Key ingredients for RNA-seq differential analysis

Neutral comparison study - The importance of modeling

M-L Martin-Magniette, E. Delannoy (and G. Rigaill)

This is not a tutorial on how to analyse RNAseq data

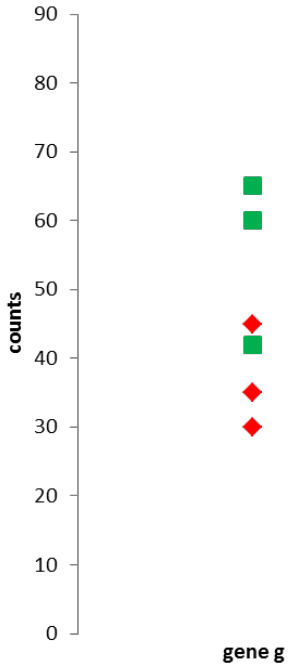
- A good tutorial can be found here: `http://www.nathalievilla.org/doc/pdf/slides-rnaseq.pdf`
- It covers
  - Exploratory analysis
  - Normalization
  - Differential Expression analysis

What this is about

- Here we will focus on the differential analysis
  - when only a few replicates are available
- What are the important parameters in ”practise” ?
  - a strategy to identify those on your own data ?
- This is part of a course on differential analysis taught by E. Delannoy and M.L Martin-Magniette to biologists and bioinformaticians at
  - IPS2 : ”Ecole chercheur SPS”
  - LIPM : ”Analyse de donnes ’omiques”
- Details can be found in the paper:
  `https://www.ncbi.nlm.nih.gov/pubmed/27742662`
- Main conclusion ”Modeling is important”!

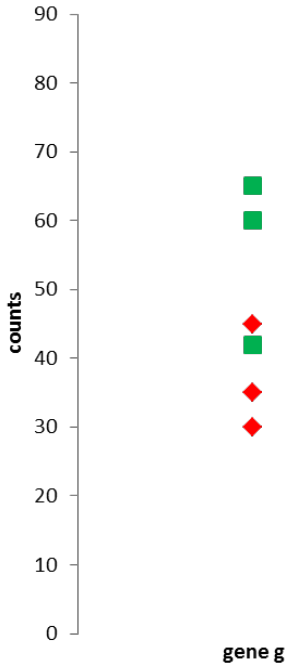
Objective of the differential analysis

- The aim is to identify a significant difference of expression between two given conditions
- It is performed with an hypothesis test based on gene expression measurements



Objective of the differential analysis

- The aim is to identify a significant difference of expression between two given conditions
- It is performed with an hypothesis test based on gene expression measurements



H_0={There is no difference}: red = green

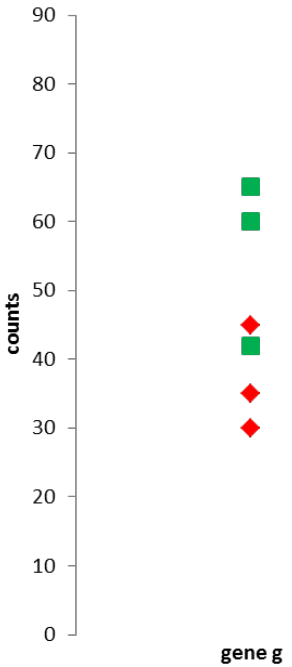
versus

H_1={There is a difference}: red $\neq$ green

Key steps for a test procedure

Construction of a test

- Formulate the two hypotheses



H_0={There is no difference}: red = green

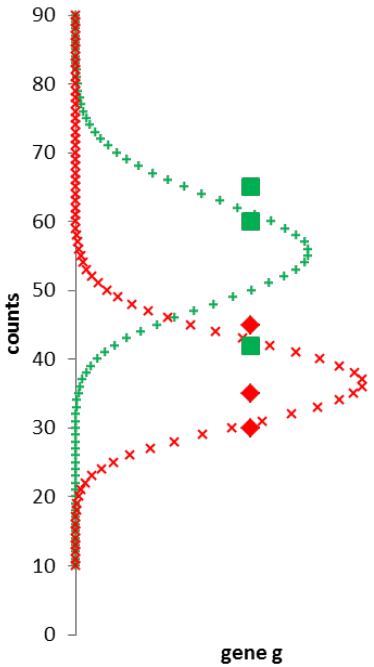
versus

H_1={There is a difference}: red $\neq$ green

Key steps for a test procedure

Construction of a test

- Formulate the two hypotheses
- Construct the test statistic
- Define its distribution under the null hypothesis



H_0={There is no difference}: red = green

versus

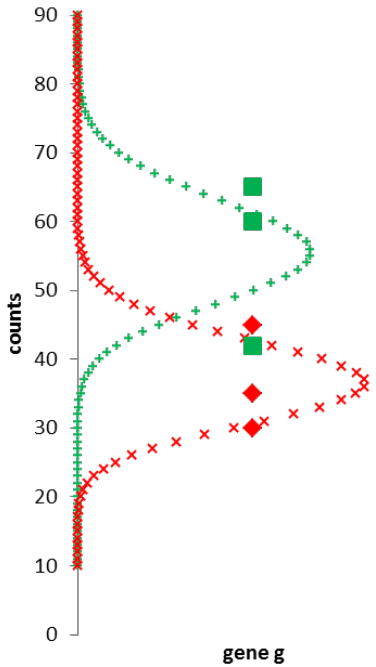
H_1={There is a difference}: red $\neq$ green

Estimation of the variance of expression for gene g

Key steps for a test procedure

Construction of a test

- Formulate the two hypotheses
- Construct the test statistic
- Define its distribution under the null hypothesis



H_0={There is no difference}: red = green

versus

H_1={There is a difference}: red $\neq$ green

Application of the test

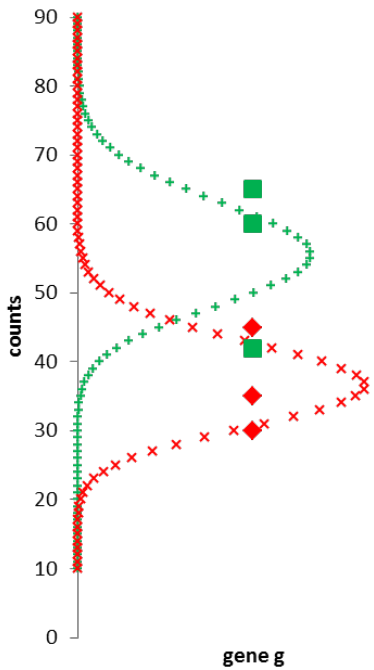
- Calculate the value of test statistic
- Calculate the p-value
- Decide if H_0 is rejected or not

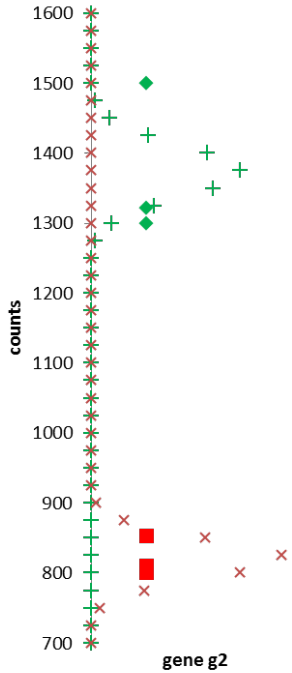
Definition of a p-value

It is the probability of seeing a result at least as extreme as the observed data, when the null hypothesis is true.

Multiple genes, multiple testing

Apply the previous procedure to every tested gene





P-value and multiple testing

- By definition, P(to be a false positive)=α
- If 10.000 tests are performed at level α, then the averaged number of false-positives is 500

Contingency table for multiple hypothesis testing

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

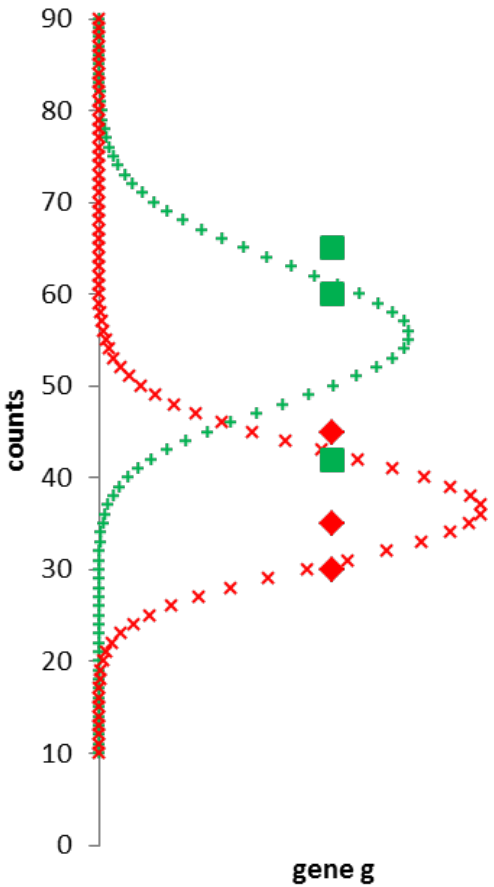
Adjustment of the raw p-values

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

Decision rule

A gene is declared differentially expressed if its adjusted p-value is lower than a given threshold

How to model RNA-seq data ?



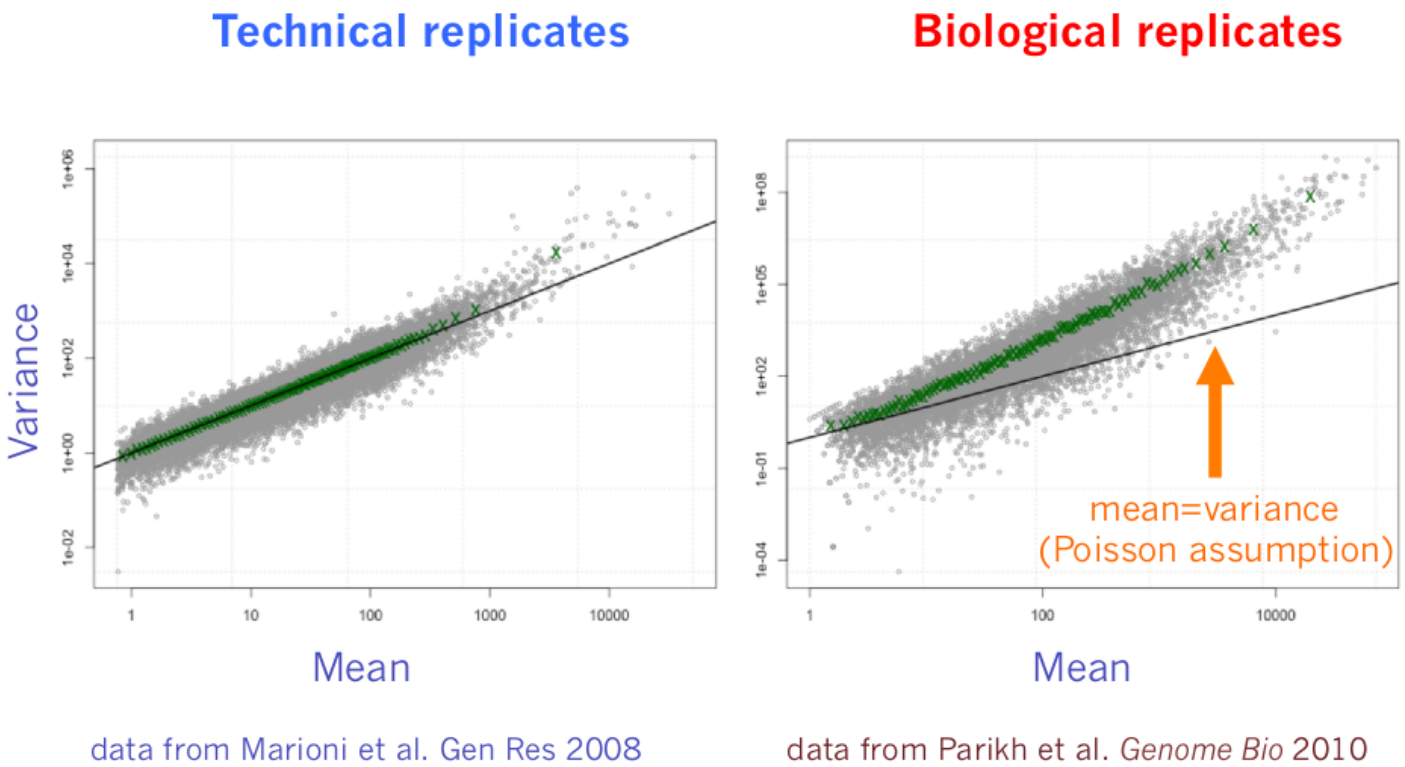
Estimation of the variance of gene g expression

Not enough measurements (replicates)

↓

Modeling of RNA-seq data

How to model RNA-seq data ?



- Overdispersion between biological replicates
- Negative binomiale distribution is often assumed: $Y \sim NB(\mu, \phi)$

$$E(Y) = \mu$$

$$V(Y) = \mu(1 + \phi\mu)$$

Three statistical frameworks

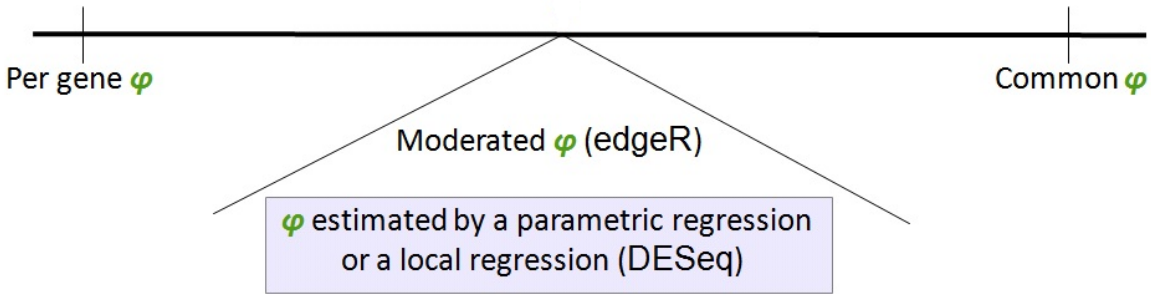
- A negative binomiale distribution (2008)
  - Expression = library size $\times \lambda_{condition}$
- A NB generalized linear model (2012)
  - allows us to decompose the expression
  - each condition is described by several factors

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

  - 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)
- Which method to estimate the variance of the gene expression (several methods)



Neutral comparison study

We want to answer these questions with a large evaluation study

- 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

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

H_0 genes

Buds vs Leaves

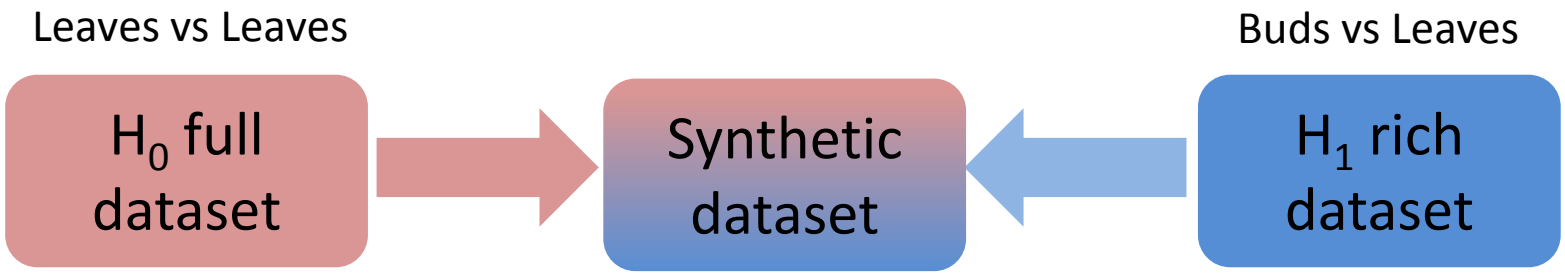
H_1 rich dataset

Unknown status

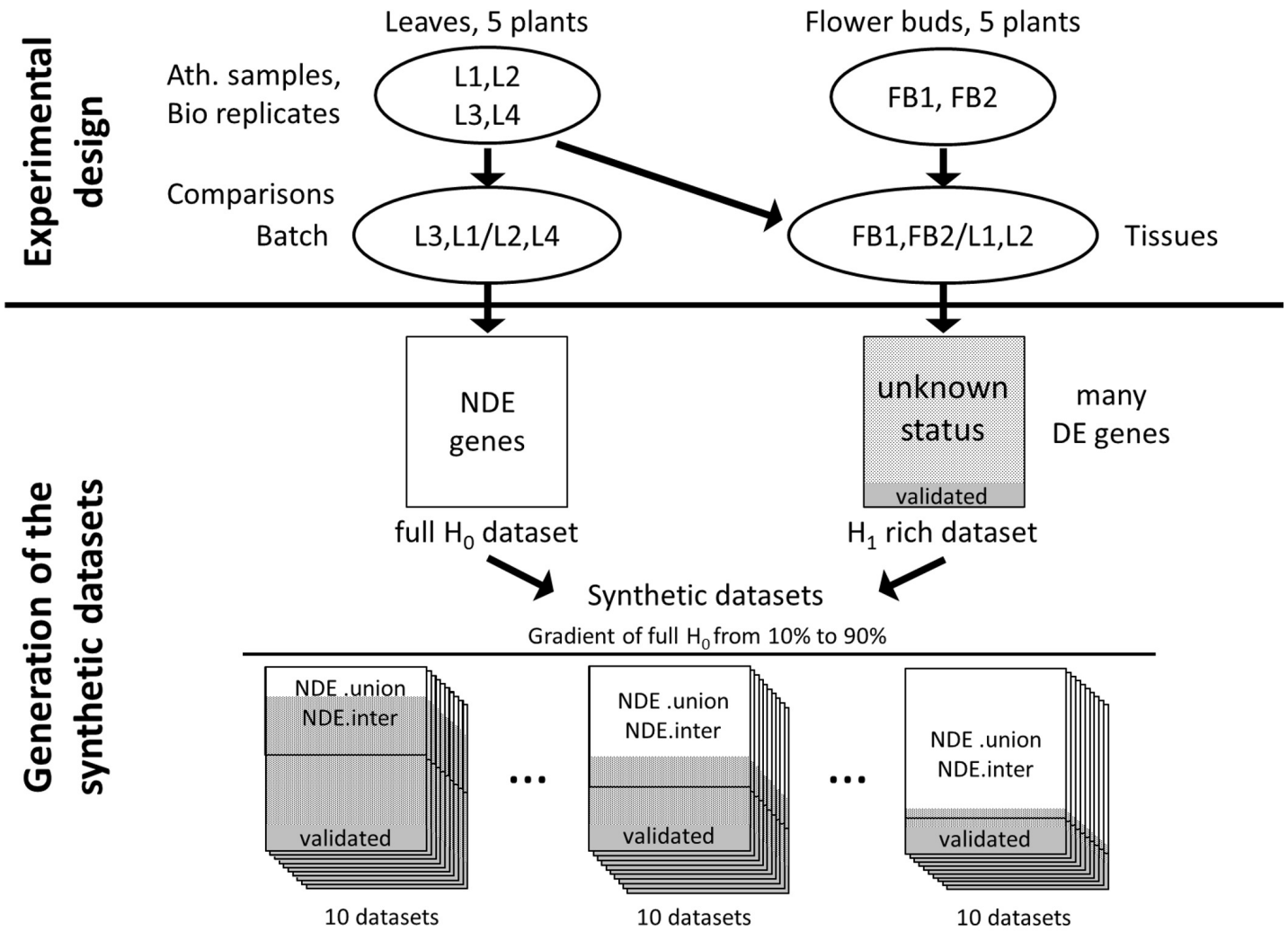
Validated

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

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

- **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

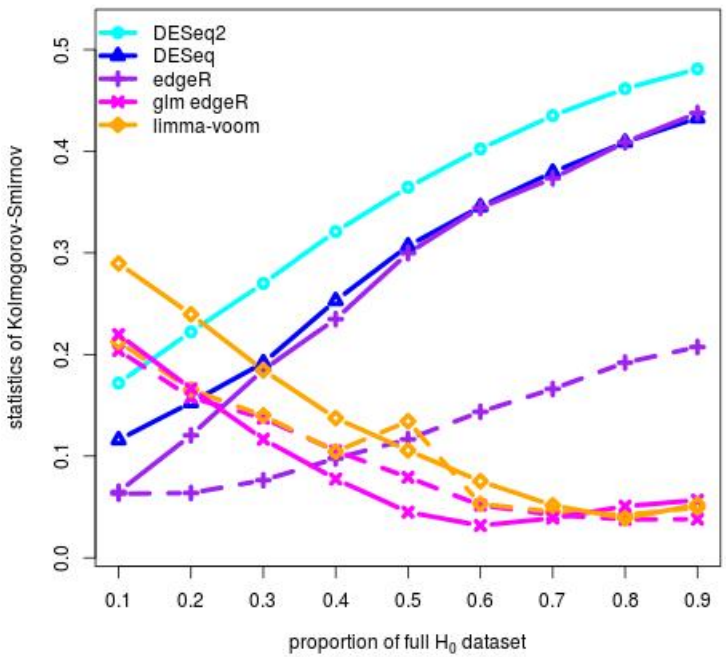
$$\text{Expression} = Cst + \alpha_{tissue} + \beta_{biological\ replicate}$$

* All methods except DESeq are also applied on filtered data
* In each method, nominal value of FDR is 5 %

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

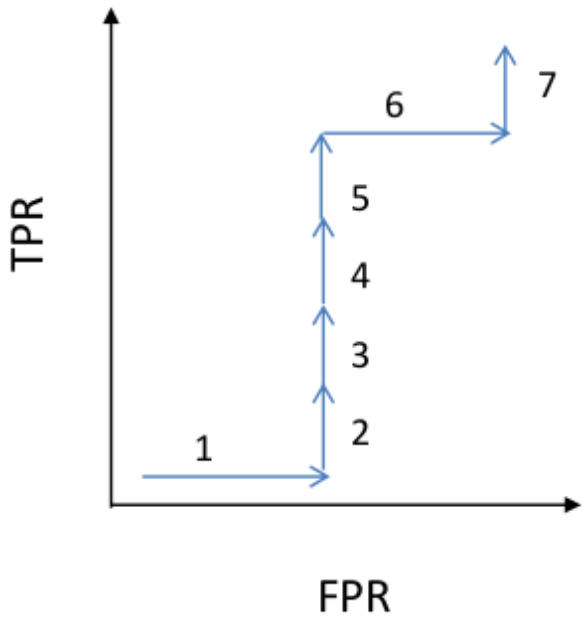
Definition of a ROC curve

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

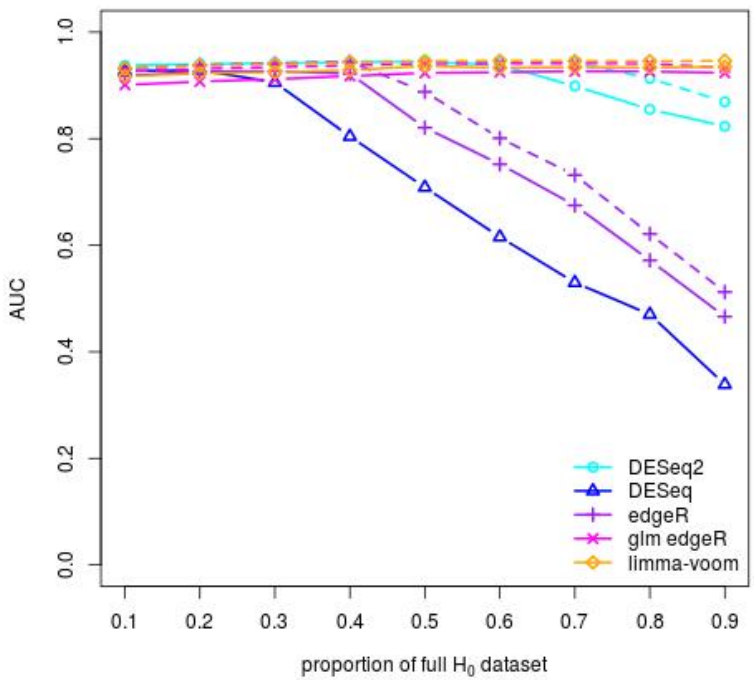
rank	gene	p-value	truth	TPR	FPR
1	G1	p1	NDE	0/5	1/2
2	G2	p2 (>p1)	DE	1/5	1/2
3	G3	p3(>p2)	DE	2/5	1/2
4	G4	p4(>p3)	DE	3/5	1/2
5	G5	p5(>p4)	DE	4/5	1/2
6	G6	p6(>p5)	NDE	4/5	2/2
7	G7	p7(>p6)	DE	5/5	2/2



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

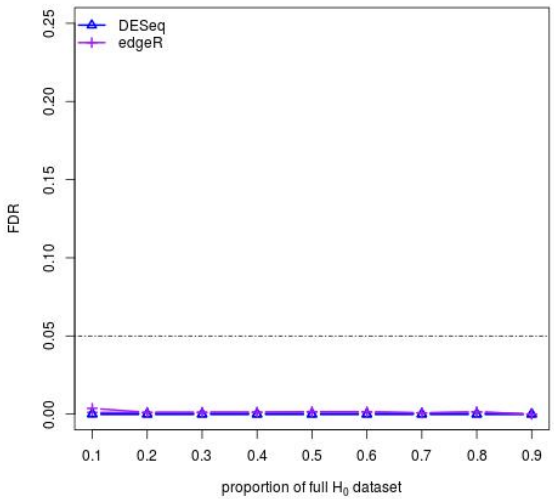


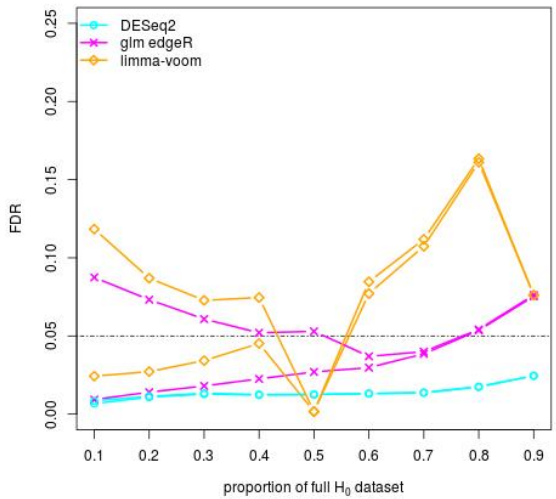
- For linear model or glm, the AUC is high and independent of the proportion of full H0 datasets
- For NB-based method, the AUC steadily decrease with the increase of the proportion of full H0 dataset when it is larger than 0.3-0.4

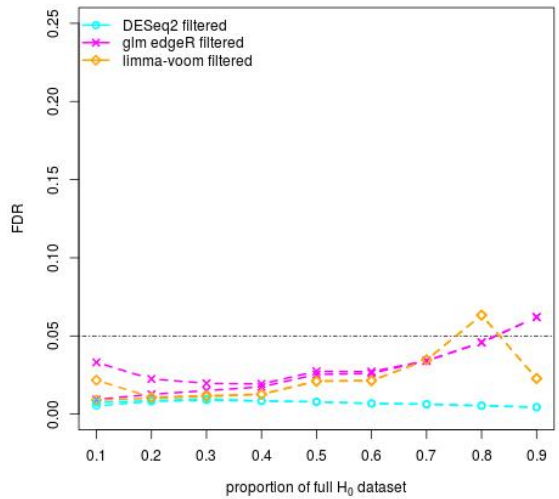
FDR estimation

Method

FDR estimation by the proportion of truly NDE among the declared DE
Comparison with the expected value 0.05





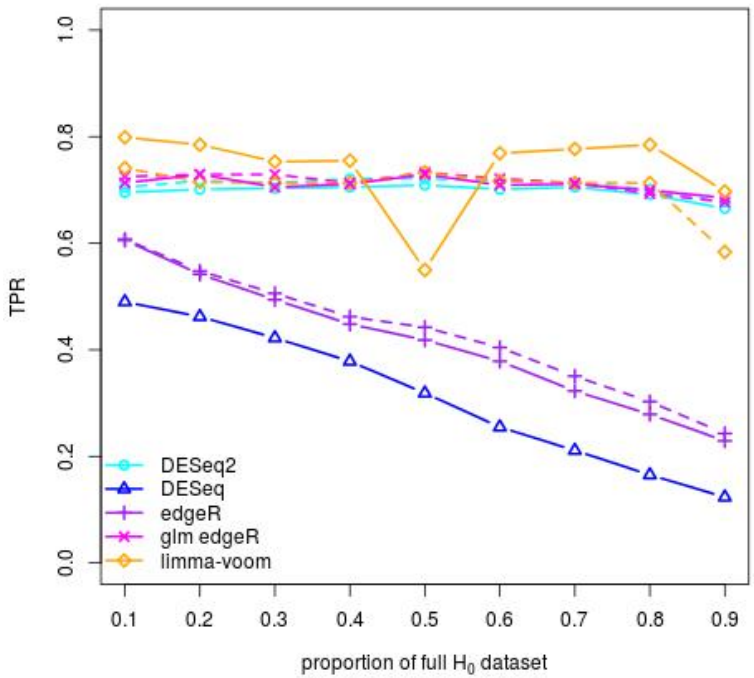


- For NB-based method, both bounds are close to 0
- For DESeq2, the FDR is always lower than 5%
- For glm edgeR, the interval generally contains 5%
- For limma-voom, the FDR control is more variable but the filtering step stabilizes its behavior

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.

Conclusions

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

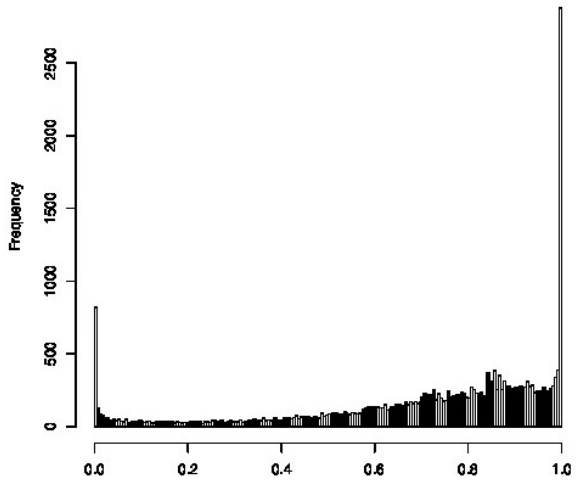
Synthetic data are a relevant framework

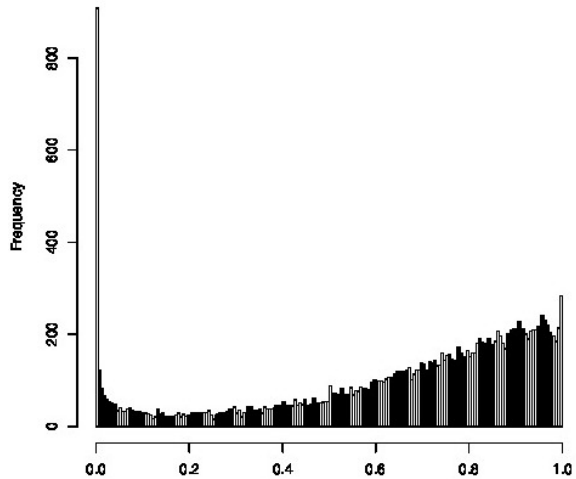
- Forget edgeR and DESeq
- use glm edgeR, DESeq2 or limma-voom
- include biological replicate as a factor
- filtering allows methods to control FDR

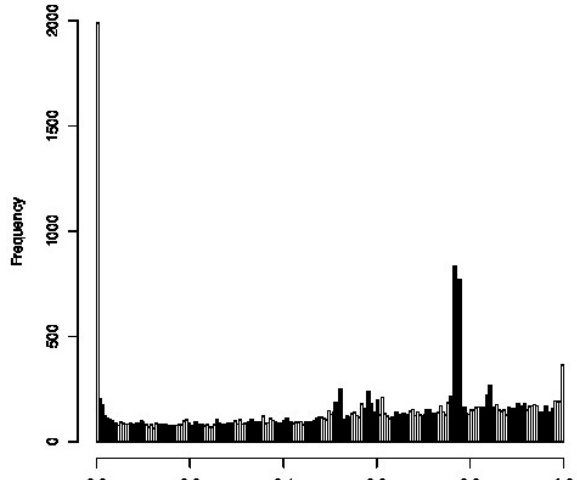
Definition of an indicator of quality

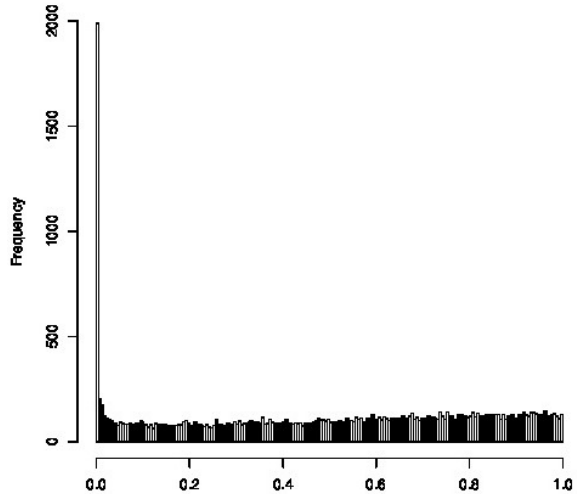
An histogram with a peak at the right side = analysis of bad quality

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









Practical case

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Biological questions

Practical case

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Biological questions

Genotype effect what genes are differentially expressed between the two genotypes ?

Practical case

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Biological questions

Genotype effect what genes are differentially expressed between the two genotypes ?

Time effect what genes are differentially expressed between two consecutive times ?

Practical case

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Biological questions

Genotype effect what genes are differentially expressed between the two genotypes ?

Time effect what genes are differentially expressed between two consecutive times ?

Genotype x Time effect what genes are impacted in their transcription by an interaction between the genotype and the time ?

Statistical translation of the context

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Factor identification

Statistical translation of the context

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$

Statistical translation of the context

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Factor identification

- Genotype $\in \{\text{wild-type, mutant}\}$
- Time $\in \{1, 2, 3\}$

Statistical translation of the context

Context

An experiment is performed to evaluate whether a given mutant behaves as the wild-type plant. Transcriptome of both plants is measured on a three-point time series at four different dates.

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$
- Time $\in \{1, 2, 3\}$
- Replicate (date) $\in \{1, 2, 3, 4\}$

2 biological factors and one technical factor

Statistical modelling

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$
- Time $\in \{1, 2, 3\}$
- Replicate (date) $\in \{1, 2, 3, 4\}$

Statistical modelling

Statistical modelling

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$
- Time $\in \{1, 2, 3\}$
- Replicate (date) $\in \{1, 2, 3, 4\}$

Statistical modelling

Let $Y_{gtr}^{\tilde{g}}$ be the expression of gene $\tilde{g}$ for the genotype g at time t in the r^{th} experiment and $\mu_{gtr}^{\tilde{g}} = \log\{E(Y_{gtr}^{\tilde{g}})\}$.

Statistical modelling

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$
- Time $\in \{1, 2, 3\}$
- Replicate (date) $\in \{1, 2, 3, 4\}$

Statistical modelling

Let $Y_{gtr}^{\tilde{g}}$ be the expression of gene $\tilde{g}$ for the genotype g at time t in the r^{th} experiment and $\mu_{gtr}^{\tilde{g}} = \log\{E(Y_{gtr}^{\tilde{g}})\}$.
Assume

$$\mu_{gtr}^{\tilde{g}} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

Statistical modelling

Factor identification

- Genotype $\in \{$wild-type, mutant$\}$
- Time $\in \{1, 2, 3\}$
- Replicate (date) $\in \{1, 2, 3, 4\}$

Statistical modelling

Let $Y_{gtr}^{\tilde{g}}$ be the expression of gene $\tilde{g}$ for the genotype g at time t in the r^{th} experiment and $\mu_{gtr}^{\tilde{g}} = \log\{E(Y_{gtr}^{\tilde{g}})\}$.
Assume

$$\mu_{gtr}^{\tilde{g}} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

$\rightsquigarrow$ interactions between technical and biological factors are not considered

First remarks

$$\mu_{gtr}^{\tilde{g}} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

Statistical modelling

The logarithm of the mean expression of gene $\tilde{g}$ is modeled as

$$\mu_{gtr}^{\tilde{g}} = \mu^{\tilde{g}} + G_g + T_t + R_r + GT_{gt}$$

First remarks

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

Statistical modelling

The logarithm of the mean expression of gene $\tilde{g}$ is modeled as

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + G_g + T_t + R_r + GT_{gt}$$

Some calculations

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

First remarks

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

Statistical modelling

The logarithm of the mean expression of gene $\tilde{g}$ is modeled as

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + G_g + T_t + R_r + GT_{gt}$$

Some calculations

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

First remarks

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + Genotype_g + Time_t + Replicate_r + Genotype_g : Time_t$$

Statistical modelling

The logarithm of the mean expression of gene $\tilde{g}$ is modeled as

$$\mu^{\tilde{g}}_{gtr} = \mu^{\tilde{g}} + G_g + T_t + R_r + GT_{gt}$$

Some calculations

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

These quantities are independent of the factor Replicate

Statistical translation of the biological questions

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype effect

- What genes are differentially expressed between the two genotypes ?

Statistical translation of the biological questions

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype effect

- What genes are differentially expressed between the two genotypes ?
- what genes show a time-averaged difference in their expression between the two genotypes independently from the experiments?

Statistical translation of the biological questions

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype effect

- What genes are differentially expressed between the two genotypes ?
- what genes show a time-averaged difference in their expression between the two genotypes independently from the experiments?

$$\Delta^{\tilde{g}}_{genotype} = \frac{1}{3}\sum_t (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr})$$

$$\Delta^{\tilde{g}}_{genotype} = (G_m - G_w) + \frac{1}{3}\sum_t (GT_{mt} - GT_{wt})$$

Statistical translation of the biological questions

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype effect

- What genes are differentially expressed between the two genotypes ?
- what genes show a time-averaged difference in their expression between the two genotypes independently from the experiments?

$$\Delta^{\tilde{g}}_{genotype} = \frac{1}{3}\sum_t (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr})$$

$$\Delta^{\tilde{g}}_{genotype} = (G_m - G_w) + \frac{1}{3}\sum_t (GT_{mt} - GT_{wt})$$

$$H_0 = \{\Delta^{\tilde{g}}_{genotype} = 0\} \text{ vs } H_1 = \{\Delta^{\tilde{g}}_{genotype} \neq 0\}$$

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

Time effect

- what genes are differentially expressed between two consecutive times ?

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

Time effect

- what genes are differentially expressed between two consecutive times ?
- For a given genotype, whatever the experiment, what genes are differentially expressed between two consecutive times ?

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

Time effect

- what genes are differentially expressed between two consecutive times ?
- For a given genotype, whatever the experiment, what genes are differentially expressed between two consecutive times ?

$$\Delta^{\tilde{g}}_{g,t,t'} = (\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r})$$

$$\Delta^{\tilde{g}}_{g,t,t'} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

$$\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

Time effect

- what genes are differentially expressed between two consecutive times ?
- For a given genotype, whatever the experiment, what genes are differentially expressed between two consecutive times ?

$$\Delta^{\tilde{g}}_{g,t,t'} = (\mu^{\tilde{g}}_{gtr} - \mu^{\tilde{g}}_{gt'r})$$

$$\Delta^{\tilde{g}}_{g,t,t'} = (T_t - T_{t'}) + (GT_{gt} - GT_{gt'})$$

$$H_0 = \{\Delta^{\tilde{g}}_{g,t,t'} = 0\} \text{ vs } H_1 = \{\Delta^{\tilde{g}}_{g,t,t'} \neq 0\}$$

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?
- what genes show a genotype difference between two time points ?

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?
- what genes show a genotype difference between two time points ?

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr}) - (\mu^{\tilde{g}}_{mt'r} - \mu^{\tilde{g}}_{wt'r})$$

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?
- what genes show a genotype difference between two time points ?

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr}) - (\mu^{\tilde{g}}_{mt'r} - \mu^{\tilde{g}}_{wt'r})$$

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype x Time interaction

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?
- what genes show a genotype difference between two time points ?

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr}) - (\mu^{\tilde{g}}_{mt'r} - \mu^{\tilde{g}}_{wt'r})$$

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype x Time interaction

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (GT_{mt} - GT_{wt}) - (GT_{mt'} - GT_{wt'})$$

Genotype x Time effect

- what genes are impacted in their transcription by an interaction between the genotype and the time ?
- what genes show a genotype difference between two time points ?

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr}) - (\mu^{\tilde{g}}_{mt'r} - \mu^{\tilde{g}}_{wt'r})$$

$$\mu^{\tilde{g}}_{mtr} - \mu^{\tilde{g}}_{wtr} = (G_m - G_w) + (GT_{mt} - GT_{wt})$$

Genotype x Time interaction

$$\Delta^{\tilde{g}}_{m,w,t,t'} = (GT_{mt} - GT_{wt}) - (GT_{mt'} - GT_{wt'})$$

$$H_0 = \{\Delta^{\tilde{g}}_{m,w,t,t'} = 0\} \text{ vs } H_1 = \{\Delta^{\tilde{g}}_{m,w,t,t'} \neq 0\}$$

Remark : when the two factors have only two modalities, it becomes to test the interaction coefficient of the model

Acknowledgements

- Marie-Laure Martin-Magniette (IPS2, Genomic networks, Paris-Saclay)
- Etienne Delannoy (IPS2, Genomic networks, Paris-Saclay)
- Christophe Ambroise (LaMME, Universit d'Evry)
- The transcriptomic platform of IPS2 (data generation and bioinformatics analysis)
- The ANR project MixStatSeq coordinated by C. Maugis (INSA /IMT, Toulouse)