# MSstats label-free preprocessing

This repo contains a script and a Rmd file for the pre-processing and normalization of MaxQuant or DIA-NN 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 for your normalization step starting from MaxQuant outpus

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 before executing the script).

**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. Modify lines between 16 to 31 to set up the parameters for both the transformation from MaxQuant format to MSstats format, and for the actual summarizaton and normalization.
4. Execute the script (click ‘Source’ on the top-right corner of the script).
5. The script should generate three .csv files: msstats\_tabular\_data\_for\_limma\_input.csv, in wide format suitable for downstream analysis with limma. And two files in long format within MSstats\_Output\_data with the un-normalized and the normalized feature intensities before and after MSstats pre-processing.

## Instructions for using the script for your normalization step starting from DIANN outputs

**BE AWARE!!**: There is a known issue with the dataProcessing function fron MSstats that makes it use a lot of RAM with big input files (> 1 million rows). If you have A big output from DIANN and have issues with your R session crashing due to RAM overload, you can execute this script up to line 105 and get the output of the MSstats formatted data from ~/MSstats\_Output\_data/MSstats\_formated\_tables/msstas\_formated\_diann\_data\_bf\_normalization.csv and continue on Galaxy, where the RAM shouldn’t be an issue.

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 MainOutput.tsv output file from DIA-NN into this folder.
4. Add your annotation\_diann.csv file into this folder.

**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).

**NOTE 2**: Check the samples folder a sample of the annotation\_diann.csv file and how it should look like.

1. Open your RStudio project by double-clicking the .Rproj file in your newly created R project folder.
2. Open the script diann\_to\_msstats\_formating\_normalization\_n\_prep\_for\_limma.R
3. Modify lines between 16 to 21 to set up the parameters for both the transformation from MaxQuant format to MSstats format, and for the actual summarizaton and normalization.
4. Execute the script (click ‘Source’ on the top-right corner of the script).
5. The script should generate three .csv files: msstats\_tabular\_data\_for\_limma\_input.csv, in wide format suitable for downstream analysis with limma. And two files in long format within MSstats\_Output\_data with the un-normalized and the normalized feature intensities before and after MSstats pre-processing.

## Creating the annotation file

You have 2 options to create your annotation file:

* Use the create\_annotation\_file.R script created for this purpuse (*RECOMENDED*). **NOTE**: Now the script only works if every sample corresponds to a different biological replicate and for label-free samples. Manually create your file if otherwise.
* Manually create your annotation.csv file in a spread sheet editor (such as MS Excel)

### Using the create\_annotation\_file.R ‘interactive’ script

1. Corroborate that you have the create\_annotation\_file.R in your R Project folder.
2. Go to the Console in your opened R Studio project session.
3. Type source("create\_annotation\_file.R")
4. Answer the questions as prompted on the Console in your R session.
5. *Important!*: please corroborate that your sample names/codes correspond with the desired experimental condition by opening the newly created annotation.csv file. It should be in the same folder of your R Project.

### Manually create your annotation.csv file in a spread sheet editor

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’.