Single Cell RNA-Seq

Carlos Torroja

Bioinformatics Unit

Single Cell RNA-Seq

- Experimental Design Considerations:
  - Replicates
  - Conditions
- Platforms: MARS-Seq, 10X, Rhapsody
- Libraries:
  - 3’ UMI
  - 5’ UMI
  - VDJ
  - ATAC-Seq
  - CITE-Seq
- Spatial Single Cell: From Imaging to Transcripts
- How the data looks like
- Data Processing: QC -> Filtering -> Normalization -> Clustering -> DR
- Data Analysis: Markers, Differential Expression -> Functional Analysis
- Cell Type Identification
- Cell-Cell Communication
- Trajectory/Pseudotime Analysis, Interactome, Regulome

Bioinformatics Unit

CNIC Bioinfo Unit

Dr. Fátima Sánchez Cabo
Dr. Fernando Martinez
Dr. Carlos Torroja
Dr. Manuel J. Gómez
Jorge de la Barrera
Dr Daniel Jiménez
Lucia Sánchez García
Juan Carlos Silla

PhD Students

Beatriz de las Mercedes (JJF Lab)
Victor Jiménez (MAP Lab)
Jon Sicilia (AH Lab)
Alvaro Serrano (AR Lab)
Inés Rivero García (MT Lab)
Diego Mañanes (DS Lab)

Embeded Bioinformatician

Jose Luis Cabrera (JLP Lab)
Marina Rosa Moreno

Master Students

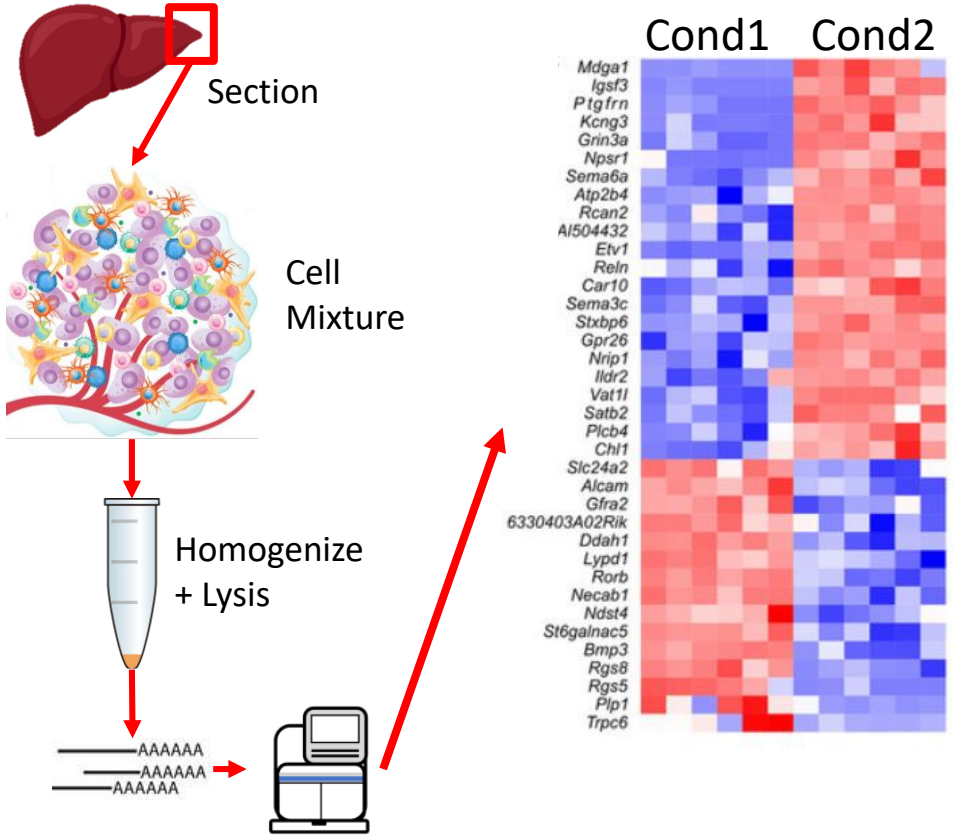
Pablo Beltrán López (UAM)
Juan Ignacio Álarez Arenas (UAM)

Link to practice data and presentation

https://www.cnic.es/nextcloud/s/beoMoMQbzR59FXb

https://drive.google.com/drive/folders/1XT6mZ2H0rqzsuPUGMpZzhZ_9Ls_gDcw6?usp=sharing

Typical Bulk RNA-Seq experiment



What is the origin of these differences?

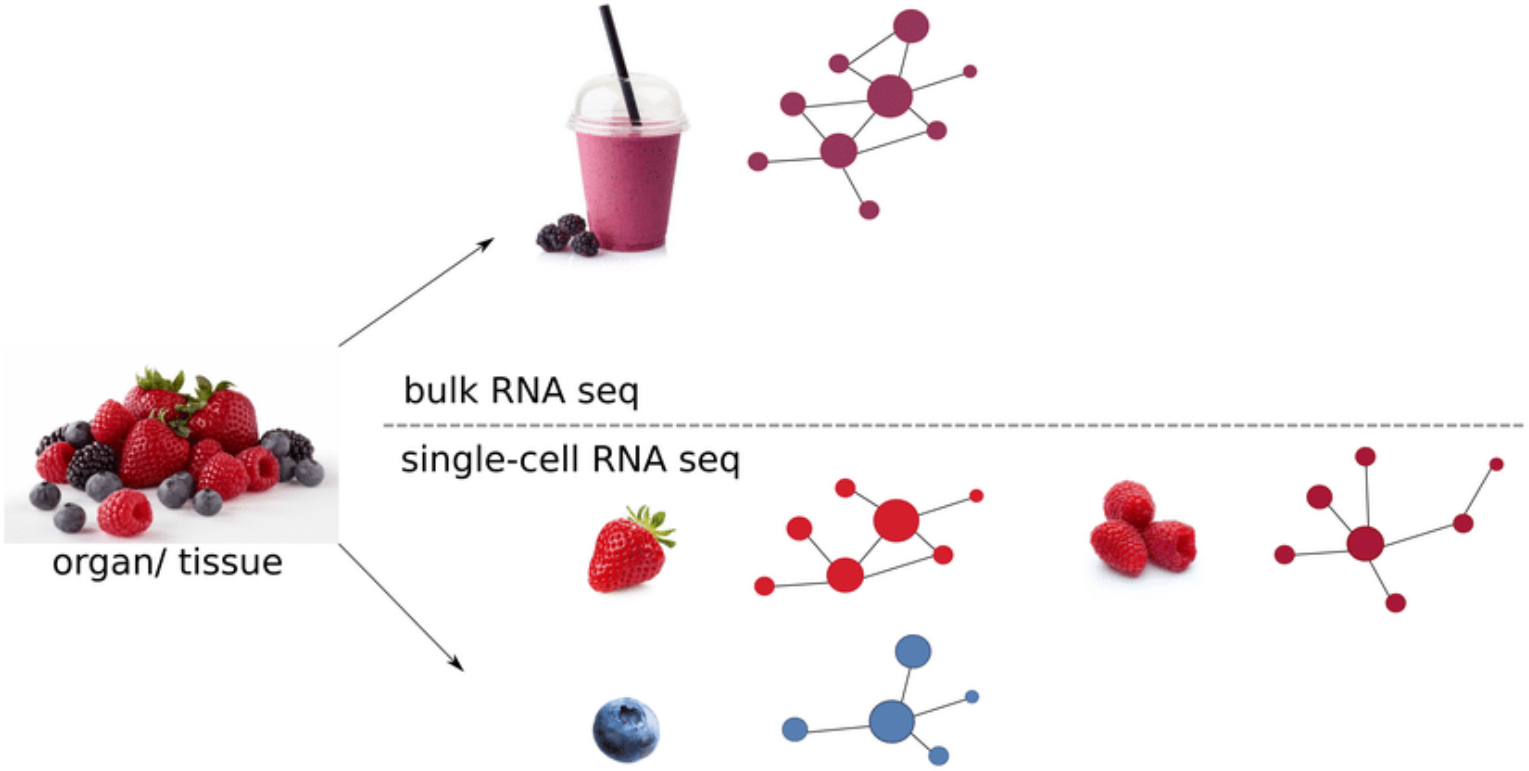
Have all cells changed their phisiological status in the same way?

Are there more cells of one type?

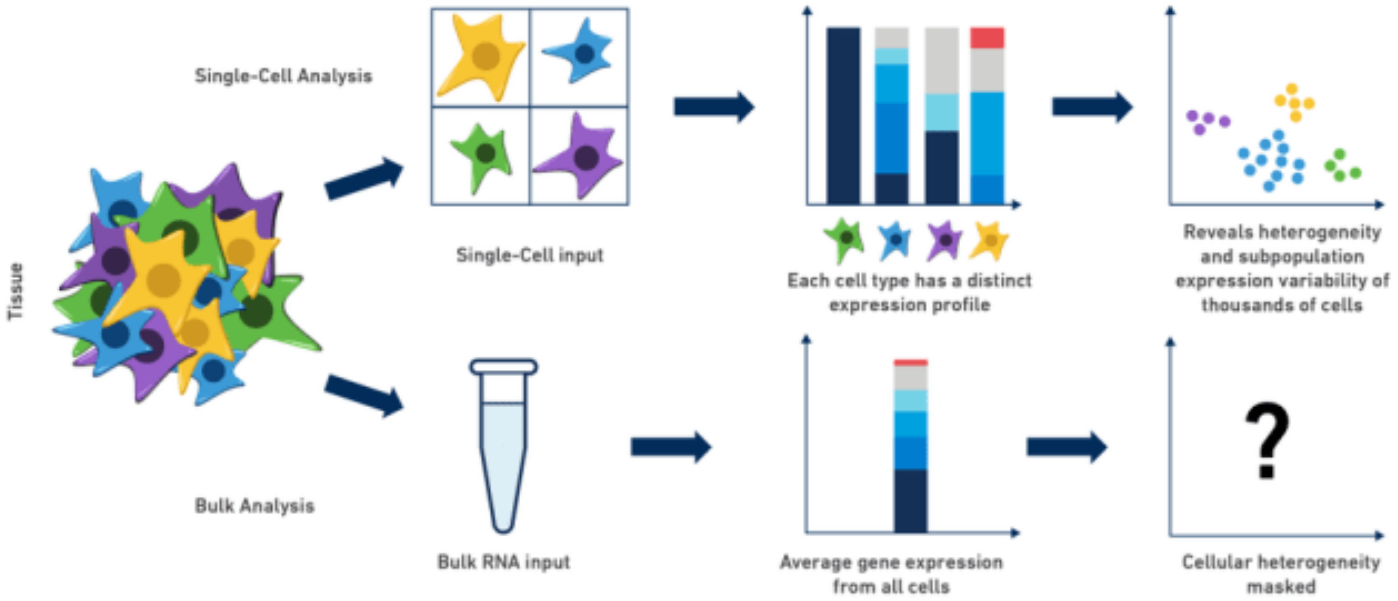
Is there a particular abundant cell that has changed?

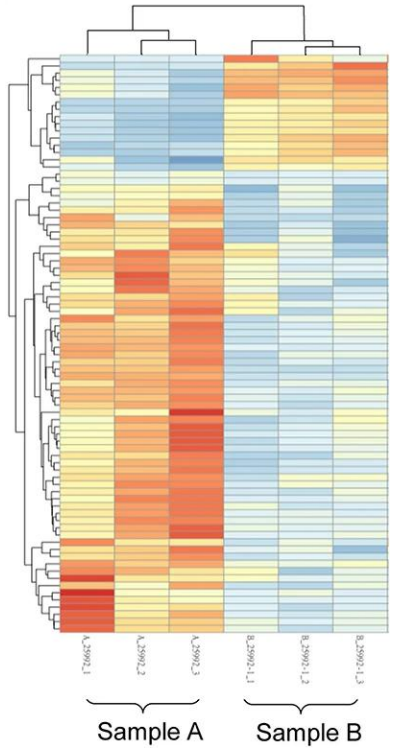
Is there a new cell type in the tissue?

From Bulk RNA-Seq to Single Cell RNA-Seq



From Bulk RNA-Seq to Single Cell RNA-Seq



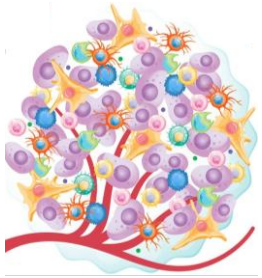


Are all changes the same in all cells?
Are changes due to changes in cell proportions?
Are changes exclusive of a particular cell type?

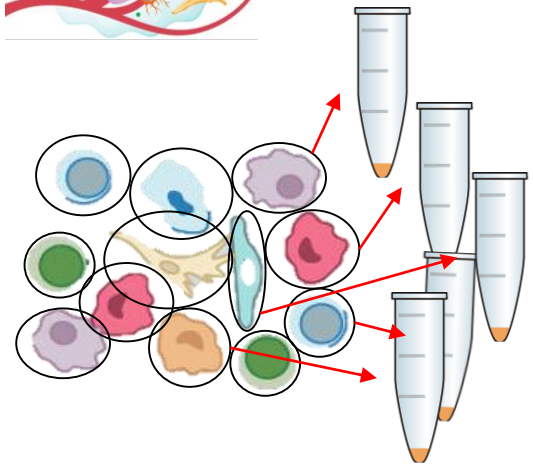
From Bulk RNA-Seq to Single Cell RNA-Seq

Singel Cell RNA-Seq experiment

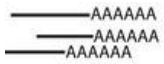
Section

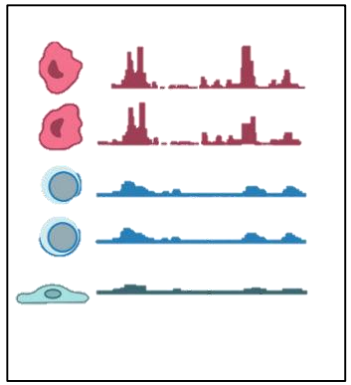


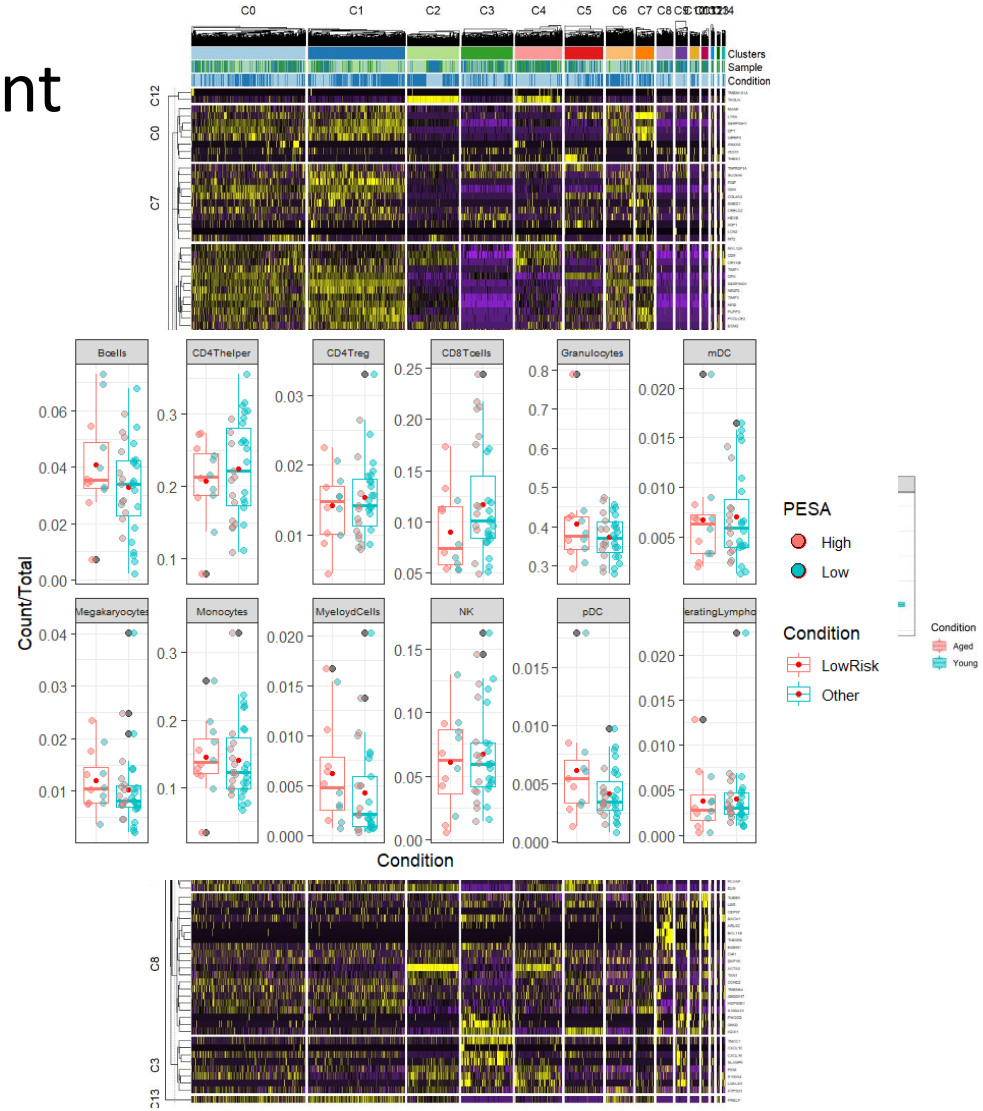
Cell
Mixture



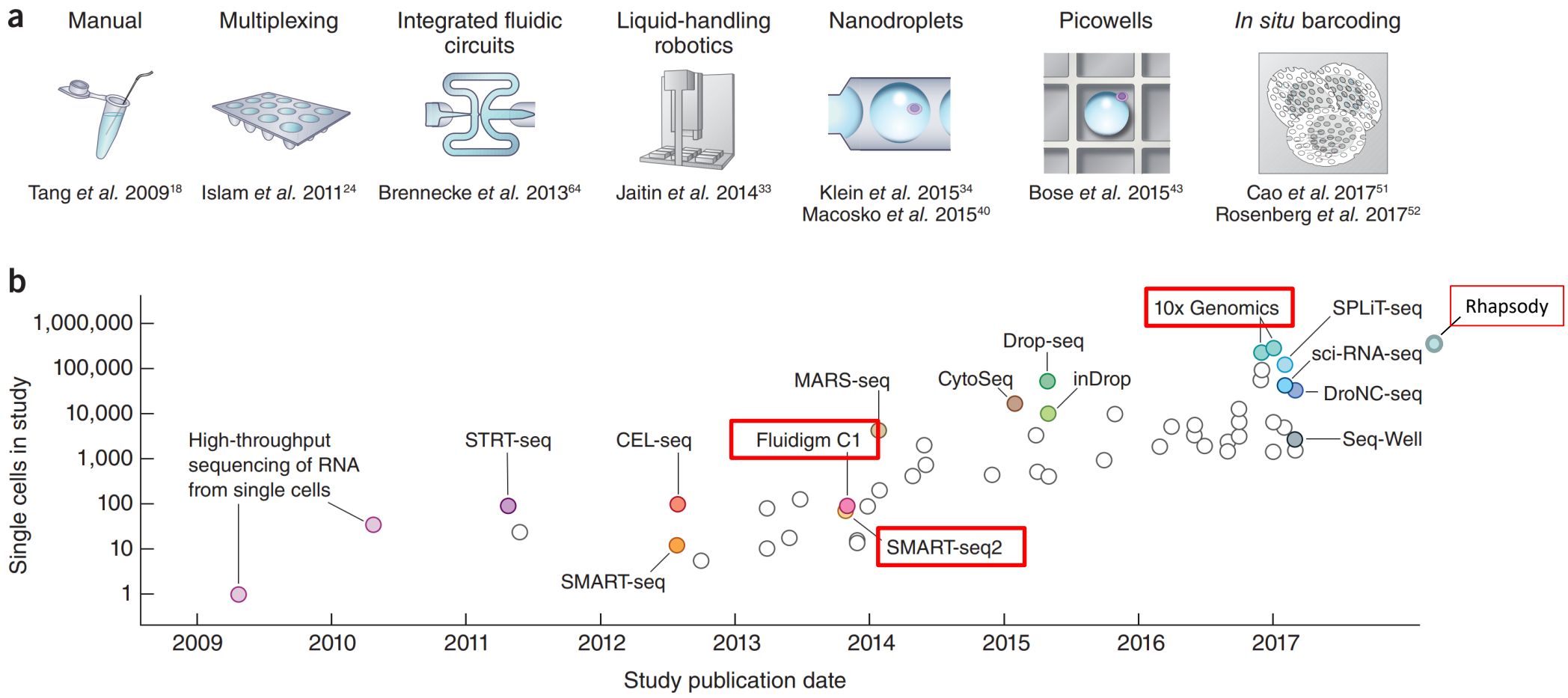
Homogenize
+ Lysis





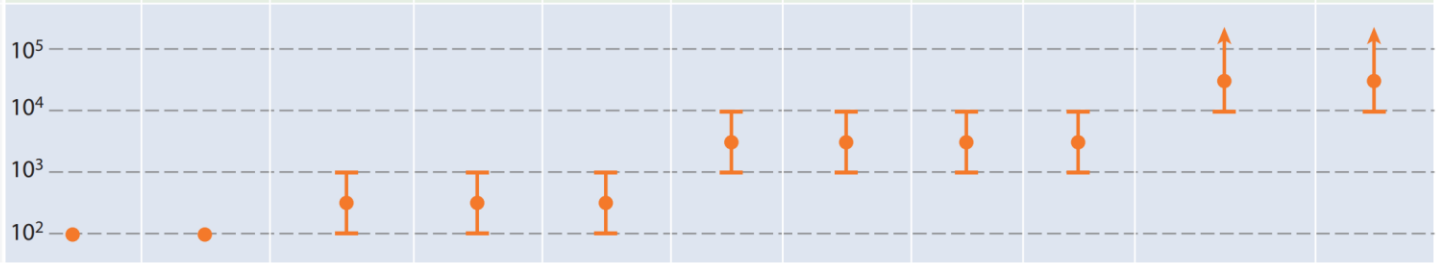


Popular single-cell RNA-seq protocols



Svensson et al. 2018. DOI:10.1038/nprot.2017.149

Popular single-cell RNA-seq protocols

	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3′	5′	3′	3′	3′	3′	3′	3′	3′	3′
Number of cells per assay											

Chen et al. 2018. DOI:10.1146/annurev-biodatasci-080917-013452

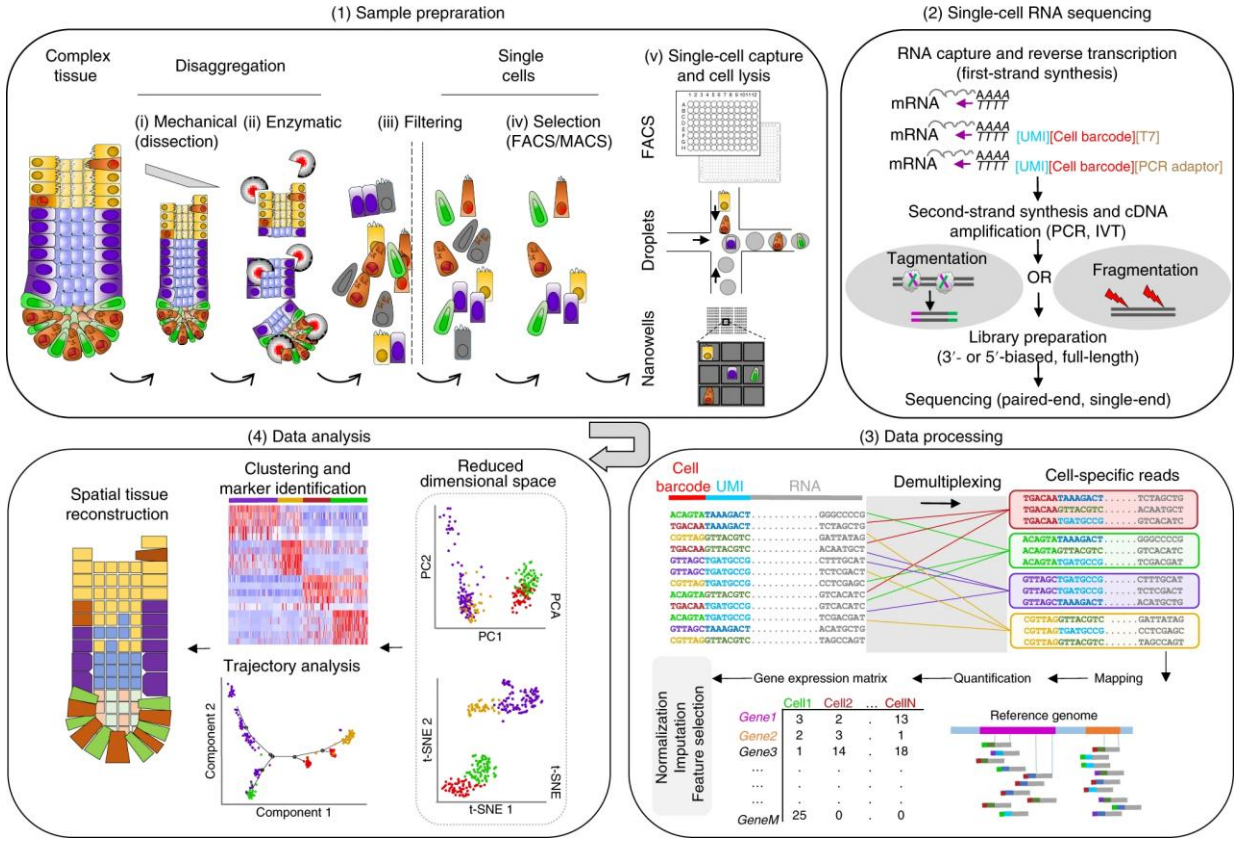
Why would you do a SC experiment?

Identify populations in a experimental model

Increase the transcriptomic knowledge of already known populations

Compare populations across conditions / development

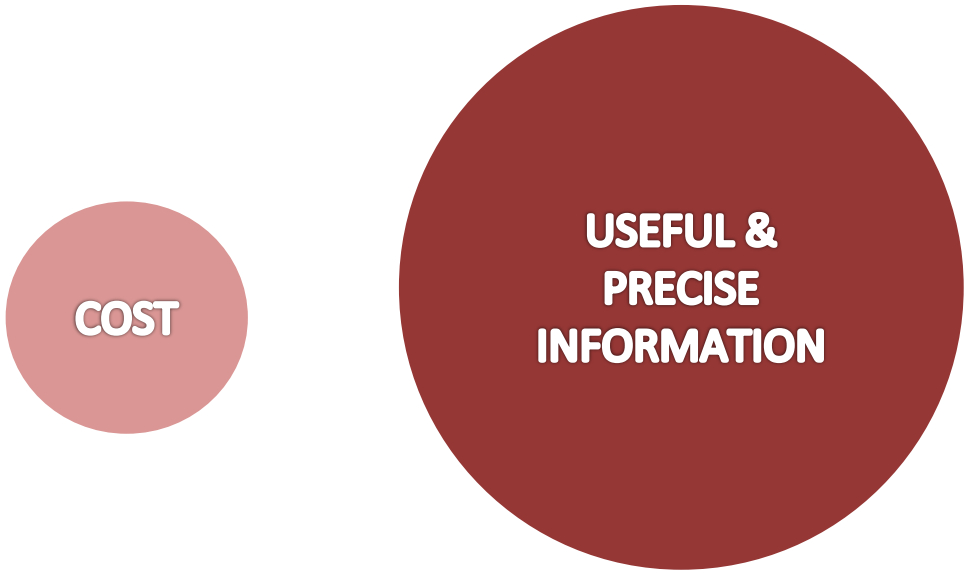
Experimental Design on Single Cell



Always use biological replicates. Not at single cell level but an individual level

Lafzi et al. 2018. https://www.nature.com/articles/s41596-018-0073-y

AIMS OF EXPERIMENTAL DESIGN



AIMS OF EXPERIMENTAL DESIGN

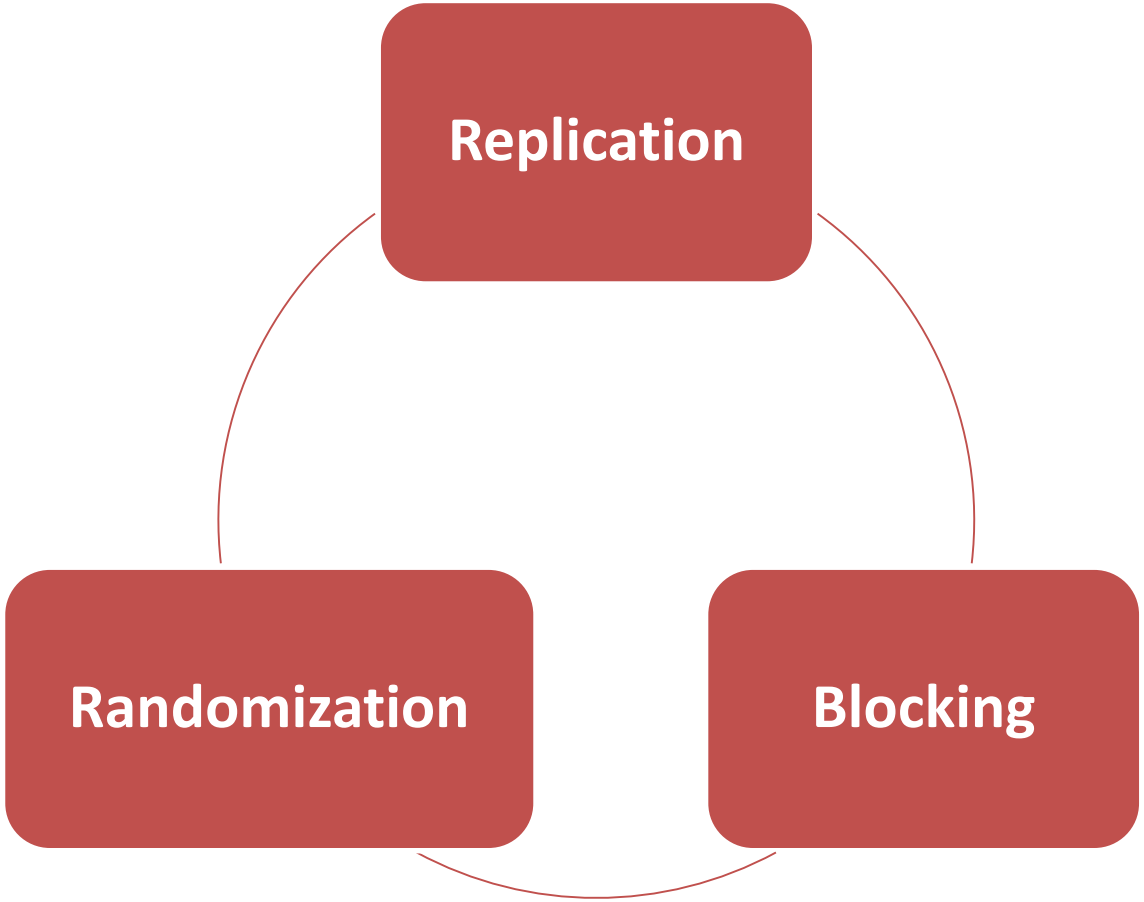
- Clear question we want to answer with the experiment:

  Data-driven vs *hypothesis-driven*

  but we know what we are looking for…

- Deep knowledge of the technology to:
  - Minimize random errors
  - Correct biases
  - Identify and control by nuisance variables

PRINCIPLES OF EXPERIMENTAL DESIGN



R.A.Fisher (1935)

REPLICATION

- As Fisher (1935a) noted, without an estimate of variability (i.e., within treatment group), there is no basis for inference (between treatment groups).
- Although we can test for differential expression between treatment groups from unreplicated data, the results of the analysis only apply to the specific subjects included in the study (i.e., the results cannot be generalized)
- Add to this, the errors that can happen during data acquisition.

TYPE OF ERRORS IN A EXPERIMENT

- Random errors
  - Not possible to calibrate
  - Minimize using repeated measurements
  - Account by nuisance variables
- Systematic errors
  - Possible to estimate and remove from the data
  - Normalization

SAMPLING AND TECHNIQUE NOISE ARE HIERARCHICAL

- **Subject sampling:** individuals are ideally drawn from a larger population to which results of the study may be generalized. (biological)
- **Cells sampling:** from each individual we recover a sample of it's cells. (biological + technical)
- **RNA sampling:** occurs during the experimental procedure when RNA is isolated from the cell. (technical)
- **Fragment sampling:** only certain fragmented RNAs that are sampled from the cells are retained for amplification. Since the sequencing reads do not represent 100% of the fragments loaded into a flow cells, this is also at play. (technical)
- **Sequencing:** At the sequencing step there is a fragment sampling step and also the sequencing process that may introduce errors

SAMPLING AND TECHNIQUE NOISE ARE HIERARCHICAL

Table 1: Replicate hierarchy in a hypothetical mouse single-cell gene expression RNA sequencing experiment

From: Replication

	Replicate type	Replicate category[a]
Animal study subjects	Colonies	B
	Strains	B
	Cohoused groups	B
	Gender	B
	Individuals	B
Sample preparation	Organs from sacrificed animals	B
	Methods for dissociating cells from tissue	T
	Dissociation runs from given tissue sample	T
	Individual cells	B
	RNA-seq library construction	T
Sequencing	Runs from the library of a given cell	T
	Reads from different transcript molecules	V[b]
	Reads with unique molecular identifier (UMI) from a given transcript molecule	T

[a]Replicates are categorized as biological (B), technical (T) or of variable type (V).

[b]Sequence reads serve diverse purposes depending on the application and how reads are used in analysis.

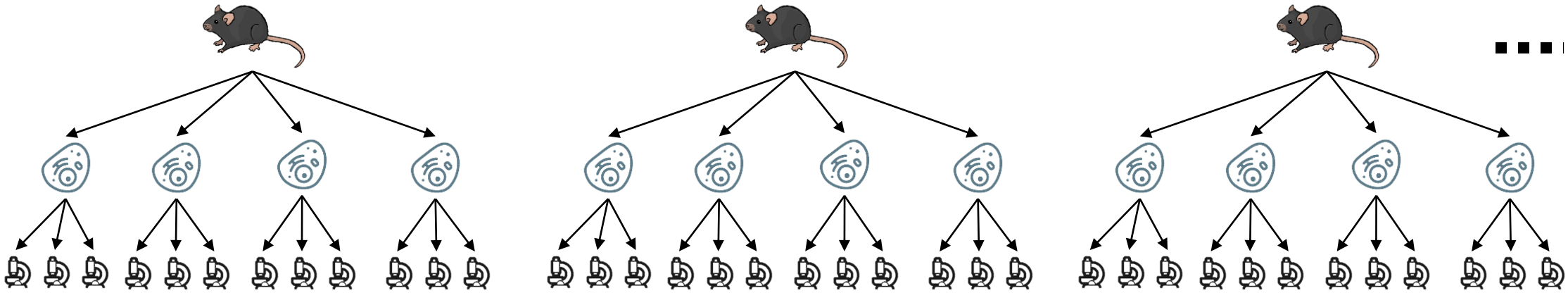
REPLICATION HELPS TO CONTROL NOISE AND ERRORS

- Technical replicates
  - Minimize random errors through averaging
  - Test technology
- Biological replicates
  - Draw conclusions about the whole population, not about the individual
  - Control variability at different experimental steps

REPLICATION

Suppose you perform an experiment in which you can measure the expression of a gene in a single cell and **you have money to perform 48 measurements**.

Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014). https://doi.org/10.1038/nmeth.3091



Which kind of replication is going on at every step?

Which sources of variation, error or noise can you think off?

NUISANCE VARIABLES IN NGS

Technical:

- RNA extraction
- Library preparation
- Flow-cell
- Lane
- Barcode
- Lab/technician

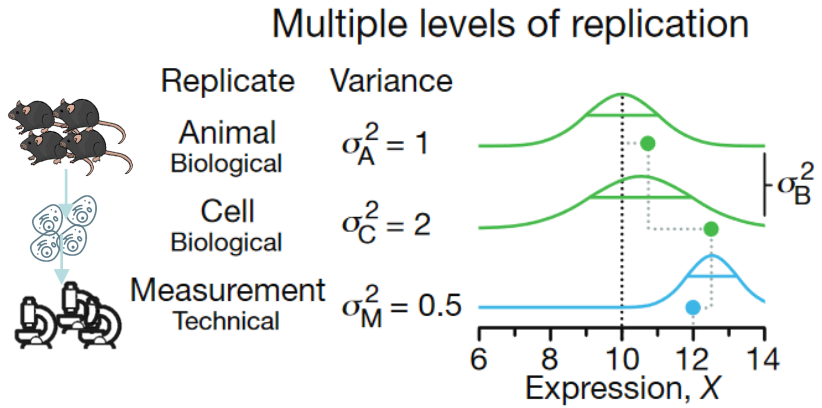
Biological:

- Sex/gender
- Litter/family
- Age
- …

Systematic Bias and Random Errors (Noise)

Experimental Design

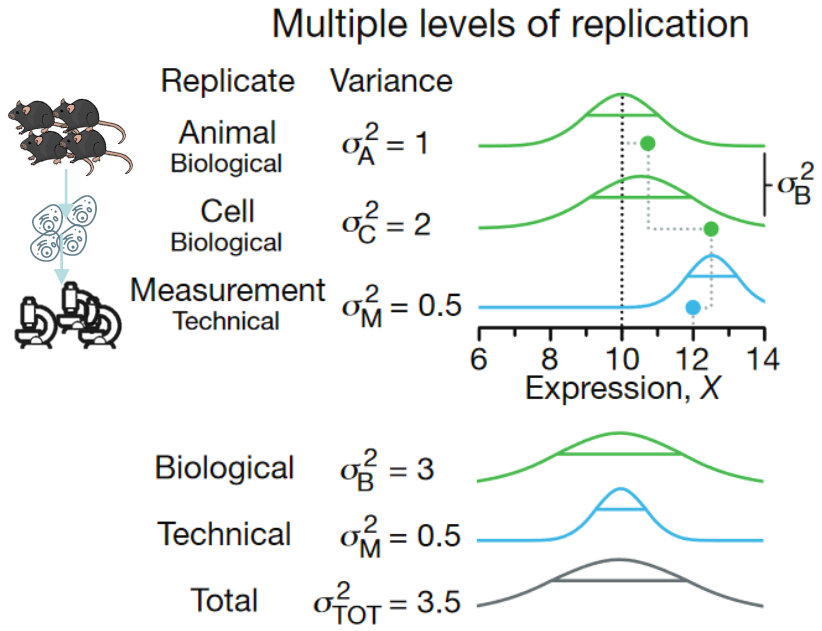
REPLICATION



Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014).
https://doi.org/10.1038/nmeth.3091

Experimental Design

REPLICATION

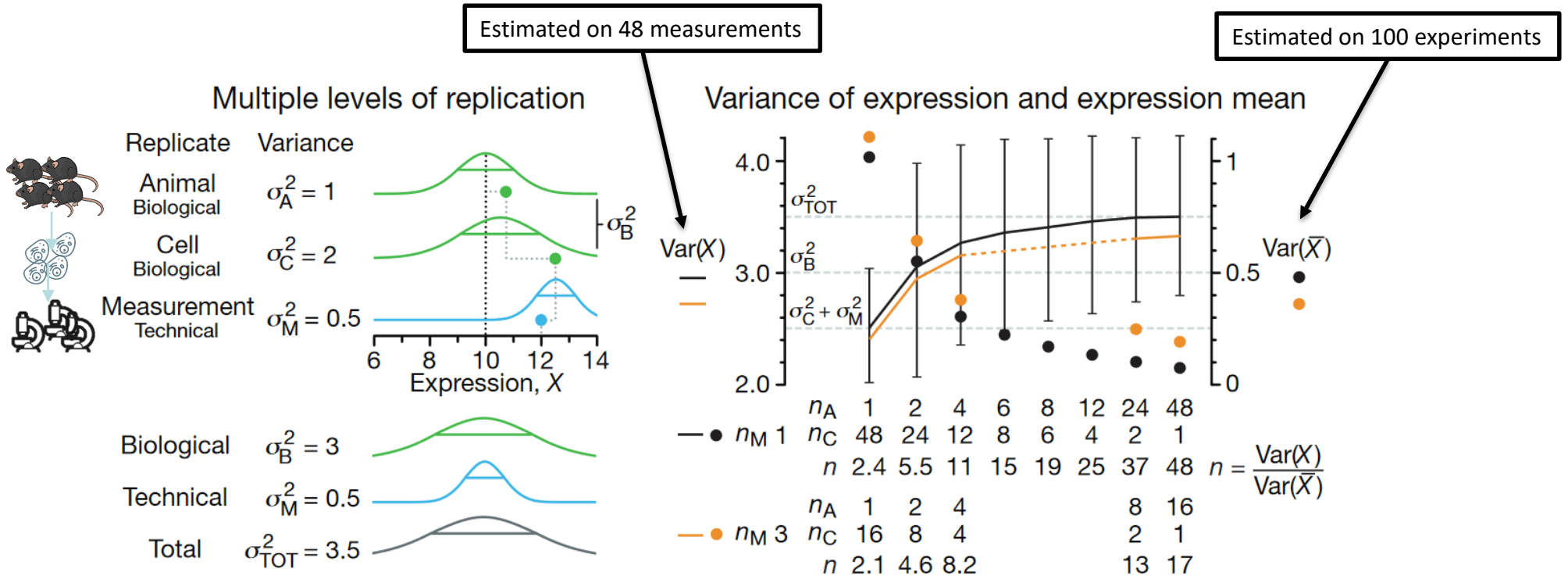


$$Var(x) = \sigma_A + \sigma_C + \sigma_M$$

Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014).
https://doi.org/10.1038/nmeth.3091

Experimental Design

REPLICATION

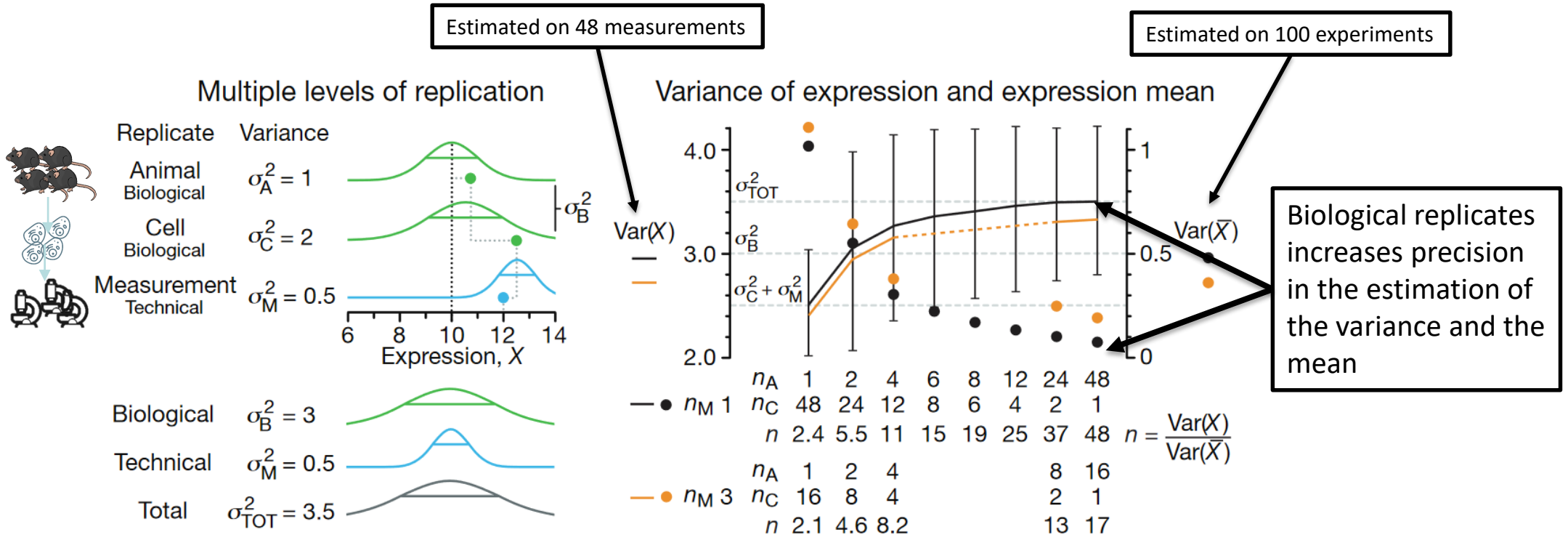


$$Var(x) = \sigma_A + \sigma_C + \sigma_M \qquad Var(\bar{X}) = \sigma_A/n_A + \sigma_C/(n_A+n_C) + \sigma_M/(n_A+n_C+n_M)$$

Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014).
https://doi.org/10.1038/nmeth.3091

Experimental Design

REPLICATION

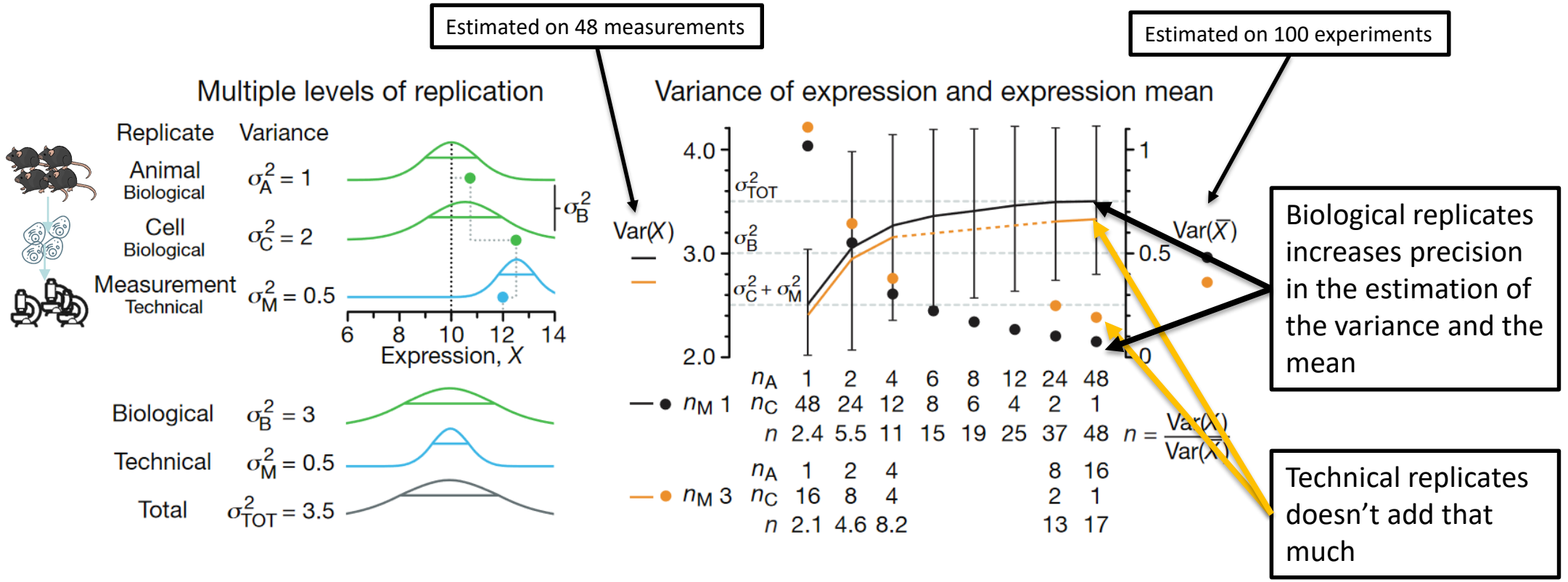


$$Var(x) = \sigma_A + \sigma_C + \sigma_M \qquad Var(\bar{X}) = \sigma_A/n_A + \sigma_C/(n_A+n_C) + \sigma_M/(n_A+n_C+n_M)$$

Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014).
https://doi.org/10.1038/nmeth.3091

Experimental Design

REPLICATION

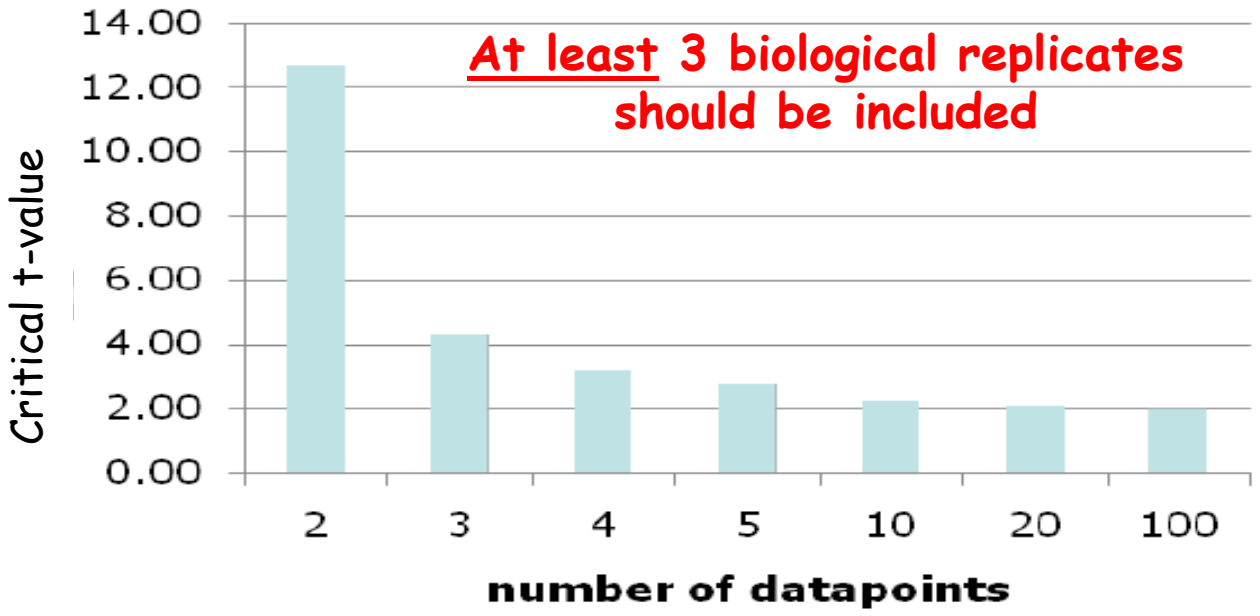


$$Var(x) = \sigma_A + \sigma_C + \sigma_M \qquad Var(\bar{X}) = \sigma_A/n_A + \sigma_C/(n_A+n_C) + \sigma_M/(n_A+n_C+n_M)$$

Blainey, P., Krzywinski, M. & Altman, N. Replication. *Nat Methods* **11**, 879–880 (2014).

REPLICATION

At least three biological replicates for omics technologies

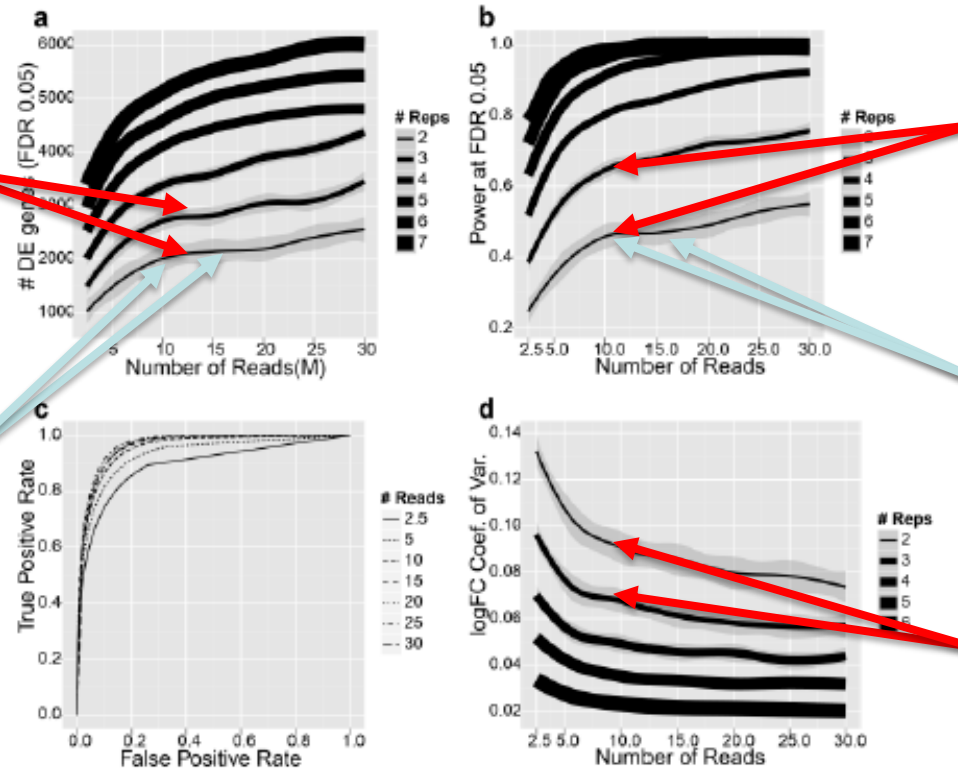


REPILCATES OR DEPTH?

- RNA-Seq experiment from MCF7 tumor cell line (mammary gland) with 7 replicates per condition (17b Estradiol and Control).
- Sequenced up to 30M reads.

There is an increase of **35%** in the DE genes (2011 -> 2709) when increasing from 2 to 3 replicates of 10M (20M -> 30M) reads.

There is only an increase of **6%** . (2011 -> 2139) in the DE genes when increasing by 50% the reads: from 2 replicates of 10M (20M) to 15M (30M) reads.



Power increases also **41%** (0.46 -> 0.65) when increasing from 2 to 3 replicates of 10M (20M -> 30M) reads.

Power increase only **19%** (0.46 -> 0.55) when increasing by 50% the reads: from 2 replicates of 10M (20M) to 15M (30M) reads.

The logFC estimation precission increases with the replicates and not so much with the sequencing depth. (Even for low expressed genes.

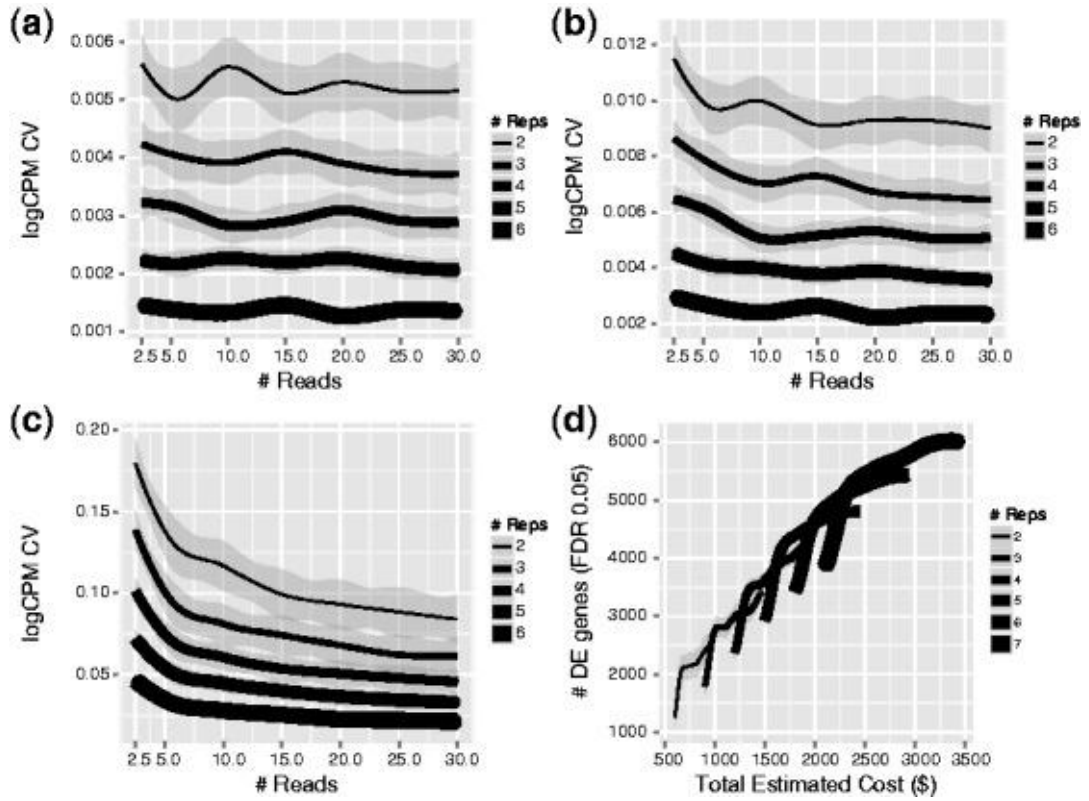
RNA-Seq differential expression studies: more sequence or more replication?
Yuwen Liu, Jie Zhou and Kevin P. White

REPILCATES OR DEPTH?

The gene expression estimation precision increases more with the replicates regardless of the level of expression.

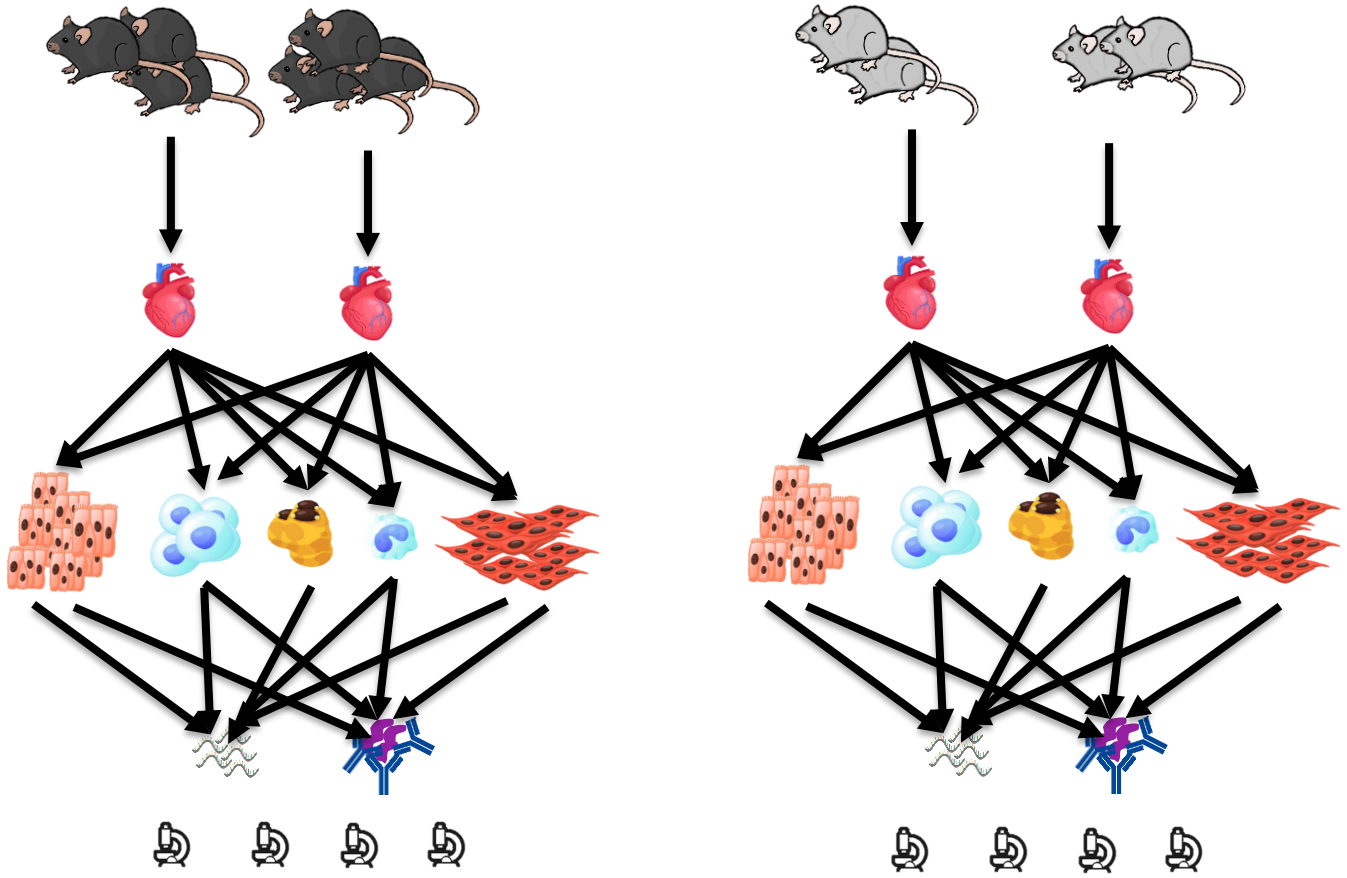
High Exp Genes

Mid Exp Genes



Low Exp Genes

Experimental Design on Single Cell



Treatment / Genotype / Stage

Tissues / Condition

Cell Types / Condition

Experimental Design on Single Cell

But how many cells should I sequence?

SCOPIT

https://alexdavisscs.shinyapps.io/scs_power_multinomial/

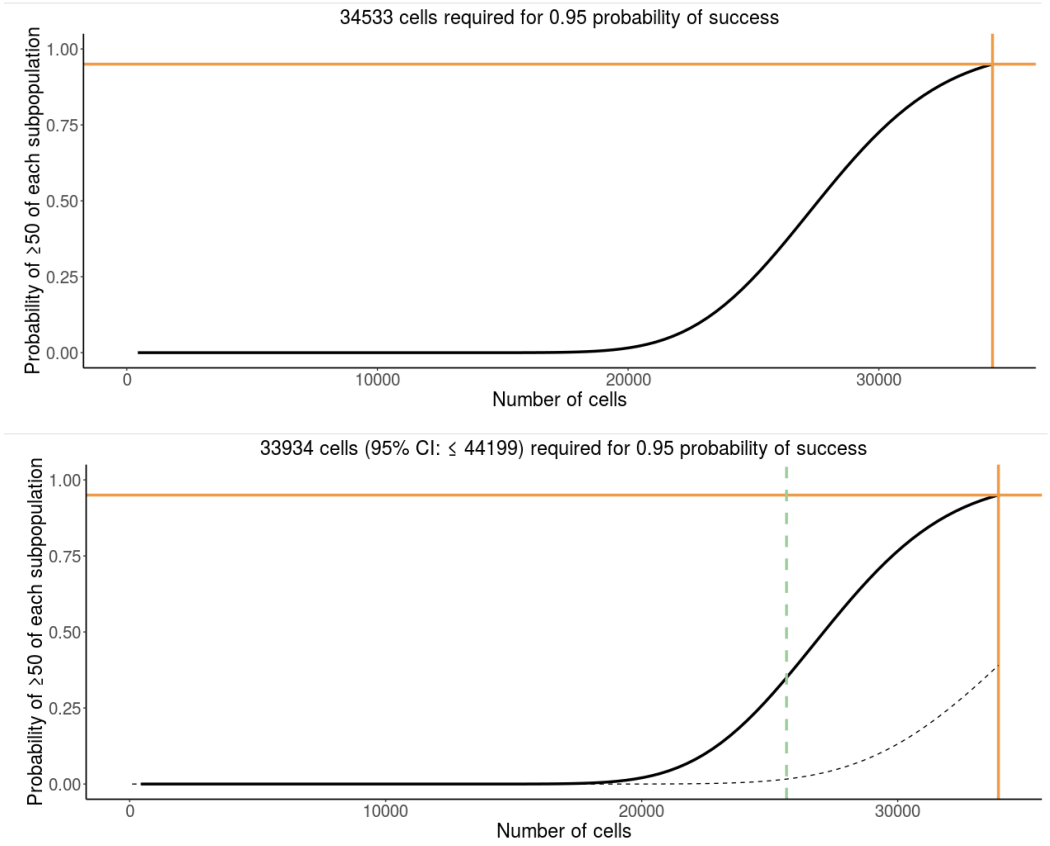
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3167-9

Experimental Design on Single Cell

Example:

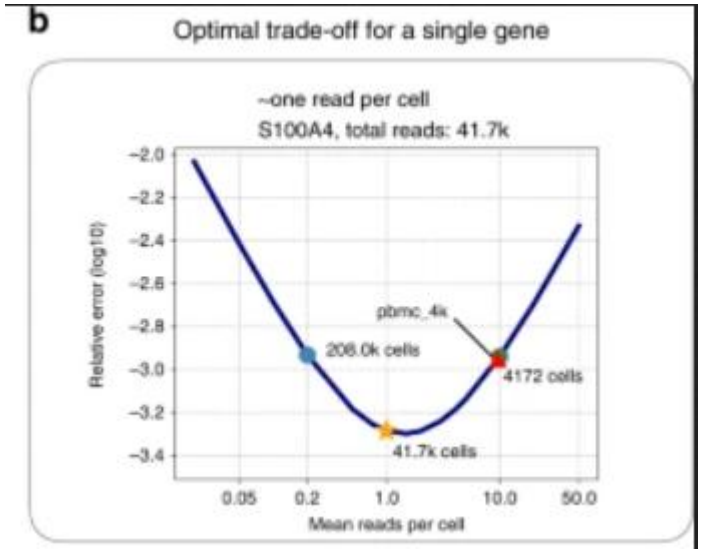
We have got an experiment with 25000 cells and 17 cell types:

Cluster	Cells	Freq
C0	5363	0.209019
C1	3327	0.129667
C2	3120	0.1216
C3	2662	0.103749
C4	2415	0.094123
C5	2393	0.093265
C6	1885	0.073466
C7	1177	0.045873
C8	1063	0.04143
C9	747	0.029114
C10	662	0.025801
C11	333	0.012978
C12	199	0.007756
C13	114	0.004443
C14	93	0.003625
C15	58	0.002261
C16	47	0.001832
Total	25658	1



Experimental Design on Single Cell

How many reads?



25000 cells * 12000 genes = 300M reads

Zhang, M.J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat Commun* **11,** 774 (2020). https://doi.org/10.1038/s41467-020-14482-y

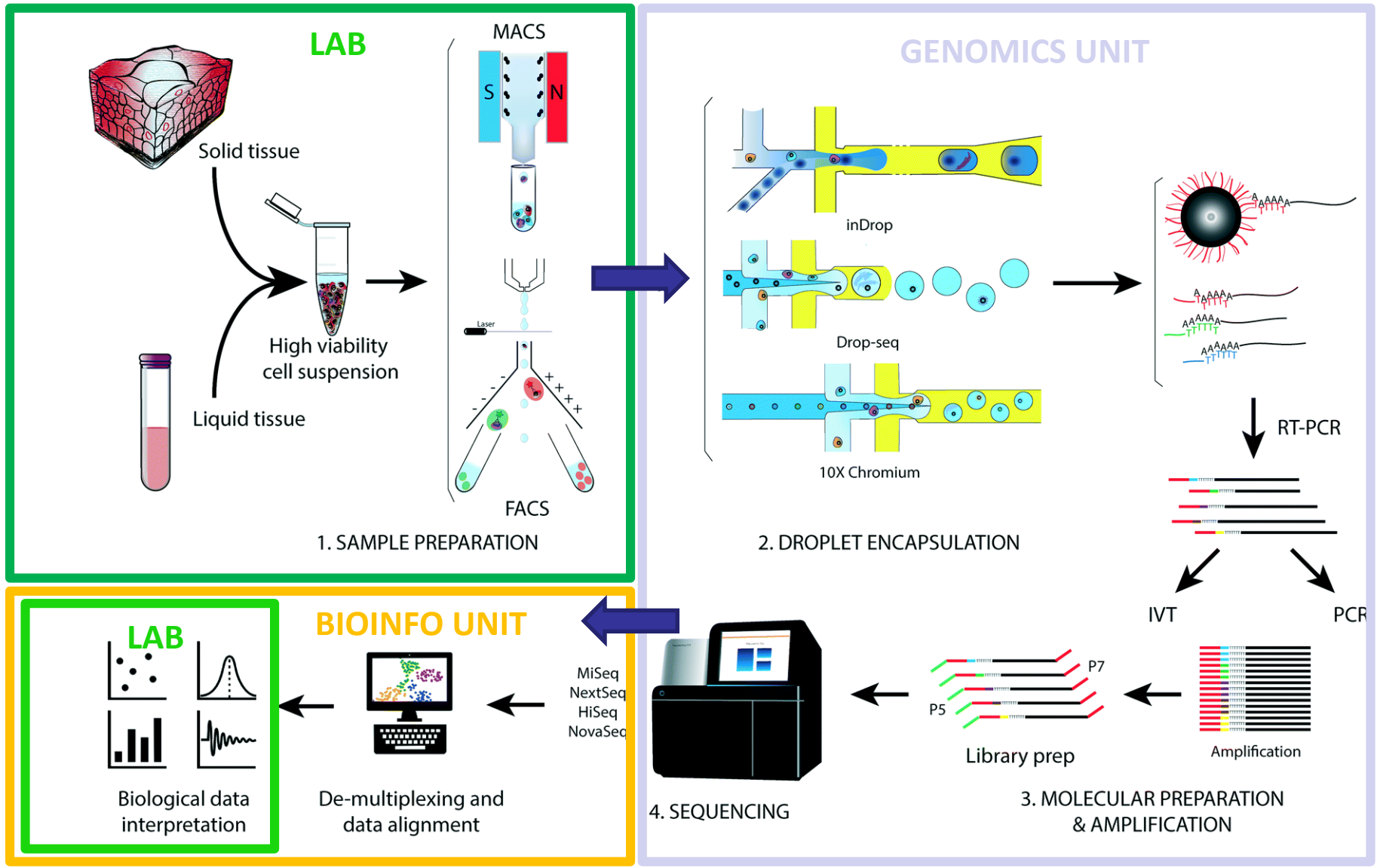
In this paper, we introduced the sequencing budget allocation problem to provide a precise answer to this question; **given a fixed budget, sequencing as many cells as possible at approximately one read per cell per gene is optimal**, both theoretically and experimentally.

Take Home Message

Always think about the experimental design before performing the experiment

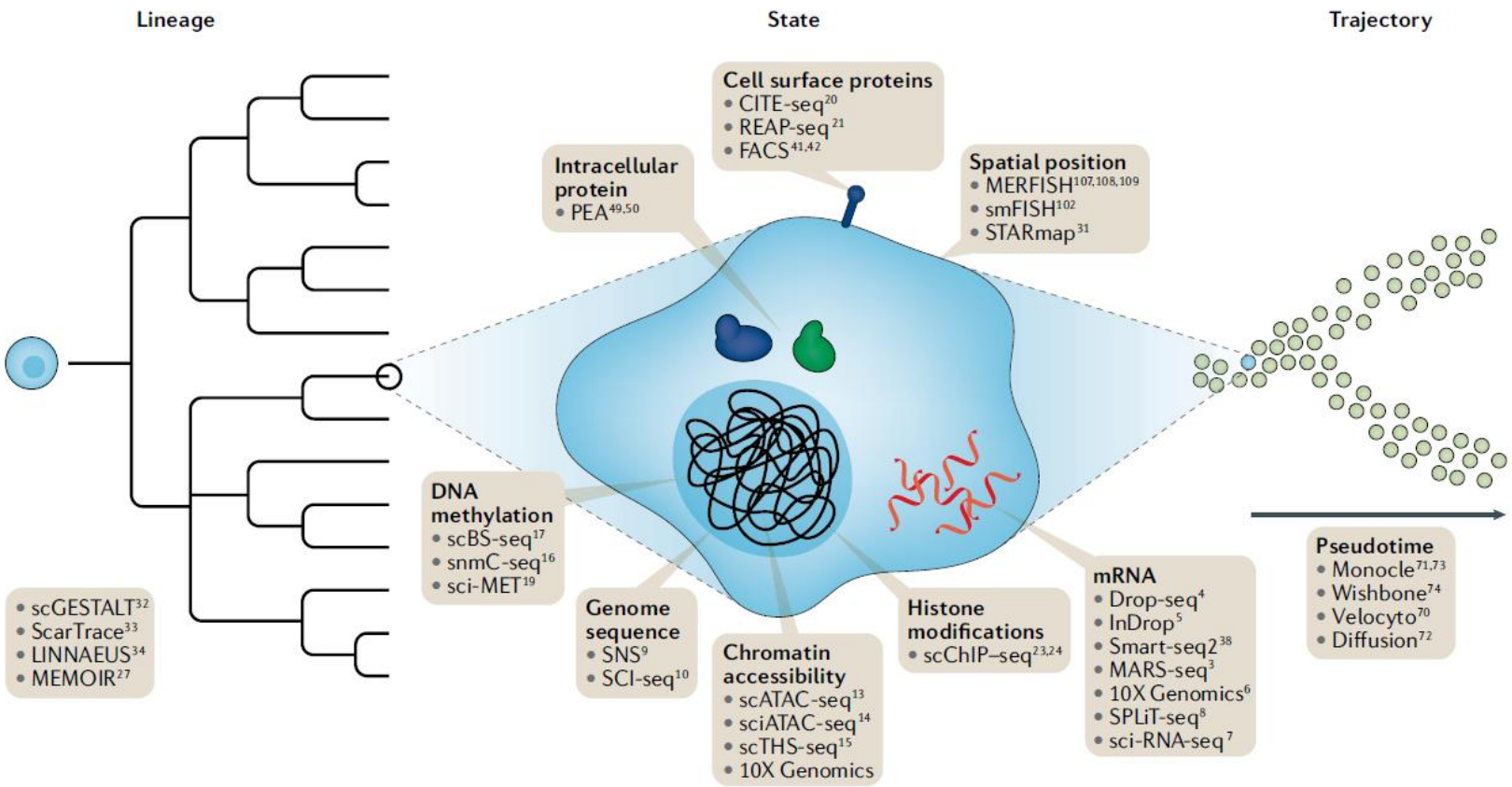
- Try to persuade the researcher to use replicates. (better replicates than more cells or sequencing depth)
- Try to reduce pooling as much as possible.
- Do not ever consider technical replicates (you will be lying to yourself).
- Estimate the minimum number of cells base on prior knowledge.

Single Cell Experiment workflow



Single Cell Feature Libraries

Not just Transcriptome



Stuart and Satija 2019. https://www.nature.com/articles/s41576-019-0093-7

Starting Material

Cell dissociation from solid tissues:

- Each tissue on each species has its own technique
- Protease digestion is required to destroy extracellular matrix (Pronase, Acutase, …)
- Needs to be adjusted to avoid cell damage
- The procedure is quite dramatic.
- In cases where the cell integrity is compromised when dissociating tissues, single nuclei can be obtained.
- Nuclei are also used from fixed or frozen tissues.

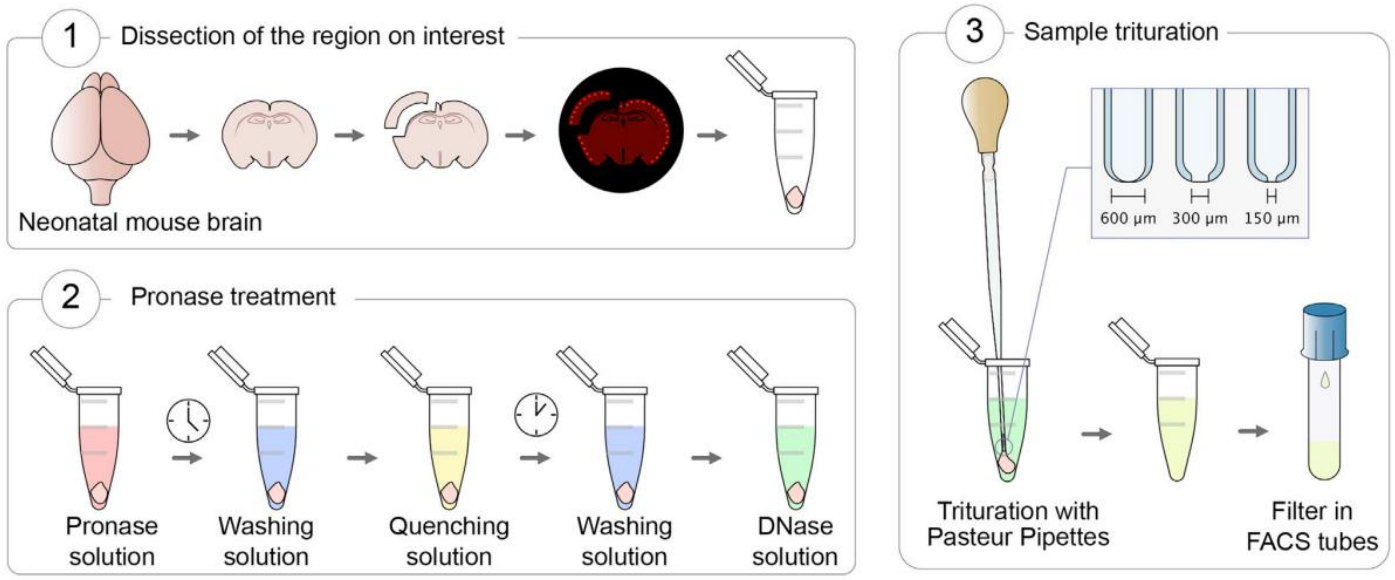
Blood:

- Straight forward after erythrocyte digestion.

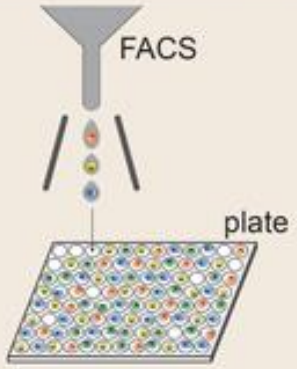
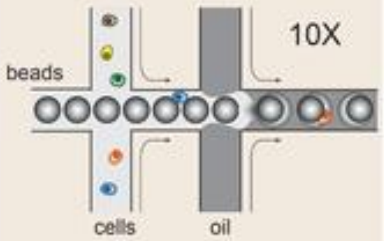
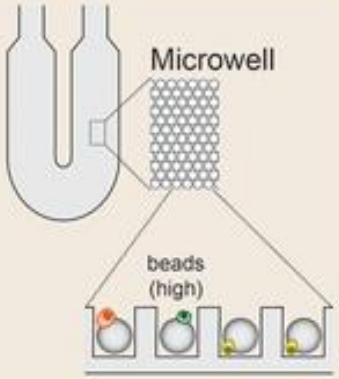
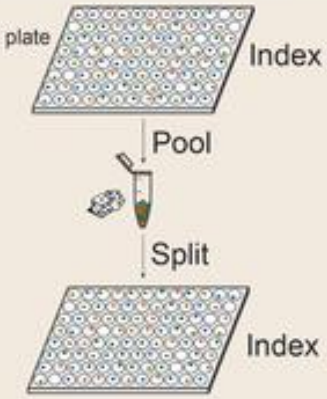
Cell Sorting by FACS:

- Cells are a bit shocked, be careful

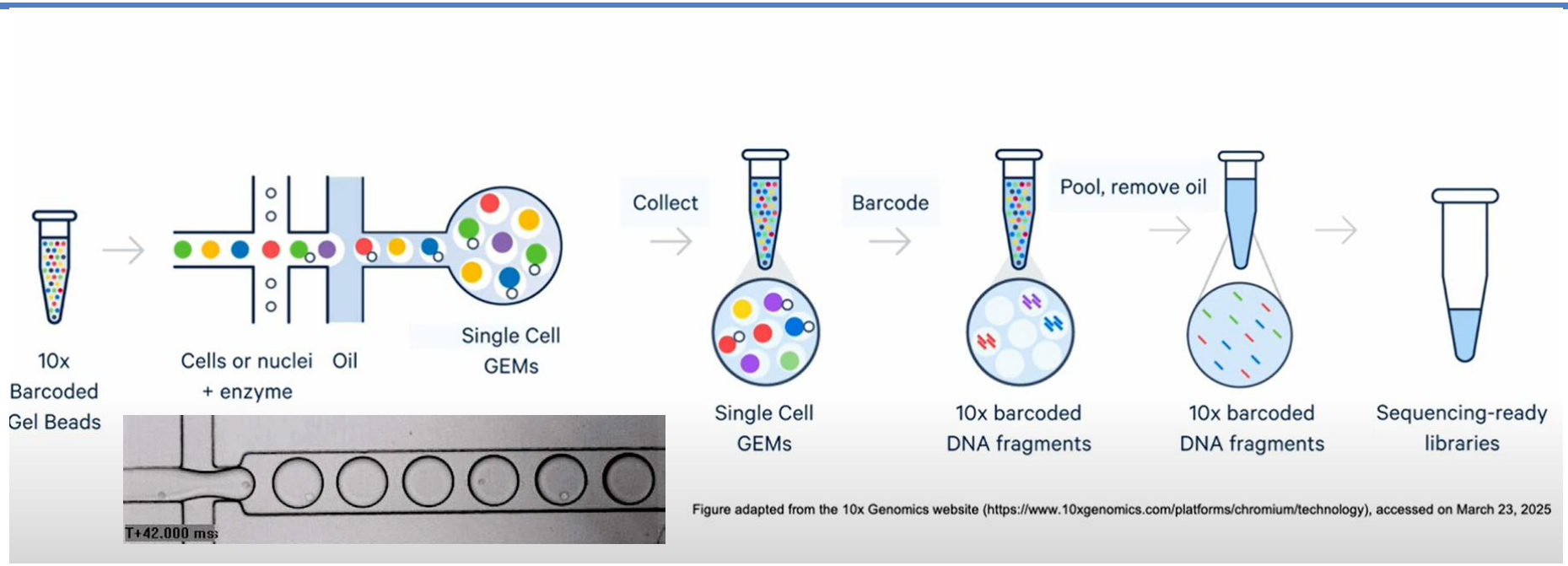
Cells Enriched by primed beads

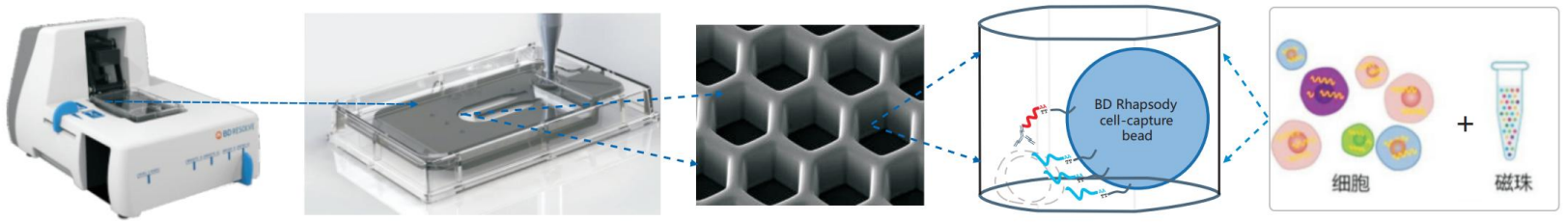


How do I Isolate The Cell?

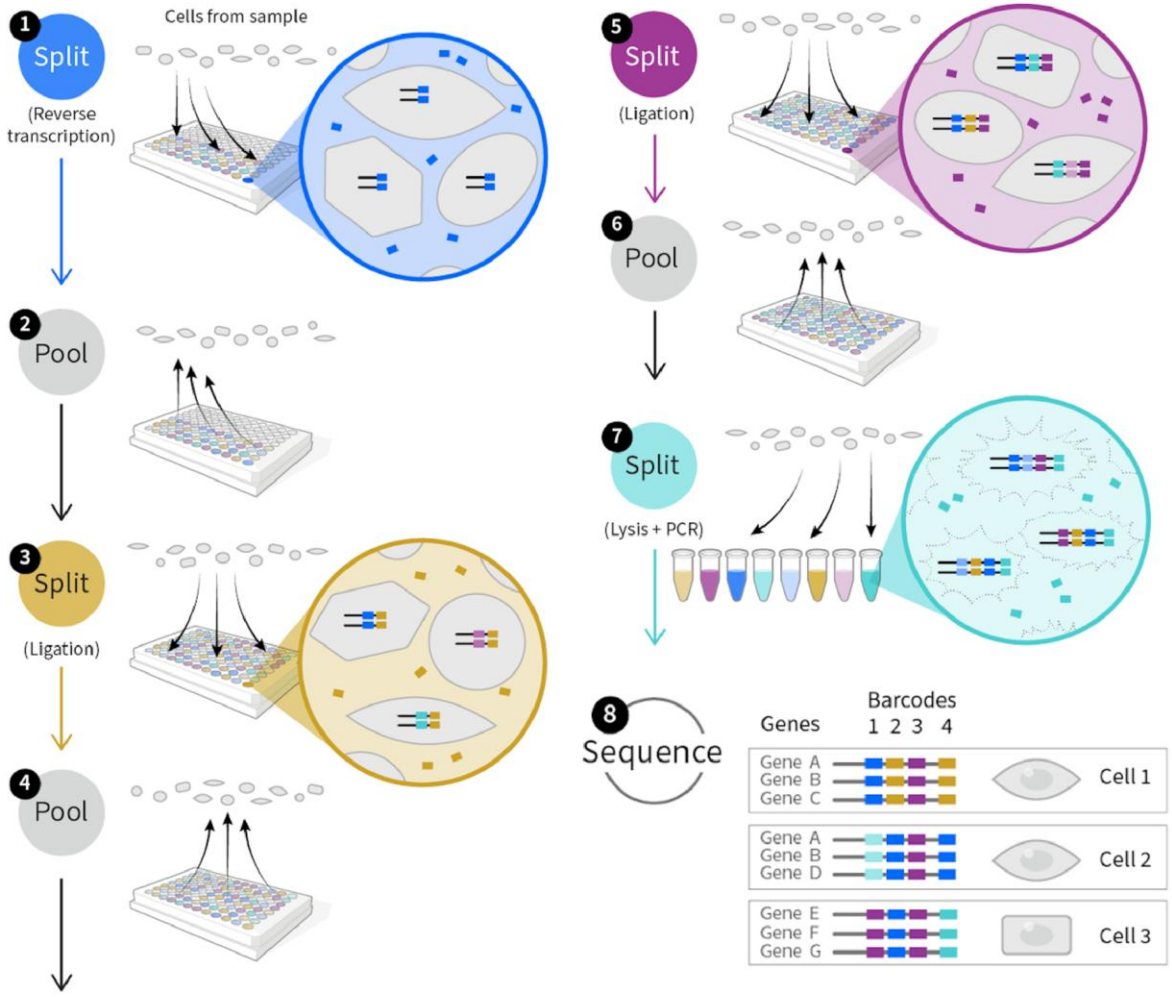
	SMART-seq	DropSeq/10X	BD Rhapsody	sci-RNA-seq
FACS-step	sc sorting into plate	bulk-sorting	bulk-sorting	sc sorting into plate
Principle	isolated cells in wells	microfluidics	microwell	split-pool indexing
Technology				

10X Chromium and Rhapsody Scheme

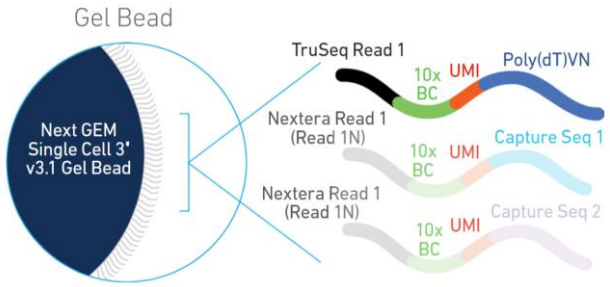




Split-Seq


1
Split
(Reverse transcription)
Cells from sample
2
Pool
3
Split
(Ligation)
4
Pool
5
Split
(Ligation)
6
Pool
7
Split
(Lysis + PCR)
8
Sequence
Genes
Barcodes
1 2 3 4
Gene A
Gene B
Gene C
Cell 1
Gene A
Gene B
Gene D
Cell 2
Gene E
Gene F
Gene G
Cell 3

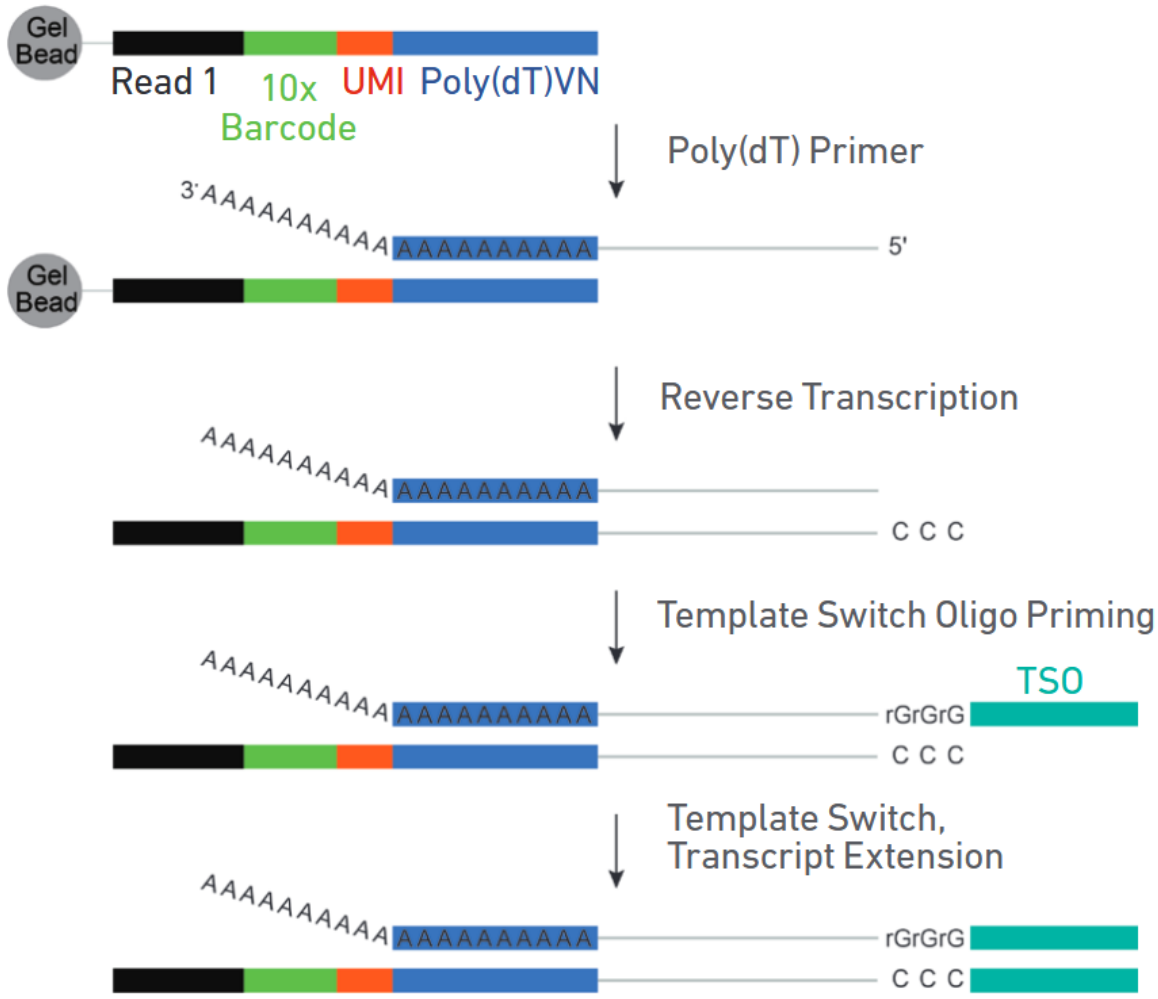
Library Preparation



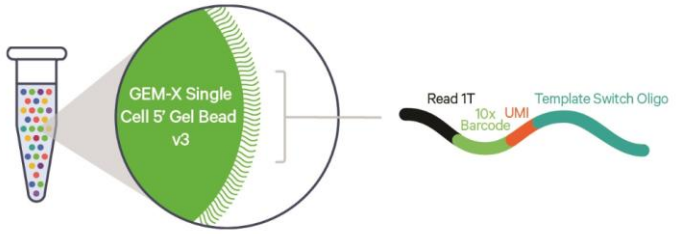
Gene Expression Library: 3’ Library

Within the drop:

- Cells lysate as soon the drop is formed
- Released cytoplasmic mRNA is captured by polyT sequences attached to the bead.
- A cDNA is produced by the RT-Pol using the TSO oligo to complete a full cDNA.
- All attached adaptors of a bead carries the same Cell Barcode. Each bead has a different Cell Barcode.
- All attached adaptors of a bead carries a different UMI (Unique Molecular Identifier).



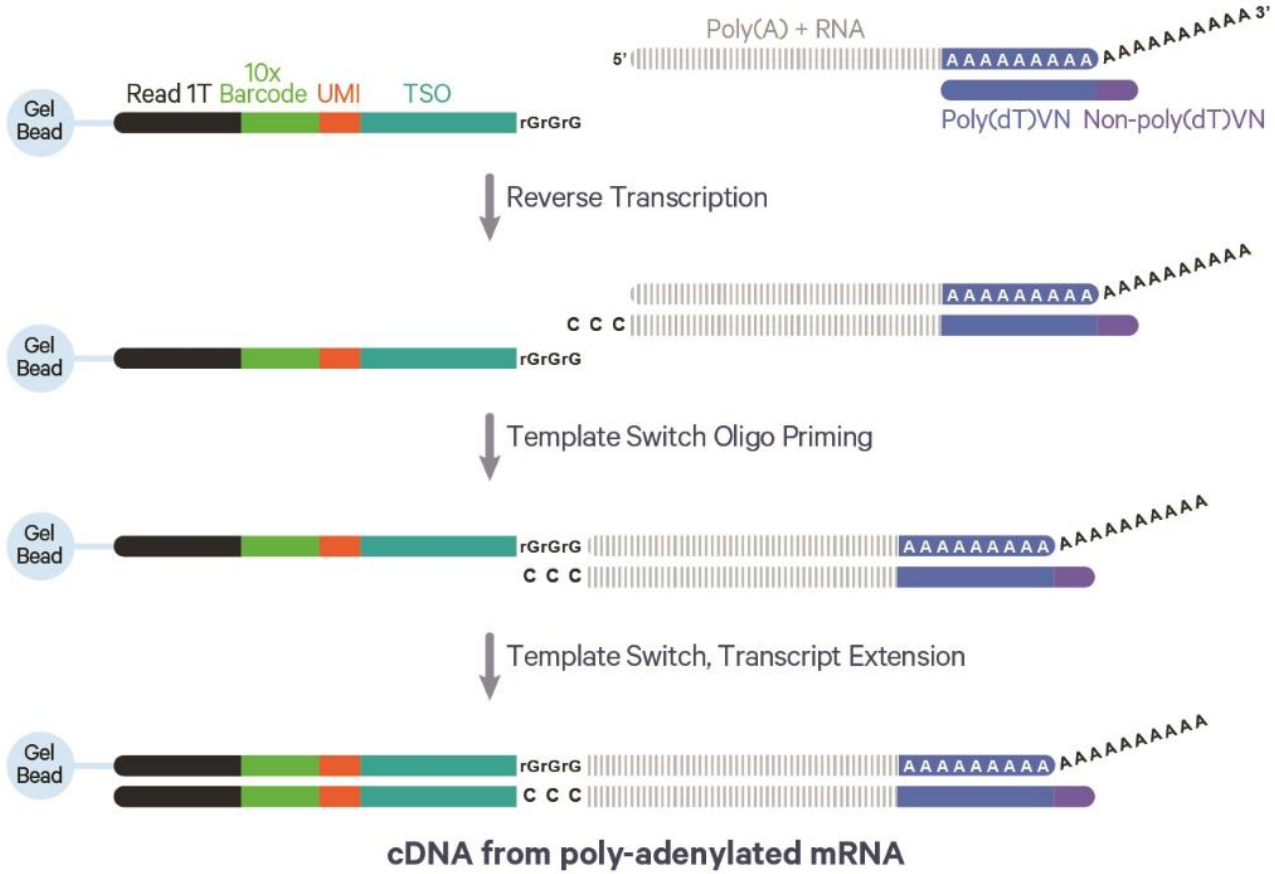
Library Preparation



Gene Expression Library: 5’ Library

Within the drop:

- Cells lysate as soon the drop is formed
- Released cytoplasmic mRNA is reverse transcribed from a free polyT oligo by the RT-Pol.
- Hybrid RNA/DNA is captured by the TSO attached to the bead.
- A DNA polymerization is produced from the TSO oligo to complete a full cDNA.
- All attached adaptors of a bead carries the same Cell Barcode. Each bead has a different Cell Barcode.
- All attached adaptors of a bead carries a different UMI (Unique Molecular Identifier).

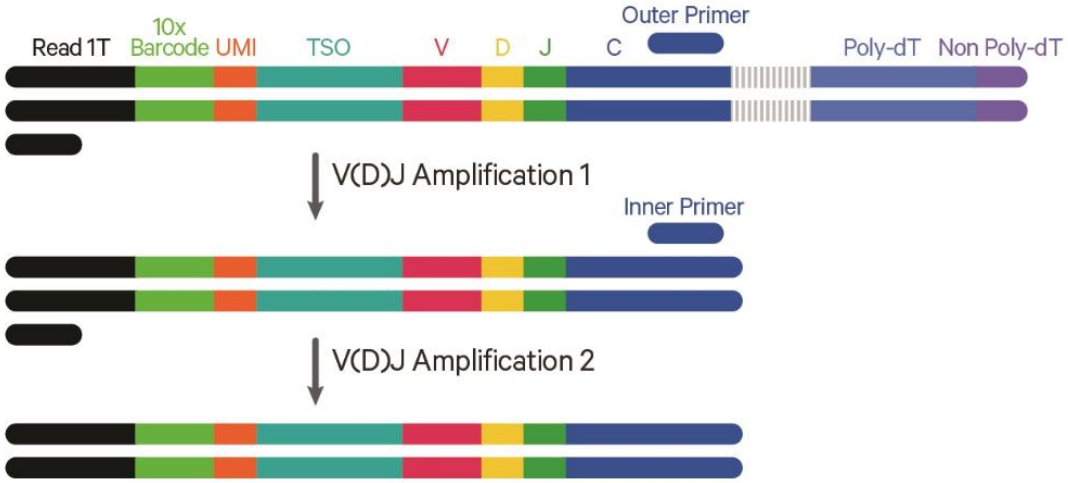


Library Preparation

VDJ Enrichment Library: 5’ Library

In bulk:

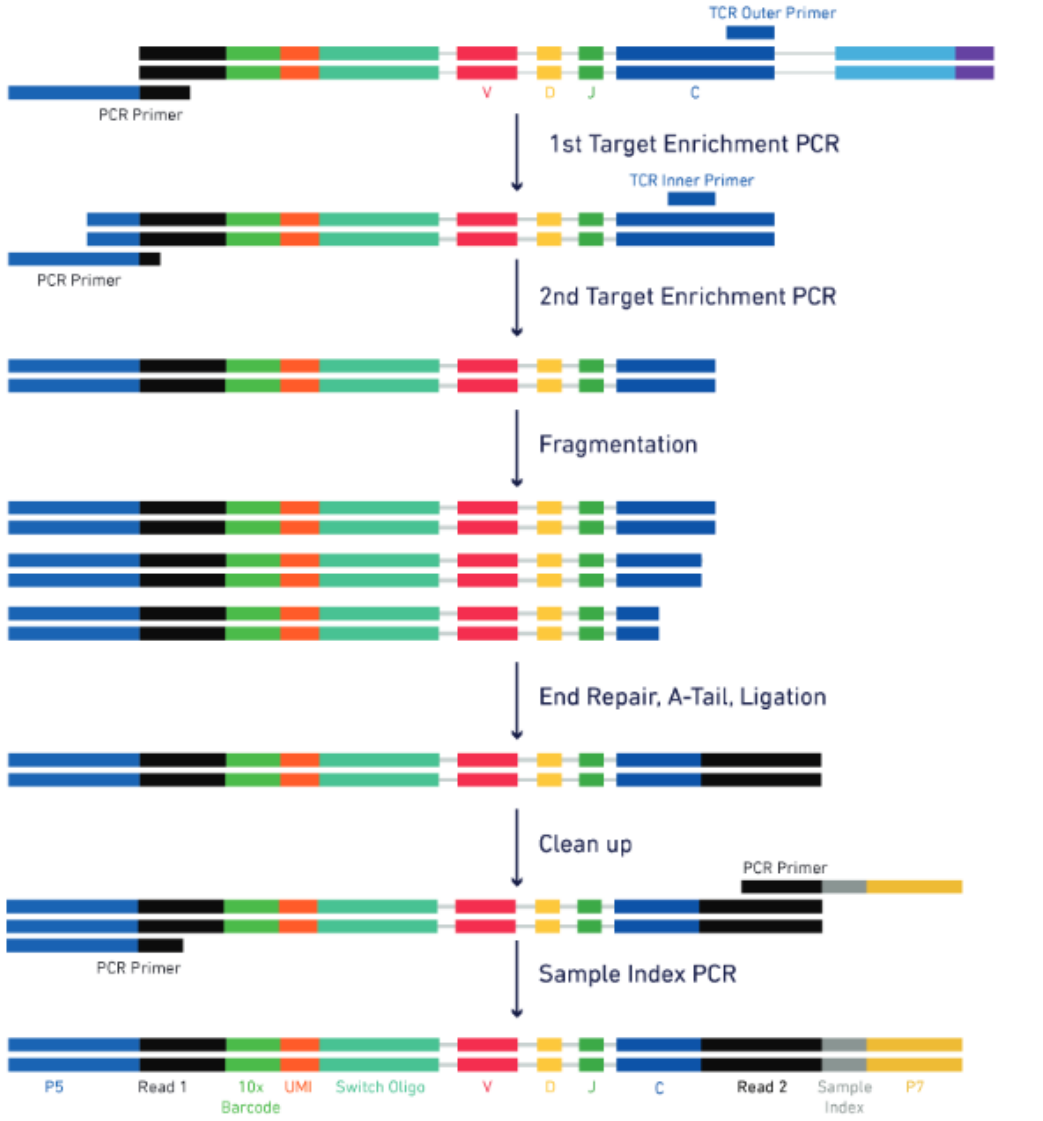
- Nested PCR amplification using same adaptor oligo and two consecutive constant region oligos



Library Preparation

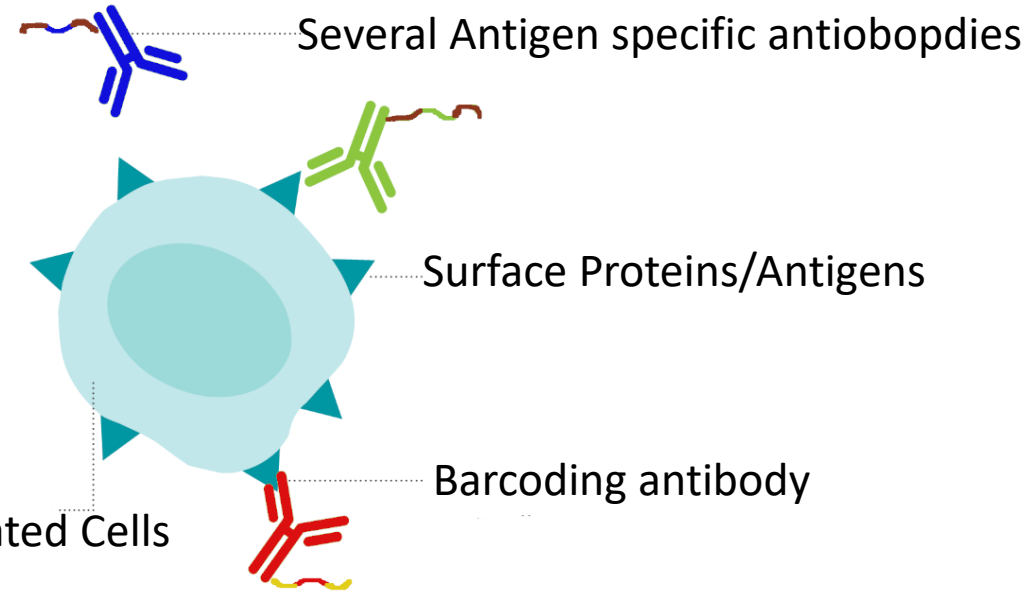
In bulk:

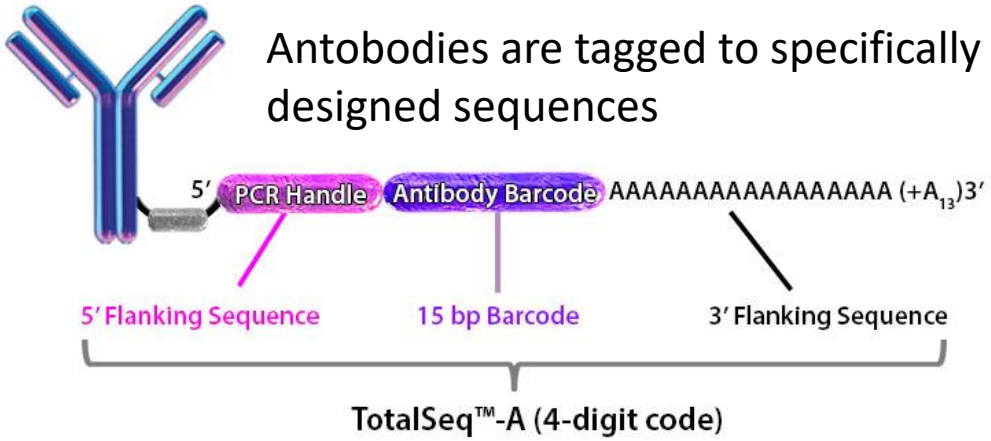
- cDNA amplification to increase the chances of sequence it.
- Enzymatic fragmentation to accommodate fragment size to the capacity of the sequencer.
- Further Illumina adaptors and library index ligation.

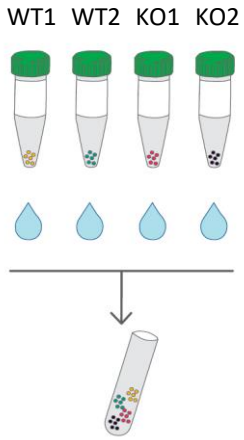


Ab/Cite-Seq + GEX

Dissociated Cells are incubated with surface antibodies to evaluate protein expression.

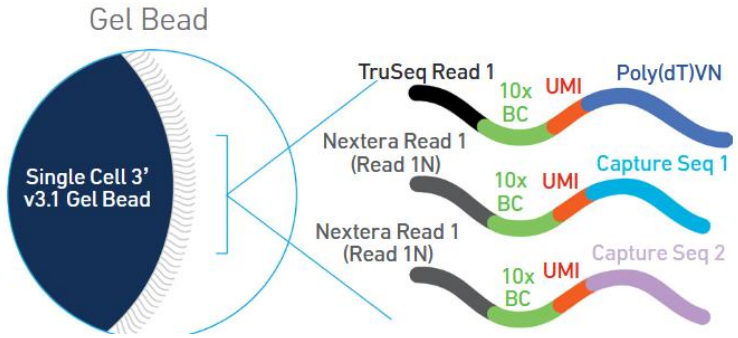


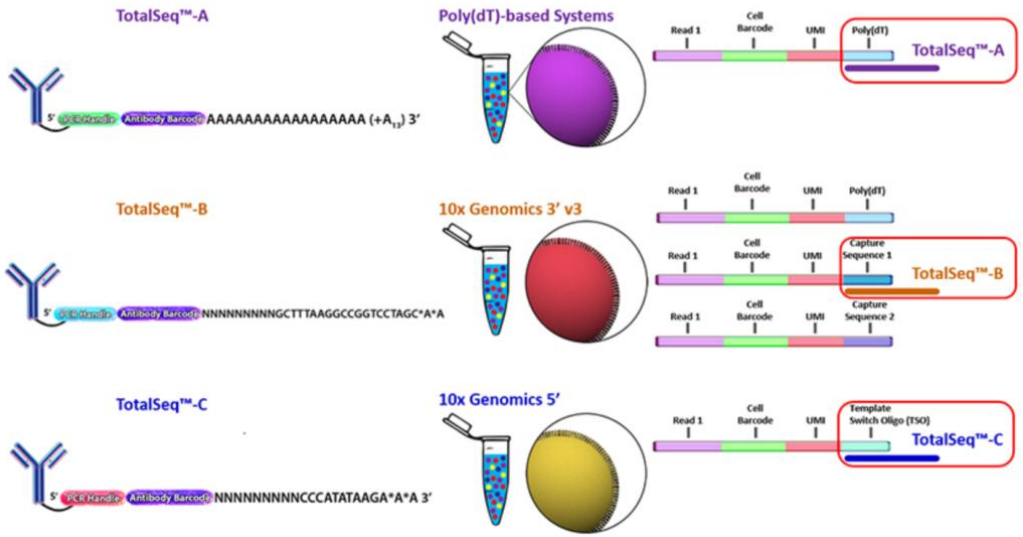




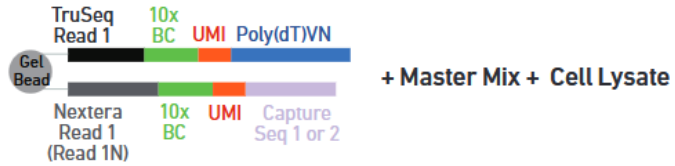
Antibodies against a general ubiquitous surface protein can be used to label samples. Same antibody but different associated barcodes for each tube.<

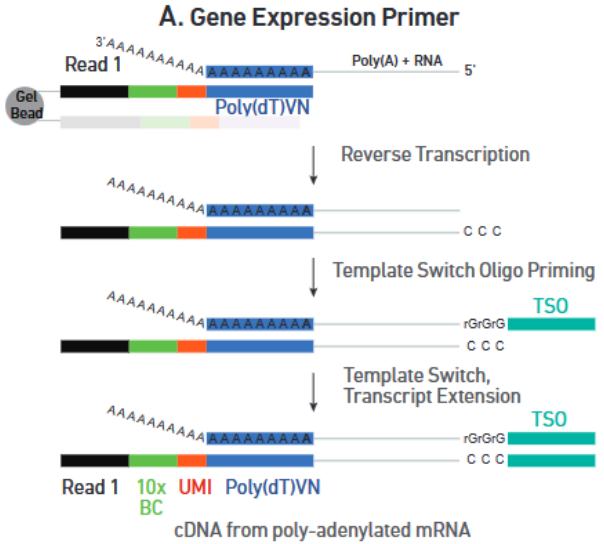
Ab/Cite-Seq + GEX

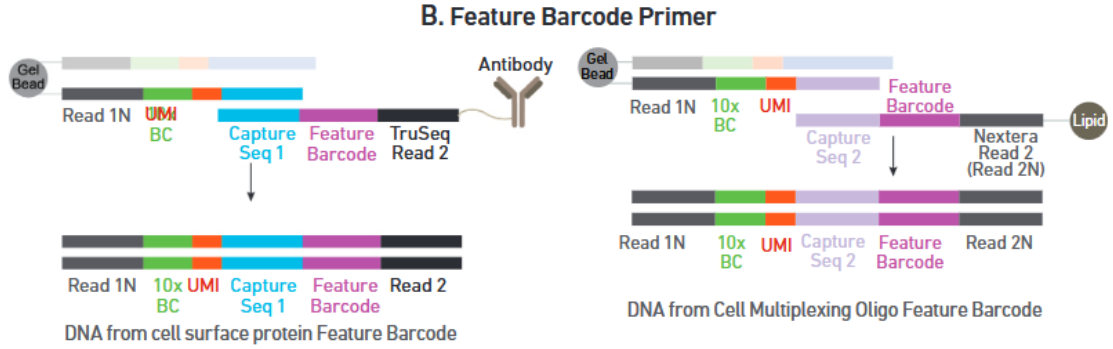




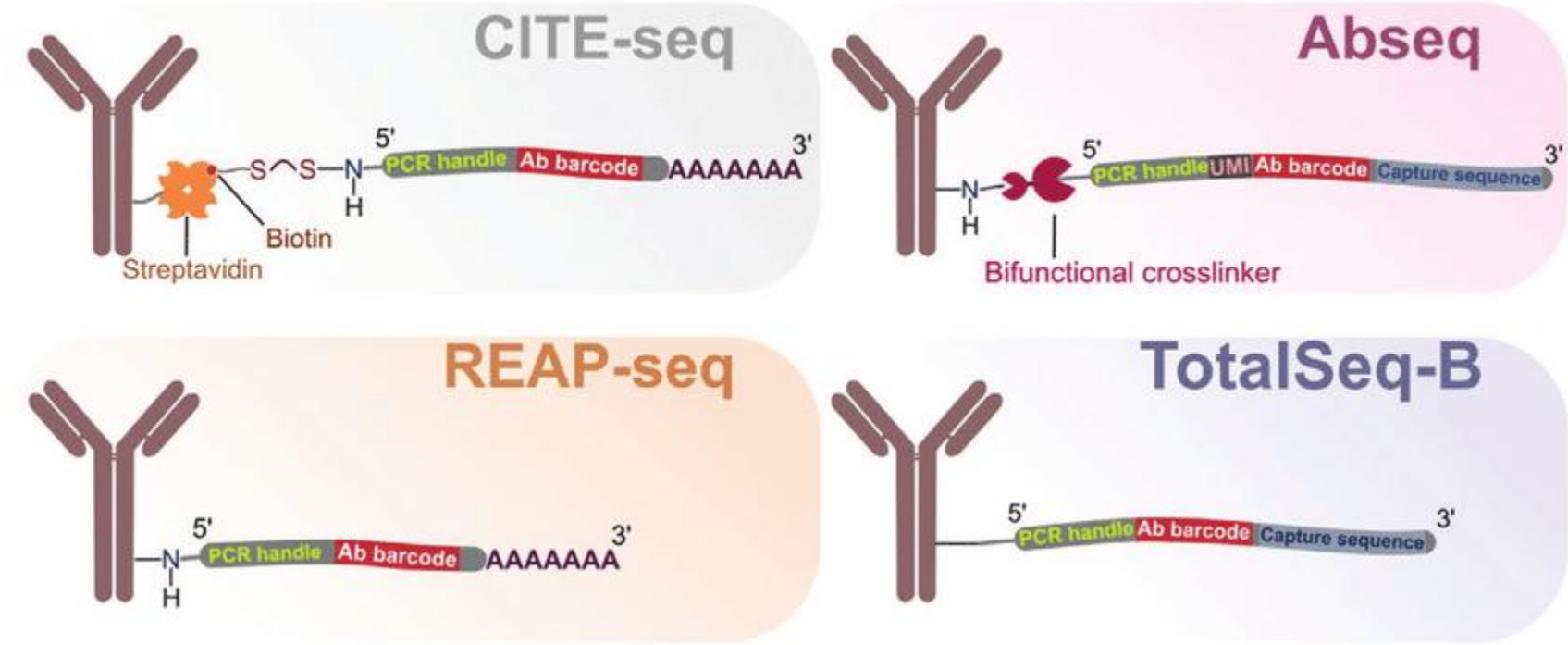
Inside individual GEMs







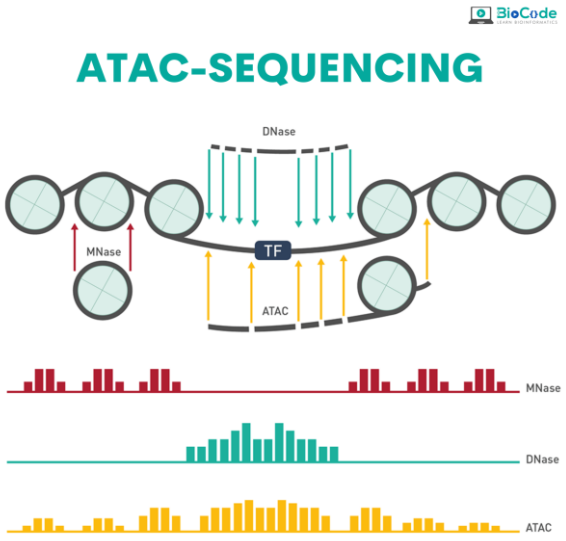
Ab/Cite-Seq

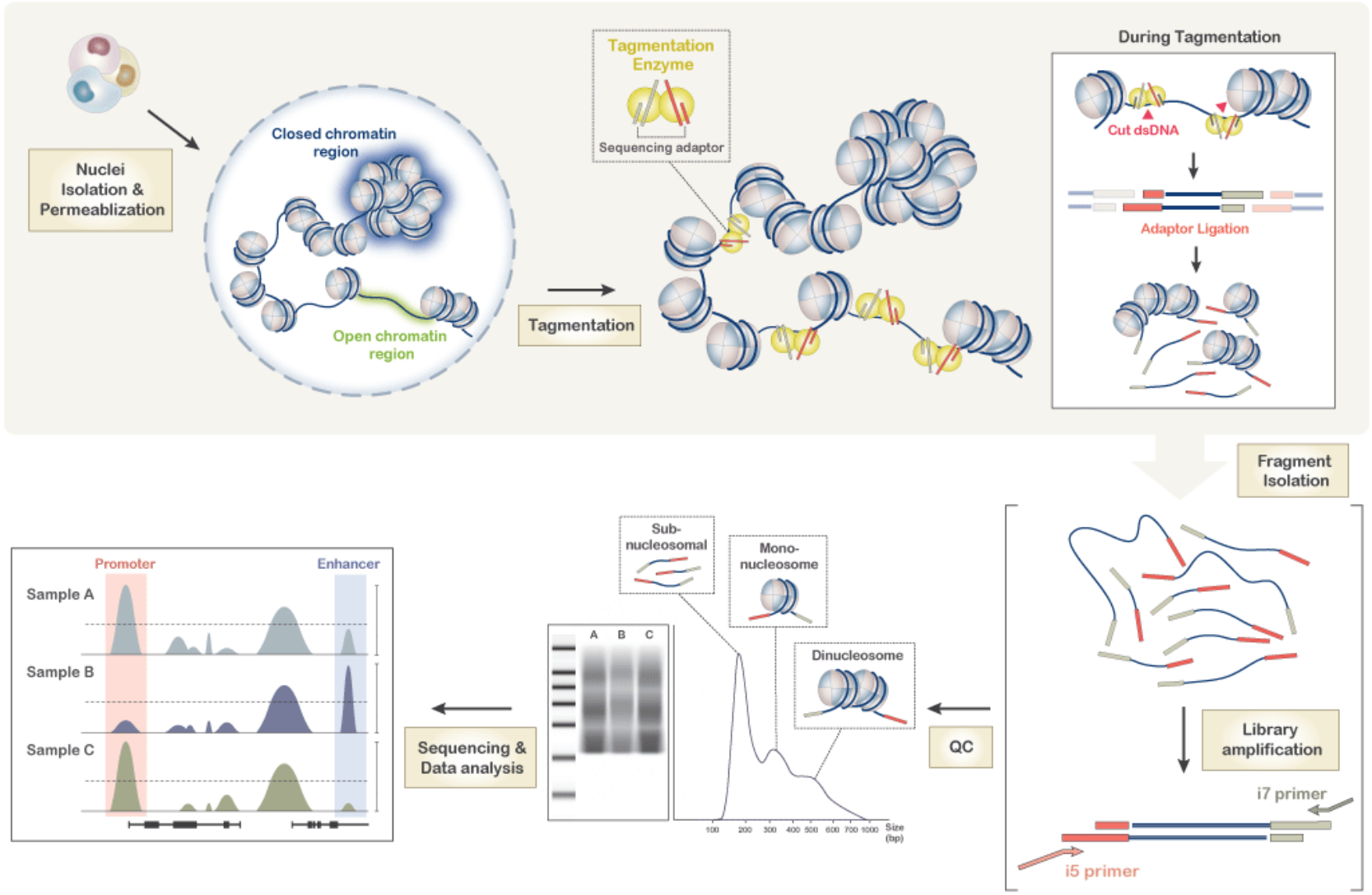

CITE-seq
5'
3'
S
S
N
H
PCR handle
Ab barcode
AAAAAAA
Biotin
Streptavidin
Abseq
5'
3'
N
H
PCR handle
UMI
Ab barcode
Capture sequence
Bifunctional crosslinker
REAP-seq
5'
3'
N
H
PCR handle
Ab barcode
AAAAAAA
TotalSeq-B
5'
3'
PCR handle
Ab barcode
Capture sequence

ATAC-Seq

Allows to identify open chromatin regions which is a proxy of a transcriptionaly active región.

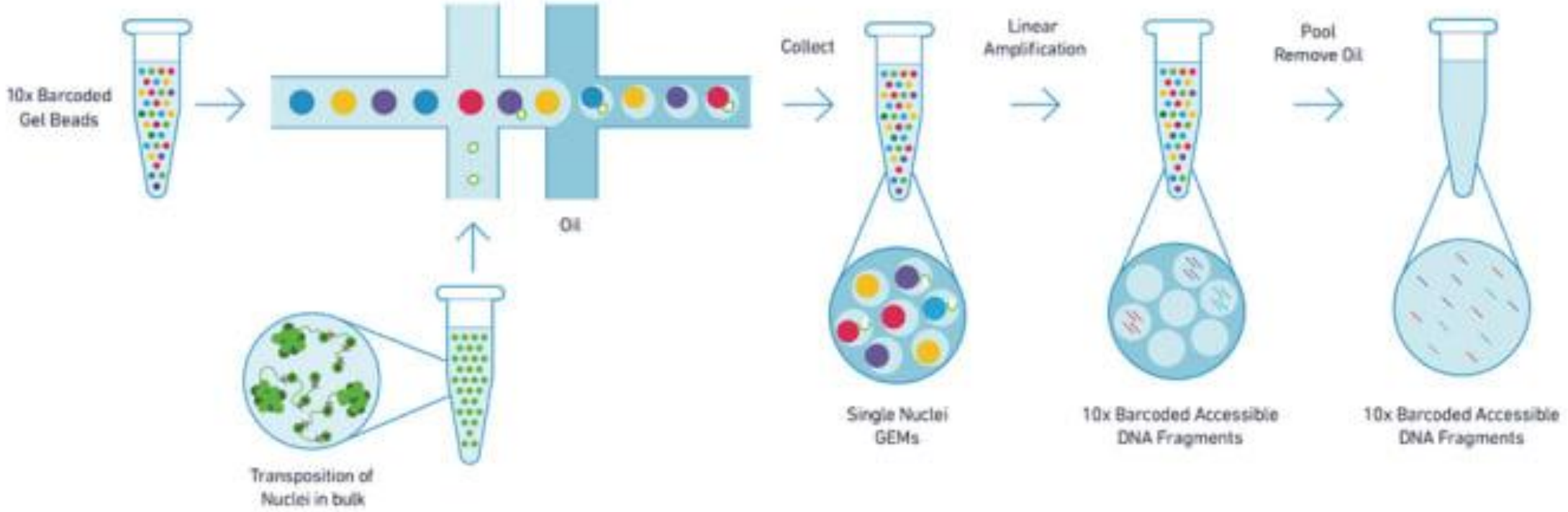
Easier, more flexible, requieres less starting material and cheaper than a ChIP-seq or 4C-Seq experiment



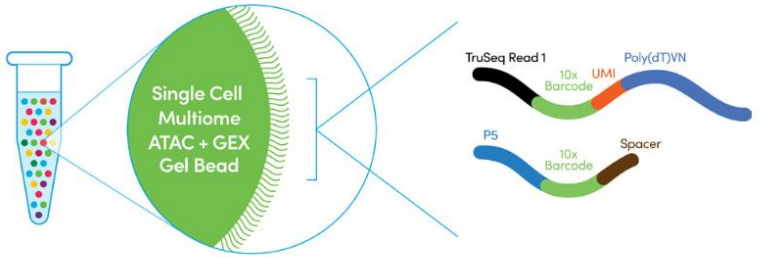


scATA-Seq + GEX

Starting material are isolated nuclei



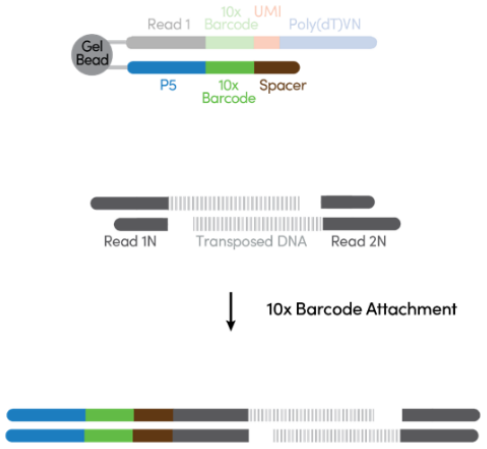
scATA-Seq + GEX

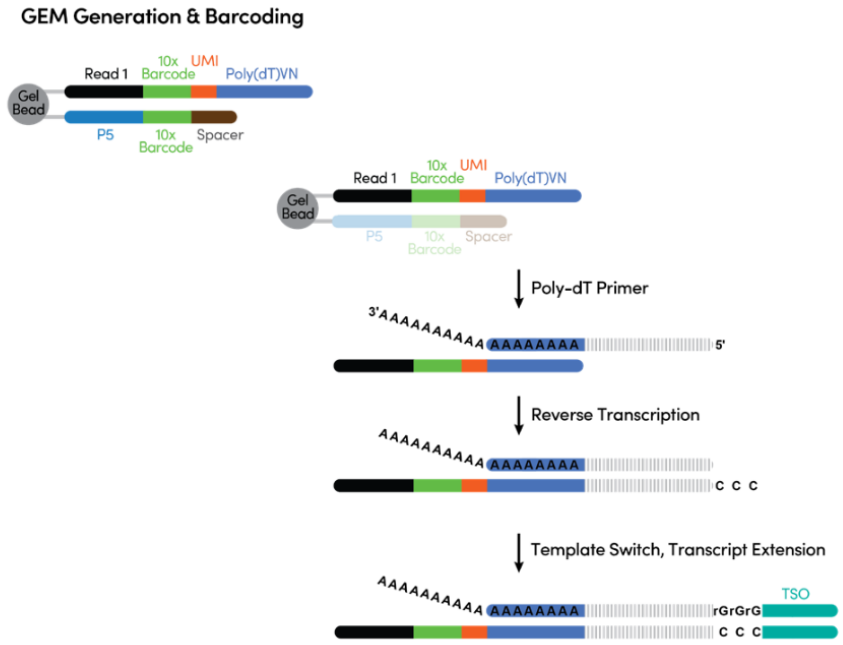


ATAC + Gene Expression Library: 3’ Library

Within the drop:

- Nuclei is lysate as soon the drop is formed.
- Released Tagmentase Fragments are released and captured by the Spacer (Tnf5 sequence complementary sequence)
- The bead barcode is attached to the fragment by PCR.
- Released nuclear mRNA is reverse transcribed from a free polyT oligo by the RT-Pol.
- A DNA polymerization is produced from the TSO oligo to complete a full cDNA.
- All attached adaptors of a bead carries the same Cell Barcode. Each bead has a different Cell Barcode.
- All attached adaptors of a bead carries a different UMI (Unique Molecular Identifier).

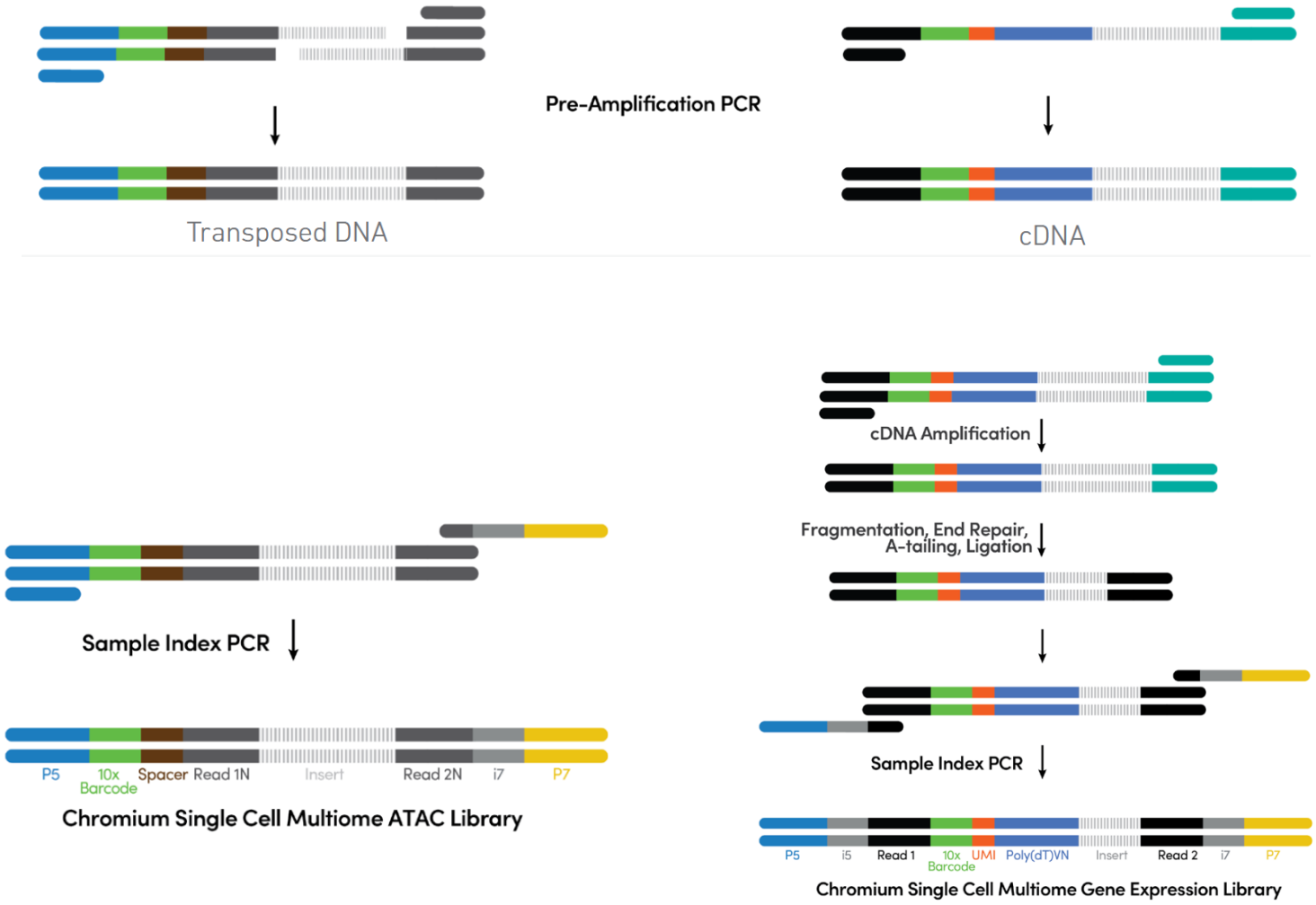




scATA-Seq + GEX

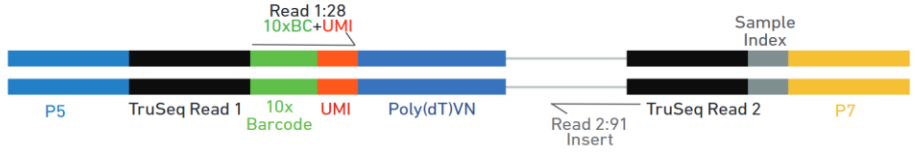
In bulk:

- cDNA amplification to increase the chances of sequence it.
- Enzymatic fragmentation to accommodate fragment size to the capacity of the sequencer.
- Further Illumina adaptors and library index ligation and amplification.

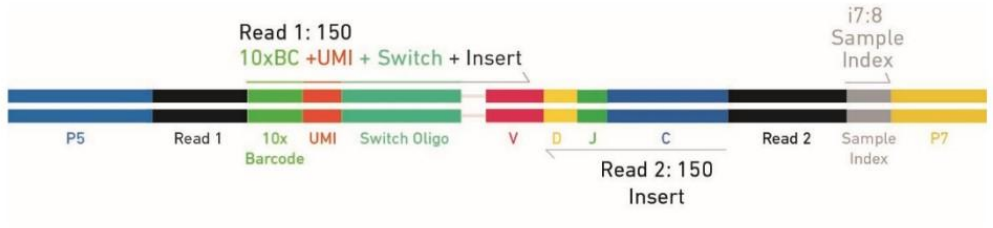


Library Outputs

Chromium Single Cell 3' Gene Expression Library (Single Index)

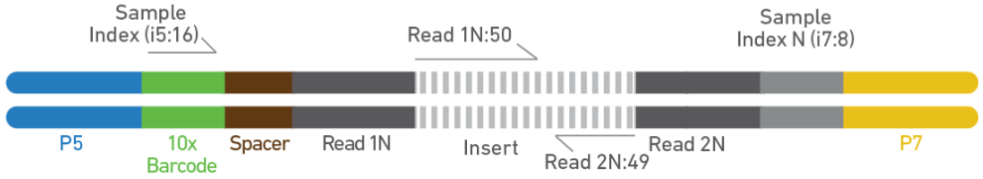


V(D)J Enriched Library Structure:

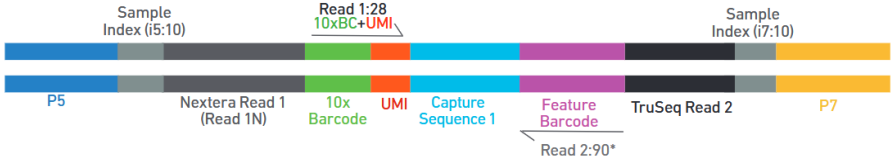


5' Gene Expression Library Structure:

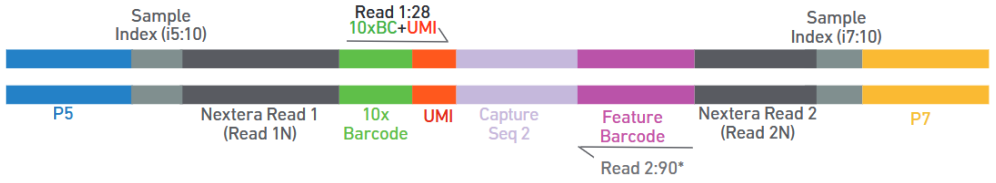
Chromium Single Cell Multiome ATAC Library



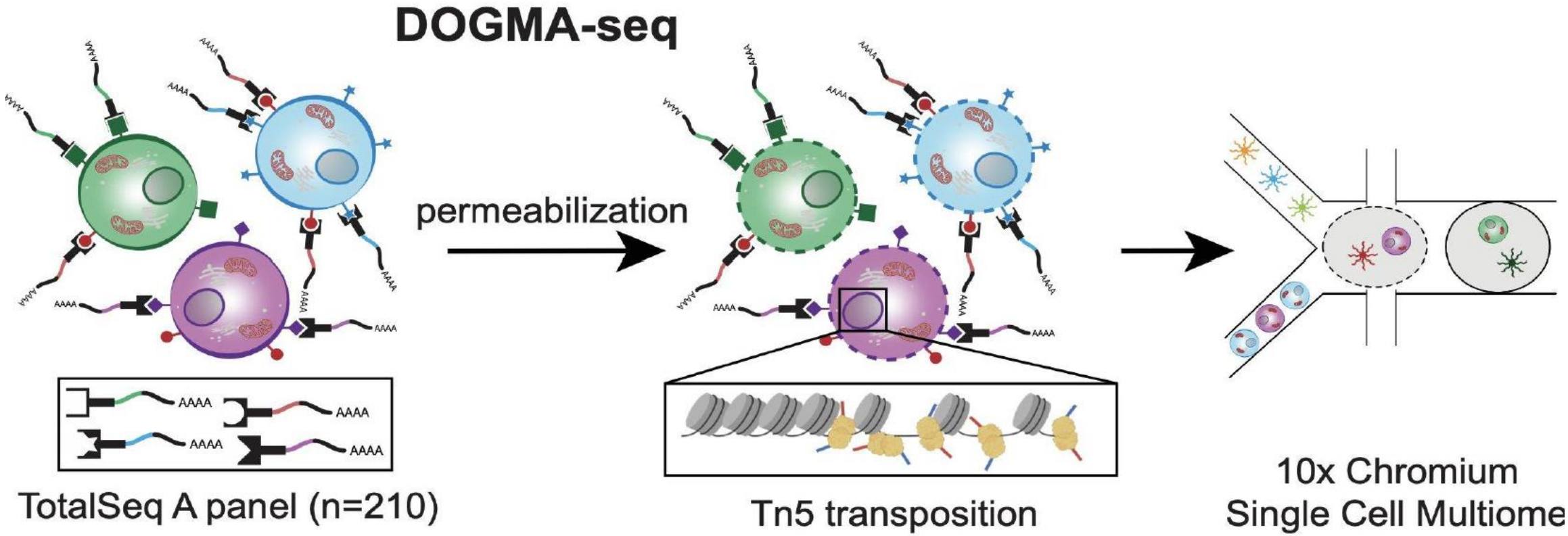
Chromium Single Cell 3' Cell Surface Protein Dual Index Library



Chromium Single Cell 3' Cell Multiplexing Dual Index Library

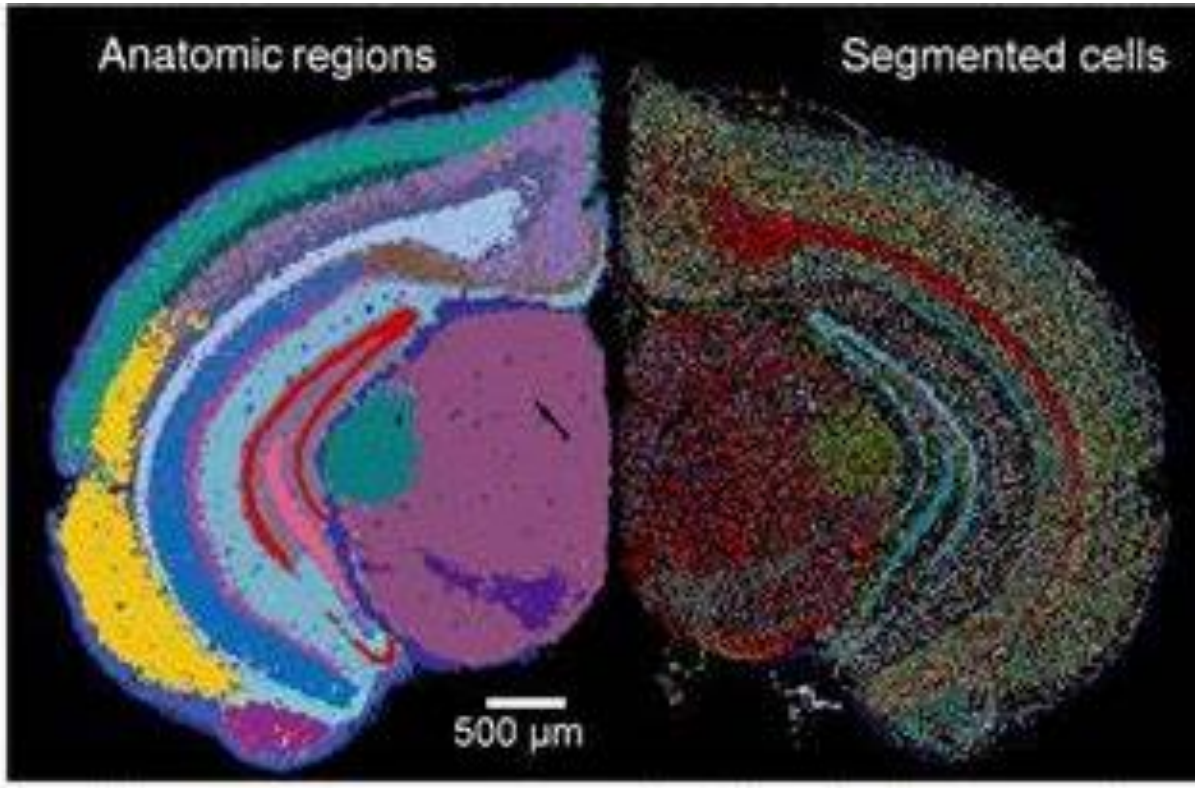


DOGMA-Seq


DOGMA-seq
permeabilization
AAAA
TotalSeq A panel (n=210)
Tn5 transposition
10x Chromium
Single Cell Multiome

Spatial Single Cell

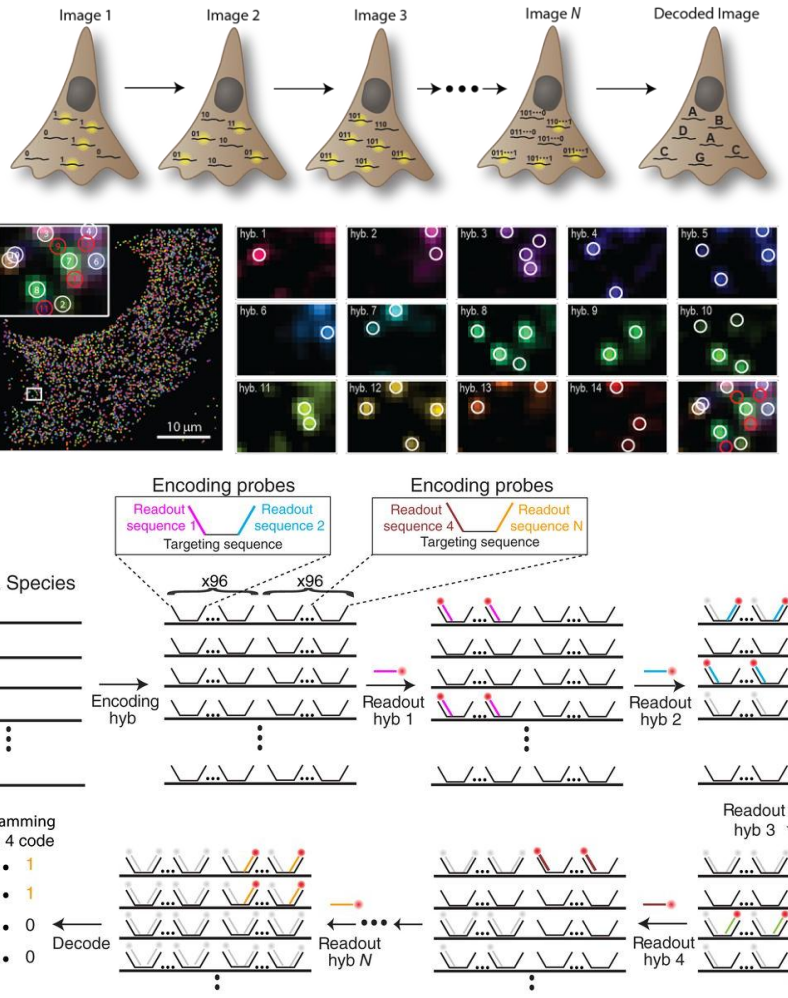
Spatial Single Cell Transcriptomics
From Cell Function to Tissue Function



Spatial SC

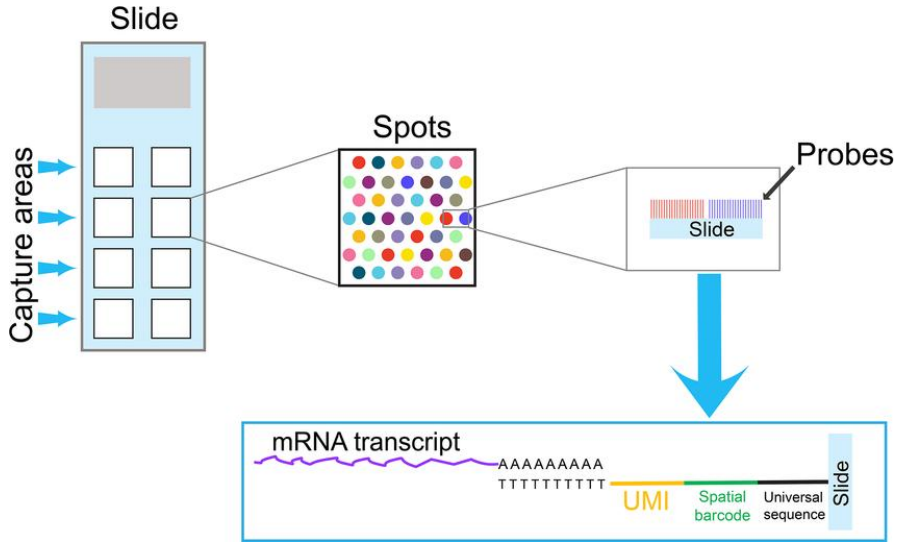
In situ Image Based Detection

MERFISH

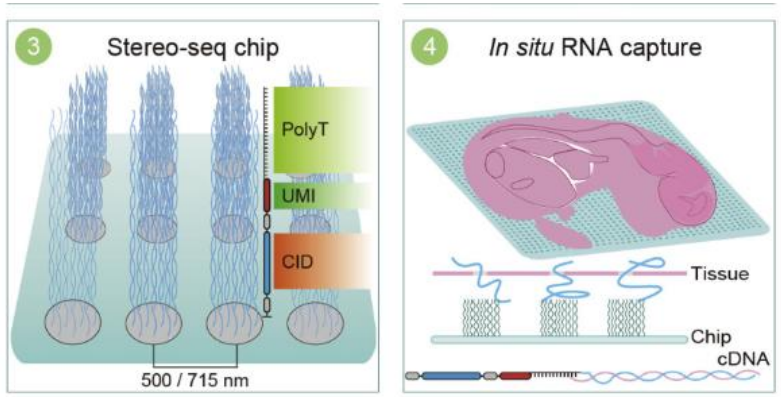


Sequencing Array Based Detection

10x Visium



Stero-Seq



Spatial SC

There are many technologies

Several Parameters are important when selecting the technology

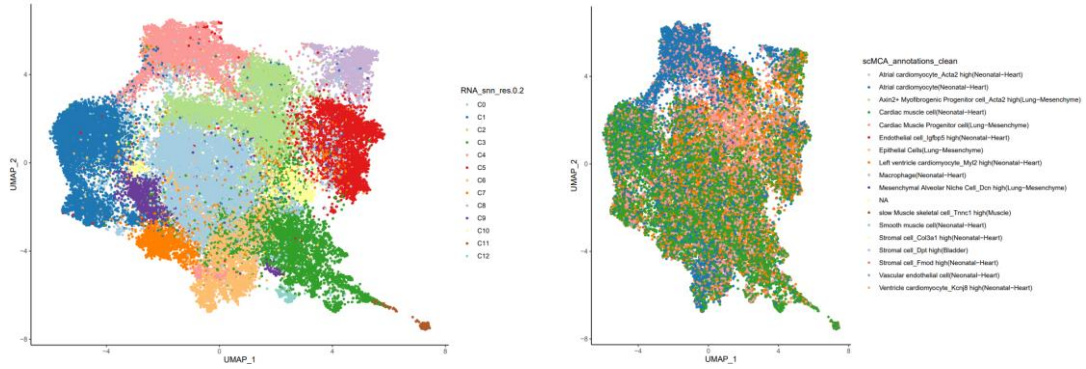
Spatial type	Spatial tech.	Data no.	Size/ radius	Spots	Genes	Sparsity of examined data
Image-based	seqFISH	5	≤1 cell	2,952	4,129	0.46
	MERFISH	4	≤1 cell	3,746	3,395	0.35
	STARmap	2	≤1 cell	1,465	593	0.79
	ISS	2	≤1 cell	6,000	120	0.93
	FISH	1	≤1 cell	3,039	84	0.72
	osmFISH	1	≤1 cell	3,405	33	0.30
	BARISTAseq	1	≤1 cell	11,426	80	0.83
	EXseq	1	≤1 cell	1,154	42	0.41
Seq-based	10× visium	21	55 μm	2,002	28,737	0.32
	Slide-seq	3	10 μm	1,9175	23,682	0.89
	ST	2	100 μm	875	17,916	0.45
	HDST	1	≤1 cell	6,000	14,199	0.99
	Seq-scope	1	≤1 cell	2,177	19,532	0.86

Spatial SC

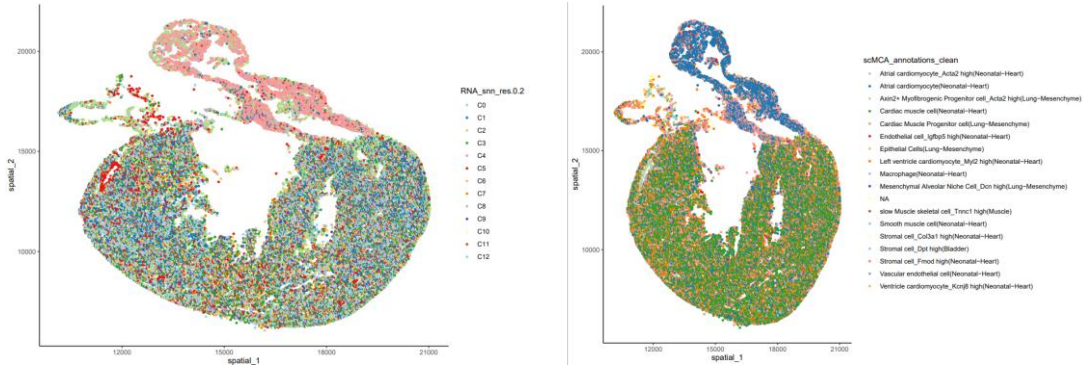
Cell Clustering based on Transcriptome Profile

Manual Cell Type Identification

UMAP



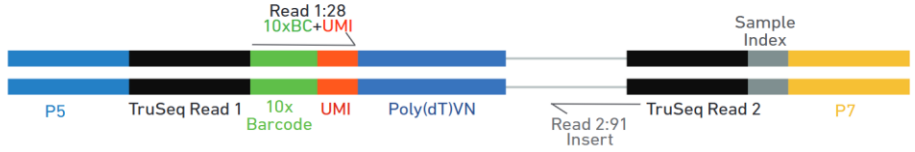
Spatial



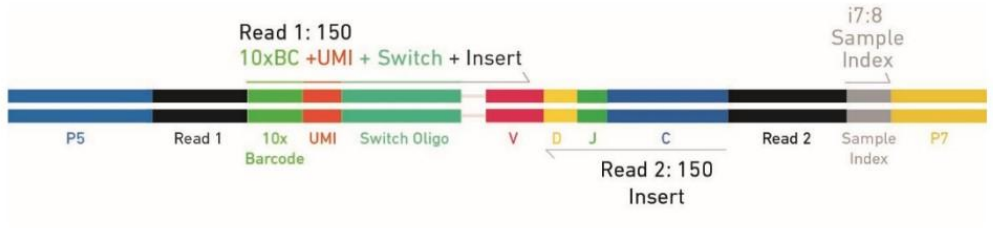
Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., ... & Wang, J. (2022). Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell, 185(10), 1777-1792. DOI: 10.1016/j.cell.2022.04.003

Library Outputs

Chromium Single Cell 3' Gene Expression Library (Single Index)

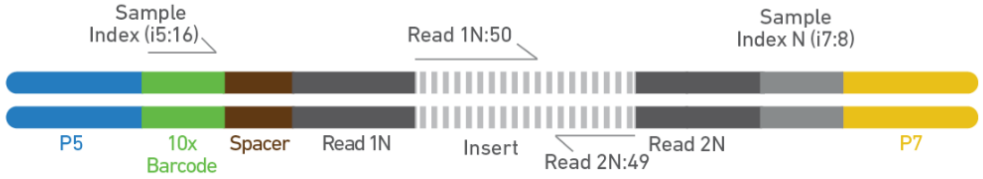


V(D)J Enriched Library Structure:

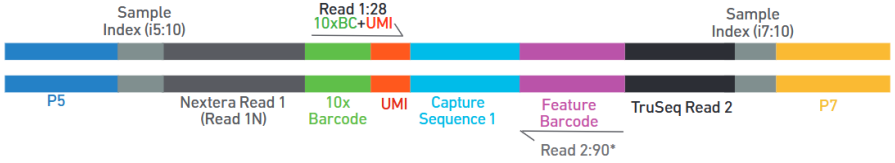


5' Gene Expression Library Structure:

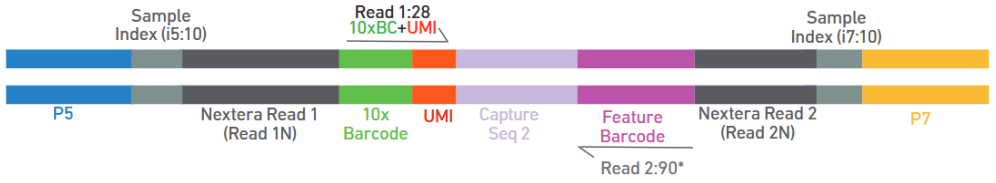
Chromium Single Cell Multiome ATAC Library



Chromium Single Cell 3' Cell Surface Protein Dual Index Library

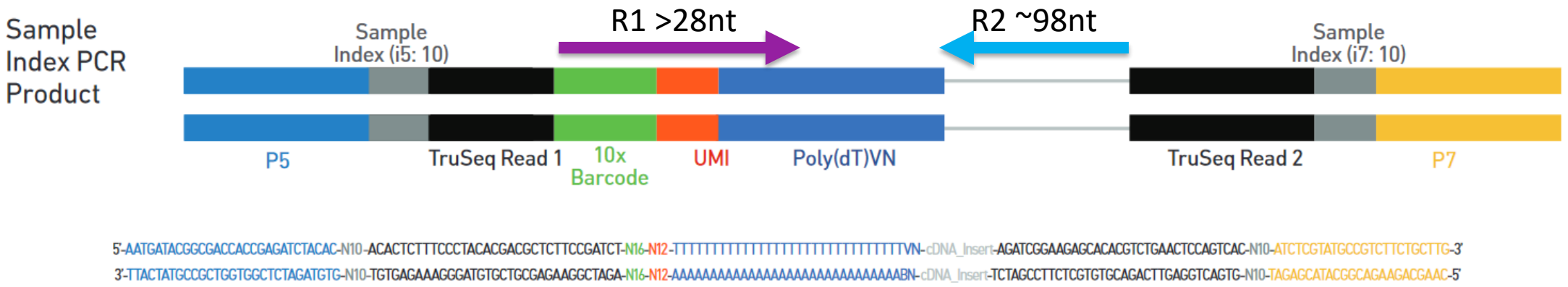


Chromium Single Cell 3' Cell Multiplexing Dual Index Library



Sequencing

- Single Cell Libraries are sequenced in paired end mode.
- Different Single cell platforms and library protocols require different lengths for both reads



- **R1** Contains Cell Barcode and UMI
- **R2** Contains sequence from the mRNA captured molecule

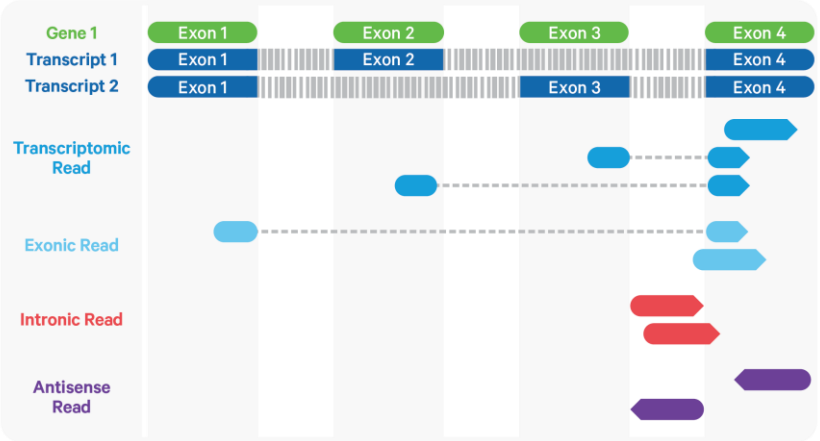
Alignment and Quantification

Pre-Processing:

- R2 is annotated with the Cell Barcode and UMI from R1
- From now only annotated R2 are used.

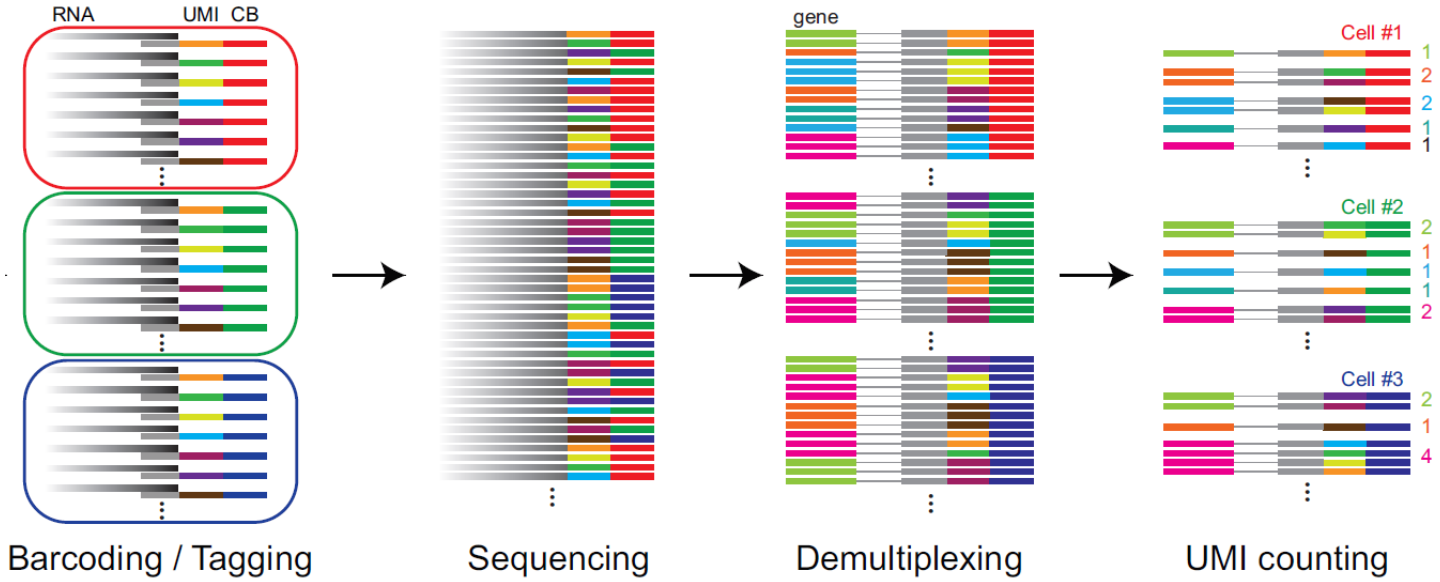
Alignment:

- R2 is mapped to the genome using STAR.
- Only transcript compatible alignments associated to a unique gene are considered for next UMI counting steps.
- Each mapped R2 is annotated with the Gene is compatible with.



UMI quantification:

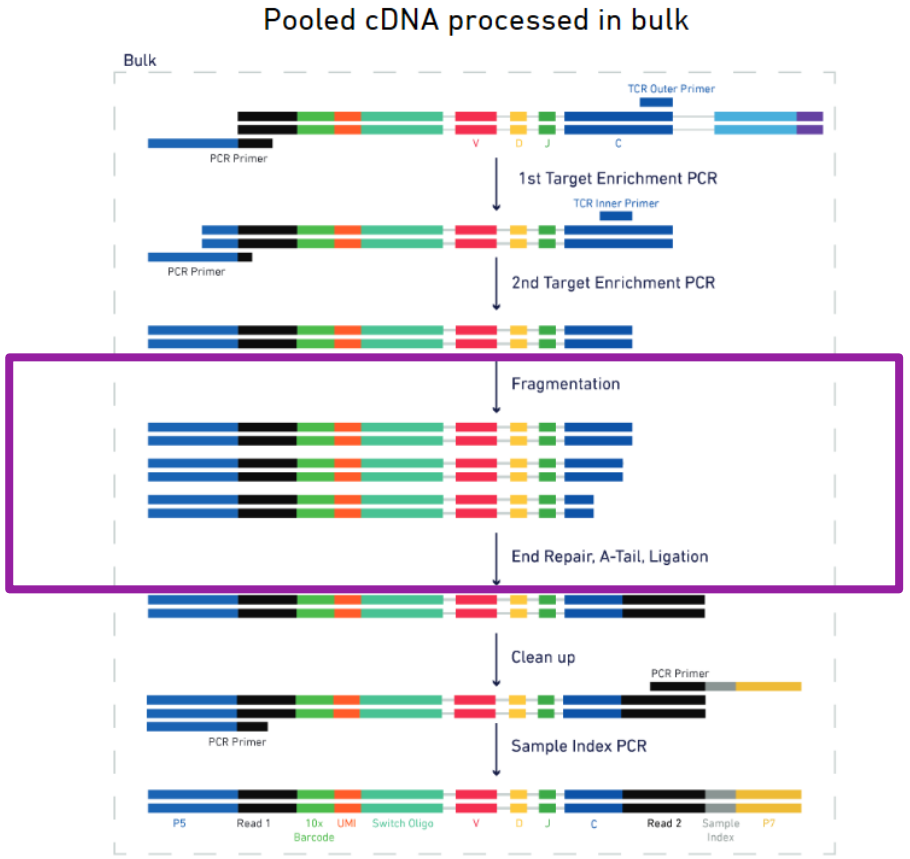
- Now we have a combination of Cell Barcode/UMI/Gene.
- All identical combinations are collapsed to one (they are PCR duplicates from one original mRNA on that cell).
- UMIs per gene per cell are computed.



Library Preparation

One UMI = One mRNA molecule

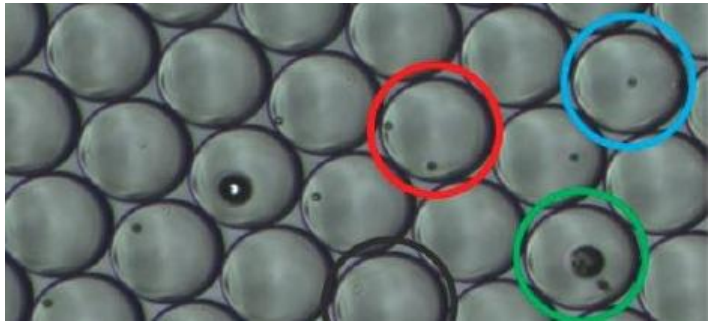
- We may observe the same UMI attached to different fragments of the same mRNA.
- Those different fragments are the product of the enzymatic fragmentation on a PCR amplified UMI-cDNA molecule.
- All identical UMIs have been originated from the same individual mRNA molecule

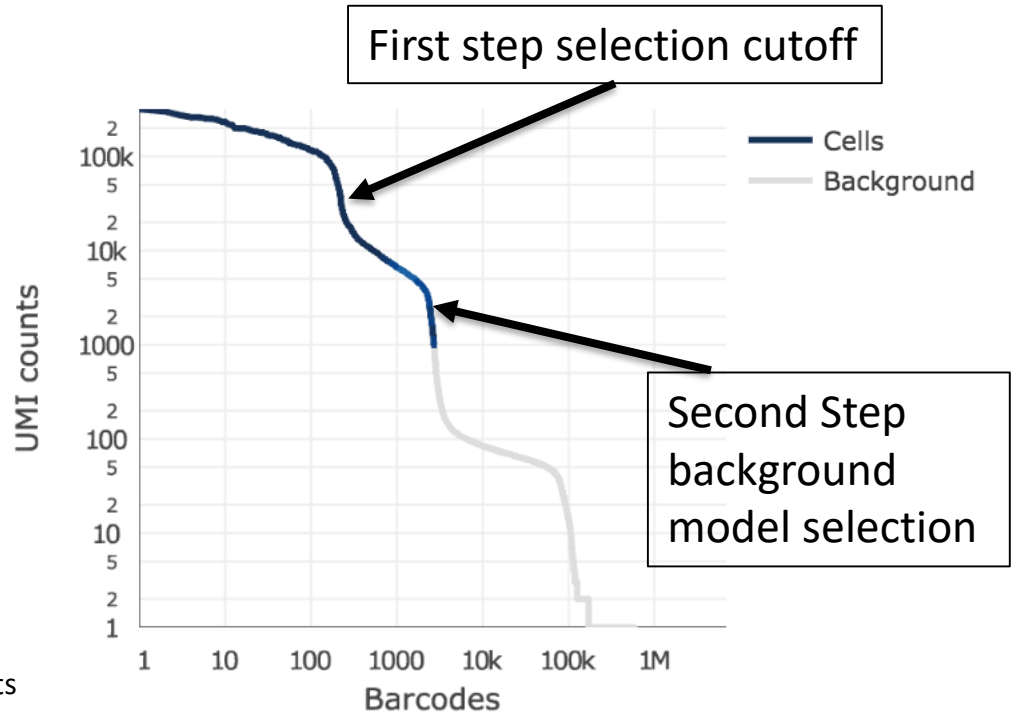


Cell Detection

Cell Detection:

- Not all beads from a cell capture step have got a cell.
- Empty beads still may capture free background RNA and produce cDNAs.
- Cell calling step tries to differentiate true cells from background
- In general, true cells produce more RNA and accounting for more genes than empty beads.
- However there are certain cell types with low RNA content that might yield library product at similar levels to background.
- Cell calling is perform in two steps:
  - First a cut-off is placed at the point in which counts per cell drops dramatically.
  - On the lower yield cells a model of the background based on the data itself is produced and cells are called when they differ substantially from that model.
- Only transcript compatible alignments associated to a unique gene are considered for next UMI counting steps.



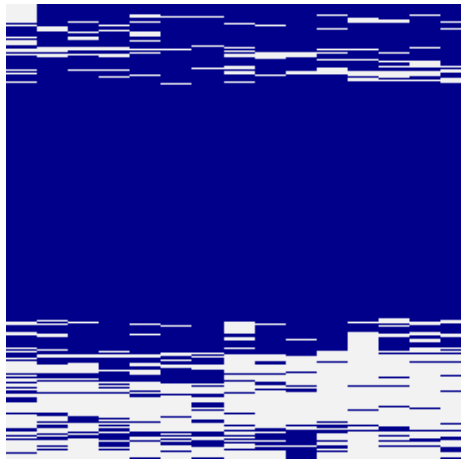


Lun, A., Riesenfeld, S., Andrews, T. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20,** 63 (2019). https://doi.org/10.1186/s13059-019-1662-y

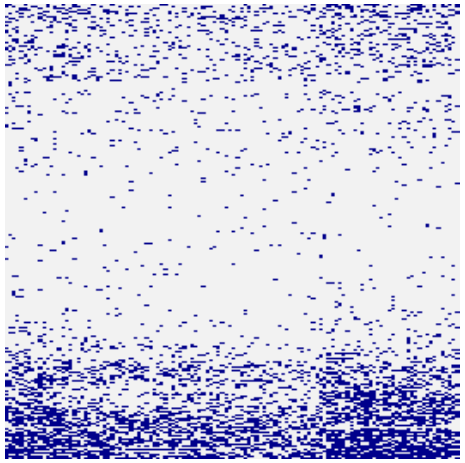
How the data looks like?

Data is very sparse

Bulk Samples



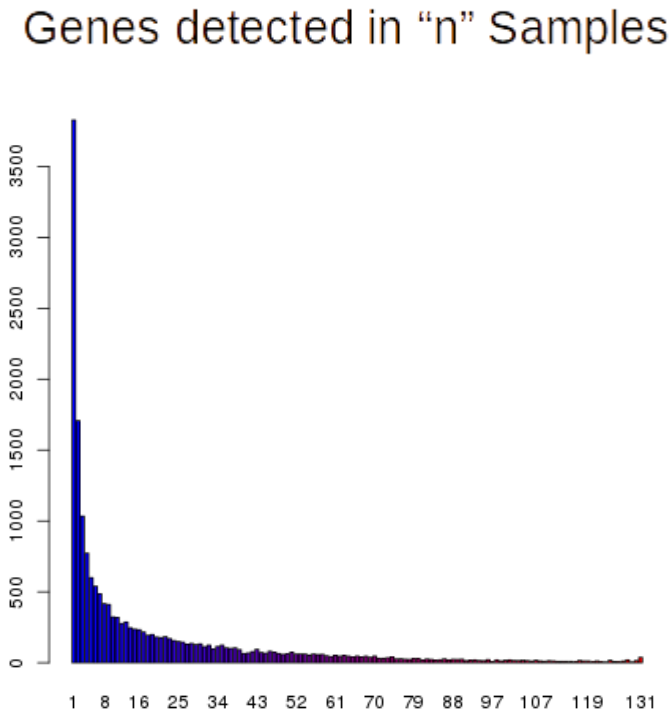
SingleCell Samples

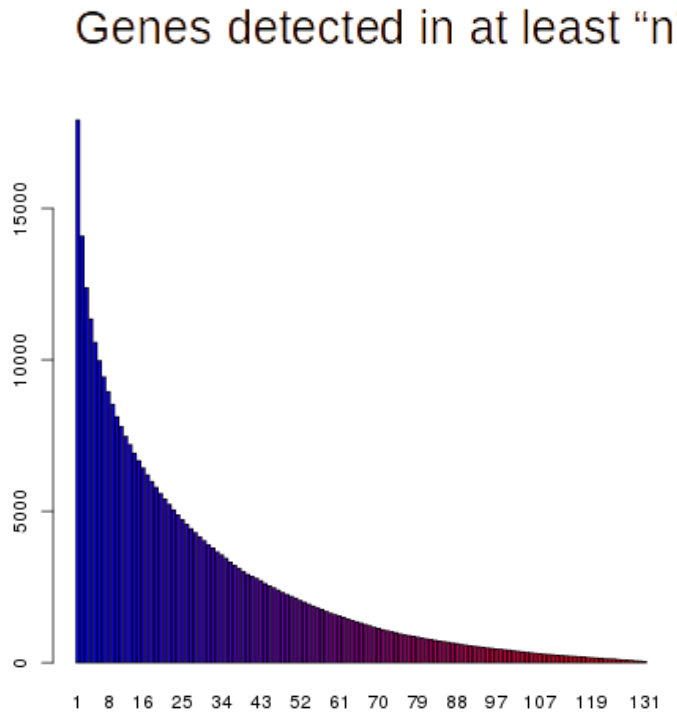


One Count in at least One Sample

Non of the assumptions for Bulk RNA-Seq applies to a scRNA-Seq experiment

How the data looks like?

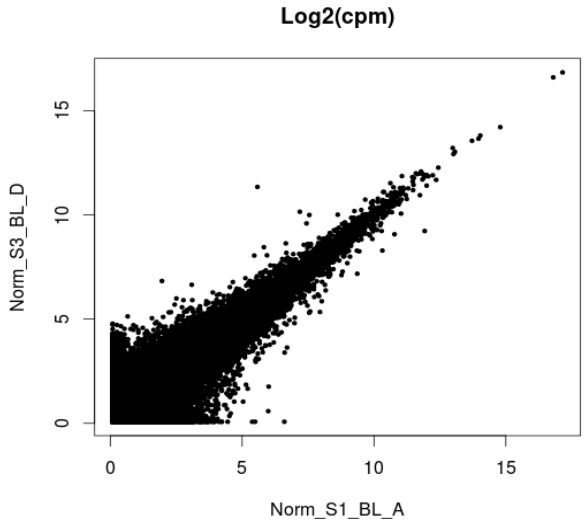




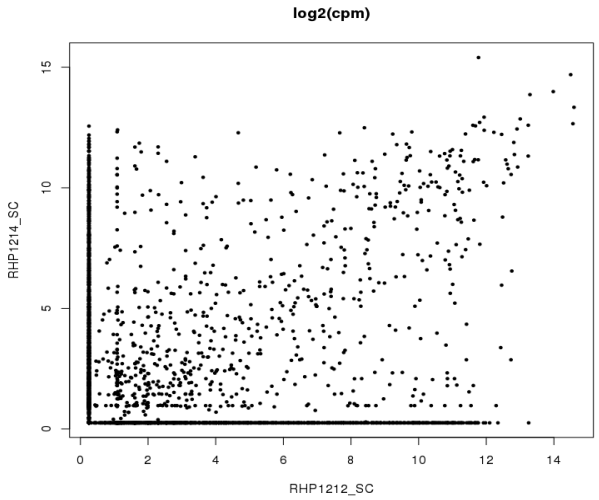
How the data looks like?

Correlations between replicates is very poor

Bulk Samples



SingleCell Samples



Tools For Singel Cell Analysis

Seurat (R) : https://satijalab.org/seurat/index.html

Scater (R): https://bioconductor.org/packages/release/bioc/html/scater.html

Scran (R): https://bioconductor.org/packages/release/bioc/html/scran.html

Scanpy (Python): https://scanpy.readthedocs.io/en/stable/

Sample Dataset Stuart 2019

Cell. 2019 June 13; 177(7): 1888–1902.e21. doi:10.1016/j.cell.2019.05.031.

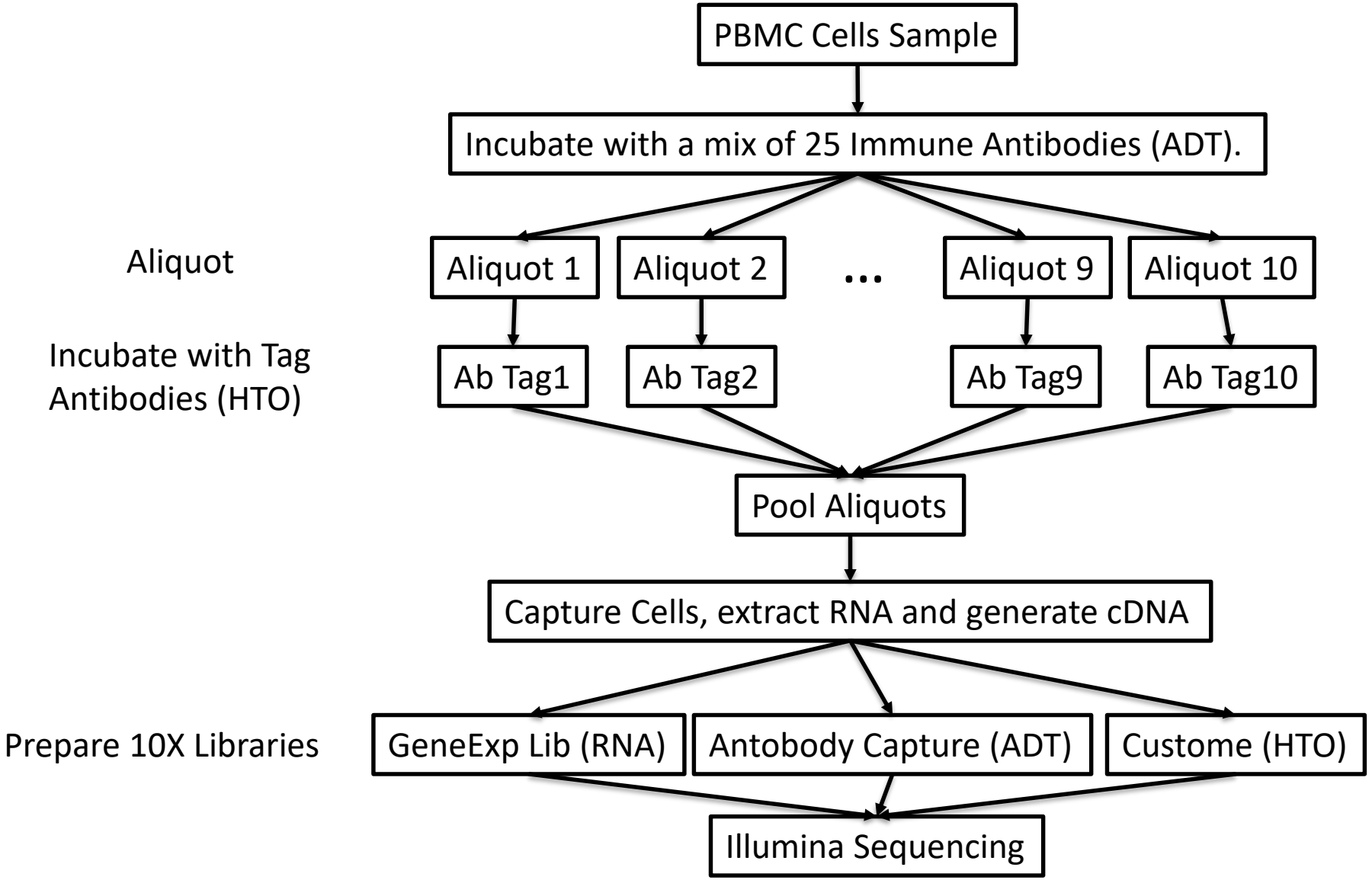
Comprehensive integration of single-cell data

Tim Stuart[1,*], Andrew Butler[1,2,*], Paul Hoffman[1], Christoph Hafemeister[1], Efthymia Papalexi[1,2], William M. Mauck III[1,2], Yuhan Hao[1,2], Marlon Stoeckius[3], Peter Smibert[3], Rahul Satija[1,2,]**

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128639

For the scRNA/CITE-seq experiment: a panel of 25 antibodies (ADTs) was used to stain the cells according to the CITE-seq protocol. To identify cell doublets, we split cells into 10 aliquots and stained each aliquot with a unique hashing (HTOs) antibody. Samples were washed and recombined and load on a 10genomics chip. 3prime 10x genomics Version 2 was ran and the manufacturers protocols was followed for cDNA library generation. To recover HTOs and ADTs we followed the CITE-seq protocol found here: https://cite-seq.com/protocol/

Experiment Protocol



Quantification

Using CellRanger pipeline versión 6. https://www.10xgenomics.com -> Support -> Software -> CellRanger

Command

```
cellranger count --id=GSE128639_10XGenomics \
        --libraries=GSE128639_10XGenomics_Sample_fastqs.csv \
        --transcriptome=refdata-gex-GRCh38-2020-A \
        --feature-ref=GSE128639_10XGenomics_HTORefTable_feature_ref.csv \
        --chemistry=threeprime --localcores=40 --localmem=96 --disable-ui --nosecondary --no-bam
```

Sample_fastqs.csv

```
fastqs,sample,library_type
HTOlib,GSE128639hto,Custom
ADTlib,GSE128639adt,Antibody Capture
GEXlib,GSE128639gex,Gene Expression
```

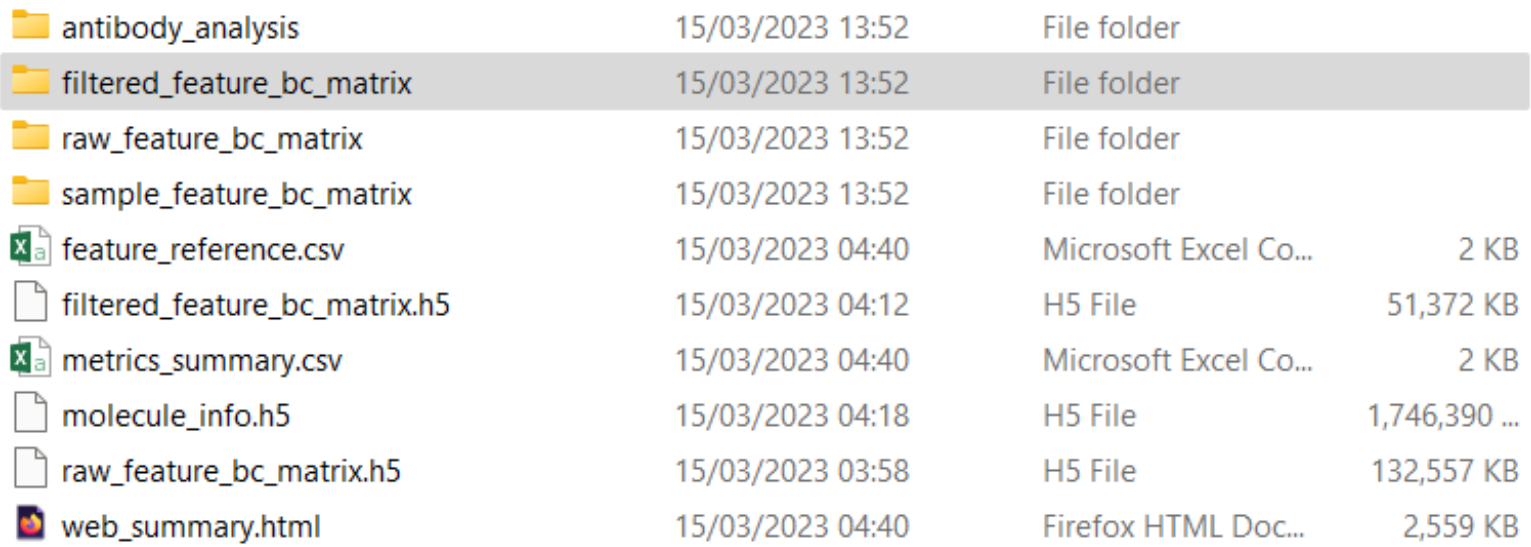
Fastq Sample Prefix

Path to folder containing fastqs of the described library

HTORefTable_feature_ref.csv

```
id,name,read,pattern,sequence,feature_type
HTO1,LNH-94a,R2,5P(BC),GTCAACTCTTTAGCG,Custom
HTO2,LNH-94b,R2,5P(BC),TGATGGCCTATTGGG,Custom
…
HTO9,LNH-94i,R2,5P(BC),CAGTAGTCACGGTCA,Custom
HTO10,LNH-94j,R2,5P(BC),ATTGACCCGCGTTAG,Custom
CD3,CD3,R2,5P(BC),CTCATTGTAACTCCT,Antibody Capture
CD56,CD56,R2,5P(BC),TCCTTTCCTGATAGG,Antibody Capture
…
```

CellRanger Output folder



antibody_analysis	15/03/2023 13:52	File folder	
filtered_feature_bc_matrix	15/03/2023 13:52	File folder	
raw_feature_bc_matrix	15/03/2023 13:52	File folder	
sample_feature_bc_matrix	15/03/2023 13:52	File folder	
feature_reference.csv	15/03/2023 04:40	Microsoft Excel Co...	2 KB
filtered_feature_bc_matrix.h5	15/03/2023 04:12	H5 File	51,372 KB
metrics_summary.csv	15/03/2023 04:40	Microsoft Excel Co...	2 KB
molecule_info.h5	15/03/2023 04:18	H5 File	1,746,390 ...
raw_feature_bc_matrix.h5	15/03/2023 03:58	H5 File	132,557 KB
web_summary.html	15/03/2023 04:40	Firefox HTML Doc...	2,559 KB

CellRanger Output: Matrix Market or h5 file

barcodes.tsv.gz

```
AAACCTGAGACAAAGG-1
AAACCTGAGACGCACA-1
AAACCTGAGAGGGCTT-1
AAACCTGAGAGTAAGG-1
AAACCTGAGAGTACAT-1
AAACCTGAGAGTGAGA-1
AAACCTGAGAGTTGGC-1
AAACCTGAGCTAACTC-1
AAACCTGAGCTGAAAT-1
AAACCTGAGCTTATCG-1
AAACCTGAGGTCGGAT-1
```

features.tsv.gz

```
ENSG00000243485 MIR1302-2HG     Gene Expression
ENSG00000237613 FAM138A Gene Expression
ENSG00000186092 OR4F5   Gene Expression
ENSG00000238009 AL627309.1      Gene Expression
ENSG00000239945 AL627309.3      Gene Expression
ENSG00000239906 AL627309.2      Gene Expression
ENSG00000241860 AL627309.5      Gene Expression
ENSG00000241599 AL627309.4      Gene Expression
```

...

```
HTO8    LNH-94h Custom
HTO9    LNH-94i Custom
HTO10   LNH-94j Custom
CD3     CD3     Antibody Capture
CD56    CD56    Antibody Capture
CD19    CD19    Antibody Capture
CD11c   CD11c   Antibody Capture
```

matrix.mtx.gz:

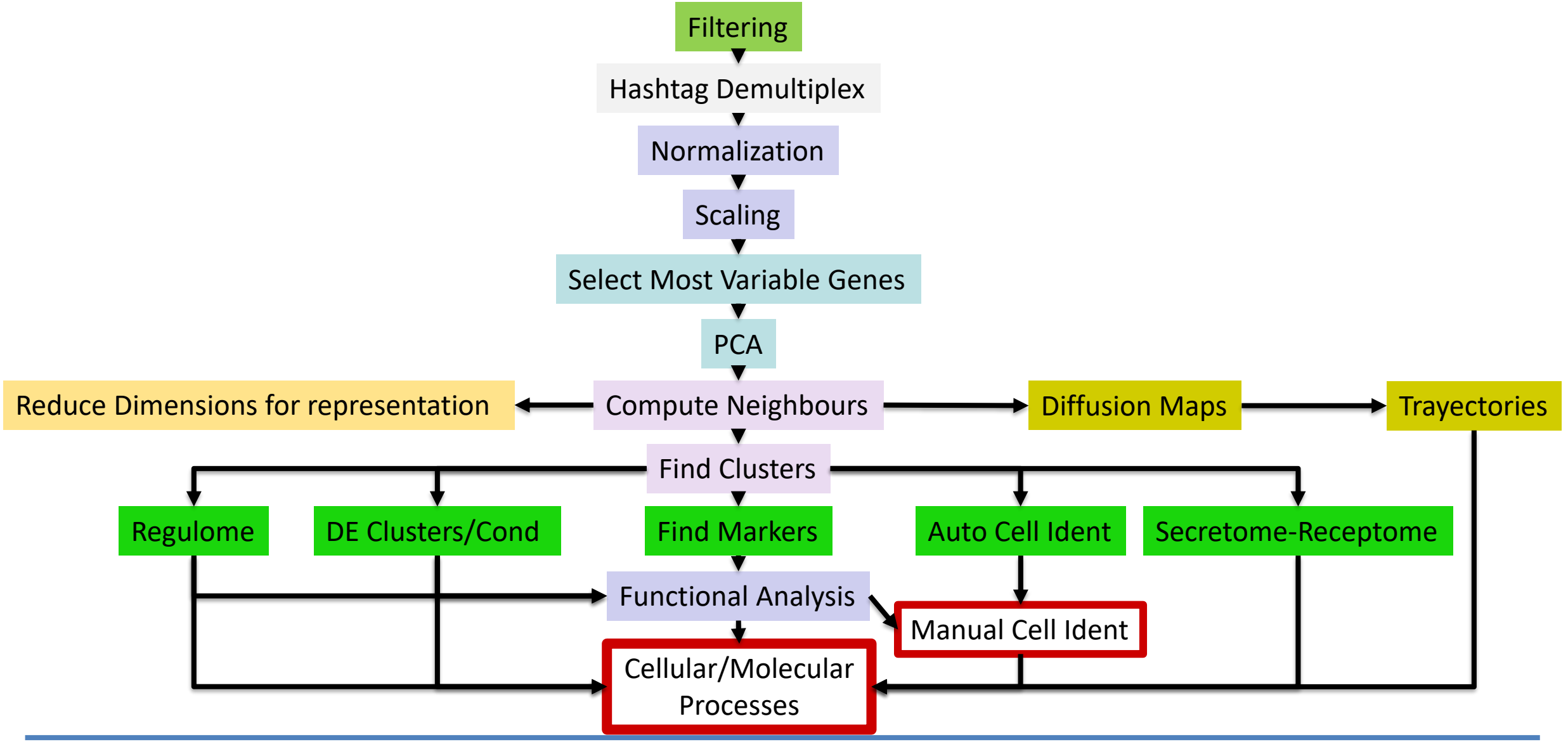
- C1 -> features (rows),
- C2 -> barcodes/cells (columns),
- C3 -> Value (value)

```
%%MatrixMarket matrix coordinate integer general
%metadata_json: {"software_version": "cellranger-6.1.1", "format_version": 2}
36636 30674 34324997
76 1 1
171 1 4
203 1 1
437 1 1
440 1 1
493 1 1
525 1 6
566 1 1
```

SPARSE MATRIX

```
ENSG00000224051 . . . . . 1 . .
ENSG00000169962 . . . . . .
ENSG00000107404 . . . . . .
ENSG00000162576 . . . . . .
ENSG00000175756 . . . 1 3 4 . . 1 . .
ENSG00000221978 . . . . . .
ENSG00000224870 1 . . . . . .
ENSG00000242485 1 . . . 6 . . 3 . . .
ENSG00000272455 . . . . . .
ENSG00000235098 . . . . . .
ENSG00000225905 . . . . . .
```

scRNA-Seq Workflow



scRNA-Seq Workflow

Filtering Cells:

Filtering of low quality cells helps to improve clustering by removing unwanted sources of variation

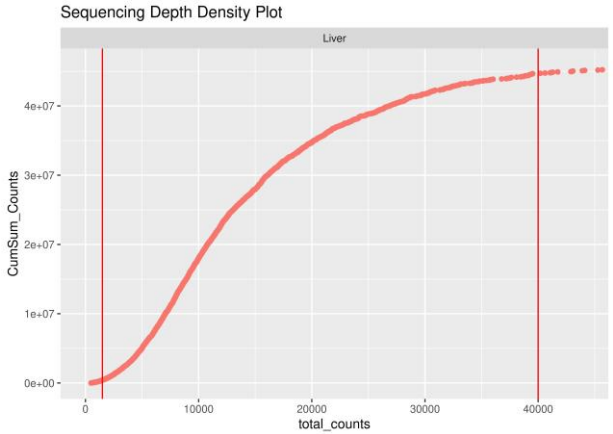
Typical filters used:

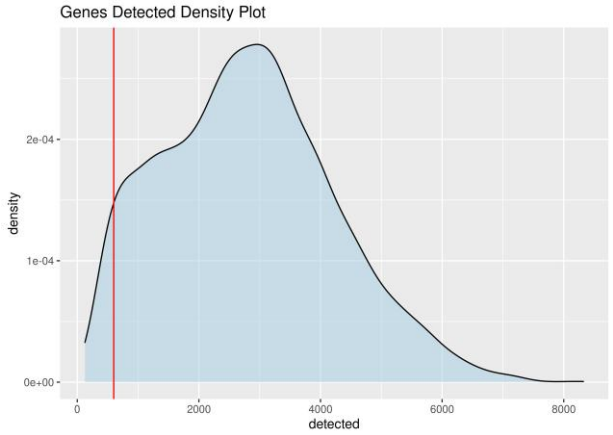
Counts:

- Low counts levels reduce the ability to profile the gene expression
- High counts may indicate multiplets (several cells captured in the same drop/well)

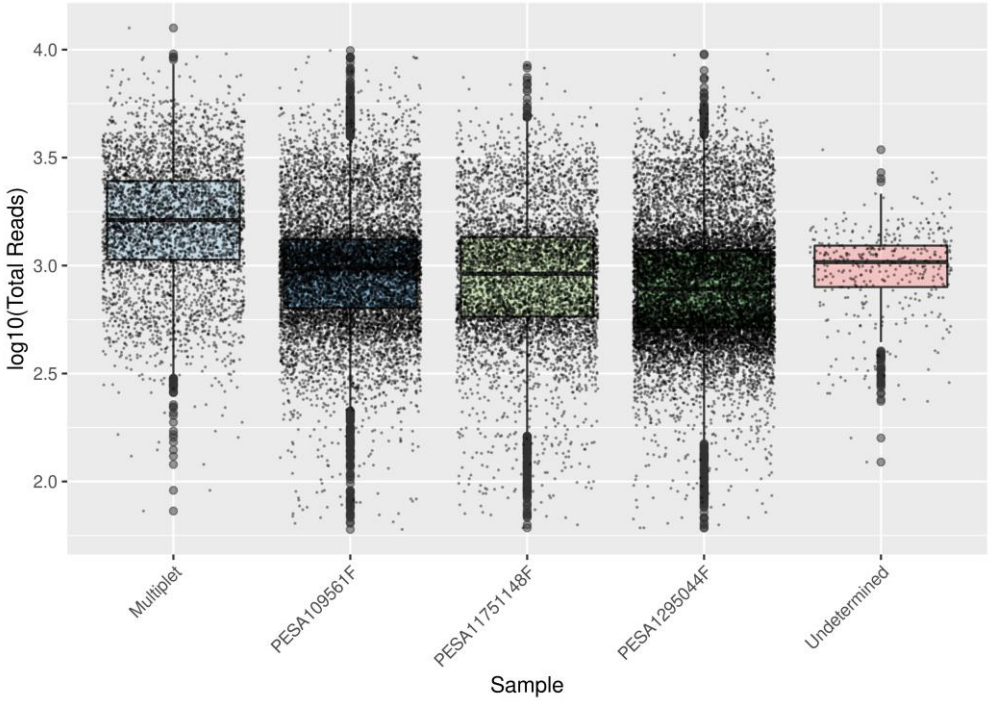
Genes:

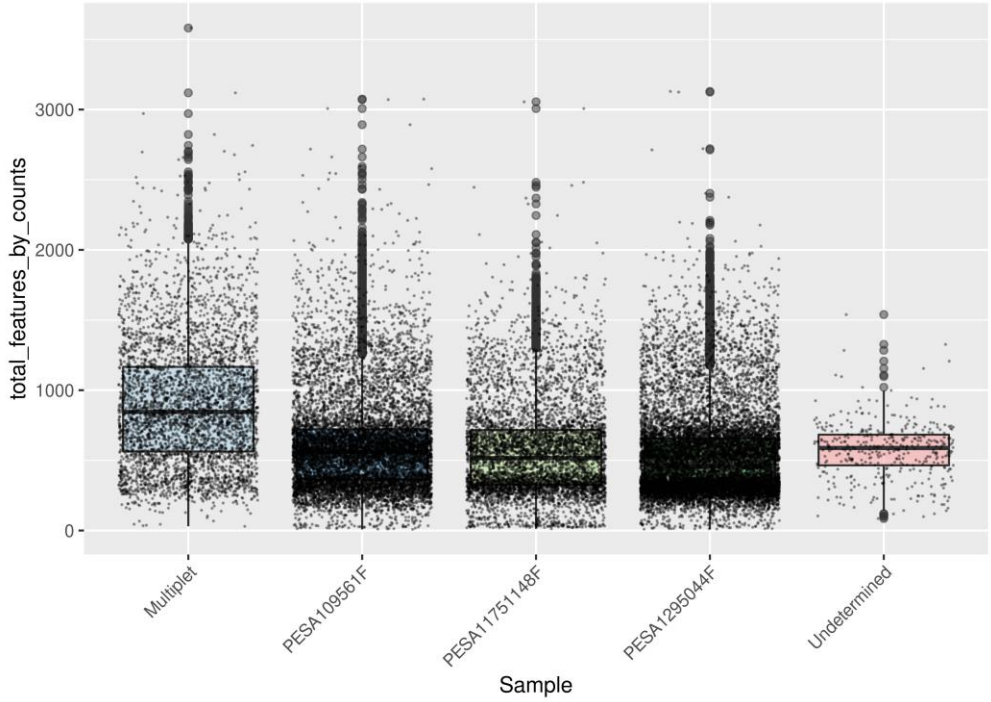
- Low level of genes detected may indicate problems in the cell viability or particular chemistry problems.
- High levels may indicate also multiplets (several cells captured in the same drop/well).
- Caution: Several cell types have particularly low levels of genes, like the neutrophils
- Caution: Several cell types have particularly high levels of genes like Undifferentiated or Stem Cells.





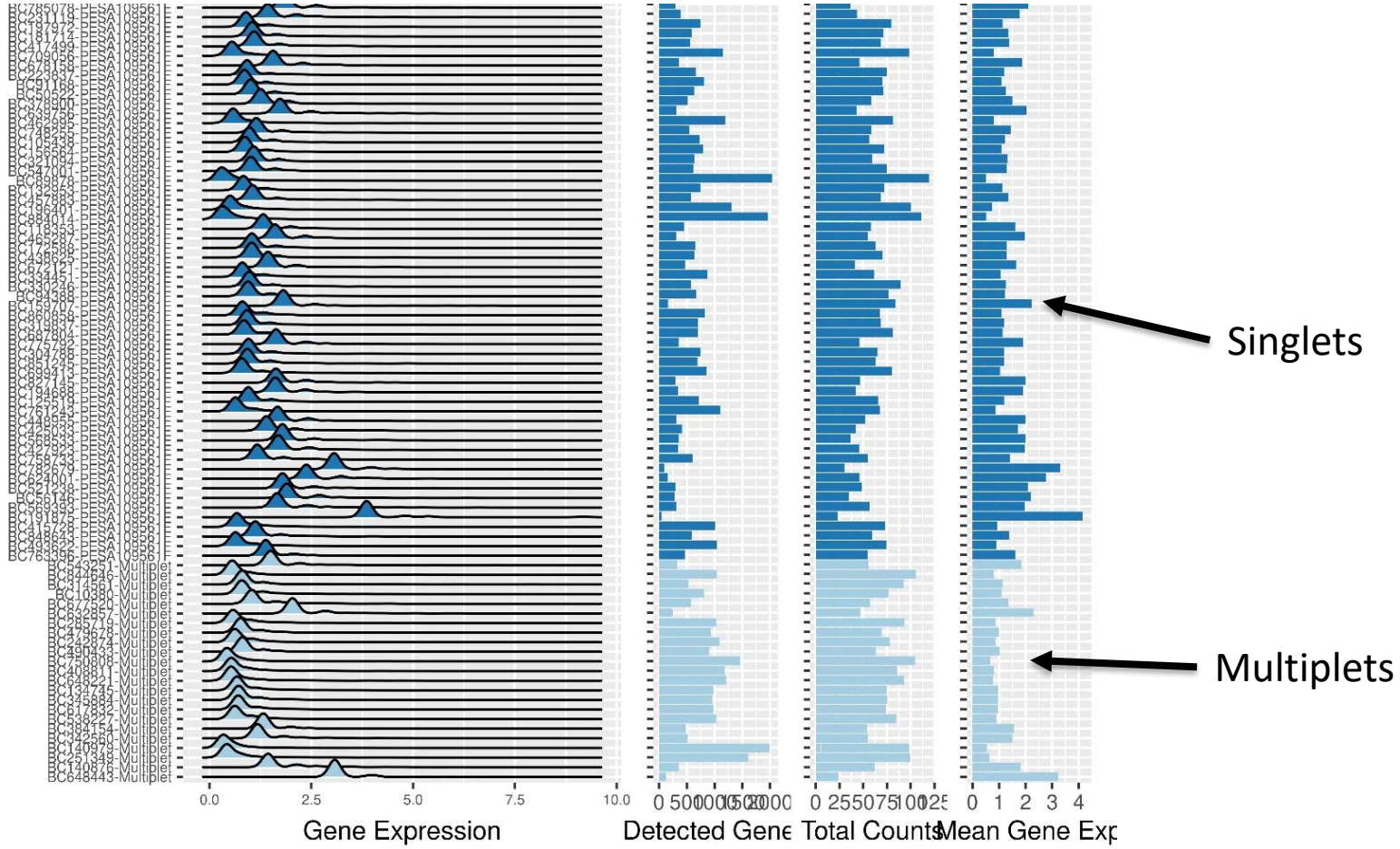
scRNA-Seq Workflow





scRNA-Seq Workflow

Counts, Genes Detection and Expression are related



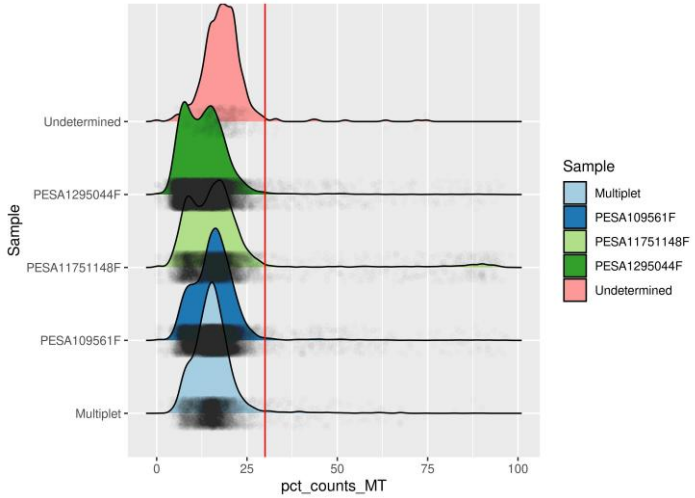
The more counts obtained from a cell, the more genes are usually captured and average gene expression decreases

scRNA-Seq Workflow

Filtering Cells:

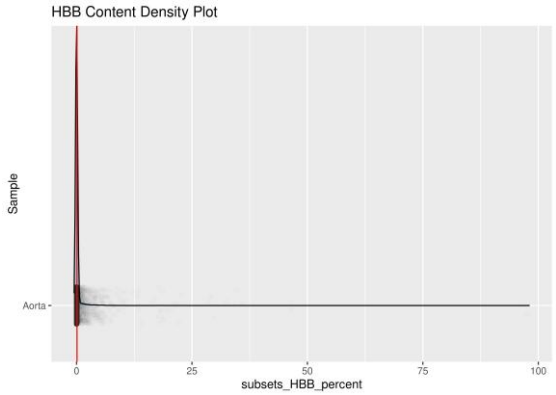
Mitochondrial Genes:

- In situations when a cell is dying or there is a problem with the chemistry, the mRNA is degraded or more sensitive than the Mitochondrial genes RNA and these becomes dominant in the profile.
- High levels of MT RNA indicates low quality cells (dying cells or low quality capture).
- Caution: Filters have to be adjusted according to the biology in study. Muscle cells have high levels of MT.



HBB Genes:

- There are experiments that involves vessels with erythrocytes that could contaminate the sample.
- Erythrocytes have high levels of HBB RNA and a few other RNAs and are very sensitive and easy to lyse.
- They may release their content and increase the background RNA to a point that an empty drop/well might be considered a cell.
- High content HBB cells are usually removed unless biologically relevant

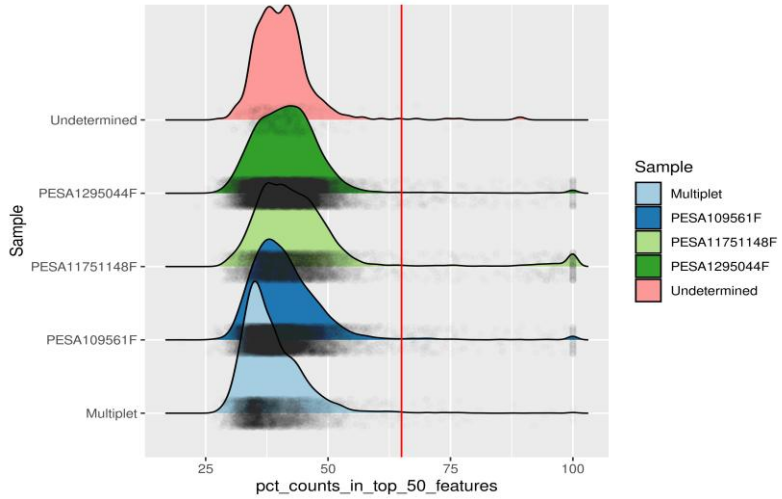


scRNA-Seq Workflow

Filtering Cells:

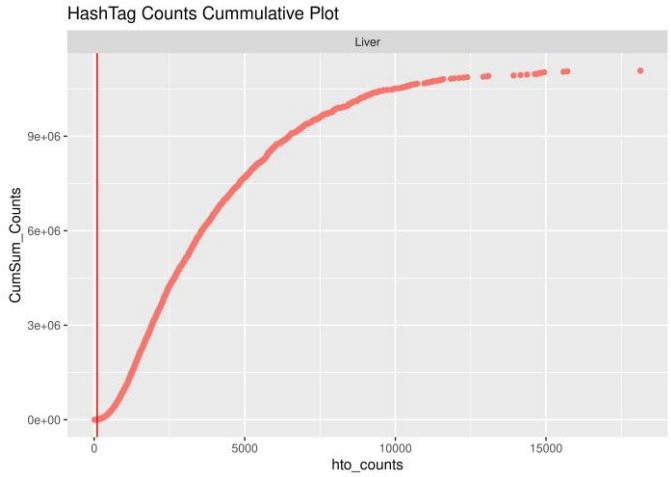
Gene Complexity:

- The idea is to detect those cells that having many reads, and sufficient genes detected, still they have very few genes with most of the reads.
- Those cells are also very low quality and usually reflects a RNA capture problem or a dying cell.

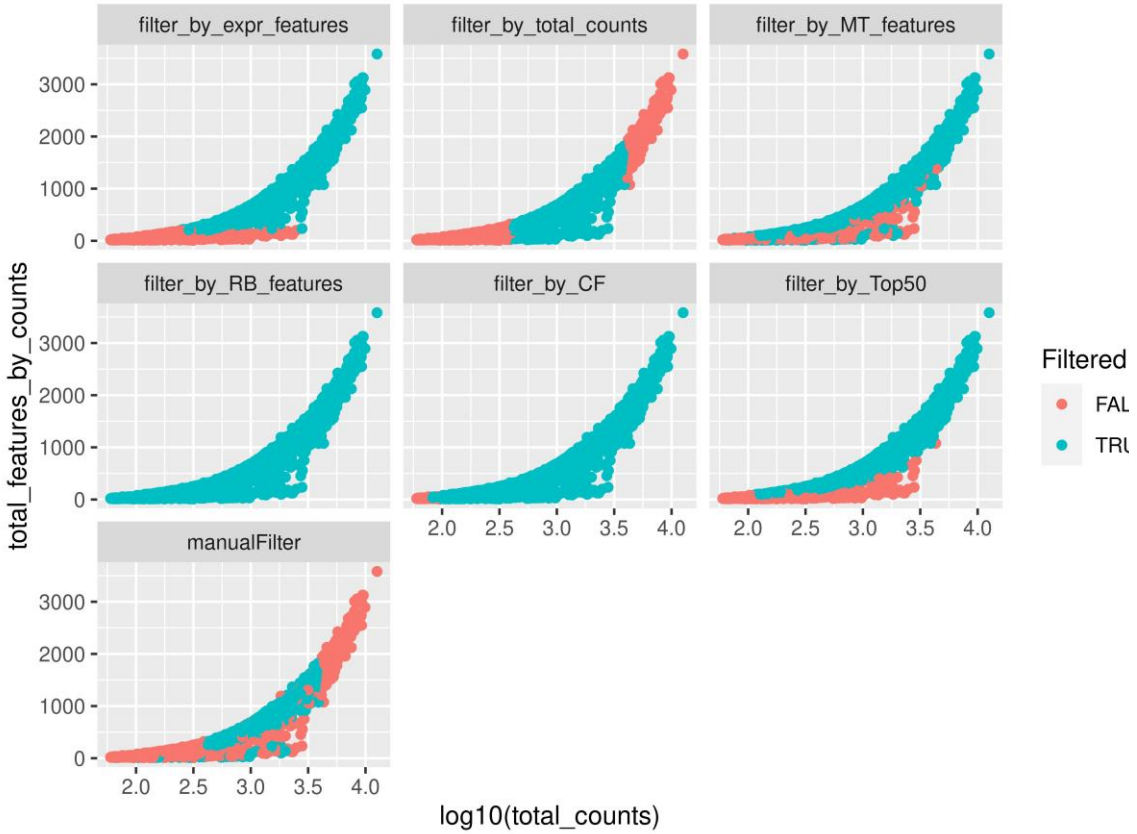


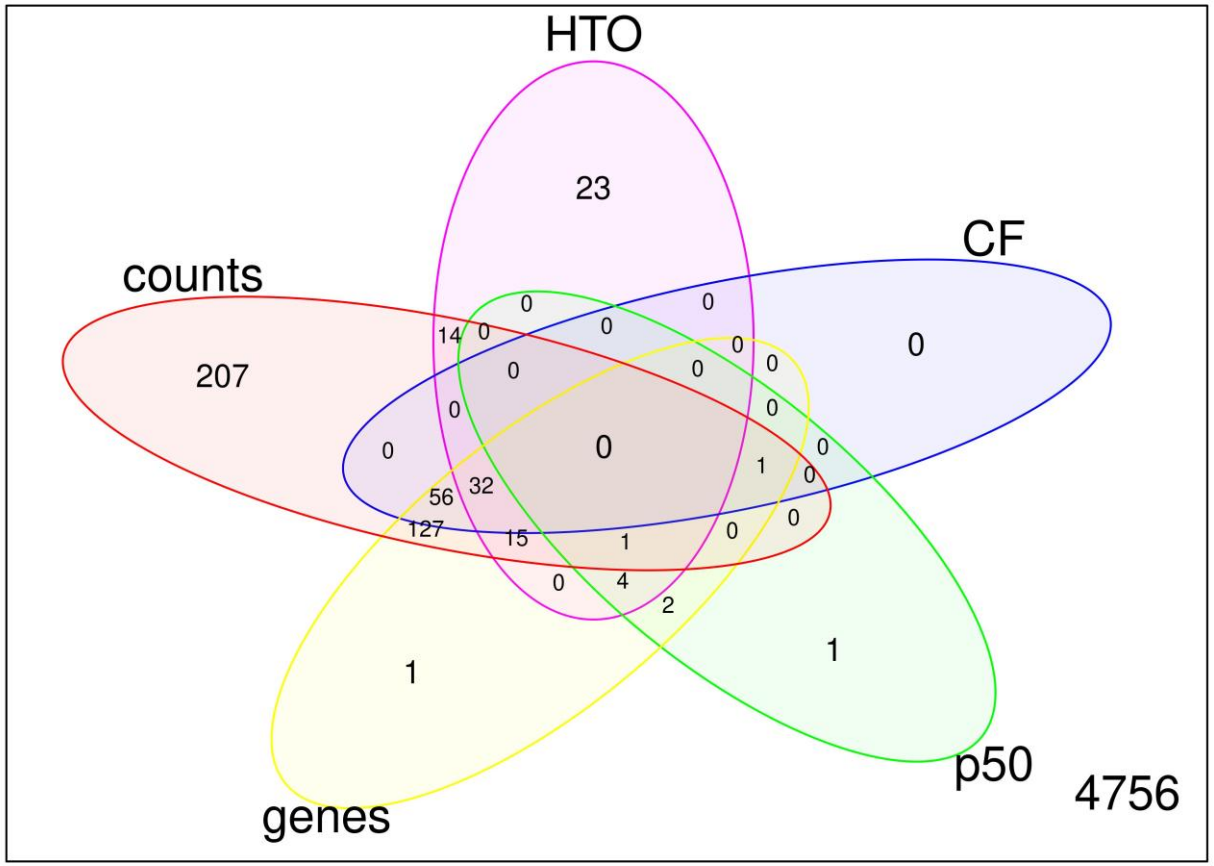
Hashtag Content:

- When there are several samples pooled in the experiment and they are labelled with antibodies/hashtags, a hashtag filter is also recommended.



Summary Of Cell Filtering Process





scRNA-Seq Workflow

Filtering Genes

- The idea is to reduce the number of genes to those more relevant to increase efficiency.
- Obviously we will remove all the genes with 0 counts across the experiment, but sometimes is also good to remove some that are only detected in a few cells.
- This will depend on the total number of cells in the experiment and minimum amount of cells expected to be part of a cluster.
- In many cases we don't know whether we are going to have small relevant clusters.
- So, to start with, just remove those present in no more than 10 cells. Later the filter can be increased or reduced

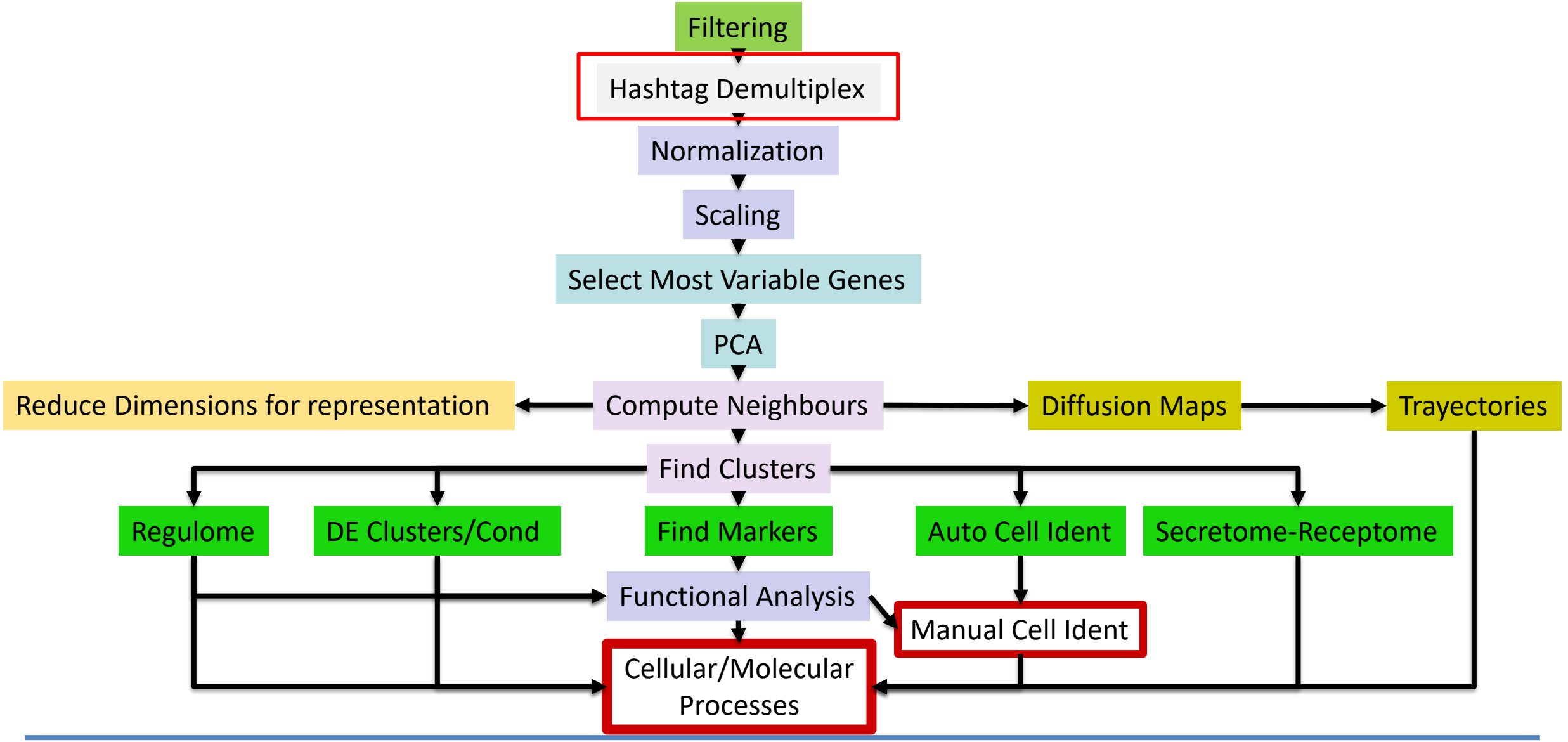
A normal situation:

Starts with 6909 cells with 32288 genes

Cell filtering ↓ Gene filtering

Ends with 5636 cells with 16567 genes

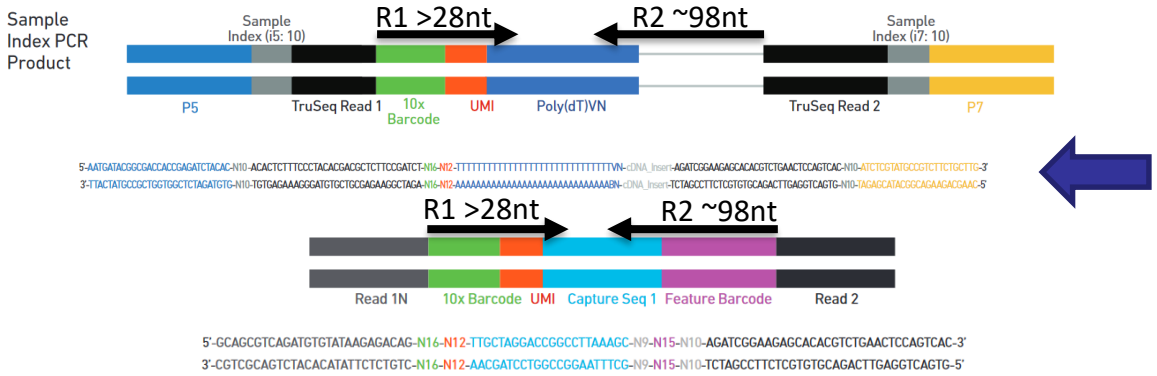
scRNA-Seq Workflow



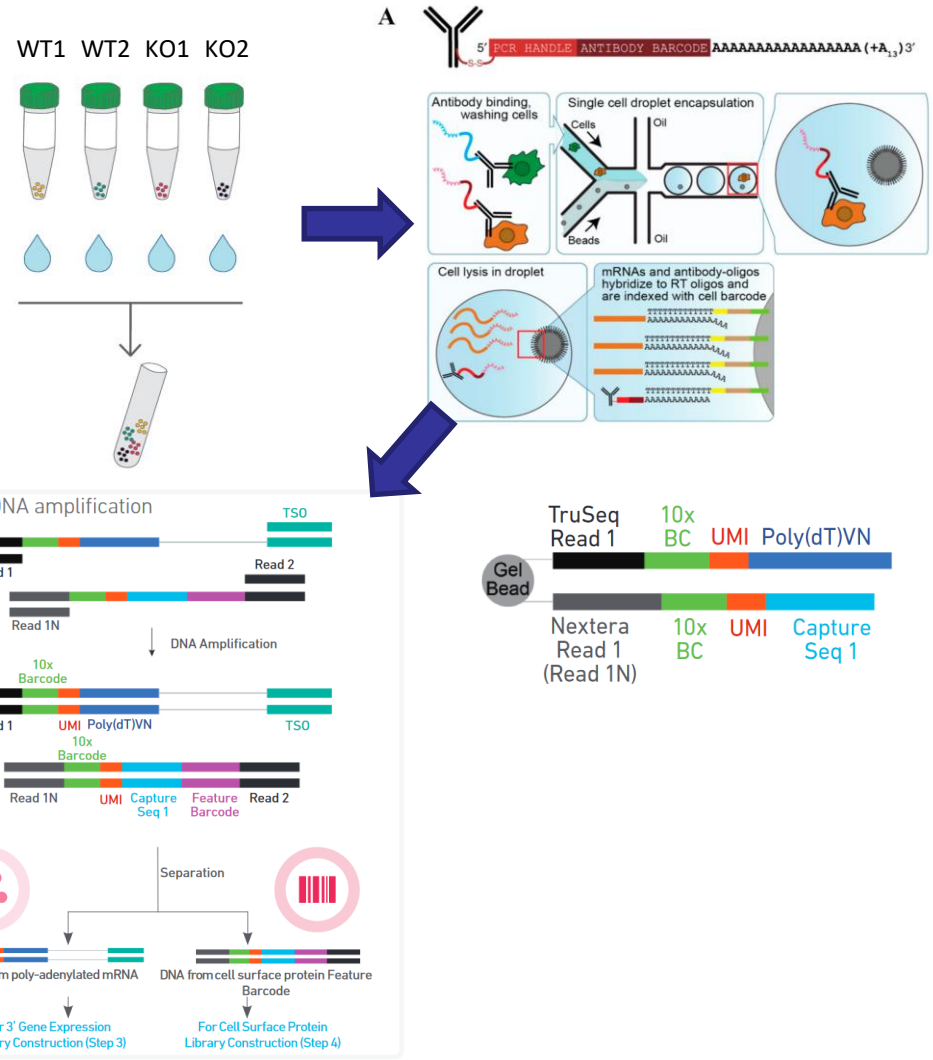
scRNA-Seq Workflow

Sample Hashtag Multiplexing:

- Several samples can be processed in one experiment by labelling the cells of each sample with an antibody against an ubiquitous membrane antigen.
- This reduces the capture and library preparation costs.
- Sample Hashtag (Sample Ab-barcode) libraries are prepared alongside but independently of mRNA libraries once cDNA is produced and amplified.
- Hashtag libraries are sequenced together with the mRNA derived libraries and sequences from both libraries are demultiplexed based on Illumina Sample barcoding system.



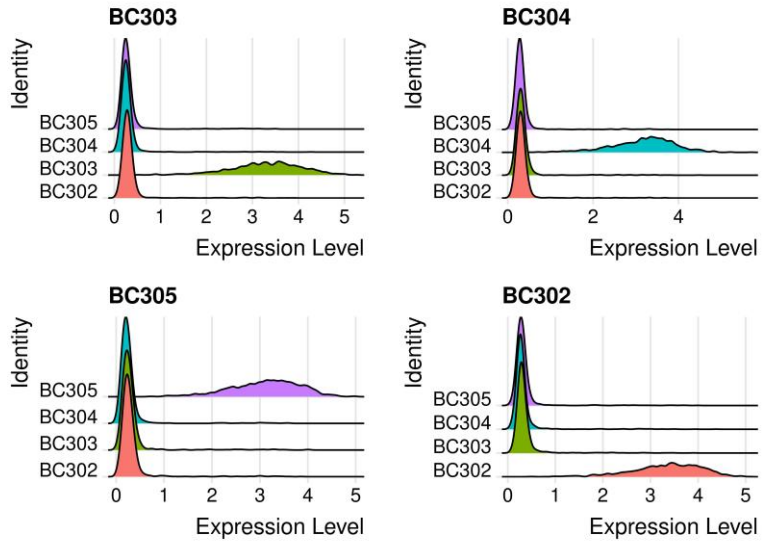
Single Cell Samples:

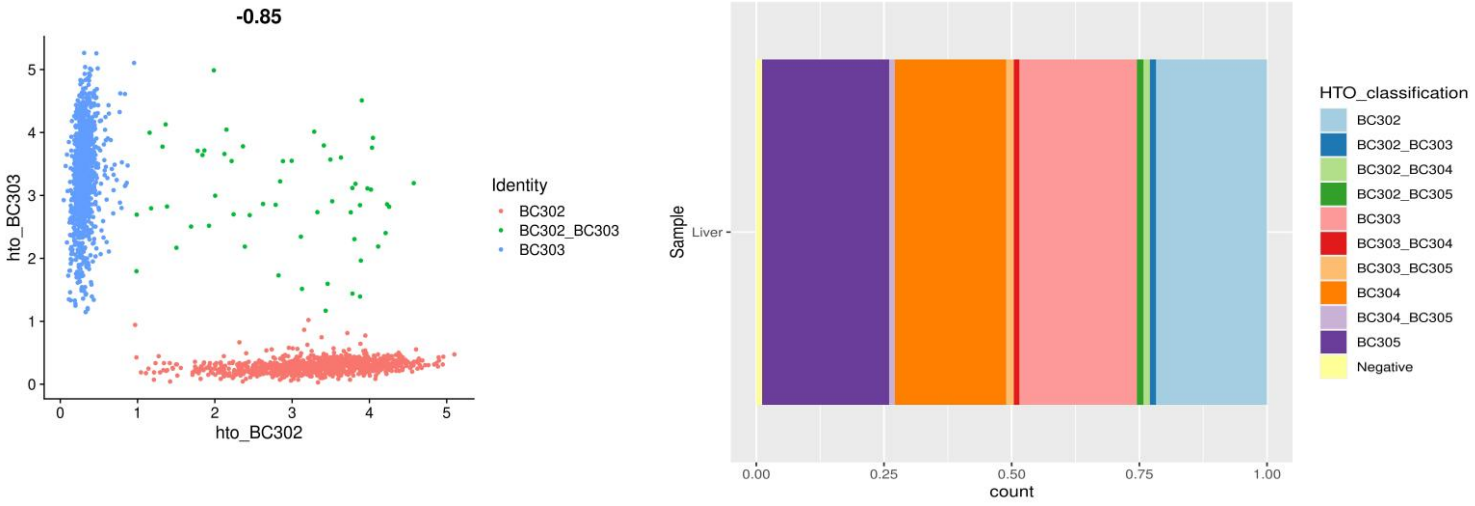


scRNA-Seq Workflow

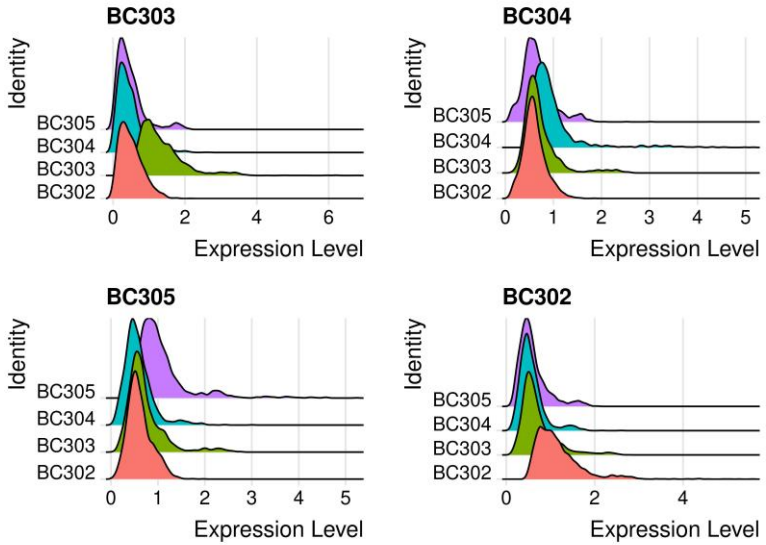
Cell Sample Assignment

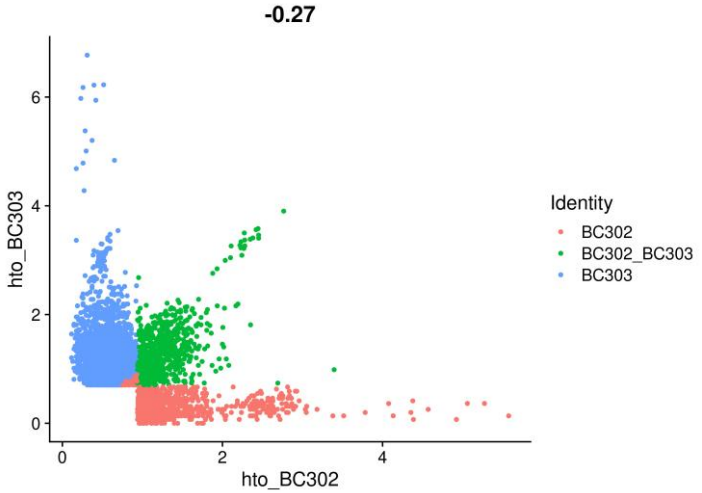
- R2 Sequence is annotated with Cell Barcode and UMIs from R1
- R2 are mapped to the Hashtag sequences used to label each sample
- Counts are generated in the same way as the genes: unique UMI/Barcode/Hashtag combination
- Only detected cell barcodes based on gene expression are kept.
- A background model is generated for each hashtag based on the expression of all the hashtags minus the highest expressed on each cell.
- Hashtags significantly far from this background model are considered as detected.
- In this way cells are labelled according to the hashtags detected

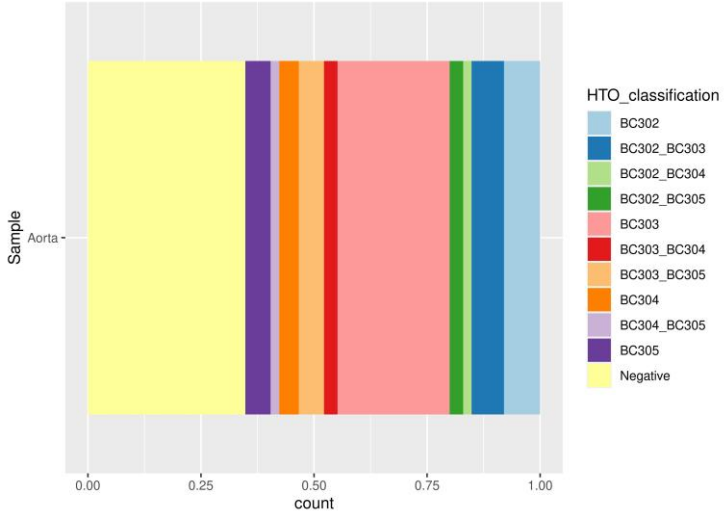




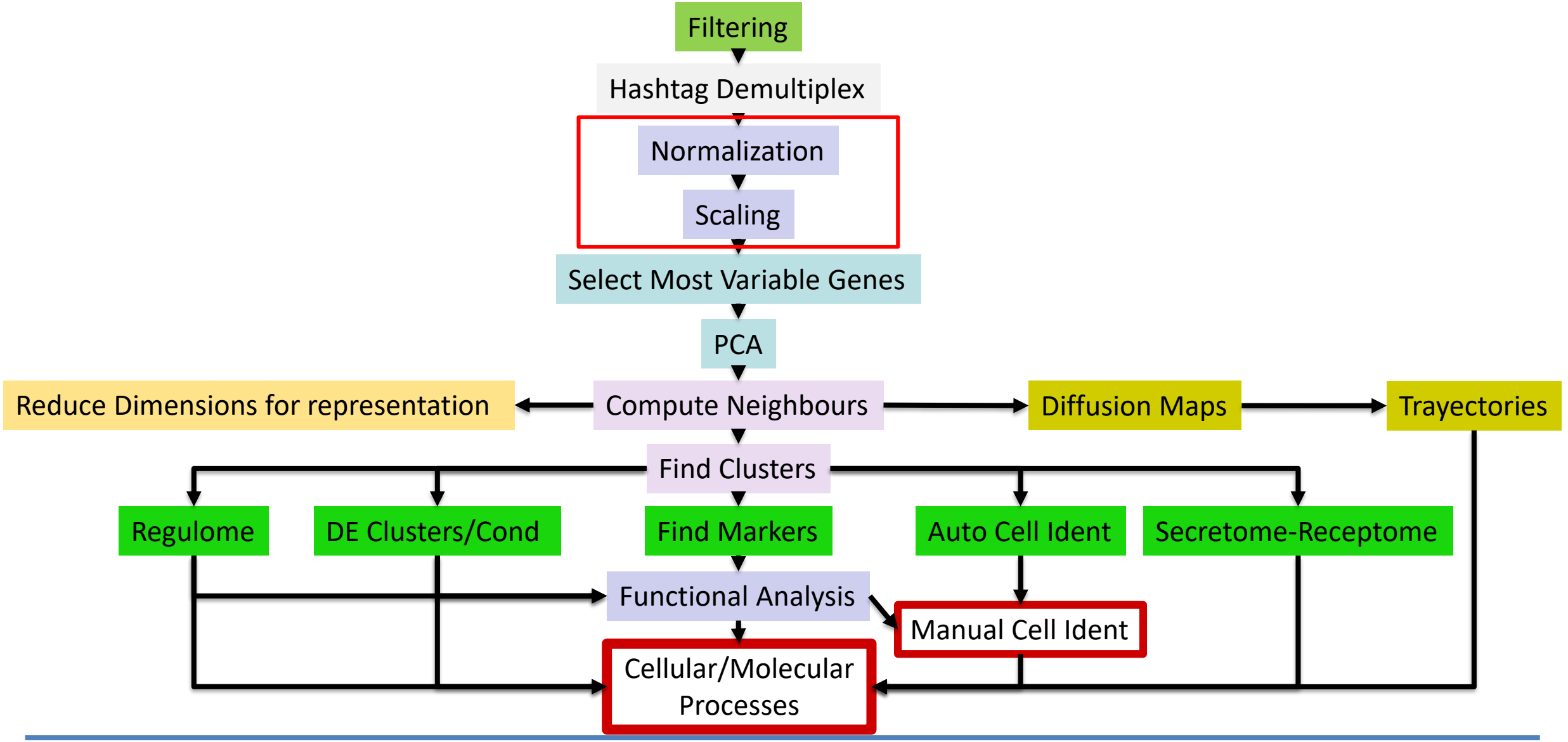
scRNA-Seq Workflow







scRNA-Seq Workflow



→ ERCC Control Sequences: Set of 96 mammal like sequences not present in model genomes used to control efficiency of the protocols

→ Normalization algorithms are able to use the information from these control sequences to estimate technical variability between replicates and batches and correct by that.

Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression

Jong Kyoung Kim[1], Aleksandra A. Kolodziejczyk[1,2], Tomislav Illicic[1,2], Sarah A

Normalization of RNA-seq data using factor analysis of control genes or samples

Davide Risso[1], John Ngai[2–4], Terence P Speed[1,5,6] & Sandrine Dudoit[1,7]

→ New normalization algorithms are designed to reduce the effect of the vast amounts 0 values like those based on pooling and deconvolution techniques.

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

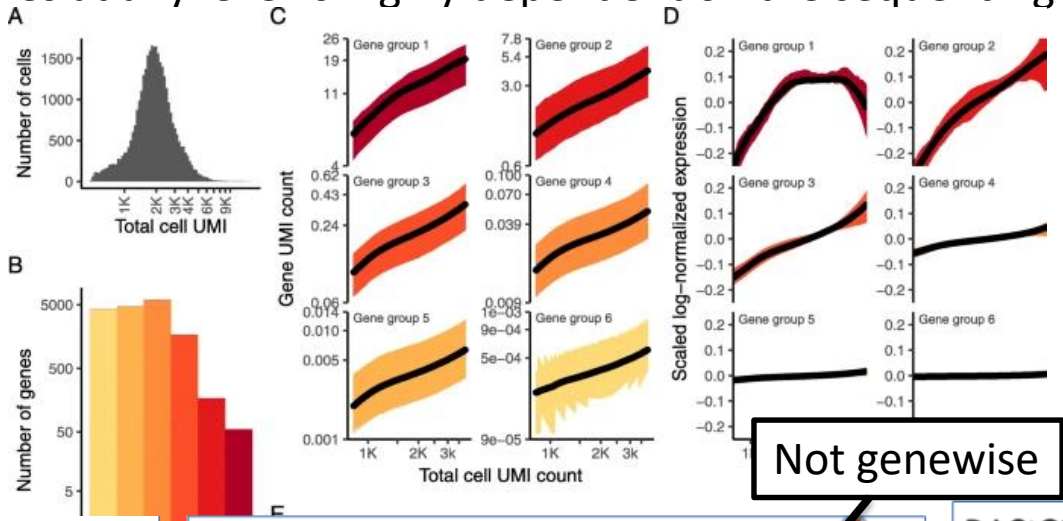
Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

BASiCS: Bayesian Analysis of Single-Cell Sequencing Data

scRNA-Seq Workflow

Normalization & Scaling

Normalization on bulk samples is based on the fact that most of the genes do not change across replicates and conditions.
In SC experiments is close to the opposite. Most of the genes detected are different between cells.
Also, the expression of genes at any level is highly dependent on the sequencing depth of the cell.



Normalization of RNA-seq data using factor analysis of control genes or samples

Davide Risso[1], John Ngai[2–4], Terence P Speed[1,5,6] & Sandrine Dudoit[1]

Need spike-ins

Pooling across cells to normalize single-cell RNA sequencing data with many zero counts

Aaron T. L. Lun[1*], Karsten Bach[2] and John C. Marioni[1,2,3*]

Not genewise

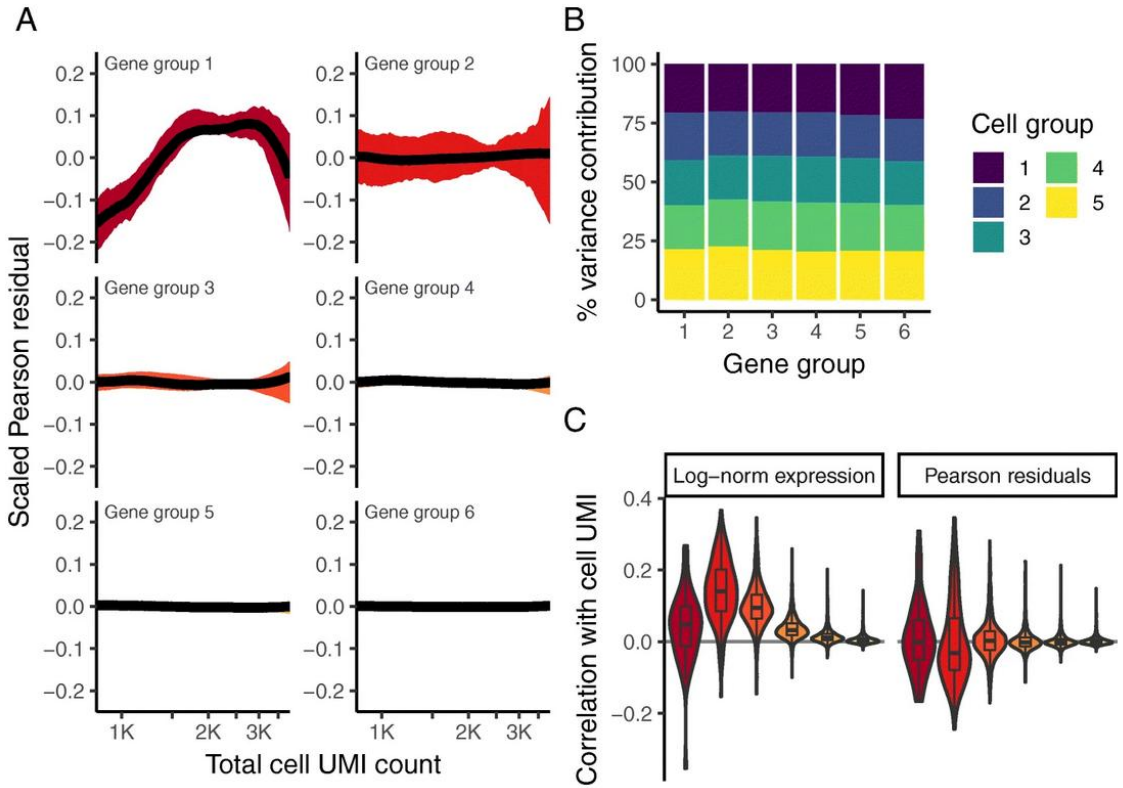
BASiCS: Bayesian Analysis of Single-Cell Sequencing Data

Overfit

scRNA-Seq Workflow

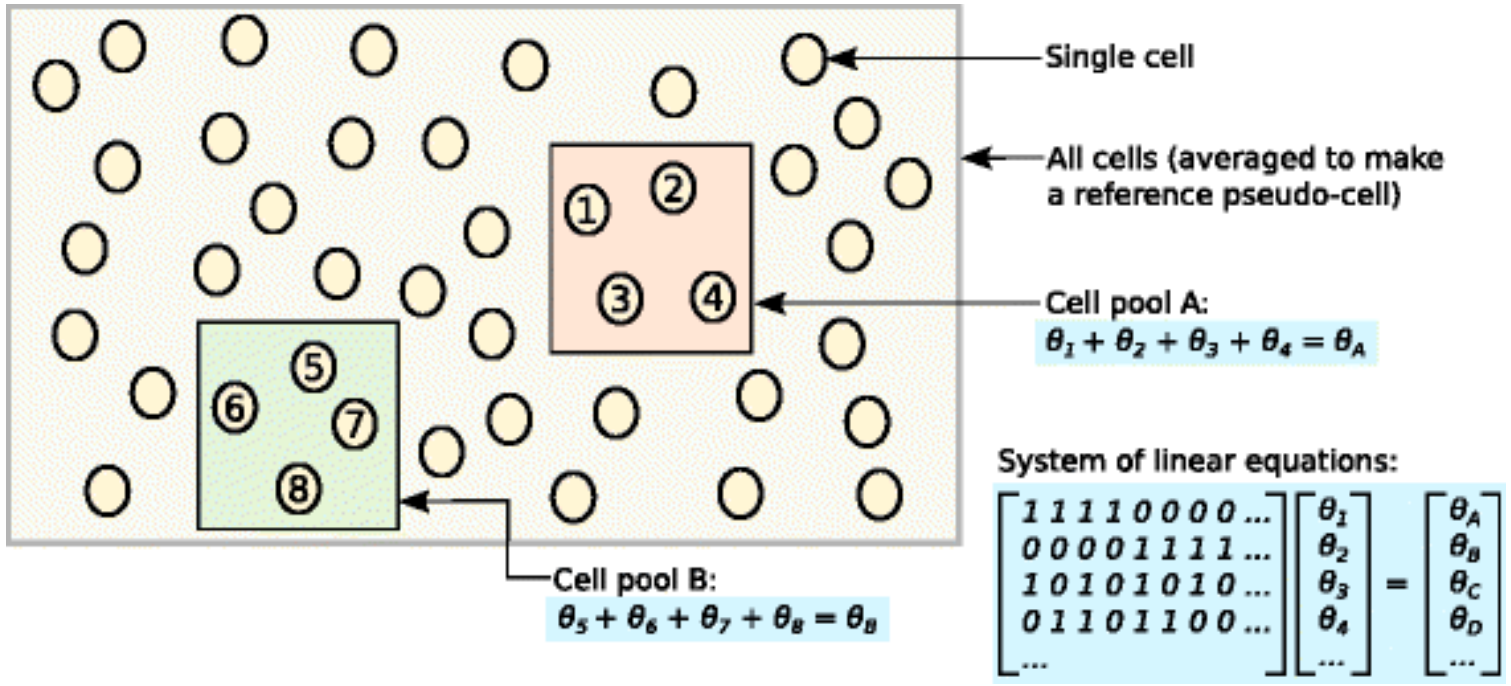
Normalization & Scaling

Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression



Hafemeister, C., Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol* **20,** 296 (2019). https://doi.org/10.1186/s13059-019-1874-1

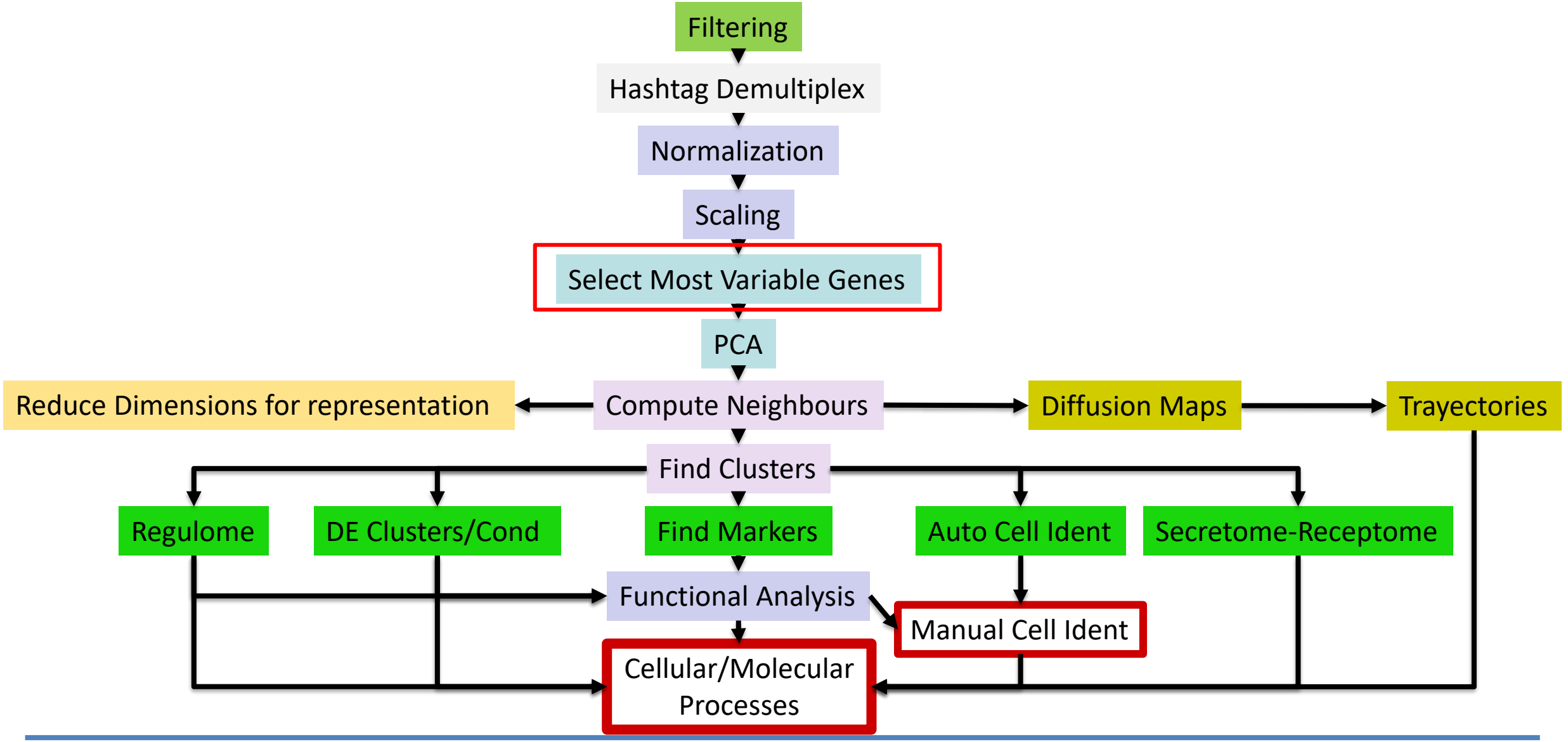
Pooled Normalization



- Each cell is pooled many times with other different cells from different experimental groups giving a set of linear equations from which deconvolute a normalization factor for each cell.
- Incorporates intrinsically a batch correction as each cell will be combined with cells from other batches in each pool.

L. Lun, A.T., Bach, K. & Marioni, J.C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* **17**, 75 (2016). https://doi.org/10.1186/s13059-016-0947-7

scRNA-Seq Workflow



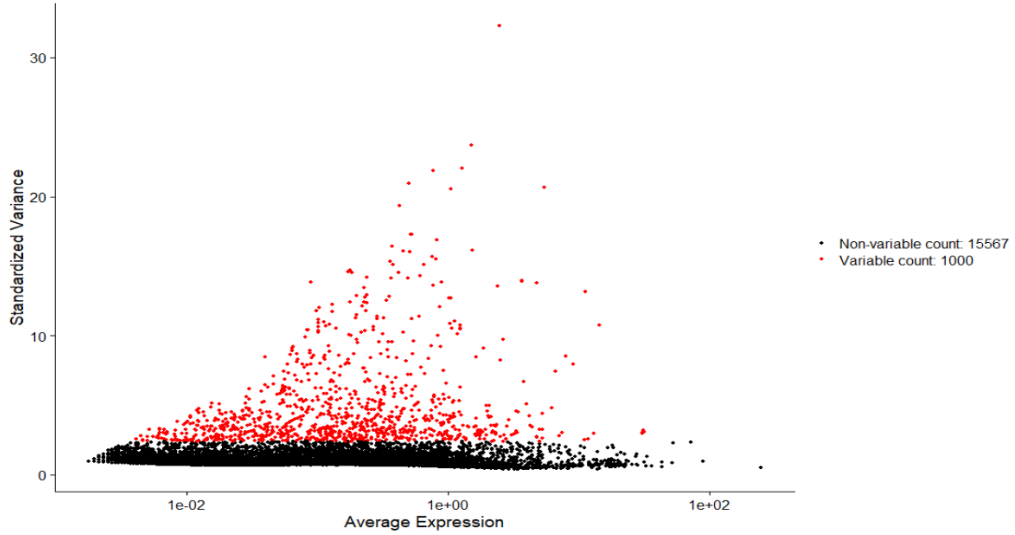
scRNA-Seq Workflow

Most Variable Genes:

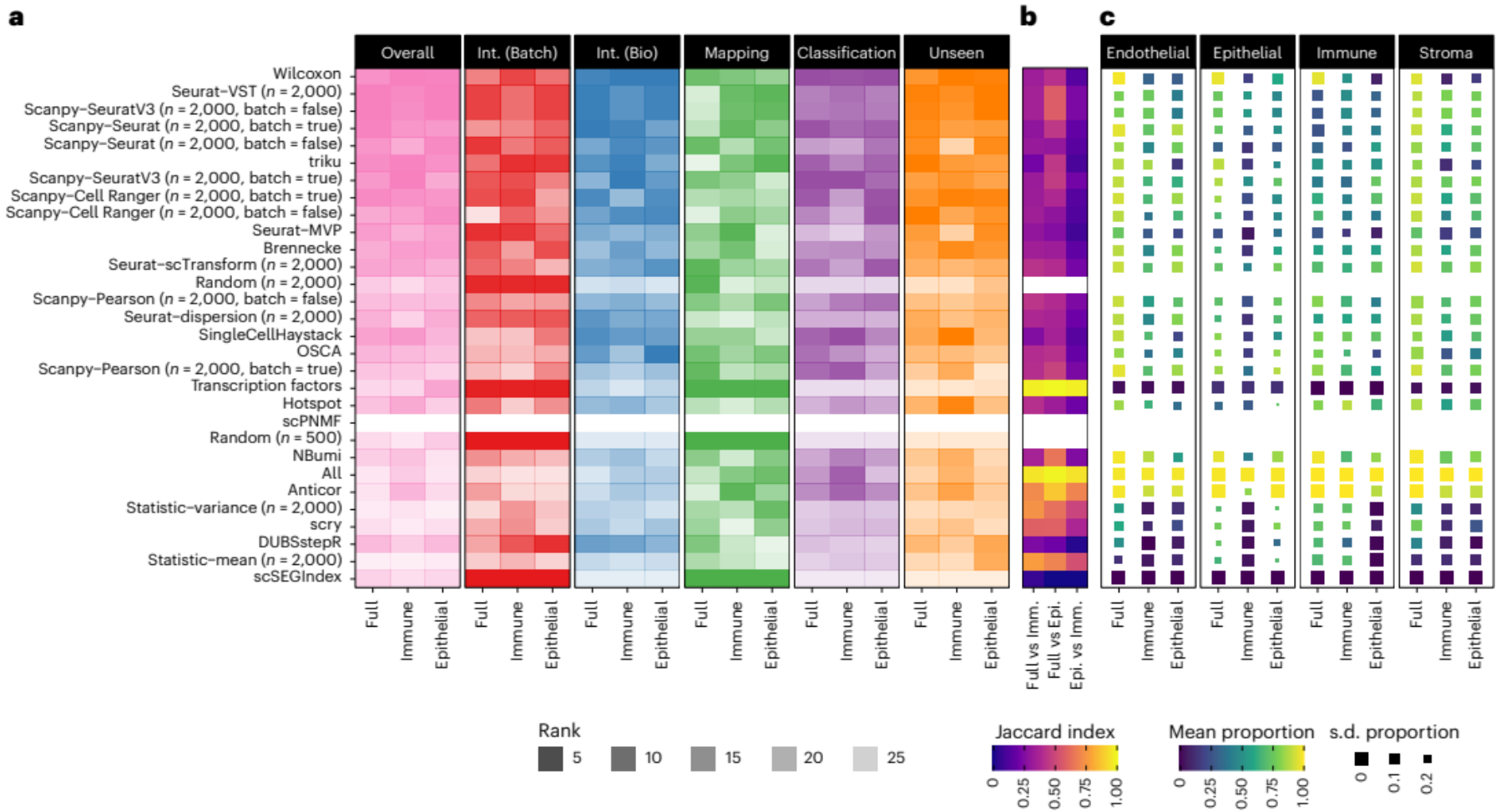
To reduce the amount of noise and increase computational efficiency there is a feature selection step to reduce the number of dimensions/features to use in the clustering analysis.

Several approaches are possible:

- Using some cut-offs in the mean expression and standard deviation
- Selecting the n top genes with the highest stdev (i.e. 2000 genes).

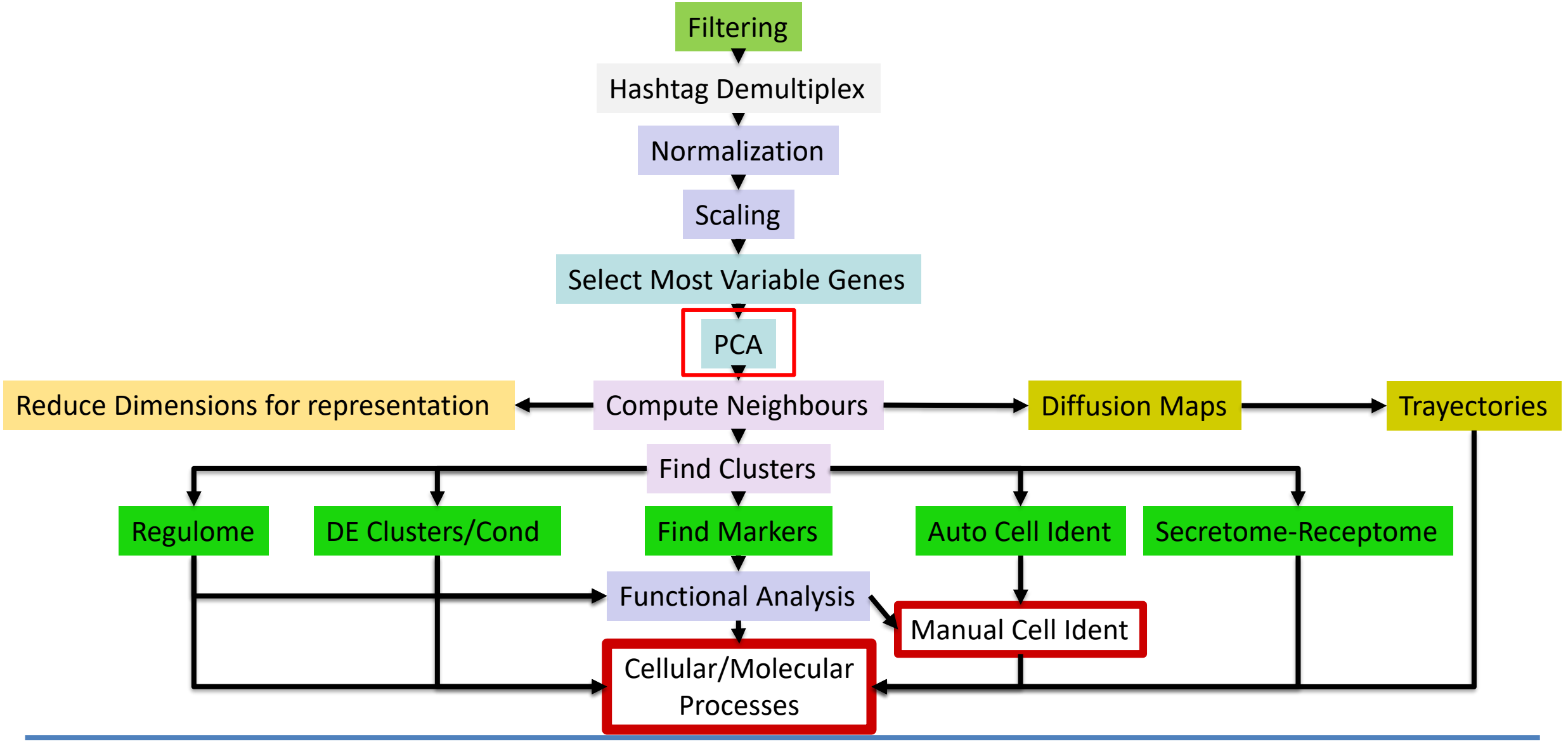


Select Most Variable Genes



Zappia L, Richter S, et. al., and Theis FJ. Feature selection methods affect the performance of scRNA-seq data integration and querying. Nat Methods. 2025 Apr;22(4):834-844. doi: 10.1038/s41592-025-02624-3. Epub 2025 Mar 13. PMID: 40082610; PMCID: PMC11978513.

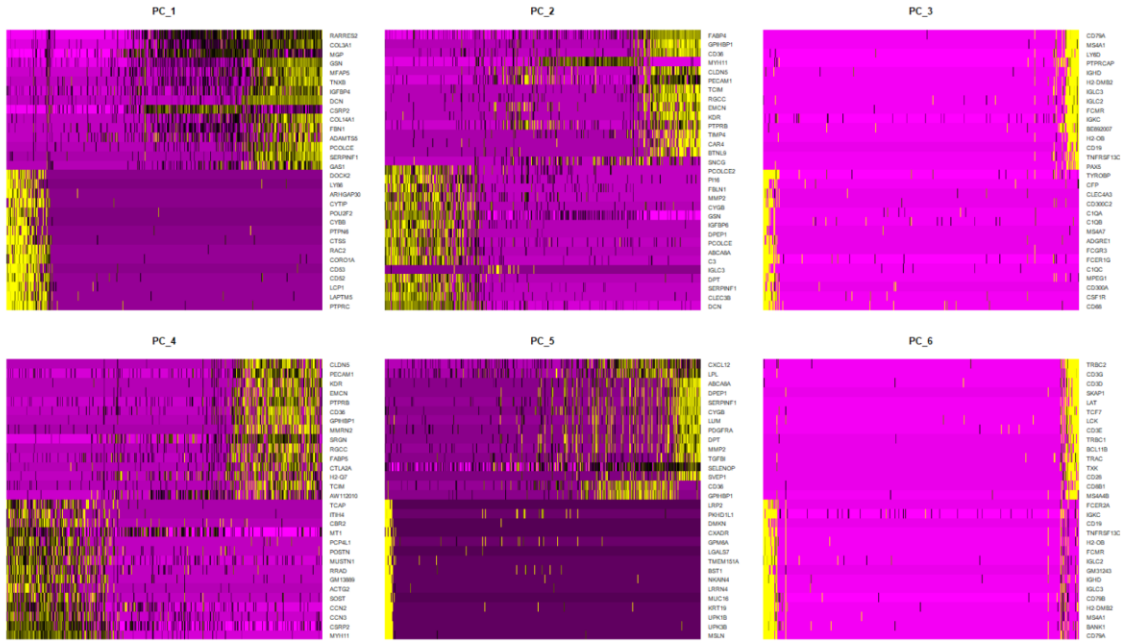
scRNA-Seq Workflow

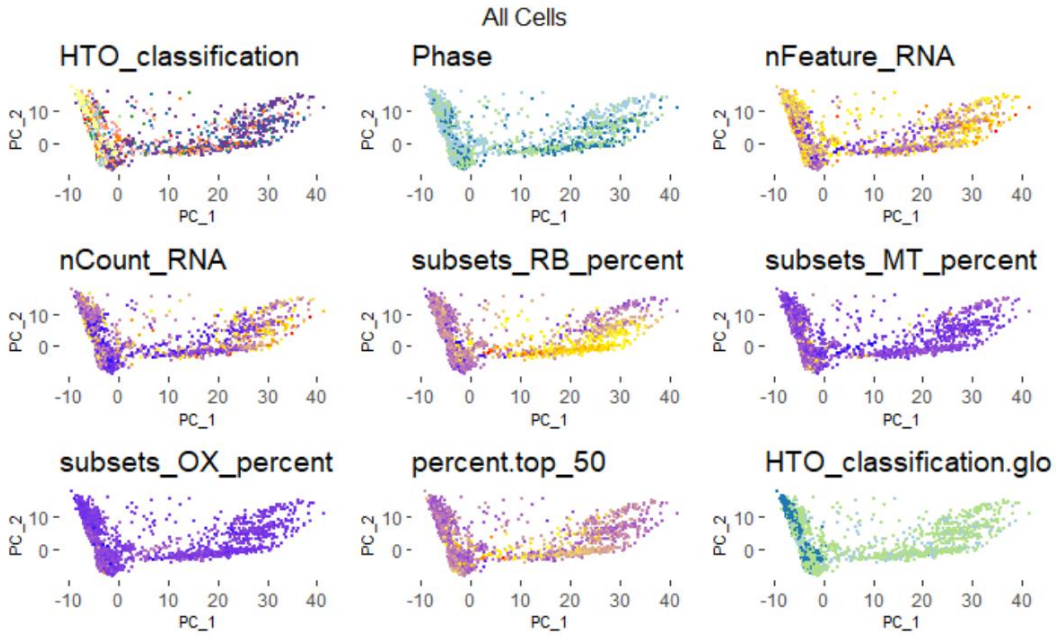


scRNA-Seq Workflow

PCA

- A Principal Component Analysis is at the basis of all following steps.
- The goal is to reduce even further the dimensions before all posterior clustering and further dimensionality reductions.
- Now the dimensions over which clustering will work are the different components. They summarise the dimensions in which the experiment express its higher variability.

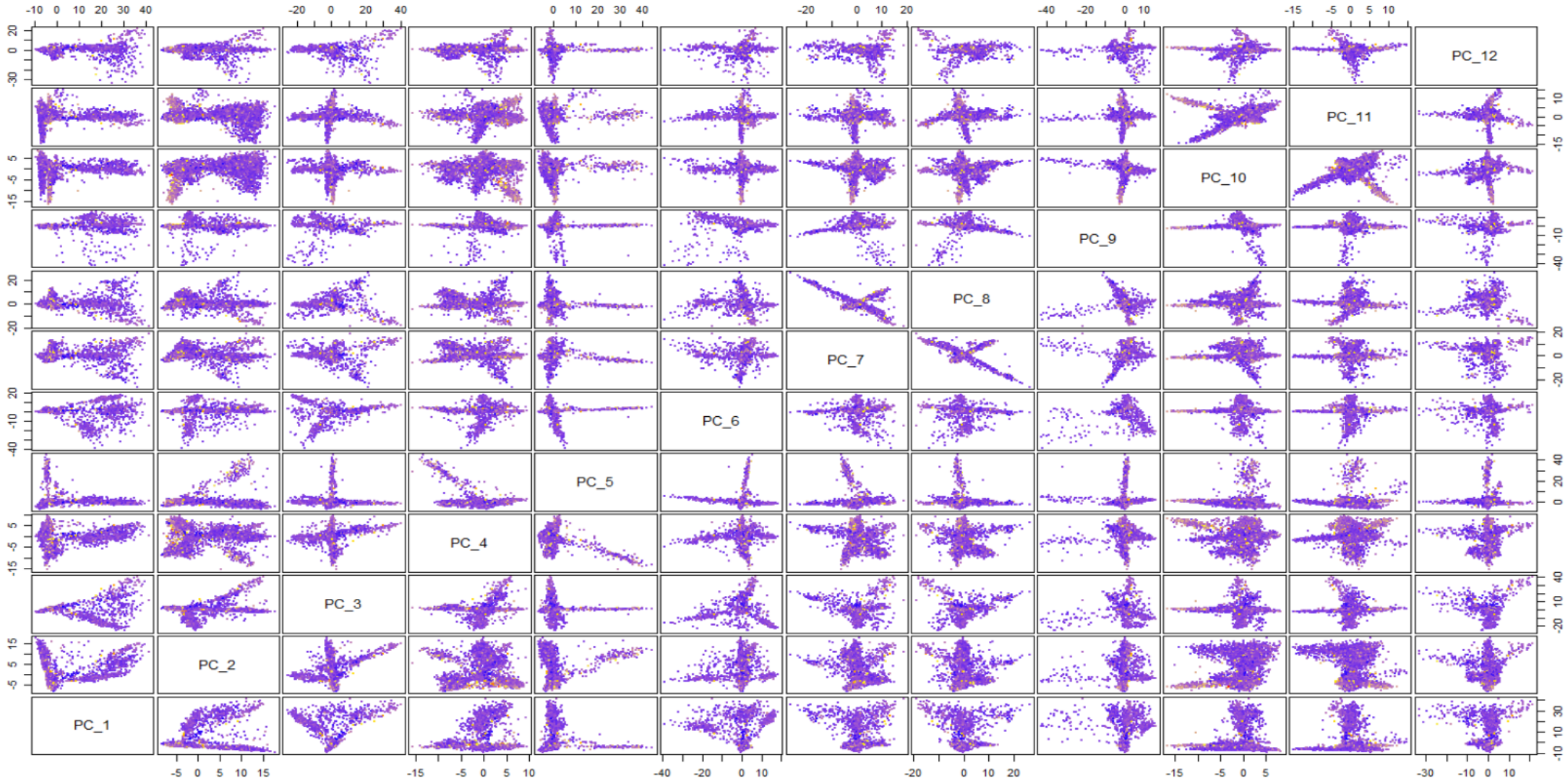




We can star exploring which genes are relevant in the experiment or maybe some thecnical variability

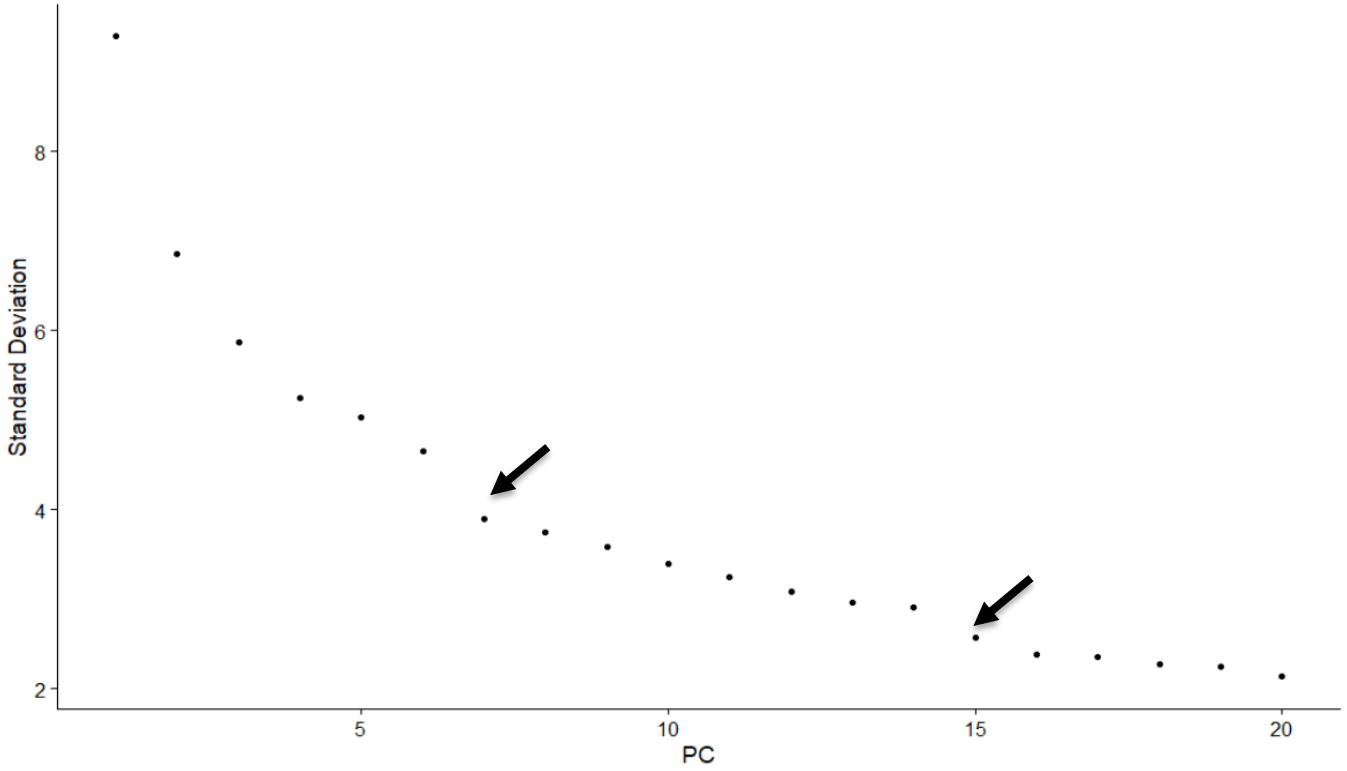
scRNA-Seq Workflow

Cell MT Content across PCAs

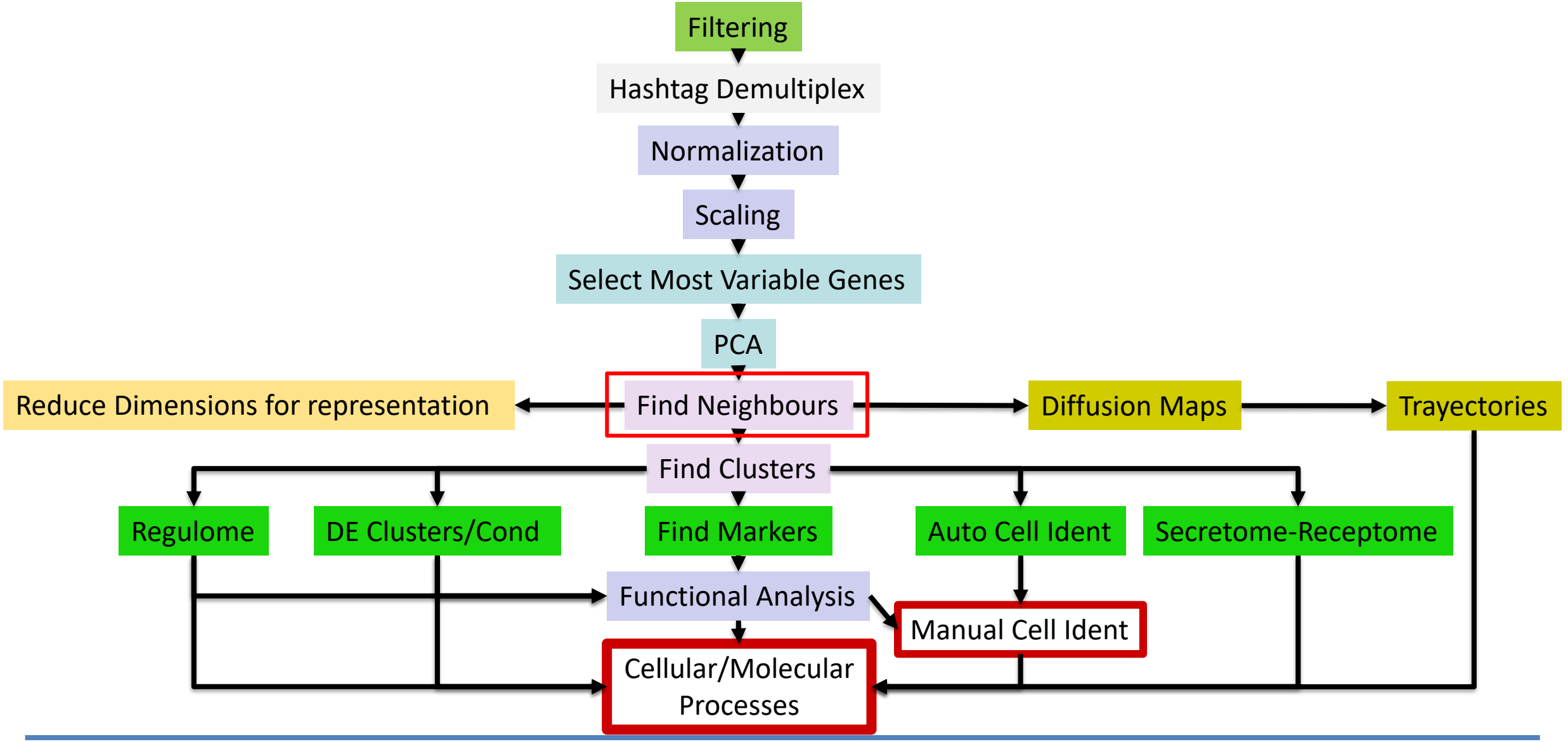


PCA

Again, to reduce spurious noise, not all PCs from the PCA are used and a PC selection is carried out again.
To select the most relevant PCs we usually evaluate the variability explained by each PC and select those PCs that accounts for most of the variability looking at places where high decreases are observed.



scRNA-Seq Workflow

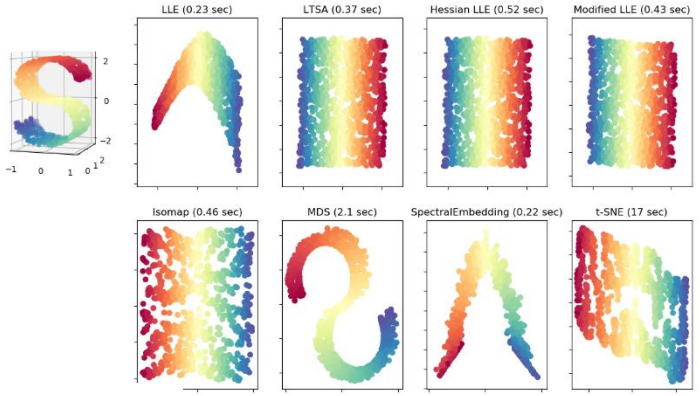


Tools for Computer-Aided Analysis

❑ Visualization

Dimensionality reduction techniques

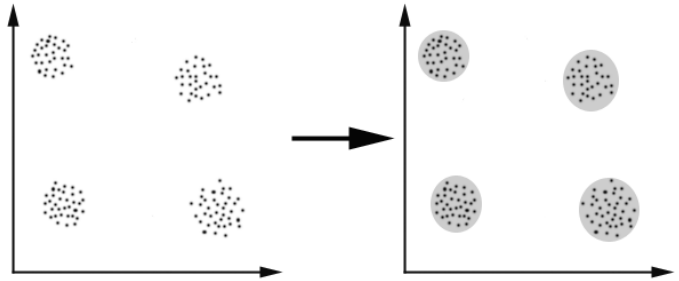
- ❑ Linear → PCA, LDA,…
- ❑ Non-linear (manifolds) → tSNE, UMAP, Isomap, LLE, MLLE, Spectral embedding, MDS…

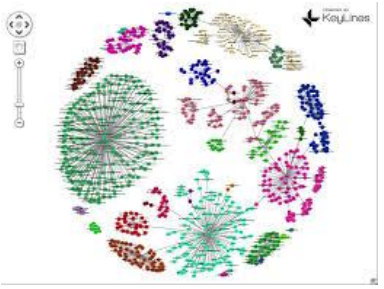


❑ Detection of Populations

Clustering Algorithms

- ❑ Supervised/Informed → Kmeans, Hierarchical, GMM,…
- ❑ Unsupervised/Uninformed → Mean-shift, X-shift, Phenograph, ClusterX



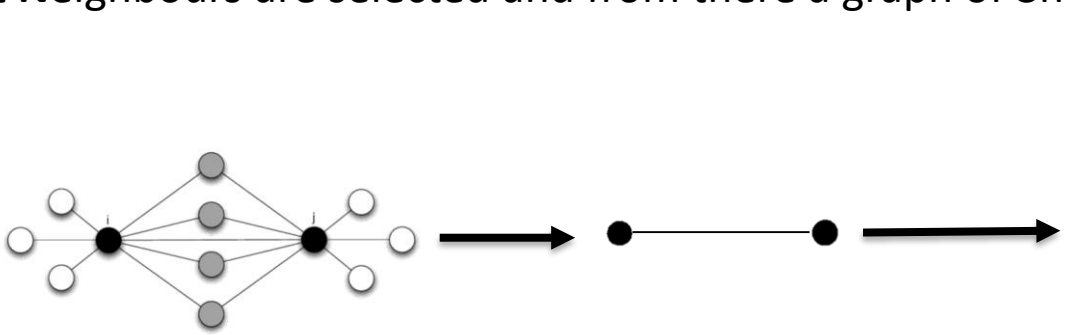


scRNA-Seq Workflow

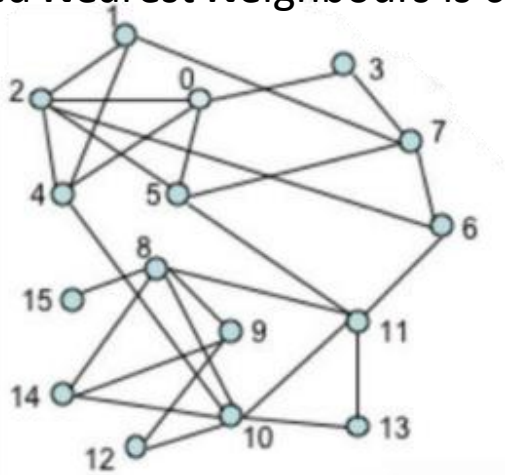
Compute Neighbours

The goal is to define in the multidimensional space which cells are next to each other. Find the neighbours of each cell.
All posterior analysis, clustering, dimensionality reduction, trajectories,… depends on this set of neighbours found for each cell.

In the first steps the distance between cells is calculated. To improve efficiency some a kernel density is applied to reduce the amount of distances to be calculated.
The **K** **N**earest **N**eighbours are selected and from there a graph of **S**hared **N**earest **N**eighbours is constructed

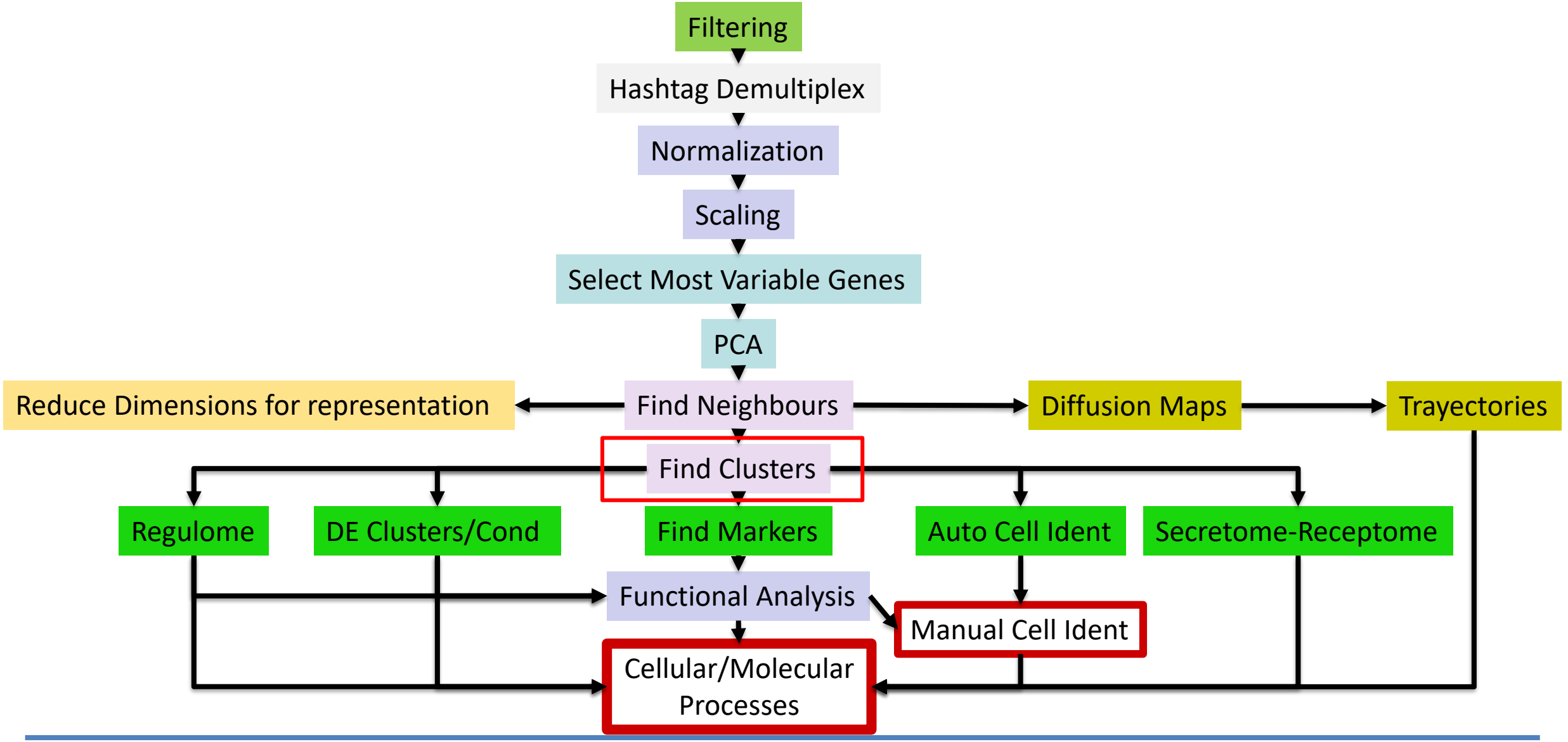


KNN SNN



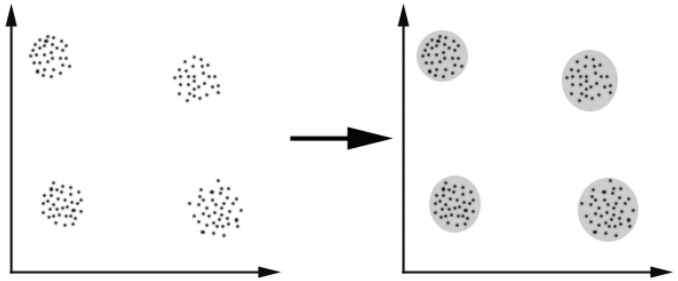
Neighbour graph

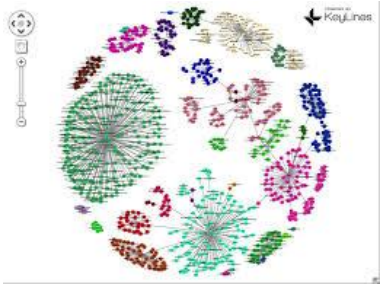
scRNA-Seq Workflow



Find Clusters

- Detection of Populations (**Clustering** Algorithms)
  - Supervised/Informed → Kmeans, Hierarchical, GMM,…
  - Unsupervised/Uninformed → Mean-shift, X-shift, Phenograph, ClusterX, Louvain



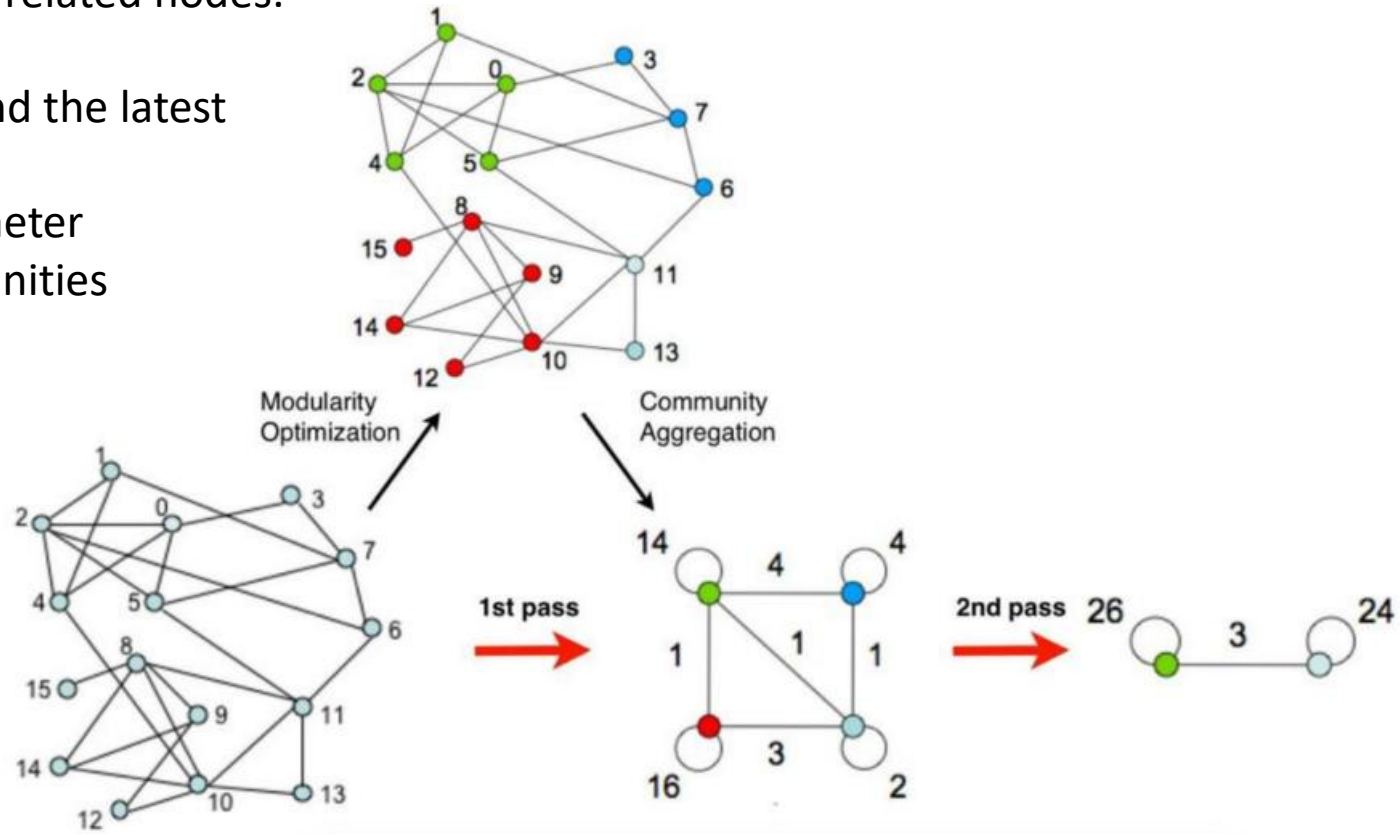


scRNA-Seq Workflow

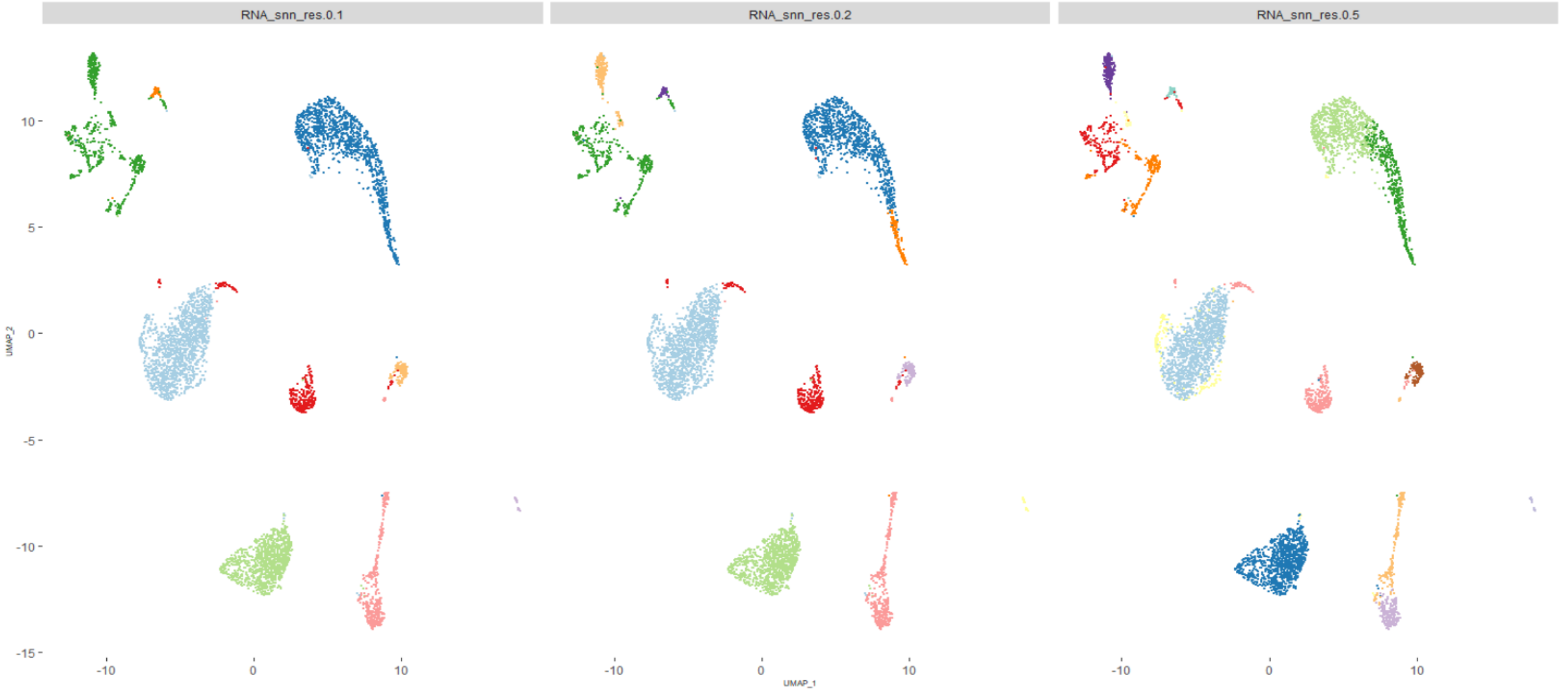
Find Clusters

- This step tries to find local communities of very related nodes.
- There are several algorithms available.
- Most widely used is Louvain based clustering and the latest Leiden improvement.
- Communities are defined by a resolution parameter
- The higher resolution, the more are the communities subdivided.

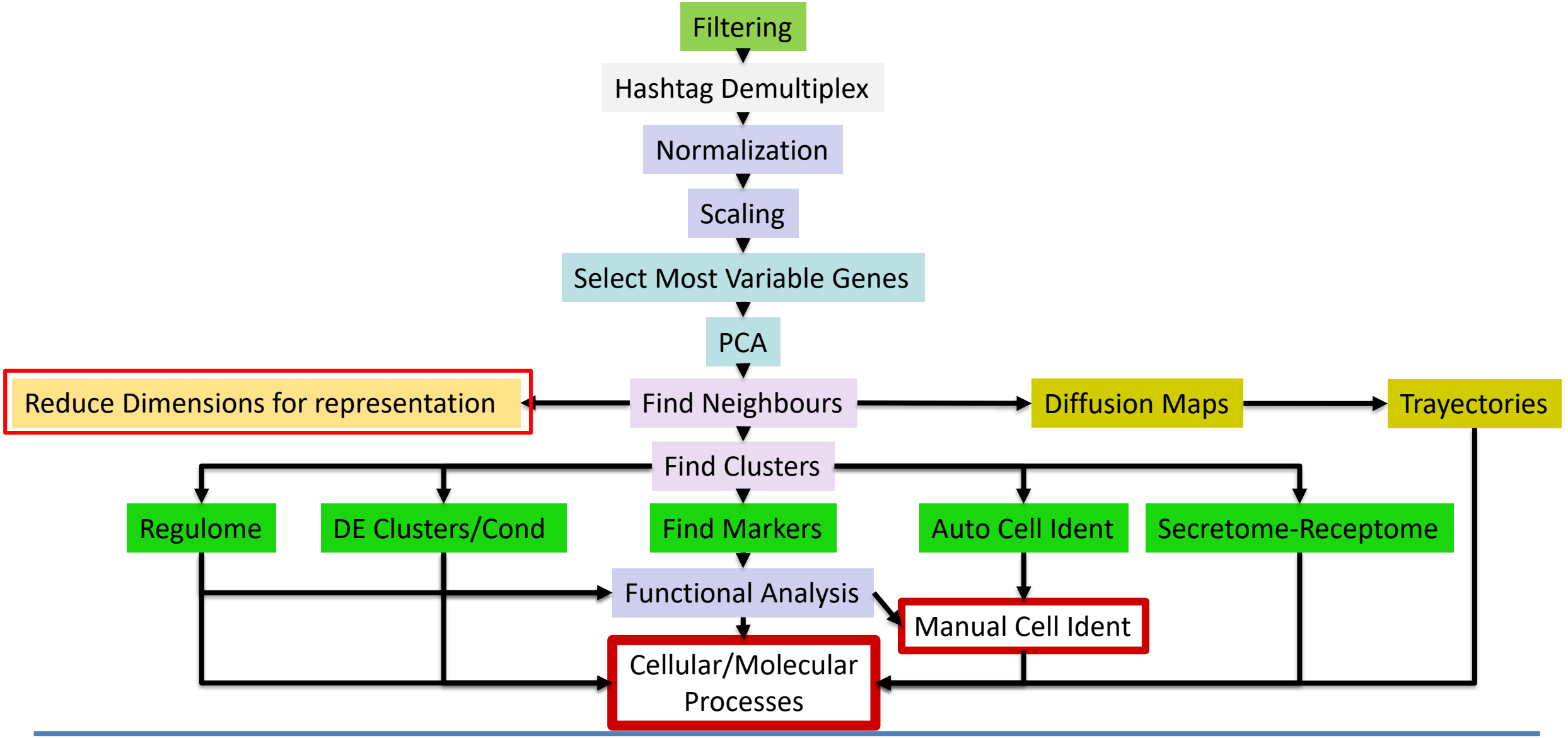
- The resolution parameter is arbitrary and here a bit of biological interpretation is needed to refine the automated clusters.



scRNA-Seq Workflow



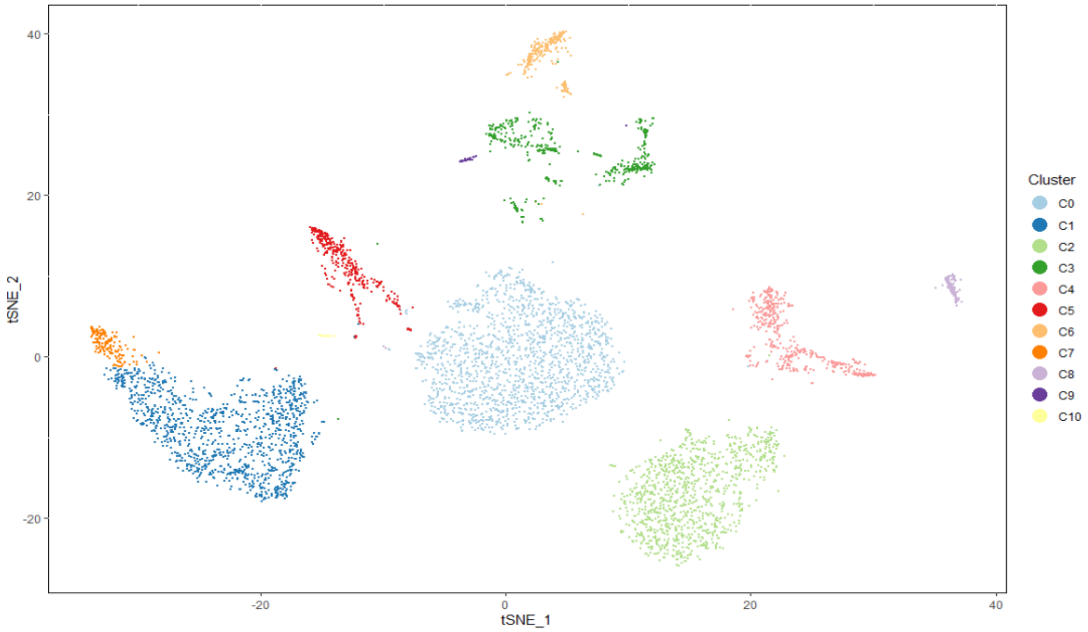
scRNA-Seq Workflow



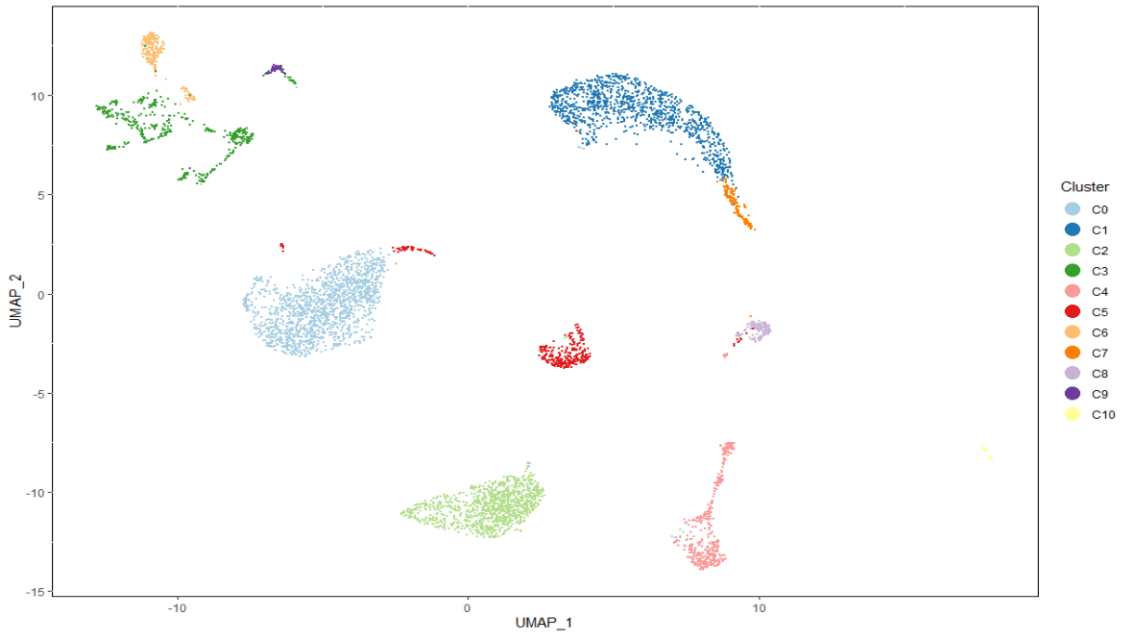
scRNA-Seq Workflow

Reduce Dimensions for representation

tSNE

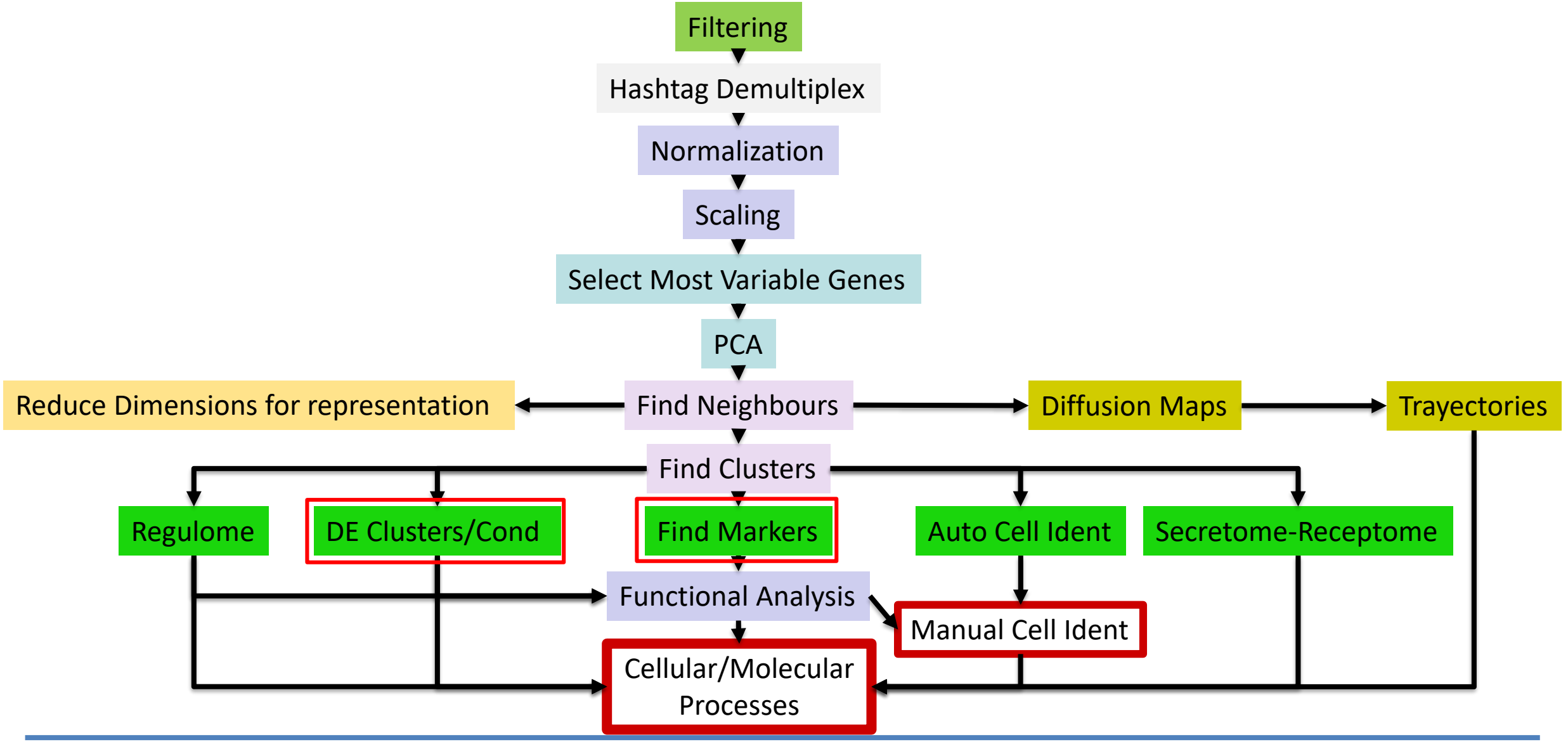


UMAP



Both are perfectly valid for representation purposes. UMAP better represents global structure although still, distances between separated clusters are not indicative.

scRNA-Seq Workflow



scRNA-Seq Workflow

Find Markers

- The goal is to define the most relevant genes that define a cluster.
- The test compares the average expression of all the genes in a particular cluster to the average expression in the rest.
- Several clusters may share marker genes.
- Several methods can be applied.

However the most robust methods are:

- Wilcoxon: A non-parametric test which is robust to the underlying distribution of the data although not very powerful.
- MAST: a two-part generalized linear model that simultaneously models the rate of expression over the background of various transcripts, and the positive expression mean. This model can accommodate complex designs.

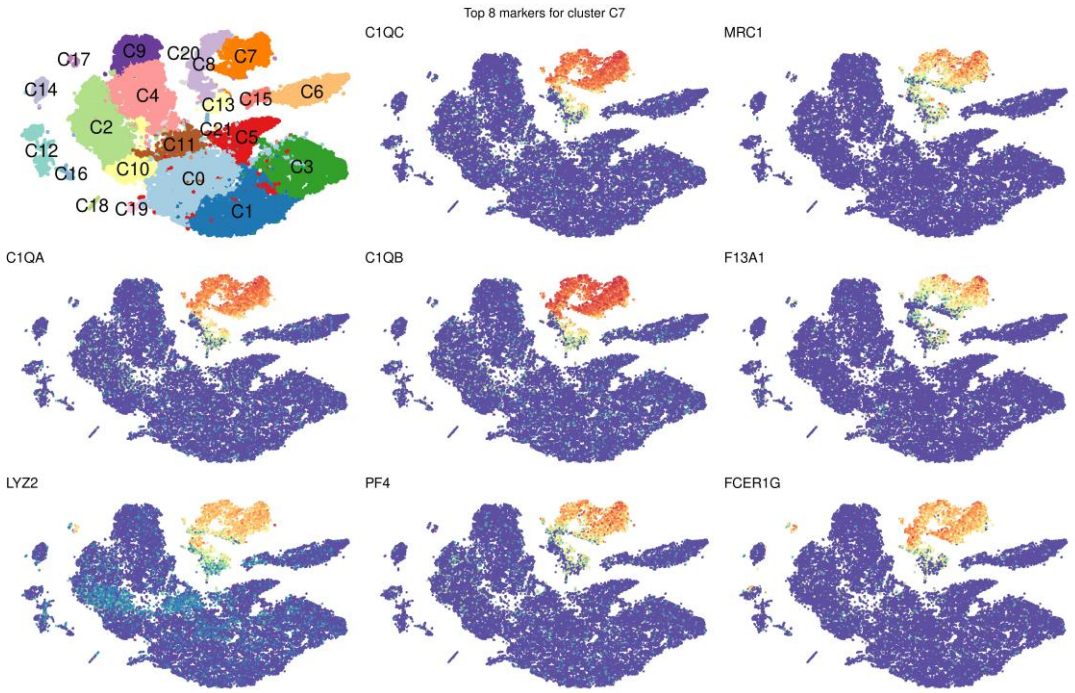
Finak, G., McDavid, A., Yajima, M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16,** 278 (2015). https://doi.org/10.1186/s13059-015-0844-5

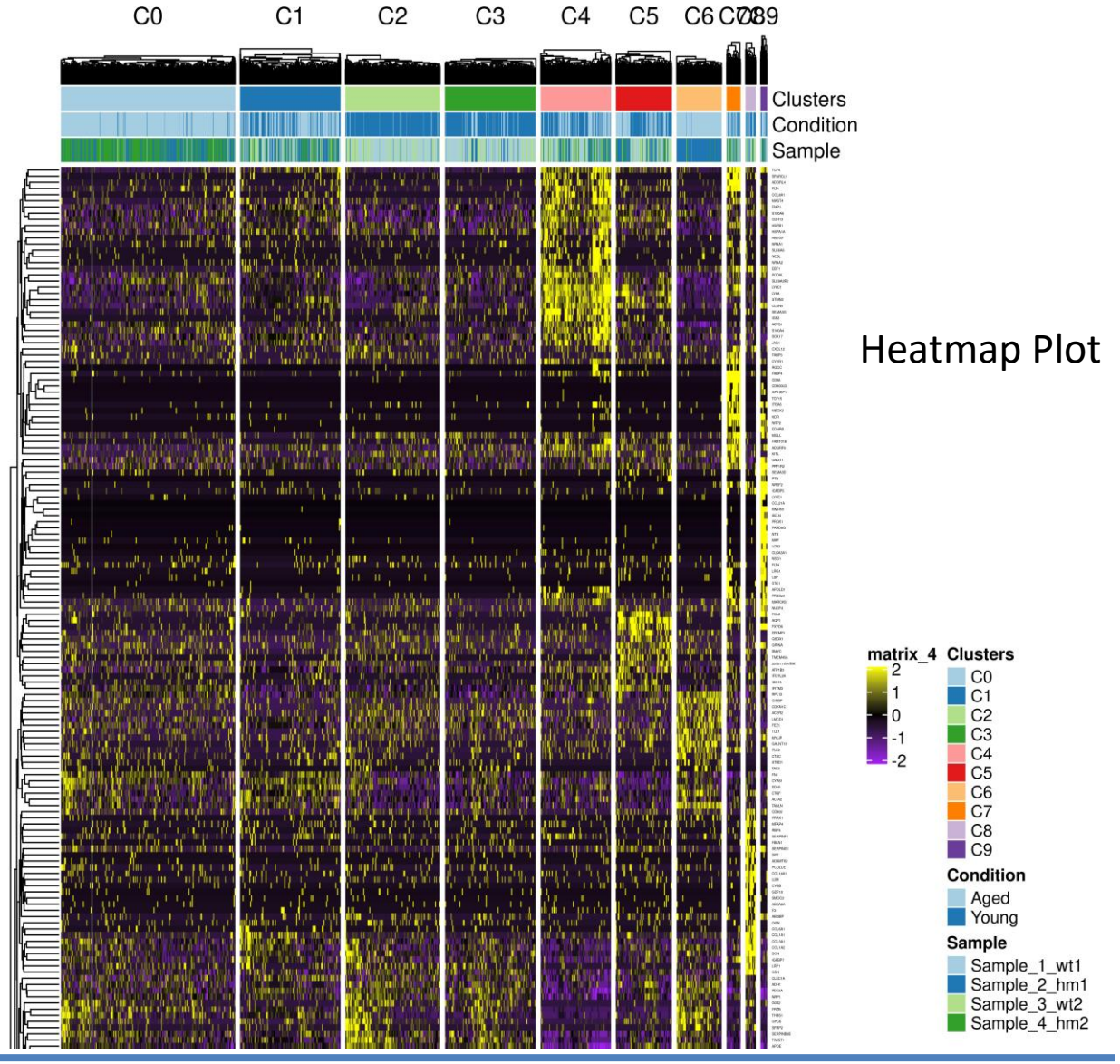
DE between cluster pairs and DE between conditions can also been computed

scRNA-Seq Workflow

Find Markers

Scatter Plot

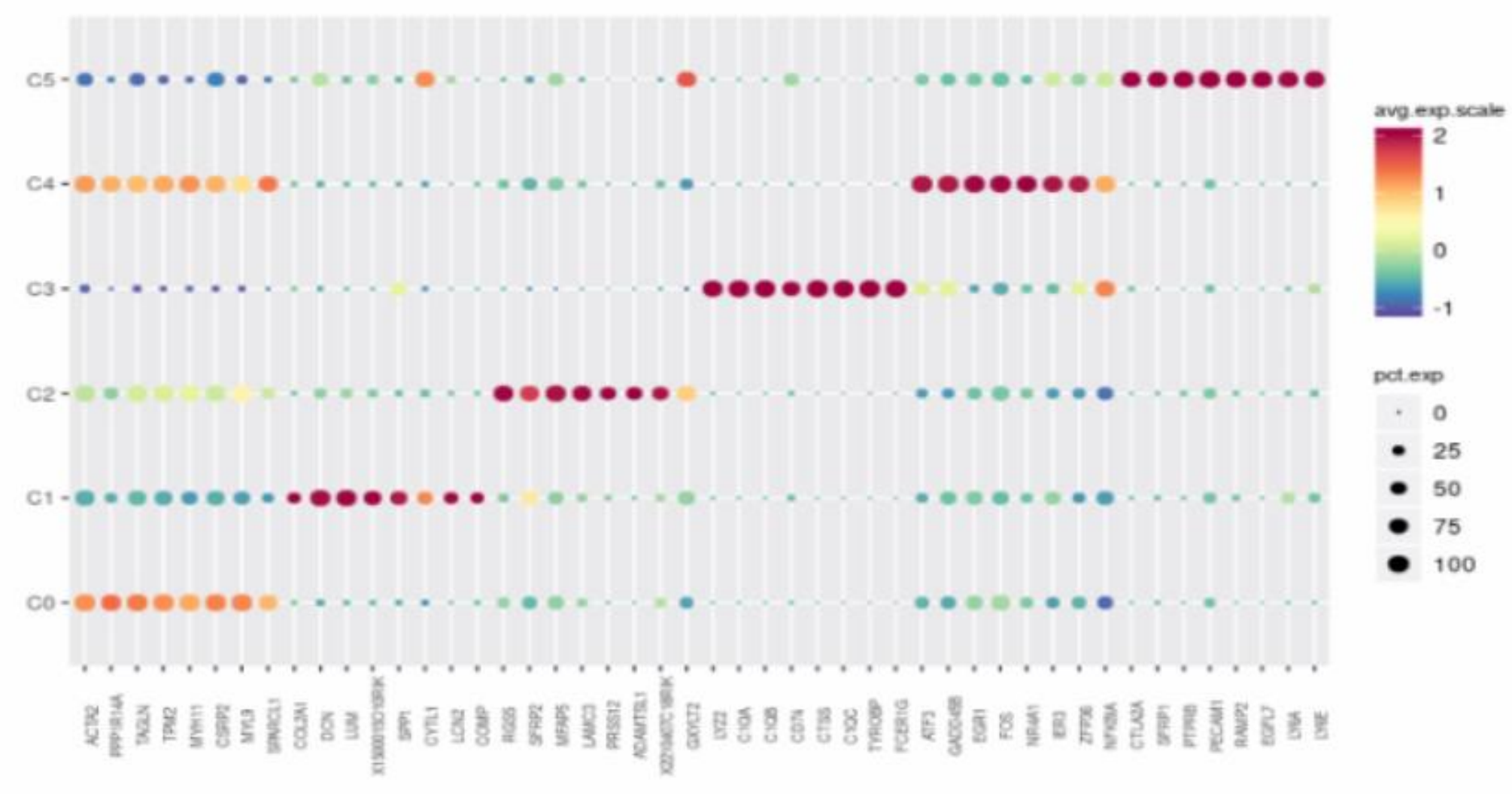




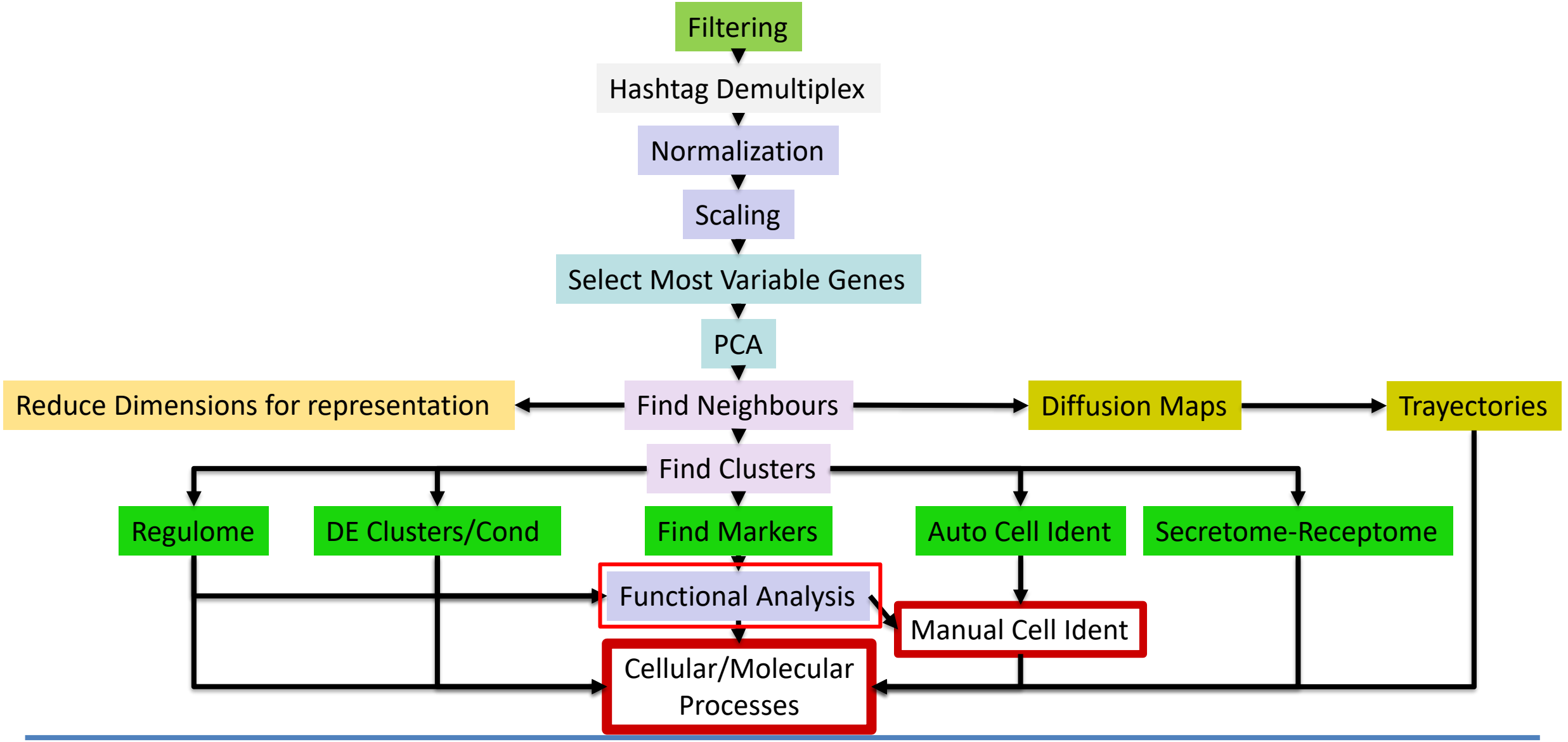
Heatmap Plot

scRNA-Seq Workflow

Top Marker Genes Dot plot

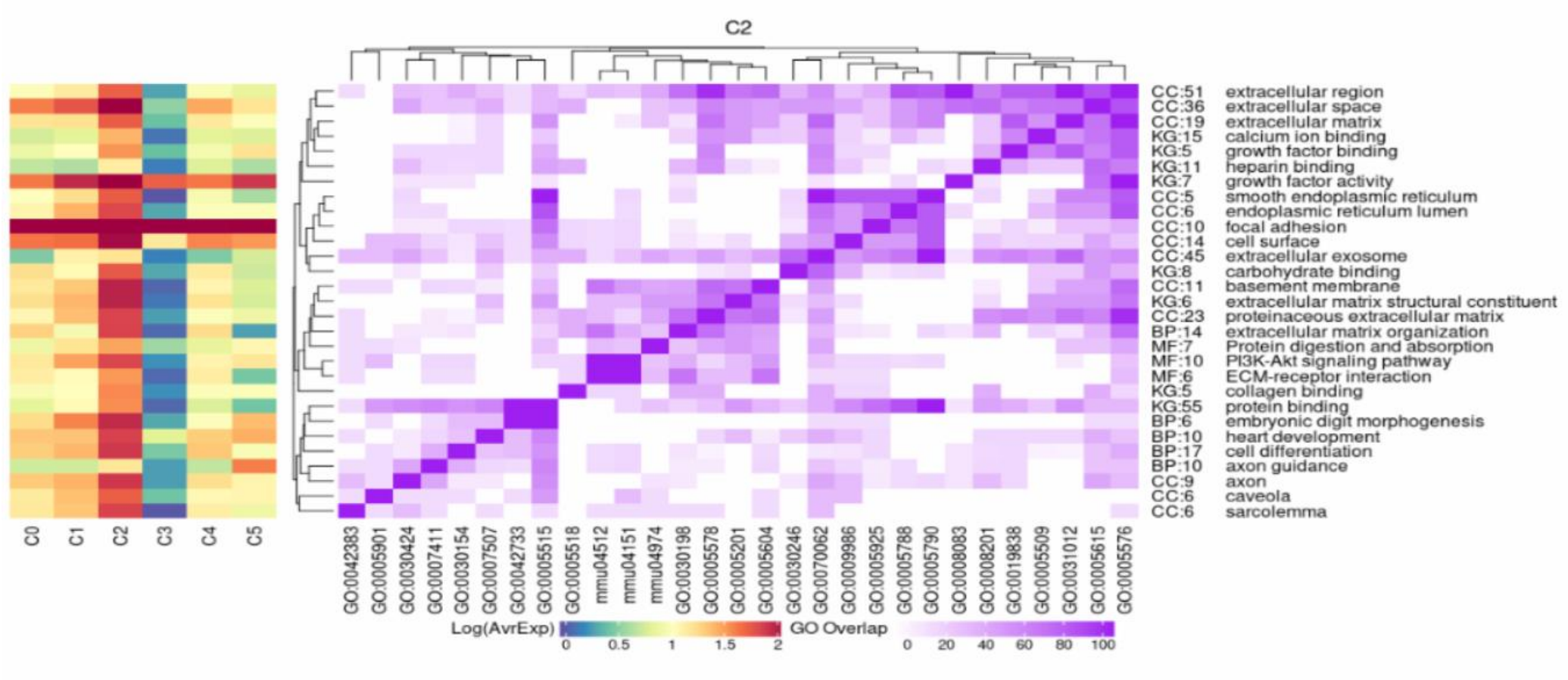


scRNA-Seq Workflow

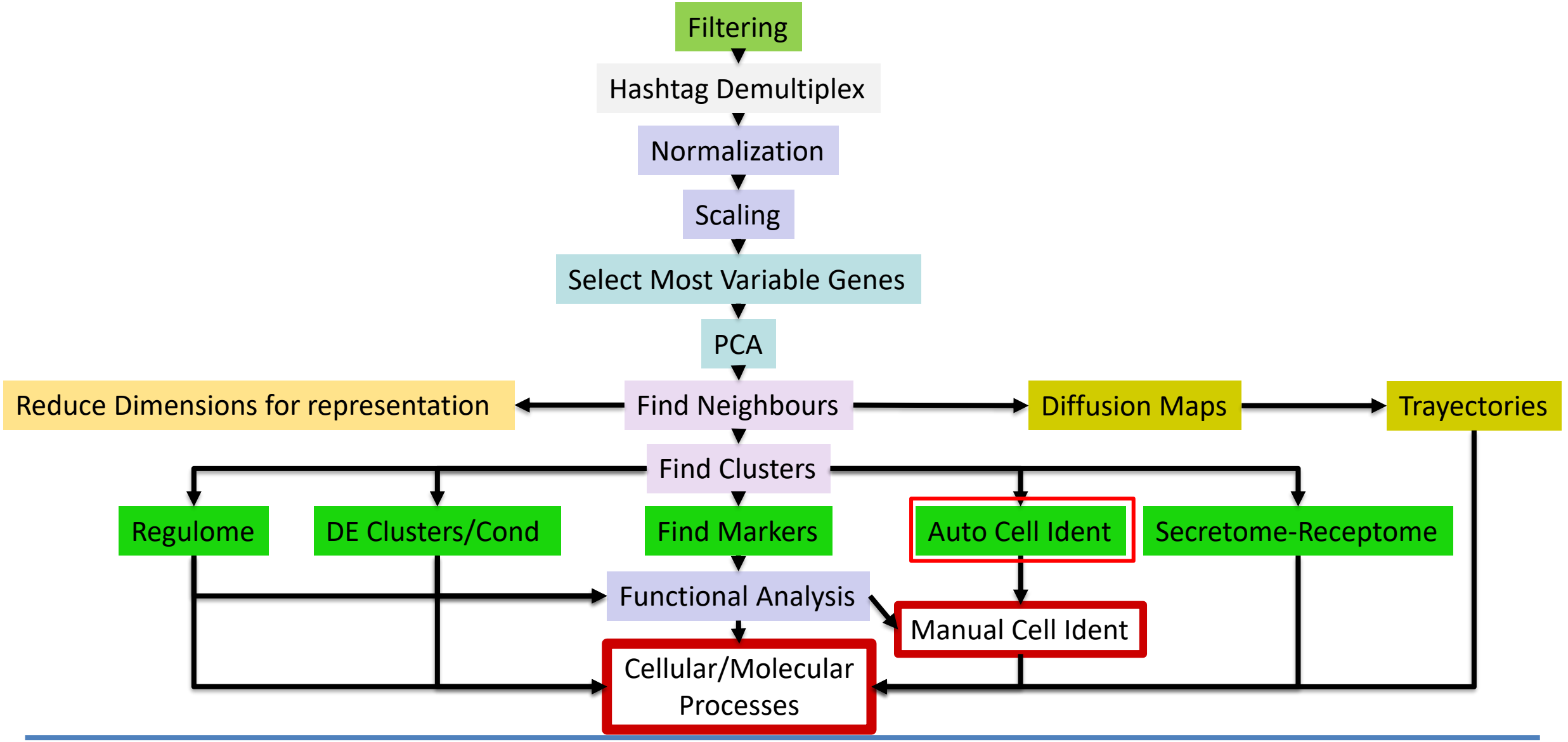


scRNA-Seq Workflow

Functional Analysis



scRNA-Seq Workflow



scRNA-Seq Workflow

Auto Cell Type Identification

SingleR: A reference correlation based cell type identification using robust variant of nearest-neighbors classification, with some tweaks to improve resolution for closely related labels

Human primary cell atlas (HPCA)

The HPCA reference consists of publicly available microarray datasets derived from human primary cells (Mabbott et al. 2013). Most of the labels refer to blood subpopulations but cell types from other tissues are also available.

Blueprint/ENCODE

The Blueprint/ENCODE reference consists of bulk RNA-seq data for pure stroma and immune cells generated by Blueprint (Martens and Stunnenberg 2013) and ENCODE projects (The ENCODE Project Consortium 2012).

Mouse RNA-seq

This reference consists of a collection of mouse bulk RNA-seq data sets downloaded from the gene expression omnibus (Benayoun et al. 2019). A variety of cell types are available, again mostly from blood but also covering several other tissues.

Novershtern hematopoietic data

The Novershtern reference (previously known as Differentiation Map) consists of microarray datasets for sorted hematopoietic cell populations from GSE24759 (Novershtern et al. 2011).

Immunological Genome Project (ImmGen)

The ImmGen reference consists of microarray profiles of pure mouse immune cells from the project of the same name (Heng et al. 2008). This is currently the most highly resolved immune reference - possibly overwhelmingly so, given the granularity of the fine labels.

Database of Immune Cell Expression/eQTLs/Epigenomics (DICE)

The DICE reference consists of bulk RNA-seq samples of sorted cell populations from the project of the same name (Schmiedel et al. 2018).

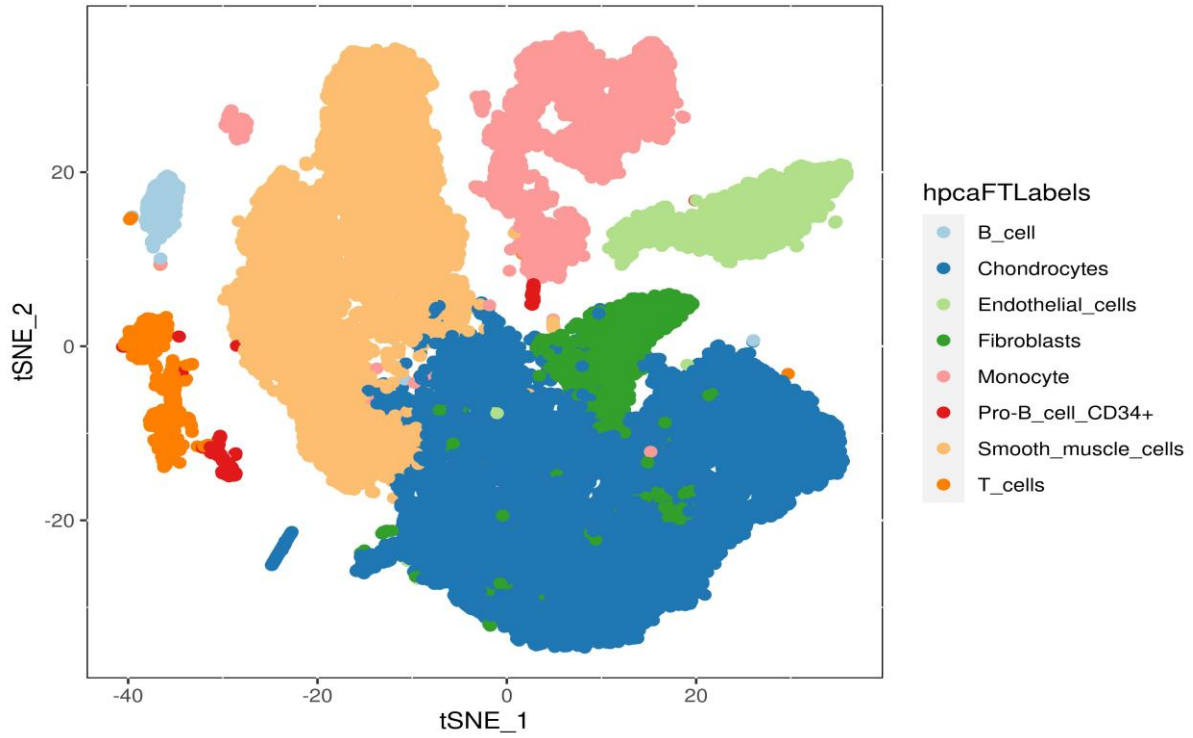
Monaco immune data

The Monaco reference consists of bulk RNA-seq samples of sorted immune cell populations from GSE107011 (Monaco et al. 2019).

scRNA-Seq Workflow

Auto Cell Type Identification

SingleR Cell Type Classification based on Human Cell Atlas Reference



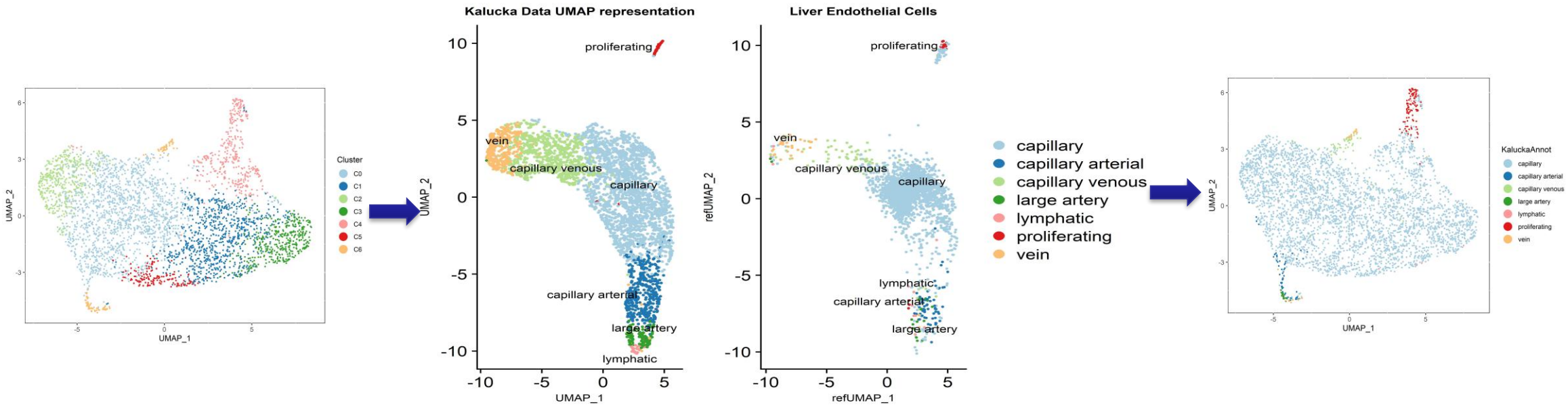
Classifications are far from perfect but they are a way to start and incorporate more knowledge using gene expression and functional analysis results to refine the annotations

scRNA-Seq Workflow

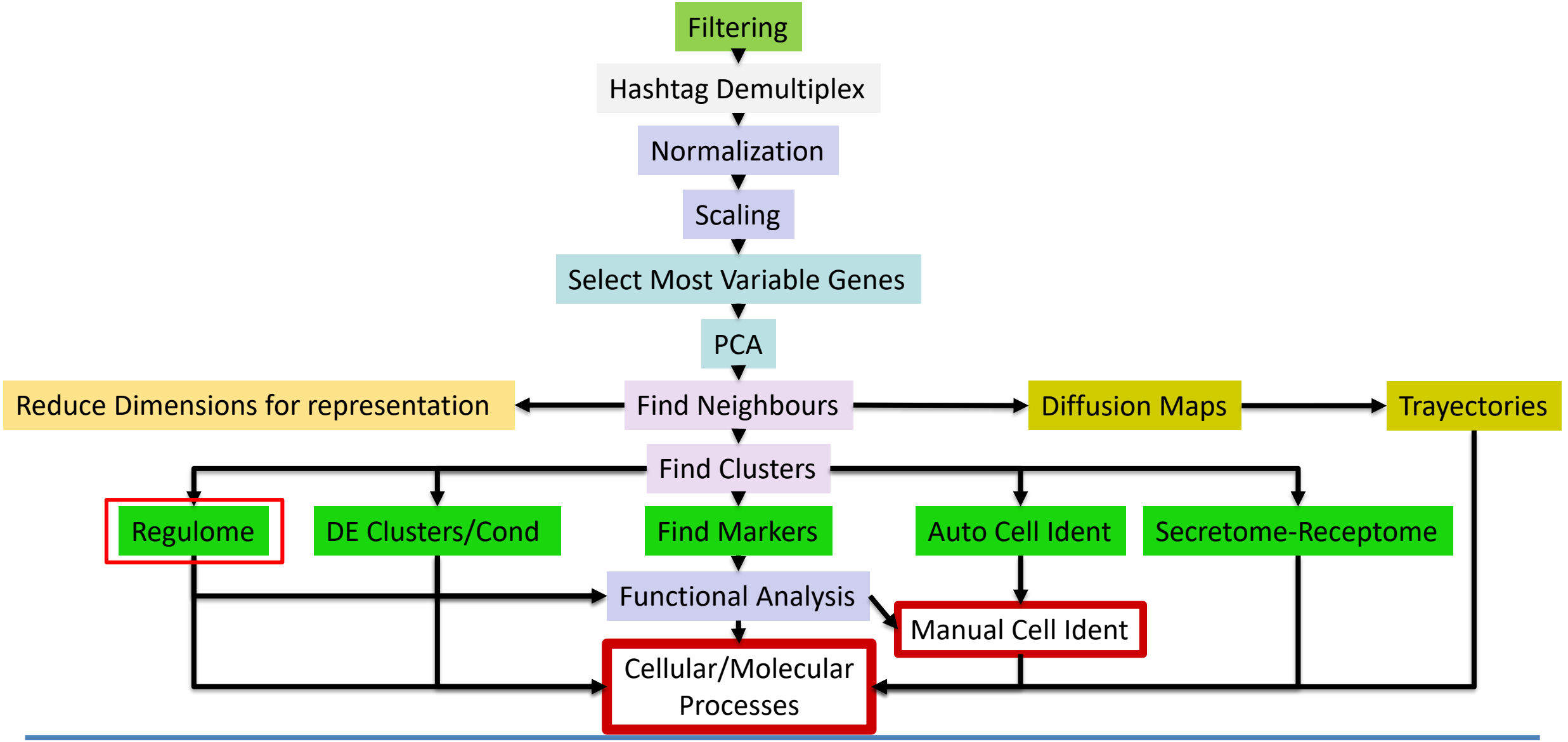
Auto Cell Type Identification

Map single cell experiment to an already annotated single cell experiment:

Using **C**anonical **C**roscorrelation **A**nalysis and **M**utual **N**earest **N**eighbours.



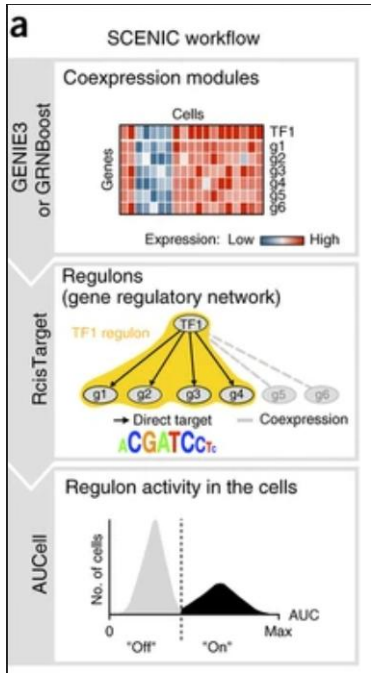
scRNA-Seq Workflow

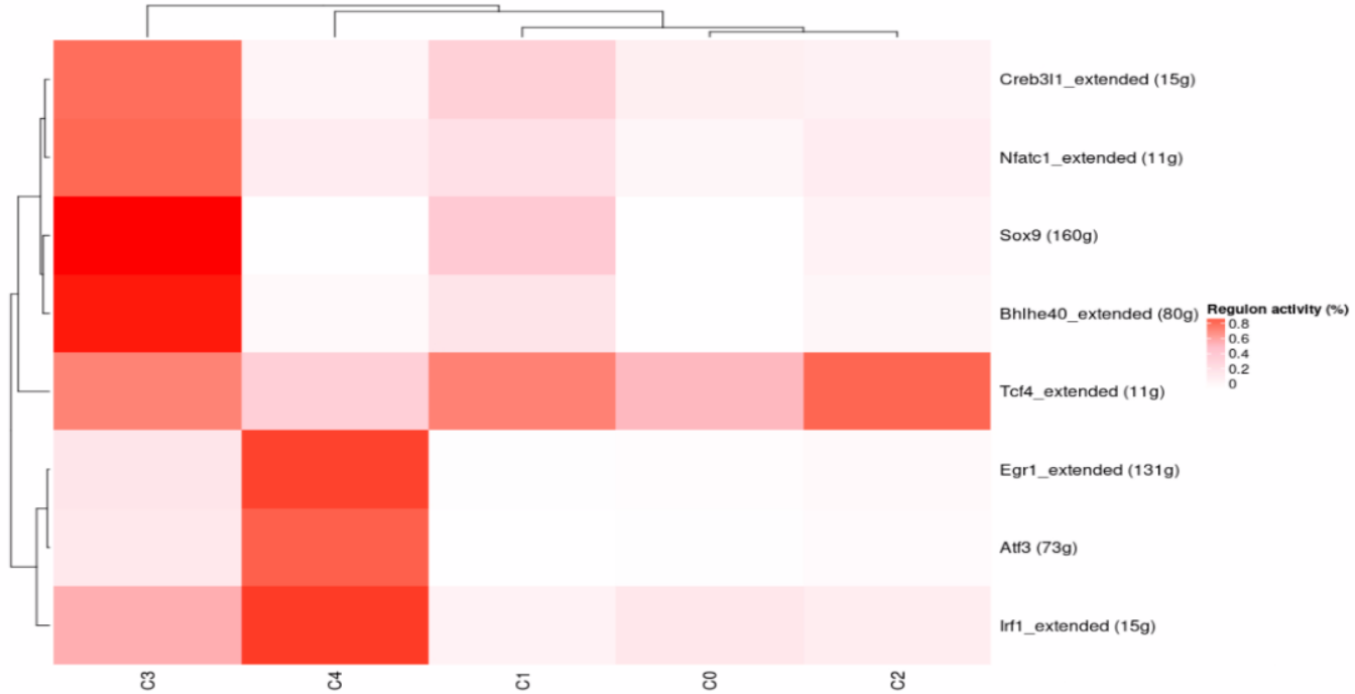


scRNA-Seq Extended

Regulome

This analysis tries to find transcription factors that might explain the gene expression variability found between clusters

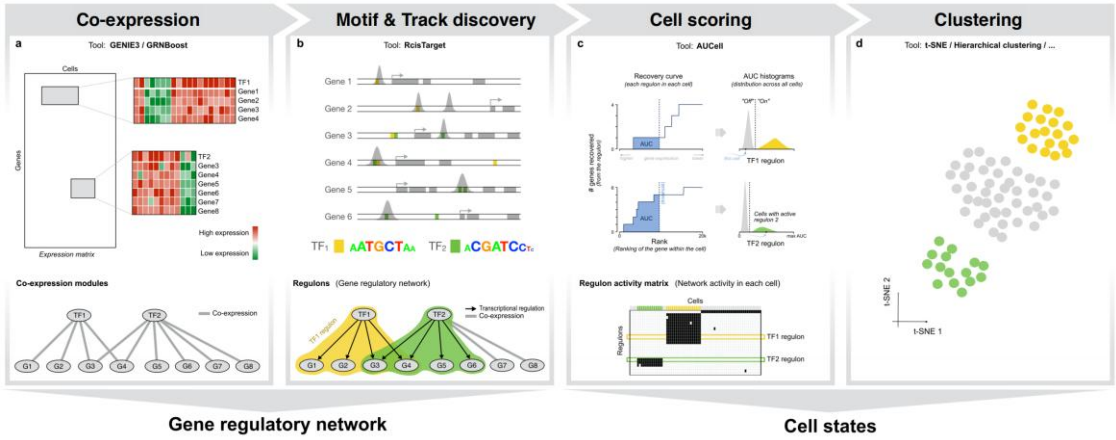




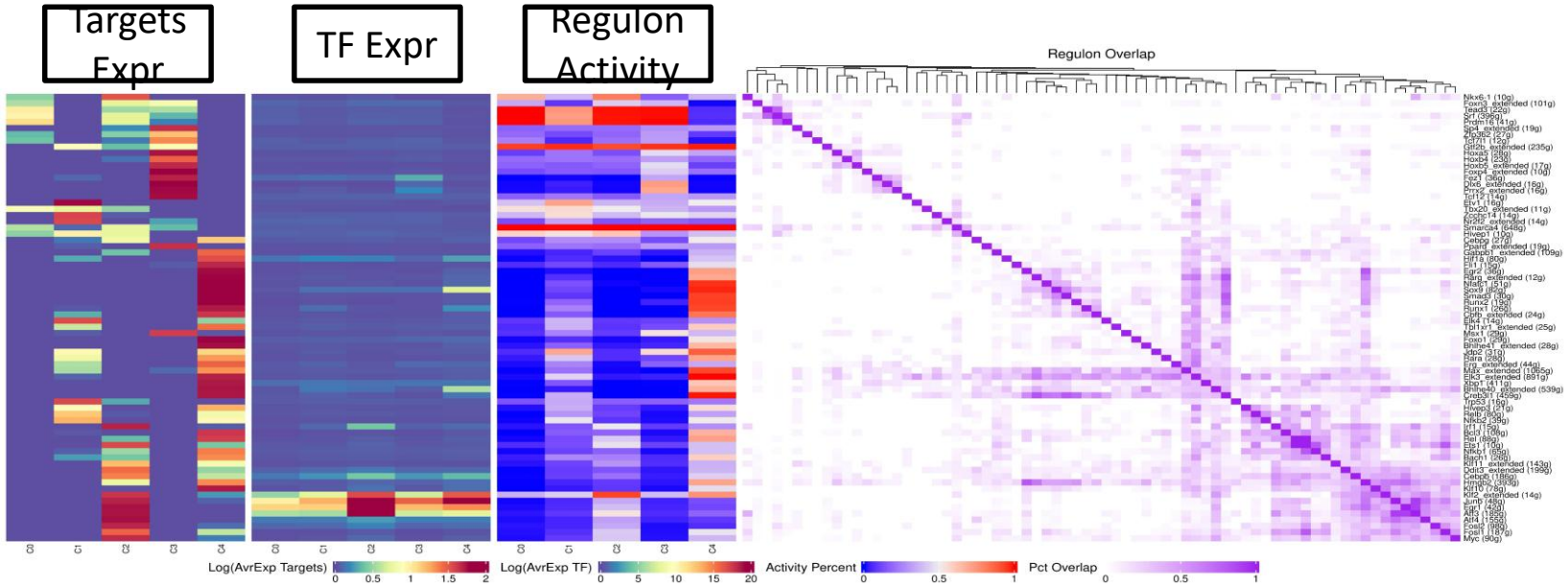
Aibar, S., González-Blas, C., Moerman, T. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14,** 1083–1086 (2017). https://doi.org/10.1038/nmeth.4463

scRNA-Seq Extended

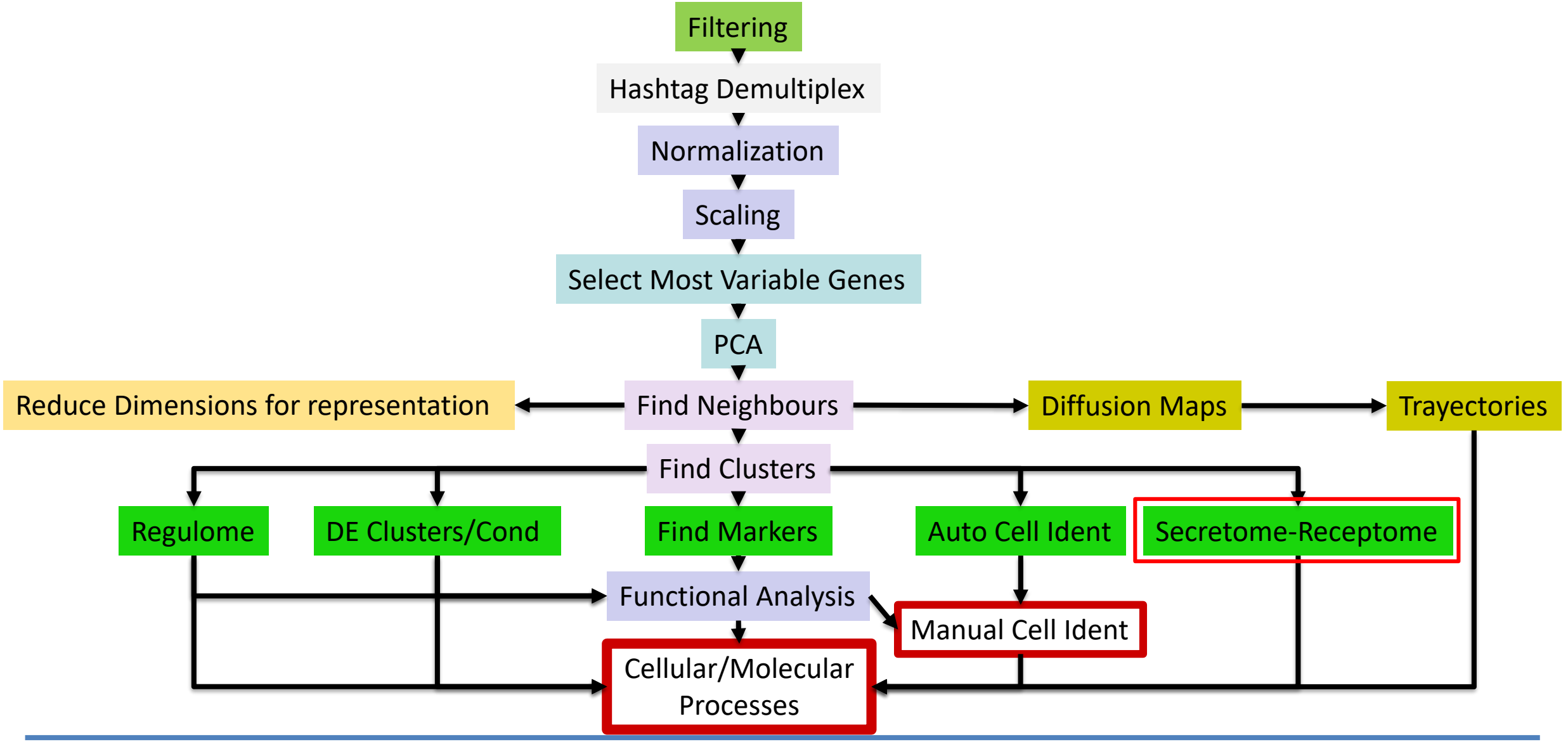
Regulome



https://scenic.aertslab.org/



scRNA-Seq Workflow

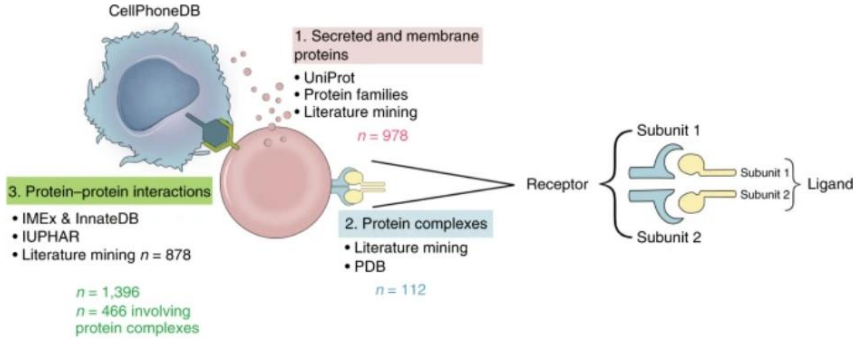


scRNA-Seq Workflow

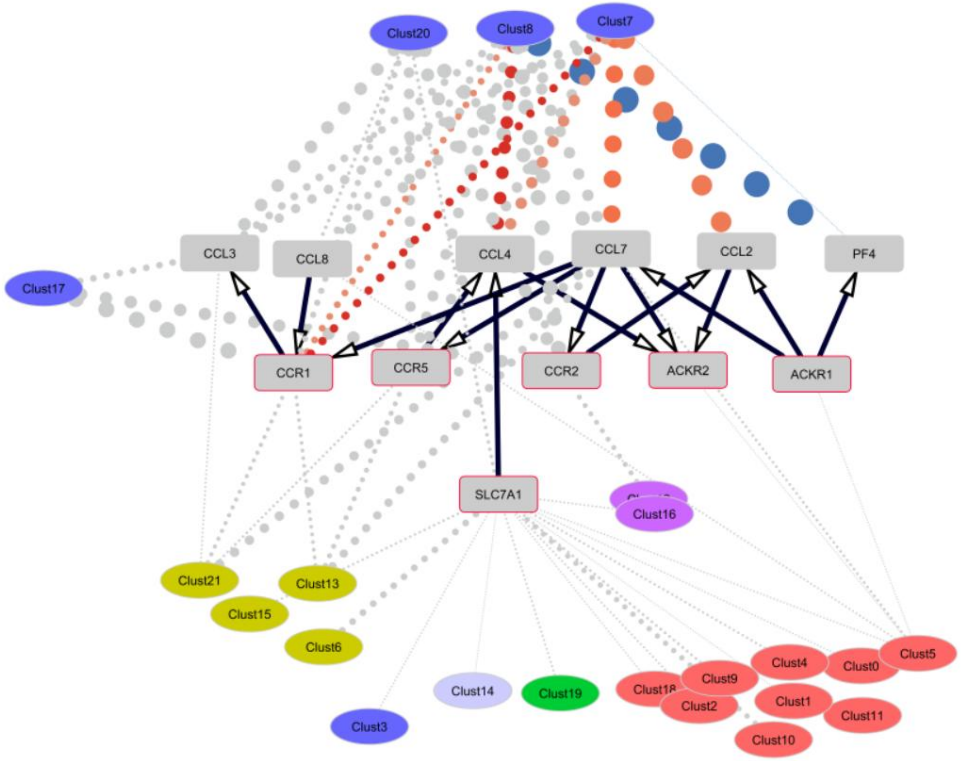
Secretome-Receptome

CellPhoneDB, CellChat, …

- Public repository of ligands, receptors and their interactions to enable a comprehensive, systematic analysis of cell–cell communication molecules
- derive enriched ligand–receptor interactions between two cell states on the basis of expression of a receptor by one cell state and a ligand by another cell state and use empirical shuffling to calculate which ligand–receptor pairs display significant cell-state specificity

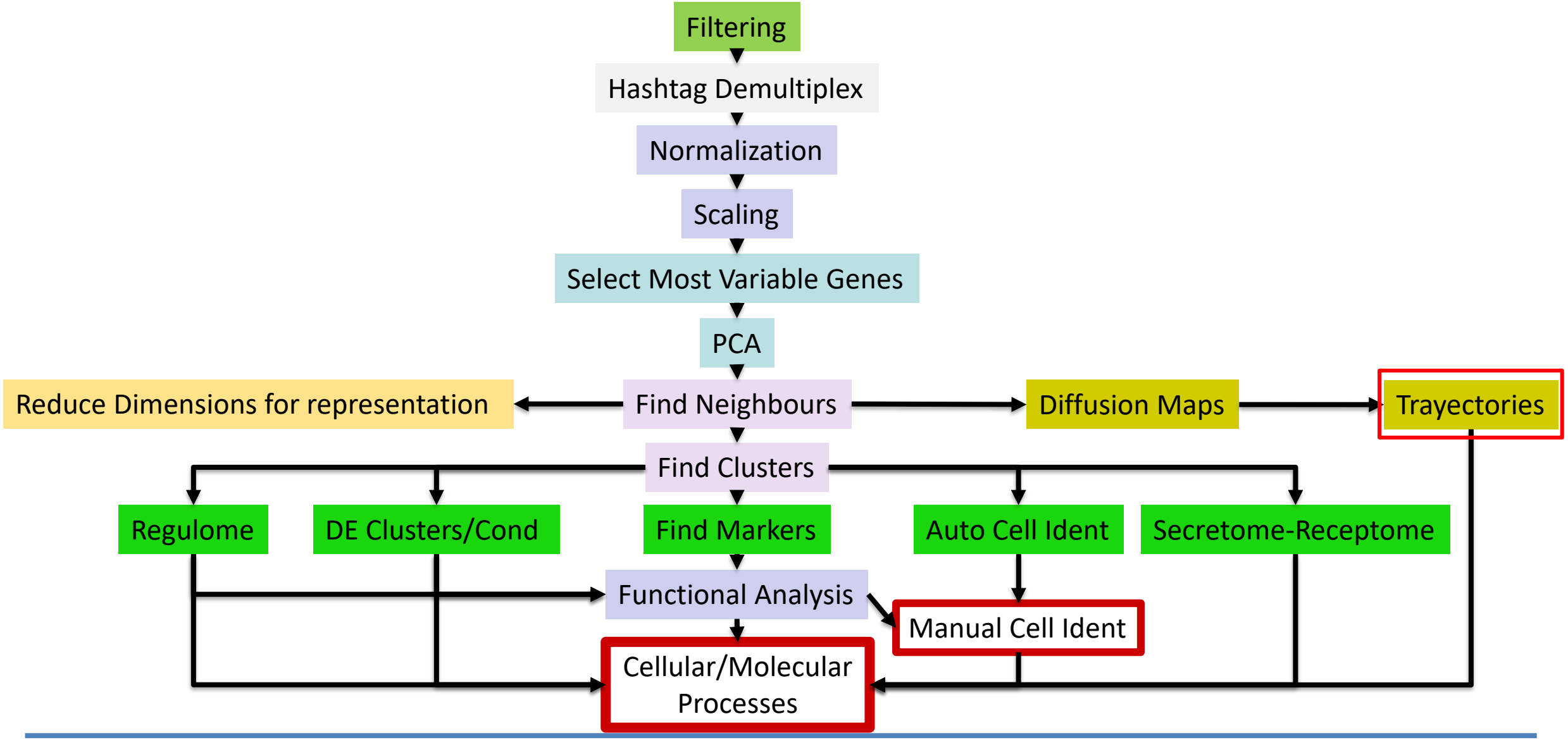


Interaction Module from Aged Aorta Samples

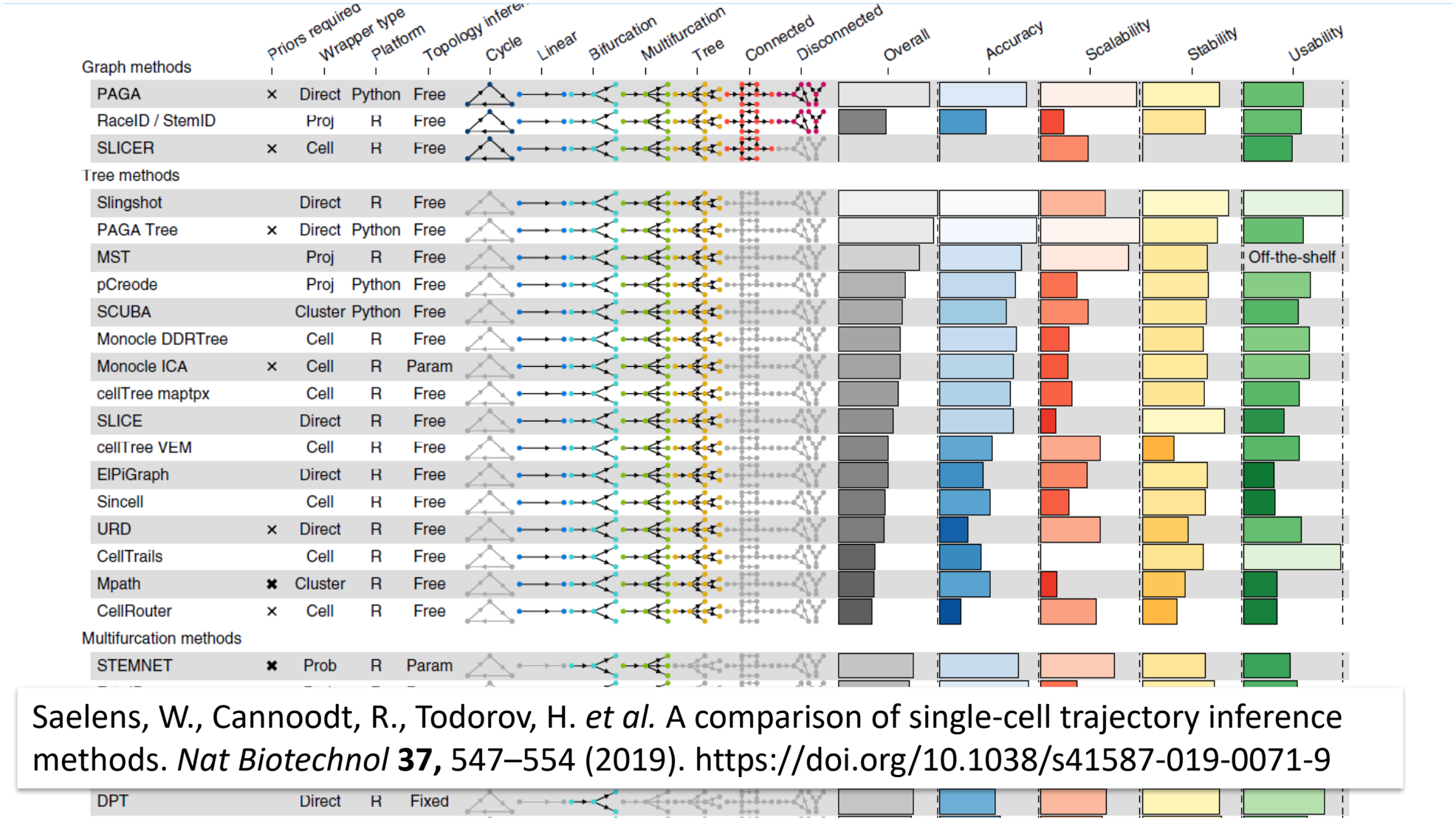


Efremova, M., Vento-Tormo, M., Teichmann, S.A. *et al.* CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* **15,** 1484–1506 (2020). https://doi.org/10.1038/s41596-020-0292-x

scRNA-Seq Workflow



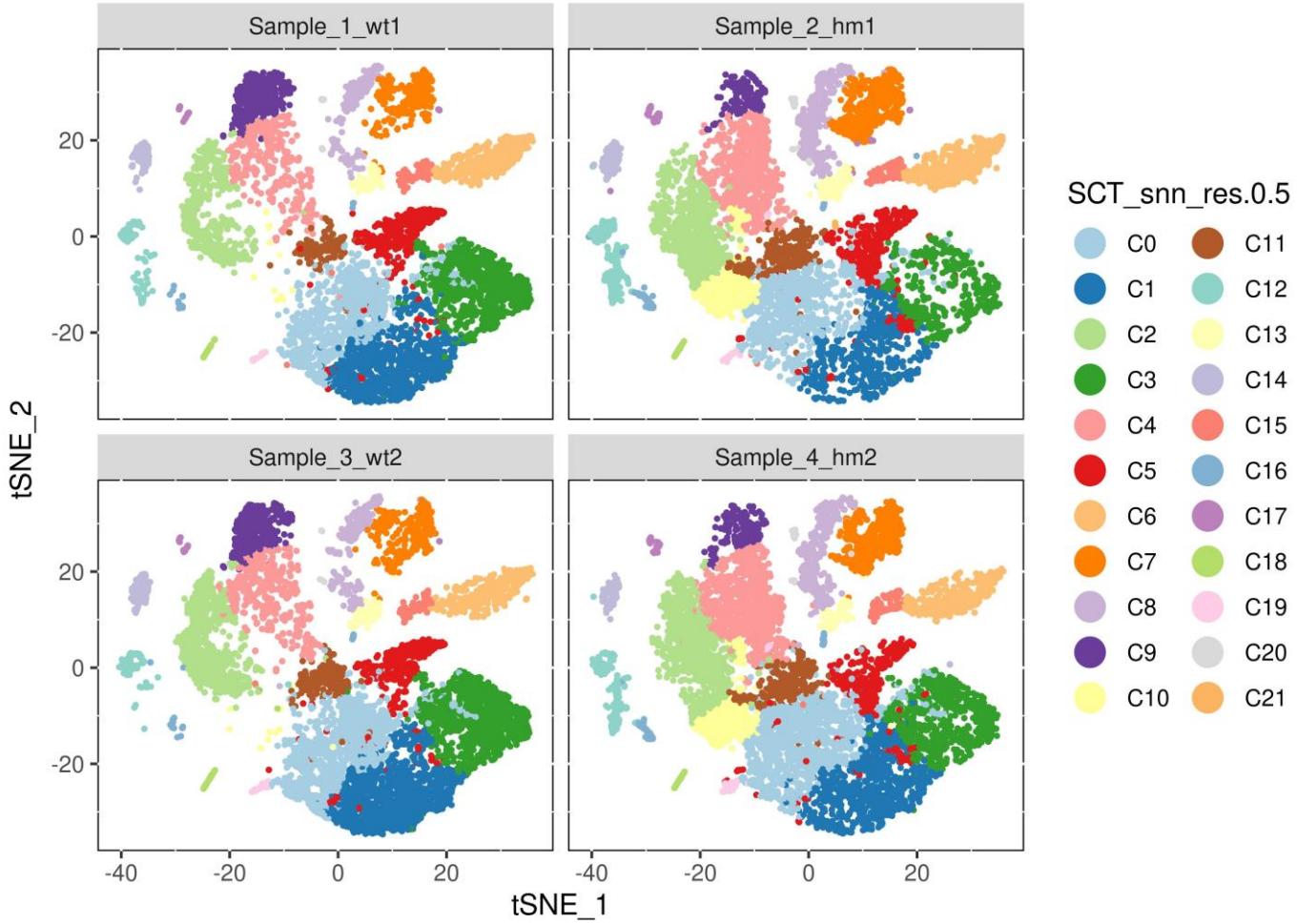
scRNA-Seq Workflow



Saelens, W., Cannoodt, R., Todorov, H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37,** 547–554 (2019). https://doi.org/10.1038/s41587-019-0071-9

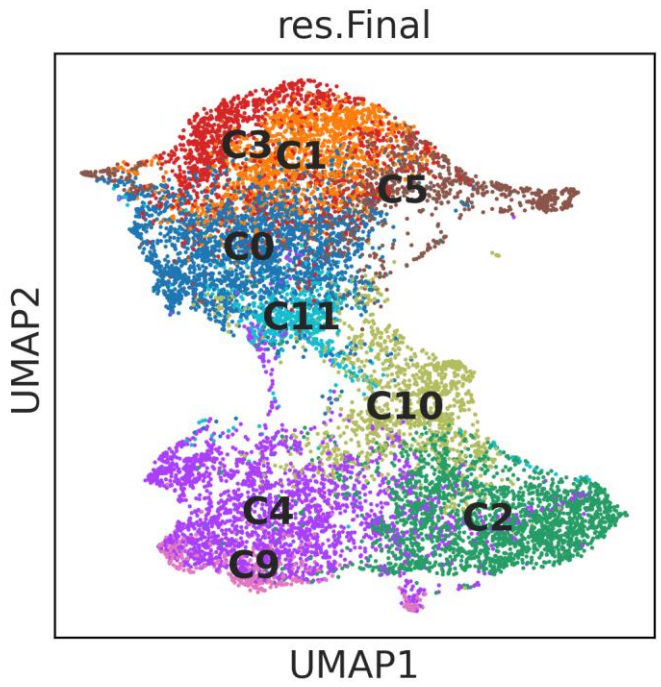
scRNA-Seq Workflow

Aorta Aged Samples from Ana Barrentino. LAB_VA



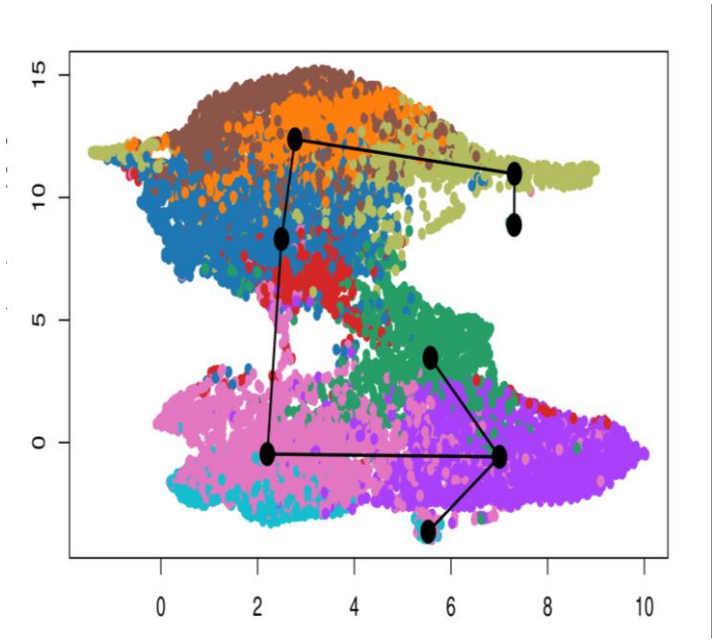
scRNA-Seq Workflow

Trajectory Analysis helped to identify the origin of the new cell population



PAGA Algorythm

Slingshot



scDAVIS: Single Cell DAta VISualisation

scDAVIS

https://bioinfo.cnic.es/scdavis/

- Explore Published Datasets
- Upload your own experiment
- Interrogate the data:
  - Manually define clusters
  - Find markers for the cluster
  - Plot results in different ways
- Download your analysis

END