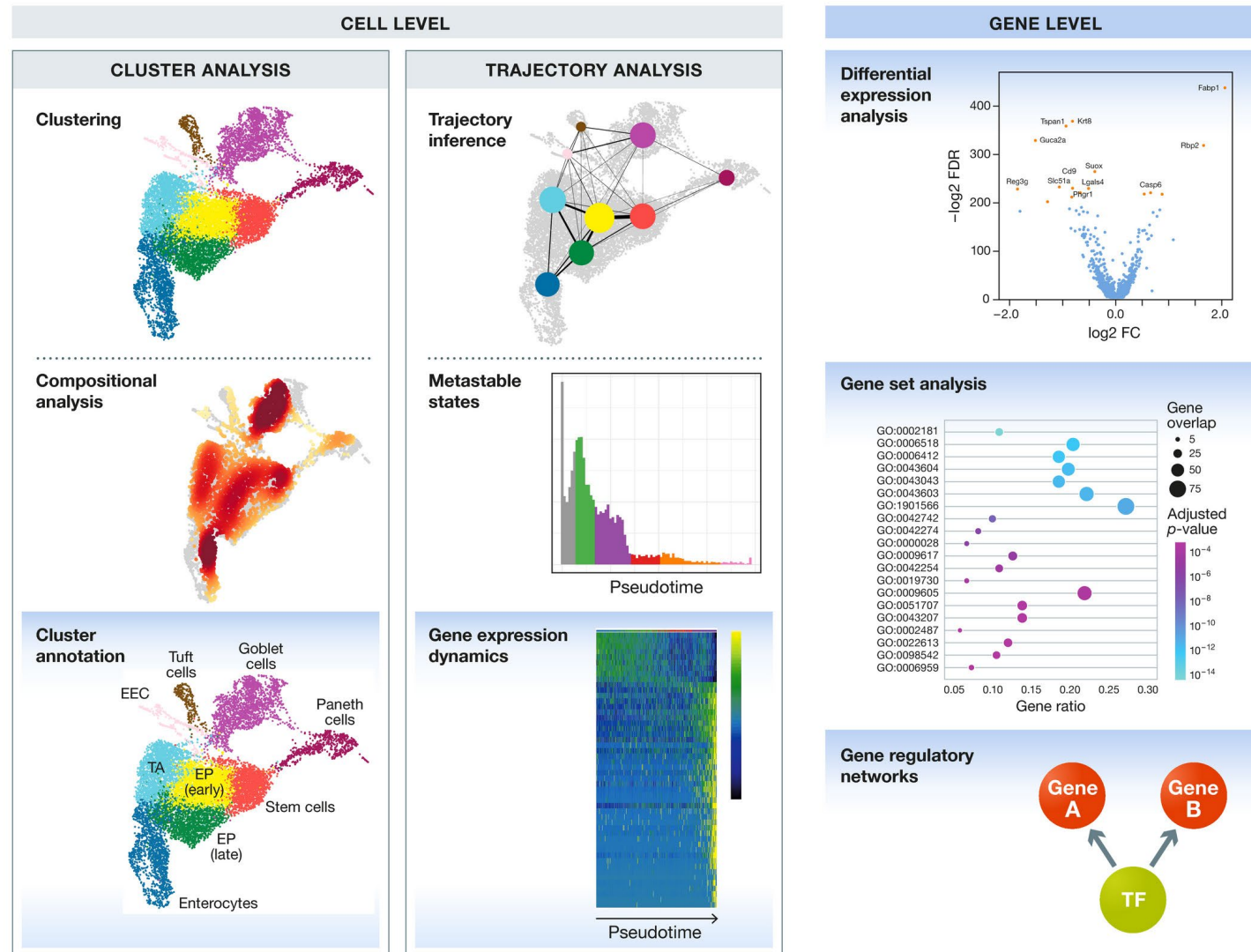


Differential expression (DE) analysis

Ahmed Mahfouz

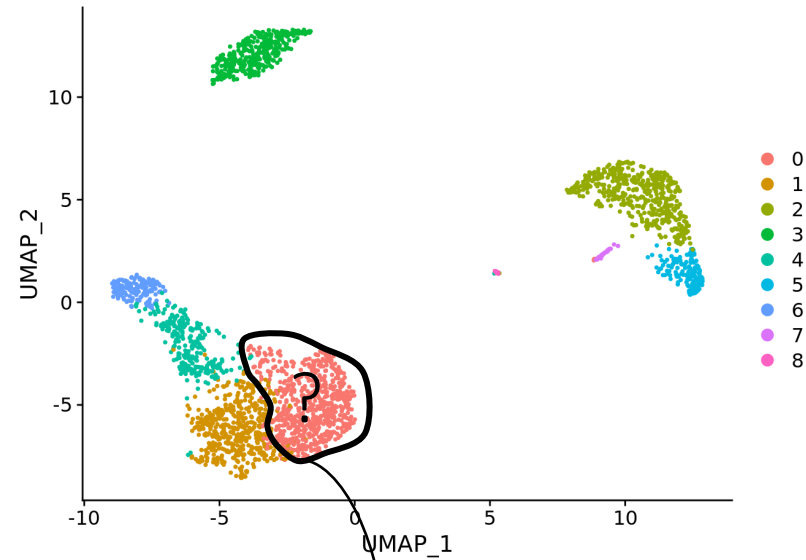
Department of Human Genetics, Leiden University Medical Center
Pattern Recognition and Bioinformatics, TU Delft

Downstream analysis of scRNA-seq data



DE for cluster annotation

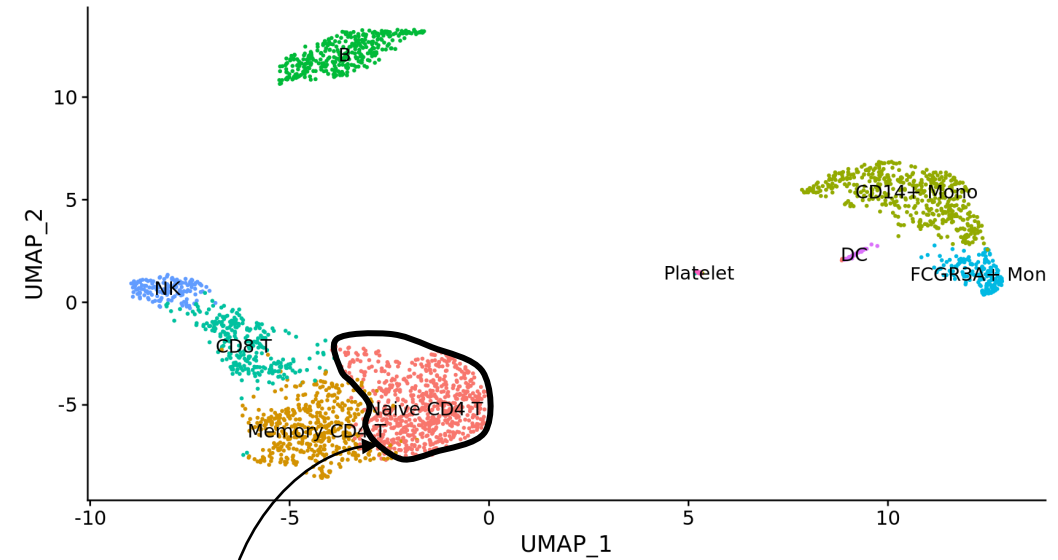
Unannotated clusters



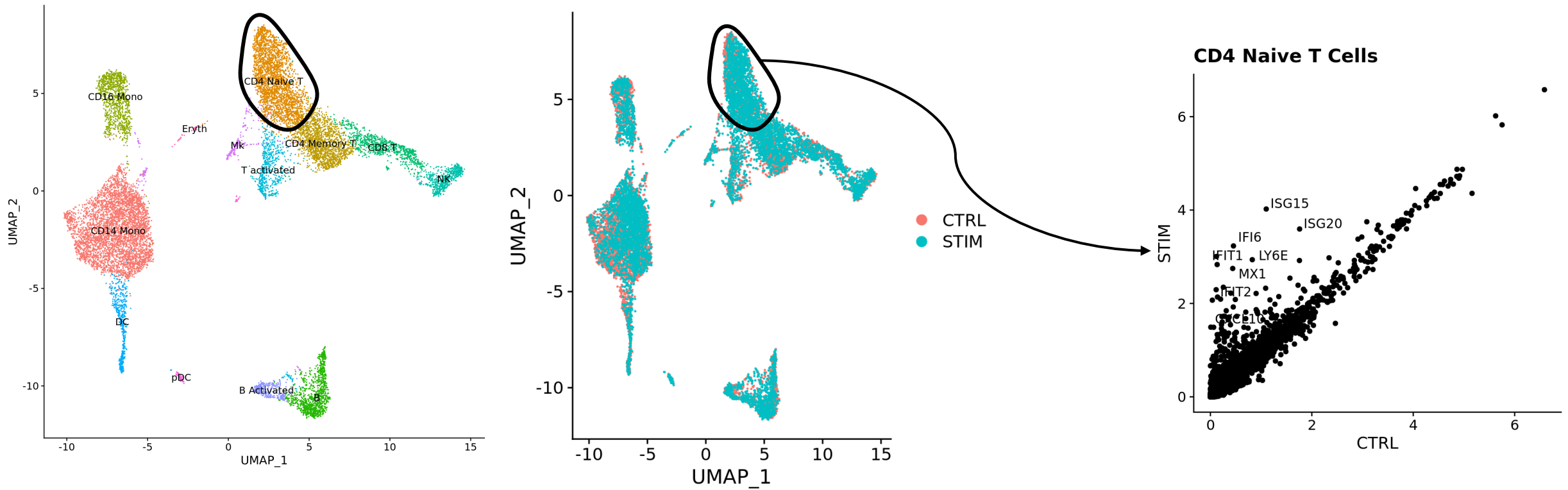
Compare *Cluster 0*
to all other cells

IL7R
CCR7

Annotated clusters



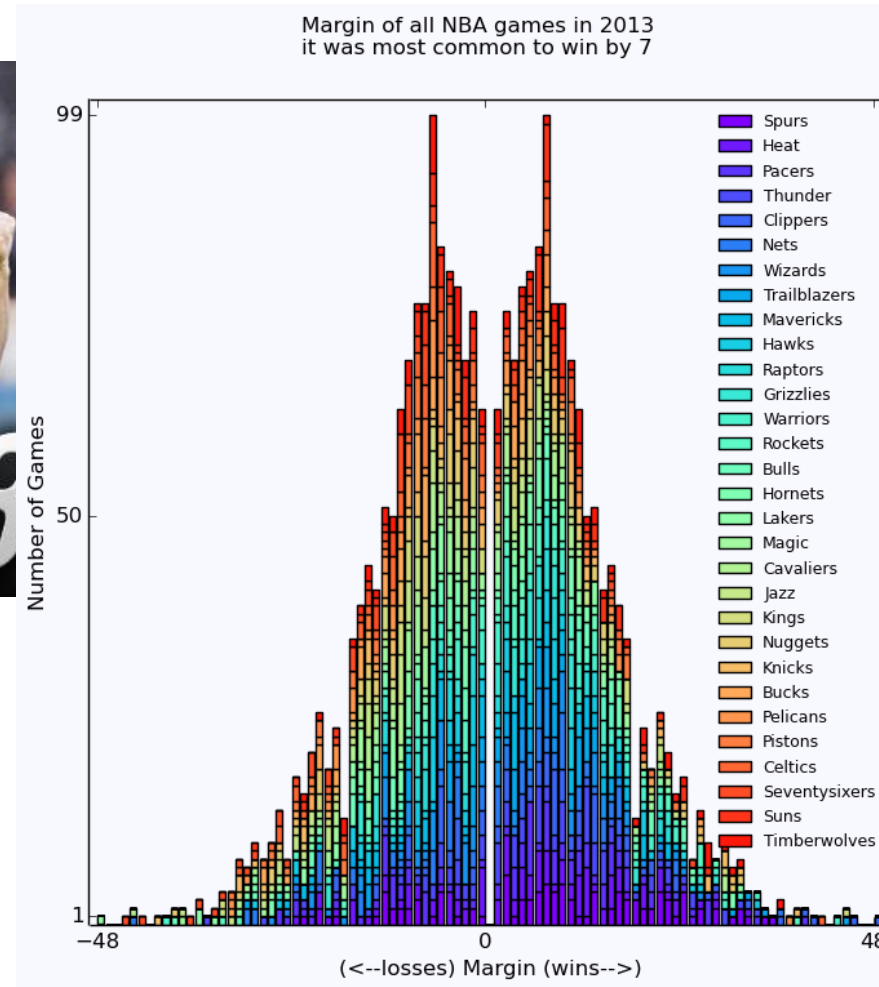
DE for comparing conditions



Outline

- Different methods for differential expression (DE) analysis for scRNA-seq data
- Single-cell DE in practice
- Working with integrated data
- Power analysis for scRNA-seq DE

Is this a large difference?



Raptors

und, Game 7 - Raptors won series 4-3

84

89

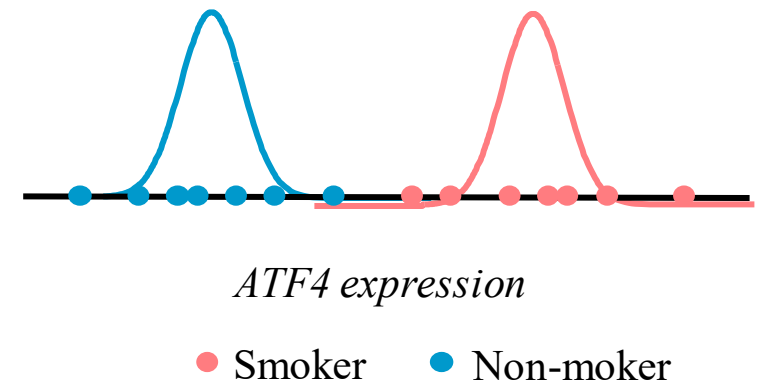


Differential gene expression

- We need to decide for every gene, if the difference in expression observed between 2 groups is significant
- Significant = greater than it would be expected just due to natural random variation

Example: Is there a difference in *ATF4* expression between smokers and non-smokers?

1. Define hypothesis
 - Null hypothesis (H_0): there is no difference in expression
 - Alternative hypothesis (H_1): there is a difference in expression
2. Measure some data
 - Expression of *ATF4*
3. Test your hypothesis
 - Compare *ATF4* expression between smokers and non-smokers

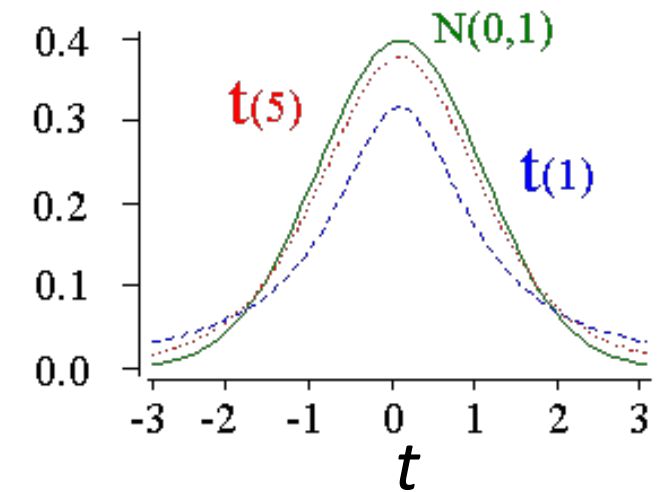
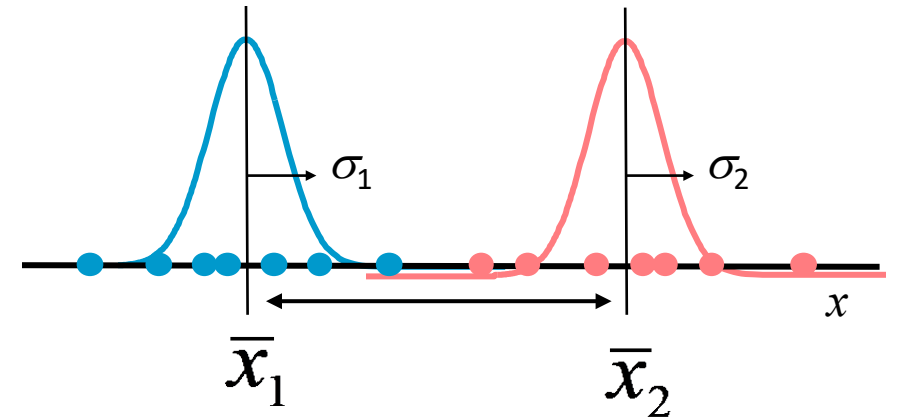


Model the data

- Model the data distribution (e.g. normal)
- Use a statistic to assess the difference (e.g. t-test)

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference in group means}}{\text{variability in groups}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(J_1 - 1)S_1^2 + (J_2 - 1)S_2^2}{J_1 + J_2 - 2} \left(\frac{1}{J_1} + \frac{1}{J_2} \right)}}$$



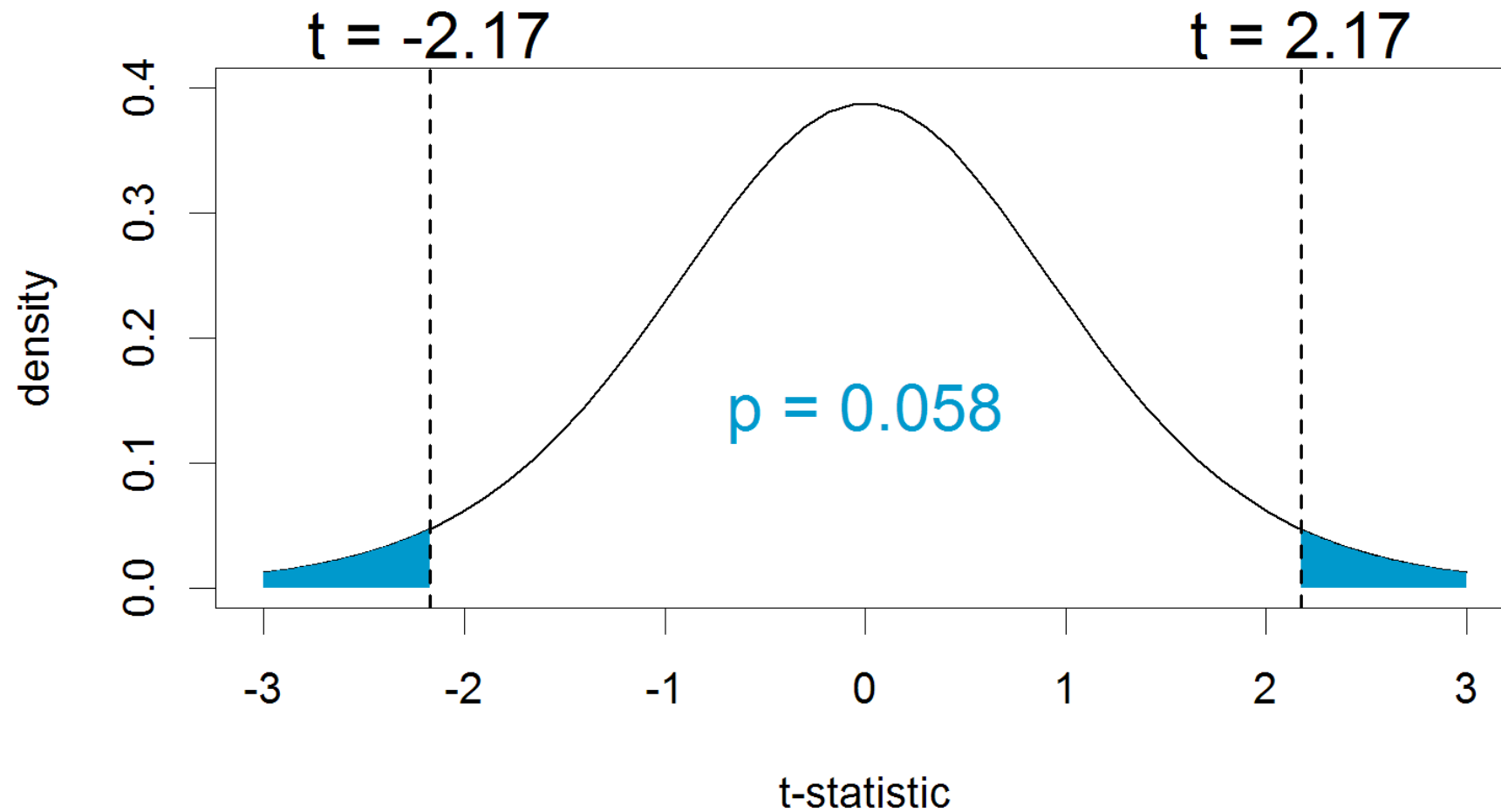
t follows a Student t -distribution with $J-1$ degrees of freedom (DOF)

P-value

Two-sided test

$t = 2.17$

DOF = 9

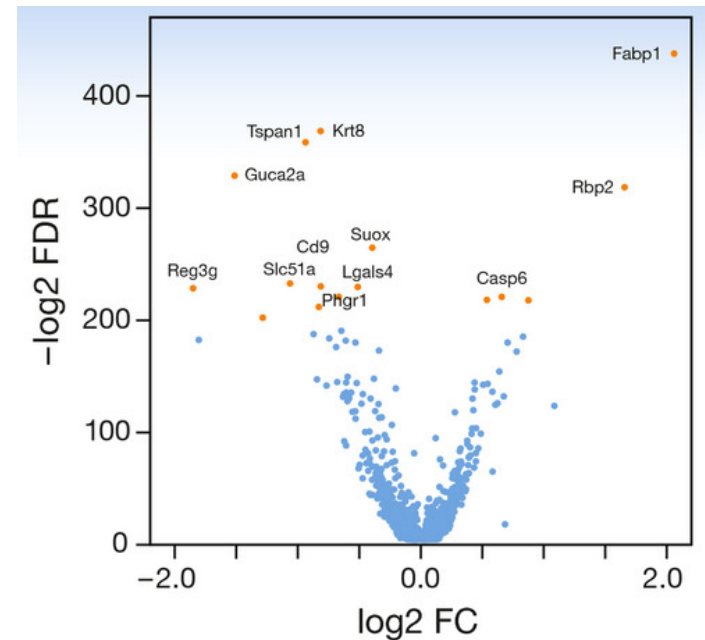


Effect size

- It is also wise to consider the effect size and not only the p-value
 - A very low p-value with a very low effect size is meaningless
- Effect size measure depends on the statistical test used
- E.g. in a t-test, the mean is compared between 2 groups (effect size = difference in the mean)
- Often represented as log fold-change (LFC)

$$lfc = \log_2 \left(\frac{\bar{X}_1}{\bar{X}_2} \right)$$

Volcano plot



Luecken and Theis (MSB 2019)

How do we model the data?

- Find an appropriate model (appropriate = fits the data better)
- Use a non-parametric test (no model assumptions)

Non-parametric tests

- Forget about modeling the data, let's use a non-parametric test.
- No assumption that expression values follow any particular distribution
- Expression values are (generally) converted to ranks and test whether the distribution of ranks for one group are significantly different from the distribution of ranks for the other group.
- Assumption: distributions have the same shape in both groups

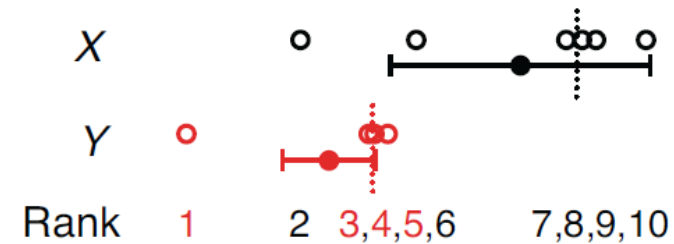
Wilcoxon rank-sum test aka Mann-Whitney U test

- H_0 : median₁ = median₂
- Start by ranking all values
- Calculate the test statistic:

$$U = W - \frac{n_Y(n_Y+1)}{2}$$

↑
sum of ranks in the
smaller-sized sample

↙
The lowest possible rank in the
sample with the lower ranks



$$\begin{aligned} W &= 1 + 3 + 4 + 5 = 13 \\ U' &= W - n_Y(n_Y + 1)/2 \\ &= 13 - 10 \\ &= 3 \end{aligned}$$

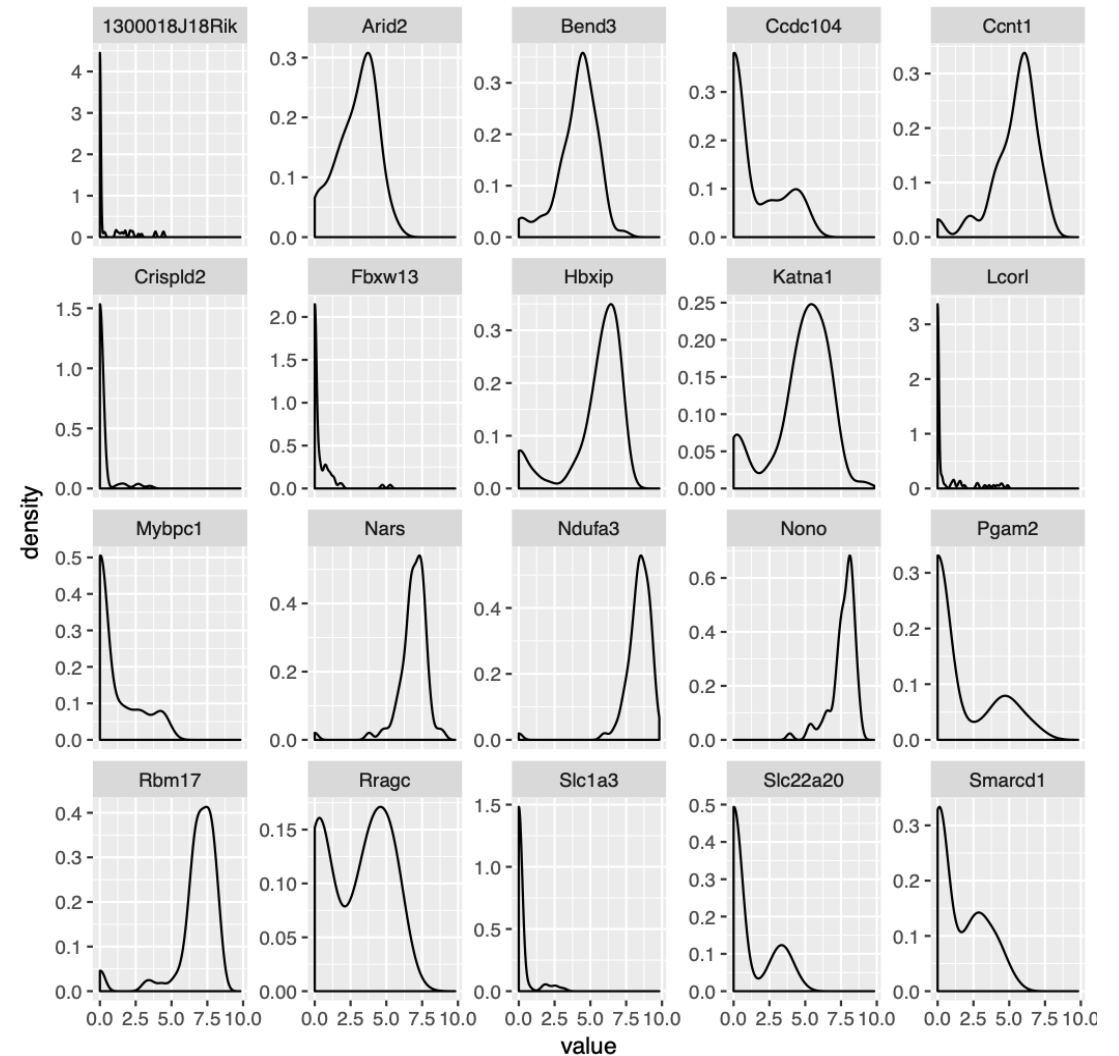
For cases in which both samples are larger than 10, the distribution of U is approximately normal

That must be the solution to everything?

- Not really...
- Wilcoxon rank sum test is not as powerful as parametric tests, i.e. it requires more data points to detect the same effects
- Might fail to deal with a large number of tied values, such as the case for zeros in single-cell RNA-seq expression data

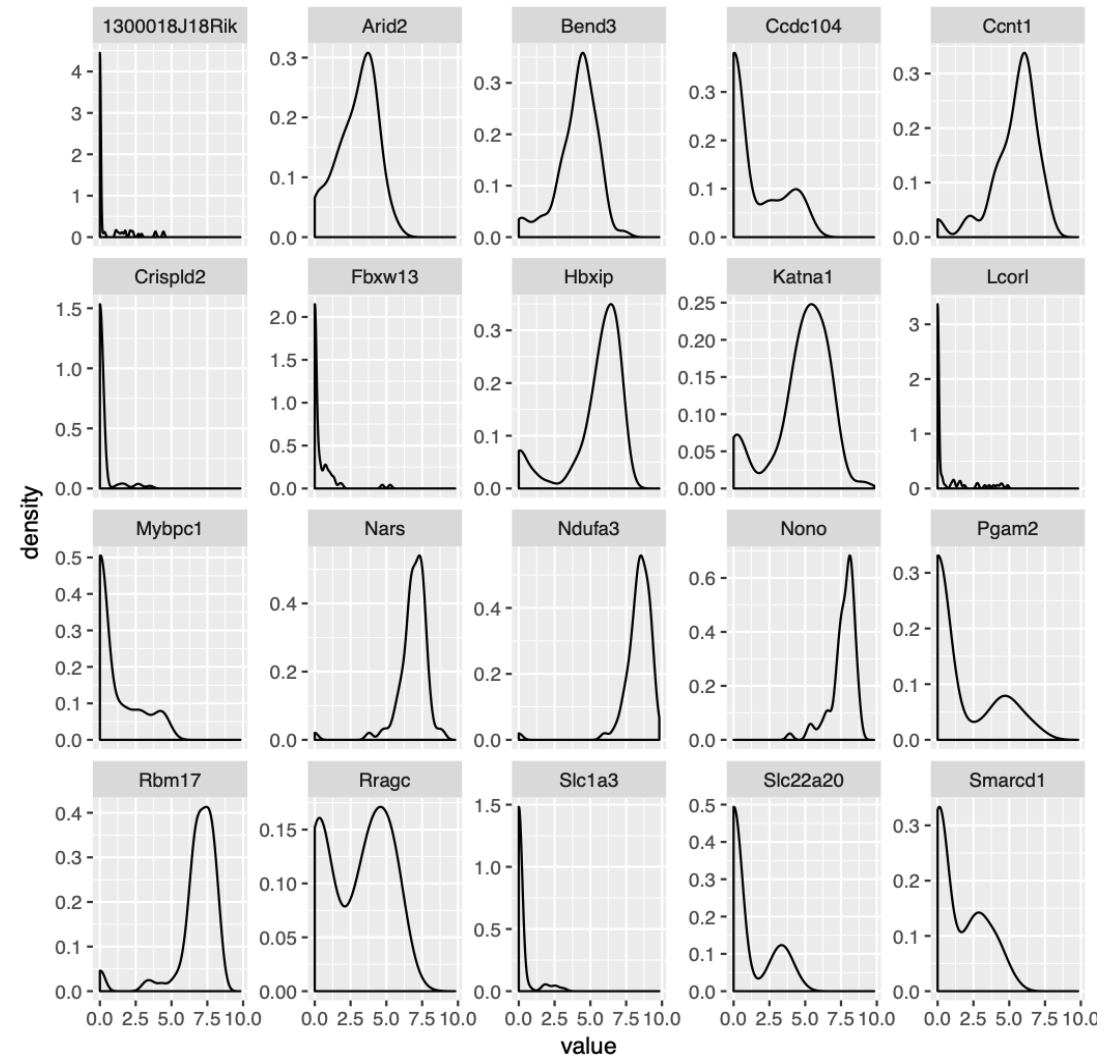
How can we model scRNA-seq data?

- Amplification bias
- Drop-out rates
- Transcriptional bursting
- Background noise
- Bias due to cell-cycle and cell size
- Often clear batch effects
- ...



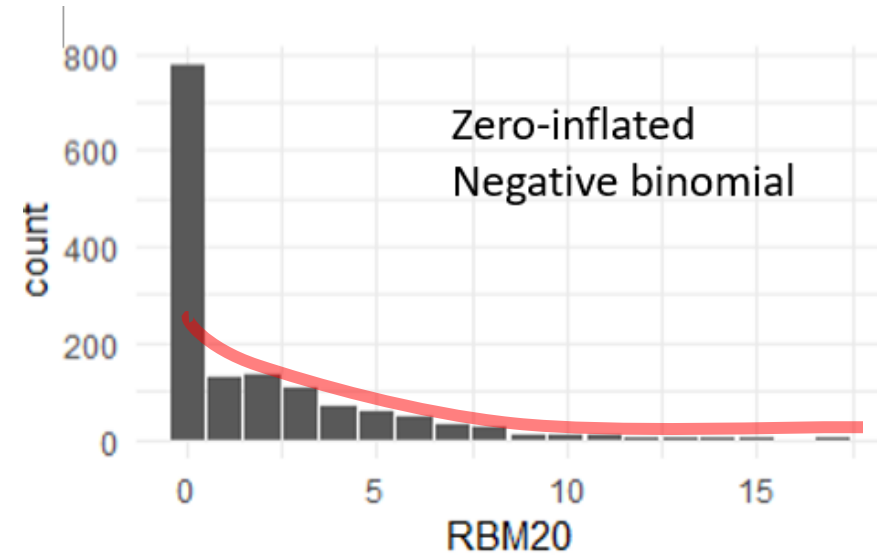
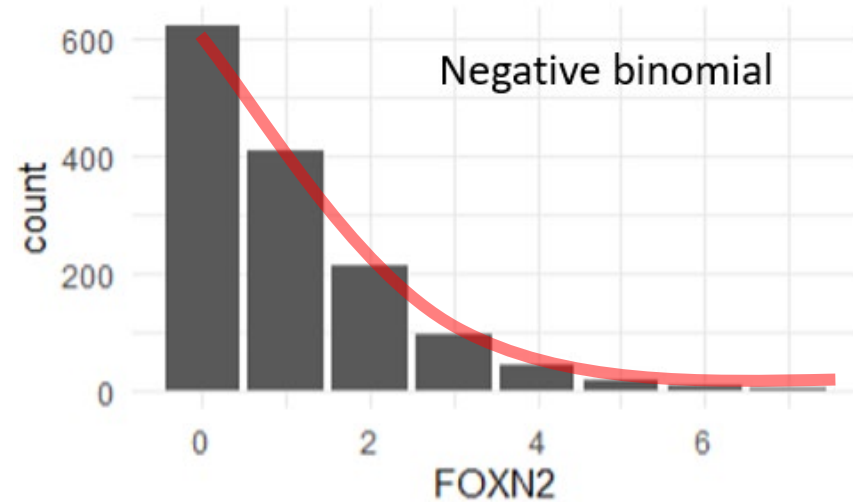
Which distribution would you use to model scRNA-seq data?

- Binomial
- Negative binomial
- Zero-inflated negative binomial
- Poisson
- ...



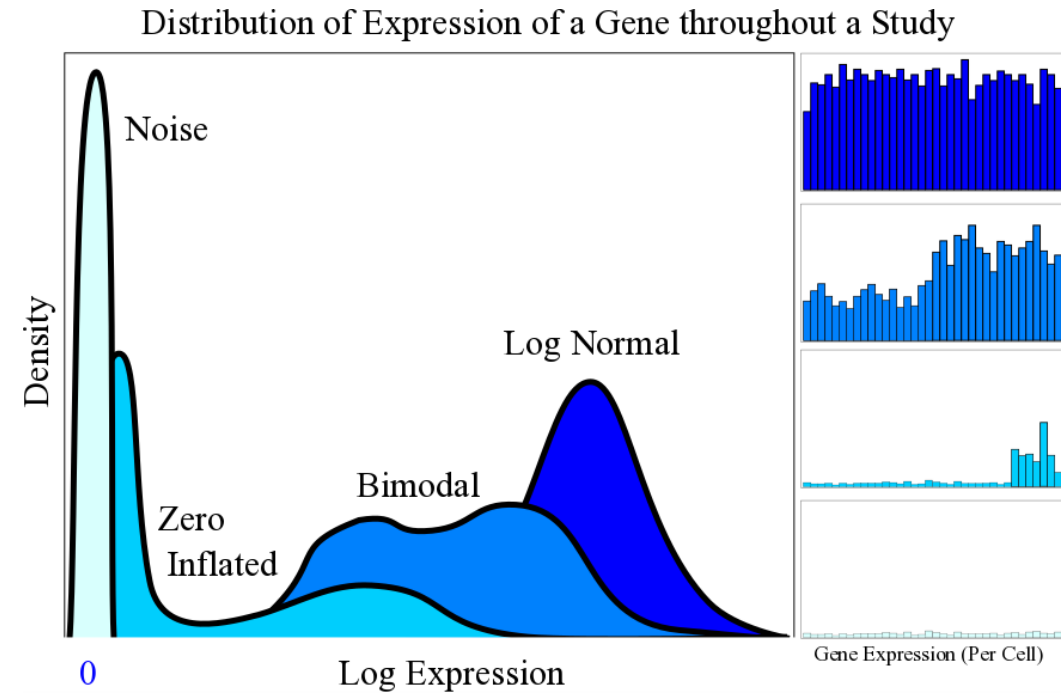
Is single-cell data zero-inflated?

- “More zeros than expected”



MAST (GLM)

- MAST uses a hurdle model (a two-part generalized linear model)
- Part 1: models the discrete expression rate of each gene across cells (is the gene expressed or not?) -> *logistic regression*
- Part 2: models the continuous expression level (conditional on the gene being expressed) -> *linear Gaussian model*
- DE is determined using a likelihood ratio test



Is single-cell data zero-inflated?

Correspondence | [Published: 14 January 2020](#)

Droplet scRNA-seq is not zero-inflated

[Valentine Svensson](#) 

[Nature Biotechnology](#) **38**, 147–150 (2020) | [Cite this article](#)

11k Accesses | **80** Citations | **89** Altmetric | [Metrics](#)

Matters Arising | [Published: 01 February 2021](#)

UMI or not UMI, that is the question for scRNA-seq zero-inflation

[Yingying Cao](#), [Simo Kitanovski](#), [Ralf Küppers](#) & [Daniel Hoffmann](#) 

[Nature Biotechnology](#) **39**, 158–159 (2021) | [Cite this article](#)

Research | [Open Access](#) | [Published: 27 July 2020](#)


Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics

[Kwangbom Choi](#), [Yang Chen](#), [Daniel A. Skelly](#) & [Gary A. Churchill](#) 

[Genome Biology](#) **21**, Article number: 183 (2020) | [Cite this article](#)

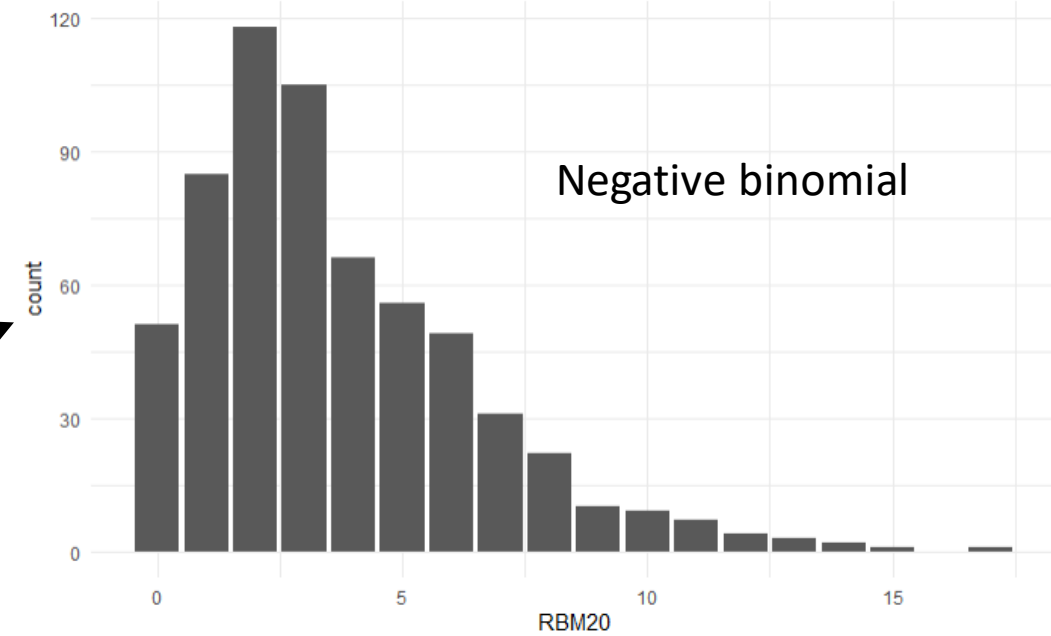
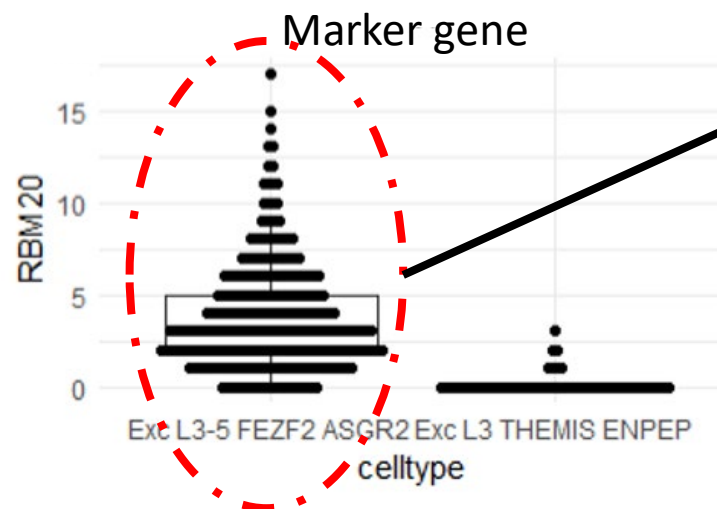
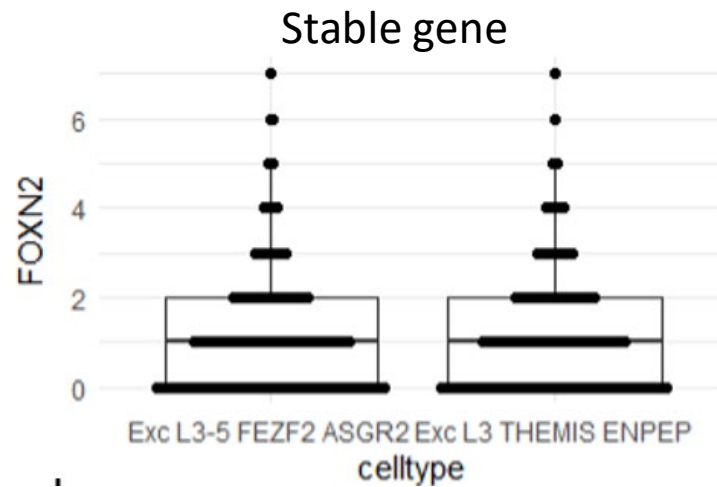
Review | [Open Access](#) | [Published: 21 January 2022](#)

Statistics or biology: the zero-inflation controversy about scRNA-seq data

[Ruochen Jiang](#), [Tianyi Sun](#), [Dongyuan Song](#) & [Jingyi Jessica Li](#) 

[Genome Biology](#) **23**, Article number: 31 (2022) | [Cite this article](#)

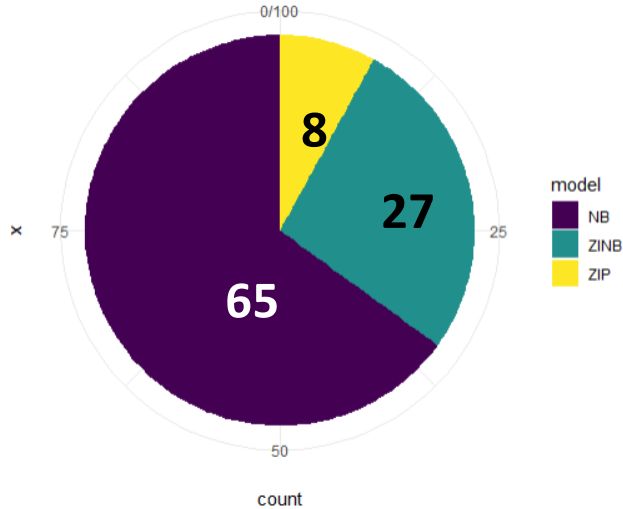
Is single-cell data zero-inflated?



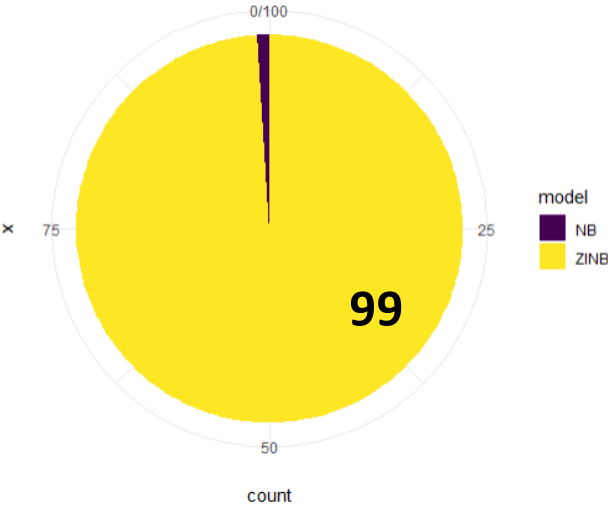
Is single-cell data zero-inflated?

Stable genes

10x (UMI)



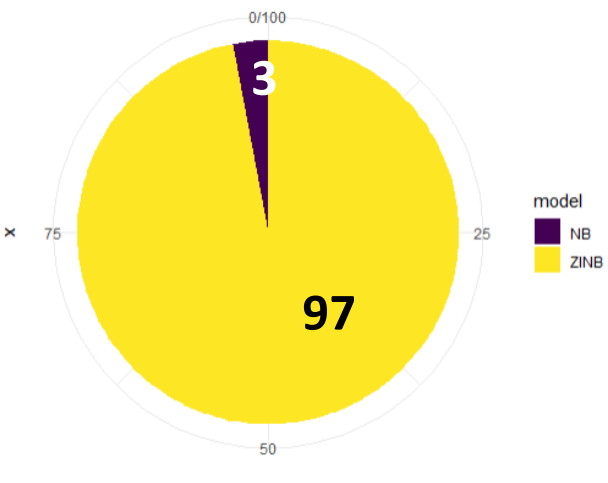
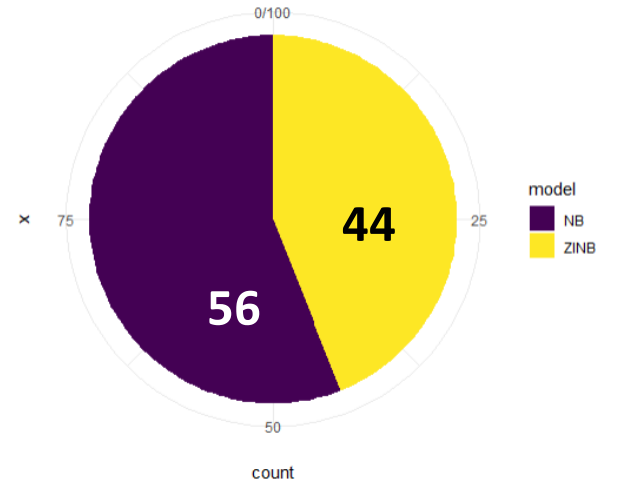
Marker genes



	Non-zero-inflated	Zero-inflated
Stable	65	35
Marker	1	99

$P = 3.02 \times 10^{-25}$
 $\log OR = 5.18$

Smart-seq (reads)



	Non-zero-inflated	Zero-inflated
Stable	56	44
Marker	3	97

$P = 5.46 \times 10^{-18}$
 $\log OR = 3.69$

Is single-cell data zero-inflated?

Some genes are zero-inflated, some are not.

The main driver of zero-inflation is biological heterogeneity.

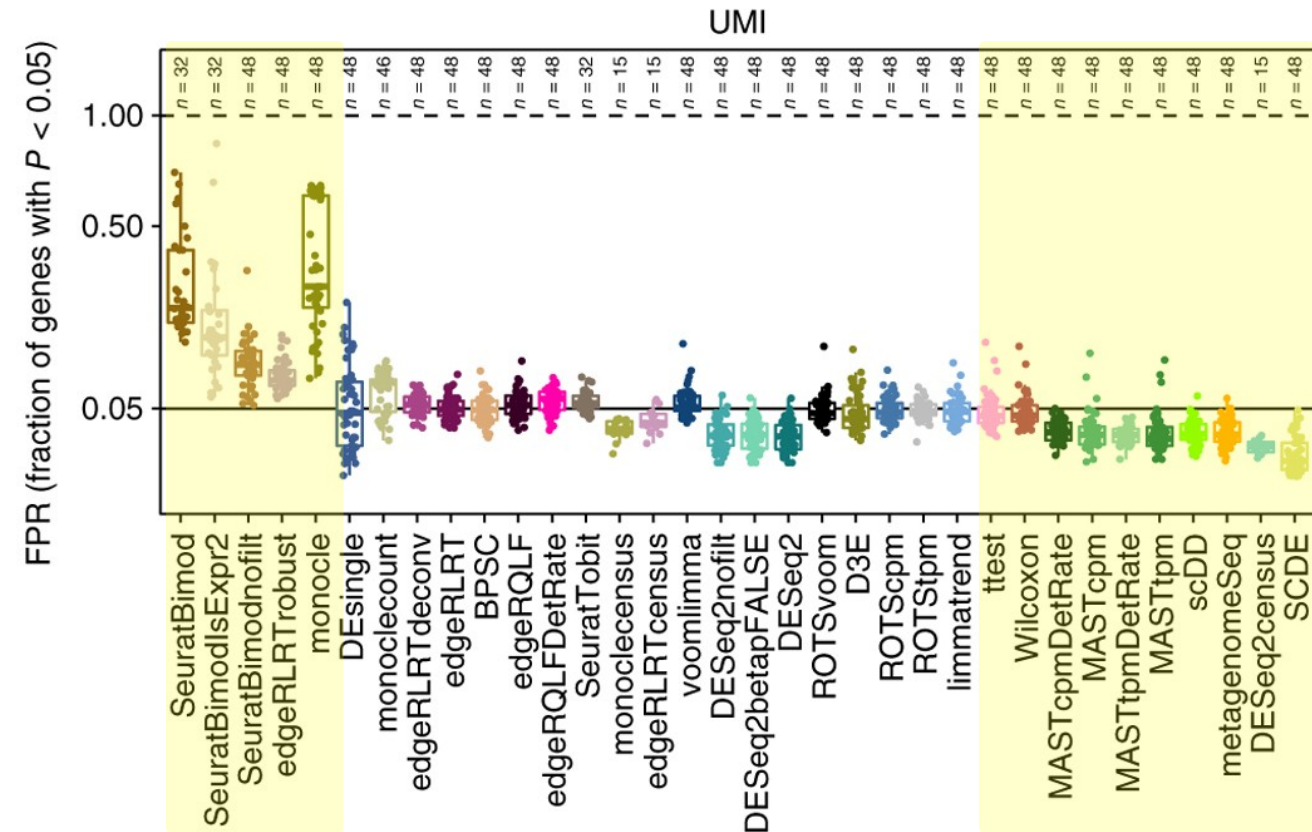
As such, the presence of a zero is mainly dictated by biology.

Which model is better for DE analysis?

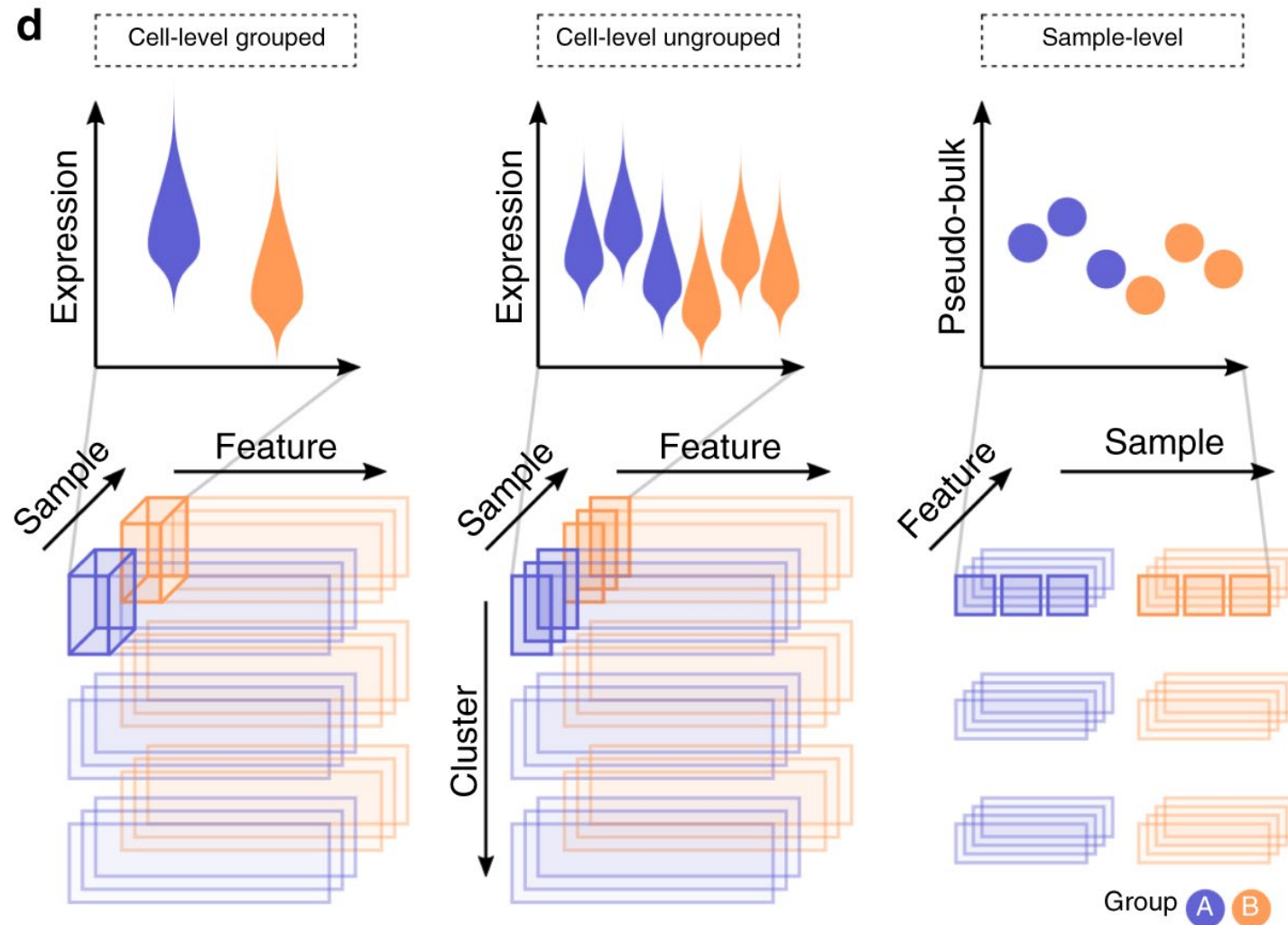
- Biological ground truth is difficult to define!
- Simulations: possible, but then results depend on the model used
- Rely on data with some known effects
 - E.g. Matched bulk RNA-seq dataset in the same purified cell type, exposed to the same perturbation under identical experimental conditions

Comparing cell types (to identify markers)

- Overall, MAST, Wilcoxon, t-test outperformed other methods
- bulk RNA-seq analysis methods do not perform worse than scRNA-seq-specific methods
- Did not consider multi-sample setups



Comparing conditions



Benchmarking based on multi-sample setups

Using simulated data

Article | [Open Access](#) | [Published: 30 November 2020](#)

***muscat* detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data**

[Helena L. Crowell](#), [Charlotte Soneson](#), [Pierre-Luc Germain](#), [Daniela Calini](#), [Ludovic Collin](#), [Catarina Raposo](#), [Dheeraj Malhotra](#) & [Mark D. Robinson](#) 

[Nature Communications](#) **11**, Article number: 6077 (2020) | [Cite this article](#)

7161 Accesses | **23** Citations | **48** Altmetric | [Metrics](#)

Article | [Open Access](#) | [Published: 02 February 2021](#)

A practical solution to pseudoreplication bias in single-cell studies

[Kip D. Zimmerman](#) , [Mark A. Espeland](#) & [Carl D. Langefeld](#) 

[Nature Communications](#) **12**, Article number: 738 (2021) | [Cite this article](#)

5078 Accesses | **4** Citations | **6** Altmetric | [Metrics](#)

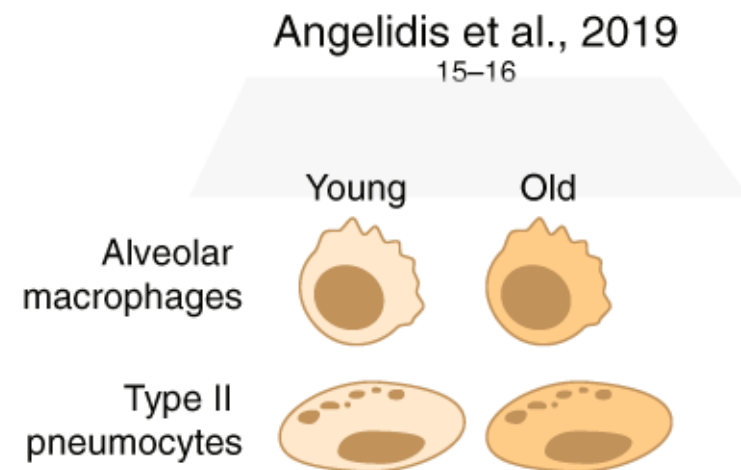
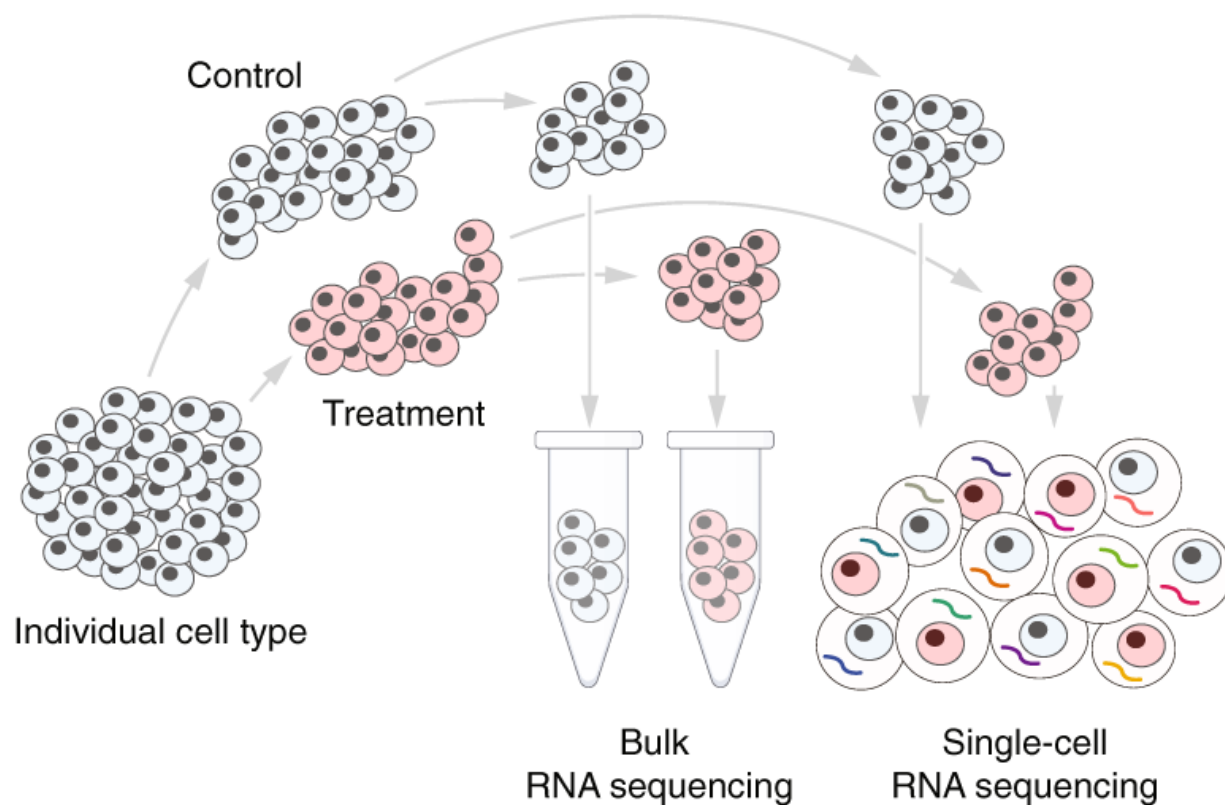
Recommends pseudobulk approaches

Recommends generalized linear
mixed models

They use different simulation models!

Benchmarking based on multi-sample setups

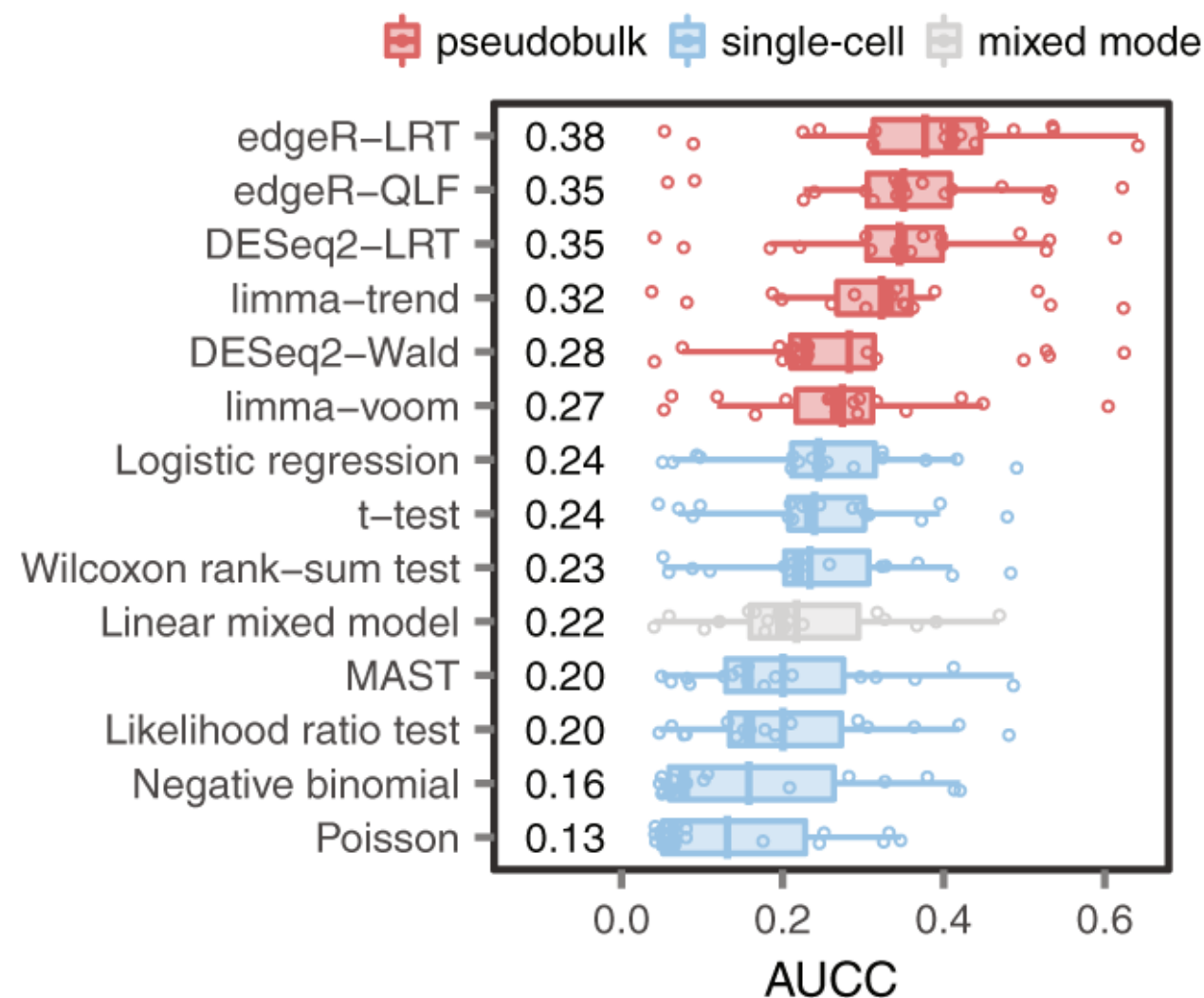
Using scRNA-seq + bulk data



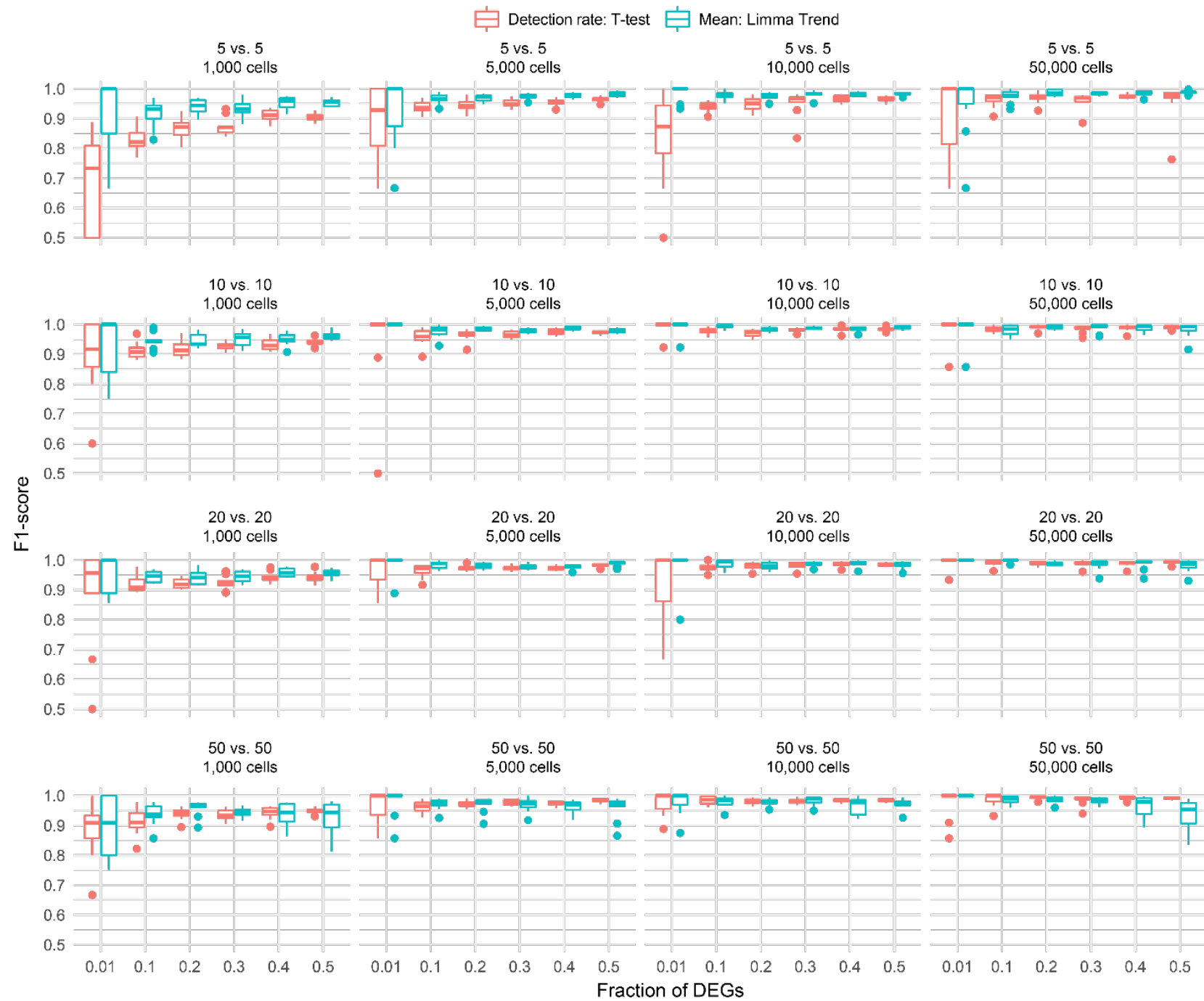
Benchmarking based on multi-sample setups

Using scRNA-seq + bulk data

- Pseudobulk methods are better
- Accounting for variation between biological replicates determines the performance of single-cell DE methods



Again here,
binarizaion
works!



Single-cell DE in practice

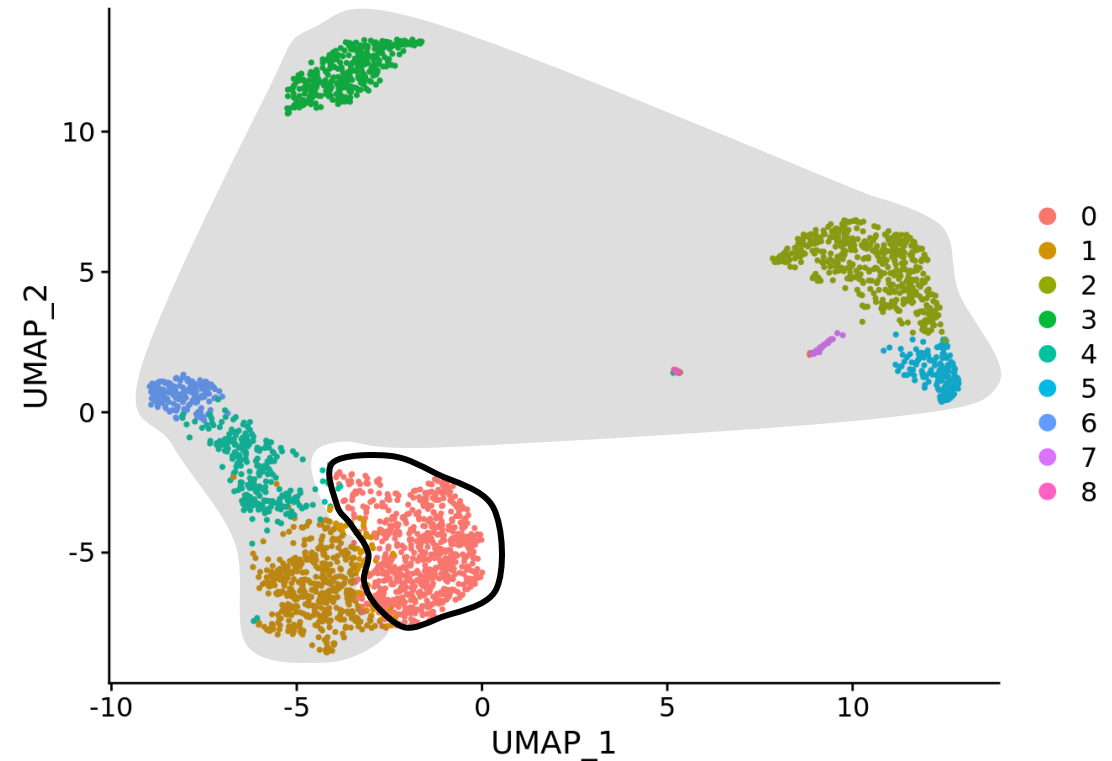
Single-cell DE in practice

Seurat

- "wilcox" : Wilcoxon rank sum test (default)
- "bimod" : Likelihood-ratio test for single cell feature expression, ([McDavid et al., Bioinformatics, 2013](#))
- "roc" : Standard AUC classifier
- "t" : Student's t-test
- "poisson" : Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets
- "negbinom" : Likelihood ratio test assuming an underlying negative binomial distribution. Use only for UMI-based datasets
- "LR" : Uses a logistic regression framework to determine differentially expressed genes. Constructs a logistic regression model predicting group membership based on each feature individually and compares this to a null model with a likelihood ratio test.
- "MAST" : GLM-framework that treats cellular detection rate as a covariate ([Finak et al., Genome Biology, 2015](#)) ([Installation instructions](#))
- "DESeq2" : DE based on a model using the negative binomial distribution ([Love et al., Genome Biology, 2014](#)) ([Installation instructions](#))

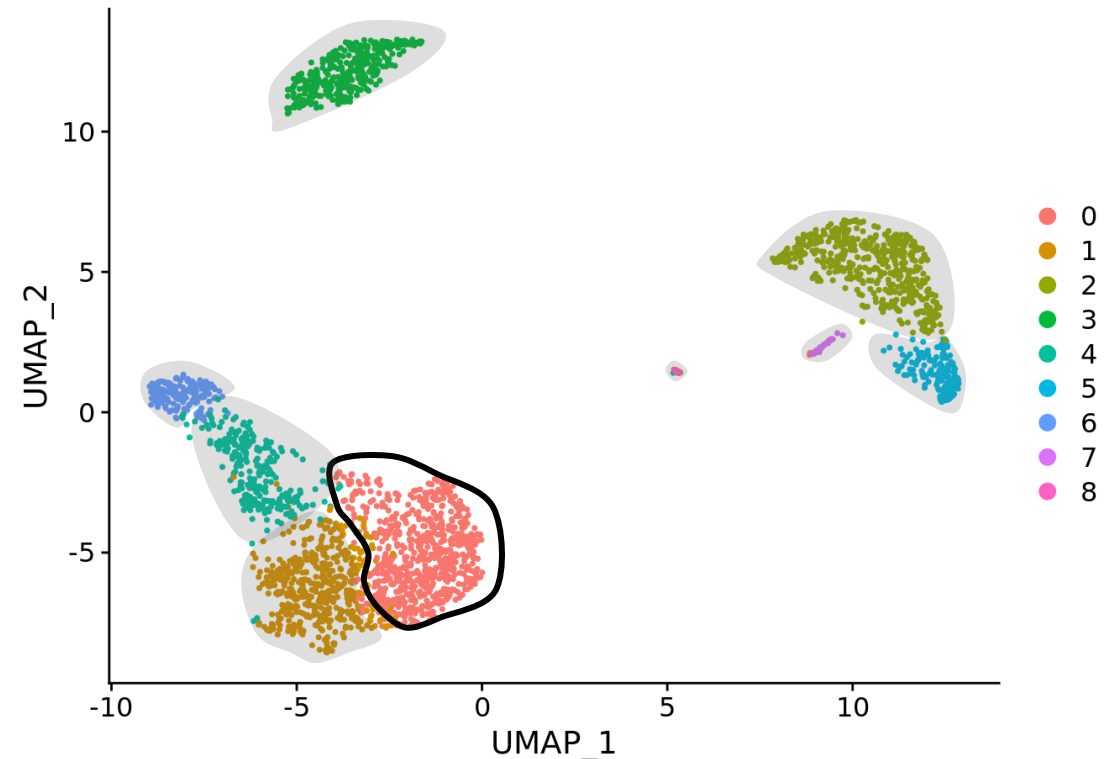
Identifying cluster markers

- Approach 1: one-vs-all (default is Seurat)
- Limitations:
 - Sensitive to the population composition (one dominant population can drive marker selection for every other cluster)



Identifying cluster markers

- Approach 2: multiple pairwise comparisons (default in scanr)
- Strategies to combine results:
 - Prioritize genes significant in *any* pairwise comparison -> focuses on combinations of genes that (together) drive separation of a cluster from the others
 - Prioritize genes significant in *all* pairwise comparisons -> explicitly favors genes that are uniquely expressed in a cluster (too stringent)
- Limitations:
 - How to combine and report results?
 - Slow

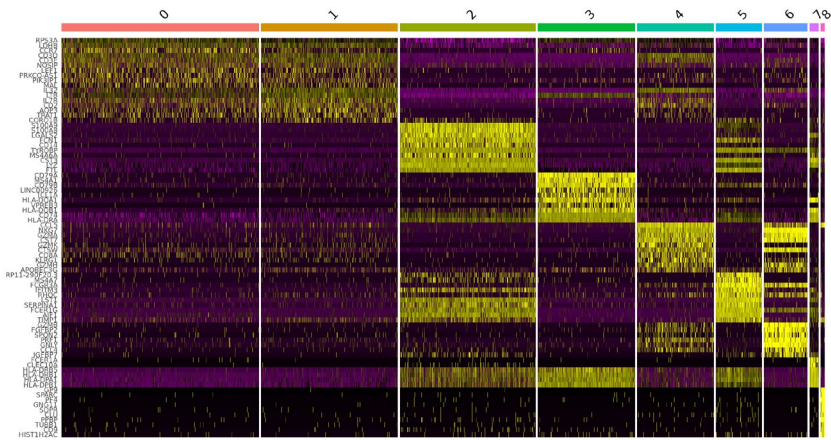


Additional (practical) considerations

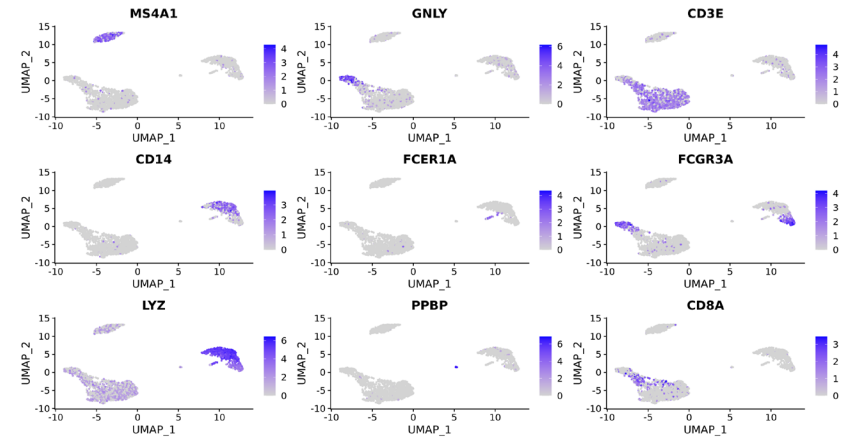
- Focus on *positive* markers only
 - It is difficult to interpret and experimentally validate the absence of expression
- Focus on genes with *large effect size* (log fold-change, LFC)
 - More biologically interesting markers (e.g. possible to validate with qPCR)
 - Faster testing (in Seurat)
- Filter genes that are very infrequently detected in either group of cells
 - Seurat: `min.pct`, `logfc.threshold`, `min.diff.pct`, `max.cells.per.ident`

Check the identified markers

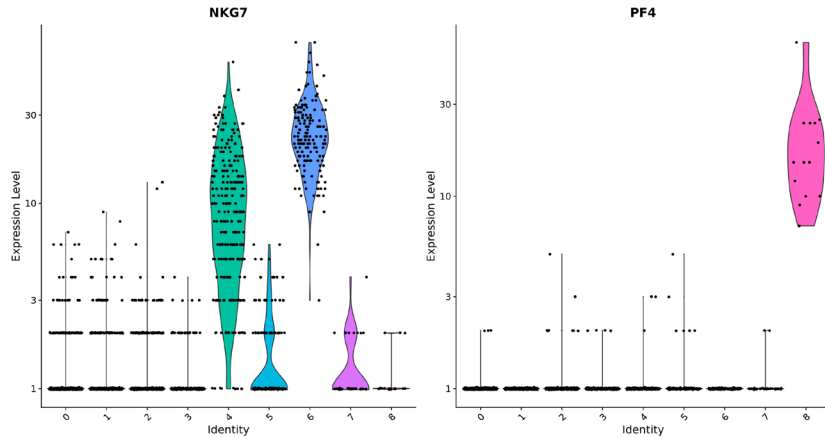
Heatmap



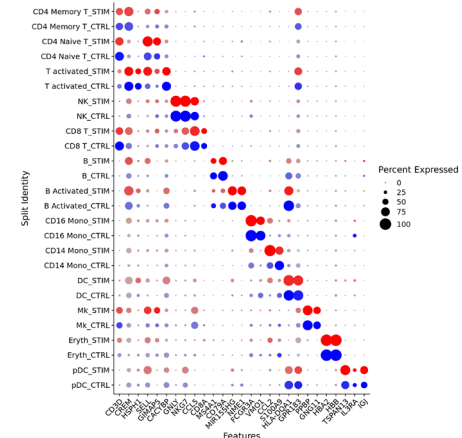
Overlap on tSNE/UMAP



Violinplot



Dotplot



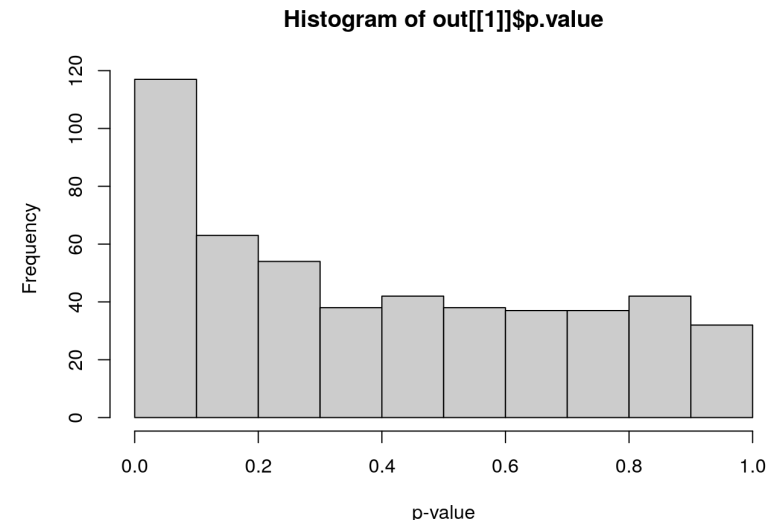
Invalidity of p-values

- DE analysis to detect marker genes between clusters is statistically flawed!

Invalidity of p-values

- Simulate i.i.d. normal values
 - perform k-means clustering
 - test for DE between clusters
 - Plot the distribution of the resulting p-values
-
- heavily skewed towards low values -> we can detect “significant” differences between clusters even in the absence of any real substructure in the data.

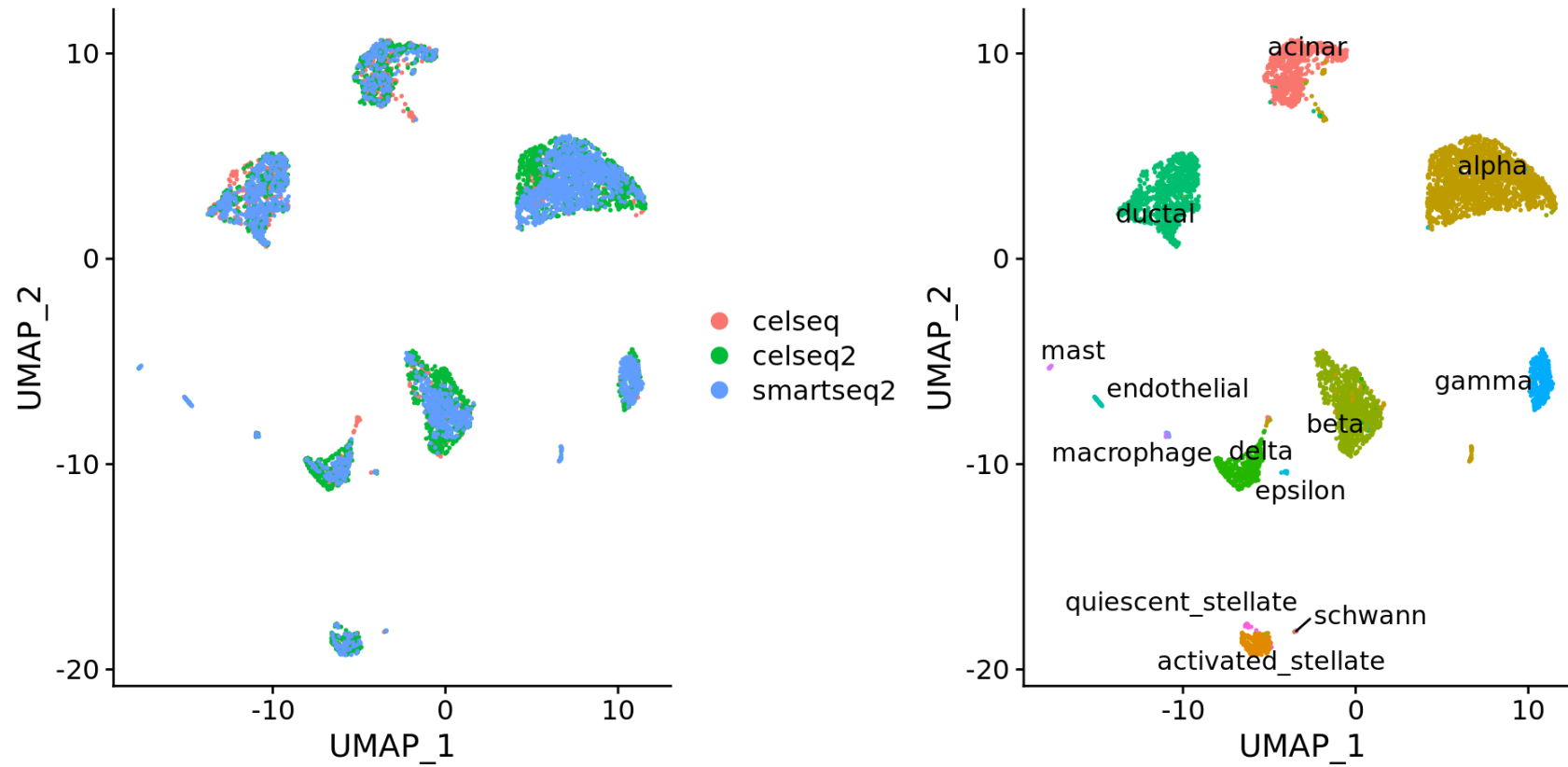
```
library(scran)
set.seed(0)
y <- matrix(rnorm(100000), ncol=200)
clusters <- kmeans(t(y), centers=2)$cluster
out <- findMarkers(y, clusters)
hist(out[[1]]$p.value, col="grey80", xlab="p-value")
```



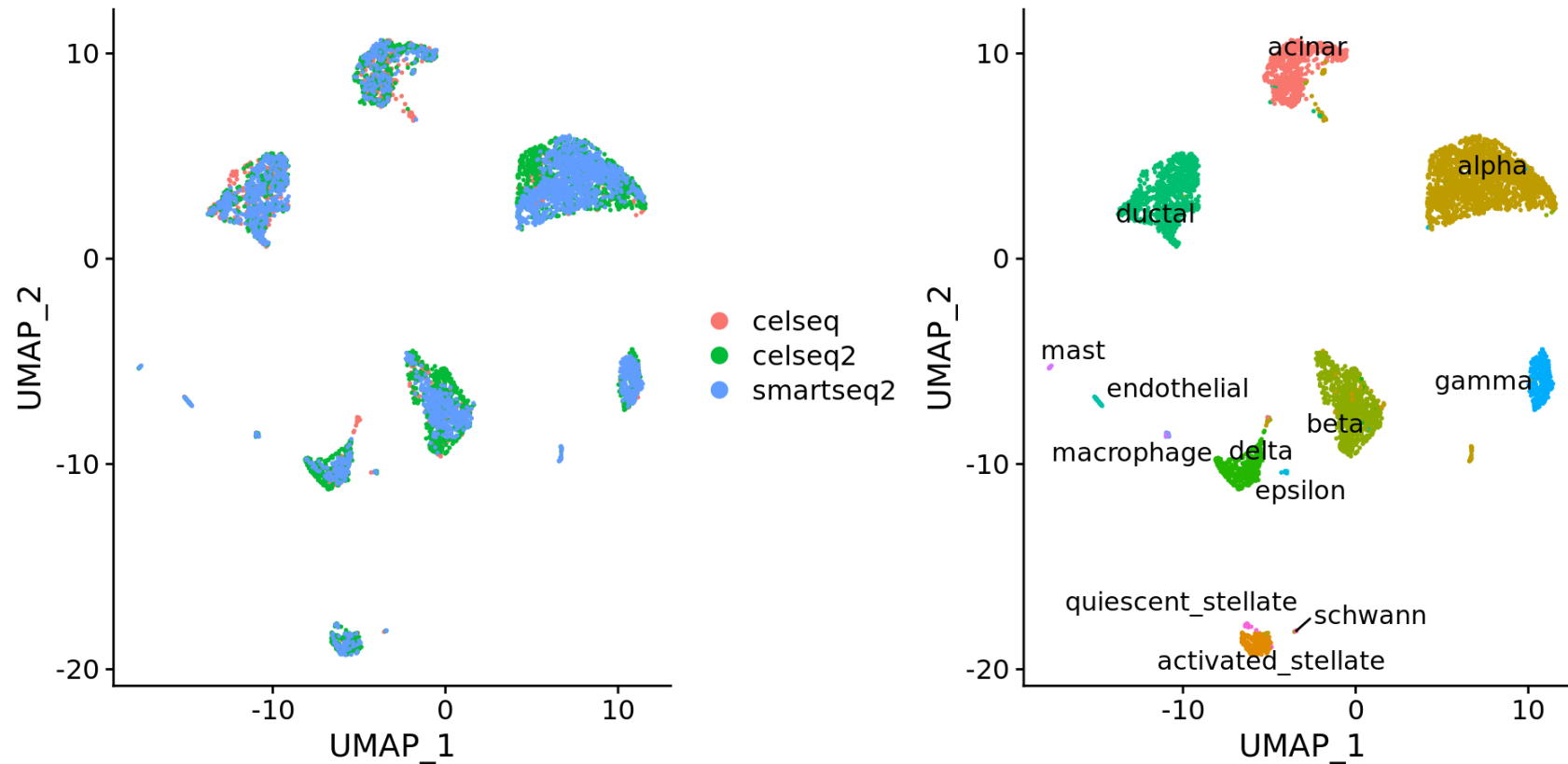
Invalidity of p-values

- DE analysis to detect marker genes between clusters is statistically flawed!
- DE analysis is performed on the same data used to obtain the clusters (data snooping) -> testing for DE genes between clusters will inevitably yield some significant results (that is how the clusters were defined).
- For marker gene detection, this effect is largely harmless as the p-values are used only for ranking.
- However, it becomes an issue when the p-values are used to define “significant differences” between clusters

DE with integrated data



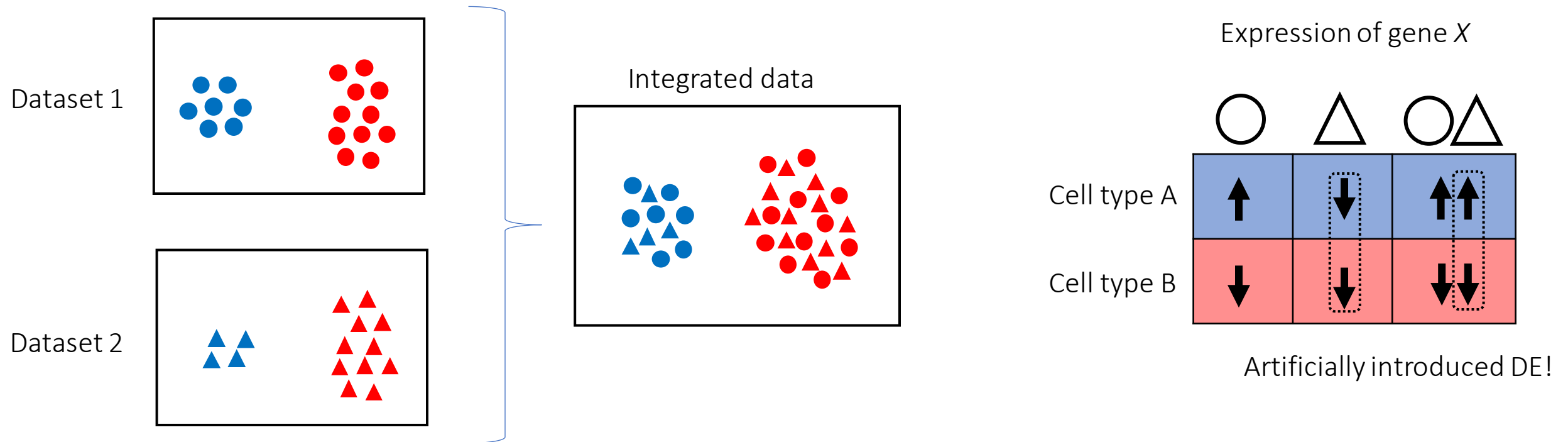
DE with integrated data



Uncorrected, measured data should be used for DE testing

Why uncorrected values?

- Correction algorithms are not obliged to preserve the magnitude or direction of differences in per-gene expression when attempting to align multiple batches.



How to perform DE with integrated data?

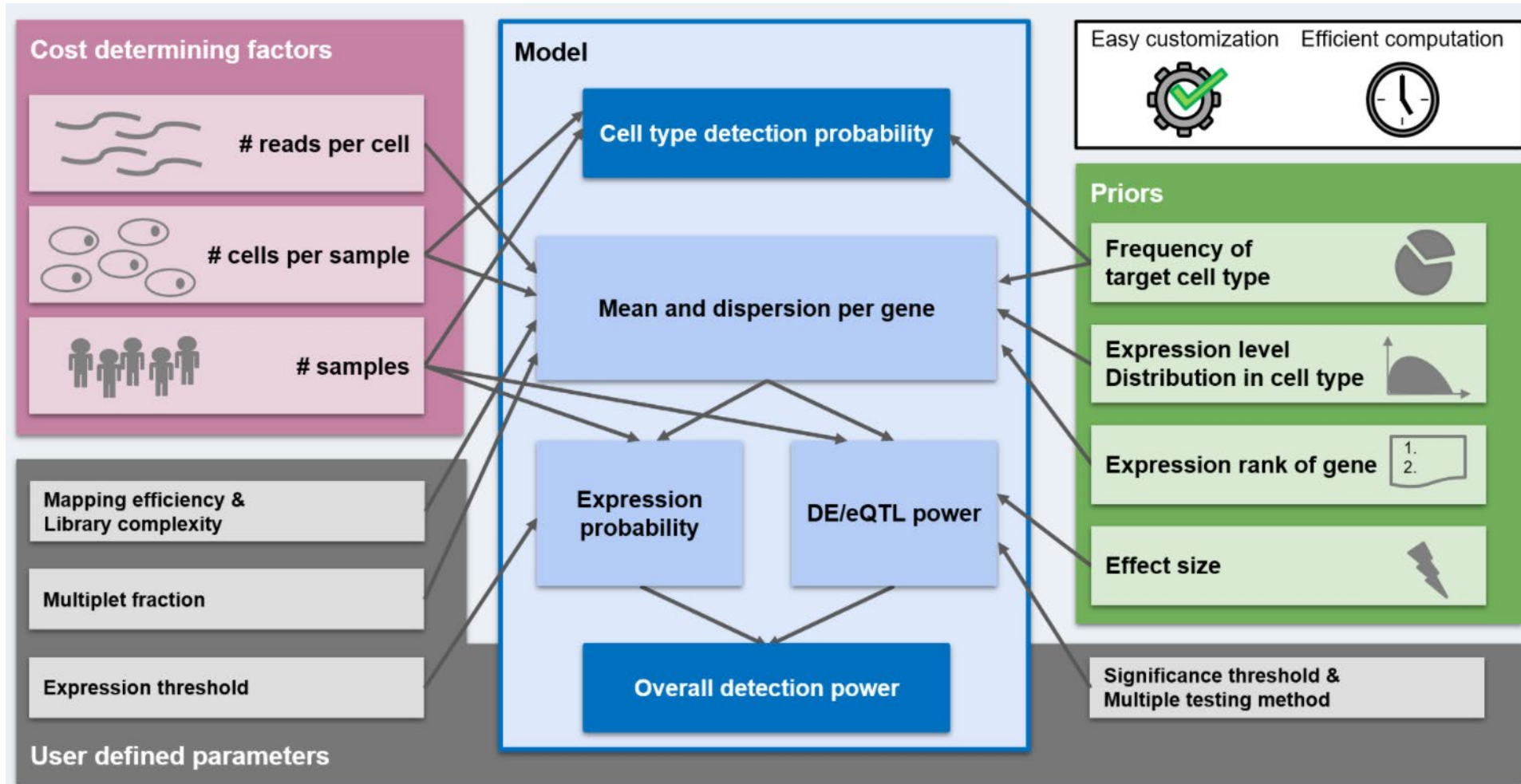
- Perform DE using the uncorrected values, separately per batch and combine p-values using meta-analysis.
- Similar to incorporating covariates in bulk DE analysis
- Penalizes genes with inconsistent DE across batches
- In practice:
 - Seurat, use the `FindConservedMarkers` function
 - scran, incorporating batches as blocks in the `findMarkers` function

Power analysis for scRNA-seq DE

Power analysis for scRNA-seq DE

- What is the best possible experiment we can do for a limited budget
- Given a certain budget, how many *samples*, how many *cells per sample*, and how many *reads per cell* are required to detect DEGs given certain assumptions about the expected effect sizes
- Available tools:
 - PowsimR
 - scDesign
 - scPower
 - ...

scPower



Website: <http://scpower.helmholtz-muenchen.de/>

R package: <https://github.com/heiniglab/scPower>

scPower

- Uses prior knowledge based on published data (website) or your own pilot data (R package)
- General recommendation: shallow sequencing of a large number of cells per individual (many DE scenarios, different platforms)
- Results based on using Negative binomial regression to detect DEGs, results might be different for other tests/models!

To summarize

- Differential expression can be used to identify cell type markers or to compare biological conditions
- For marker gene identification (comparing cell types), MAST and Wilcoxon rank-sum test perform well
- P-values obtained from cell type comparisons are statistically invalid
- For comparisons between conditions, it is better to use pseudobulk approaches to account for variation between biological replicates
- DE testing should not be performed on batch-corrected data, but instead on measured data with technical covariates included in the model

Thank You!

 a.mahfouz@lumc.nl

 mahfouzlab.org

 @ahmedElkoussy