Yeast RNA-seq: differential analysis with 48-replicates

Jacques van Helden

Last update: 2019-02-10

	Parameter value
verbosity	1
parallel	TRUE
cores	4
epsilon	0.01
alpha	0.05
lambda	0.5
compute.all	FALSE
save.results	FALSE
reload.previous.session	TRUE
dir.main	/Users/jvanheld/Documents/enseignement/bioinformatics_courses/statistics_bioinformatics/R-files/RNA-seq/Gierlinski_2015_yeast_Snf2_48replicates
dir.data	/Users/jvanheld/Documents/enseignement/bioinformatics_courses/statistics_bioinformatics/R-files/RNA-seq/Gierlinski_2015_yeast_Snf2_48replicates/data
save.image.file	/Users/jvanheld/Documents/enseignement/bioinformatics_courses/statistics_bioinformatics/R-files/RNA-seq/Gierlinski_2015_yeast_Snf2_48replicates/data/Gierlinski_2015_yeast_Snf2_48replicates.RData

1 Introduction

We explore here different methods for the detection of differentially expressed genes (**DEG**) with an exceptional data set comprised of 48 replicates of 2 yeast strains: wild-type (WT) and Snf2 mutant.

This dataset was published in 2015 by Gierlinski and co-workers (Geoff Barton’s team) and was precisely designed for the comparative evaluation of RNA-seq analysis tools.

This document contains an exploratory analysis, which will serve as basis to write the corresponding chapter of the book Statistics for Bioinformatics. I don’t at all intend to reproduce the detailed methodological work performed by Geoff Barton and his team.

2 Data sources

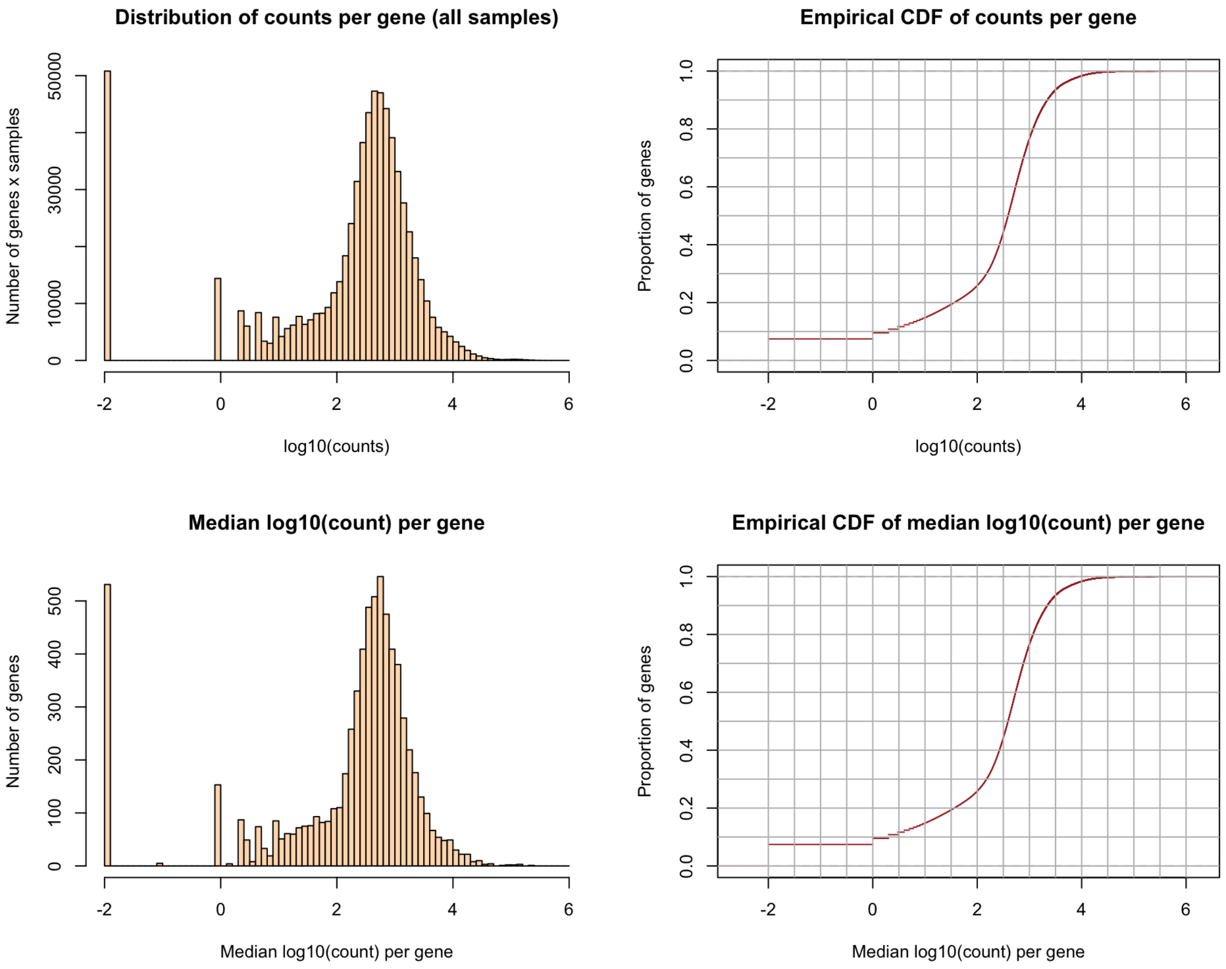
The data was kindly shared by Geoff Barton and Christian Cole on the figshare Web site: https://figshare.com/articles/Metadata_for_a_highly_replicated_two_condition_yeast_RNAseq_experiment_/1416210 (https://figshare.com/articles/Metadata_for_a_highly_replicated_two_condition_yeast_RNAseq_experiment_/1416210)

- The full data sets (raw reads) are available at ENA http://www.ebi.ac.uk/ena/data/view/ERP004763 (http://www.ebi.ac.uk/ena/data/view/ERP004763)
- Count tables (generated with htseq-count) were downloaded from figshare:
  - Wild-type: https://dx.doi.org/10.6084/m9.figshare.1425503 (https://dx.doi.org/10.6084/m9.figshare.1425503)
  - SNF2 knock-out: https://dx.doi.org/10.6084/m9.figshare.1425502 (https://dx.doi.org/10.6084/m9.figshare.1425502)

3 Data laoding

```
[1] "ERP004763_sample_mapping.tsv"
[2] "gene_descriptions.tab"
[3] "gene_ids.txt"
[4] "Gierlinski_2015_yeast_Snf2_48replicates.RData"
[5] "README.html"
[6] "README.md"
[7] "Snf2_raw_counts.tsv.gz"
[8] "WT_raw_counts.tsv.gz"
```

3.1 Read count statistics



Distribution of log10(counts) per gene. Left: histogram of read counts per gene (all WT + Snf2 samples together). Right: empirical cumulative distribution function (eCDF). Top: Number of measures (genes * samples). Bottom: median count per gene.

4 Detection of differentially expressed genes

4.1 DEG detection with DESeq2

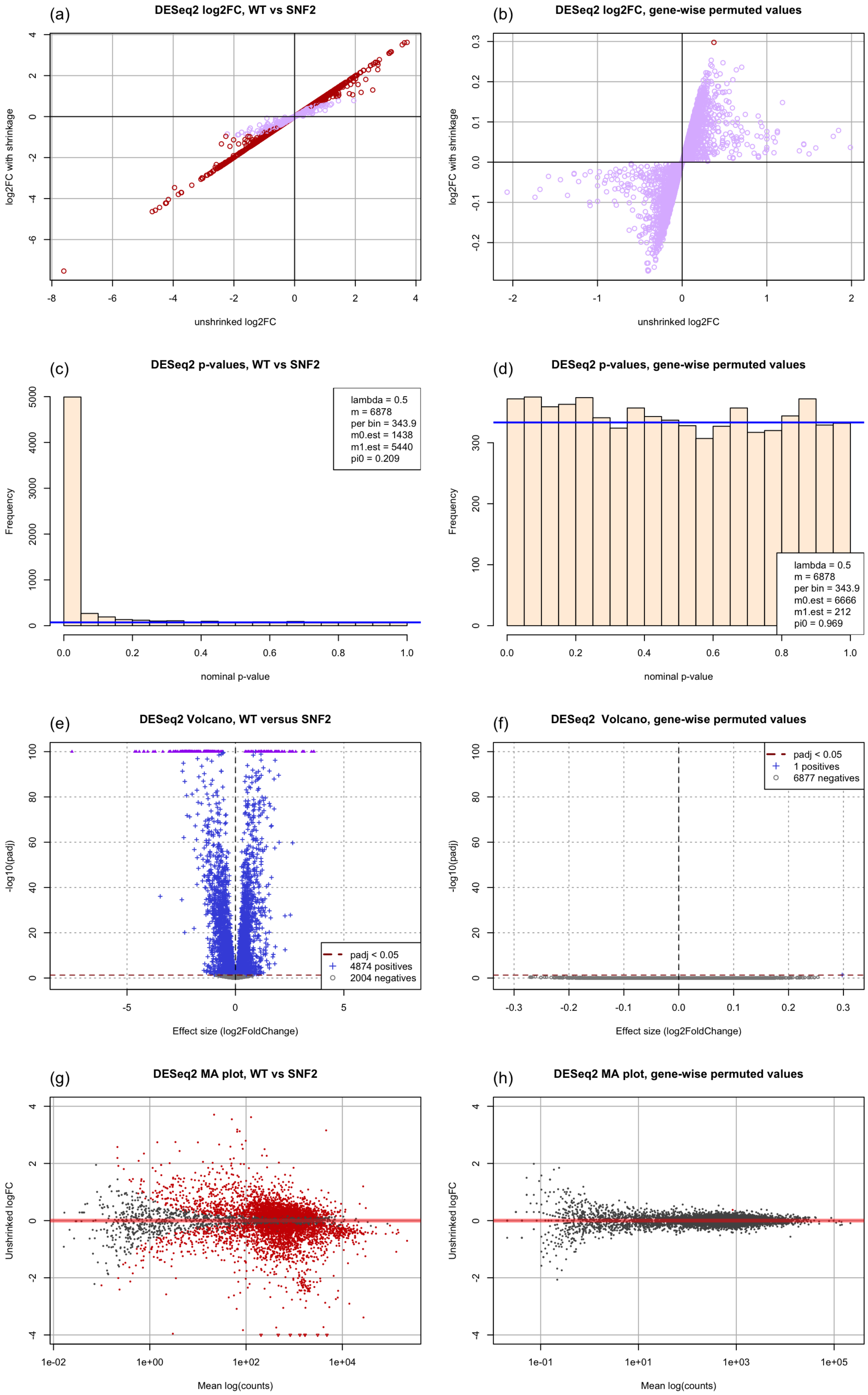
4.2 Negative control: gene-wise permuted dataset

As a negative control, we generate fake expression matrix by permuting the counts of each gene (row) between the different samples (columns). Since the permutation encompasses the 48 replicates of the two strains (WT and SNF2), each group should consist of a mixture between the two groups, so that differential analysis should return no positive gene.

4.3 Dagnostic plots

For each analysis (WT versus SNF2, gene-wise permuted data) we generate a series of diagnostic plots that allow us to grasp the results and compare the observed effect (WT versus SNF2 strains) with the random expectation (gene-wise permuted values).

1. **log2FC plots**. DESeq2 relies on a shrinkage strategy (Love et al., 2014) to estimate the variance of each gene, which impacts the log2FC. The comparison between shrinkage-based and “unshrinked” log2FC gives an ida of the impact of the shrinkage.
2. **P-value histograms**. Under the null hypothesis (no differential expression at all), the p-value distribution should be uniform (by definition of the p-value). P-value histograms of gene-wise permuted values should thus be flat. In contrast, with real gene count data sets, the histogram should present an enrichment of low p-values, corresponding to the supposedly differentially expressed genes. Moreover, Storey and Tibshrani’s method enables to estimate the respective numbers of genes under null (m_0) or alternative (m_1) hypothesis, as well as the proportion of null features (π_0), by comparing the left and right sides of a p-value histogram.
3. **Volcano plots** provide a simultaneous view of the effect size (log2 fold change, in abcsissa) and its significance ($-log_{10}(FC)$, in ordinate). These two criteria are both relevant to detect relevant genes for differential expression.
4. **MA plots** indicate the relationship between the level of regulation ($log(FC)$, in ordinate) and the mean level of expression (mean of the log(counts), in abcsissa).


(a)
DESeq2 log2FC, WT vs SNF2
log2FC with shrinkage
unshrinked log2FC
-8
-6
-4
-2
0
2
4
(b)
DESeq2 log2FC, gene-wise permuted values
log2FC with shrinkage
unshrinked log2FC
-0.2
-0.1
0.0
0.1
0.2
0.3
-2
-1
0
1
2
(c)
DESeq2 p-values, WT vs SNF2
Frequency
nominal p-value
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
lambda = 0.5
m = 6878
per bin = 343.9
m0.est = 1438
m1.est = 5440
pi0 = 0.209
(d)
DESeq2 p-values, gene-wise permuted values
Frequency
nominal p-value
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
lambda = 0.5
m = 6878
per bin = 343.9
m0.est = 6666
m1.est = 212
pi0 = 0.969
(e)
DESeq2 Volcano, WT versus SNF2
-log10(padj)
Effect size (log2FoldChange)
0
20
40
60
80
100
-5
0
5
padj < 0.05
4874 positives
2004 negatives
(f)
DESeq2 Volcano, gene-wise permuted values
-log10(padj)
Effect size (log2FoldChange)
0
20
40
60
80
100
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
padj < 0.05
1 positives
6877 negatives
(g)
DESeq2 MA plot, WT vs SNF2
Unshrinked logFC
Mean log(counts)
-4
-2
0
2
4
1e-02
1e+00
1e+02
1e+04
(h)
DESeq2 MA plot, gene-wise permuted values
Unshrinked logFC
Mean log(counts)
-4
-2
0
2
4
1e-01
1e+01
1e+03
1e+05

MA plots for differential expression analysis based on the 48-replicate dataset from Gierlinski et al. (2015). **Left panels:** WT versus SNF2. **Right panels:** gene-wise permuted values (negative control). (a-b): log2FC with and without shrinkage. (c-d): P-value histograms. (e-f): Volcano plots. (g-h): MA plots with unshrinked logFC.

5 References

1. Gierliński,M., Cole,C., Schofield,P., Schurch,N.J., Sherstnev,A., Singh,V., Wrobel,N., Gharbi,K., Simpson,G., Owen-Hughes,T., et al. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics, 10.1093/bioinformatics/btv425. http://bioinformatics.oxfordjournals.org/content/31/22/3625 (http://bioinformatics.oxfordjournals.org/content/31/22/3625)

2. Schurch,N.J., Schofield,P., Gierliński,M., Cole,C., Sherstnev,A., Singh,V., Wrobel,N., Gharbi,K., Simpson,G.G., Owen-Hughes,T., et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 10.1261/rna.053959.115. http://rnajournal.cshlp.org/content/early/2016/03/30/rna.053959.115 (http://rnajournal.cshlp.org/content/early/2016/03/30/rna.053959.115)

3. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15, 550.