scRNA-seq

Differential expression analyses

Olga Dethlefsen

NBIS, National Bioinformatics Infrastructure Sweden

May 2018

Outline

Outline

- Introduction: what is so special about scRNA-seq DE?

Outline

- Introduction: what is so special about scRNA-seq DE?
- Common methods: what is out there?

Outline

- Introduction: what is so special about scRNA-seq DE?
- Common methods: what is out there?
- Performance: how do we know what is best?

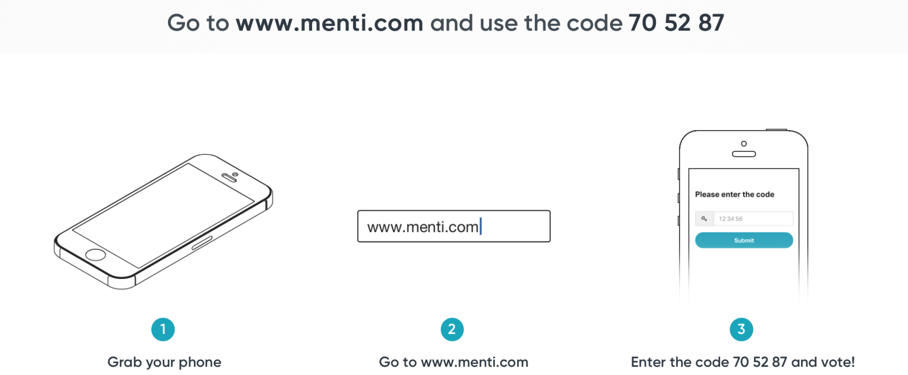
Outline

- Introduction: what is so special about scRNA-seq DE?
- Common methods: what is out there?
- Performance: how do we know what is best?
- Practicalities: what to do in real life?

Outline

- Introduction: what is so special about scRNA-seq DE?
- Common methods: what is out there?
- Performance: how do we know what is best?
- Practicalities: what to do in real life?
- Summary: what to remember from this hour?

Let’s get to know each other



https://www.menti.com

Introduction

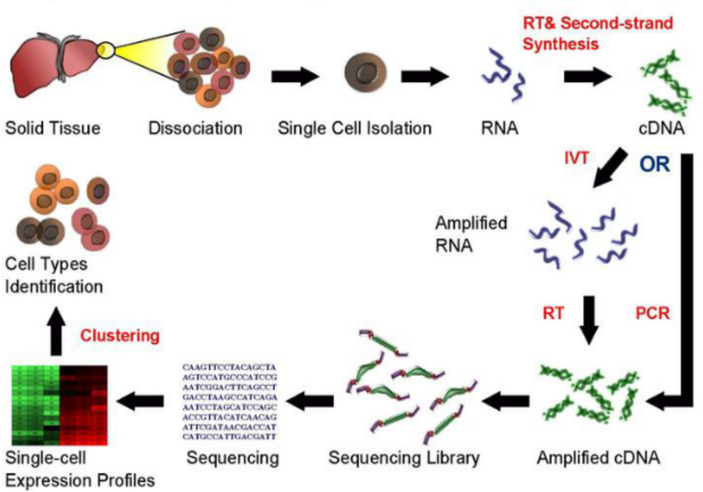


Figure: Simplified scRNA-seq workflow [adapted from Wikipedia]

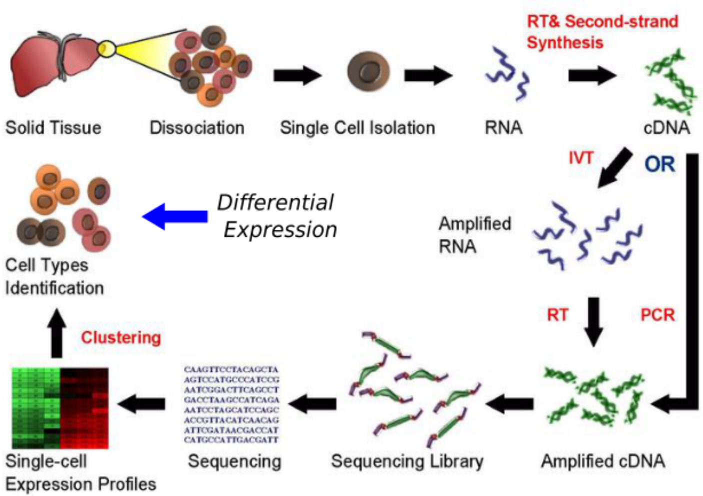
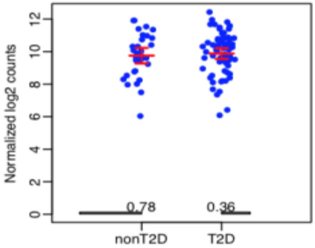
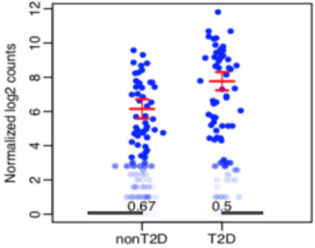


Figure: Simplified scRNA-seq workflow [adapted from Wikipedia]

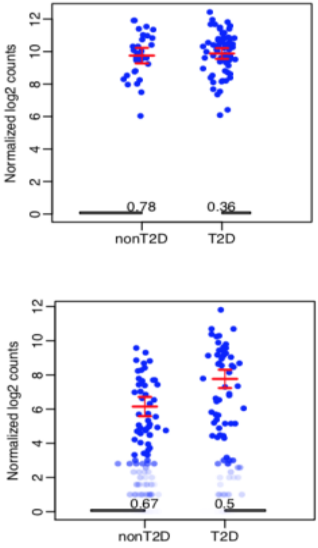




adapted from Wu et al. 2017

Differential expression means

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



adapted from Wu et al. 2017

Differential expression means

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

Differential expression is an old "problem"

- known from bulk RNA-seq and microarray studies
- in fact building on one of the most common stastistical problems, i.e comparing groups for statistical differences

Differential expression is an old problem.

So what is all the commotion about?

https://www.menti.com & 70 52 87

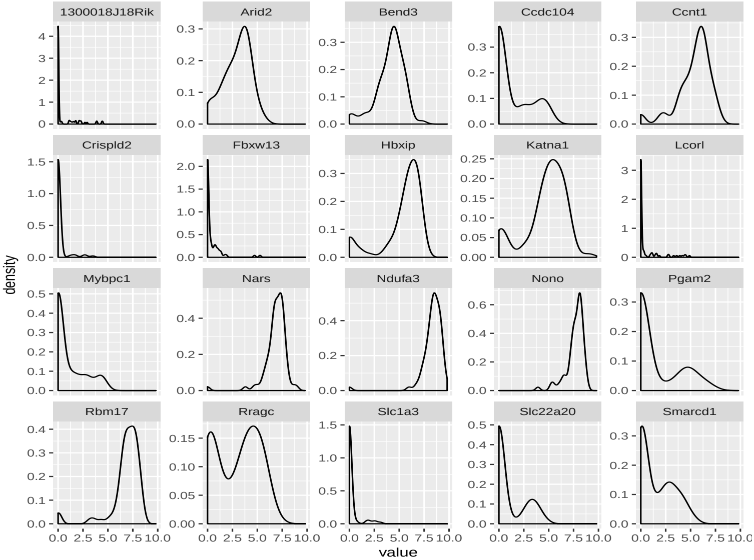
Differential expression is an old problem.

So what is all the commotion about?

https://www.menti.com & 70 52 87

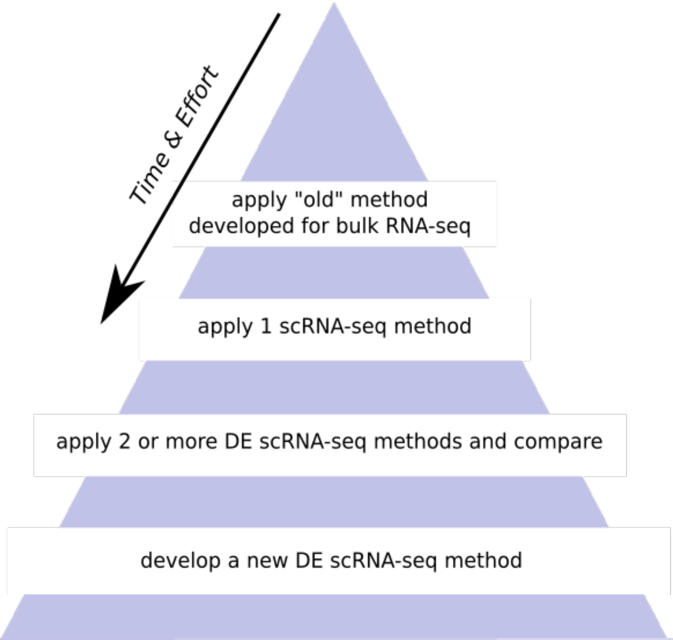
scRNA-seq: special characteristics

- 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 and uneven depth (technical)
- stochastic nature of transcription (biological)
- multimodality in gene expression; presence of multiple possible cell states within a cell population (biological)



Based on tutorial data

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

Generic

- parametric tests, e.g. t-test
- non-parametric tests, e.g. Kruskal-Wallis

RNA-seq based

- edgeR
- limma
- DEseq2

scRNA-seq specific

- MAST, SCDE, Monocle
- D^3E, Pagoda

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.

Miao and Zhang 2016

	Short name	Method	Software version	Input	Available from	Reference
	BPSC	BPSC	BPSC 0.99.0/1	CPM	GitHub	[11]
	D3E	D3E	D3E 1.0	raw counts	GitHub	[12]
	DESeq2	DESeq2	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2betapFALSE	DESeq2 without beta prior	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DESeq2census	DESeq2	DESeq2 1.14.1	Census counts	Bioconductor	[13]
	DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	Bioconductor	[13]
	DEsingle	DEsingle	DEsingle 0.1.0	raw counts	GitHub	[14]
	edgeRLRT	edgeR/LRT	edgeR 3.19.1	raw counts	Bioconductor	[15–17]
	edgeRLRTcensus	edgeR/LRT	edgeR 3.19.1	Census counts	Bioconductor	[15–17]
	edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.19.1, scran 1.2.0	raw counts	Bioconductor	[15, 17, 18]
	edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.19.1	raw counts	Bioconductor	[15–17, 19]
	edgeRQLF	edgeR/QLF	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	edgeRQLFDetRate	edgeR/QLF with cellular detection rate as covariate	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
	limmatrend	limma-trend	limma 3.30.13	log_2(CPM)	Bioconductor	[21, 22]
	MASTcpm	MAST	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTcpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(CPM+1)	Bioconductor	[23]
	MASTtpm	MAST	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	MASTtpmDetRate	MAST with cellular detection rate as covariate	MAST 1.0.5	log_2(TPM+1)	Bioconductor	[23]
	metagenomeSeq	metagenomeSeq	metagenomeSeq 1.16.0	raw counts	Bioconductor	[24]
	monocle	monocle (tobit)	monocle 2.2.0	TPM	Bioconductor	[25]
	monoclecensus	monocle (Negative Binomial)	monocle 2.2.0	Census counts	Bioconductor	[25, 26]
	monoclecount	monocle (Negative Binomial)	monocle 2.2.0	raw counts	Bioconductor	[25]
	NODES	NODES	NODES 0.0.0.9010	raw counts	Author-provided link	[27]
	ROTScpm	ROTS	ROTS 1.2.0	CPM	Bioconductor	[28, 29]
	ROTStpm	ROTS	ROTS 1.2.0	TPM	Bioconductor	[28, 29]
	ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	Bioconductor	[28, 29]
	SAMseq	SAMseq	samr 2.0	raw counts	CRAN	[30]
	scDD	scDD	scDD 1.0.0	raw counts	Bioconductor	[31]
	SCDE	SCDE	scde 2.2.0	raw counts	Bioconductor	[32]
	SeuratBimod	Seurat (bimod test)	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodnofilt	Seurat (bimod test) without the internal filtering	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratBimodIsExpr2	Seurat (bimod test) with internal expression threshold set to 2	Seurat 1.4.0.7	raw counts	GitHub	[33, 34]
	SeuratTobit	Seurat (tobit test)	Seurat 1.4.0.7	TPM	GitHub	[25, 33]
	ttest	t-test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 35]
	voomlimma	voom-limma	limma 3.30.13	raw counts	Bioconductor	[21, 22]
	Wilcoxon	Wilcoxon test	stats (R v 3.3)	TMM-normalized TPM	CRAN	[16, 36]

Soneson and Robinson 2018

More detailed examples

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):

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

- 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

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

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)

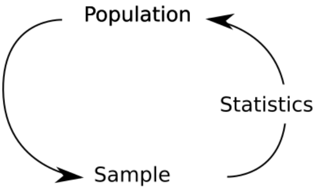
- The observable expression level Y is then modeled using GAM,

$E(Y) = s(\varphi_t(b_x, s_i)) + \epsilon$ 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...

The key

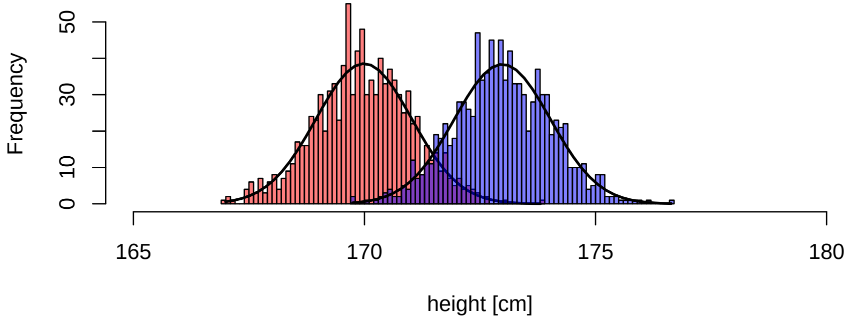


$$Outcome_i = (Model_i) + error_i$$

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

The key

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



Generic recipe

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

MAST (revisted)

- 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):

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

- 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

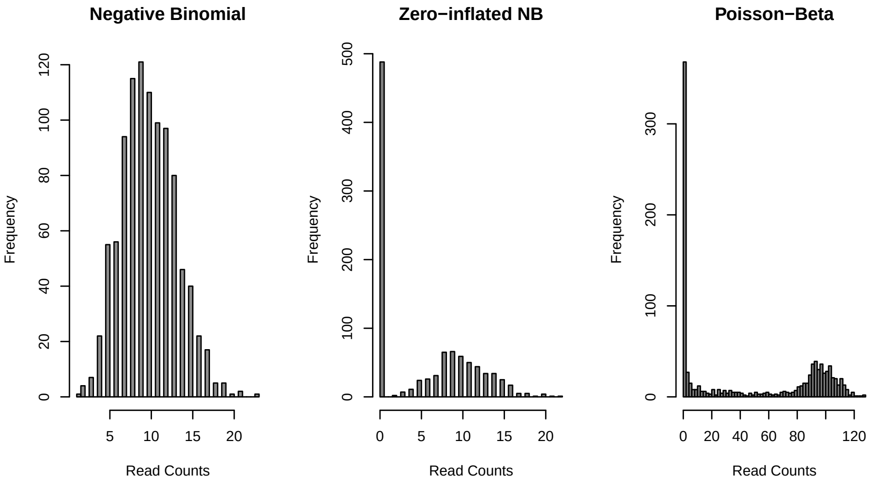
Generic recipe

- 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

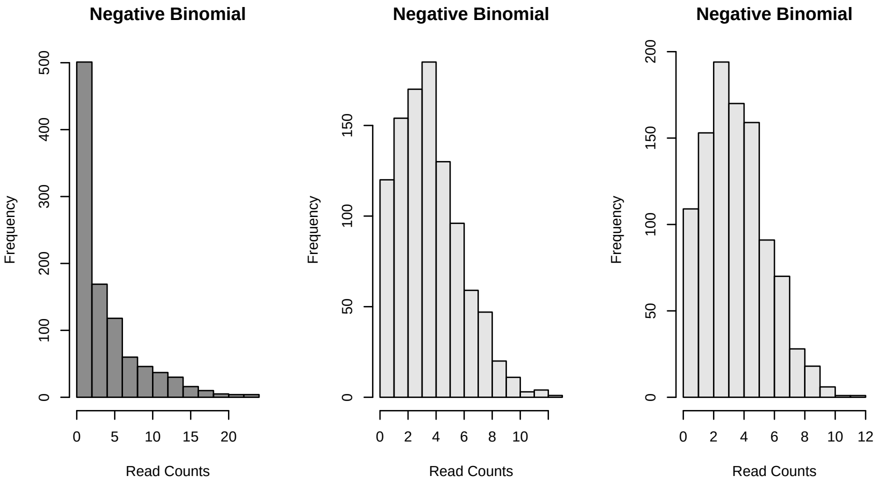
Important implication

the better model fits to the data the better statistics

Common distributions

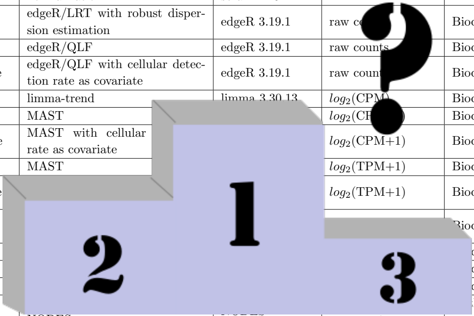


Common distributions



Performance

BPSC	BPSC	BPSC 0.99.0/1	CPM	GitHub	[11]
D3E	D3E	D3E 1.0	raw counts	GitHub	[12]
DESeq2	DESeq2	DESeq2 1.14.1	raw counts	Bioconductor	[13]
DESeq2betapFALSE	DESeq2 without beta prior	DESeq2 1.14.1	raw counts	Bioconductor	[13]
DESeq2census	DESeq2	DESeq2 1.14.1	Census counts	Bioconductor	[13]
DESeq2nofilt	DESeq2 without the built-in independent filtering	DESeq2 1.14.1	raw counts	Bioconductor	[13]
DEsingle	DEsingle	DEsingle 0.1.0	raw counts	GitHub	[14]
edgeRLRT	edgeR/LRT	edgeR 3.19.1	raw counts	Bioconductor	[15–17]
edgeRLRTcensus	edgeR/LRT	edgeR 3.19.1	Census counts	Bioconductor	[15–17]
edgeRLRTdeconv	edgeR/LRT with deconvolution normalization	edgeR 3.19.1, scran 1.2.0	raw counts	Bioconductor	[15, 17, 18]
edgeRLRTrobust	edgeR/LRT with robust dispersion estimation	edgeR 3.19.1	raw c	Bioconductor	[15–17, 19]
edgeRQLF	edgeR/QLF	edgeR 3.19.1	raw counts	Bioconductor	[15, 16, 20]
edgeRQLFDetRate	edgeR/QLF with cellular detection rate as covariate	edgeR 3.19.1	raw coun	Bioconductor	[15, 16, 20]
limmatrend	limma-trend	limma 3.30.13	log_2(CPM)	Bioconductor	[21, 22]
MASTcpm	MAST		log_2(C	Bioconductor	[23]
MASTcpmDetRate	MAST with cellular rate as covariate		log_2(CPM+1)	Bioconductor	[23]
MASTtpm	MAST		log_2(TPM+1)	Bioconductor	[23]
MASTtpmDetRate			log_2(TPM+1)	Bioconductor	[23]
metagenomeSeq				ioconductor	[24]
monocle				conductor	[25]
monoclecensus				conductor	[25, 26]
monoclecount				conductor	[25]
NODES	NODES	0.0.0.9010	raw counts	hor- provided link	[27]
ROTScpm	ROTS	ROTS 1.2.0	CPM	Bioconductor	[28, 29]
ROTStpm	ROTS	ROTS 1.2.0	TPM	Bioconductor	[28, 29]
ROTSvoom	ROTS	ROTS 1.2.0	voom-transformed raw counts	Bioconductor	[28, 29]
SAMseq	SAMseq	samr 2.0	raw counts	CRAN	[30]
scDD	scDD	scDD 1.0.0	raw counts	Bioconductor	[31]
SCDE	SCDE	scde 2.2.0	raw counts	Bioconductor	[32]



No ground truth, i.e. no independently validated truth is available for testing

Known data

using data we know something about to get "positive controls"

No ground truth, i.e. no independently validated truth is available for testing

Known data

using data we know something about to get "positive controls"

Simulated data

null-data sets by re-sampling, modeling data sets based on various distributions

No ground truth, i.e. no independently validated truth is available for testing

Known data

using data we know something about to get "positive controls"

Simulated data

null-data sets by re-sampling, modeling data sets based on various distributions

Comparing between methods and scenarios

Comparing numbers of DEs incl. as a function of group size

No ground truth, i.e. no independently validated truth is available for testing

Known data

using data we know something about to get "positive controls"

Simulated data

null-data sets by re-sampling, modeling data sets based on various distributions

Comparing between methods and scenarios

Comparing numbers of DEs incl. as a function of group size

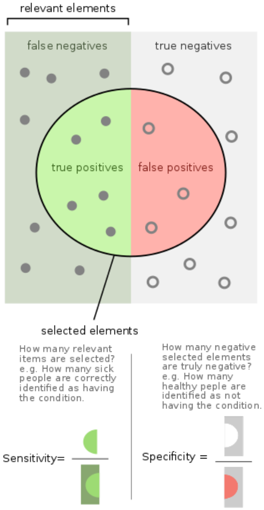
Investigating results

How does the expression and distributions of detected DEs look like?

False positives (type I error) vs. false negatives (type II error)

Sensitivity and specificity

Precision and recall



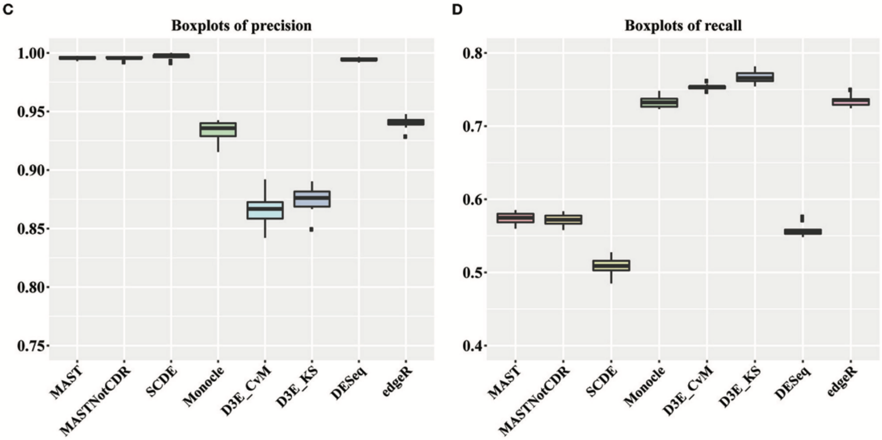
		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

adapted from Wikipedia

False positives (type I error) vs. false negatives (type II error)

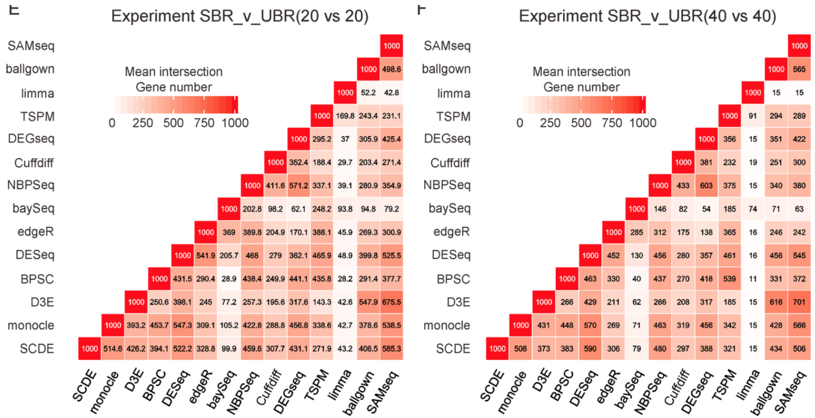
Sensitivity and specificity

Precision and recall



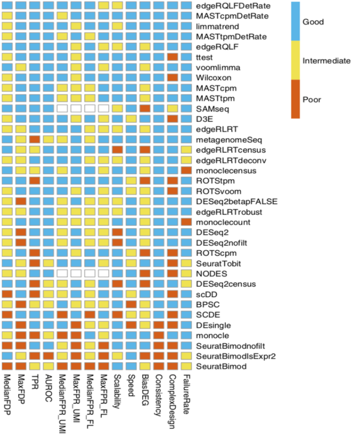
Dal Molin, Baruzzo, and Di Camillo 2017: 2 conditions of 100 cells each simulated with 10 000 genes, out of which 2 000 set to DEs (based on NB and bimodal distributions)

Consistency



Miao et al. 2017

And so much more...



Soneson and Robinson 2018

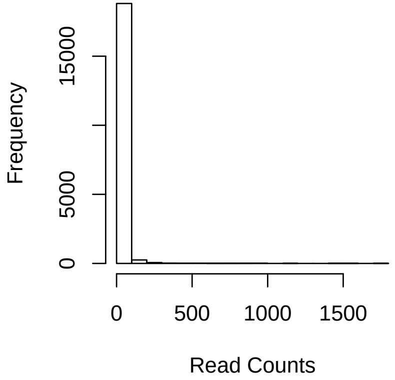
Bias, robustness and scalability in single-cell differential expression analysis

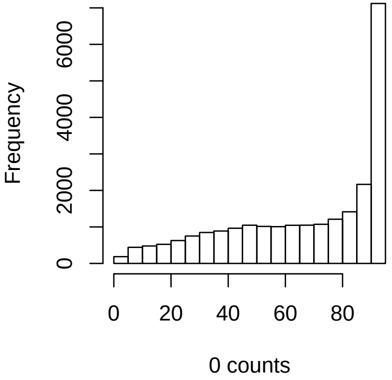
- 36 statistical approaches for DE analysis to compare the expression levels in the two groups of cells
- based on 9 datasets, with 11 - 21 separate instances (sample size effect)
- extensive evaluation metrics incl. number of genes found, characteristics of the false postivie detections, robutsness of methods, similarities between methods etc.
- *conquer*, a collection of consistently processed, analysis-ready public scRNA-seq data sets

Practicalities

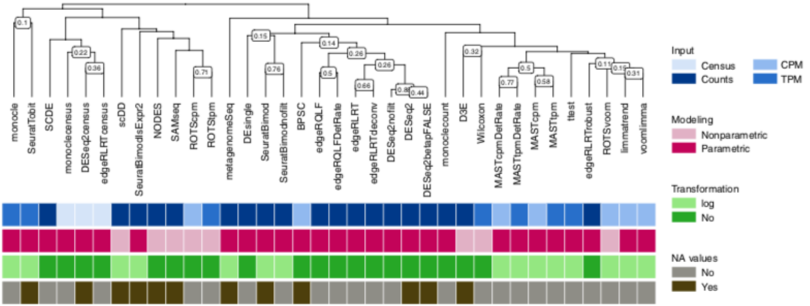
Getting to know your data

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



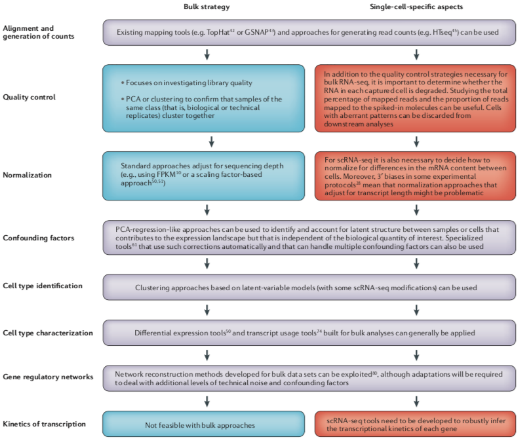


Choosing DE methods



Soneson and Robinson 2018

Rembering the bigger picture



Stegle, Teichmann, and Marioni 2015

QC filtering

Cell-cycle phase

Normalization of cell-specific biases

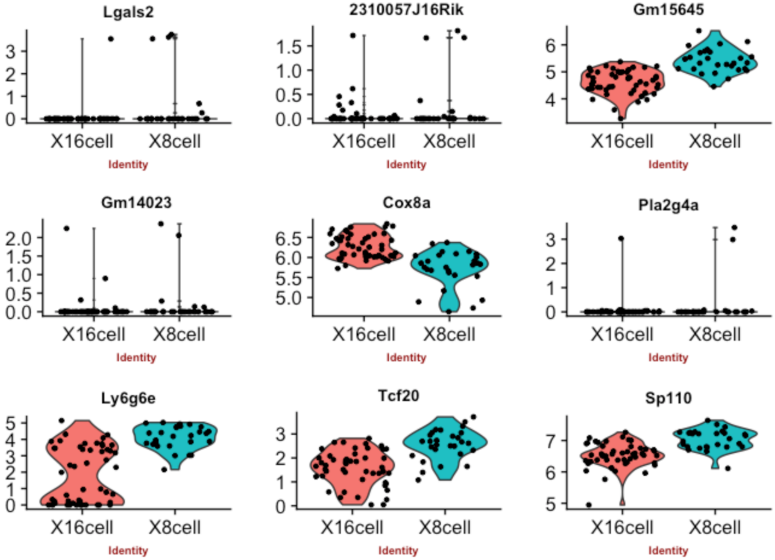
Confouding factors, incl. batch effects

Detection rate, i.e the fraction of detected genes per cell

Impututations strategies for dropout values

What is pragmatic: programming language, platform, speed, collaborative workflows etc.

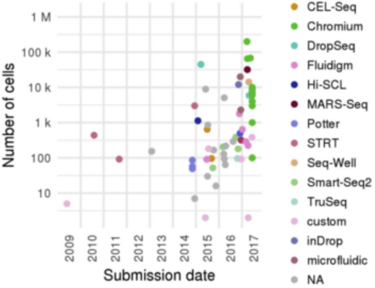
Staying critical



What to remember from this hour?

https://www.menti.com & 70 52 87

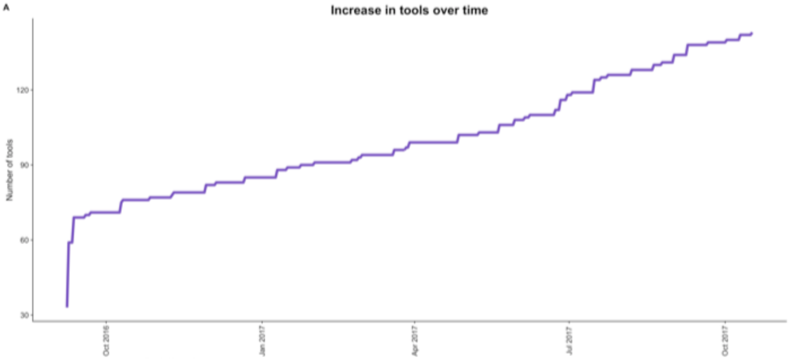
Growing field



Angerer et al. 2017

Growing field

https://www.scrna-tools.org/tools



Zappia, Phipson, and Oshlack 2018

Summary

- scRNA-seq is a rapidly growing field
- DE is a common task so many newer and better methods will be developed
- understanding basic statistical concepts enables one to think more like a statistician: to choose and evaluate methods given data set
- staying critical, staying updated, staying connected

Wu, Zhijin, et al. 2017. "Two-phase differential expression analysis for single cell RNA-seq". *Bioinformatics* 00 (00): 1–9. ISSN: 1367-4803. doi:10.1093/bioinformatics/bty329.

Miao, Zhun, and Xuegong Zhang. 2016. "Differential expression analyses for single-cell RNA-Seq: old questions on new data". *Quantitative Biology* 4 (4): 243–260. ISSN: 20954697. doi:10.1007/s40484-016-0089-7.

Soneson, Charlotte, and Mark D. Robinson. 2018. "Bias, robustness and scalability in single-cell differential expression analysis". *Nature Methods* 15 (4): 255–261. ISSN: 15487105. doi:10.1038/nmeth.4612. http://dx.doi.org/10.1038/nmeth.4612.

Dal Molin, Alessandra, Giacomo Baruzzo, and Barbara Di Camillo. 2017. "Single-cell RNA-sequencing: Assessment of differential expression analysis methods". *Frontiers in Genetics* 8 (MAY). ISSN: 16648021. doi:10.3389/fgene.2017.00062.

Miao, Zhun, et al. 2017. "DEsingle for detecting three types of differential expression in single-cell RNA-seq data", no. May: 1–2. ISSN: 1367-4803. doi:10.1093/bioinformatics/bty332. arXiv: 103549.

Stegle, Oliver, Sarah A Teichmann, and John C Marioni. 2015. "Computational and analytical challenges in single-cell transcriptomics." *Nature reviews. Genetics* 16 (January 2014): 133–145.

Angerer, Philipp, et al. 2017. "Single cells make big data: New challenges and opportunities in transcriptomics". *Current Opinion in Systems Biology* 4:85–91. ISSN: 24523100. doi:10.1016/j.coisb.2017.07.004. http://dx.doi.org/10.1016/j.coisb.2017.07.004.