Data specificities and normalization

Etienne Delannoy[1] and Marie-Laure Martin-Magniette[1,2]

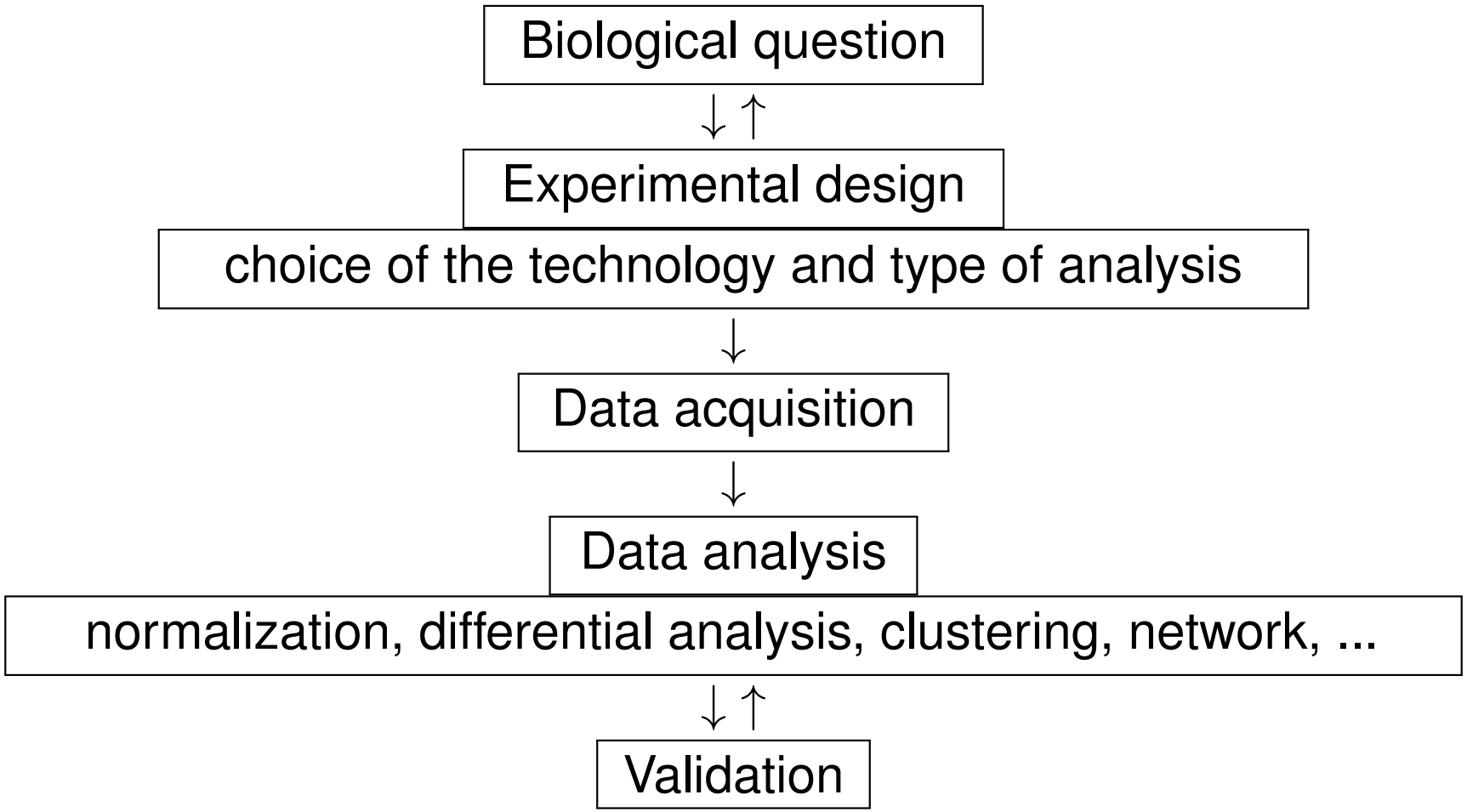
1- IPS2 Institut des Sciences des Plantes de Paris-Saclay

2- UMR AgroParisTech/INRA Mathematique et Informatique Appliquees

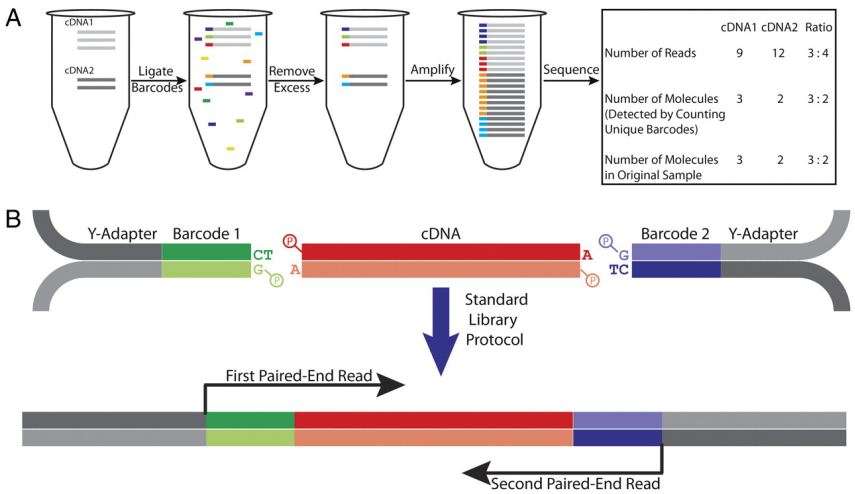
Aims of the talk

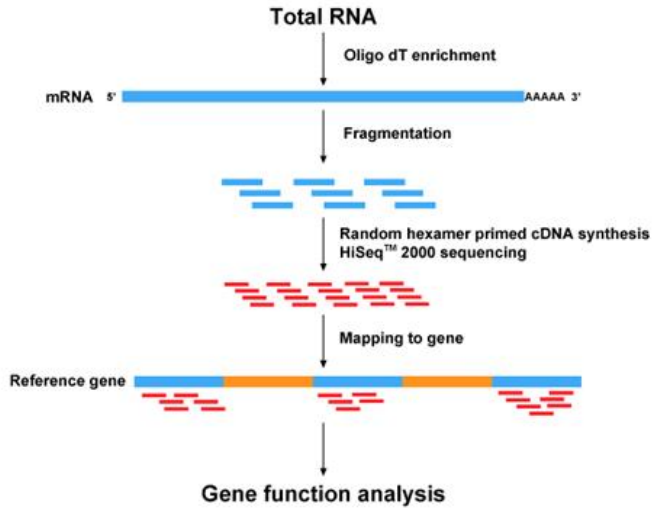
- Quantitative analysis of gene expression
- Overview of the different methods to normalize RNA-seq data before a differential analysis
- It is not exhaustive

Design of a transcriptomic project



High-throughput transcriptome sequencing (HTS) data





- Reads aligned or directly mapped to the genome to get counts (discrete data) $\Rightarrow$ digital measures of gene expression

Mapping step

Statistical analyses begin with count tables

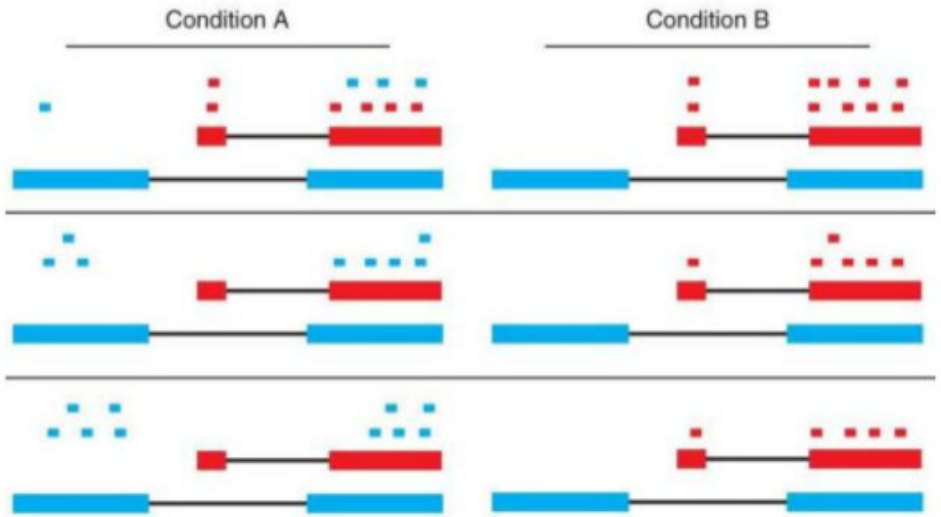
Mapping step

Statistical analyses begin with count tables

Reads → (Mapping and counting) → count tables

Not trivial

- Mapping and counting require a reference and a good annotation.
- Mapping parameters : sequencing errors vs polymorphism
- Pb with ambiguous mapping (Gene families, isoforms)

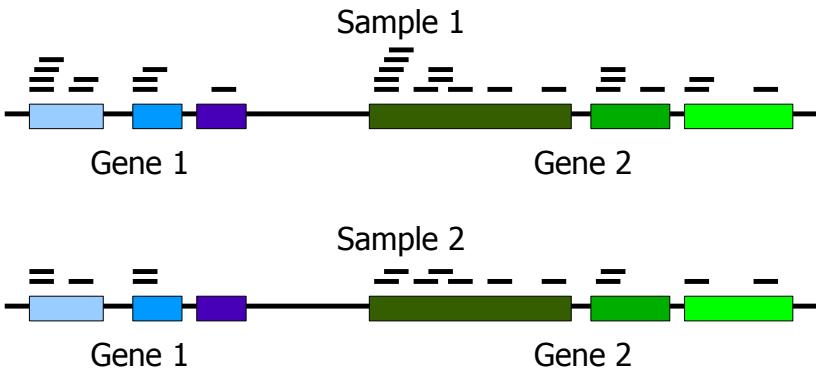


- Counts per gene? isoform? exon? base?

HTS data characteristics

Some statistical challenges of HTS data

- Discrete, non-negative, and skewed data with very large dynamic range (up to 5+ orders of magnitude)
- Sequencing depth (= "**library size**") varies among experiments
- Total number of reads for a gene $\propto$ expression level $\times$ length



Gene	E1	E2	E3
13CDNA73	4	0	6
A2BP1	19	18	20
A2M	2724	2209	13
A4GALT	0	0	48
AAAS	57	29	224
AACS	1904	129	4
AADACL1	3	13	239
[...]			

Normalization

Definition

- Normalization is a process designed to identify and correct **technical biases**.
- Two types of bias

  controlable biases: the construction of cDNA libraries
  uncontrolable biases: sequencing process

Between and within normalization

Within-sample normalization

- Enabling comparisons of genes from a same sample
- Not required for a differential analysis
- Not really relevant for the data interpretation
- Sources of variability: gene length and sequence composition (GC content)

Between-sample normalization

- Enabling comparisons of genes from different samples
- Sources of variability: library size, presence of majority fragments, sequence composition due to PCR-amplification step in library preparation‘(Pickrell et al. 2010, Risso et al. 2011)

Which normalization method ?

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 ?

French StatOmique Consortium; 2012. doi : 10.1093./bib/bbs046

Three types of methods

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

Distribution adjustment

TC (Marioni et al. 2008), Quantile FQ (Robinson and Smyth 2008), Upper Quartile UQ (Bullard et al. 2010), Median

Method taking length into account

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

The Effective Library Size concept

Trimmed Mean of M-values TMM (Robinson et al. 2010, package edgeR), RLE (Anders and Huber 2010, package DESeq2)

Distribution adjustement

For sample j, raw counts of gene g divided by a scaling factor

$$\frac{Y_{gj}}{\hat{s}_j}$$

- Total read count normalization (Marioni et al. 2008)

$$\hat{s}_j = \frac{N_j}{\frac{1}{n}\sum_\ell N_\ell}, \text{ where } N_j = \sum_g Y_{gj}$$

- Upper Quartile normalization (Bullard et al. 2010)

$$\hat{s}_j = \frac{Q3_j}{\frac{1}{n}\sum_\ell Q3_\ell}$$

$Q3_j$ is computed after exclusion of transcripts with no read count

- Median

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

Reads Per Kilobase per Million mapped reads

$$\frac{Y_{gj}}{N_j * L_g} * 10^3 * 10^6$$

- RPKM method is an adjustment for library size and transcript length
- Allows to compare expression levels between genes of the same sample
- Unbiased estimation of number of reads but affect the variability. (Oshlack et al. 2009)

Method based on the Effective Library Size

Relative Log Expression (RLE)

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

$$(\prod_{\ell=1}^{n} Y_{g\ell})^{1/n}$$

- calculate normalization factor

$$\tilde{s}_j = median_g \frac{Y_{gj}}{(\prod_{\ell=1}^{n} Y_{g\ell})^{1/n}}$$

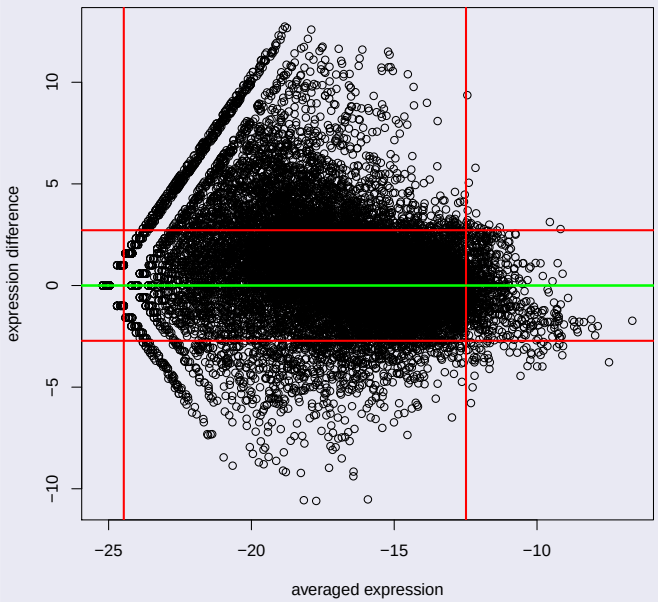
- normalize them such that their product equals 1

$$s_j = \frac{\tilde{s}_j}{exp[\frac{1}{n}\sum_\ell \log \tilde{s}_\ell]}$$

Method based on the Effective Library Size

Trimmed Mean of M-values (TMM)

Assumption: the majority of the genes are not differentially expressed



- Filter on genes with nul counts
- Filter on the resp. 30% and 5% more extreme values of M^r_{gj} and A^r_{gj}

where

$$M^r_{gj} = log2(\frac{Y_{gj}/N_j}{Y_{gr}/N_r})$$

$$A^r_{gj} = [log2(\frac{Y_{gj}}{N_j}) + log2(\frac{Y_{gr}}{N_r})]/2$$

TMM normalization

Algorithm

- Select the reference r as the library whose upper quartile is closest to the mean upper quartile.
- Compute weights $w_{gj}^r = (\frac{N_j - Y_{gj}}{N_j Y_{gj}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}})$
- Compute $TMM_j^r = \frac{\sum_{g \in G^\star} w_{gj}^r M_{gj}^r}{\sum_{g \in G^\star} w_{gj}^r}$
- Define

$$\tilde{s}_j = 2^{TMM_j^r}$$

- Normalize them such that their product equals 1

$$s_j = \frac{\tilde{s}_j}{exp[\frac{1}{n} \sum_\ell \tilde{s}_\ell]}$$

Comparison of 7 normalization methods

Differential analyses on 4 real datasets (RNA-seq or miRNA-seq) and one simulated dataset
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.

Comparison indicators

Distribution and properties of normalized datasets

Boxplots, variability between biological replicates

Comparison of DE genes

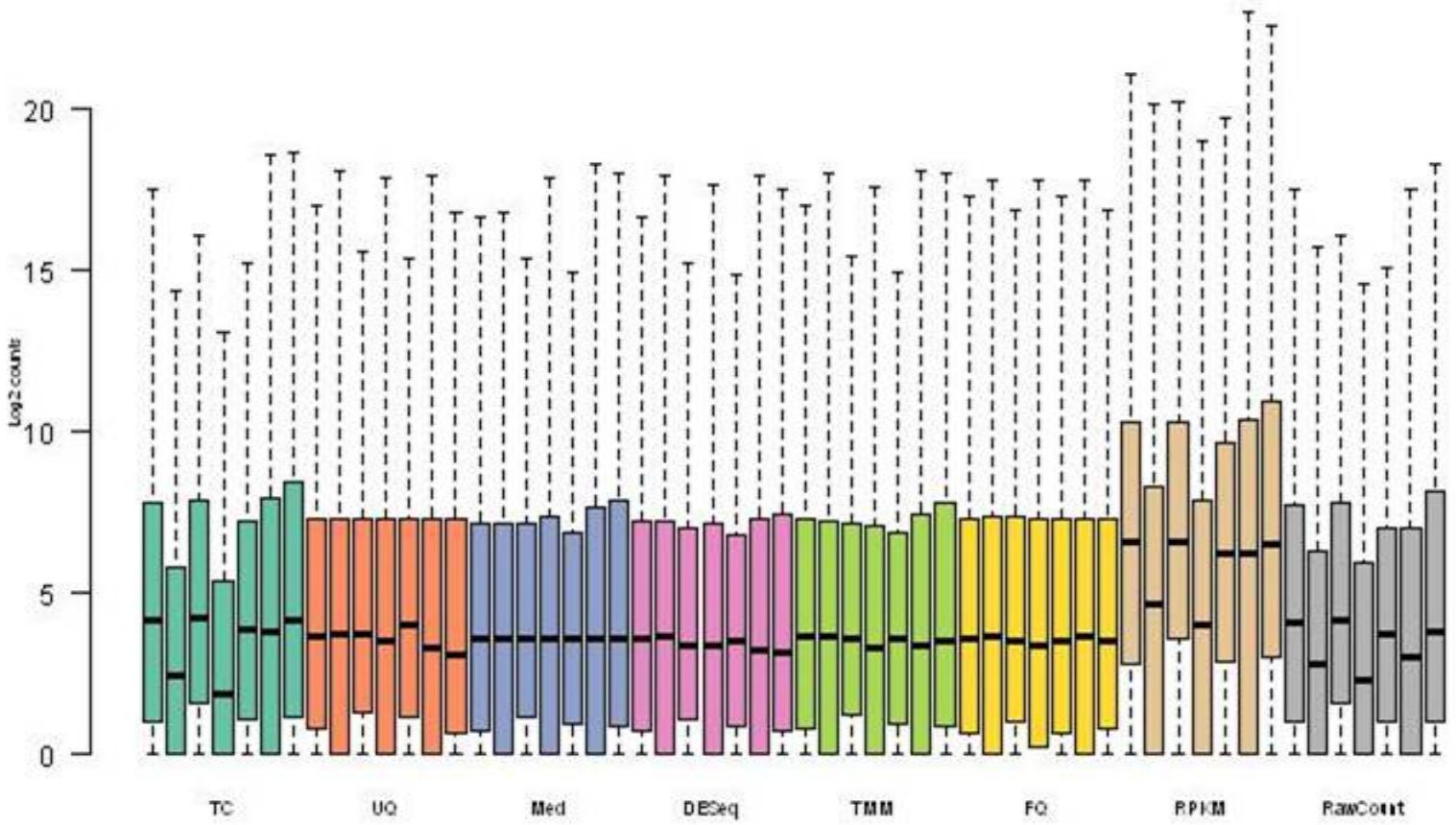
- Differential analysis: DESeq v1.6.1, default parameters
- Number of common DE genes, similarity between list of genes (dendrogram - binary distance and Ward linkage)

Power and control of the Type-I error rate

- simulated data
- non equivalent library sizes
- presence of majority genes

Normalized data distribution

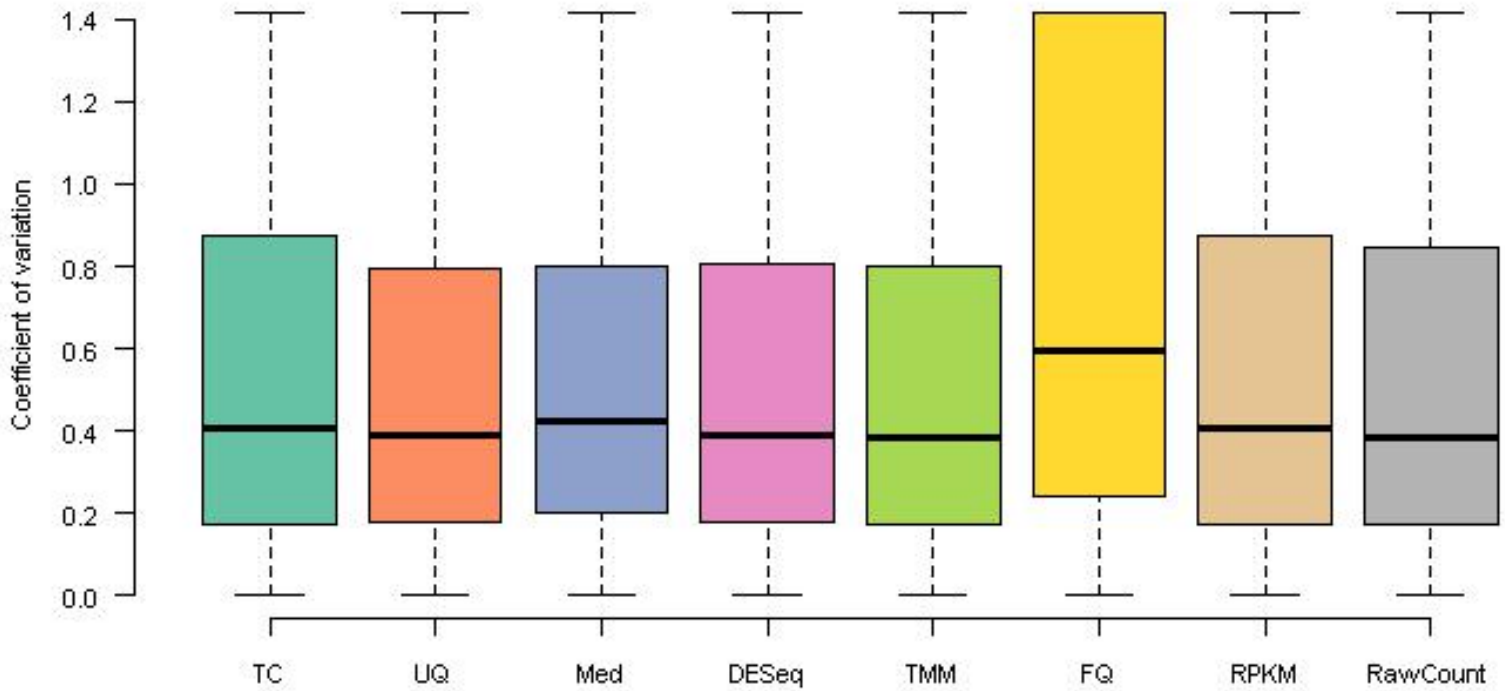
When large diff. in lib. size, TC and RPKM do not improve over the raw counts.



Example: *Mus musculus* dataset

Within-condition variability

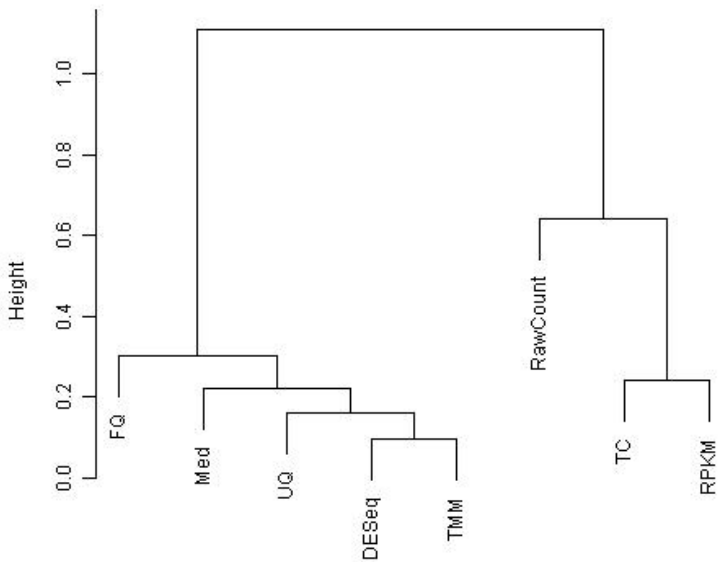
Example: *Mus musculus, condition D* dataset



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

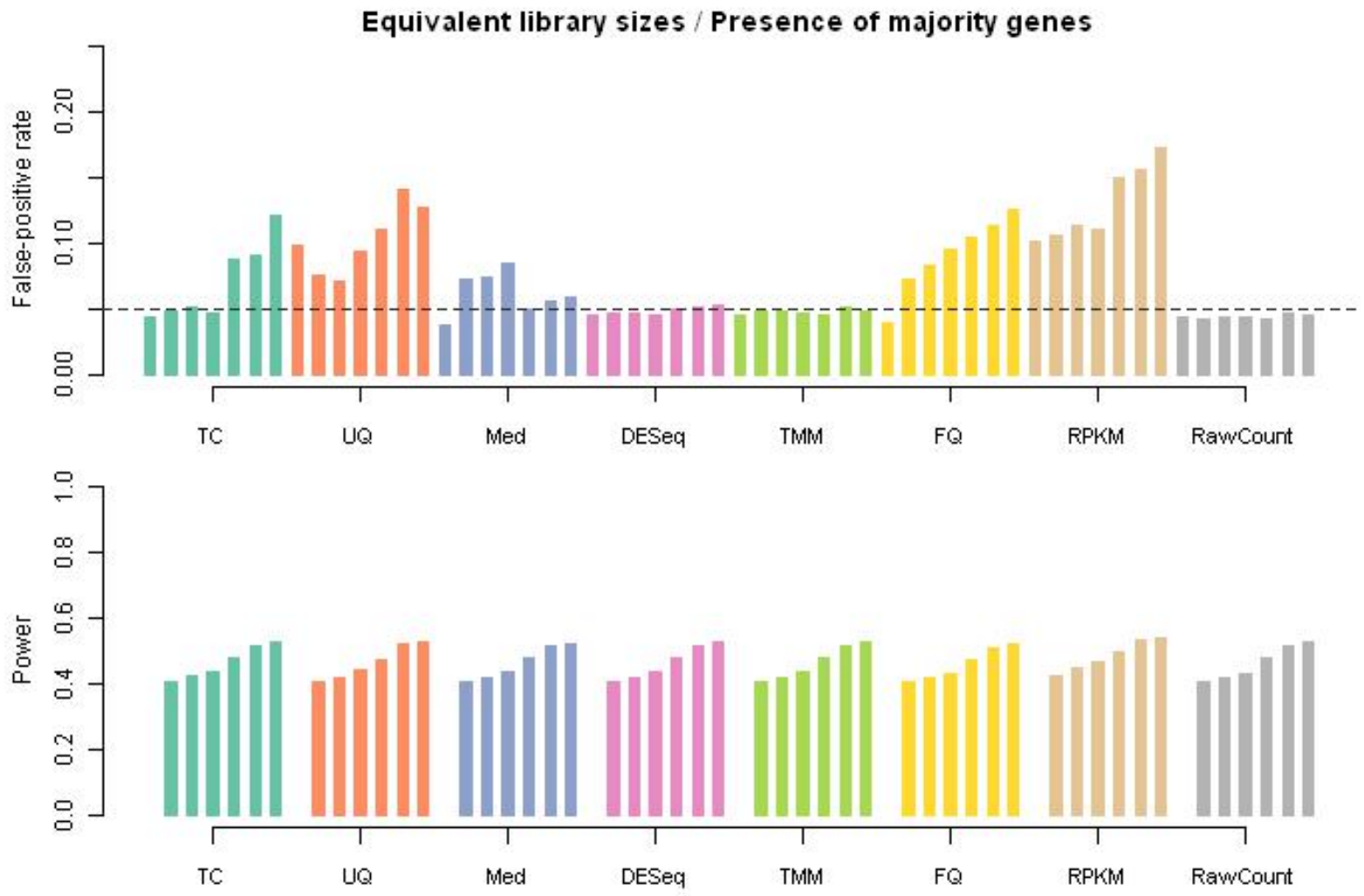


Consensus matrice

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



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	–	+	+	–	–

Conclusions on normalization before differential analysis

- Normalisation is necessary and not trivial
- Hypothesis : the majority of genes is invariant between samples.
- Differences between normalisation methods when genes with large number of reads and very different library depths.
- TMM and RLE : performant and robust methods in a DE analysis context on the gene scale
- Risso et al (2014) proposed the method RUVSeq, which is based on a factor analysis. The aim is to remove effects of unobservable covariates.

Normalisation TMM or DESeq is specific of the group of samples considered

gene	normalisation 1	normalisation 2
AT1G01010.1	137.8	117.2
AT1G01020.1	70.9	60.3
AT1G01030.1	126.0	107.1
AT1G01040.2	561.8	477.6
AT1G01050.1	1153.9	980.8
AT1G01060.1	3296.2	2801.7
AT1G01070.1	168.0	142.8
AT1G01080.2	876.9	745.3
AT1G01090.1	4733.7	4023.5
AT1G01100.1	3384.2	2876.5
AT1G01110.2	56.4	48.0
AT1G01120.1	1739.4	1478.4
AT1G01130.1	10.5	8.9
AT1G01140.3	938.6	797.8
AT1G01160.2	308.5	262.2
AT1G01170.1	535.6	455.2
AT1G01180.1	325.6	276.7
.....	.....	.....