Differential expression analysis for transcriptomics data

Recent advances in a rapidly evolving field

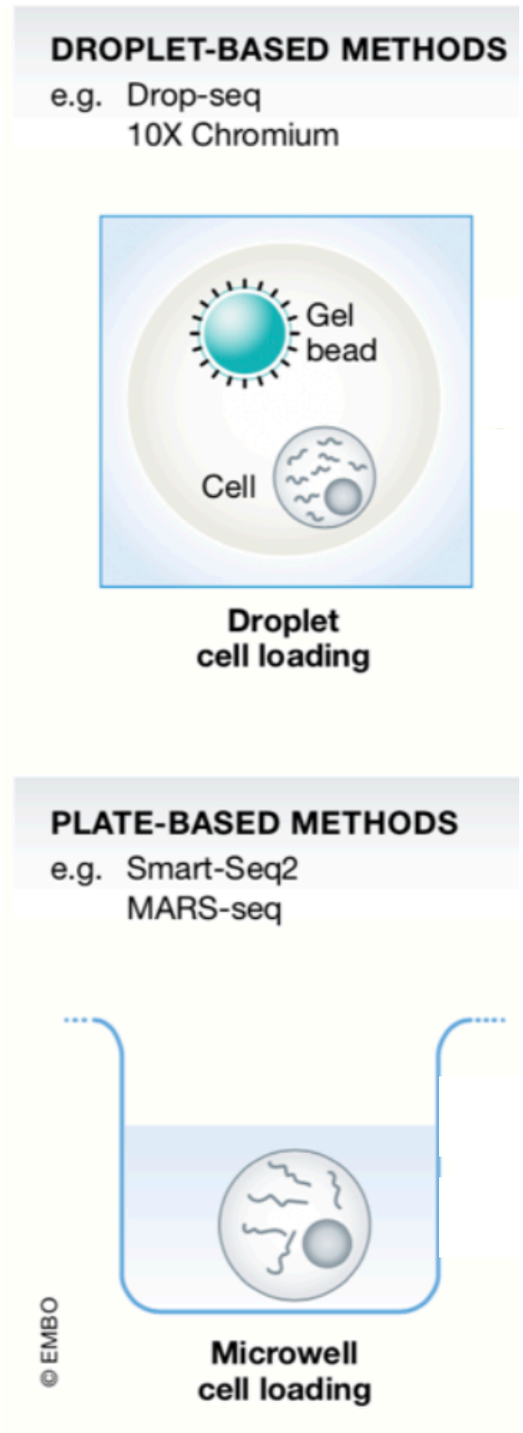
Outline

- **Single-cell transcriptomics:** recent advances in protocols and data
- **Muscat:** multi-patient multi-condition differential expression analyses
- **satuRn:** transcript-level inference for single-cell data

Outline

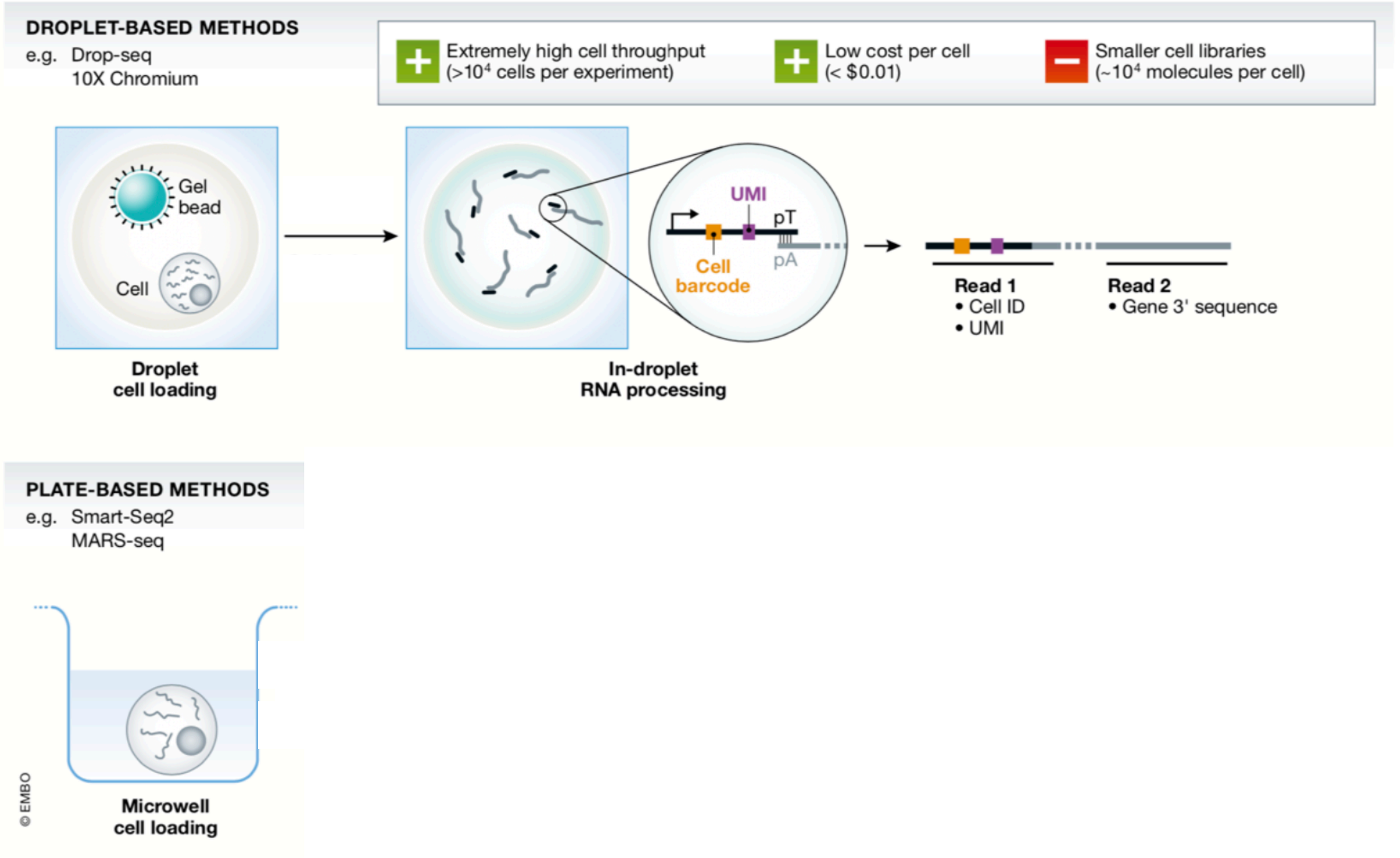
- **Single-cell transcriptomics:** recent advances in protocols and data
- **Muscat:** multi-patient multi-condition differential expression analyses
- **satuRn:** transcript-level inference for single-cell data

Single-cell transcriptomics protocols



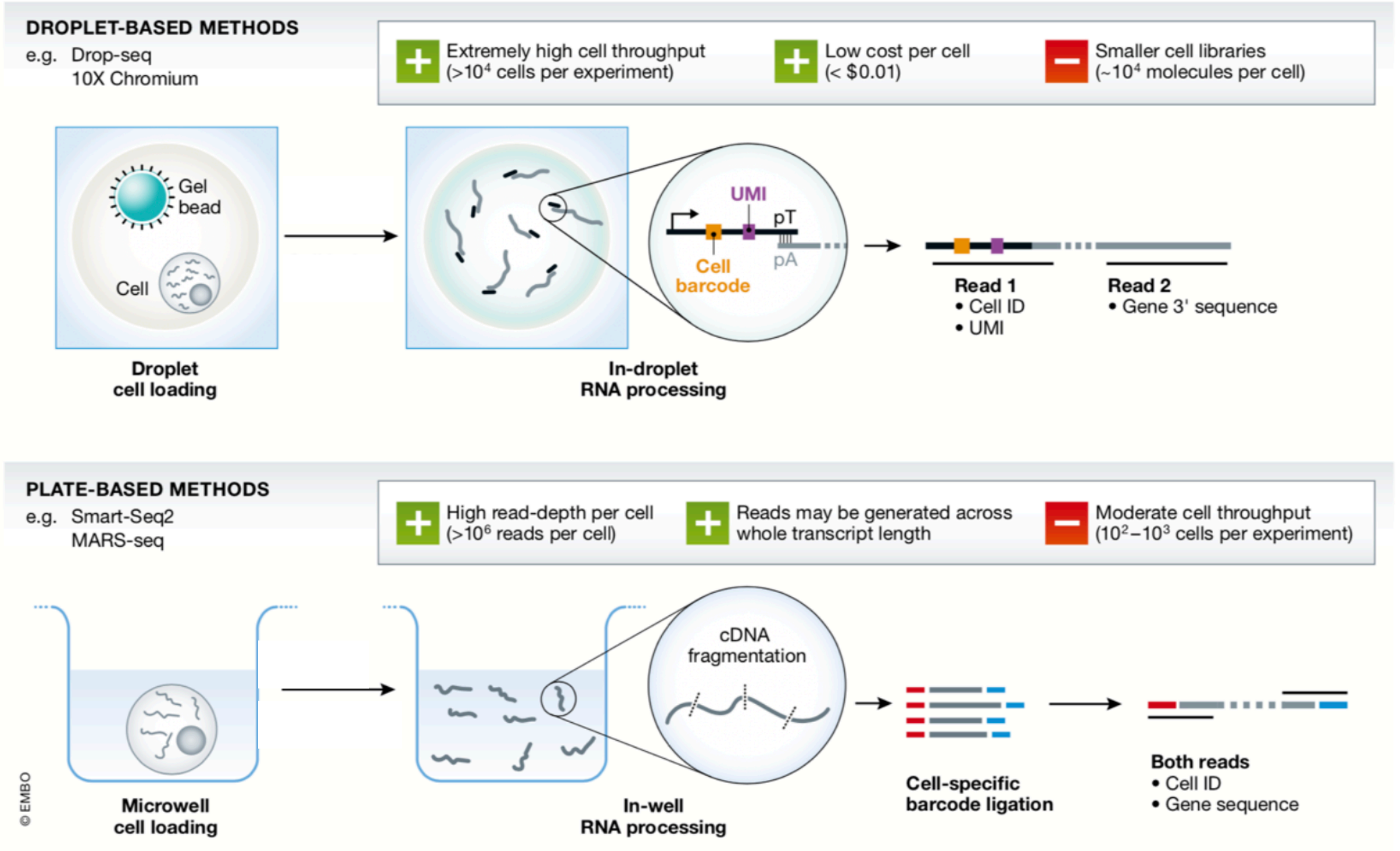
From Griffiths et al. (2018), doi: 10.15252/msb.20178046

Single-cell transcriptomics protocols



From Griffiths et al. (2018), doi: 10.15252/msb.20178046

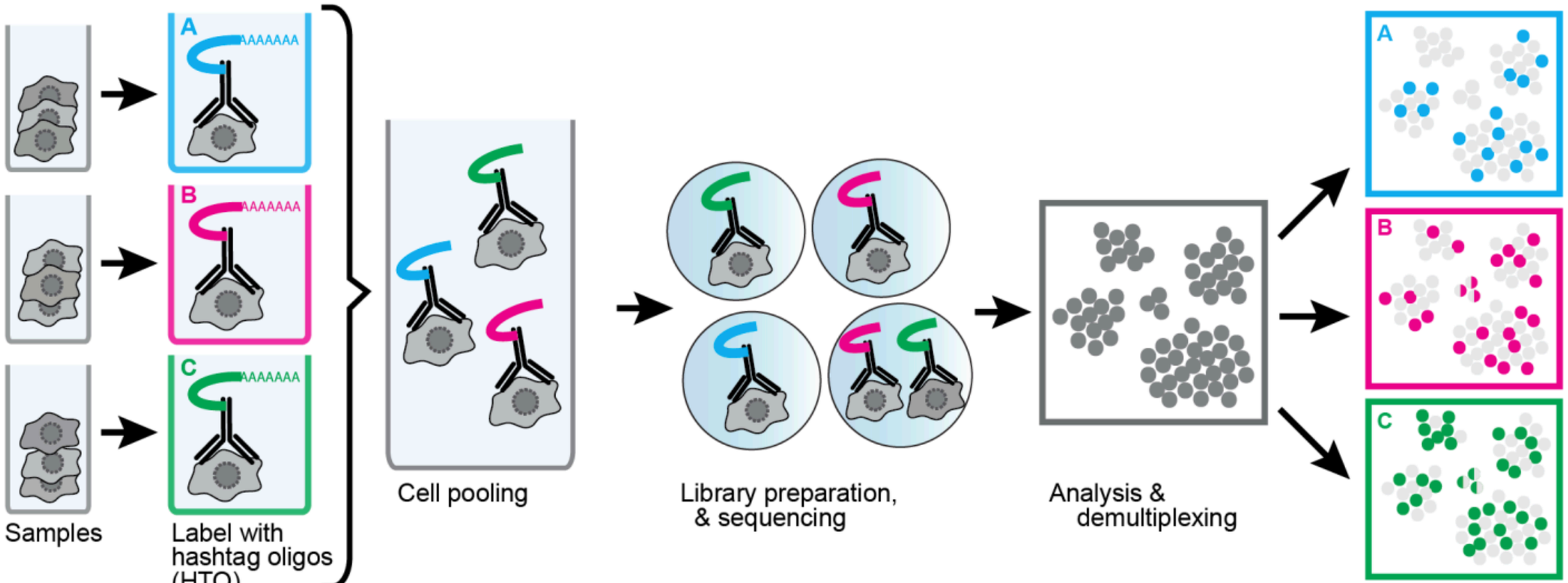
Single-cell transcriptomics protocols



From Griffiths et al. (2018), doi: 10.15252/msb.20178046

Single-cell transcriptomics - Advanced protocols

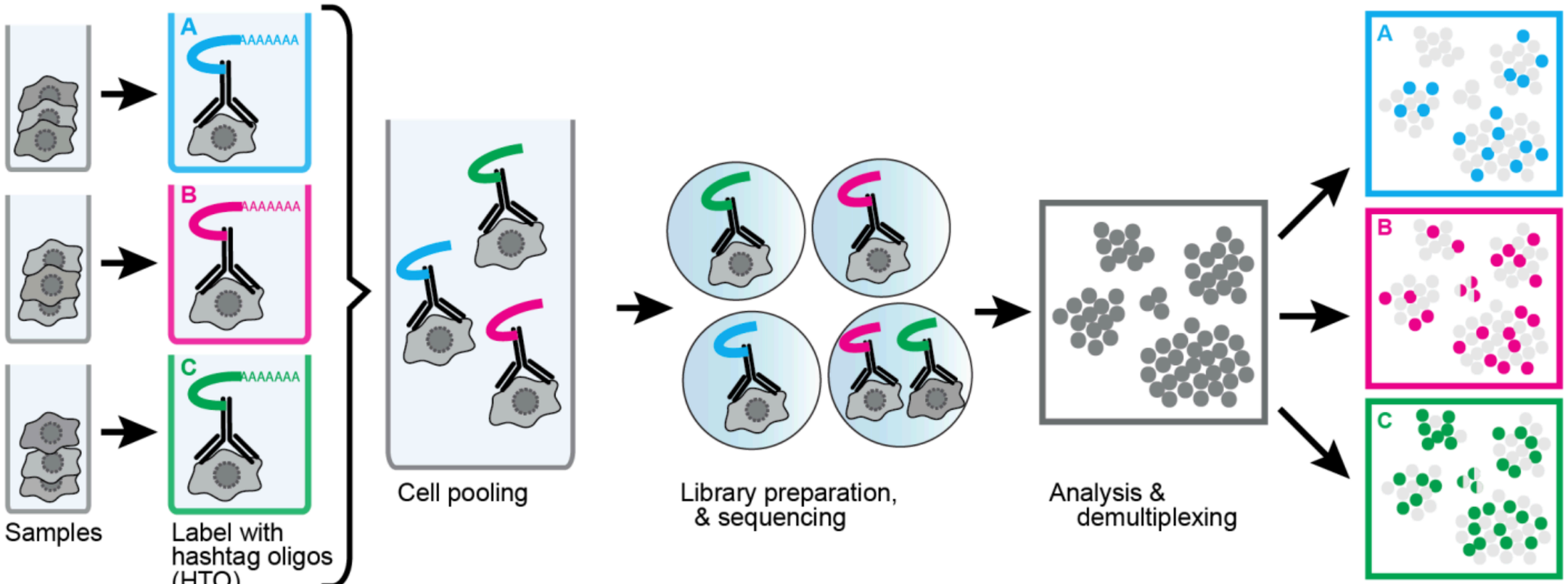
- Cell hashing - sample multiplexing



From: https://cite-seq.com/cell-hashing/

Single-cell transcriptomics - Advanced protocols

- Cell hashing - sample multiplexing



From: https://cite-seq.com/cell-hashing/

- Spatially resolved transcriptomics (Visium)
- CITE-seq
- ASAP-seq

Bulk versus single-cell data

Major differences:

1. Higher technical variation in single-cell data

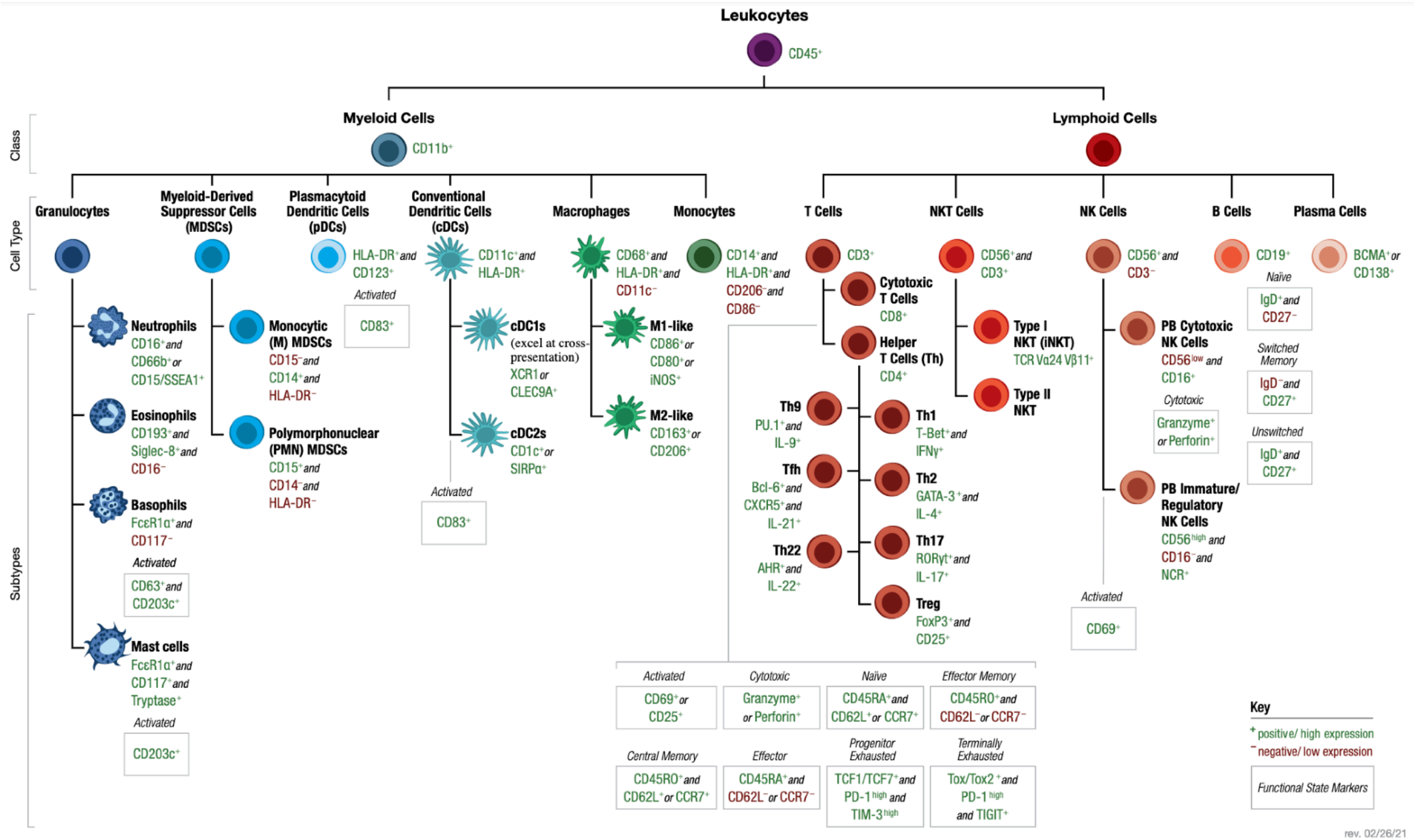
2. Higher biological variation in single-cell data

3. Single-cell data is very sparse

-> lower signal-to-noise ratio

Hierarchical data structure

- Single-cell data is hierarchical/clustered in nature
- Resolution of inference depends on research hypothesis and quality of data



From: https://www.cellsignal.com/pathways/immune-cell-markers-human

Hierarchical data structure

- Single-cell data is hierarchical/clustered in nature
- Resolution of inference depends on research hypothesis and quality of data

- With hashed (multi-patient) data, an addition level of hierarchy appears

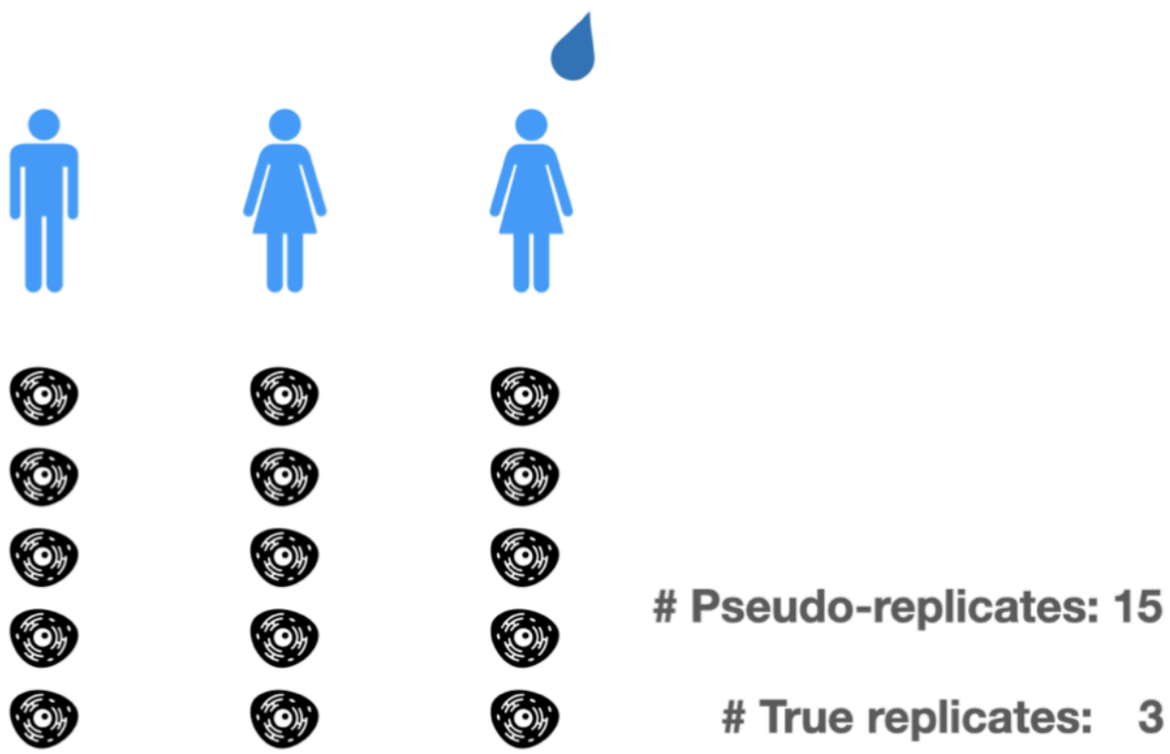
Hierarchical data structure

- Single-cell data is hierarchical/clustered in nature
- Resolution of inference depends on research hypothesis and quality of data

- With hashed (multi-patient) data, an addition level of hierarchy appears

  -> cells of the same patient are more similar than cells of different patients

  -> individual cells can be considered **pseudo replicates**



Outline

- **Single-cell transcriptomics:** recent advances in protocols and data
- **Muscat:** multi-patient multi-condition differential expression analyses
- **satuRn:** transcript-level inference for single-cell data

Muscat

- Published by the Mark Robinson group in Nature Communications (2020)
- Bioconductor package

- Method for **multi-patient**, multi-condition differential expression (DE) analysis

Muscat

- Published by the Mark Robinson group in Nature Communications (2020)
- Bioconductor package

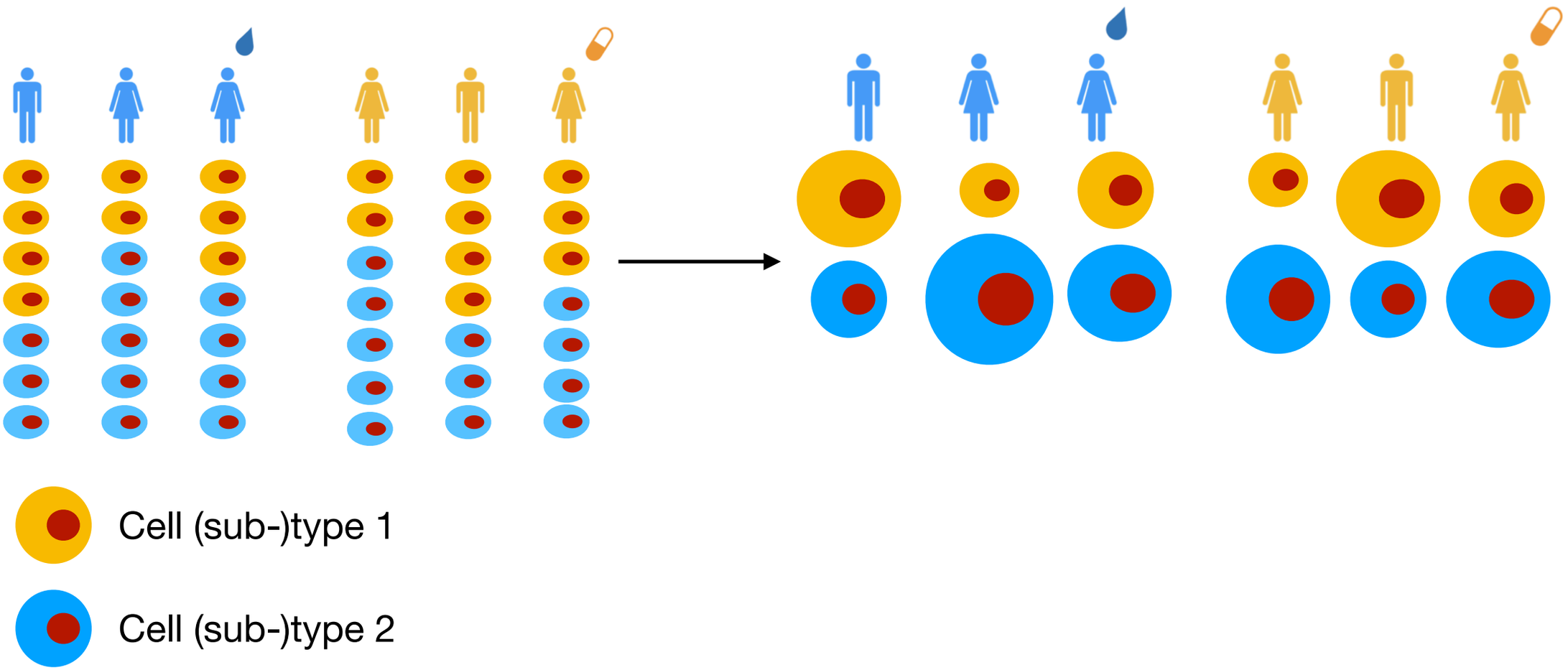
- Method for **multi-patient**, multi-condition differential expression (DE) analysis
- **Aggregates single-cell data to pseudo-bulk**
- Applies edgeR on pseudo-bulk data

Muscat

- Aggregates single-cell data to pseudo-bulk

  -> summation of the counts of individual cells to some higher hierarchical level

  -> a single count per cell (sub-)type, per patient



Muscat

- Aggregates single-cell data to pseudo-bulk

  -> summation of the counts of individual cells to some higher hierarchical level

  -> a single count per cell (sub-)type, per patient

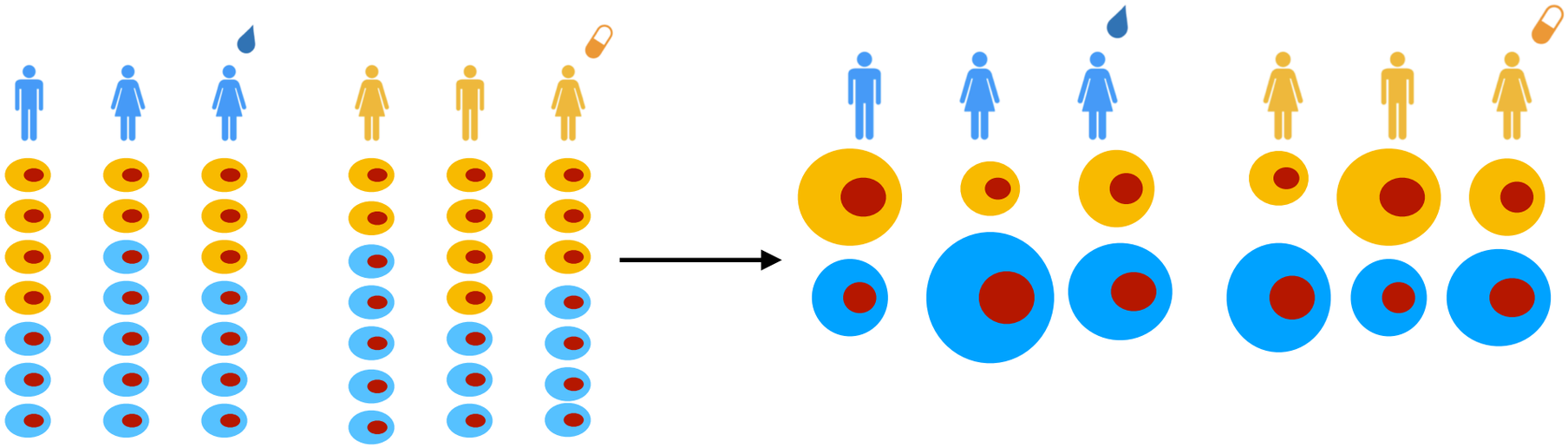
- Pseudo-bulk data != bulk data

- Still able to differentiate between cell (sub-)types

Muscat

Advantages

- Fast
- Data less sparse -> negative binomial assumption
- Avoids pseudoreplication bias issues



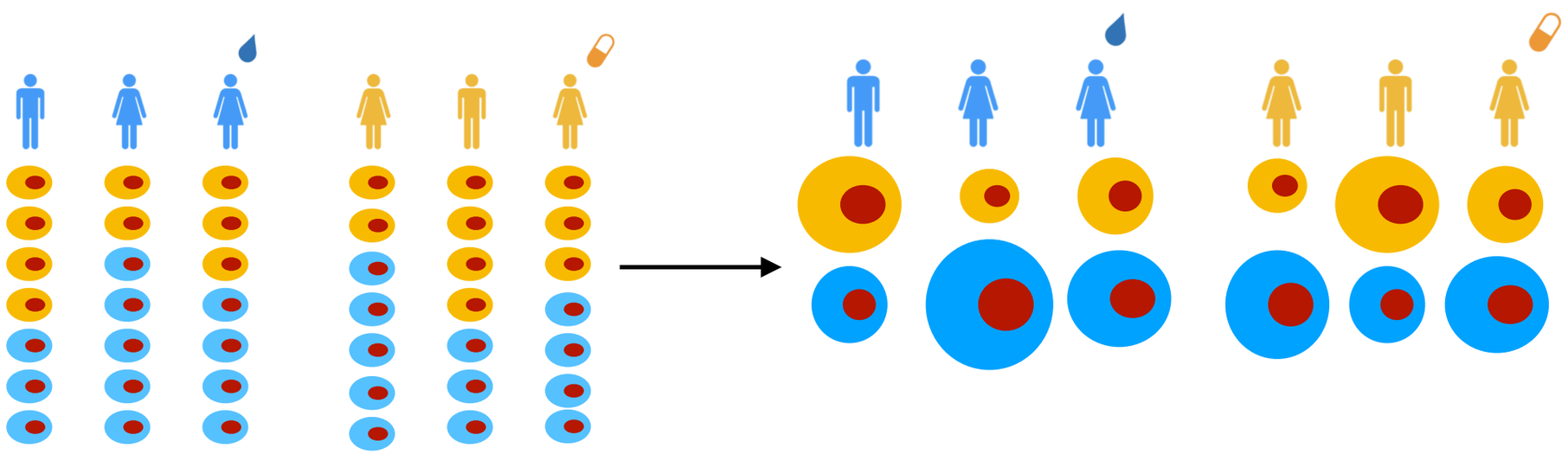
Muscat

Advantages

- Fast
- Data less sparse -> negative binomial assumption
- Avoids pseudoreplication bias issues

Disadvantages

- Few replicates -> low power
- Sensitive to imbalances in the number of aggregated cells



Muscat

Advantages

- Fast
- Data less sparse -> negative binomial assumption
- Avoids pseudoreplication bias issues

Disadvantages

- Few replicates -> low power
- Sensitive to imbalances in the number of aggregated cells

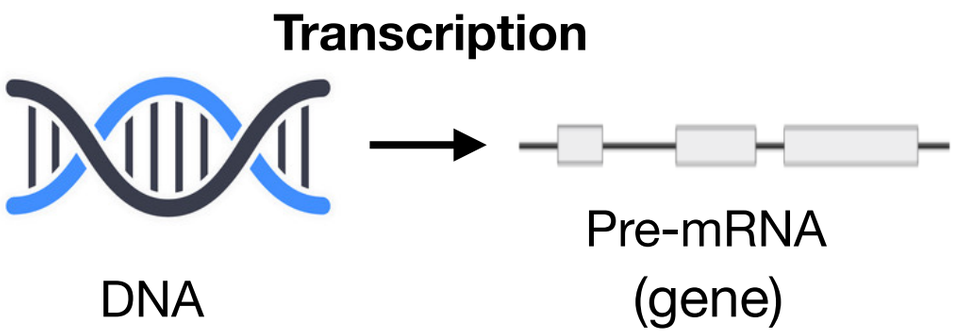
Alternatives

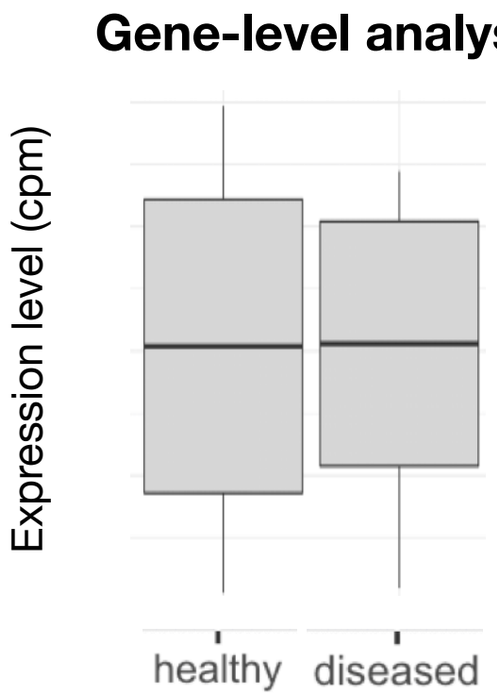
- *Distinct* R package
- Methods that specifically account for hierarchical nature of single-cell data

Outline

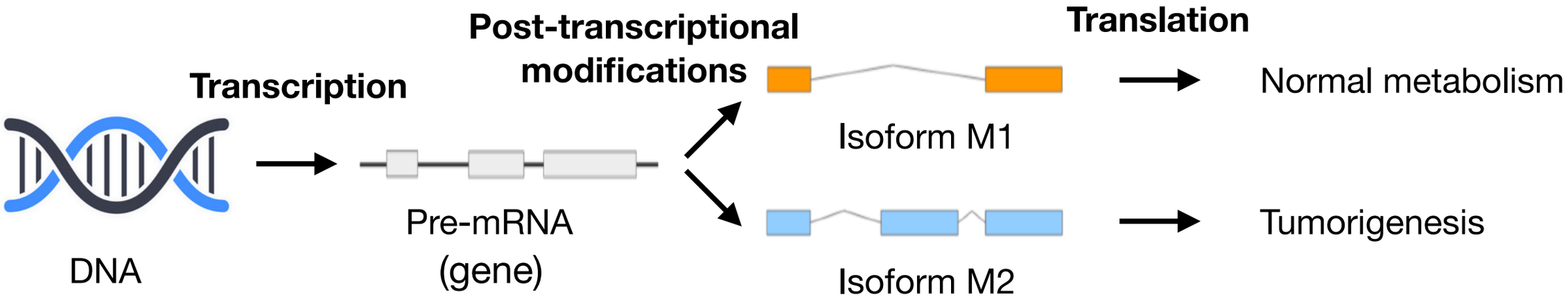
- **Single-cell transcriptomics:** recent advances in protocols and data
- **Muscat:** multi-patient multi-condition differential expression analyses
- **satuRn:** transcript-level inference for single-cell data

Differential Transcript Usage (DTU)

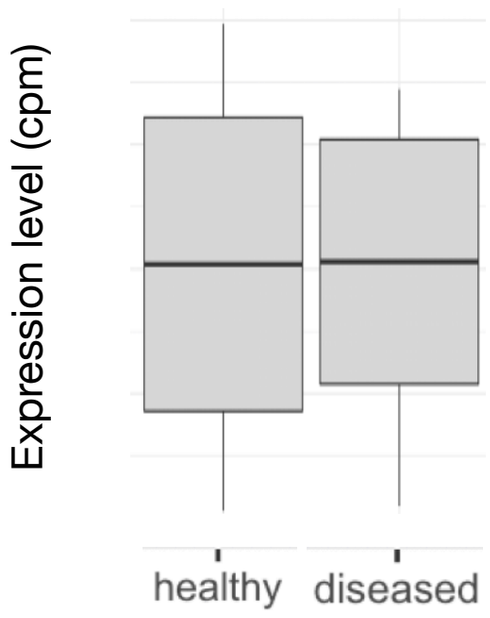




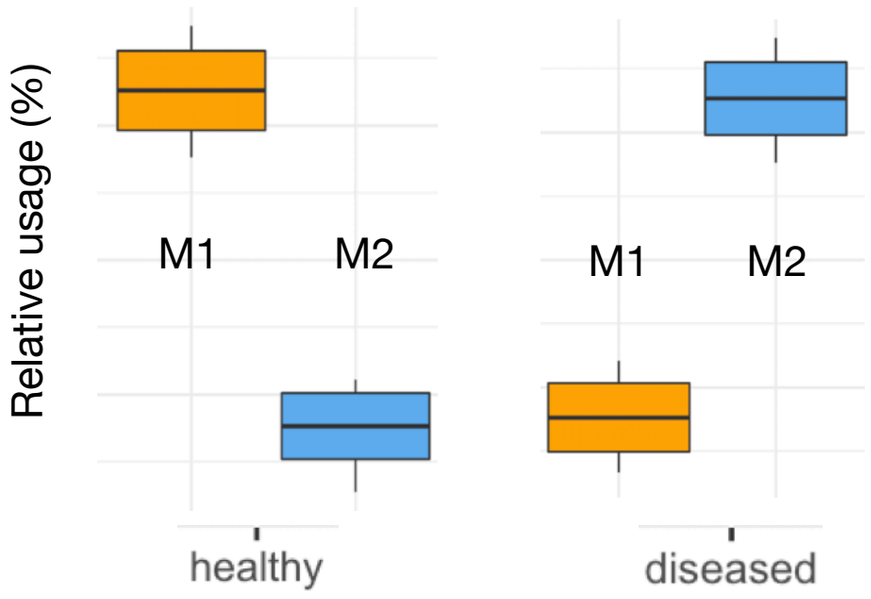
Differential Transcript Usage (DTU)



Gene-level analysis



Transcript-level DTU analysis



Prerequisites for DTU analysis

- **Full-length** RNA-seq data
  -> Transcript-level abundances require sequencing reads from both 3’ and 5’ end
  - **SMART-seq**, SMARTer, Quartz-seq
  - **Long read** RNA protocols (PacBio, Oxford Nanopore)
  - Not* 10X, Visium, Drop-seq, CEL-seq, InDrop, MARS-seq

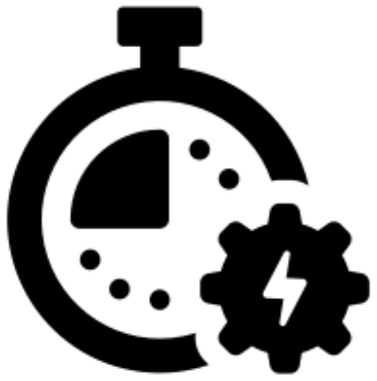
Prerequisites for DTU analysis

- **Full-length** RNA-seq data
  -> Transcript-level abundances require sequencing reads from both 3’ and 5’ end
  - **SMART-seq**, SMARTer, Quartz-seq
  - **Long read** RNA protocols (PacBio, Oxford Nanopore)
  - Not* 10X, Visium, Drop-seq, CEL-seq, InDrop, MARS-seq

- **Splice-aware** alignment
  - Ambiguity in assigning reads to transcripts
  - **Pseudo-alignment** tools like kallisto, salmon and sailfish
  - STAR, HISAT2
  - Bowtie

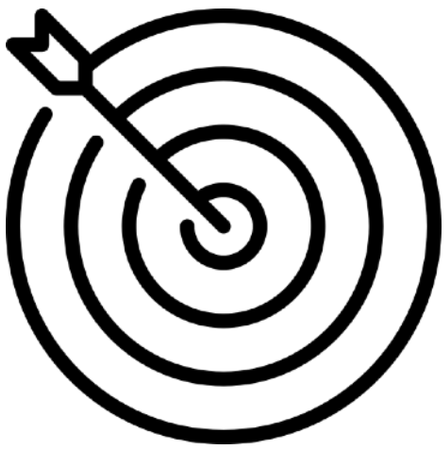
What makes a for good DTU analysis method?

Scalability

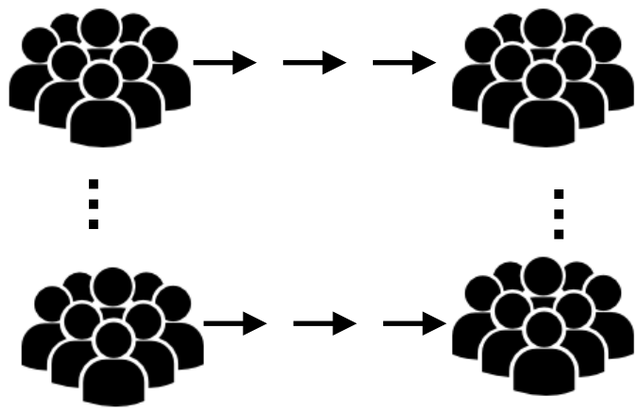


Performance

Type 1 error control



Complex designs



Sparse data

0

Scalable **a**nalysis of differential **t**ranscript **u**sage for **RN**a-seq data

Software development

- Denote the expression of transcript *t* of gene *g* in sample *i* as Y_{gti}
- Denote the usage of transcript *t* of gene *g* in sample *i* as*:*

$$U_{gti} = \frac{Y_{gti}}{Y_{g.i}}$$

Software development

- Denote the expression of transcript *t* of gene *g* in sample *i* as Y_{gti}
- Denote the usage of transcript *t* of gene *g* in sample *i* as:

$$U_{gti} = \frac{Y_{gti}}{Y_{g.i}}$$

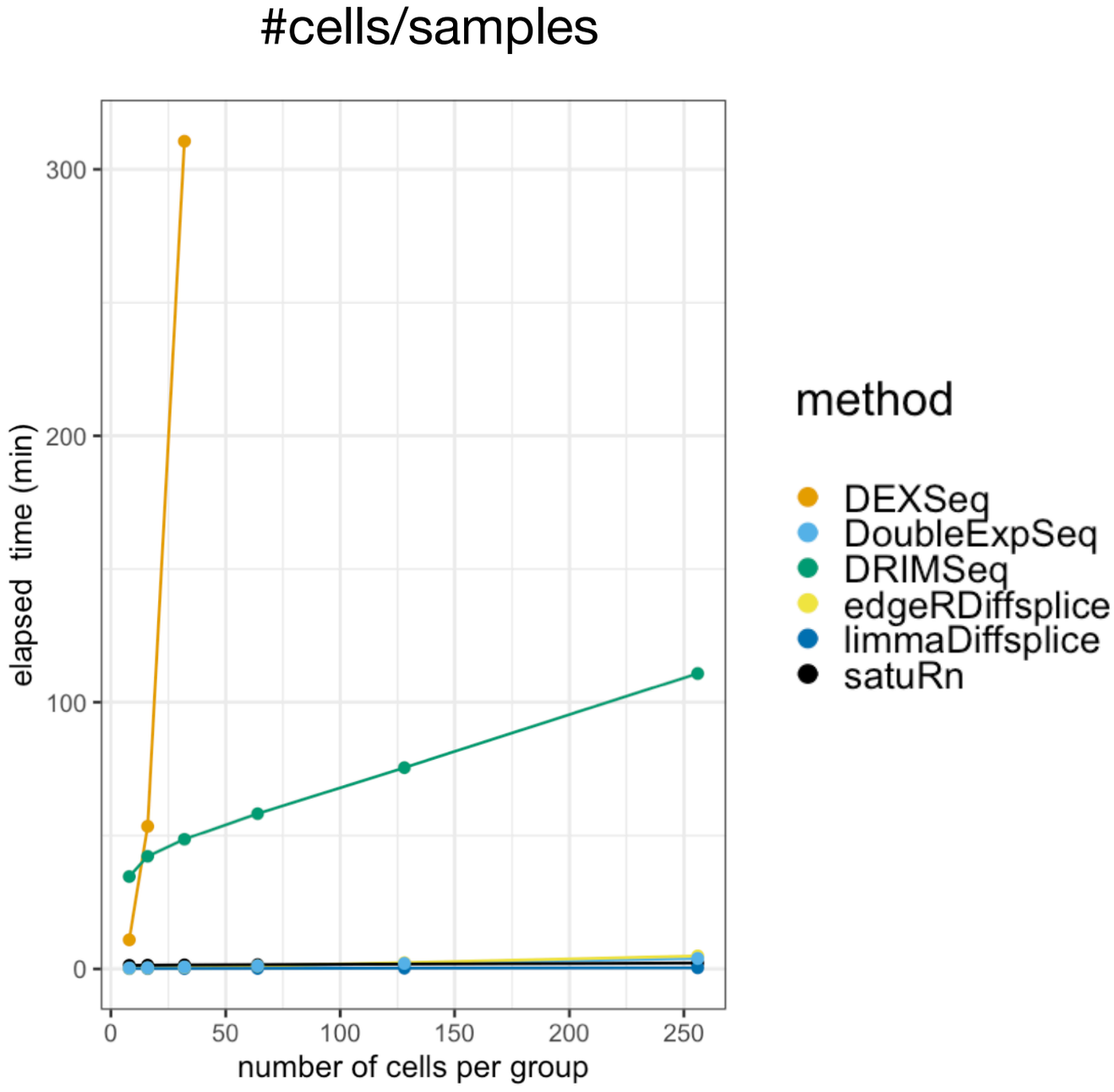
- Describe the **quasi-binomial** GLM:

$$\begin{cases} E\left[U_{gti} \middle| \boldsymbol{X_i}, Y_{g.i}\right] = \pi_{gti} \\ \log\left(\frac{\pi_{gti}}{1-\pi_{gti}}\right) = \eta_{gti} \\ \eta_{gti} = \boldsymbol{X_i^T \beta_{gt}} \end{cases}$$

- With variance:

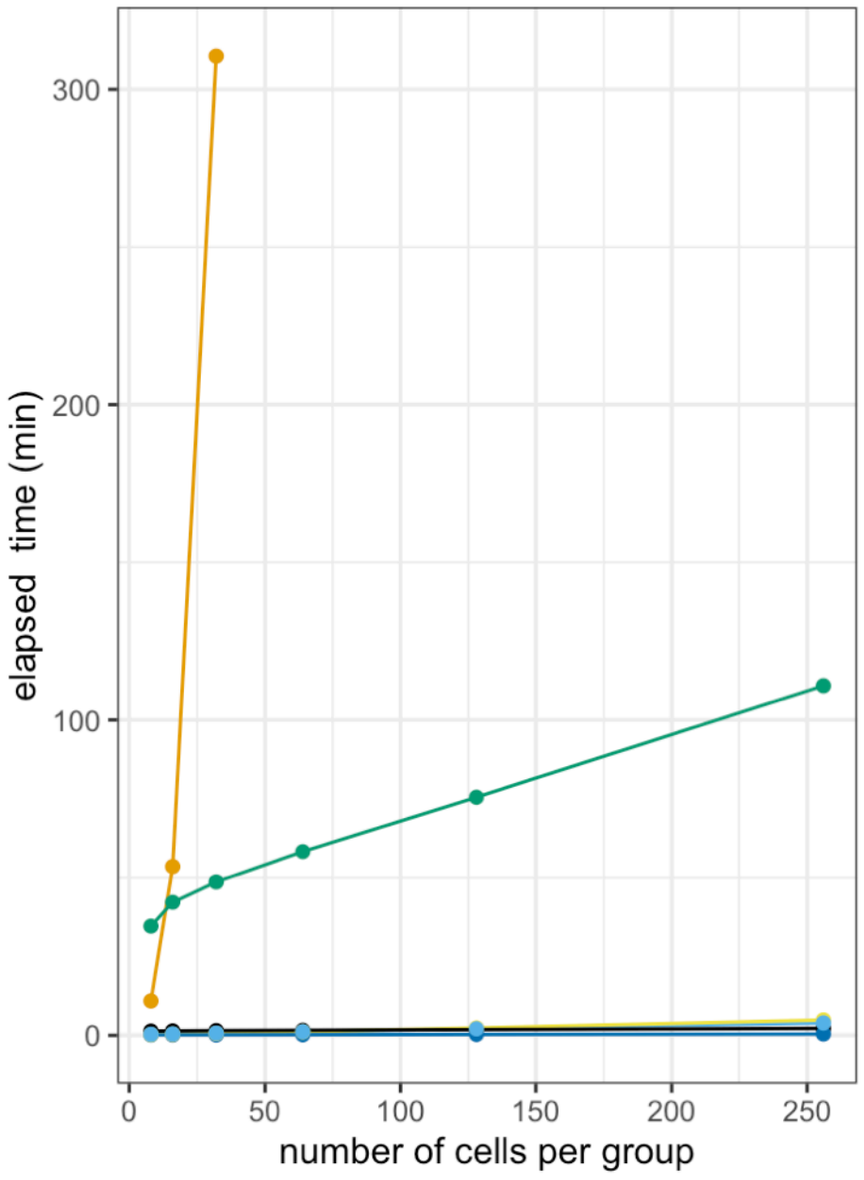
$$Var\left[U_{gti} \middle| \boldsymbol{X_i}, Y_{g.i}\right] = \frac{\pi_{gti} * (1-\pi_{gti})}{Y_{g.i}} * \phi_{gt}$$

Scalability

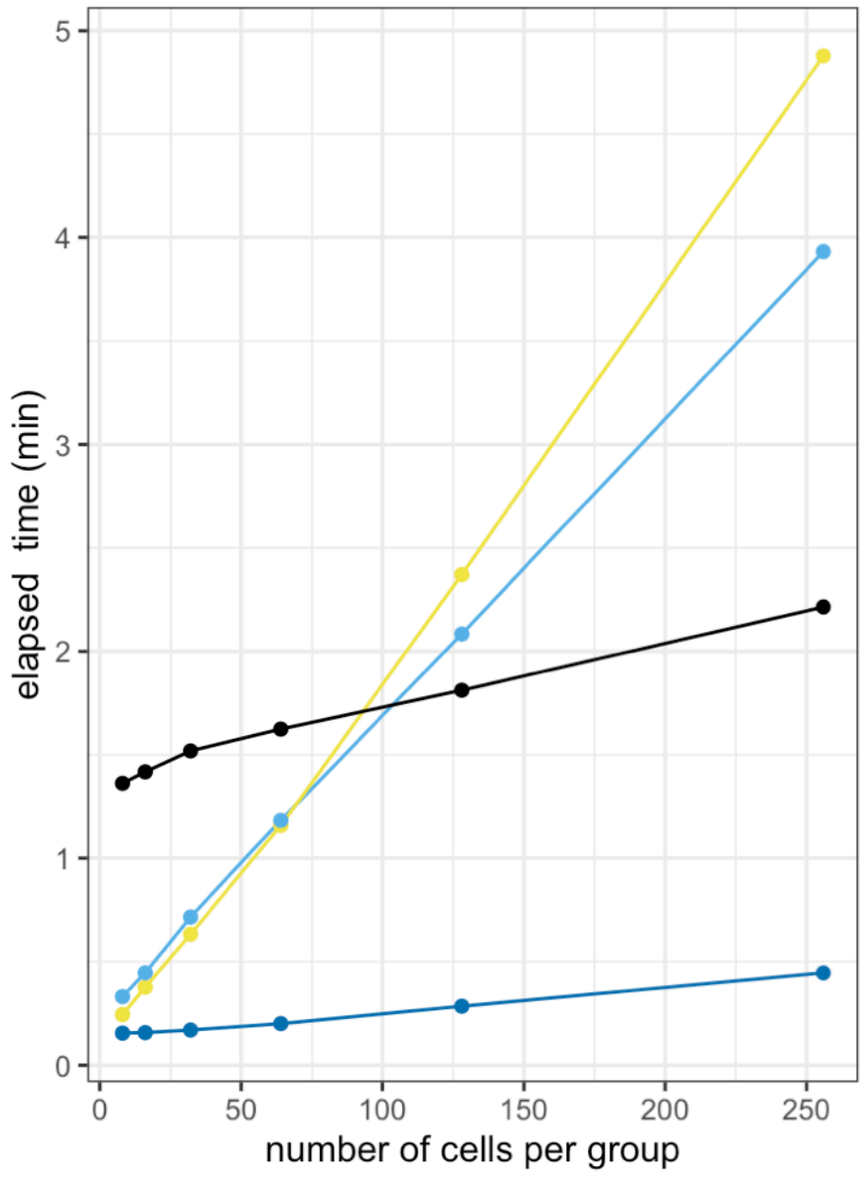


Scalability

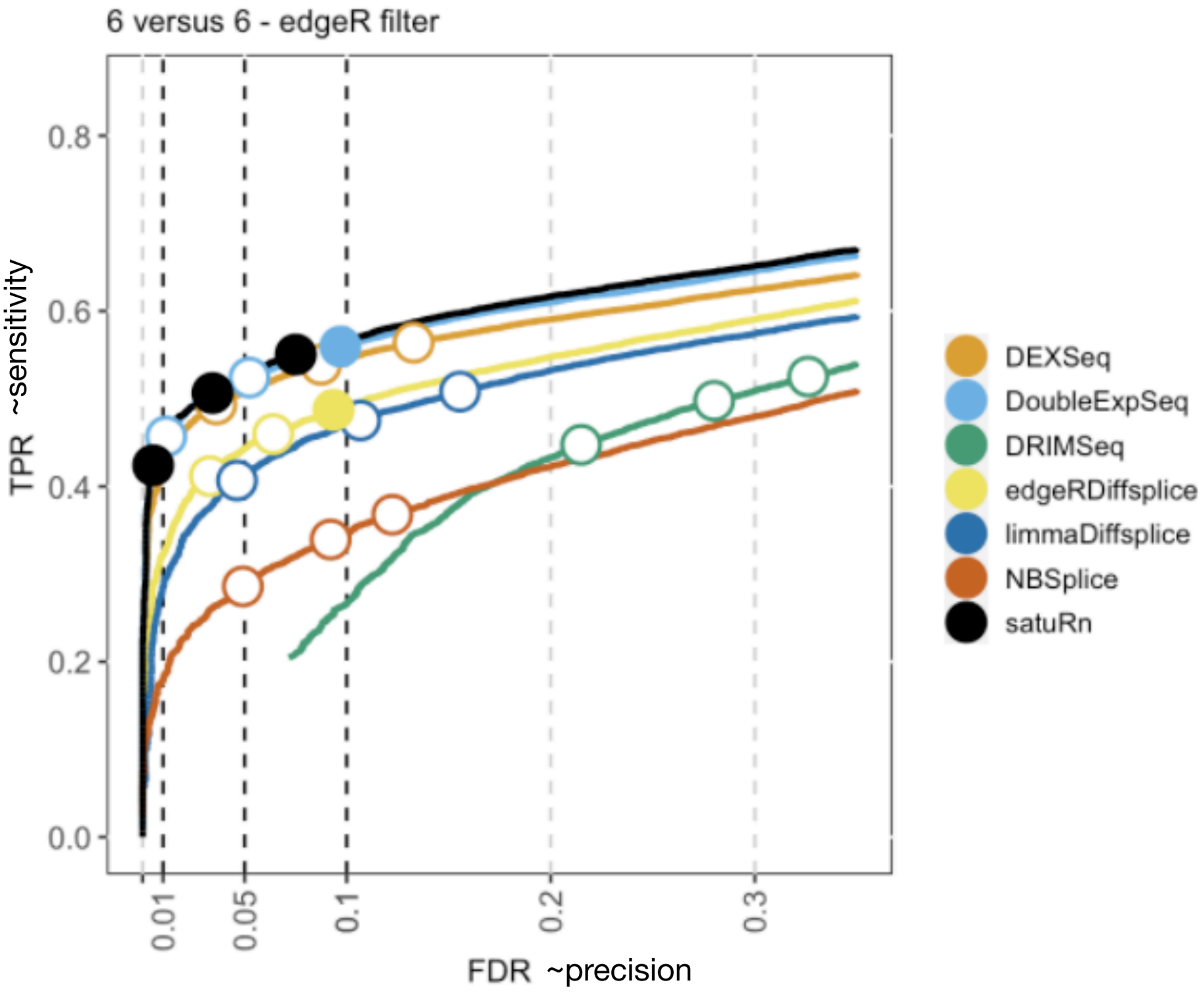
#cells/samples



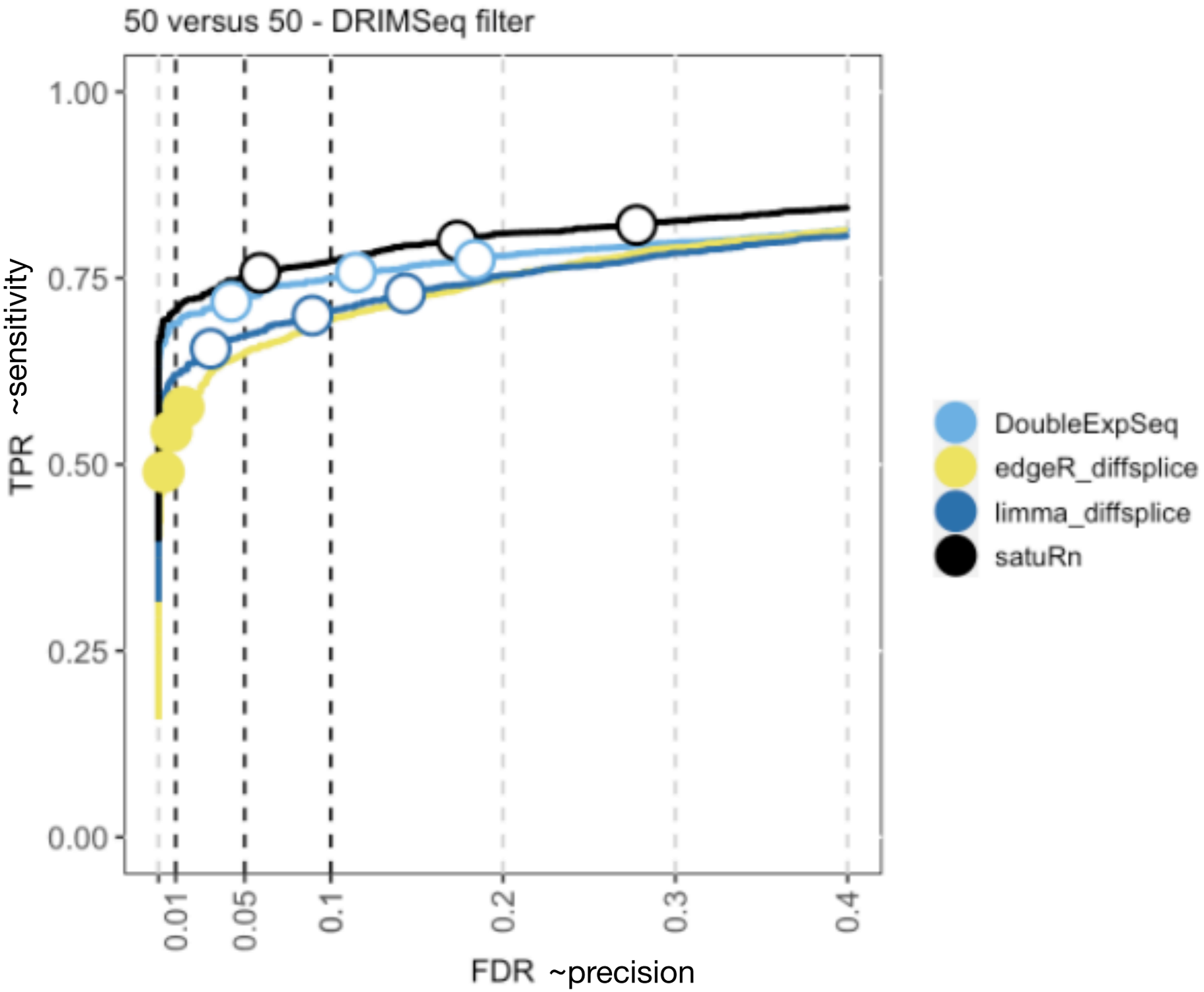
#cells/samples (zoom)



Good performance in bulk RNA-Seq



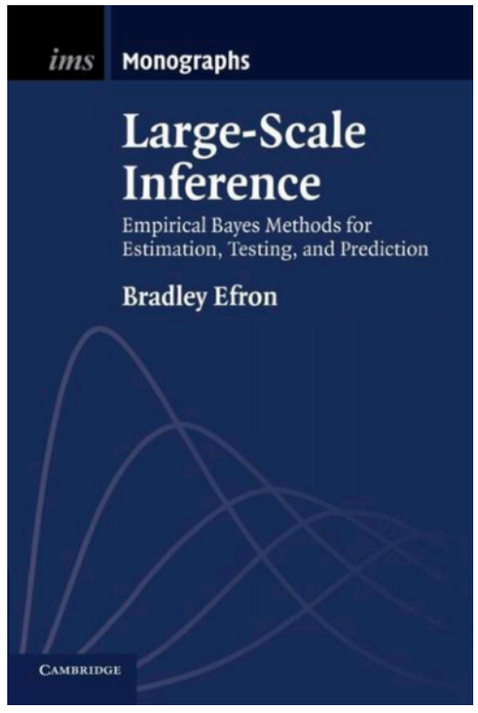
Poor FDR control in scRNA-Seq

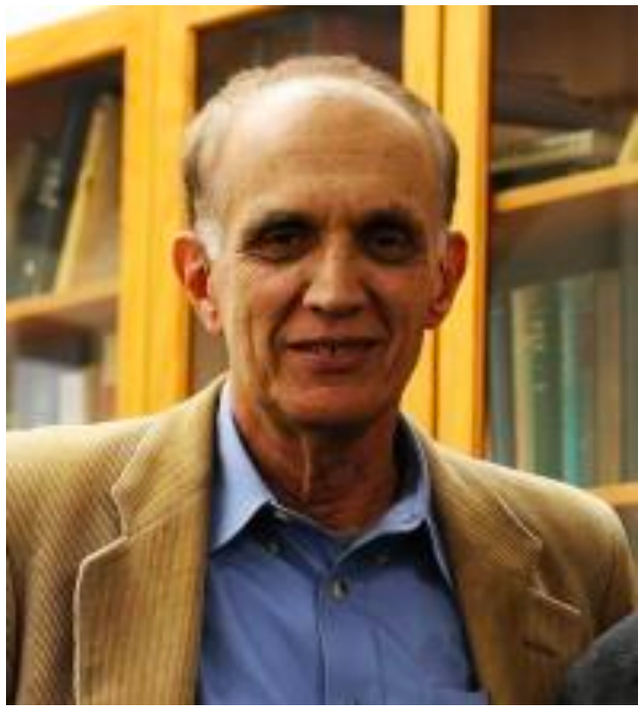


FDR control

Potential issues:

- Transcript-transcript correlation
- Cell-cell correlation
- Unobserved confounders





Solution: empirical null distribution

In practice:

1. Take p-values p_{gt} and convert to z-scores (inverse CDF)

$$z_{gt} = \Phi^{-1}\left(\frac{p_{gt}}{2}\right) * sign(S)$$

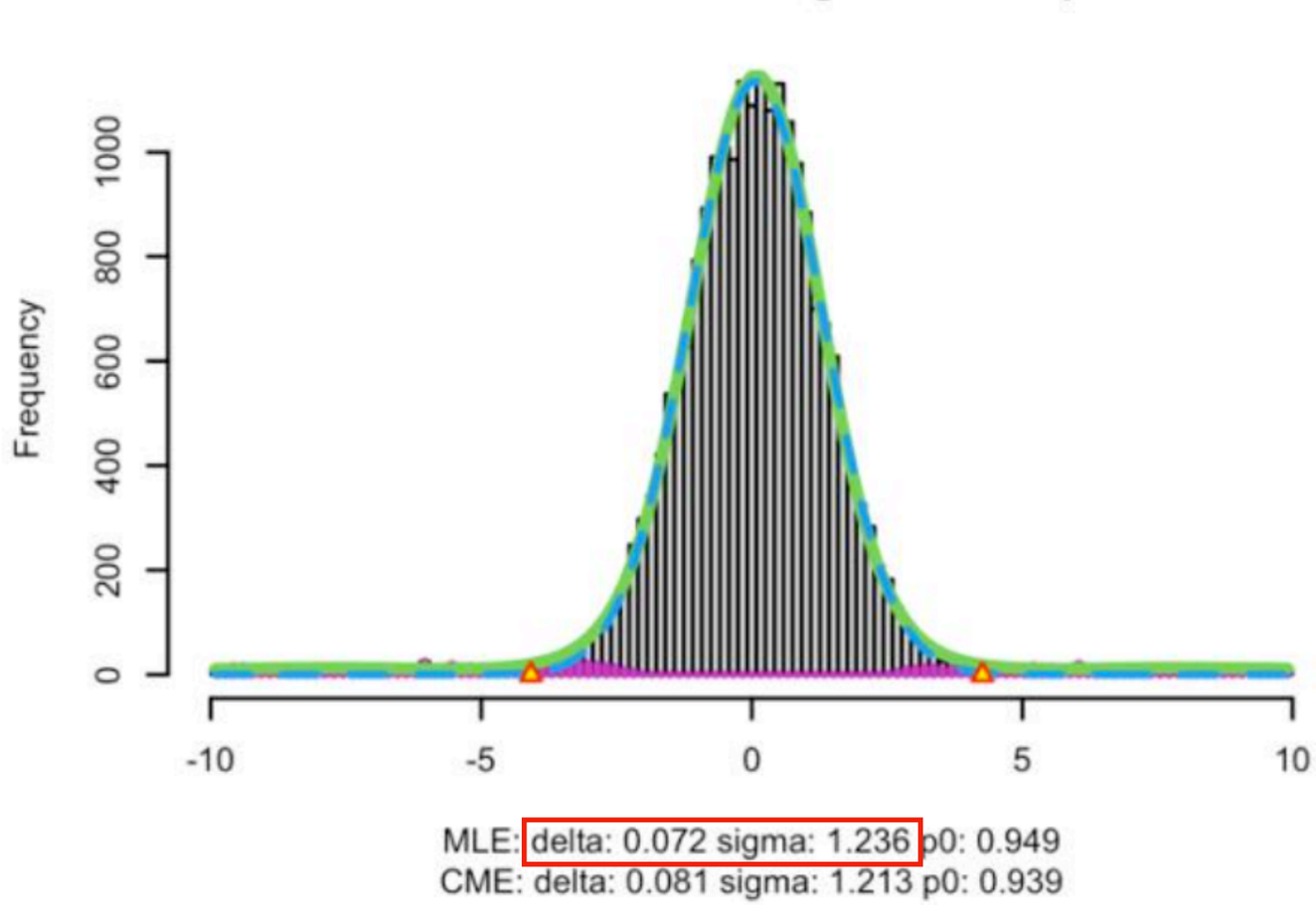
Solution: empirical null distribution

In practice:

1. Take p-values p_{gt} and convert to z-scores (inverse CDF)

$$z_{gt} = \Phi^{-1}\left(\frac{p_{gt}}{2}\right) * sign(S)$$

2. Empirically determine how the null tests (mid 50%) are distributed



Solution: empirical null distribution

In practice:

1. Take p-values p_{gt} and convert to z-scores (inverse CDF)

$$z_{gt} = \Phi^{-1}\left(\frac{p_{gt}}{2}\right) * sign(S)$$

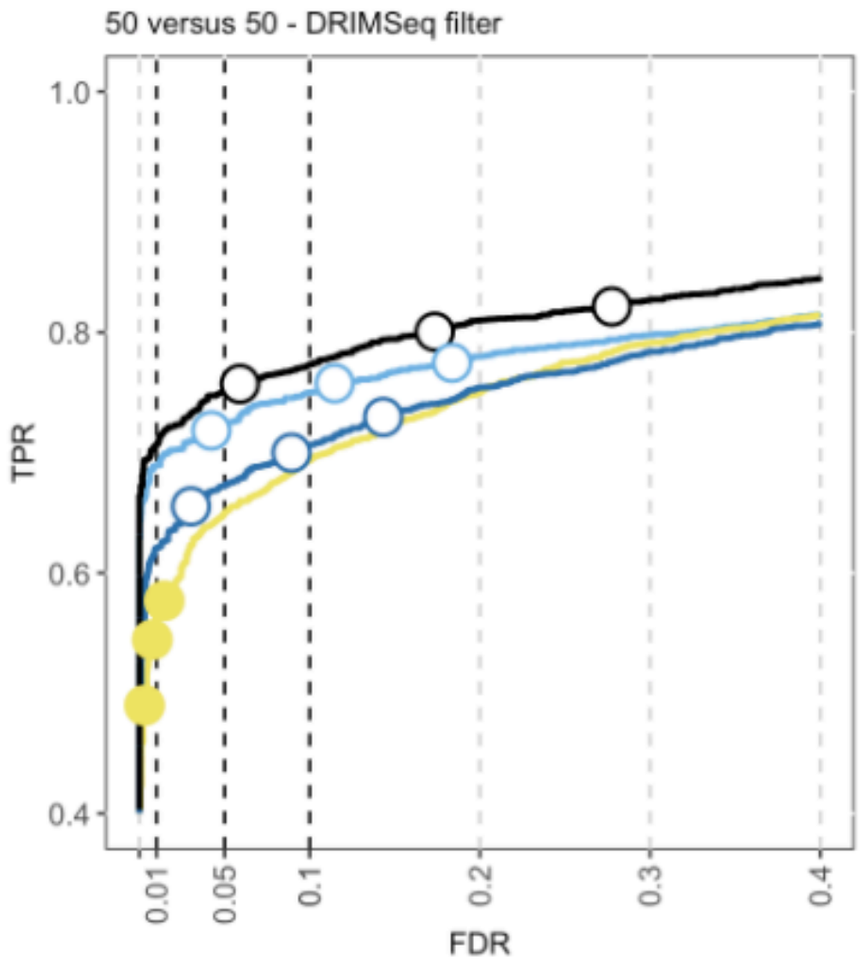
2. Empirically determine how the null tests (mid 50%) are distributed

3. Recompute p-values given the new null

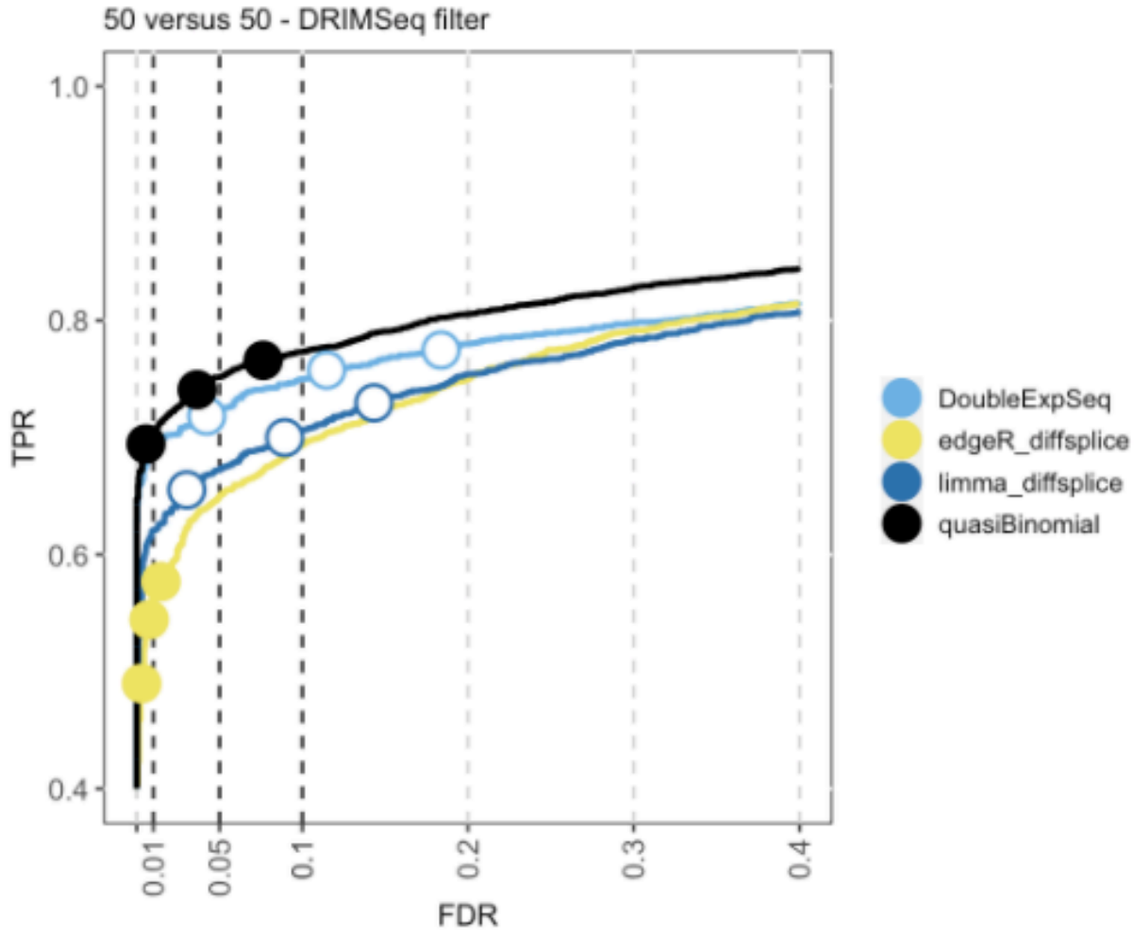
$$z^*_{gt} = \frac{(z_{gt} - \mu^*)}{\sigma^*}$$

$$p^*_{gt} = 2 * \Phi\left(-abs\left(z^*_{gt}\right)\right)$$

FDR control in scRNA-Seq restored

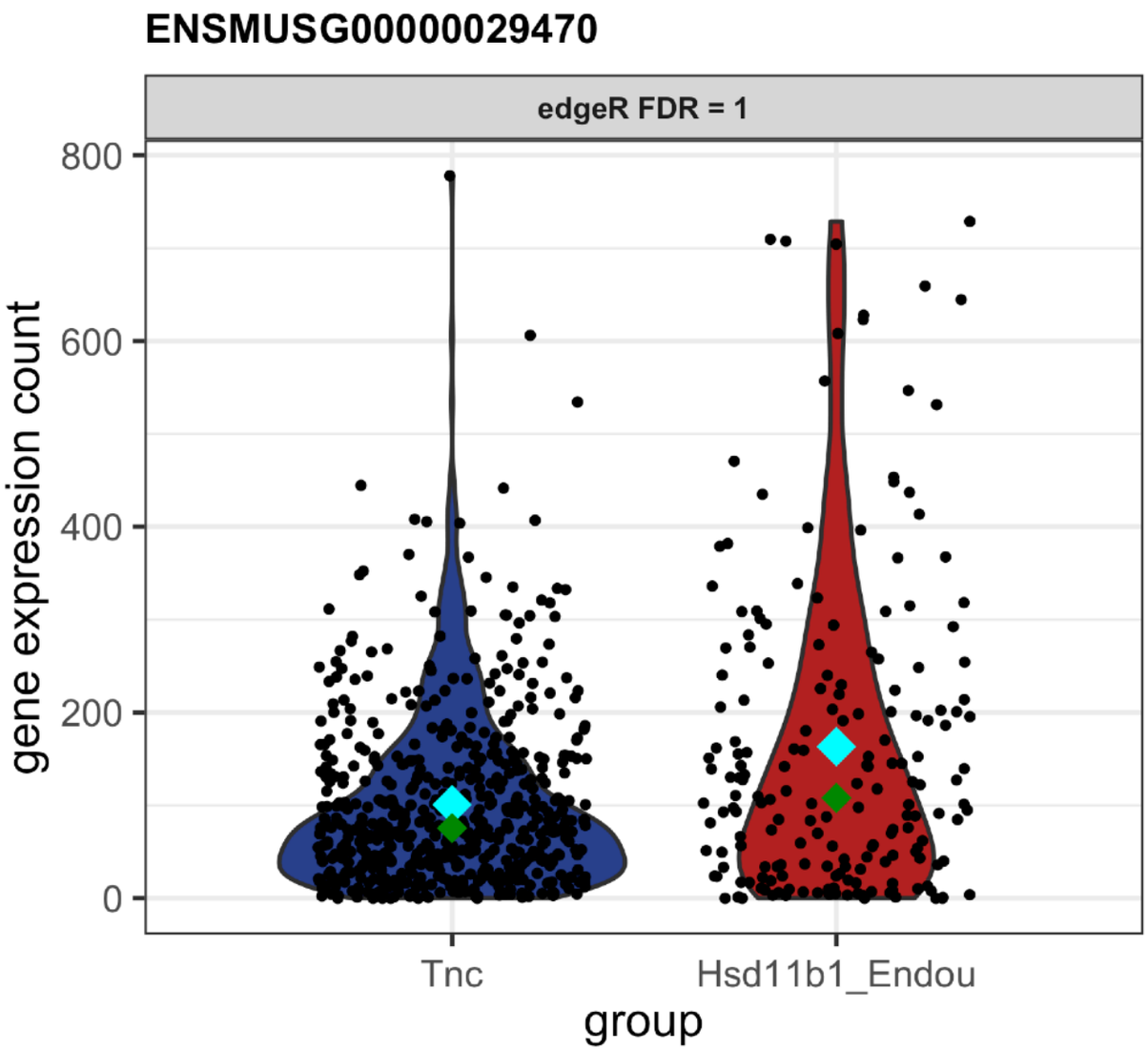


Empirical null estimation



Case study

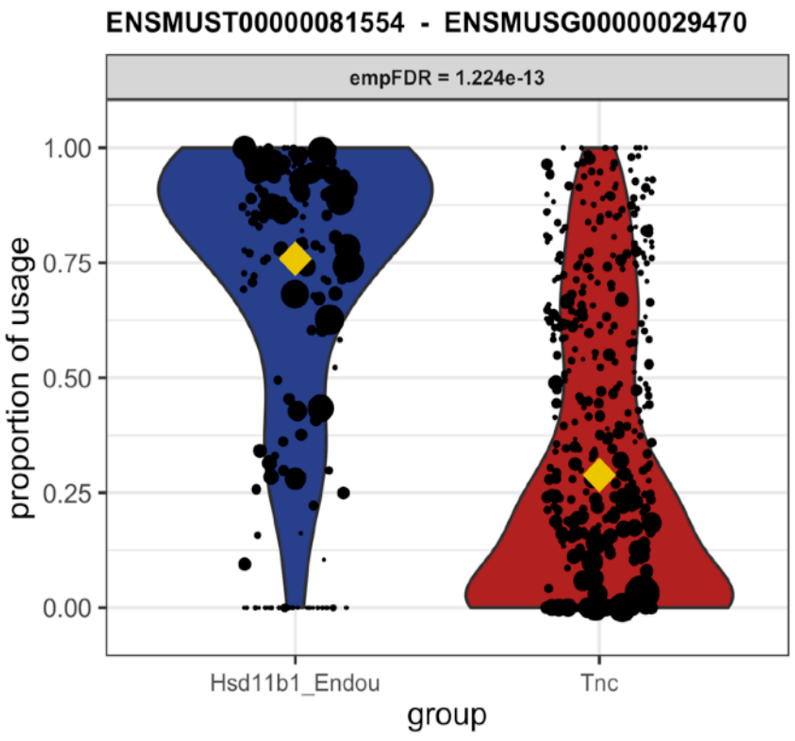
<u>No evidence for differential gene expression</u>

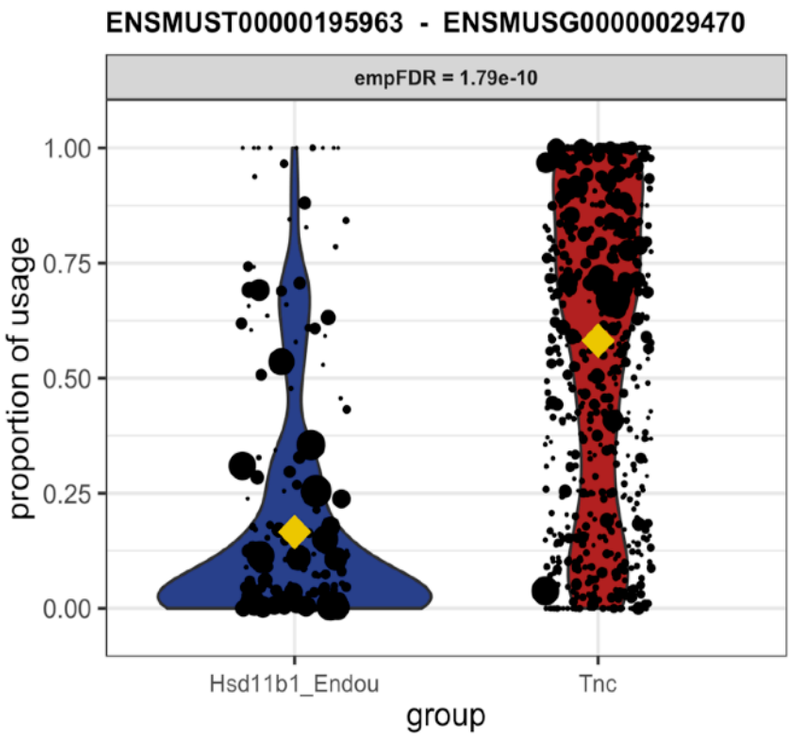


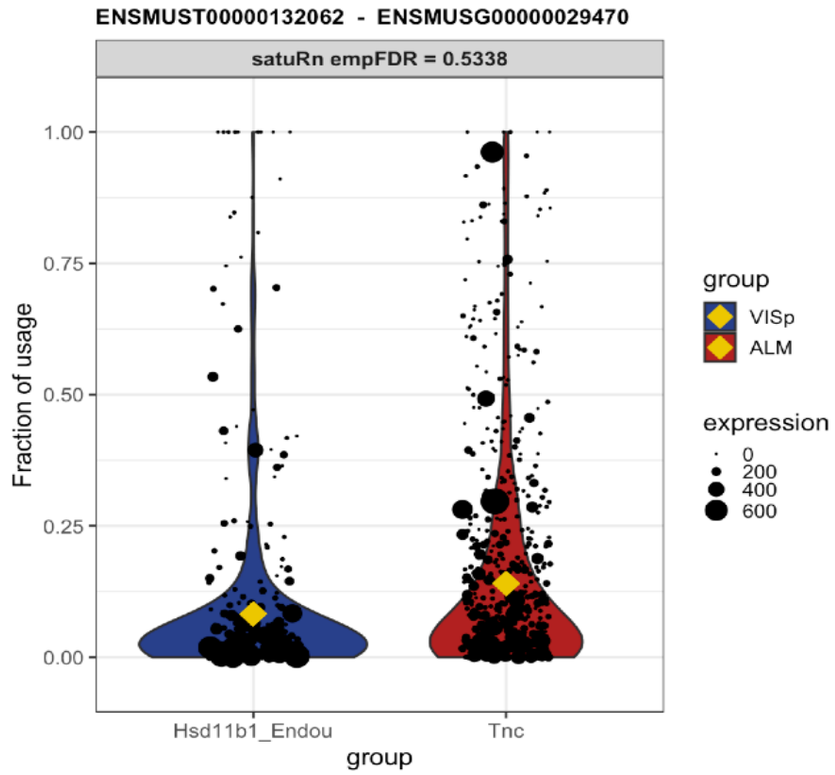
Dataset obtained from Tasic et al. (2018), Nature 563, 72–78

Case study

Strong evidence for differential transcript usage







Crucially, the left isoform is protein coding, while the middle isoform is not

Dataset obtained from Tasic et al. (2018), Nature 563, 72–78

Case study

- DGE and DTU between different cell types
- Number of DGE genes associated with number of genes with DTU transcripts
- **Limited overlap: orthogonal information**

Comparison	Cell type 1 (ALM)	Cell type 2 (VISp)	DGE	DTU Gene	Overlap
1	Cpa6 Gpr88	Batf3	203	15	1
2	Cbln4 Fezf2	Col27a1	281	53	3
3	Cpa6 Gpr88	Col6a1 Fezf2	154	5	0
4	Gkn1 Pcdh19	Col6a1 Fezf2	231	22	1
5	Lypd1 Gpr88	Hsd11b1 Endou	331	69	4
6	Tnc	Hsd11b1 Endou	595	112	10
7	Tmem163 Dmrtb1	Hsd11b1 Endou	471	53	7
8	Tmem163 Arhgap25	Whrn Tox2	197	40	1

Case study

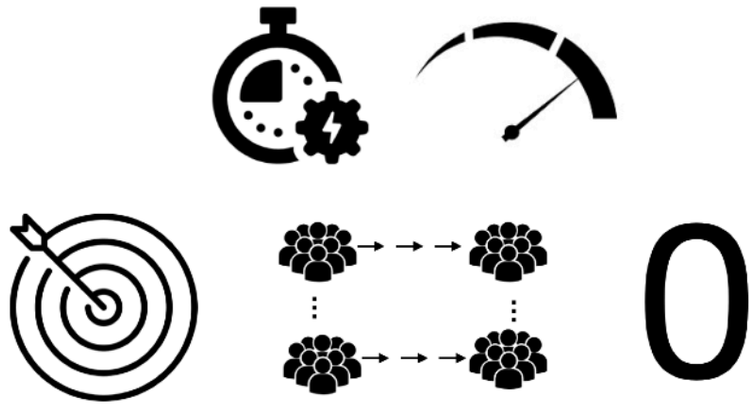
- DGE and DTU between different cell types
- Number of DGE genes associated with number of genes with DTU transcripts
- **Limited overlap: orthogonal information**

Comparison	Cell type 1 (ALM)	Cell type 2 (VISp)	DGE	DTU Gene	Overlap
1	Cpa6 Gpr88	Batf3	203	15	1
2	Cbln4 Fezf2	Col27a1	281	53	3
3	Cpa6 Gpr88	Col6a1 Fezf2	154	5	0
4	Gkn1 Pcdh19	Col6a1 Fezf2	231	22	1
5	Lypd1 Gpr88	Hsd11b1 Endou	331	69	4
6	Tnc	Hsd11b1 Endou	595	112	10
7	Tmem163 Dmrtb1	Hsd11b1 Endou	471	53	7
8	Tmem163 Arhgap25	Whrn Tox2	197	40	1

- **GSEA analysis:** similar gene sets from DGE and DTU

satuRn take-home

- *satuRn* is:



- Detects biologically relevant DTU signal in a case study

- Published in F1000Research (https://f1000research.com/articles/10-374)

- **Available from Bioconductor (https://bioconductor.org/packages/release/bioc/html/satuRn.html)**

Differential expression analysis for transcriptomics data

Recent advances in a rapidly evolving field

statOmics research group - Ghent University

Team leader
Prof. Lieven Clement

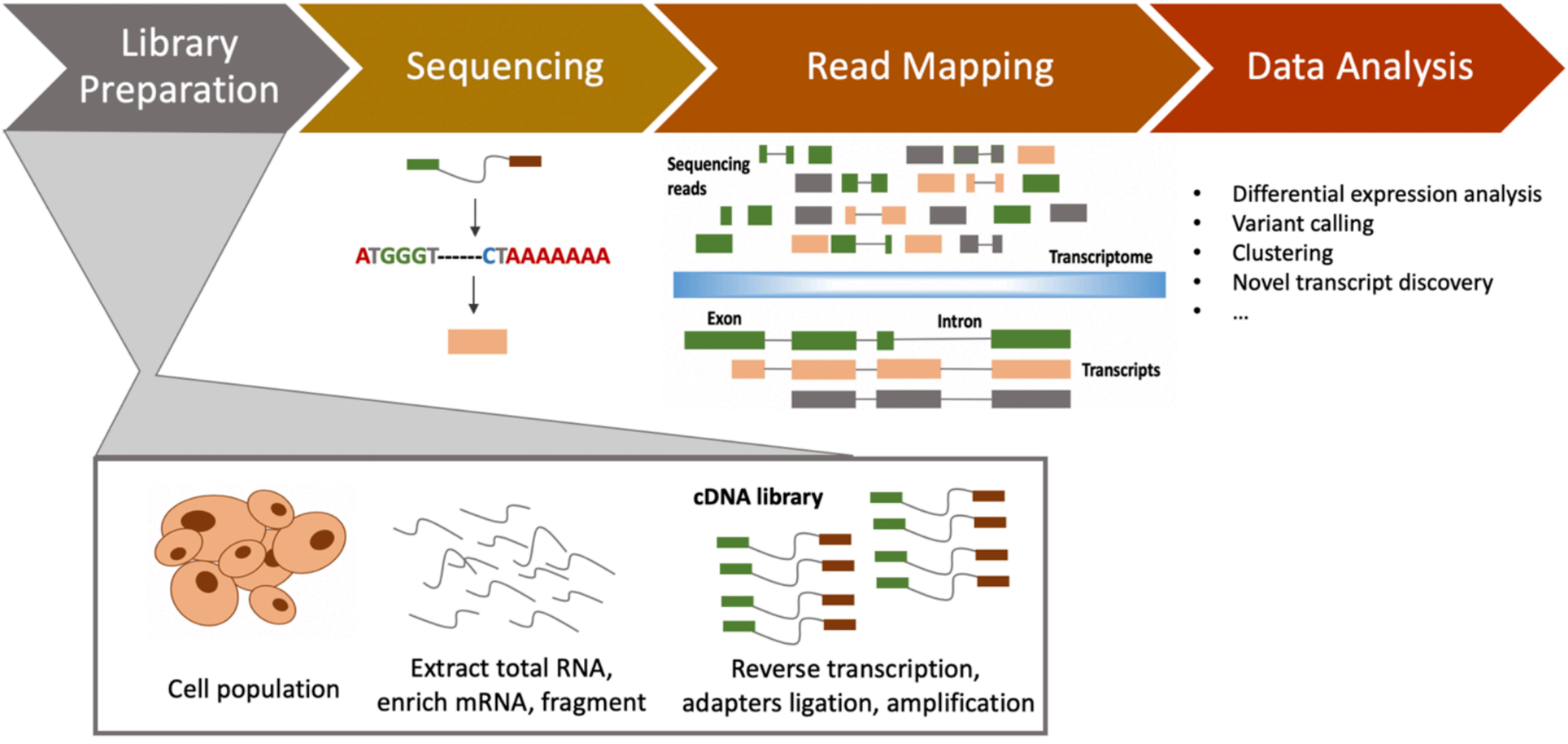
Transcriptomics and single-cell omics

Proteomics

Meta-omics

UNIVERSITEIT GENT

Bulk transcriptomics protocols

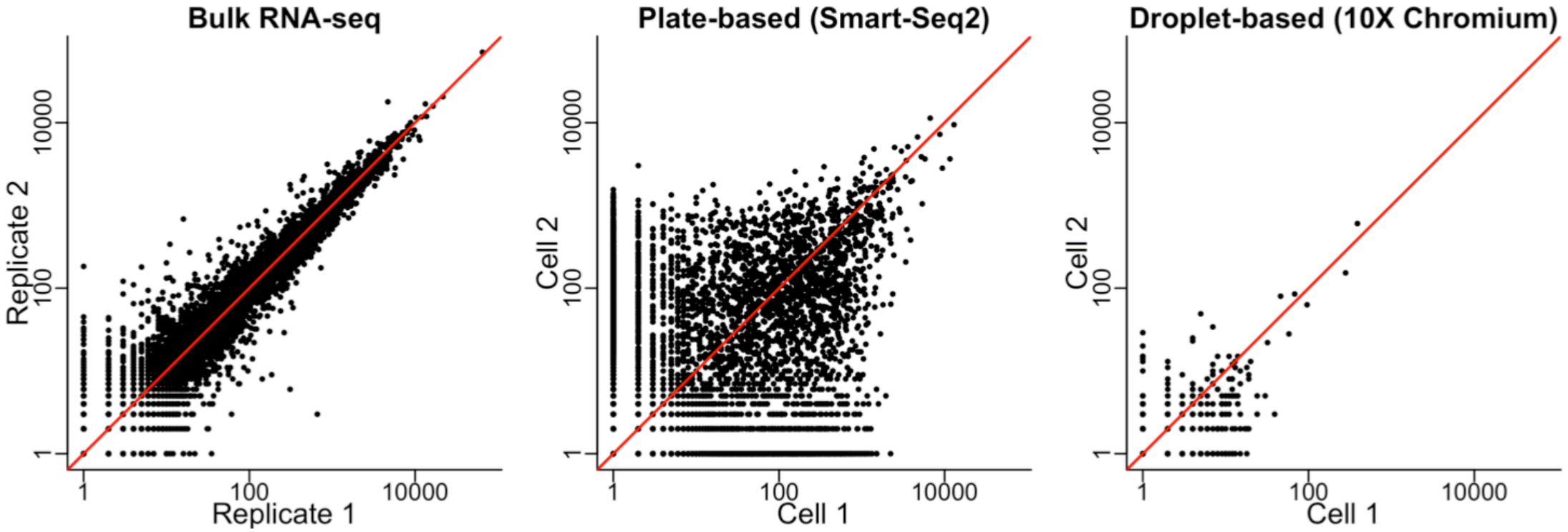


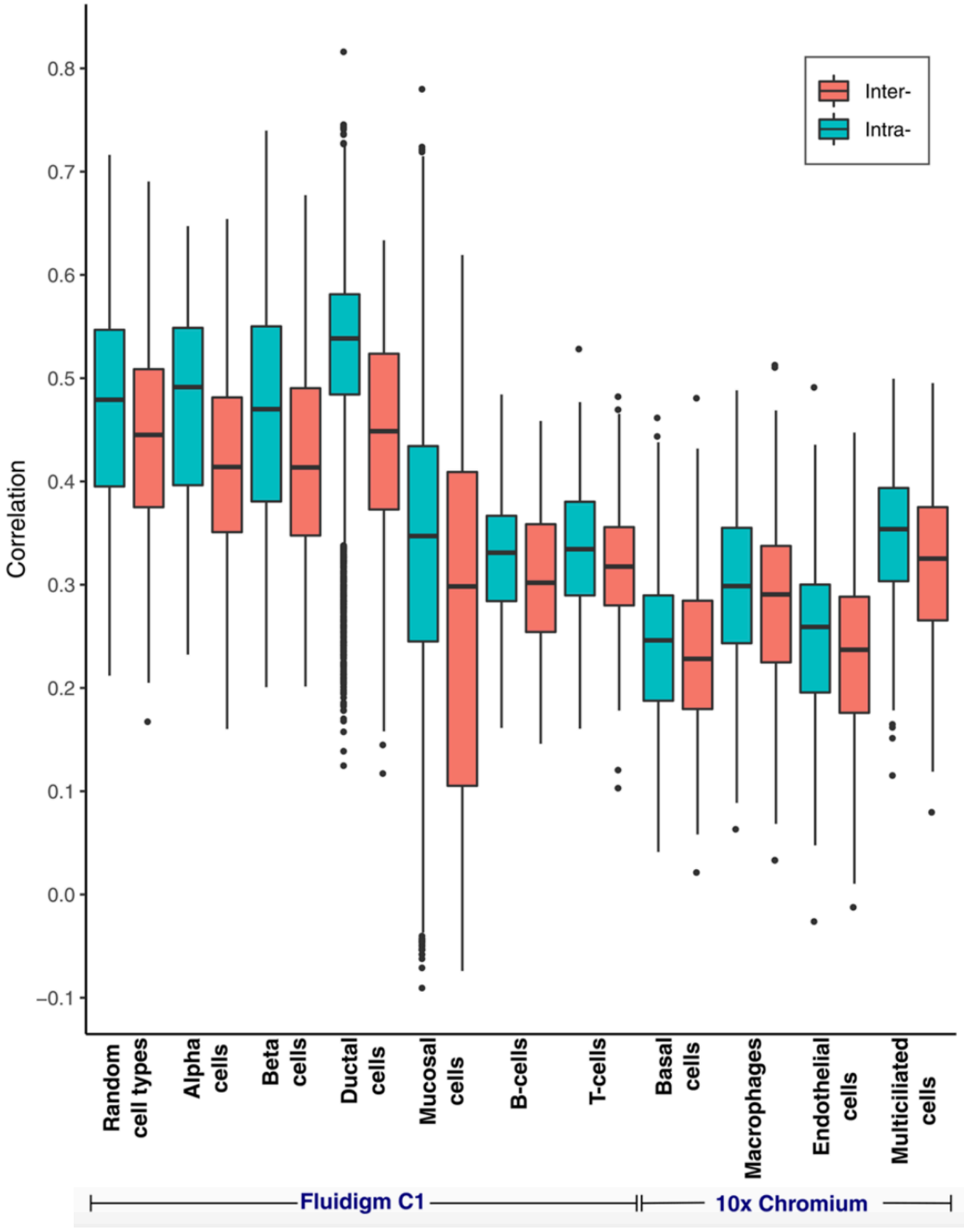
Bulk versus single-cell data

1. Higher technical variation in single-cell data

2. **Higher biological variation in single-cell data**

3. **Single-cell data is very sparse**





From Zimmerman et al. (2021), https://doi.org/10.1038/s41467-021-21038-1