SCRIPT TWO: PATHWAY AND MATRICES (SEMANTIC SIMILARITY)

by Eugenio Del Prete

1. The work folder is important to avoid mixing unnecessary results and to collect the final .txt files for this script (mandatory). All the 'XXX\_correspondence.txt' files created from the script one must be in the same folder (in this example, the folder is 'Output\_Txt').

2. The most important packages are in the list. All the dependencies will be automatically loaded. If they lack, you need to download them from R-CRAN or BIOCONDUCTOR repositories. The used versions are reported in the manuscript.

3. This step generates the list of the files in the folder with the correspondence GO terms - differential expressed genes.

4. The dataset are separated for list of genes. In this example, only the first 5 GO terms are taken into account. If you want to change this value (X):

go\_term\_sub <- lapply(go\_term, function(x) x[[1]][1:5]) -> change 5 into X

aux\_3 <- strsplit(aux\_2[[1]][2:6],",") -> change 6 into X+1

and then you can save the result for KEGG enrichment (see below, 8). Pay attention to the name of the file: remember to put the number of GO terms at the end.

5. This is the preparation of the annotation for the semantic similarity evaluation.

6. The semantic similarity matrices are generated, one for the genes and one for GO terms. The Wang method is used to calculate the semantic similarity and the Best-Match-Average method is used to calculate the semantic similarity among groups. These methods are suitable for highlighting the importance of the gene ontology tree. Different methods provides different results.

7. It is possible to plot the semantic similarity matrices, for GO terms, genes and DO terms (disease ontology). As shown, DO ID are preselected and available online: they depends on the pathologies that you want to study. Pay attention to the position of ‘library(DOSE)’: some function can have a conflict (overwriting), so it is better not to move this line from its positon in the script.

8. The final step is the evaluation of the KEGG Enrichment, with a graph, taking into account the variable generated in the step 4. ‘library(clusterProfiler)’ has the same issue of the previous one, so it is better not to move this line from its positon in the script.