Key ingredients for RNA-seq differential analysis

Neutral comparison study

Etienne Delannoy & Marie-Laure Martin-Magniette

Plant Science Institut of Paris-Saclay (IPS2)

Applied Mathematics and Informatics Unit at AgroParisTech

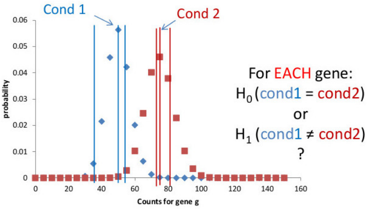
Objective of the differential analysis

- 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

$$H_0 = \{\text{There is no difference}\}$$

versus

$$H_1 = \{\text{There is a difference}\}$$



For EACH gene:
H_0 (cond1 = cond2)
or
H_1 (cond1 ≠ cond2)
?

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
- Calculate the p-value
- Decide if the null hypothesis is rejected or not with respect to the value of the test statistic

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

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
- 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

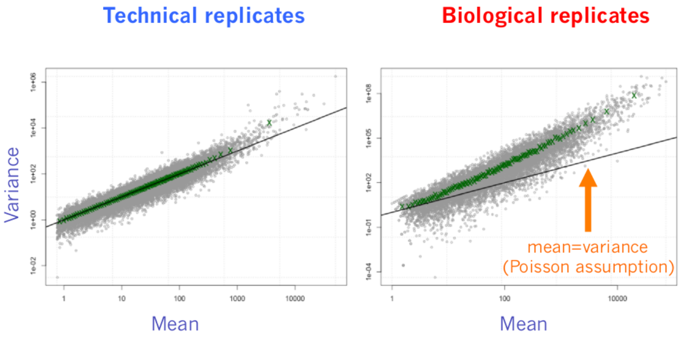
Adjustment of the raw p-values

- $FWER = P(FP > 0)$ (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 ?



Technical replicates

Biological replicates

mean=variance (Poisson assumption)

data from Marioni et al. Gen Res 2008

data from Parikh et al. *Genome Bio* 2010

- 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

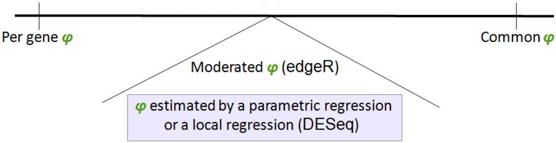
- A negative binomiale distribution (2008)
  - Expression = library size $\times \lambda_{condition}$
- A NB generalized linear model (2012)
  - allows us to decompose the expression
  - each condition is described by several factors

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

  - 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)



Per gene φ

Common φ

Moderated φ (edgeR)

φ estimated by a parametric regression or a local regression (DESeq)

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)

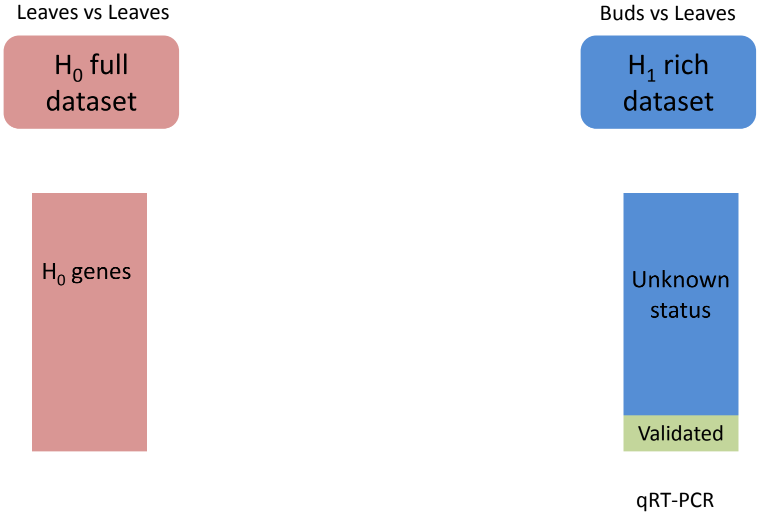
→ proportion of truly DE declared DE

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

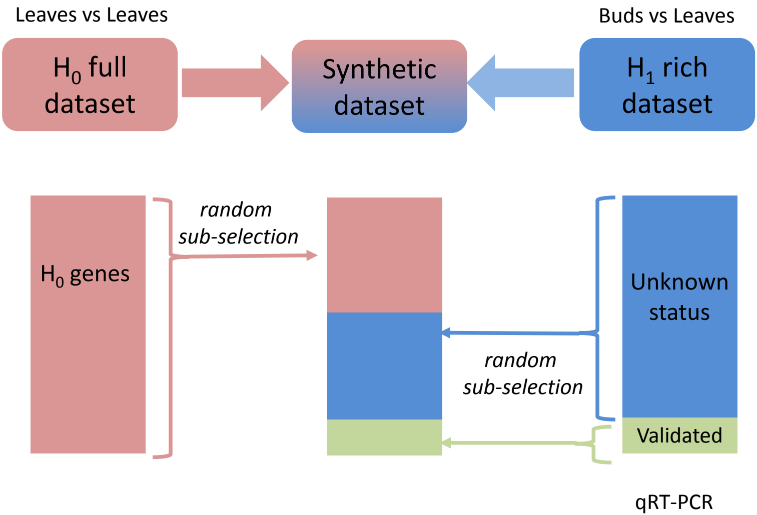
H_1 rich dataset

Unknown status

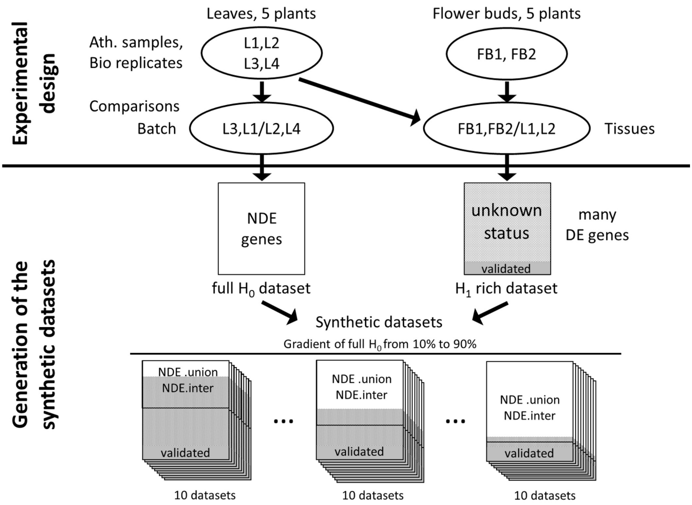
Validated

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

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

- **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

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

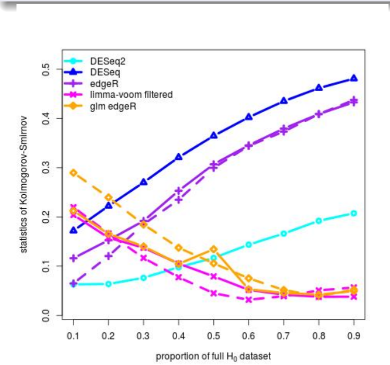
* All methods except DESeq are also applied on filtered data

* 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

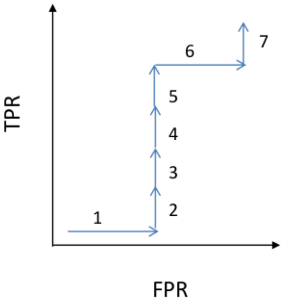
Definition of a ROC curve

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

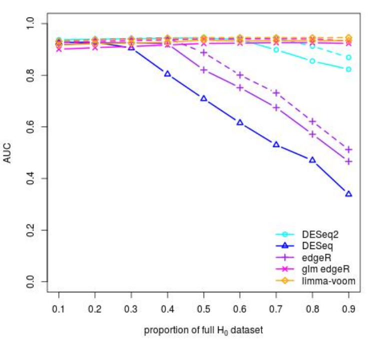
rank	gene	p-value	truth	TPR	FPR
1	G1	p1	NDE	0/5	1/2
2	G2	p2 (>p1)	DE	1/5	1/2
3	G3	p3(>p2)	DE	2/5	1/2
4	G4	p4(>p3)	DE	3/5	1/2
5	G5	p5(>p4)	DE	4/5	1/2
6	G6	p6(>p5)	NDE	4/5	2/2
7	G7	p7(>p6)	DE	5/5	2/2



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

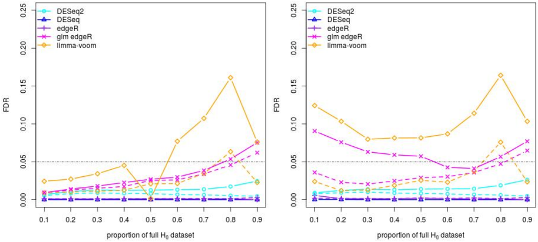


- For linear model or glm, the AUC is high and independent of the proportion of full H0 datasets
- For NB-based method, the AUC steadily decrease with the increase of the proportion of full H0 dataset when it is larger than 0.3-0.4

FDR estimation

Method

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

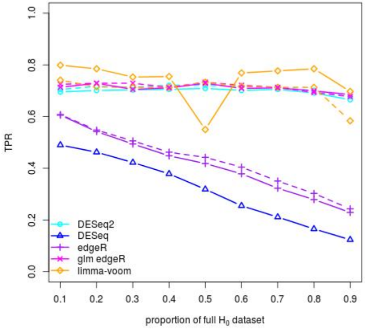


- For NB-based method, both bounds are close to 0
- For DESeq2, the FDR is always lower than 5%
- For glm edgeR, the interval generally contains 5%
- For limma-voom, the FDR control is more variable but the filtering step stabilizes its behavior

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.

Conclusions

modeling $\geq$ filtering $\geq$ dispersion

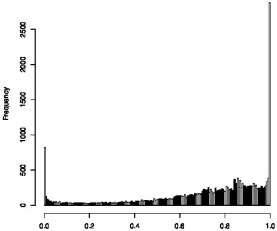
Synthetic data are a relevant framework

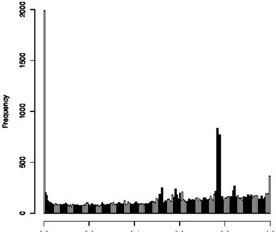
- Forget edgeR and DESeq
- use glm edgeR, DESeq2 or limma-voom
- include biological replicate as a factor
- filtering allows methods to control FDR

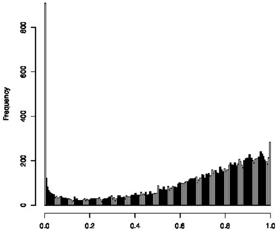
Definition of an indicator of quality

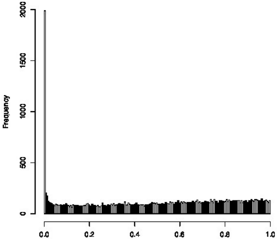
An histogram with a peak at the right side = analysis of bad quality

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









Acknowledgements

POP

Plateforme transcriptOmique IPS2

- Guillem Rigaill (IPS2, Genomic networks, Paris-Saclay)
- The transcriptomic platform of IPS2 (data generation and bioinformtics analysis)
- The ANR project MixStatSeq coordinated by C. Maugis (IMT, Toulouse) and involving A. Rau (GABI, INRA) and G. Celeux (INRIA, Saclay)

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

ANR