Comparative analysis of RNA-Seq data with DESeq2

Simon Anders
EMBL Heidelberg
Two applications of RNA-Seq

Discovery
- find new transcripts
- find transcript boundaries
- find splice junctions

Comparison
Given samples from different experimental conditions, find effects of the treatment on
- gene expression strengths
- isoform abundance ratios, splice patterns, transcript boundaries
# Sequencing count data

<table>
<thead>
<tr>
<th>Gene</th>
<th>control-1</th>
<th>control-2</th>
<th>control-3</th>
<th>treated-1</th>
<th>treated-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBgn00000008</td>
<td>78</td>
<td>46</td>
<td>43</td>
<td>47</td>
<td>89</td>
</tr>
<tr>
<td>FBgn00000014</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FBgn00000015</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FBgn00000017</td>
<td>3187</td>
<td>1672</td>
<td>1859</td>
<td>2445</td>
<td>4615</td>
</tr>
<tr>
<td>FBgn00000018</td>
<td>369</td>
<td>150</td>
<td>176</td>
<td>288</td>
<td>383</td>
</tr>
</tbody>
</table>

[...]
Counting rules

• Count reads, not base-pairs
• Count each read at most once.
• Discard a read if
  • it cannot be uniquely mapped
  • its alignment overlaps with several genes
  • the alignment quality score is bad
  • (for paired-end reads) the mates do not map to the same gene
Why we discard non-unique alignments

gene A

gene B

treatment condition

control condition
Normalization for library size

• If sample A has been sampled deeper than sample B, we expect counts to be higher.

• Naive approach: Divide by the total number of reads per sample

• Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.
Normalization for library size

actual expression

sequenced reads

naively normalized
Normalization for library size

Histogram of log2(sample2/sample1)
Normalization for library size

To compare more than two samples:

• Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples
• Normalize each sample to this reference, to get one scaling factor (“size factor”) per sample.

Anders and Huber, 2010
similar approach: Robinson and Oshlack, 2010
Counting noise

In RNA-Seq, noise (and hence power) depends on count level.

Why?
The Poisson distribution

• This bag contains very many small balls, 10% of which are red.

• Several experimenters are tasked with determining the percentage of red balls.

• Each of them is permitted to draw 20 balls out of the bag, without looking.
$\frac{7}{100} = 7\%$

$\frac{10}{100} = 10\%$

$\frac{8}{100} = 8\%$

$\frac{11}{100} = 11\%$
**Poisson distribution: Counting uncertainty**

<table>
<thead>
<tr>
<th>expected number of red balls</th>
<th>standard deviation of number of red balls</th>
<th>relative error in estimate for the fraction of red balls</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\sqrt{10} = 3$</td>
<td>$1 / \sqrt{10} = 31.6%$</td>
</tr>
<tr>
<td>100</td>
<td>$\sqrt{100} = 10$</td>
<td>$1 / \sqrt{100} = 10.0%$</td>
</tr>
<tr>
<td>1,000</td>
<td>$\sqrt{1,000} = 32$</td>
<td>$1 / \sqrt{1000} = 3.2%$</td>
</tr>
<tr>
<td>10,000</td>
<td>$\sqrt{10,000} = 100$</td>
<td>$1 / \sqrt{10000} = 1.0%$</td>
</tr>
</tbody>
</table>
The negative binomial distribution

A commonly used generalization of the Poisson distribution with *two* parameters

\[
\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots
\]
Biological sample with mean $\mu$ and variance $\nu$.

Poisson distribution with mean $q$ and variance $q$.

Negative binomial with mean $\mu$ and variance $q+\nu$. 

The NB from a hierarchical model
Testing: Generalized linear models

Two sample groups: treatment and control.

Model:

- Count value $K_{ij}$ for a gene in sample $j$ is generated by NB distribution with mean $s_j \mu_j$ and dispersion $\alpha$.

- The expected expression strength is:
  \[
  \log \mu_j = \beta_{i0} + x_j \beta_{iT}
  \]

  $x_j = 0$ if $j$ is control sample
  $x_j = 1$ if $j$ is treatment sample

Null model:
\[
\beta_{iT} = 0, \text{ i.e., expectation is the same for all samples}
\]

Alternative model:
\[
\beta_{iT} \neq 0, \text{ i.e., expected expression changes from control to treatment, with log fold change (LFC) } \beta_T
\]
Testing: Generalized linear models

\[ K_{ij} \sim \text{NB}(s_j \mu_{ij}, \alpha_i) \]

\[ \log \mu_{ij} = \beta_{i0} + x_j \beta_{iT} \]

\[ x_j = 0 \text{ for if } j \text{ is control sample} \]
\[ x_j = 1 \text{ for if } j \text{ is treatment sample} \]

Calculate the coefficients \( \beta \) that fit best the observed data \( K \).

Is the value for \( \beta_{iT} \) significantly different from null?

Can we reject the null hypothesis that it is merely cause by noise (as given by the dispersion \( \alpha_i \))?

We use a Wald test to get a \( p \) value.
Tasks in comparative RNA-Seq analysis

• Estimate fold-change between control and treatment

• Estimate variability within groups

• Determine significance

the hard part
Dispersion

- Minimum variance of count data:
  \[ v = \mu \quad \text{(Poisson)} \]

- Actual variance:
  \[ v = \mu + \alpha \mu^2 \]

- \( \alpha \): “dispersion” \[ \alpha = (\mu - v) / \mu^2 \]
  (squared coefficient of variation of extra-Poisson variability)
Shrinkage estimation of dispersion (within-group variability)
Shrinkage estimation of dispersion (within-group variability)
Shrinkage estimation of dispersion (within-group variability)
Shrinkage estimation of effect sizes

without shrinkage

with shrinkage
Complex designs

Simple: Comparison between two groups.

More complex:
• paired samples
• testing for interaction effects
• accounting for nuisance covariates
• …
GLMs: Blocking factor

<table>
<thead>
<tr>
<th>Sample</th>
<th>treated</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S2</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S3</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S4</td>
<td>no</td>
<td>female</td>
</tr>
<tr>
<td>S5</td>
<td>no</td>
<td>female</td>
</tr>
<tr>
<td>S6</td>
<td>yes</td>
<td>male</td>
</tr>
<tr>
<td>S7</td>
<td>yes</td>
<td>male</td>
</tr>
<tr>
<td>S8</td>
<td>yes</td>
<td>female</td>
</tr>
<tr>
<td>S9</td>
<td>yes</td>
<td>female</td>
</tr>
<tr>
<td>S10</td>
<td>yes</td>
<td>female</td>
</tr>
</tbody>
</table>
GLMs: Blocking factor

\[ K_{ij} \sim NB(s_j \mu_{ij}, \alpha_{ij}) \]

full model for gene \( i \):

\[ \log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T \]

reduced model for gene \( i \):

\[ \log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S \]
GLMs: Interaction

\[ K_{ij} \sim NB(s_j \mu_{ij}, \alpha_{ij}) \]

full model for gene \( i \):

\[ \log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T + \beta_i^I x_j^S x_j^T \]

reduced model for gene \( i \):

\[ \log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T \]
GLMs: paired designs

- Often, samples are paired (e.g., a tumour and a healthy-tissue sample from the same patient)
- Then, using pair identity as blocking factor improves power.

full model:
\[
\log \mu_{ijl} = \beta_i^0 + \begin{cases} 
0 & \text{for } l = 1(\text{healthy}) \\
\beta_i^T & \text{for } l = 2(\text{tumour}) 
\end{cases}
\]

reduced model:
\[
\log \mu_{ij} = \beta_i^0
\]

i gene
j subject
l tissue state
GLMs: Dual-assay designs

How does the affinity of an RNA-binding protein to mRNA change under some drug treatment?

Prepare control and treated samples (in replicates) and perform on each sample RNA-Seq and CLIP-Seq.

For each sample, we are interested in the ratio of CLIP-Seq to RNA-Seq reads.

How is this ratio affected by treatment?
GLMs: CLIP-Seq/RNA-Seq assay

full model:
  count ~ assayType + treatment + assayType:treatment

reduced model:
  count ~ assayType + treatment
GLMs: CLIP-Seq/RNA-Seq assay

full model:
  \[ \text{count} \sim \text{sample} + \text{assayType} + \text{assayType: treatment} \]

reduced model:
  \[ \text{count} \sim \text{sample} + \text{assayType} \]
Genes and transcripts

• So far, we looked at read counts _per gene_.

A gene’s read count may increase
• because the gene produces _more_ transcripts
• because the gene produces _longer_ transcripts

How to look at gene sub-structure?
Assigning reads to transcripts

100 reads from A
30 reads from B

10 reads from A
30 reads from B
Assigning reads to transcripts

200 reads

(50 from A,
150 from B?)

5 reads

from A

15 reads

from B

total: A: 55 reads
B: 165 reads (accuracy?)
One step back: Differential exon usage

Our tool, DEXSeq, tests for differential usage of exons.

Usage on an exon =

\[
\frac{\text{number of reads mapping to the exon}}{\text{number of reads mapping to any other exon of the same gene}}
\]
Differential exon usage -- Example
Differential exon usage -- Example
Differential usage of exons or of isoforms?

cassette exon with well-understood function
cassette exon with uncharacterized function
Summary

• Estimating fold-changes without estimating variability is pointless.
• Estimating variability from few samples requires information sharing across genes (shrinkage)
• Shrinkage can also regularize fold-change estimates. (New in DESeq2)
Acknowledgements

Co-authors:
• Wolfgang Huber
• Alejandro Reyes
• Mike Love (MPI-MG Berlin)

Thanks also to
• the rest of the Huber group
• all users who provided feedback

Funding:
• EMBL
• European Union: FP7-health Project Radiant
Fisher’s exact test between two samples

Example data: fly cell culture, knock-down of pasilla
(Brooks et al., Genome Res., 2011)

knock-down sample T2
versus
control sample U3
control sample U2
versus
control sample U3

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