Statistical analysis of RNA-seq data

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

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

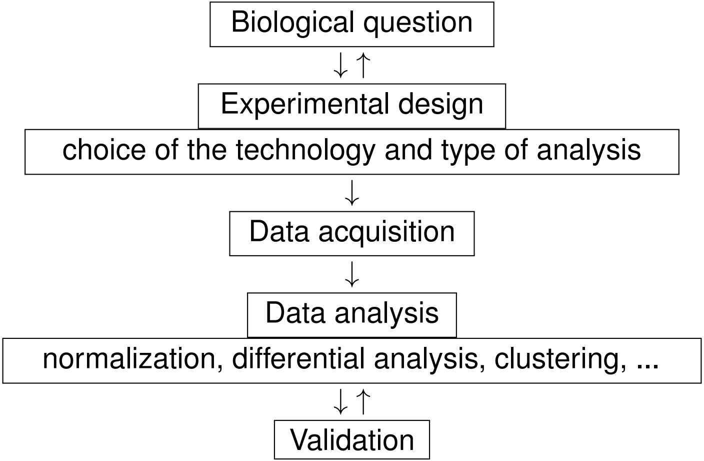
2- UMR AgroParisTech/INRA Mathematique et Informatique Appliquees

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



Biological question

↓↑

Experimental design

choice of the technology and type of analysis

↓

Data acquisition

↓

Data analysis

normalization, differential analysis, clustering, ...

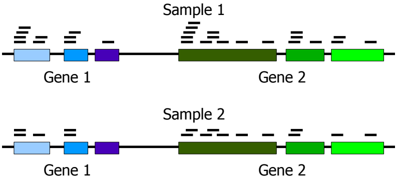
↓↑

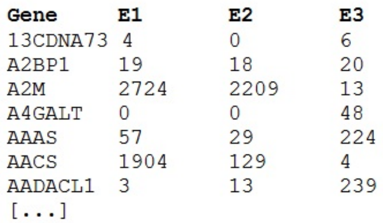
Validation

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 × 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

1. Introduction

2. Normalization
    - Different types of normalization
    - Comparaison of 7 normalization methods
    - Conclusions on normalization

3. Differential analysis

4. Clustering of transcriptomic data

5. Experimental design

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)

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

- Upper Quartile normalization (Bullard et al. 2010)

$$\hat{s}_j = \frac{Q3_j}{\frac{1}{n}\sum_{\ell=1} Q3_\ell}, \text{ where } Q3_j = Y_{(\frac{3}{4}[G+1])j}$$

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

- 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)

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

- 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

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

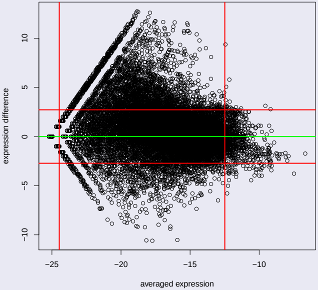
- normalize them such that their product equals 1

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

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 $w^r_{gj} = (\frac{N_j - Y_{gj}}{N_j Y_{gj}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}})$
- Compute $TMM^r_j = \frac{\sum_{g \in G^\star} w^r_{gj} M^r_{gj}}{\sum_{g \in G^\star} w^r_{gj}}$
- Define
$$\tilde{s}_j = 2^{TMM^r_j}$$
- 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

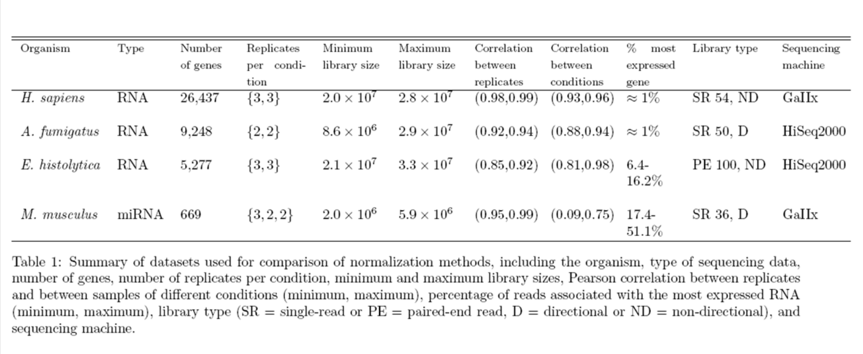
Questions

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

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis
French StatOmique Consortium (2012) doi : 10.1093./bib/bbs046

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)	≈ 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)	≈ 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 procedures

Distribution and properties of normalized datasets

Boxplots, variability between biological replicates

Comparison of DE genes

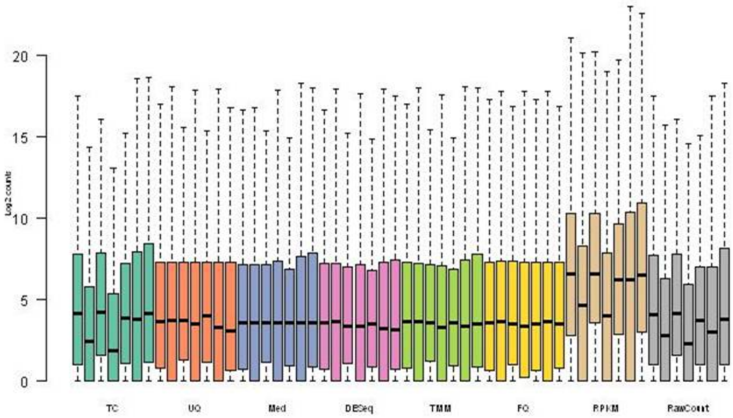
- Differential analysis by exact test: DESeq v1.6.1, default parameters
- Number of common DE genes, similarity between list of genes (dendrogram - binary distance and Ward linkage)

Power and control of the Type-I error rate

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

Normalized data distribution

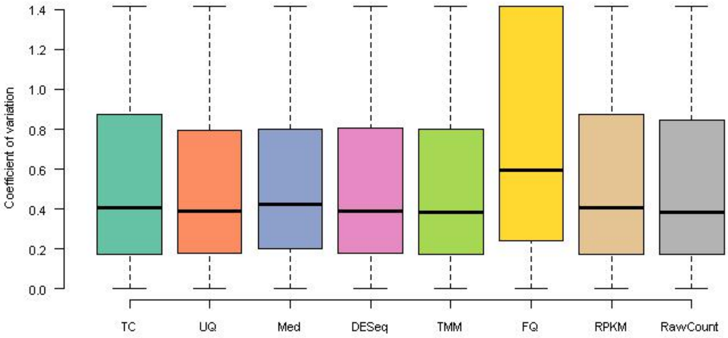
When large diff. in lib. size, TC and RPKM do not improve over the raw counts.



Example: *Mus musculus* dataset

Within-condition variability

Example: *Mus musculus, condition D* dataset



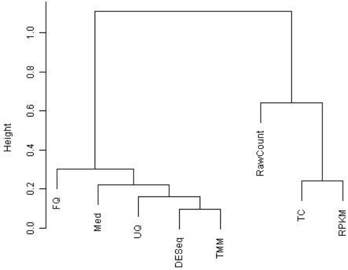
Lists of differentially expressed (DE) genes

For each dataset

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

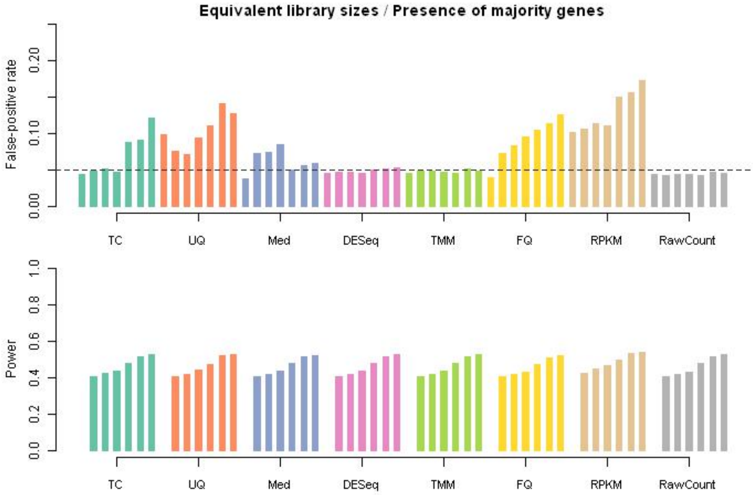
Consensus matrice

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



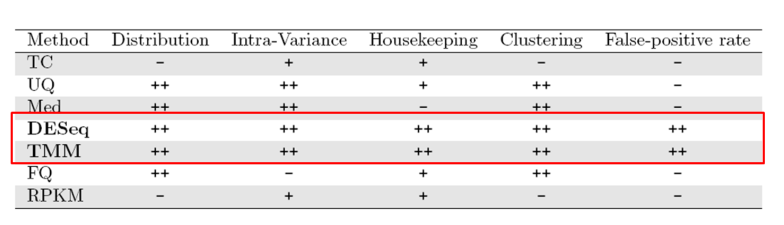
So the Winner is ... ?

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

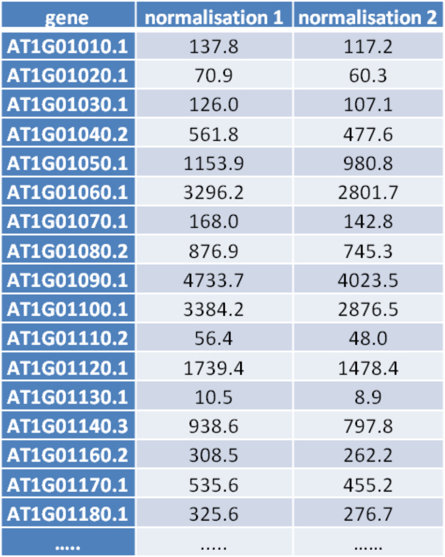
- RNA-seq data are affected by technical biaises (total number of mapped reads per lane, gene length, composition bias)

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.

Normalisation step 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
.....	.....	......

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

TP session

- Directory Script, look at NormalizationMethods.R
- Read the code
- Run the code and observe the normalized count distribution

Table of contents

1 Introduction

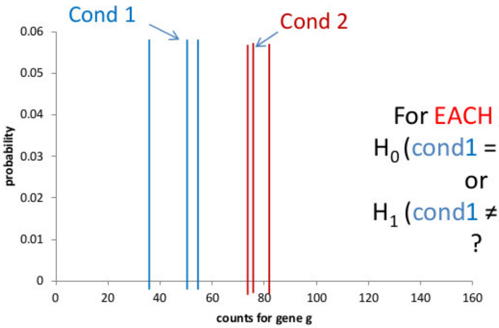
2 Normalization

3 **Differential analysis**
- Introduction
- Test for RNA-seq data
- Method evaluation

4 Clustering of transcriptomic data

5 Experimental design

Statistical hypothesis test



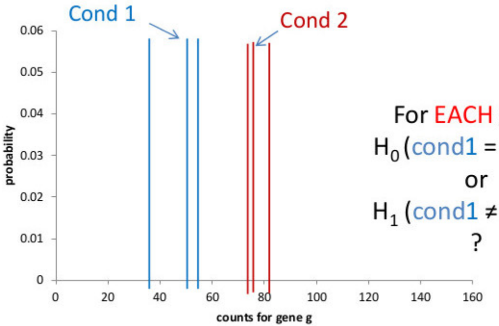
For EACH gene:
H_0 (cond1 = cond2)
or
H_1 (cond1 ≠ cond2)
?

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)
?

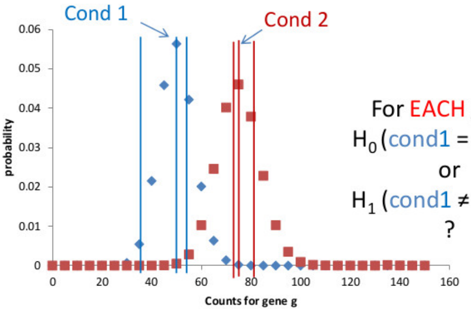
Technical and biological variabilities

Cond1 and Cond2 measurements are variable

Not enough replicates

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

Statistical hypothesis test



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

Principle of a hypothesis test

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

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.

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)

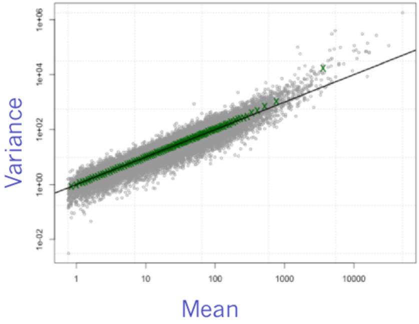
Calculating adjusted p-values

- 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 $\tilde{p}_{(i)} = a_i p_{(i)}$
  4. Set $\tilde{p}_{(i)}$ = min($\tilde{p}_{(i)}$, 1) for all i

Test for RNA-seq data

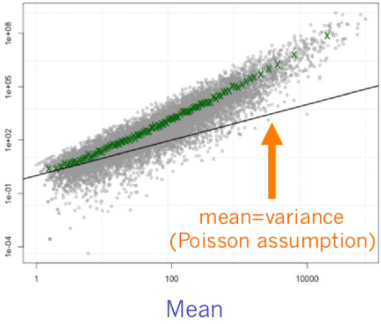
Distribution for count data

Technical replicates



data from Marioni et al. Gen Res 2008

Biological replicates



mean=variance
(Poisson assumption)

data from Parikh et al. *Genome Bio* 2010

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

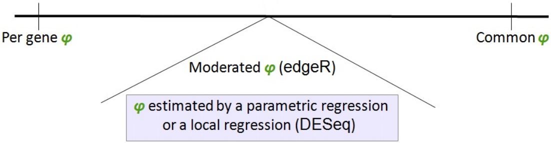
$$H_0 = \{\lambda_{g1} = \lambda_{g2}\} \text{ vs } H_1 = \{\lambda_{g1} \neq \lambda_{g2}\}$$

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

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!


Per gene φ
Common φ
Moderated φ (edgeR)
φ estimated by a parametric regression or a local regression (DESeq)

- 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*

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

GLM framework for RNAseq data

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

We assume

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

with

$$E(log(Y_{gjv})) = \log(s_{jv}) + log(\lambda_{gjv})$$

where

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

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

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

Linear model framework for RNAseq data

- Let Y_{gjv} be the counts of reads for gene g in the sample described by the uplet (j, v)
- Data are transformed so that a linear model can be considered
- Linear Model allows to decompose a function of the mean of the observations

We assume

$$\tilde{Y}_{gjv} \sim \mathcal{N}(\mu_{gjv}, \sigma_g^2)$$

with

$$E(\tilde{Y}_{gjv}) = Intercept + s_{jv} + \alpha_{gj} + \beta_{gv} + \gamma_{gjv}$$

Example

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

Conclusions on the differential analysis

- The discrete nature as well as the extreme precision of RNA-seq measures are a challenge for statistical analyses
- Differential analysis is based on several assumptions: read distribution and dispersion modeling
- Some methods proposed to filter low counts
- Exact test is not tailored for experiments with more than one factor
- GLM is a more flexible framework
- Another alternative is to transform the data and to use a linear model (limma-voom)
- In the latter two cases, it is important to determine which factors are important to include in the expectation decomposition ?

In practice

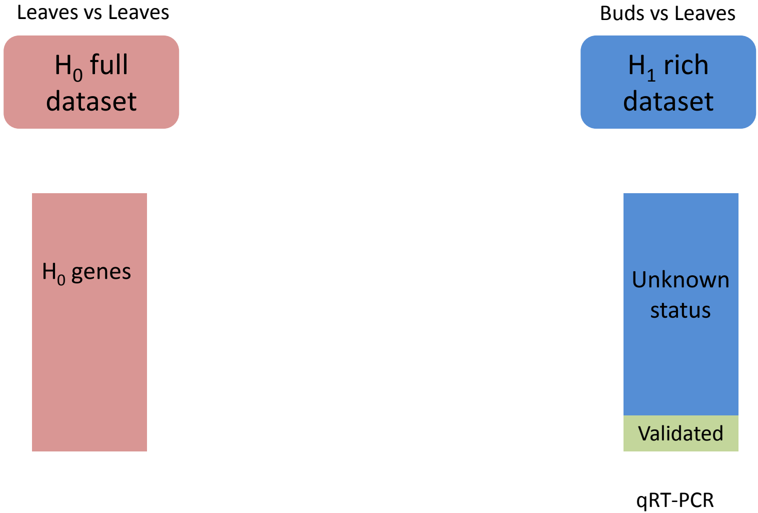
To perform a differential analysis, we have to make a decision about

- the filtering (yes or no)
- the modeling (NB, GLM or LM and which factors are important to consider ?)
- the dispersion estimation (several methods)

How to evaluate methods for the differential analysis of gene expression?

- 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?
- Another solution: synthetic data

Synthetic data simulations



Leaves vs Leaves

H_0 full dataset

H_0 genes

Buds vs Leaves

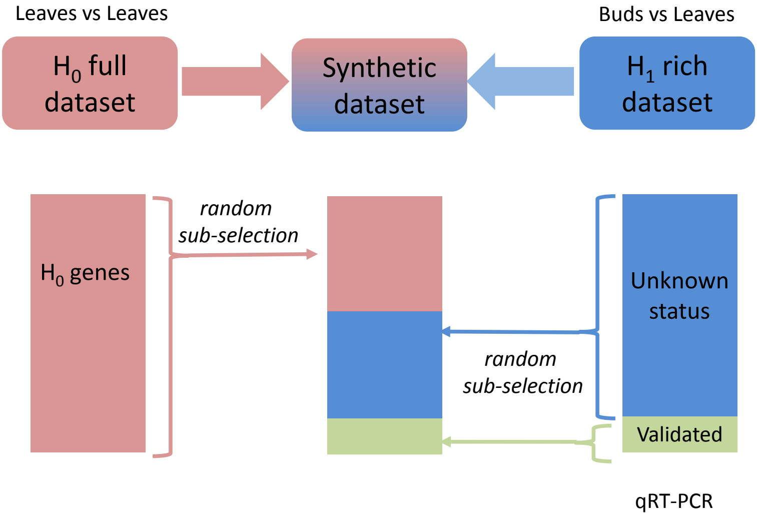
H_1 rich dataset

Unknown status

Validated

qRT-PCR

Synthetic data simulations



Fifteen evaluated methods

Information on the experimental design

- Two conditions: buds and leaves
- Two biological replicates are available for each tissu
- It means that gene expression is measurement for each tissu from plants grown in the same growth chamber but at two different dates

Question about the differential analysis

- the filtering (yes or no)
- Count modeling
  - NB model
  - GLM with or without batch effect
  - Data transformation and linear model with or without batch effect
- dispersion estimation: edgeR, DESeq, DESeq2
- limma for linear model

Definition of a ROC curve

Drawing a ROC curve:

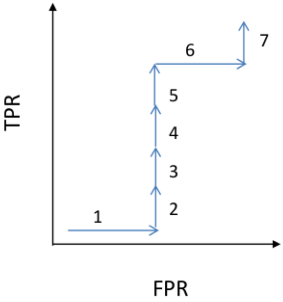
1- sort genes by increasing raw p-value

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



The sets of truly DE genes and truly NDE genes

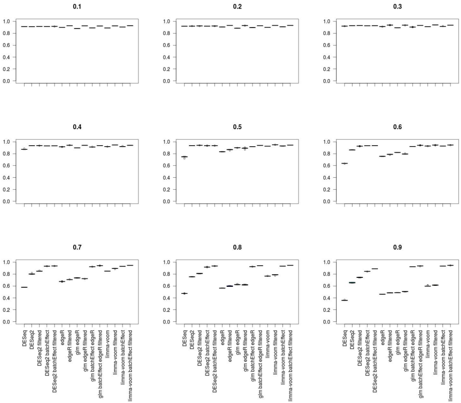
the set of truly DE genes

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

the set of truly NDE genes

- The proper identification is not straightforward
- NDE.union: genes declared at least once not differentially expressed by DESeq2, glm edgeR and limma-voom taking into account a batch effect.
- NDE.union may include some genes that are not truly NDE
- NDE.inter: genes always declared not differentially expressed by the three methods.
- NDE.inter may exclude some truly NDE genes.
- Consequently both should be considered for AUC and FDR evaluations.

Discrimination of DE and NDE genes

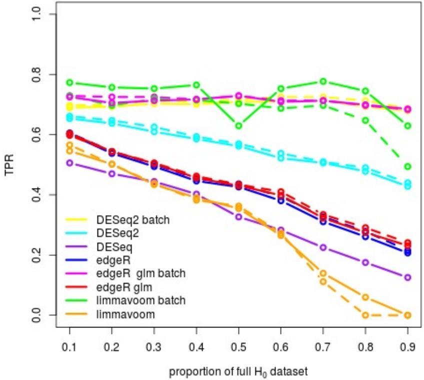


- Data filtering has a slight effect
- For a proportion of full H0 dataset above 0.6 (implying a smaller proportion of DE genes), linear modeling after data transformation or glm modeling improves the AUC
- This increase is even greater when a batch effect is considered
- The variance-mean relationship modeling seems to have a limited impact

Similar results with NDE.union and NDE.inter

Evaluation of the TPR

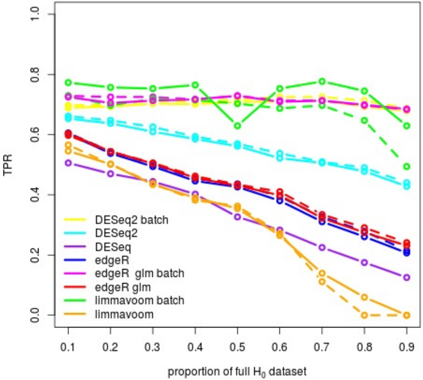
Tests were performed at FDR level 5%



- Two groups : methods taking into account a batch effect and all the other methods.

Evaluation of the TPR

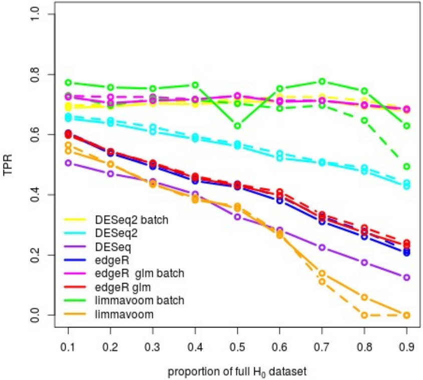
Tests were performed at FDR level 5%



- Two groups : methods taking into account a batch effect and all the other methods.
- For the first group, methods show a high TPR
- For the second group, the TPR is a function of the full H0 dataset proportion.

Evaluation of the TPR

Tests were performed at FDR level 5%



- Two groups : methods taking into account a batch effect and all the other methods.
- For the first group, methods show a high TPR
- For the second group, 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.

Evaluation of the FDR

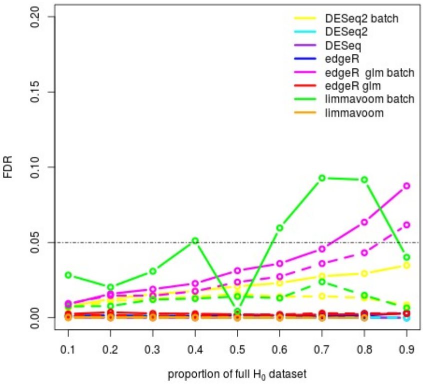
- Requires to identify the set of genes non-differentially expressed (NDE)
- Not easy
- Two sets are defined :

  (i) intersection of the genes declared NDE by edgeR, DESeq2 and limma-voom that take a batch effect into account (lower bound)

  (ii) union of the genes declared NDE by edgeR, DESeq2 and limma-voom that take a batch effect into account (upper bound)

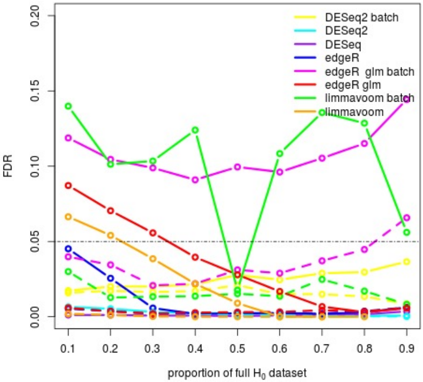
Evaluation of the lower bound

Tests were performed at FDR level 5%



- For the methods not taking a bach effect into account, the lower bound is close to 0
- For the methods taking a bach effect into account, the lower bound is higher and increases with the proportion of full H0 dataset
- The filtering procedure seems to stabilize the lower bound across the proportion of full H0 dataset.

Evaluation of the upper bound



- A strong effect of the data filtering
- The upper bound of all filtered methods was lower than 0.05 except for glm edgeR with a batch effect at 90% of full H0 dataset
- For the methods not taking a batch effect into account, the upper bound was very close to 0, suggesting that they were very conservative

Evaluation of the p-values

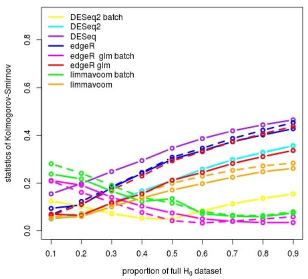
Recall

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

Evaluation of the p-values

Recall

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



- 73% of tests are rejected after a Bonferroni adjustement
- KS statistic values are smaller for models taking a batch effect into account than for model without batch effect
- Data filtering has a slight effect

Conclusions

Synthetic data are a relevant framework

- Our analysis suggests that a well-modeling of the expectation of the counts is crucial
- To date, biological replicates of plants are not considered as a factor
- Impossible to observe this fact on simulated data

the raw p-value distribution = an indicator of quality

- It should be used to evaluate the fit between the counts and the model
- an histogram with a peak at the right side often indicates that the modeling is incorrect

modeling $\geq$ filtering $\geq$ dispersion

Table of contents

1 Introduction

2 Normalization

3 Differential analysis

4 Clustering of transcriptomic data
- Model-based clustering
- Mixture models for transcriptomic data

5 Experimental design

Gene co-expression for gene function prediction

- Transcriptome data: main source of 'omic information available for living organisms
  - Microarrays (∼1995 -)
  - High-throughput sequencing: RNA-seq (∼2008 -)
- Comparison of two conditions (hypothesis tests) → Differential expression analysis

Co-expression (clustering) analysis

- Study gene expression behavior across several conditions
- Co-expressed genes may be involved in similar biological process(es) (Eisen *et al.*, 1998))
- ⇒ Co-expression is a tool to study genes without known or predicted function (orphan genes)
- ⇒ It is also the first step to build a regulatory network

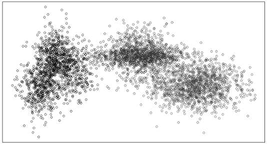
Model-based clustering

- Probabilistic clustering models : data assumed to come from distinct subpopulations, each modeled separately
- Rigourous framework for parameter estimation and model selection
- **Output**: each gene assigned a probability of cluster membership

What are the key ingredients to define a mixture model ?

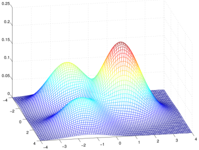
Key ingredients of a mixture model

what we observe

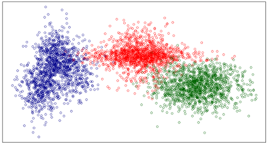


$Z = ?$

the model



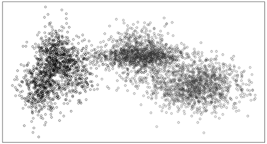
the expected results



Z: 1 = ●, 2 = ●, 3 = ●

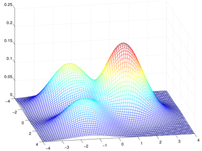
Key ingredients of a mixture model

what we observe

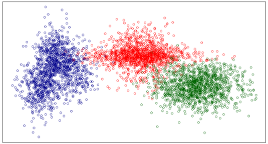


$Z = ?$

the model



the expected results



Z: 1 = ●, 2 = ●, 3 = ●

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote n observations described by Q variables
Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be the latent vector.

1) Distribution of Z: $\{Z_i\}$ are assumed to be independent and

$$P(Z_i = k) = \pi_k \text{ with } \sum_{k=1}^{K} \pi_k = 1 \qquad \rightarrow \mathbf{Z} \sim \mathcal{M}(n; \pi_1, \ldots, \pi_K)$$

K is the number of components of the mixture

2) Distribution of $(\mathbf{Y}_i | Z_i = k)$ is a parametric distribution $f(\bullet; \gamma_k)$

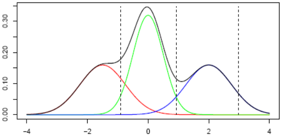
Questions around the mixtures

- Modeling: what distribution for each component ?
  $\rightsquigarrow$ it depends on observed data.
- Inference: how to estimate the parameters ?
  $\rightsquigarrow$ it is usually done with an EM-like algorithm (Dempster et al., 77)
- Model selection: how to choose the number of components ?
  - A collection of mixtures with **a varying number of components** is usually considered
  - A criterion is used to select the best model of the collection

Outputs of the model and data classification

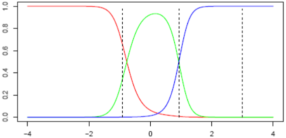
Distribution:

$$g(y_i) = \pi_1 f(y_i; \gamma_1) + \pi_2 f(y_i; \gamma_2) + \pi_3 f(y_i; \gamma_3)$$



Conditional probabilities:

$$\tau_{ik} = P(Z_i = k | y_i) = \frac{\pi_k f(y_i; \gamma_k)}{g(y_i)}$$



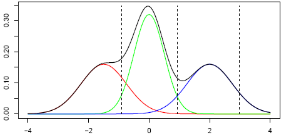
τ_{ik}	$i = 1$	$i = 2$	$i = 3$
$k = 1$	0.658	0.007	0.0
$k = 2$	0.342	0.478	0.0
$k = 3$	0.0	0.515	1.0

→ These probabilities enables the classification of the observations into the subpopulations

Outputs of the model and data classification

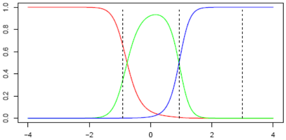
Distribution:

$$g(y_i) = \pi_1 f(y_i;\gamma_1) + \pi_2 f(y_i;\gamma_2) + \pi_3 f(y_i;\gamma_3)$$



Conditional probabilities:

$$\tau_{ik} = P(Z_i = k|y_i) = \frac{\pi_k f(y_i;\gamma_k)}{g(y_i)}$$



τ_{ik}	$i = 1$	$i = 2$	$i = 3$
$k = 1$	0.658	0.007	0.0
$k = 2$	0.342	0.478	0.0
$k = 3$	0.0	0.515	1.0

Maximum A Posteriori rule: Classification in the component for which the conditional probability is the highest.

Finite mixture models

Assume data **y** come from K distinct subpopulations, each modeled separately:

$$g(\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \gamma_k)$$

- $(\pi_1, \ldots, \pi_K)'$ are the mixing proportions, where $\sum_{k=1}^{K} \pi_k = 1$
- $f(\cdot; \gamma_k)$ is the density of the k^{th} component

- For microarray data, we often assume $\mathbf{y}_i | Z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k)$
- For RNA-seq data, we need to choose the family and parameterization of $f(\cdot; \gamma_k)$

Question

Let $y_{ij\ell}$ be the observed count for gene i in condition j and replicate ℓ. Propose a distribution for $f(\cdot; \gamma_k)$

Poisson mixture model

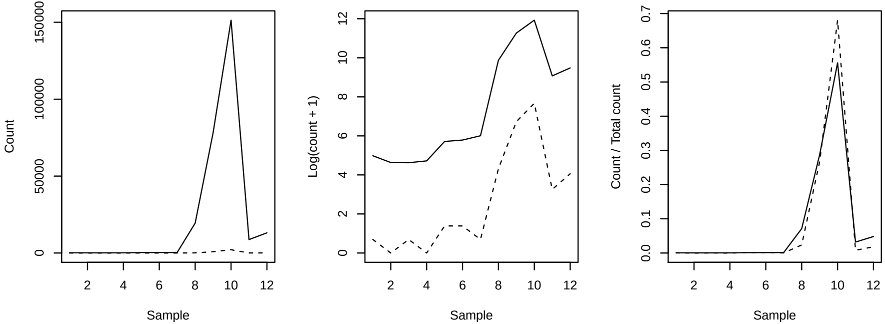
Let $y_{ij\ell}$ be the observed count for gene i in condition j and replicate ℓ
Assume

$$\mathbf{y}_i | Z_i = k \sim \prod_{j=1}^{J} \prod_{\ell=1}^{L_j} \mathcal{P}(y_{ij\ell}; \mu_{ij\ell k})$$

for $i = 1, \ldots, n$ independently, where variables are independent conditionally on the components.

Question: How to parameterize the mean $\mu_{ij\ell k}$ to obtain meaningful clusters of co-expressed genes?

Model parameterization: Which genes should be clustered?



Parameterization of the PMM

Consider the following parameterization:

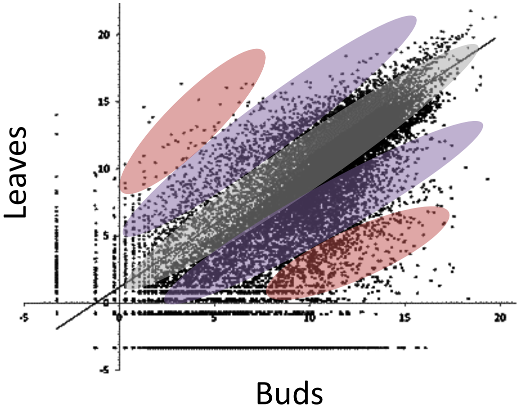
$$\mu_{ij\ell k} = w_i \lambda_{jk} s_{j\ell}$$

- w_i : overall expression level of observation i ($y_{i..}$)
- $\lambda_k = (\lambda_{jk})$: clustering parameters that define the profiles of genes in cluster k (variation around w_i)
- $s_{j\ell}$: normalized library size for replicate ℓ of condition j

- **Constraint**: $\sum_j \sum_\ell s_{j\ell} = 1$ for all $k = 1, \ldots, K$

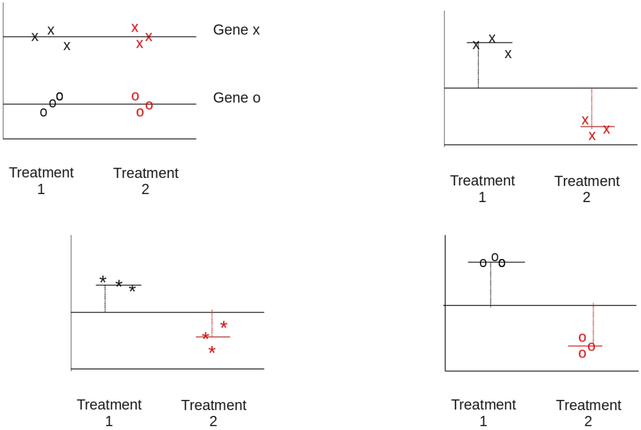
Interpretation

- ⇒ Genes are assigned to the same cluster if they share the same **profile of variation** around their mean count across all conditions



Interpretation

- ⇒ Genes are assigned to the same cluster if they share the same **profile of variation** around their mean count across all conditions



Inference: $\mu_{ij\ell k} = w_i \lambda_{jk} s_{j\ell}$

Expression level

$$\hat{w}_i = y_{i\cdot\cdot}$$

Library size effect

- The MLE estimator of $s_{j\ell}$ is the "total count" scaling factor:

$$\hat{s}_{j\ell} = y_{\cdot jl} / y_{\cdot\cdot\cdot}$$

- Other estimators possible: Trimmed Mean of M-values, quantile, DESeq, ...
- After estimating s_{jl} from the data, we consider this parameter to be fixed.

Parameter estimation

π_k's and λ's are estimated with an EM algorithm

Model selection

- BIC aims at finding a good number of components to a global fit of the data distribution

$$BIC(m) = \log P(\mathbf{Y}|m, \widehat{\theta}) - \frac{\nu_m}{2} \log(n).$$

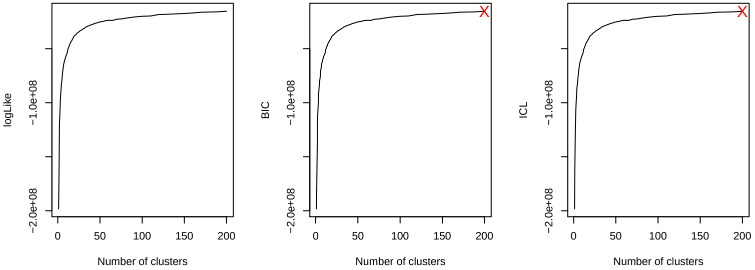
  where

  - ν_m is the number of free parameters of the model m
  - $P(\mathbf{Y}|m, \widehat{\theta})$ is the maximum likelihood under this model.

- ICL is dedicated to a classification purpose. The penality has an entropy term that penalizes stronger models for which the classification is uncertain.

$$ICL(m) = BIC(m) - \left\{ -\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik} \right\}$$

Behavior of BIC and ICL in practice for RNA-seq data



Slope heuristics (Birgé and Massart, 2006)

- Non-asymptotic framework: construct a penalized criterion such that the selected model has a risk close to the oracle model
- Theoretically validated in Gaussian framework, but encouraging applications in other contexts (Baudry et al., 2012)

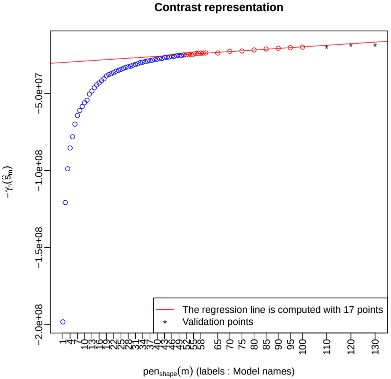
$$SH(m) = \log P(\mathbf{Y}|m, \hat{\theta}) + \kappa pen_{shape}(m)$$

In large dimensions:

- Stabilization of bias
- Linear behavior of $\frac{D}{n} \mapsto -\gamma_n(\hat{s}_D)$
- $\Rightarrow$ Estimation of slope to calibrate $\hat{\kappa}$ in a data-driven manner (Data-Driven Slope Estimation = DDSE), `capushe` R package

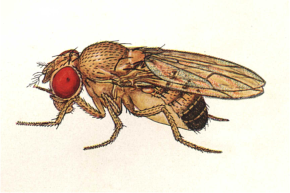
`HTSCluster`: Embryonic fly development

- Using slope heuristics, selected model is $\hat{K} = 48$ (selected model via BIC and ICL is $\hat{K} = 130$)

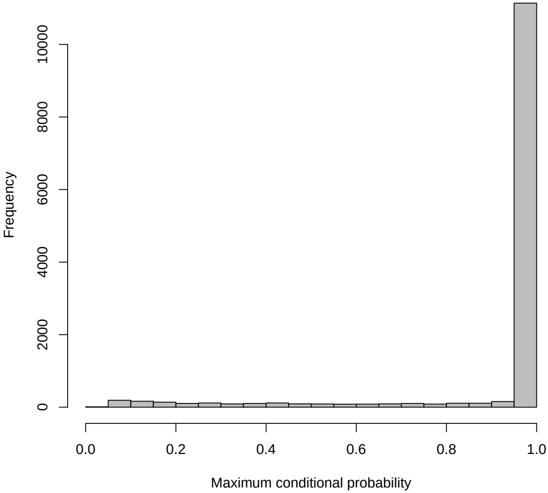


Real data analysis: Embryonic fly development

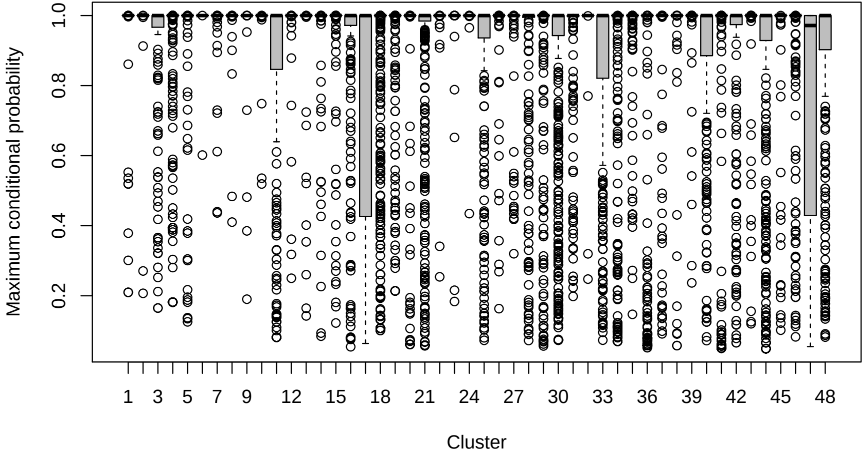
- modENCODE project to provide functional annotation of Drosophila (Graveley et al., 2011)
- Expression dynamics over 27 distinct stages of development during life cycle studied with RNA-seq
- 12 embryonic samples (collected at 2-hr intervals over 24 hrs) for 13,164 genes downloaded from ReCount database (Frazee et al., 2011)



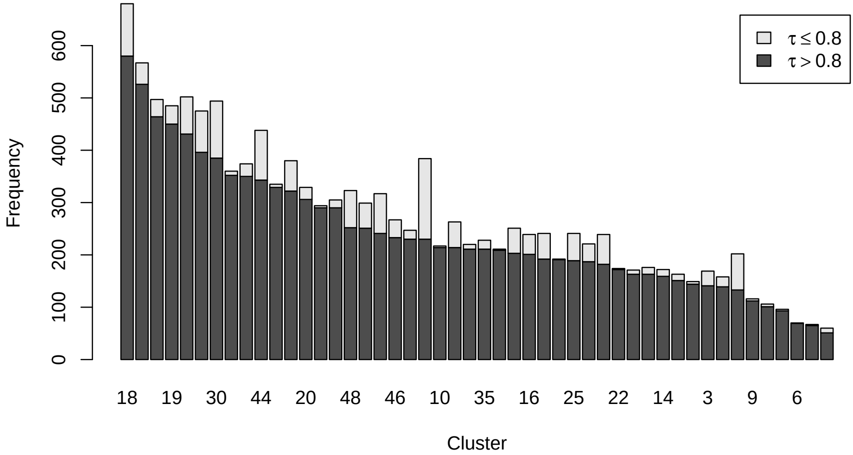
HTSCluster: Embryonic fly development



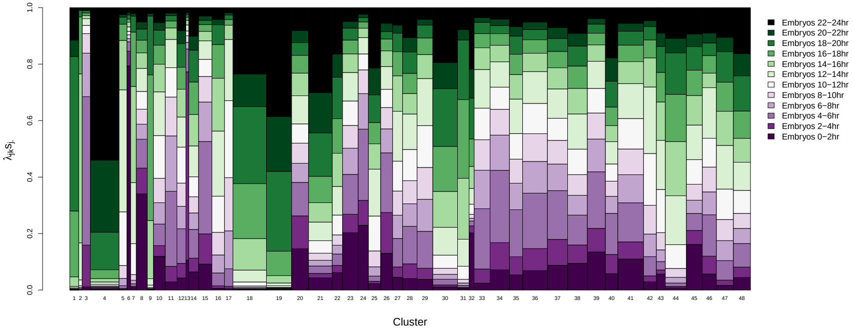
`HTSCluster`: Embryonic fly development



HTSCluster: Embryonic fly development



HTSCluster: Embryonic fly development



- Functional enrichment analysis: 33 of 48 clusters associated with at least one Gene Ontology Biological Process term (e.g., cluster 6 associated with muscle attachment)

Competing models

1. PoisL (Cai et al., 2004): K-means type algorithm using Poisson loglinear model
   - Equivalent to `HTSCluster` when equal library sizes, unreplicated data, equiprobable Poisson mixtures, and parameter estimation via the Classification EM (CEM) algorithm
2. Witten (2011): hierarchical clustering of dissimilarity measure based on a Poisson loglinear model
   - Originally intended to cluster samples
3. Si et al. (2014): model-based hierarchical algorithm using Poisson and negative binomial models
4. Classic K-means algorithm on expression profiles ($y_{ij\ell}/y_{i\cdot\cdot}$)

- Number of clusters not addressed by any of the above
- Functional enrichments of Si et al. (2014) seem to be weaker than ours

Conclusions: Model-based clustering for HTS data

`HTSCluster` for clustering count-based RNA-seq profiles:

- Poisson mixture model to directly model counts (i.e., no data transformations, etc.) from HTS experiments
- Interpretable parameter constraints lead to straightforward parameter estimation (EM algorithm), model selection (slope heuristics)
- Similar or better performance than previously proposed approaches on simulated data
- Analyses of real datasets are promising
- R package on CRAN available HTSCluster
- Recently published in Bioinformatics

Table of contents

1. Introduction
2. Normalization
3. Differential analysis
4. Clustering of transcriptomic data
5. **Experimental design**
   - Theoretical concepts
   - Conclusion and discussion

A statistical model: what for?

Aim of an experiment: answer to a biological question.

Results of an experiment: (numerous, numerical) measurements.

Model: mathematical formula that relates the experimental conditions and the observed measurements (response).

(Statistical) modelling: translating a biological question into a mathematical model ($\neq$ PIPELINE!)

Statistical model: mathematical formula involving

- the experimental conditions,
- the biological response,
- the parameters that describe the influence of the conditions on the (mean, theoretical) response,
- and a description of the (technical, biological) variability.

Experimental Design

Definition

A good design is dedicated to the **asked question** and facilitates data analyses and interpretation of the results. It maximizes collected information and proposes experiments with respect to the financial and material constraints.

Ronald A. Fisher (1890-1962)

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of

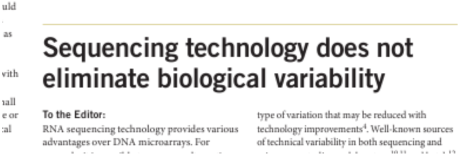
Basic principles - Fisher (1935)

- (technical and biological) replications
  Replication (independent obs.) $\neq$ Repeated measurements
- Randomization : randomize as much as is practical, to protect against unanticipated biases
- Blocking : dividing the observations into homogeneous groups. Isolating variation attributable to a nuisance variable (e.g. lane)

Basic principles - Fisher (1935)

- (technical <u>and</u> biological) replications
  Replication (independent obs.) $\neq$ Repeated measurements
- Randomization : randomize as much as is practical, to protect against unanticipated biases
- Blocking : dividing the observations into homogeneous groups. Isolating variation attributable to a nuisance variable (e.g. lane)

Correspondence Nature Biotechnology (July 2011)



Sequencing technology does not eliminate biological variability

To the Editor:
RNA sequencing technology provides various advantages over DNA microarrays. For

type of variation that may be reduced with technology improvements[4]. Well-known sources of technical variability in both sequencing and

Steps of experiment designing

1. Formulate a broadly stated research problem in terms of explicit, addressable questions.
2. Considering the population under study, identifying appropriate sampling or experimental units, defining relevant variables, and determining how those variables will be measured.
3. Describe the data analysis strategy
4. Anticipate eventual complications during the collection step and propose a way to handle them

source : Northern Prairie Wildlife Research Center, *Statistics for Wildlifers: How much and what kind?*

How to Design a good RNA-Seq experiment in an interdisciplinary context?

Some basic rules

- Rule 1 Share a minimal common language
- Rule 2 Well define the biological question
- Rule 3 Anticipate difficulties with a well designed experiment
- Make good choices : Replicates vs Sequencing depth

Rule 2: Well define the biological question

- Choose scientific problems on feasibility and interest
- Order your objectives (primary and secondary)
- Ask yourself if RNA-seq is better than microarray regarding the biological question

Recall that RNA-Seq technology is useful to

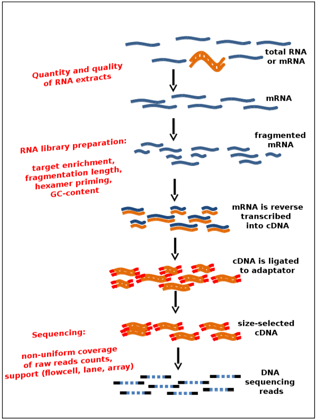
- Study all the transcribed entities
- Detect and estimate isoforms
- Construct and study a *de novo* transcriptome

Rule 3: Anticipate difficulties with a well designed experiment

1. Prepare a checklist with all the needed elements to be collected,
2. Collect data and determine all factors of variation,
3. Choose bioinformatics and statistical models,
4. Draw conclusions on results.

Be aware of different types of bias

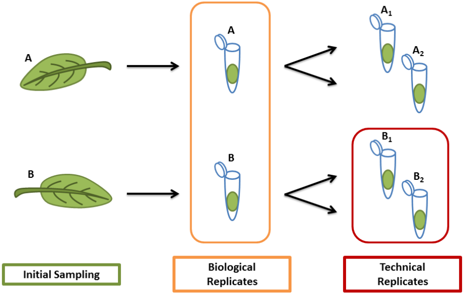
Identify controllable biases / technical specificities


total RNA or mRNA
Quantity and quality of RNA extracts
mRNA
fragmented mRNA
RNA library preparation:
target enrichment, fragmentation length, hexamer priming, GC-content
mRNA is reverse transcribed into cDNA
cDNA is ligated to adaptator
size-selected cDNA
Sequencing:
non-uniform coverage of raw reads counts, support (flowcell, lane, array)
DNA sequencing reads

Keep in mind the influence of effects on results:
lane $\leq$ run $\leq$ RNA library preparation $\leq$ biological
(Marioni, 2008), (Bullard, 2010)

$\Rightarrow$ Increase biological replications !

Biological and technical replicates



Biological replicate : sampling of individuals from a population in order to make inferences about that population

Technical replicate adresses the measurement error of the assay.

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate to a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)

Why increasing the number of biological replicates?

- To generalize to the population level
- To estimate to a higher degree of accuracy variation in individual transcript (Hart, 2013)
- To improve detection of DE transcripts and control of false positive rate: TRUE with at least 3 (Sonenson 2013, Robles 2012)

McIntyre et al. (2011) BMC Genomics

Technical variability $=>$ inconsistent detection of exons at low levels of coverage ($<$5reads per nucleotide)
Doing technical replication may be important in studies where low abundant mRNAs are the focus.

More biological replicates or increasing sequencing depth?

It depends! (Haas, 2012), (Liu, 2014)

- DE transcript detection: (+) biological replicates
- Construction and annotation of transcriptome: (+) depth and (+) sampling conditions
- Transcriptomic variants search: (+) biological replicates and (+) depth

A solution: multiplexing.

Tag or bar coded with specific sequences added during library construction and that allow multiple samples to be included in the same sequencing reaction (lane)

Decision tools available: Scotty (Busby et al. 2013),
Library RNAseqPower in Bioconductor (Hart et al., 2013)

To summarize

The scientific question of interest drives the experimental choices

- Collect informations before planning
- All skills are needed to discussions right from project construction
- Optimum compromise between replication number and sequencing depth depends on the question
- Biological replicates are important in most RNA-seq experiments
- Wherever possible apply the three Fisher's principles of randomization, replication and local control (blocking)

And do not forget: budget also includes cost of biological data acquisition, sequencing data backup, bioinformatics and statistical analysis.