

ELIXIR course

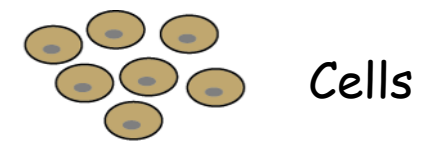
Single Cell Data: Visualization & Clustering

Gennaro Gambardella
Ph.D. in Computational Biology and Bioinformatics
di Bernardo's lab

FONDAZIONE

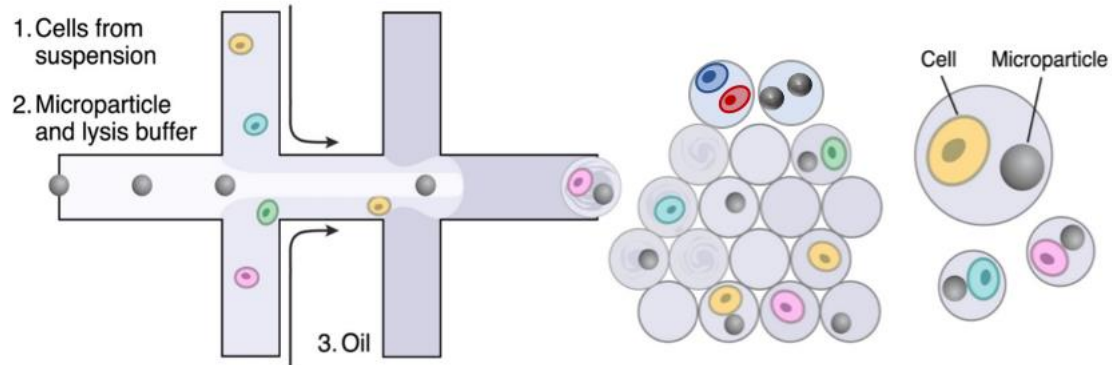


How does single cell work?

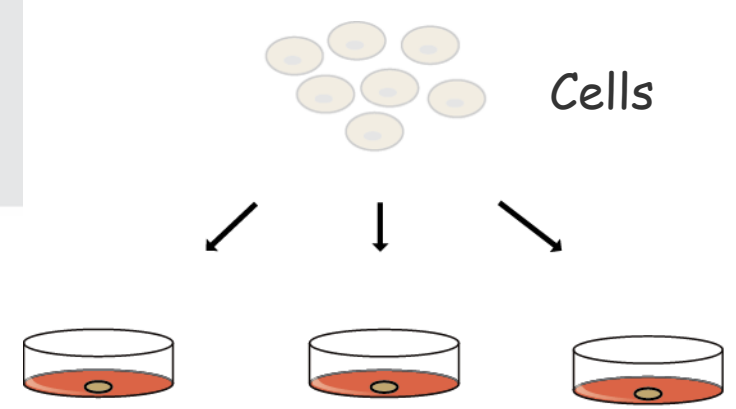


How does single cell work?

1. Cell Separation: Dropseq is a microfluidics-based single cell RNA-sequencing platform that relies on passive co-flow of cells and microparticles (*i.e.* beads) to generate aqueous droplets within oil that contain exactly one cell and one microparticle.

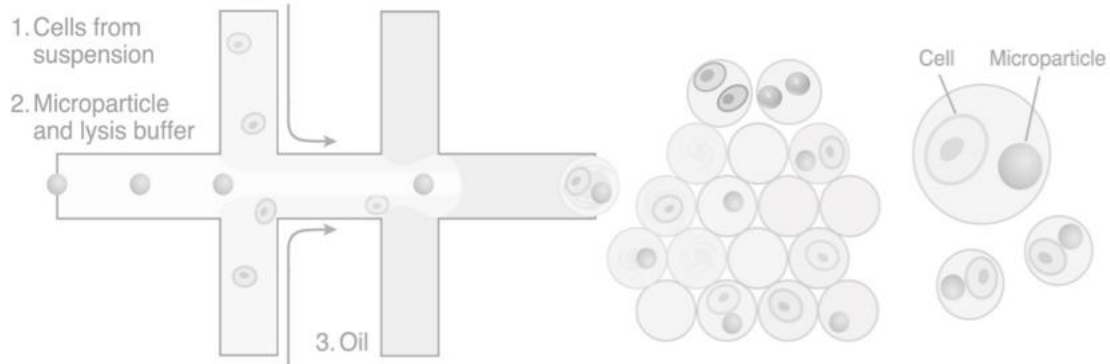
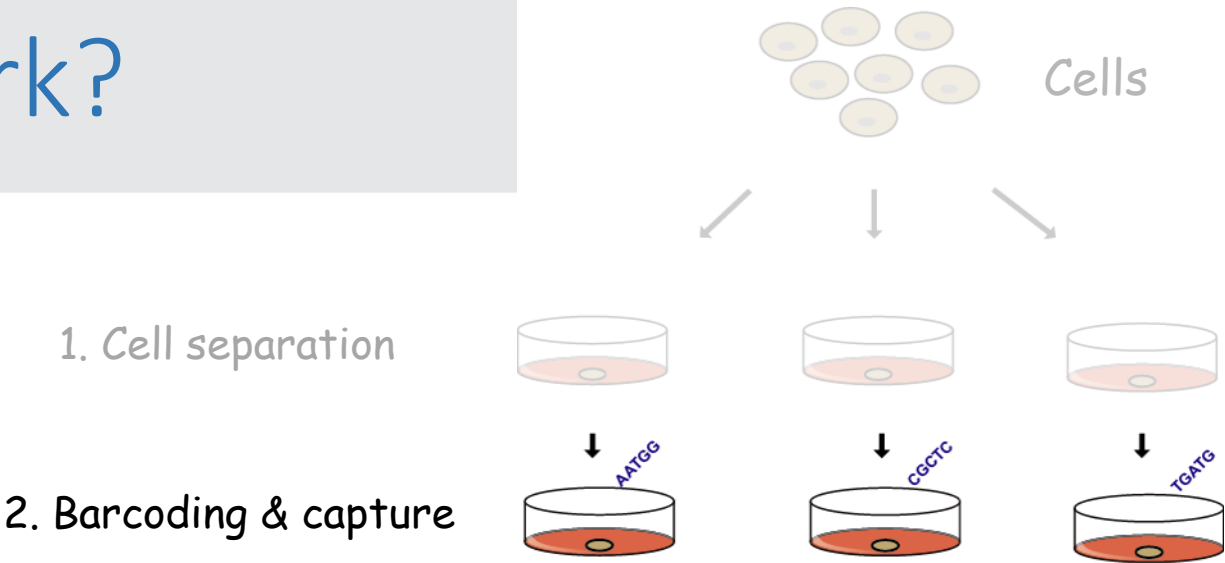


1. Cell separation

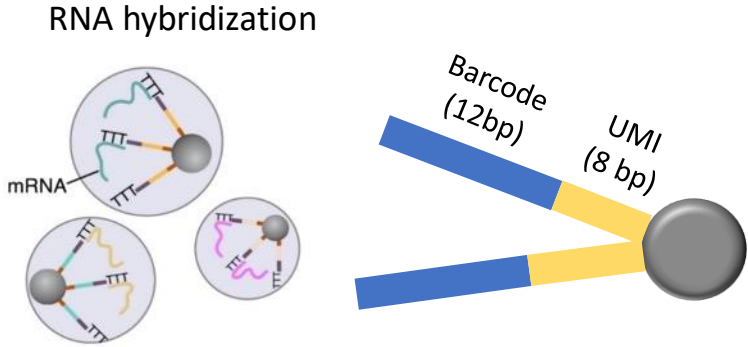


How does single cell work?

2. Barcoding and capture: Each microparticle contains a **unique barcode** and about 10^8 primers that have a poly-T tail to capture RNA content of a cell and a **unique molecular identifier (UMI)**.



Cell Lysis →

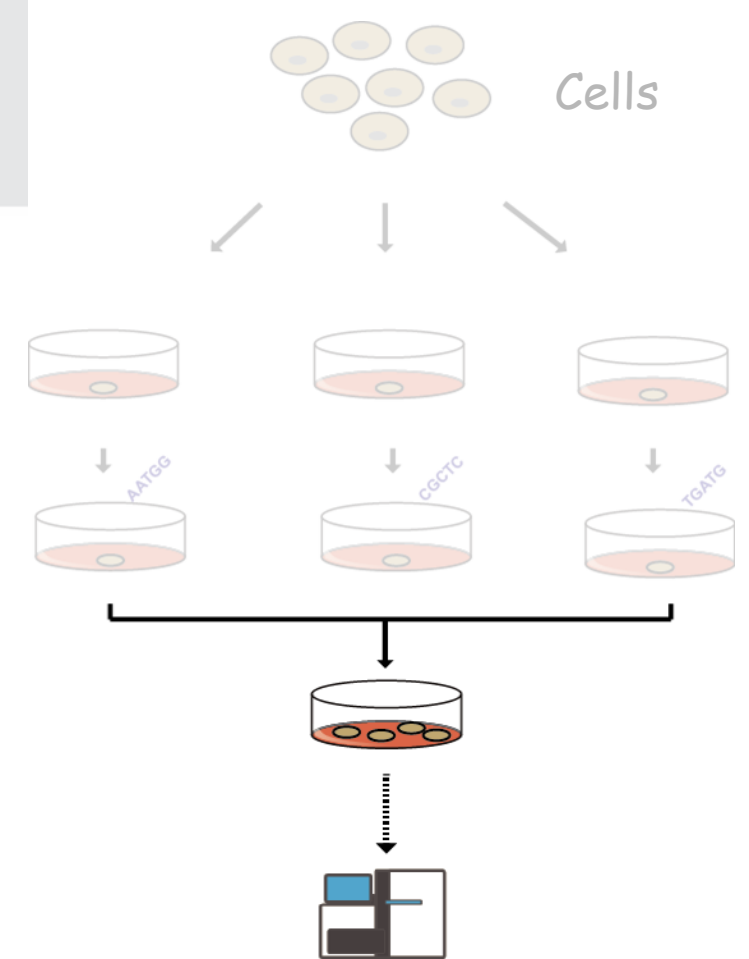


Adapted from Macosko et al. - Cell, 2015

How does single cell work?

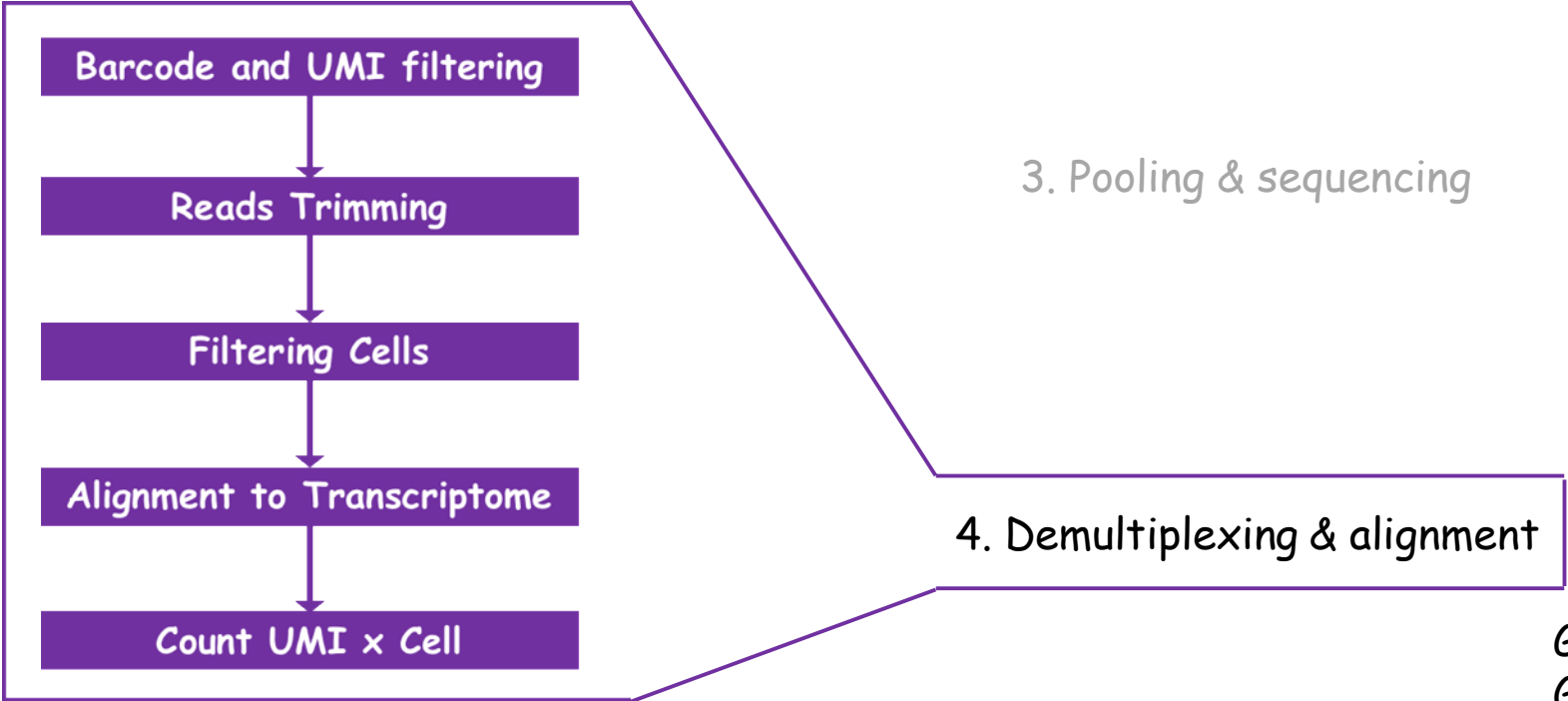
3. Pooling: Captured mRNA transcripts from all the droplets are then collected, reverse-transcribed and amplified in pools in order to be used for standard population-based RNA-sequencing platform and to profile any desired number of cells

1. Cell separation
2. Barcoding & capture
3. Pooling & sequencing

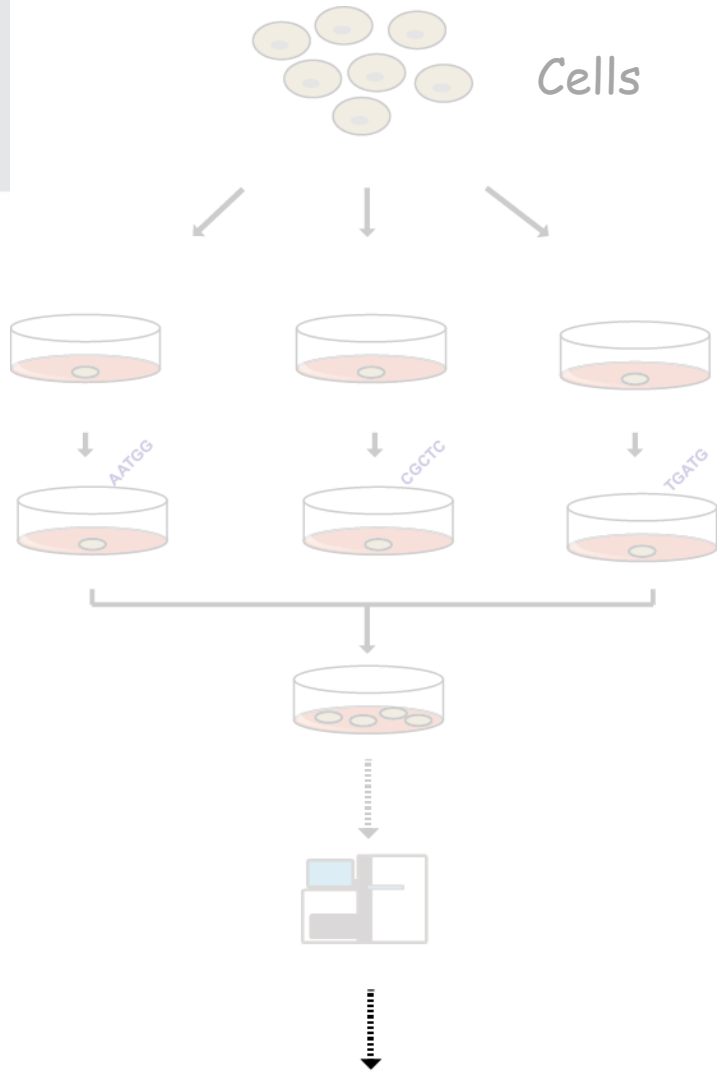


How does single cell work?

4. alignment: Fragments (reads) from the sequencer are then assigned to the right cell of origin with a bioinformatics approach.



- 1. Cell separation
- 2. Barcoding & capture
- 3. Pooling & sequencing



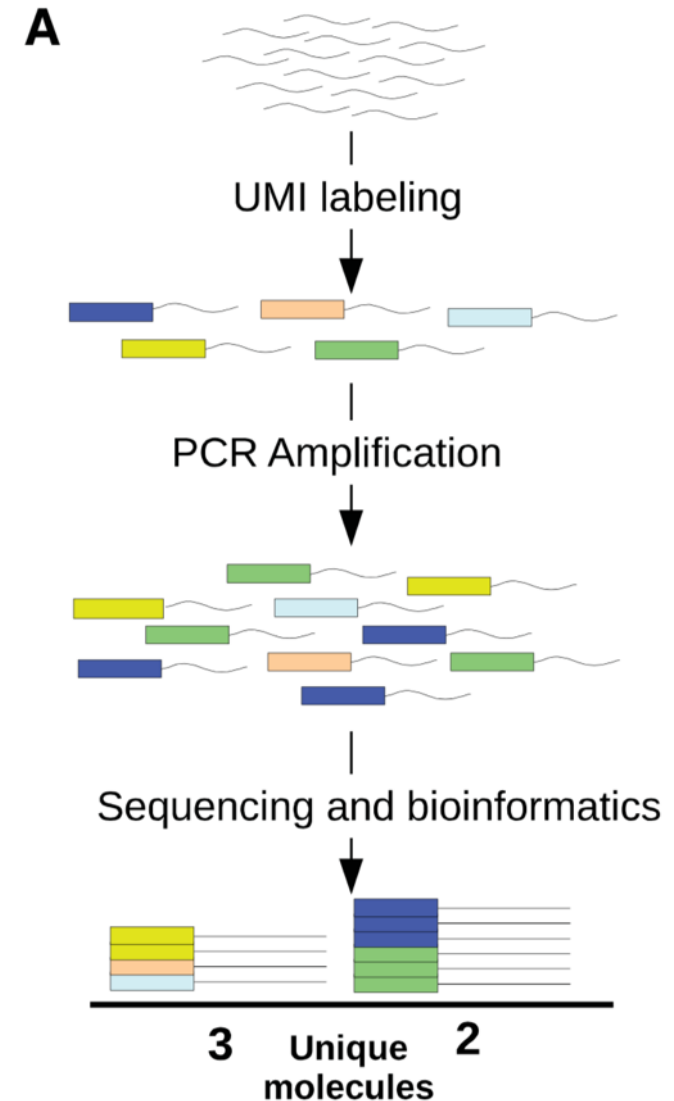
4. Demultiplexing & alignment

	AATGG	CGCTC	CGCTC
Gene 1	3	1	1
Gene 2	2	1	2
Gene 3	1	3	0

Barcodes and UMI

Barcodes: sequencing barcodes are used to assign reads to the cell of origin.

UMI: By incorporating a UMI into the same location in each fragment during library preparation, but prior to PCR amplification, it is possible to accurately identify true PCR duplicates because they have both identical alignment coordinates and identical UMI sequences



In silico reconstruction of thousands of single-cell transcriptomes

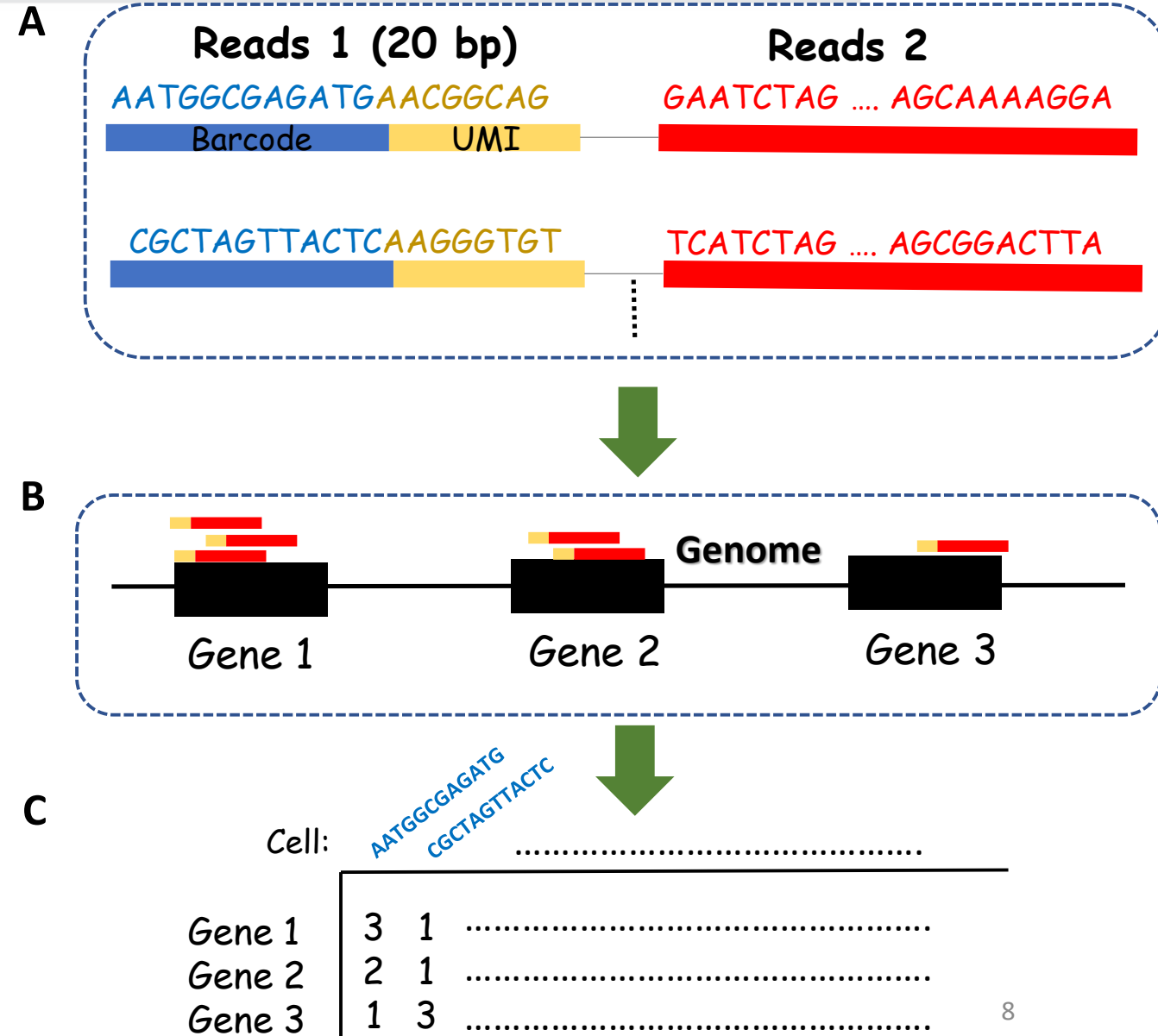


(A) Millions of paired-end reads are generated from a Drop-seq library on a high-throughput sequencer:

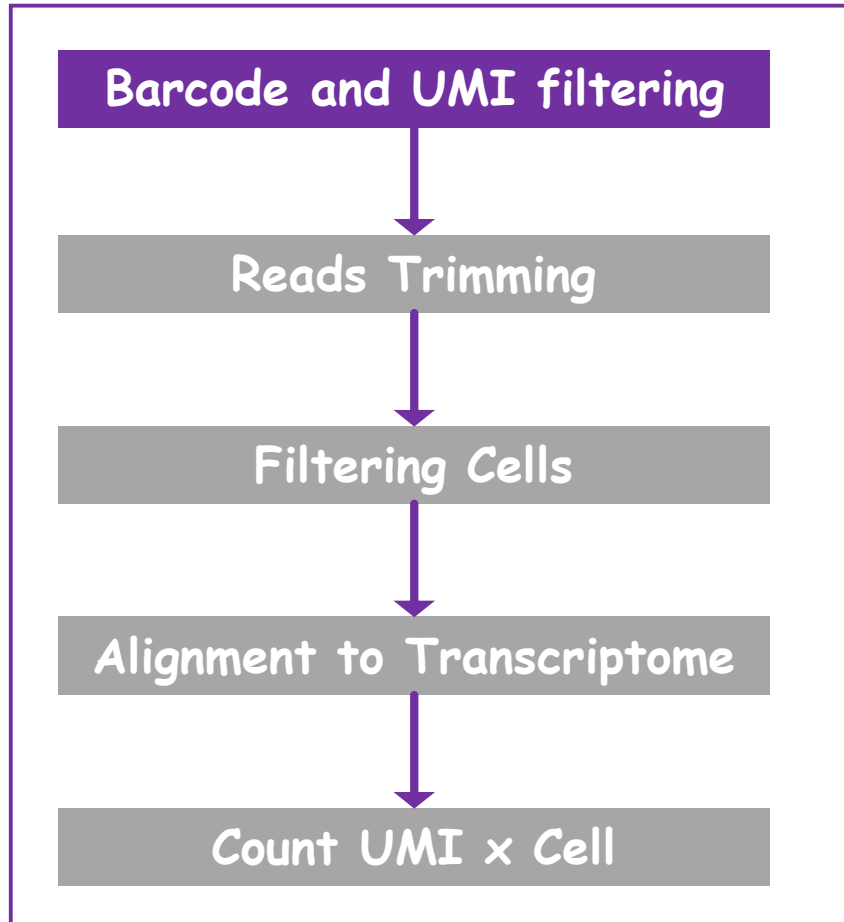
- The first read yields **the cell barcode and UMI**.
- The second, paired read interrogates sequence from the cDNA.

(B) The **second reads are aligned to a reference genome** to identify the gene-of-origin of the cDNA.

(C) Next, **reads are organized by their cell barcodes**, and individual UMIs are counted for each gene in each cell.



Computational Pipeline: Steps

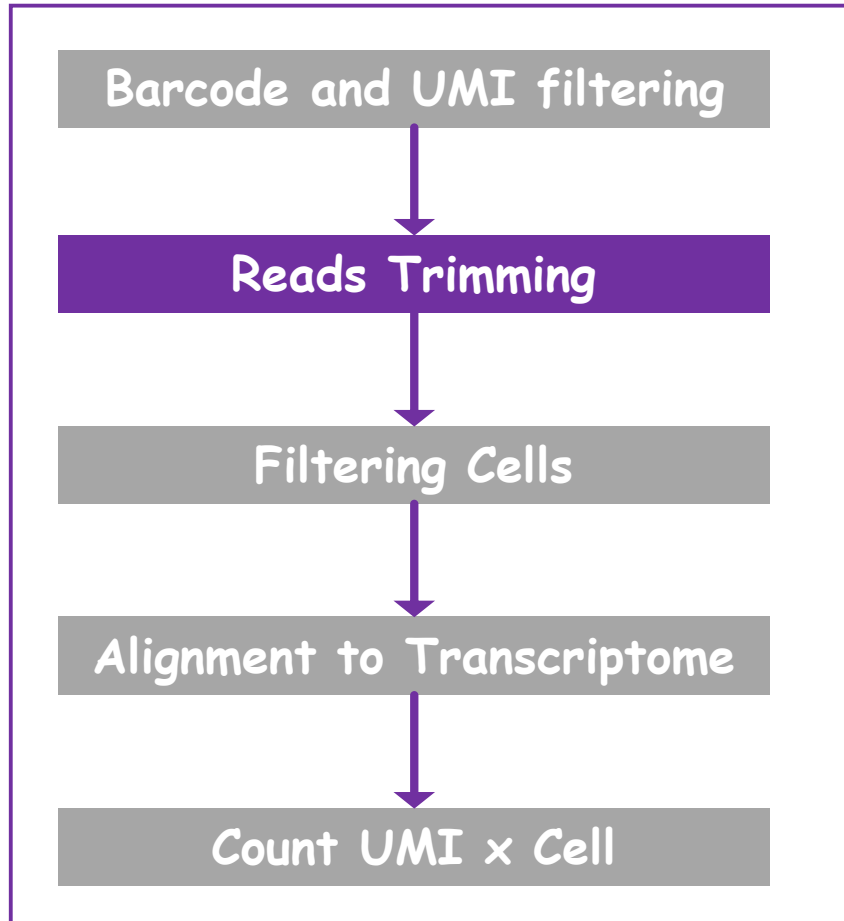


Cell Barcodes and UMIs

- Must not contain N
- Must not contain bases with base quality < 10

```
AACGTCGTGAGT  
CCGCTGACTGAG  
ANTGGCGAGATG
```

Computational Pipeline: Steps

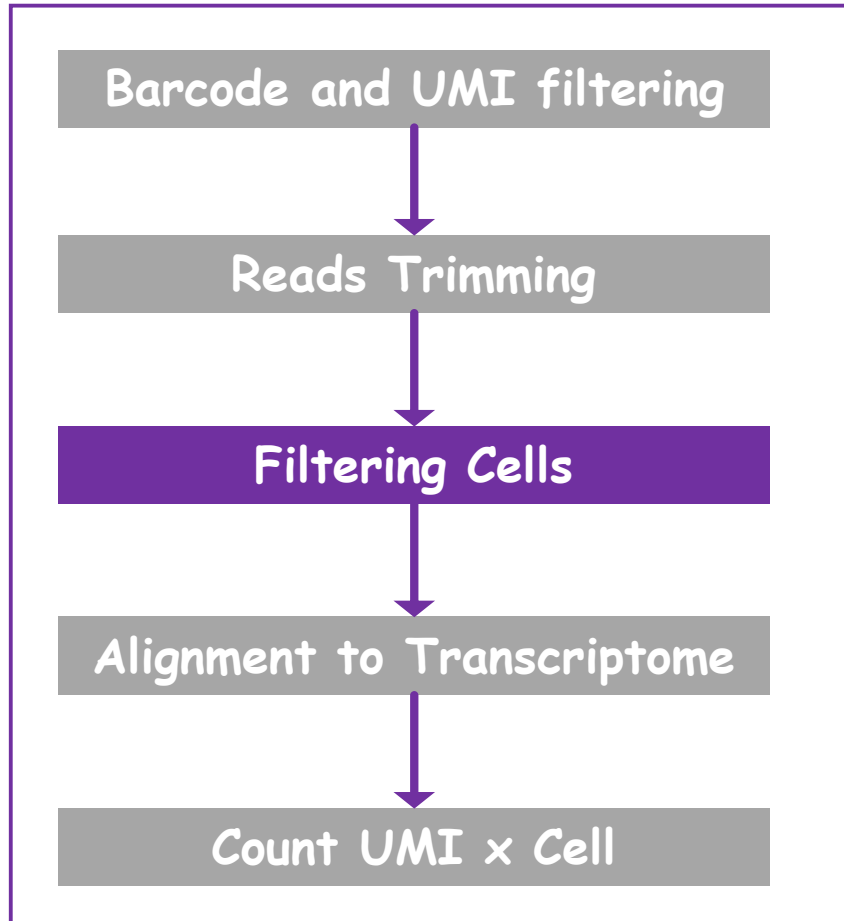


Extra sequence that might have snuck it's way into the reads 2 are trimmed away.

1. SMART Adapter that can occur 5' of the read 2.
 - We search for 5 contiguous bases of the SMART adapter at the 5' end of the read with no mismatches .
2. PolyA tails from reads:
 - We search for at least 6 contiguous A's in the 3' of read 2 with no mismatches.

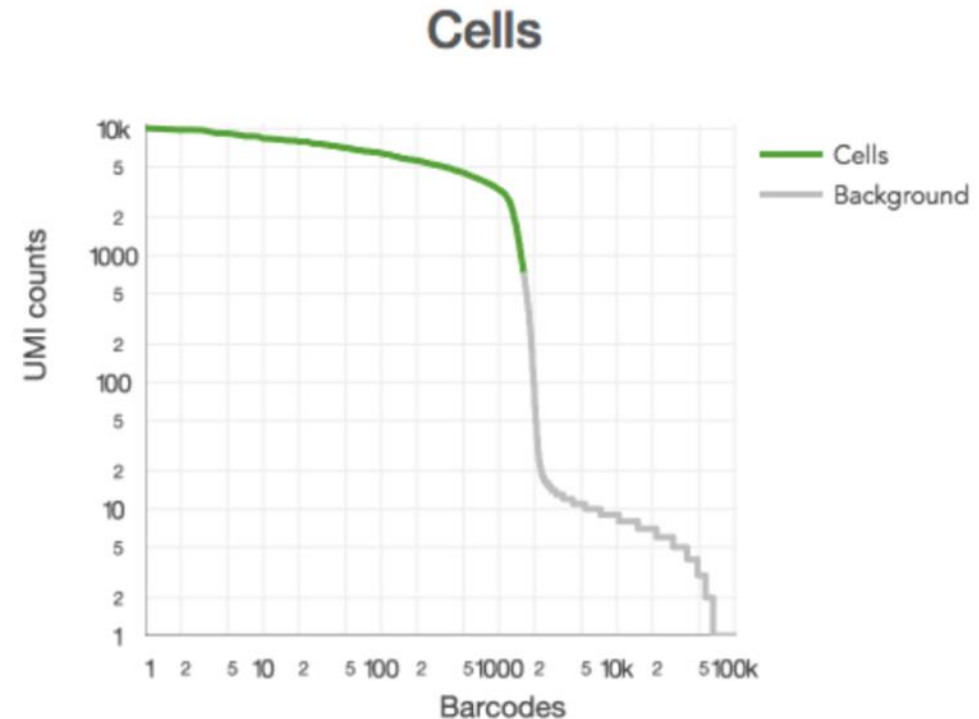
SMART Adapter: AAGCAGTGGTATCAACGCAGAGTGAATGGG

Computational Pipeline: Steps

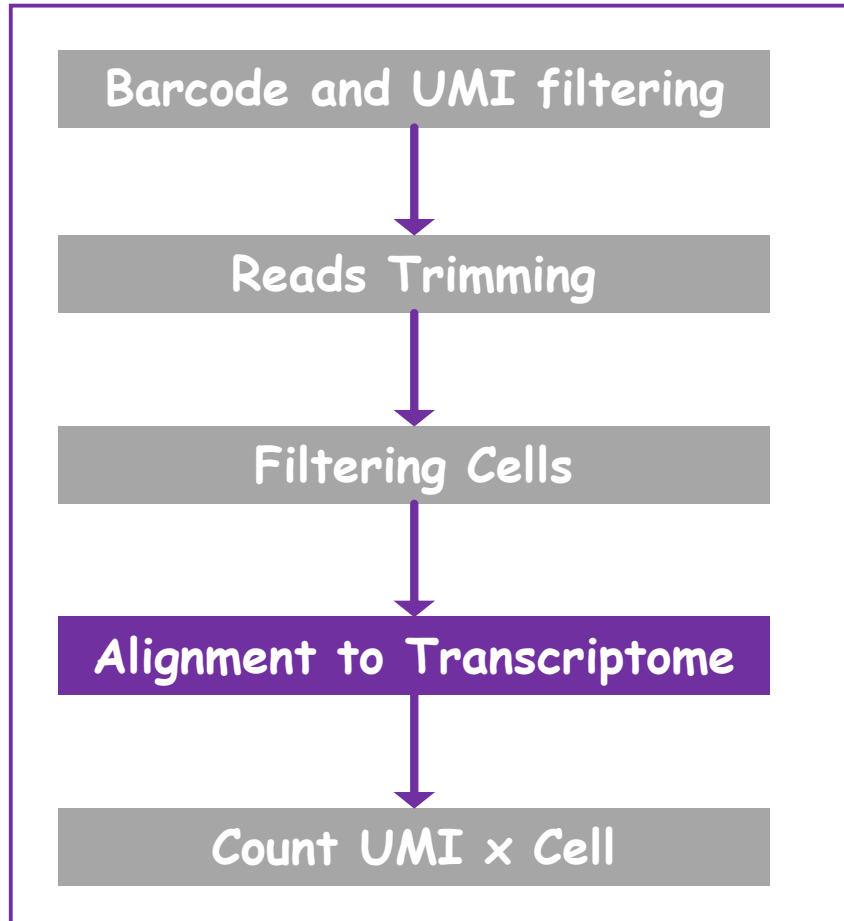


Select BARCODES that likely contain cells

- Sum UMI counts for each barcode
- Select barcodes with total UMI count > 10% of the 99th percentile of the expected recovered cells.



Computational Pipeline: Steps

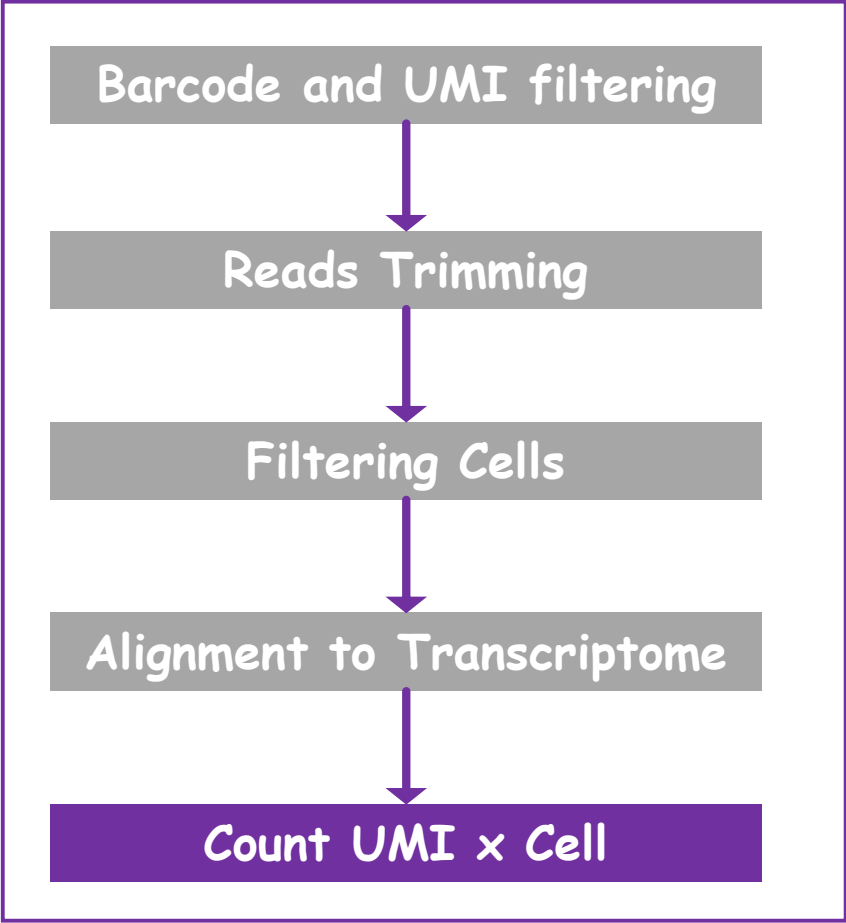


Alignment is done via STAR (Spliced Transcripts Alignment to a Reference)

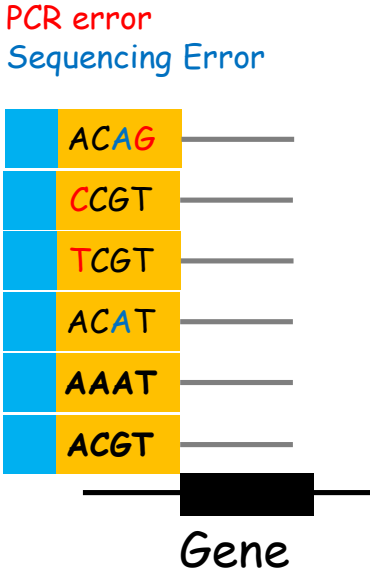
- Robust, open-source, junction-aware RNA-seq aligner
- **Aligns reads to the genome** instead of using only the transcriptome

We only retain uniquely mapped reads, thus reads mapping on multiple area of the genome are discarded.

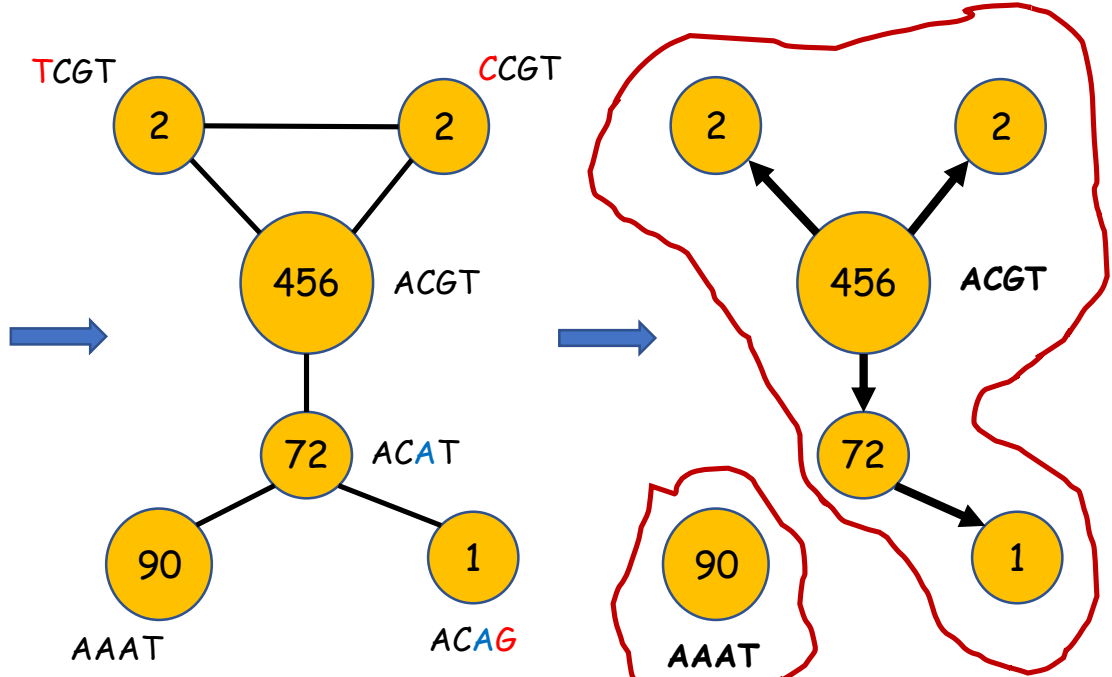
Computational Pipeline: Steps



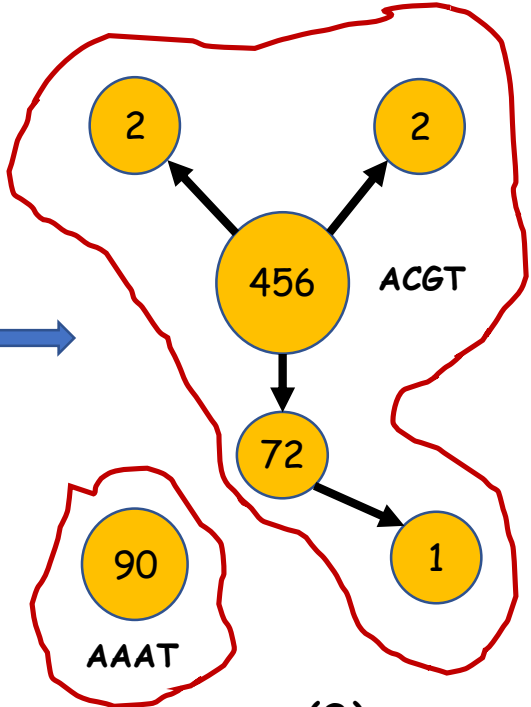
UMIs that are 1 nucleotide mismatch away from a higher-count UMI are corrected to that UMI using a network approach (Smith et al. 2017, Genome Research).



(1)



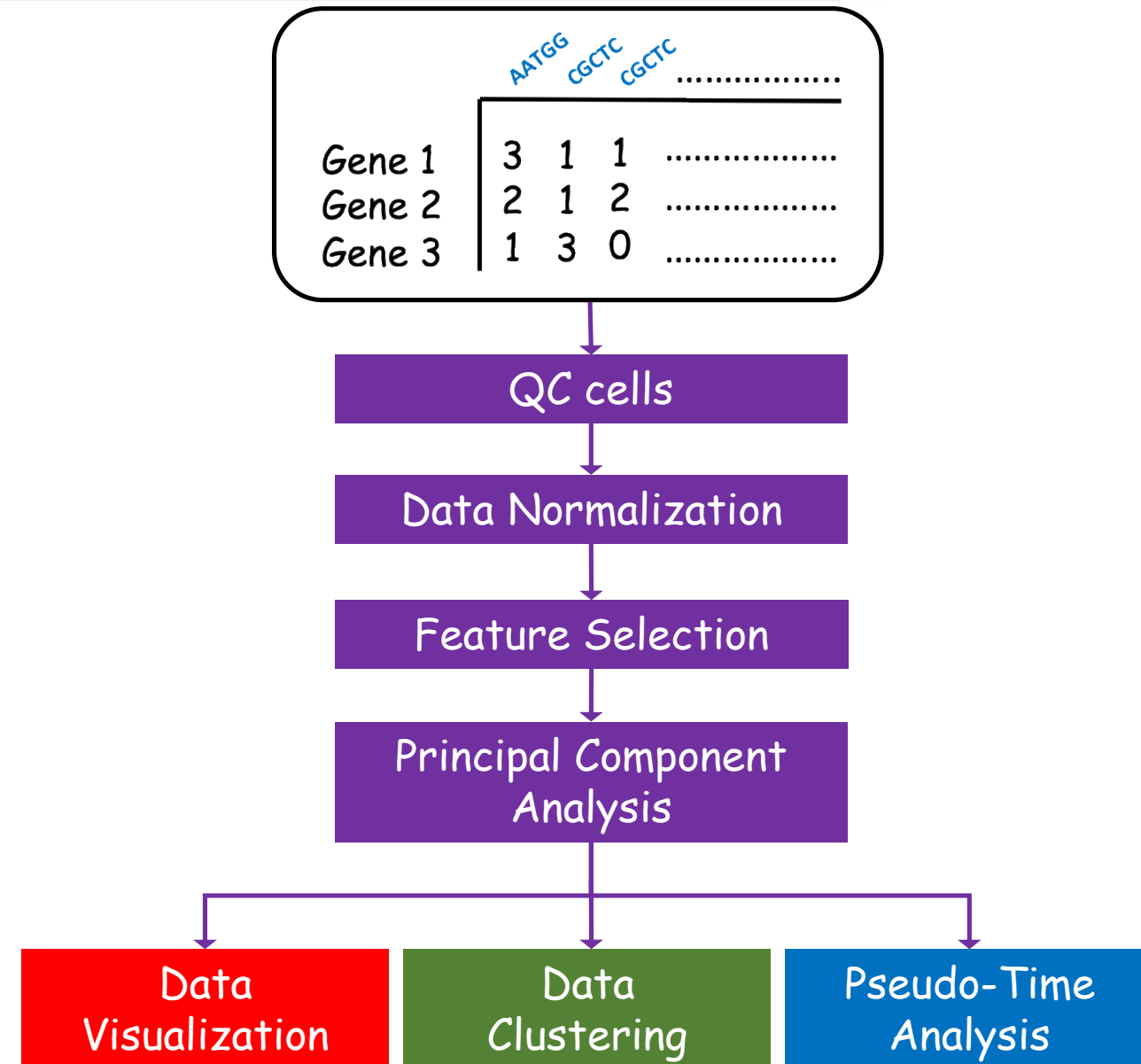
(2)



(3)

Downstream Analyses

- Once we have aligned reads and obtained the matrix containing the raw gene transcripts per cell. We can proceed normalizing the data and removing confounder factors present in the data.
- Then, we can carry out analyses that are relevant to answer our biological question.
- The exact nature of the analysis depends on the dataset and the aim of the project.

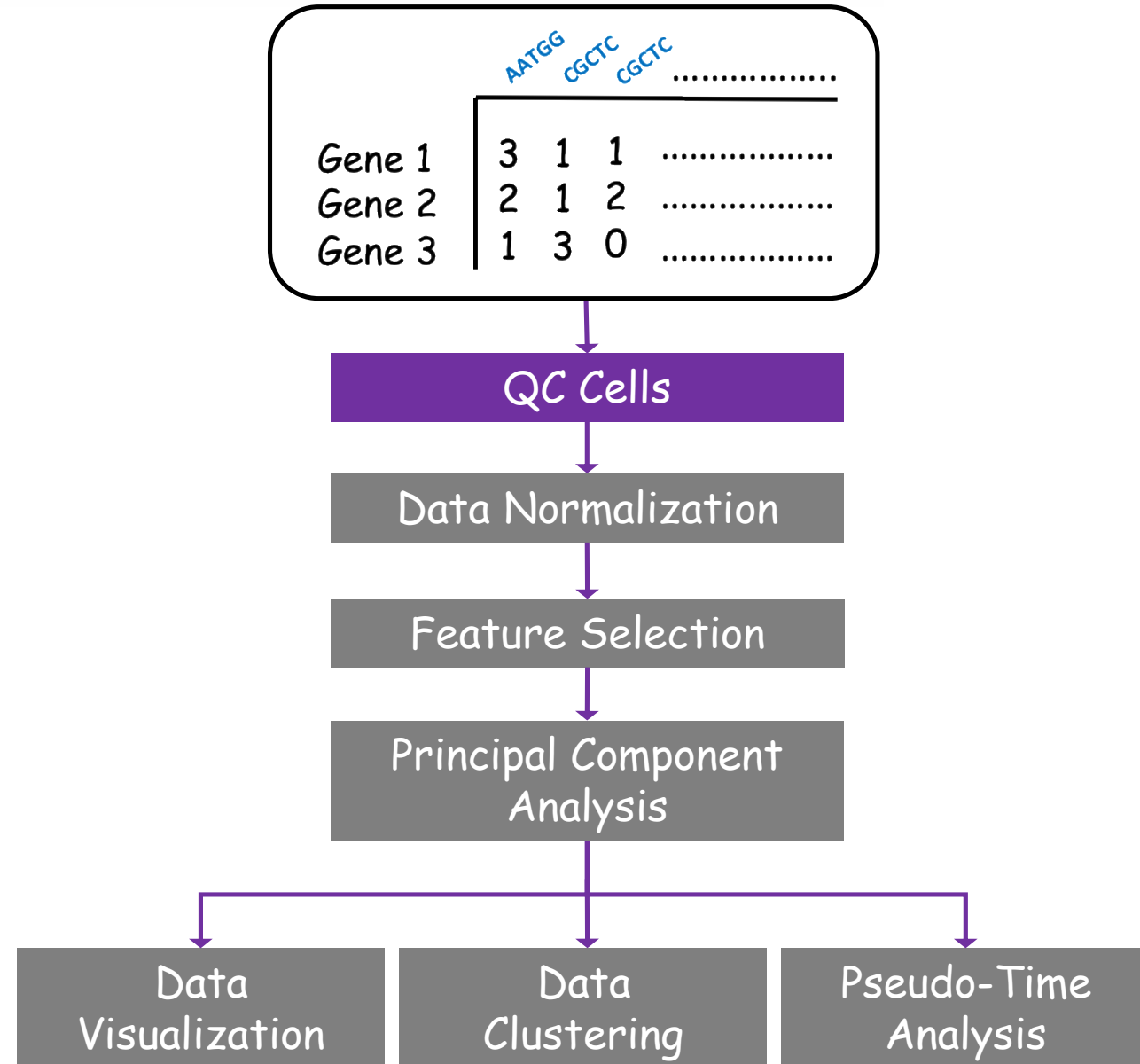


Filter Out not “good” Cells



Quality control metrics commonly used by the community are actually based on:

1. The number of unique genes detected in each cell.
 - Low-quality cells or empty droplets will often have very few genes
 - Cell doublets or multiplets may exhibit an aberrantly high gene count
2. The total number of UMI detected within a cell.
3. The percentage of reads that map to the mitochondrial genome.
 - Low-quality / dying cells often exhibit extensive mitochondrial contamination



An example of very simple normalization method

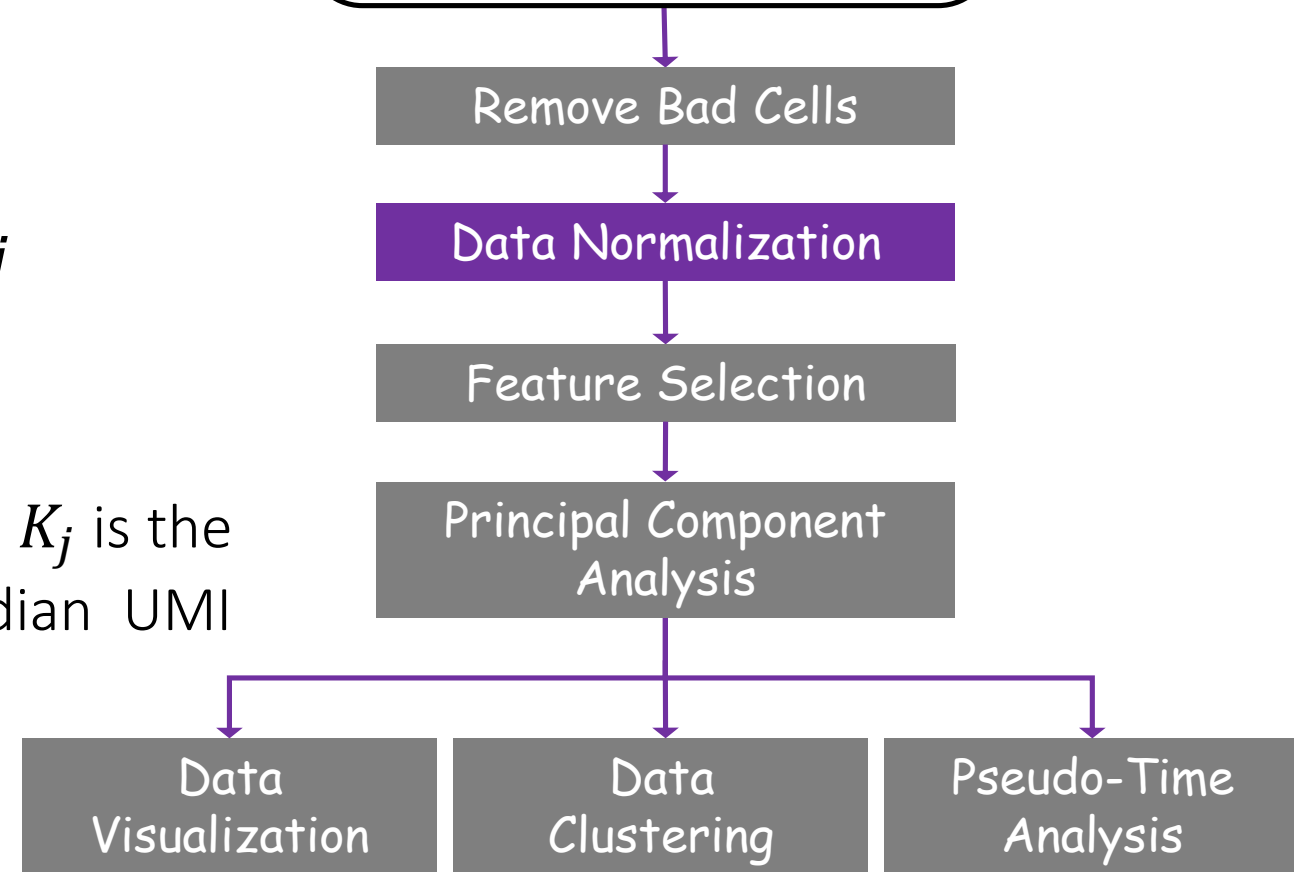


UMI normalization is performed by first dividing UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across cells.

$$\hat{k}_{i,j} = k_{i,j} \cdot \frac{\bar{K}}{K_j} \quad K_j = \sum_i k_{i,j}$$

Where $k_{i,j}$ is the UMI count of gene i in cell j , K_j is the total UMI count of cell j and \bar{K} is the median UMI count among the cell population.

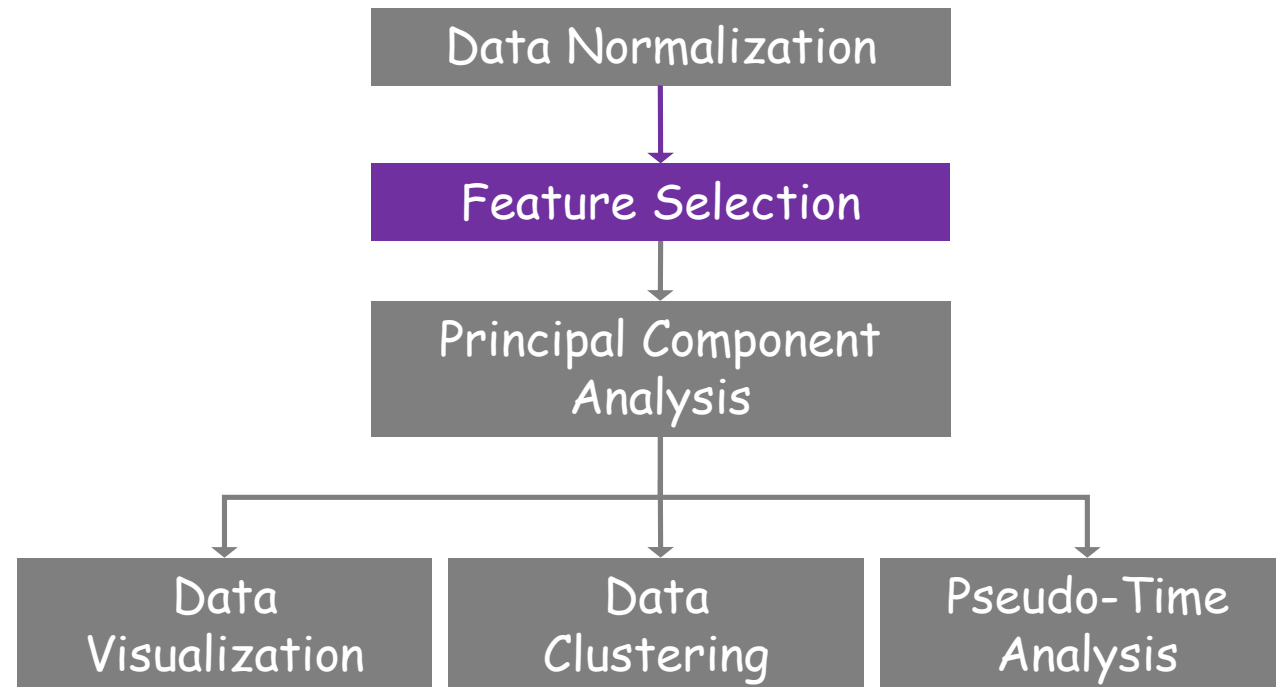
	AATGG	CGCTC	CGCTC
Gene 1	3	1	1
Gene 2	2	1	2
Gene 3	1	3	0



Feature Selection



- Single-cell RNASeq is capable of measuring the expression of many thousands of genes in every cell.
- Only a portion of those will show a response to the biological condition of interest, e.g. differences in cell-type, drivers of differentiation and so on...



Feature Selection: Highly Variable Genes (HVG)



- HVG **assumes that** if genes have large differences in expression across cells some of those **differences are due to biological difference between the cells rather than technical noise.**
- HGV consist in **identifying genes that exhibit high cell-to-cell variation in the dataset** (i.e, they are highly expressed in some cells, and lowly expressed in others).
- However, because of the nature of count data, **there is a positive relationship between the mean expression of a gene and the variance in the read counts across cells.** This relationship must be corrected for to properly identify HVGs.

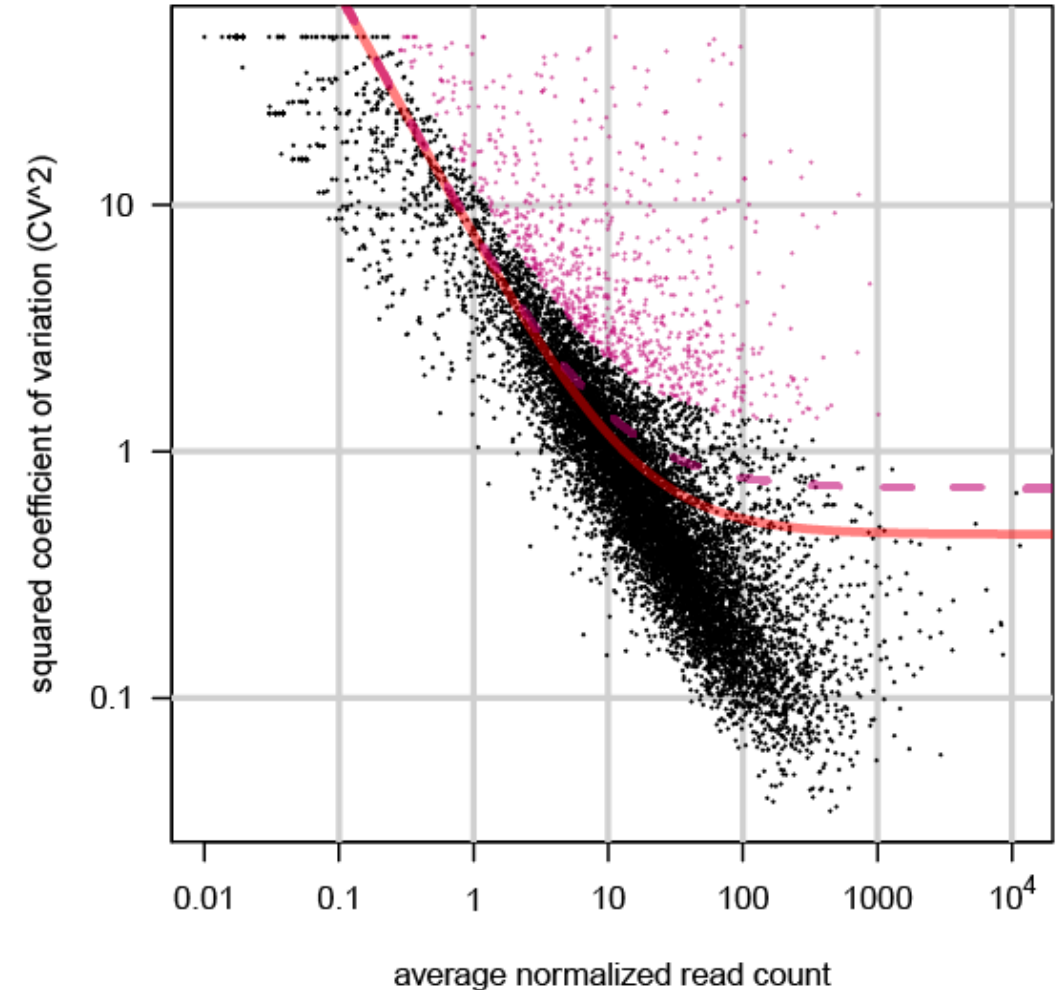
Feature Selection: Highly Variable Genes (HVG)

A popular method to correct for the relationship between variance and mean expression was proposed by [Brennecke et al.](#)

Brennecke method consist in:

1. normalize data for library size
2. calculate the mean and the square coefficient of variation (CV^2) for each gene.
3. a quadratic curve is fit to the relationship between these two variables
4. a chi-square test is used to find genes significantly above the curve.

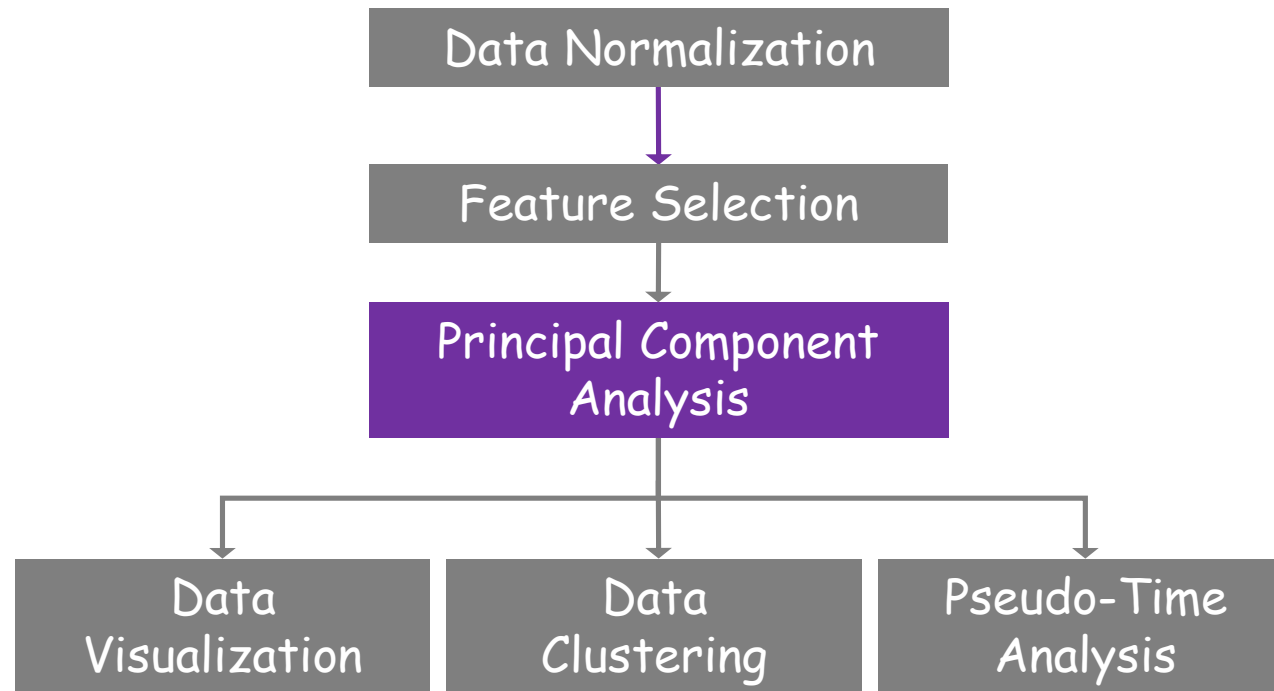
$$CV^2 = \frac{\delta^2}{\mu^2}$$



Dimensionality Reduction: PCA

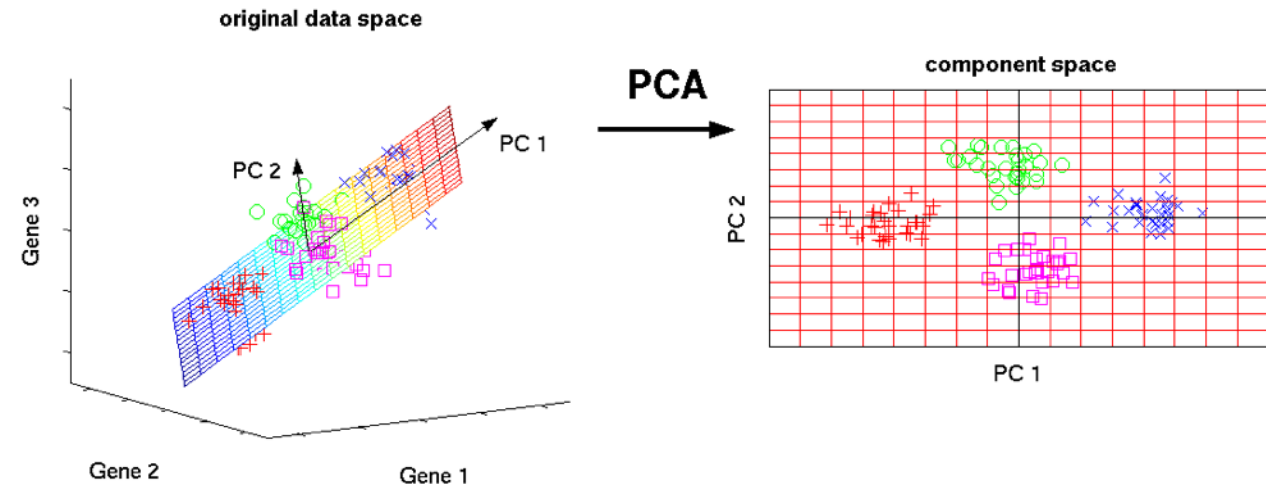


- Low variance can often be assumed to represent undesired background noise.
- The dimensionality of the data can therefore be reduced, without loss of relevant information, by extracting a lower dimensional component space covering the highest variance.
- Using a lower number of principal components instead of the high-dimensional original data is a common pre-processing step that often improves results of subsequent analyses such as classification and clustering.



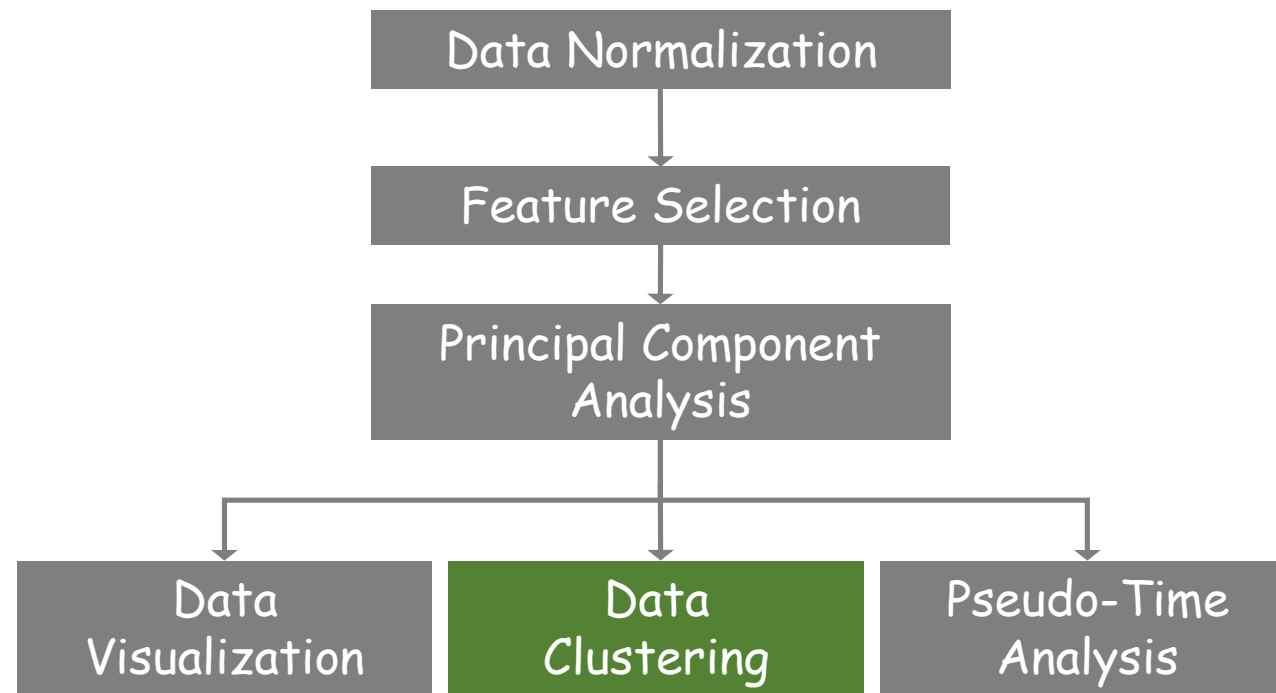
Dimensionality Reduction: PCA

- **Principal component analysis (PCA)** is a statistical procedure that uses a transformation to convert a set of observations into a set of values of linearly uncorrelated variables called principal components (PCs).
- The number of PCs is less than or equal to the number of original variables.
- Mathematically, the PCs correspond to the eigenvectors of the covariance matrix.
- The eigenvectors are sorted by eigenvalue so that **the first principal component accounts for as much of the variability in the data as possible**, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.



Data Clustering

- One of the most promising applications of scRNA-seq is *de novo discovery and annotation of cell-types* based on transcription profiles.
- Computationally, this is a hard problem as it amounts to unsupervised clustering.
- We need to identify groups of cells based on the similarities of the transcriptomes without any prior knowledge of the labels and the number of groups.



Data Clustering: Clustering methods



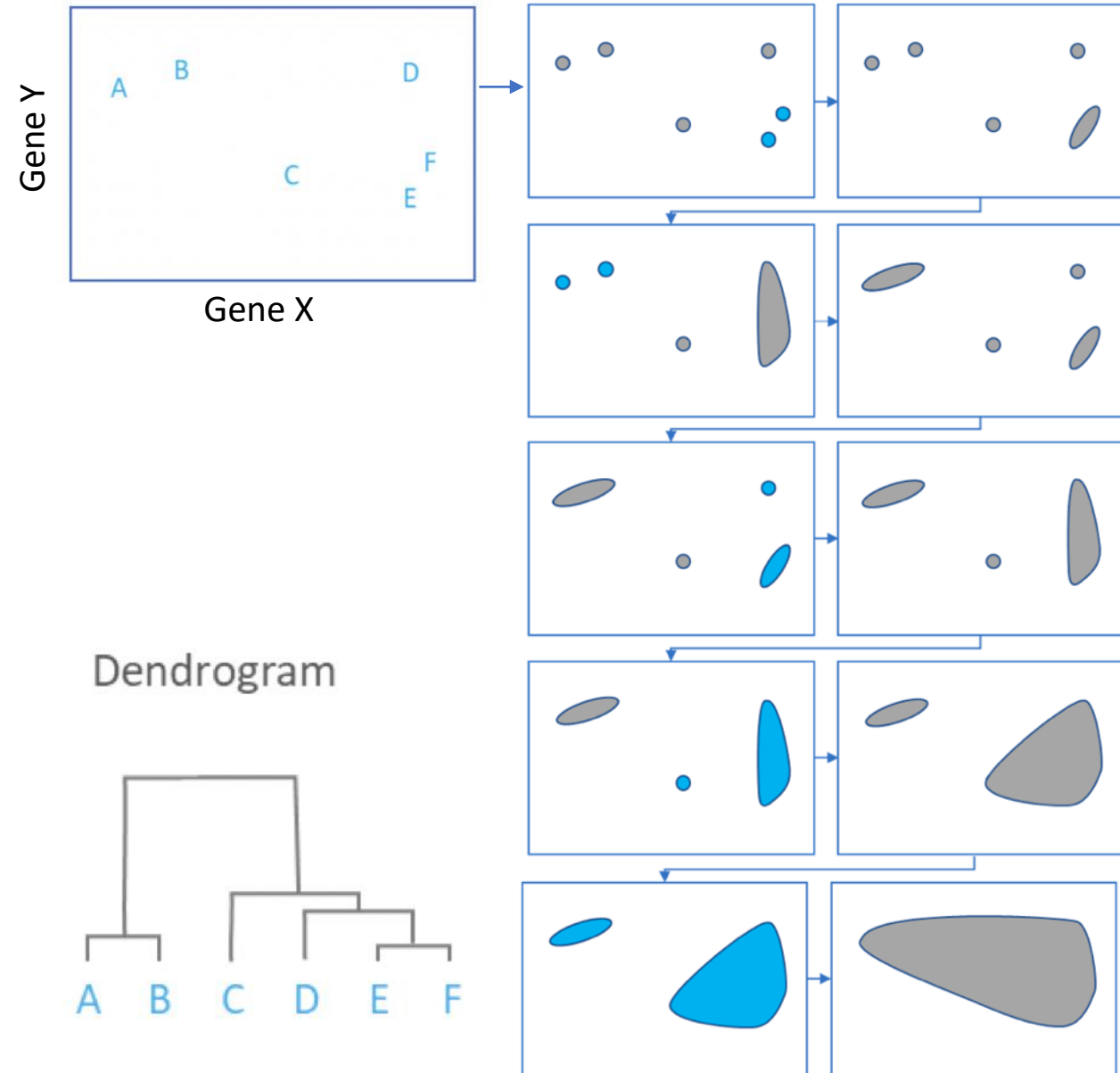
- **Unsupervised clustering** is useful in many different applications and it has been widely studied in machine learning.
- The most popular clustering approaches are the:
 - hierarchical clustering
 - k-means clustering
 - graph-based clustering

Data Clustering: Hierarchical clustering

- Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters generally divided into two types:

1. **Agglomerative**: Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
2. **Divisive**: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

- Merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.



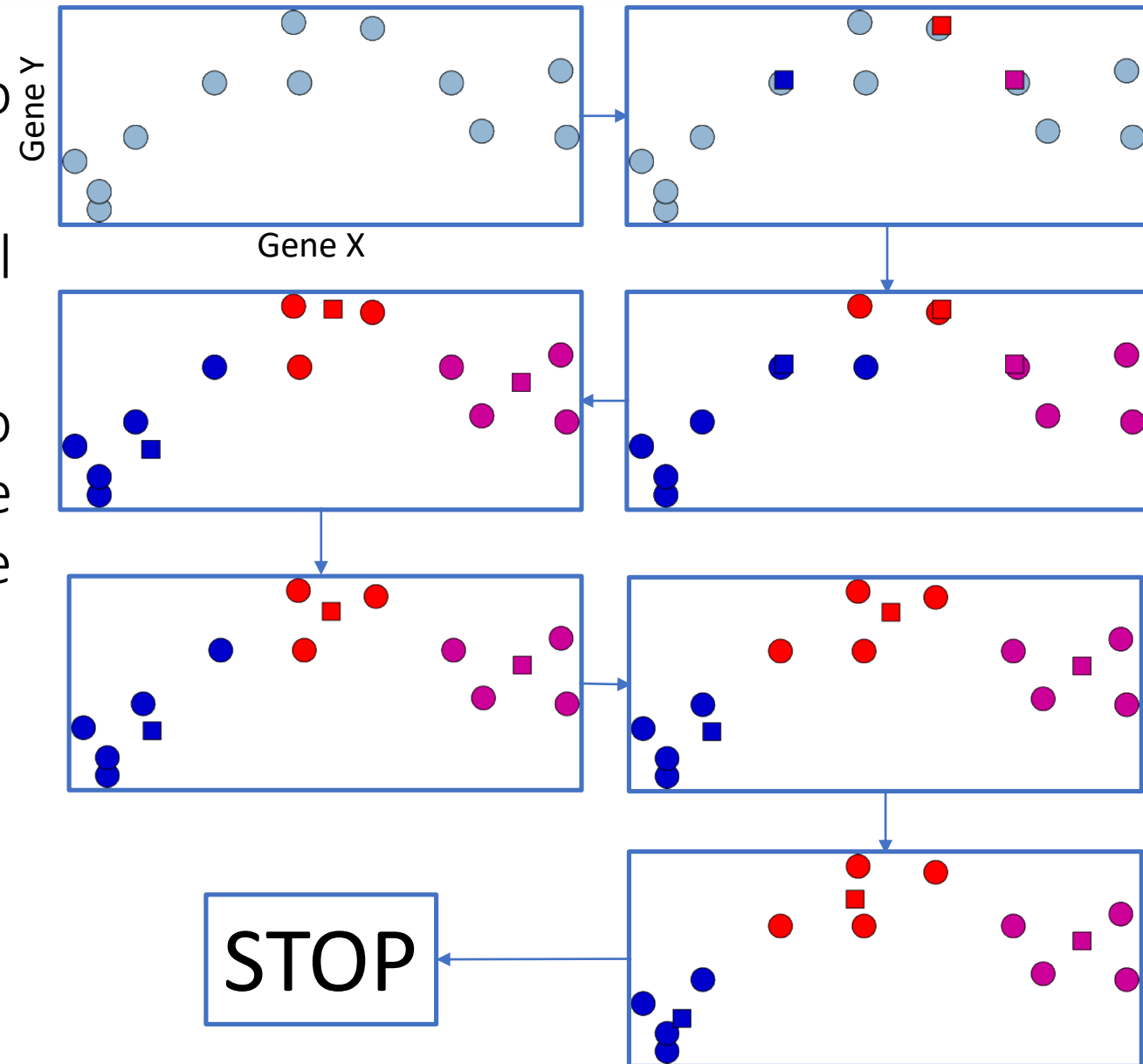
Data Clustering: k-means

In k-means clustering, the goal is to partition N cells into k different clusters.

Initialization Step: Start with some initial cluster centers

At each iteration: assign each point to the closest center and recalculate centers as the mean of the points in the new cluster.

Needs to know the k expected clusters



Data Clustering: Graph-Based methods

Are methods able to identify groups or modules of nodes in a network (or graph).

Graph-based methods are very efficient and can be applied to networks containing millions of nodes.

Definitions of Graph:

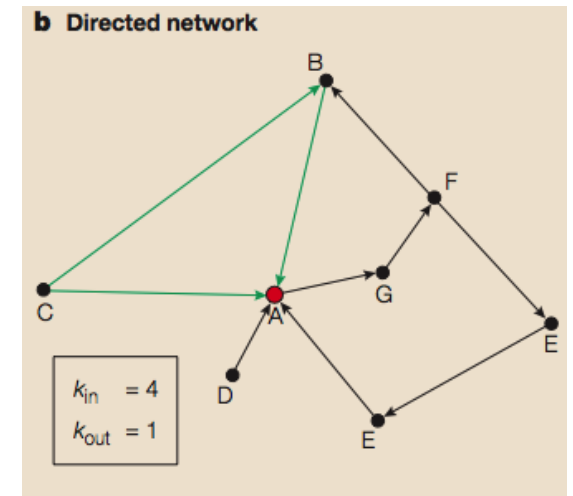
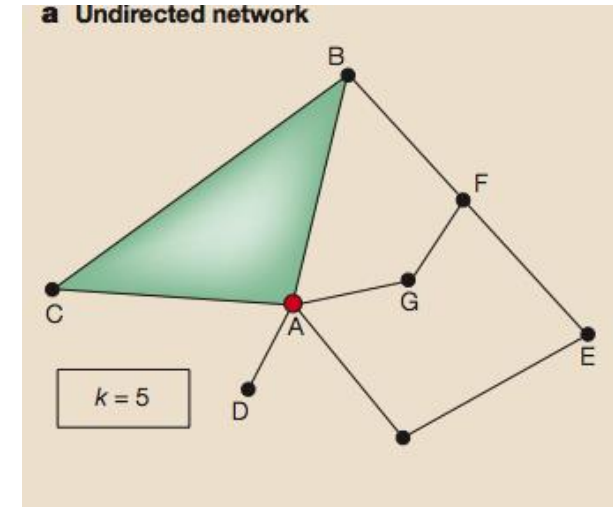
graph $G=\{V,E\}$ where V is a set of **vertices or nodes**, and E is a set of **edges**

degree k : number of edges connected to a node

digraph: the edges have a direction

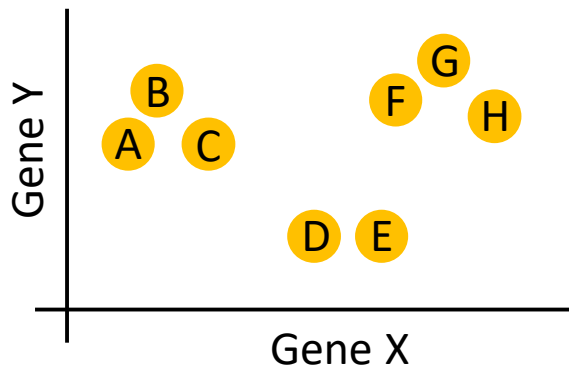
$P(k)$ degree distribution: probability that a node has degree k : $P(k)=N(k)/N$

How can I obtain a weighted graph of cells?



Obtaining a network of cell

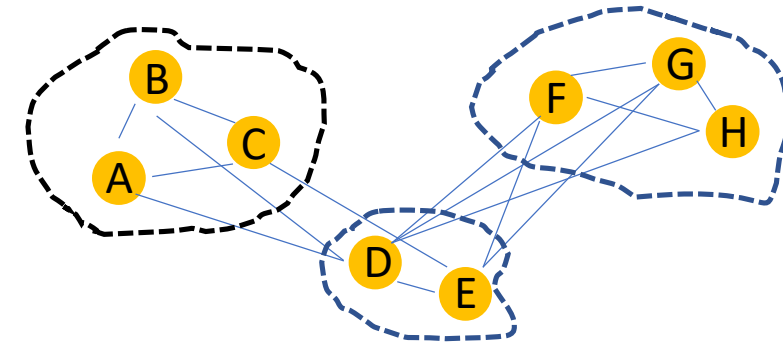
1. We construct a K-nearest neighbour (KNN) graph based on the Euclidean distance among cells
2. Edge weights between any two cells is based on the shared overlap in their local neighbourhood (Jaccard similarity)
3. Once Network of cell is built you can apply any community algorithm detection you want. The one actually most used is called Louvian



1

	A	B	C	D	E	F	G	H
A	0	.5	.7	2	3	4	5	6
B	.5	0	.6	2	3	4	5	6
C	.7	.6	0	1	2	3	4	5
D	2	2	1	0	.5	1	2	3
E	3	3	2	.5	0	.8	1	.9
F	4	4	5	1	.8	0	.5	.7
G	5	5	4	2	1	.5	0	.7
H	6	6	5	3	.9	.7	.7	0

2



$$Jaccard(A, b) = \frac{\# \text{ of shared neighb.}}{\text{total number of neighb.}}$$

Euclidean distance among cells

Single cell Data Visualization

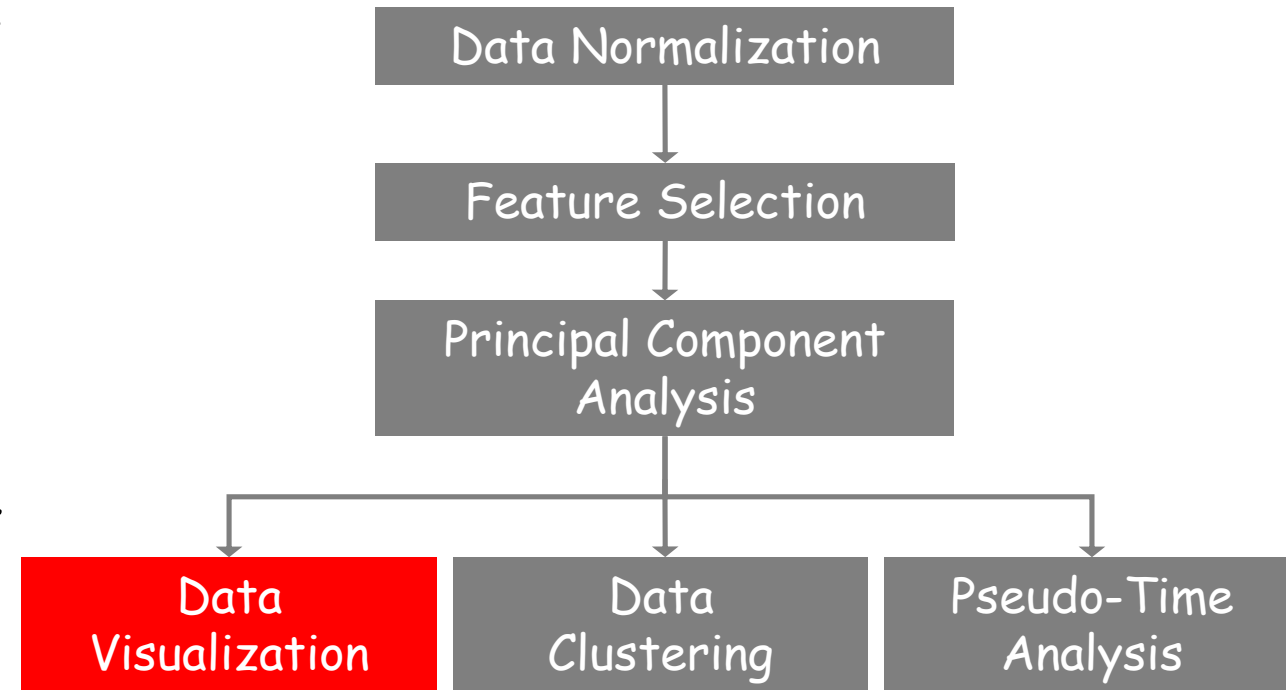


To visualize and explore single cell dataset two popular non-linear dimensional reduction techniques are actually used:

- tSNE (t-Distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)

The goal of these algorithms is to learn the underlying manifold of the data in order to *place similar cells together in low-dimensional space*.

Cells within the graph-based clusters determined as shown before usually co-localize on these dimension reduction plots.

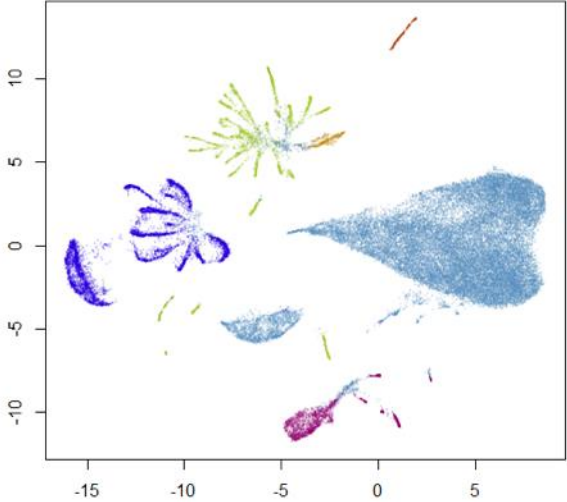


Umap & t-SNE example

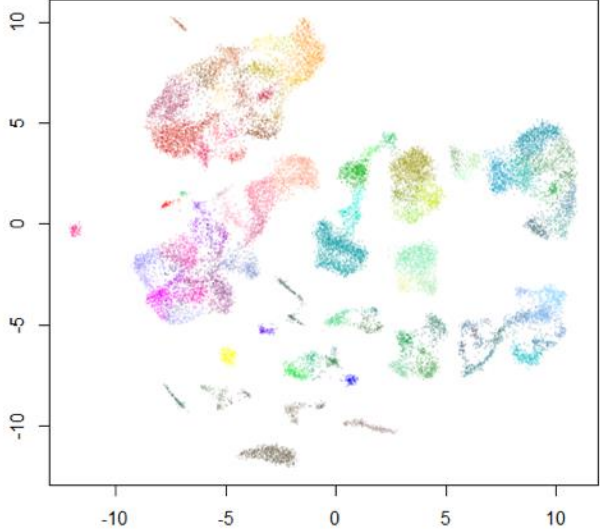


They usually perform similarly although UMAP tend stretch the distance between cluster of cells.

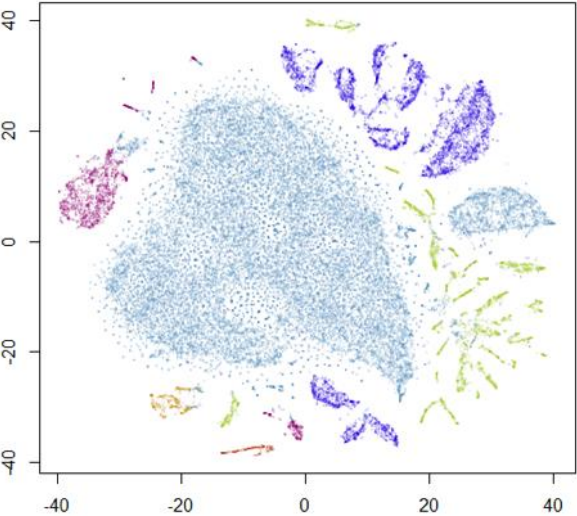
macosko2015 UMAP (Z-scaled)



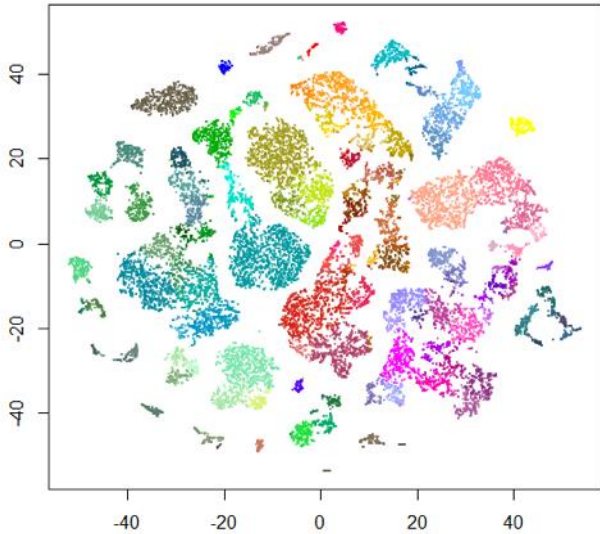
tasic2018 UMAP (a = 2, b = 2)



macosko2015 t-SNE (Z-scaled)



tasic2018 t-SNE



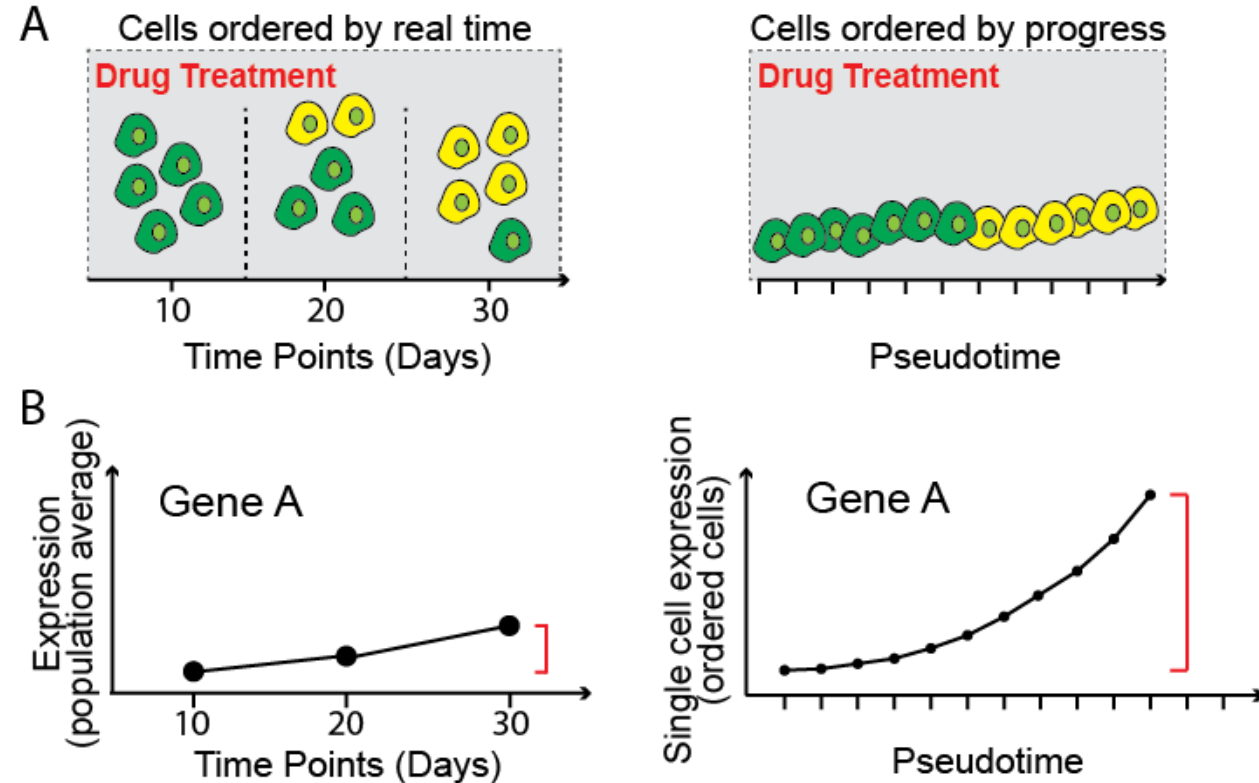
Pseudo-Time Analysis: Concepts

Problem:

- In many situations, one is studying a process where cells change continuously from one type to another (e.g. processes taking place during development)
- Ideally, we would like to monitor the expression levels of an individual cell over time.
- Unfortunately, such monitoring is not possible with scRNA-seq since the cell is lysed (destroyed) when the RNA is extracted.

Solution:

- We can sample cells at multiple time-points and obtain snapshots of the gene expression profiles.
- Since some of the cells will proceed faster along the differentiation than others, each snapshot may contain cells at varying points along the developmental progression.
- We use statistical methods to order the cells along one or more trajectories which represent the underlying developmental trajectories, this ordering is referred to as “pseudotime”.



THE END