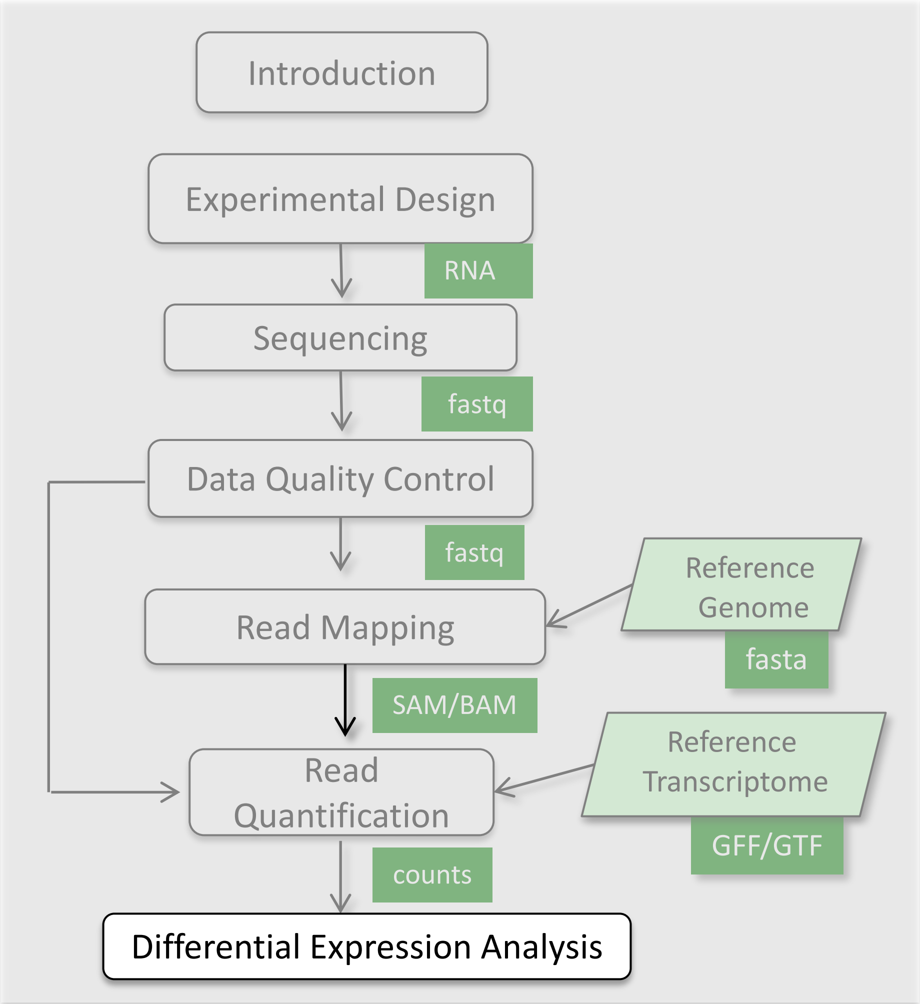
Differential Expression Analysis

Falko Noé

University of Zurich UZH | ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich



Differential expression analysis

- General idea
- Tools
- Exploration of differential expression results

What is differential expression?

Differential expression is the **assessment** of differences in read counts between two or more experimental conditions. Genes are differentially expressed if this difference is **statistically significant**.

Differential expression testing: General Idea

- **Each** gene that has been identified above a certain threshold of expression is **independently** tested for difference between two groups
- The test usually relies on the count data of the gene in the samples involved in the testing (e.g., 3 replicates per group)
- The normalization is applied when the comparison is performed
- The output is a table which associates to each gene a series of statistics, notably the **p-value** of the test and **log2ratio** (or fold-change)

Hypothesis testing

Is gene g differentially expressed between the conditions?

1. Formulate hypotheses:

H0: gene g is NOT differentially expressed between the conditions

H1: gene g is differentially expressed between the conditions

Hypothesis testing

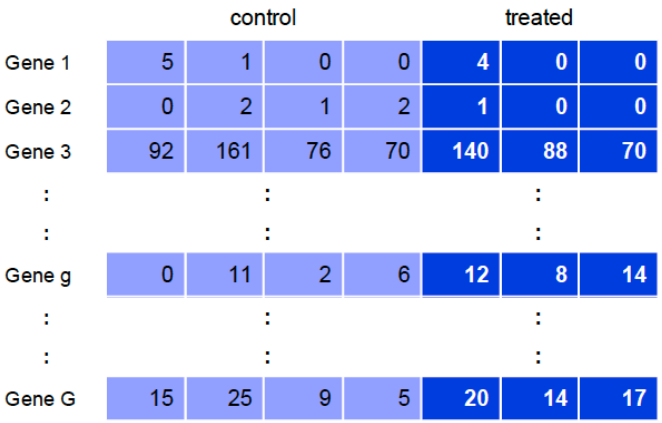
Is gene g differentially expressed between the conditions?

1. Formulate two hypotheses:

H0: gene g is NOT differentially expressed between the conditions

H1: gene g is differentially expressed between the conditions

2. Collect appropriate data



	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:		:				:	
:		:				:	
Gene g	0	11	2	6	12	8	14
:		:				:	
:		:				:	
Gene G	15	25	9	5	20	14	17

Hypothesis testing

Is gene g differentially expressed between the conditions?

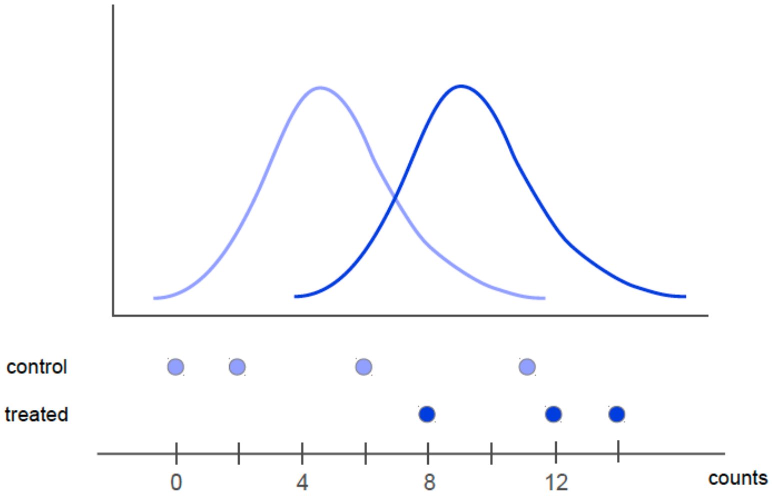
1. Formulate two hypotheses:

   H0: gene g is NOT differentially expressed between the conditions

   H1: gene g is differentially expressed between the conditions

2. Collect appropriate data

3. Fit the model for each gene



Hypothesis testing

Is gene g differentially expressed between the conditions?

1. Formulate two hypotheses:

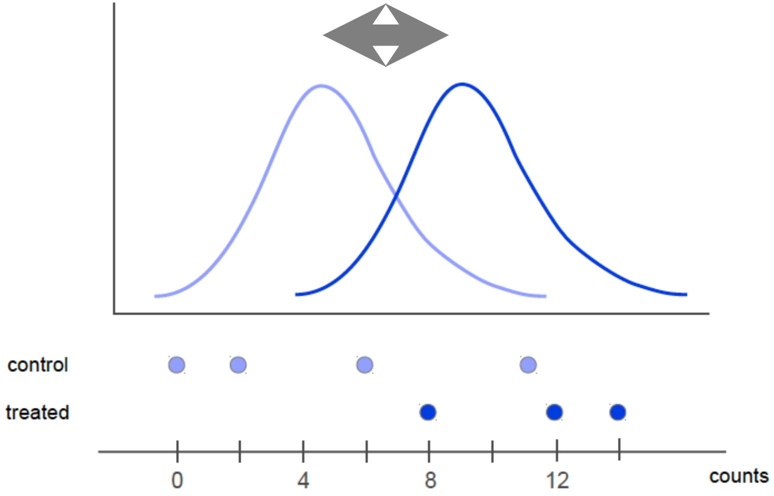
   H0: gene g is NOT differentially expressed between the conditions

   H1: gene g is differentially expressed between the conditions

2. Collect appropriate data

3. Fit the model for each gene

4. Use statistics to quantify the difference



Hypothesis testing

Is gene g differentially expressed between the conditions?

1. Formulate two hypotheses:

   H0: gene g is NOT differentially expressed between the conditions

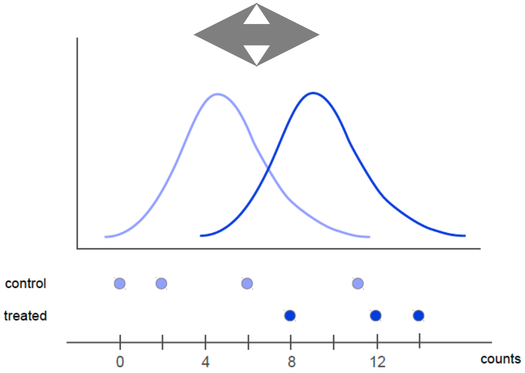
   H1: gene g is differentially expressed between the conditions

2. Collect appropriate data

3. Fit the model for each gene

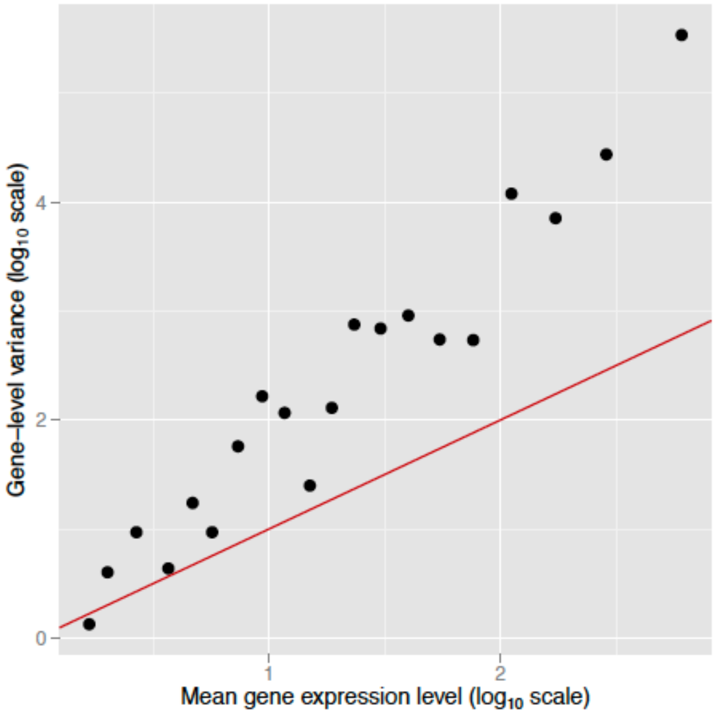
4. Use statistics to quantify the difference

- Small p-value: small chance that the observed result is due to pure coincidence
- Large p-value: large chance the the observed result is caused by random noise



Testing for differential expression

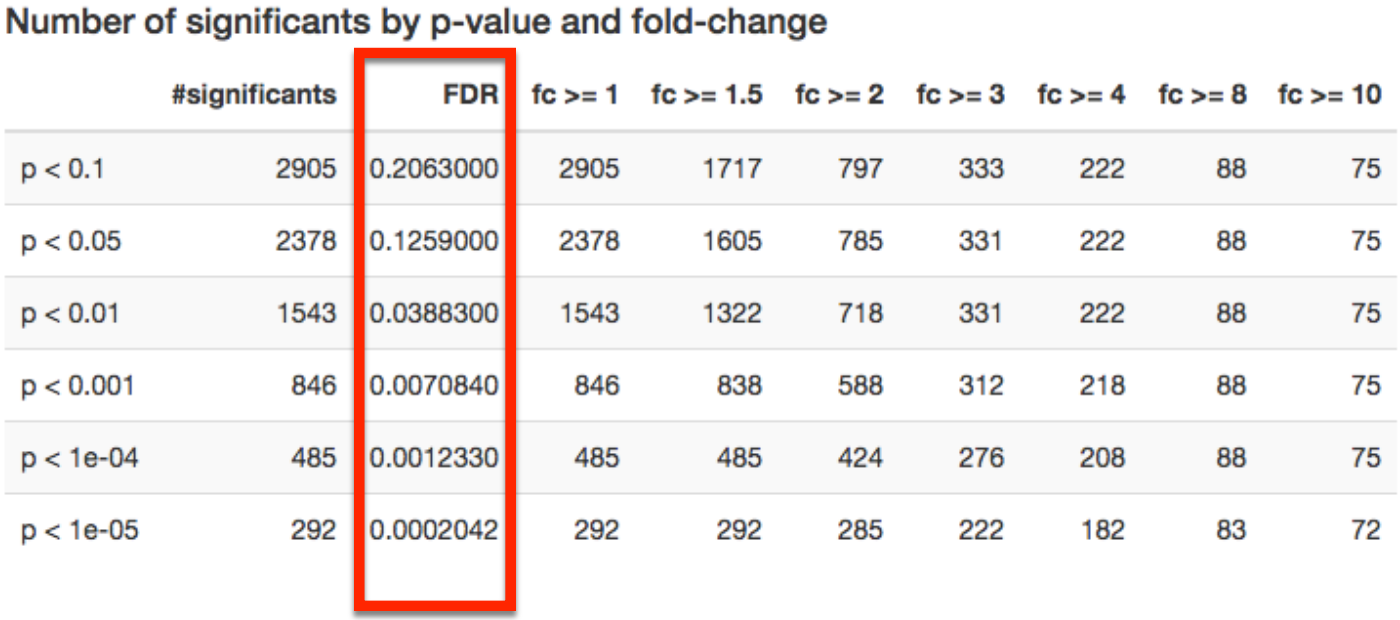
- The variance for the counts for a given gene between samples is often much larger than the mean, making the Poisson assumption restrictive
  - Binomial distribution
- Biological replicates are crucial (no replicates, no statistics, no p-Value)
- The test is done as many times as there are genes
  - Adjust the p-values for multiple hypothesis testing (estimate the False Discovery Rate according to the Benjamini-Hochberg rule)



Multiple testing

- Standard equations for significance and estimates of error rates (false positives and false negatives) apply **only to single comparisons**
- If multiple comparisons are performed (e.g. >1 gene) this must be reflected in the computations for significance and error rates.
- When multiple tests are performed, error rates, statistics and associated p-values **must be adjusted**
  - The strictest correction is Bonferroni (all tests are independent)
  - In the context of differential gene expression, a **F**alse **D**iscovery **R**ate is applied to allow for partial interdependence of the genes

A summary of the results of differentially expressed genes (DEG)

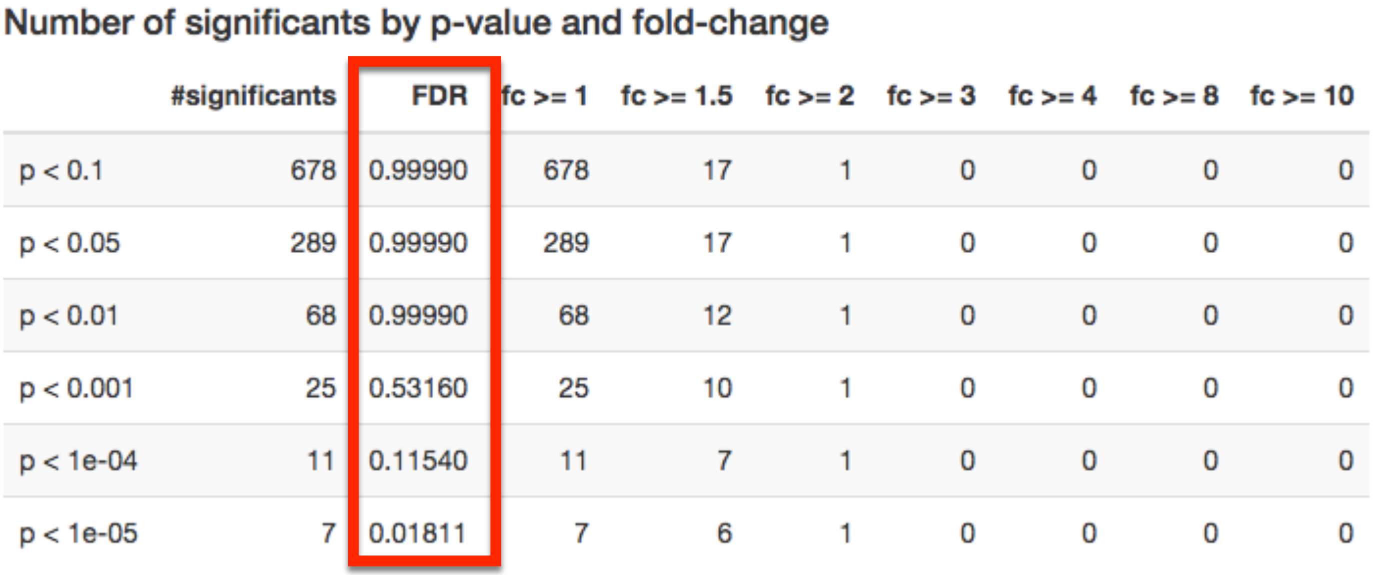


Number of significants by p-value and fold-change

	#significants	FDR	fc >= 1	fc >= 1.5	fc >= 2	fc >= 3	fc >= 4	fc >= 8	fc >= 10
p < 0.1	2905	0.2063000	2905	1717	797	333	222	88	75
p < 0.05	2378	0.1259000	2378	1605	785	331	222	88	75
p < 0.01	1543	0.0388300	1543	1322	718	331	222	88	75
p < 0.001	846	0.0070840	846	838	588	312	218	88	75
p < 1e-04	485	0.0012330	485	485	424	276	208	88	75
p < 1e-05	292	0.0002042	292	292	285	222	182	83	72

- How many features are identified above certain thresholds ?
- What are the trends at fixed p-values and fold-change?
- Does the false discovery rate decreases with p-value?

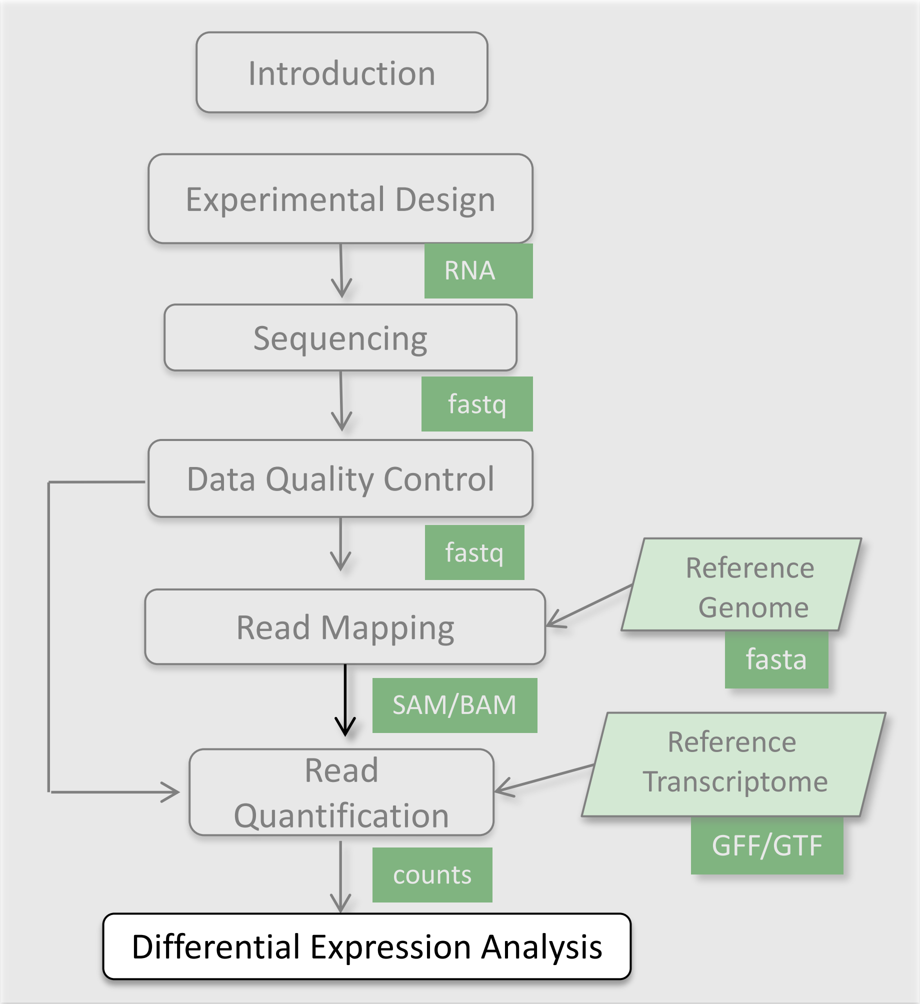
A summary of the results of differentially expressed genes (DEG)



Number of significants by p-value and fold-change

	#significants	FDR	fc >= 1	fc >= 1.5	fc >= 2	fc >= 3	fc >= 4	fc >= 8	fc >= 10
p < 0.1	678	0.99990	678	17	1	0	0	0	0
p < 0.05	289	0.99990	289	17	1	0	0	0	0
p < 0.01	68	0.99990	68	12	1	0	0	0	0
p < 0.001	25	0.53160	25	10	1	0	0	0	0
p < 1e-04	11	0.11540	11	7	1	0	0	0	0
p < 1e-05	7	0.01811	7	6	1	0	0	0	0

- Something like this can indicate
  - very small difference in the groups (down to a handful of genes)
  - batch effects (i.e. , hidden confounders)
  - wrong group assignment (i.e., mislabeling)



Differential expression analysis

- General idea
- Tools
- Exploration of differential expression results

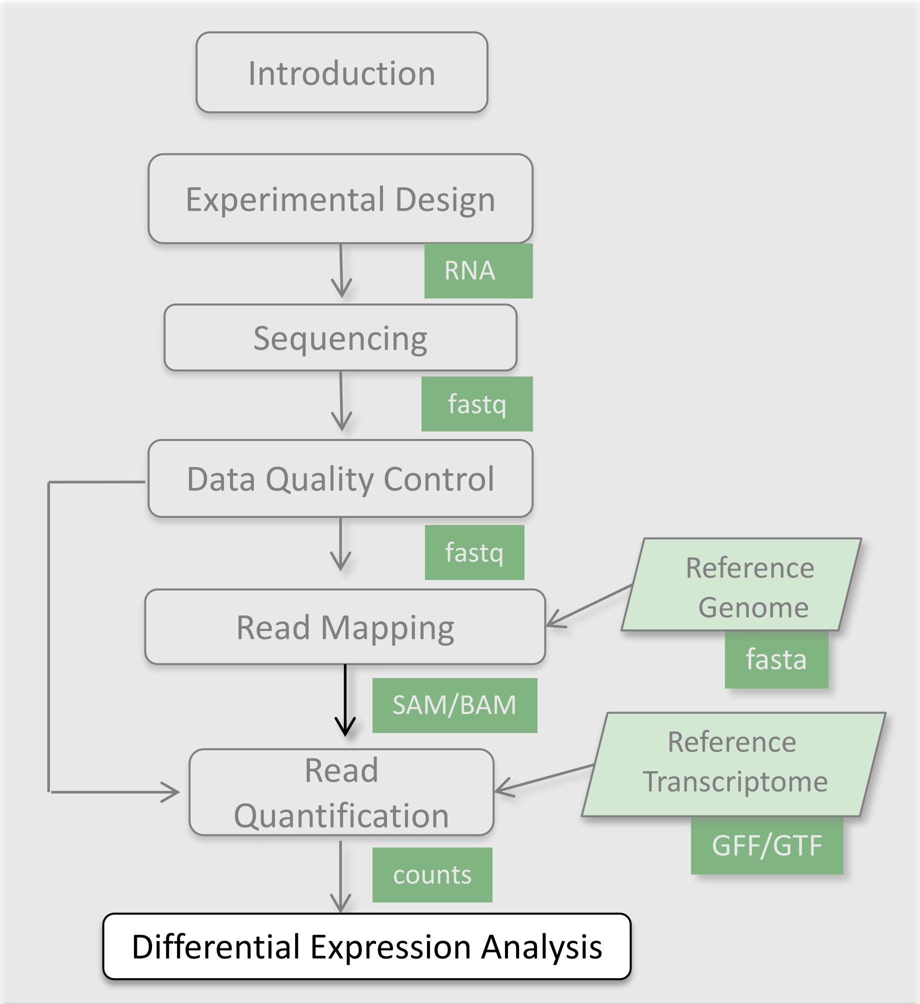
Popular tools

- R/Bioconductor packages: DESeq2 and EdgeR
- Start with un-normalized integer read counts
  - EdgeR: TMM
  - DESeq2: median of ratios
- Similar performance
  - DESeq2 moderates log ratios (genes with very low and highly variable counts)
  - DESeq2 has slightly more conservative p-values
  - EdgeR handles outliers better/more conservative logFC

Importance of biological replicates

- The fundamental problem with generalizing results gathered from unreplicated data is a complete lack of knowledge about biological variation. Without an estimate of variability within the groups, there is no sound statistical basis for inference of differences between the groups.

- When biological replicates are impossible
  - EdgeR
    - Ignore P values
    - Rank genes/isoforms by fold changes



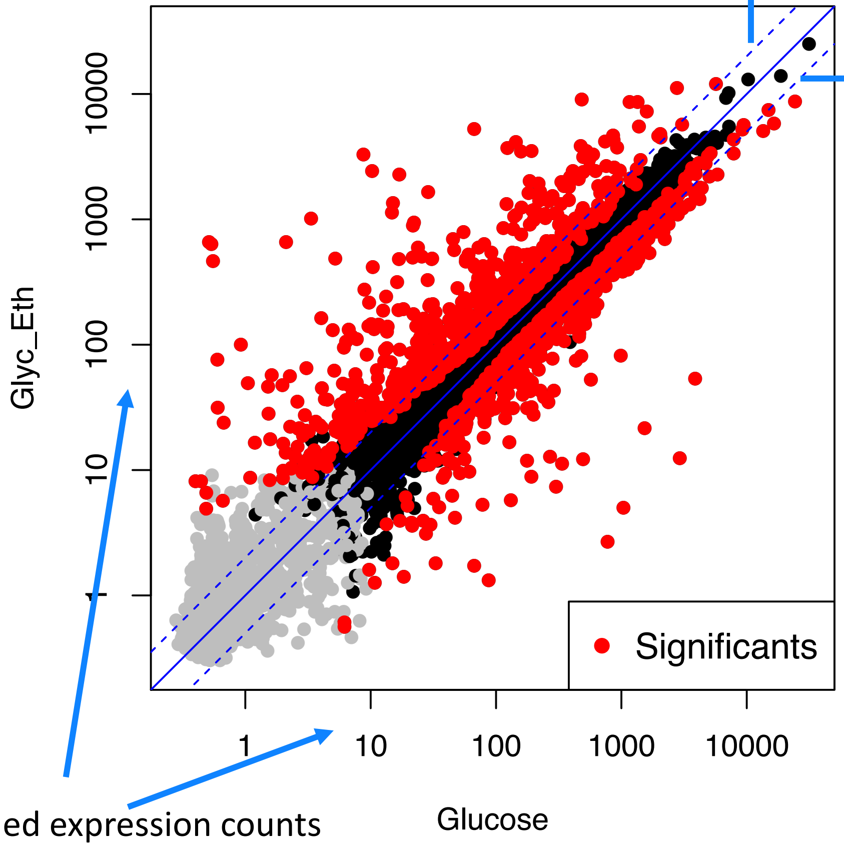
Differential expression analysis

- General idea
- Tools
- Exploration of differential expression results

Comparison of average expression between sample groups

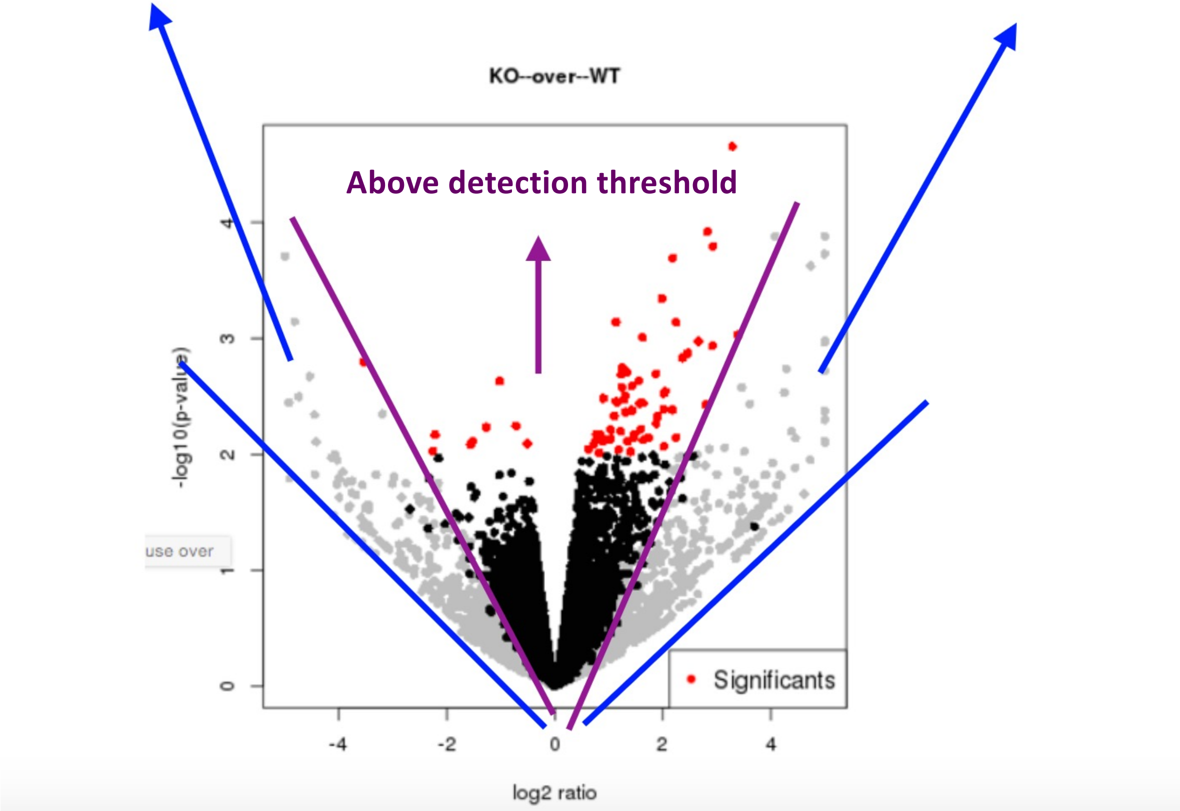
1.5x up in Glyc_Eth

1.5x up in Glucose



Log-scaled expression counts

Distribution of significantly DEGs – volcano plot

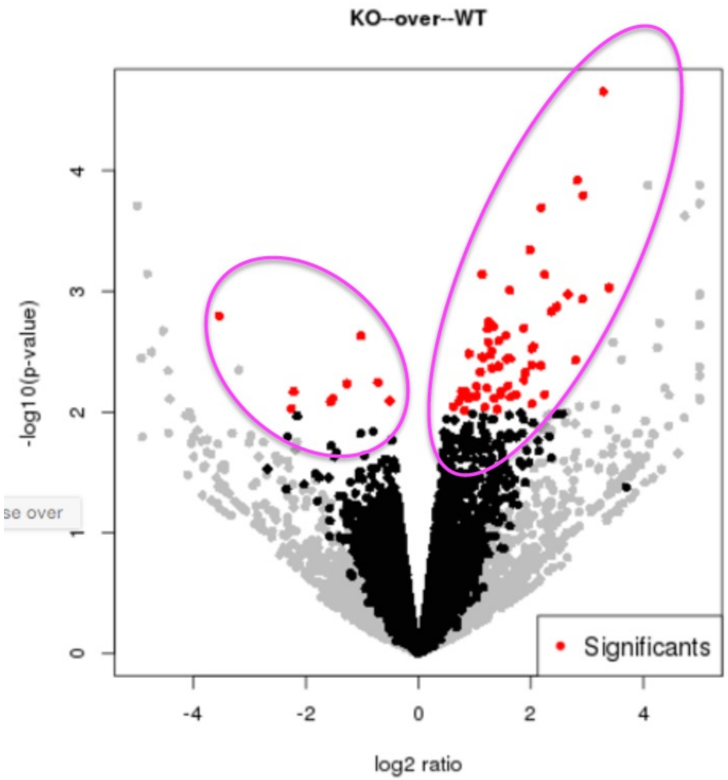


Below expression detection threshold

Above detection threshold

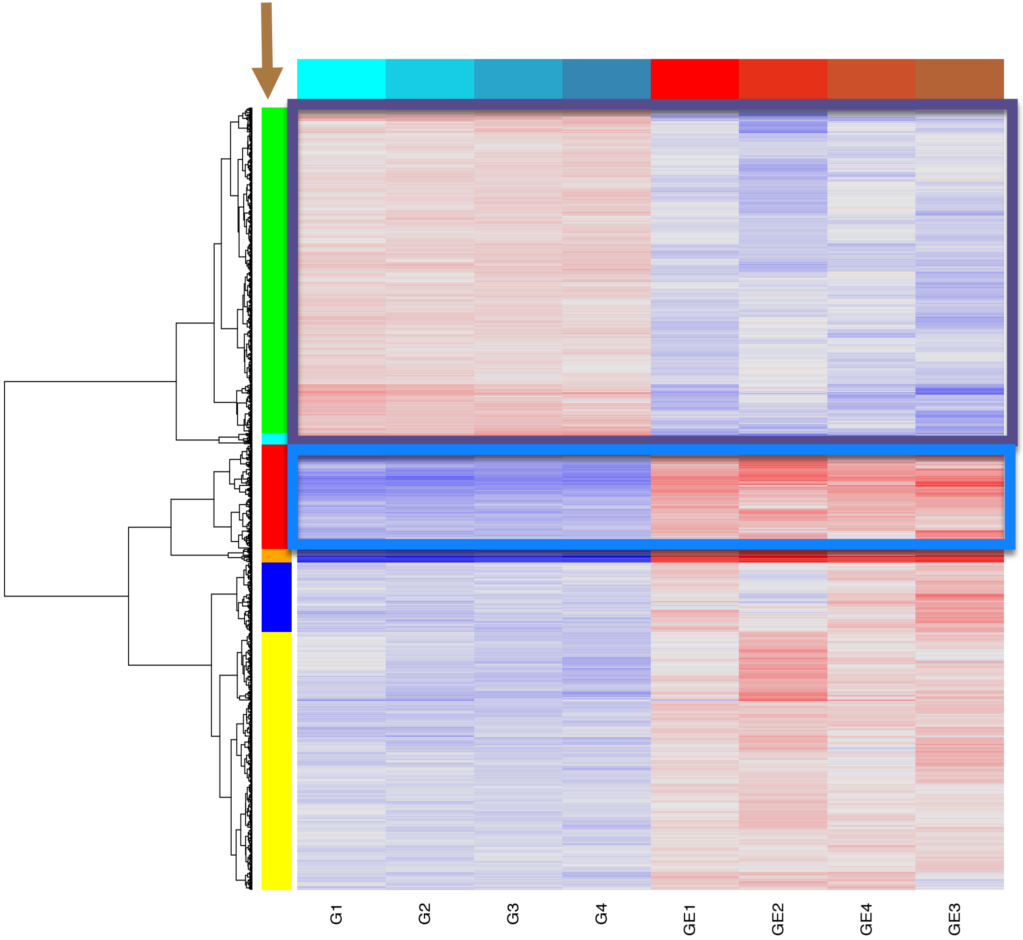
Below expression detection threshold

Distribution of significantly DEGs - volcano plot



- Is the distribution symmetric?
- If not, is there a biological reason for that?

Heatmap of significant DEGs



- How do the genes cluster?
- Is there a dominant cluster?
- Are there small clusters which might represent very strong effects?

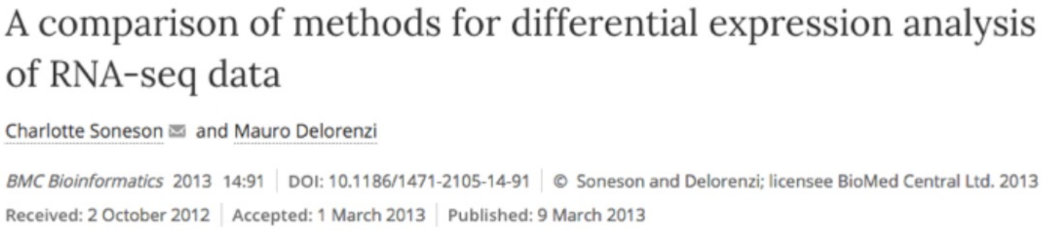
Differential expression analysis: summary

- Biological replicates are crucial
- Variance of count data follows Binomial distribution

- Correction for multiple testing (FDR)
  - Allows partial interdependency of genes
  - Decrease with decreasing p-values

- Look at the summary of p-values and FDR to spot unusual trends
- Compare distribution of mis-regulated genes with expectation

Further reading

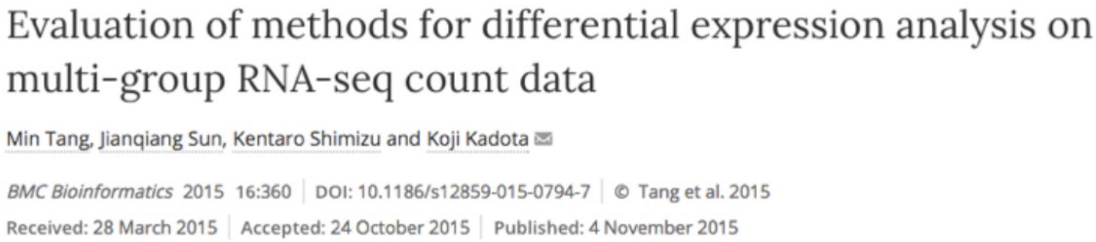


A comparison of methods for differential expression analysis of RNA-seq data

Charlotte Soneson and Mauro Delorenzi

BMC Bioinformatics 2013 14:91 | DOI: 10.1186/1471-2105-14-91 | © Soneson and Delorenzi; licensee BioMed Central Ltd. 2013

Received: 2 October 2012 | Accepted: 1 March 2013 | Published: 9 March 2013

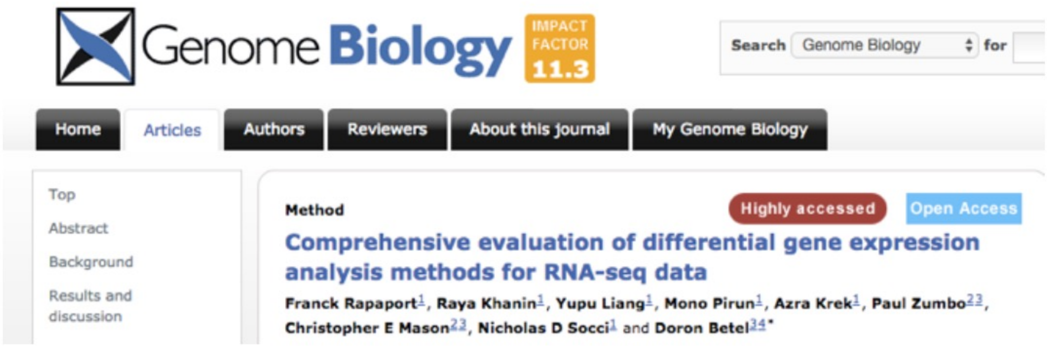


Evaluation of methods for differential expression analysis on multi-group RNA-seq count data

Min Tang, Jianqiang Sun, Kentaro Shimizu and Koji Kadota

BMC Bioinformatics 2015 16:360 | DOI: 10.1186/s12859-015-0794-7 | © Tang et al. 2015

Received: 28 March 2015 | Accepted: 24 October 2015 | Published: 4 November 2015

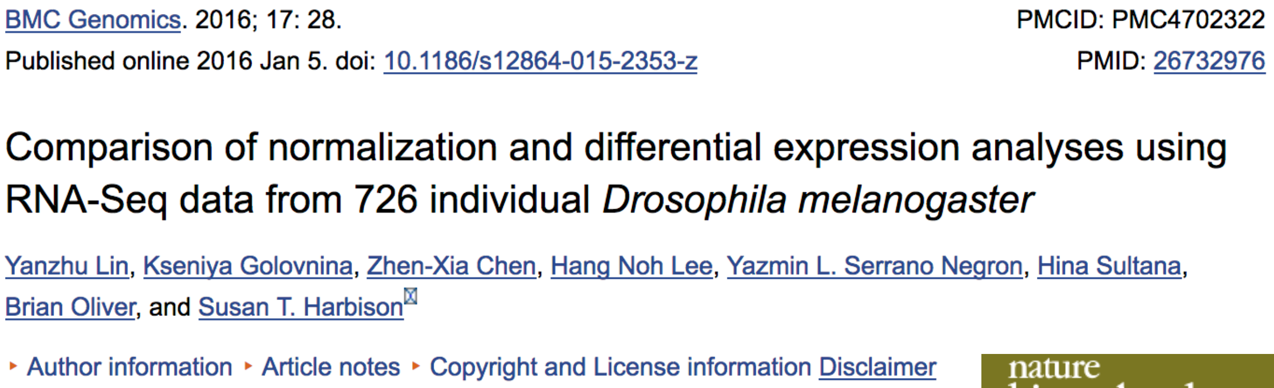


Method

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]*

Further reading



BMC Genomics. 2016; 17: 28.

Published online 2016 Jan 5. doi: 10.1186/s12864-015-2353-z

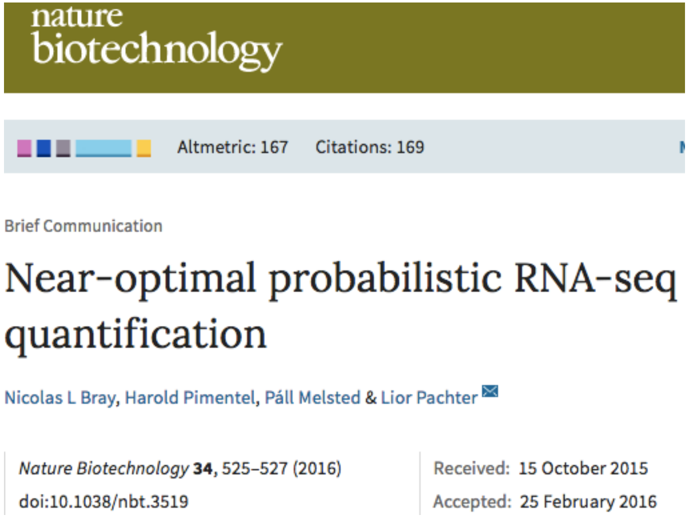
PMCID: PMC4702322

PMID: 26732976

Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster*

Yanzhu Lin, Kseniya Golovnina, Zhen-Xia Chen, Hang Noh Lee, Yazmin L. Serrano Negron, Hina Sultana, Brian Oliver, and Susan T. Harbison

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer



nature biotechnology

Altmetric: 167 Citations: 169

Brief Communication

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray, Harold Pimentel, Páll Melsted & Lior Pachter

Nature Biotechnology **34**, 525–527 (2016)

doi:10.1038/nbt.3519

Received: 15 October 2015

Accepted: 25 February 2016