Enrichment analysis

## Material

* [MSigDB](http://www.gsea-msigdb.org/gsea/msigdb/index.jsp)
* clusterProfiler [vignette](https://bioconductor.org/packages/release/bioc/vignettes/clusterProfiler/inst/doc/clusterProfiler.html)
* [Revigo](http://revigo.irb.hr/)
* [Signaling Pathway Impact Analysis (SPIA)](https://bioconductor.org/packages/release/bioc/html/SPIA.html)
* Original [paper](https://www.pnas.org/content/102/43/15545) on GSEA
* [STRING](https://string-db.org/) for protein-protein interactions
* [GO figure!](https://gitlab.com/evogenlab/GO-Figure) for plotting GO terms and the associated [paper](https://www.frontiersin.org/articles/10.3389/fbinf.2021.638255/full)

## Exercises

Load the following packages:

If the FindMarkers or FindAllMarkers functions were used, we obtained a table listing only the significant genes, but we don’t have any information of fold change for the non-significant genes. Therefore, we can use the over-representation analysis which is a threshold-based method. Using our list of significant genes, we can test if any gene set is over-represented among significant genes or not using a test similar to a Fisher test to compare differences in proportions.

The clusterProfiler package provides functions for over-representation analysis of Gene Ontology gene sets (among other functions, including functions for actual GSEA) or KEGG gene sets.

Genes can be labeled using different types of labels, eg symbol, Ensembl ID, Entrez ID. To list the allowed label types use:

## About OrgDb

For other organisms, you can find available OrgDbs at [bioconductor](http://bioconductor.org/packages/release/BiocViews.html#___OrgDb)

Let’s select a set of genes that are downregulated in the tumor cells compared to normal:

We can do a Gene Ontology term over-representation analysis based on this set of genes. Make sure you check out the help of this function to understand its arguments:

The results are stored in the @result slot:

|  |
| --- |
| The columns GeneRatio and BgRatio |
| The columns GeneRatio and BgRatio that are in the enrichResult object represent the numbers that are used as input for the Fisher’s exact test.  The two numbers (M/N) in the GeneRatio column are:   * M: Number of genes of interest (in our case tum\_down\_genes) that are in the GO set * N: Number of genes of interest with any GO annoation.   The two numbers (k/n) in the BgRatio column are:   * k: Number of genes in the universe that are in the GO set * n: Number of genes in the universe with any GO annoation   A low p-value resulting from the Fisher’s exact means that M/N is signficantly greater than k/n. |

Some GO terms seem redundant because they contain many of the same genes, which is a characteristic of Gene Ontology gene sets. We can simplify this list by removing redundant gene sets:

We can quite easily generate a plot called an enrichment map with the enrichplot package:

Instead of testing for Gene Ontology terms, we can also test for other gene set collections. For example the Hallmark collection from [MSigDB](http://www.gsea-msigdb.org/gsea/msigdb/index.jsp):

### Clear environment

Clear your environment:

rm(list = ls())  
gc()  
.rs.restartR()