Statistical challenges in RNA-Seq data analysis

Julie Aubert
UMR 518 AgroParisTech-INRA Mathématiques et Informatique Appliquées

Ecole de bioinformatique, Station biologique de Roscoff, 2013 Nov. 18

A statistical model : what for ?

Aim of an experiment : answer to a biological question.

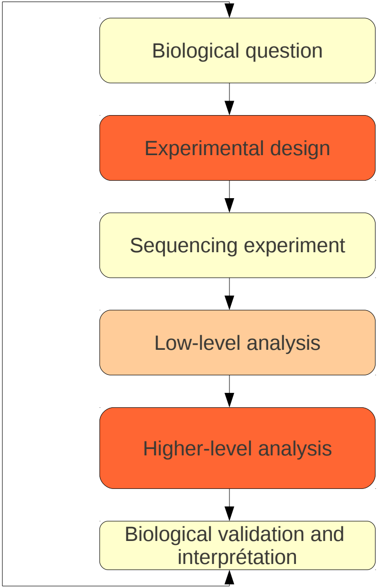
Results of an experiment : (numerous, numerical) measurements.

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.



Exploratory Data Analysis, image analysis, base calling, read mapping, metadata integration

Exploratory Data Analysis, normalization and expression quantification, differential analysis, metadata integration

*Adapted from S. Dudoit, Berkeley

Outline

1. Experimental design
2. Normalization
3. Differential analysis
4. Conclusions

Experimental Design - checklist (Dean and Voss 1999)

1. Define the objectives of the experiment.
2. Identify all sources of variation including treatment factors and their levels, experimental units, blocking factors, noise factors and covariates
3. Choose a rule for assigning the experimental units to the treatment.
4. Specify the measurements to be made, the exp. procedure and the anticipated diff.
5. Run a pilot exp.
6. Specify the model.
7. Outline the analysis.
8. Calculate the number of obs. that need to be taken.
9. Review the above decisions. Revise if necessary.

Experimental Design

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

Application to NGS

- Identify controllable biases / technical specificities
- lane effect< run effect< library prep effect<<biological effect [Marioni et al 2008, Bullard et al 2010]
  $\Rightarrow$ Increase biological replications !

"Sequencing technology does not eliminate biological variability", Correspondence Nature Biotechnology (July 2011)

Experimental Design

Definition

A good design is a list of experiments to conduct in order to answer to the **asked question** which maximize collected information and minimize the number of experiments (or the experiments cost) with respect to constraints.

Objectives RNA-seq

Ex : To find genes or transcripts differentially expressed between several conditions.

Experimental Design

Technical choices

Choice of sequencing technology, type of reads (paired-end ?), type of sequencing (directional ?), library preparation protocol.

Sequencing depth

Barcoding (*attaching a known sequence of nucleotides to the 3' ends of the NGS technology adapter sequences identifing a sample*) or not

Pooling* of barcoded sample for a simultaneous sequencing and number of samples.

Technical challenge : combining approximately equal ratios of cDNA preparations to achieve approximately similar depths of sequencing for all samples

Replicate number and sample allocation to runs/lanes

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.

Technical vs biological replicates

- Increasing the number of bio. replicates increases the precision and generalizability of the results
- Technical variability => inconsistent detection of exons at low levels of coverage (<5reads per nucleotide) (McIntyre et al. 2011)
- Doing technical replication may be important in studies where low abundant mRNAs are the focus.

Experimental Design : illustration (1)

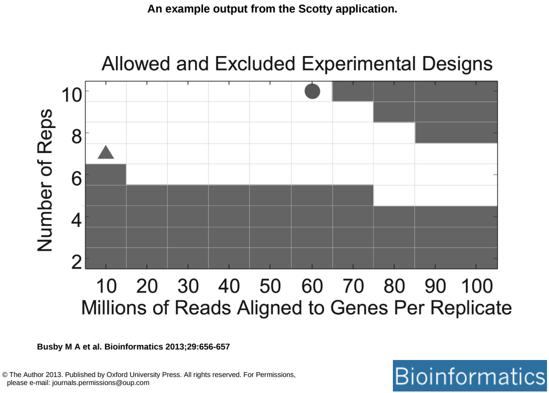
José A Robles et al. BMC Genomics 2012

- Increasing the number of bio. replicates increases True Positive Rate (stable False Positive Rate).
- Increasing coverage depth at fixed number of bio. replicates increases slightly TPR with a stable FPR.
- Increasing the number of bio. replicates at 1/n x 100% seq. depth (multiplexing) increases TPR.

Increasing the number of replicates sample more powerful than increasing sequencing depth (Rapaport et al. 2013)

Experimental Design : illustration (2)

It's a balance : cost, precision ⟺ nb bio. replicates, sequencing depth.



An example output from the Scotty application.

Busby M A et al. Bioinformatics 2013;29:656-657

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Bioinformatics

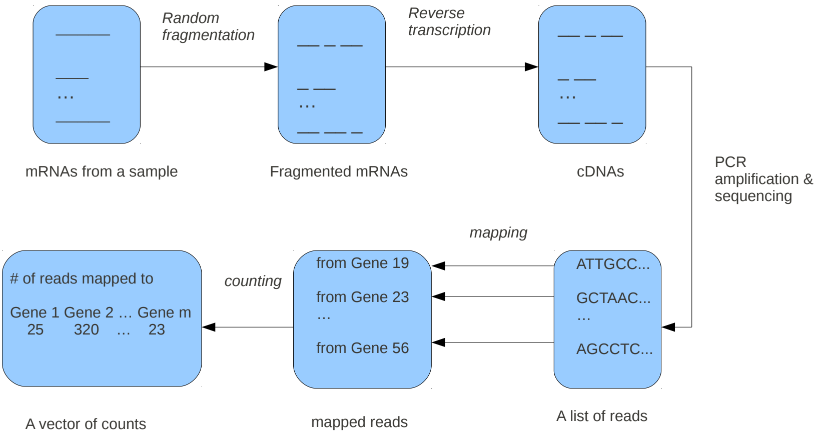
This figure shows the user which of the tested experimental configurations do (white) and do not (shaded) conform to the user-defined constraints. Scotty then indicates the optimal configuration based on cost (filled triangle) and power (filled circle).

Busby M A et al. Bioinformatics 2013

Experimental Design : conclusions

- The scientific question of interest drives the experimental choices
- Collect informations before planning
- A good design is a balance between nb of bio. replicates and sequencing depth

RNA-sequencing



Adapted from Li et al. (2011)

Exploratory data analysis

Conducting data analysis is like drinking a fine wine. It is important to swirl and sniff the wine, to unpack the complex bouquet and to appreciate the experience. Gulping the wine doesn't work. (Daniel B. Wright - 2003)

Normalization

Definition

Normalization is a process designed to identify and correct **technical biases** removing the least possible biological signal. This step is technology and platform-dependant.

Within-lane normalization

Normalisation enabling comparisons of fragments (genes) from a same sample.
No need in a differential analysis context.

Between-lane normalization

Normalisation enabling comparisons of fragments (genes) from different samples.

Sources of variability

Within-sample

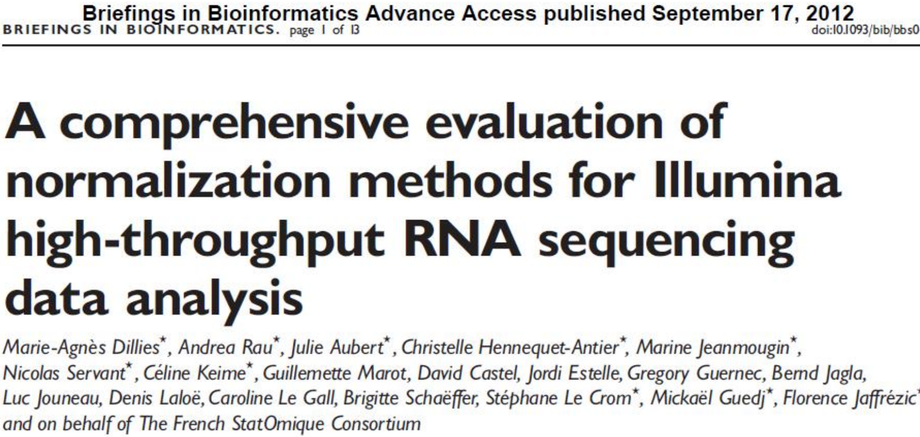
- Gene length
- Sequence composition (GC content)

Between-sample

- Depth (total number of sequenced and mapped reads)
- Sampling bias in library construction ?
- Presence of majority fragments
- Sequence composition du to PCR-amplification step in library preparation‘(Pickrell et al. 2010, Risso et al. 2011)

StatOmique workshop

http://vim-iip.jouy.inra.fr:8080/statomique/



Briefings in Bioinformatics Advance Access published September 17, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13
doi:10.1093/bib/bbs0

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic*
and on behalf of The French StatOmique Consortium*

Comparison of normalization methods

At lot of different normalization methods...

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

Which one is the best ?

- How to and on which criteria choice a normalisation adapted to our experiment ?
- What impact of the bioinformatics, normalisation step or differential analysis method on lists of DE genes ?

Normalisation methods

Global methods : normalised counts are raw counts divided by a scaling factor calculated for each sample

Distribution adjustment

Assumption (TC, UQ, Median) : read counts are prop. to expression level and sequencing depth
Total number of reads : TC (Marioni et al. 2008), Quantile : FQ (Robinson and Smyth 2008), Upper Quartile : UQ (Bullard et al. 2010), Median

Method taking length into account

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

The Effective Library Size concept

Trimmed Means of M-values TMM (Robinson et Oschlack 2010, *edgeR*)
DESeq (Anders et Huber 2010, *DESeq*)

Notations

- x_{ij} : number of reads for gene i in sample j
- N_j : number of reads in sample j (library size of sample j)
- n : number of samples in the experiment
- $\hat{s}_j$: normalization factor associated with sample j
- L_i : length of gene i

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

Adjust for lane sequencing depth (library size)

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

$$\frac{x_{ij}}{N_j} = \frac{x_{ij}}{\hat{s}_j}, \qquad \hat{s}_j = N_j \tag{1}$$

- Problem makes *frequencies* comparable between lanes, *not read counts*
- Solution rescale scaling factors so that their sum across lanes is equal to, e.g., the number of lanes
- Makes normalization procedures comparable

$$\hat{s}_j = \frac{N_j}{\frac{1}{n}\sum_l N_l} \tag{2}$$

RPKM normalization

Reads Per Kilobase per Million mapped reads
Adjust for lane sequencing depth (library size) and gene length

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

$$\frac{x_{ij}}{N_j * L_i} * 10^3 * 10^6 \tag{3}$$

- Allows to compare expression levels between genes of the same sample
- Unbiased estimation of number of reads but affect the variance. (Oshlack et al. 2009)

The Effective Library Size concept

Motivation

Different biological conditions express different RNA repertoires, leading to different total amounts of RNA

Assumption

A majority of transcripts is not differentially expressed

Aim

Minimizing effect of (very) majority sequences

Methods based on the Effective Library Size Concept

Trimmed Mean of M-values Robinson et al. 2010 (edgeR)

Filter on transcripts with nul counts, on the resp. 30% and 5% more extreme $M_i = log2(\frac{Y_{ik}/N_k}{Y_{ik'}/N'_k})$ and A values
Hyp : We may not estimate the total ARN production in one condition but we may estimate a global expression change between two conditions from non extreme M_i distribution.

Calculation of a scaling factor between two conditions and normalization to avoid dependance on a reference sample

Anders and Huber 2010 - Package DESeq

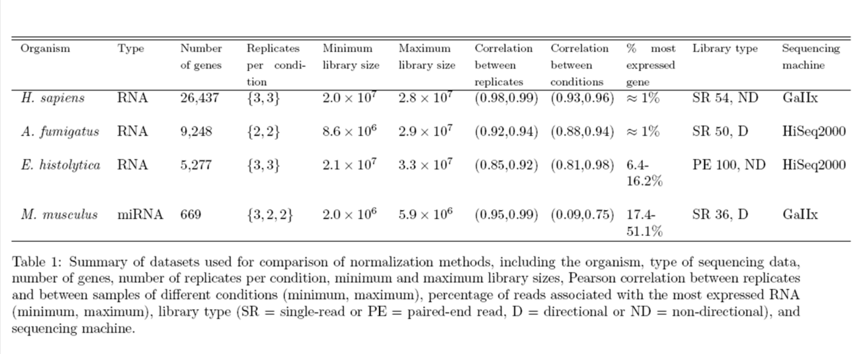
$$\hat{s}_j = median_i(\frac{k_{ij}}{(\pi_{v=1}^{m} k_{iv})^{1/m}})$$

k_{ij} : number of reads in sample j assigned to gene i

denominator : pseudo-reference sample created from geometric mean across samples

4 real datasets and one simulated dataset

RNA-seq or miRNA-seq, DE, at least 2 conditions, at least 2 bio. rep., no tech. rep.



Organism	Type	Number of genes	Replicates per condition	Minimum library size	Maximum library size	Correlation between replicates	Correlation between conditions	% most expressed gene	Library type	Sequencing machine
H. sapiens	RNA	26,437	$\{3,3\}$	2.0×10^7	2.8×10^7	(0.98,0.99)	(0.93,0.96)	$\approx 1\%$	SR 54, ND	GaIIx
A. fumigatus	RNA	9,248	$\{2,2\}$	8.6×10^6	2.9×10^7	(0.92,0.94)	(0.88,0.94)	$\approx 1\%$	SR 50, D	HiSeq2000
E. histolytica	RNA	5,277	$\{3,3\}$	2.1×10^7	3.3×10^7	(0.85,0.92)	(0.81,0.98)	6.4-16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	$\{3,2,2\}$	2.0×10^6	5.9×10^6	(0.95,0.99)	(0.09,0.75)	17.4-51.1%	SR 36, D	GaIIx

Table 1: Summary of datasets used for comparison of normalization methods, including the organism, type of sequencing data, number of genes, number of replicates per condition, minimum and maximum library sizes, Pearson correlation between replicates and between samples of different conditions (minimum, maximum), percentage of reads associated with the most expressed RNA (minimum, maximum), library type (SR = single-read or PE = paired-end read, D = directional or ND = non-directional), and sequencing machine.

Comparison procedures

Distribution and properties of normalized datasets

Boxplots, variability between biological replicates

Comparison of DE genes

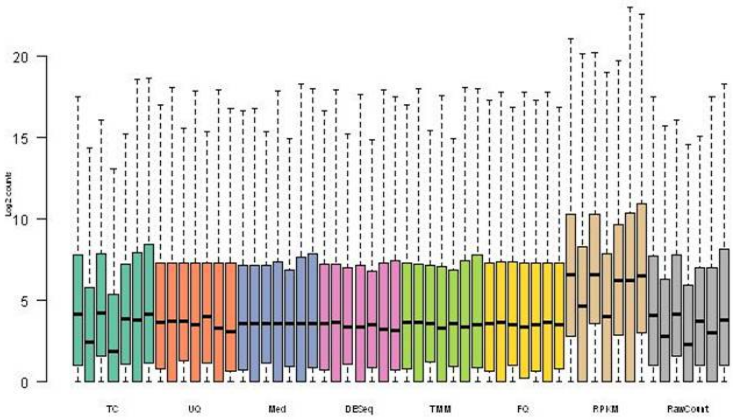
- Differential analysis : DESeq v1.6.1 (Anders and Huber 2010), default param.
- Number of common DE genes, similarity between list of genes (dendrogram - binary distance and Ward linkage)

Power and control of the Type-I error rate

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

Normalized data distribution

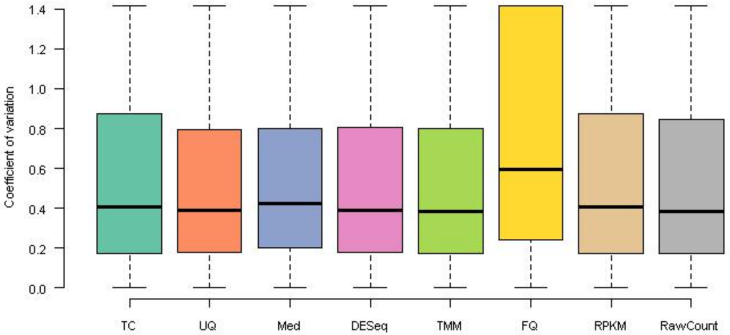
When large diff. in lib. size, TC and RPKM do not improve over the raw counts.



Example : *Mus musculus* dataset

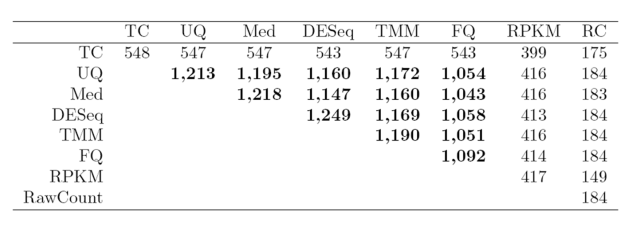
Within-condition variability

Example : *Mus musculus, condition D* dataset



Number of DE genes

- DESeq v1.6.0, default parameters
- Input data : raw counts + scaling factors $\hat{s}_j$ (except RPKM)
- RPKM : normalized data **non rounded** and normalization parameter $\hat{s}_j = 1$

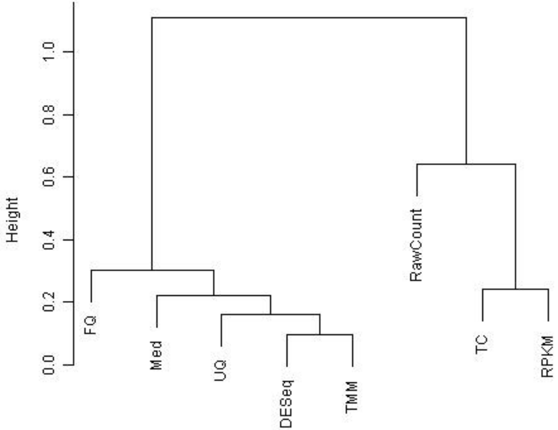


	TC	UQ	Med	DESeq	TMM	FQ	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1,213**	**1,195**	**1,160**	**1,172**	**1,054**	416	184
Med			**1,218**	**1,147**	**1,160**	**1,043**	416	183
DESeq				**1,249**	**1,169**	**1,058**	413	184
TMM					**1,190**	**1,051**	416	184
FQ						**1,092**	414	184
RPKM							417	149
RawCount								184

Example : *E. histolytica* dataset, common genes

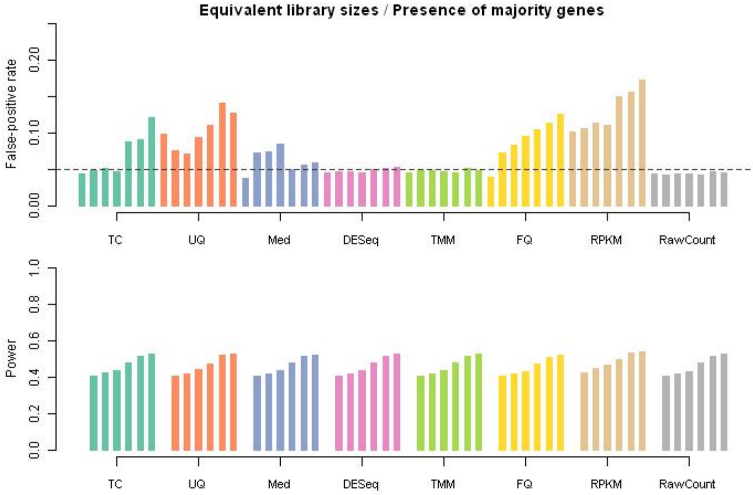
Consensus dendrogram - Ward linkage algo.

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



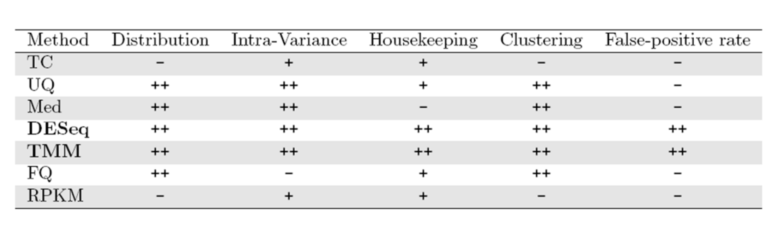
So the Winner is ... ?

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	–	+	+	–	–

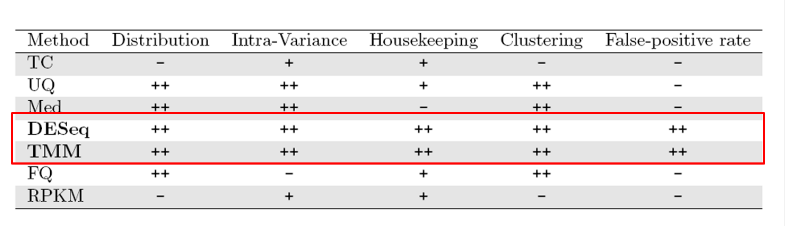
So the Winner is ... ?

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	–	+	+	–	–

Interpretation

- RawCount Often fewer differential expressed genes (*A. fumigatus* : no DE gene)
- TC, RPKM
  - Sensitive to the presence of majority genes
  - Less effective stabilization of distributions
  - Ineffective (similar to RawCount)
- FQ
  - Can increase between group variance
  - Is based on an very (too) strong assumption (similar distributions)
- Median High variability of housekeeping genes
- TC, RPKM, FQ, Med, UQ Adjustment of distributions, implies a similarity between RNA repertoires expressed

Conclusions on “StatOmique” workshop

- Hypothesis : the majority of genes is invariant between two samples.
- Differences between methods when presence of majority sequences, very different library depths.
- TMM and DESeq : performant and robust methods in a DE analysis context on the gene scale.
- Normalisation is necessary and not trivial.

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
- A normalization is needed and has a great impact on the DE genes (Bullard et al 2010), (Dillies et al 2012)

- Detection of differential expression in RNA-seq data is inherently biased (more power to detect DE of longer genes)
- Do not normalise by gene length in a context of differential analysis.

Differential analysis

Aim : To detect differentially expressed genes between two conditions

- Discrete quantitative data
- Few replicates
- Overdispersion problem

Challenge : method which takes into account overdispersion and few number of replicates

- Proposed methods : edgeR, DESeq for the most used and known *Anders et al. 2013, Nature Protocols*
- An abundant littérature
- Comparison of methods : Pachter et al. (2011), Kvam et Liu (2012), Soneson et Delorenzi (2013), Rapaport el al. (2013)

Differential analysis gene-by-gene- with replicates

For each gene i

Is there a significant difference in expression between condition A and B ?

- Statistical model (definition and parameter estimation) - Generalized linear framework
- Test : Equality of relative abundance of gene i in condition A and B vs non-equality

The Poisson Model

Let be Y_{ijk} the count for replicate j in condition k from gene i

- Y_{ijk} follows a Poisson distribution (μ_{ijk}).
- Property : $Var(Y_{ijk}) = Mean(Y_{ijk}) = \mu_{ijk}$

Overdispersion in RNA-seq data

Counts from biological replicates tend to have variance exceeding the mean (= overdispersion relative to the Poisson distribution)

What causes this overdispersion ?

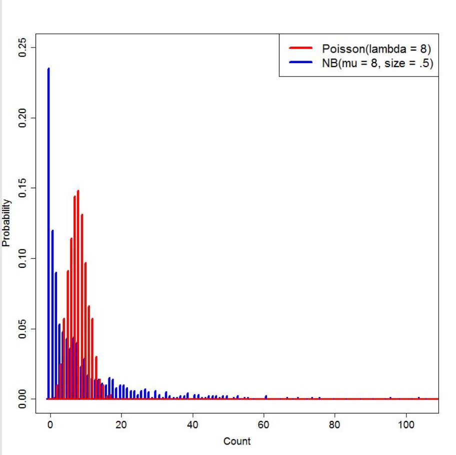
- Correlated gene counts
- Clustering of subjects
- Within-group heterogeneity
- Within-group variation in transcription levels
- Different types of noise present...

In case of overdispersion, ↑ of the type I error rate (prob. to declare incorrectly a gene DE).

Alternative : Negative Binomial Models

A supplementary dispersion parameter ϕ to model the variance

Poisson vs Negative Binomial Models



Types of noise in data

1. Shot noise : unavoidable noise inherent in counting process (dominant for weakly expressed genes)
2. Technical noise : from sample preparation and sequencing, hopefully negligable
3. Biological noise : unaccounted for differences between samples (dominant for strongly expressed genes)

ϕ estimation

Many genes, very few biological samples - difficult to estimate ϕ on a gene-by-gene basis

Some proposed solutions

Method	Variance	Reference
DESeq	$\mu(1+\phi_\mu \mu)$	Anders et Huber (2010)
edgeR	$\mu(1+\phi\mu)$	Robinson et Smyth (2009)
NBPseq	$\mu(1+\phi\mu^{\alpha-1})$	Di et al. (2011)

- **edgeR** : borrow information across genes for stable estimates of ϕ 3 ways to estimate ϕ (common, trend, moderated)
- **DESeq** : data-driven relationship of variance and mean estimated using parametric or local regression for robust fit across genes
- **NBPseq** : ϕ and α estimated by LM based on all the genes.

Negative Binomial Models

▸ $\mu_{ijk} = \lambda_{ij} m_{jk}$ where $m_j k$: size factor (library size)
Test : H_{0i} : $\lambda_{iA} = \lambda_{iB}$ vs H_{1i} : $\lambda_{iA} \neq \lambda iB$

edgeR

- Adjust observed counts up or down depending on whether library sizes are below or above the geometric mean => Creates approximately identically distributed pseudodata
- Estimation of ϕ_i by conditional ML conditioning on the TC for gene i
- Empirical Bayes procedure to shrink dispersions toward a consensus value
- An exact test analogous to Fisher's exact test but adapted to overdispersed data (Robinson and Smyth 2008)

DESeq

Test similar to Fisher's exact test (calculation has changed)

Negative Binomial Models - DESeq

Assumptions :

1. $Y_{ijk} \sim NB(\mu_{ijk}, \sigma_{ijk})$, where μ_{ijk} is the mean, and σ_{ijk} is the variance
2. The mean μ_{ijk} is the product of a condition-dependent per-gene value λ_{ij} and a size factor (library size) m_{jk} :

$$\mu_{ijk} = \lambda_{ij} m_{jk}$$

3. Variance decomposition : The variance σ_{ijk} is the sum of a shot noise term and a raw variance term : $\sigma_{ijk} = \mu_{ijk} + \alpha_i \mu^2$ where α_i the dispersion value.
4. Per-gene dispersion α_i or pooled α is a smooth function of the mean :

$$\alpha_i = f_j(\lambda_{ij})$$

DESeq Bioconductor package

Three sets of parameters need to be estimated :

1. Size factors $m_j k$ (normalization factors) *(see normalization part)*
2. For each experimental condition j, n expression strength parameters λ_{ij} estimated by average of counts from the replicates for each condition, transformed to the common scale :

$$\hat{\lambda_{ij}} = \frac{1}{r_j} \sum_k \frac{y_{ijk}}{\hat{m_j k}}$$

3. The smooth functions f_j for each condition j to model dependence of α_i on the expected mean λ_{ij} : local or gamma GLM estimation (*fit='local'* or *fit='parametric'*)

Practical considerations

Input Data = raw counts
normalization offsets are included in the model

- Version matters : edgeR 2.6.7 et DESeq 1.6.1 (Bioconductor 2.9)

edgeR

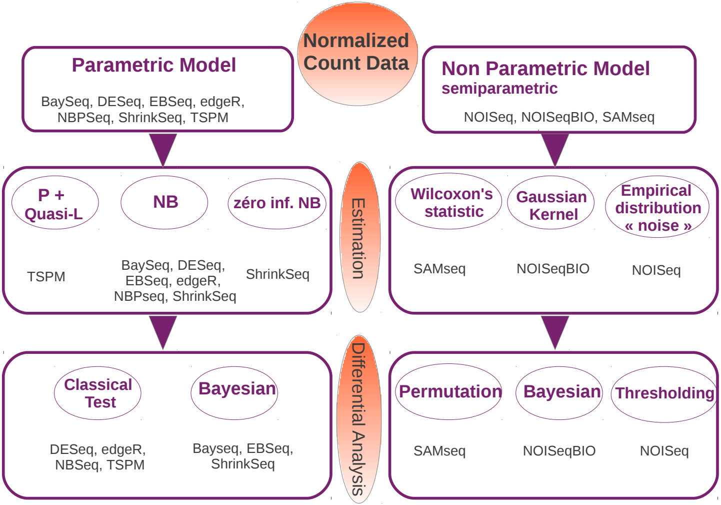
TMM normalization Common dispersion must be estimated before tagwise dispersions GLM functionality (for experiments with multiple factors) now available

DESeq

Two possibilities to obtain a smooth functions fj $(\cdot)$

- Conservative estimates : genes are assigned the maximum of the fitted and empirical values of α_i (sharingMode = "maximum")
- Local fit regression (as described in paper) is no longer the default
- Each column = independent biological replicate

A lot of statistical methods...still developped



Differential Analysis between two conditions

edgeR or DESeq or another method ?

- None is perfect !
- results obtained by edgeR and DESeq are mostly the same

Robles et al. 2011

edgeR

- a slightly inflated FPR from edgeR (small values of n or only high-counts transcripts)
- Performance improves as number of replicates increases

DESeq

- conservative whatever n
- over-conservative behaviour when only low-counts (<100)

Remark : non-parametric methods not enough detection power (Kvam and Liu 2012).

Comparaison of differential analysis methods

Soneson et Delorenzi (2013)

Evaluation of 11 methods on both simulated and real data.

- Very small sample sizes => pb for all methods : be caution in your interpretation
- For larger sample size, a variance-stabilizing transformation with limma or SAMseq method (min. 5) quite good results

Rapaport et al. (2013)

Evalution on methods using SEQC benchmark dataset and ENCODE data.

- Significant differences between methods.
- Array-based methods adapted perform comparably to specific methods.
- Increasing the number of replicates samples significantly improves detection power over increased sequencing depth.

DESeq2 Love and Huber (2013)

Differences with DESeq.

- Dispersion shrinkage
- Fold Change shrinkage (for CPA and Gene Set Enrichissment Analysis)
- Detection of outliers

- Improve power
- Only one command line
- Not publied ?

Multiple Testing

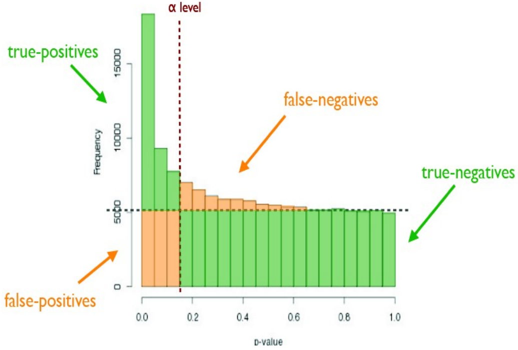
False positive (FP) (**type I error** : α) : A not DE gene which is declared DE.

For all 'genes', we test H_0 (gene i is not DE) vs H_1 (the gene is DE)

Pb :

If all the genes are not DE et each test is realised at α level
then for 10000 genes and $\alpha = 0.05$ we have $E(FP) = 500$ genes.

Multiple testing



Source : M. Guedj, Pharnext

The procedure of Benjamini-Hochberg (95) controls the False Discovery Rate FDR$=E(FP/P)$ si$P>0$.

The Bonferroni procedure controls the Family-Wise Error Rate

Conclusions on differential analysis

- Methods dedicated to microarrays are not applicable to RNA-seq
- Adaptation of these methods quite good behaviour when the number of replicates increases
- Negative binomial (NB) model framework
- NB : distinction of methods on the information sharing for modelisation of the dispersion parameter (needed when n small)
- Negative binomial model not enough flexible ? (how to take into account zero-inflation and heavy tail) : ZI-BN, Tweedie-Poisson

Adapt the method to your data (nb of rep.)

Specific methods developped for few replicates.The need for 'sophisticated' methods decreases when the number of replicates increases.

Other questions

Gene-Set Enrichissment Analysis

These tests assume that genes have the same chance to be declared DE.
But sometimes over-detection of longer and mare expressed genes
GOSeq (Young et al. 2011)

Filter or not

General conclusions

Pratical conclusions

- Need to collaborate between biologists, bioinformaticians et statisticians
- and in a ideal world since the project construction
- Adaptation of methods and tools to the asked question (no pipeline)
- Check all the steps of the data analysis (quality, normalization, differential analysis . . .)

Statistics not only useful for differential analysis with RNA-seq

Aknowledgements - StatOmique

- All the participants of the StatOmique workshop :***M.-A. Dillies***, *B. Jagla*, ***A. Rau***, *J. Estelle, G. Guernec, L. Jouneau, B. Schaeffer, D. Laloe, C. Hennequet-Antier, M. Jeanmougin, M. Guedj, N. Servant, C. Keime, D. Castel, S. Le Crom, F. Jaffrezic, G. Marot, C. Le Gall, D. Charif*
- The biologists who annotated or accepted their data be included in the study : *C. Chau Hon, T. Strub, I. Davidson, G. Janbon*

`http://vim-iip.jouy.inra.fr:8080/statomique/`

References

- Anders, S and Huber, W. (2010) **Differential expression analysis for sequence count data**, *Genome Biology*,11 :R106.
- Anders, S, McCarthy, DJ, Chen, Y, Okoniewski, M, Smyth GK, Huber, W and Robinson, MD (2013) *Count-based differential expression analysis of RNA sequencing data using R and Bioconductor*, *Nature Protocols*, doi :10.1038.
- Bullard JH, Purdom E, Hansen KD, Dudoit S. (2010) **Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments**, *BMC Bioinformatics*, 11 :94
- Di Y, Schaef, DW, Cumbie JS, Chang JH (2011) **The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq**, *Statistical Applications in Genetics and Molecular Biology*, 10(1), Article 24.

References

- Dillies M-A et al. on behalf of The French StatOmique Consortium (2012) **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis**, *Briefing in Bioinformatics*.
- Kvam V, Liu P (2012) **A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data**
- Li J, Jiang H, Wong WH, (2010) **Modeling non-uniformity in short read rates in RNA-Seq data**, *Genome Biology*, 11 :R50
- Li J, Witten DM, Johnstone IM, Tisbhirani R (2011) **Normalization, testing, and false discovery rate estimation for RNA-sequencing data**, *Biostatistics*, 1-16
- Marioni J.C., Mason C.E. et al. (2008) **RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays**, *Genome Research*, 18 : 1509-1517.

References

- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008) **Mapping and quantifying mammalian transcriptomes by RNA-seq.** *Nature Methods*, 5(7), 621-628
- Pachter L (2011) **Models for transcript quantification from RNA-seq**
- Rapaport et al. (2013) **Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data**, *Genome Biology*,14 :R95
- Robles et al (2012) **Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing**, preprint
- Robinson MD, Oshlack A. (2010) **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biology*, 11 :R25
- Robinson MD and Smyth, GK. (2008) **Small-sample estimation of negative binomial dispersion, with applications to SAGE data**

References

- Robinson MD, McCarthy DJ, Smyth, GK. (2009) **edgeR : a Bioconductor package for differential expression analysis of digital gene expression data**, *Bioinformatics*
- Soneson, C, Delorenzi, M. (2013) **A comparison of methods for differential expression analysis of RNA-seq data**. *BMC Bioinformatics*,14 :91
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. (2011) **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**, *Nature Biotechnology*, 28(5) : 511 ?515.
- Young MD, Wakefiled MJ, Smyth GK., Oshlack A. (2011) **Gene ontology analysis for RNA-seq :accounting for selection bias** , *Genome Biology*

Length bias (Oshlack 2009, Bullard et al. 2010)

At same expression level, a long transcript will have more reads than a shorter transcript. Number of reads $\neq$ expression level

$$\mu = E(X) = cNL = Var(X)$$

- X mesured number of reads in a library mapping a specific transcript, Poisson r.v.
- c proportionnality constant
- N total number of transcripts
- L gene length

Test :

$$t = \frac{X_1 - X_2}{\sqrt(cN_1L + cN_2L)}$$

Power of test depends on a parameter prop. to $\sqrt(L)$.

Identical result after normalization by gene length (but out of the Poisson framework).

First commands

Installation des packages :

```
source("http ://www.bioconductor.org/biocLite.R")
biocLite(c("DESeq", "edgeR"))
```

Chargement des packages :

```
library(DESeq)
library(edgeR)
```

edgeR main commands

generate raw counts from NB, create list object

```
y <- matrix(rnbinom(80,size=1/0.2,mu=10),nrow=20,ncol=4)
rownames(y) <- paste("Gene",1 :nrow(y),sep=".")
group <- factor(c(1,1,2,2))
```

perform DA with edgeR

```
y <- DGEList(counts=y,group=group)
y <- calcNormFactors(y,method="TMM")
y <- estimateCommonDisp(y)
y <- estimateTagwiseDisp(y)
result <- exactTest(y,dispersion="tagwise")
```

Observe some results - DGE with FDR BH

```
topTags(result)
summary(decideTestsDGE(result),p.value=0.05)
```

DESeq main commands

```
cds <- newCountDataSet(y, group)
cds <- estimateSizeFactors(cds)
sizeFactors(cds)
cds <- estimateDispersions(cds)
res <- nbinomTest( cds, "1", "2" )
```

Quelques références pour débuter

- http://www.r-project.org/ : manuel, FAQ, RJournal, etc...
- http://www.bioconductor.org/help/publications/
- cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf
- G. Millot, (2009), Comprendre et réaliser les tests statistiques à l'aide de R, Editions De Boeck, 704 p.