scRNA-seq

Differential expression analysis methods

Olga Dethlefsen

NBIS, National Bioinformatics Infrastructure Sweden

October 2017

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Outline

- Introduction: what is so special about DE with scRNA-seq
- Common methods: what is out there
- Performance: how to choose the best method
- Summary
- DE tutorial

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

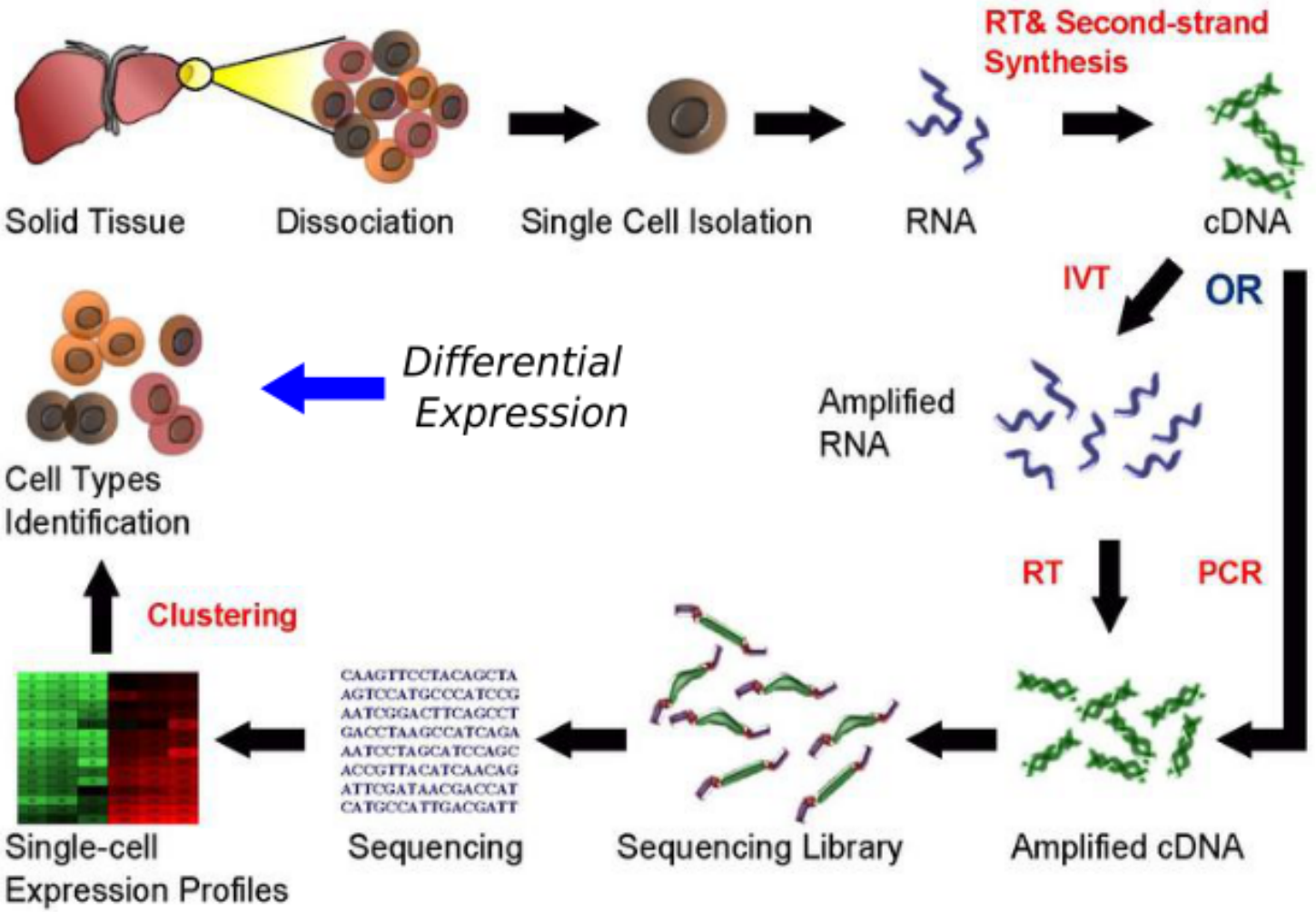


Figure: Simplified scRNA-seq workflow [adopted from http://hemberg-lab.github.io/

Differential expression is an old problem...so

why is DE scRNA-seq different to RNA-seq?

- ?
- ?
- ?
- ?
- ?

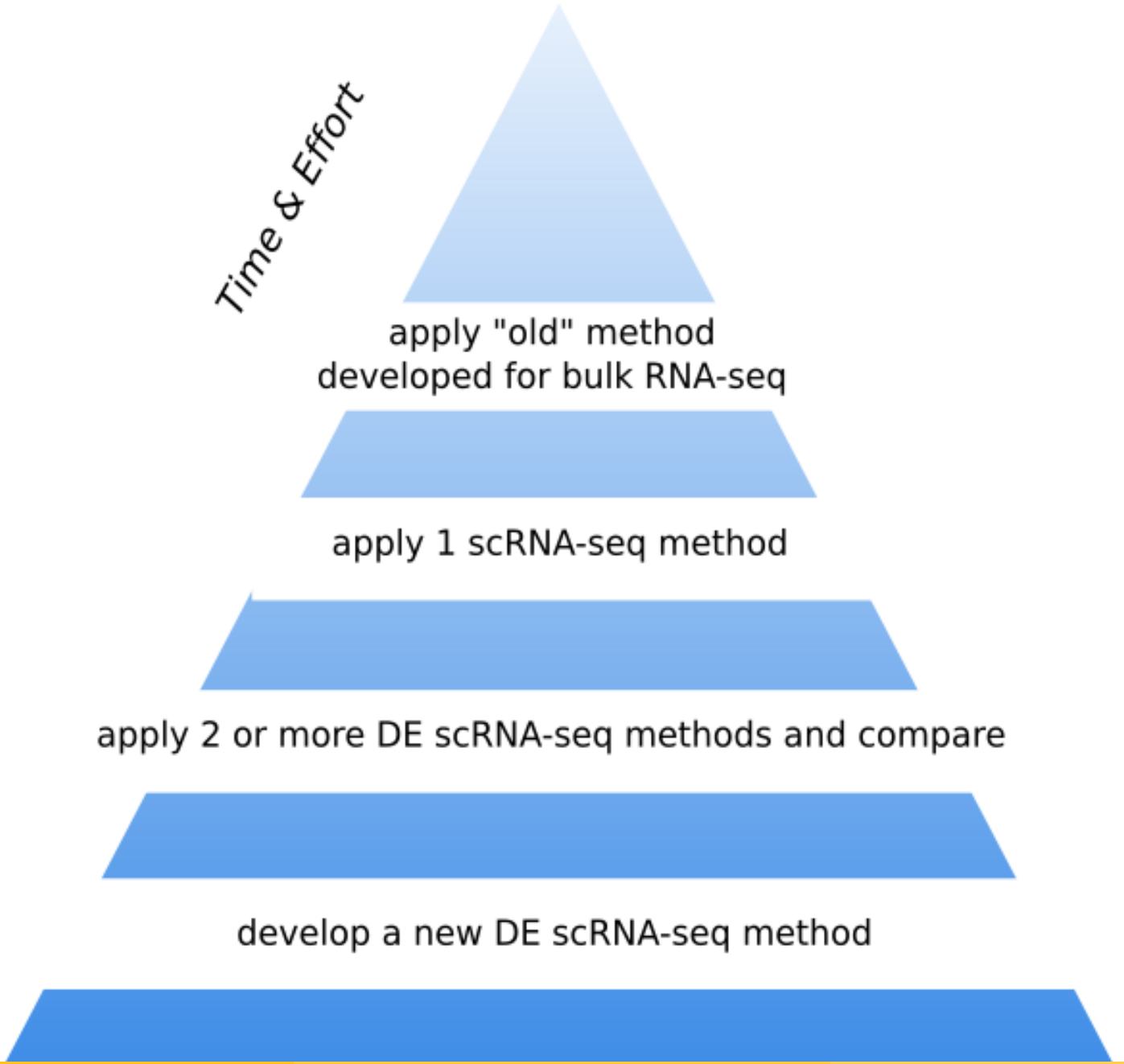
Differential expression is an old problem...so

why is DE scRNA-seq different to RNA-seq?

- scRNA-seq are affected by higher noise (technical and biological factors)
- low amount of available mRNAs results in amplification biases and "dropout events" (technical)
- 3' bias, partial coverage and uneven depth (technical)
- stochastic nature of transcription (biological)
- multimodality in gene expression; presence of multiple possible cell states within a cell population (biological)

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Common methods


Time & Effort
apply "old" method
developed for bulk RNA-seq
apply 1 scRNA-seq method
apply 2 or more DE scRNA-seq methods and compare
develop a new DE scRNA-seq method

Common methods

- non-parametric test e.g. Kruskal-Wallis (generic)
- edgeR, limma (bulk RNA-seq)
- MAST, SCDE, Monocle (scRNA-seq)
- D^3E, Pagoda (scRNA-seq)

Method	Model	Input	Platform	Threshold	Run time	Ref.
SCDE	Poisson and negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[13]
monocle	Generalized additive models	Read counts matrix	R(package)	*p*-value	Minutes	[14]
D3E	Non-parametric (test of distribution)	Read counts matrix	Python(package)	*p*-value	1 hour	[15]
BPSC	Beta-Poisson model	Read counts matrix	R(package)	*p*-value	1 hour	[16]
DESeq	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[10]
edgeR	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[11]
baySeq	Negative binomial model	Read counts matrix	R(package)	Likelihood	12 hours	[24]
NBPSeq	Negative binomial model	Read counts matrix	R(package)	*p*-value	Minutes	[25]
Cuffdiff	Beta negative binomial model	Sam file	Linux	*p*-value	13 hours	[26]
DEGseq	Poisson model	Read counts matrix	R(package)	*p*-value	Minutes	[12]
TSPM	Poisson model	Read counts matrix	R(script)	*p*-value	1 hour	[27]
limma	Linear models	Read counts matrix	R(package)	*p*-value	Seconds	[28]
ballgown	Nested linear models	Read counts matrix /ctab file	R(package)	*p*-value	Seconds	[29]
SAMseq	Non-parametric (resampling)	Read count matrix	R(package)	*p*-value	Minutes	[30]

Run time is measured by one experiment of 40 samples vs 40 samples, and the used parameters and settings are shown in the materials and methods part.

Table: Information of gene differential expression analysis methods used [Miao and Zhang, 2017, Quantitative Biology 2016, 4]

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

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):

$$logit(P_r(Z_{ig} = 1)) = X_i \beta_g^D$$

$P_r(Y_{ig} = Y|Z_{ig} = 1) = N(X_i \beta_g^C, \sigma_g^2)$, where X_i 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

SCDE

- models the read counts for each gene using a mixture of a NB, negative binomial, and a Poisson distribution
- NB distribution models the transcripts that are amplified and detected
- Poisson distribution models the unobserved or background-level signal of transcripts that are not amplified (e.g. dropout events)
- subset of robust genes is used to fit, via EM algorithm, the parameters to the mixture of models
- For DE, the posterior probability that the gene shows a fold expression difference between two conditions is computed using a Bayesian approach

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Monocole

- Originally designed for ordering cells by progress through differentiation stages (pseudo-time)
- The mean expression level of each gene is modeled with a GAM, generalized additive model, which relates one or more predictor variables to a response variable as

$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_m(x_m)$ where Y is a specific gene expression level, x_i are predictor variables, g is a link function, typically log function, and f_i are non-parametric functions (e.g. cubic splines)

- The observable expression level Y is then modeled using GAM,

$E(Y) = s(\varphi_t(b_x, s_i)) + \epsilon$ where $\varphi_t(b_x, s_i)$ is the assigned pseudo-time of a cell and s is a cubic smoothing function with three degrees of freedom. The error term ϵ is normally distributed with a mean of zero

- The DE test is performed using an approx. χ^2 likelihood ratio test

Let's stop for a minute...

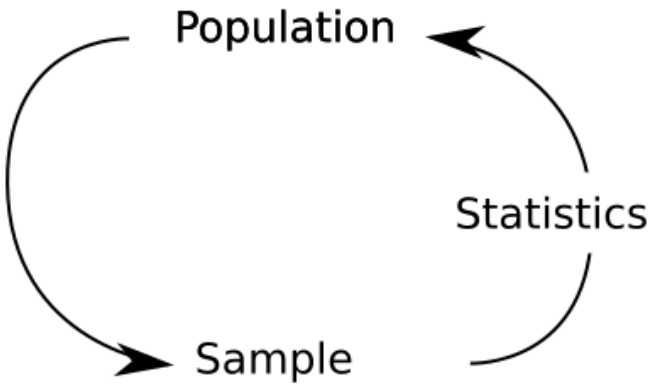
Differential expression

Differential expression analysis

- means taking the normalized read count data &
- performing statistical analysis to discover quantitative changes in expression levels between experimental groups.
- e.g. to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.
- or simply: checking for differences in distributions

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

The key

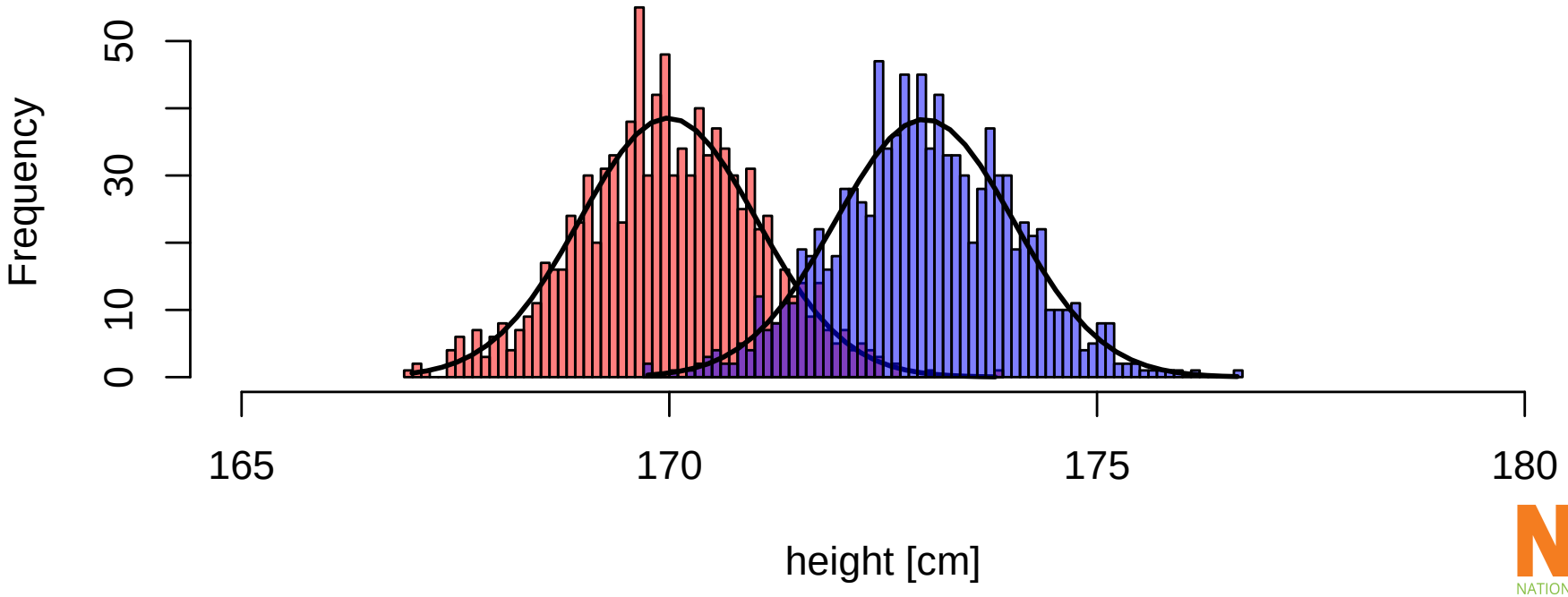


$$Outcome_i = (Model_i) + error_i$$

- we collect data on a *sample* from a much larger *population*. *Statistics* lets us to make inferences about the population from which it was derived
- we try to predict the outcome given a model fitted to the data

The key

$$t = \frac{x_1 - x_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



The key

Simple recipe

- model e.g. gene expression with random error
- fit model to the data and/or data to the model, estimate model parameters
- use model for prediction and/or inference

The key: MAST (again)

- 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):

$$logit(P_r(Z_{ig} = 1)) = X_i \beta_g^D$$

$P_r(Y_{ig} = Y | Z_{ig} = 1) = N(X_i \beta_g^C, \sigma_g^2)$, where X_i 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

The key: SCDE (again)

- models the read counts for each gene using a mixture of a NB, negative binomial, and a Poisson distribution
- NB distribution models the transcripts that are amplified and detected
- Poisson distribution models the unobserved or background-level signal of transcripts that are not amplified (e.g. dropout events)
- subset of robust genes is used to fit, via EM algorithm, the parameters to the mixture of models
- For DE, the posterior probability that the gene shows a fold expression difference between two conditions is computed using a Bayesian approach

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

The key: Monocole (again)

- Originally designed for ordering cells by progress through differentiation stages (pseudo-time)
- The mean expression level of each gene is modeled with a GAM, generalized additive model, which relates one or more predictor variables to a response variable as

$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_m(x_m)$ where Y is a specific gene expression level, x_i are predictor variables, g is a link function, typically log function, and f_i are non-parametric functions (e.g. cubic splines)

- The observable expression level Y is then modeled using GAM,

$E(Y) = s(\varphi_t(b_x, s_i)) + \epsilon$ where $\varphi_t(b_x, s_i)$ is the assigned pseudo-time of a cell and s is a cubic smoothing function with three degrees of freedom. The error term ϵ is normally distributed with a mean of zero

- The DE test is performed using an approx. χ^2 likelihood ratio test

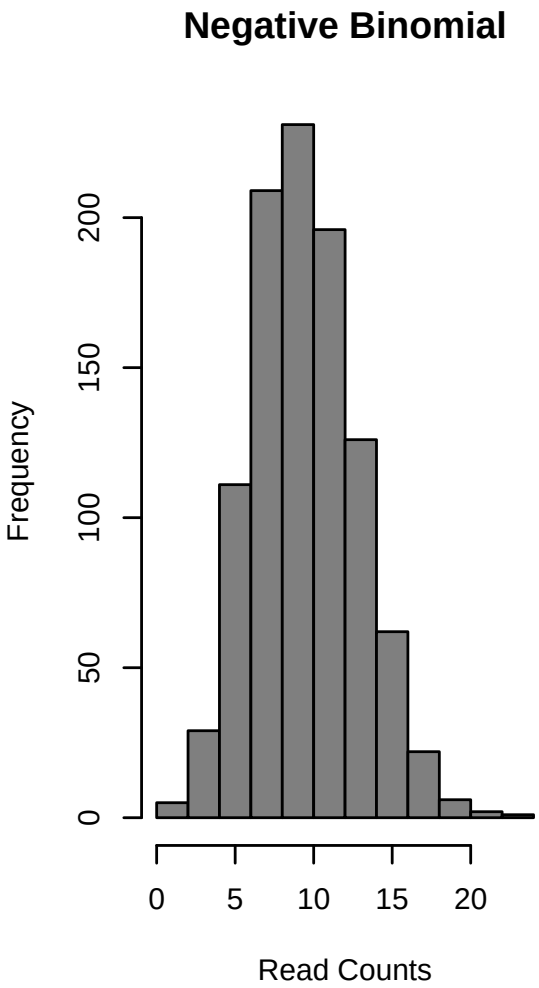
They key: implication

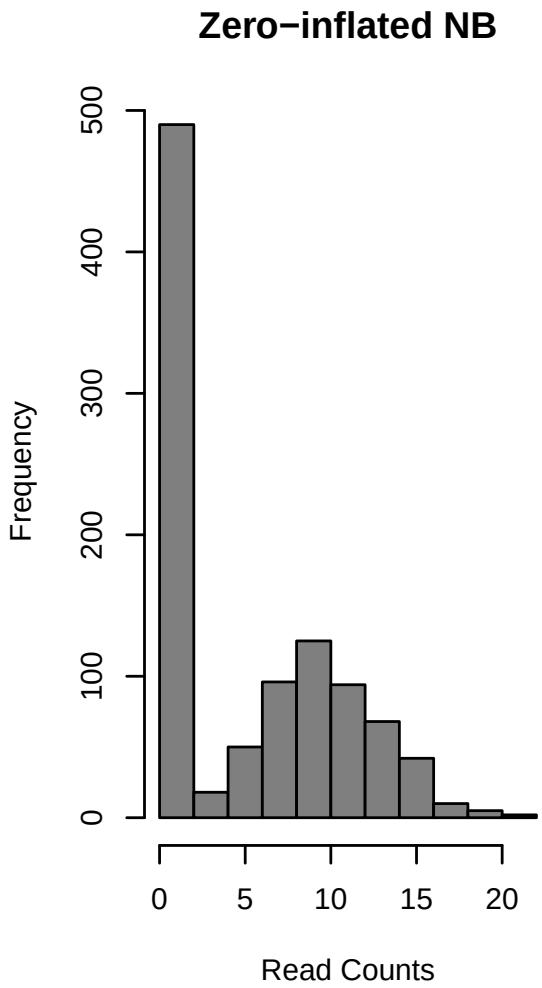
Simple recipe

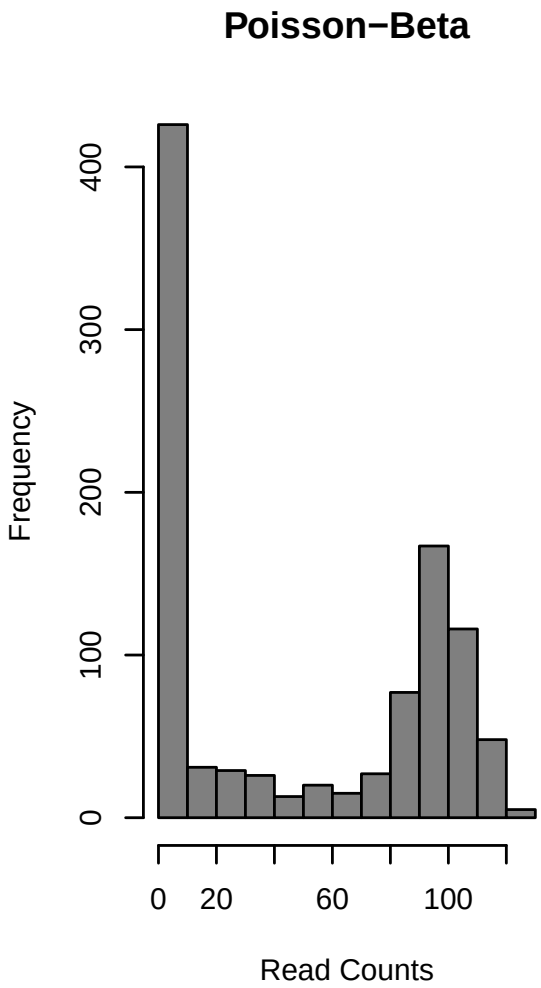
- model e.g. gene expression with random error
- fit model to the data and/or data to the model, estimate model parameters
- use model for prediction and/or inference

Implication

- the better model fits to the data the better statistics


Negative Binomial
Frequency
0
50
100
150
200
0
5
10
15
20
Read Counts


Zero−inflated NB
Frequency
0
100
200
300
400
500
0
5
10
15
20
Read Counts


Poisson−Beta
Frequency
0
100
200
300
400
0
20
60
100
Read Counts

Performance

No golden standard

There is no golden standard, no single best solution

...so what do we do?

No golden standard

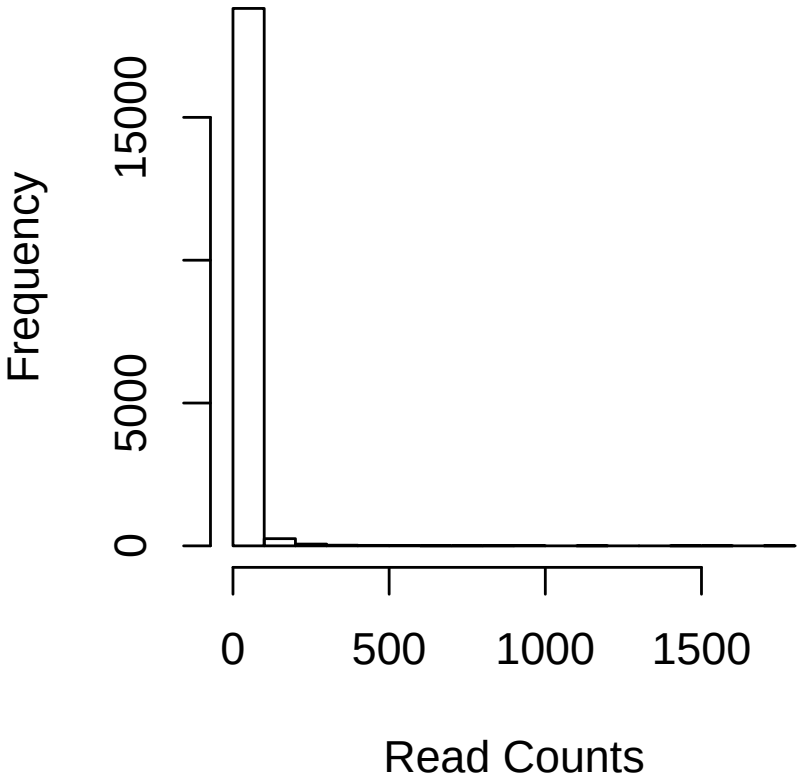
There is no golden standard, no single best solution

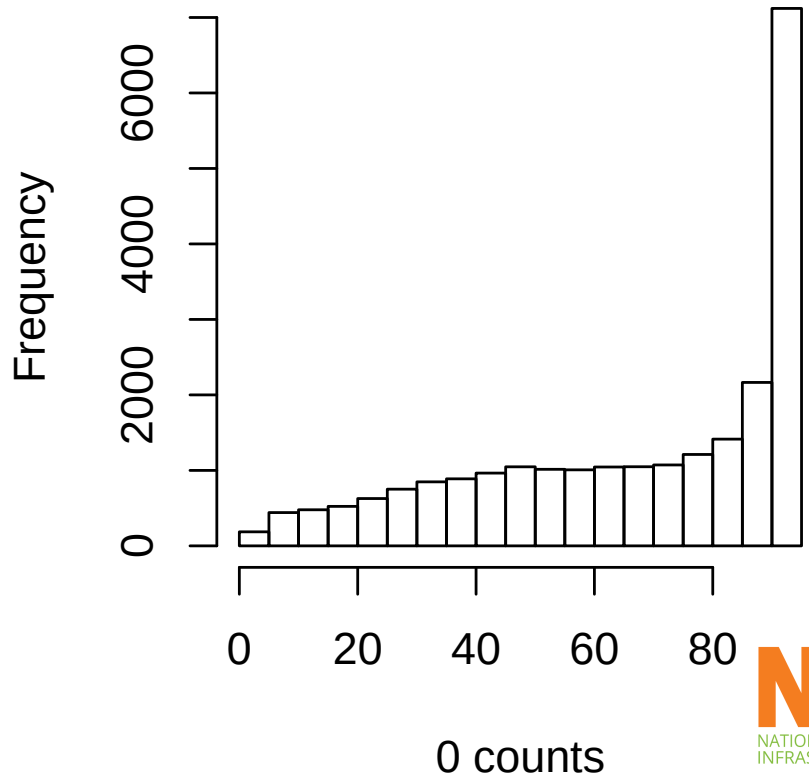
...so what do we do?

we gather as much evidence as possible

Get to know your data & wisely choose DE methods

Example data: 46,078 genes x 96 cells
22,229 genes with no expression at all





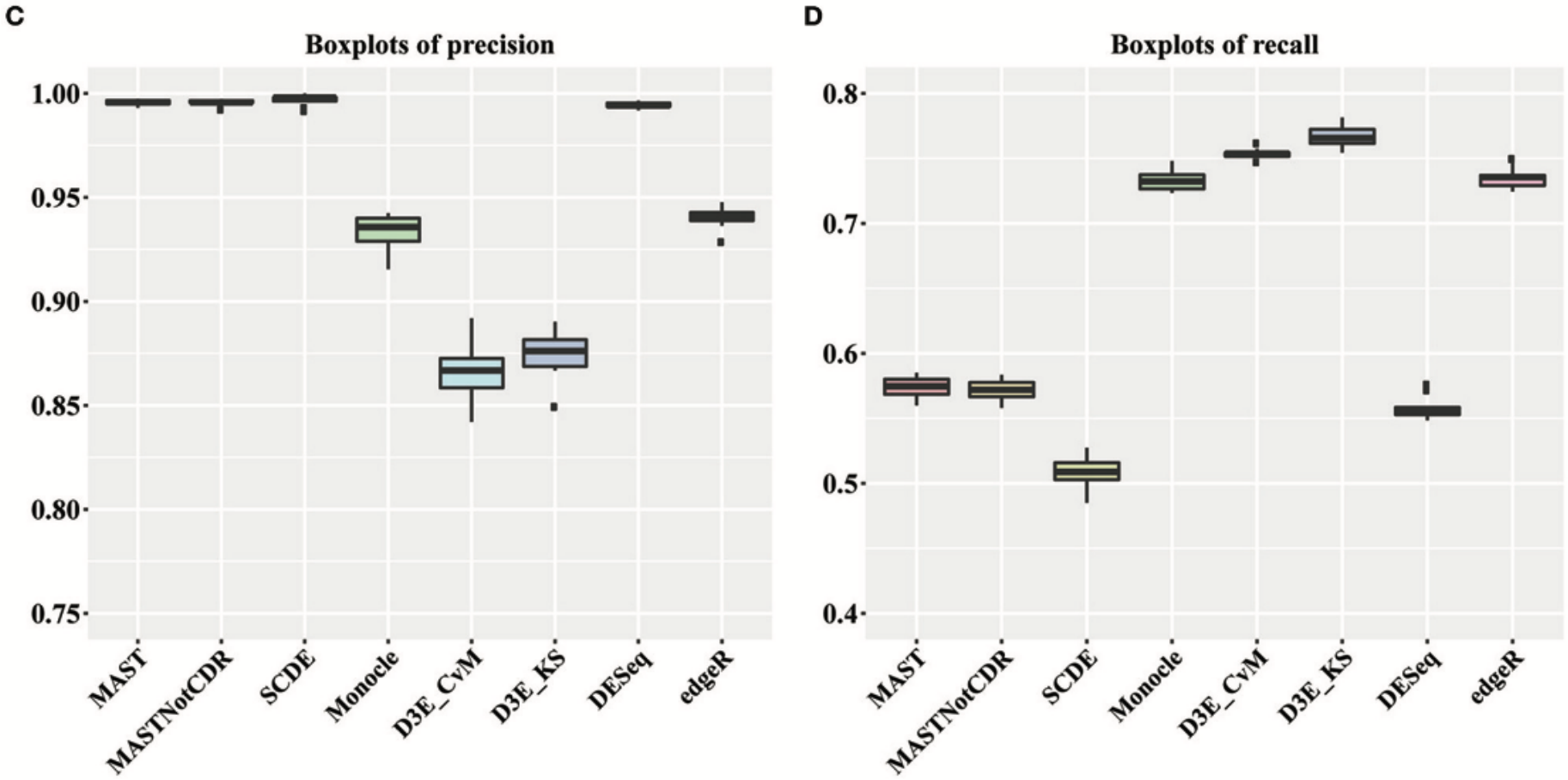
Learn from methodological papers and/or past studies

e.g. Dal Molin, Barruzo and Di Camillillo, frontiers in Genetics 2017, Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods

- 10,000 genes simulated for 2 conditions with sample size of 100 cells each
- 8,000 genes were simulated as not differentially expressed using the same distribution (unimodal: NB and bimodal: two-component NB mixture)
- 2,000 genes were simulated as differentially expressed according to four types of differential expressions
- real dataset: 44 mouse Embryonic Stem Cells and 44 Embryonic Fibroblsts for positive control
- real dataset: 80 single cells as negative control

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

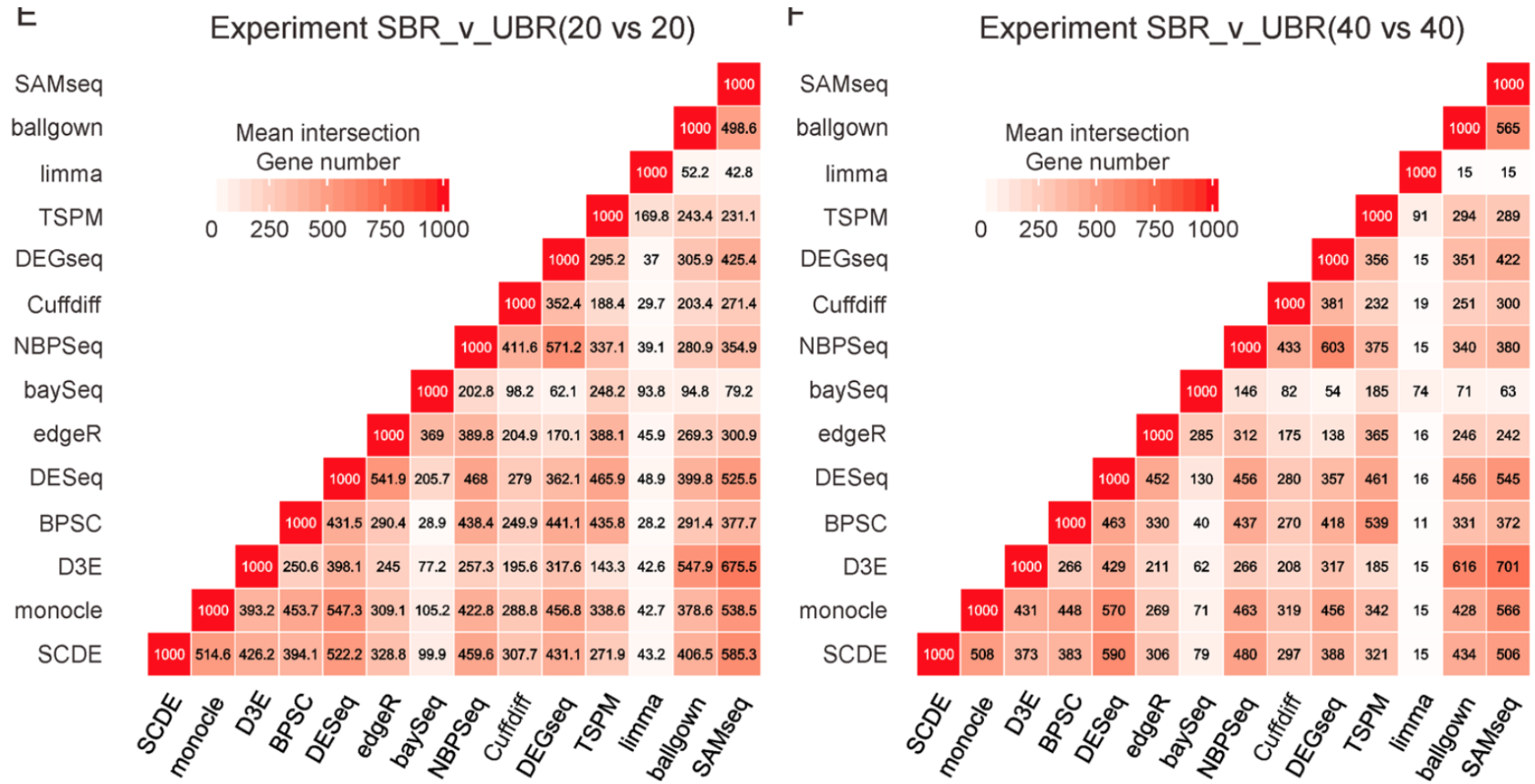
Learn from methodological papers and/or past studies



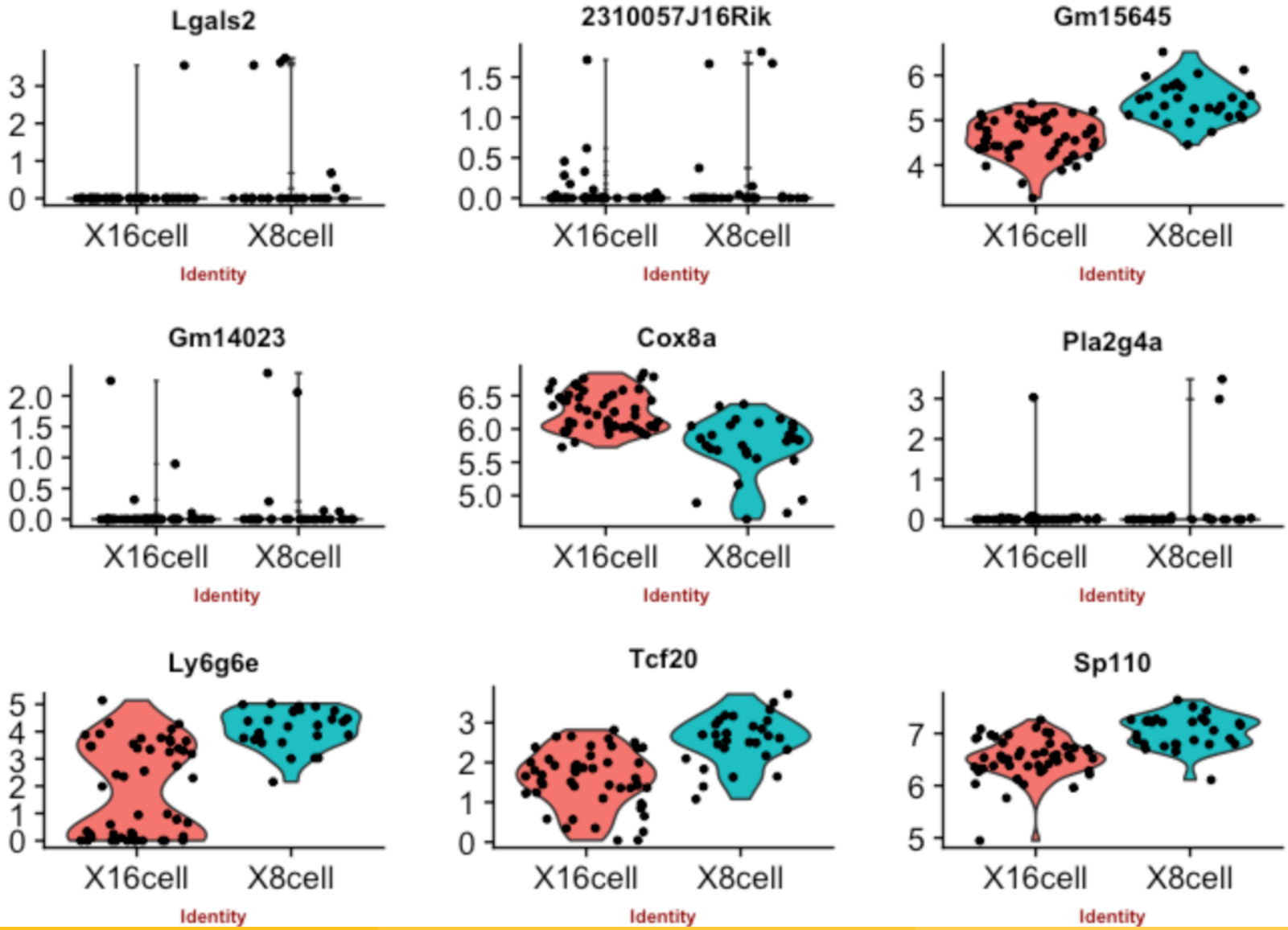
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Compare methods

e.g. Miao and Zhang, Quantitative Biology 2016,4: Differential expression analyses for single-cell RNA-Seq: old questions on new data



Stay critical



Summary

Summary

- scRNA-seq is a rapidly growing field
- DE is a common task so many newer and better methods will be developed
- think like a statistician: get to know your data, think about distributions and models best for your data. Avoid applying methods blindly
- comparing methods is good as long as you are aware what you are comparing and why
- stay critical

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

DE tutorial

DE tutorial

Based on the dataset used is single-cell RNA-seq data (SmartSeq) from mouse embryonic development from Deng. et al. Science 2014, Vol. 343 no. 6167 pp. 193-196, "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells".

- check for differentially expressed genes between 8-cell and 16-cell stage embryos
- with many methods incl. SCDE, MAST, SC3 package, Pagoda, Seurat
- and compare the results, trying to decide on the best DE method for the dataset

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Thank you for attention

Questions?

Enjoy the rest of the course

olga.dethlefsen@nbis.se

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN