# FDR correction on a subset of features / proteins after limma

This repo constains a main script and a series of sample scripts that would help you execute a limma-based differential abundance analysis and perform multiple testing correction only on a subset of interesting features (i.e. semi-tryptic peptides or bacterial proteins from metaproteomics studies, etc).

## What does it do?

The main script is set up to run a limma-based differential abundance analysis on two groups, and generate:

1. A tabular output with FCs, non-adjusted (moderated) p-values, and adjusted p-values.
2. The tabular output will contain a column fdr\_correction that states if the multiple-testing correction was performed on the subset (‘feature-specific’) or if it is showing the globally-calculated FDR.
3. A short HTML report with an interactive volcano plot of the global comparison and the comparison based on the interesting/selected features.

## How to use it?

First of all: download this repo and set it up as a separate RStudio project for your particular analysis.

You might need to delete or move the example files found in the data/ folder.

I am now describing the required input files and below I’ll describe which of the downloaded script you would need to open and modify and how.

### Required input files (should be placed in the folder data, that will be cloned together with this repo):

Refer to the data/ folder in this repo for examples.

* input\_limma.txt
  + First column should be called ID. It contains the ‘names’ of the proteins or peptide sequences that you will investigate.
  + Each other column represents a sample, and should contain log2-transformed and normalized protein/feature abundance values.
  + Note: Right now you should only have two groups of samples in your input\_limma.txt, but you can refer to the [limma\_multigroup\_script](https://github.com/MiguelCos/limma_multi_group) to be able to modify this script in order to do multiple paired comparisons if necessary.
* annotation.txt
  + First column: Sample\_ID
  + Second column: Group
  + Info: this table will associated the sample ID with each of your experimental conditions. The IDs in the Sample\_ID column should match **exacty** the sample names represented in the columns of input\_limma.txt
* annotation\_features.txt
  + First column: ID. Should contain **exactly** the same protein/features IDs that you have in the ID column of the input\_limma.txt file.
  + Second column: feature\_category. Should contain a characteristic which is interesting for each feature that you want to evaluate.
    - For example: you have a list of peptides that you have already characterized as ‘semi-specific’ or not (for example with [this script](https://github.com/MiguelCos/mapping_peptides_to_proteins_from_fasta_file)). Then you can select to keep only the column of your feature IDs (i.e. peptide sequences) and the interesting feature type (i.e. tryptic, semi\_Nterm, semi\_Cterm).
    - In this repo I am also including an example script on how to generate this kind of tabular annotation if you have, for example: bacterial vs human proteins. For this, check the link in the section “How to create annotation\_features.txt” below.

### Modifying the script:

Provided that the user has placed the required input files in the data/ folder, the user should:

1. Open the script fdr\_correct\_on\_limma\_subset.R.
2. Give the analysis a ‘code name’ (lines 5 -> 7).
3. Define if doing a robust or least-squares regression with limma (lines 9 -> 11).
4. Define the interesting feature type (lines 13 -> 17): in this example, by default, this is defined as “semi\_Nterm” because that is the interesting feature of the peptides for which we want to apply the ‘selected’ FDR correction. Notice that the interesting feature should match to one of the ones that you have listed in the annotation\_features.txt file.
5. Define which/how the two groups should be compared (line 72): you should modify the makeContrasts arguments to match your experimental groups as defined in the annotation.txt. Note: The group at the left of the - sign represents the numerator of the comparison (everything ‘upregulated’ would be increased in the group at the left of -.
6. Then you can execute the code line by line or just click Source in the top right corner of the script in R Studio.
7. You will get an Output/ folder containing the tabular results of the limma comparison an a short HTML report as described in the first section of this documentation.

## How to create annotation\_features.txt

Please refer to the example on how to create an annotation\_features.txt file in [this link](https://github.com/MiguelCos/limma_FDR_on_feature_subset/blob/main/Rmds/how_to_generate_annotation_features.md).

The first example refers to working from an annotated set of peptides from [this script](https://github.com/MiguelCos/mapping_peptides_to_proteins_from_fasta_file).

The second example is more generic when you can use the protein name or header to identify a protein as human or not (i.e. for example, in metaproteomics studies).

## Two other approaches for differential weighting for multiple testing were tested

The short report for those tests can be found in this repo at: [Testing different hypothesis weighting approaches](https://github.com/MiguelCos/limma_FDR_on_feature_subset/blob/main/Rmds/test_fdr_correction_approaches_semi_tryptic_peptides_tmt16plex_dataset.md#distribution-of-adjusted-p-values-per-approach)