Single-cell RNA-Seq Bioinformatics

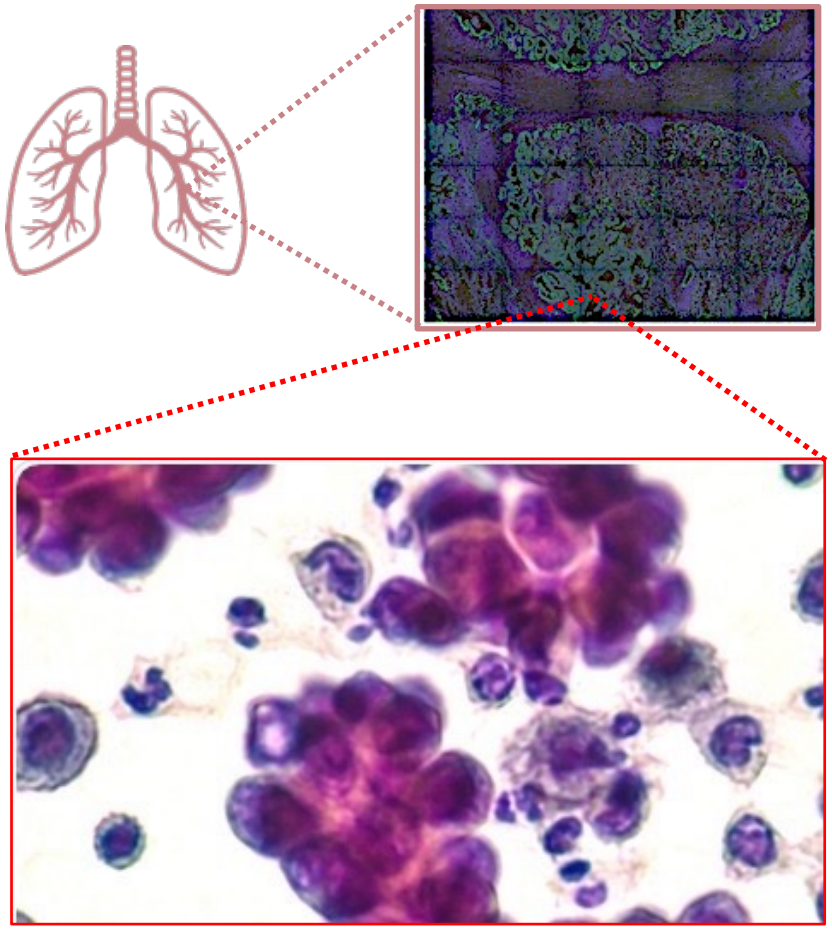
Day 1: Quality Control, Clustering and Integration

DISC Single-cell series Oct-Nov 2023

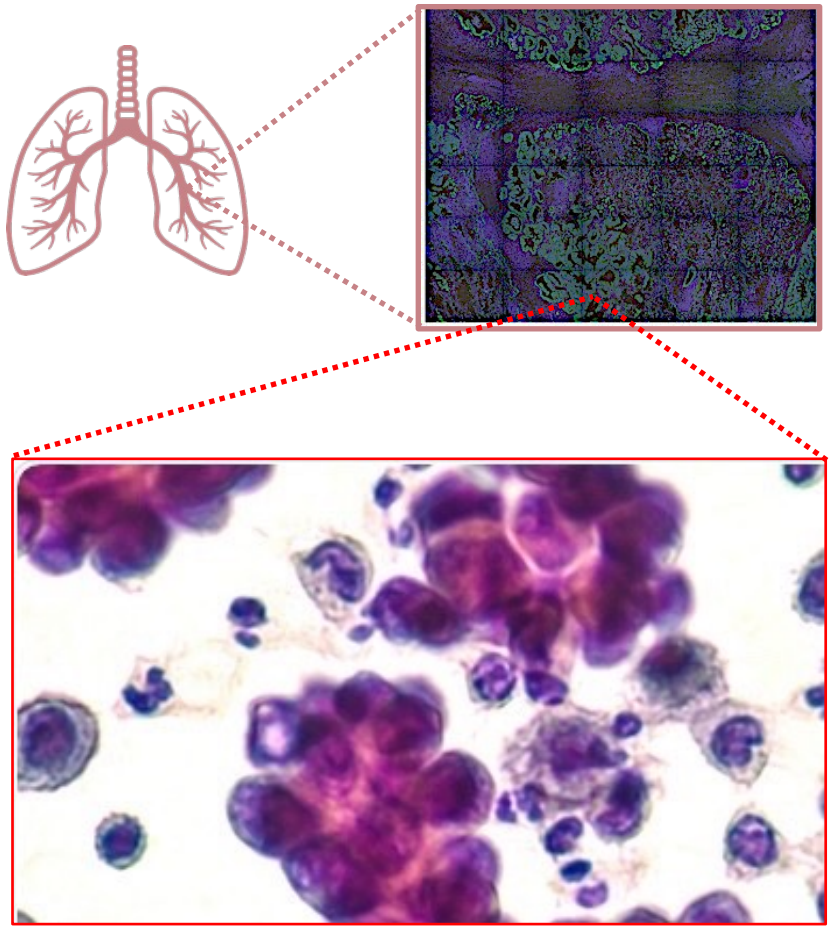
Eric Reed, Data Scientist, Data Intensive Studies Center: Eric.Reed@tufts.edu

Rebecca Batorsky, Data Scientist, Data Intensive Studies Center**:** Rebecca.Batorsky@tufts.edu

What is transcriptomics?



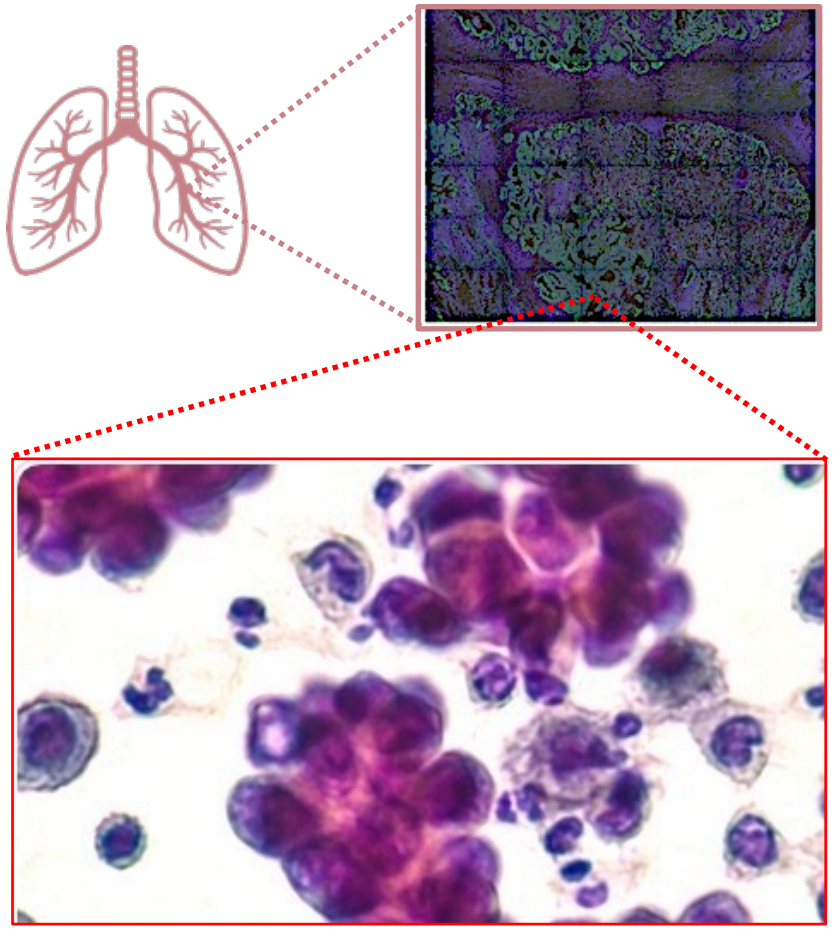
What is transcriptomics?



What are these cells?

What are these cells doing?

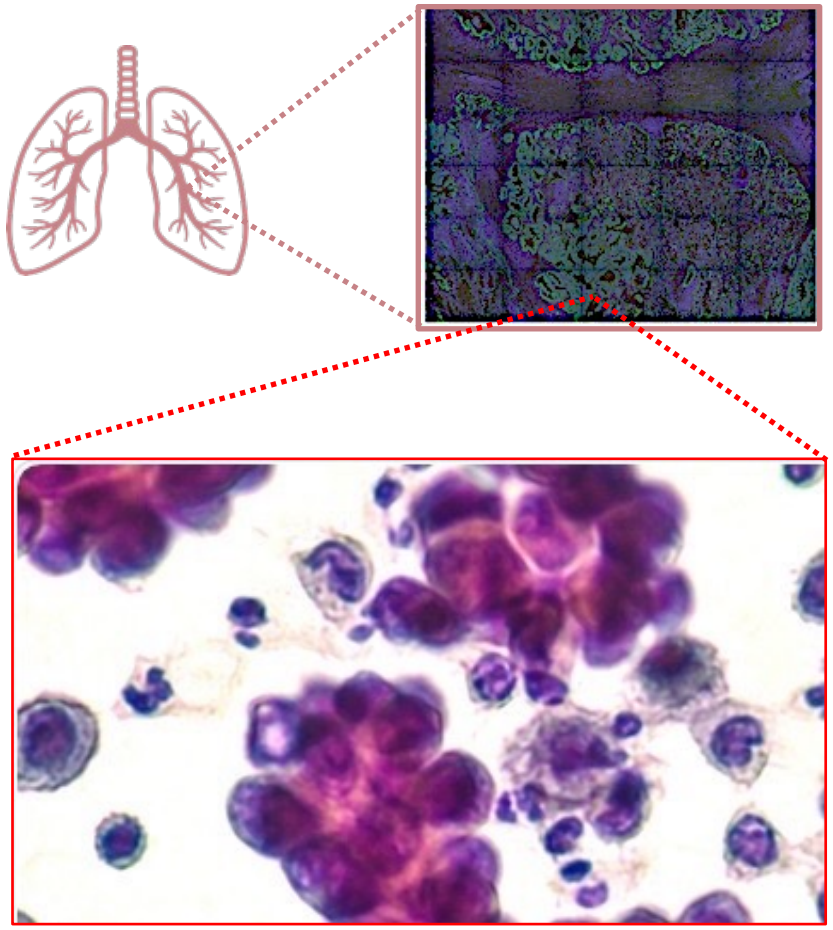
What is transcriptomics?

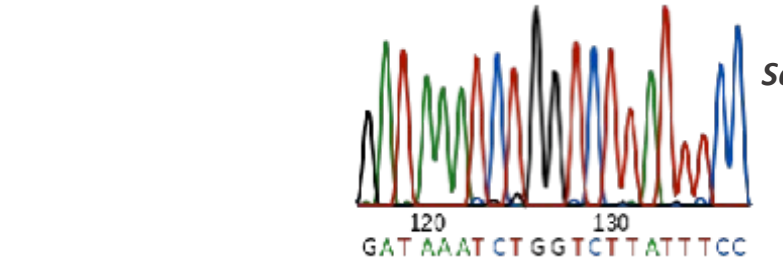


What are these cells?

What are these cells doing?

What is transcriptomics?



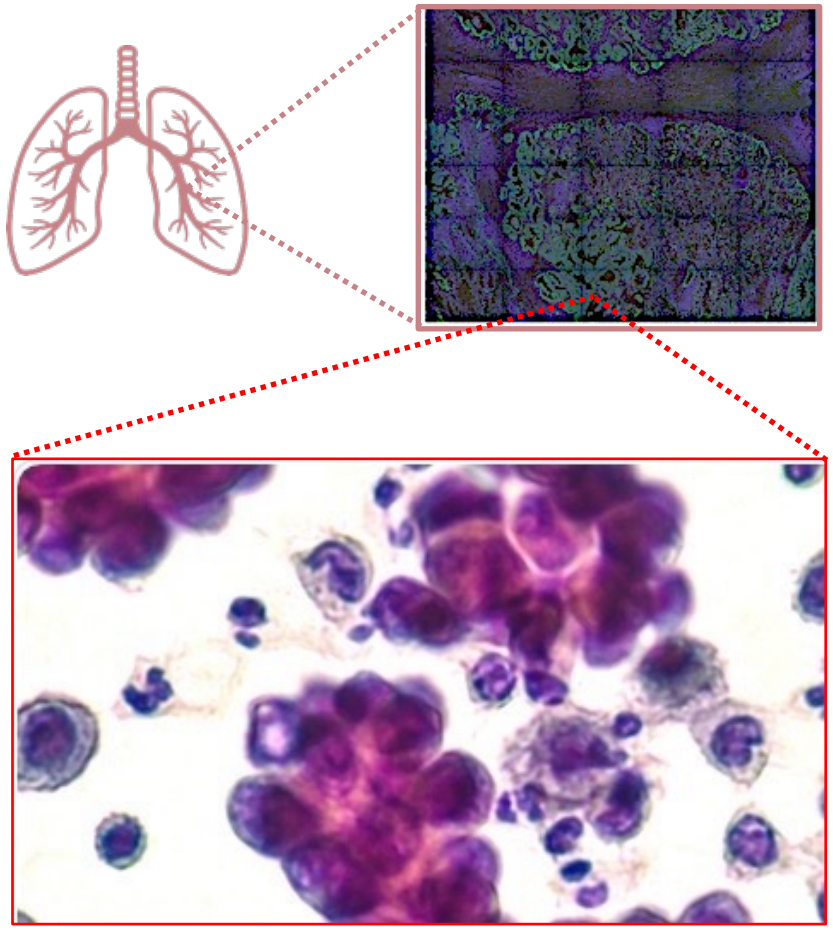


Sequencing

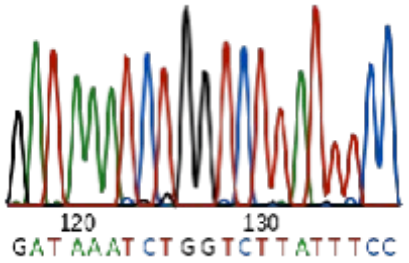
What are these cells?

What are these cells doing?

What is transcriptomics?



What are these cells?
What are these cells doing?

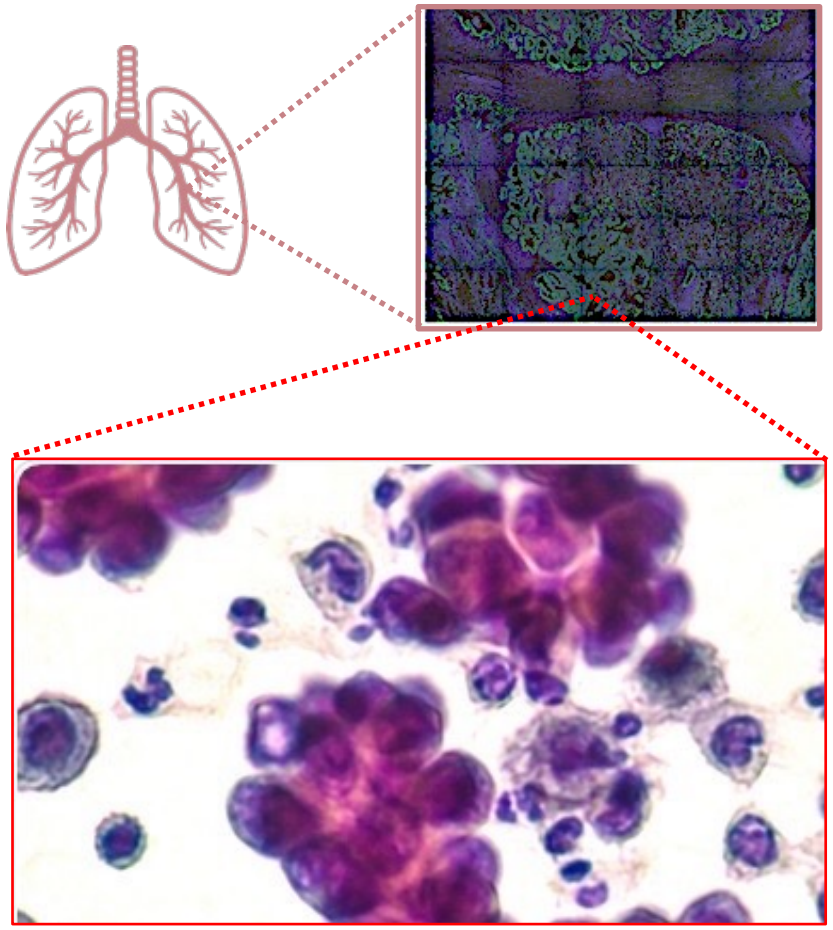


Sequencing

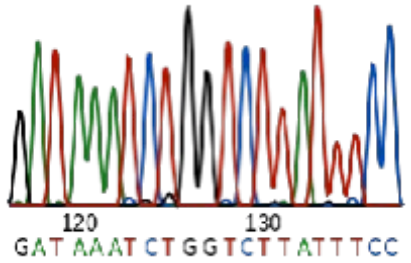
DNA → Genome

~ 1 Genome

What is transcriptomics?



What are these cells?
What are these cells doing?



Sequencing

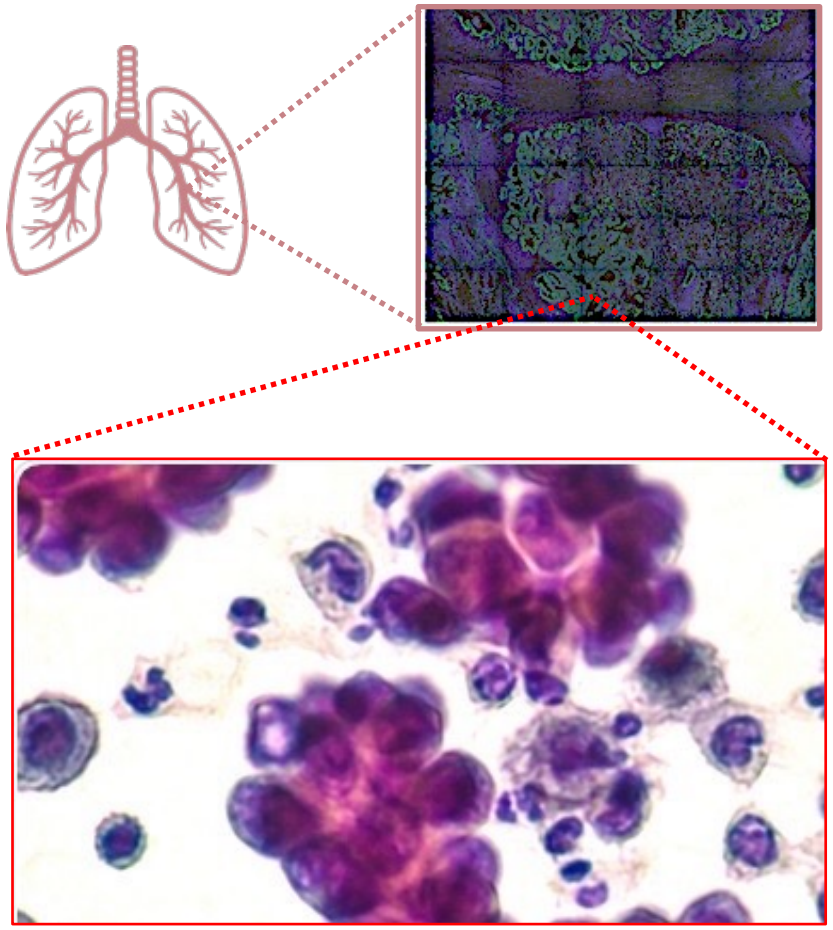
DNA → Genome

~ 1 Genome

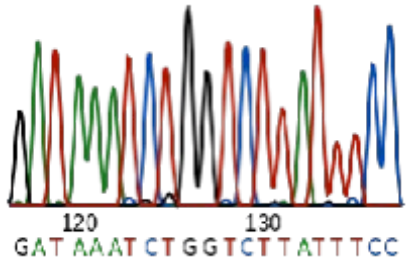
Protein → Proteome

~20,000 protein coding genes

What is transcriptomics?



What are these cells?
What are these cells doing?



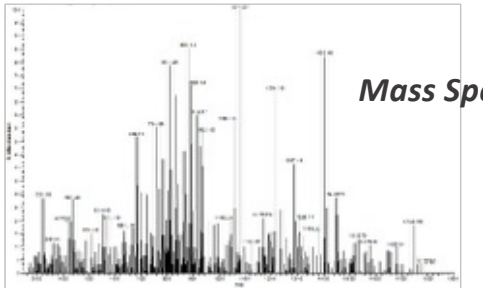
Sequencing

DNA → Genome

~ 1 Genome

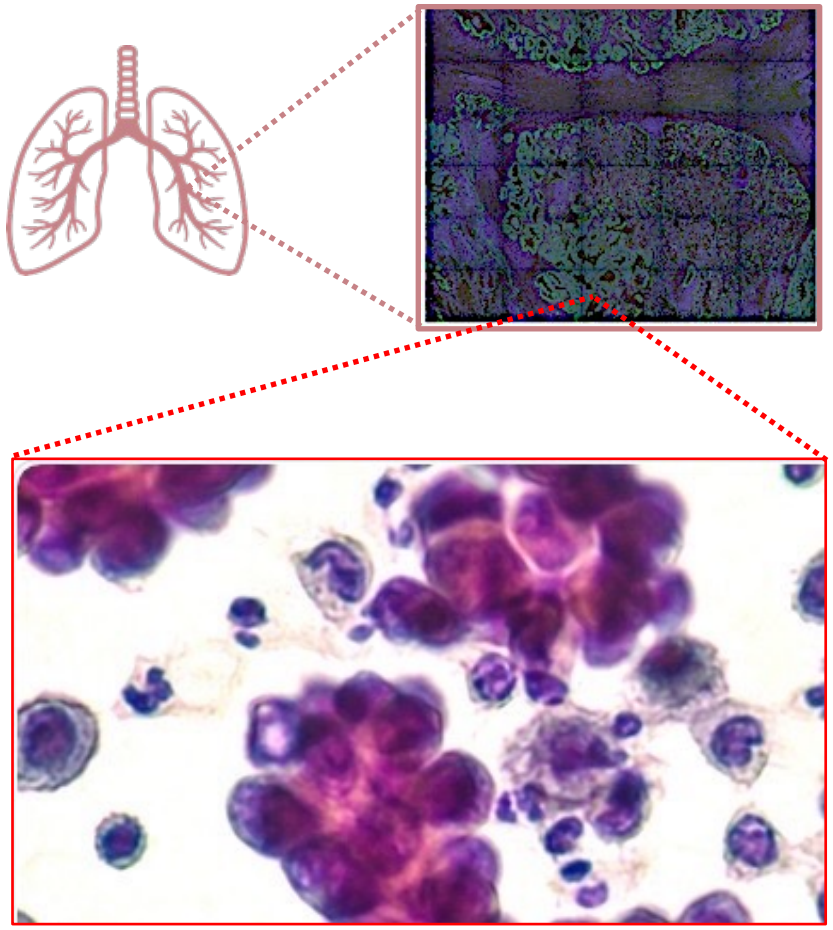
Protein → Proteome

~20,000 protein coding genes

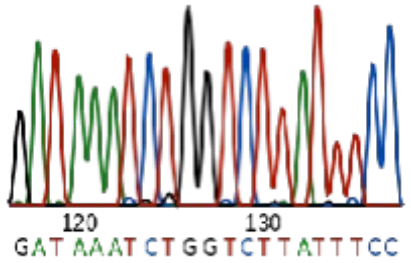


Mass Spectrometry

What is transcriptomics?



What are these cells?
What are these cells doing?



Sequencing

DNA → Genome

~ 1 Genome

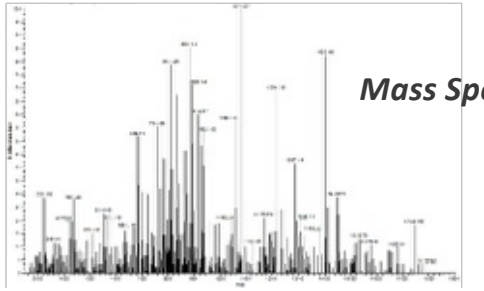
Transcription

mRNA → Transcriptome

Translation

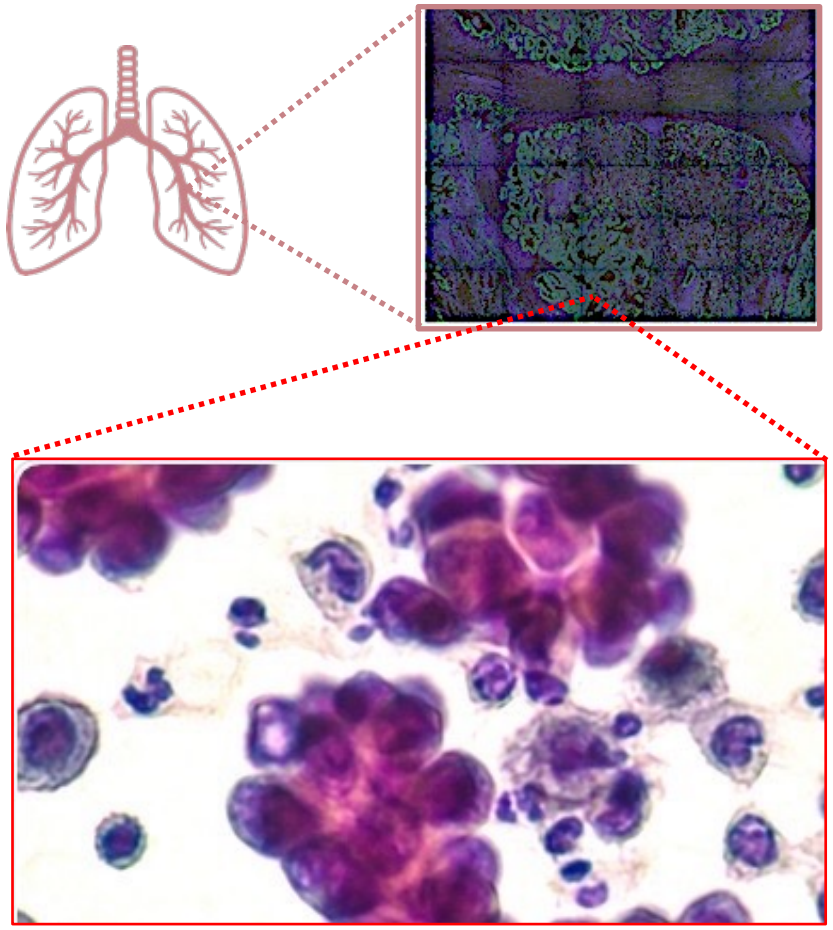
Protein → Proteome

~20,000 protein coding genes

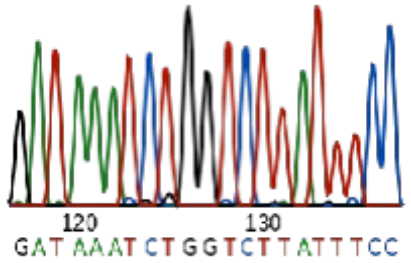


Mass Spectrometry

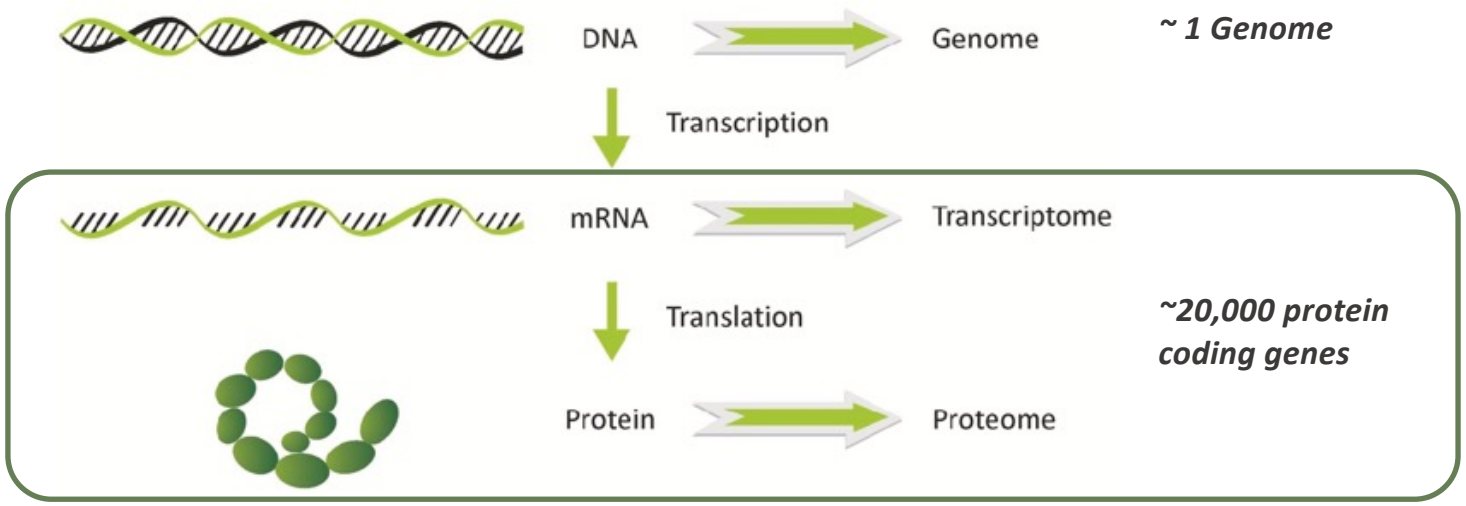
What is transcriptomics?

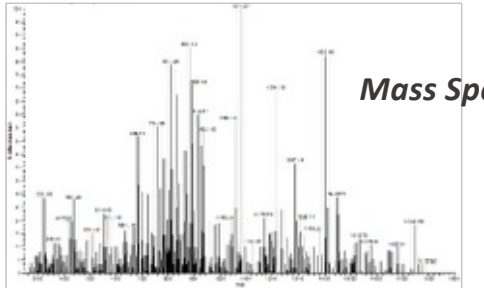


What are these cells?
What are these cells doing?



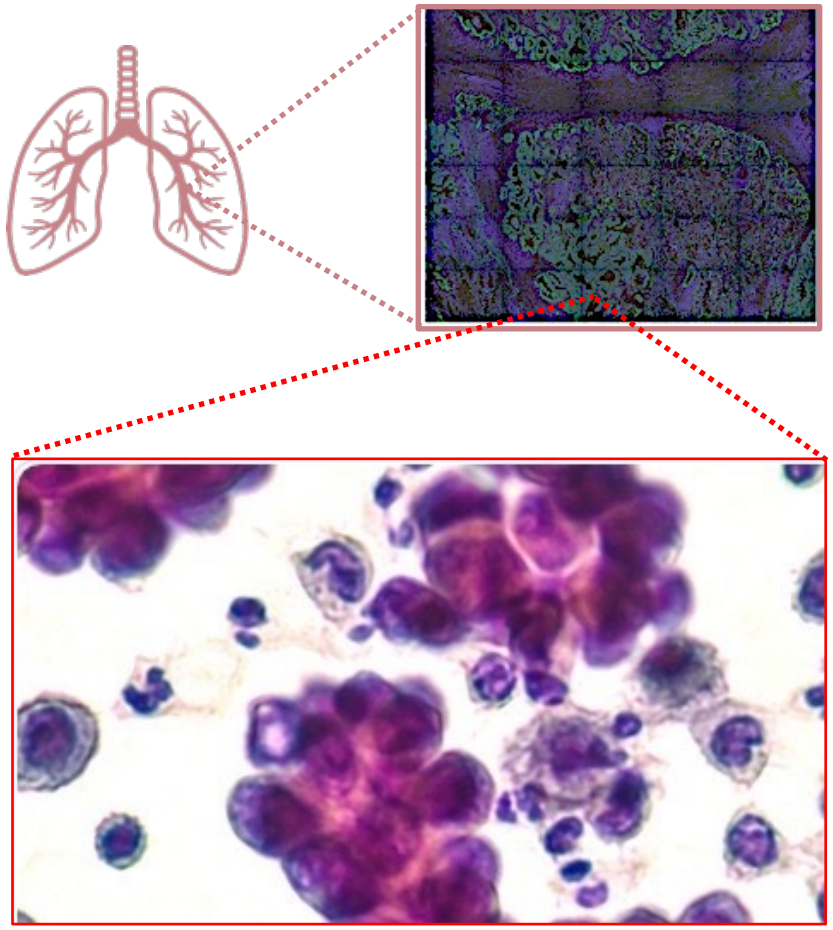
Sequencing



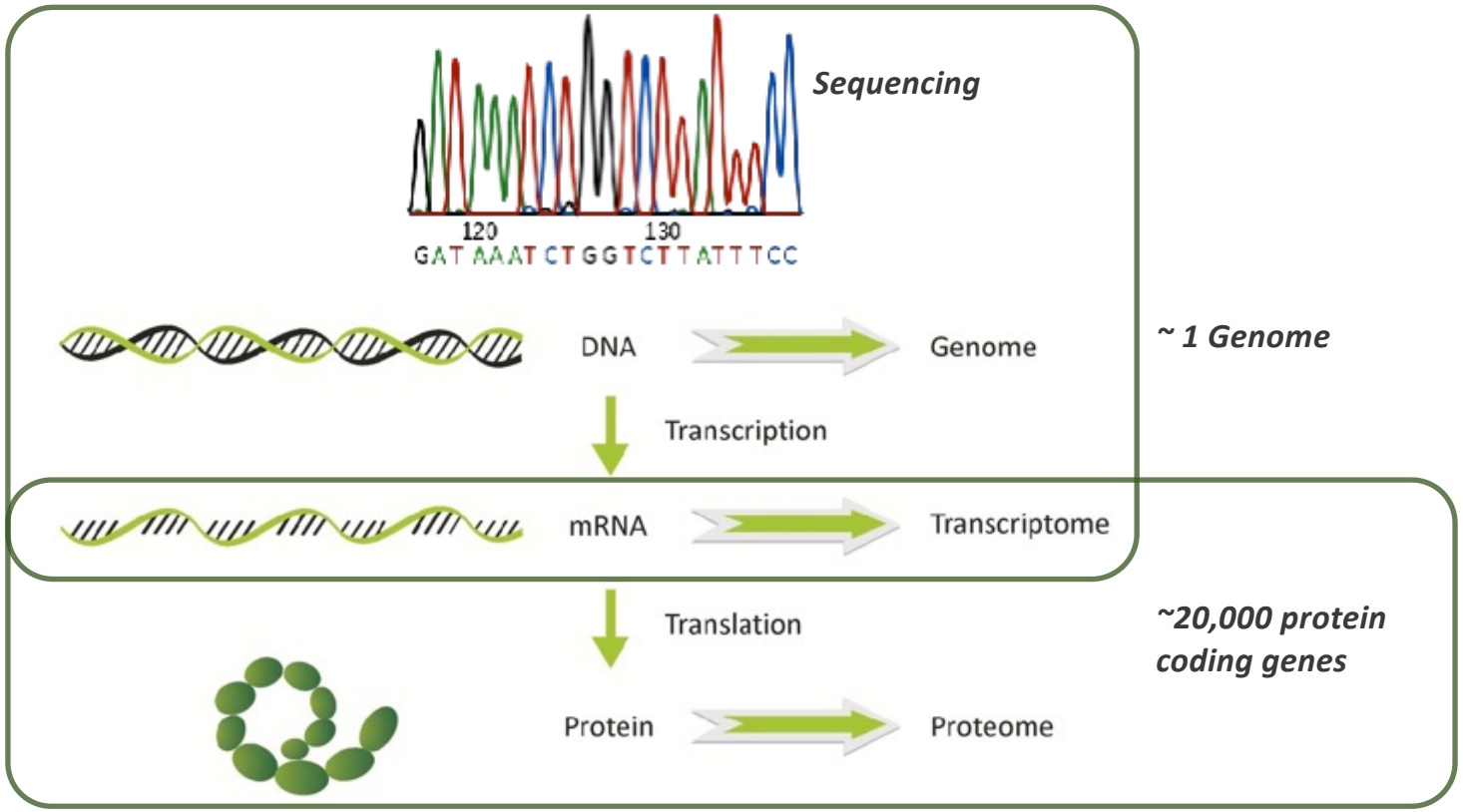


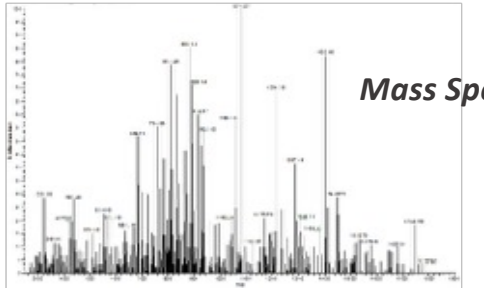
Mass Spectrometry

What is transcriptomics?



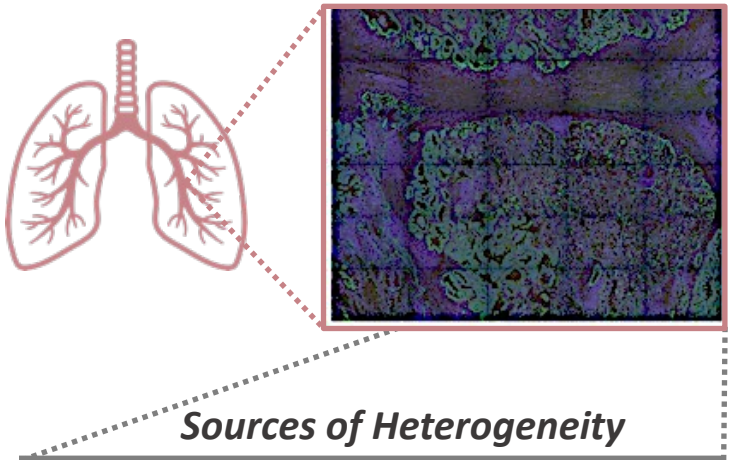
What are these cells?
What are these cells doing?





What is transcriptomics?

Domains of Transcriptomics

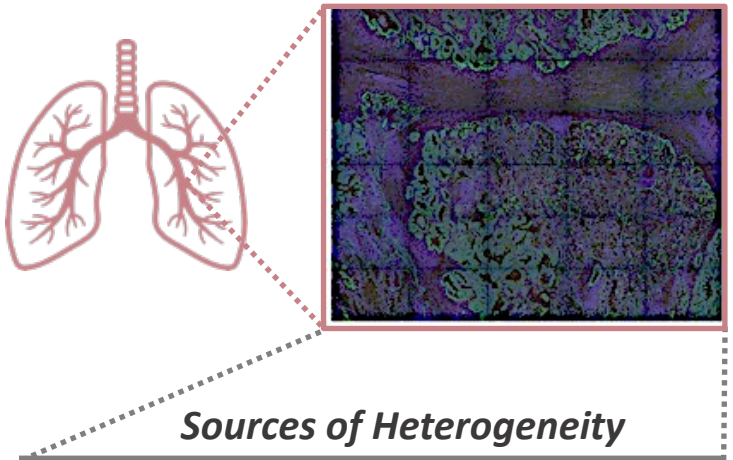


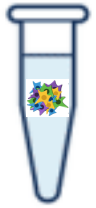
Sources of Heterogeneity

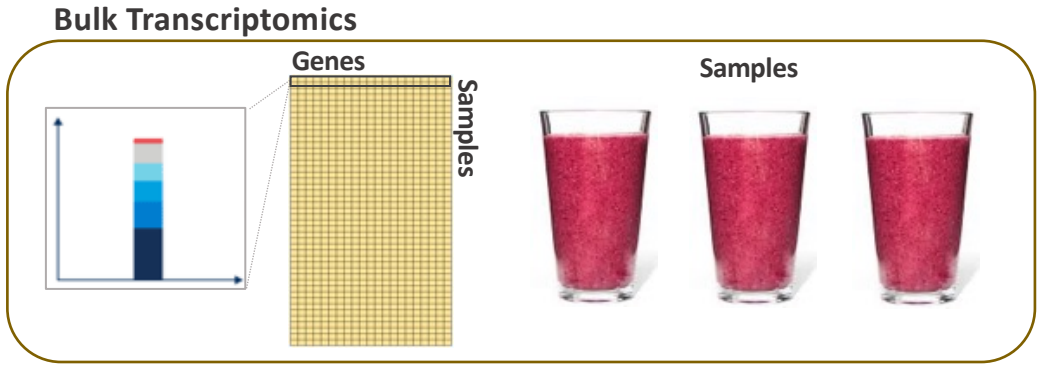
- Cell states/activity
- Cell type population
- Microstructures

https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Domains of Transcriptomics





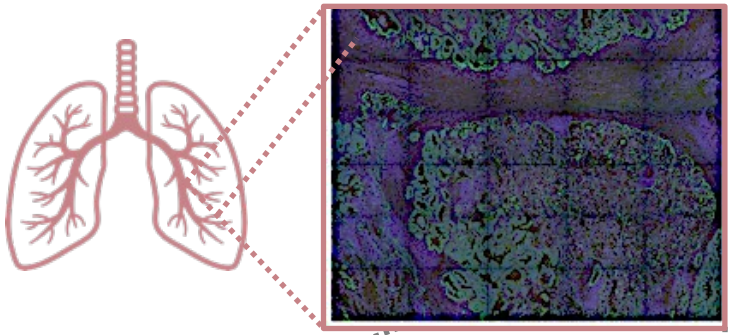


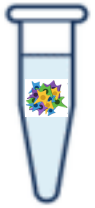
Sources of Heterogeneity

- Cell states/activity
- Cell type population
- Microstructures

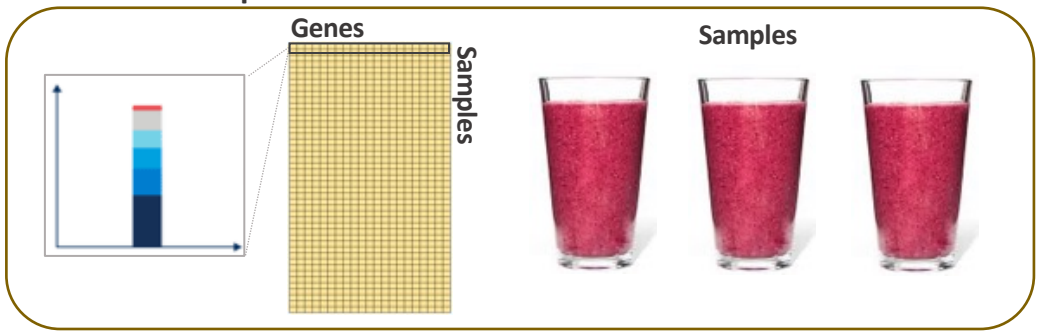
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Domains of Transcriptomics

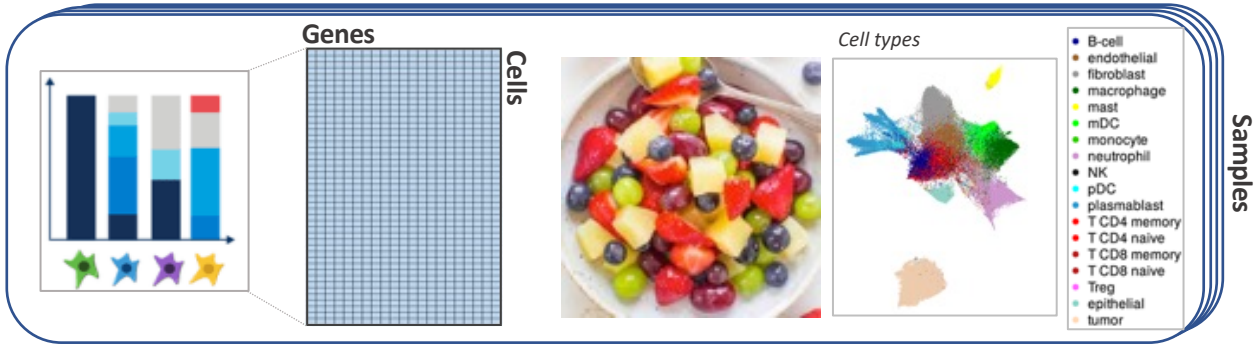




Bulk Transcriptomics



Single-cell Transcriptomics

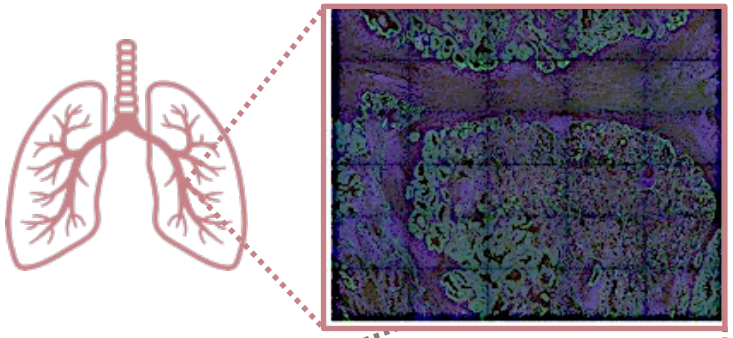


Sources of Heterogeneity

- Cell states/activity
- Cell type population
- Microstructures

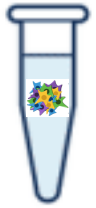
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

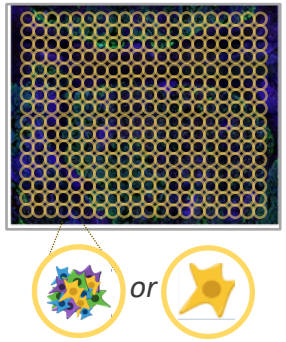
Domains of Transcriptomics



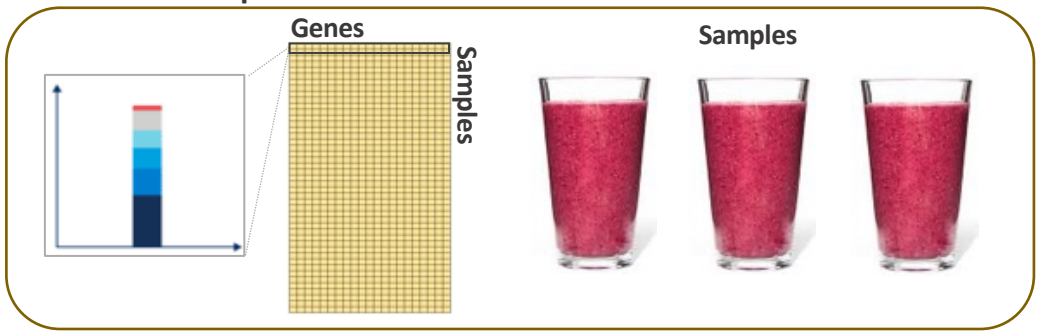
Sources of Heterogeneity

- Cell states/activity
- Cell type population
- Microstructures

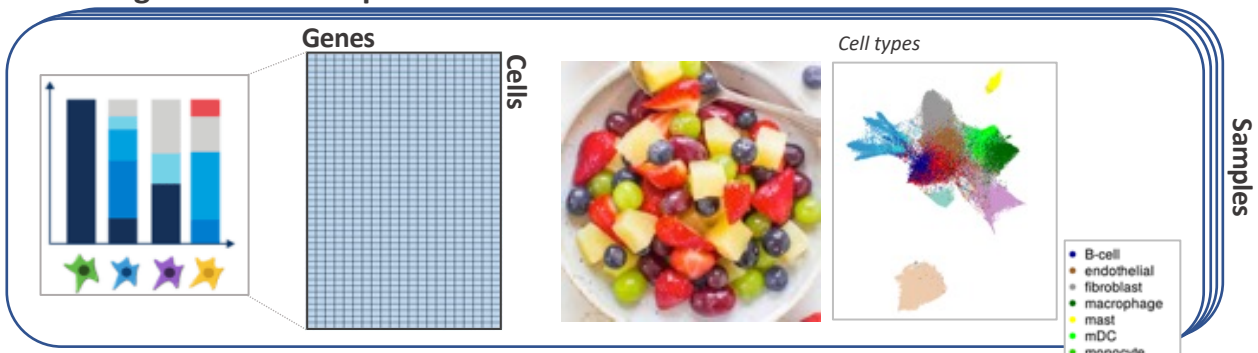




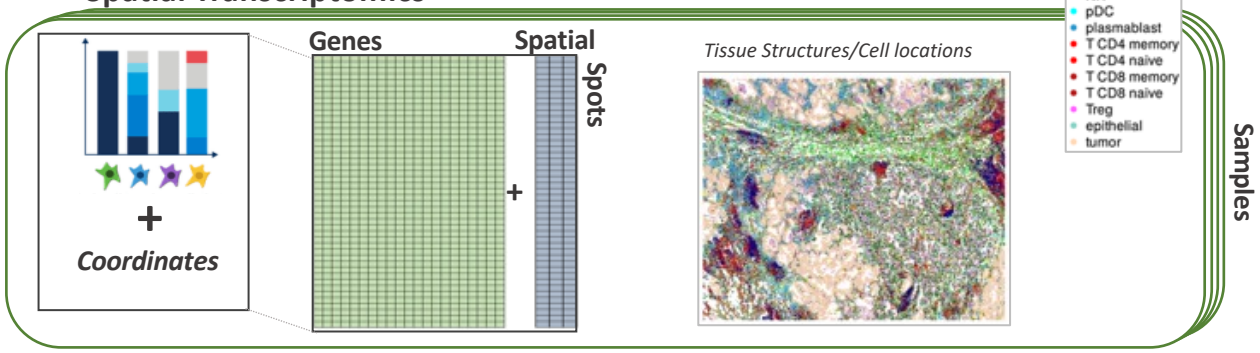
Bulk Transcriptomics



Single-cell Transcriptomics

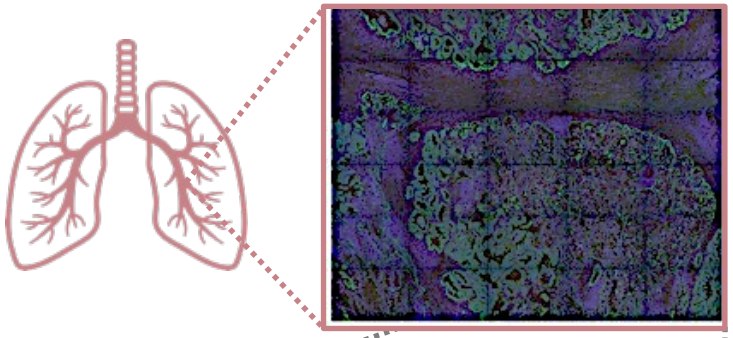


Spatial Transcriptomics



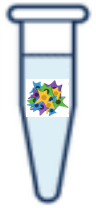
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

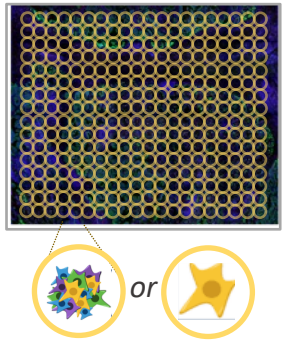
Domains of Transcriptomics



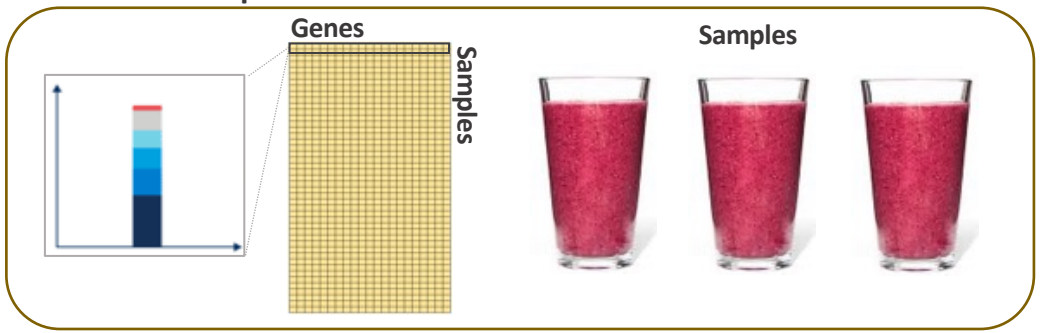
Sources of Heterogeneity

- Cell states/activity
- Cell type population
- Microstructures

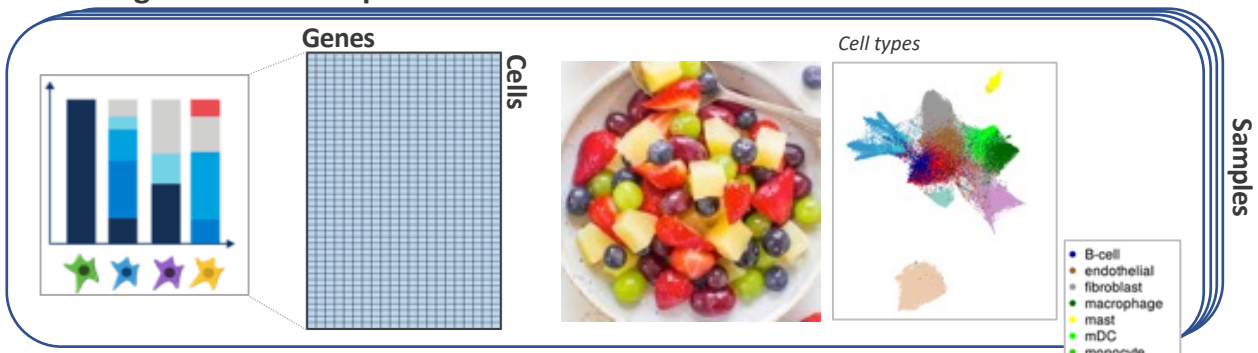




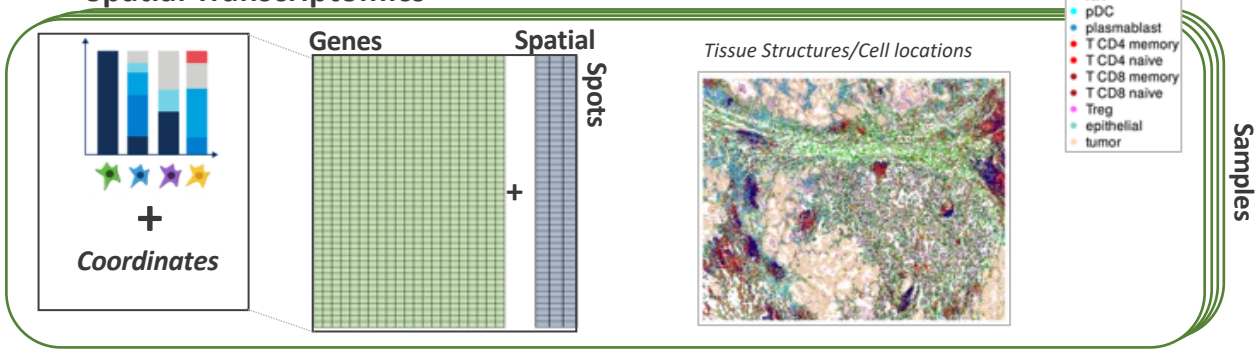
Bulk Transcriptomics



Single-cell Transcriptomics



Spatial Transcriptomics



https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

Bulk and Single-cell Transcriptomics Analyses

Gene-level Methods

- **Bulk: Tissue-level**
- **Single-cell: Cell type level**

Differential gene expression analysis (DGE)

- Identify differences in gene expression between groups

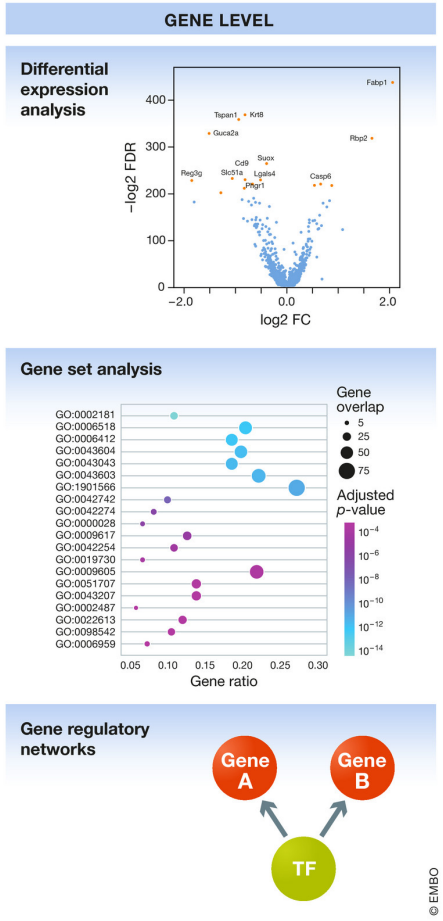
Gene set analyses

- Annotate DGE signatures based on their presentation in known biological pathways

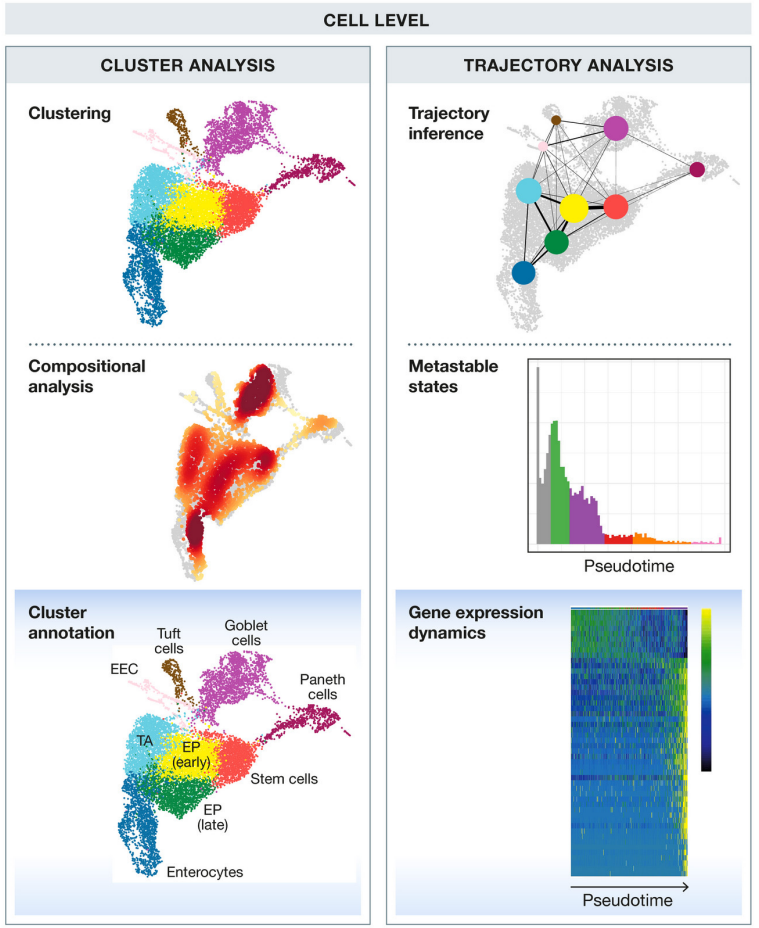
Gene regulatory networks

- Infer patterns of gene co-expression and their likely regulatory mechanisms

Bulk and Single Cell



Single-cell Only



https://doi.org/10.15252/msb.20188746

Single-cell Transcriptomics Analyses

Cell-level methods

- **Single-cell only**

Cellular Clustering analysis

- Delineate cell profiles into clusters of similar gene expression profiles, indicative of cell type and biological states

Cell type annotation

- Infer identity of cells within each cluster based on their gene expression patterns

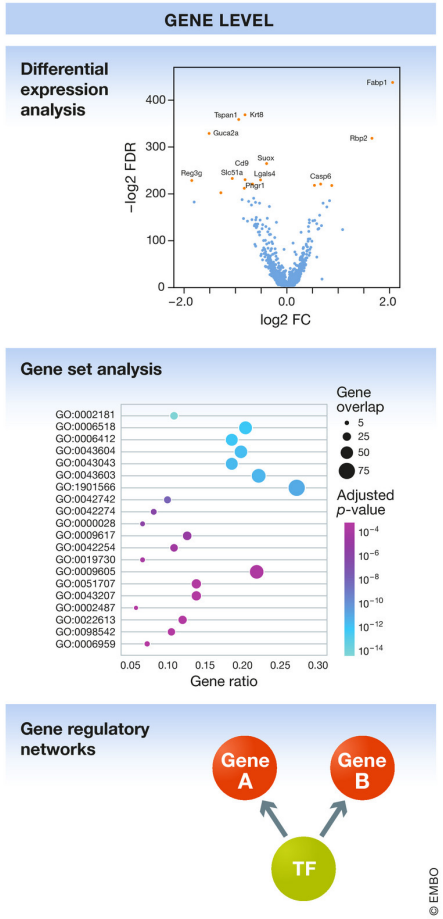
Compositional analysis

- Infer differences in the abundance of one-or-more cell types between groups

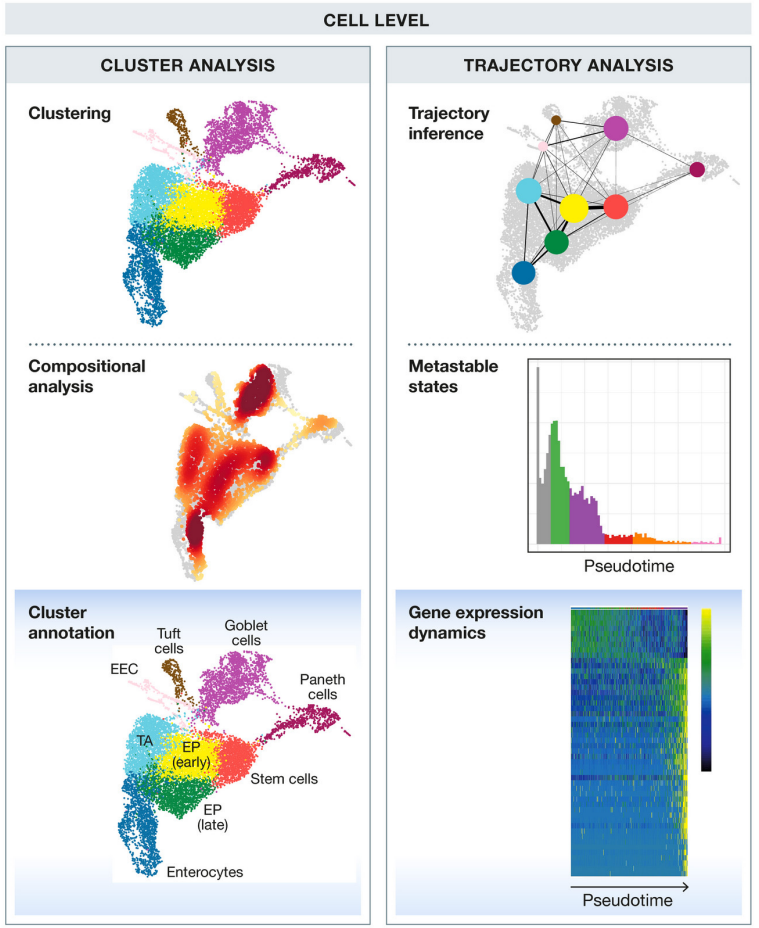
Trajectory analysis

- Infer temporal relationships between cell profiles
  - Differentiation trajectory
  - Transitions between biological states

Bulk and Single Cell

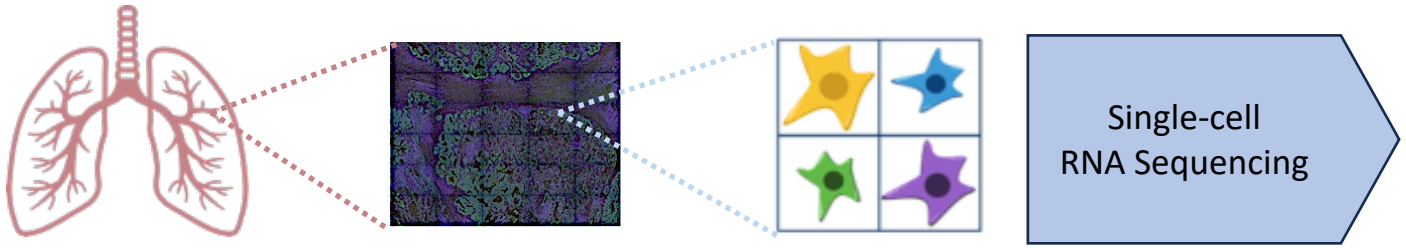


Single-cell Only



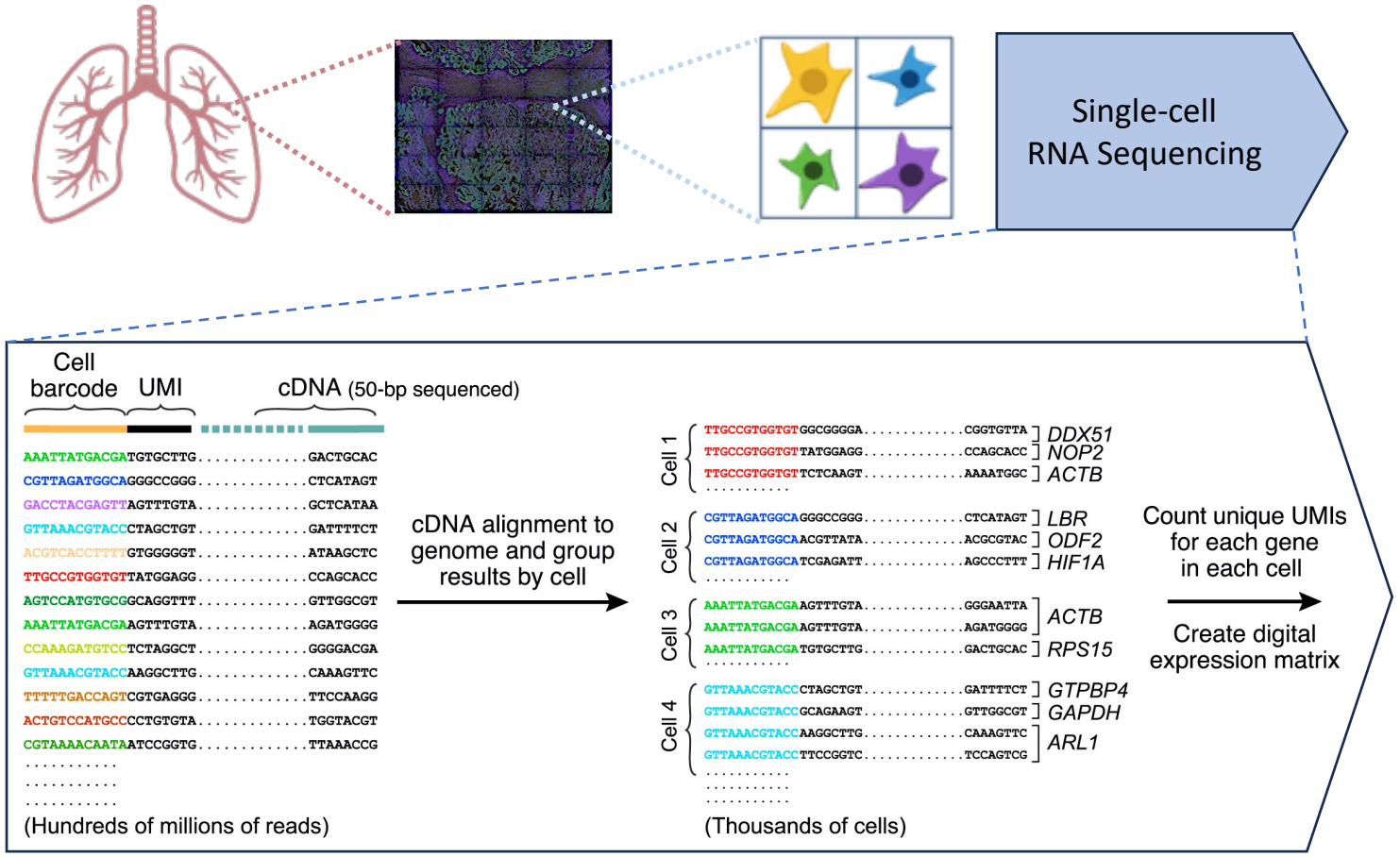
https://doi.org/10.15252/msb.20188746

Single-cell RNA Sequencing Data



DOI:https://doi.org/10.1038/s41591-020-0901-9

Single-cell RNA Sequencing Data



Cell barcode

Identifies cell from which the transcript originated

Unique Molecular Identifier (UMI)

Uniquely identifies the RNA fragment a read came from.

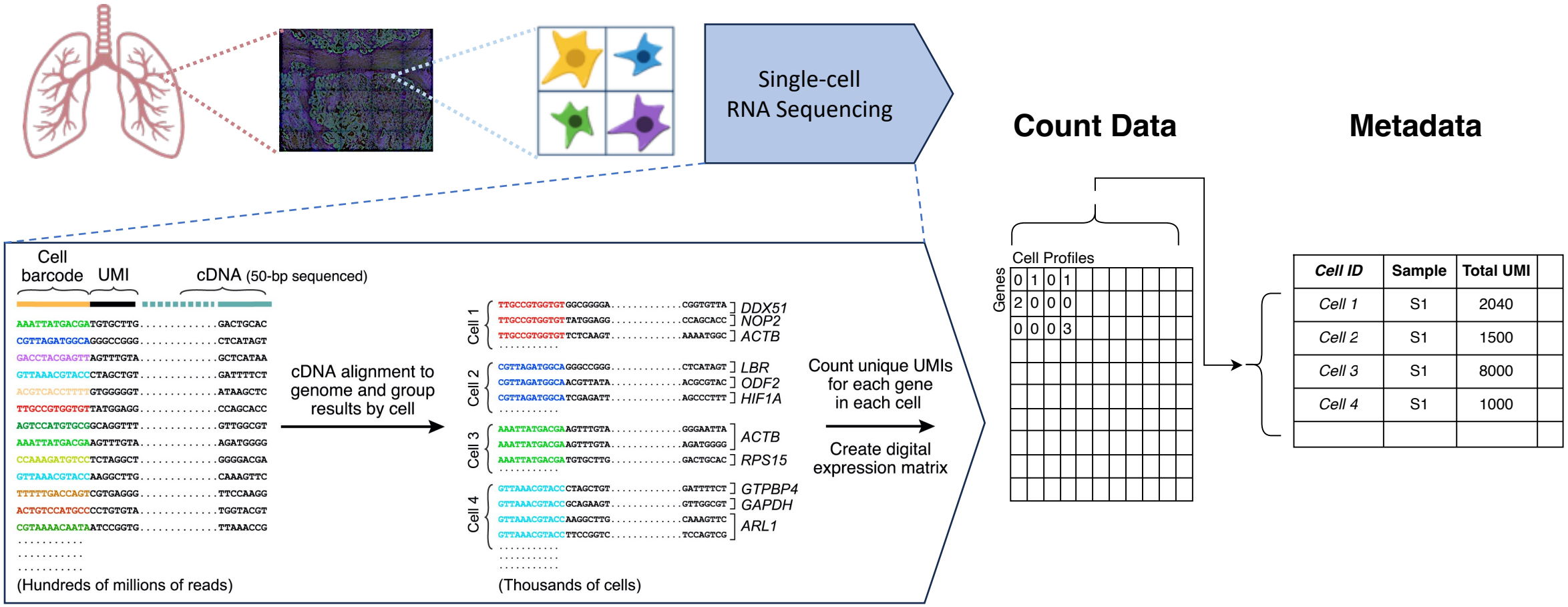
- *"*UMIs" and "counts" are often used interchangeably

cDNA sequence

The sequence of a captured RNA transcript

DOI:https://doi.org/10.1038/s41591-020-0901-9

Single-cell RNA Sequencing Data



Cell ID	Sample	Total UMI	
Cell 1	S1	2040	
Cell 2	S1	1500	
Cell 3	S1	8000	
Cell 4	S1	1000	

Cell barcode

Identifies cell from which the transcript originated

Unique Molecular Identifier (UMI)

Uniquely identifies the RNA fragment a read came from.

- *"*UMIs" and "counts" are often used interchangeably

cDNA sequence

The sequence of a captured RNA transcript

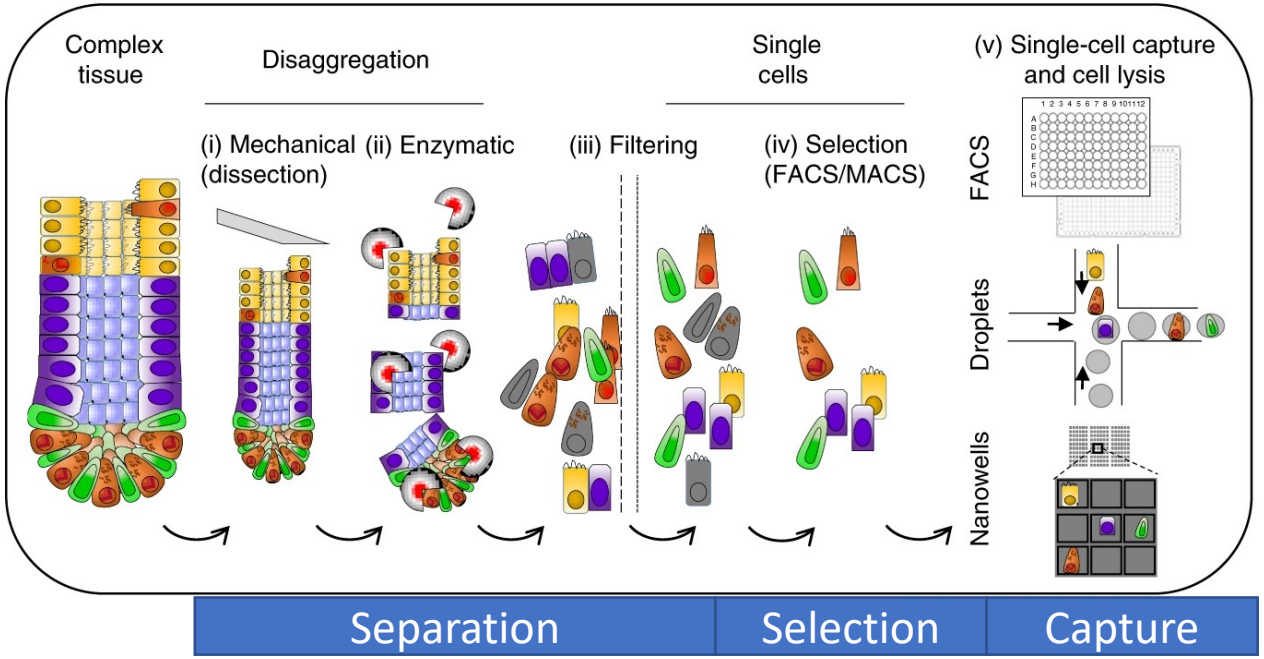
DOI:https://doi.org/10.1038/s41591-020-0901-9

Sample Preparation

Separation

Dissociate tissue into individual cells

- **Stressful**
  - **Induce activation of stress and apoptotic genes**
    - **Especially mitochondria content**
  - **Some-to-most cells**



https://doi.org/10.1038/s41596-018-0073-y

Sample Preparation

Separation

Dissociate tissue into individual cells

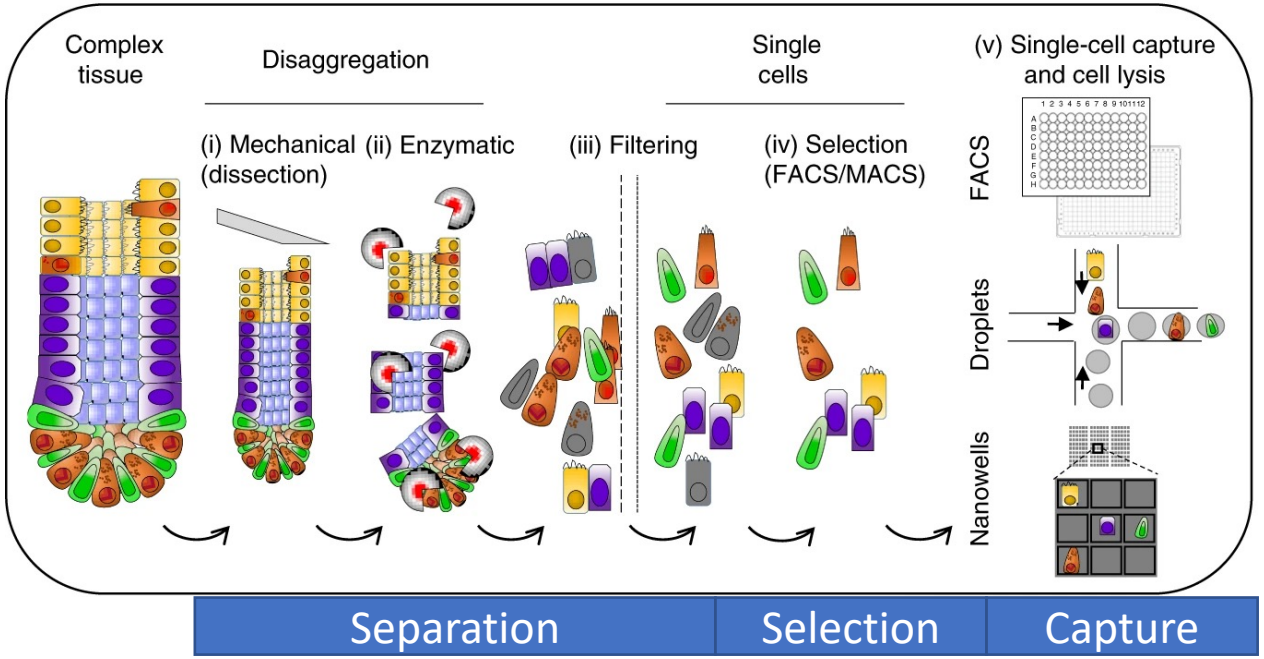
- **Stressful**
  - **Induce activation of stress and apoptotic genes**
    - **Especially mitochondria content**
  - **Some-to-most cells**

Selection (Optional)

Filter of cells for types of interest

- Example: **$CD45^+$ selection for immune cells**

- **Varying efficacy**
  - **May distort cell type composition analyses**



https://doi.org/10.1038/s41596-018-0073-y

Sample Preparation

Separation

Dissociate tissue into individual cells

- **Stressful**
  - **Induce activation of stress and apoptotic genes**
    - **Especially mitochondria content**
  - **Some-to-most cells**

Selection (Optional)

Filter of cells for types of interest

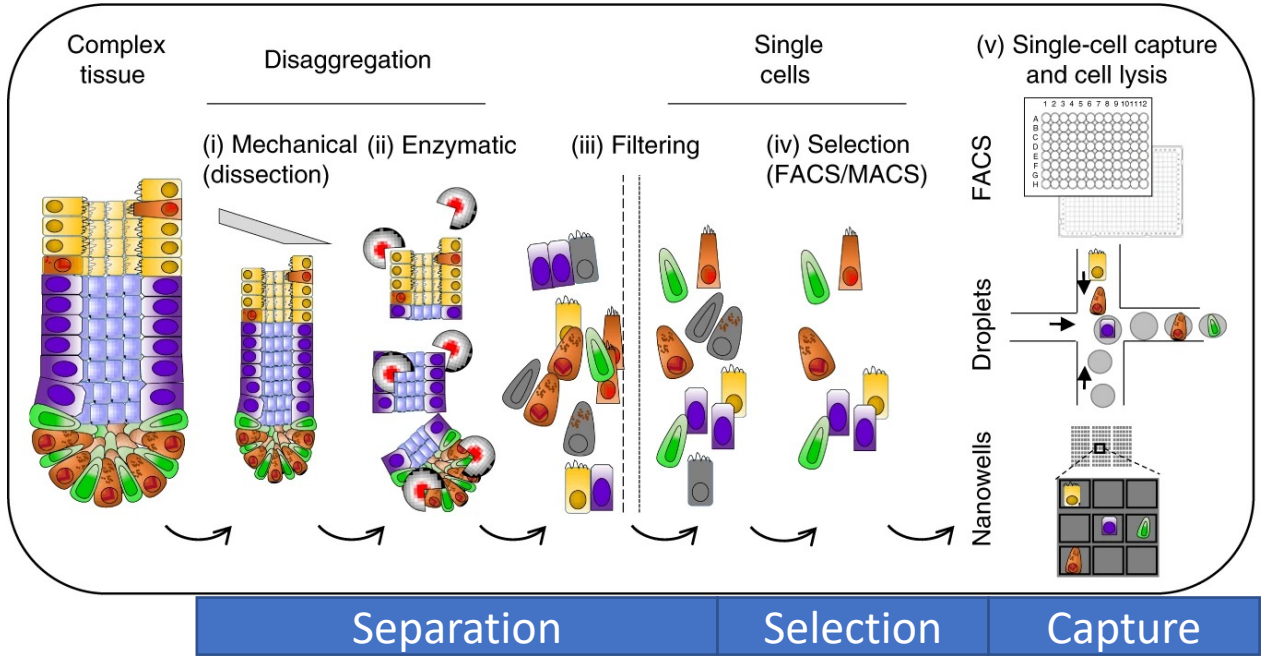
- Example: **$CD45^+$ selection for immune cells**

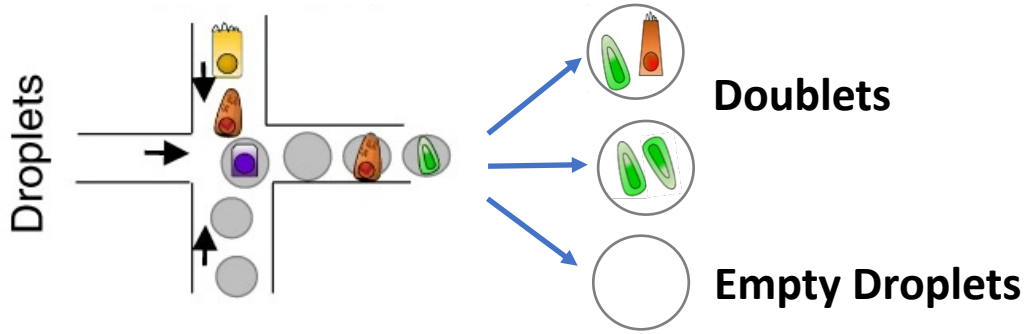
- **Varying efficacy**
  - **May distort cell type composition analyses**

Capture

Suspending individual cells for sequencing

- **Doublets/Empty Droplets**
- Ambient RNA





https://doi.org/10.1038/s41596-018-0073-y

Sample Preparation

Separation

Dissociate tissue into individual cells

- **Stressful**
  - **Induce activation of stress and apoptotic genes**
    - **Especially mitochondria content**
  - **Some-to-most cells**

Selection (Optional)

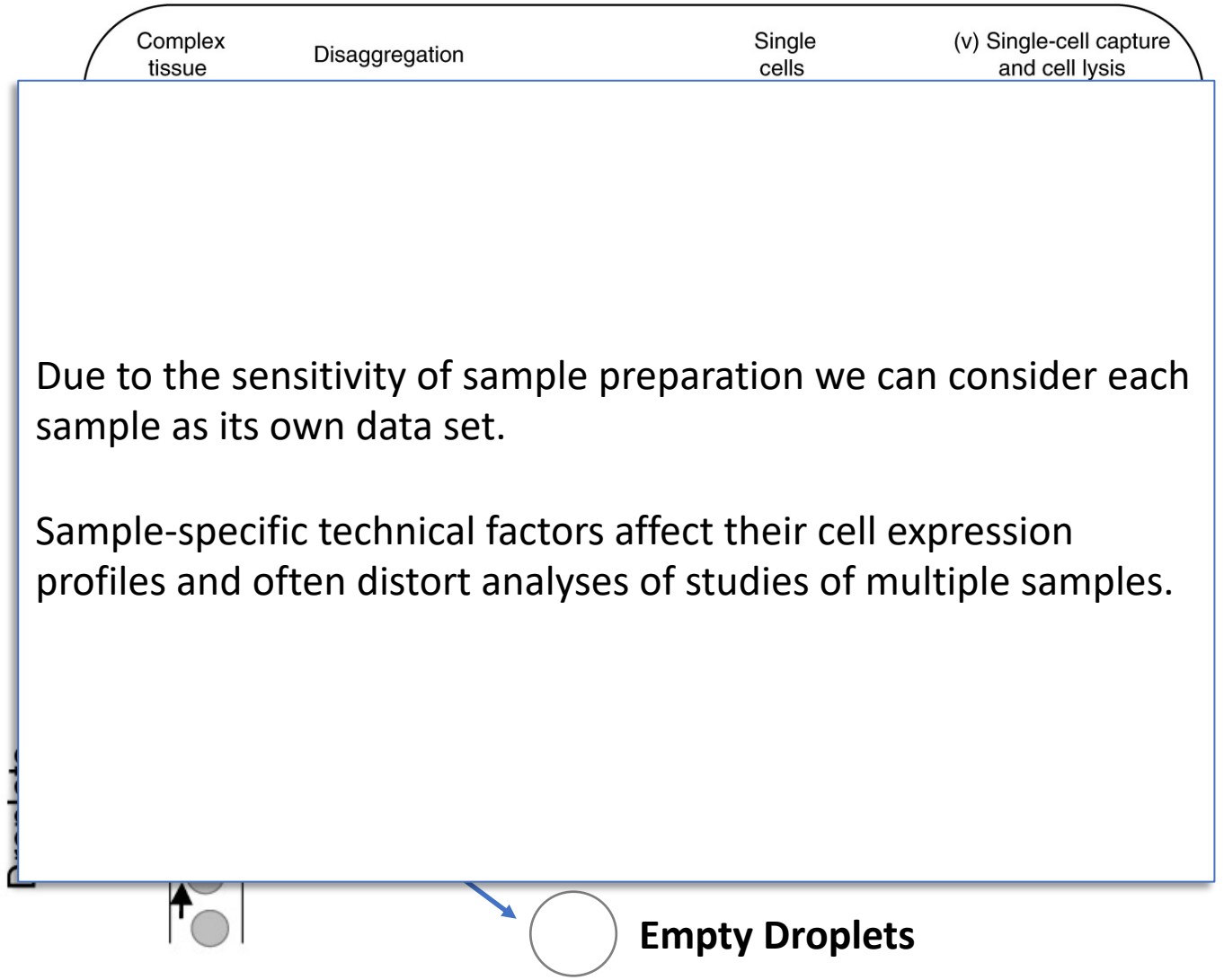
Filter of cells for types of interest

- Example: **$CD45^+$ selection for immune cells**
- **Varying efficacy**
  - **May distort cell type composition analyses**

Capture

Suspending individual cells for sequencing

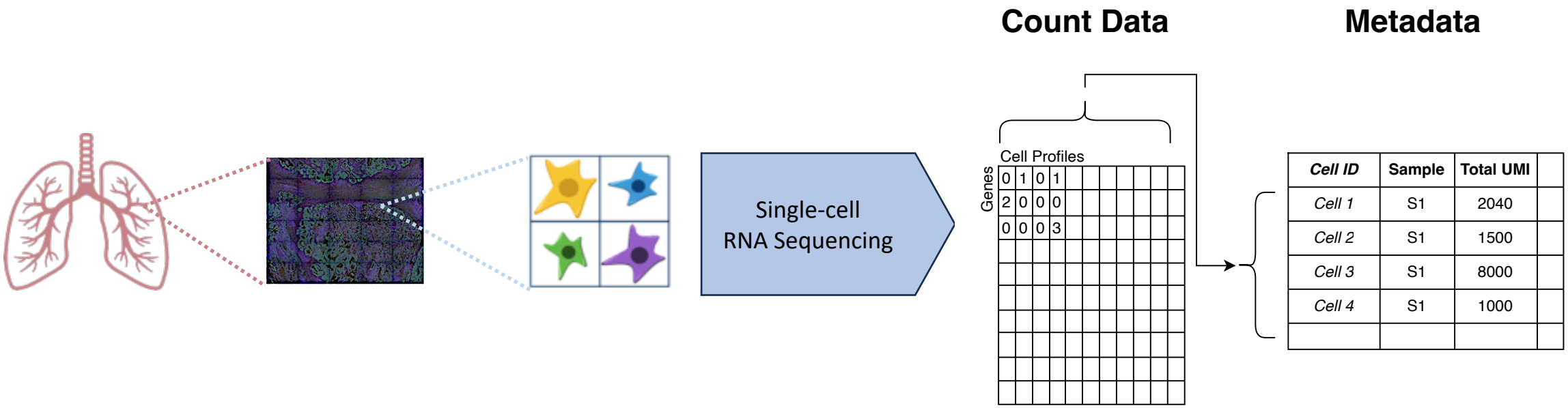
- **Doublets/Empty Droplets**
- Ambient RNA



https://doi.org/10.1038/s41596-018-0073-y

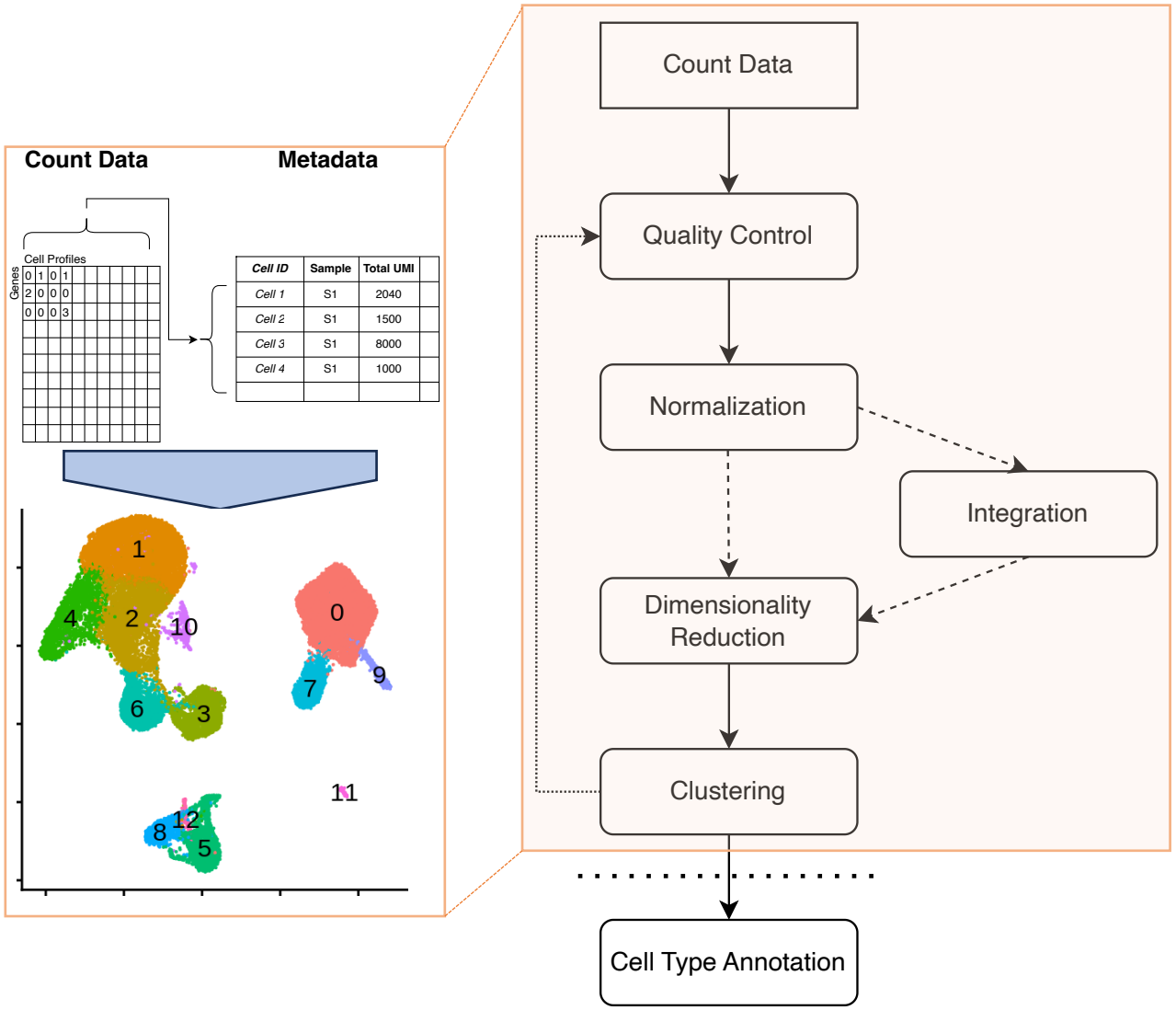
What we are covering today

- The majority of downstream scRNAseq analyses require cell type labels, which are unknown in the raw data
- To this end, we will cover a common scRNAseq clustering workflow to facilitate cell type identification



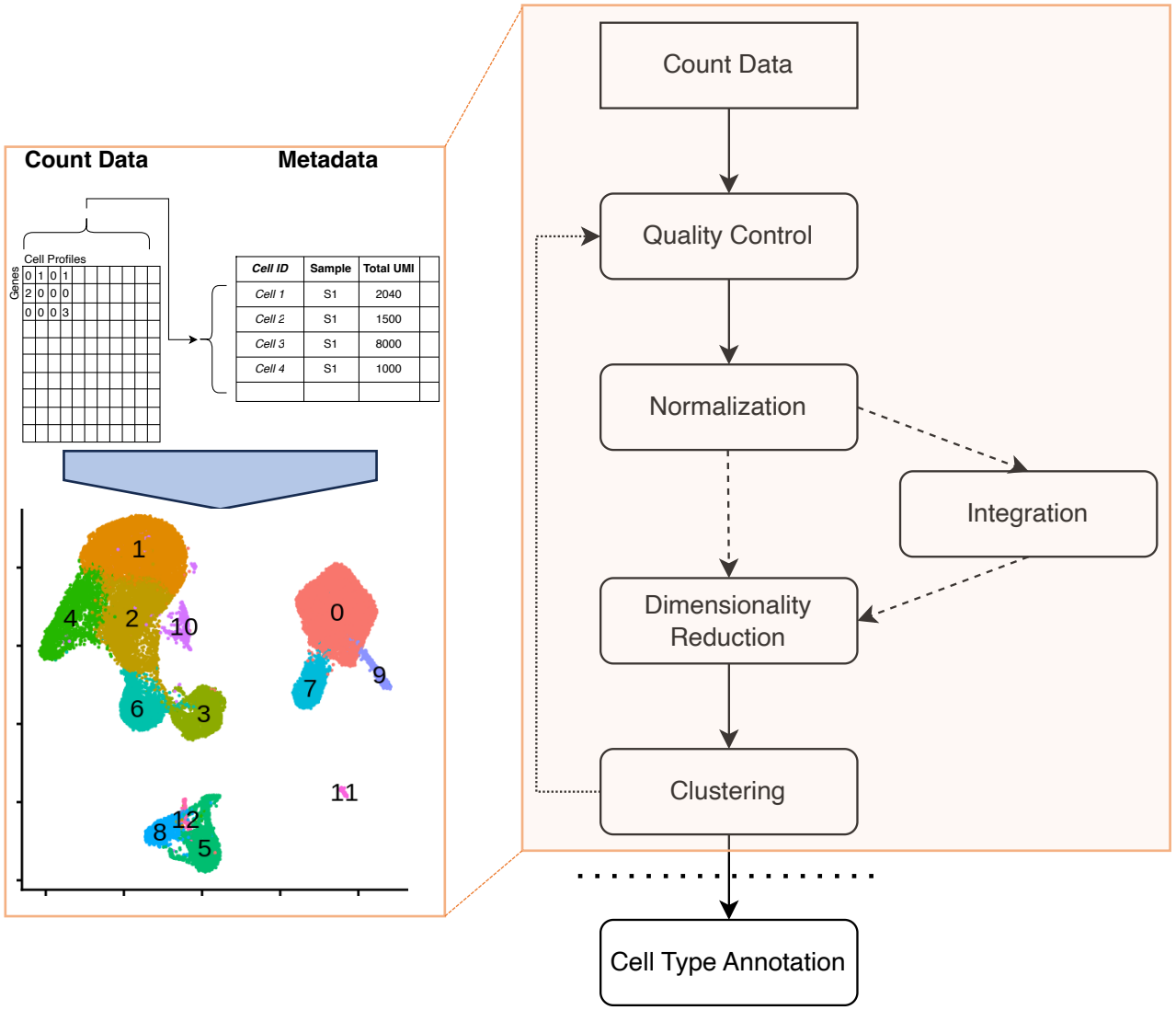
What we are covering today

scRNAseq Clustering Workflow



What we are covering today

scRNAseq Clustering Workflow



Case Study

scRNAseq Data

Peripheral Blood Mononuclear Cells (PBMCs)

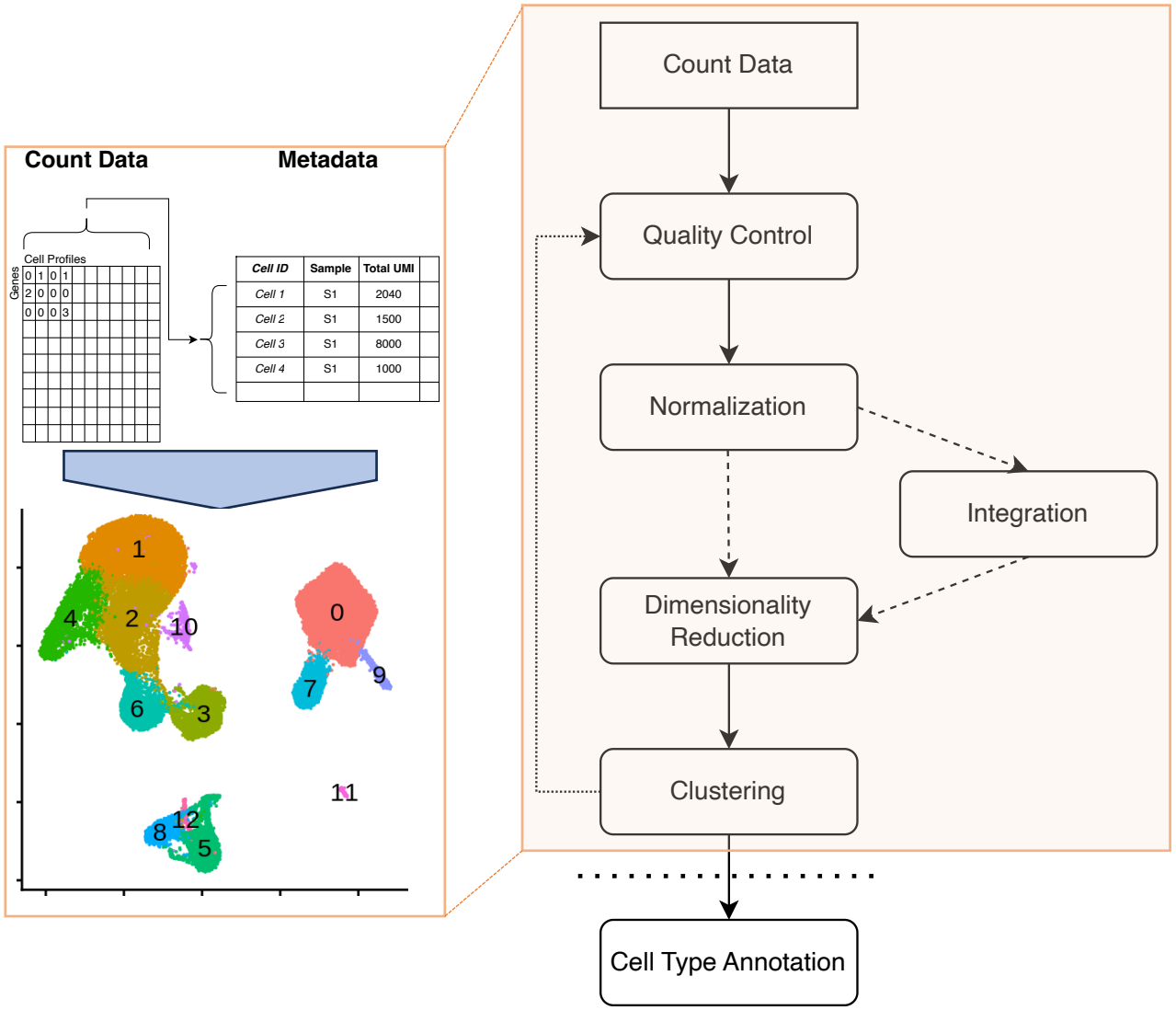
Includes:
- B cells
- T cells
- Natural killer cells
- Monocytes
- Macrophages

Doesn't include:
- Neutrophils
- Platelets
- Red blood cells

Two samples
- Control: "ctrl"
- Interferon beta-treated (stimulated): "stim"

What we are covering today

scRNAseq Clustering Workflow



Case Study

scRNAseq Data

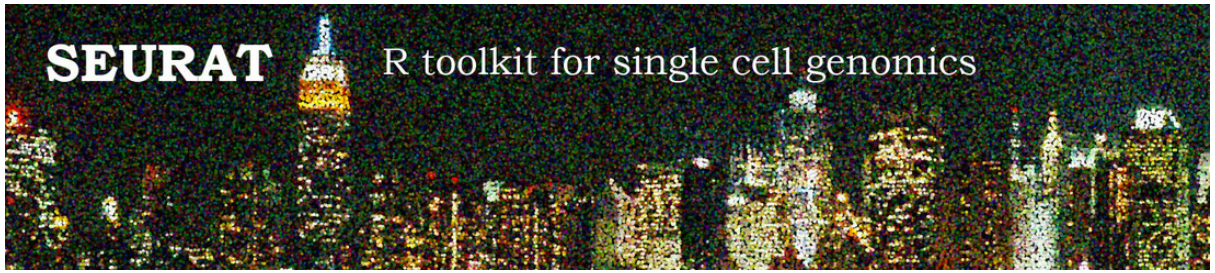
Peripheral Blood Mononuclear Cells (PBMCs)

Includes:
- B cells
- T cells
- Natural killer cells
- Monocytes
- Macrophages

Doesn't include:
- Neutrophils
- Platelets
- Red blood cells

Two samples
- Control: "ctrl"
- Interferon beta-treated (stimulated): "stim"

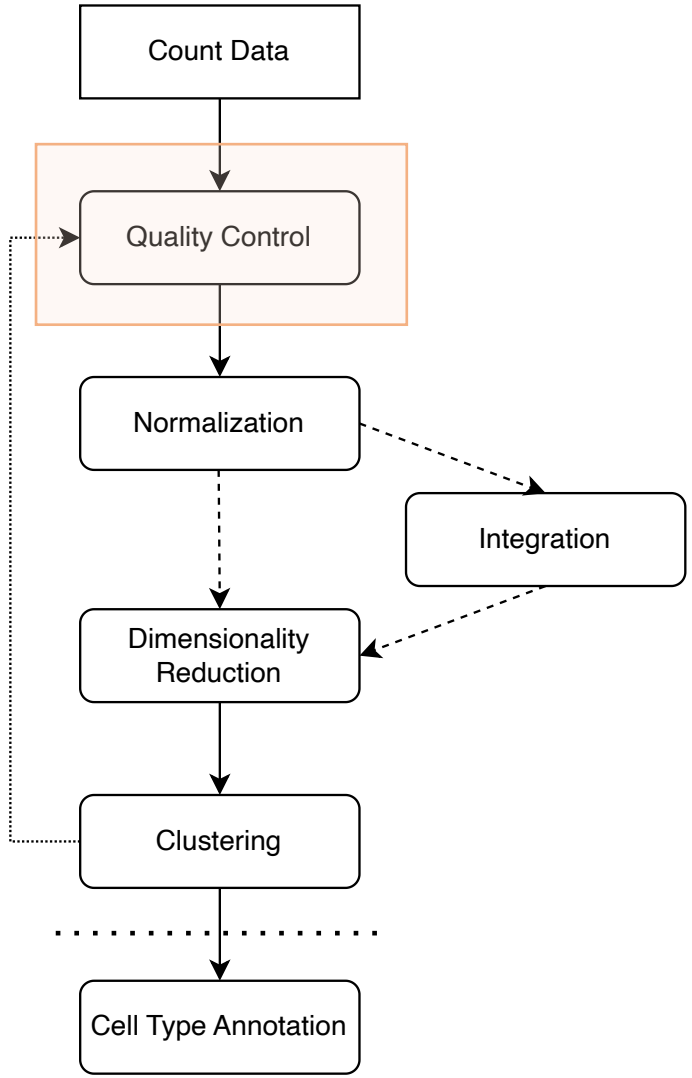
Specific Workflow



https://satijalab.org/seurat/

Quality Control

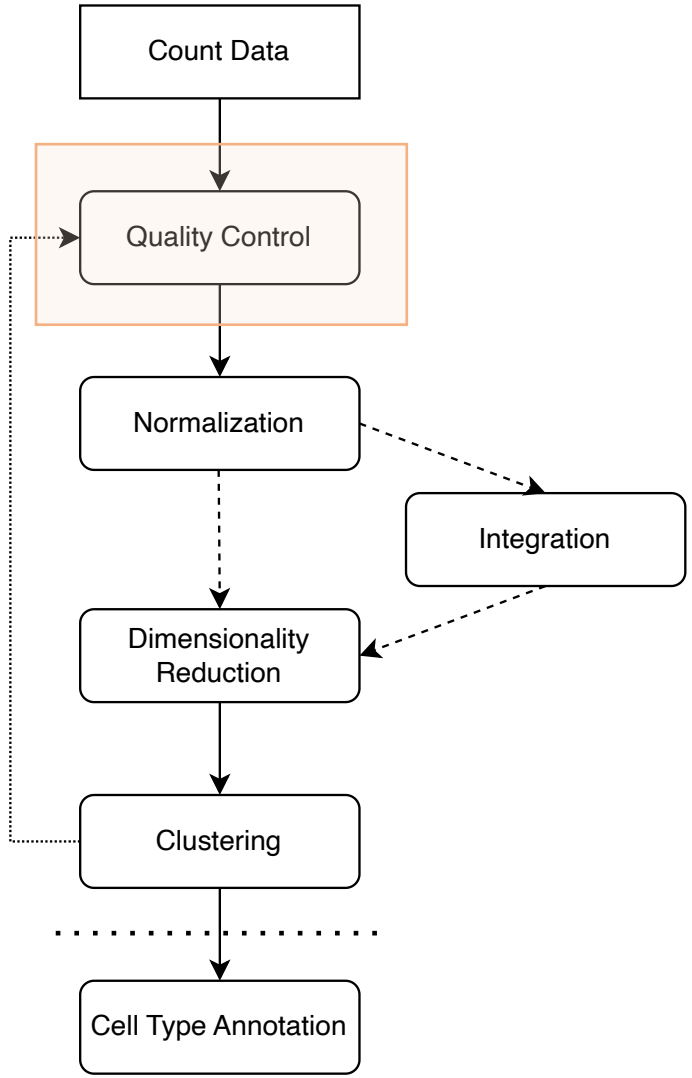
Low quality samples are a sources of unwanted variability in the data



Quality Control

Low quality samples are a sources of unwanted variability in the data

- Low resolution profiles have low information
- Low resolution profiles could be empty droplets containing ambient RNA
  - Low UMI count
  - Few genes detected
  - Low complexity
    - Novelty Score
      - Log10(# Genes) / Log10(# UMIs)

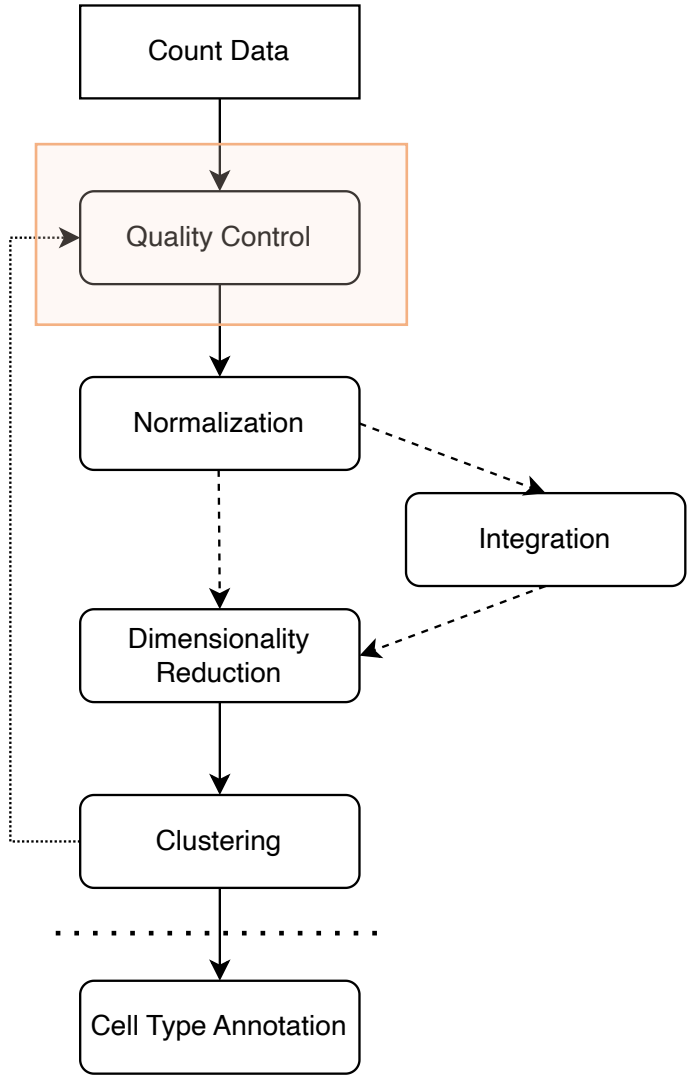


Quality Control

Low quality samples are a sources of unwanted variability in the data

- Low resolution profiles have low information
- Low resolution profiles could be empty droplets containing ambient RNA
  - Low UMI count
  - Few genes detected
  - Low complexity
    - Novelty Score
      - Log10(# Genes) / Log10(# UMIs)

- Stressed or dying cells can be artifacts of sample preparation
  - **High mitochondrial content**



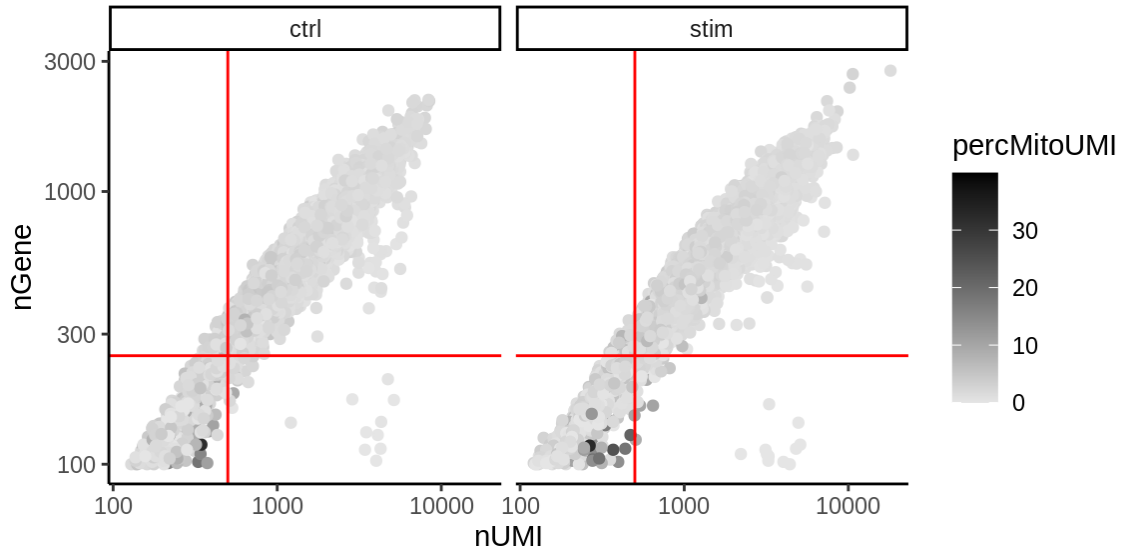
Quality Control

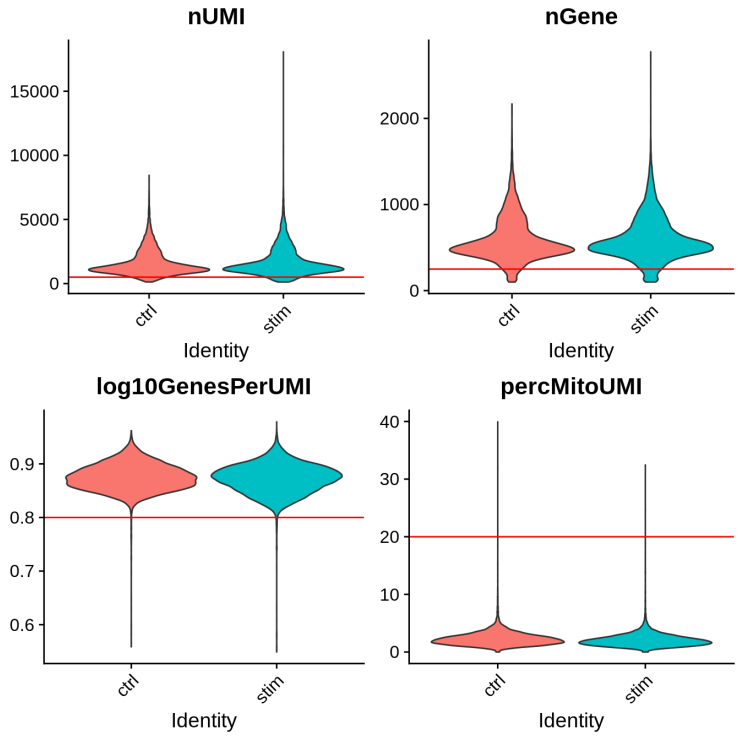
PBMC Data: 31,444 Total Cell Profiles

Keep Cell Profiles:

- Total UMI count (nUMI) > 500
- Number of counted genes (nGene) > 250
- Novelty Score (log10GenePerUMI) > 0.8
- Percent Mitochondrial UMI (percMitoUMI) < 0.2

	Fail	Pass
nUMI > 500	1609	29835
nGene > 250	1408	30036
log10GenesPerUMI > 0.8	112	31332
percMitoUMI > 0.2	154	31290
All filters	1860	29584





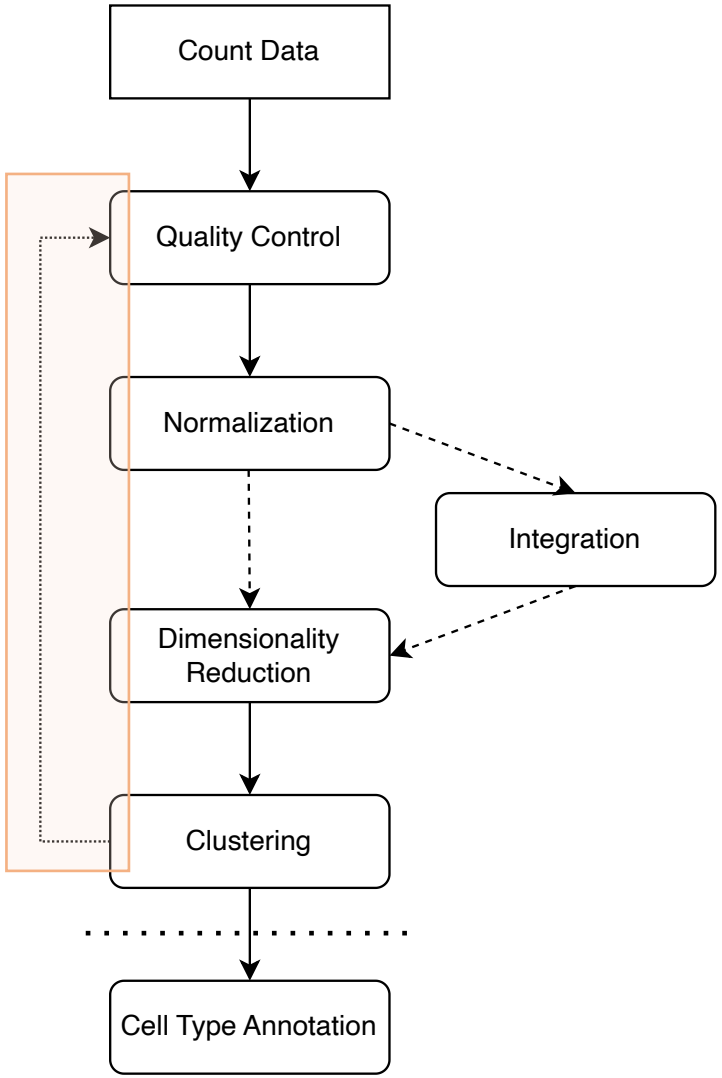
Quality Control

Generally, these workflows are iterative

1. **Start by using less stringent quality control filters**
2. **Evaluate these quality control metrics at later point in the workflow.**
   - *Are specific cell clusters characterized primarily based on mitochondrial content, low total UMI?*
3. **If necessary, reperform quality control filtering with more stringent thresholds**

Biological understanding of our samples is helpful

Do we expect specific cell types to have elevated mitochondrial content, or few expressed genes?

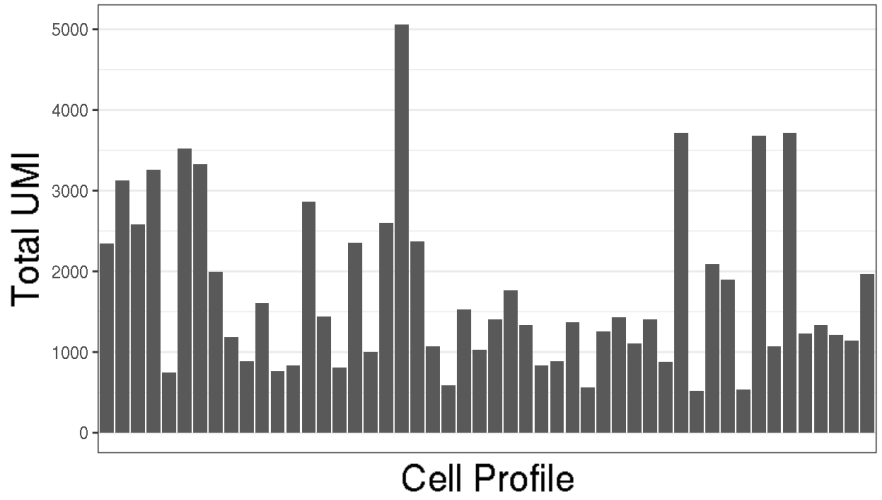


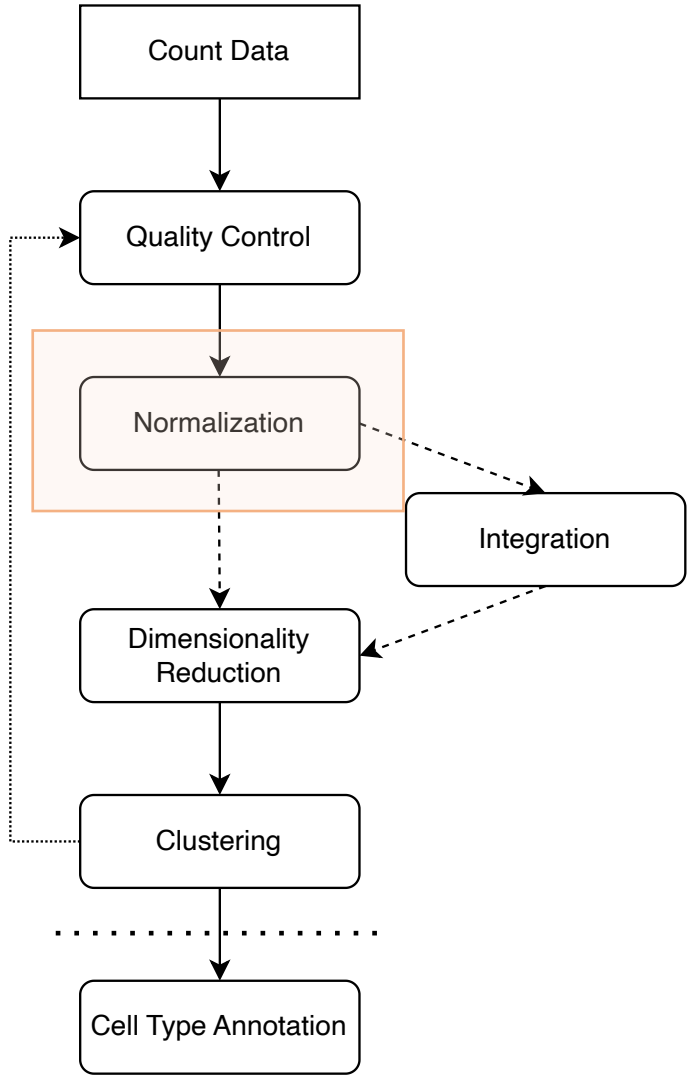
Data Normalization

The performance of statistical and machine learning methods is dependent on the properties of our data

Count data is sub-optimal as input for our clustering workflow

- **Gene UMI counts are dependent on the total number of UMIs in a sample**
- **We cannot immediately compare UMI counts across samples**



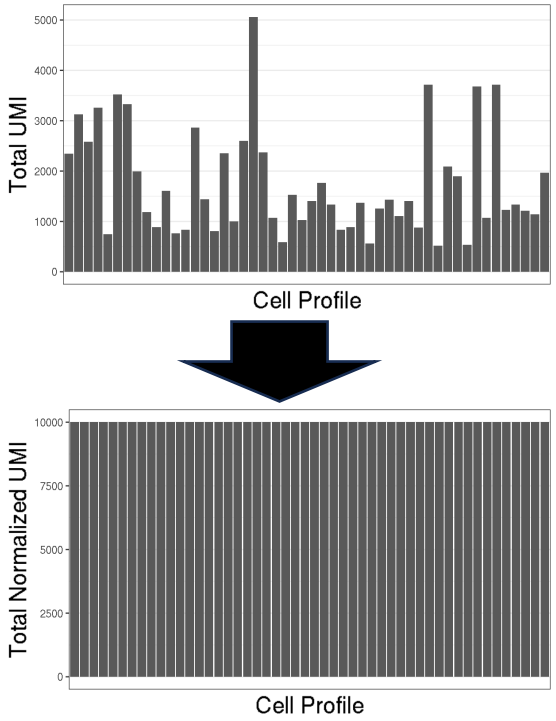


Data Normalization

Data Normalization

Normalization → Transformation → Scaling/Centering

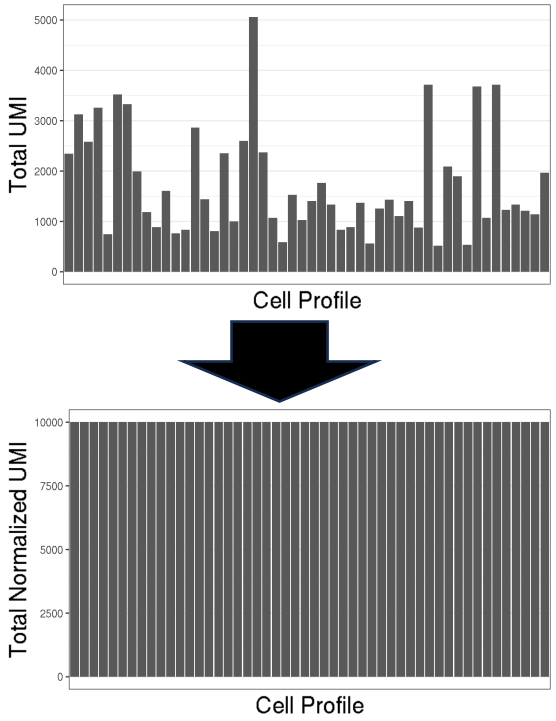
Gene counts are relative to the UMI counts per samples

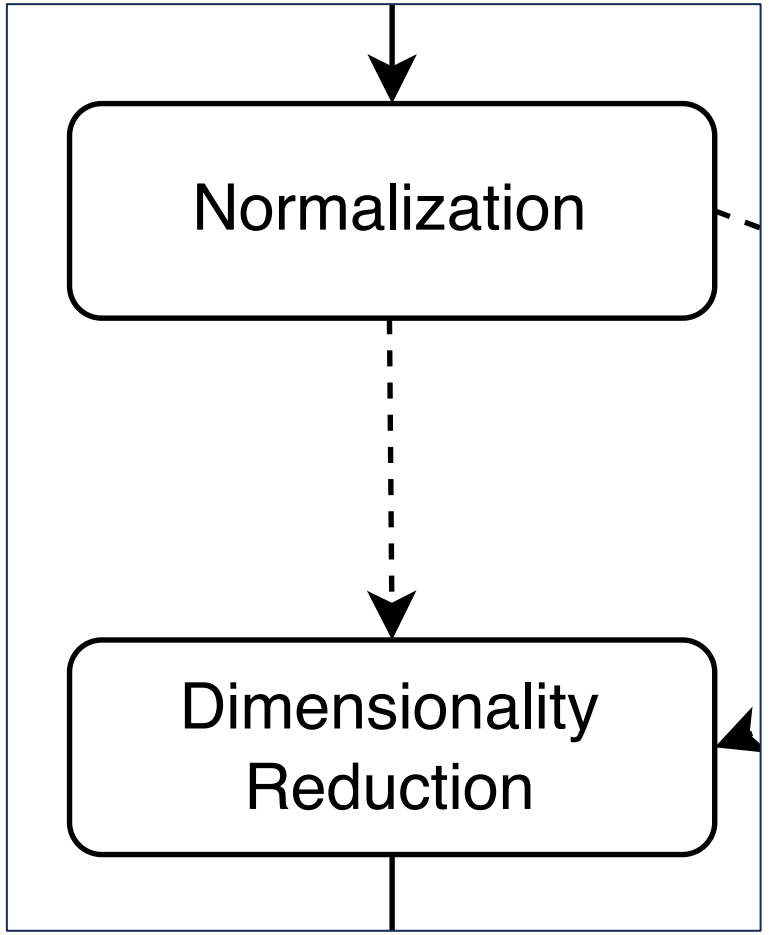


Data Normalization

Normalization → Transformation → Scaling/Centering

Gene counts are relative to the UMI counts per samples





Dimensionality Reduction

1. **Feature Extraction**
   - **Reducing the dimensionality of the data, while preserving sources of variability**
   - **Improves the performance of machine learning methods speeds up computational time for modeling the data**
     - **Requirement for many scRNAseq clustering algorithms**

2. Data Visualization
   - Reducing the dimensionality of the data, while preserving the relative distances between cells
   - Allows us to better visualize the relatively similarity between cell profiles

Dimensionality Reduction

Feature Extraction

Method

- *Principal Component Analysis (PCA)*

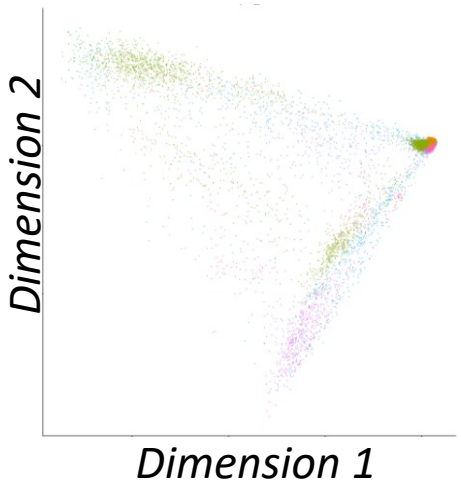
Purpose

- Reduce dimension of data set, while preserving most of its information

Input Data

- High dimensional data set

PCA



Visualization

Method

- *Uniform Manifold Approximation (UMAP)*
- *t-distributed Stochastic Neighbor Embedding (tSNE)*

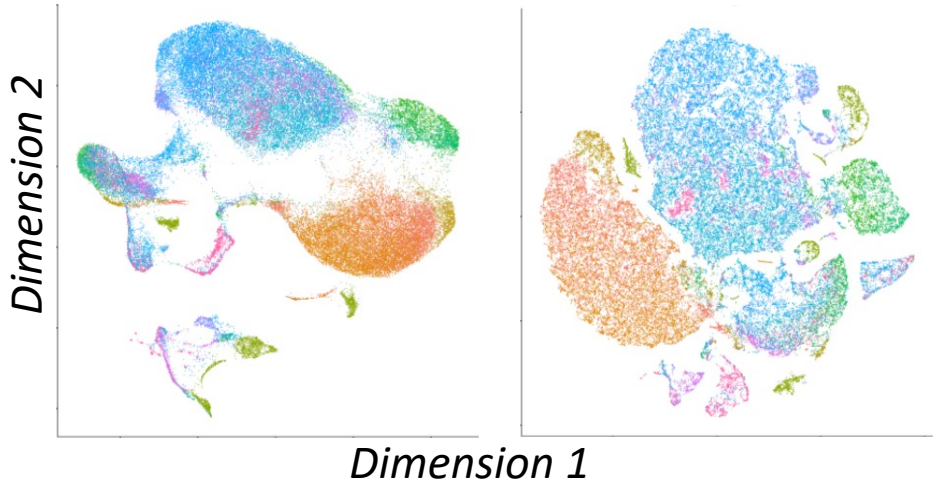
Purpose

- Visualize relationships between cells in 2-dimensions

Input Data

- PCA matrix

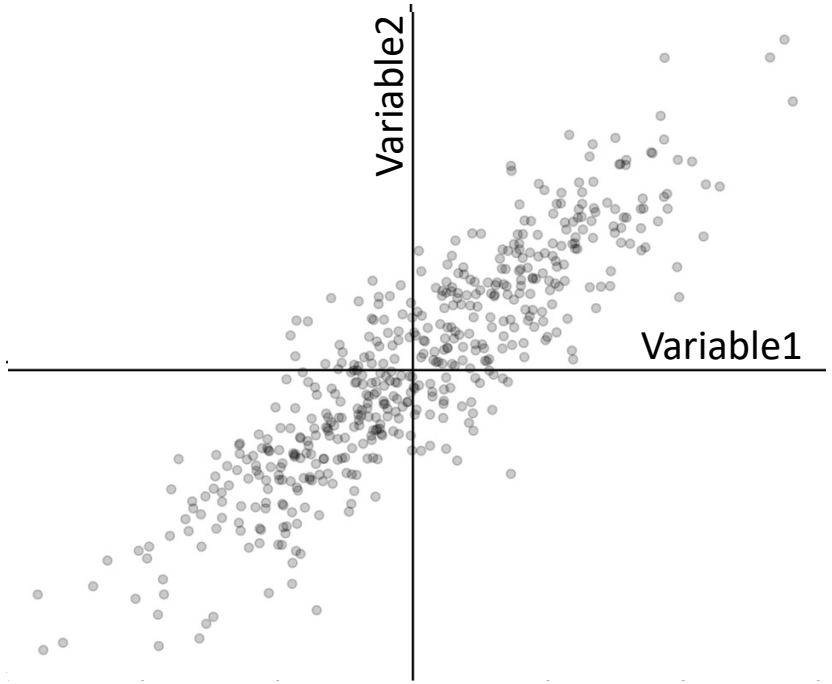
UMAP *tSNE*



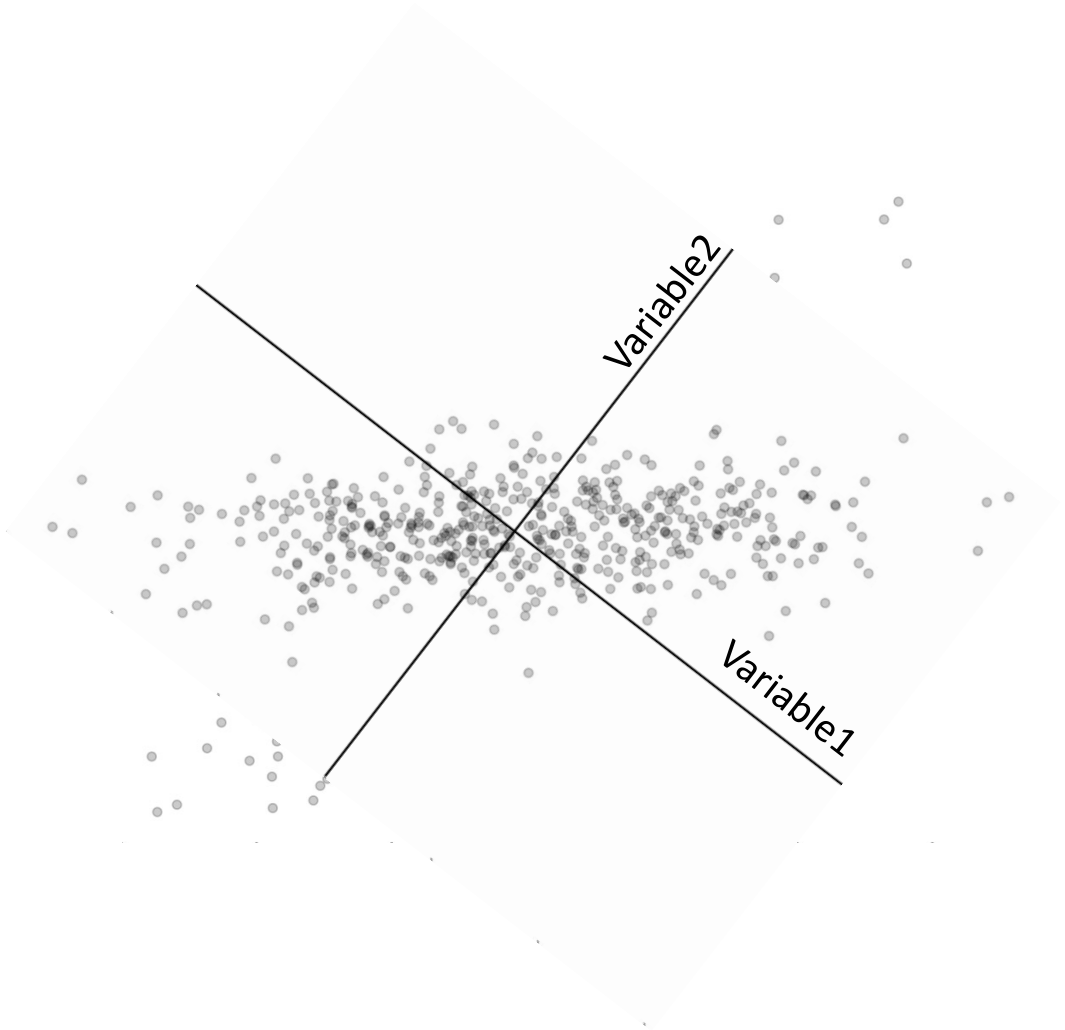
Principal Component Analysis

2-Dimensional Example

Variance(Variable 1) = 1
Variance(Variable 2) = 1



Principal Component Analysis

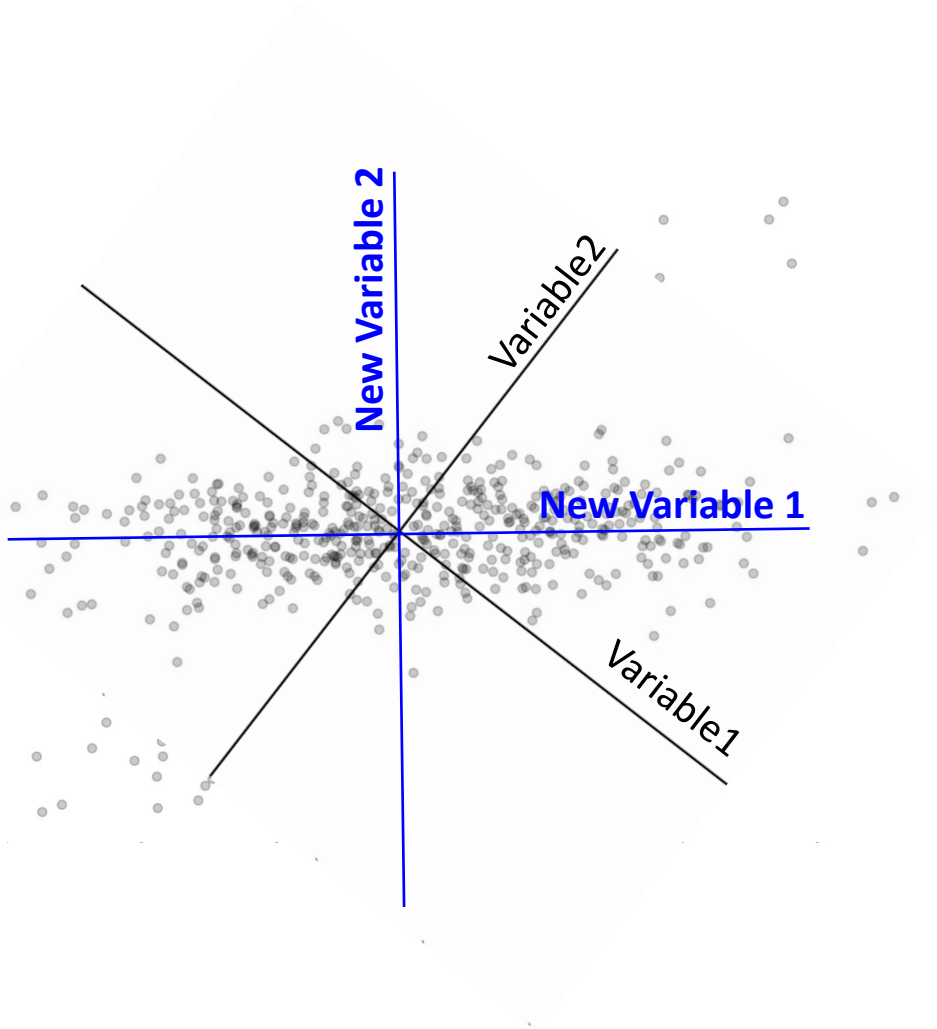


Principal Component Analysis

2-Dimensional Example

Variance(Variable 1) = 1
Variance(Variable 2) = 1

New Variable 1 ~ Variance = > 1
New Variable 2 ~ Variance = < 1

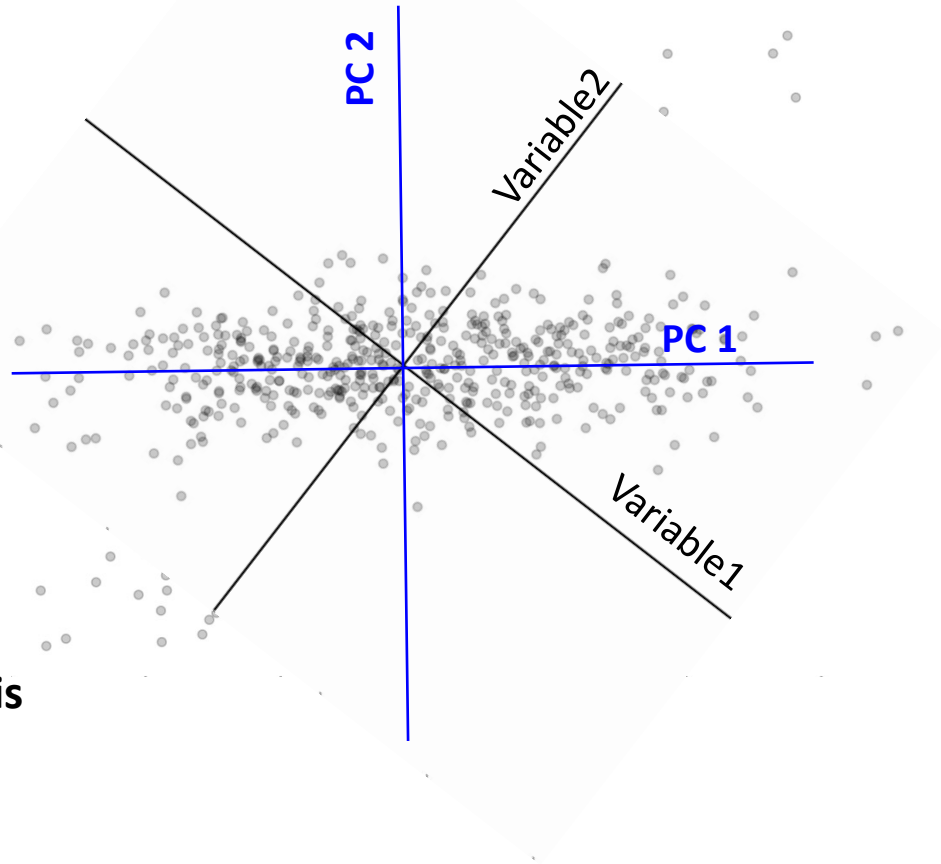


Principal Component Analysis

2-Dimensional Example

Variance(Variable 1) = 1
Variance(Variable 2) = 1

Prin. Comp. 1 ~ Variance = > 1
Prin. Comp. 2 ~ Variance = < 1



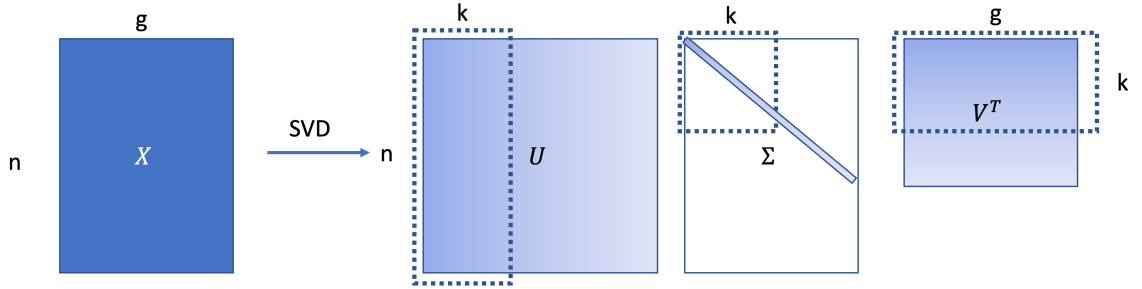
- **PC1 and PC2 preserve all variability**
- **The majority of the variance in the data is explained by PC1**
- **PC1 and PC2 are orthogonal (90° angle)**

Principal Component Analysis (Singular Value Decomposition)

https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

Principal Component Analysis (Singular Value Decomposition)

$$X = U\Sigma V^T$$



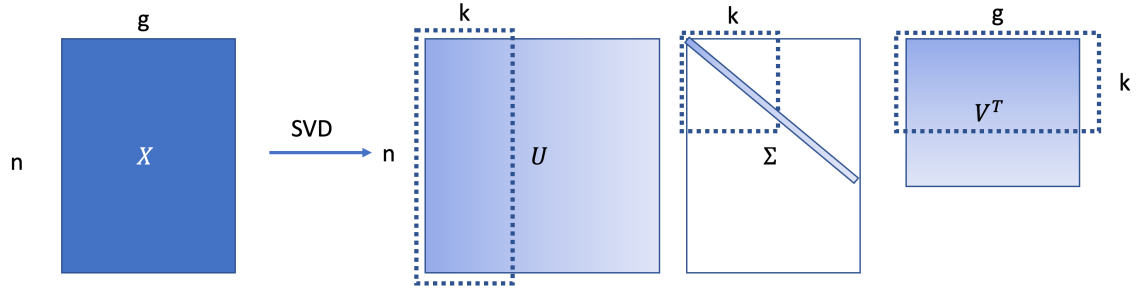
n: Number of
g: Number of genes
k: Number of PCs

X: Data Matrix
U: Matrix of cell profile PCs
V: Matrix of gene “loadings”
Σ: Standard deviation of each PC

https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

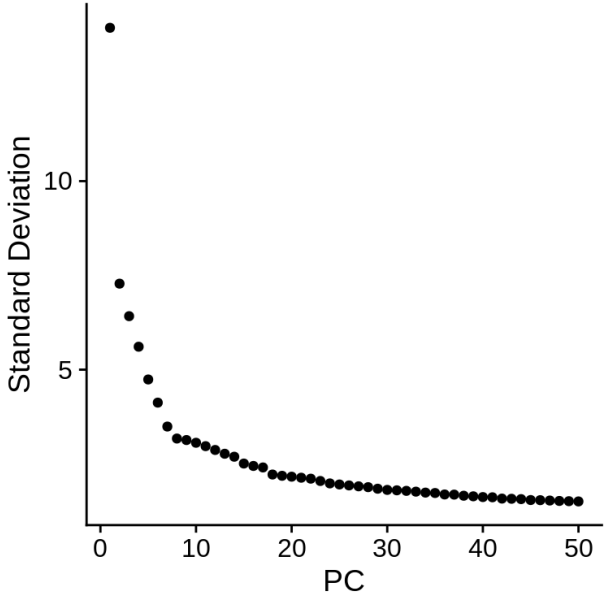
Principal Component Analysis (Singular Value Decomposition)

$$X = U\Sigma V^T$$



n: Number of
g: Number of genes
k: Number of PCs

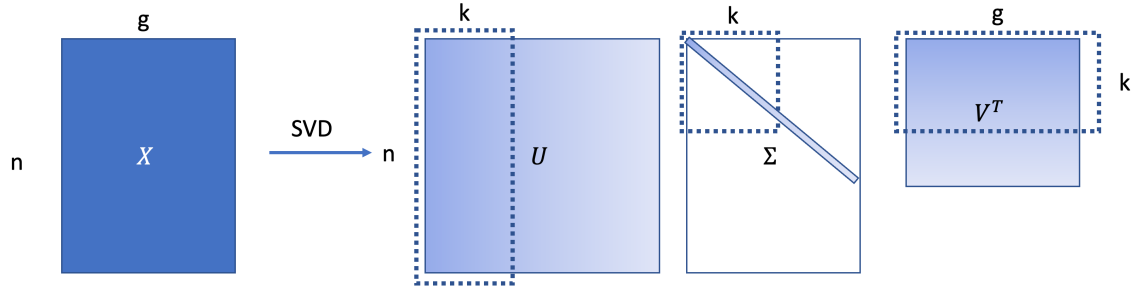
X: Data Matrix
U: Matrix of cell profile PCs
V: Matrix of gene “loadings”
Σ: Standard deviation of each PC



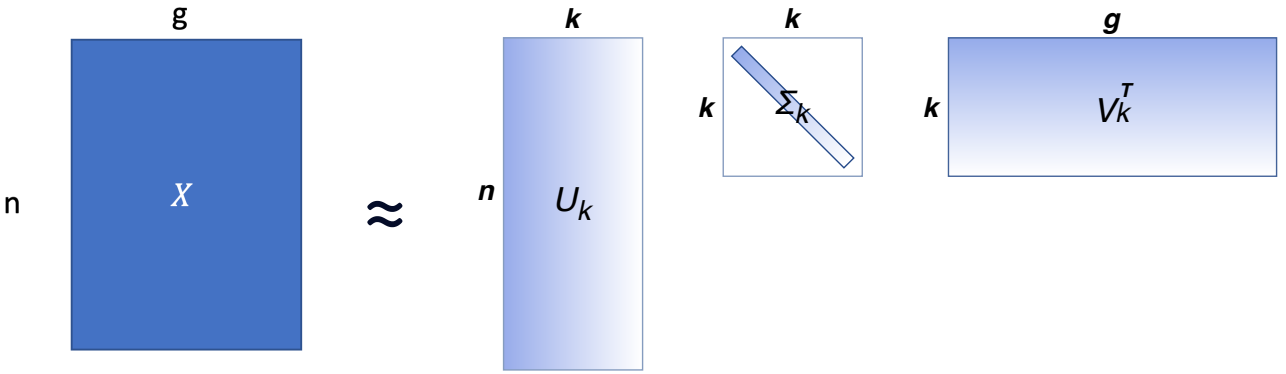
https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

Principal Component Analysis (Singular Value Decomposition)

$$X = U\Sigma V^T$$

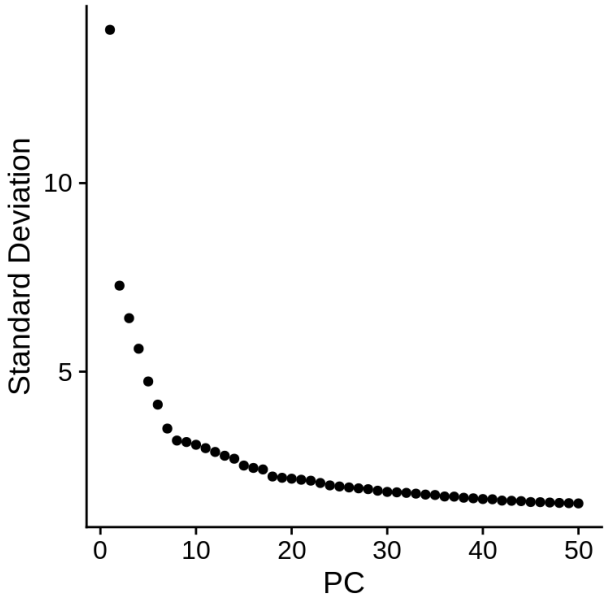


$$X \approx U_k \Sigma_k V_k^T$$



n: Number of
g: Number of genes
k: Number of PCs

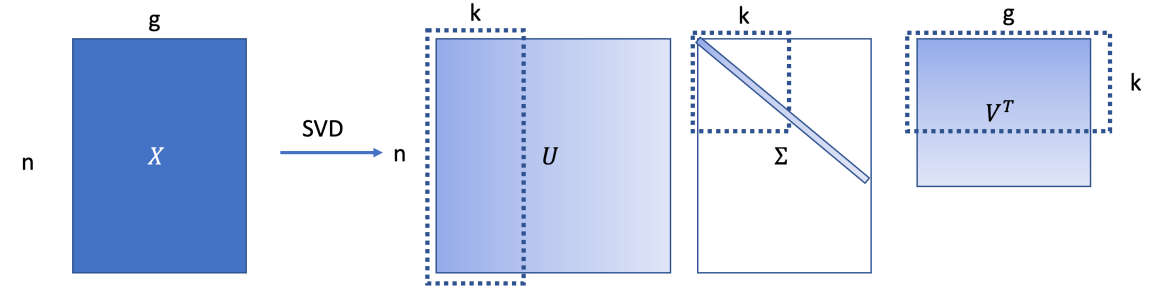
X: Data Matrix
U: Matrix of cell profile PCs
V: Matrix of gene “loadings”
Σ: Standard deviation of each PC



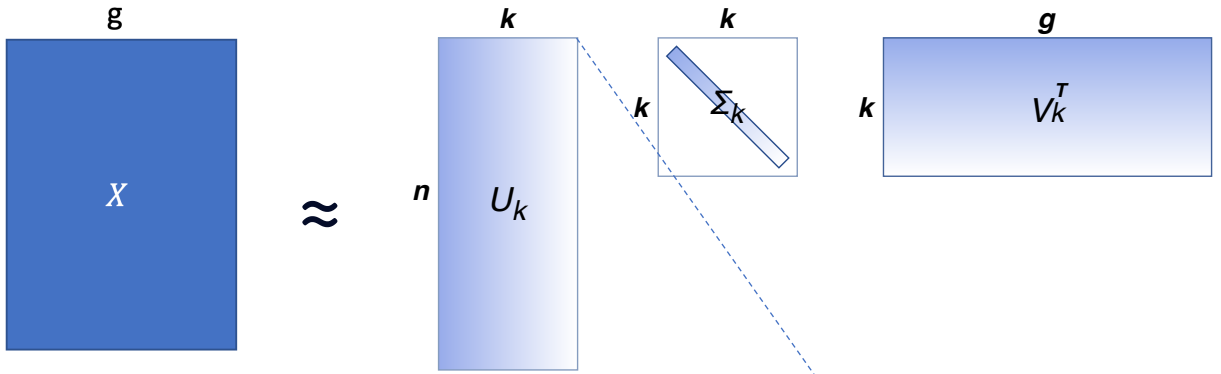
https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

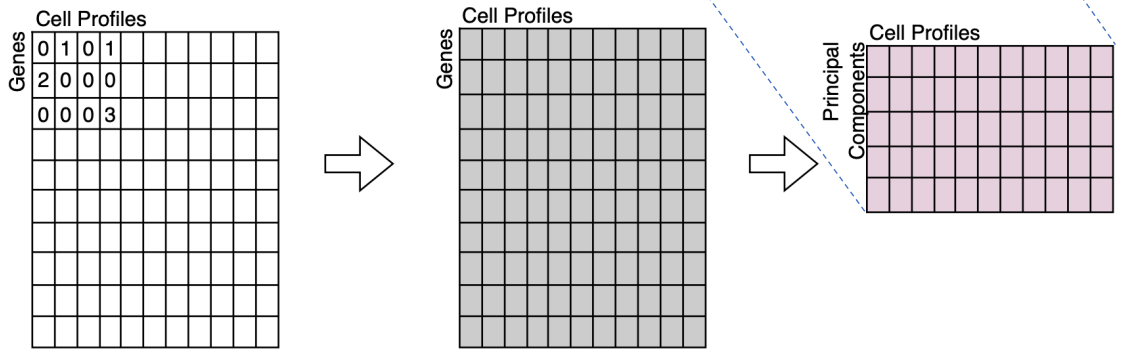
Principal Component Analysis (Singular Value Decomposition)

$$X = U\Sigma V^T$$



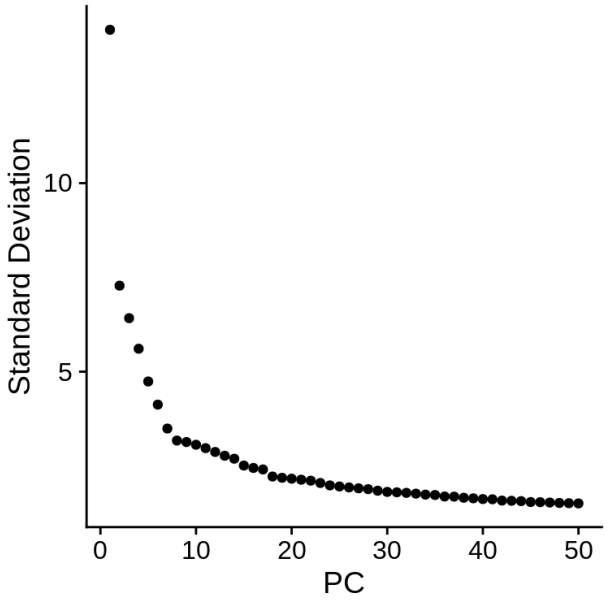
$$X \approx U_k \Sigma_k V_k^T$$





n: Number of
g: Number of genes
k: Number of PCs

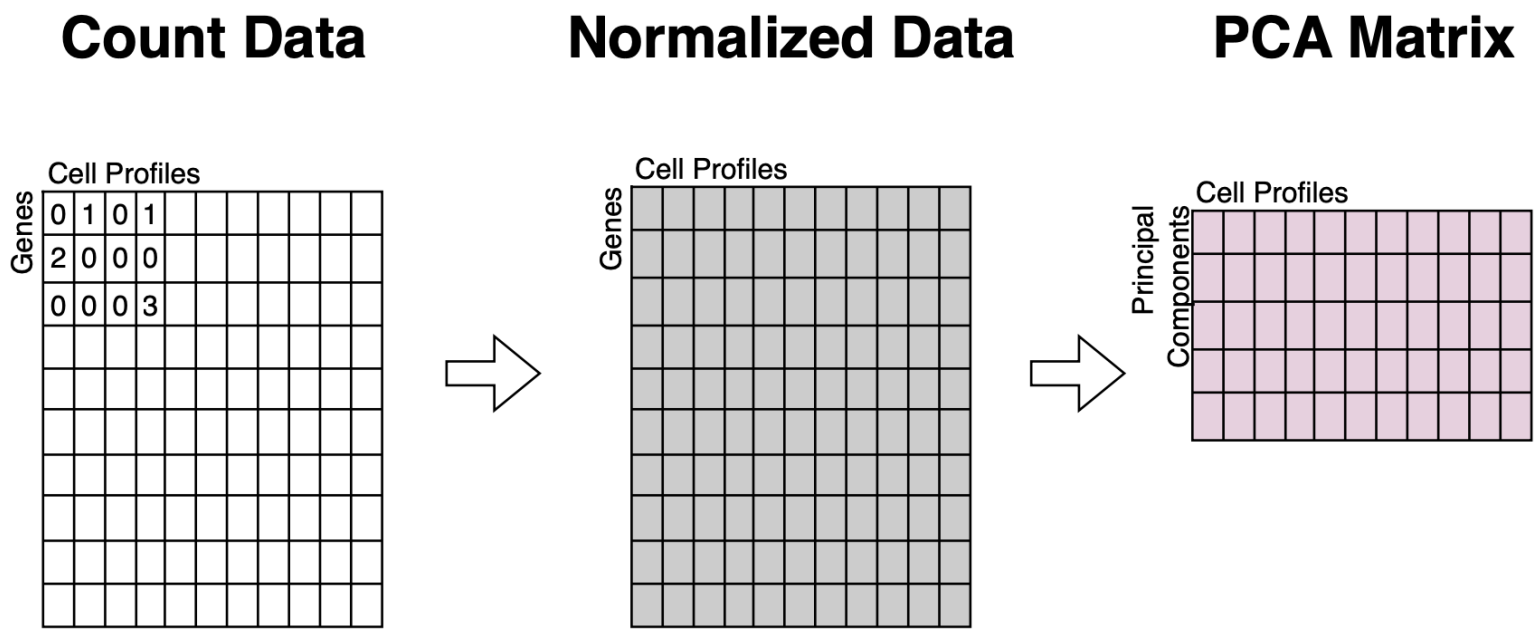
X: Data Matrix
U: Matrix of cell profile PCs
V: Matrix of gene "loadings"
Σ: Standard deviation of each PC



https://xinmingtu.cn/blog/2022/CCA_dual_PCA/

Principal Component Analysis

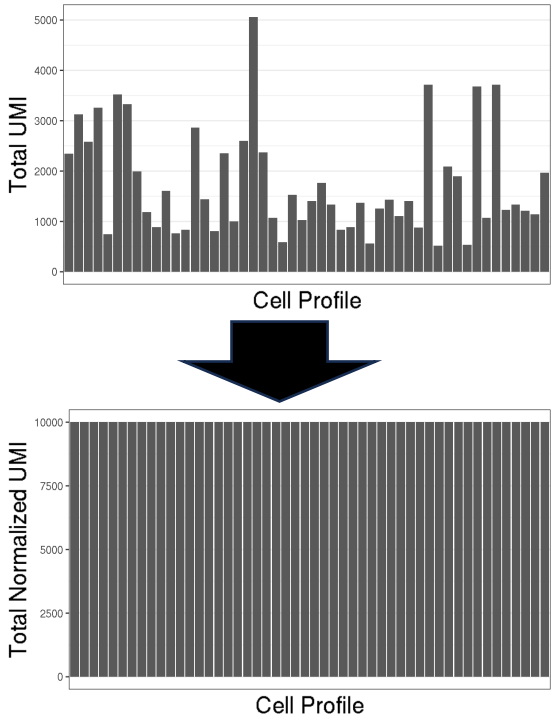
- Principal Component Analysis requires that input data is centered at mean 0
  - The “scale” of the data matters
    - Genes with higher variance have greater influence on PCA

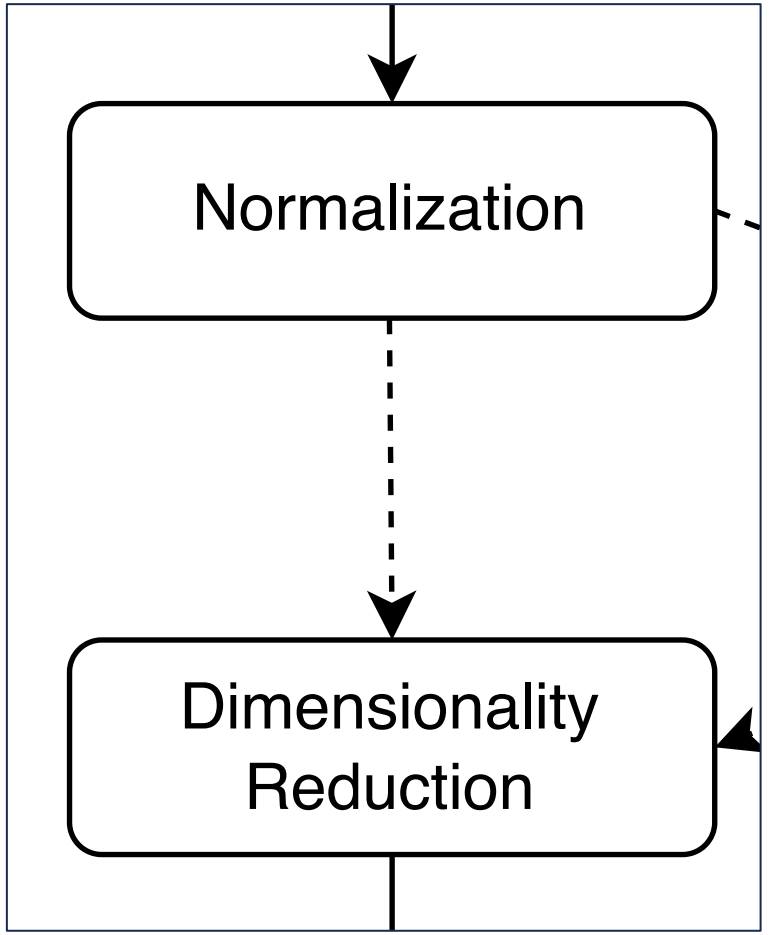


Data Normalization

Normalization → Transformation → Scaling/Centering

Gene counts are relative to the UMI counts per samples

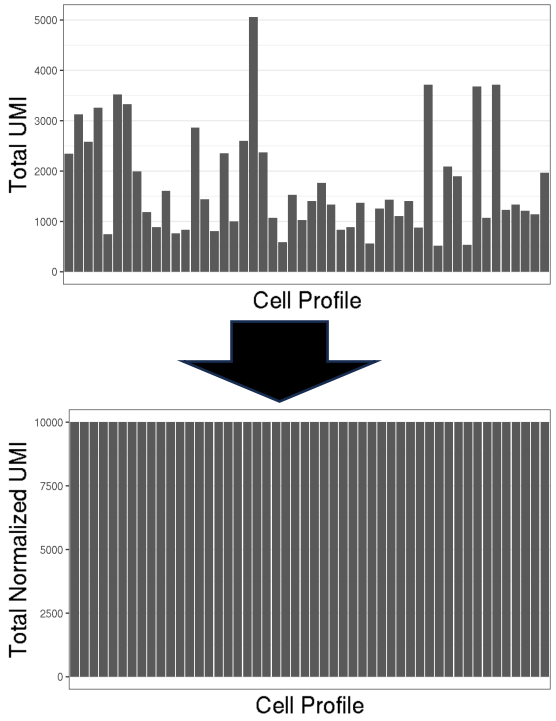




Data Normalization (Simple)

Normalization → Transformation → Scaling/Centering

Gene counts are relative to the UMI counts per samples

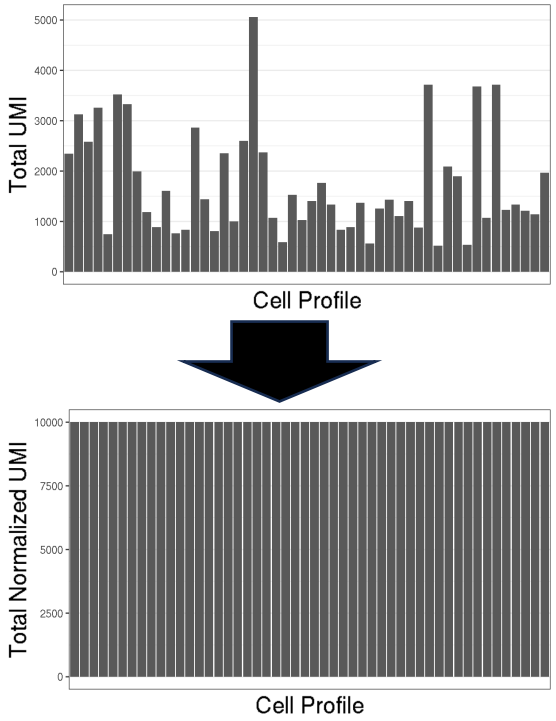


$$\frac{\text{UMI Gene Count}}{Total\ UMI\ Per\ Sample} * 10000$$

Data Normalization (Simple)

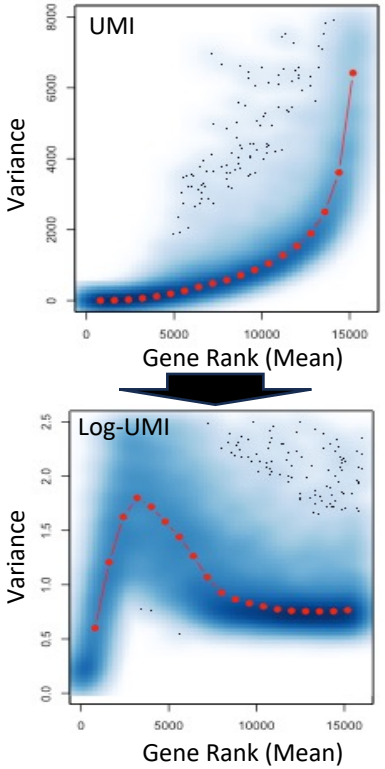
Normalization → Transformation → Scaling/Centering

Gene counts are relative to the UMI counts per samples



$$\frac{\text{UMI Gene Count}}{Total\ UMI\ Per\ Sample} * 10000$$

The variance of UMI counts depend on their value

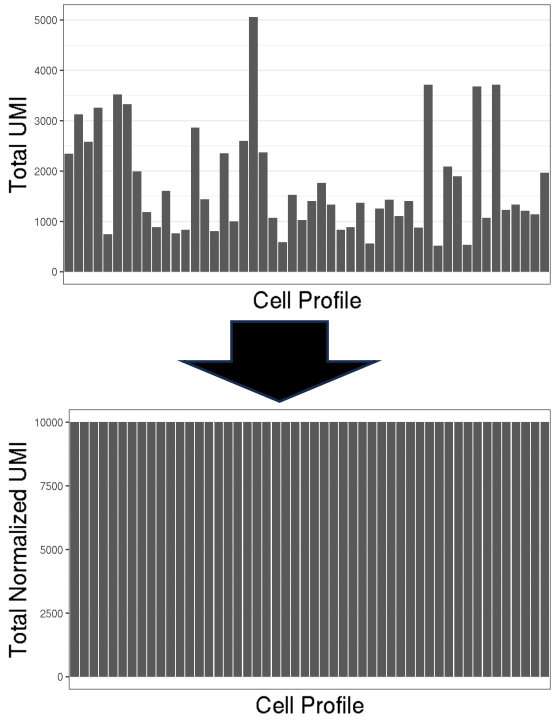


$$Log - Transform$$

Data Normalization (Simple)

Normalization

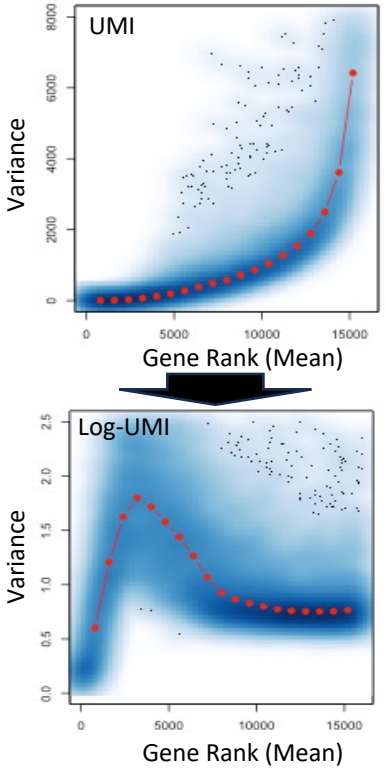
Gene counts are relative to the UMI counts per samples



$$\frac{\text{UMI Gene Count}}{Total\ UMI\ Per\ Sample} * 10000$$

Transformation

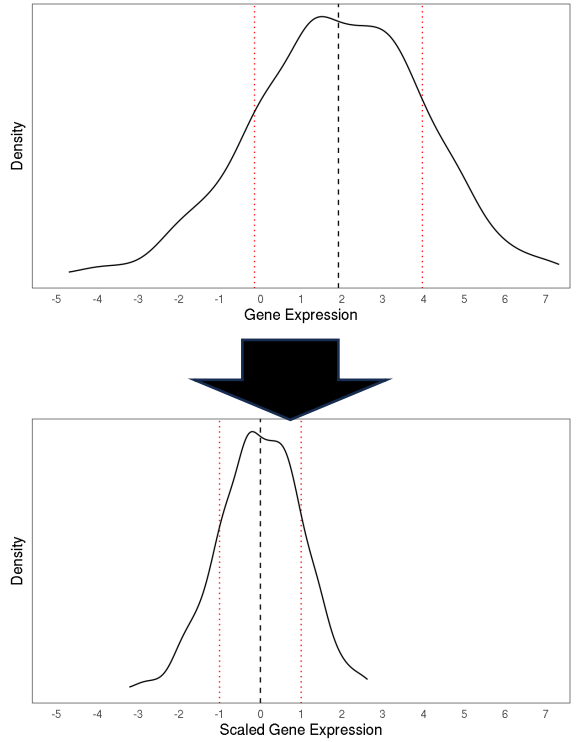
The variance of UMI counts depend on their value



$$Log - Transform$$

Scaling/Centering

- PCA requires features have mean of 0
- Force equal influence of each feature



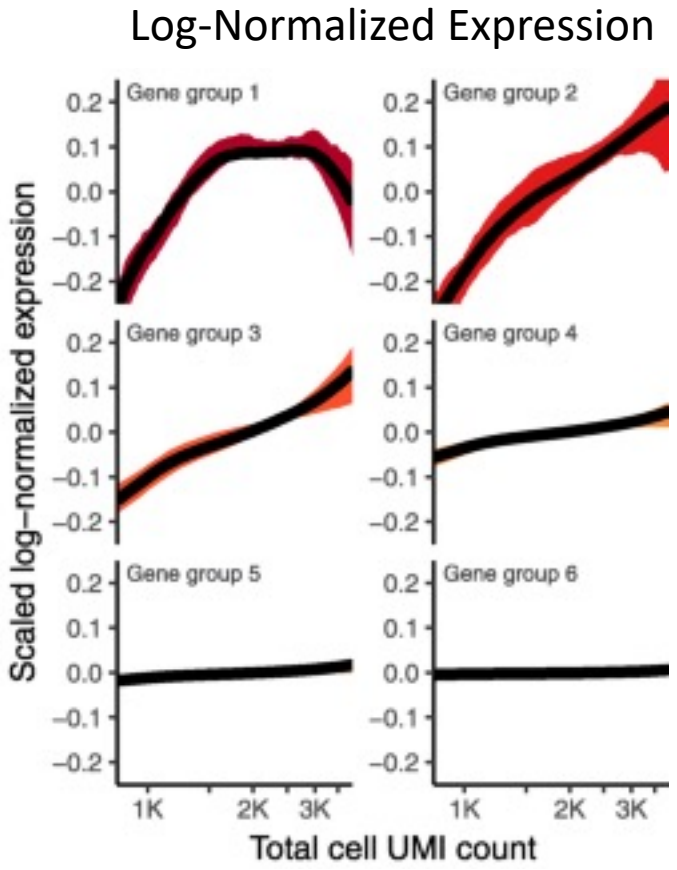
$$Center\ and\ scale\ to$$
$$mean = 0$$
$$standard\ deviation = 1$$

Data Normalization (Regularized Negative Binomial Regression)

Hafemeister and Satija, 2019

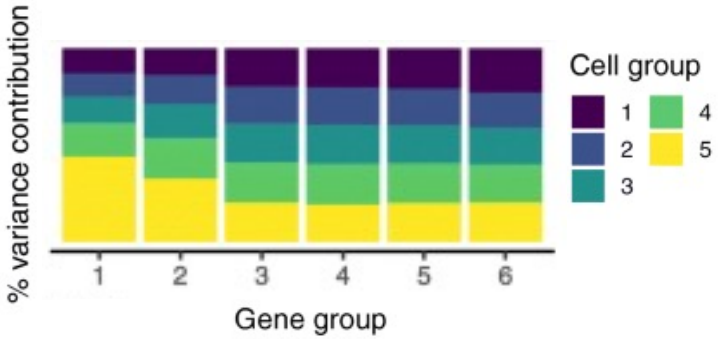
After our simple log-normalization

- Gene expression is correlated to total UMI for highly expressed genes
- Highly expressed genes disproportionately contribute to the variance of cells with low UMI counts



Gene Group 1: Highly Expressed
Gene Group 2: Lowly Expressed

Cell Group 1: High Total UMI
Cell Group 5: Low Total UMI



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1

Data Normalization (Regularized Negative Binomial Regression)

Gene Group 1: Highly Expresse
Gene Group 2: Lowly Expressed

Cell Group 1: High Total UMI
Cell Group 5: Low Total UMI

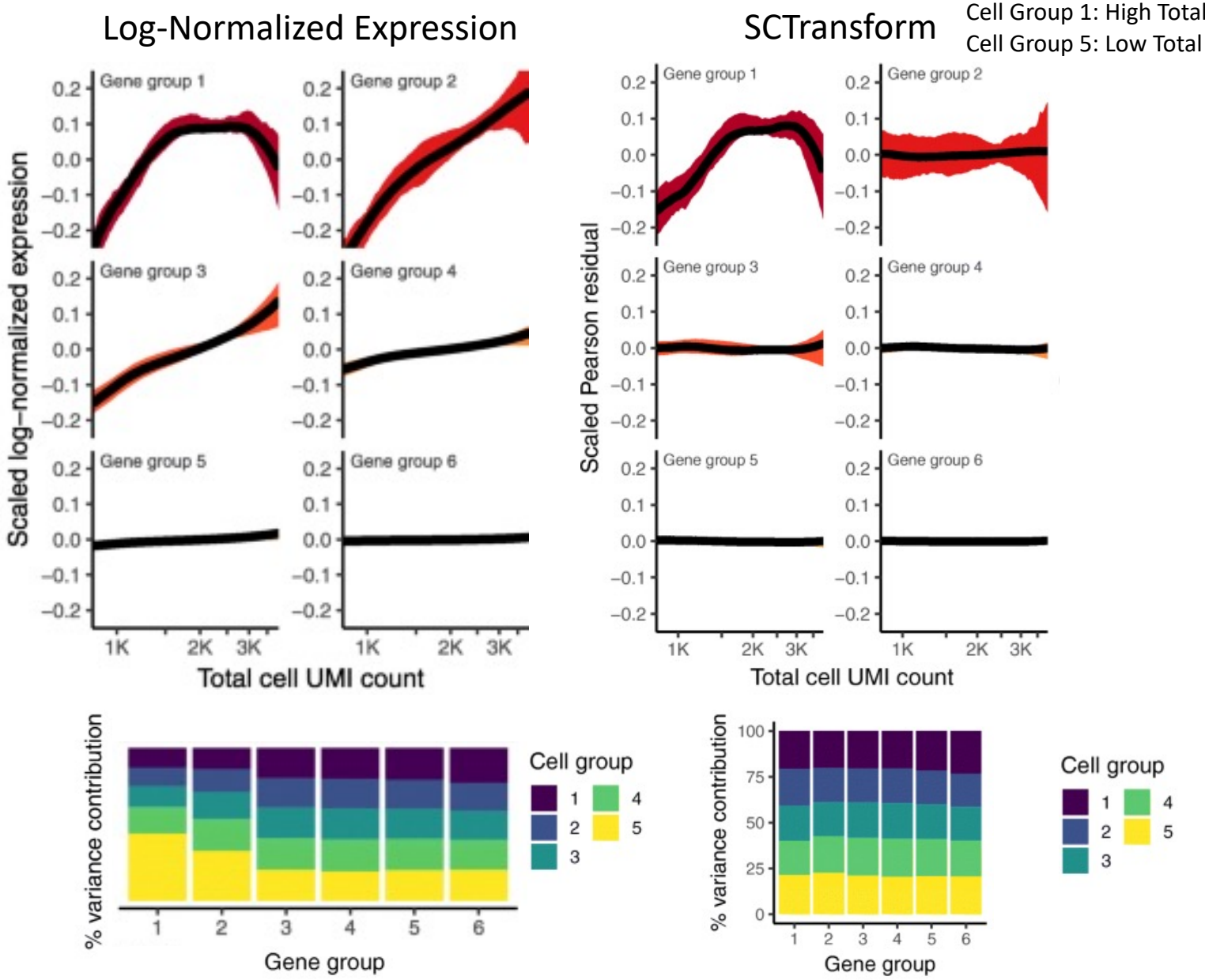
Hafemeister and Satija, 2019

After standard log-normalization

- Gene abundance is correlated to total UMI for highly expressed genes
- Highly expressed genes disproportionately contribute to the variance of cells with low UMI counts

Proposed "Regularized Negative Binoimal Regression"

- Implemented by SCTransform() Seurat function
- Does not assume fixed expected total counts per cell



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1

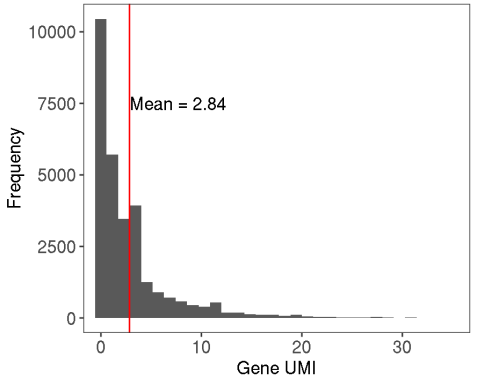
Data Normalization (Regularized Negative Binomial Regression)

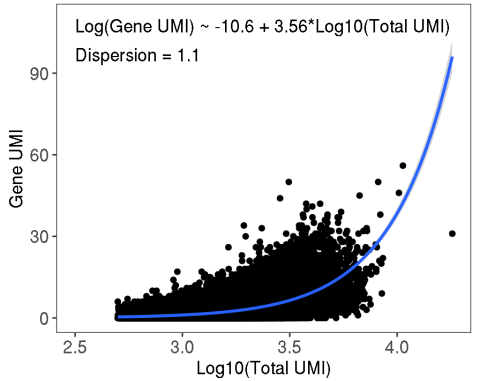
Data Normalization (Regularized Negative Binomial Regression)

1. For each gene, fit Negative Binomial Model

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m$$

- x_i = Vector of UMI counts for gene, i,
- m = Vector of UMI counts for each sample





https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1#Sec9

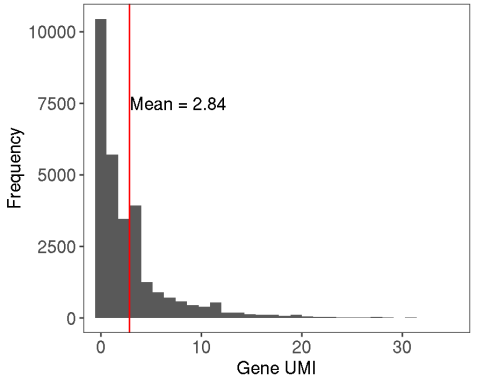
Data Normalization (Regularized Negative Binomial Regression)

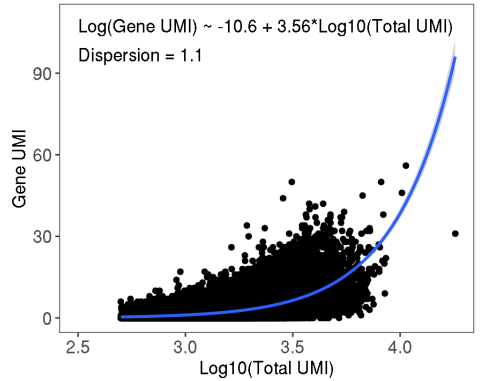
1. For each gene, fit Negative Binomial Model

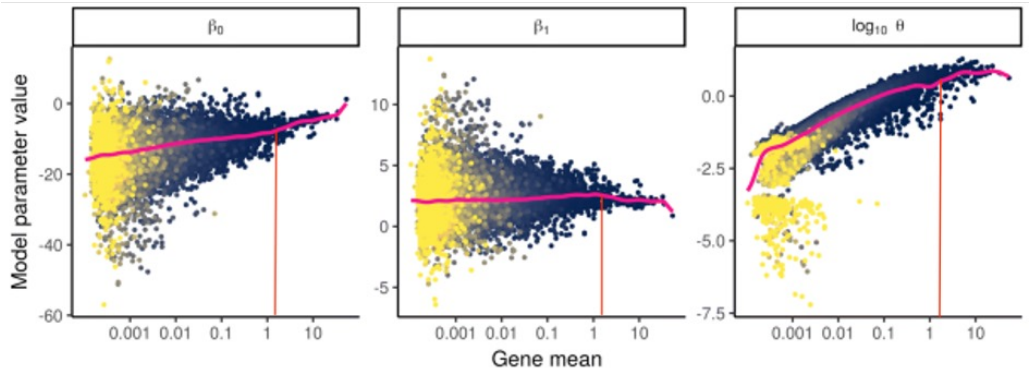
$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m$$

- X_i = Vector of UMI counts for gene, i,
- m = Vector of UMI counts for each sample

2. "Regularize" fitted parameters
   - Estimate parameters, β_0, β_1, and θ, for each gene based on their mean







https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1#Sec9

Data Normalization (Regularized Negative Binomial Regression)

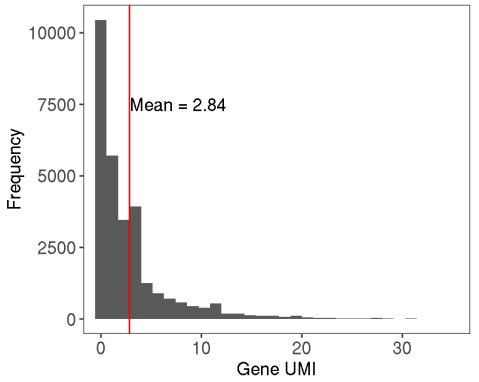
1. For each gene, fit Negative Binomial Model

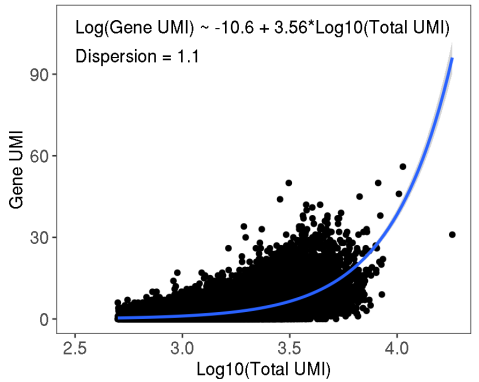
$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m$$

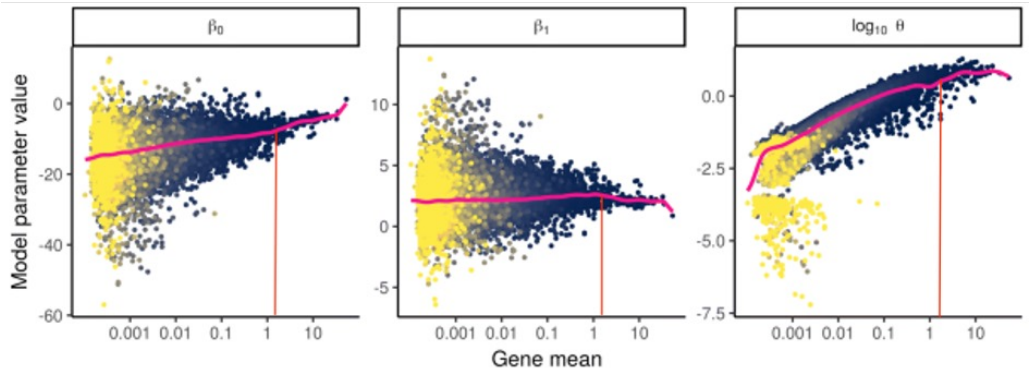
- X_i = Vector of UMI counts for gene, i,
- m = Vector of UMI counts for each sample

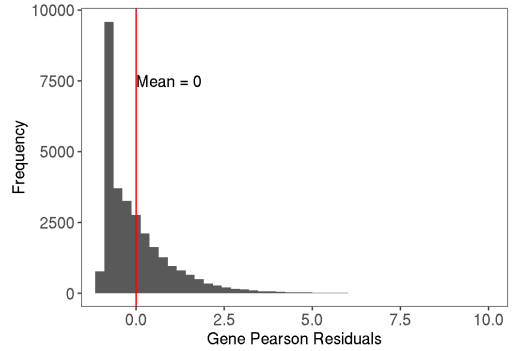
2. "Regularize" fitted parameters
   - Estimate parameters, β_0, β_1, and θ, for each gene based on their mean

3. Obtain Pearson residuals for each gene based on negative binomial model with "regularized" parameter estimates





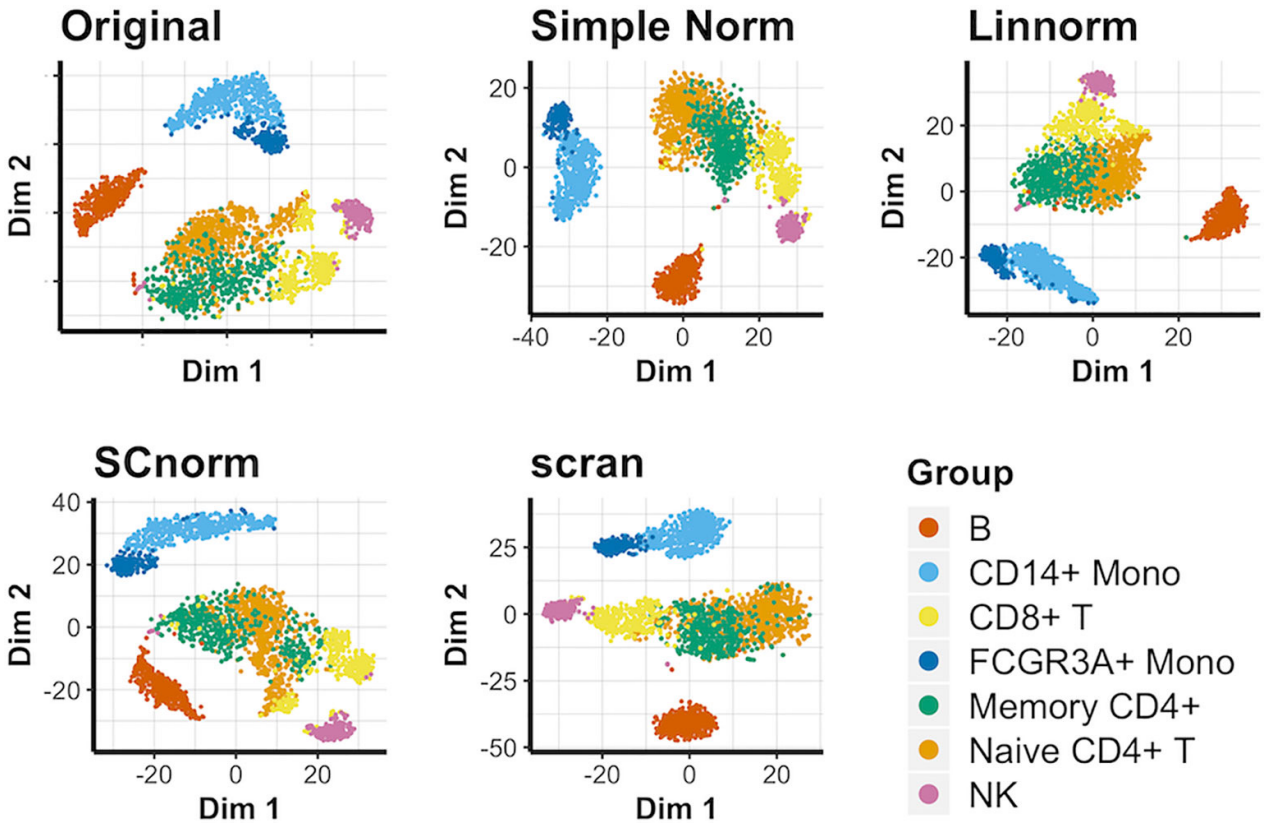




https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1874-1#Sec9

Data Normalization

t-SNE plots for Human PBMC scRNAseq Data



frontiers in Genetics

Front Genet. 2020; 11: 41.
Published online 2020 Feb 7. doi: 10.3389/fgene.2020.00041
PMCID: PMC7019105
PMID: 32117453

Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey

Nicholas Lytal, [1] Di Ran, [2] and Lingling An [1,2,3,*]

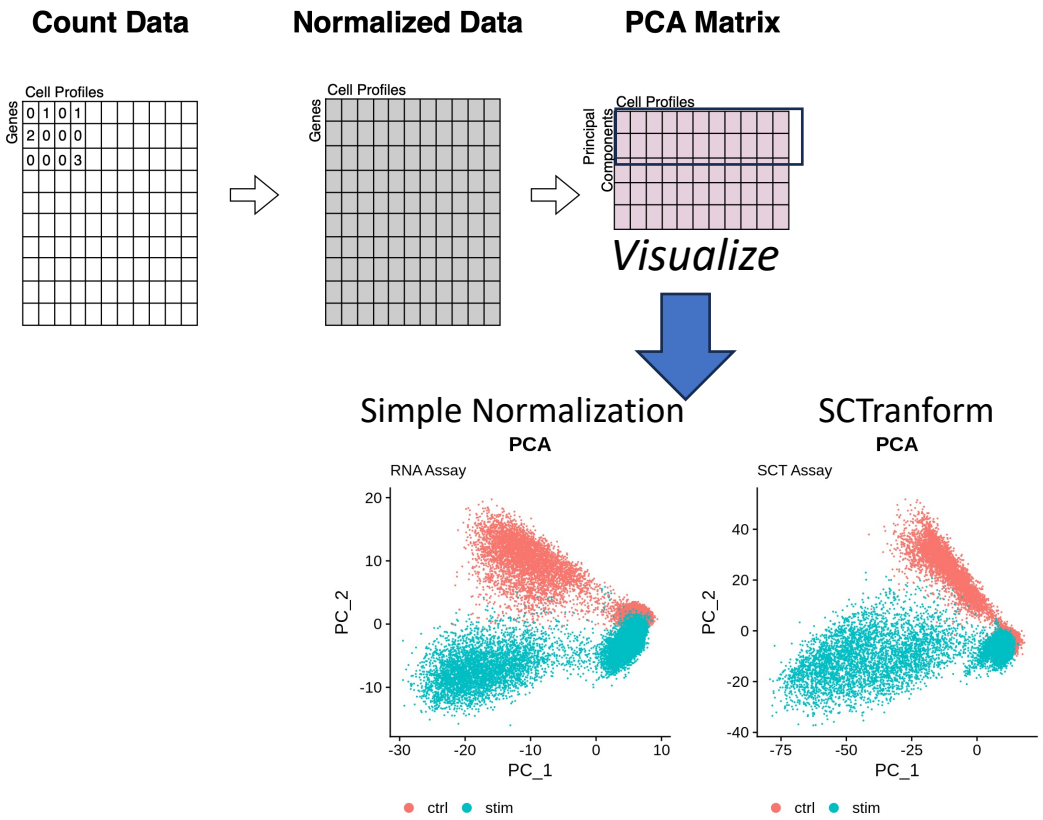
"Despite these slight advantages, comparisons to the simple normalization process built in to Seurat reveals that even exceptional methods do not greatly distinguish themselves from a more straightforward normalization approach."

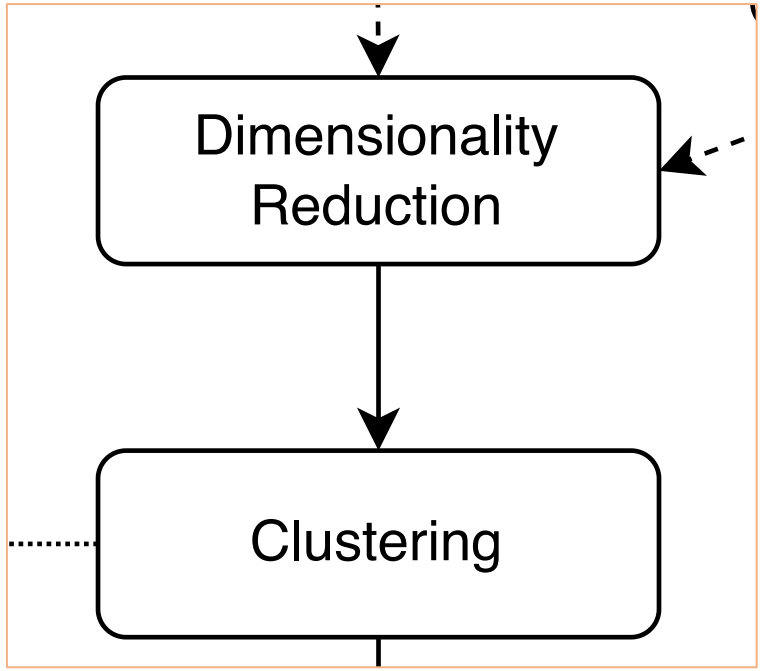
https://doi.org/10.3389/fgene.2020.00041

Dimensionality Reduction (Visualization)

A PCA matrix will be the input data for our clustering procedure.

Before clustering, we can use dimensionality reduction to inspect our data for sources of unwanted variability





Dimensionality Reduction

1. Feature Extraction
   - Reducing the dimensionality of the data, while preserving sources of variability
   - Improves the performance of machine learning methods speeds up computational time for modeling the data
     - Requirement for many scRNAseq clustering algorithms

2. **Data Visualization**
   - **Reducing the dimensionality of the data, while preserving the relative distances between cells**
   - **Allows us to better visualize the relatively similarity between cell profiles**

Uniform Manifold Approximation (UMAP) and **t-distributed Stochastic Neighbor Embedding (t-SNE)**

P x N

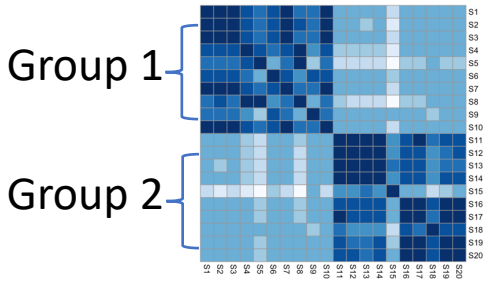
Uniform Manifold Approximation (UMAP) and **t-distributed Stochastic Neighbor Embedding (t-SNE)**

PCA
Data Set

P x N

Calculate sample-wise
distance matrix

N x N



UMAP
k-Nearest Neighbor Graph

tSNE
Euclidean Distance

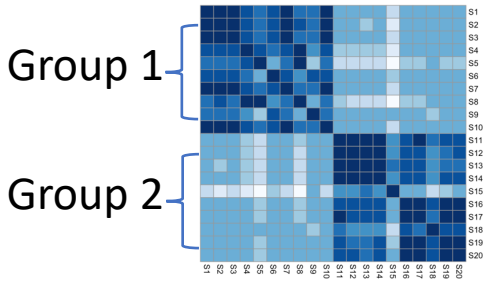
Uniform Manifold Approximation (UMAP) and **t-distributed Stochastic Neighbor Embedding (t-SNE)**

PCA
Data Set

P x N

Calculate sample-wise
distance matrix

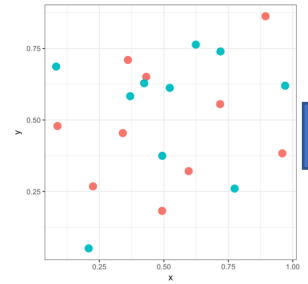
N x N

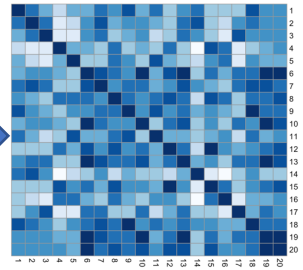


Find 2D data set with a similar
target distance matrix

2 x N

Initialization





UMAP
k-Nearest Neighbor Graph

tSNE
Euclidean Distance

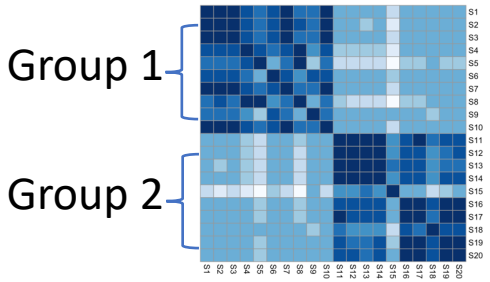
Uniform Manifold Approximation (UMAP) and **t-distributed Stochastic Neighbor Embedding (t-SNE)**

PCA
Data Set

P x N

Calculate sample-wise
distance matrix

N x N



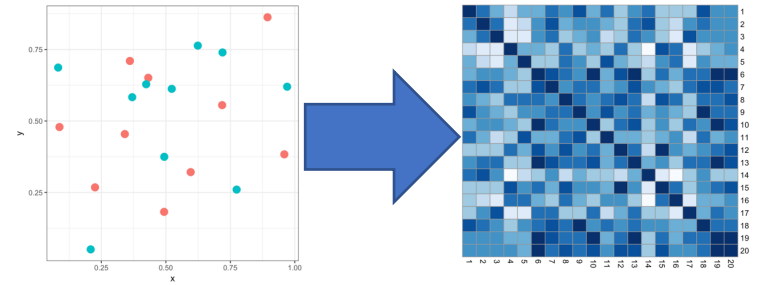
UMAP
k-Nearest Neighbor Graph

tSNE
Euclidean Distance

Find 2D data set with a similar
target distance matrix

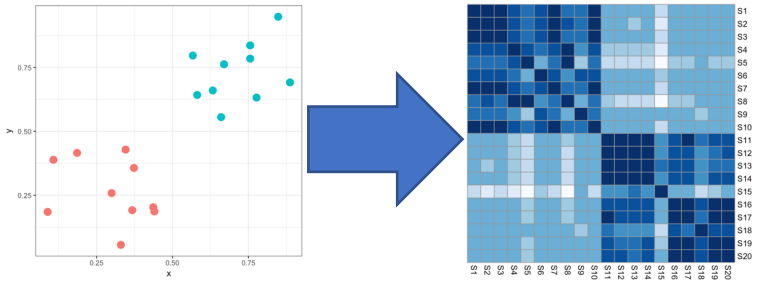
2 x N

Initialization



Optimization

Final model



Dimensionality Reduction (Visualization)

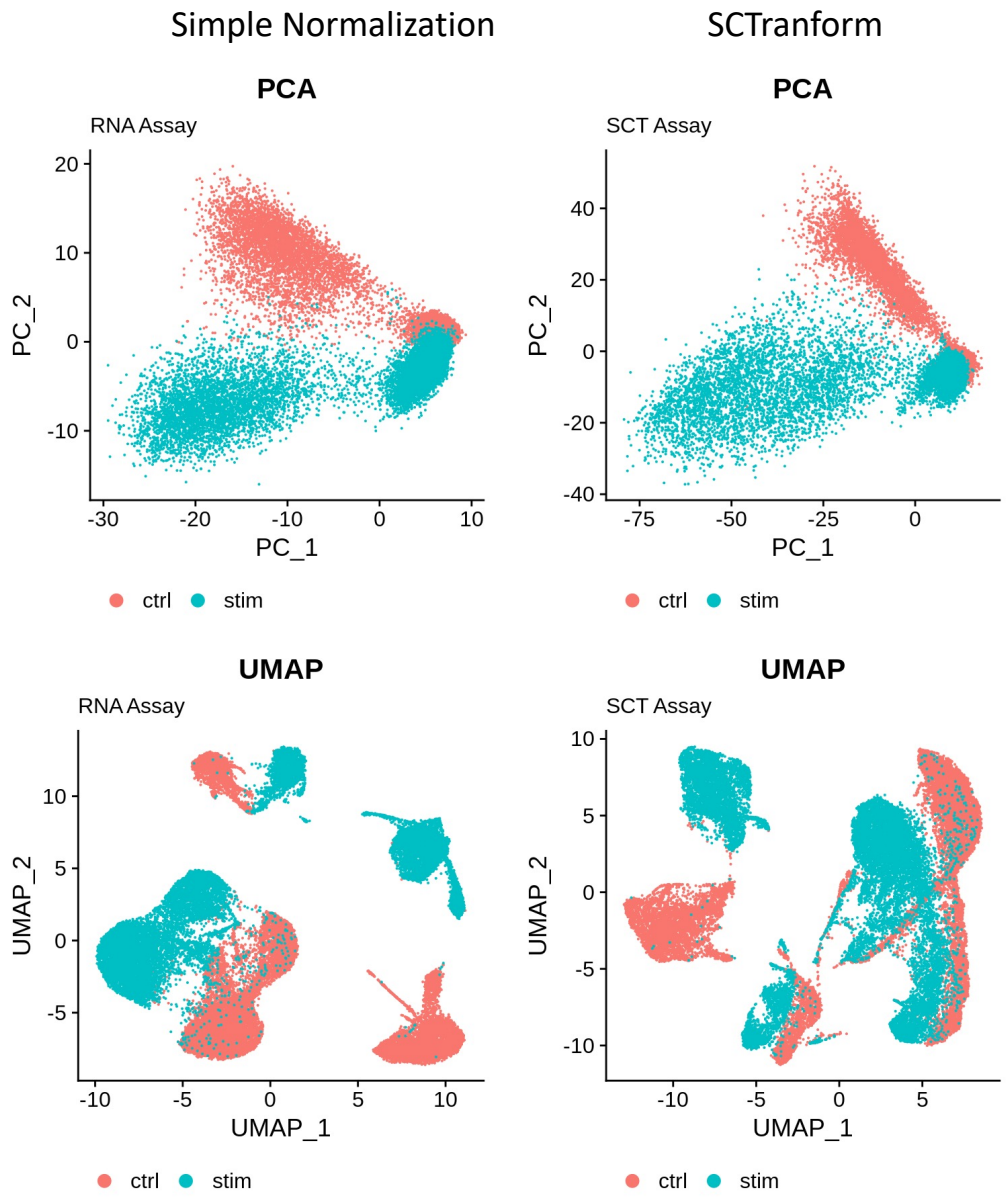
Regardless of normalization method, the majority of variability in our data is associated with differences between control and stimulated PBMCs.

- Biologically meaningful
- Technical artifact of sample preparation

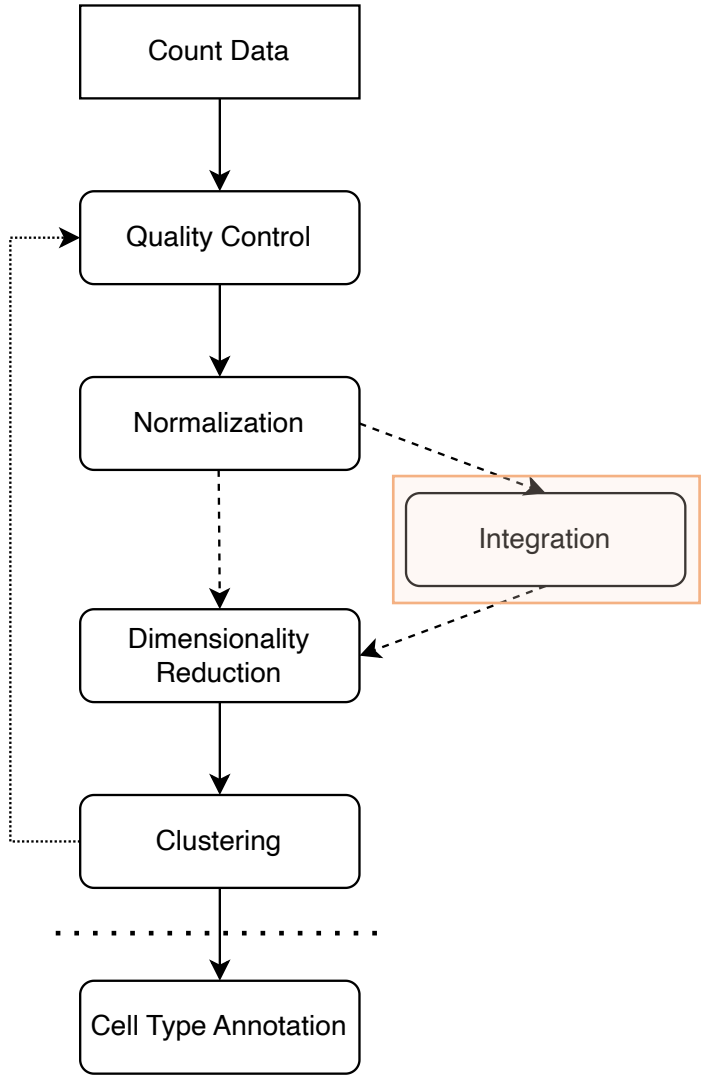
We expect many like cell types represented in both data sets

- B cells, T cells etc

What can we do to our data to make identification of these co-represented cell types easier?

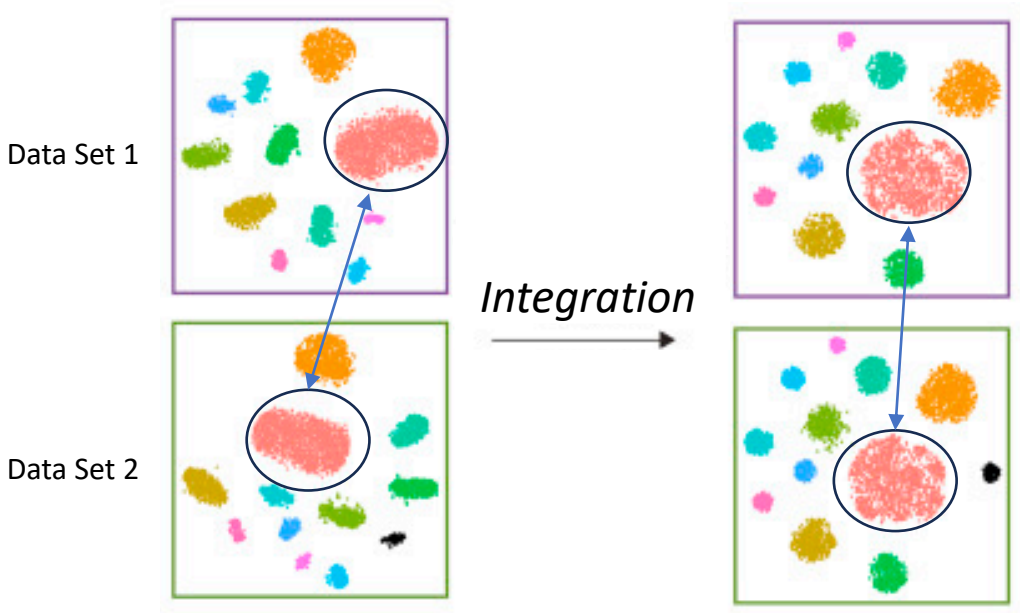


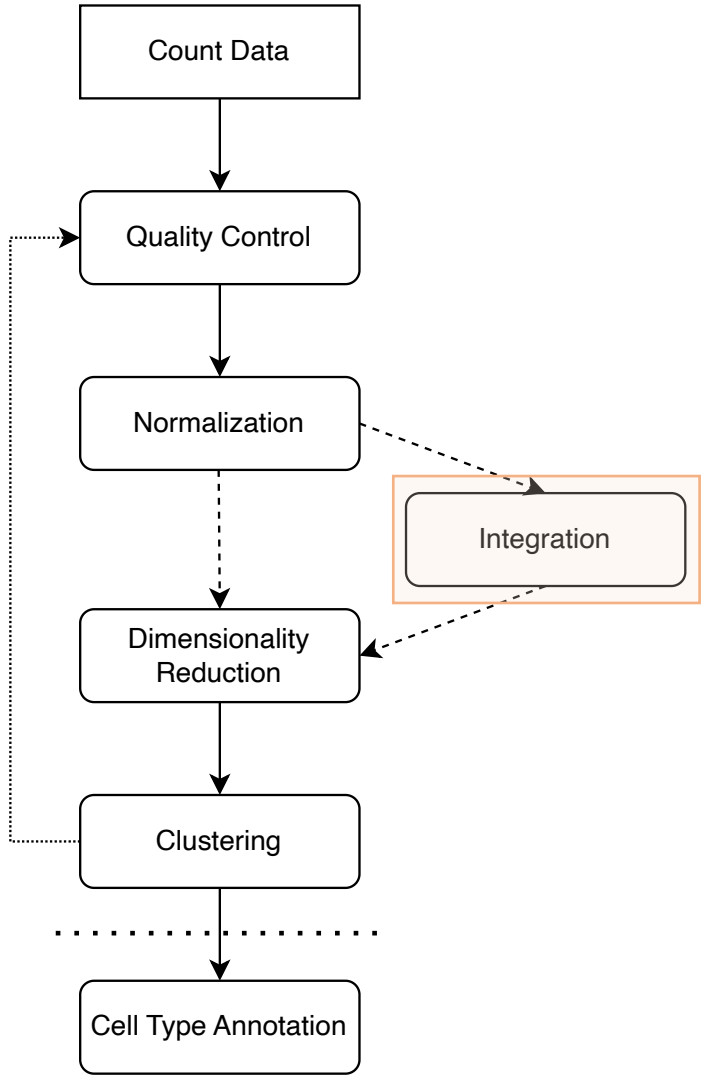
Data Integration



Data Integration

Transform our data such that the profiles of the same cell types are more similar across data sets

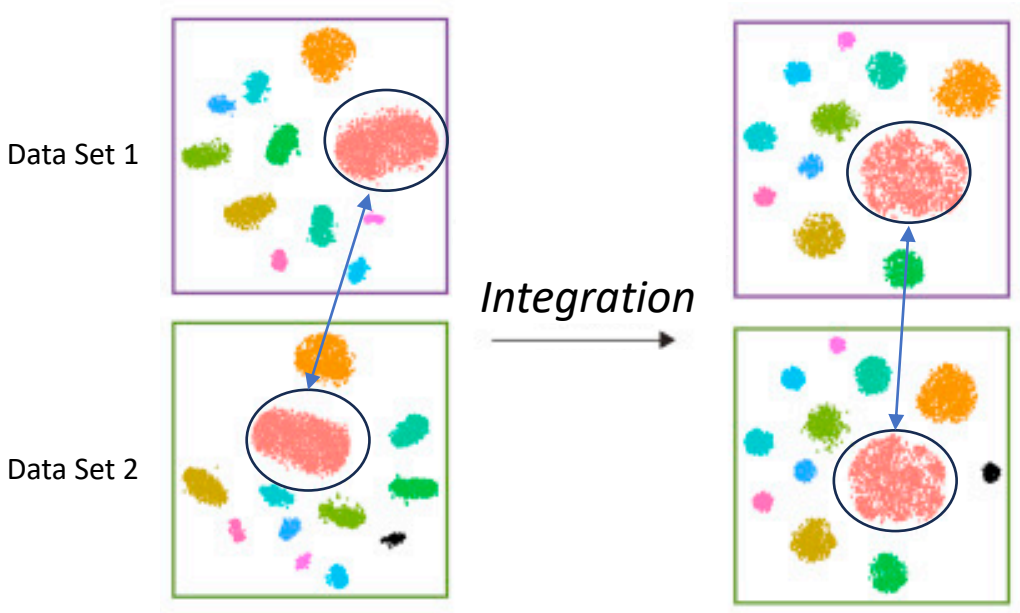


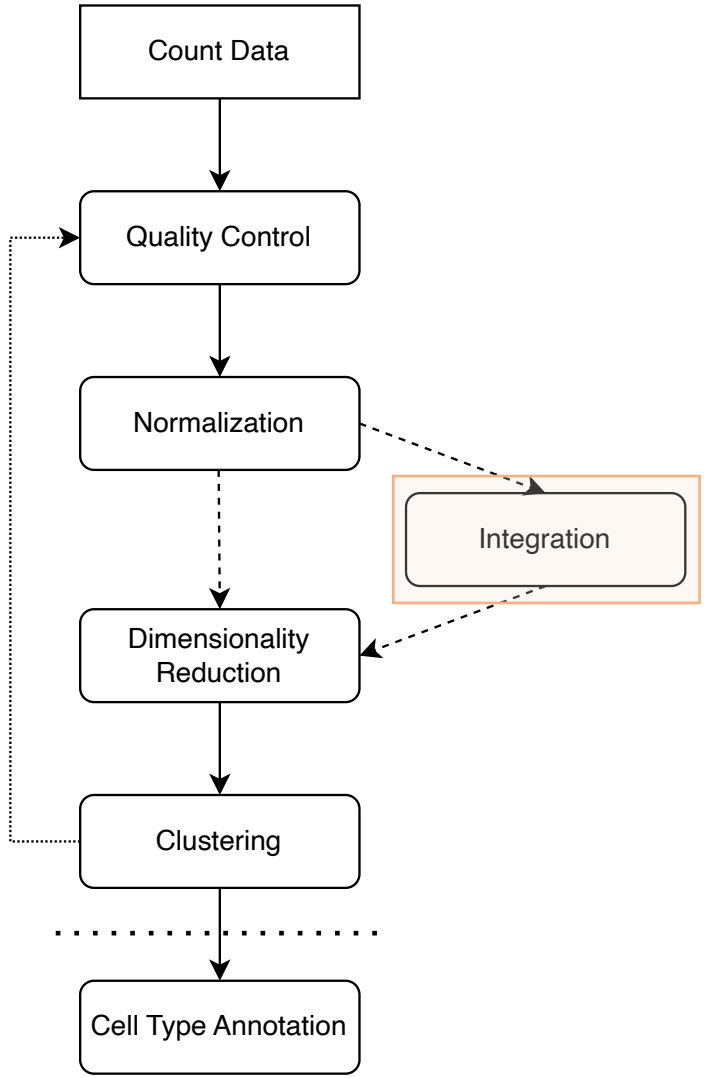


Data Integration

Transform our data such that the profiles of the same cell types are more similar across data sets

- **Seurat this is performed with two main procedures**
  - **Canonical Correlation Analysis**
  - **Mutual Nearest Neighbors**



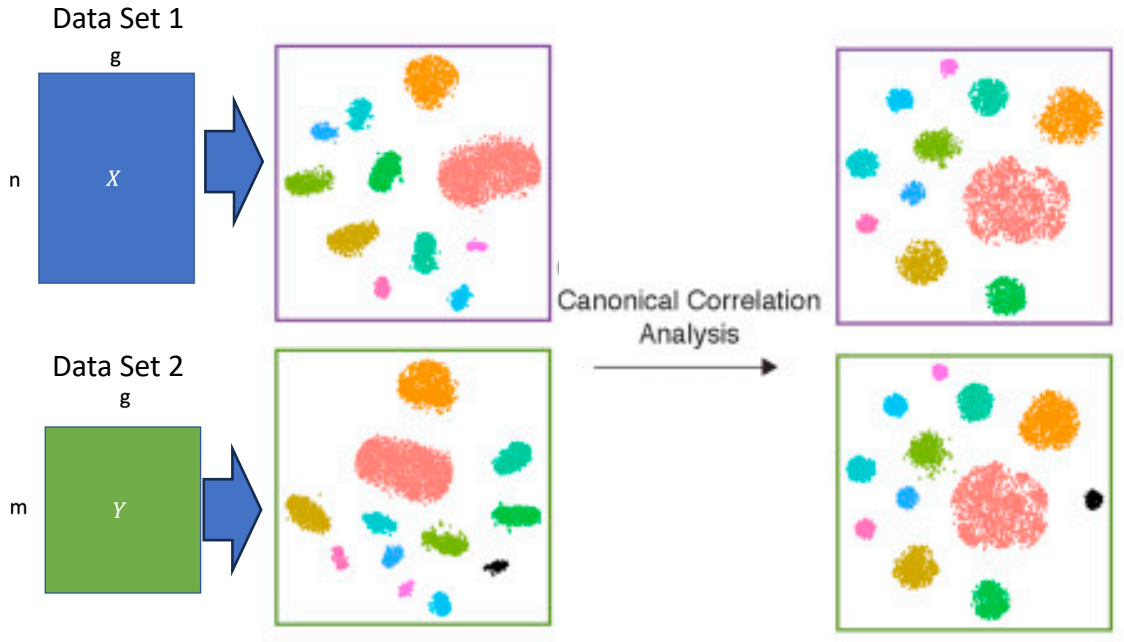


Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability

Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability

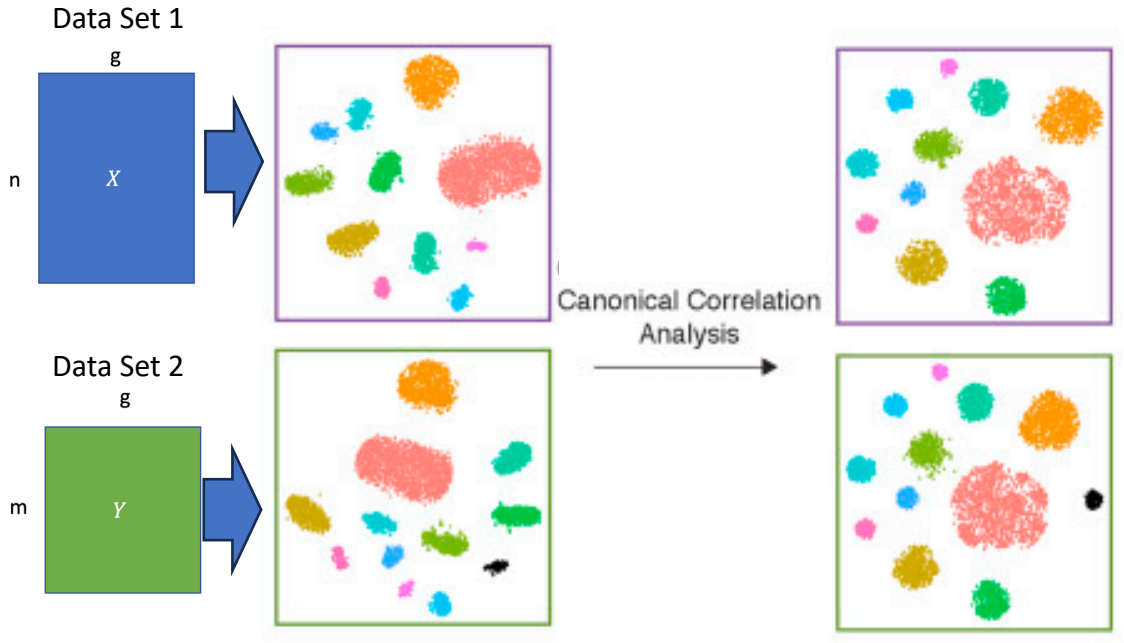


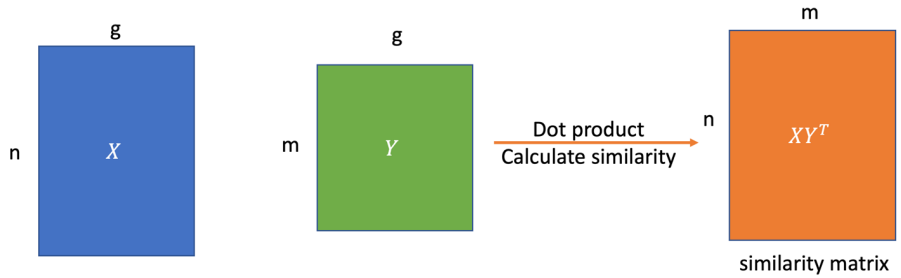
n: Number of cell profiles in dataset 1
m: Number of cell profiles in dataset 2
g: Number of genes

https://doi.org/10.1038/nbt.4091

Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability

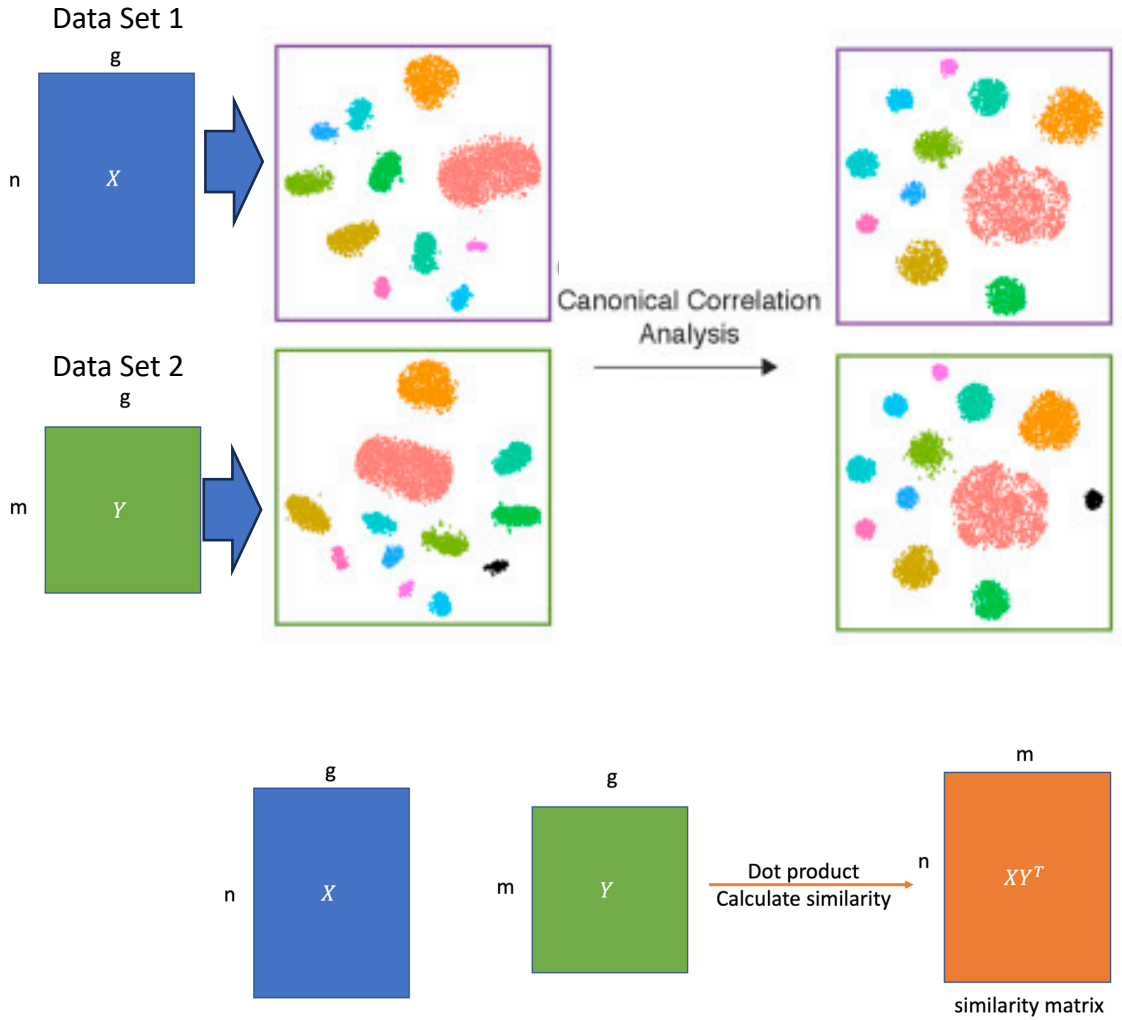




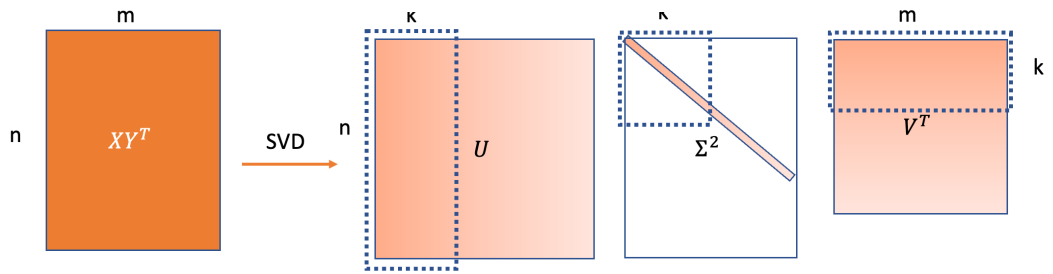
https://doi.org/10.1038/nbt.4091

Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability



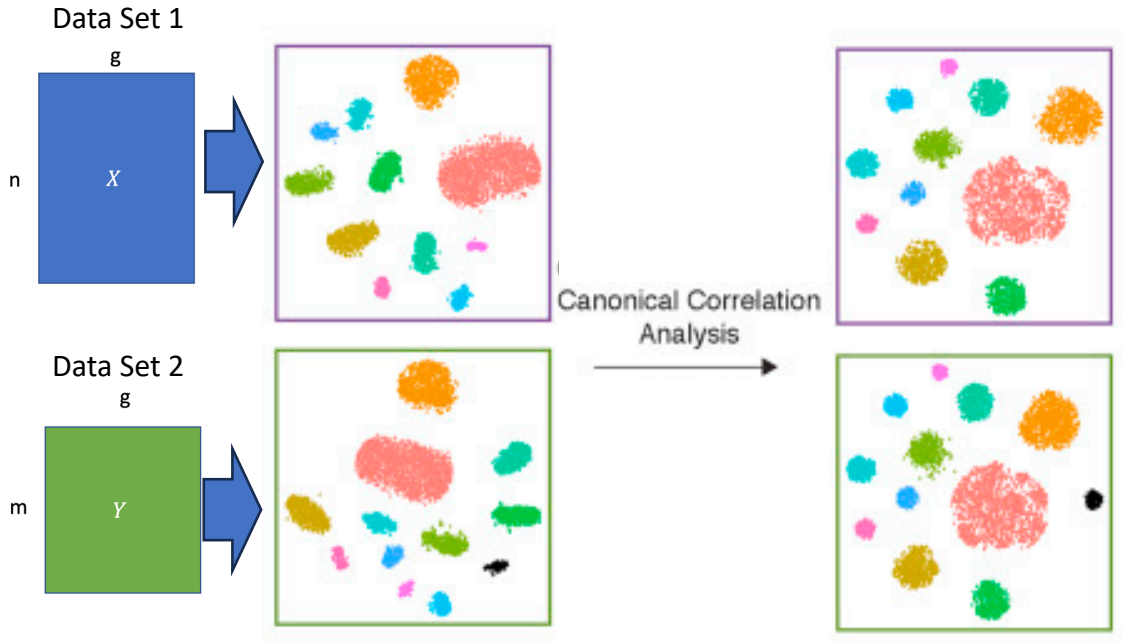
$$XY^T = U\Sigma^2 V^T$$

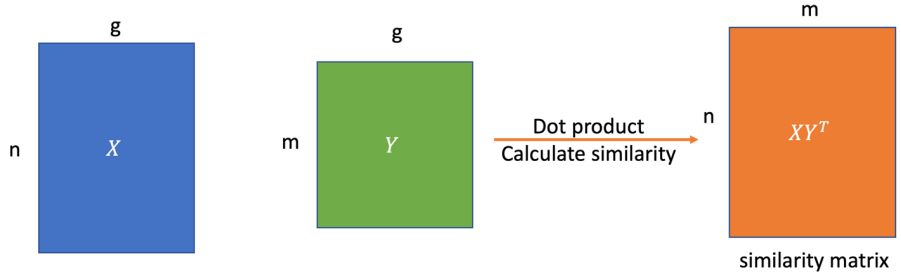


https://doi.org/10.1038/nbt.4091

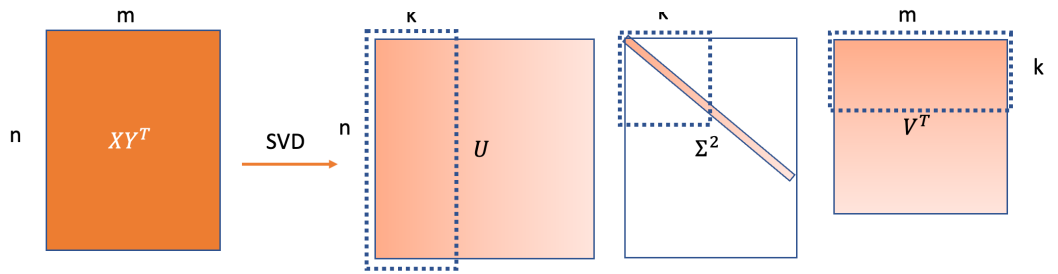
Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability

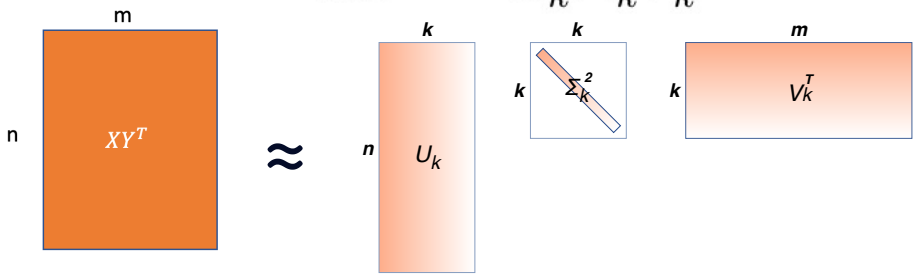




$$XY^T = U\Sigma^2 V^T$$



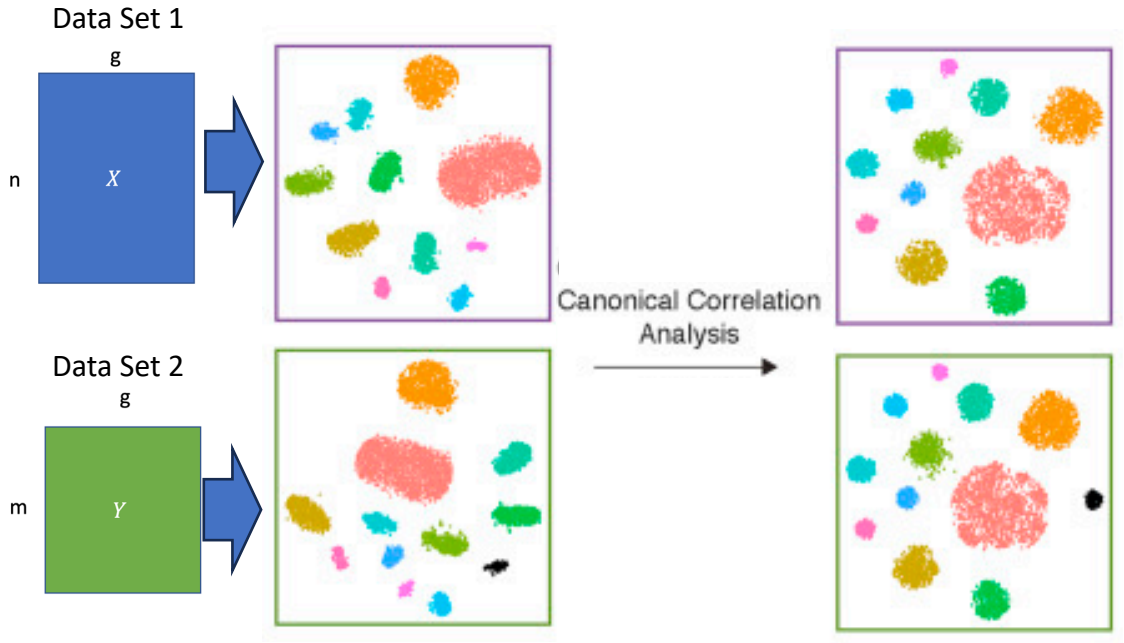
$$XY^T \approx U_k^2 \Sigma_k^2 V_k^T$$

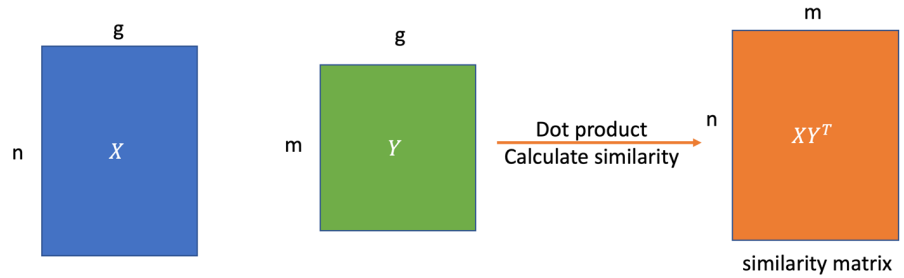


https://doi.org/10.1038/nbt.4091

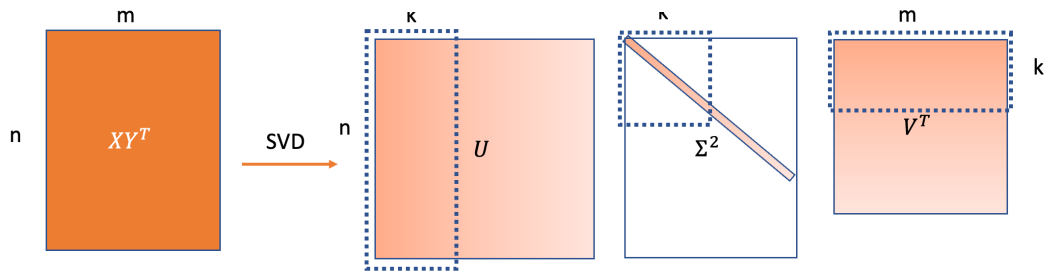
Data Integration (Canonical Correlation Analysis)

Project two data sets into reduced dimension space that captures most of their **shared** variability

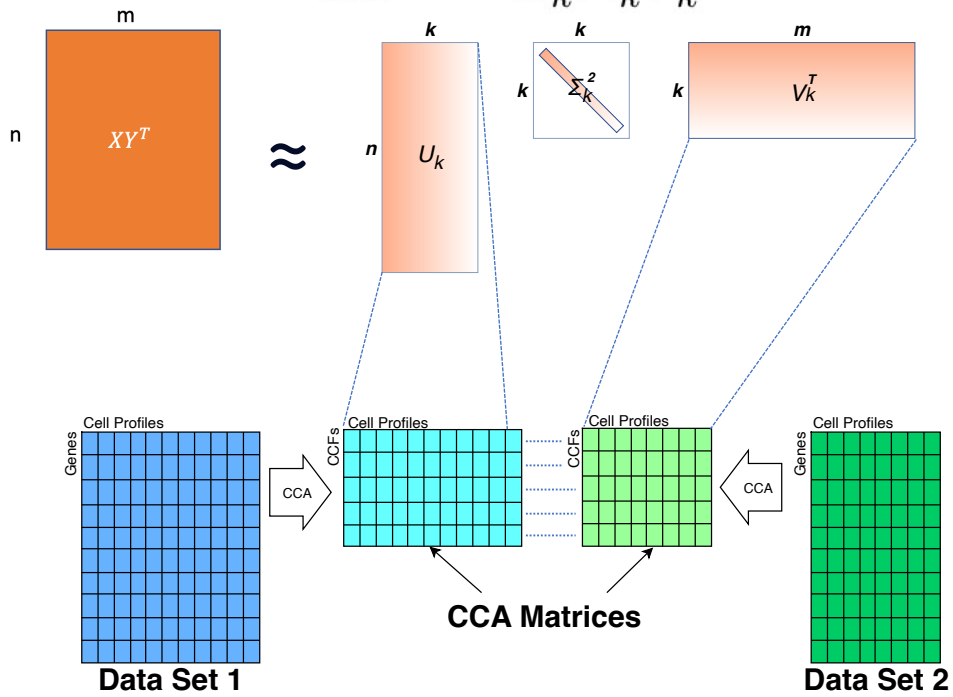




$$XY^T = U\Sigma^2 V^T$$



$$XY^T \approx U_k^2 \Sigma_k^2 V_k^T$$



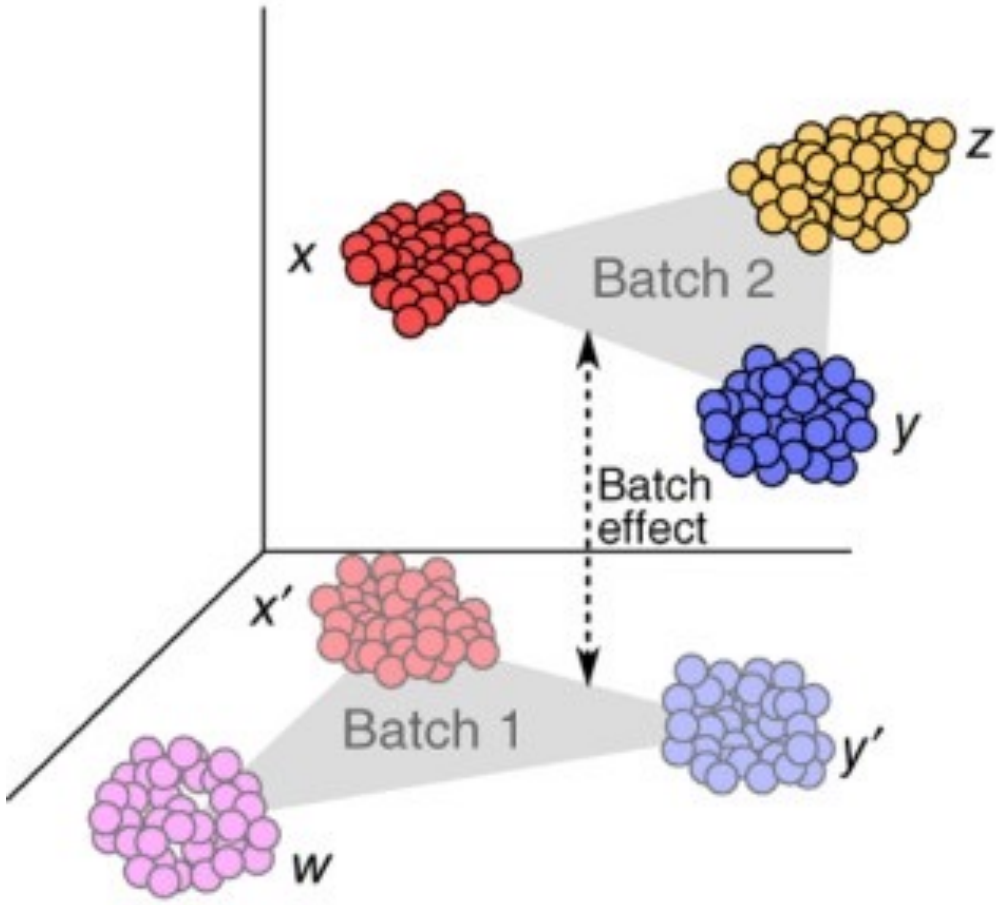
https://doi.org/10.1038/nbt.4091

Data Integration (Mutual Nearest Neighbor)

Characterize cells with similar CCA profiles across data sets.

Hypothetical sets set containing two data sets (batched)

- **Each batch has three cell types**
  - **Two pairs of cell types are shared across data sets**
    - **X and X', Y and Y'**
  - **Two cell types are specific to either data set**
    - **W and Z**



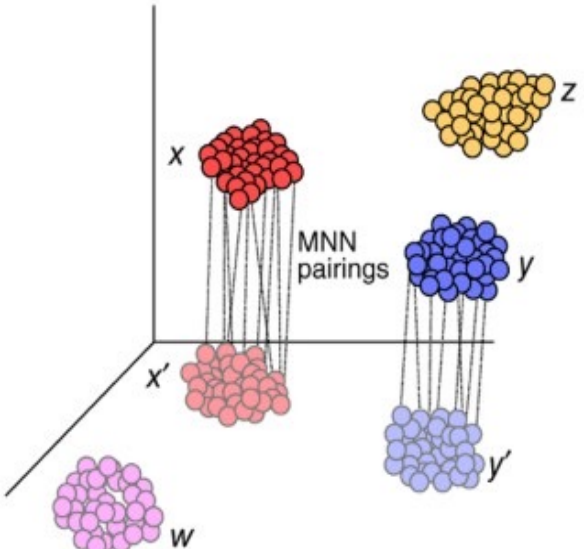
https://doi.org/10.1038/nbt.4091

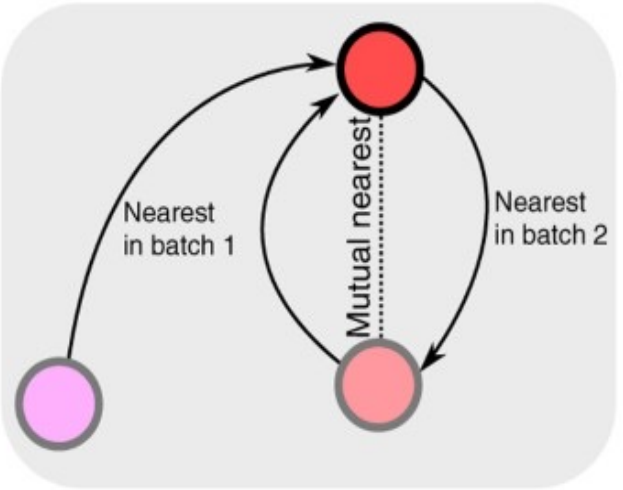
Data Integration (Mutual Nearest Neighbor)

Mutual Nearest Neighbors estimates pairings of cells across batches that are likely to be the same cell type

- **Cells in X' of data set 1 are nearest to cells in X of data set 2**
- **Cells in X of data set 2 are nearest to cells in X of data set 1**
  - **Mutual Nearest Neighbors**

- **Cells in W and Z have lack MNN across data sets**

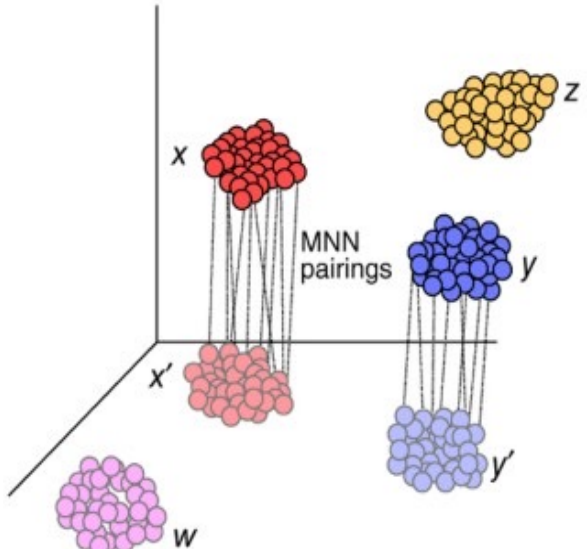


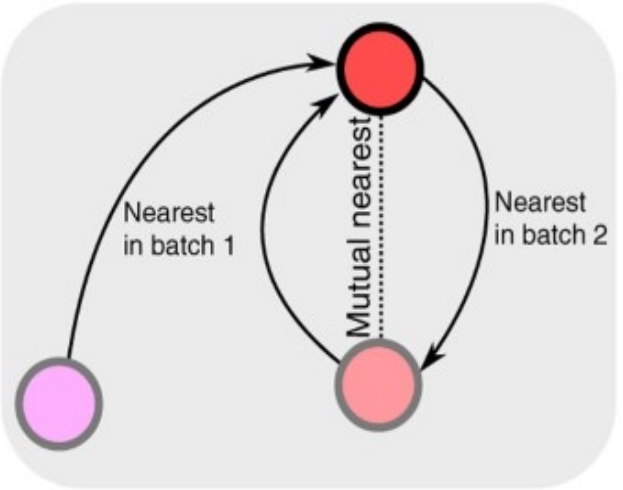


https://doi.org/10.1038/nbt.4091

Data Integration (Mutual Nearest Neighbor)

Transform cell expression profiles closer to their MNN

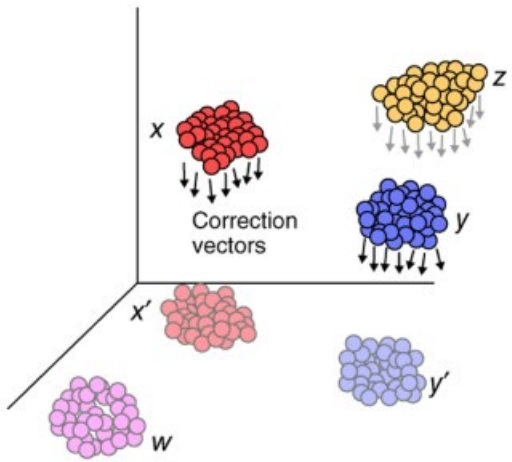


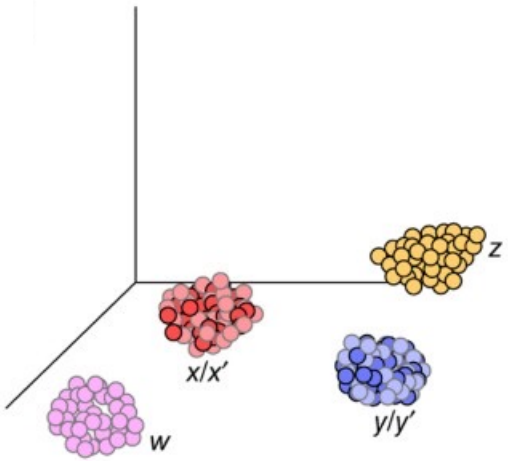


Using these mutual nearest neighbor estimations we then transform

- **X cells to be closer to X' cells**
- **Y cells to be closer to Y' cells**

- **Disparate cell types Z and W remain distant**



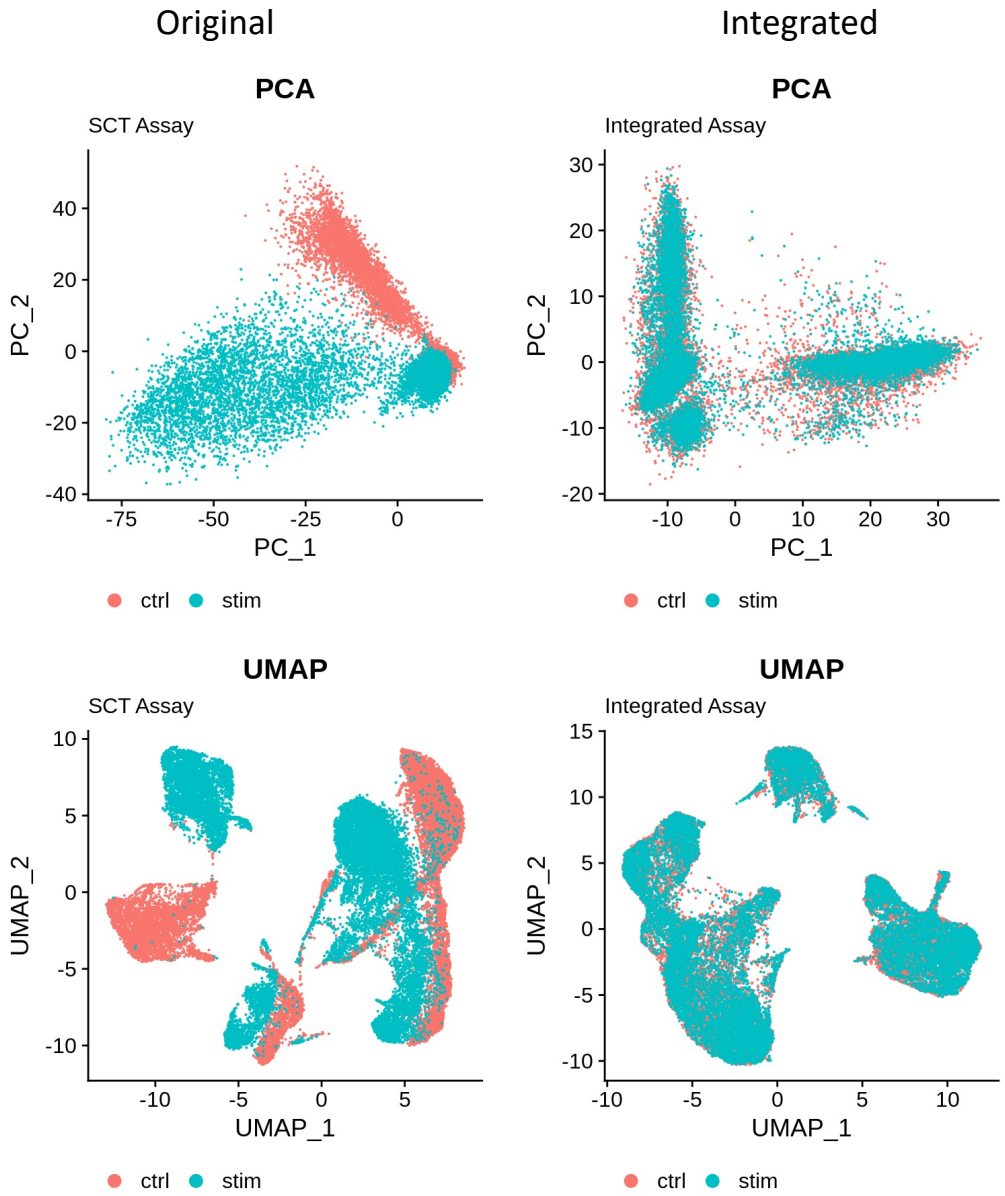


https://doi.org/10.1038/nbt.4091

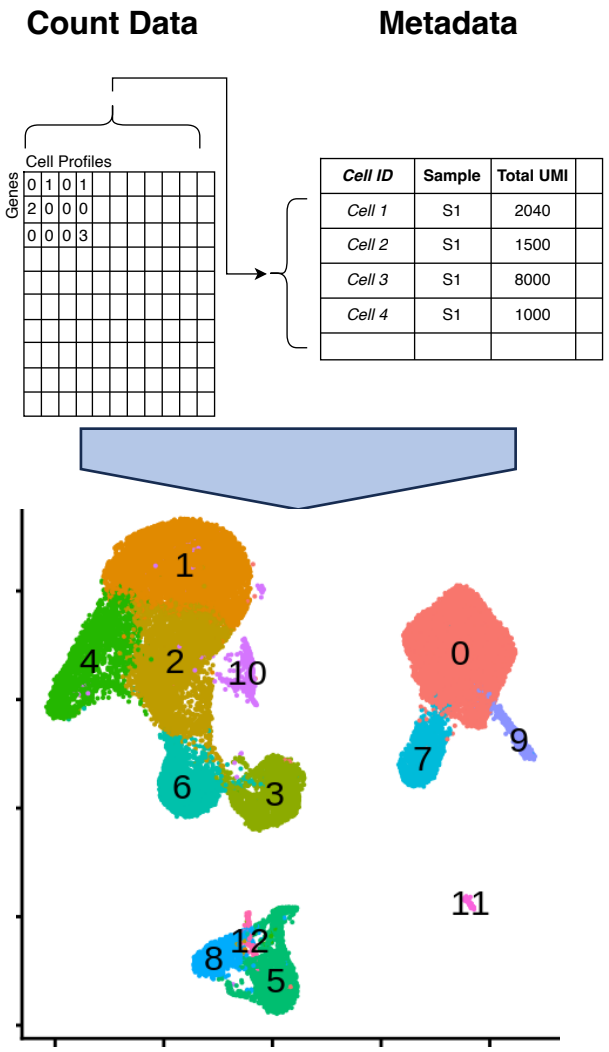
Data Integration

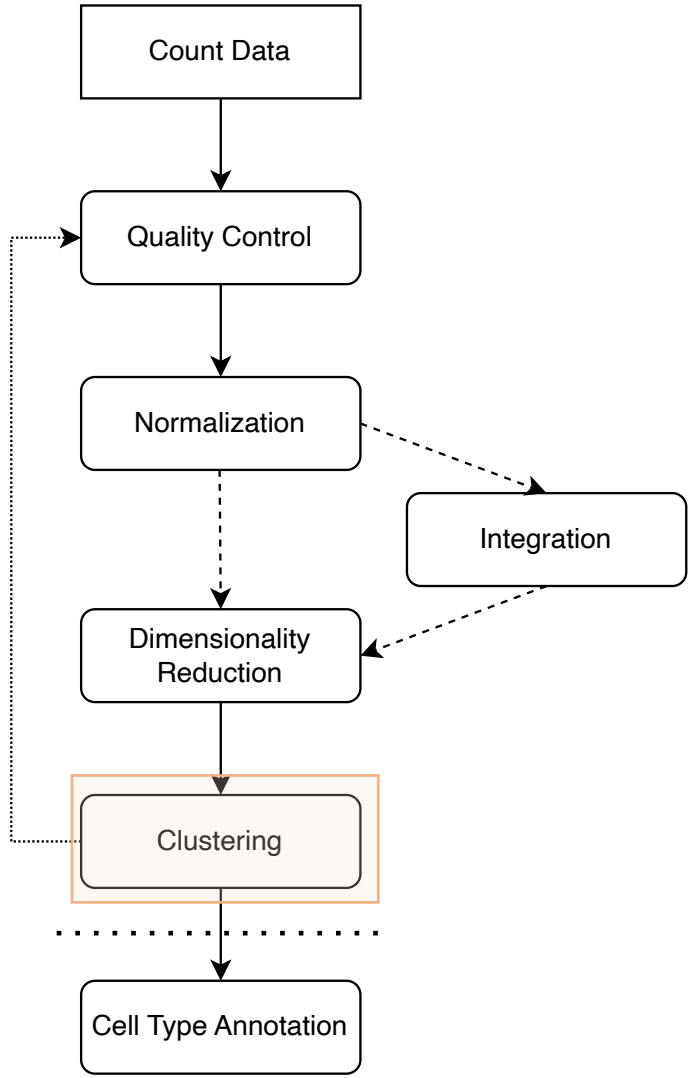
Our control and stimulated samples appear well integrated in two dimensions

- **Best to inspect the data prior to integration**
- **Consider how similar we expect data sets to be**

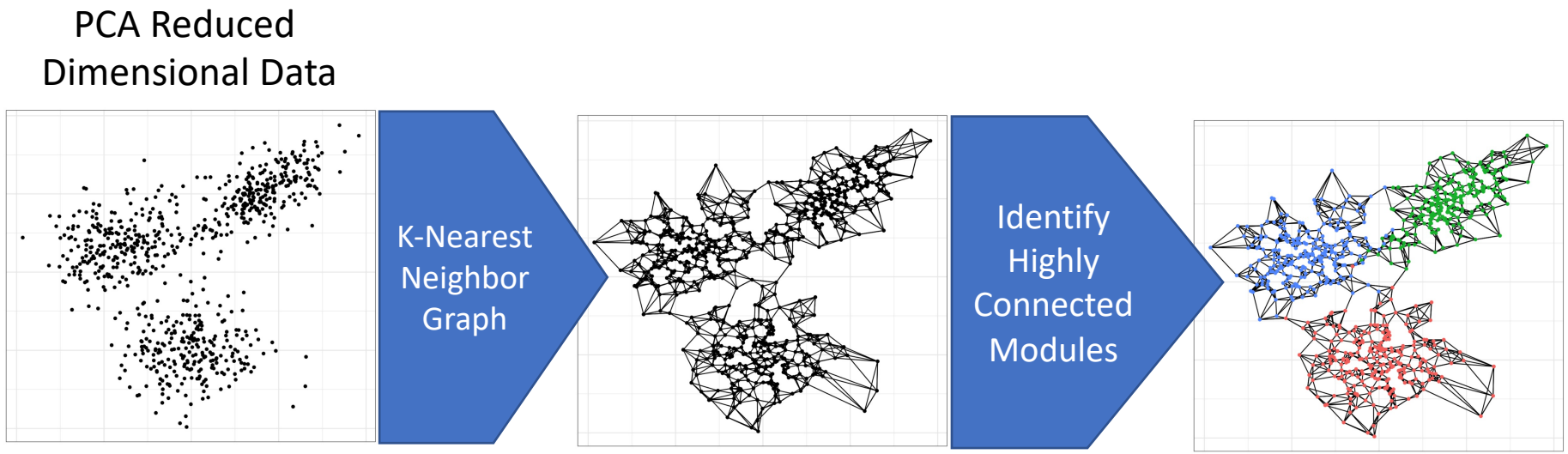


Clustering


Count Data
Metadata
Cell Profiles
Genes
0 1 0 1
2 0 0 0
0 0 0 3
Cell ID | Sample | Total UMI
Cell 1 | S1 | 2040
Cell 2 | S1 | 1500
Cell 3 | S1 | 8000
Cell 4 | S1 | 1000
1
2
3
4
5
6
7
8
9
10
11
12
0

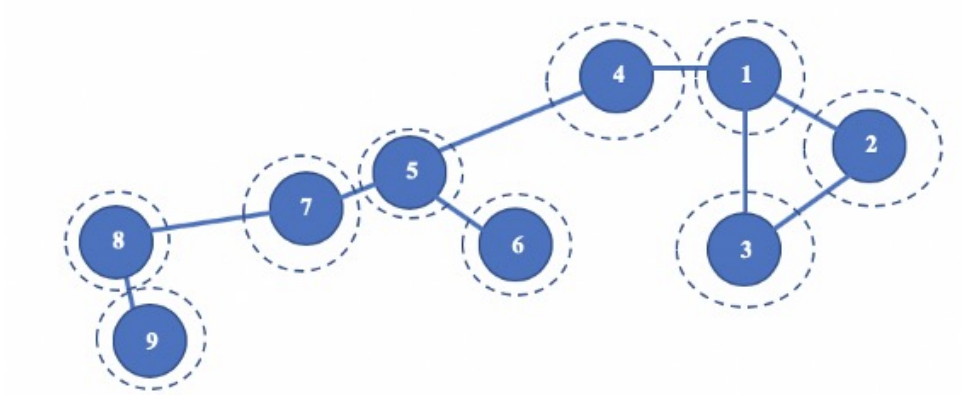

Count Data
Quality Control
Normalization
Integration
Dimensionality Reduction
Clustering
Cell Type Annotation

Graph-based Clustering



Clustering (Louvain Algorithm)

1. Create k-nearest neighbor graph

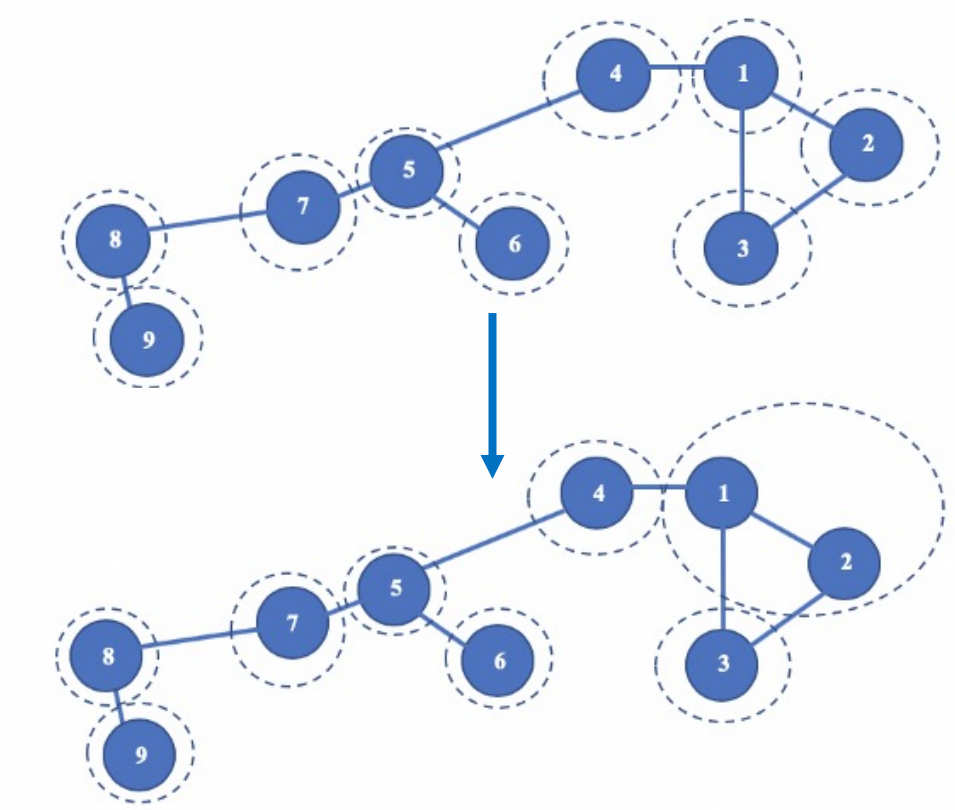


https://medium.com/walmartglobaltech/demystifying-louvains-algorithm-and-its-implementation-in-gpu-9a07cdd3b010

Clustering (Louvain Algorithm)

1. Create k-nearest neighbor graph

2. Merge cells into communities that most increase the *modularity* of the graph



Modularity

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

A_{ij} = Weight of edge between cells *i* and *j*

k_i, k_j = *Degree of cells i and j*

$\delta(c_i,c_j)$ =

- 1 if *cells i and j are in the same community*
- *0 otherwise*

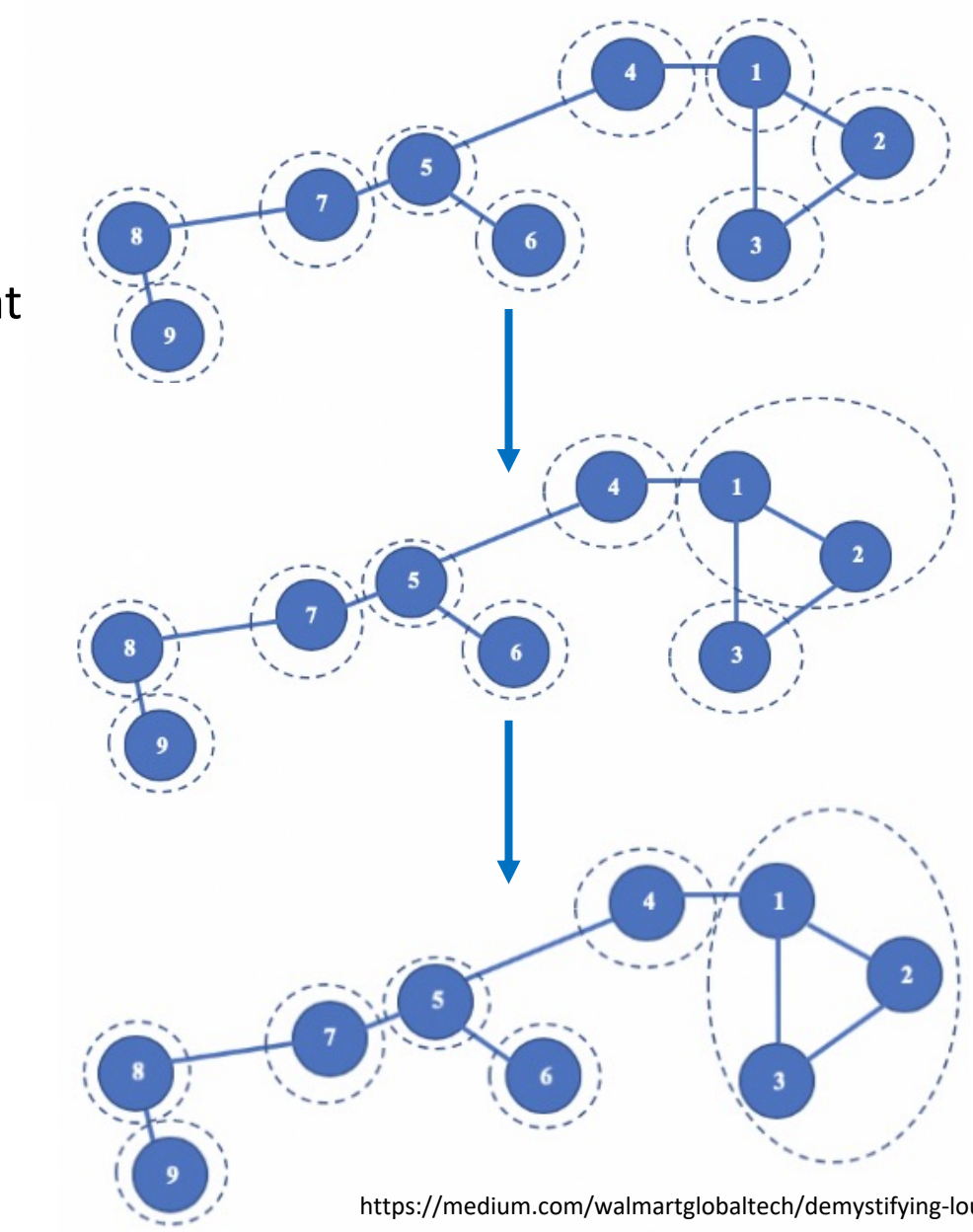
https://medium.com/walmartglobaltech/demystifying-louvains-algorithm-and-its-implementation-in-gpu-9a07cdd3b010

Clustering (Louvain Algorithm)

1. Create k-nearest neighbor graph

2. Merge cells into communities that most increase the *modularity* of the graph

3. Repeat (2) until there is modularity increase is below a specified threshold
   - This threshold controls the resolution
     - High resolution -> More clusters



Modularity

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

A_{ij} = Weight of edge between cells *i* and *j*

k_i, k_j = *Degree of cells i and j*

$\delta(c_i,c_j)$ =
- 1 if *cells i and j are in the same community*
- *0 otherwise*

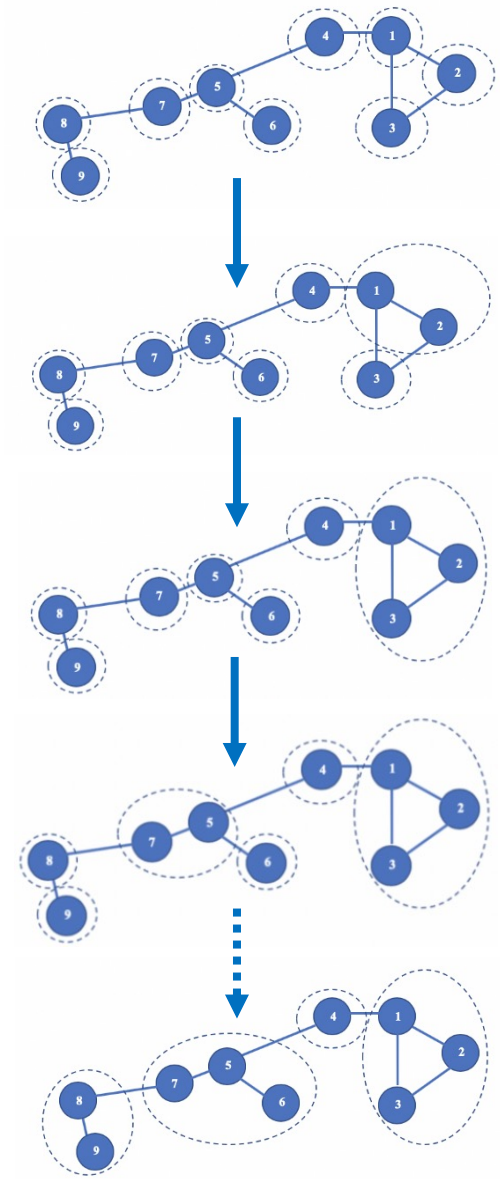
https://medium.com/walmartglobaltech/demystifying-louvains-algorithm-and-its-implementation-in-gpu-9a07cdd3b010

Clustering (Louvain Algorithm)

1. Create k-nearest neighbor graph

2. Merge cells into communities that most increase the *modularity* of the graph

3. Repeat (2) until there is modularity increase is below a specified threshold
   - This threshold controls the resolution
     - High resolution -> More clusters



Modularity

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

A_{ij} = Weight of edge between cells *i* and *j*

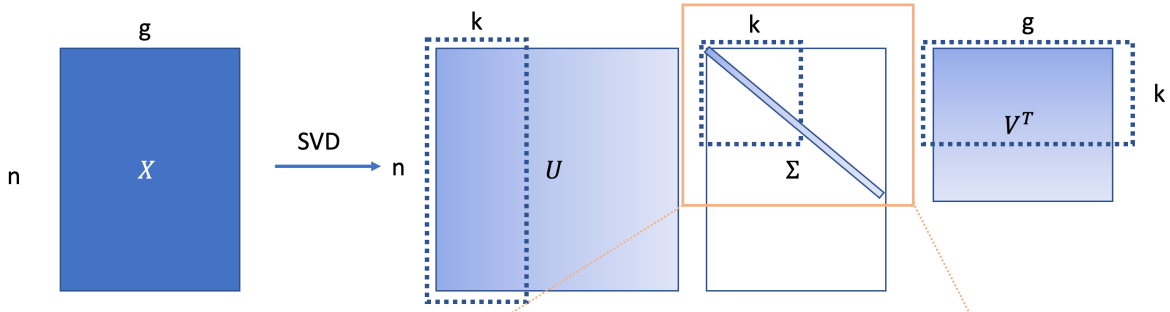
k_i, k_j = *Degree of cells i and j*

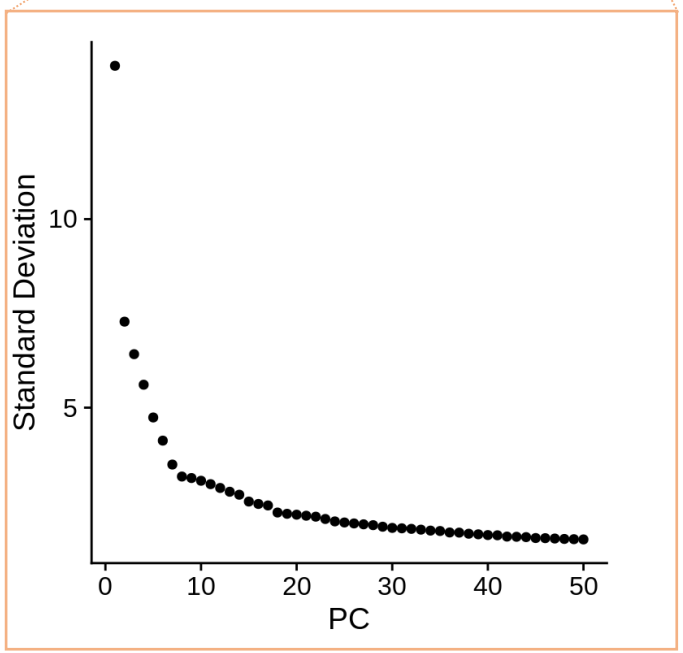
$\delta(c_i,c_j)$ =
- 1 if *cells i and j are in the same community*
- *0 otherwise*

https://medium.com/walmartglobaltech/demystifying-louvains-algorithm-and-its-implementation-in-gpu-9a07cdd3b010

Clustering (Louvain Algorithm)

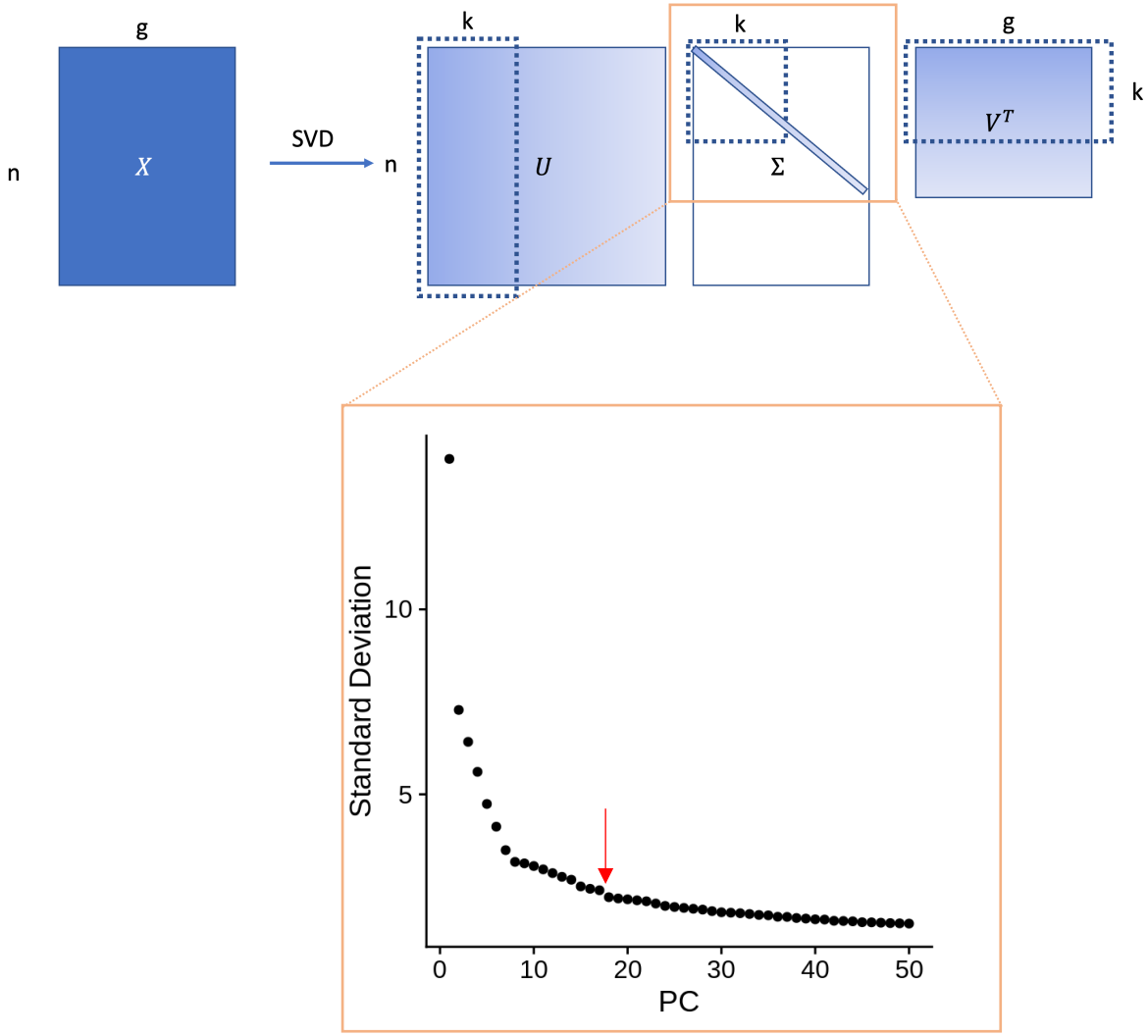
Choosing the number of principal components to use for clustering

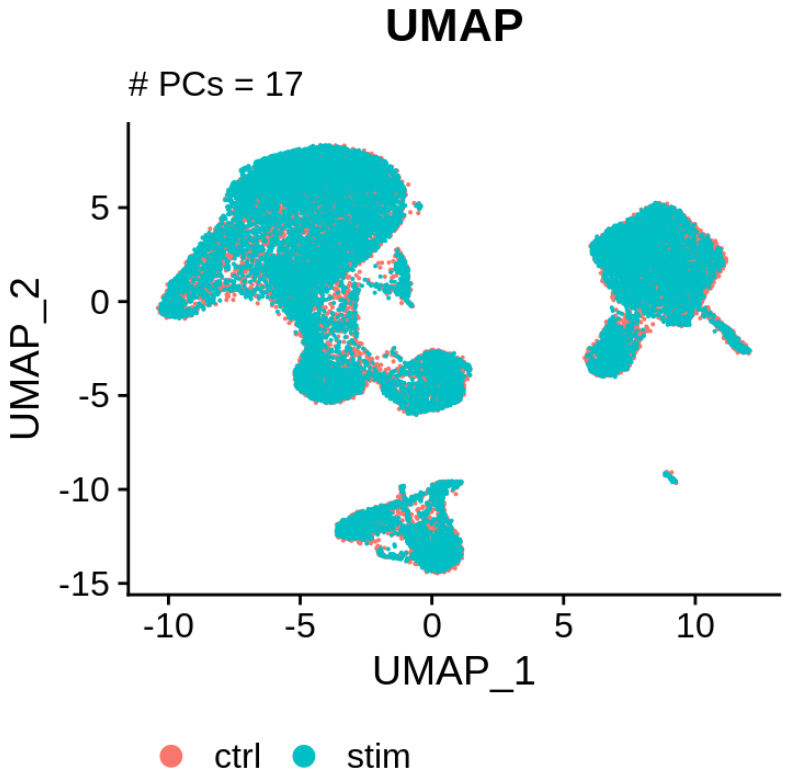




Clustering (Louvain Algorithm)

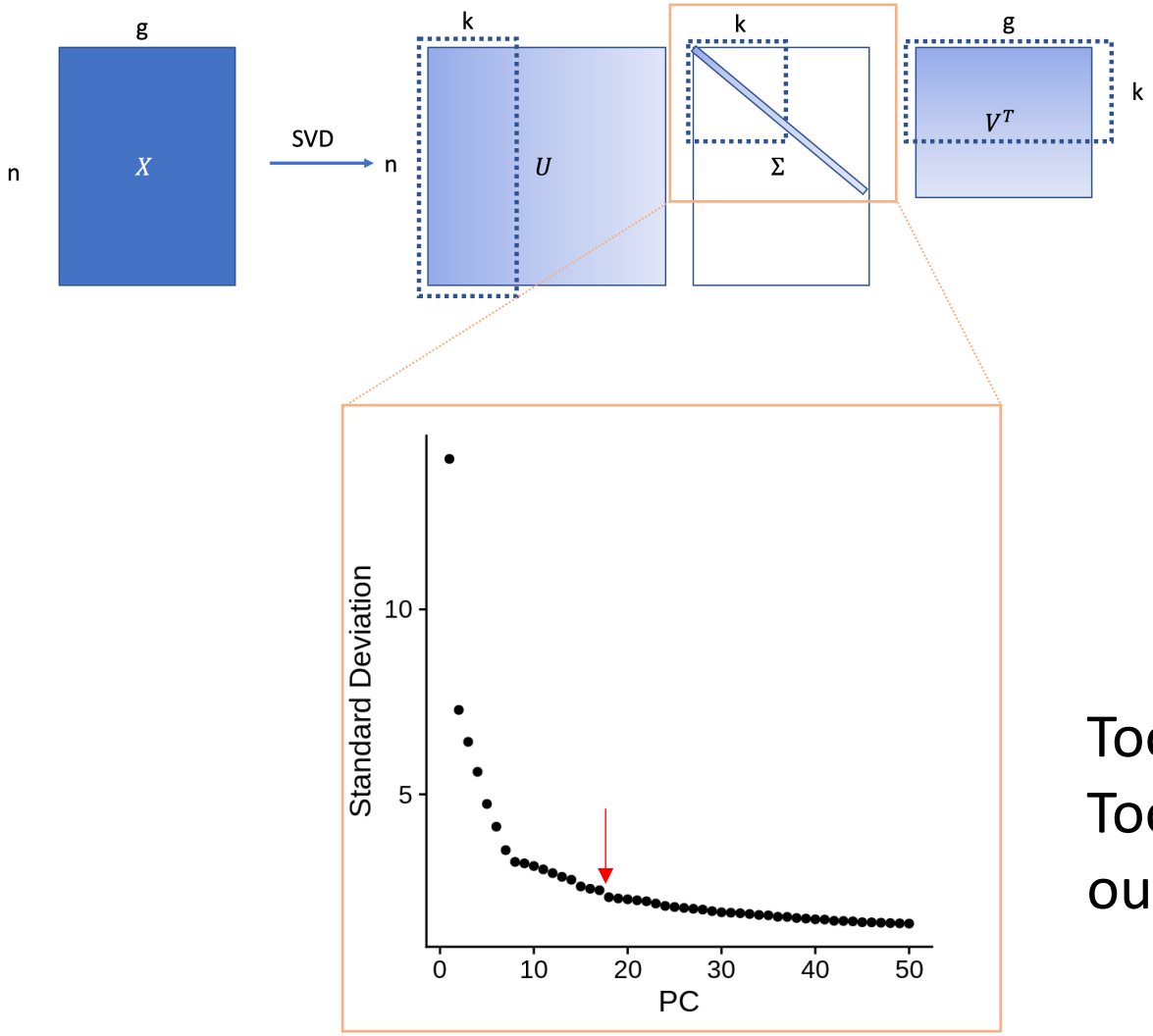
Choosing the number of principal components to use for clustering

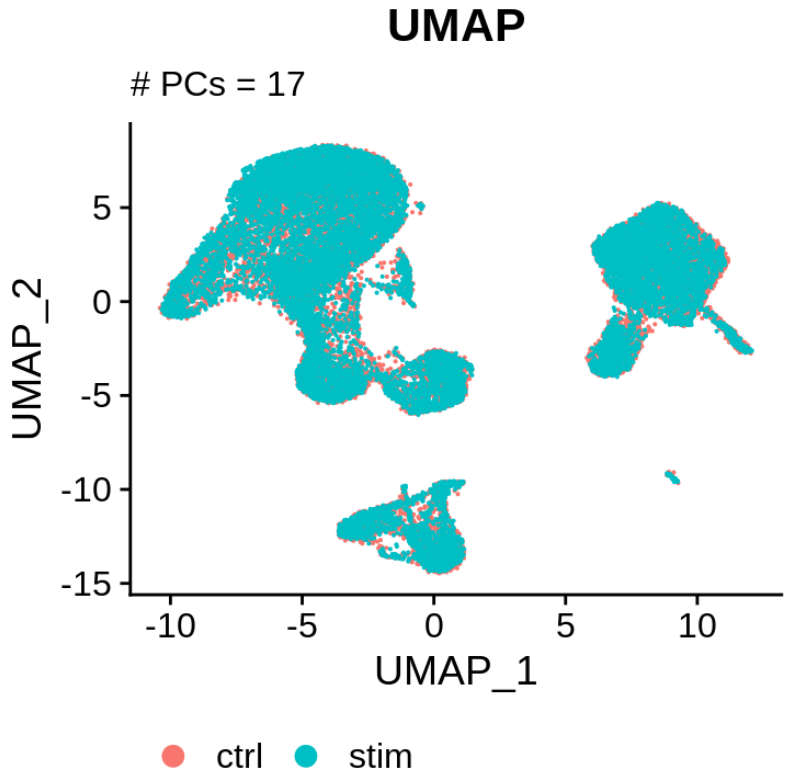




Clustering (Louvain Algorithm)

Choosing the number of principal components to use for clustering





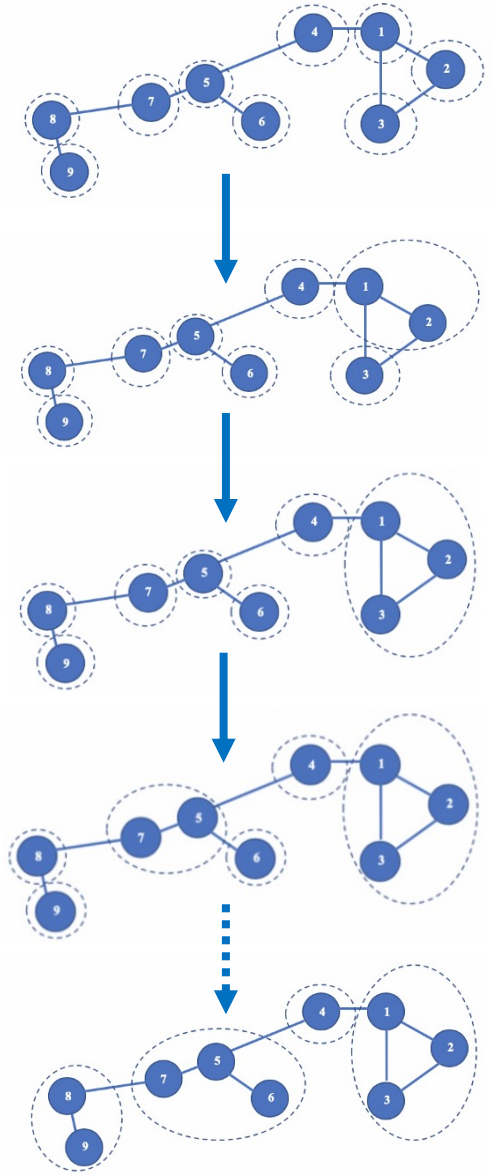
Too Few PCs: Don't include important information

Too Many PCs: We include extraneous variability in our data

Clustering (Louvain Algorithm)

Resolution Parameter

High Resolution

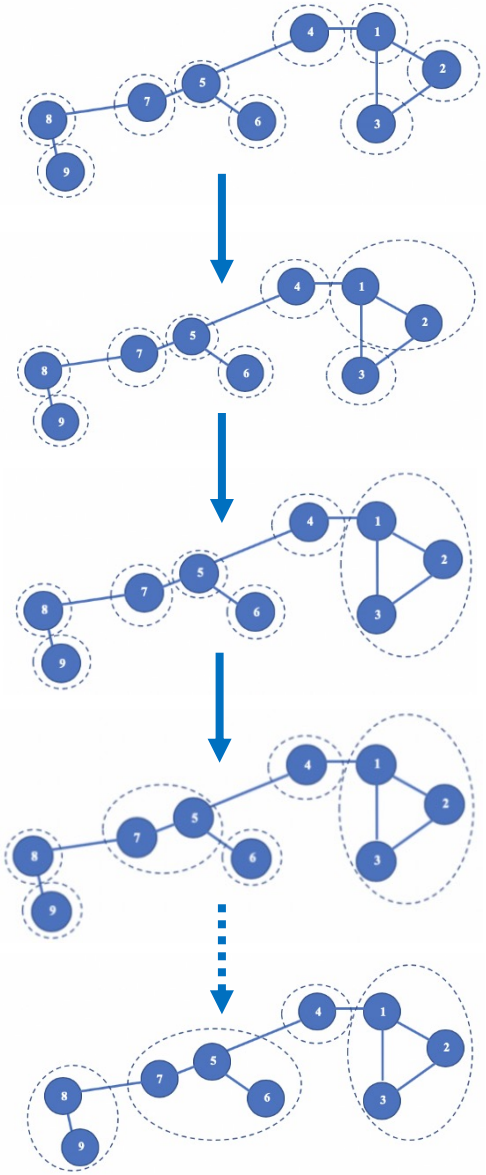


Low Resolution

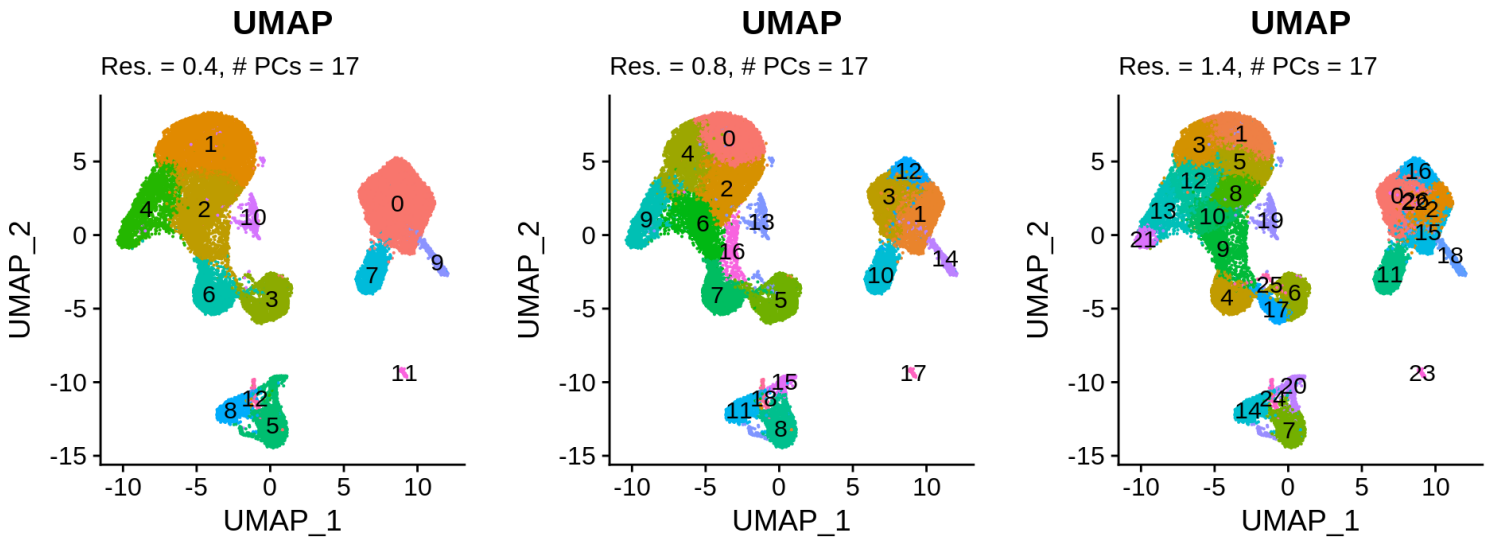
Clustering (Louvain Algorithm)

Resolution Parameter

High Resolution



Low Resolution

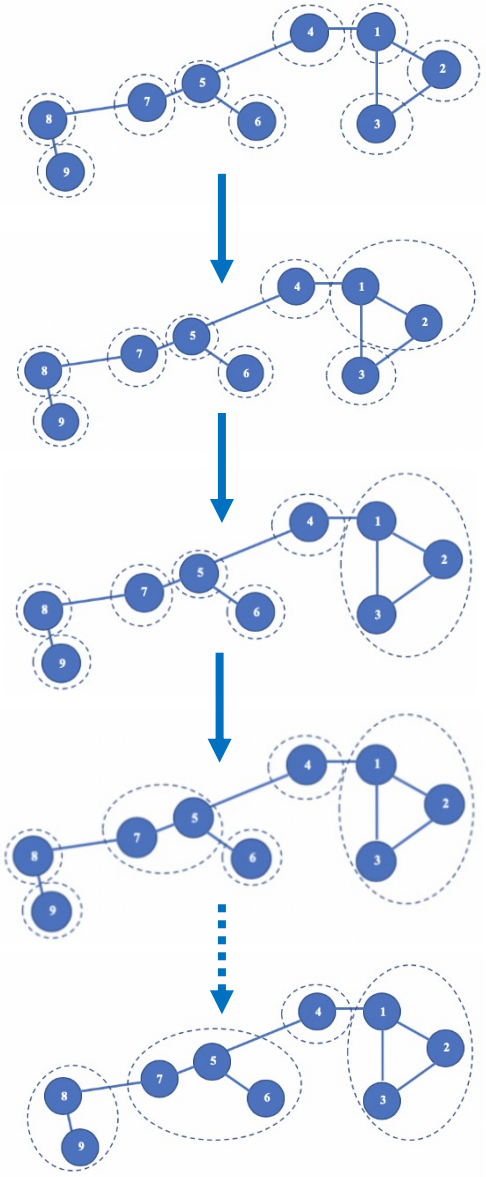


Low Resolution ← High Resolution

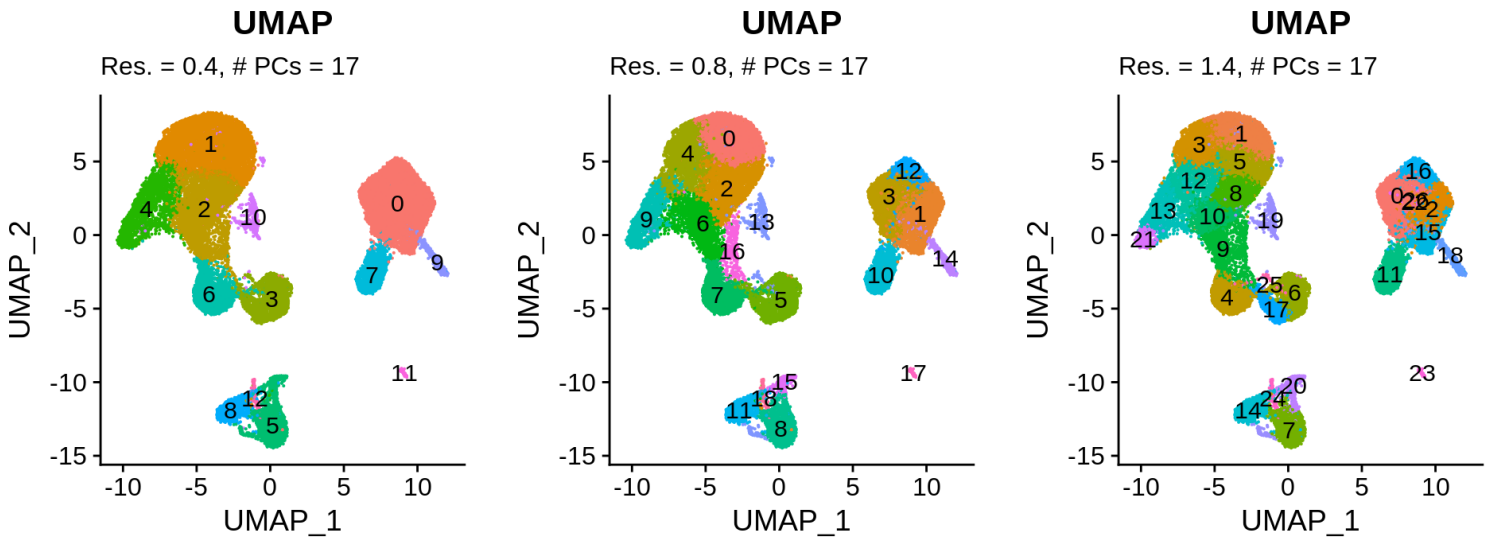
Clustering (Louvain Algorithm)

Resolution Parameter

High Resolution



Low Resolution



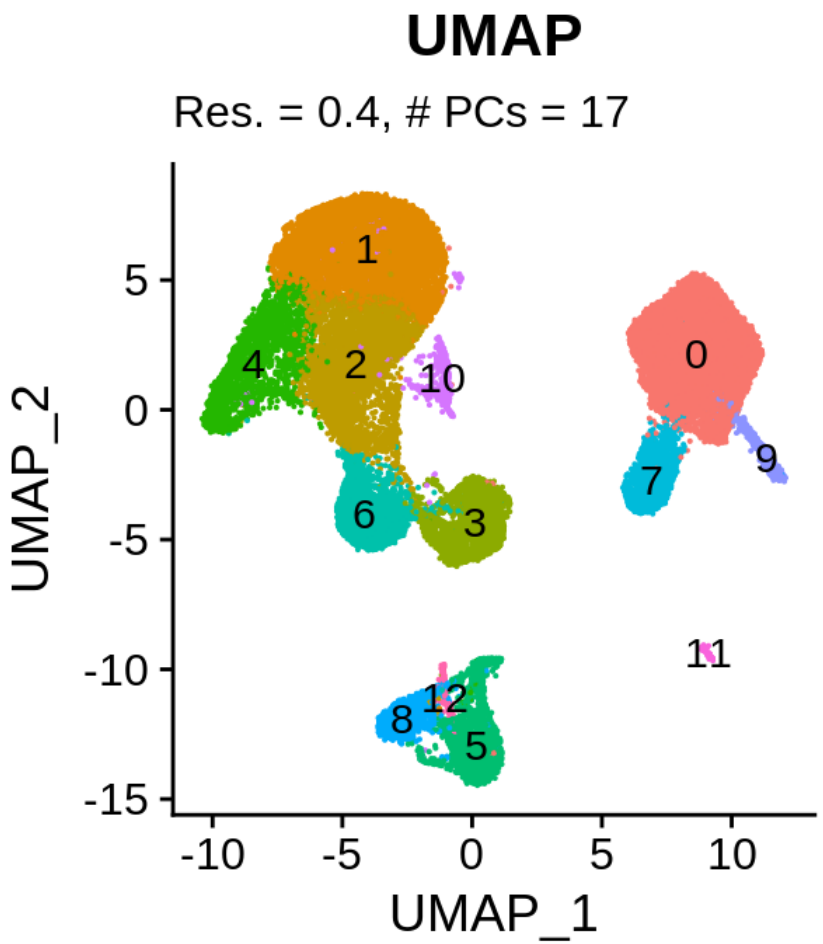
Low Resolution ← High Resolution

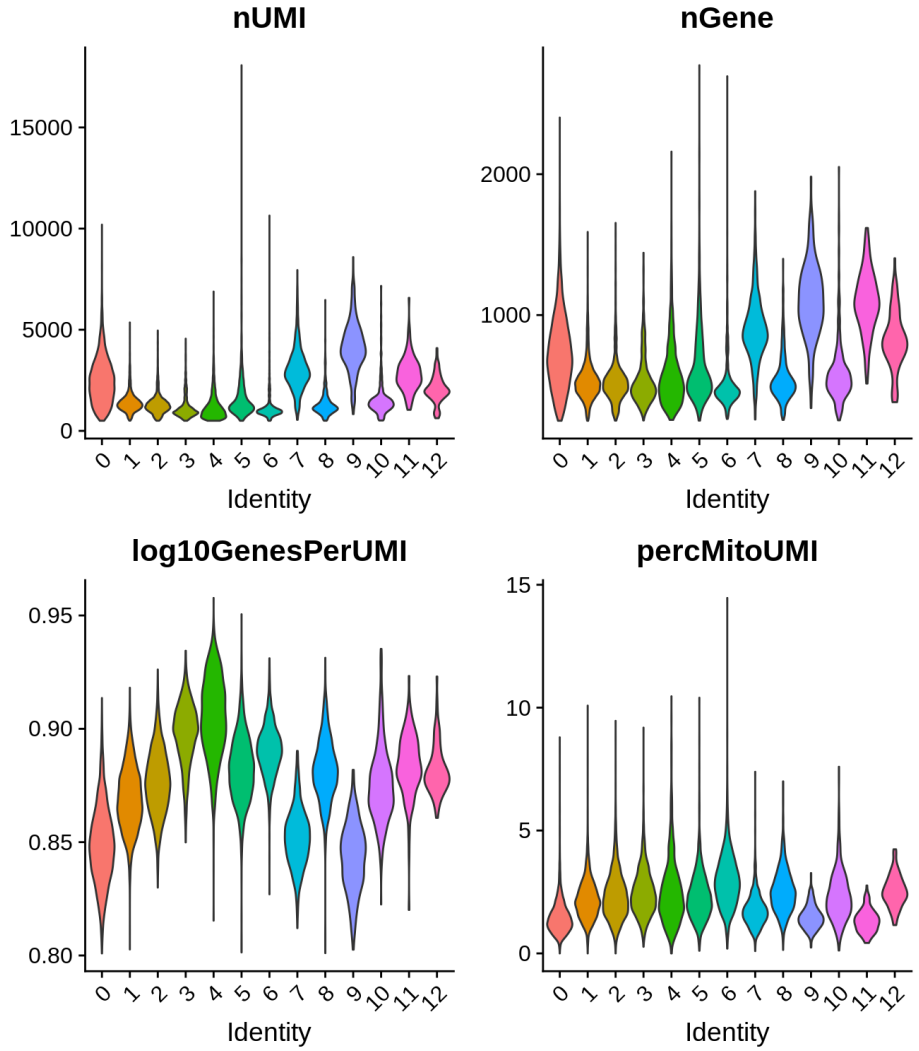
Too Low Resolution: Can't distinguish finer cell subtypes
Too High Resolution: Cell clusters aren't distinguishable

A common strategy is to start at low resolution. Re-cluster specific cell types to achieve finer resolution.

Inspecting Cluster Results

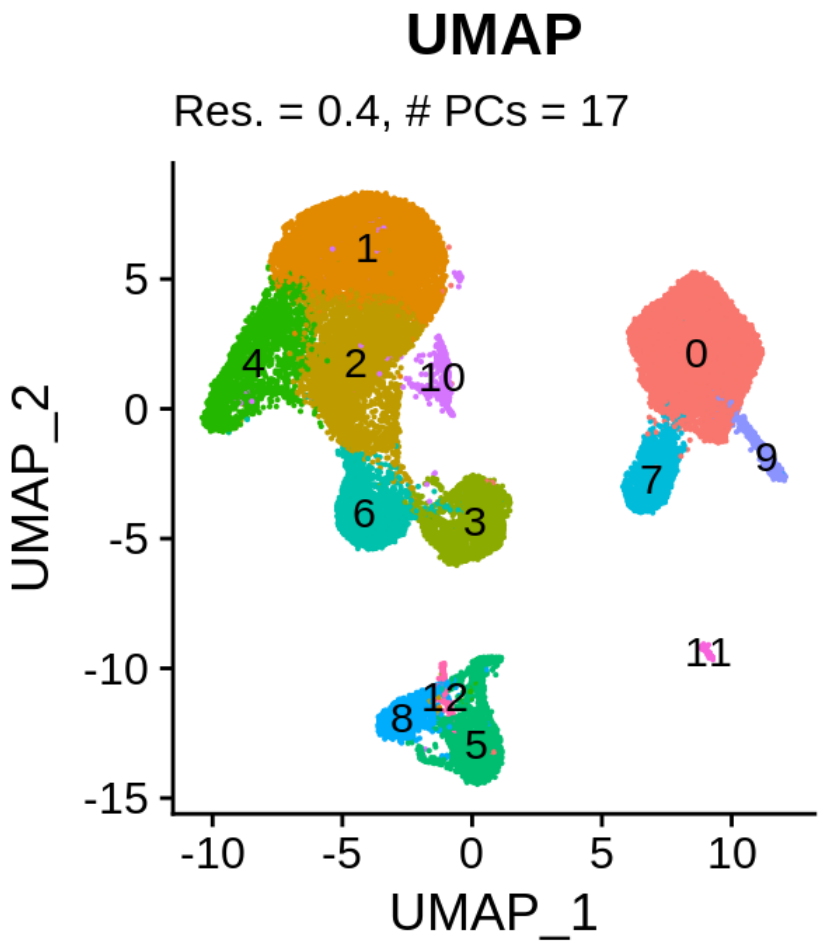
- Total UMI count (nUMI)
- Number of counted genes (nGene)
- Novelty Score (log10GenePerUMI)
- Percent Mitochondrial UMI (percMitoUMI)

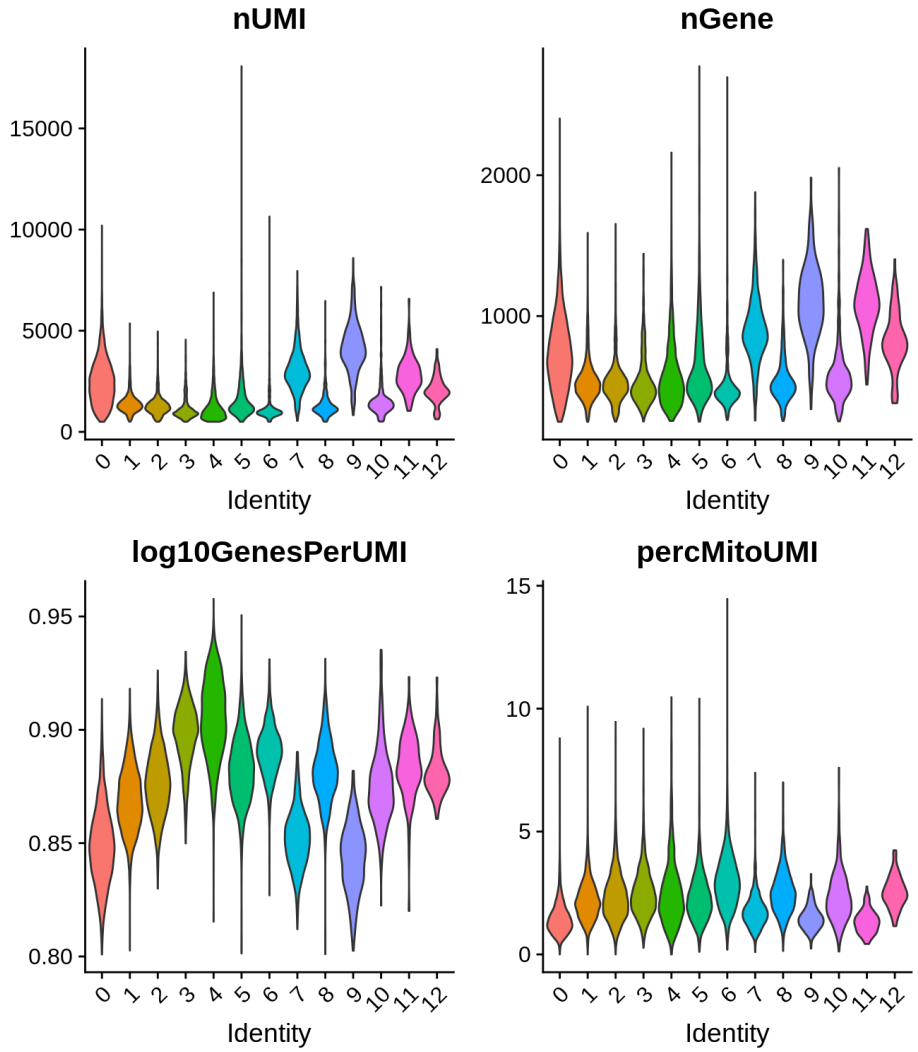




Inspecting Cluster Results

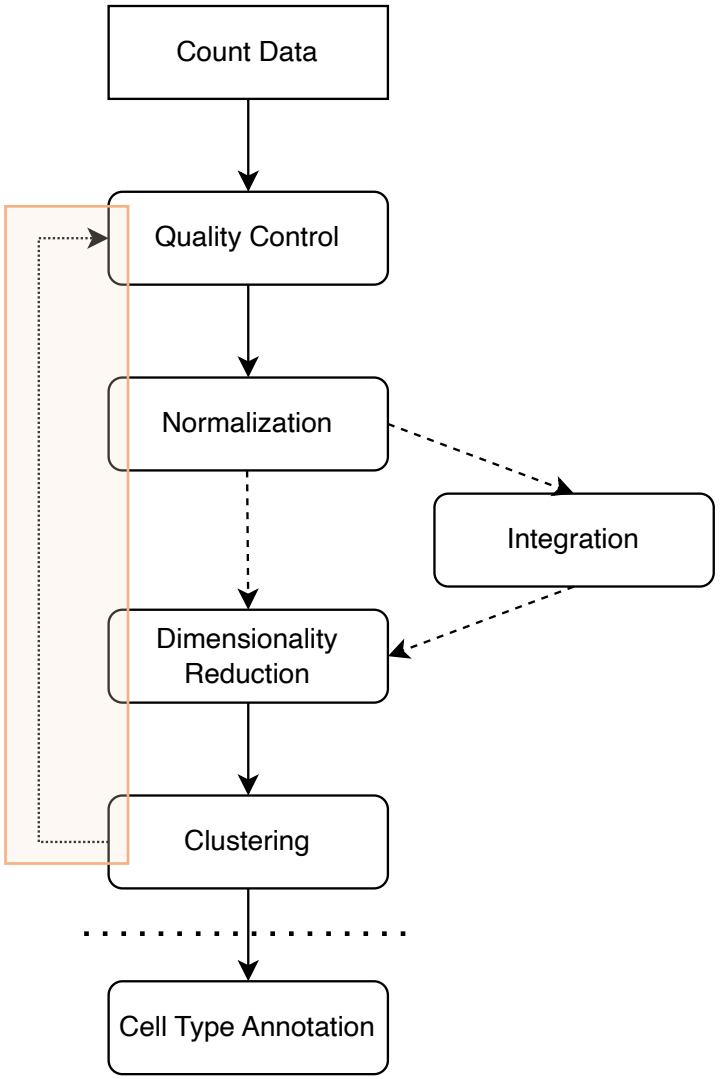
- Total UMI count (nUMI)
- Number of counted genes (nGene)
- Novelty Score (log10GenePerUMI)
- Percent Mitochondrial UMI (percMitoUMI)





- No clusters “jump” out as potential outliers
- We do observe from cluster-wise differences
  - E.g. high mitochondrial content in cluster 12

Discussion

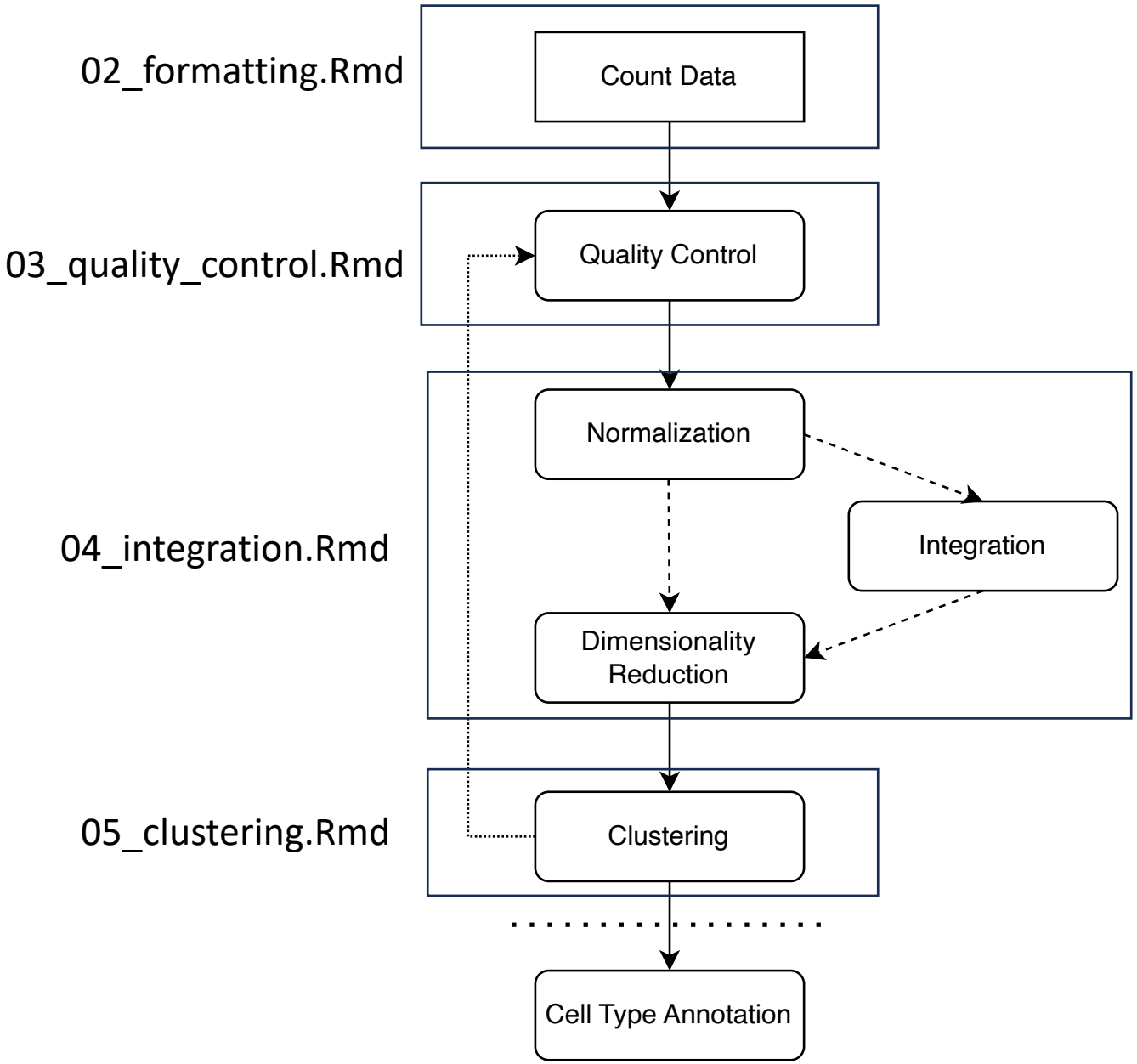


Like with other “omics”, analyses of scRNAseq data are **iterative**

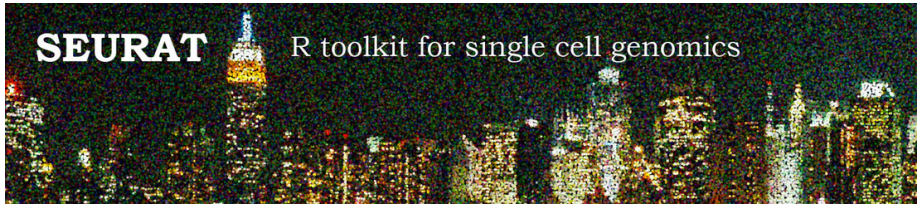
- Relies heavily on heuristic decisions
- There are many tuning parameters
  - Data filtering thresholds
  - Feature selection
  - Number of principal component to use for clustering
  - Stopping parameter for Louvain Algorithm
    - High resolution -> More clusters
- Critical to have accurate cell labels before moving onto downstream analyses

- **Cell type annotation and differential gene expression analysis will be covered in next week’s workshop**

Workshop Portion



We are going to cover this workflow using the…



https://satijalab.org/seurat/