Statistical challenges in (RNA-Seq) data analysis

Julie Aubert
UMR 518 AgroParisTech-INRA Mathématiques et Informatique Appliquées

Ecole de bioinformatique, Station biologique de Roscoff, 2014 Oct. 7
Statistics = study of the collection, analysis, interpretation, presentation and organization of data (source: Wikipedia)
Biological question

Experimental design

Sequencing experiment

Low-level analysis

Higher-level analysis

Biological validation and interpretation

*Adapted from S. Dudoit, Berkeley

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 Normalisation and Differential analysis
3 Conclusions
Another definition

Statistics consists of trying to understand data and to obtain more understandable data. Savage (1977)

Key of a good data analysis: having good data to analyze. Requisite: a clearly defined research objective.

The statistician who supposes that his main contribution to the planning of an experiment will involve statistical theory, finds repeatedly that he makes his most valuable contribution simply by persuading the investigator to explain why he wishes to do the experiment. Gertrude M Cox
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 (≠ 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.
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?
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.

Basic principles - Fisher (1935)

- (technical and biological) replications
  Replication (independent obs.) ≠ 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)
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

From Alon, 2009
Objectives - RNA-Seq

- Identify differentially expressed (DE) genes?
- Detect and estimate isoforms?
- Construct 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

Keep in mind the influence of effects on results:

\[ \text{lane} \leq \text{run} \leq \text{RNA library preparation} \leq \text{biological} \]

(Marioni, 2008), (Bullard, 2010)
RNA-seq experiment analysis: from A to Z

Experimental design

Choose bioinformatics & statistical models

Answer your biological question

* function only implemented in spliced read mapping tools

Adapted from Mutz, 2013
Make good choices
How many reads?

- $100M$ to detect $90\%$ of the transcripts of $81\%$ of human genes (Toung, 2011)
- $20M$ reads of $75bp$ can detect transcripts of medium and low abundance in chicken (Wand, 2011)
- $10M$ to cover by at least $10$ reads $90\%$ of all (human and zebrafish) genes (Hart, 2013)
Biological and technical replicates

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

**Technical replicate**: addresses the measurement error of the assay.
Experimental design

Why increasing the number of biological replicates?

- Technical variability $\Rightarrow$ inconsistent detection of exons at low levels of coverage ($<5$ reads per nucleotide) (McIntyre et al. 2011)
- Doing technical replication may be important in studies where low abundant mRNAs are the focus.

- 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)
Replication: quality rather than quantity

*NATURE REVIEWS GENETICS | PERSPECTIVES | OPINION*

**ARTICLE SERIES:** Applications of next-generation sequencing

**The role of replicates for error mitigation in next-generation sequencing**

Kimberly Robasky, Nathan E. Lewis & George M. Church

**Affiliations | Corresponding author**

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, 2013), RNaseqPower (Hart, 2013)
Experimental Design: Scotty

It’s a balance: cost, precision $\iff$ nb bio. replicates, sequencing depth.

An example output from the Scotty application.

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).
Experimental Design: key points

- The **scientific question** of interest drives the experimental choices

Collect informations before planning
All skills are needed to discussions right from project construction
Sequencing and other technical biases potentially increase the required sample size and sequencing depth
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.
Normalisation and Differential analysis

RNA-sequencing

Random fragmentation

Reverse transcription

mRNAs from a sample → Fragmented mRNAs → cDNAs

# of reads mapped to
Gene 1 Gene 2 ... Gene m
25  320  ...  23

A vector of counts

counting

mapping

from Gene 19
from Gene 23 ...

from Gene 56

ATTGCC...
GCTAAC...
...
AGCCTC...

A list of reads

PCR amplification & sequencing

Adapted from Li et al. (2011)
Isoform detection and quantification

RADIANT
Rapid Development and Distribution of Statistical Tools for High-Throughput Sequencing Data

From E. Bernard

ECCB'14 RADIANT Workshop
Analysis of differential isoform usage by RNA-seq: statistical methodologies and open software, Strasbourg.

Latest News
RADIANT workshop at ECCB'14...read more
Differential Analysis
Identification of differentially expressed genes (DE)

A gene is declared differentially expressed (DE) between two conditions if the observed difference is statistically significant, i.e., more than only due to natural random variation.

- Statistical tools are necessary to take this decision.
- The main steps are: experimental design, normalisation and differential analysis, multiple testing.
Fold Change approach and ideal cut-off values

Cut-off values for gene expression fold change when performing RNA seq

I would like to know what the general consensus is regarding cut-off values for gene expression fold changes (is it mainly >2 up and down-regulated?). Also, is this cut-off applied together with the cut-off for p-value which is p<0.05?

I think the general consensus is > and < than 2-fold, however, we should all justify our rationale for using 2-fold. In our specific case, a difference

> for most gene expression change, people always use fold change 2 as a cutoff for microarray or qPCR. As for RNAseq, since the method is much more sensitive, I guess it must lose some specificity, so I think it may need a higher cutoff number than 2.
Fold Change approach and ideal cut-off values

\[ FC_i = \frac{x_i}{y_i}. \]

<table>
<thead>
<tr>
<th>Gene</th>
<th>CondA1</th>
<th>CondA2</th>
<th>CondB1</th>
<th>CondB2</th>
<th>FC</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene1</td>
<td>5.00</td>
<td>7.00</td>
<td>2.00</td>
<td>2.00</td>
<td>3.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Gene2</td>
<td>800.00</td>
<td>1000.00</td>
<td>350.00</td>
<td>250.00</td>
<td>3.00</td>
<td>0.03</td>
</tr>
<tr>
<td>Gene3</td>
<td>700.00</td>
<td>1100.00</td>
<td>350.00</td>
<td>250.00</td>
<td>3.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Gene4</td>
<td>500.00</td>
<td>1300.00</td>
<td>550.00</td>
<td>50.00</td>
<td>3.00</td>
<td>0.33</td>
</tr>
</tbody>
</table>

FC does not take the variance of the samples into account. Problematic since variability in gene expression is partially gene-specific.
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 due to PCR-amplification step in library preparation (Pickrell et al. 2010, Risso et al. 2011)
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?

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

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)
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.

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

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
So the Winner is ...?

In most cases
The methods yield similar results

However ...
Differences appear based on data characteristics

<table>
<thead>
<tr>
<th>Method</th>
<th>Distribution</th>
<th>Intra-Variance</th>
<th>Housekeeping</th>
<th>Clustering</th>
<th>False-positive rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TC</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UQ</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>Med</td>
<td>++</td>
<td>++</td>
<td>-</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>DESeq</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>TMM</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>FQ</td>
<td>++</td>
<td>-</td>
<td>+</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>RPKM</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
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

- 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.
Normalization : key points

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

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 ($\mu_{ijk}$).
- Property: $\text{Var}(Y_{ijk}) = \text{Mean}(Y_{ijk}) = \mu_{ijk}$
Mean-Variance Relationship

Technical replicates

\[
\text{Variance} \quad \text{Mean}
\]


Biological replicates

\[
\text{Variance} \quad \text{Mean}
\]

data from Parikh et al. Genome Bio 2010

From D. Robinson and D. McCarthy
Overdispersion in RNA-seq data

Counts from biological replicates tend to have variance exceeding the mean (= overdispersion relative to the Poisson distribution). Poisson describes only technical variation.

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 $\phi$ to model the variance

Poisson vs Negative Binomial Models
Model assumptions

- **Poisson** describes technical variation:

  \[ Y_{ij} \sim \text{Pois}(M_j \cdot \lambda_{ij}) \]

  \[ \text{mean}(Y_{ij}) = \text{variance}(Y_{ij}) = M_j \cdot \lambda_{ij} \]

- **Negative binomial** models **biological** variability using the dispersion parameter \( \varphi \):

  \[ Y_{ij} \sim \text{NB}(\mu_{ij} = M_j \cdot \lambda_{ij}, \varphi_i) \]

- Same mean, variance is quadratic in the mean:

  \[ \text{variance}(Y_{ij}) = \mu_{ij} (1 + \mu_{ij} \varphi_i) \]

Critical parameter to estimate: **dispersion**
**φ estimation : crucial!**

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

Some proposed solutions

<table>
<thead>
<tr>
<th>Method</th>
<th>Variance</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESeq</td>
<td>$\mu(1 + \phi \mu \mu)$</td>
<td>Anders et Huber (2010)</td>
</tr>
<tr>
<td>edgeR</td>
<td>$\mu(1 + \phi \mu)$</td>
<td>Robinson et Smyth (2009)</td>
</tr>
<tr>
<td>NBPseq</td>
<td>$\mu(1 + \phi \mu^{\alpha-1})$</td>
<td>Di et al. (2011)</td>
</tr>
</tbody>
</table>

- **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** : $\phi$ and $\alpha$ estimated by LM based on all the genes.
A lot of statistical methods...still developed

Normalized Count Data

Parametric Model
- BaySeq, DESeq, EBSseq, edgeR, NBSeq, ShrinkSeq, TSPM

Non Parametric Model
- NOISeq, NOISeqBIO, SAMseq

Normalized Data

Estimation
- Wilcoxon's statistic
- Gaussian Kernel
- Empirical distribution « noise »

Differential Analysis

Classical Test
- DESeq, edgeR, NBSeq, TSPM

Bayesian
- Bayseq, EBSseq, edgeR, NBSeq, ShrinkSeq

Differential Analysis between two conditions
Comparaison of differential analysis methods
Soneson et Delorenzi (2013)

Evaluation of 11 methods on both simulated and real data.

- Obs 1 : The number of replicates matters! (Differently for different methods)
- Obs 2 : Results are more accurate and less variable between methods if DE genes are regulated in both directions.
- Obs 3 : Outlier counts affect different methods in different ways
- The dispersion estimation method matters!
Comparaison of differential analysis methods
Soneson et Delorenzi (2013)

- Small number of replicates (2-3) or low expression $\rightarrow$ be careful!!
- Large number of replicates (10 or so) or very high expression $\rightarrow$ method choice does not matter much.
- Removing genes with outlier counts or using non-parametric methods reduce the sensitivity to outliers
- Allow tagwise dispersion values
- Normalization methods have problems when all DE genes are regulated in one direction. Iterative approaches like TCC improve performance
Comparaison of differential analysis methods
Rapaport et al. (2013)

Evaluation 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 PCA and Gene Set Enrichissement Analysis)
- Detection of outliers

- Improves power
- Only one command line
Outliers

Challenge: **edgeR** can be sensitive to outliers

![Graph showing sensitivity of edgeR to outliers](image)

**Li and Tibshirani, 2011**

A) Type I error rate at $p_{nom} < 0.05$, $B_0$

- **no outliers**

B) Type I error rate at $p_{nom} < 0.05$, $S_0$

- **presence of outliers**

**Soneson and Delorenzi, 2013**
Why is robustness needed?

Random split of dataset: $n_1=5; \ n_2=5 \rightarrow$ Very little true differential expression

Results driven by outliers

<table>
<thead>
<tr>
<th></th>
<th>NA19222</th>
<th>NA12287</th>
<th>NA19172</th>
<th>NA11881</th>
<th>NA18871</th>
<th>NA12872</th>
<th>NA18916</th>
<th>NA18856</th>
<th>NA19193</th>
<th>NA19140</th>
</tr>
</thead>
<tbody>
<tr>
<td>4004</td>
<td>0.0</td>
<td>1.9</td>
<td>178.1</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2538</td>
<td>2.0</td>
<td>0.6</td>
<td>235.5</td>
<td>6.8</td>
<td>60.2</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.5</td>
<td>1.3</td>
</tr>
<tr>
<td>4962</td>
<td>3.5</td>
<td>0.6</td>
<td>429.5</td>
<td>1.0</td>
<td>35.9</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>4.7</td>
</tr>
<tr>
<td>7921</td>
<td>1.0</td>
<td>5.1</td>
<td>78.9</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>6115</td>
<td>0.0</td>
<td>1.3</td>
<td>0.0</td>
<td>1.9</td>
<td>0.0</td>
<td>0.5</td>
<td>46.1</td>
<td>0.0</td>
<td>100.1</td>
<td>1.3</td>
</tr>
<tr>
<td>5156</td>
<td>13.8</td>
<td>1.3</td>
<td>30.7</td>
<td>0.0</td>
<td>7.1</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.3</td>
</tr>
<tr>
<td>2527</td>
<td>23.7</td>
<td>111.0</td>
<td>228.8</td>
<td>77.0</td>
<td>129.5</td>
<td>10.0</td>
<td>45.3</td>
<td>27.4</td>
<td>26.3</td>
<td>19.1</td>
</tr>
<tr>
<td>1115</td>
<td>2.0</td>
<td>15.2</td>
<td>1074.8</td>
<td>19.5</td>
<td>13.2</td>
<td>10.0</td>
<td>29.6</td>
<td>0.0</td>
<td>1.3</td>
<td>5.5</td>
</tr>
<tr>
<td>3175</td>
<td>3.0</td>
<td>6.3</td>
<td>181.0</td>
<td>7.8</td>
<td>7.6</td>
<td>0.0</td>
<td>5.5</td>
<td>3.0</td>
<td>3.1</td>
<td>2.5</td>
</tr>
<tr>
<td>7951</td>
<td>1.0</td>
<td>12.1</td>
<td>35.9</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>7631</td>
<td>0.0</td>
<td>1.9</td>
<td>0.4</td>
<td>1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>29.6</td>
<td>0.0</td>
<td>24.4</td>
<td>5.5</td>
</tr>
<tr>
<td>3437</td>
<td>24.6</td>
<td>31.1</td>
<td>167.0</td>
<td>4.9</td>
<td>21.2</td>
<td>4.5</td>
<td>8.3</td>
<td>10.1</td>
<td>8.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

CPMs (counts per million)

logFC  logCPM  LR   PValue  FDR

4004  -10.413038  4.186203  30.07924  4.147469e-08  0.0002239513
2538  -5.942865  4.963086  29.60406  5.299369e-08  0.0002239513
4962  -6.387682  5.576979  26.06085  3.308237e-06  0.0064003595
7921  -5.803793  3.183079  22.51927  2.080466e-06  0.0043960241
6115  5.746084  3.921353  21.37010  3.786299e-06  0.0211672325
5156  0.000000  0.000000  0.000000  0.0000000000  0.0000000000
2527  -6.387682  5.576979  26.06085  3.308237e-06  0.0064003595
1115  -4.573655  2.512035  20.13483  7.217026e-06  0.0101663841
3175  -2.154480  6.128702  18.44343  1.750229e-05  0.0211327628
7951  -4.575394  6.873996  18.14127  2.051076e-05  0.0211672325
7631  4.311717  2.683367  17.57990  2.754846e-05  0.0211672325
3437  -3.014484  4.821100  17.05690  3.627624e-05  0.0255505626

Transcriptome genetics using second generation sequencing in a Caucasian population

Stephen B. Montgomery1,2, Micha Sammeth3, Maria Gutierrez-Arceiul3, Radoslaw P. Lach1, Catherine Ingle1, James Nisbett1, Roderic Guigo3 & Emmanuel T. Dermitzakis1,2

Nature, 2010
Current policies (robustness)

- **edgeR**: one option: moderate dispersion less towards trend
  Allows dispersions to be driven more by the data

- **DESeq**: take the maximum of the fit (trended) or the feature-specific dispersion
  Very robust, but many genes pay a penalty, less powerful.

- **DESeq2**: calculate Cook’s distance and filter genes with outliers
  Can inadvertently filter interesting genes

Goal (Robinson and co.): achieve a middle ground between protection against outliers while maintaining high power with observation weights
Robinson’s simulations

- Robust edgeR suffers a tiny bit in power with no outliers, but has good capacity to dampen their effect if present.
- DESeq’s policy on outliers has a global effect, resulting in (sometimes drastic) drop in power.
- DESeq2 is very powerful in the absence of outliers, but policy to filter outliers results in loss of power.
- edgeR and edgeR robust are a bit liberal (5% FDR might mean 6% or 7%).
Multiple Testing

False positive (FP) (**type I error** : $\alpha$) : 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) using a statistical test (calcul of a score)

**Pb** :

Let assume all the G genes are not DE. Each test is realized at $\alpha$ level

Ex : $G = 10000$ genes and $\alpha = 0.05 \rightarrow E(FP) = 500$ genes.
The Family Wise Error Rate (FWER)

Definition

\[ FWER = P(FP > 0) \]

Probability of having at least one false positive, of declaring DE at least one non DE gene.

The Bonferroni procedure

Either each test is realized at \( \alpha = \alpha^*/G \) level or use of adjusted p-value

\[ p_{Bonf_i} = \min(1, p_i \times G) \]

and FWER \( \leq \alpha^* \).

For \( G = 2000 \), \( \leq \alpha^* = 0.05 \), \( \alpha = 2.510^{-5} \).

Easy but conservative and not powerful.

When the number of tests increases, the FWER \( \rightarrow 1 \) with constant FP.
The procedure of Benjamini-Hochberg (95) is one of the procedures which controls the False Discovery Rate FDR = \( E(\frac{FP}{P}) \) si \( P > 0 \).
False Discovery Rate (FDR) Benjamini et Hochberg (95)

\[
\text{FDR} = \frac{E(FP)}{P} \quad \text{si } P > 0 \\
= 1 \quad \text{sinon}
\]

Idea : Do not control the error rate ut the proportion of error ⇒ less conservative than control of the FWER.

Prop

\[
\text{FDR} \leq \text{FWER}
\]
FDR

Principle: The number of declared positive elements $P$ is given by the greater $i p(i) \leq i \alpha^* / G$.

**Prop**

In case of independent tests, $FDR \leq (G_0 / G) \alpha^* \leq \alpha^*$

**Prop**

FDR Benjamini-Hochberg: $\pi_0 = \frac{G_0}{G} = 1$
Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

Controlling the FWER :
Having a great confidence on the DE elements (strong control). Accepting to not detect some elements (lack of power $\Leftrightarrow$ a few DE elements)

Controlling the FDR :
Accepting a proportion of FP among DE elements. Very interesting in exploratory study.
Interpretation - Statistical and practical significance

- Practical significance (importance) and statistical significance (detectability) have little to do with each other.
- An effect can be important, but undetectable (statistically insignificant) because the data are few, irrelevant, or of poor quality.
- An effect can be statistically significant (detectable) even if it is small and unimportant, if the data are many and of high quality.
Differential Analysis: key points

- Methods dedicated to microarrays are not applicable to RNA-seq.
- Small number of replicates (2-3) or low expression → be careful!!
- Large number of replicates (10 or so) or very high expression → method choice does not matter much.
- Removing genes with outlier counts or using non-parametric methods reduce the sensitivity to outliers.
- Don’t forget to correct for multiple testing!

Adapt the method to your data (nb of rep.)
Specific methods developed 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 more expressed genes

GOSeq (Young et al. 2011)

Filter or not

Rau et al. 2013; Huber et al.
General conclusions

Practical conclusions

- Need to collaborate between biologists, bioinformaticians and statisticians
- and in an 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
Conclusions

From Kumamaru, 2012
Aknowledgements

StatOmique (in particular M.-A. Dillies, A. Rau, C. Hennequet-Antier)

PEPI IBIS (INRA) Pôle planification expérimentale et RNA-Seq

David Robinson, Charlotte Soneson
References


Li J, Witten DM, Johnstone IM, Tisbhirani R (2011) *Normalization, testing, and false discovery rate estimation for RNA-sequencing data*, *Biostatistics*, 1-16


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


Robinson MD and Smyth, GK. (2008) *Small-sample estimation of negative binomial dispersion, with applications to SAGE data*


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 $i$
**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 $\Rightarrow$ 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)

$$\frac{X_{ij}}{N_j \times L_i} \times 10^3 \times 10^6 \quad (1)$$

- 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
\[ M_i = \log_2 \left( \frac{Y_{ik}/N_k}{Y_{ik}'/N_k'} \right) \] 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

\[ \hat{s}_j = \text{median}_i \left( \frac{k_{ij}}{\left( \prod_{v=1}^{m} k_{iv} \right)^{1/m}} \right) \]

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

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

\[
\mu = E(X) = cNL = Var(X)
\]

- \( X \) measured number of reads in a library mapping a specific transcript, Poisson r.v.
- \( c \) proportionality 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).
Negative Binomial Models

\[ \mu_{ijk} = \lambda_{ij}m_{jk} \text{ where } m_{jk} : \text{size factor (library size)} \]

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 \( \Rightarrow \) Creates approximately identically distributed pseudodata
- Estimation of \( \phi_i \) 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 $\mu_{ijk}$ is the mean, and $\sigma_{ijk}$ is the variance.

2. The mean $\mu_{ijk}$ is the product of a condition-dependent per-gene value $\lambda_{ij}$ and a size factor (library size) $m_{jk}$:

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

3. Variance decomposition: The variance $\sigma_{ijk}$ is the sum of a shot noise term and a raw variance term:

   $$\sigma_{ijk} = \mu_{ijk} + \alpha_i \mu^2$$

   where $\alpha_i$ is the dispersion value.

4. Per-gene dispersion $\alpha_i$ or pooled $\alpha$ 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 $m_{jk}$ (normalization factors) (see normalization part)

2. For each experimental condition $j$, $n$ expression strength parameters $\lambda_{ij}$ estimated by average of counts from the replicates for each condition, transformed to the common scale:

$$\hat{\lambda}_{ij} = \frac{1}{r_j} \sum_k \frac{y_{ijk}}{\hat{m}_{jk}}$$

3. The smooth functions $f_j$ for each condition $j$ to model dependence of $\alpha_i$ on the expected mean $\lambda_{ij}$: 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 (·)
- Conservative estimates: genes are assigned the maximum of the fitted and empirical values of $\alpha_i$ (sharingMode = "maximum")
- Local fit regression (as described in paper) is no longer the default
Each column = independent biological replicate
First commands

Installation des packages :

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

Chargement des packages :

```r
library(DESeq)
library(edgeR)
```
edgeR main commands

generate raw counts from NB, create list object

```r
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

```r
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

```r
topTags(result)
summary(decideTestsDGE(result), p.value=0.05)
```
DESeq main commands

```r
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...
- cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf