Differential gene expression analysis using RNA-seq

Applied Bioinformatics Core, September/October 2018

Friederike Dündar with Luce Skrabanek & Paul Zumbo
DIFFERENTIAL GENE EXPRESSION
Bioinformatics workflow of RNA-seq analysis

1. **Images** (.tif)
2. **Raw reads** (.fastq) → **FASTQC**
3. **Aligned reads** (.sam/.bam) → **RSeQC**
4. **Read count table** (.txt)
5. **Normalized read count table** (.Robj)
6. **List of fold changes & statistical values** (.Robj, .txt)
7. **Downstream analyses on DE genes**

**Base calling & demultiplexing**
- Bustard/RTA/OLB, CASAVA

**Mapping**
- STAR

**Counting**
- HTSeq, featureCounts

**Normalizing**
- DESeq2, edgeR

**Descriptive plots**
Bioinformatics workflow of RNA-seq analysis

Base calling & demultiplexing
Bustard/RTA/OLB, CASAVA

Mapping
STAR

Counting
HTSeq, featureCounts

Normalizing
DESeq2, edgeR

DE test & multiple testing correction
DESeq2, edgeR, limma

Descriptive plots

List of fold changes & statistical values
.Robj, .txt

Downstream analyses on DE genes

Images
.tif

Raw reads
.fastq

Aligned reads
.sam/.bam

Read count table
.txt

Normalized read count table
.Robj
# Read count table

<table>
<thead>
<tr>
<th></th>
<th>SNF2_1</th>
<th>SNF2_2</th>
<th>SNF2_3</th>
<th>SNF2_4</th>
<th>SNF2_5</th>
<th>WT_1</th>
<th>WT_2</th>
<th>WT_3</th>
<th>WT_4</th>
<th>WT_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL012W</td>
<td>7347</td>
<td>7170</td>
<td>7643</td>
<td>8111</td>
<td>5943</td>
<td>4309</td>
<td>3769</td>
<td>3034</td>
<td>5601</td>
<td>4164</td>
</tr>
<tr>
<td>YAL068C</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>YAL067C</td>
<td>103</td>
<td>51</td>
<td>44</td>
<td>90</td>
<td>53</td>
<td>12</td>
<td>23</td>
<td>21</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>YAL066W</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL065C</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>YAL064W-B</td>
<td>13</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

![Dendrogram](image)
DE basics

1. Estimate **magnitude** of DE taking into account differences in sequencing depth, technical, and biological read count variability.

2. Estimate the **significance** of the difference accounting for performing thousands of tests.

**H0**: no difference in the read distribution between two conditions

Garber et al. (2011) Nature Methods, 8(6), 469–477. doi:10.1038/nmeth.1613
Modelling gene expression values with linear models

Y = b₀ + b₁ * x + e

b₀: intercept, i.e. average of the baseline group
b₁: difference between baseline and non-reference group
x : 0 if genotype == “SNF2”, 1 if genotype == “WT”

both betas are estimates! (they’re right on spot because the data is so clear for this example and the model is so simple)

rlog.norm ~ genotype

> coef(lmfit)

## Coefficients:
## (Intercept)  genotype
## 6.666 3.111
Modeling read counts and estimating the log2-fold-change (DESeq2)

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

\[ \mu_{ij} = s_j q_{ij} \]

Once the coefficients are estimated, the significance tests need to test how far away from zero they are since zero would mean “no difference”.

H0: no difference in the read distribution between two conditions

Let’s do this!
From read counts to DE

<table>
<thead>
<tr>
<th></th>
<th>SNF2_1</th>
<th>SNF2_2</th>
<th>SNF2_3</th>
<th>SNF2_4</th>
<th>SNF2_5</th>
<th>WT_1</th>
<th>WT_2</th>
<th>WT_3</th>
<th>WT_4</th>
<th>WT_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL012W</td>
<td>7347</td>
<td>7170</td>
<td>7643</td>
<td>8111</td>
<td>5943</td>
<td>4309</td>
<td>3769</td>
<td>3034</td>
<td>5601</td>
<td>4164</td>
</tr>
<tr>
<td>YAL068C</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>YAL067C</td>
<td>103</td>
<td>51</td>
<td>44</td>
<td>90</td>
<td>53</td>
<td>12</td>
<td>23</td>
<td>21</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>YAL066W</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAL065C</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>YAL064W-B</td>
<td>13</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL012W</td>
<td>5538.0476736</td>
<td>-0.3049727</td>
<td>0.1564379</td>
<td>1.9494807</td>
<td>5.123804e-02</td>
</tr>
<tr>
<td>YAL068C</td>
<td>0.9677468</td>
<td>-0.1306360</td>
<td>0.3922204</td>
<td>-0.3330679</td>
<td>7.390830e-01</td>
</tr>
<tr>
<td>YAL067C</td>
<td>40.8756727</td>
<td>-1.0144579</td>
<td>0.2128597</td>
<td>-4.7658520</td>
<td>1.880572e-06</td>
</tr>
<tr>
<td>YAL066W</td>
<td>0.1403184</td>
<td>-0.1343829</td>
<td>0.1806512</td>
<td>-0.7438804</td>
<td>4.569489e-01</td>
</tr>
<tr>
<td>YAL065C</td>
<td>5.1638597</td>
<td>0.3447455</td>
<td>0.4060259</td>
<td>0.8490726</td>
<td>3.958409e-01</td>
</tr>
<tr>
<td>YAL064W-B</td>
<td>8.4455750</td>
<td>0.1250101</td>
<td>0.3437285</td>
<td>0.3636887</td>
<td>7.160905e-01</td>
</tr>
</tbody>
</table>

average norm. count
standard error estimate for the logFC
What next?

• Do your results make sense?
• Are the results robust?
  • do **multiple tools** agree on the majority of the genes?
  • are **the fold changes** strong enough to explain the phenotype you are seeing?
  • have **other experiments** yielded similar results?
• Downstream analyses: mostly exploratory

**How to decide which tool(s) to use?**
• function/content of original publication
  • code maintained?
  • well documented?
  • used by others?
  • efficient?
RNACocktail tries to implement all (current!) best performers for various RNA-seq analyses


https://bioinform.github.io/rnacocktail/
Where to get help and inspiration?

bioconductor.org/help/workflows

**F100Research Software Tool Articles**

Periodic Table of Bioinformatics:
http://elements.eaglegenomics.com/

mailing lists/github issues of the individual tools

biostars.org
seqanswers.com
stackoverflow.com


supplemental material of publications based on HTS data

**WALK-IN CLINICS**

@ WCM:
Thursdays, 1:30 – 3 pm,
LC-504 (1300 York Ave)

abc.med.cornell.edu

@ MSKCC:
https://www.mskcc.org/research-advantage/core-facilities/bioinformatics

https://github.com/abcdbug/dbug
Everything’s connected…

Sample type & quality
- Low input?
- Degraded?

Experimental design
- Controls
- No. of replicates
- Randomization

Library preparation
- Poly-A enrichment vs. ribo minus
- Strand information

Sequencing
- Read length
- PE vs. SR
- Sequencing errors

Biological question
- Expression quantification
- Alternative splicing
- De novo assembly needed
- mRNAs, small RNAs
- ….

Bioinformatics
- Aligner
- Annotation
- Normalization
- DE analysis strategy