

RNA-seq with R- bioconductor

Part 2: Dealing with Unwanted Variation

MAARTEN LEERKES PHD

1

Refresher: Three Aspects

What is R?

What is Bioconductor?

What is RNAseq?

2

Sequencing by Synthesis

[Intro to Sequencing by Synthesis: Industry-leading Data Quality](#)

<https://www.youtube.com/watch?v=HMyCqWhwB8E>

3

Numerous Possible Analysis Strategies

There is no one ‘correct’ way to analyze RNA-seq data

Two major branches

- Direct alignment of reads (spliced or unspliced) to genome or transcriptome
- Assembly of reads followed by alignment*

*Assembly is the only option when working with a creature with no genome sequence, alignment of contigs may be to ESTs, cDNAs etc

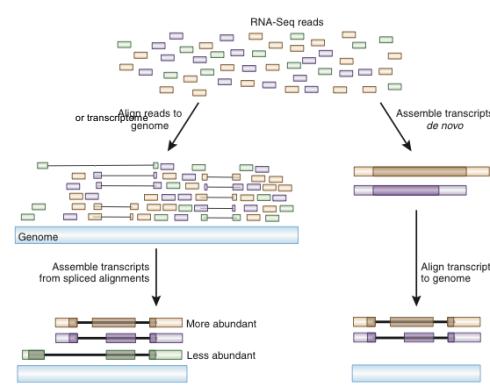
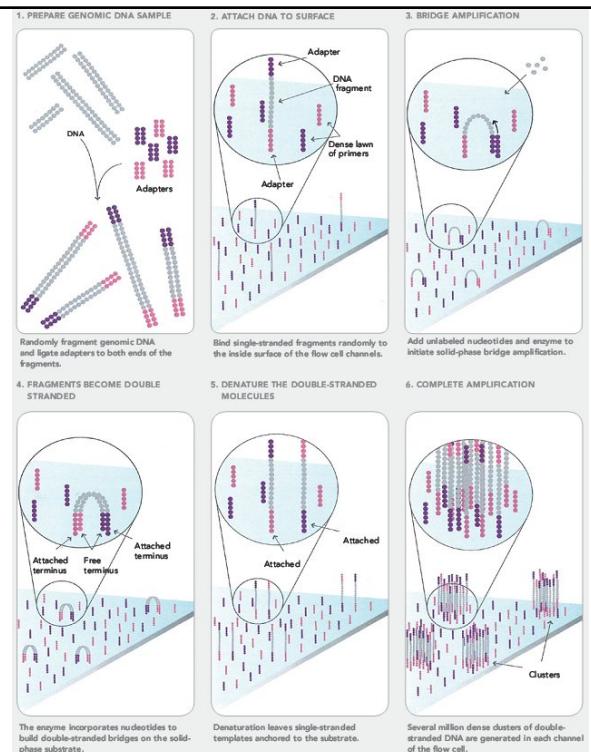
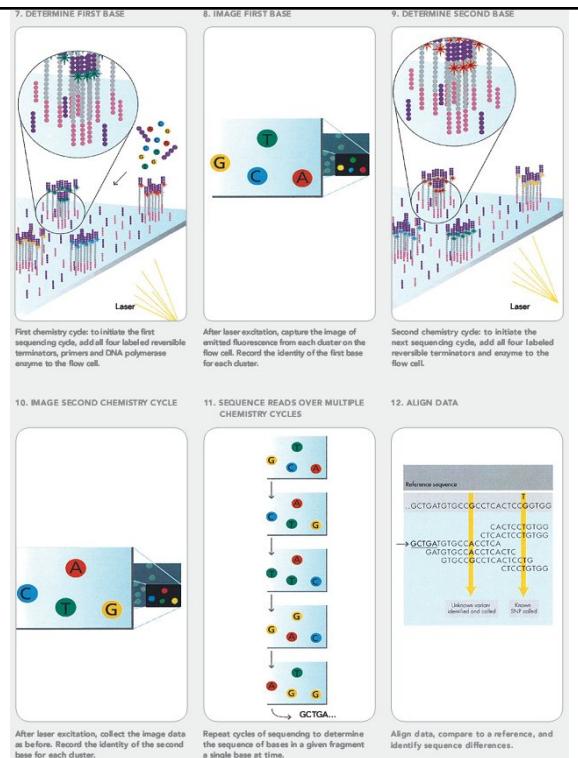


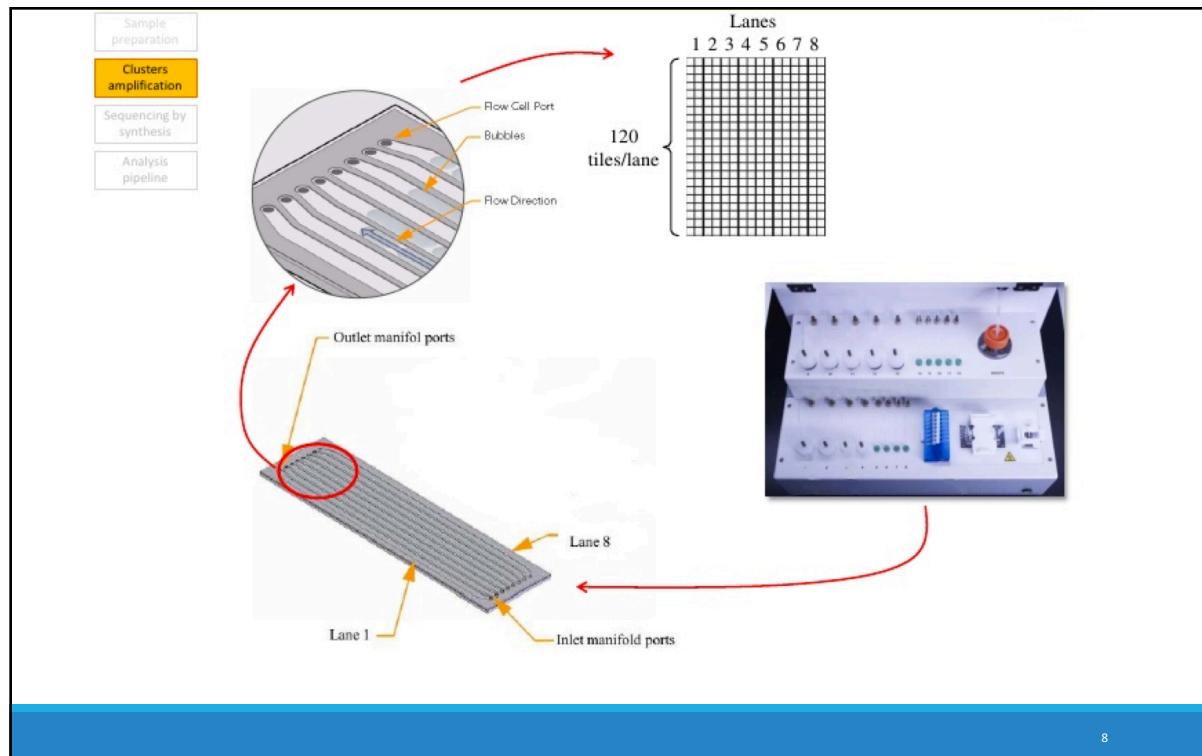
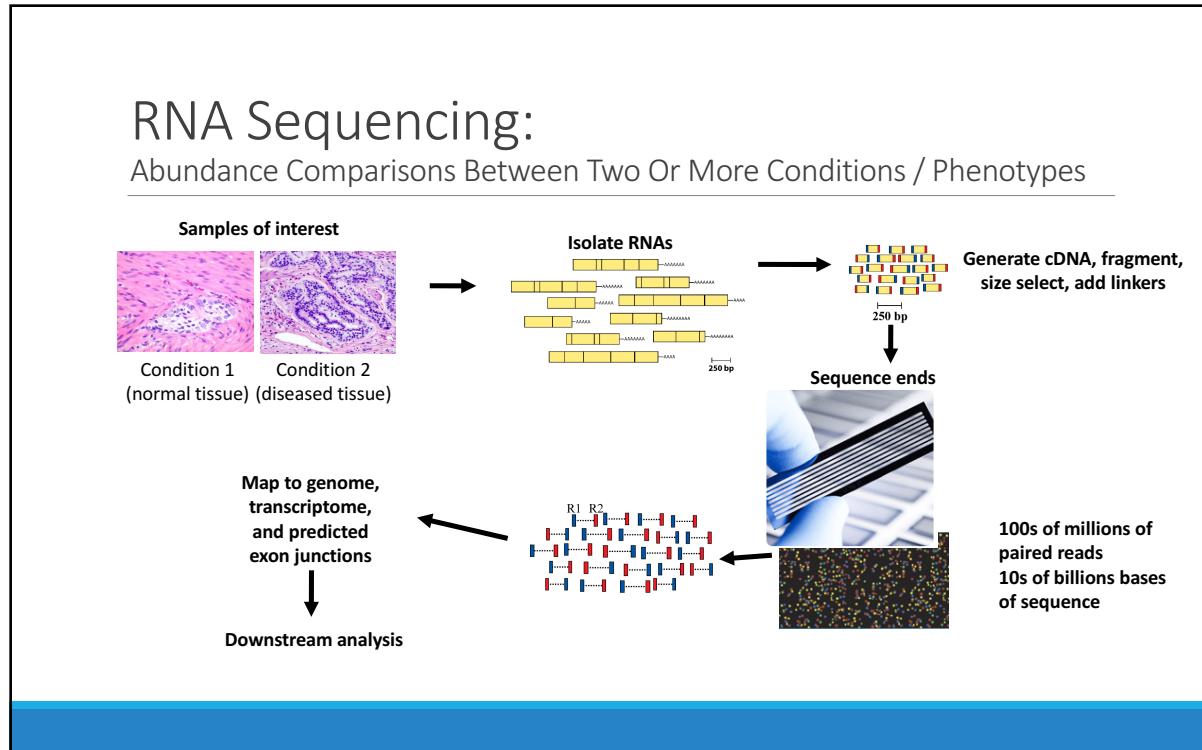
Image from Haas & Zody, 2010

The Basis Of Illumina Sequencing On A Flow Cell



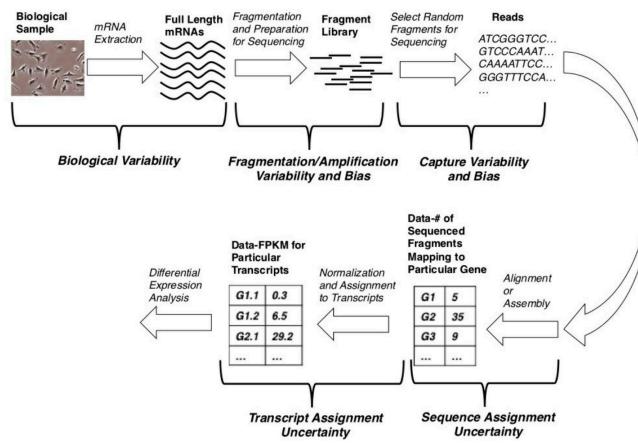
Illumina Clonal Expansion Followed By Image Processing





Potential Sources Of Bias And Variability

Figure 1. The RNA-seq pipeline. This schematic illustrates the process of going from cells to RNA-seq data, with potential sources of bias and variability noted along the way. See the Introduction for details.



9

What is Normalization

The output of the Tophat-Cufflinks pipeline for every condition are estimates of the mean and variance of At for each transcript t (\bar{A}_t and V^*A_t), which specify an associated normal distribution.

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4263583/pdf/biosensors-03-00238.pdf>

Biosensors 2013, 3, 238–258; doi:10.3390/bios3030238

OPEN ACCESS

biosensors
ISSN 2079-6374
www.mdpi.com/journal/biosensors

Article

A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach

Sergiusz Wesołowski ¹, Marc R. Birtwistle ² and Grzegorz A. Rempała ^{3,*}

¹ Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA; E-Mail: wesew@math.fsu.edu

² Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY 10029, USA; E-Mail: marc.birtwistle@mssm.edu

³ Division of Biostatistics, The Ohio State University, Columbus, OH 43210, USA

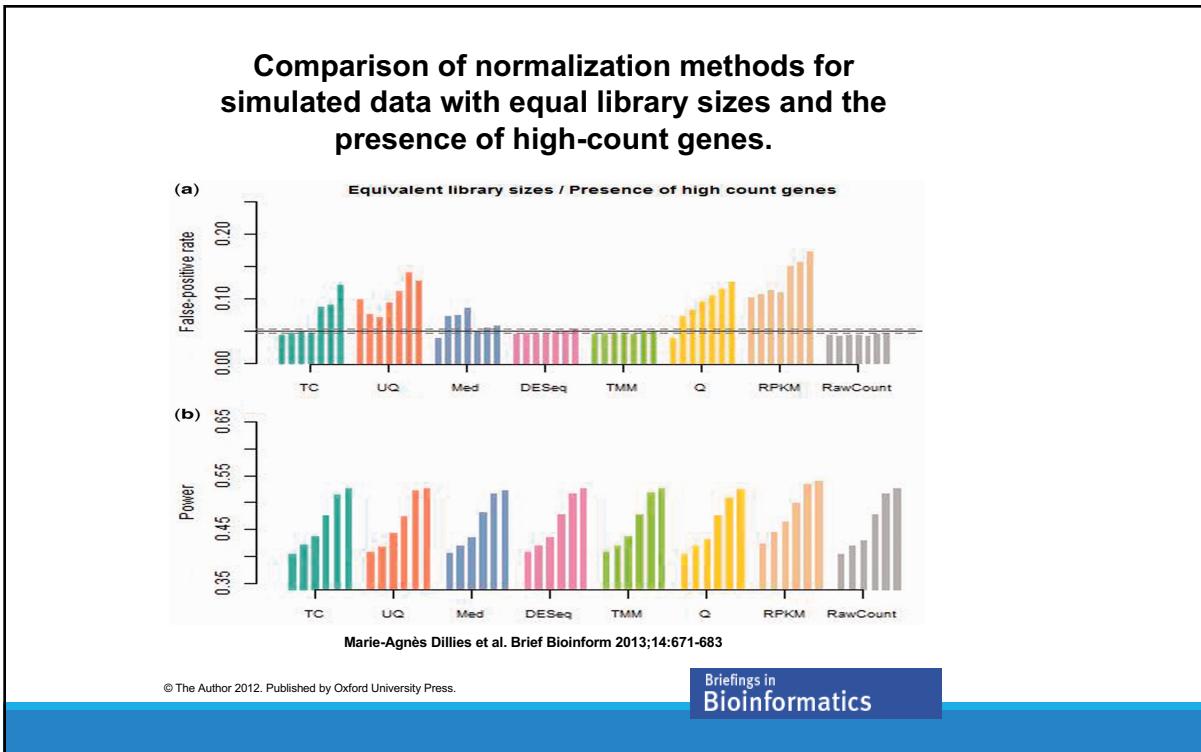
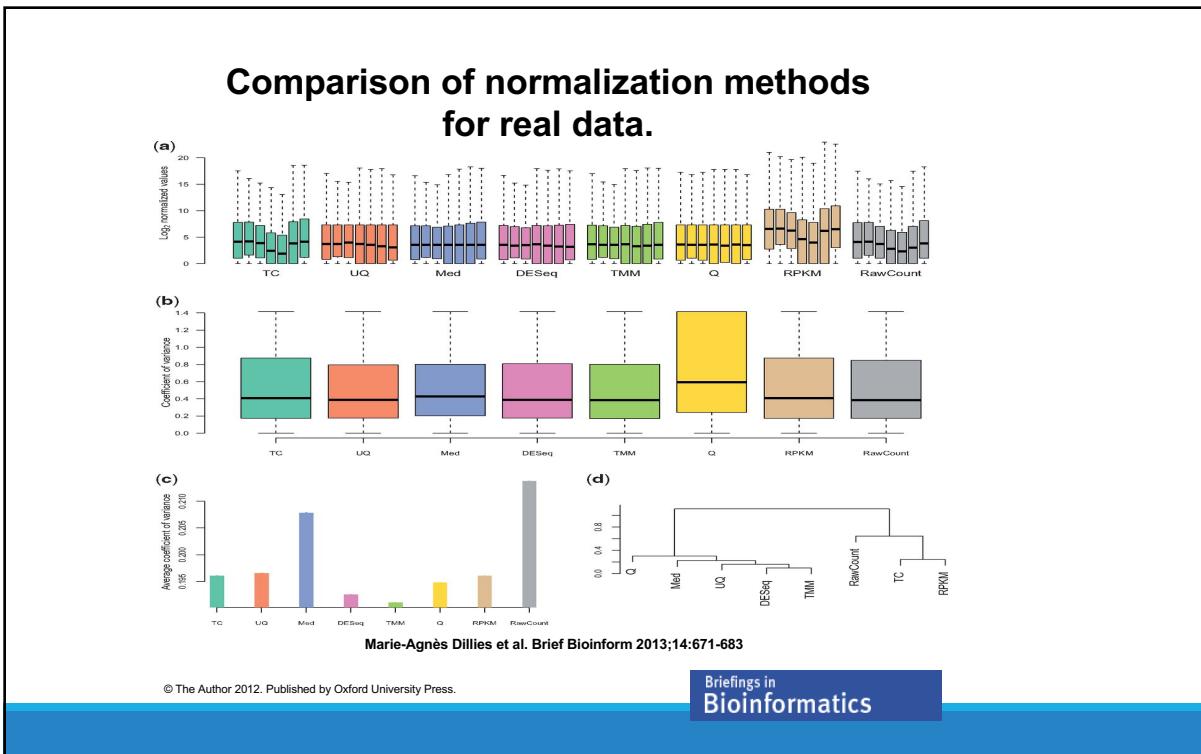
* Author to whom correspondence should be addressed; E-Mail: grempala@cph.ohio.edu.

Received: 12 May 2013; in revised form: 9 June 2013 / Accepted: 12 June 2013 /

Published: 28 June 2013

Abstract: Transcriptome-based biosensors are expected to have a large impact on the future of biotechnology. However, a central aspect of transcriptomics is differential expression analysis, where, currently, deep RNA sequencing (RNA-seq) has the potential to replace the microarray as the standard assay for RNA quantification. Our contributions here to RNA-seq differential expression analysis are two-fold. First, given the high cost of an RNA-seq run, biological replicates are rare, and therefore, information sharing across genes to obtain variance estimates is crucial. To handle such information sharing in a rigorous manner, we propose an hierarchical, empirical Bayes approach (R-EBSeq) that combines the Cufflinks model for generating relative transcript abundance measurements, known as FPKM

10



Normalization

One often overlooked aspect is **normalization**, which is the **transformation of values that allows comparisons between samples** in a way that eliminates the effects of sources of variability that are not of interest.

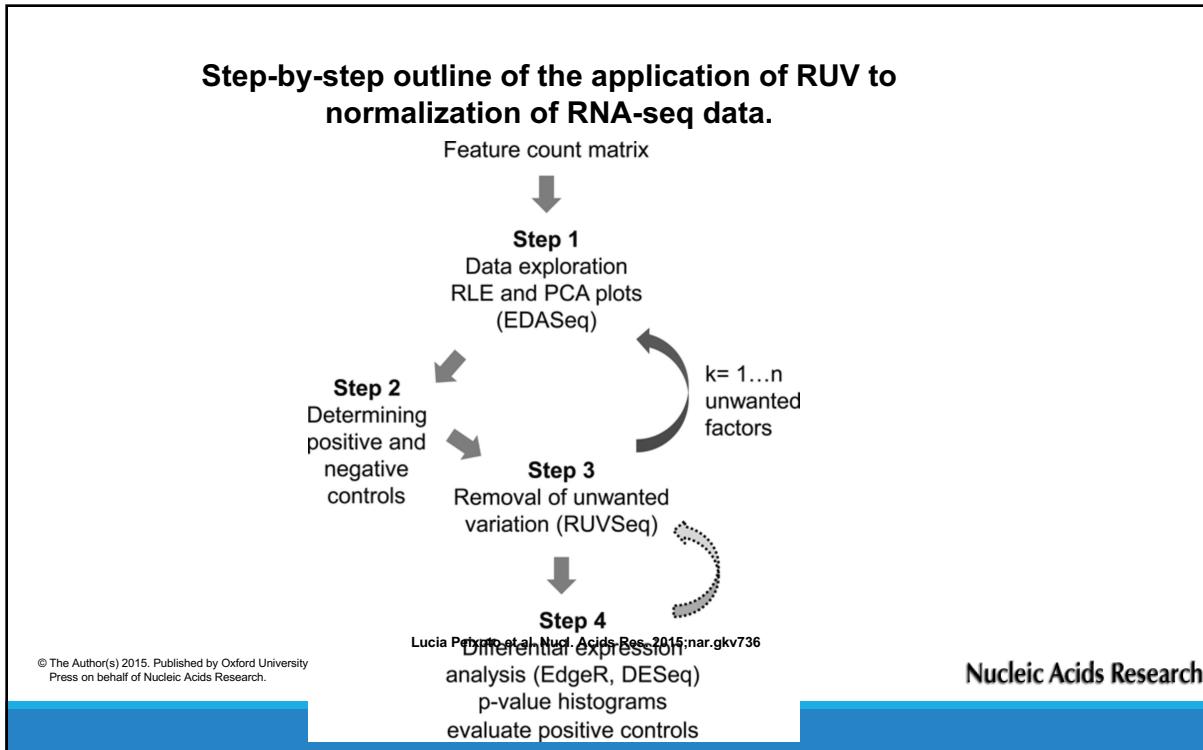
We refer to those effects as ‘unwanted variation’. A variety of technical and biological factors, collectively known as ‘batch effects’, contribute unwanted variation to genome-wide gene expression data. These factors include differences in amount of RNA, library preparation, equipment, operators, and procedures for sample extraction, preservation, or storage. Proper normalization, or removal of these factors, has been shown to critically impact the analysis of high-throughput data (1–3). In spite of this, commonly used methods for RNA-seq normalization, such as upper quartile scaling (UQ)(2), trimmed mean of M values (TMM)(4) and FPKM (5), account only for global differences in sequencing depth between libraries (6).

13

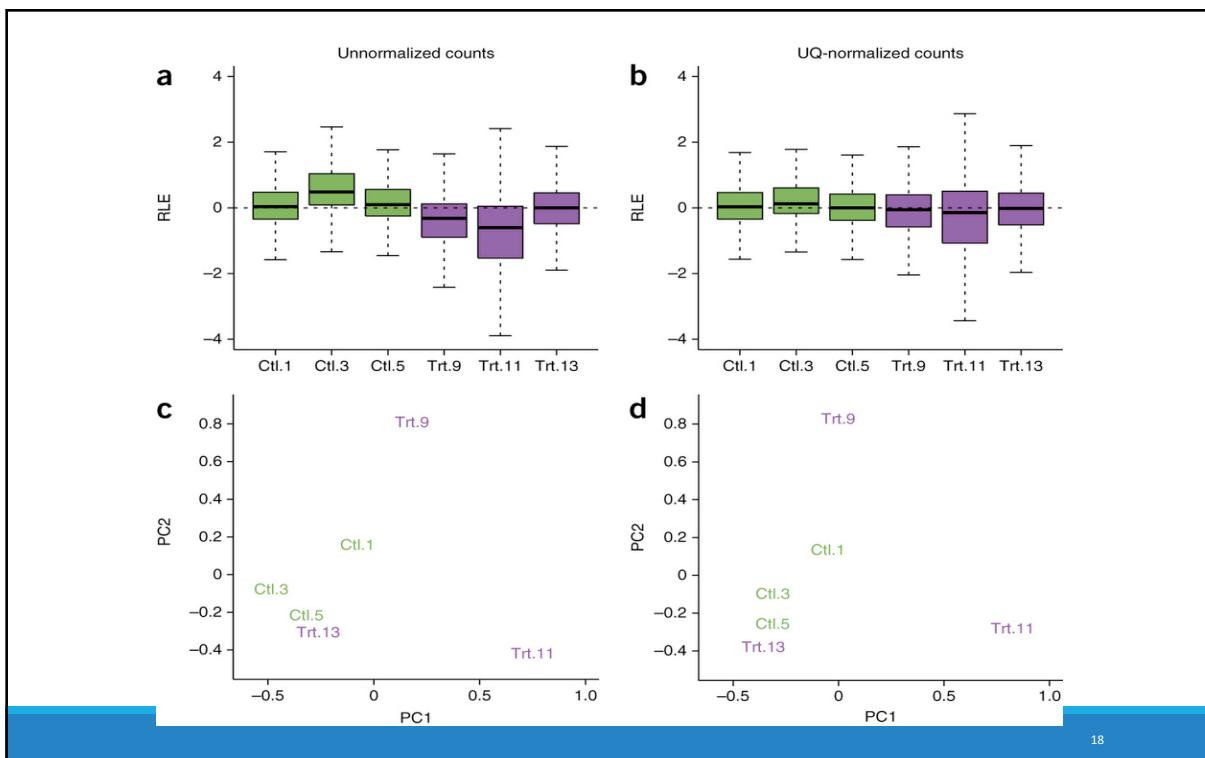
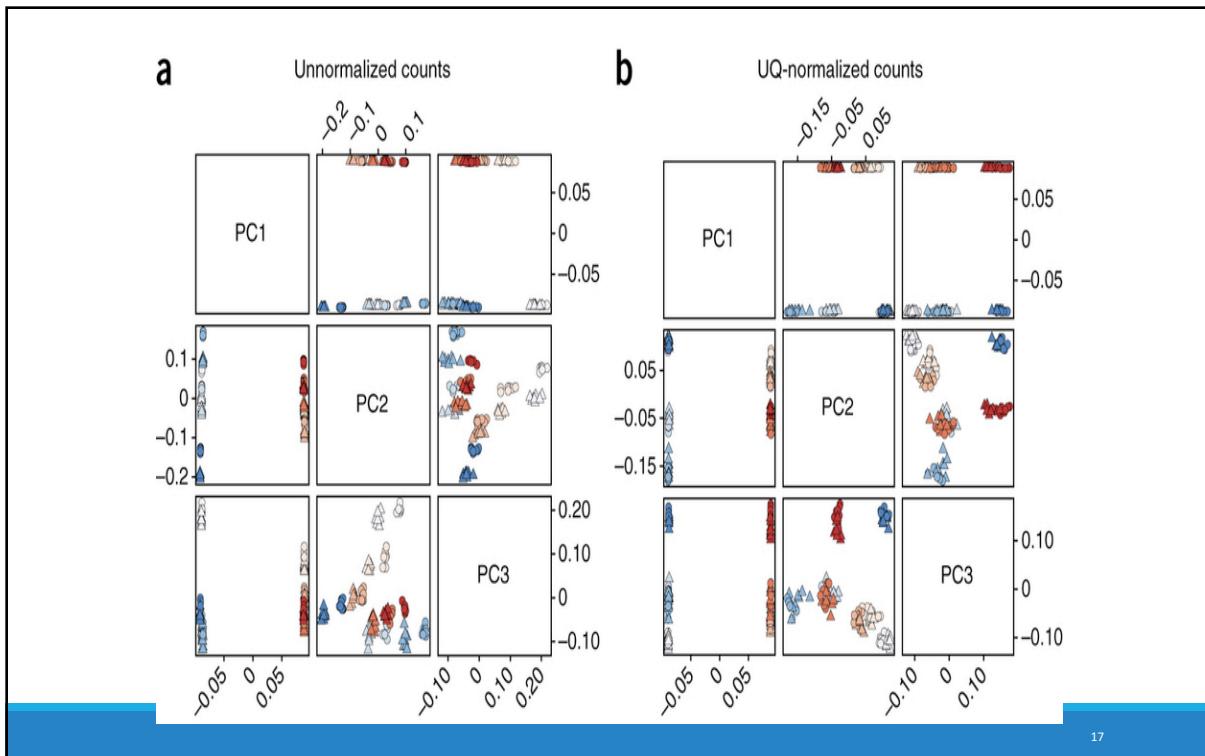
Quantitative and qualitative effects of the choice of normalization method in combined analysis of gene expression changes following FC and OLM. (A) Number of genes and enriched KEGG pathways for OLM and FC relative to combined controls following UQ normalization.

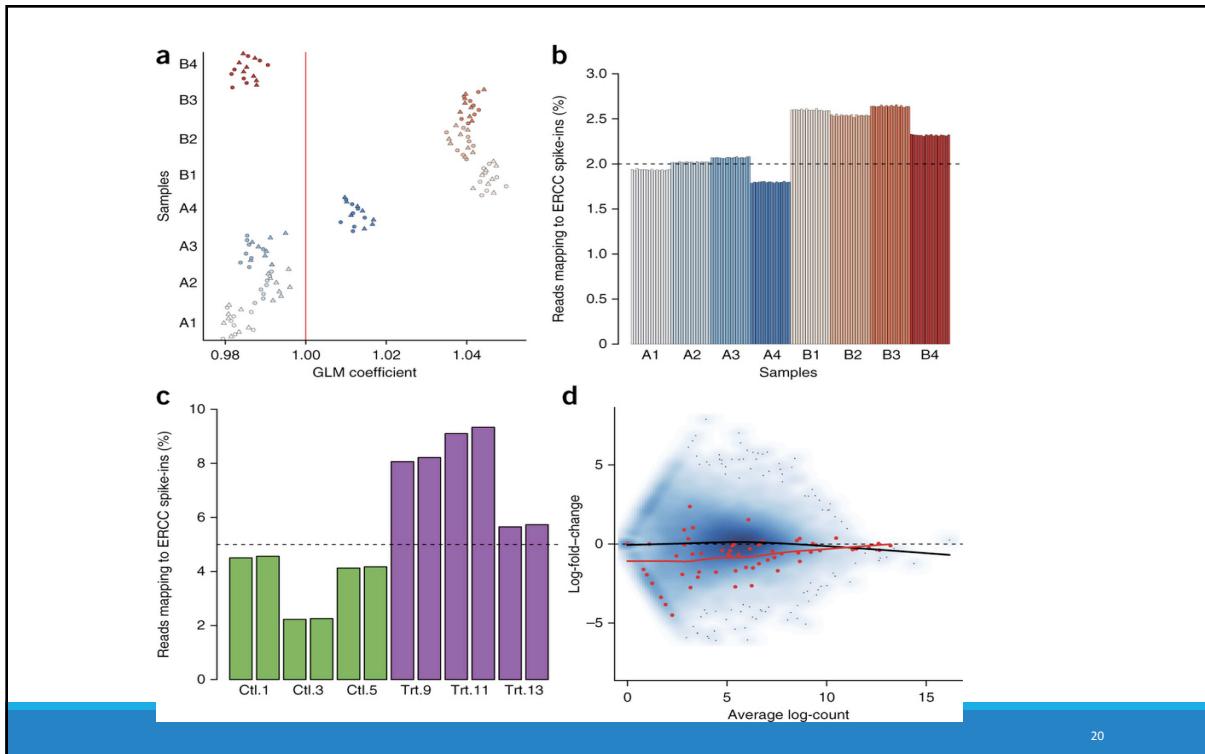
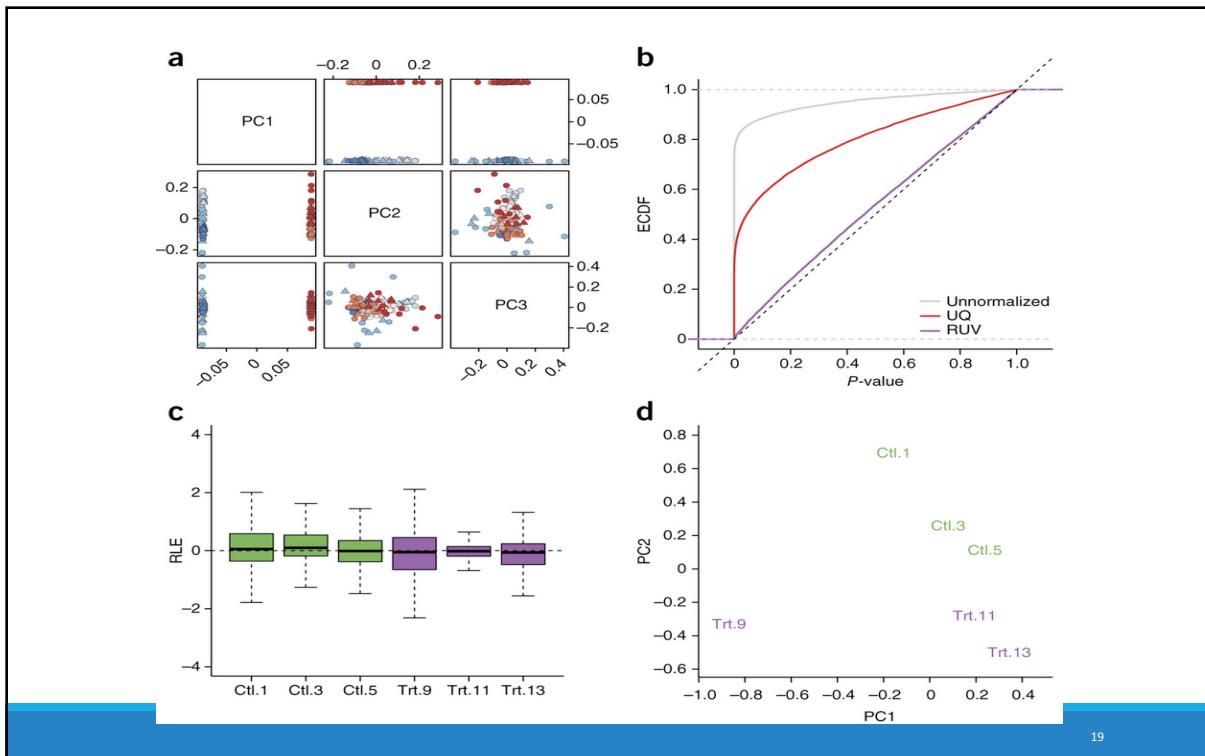
A UQ-normalization		B RUV-normalization		
		Upregulated		
	Upregulated		genes	Enriched pathways
OLM vs HC+CC	34	MAPK	OLM vs HC+CC	117
FC vs HC+CC	52	MAPK	FC vs HC+CC	210
			MAPK, p53, cell cycle, circadian rhythms, cancer	
			MAPK, T-cell receptor signaling	
Downregulated		Downregulated		
	genes	Enriched pathways	genes	Enriched pathways
OLM vs HC+CC	554	Ribosome, glycolysis, Lupus	OLM vs HC+CC	43
FC vs HC+CC	166	None	FC vs HC+CC	118
			None	
			Chemokine signaling, JAK-STAT, Toll-like receptor, RIG-I signaling	

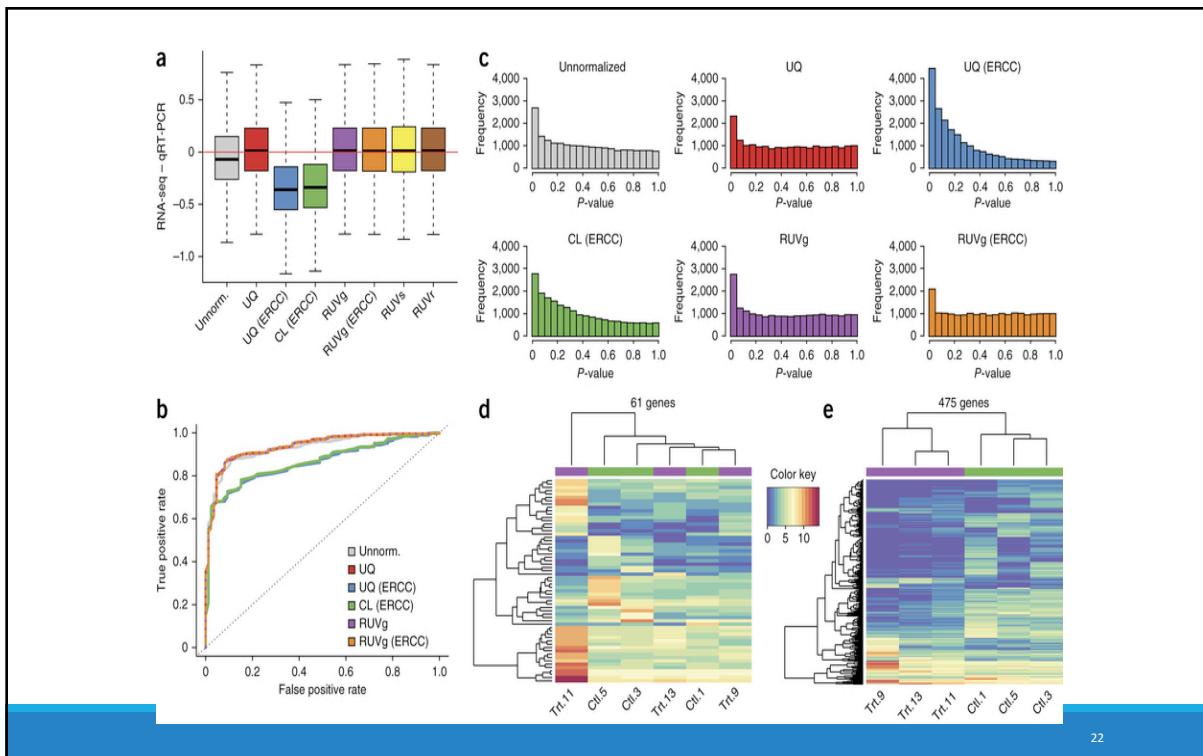
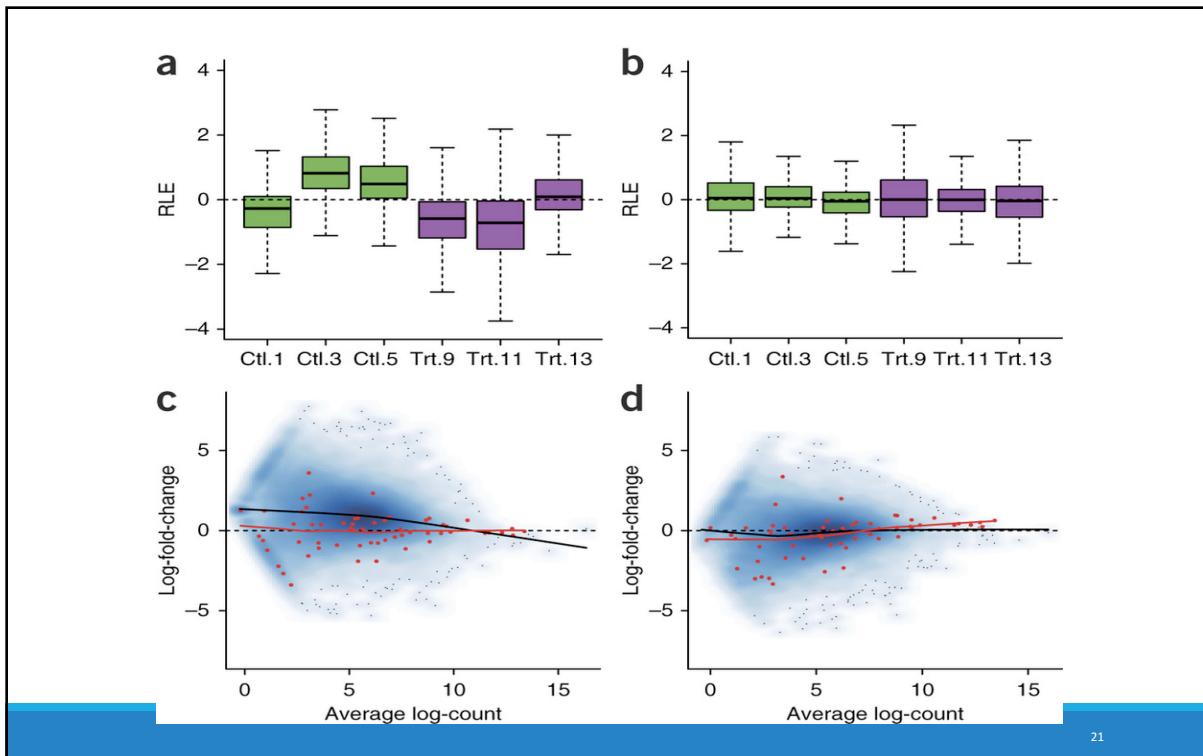
Lucia Peixoto et al. *Nucl. Acids Res.* 2015;nar.gkv736



Using Controls For The Normalization Of Rna-seq Data







Normalization

One often overlooked aspect is **normalization**, which is the **transformation of values that allows comparisons between samples** in a way that eliminates the effects of sources of variability that are not of interest.

We refer to those effects as ‘unwanted variation’. A variety of technical and biological factors, collectively known as ‘**batch effects**’, contribute unwanted variation to genome-wide gene expression data.

These factors include differences in amount of RNA, library preparation, equipment, operators, and procedures for sample extraction, preservation, or storage.

Proper normalization, or removal of these factors, has been shown to critically impact the analysis of high-throughput data (1–3). In spite of this, commonly used methods for RNA-seq normalization, such as upper quartile scaling (UQ)(2), trimmed mean of M values (TMM)(4) and FPKM (5), account only for global differences in sequencing depth between libraries (6).

The R code to reproduce all the main figures and tables of the article is available as tutorials in the supplementary material and downloadable from GitHub (github.com/drisso/peixoto2015_tutorial).

23

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

Table 1

Software packages for detecting differential expression

Method	Version	Reference	Normalization ^a	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[⁴]	TMM/Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[⁵]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[⁶]	Scaling factors (<u>quantile</u> /TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[⁷]	<u>RPKM</u> /TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[⁸]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[⁹]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[¹⁰]	<u>Geometric</u> (DESeq-like)/quartile /classic-fpkpm	Beta negative binomial distribution	t-test
EBSeq	1.1.7	[¹¹]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

^aIn case of availability of several normalization methods, the default one is underlined.

24

Count Based Statistics

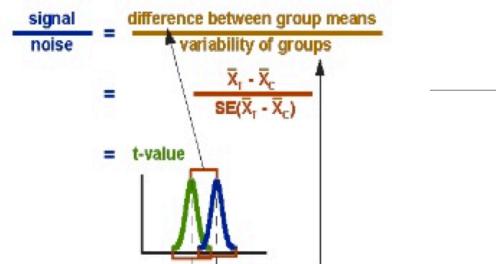
People often use discrete distributions (Poisson, negative binomial etc.) rather than continuous (e.g. normal) distributions for modeling RNA-seq data.

This is natural when you consider the way data are generated.

25

Problems associated with a t test

Couldn't we just use a Student's t test for each gene?



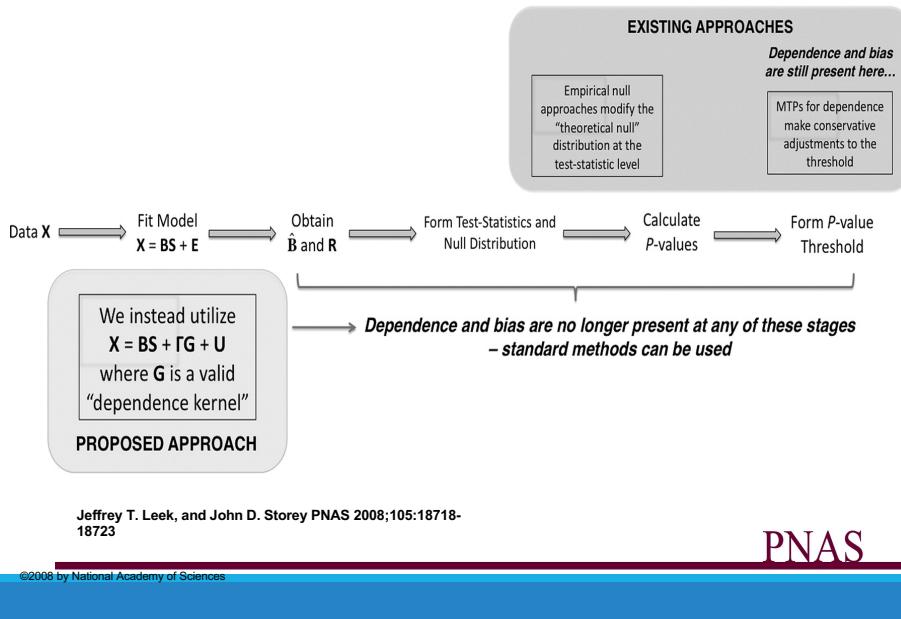
http://www.socialresearchmethods.net/kb/stat_t.php

Problems with this approach:

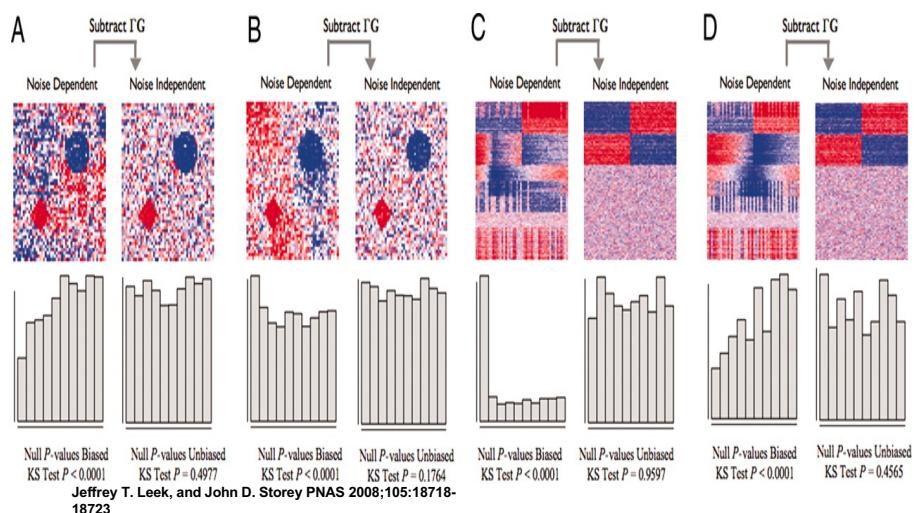
- May have **few replicates**
- Distribution is **not normal**
- **Multiple testing** issues

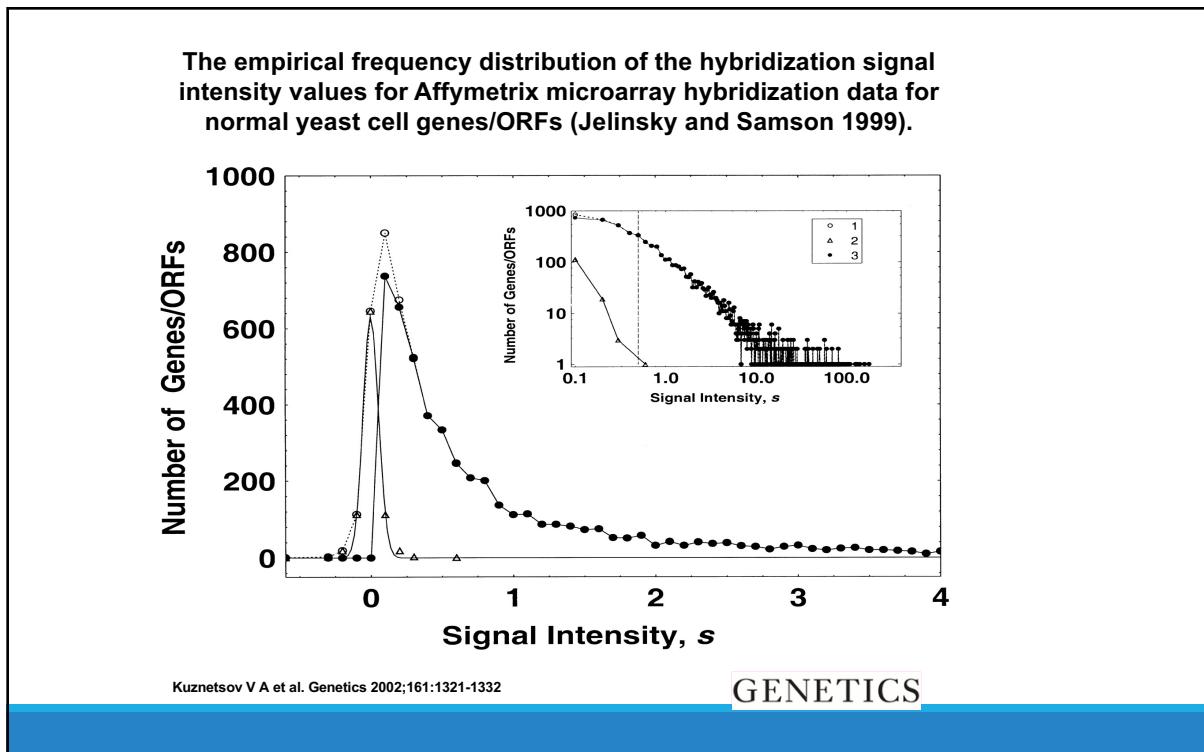
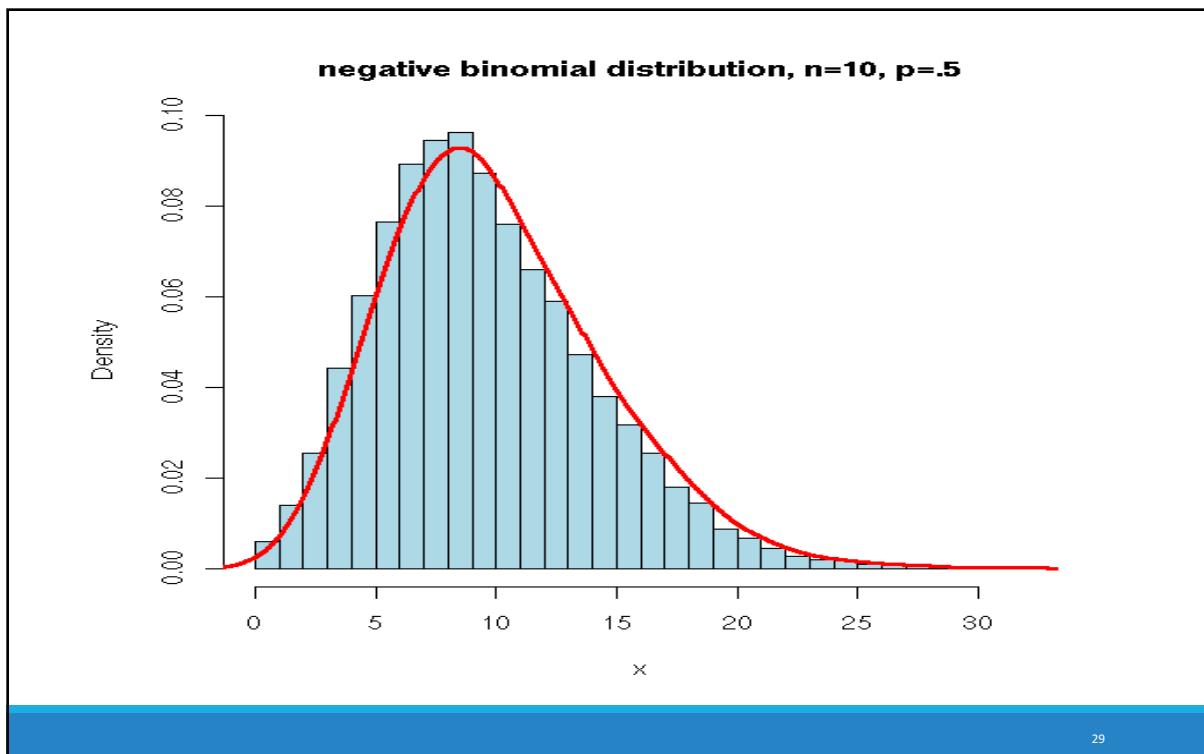
26

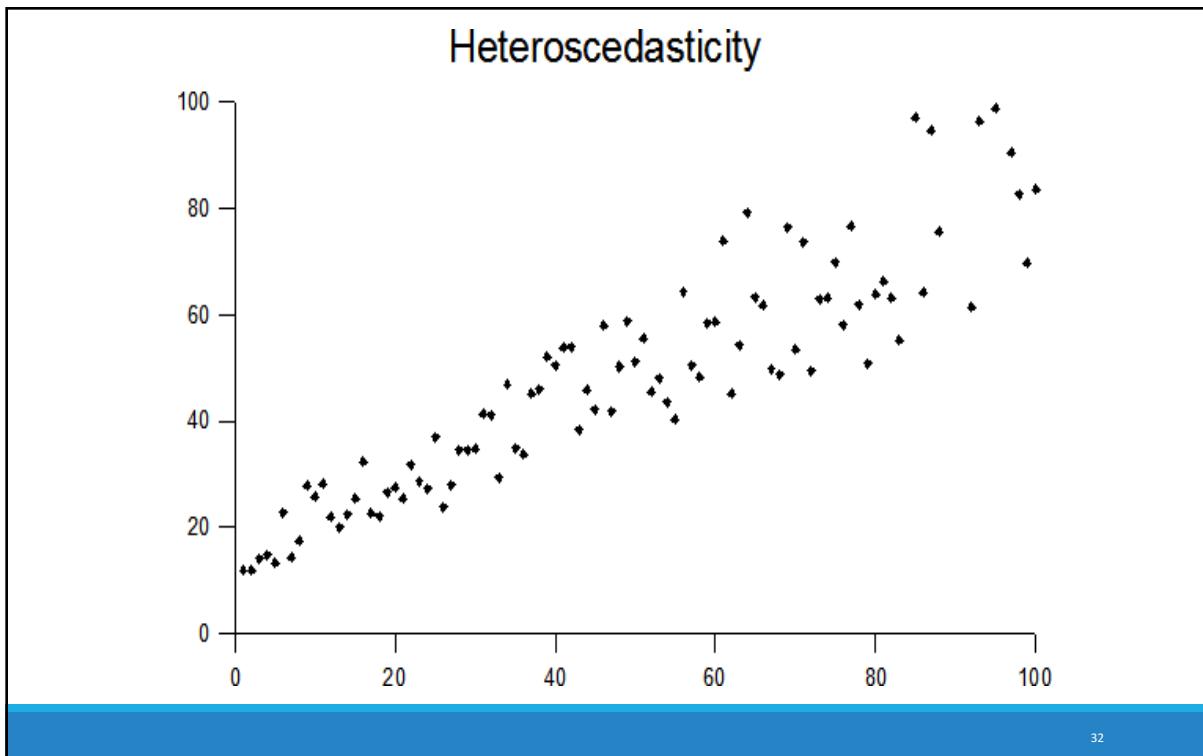
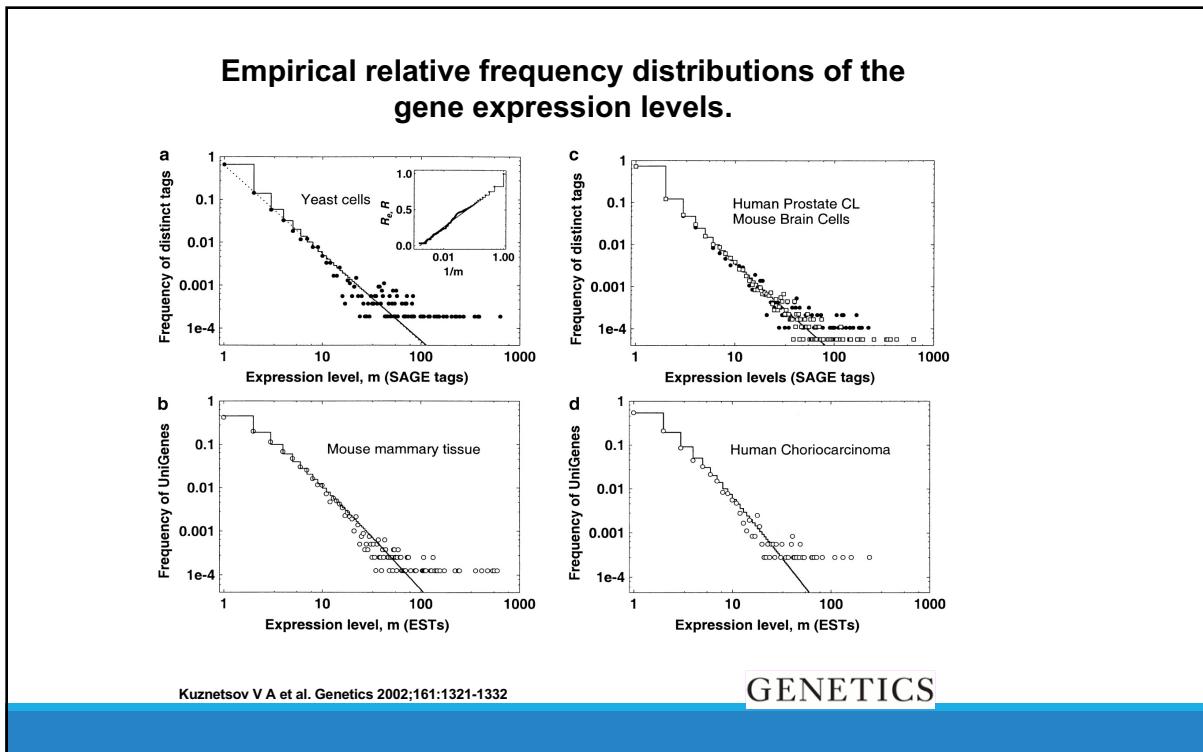
A schematic of the general steps of multiple hypothesis testing.



Simulated examples of multiple testing dependence.







 **Bioconductor**
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

[Home](#) [Install](#) [Help](#)

Home » Bioconductor 3.2 » Software Packages » sSeq

sSeq

platforms all | downloads top 50% | posts 0 | in BioC 2.5 years
 build ok | commits 0.33 | test coverage unknown

Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

Bioconductor version: Release (3.2)

The purpose of this package is to discover the genes that are differentially expressed between two conditions in RNA-seq experiments. Gene expression is measured in counts of transcripts and modeled with the Negative Binomial (NB) distribution using a shrinkage approach for dispersion estimation. The method of moment (MM) estimates for dispersion are shrunk towards an estimated target, which minimizes the average squared difference between the shrinkage estimates and the initial estimates. The exact per-gene probability under the NB model is calculated, and used to test the hypothesis that the expected expression of a gene in two conditions identically follow a NB distribution.

Author: Danni Yu <dyu@purdue.edu>, Wolfgang Huber <whuber@embl.de> and Olga Vitek <ovitek@purdue.edu>

Maintainer: Danni Yu <dyu@purdue.edu>

Citation (from within R, enter `citation("sSeq")`):

Yu D, Huber W and Vitek O (2013). *sSeq: Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size*. R package version 1.8.0.

33

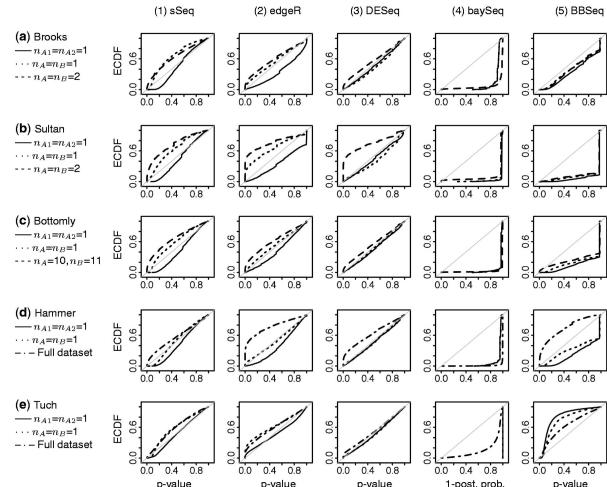
Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

Table 2.
 Areas under the ROC curves of detecting differentially expressed genes for the datasets with an external 'gold standard' while varying the FDR-adjusted *P*-value or posterior probability cutoff

Methods	Simulation1	Simulation2	Simulation3	MAQC Project	Griffith et al.				
$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_A = 3, n_B = 2$	$n_i = 1$			
Proposed sSeq	0.947	0.962	0.951	0.967	0.856	0.888	0.585	0.911	0.689
Existing edgeR	0.918	0.948	0.938	0.951	0.840	0.833	0.558	0.850	0.557
DESeq	0.932	0.940	0.937	0.949	0.842	0.816	0.577	0.867	0.596
baySeq	0.568	0.711	0.548	0.714	0.558	0.628	0.551	0.852	0.702
BBSeq	0.675	0.672	0.669	0.674	0.578	0.619	0.560	0.617	0.544
SAMseq		0.964		0.968		0.882		0.563	

Sub-columns are subsets of the data with one randomly selected replicate per condition and the full datasets. Values closer to 1 indicate higher sensitivity and specificity.

34



Danni Yu et al. Bioinformatics 2013;29:1275-1282

© The Author 2013. Published by Oxford University Press

Bioinformatics

Multiple Biological Replicates Are Necessary

sSeq can produce meaningful results in under-replicated RNA-seq screens.

However, we stress that RNA-seq screens do not eliminate the biological variation in gene expression [Equation \(12\)](#).

As evidenced by [Table 2](#) and [Figure 2](#), the under-replicated screens have lower reproducibility as compared with the replicated studies.

Multiple biological replicates are necessary to adequately assess the full extent of the variation in the biological system.

Therefore, the under-replicated screens can only be conducted when followed by a rigorous experimental validation with complementary technologies and adequate sample size.

Experimental Design: Recommended Software

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > q-bio > arXiv:1505.02017

Search or Article-id (Help | Advanced search) All papers Go!

Quantitative Biology > Genomics

Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment

Nicholas J. Schurch, Pieta Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Charbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, Geoffrey J. Barton

Submitted on 8 May 2015 (v1), last revised 8 Jun 2015 (this version, v2)

An RNA-seq experiment with 48 biological replicates in each of 2 conditions was performed to determine the number of biological replicates (n_r) required, and to identify the most effective statistical analysis tools for identifying differential gene expression (DGE). When $n_r = 3$, seven of the nine tools evaluated give true positive rates (TPR) of only 20 to 40 percent. For high fold-change genes ($\log_2(FC) > 2$) the TPR is > 85 percent. Two tools performed poorly; over- or under-predicting the number of differentially expressed genes. Increasing replication gives a large increase in TPR when considering all DE genes but only a small increase for high fold-change genes. Achieving a TPR > 85% across all fold-changes requires $n_r > 20$. For future RNA-seq experiments these results suggest $n_r > 6$, rising to $n_r > 12$ when identifying DGE irrespective of fold-change is important. For $6 < n_r < 12$, superior TPR makes edgeR the leading tool tested. For $n_r \geq 12$, minimizing false positives is more important and DESeq outperforms the other tools.

Comments: 21 Pages and 4 Figures in main text. 9 Figures in Supplement attached to PDF. Revision to correct a minor error in the abstract
Subjects: Genomics (q-bio.GN)
Cite as: arXiv:1505.02017 [q-bio.GN]
(or arXiv:1505.02017v2 [q-bio.GN] for this version)

Download:
• PDF only (license)
Current browse context:
q-bio.GN
< prev | next >
new | recent | 1505
Change to browse by:
q-bio
References & Citations
• NASA ADS
Bookmark what is this?

37

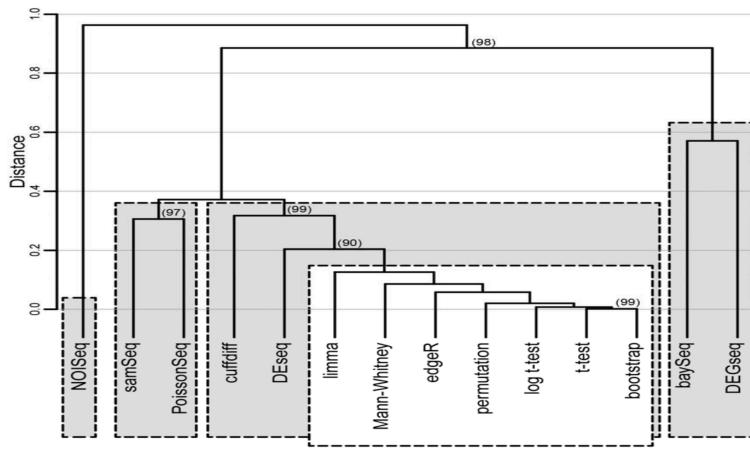
Experimental Design: Recommended Software

Table 1: RNA-seq differential gene expression tools and statistical tests.

Name	Assumed Distribution	Normalization	Description	Version	Citations ^a	Reference
<i>t-test</i>	normal	DESeq ^b	two-sample t-test for equal variances	-	-	-
<i>log t-test</i>	log-normal	DESeq ^b	log-ratio t-test	-	-	-
<i>Mann-Whitney</i>	none	DESeq ^b	Mann-Whitney test	-	-	Mann and Whitney (1947)
<i>Permutation</i>	none	DESeq ^b	permutation test	-	-	Efron and Tibshirani (1993)
<i>Bootstrap</i>	normal	DESeq ^b	bootstrap test	-	-	Efron and Tibshirani (1993)
<i>baySeq^c</i>	negative binomial	Internal	Empirical Bayesian estimate of posterior likelihood	1.8	109	Hardcastle and Kelly (2010)
<i>Cuffdiff</i>	negative binomial	Internal	unknown	2.1.1	481	Trapnell et al. (2012)
<i>DEGseq^d</i>	binomial	None	random sampling model using Fisher's exact test and the likelihood ratio test	1.10.0	215	Wang et al. (2010)
<i>DESeq^d</i>	negative binomial	DESeq ^b	Shrinkage variance	1.4.1	1204	Anders and Huber (2010)
<i>edgeR^d</i>	negative binomial	TMM ^e	Empirical Bayes estimation & an exact test analogous to Fisher's exact test but adapted to over-dispersed data	2.2.5	822	Robinson et al. (2010)
<i>Limma^f</i>	Log-normal	TMM ^e	Generalised linear model	3.4.4	15	Law et al. (2014)
<i>NOISEq^f</i>	None	RPKM	Non-parametric test based on signal-to-noise ratio	29/04/2011	113	Tarazona et al. (2011)
<i>PoissonSeq^f</i>	Poisson log-linear model	Internal	Score statistic	1.1	25	Li et al. (2012)
<i>SAMSeq^f</i>	None	Internal	Mann-Whitney test with Poisson resampling	2.0	26	Li and Tibshirani (2013)

38

Experimental Design: Recommended Software

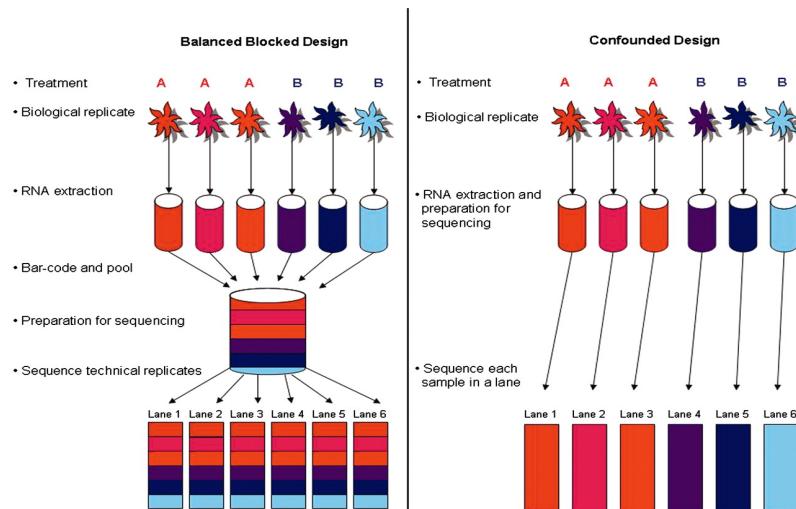


39

	Agreement with other tools ¹	WT v WT FPR ²	Fold-change Threshold (T) ³	Tool recommended for: (# good replicates per condition) ⁴		
				<=3	<=12	>12
<i>BaySeq</i>	inconsistent	Pass				
<i>cuffdiff</i>	consistent	Fail				
<i>DEGSeq</i>	inconsistent	Fail				
<i>DESeq</i>	consistent	Pass	0			Yes
			0.5		Yes	Yes
			2	Yes	Yes	Yes
<i>edgeR</i>	consistent	Pass	0			Yes
			0.5	Yes	Yes	Yes
			2	Yes	Yes	Yes
<i>Limma</i>	consistent	Pass	0			Yes
			0.5		Yes	Yes
			2	Yes	Yes	Yes
<i>NOISEq</i>	inconsistent	Pass				
<i>PoissonSeq</i>	inconsistent	Fail				
<i>SAMSeq</i>	inconsistent	Fail				

40

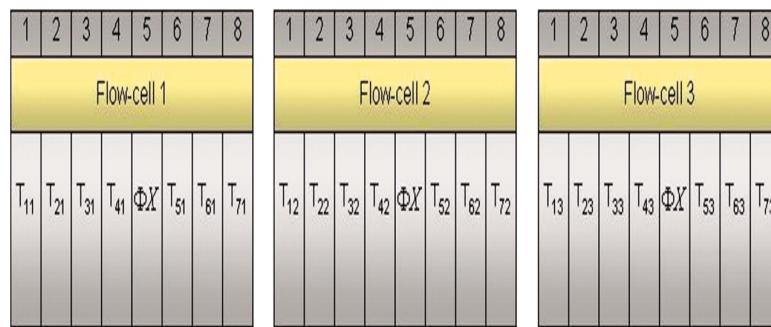
Comparison of two designs for testing differential expression between treatments A and B. Treatment A is denoted by red tones and treatment B by blue tones.



Paul L. Auer, and R. W. Doerge Genetics 2010;185:405-416

GENETICS

A multiple flow-cell design based on three biological replicates within seven treatment groups.



Paul L. Auer, and R. W. Doerge Genetics 2010;185:405-416

GENETICS

Technical vs biological replicates

Technical replicates:

Assess variability of measurement technique
 Typically low for bulk RNA-seq (not necessarily single-cell RNA-seq)
 Poisson distribution can model variability between RNA-seq technical replicates rather well

Biological replicates:

Assess variability between individuals / “normal” biological variation
 Necessary for drawing conclusions about biology
 Variability across RNA-seq biological replicates not well modelled by Poisson – usually negative binomial (“overdispersed Poisson”) is used

43

Replicates and differential expression

Intuitively, the variation **between** the groups that you want to compare should be large compared to the variation **within** each group to be able to say that we have differential expression.

The more biological replicates, the better you can estimate the variation. But how many replicates are needed?

Depends:

Homogeneous cell lines, inbred mice etc: maybe 3 samples / group enough.
 Clinical case-control studies on patients: can need a dozen, hundreds or thousands, depending on the specifics

44

7 Recommendations for RNA-seq experiment design

The results of this study suggest the following should be considered when designing an RNA-seq experiment for DGE:

- 1) At least 6 replicates per condition for all experiments.
- 2) At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.
- 3) For experiments with < 12 replicates per condition; use *edgeR*.
- 4) For experiments with > 12 replicates per condition; use *DESeq*.

12

- 5) Apply a fold-change threshold appropriate to the number of replicates per condition between $0.1 \leq T \leq 0.5$ (see Figure 2 and the discussion of tool performance as a function of replication).

45

McIntyre et al. BMC Genomics 2011, 12:293
<http://www.biomedcentral.com/1471-2164/12/293>

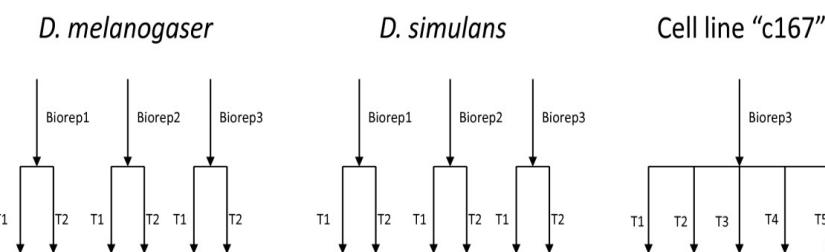


RESEARCH ARTICLE

Open Access

RNA-seq: technical variability and sampling

Lauren M McIntyre^{1*}, Kenneth K Lopiano², Alison M Morse¹, Victor Amin¹, Ann L Oberg³, Linda J Young² and Sergey V Nuzhdin⁴



46

A “Good” Lane Has ~30 Million Clusters. This Represents 0.0013 % Of The Starting Library

The number of molecules present in the library is estimated to be 2.408×10^{12} .

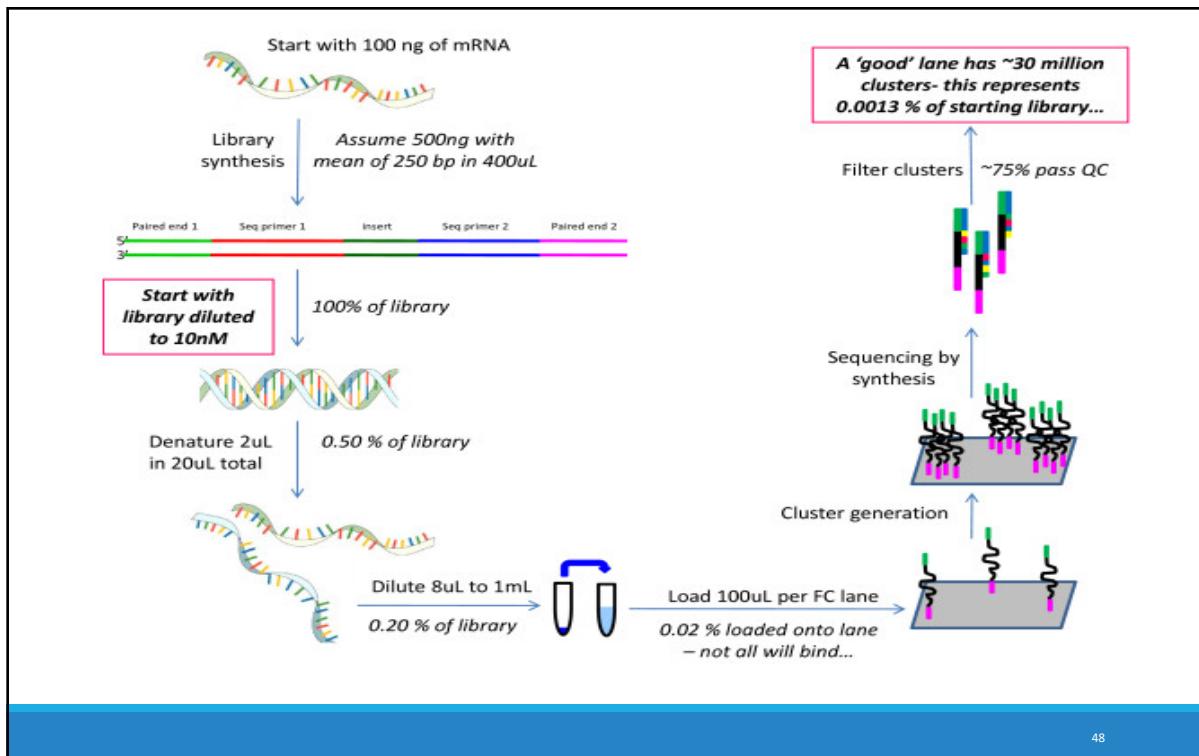
This estimate was derived using the Illumina protocol. Specifically, the starting material was assumed to be 100 ng of mRNA that then resulted in 500 ng of library with an insert size of 250bp in a volume of 400 uL for a final library concentration of 10 nM.

The number of pmoles is therefore $4 [(10 \text{ nmol}/1\text{L}) (1\text{L}/1000 \text{ mL})(1 \text{ mL}/1000 \text{ uL})(400 \text{ uL})(1000 \text{ pmol}/1 \text{ nmol})]$.

The number of molecules is found by multiplying Avogadro's number which is molecules/moles and adjusting for units [$4 \text{ pmole}/1000/1000/1000*6.02E + 023$] to give 2.408×10^{12} molecules.

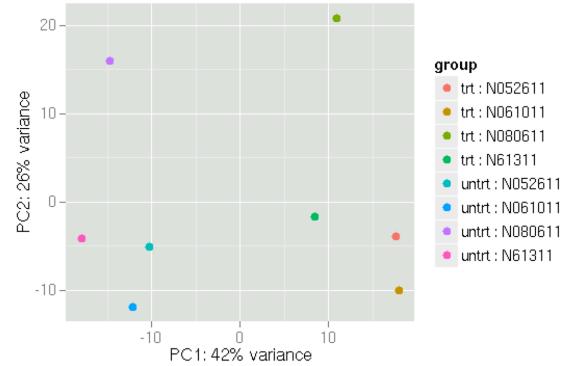
On the current Solexa/Illumina technology, the GAIIx, approximately 30 million of the total possible molecules are sampled in a given lane. This represents approximately 0.0013% ($30,000,000/2.408 \times 10^{12}$) of the total number of available molecules for analysis (Figure 2).

47



48

Intro to SVA

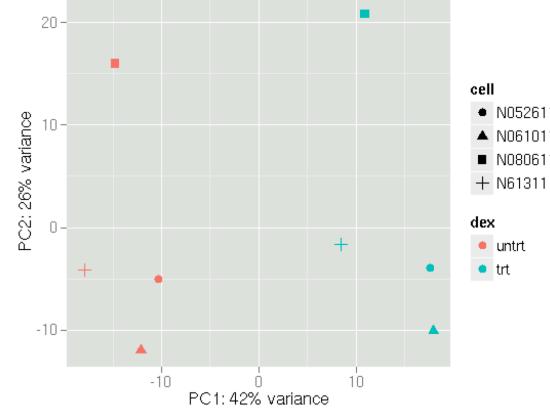


PCA plot using the rlog-transformed values. Each unique combination of treatment and cell line is given its own color.

Here, we have used the function `plotPCA` that comes with `DESeq2`. The two terms specified by `intgroup` are the interesting groups for labeling the samples; they tell the function to use them to choose colors. We can also build the PCA plot from scratch using the `ggplot2` package (Wickham 2009). This is done by asking the `plotPCA` function to return the data used for plotting rather than building the plot. See the [ggplot2 documentation](#) for more details on using `ggplot2`.

49

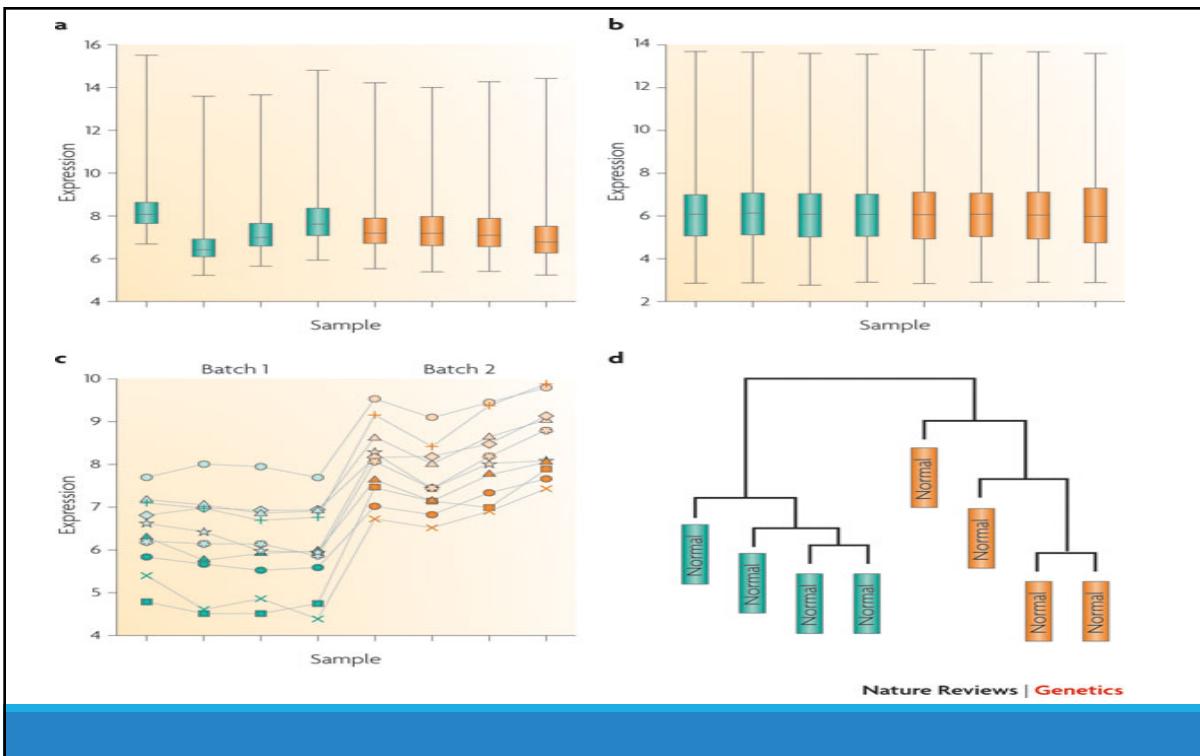
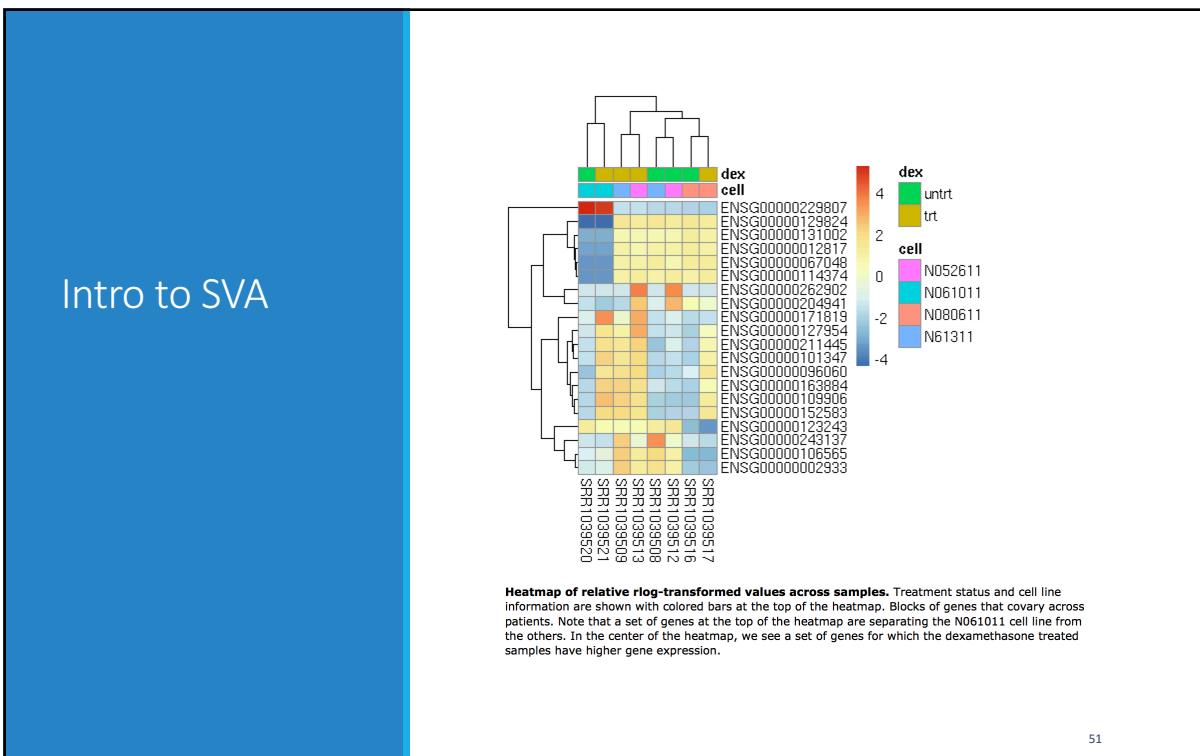
Intro to SVA

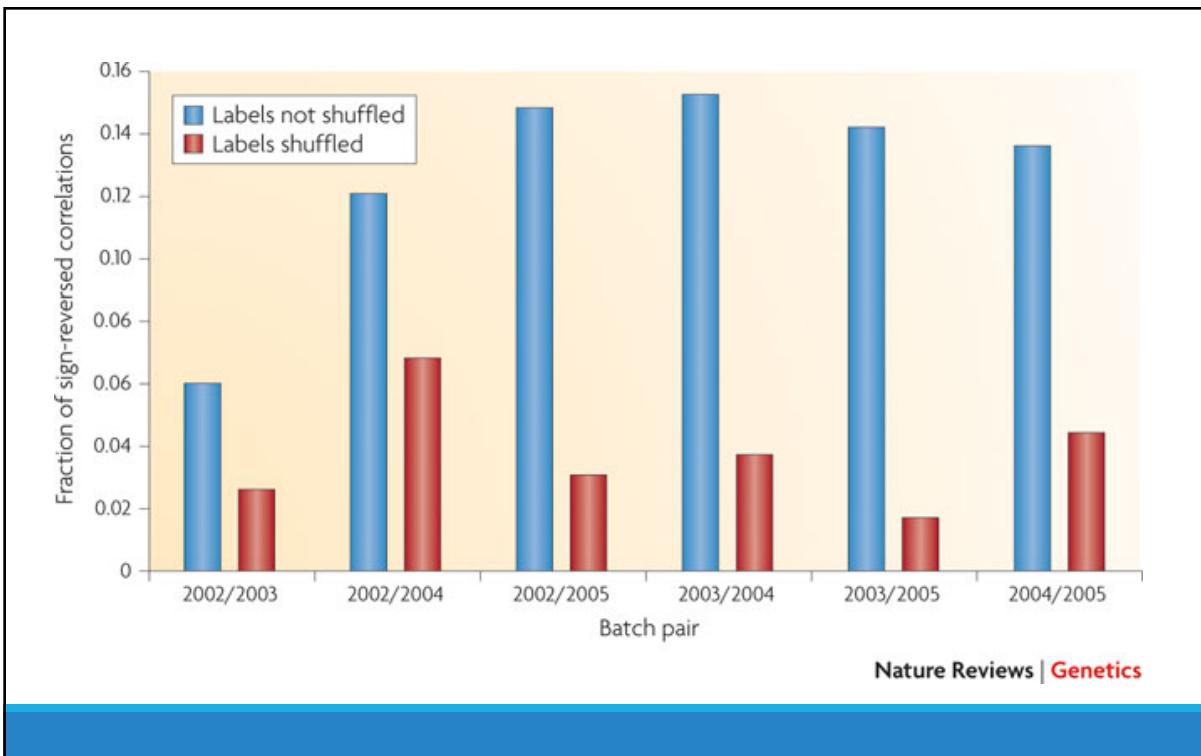
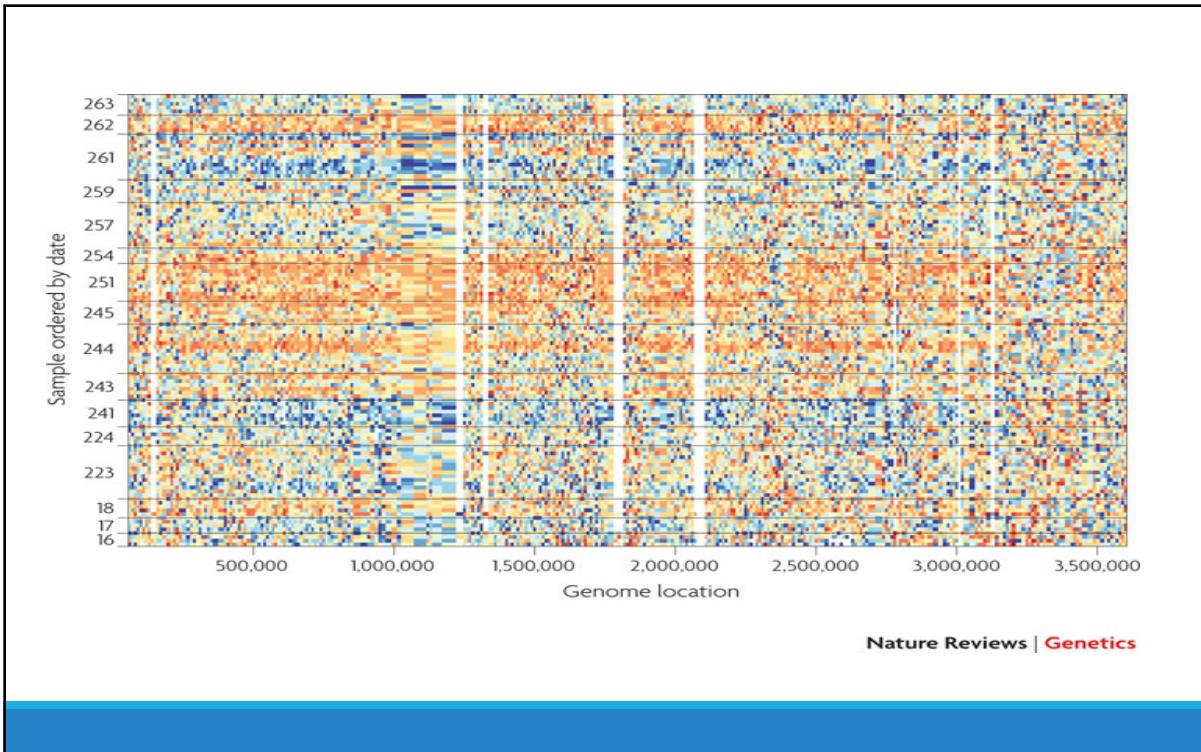


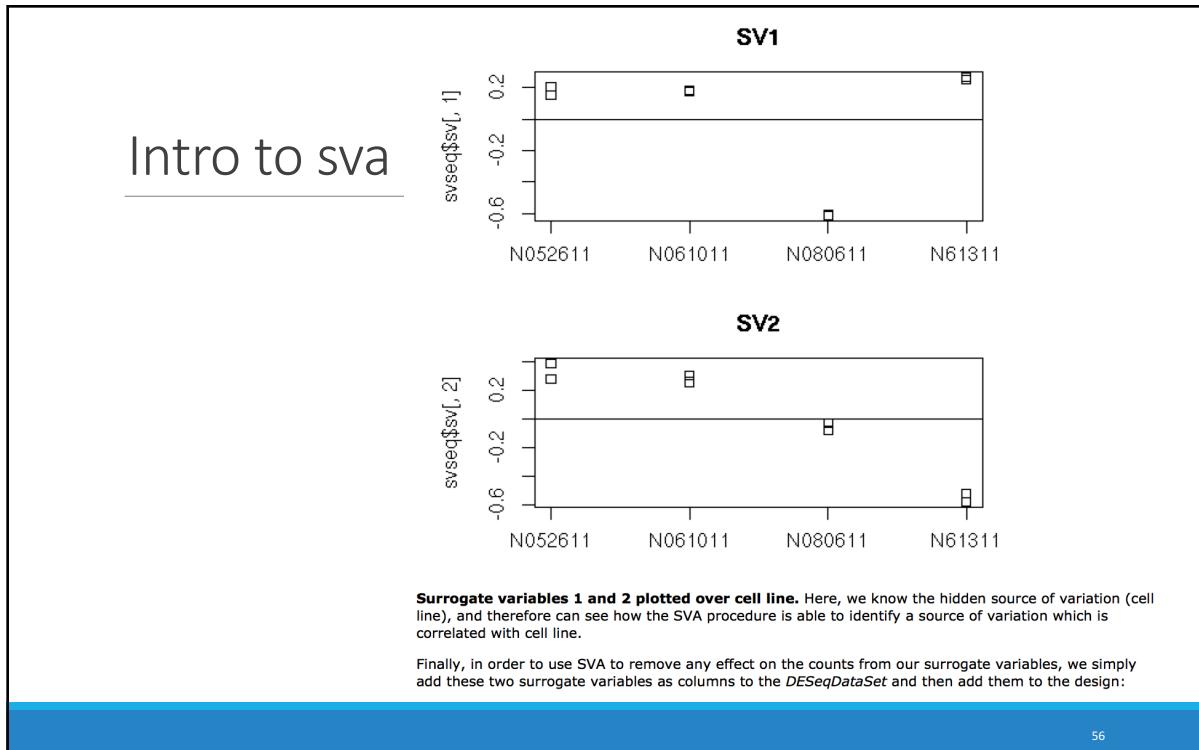
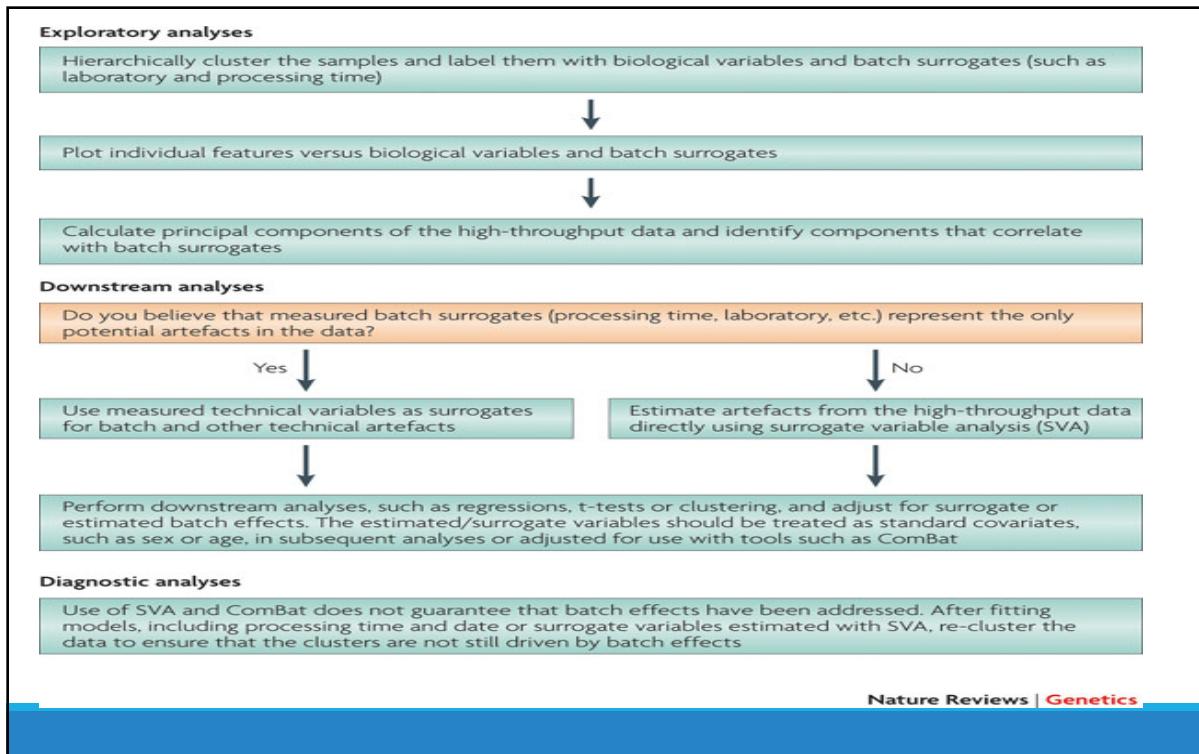
PCA plot using the rlog-transformed values with custom ggplot2 code. Here we specify cell line (plotting symbol) and dexamethasone treatment (color).

From the PCA plot, we see that the differences between cells (the different plotting shapes) are considerable, though not stronger than the differences due to treatment with dexamethasone (red vs blue color). This shows why it will be important to account for this in differential testing by using a paired design ("paired", because each dex treated sample is paired with one untreated sample from the same cell line). We are already set up for this design by assigning the formula `~ cell + dex` earlier.

50







PLOS | GENETICS

Browse | Publish | About

OPEN ACCESS PEER-REVIEWED

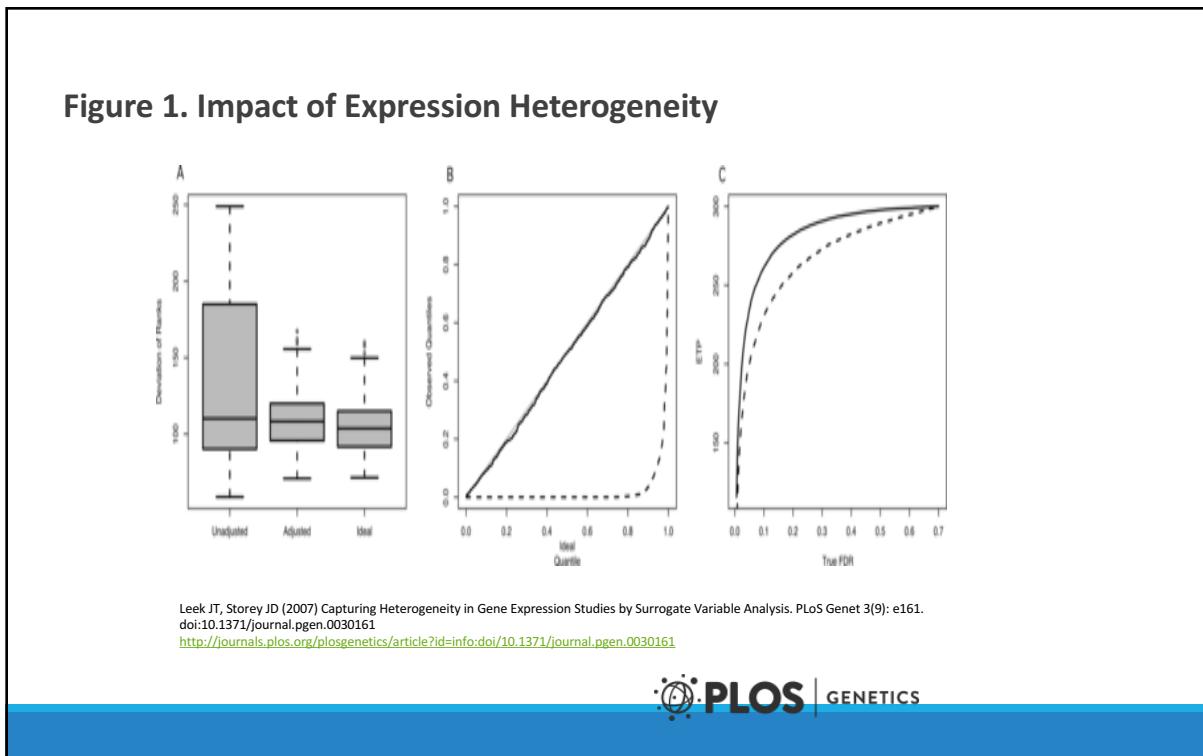
RESEARCH ARTICLE

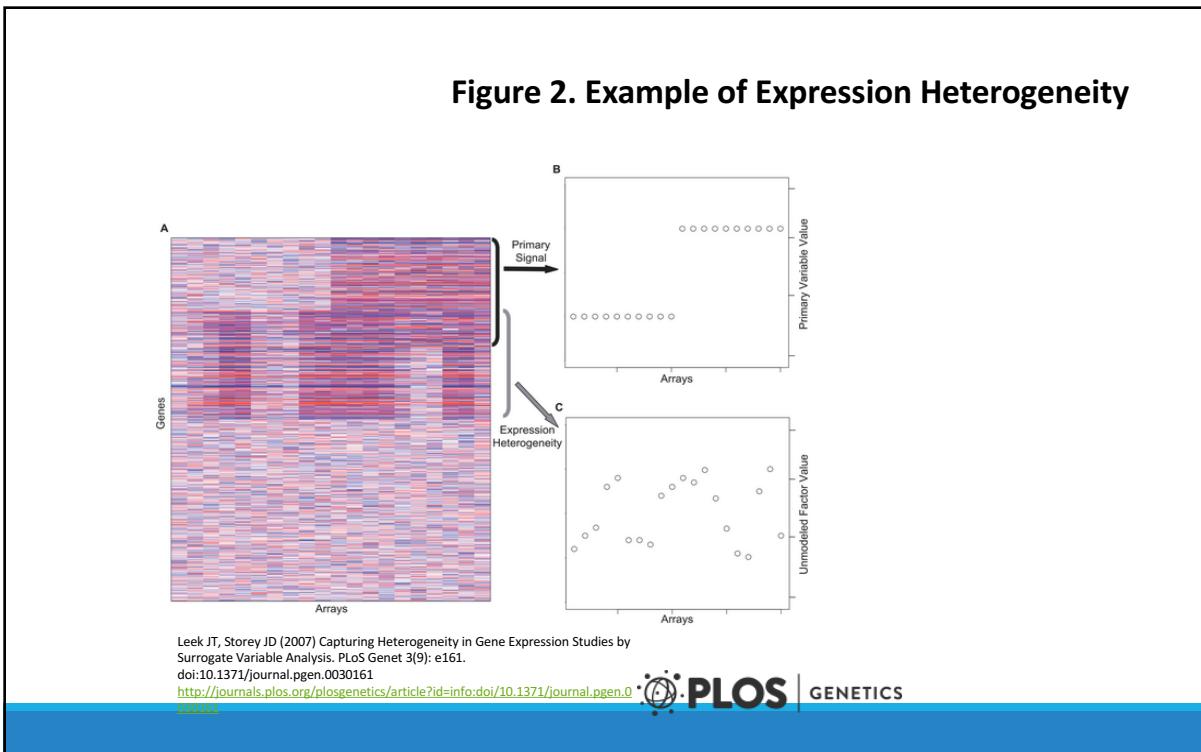
Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

Jeffrey T Leek, John D Storey 

Published: September 28, 2007 • DOI: 10.1371/journal.pgen.0030161

57





Study	Analysis Type	q-Value Threshold			
		0.01	0.025	0.05	0.10
Genetics of gene expression	Unadjusted	1,063	1,343	1,604	1,951
	SVA adjusted	1,428	1,676	1,894	2,292
Disease Class	Unadjusted	1	19	96	274
	SVA adjusted	1	1	52	218
Time course	Unadjusted	161	273	422	823
	Tissue adjusted	270	482	795	1,548
	SVA Adjusted	196	367	563	991

The results of the significance analysis in the three real gene expression studies. The results of the genetics of gene expression study include the number of significant *cis*-linkages before and after adjusting for surrogate variables. The disease class results report the number of genes differentially expressed between *BRCA1* and *BRCA2* before and after adjusting for surrogate variables. For the time-course study, the number of genes differentially expressed with respect to age are shown for an unadjusted analysis, an analysis adjusted for tissue type, and an SVA-adjusted analysis. An SVA-adjusted analysis may result in an increase or decrease in the number of significant results depending on the direction and degree to which the unmodeled factors (now captured by surrogate variables) were confounded with the primary variables.

doi:10.1371/journal.pgen.0030161.t001

Nucleic Acids Research Advance Access published October 7, 2014

Nucleic Acids Research, 2014 1
doi: 10.1093/nar/gku864

svaseq: removing batch effects and other unwanted noise from sequencing data

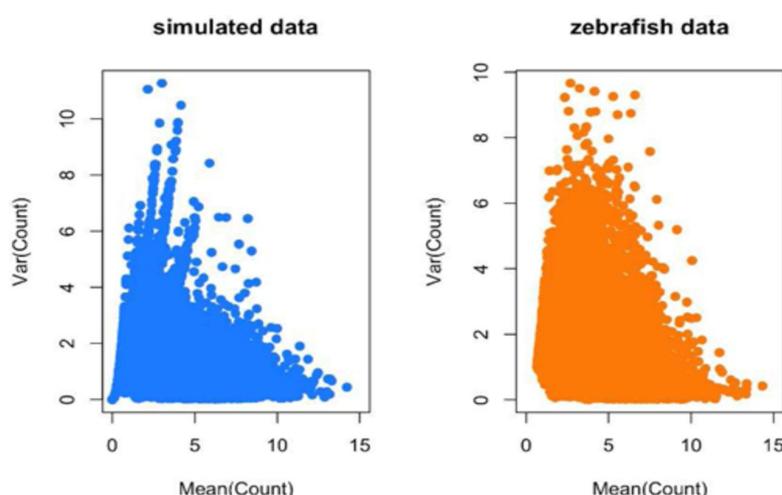
Jeffrey T. Leek^{*}

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health Baltimore, MD 21212, US

Received June 24, 2014; Revised August 20, 2014; Accepted September 8, 2014

61

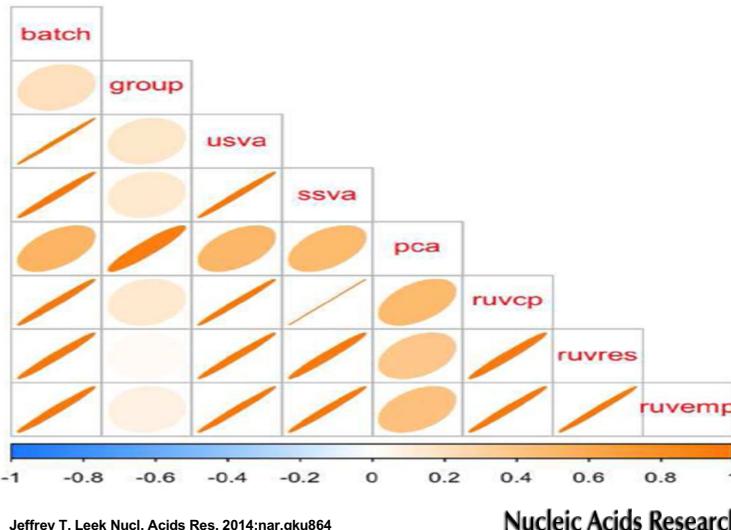
Distribution of means and variances for simulated and real Zebrafish data.



Jeffrey T. Leek *Nucl. Acids Res.* 2014;nar.gku864

Nucleic Acids Research

Correlation between simulated batch and group variables and various batch estimates.



Jeffrey T. Leek Nucl. Acids Res. 2014;nar.gku864

Nucleic Acids Research

Normalization

One often overlooked aspect is **normalization**, which is the **transformation of values that allows comparisons between samples** in a way that eliminates the effects of sources of variability that are not of interest.

We refer to those effects as ‘unwanted variation’. A variety of technical and biological factors, collectively known as ‘**batch effects**’, contribute unwanted variation to genome-wide gene expression data.

These factors include differences in amount of RNA, library preparation, equipment, operators, and procedures for sample extraction, preservation, or storage.

Proper normalization, or removal of these factors, has been shown to critically impact the analysis of high-throughput data (1–3). In spite of this, commonly used methods for RNA-seq normalization, such as upper quartile scaling (UQ)(2), trimmed mean of M values (TMM)(4) and FPKM (5), account only for global differences in sequencing depth between libraries (6).

The R code to ***reproduce all the main figures*** and tables of the article is available as tutorials in the supplementary material and downloadable from GitHub (github.com/drisso/peixoto2015_tutorial).

The screenshot shows the GitHub repository page for `drisso/peixoto2015_tutorial`. The repository title is "Tutorial to reproduce the analysis of Peixoto et al. (2015)". It has 23 commits, 1 branch, 0 releases, and 1 contributor. The latest commit was made on Nov 25, 2015. The commit history includes changes to README.md, input files, and additional files like Rmd, html, and pdf.

File	Description	Date
Peixoto_Input_for_Additional_file_1	Changed input files location	8 months ago
Peixoto_Input_for_Additional_file_2	Changed input files location	8 months ago
.gitignore	Add gitignore	9 months ago
Peixoto_Additional_File_1.R	Changed input files location	8 months ago
Peixoto_Additional_File_1.Rmd	Changed input files location	8 months ago
Peixoto_Additional_File_1.html	Changed input files location	8 months ago
Peixoto_Additional_File_1.pdf	Changed input files location	8 months ago
Peixoto_Additional_File_2.R	Added R and PDF file for second tutorial	8 months ago
Peixoto_Additional_File_2.Rmd	Changed input files location	8 months ago
Peixoto_Additional_File_2.html	Changed input files location	8 months ago
Peixoto_Additional_File_2.pdf	Added R and PDF file for second tutorial	8 months ago
README.md	Update README.md	3 months ago
biblio.bb	Initial commit	9 months ago

65

The screenshot shows the content of the `README.md` file. It starts with a section titled "Tutorial to reproduce the analysis of Peixoto et al. (2015)". Below it is a heading "Analysis of FC and OLM datasets". It mentions that all needed input files are in the `Peixoto_Input_for_Additional_file_1` directory. It also notes that the tutorial has been tested with R 3.2.0. The `Peixoto_Additional_File_1.Rmd` file contains R Markdown and can be run to reproduce the full report. The `Peixoto_Additional_File_1.R` file contains a pure R script and can be sourced to any R session. The `Peixoto_Additional_File_1.html` and `Peixoto_Additional_File_1.pdf` files contain the results of the analysis. A link to the "Session Info" section is provided for details on package versions used. The next section is "Analysis of publicly available data", which discusses the processed expression data from GEO available in the `Peixoto_Input_for_Additional_file_2` directory. The `Peixoto_Additional_File_2.Rmd` file contains R Markdown and can be run to reproduce the analysis. The `Peixoto_Additional_File_2.html` file contains the results.

66

The End
