**Guide to the results folder**

***(dataset with replicates)***

Each output project folder is structured as three types of subfolders:

**1\_MS\_files**

Raw data acquired by the mass spectrometer.

*Keep these files preciously for future reference, re-analysis or publication!* It is essential to make the raw data available when publishing your results. The need to make supporting data publicly available is increasingly being recognized and enforced by scientific journals. See <https://www.proteomexchange.org/> for the most common repositories for proteomics data.

**2\_“SEARCH-ENGINE-NAME”\_search**

The raw data was searched in one of several search software (currently, DiaNN [[1]](#endnote-1), FragPipe [[2]](#endnote-2) and MaxQuant [[3]](#endnote-3) are supported. It is important to keep these files, not just the raw files, as some repositories also require search results to be uploaded. *Please keep these files for future reference/re-analysis/publication!*

For post-processing the output of search engines (see next section), we use the most basal level of output, i.e. the identifications tables, also known as “PSMs” (peptide-spectrum matches):

* For DiaNN, these are contained in a report.tsv or report.parquet file. We also need the report.log.txt file to read search parameters.
* For FragPipe, we need the output folder: we read the workflow and samples manifest files which FragPipe saves into the output folder, and use the information to parse the psm.tsv output file(s).
* For MaxQuant, we use the evidence.txt file (which contains PSMs per se as well as indirect, match-between-runs based identifications) and also parse parameters from the mqpar.xml file.

**3(.#?)\_Post\_processing\_YYYY\_MM\_DD**

Detailed results of post-processing the *PSMs table*. *If several analyses were delivered, there are several folders with incremental list numbers and analysis date, e.g.:*

*…/3.1\_Post\_processing\_2025\_03\_25*

***In each such folder, the main Excel report table is located in sub-folder …/Tables!***

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 correct behaviour of normalizations, etc…
* Tables: All result tables are saved here, including the main Excel report.
* Protein plots: If applicable, these include, for proteins of interest:
  + sequence coverage plots,
  + plots of correlation between samples,
  + ratio plots, where the sequence ratio of intensity between sample pairs is plotted for each observed part (peptide) of the protein sequence.
* Heatmaps and Coverage: Sequence coverage plots and heatmaps for proteins of interest. A*lso includes input files for visualizing observed peptides overlapped over a 3D structure.*
* Sorting plots and Profile plots are different ways of looking at protein group abundance, coverage or spectral counts per sample or across all samples, respectively[[4]](#footnote-1).
* Pearson correlation map: Comparison of all samples using Pearson correlation.
* Dimensionality red. plots: dimensionality reduction plots such as PCA, t-SNE or U-MAP.
* Clustering: Heatmaps with hierarchical clustering.
* Reg. analysis: DEP (Differential Expression of Proteins) analysis. Usually the **most interesting folder**. Volcano plots (x = log2 fold change, y = unmodified -log10 P-value) are drawn for several different tests. Un-corrected 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 (decided before the analysis, usually 20%). Entities are deemed “regulated” if their P-value is significant AND their logFC is higher than either a fixed value (e.g. +/- 1 log2) or a fixed percentage of inter-sample group fold changes (e.g. 5%).
  + t-tests: by default a moderated t-tests from *limma* [[5]](#endnote-4) is used for downstream analysis, but a Welch’s t-tests, a permutations t-tests or the modified version of *limma*’s from package *DEqMS* [[6]](#endnote-5) are also run. The decision as to which test to used is informed by a comparison of the relative statistical power of the tests performed.
  + We also run LRT and ODP tests from package *edge [[7]](#endnote-6)*.
  + We also perform if relevant ANOVA analysis using *limma*’s moderated F-test. In this case, the decision as to what is regulated is based off *limma*’s decideTests() function.
  + For pull-downs, the SAINTexpress algorithm can also be run ([*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. Normalized protein cross-fraction profiles are compared across conditions and tested statistically using a one-way Welch’s t-test.
  + GO enrich: GO terms enrichment analyses using *topGO [[8]](#endnote-7)* and CytoScape’s *clueGO* [[9]](#endnote-8) add-on. The following analyses are made:
    - For each statistical test, a filter is created corresponding to regulated protein groups for each contrast. Filters also include all-or-nothing responders. GO term enrichment analyses are then performed, comparing protein groups in each filters to all observed protein groups in the parent dataset or sub-dataset.
    - 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, but on the modified peptides table. By default, each peptide’s logFC is re-normalized to correct for the logFC of its parent protein.
* Amica: Protein-groups results table re-formatted for visualization/analysis into Amica [[10]](#endnote-9), an interactive tool created by the Max Perutz Labs Proteomics Facility at the Vienna Biocenter: <https://bioapps.maxperutzlabs.ac.at/app/amica>
* Time profile plots: Plots of changes in protein groups expression across time, only applicable if the experiment has a time dimension.
* Venn diagrams: If possible (sample groups number larger than 1 and not greater than 7), Venn diagrams representing the overlap between protein groups or modified-peptides defined as regulated (up-, down-, or both) by the filters defined in the regulation analysis (see *…/Reg. analysis* subfolder).
* STRINGdb: STRINGdb [[11]](#endnote-10) interaction networks based on the filters defined in the regulation analysis (see *…/Reg. analysis* subfolder).
* pRoloc: Optional, for samples including subcellular fraction. For each series of fractions, *pRoloc [[12]](#endnote-11)* analysis uses pre-defined compartment markers to predict the most likely protein localization.

1. Full references: <https://github.com/vdemichev/DiaNN?tab=readme-ov-file#key-publications> [↑](#endnote-ref-1)
2. Full references: <https://github.com/Nesvilab/FragPipe> [↑](#endnote-ref-2)
3. Full references: <https://www.maxquant.org/Publications/> [↑](#endnote-ref-3)
4. *If your experience includes TMT labelling, it is to be expected that coverage/spectral 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.* [↑](#footnote-ref-1)
5. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47. [↑](#endnote-ref-4)
6. Zhu Y (2024). DEqMS: a tool to perform statistical analysis of differential protein expression for quantitative proteomics data. [↑](#endnote-ref-5)
7. Storey JD, Leek JT, Bass AJ (2024). edge: Extraction of Differential Gene Expression. [↑](#endnote-ref-6)
8. Alexa A, Rahnenfuhrer J (2024). topGO: Enrichment Analysis for Gene Ontology. [↑](#endnote-ref-7)
9. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009 Apr 15;25(8):1091-3. doi: 10.1093/bioinformatics/btp101. Epub 2009 Feb 23. PMID: 19237447; PMCID: PMC2666812. [↑](#endnote-ref-8)
10. Didusch S, Madern M, Hartl M, Baccarini M. amica: an interactive and user-friendly web-platform for the analysis of proteomics data. BMC Genomics. 2022 Dec 9;23(1):817. doi: 10.1186/s12864-022-09058-7. PMID: 36494623; PMCID: PMC9733095. [↑](#endnote-ref-9)
11. See all STRINGdb references here: <https://string-db.org/cgi/about?footer_active_subpage=references> [↑](#endnote-ref-10)
12. Gatto L, Breckels LM, Wieczorek S, Burger T, Lilley KS (2014). “Mass-spectrometry based spatial proteomics data analysis using pRoloc and pRolocdata.” Bioinformatics.

    Breckels LM, Gatto L, Christoforou A, Groen AJ, Lilley KS, Trotter MW (2013). “The effect of organelle discovery upon sub-cellular protein localisation.” J Proteomics.

    Gatto L, Breckels LM, Burger T, Nightingale DJ, Groen AJ, Campbell C, Mulvey CM, Christroforou A, Ferro M, Lilley KS (2014). “A foundation for reliable spatial proteomics data analysis.” Mol Cell Proteomics.

    Breckels LM, Holden S, Wonjar D, Mulvey CM, Christoforou A, Groen A, Trotter MW, Kohlbacker O, Lilley KS, Gatto L (2016). “Learning from heterogeneous data sources: an application in spatial proteomics.” PLoS Comput Biol.

    Breckels LM, Mulvey CM, Lilley KS, Gatto L (2016). “A Bioconductor workflow for processing and analysing spatial proteomics data.” F1000Research.

    Crook OM, Breckels LM, Lilley KS, Kirk PD, Gatto L (2019). “A Bioconductor workflow for the Bayesian analysis of spatial proteomics.” F1000Research. [↑](#endnote-ref-11)