Some statistical methods for RNA-seq data analysis

Elsa Bernard

Institut Curie U900 / Mines ParisTech

Juin 2015

* Slides inspired from Marine Jeanmougin, Julie Aubert, Laurent Jacob, Simon Anders, Michael Love and Peter N. Robinson
RNA-seq: many applications

1. **Gene/exon quantification** or **Estimation of transcript expression**
   - need for normalization
   - previous to differential expression analysis

Reference genome sequence

Detection of (novel) **alternative splicing** isoforms

3. **Fusion genes** identification
RNA-seq: many applications

1. **Gene/exon quantification** or **Estimation of transcript expression**
   - need for normalization
   - previous to differential expression analysis

2. Detection of (novel) **alternative splicing** isoforms

3. **Fusion genes** identification
RNA-seq: many applications

1. **Gene/exon quantification** or **Estimation of transcript expression**
   - need for normalization
   - previous to differential expression analysis

2. Detection of (novel) **alternative splicing** isoforms

3. **Fusion genes** identification
This morning

1. **Gene/exon quantification** or **Estimation of transcript expression**
   - need for **normalization**
   - previous to **differential expression analysis**

2. Detection of (novel) **alternative splicing** isoforms

3. **Fusion genes** identification
RNA-seq pipeline

**Pre-processing**

- Reads alignment

**Quantification / Discovery**

- Statistical analysis

**Biological validation**

- Interpretation

Alignment with:

- samp1.bam
- samp2.bam
- ...

Condition A

Condition B
RNA-seq pipeline

Pre-processing

 Reads alignment

 Quantification / Discovery

 Statistical analysis

 Biological validation Interpretation

Quality control
- total number of reads
- number of reads per barcode
- platform-specific quality scores

Removal of poor quality reads
RNA-seq pipeline

Pre-processing

Reads alignment

Quantification / Discovery

Statistical analysis

Biological validation Interpretation

TopHat / BWA ...
- sample1.bam, sample2.bam ....

Alignment quality
- alignment score
- uniquely mapped reads
RNA-seq pipeline

Pre-processing

Reads alignment

Quantification / Discovery

Counting reads
- HTseq
Gene quantification
- normalization
Transcript estimation
Alternative splicing
Fusion genes

Statistical analysis

Biological validation
Interpretation
RNA-seq pipeline

**Pre-processing**

- Reads alignment

**Quantification / Discovery**

- Statistical analysis

**Biological validation Interpretation**

**Exploratory data analysis**
- PCA, clustering ...

**Differential analysis**

**Pathway analysis**
- ...

**Condition A**

**Condition B**
RNA-seq data

- Matrix of **counts** (non-negative integer values)
- Each column: one experimental unit (sample)
- Each row: one variable (gene, exon)

**Pasilla data**

Study of the transcriptomic effect of RNAi knockdown on the Pasilla gene in *Drosophila melanogaster*

```r
> require(pasilla)
> data("pasillaGenes")
> head(counts(pasillaGenes))
```

<table>
<thead>
<tr>
<th></th>
<th>treated1fb</th>
<th>treated2fb</th>
<th>treated3fb</th>
<th>untreated1fb</th>
<th>untreated2fb</th>
<th>untreated3fb</th>
<th>untreated4fb</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn0000003</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FBgn0000008</td>
<td>118</td>
<td>139</td>
<td>77</td>
<td>89</td>
<td>142</td>
<td>84</td>
<td>76</td>
</tr>
<tr>
<td>FBgn0000014</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FBgn0000015</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>FBgn0000017</td>
<td>4852</td>
<td>4853</td>
<td>3710</td>
<td>4640</td>
<td>7754</td>
<td>4026</td>
<td>3425</td>
</tr>
<tr>
<td>FBgn0000018</td>
<td>572</td>
<td>497</td>
<td>322</td>
<td>552</td>
<td>663</td>
<td>272</td>
<td>321</td>
</tr>
</tbody>
</table>
Outline

1. Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2. Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
1. **Normalization approaches**
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2. **Differential expression**
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
Normalization

An **essential step** in the analysis of gene expression:

- to compare gene expressions from a same sample
- to compare genes from different samples (**differential analysis**)

**Definition**

Normalization is a process designed to **identify and correct technical biases** removing the least possible biological signal.

- **batch effects** (library prep, sequencing technology, ...)

**Goals**

- accurate estimation of gene expression levels
- reliable differential expression analysis

*Normalization has a great impact on DE results!* (Bullard et al 2010, Dillies et al 2012)
Normalization

An **essential step** in the analysis of gene expression:

- to compare gene expressions from a same sample
- to compare genes from different samples (**differential analysis**)

**Definition**

Normalization is a process designed to **identify and correct technical biases** removing the least possible biological signal.

- **batch effects** (library prep, sequencing technology, ...)

**Goals**

- accurate estimation of gene expression levels
- reliable differential expression analysis

*Normalization has a great impact on DE results!* (Bullard et al 2010, Dillies et al 2012)
## Sources of variability

### Within-sample
- Gene length
- Nucleotide composition (GC content)

### Between-sample
- Library size (number of mapped reads)
- Batch effects

---

A lot of different normalization methods...

- Some are part of models for DE, others are ’stand-alone’
- They do not rely on similar hypotheses
## Sources of variability

### Within-sample
- Gene length
- Nucleotide composition (GC content)

### Between-sample
- Library size (number of mapped reads)
- Batch effects

### A lot of different normalization methods...
- Some are part of models for DE, others are ’stand-alone’
- They do not rely on similar hypotheses
Notations

- $k_{ij}$: number of reads for gene $i$ in sample $j$ (observed)
- $L_i$: length of gene $i$
- $q_{ij}$: expression level of gene $i$ in sample $j$ (quantity of interest, unobserved)
- $N_j$: library size of sample $j$
- $s_j$: scaling factor associated with sample $j$
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
At the same expression level, a long gene will have more reads than a shorter one!

\[ k_{ij} \propto L_i q_{ij} \]
The higher sequencing depth, the higher counts!

\[ k_{ij} \propto N_j q_{ij} \]
A very intuitive approach to try to correct for length + depth biases

**RPKM (Reads per Kilo base per Million mapped reads)**

Normalization for RNA length and for library size:

\[
\text{RPKM}_{ij} = \frac{10^9 \times k_{ij}}{N_j \times L_i},
\]

where:

- \(k_{ij}\): number of reads for gene \(i\) in sample \(j\)
- \(N_j\): library size for sample \(j\) (in millions)
- \(L_i\): length of gene \(i\) in base pair

Mortazavi, A. et al. (2008) *Nature Methods*
**RPKM calculation**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Length (bases)</th>
<th>RPKM Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>600</td>
<td>$\frac{12}{0.6\times6} = 3.33$</td>
</tr>
<tr>
<td>B</td>
<td>1100</td>
<td>$\frac{24}{1.1\times6} = 3.64$</td>
</tr>
<tr>
<td>C</td>
<td>1400</td>
<td>$\frac{11}{1.4\times6} = 1.31$</td>
</tr>
</tbody>
</table>

**Sample 1**
- **Gene A**: C = 12
- **Gene B**: C = 24
- **Gene C**: C = 11

**Sample 2**
- **Gene A**: C = 19
- **Gene B**: C = 28
- **Gene C**: C = 16

**RPKM Calculation**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Length (bases)</th>
<th>RPKM Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>600</td>
<td>$\frac{19}{0.6\times8} = 3.96$</td>
</tr>
<tr>
<td>B</td>
<td>1100</td>
<td>$\frac{28}{1.1\times8} = 1.94$</td>
</tr>
<tr>
<td>C</td>
<td>1400</td>
<td>$\frac{16}{1.4\times8} = 1.43$</td>
</tr>
</tbody>
</table>

**Figure**: RPKM calculation
Outline

1 Normalization approaches
   ▪ Within-sample biases
   ▪ Between-sample biases
   ▪ Comparison of normalization methods

2 Differential expression
   ▪ Introduction to differential analysis
   ▪ Fisher’s exact test
   ▪ The poisson model and its limitations
   ▪ Negative Binomial alternative
Zebrafish data analysis from Risso et al., 2014.
Green: control samples. Purple: treated samples.
RLE: relative log expression (comparable samples should have similar RLE distributions centered around 0)

Risso, D. et al. (2014) Nature Biotech
Between-sample normalization

1. global scaling factor (using one sample)
   \[ \mathbb{E} K_{ij} = s_j q_{ij} \]
   - \( \hat{s}_j \)
   - Total number of reads: TC (Marioni et al. 2008)
   - Upper Quartile: UQ (Bullard et al. 2010)

2. global scaling factor (using several samples)
   - more robust
   - Anders and Huber 2010 - Package DESeq
   - Trimmed Mean of M-values TMM - Package edgeR

3. additive effects (regression-based)
   - estimate technical effects with control genes
   - Remove Unwanted Variation - Package RUVseq
Between-sample normalization

1. **global scaling factor (using one sample)**
   - \[
     E K_{ij} = s_j q_{ij}
   \]
   - \(\hat{s}_j??\)
   - Total number of reads: TC (Marioni et al. 2008)
   - Upper Quartile: UQ (Bullard et al. 2010)

2. **global scaling factor (using several samples)**
   - more robust
   - Anders and Huber 2010 - Package DESeq
   - Trimmed Mean of M-values TMM - Package edgeR

3. **additive effects (regression-based)**
   - estimate technical effects with control genes
   - Remove Unwanted Variation - Package RUVseq
Between-sample normalization

1. **global scaling factor (using one sample)**
   - $\mathbb{E}K_{ij} = s_j q_{ij}$
   - $\hat{s}_j$ ??
   - Total number of reads: TC (Marioni et al. 2008)
   - Upper Quartile: UQ (Bullard et al. 2010)

2. **global scaling factor (using several samples)**
   - more robust
   - Anders and Huber 2010 - Package DESeq
   - Trimmed Mean of M-values TMM - Package edgeR

3. **additive effects (regression-based)**
   - estimate technical effects with control genes
   - Remove Unwanted Variation - Package RUVseq
Between-sample normalization

1. **Global scaling factor (using one sample)**
   - $E K_{ij} = s_j q_{ij}$
   - $\hat{s}_j$ ??
   - Total number of reads: TC (Marioni et al. 2008)
   - Upper Quartile: UQ (Bullard et al. 2010)

2. **Global scaling factor (using several samples)**
   - More robust
   - Anders and Huber 2010 - Package DESeq
   - Trimmed Mean of M-values TMM - Package edgeR

3. **Additive effects (regression-based)**
   - Estimate technical effects with control genes
   - Remove Unwanted Variation - Package RUVseq
Between-sample normalization

1. **global scaling factor (using one sample)**
   - \( EK_{ij} = s_j q_{ij} \)
   - \( \hat{s}_j \)
   - Total number of reads: TC (Marioni et al. 2008)
   - Upper Quartile: UQ (Bullard et al. 2010)

2. **global scaling factor (using several samples)**
   - more robust
   - Anders and Huber 2010 - Package DESeq
   - Trimmed Mean of M-values TMM - Package edgeR

3. **additive effects (regression-based)**
   - estimate technical effects with control genes
   - Remove Unwanted Variation - Package RUVseq
Global scaling - I

global scaling factor (using one sample)

- $\mathbb{E}K_{ij} = s_j q_{ij}$

- $\hat{s}_j$

1. Total number of reads TC $\longrightarrow \hat{S}_j = \frac{N_j}{\frac{1}{n} \sum_i N_i}$
   - intuitive but total read count is strongly dependent on a few highly expressed transcripts

2. Upper Quartile UQ $\longrightarrow \hat{S}_j = \frac{Q3_j}{\frac{1}{n} \sum_i Q3_i}$ with Q3 the 75-th quantile.
   - calculate Q3 after exclusion of genes with no read count
   - more robust to highly express genes
Global scaling - I

```r
> dim(counts(pasillaGenes))
14470 7

> # Upper Quartile normalization
> sc = apply(counts(pasillaGenes), 2,
    FUN=function(x) quantile(x[x!=0],probs=3/4))
> scaling.factor = sc / mean(sc)

> print(scaling.factor)
  treated1fb  treated2fb  treated3fb
1.3120821 0.7722063 0.8825215
  untreated1fb  untreated2fb  untreated3fb  untreated4fb
1.0195798 1.4925979 0.7320917 0.7889207

> counts.normalized = t(t(counts(pasillaGenes))/scaling.factor)
```
Global scaling - II

global scaling factor (using several samples)

\[ \mathbb{E}K_{ij} = s_j q_{ij} \]

\[ \hat{s}_j \]

1. DESeq (Anders and Huber 2010)
2. Trimmed Mean of M-values TMM (Robinson et al. 2010)

Motivation

- A few highly differentially expressed genes have a strong influence on read count
  -\[ \Rightarrow \]
  - highly differentially expressed genes may distort the ratio of total reads
  -\[ \Rightarrow \]
  - the total number of read is not a reasonable choice for \( s_j \)
- Aim: minimizing effect of such genes

Assumption

A majority of transcripts is not differentially expressed
General idea

Let us consider two replicated samples, indexed with $j = 1$ and $j = 2$. Given that the samples are replicates we expect the ratio of counts to be the "same" for all genes:

- $\forall i$, $\frac{k_{i1}}{k_{i2}}$ should be the same
- of course not exactly constant! but narrow distribution around its mode
- $\hat{s} = \text{median}_i \frac{k_{i1}}{k_{i2}}$: a good estimate of the sequencing depth ratio
- if $j = 1$ and $j = 2$ are not replicates the median should still be a good estimate as long as few genes are DE.

→ Need to be generalized to more than 2 samples:

- need to compare all samples to a same reference
- definition of a fictive "reference sample" against which to compare everything:

$$k_{ij}^{\text{ref}} = \left( \prod_{j=1}^{m} k_{ij} \right)^{1/m}$$
DESeq (Anders and Huber 2010)

General idea

Let us consider two **replicated samples**, indexed with $j = 1$ and $j = 2$. Given that the samples are replicates we expect the ratio of counts to be the "same" for all genes:

- $\forall i$, $\frac{k_{i1}}{k_{i2}}$ should be the same
- of course not exactly constant! but narrow distribution around its mode
- $\hat{s} = \text{median}_i \frac{k_{i1}}{k_{i2}}$: a good estimate of the sequencing depth ratio
- if $j = 1$ and $j = 2$ are not replicates the median should still be a good estimate as long as few genes are DE.

$\Rightarrow$ Need to be **generalized to more than 2 samples**:

- need to compare all samples to a same *reference*
- definition of a fictive "*reference sample*" against which to compare everything:

$$k_{i}^{\text{ref}} = \left( \prod_{j=1}^{m} k_{ij} \right)^{1/m}$$
Generalization

Calculation of the scaling factor:

\[ \hat{s}_j = \text{median}_i \frac{k_{ij}}{k_{i}^{\text{ref}}} \]

where:

- \( k_{ij} \): number of reads in sample \( j \) assigned to gene \( i \)
- denominator: reference sample created from geometric mean across samples

R package **DESeq**:

- `estimateSizeFactors()` : estimate the size factors for a "CountDataSet" object
> require(DESeq)

> # estimate the size factors:
> pasillaGenes <- estimateSizeFactors( pasillaGenes )

> print( sizeFactors(pasillaGenes) )

treated1fb  treated2fb  treated3fb  
  1.5116926  0.7843521  0.8958321  
untreated1fb  untreated2fb  untreated3fb  untreated4fb  
  1.0499961  1.6585559  0.7117763  0.7837458 

> # understand what happen!
> # calculate the gene-wise geometric means
> geomeans <- exp( rowMeans( log( counts(pasillaGenes) ) ) )

> # Plot a histogram of the ratios
> # ratio of sample 1 over the reference
> hist(log2( counts(pasillaGenes)[,1] / geomeans ), breaks=100)
> abline(v=log2( sizeFactors(pasillaGenes)[ j ] ), col="red")
> # Plot a histogram of the ratios
> # ratio of sample 1 over the reference
> hist(log2( counts(pasillaGenes)[,1] / geomeans ), breaks=100)
> abline(v=log2( sizeFactors(pasillaGenes)[ j ] ), col="red")
Remove unwanted variation (RUV)

**additive effect (regression-based)**

- uses **control genes** (housekeeping genes, spike-in) to estimate technical noise
- estimate a **gene-specific nuisance effect**

1. RUVSeq (Risso et al. 2014). R package `RUVSeq`

### Framework

- **factor of interest** $x_j$ (e.g. outcome) for sample $j$, its effect $\beta_i$ on gene $i$.
- **unwanted factor** $w_j$ (e.g. batch) for sample $j$, its effect $\alpha_i$ on gene $j$.
- $\log k_{ij} = x_j \beta_i + w_j \alpha_i + \epsilon_{ij}$
- **control genes** are not affected by the **factor of interest** $X$
  - $\beta_c = 0$ for control gene $c$. Hence $\log k_c = W \alpha_c + \epsilon_c$.
  - estimate $\hat{W}$ by PCA.
    - plug back $\hat{W}$ in the model and do a regression to get $\hat{\beta}$ and $\hat{\alpha}$
  - remove $\hat{W} \hat{\alpha}$ from $\log k$
Remove unwanted variation (RUV)

**additive effect (regression-based)**

- uses **control genes** (housekeeping genes, spike-in) to estimate technical noise
- estimate a **gene-specific nuisance effect**

1. RUVSeq (Risso et al. 2014). R package **RUVSeq**

**Framework**

- **factor of interest** $x_j$ (e.g. outcome) for sample $j$, its effect $\beta_i$ on gene $i$.
- **unwanted factor** $w_j$ (e.g. batch) for sample $j$, its effect $\alpha_i$ on gene $j$.
- $\log k_{ij} = x_j \beta_i + w_j \alpha_i + \epsilon_{ij}$
- **control genes** are not affected by the **factor of interest** $X$
  1. $\beta_c = 0$ for control gene $c$. Hence $\log k_c = W \alpha_c + \epsilon_c$.
  2. estimate $\hat{W}$ by PCA.
     plug back $\hat{W}$ in the model and do a regression to get $\hat{\beta}$ and $\hat{\alpha}$
  3. remove $\hat{W} \hat{\alpha}$ from $\log k$
RUVM result

**Zebrafish data analysis from Risso et al., 2014.**

Green: control samples. Purple: treated samples.

RLE: relative log expression (comparable samples should have similar RLE distributions centered around 0)

Risso, D. *et al.* (2014) *Nature Biotech*
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
Which method should you use for normalization of RNA-Seq data?

- How to choose a normalization adapted to your experiment?
- What is the impact of the normalization step on the downstream analysis?
A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis


Submitted: 12th April 2012; Received (in revised form): 29th June 2012
An effective normalization should result in a *stabilization of read counts* across samples

**Figure**: Effects of normalization on *E*. histolytica data.

**Results**

- most of the methods yield comparable results
- RPKM and TC that do not improve over the raw counts (sensitive to high count genes)
Effect of normalization on housekeeping genes

**Method**

Assumption: housekeeping genes are *similarly expressed across samples*

- 30 housekeeping genes selected from a list previously described in Eisenberg et Levanon (2003)
- average the coefficient of variation of housekeeping genes

![Graph showing variation in expression among a set of housekeeping genes](image)

**Results**

DESeq and TMM normalization methods lead to smallest coefficient of variation
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>
Conclusions on normalization

- RNA-seq data are affected by biases (total number of mapped reads per sample, 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)
- TC, RPKM, UQ Adjustment of distributions, implies a similarity between RNA repertoires expressed
- DESeq, TMM More robust ratio of counts using several samples, suppose that the majority of the genes are not DE.
- RUVSeq Powerful when a large set of control genes can be identified
Conclusions on normalization

- RNA-seq data are affected by biases (total number of mapped reads per sample, 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)
- TC, RPKM, UQ Adjustment of distributions, implies a similarity between RNA repertoires expressed
- DESeq, TMM More robust ratio of counts using several samples, suppose that the majority of the genes are not DE.
- RUVSeq Powerful when a large set of control genes can be identified
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
Differential analysis

What is differential gene expression?

A gene is declared differentially expressed (DE) if an observed difference or change in expression between two experimental conditions is statistically significant.  

How to determine the level of significance?

⇝ Statistical tools (hypothesis testing)

⇝ Statistical tools for RNA-seq need to analyze read-count distributions

^greater than expected just due to natural random variation

Often used to compare expression levels in different conditions:

■ Tissue: liver vs. brain

■ Treatment: drugs A, B, and C

■ State: healthy controls vs. patient

■ Across time
Differential analysis

What is differential gene expression?

A gene is declared differentially expressed (DE) if an observed difference or change in expression between two experimental conditions is statistically significant.

How to determine the level of significance?

- Statistical tools (hypothesis testing)
- Statistical tools for RNA-seq need to analyze read-count distributions

\[ a \] greater than expected just due to natural random variation

Often used to compare expression levels in different conditions:

- Tissue: liver vs. brain
- Treatment: drugs A, B, and C
- State: healthy controls vs. patient
- Across time
Hypothesis testing

The key notions are:

1. formulate the **testing hypothesis**: null hypothesis versus alternative
2. **p-value** computation: probability of observing the data given that a hypothesis is true
3. **type I and type II errors**
4. **multiple-testing**: control of the FDR (false discovery rate)
Null hypothesis

Formulate the null hypothesis

⇝ The statement being tested in a test of statistical significance is called the null hypothesis

⇝ The null hypothesis is usually a statement of *no effect* or *no difference*

Example

Let $q_i$ be the expression level of gene $i$. We have access to measurements of $q_i$ in two groups $A$ and $B$. ie we observe $(q_{i1}^A, q_{i2}^A, \ldots)$ and $(q_{i1}^B, q_{i2}^B, \ldots)$ (biological replicates in both groups).

⇝ $H_0$: $q_i^A$ and $q_i^B$ follow the same distribution

⇝ $H_0$: $q_i^A$ and $q_i^B$ have equal mean
Formulate the null hypothesis

The statement being tested in a test of statistical significance is called the null hypothesis.

The null hypothesis is usually a statement of *no effect* or *no difference*.

Example t-test

We suppose:

\[ q_i^A \sim N(\mu_i^A, \sigma) \]
\[ q_i^B \sim N(\mu_i^B, \sigma) \]

\[ H_0: \mu_i^A = \mu_i^B \]
p-value is the probability of an observed (or more extreme) result assuming that the null hypothesis is true

\[ p = P(\text{observation} \mid H_0 \text{ is true}) \]

- p "small" means that \( H_0 \) is likely to be "false"
p-value is the probability of an observed (or more extreme) result assuming that the null hypothesis is true

\[ p = P(\text{observation} \mid H_0 \text{ is true}) \]

- p "small" means that \( H_0 \) is likely to be "false"

**Example t-test**

We suppose:

- \( q_i^A \sim N(\mu_i^A, \sigma) \) and \( q_i^B \sim N(\mu_i^B, \sigma) \)

- t-statistic

\[ t = \frac{\bar{q}_i^A - \bar{q}_i^B}{s / \sqrt{n}} \]

- obtained p-value (the t-statistic follows a Student law under \( H_0 \))

\[ p = P(T \geq t \mid H_0) \]
p-value is the probability of an observed (or more extreme) result assuming that the null hypothesis is true

\[ p = P(\text{observation} \mid H_0 \text{ is true}) \]

- p "small" means that \( H_0 \) is likely to be "false"
"uncorrecting testing" reject $H_0$ if $p \leq \alpha$ (eg, $\alpha = 0.05$)

- Type I error = false positive = FP
- Type II error = false negative = FN
Multiple testing

1 "Uncorrected testing"
   - Gives $P(FP_i) \leq \alpha$ for all $1 \leq i \leq n$
   - Many type I errors (FP)
   - eg: 10000 genes that are not DE. Significance level $\alpha = 0.05$. But $10000 \times 0.05 = 500$ genes will be call DE "by chance".

2 Control of the type I error
   - e.g.: Bonferroni: use per-comparison significance level $\alpha/n$
   - Guarantees $P(FP) \leq \alpha$
   - Very conservative

3 Control of the FDR false discovery rate
   - first defined by Benjamini-Hochberg (BH, 1995, 2000)
   - Guarantees $FDR = \mathbb{E} \left( \frac{FP}{FP+TP} \right) \leq \alpha$
   - finding 100 DE genes with only 2 FP seems better than finding 6 DE genes with 2 FP ...
Multiple testing

1. "Uncorrected testing"
   - Gives $P(FP_i) \leq \alpha$ for all $1 \leq i \leq n$
   - Many type I errors (FP)
   - eg: 10000 genes that are not DE. Significance level $\alpha = 0.05$. But $10000 \times 0.05 = 500$ genes will be called DE "by chance".

2. Control of the type I error
   - e.g.: Bonferroni: use per-comparison significance level $\alpha / n$
   - Guarantees $P(FP) \leq \alpha$
   - Very conservative

3. Control of the FDR false discovery rate
   - first defined by Benjamini-Hochberg (BH, 1995, 2000)
   - Guarantees $FDR = \mathbb{E}\left(\frac{FP}{FP+TP}\right) \leq \alpha$
   - finding 100 DE genes with only 2 FP seems better than finding 6 DE genes with 2 FP ...
Multiple testing

1. "Uncorrected testing"
   - Gives $P(FP_i) \leq \alpha$ for all $1 \leq i \leq n$
   - Many type I errors (FP)
   - eg: 10000 genes that are not DE. Significance level $\alpha = 0.05$. But $10000 \times 0.05 = 500$ genes will be call DE "by chance".

2. Control of the type I error
   - e.g.: Bonferroni: use per-comparison significance level $\alpha/n$
   - Guarantees $P(FP) \leq \alpha$
   - Very conservative

3. Control of the FDR false discovery rate
   - first defined by Benjamini-Hochberg (BH, 1995, 2000)
   - Guarantees $FDR = \mathbb{E}\left(\frac{FP}{FP+TP}\right) \leq \alpha$
   - finding 100 DE genes with only 2 FP seems better than finding 6 DE genes with 2 FP ...
Multiple testing

1. "Uncorrected testing"
   - Gives $P(FP_i) \leq \alpha$ for all $1 \leq i \leq n$
   - Many type I errors (FP)
   - eg: 10000 genes that are not DE. Significance level $\alpha = 0.05$. But $10000 \times 0.05 = 500$ genes will be call DE "by chance".

2. Control of the type I error
   - e.g.: Bonferroni: use per-comparison significance level $\alpha/n$
   - Guarantees $P(FP) \leq \alpha$
   - Very conservative

3. Control of the FDR false discovery rate
   - first defined by Benjamini-Hochberg (BH, 1995, 2000)
   - Guarantees $FDR = \mathbb{E} \left( \frac{FP}{FP + TP} \right) \leq \alpha$
   - finding 100 DE genes with only 2 FP seems better than finding 6 DE genes with 2 FP ...
## Strategy

Differential expression gene-by-gene:

**For each gene \( i \), is there a significant difference in expression between the condition 1 and condition 2?**

- Statistical model (definition and parameter estimation)
- Testing for differential expression:

\[ H_{0i} : \mu_{i1} = \mu_{i2} \]

## State of the art

- An abundant literature
  - Fisher’s exact test
  - Poisson model
  - Negative Binomial model (\texttt{DESeq}, \texttt{edgeR})
- Comparison of methods (Pachter et al. 2011, Kvam and Liu 2012, Soneson and Delorenzi 2013)
1 Normalization approaches
- Within-sample biases
- Between-sample biases
- Comparison of normalization methods

2 Differential expression
- Introduction to differential analysis
- Fisher’s exact test
- The poisson model and its limitations
- Negative Binomial alternative
Fisher’s exact test

can be used for RNA-seq without replicates, on a gene-by-gene basis, organizing the data in a 2 x 2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>condition 1</th>
<th>condition 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene i</td>
<td>$x_{i1}$</td>
<td>$x_{i2}$</td>
<td>$x_i.$</td>
</tr>
<tr>
<td>Remaininggenes</td>
<td>$\sum_{g \neq i} x_{g1}$</td>
<td>$\sum_{g \neq i} x_{g2}$</td>
<td>$\sum_{g \neq i} x_{g.}$</td>
</tr>
<tr>
<td>Total</td>
<td>$x_{.1}$</td>
<td>$x_{.2}$</td>
<td>$x_{..}$</td>
</tr>
</tbody>
</table>

Null hypothesis

The proportion of counts for some gene $i$ amongst two samples is the same as that of the remaining genes:

$H_{0i} : \frac{\pi_{i1}}{\pi_{i2}} = \frac{\pi_{g1}}{\pi_{g2}}$

where $\pi_{i1}$ is the true (unknown) proportion of counts in sample 1

$\Rightarrow$ we can calculate the p-value $p = P(\text{readcount} \geq x_{i1} | H_0)$ exactly using the hypergeometric law (one or two-sided Fisher exact test)
Fisher’s exact test

- can be used for RNA-seq without replicates, on a gene-by-gene basis, organizing the data in a 2 x 2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>condition 1</th>
<th>condition 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene i</td>
<td>$x_{i1}$</td>
<td>$x_{i2}$</td>
<td>$x_i.$</td>
</tr>
<tr>
<td>Remaining genes</td>
<td>$\sum_{g \neq i} x_g$</td>
<td>$\sum_{g \neq i} x_g$</td>
<td>$\sum_{g \neq i} x_g.$</td>
</tr>
<tr>
<td>Total</td>
<td>$x_{.1}$</td>
<td>$x_{.2}$</td>
<td>$x_{..}$</td>
</tr>
</tbody>
</table>

Null hypothesis

The proportion of counts for some gene $i$ amongst two samples is the same as that of the remaining genes:

$$H_{0i} : \frac{\pi_{i1}}{\pi_{i2}} = \frac{\pi_{g1}}{\pi_{g2}}$$

where $\pi_{i1}$ is the true (unknown) proportion of counts in sample 1

〜 we can calculate the p-value $p = P(\text{readcount} \geq x_{i1}|H_0)$ exactly using the hypergeometric law (one or two-sided Fisher exact test)
Fisher’s exact test

> countTable

<table>
<thead>
<tr>
<th></th>
<th>condition 1</th>
<th>condition 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>216</td>
<td>160</td>
<td>376</td>
</tr>
<tr>
<td>Remaining genes</td>
<td>28,351,805</td>
<td>21,934,509</td>
<td>50,286,314</td>
</tr>
<tr>
<td>Total</td>
<td>28,352,021</td>
<td>21,934,669</td>
<td>50,286,690</td>
</tr>
</tbody>
</table>

> fisher.test(countTable)
Fisher’s Exact Test for Count Data
data: countTable
p-value = 0.7159
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.847 1.289
sample estimates:
odds ratio
  1.04

~~ if test for many genes, need to adjust p-value for multiple-testing!~~
Fisher’s exact test

Need for replicates!

Without replication:

- **complete lack of knowledge about biological variation.**
- **no sound statistical basis** for inference of differences between the groups.

Tarazona, S. *et al.* (2011) *Genome Research*

"We propose a novel methodology for the assessment of differentially expressed features, NOISeq, that empirically models the noise in count data, is reasonably robust against the choice of SD, and can work in the absence of replication."
Fisher’s exact test

Need for replicates!

knock-down sample T2
versus
control sample U3

count sample U2
versus
control sample U3

red: significant genes according to Fisher test (at 10% FDR)

Figure: Fly cell culture, knock-down of pasilla (from Simon Anders)
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
Poisson model – Intuition

Need to model:

- non-negative integer values (count data)

From the Binomial law to the Poisson distribution:

- e.g., a series of $n = 10$ coin flips, each of which has a probability of $p = 5$ of heads
- The binomial distribution gives us the probability of observing $k$ heads

\[
p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}\]

Event: An RNAseq read “lands” in a given gene (success) or not (failure)

- $B(n, p)$ converges to $\mathcal{P}(\lambda = np)$ when $N >> p$

Marioni, J. et al. (2008) Genome Research
The number of reads that are mapped into a gene was first modeled using a Poisson distribution

\[ P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \]

with \( \lambda > 0 \).

⇒ only one parameter is needed to determine the probability of an event

- Poisson distribution naturally appears for count data
- It assumes that mean and variance are the same:
  \[ \lambda = E(X) = Var(X) \]
- no need to estimate the variance (convenient!)
Poisson (mean = variance)

For $X \sim \text{Poisson}(\lambda)$, both the mean and the variance are equal to $\lambda$.

\[ P(X=k) = \frac{e^{-\lambda} \lambda^k}{k!} \]

\[ \lambda = 1, 3, 6, 9 \]

Poisson distribution
Limitations of the poisson model

The variance grows faster than the mean in RNAseq data.

\[
\text{Poisson: } v \sim \mu^1 \\
\text{NB: } v \sim \mu^2
\]

Data: Nagalakshmi et al. Science 2008

Figure: Mean count vs variance of RNA seq data. Orange line: the fitted observed curve. Purple: the variance implied by the Poisson distribution.

Limitations of the poisson model

The variance grows faster than the mean in RNAseq data.

For Poisson-distributed data, the variance is equal to the mean. No need to estimate the variance. This is convenient.

E.g. Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010), ...

\[
\text{Poisson: } v \sim \mu^1 \\
\text{NB: } v \sim \mu^2
\]

Data: Nagalakshmi et al. Science 2008

Figure: Mean count vs variance of RNA seq data. Orange line: the fitted observed curve. Purple: the variance implied by the Poisson distribution.

Overdispersion in RNA-seq data! \( \sim \) counts from biological replicates tend to have variance exceeding the mean.
What is the impact of overdispersion?

⇒ overdispersion ⇒ underestimation of the biological variance

Let us consider this question using the t-distribution:

\[ t_i = \frac{\bar{x}_i^{(1)} - \bar{x}_i^{(2)}}{S \sqrt{\frac{1}{n}}}, \]

where

- \( S \) is the sample standard deviation,
- \( n \) is the sample size

⇒ Underestimation of the variance ⇒ overestimation of \( t_i \)

Figure: Empirical cumulative distribution functions (ECDFs) of p-values. No genes are truly differentially expressed, and the ECDF curves (blue) should remain below the diagonal (gray)

⇒ Overdispersion will lead to an increased type I error rate (probability to falsely declare a gene DE)
What is the impact of overdispersion?

\[ t_i = \frac{\bar{x}^{(1)}_i - \bar{x}^{(2)}_i}{\frac{S}{\sqrt{n}}} , \]

where
- \( S \) is the sample standard deviation,
- \( n \) is the sample size

\[ \Rightarrow \text{Underestimation of the variance} \Rightarrow \text{overestimation of } t_i \]

\[ \Rightarrow \text{Overdispersion will lead to an increased type I error rate} \ (\text{probability to falsely declare a gene DE}) \]

**Figure:** Empirical cumulative distribution functions (ECDFs) of *p*-values. No genes are truly differentially expressed, and the ECDF curves (blue) should remain below the diagonal (gray).
1 Normalization approaches
   - Within-sample biases
   - Between-sample biases
   - Comparison of normalization methods

2 Differential expression
   - Introduction to differential analysis
   - Fisher’s exact test
   - The poisson model and its limitations
   - Negative Binomial alternative
### Alternative approaches

#### Parametric approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>baySeq</td>
<td>NB</td>
<td>Hardcastle TJ and Kelly KA (2010)</td>
</tr>
<tr>
<td>EBSeq</td>
<td>NB</td>
<td>Leng N (2012)</td>
</tr>
<tr>
<td>ShrinkSeq</td>
<td>NB (zero-inflated)</td>
<td>Van de Wiel MA et al. (2012)</td>
</tr>
<tr>
<td>edgeR</td>
<td>NB</td>
<td>Robinson MD et al. (2010)</td>
</tr>
<tr>
<td>DESeq</td>
<td>NB</td>
<td>Anders S and Huber W (2010)</td>
</tr>
<tr>
<td>NBPSq</td>
<td>over-parameterized NB</td>
<td>Di Y et al. (2011)</td>
</tr>
<tr>
<td>TSPM</td>
<td>poisson</td>
<td>Auer PL and Doerge RW (2011)</td>
</tr>
</tbody>
</table>

#### Non-parametric strategies

- NOISeq (Tarazona S et al. 2011)
- SAMseq (Li J and Tibshirani R 2011)

#### Transformation-based methods

→ aim to find a transformation for counts to analyze them by traditional methods

- voom + limma
- vst + limma
The negative binomial distribution can be used as an alternative to the Poisson distribution:

\[ X_{ij} \sim NB(\mu_{ij}, \phi_i) \]

where:

- \( E(X_{ij}) = \mu_{ij} \)
- \( Var(X_{ij}) = \mu_{ij} + \phi_i \mu_{ij}^2 \)
- \( \phi_i \) is the dispersion parameter

The variance is always larger than the mean for the negative binomial \( \Rightarrow \) suitable for RNA-seq data
Negative binomial distribution

\[ NB(20, 0.25) \]

\[ NB(10, 0.25) \]

\[ NB(20, 0.5) \]

\[ NB(10, 0.5) \]
Many genes, few biological samples - difficult to estimate $\phi$ on a gene-by-gene basis

Some proposed solutions:

<table>
<thead>
<tr>
<th>Method</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESeq</td>
<td>$\mu(1 + \phi \mu \mu)$</td>
</tr>
<tr>
<td>edgeR</td>
<td>$\mu(1 + \phi \mu)$</td>
</tr>
<tr>
<td>NBPseq</td>
<td>$\mu(1 + \phi \mu^{\alpha-1})$</td>
</tr>
</tbody>
</table>

**DESeq**

data-driven relationship of variance and mean estimated using local regression for robust fit across genes

**edgeR**

Borrow information across genes for stable estimates of $\phi$. 3 ways to estimate $\phi$: common, trended, tagwise (moderated)

**NBPSseq**

NBP includes two parameters $\phi$ and $\alpha$, estimated from all the genes
Some practical considerations

- Data must be input as **raw counts** (and not RPKM or FPKM values): normalization offsets are included in the model.

- Each column should be an **independent biological replicate**

- Multi-factor designs now included

- Check out the DESeq Users' Guide for examples
  

- **Version matters**!
DESeq: Main commands

Step 1: creation of a CountDataSet object

```r
> head(countTable)
   untreated3 untreated4 treated2 treated3
FBgn00000003 0  0  0  1
FBgn00000008 76 70  88 70
FBgn0000014  0  0  0  0
FBgn0000015  1  2  0  0

> condition
[1] untreated untreated treated treated treated
Levels: treated untreated

# We can now instantiate a CountDataSet (central data structure in the DESeq package)
> cds = newCountDataSet( countTable, condition )
```
### Step 2: Normalization

```r
# estimate the effective library size
> cds <- estimateSizeFactors(cds)

> sizeFactors(cds)
  treated2fb  treated3fb untreated3fb untreated4fb
  1.297       1.042      0.818      0.911

# If we divide each column of the count table by the size factor for this column, the count values are brought to a common scale

> head(counts(cds, normalized=TRUE))

<table>
<thead>
<tr>
<th></th>
<th>untreated3</th>
<th>untreated4</th>
<th>treated2</th>
<th>treated3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn00000003</td>
<td>0.00</td>
<td>0.00</td>
<td>0.0</td>
<td>0.897</td>
</tr>
<tr>
<td>FBgn00000008</td>
<td>87.05</td>
<td>69.27</td>
<td>86.1</td>
<td>62.803</td>
</tr>
<tr>
<td>FBgn00000014</td>
<td>0.00</td>
<td>0.00</td>
<td>0.0</td>
<td>0.000</td>
</tr>
<tr>
<td>FBgn00000015</td>
<td>1.15</td>
<td>1.98</td>
<td>0.0</td>
<td>0.000</td>
</tr>
</tbody>
</table>
```
DESeq: Main commands

Step 3: Differential analysis

> cds <- estimateDispersions(cds)
# estimates a dispersion value for each gene
# fits a curve through the estimates
# assigns to each gene a dispersion value
   (choice between the per-gene estimate and the fitted value)

> plotDispEsts(cds) # estimates against the mean normalized counts

> res <- nbinomTest( cds, "untreated", "treated" )
DESeq: Main commands

Results:

```r
> head(res)
   id  baseMean baseMeanA baseMeanB foldChange log2FoldChange  pval  padj
FBgn0000008  91.78    93.05    90.51   0.973      -0.0399 1.000 1.000
FBgn0000014   1.93     0.00     3.85      Inf       Inf    0.378 0.913
FBgn0000017 3995.15  4340.18 3650.11   0.841      -0.2498 0.378 0.845
FBgn0000018  344.22  342.43  346.01   1.010       0.0150 0.896 1.000
FBgn0000024   5.65     4.09     7.21    1.763       0.8180 0.525 0.972
FBgn0000032 1025.52 1038.25 1012.79   0.975      -0.0358 0.801 1.000
```

> plot(res$baseMean, res$log2FoldChange, log="x", pch=20, cex=.3,
  + col = ifelse( res$padj < .1, "red", "black"))

multiple testing correction: here genes are called DE if adjusted p-value are below 10% FDR
**DESeq**: Main commands

## Results:

```r
> head(res)

   id  baseMean baseMeanA baseMeanB  foldChange log2FoldChange  pval    padj
FBgn0000008  91.78    93.05    90.51    0.973    -0.0399  1.0000  1.0000
FBgn0000014   1.93     0.00     3.85   Inf        Inf    0.3780  0.913
FBgn0000017  3995.15  4340.18  3650.11  0.841   -0.2498  0.3780  0.2760
FBgn0000018  344.22   342.43   346.01  1.010    0.0150  0.8960  1.0000
FBgn0000024   5.65     4.09     7.21   1.763    0.8180  0.5250  0.9720
FBgn0000032 1025.52  1038.25  1012.79  0.975   -0.0358  0.8010  1.0000
```

```r
> > plot(res$baseMean, res$log2FoldChange, log="x", pch=20, cex=.3,
+     col = ifelse( res$padj < .1, "red", "black"))
```

**multiple testing correction**: here genes are called DE if adjusted p-value are below 10% FDR
What’s next?

What happens after a differential analysis?

**Further analysis**

- Test for enriched functional categories (i.e., do differentially expressed genes tend to share the same function?)
- Clustering of genes (i.e., co-expression analysis)
- Inference of gene networks
- Integration with other data (epigenomic, metabolomic, proteomic, ...)

**Biological validation**

- Gene knock-down experiments
- qPCR validation
Thank you!

* Slides inspired from Marine Jeanmougin, Julie Aubert, Laurent Jacob, Simon Anders, Michael Love and Peter N. Robinson