# MSstats label-free preprocessing

This repo contains a script and a Rmd file for the pre-processing and normalization of MaxQuant output files through the MSstats R package. The output is a tabular file in wide format (1 row per protein, 1 column per sample/condition) that could be used as an input to run statistics with Limma or similar.

## Instructions for using the script and to create a reproducible report of your normalization step

1. Download/clone the contents of this repo into your local computer. This should create a R project folder with the script to run the preprocessing.
2. Delete the MSstats\_Output\_data/ folder and its contents from your local computer.
3. Add the next three MaxQuant output files into this folder

* evidence.txt
* proteinGroups.txt
* annotation.csv (not included in the MaxQuant txt folder, see below how to create this one).

**NOTE**: These files should be in the same folder as the R script, and this folder should be an initiated RStudio project (There should be a .Rproj file in the same folder).

1. Open your RStudio project by double-clicking the .Rproj file in your newly created R project folder.
2. Open the script mq\_to\_msstats\_formating\_normalization\_n\_prep\_for\_limma.R
3. Set the options to run the pre-processing and normalization with MSstats in the ‘Set conditions’ section of the script between the lines 9 and 32
4. Click ‘Source’ in your R Studio session
5. The script should generate two .csv files: one in long MSstats-format and one in wide format, suitable as an input for Limma.

### Creating the annotation file (provisional)

At some point I will add something in the processing script (or create a new script) for trying to automotize this, but for now, the annotation.csv file can be created manually as described now.

1. Open a new spread sheet (i.e. in MS Excel).
2. The first row should be your column names as follows: “Raw.file”, “Condition”, “BioReplicate”, “IsotopeLabelType”
3. Fill the rows with the required information for each of the required sample.

* For Raw.file: give the name of your Thermo RAW file as it was named when processed by MaxQuant.
* For Condition: give the Experimental or Biological condition of the sample.
* For BioReplicate: give the number of the biological replicate associated with this sample. If every sample came from a different biological source, then you can give a different (any) number for each sample.
* For IsotopeLabelType: Type of labelling. Since in this case we are working with label-free quantification, set all rows in this column to ‘L’.