Appendix for "A Bioconductor workflow for processing, evaluating, and interpreting expression proteomics data"

Charlotte Hutchings, Cambridge Centre for Proteomics, University of Cambridge, UK

Appendix

This Appendix accompanies the paper "A Bioconductor workflow for processing, evaluating, and interpreting expression proteomics data" by Hutchings et al, submitted to F1000Research in August 2023. Associated data can be found on Zenodo at http://doi.org/10.5281/zenodo.7837375 and also in the Github repository https://github.com/CambridgeCentreForProteomics/f1000_expression_proteomics/.

Identification search with Proteome Discoverer

The use-case data analyzed in this workflow was initially processed using Proteome Discoverer version 2.5. Whilst much of the identification and quantification takes place out of sight of the user, Proteome Discoverer incorporates several user-defined search parameters which must be specified according to the sample preparation methods and MS instrumentation used. There is also the option to apply both basic and advanced data filtering parameters during the search. Users must be aware of these parameters as they will directly influence the data output and downstream processing.

Whilst an in-depth discussion of identification searches is outside of the scope of this workflow, a few key parameters are discussed to put the data into context. During sample preparation, TMT-labelled cell pellets were combined and separated into 8 fractions using a Pierce High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific). After being analyzed by MS, the 8 resulting raw files were uploaded to Proteome Discoverer 2.5 and processed using a single processing and consensus workflow. LFQ supernatant fractions were each analyzed on a separate mass spectrometry run resulting in 6 raw files. These files were imported into Proteome Discoverer with each sample having its own independent processing step followed by a single multi-consensus step. All processing and consensus workflow templates are provided in the supplementary materials.

For both TMT and LFQ workflows, SequestHT was selected as the search engine and trypsin specified as the enzyme used for proteolytic digestion. Since the digestion was carried out overnight with a 1:20 w/w ratio of trypsin:protein, digestion was expected to be complete and a low threshold of 2 missed cleavages was allowed. For MS analysis, a Fourier Transform orbitrap with a resolving power of 120,000 m/z was used as the mass analyzer for precursor ion mass, and a linear ion trap was used to measure fragment ion mass. This information determined the thresholds for precursor and fragment mass tolerances, two key parameters for the identification search. The precursor mass tolerance determines which mass range of peptide sequences are considered for each observed spectrum, whilst the fragment mass tolerance specifies how similar the observed and theoretical peptide fragment spectra should be for a match. If these tolerances are too narrow then the correct peptide sequence may be omitted and true positives are lost. However, if thresholds are set too wide then incorrect peptide sequences are considered and false positives arise. Based on the instrumentation used in this experiment, standard mass tolerances of 10 ppm and 0.5 Da were allowed for precursors and fragments, respectively. Given the intrinsic variability of LFQ between MS runs, RT alignment was used for the label-free samples with a 10-minute retention time window.

In addition to the parameters based on the experimental protocol, we also applied some basic non-specific filtering. We only retained high confidence PSMs from the identification search. Such filtering is necessary because only a fraction of the PSMs outputted by any given search engine will be genuine matches, or true discoveries, whilst the remainder are incorrect false discoveries. To deal with this problem, PSM confidence level (high, medium or low) is determined via the Proteome Discoverer Percolator node (Käll et al. 2007) which estimates each PSM's false discovery rate (FDR). The raw spectra are searched against the database of interest as well as a decoy database containing randomised peptide sequences, often generated by shuffling or reversing the original peptide sequences. False discovery rate is then defined as the proportion of total PSMs that are matched to the decoy database, and, therefore, are known false discoveries. This is done for all spectra and we considered a PSM to be of 'high confidence' if it had a false discovery rate <1 %, 'medium confidence' if <5 %, and 'low confidence' if the false discovery rate exceeded 5 %. Only PSMs annotated as high confidence were kept.

Whilst the basic filtering steps completed during this identification search could just have easily been carried out in R using the `SummarizedExperiment` and `QFeatures` infrastructure, applying them here saves time later on and reduces the burden of storing large data files. These steps are also relatively standard and non-specific so we do not need to assess the data prior to their implementation. However, Proteome Discoverer also provides the option to carry out more in-depth filtering through the use of parameters such as the SPS Mass Match %, co-isolation interference % and signal-to-noise thresholds. We advise against implementing such filtering at this stage since decisions regarding thresholds will likely be influenced by the quality of data output, as demonstrated later in this workflow. Instead, thresholds for the three aforementioned parameters were set to 0 during the identification search.

Using this workflow with MaxQuant data

This workflow was written for proteomics data processed using the Proteome Discoverer software. Nevertheless, the workflow and basic principles discussed are also applicable to the output of any similar proteomics raw data processing software, including MaxQuant. Below we outline the differences to be aware of when following this workflow using MaxQuant output text files. The code as written will require some minor modifications to work properly with MaxQuant formatted data.

1. The rough equivalent of the PSMs.txt file output by Proteome Discoverer is the evidence.txt file output by MaxQuant.
2. Decoy PSMs (known false discoveries which are used to calculate false discovery rate) are automatically filtered out by Proteome Discoverer, but this is not the case with MaxQuant. Hence when working with MaxQuant outputs it is important to filter out rows with '+' in the `Reverse` column.
3. Equivalent column names and the type of data contained are described here. Ellipses are put where there no equivalent column exists. (PD PSMs.txt column = MaxQuant evidence.txt column):

- `Abundance` (float) = `Reporter.intensity.corrected` (integer)
- `Sequence` (string) = `Sequence` (string)
- `Master.Protein.Accessions` (string) = `Leading.proteins` (string)
- `Master.Protein.Descriptions` (string) = ...
- `Contaminants` (string, True or False) = `Potential.contaminant` (string, + or blank)
- ... = `Reverse` (string, + or blank)
- `Rank` (integer) = ...
- `Search.Engine.Rank` (integer) = ...
- `PSM.Ambiguity` (string) = ...
- `Number.of.Protein.Groups` (integer) = ... (you might calculate this by counting the number of ; in the Leading.proteins column and adding 1)
- `Average.Reporter.SN` (float) = ... (you might calculate the average reporter ion intensity and threshold based on that instead)

- `Isolation.Interference.in.Percent` (float) = `PIF` (float, to get the data in exactly the same format you have to calculate (1 - PIF)*100)
- `SPS.Mass.Matches.in.Percent` (integer) = ...

(PD Proteins.txt column = MaxQuant proteinGroups.txt column):

- `Accession` (string) = `Majority.protein.IDs` (string)
- `Protein.FDR.Confidence.Combined` (string; High, Medium, or Low) = `Q.value` (float, a Proteome Discoverer protein FDR of 'High' is equivalent to a Q.value < 0.01)

References

Käll, Lukas, Jesse D Canterbury, Jason Weston, William Stafford Noble, and Michael J MacCoss. 2007. "Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets." *Nature Methods* 4 (11): 923–25. https://doi.org/10.1038/nmeth1113.