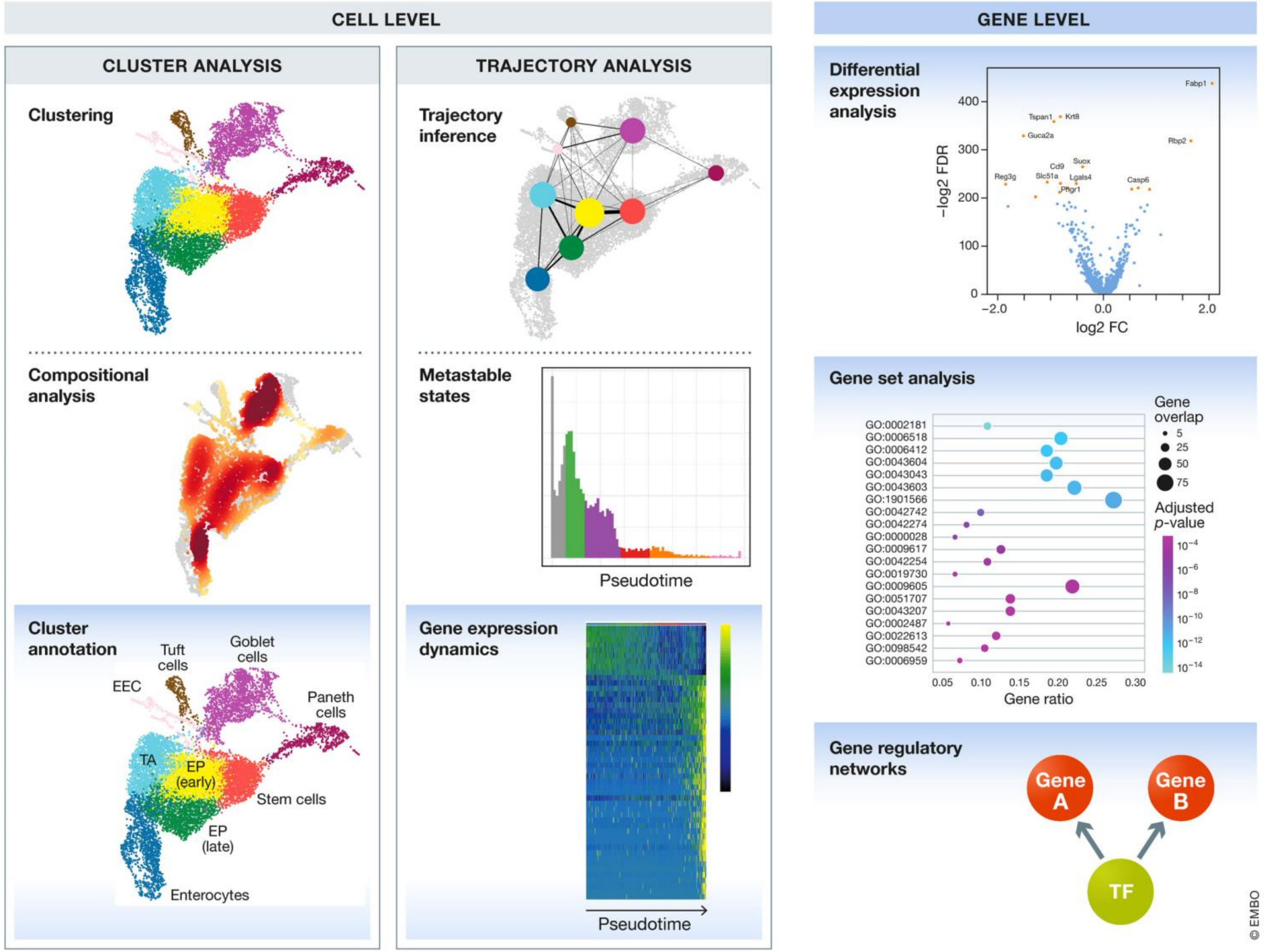
Differential expression analysis

Ahmed Mahfouz

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

@ahmedElkoussy

Downstream analysis of scRNA-seq data



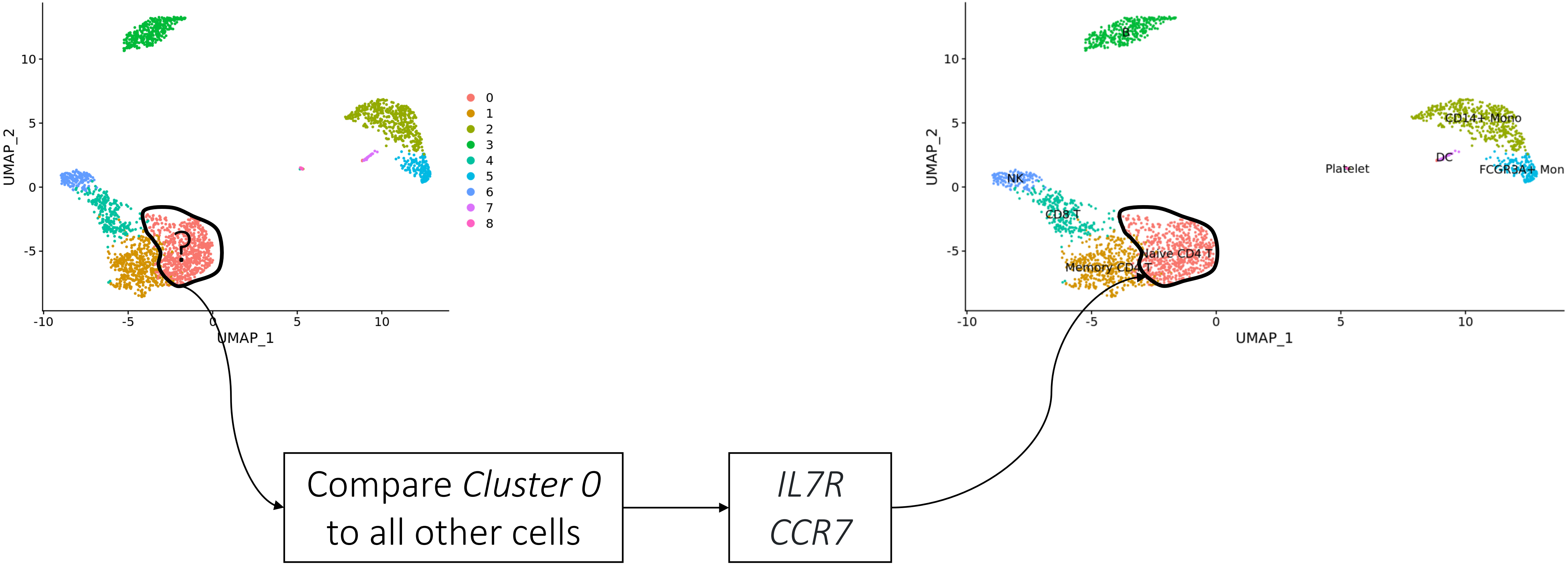
© EMBO

Luecken and Theis (MSB 2019)

DE for cluster annotation

Unannotated clusters

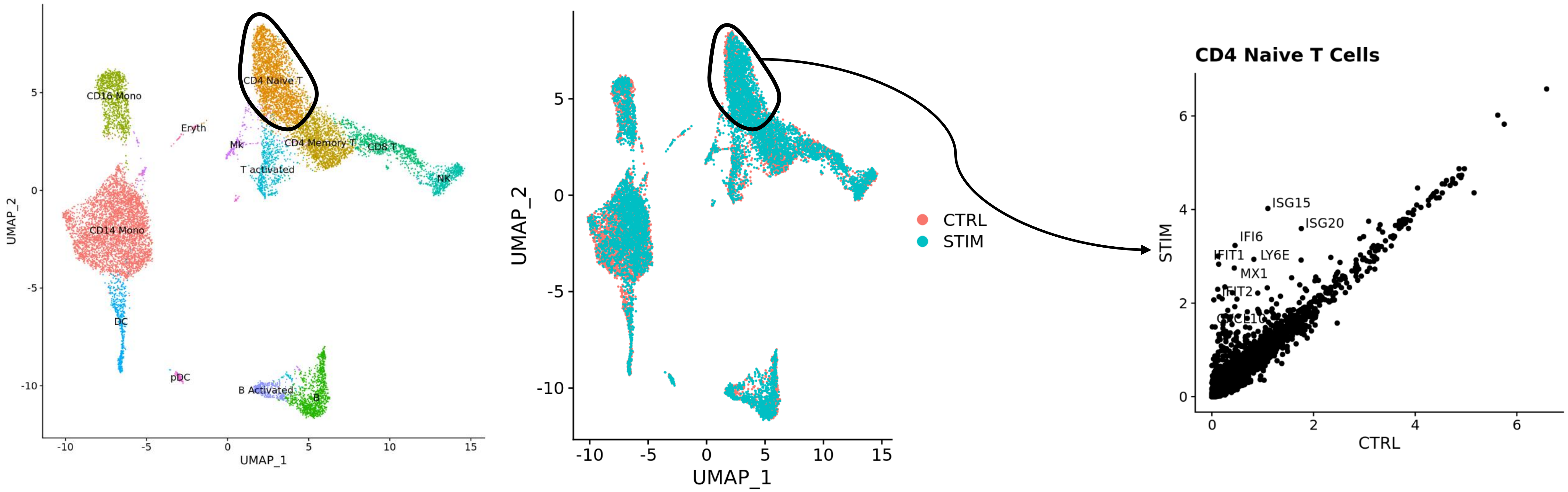
Annotated clusters



Compare *Cluster 0* to all other cells

IL7R
CCR7

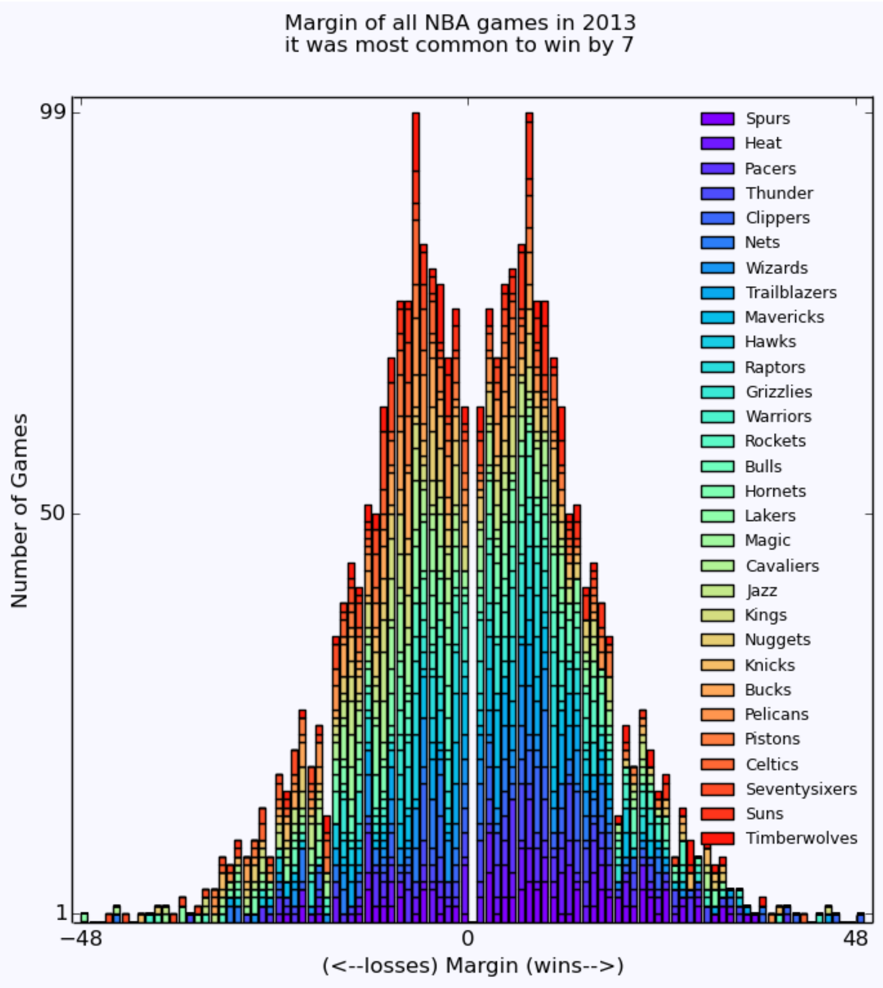
DE for comparing conditions

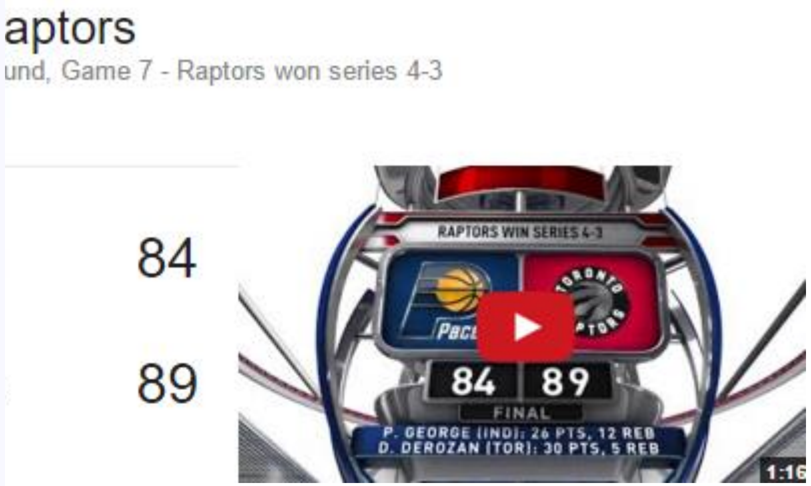


Outline

- Bulk DE analysis
- DE analysis for scRNA-seq data
- Single-cell DE in practice
- Working with integrated data

Is this a large difference?


Margin of all NBA games in 2013
it was most common to win by 7


aptors

und, Game 7 - Raptors won series 4-3

84

89

RAPTORS WIN SERIES 4-3
Pacers
TORONTO RAPTORS
84
89
FINAL
P. GEORGE (IND): 26 PTS, 12 REB
D. DEROZAN (TOR): 30 PTS, 5 REB
1:16

Hypothesis testing

1. An idea -> *hypothesis*

2. Measure something -> *data*

3. Analyse the data -> *hypothesis test*

Hypothesis testing

1. An idea -> *hypothesis*

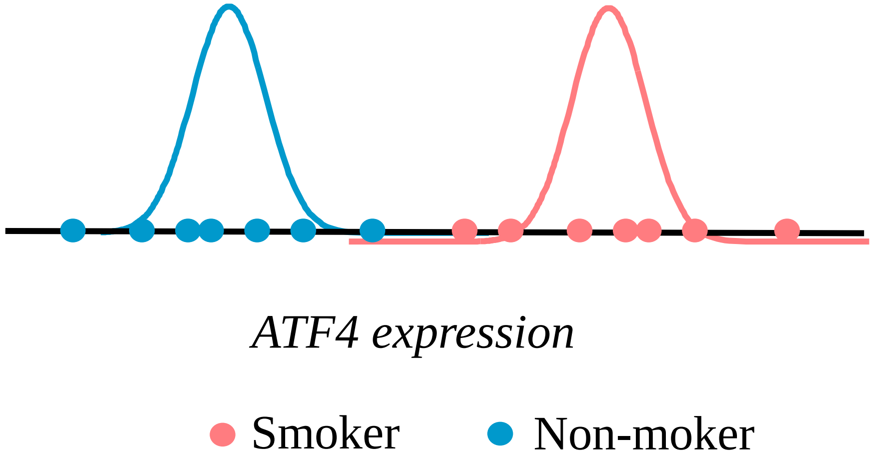
   smoking increases ATF4 expression

2. Measure something -> *data*

   RNA-seq on blood of smokers

3. Analyse the data -> *hypothesis test*

   Compare ATF4 expression between smokers and non-smokers



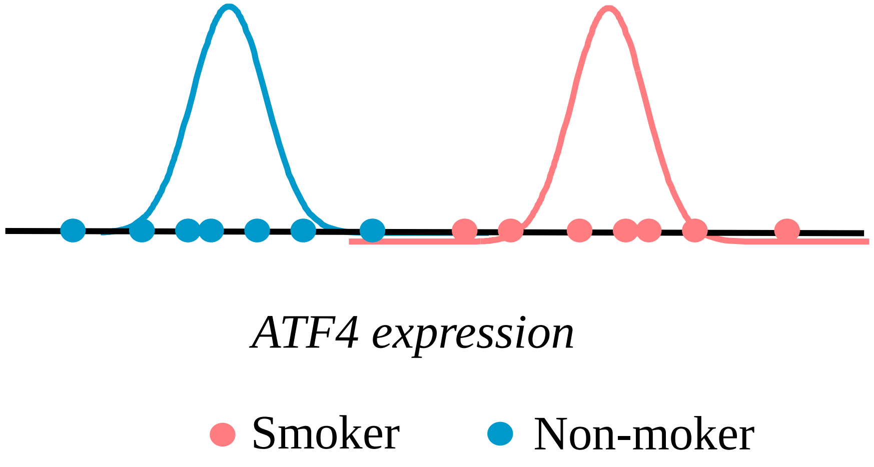
Variation in data

Sources of variation:

- Personal background and environment
- Random variation in measurements

OR

- Our hypothesis (e.g. smoking)



How do we know if an observed difference is "real"?

Statistically significant = too unlikely to be a *coincidence*

➢ But what do you expect if it is a coincidence?

Hypothesis testing

1) Assuming ***there is no real difference*** between conditions

1. null hypothesis

2) What is the probability of finding a difference in the data (population) *by chance*?

2. *p*-value

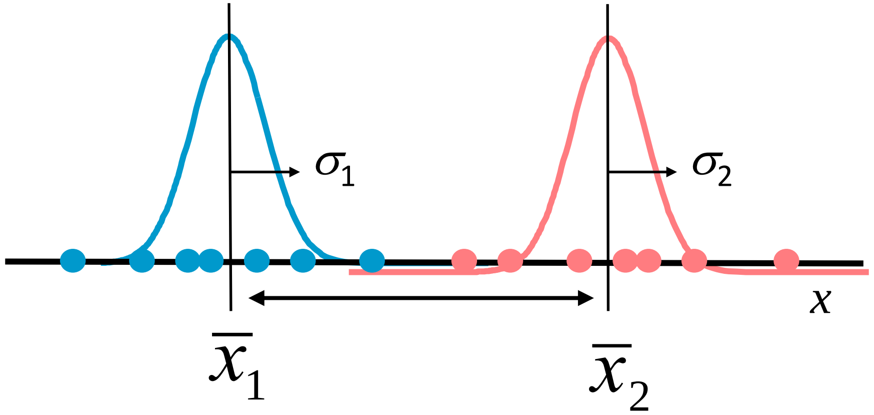
3) If this is probability is low, the *assumption is likely incorrect* : **there is a difference**

3. reject the null hypothesis?

Model the data

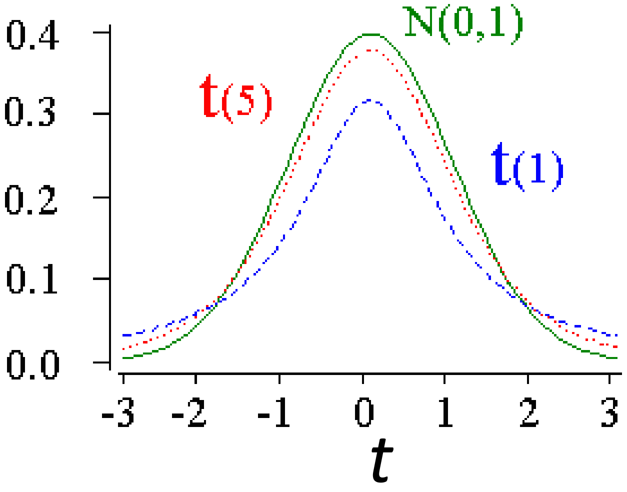
- Model the data distribution (e.g. normal)

- Use a statistic to assess the difference (e.g. t-test)



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

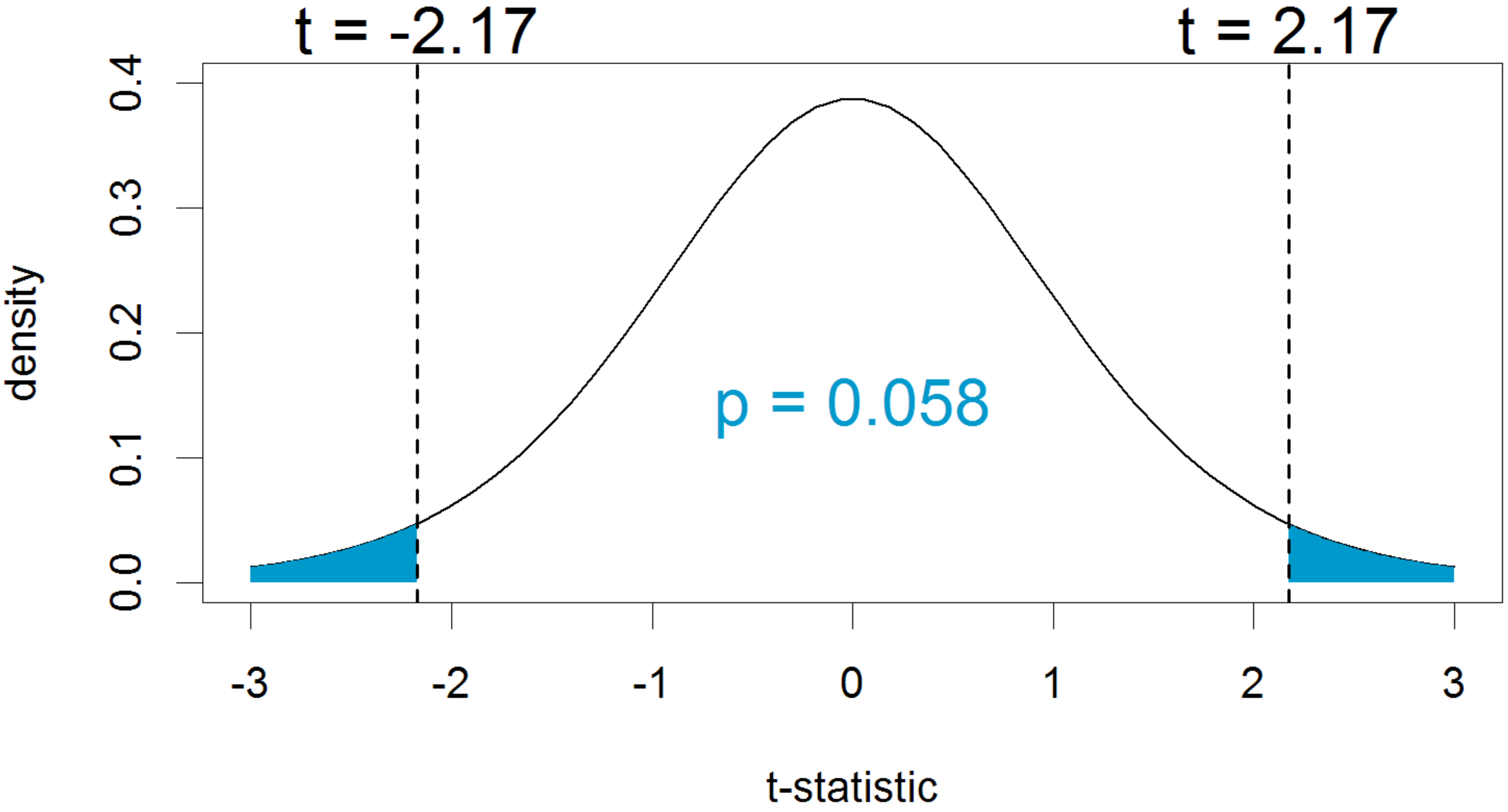
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(J_1-1)S_1^2 + (J_2-1)S_2^2}{J_1+J_2-2}\left(\frac{1}{J_1}+\frac{1}{J_2}\right)}}$$



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

P-value

Two-sided test
t = 2.17
DOF = 9

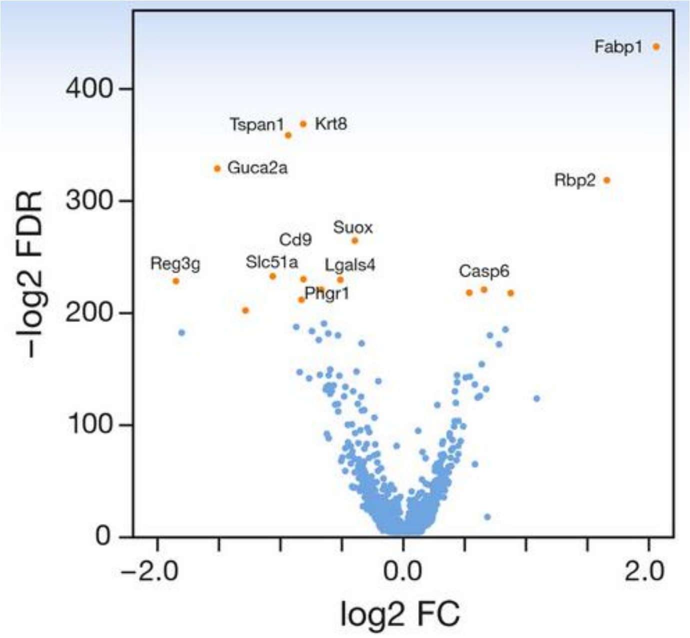


Effect size

- 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)

Can we just use a Student’s t-test for DE analysis?

- Not really:
  - ➢ *Few replicates*: wrong estimate of variance

    “borrow” information across genes to get a better variance estimate.

  - ➢ *Data distribution is not normal*:

    use discrete distributions (Poisson, negative binomial etc.) rather than continuous (e.g. normal) distributions for modeling RNA-seq data (count data)

  - ➢ *Non-symmetric wrt differences in group size*: favor genes where the larger group has the higher relative variance as this increases the estimated degrees of freedom and decreases the resulting p-value

- DESeq2 and edgeR solve these issues for bulk RNA-seq data. Can we also use them for scRNA-seq data?

What is special about scRNA-seq

Bulk

Estimate gene variance from few samples

No drop-outs

Single-cell

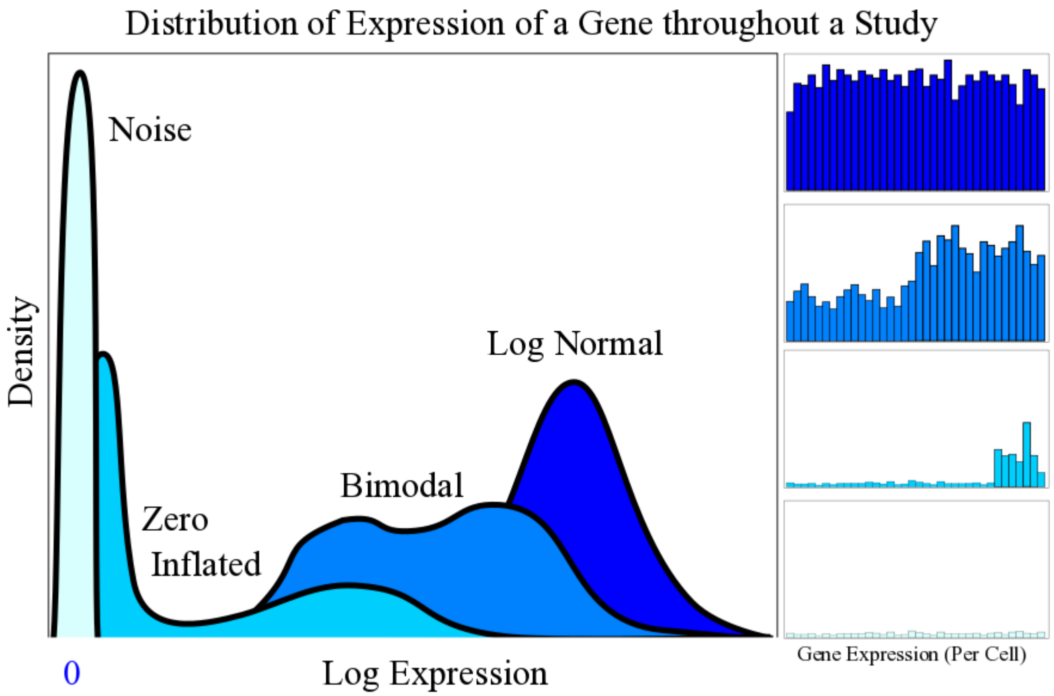
Many samples

Drop-outs

- These artefacts are taken into account in DE methods designed specifically for single-cell data
  - SCDE (Kharachenko, Nature Methods 2014)
  - MAST (Finak, Genome Biology 2015)
  - …

MAST

- 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*

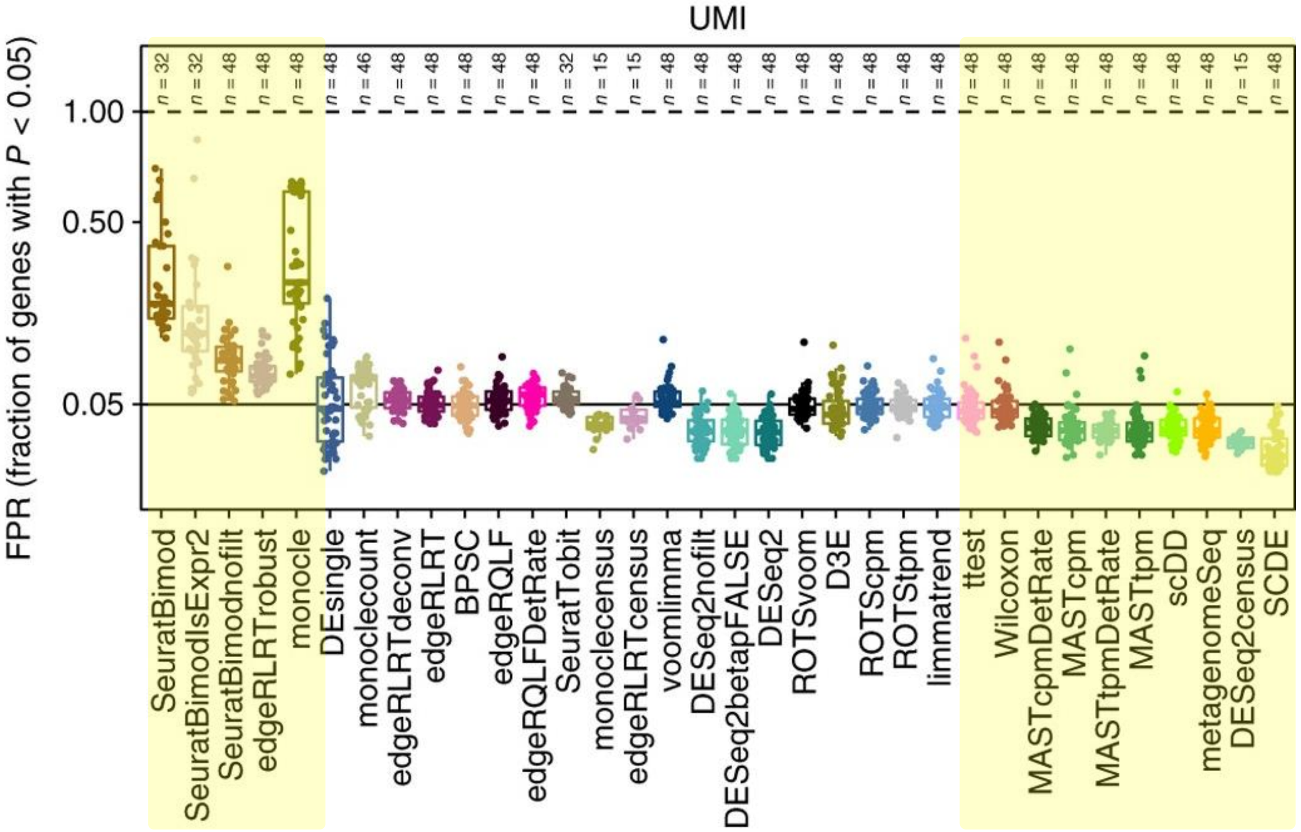


Finak et al. (Genome Biol 2015)

Comparing different methods

- Benchmark study (Soneson & Robinson, Nature Methods 2018)

- Overall, MAST, Wilcoxon, t-test outperformed other methods



Non-parametric tests

- Forget about modeling the data (it seems difficult), let’s use a non-parametric test.
    - ➢ Svensson, ***Droplet scRNA-seq is not zero-inflated***, Nature Biotechnology 2020

- 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: median$_1$ = median$_2$

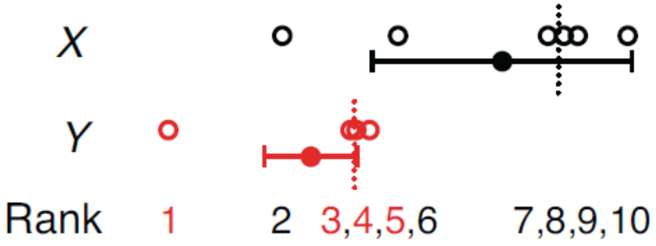
- Start by ranking all values

- 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

Example from: Krzywinski M, Altman N (2014) Nat Methods 11:467–469.

That must be the solution to everything?

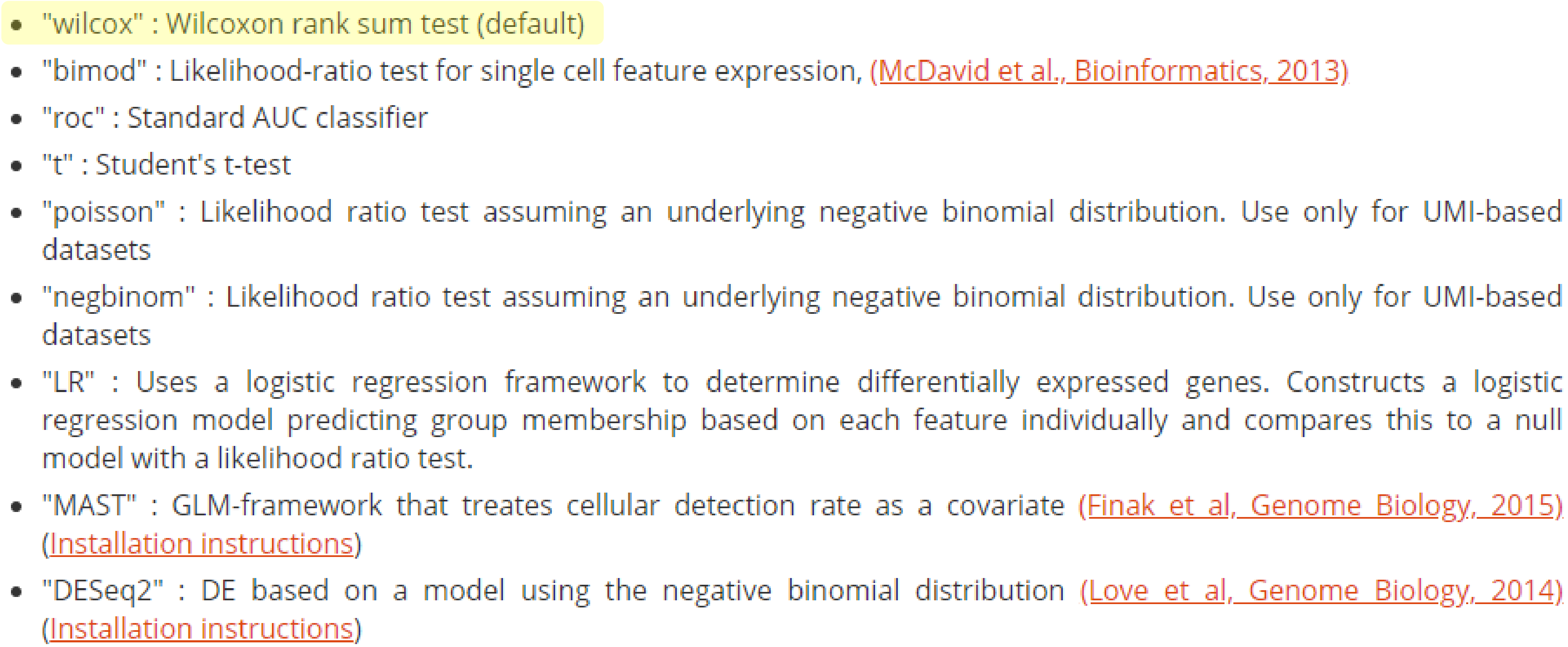
- Not really…

- 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.

Single-cell DE in practice

Seurat



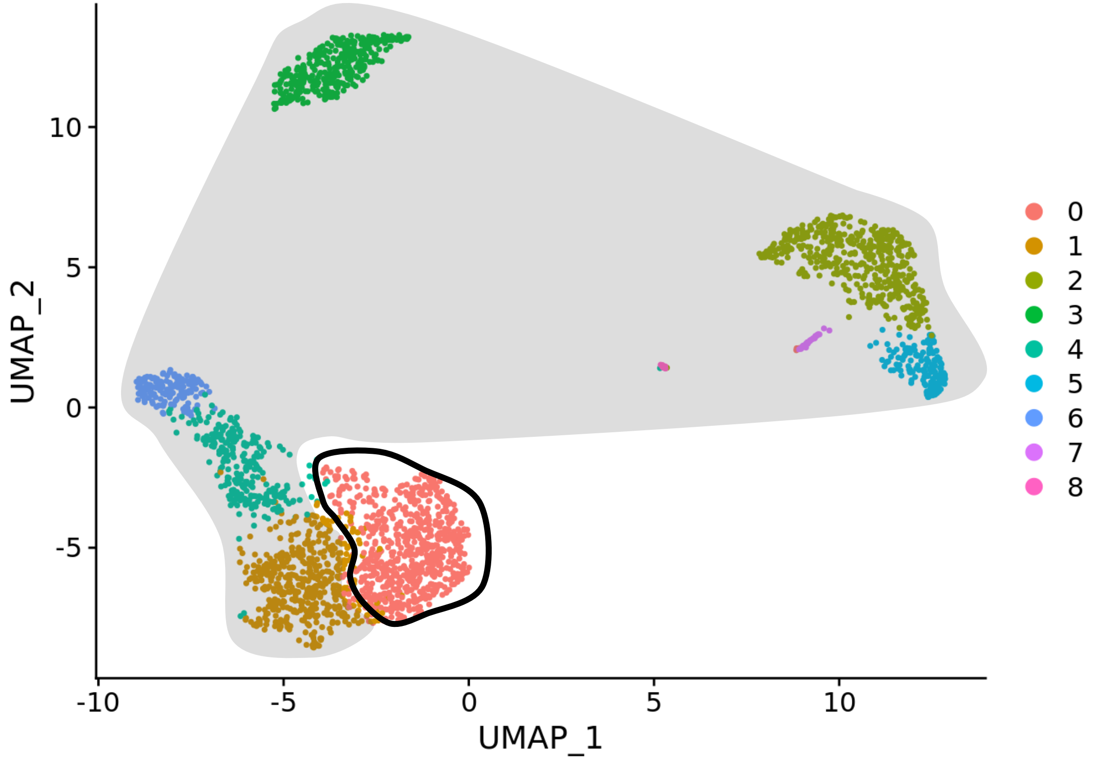
- "wilcox" : Wilcoxon rank sum test (default)
- "bimod" : Likelihood-ratio test for single cell feature expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Standard AUC classifier
- "t" : Student's t-test
- "poisson" : Likelihood ratio test assuming an underlying negative binomial distribution. 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 treates 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)

https://satijalab.org/seurat/v3.1/de_vignette.html

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)

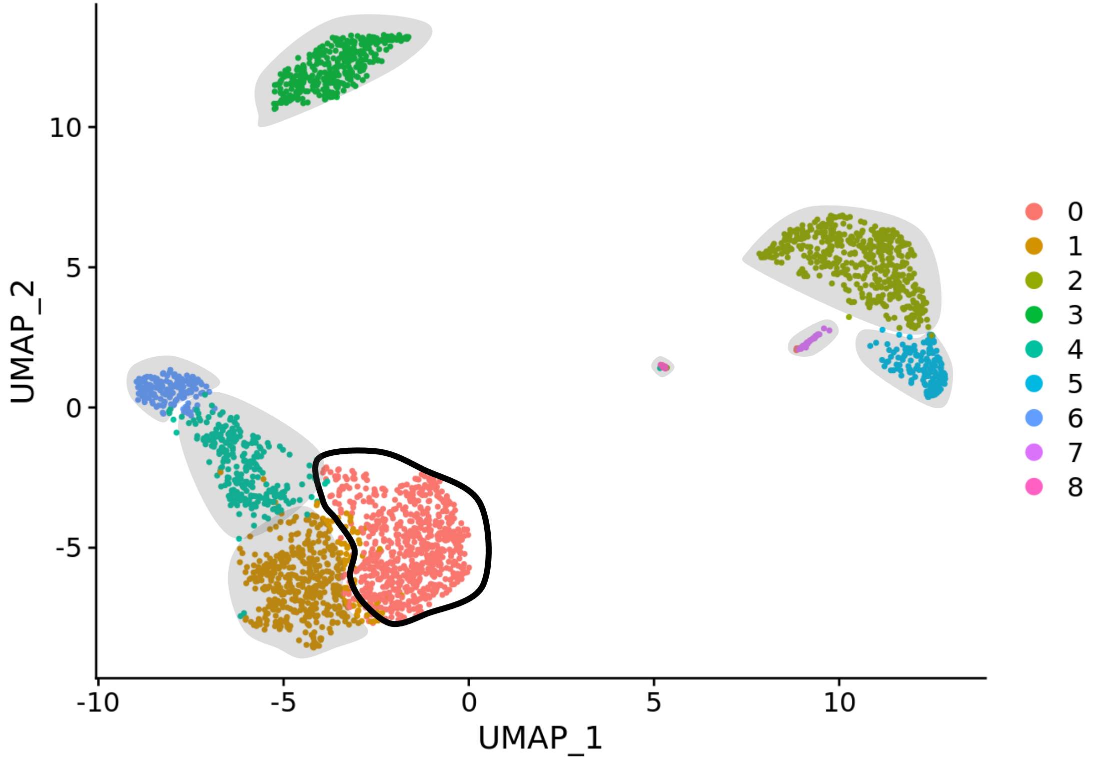


Identifying cluster markers

- Approach 2: multiple pairwise comparisons (default in scran)

- 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

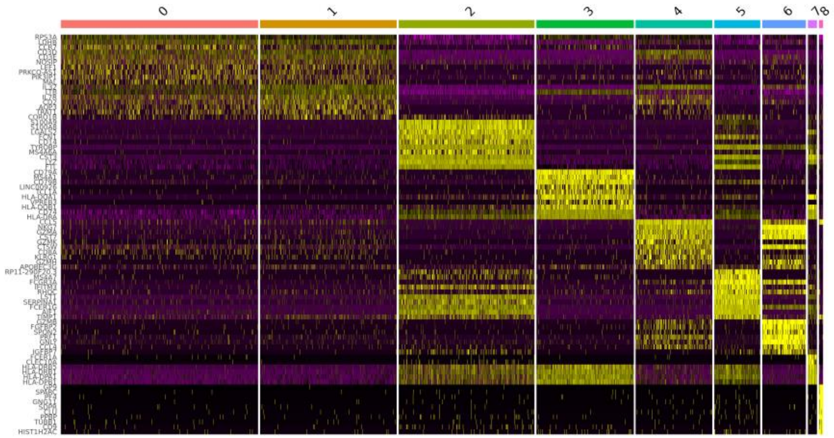
- Focus on ***positive*** markers only
  - It is difficult to interpret and experimentally validate the absence of expression

- 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)

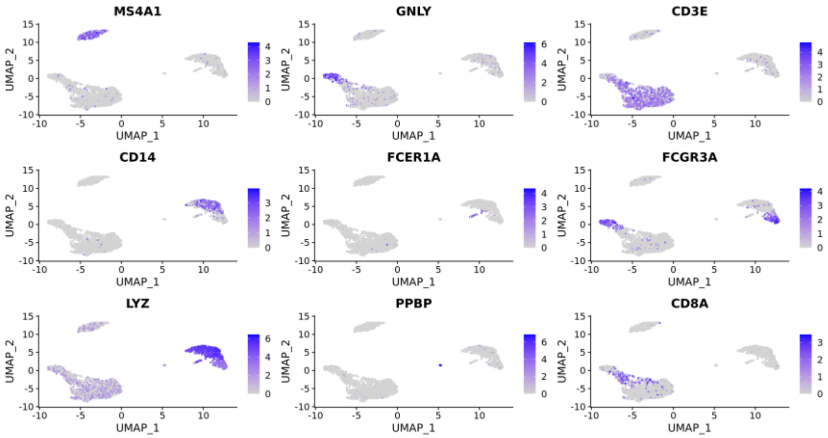
- Filter genes that are very infrequently detected in either group of cells
  - Seurat: `min.pct, logfc.threshold, min.diff.pct, max.cells.per.ident`

Check the identified markers

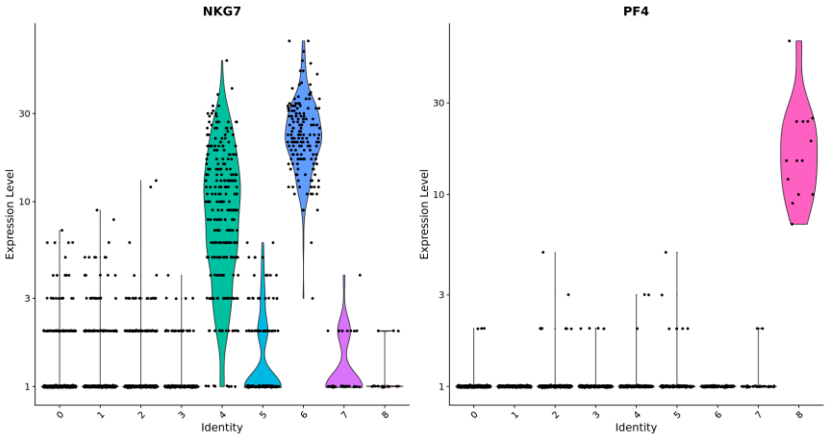
Heatmap



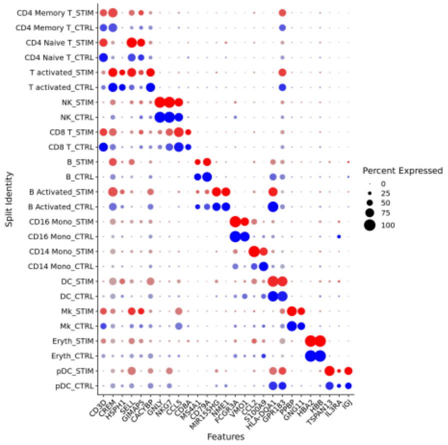
Overlap on tSNE/UMAP



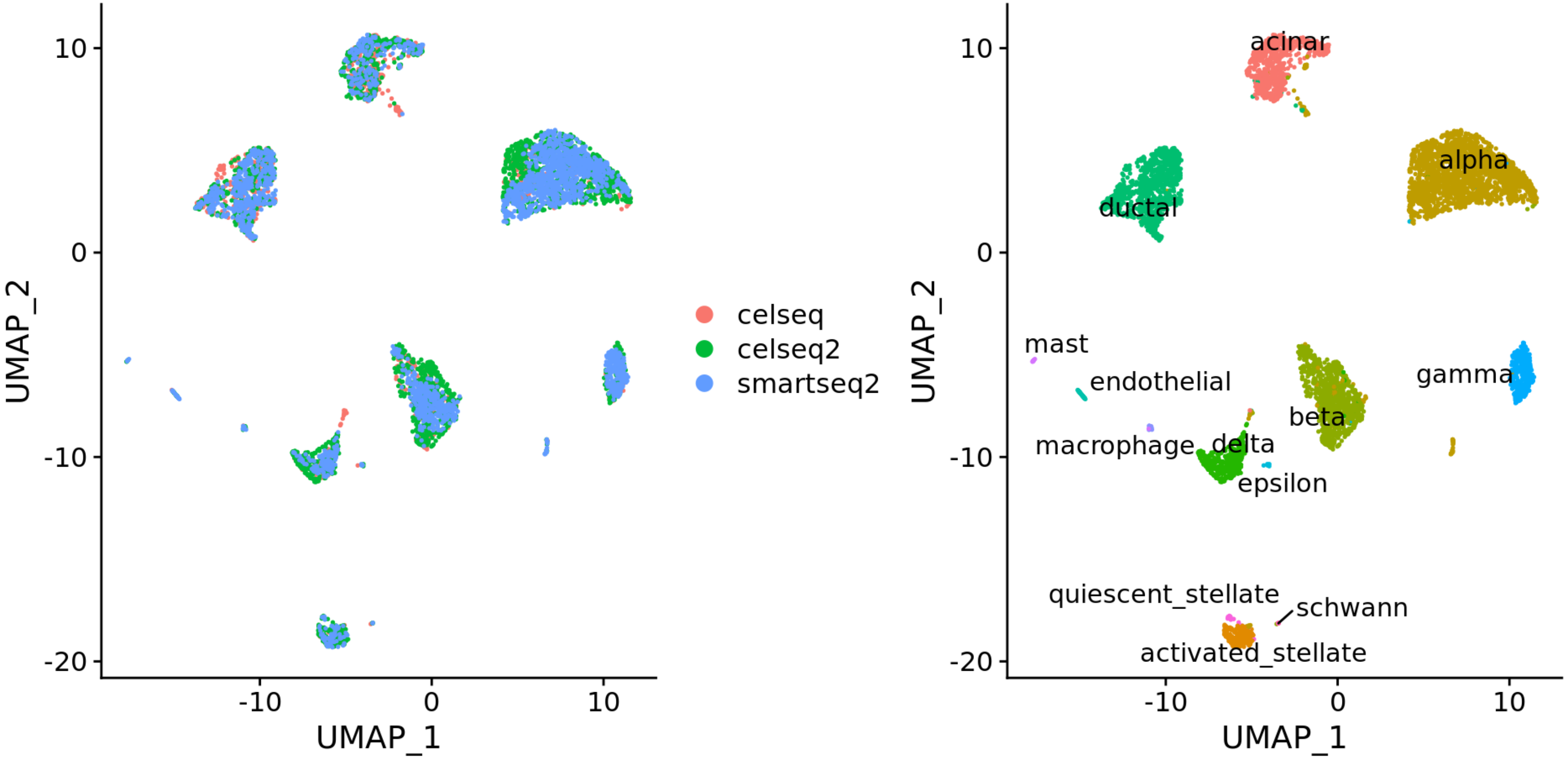
Violinplot



Dotplot



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.

- Example
  - Consider a dataset (first batch) with two cell types, *A* and *B*. Consider a second batch with the same cell types, denoted as *A′* and *B′*. Assume that, gene *X* is expressed in *A* but not in *A′*, *B* or *B′* .
  - We then merge the batches together based on the shared cell types. This yields a result where *A* and *A′* cells are intermingled and the difference due to *X* is eliminated.
  - Now, if we corrected the second batch to the first, we must have coerced the expression values of *X* in *A′* to non-zero values to align with those of *A*, while leaving the expression of *X* in *B′* and *B* at zero. Thus, we have artificially introduced DE between *A′* and *B′* for *X* in the second batch to align with the DE between *A* and *B* in the first batch.

Amezquita et al. (Nature Methods 2019)

How to perform DE with integrated data?

- Perform DE using the uncorrected values, separately per batch and combine p-values using meta-analysis.

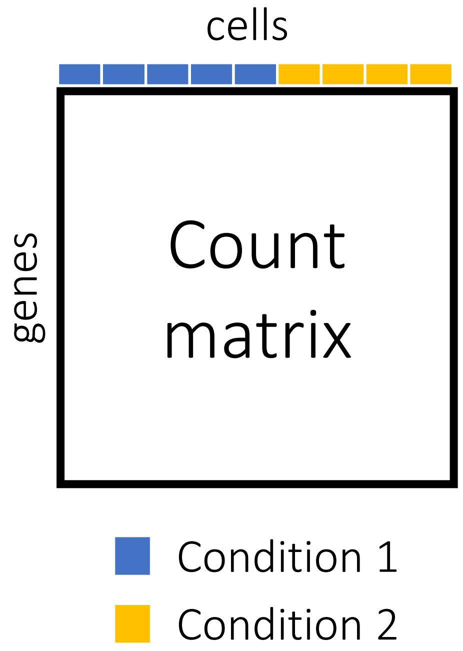
- Similar to incorporating covariates in bulk DE analysis

- Penalizes genes with inconsistent DE across batches

- In practice:
  - Seurat, use the `FindConservedMarkers` function
  - scran, incorporating batches as blocks in the `findMarkers` function

DE between conditions

1. Assembled sample-level data by aggregating measurements for each cell population (for each sample) to obtain ***pseudobulk*** data
2. Use standard DE analysis pipelines designed for bulk RNA-seq data (edgeR, limma,…)

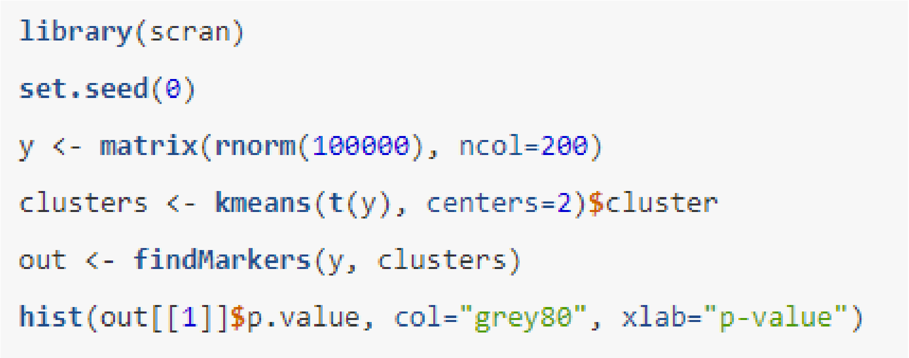


- Why?
  - Normalization is more straightforward.
  - Each sample is represented no more than once for each condition, avoiding problems from unmodelled correlations between samples.
  - Variance between cells within each sample is masked. This avoids penalizing DEGs that are not uniformly up- or down-regulated for all cells in all samples of one condition

Crowell et al. (bioRxiv 2020)

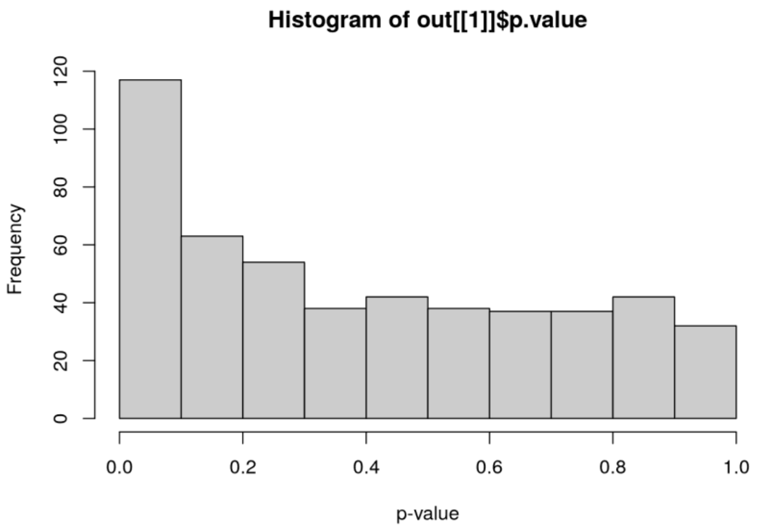
Invalidity of p-values

- Simulate i.i.d. normal values
- perform k-means clustering
- test for DE between clusters
- Plot the distribution of the resulting p-values



```r
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")
```

- heavily skewed towards low values -> we can detect “significant” differences between clusters even in the absence of any real substructure in the data.



Amezquita et al. (Nature Methods 2019)

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

Amezquita et al. (Nature Methods 2019)

To summarize

- MAST and Wilcoxon rank-sum test perform well on scRNA-seq data

- DE testing should not be performed on batch-corrected data, but instead on measured data with technical covariates included in the model

- DE between conditions is better done using aggregated pseudobulk data

Mini-symposium (Friday 23 October 2020)

9:00	Anna Alemany *Hubrecht Institute*	Single-cell and Spatial transcriptomics reveal somitogenesis in mouse gastruloids
9:45	Jop Kind *Hubrecht Institute*	Simultaneous quantifications of epigenetics and transcriptomics in the same cell with scDam&T
		Break
11:00	Ruben Boers *Erasmus MC*	Whole genome cell state tracing of gene and enhancer activity in the small intestine
11:45	Stefan Semrau *Leiden University*	Single-cell RNA-seq unravels developmental dynamics in vivo and in vitro

Before you go…

- Rstudio Cloud will be accessible until 19 November 2020.

- All course materials (lectures, markdown files, data,…) is available:

https://github.com/LeidenCBC/MGC-BioSB-SingleCellAnalysis2020

- Don’t forget to return the evaluation forms after the mini-symposium.

Thank You!

✉ a.mahfouz@lumc.nl

🔗 https://www.lcbc.nl/

@ahmedElkoussy