Analysis of bulk RNA-seq III: DGE and beyond

Analysis of Next-Generation Sequencing Data

Friederike Dündar

Applied Bioinformatics Core

Slides at https://bit.ly/2T3sjRg[1]

March 10, 2020

Weill Cornell Medicine

[1] https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/schedule_2020/

1 Re-cap

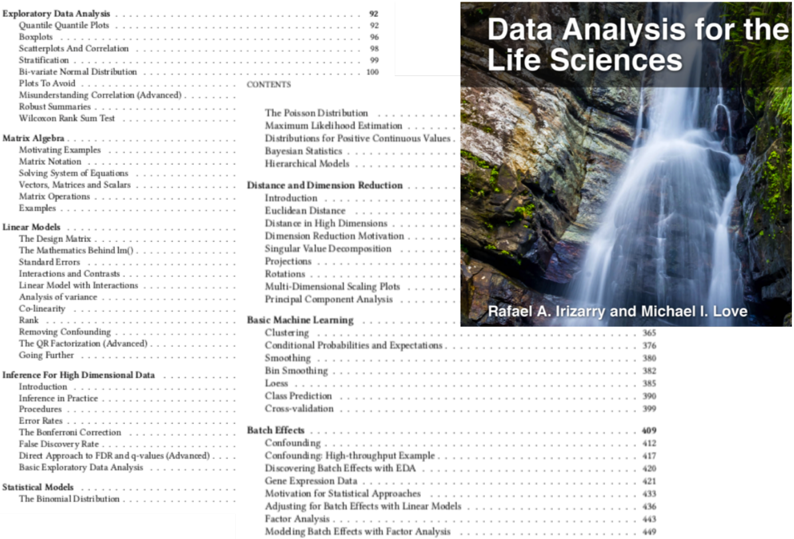
2 Post-p-value calculations

3 Downstream analyses

4 References

Re-cap

https://leanpub.com/dataanalysisforthelifesciences



Exploratory Data Analysis 92
Quantile Quantile Plots 92
Boxplots 96
Scatterplots And Correlation 98
Stratification 99
Bi-variate Normal Distribution 100
Plots To Avoid
Misunderstanding Correlation (Advanced)
Robust Summaries
Wilcoxon Rank Sum Test

Matrix Algebra
Motivating Examples
Matrix Notation
Solving System of Equations
Vectors, Matrices and Scalars
Matrix Operations
Examples

Linear Models
The Design Matrix
The Mathematics Behind lm()
Standard Errors
Interactions and Contrasts
Linear Model with Interactions
Analysis of variance
Co-linearity
Rank
Removing Confounding
The QR Factorization (Advanced)
Going Further

Inference For High Dimensional Data
Introduction
Inference in Practice
Procedures
Error Rates
The Bonferroni Correction
False Discovery Rate
Direct Approach to FDR and q-values (Advanced)
Basic Exploratory Data Analysis

Statistical Models
The Binomial Distribution

CONTENTS

The Poisson Distribution
Maximum Likelihood Estimation
Distributions for Positive Continuous Values
Bayesian Statistics
Hierarchical Models

Distance and Dimension Reduction
Introduction
Euclidean Distance
Distance in High Dimensions
Dimension Reduction Motivation
Singular Value Decomposition
Projections
Rotations
Multi-Dimensional Scaling Plots
Principal Component Analysis

Basic Machine Learning
Clustering 365
Conditional Probabilities and Expectations 376
Smoothing 380
Bin Smoothing 382
Loess 385
Class Prediction 390
Cross-validation 399

Batch Effects 409
Confounding 412
Confounding: High-throughput Example 417
Discovering Batch Effects with EDA 420
Gene Expression Data 421
Motivation for Statistical Approaches 433
Adjusting for Batch Effects with Linear Models 436
Factor Analysis 443
Modeling Batch Effects with Factor Analysis 449

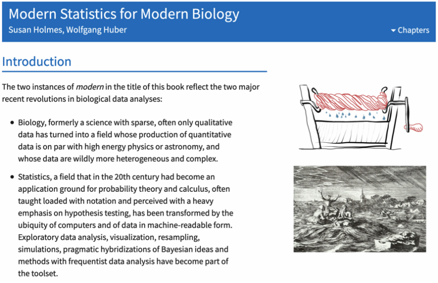
Data Analysis for the Life Sciences

Rafael A. Irizarry and Michael I. Love

I do not get a commission; I honestly believe this book is a great resource and should be on every bioinformatician's desk and/or hard drive.

Additional recommendations from Merv

- https://web.stanford.edu/class/bios221/book/introduction.html
- https://onlinelibrary.wiley.com/doi/book/10.1002/0470114754



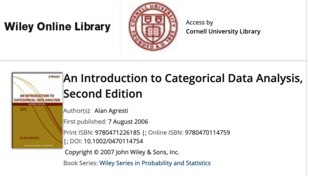
Modern Statistics for Modern Biology

Susan Holmes, Wolfgang Huber

Introduction

The two instances of *modern* in the title of this book reflect the two major recent revolutions in biological data analyses:

- Biology, formerly a science with sparse, often only qualitative data has turned into a field whose production of quantitative data is on par with high energy physics or astronomy, and whose data are wildly more heterogeneous and complex.
- Statistics, a field that in the 20th century had become an application ground for probability theory and calculus, often taught loaded with notation and perceived with a heavy emphasis on hypothesis testing, has been transformed by the ubiquity of computers and of data in machine-readable form. Exploratory data analysis, visualization, resampling, simulations, pragmatic hybridizations of Bayesian ideas and methods with frequentist data analysis have become part of the toolset.



Wiley Online Library

Access by
Cornell University Library

An Introduction to Categorical Data Analysis, Second Edition

Author(s): Alan Agresti

First published: 7 August 2006

Print ISBN: 9780471226185 |; Online ISBN: 9780470114759 |; DOI: 10.1002/0470114754

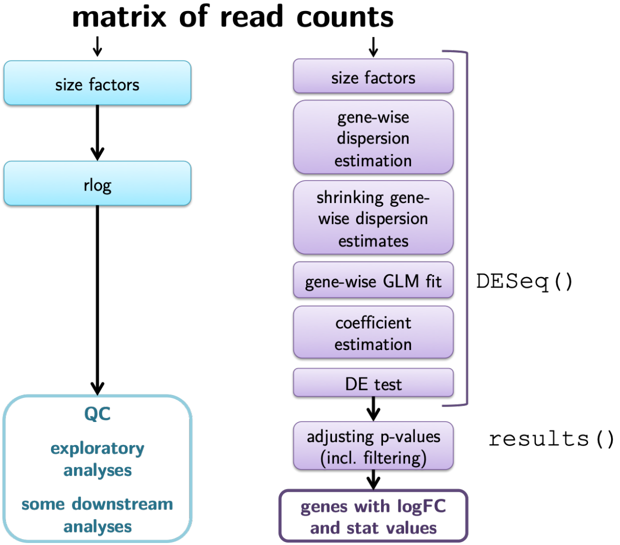
Copyright © 2007 John Wiley & Sons, Inc.

Book Series: Wiley Series in Probability and Statistics

Re-cap: properties of read count data

Property	Relevant for	How it's addressed
Discrete, **non-negative** measurements with greater variability than what could be handled by a Poisson distribution	Estimating robust changes of expression values between different condition	The gene-wise read counts are modeled with a negative binomial distribution; the variances are estimates based on all genes in a given matrix to reduce the noise
Heteroskedasticity (lower read counts often have greater variance than higher read counts)	Obtaining robust transcript abundance measurements that roughly follow a normal distribution, which is often expected for exploratory analyses	Variance shrinkage using the variances from all genes in a given matrix
Large dynamic range		log-transformation
Not an immediate reflection of true transcript abundance	Interpretation and comparison of transcript abundances	Normalization for gene length, sequencing depths, GC content and the overall RNA universe

Summary: from read counts to DGE et al.



Post-p-value calculations

Adjusting for multiple hypothesis testing with independent filtering

- thousands of genes = thousands of tests ⇒ the absolute number of false positives becomes a troublesome burden even at p-values of 1%
- the **adjustment** of the p-values for the abundant hypothesis testing is typically done via the **false discovery rate** as described by Benjamini and Hochberg [2]
    - the more tests we perform, the more strongly the individual p-values will be “punished”
- Love et al. [2014] and others have repeatedly argued that genes with very low read counts can be ignored for downstream analyses and statistical tests are their read counts are often too unreliable and variable to be accurately assessed with only 3-5 replicates

How low is too low?

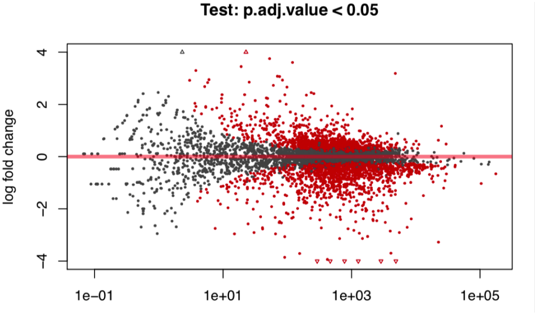
The `results()` function of `DESeq2` will try to find the optimal expression cut-off to maximize the absolute number of genes that pass the adjusted p-value threshold.

[2] see `?p.adjust()`

Shrinking the logFC values

- visualizations and downstream analyses may sometimes benefit from using the **fold changes** instead of the normalized read count values per gene

- Normalized read counts ⇒ transcript abundances per gene per sample
- logFC ⇒ magnitude of the **difference** between multiple samples and conditions



Comparison of additional tools for DGE analysis

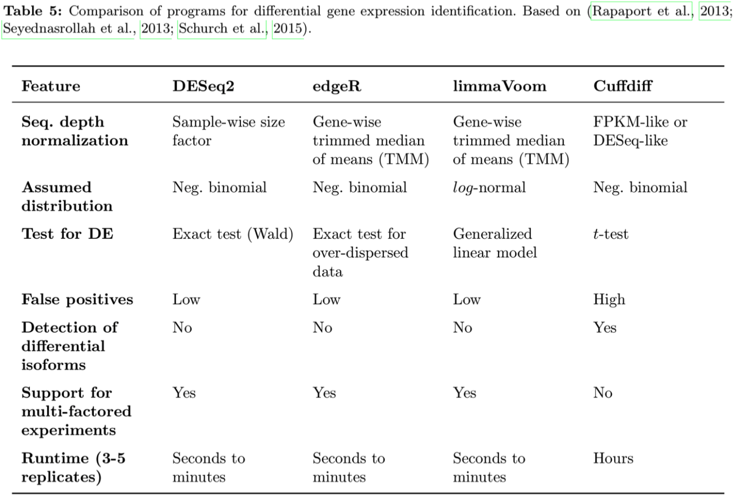


Table 5: Comparison of programs for differential gene expression identification. Based on (Rapaport et al., 2013; Seyednasrollah et al., 2013; Schurch et al., 2015).

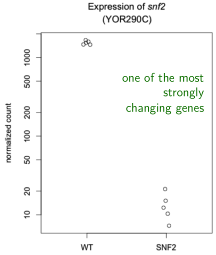
Feature	**DESeq2**	**edgeR**	**limmaVoom**	**Cuffdiff**
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

When in doubt, compare the results of `limma`, `edgeR`, and `DESeq2` to get a feeling for how robust your favorite DE genes are. All packages can be found at Bioconductor.

Downstream analyses

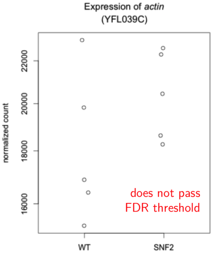
Understanding the RESULTS of the DGE analysis

- Investigate the `results()` output:
  - How many DE genes? (FDR/q-value!)
  - How strongly do the DE genes change?
  - Directions of change?
  - Are your favorite genes among the DE genes?



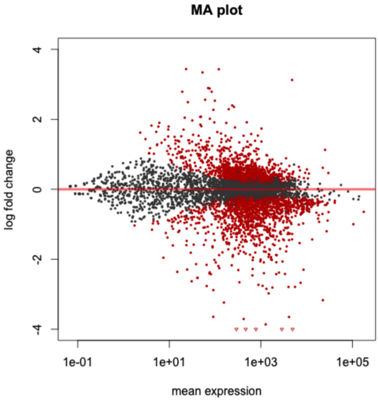
Expression of *snf2*
(YOR290C)

one of the most strongly changing genes



Expression of *actin*
(YFL039C)

does not pass FDR threshold



MA plot

Understanding the FUNCTIONS of your DE genes

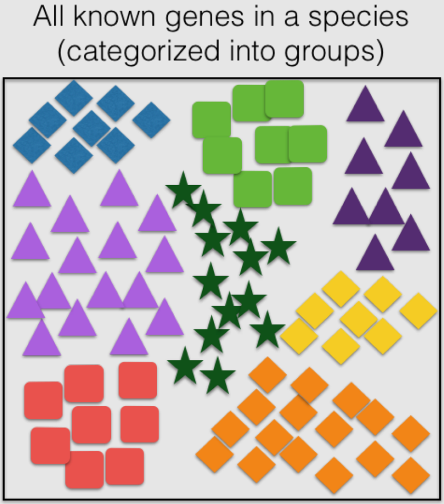
There are myriad tools for this – many are web-based, many are R packages, many will address very specific questions. Typical points of interest are:

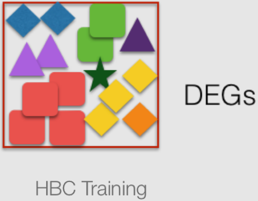
- enriched gene ontology (GO) terms
  - ontology = standardized vocabulary
  - 3 classes of gene ontologies are maintained:
    - biological processes (BP), cell components (CC), and molecular functions (MF)
- enriched pathways
  - gene sets: e.g. from MSigDB [Liberzon et al., 2015]
  - physical interaction networks: e.g. from STRING [Szklarczyk et al., 2017]
  - metabolic (and other) pathways: e.g. from KEGG [Kanehisa et al., 2017]
- upstream regulators

None (!) of these methods should lead you to make definitive claims about the role of certain pathways for your phenotype. These are **hypothesis-generating** tools! Also: make sure you use **shrunken logFC** values [Zhu et al., 2019].

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)





Category	Background	DE list	Over-represented?
A	35/6600	25/500	likely
B	56/6600	2/500	unlikely
C	10/6600	9/500	likely

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)

- "2x2 table method"
- assessing overlap of DE genes with genes of a given pathway
- statistical test: e.g. hypergeometric test
- limitations:
  - direction of change is ignored
  - magnitude of change is ignored
  - interprets genes as well as pathways as independent entities

See Khatri et al. [2012] for details!

Two typical approaches of enrichment analyses

1. Over-representation analysis (ORA)

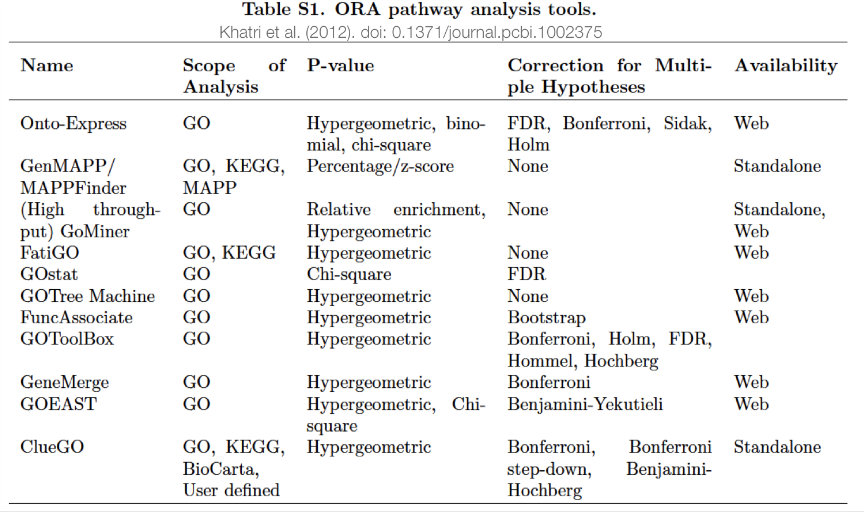


Table S1. ORA pathway analysis tools.

Khatri et al. (2012). doi: 0.1371/journal.pcbi.1002375

Name	Scope of Analysis	P-value	Correction for Multiple Hypotheses	Availability
Onto-Express	GO	Hypergeometric, binomial, chi-square	FDR, Bonferroni, Sidak, Holm	Web
GenMAPP/ MAPPFinder	GO, KEGG, MAPP	Percentage/z-score	None	Standalone
(High throughput) GoMiner	GO	Relative enrichment, Hypergeometric	None	Standalone, Web
FatiGO	GO, KEGG	Hypergeometric	None	Web
GOstat	GO	Chi-square	FDR	
GOTree Machine	GO	Hypergeometric	None	Web
FuncAssociate	GO	Hypergeometric	Bootstrap	Web
GOToolBox	GO	Hypergeometric	Bonferroni, Holm, FDR, Hommel, Hochberg	
GeneMerge	GO	Hypergeometric	Bonferroni	Web
GOEAST	GO	Hypergeometric, Chi-square	Benjamini-Yekutieli	Web
ClueGO	GO, KEGG, BioCarta, User defined	Hypergeometric	Bonferroni, Bonferroni step-down, Benjamini-Hochberg	Standalone

Two typical approaches of enrichment analyses

2. Functional Class Scoring ("Gene set enrichment")

- gene-level statistics for all genes in a pathway are aggregated into a single pathway-level statistic
- score will depend on size of the pathway, and the amount of correlation between genes in the pathway
- all genes are used
- direction and magnitude of change matter
- coordinated changes of genes within the same pathway matter, too

Two typical approaches of enrichment analyses

2. Functional Class Scoring ("Gene set enrichment")

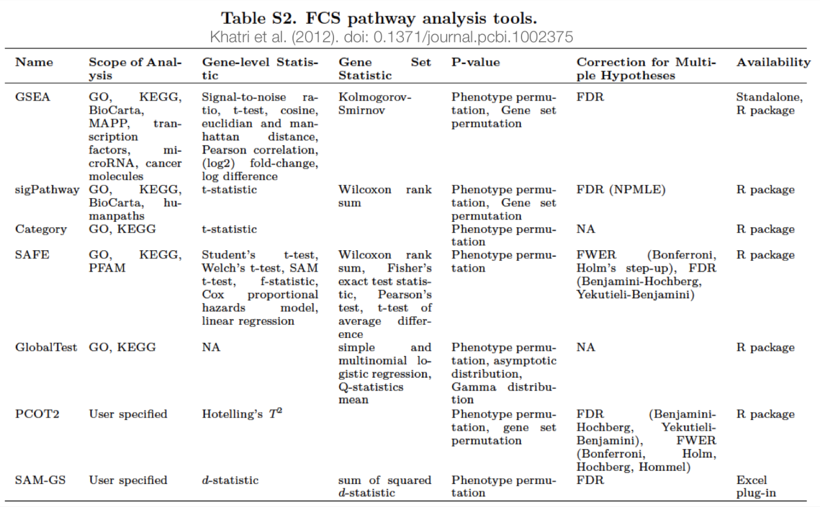


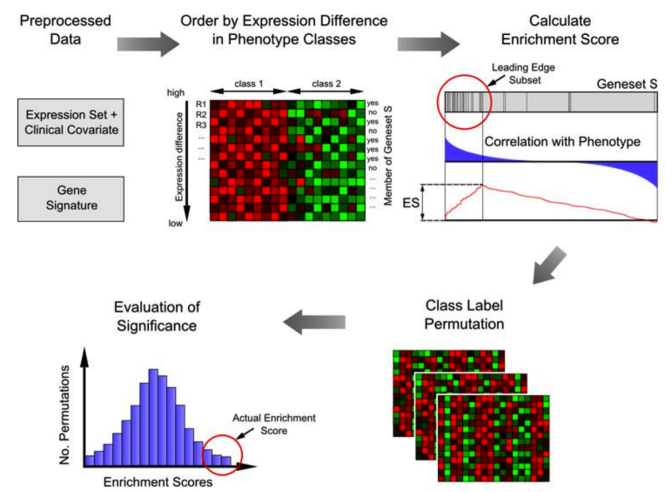
Table S2. FCS pathway analysis tools.

Khatri et al. (2012). doi: 0.1371/journal.pcbi.1002375

Name	Scope of Analysis	Gene-level Statistic	Gene Set Statistic	P-value	Correction for Multiple Hypotheses	Availability
GSEA	GO, KEGG, BioCarta, MAPP, transcription factors, microRNA, cancer molecules	Signal-to-noise ratio, t-test, cosine, euclidian and manhattan distance, Pearson correlation, (log2) fold-change, log difference	Kolmogorov-Smirnov	Phenotype permutation, Gene set permutation	FDR	Standalone, R package
sigPathway	GO, KEGG, BioCarta, humanpaths	t-statistic	Wilcoxon rank sum	Phenotype permutation, Gene set permutation	FDR (NPMLE)	R package
Category	GO, KEGG	t-statistic		Phenotype permutation	NA	R package
SAFE	GO, KEGG, PFAM	Student's t-test, Welch's t-test, SAM t-test, f-statistic, Cox proportional hazards model, linear regression	Wilcoxon rank sum, Fisher's exact test statistic, Pearson's test, t-test of average difference	Phenotype permutation	FWER (Bonferroni, Holm's step-up), FDR (Benjamini-Hochberg, Yekutieli-Benjamini)	R package
GlobalTest	GO, KEGG	NA	simple and multinomial logistic regression, Q-statistics mean	Phenotype permutation, asymptotic distribution, Gamma distribution	NA	R package
PCOT2	User specified	Hotelling's T^2		Phenotype permutation, gene set permutation	FDR (Benjamini-Hochberg, Yekutieli-Benjamini), FWER (Bonferroni, Holm, Hochberg, Hommel)	R package
SAM-GS	User specified	*d*-statistic	sum of squared *d*-statistic	Phenotype permutation	FDR	Excel plug-in

Two typical approaches of enrichment analyses

2. Functional Class Scoring: Example GSEA

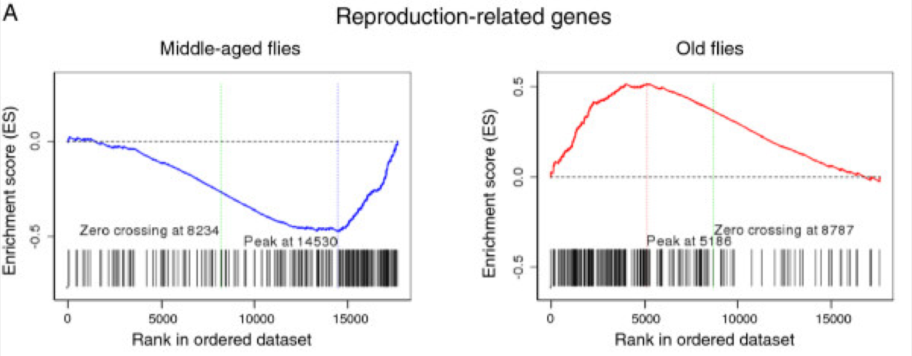


http://slideplayer.biz.tr/slide/2738467/10/images/20/Gene+Set+Enrichment+Analysis+(GSEA).jpg

Two typical approaches of enrichment analyses

2. Functional Class Scoring ("Gene set enrichment")

Example GSEA results for positive and negative correlation



Doroszuk et al. (2012) doi: 10.1186/1471-2164-13-167

Summary – downstream analyses

Know your biological question(s) of interest!

- all enrichment methods potentially suffer from **gene length bias**
  - long genes will get more reads
- for **GO terms**:
  - use goseq to identify enriched GO terms [Young et al., 2010]
  - use additional tools, such as GOrilla, REVIGO [Eden et al., 2009, Supek et al., 2011] to summarize the often redundant GO term lists
- for **KEGG pathways**:
  - e.g. GAGE and PATHVIEW [Luo and Brouwer, 2013, Luo et al., 2017] [3]
- miscellaneous including attempts to predict upstream regulators
  - Enrichr [Chen et al., 2013]
  - RegulatorTrail [Kehl et al., 2017]
  - Ingenuity Pathway Analysis Studio (proprietory software!)

See the additional links and material on our course website!

[3] https://www.r-bloggers.com/tutorial-rna-seq-differential-expression-pathway-analysis-with-sailfish-deseq2-gage-and-pathview/

References

Edward Y. Chen, Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela V. Meirelles, Neil R. Clark, and Avi Ma'ayan. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 2013. doi: 10.1186/1471-2105-14-128. URL http://amp.pharm.mssm.edu/Enrichr.

Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, jan 2009. doi: 10.1186/1471-2105-10-48. URL http://cbl-gorilla.cs.technion.ac.il.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkw1092.

Tim Kehl, Lara Schneider, Florian Schmidt, Daniel Stöckel, Nico Gerstner, Christina Backes, Eckart Meese, Andreas Keller, Marcel H. Schulz, and Hans Peter Lenhof. RegulatorTrail: A web service for the identification of key transcriptional regulators. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkx350. URL https://regulatortrail.bioinf.uni-sb.de/.

Purvesh Khatri, Marina Sirota, and Atul J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 2012. doi: 10.1371/journal.pcbi.1002375.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 2015. doi: 10.1016/j.cels.2015.12.004.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, December 2014. doi: 10.1186/s13059-014-0550-8.

Weijun Luo and Cory Brouwer. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 2013. doi: 10.1093/bioinformatics/btt285.

Weijun Luo, Gaurav Pant, Yeshvant K. Bhavnasi, Steven G. Blanchard, and Cory Brouwer. Pathview Web: User friendly pathway visualization and data integration. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkx372. URL https://pathview.uncc.edu/.

Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6 (7):e21800, jan 2011. doi: 10.1371/journal.pone.0021800. URL http://revigo.irb.hr/.

Damian Szklarczyk, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T. Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian Von Mering. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkw937.

Matthew D. Young, Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 2010. doi: 10.1186/gb-2010-11-2-r14.

Anqi Zhu, Joseph G. Ibrahim, and Michael I. Love. Heavy-Tailed prior distributions for sequence count data: Removing the noise and preserving large differences. *Bioinformatics*, 35(12):2084–2092, 2019. doi: 10.1093/bioinformatics/bty895.