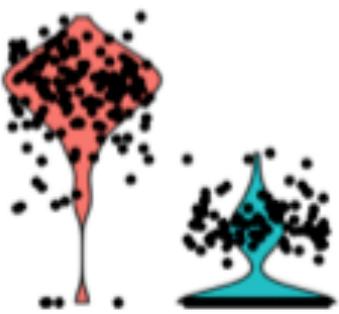


# Differential expression analysis



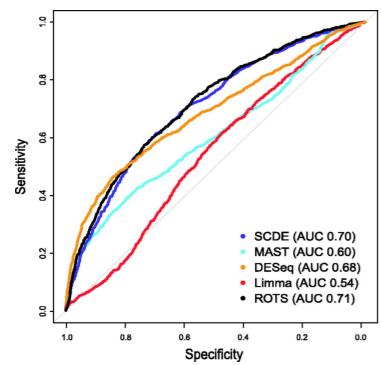


## ❖ What is “differential expression analysis”

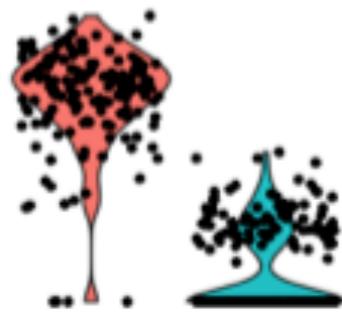


## ❖ Common methods

- ▶ intro to statistical inference

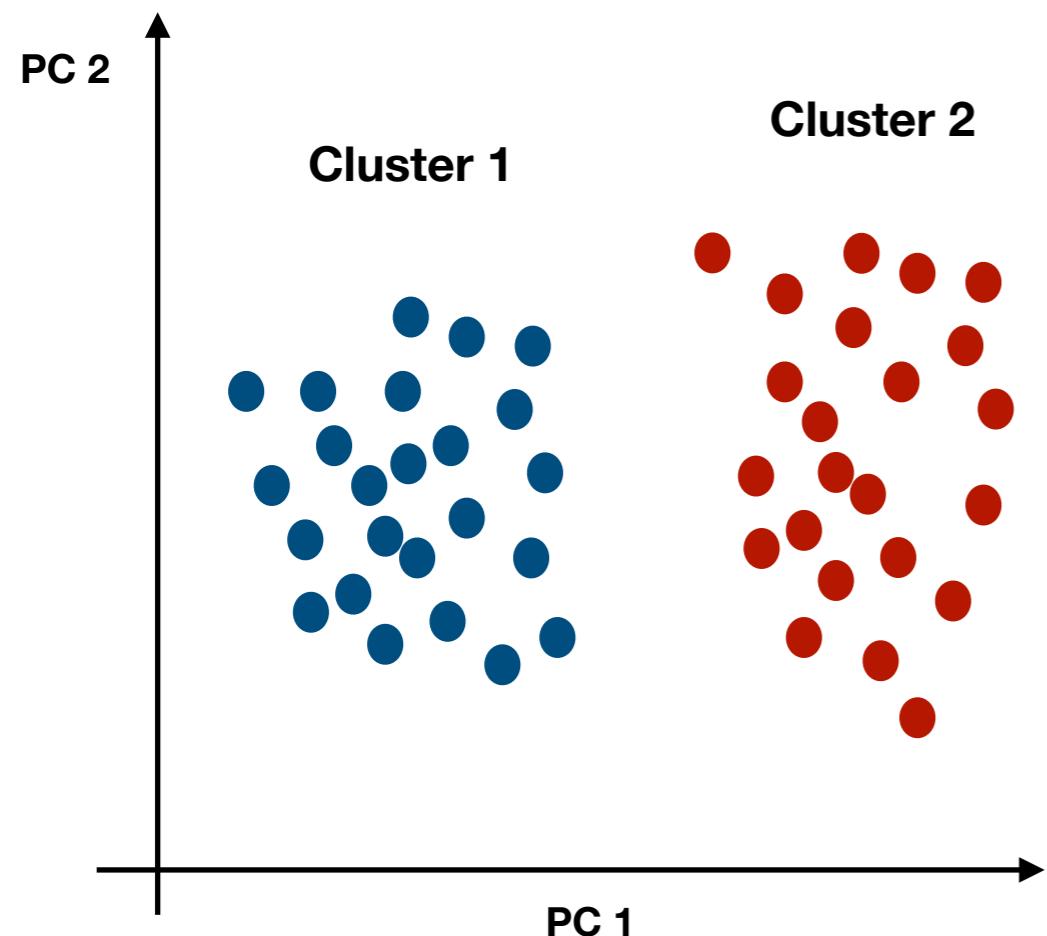


## ❖ Performance

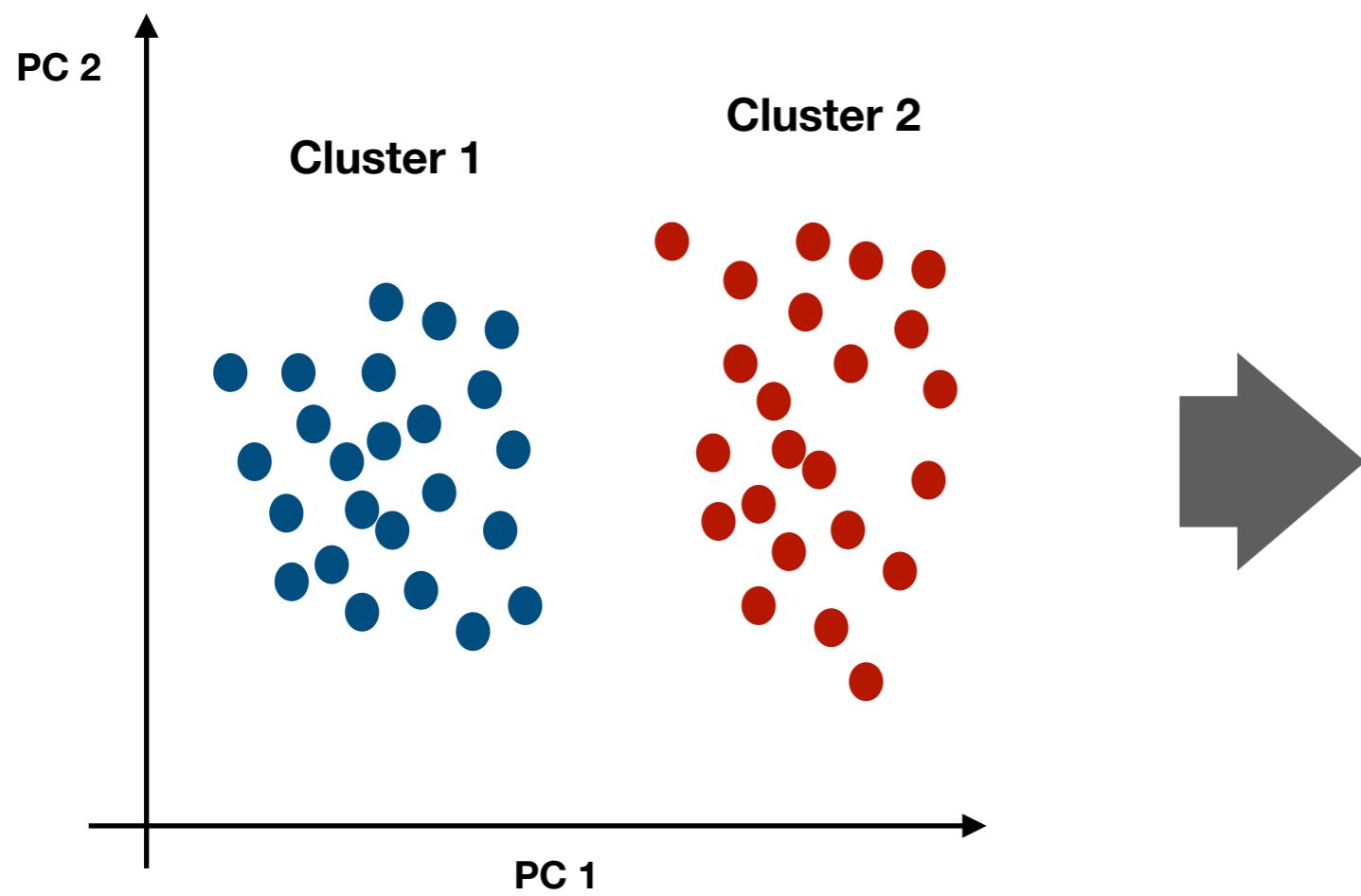


❖ What is “differential expression analysis” ?

❖ What is “differential expression analysis”



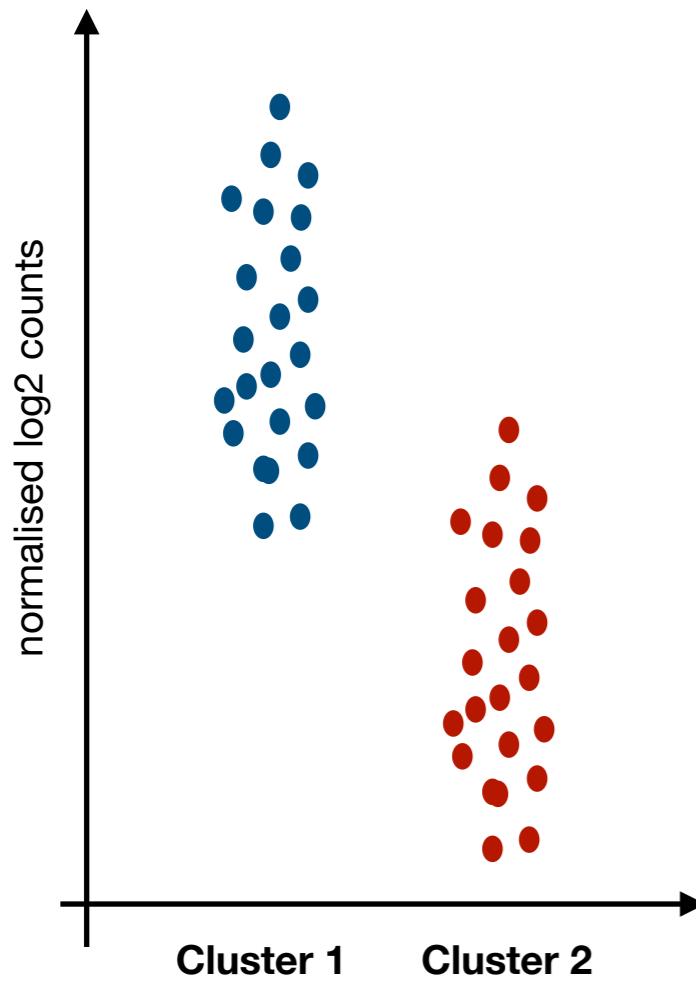
## ❖ What is “differential expression analysis”



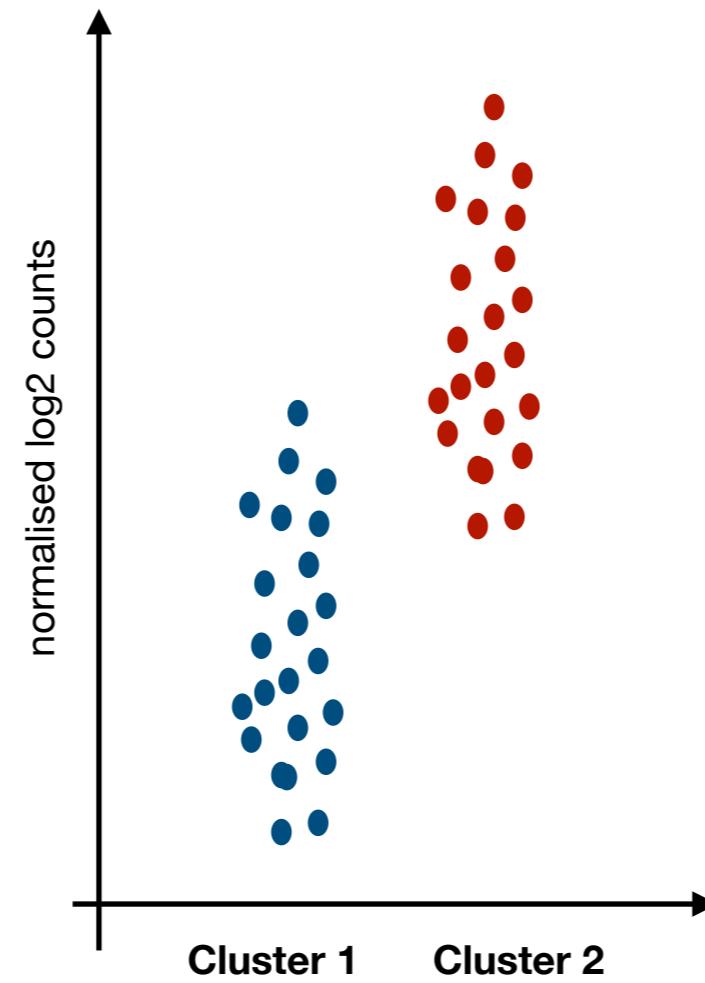
gene	logFC (avg)	p-value
CD79A	2.82	$4.73 \times 10^{-20}$
CD79B	2.23	$4.07 \times 10^{-19}$
MS4A1	-2.44	$4.67 \times 10^{-19}$
CD74	2.07	$2.56 \times 10^{-17}$
HLA-DRB1	1.53	$5.04 \times 10^{-17}$
IGHM	-3.7	$6.00 \times 10^{-17}$
HLA-DPA1	1.45	$1.11 \times 10^{-16}$
HLA-DQB1	1.73	$2.35 \times 10^{-16}$
...		
...		
...		

HLA-DQA1	1.83	$3.01 \times 10^{-16}$
HLA-DRA	1.49	$4.66 \times 10^{-16}$

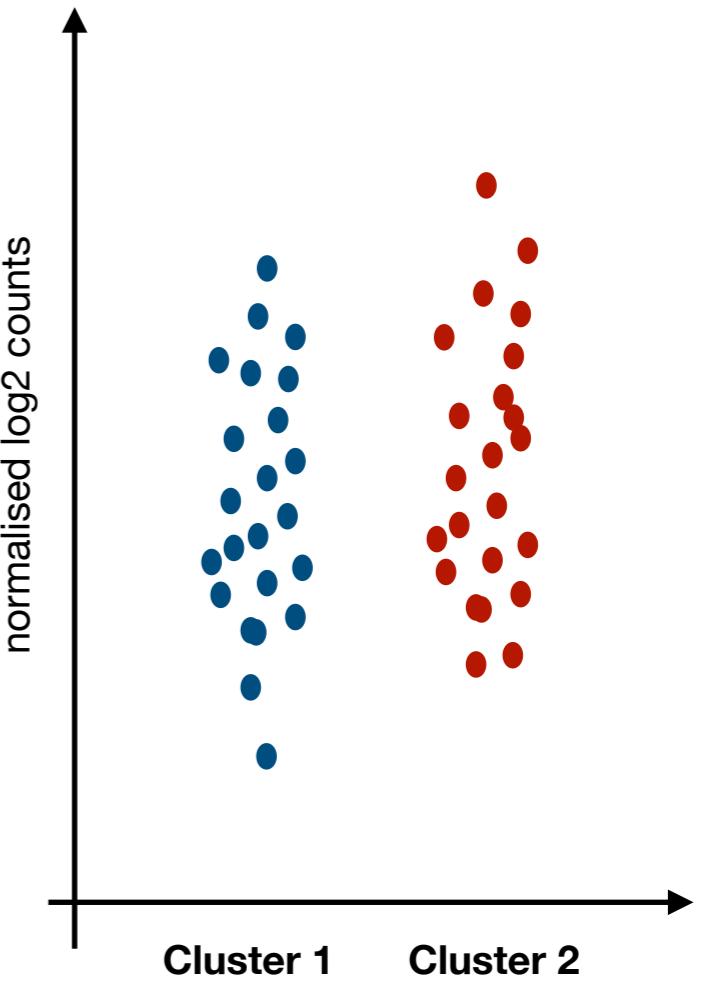
**CD79A**



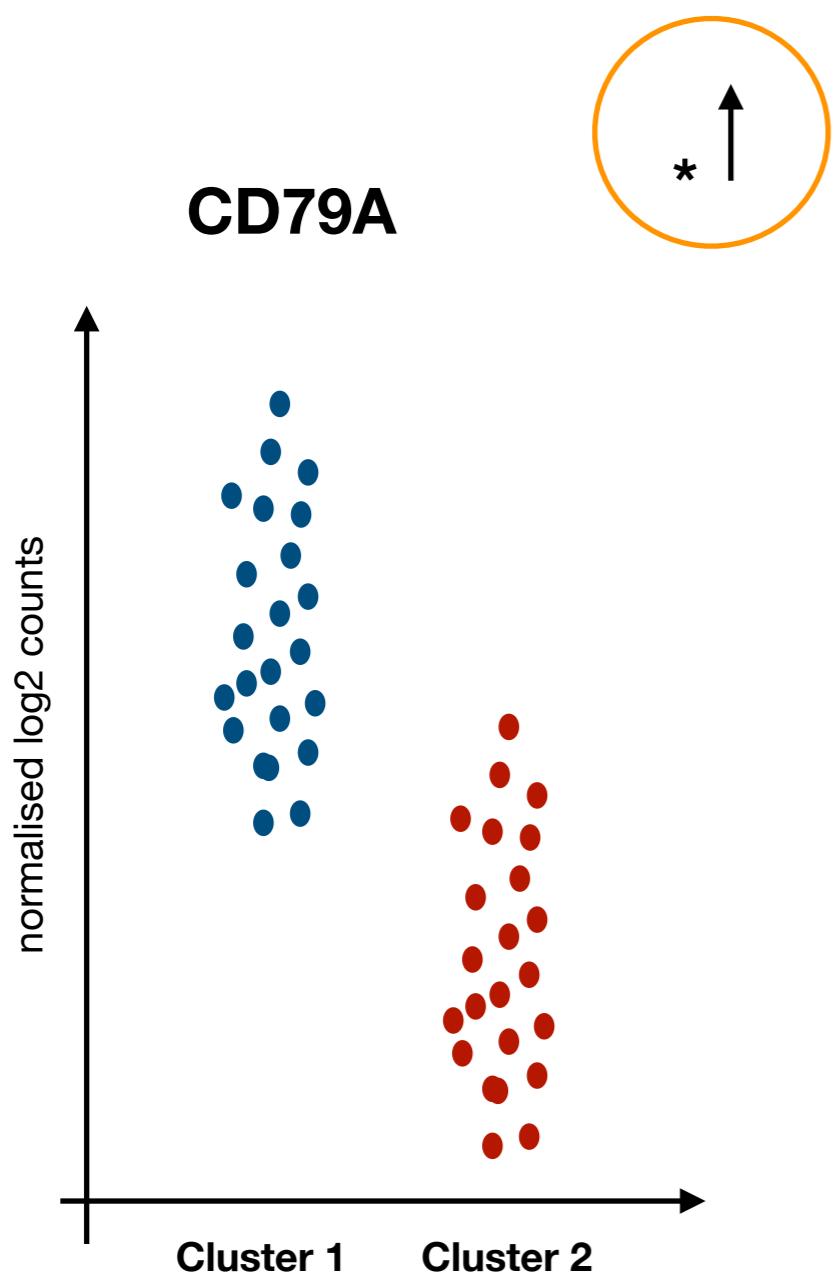
**MS41A**



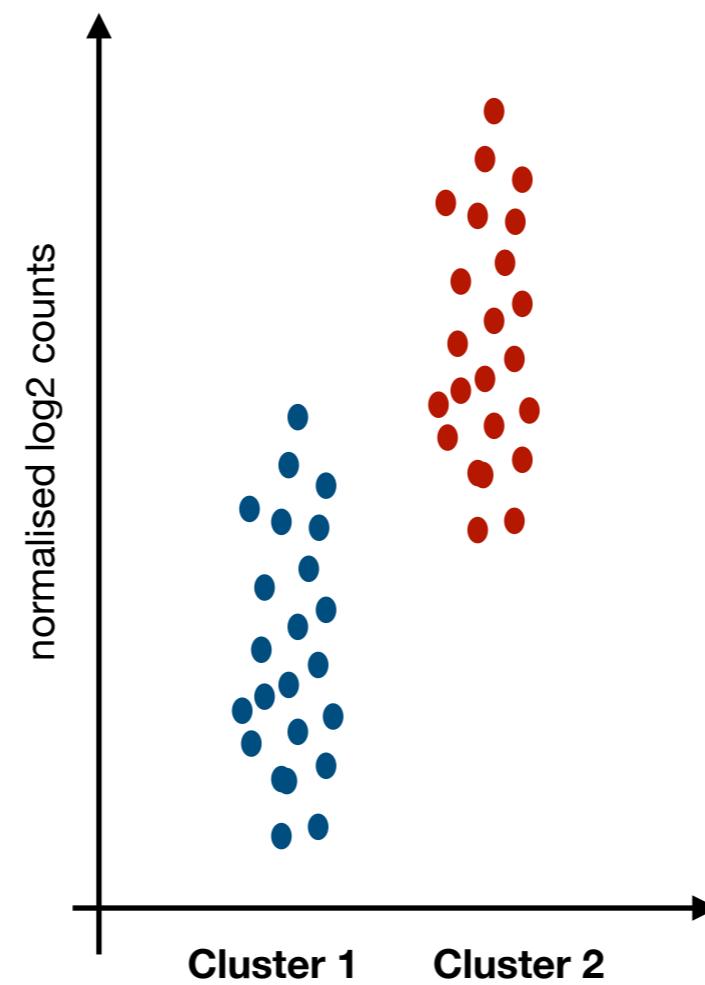
**FOX1**



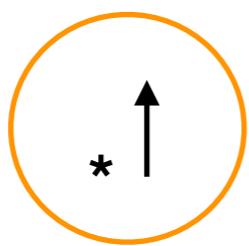
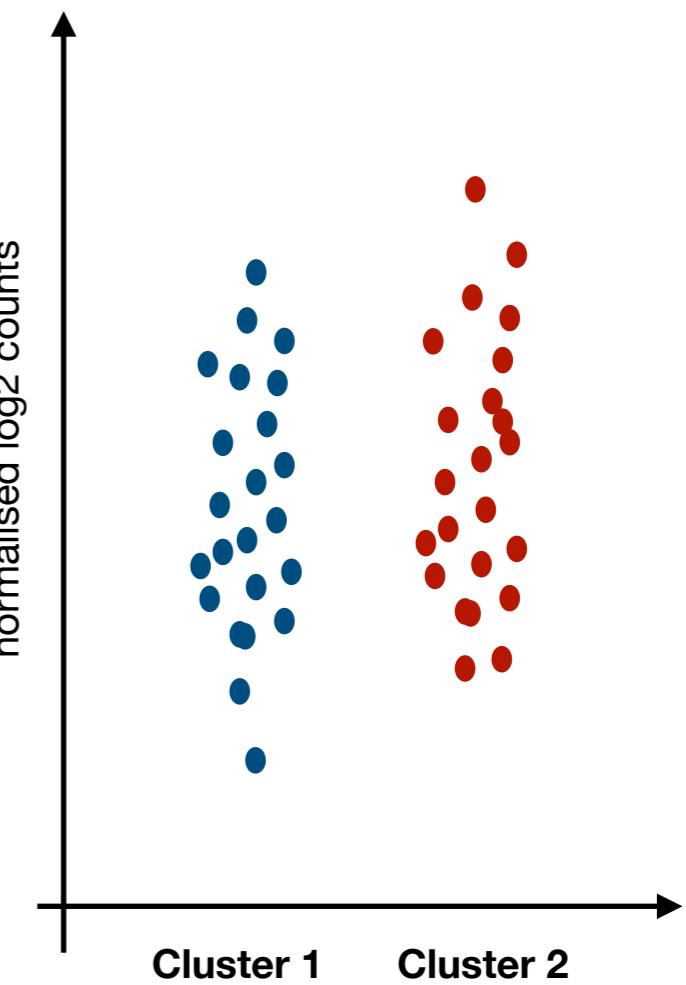
**CD79A**

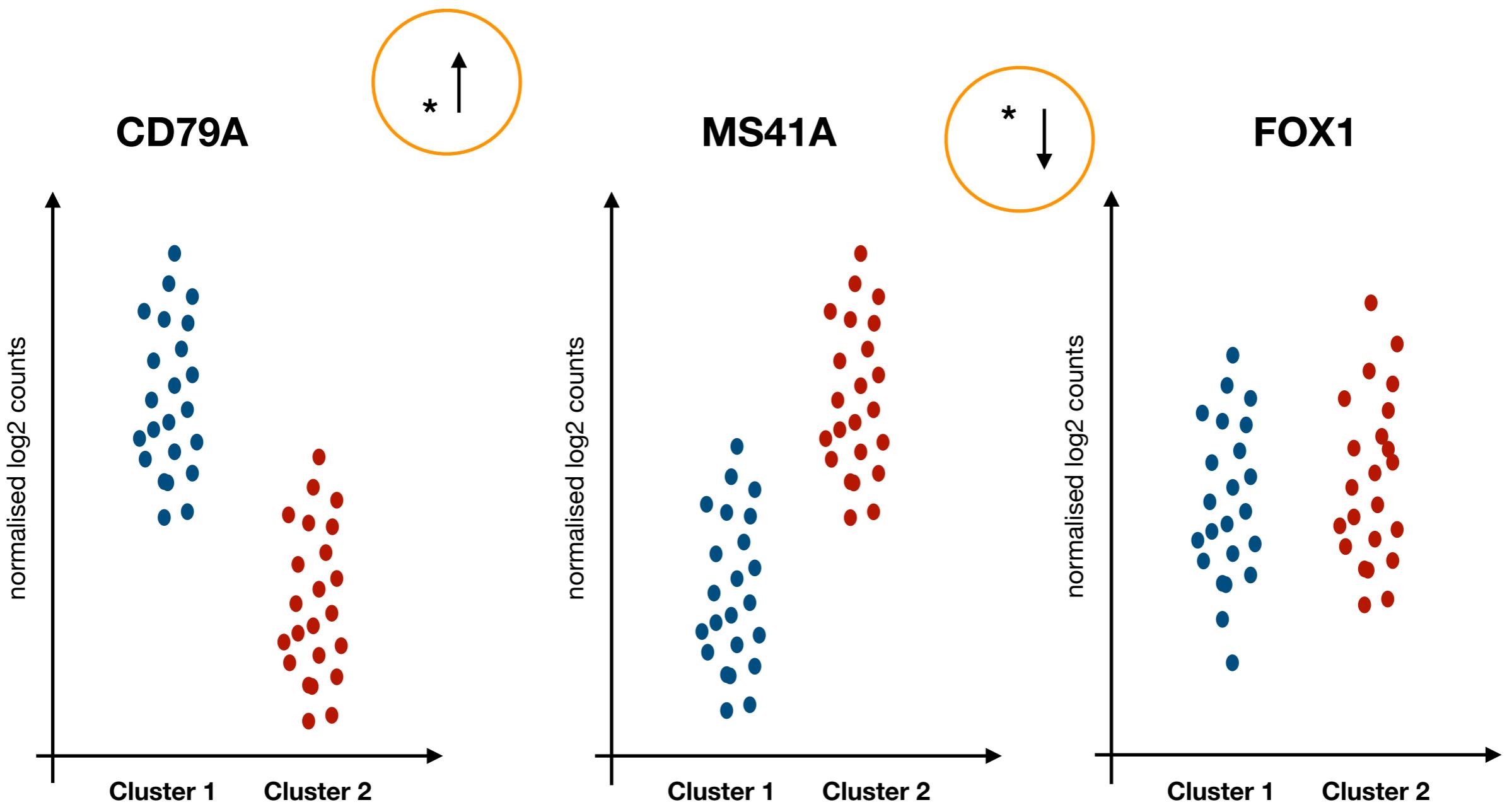


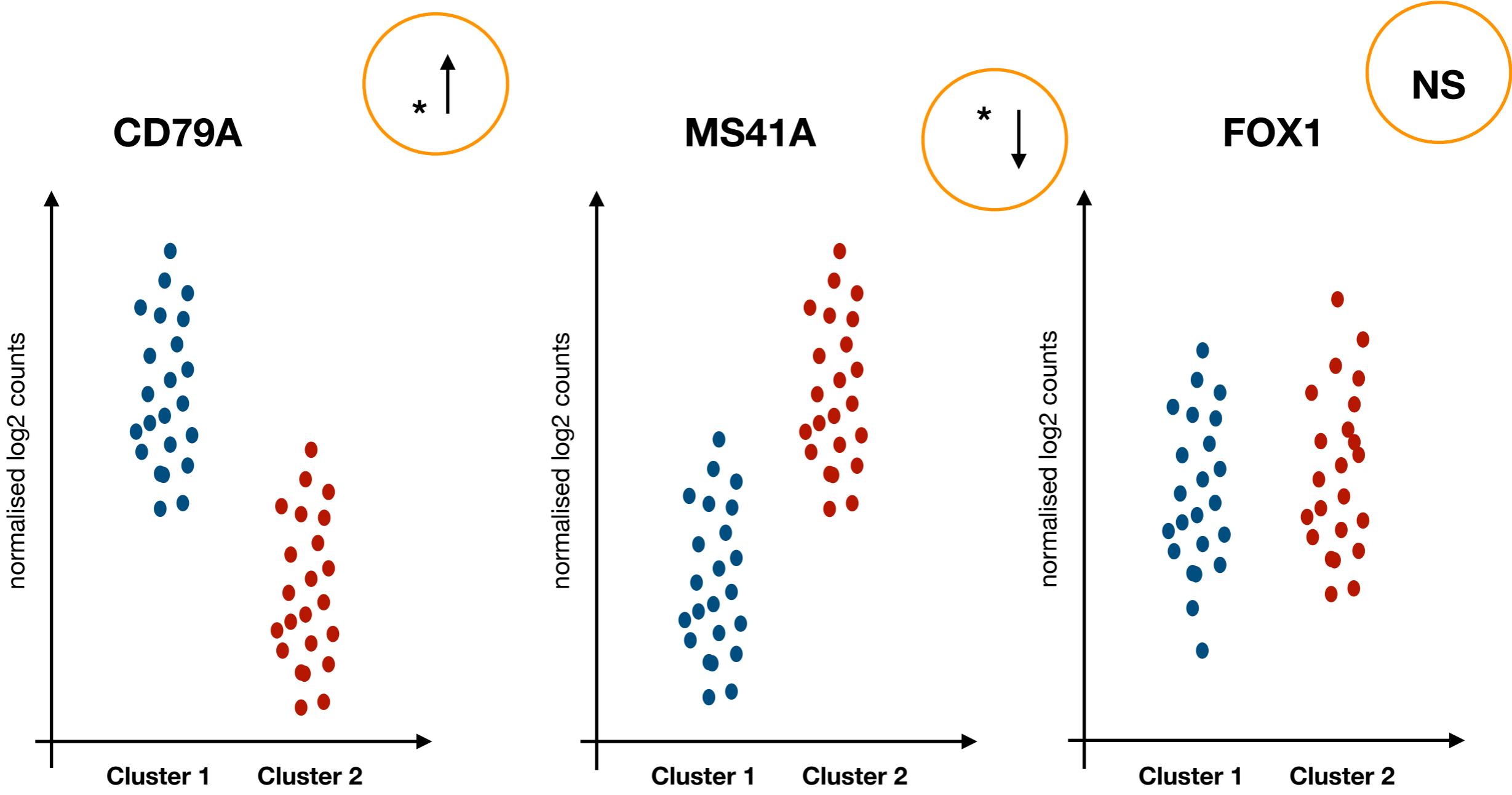
**MS41A**

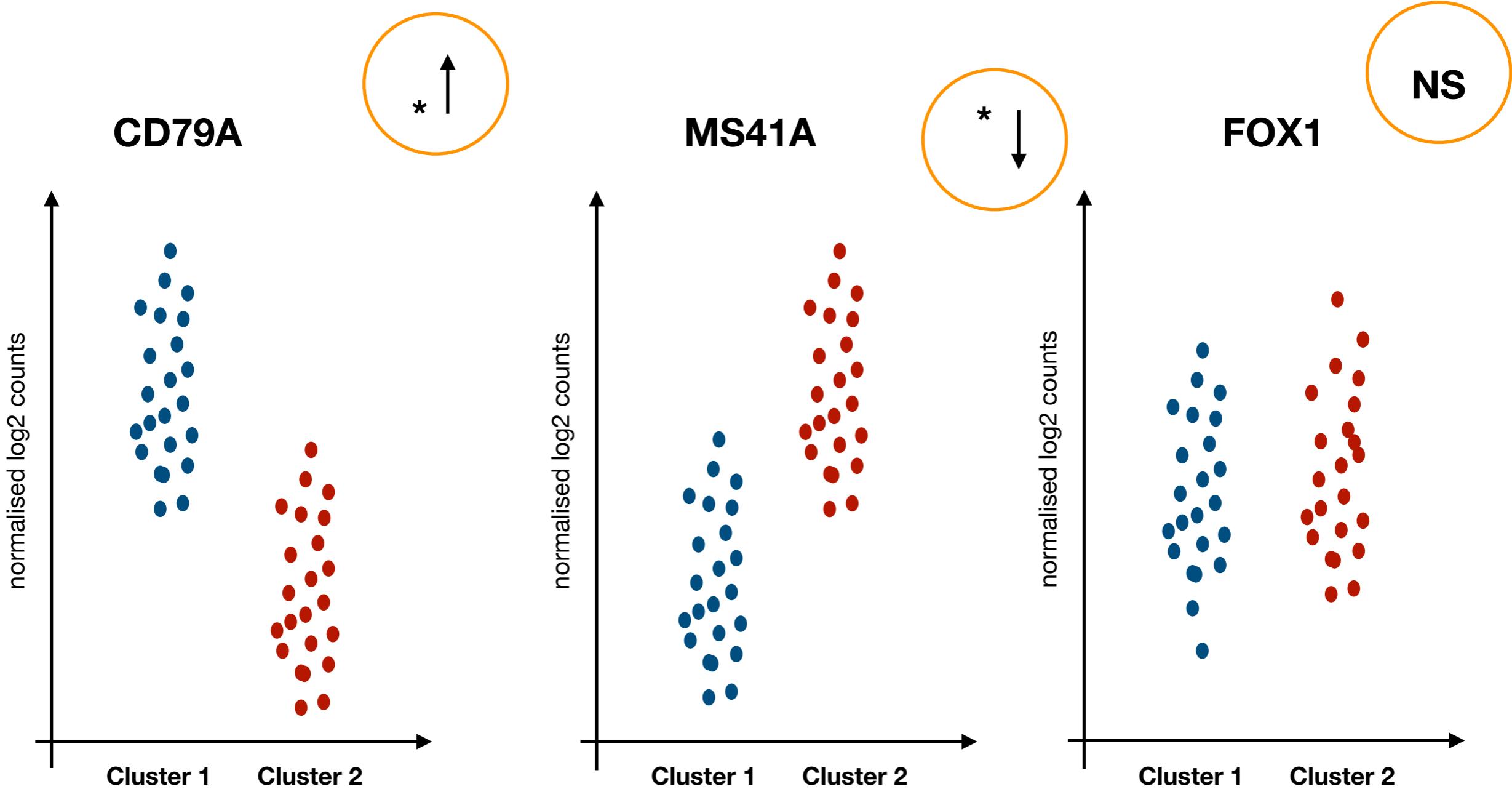


**FOX1**



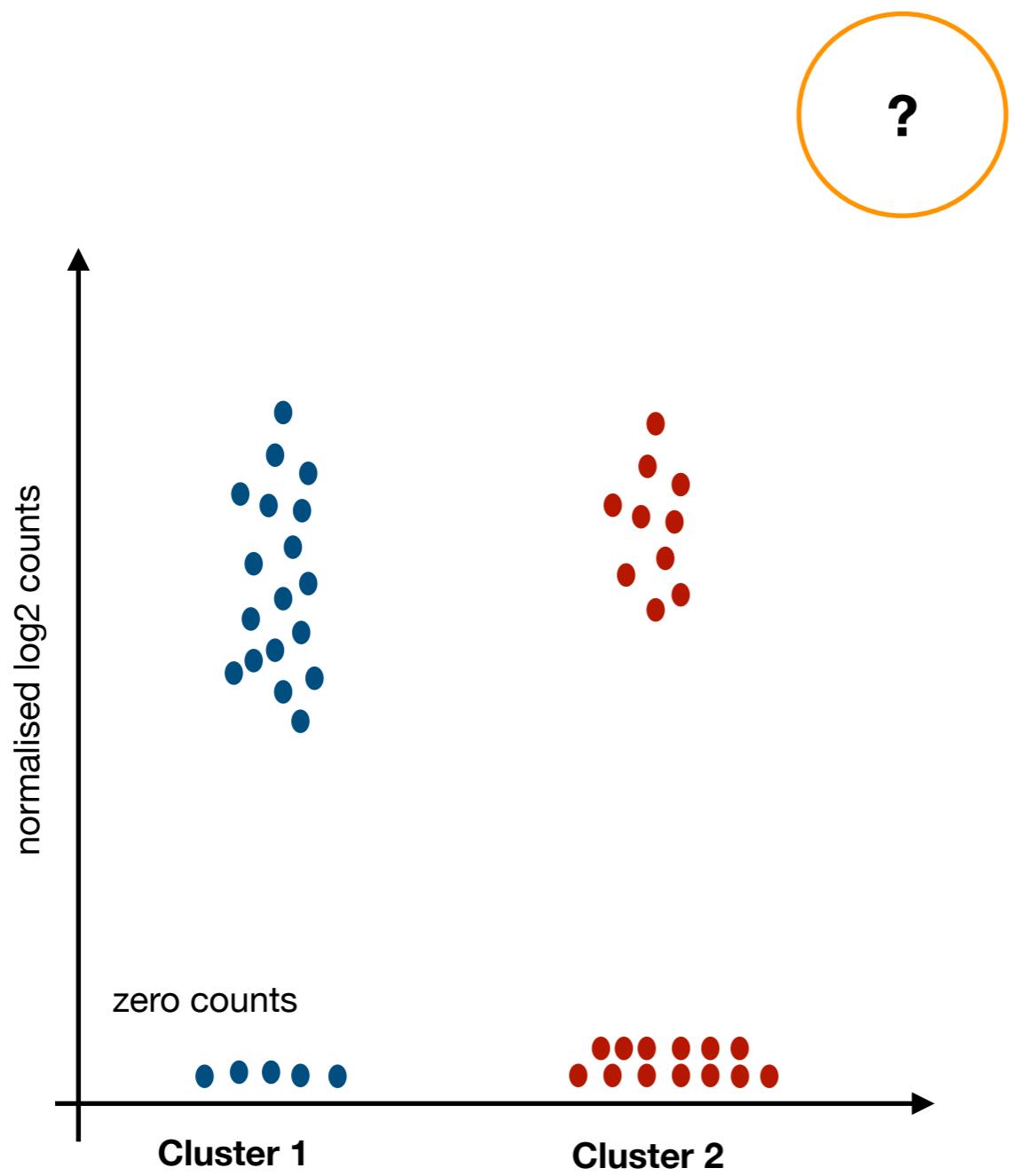


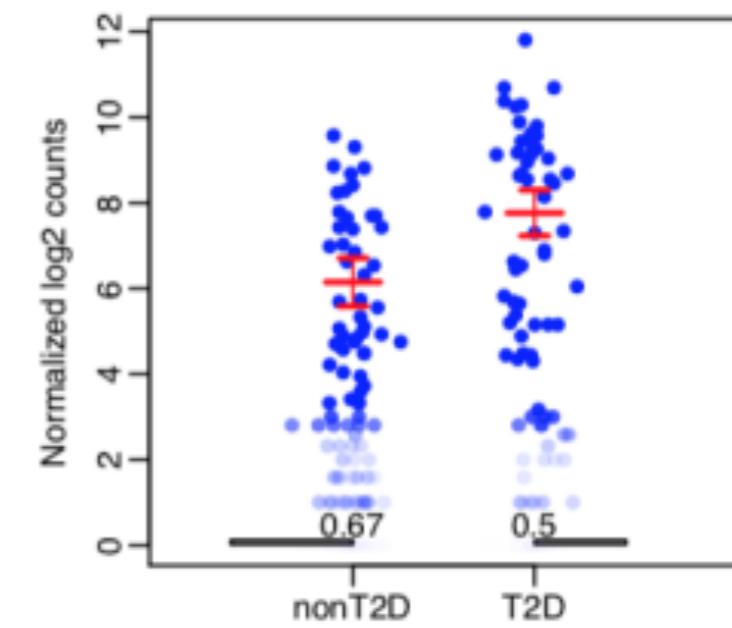
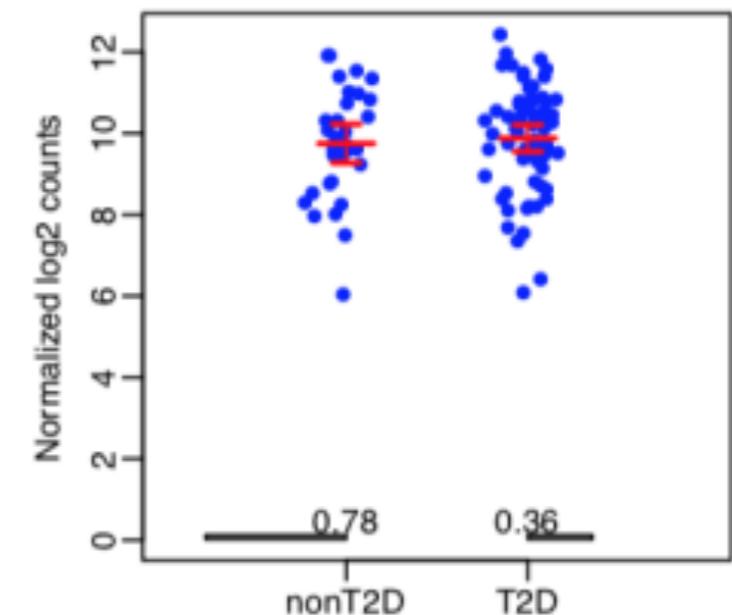
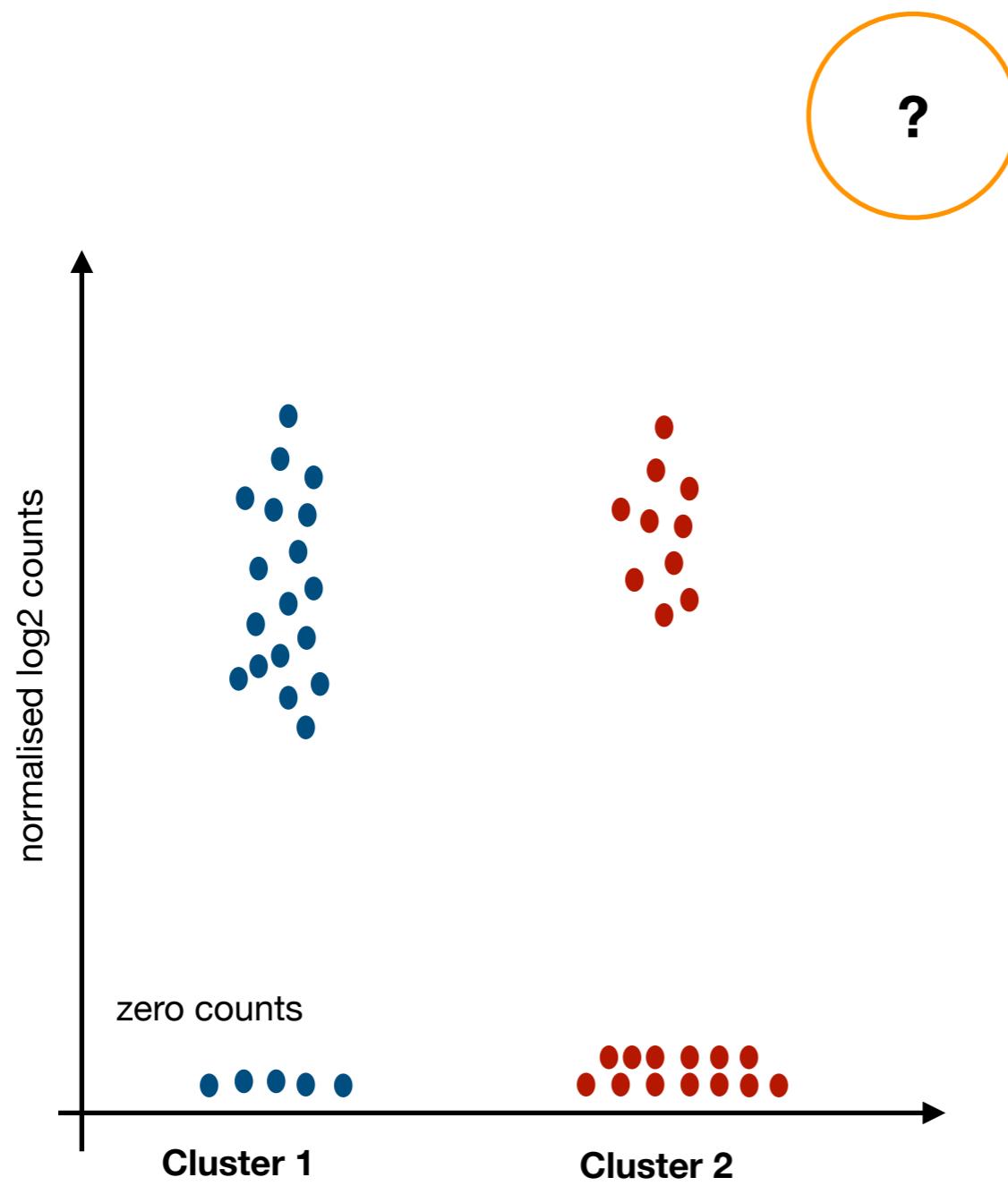




## Differential expression means:

- ❖ taking read count data & performing statistical analysis to discover quantitative changes in expression levels between experimental groups (e.g. clusters)
- ❖ i.e. to decide whether, for a given gene, an observed difference in read counts is significant (greater than it would be expected just due to natural random variation)





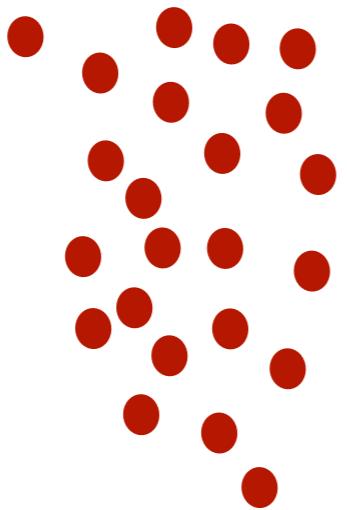
Wu *et al.* (Bioinformatics 2017):  
Two phase differential expression

“...most computational methods still stick with the old mentality of viewing differential expression as a simple “up or down” phenomenon. We advocate that we should fully embrace the features of single cell data, which allows us to observe binary (from Off to On) as well as continuous (the amount of expression) regulations.”

**Cluster 1**



**Cluster 2**



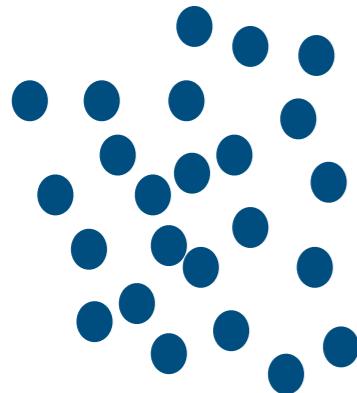
**Cluster 3**



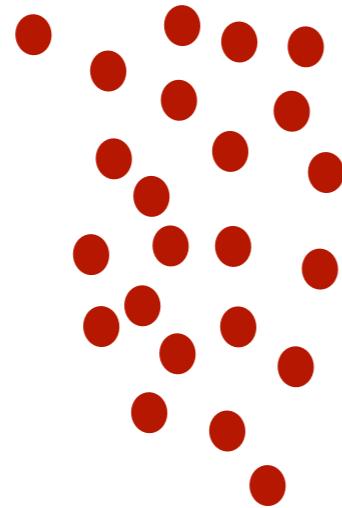
**Cluster 4**



**Cluster 1**



**Cluster 2**



**Cluster 3**

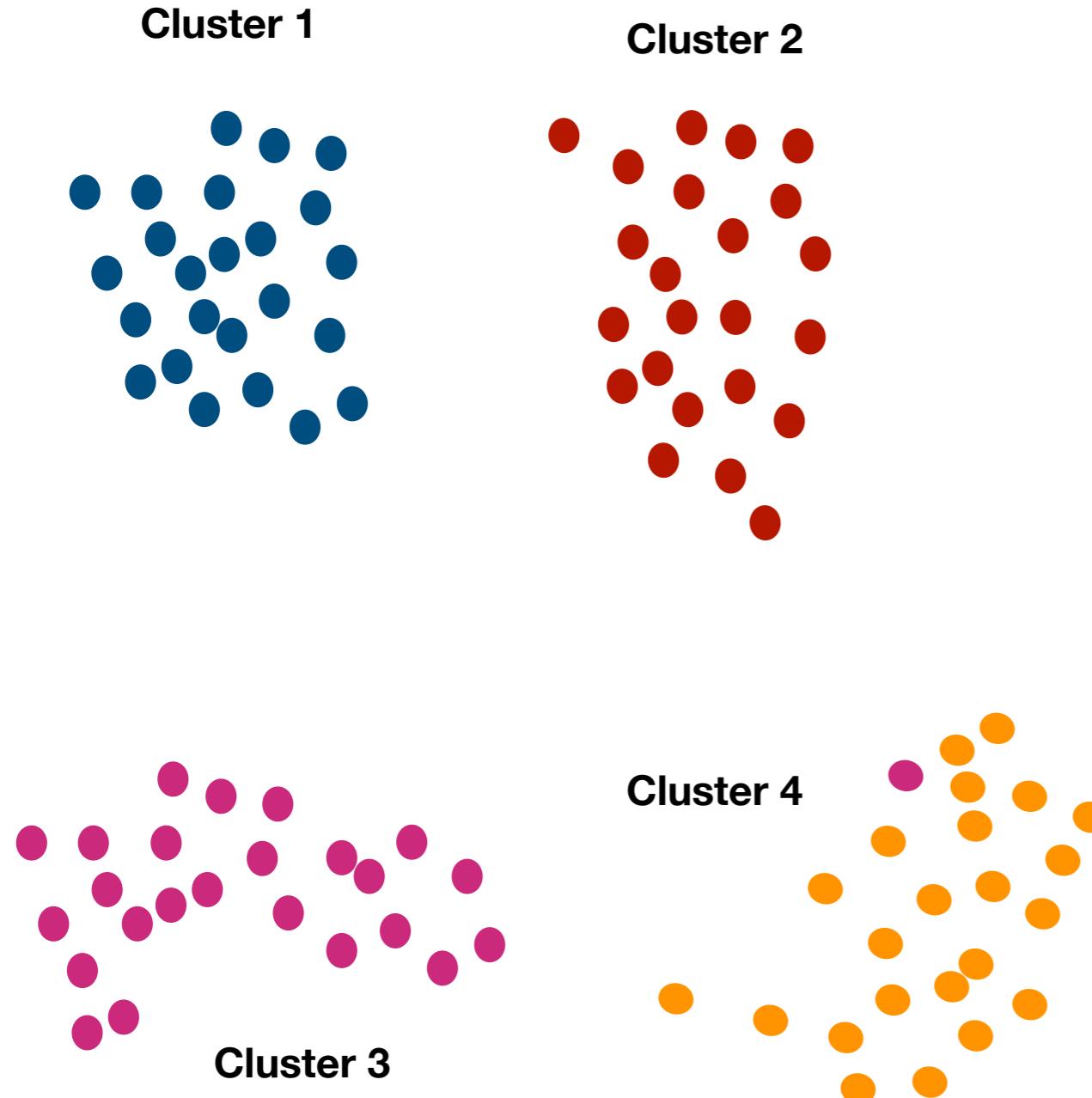


**Cluster 4**



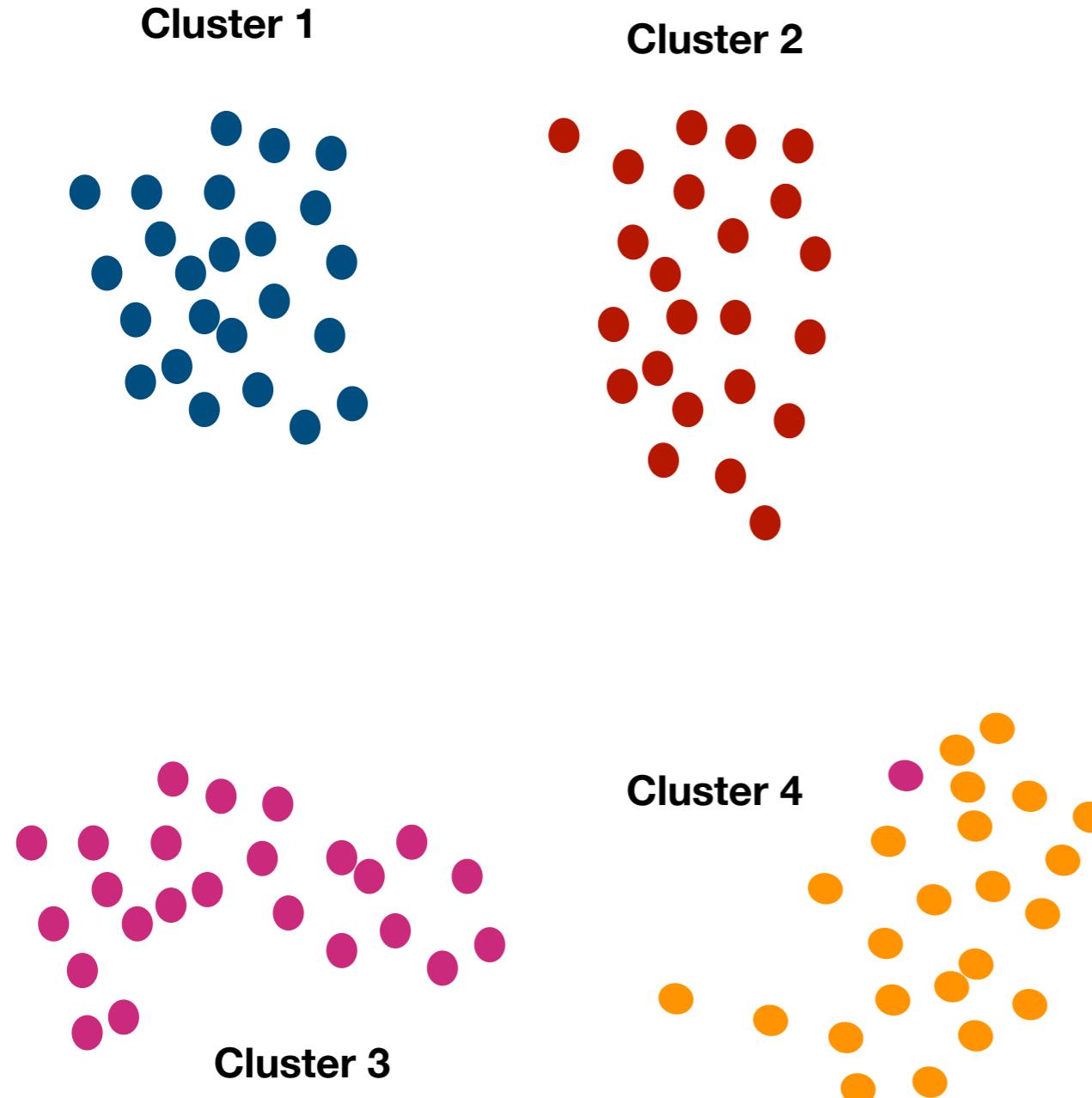
**Differential expression is comparative.  
Common comparisons include:**

- ❖ pairwise cluster comparisons,  
e.g. c1 vs. c2, c2 vs. c3 etc.



**Differential expression is comparative.  
Common comparisons include:**

- ❖ pairwise cluster comparisons,  
e.g. c1 vs. c2, c2 vs. c3 etc.
- ❖ for a given cluster finding ‘marker genes’ that :
  - ❖ DE compared to all cells outside of the cluster
  - ❖ DE compared to at least one other cluster
  - ❖ DE compared to each of the other clusters
  - ❖ DE compare to “most” of the other clusters
  - ❖ DE and up-regulated (up-regulated markers are somehow easier to interpret)



**Differential expression is comparative.  
Common comparisons include:**

- ❖ pairwise cluster comparisons,  
e.g. c1 vs. c2, c2 vs. c3 etc.
- ❖ for a given cluster finding ‘marker genes’ that :
  - ❖ DE compared to all cells outside of the cluster
  - ❖ DE compared to at least one other cluster
  - ❖ DE compared to each of the other clusters
  - ❖ DE compare to “most” of the other clusters
  - ❖ DE and up-regulated (up-regulated markers are somehow easier to interpret)
- ❖ cell-type comparisons (if cell type is known) (with and without clustering)



❖ **Common methods**

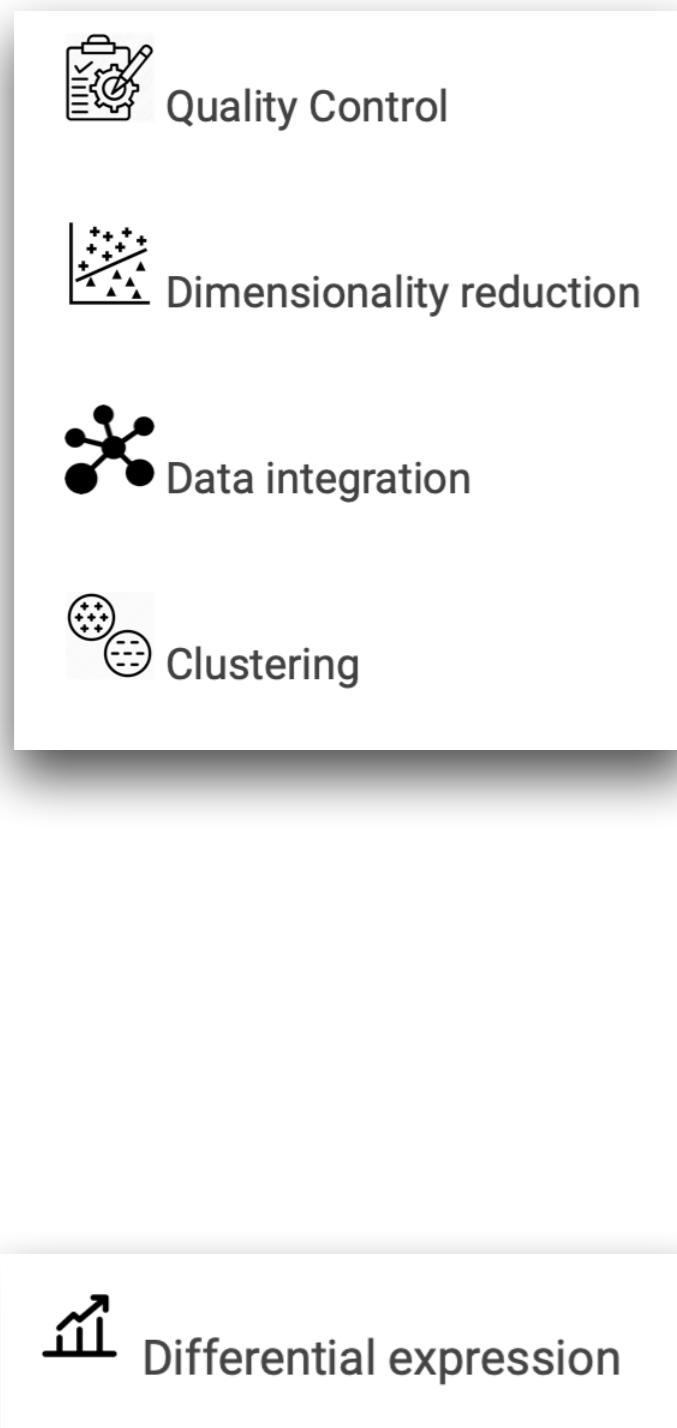
# Context

# Context



- ❖ Setting-up data
- ❖ Quality control and removal of “problematic “ cells
- ❖ Classification of cell cycle phase
- ❖ Normalization
- ❖ Imputations
- ❖ Selection of highly variable genes
- ❖ Data integration
- ❖ K-means / HCL / graph based clustering

# Context



- ❖ Setting-up data
- ❖ Quality control and removal of “problematic “ cells
- ❖ Classification of cell cycle phase
- ❖ Normalization
- ❖ Imputations
- ❖ Selection of highly variable genes
- ❖ Data integration
- ❖ K-means / HCL / graph based clustering



# Functions

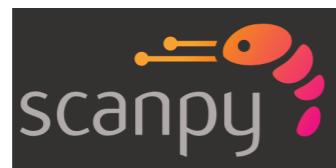
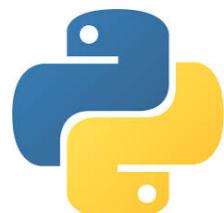


**SEURAT**

`FindAllMarkers()`



`findMarkers()`



`scanpy.tl.rank_genes_groups()`

# FindAllMarkers

From [Seurat v3.1.2](#)  
by [Paul Hoffman](#)

99.99th  
Percentile

## Gene Expression Markers For All Identity Classes

Finds markers (differentially expressed genes) for each of the identity classes in a dataset

## Usage

```
FindAllMarkers(  
  object,  
  assay = NULL,  
  features = NULL,  
  logfc.threshold = 0.25,  
  test.use = "wilcox",  
  slot = "data",  
  min.pct = 0.1,  
  min.diff.pct = -Inf,  
  node = NULL,  
  verbose = TRUE,  
  only.pos = FALSE,  
  max.cells.per.ident = Inf,  
  random.seed = 1,  
  latent.vars = NULL,  
  min.cells.feature = 3,  
  min.cells.group = 3,  
  pseudocount.use = 1,  
  return.thresh = 0.01,  
  ...  
)
```

**test.use**

Denotes which test to use. Available options are:

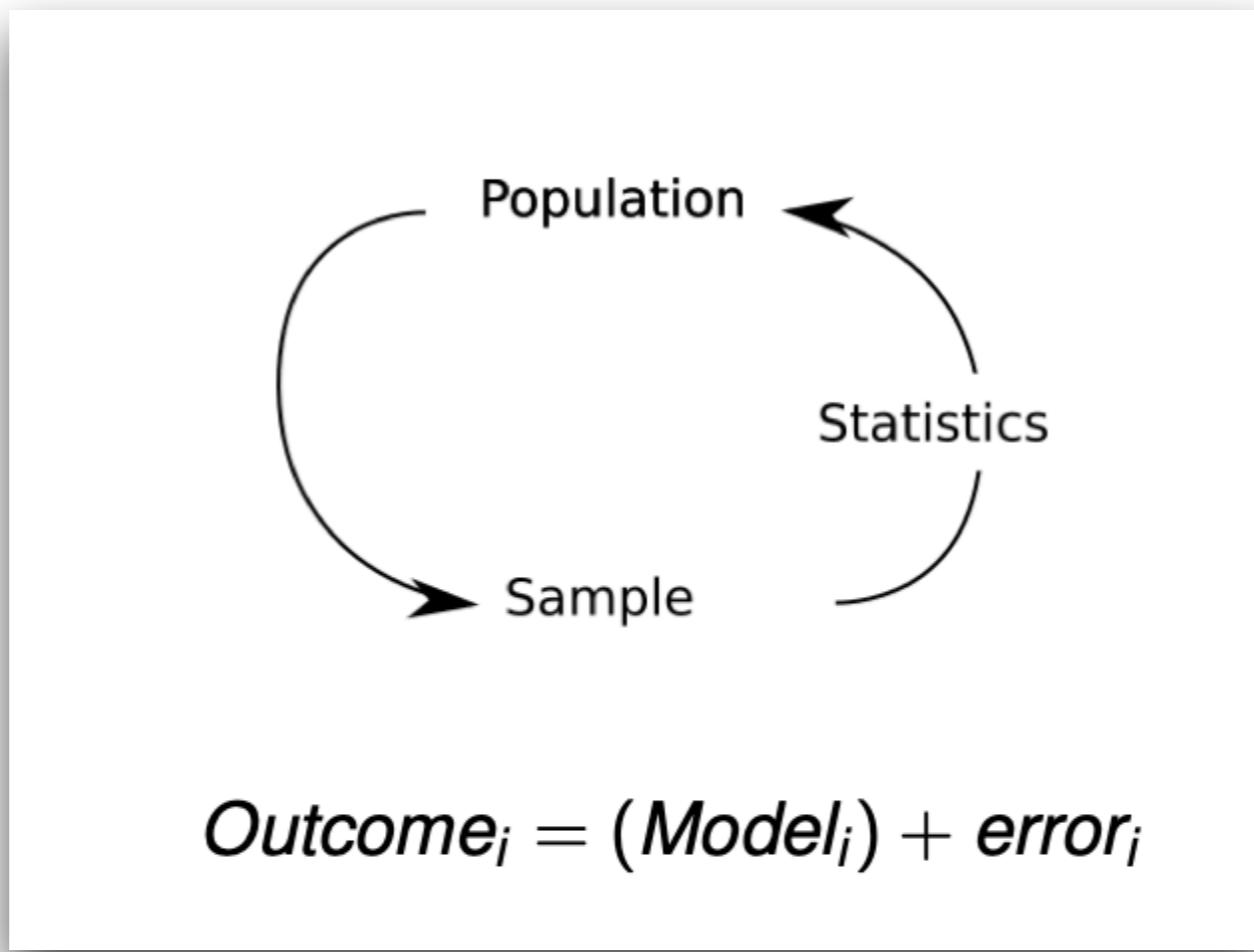
- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.
- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.
- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets
- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. 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" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.
- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

## ► intro to statistical inference

- ❖ i.e. to decide whether, for a given gene, an observed difference in read counts is significant (greater than it would be expected just due to natural random variation)

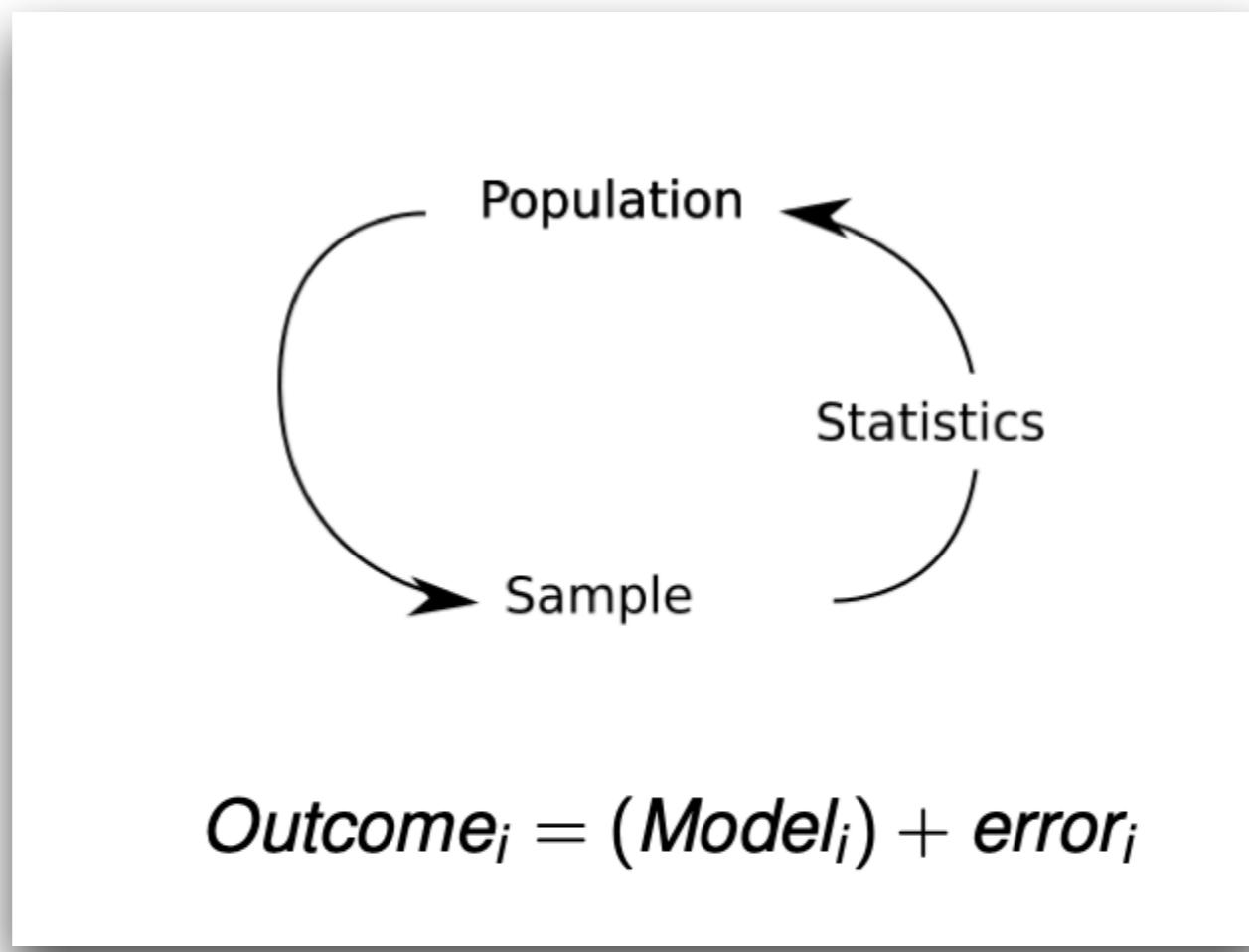
## ► intro to statistical inference

- ❖ i.e. to decide whether, for a given gene, an observed difference in read counts is significant (greater than it would be expected just due to natural random variation)



## ► intro to statistical inference

- ❖ i.e. to decide whether, for a given gene, an observed difference in read counts is significant (greater than it would be expected just due to natural random variation)



- ❖ we collect data on a sample from a much larger population
- ❖ summary statistics lets us to make inferences (conclusions) about the population from which samples was derived
- ❖ as well as predict the outcome given a model fitted to the data

e.g.

**Is there a difference in height between students taking scRNA-seq course in 2019 and 2020?**

e.g.

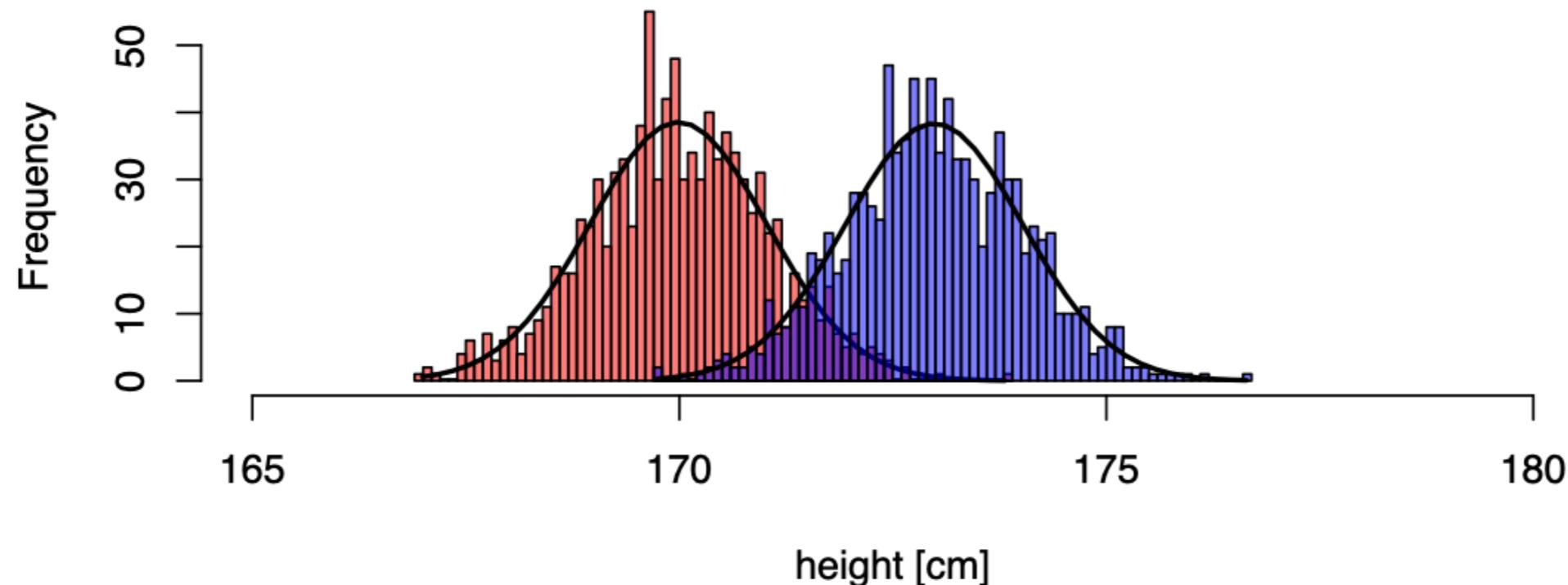
**Is there a difference in height between students taking scRNA-seq course in 2019 and 2020?**

- **H0: null hypothesis: there is no difference in height**
- **H1: alternative hypothesis: difference of means is not equal to 0**

e.g.

**Is there a difference in height between students taking scRNA-seq course in 2019 and 2020?**

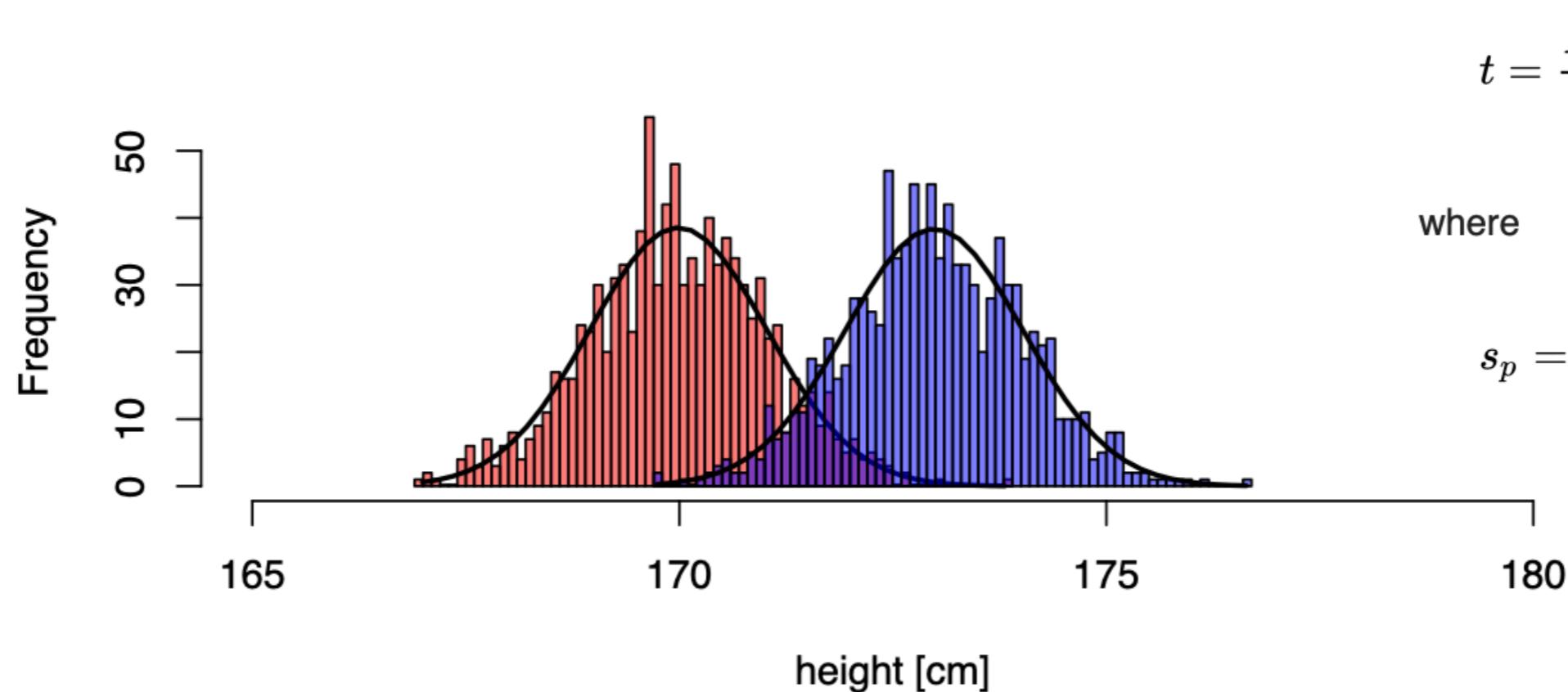
- **H<sub>0</sub>: null hypothesis: there is no difference in height**
- **H<sub>1</sub>: alternative hypothesis: difference of means is not equal to 0**



e.g.

Is there a difference in height between students taking scRNA-seq course in 2019 and 2020?

- **H<sub>0</sub>: null hypothesis: there is no difference in height**
- **H<sub>1</sub>: alternative hypothesis: difference of means is not equal to 0**



$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$$

where

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}.$$

The observed value, here of mean difference, form basis of observed test statistics. A test statistics enables us to carry out a hypothesis test, which is a formal procedure to decide between the null and alternative hypotheses.

**Important implication:**

**The better model fits to the data the better (more accurate) statistics**

## **Generic non-parametric methods**

**when we cannot fit a model to our data**

## **Generic non-parametric methods**

**when we cannot fit a model to our data**

**e.g. Wilcoxon rank-sum test, Kruskal-Wallis, Kolmogorov-Smirnov test**

## **Generic non-parametric methods**

**when we cannot fit a model to our data**

**e.g. Wilcoxon rank-sum test, Kruskal-Wallis, Kolmogorov-Smirnov test**

- ✓ non-parametric test generally convert observed expression values to ranks
- ✓ they test whether the distribution of ranks for one group are significantly different from the distribution of ranks for the other group

## Generic non-parametric methods

when we cannot fit a model to our data

e.g. Wilcoxon rank-sum test, Kruskal-Wallis, Kolmogorov-Smirnov test

- ✓ non-parametric test generally convert observed expression values to ranks
  - may fail in presence of large number of tied values, such as the case of dropouts (zeros) in scRNA-seq
  - if the conditions for a parametric test hold, then it will be typically more powerful than a non-parametric test
- ✓ they test whether the distribution of ranks for one group are significantly different from the distribution of ranks for the other group

## Generic non-parametric methods

when we cannot fit a model to our data

e.g. Wilcoxon rank-sum test, Kruskal-Wallis, Kolmogorov-Smirnov test

- ✓ non-parametric test generally convert observed expression values to ranks
  - may fail in presence of large number of tied values, such as the case of dropouts (zeros) in scRNA-seq
  - if the conditions for a parametric test hold, then it will be typically more powerful than a non-parametric test
- ✓ they test whether the distribution of ranks for one group are significantly different from the distribution of ranks for the other group

Gene-wise null hypothesis:

it is equally likely that a randomly selected cell from group 1 will have higher or lower expression of the gene than a randomly selected cell from group 2

**test.use**

Denotes which test to use. Available options are:



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.



- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.
- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets
- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. 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" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.
- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

**test.use**

Denotes which test to use. Available options are:



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.



- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.

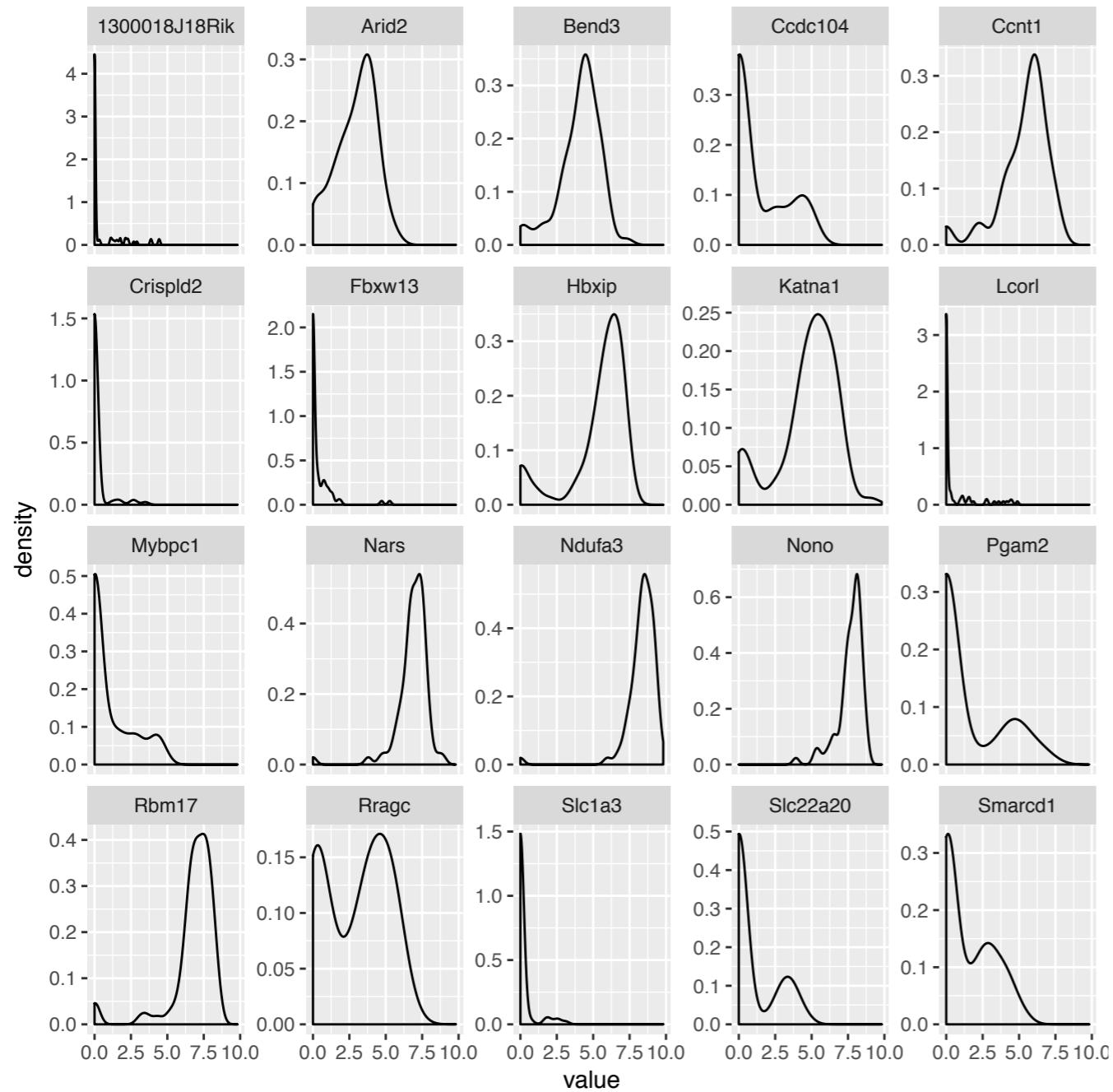


- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets

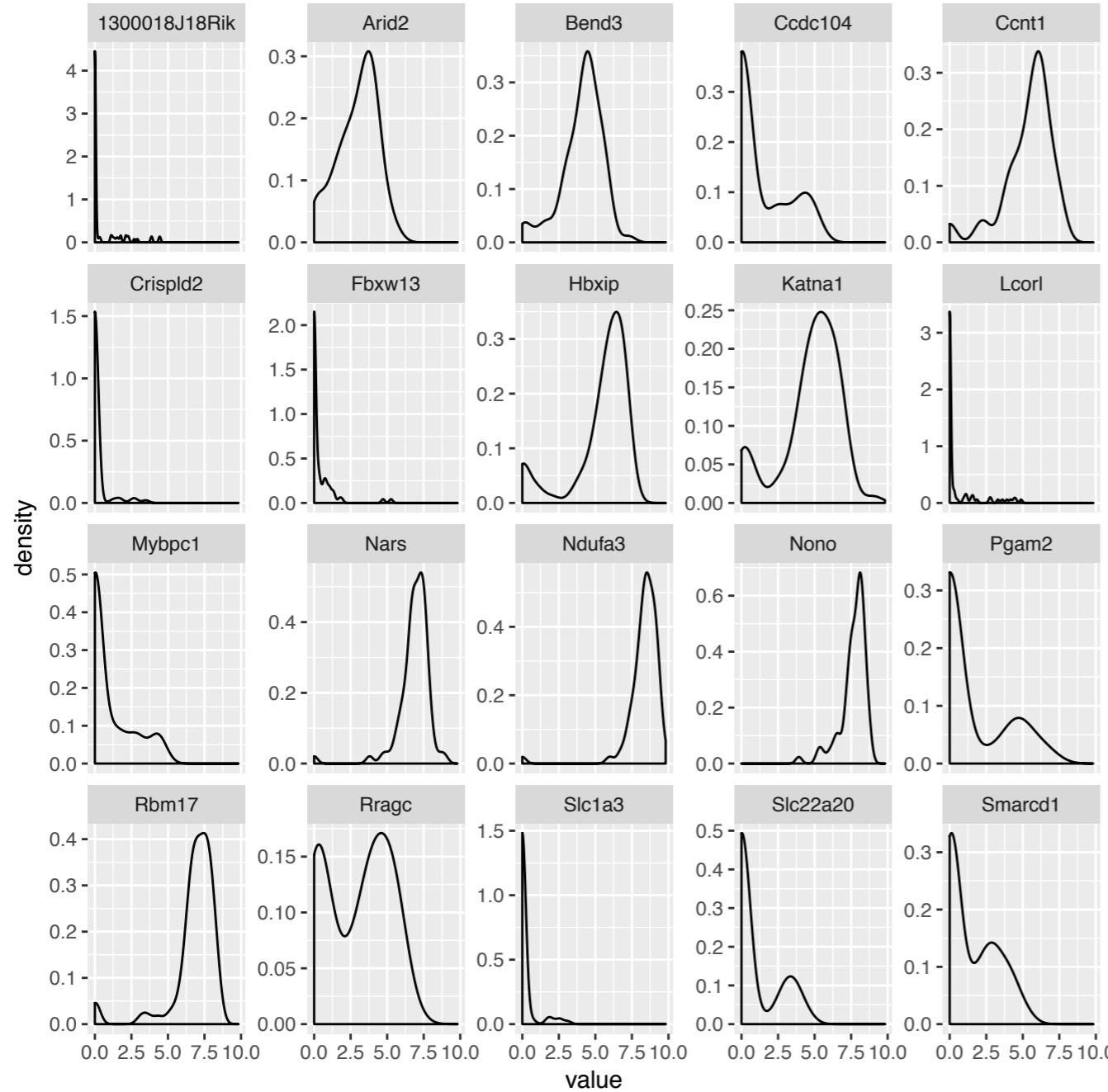


- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. Use only for UMI-based datasets
- "LR" : Uses a logistic regression framework to determine differentially expressed genes. Constructs a logistic regression model predicting group membership based on each feature individually and compares this to a null model with a likelihood ratio test.
- "MAST" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.
- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

## Why special distributions?



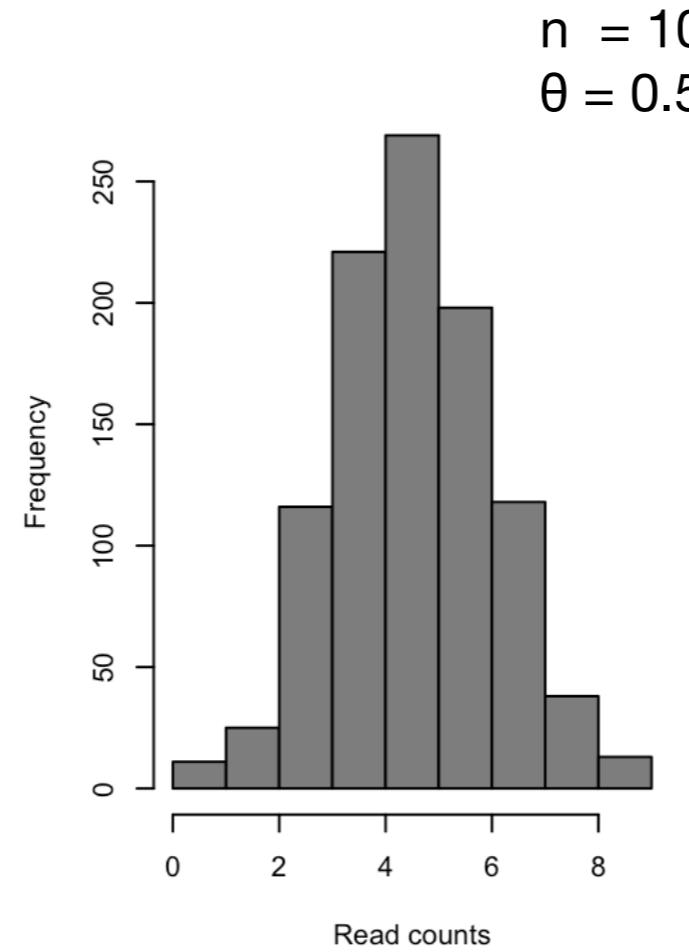
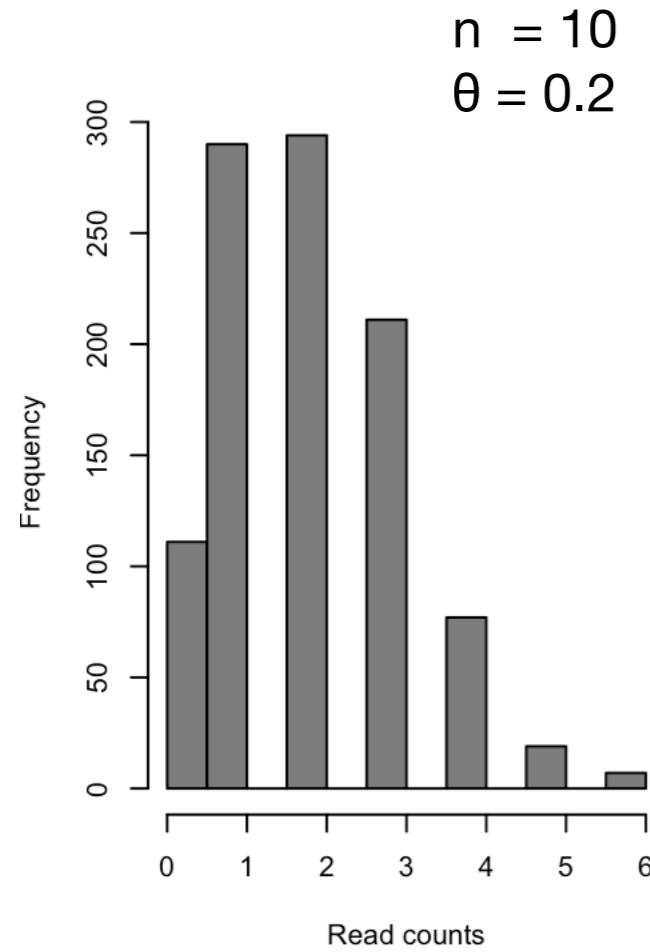
## Why special distributions?



- ❖ high noise levels (technical and biological factors)
- ❖ low library sizes
- ❖ low amount of available mRNAs results in amplification biases and “dropout events”
- ❖ 3' bias, partial coverage, uneven depth
- ❖ stochastic nature of transcription
- ❖ multimodality in gene expression; presence of multiple possible cell states within a cell population

## **What kind of distributions?**

## Binomial



$Bi(n, \theta)$

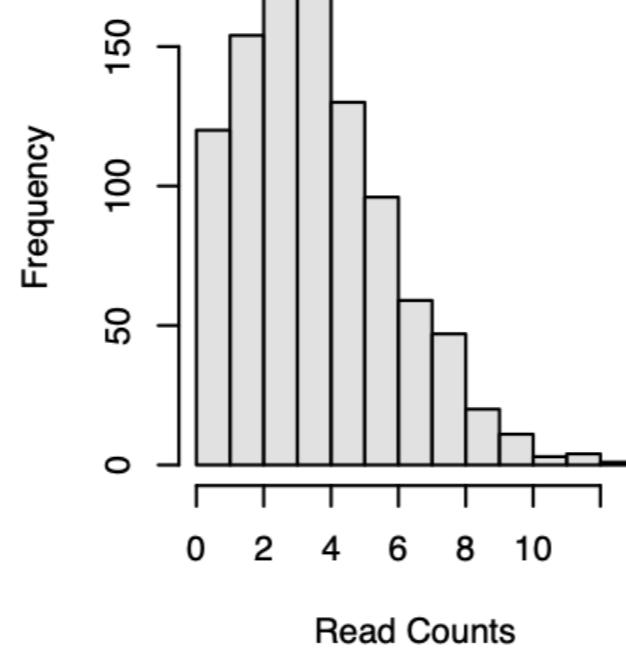
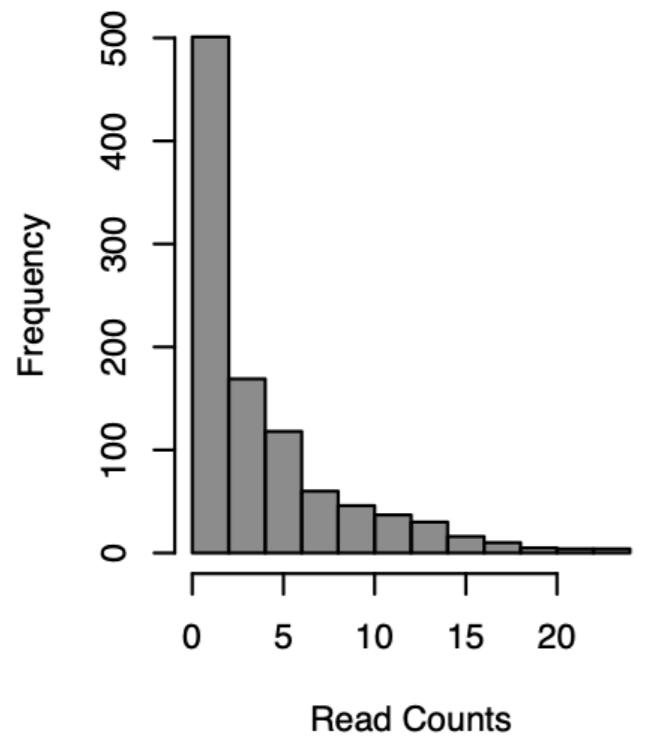
discrete probability distribution of the number of success in a sequence of  $n$  independent experiments;  
 $\theta$  - probability of success

Used to compare proportions of zeros.

Gene-wise null hypothesis:  
probability of being expressed is the same in group 1 and group 2

avail in scran

## Negative binomial



$$NeBi = (\mu, \delta^2)$$

$$\mu = mu$$

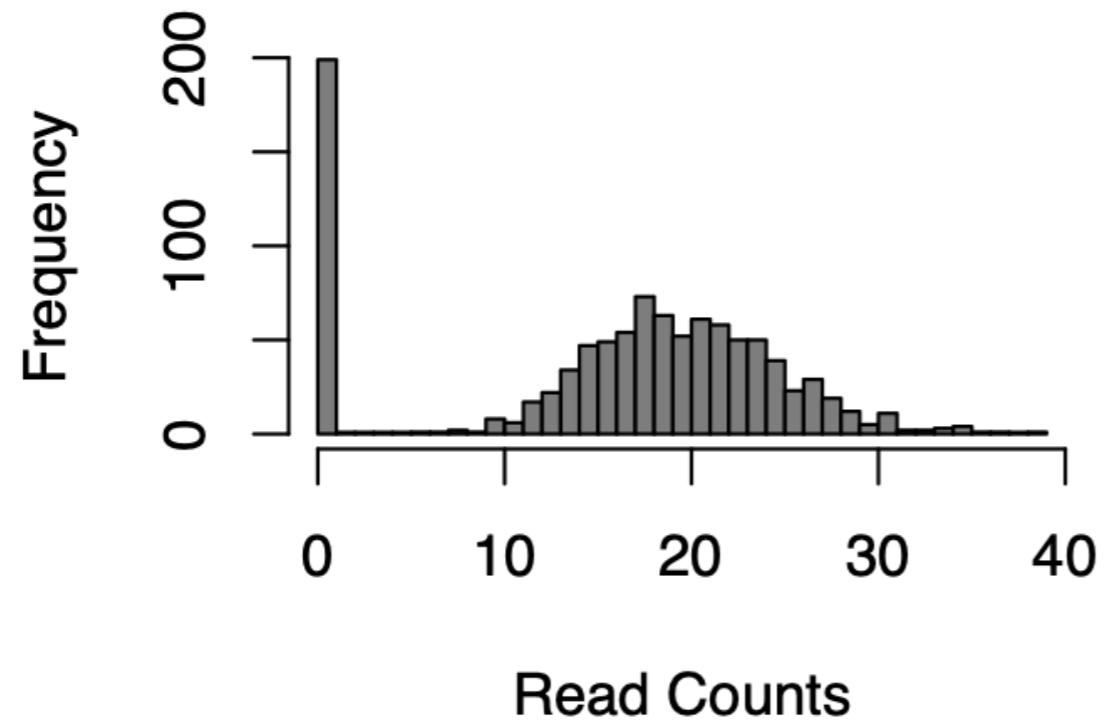
$$\delta^2 = mu + mu^2 / size$$

$\mu$  mean expression

$\delta^2$  dispersion, which is inversely related to the variance

NeBi fits bulk RNA-seq data very well and it is used for most statistical methods designed for such data. In addition, it has been shown to fit the distribution of molecule counts obtained from data tagged by unique molecular identifiers (UMIs) quite well (Grun et al. 2014, Islam et al. 2011).

## zero-inflated negative binomial



$$NeBi = (\mu, \delta^2)$$

$$\mu = mu * (1 - d)$$

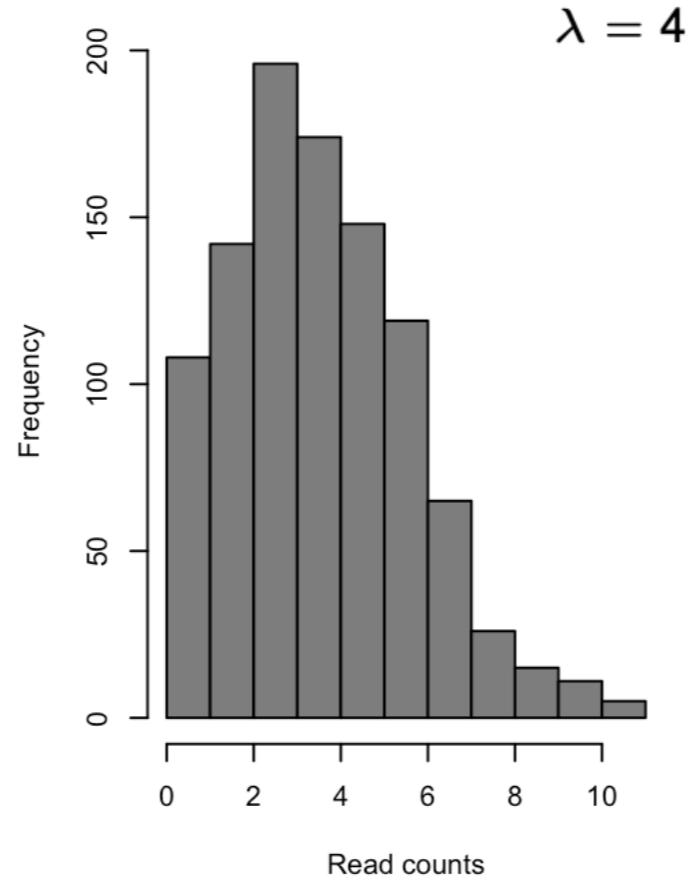
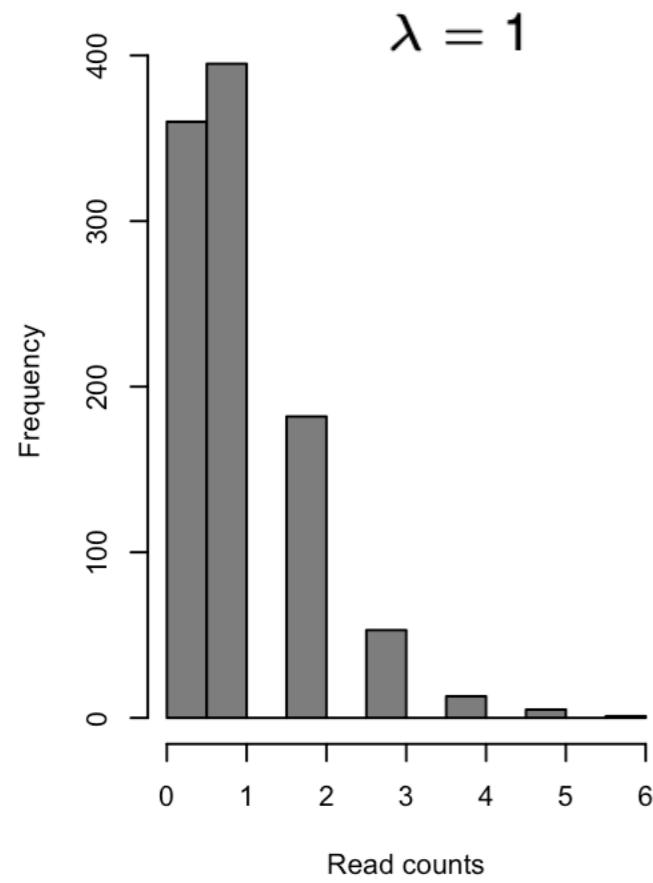
$$\delta^2 = \mu * (1 - d) * (1 + d * \mu + \mu / size)$$

d, dropout rate.

The dropout of a gene is strongly correlated with the mean expression of the gene. Different zero-inflated negative binomial models use different relationships between mean expression and dropout rate.

Implemented in MAST, SCDE

## Poisson



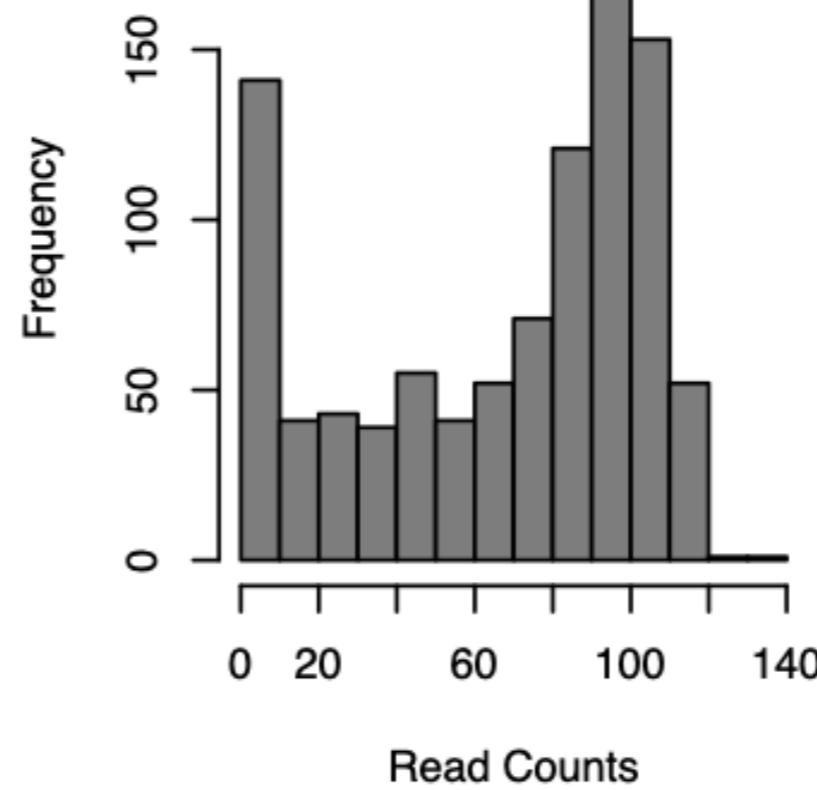
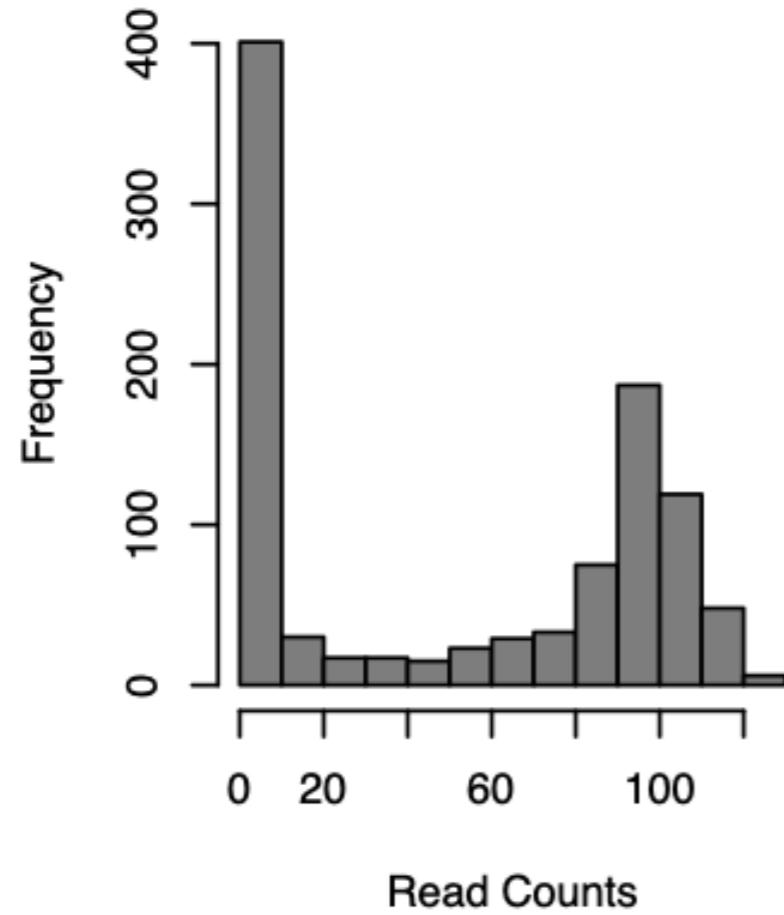
$\text{Poi}(\lambda)$

discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate, *lambda*, and independently of the time since the last event

## Poisson-Beta

$$\mu = g * a / (a + b)$$

$$\delta^2 = g^2 * a * b / ((a + b + 1) * (a + b)^2)$$



- a: the rate of activation of transcription
- b: the rate of inhibition of transcription
- g: the rate of transcript production while transcription is active at the locus

implemented in BPSC

**test.use**

Denotes which test to use. Available options are:



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.



- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.



- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets



- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. 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" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.
- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

**test.use**

Denotes which test to use. Available options are:



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.



- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.



- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets



- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. 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" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.



- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>

## MAST

- uses generalized linear hurdle model
- designed to account for stochastic dropouts and bimodal expression distribution in which expression is either strongly non-zero or non-detectable
- The rate of expression  $Z$ , and the level of expression  $Y$ , are modeled for each gene  $g$ , indicating whether gene  $g$  is expressed in cell  $i$  (i.e.,  $Z_{ig} = 0$  if  $y_{ig} = 0$  and  $z_{ig} = 1$  if  $y_{ig} > 0$ )
- A logistic regression model for the discrete variable  $Z$  and a Gaussian linear model for the continuous variable ( $Y|Z=1$ ):

$$\text{logit}(\Pr(Z_{ig} = 1)) = X_i \beta_g^D$$
$$\Pr(Y_{ig} = Y|Z_{ig} = 1) = N(X_i \beta_g^C, \sigma_g^2), \text{ where } X_i \text{ is a design matrix}$$

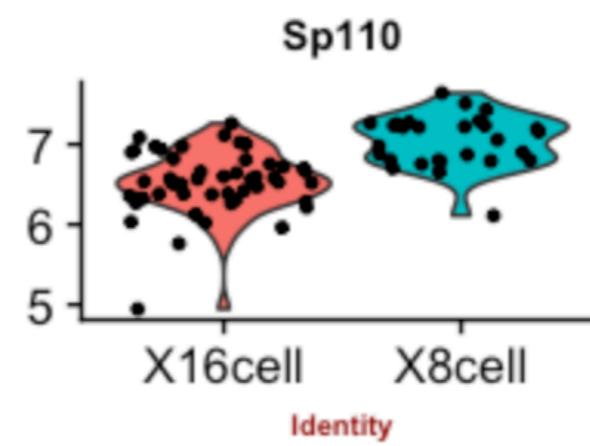
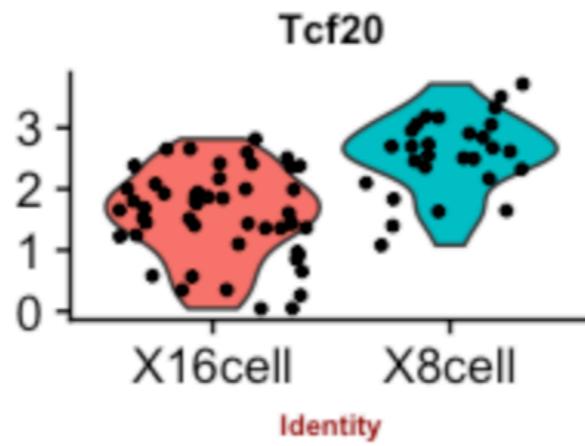
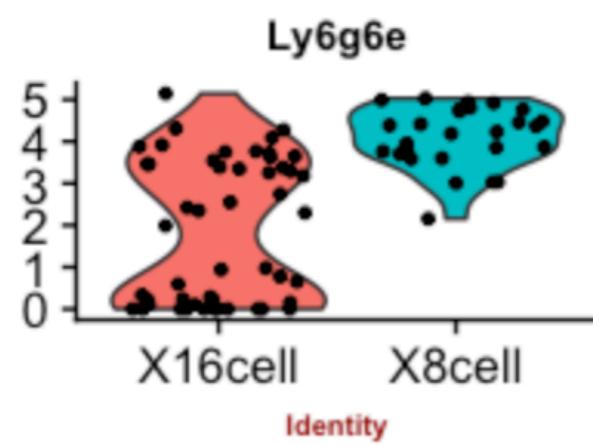
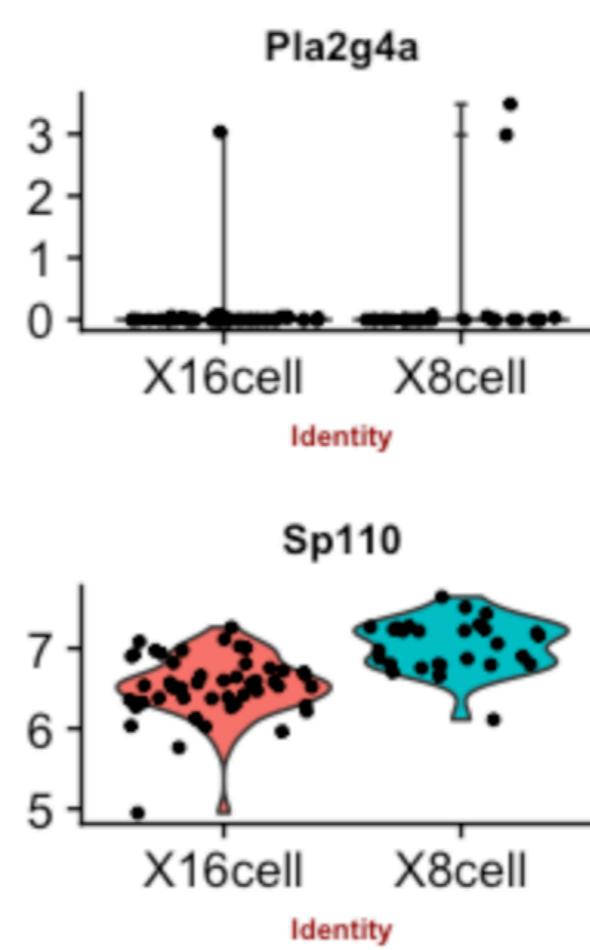
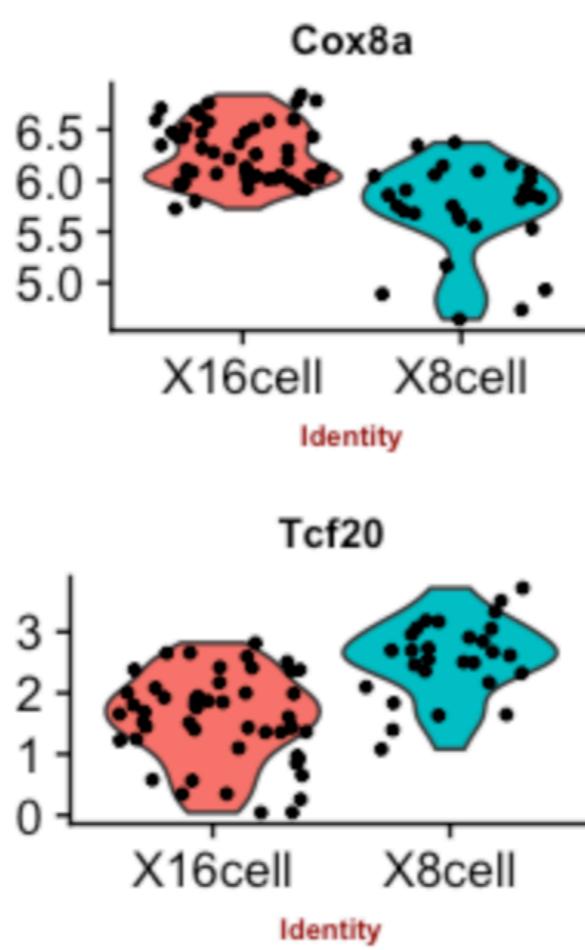
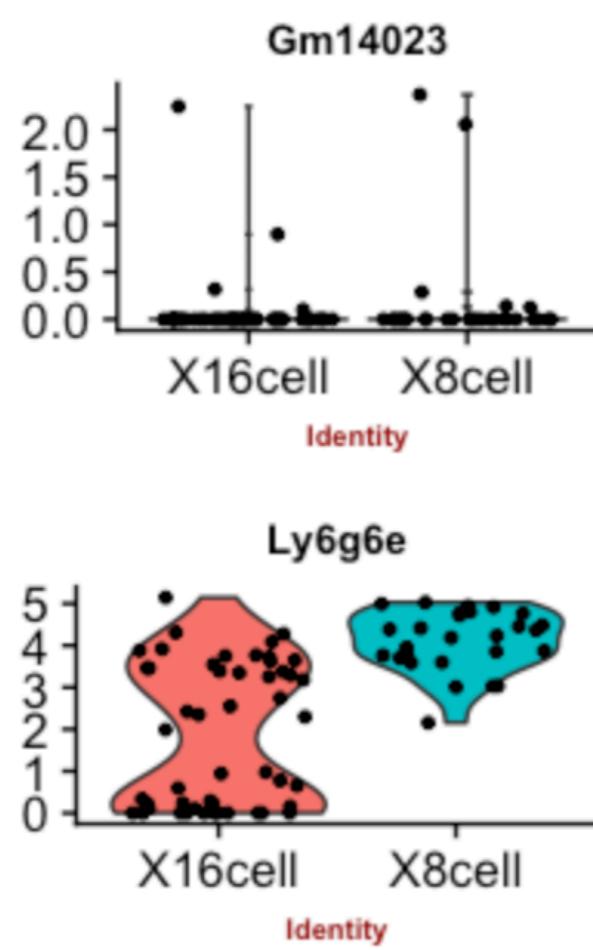
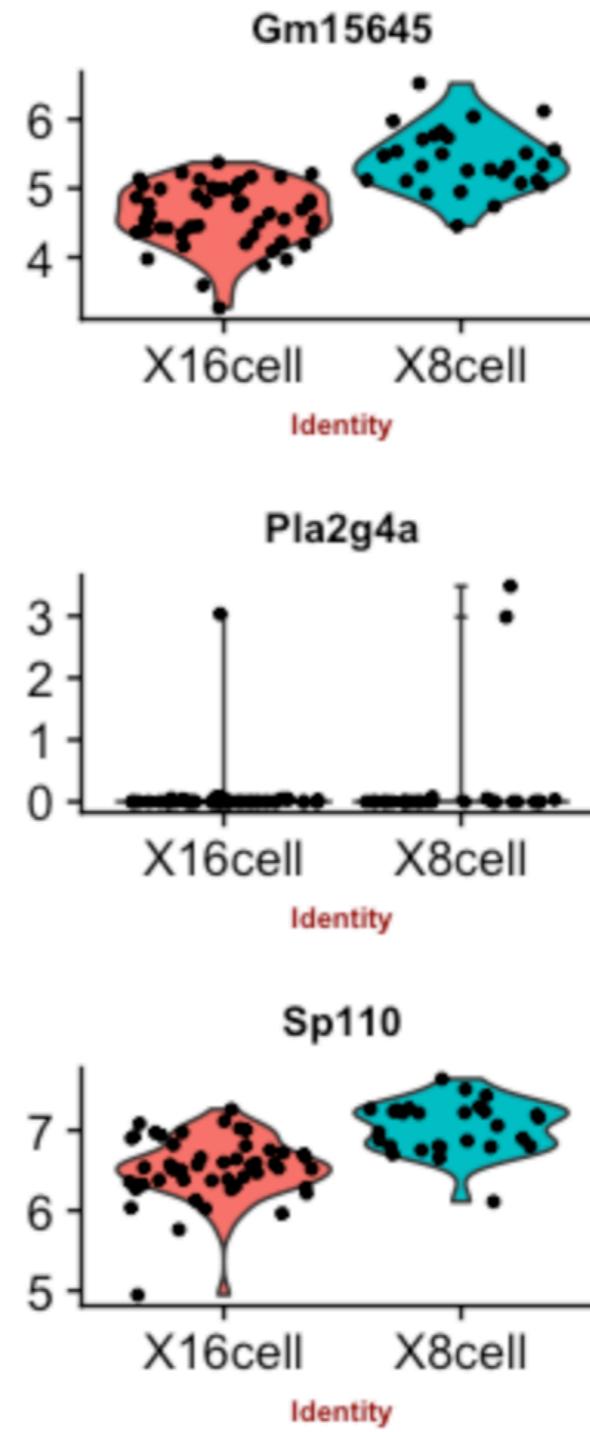
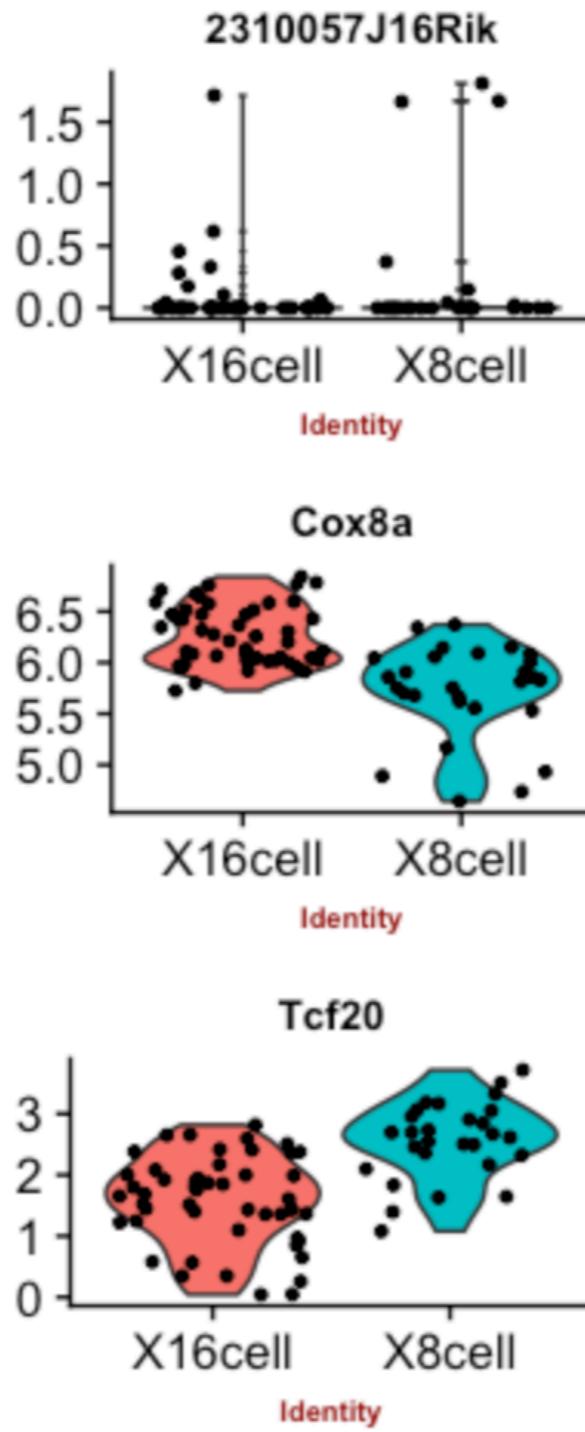
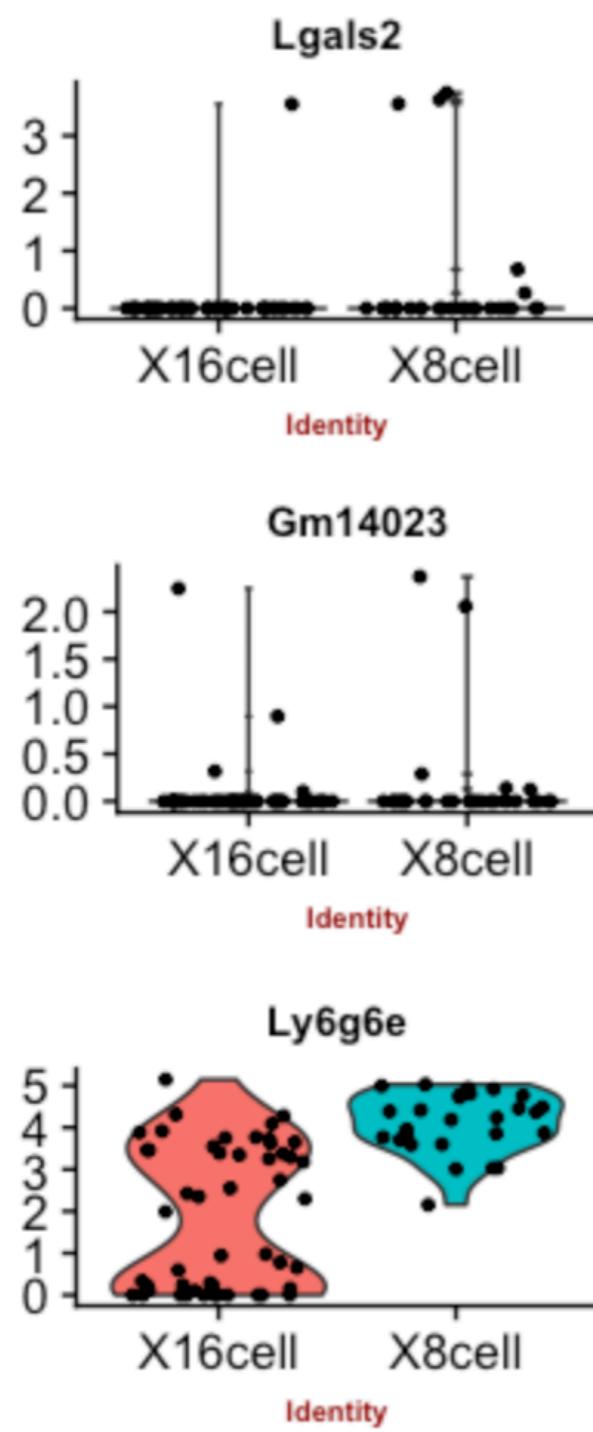
- Model parameters are fitted using an empirical Bayesian framework
  - Allows for a joint estimate of nuisance and treatment effects
  - DE is determined using the likelihood ratio test
-



## So what's really important?

- ❖ Models can get quite complicated but in Seurat / Scran / Scanpy default methods are set to t-test and wilcox
- ❖ It's important to understand what are we trying to compare, e.g. mean expressions, or probability of being expressed
- ❖ It's important to understand the data
- ❖ It's important to assess and validate the results

## What's important: assessing results



## ❖ Performance

**Why is it hard to say which method is best?**

## Performance

### No ground truth data available

- ❖ **Known data:**
  - ❖ using data we know something about to get “positive controls”
- ❖ **Simulated data:**
  - ❖ null-data sets by re-sampling, modelling datasets based on various distributions
- ❖ **Compare:**
  - ❖ comparing between numbers and ranks of DEs
- ❖ **Investigating results:**
  - ❖ how does the expression and distributions of detected DEs look like?

# Performance

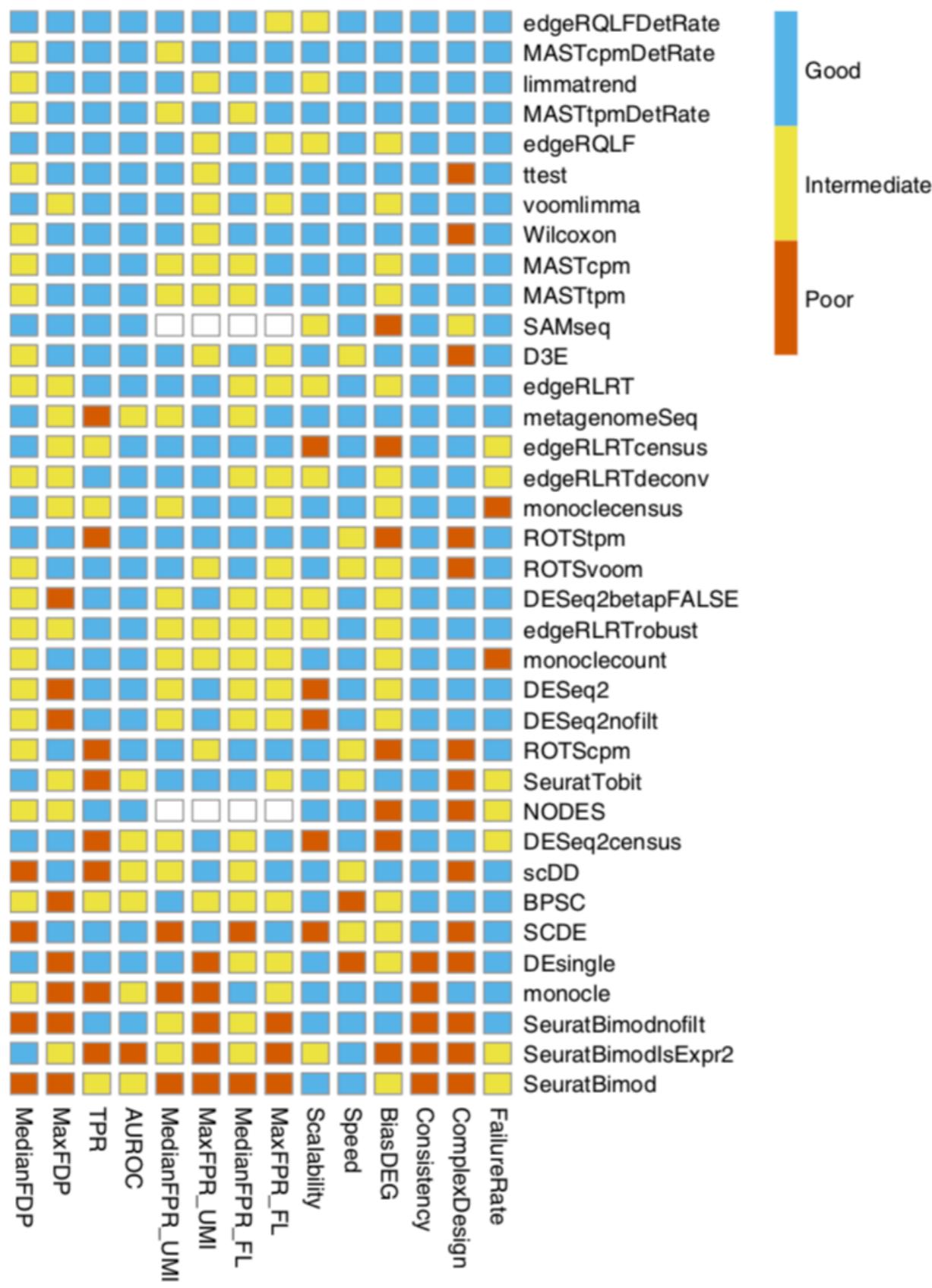
	Short name	Method	Software version	Input
■	BPSC	BPSC	BPSC 0.99.0/1	CPM
■	D3E	D3E	D3E 1.0	raw counts
■	DESeq2	DESeq2	DESeq2 1.14.1	raw counts
■	DESeq2betaPFALSE	DESeq2 without beta prior	DESeq2 1.14.1	raw counts
■	DESeq2census	DESeq2	DESeq2 1.14.1	Census counts
■	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts
■	DEsingle	DEsingle	DEsingle 0.1.0	raw counts
■	edgeRLRT	edgeR/LRT	edgeR 3.19.1	raw counts
■	edgeRLRTcensus	edgeR/LRT	edgeR 3.19.1	Census counts
■	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.19.1, scran 1.2.0	raw counts
■	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.19.1	raw counts
■	edgeRQLF	edgeR/QLF	edgeR 3.19.1	raw counts
■	edgeRQLFDetRate	edgeR/QLF with cellular detection rate as covariate	edgeR 3.19.1	raw counts
■	limmatrend	limma-trend	limma 3.30.13	$\log_2(\text{CPM})$
■	MASTcpm	MAST	MAST 1.0.5	$\log_2(\text{CPM}+1)$
■	MASTcpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	$\log_2(\text{CPM}+1)$
■	MASTtpm	MAST	MAST 1.0.5	$\log_2(\text{TPM}+1)$
■	MASTtpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	$\log_2(\text{TPM}+1)$
■	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts
■	monocle	monocle (tobit)	monocle 2.2.0	TPM
■	monoclecensus	monocle (Negative Binomial)	monocle 2.2.0	Census counts
■	monoclecount	monocle (Negative Binomial)	monocle 2.2.0	raw counts
■	NODES	NODES	NODES 0.0.0.9010	raw counts
■	ROTScpm	ROTS	ROTS 1.2.0	CPM
■	ROTStpm	ROTS	ROTS 1.2.0	TPM
■	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts
■	SAMseq	SAMseq	samr 2.0	raw counts
■	scDD	scDD	scDD 1.0.0	raw counts
■	SCDE	SCDE	scde 2.2.0	raw counts
■	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts
■	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts
■	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts
■	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM
■	ttest	t-test	stats (R v 3.3)	TMM-normalized TPM
■	voomlimma	voom-limma	limma 3.30.13	raw counts
■	Wilcoxon	Wilcoxon test	stats (R v 3.3)	TMM-normalized TPM

Bias, robustness and scalability in single-cell differential expression and analysis:

- ❖ 36 statistical approaches for DE analysis to compare the expression levels in the two groups of cells
- ❖ based on 9 data sets, with 11 - 21 separate instances (sample size effect)
- ❖ extensive evaluation of metrics incl. number of genes found, characteristics of the false positive detections, robustness of methods, similarities between methods

Soneson & Robinsons,  
Nature Methods, 2018

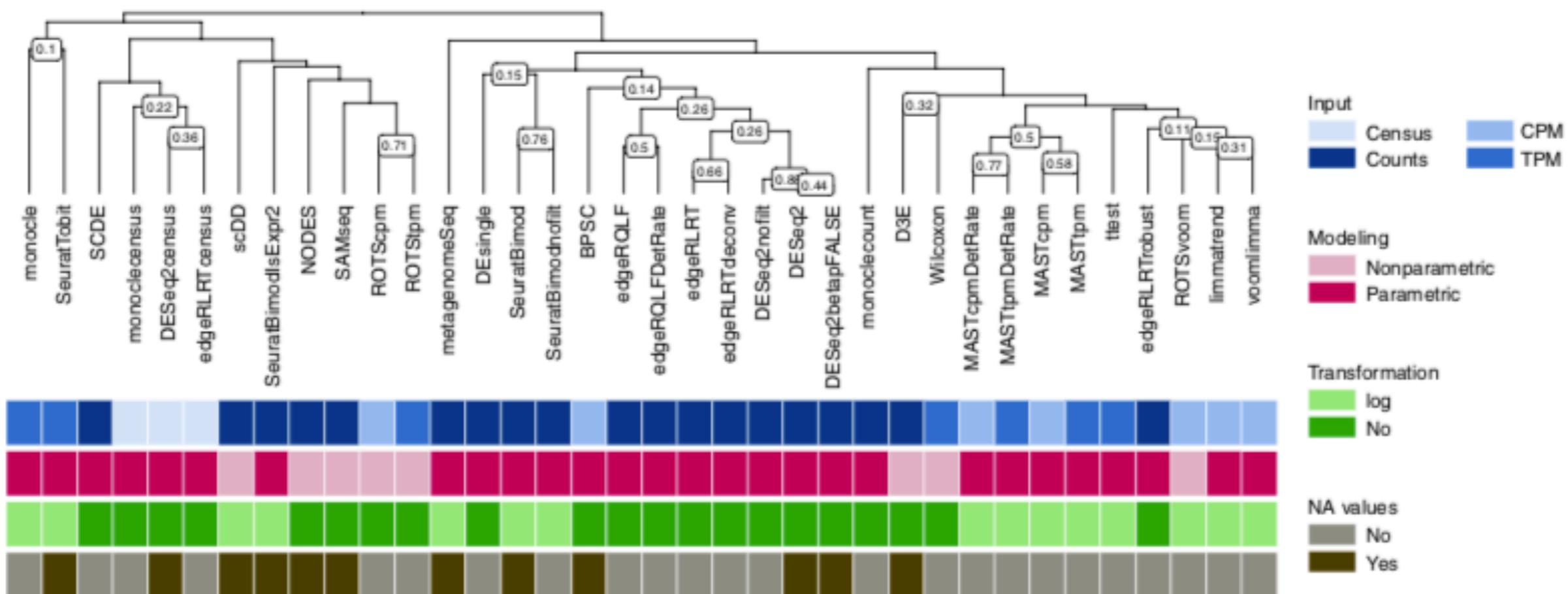
# Performance



## Some highlights:

- ❖ t-test and Wilcoxon work well, given at least few dozens cells to compare
- ❖ bulk RNA-seq analysis methods do not generally perform worse than those specifically developed for scRNA-seq
- ❖ Filtering out lowly expressed genes is quite important for good performance of bulk methods (edgeR, DEseq2)

# Performance



# Droplet scRNA-seq is not zero-inflated

To the Editor — Potential users of single-cell RNA-sequencing (scRNA-seq)<sup>1</sup> often encounter a choice between high-throughput droplet-based methods and high-sensitivity plate-based methods. There is a widespread belief that scRNA-seq will often fail to generate measurements for some genes from some cells owing to technical molecular inefficiencies. It is believed that this causes data to have an overabundance of zero values compared to what is expected from random sampling and that this effect is particularly pronounced in droplet-based methods. Here I present an investigation of published data for technical controls in droplet-based scRNA-seq experiments that demonstrates that the number of zero values in the data is consistent with common distributional models of molecule sampling counts. Thus, any additional zero values in biological data likely result from biological variation or may reflect variation in gene abundance among cell types or cell states.

dimensionality reduction methods with special handling of zero values had been introduced<sup>8,9</sup>. All these approaches share two common themes (which deviate from SCDE): first, expression data are considered continuous with additional zero values, and second, a proportional relation is identified between the number of zero values and the average expression level of a gene<sup>10</sup>.

In the field of computational methods for scRNA-seq analysis, many methods have been designed to correct zero values in data, with the aim of allowing users to predict what the expression level of a gene in a cell would have been, had there been no zero-inflation or dropouts<sup>11–17</sup>, referred to as ‘imputation’.

High-throughput droplet-based scRNA-seq methods measure discrete counts of unique molecular identifiers (UMIs). The stochastic sampling of counts can be modeled using a Gamma-Poisson distribution, more colloquially known as

platforms, it is possible to use negative control data with no biological variation. This will answer whether technical shortcomings in scRNA-seq methods produce an excess of technical zeros compared to expectations.

Negative-control datasets have been generated by adding a solution of RNA to the fluid in microfluidic systems, making the RNA content in each droplet identical. Five such datasets have been published: one to benchmark Drop-seq<sup>24</sup>, one to benchmark InDrops<sup>25</sup>, one to benchmark an early version of the commercial scRNA-seq platform from 10x Genomics<sup>26</sup> and two to benchmark a later version of the commercial platform from 10x Genomics<sup>27</sup> (Supplementary Note 1). The negative control data differ in the RNA solution added: bulk RNA from cells or tissues or artificial External RNA Control Consortium (ERCC) RNA (Table 1). For simplicity, here I refer to each RNA species, regardless

# Finally

## test.use

Denotes which test to use. Available options are:



- "wilcox" : Identifies differentially expressed genes between two groups of cells using a Wilcoxon Rank Sum test (default)
- "bimod" : Likelihood-ratio test for single cell gene expression, (McDavid et al., Bioinformatics, 2013)
- "roc" : Identifies 'markers' of gene expression using ROC analysis. For each gene, evaluates (using AUC) a classifier built on that gene alone, to classify between two groups of cells. An AUC value of 1 means that expression values for this gene alone can perfectly classify the two groupings (i.e. Each of the cells in cells.1 exhibit a higher level than each of the cells in cells.2). An AUC value of 0 also means there is perfect classification, but in the other direction. A value of 0.5 implies that the gene has no predictive power to classify the two groups. Returns a 'predictive power' ( $\text{abs}(\text{AUC}-0.5) * 2$ ) ranked matrix of putative differentially expressed genes.



- "t" : Identify differentially expressed genes between two groups of cells using the Student's t-test.



- "negbinom" : Identifies differentially expressed genes between two groups of cells using a negative binomial generalized linear model. Use only for UMI-based datasets



- "poisson" : Identifies differentially expressed genes between two groups of cells using a poisson generalized linear model. 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" : Identifies differentially expressed genes between two groups of cells using a hurdle model tailored to scRNA-seq data. Utilizes the MAST package to run the DE testing.



- "DESeq2" : Identifies differentially expressed genes between two groups of cells based on a model using DESeq2 which uses a negative binomial distribution (Love et al, Genome Biology, 2014). This test does not support pre-filtering of genes based on average difference (or percent detection rate) between cell groups. However, genes may be pre-filtered based on their minimum detection rate (min.pct) across both cell groups. To use this method, please install DESeq2, using the instructions at <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>



**Thank you for your attention!**