RNA Seq analysis

Differential Gene Expression

Platform ABiMS
SouthGreen

Recurrent sources

- Gildas Le Corguillé - ABiMS
- Julie Aubert - AgroParisTech/INRA
- Matthias Zytnicki - MIA Toulouse/INRA
- Hugo Varet - Institut Pasteur

Designing the experiment

EXPLAIN THE VARIABILITY

Build an experimental design

Build an experimental design

- to control the variability during the experiment in order to address the biological question:
  1. What is the biological question?
  2. How to estimate the associated biological variabilities?
  3. How to control the technical variabilities (day, lane, run, etc.)?

Biological or technical uncontrolled effects could:

- Hide/cancel the biological effect of interest
- Wrongly increase the biological effect of interest

Experimental design

Basic

id	state
c1	control
c2	control
c3	control
t1	treated
t2	treated
t3	treated

Experimental design

Paired samples

id	state	date
control-t1	control	12/06/2016
control-t2	control	20/06/2016
control-t3	control	25/06/2016
treated-t1	treated	12/06/2016
treated-t2	treated	20/06/2016
treated-t3	treated	25/06/2016

Experimental design

Paired samples

id	state	sample
sample1-control	control	sample1
sample1-treated	treated	sample1
sample2-control	control	sample2
sample2-treated	treated	sample2
sample3-control	control	sample3
sample3-treated	treated	sample3

Experimental design

Paired samples

id	tissue	sample
sample1-skin	skin	sample1
sample1-muscle	muscle	sample1
sample2-skin	skin	sample2
sample2-muscle	muscle	sample2
sample3-skin	skin	sample3
sample3-muscle	muscle	sample3

Experimental design

Time course experiment

id	state	sample	time
sample1-0h	treated	sample1	0h
sample2-0h	treated	sample2	0h
sample3-0h	treated	sample3	0h
sample1-4h	treated	sample1	4h
sample2-4h	treated	sample2	4h
sample3-4h	treated	sample3	8h
sample1-8h	treated	sample1	8h
sample2-8h	treated	sample2	8h
sample3-8h	treated	sample3	8h

Experimental design

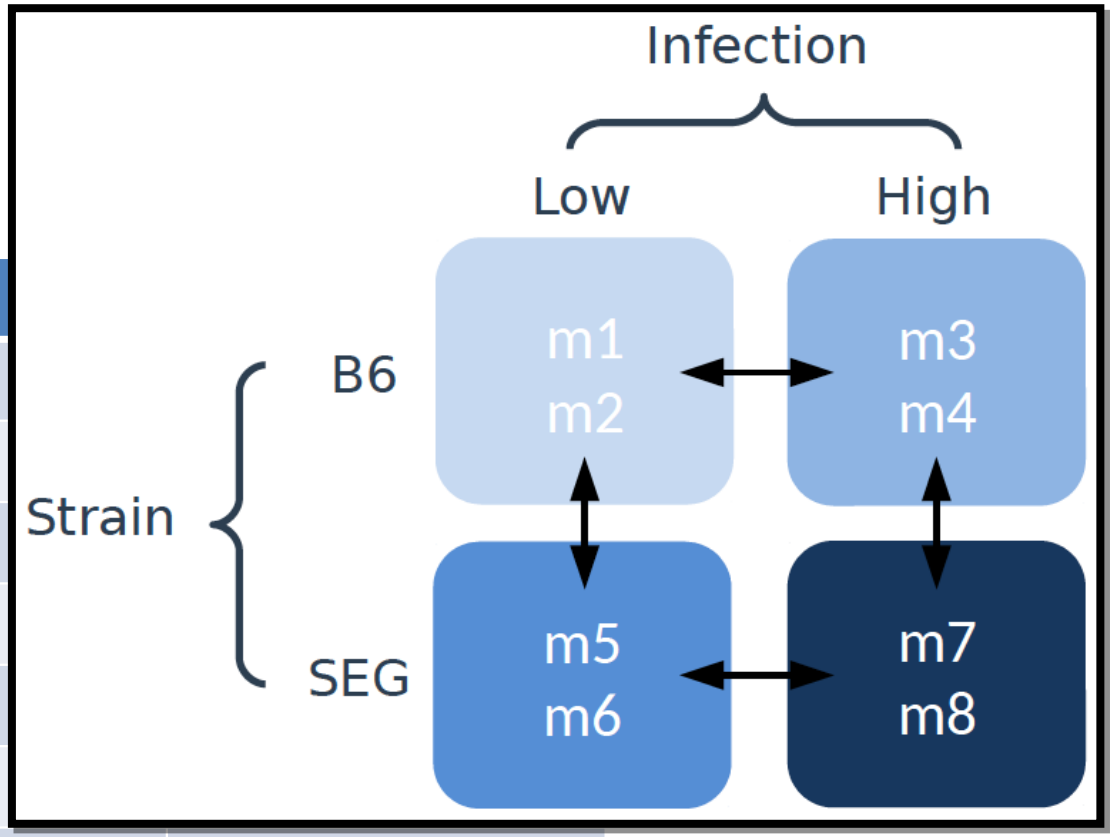
Complex design

id	strain	infection
m1	B6	low
m2	B6	low
m3	B6	high
m4	B6	high
m5	SEG	low
m6	SEG	low
m7	SEG	high
m8	SEG	high

Experimental design

Complex design

id	strain	infection
m1	B6	low
m2	B6	low
m3	B6	high
m4	B6	high
m5	SEG	low
m6	SEG	low
m7	SEG	high
m8	SEG	high



Experimental design

Which effect?

Confounding effect

id	state
c1	control
c2	control
c3	control
t1	treated
t2	treated
t3	treated

Experimental design

Which effect?

Confounding effect

id	state	age	gender	date	exp
c1	control	45	female	09/06/15	Louis
c2	control	52	female	11/06/15	Louis
c3	control	48	female	13/06/15	Louis
t1	treated	31	male	21/02/15	François
t2	treated	25	male	23/02/15	François
t3	treated	27	male	25/02/15	François

Experimental design

Which effect?

~~Confounding effect~~

id	state	age	gender	date	exp
c1	control	45	female	09/06/15	Louis
c2	control	36	female	11/03/15	François
c3	control	48	male	23/11/15	Louis
c4	control	22	male	15/02/15	François
t1	treated	31	female	21/02/15	François
t2	treated	25	female	03/12/15	François
t3	treated	27	male	25/07/15	Louis
t4	treated	45	male	01/01/16	Louis

HOW DEEP IS ENOUGHT?

How deep is enought ?

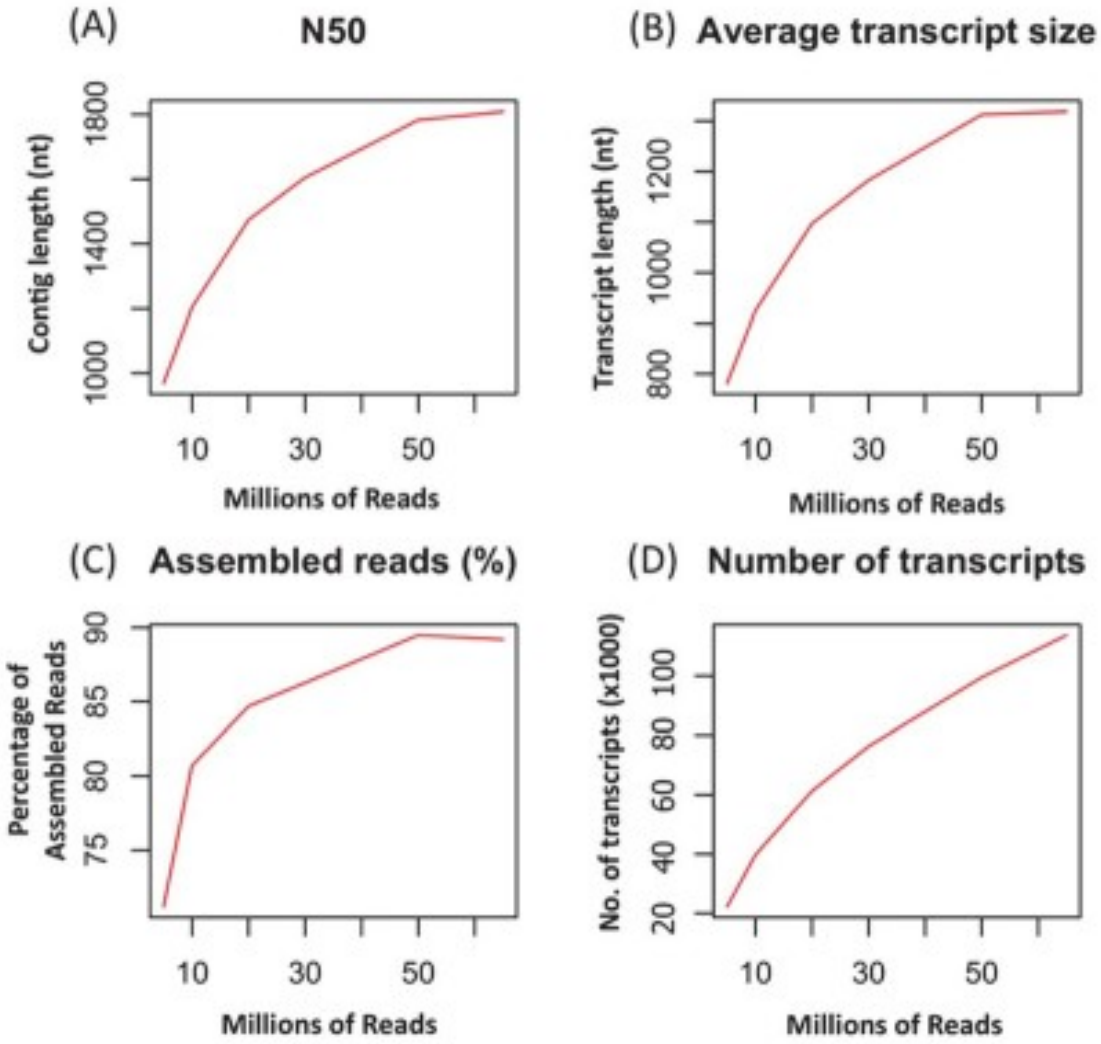


Fig. 6 Effect of sequencing depth on a transcriptome assembly. Four Paired-End assemblies using 5, 10, 20, 30, 50 and 65 million reads were generated using Oases.[37] The N50 contig size (A), average transcript size (B), percentage of reads used in the assembly (C), and number of transcripts (D) *versus* number of reads used in the assembly are shown.

Góngora-Castillo, E., & Buell, C. R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Natural Product Reports. doi:10.1039/c3np20099j

How deep is enought ?

Human

Majority of expressed genes and AS events can be detected with modest sequencing depths (~100 M filtered reads), the estimated gene expression levels and exon/intron inclusion levels were less accurate

- To detect expressed genes and AS events, ~100 to 150 million (M) filtered reads were needed.
- For a DE analysis and detect 80% of events, ~300 M filtered reads were needed
- For detecting differential AS and detect 80% of events, at least 400 M filtered reads were necessary

Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. Yichuan Liu et al., 2013.

How deep is enought ?

Depends on the purpose of the experiment and the nature of the samples (ENCODE).

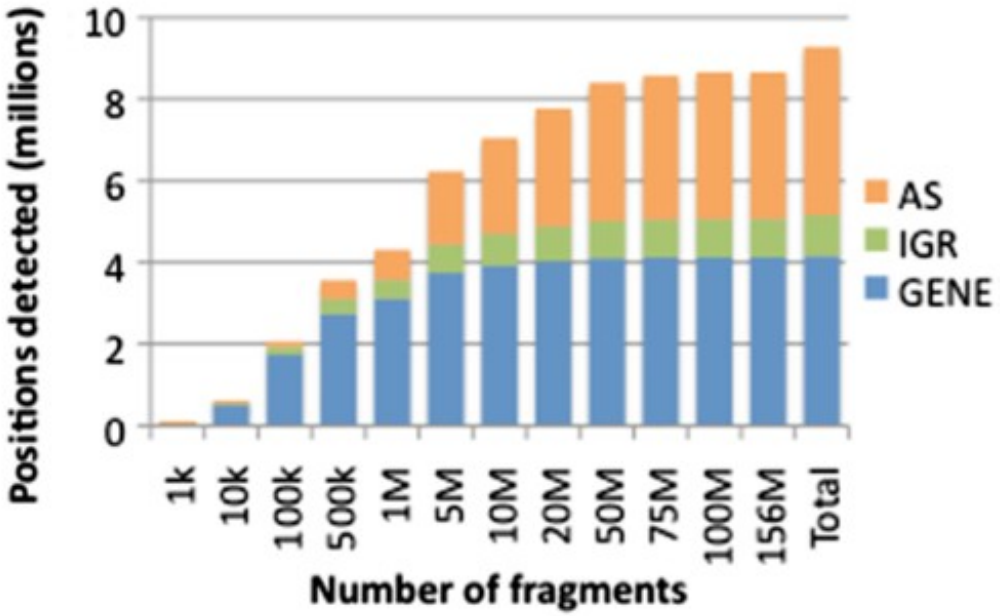
- 100M of reads is sufficient to detect 90% of the transcripts and 81% of the genes of the human transcriptome. (Tung et al. 2011)
- 20M reads (75bp) is sufficient to detect transcripts expressed at a medium or low level in the chicken. (Wang et al. 2011)
- 10 M of reads allow 90% of transcripts (human, zebrafish) to be covered by an average of 10 reads. (Hart et al. 2013)
- Between 30M and 100M reads per sample depending on the study.

NB.http://encodeproject.org/ENCODE/dataStandards.html

How deep is enought ?

Bacteria

E. Coli : 5000 genes
intergenic (IGR)
antisense to ORFs or ncRNAs (AS)



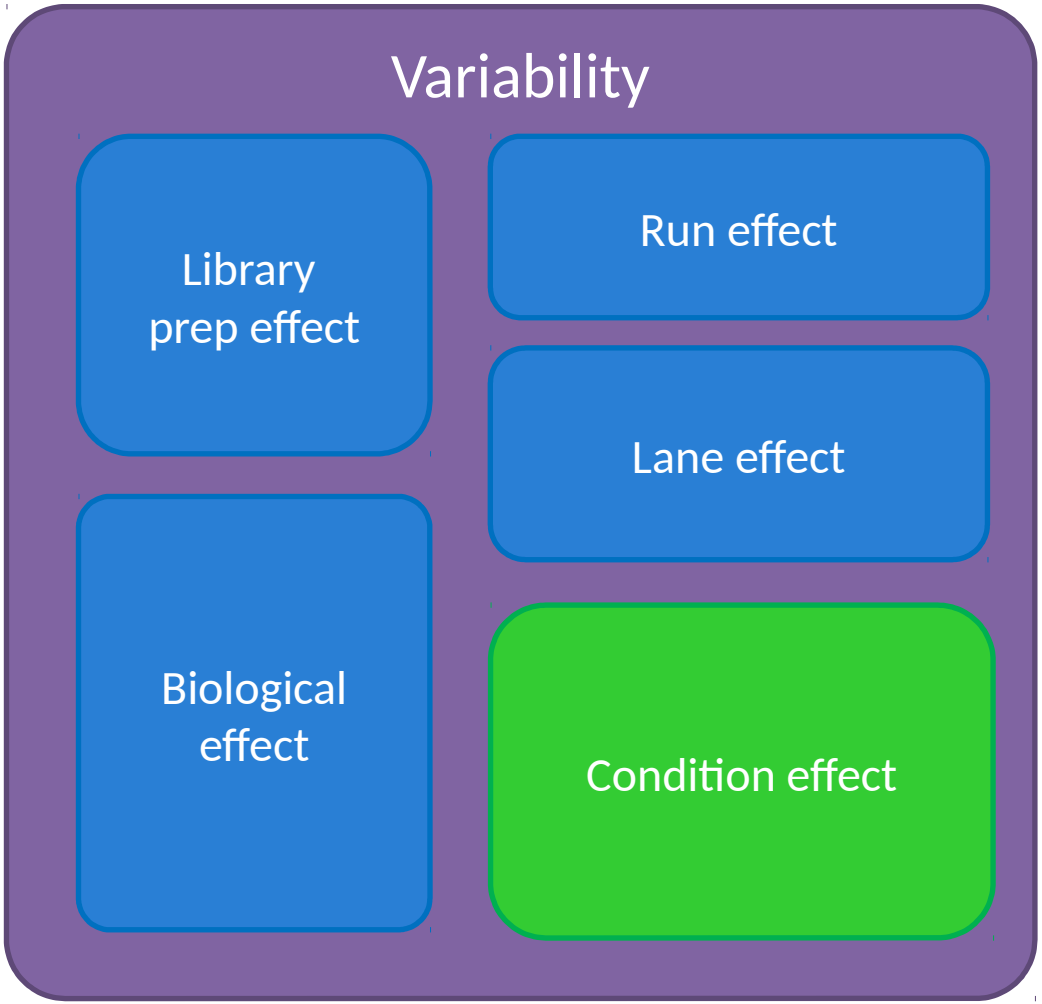
« A sequencing depth of **5-10 million** non- rRNA fragments enables profiling of the vast majority of transcriptional activity in diverse species grown under diverse culture conditions. »

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., & Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? BMC genomics, 13, 734. doi:10.1186/1471-2164-13-734

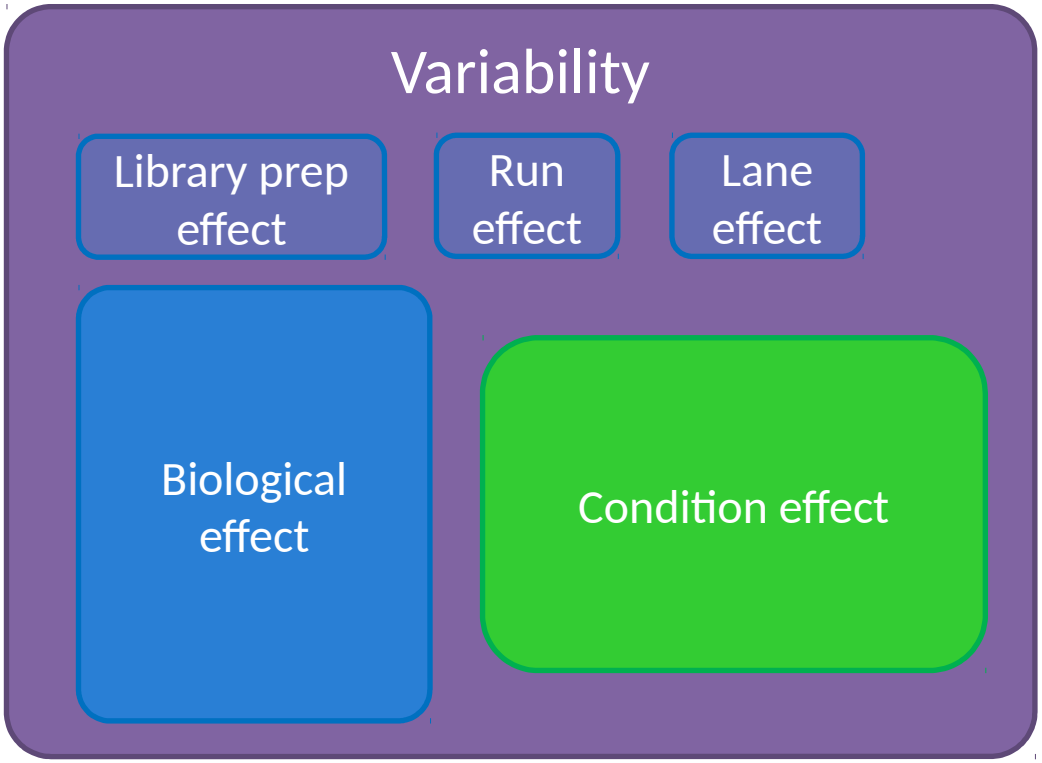
Bias

EXPLAIN THE VARIABILITY

The variability

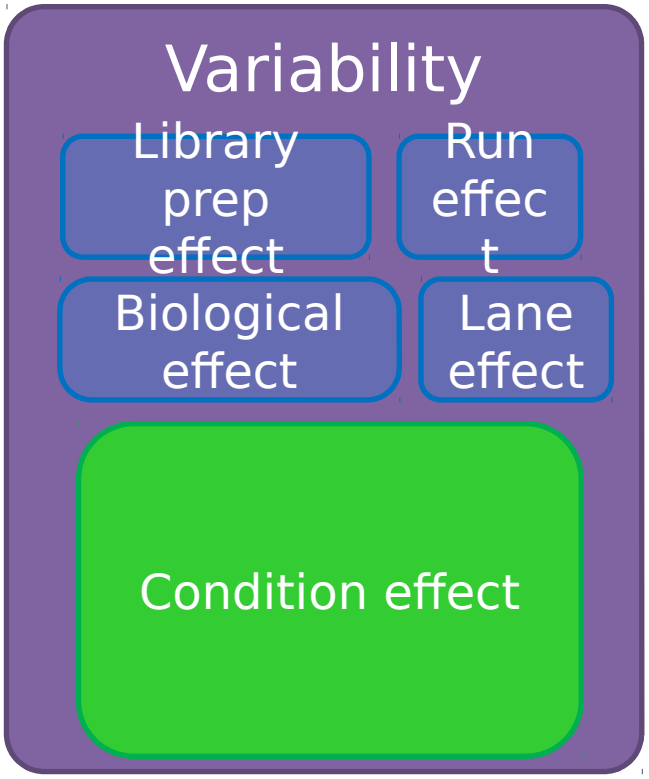


The variability



Technical replicates
+ normalization
+ statistics

The variability



Technical replicates
+ normalization
+ statistics

+

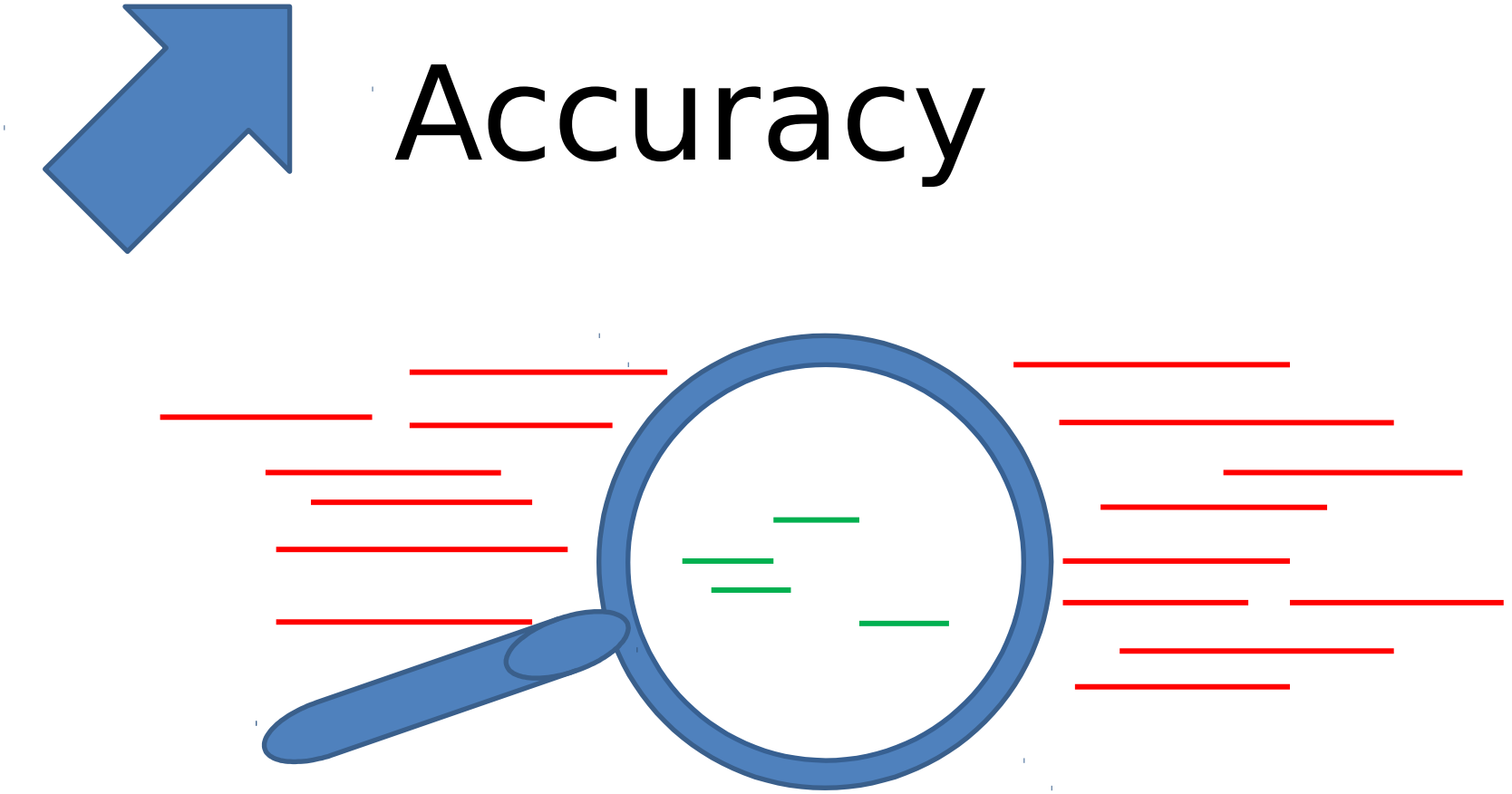
Biological replicates
+ statitics

REPLICATES

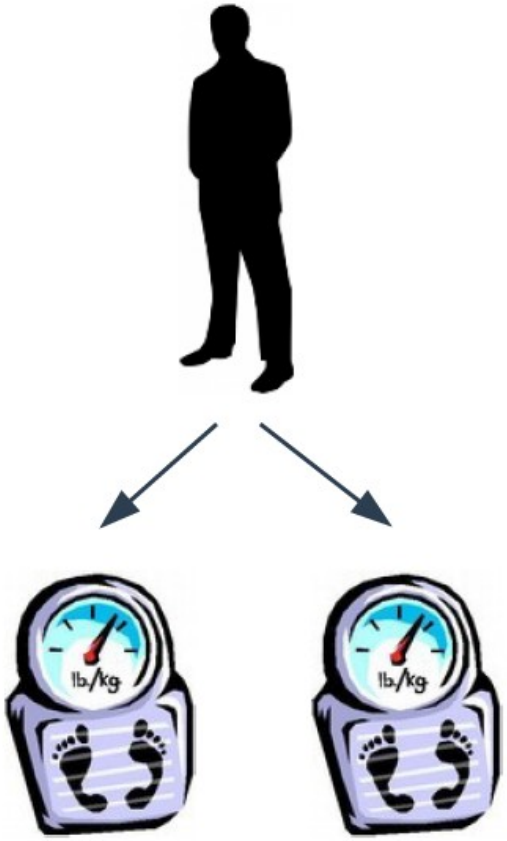
Replicates

RNA-Seq is not a mature technology.

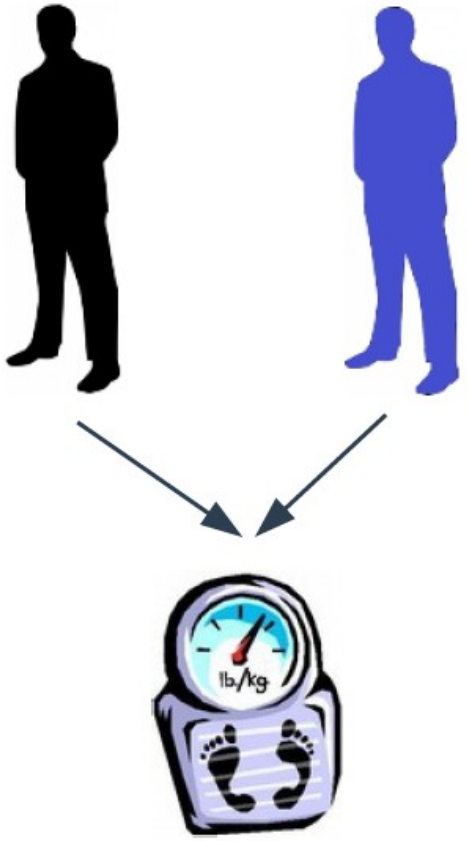
Experiments should be performed with **three or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful



Biological vs. technical replicates

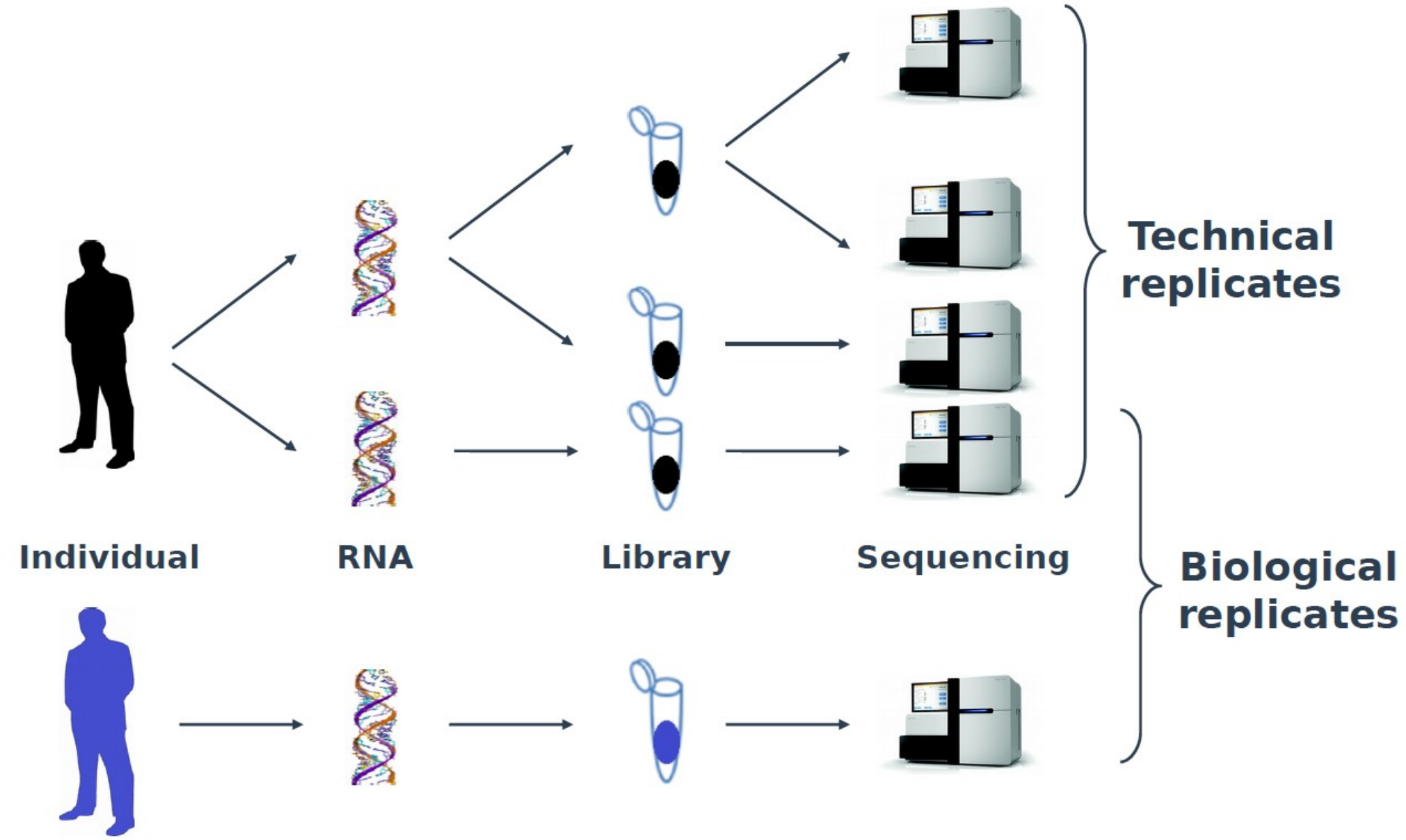


Technical

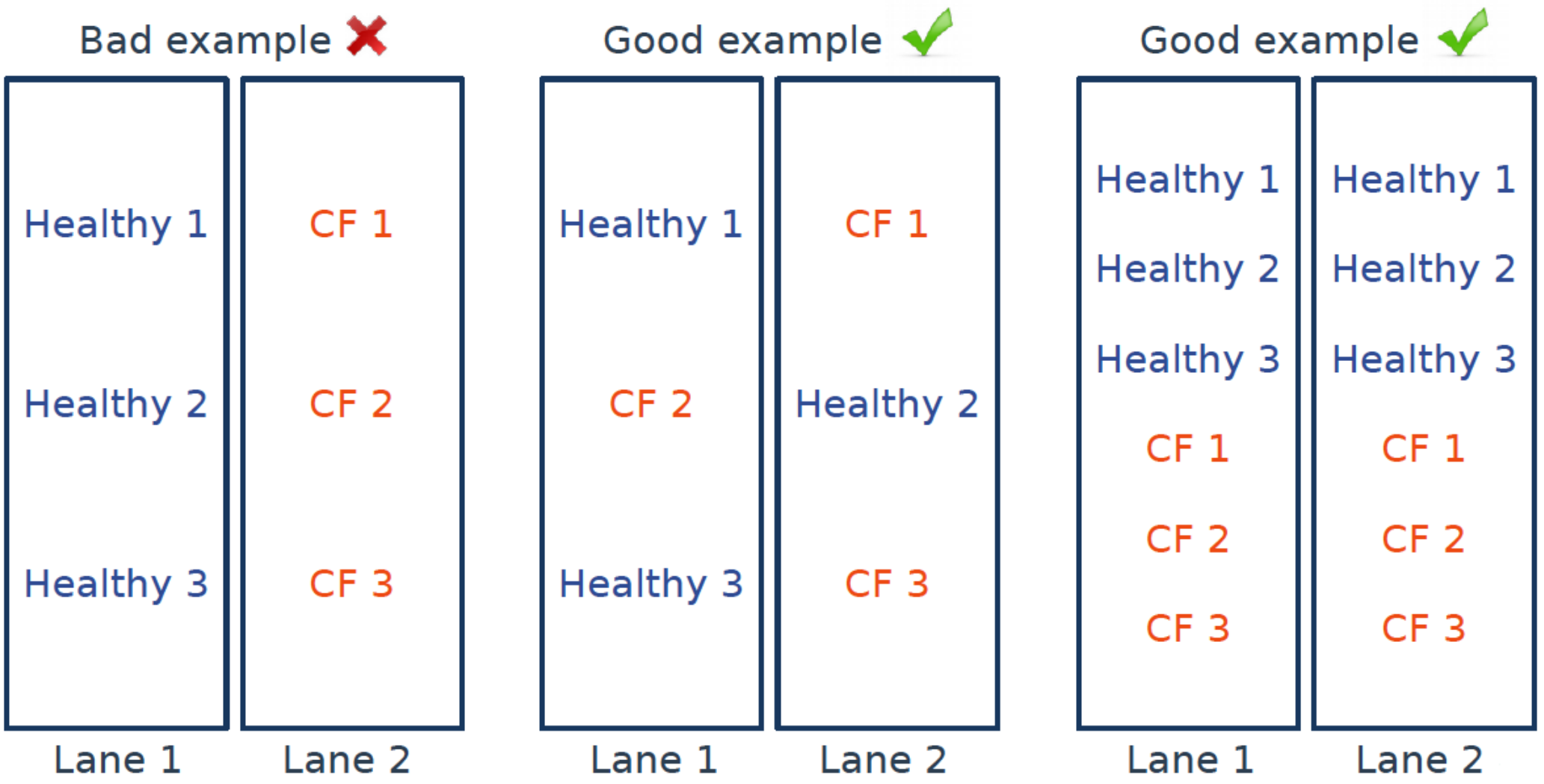


Biological

Biological vs. technical replicates



Technical replicates



Technical vs biological replicates

- Increasing the number of **bio. replicates** increases the **precision and generalizability** of the results

- Doing **technical replication** may be important in studies where **low abundant mRNAs** are the focus.

- Technical variability => inconsistent detection of exons at low levels of coverage (<5reads per nucleotide) (McIntyre et al. 2011)

Guidelines from the Encyclopedia of DNA Elements (ENCODE) consortium (June 2011)

Experiment design : ENCODE recommandations

- A typical R2 (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between 0.92 to 0.98.
- Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.

- Correlation of >0.9 between isogenic replicates (replicates from biosamples derived from the same model organism strain) and >0.8 between anisogenic replicates (replicates from biosamples derived from different model organism strain) .

Correlation analysis : PtR Perl-to-R

Raw count matrix -> counts-per-million (CPM) data transformation followed by a log2 transform

Compare replicate

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2
--min_rowSums 10 --compare_replicates
```

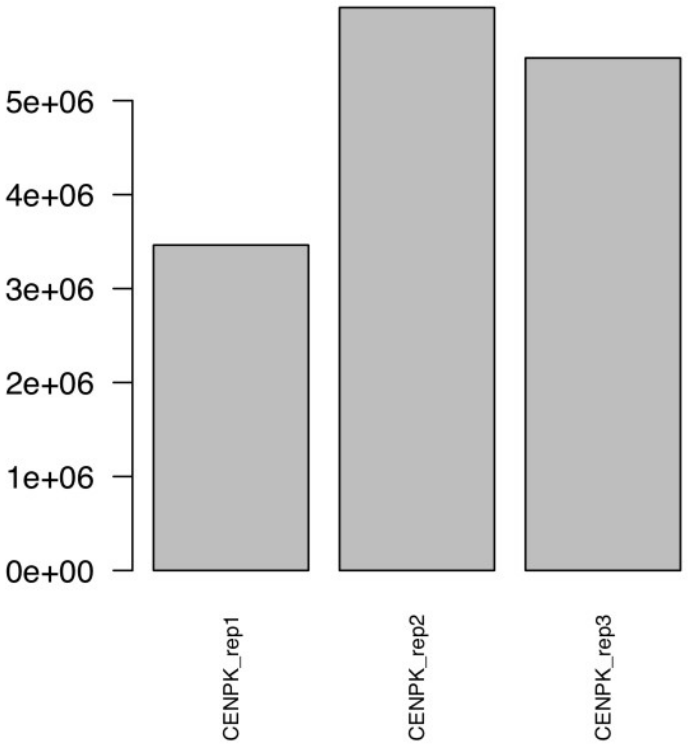
Correlation matrix

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2
--min_rowSums 10 --CPM --sample_cor_matrix
```

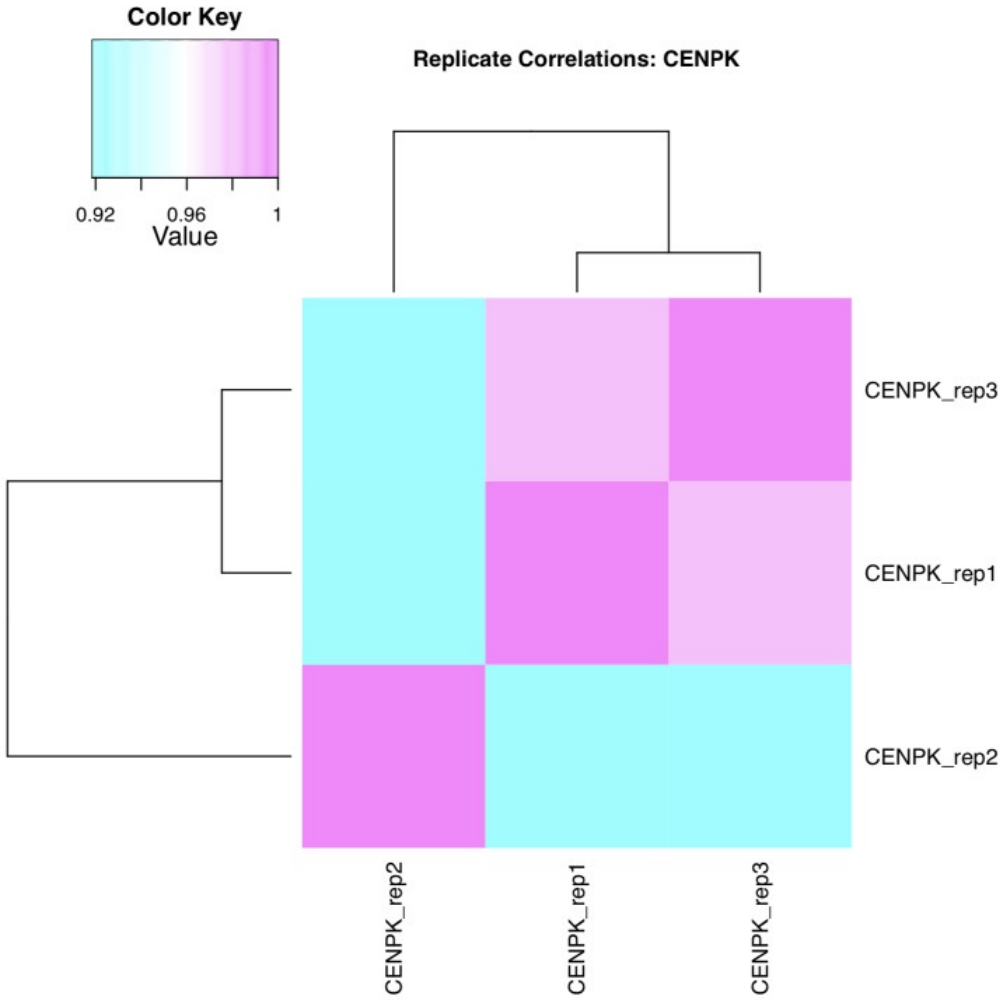
Principal composant analysis

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2
--min_rowSums 10 --CPM --center_rows --prin_comp 3
```

Compare replicates for each of your samples



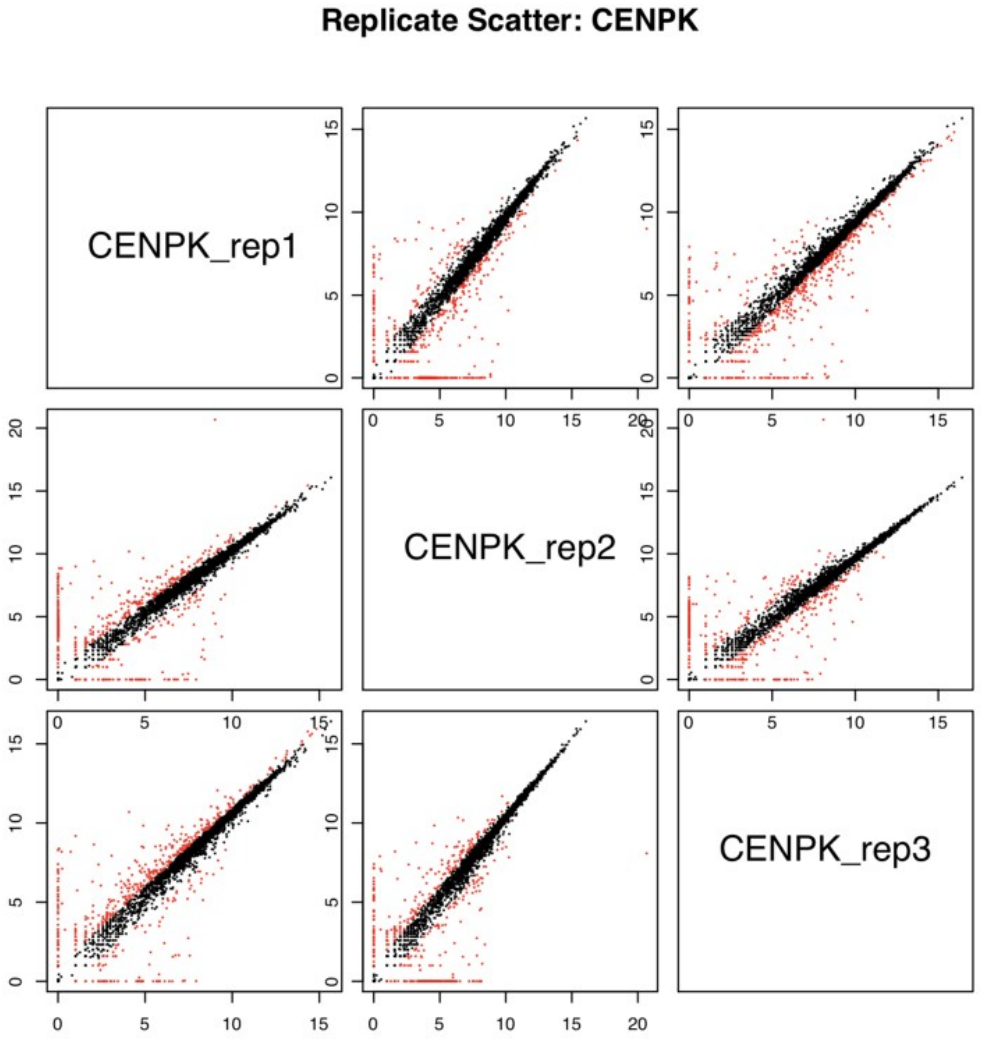
Sequencing depth



Correlation matrix

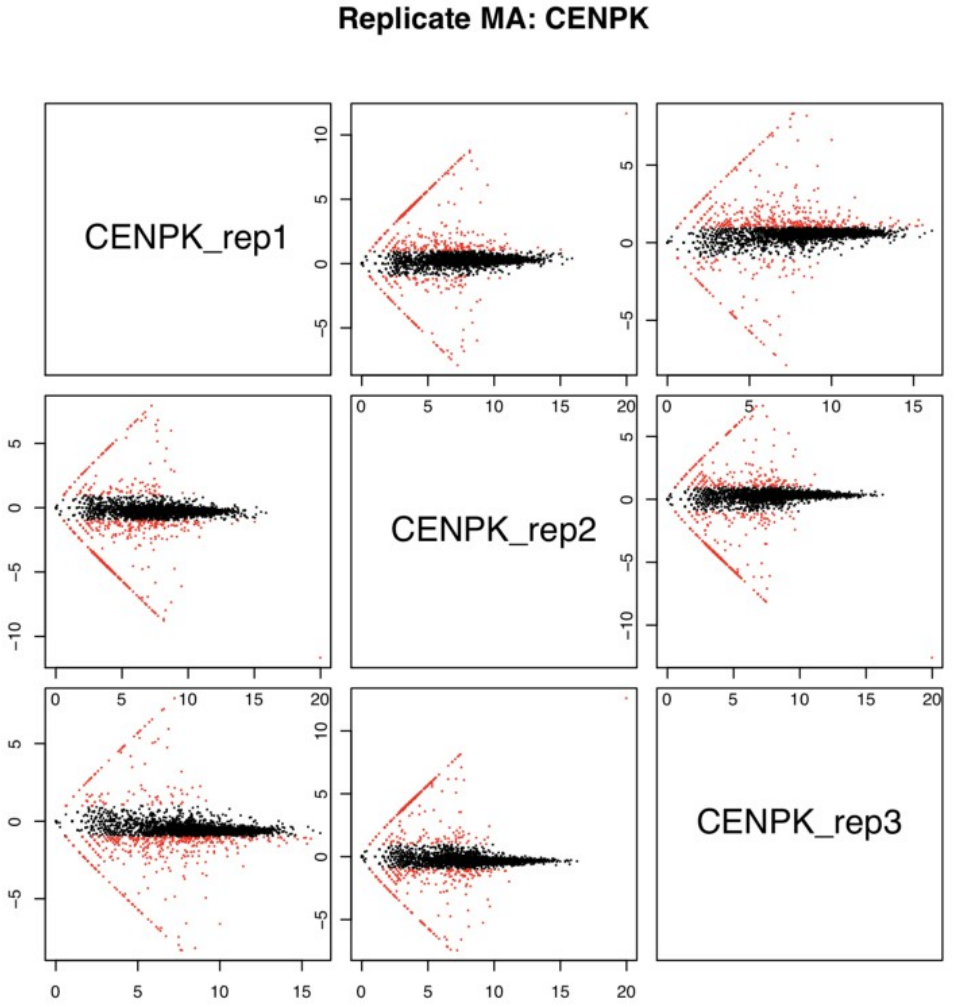
Pearson analysis

Compare replicates for each of your samples



Scatter plot

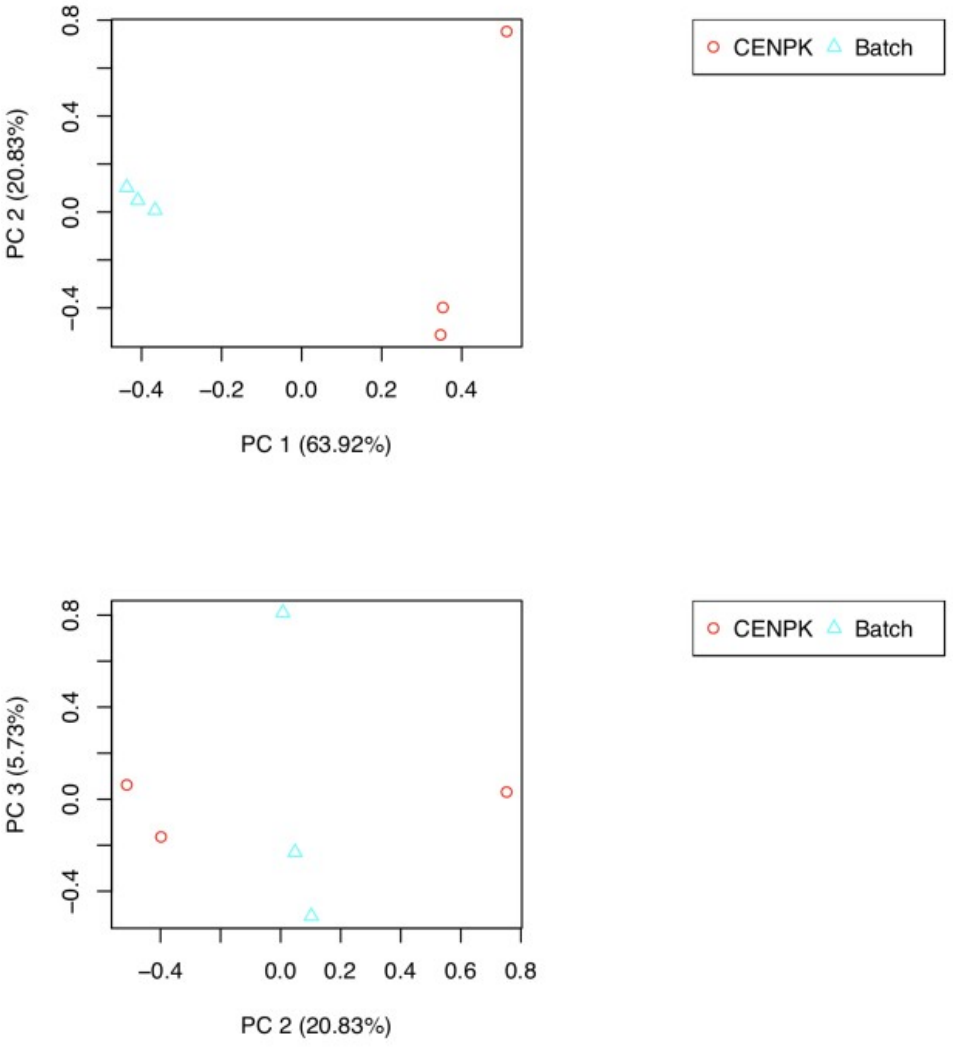
Pairwise comparisons of replicate log(CPM) values.
2-fold different are highlighted in red:

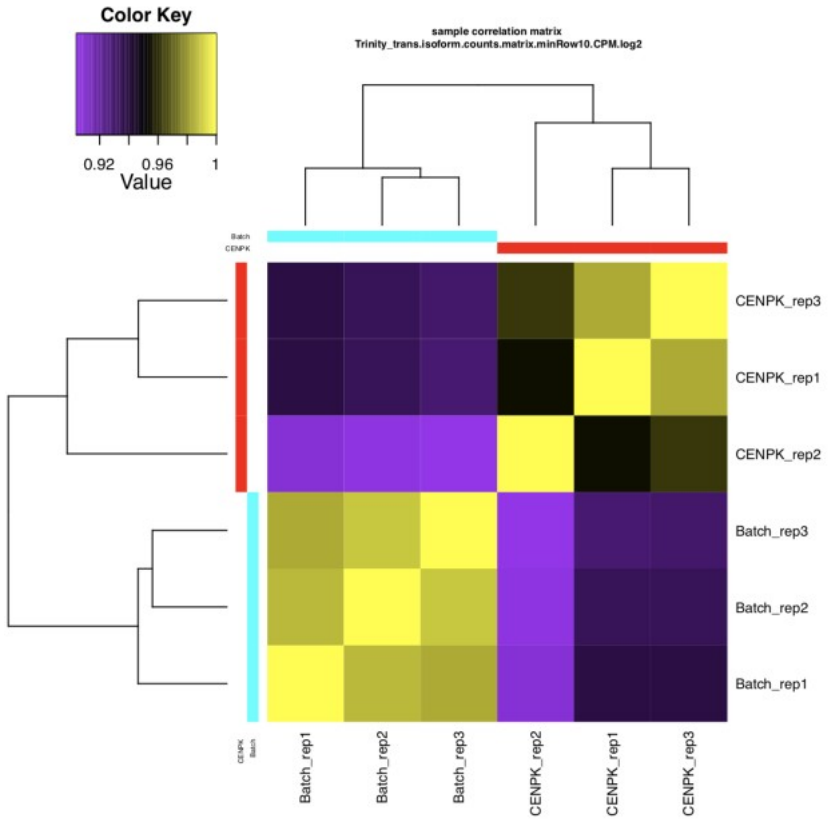


MA plot

x-axis: mean log(CPM), y-axis log(fold_change).
2-fold different are highlighted in red:

Compare Replicates Across Samples





PCA

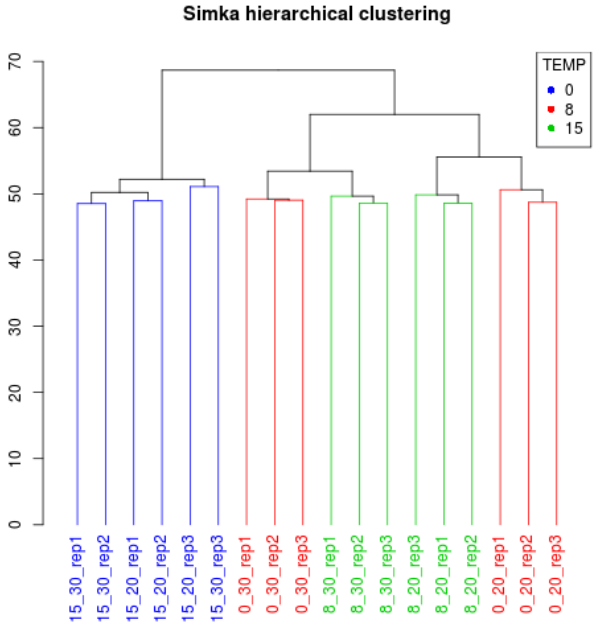
Correlation matrix

Simka

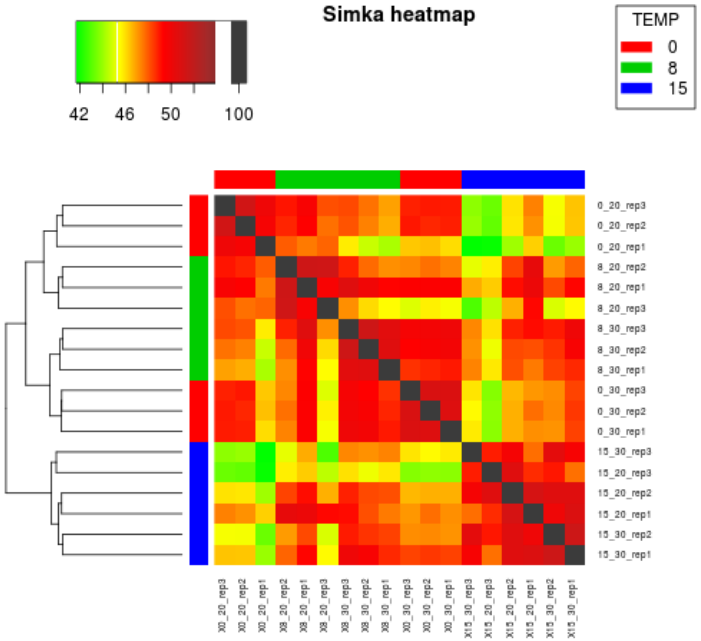
https://github.com/GATB/simka

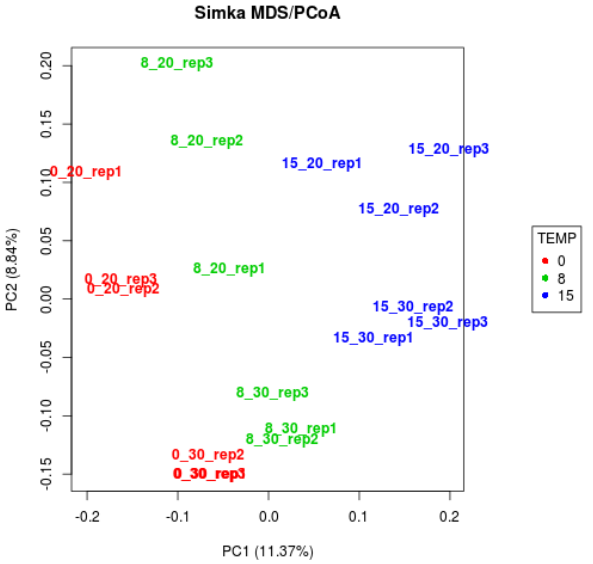
Simka is a *de novo* comparative metagenomics tool.
Simka represents each dataset as a k-mer spectrum and compute several classical ecological distances between them.

Presence/absence Jaccard index



Abundance BrayCurtis index



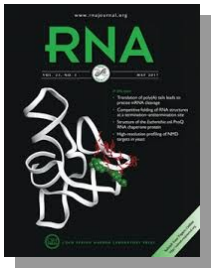


Sample size

Why increase the number of biological replicates?

- Generalizing the results to the population
- Estimate more accurately the variation of each transcript individually (Hart et al. 2013)
- Improve the detection of differential transcripts and rate control false positives: TRUE from 3 (Sonenson et al, 2013, Robles et al 2012.)

Sample size



How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? Schurch et al. RNA. 2016 Jun; 22(6): 839–851.

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 »:

- At least six replicates per condition for all experiments.
- At least 12 replicates per condition for experiments where identifying the majority of all DE genes is important.
- For experiments with <12 replicates per condition; use *edgeR (exact)* or *DESeq2*.
- For experiments with >12 replicates per condition; use *DESeq*.
- Apply a fold-change threshold appropriate to the number of replicates per condition between $0.1 \leq T \leq 0.5$ (see Fig. 2 and the discussion of tool performance as a function of replication).

Sample size

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, “A survey of best practices for RNA-seq data analysis.,” *Genome Biol.*, vol. 17, p. 13, 2016.

Sample size

TABLE 2. A summary of the recommendations of this paper

	Agreement with other tools[a]	WT vs. WT FPR[b]	Fold-change threshold (T)[c]	Tool recommended for: (# good replicates per condition)[d] ≤3	≤12	>12
DESeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
DESeq2	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
EBSeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
edgeR (exact)	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
Limma	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
cuffdiff	Consistent	Fail				
BaySeq	Inconsistent	Pass				
edgeR (GLM)	Inconsistent	Pass				
DEGSeq	Inconsistent	Fail				
NOISeq	Inconsistent	Fail				
PoissonSeq	Inconsistent	Fail				
SAMSeq	Inconsistent	Fail				

[a]Full clean replicate data set, see section "Tool Consistency with High Replicate Data" and Figure 3.
[b]See section "Testing Tool False Positive Rates" and Figure 4.
[c]See section "Differential Expression Tool Performance as a Function of Replicate Number."
[d]See Figure 2.

Schurch et al. RNA. 2016

A statistical answer : Conclusions

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics, 13, 484.

This work quantitatively explores comparisons between contemporary analysis tools and experimental design choices for the detection of differential expression using RNA-Seq. …

- With regard to testing of various experimental designs, this work strongly suggests that greater power is gained through the use of biological replicates relative to library (technical) replicates and sequencing depth.

- Strikingly, sequencing depth could be reduced as low as 15% without substantial impacts on false positive or true positive rates.

Sample size vs depth

It's up to you! (Haas et al., 2012, Liu Y. et al 2013)

- Detection of differential transcripts:
  - (+) biological replicates
- Construction / transcriptome annotation:
  - (+) depth & (+) conditions
- Search variants:
  - (+) biological replicates & (+) depth

Pool

? Are those pooling are the same?

Pool

? Are those pooling are the same?

	samp1			samp2			FC
gene1	1000	1000	1000	2000	2000	2000	1/2
gene2	0	0	3000	0	0	6000	1/2
count			300			600	

- One sample with one over expressed gene can flood the count

The Fold change w/out replicates

? Is FC enough to describe the variability?

The Fold change w/out replicates

? Is FC enough to describe the variability?

	samp1	samp2	FC
gene1	1	2	1/2
gene2	1000	2000	1/2

are the same ?

- FC can mask genes with large differences (B-A) but small ratios (A/B)

HOW TO PERFORM A DEG ANALYSIS

INPUTS

Inputs

1. Raw count table

id	LL06_1	LL06_2	LL09_1	LL09_2
comp3130_seq1	12	6	9	15
comp3131_seq2	167	233	987	856
comp4523_seq1	685	785	648	458
comp6984_seq3	87	68	354	591

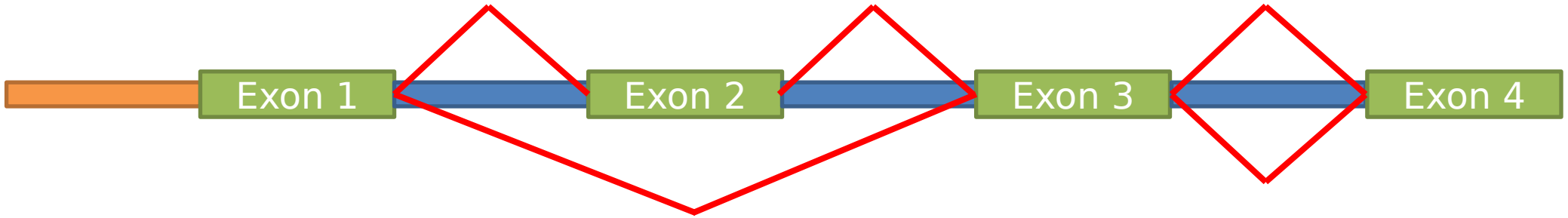
Inputs

2. Samples metadata / Samples info

samplename	batch	light	hour	…
LL06_1	1	LL	06	
HL06_1	1	HL	06	
LL09_1	1	LL	09	
HL09_1	1	HL	09	
LL12_1	1	LL	12	
HL12_1	1	HL	12	
LL06_2	2	LL	06	
HL06_2	2	HL	06	
LL09_2	2	LL	09	

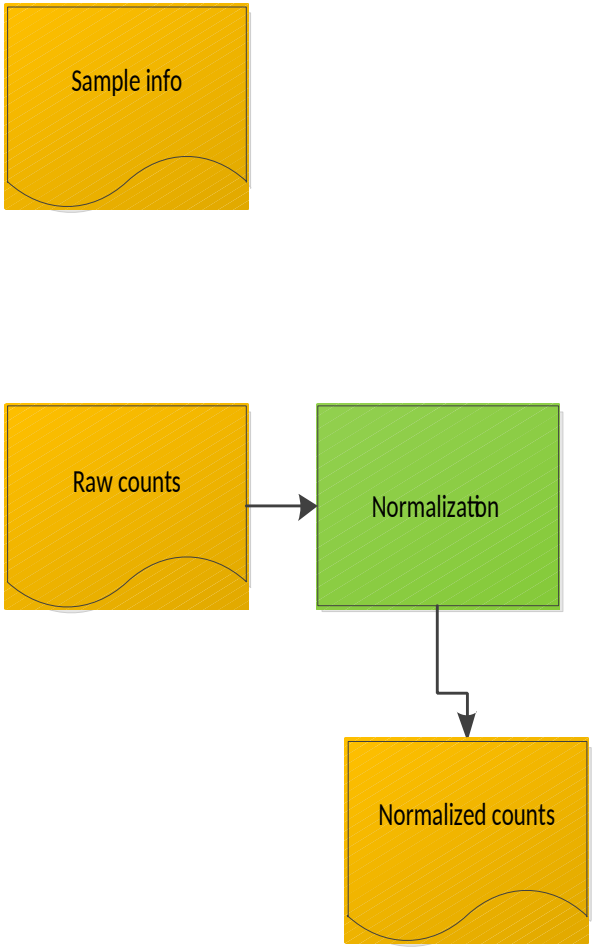
Inputs

- The scale
  - Exon level -> DEXSeq
  - Gene level
  - Isoform level

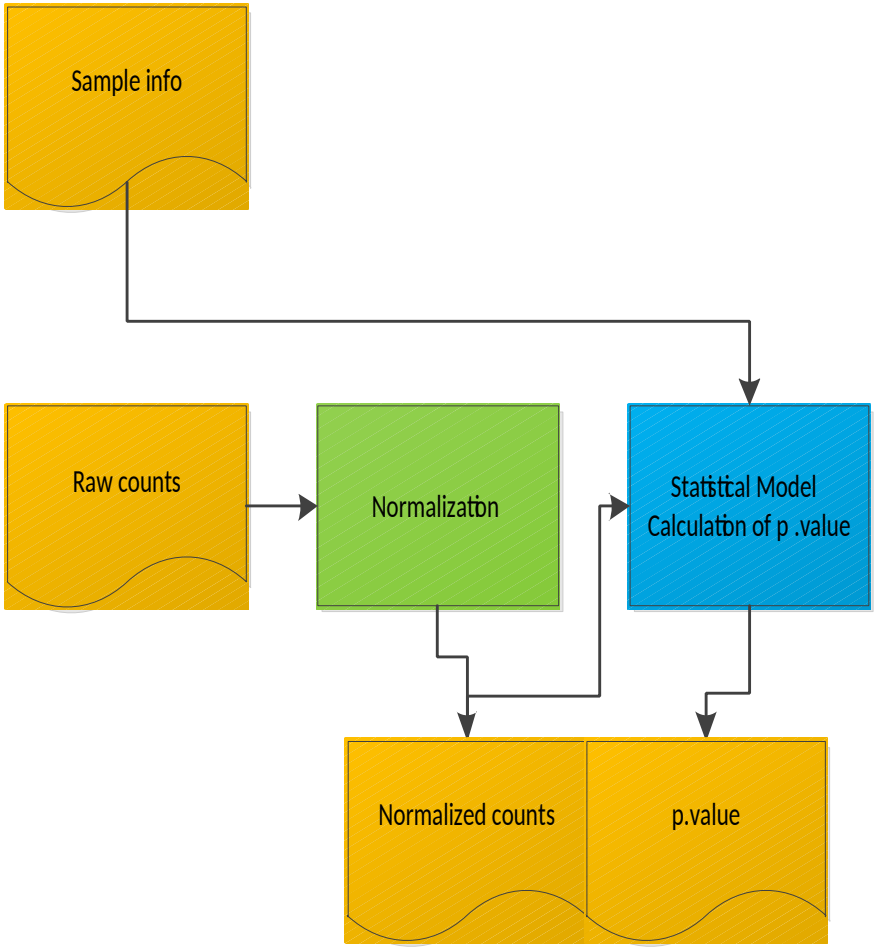


THE WORKFLOW

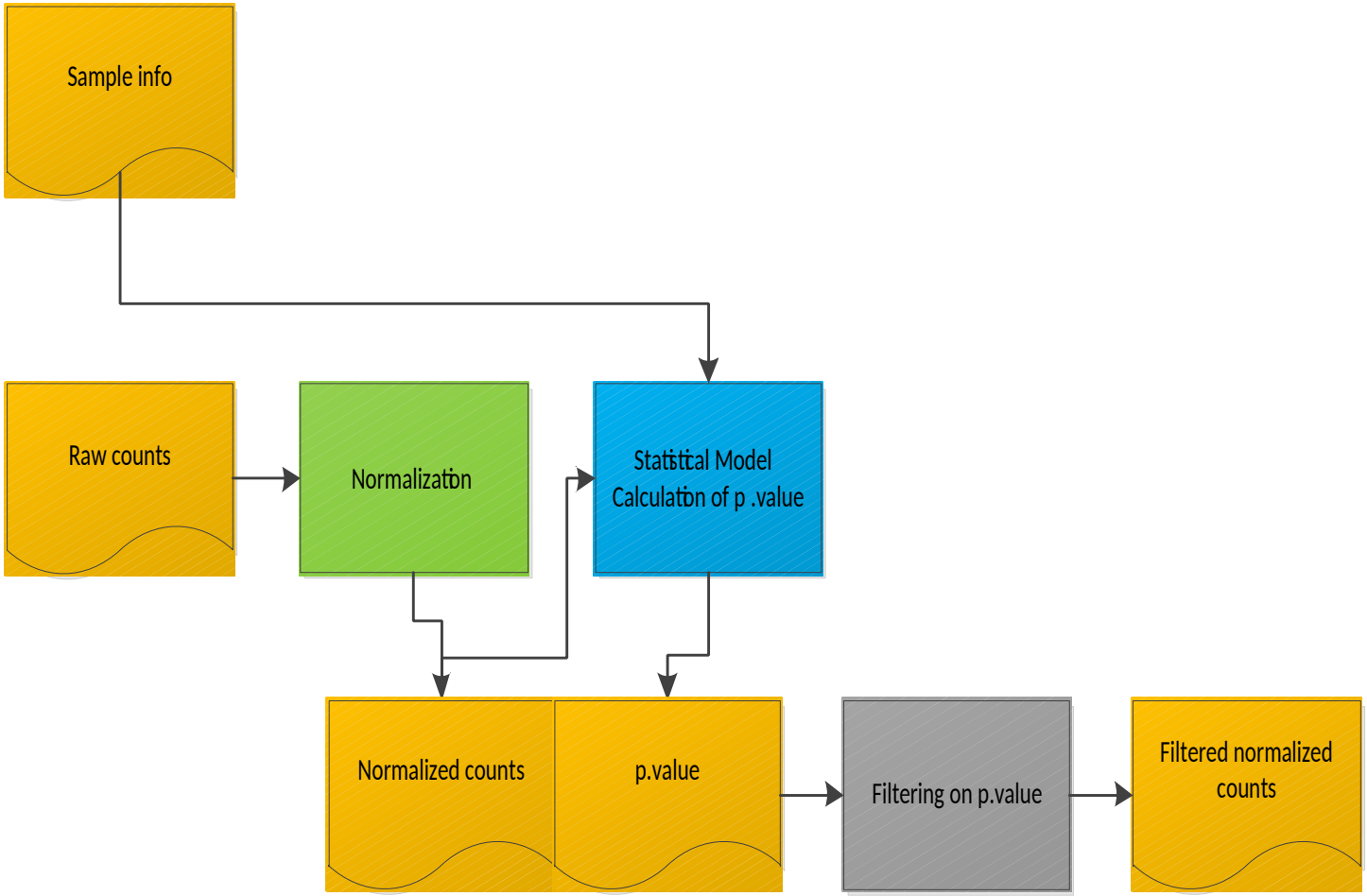
The workflow



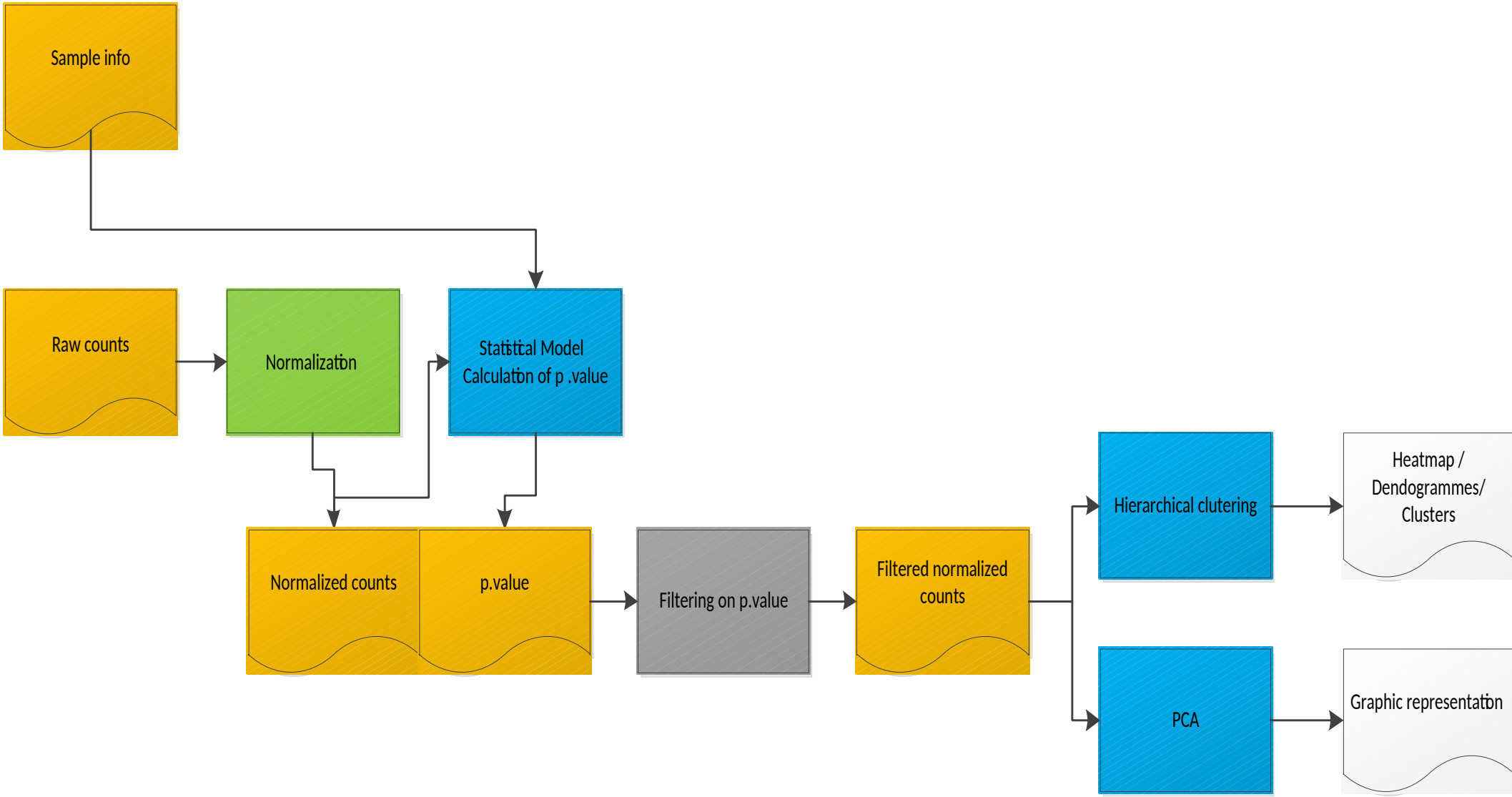
The workflow



The workflow

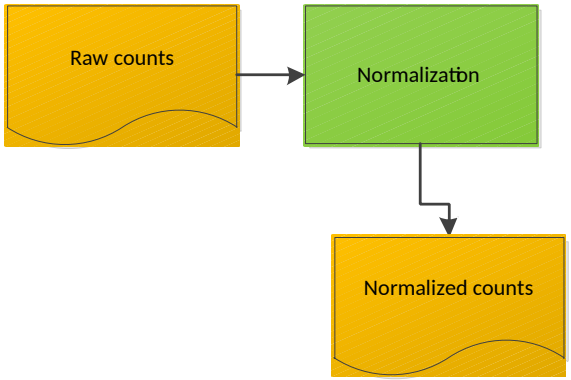


The workflow



NORMALIZATION

Normalization



Normalization

WARNING

- It is important to recognize that the number of reads which overlap a gene is not a direct measure of the gene's expression.

  => Genes length bias

  => One effect of this bias is to reduce the ability to detect differential expression among shorter genes simply from the lack of coverage since the power of statistical tests involving count data decreases with lower number of count

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci and Doron

Normalization

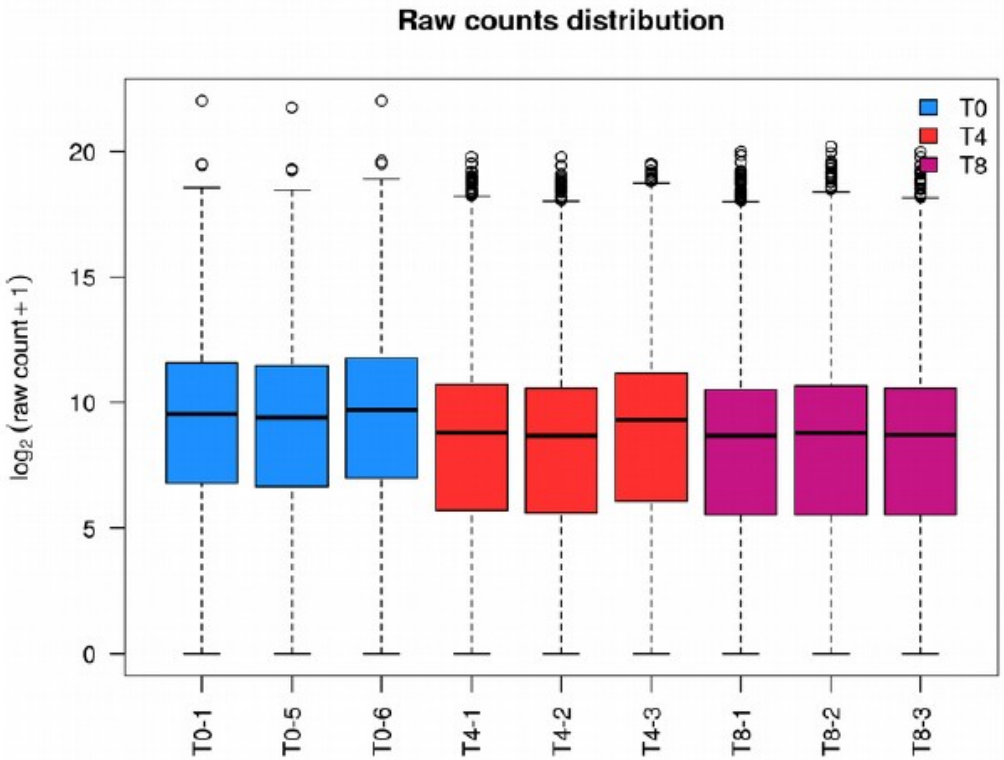
Why performing normalisation ?

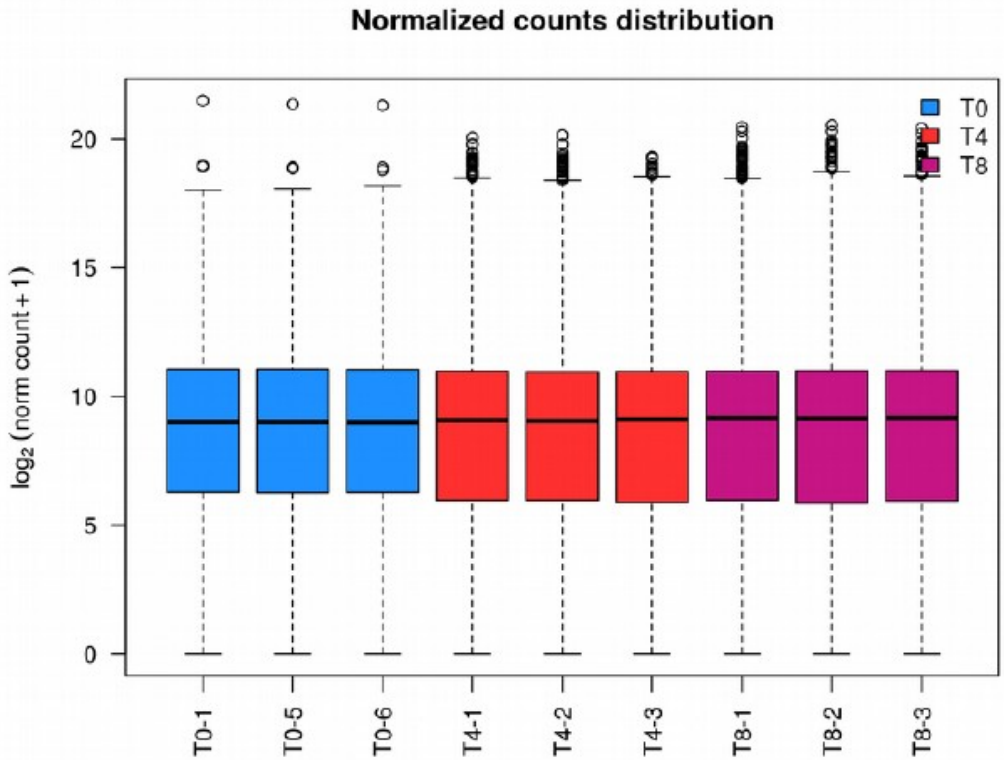
- Between-sample ⎵ compare a gene in different sample
  - Depth of sequencing == library size
  - Sampling bias during the libraries construction == batch effect
  - Presence of majority fragments == saturation
  - Sequence composition du to PCR-amplification step (GC content)

- Within-sample ⎵ compare genes in a sample
  - Gene length
  - Sequence composition (GC content)

Normalization

Why





Normalization

How ?

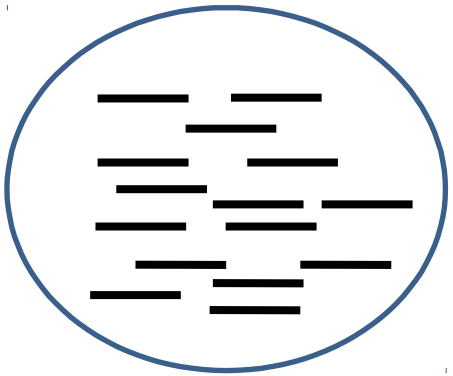
- Between-lane ⎵ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using housekeeping genes

- Within-lane ⎵ compare genes in a sample
  - Normalize on gene lengths

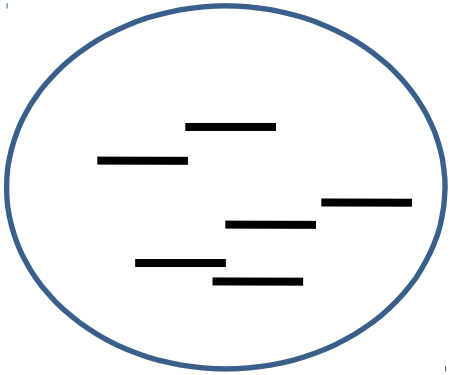
Normalization

How ?

- Between-lane ␣ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using housekeeping genes
    - When :



Condition A

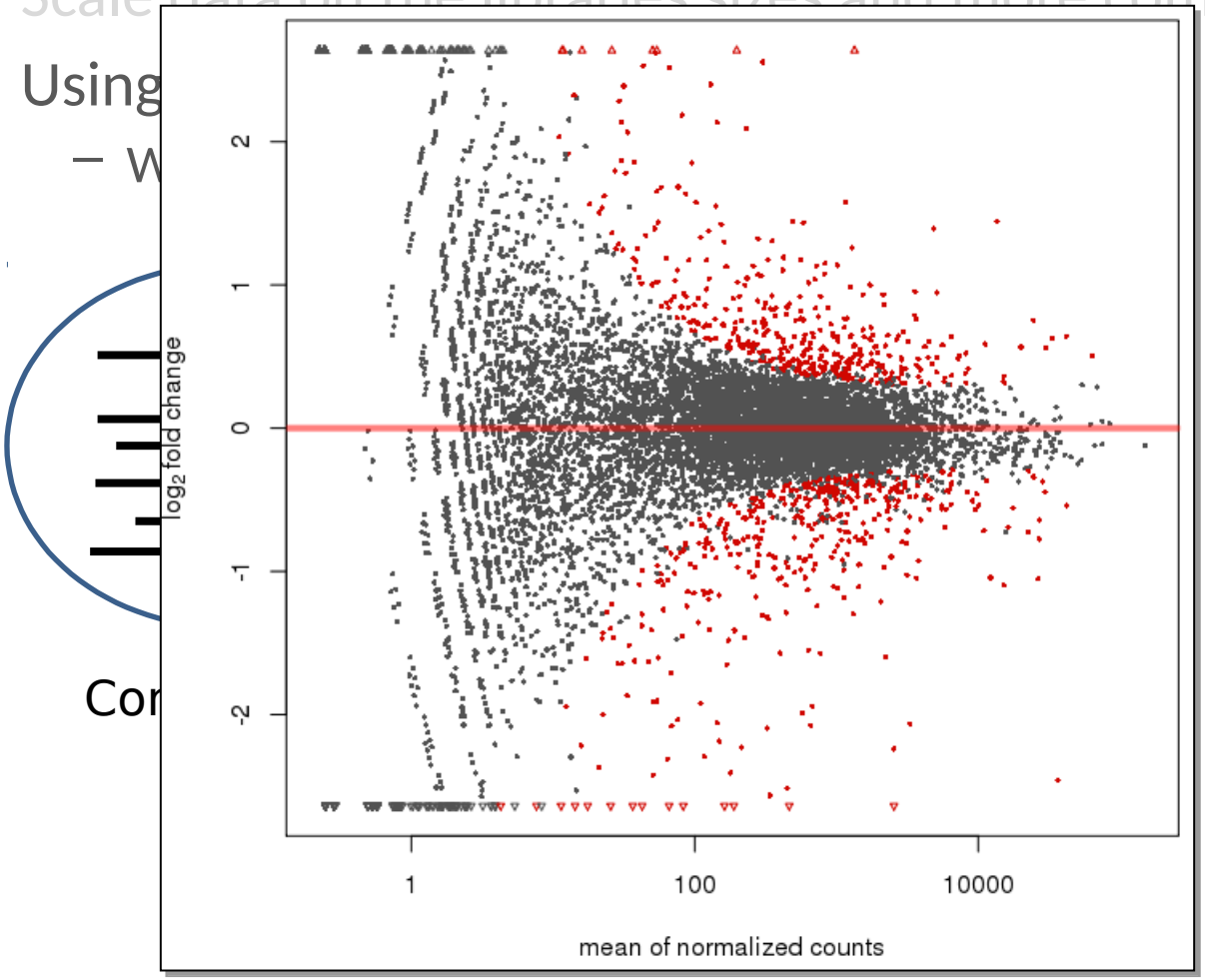


Condition B

Normalization

How ?

- Between-lane ⎵ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using
    - W

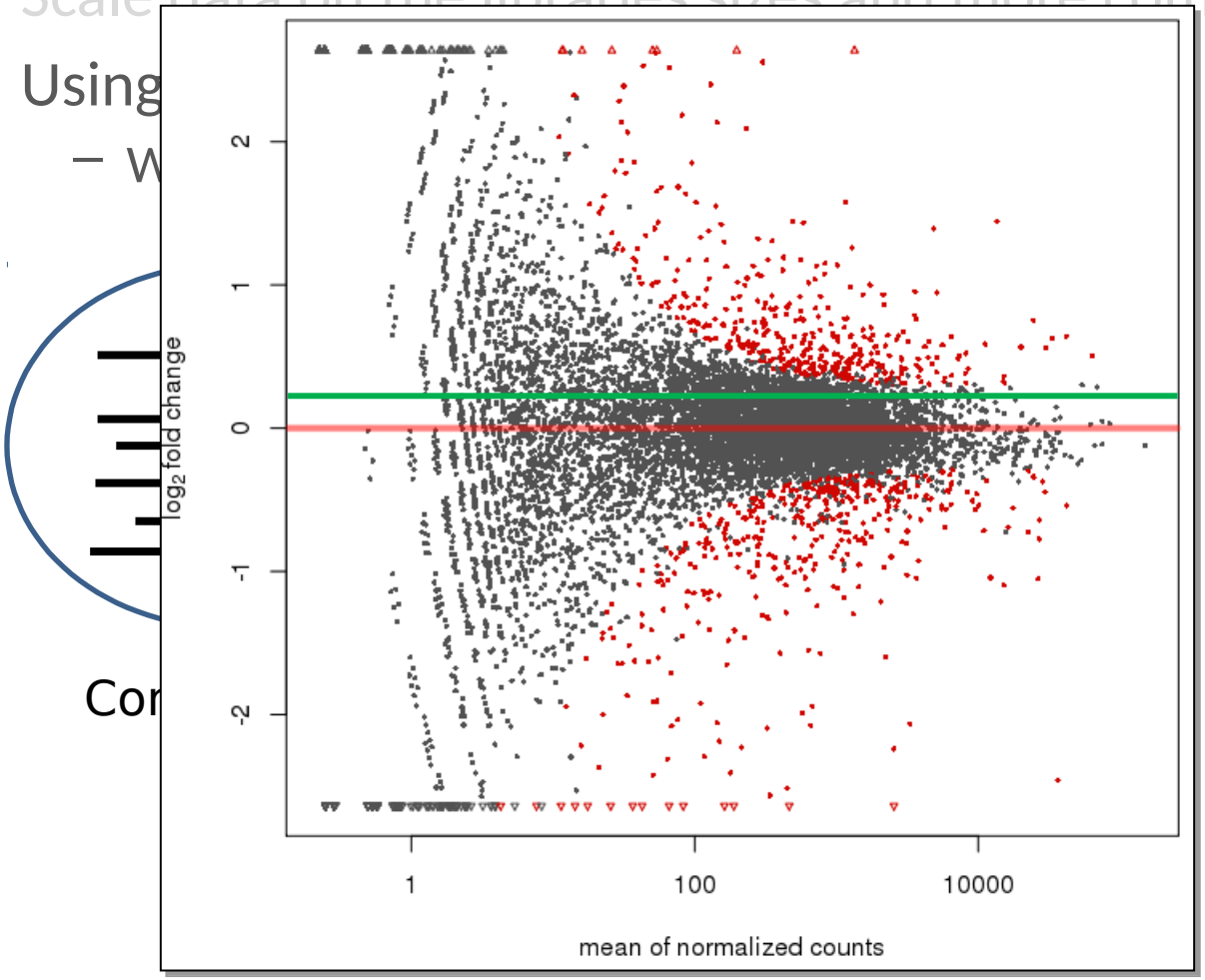


Cor

Normalization

How ?

- Between-lane ⎵ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using
    - W

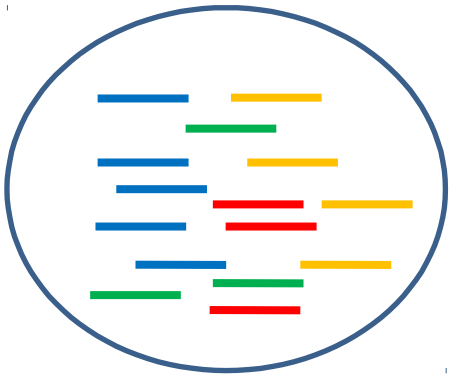


Cor

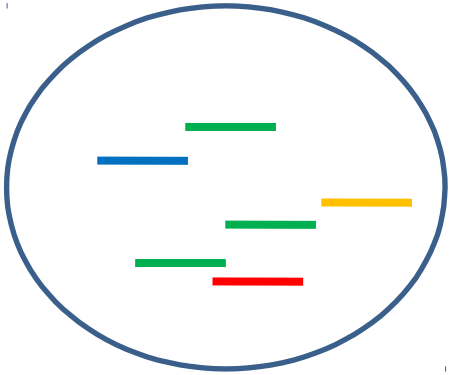
Normalization

How ?

- Between-lane ⎵ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using housekeeping genes
    - When :



Condition A



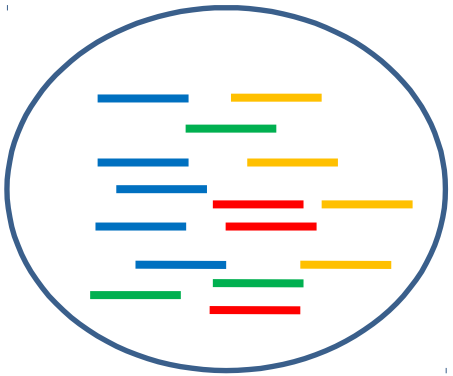
Condition B

- Examples : actin, GAPDH, ubiquitin, HSP90, Histone, rRNA, tRNA

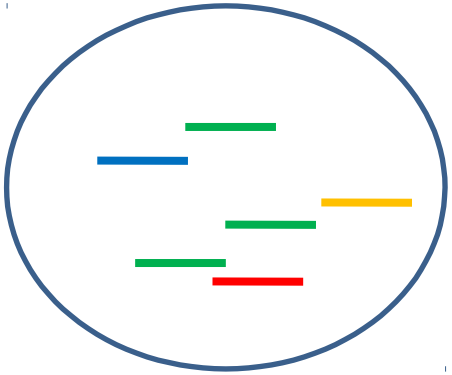
Normalization

How ?

- Between-lane ⎵ compare a gene in different sample
  - Scale data on the libraries sizes and more complex methods
  - Using housekeeping genes
    - When :



Condition A



Condition B

It depends

    - Examples : actin, GAPDH, ubiquitin, HSP90, Histone, rRNA, tRNA

Normalization

Normalization methods

Total Counts (TC)

- Motivation: greater lane sequencing depth => greater counts

- Assumption: read counts are proportional to expression level and sequencing depth (same RNAs in equal proportion)

- Method: divide transcript read count by total number of reads

Problem:

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization methods

Upper Quartile normalization (UQ) or Median (Med)

- Motivation: total read count is strongly dependent on a few highly expressed transcripts

- Assumption: read counts are proportional to expression level and sequencing depth

- Method: divide transcript read count by, e.g., upper quartile

- Problem: Sensitive to the presence of majority genes

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

Normalization methods

Reads Per Kilobase per Million mapped reads (RPKM / FPKM)

- Motivation: greater lane sequencing depth and gene length => greater counts whatever the expression level
  Allow comparaison of expression of different genes in a sample
- Assumption: read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportion)
- Method: divide gene read count by total number of reads (in million) and gene length (in kb)

Problem:

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/Julie_AUBERT.pdf

Normalization

- RPFM / FPFM
  - Pro
    - Simple, easy to understand
    - Comparable between different genes within the same dataset
  - Cons
    - Small changes in highly expressed genes (especially differences in rRNA contamination) cause a global shift in all other values
    - Small changes across lowly expressed genes (especially differences in DNA contamination) cause differences across a wide number of genes.
    - Mixing of noise levels
    - Noise is generally linked to the number of observations
    - The same RPKM value could come from
      - A small lowly observed gene with high noise
      - A large well observed gene with low noise

Normalization

- RPFM / FPFM
  - Pro
    - S
    - C
  - Con
    - S
    - S
    - N
    - N
    - The same
      - A sma
      - A larg

Conspiracy

Normalization

Normalization methods

The Effective Library Size concept : TMM (edgeR) and DESeq

- Motivation: Different biological conditions express different RNA repertoires, leading to different total amounts of RNA

- Assumption: A majority of transcripts is not differentially expressed

As many down- as up-regulated genes

- Method: Minimizing effect of (very) majority sequences

Problem:

http://www.crrs.fr/inee/recherche/fichiers/EPEGE/Communications/

Normalization

Normalization methods

The Effective Library Size concept : TMM (edgeR) and DESeq

- Motivation: Different biological conditions express different RNA repertoires, leading to different total amounts of RNA

- Assumption: A majority of transcripts is not differentially expressed

As many down- as up-regulated genes

- Method: Minimizing effect of (very) majority sequences

Problem:

http://www.cnrs.fr/inee/recherche/fichiers/EPEGE/Communications/

Normalization

The Effective Library Size

- TMM / edgeR

  Uses the number of mapped reads (i. e., count table column sums) and estimates an additional normalization factor to account for sample-specific effects (e. g., diversity); these two factors are combined and used as an offset in the NB model.

- DESeq

  Defines a virtual reference sample by taking the median of each gene's values across samples, and then computes size factors as the median of ratios of each sample to the reference sample.

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor
Simon Anders, Davis J. McCarthy, Yunshen Chen, Michal Okoniewski, Gordon K.Smyth, Wolfgang Huber & Mark D

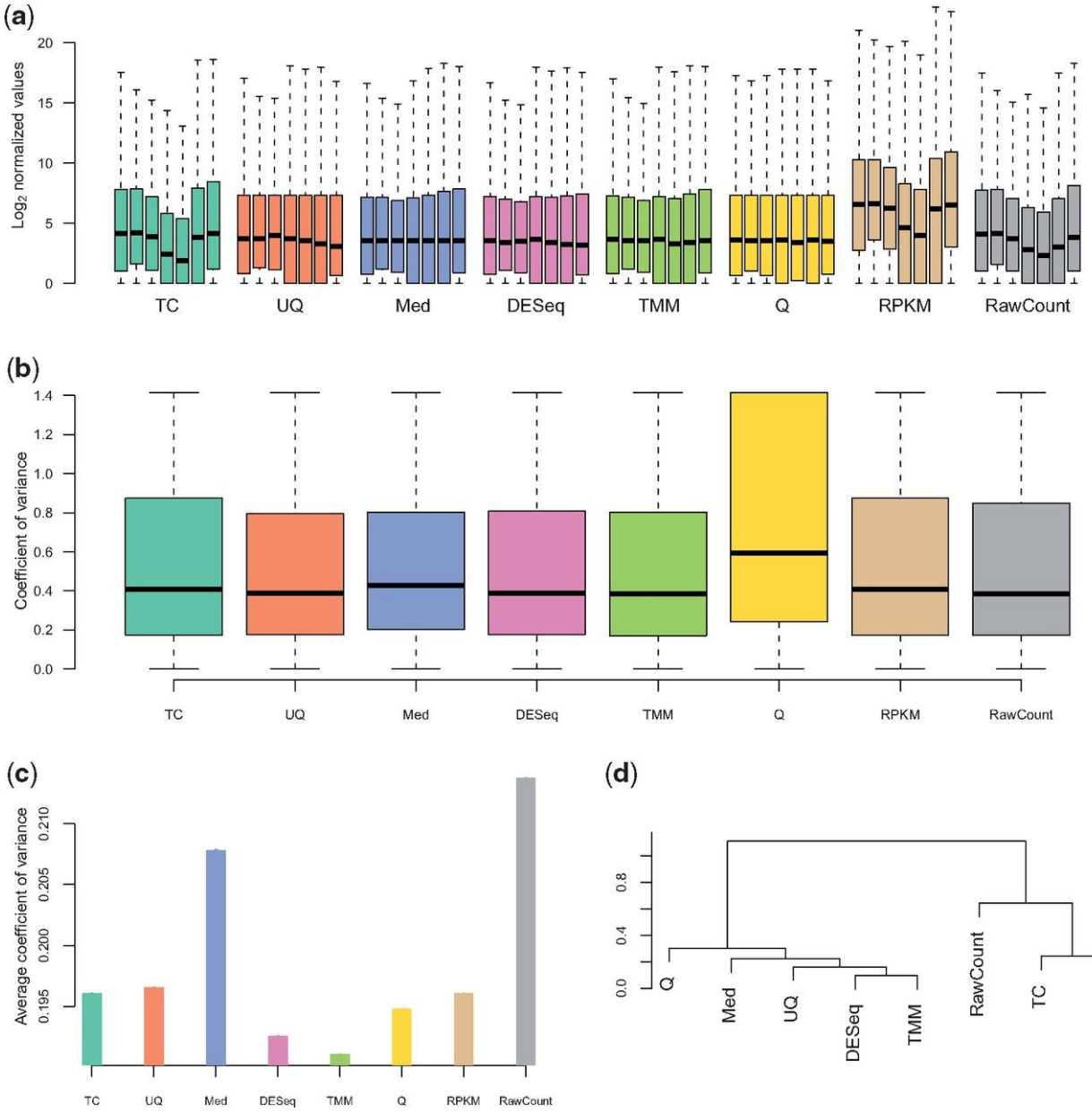


Figure 1:

Comparison of normalization methods for real data.

(A) Boxplots of log2(counts + 1) for all conditions and replicates in the M. musculus data, by normalization method.

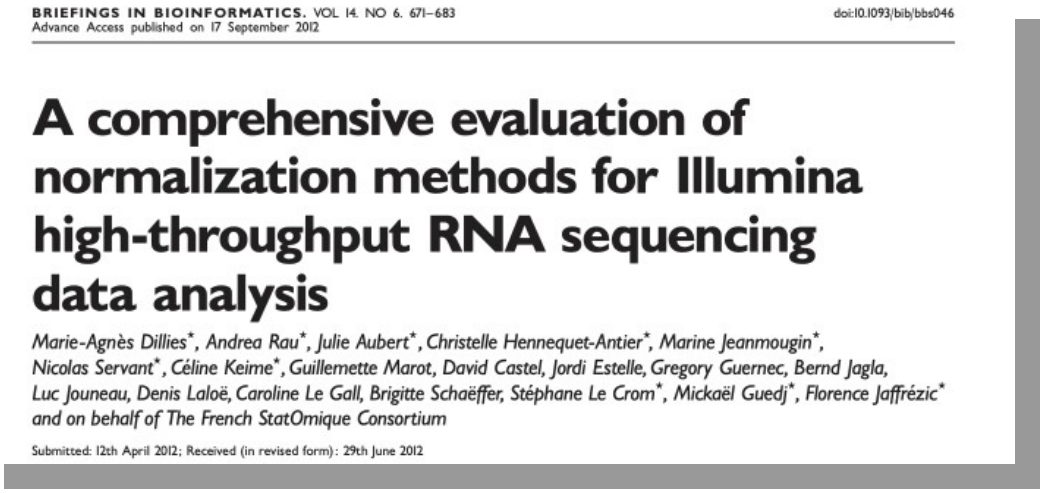
(B) Boxplots of intra-group variance for one of the conditions in the M. musculus data, by normalization method.

(C) Analysis of housekeeping genes for the H. sapiens data.

(D) Consensus dendrogram of differential analysis results, using the DESeq Bioconductor package, for all normalization methods across the four datasets under consideration.

MA Dillies, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform (2013) 14 (6): 671-683 :480

Normalization



BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012
doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

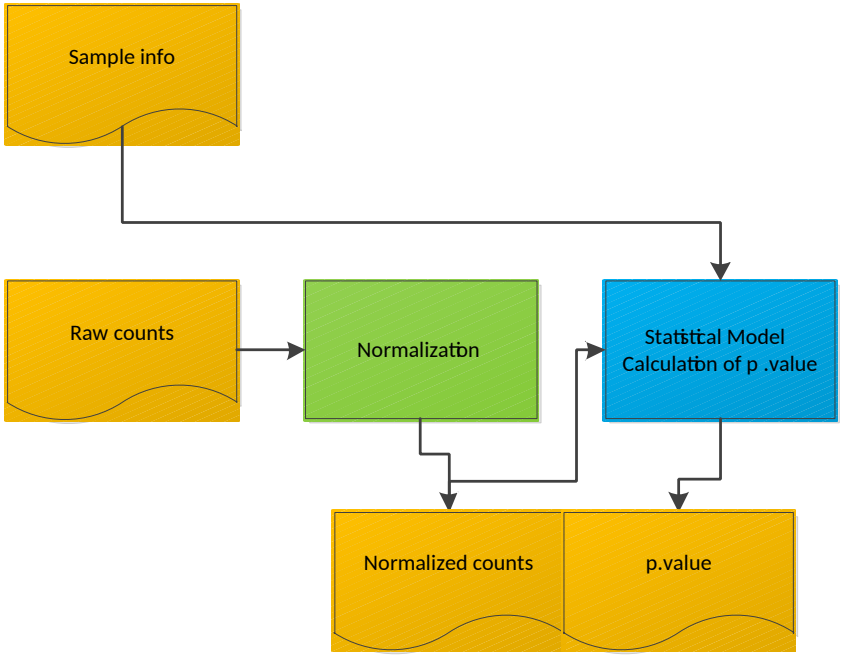
"Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions..."

Dillies et al., Brief Bioinf, 2013. doi:10.1093/bib/bbs046

Models

STATISTICS

Statistics


Sample info
Raw counts
Normalization
Statistical Model
Calculation of p .value
Normalized counts
p.value

- Rappel

$$\mathbb{P}(X=k) = \int_0^{+\infty} \frac{\lambda^k e^{-\lambda}}{k!} \frac{\lambda^{r-1} e^{-\lambda/\theta}}{\Gamma(r)\theta^r} \mathrm{d}\lambda$$

$$\begin{aligned}
\mathbb{P}(X_n \leq k) &= I_p(n, k+1) \\
&= 1 - I_{1-p}(k+1, n) \\
&= 1 - I_{1-p}((k+n)-(n-1), (n-1)+1) \\
&= 1 - \mathbb{P}(Y_{k+n} \leq n-1) \\
&= \mathbb{P}(Y_{k+n} \geq n)
\end{aligned}$$

$$\mathbb{P}(X=k) = \int_0^{+\infty} \frac{\lambda^{r+k-1} e^{-\lambda}}{\Gamma(r)\theta^r} \left(\frac{1+\theta}{\theta}\right)^{r+k} \mathrm{d}t$$

$$f(k; r, p) = \int_0^{\infty} f_{\mathrm{Poisson}(\lambda)}(k) \cdot f_{\mathrm{Gamma}\left(r, \frac{p}{1-p}\right)}(\lambda)\, \mathrm{d}\lambda$$

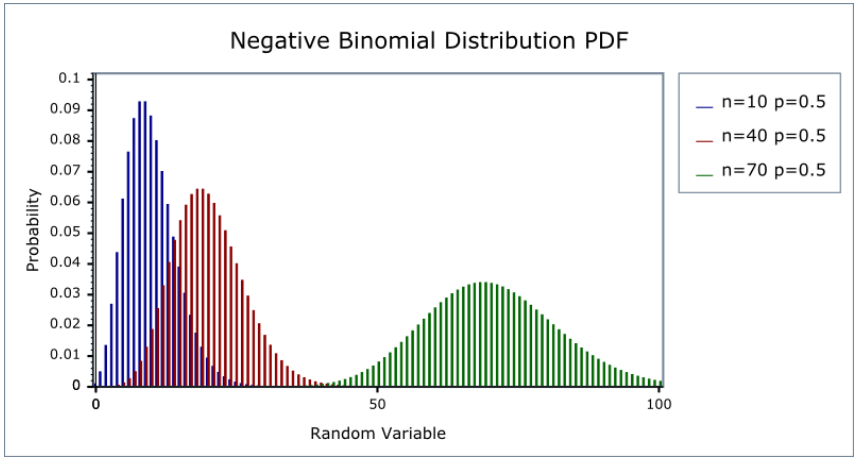
$$= \int_0^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r \Gamma(r)}\, \mathrm{d}\lambda$$

$$= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)} \int_0^{\infty} \lambda^{r+k-1} e^{-\lambda/p}\, \mathrm{d}\lambda$$

$$= \frac{(1-p)^r p^{-r}}{k!\,\Gamma(r)}\, p^{r+k}\, \Gamma(r+k)$$

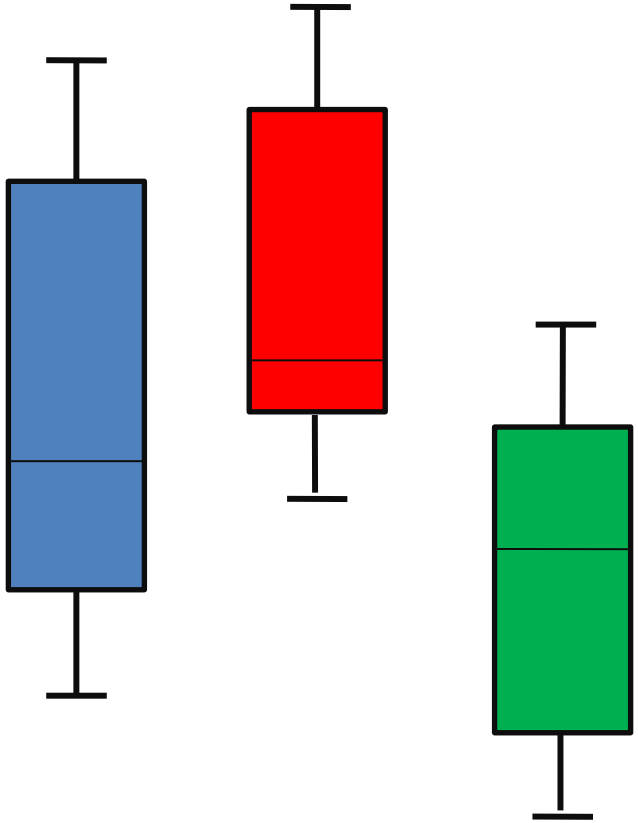
$$= \frac{\Gamma(r+k)}{k!\,\Gamma(r)} (1-p)^r p^k.$$

$$\begin{aligned}
\binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-(k+1))!} \\
&= \frac{n!(k+1)}{k!(k+1)(n-k)!} + \frac{n!(n-k)}{(k+1)!(n-k-1)!(n-k)} \\
&= \frac{n!(k+1)}{(k+1)!(n-k)!} + \frac{n!(n-k)}{(k+1)!(n-k)!} \\
&= \frac{n!((k+1)+(n-k))}{(k+1)!(n-k)!} \\
&= \frac{n!(n+1)}{(k+1)!(n-k)!} \\
&= \frac{(n+1)!}{(k+1)!((n+1)-(k+1))!} \\
&= \binom{n+1}{k+1}.
\end{aligned}$$



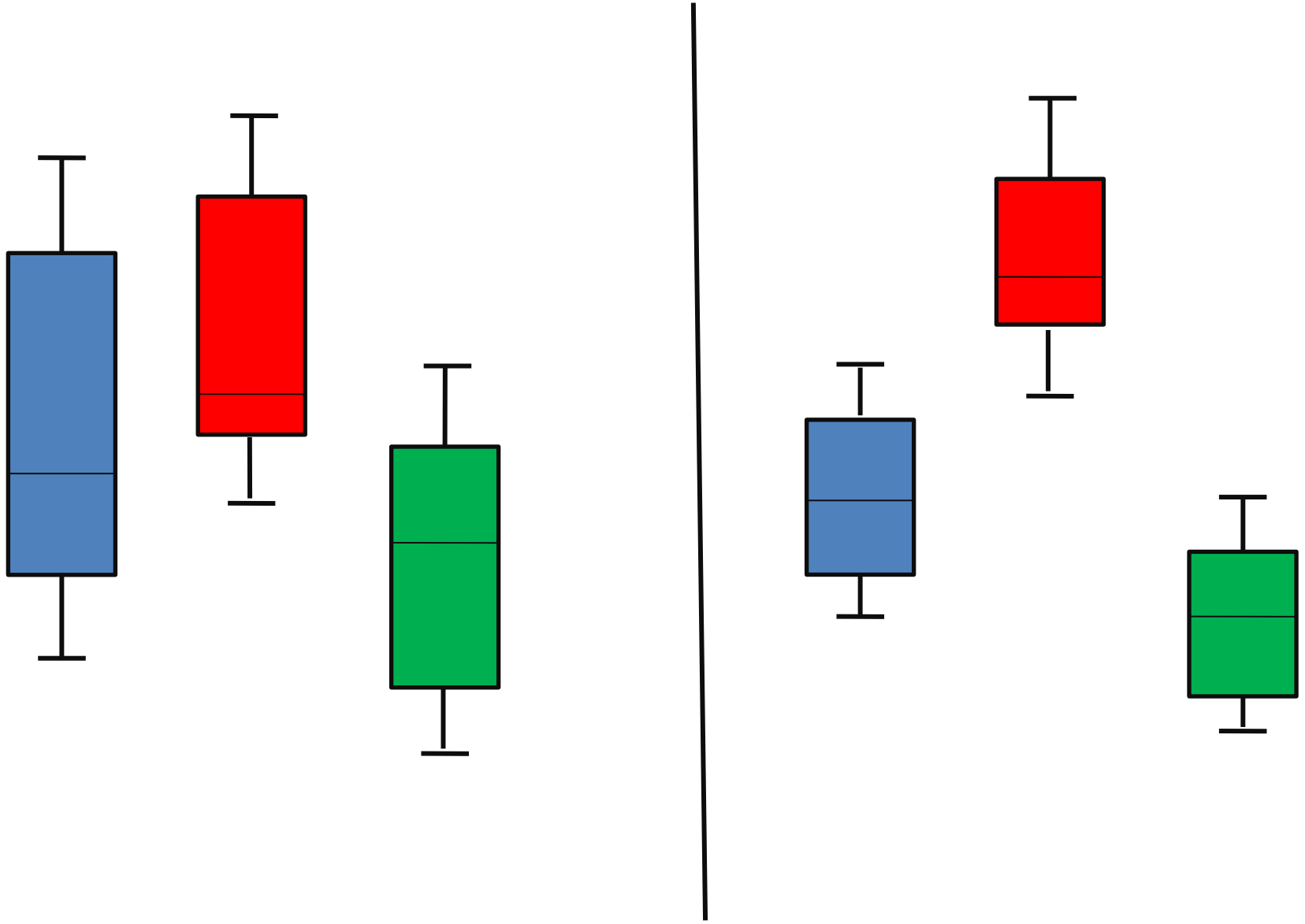
Statistics

Significant?



Statistics

Significant?



The model

Poisson distribution
- Motivation: Poisson distribution appears when things are counted
- Assumption: mean and variance are the same
- Method: Poisson distribution has only one parameter λ (expected number of reads)

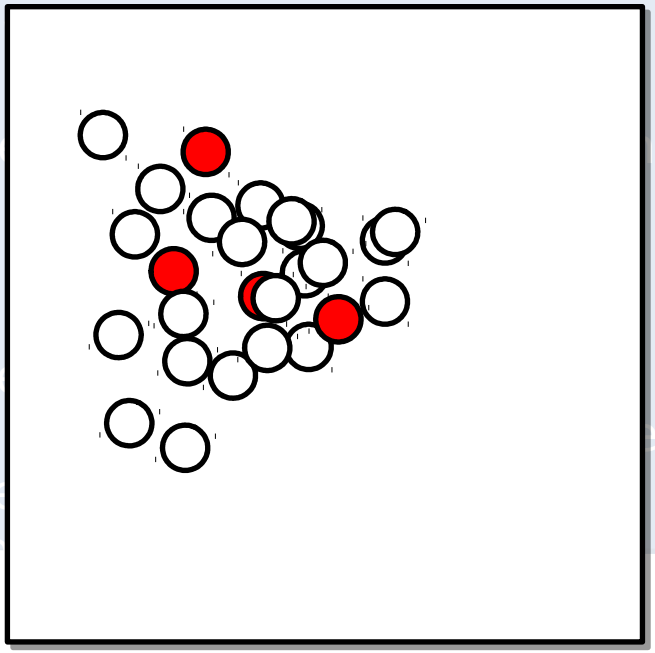
- Problem:
  Good distribution for technical replicates
  But biological variability of RNA-seq count data cannot be capture using the Poisson distribution because data present overdispersion (i.e. variance of counts larger than mean)

http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf

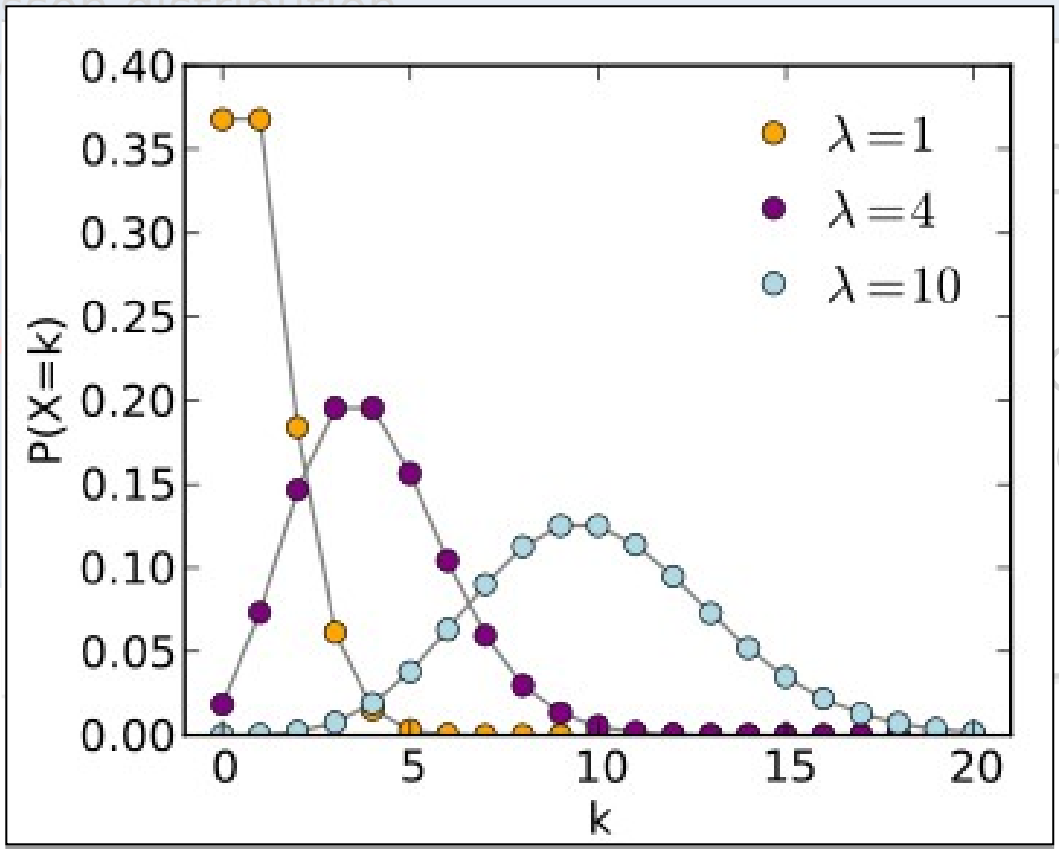
The model

Poisson distribution

- M… …on appears
- A… …nce are the
- M… …has only on… …umber of rea…
- Pr…
- …ical replicat…
- … RNA-seq co… …e using the Pois… …se data pres…
- …ger than me…



The model



Mean λ
Variance λ

Wikipedia

The model

- Consider this situation:
  - Several flow cell lanes are filled with aliquots of the same prepared library.
  - The concentration of a certain transcript species is exactly the same in each lane.
  - We get the same total number of reads from each lane.

- For each lane, count how often you see a read from the transcript. Will the count all be the same?

- No! Even for equal concentration, the counts will vary. This theoretically unavoidable noise is called shot noise.

The model

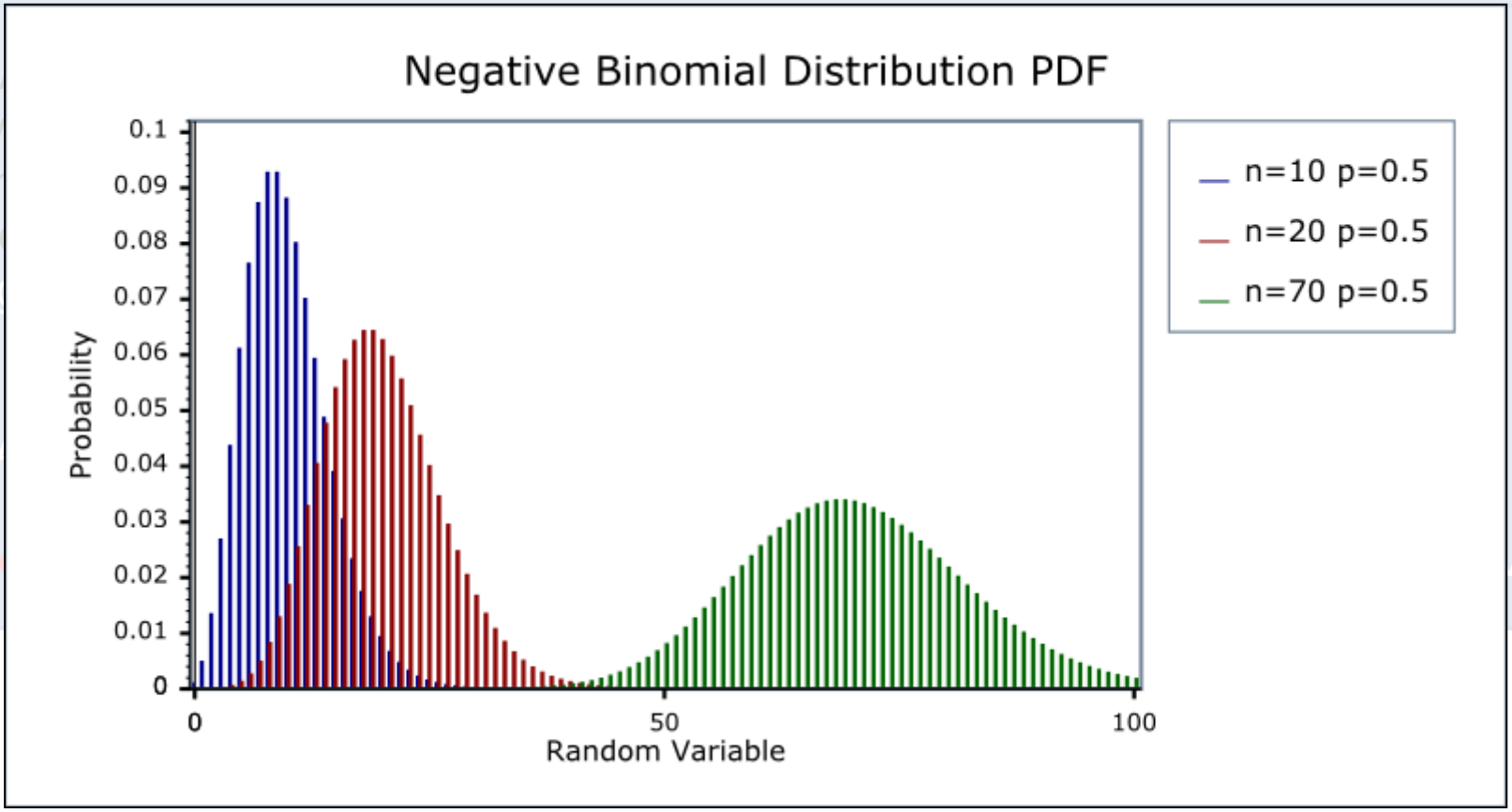
Negative Binomial (NB): edgeR and DESeq

- Motivation: distribution takes into account Overdispersion
- Assumption:
- Method: NB is a two-parameter distribution
  - Origin: $Y \sim NB(p, m)$
    - Y ... number of successes in a sequence of Bernoulli trials with probability p before r failures occur
  - RNASeq case: λ (mean) and φ (overdispersion)

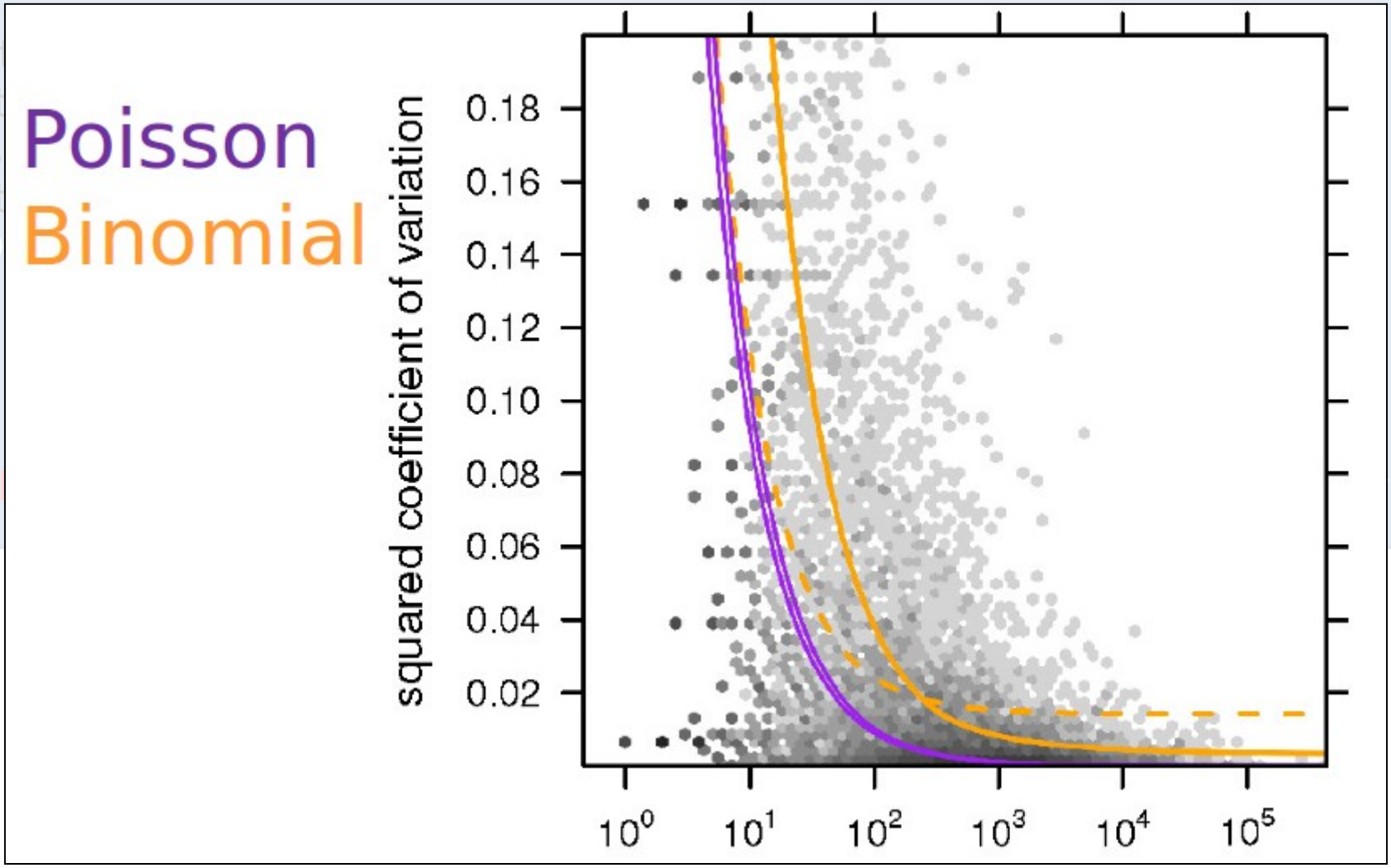
- Problem: φ_i / gene cannot be estimated due to the small number of individuals

http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf

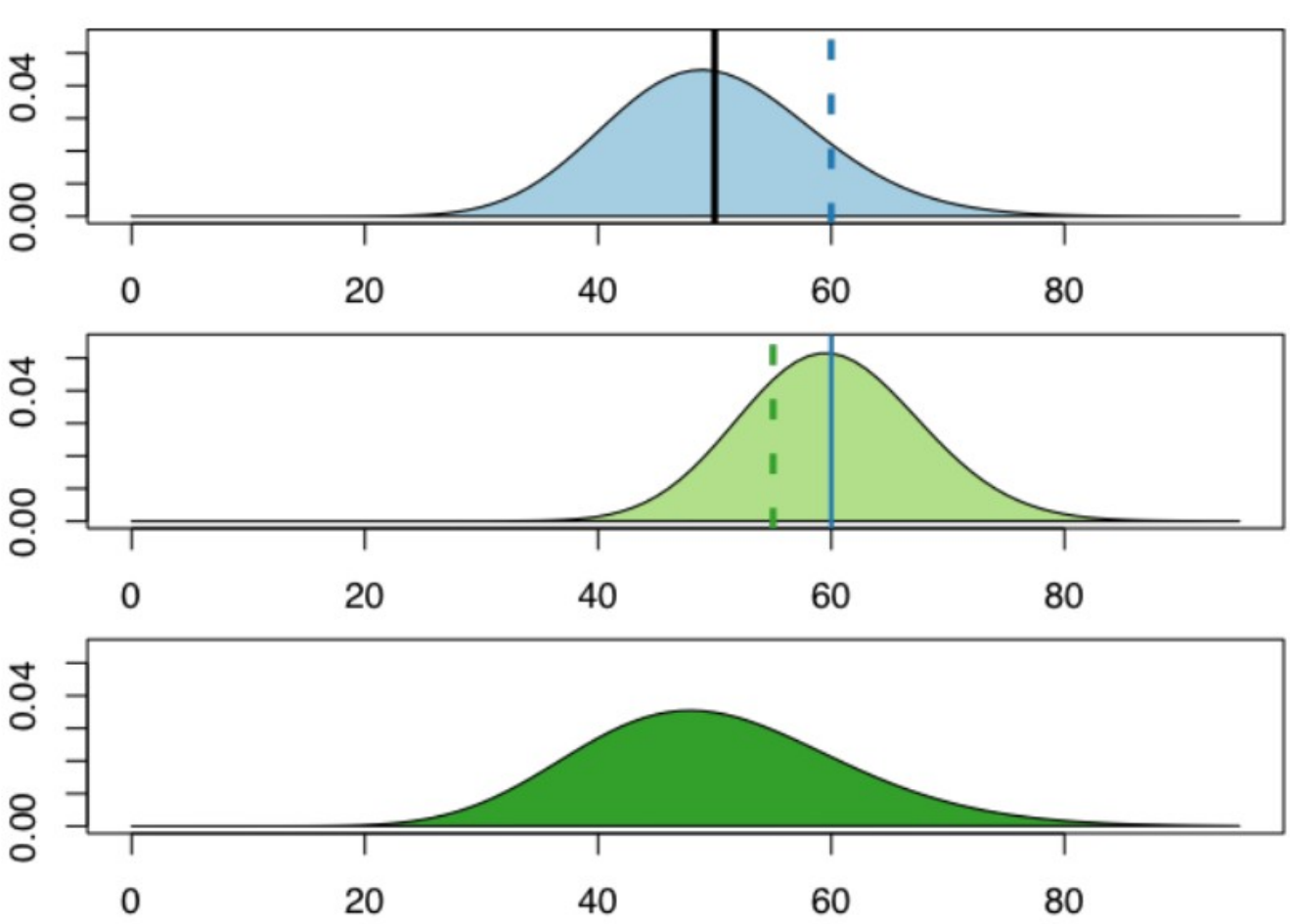
The model


Negative Binomial Distribution PDF
Probability
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
50
100
Random Variable
n=10 p=0.5
n=20 p=0.5
n=70 p=0.5

The model


Poisson
Binomial
squared coefficient of variation
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
10^0
10^1
10^2
10^3
10^4
10^5

The model



Biological sample with mean μ and variance *v*

Poisson distribution with mean *q* and variance *q*.

Negative binomial with mean *μ* and variance *q*+*v*.

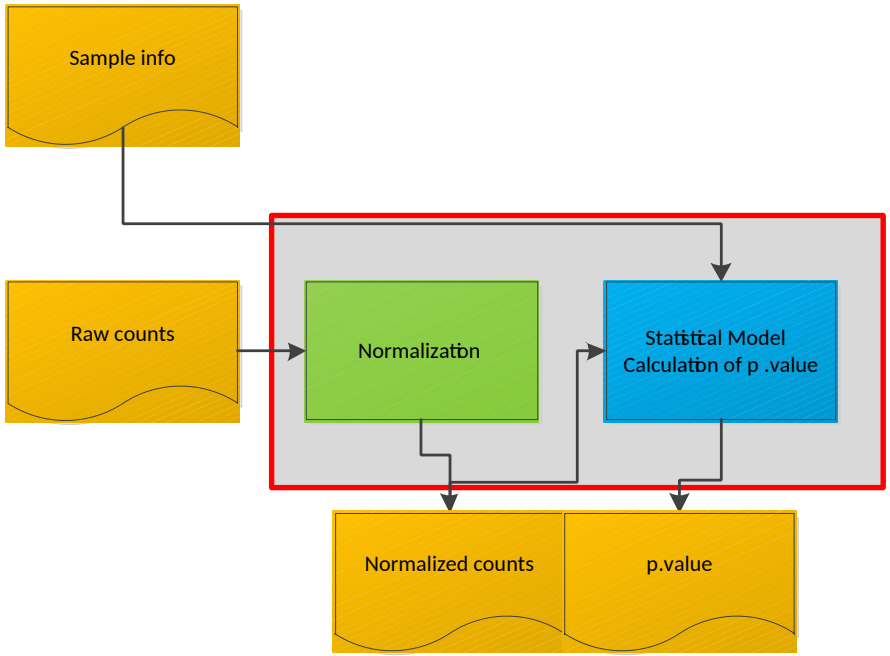
The model

Negative Binomial (NB): edgeR and DESeq

- Motivation: distribution takes into account Overdispersion
- Assumption:
- Method: NB is a two-parameter distribution
  - Origin: $Y \sim NB(p, m)$
    - Y ... number of successes in a sequence of Bernoulli trials with probability p before r failures occur
  - RNASeq case: λ (mean) and φ (overdispersion)

- Problem: φ_i / gene cannot be estimated due to the small number of individuals

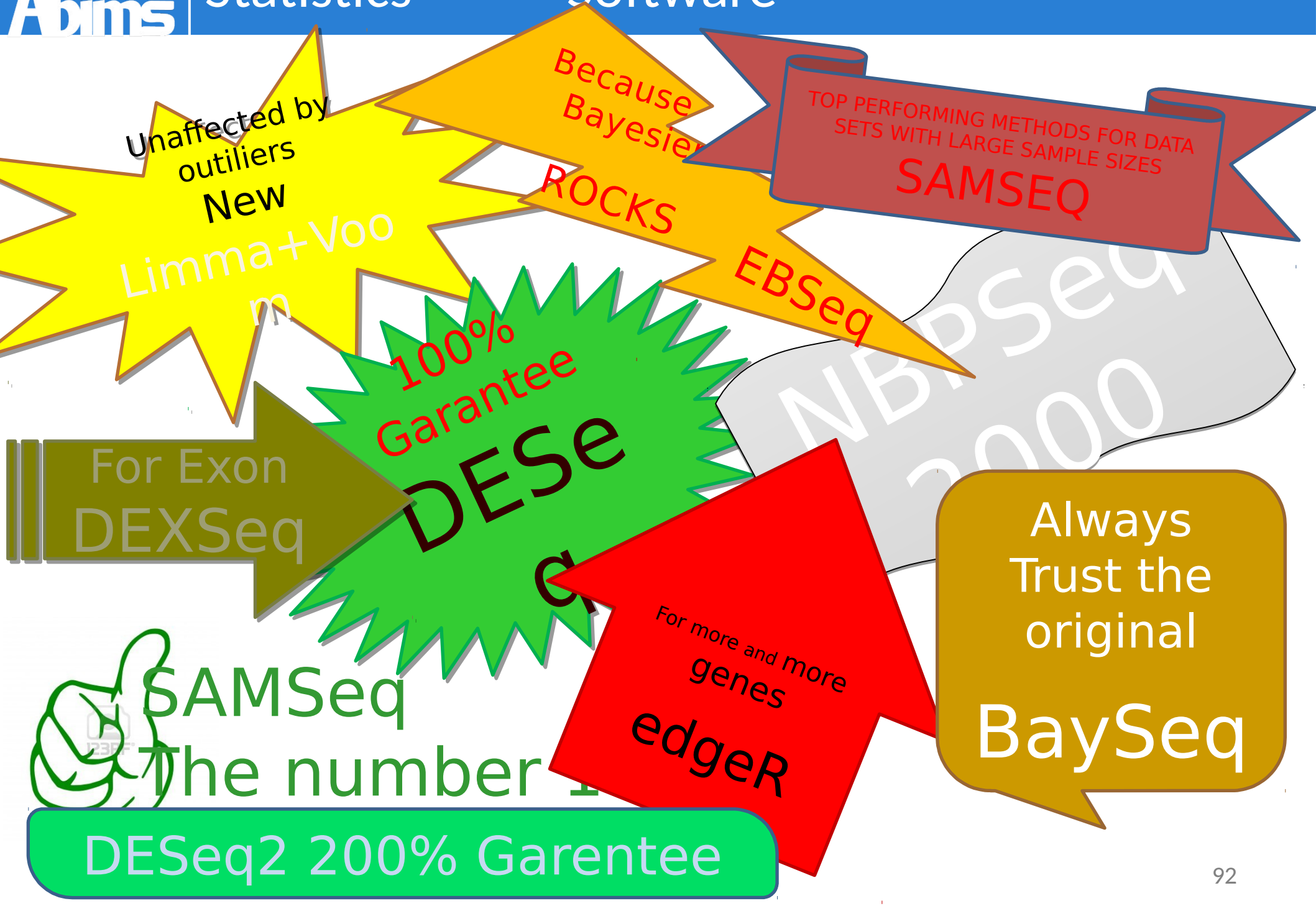
http://smpgd2012.univ-lyon1.fr/IMG/pdf/NGS/gonzalez-smpgd2012.pdf



Software

STATISTICS

Statistics Software



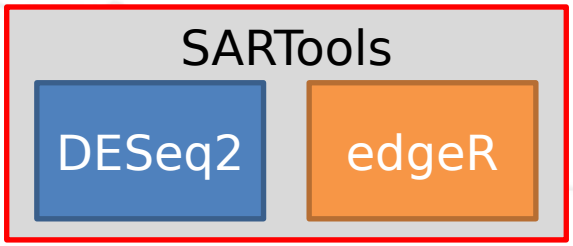
- In this paper, we have evaluated and compared eleven methods for differential expression analysis of RNA-seq data. Table 2 summarizes the main findings and observations. No single method among those evaluated here is optimal under all circumstances, and hence the method of choice in a particular situation depends on the experimental conditions. Among the methods evaluated in this paper, those based on a variance-stabilizing transformation combined with limma (i.e., voom+limma and vst+limma) performed well under many conditions, were relatively unaffected by outliers and were computationally fast, but they required at least 3 samples per condition to have sufficient power to detect any differentially expressed genes. As shown in the supplementary material (Additional file 1), they also performed worse when the dispersion differed between the two conditions. The non-parametric SAMseq, which was among the top performing methods for data sets with large sample sizes, required at least 4-5 samples per condition to have sufficient power to find DE genes. For highly expressed genes, the fold change required for statistical significance by SAMseq was lower than for many other methods, which can potentially compromise the biological significance of some of the statistically significantly DE genes. The same was true for ShrinkSeq, which however has an option for imposing a fold change requirement in the inference procedure.
- Small sample sizes (2 samples per condition) imposed problems also for the methods that were indeed able to find differentially expressed genes, there leading to false discovery rates sometimes widely exceeding the desired threshold implied by the FDR cutoff. For the parametric methods this may be due to inaccuracies in the estimation of the mean and dispersion parameters. In our study, TSPM stood out as the method being most affected by the sample size, potentially due to the use of asymptotic statistics. Even though the development goes towards large sample sizes, and barcoding and multiplexing create opportunities to analyze more samples at a fixed cost, as of today RNA-seq experiments are often too expensive to allow extensive replication. The results conveyed in this study strongly suggest that the differentially expressed genes found between small collections of samples need to be interpreted with caution and that the true FDR may be several times higher than the selected FDR threshold.
- DESeq, edgeR and NBPSeq are based on similar principles and showed, overall, relatively similar accuracy with respect to gene ranking. However, the sets of significantly differentially expressed genes at a pre-specified FDR threshold varied considerably between the methods, due to the different ways of estimating the dispersion parameters. With default settings and for reasonably large sample sizes, DESeq was often overly conservative, while edgeR and in particular NBPSeq often were too liberal and called a larger number of false (and true) DE genes. In the supplementary material (Additional file 1) we show that varying the parameters of edgeR and DESeq can have large effects on the results of the differential expression analysis, both in terms of the ability to control type I error rates and false discovery rates and in terms of the ability to detect the truly DE genes. These results also show that the recommended parameters (that are used in the main paper) are indeed well chosen and often provide the best results.
- EBSeq, baySeq and ShrinkSeq use a different inferential approach, and estimate the posterior probability of being differentially expressed, for each gene. baySeq performed well under some conditions but the results were highly variable, especially when all DE genes were upregulated in one condition compared to the other. In the presence of outliers, EBSeq found a lower fraction of false positives than baySeq for large sample sizes, while the opposite was true for small sample sizes.

- limma (i.e., voom+limma and vst+limma)
  - unaffected by outliers
  - but they required at least 3 samples per condition
- SAMseq, ShrinkSeq (The non-parametric)
  - top performing methods for data sets with large sample sizes
  - required at least 4-5 samples per condition
  - fold change required for statistical significance was lower ␣ compromise the biological significance
  - Small sample sizes inaccuracies in the estimation of the mean and dispersion parameters
- TSPM
  - most affected by the sample size
- DESeq, edgeR and NBPSeq
  - showed, overall, relatively similar accuracy with respect to gene ranking
  - recommended parameters well chosen and often provide the best results
  - pre-specified FDR threshold varied considerably between the methods
  - DESeq : overly conservative
  - edgeR, NBPSeq : too liberal and called a larger number of false (and true) DE genes.
  - edgeR, DESeq : varying the parameters of can have large effects on the results
- EBSeq, baySeq and ShrinkSeq (posterior probability)
  - baySeq performed well under some conditions ; results were highly variable, especially when all DE genes were upregulated in one condition
  - EBSeq In the presence of outliers, found a lower fraction of false positives for large sample sizes not fot small sample sizes
  - baySeq In the presence of outliers, found a lower fraction of false positives true for small sample sizes not fot large sample sizes

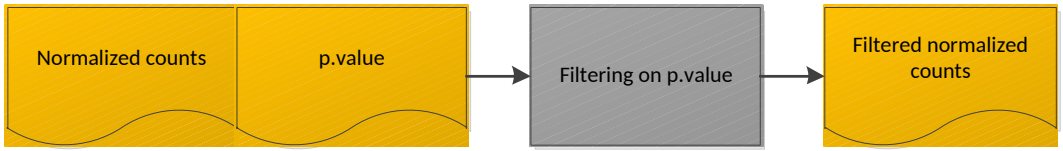
A comparison of methods for differential expression analysis of RNA-seq data. Charlotte Soneson1 and Mauro Delorenzi
BMC Bioinformatics 2013, 14:91 doi:10.1186/1471-2105-14-91

SARTools = Statistical Analysis of RNA-Seq Tools

1. Perform a systematic quality control of the data
2. Avoid misusing the DESeq2 or edgeR packages
3. Keep track of all the parameters used: reproducible research
4. Provide a HTML report containing all the results of the analysis

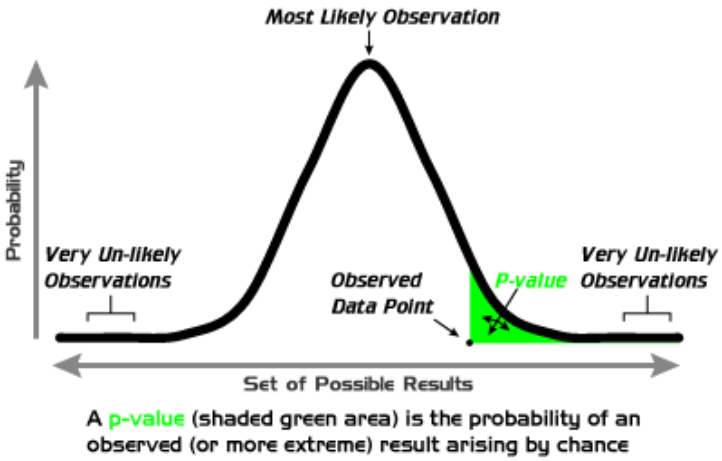


Statistics Outputs

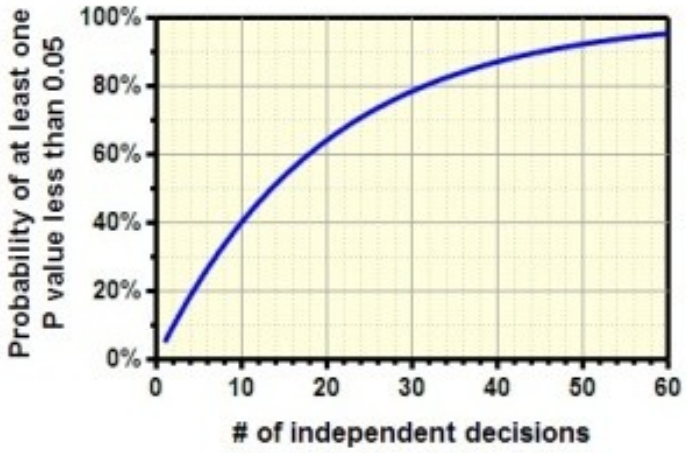


The results

- p.value
  - The p-value of the test statistic is a way of saying how extreme that statistic is for our sample data. The smaller the p-value, the more unlikely the observed sample.

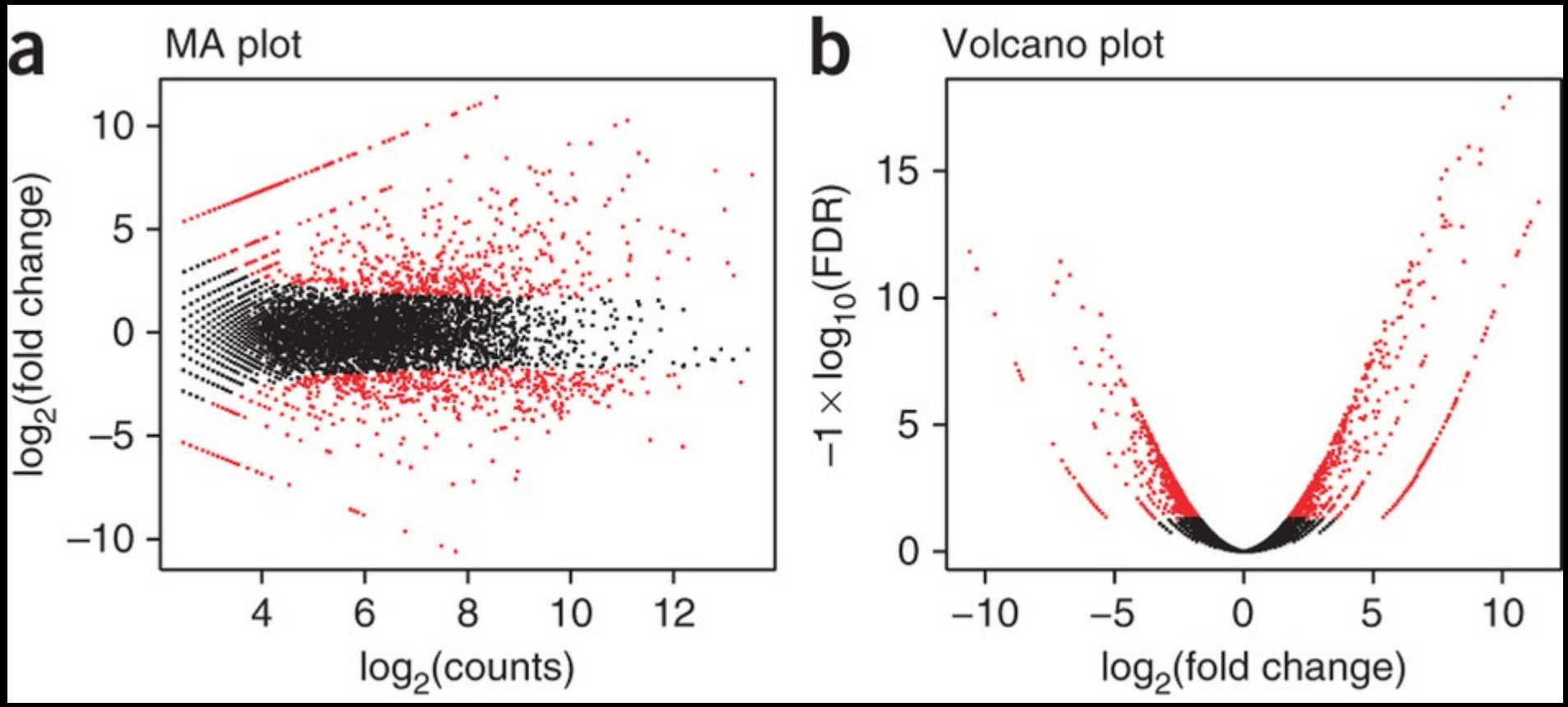


- adjusted p.value / False Discovery Rate
  - Used in multiple hypothesis testing
  - Corrections
    - Bonferroni
    - Benjamini-Hochberg (BH)



Filtering

- alpha risk
  - The number alpha is the threshold value that we measure p-values against. It tells us how extreme observed results must be in order to reject the null hypothesis of a significance test.
  - Must be set in advance ! 
  - Ex:
    - For results with a 90% level of confidence, the value of alpha is 1 - 0.90 = 0.10.
    - For results with a 95% level of confidence, the value of alpha is 1 - 0.95 = 0.05.
    - For results with a 99% level of confidence, the value of alpha is 1 - 0.99 = 0.01.
  - So:
    - alpha > pvalue ▭ H0 is rejected ▭ 


a
MA plot
log2(fold change)
10
5
0
−5
−10
4
6
8
10
12
log2(counts)
b
Volcano plot
−1 × log10(FDR)
15
10
5
0
−10
−5
0
5
10
log2(fold change)

Log Fold Change - LogFC

`log2(cond2/cond1)`

cond1	cond2	FC 2/1	logFC
100	800	8	3
100	400	4	2
100	200	2	1
200	100	0.500	-1
400	100	0.250	-2
800	100	0.125	-3