Comparison of differential expression analysis tools for RNA-seq

Laura Elo, PhD, Adjunct Professor, Group Leader

Computational Biomedicine Group
Turku Centre for Biotechnology and
Department of Mathematics and Statistics

Turun yliopisto
University of Turku

Background

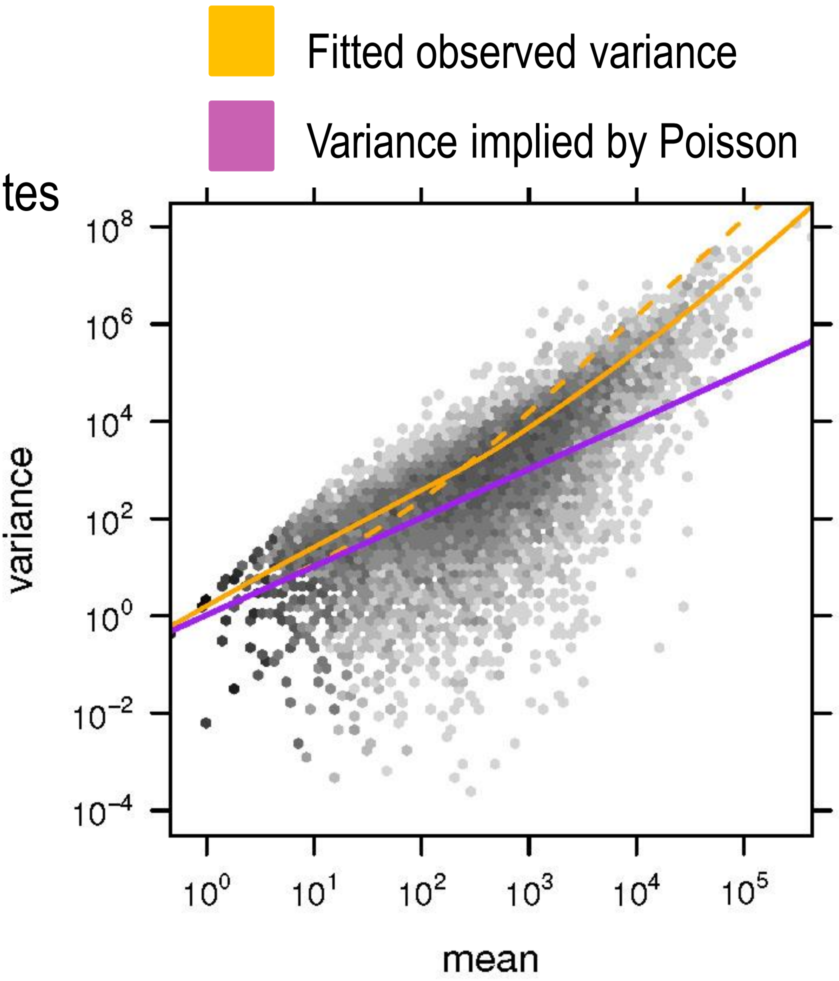
- A fundamental research problem in many RNA-seq studies is the identification of reliable molecular markers showing differential expression between sample groups (e.g. healthy and disease)

- A number of data analysis methods and pipelines have already been developed for this task

- BUT… there is no clear consensus about the best practices, which makes the choice of an appropriate method a daunting task

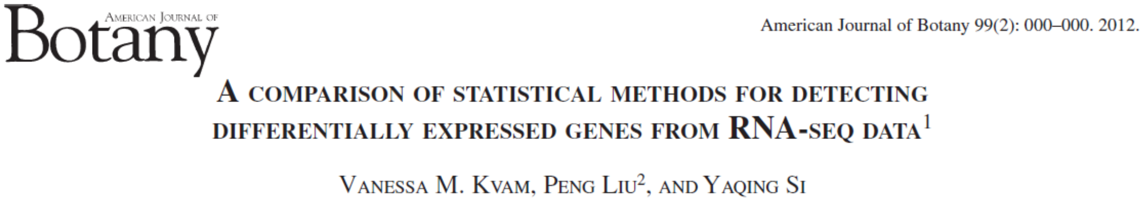
Data analysis challenges

- Normalization
  - Remove technical biases
  - Sequencing depth varies between replicates
- Small numbers of replicates
  - Accuracy of dispersion estimation
  - Permutation methods not effective
- Statistical model
  - Overdispersion



Anders and Huber, Genome Biol. 11:R106, 2010

Previous comparison studies

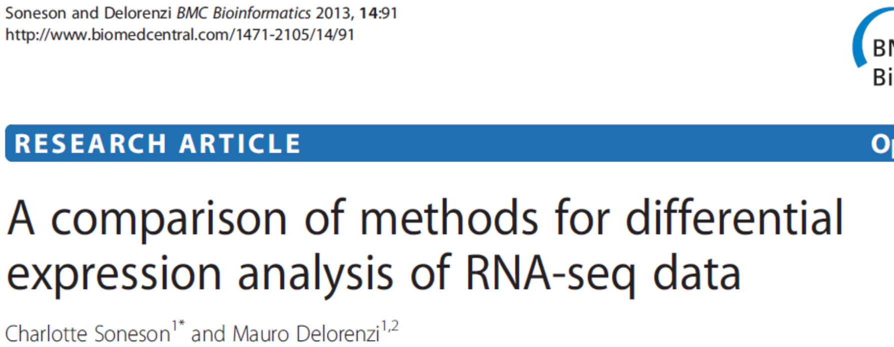


Botany American Journal of

American Journal of Botany 99(2): 000–000. 2012.

A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA[1]

VANESSA M. KVAM, PENG LIU[2], AND YAQING SI



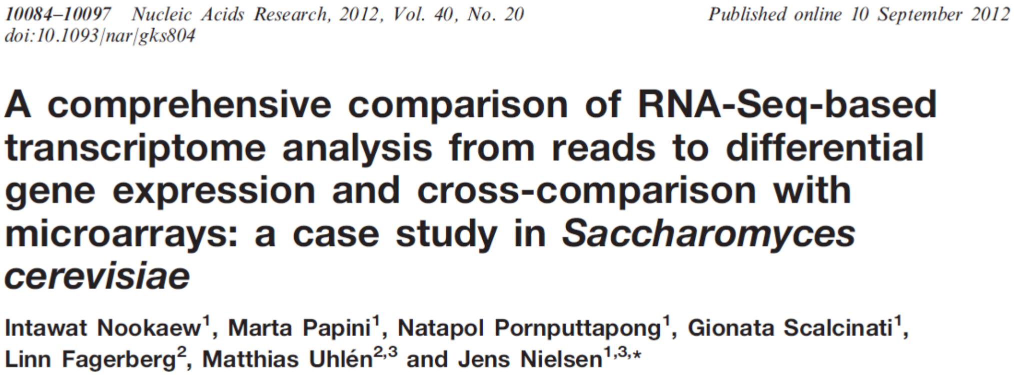
Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

A comparison of methods for differential expression analysis of RNA-seq data

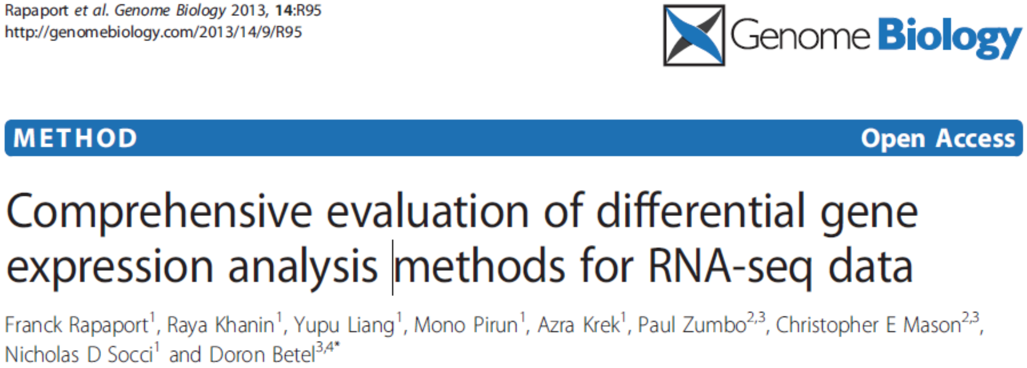
Charlotte Soneson[1*] and Mauro Delorenzi[1,2]



10084–10097 Nucleic Acids Research, 2012, Vol. 40, No. 20 *Published online 10 September 2012*
doi:10.1093/nar/gks804

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*

Intawat Nookaew[1], Marta Papini[1], Natapol Pornputtapong[1], Gionata Scalcinati[1], Linn Fagerberg[2], Matthias Uhlén[2,3] and Jens Nielsen[1,3,*]



Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

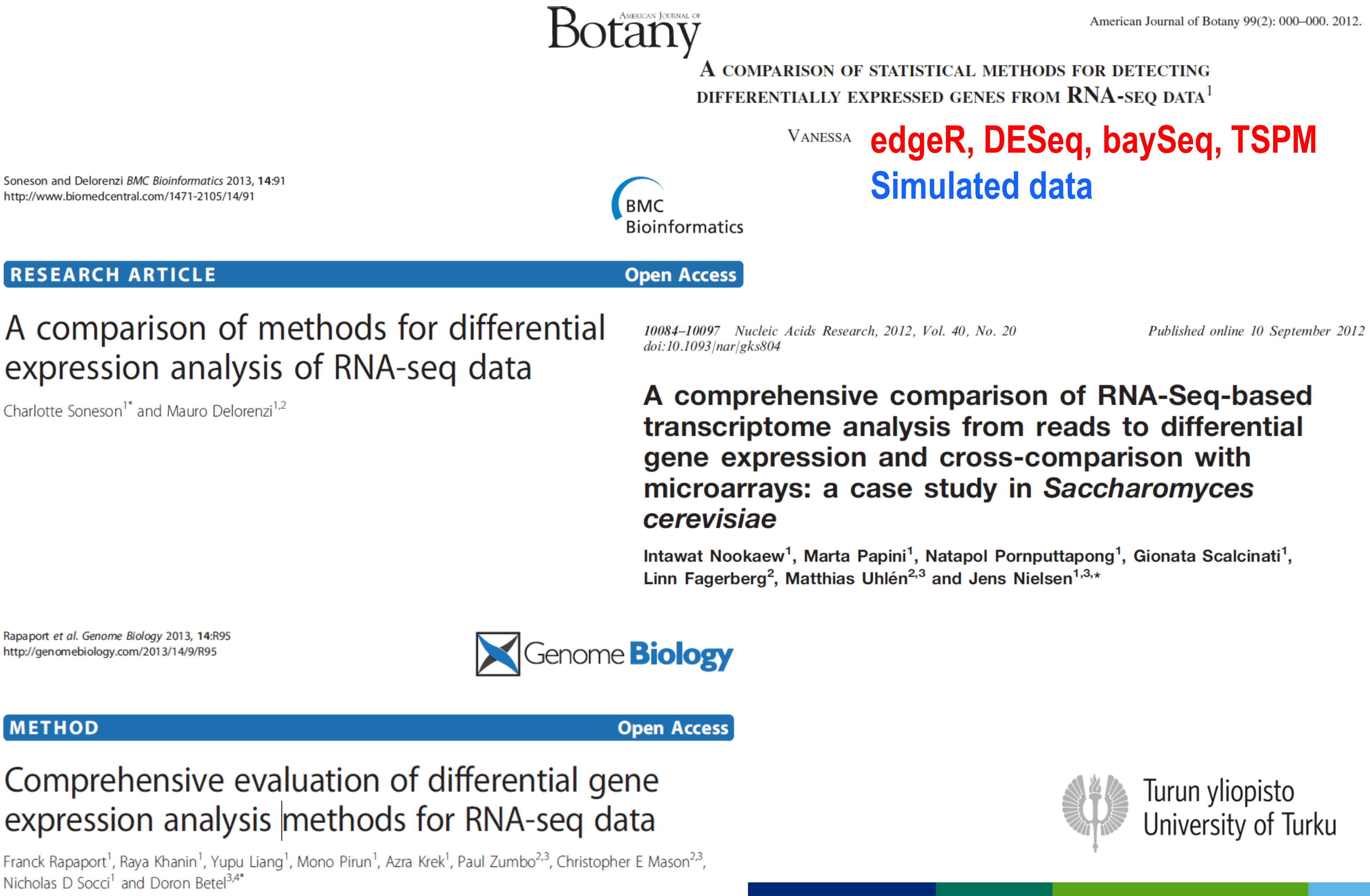
Genome Biology

METHOD **Open Access**

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Previous comparison studies



Botany AMERICAN JOURNAL OF

American Journal of Botany 99(2): 000–000. 2012.

A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA[1]

VANESSA

edgeR, DESeq, baySeq, TSPM

Simulated data

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

10084–10097 Nucleic Acids Research, 2012, Vol. 40, No. 20 *Published online 10 September 2012*
doi:10.1093/nar/gks804

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*

Intawat Nookaew[1], Marta Papini[1], Natapol Pornputtapong[1], Gionata Scalcinati[1], Linn Fagerberg[2], Matthias Uhlén[2,3] and Jens Nielsen[1,3,*]

Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

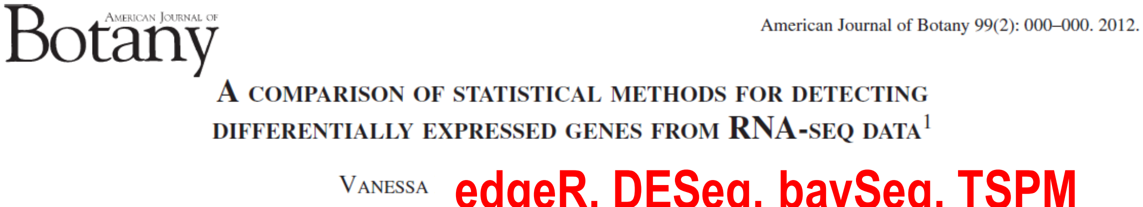
Genome Biology

METHOD **Open Access**

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Previous comparison studies



Botany AMERICAN JOURNAL OF

American Journal of Botany 99(2): 000–000. 2012.

A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA[1]

VANESSA

edgeR, DESeq, baySeq, TSPM

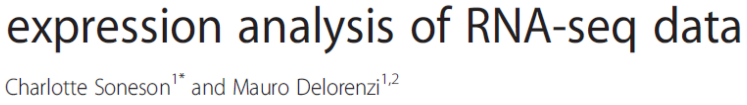
Simulated data

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

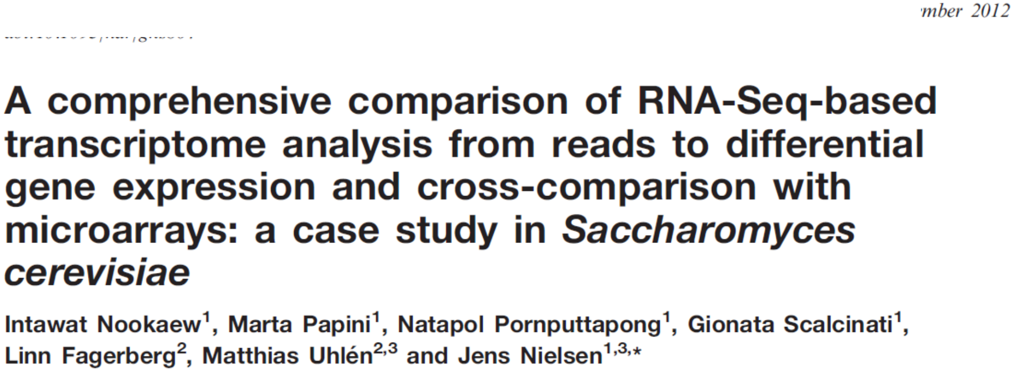
edgeR, DESeq, baySeq, TSPM, NOIseq, limma, EBSeq, SAMseq, NBPSeq, ShrinkSeq

Simulated data mainly



expression analysis of RNA-seq data

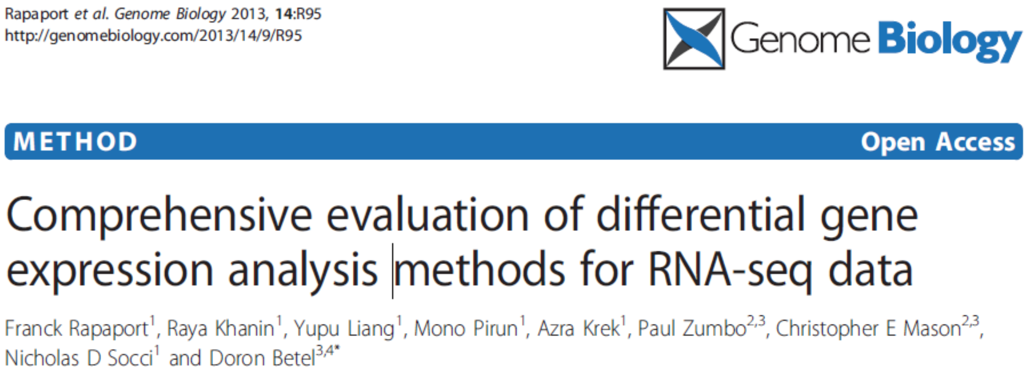
Charlotte Soneson[1*] and Mauro Delorenzi[1,2]



mber 2012

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*

Intawat Nookaew[1], Marta Papini[1], Natapol Pornputtapong[1], Gionata Scalcinati[1], Linn Fagerberg[2], Matthias Uhlén[2,3] and Jens Nielsen[1,3,*]



Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

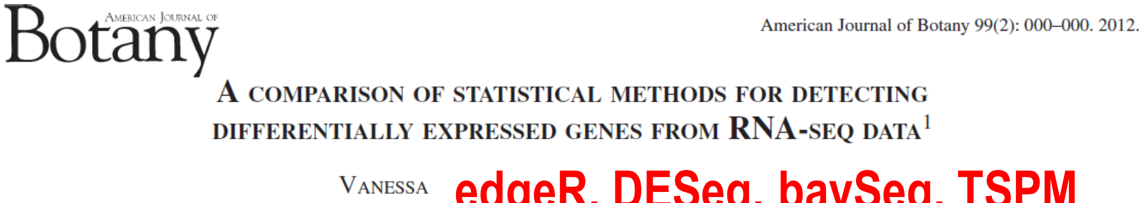
Genome Biology

METHOD **Open Access**

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Previous comparison studies



Botany
AMERICAN JOURNAL OF

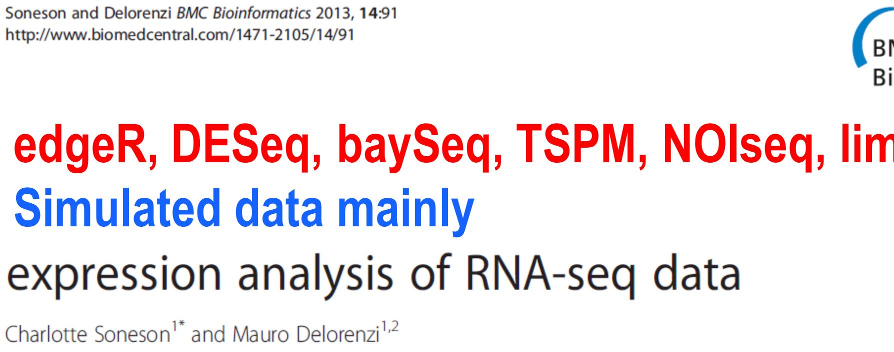
American Journal of Botany 99(2): 000–000. 2012.

A COMPARISON OF STATISTICAL METHODS FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA[1]

VANESSA

edgeR, DESeq, baySeq, TSPM

Simulated data



Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

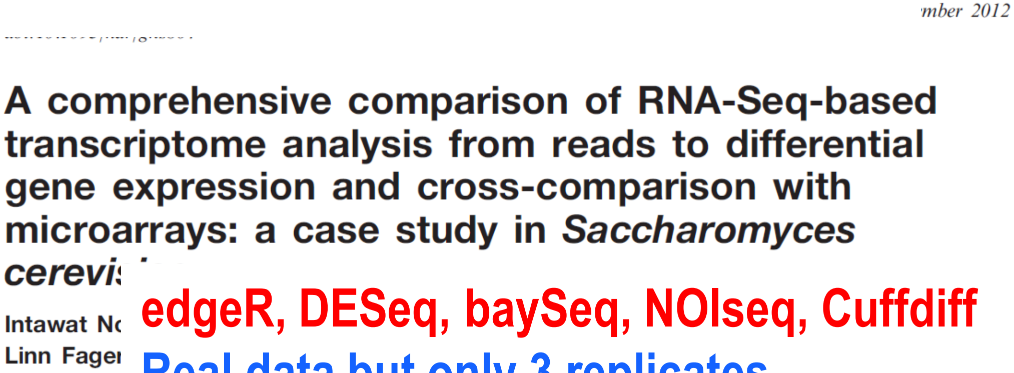
BMC Bioinformatics

expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

edgeR, DESeq, baySeq, TSPM, NOIseq, limma, EBSeq, SAMseq, NBPSeq, ShrinkSeq

Simulated data mainly



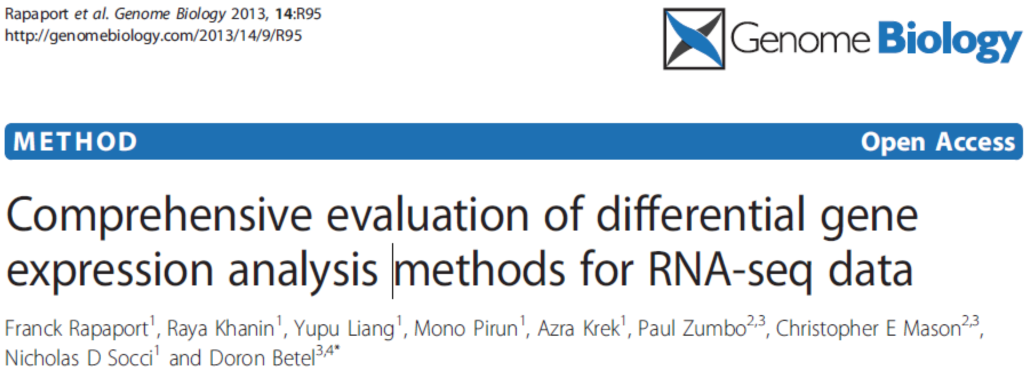
mber 2012

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevi:*

Intawat Nc
Linn Fager

edgeR, DESeq, baySeq, NOIseq, Cuffdiff

Real data but only 3 replicates



Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

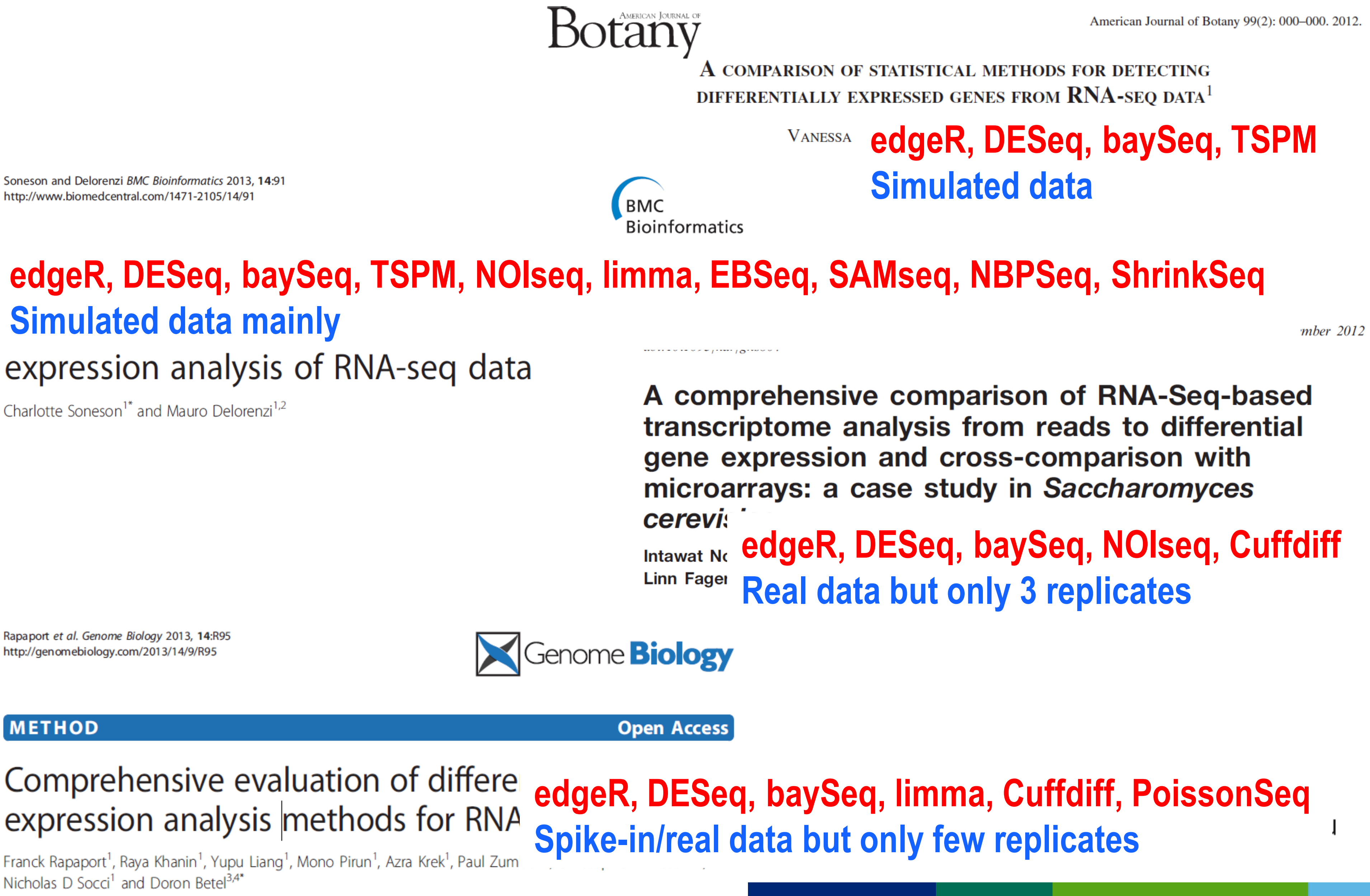
Genome **Biology**

METHOD **Open Access**

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]

Previous comparison studies



American Journal of Botany

American Journal of Botany 99(2): 000–000. 2012.

A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data[1]

Vanessa

edgeR, DESeq, baySeq, TSPM

Simulated data

Soneson and Delorenzi *BMC Bioinformatics* 2013, **14**:91
http://www.biomedcentral.com/1471-2105/14/91

BMC Bioinformatics

edgeR, DESeq, baySeq, TSPM, NOIseq, limma, EBSeq, SAMseq, NBPSeq, ShrinkSeq

Simulated data mainly

expression analysis of RNA-seq data

Charlotte Soneson[1*] and Mauro Delorenzi[1,2]

mber 2012

A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevi*

Intawat N

Linn Fage

edgeR, DESeq, baySeq, NOIseq, Cuffdiff

Real data but only 3 replicates

Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

Genome Biology

METHOD Open Access

Comprehensive evaluation of differe expression analysis methods for RNA

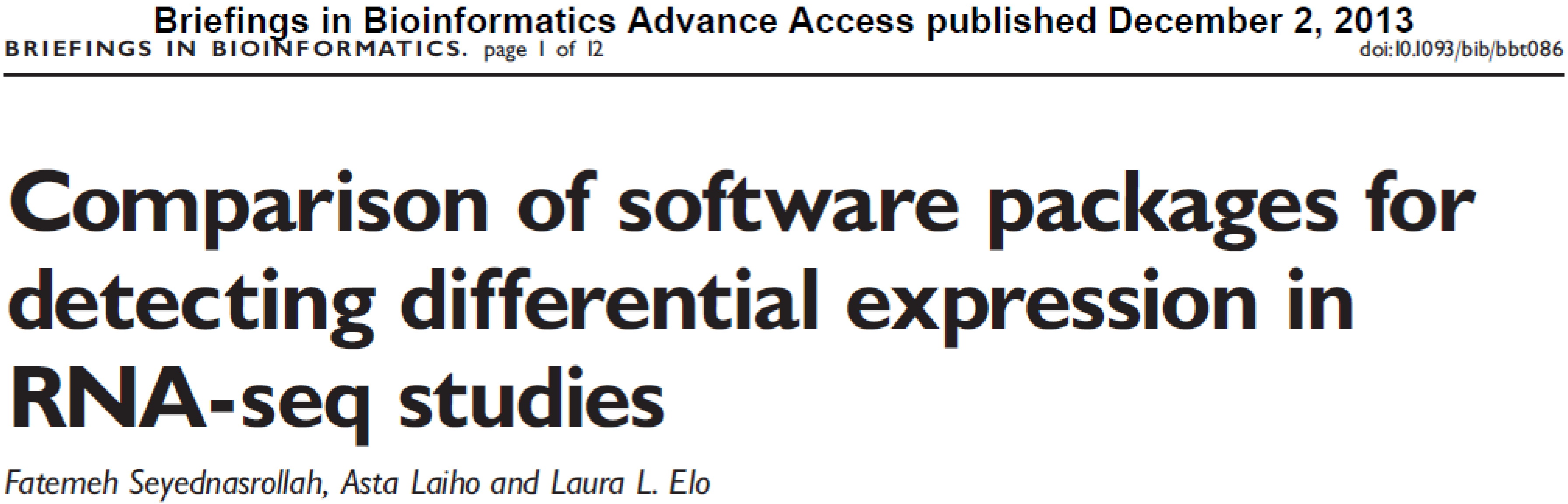
Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zum Nicholas D Socci[1] and Doron Betel[3,4*]

edgeR, DESeq, baySeq, limma, Cuffdiff, PoissonSeq

Spike-in/real data but only few replicates

Goal of this study

- To assist the choice of a robust pipeline for detecting differential expression between sample groups in a practical research setting



Briefings in Bioinformatics Advance Access published December 2, 2013

BRIEFINGS IN BIOINFORMATICS. page 1 of 12

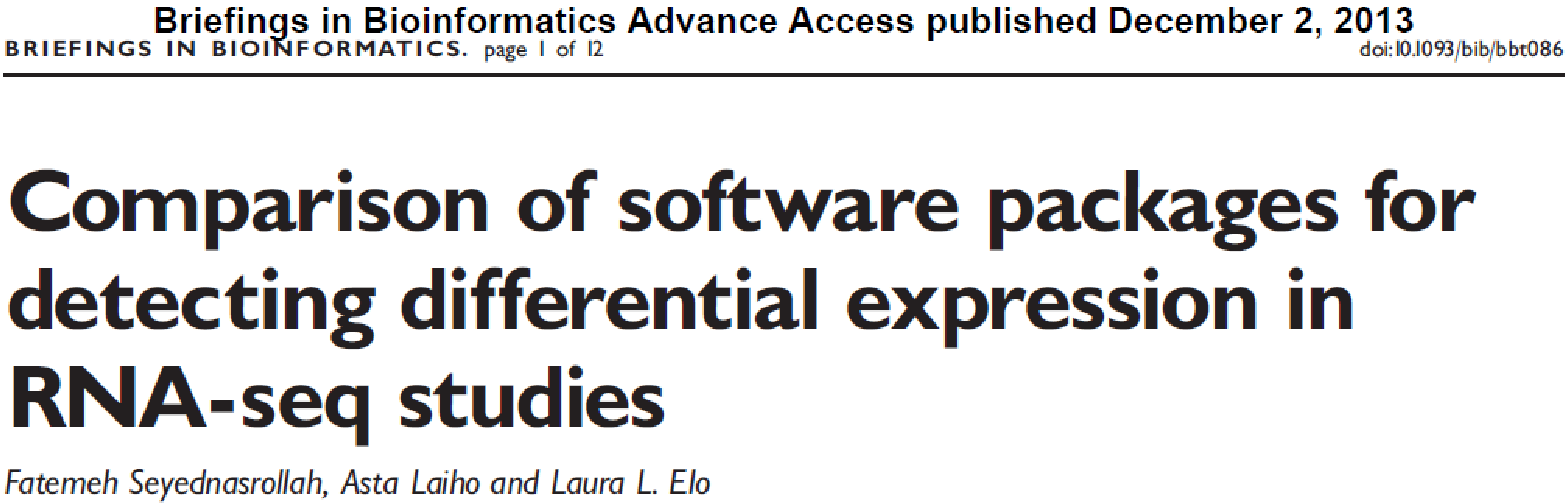
doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

Goal of this study

- To assist the choice of a robust pipeline for detecting differential expression between sample groups in a practical research setting



Briefings in Bioinformatics Advance Access published December 2, 2013

BRIEFINGS IN BIOINFORMATICS. page 1 of 12

doi:10.1093/bib/bbt086

Comparison of software packages for detecting differential expression in RNA-seq studies

Fatemeh Seyednasrollah, Asta Laiho and Laura L. Elo

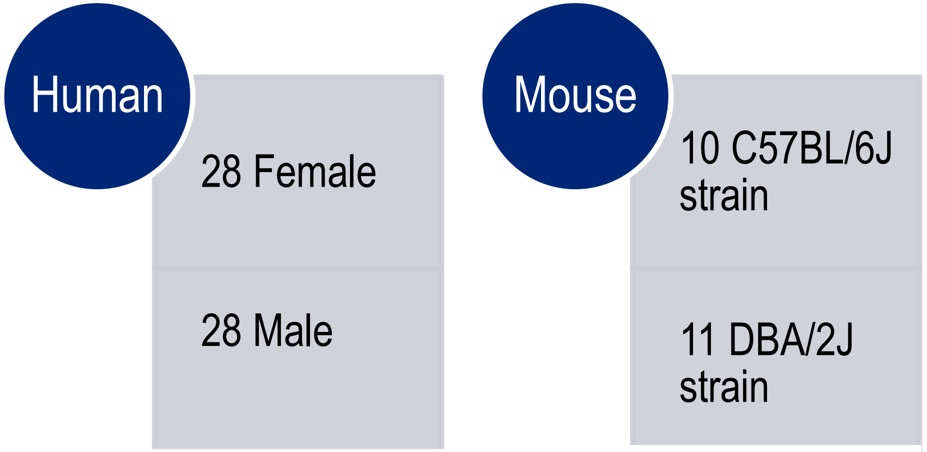
edgeR, DESeq, baySeq, NOIseq, limma, EBSeq, SAMseq, Cuffdiff 2

Real data with 2 to 28 replicates per group

Turun yliopisto
University of Turku

Datasets

- Two publicly available datasets generated by Illumina Genome Analyzer II platform
  - Publicly available to make the analysis reproducible
  - Large number of samples
  - Different level of heterogeneity
  - Different organisms



Human

28 Female

28 Male

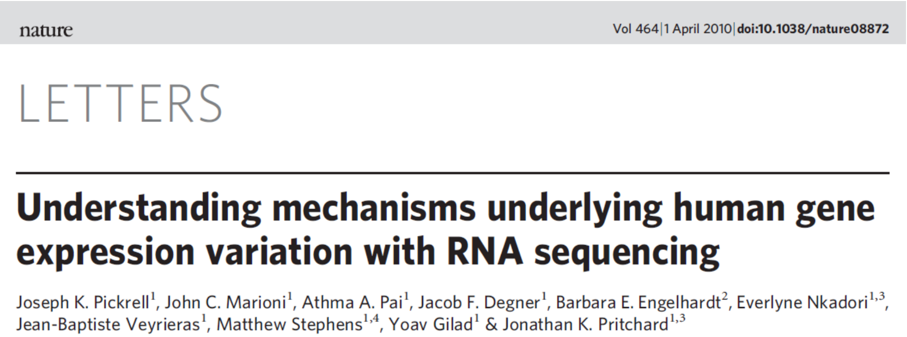
lymphoblastoid cell lines of unrelated Nigerian individuals

Mouse

10 C57BL/6J strain

11 DBA/2J strain

striatum samples

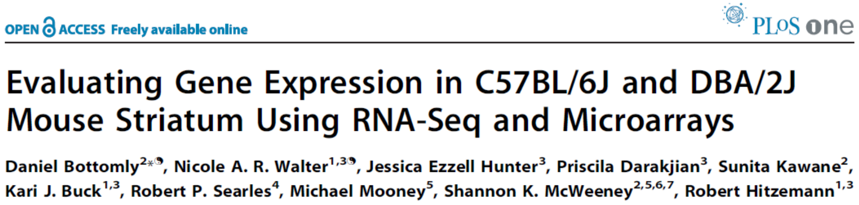


nature Vol 464|1 April 2010|doi:10.1038/nature08872

LETTERS

Understanding mechanisms underlying human gene expression variation with RNA sequencing

Joseph K. Pickrell[1], John C. Marioni[1], Athma A. Pai[1], Jacob F. Degner[1], Barbara E. Engelhardt[2], Everlyne Nkadori[1,3], Jean-Baptiste Veyrieras[1], Matthew Stephens[1,4], Yoav Gilad[1] & Jonathan K. Pritchard[1,3]



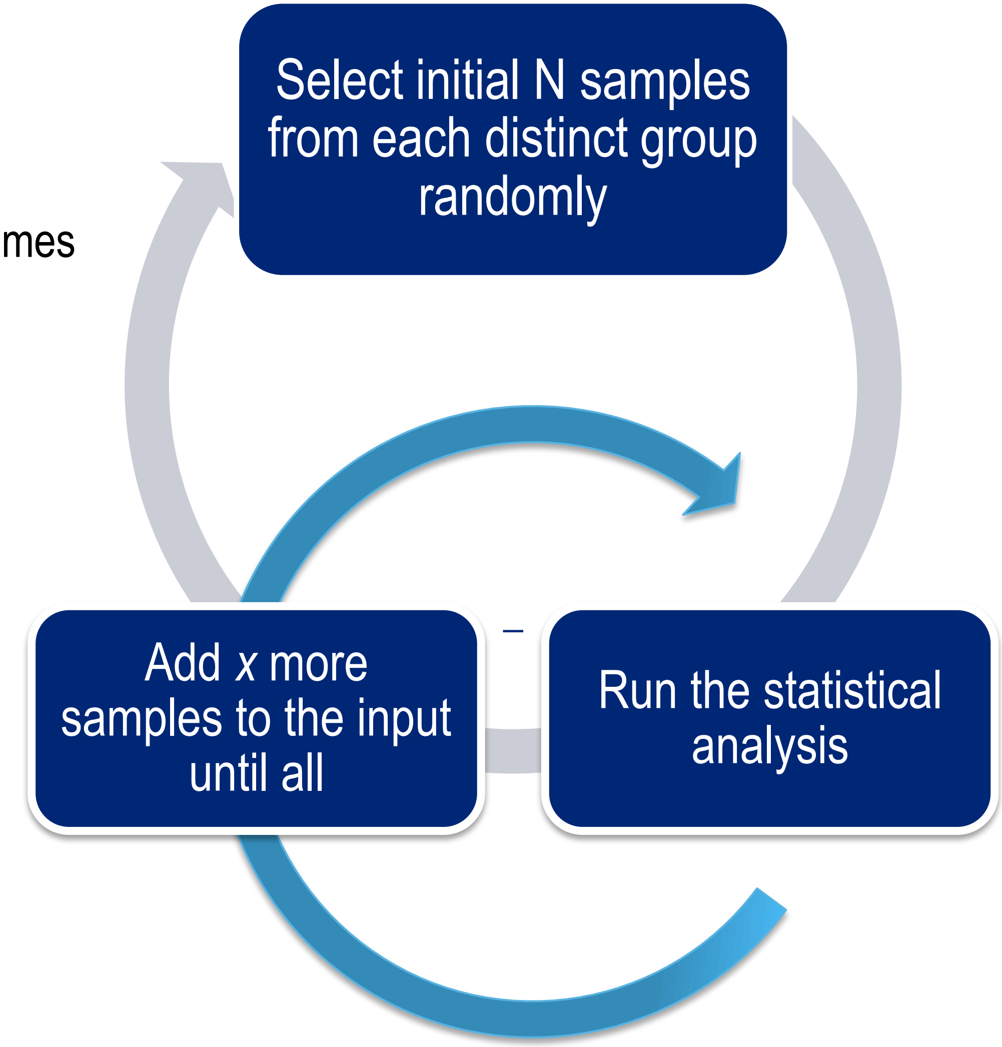
OPEN ACCESS Freely available online PLoS one

Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays

Daniel Bottomly[2*], Nicole A. R. Walter[1,3], Jessica Ezzell Hunter[3], Priscila Darakjian[3], Sunita Kawane[2], Kari J. Buck[1,3], Robert P. Searles[4], Michael Mooney[5], Shannon K. McWeeney[2,5,6,7], Robert Hitzemann[1,3]

Experimental Design

Repeat ten times

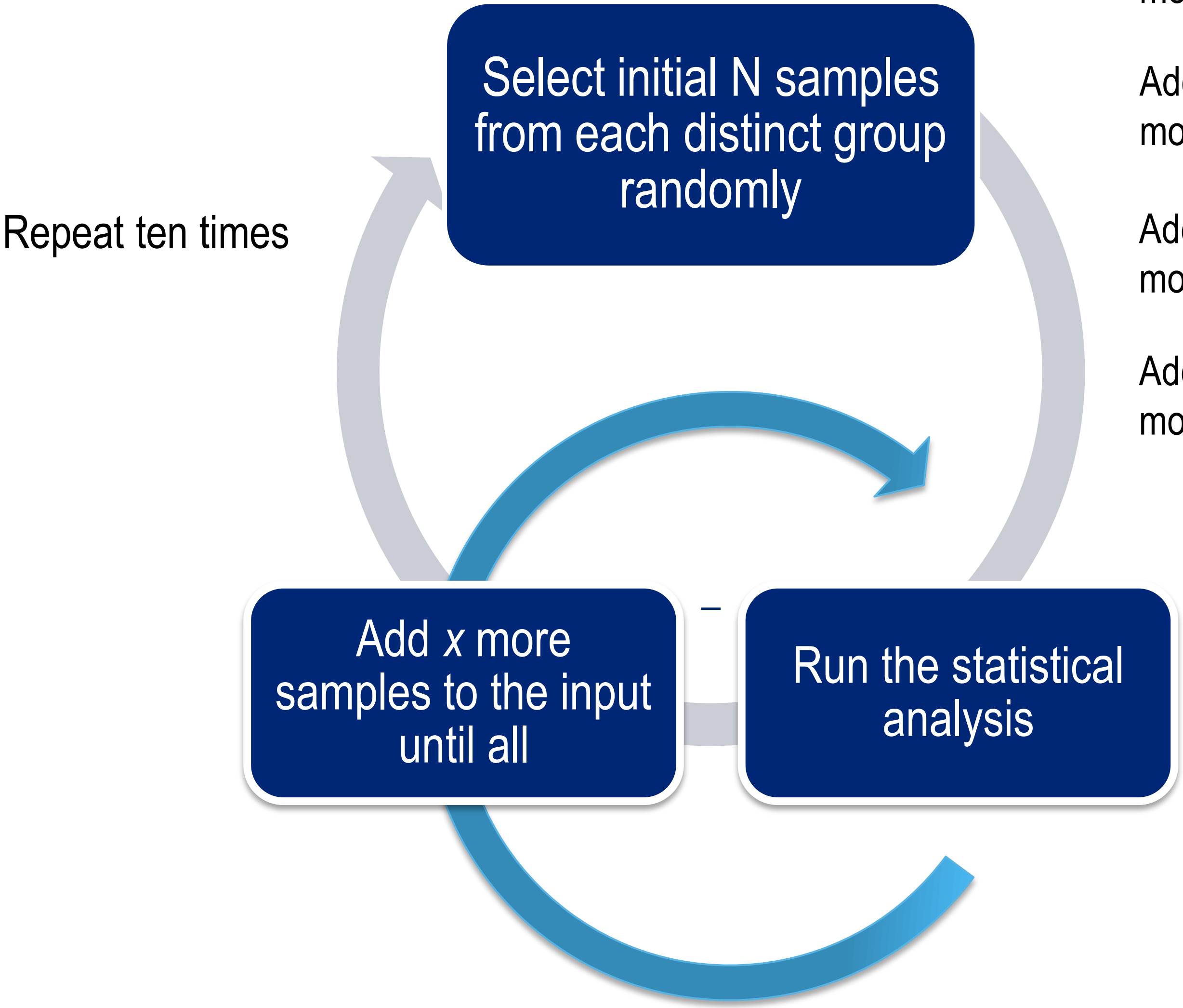


Select initial N samples from each distinct group randomly

Run the statistical analysis

Add *x* more samples to the input until all

Experimental Design

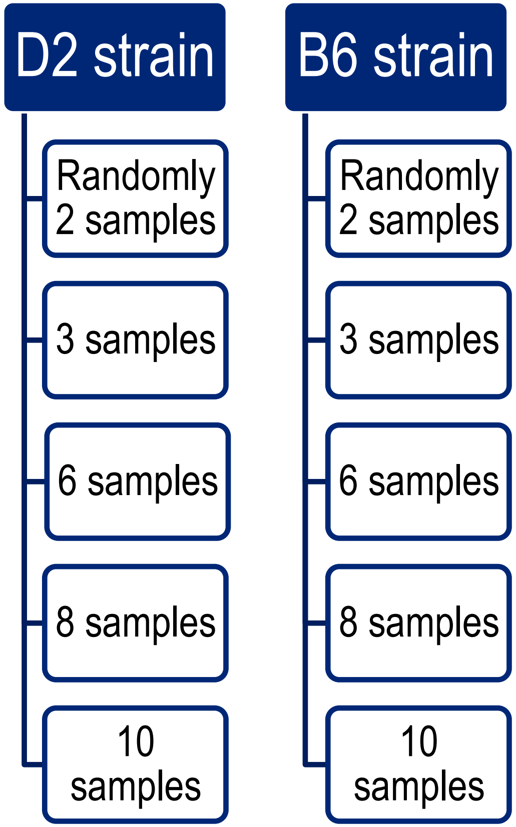


Repeat ten times

Select initial N samples from each distinct group randomly

Run the statistical analysis

Add *x* more samples to the input until all



D2 strain

Randomly 2 samples

3 samples

6 samples

8 samples

10 samples

B6 strain

Randomly 2 samples

3 samples

6 samples

8 samples

10 samples

Add more

Add more

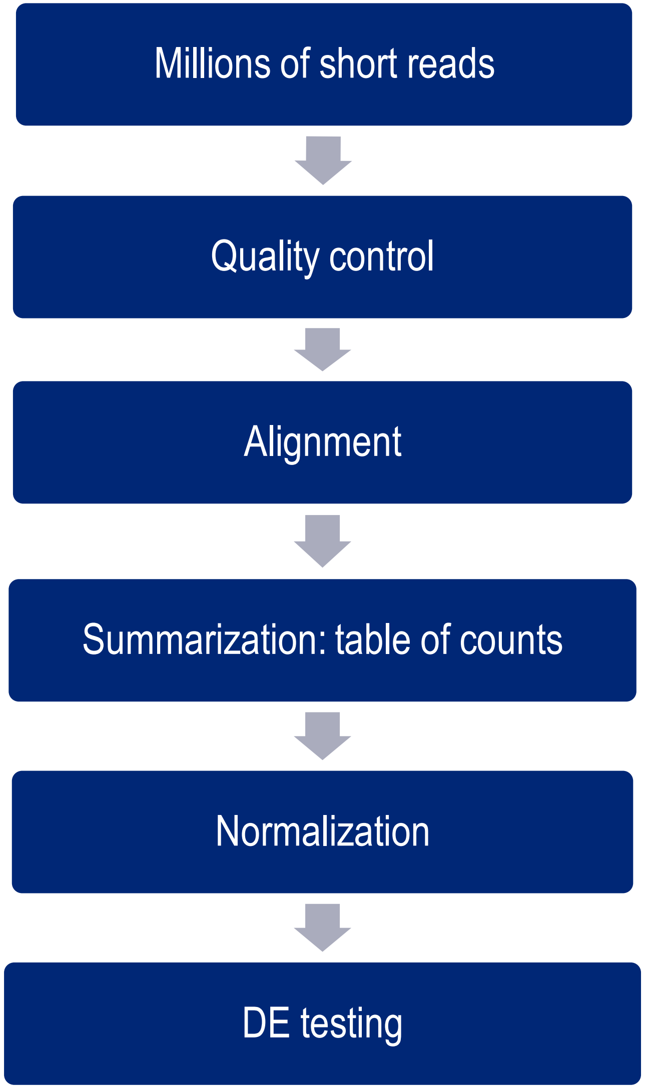
Add more

Add more

To estimate the false discoveries, we repeated the same procedure but within the groups (e.g., sampling within the group of female samples)

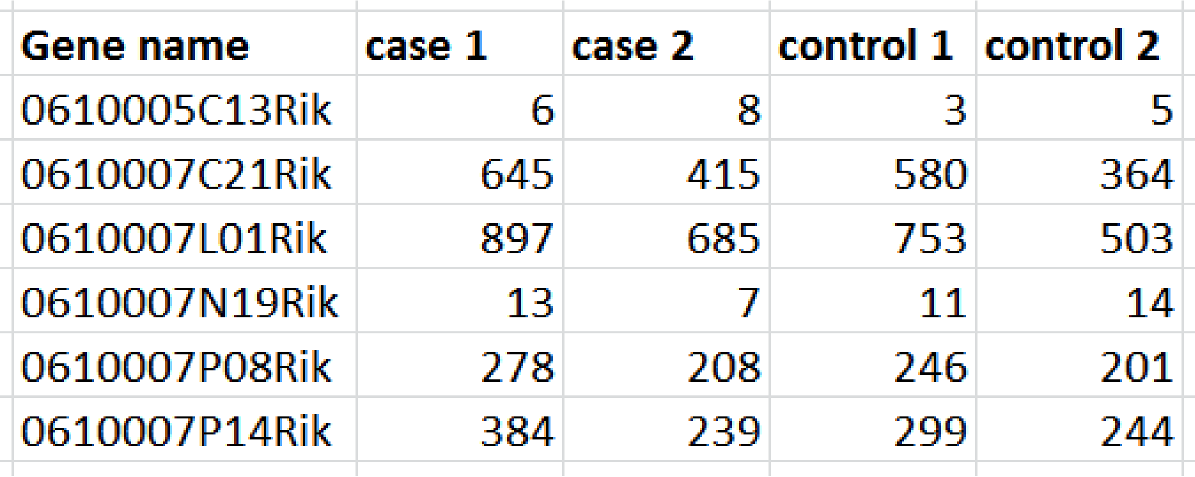
RNA-seq data analysis pipeline

- Quality control (fastq files)
  - FastQC
- Alignment
  - TopHat2 (RefSeq references)
  - Alignment rate in human 89% and mouse 86%
- Expression level quantification
  - HTSeq
  - Table of counts
- Normalization
  - Package default/TMM
  - TMM: Trimmed Mean of M values
- Statistical analysis
  - Eight state-of-the-art methods



Count tables

- Matrix of data with genomic features as rows and experiment samples as columns
- Is the difference between the conditions greater than what we expect taking into account normal biological variation? Can we detect reliable differentially expressed biomarkers?



Gene name	case 1	case 2	control 1	control 2
0610005C13Rik	6	8	3	5
0610007C21Rik	645	415	580	364
0610007L01Rik	897	685	753	503
0610007N19Rik	13	7	11	14
0610007P08Rik	278	208	246	201
0610007P14Rik	384	239	299	244

Software packages

Method	Normalization	Read counts distribution	Differential Expression Test
edgeR	TMM	Negative Binomial distribution	Exact test
DESeq	DESeq sizeFactors	Negative Binomial distribution	Exact test
Limma	TMM	Voom transformation of counts	Empirical Bayes method
NOISeq	RPKM/TMM/Upper Quantile	Non parametric method	compares the observed differences to null distribution (Contrasts fold changes and absolute differences within a condition)
baySeq	Scaling factors/TMM	Negative Binomial distribution	Empirical Bayesian Analysis
SAMseq	Method based on the mean read count over the null features of the data set	Non parametric method	Wilcoxon rank statistic and a resampling strategy
Cuffdiff2	DESeq like normalization	Beta Negative Binomial distribution	t-test
EBSeq	Median normalization	Negative Binomial distribution	Empirical Bayesian Analysis

Performance criteria

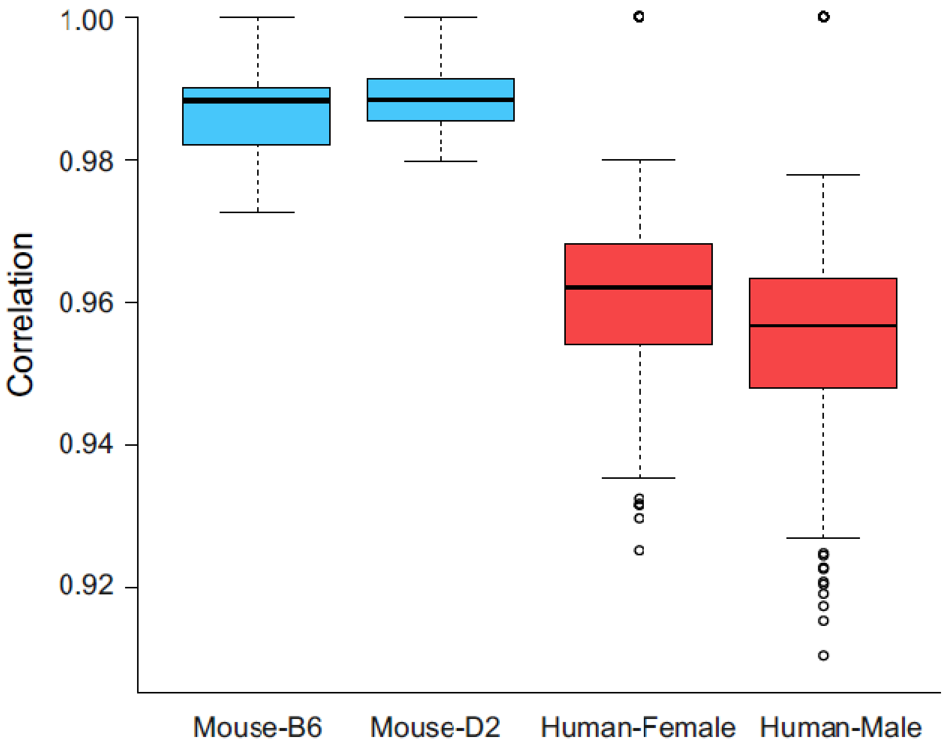
- Number of detections and their consistency
- False discoveries
- Correlation between methods
- Runtimes

- False discovery rate control FDR < 0.05
  - NOIseq did not report any FDR estimate (probability of differential expression > 0.8)

- Focus on default parameters and recommendations provided in the software manuals which are likely used by an average user

Data set intrinsic properties

- The mouse data are more homogenous than the human data

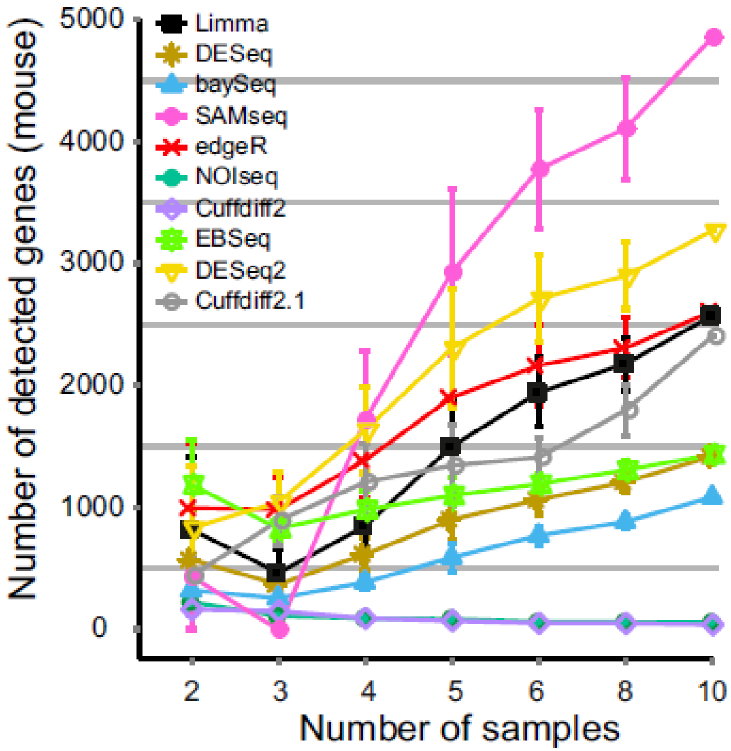


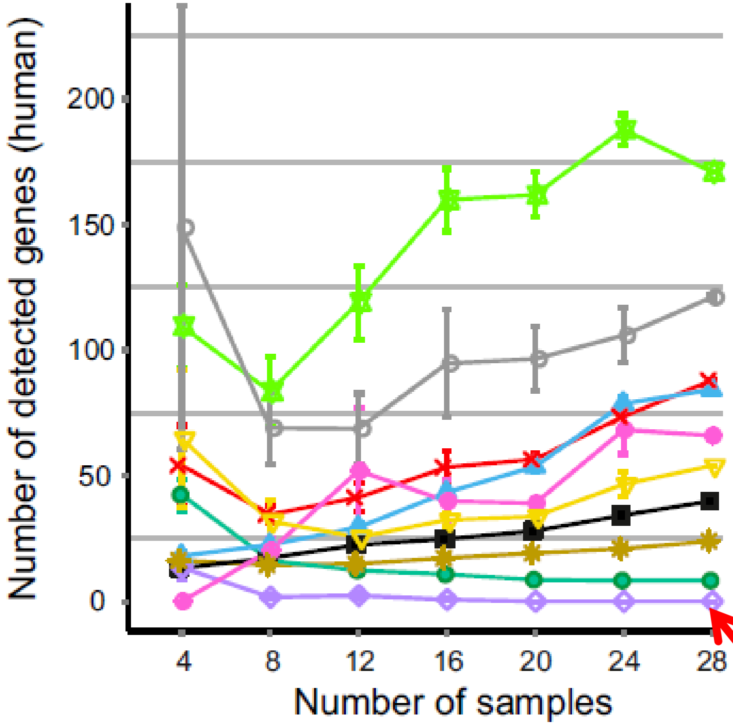
Turun yliopisto
University of Turku

Seyednasrollah et al. To appear in Brief Bioinf

Results: Number of detections

- Number of detections increased as the number of replicates increased, except for **NOIseq** and **Cuffdiff 2** (low power)





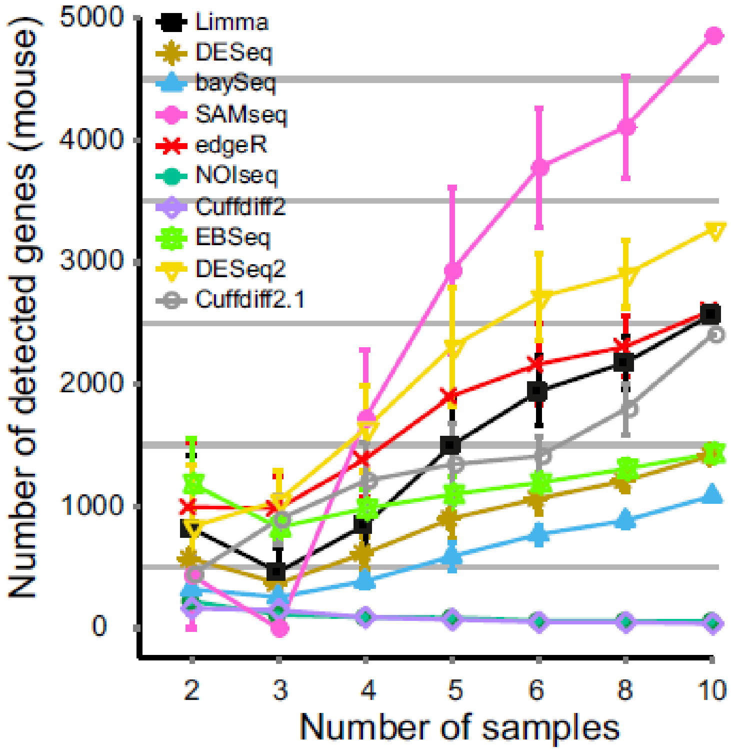
Cuffdiff 2.1

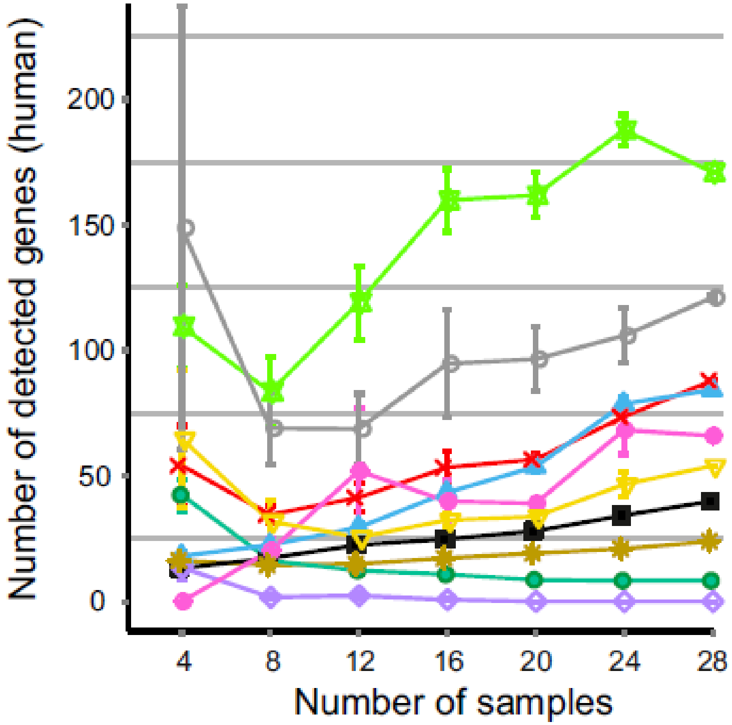
Cuffdiff 2: No detections in the complete human data

Seyednasrollah et al. To appear in Brief Bioinf

Results: Number of detections

- Moderate: **DESeq** (more conservative) and **Limma**
- Liberal: **edgeR** and **SAMseq** (except for smallest numbers of replicates)
- Data dependent: **baySeq** and **EBseq**

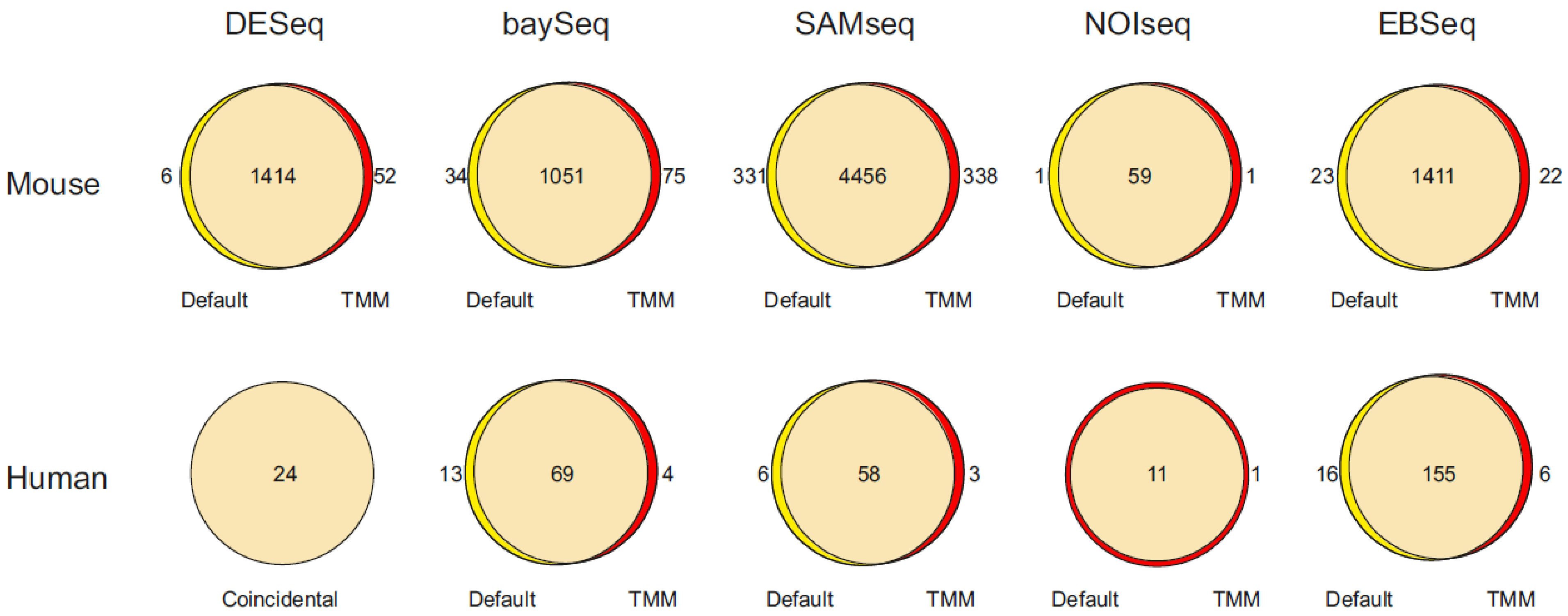




Seyednasrollah et al. To appear in Brief Bioinf

Effect of normalization on the detections

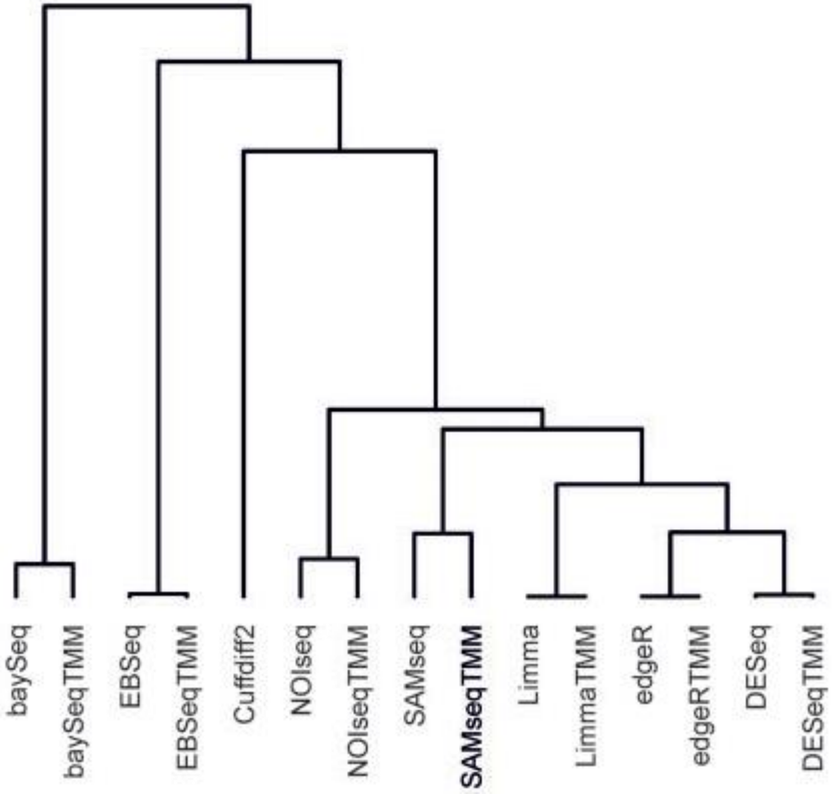
- The package default normalization and the TMM normalization produced highly overlapping detections (>80%)



Seyednasrollah et al. To appear in Brief Bioinf

Effect of normalization on the detections

- Comparison of the gene rankings confirmed the overall similarity of the results

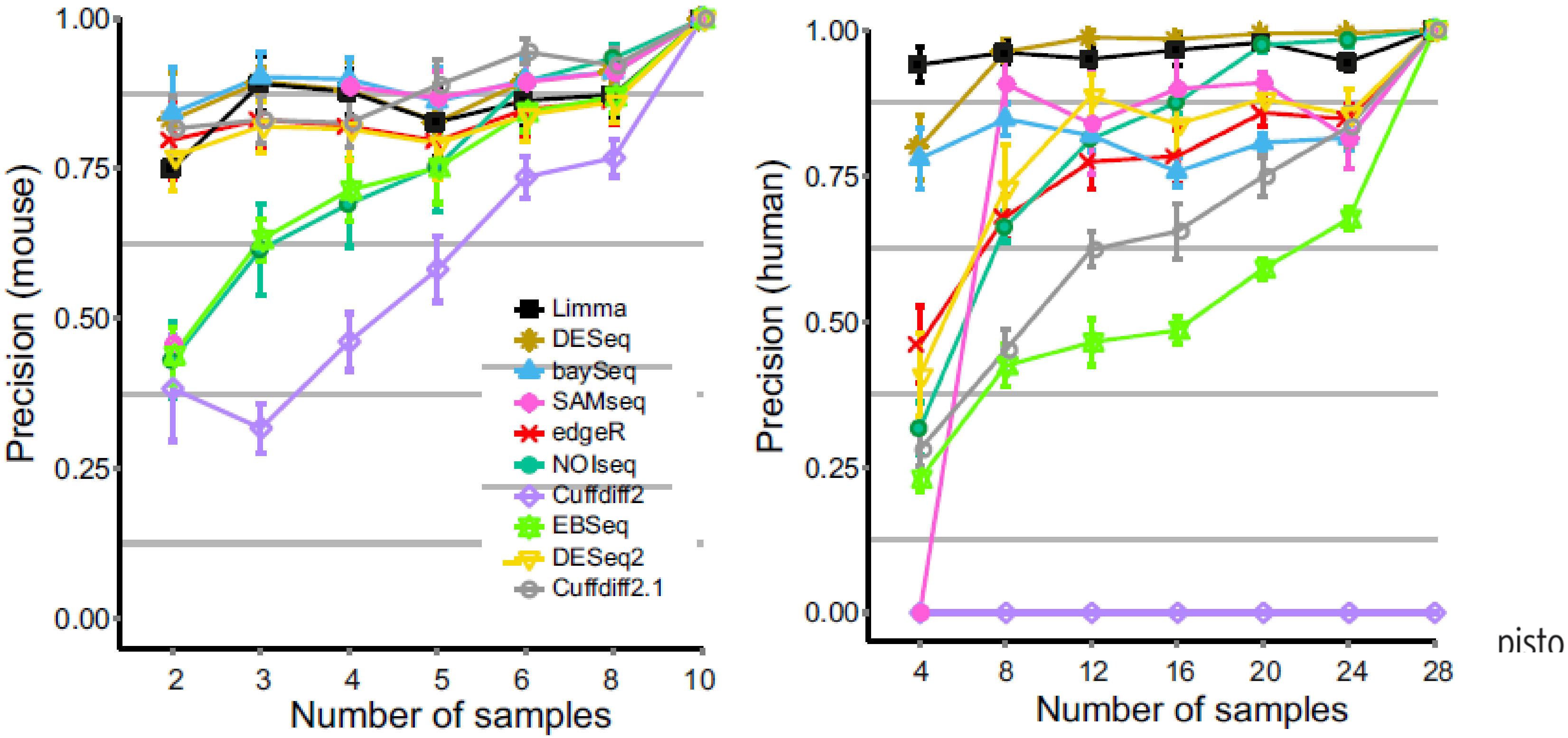


Turun yliopisto
University of Turku

Seyednasrollah et al. To appear in Brief Bioinf

Results: Consistency of detections

- Overlap of detections between the subdatasets and the complete data
  - Generally highest with **DESeq** and **Limma**
  - Generally lowest with **NOIseq**, **Cuffdiff 2** and **EBseq**

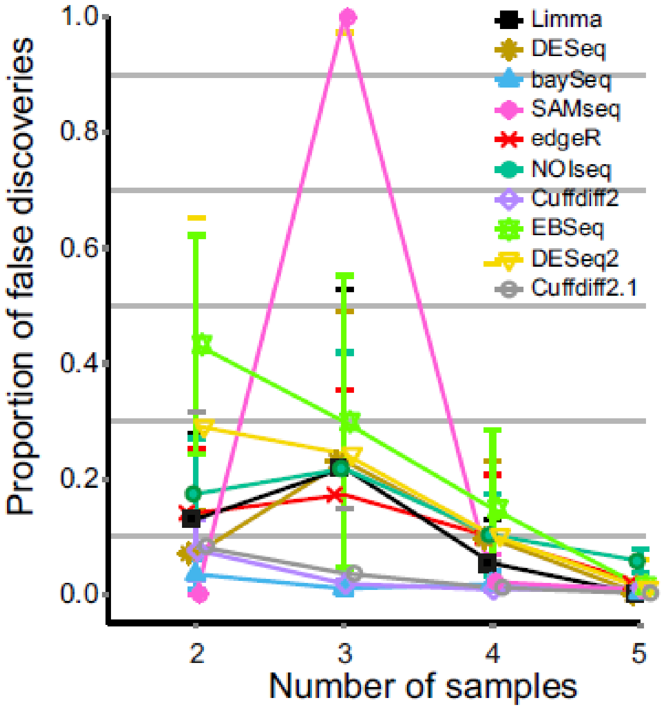


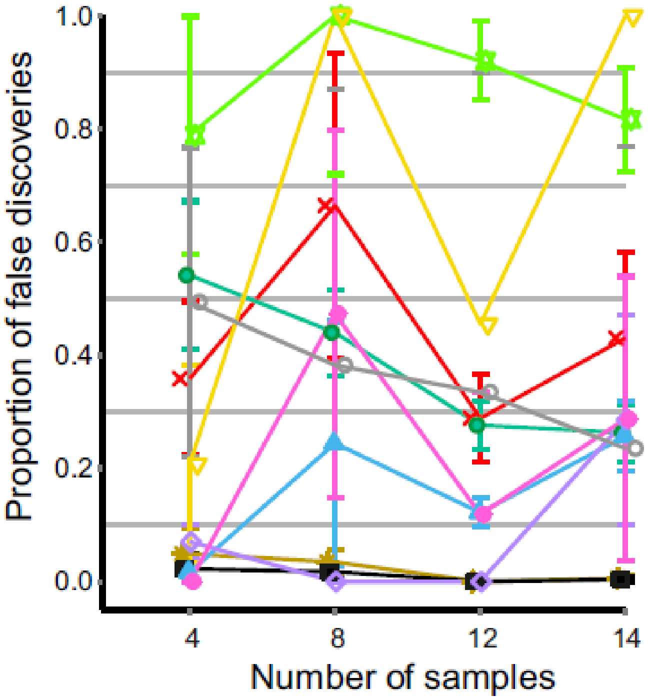
nisto

Seyednasrollah et al. To appear in Brief Bioinf

Results: False discoveries

- Number of false discoveries decreased when the number of replicates was increased, especially in less heterogeneous data (mouse)
  - In general, **Limma**, **DESeq** and **baySeq** performed well
  - **EBseq**, **SAMseq**, **edgeR** and **NOIseq** identified relatively many false positives





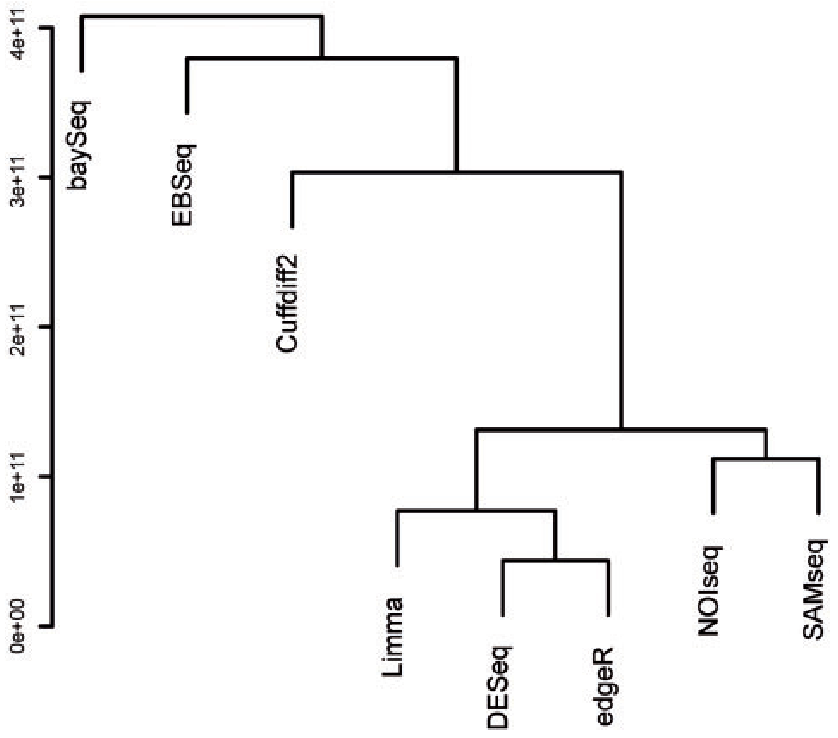
ın

ve

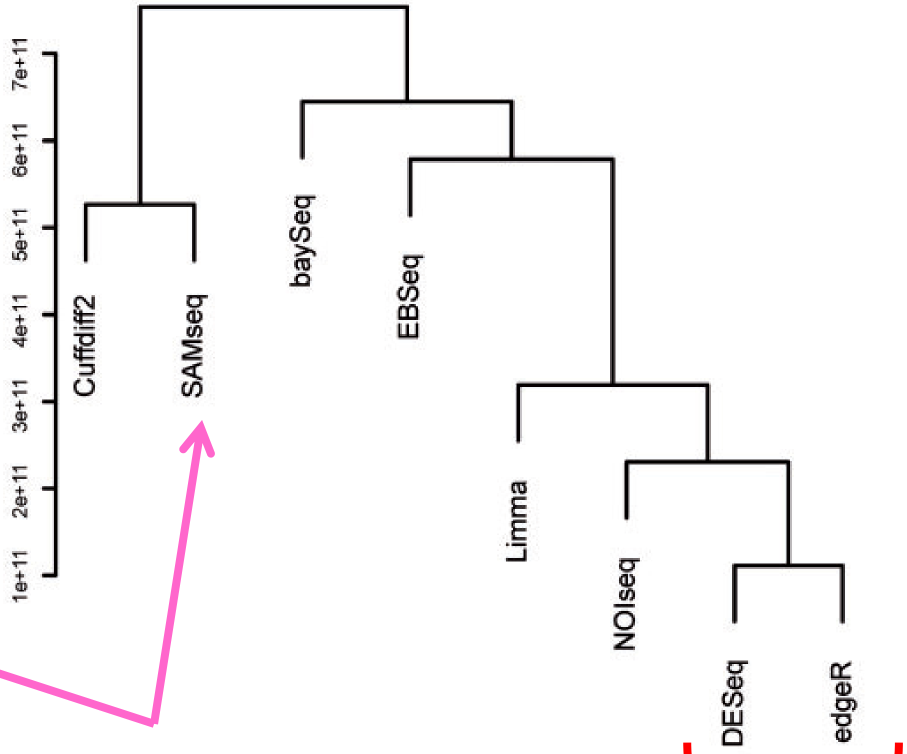
Seyednasrollah et al. To appear in Brief Bioinf

Results: Similarity between the methods

Mouse



Human



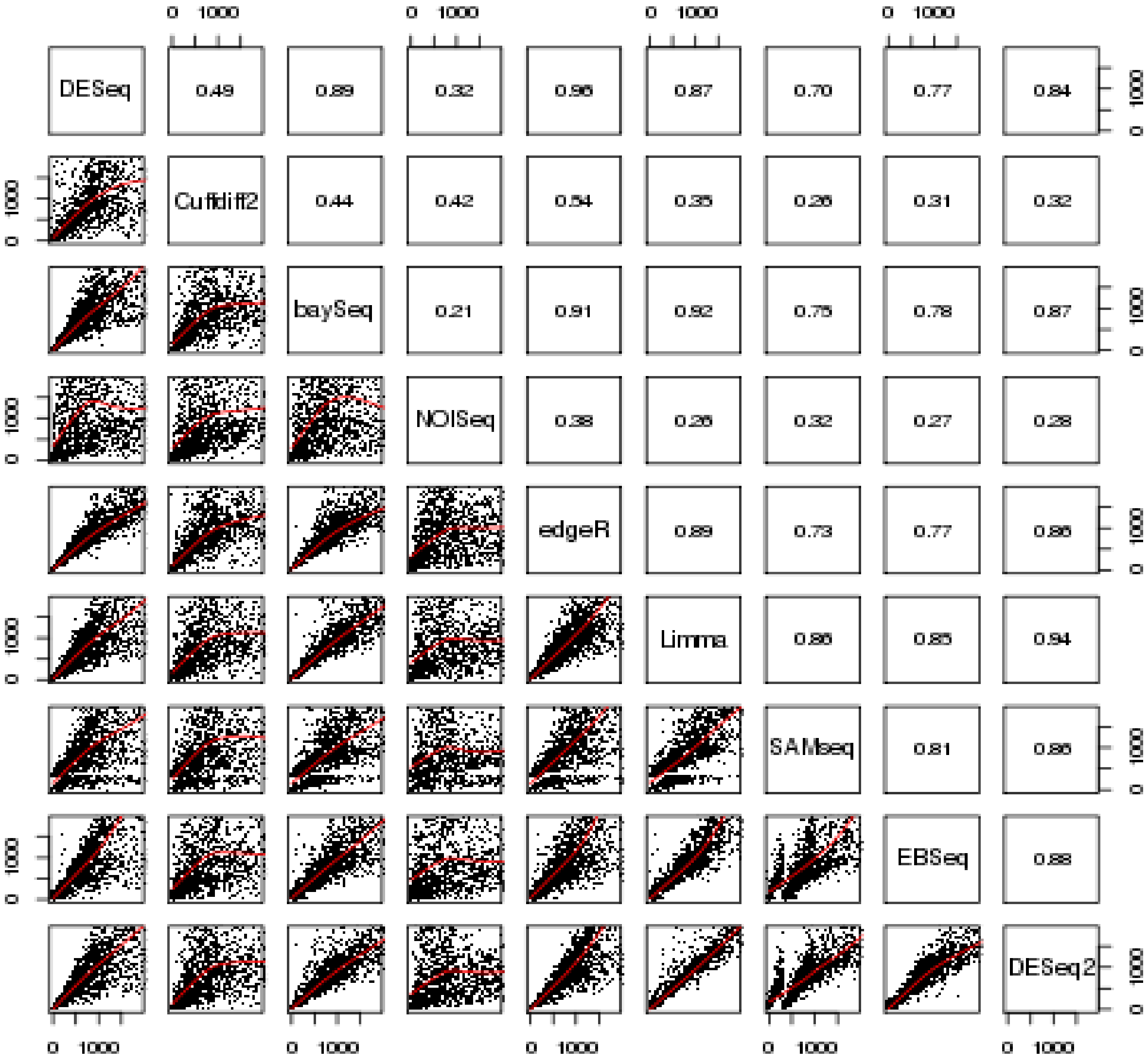
Same underlying statistical model

Dependent
on the data

Seyednasrollah et al. To appear in Brief Bioinf

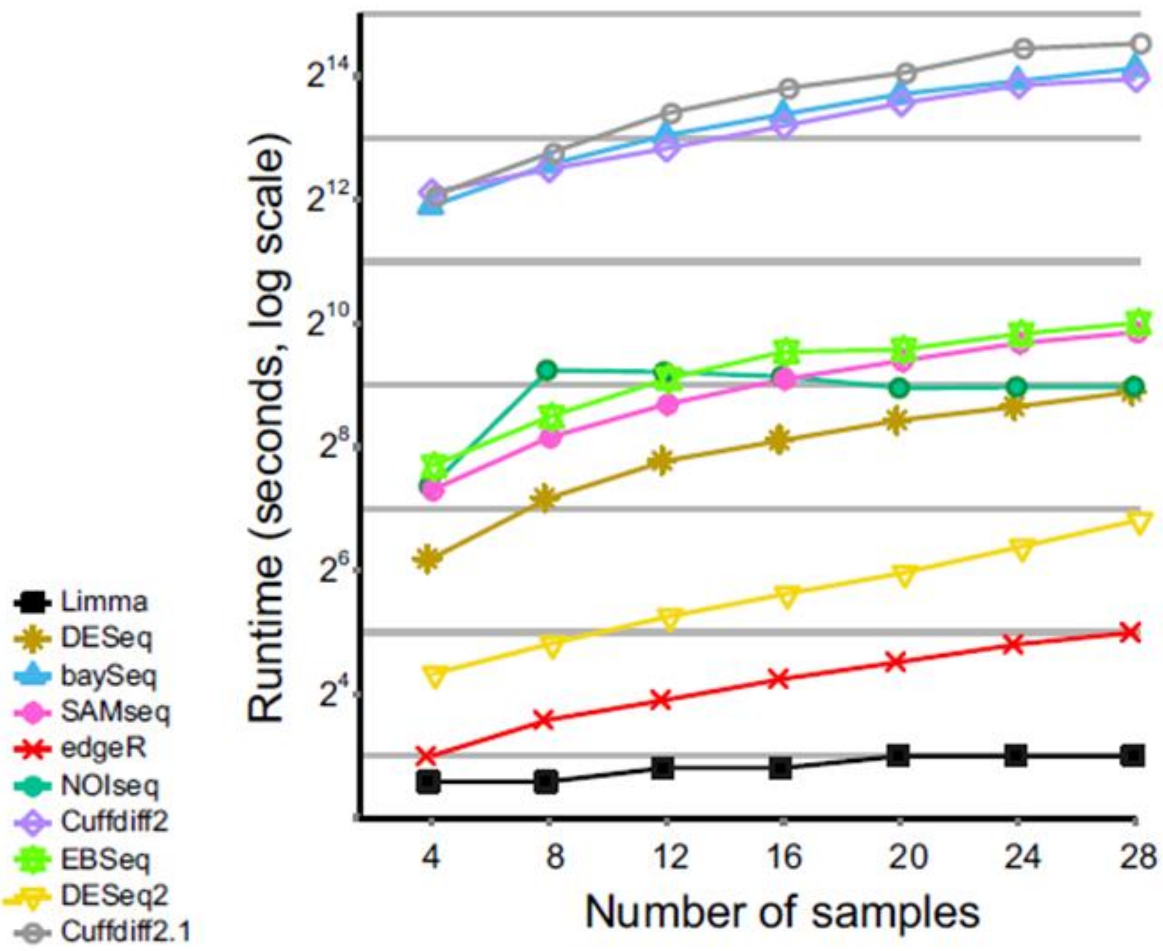
Results: Similarity between the methods

Those 1952 genes that were among the top 1000 ranked genes within any of the methods in the mouse data and the corresponding Spearman rank correlations



Seyednasrollah et al. To appear in Brief Bioinf

Results: Runtimes



Hours
Cuffdiff 2.1
Cuffdiff 2
baySeq

Seconds
Limma
edgeR

Turun yliopisto
University of Turku

Seyednasrollah et al. To appear in Brief Bioinf

Conclusions

- There can be large differences in the results obtained with the different software packages
- The choice of the normalization method had surprisingly little influence on the outcome
- Differences between the results obtained using different versions of the software packages can be significant
- No single method is likely to be optimal under all circumstances
- Marked differences in the quality and detail of the documentation of the pipelines

Relation to other comparison studies

- Overall, our observations in real data complemented well the previous observations by Soneson and Delorenzi in simulated data

- DESeq was often relatively conservative
- edgeR and EBSeq were often too liberal
- SAMseq performed well only when the number of replicates was relatively large
- Performance of baySeq was highly variable depending on the data
- Limma performed generally well under many circumstances

General guidelines

- Robust performance under many circumstances?
  - Limma and DESeq (more conservative)

- Do you have small number of biological replicates (say <5)?
  - Take the results with caution
  - It may be informative to consider more than one software package
  - We do not recommend non-parametric approaches like SAMseq
- Do you have more than five replicates?
  - Avoid using NOIseq and Cuffdiff 2
  - With relatively large numbers of replicates (say >10) non-parametric methods like SAMseq may be useful

- Investigate the properties of the data in advance

Acknowledgements

- Computational Biomedicine Group, Turku Centre for Biotechnology
  - An Le Thi Thanh, PhD
  - Tomi Suomi, MSc
  - Daniel Laajala, MSc
  - Anna Pursiheimo, MSc
  - Maria Jaakkola, MSc
  - Kalaimathy Singaravelu, MSc
  - Deepankar Chakroborty, BSc
  - Bishwa Ghimire, MSc (FMSC)

Fatemeh Seyednasrollah

Asta Laiho, MSc Tech
FMSC

- Academy of Finland, JDRF, Päivikki and Sakari Sohlberg Foundation, Yrjö Jahnsson Foundation

Turun yliopisto
University of Turku