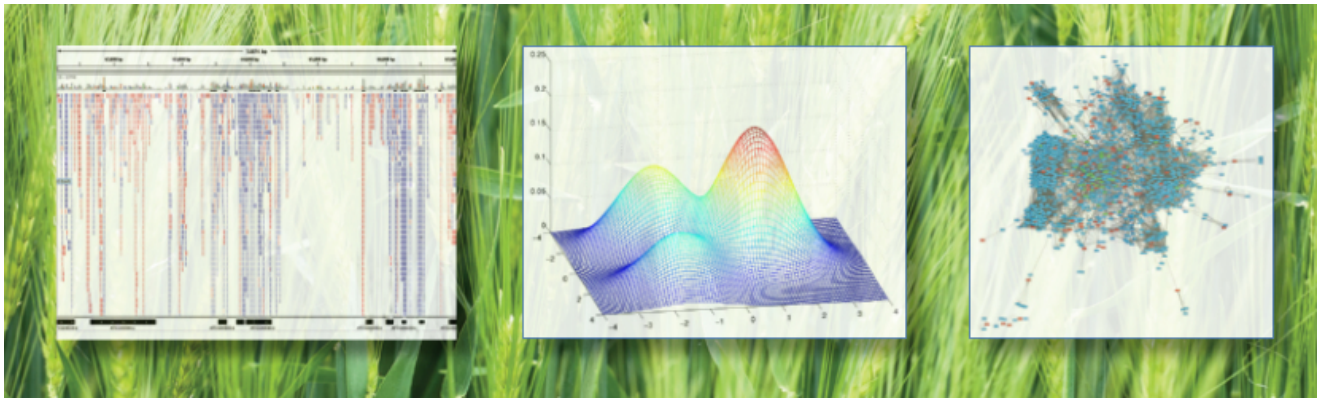
Course 3: Normalization and differential analysis from mRNA-Seq experiments

Normalization

Julie Aubert

UMR518 AgroParisTech/INRA Mathématiques et Informatique Appliquées-Paris



Normalization

Introduction

Overview of different normalization methods

Comparison of different normalization methods

Outline

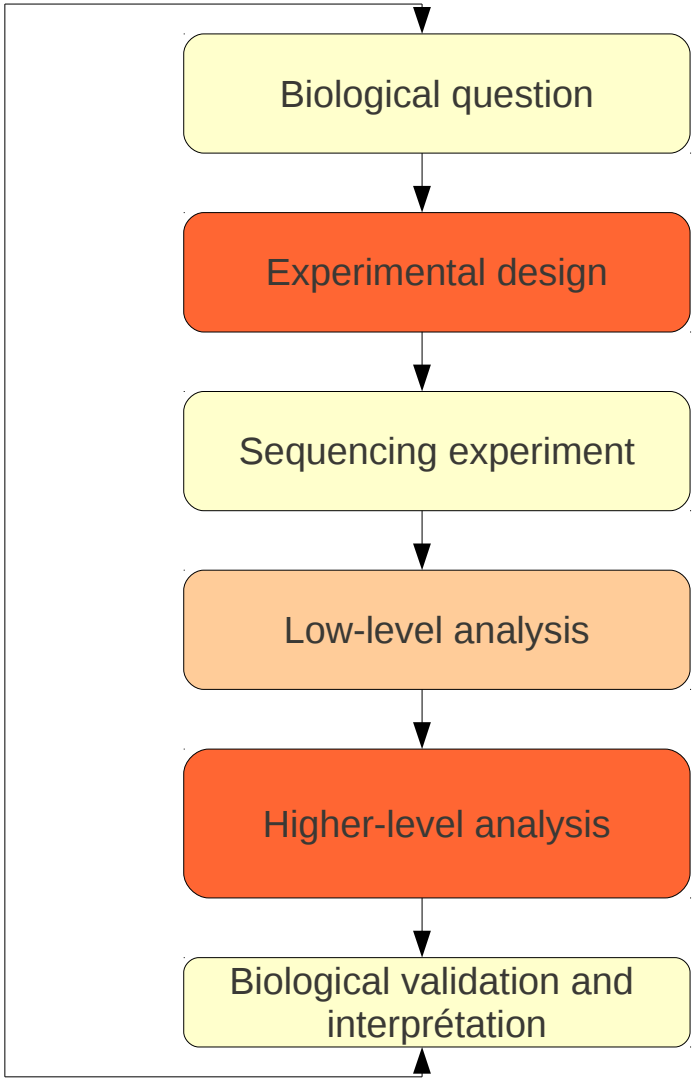
Transcriptome

- **Transcriptome** = complete set of transcripts in a cell and their quantity, for a specific developmental stage or physiological condition (Zhang et al. 2009).
  - varies across environmental conditions.
  - reflects actively expressed genes at a given time.
- mRNA expression level in a given population.

Key aims of transcriptomics

- to catalogue all species of transcript; to determine the transcriptional structure of genes
- **differential analysis:** to quantify the **changing expression levels** of each transcript during development and under different condition.

A typical RNA-Seq experiment

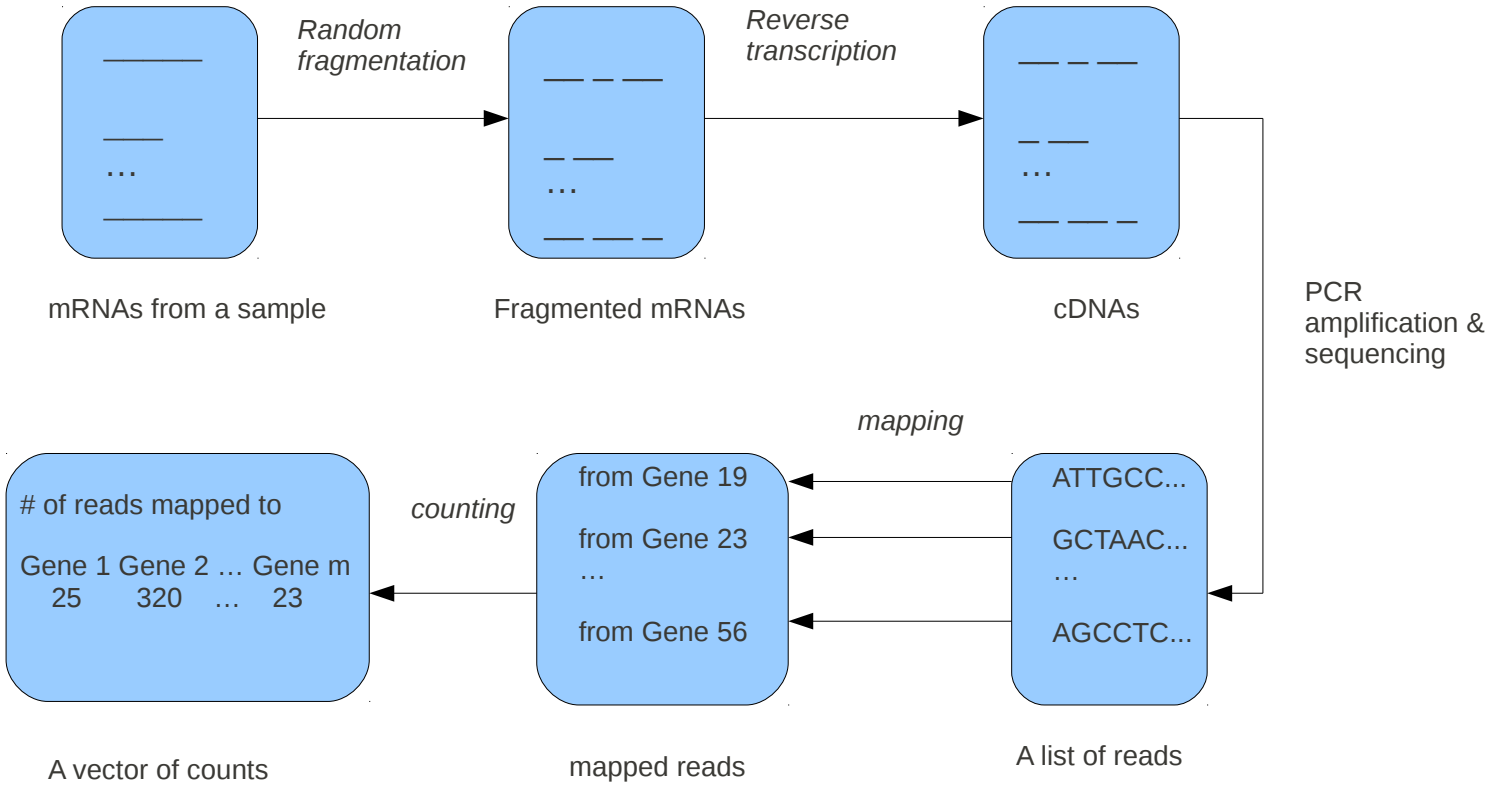


Exploratory Data Analysis, image analysis, base calling, read mapping, metadata integration

Exploratory Data Analysis, normalization and expression quantification, differential analysis, metadata integration

*Adapted from S. Dudoit, Berkeley

RNA-sequencing



Adapted from Li et al. (2011)

A typical raw dataset

	S_1	S_2	...	S_j	...	S_n
Gene 1	16	9	...	y_{1j}	...	15
Gene 2	4448	3973	...	y_{2j}	...	3964
...	...	...	...	...	...	...
Gene i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{in}
...	...	...	...	...	...	...
Gene G	59	164	...	y_G	...	143
Seq. depth	6865057	11127087	...	$N_j = \sum_{i=1}^{G} y_{ij}$	...	11320226

y_{ij} = number of sequences from sample j assigned to gene i.

Remark: one row = one region of interest (gene, exon, transcript, $\cdots$).

Differential analysis

Identification of differentially expressed (DE) genes

A gene is declared **differentially expressed** (DE) between two conditions if the observed difference is statistically significant, i.e. greater than a natural random variation.

- Need of statistical tools to make a decision.
- Main steps of the analysis: experimental design, normalization, differential analysis, multiple testing.

Statistical issues of gene expression analysis from RNA-Seq experiment

- A large number of genes and few replicates
- Discrete, positive and skewed data
- Large dynamic range with presence of 0 counts
- The total number of sequences (= **library size**) is not the same for all the samples

Normalization or how to make measurements comparable

Definition

Normalization is a process designed to identify and correct **technical biases** removing the least possible biological signal. This step is technology and platform-dependant.

Technical biases

Some biases may be **controlled** by an adapted experimental design or a good experimental protocol.

Normalization aims to correct systematic **uncontrollable** biases such as those induced by sequencing process.

Within and between normalization

Within-sample normalization enabling comparisons of fragments (genes) from a **same** sample.

Between-sample normalization enabling comparisons of fragments (genes) from **different** samples.

Sources of variability

Read counts are proportional to expression level, gene length and sequencing depth (same RNAs in equal proportion).

Within-sample

- Gene length
- Sequence composition (GC content)

Between-sample

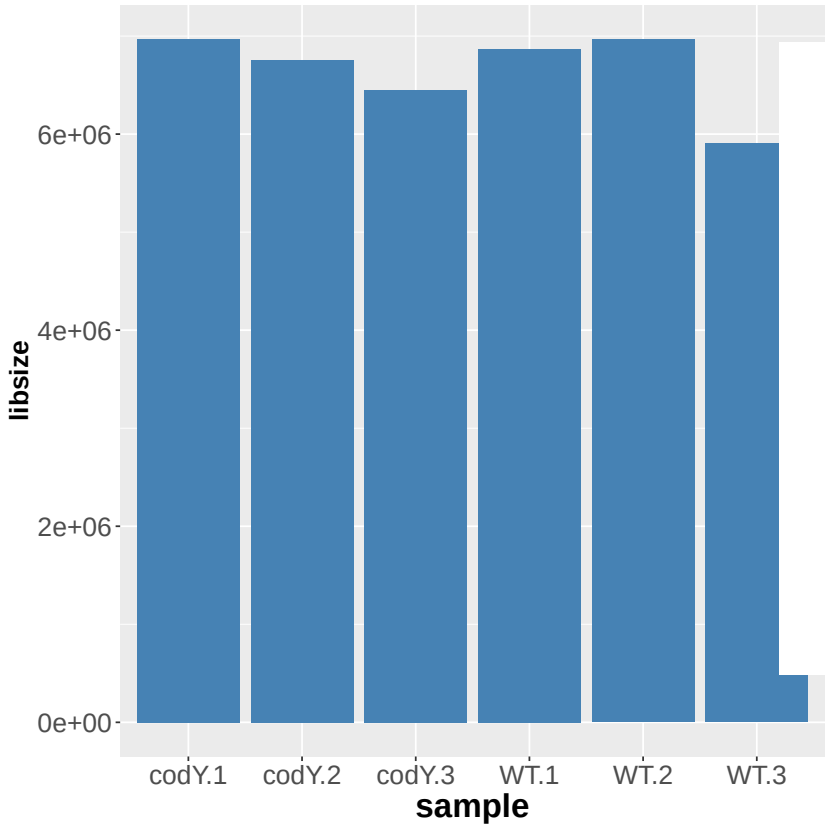
- Depth (total number of sequenced and mapped reads)
- RNA-composition or presence of majority fragments
- Sequence composition du to PCR-amplification step in library preparation (Pickrell et al. 2010, Risso et al. 2011)

Normalization and differential expression (DE) analysis

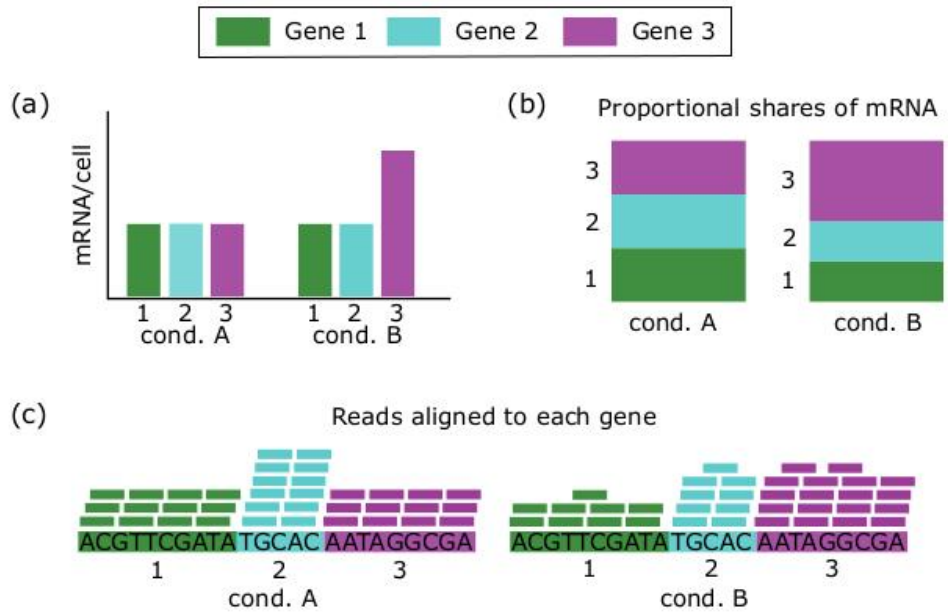
DE analysis concerned with **relative changes** in expression levels between conditions rather than estimating absolute expression levels.

Technical effects to correct are those related to the experimental conditions (sample-specific effects).

Sequencing depth



RNA composition



from Evans et al. (2017)

Outline

Typology of normalization methods

Typology according to the underlying assumptions (Evans et al. 2017).

Normalization by library size

Same total expression, same amount of mRNA/cell for each experimental condition.

Normalization by distribution or testing

- DE and non-DE genes have the same behaviour.
- Balanced expression.

Normalization by controls

- Existence of control (invariant set of genes).
- Control genes behave like non-control genes (same technical effects).

Normalization by library size

How does it work ?

Normalized counts are raw counts divided by a scaling factor calculated for each sample.

Total Count (TC) (Marioni et al. 2008)

The scaling factor depends on the total number of reads in each sample.

$$\frac{y_{ij}}{\hat{s}_j}, \qquad \hat{s}_j = \frac{N_j}{\frac{1}{n}\sum_\ell N_\ell}$$

y_{ij} number of reads for gene i in sample j, N_j number of reads in sample j (library size of sample j), n number of samples in the experiment, $\hat{s}_j$ normalization factor associated with sample j

Reads Per KiloBase Per Million Mapped (Mortazavi et al. 2008)

$$\frac{y_{ij}}{\hat{s}_{ij}}, \qquad \hat{s}_{ij} = N_j * L_i * 10^3 * 10^6$$

L_i: length of gene i

Normalization by library size: some remarks

Reads Per KiloBase Per Million Mapped (RPKM) and its variants (FPKM, ERPKM)

- Allows to compare expression levels between genes of the same sample
- Unbiased estimation of number of reads but affect the variance. (Oshlack et al. 2009)

Total Count normalization may be driven by a small number of highly expressed genes.

Normalization by distribution: variants of TC normalization

Upper Quartile normalization (Bullard et al. 2010)

$$\hat{s}_j = \frac{Q_{3j}}{\frac{1}{n}\sum_\ell Q_{3\ell}}$$

Median normalization

$$\hat{s}_j = \frac{median_j}{\frac{1}{n}\sum_\ell median_\ell}$$

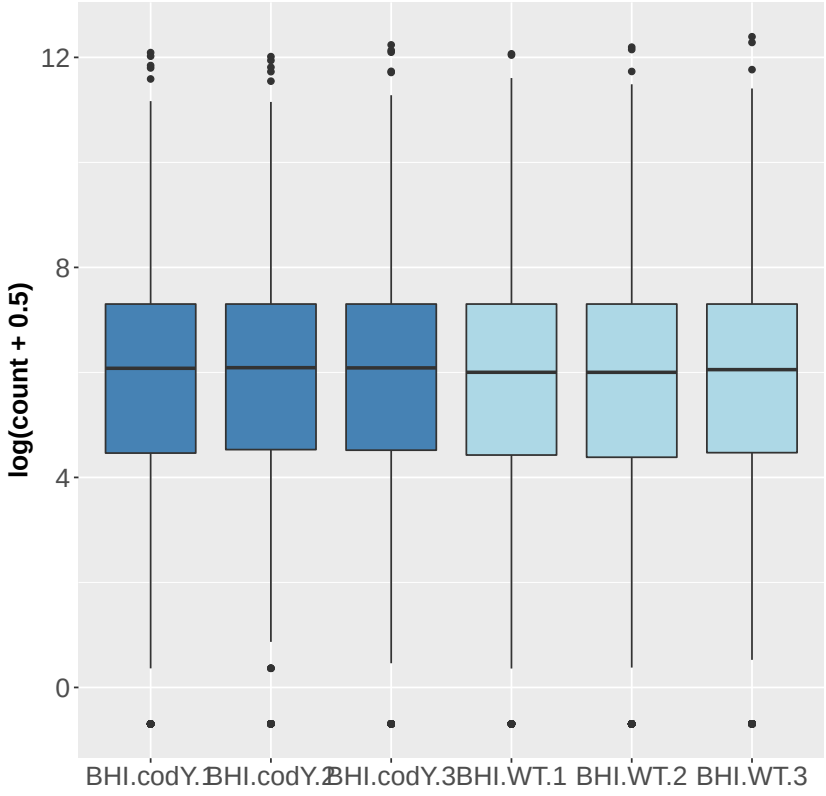
Q_{3j} and $median_j$ are calculated after exclusion of genes with no read counts.

Quantile: FQ (Robinson and Smyth 2008)

Force all samples to have the same distribution.

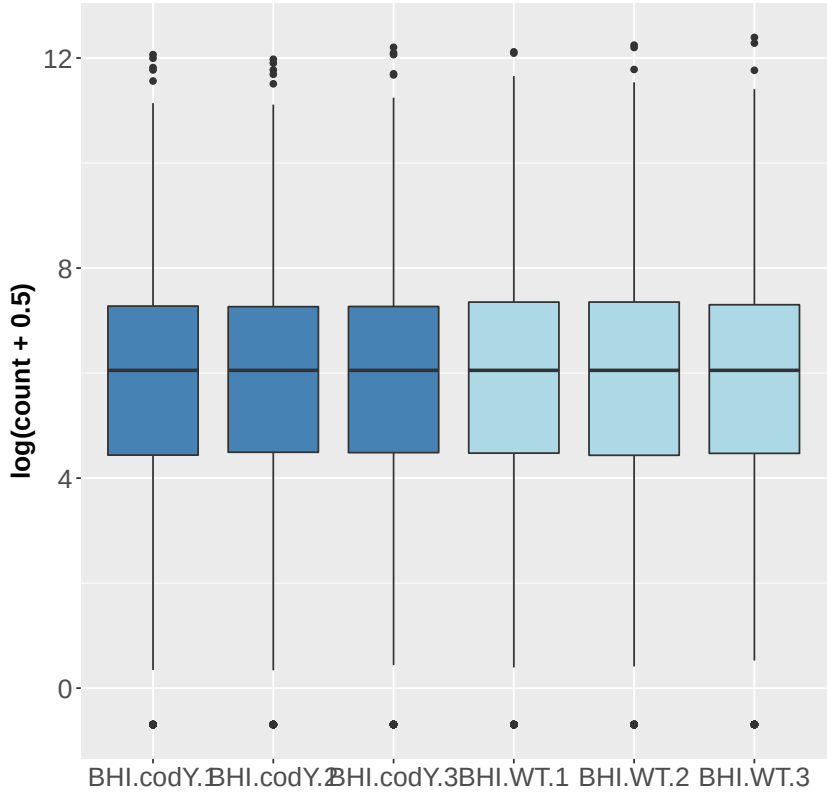
Upper quartile and median normalization

Upper quartile normalization



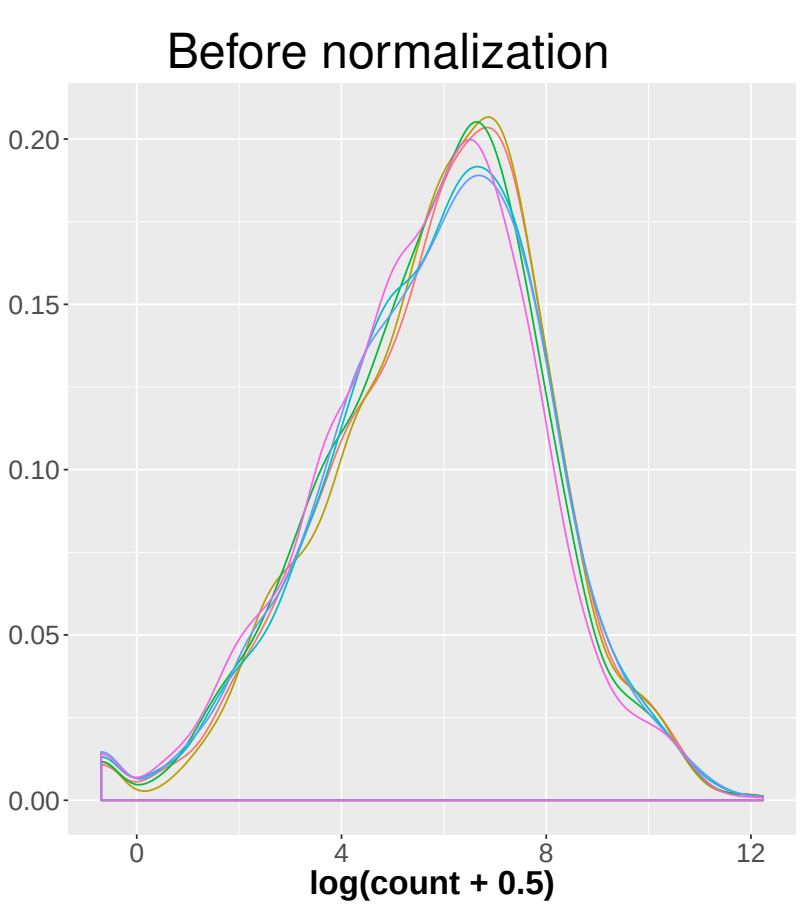
Third quartile is equal across samples.

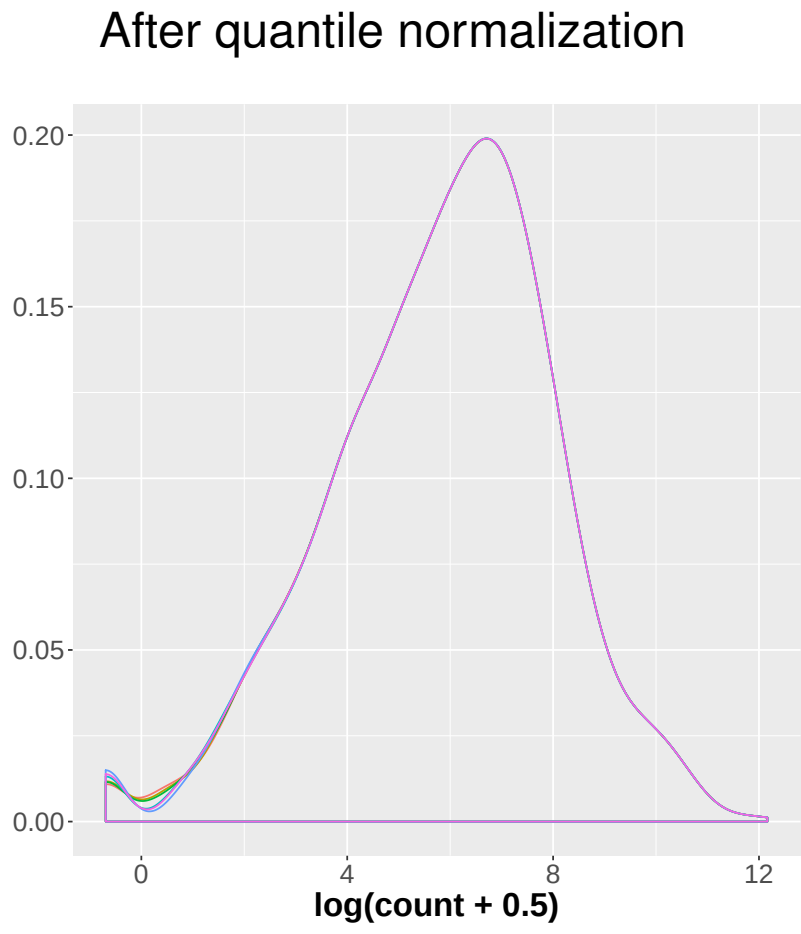
Median normalization



Median is equal across samples.

Quantile normalization





The Effective Library Size concept

Motivation

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

Assumption

A majority of transcripts is not differentially expressed

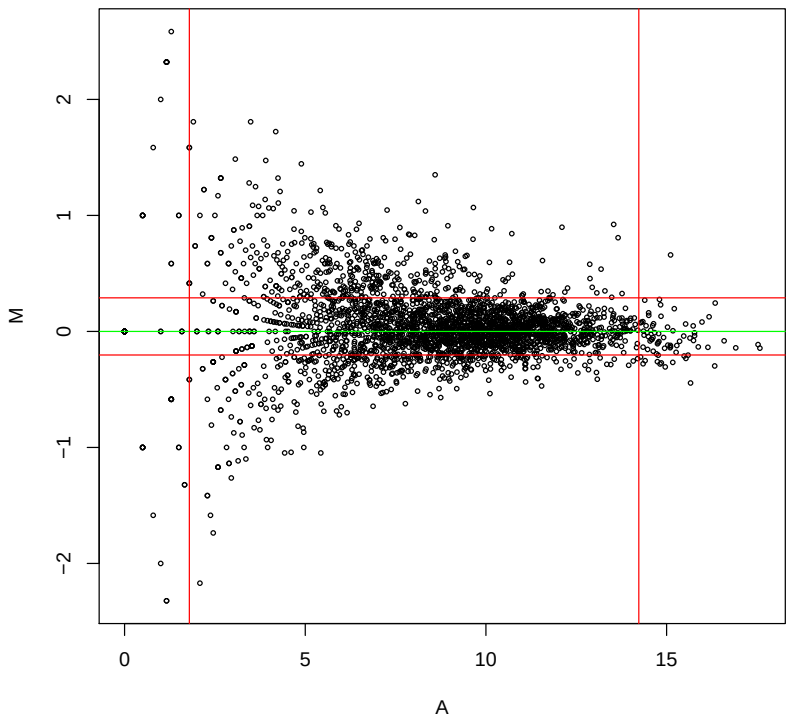
Aim

Minimizing effect of (very) majority sequences

- Trimmed Mean of M-values, Robinson and Oshlack 2010 (edgeR)
- Relative Log-Expression, Anders and Huber 2010 (DESeq2)

Trimmed Mean of M-values (TMM)

Idea: we may not estimate the total ARN production in one condition but we may estimate a global expression change between two conditions from non extreme M_i distribution.



Filter on:

- transcripts with nul counts,
- the 30% more extreme $M_{ij}^r = log_2(\frac{y_{ij}/N_j}{y_{ir}/N_r})$ values,
- the 5% more extreme $A_{ij}^r = 0.5 \times [log_2(\frac{y_{ij}}{N_j}) + log_2(\frac{y_{ir}}{N_r})]$ values.

Trimmed Mean of M-values

1. Select the reference sample r
2. Define a set of genes G^* for which neither the M_{ij}^r or the A_{ij}^r value was trimmed
3. Calculate the scaling factors $TMM_j^{(r)}$ such as

$$log_2(TMM_j^{(r)}) = \frac{\sum_{i \in G^*} w_{ij}^r M_{ij}^r}{\sum_{i \in G^*} w_{ij}^r}$$

with $w_{ij}^r = \frac{N_j - y_{ij}}{N_j y_{ij}} - \frac{N_r - y_{ir}}{N_r y_{ir}}$

4. Rescale the factors to avoid dependance on a specific reference sample

$$\hat{s}_j = \frac{TMM_j^{(r)}}{exp(\sum_\ell TMM_\ell^{(r)}/n)}$$

The Relative Log-Expression method (RLE, DESeq)

1. Compute a pseudo-reference sample: geometric mean across samples (less sensitive to extreme value than standard mean)

$$y_{ij}^{r} = \left(\pi_{j=1}^{n} y_{ij}^{1/n}\right)$$

with y_{ij} number of reads in sample j assigned to gene i, n number of samples in the experiment.

2. Calculate scaling factors

$$\hat{s}_j = \underset{i:y_{ij}^{r} \neq 0}{median} \frac{y_{ij}}{y_{ij}^{r}}$$

Some remarks about TMM and RLE normalization

Interpretation of the scaling factors

- The normalization factors of all the libraries multiply to 1.
- $\hat{s}_j < 1$: a small number of high count genes are monopolizing the sequencing. $\Rightarrow$ Need of downscaling.

	WT.1	WT.2	WT.3	codY.1	codY.2	codY.3
RLE	1.05	1.05	0.87	1.06	1.06	0.93
TMM	1.02	1.00	0.97	1.01	1.05	0.95

Model-based normalization, not transformation

In edgeR and DESeq2, normalization factors $=$ correction factors that enter into the model.

Normalization by testing

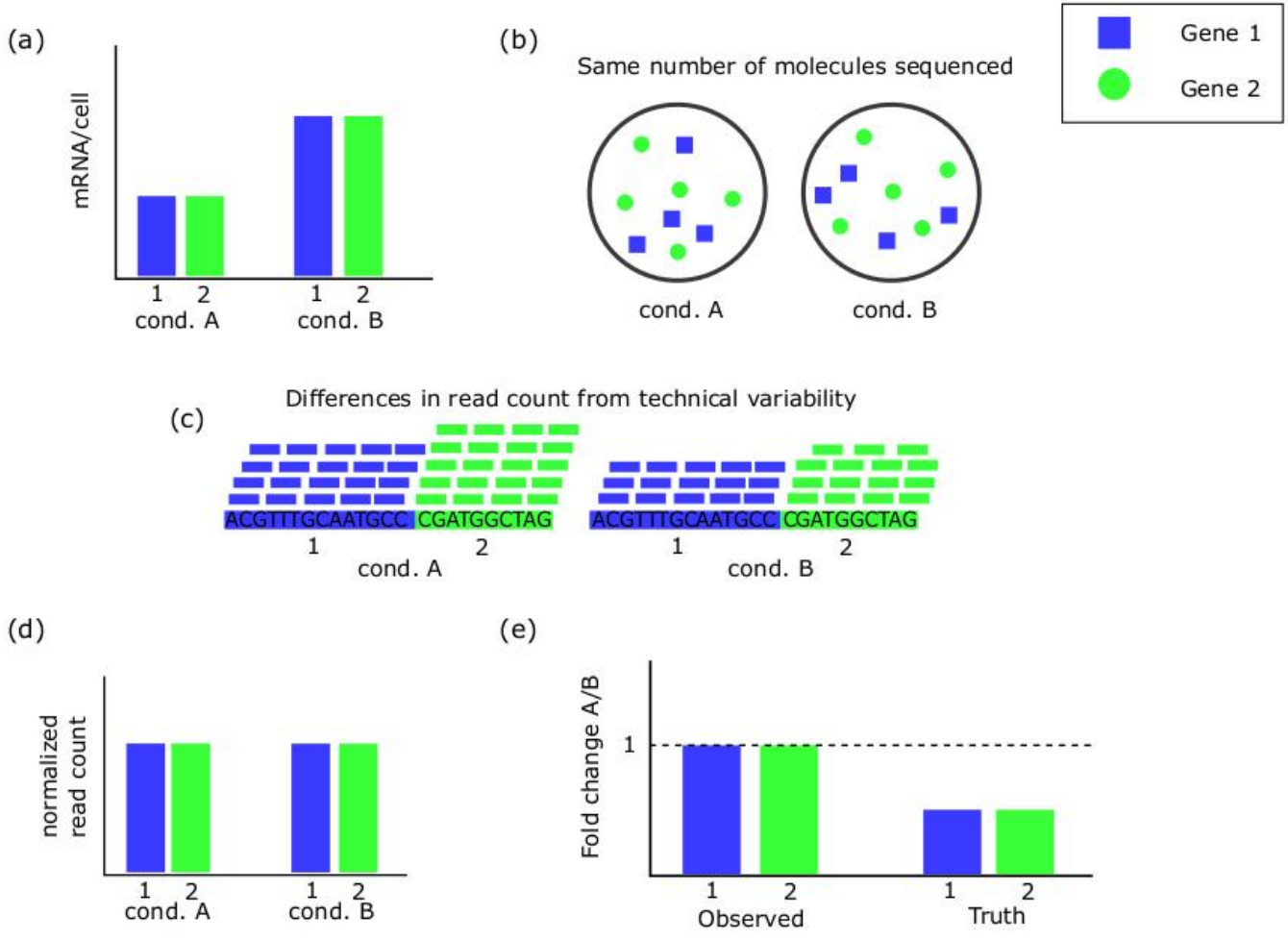
Iterations between two steps

1. Estimation of a set of non DE-genes.
2. Estimation of a scaling factor for each sample using this defined set.

Methods

- PoissonSeq (Li et al. 2012)
- DEGES (Kadota et al. 2012)

Where conventional methods fail



from Evans et al. (2017)

Normalization by controls

Assumptions

- Existence of controls and behavior as expected (negative controls = non-DE)
- Controls behave like non-control genes (affected by same technical effects)

Methods

- Housekeeping genes
- "Conventional normalization" with Spike-ins
- Factor analysis of controls: Remove Unwanted Variation (RUV) (Risso et al., 2014)

RUVSeq: Remove Unwanted sources of Variation
Risso et al., 2014

Motivation

Most methods fail to correct for **complex** unwanted technical effects.

Aim

To remove variation between samples that is not the result of the biological covariates of interest.

Three variants

- RUVg: existence of negative controls (non-DE across conditions)
- RUVs: existence of negative controls and negative controls samples (expression not related to biological conditions)
- RUVr: does not require existence of controls. Factors of wanted variation are known (design matrix) and the factors of unwanted variation are not correlated with experimental conditions.

RUVSeq principle

Estimate unwanted technical effects using a Generalized Linear Model

$$\log \mathbb{E}[Y|W, X, O] = W\alpha + X\beta + O.$$

- Y: observed read count matrix.
- X: known design matrix for the experiment.
- W: matrix related to k factors of unwanted variance (k must be fixed beforehand).
- O: optional matrix of sequencing depth offsets.

Estimation of W

- RUVg: use a set of J negative control genes ($\beta_j = 0; j \in [1; J]$).
- RUVs: use a set of negative control samples (technical replicates) ($\beta = 0$).
- RUVr: use the residuals from the first-pass GLM regression of Y on X without W.

Outline

Comparison of normalization methods

At lot of different normalization methods...

- Some are part of models for DE, others are 'stand-alone'.
- They do not rely on similar hypotheses.
- But all of them claim to remove technical bias associated with RNA-seq data.

Which one is the best ?

- How to and on which criteria choice a normalisation adapted to our experiment ?
- What impact of the bioinformatics, normalisation step or differential analysis method on lists of DE genes ?

Briefings in Bioinformatics Advance Access published September 17, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 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*

Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions

Ciaran Evans ✉, Johanna Hardin, Daniel M. Stoebel

Briefings in Bioinformatics, bbx008, https://doi.org/10.1093/bib/bbx008

4 real datasets

At least 2 conditions, at least 2 bio. rep., no tech. rep.

Organism	Type	Number of genes	Replicates per condition	Minimum library size	Maximum library size	Correlation between replicates	Correlation between conditions	% most expressed gene	Library type	Sequencing machine
H. sapiens	RNA	26,437	$\{3,3\}$	2.0×10^7	2.8×10^7	(0.98,0.99)	(0.93,0.96)	$\approx 1\%$	SR 54, ND	GaIIx
A. fumigatus	RNA	9,248	$\{2,2\}$	8.6×10^6	2.9×10^7	(0.92,0.94)	(0.88,0.94)	$\approx 1\%$	SR 50, D	HiSeq2000
E. histolytica	RNA	5,277	$\{3,3\}$	2.1×10^7	3.3×10^7	(0.85,0.92)	(0.81,0.98)	6.4-16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	$\{3,2,2\}$	2.0×10^6	5.9×10^6	(0.95,0.99)	(0.09,0.75)	17.4-51.1%	SR 36, D	GaIIx

Table 1: Summary of datasets used for comparison of normalization methods, including the organism, type of sequencing data, number of genes, number of replicates per condition, minimum and maximum library sizes, Pearson correlation between replicates and between samples of different conditions (minimum, maximum), percentage of reads associated with the most expressed RNA (minimum, maximum), library type (SR = single-read or PE = paired-end read, D = directional or ND = non-directional), and sequencing machine.

Simulated dataset (from the mouse dataset)

- Proportion of DE genes: from 0 to 30%
- equivalent / not equivalent library sizes
- presence / absence of high count genes

Comparison procedures

Distribution and properties of normalized datasets

Boxplots, variability between biological replicates

Comparison of DE genes

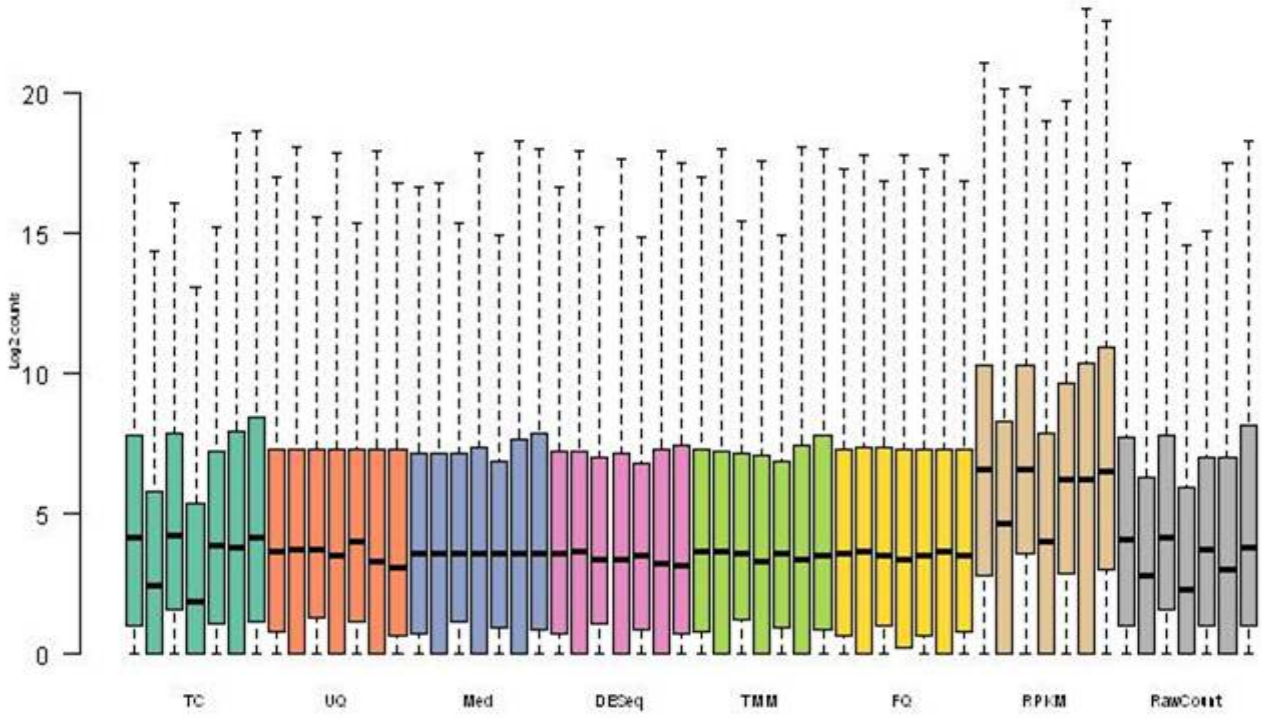
- Differential analysis: DESeq v1.6.1 (Anders and Huber 2010), default param.
- Number of common DE genes, similarity between list of genes (dendrogram - binary distance and Ward linkage)

Power and control of the Type-I error rate

Simulation study

Normalized data distribution

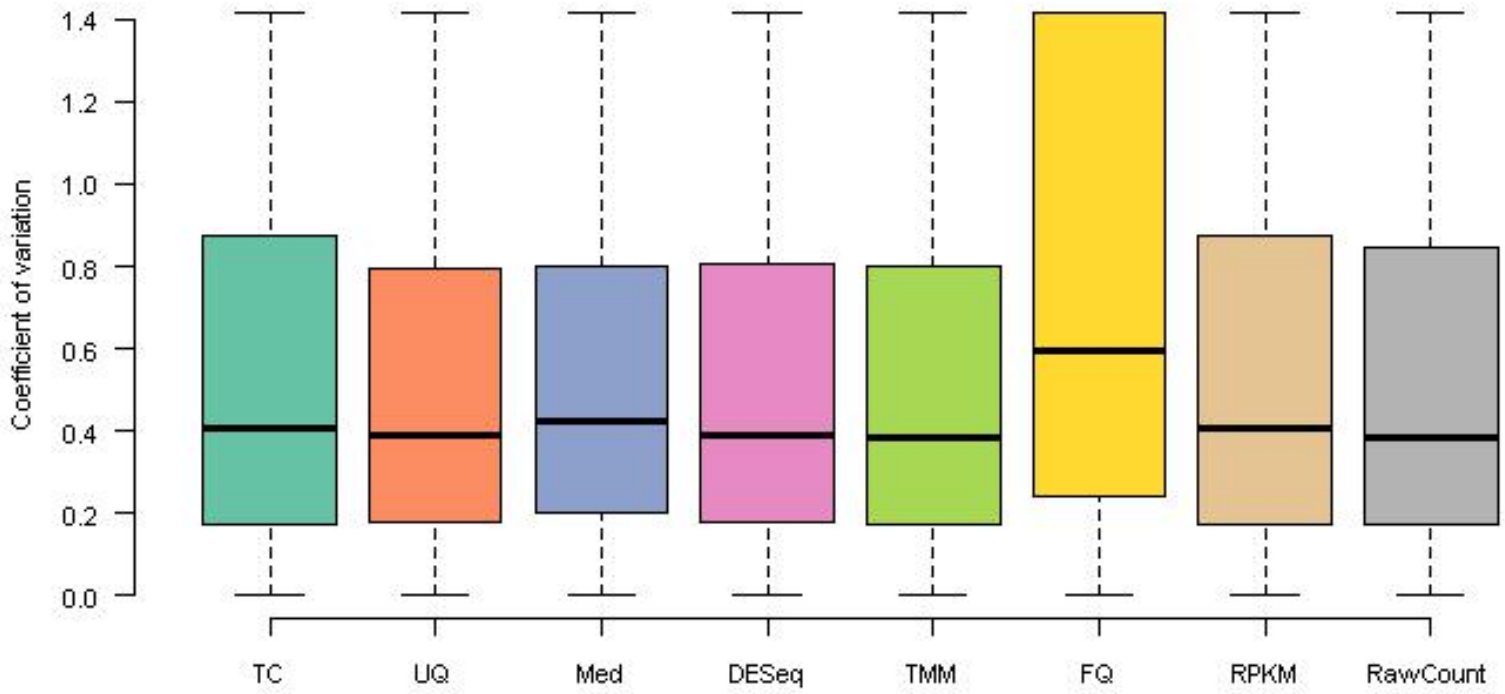
When large diff. in lib. size, TC and RPKM do not improve over the raw counts.



Example: *Mus musculus* dataset

Within-condition variability

Example: *Mus musculus, condition D* dataset



Number of DE genes

- DESeq v1.6.0, default parameters
- Input data: raw counts + scaling factors $\hat{s}_j$ (except RPKM)
- RPKM: normalized data **non rounded** and normalization parameter $\hat{s}_j = 1$

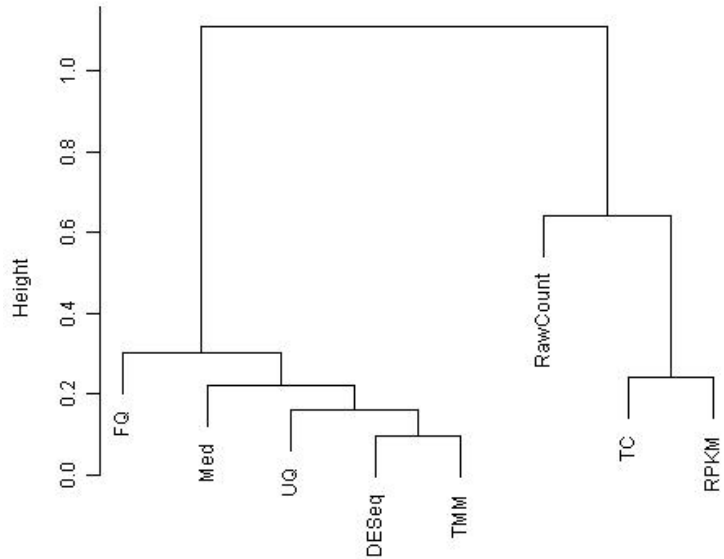
	TC	UQ	Med	DESeq	TMM	FQ	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1,213**	**1,195**	**1,160**	**1,172**	**1,054**	416	184
Med			**1,218**	**1,147**	**1,160**	**1,043**	416	183
DESeq				**1,249**	**1,169**	**1,058**	413	184
TMM					**1,190**	**1,051**	416	184
FQ						**1,092**	414	184
RPKM							417	149
RawCount								184

Example: *E. histolytica* dataset, common genes

Lists of differentially expressed (DE) genes

For each dataset

- (gene x method) binary matrice:
  - 1: DE gene
  - 0: non DE gene
- Jaccard distance between methods
- dendrogramm, Ward linkage algorithm

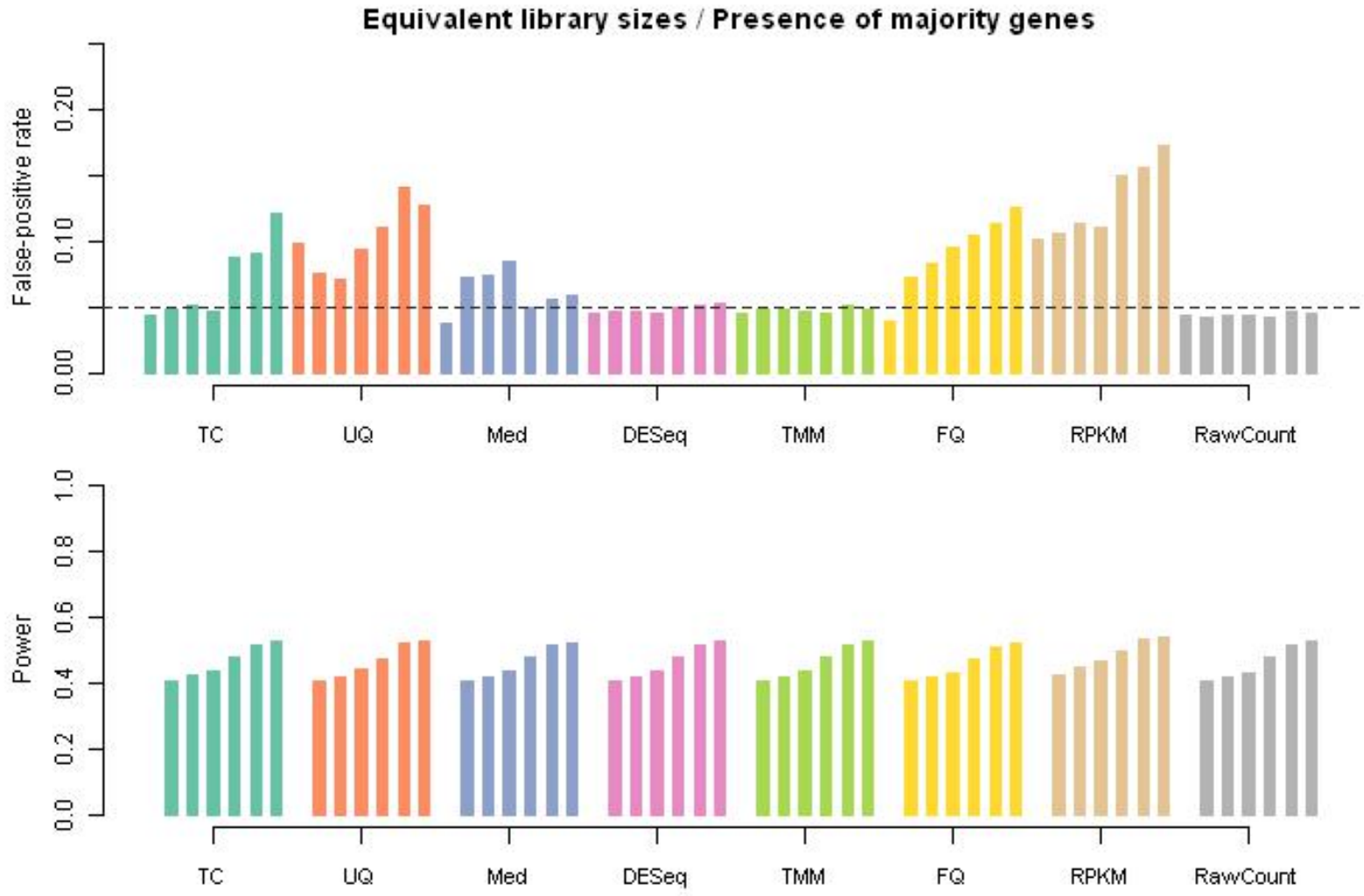


Consensus matrice

Mean of the distance matrices obtained from each dataset

Type-I Error Rate and Power (Simulated data)

Inflated FP rate for all the methods except TMM and DESeq (RLE)



So the Winner is ... ?

In most cases

The methods yield similar results

However ...

Differences appear based on data characteristics

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
FQ	++	–	+	++	–
RPKM	–	+	+	–	–

Interpretation

- RawCount Often fewer differential expressed genes (*A. fumigatus*: no DE gene)
- TC, RPKM
  - Sensitive to the presence of majority genes
  - Less effective stabilization of distributions
  - Ineffective (similar to RawCount)
- FQ
  - Can increase between group variance
  - Is based on an very (too) strong assumption (similar distributions)
- Median High variability of housekeeping genes
- TC, RPKM, FQ, Med, UQ Adjustment of distributions, implies a similarity between RNA repertoires expressed

Conclusions of the study of StatOmique (2013)

- Hypothesis: the majority of genes is invariant between two samples.
- Differences between methods when presence of majority sequences, very different library depths.
- TMM and RLE: performant and robust methods in a DE analysis context on the gene scale.
- Normalisation is necessary and not trivial.
- Do not normalise by gene length in a context of differential analysis.

Comparaison study of Evans et al. (2017)

Simulation parameters

- 2 datasets: 10000 genes in 4 samples, or 1000 genes in 10 samples
- same or different amount of mRNA / cell

Criteria for comparison

- Mean Square Error (MSE) of the log fold change for non DE genes (should be close to 0) samples
- Empirical False Discovery Rate (eFDR)

Normalization methods

DESeq (RLE), TMM, TC, DEGES, PoissonSeq

Conclusions of the study of Evans et al. 2017

- The correct normalization method to use depends on which assumptions are valid for the biological experiment:
  - same / different amount of mRNA / cell
  - symmetry of differential expression
  - low number of DE genes
- Incorrect normalization leads to problem in downstream analysis, such as inflated FP.
- There are examples of global shifts in expression that violate assumptions of conventional normalization methods, requiring controls.

Normalization: key points

Detection of differential expression in RNA-seq data is inherently biased (more power to detect DE of longer genes).

Do not normalize by gene length in a context of differential analysis.

RNA-seq data are affected by technical biaises (total number of mapped reads per lane, gene length, composition bias)
$\Rightarrow$ A normalization is needed and has a great impact on the DE genes.

The correct normalization method to use depends on which assumptions are valid for the biological experiment.

No normalization method is perfect, and for every method there exists cases for which the assumptions are violated.

References

- Anders S, Huber W. (2010) **Differential expression analysis for sequence count data.** *Genome Biology*;11:R106.
- Bullard JH, Purdom E, Hansen KD, et al. (2010) **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics*;11(1):1-13.
- Kadota K, Nishiyama T, Shimizu K. (2012) **A normalization strategy for comparing tag count data.** *Algorithms Mol Biol*;7(1):1-13.
- The French StatOmique Consortium; Dillies MA, Rau A, Aubert J, Hennequet-Antier C, et al. (2013) **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.** *Briefings in Bioinformatics*;14(6):671?83.
- Evans C, Hardin J, Stoebel DM (2017) **Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions.** *Briefings in Bioinformatics*, doi.org/10.1093/bib/bbx008.
- Li J, Witten D, Johnstone I, et al. (2012) **Normalization, testing, and false discovery rate estimation for RNA-sequencing data.** *Biostatistics*;13(3):523?38.
- Marioni JC, Mason CE et al. (2008) **RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Research*,18:1509-1517.

References

- Mortazavi A, Williams BA, McCue K, et al. (2008) **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nature Methods*;5(7):621-8.
- Oshlack A, Wakefield MJ (2009) **Transcript length bias in RNA-seq data confounds systems biology.***Biology Direct*;4(1):1-10.
- Robinson MD, Oshlack A. (2010) **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biology*,11:R25
- Robinson MD and Smyth GK (2008). **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics*, 9, 321-332.
- Robinson MD, McCarthy DJ and Smyth GK (2010). **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 26,139-140.
- Risso D, Schwartz K, Sherlock G, et al. (2011). **GC-content normalization for RNA-Seq data.** *BMC Bioinformatics*;12(1):1-17.
- Risso D, Ngai J, Speed T and Dudoit S (2014). **Normalization of RNA-seq data using factor analysis of control genes or samples.** *Nature Biotechnology*, 32(9):896-902.