b
UNIVERSITÄT
BERN

Bulk RNA-seq

Differential gene expression analysis

Interfaculty Bioinformatics Unit, University of Bern

03.06.2021

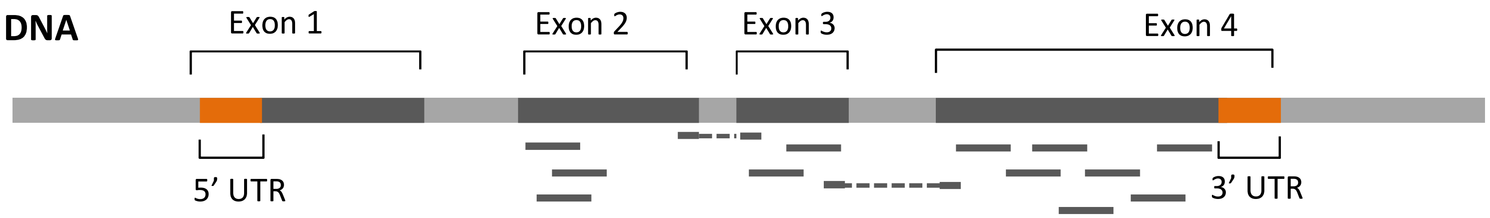
SIB
Swiss Institute of
Bioinformatics

RNA-seq data processing

Step 1: Assess quality and quantity of reads

Step 2: Map reads to reference genome

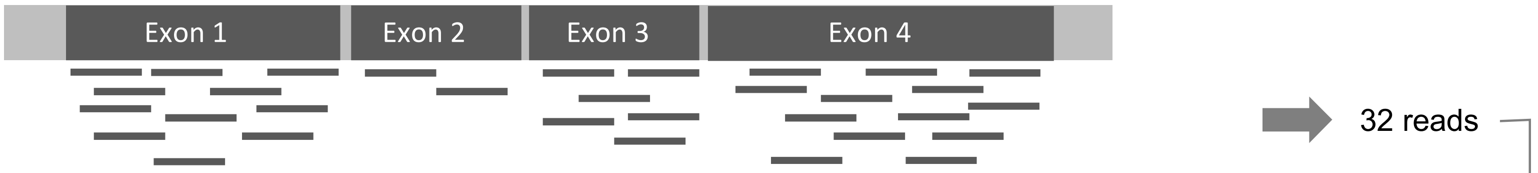
The majority of reads come from mature transcripts which lack introns but we map to the reference genome which contains introns → We use an alignment tool that can handle large gaps (e.g. **Hisat2**)



RNA-seq data processing

Step 3: Count the number of reads mapping to each gene

In each sample, we count how many reads overlap with each genes (using a tool like **featureCounts**). This requires information on where each gene is located in the genome, available for example from Ensembl (http://www.ensembl.org/index.html)



We end up with a table of read counts for each sample and gene:

	Sample C1	Sample C2	Sample C3	Sample T1	Sample T2	Sample T3
Gene1	0	2	1	18	55	32
Gene2	10256	8953	9665	15846	7546	5482

Test for differential gene expression

For each gene, we test for differential expression between 2 experimental groups (in this example C vs T). Each group has to contain biological replicates (in this example 3 samples per group).

*** (?)

	Sample C1	Sample C2	Sample C3	Sample T1	Sample T2	Sample T3
Gene1	32	55	18	0	1	0
Gene2	10256	8953	9665	15846	7546	5482

We use DESeq2 for this task, and the analyses involves the following steps:

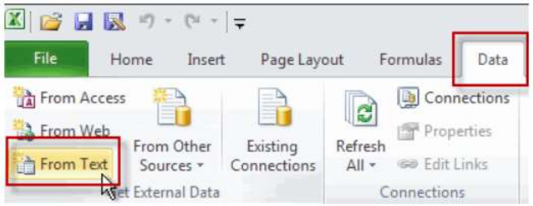
1. **Normalisation**: Correct for differences in the total number of reads between samples
2. **Estimate the variance** between replicates: Because RNA-seq experiments often have relatively few replicates within experimental groups, DESeq2 incorporates information from other genes with similar overall expression level into the estimation.
3. **Adjust log-fold change (LFC)**: This step takes into account the evidence based on which the LFC is estimated. If it is weak (e.g. because the gene is lowly expressed, the variance between replicates is high or we have few replicates), the LFC is shrunk toward zero.
4. Using the adjusted LFC and the variance estimate, we calculate a **test statistic** and compare it to the normal distribution to obtain a **P-value.**
5. **Multiple test correction**: To take into account the fact that we perform many tests (one per gene), DESeq2 applies a false discovery rate correction based on the Benjamini-Hochberg procedure. However, the multiple test correction considers only genes that could potentially be detected as differentially expressed. Only these genes will have an adjusted P-value. The mean read count across all samples is used to decide if a gene should be included or not.

For details, please refer to DESeq2 documentation available at http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html

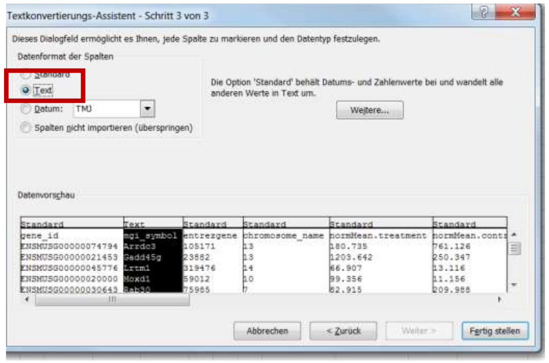
Overview of output files

For each comparison, you obtain a file where the format of the file name is: **Condition1.Condition2.DEResults.original.rlog.txt**

You can easily import these files into Excel:



It is best to select “Text” format for the column containing the gene symbols. There are some rare cases, where Excel will interpret a gene name as a date and convert it!



Textkonvertierungs-Assistent - Schritt 3 von 3

Dieses Dialogfeld ermöglicht es Ihnen, jede Spalte zu markieren und den Datentyp festzulegen.

Datenformat der Spalten

○ Standard
● Text
○ Datum: TMJ
○ Spalten nicht importieren (überspringen)

Die Option 'Standard' behält Datums- und Zahlenwerte bei und wandelt alle anderen Werte in Text um.

Datenvorschau

Standard	Text	Standard	Standard	Standard	Standard
gene_id	mgi_symbol	entrezgene	chromosome_name	normMean.treatment	normMean.contr
ENSMUSG00000074794	Arrdc3	105171	13	180.735	761.126
ENSMUSG00000021453	Gadd45g	23882	13	1203.642	250.347
ENSMUSG00000045776	Lrtm1	319476	14	66.907	13.116
ENSMUSG00000020000	Moxd1	59012	10	99.356	11.156
ENSMUSG00000030643	Rab30	75985	7	82.915	209.988

Output file format

GENE INFO

First couple of columns contain **information on genes**, e.g. various IDs

gene_id = Ensembl ID
symbol = Official gene symbol
entrezgene = Entrez ID

gene_id	symbol	entrezgene
ENSMUSG00000074794	Arrdc3	105171
ENSMUSG00000021453	Gadd45g	23882
ENSMUSG00000035805	Mlc1	170790

COUNTS

This is followed by the **mean normalised number of reads** (counts) in each experimental group,

normMean.expGroup1	normMean.expGroup2
180.735	761.126
1203.642	250.347
0.334	0

and many columns with the **counts** in each sample in the following forms:

A) Header = sampleID → original counts (as in table on slide 3)

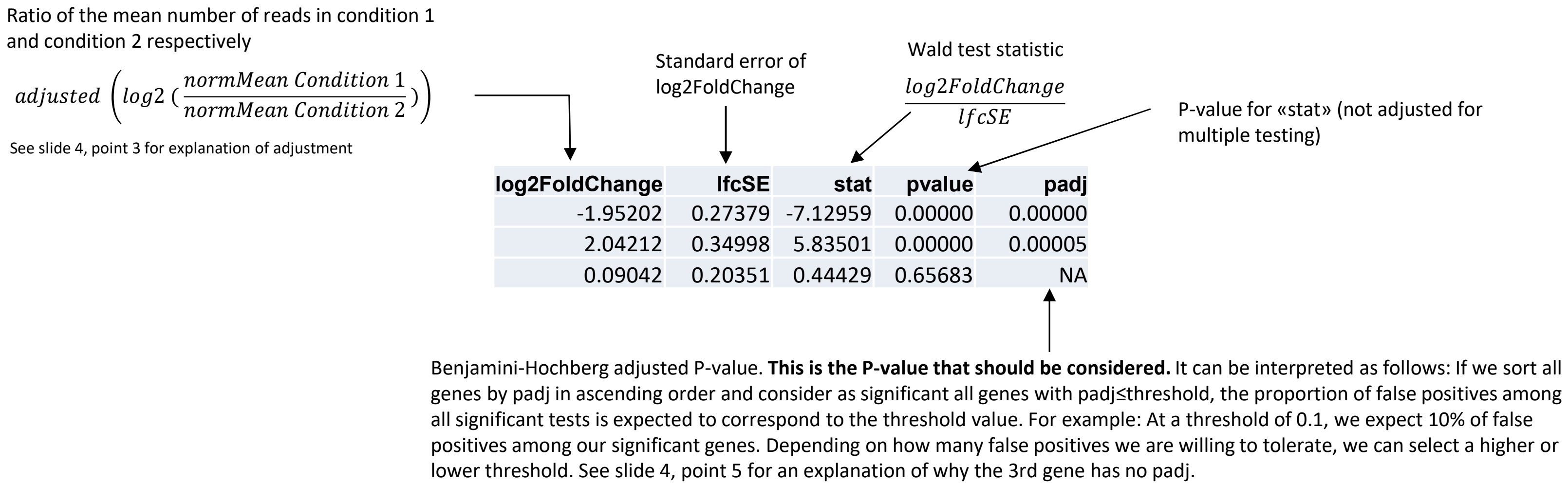
B) sampleID.norm → **normalised counts**. These have been adjusted to account for differences in sequencing depth between samples but NOT for differences in gene length! This means that values can be compared between samples but not between genes. Longer genes will tend to have higher counts.

C) sampleID.rlog → counts after regularized log transformation (see DESeq2 documentation). May be useful e.g. for visualisation.

The normalised counts will typically be the most useful.

Output file format

STATISTICAL TEST RESULTS (DESEQ2)



Ratio of the mean number of reads in condition 1 and condition 2 respectively

$$adjusted\left(log2\left(\frac{normMean\ Condition\ 1}{normMean\ Condition\ 2}\right)\right)$$

See slide 4, point 3 for explanation of adjustment

Standard error of log2FoldChange

Wald test statistic

$$\frac{log2FoldChange}{lfcSE}$$

P-value for «stat» (not adjusted for multiple testing)

log2FoldChange	lfcSE	stat	pvalue	padj
-1.95202	0.27379	-7.12959	0.00000	0.00000
2.04212	0.34998	5.83501	0.00000	0.00005
0.09042	0.20351	0.44429	0.65683	NA

Benjamini-Hochberg adjusted P-value. **This is the P-value that should be considered.** It can be interpreted as follows: If we sort all genes by padj in ascending order and consider as significant all genes with padj≤threshold, the proportion of false positives among all significant tests is expected to correspond to the threshold value. For example: At a threshold of 0.1, we expect 10% of false positives among our significant genes. Depending on how many false positives we are willing to tolerate, we can select a higher or lower threshold. See slide 4, point 5 for an explanation of why the 3rd gene has no padj.