Introduction to scRNAseq & experimental considerations

Jules GILET - ELIXIR France (Institut Curie, Paris)

Single cell RNAseq data analysis with R - european course
ELIXIR EXCELERATE project

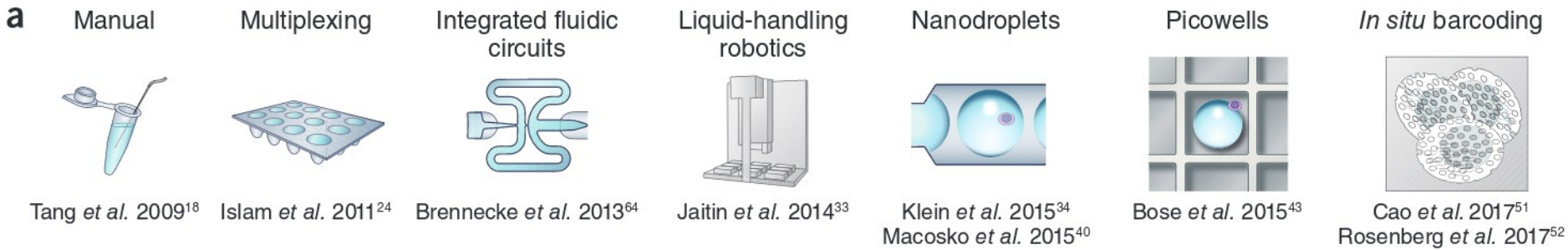
2019-05-27, Espoo, Finland

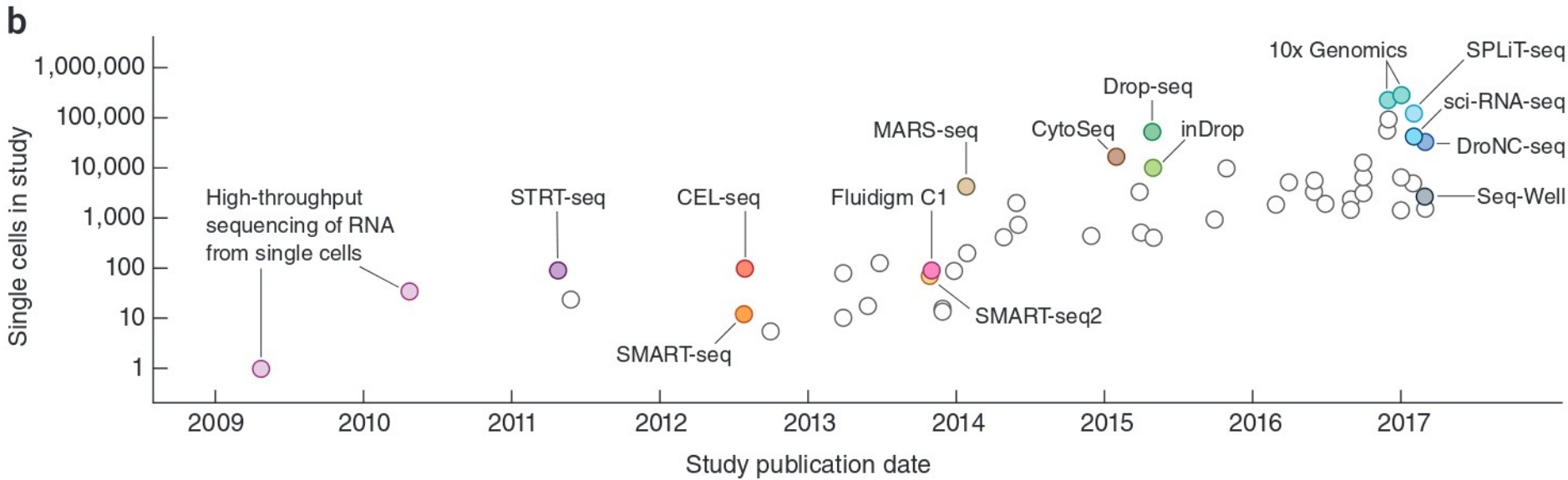
Outline

- Technology overview
- Primary processing
- Example of downstream applications
- Experimental design
- Technical biases

technologies & libraries

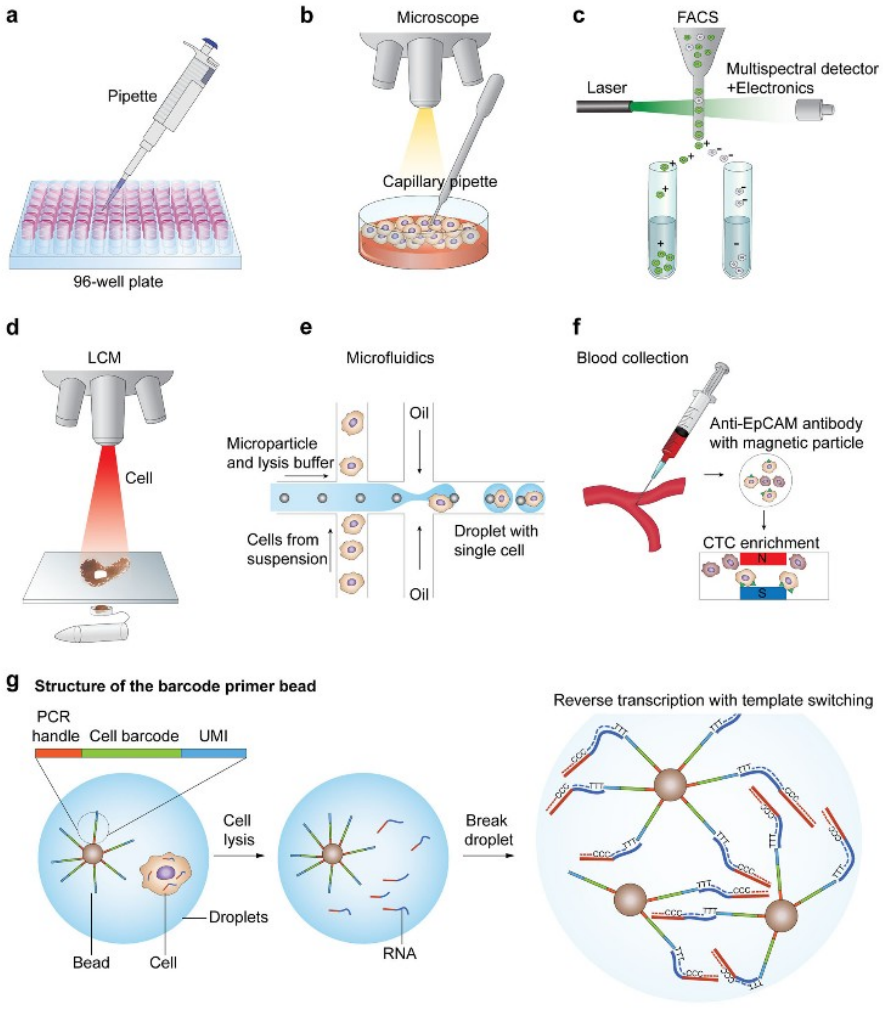
Evolution of scRNAseq techniques





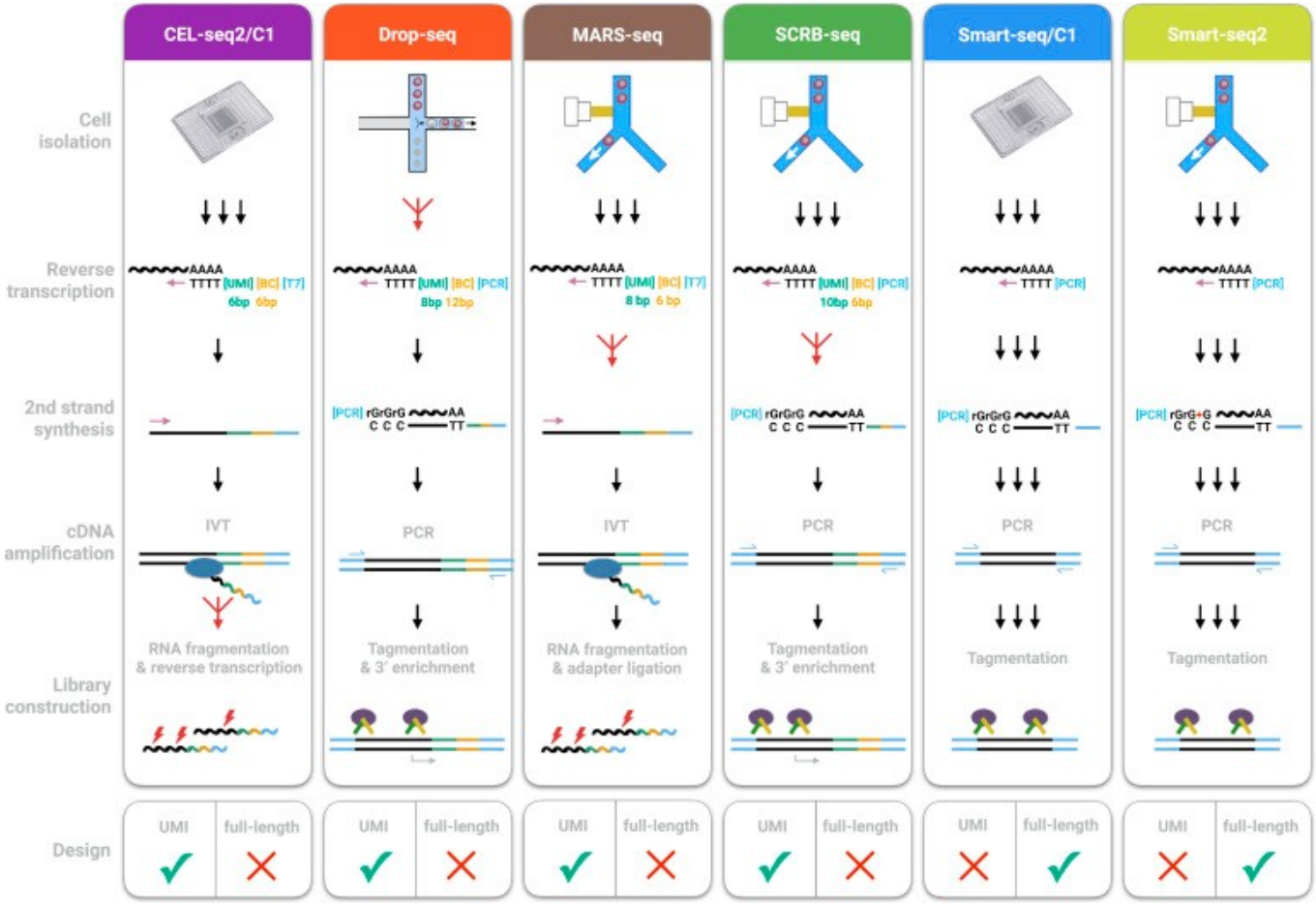
Svensson et al. Nature Prorocols (2018)

Methods for single cell isolation



Hwang et al. Experimental & Molecular Medicine (2018)

Some scRNAseq strategies

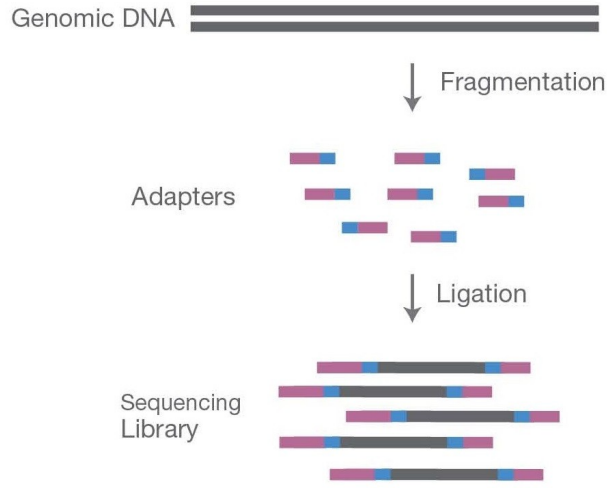


Ziegenhain et al. Molecular Cell (2017)

Sequencing cDNA: length limitations

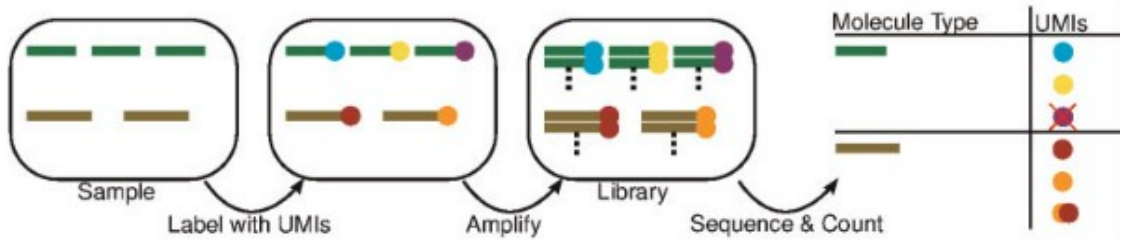
NGS max sequencing capabilities:

HiSeq2500 : 2 x 300 bp (rapid run v2)
NovaSeq : 2 x 250 bp

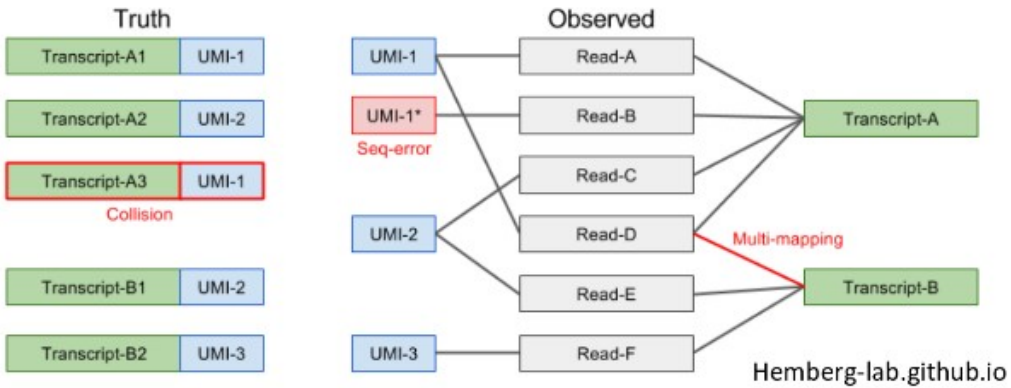


NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Unique Molecular Identifiers



Pflug et al. Bioinformatics (2018)



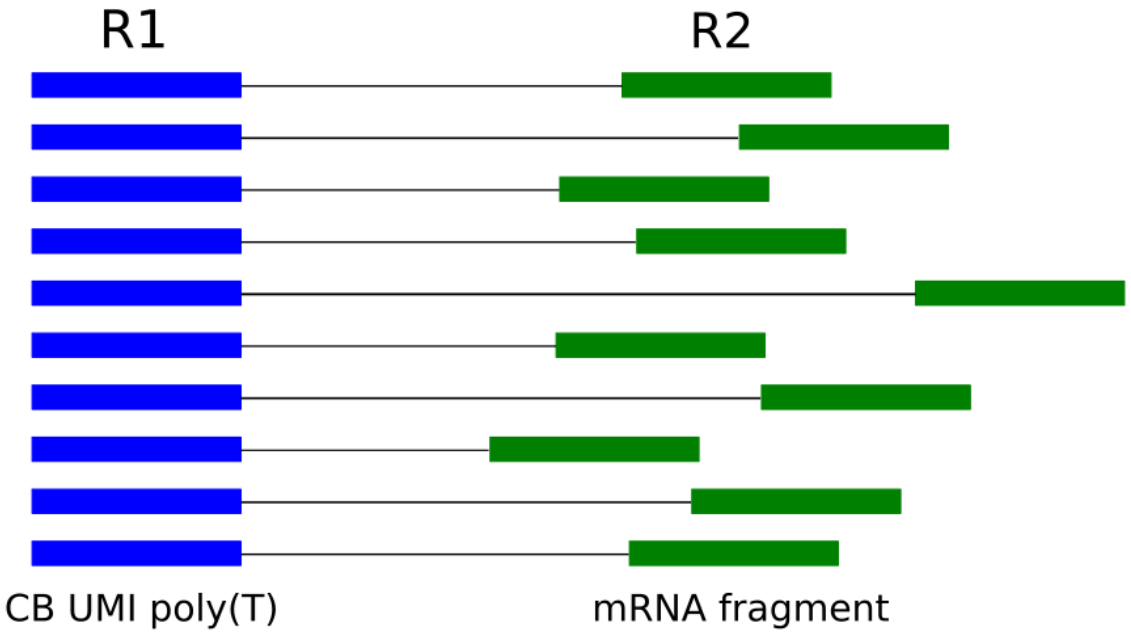
<u>UMI correction:</u>

1 edit distance can be confidently corrected

Different strategies exists, integration of UMI + CB + mapped read, network based methods.

UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy (Smith et al. Genome Research 2017)

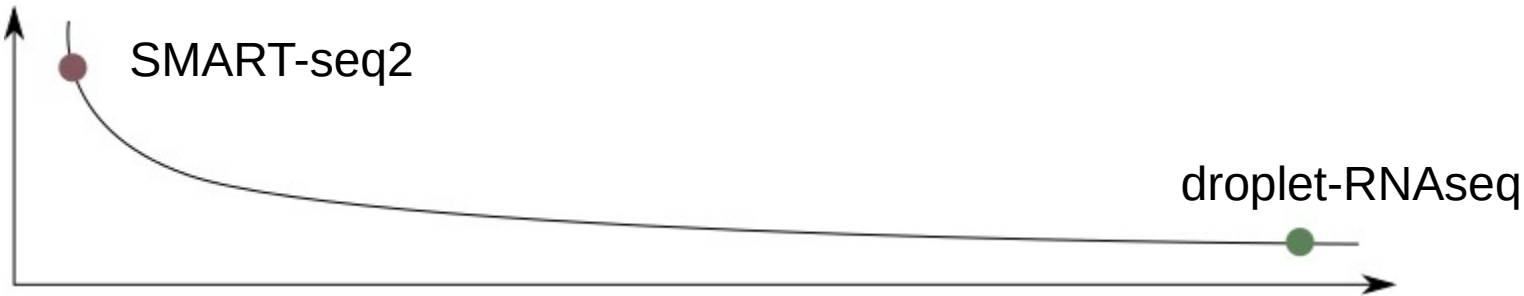
fragmentation associated to UMI increases coverage for a given mRNA



In 3’ libraries, actual coverage vary according to the level of duplication of a given cDNA.

A dichotomous overview of scRNAseq technologies

number of genes



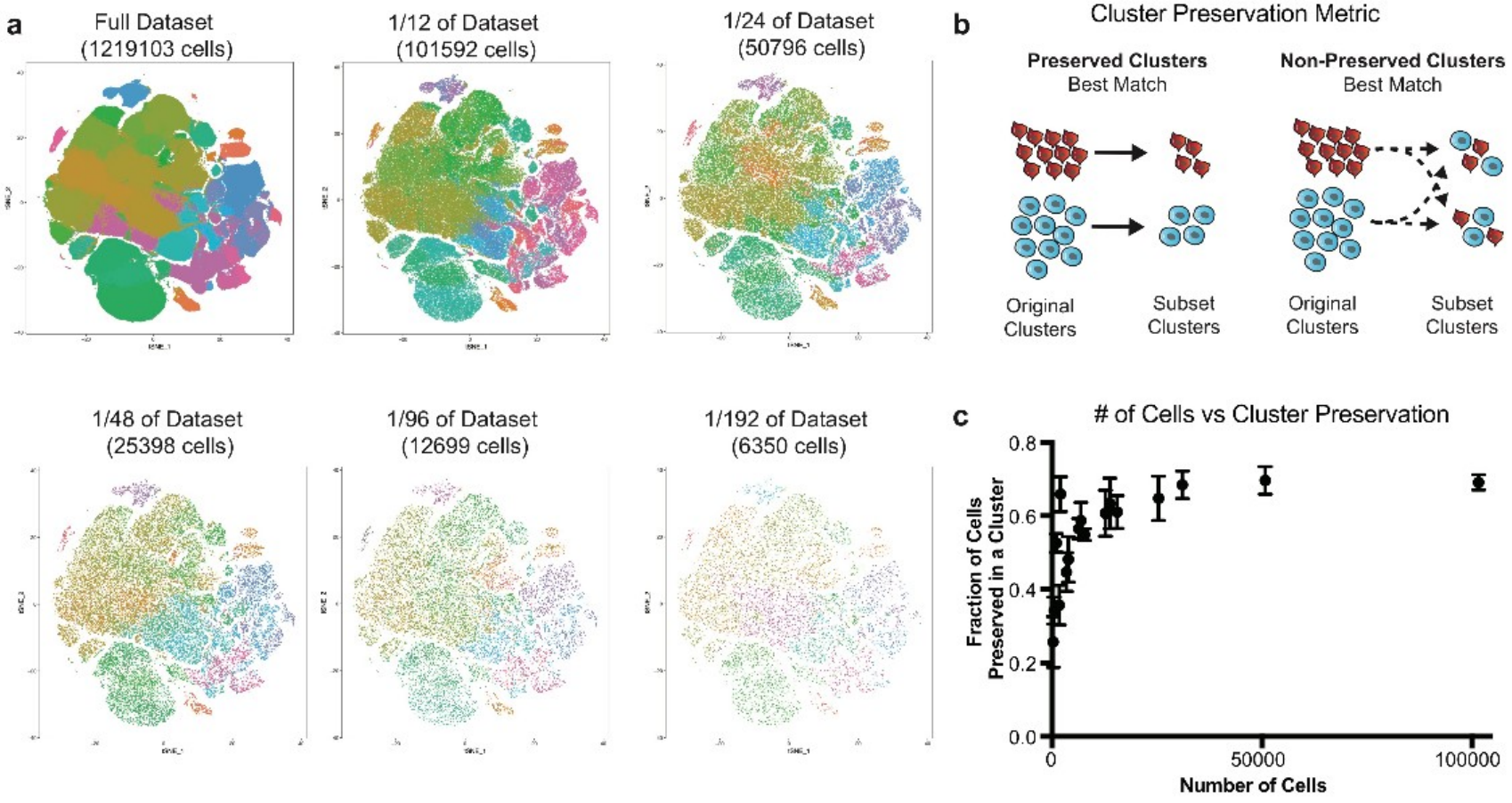
SMART-seq2:
~ 100 cells
~ 1 M RPC

full-length libraries

Droplet-based
(eg. 10x):
~ 10000 cells
~ 50 k RPC

3' librairies
UMI

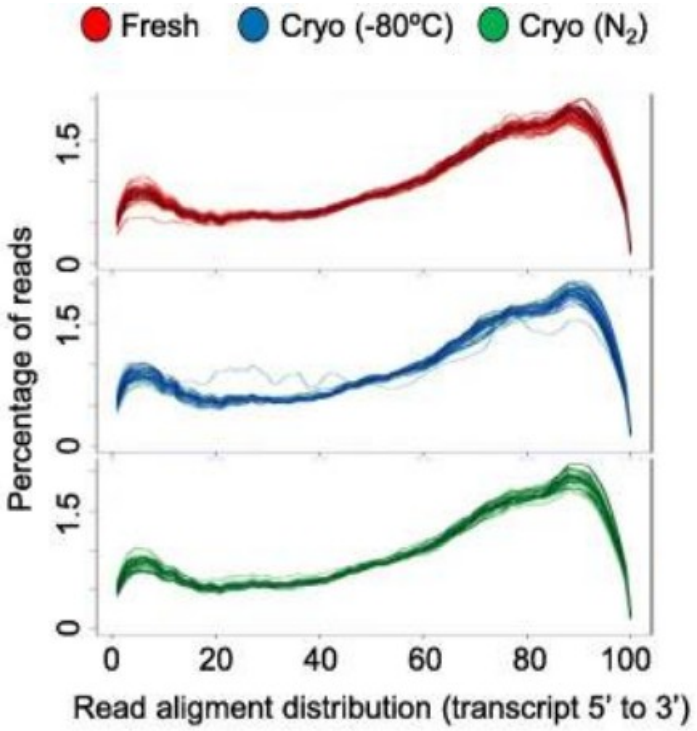
More cells, or more genes?



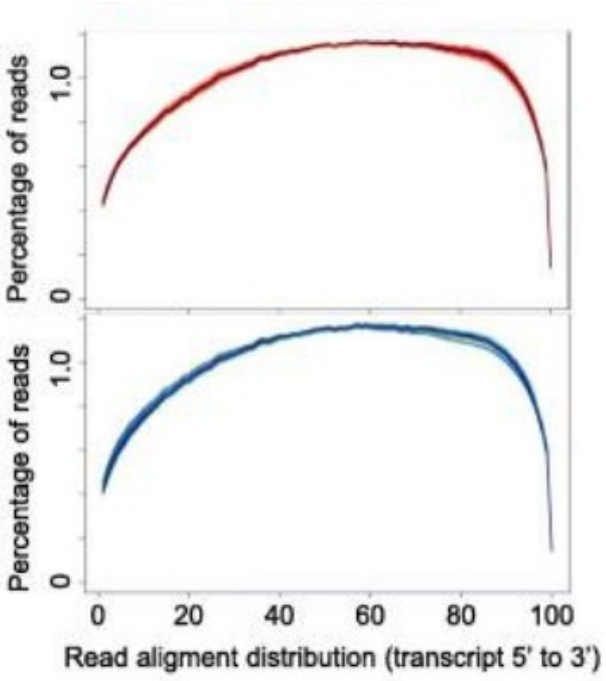
Bhaduri et al. biorXiv (2017)

Transcript coverage

droplet-based (10x):



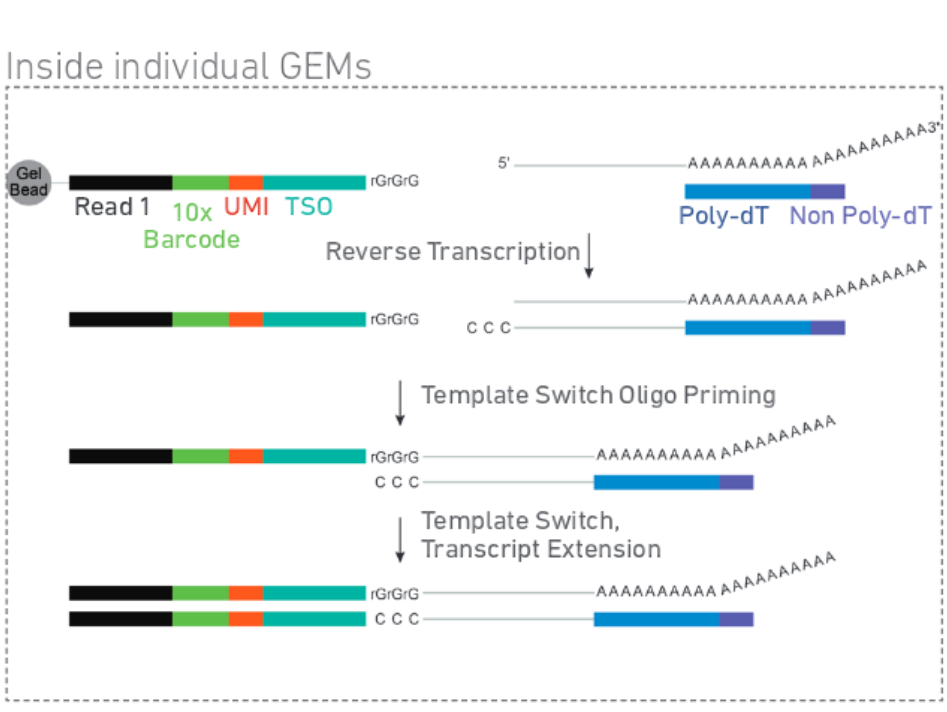
SMARTseq2:

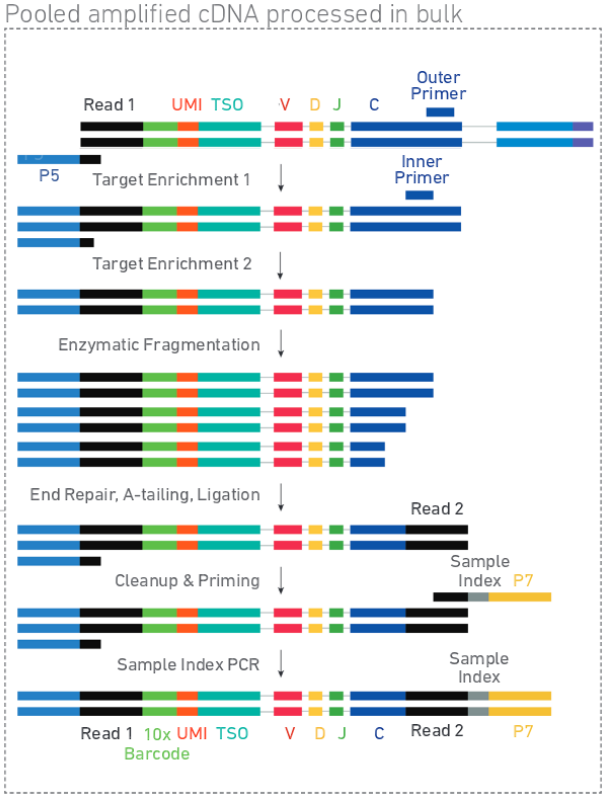


Guillaumet-Adkins et al.
Genome Biol (2017)

Application: TCRseq

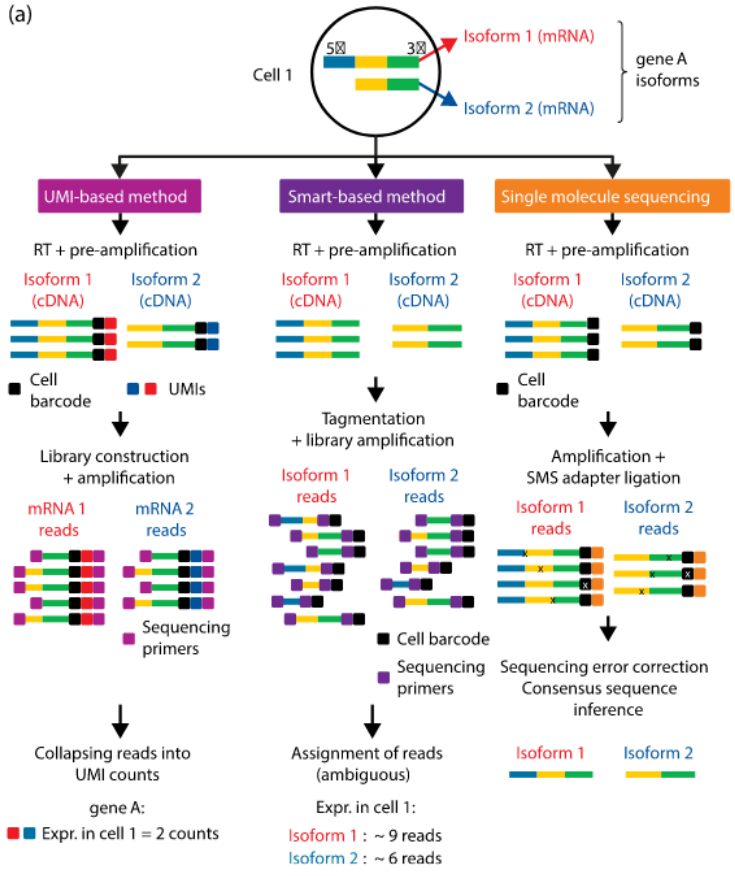
- 3’ libraries : detection of rearranged TCR is possible in 1-2 % or enriched T cells
- 5’ libraries : detection is possible in 100 % of the cells.

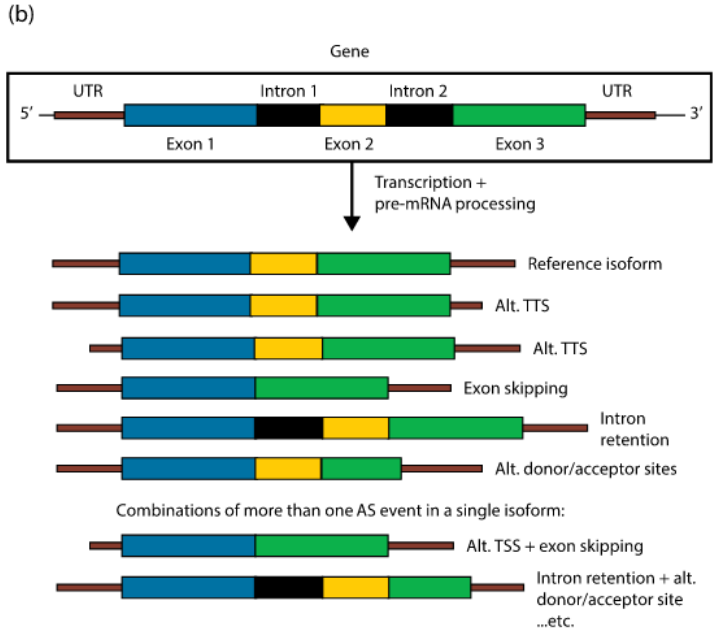




Application: splicing variants

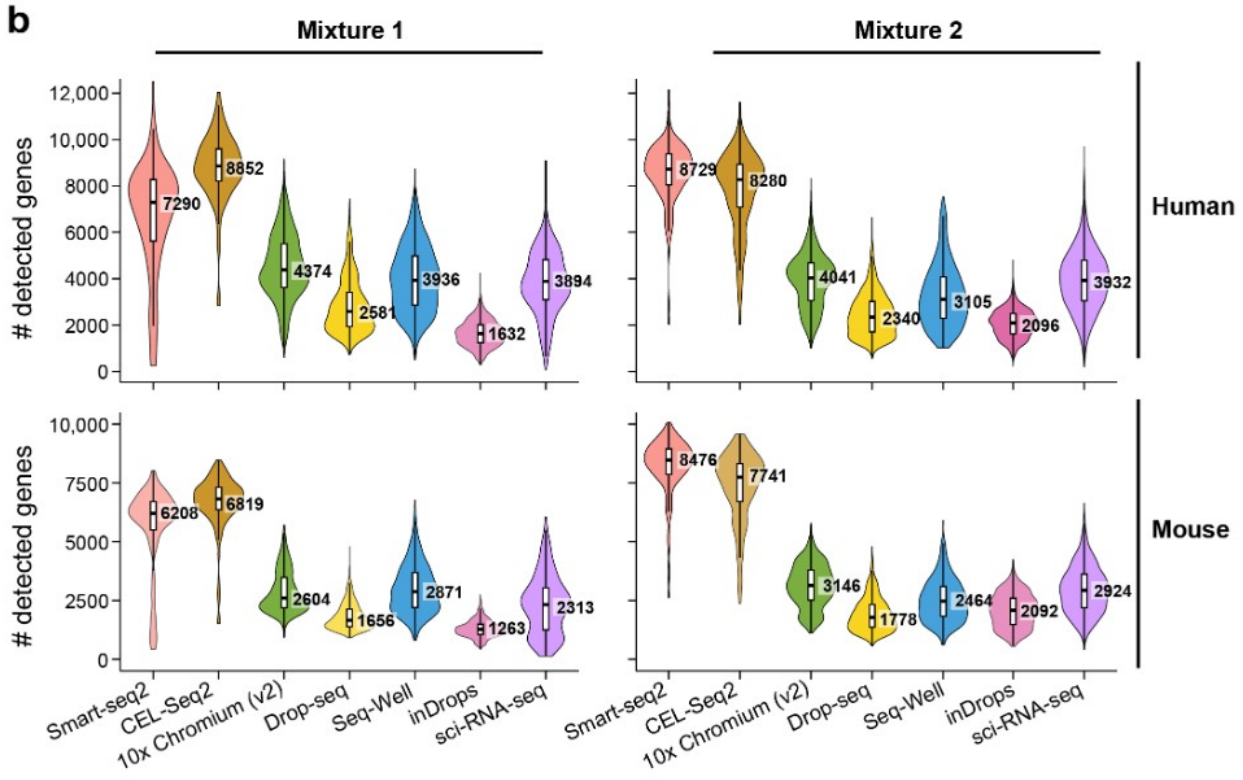
Depending on the location of the splice locus + the transcript coverage, isoforms can be detected (see velocyto for specific applications).





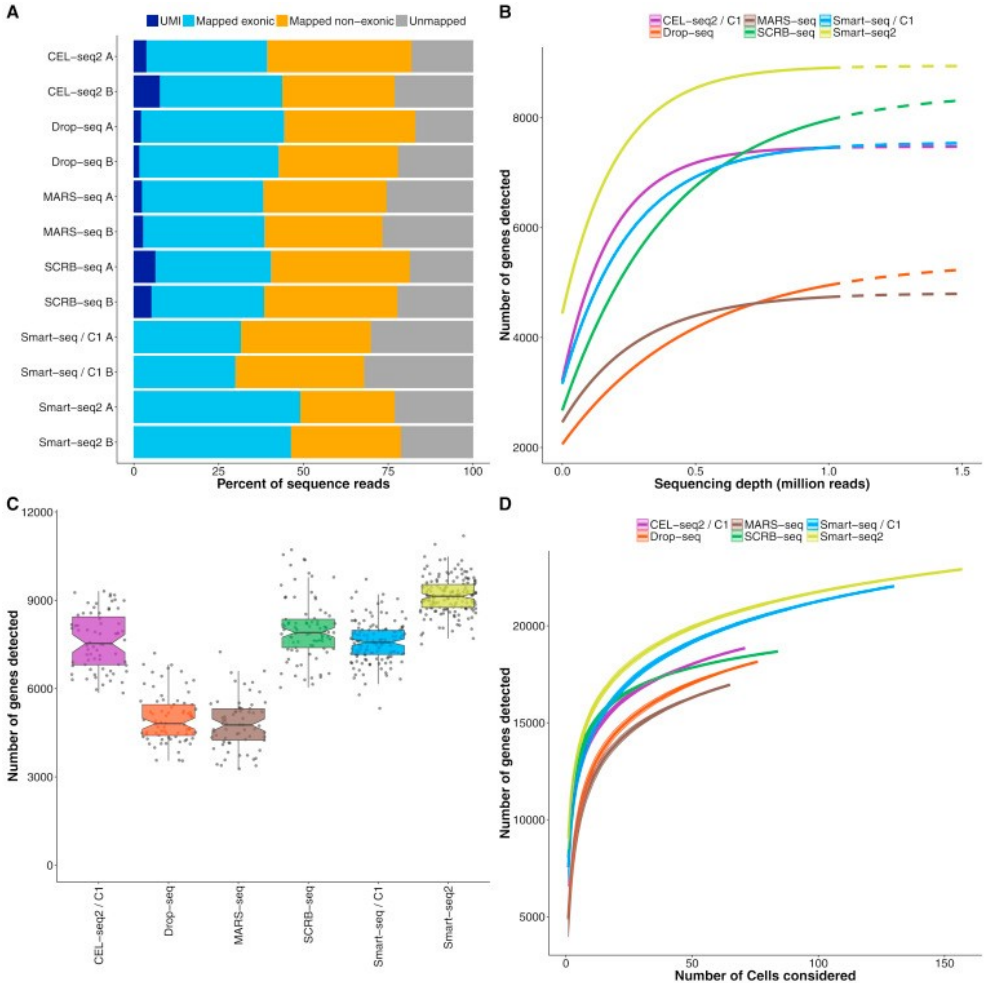
Arzalluz-Luque et al., Genome Biology (2018)

Comparative sensitivity of scRNAseq technologies



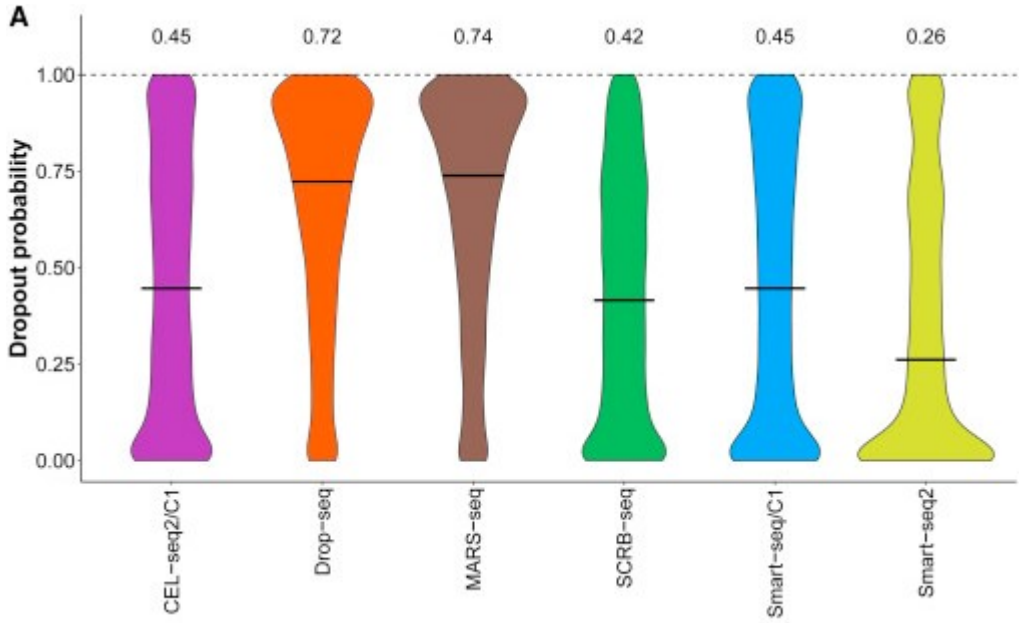
Ding, Adiconis, Simmons et al., biorXiv (2019)

Comparative sensitivity of scRNAseq technologies



Ziegenhain et al. Molecular Cell (2017)

Drop-out across technologies



Ziegenhain et al. Molecular Cell (2017)

Key point : whatever the sc technology, not detecting any transcript is not a proof the gene isn't expressed.

A practical consideration

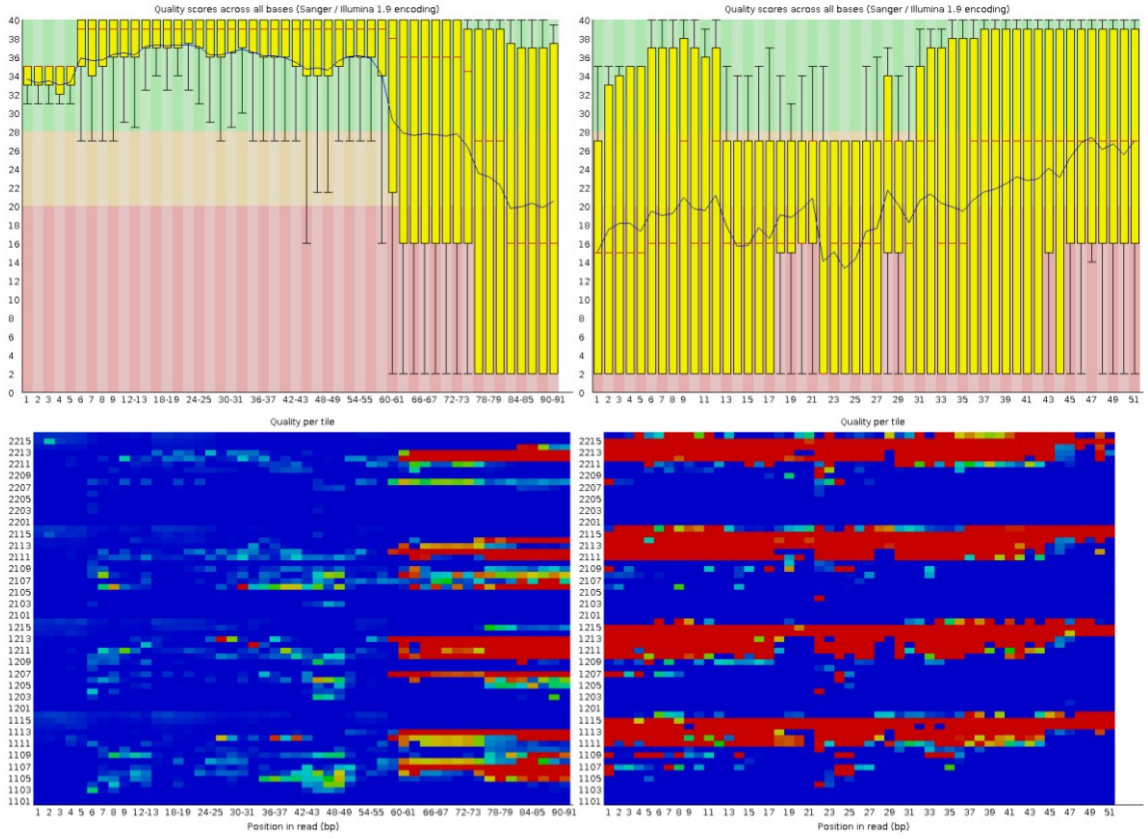
Table 1. Cost Efficiency Extrapolation for Single-Cell RNA-Seq Experiments

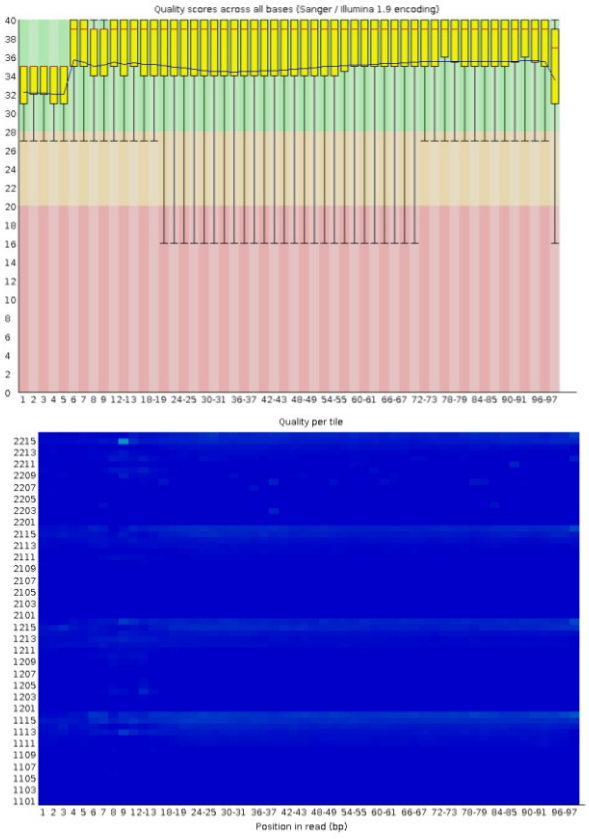
Method	TPR[a]	FDR[a] (%)	Cell per Group[b]	Library Cost ($)	Minimal Cost[c] ($)
CEL-seq2/C1	0.8	~6.1	86/100/110	~9	~2,420/2,310/2,250
Drop-seq	0.8	~8.4	99/135/254	~0.1	~1,010/700/690
MARS-seq	0.8	~7.3	110/135/160	~1.3	~1,380/1,030/820
SCRB-seq	0.8	~6.1	64/90/166	~2	~900/810/1,080
Smart-seq/C1	0.8	~4.9	150/172/215	~25	~9,010/9,440/11,290
Smart-seq2 (commercial)	0.8	~5.2	95/105/128	~30	~10,470/11,040 /13,160
Smart-seq2 (in-house Tn5)	0.8	~5.2	95/105/128	~3	~1,520/1,160/1,090

Ziegenhain et al. Molecular Cell (2017)

primary analysis & data generation

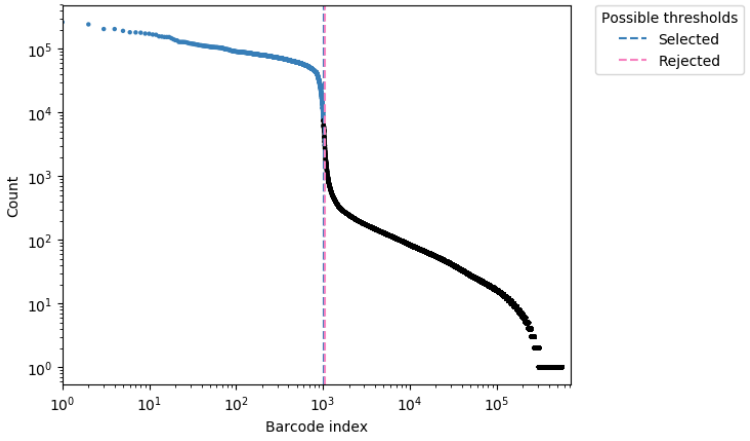
QC pre-check: quality of sequenced reads

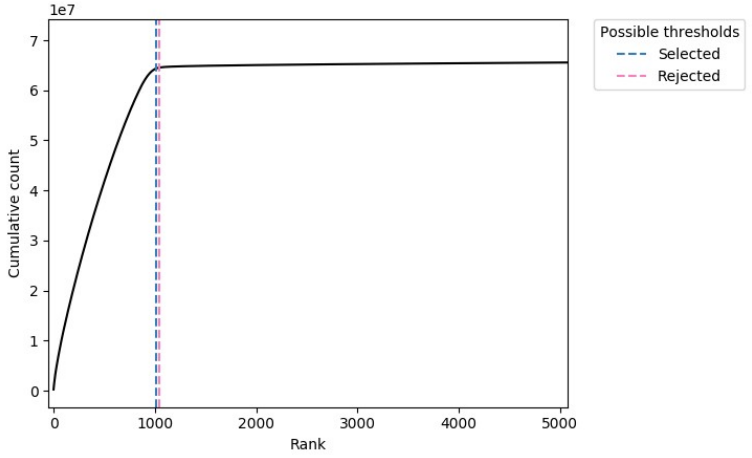


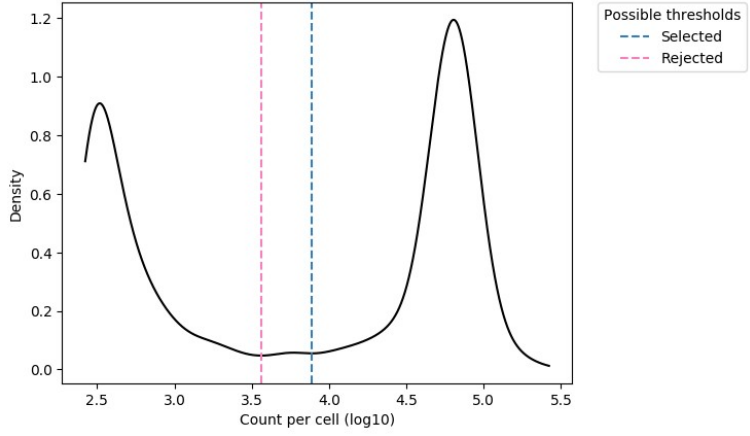


Positional quality of the sequenced reads (Phred scores). Bottom-left: experiment with a flowcell issue. Inspecting the quality of the sequencing (eg. fastqc, reads above Q30 in CR report…) is recommended.

Cell calling in droplet-based technologies







Mapping or transcript quantification

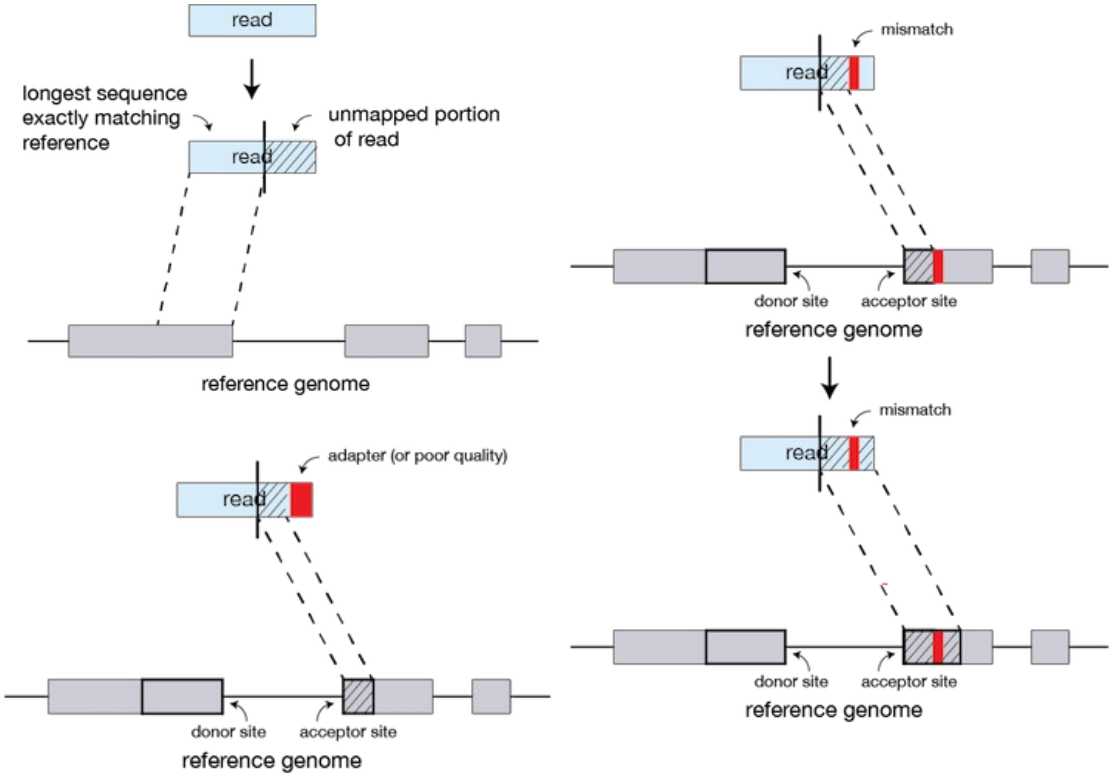
mapping engines:

- tophat, bowtie2, STAR

alignment-free transcript quantification:

- RNASkim, eXpress, kallisto, salmon

Transcript mapping (eg. STAR)



	union	intersection_strict	intersection_nonempty
read / gene_A	gene_A	gene_A	gene_A
read / gene_A	gene_A	no_feature	gene_A
read / gene_A gene_A	gene_A	no_feature	gene_A
read read / gene_A gene_A	gene_A	gene_A	gene_A
read / gene_A gene_B	gene_A	gene_A	gene_A
read / gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read / gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? / gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

1. Sequenced reads (fastq file) + reference genome = alignements (SAM/BAM file)
2. Feature quantification (eg. FeatureCounts, HTseq)

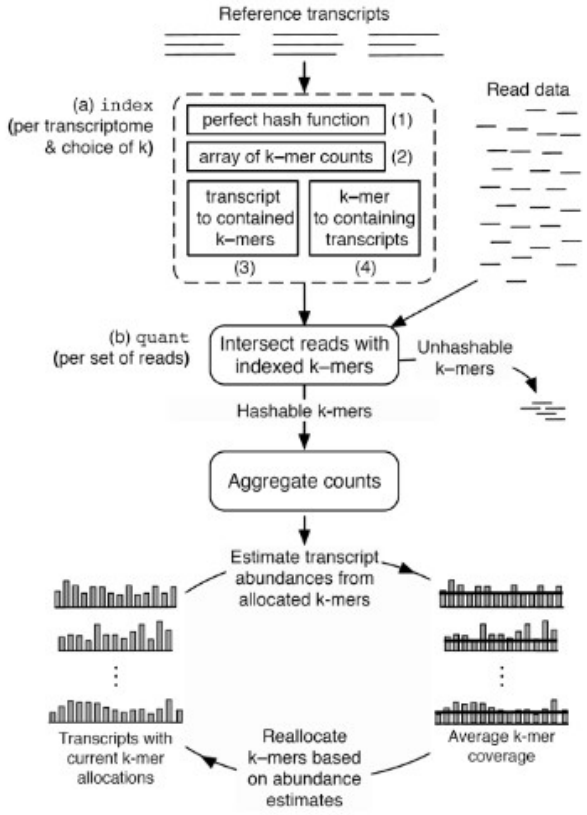
Transcript quantification, quasi-mapping (eg. Salmon)

sequence ATGGAAGTCGCGGAATC

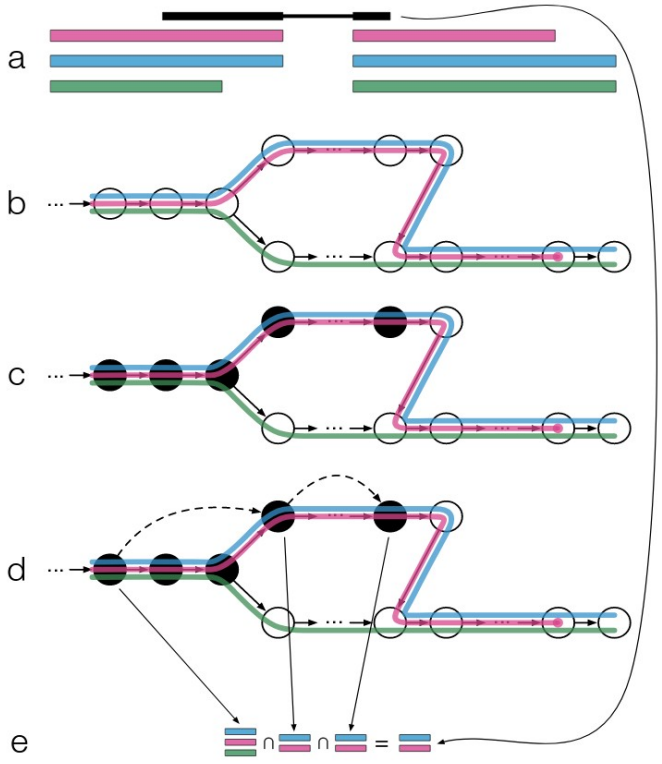
7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

Start : fastq + reference transcriptome
result :



Patro et al., Nature Biotechnology, 2014



Bray et al., arXiv, 2015

1. Sequenced reads (fastq file) + reference transcriptome = count matrix (usually TPM)

RNAseq expression units

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

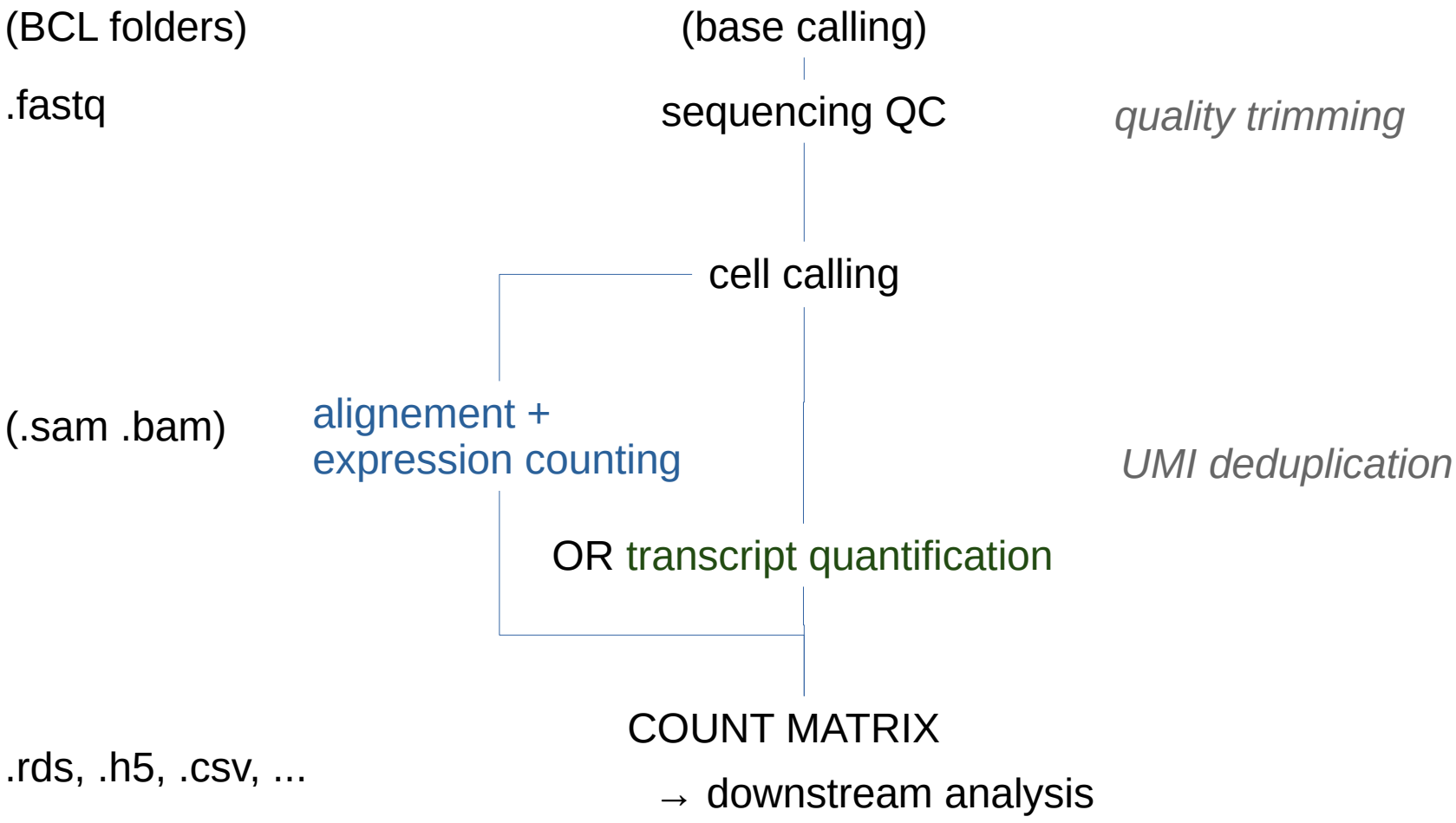
$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}}\right) \cdot 10^6$$

With:

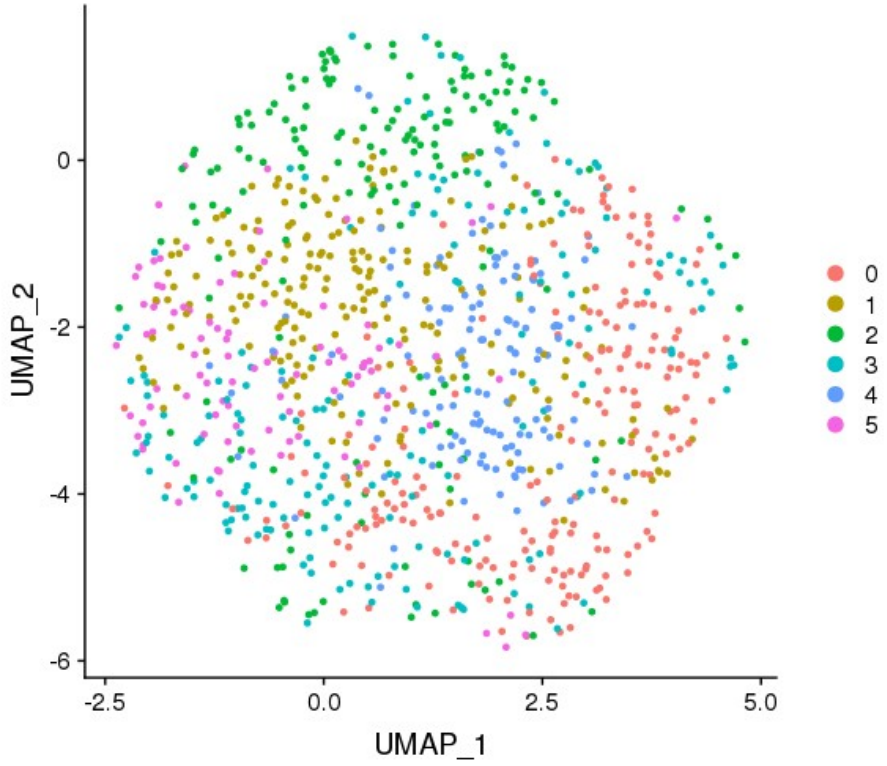
- Xi: observed count
- li: length of the transcript
- N number of fragments sequenced

Summary of primary analysis



downstream applications

Data partitioning and cell clustering



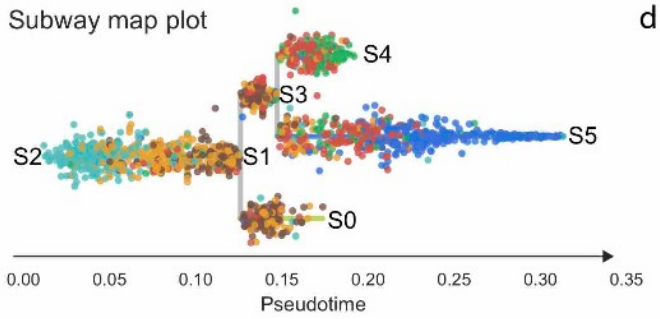
graph-based clustering, Seurat v3, resolution=0.8

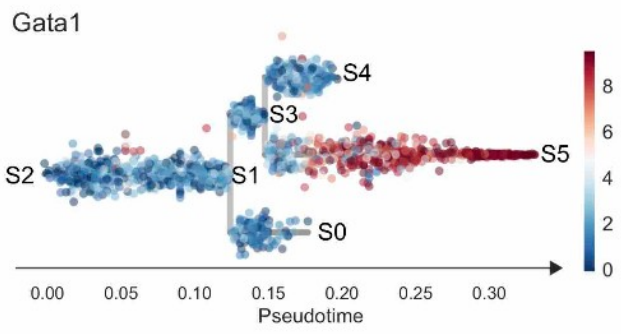
```
emat <- Matrix::Matrix(data=extraDistr::rzinb(25000*1000, 50, 0.95, 0.75) \
                                                    , nrow=25000, ncol=1000, sparse=TRUE)


emat[1:10,1:5]

## 10 x 5 sparse Matrix of class "dgCMatrix"

##        cell1 cell2 cell3 cell4 cell5
## gene1      .     2     .     .     .
## gene2      .     2     .     3     .
## gene3      .     .     .     .     2
## gene4      7     .     .     .     3
## gene5      1     .     .     .     1
## gene6      .     .     .     .     6
## gene7      .     .     .     .     3
## gene8      .     .     2     .     .
## gene9      .     3     .     .     .
## gene10     .     3     .     .     6
```

Application: cell heterogeneity

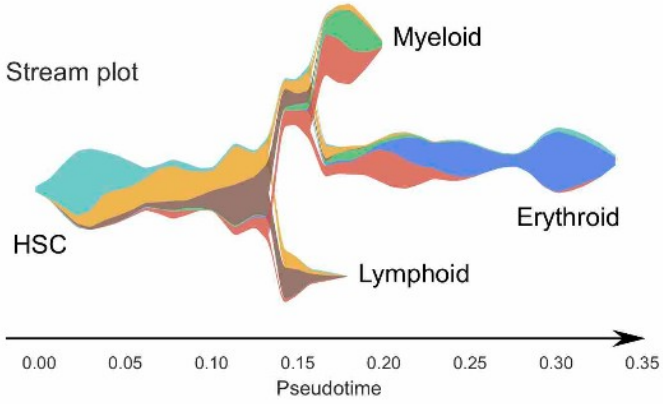
- How to define a cell subset? Correlation with a cell cluster?
- Any matrix can be mathematically partitioned
- A discrete partitioning of the data is not always desirable: continuous scales are more adapted to dynamic processes such as cell differentiation.

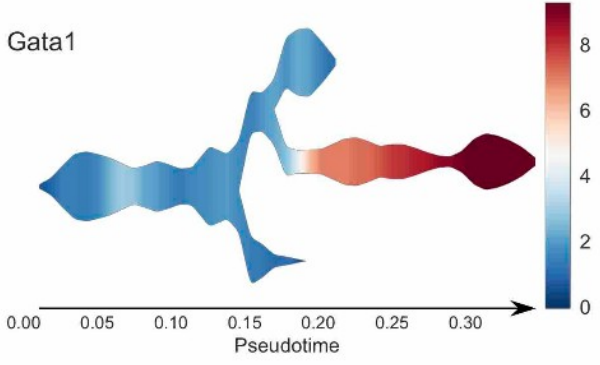
Application: transcriptional dynamics and differentiation processes

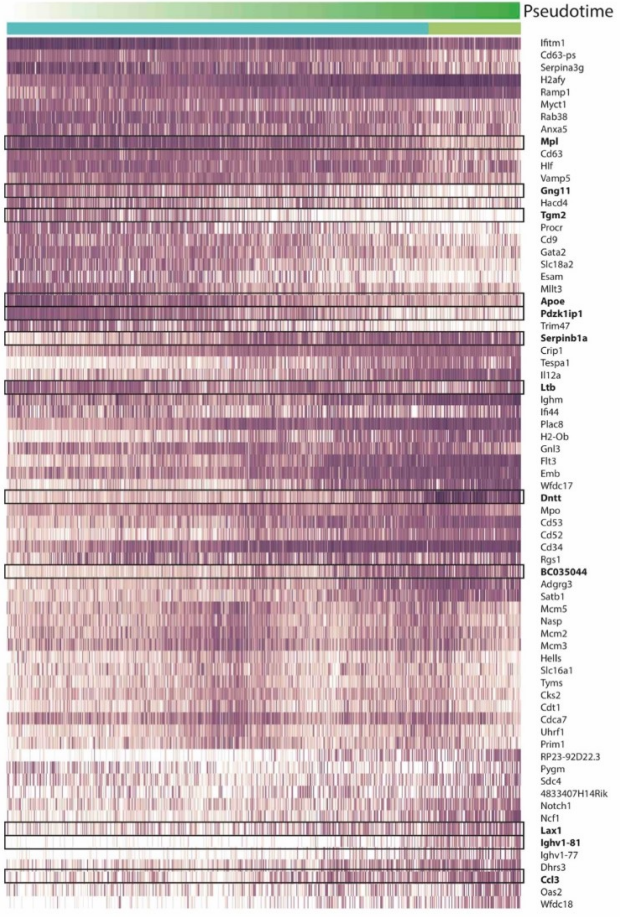




d

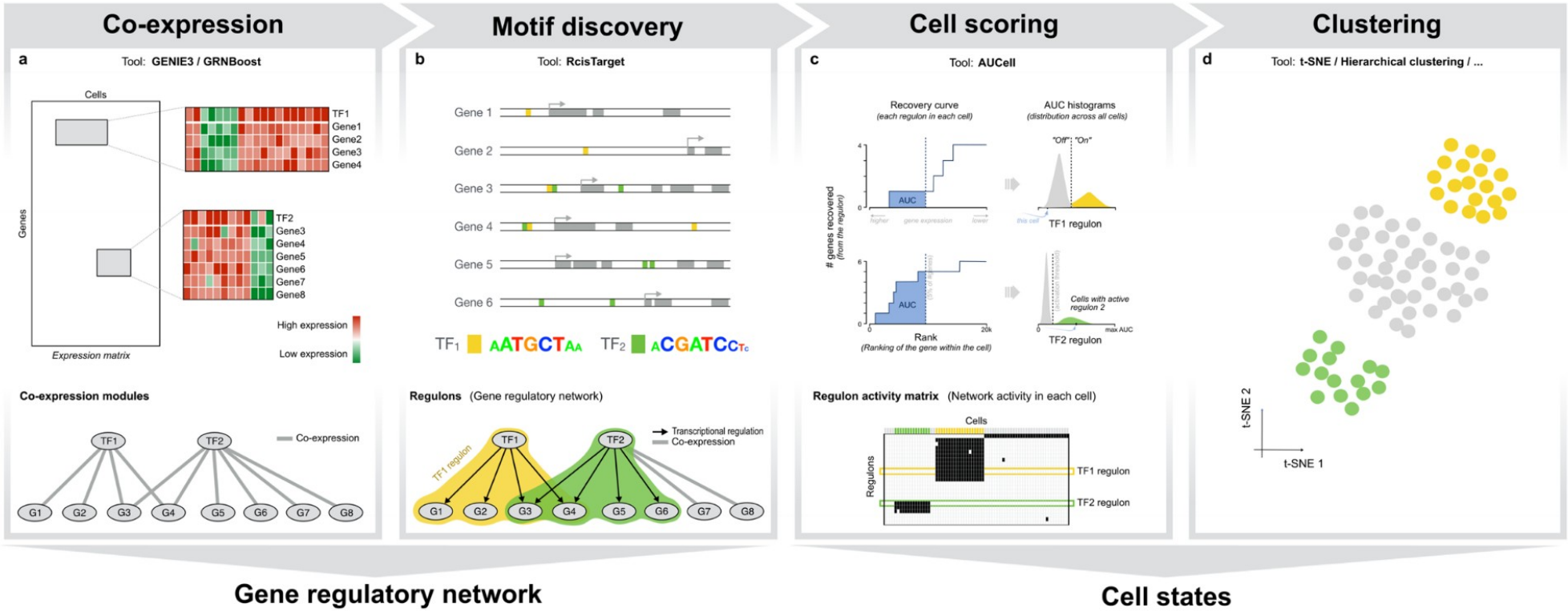






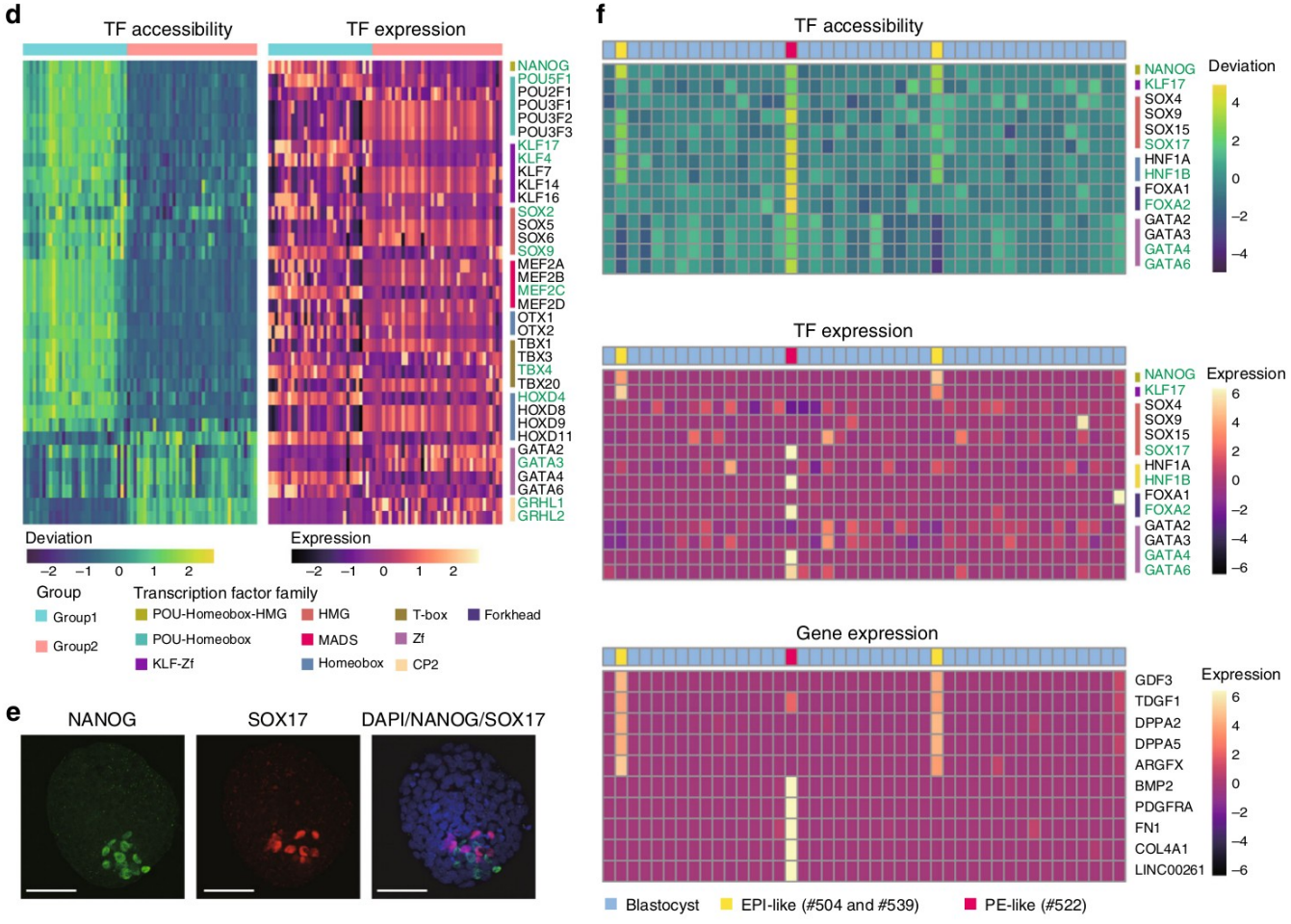
Chen, Albergante et al. Nature Communications (2019)

Application: identification of gene regulatory modules (SCENIC, Aerts lab)



Aibar et al. Nature Methods (2017)

Application: scRNAseq & scATACseq

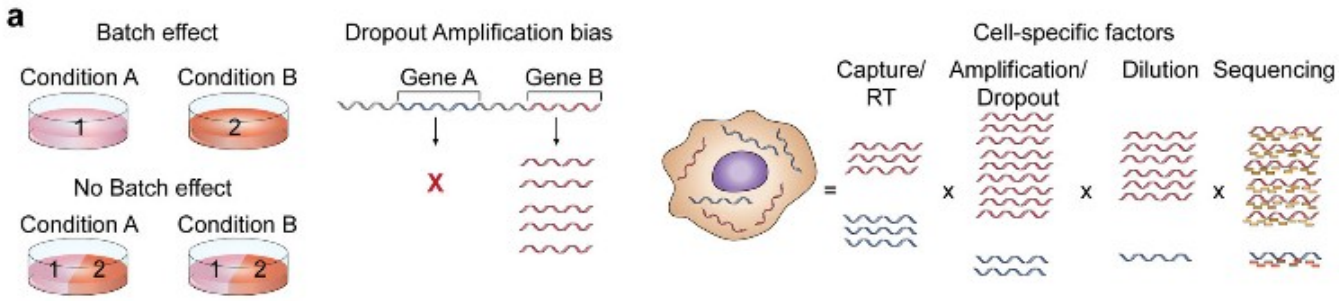


scCAT-seq : mild lysis approach and a physical dissociation strategy to separate the nucleus and cytoplasm of each single cell.
The supernatant cytoplasm component is subjected to the Smart-seq2 method.
The precipitated nucleus is then subjected to a Tn5 transposase-based and carrier DNA-mediated protocol to amplify the fragments within accessible regions.

Liu, Liu, Quintero, Wu, Yuan et al. Nature Communication (2019)

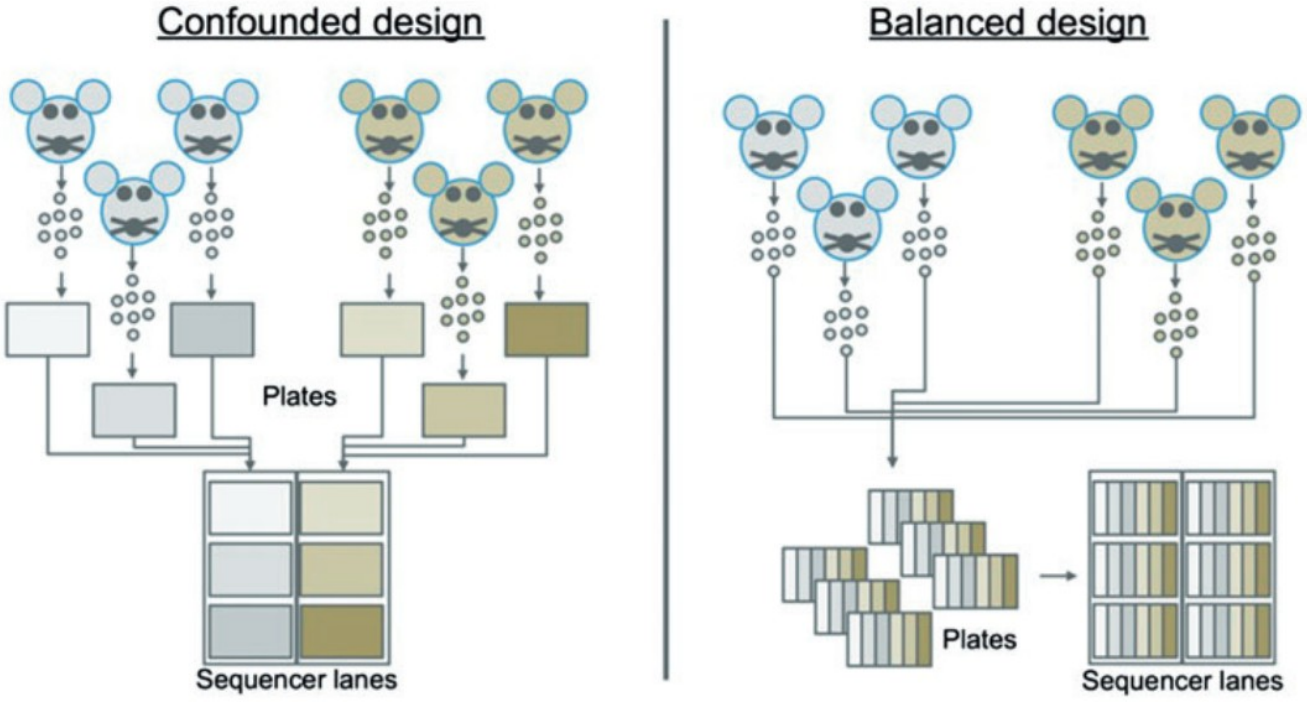
experimental and technical biases

Observed transcript counts are the combination of factors



Hwang et al. Experimental & Molecular Medicine (2018)

Confounded designs in scRNAseq



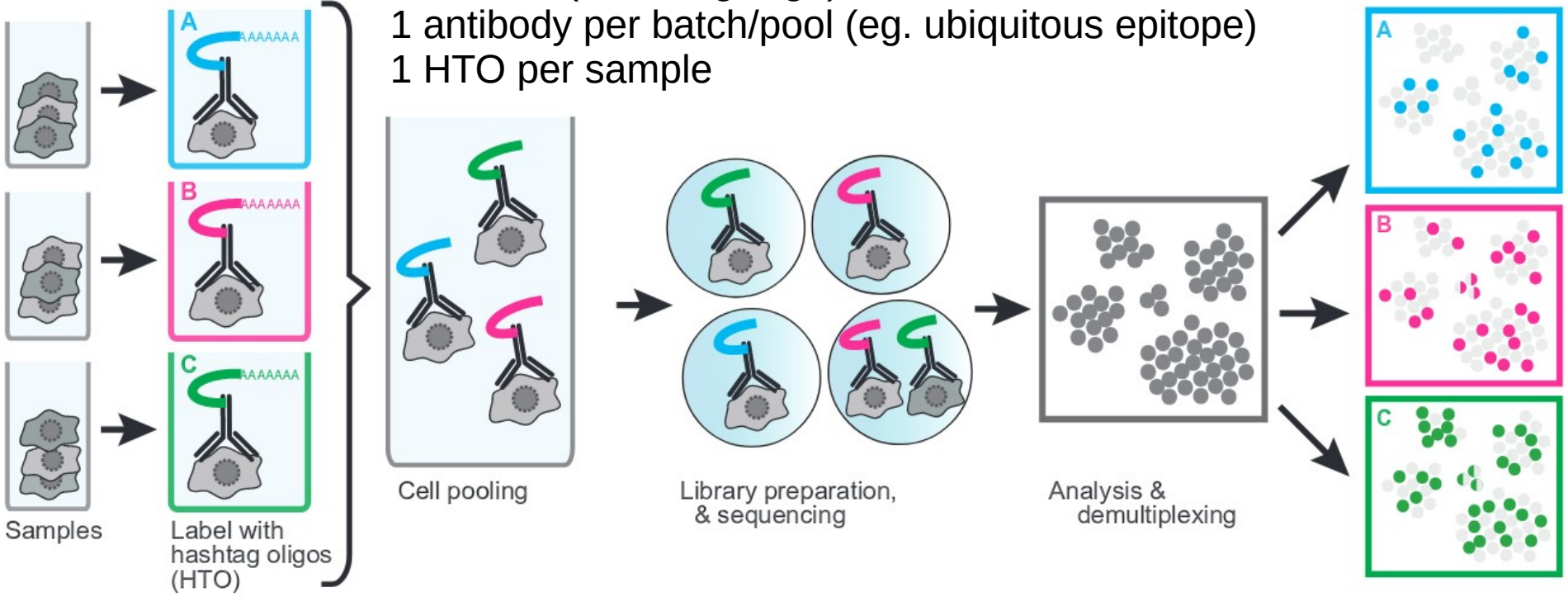
Baran-Gale et al. Briefings in Functional Genomics (2018)

Experiments on human samples can hardly be pooled.

Due to the costs and experimental constrains, droplet-seq experiments are frequently confounded in their design.

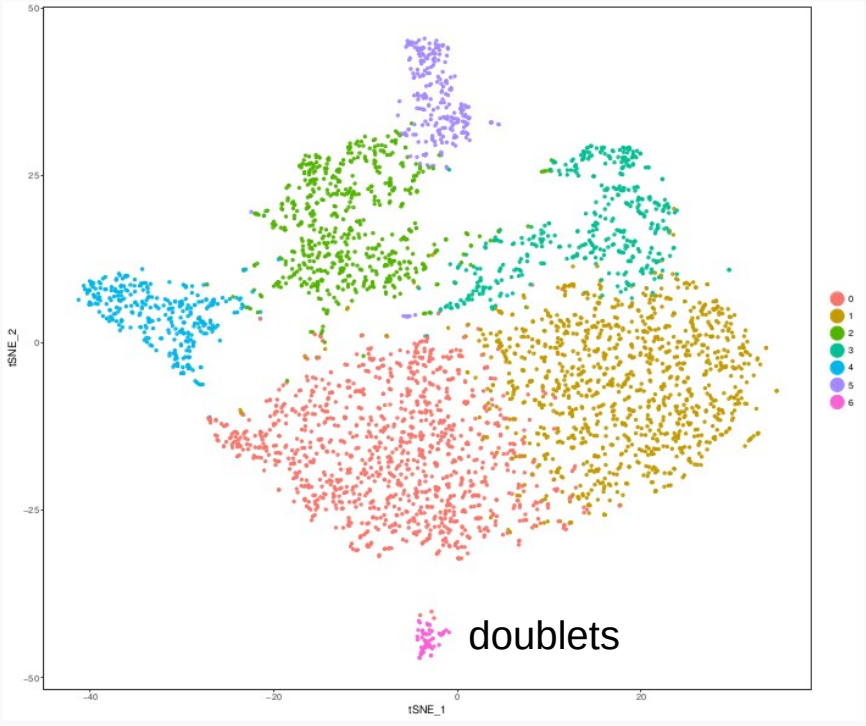
Using cell hashing to resolve confounding experimental designs

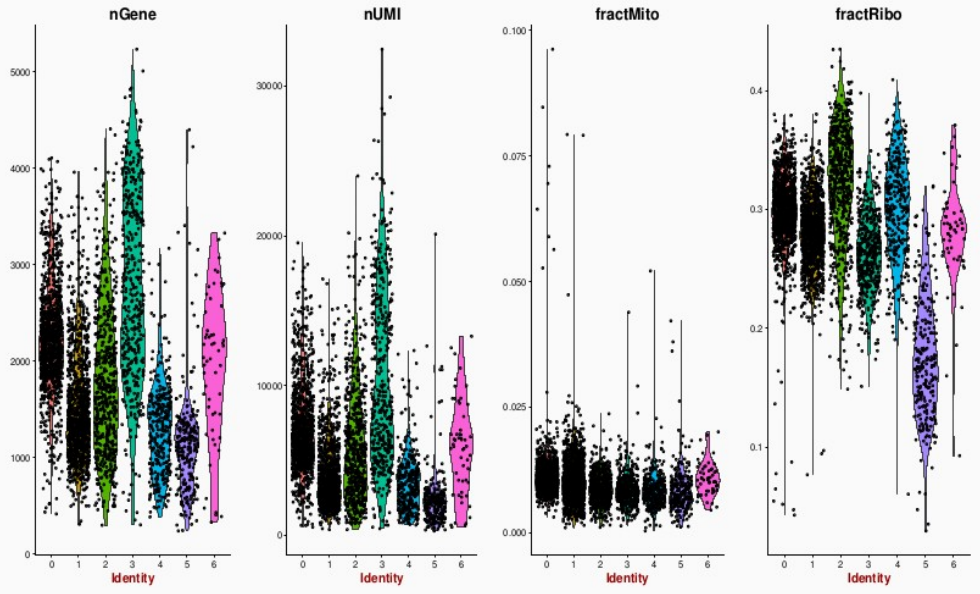
barcoded (HashTagOligo)
1 antibody per batch/pool (eg. ubiquitous epitope)
1 HTO per sample



Stoeckius & Satija, biorXiv, 2017

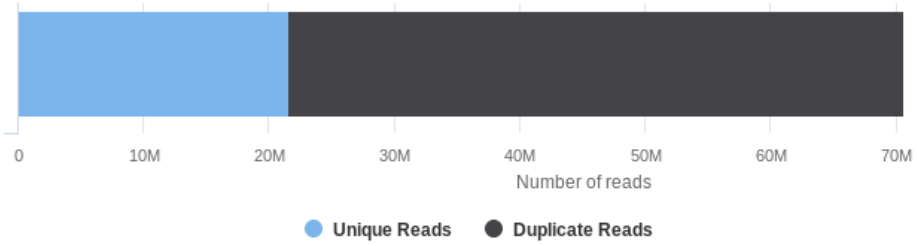
Doublets in heterogenous samples

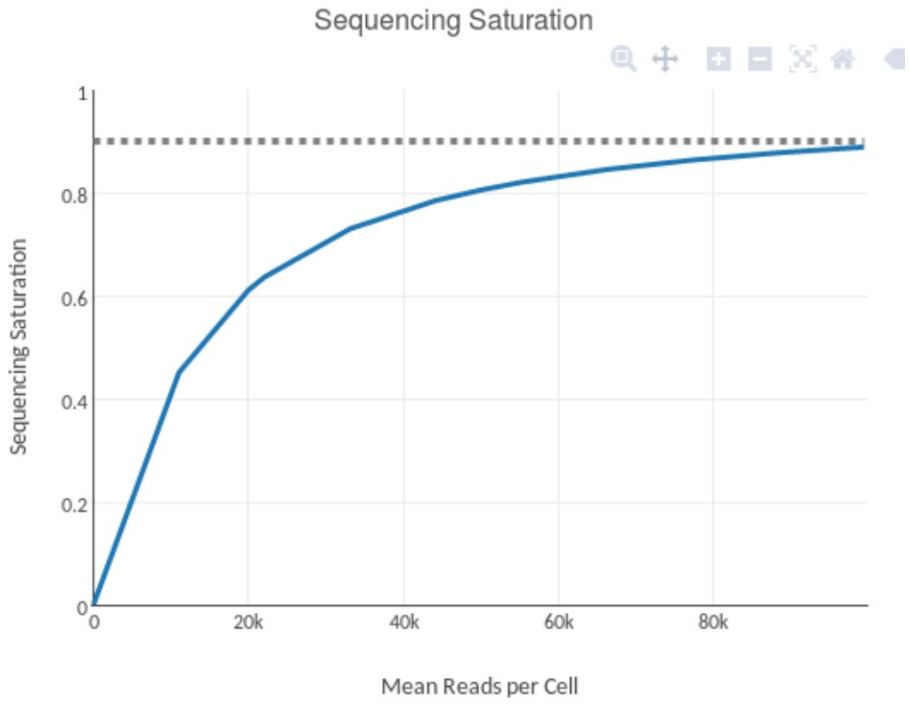




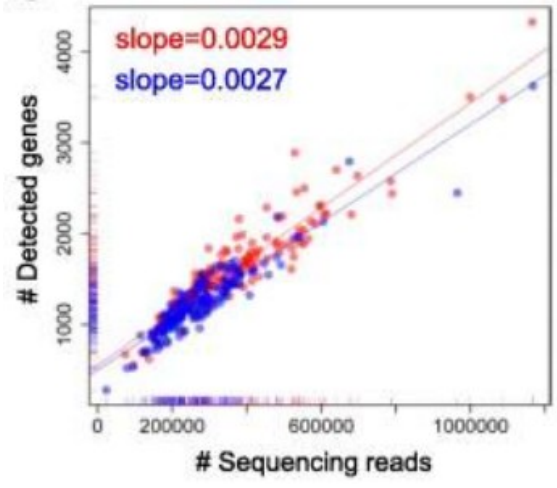
doublets cells are defined by co-expression of both T- and APC- restricted genes (an immune synapse has been captured)

Estimating the appropriate sequencing depth



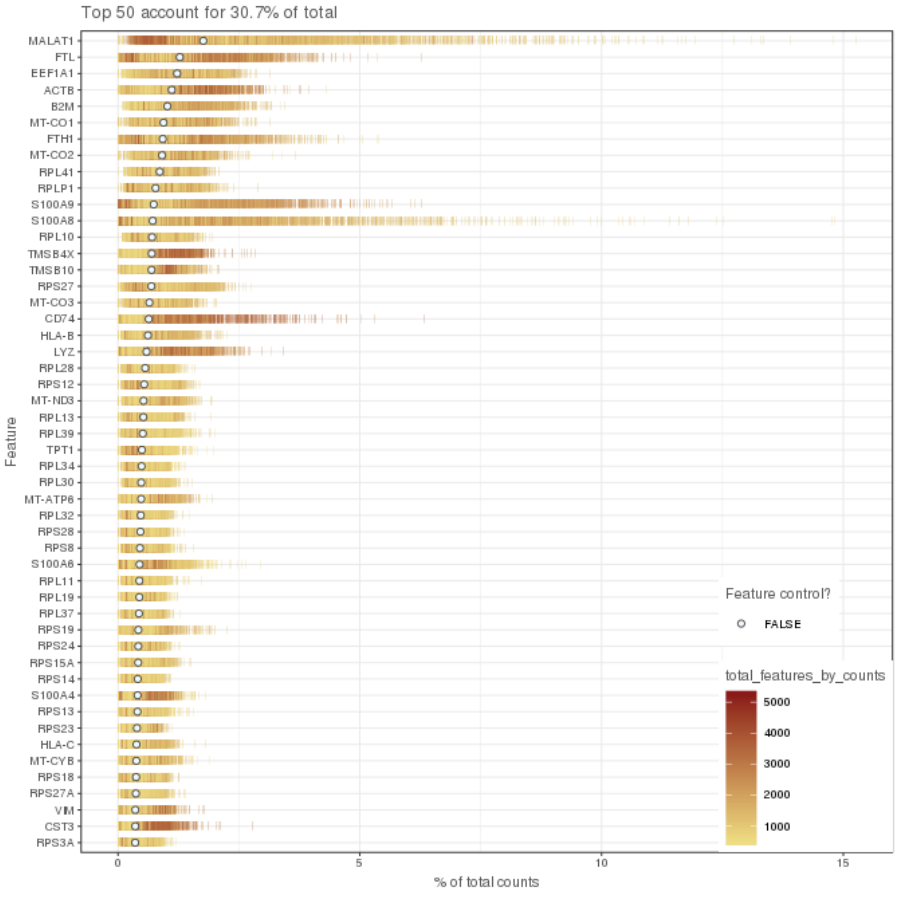


Saturation point is never achieved in scRNAseq

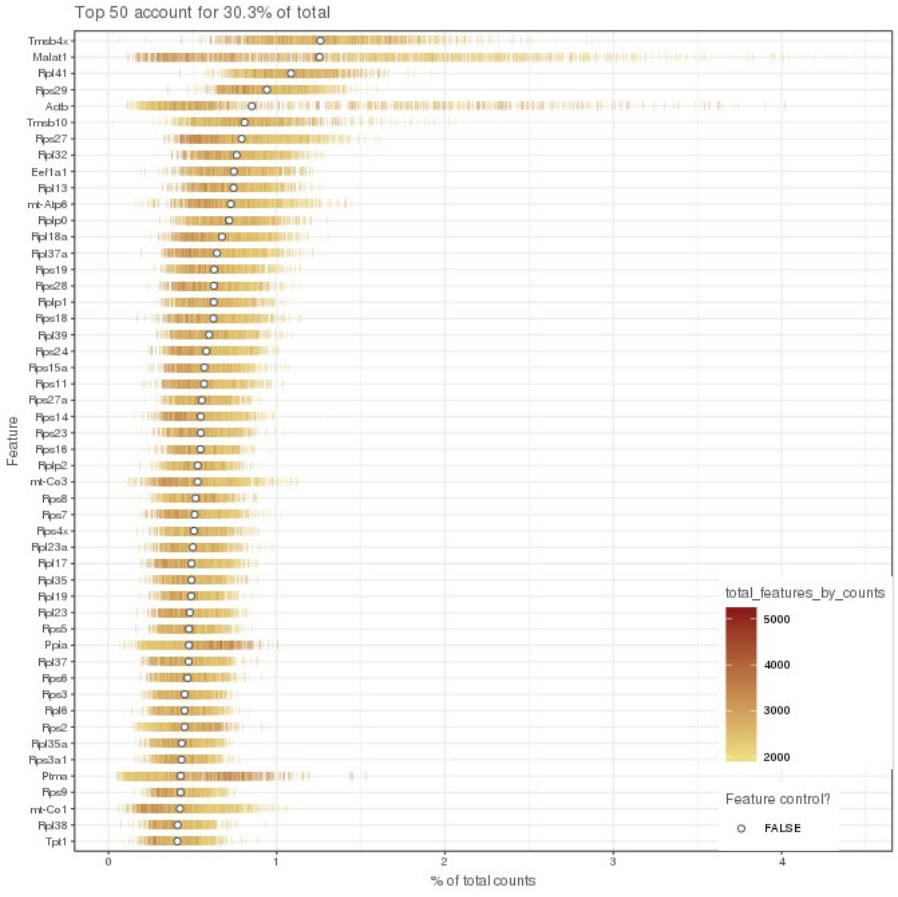


Transcripts coding for Ribosomal Proteins are abundant in cells

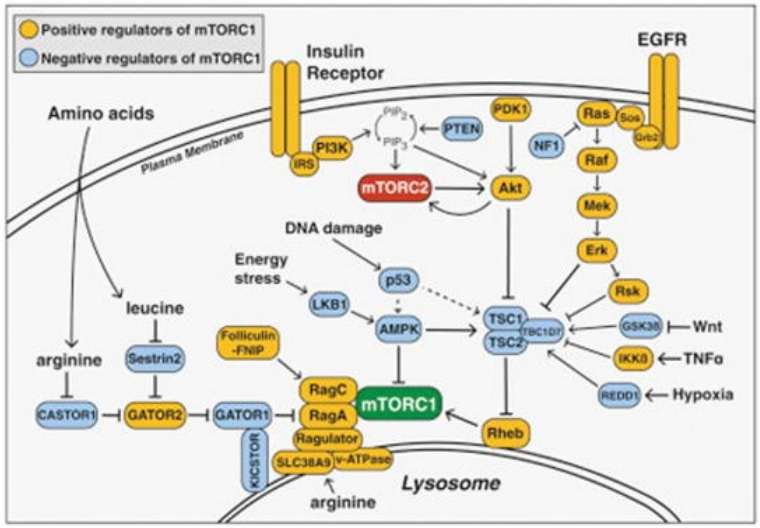
human sample - myeloid cells

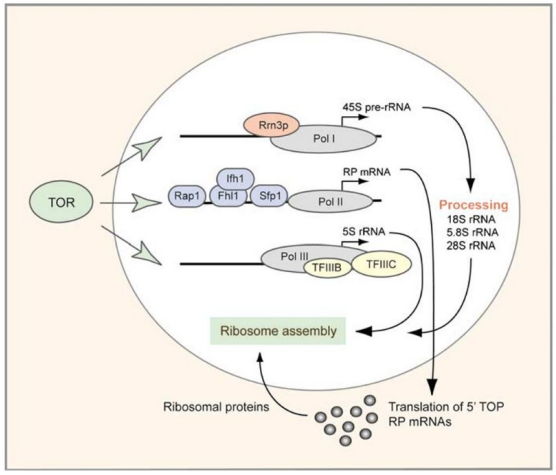


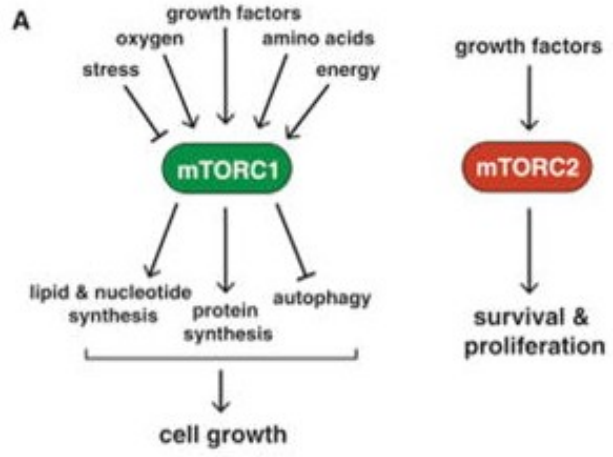
murine sample - lymphoid cells

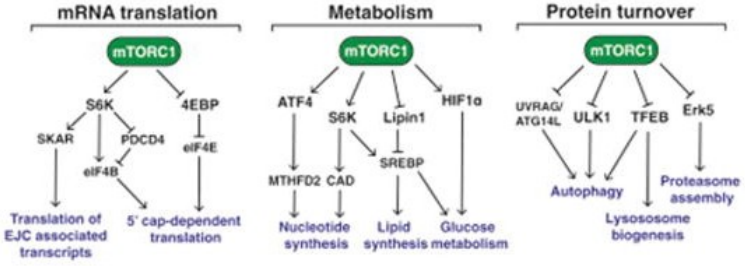


Ribosome biogenesis is quickly regulated by the cellular environment







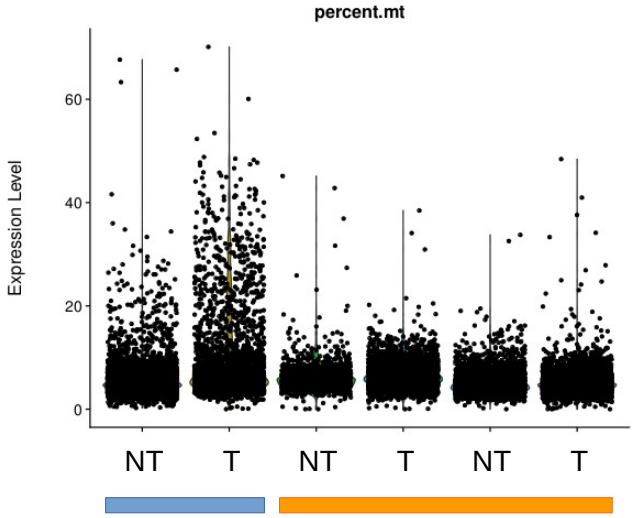


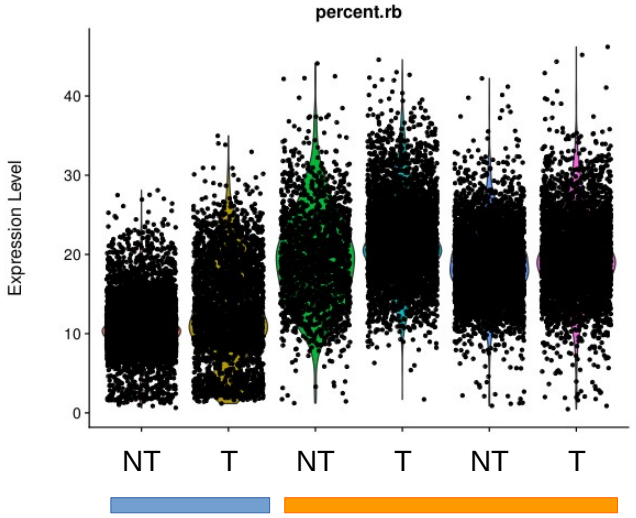
Induction of RP genes: Hi-glucose, Insulin, GFs (culture medium+SVF)

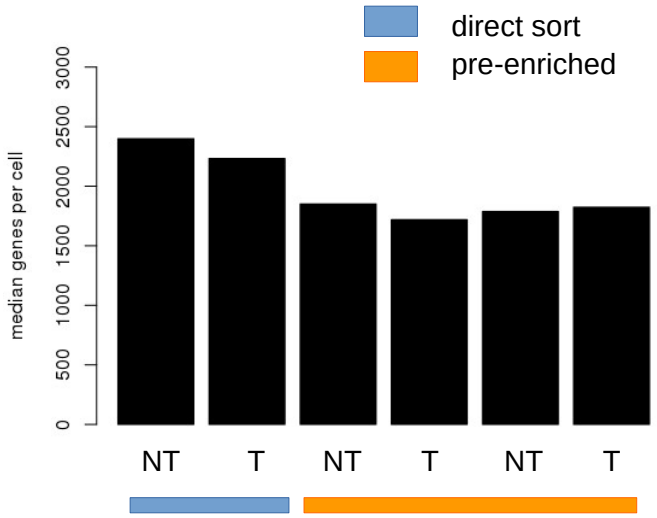
Inhibition of RP genes: nutrient deprivation, hypoxia, DNA damage

Saxton et al. Cell (2018)

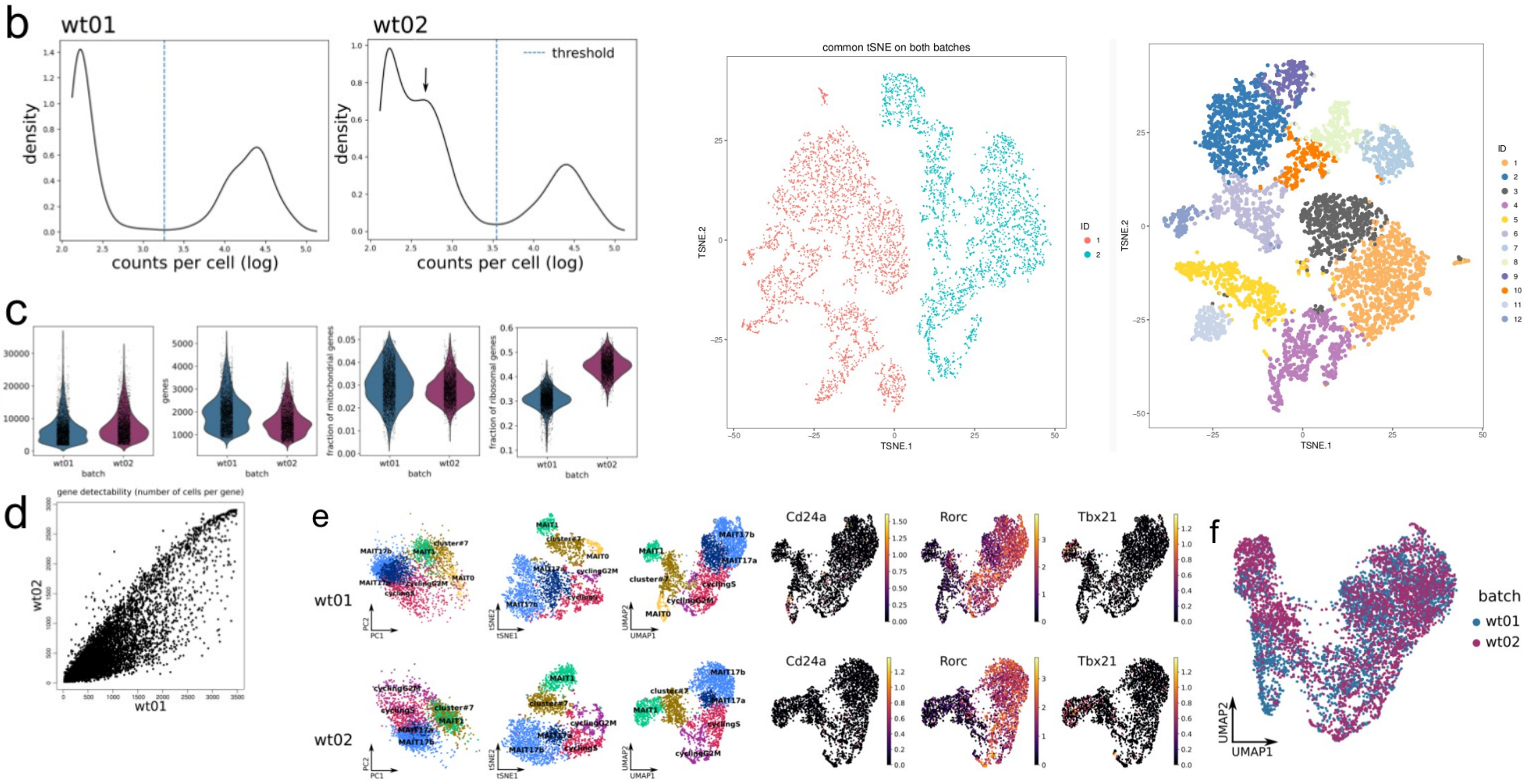
Technical artifacts: effect(s) of sample processing on gene detection



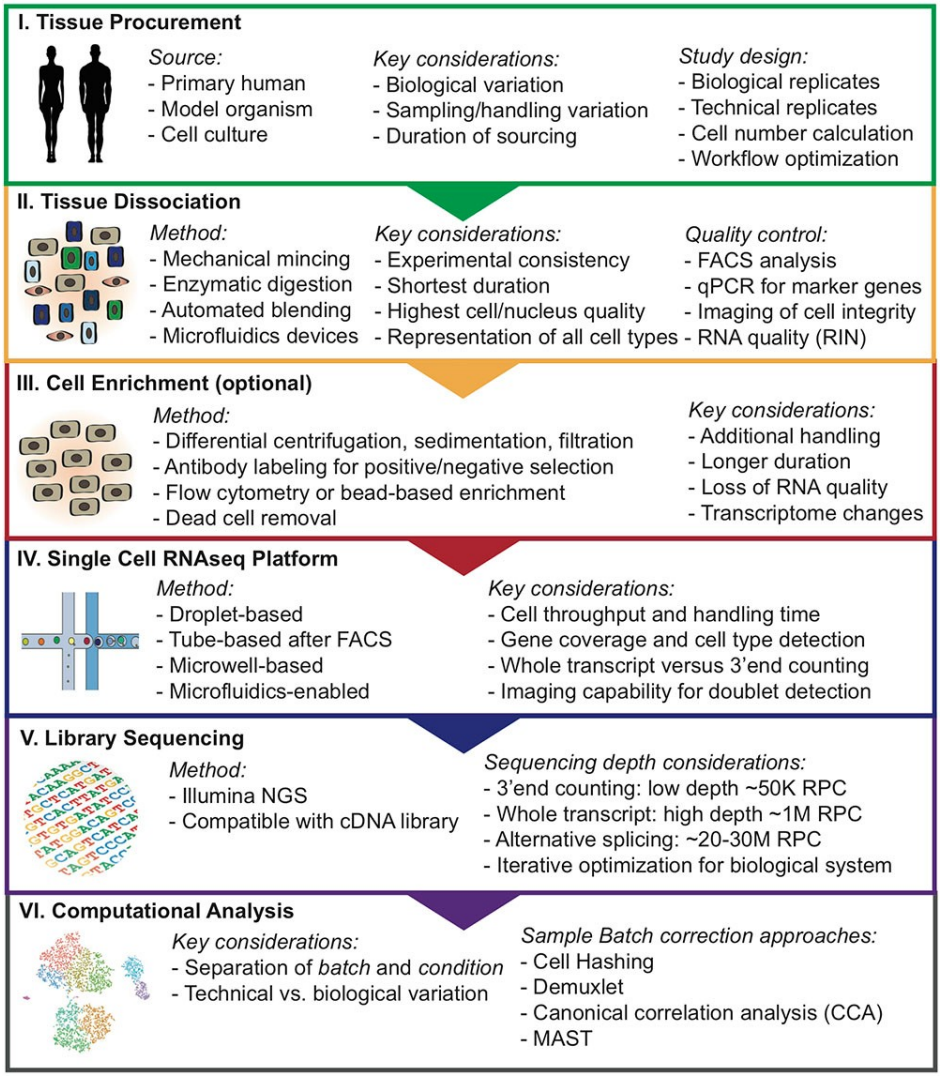




Batch effect in technical replicates (mouse littermates)



Artifacts, variations and technical limitations in scRNAseq experiments



Nguyen et al. Frontiers in Cell Developmental Biology (2018)

Summary (1)

- scRNAseq has inherent technological limitations:
  - data are noisy (dropouts)
  - lowly expressed genes can remains undetected
  - samples can be contaminated by unexpected cell types
  - samples will contain (homotypic and heterotypic) doublets
  - only specific experimental set-ups can resolve confounding design
  - replicates without any technical/batch effect are (very) unlikely

Summary (2)

- key points to consider during pre-processing of scRNAseq:
  - a good understanding of the nature of the sample is essential (sampling conditions, preparation, purity)
  - identifying the source of technical effects helps resolving them
  - before any correction of multiple batches, an individual exploration of single samples is highly recommended