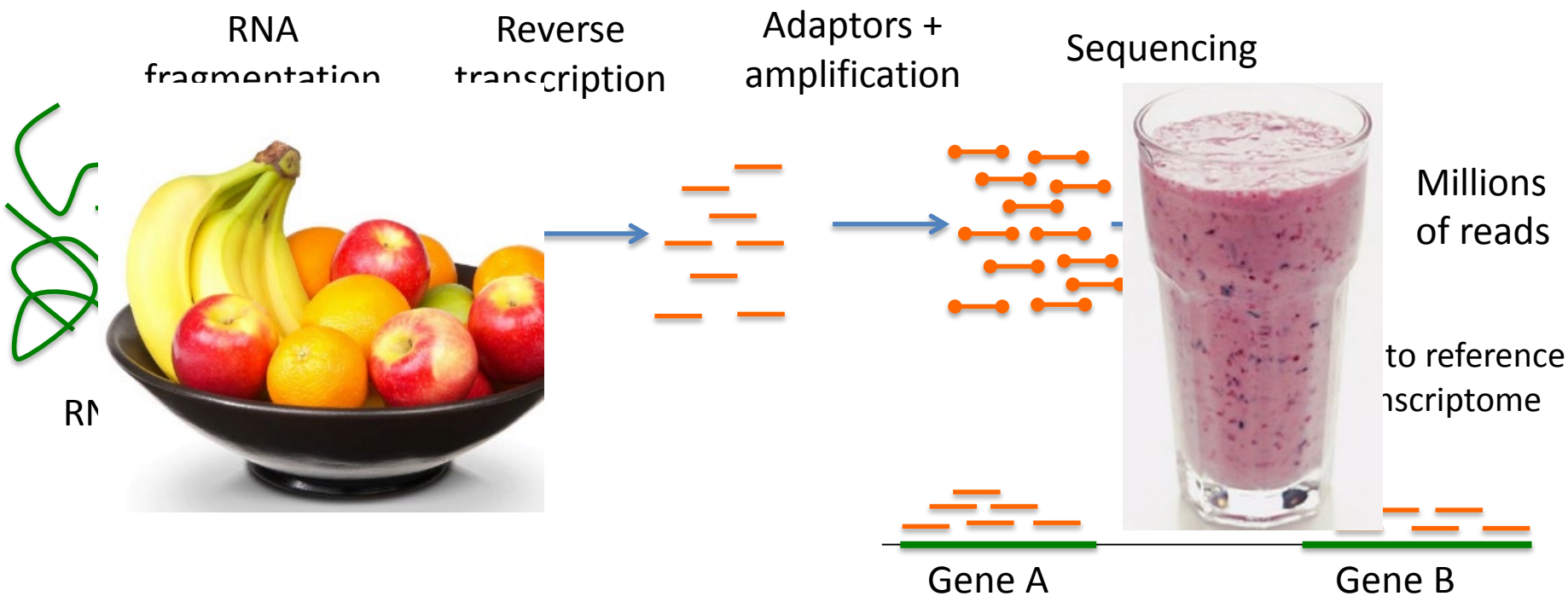
Single-cell RNA-Seq data: a (very) short overview

March 13, 2020

Dr. ir. Perry D. Moerland
Bioinformatics Laboratory
Amsterdam UMC
p.d.moerland@amsterdamumc.nl
www.bioinformaticslaboratory.nl

RNA-Seq



Gene expression => Quite easy (count the reads)

Transcript expression => More difficult (but can be done probabilistically)

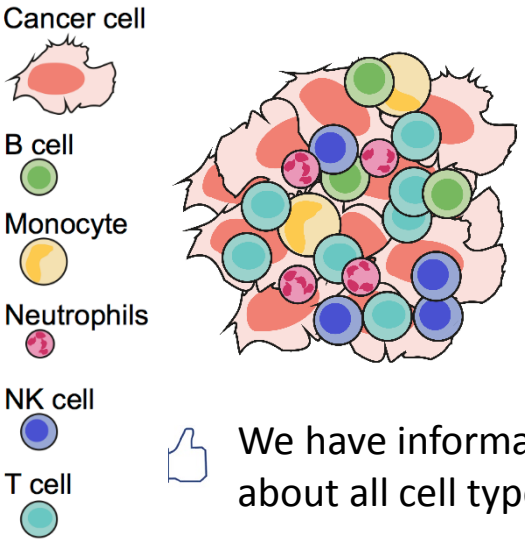
Gene fusion => More difficult (esp. for new fusion events)

Splicing => More difficult (esp. for poorly annotated isoforms)

Measured expression is the average over all cell types

Single-cell RNA-Seq: motivation

Bulk sample gene expression analysis

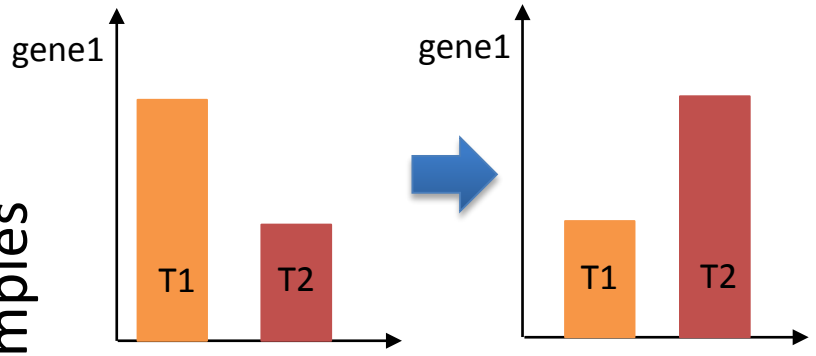


We have information about all cell types

The signal is mixed with other cell types

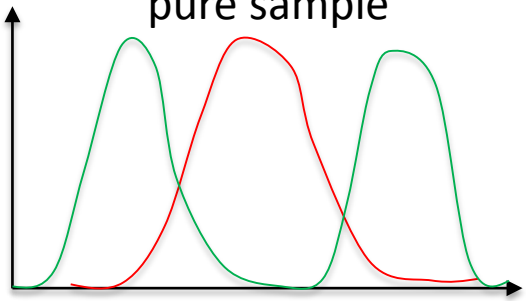
Examples

Different expression in different cell types in mixed sample



Same average

Bimodal expression in a pure sample

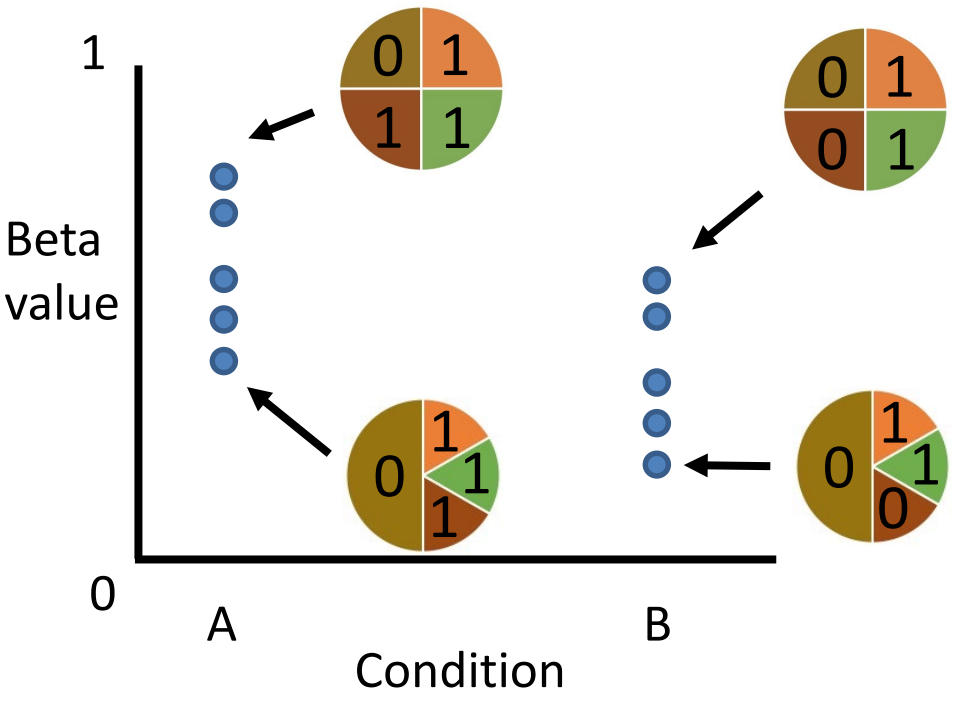


Same average

Possible in some cases to infer the fraction of the different cell types, but (almost) impossible to infer their actual gene expression profile

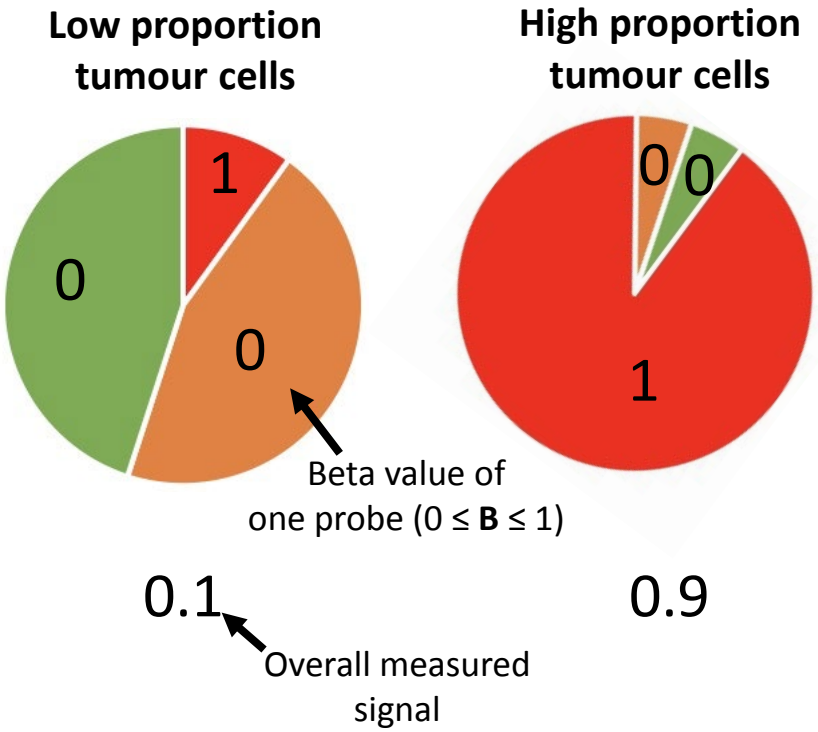
Influence of heterogeneity

Heterogeneous composition



Signal dilution → false negatives

Associated with phenotype



Signal inflation → false positives

2009: n=1

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang[1,3], Catalin Barbacioru[2,3], Yangzhou Wang[2], Ellen Nordman[2], Clarence Lee[2], Nanlan Xu[2], Xiaohui Wang[2], John Bodeau[2], Brian B Tuch[2], Asim Siddiqui[2], Kaiqin Lao[2] & M Azim Surani[1]

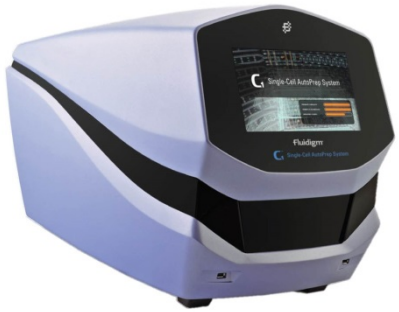
Next-generation sequencing technology is a powerful tool for transcriptome analysis. However, under certain conditions, only a small amount of material is available, which requires more sensitive techniques that can preferably be used at the single-cell level. Here we describe a single-cell digital gene expression profiling assay. Using our mRNA-Seq assay with only a single mouse blastomere, we detected the expression of 75% (5,270) more genes than microarray techniques and identified 1,753 previously unknown splice junctions called by at least 5 reads. Moreover, 8–19% of the genes with multiple known transcript isoforms expressed at least two isoforms in the same blastomere or oocyte, which unambiguously demonstrated the complexity of the transcript variants at whole-genome scale in individual cells. Finally, for *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ (*Eif2c2*$^{-/-}$) oocytes, we found that 1,696 and 1,553 genes, respectively, were abnormally upregulated compared to wild-type controls, with 619 genes in common.

function[14,15]. Therefore, a more sensitive mRNA-Seq assay, ideally an assay capable of working at single cell resolution, is needed to meaningfully study crucial developmental processes and stem cell biology.

Here we modified a widely used single-cell whole-transcriptome amplification method to generate cDNAs as long as 3 kilobases (kb) efficiently and without bias[16,17]. With Applied Biosystems' next-generation sequencing SOLiD system, we found that it is feasible to get digital gene expression profiles at single-cell resolution. Using our mRNA-Seq assay with only a single mouse blastomere, we detected expression of 5,270 more genes than microarrays using hundreds of blastomeres. Using only a single blastomere, we also identified 1,753 previously unknown splice junctions, which have never been detected by microarrays at single-cell resolution. We found that hundreds of genes expressed two or more transcript variants in the same cell. We also found that in *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ mature oocytes, 1,696 and 1,553 genes, respectively, were abnormally upregulated, and 1,571 and 1,121 genes, respectively,

Tang et al., Nature Methods (2009)

Technological breakthroughs



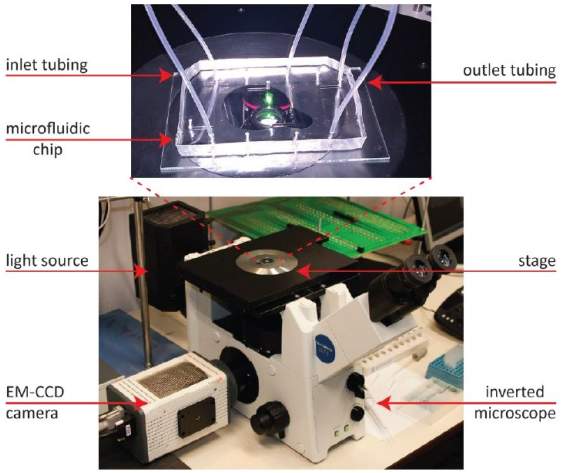
FLUIDIGM

10X GENOMICS®

WAFERGEN BIOSYSTEMS

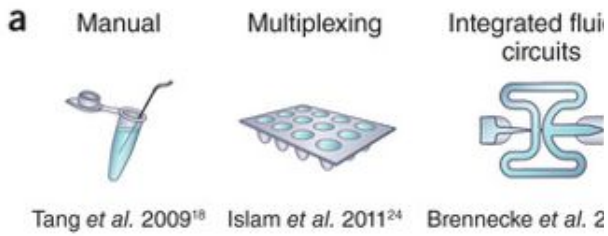
BIO-RAD

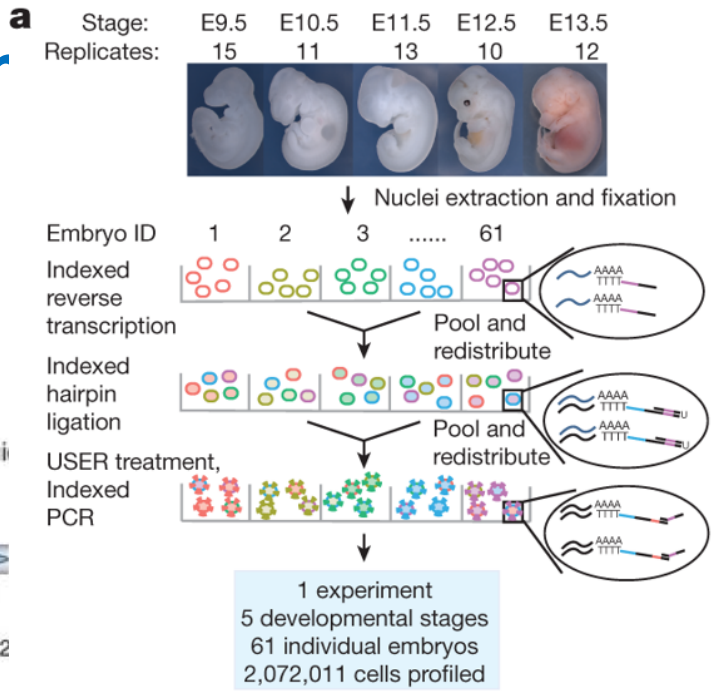
dolomite bio

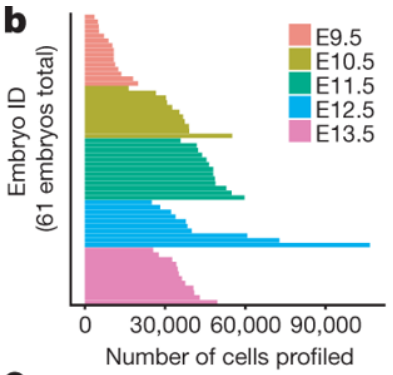


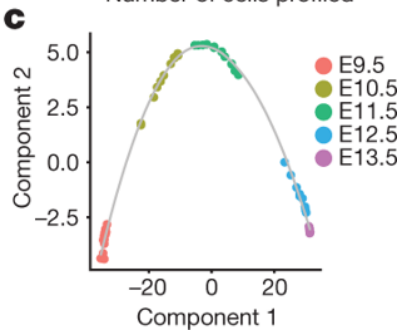
DIY

Exponential scalir the past decade

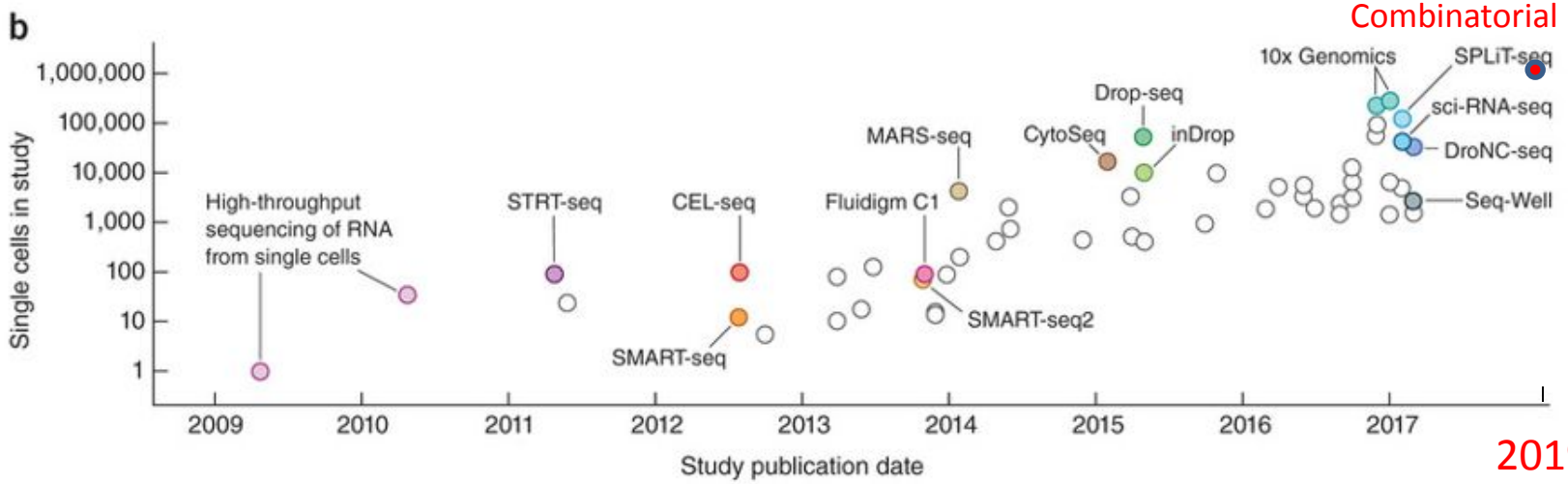








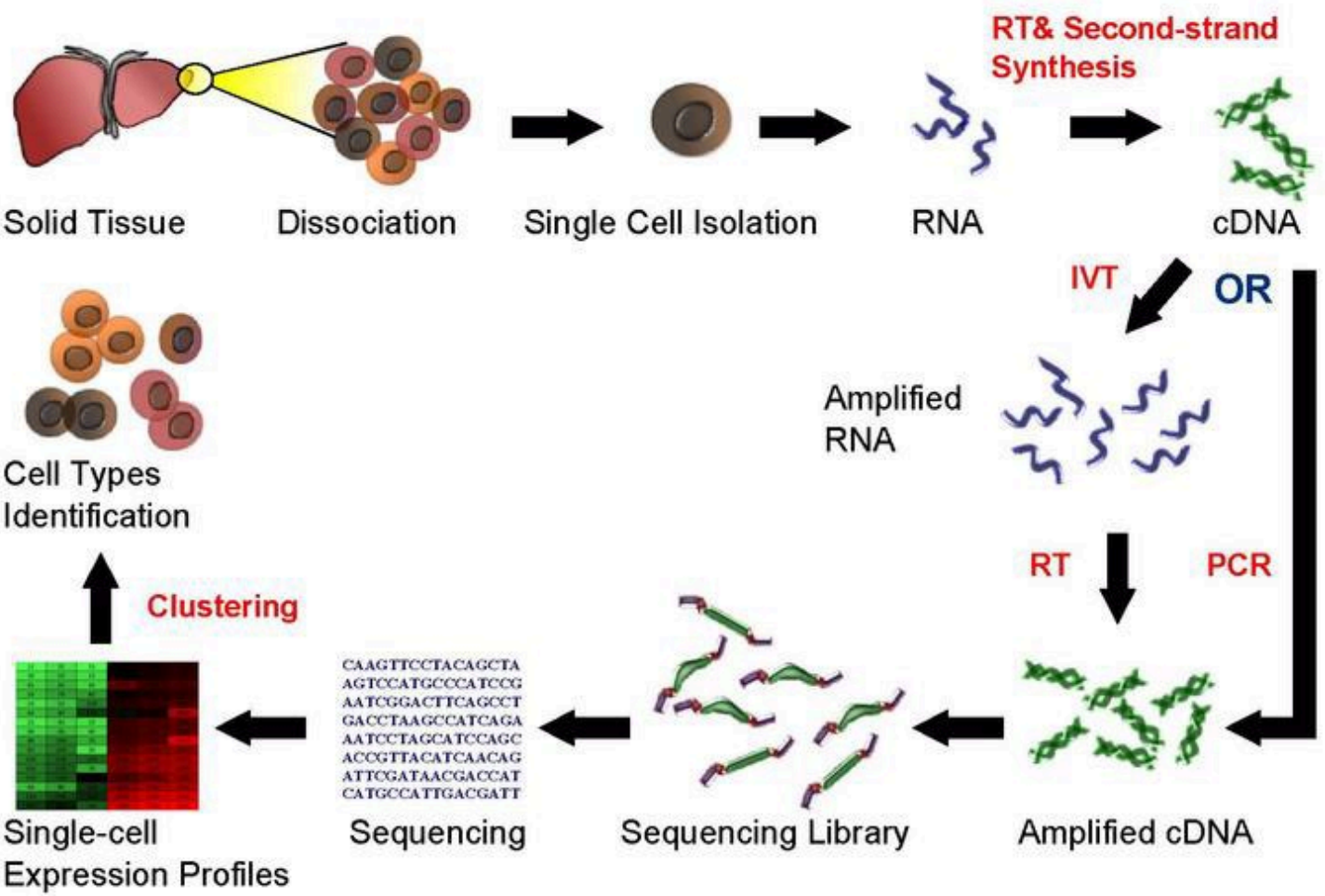
Combinatorial indexing



2019
Cao et al., Nature

Svensson et al., Nature Methods (2018)

scRNA-Seq: workflow



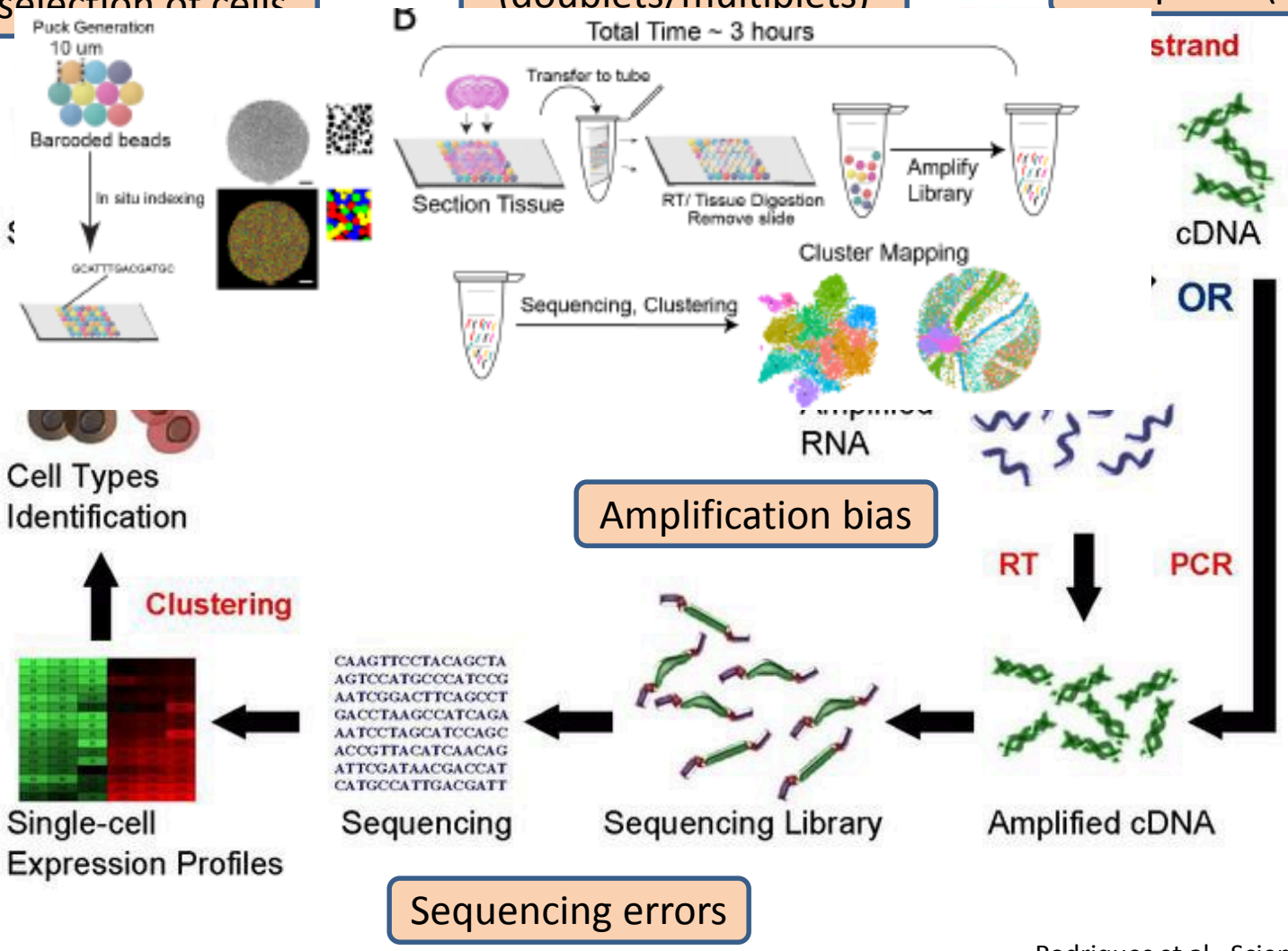
https://en.wikipedia.org/wiki/Single_cell_sequencing

scRNA-Seq: what can go wrong?

Biased selection of cells

Duplicates (doublets/multiplets)

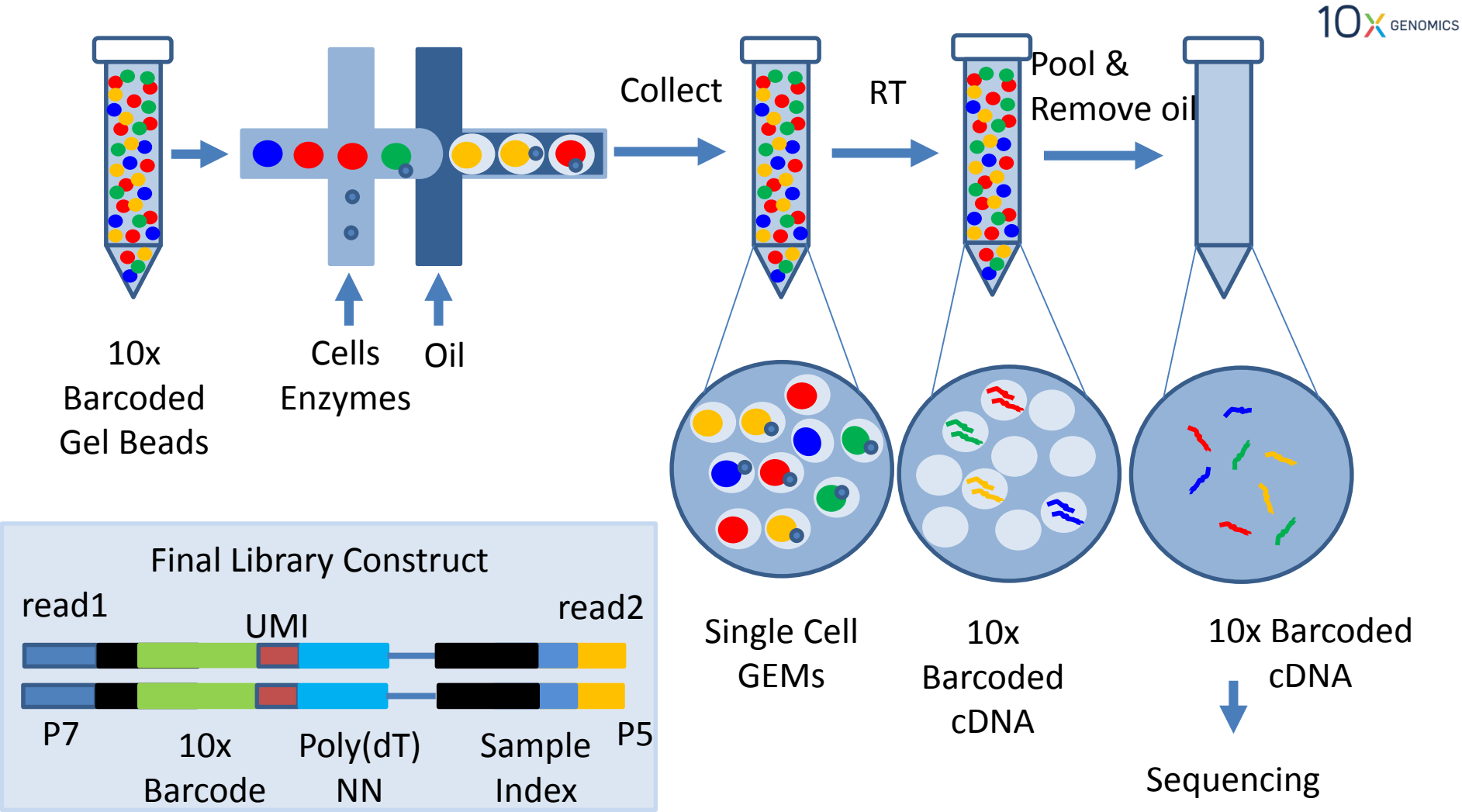
Dropouts (60-90%)?



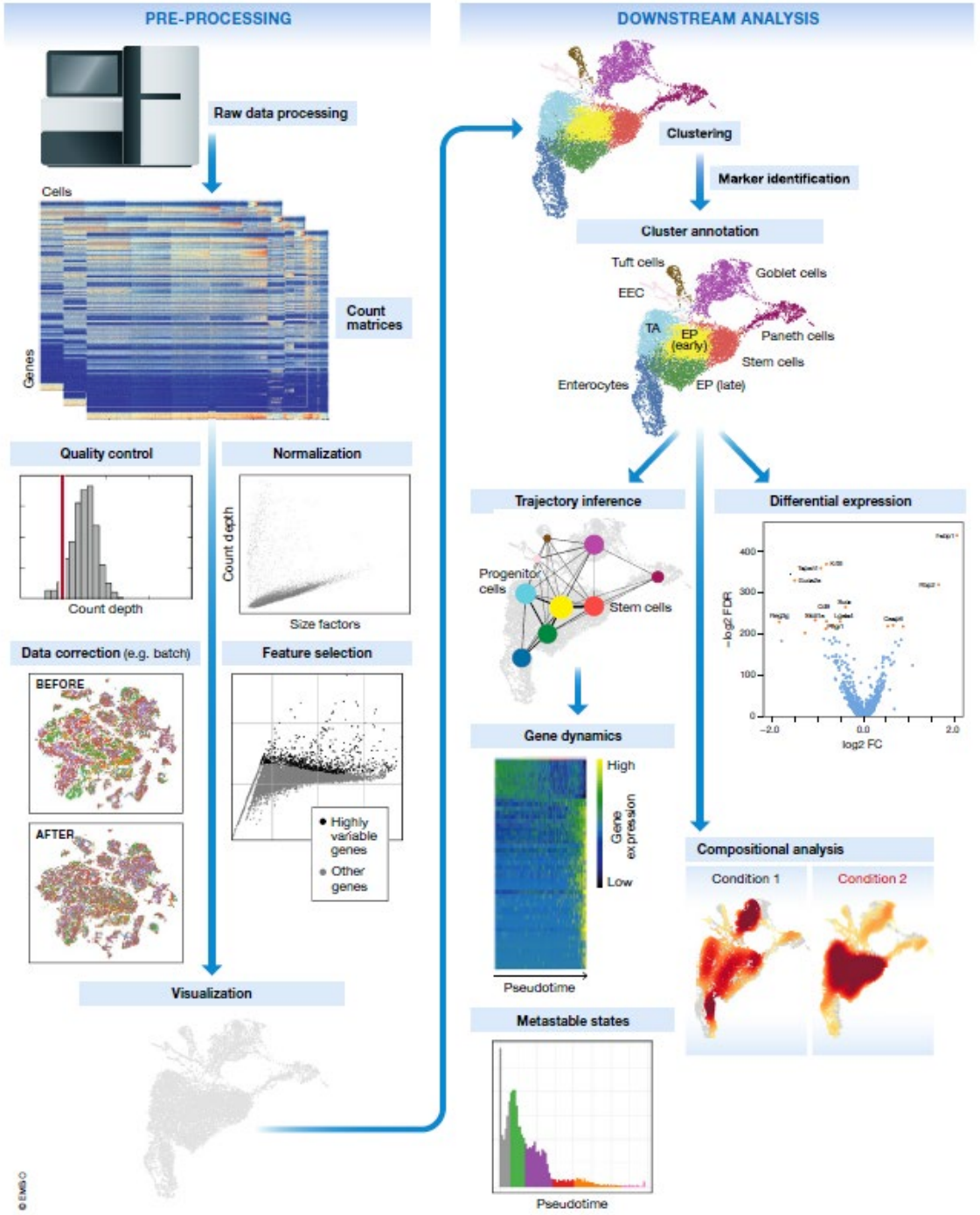
Sequencing errors

Rodriques et al., Science (2019): Slide-seq

https://en.wikipedia.org/wiki/Single_cell_sequencing

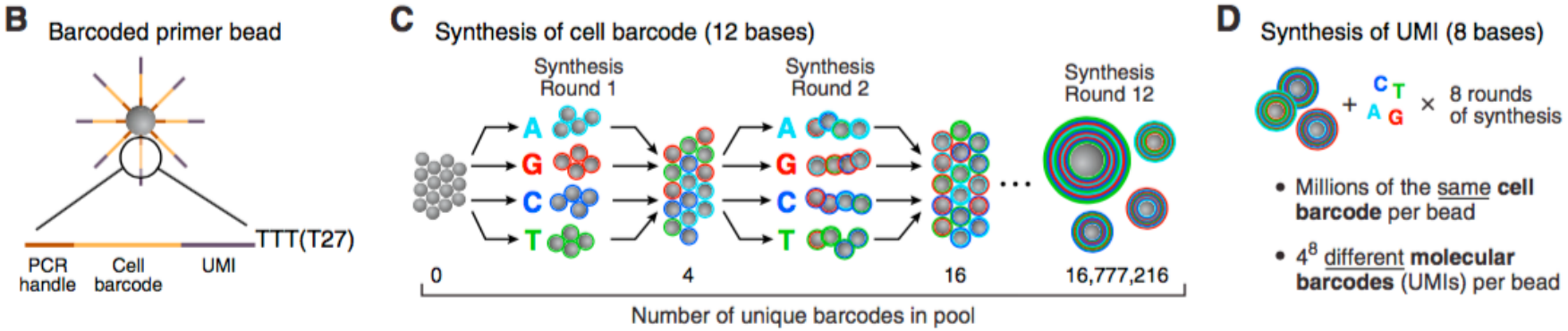

10x GENOMICS
10x Barcoded Gel Beads
Cells Enzymes
Oil
Collect
RT
Pool & Remove oil
Single Cell GEMs
10x Barcoded cDNA
10x Barcoded cDNA
Sequencing
Final Library Construct
read1
UMI
read2
P7
10x Barcode
Poly(dT) NN
Sample Index
P5

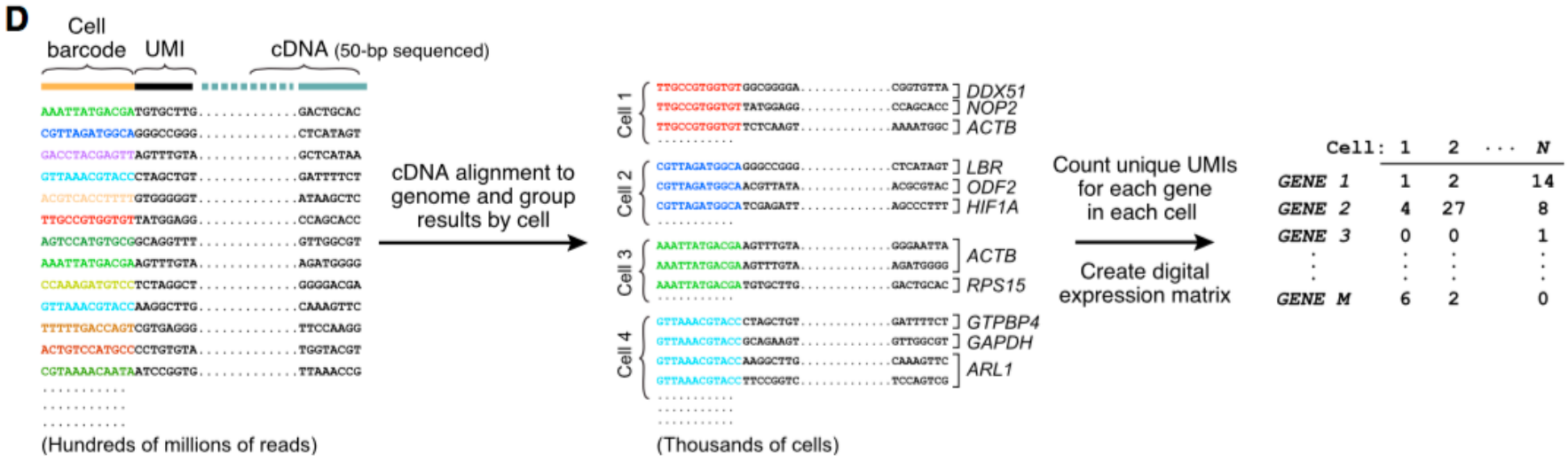
Typical single-cell RNA-seq analysis workflow



Luecken et al., Molecular Systems Biology (2019)

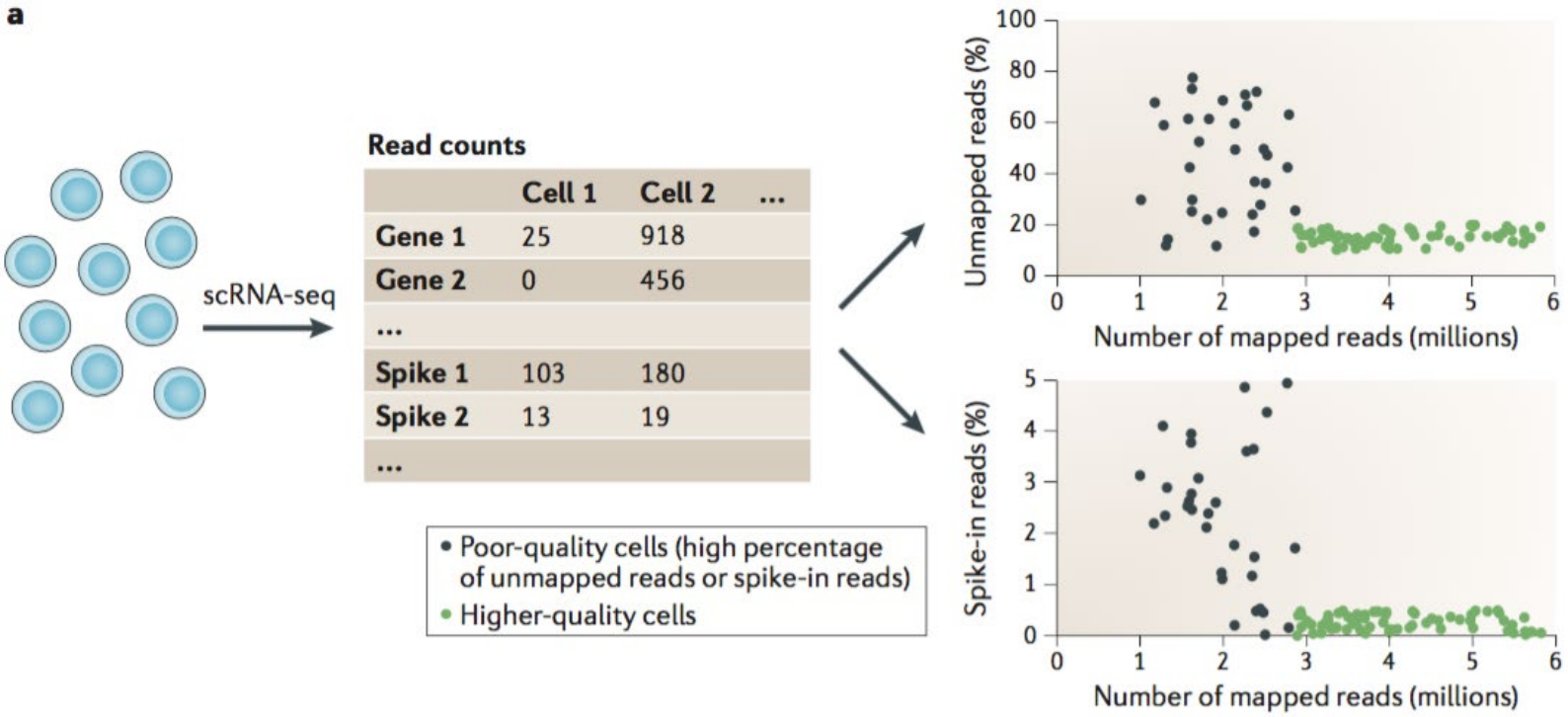
Amplification bias: unique molecular identifiers (I)





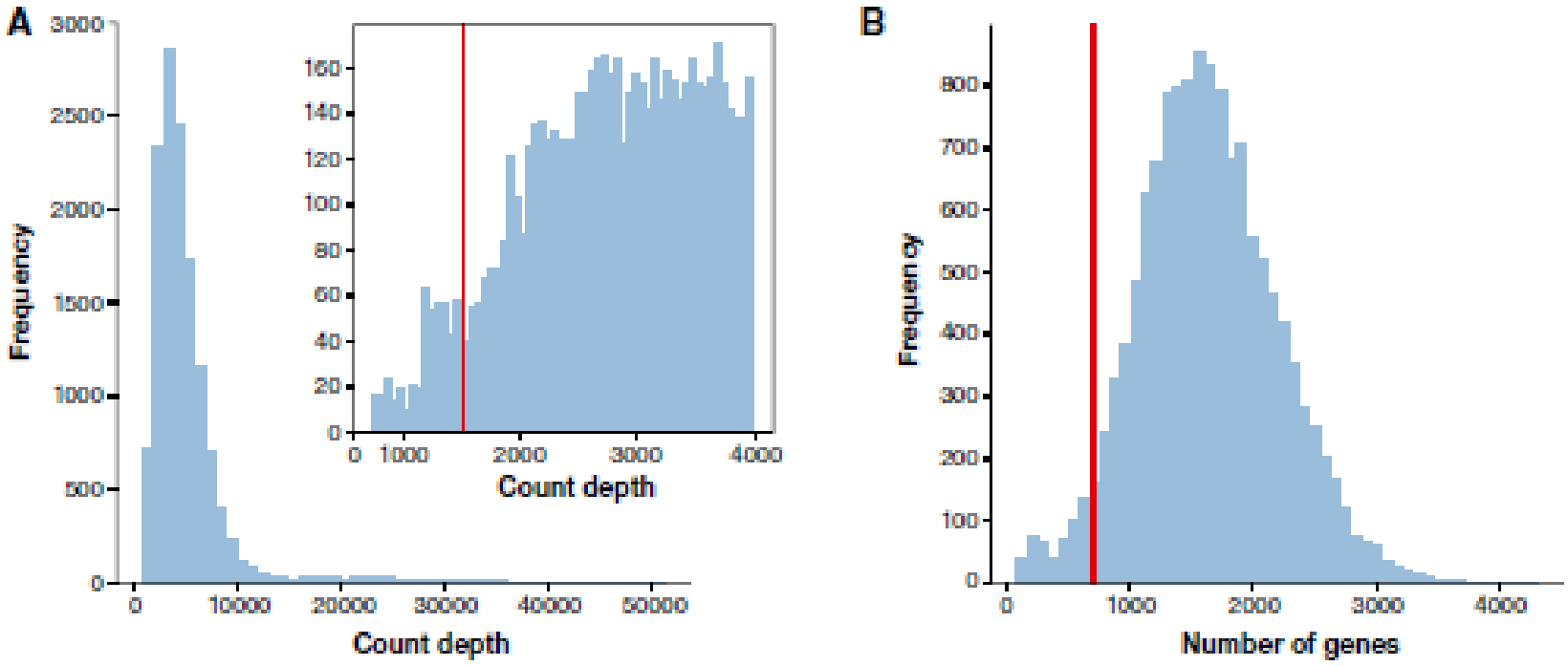
http://mccarrolllab.com/dropseq/

QC at cell level



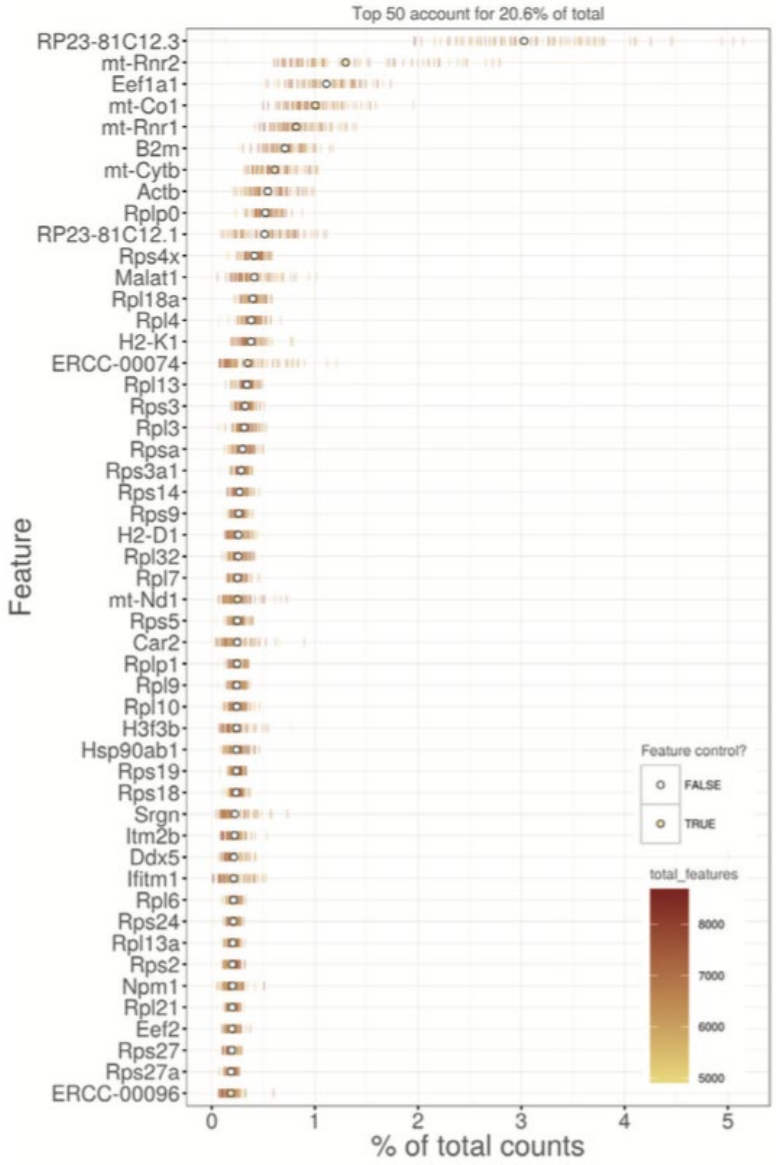
Stegle et al., Nature Reviews Genetics (2015)

QC at cell level (II)



Luecken et al., Molecular Systems Biology (2019)

QC at gene level



Percentage of total counts assigned to the top 50 most highly-abundant features

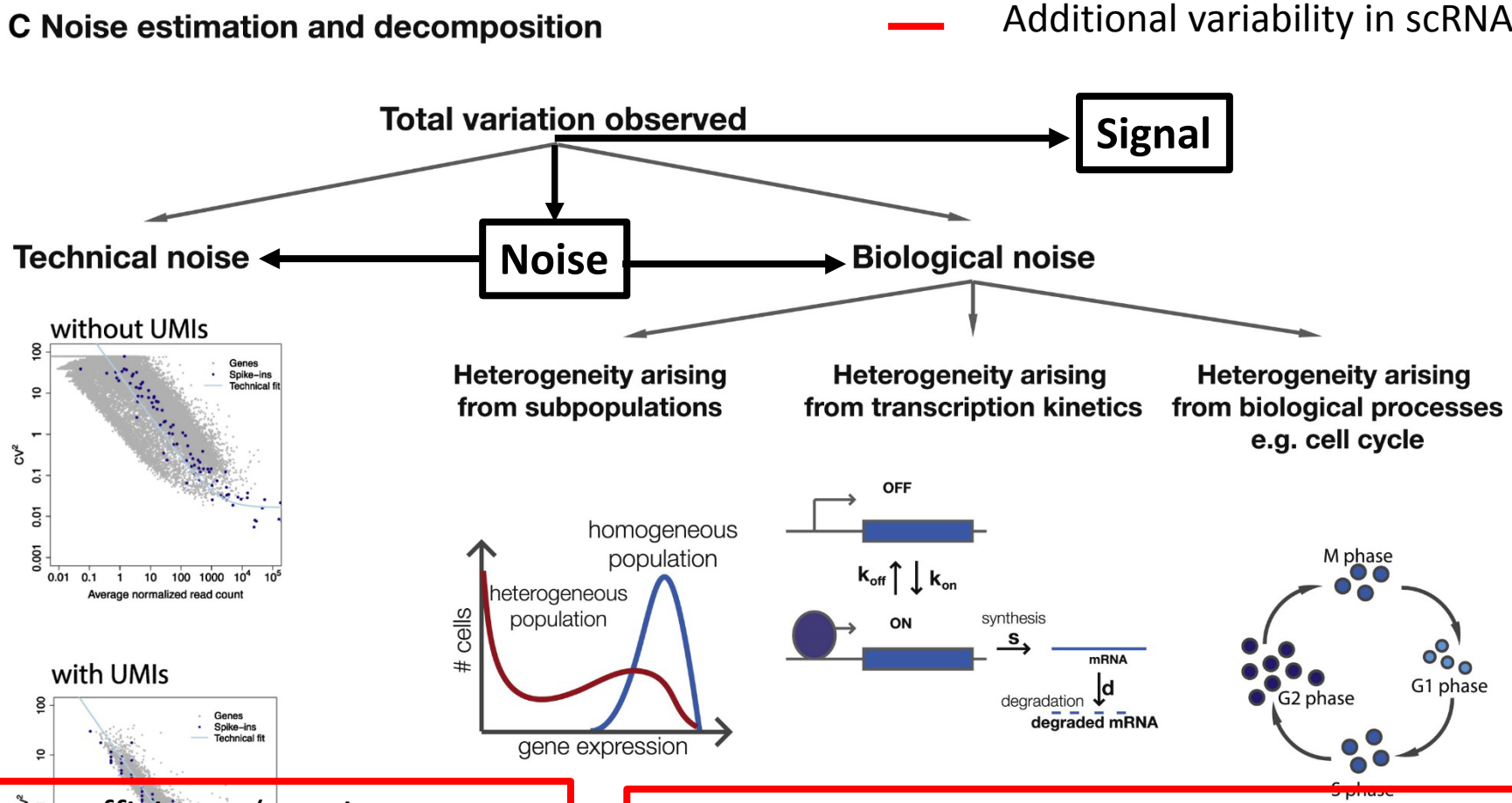
→ Does it match expected biology?

Lun et al., F1000Research (2016)

Variability in scRNA-Seq data (I)

Variability in bulk RNA-Seq

Additional variability in scRNA-seq



capture efficiency (starting amount of RNA)

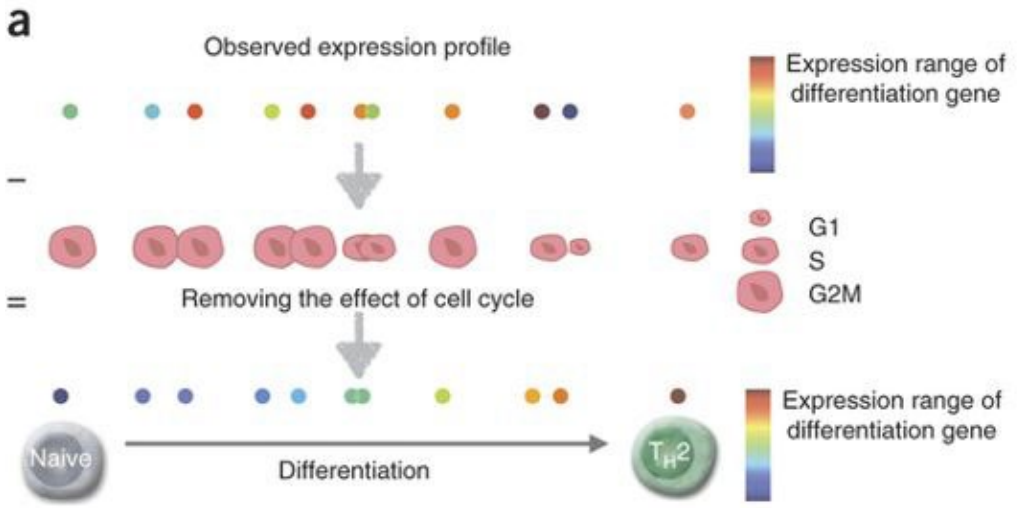
Extrinsic noise (regulation by transcription factors) vs **intrinsIc** noise (stochastic bursting/firing, cell cycle)

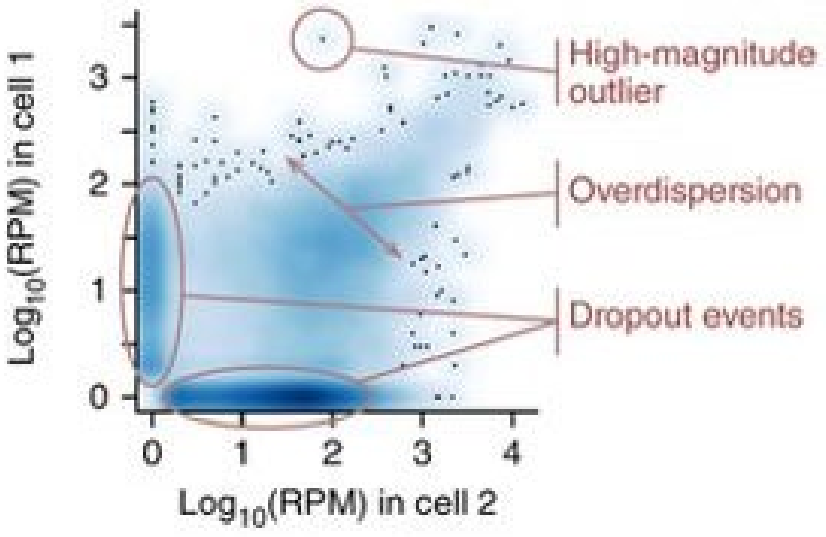
Overdispersion, batch effects, sequencing depth, GC bias, amplification bias

Kolodziejczyk et al., Molecular Cell (2015)

Variability in scRNA-Seq data (II)

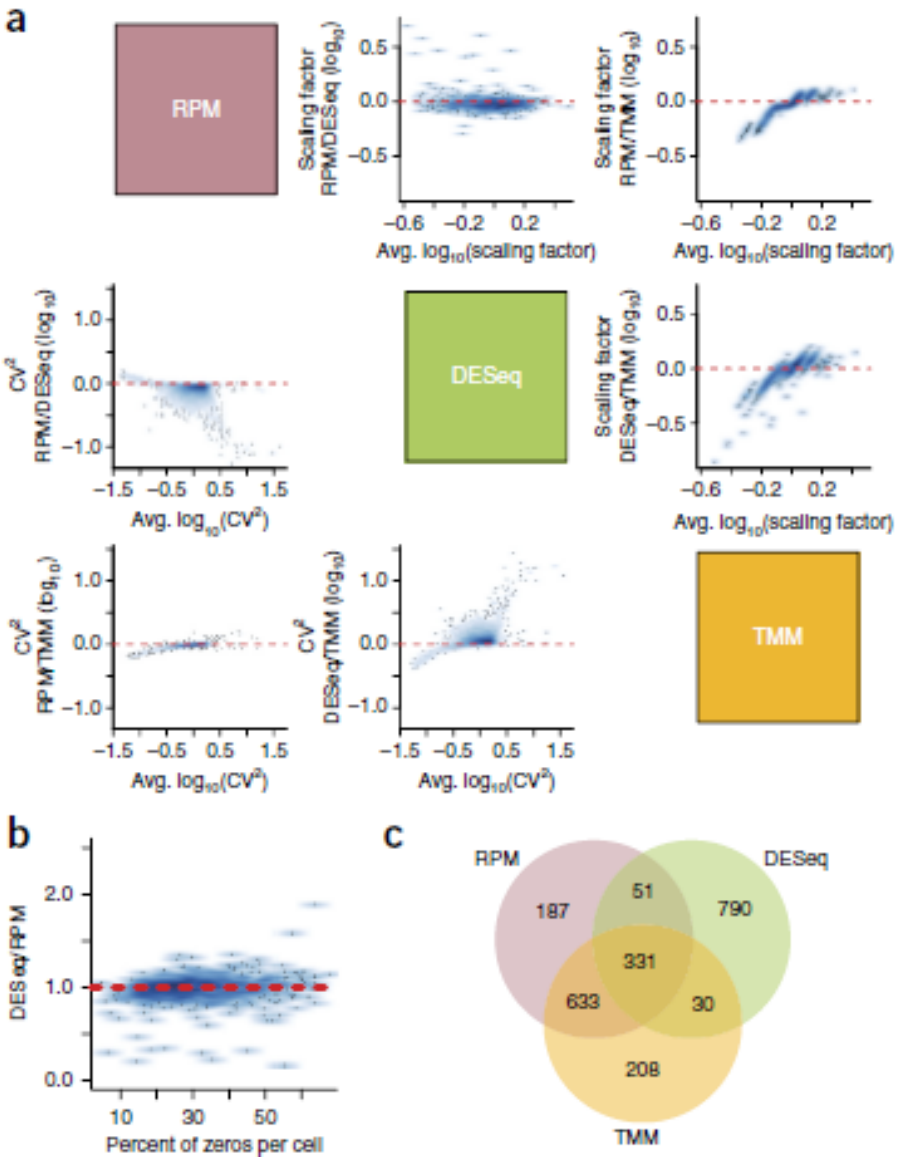
Cell cycle





Karchenko et al., Nature Methods (2014)

Normalization



Methods developed for bulk samples commonly used, but poor fit for scRNA-Seq data…

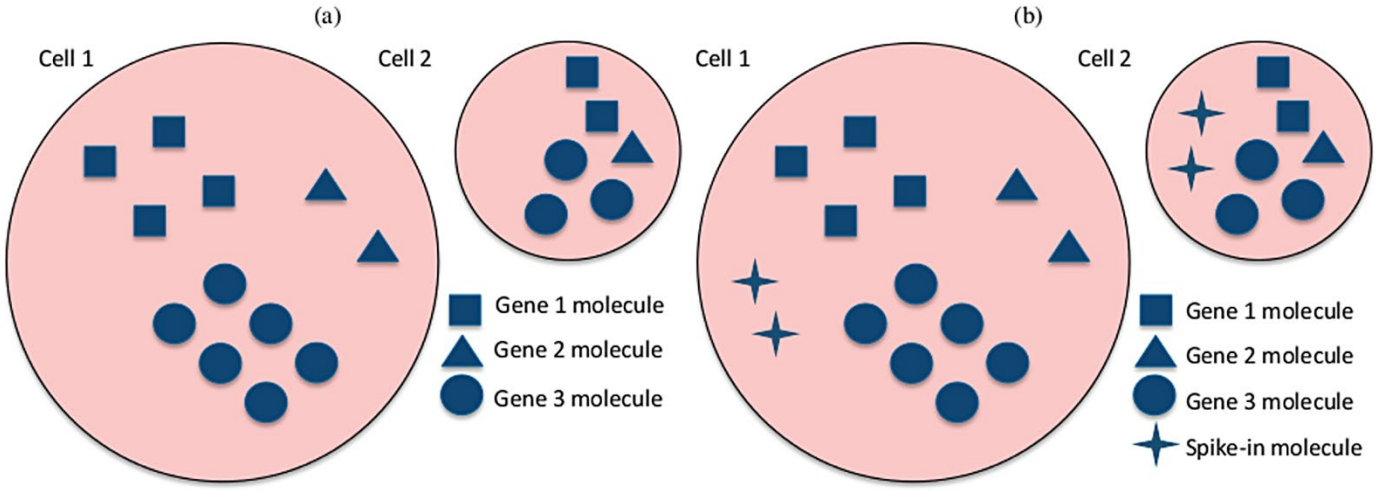
Spike-ins (I)

- Addition of external controls
- ERCC spike-ins most widely used, mix consists of 92 mRNAs at different concentrations
- Important to add equal amounts to each cell preferably in the lysis buffer

Spike-ins (II)

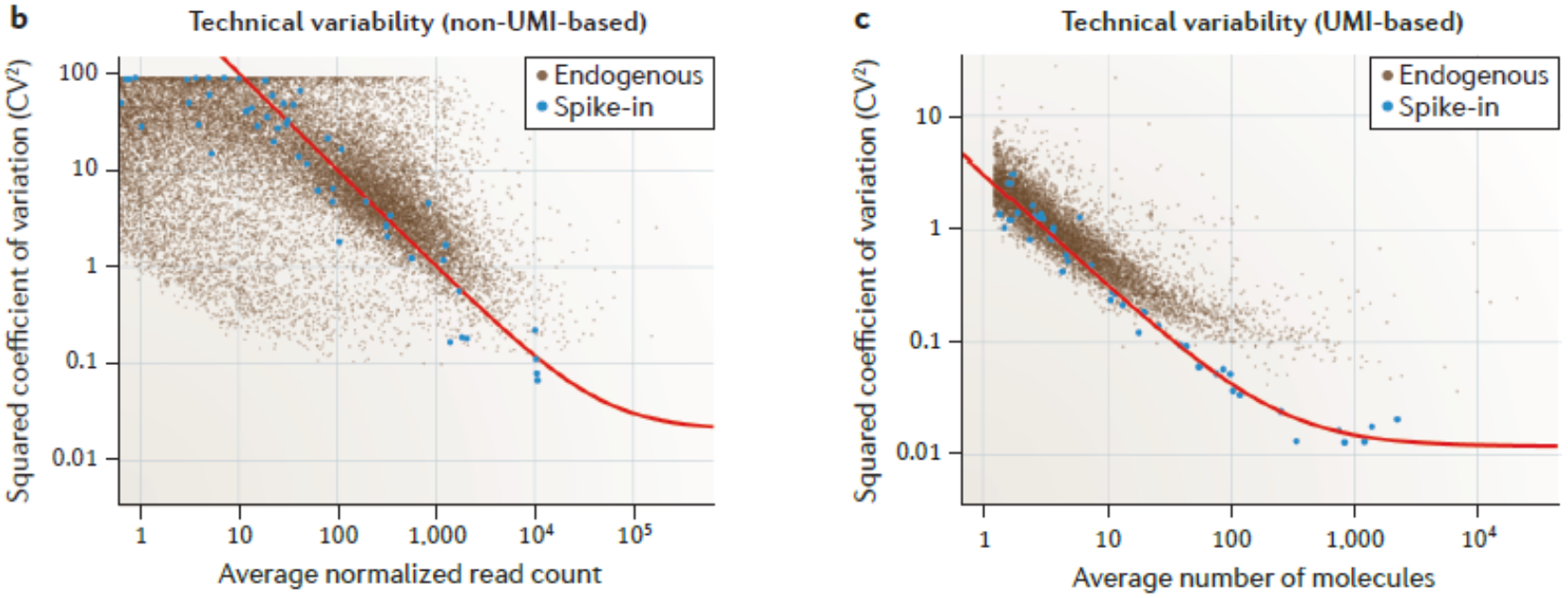
Can be used to model

- Technical noise
- Drop-out rates
- Starting amount of RNA in the cell
- Data normalization



Two cells from a homogeneous population but with different total mRNA content

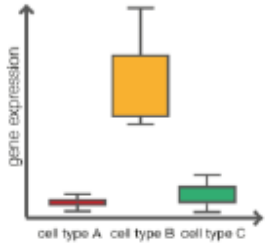
Amplication bias: unique molecular identifiers (II)



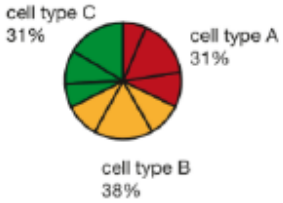
scRNA-Seq: common applications

Characterization of cell type populations

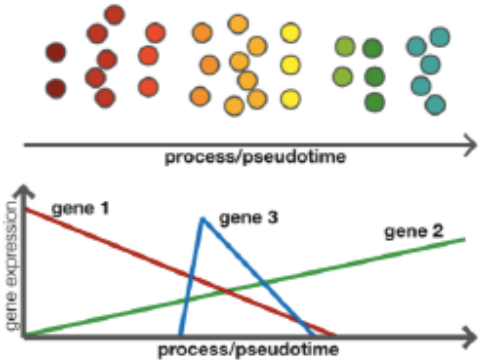
Finding markers of cell type



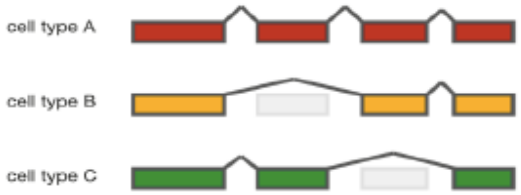
Frequency of cell type in the population



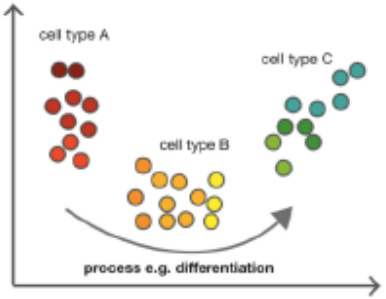
Identify genes that drive a process across time



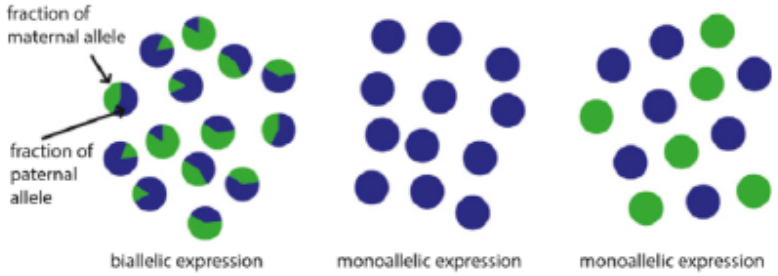
Differential splicing between populations

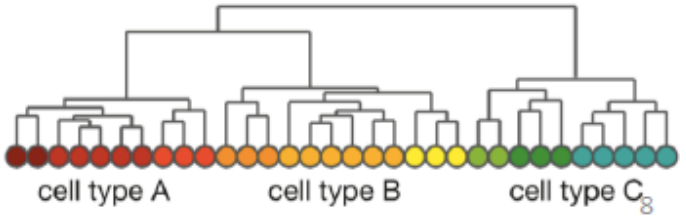


Identify cell type populations (e.g. dim reduction or clustering)



Identify allele-specific expression



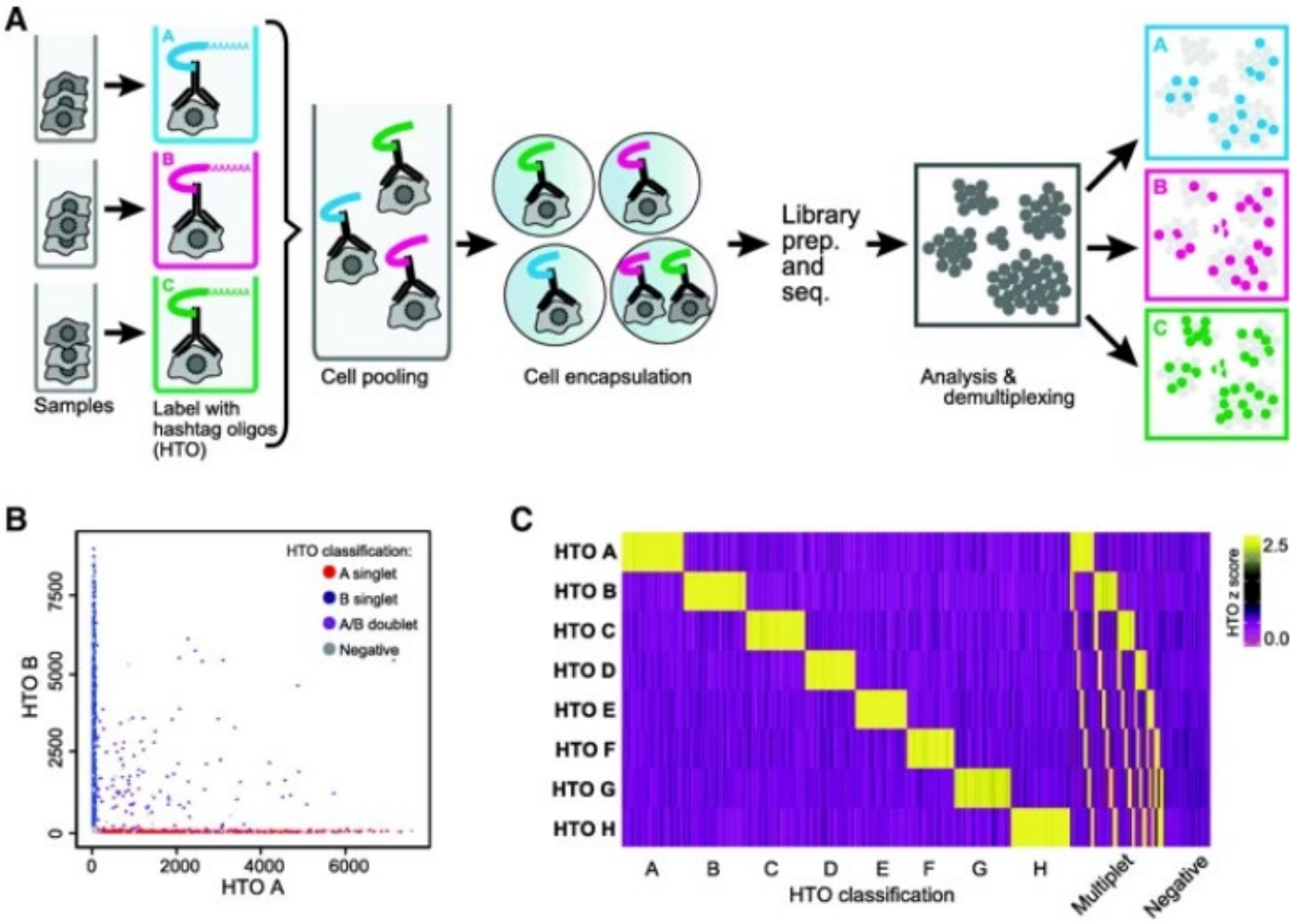


Adapted from Kolodziejczyk et al. (2015). *Molecular Cell* **58**

Capita selecta

- Trade-off between # cells vs. #reads
- # cells
- Batch effects, multiplexing
- Multi-omics

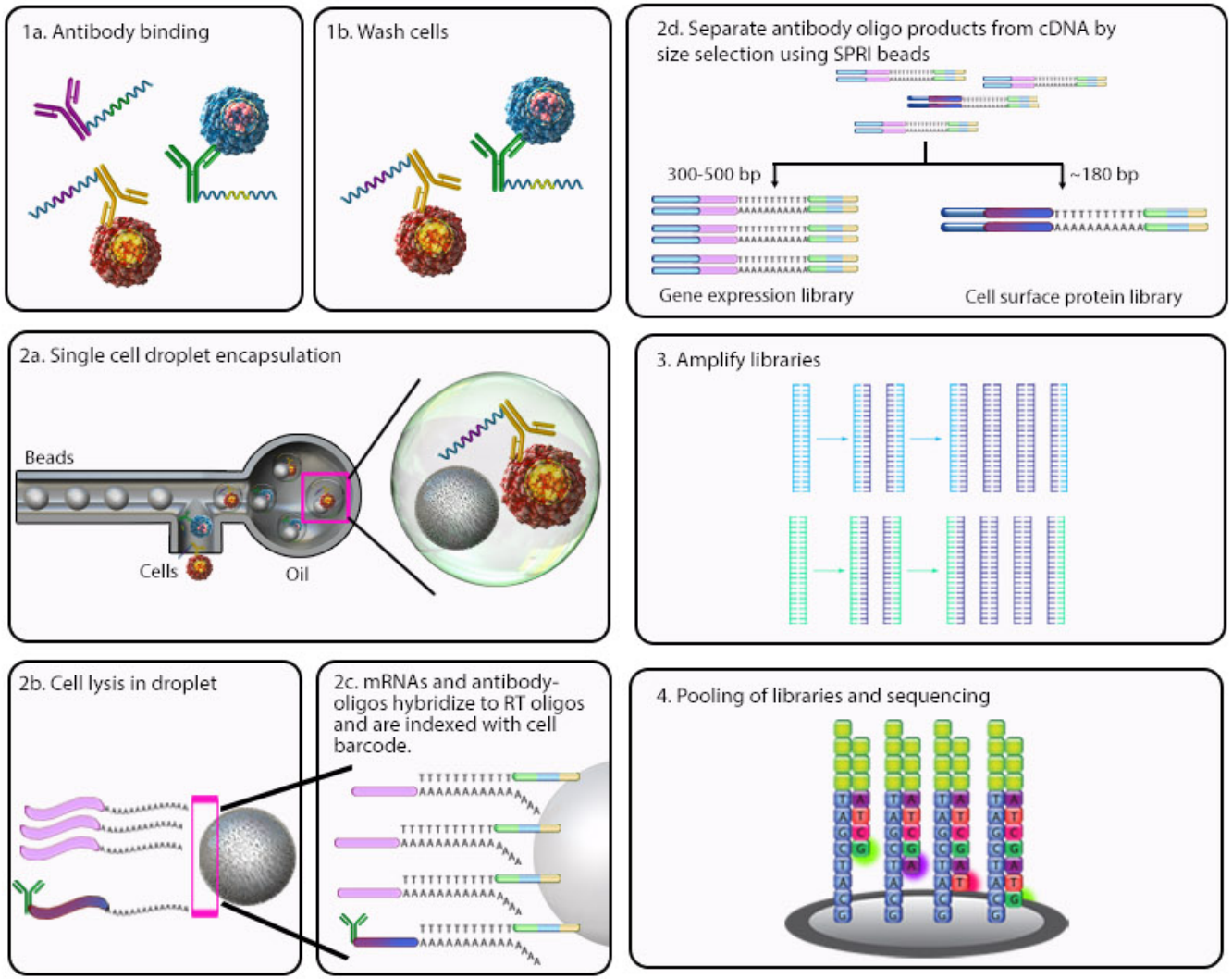
Fig. 1

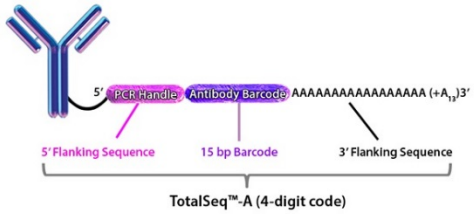


Hash
Tag
Oligonucleotide

Stoekius *et al.* Gen. Biol., 2018

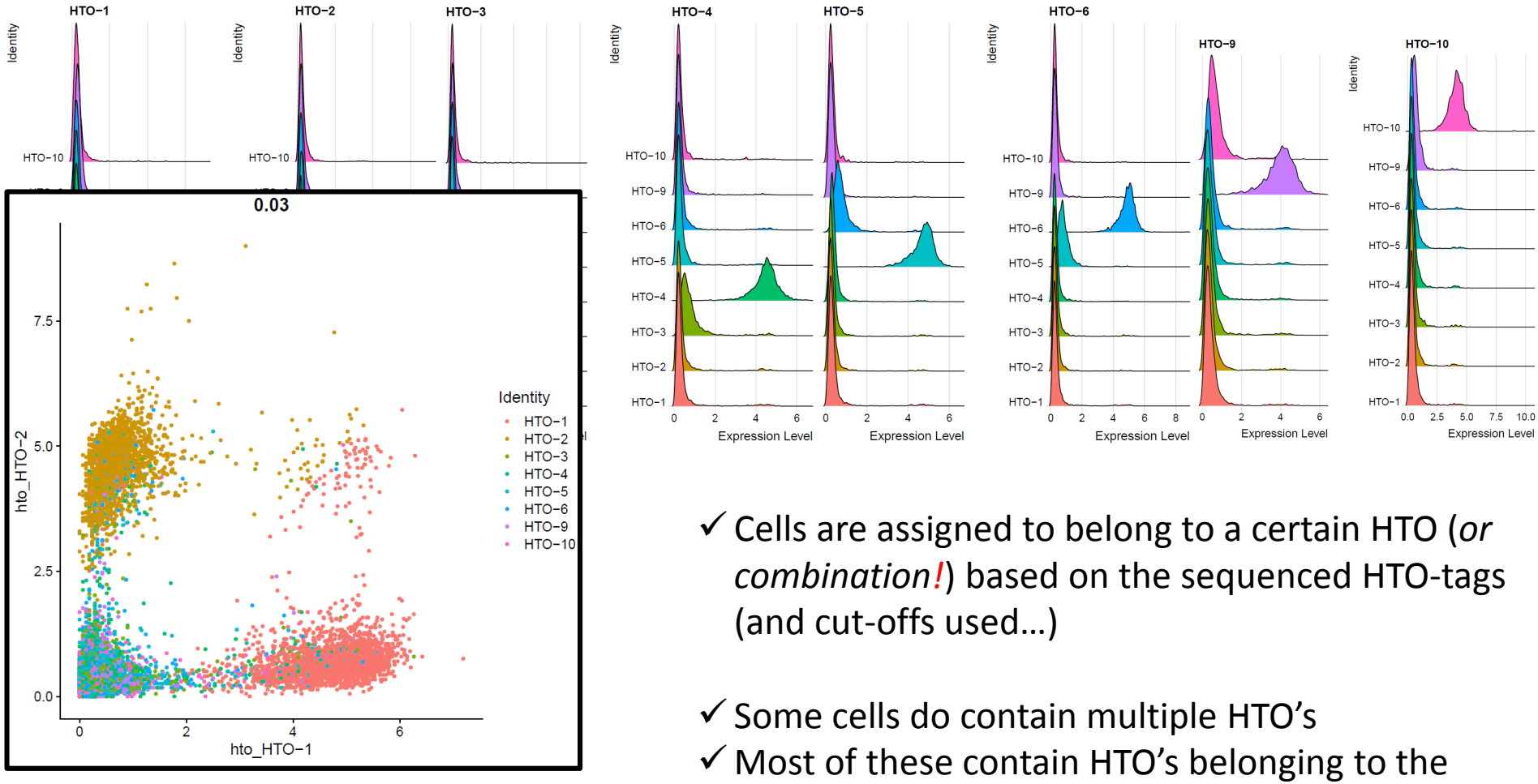
Antibody
Derived
Tags





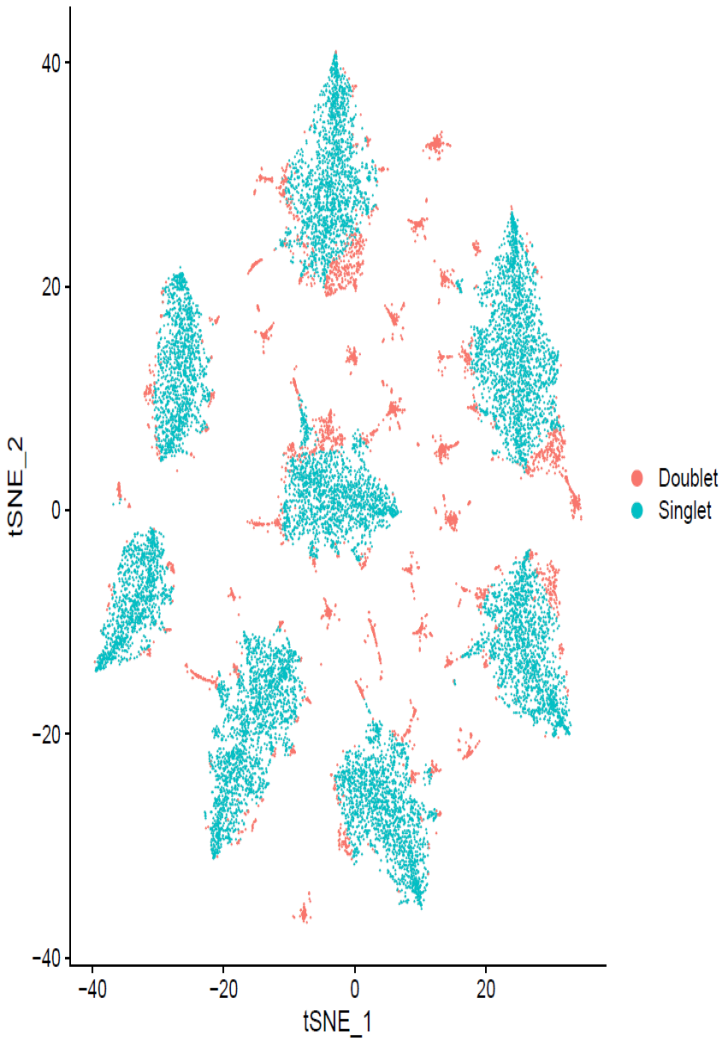
https://www.biolegend.com/en-us/totalseq

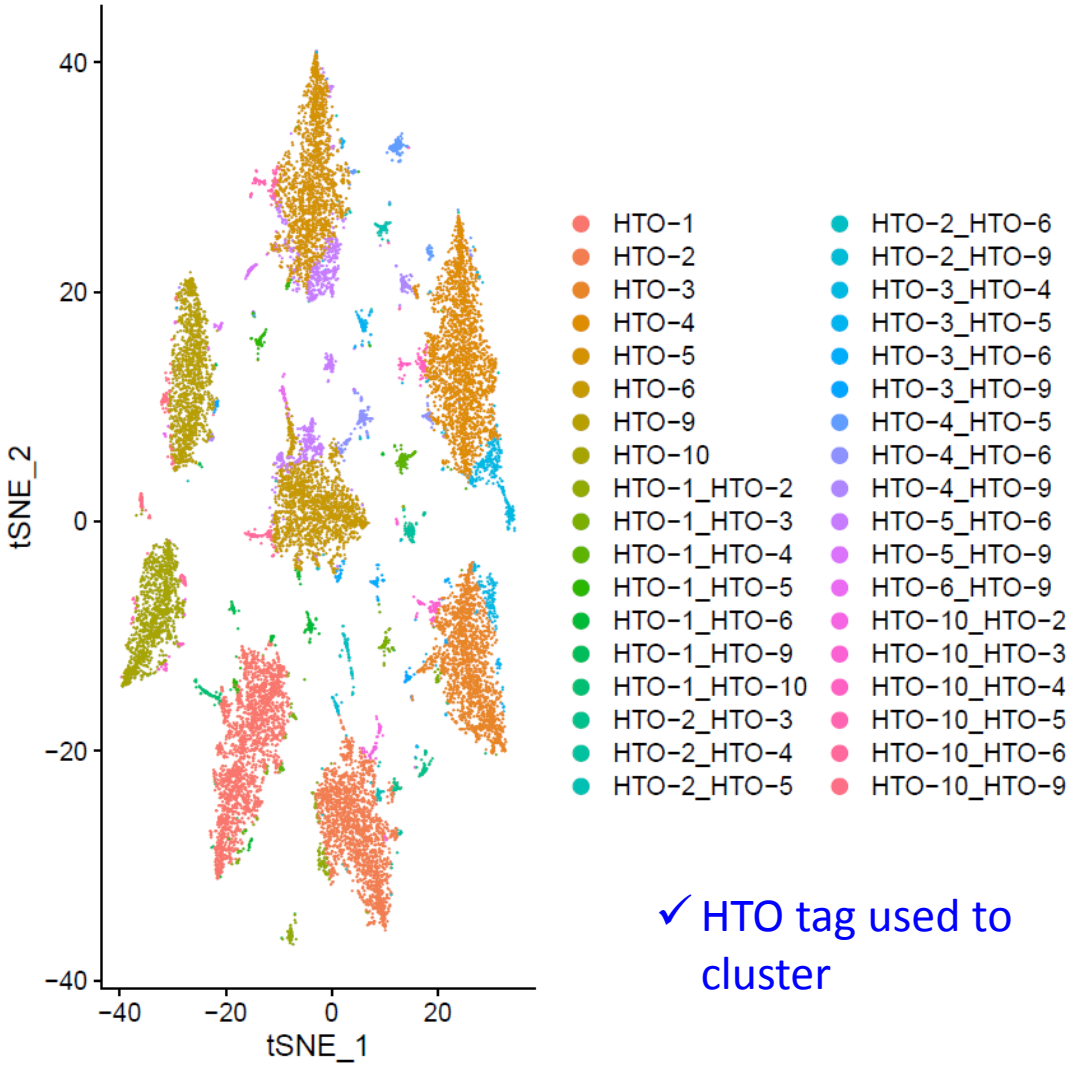
Demultiplexing HTO



- ✓ Cells are assigned to belong to a certain HTO (*or combination!*) based on the sequenced HTO-tags (and cut-offs used…)

- ✓ Some cells do contain multiple HTO's
- ✓ Most of these contain HTO's belonging to the same subject, i.e. HTO_9/HTO_10 etc. -> Doublets





✓ HTO tag used to cluster

Antibody Derived Tags (ADTs)