Tutorial on data processing and msqrob2 analysis of experiments with simple designs

Lieven Clement

The result of a quantitative analysis is a list of precursor, peptide and/or protein abundances for every protein in different samples. In this tutorial we introduce a generic workflow for differential analysis of quantitative datasets with simple experimental designs.

In order to extract relevant information from these high troughput datasets, we will use our msqrob2 software tool.

Staes spike-in study (DIA-NN output)

We will use a publicly available spike-in study published by Staes et al. (Staes et al. 2024). They spiked digested UPS proteins in a yeast digested background at the following ratios (yeast:ups ratio 10:1, 10:2, 10:4, 10:8, 10:10). Here we will use a subset of the data, i.e. dilutions 10:2 and 10:4.

We will use output of the search engine DIA-NN 2.2.0. The main search output for this DIA-NN version was stored in the report.parquet file in the DIA-NN output directory, which can be found under data/spikein24-staesetal2024.parquet

DIA-NN provides multiple quantifications, e.g. derived from the MS1 or MS2 spectra, and at precursor or protein (protein group) level. The term 'precursor' refers to a charged peptide species and is the basic unit of identification and quantification in DIA. Hence, in the context of DIA we refer to a precursor table, instead of to a PSM table in DDA.

Examples of different quantities are:

- raw MS1 area: Ms1.Area, normalised MS1 Area: Ms1.Normalised, MS2 Precursor quantities: Precursor.Quantity, Normalised MS2 Precursor quantities: Precursor.Normalised, etc., which are all at the precursor level
- MS2 based summary at the protein (protein group)-level: PG.MaxLFQ

[1.a] Participants can perform an analysis using the staes rmarkdown script or the msqrob2GUI. Here, we will use the `Precursor.Quantity` column. Follow the steps in the script or gui and try to understand each of the analysis steps. We know the real FC for the spike in proteins and the yeast proteins (see description of the data). What do you observe?

[1.b] Repeat the analysis and change the normalisation to `diff.median` that centers all samples (columns) so that they all match the grand median by subtracting, from each column, the difference between that column's median and the grand median (code in script: `qf <- normalize(qf,  i = "precursors_log", name = "precursors_norm", method = "diff.median")`. What do you observe and try to explain this.

[1.c] Repeat the analysis starting from the `Precursor.Normalised` column. What do you observe? Are all data processing steps needed?

[1.d] Repeat the analysis, again use the GUI or the basis script with median of ratios normalisation factors (`nfLogMedianOfRatios` function). First change the summarisation by replacing maxLFQ with median polish (in scripts: `fun = MsCoreUtils::medianPolish` in the `aggregateFeatures` function), then use simple median summarisation (in scripts: `fun = matrixStats::colMedians`). What do you observe and try to explain this. (Note, that if you use scripts, you also have to add an additional argument for median polish and median summarisation to handle missing values, i.e. add an additional argument `na.rm = TRUE`, i.e. replace `fun = function(X) iq::maxLFQ(X)$estimate` with `fun = function(X) MsCoreUtils::medianPolish(X, na.rm = TRUE)`.

Breast cancer example

Eighteen Estrogen Receptor Positive Breast cancer tissues from patients treated with tamoxifen upon recurrence have been assessed in a proteomics study. Nine patients had a good outcome (or) and the other nine had a poor outcome (pd). The proteomes have been assessed using an LTQ-Orbitrap and the thermo output .RAW files were searched with MaxQuant (version 1.4.1.2) against the human proteome database (FASTA version 2012-09, human canonical proteome).

The data can be found in the folder dda/cancer after downloading and unzipping all data locally. Download data

Three peptides txt files are available:

1. For a 3 vs 3 comparison
2. For a 6 vs 6 comparison
3. For a 9 vs 9 comparison

Note, that the data are from data dependent acquisition and searched with MaxQuant, so the data processing will be slightly different. Users who work with scripts can modify the cptac Rmarkdown script.

```
https://github.com/statOmics/msqrob2data/raw/refs/heads/main/dda/cancer/peptides3vs3.txt
https://github.com/statOmics/msqrob2data/raw/refs/heads/main/dda/cancer/peptides6vs6.txt
https://github.com/statOmics/msqrob2data/raw/refs/heads/main/dda/cancer/peptides9vs9.txt
```

References

Staes, An, Teresa Mendes Maia, Sara Dufour, et al. 2024. "Benefit of in Silico Predicted Spectral Libraries in Data-independent Acquisition Data Analysis Workflows." *Journal of Proteome Research* 23 (6): 2078–89. https://doi.org/10.1021/acs.jproteome.4c00048.