# RNA-seq with R-bioconductor

# Dealing with unwanted variation


- Maarten Leerkes PhD

# Refresher: three aspects

- What is R
- What is Bioconductor
- What is RNAseq

# Sequencing by synthesis

- Intro to Sequencing by Synthesis: Industry-leading Data Quality
- https://www.youtube.com/watch?v=HMyCqWhwB8E

# 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*

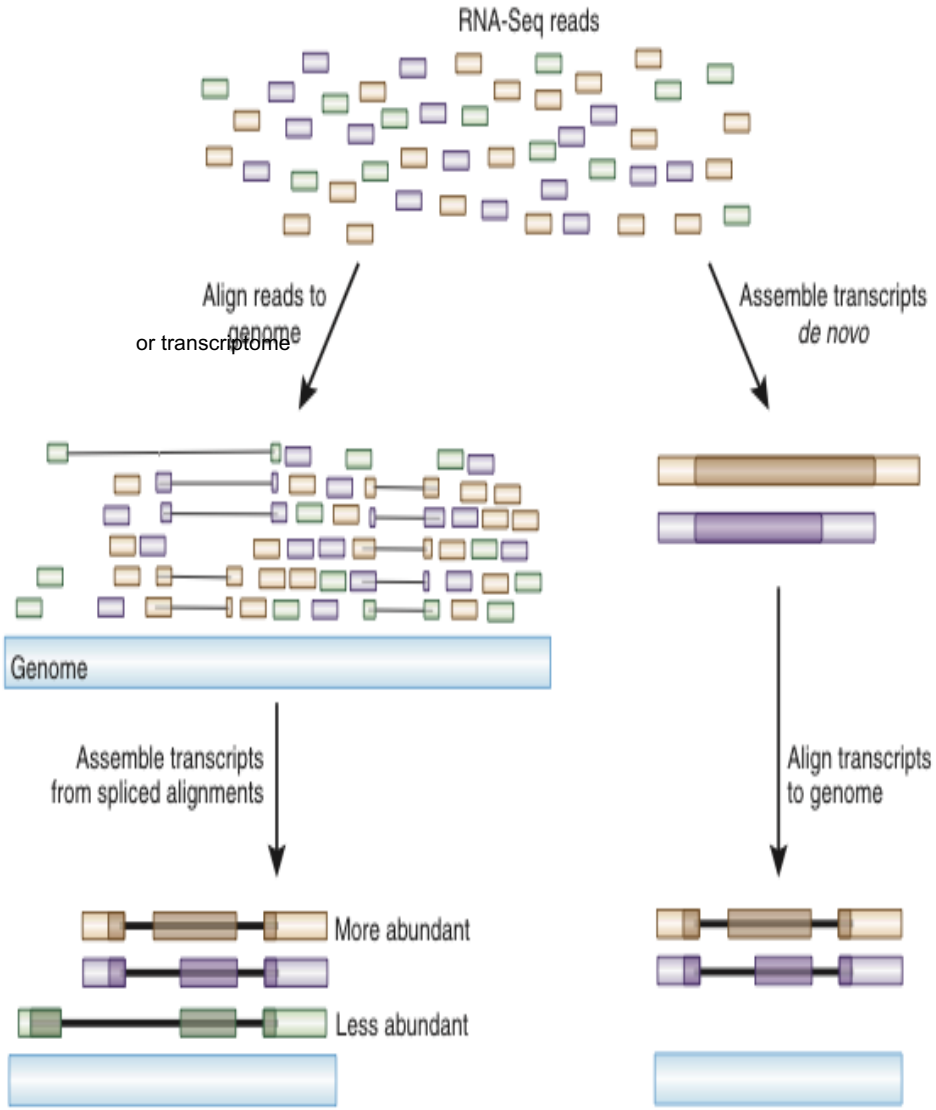
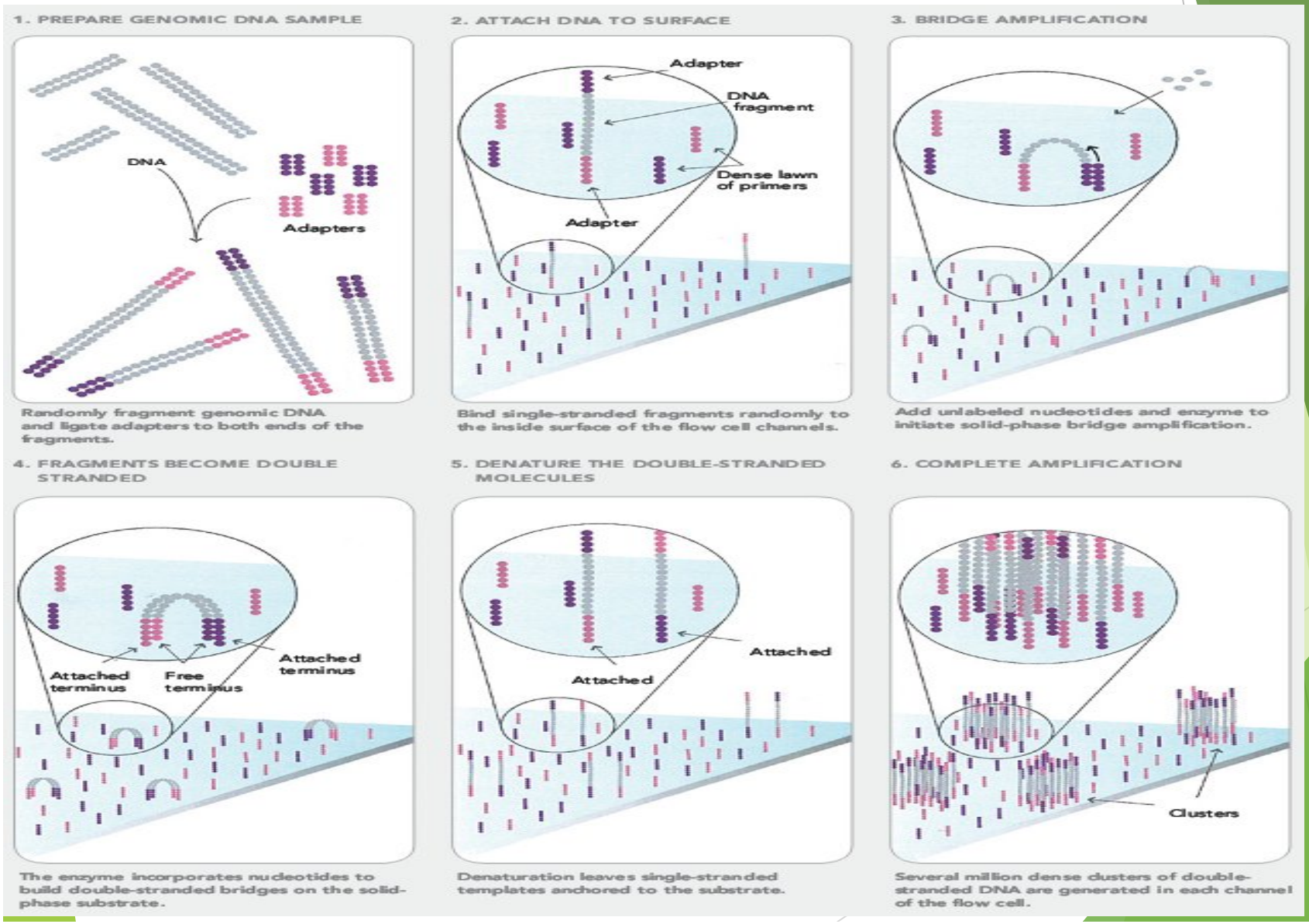


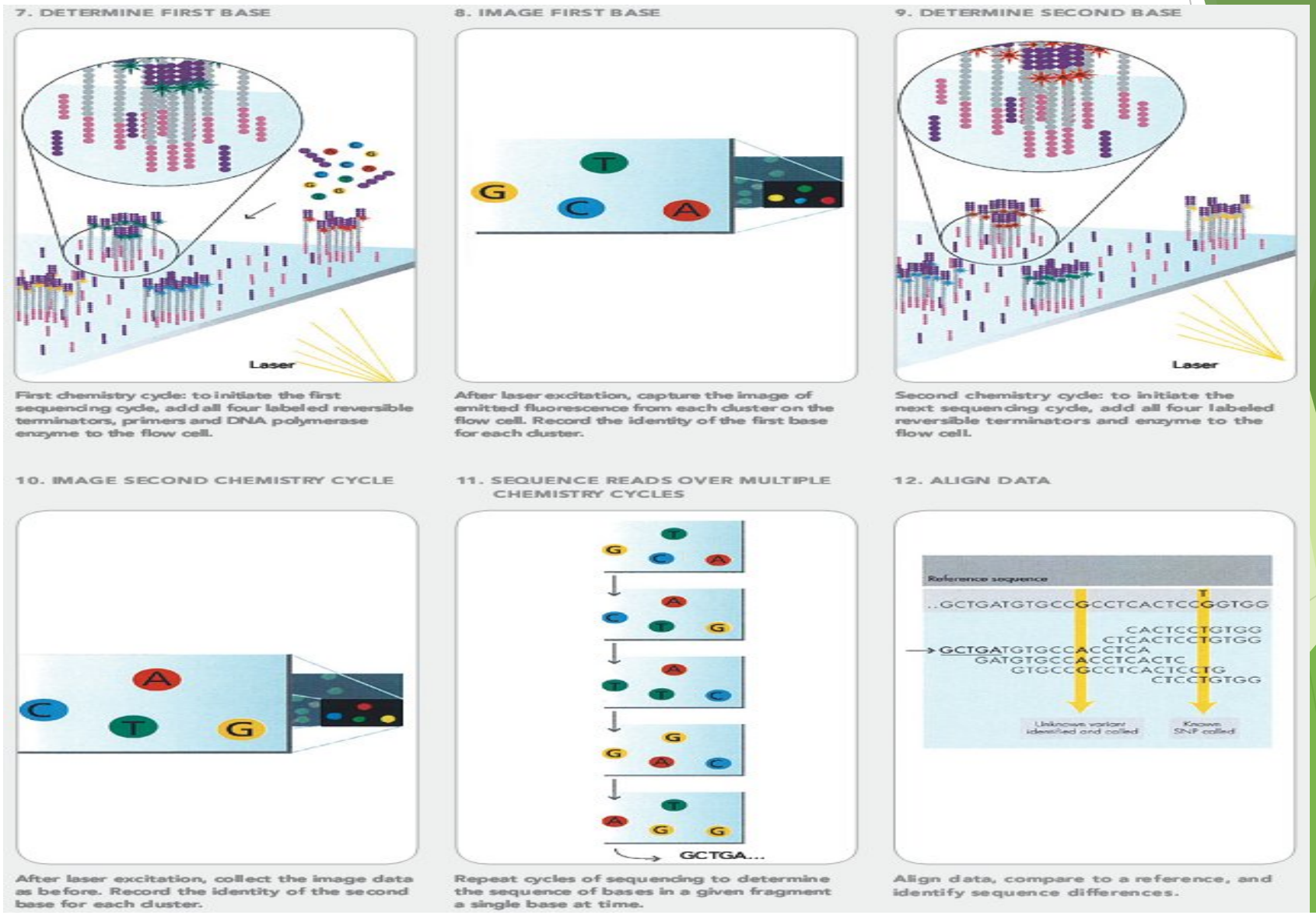

*Image from Haas & Zody, 2010*

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

# The basis of Illumina sequencing on a flow cell

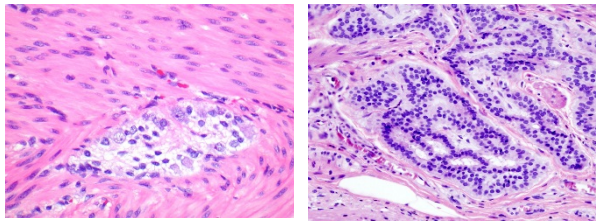


# Illumina clonal expansion followed by image processing



# RNA sequencing: abundance comparisons between two or more conditions / phenotypes

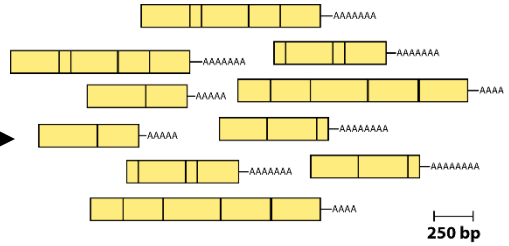
**Samples of interest**



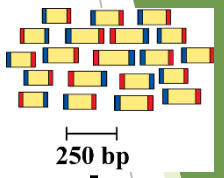
Condition 1 (normal tissue)

Condition 2 (diseased tissue)

**Isolate RNAs**

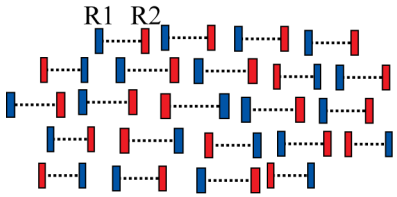



**Generate cDNA, fragment, size select, add linkers**




**Sequence ends**

**100s of millions of paired reads**
**10s of billions bases of sequence**




**Map to genome, transcriptome, and predicted exon junctions**

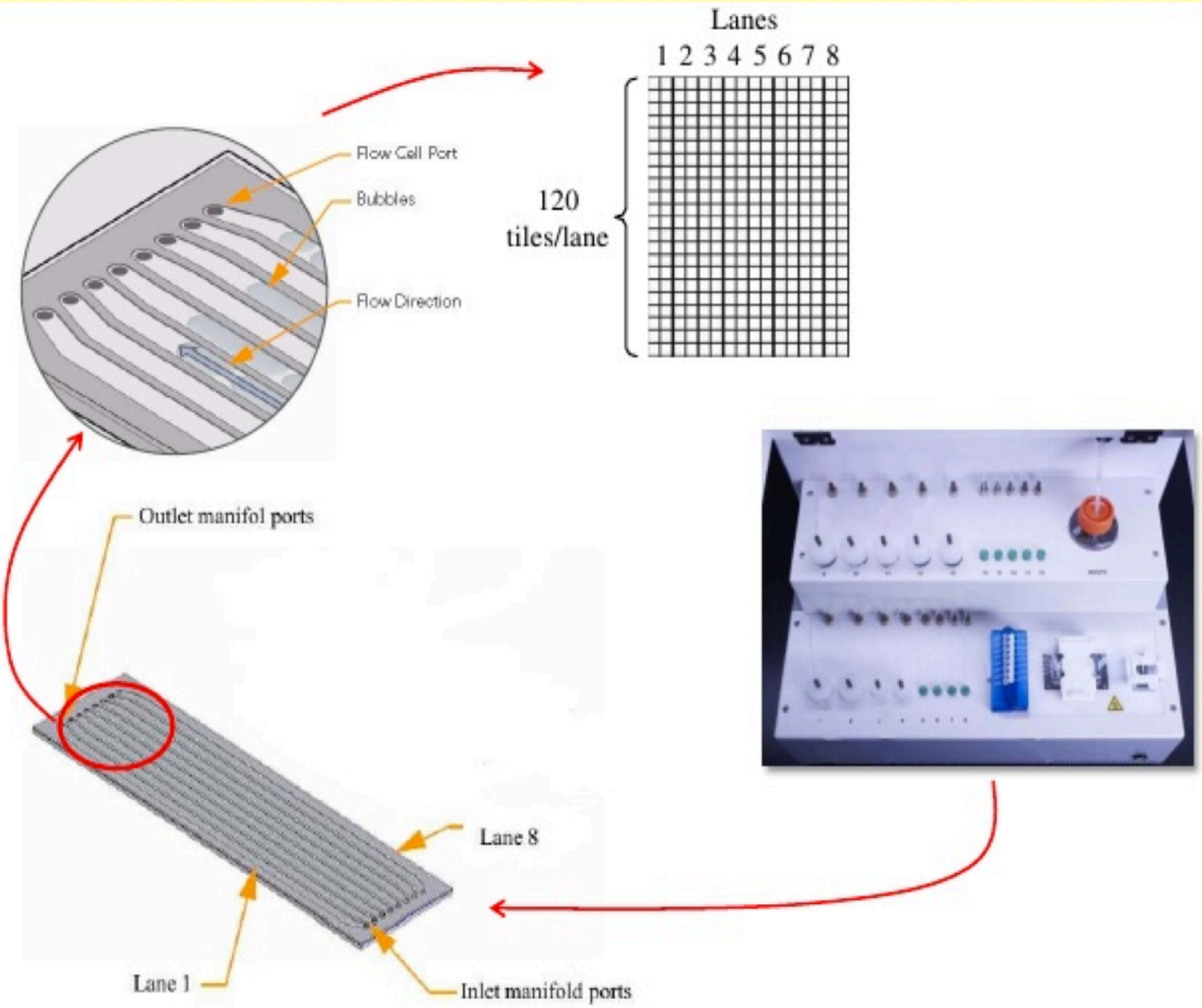
**Downstream analysis**

Sample preparation

**Clusters amplification**

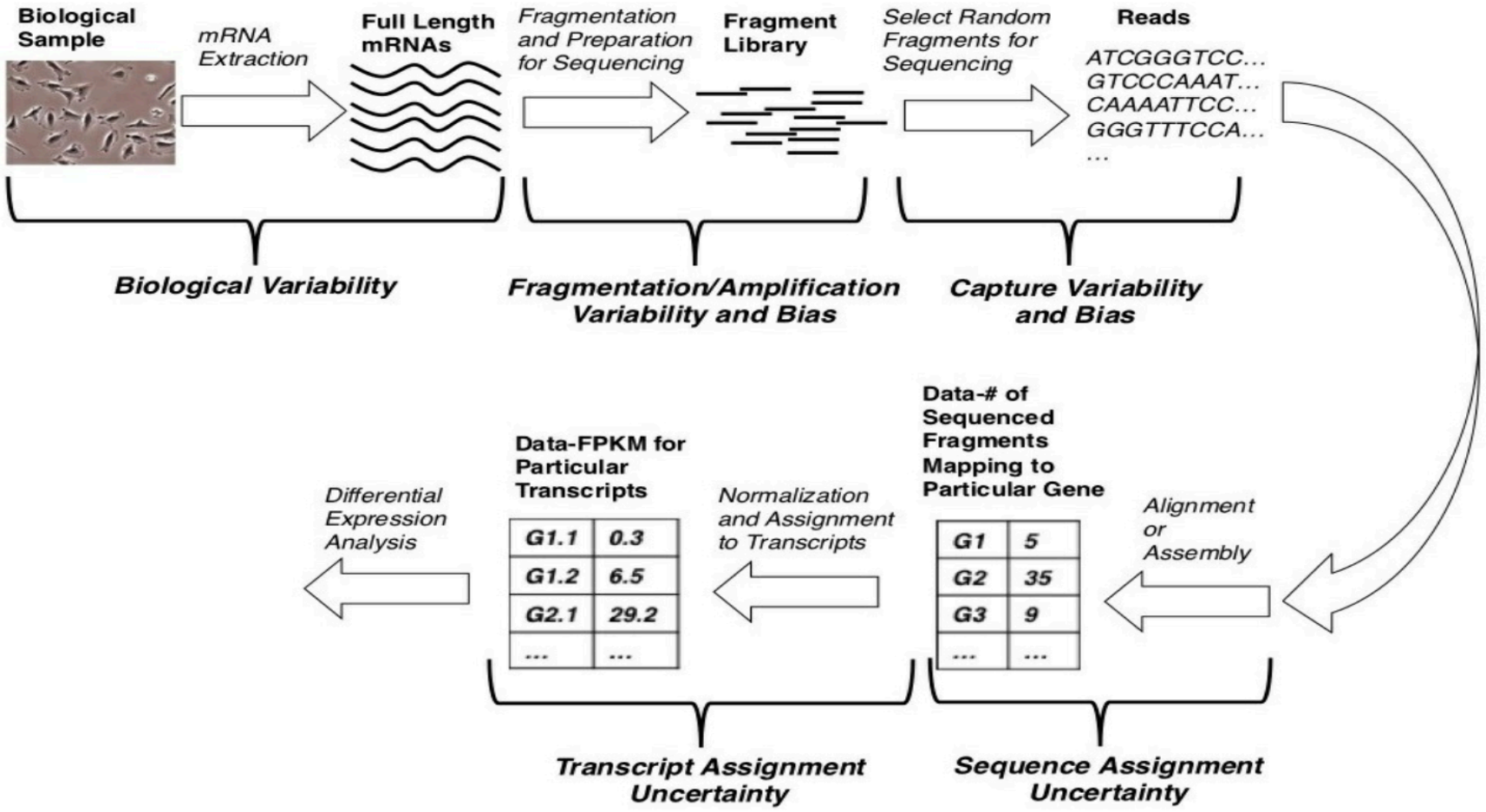
Sequencing by synthesis

Analysis pipeline



# potential sources of bias and

**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.

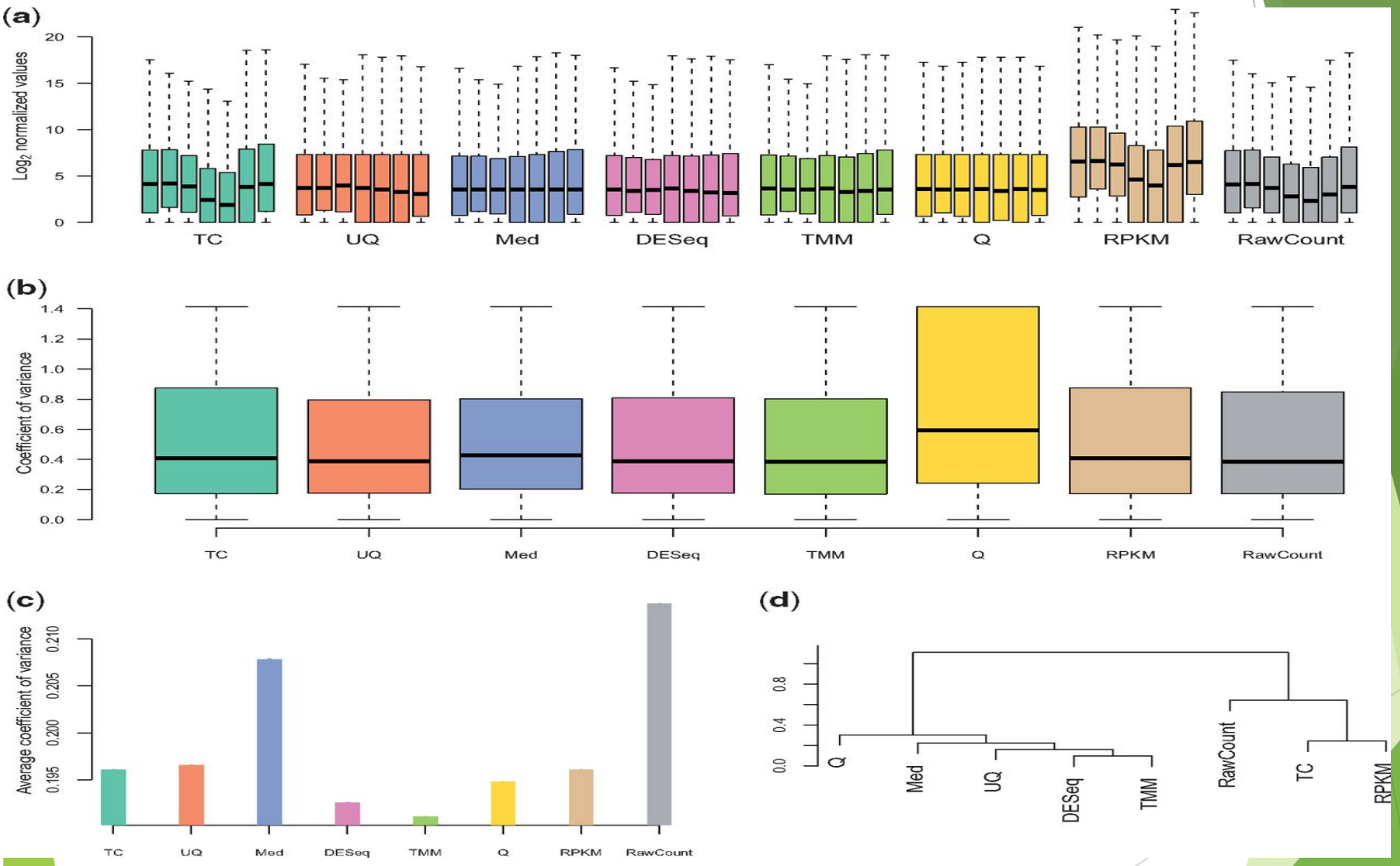


# What is normalization

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

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

# Comparison of normalization methods for real data.

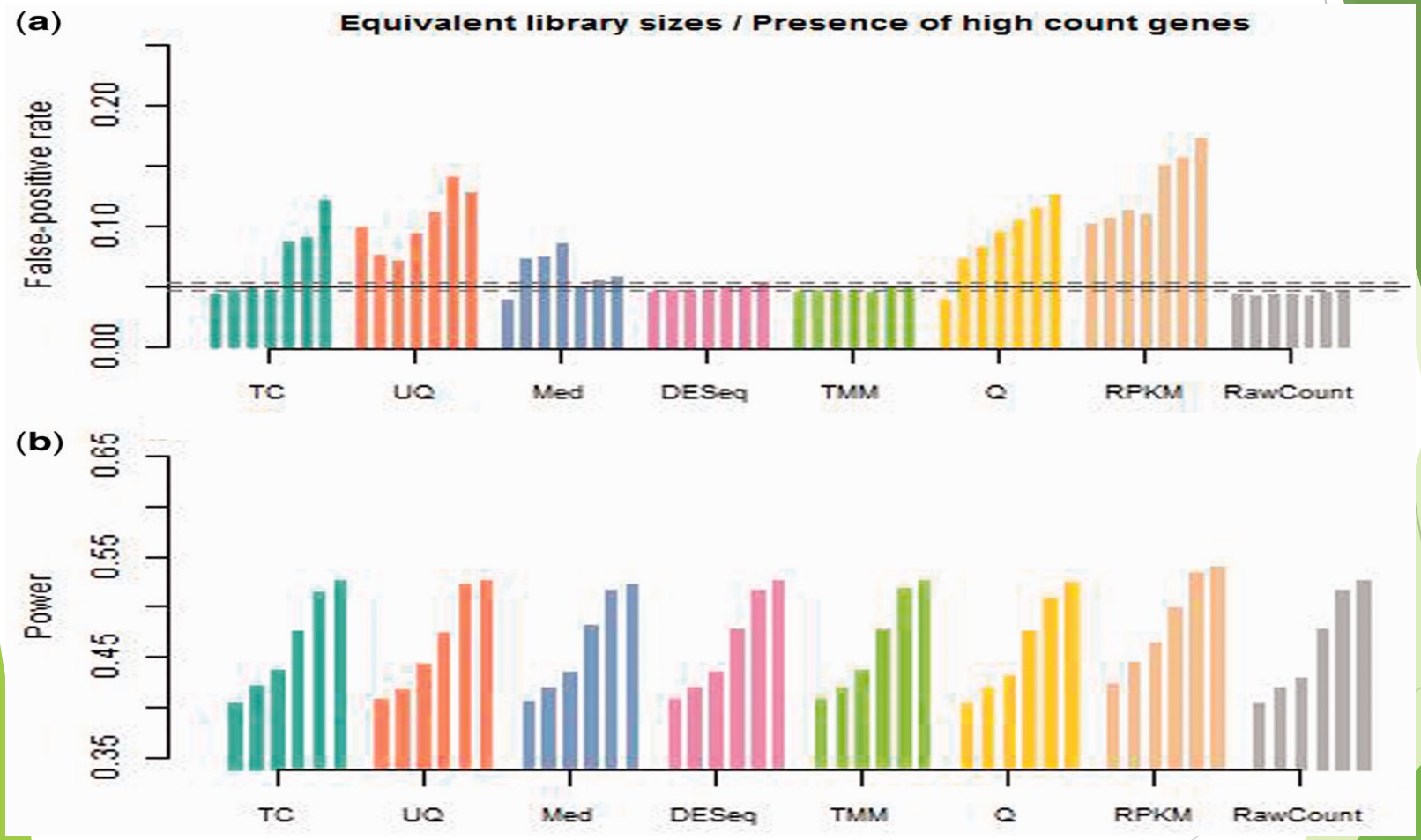



Marie-Agnès Dillies et al. Brief Bioinform 2013;14:671-683



Briefings in Bioinformatics

# Comparison of normalization methods for simulated data with equal library sizes and the presence of high-count genes.




Marie-Agnès Dillies et al. Brief Bioinform 2013;14:671-683



Briefings in Bioinformatics

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).

**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**

***Upregulated***

| | genes | Enriched pathways |
|---|---|---|
| **OLM vs HC+CC** | 34 | MAPK |
| **FC vs HC+CC** | 52 | MAPK |

***Downregulated***

| | genes | Enriched pathways |
|---|---|---|
| **OLM vs HC+CC** | 554 | Ribosome, glycolysis, Lupus |
| **FC vs HC+CC** | 166 | None |

**B** **RUV-normalization**

***Upregulated***

| | genes | Enriched pathways |
|---|---|---|
| **OLM vs HC+CC** | 117 | MAPK, p53, cell cycle, circadian rhythms, cancer |
| **FC vs HC+CC** | 210 | MAPK, T-cell receptor signaling |

***Downregulated***

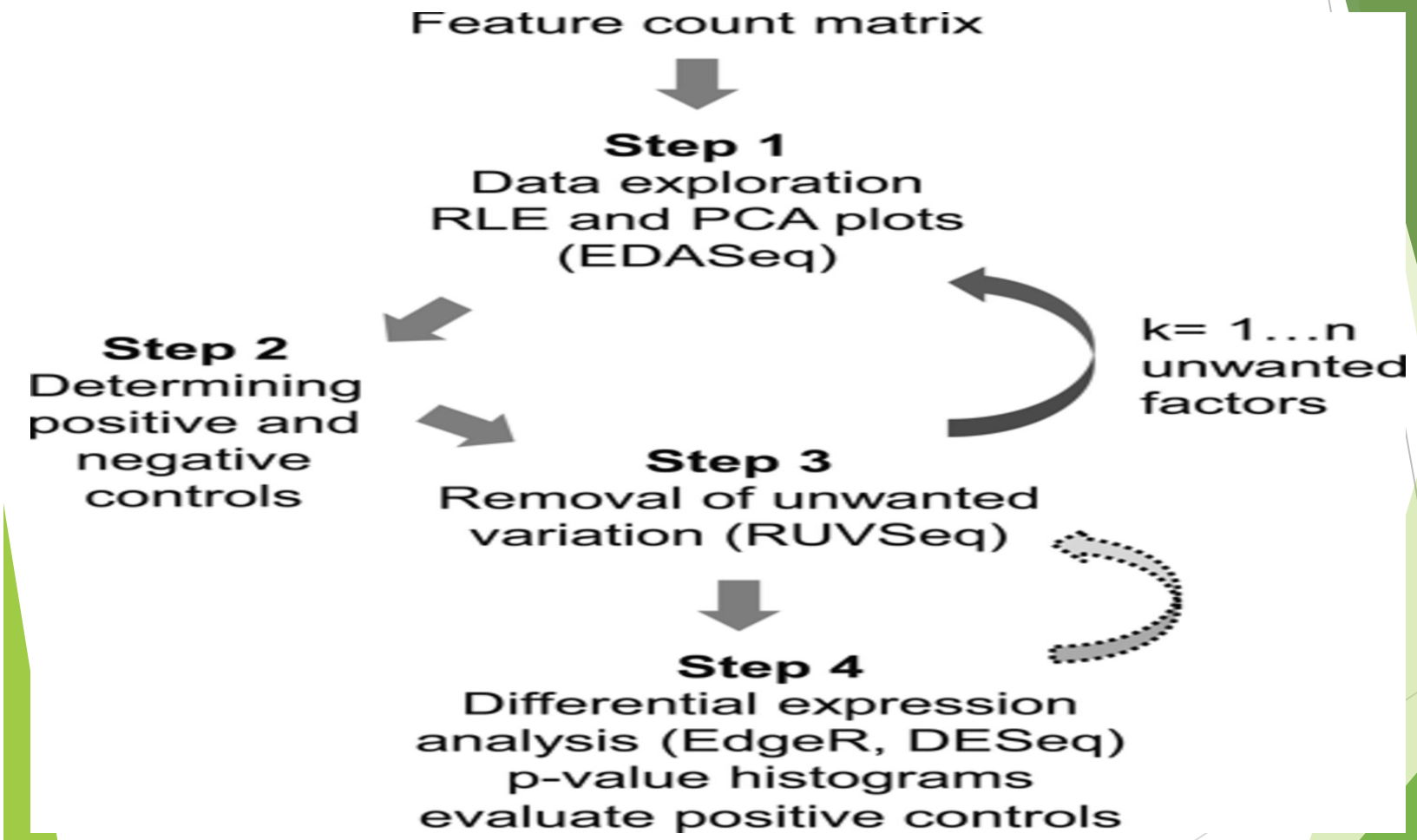
| | genes | Enriched pathways |
|---|---|---|
| **OLM vs HC+CC** | 43 | None |
| **FC vs HC+CC** | 118 | Chemokine signaling, JAK-STAT, Toll-like receptor, RIG-I signaling |

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



Nucleic Acids Research

# Step-by-step outline of the application of RUV to normalization of RNA-seq data.

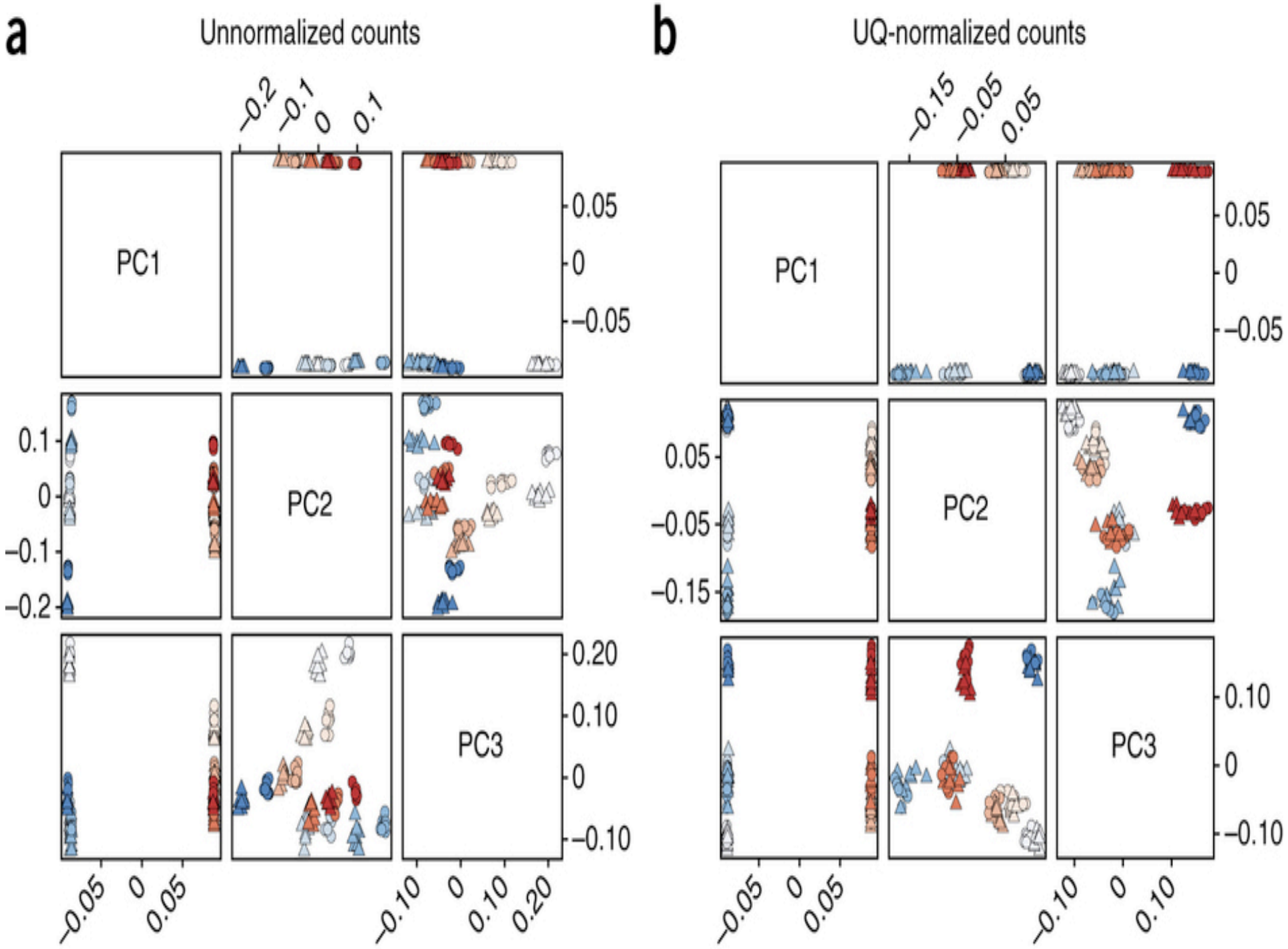


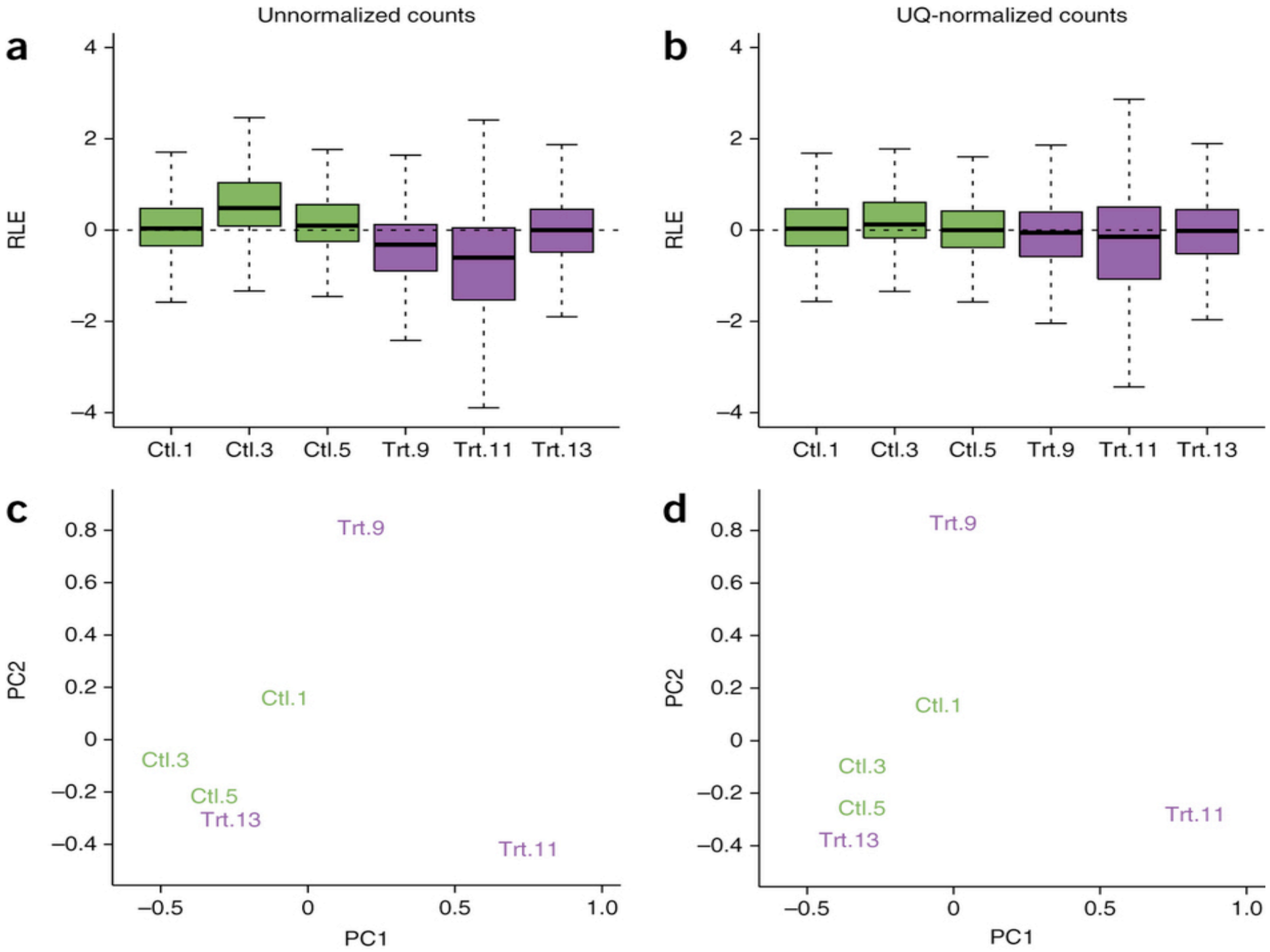

**Lucia Peixoto et al. Nucl. Acids Res. 2015;nar.gkv736**

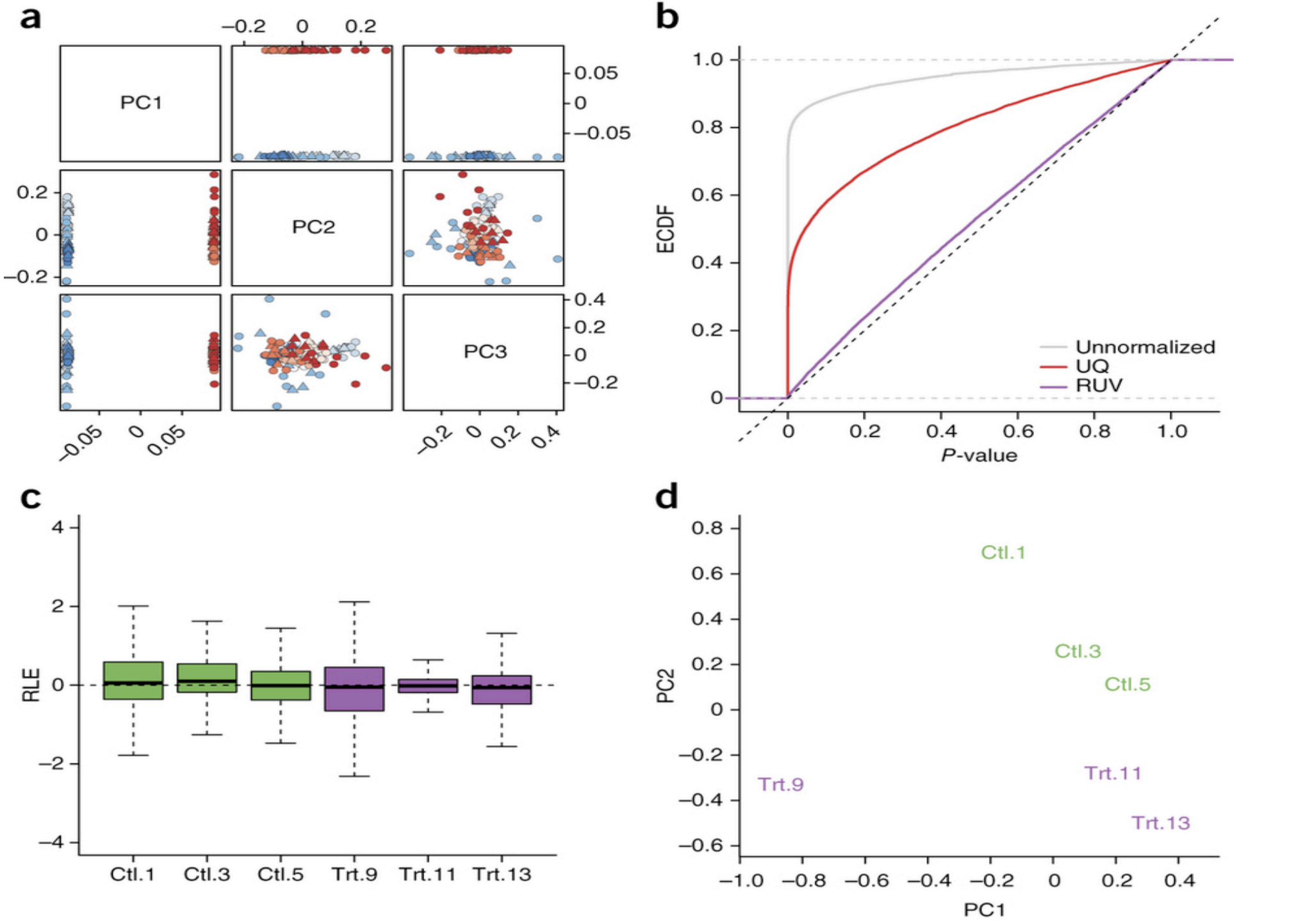


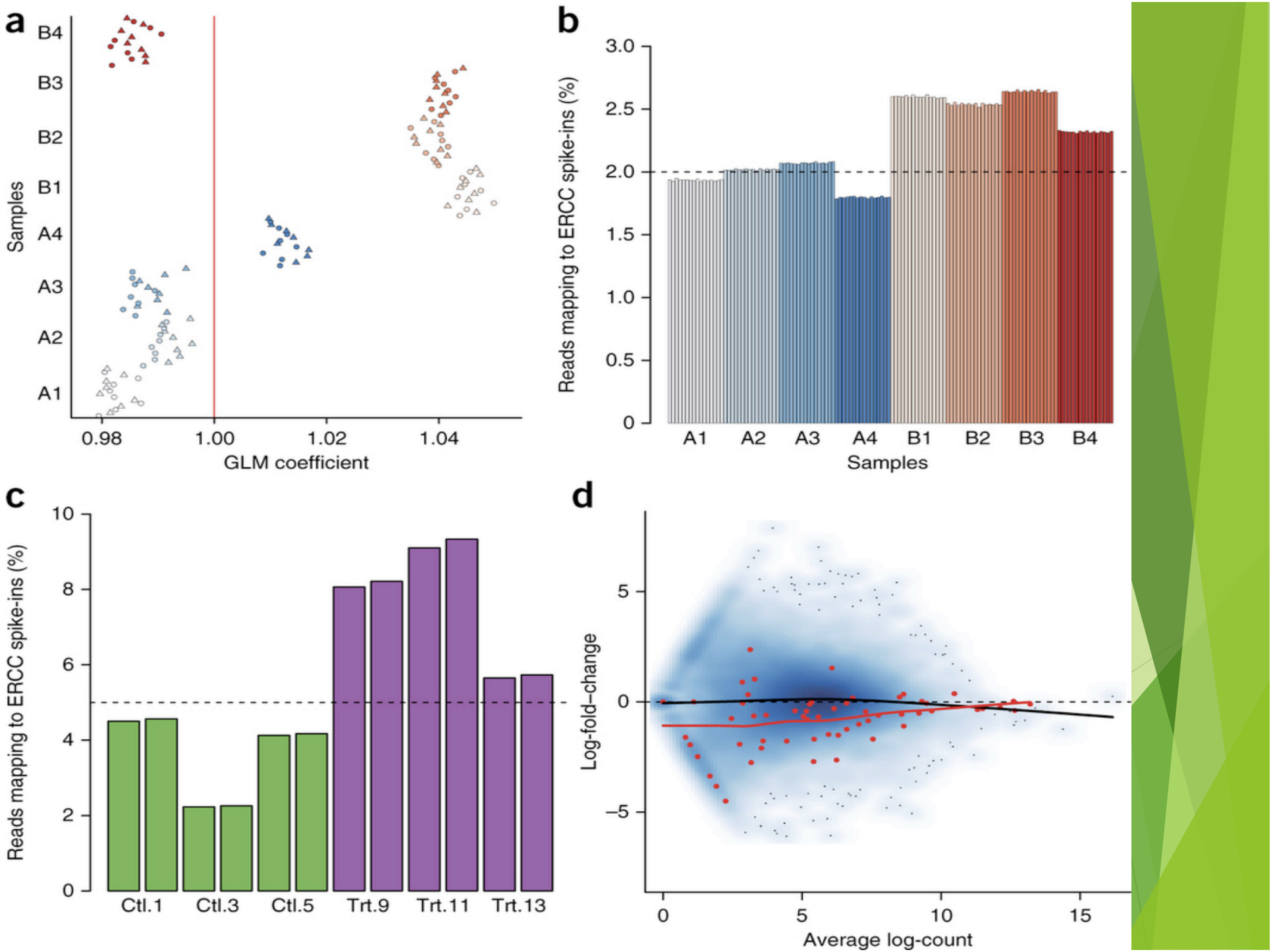
Nucleic Acids Research

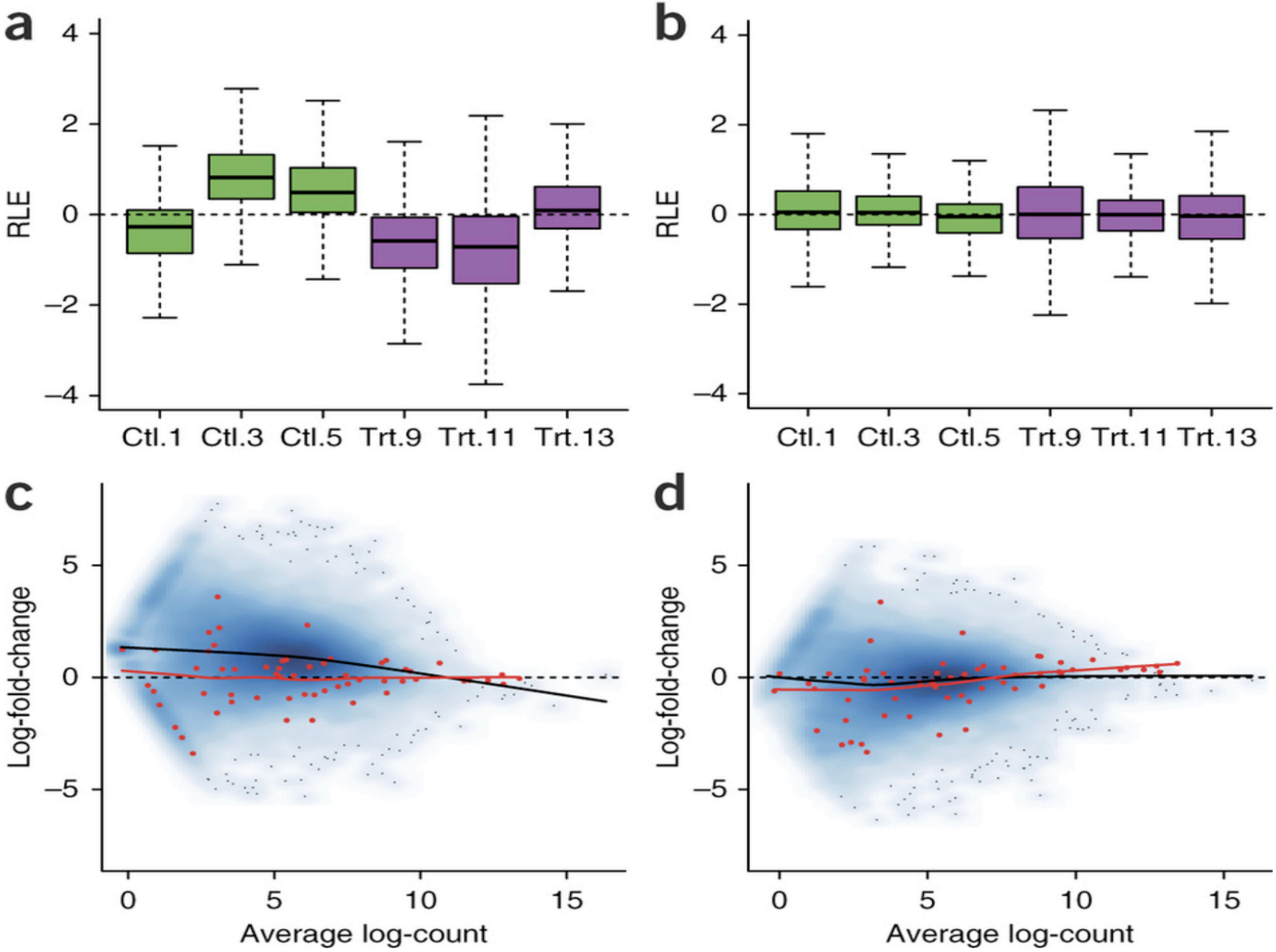
# Using controls for the normalization of RNA-seq data

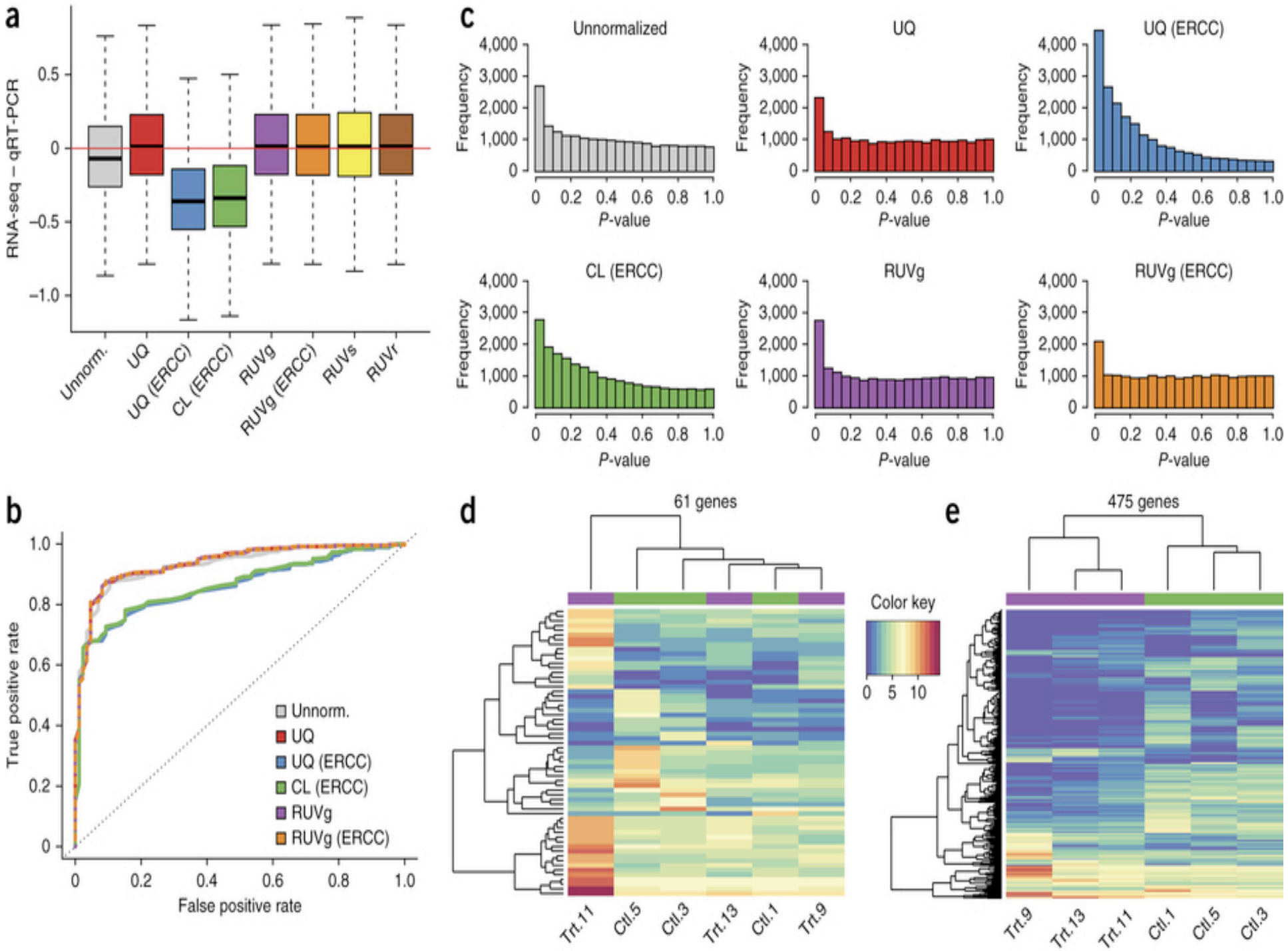

a
Unnormalized counts
PC1
PC2
PC3
b
UQ-normalized counts
PC1
PC2
PC3


a
Unnormalized counts
RLE
4
2
0
−2
−4
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13
b
UQ-normalized counts
RLE
4
2
0
−2
−4
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13
c
PC2
0.8
0.6
0.4
0.2
0
−0.2
−0.4
Trt.9
Ctl.1
Ctl.3
Ctl.5
Trt.13
Trt.11
−0.5
0
0.5
1.0
PC1
d
PC2
0.8
0.6
0.4
0.2
0
−0.2
−0.4
Trt.9
Ctl.1
Ctl.3
Ctl.5
Trt.13
Trt.11
−0.5
0
0.5
1.0
PC1


a
PC1
PC2
PC3
b
ECDF
P-value
Unnormalized
UQ
RUV
c
RLE
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13
d
PC2
PC1
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13


a
B4
B3
B2
B1
A4
A3
A2
A1
Samples
0.98
1.00
1.02
1.04
GLM coefficient
b
Reads mapping to ERCC spike-ins (%)
3.0
2.5
2.0
1.5
1.0
0.5
0
A1
A2
A3
A4
B1
B2
B3
B4
Samples
c
Reads mapping to ERCC spike-ins (%)
10
8
6
4
2
0
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13
d
Log-fold-change
5
0
−5
0
5
10
15
Average log-count


a
b
c
d
RLE
Log-fold-change
Average log-count
Ctl.1
Ctl.3
Ctl.5
Trt.9
Trt.11
Trt.13
4
2
0
−2
−4
5
−5
0
5
10
15


a
RNA-seq − qRT-PCR
0.5
0
−0.5
−1.0
Unnorm.
UQ
UQ (ERCC)
CL (ERCC)
RUVg
RUVg (ERCC)
RUVs
RUVr
b
True positive rate
False positive rate
0
0.2
0.4
0.6
0.8
1.0
Unnorm.
UQ
UQ (ERCC)
CL (ERCC)
RUVg
RUVg (ERCC)
c
Unnormalized
UQ
UQ (ERCC)
CL (ERCC)
RUVg
RUVg (ERCC)
Frequency
P-value
4,000
3,000
2,000
1,000
0
d
61 genes
Color key
0 5 10
Trt.11
Ctl.5
Ctl.3
Trt.13
Ctl.1
Trt.9
e
475 genes
Trt.9
Trt.13
Trt.11
Ctl.1
Ctl.5
Ctl.3

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 form GitHub (github.com/drisso/peixoto2015_tutorial).

# 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 (quantile/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] | RPKM/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] | Geometric (DESeq-like)/quartile /classic-fpkm | 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 |

[a]In case of availability of several normalization methods, the default one is underlined.

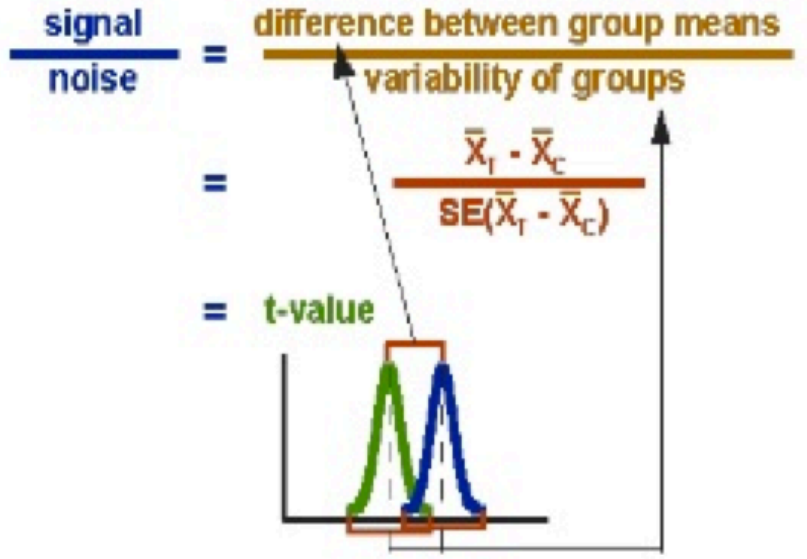
# Count-based statistics

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.

# 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

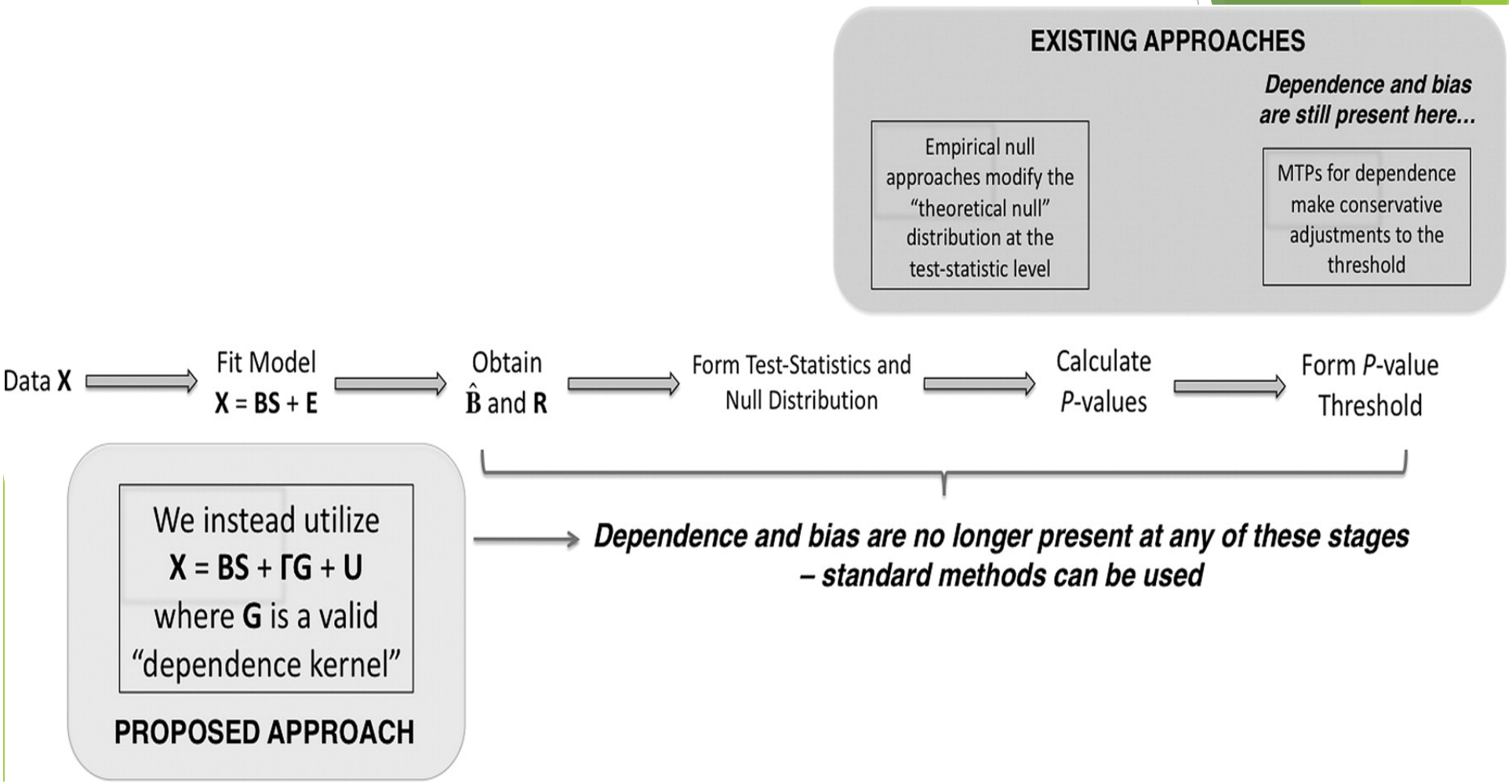
Problems with this approach:

- May have **few replicates**

- Distribution is **not normal**

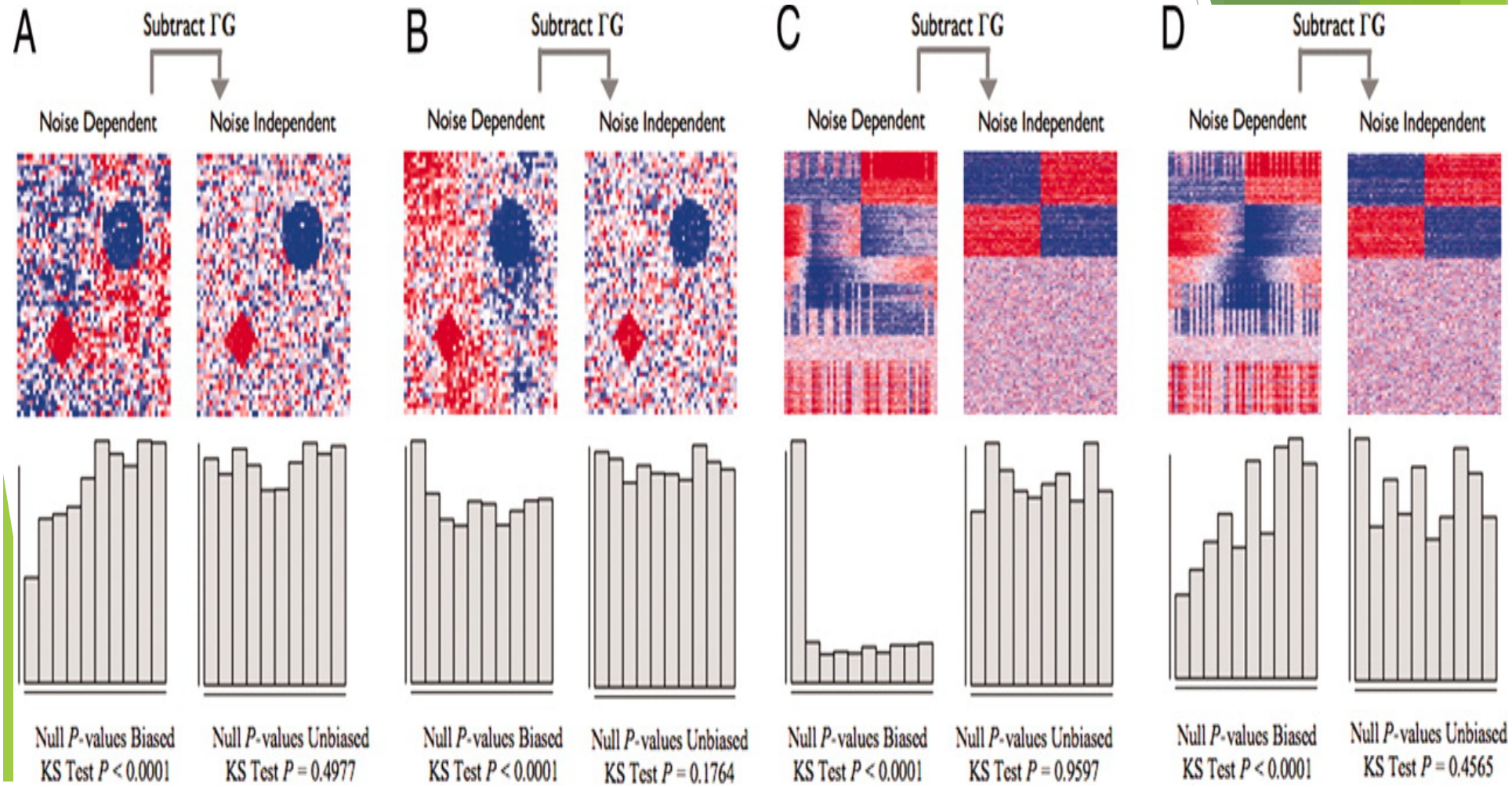
- **Multiple testing** issues

# A schematic of the general steps of multiple hypothesis testing.

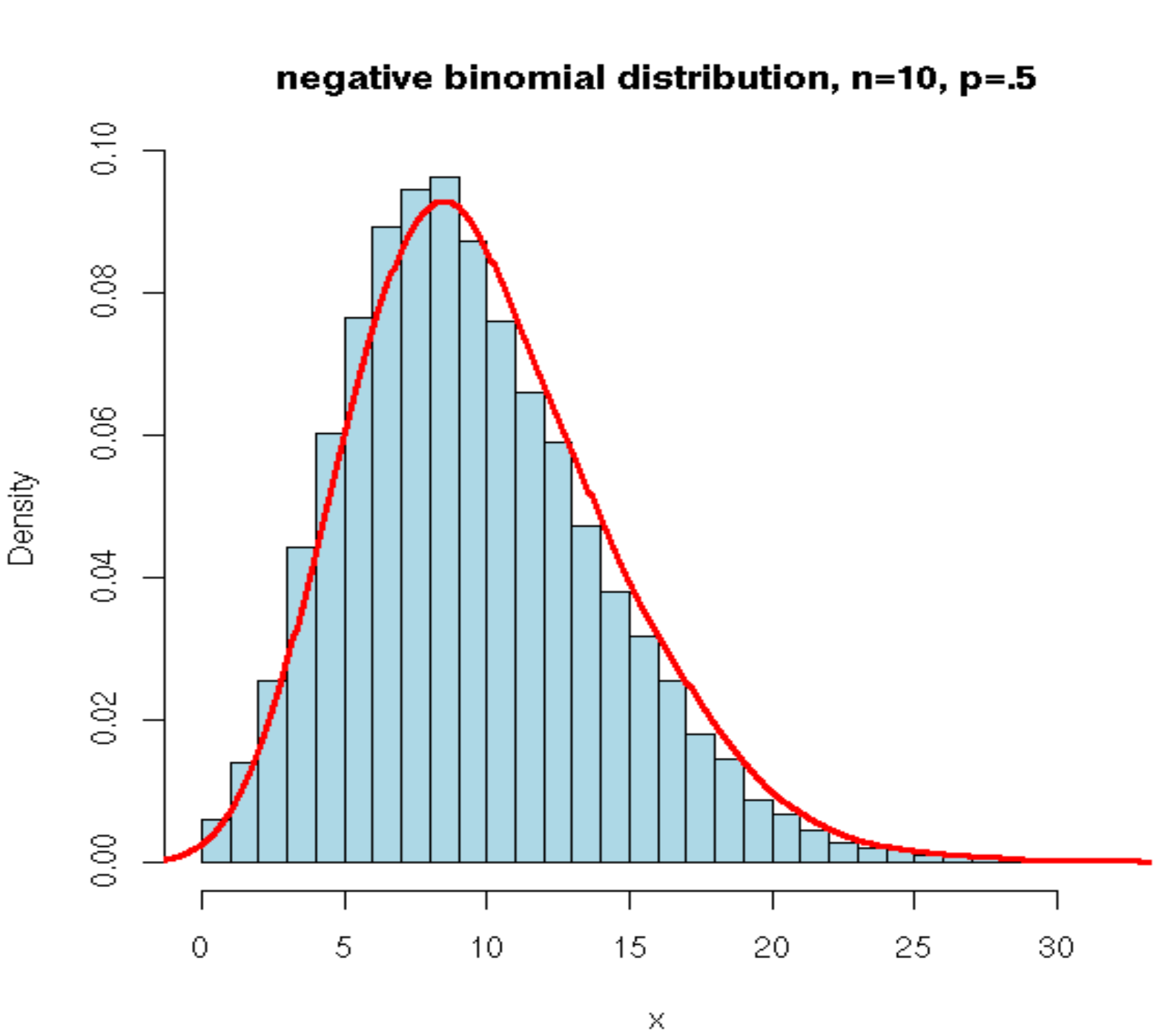



**Jeffrey T. Leek, and John D. Storey PNAS 2008;105:18718-18723**

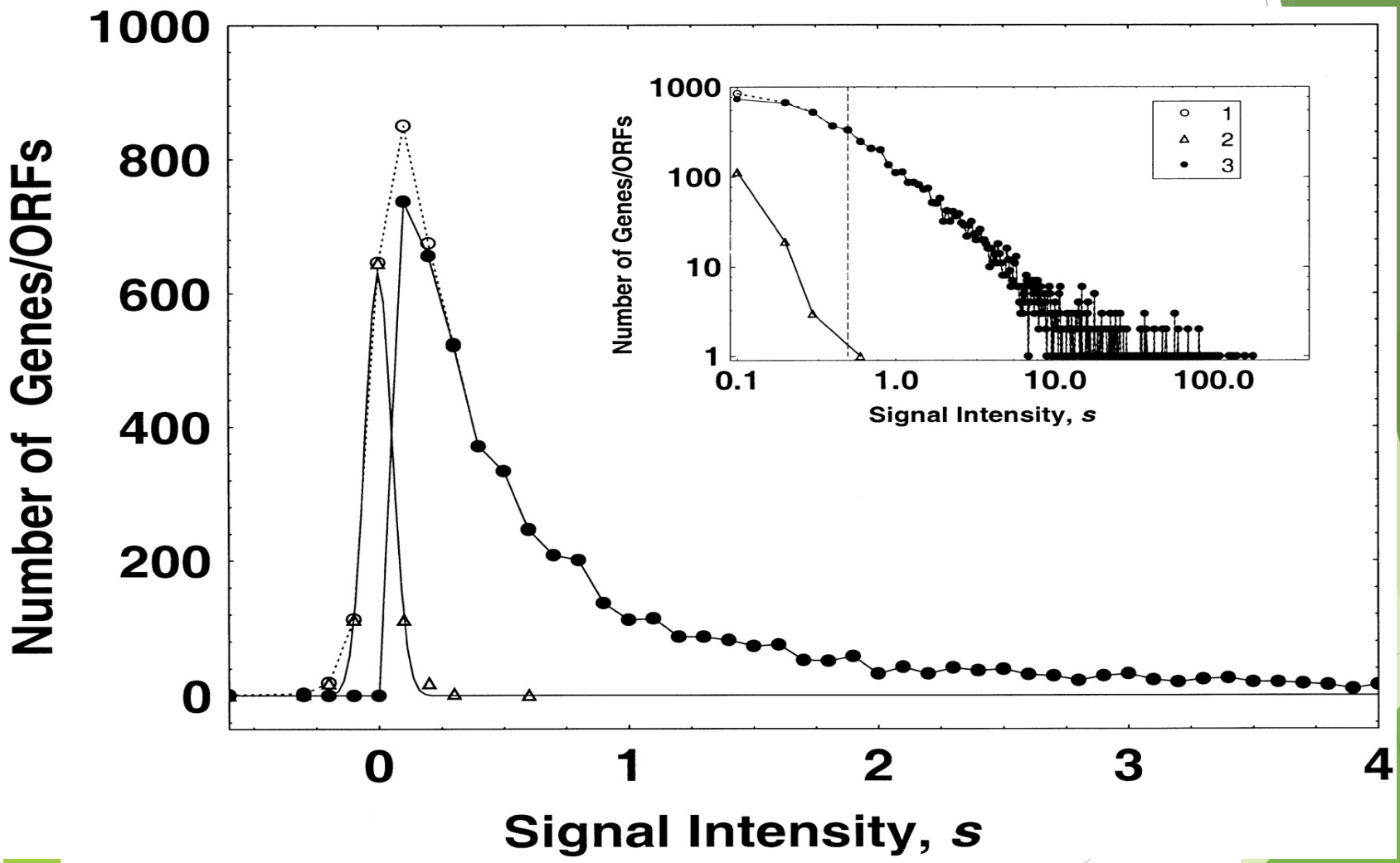
# Simulated examples of multiple testing dependence.




**Jeffrey T. Leek, and John D. Storey PNAS 2008;105:18718-18723**


negative binomial distribution, n=10, p=.5
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
x

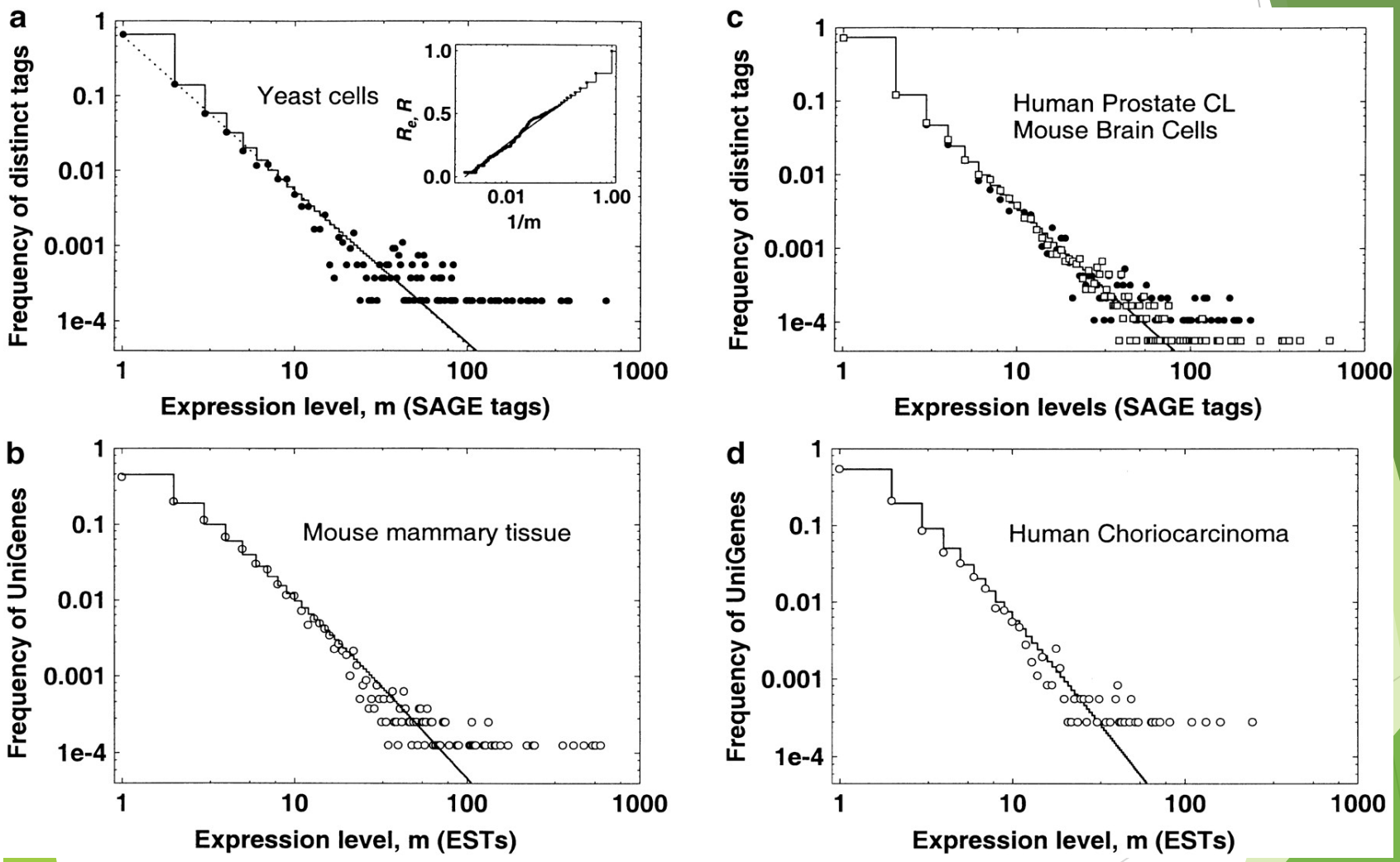
**The empirical frequency distribution of the hybridization signal intensity values for Affymetrix microarray hybridization data for normal yeast cell genes/ORFs (Jelinsky and Samson 1999).**




Kuznetsov V A et al. Genetics 2002;161:1321-1332

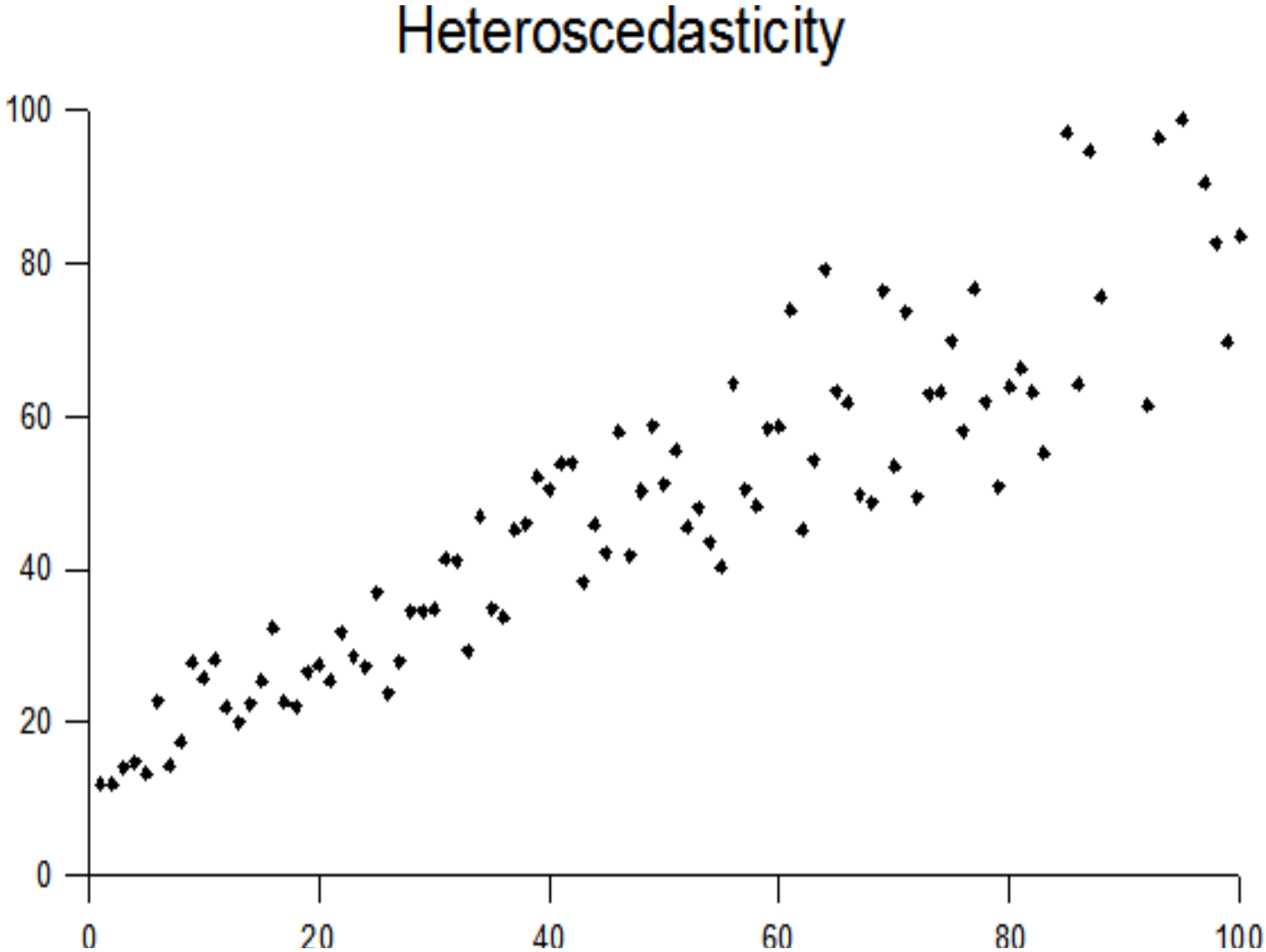
GENETICS

# Empirical relative frequency distributions of the gene expression levels.




Kuznetsov V A et al. Genetics 2002;161:1321-1332

GENETICS


Heteroscedasticity
100
80
60
40
20
0
0
20
40
60
80
100

# 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 at purdue.edu>, Wolfgang Huber <whuber at embl.de> and Olga Vitek <ovitek at purdue.edu>

Maintainer: Danni Yu <dyu at 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.

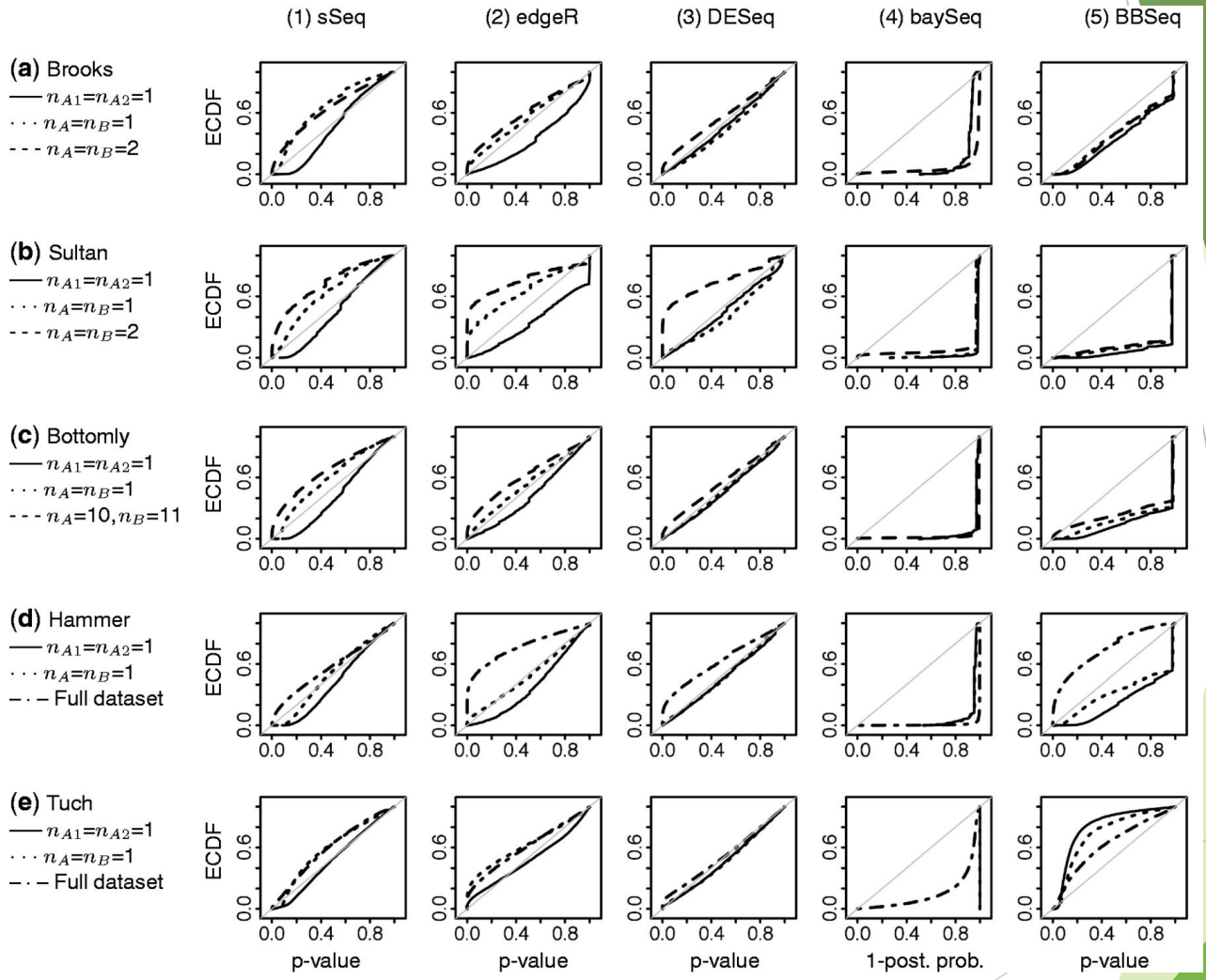
# Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size

**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.* |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $n_i = 1$ | $n_i = 2$ | $n_i = 1$ | $n_i = 2$ | $n_i = 1$ | $n_i = 2$ | $n_i = 1$ | $n_A = 3, n_B = 2$ | $n_i = 1$ |
| 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.

# The ECDF curves of detecting differential expression for the datasets with no external ‘gold standard’.




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



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:



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 Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, Geoffrey J. Barton

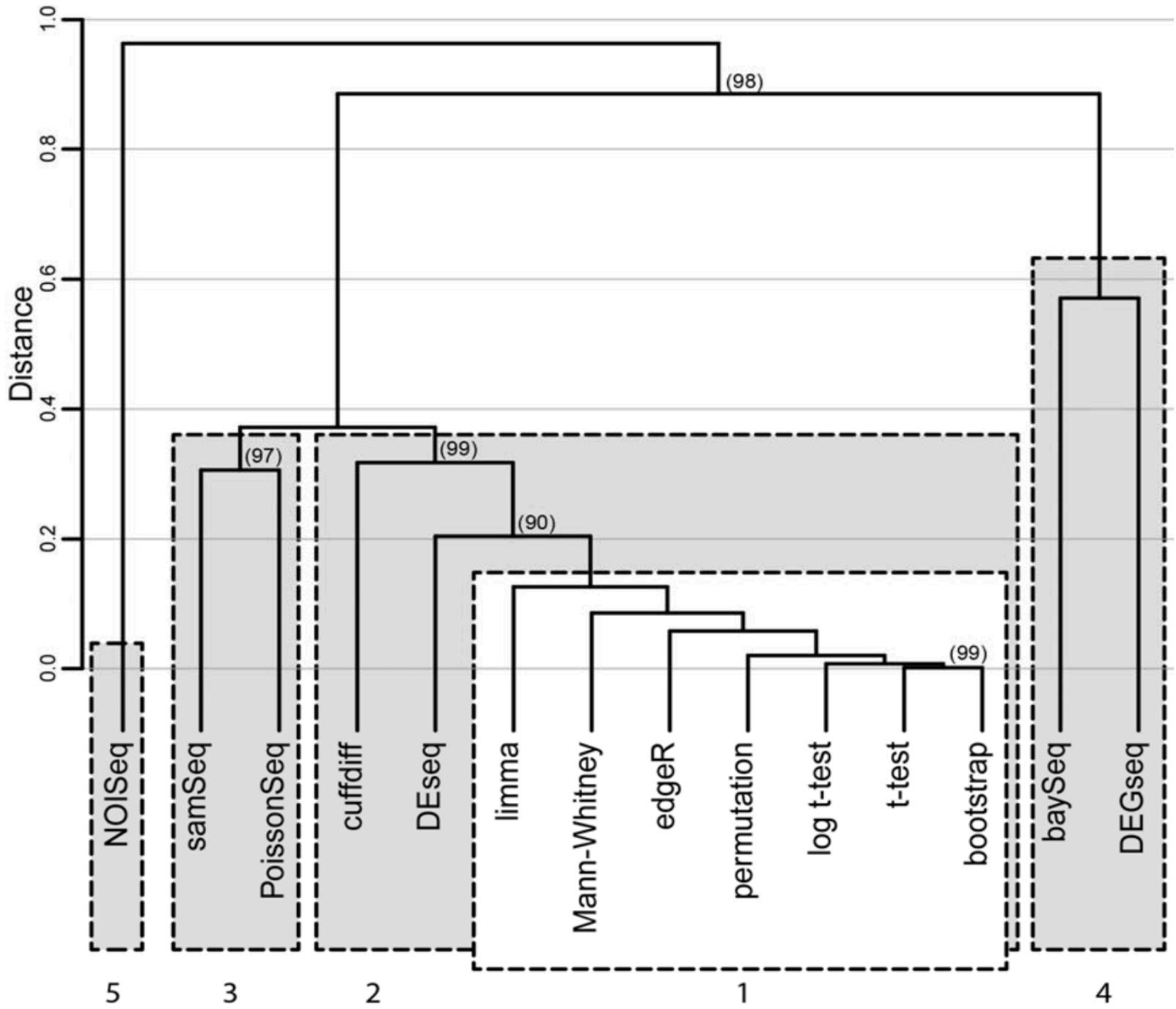
*(Submitted on 8 May 2015 (v1), last revised 8 Jun 2015 (this version, v2))*

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)

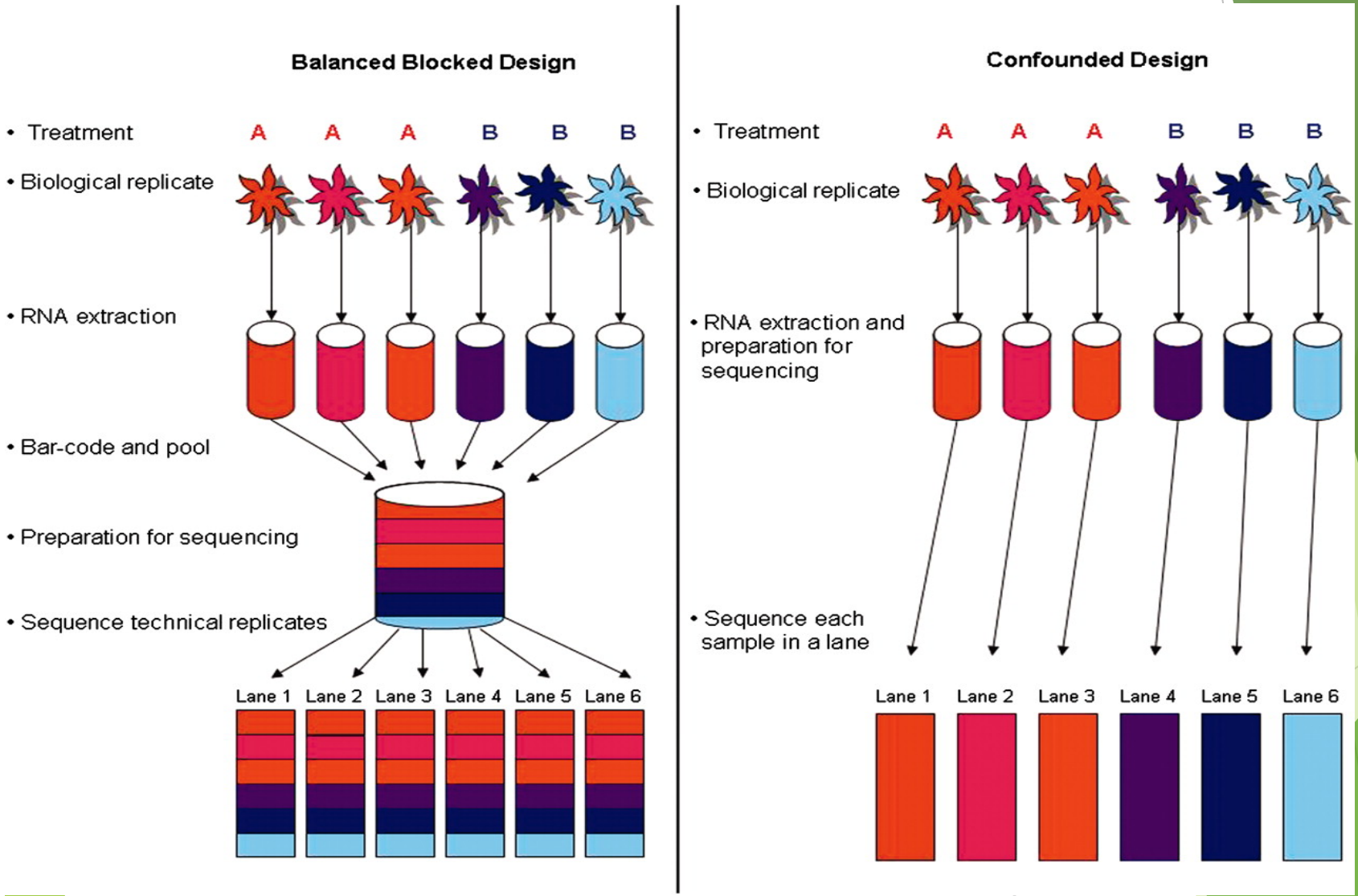
**Table 1:** RNA-seq differential gene expression tools and statistical tests.

| Name | Assumed Distribution | Normalization | Description | Version | Citations[4] | Reference |
|---|---|---|---|---|---|---|
| *t-test* | normal | DEseq[1] | two-sample t-test for equal variances | - | - | - |
| *log t-test* | log-normal | DEseq[1] | log-ratio t-test | - | - | - |
| *Mann-Whitney* | none | DEseq[1] | Mann-Whitney test | - | - | Mann and Whitney (1947) |
| *Permutation* | none | DEseq[1] | permutation test | - | - | Efron and Tibshirani (1993) |
| *Bootstrap* | normal | DEseq[1] | bootstrap test | - | - | Efron and Tibshirani (1993) |
| *baySeq*[3] | negative binomial | Internal | Empirical Bayesian estimate of posterior likelihood | 1.8 | 109 | Hardcastle and Kelly (2010) |
| *Cuffdiff* | negative binomial | Internal | unknown | 2.1.1 | 481 | Trapnell et al. (2012) |
| *DEGseq*[3] | binomial | None | random sampling model using Fisher's exact test and the likelihood ratio test | 1.10.0 | 215 | Wang et al. (2010) |
| *DESeq*[3] | negative binomial | DEseq[1] | Shrinkage variance | 1.4.1 | 1204 | Anders and Huber (2010) |
| *edgeR*[3] | negative binomial | TMM[2] | 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) |
| *Limma*[3] | Log-normal | TMM[2] | Generalised linear model | 3.4.4 | 15 | Law et al. (2014) |
| *NOISeq*[3] | None | RPKM | Non-parametric test based on signal-to-noise ratio | 29/04/2011 | 113 | Tarazona et al. (2011) |
| *PoissonSeq*[3] | Poisson log-linear model | Internal | Score statistic | 1.1 | 25 | Li et al. (2012) |
| *SAMSeq*[3] | None | Internal | Mann-Whitney test with Poisson resampling | 2.0 | 26 | Li and Tibshirani (2013) |


Distance
1.0
0.8
0.6
0.4
0.2
0.0
(98)
(97)
(99)
(90)
(99)
NOISeq
samSeq
PoissonSeq
cuffdiff
DEseq
limma
Mann-Whitney
edgeR
permutation
log t-test
t-test
bootstrap
baySeq
DEGseq
5
3
2
1
4

| | 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 |
| *BaySeq* | inconsistent | Pass | | | | |
| *cuffdiff* | consistent | Fail | | | | |
| *DEGSeq* | inconsistent | Fail | | | | |
| ***DESeq*** | consistent | Pass | 0 | | | Yes |
| | | | 0.5 | | Yes | Yes |
| | | | 2 | Yes | Yes | Yes |
| ***edgeR*** | consistent | Pass | 0 | | | Yes |
| | | | 0.5 | Yes | Yes | Yes |
| | | | 2 | Yes | Yes | Yes |
| ***Limma*** | consistent | Pass | 0 | | | Yes |
| | | | 0.5 | | Yes | Yes |
| | | | 2 | Yes | Yes | Yes |
| *NOISeq* | inconsistent | Pass | | | | |
| *PoissonSeq* | inconsistent | Fail | | | | |
| *SAMSeq* | inconsistent | Fail | | | | |

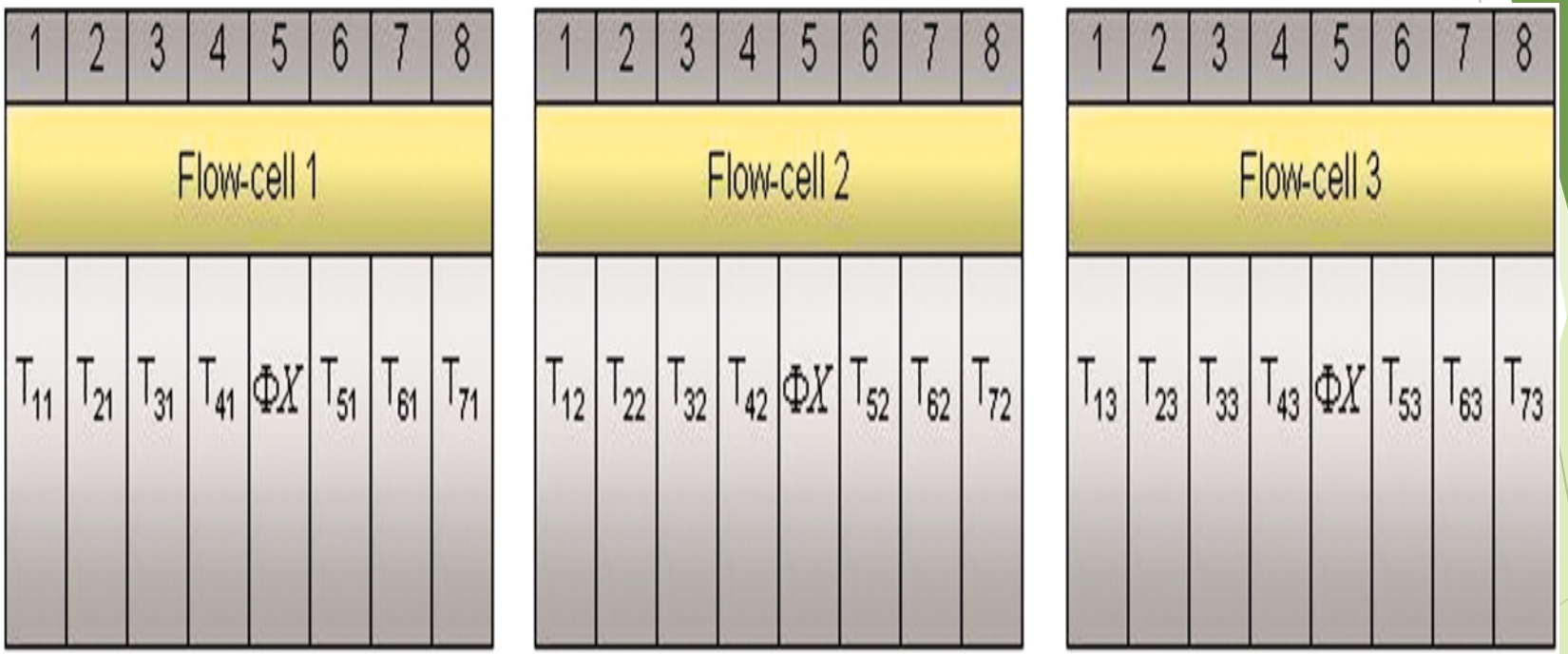
# Comparison of two designs for testing differential expression between treatments A and B. Treatment A is denoted by red tones and treatment B by blue tones.




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

# 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 ....

## 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*.



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).

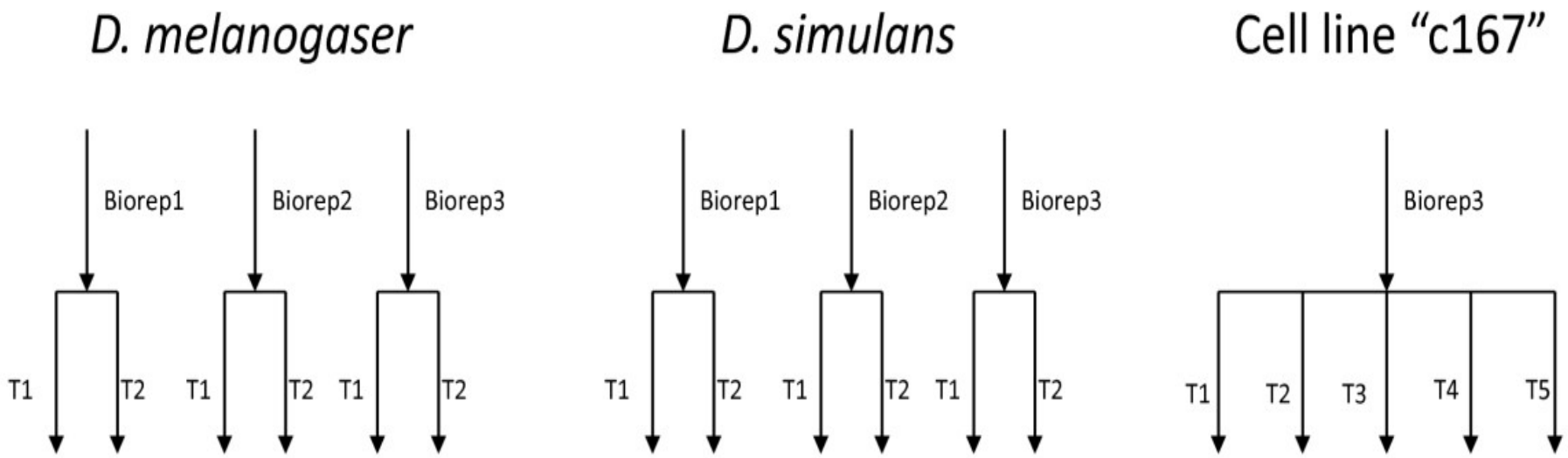
McIntyre *et al. BMC Genomics* 2011, **12**:293
http://www.biomedcentral.com/1471-2164/12/293

BMC Genomics

RESEARCH ARTICLE Open Access

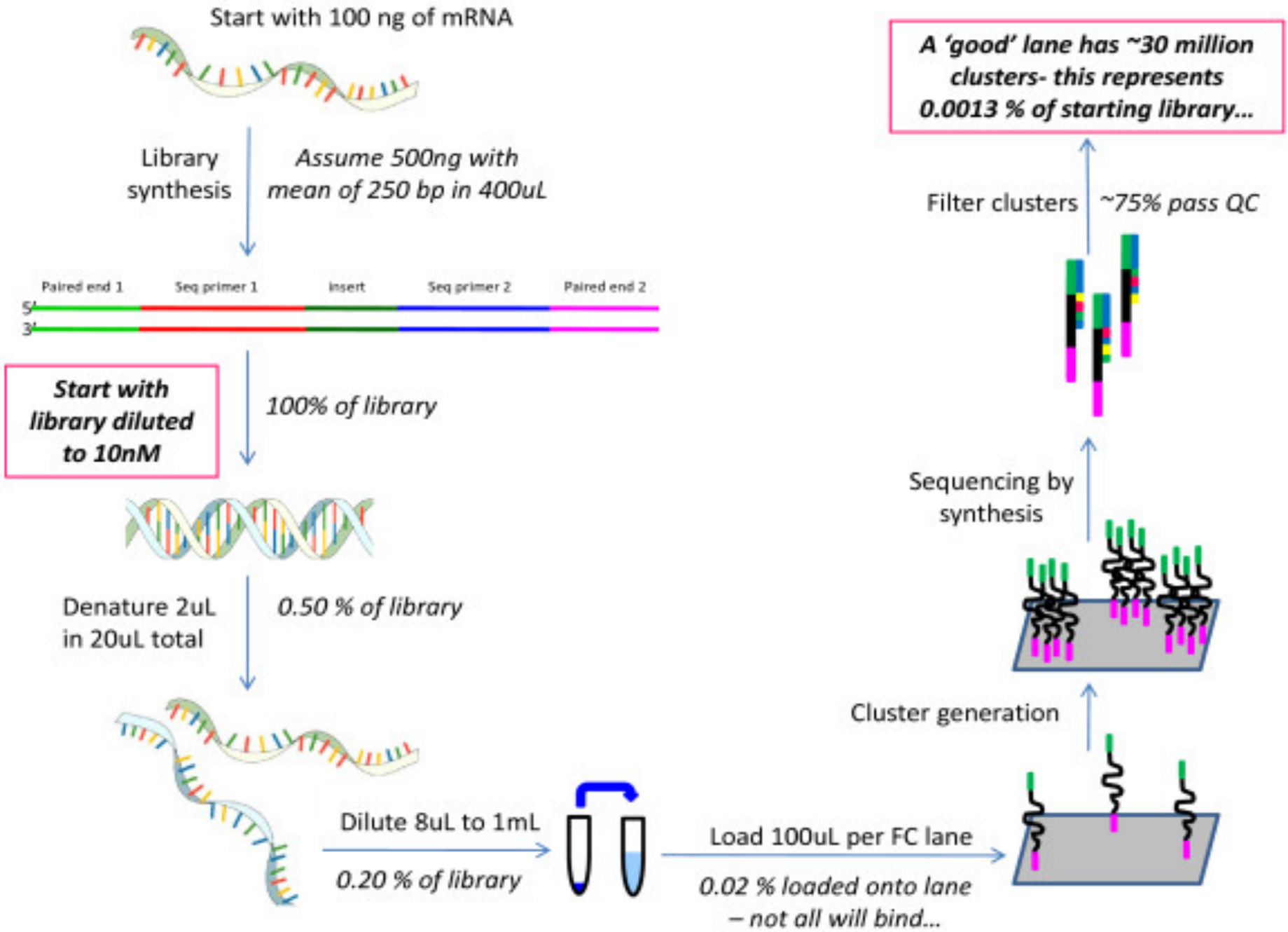
# RNA-seq: technical variability and sampling

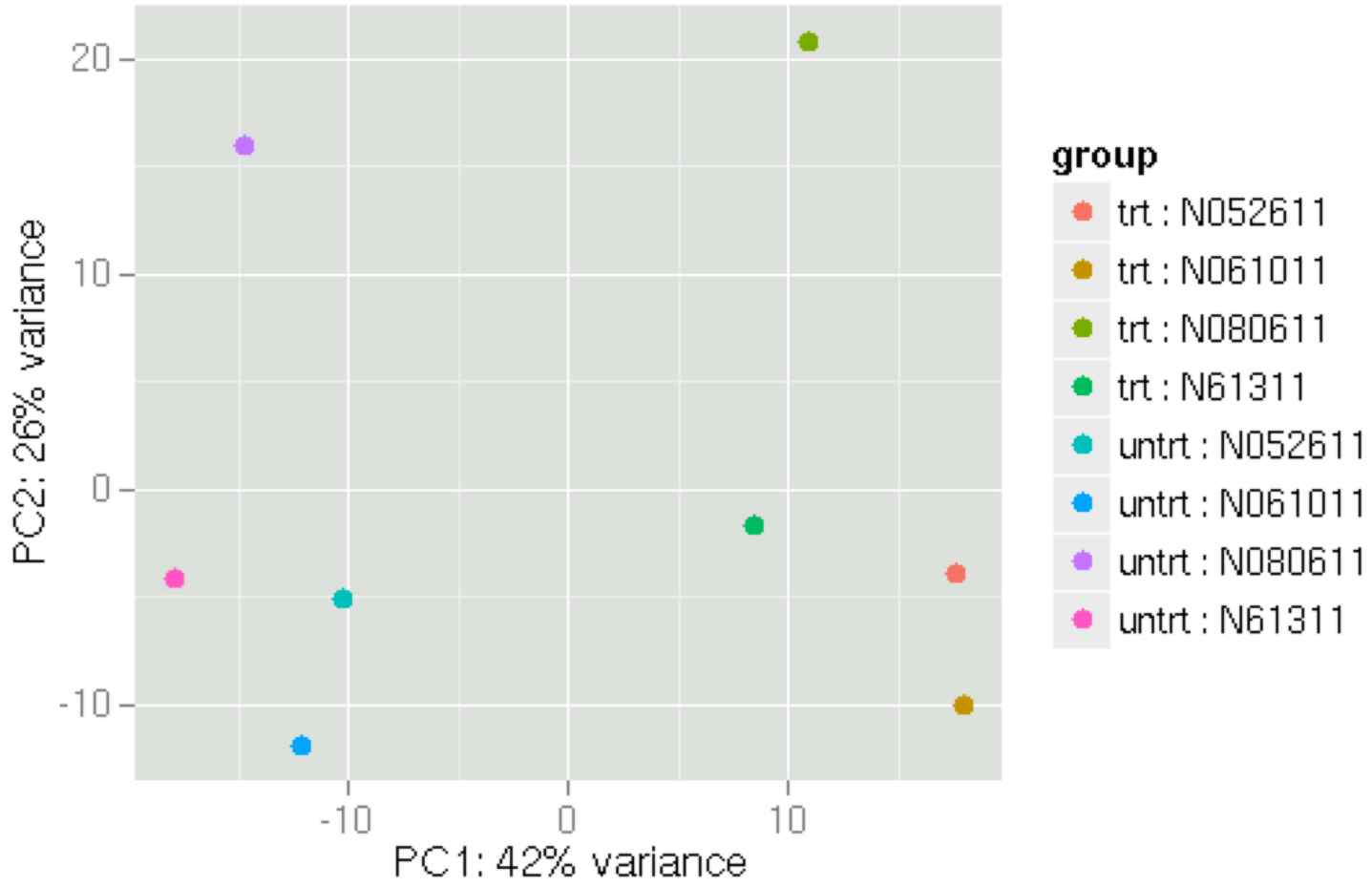
Lauren M McIntyre[1*], Kenneth K Lopiano[2], Alison M Morse[1], Victor Amin[1], Ann L Oberg[3], Linda J Young[2] and Sergey V Nuzhdin[4]



# A “good” lane has ~30 million clusters. This represents 0.0013 % of the starting library

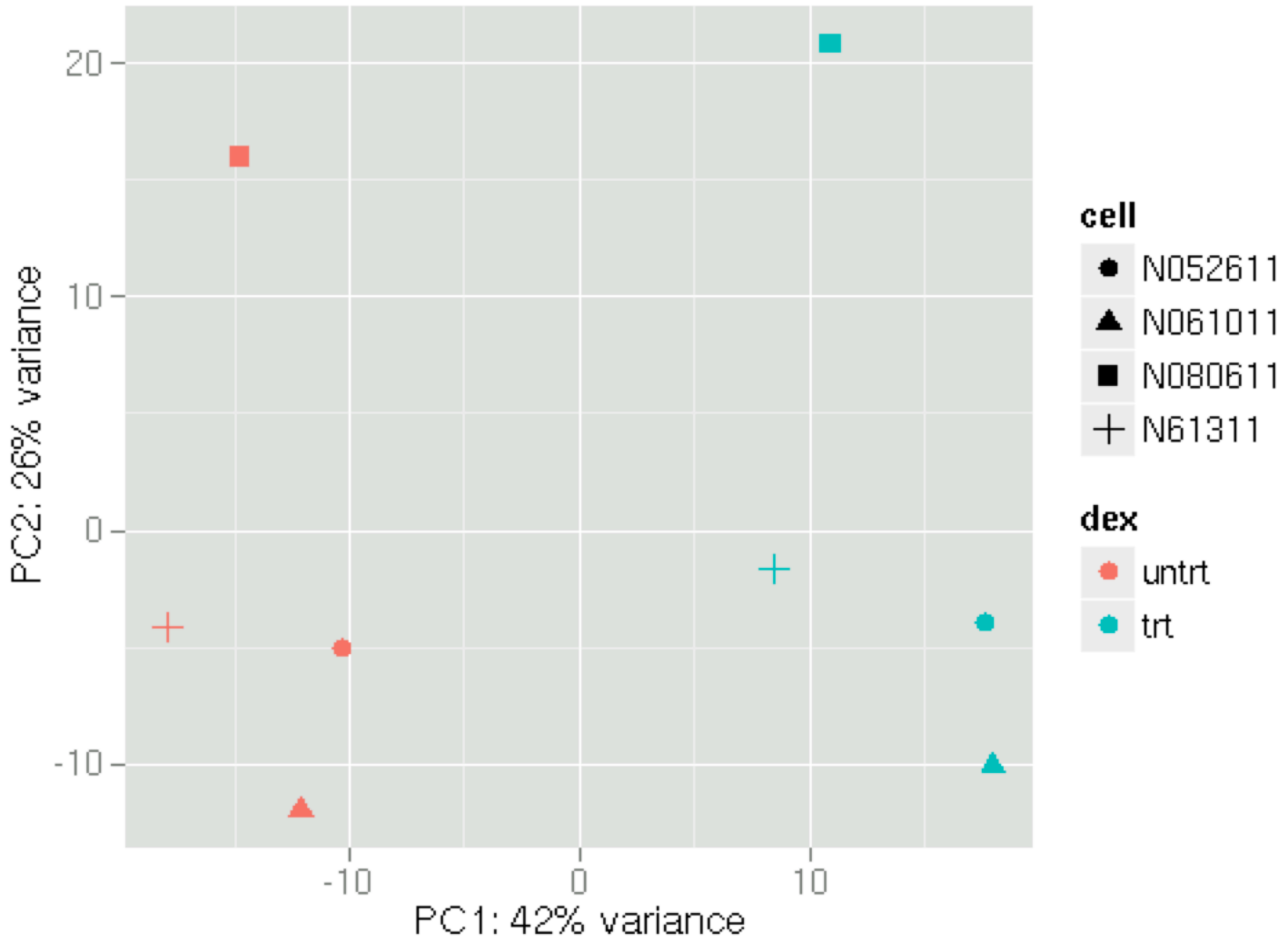
- The number of molecules present in the library is estimated to be $2.408 \times 10^{12}$.
- This estimate was derived using the Illumina protocol. Specifically, the starting material was assumed to be 100 ng of mRNA that then resulted in 500 ng of library with an insert size of 250bp in a volume of 400 uL for a final library concentration of 10 nM.
- The number of pmoles is therefore 4 [(10 nmol/1L) (1L/1000 mL)(1 mL/1000 uL)(400 uL)(1000 pmol/1 nmol)].
- The number of molecules is found by multiplying Avogadro's number which is molecules/moles and adjusting for units [4 pmole)/1000/1000/1000/1000*6.02E + 023] to give $2.408 \times 10^{12}$ 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).


Start with 100 ng of mRNA
Library synthesis
Assume 500ng with mean of 250 bp in 400uL
Paired end 1
Seq primer 1
Insert
Seq primer 2
Paired end 2
5'
3'
Start with library diluted to 10nM
100% of library
Denature 2uL in 20uL total
0.50 % of library
Dilute 8uL to 1mL
0.20 % of library
Load 100uL per FC lane
0.02 % loaded onto lane – not all will bind…
Cluster generation
Sequencing by synthesis
Filter clusters
~75% pass QC
A 'good' lane has ~30 million clusters- this represents 0.0013 % of starting library…



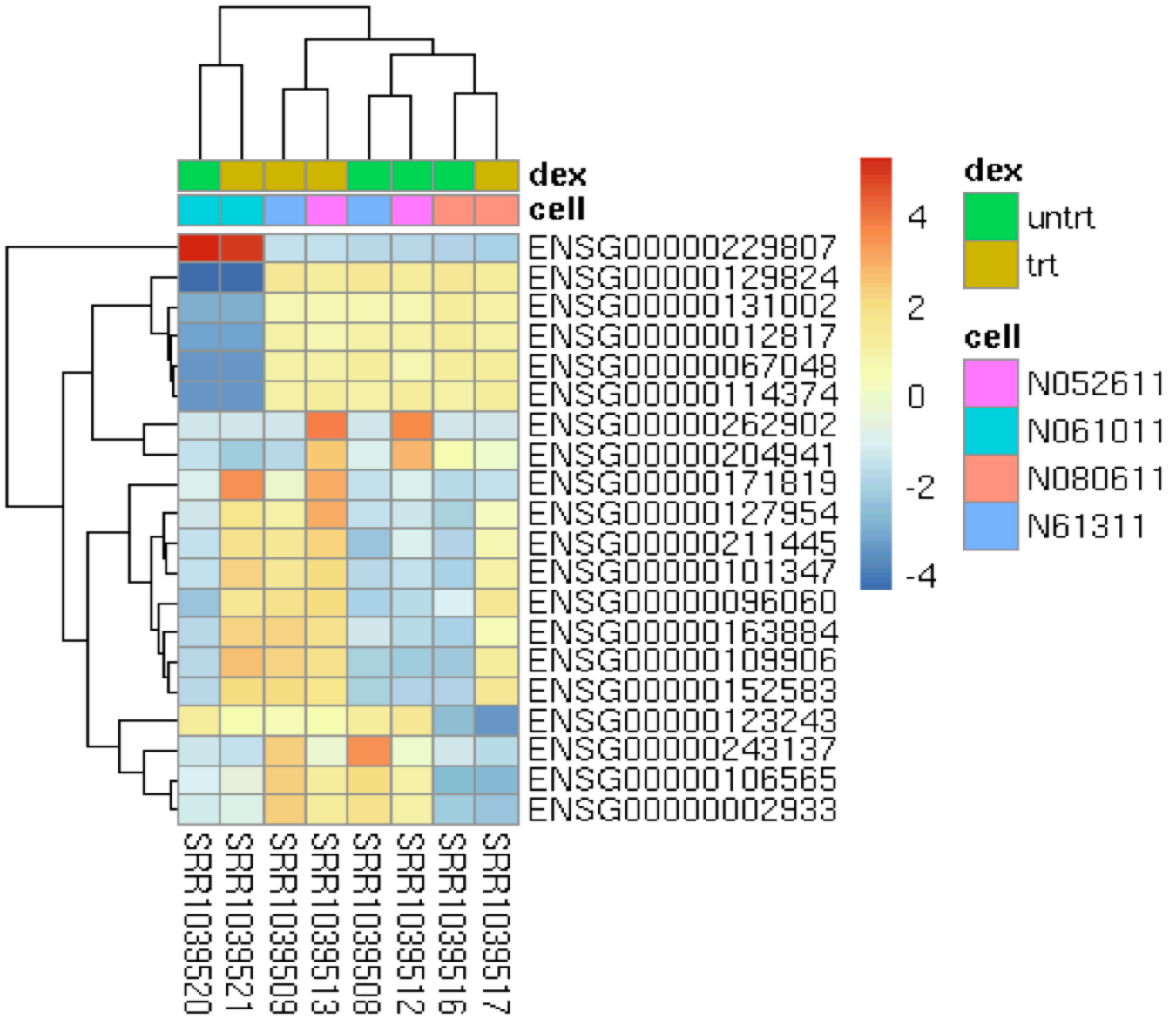

**PCA plot using the rlog-transformed values.** Each unique combination of treatment and cell line is given its own color.

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 *ggplot*.

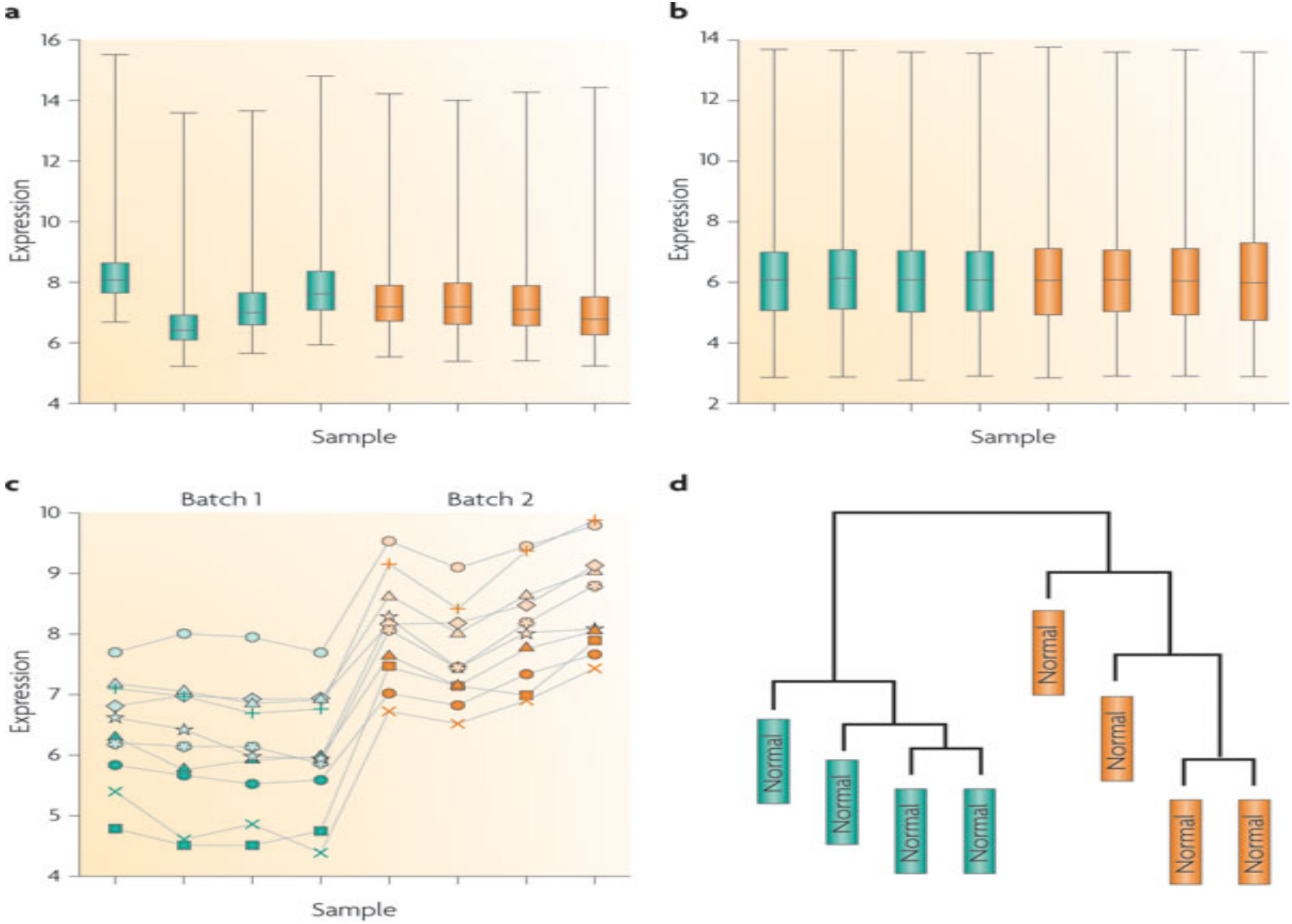



**PCA plot using the rlog-transformed values with custom *ggplot2* code.** Here we specify cell line (plotting symbol) and dexamethasone treatment (color).

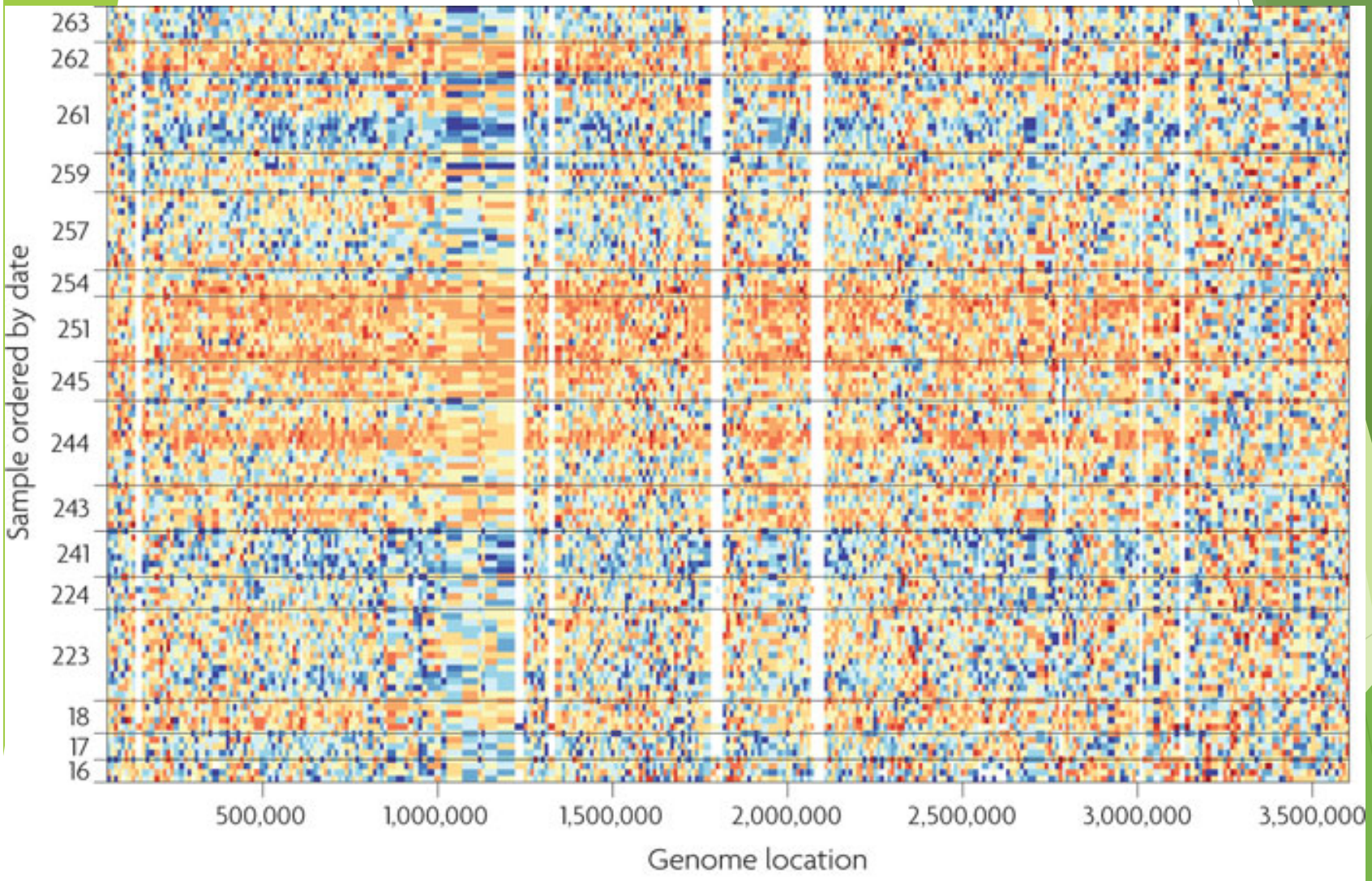
From the PCA plot, we see that the differences between cells (the different plotting shapes) are considerable, though not stronger than the differences due to treatment with dexamethasone (red vs blue color). This shows why it will be important to account for this in differential testing by using a paired 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.



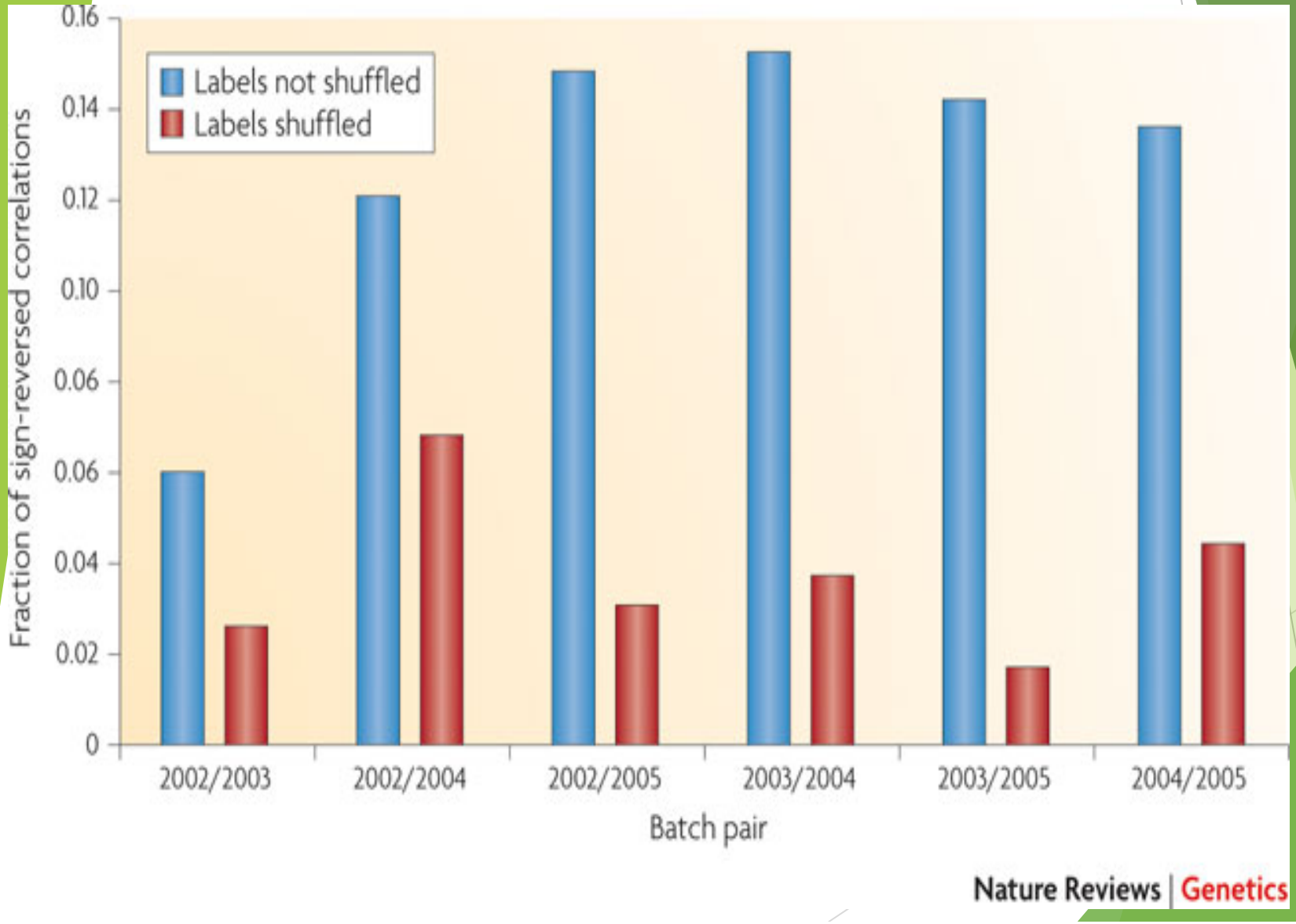

**Heatmap of relative rlog-transformed values across samples.** Treatment status and cell line information are shown with colored bars at the top of the heatmap. Blocks of genes that covary across patients. Note that a set of genes at the top of the heatmap are separating the N061011 cell line from the others. In the center of the heatmap, we see a set of genes for which the dexamethasone treated samples have higher gene expression.

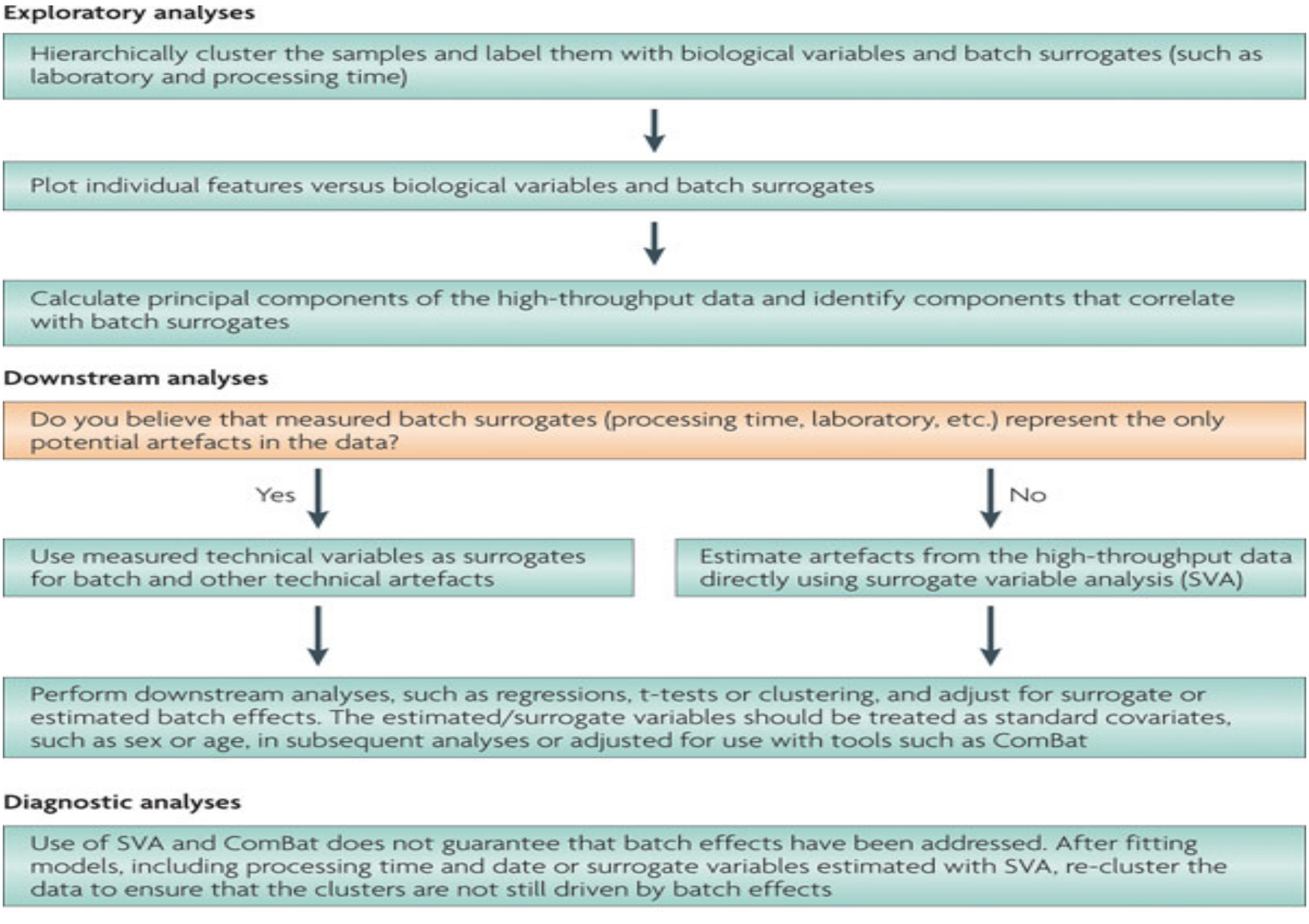


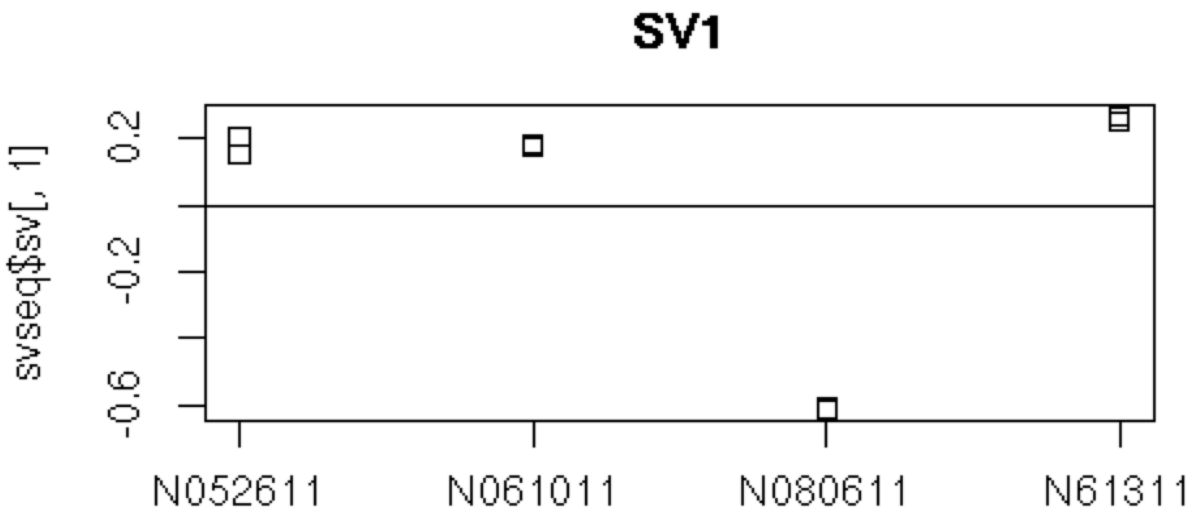

Nature Reviews | Genetics

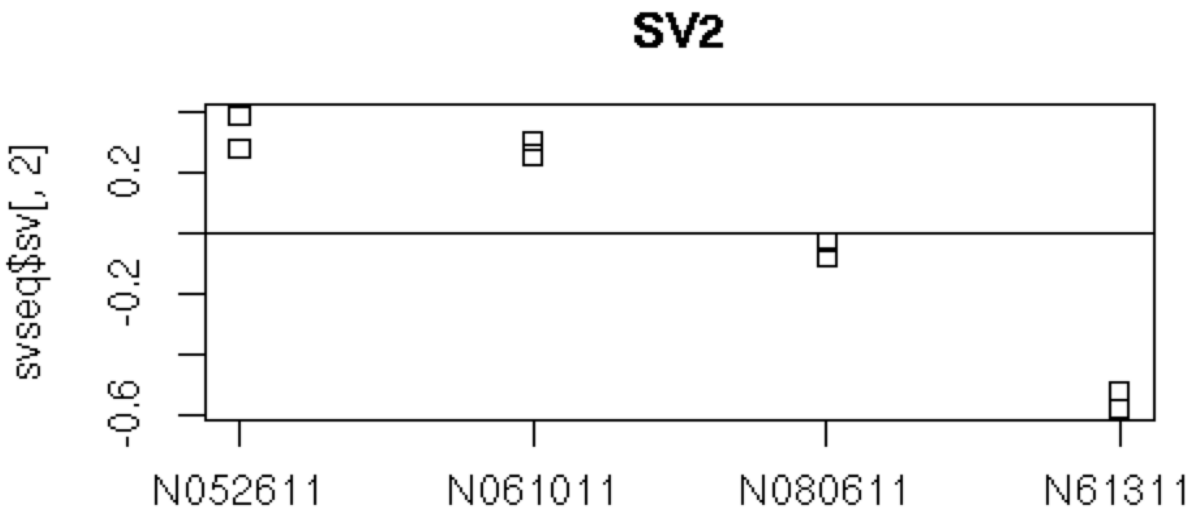



Nature Reviews | Genetics


Labels not shuffled
Labels shuffled
Fraction of sign-reversed correlations
0.16
0.14
0.12
0.10
0.06
0.06
0.04
0.02
0
2002/2003
2002/2004
2002/2005
2003/2004
2003/2005
2004/2005
Batch pair
Nature Reviews | Genetics


Exploratory analyses
Hierarchically cluster the samples and label them with biological variables and batch surrogates (such as laboratory and processing time)
Plot individual features versus biological variables and batch surrogates
Calculate principal components of the high-throughput data and identify components that correlate with batch surrogates
Downstream analyses
Do you believe that measured batch surrogates (processing time, laboratory, etc.) represent the only potential artefacts in the data?
Yes
No
Use measured technical variables as surrogates for batch and other technical artefacts
Estimate artefacts from the high-throughput data directly using surrogate variable analysis (SVA)
Perform downstream analyses, such as regressions, t-tests or clustering, and adjust for surrogate or estimated batch effects. The estimated/surrogate variables should be treated as standard covariates, such as sex or age, in subsequent analyses or adjusted for use with tools such as ComBat
Diagnostic analyses
Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects




**Surrogate variables 1 and 2 plotted over cell line.** Here, we know the hidden source of variation (cell line), and therefore can see how the SVA procedure is able to identify a source of variation which is correlated with cell line.

Finally, in order to use SVA to remove any effect on the counts from our surrogate variables, we simply add these two surrogate variables as columns to the *DESeqDataSet* and then add them to the design:

sva

PLOS | GENETICS

Browse Publish About

OPEN ACCESS PEER-REVIEWED

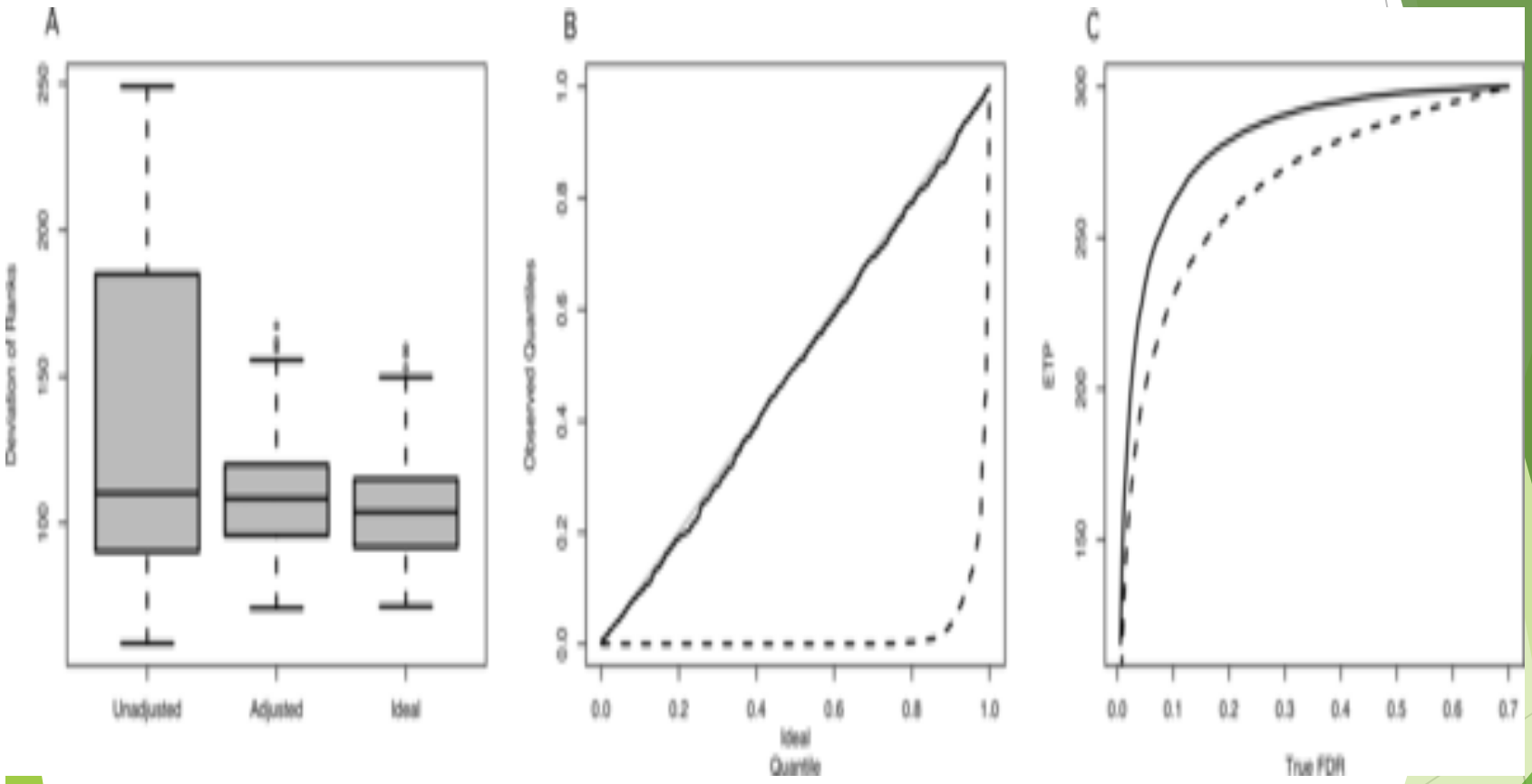
RESEARCH ARTICLE

# 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

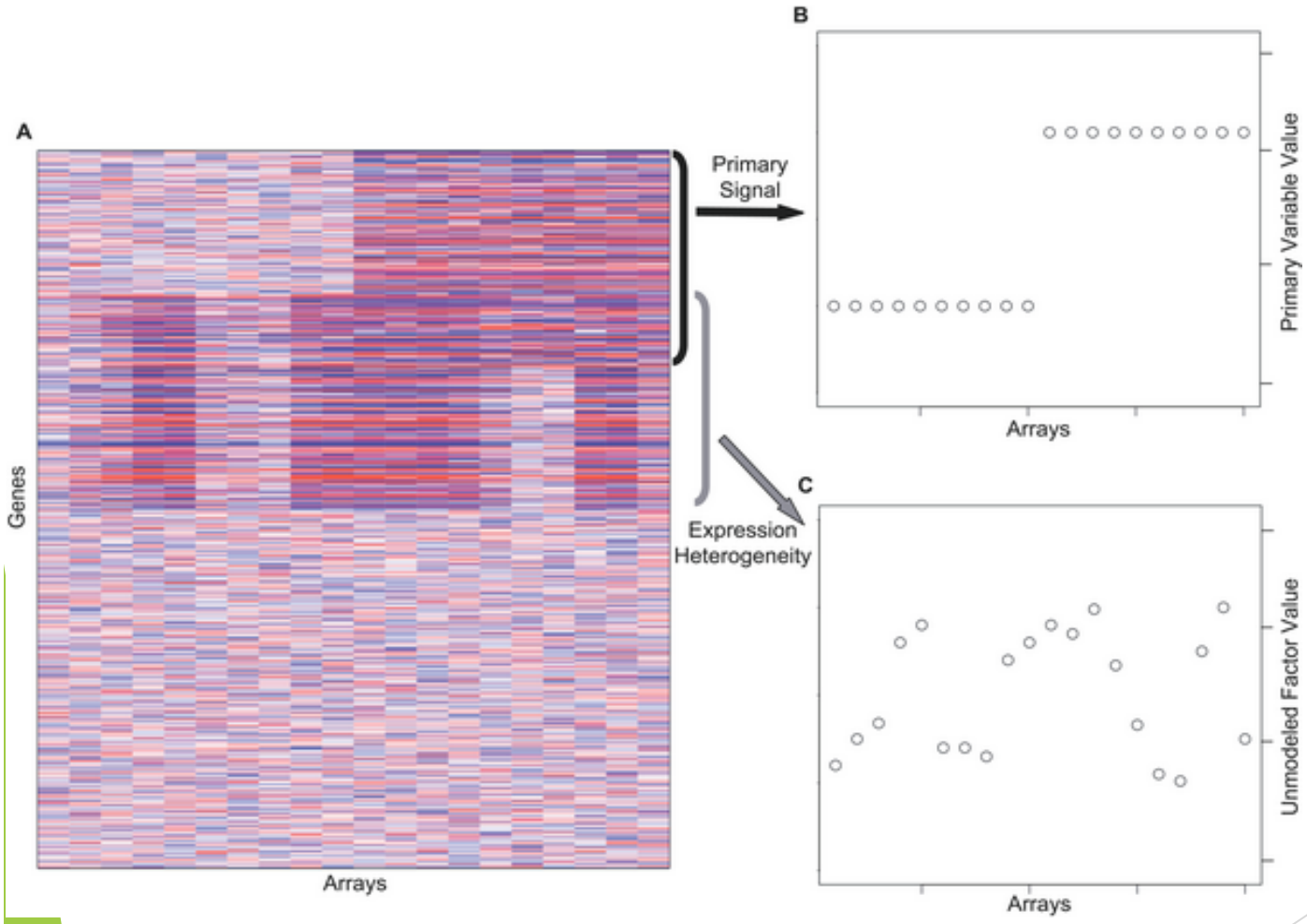
# Figure 1. Impact of Expression Heterogeneity




Leek JT, Storey JD (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet 3(9): e161. doi:10.1371/journal.pgen.0030161
http://journals.plos.org/plosgenetics/article?id=info:doi/10.1371/journal.pgen.0030161

PLOS | GENETICS

# Figure 2. Example of Expression Heterogeneity




Leek JT, Storey JD (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet 3(9): e161. doi:10.1371/journal.pgen.0030161
http://journals.plos.org/plosgenetics/article?id=info:doi/10.1371/journal.pgen.0030161

PLOS | GENETICS

| 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*

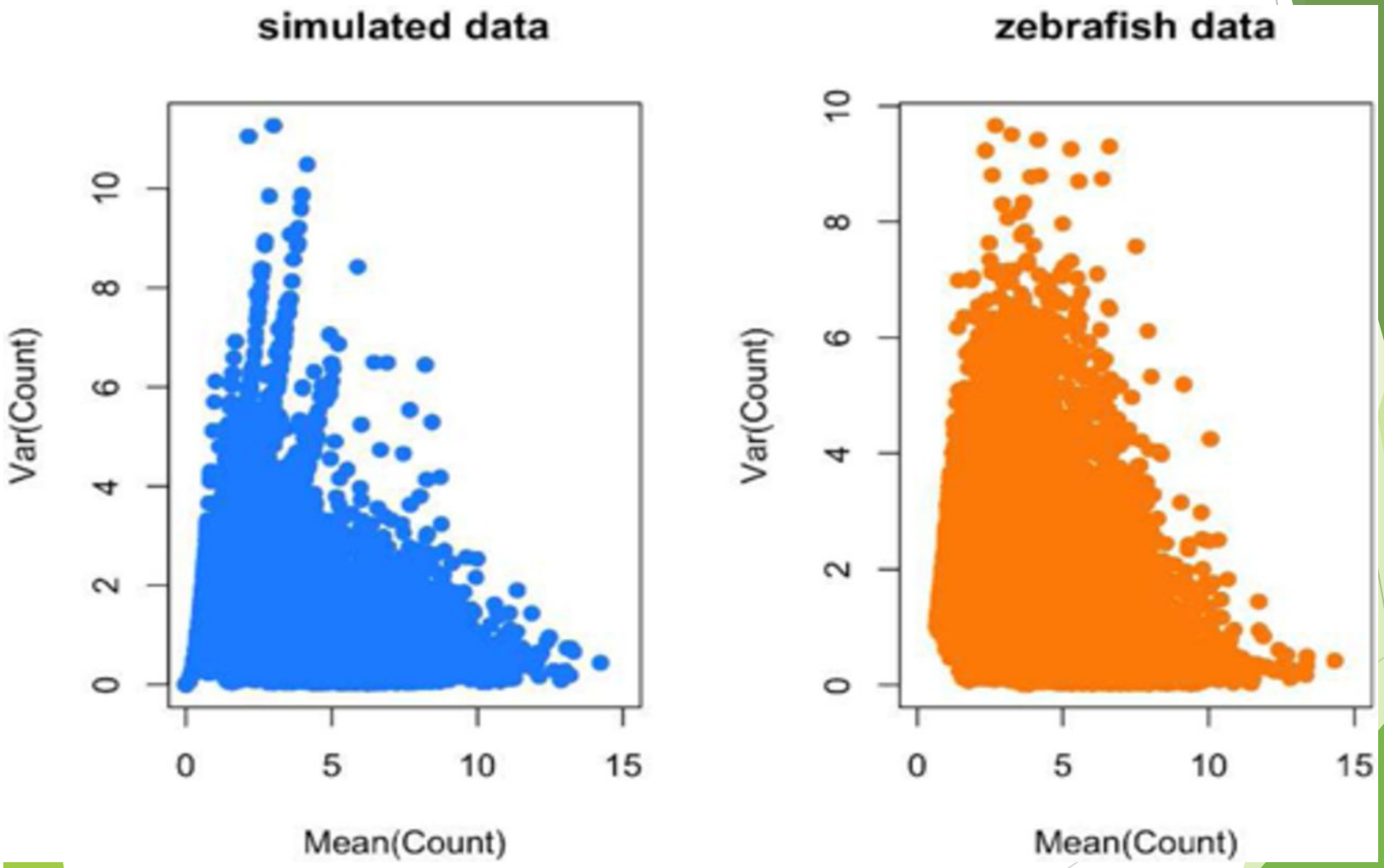
# svaseq: removing batch effects and other unwanted noise from sequencing data

**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

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



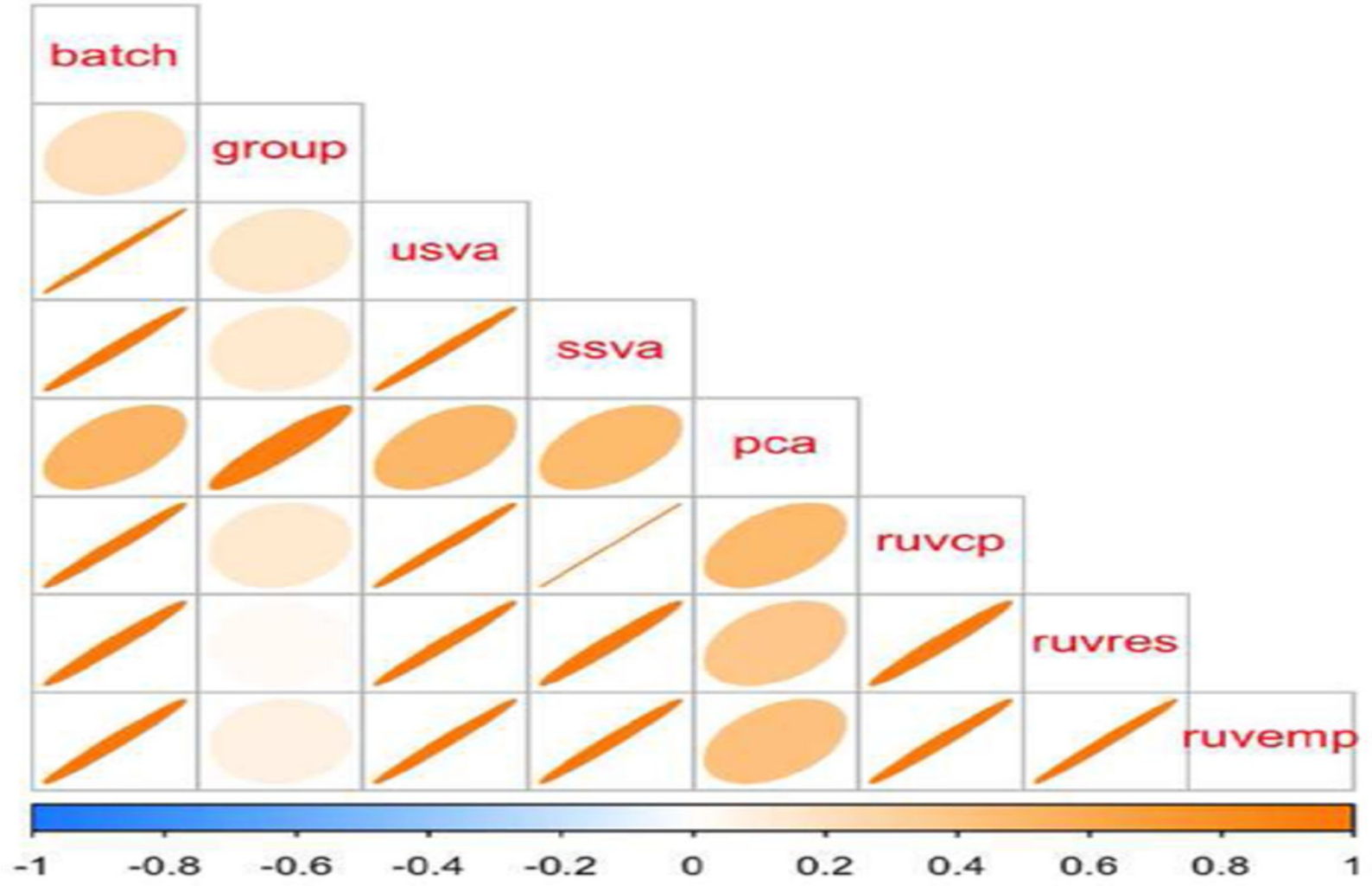




Jeffrey T. Leek Nucl. Acids Res. 2014;nar.gku864

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

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 form GitHub (github.com/drisso/peixoto2015_tutorial)

# drisso / peixoto2015_tutorial





Tutorial to reproduce the analysis of Peixoto et al. (2015)



drisso Update README.md — Latest commit 40878fb on Nov 25, 2015

| | | |
|---|---|---|
| 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.bib | Initial commit | 9 months ago |

README.md

# Tutorial to reproduce the analysis of Peixoto et al. (2015)

## Analysis of FC and OLM datasets

All the needed input files are in the Peixoto_Input_for_Additional_file_1 directory.

This tutorial has been tested with R 3.2.0.

Peixoto_Additional_File_1.Rmd contains the Rmarkdown and can be run to reproduce the full report.

Peixoto_Additional_File_1.R contains a pure R script and can be sourced to any R session.

Peixoto_Additional_File_1.html and Peixoto_Additional_File_1.pdf contain the results of the analysis. Please refer to the Section "Session Info" for details on the package versions used.

## Analysis of publicly available data

The processed expression data from GEO are available in the Peixoto_Input_for_Additional_file_2 directory.

Peixoto_Additional_File_2.Rmd contains the Rmarkdown and can be run to reproduce the analysis.

Peixoto_Additional_File_2.html contains the results.

# End