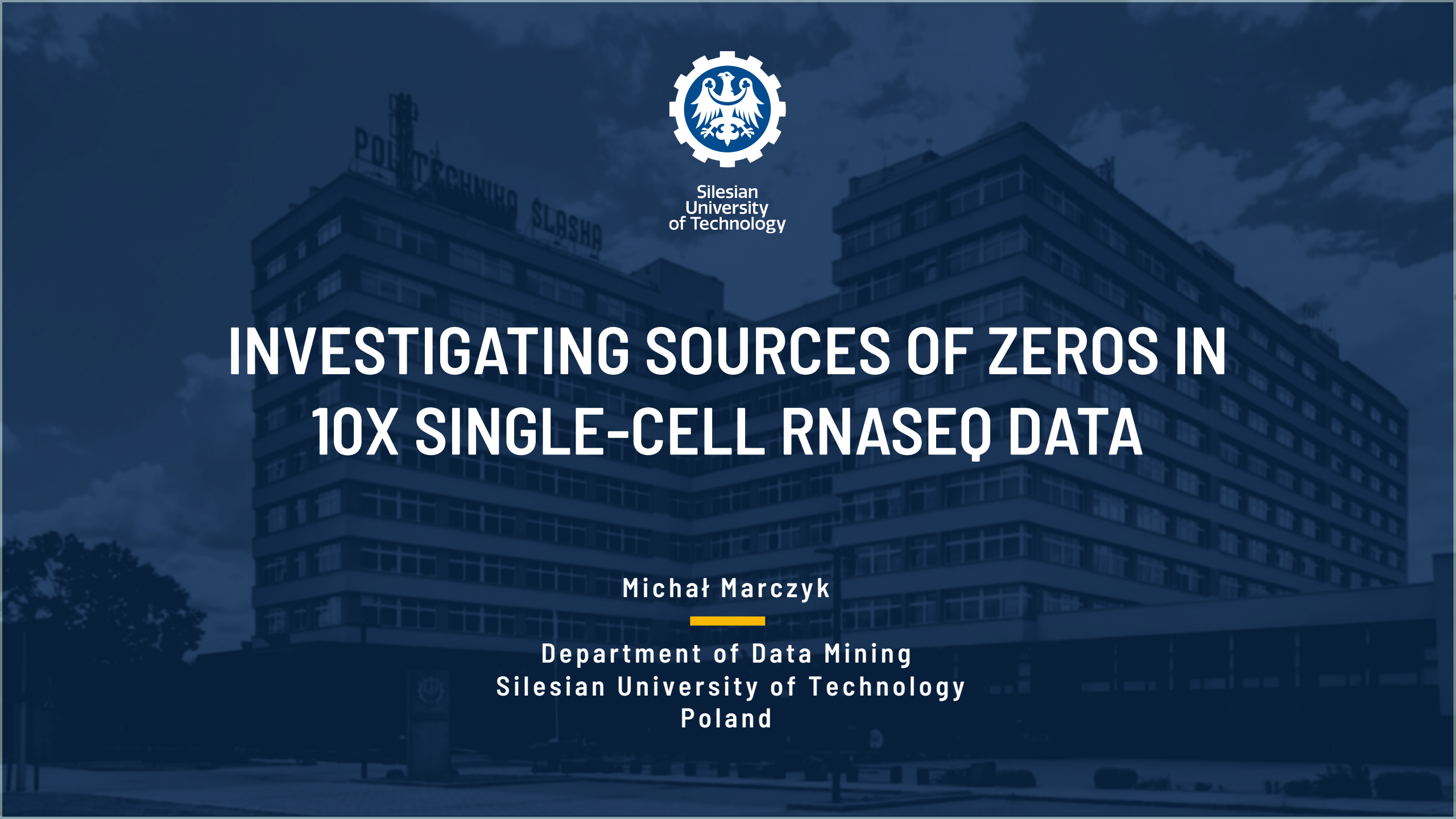


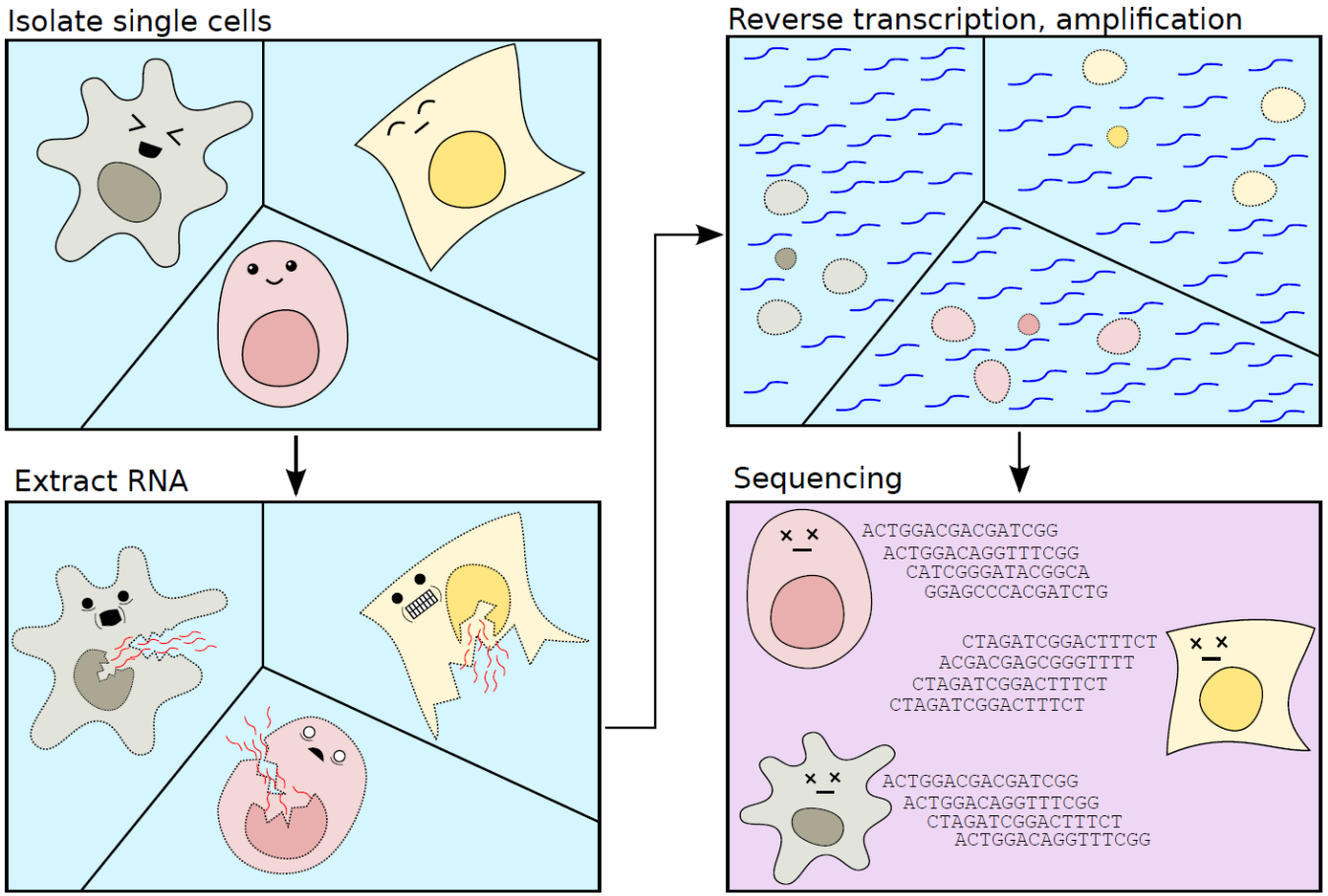
INVESTIGATING SOURCES OF ZEROS IN 10X SINGLE-CELL RNASEQ DATA

Michał Marczyk

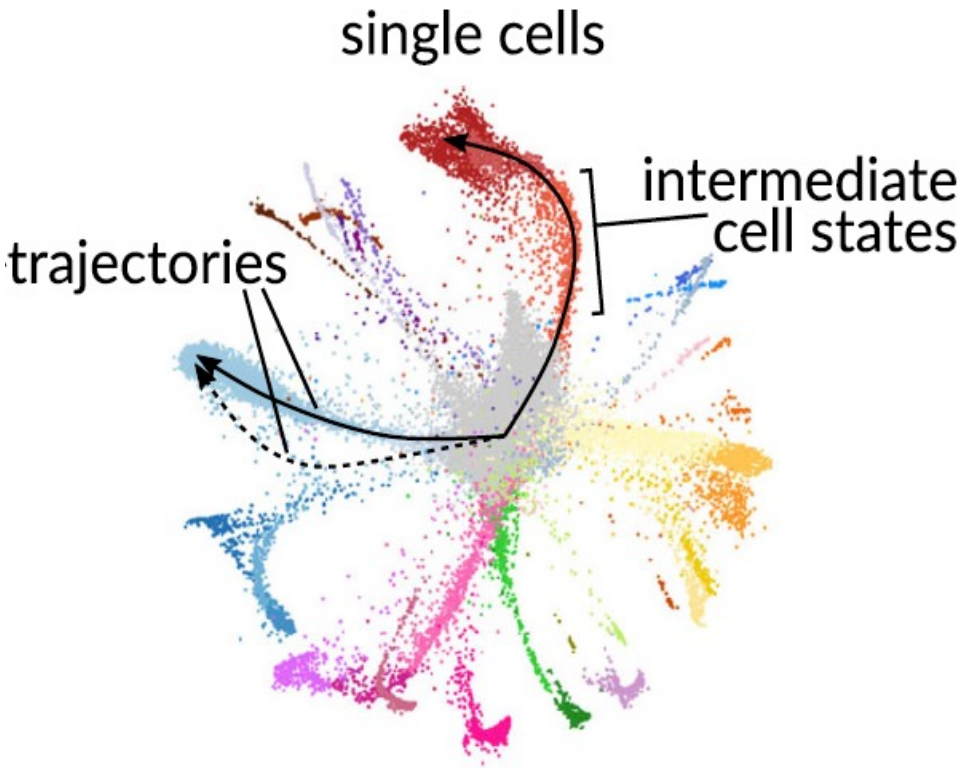
Department of Data Mining
Silesian University of Technology
Poland

Single-cel RNA sequencing

Single-cell sequencing is a resolution revolution in transcriptomics and genomics.



Lun et al. 2017

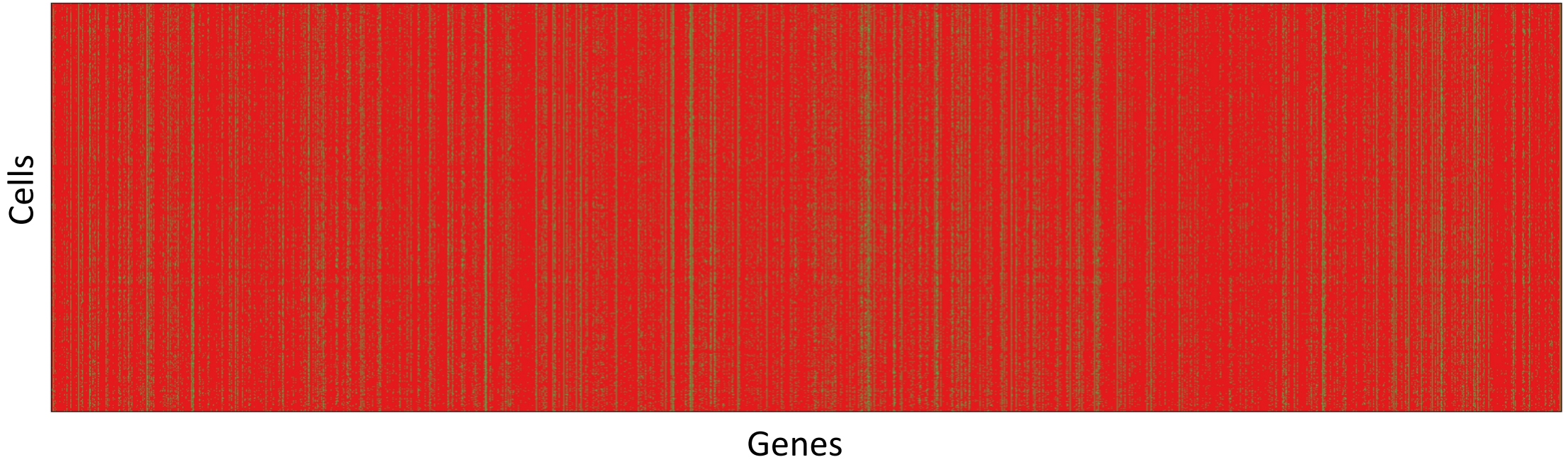


Lahnemann et al., 2020

“Zeros” in single-cell RNA sequencing data

- Some genes could be highly expressed in one cell but not expressed in another.

- For some genes, we see more cells with expression (green) than not (red), but overall, there are much more zeros in the data.
- Only 50 genes have non-zero values in all cells.

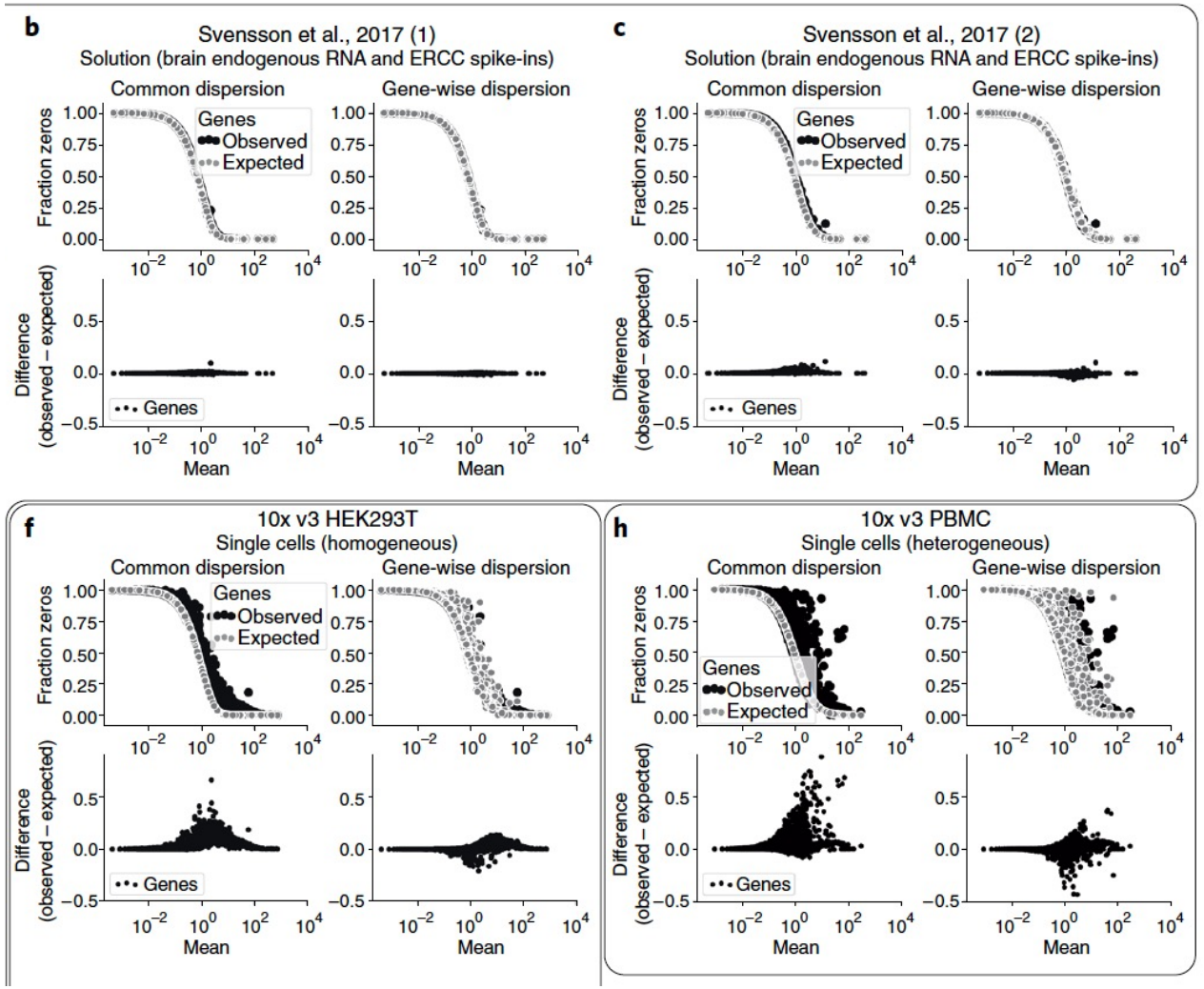


“Droplet scRNA-seq is not zero-inflated”

- The number of zeros observed is consistent with the theoretical models.
- Additional zero values in the data are likely due to biological variability.

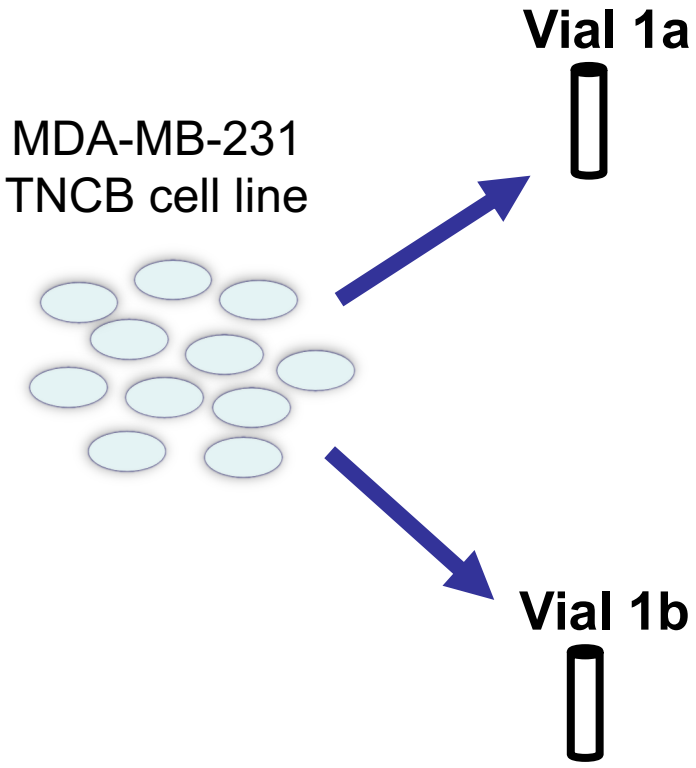
- The most important factor that determines the number of zeros in the scRNA-seq data is the depth of sequencing (total UMI) per cell.

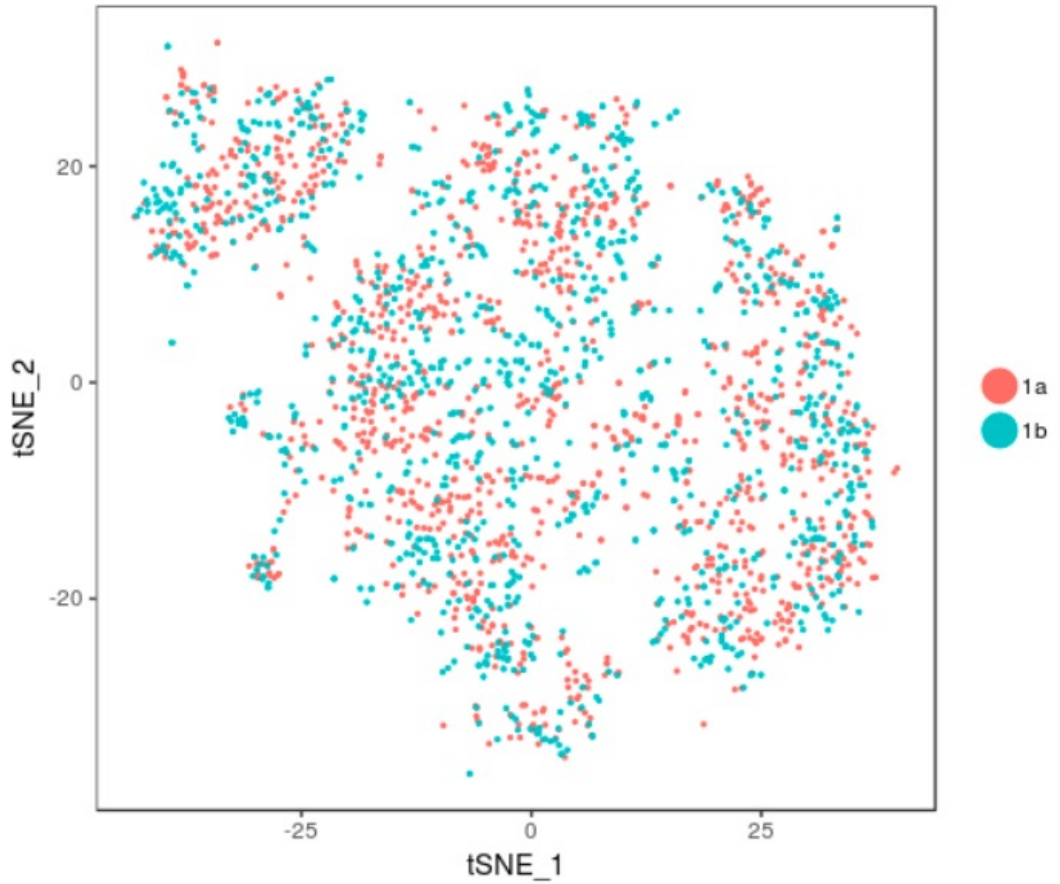
But other important factors could also exist!



Svensson et al., 2020

Data

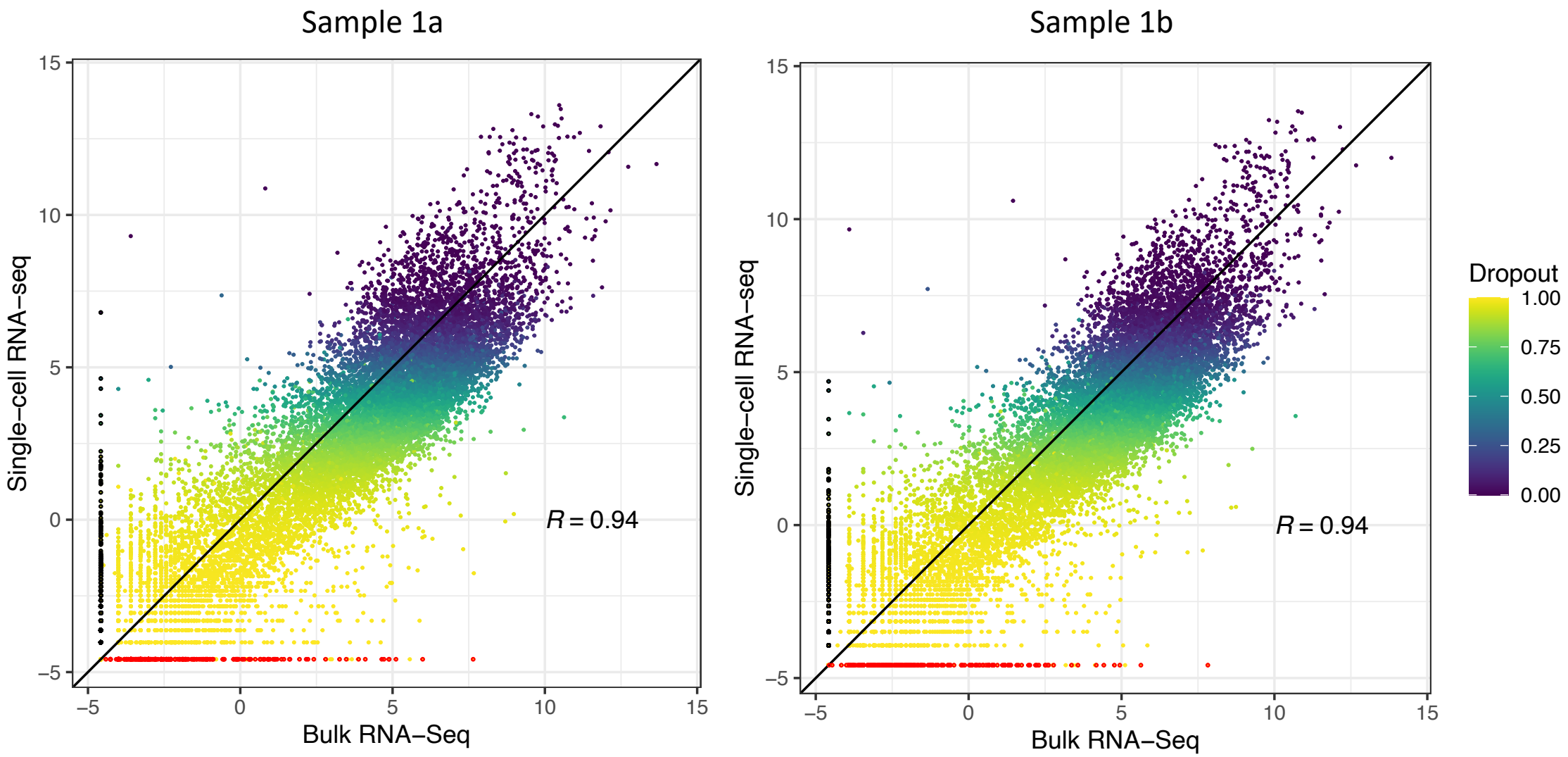




Platform	Type of data
scRNAseq	Single-cell mRNA expression
bRNAseq	Bulk mRNA expression
ATACseq	Chromatin accessibility
DNAm	DNA methylation

- 2 biological replicates (1a and 1b).
- DNA and RNA for all experiments was extracted on the same day.

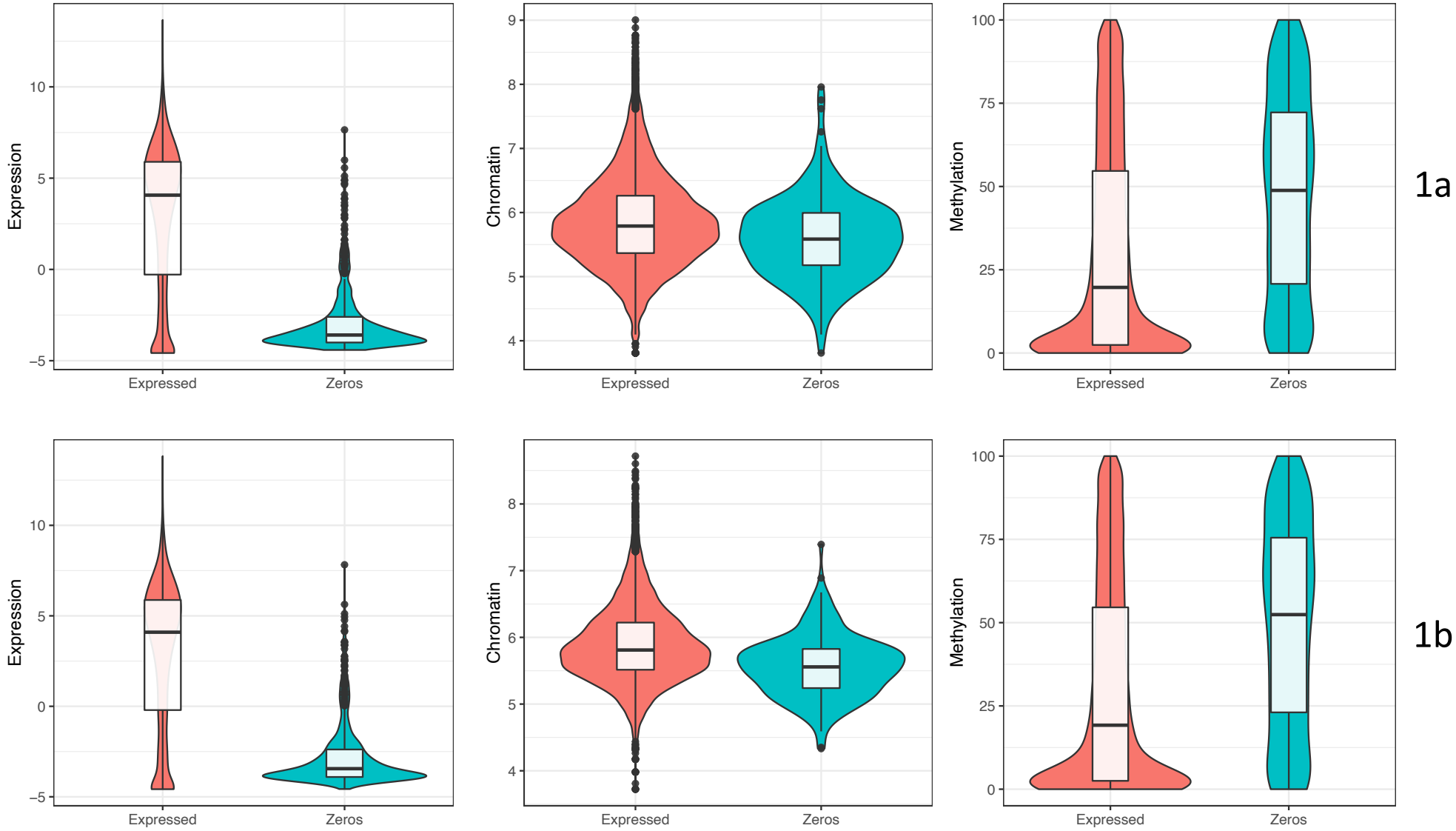
Single-cell vs bulk RNA-seq



The technological differences are greater than the biological differences between samples.

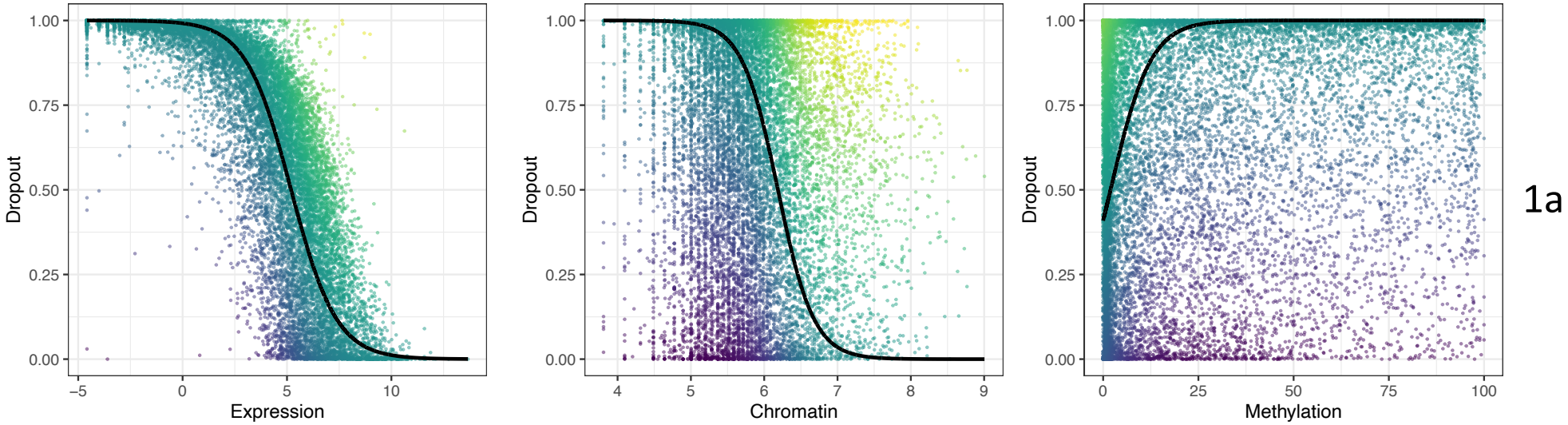
“Zeros” from scRNA-seq in other modalities

In bRNAseq and ATACseq, genes with zeros have lower signal, while in DNAm genes with zeros have higher signal.

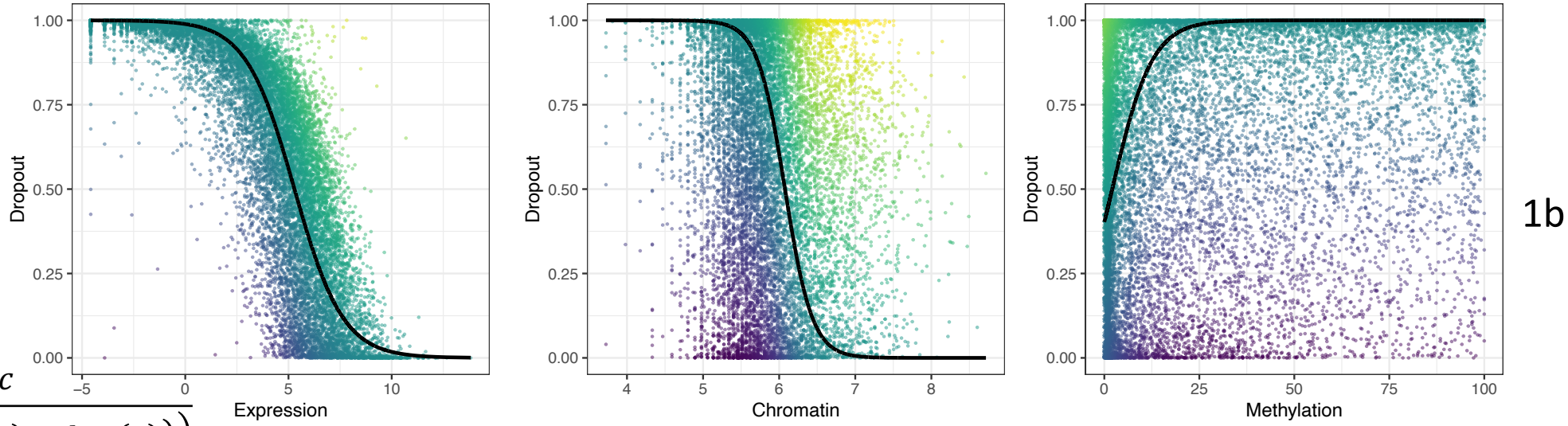


Dropout rate vs signal from 3 different modalities

Higher expression and chromatin openness, and lower methylation level relate to lower dropout rate.



1a



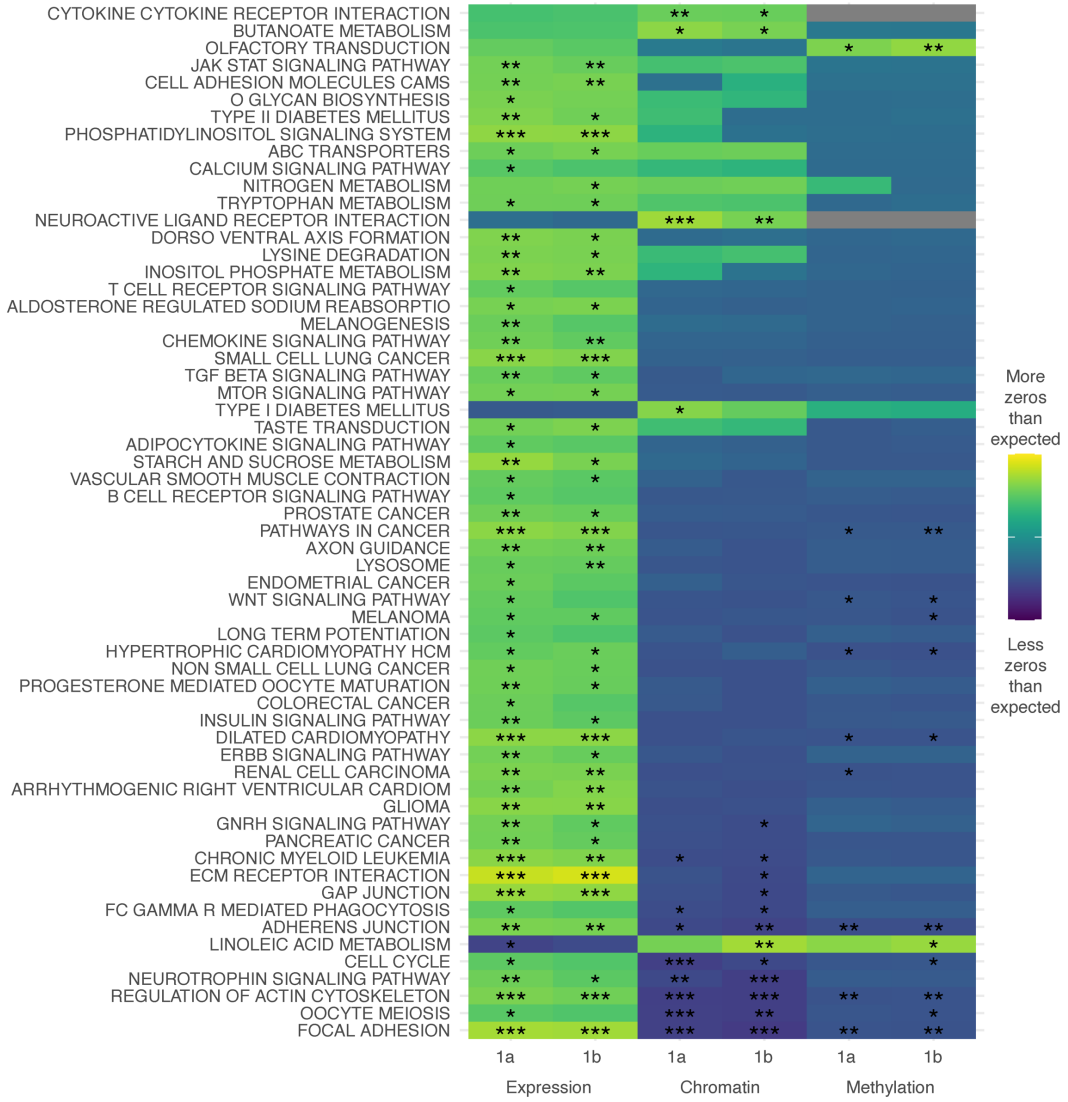
1b

5PL model:

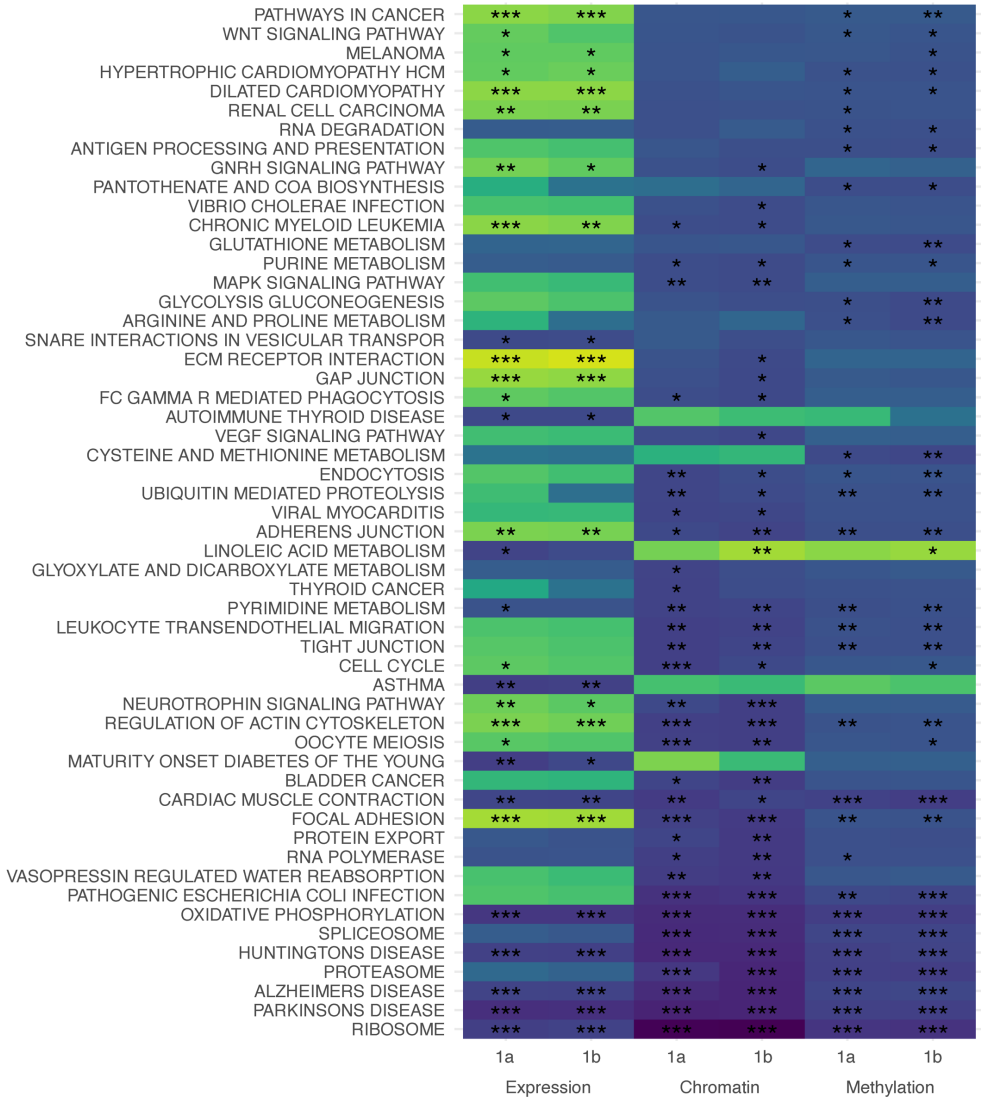
$$f(x) = c + \frac{d - c}{1 + exp\left(b\left(log(x) - log(e)\right)\right)}$$

Functional analysis of 5PL model residuals

Paths with genes with more zeros than expected



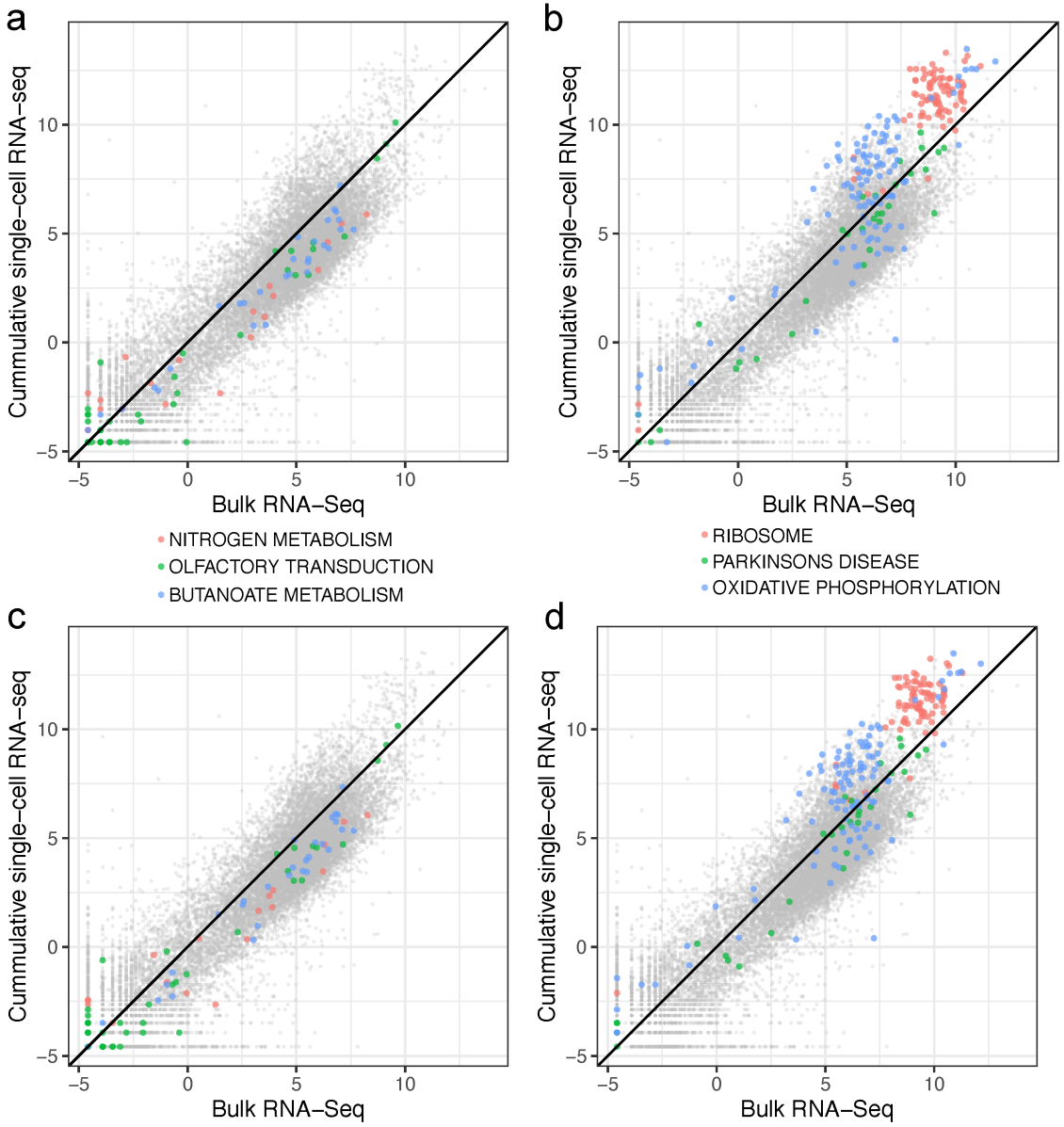
Paths with genes with fewer zeros than expected



Functional analysis of 5PL model residuals

- Genes included in top3 up- and down-regulated KEGG pathways.

- The mechanism of reduced expression of some genes on single-cell level might be different to what was expected: some groups of genes, like ribosomal genes, have higher expression on a single-cell level than bulk level, so low expression genes could not be captured due to limited total number of sequencing read counts that are measured in one experiment.



Potential technical sources of “zeros”

- Biased reads coverage at 3’ end of transcripts (TIN score).
- Difference in %GC content of the gene sequence.
- Length of the gene.
- No. of transcripts per gene.
- Lower mappability in a gene region.

Odds ratio with 95% CI

Factor	Sample 1a		Sample 1b	
	Full	Reduced	Full	Reduced
GC	1.011 (0.988;1.034)	-	1.014 (0.991;1.037)	-
Gene_length	0.999 (0.998;1)	-	0.999 (0.998;1)	-
N_transcripts	1.029 (0.979;1.078)	1.082 (1.046;1.119)*	1.067 (1.009;1.124)	1.11 (1.071;1.149)*
Mappability	0.978 (0.906;1.05)	1.018 (1.013;1.022)*	1.006 (0.972;1.04)*	1.016 (1.012;1.021)*
TIN	1.039 (1.025;1.05)*	1.011 (1.006;1.017)*	1.062 (1.046;1.078)*	1.017 (1.012;1.023)*
Expression	1.304 (1.168;1.440)*	1.325 (1.269;1.382)*	1.107 (0.978;1.237)	1.243 (1.186;1.3)*
Chromatin	1.119 (0.866;1.371)	-	1.554 (1.224;1.884)*	-
Methylation	1 (0.994;1.006)	-	1.001 (0.996;1.008)	-

A low gene expression level in the analysed sample is the main contributor for zeros, **but not only.**

Conclusions

- The differences between biological / technical replicates are minimal on all platforms.
- The differences in gene expression between platforms are greater than when comparing replicates on the same platform.

- There are some potential biological factors that may indicate why some genes have a higher dropout rate than others.

- There are various technical factors that can potentially affect the dropout rate, the most important of which are read coverage across the transcript and mappability in the gene region.

Acknowledgments

POLSL:

- Joanna Polańska
- Joanna Żyła
- Anna Papież
- Hanna Słowik

Yale CANCER CENTER

A Comprehensive Cancer Center Designated by the National Cancer Institute

Yale:

- Jun Zhao
- Rihao Qu
- Xiaotong Li
- Yuval Kluger
- Christos Hatzis