The project folder is structured as follows (new features in red):

**1 – Raw files**

Raw data acquired by the mass spectrometer. We searched this data with either MaxQuant (files acquired in Data Dependent Acquisition mode) or increasingly often with DIA-NN (Data Independent Acquisition mode). Both software identify peptides by matching acquired spectra to theoretically expected peptides.

*Please keep these files for future reference/re-analysis/publication.*

**2 – DIA-NN/MaxQuant\_search (*v1)***

***This folder contains the results from the search software (MaxQuant or DIA-NN; if several analyses were delivered, they have version numbers and incremental list numbers, e.g.* 3 – Analysis\_*YYYY-MM-DD\_v2)*** Some of these files (especially the .mzTab files) are important for submitting search results to ProteomeXchange (e.g. PRIDE) repositories prior to publication.

* **MaxQuant**

For MaxQuant, we delete most temporary files, keeping only the “andromeda” and “txt” subfolders in the “combined” folder as well as the MaxQuant search parameters file. The “txt” folder contains MaxQuant’s output tables (despite the extension, tables are formatted as tab-separated values = .tsv). Of particular interest is *evidence.txt*, which is a long table of all individual peptidoform (= modified peptide sequence) observations, including direct peptide-to-spectrum matches (PSMs) and indirect match-between runs identifications. This is the main file which is used for reprocessing the MaxQuant output.

* **DIA-NN search**

For DIA-NN, we keep all results .tsv file. Table names vary depending on DIA-NN’s parameters. The equivalent of the *evidence.txt* table is usually the .tsv file with the shortest name in the folder. When in doubt, its precise name can be found by opening the log file in the folder and looking up the name of the file after the “*--out*” (= output) argument in the latest “*diann.exe*” call.

**3 – Post-processing**

***This folder contains the results of Post-processing (complementary information, tables, and plots).***

Detailed results of post-processing the *evidence.txt* (MaxQuant) or *REPORTNAME.tsv* (DIA-NN) file found in subfolder “a – *Search…”*. Contains several subfolders, and up to 3 protein groups-level Excel results tables. Each table’s first tab describes the types of columns you can find in the others. (There is also in subfolder *…/Tables/* a larger Excel table called *proteinGroups - Full.xlsx* with all of the rows and columns; these other 3 tables are thematic extracts from the information in that larger, single tab one.)

* *proteinGroups - Description.xlsx*: Contains all information relative to what the protein groups are (accessions, annotations etc…)
* *proteinGroups - Reg. analysis.xlsx*: Contains ones tab per statistical test. Row are re-*ordered* from high to low ratios for samples relevant for that test.
* *proteinGroups - GO terms.xlsx*: If applicable, one tab per GO term of interest showing protein groups annotated directly with it or with one of its offspring terms.

Current subfolders include:

* Summary plots: Plots which we create to monitor the quality of the MS runs.
* Workflow control: Plots created during data processing to check the quality of the data, its distribution, the behavior of normalizations, etc…
* Tables: All results table in .csv format are saved here. Excel versions of these, with formatting/tabs/etc… are saved in the main “2 – Analysis” folder for easier access.
* Amica: Protein-groups results table re-formatted for visualization/analysis into Amica, an interactive tool created by the Max Perutz Labs Proteomics Facility at the Vienna Biocenter: <https://bioapps.maxperutzlabs.ac.at/app/amica>
* Coverage: Sequence coverage plots for proteins of interest. Now also includes input for visualizing observed peptides overlapped over a 3D structure.
* Protein Groups sorted by and Profile plots for are different ways of looking at protein group relative and absolute abundance, coverage and spectral counts per sample or across all samples, respectively. *N.B.: if your experience includes TMT labelling, it is to be expected that coverage/spectra counts will be the same across all samples which were combined in a same TMT sample: that is, so to speak, the point of the whole thing.*
* Time profile plots: Only applicable if the experiment has a time dimension, plots of the behavior of protein groups across time.
* Pearson correlation map: Comparison of all samples using Pearson correlation.
* PCA and t-SNE plots: Two dimensionality reduction techniques, allowing the visualization the relationship between samples (PCA only) or protein groups (both).
* Reg. analysis: Regulation analysis (aka DEP, or Differential Expression of Proteins analysis). Usually the **most interesting folder**. Includes:
* t-tests: Volcano plots (x = log2(ratio), aka logFC; y = -log10(P-value)) for each moderated t-test performed (using *limma*). P-values are deemed significant if they exceed the threshold calculated using the Benjamin-Hochberg procedure for the highest (= most permissive) acceptable False Discovery Rate level (these are decided before the analysis, and are usually 10, 20 or 30%). Protein groups are deemed “regulated” if their P-value is significant AND their logFC is higher than (typically) 95% of control-to-control (nested replicates) or control-to-average-control (non-nested) logFCs.
* F-tests: Same general principle as above, except that an F-test is performed. The F-test allows the testing of several different single (e.g. treatments A and B vs ctrl) and double contrasts (e.g. [A vs ctrl] vs [B vs ctrl]) in a single test. In general, this is considered more powerful.
* SAINTexpress: Results from the SAINTexpress algorithm , specifically designed for analysis of pull-downs ([*http://saint-apms.sourceforge.net/Main.html*](http://saint-apms.sourceforge.net/Main.html)).
* Localisation: Optional, for datasets including subcellular fractions for each replicate/condition. Normalised protein cross-fraction profiles are compared across conditions and tested statistically using a one-way t-test.
* GO enrich: GO terms enrichment analyses (using *topGO*). The following analyses are made:
* t-tests/F-tests/localisation t-tests: For each contrast from each relevant test, a filter is created corresponding to regulated protein groups. GO term enrichment analysis is performed comparing the protein groups in each filters to all observed protein groups in the dataset. The P-value is calculated using Fisher’s exact test. On the bubble plots generated, the X-axis corresponds to the average logFC for that particular term for ALL proteins in the dataset, and thus represents the general trend for that particular term in the relevant sample.
* Dataset: This enrichment analysis is a comparison of the observed dataset versus the theoretical proteome of the parent organism. This analysis is useful to check for expected sample biases (e.g. tissue-specificity).
* Phospho (STY) / other PTM: For any post-translational modification which is expected to vary between samples, we include a folder named after the modification. The analysis is similar to that done for protein groups (t/F-tests, GO terms enrichment), but on the modified peptides table. By default, each peptide’s logFC is re-normalized to correct for the logFC of its parent protein.
* Clustering: Heatmaps of protein-groups level quantitative data with hierarchical clustering.
* Venn diagrams: If possible (sample number larger than 1 and not greater smaller than 5), Venn diagrams representing the overlap between protein groups defined as regulated (up-, down-, or both) by the filters defined in the regulation analysis (see *…/Reg. analysis* subfolder).
* STRINGdb: STRINGdb interaction network based on the filters defined in the regulation analysis (see *…/Reg. analysis* subfolder).
* Heatmaps and Coverage: If proteins “of interest” were defined prior to analysis, then a heatmap and a peptides coverage map are created for them.
* pRoloc: Optional, for samples including subcellular fraction. For each series of fraction, pRoloc analysis uses pre-defined compartment markers to predict the most likely protein localisation.