

Bulk RNA-seq

Differential gene expression analysis

Interfaculty Bioinformatics Unit | IBU, University of Bern

12.01.2022

Guidelines for authorship in collaborations

Service Mode

These are usually comprised of straightforward analysis with simple experimental designs using one or more of our semi-automated and highly reproducible analysis pipelines.

This also includes standard assistance with the manuscript writing (i.e. related Materials & Methods section). In such cases, usually an acknowledgement to IBU in general and, potentially but not necessarily, to the responsible bioinformatician(s) is appropriate.

Example sentence for acknowledgements:

The Interfaculty Bioinformatics Unit (IBU), University of Bern provided computational infrastructure and support with bioinformatic analyses.

→ For more details please have a look at our website <https://www.bioinformatics.unibe.ch>

Guidelines for authorship in collaborations

Research Mode

These apply to all analyses not covered in the previous section. It includes

- custom pipelines due to specific data or analysis needs
- analysis of data from emerging technologies not covered in our bundles that will require the design and implementation of a workflow
- custom downstream analysis and visualization of data from service mode.

Usually, this is considered to be a substantial contribution to the study upon publication and as such leads to a co-authorship. Exceptions to this rule need to be discussed on a per-project basis.

Example Affiliation:

Interfaculty Bioinformatics Unit (IBU) and Swiss Institute of Bioinformatics (SIB), University of Bern, Bern, Switzerland

→ For more details please have a look at our website <https://www.bioinformatics.unibe.ch>

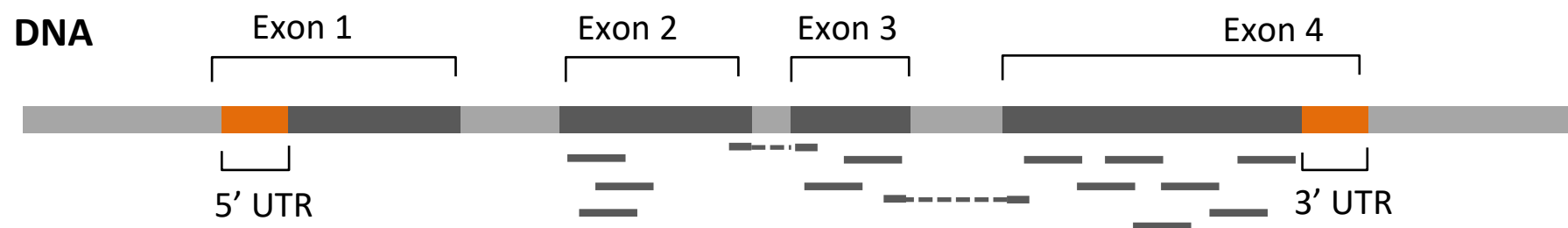
RNA-seq data processing

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>)

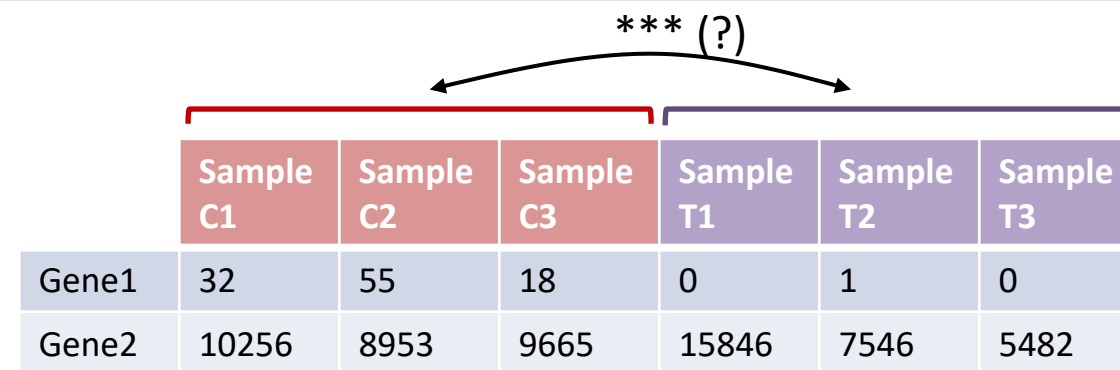


We end up with a table of read counts for each sample and gene:

	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

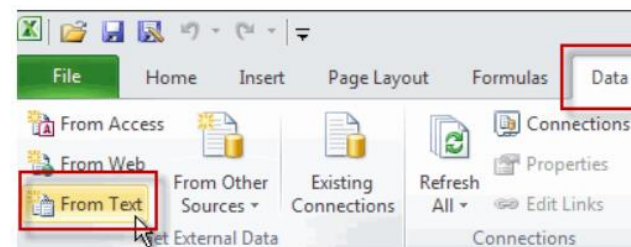
We use DESeq2 for this task, and the analysis involves the following steps:

- 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.
- 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.

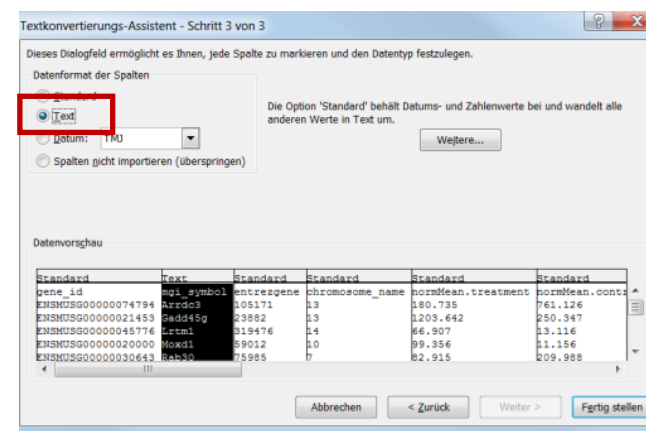
Overview of output files

For each comparison, you obtain a file where the format of the file name is:
Condition1.Condition2.DERResults.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!



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

$$\text{adjusted} \left(\log_2 \left(\frac{\text{normMean Condition 1}}{\text{normMean Condition 2}} \right) \right)$$

See slide 4, point 3 for explanation of adjustment

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

Standard error of
log2FoldChange

Wald test statistic

$$\frac{\log_2 \text{FoldChange}}{\text{lfcSE}}$$

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

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 $\text{padj} \leq \text{threshold}$, 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.

Bulk RNA-seq

Gene ontology (GO) enrichment analysis

Interfaculty Bioinformatics Unit | IBU, University of Bern

12.01.2022

The Gene Ontology (GO)

<http://geneontology.org/>

The Gene Ontology project provides controlled vocabularies of defined terms representing gene product properties. These cover three domains:

1) Cellular component (CC):

These terms describe a component of a cell that is part of a larger object, such as an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

2) Biological Process (BP): → Often tends to be the most interesting category

A biological process term describes a series of events accomplished by one or more organized assemblies of molecular functions. Examples of broad biological process terms are "cellular physiological process" or "signal transduction". Examples of more specific terms are "pyrimidine metabolic process" or "alpha-glucoside transport". The general rule to assist in distinguishing between a biological process and a molecular function is that a process must have more than one distinct steps.

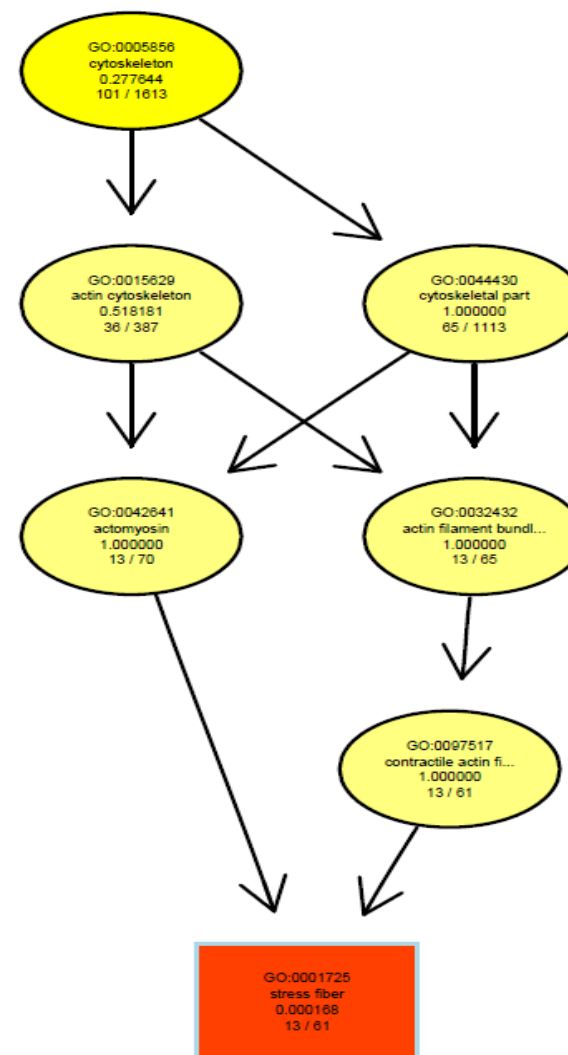
3) Molecular Function (MF):

Molecular function terms describes activities that occur at the molecular level, such as "catalytic activity" or "binding activity". GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are "catalytic activity" and "transporter activity"; examples of narrower functional terms are "adenylate cyclase activity" or "Toll receptor binding".

The GO as a graph

The structure of the GO can be described in terms of a graph (see example on right), where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms, but unlike a strict hierarchy, a term may have more than one parent term.

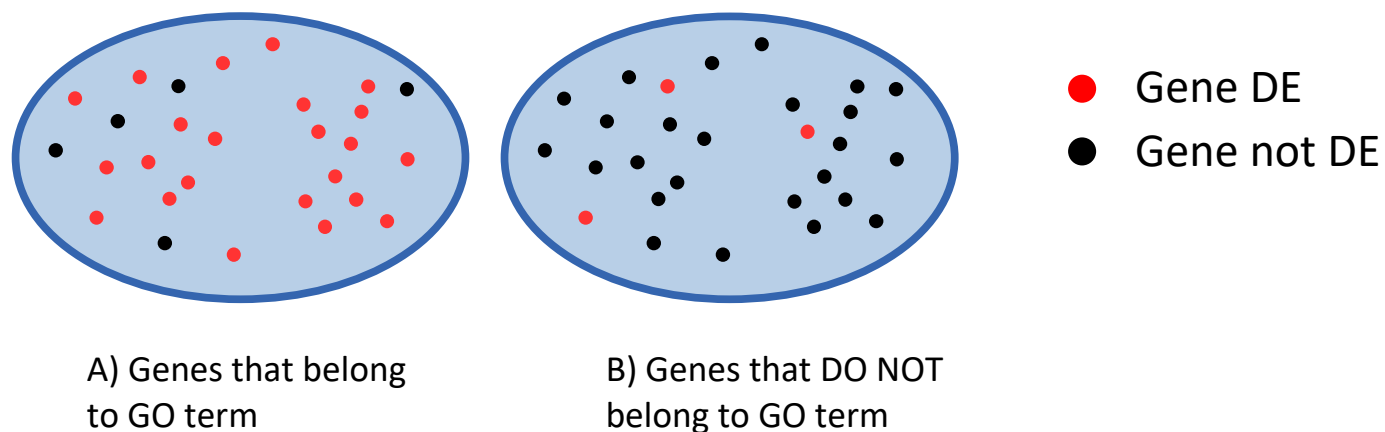
<http://geneontology.org/page/ontology-structure>



GO enrichment analysis

Goal: Detect which processes are affected by the experiment

Principle: Test if differentially expressed (DE) genes are significantly overrepresented within a particular GO term



Is the proportion of red genes higher in A than in B?

GO enrichment analysis

There are many different tools to perform GO enrichment analysis, and currently there is no consensus in the literature as to which one is the best. It is also very well possible that there is no single tool that always performs best in all datasets.

Your results were produced with **topGO** which is a widely used R Bioconductor package. An advantage of topGO is that it can take into account the hierarchical structure of the GO (i.e. the parent-child relationships).

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer

Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006

Advance Access publication April 10, 2006

Associate Editor: Martin Bishop

topGO results file

For each pairwise comparison between experimental groups, you will receive a text file.
The file name is set up like this: **Condition1.Condition2.topGoResults.txt**

Identifier and description of the GO term

Position the GO term would have if table were ranked by classic.Fisher P-value rather than weight01.Fisher

See slide 2

GO.ID	Term	Annotated	Significant	Expected	Rank in classic.Fisher	weight01.Fisher	weight01.KS	classic.Fisher	ontology
GO:0008201	heparin binding	118	61	28.63	5	1.10E-10	2.60E-09	1.10E-10	MF
GO:0008013	beta-catenin binding	75	39	18.2	10	2.00E-07	5.00E-06	2.00E-07	MF
....									
GO:0090090	negative regulation of canonical Wnt sig...	92	49	22.42	174	2.20E-09	2.00E-09	2.20E-09	BP
GO:0002053	positive regulation of mesenchymal cell ...	32	23	7.8	213	2.00E-08	1.90E-07	2.00E-08	BP
....									

Total number of genes assigned to GO term and actually detected in our dataset

Number of these genes that are detected as differentially expressed in our DE analysis (adjusted-P<0.05)

Number of differentially expressed (DE) genes we would expect to see if DE genes were randomly distributed across GO terms

P-values from three different ways to perform the enrichment test → see next slide for details

The results are ranked by weight01.Fisher and the **top 50** terms are output for each of the three subontologies, **provided at least one of the three P-values is below 0.05. No correction for multiple testing is applied** as it is not clear how to correctly do this (See Section 6.2. of <http://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf>)

topGO results file

The table contains P-values from 3 different enrichment tests. This lets you assess how consistently a particular term is detected as significant.

weight01.Fisher	weight01.KS	classic.Fisher
1.10E-10	2.60E-09	1.10E-10
2.00E-07	5.00E-06	2.00E-07

The three analyses differ in

1) Whether or not they consider the **hierarchical structure of the GO**

- **weight01**: Considers the GO graph and tries to find the most interesting term in a particular region of the graph (see Alexa et al. 2006 for details). It tends to prioritise more specific terms (i.e. children) over more general terms.
- **classic**: Performs a separate test for each GO term, ignoring the overlap between terms. It tends to favour larger terms

2) How they **rank the genes** within a GO term

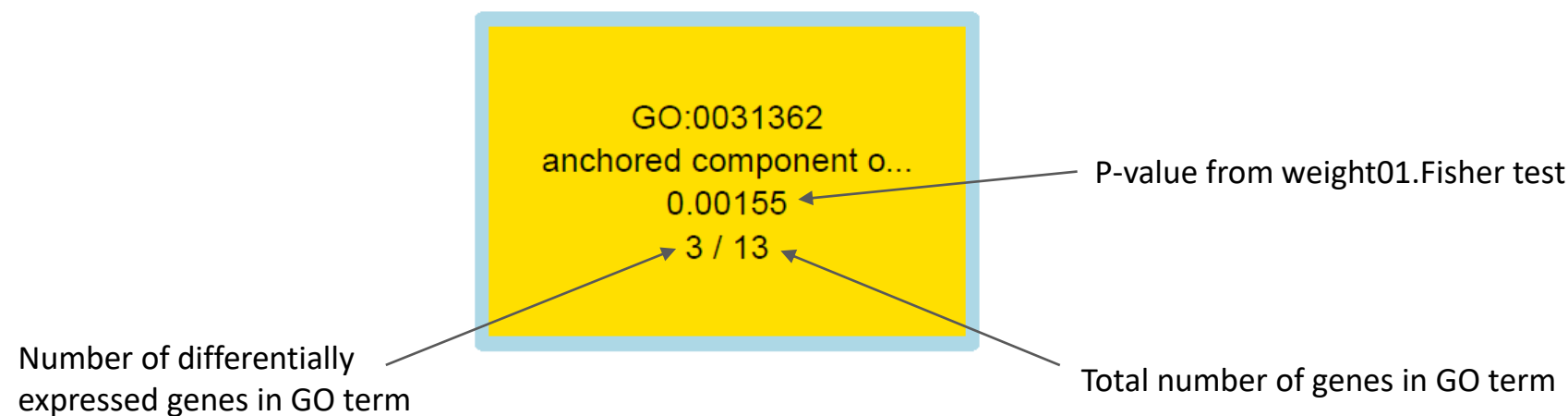
- **Fisher**: Performs a Fisher's exact test which compares the proportion of differentially expressed (DE) genes among all genes assigned to the GO term and all other genes (→ Slide 4). This approach relies on a fixed threshold that determines if a gene is considered DE or not.
- **KS** (Kolmogorov-Smirnov test): Orders all genes by P-value and tests if the genes assigned to a particular GO term are enriched at the top or the bottom of this table (Subramanian et al. 2005 PNAS)

topGO results file

For each pairwise comparison between experimental groups, you will also receive three pdf files which show how the detected terms are distributed across the GO graph. These plots include the **TOP 10** terms based on weight01.Fisher, **without applying a P-value cut-off**.

The file name is set up like this: **Condition1.Condition2.subontology_weight01_10_all.pdf**.
Subontology is one of CC, MF, BP.

The top 10 terms are shown as rectangles, all other terms as ovals. The colour of the box indicates the relative significance (red=most significant).



Bulk RNAseq

Gene Set Enrichment Analysis (GSEA)

Interfaculty Bioinformatics Unit | IBU, University of Bern

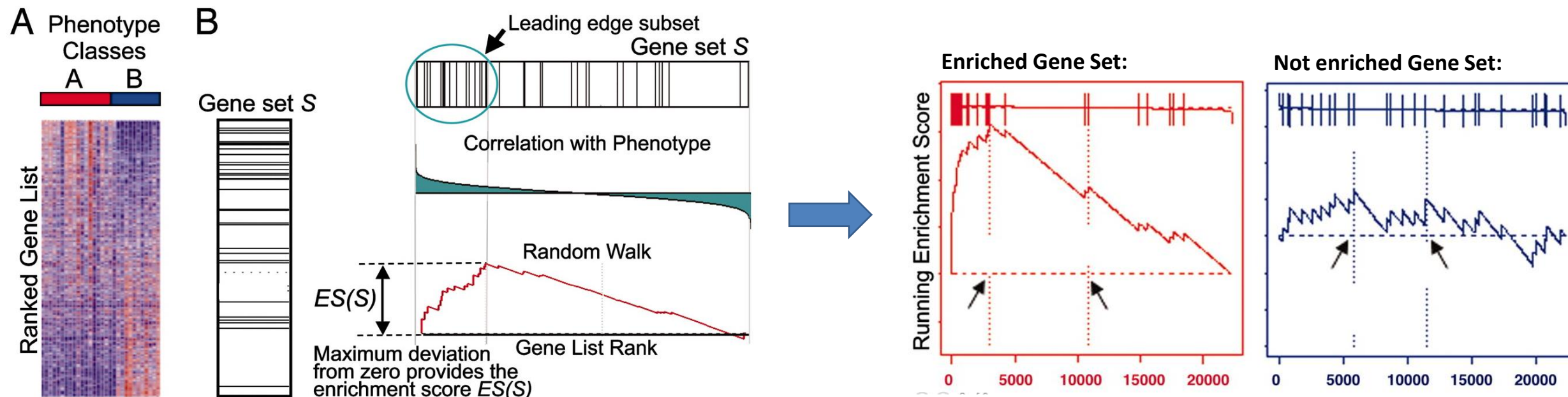
12.01.2022

Gene Set Enrichment Analysis (GSEA)

Goal: Detect which pathways/processes are affected by the experiment

Principle: GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states

(Subramanian et al. 2005, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles (PNAS))



Gene Set Enrichment Analysis (GSEA)

Your results were produced with R Bioconductor package **clusterProfiler**.

(Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu x, Liu S, Bo X, Yu G (2021). “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.” *The Innovation*, 2(3), 100141. doi: 10.1016/j.xinn.2021.100141.)

GSEA are performed based on two different databases (if available):

- **KEGG pathways:**
Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (<https://www.genome.jp/kegg/>)
- **MSigDB:**
The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. We use the hallmark gene sets, which are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes. (<https://www.gsea-msigdb.org/gsea/msigdb>)

<https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>

Result files

gseKEGG.txt

For each pairwise comparison between experimental groups, you will receive a text file with enriched KEGG pathways. The file name is set up like this: **Condition1.Condition2.gseKEGG.txt**

KEGG pathway ID Number of Genes in KEGG pathway p-value of the enrichmentScore (ES) is calculated using permutation test

ID	Description	setSize	enrichmentScore	pvalue	p.adjust
mmu05165	Human papillomavirus infection	323	-0.35271768	0.000179211	0.002511479
mmu04020	Calcium signaling pathway	237	-0.379500422	0.000181917	0.002511479
mmu04360	Axon guidance	179	-0.388358414	0.000182916	0.002511479
mmu04510	Focal adhesion	198	-0.399113881	0.000183016	0.002511479
...					

Pathway description represent the degree to which a set S is over-represented at the top or bottom of the ranked list L adjust the estimated significance level to account for multiple hypothesis testing (Benjamini-Hochberg). **This is the p-value that should be considered**

Result files

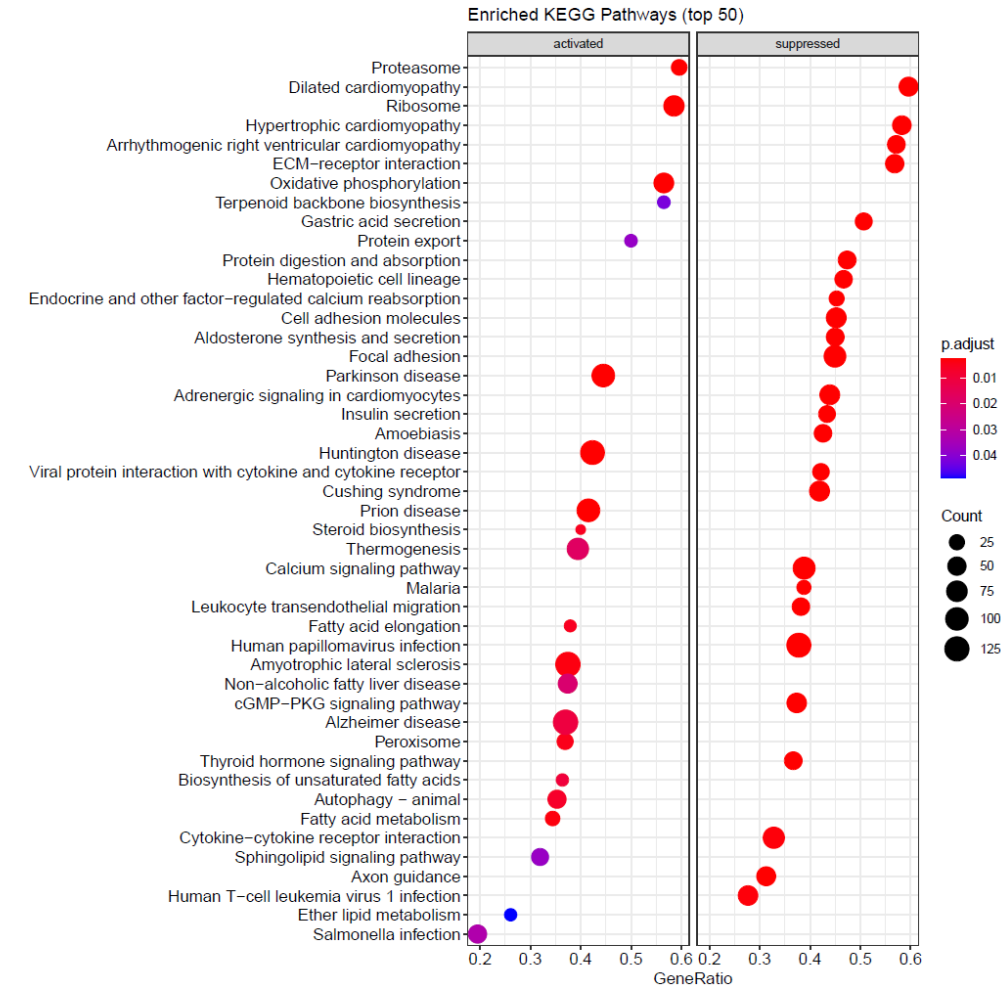
gseKEGG.pdf

For each pairwise comparison between experimental groups, you will receive a pdf file with a dotplot of top 50 significant KEGG pathways.

The colors correspond to the adjusted p-value and the point size to the number of genes in the KEGG pathway.

The file name is set up like this:

Condition1.Condition2.gseKEGG.pdf



Result files

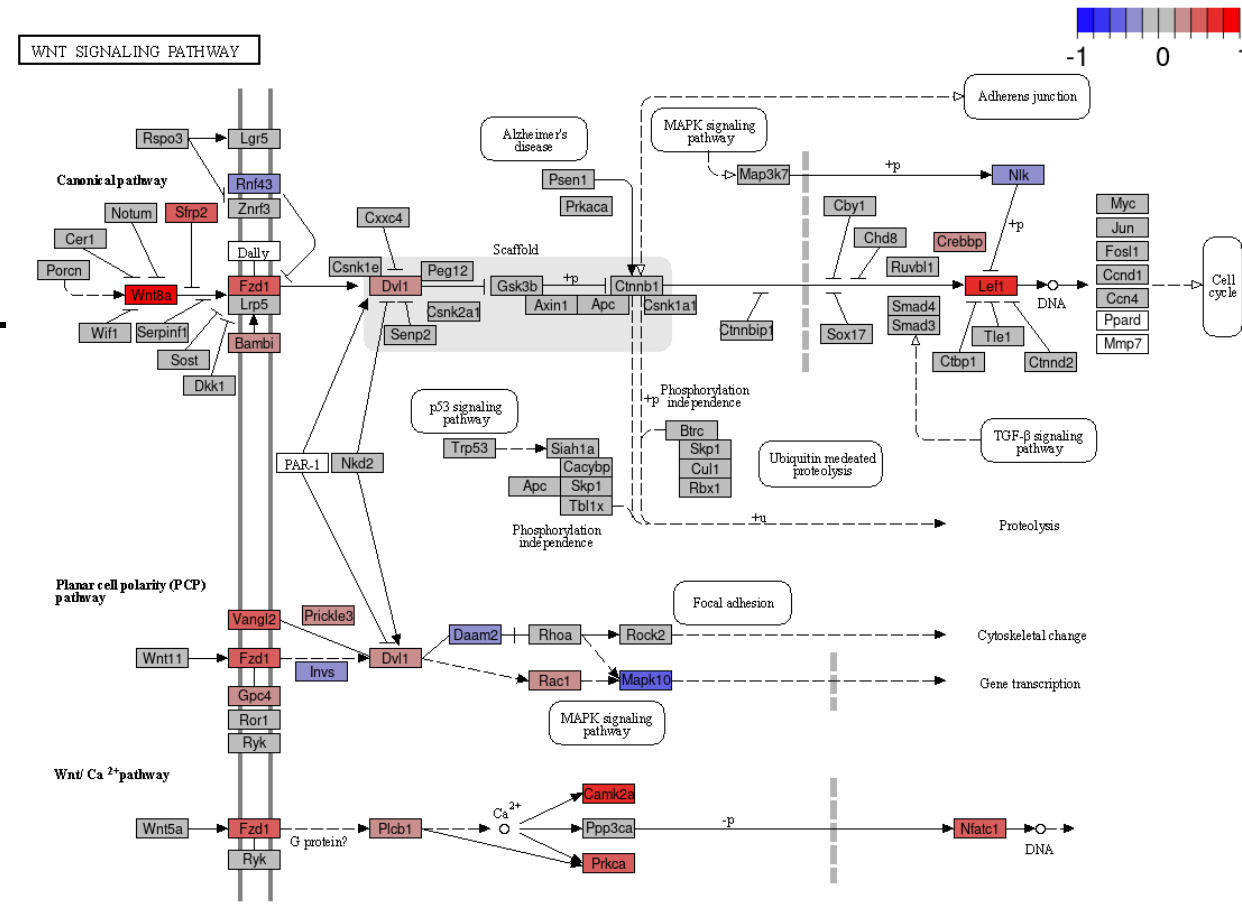
gseKEGG.png

For each pairwise comparison between experimental groups, you will receive pdf files with significant KEGG pathway plots with integrated log₂FoldChanges. The plots were produced with R Bioconductor package pathview. (Luo, Weijun, Brouwer, Cory (2013). "Pathview: an R/Bioconductor package for pathway-based data integration and visualization." *Bioinformatics*, 29(14), 1830-1831. doi: 10.1093/bioinformatics/btt285.)

The color of the boxes represents log₂FoldChanges of gene expression between the two conditions (blue: downregulated, red: upregulated).

The file name is set up like this:

KEGGID.Condition1.Condition2.gseKEGG.png



Data on KEGG graph
Rendered by Pathview

<https://bioconductor.org/packages/release/bioc/html/pathview.html>

Result files

gseMSigDB_hallmark.txt

For each pairwise comparison between experimental groups, you will receive a text file with enriched MSigDB hallmark gene sets. The file name is set up like this:

Condition1.Condition2.gseMSigDB_hallmark.txt

MSigDB hallmark
gene set identifier

Number of Genes in
gene set

p-value of the enrichmentScore (ES) is
calculated using permutation test

ID	setSize	enrichmentScore	pvalue	p.adjust
HALLMARK_TNFA_SIGNALING_VIA_NFKB	10	0.83002413	0.000317172	6.98E-03
HALLMARK_IL6_JAK_STAT3_SIGNALING	12	0.803994237	0.000655058	7.21E-03

represent the degree to which a set S is over-represented at the top or bottom of the ranked list L

adjust the estimated significance level to account for multiple hypothesis testing (Benjamini-Hochberg).
This is the p-value that should be considered

Result files

gseMSigDB_hallmark.pdf

For each pairwise comparison between experimental groups, you will receive a pdf file with a dotplot of top 50 significant MSigDB hallmark gene sets.

The color correspond to the adjusted p-value and the point size to the number of genes in the MSigDB hallmark gene set.

The file name is set up like this:

Condition1.Condition2.gseMSigDB_hallmark.pdf

