UNIVERSITÄT
BERN

Bulk RNAseq

Gene Set Enrichment Analysis (GSEA)

Interfaculty Bioinformatics Unit, University of Bern

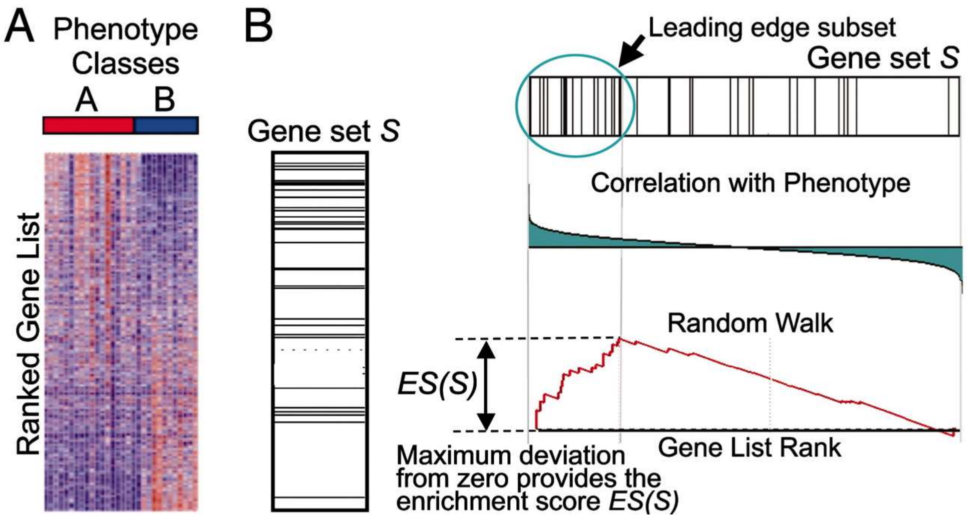
09.09.2021

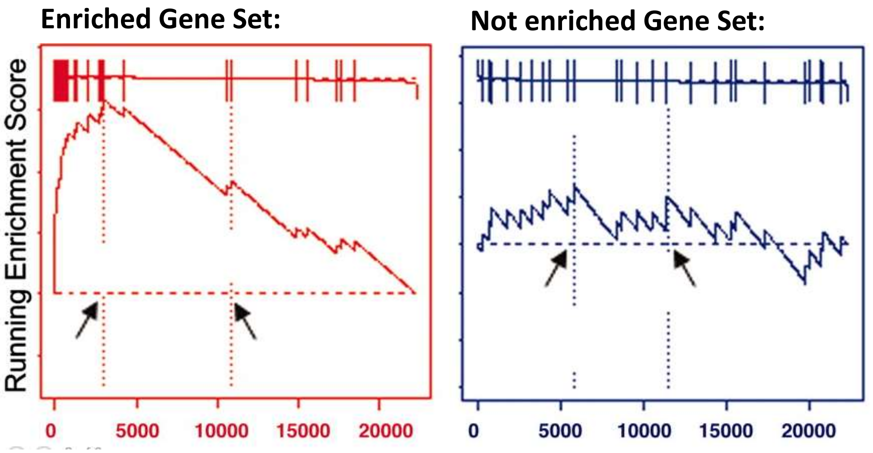
SIB
Swiss Institute of Bioinformatics

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

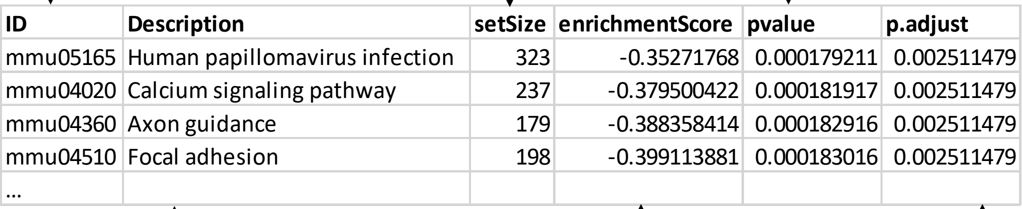
gseKEGG.txt

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

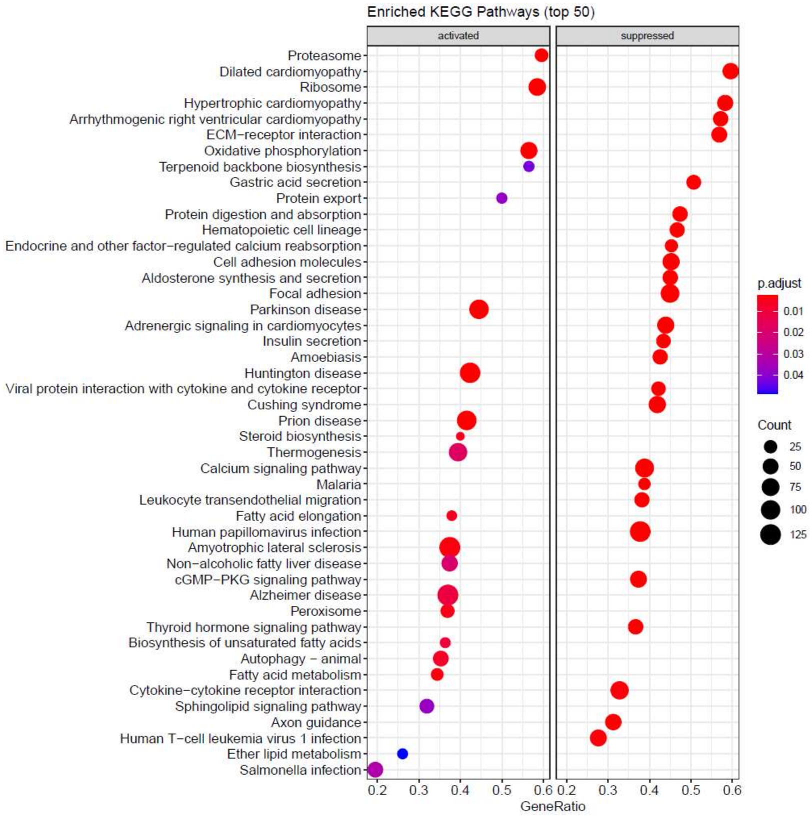
Result files

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



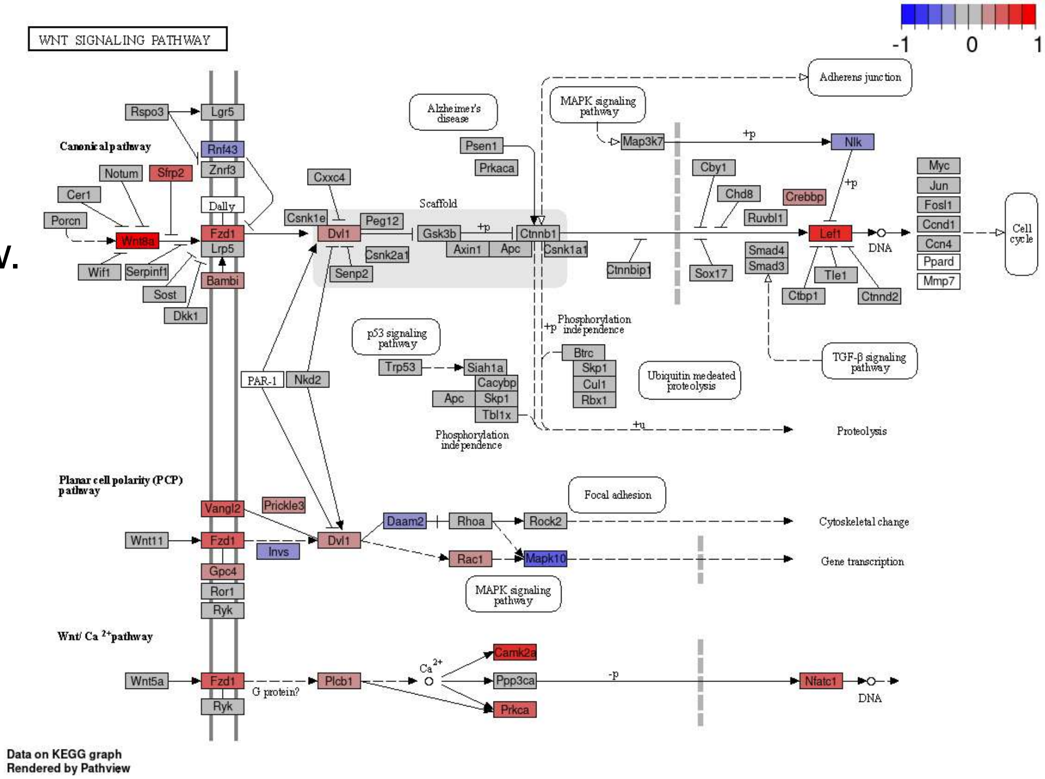
Result files

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**



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

Result files

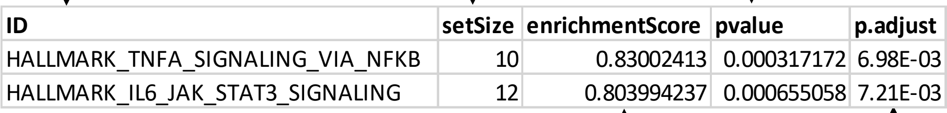
gseMSigDB_hallmark.txt

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

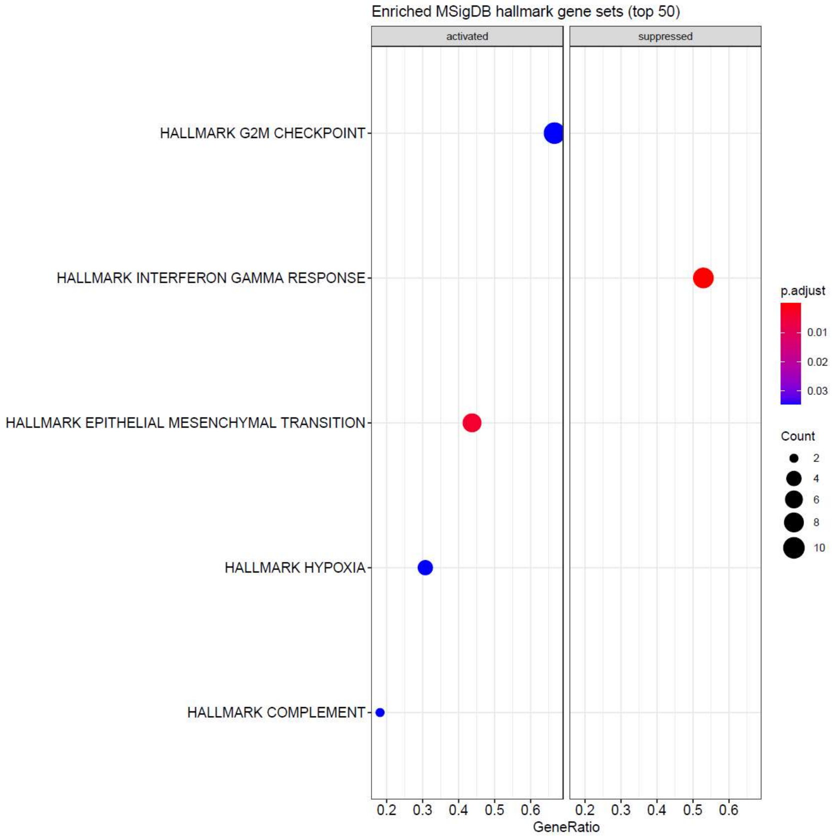
Result files

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


Enriched MSigDB hallmark gene sets (top 50)
activated
suppressed
HALLMARK G2M CHECKPOINT
HALLMARK INTERFERON GAMMA RESPONSE
HALLMARK EPITHELIAL MESENCHYMAL TRANSITION
HALLMARK HYPOXIA
HALLMARK COMPLEMENT
GeneRatio
p.adjust
Count