SCRIPT ONE: COELIAC DISEASE EXAMPLES (GENE SET ENRICHMENT ANALYSIS)

by Eugenio Del Prete

There are two example in this script: GSE11501 (GPL6104) and GSE87629 (GPL694). We suggest to run one example at a time, collect the result and erase the R global environment for the next example, otherwise you need to rename the variable for each example. With reference to the numbers in the script:

1. The work folder is important to avoid mixing unnecessary results and to collect the final .txt file for the script two (mandatory).

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. In this way, all the couple experiment-platform are downloaded. Pay attention to the correct platform for the next step.

4. The data are converted in an ExpressionSet class for the calculation of the differential expression. 'getGEO' is a function from the package 'GEOquery': pay attention, because last version seems to have some issues;

5. The matrix design is not always straightforward. You need to check which GSM experiments (health and sick patients) are available. For instance, 110 sick patients and 22 healthy control are consecutive in the example A: in this case, the matrix design is simple. For the example B, you need to eliminate the experiment GSM2335908 and create a matrix design with the alternation of sick patients and healthy control. Unfortunately, a previous control of GSM experiments on GEO NCBI is compulsory, because the choice is up to the user and very difficult to automate.

6. The name of the features in the table are not standardised. For the creation of the correct subtable, you should check the name of the features (column name): usually, the issue is in the column of the genes ('ID', 'adj.P.Val', 'P.Val', 't', 'B', 'logFC' are unchanged);

7. You can save the subtable in Excel format for other analyses.

8. This short function is important for the creation of the ‘topGO class’, because you can select the threshold of LogFC value. Different thresholds provide different results.

9. This is the creation of the ‘topGO class’, with the choice of the gene ontology class (in this case, BP), the list of genes, the previous selection function and the annotation;

10. Some value are collected: number and name of the genes, GO terms used for the analysis (not only the final ones).

11. Fisher's Exact Test and Kolmogorov-Smirnov Test are performed. The second one is only for evaluation purpose.

12. If you are interested in comparing the two test, you can use this step.

13. The GO terms tree is created, with priority on the first 5 GO terms and Fisher's Exact Test, but using all the available information.

14. This step is very important. It generates a text file with the correspondence GO terms - differential expressed genes. Pay attention to the name of the file. For example, CD1\_GSE11501 means

* CD -> pathology code,
* 1 -> LogFC threshold,
* GSE11501 -> dataset.

This file is mandatory for the script two.