Lorenz Gerbeth
13.06.2020

Course Project – Reproducible Research in R

Finding common genes in preselected gene lists

Suggested steps:

1. Install the tmod package
2. Download the full gene set from the GSEA homepage and load it into R using the tmodImportMSigDB function. This results in a list containing a dataframe which encodes the IDs of the gene lists and link it to their names, a dataframe containing the gene lists, and a list with all genes.
3. Another list containing the gene lists of interest will be prepared beforehand by me. Using this, the right lists can be selected from the GSEA data set (object class and structure has to be figured out).
4. A new dataframe is created containing all genes in its first column (extracted from the GSEA data)
5. Now every gene list of interest isolated in step 3 is added to this dataframe in a way that if the gene written in a particular row of column 1 (all genes) is present in this gene list, the observation will be denoted as 1 and otherwise as 0.
6. Subsequently, a last column can be added to the data frame, which contains the sum of every row. The observations are then sorted to display the genes occurring in most lists at the top and the ones with fewest occurences at the bottom.

Example of final output:

Gene	**List1**	**List2**	**List3**	**List4**	**List5**	**…**	**SumOfOcc**
G1	1	1	0	1	1	…	16
G2	0	1	1	0	1	…	15
G3	0	0	1	1	1	…	15
G4	1	0	0	1	1	…	12
G5	1	0	1	0	0	…	10
G6	0	0	0	1	0	…	10
G7	0	1	0	0	0	…	10
…	…	…	…	…	…	…	…