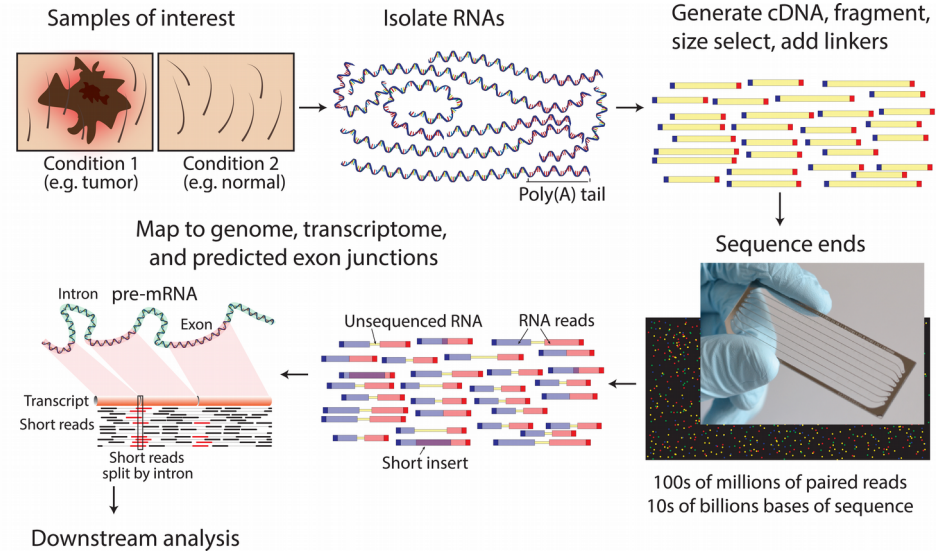
SINGLE-CELL RNA-SEQ

Presentation to initiatize a project

Bioinformatics Core Facility
bigr@gustaveroussy.fr
B2M+1, Gustave-Roussy

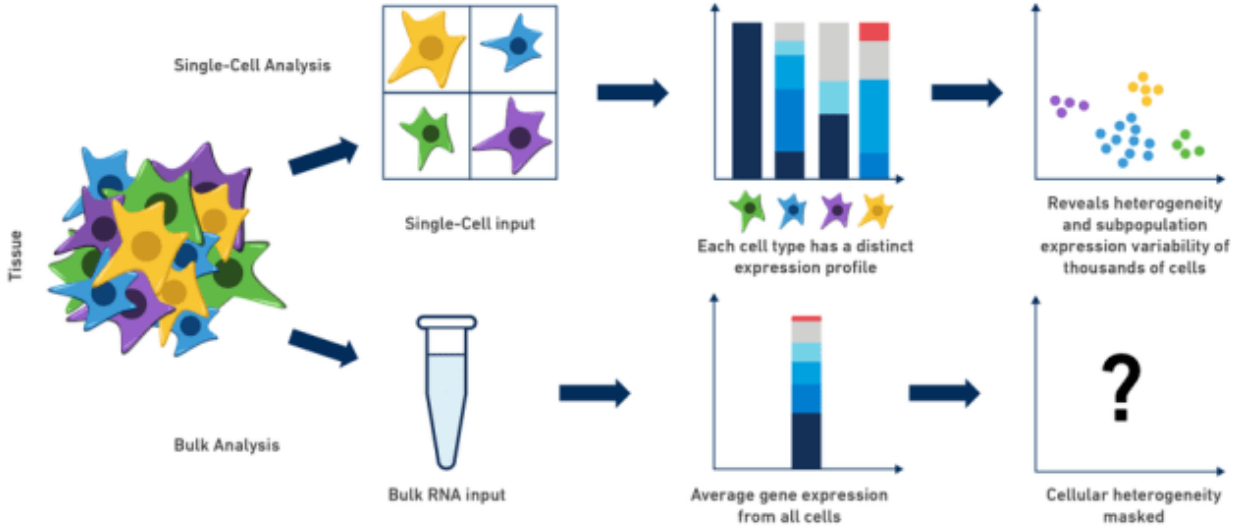
GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Bulk RNA-seq vs Single-cell RNA-seq



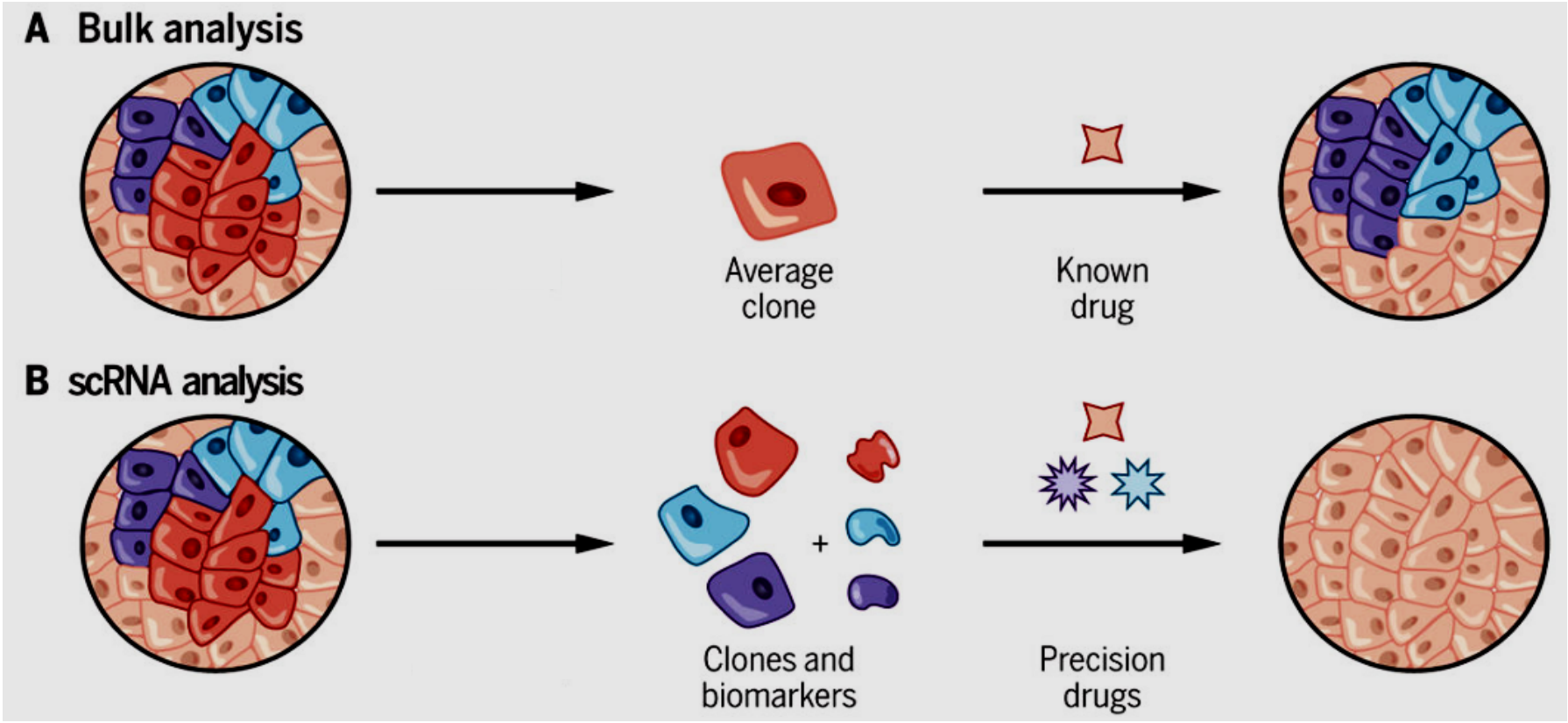
Griffith et al. (2015) PLOS

Measures the **average expression level** of each gene in a **population** of cells.



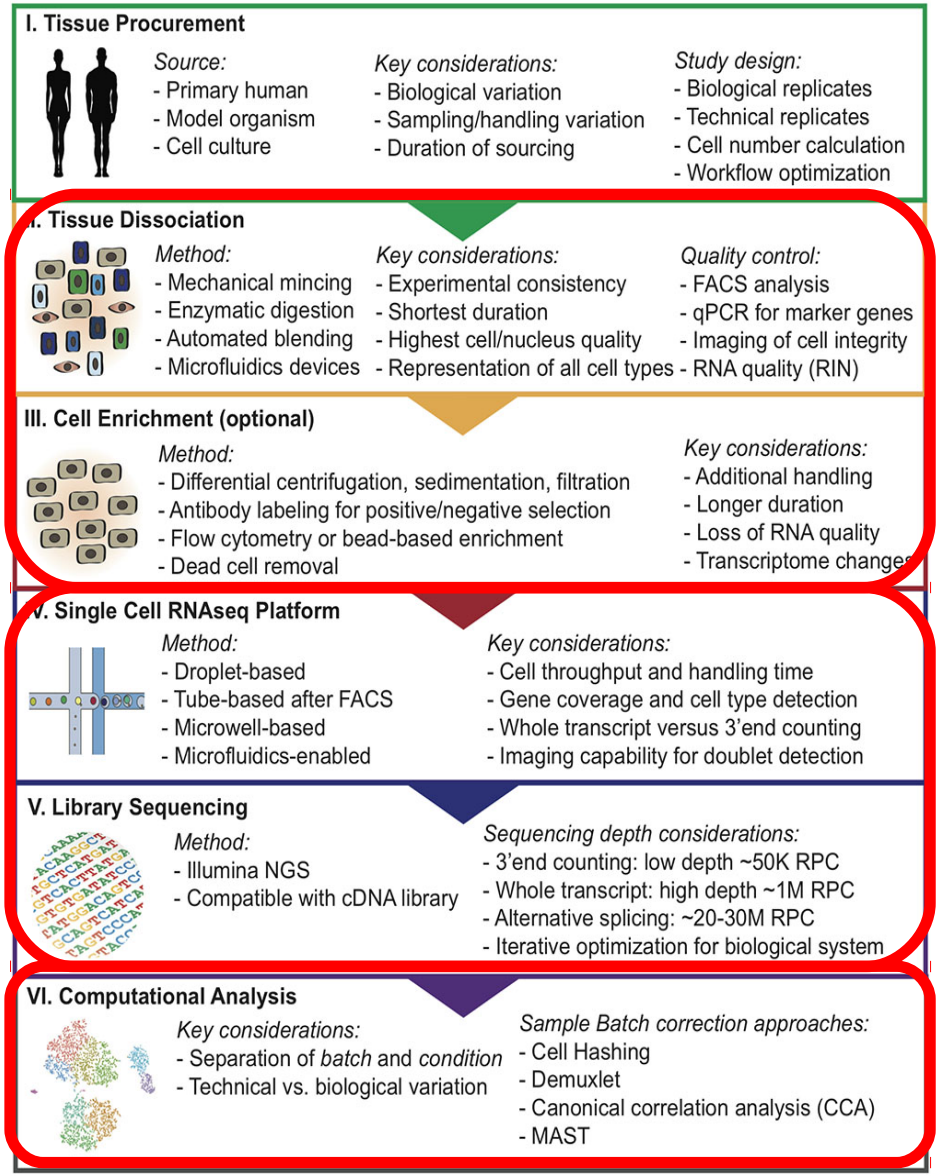
Source: 10XGenomics

Bulk RNA-seq vs Single-cell RNA-seq



Shalek and Benson (2017) Science Translational Medicine

Pipeline



Nguyen et al. (2019) Front. Cell Dev. Biol.

Definition:
Pipeline: a series of steps of an analysis.

From broad tissue to cell mix

From cell mix to nucleotide sequences

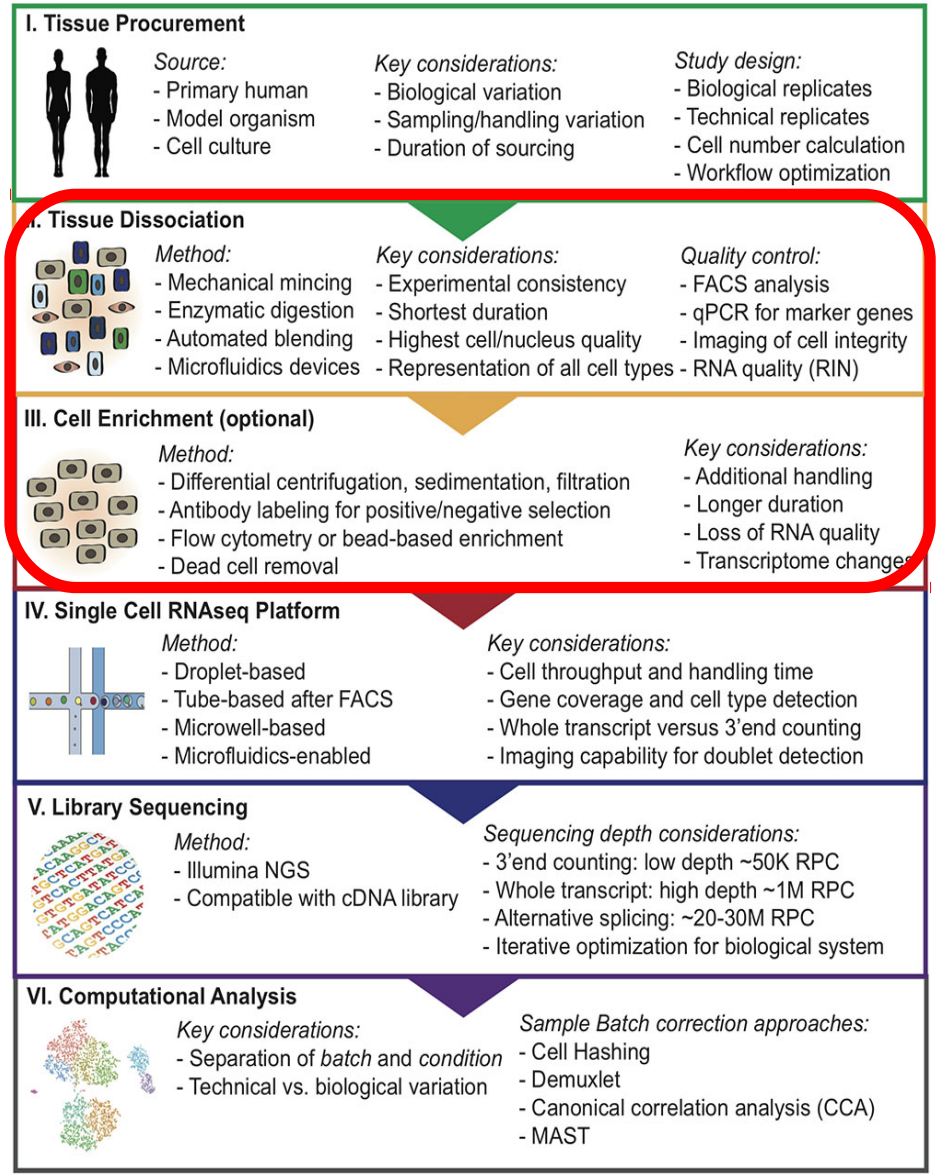
Biologic experimentation

From nucleotide sequences to results of interest

Bioinformatics analysis

Biologic experimentation

Biologic experimentation



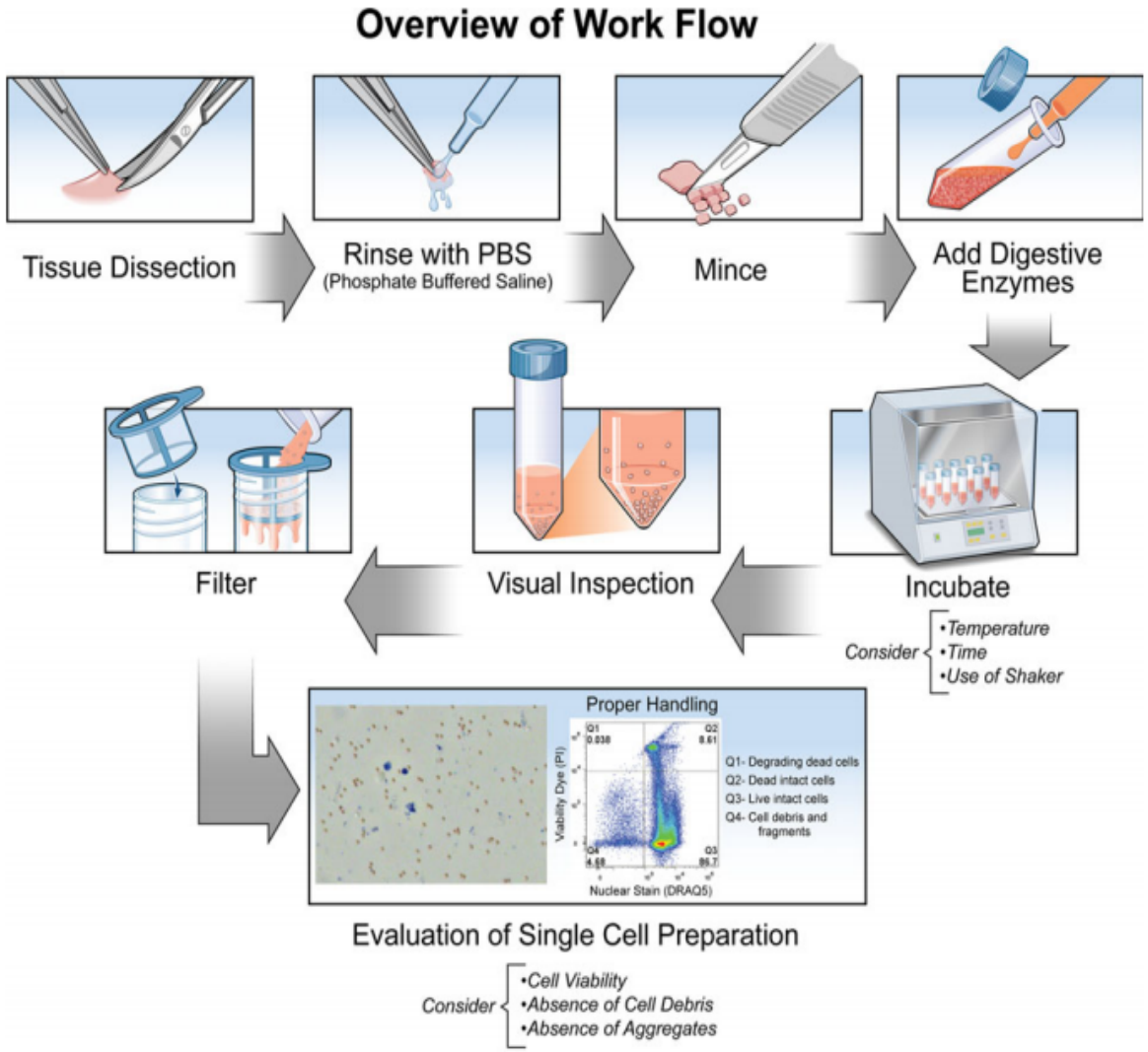
Nguyen et al. (2019) Front. Cell Dev. Biol.

From broad tissue to cell mix

Definition:
Pipeline: a series of steps of an analysis.

Tissue dissociation

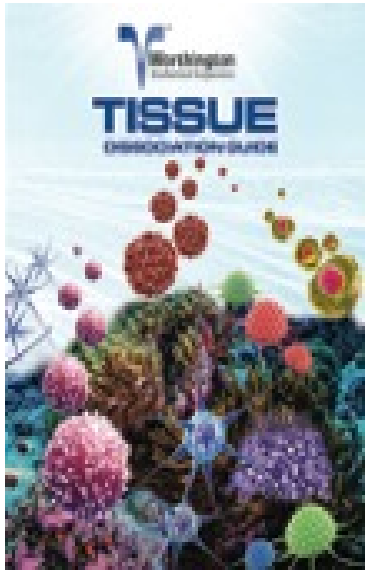
- Crucial starting points
- **Viability** >70% +++
- Minimize the cell preparation (faster is better)
- Remove debris, fibers and clump
- Cell preparation optimization can take a while



Reichard & Asosingh (2018) Cytometry

Tissue dissociation

1. Type of tissue
2. Species of origin
3. Age of the animal
4. Genetic modification(s) (knockouts, etc.)
5. Dissociation medium used
6. Enzyme(s) used
7. Impurities in any crude enzyme preparation used
8. Concentration(s) of enzyme(s) used
9. Temperature
10. Incubation times



Tissue Tables (references, grouped by tissue type and species)

Adipose/Fat	Adrenal	Bone	Brain
Cartilage	Colon	Endothelial	Epithelial
Eye	Heart	Intestine	Kidney
Liver	Lung	Lymph nodes	Mammary
Miscellaneous	Muscle	Neural	Pancreas
Parotid	Pituitary	Prostate	Reproductive
Scales	Skin	Spleen	Stem
Thymus	Thyroid/Parathyroid	Tonsil	Tumor

Table of Contents

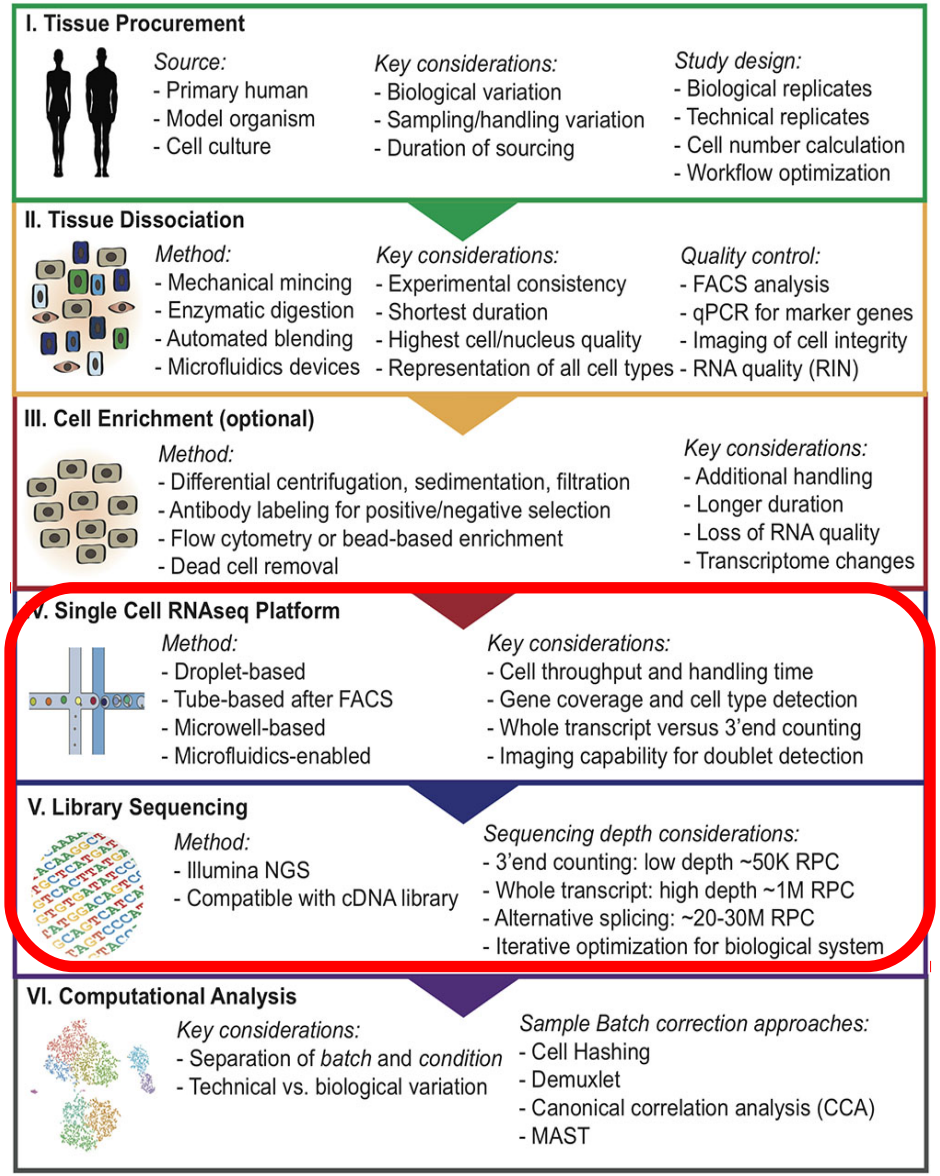
I. **Introduction**
II. Cell Isolation Theory
- Tissue Types
  - Epithelial Tissue
  - Connective Tissue
- Dissociating Enzymes
  - Collagenase
  - Trypsin
  - Elastase
  - Hyaluronidase
  - Papain
  - Chymotrypsin
  - Deoxyribonuclease I
  - Neutral Protease (Dispase)
  - Trypsin Inhibitor
  - Animal Origin Free (AOF) Enzymes
  - Celase® GMP

III. Cell Isolation Techniques
- Methods & Materials
  - Working With Enzymes
  - Basic Primary Cell Isolation
  - Equilibration with 95%O_2:5%CO_2
  - Trituration
  - Enzymatic Cell Harvesting
  - Cell Adhesion and Harvesting
  - Trypsin for Cell Harvesting
  - Cell Release Procedure
- Optimization Techniques
  - General Guidelines
  - Optimization Strategy
  - Cell Quantitation
  - Measure of Viability

IV. Use-Tested Cell Isolation Systems
V. Tissue Culture Glossary
VI. Stem Cell Glossary
VII. How To Cite Worthington Literature

http://www.worthington-biochem.com/tissuedissociation/default.html

Biologic experimentation

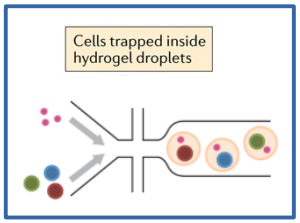


Nguyen et al. (2019) Front. Cell Dev. Biol.

Definition:
Pipeline: a series of steps of an analysis.

From cell mix to nucleotide sequences

Barcode and UMI in 10X



The 16bp 10x barcode is unique to each Gel Bead and tells you which cell the transcript is from.

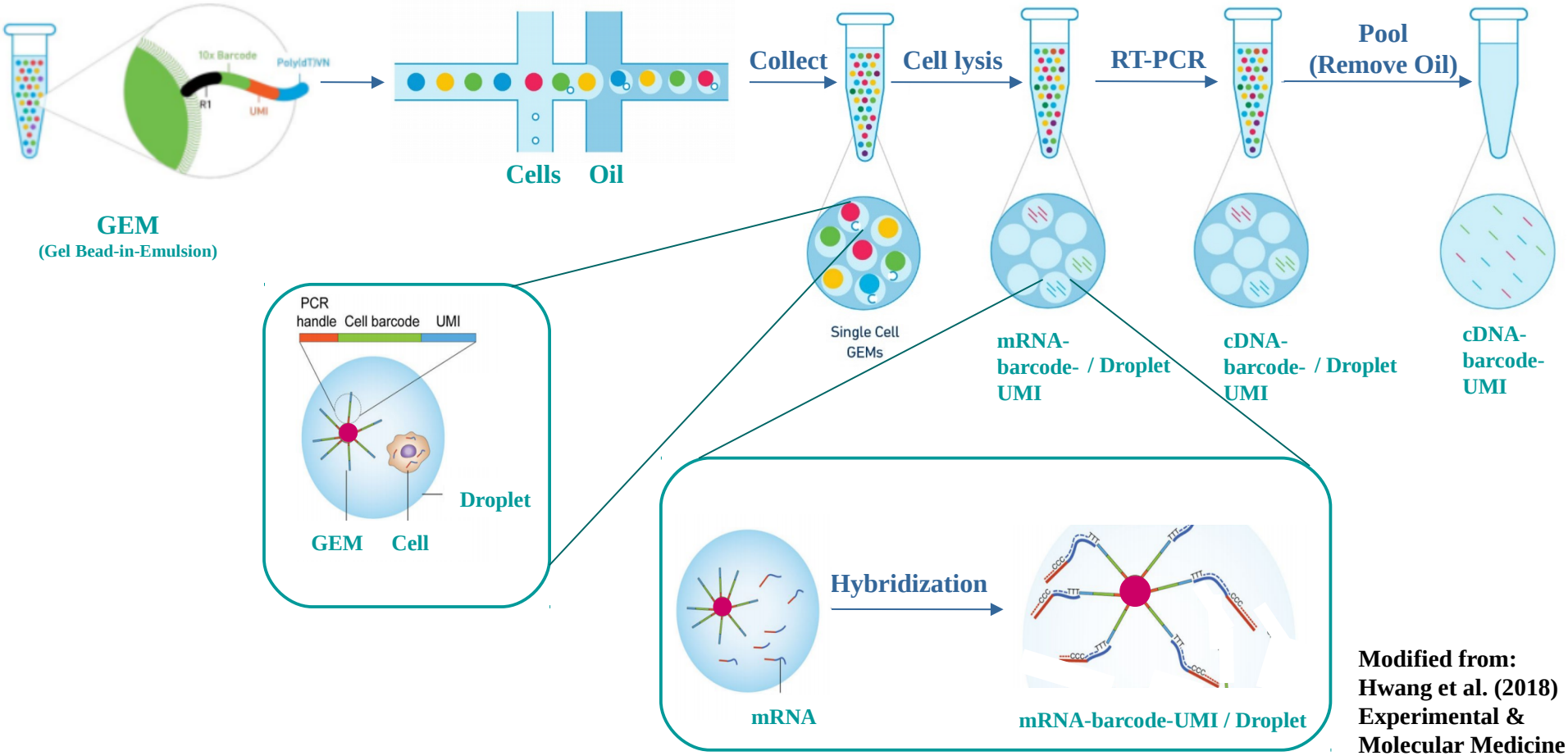
~ 737 000 barcodes

The 10bp UMI enables accurate quantitation of cell expression levels.

~ 1 million UMIs

The 8bp sample index allows to assign each barcoded read to its sample of origin.

4 sample index



Modified from:
Hwang et al. (2018)
Experimental & Molecular Medicine

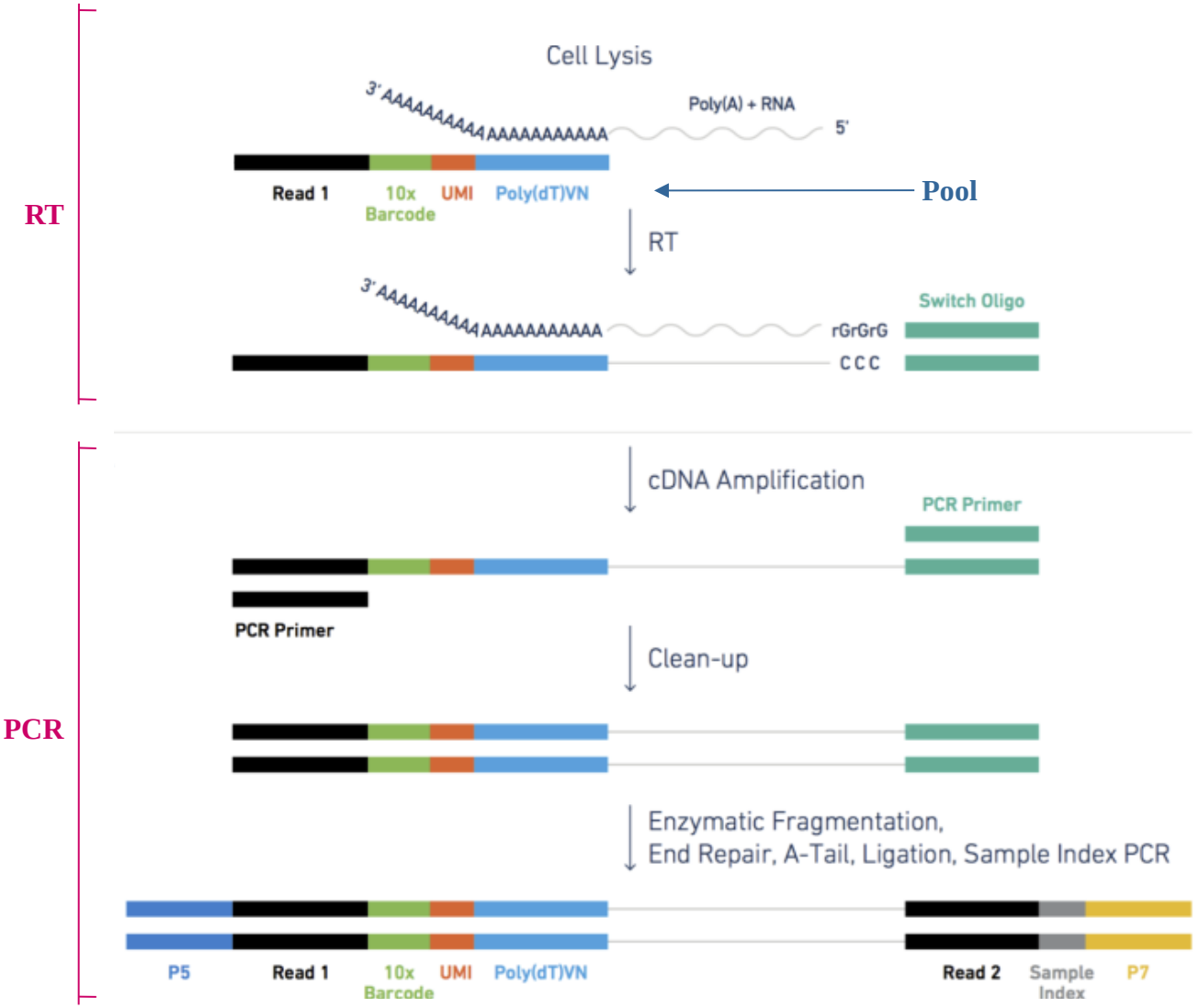
Sample index, Barcode and UMI

- Barcode (12-16nt): cell identification
- Limit:
  some reads are copies of PCR from the same transcript
- Solution:
  UMI (Unique Molecular Identifiers) (8-10nt): molecular identification

  ↓

  Improved accuracy of gene expression measures

- Sample index (8nt): sample identification

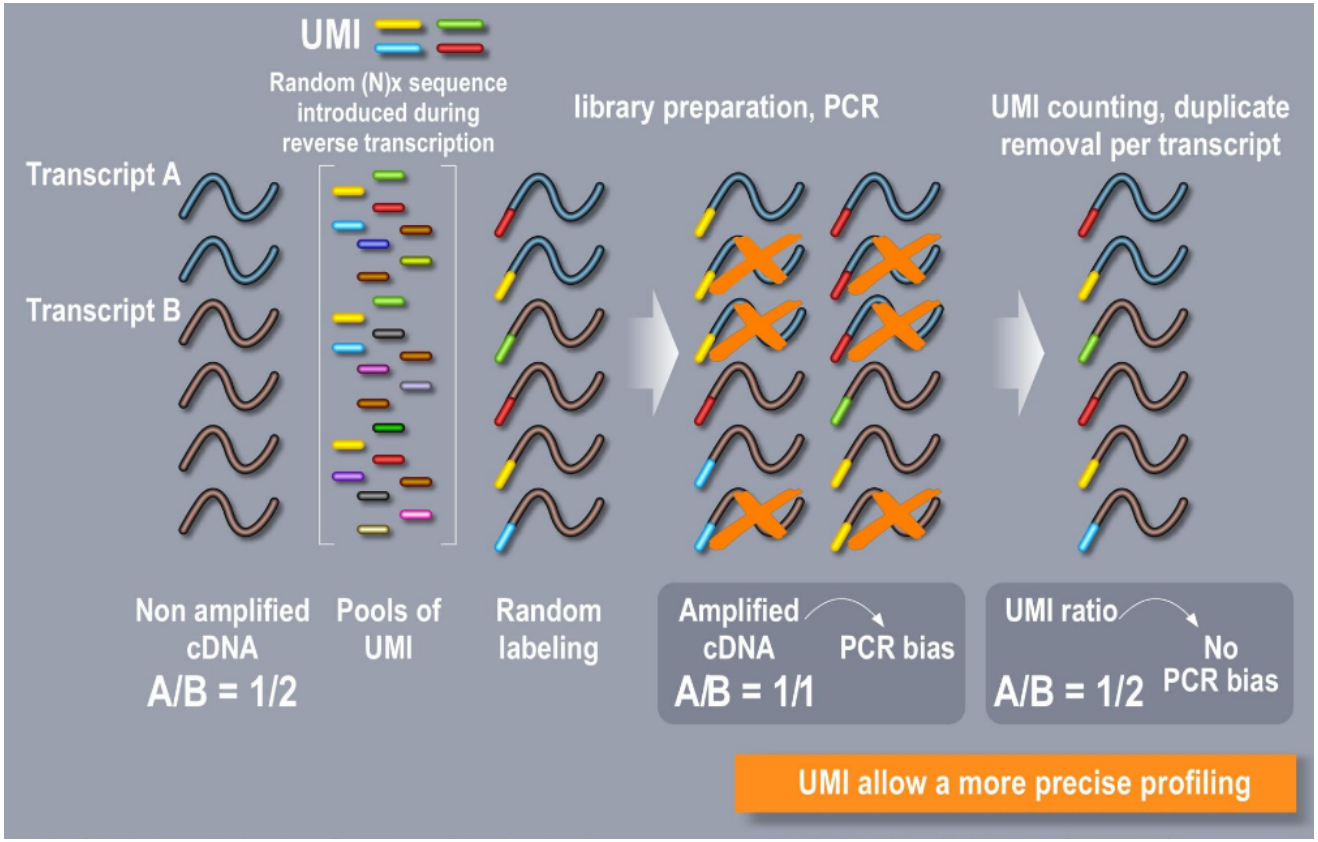


Sample index, Barcode and UMI

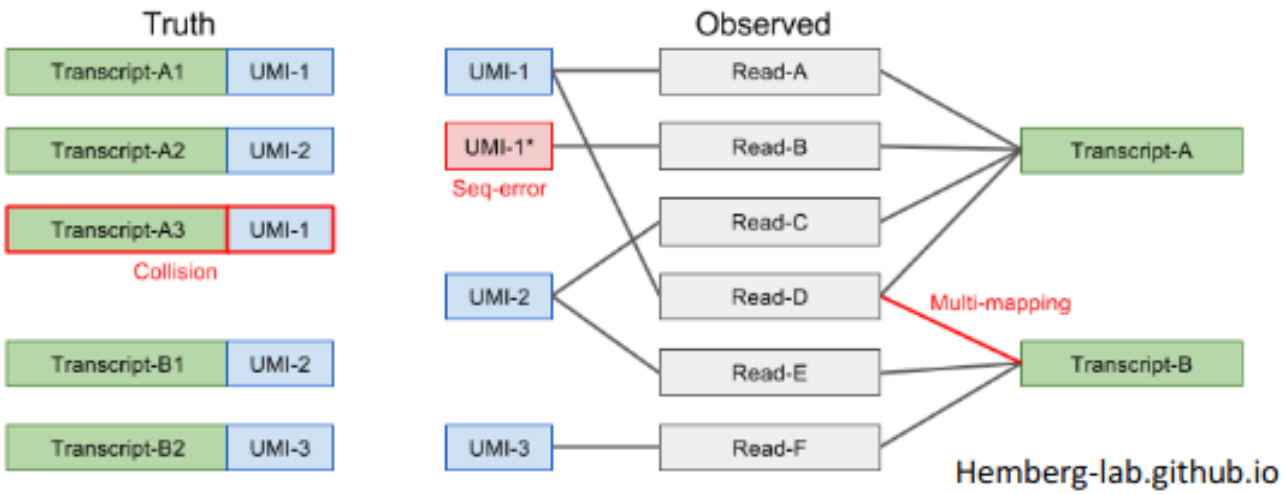
- Barcode (12-16nt): cell identification
- Limit:
  some reads are copies of PCR from the same transcript
- Solution:
  UMI (Unique Molecular Identifiers) (8-10nt): molecular identification

↓

Improved accuracy of gene expression measures



Sample index, Barcode and UMI



Limits:

- Same UMI does not necessarily mean same molecule
  - Biases in UMI frequency and short UMIs
- Difference UMI does not necessarily mean different molecule
  - Sequencing errors (~ 7-10% of UMI have at least 1 error)
- Different transcript does not necessarily mean different molecule
  - Mapping errors / multi-mapping

→ System of correction when 1 mismatch occurs

Single-cell technologies

3’ & 5’ (90 bases)

Pros:
- a lot of cells (identification of rare cell types, ...),
- cost

Cons:
- lows transcript coverage
- detecting certain lowly expressed genes/transcripts is difficult (**~15% ARN are captured by 10X**)

→ **Cells level**

Long reads (full-length)

Pros:
- full length information (study of alternative splices, SNV analysis, …)
- detect more mARN diversity

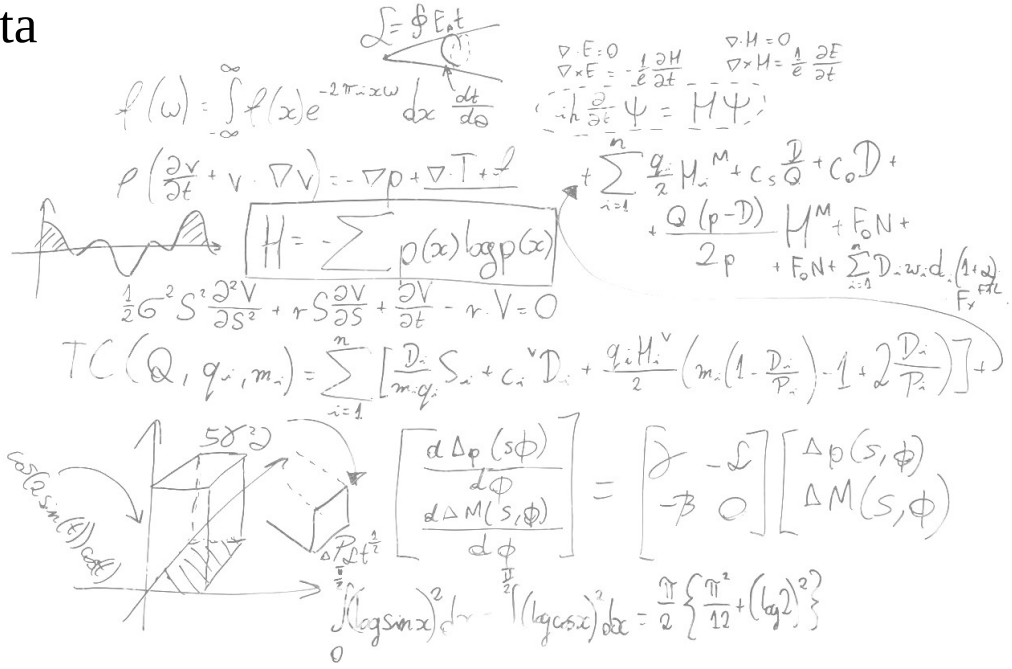
Cons:
- error rate of sequencing
- low througthput (cost)

→ **RNA level**

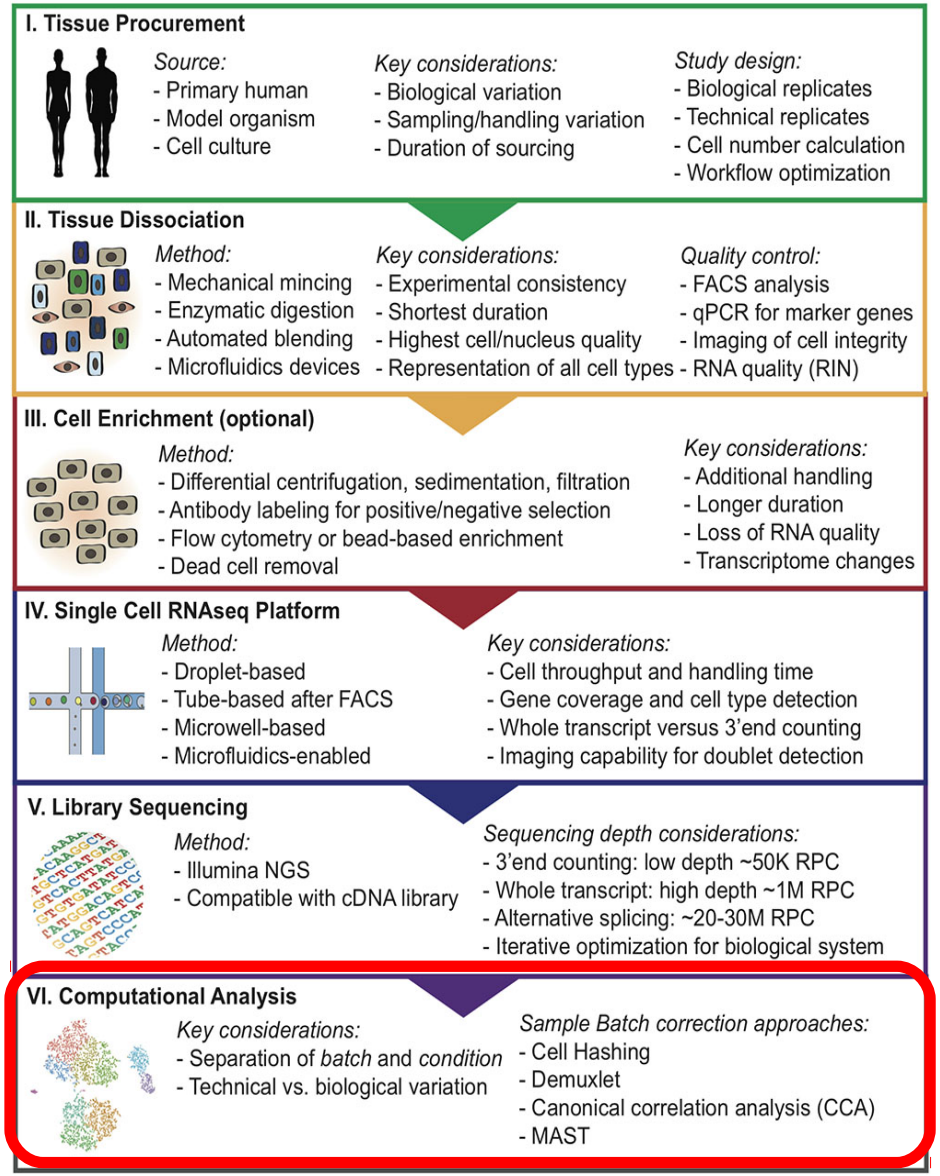
Bioinformatics analysis

Computational challenges

- Lower coverage/depth than bulk RNA-seq
- Technical & biological noise
- High dimensionality
- High variability
- Dropouts => Zero-inflated data
- Multimodality
- …



Bioinformatics analysis

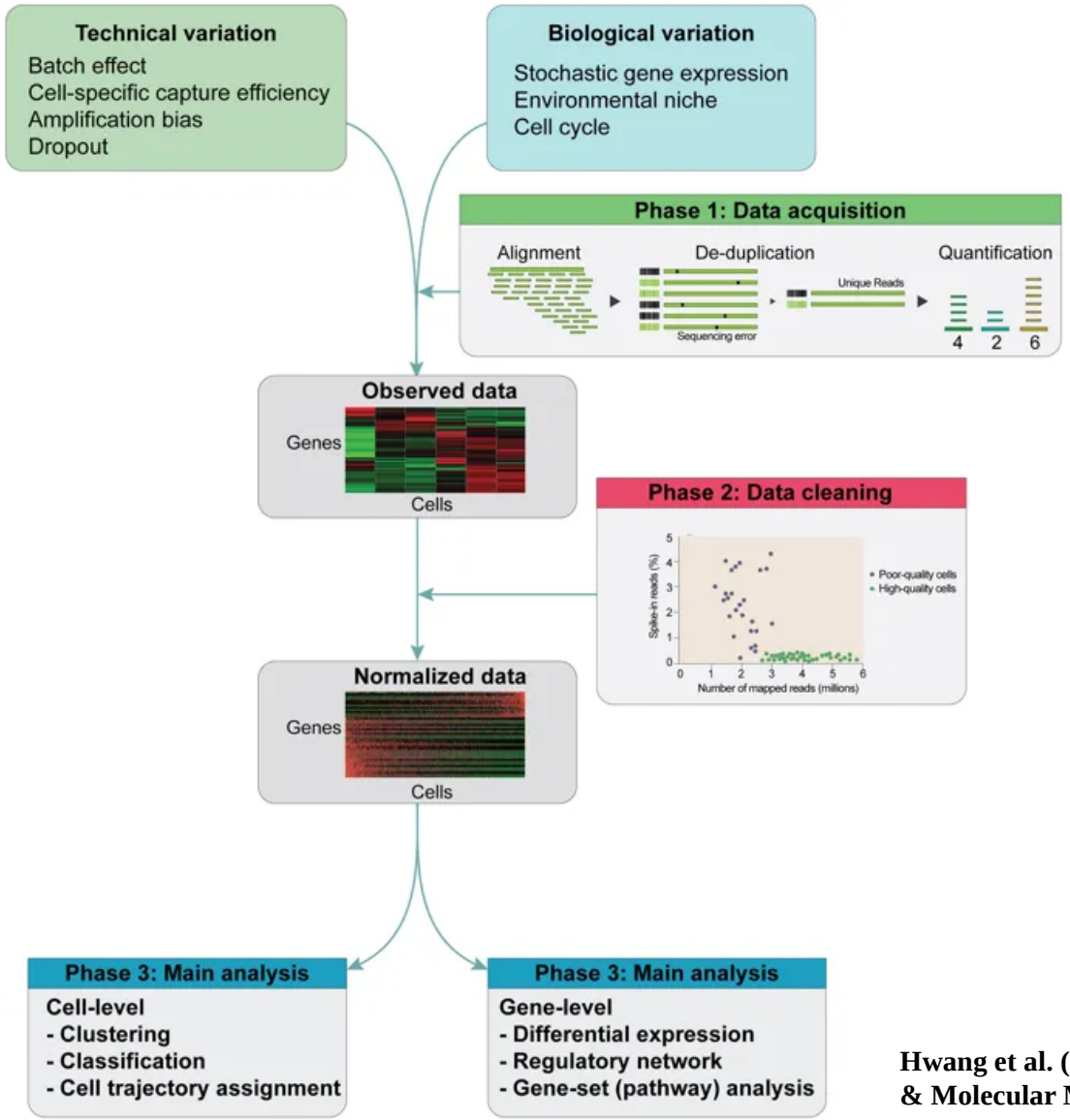


Nguyen et al. (2019) Front. Cell Dev. Biol.

Definition:
Pipeline: a series of steps of an analysis.

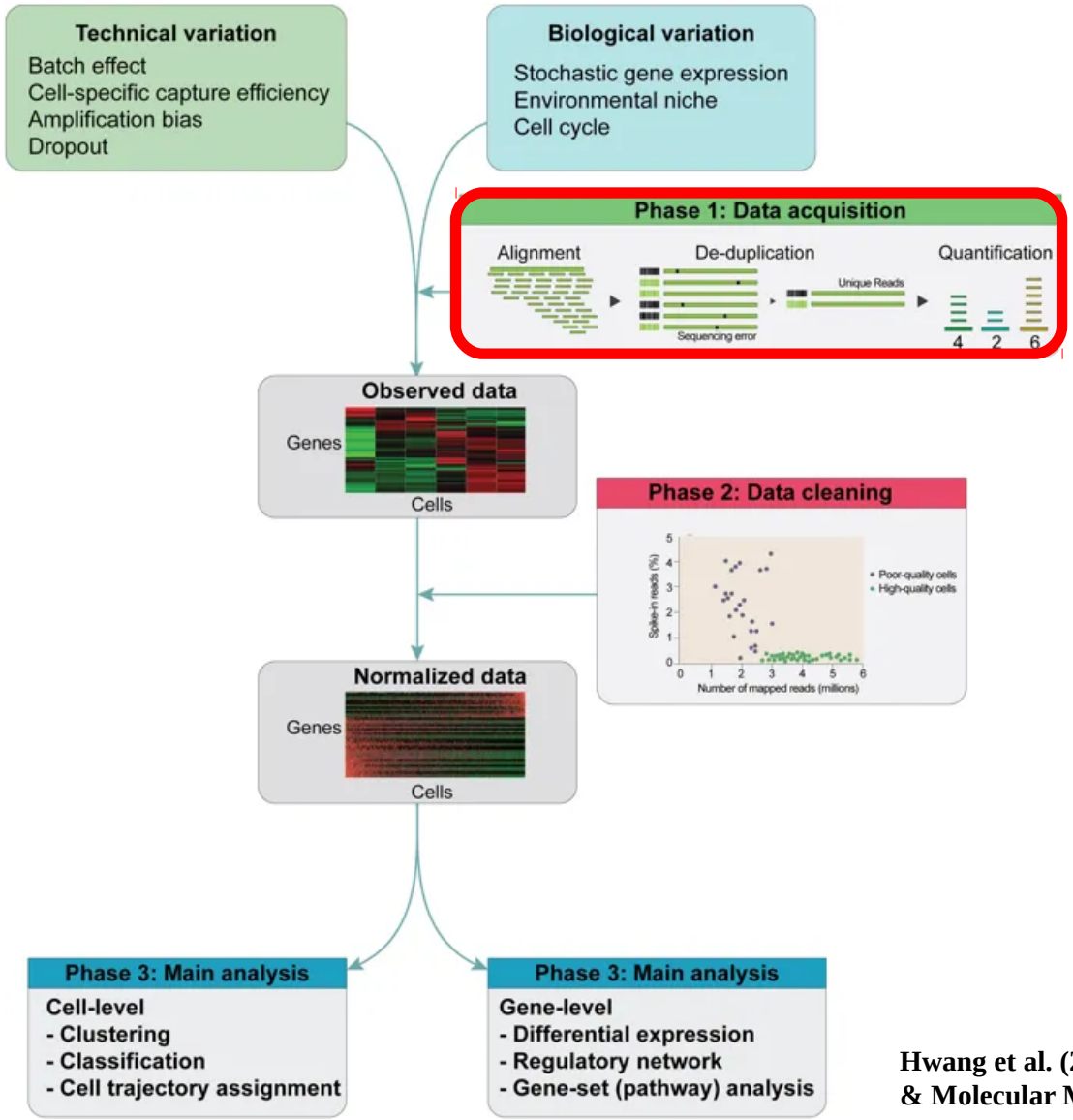
From nucleotide sequences to results of interest

Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

Computational pipeline



From sequencer to counts matrix

Hwang et al. (2018) Experimental & Molecular Medicine

Demultiplexing

cDNA-barcode-UMI —sequencing→ **File.bcl** —converter→ **File.fastq**

Illumina sequencer's **b**ase **c**all **f**iles

R1 : Barcodes + UMIs

```
@A00461:74:HF2K2DRXX:1:2101:1551:1016 1:N:0:CCACTTAT
TNGTGCTTCGCTACAAAATTCCTCGGAG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:1696:1016 1:N:0:CCACTTAT
TNCAATCAGCCACCGTTCCGGGCATTTG
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:2817:1016 1:N:0:CCACTTAT
ANCCCTGGTTGAAGTATAGTTTTAGTTA
```

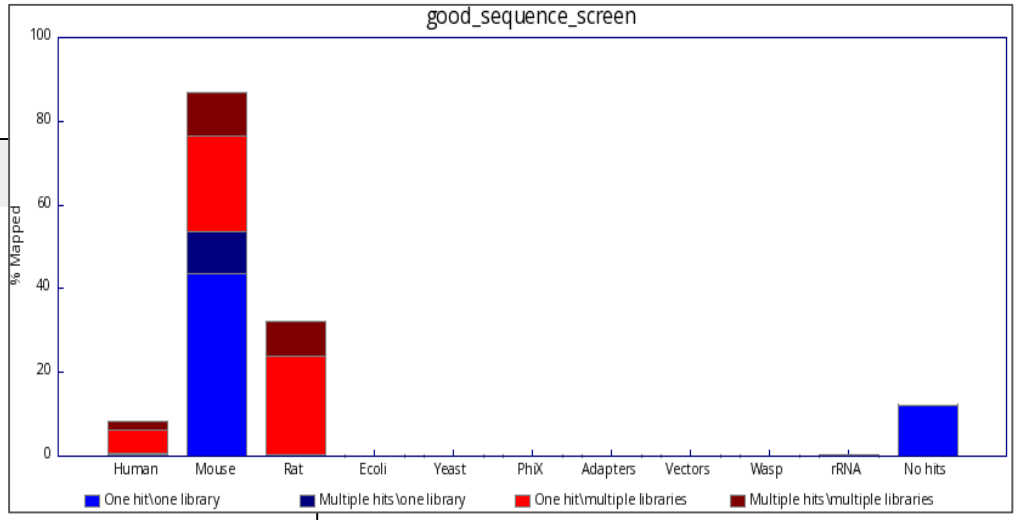
I1 : Sample Index

```
@A00461:74:HF2K2DRXX:1:2101:1551:1016 1:N:0:CCACTTAT
CCACTTAT
+
FFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:1696:1016 1:N:0:CCACTTAT
CCACTTAT
+
FFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:2817:1016 1:N:0:CCACTTAT
CCACTTAT
```

R2 : Real sequences

```
@A00461:74:HF2K2DRXX:1:2101:1551:1016 2:N:0:CCACTTAT
NAAATGTTTGCTACCAGGAAATATCAGATGGGAAGAAAAGGAATTAGACCTTGTTCTTATCTTGCTTGGGAGCATAGAACTGCCCCTGCTG
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:1696:1016 2:N:0:CCACTTAT
NGTCTGGATTTAGACTCTTCCACACCTAAGAACTGGTATCTCCTGACTCTTTCCAGGCTGTGTCCCTTTGATCACATTATTTCTCTTTGTC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00461:74:HF2K2DRXX:1:2101:2817:1016 2:N:0:CCACTTAT
NTGGAATTCATGGACGACACGAGCCGATCCATCATCCGCAATGTAAAAGGCCCCGTGCGCGAGGGCGACGTGCTCACCCTTTTGGAGTCAG
```

Reads Quality-control



FastQC Report

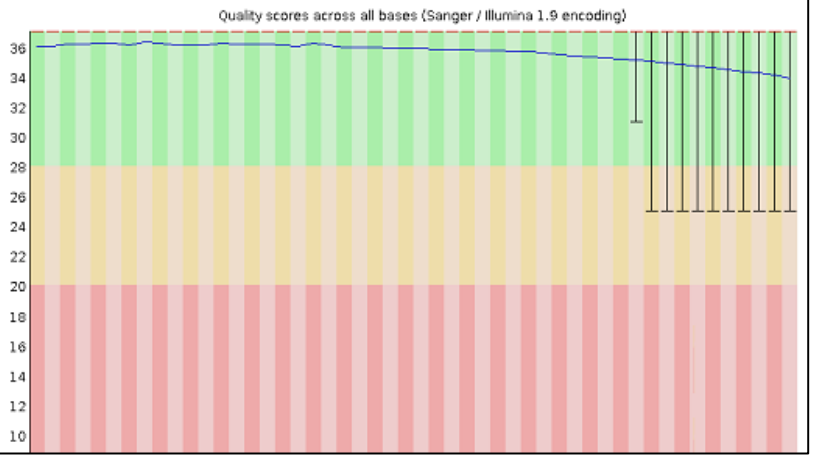
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

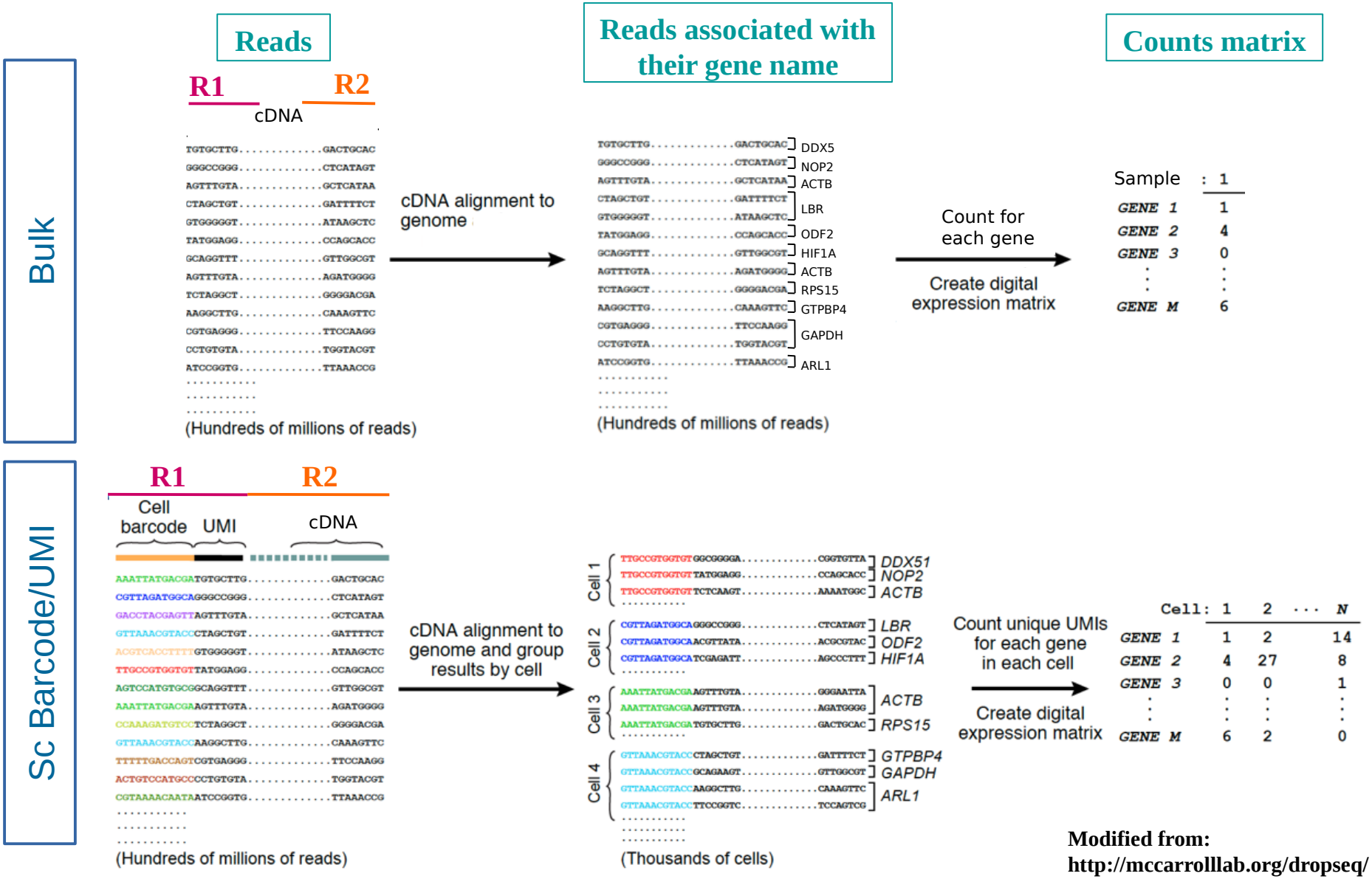
Measure	Value
Filename	BC_392_1_S29_R2_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	109443265
Sequences flagged as poor quality	0
Sequence length	91
%GC	43

Per base sequence quality



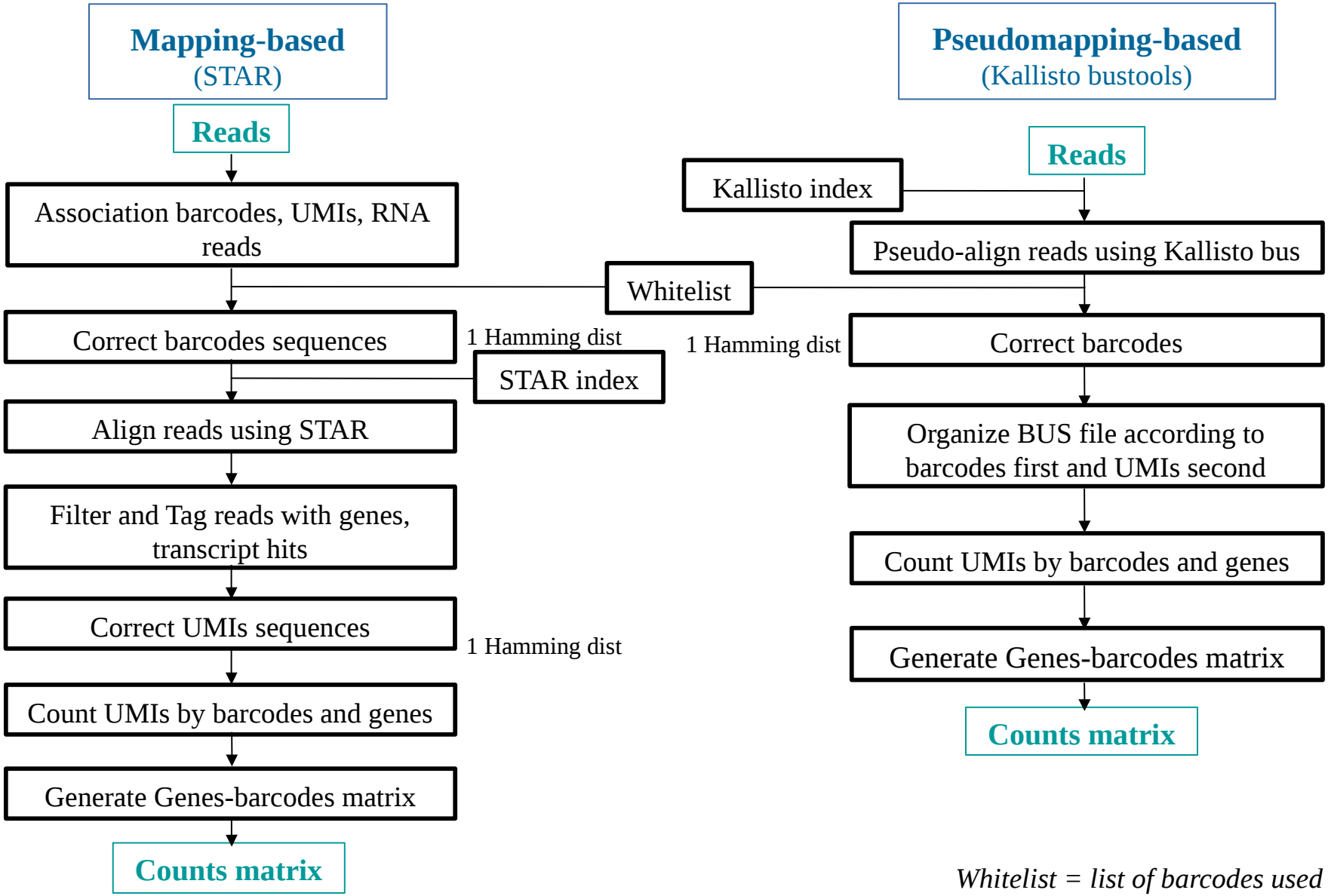
- As usual : FASTQC, FastqScreen, …
- If QC is bad at the end
  => trimming
  => re-check quality-control
- For sc Barcode/UMI: Trimming just For R2
  (R1= Cell barcode + UMI)

Reads processing pipeline



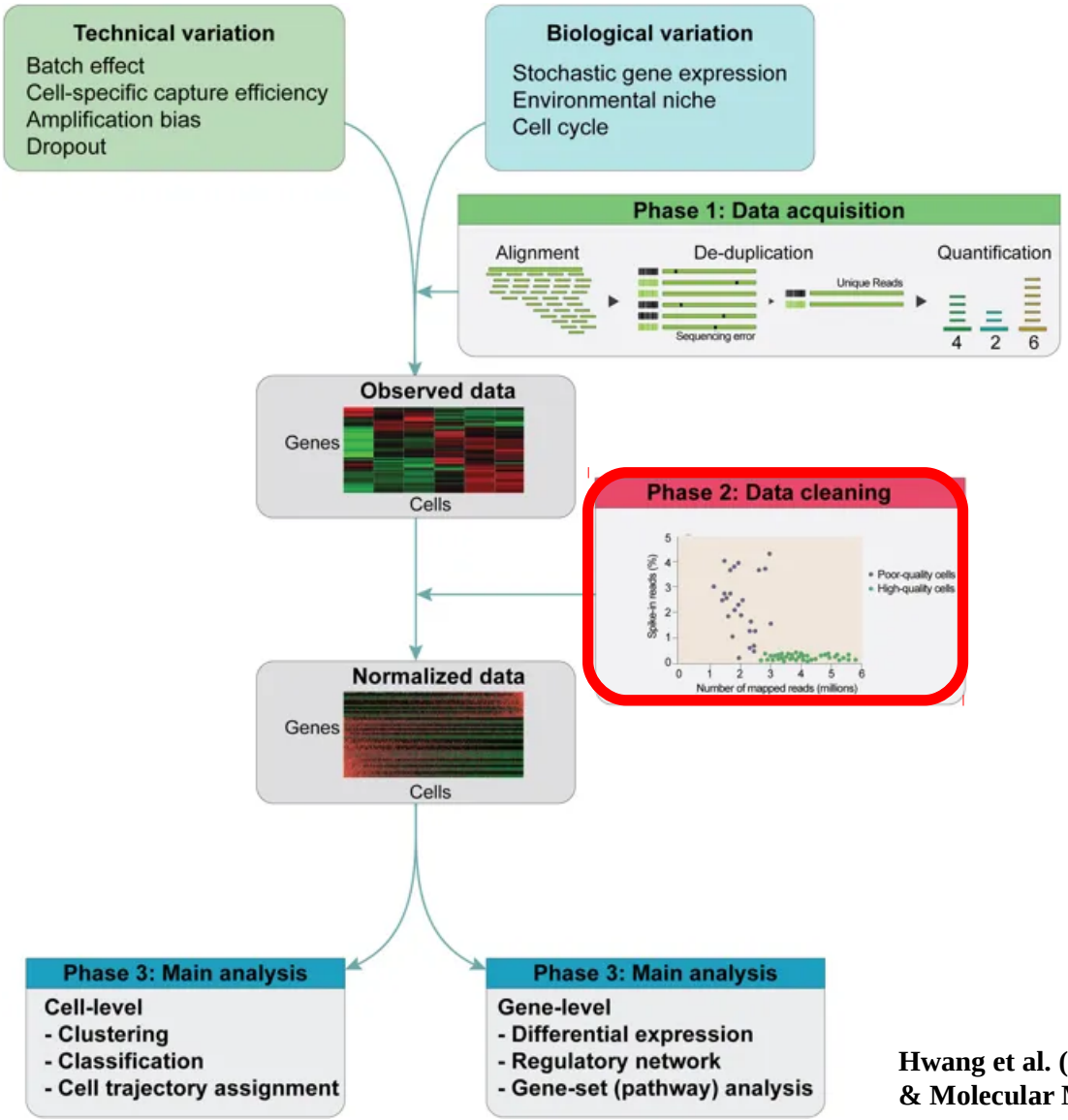
Modified from:
http://mccarrolllab.org/dropseq/

Reads processing pipeline



Whitelist = list of barcodes used

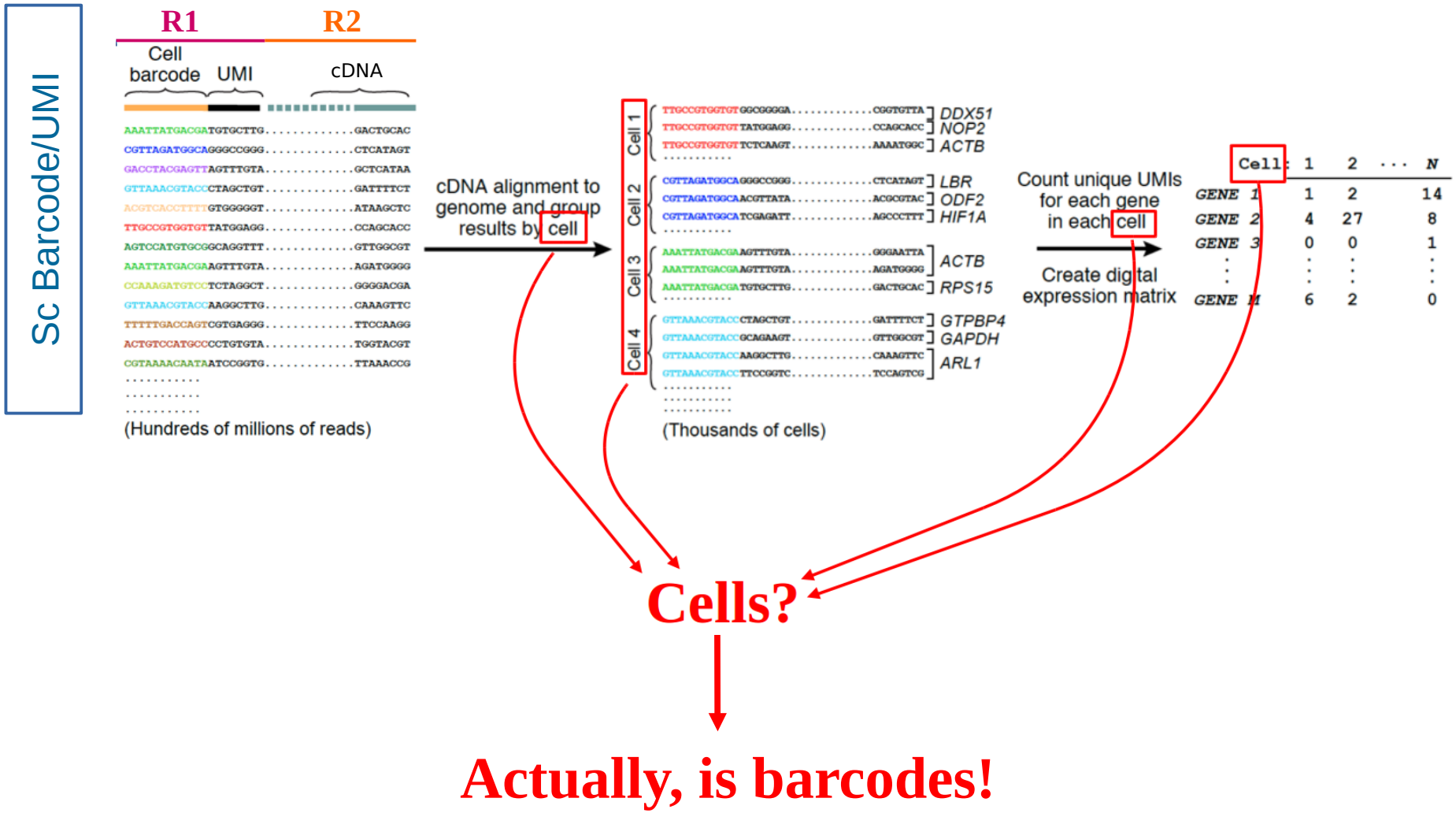
Computational pipeline



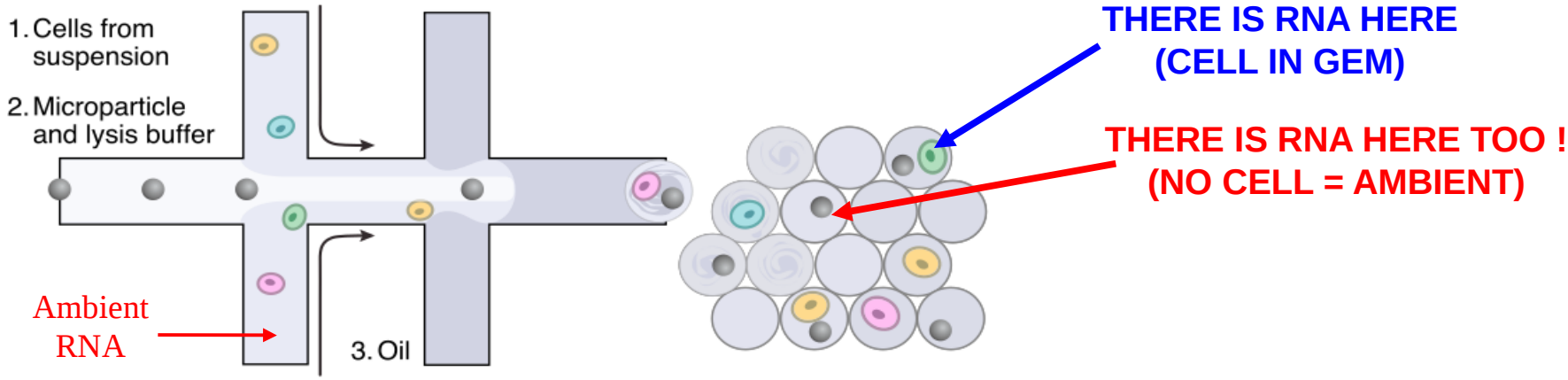
From raw counts matrix to normalized data

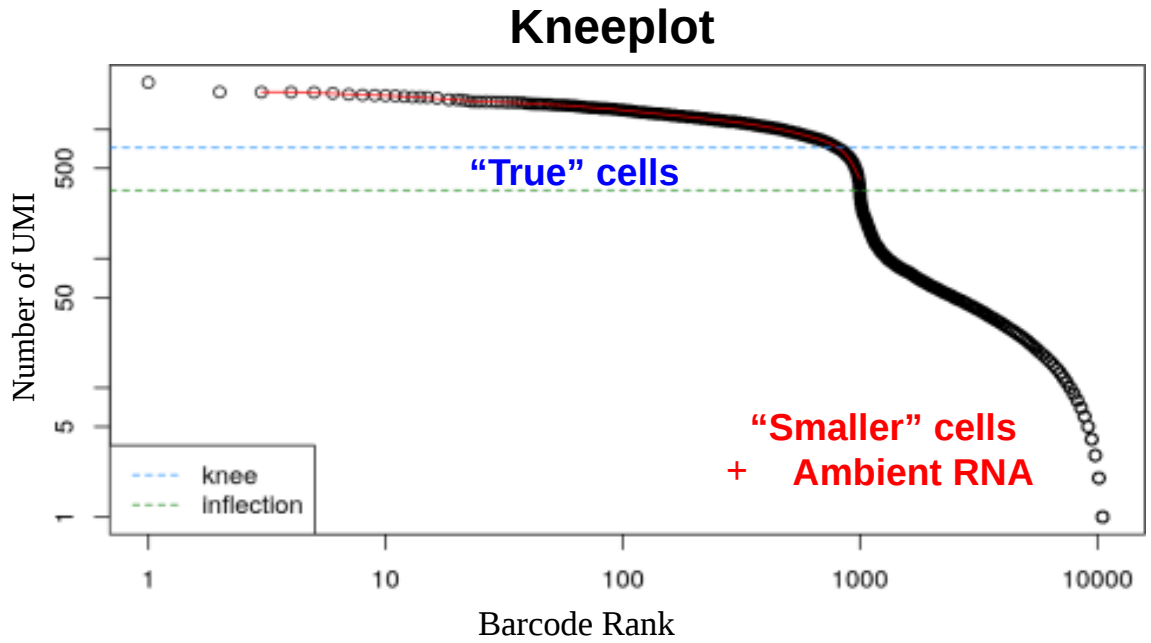
Hwang et al. (2018) Experimental & Molecular Medicine

Reads processing pipeline



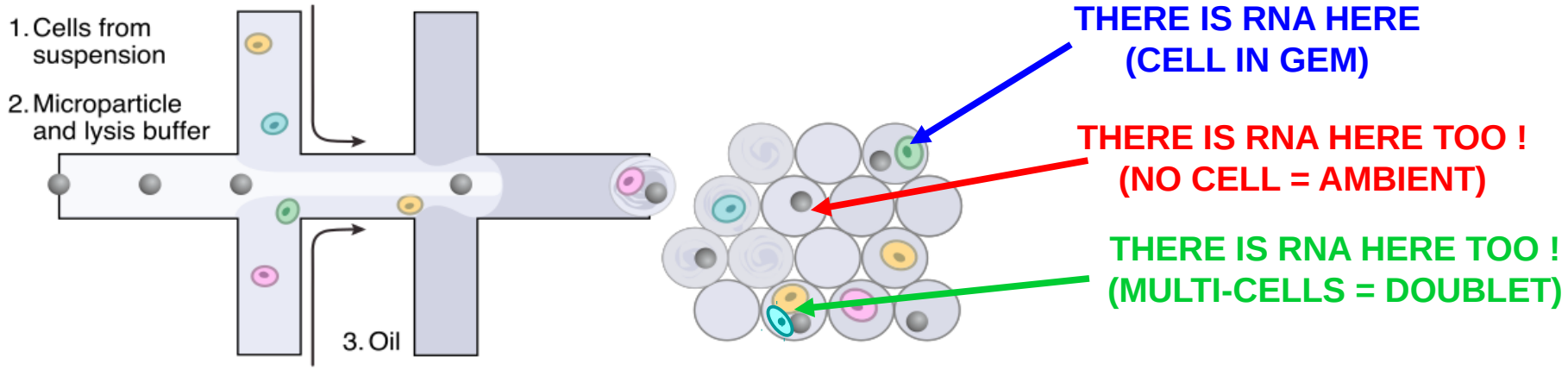
Filtering droplets: empty droplets

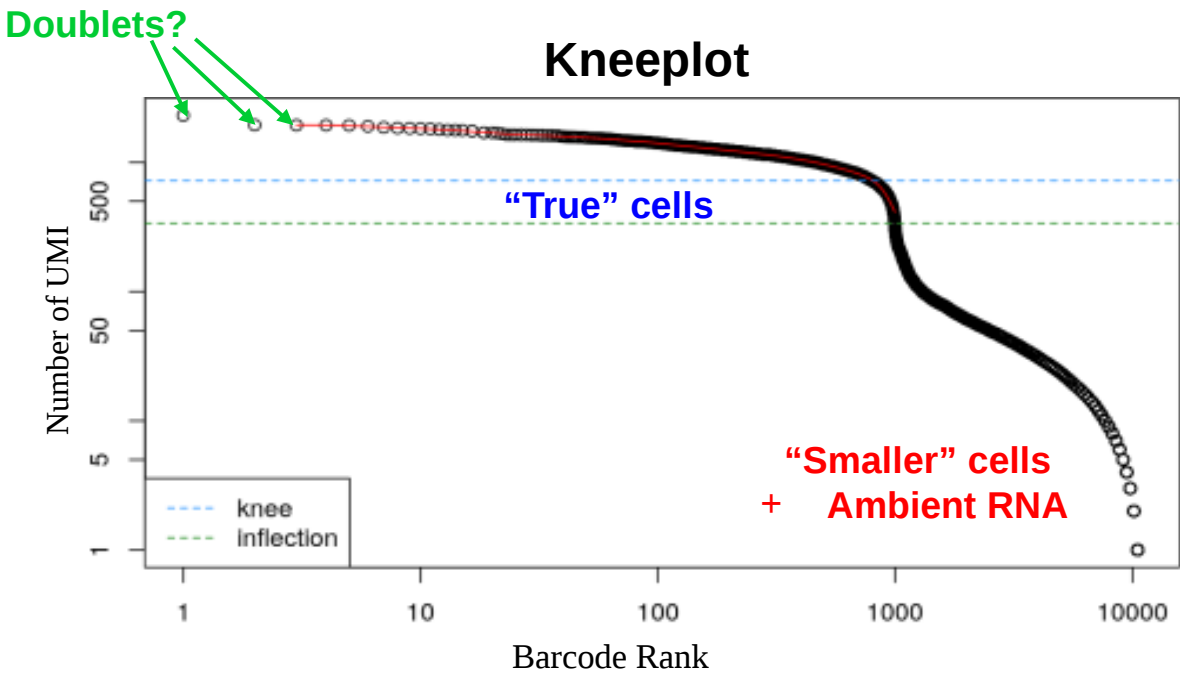




- Viability = 70%
  → 30% dead cells
  → ambient RNA
  → noise in empty droplet
  + noise in droplet with cell
- Package R: EmptyDrops

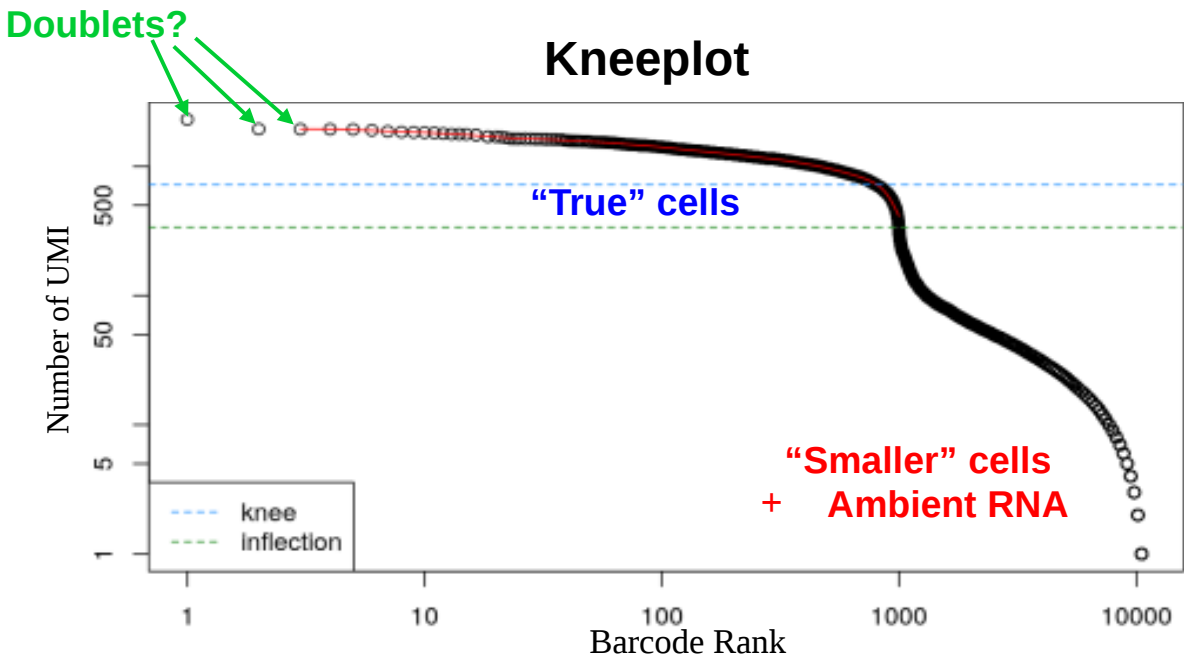
Filtering droplets: doublets or multiplets

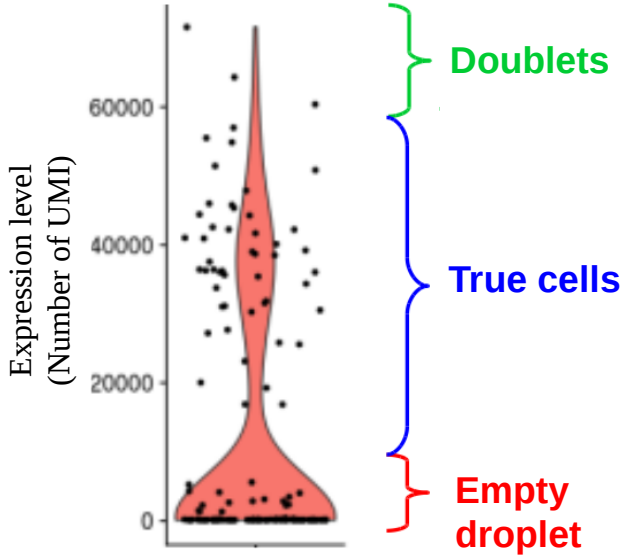


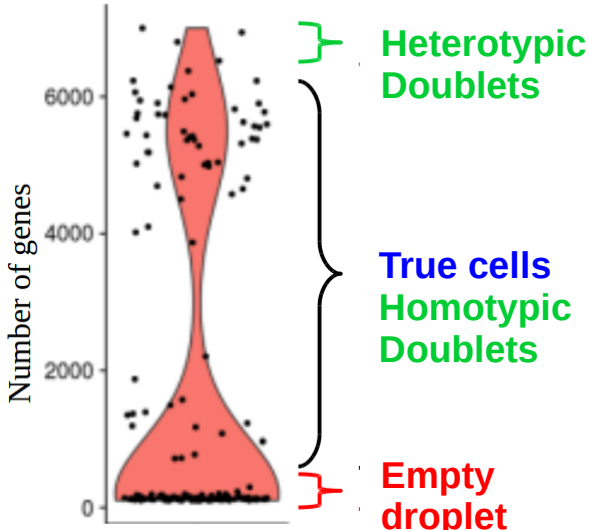


- Doublet types:
  - Homotypic: same cell type in droplet
  - Heterotypic: different cell types in droplet
- Doublet rate:
  - 1% pour 1000 cells
  - 5% for 10 000 cells

Filtering droplets: doublets or multiplets



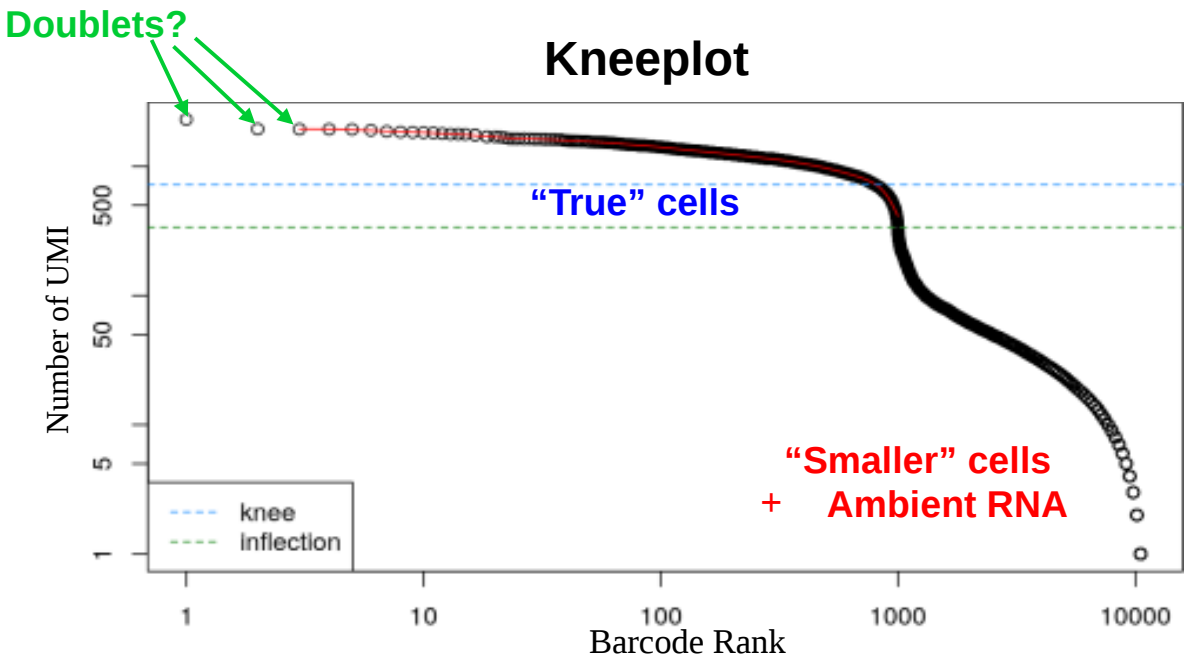


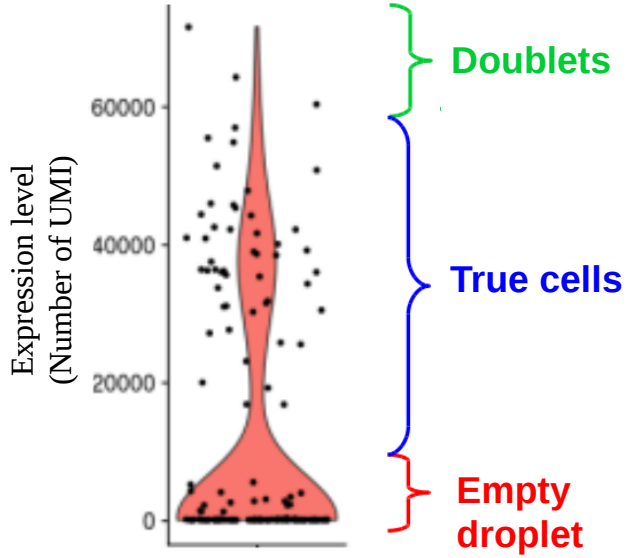


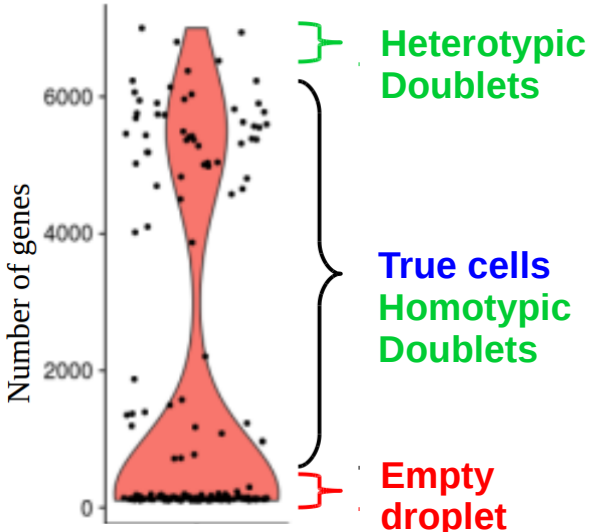
What is a expressed gene?

- Minimum number of UMI?
  (Be careful: droplet technology = low depth!)
- Minimum number of cell expressing it?
- Minimum mean gene expression?

Filtering droplets: doublets or multiplets





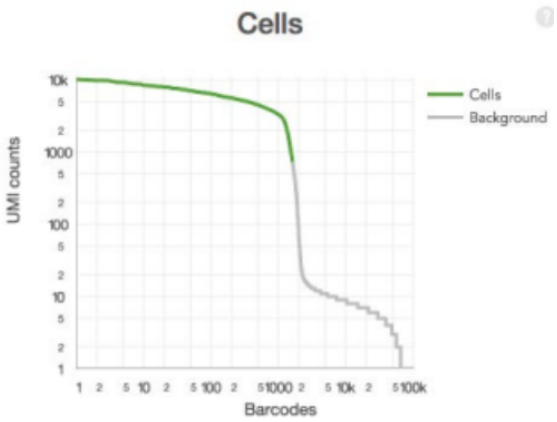


How do we choose the thresholds?

- Genes expressed > 5 cells
- 200 genes < kept cells
- Expected number of cells
- According to the biological knowledge of the sample
- R packages: EmptyDrops, DropletUtils, scDblFinder, Scran, scds, ...

Kneeplot: Diagnosis

Typical Sample Profile

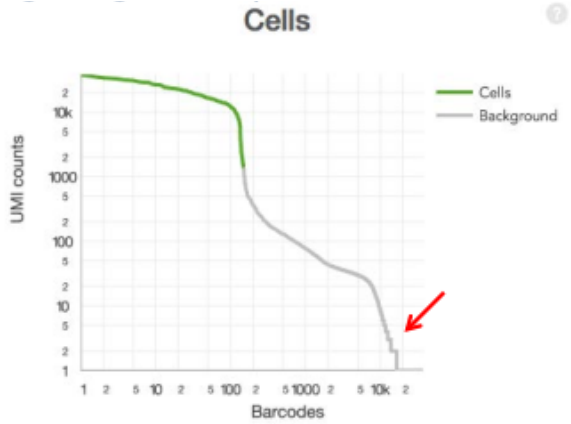


Defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	> 1,000
UMIs	> 10,000

Good!

Low Barcode Counts



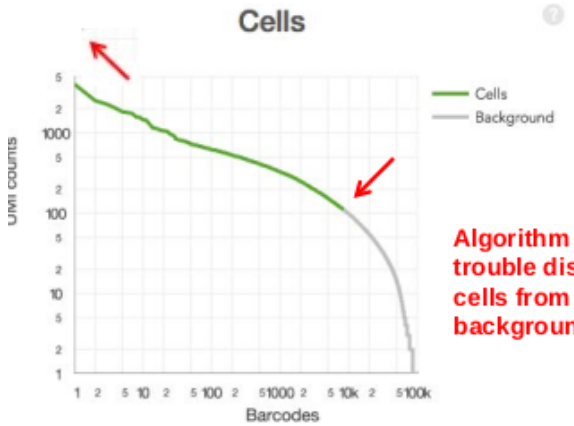
Low number of barcodes detected

Metric	Value
Barcodes	~ 15,000
Cell Barcodes	> 100
UMIs	> 10,000

Bad!

Too low a sequencing depth, so no sequencing of the ambient RNA, but also no sequencing of part of the signal.

Loss of Single Cell Behaviour



Algorithm has trouble discerning cells from the background

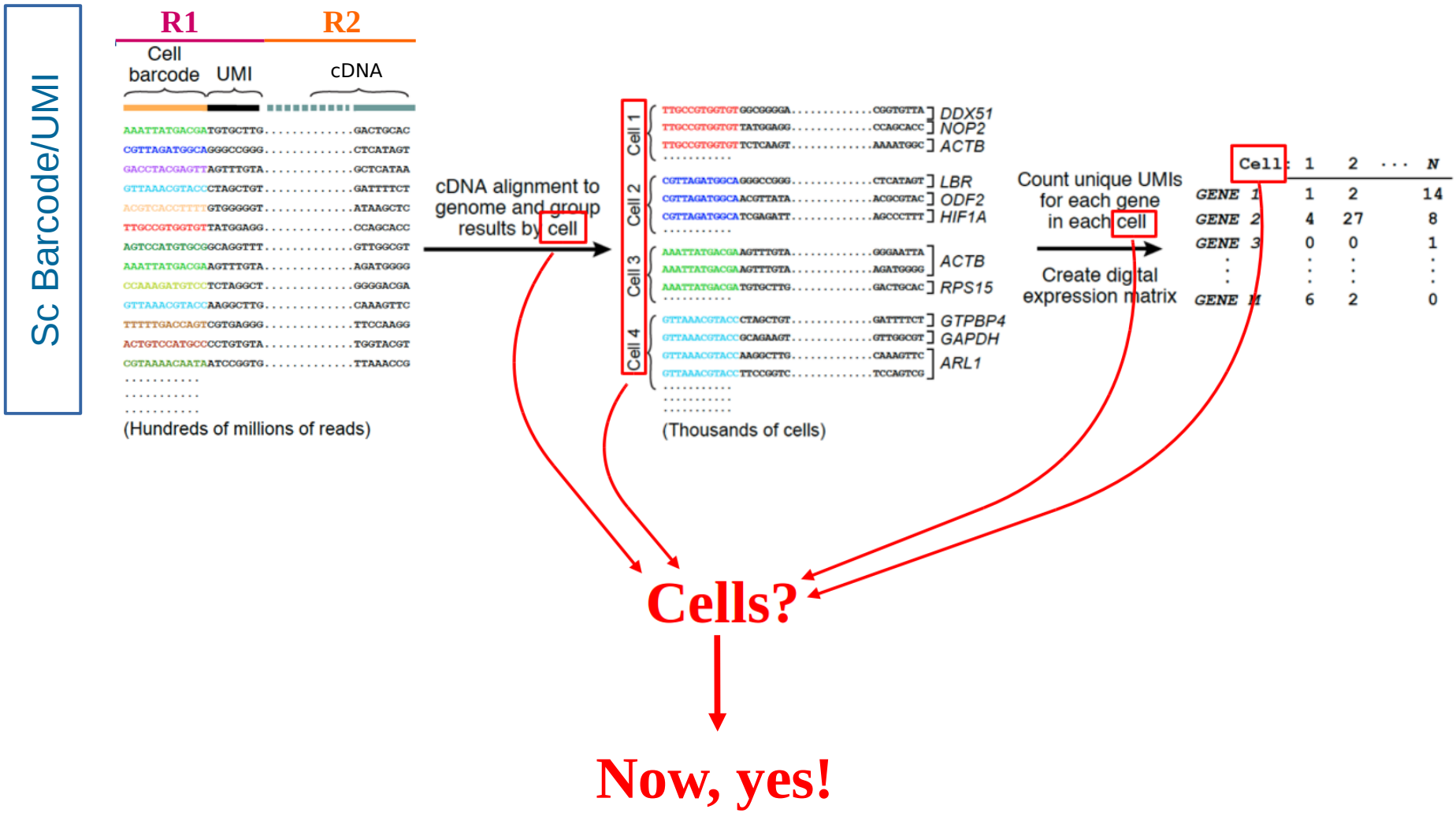
Lack of defined cliff and knee

Metric	Value
Barcodes	> 90,000
Cell Barcodes	~ 10,000
UMIs	> 10,000

Bad!

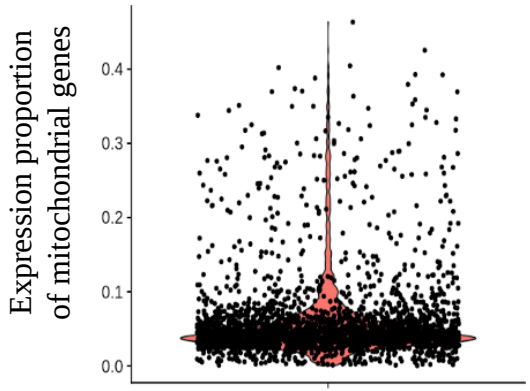
Problem in cell lysis, their RNA has not been released into the droplet, so there is almost only background noise so little UMI.
(Corresponds to the bottom knee on the first.)

Reads processing pipeline



Filtering cells

Mitochondrial genes expression

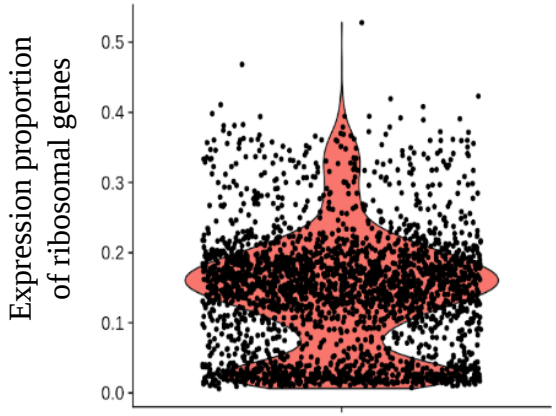


High percentage of mitochondrial gene expression may be due to apoptotic cells:

Kept cells < 5-20% mtRNAs

Genes names beginning by “MT-”.

Ribosomal protein genes expression



Linked to: cellular activity? cell cycle? Not very clear! Community debate, hard to say if it matters or not.

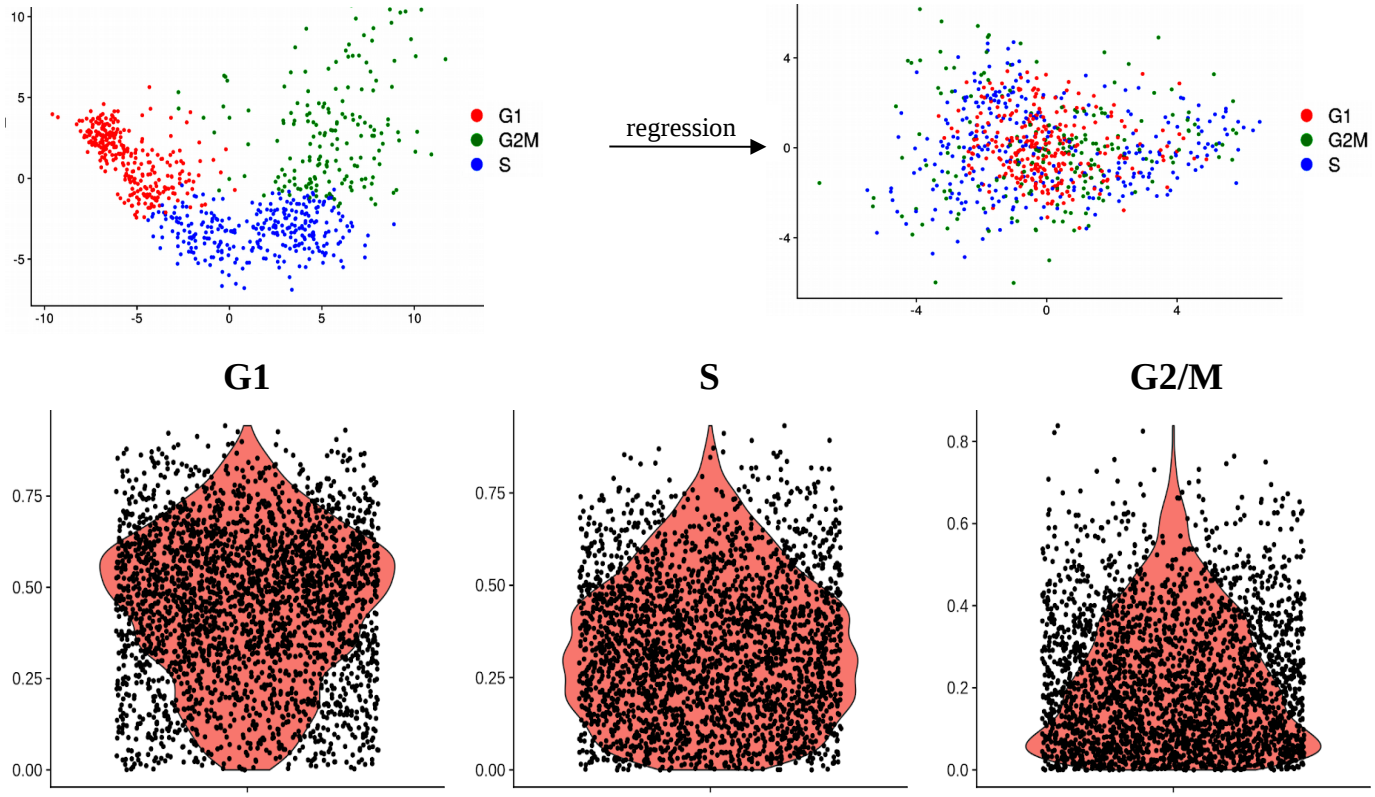
Kept cells < 25% rbRNAs ?
10% rbRNAs < Kept cells ?

Genes names beginning by “RP-”.

This thresholds are subjectives!
Need to be adjusted according to the biological knowledge of the sample!

R packages: Scater, scRNAseq

Filtering cells



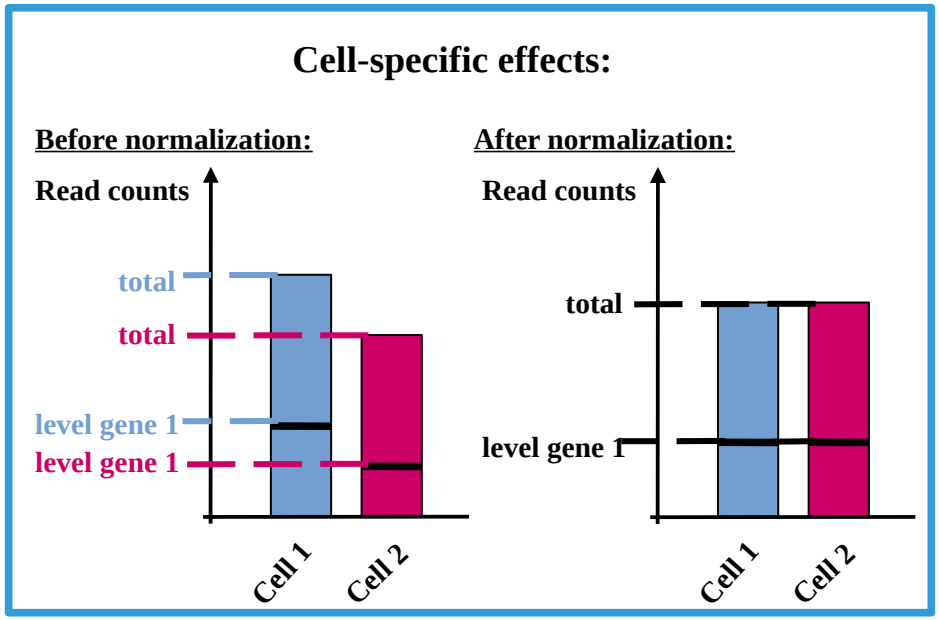
Expression varies according to the cell cycle, so the same cell type will have a variation in the expression of these genes.

1) Identification of the cell cycle stage of cells.
2) Normalization.

R packages: Scran (function cyclone), seurat

Normalization: different levels

- Process of identifying and removing systematic variation NOT due to real biological differences.
- 2 levels:
  - Sample/Technology-specific effects : batch effect → data integration.
  - Cell-specific effects: sequencing depth (library size) → make the distributions comparable.



	Cell-specific effects	Gene-specific	Corrections
Sequencing depth	✓		Depth Normalizatio
Amplification	✓	✓	UMIs

Modified from: Vallejos et al. (2017) Nature Methods

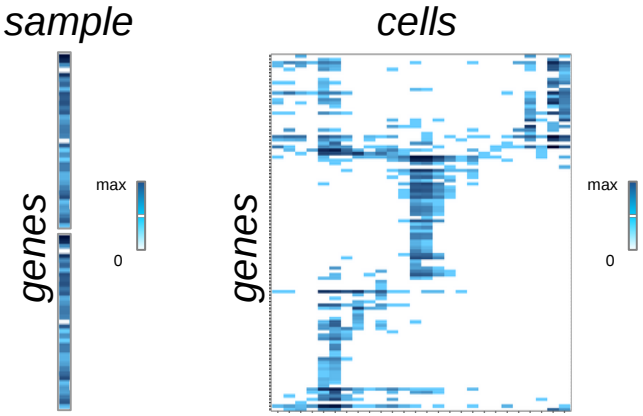
Normalization: sequencing depth

The problem of zeros-inflation:

	BULK	SINGLE-CELL
Total RNA	100 ng (~10.000 cells)	10 pg (per cell)
mRNA	~ 5 ng (~10.000 cells)	<< 1 pg (per cell)
Reads	~100 million	~ 50 k (per cell)

Sample	1
GENE 1	1
GENE 2	4
GENE 3	0
⋮	⋮
GENE M	6

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0



SC MATRIX IS SPARSE !
(ie, mostly filled with zeros)

Normalization: sequencing depth

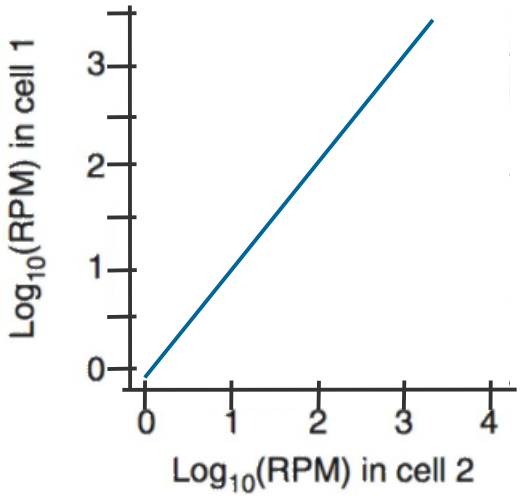
The problem of zeros-inflation:

Origin of zeros: - the gene not expressed in the cell
- the gene expressed, but not detected due to the limitations of current experimental protocols (capture/RT efficiency or low sequencing depth)

→ **dropouts**.

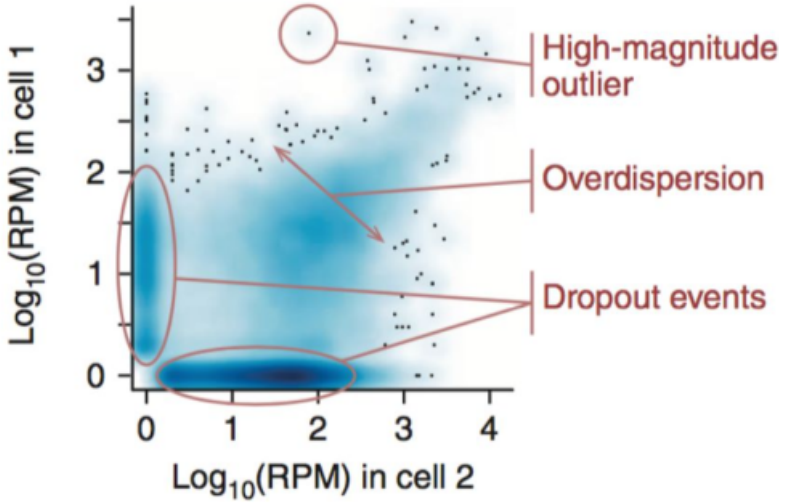
Example of 2 cells with the same cell type:

In theory:



Same level of expression in all cells of the same cell type

In practice:



Kharchenko et al. (2014) Nature Methods

Dropout : gene expressed in one cell and not in the other.
Overdispersion : gene expressed in both cells but not at the same level.
Outlier : High overdispersion.

Normalization: sequencing depth

The solution of dropouts: the Imputation

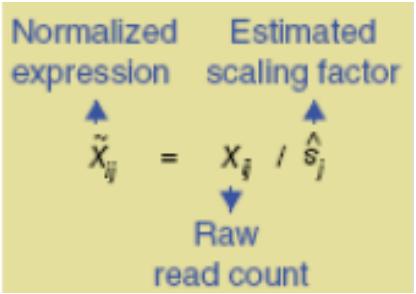
Imputation : estimation of real dropouts values.
Limit : - difficult to identify dropouts events from real biological zeros.
- a lot of zero values so very little information for the estimation.
Danger : if errors in the imputation → consequences on downstream analyzes!

Community opinion :
In theory : we have to do it.
In practice : **there isn't much improvement in the results**, so many researchers don't do it.
Svensson, jan. 2020, Nature Biotechnology: there is no dropouts in droplet techniques.
Hou et al., aug. 2020, Genome Biology: no improvement in downstream analyzes.
But the different R packages above demonstrate their usefulness.

Normalization: sequencing depth

Bulk normalization methods can't be used for the single cell:

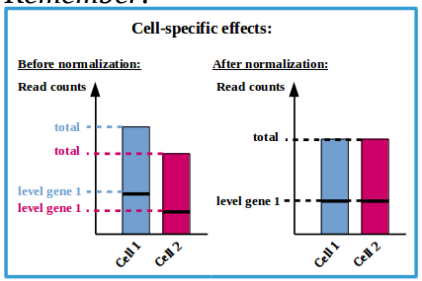
- **RPKM/FPKM** (Reads/Fragments per kilobase of transcript per million reads of library) :
  Normalize for sequencing depth and transcript length at the same time → **KO if NOT full-length data**
- **TPM/CPM** (Transcripts/Counts Per Million): **KO if a small number of genes carry most of the signal**
- **Included in statistical model** :
  DESeq2, edgeR: suppose that **>=50% of genes are NOT DE** (**KO with different cell types**) & **uses** only genes with **non-zero expression** values across all cells (**KO if too many zeros**).
- **Global scaling**:
  LogNormalize: normalizes the gene expression measurements of each cell by the total expression (represented by a scale factor) and transforms (log) the result.



- hypothesis: cell populations are homogenous & the RNA level is similar in all cells.
- choice of the scaling factors: - Upper Quartile UMI counts
  - median UMI counts (CellRanger)
  - quantile 99% (Scater)
  - excluding zeros prior to calculating
  - 10 000 default (Seurat)

KO if you have too many zeros

Remember:



- in practice: **hypotheses are not always verified**, but people use this method

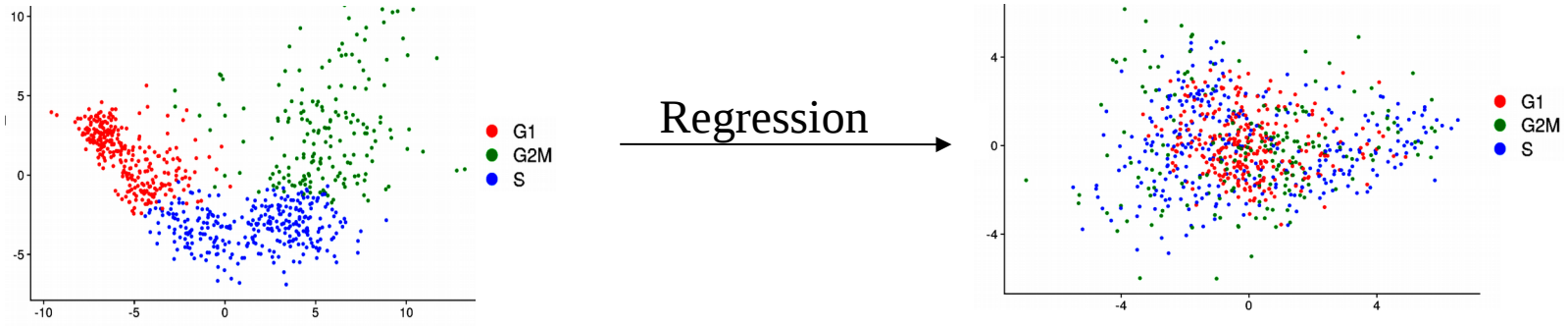
  Normalization included in statistical model:
  - tools: SCDE, Monocle, MAST, SCTransform (Seurat), ...

There is no perfect normalization, it depends on data!

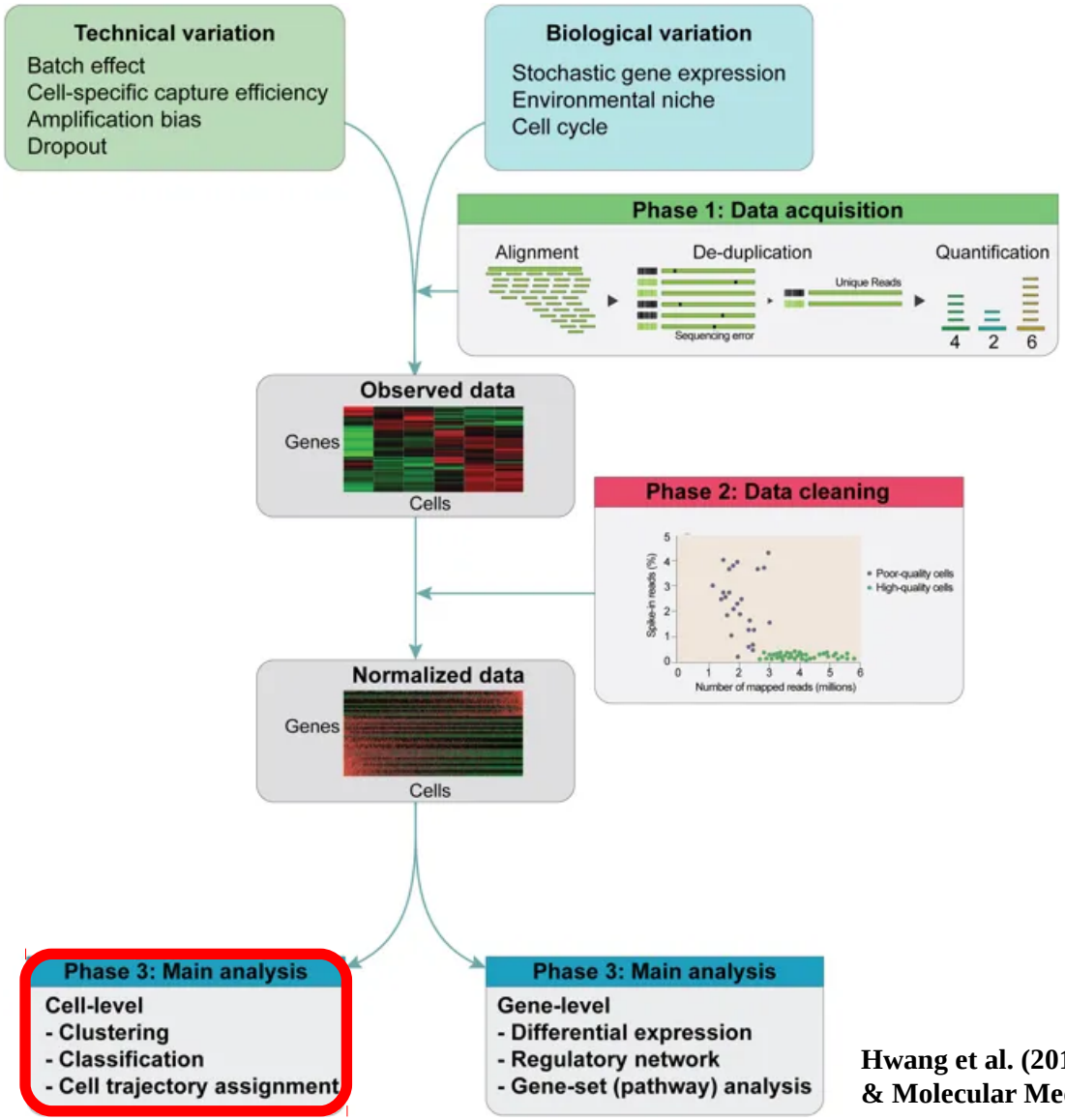
Rough solution : global log-normalization

Normalization: biological factors

- Regression methods provided to account for know factors:
  - Cell cycle, % mitochondrial RNA, % ribosomal protein RNA, …
  - R packages : Seurat (at the same time as depth correction).



Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

Goal:
Identify cell types

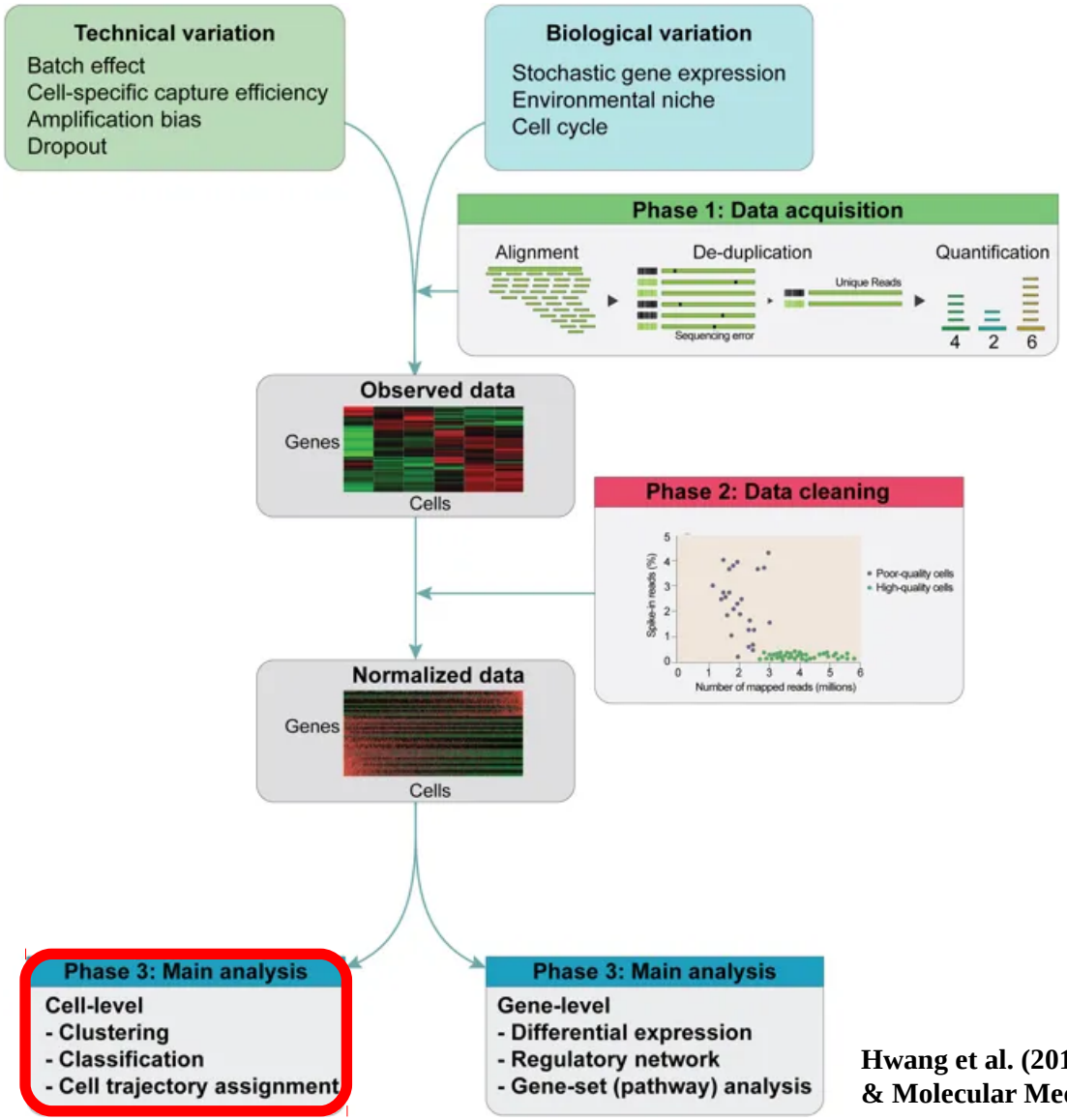
Step 1:
From a normalized matrix to a reduced space (Dimension reduction)

1) Feature selection
2) Feature extraction

Step 2:
Identify similar cells in reduced space (Clustering)

Step 3:
Identify cells types of similar cells (Classification)

Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

Goal:
Identify cell types

Step 1:
From a normalized matrix to a reduced space (Dimension reduction)

1) Feature selection
2) Feature extraction

Step 2:
Identify similar cells in reduced space (Clustering)

Step 3:
Identify cells types of similar cells (Classification)

High dimension: problems

Problem:

BULK

Sample	1
GENE 1	1
GENE 2	4
GENE 3	0
⋮	⋮
GENE M	6

Data * *N* →

SINGLE-CELL

Cell:	1	2	···	N
GENE 1	1	2		14
GENE 2	4	27		8
GENE 3	0	0		1
⋮	⋮	⋮		⋮
GENE M	6	2		0

1 bulk = 1 dimension

1 cell = 1 dimension

N cells = *N* dimensions

High Dimension

- Increase storage & computational efficiency
- Visualize Data in 2D or 3D problem
- Clean noise background
- Curse of Dimensionality problems

Solution:

Goal 1:

Identify cell types

WE ARE HERE

Step 1:

From a normalized matrix to a reduced space (Dimension reduction)

1) Feature selection
2) Feature extraction

Feature selection: filter genes

- Goal: To **select genes that contain useful information about the biology** of the system while **removing genes that contain random noise.**

- How:

  - Keep only highly expressed genes?

  - Keep only Highly Variable Genes (500 < HVG < 3000) → several ways to calculate variability

Feature extraction

Goal: To **combine genes that contain useful information about the biology** of the system**.**

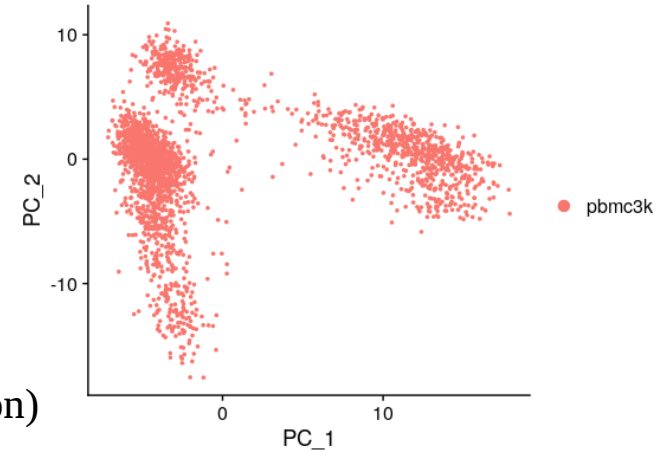
How (popular methods used):

For calculation:

- PCA (Principal Component Analysis)
- scBFA (single-cell Binary Factor Analysis)

For representation:

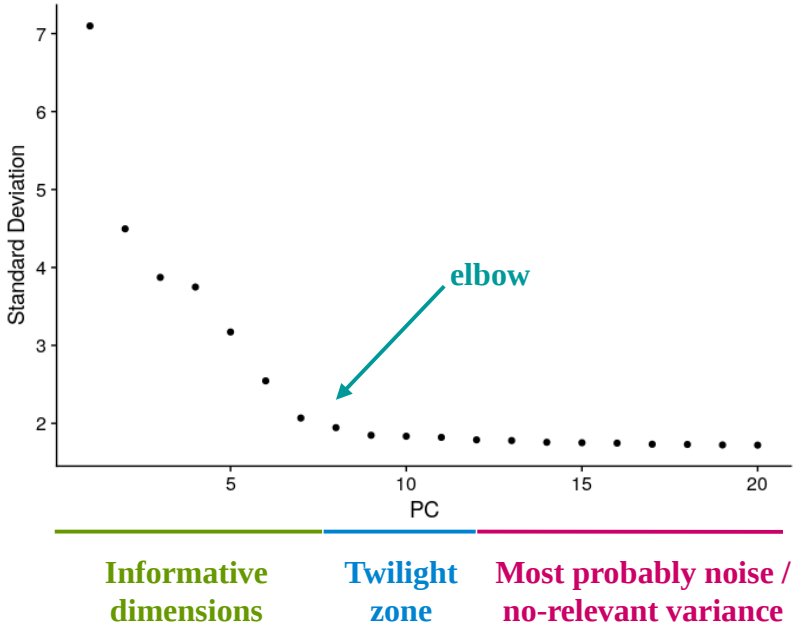
- t-SNE (t-distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection)



Feature extraction: PCA

How to determine the number of PCs to keep? (Often between 5 and 50 PCs)

- Graphic: the Elbow plot



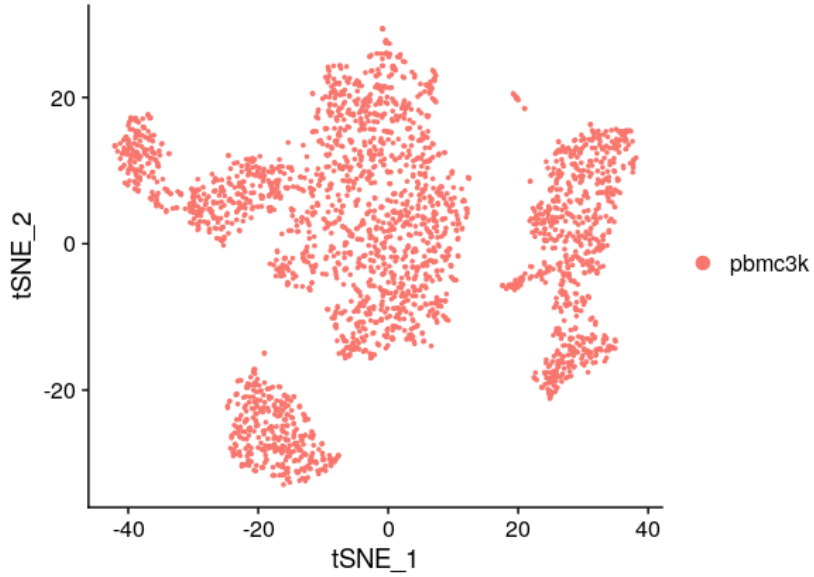
- Algorithms:
  - JackStraw (Macosko et al (2015) Cell): use a resampling test by randomly permutations of 1% of genes and rerun PCA, several times. 'Significant' PCs as those who have a strong enrichment of genes with low p-values.
  - To retain all PCs until the percentage of total variation explained reaches some threshold (threshold is determined by the proportion of variance attributed to the biology versus technical noise).
  - Assumes: 1 sub-population = 1 type of variation = 1 PC; if expected 10 sub-populations, number of PCs is average 10.

Feature extraction: t-SNE

t-SNE: t-Stochastic Neighbor Embedding **(Van der Maaten and Hinton (2008) JMLR)**

- Goal:
  To find an optimal low-dimensional representation of the data (2-3 dimensions) that preserves the neighborhood of each point from the high-dimensional space (ie. similar points remain similar) based on a probabilistic interpretation of the proximities => **preserve local structure.**

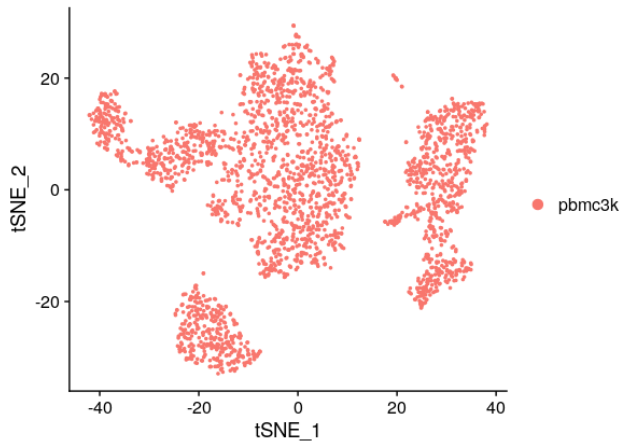
- How:
  To minimize the divergence between two distributions: a distribution that measures pairwise similarities of the input objects and a distribution that measures pairwise similarities of the corresponding low-dimensional points in the embedding.

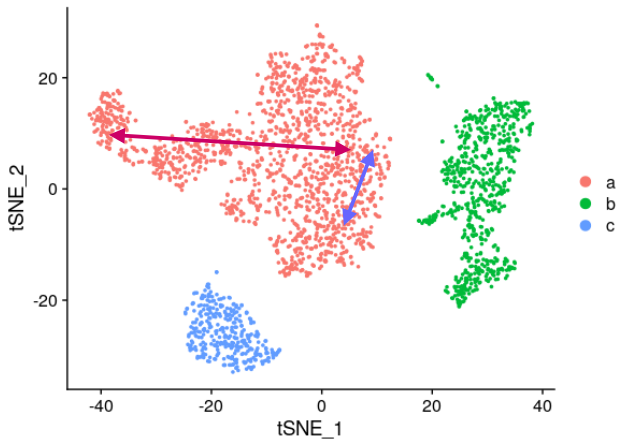


Feature extraction: t-SNE

- Advantages:
  - not restricted to linear transformations.
  - **neighborhood is important but distance from neighbors is not important** => enabling it to separate many distinct clusters in a complex population (easier than PCA), but *

- Limits:
  - * we cannot use positions of populations to determine relationships between them.





Cells of pop a are similar.
This cells are more similar than this cells.

~~pop a and pop b are closer than pop a and pop c.~~

  - * coordinates after embedding have no meaning => just for visualization.

Feature extraction: UMAP

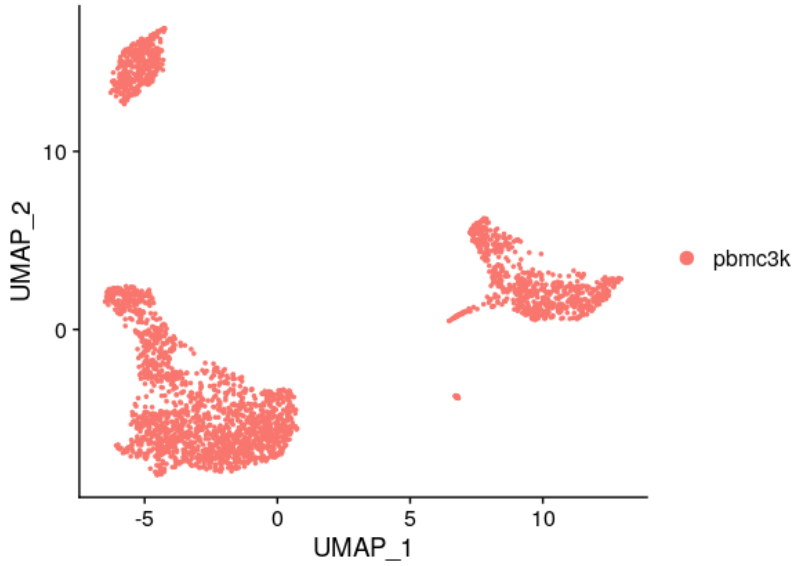
UMAP: Uniform Manifold Approximation and Projection **(McInnes, Healy, and Melville (2018) arXiv)**

- Goal:
  To find an optimal low-dimensional representation of the data (2-3 dimensions) that preserves the k neighborhood of each point from the high-dimensional space (ie. similar points remain similar **and** distinct points remain distinct) based on a probabilistic interpretation of the proximities => **preserve local structure and global structure!**

- How:
  1- Construct a topological presentation of the high-dimensional data (in this case a weighted k-NN graph).

  2- Given a low-dimensional data, construct a graph in the similar way.

  3- Minimize the dissimilarity between the two graphs: look for the low-dimensional data whose graph is the closest to that of the high-dimensional data.



- Definition:
  k-NN graph: is a neighborhood graph in which two vertices (i.e. cells) p and q are connected by an edge, if the distance between p and q is among the k smallest distances from p to the other vertices.

Feature extraction: UMAP

- Advantages:
  - not restricted to linear transformations.
  - to have more compact visual clusters with more empty space between them.
  - to preserve the local and the global data structure.
  - much faster method for visualization (more and more used).

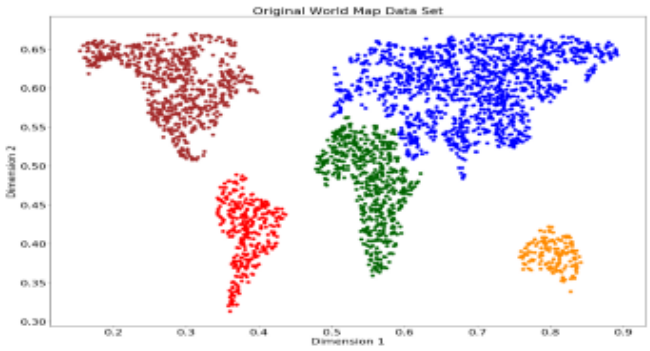
- Limits:

  - require specification of very important parameters : the number of neighbors and the minimum distance between embedded points.

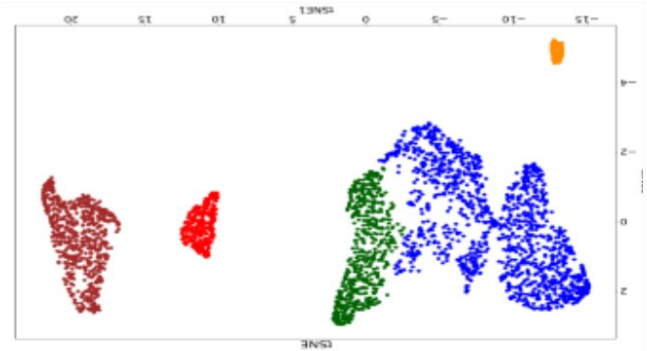
  If these values are:
  - too low: noise will be treated as structure,
  - too high: discard fine structure.

Feature extraction: Conclusion

RAW



t-SNE



UMAP

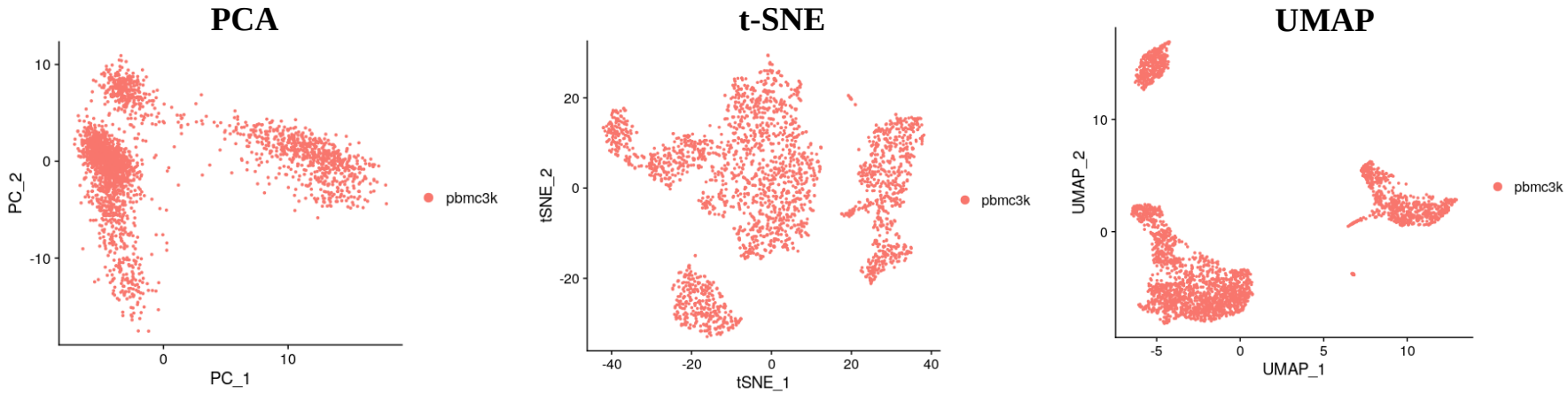
https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17

Comparisons:

t-SNE: preserve local structure.

UMAP: preserve local **and** global structure.

Feature extraction: Conclusion



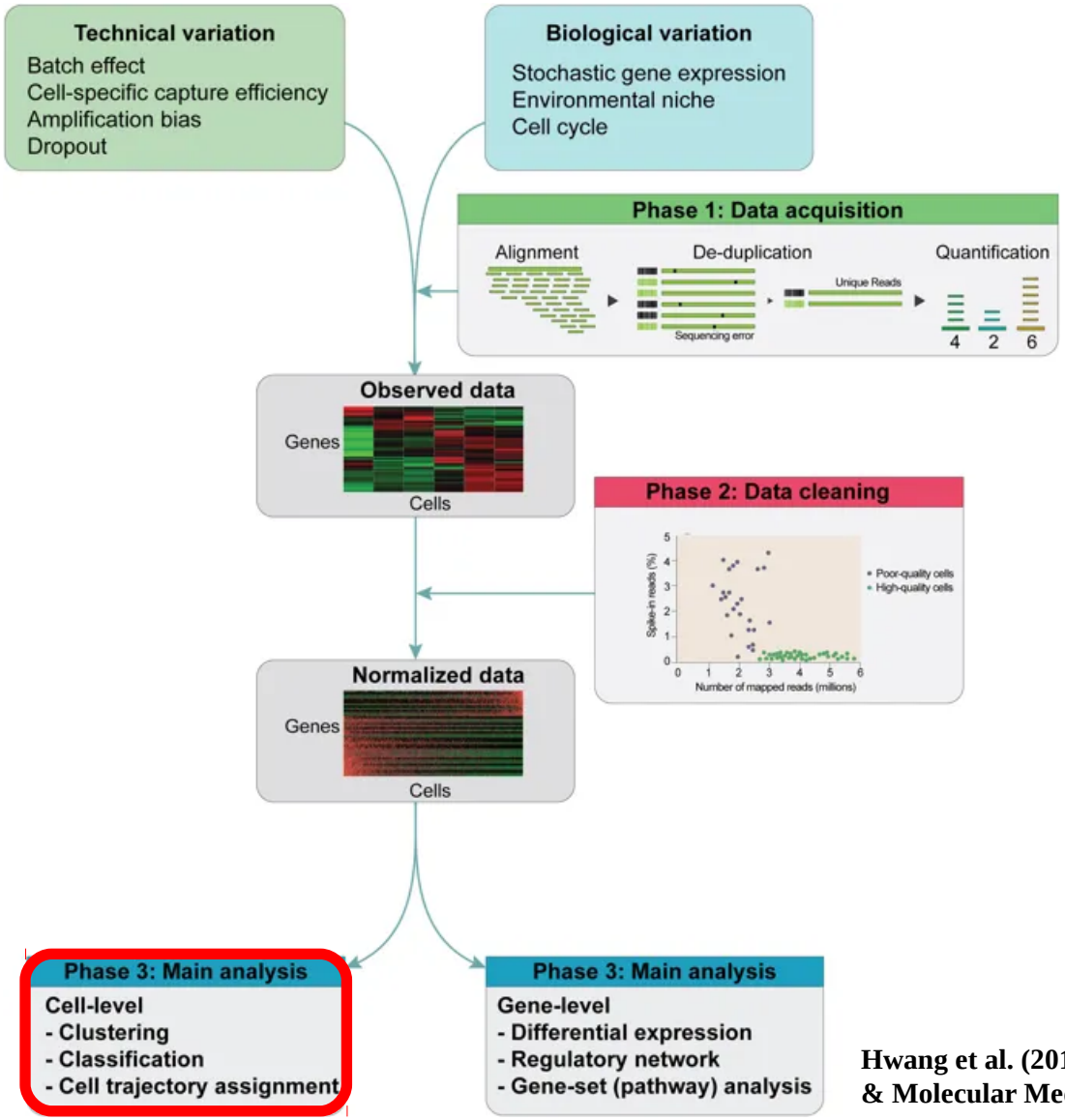
General rules :

Dimensionality reduction necessarily involves **discarding information** in order to fit high-dimensional data into a lower dimensional space (worse in 2 dimensions).

We would **not perform downstream analysis on the t-SNE/UMAP** coordinates, but **perform it on** the first 5-50 PCs from **PCA** and then visualize on the t-SNE/UMAP plot. This ensures that downstream analysis makes use of the information that was lost during compression into two dimensions for visualization.

To **choose the parameters**, we must **repeat the visualization** several times to ensure that the results are representative by testing a **range of value**s for these parameters.

Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

Goal:
Identify cell types

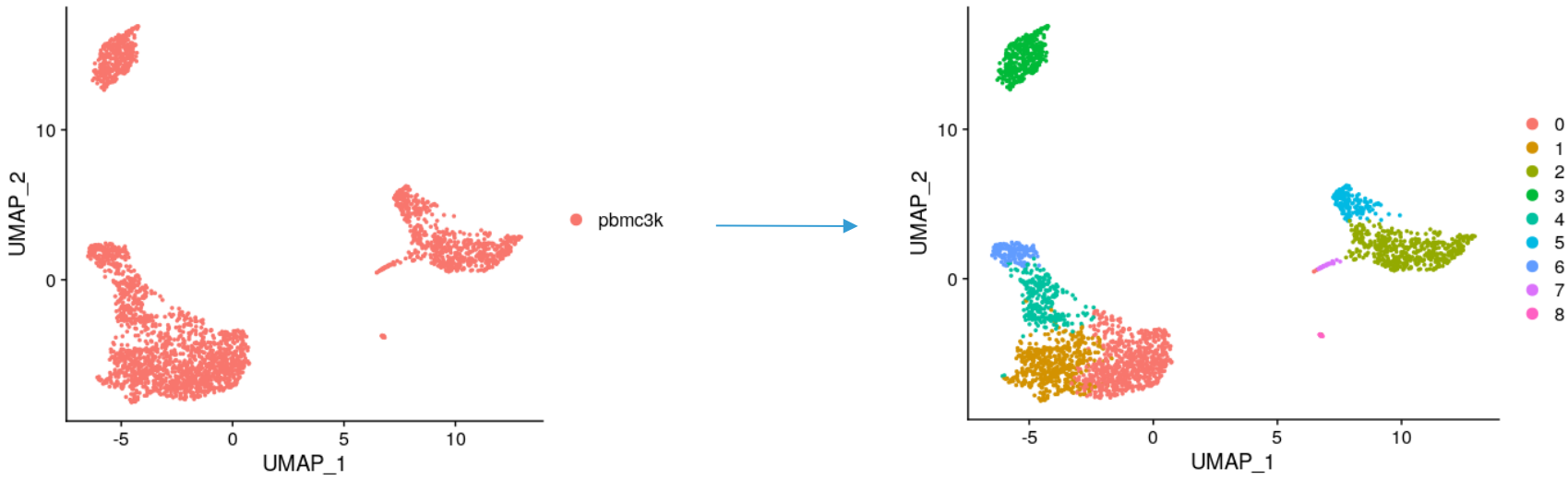
Step 1:
From a normalized matrix to a reduced space (Dimension reduction)

1) Feature selection
2) Feature extraction

Step 2:
Identify similar cells in reduced space (Clustering)

Step 3:
Identify cells types of similar cells (Classification)

Clustering



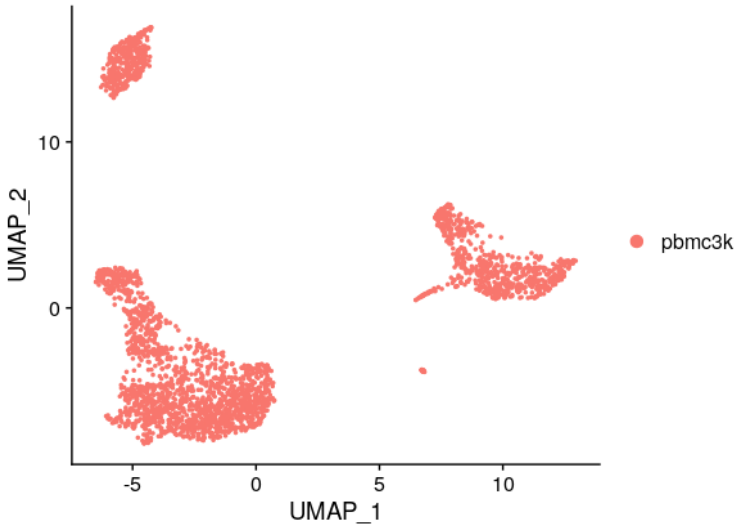
What cell types / states do these similar cell groups belong to?

Clustering

- Goal:
  To empirically **define groups of cells** with similar expression profiles => a critical step for extracting biological insights.
  Unsupervised Clustering:
  To identify groups of cells based on the similarities of the transcriptomes **without any prior knowledge of the labels.**

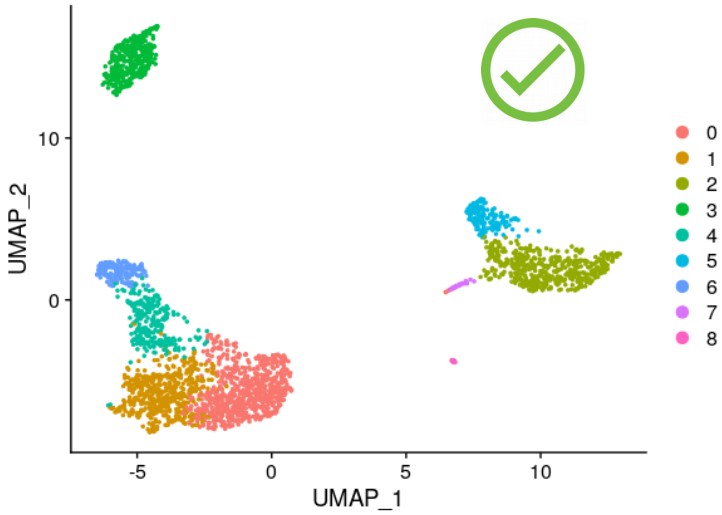
- Challenges :
  What is the number of clusters k?
  What defines a good clustering?
  What is a cell type? State type?
  Scalability: in the last few years the number of cells in scRNA-seq experiments has been multiplied to 10^4.

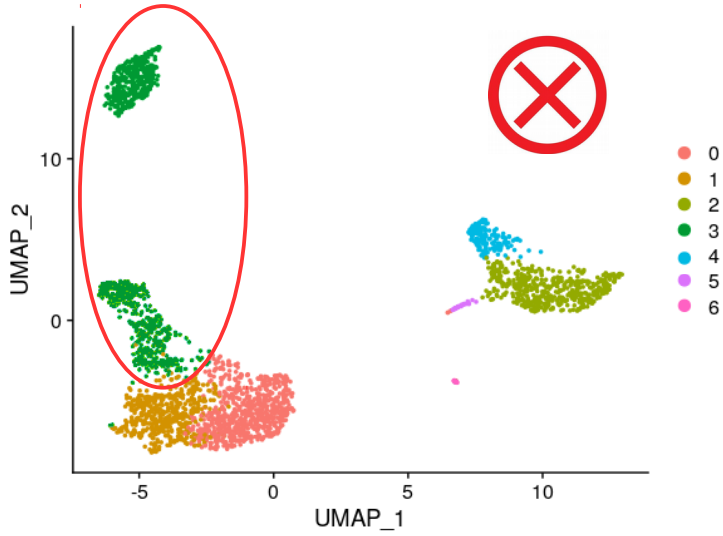
Resolution (granularity)

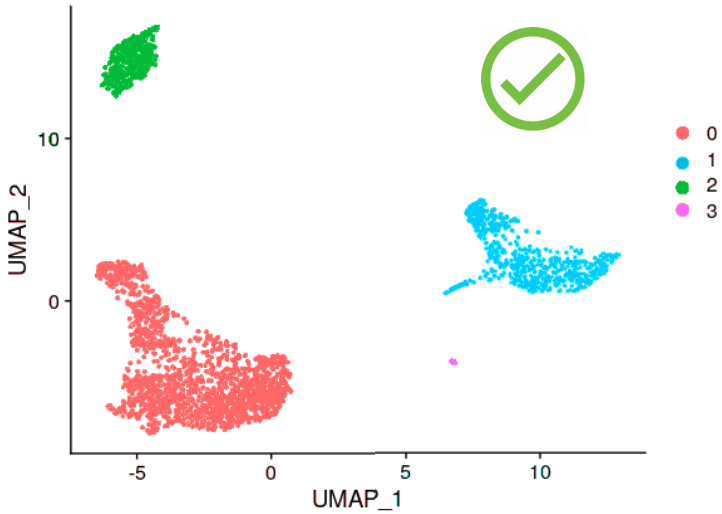


- Methods:
  Hierarchical clustering, k-means clustering, Graph-based methods, Model-based, Kmedoids, ...

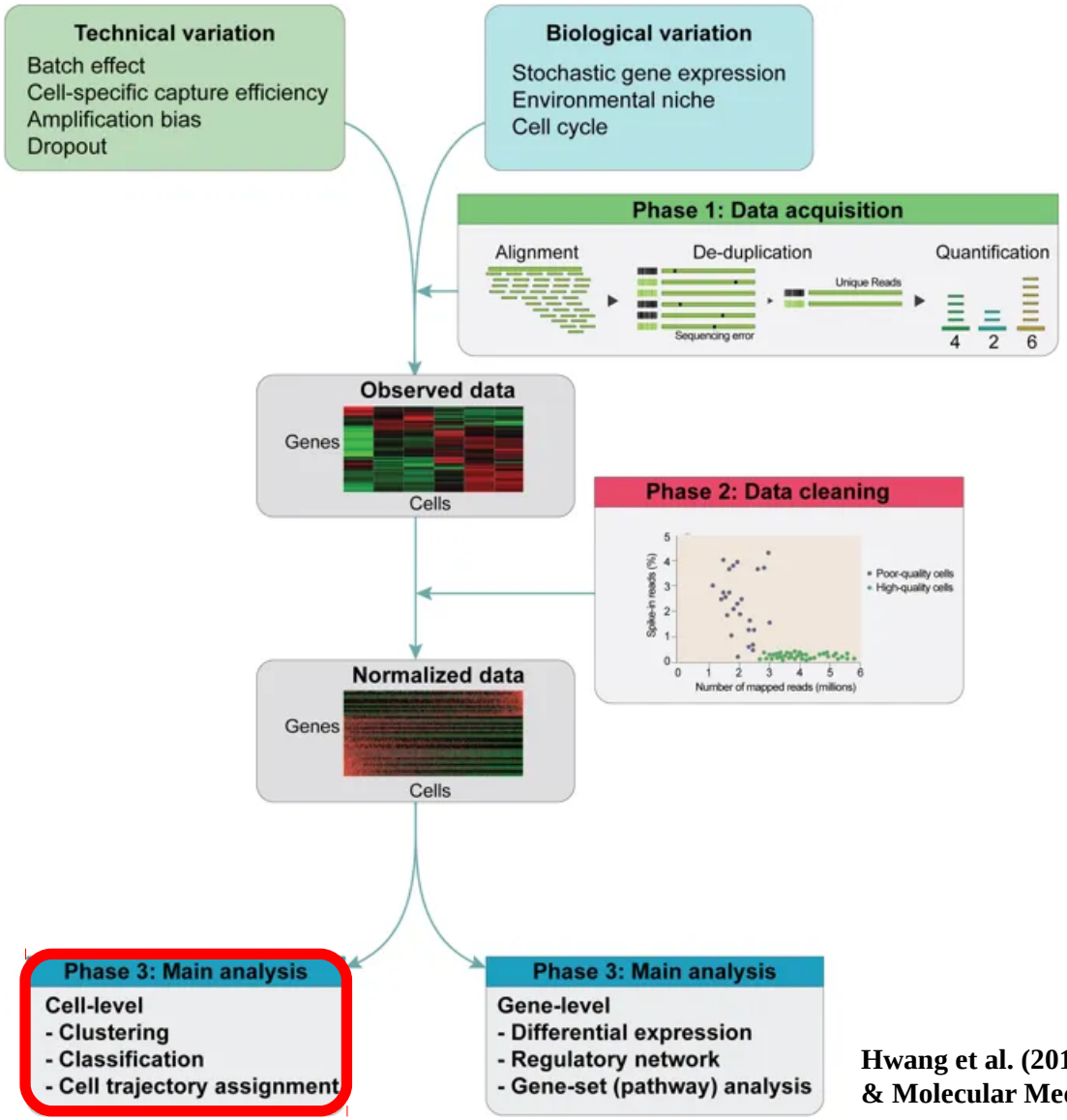
Clustering


UMAP_2
10
0
-5
0
5
10
UMAP_1
0
1
2
3
4
5
6
7
8


UMAP_2
10
0
-5
0
5
10
UMAP_1
0
1
2
3
4
5
6


UMAP_2
10
0
-5
0
5
10
UMAP_1
0
1
2
3

Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

Goal:
Identify cell types

Step 1:
From a normalized matrix to a reduced space (Dimension reduction)

1) Feature selection
2) Feature extraction

Step 2:
Identify similar cells in reduced space (Clustering)

Step 3:
Identify cells types of similar cells (Classification)

Cluster annotation (Classification)

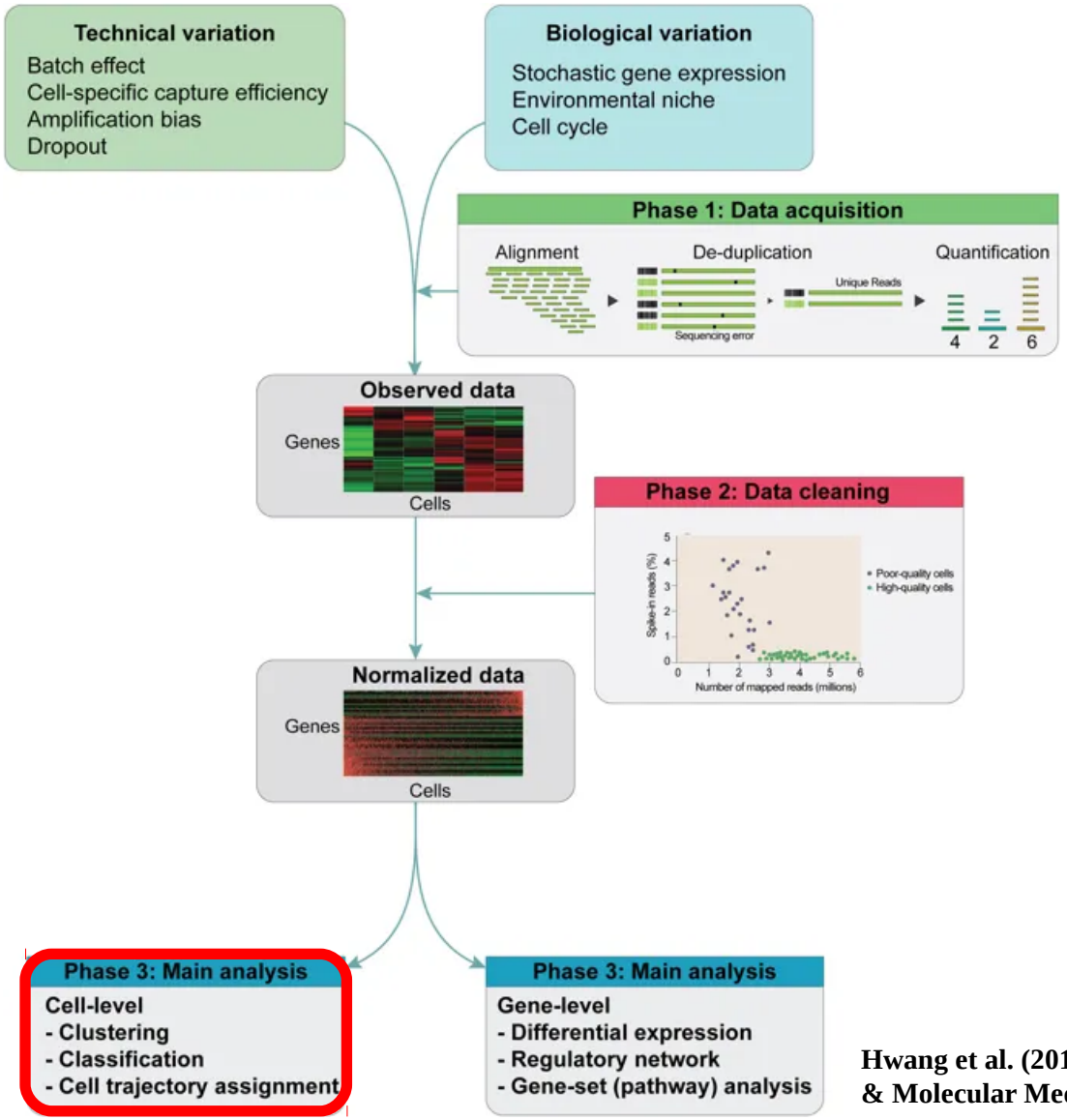
- Goal:
  Identify the cell types/states, corresponding to the clusters, that make up our sample.
  This is one of the most difficult and crucial steps in single-cell analysis.

- Automatic annotation:
  Compare the expression profiles to published reference datasets where each sample or cell has already been annotated with its putative biological state by domain experts.
  References: microarray, RNA-seq bulk, sc RNA-seq.
  Limits:
    - missing references (Human Cell Altas soon?)
    - the reliability of the reference depends greatly on the expertise of the original authors who assigned the labels in the first place (potential error transmission)
    - some tools annotate clusters even if their annotation is probably wrong

  R Packages: SingleR, clustifyr … no tool better than the others.

- Cells annotation (we can also annotate cells individually, without taking clusters into account):
  Limits:
    - heterogeneity between cells of the same type + noise = bad results. For the annotation on clusters, the heterogeneity and the noise are smoothed because we make a kind of average of the expression of the cells of each cluster.
    - Slow: annotation of several thousand cells, against around ten clusters.

Computational pipeline



Hwang et al. (2018) Experimental & Molecular Medicine

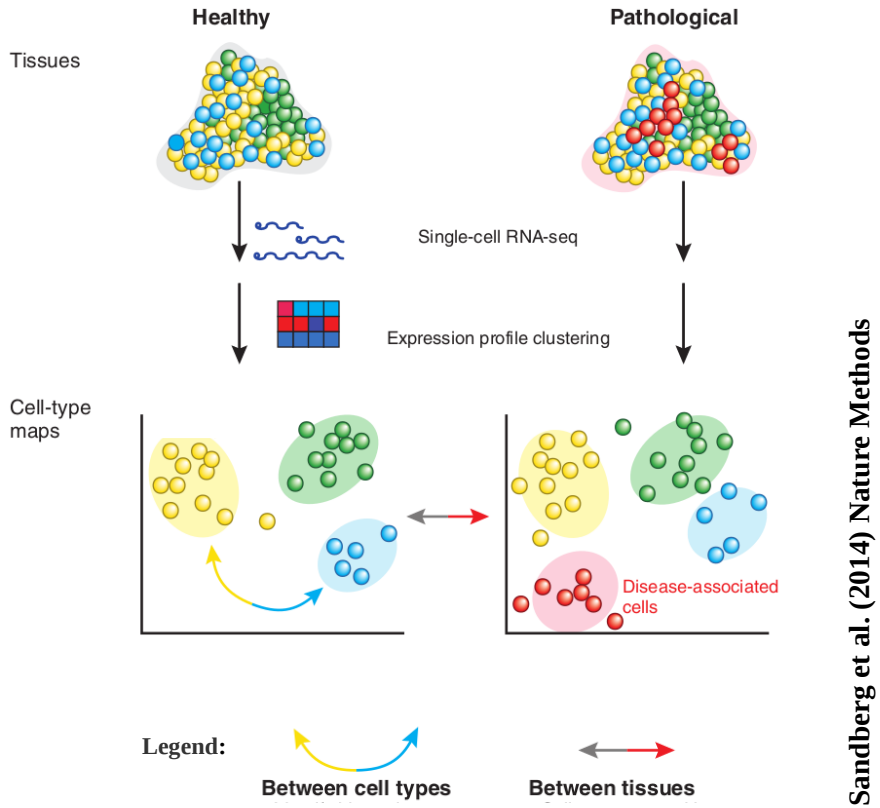
Goal :
Find differential genes expression

Step 1:
Make the **Goal 1 of cell-level analysis** to identify cell types

Step 2:
Use a statistical test to make differential expression analysis between cell types, conditions, tissues.

Differential Gene Expression Analysis

- Goal:
  To identify genes whose **expression differs** under:
  - different cell types (markers genes).
  - different conditions / tissues.



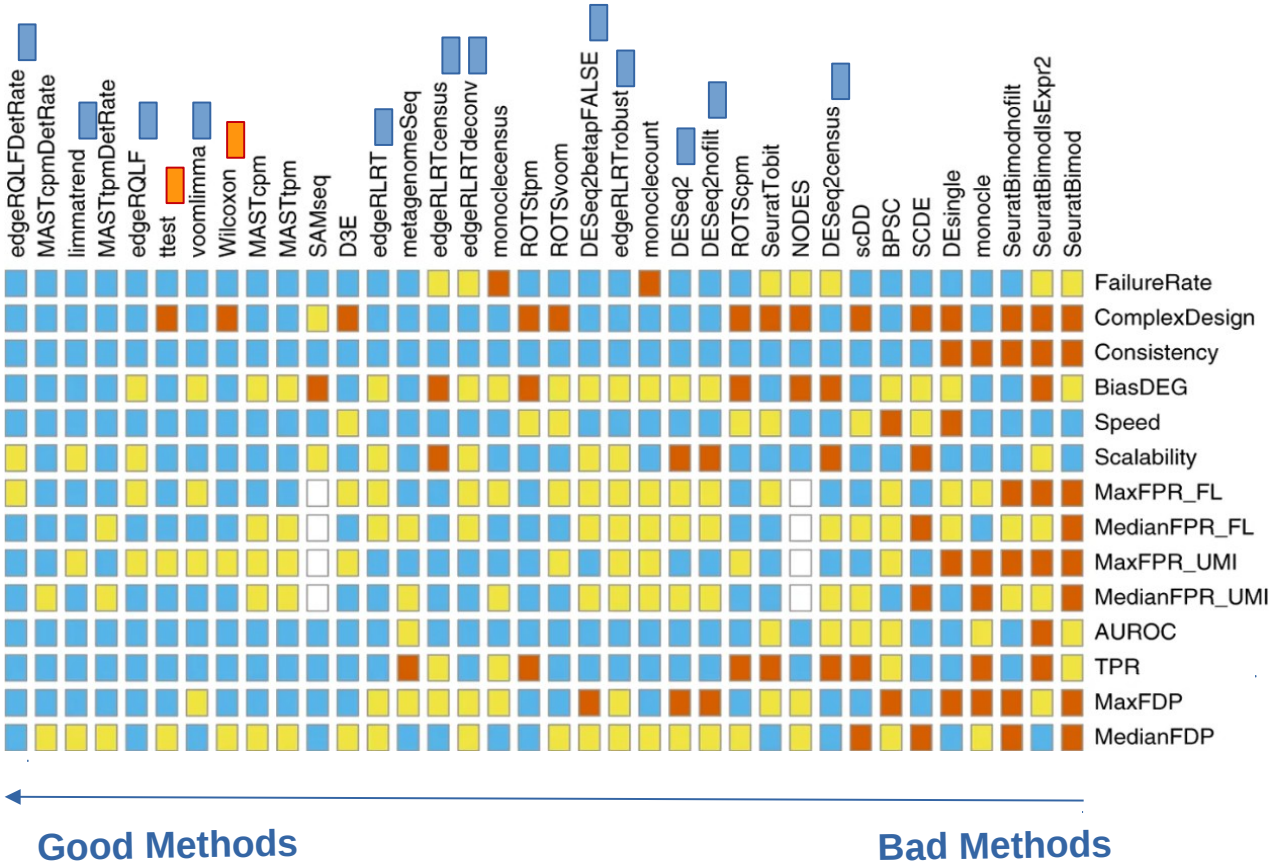
Sandberg et al. (2014) Nature Methods

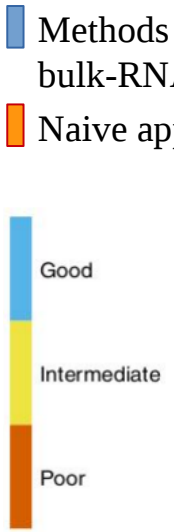
Differential Gene Expression Analysis

- Choice of tools:

Soneson & Robinson (2018) Nature Methods

- 36 approaches on 9 datasets (6 full-length + 3 UMI) + simulations





Methods borrowed from bulk-RNA-seq

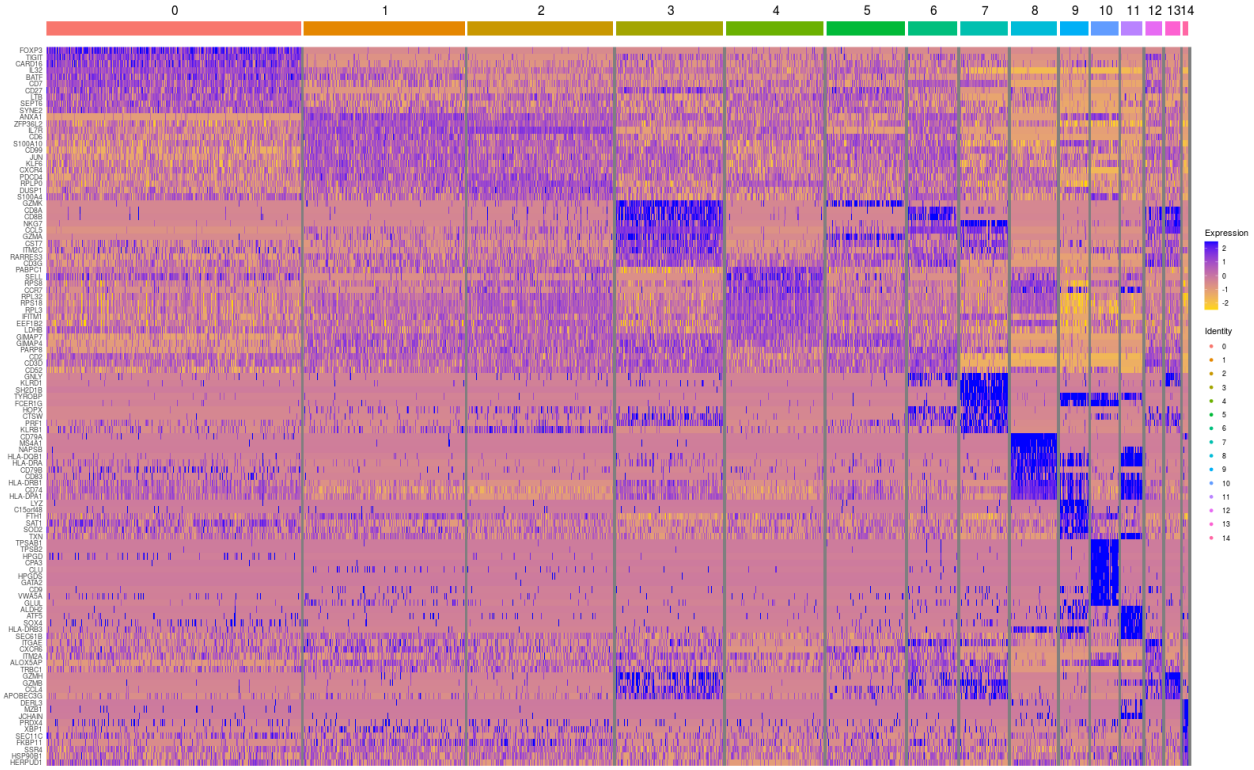
Naive approaches

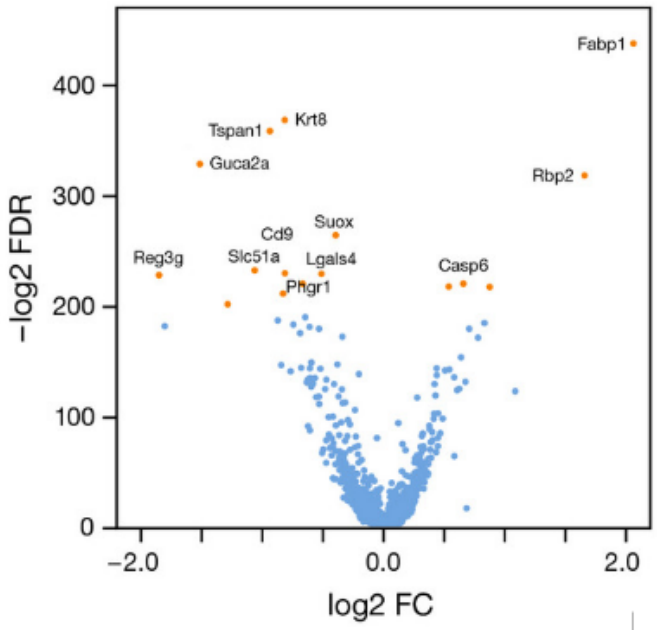
- bulk and naive methods perform well
- edgeR detect a lot FP when expression is weak

Differential Gene Expression Analysis

- Conclusion:

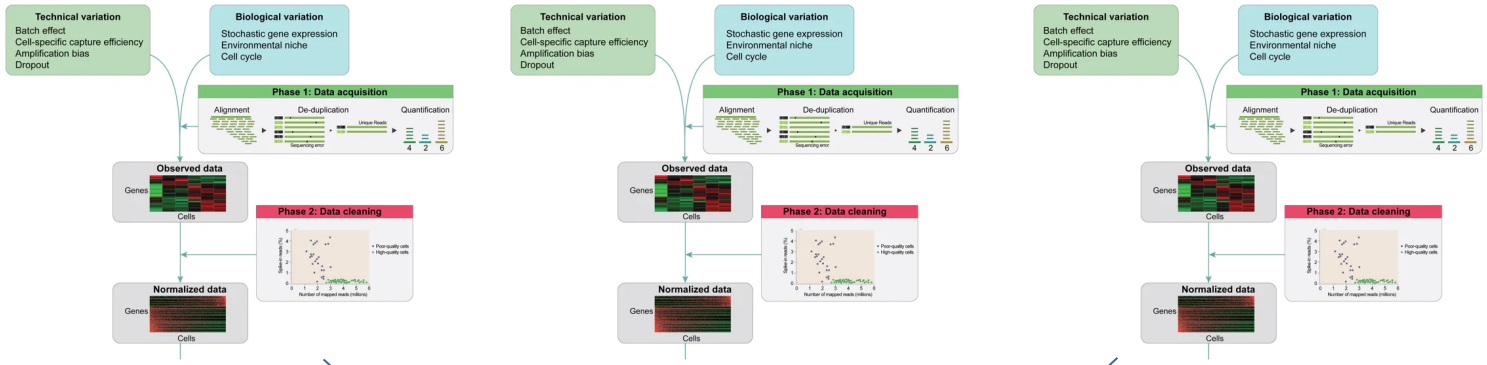
  - methods developed specifically for scRNAseq → not better performance than bulk RNA-seq methods.





Multiple samples

Integration



Integration

Goal :
Grouping several samples

Phase 3: Main analysis

Cell-level
- Clustering
- Classification
- Cell trajectory assignment

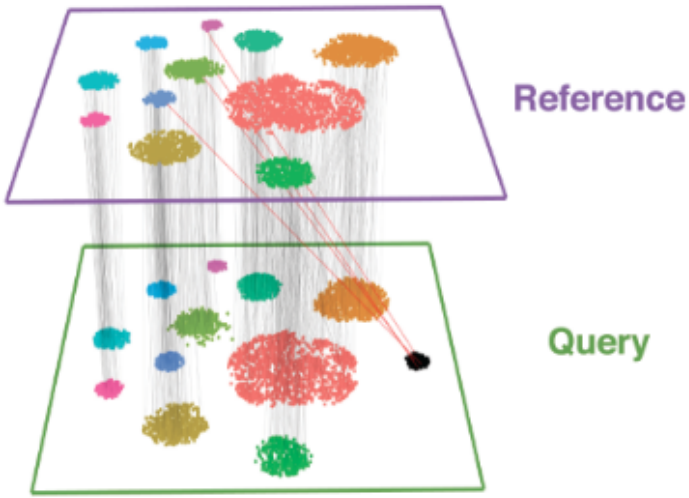
Phase 3: Main analysis

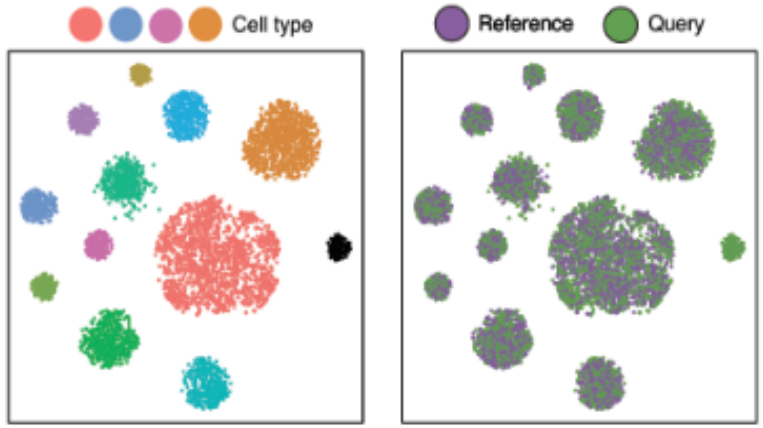
Gene-level
- Differential expression
- Regulatory network
- Gene-set (pathway) analysis

↔ **Comparison with individual analysis**

Integration

Most batch correction algorithms require **at least one identical cell type to be shared between any pair of data batches**, to guide the data alignment.



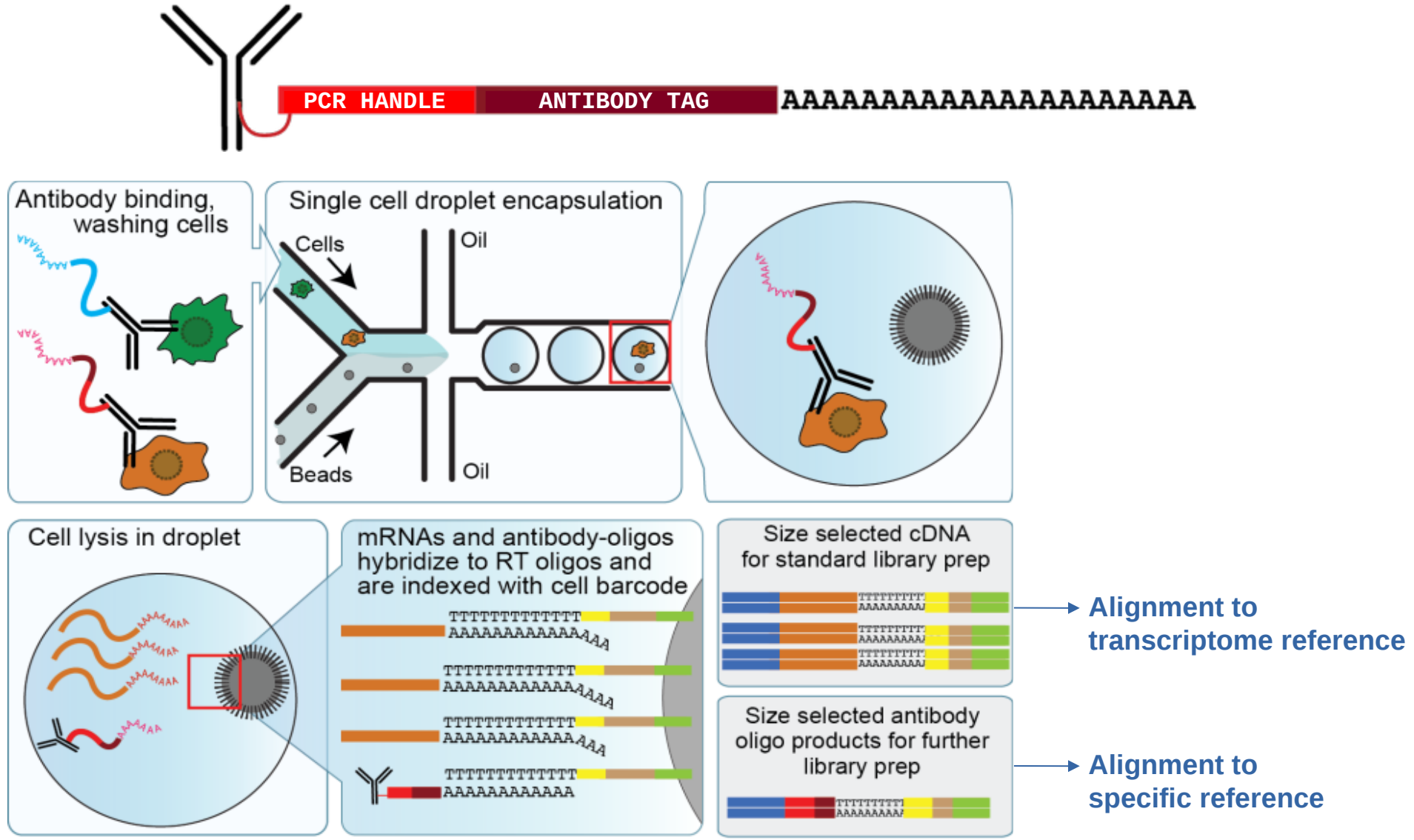


Stuart et al. (2019) Cell

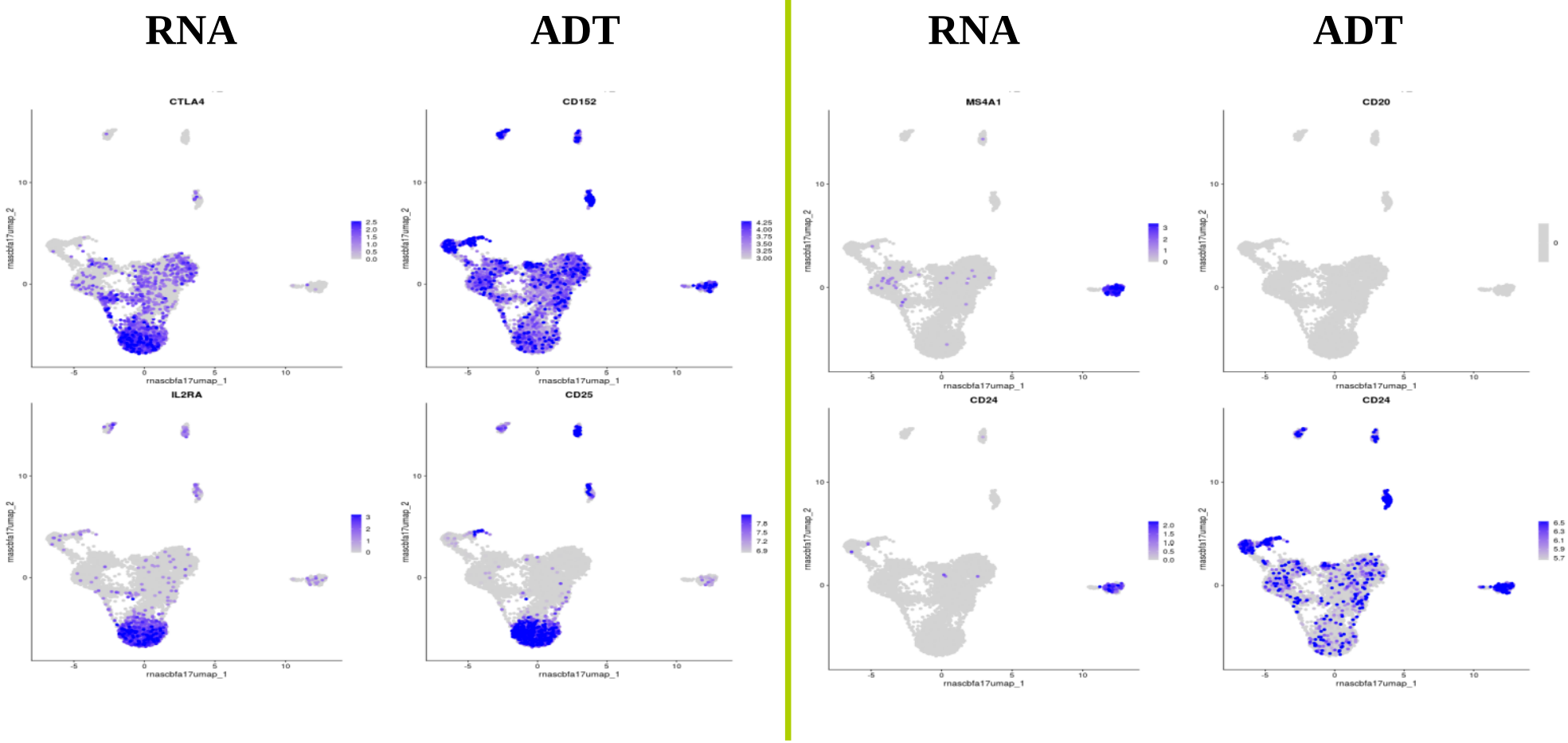
Tools: Seurat, LIGER, Harmony, ...

Profiling of Cell surface proteins: CITE-seq

CITE-seq: ADT (Antibody-Derived Tags)



CITE-seq: ADT (Antibody-Derived Tags)

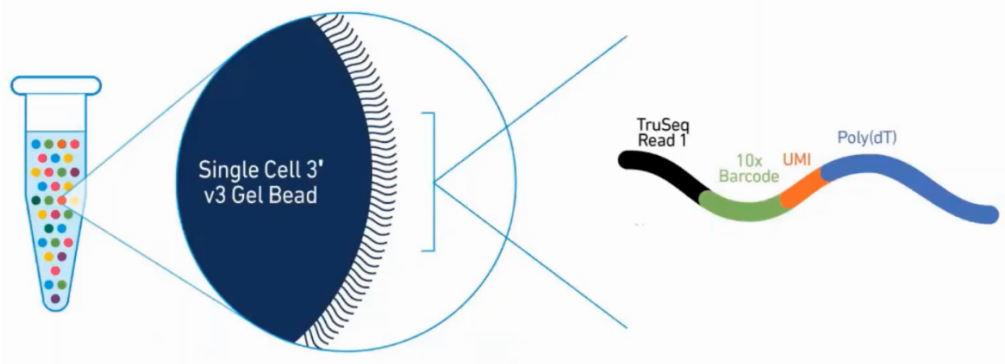


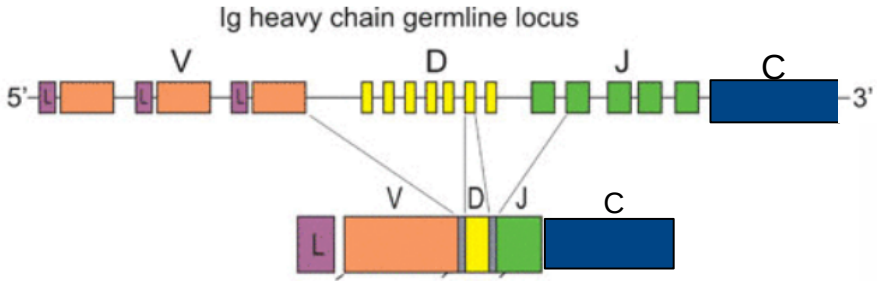
Profiling of Immune Diversity: TCR/BCR

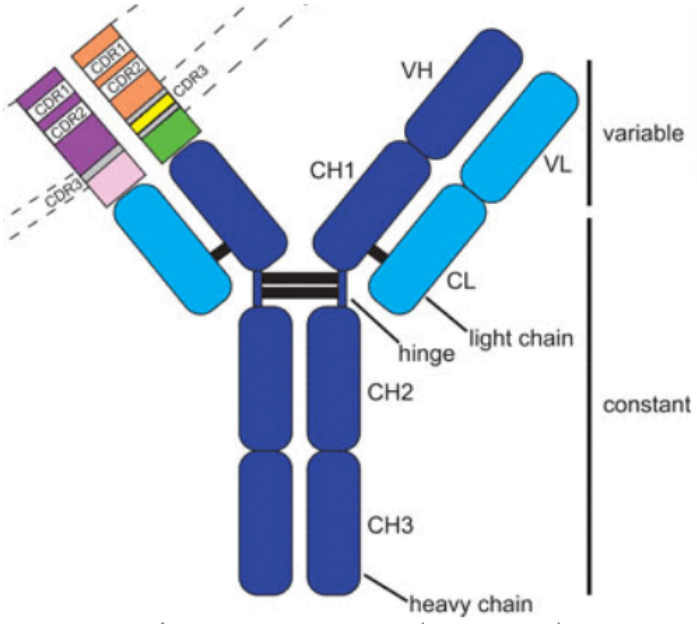
TCR/BCR Profiling

www.10xgenomics.com

scRNA-seq 3’



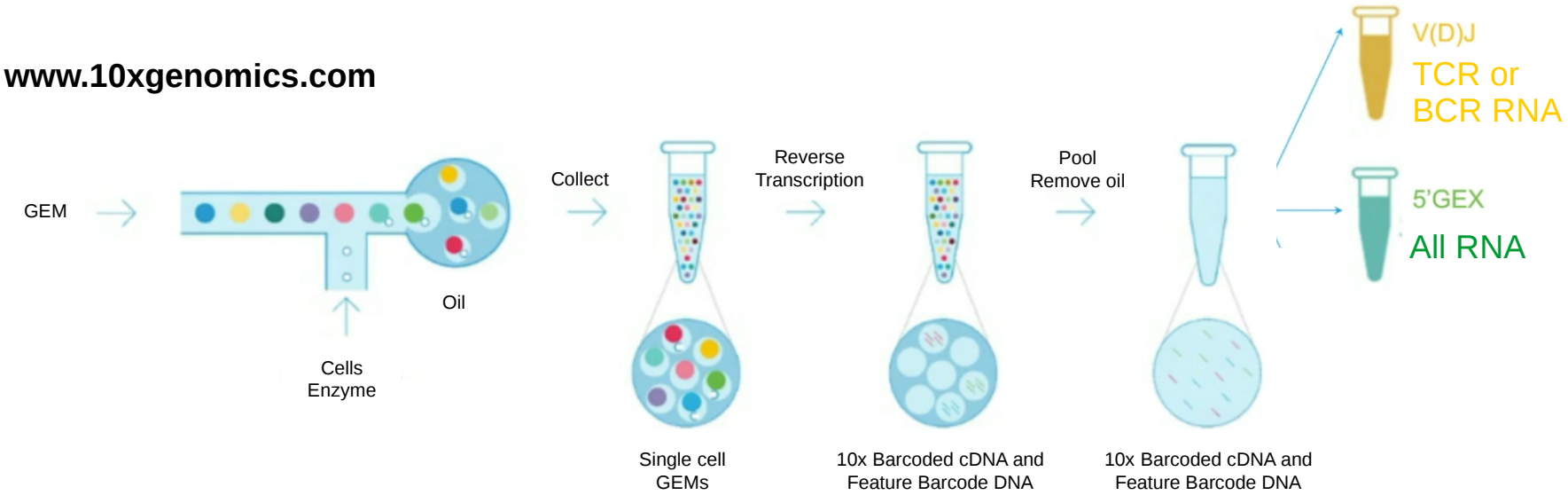


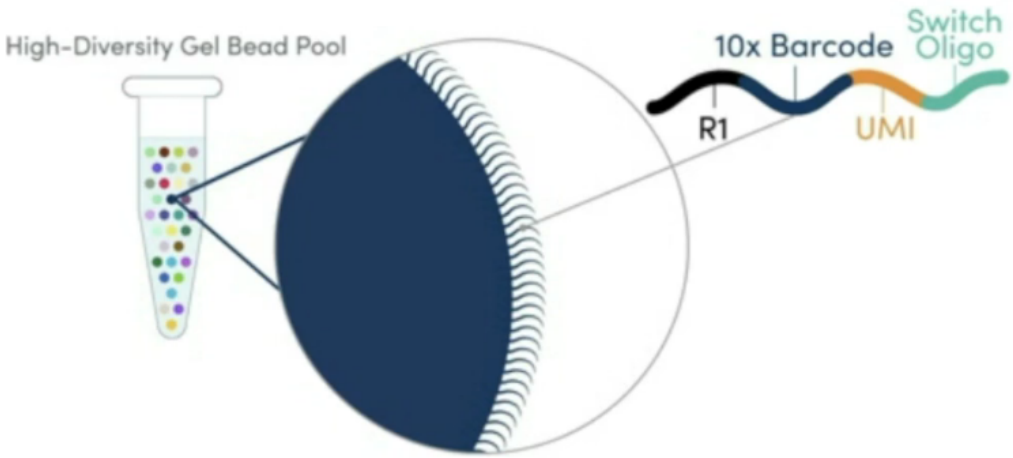


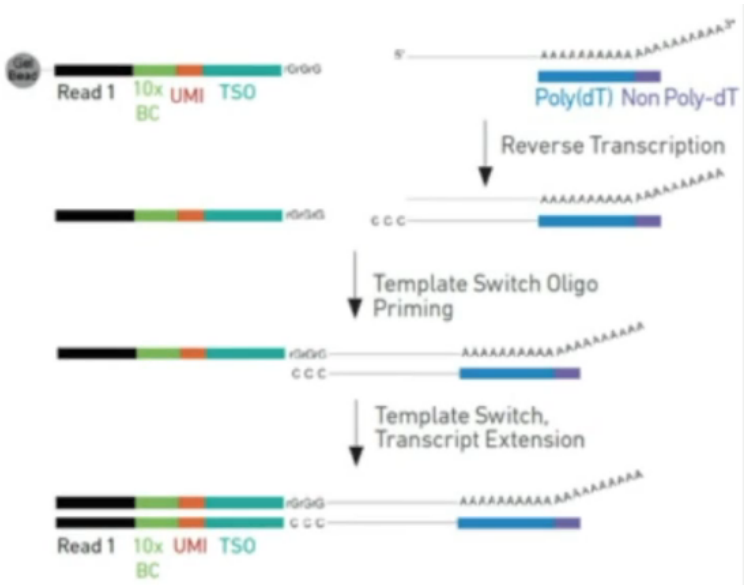
Modified from
Boyd and Joshi. 2014, Microbiol. Spectr.

TCR/BCR Profiling

www.10xgenomics.com

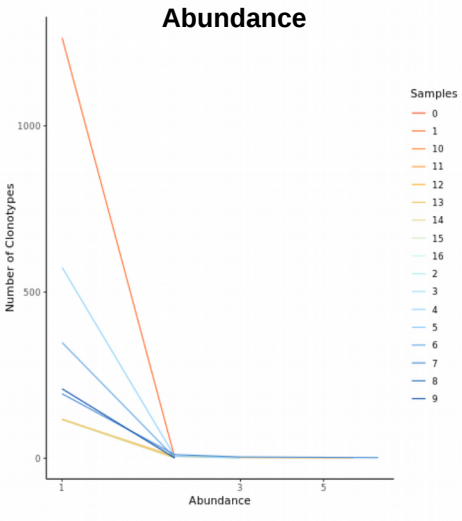




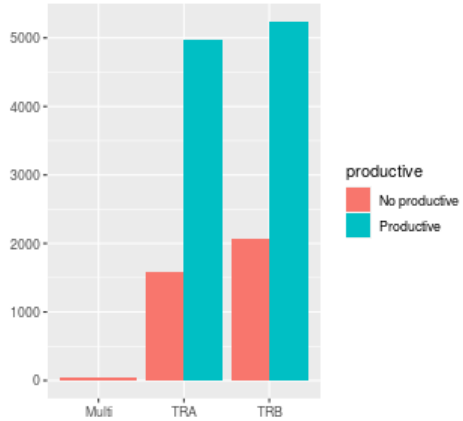


TCR/BCR Profiling

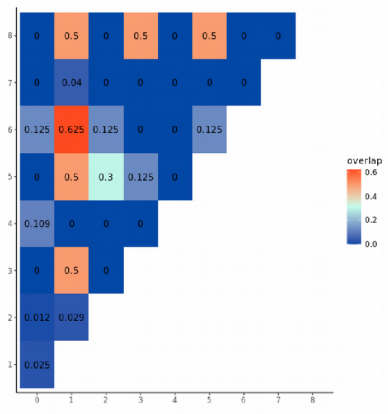
Abundance



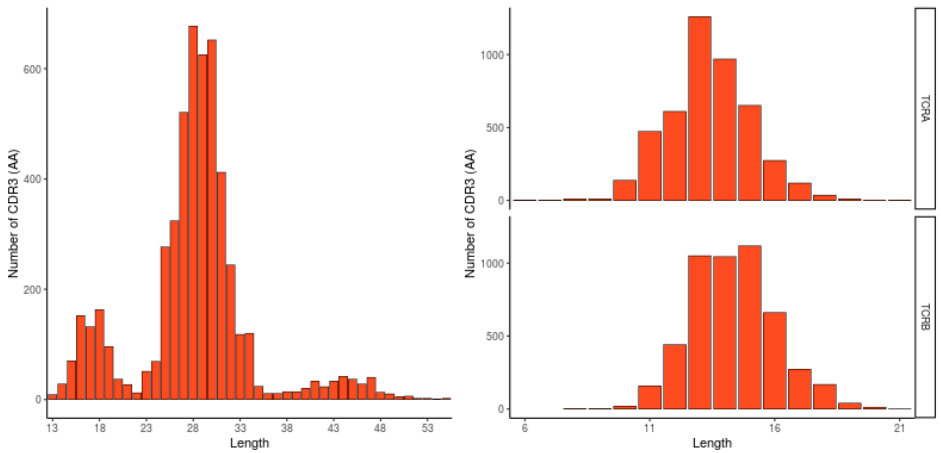
Productivity



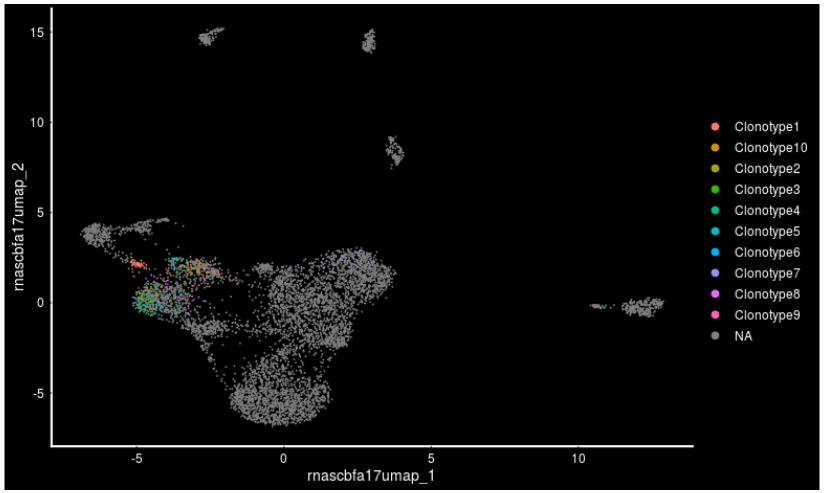
Clonotypes overlap between clusters



Length

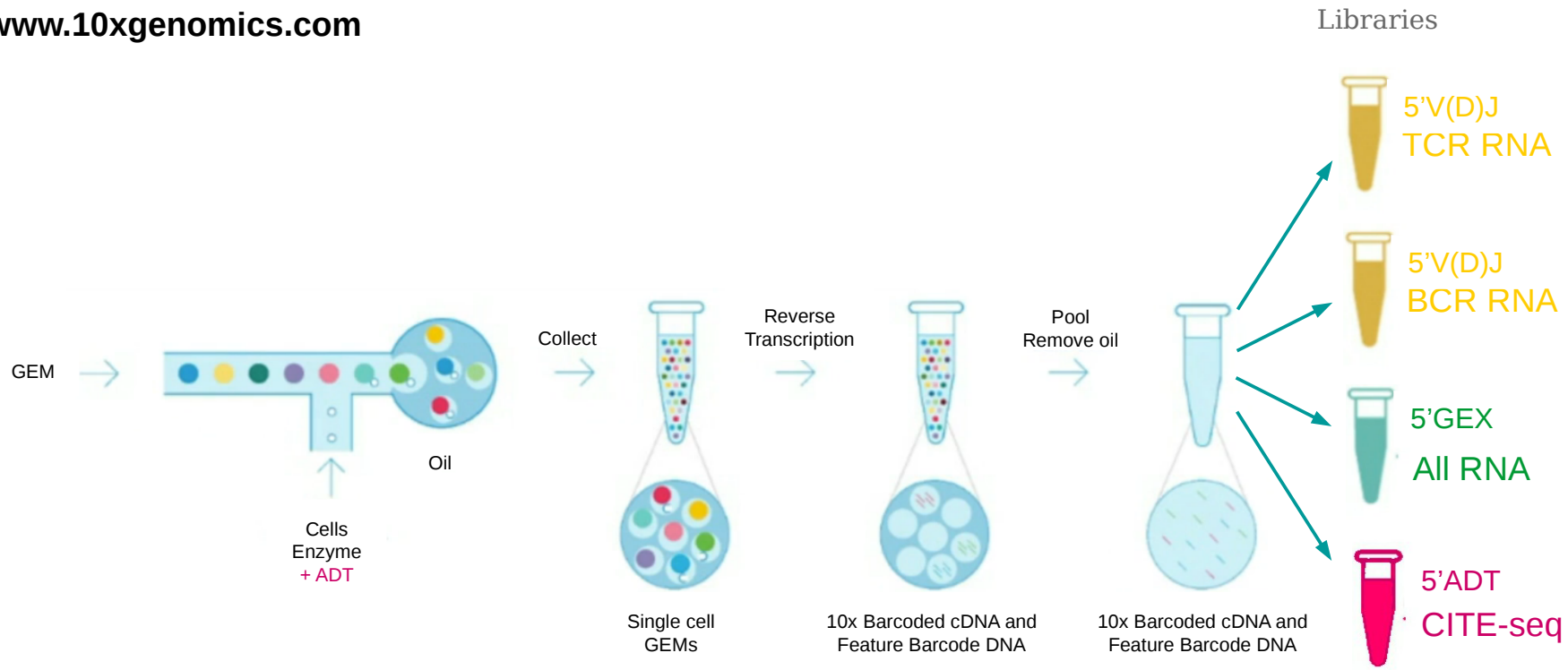


Top 10 Clonotypes (by frequencies)



GE + TCR + BCR + CITE-seq on the same cells?

www.10xgenomics.com



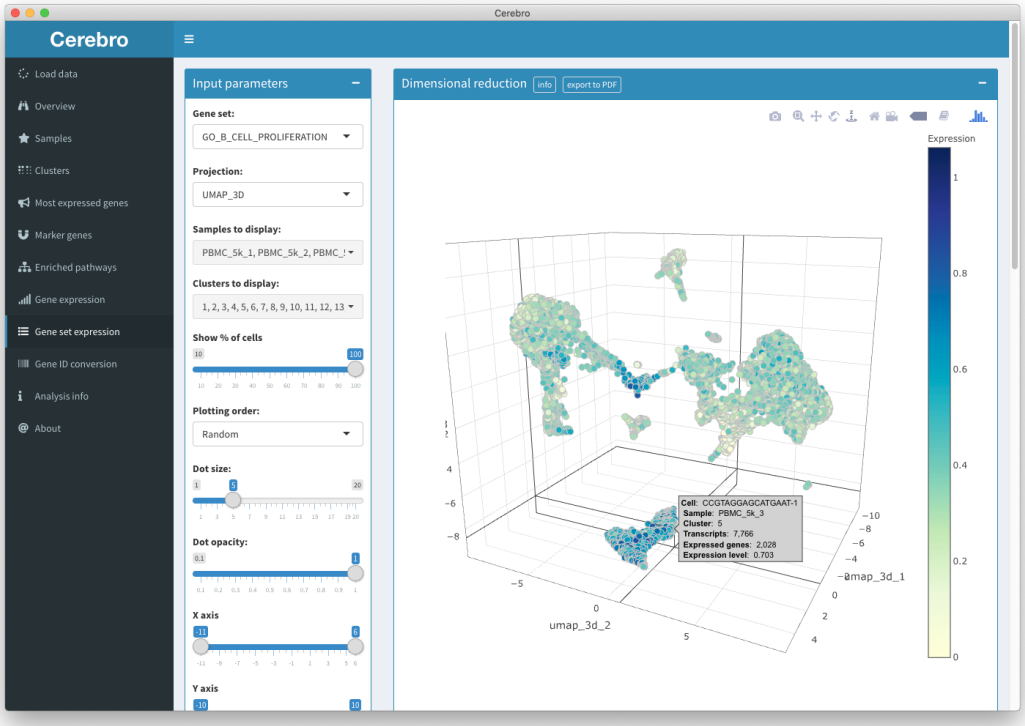
CEREBRO

CEREBRO

Results visualization tool:

https://github.com/romanhaa/Cerebro

App R Shiny



Load data

Overview

Samples

Clusters

Most expressed genes

Marker genes

Enriched pathways

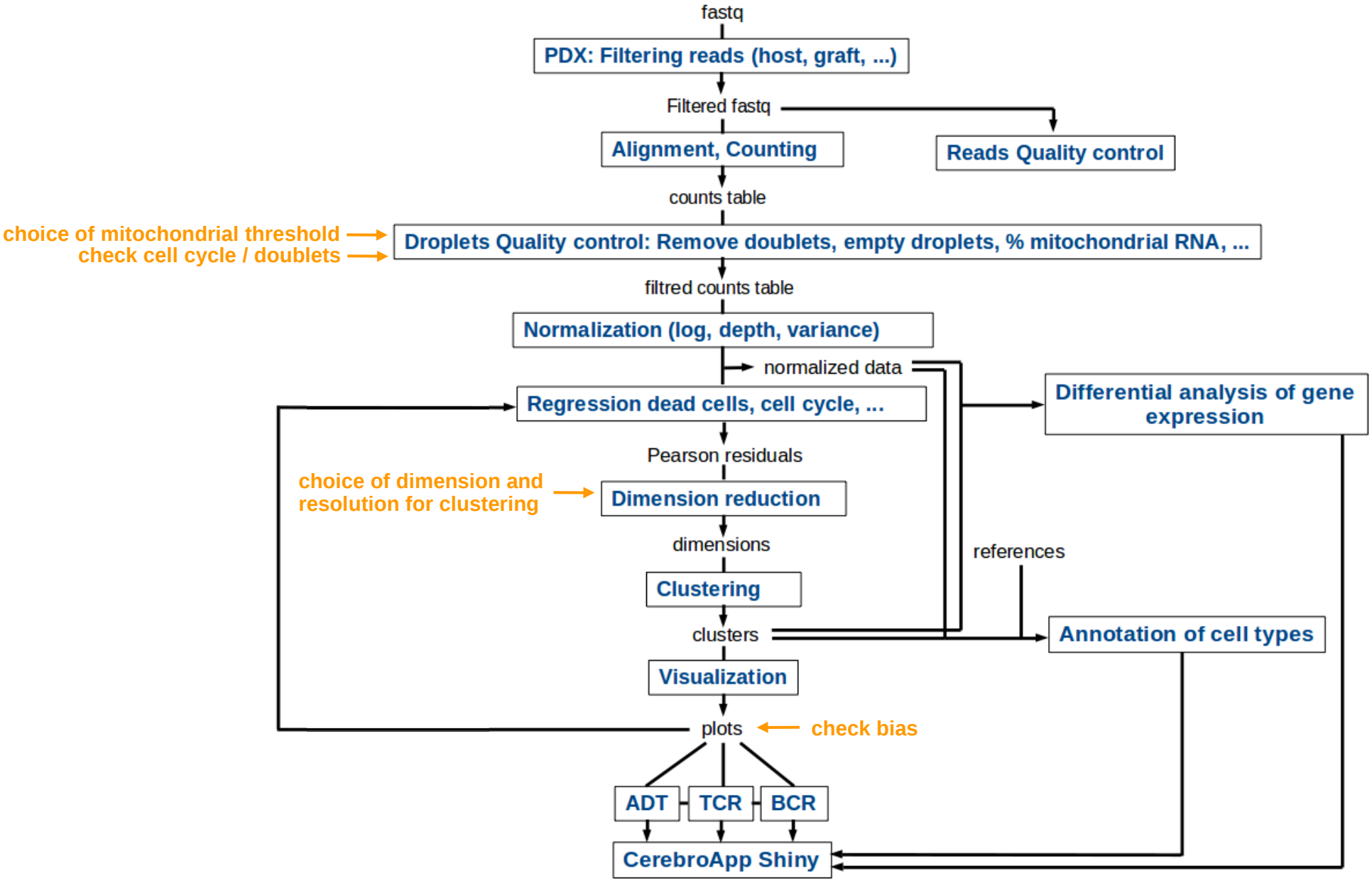
Gene expression

Gene set expression

+ add all the metadata

Our Pipeline

Our pipeline



THANK YOU FOR YOUR ATTENTION