Statistical analysis of RNA-seq data

Marie-Laure Martin-Magniette

IPS2 Institut des Sciences des Plantes de Paris-Saclay

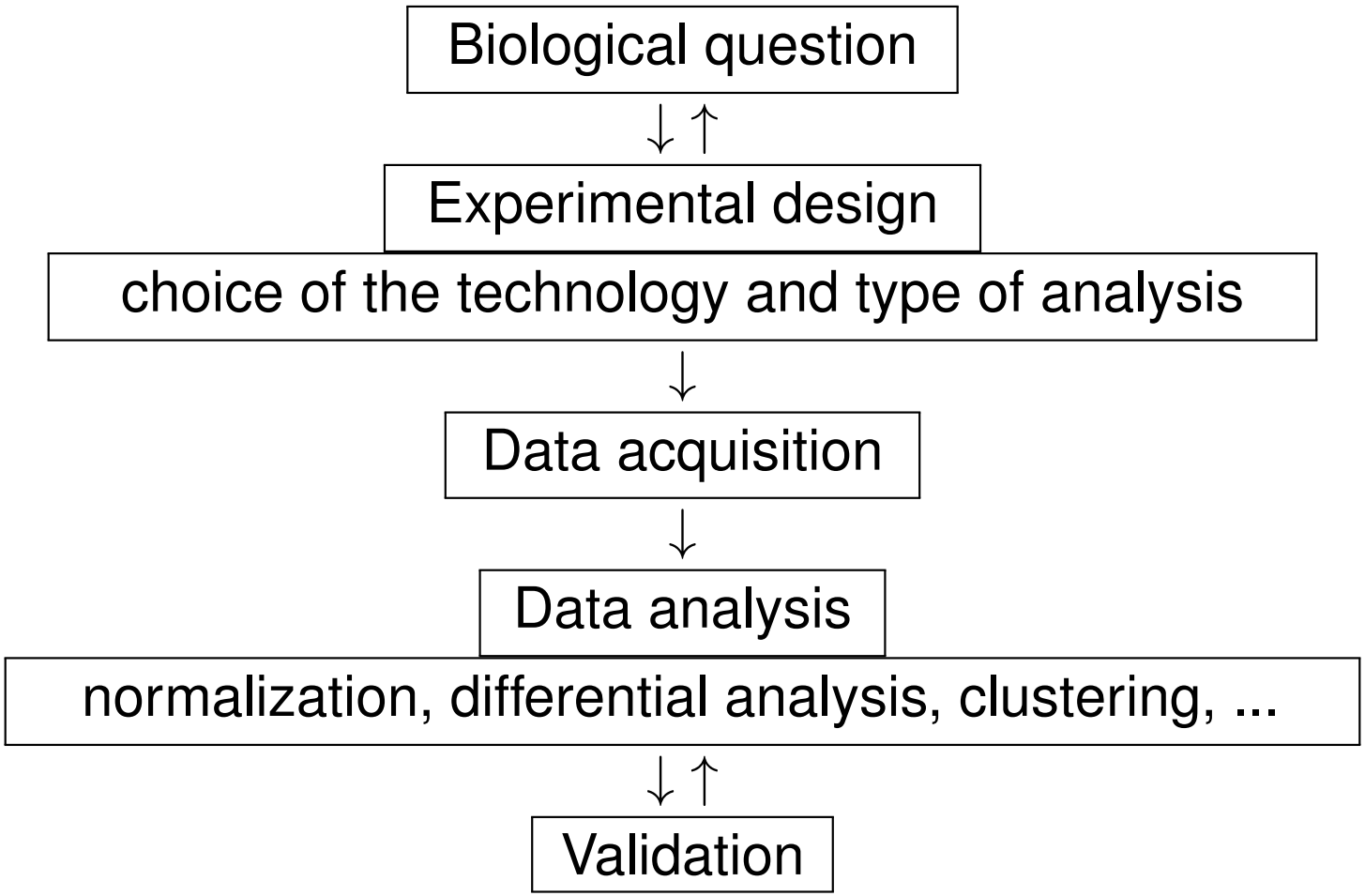
UMR AgroParisTech/INRA Mathematique et Informatique Appliquees

Table of contents

Aims of the lecture

- Quantitative analysis of gene expression
- Overview of the different steps of the analysis
- It is not exhaustive

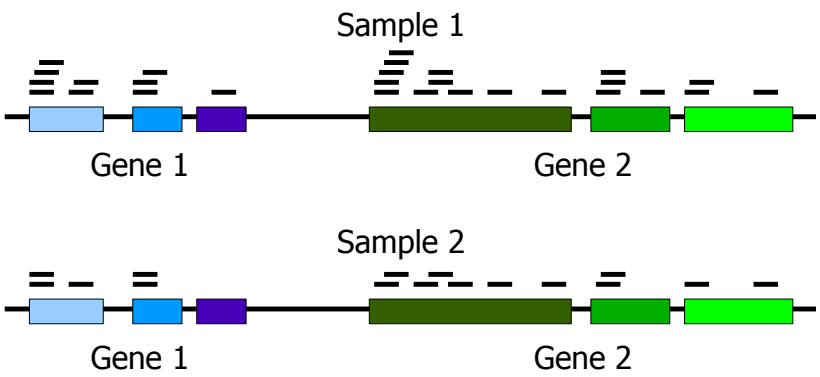
Design of a transcriptomic project



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
[...]			

To date, most methodological developments are for experimental design, normalization, and differential analysis...

Table of contents

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

Two types of 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)

Between-sample normalization: the scaling factor

Definition

For sample j, let Y_{gj} be the raw count for gene g.
The normalized count is defined by:

$$\frac{Y_{gj}}{s_j},$$

where s_j is the scaling factor for the sample j.

Three types of methods:

- Distribution adjustment
- Method taking length into account
- The Effective Library Size concept

Distribution adjustment

Let n be the number of samples in the project

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

$$s_j = \frac{N_j}{\frac{1}{n}\sum_{\ell=1}^{n} N_\ell}, \text{ where } N_j = \sum_g Y_{gj}$$

- Median

$$s_j = \frac{median_g\, Y_{gj}}{\frac{1}{n}\sum_{\ell=1}^{n} median_g\, Y_{g\ell}}$$

Method taking length into account

RPKM: Reads Per Kilobase per Million mapped reads

- Motivation greater lane sequencing depth and transcript length $=>$ greater counts whatever the expression level
- Assumption read counts are proportional to expression level, transcript length and sequencing depth (same RNAs in equal proportion)
- Method divide gene read count by total number of reads (in million) and transcript length (in kilobase)

$$\frac{Y_{gj}}{N_j * L_g} * 10^3 * 10^6 \quad (1)$$

- 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

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

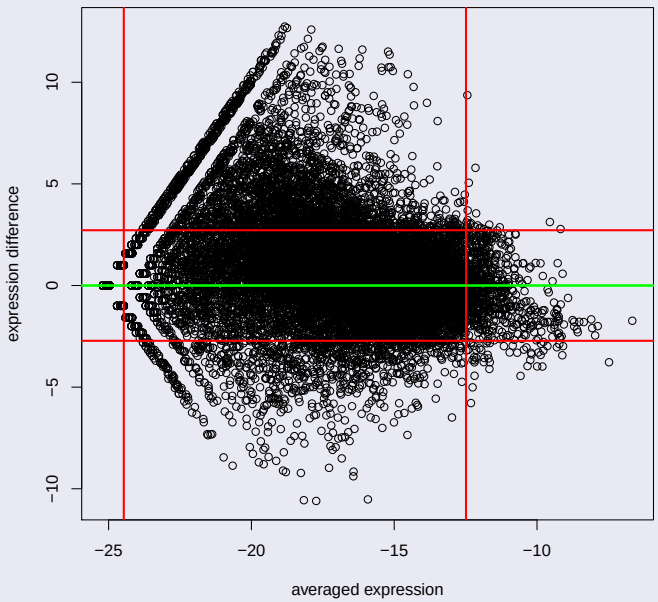
- normalize them such that their product equals 1

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

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 $w_{gj}^r = \left(\frac{N_j - Y_{gj}}{N_j Y_{gj}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}}\right)$
- 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

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

Which normalization method ?

At lot of different normalization methods...

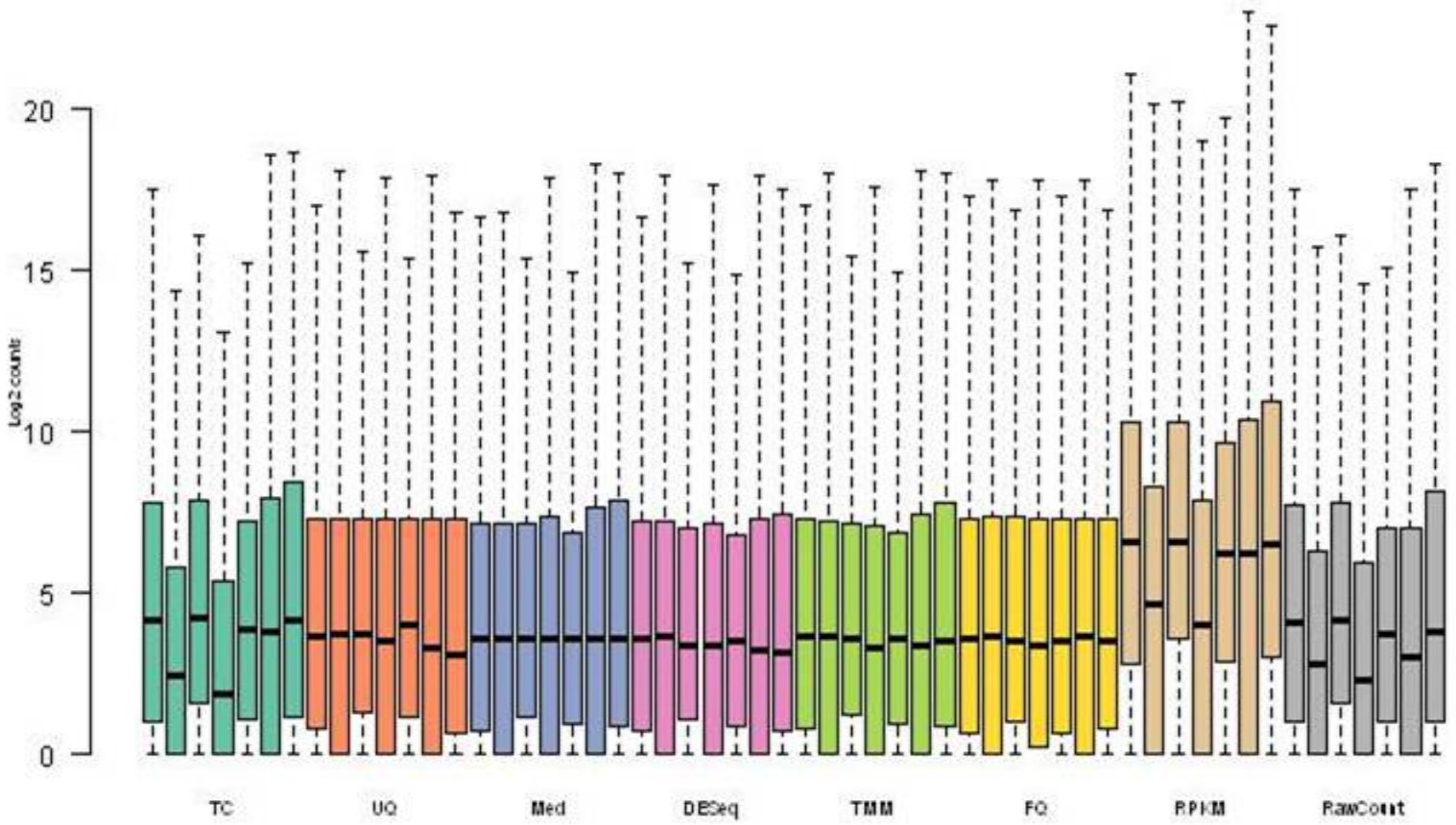
- Some are part of models for DE, others are 'stand-alone'
- They do not rely on similar hypotheses
- But all of them claim to remove technical bias associated with RNA-seq data

Questions

- Which one is the best ?
- Which criteria are relevant for this choice ?

Normalized data distribution

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



Example: *Mus musculus* dataset

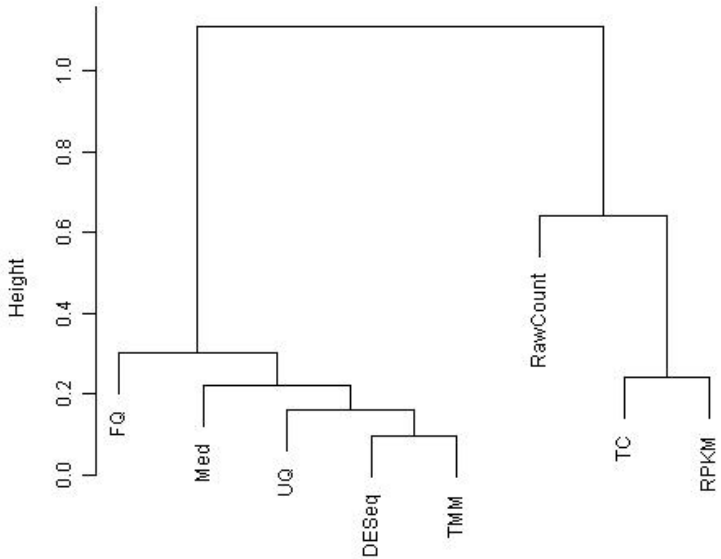
Lists of differentially expressed (DE) genes

For each dataset

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

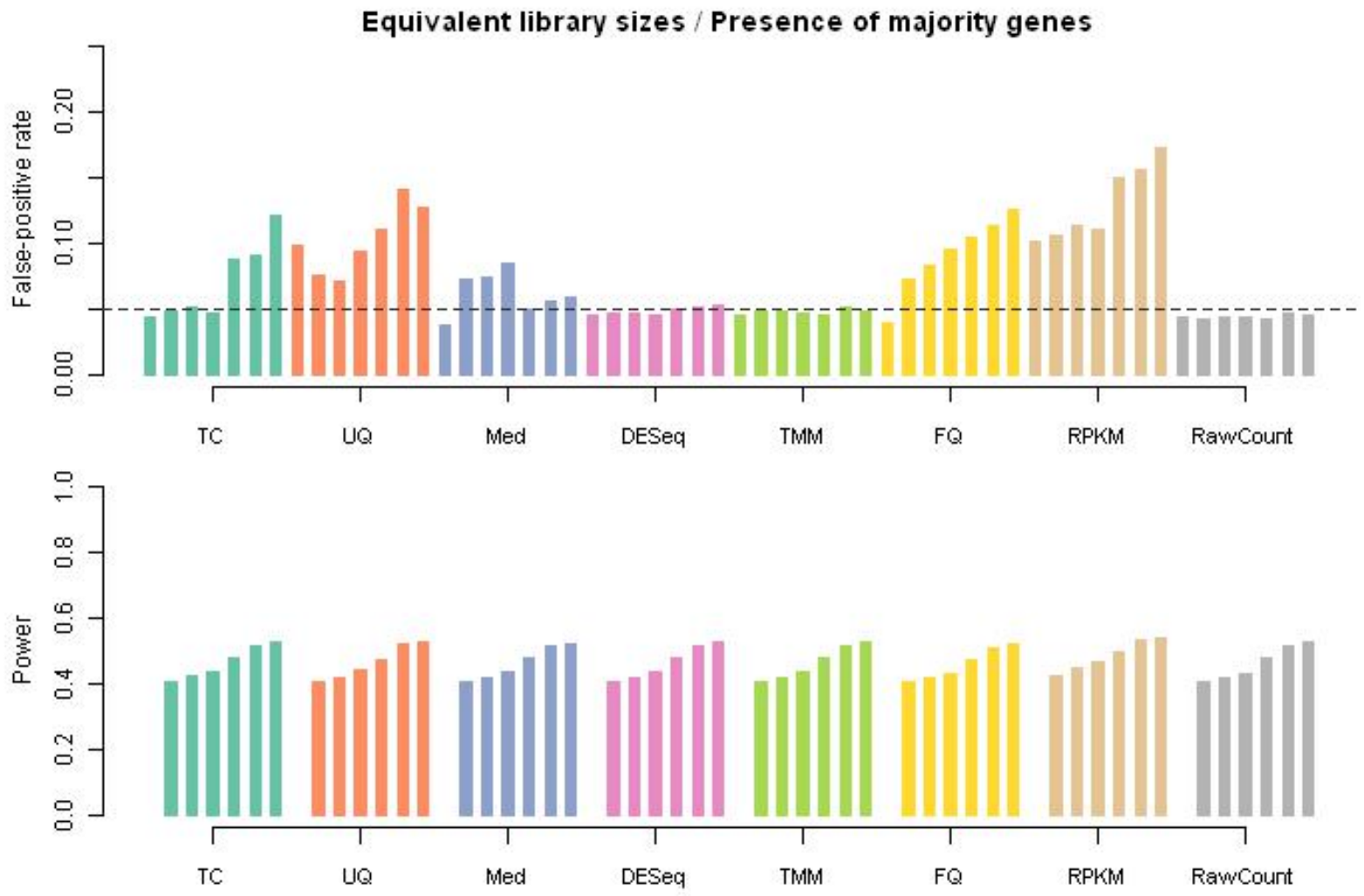
Consensus matrice

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

- RNA-seq data are affected by technical biaises (total number of mapped reads per lane, gene length, composition bias)
- Csq1 : non-uniformity of the distribution of reads along the genome
- Csq2 : technical variability within and between-sample
- Normalization by gene length isn't required for the differential analysis.
- Normalisation is necessary and not trivial.

Conclusions on normalization

- Differences between normalisation methods when genes with large number of reads and very different library depths
- TMM and DESeq : performant and robust methods in a differential analysi context
- Risso et al (2014) proposed a new method (RUVSeq). It is a factor analysis based on a suitably-chosen subset of negative control genes

Table of contents

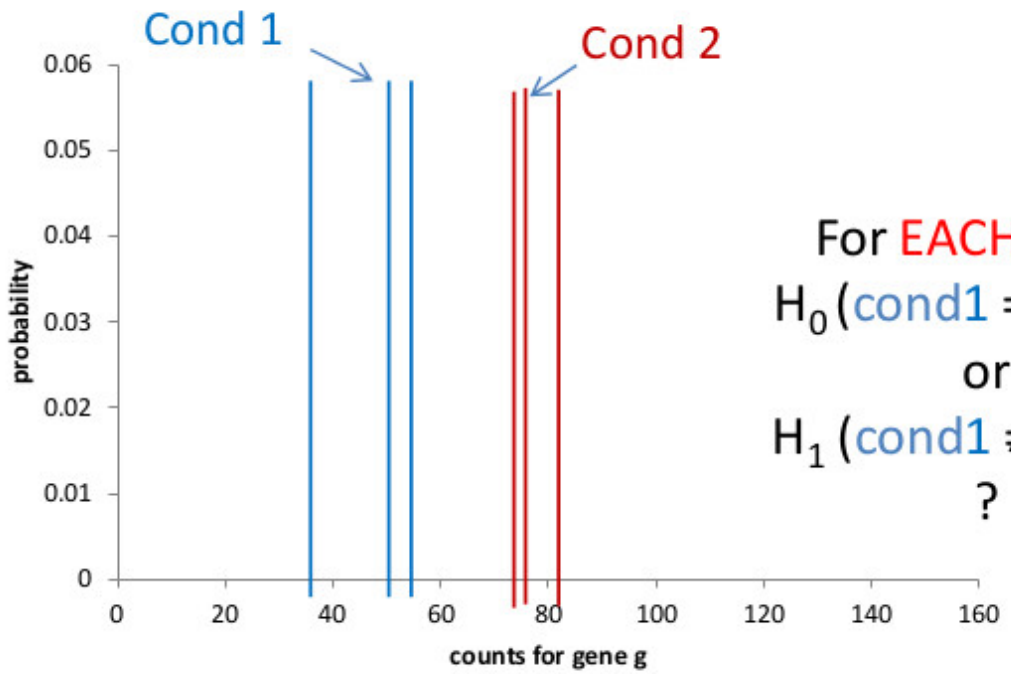
Key steps for a test procedure

- The aim is to identify a significant difference between two means
- Required data are measurements in the two conditions
- It is performed with an hypothesis test

$$H_0 = \{\text{There is no difference}\}$$
$$\text{versus}$$
$$H_1 = \{\text{There is a difference}\}$$

- Construct a test statistic (a function of the observations)
- Decide the most relevant hypothesis with respect to the test statistic

Statistical hypothesis test



For EACH gene:

H_0 (cond1 = cond2)

or

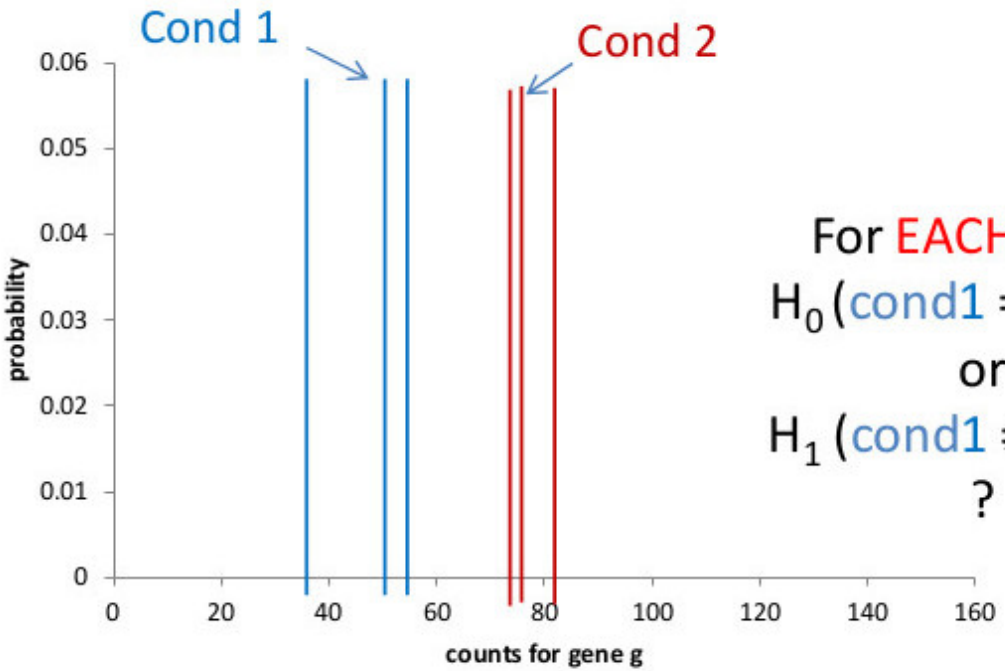
H_1 (cond1 ≠ cond2)

?

Technical and biological variabilities

Cond1 and Cond2 measurements are variable

Statistical hypothesis test



For EACH gene:

H_0 (cond1 = cond2)

or

H_1 (cond1 ≠ cond2)

?

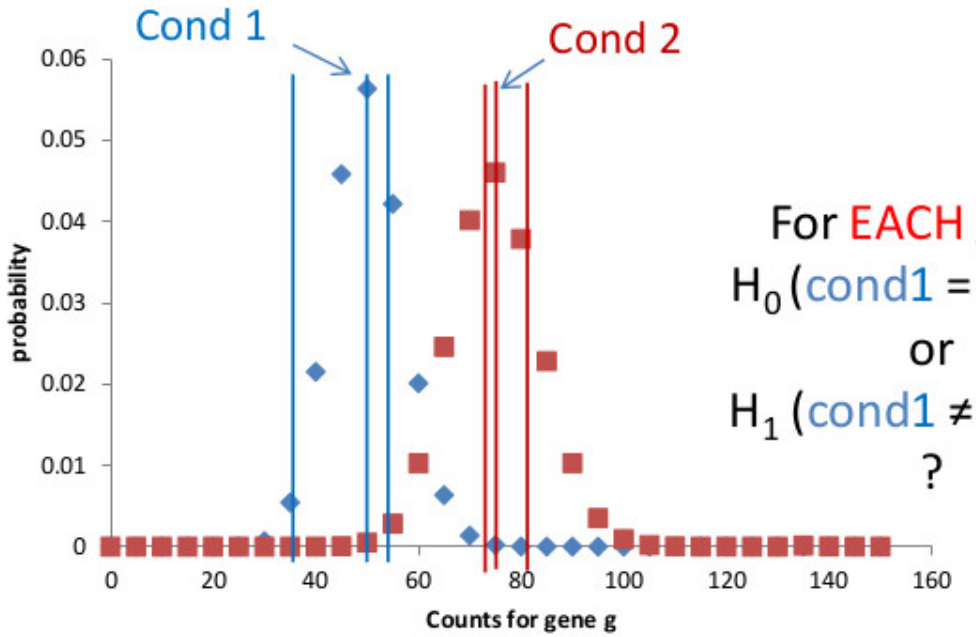
Technical and biological variabilities

Cond1 and Cond2 measurements are variable

Not enough replicates

Estimating the variability of cond1 and cond2 for each gene by modeling

Hypothesis test: modeling the variability



For EACH gene:

H_0 (cond1 = cond2)

or

H_1 (cond1 ≠ cond2)

?

Two approaches: exact test or test within a Generalized Linear Models (GLM)

Both required to estimate the mean and the dispersion

Decision table

	H_0 true no difference	H_0 false difference
Do not reject H_0	Right decision	Wrong decision type II error
Reject H_0	Wrong decision type I error	Right decision

Acceptable errors

- In a type I error, the null hypothesis is really true but the statistical test has led you to believe that it is false. It is a false positive.
- In a type II error, the null hypothesis is really false but the test has not picked up this difference. It is a false negative.

Decision rule

Construction of a test

- Formulate the two hypotheses
- Construct the test statistic
- Define its distribution under the null hypothesis
- Calculate the p-value
- Decide if the null hypothesis is rejected or not

Definition of a p-value

It is the probability of seeing a result as extreme or more extreme than the observed data, when the null hypothesis is true

Decision rule

The null hypothesis is rejected if the p-value is lower than a given threshold

Multiple testing

- The result of a test can be viewed as a random variable:

  0 if the result is a true positive
  1 if the result is a false positive

- By definition, P(to be a false positive)=α

Question

- Perform G=10000 tests at level α
- What is the expected number of false-positive ?

Contingency table for multiple hypotesis testing

	True null hypotheses	False null hypotheses	
Declared non-significant	True Negatives	False Negatives	Negatives
Declared significant	False Positives	True Positives	Positives

P-value adjustment

Adjust the raw p-values to control

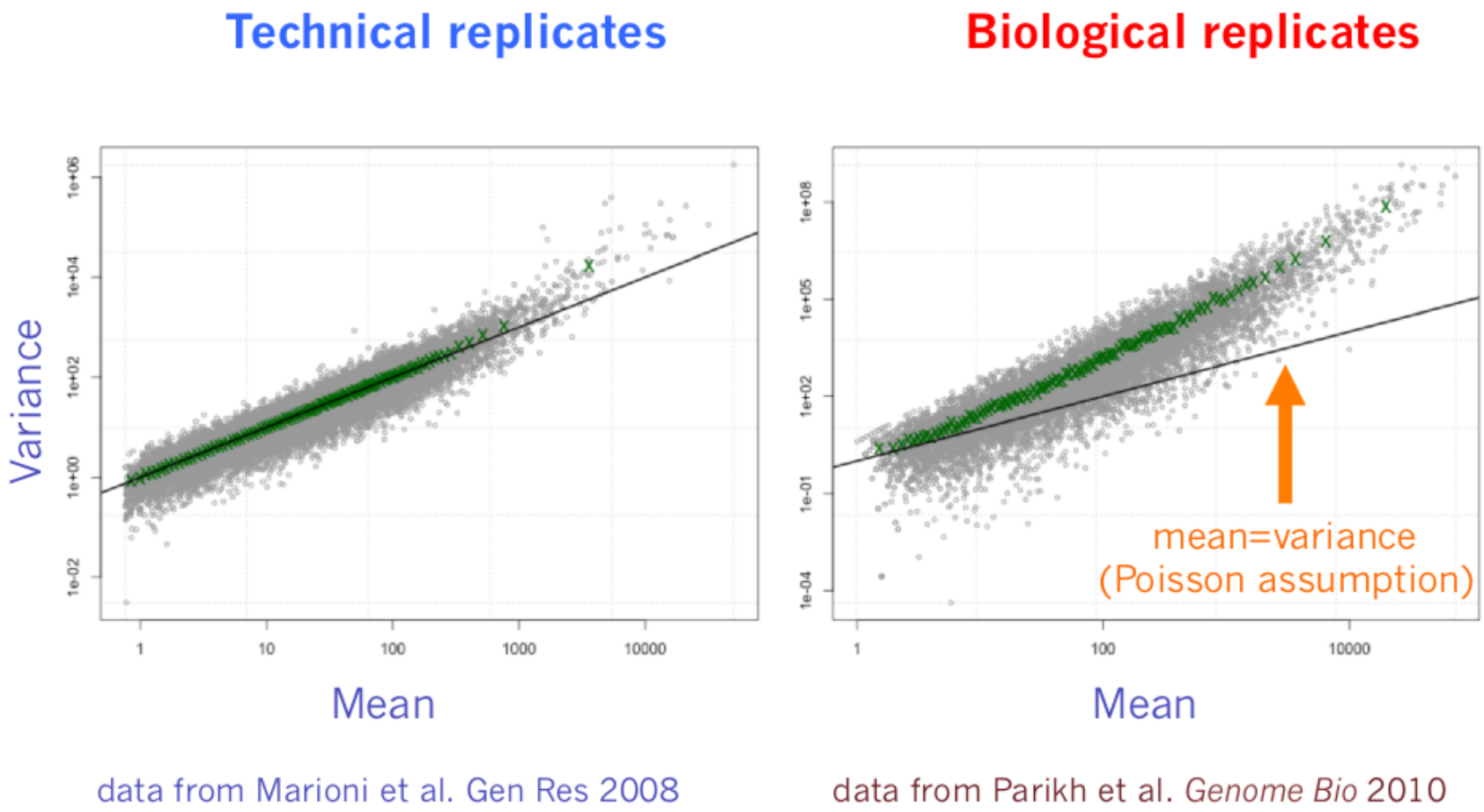
- $FWER = P(FP > 0)$ (Bonferroni procedure)
- $FDR = E(FP/P) if P > 0$ or 1 otherwise (Benjamini-Hochberg procedure)

Calculating adjusted p-values

- Start with (unadjusted) p-values for G hypotheses
  1. Order the p-values $p_{(1)} < \ldots < p_{(G)}$
  2. Multiply each $p_{(i)}$ by its adjustment factor
     - Bonferroni: $a_i = G$
     - Benjamini-Hochberg : $a_i = \frac{G}{i}$
  3. Let $p_{(i)} = a_i p_{(i)}$
  4. Set $p_{(i)}$ = min($p_{(i)}$, 1) for all i

Test for RNA-seq data

Distribution for count data



From D. Robinson and D. McCarthy

Notations and framework

- Let $Y_{g11}, \ldots, Y_{g1n}$ be independent counts from condition 1,

$$Y_{g1\ell} \sim NB(s_\ell \lambda_{g1}, \phi_g)$$

- Let $Y_{g21}, \ldots, Y_{g2n}$ be independent counts from condition 2,

$$Y_{g2r} \sim NB(s_r \lambda_{g2}, \phi_g)$$

- s_ℓ the library size of sample ℓ
- λ_{gj} the proportion of the library for this particular gene g in condition j.
- We want to test

$$H_0 = \{\lambda_{g1} = \lambda_{g2}\} \text{ vs } H_1 = \{\lambda_{g1} \neq \lambda_{g2}\}$$

- Need to estimate λ_{g1}, λ_{g2} and ϕ_g

Exact Negative Binomial Test
Robinson & Smyth (2008) Biostatistics

If librairie sizes are equal (Anderson & Boullion, 1972)

- $s_\ell = s$ for $\ell = 1, \ldots, n$
- $Y_{g11} \ldots + Y_{g1n} \sim NB(ns\lambda_{g1}, \phi_g/n)$
- $Y_{g21} \ldots + Y_{g2n} \sim NB(ns\lambda_{g2}, \phi_g/n)$
- There exists a sufficient statistic for $\lambda_{g1}, \lambda_{g2}$
- ϕ_g can be estimated independently from $\lambda_{g1}, \lambda_{g2}$

The normalisation is performed to get librairies with equal size

Limitations of the exact test

- It is assumed that $m_{g1\ell} = s_\ell \times \lambda_{g1}$
- Consequently it is only developed when two conditions are available
- However reality is more complex
- Even in simple design, replicates can vary with others factors such as time of sampling, ...

GLM framework for RNAseq data

- Let Y_{gjvr} be the counts of reads for gene g in the sample described by the uplet (j, v, r)
- Generalized Linear Model allows to decompose a function of the mean of the observations

We assume

$$Y_{gjvr} \sim NB(\mu_{gjvr}, \phi_g)$$

with

$$\log(\mu_{gjvr}) = \log(s_{jvr}) + log(\lambda_{gjvr})$$

where

- s_{jvr} is the library size for sample described by (j, v, r)
- $\log(\lambda_{gjvr}) = f(g, j, v, r)$ for example

$$\log_2(\lambda_{gjvr}) = Intercept + \alpha_{gr} + \beta_{gj} + \gamma_{gv} + \delta_{gjv}$$

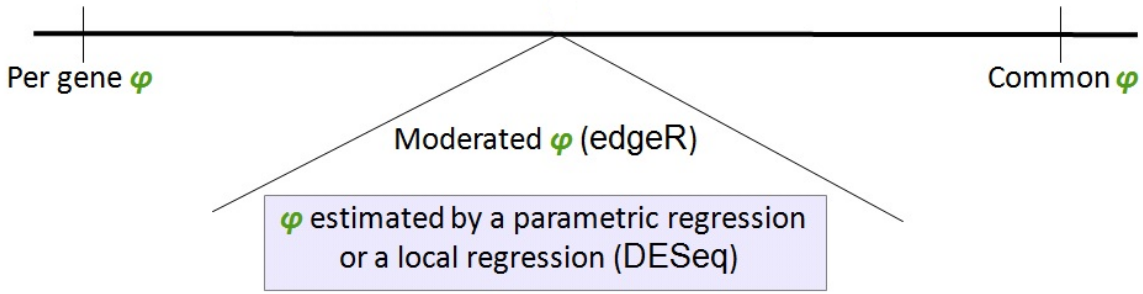
Inference and test

- Parameters are those describing the mean and the dispersion ϕ_g.
- They are estimated by the quasi-likelihood estimators: ϕ_g are estimated and parameters describing the mean are then estimated by maximizing a function of the data and the dispersion
- Test based on the likelihood ratio test or the Wald test

Dispersion estimation

Few replicates to accurately estimate the dispersion parameter

- DESeq: ϕ_g is a smooth function of $\lambda_g = \lambda_{g1} = \lambda_{g2}$
- edgeR: empirical Bayesian procedure to estimate ϕ_g
- ... and many, many more methods!



- Soneson & Delorenzi (2013, BMC Bioinf) compared 11 methods using simulations:
  No single method is optimal under all circumstances ...
- Nookaew et al. (2012, NAR) compared microarray & RNA-seq differential analyses:
  Importance of mapping to estimate gene expression level

Example

- Consider a project where a wild-type plant and three mutants are studied
- Available data are three biological replicates of the gene expression of each type of plant
- The aim is to find genes affected by each mutation
- How can you answer this question ?

Three statistical frameworks

- A negative binomiale distribution (2008)
  - Expression = library size $\times \lambda_{condition}$
- A Generalized Linear Model (2012)
  - allows us to decompose the expression
  - each condition is described by several factors

$$\log(\lambda_{condition}) = Cst + \alpha_{genotype} + \beta_{stress} + \gamma_{genotype,stress}$$

  - Effect of each factor is tested
- A linear model (2014)
  - data are transformed to work with a Gaussian
  - allows us to decompose the expression

In practice

- Do we filter genes with low expression (yes or no)
- How to model the gene expression (NB, GLM or LM and which factors are important to consider ?)
- Which method to estimate the variance of the gene expression (several methods)

Large evaluation study

We want to answer these questions

- How the statistical models fit RNA-seq data ?
- Are DE genes discriminated from NDE genes ?
- Are the false-positives controlled ?
- Are the methods powerful (able to find the truly DE genes)

Which kind of data is relevant for an evaluation ?

- **Real data**:
  - More realistic
  - ... but no extensively validated data yet available
- **Simulated data**:
  - Truth is well-controlled
  - ... but what model should be used to simulate data? How realistic are the simulated data? How much do results depend on the model used?

Which kind of data is relevant for an evaluation ?

- **Real data**:
  - More realistic
  - ... but no extensively validated data yet available
- **Simulated data**:
  - Truth is well-controlled
  - ... but what model should be used to simulate data? How realistic are the simulated data? How much do results depend on the model used?

Our idea was to create synthetic data

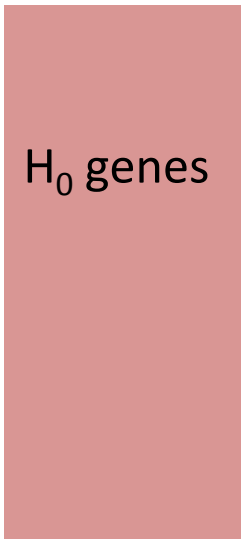
Creation of synthetic datasets

Leaves vs Leaves

H_0 full dataset

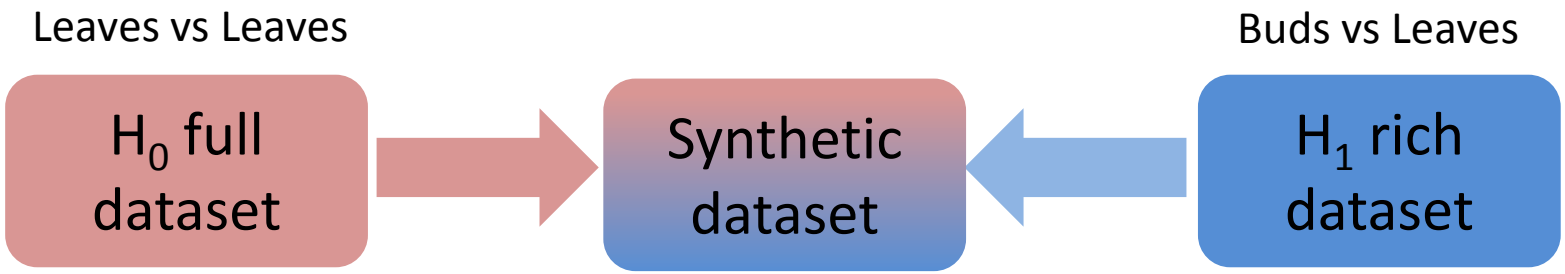
Buds vs Leaves

H_1 rich dataset

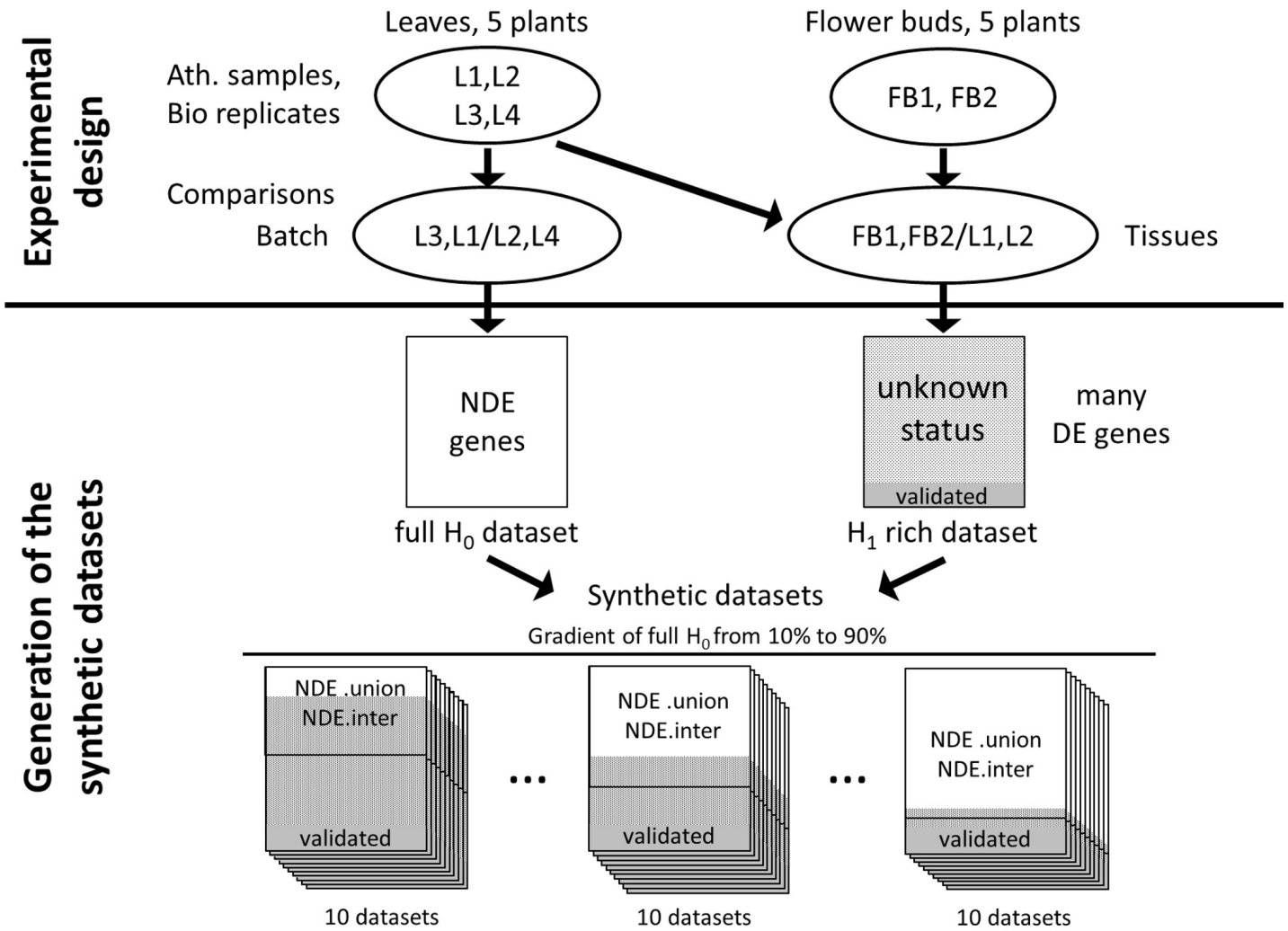


qRT-PCR

Creation of synthetic datasets



Creation of synthetic datasets



Definition of the truth

the set of truly DE genes

251 DE genes identified by qRT-PCR among 332 randomly chosen genes

the set of truly NDE genes

- The proper identification is not straightforward
- Definition of two sets
- NDE.union: may include some genes that are not truly NDE
- NDE.inter: may exclude some truly NDE genes.

The 3 frameworks described by 9 methods

- **edgeR** and **DESeq** are NB-based method

$$\text{Expression} = \text{library size} \times \lambda_{tissue}$$

- **glm edgeR** and **DESeq2** are GLM approaches

$$\log(\lambda_{condition}) = Cst + \alpha_{tissue} + \beta_{biological\ replicate}$$

- **limma-voom** is a linear model

  Data are transformed with the voom method

$$\text{Expression} = Cst + \alpha_{tissue} + \beta_{biological\ replicate}$$

- All methods except DESeq are also applied on filtered data
- In each method, FDR is fixed at 5 %

Evaluation criteria

- How the statistical models fit RNA-seq data ?

→ study of the p-value distribution

- Do p-values well discriminate DE and NDE genes ?

→ ROC curves

- Are the false-positives controlled ?

→ proportion of truly NDE declared DE

- Are the methods powerful (able to find the truly DE genes)

→ proportion of truly DE declared DE

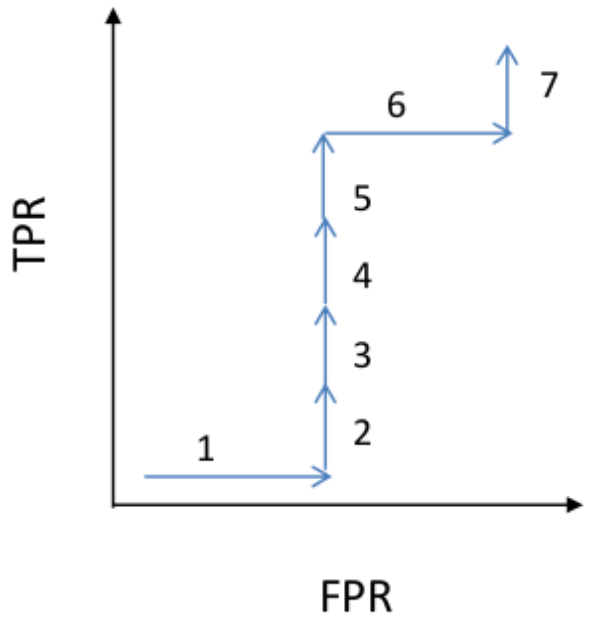
Definition of a ROC curve

Drawing a ROC curve:
1- sort genes by increasing raw p-value
2- knowing the truth (DE or NDE) for each gene, go down the sorted list counting the proportion of all the DE genes encountered so far (TPR) and the proportion of all the NDE genes encountered so far in the list (FPR)

Example:

7 genes: 5 DE and 2 NDE

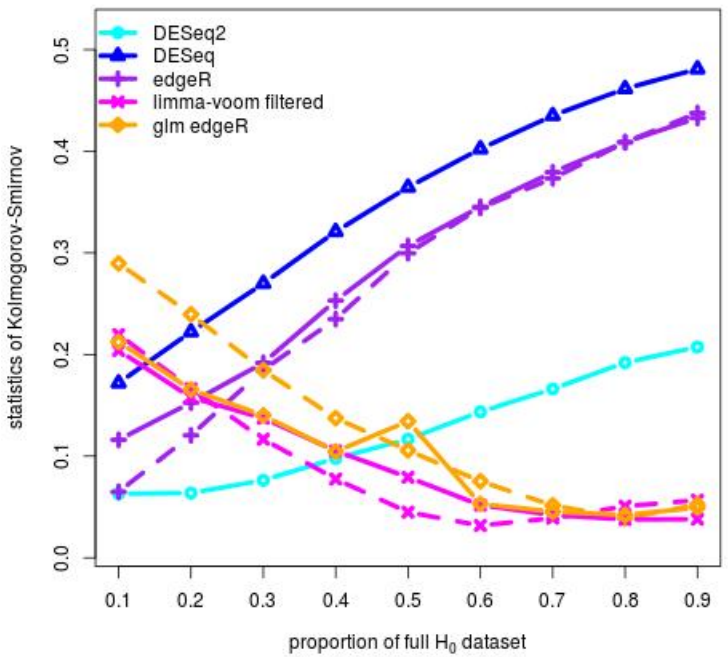
rank	gene	p-value	truth	TPR	FPR
1	G1	p1	NDE	0/5	1/2
2	G2	p2 (>p1)	DE	1/5	1/2
3	G3	p3(>p2)	DE	2/5	1/2
4	G4	p4(>p3)	DE	3/5	1/2
5	G5	p5(>p4)	DE	4/5	1/2
6	G6	p6(>p5)	NDE	4/5	2/2
7	G7	p7(>p6)	DE	5/5	2/2



Distribution of the p-values

Method

- When no difference is expected, histogram of the p-values are expected to be uniform histogram
- For each synthetic dataset, 100 evaluations of the uniform distribution of 1000 genes randomly chosen in the full H_0 dataset are performed

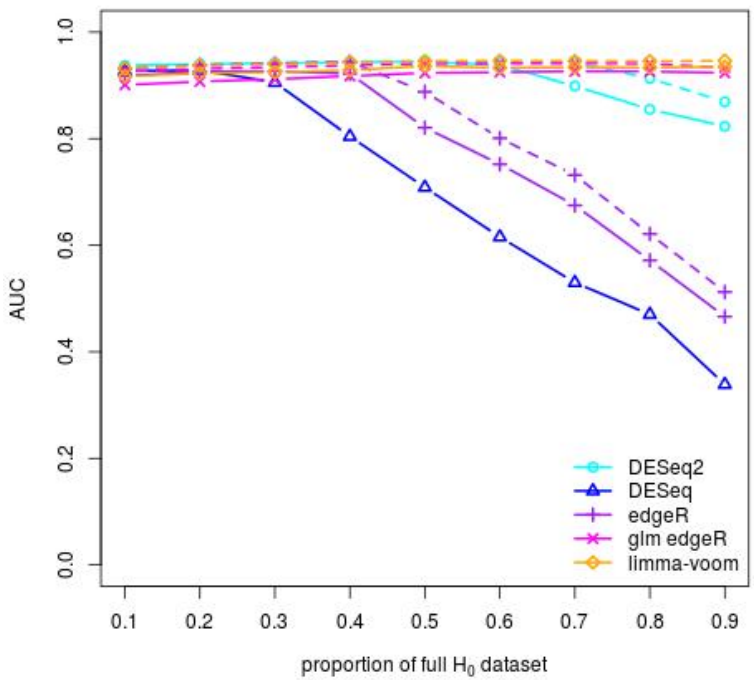


- the raw p-values are not properly calculated (67% of tests are rejected after a strict FP control)
- test statistic values are smaller for linear or generalized linear models

Discrimination of DE and NDE genes

Method

- sort raw p-values into ascending order
- compare them with the truth
- construct a ROC curve and calculate AUC
- AUC close to 1 indicates a good discrimination

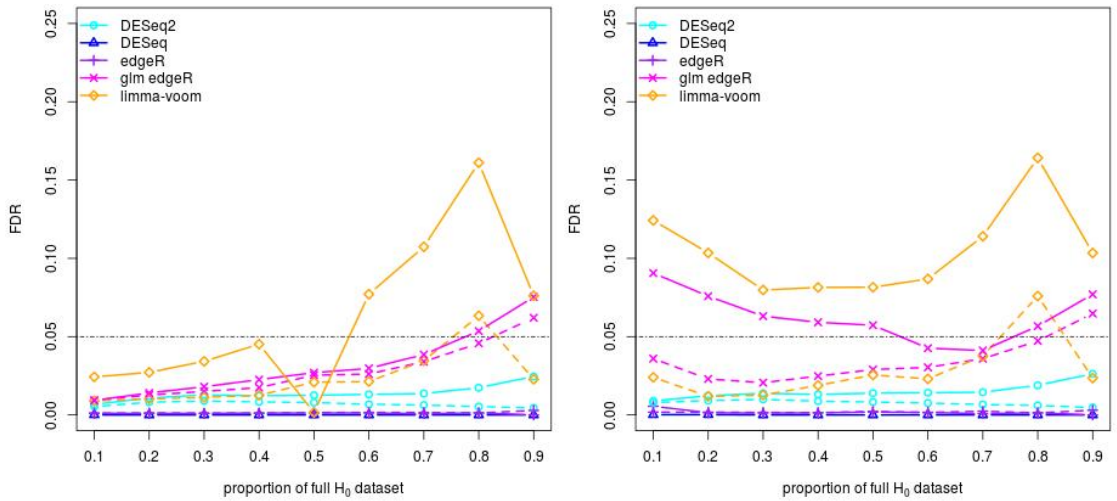


- For linear model or glm, the AUC is high and independent of the proportion of full H0 datasets
- For NB-based method, the AUC steadily decrease with the increase of the proportion of full H0 dataset when it is larger than 0.3-0.4

FDR estimation

Method

Proportion of truly NDE among the declared DE
Expected value : 5%

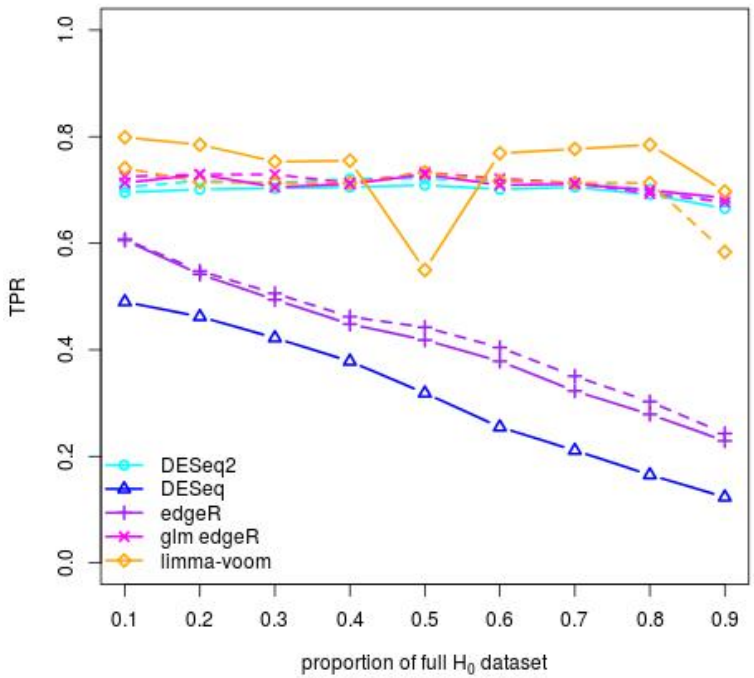


- For NB-based method, both bounds are close to 0
- For DESeq2, the FDR is always lower than 5%
- For glm edgeR, the interval generally contains 5%
- For limma-voom, the FDR control is more variable but the filtering step stabilizes its behavior

Are truly DE declared DE ?

Method

Proportion of truly DE genes among the declared DE genes

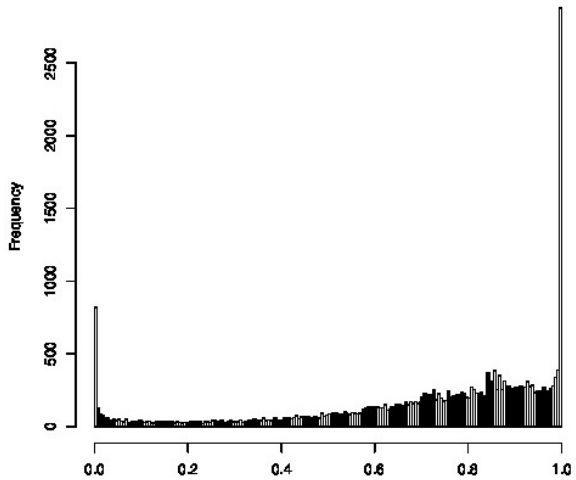


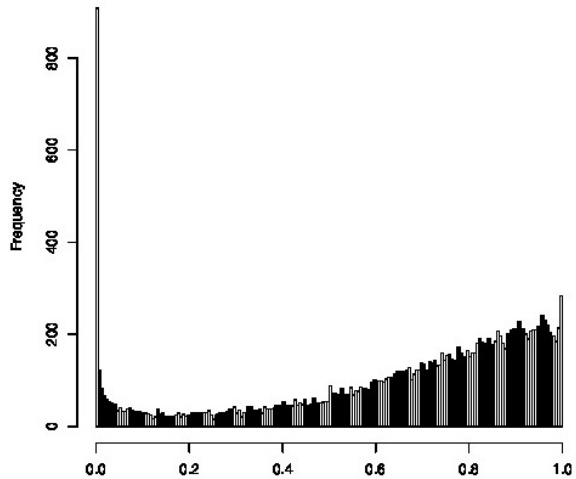
- LM or GLM based-methods show a high TPR
- For NB-based methods, the TPR is a function of the full H0 dataset proportion.
- The variance-mean relationship modeling and the data filtering seem to have only a limited impact.

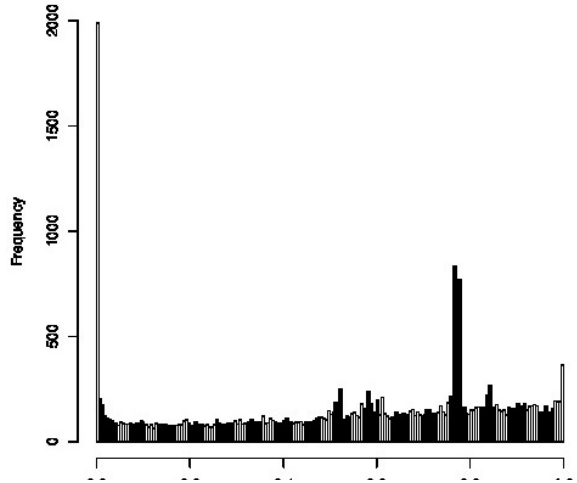
Definition of an indicator of quality

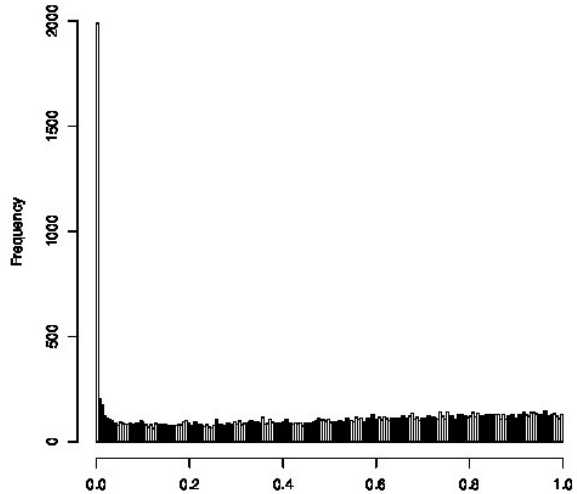
An histogram with a peak at the right side = analysis of bad quality

Let's play a game : which analysis is correct ?









Conclusions

$$\text{modeling} \geq \text{filtering} \geq \text{dispersion}$$

Synthetic data are a relevant framework

- Forget edgeR and DESeq
- use glm edgeR, DESeq2 or limma-voom
- include biological replicate as a factor
- filtering allows methods to control FDR