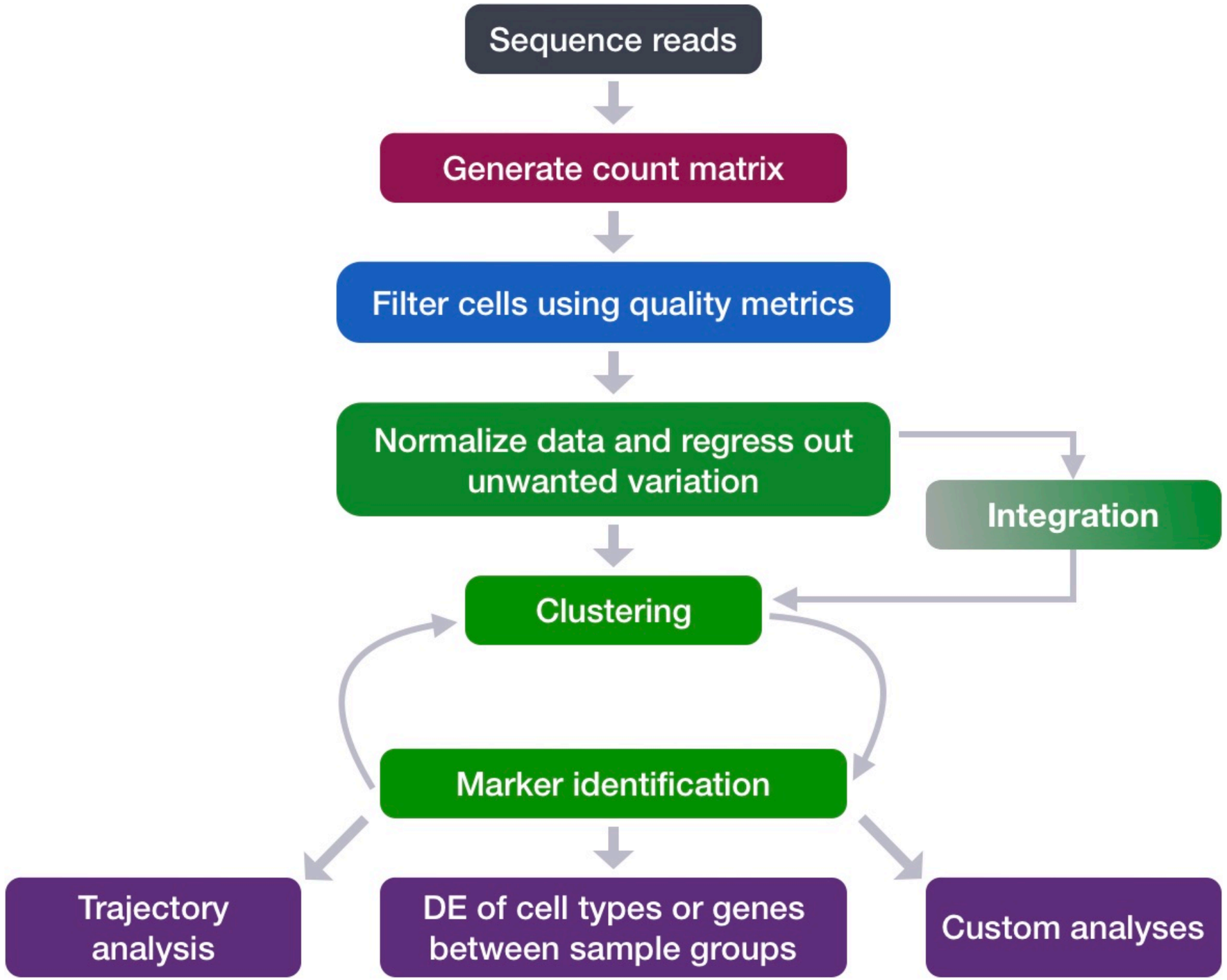
To integrate or not to integrate?



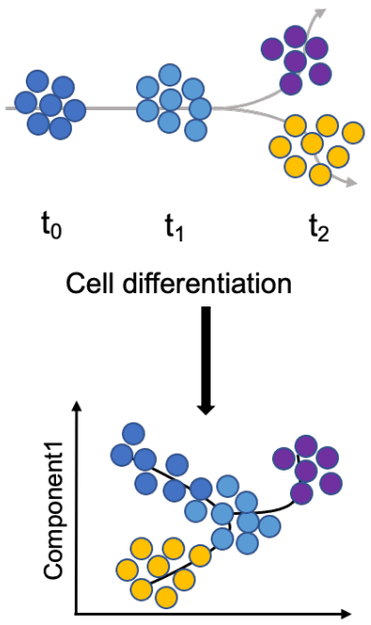
Clusters will now be a representation of cells from both conditions allowing for more interpretable results downstream.



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

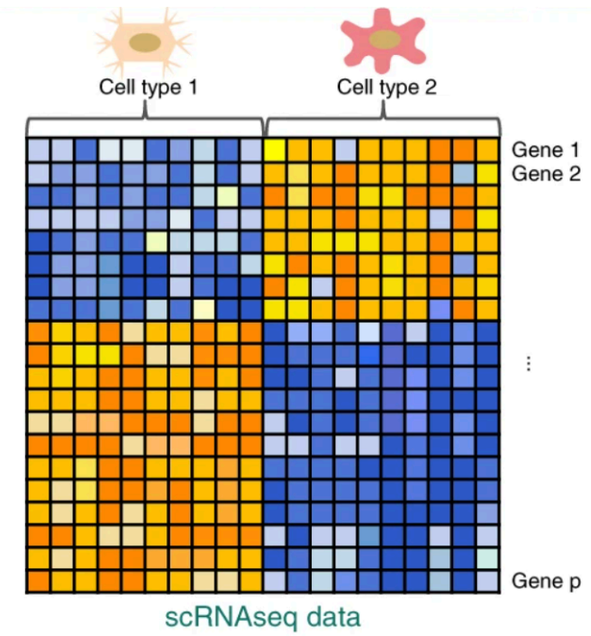
Trajectory analysis



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

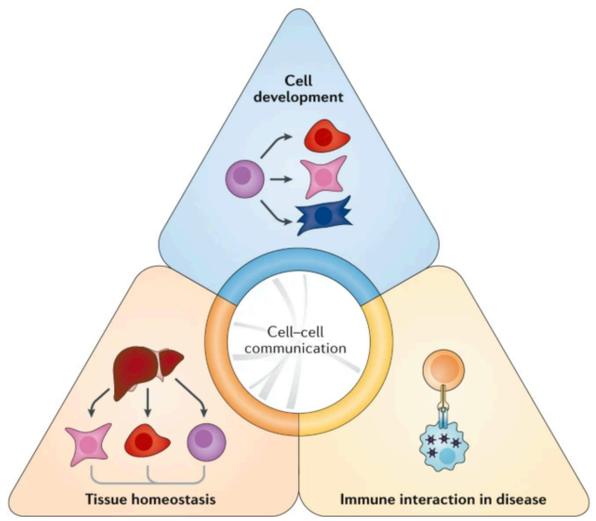
DE of cell types or genes between sample groups



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