**Capseome Project Team 2**

1. **Introduction**

This report details the processing and analysis of proteomics data from the PRIDE dataset **PXD061542**, which is part of the project *"SOX9-dependent fibrosis drives renal function in nephronophthisis" [1]*.

Briefly, the study investigates the molecular mechanisms underlying renal fibrosis, a pathological feature and important outcome prediction marker across several kidney diseases. More importantly, there is a marked correlation between the severity of fibrosis and kidney function decline, yet the molecular mechanisms underlying fibrosis are not yet understood.

The authors employed a mouse model where Fbxw7, a recognition receptor that controls the protein level of many target genes, therefore being crucial to cellular proteostasis and regulation of many basic cell functions, is deleted in the renal epithelium. Fbxw7 deletion in the mouse kidney recapitulates a juvenile-adult nephronophthisis (NPHP) pathology, a genetic disorder characterized by extensive kidney fibrosis and cyst formation. The study showed that following Fbxw7 knock-out, induction of the Sox9-Wnt axis is an important driver of renal fibrosis.

The dataset comprises mass spectrometry-based proteomic analyses of Fbxw7-null mIMCD-3 cell lines (isolated from the kidney of an adult mouse), aiming to identify proteomic changes associated with the loss of FBW7. They followed a bottom-up approach, using data-independent acquisition (DIA), where they processed the total protein from each sample. Briefly, the total protein suspension was reduced, alkylated and purified, digested with trypsin and separated by reverse phase. Eluted peptides were ionised by electrospray followed by mass spectrometric analysis in an Orbitrap Exploris 480 mass spectrometer. The resulting data were deposited in the PRIDE repository, a component of the ProteomeXchange consortium that facilitates the sharing of proteomics data.

1. **Dataset Acquisition and Preparation**

The dataset was accessed using the rpx R package. While the standard workflow specified the use of .mzID files, none were provided in the PRIDE repository for this project. Instead, the available .mzML raw files were downloaded and processed locally using **SearchGUI**, a widely used search engine frontend for proteomics data analysis.

Out of twelve .mzML files, **two were found to be corrupt and could not be analyzed**. The remaining ten were successfully processed in SearchGUI, and search results were exported as .mzID files.

Due to the large size of the .mzID files (approximately 1 GB each), direct processing in R was impractical. To address this, the .mzID files were converted to .csv format using **Python in an Anaconda environment**, enabling lightweight and flexible downstream analysis in R.

Additionally, to allow for a direct comparison with the publication results and to compare differences between two peptide and protein identification approaches, we used EncyclopeDIA. EncyclopeDIA is a library search engine for peptide identification that counts with several algorithms for DIA data analysis [2]. Briefly, these kinds of tools search the spectra of peaks against a FASTA database containing protein sequences to match peaks and peptides and/or proteins. EncyclopeDIA allows for the use of various workflows. In this project, we generated an empiricarlly-corrected DIA chromatogram library [3](.elib format) from the raw .mzML files available; together with a Prosit-generated library for *Mus musculus* that is used as scaffold for the analysis, and the corresponding protein FASTA file. The software first creates the chromatogram library, which is then used for peptide and protein identification.

The software outputs several graphs that can be used to assess the quality of the analysis. Here are some examples from one of the runs (we first run each sample independently):



Figure Delta RT from library. The histogram shows the difference in retention time (RT) between the observed peptides in the sample and the expected RTs from the spectral library. The x-axis represents the D(RT) in minutes, while the y-axis shows the frequency.

Figure & 2: MS1 & MS2 Mass error (PPM): mass accuracy of the fragment ions matched to the spectral library. In the x-axis, the mass error in parts per million.

The y-axis displays the relative frequency of the mass errors.



Due to the nature of the experiment, the quantification of peptides and proteins is more challenging. Briefly, in a DIA experiment the quantification engine selects the best ion to quantify. Hence, it does not make sense to perform quantitation per sample because the engine might select a different ion in each of the samples. Meaning, the quantification results are only generated by integrating the previous analysis and generating a “shared” quantitation report that could then be used quantitatively. Unfortunately, because of the heavy nature of the analysis, we did not have time to generate the integrated quantification results.

1. **PSM Object Creation and Preprocessing**

Searchgui output:

The ten .csv files were read and merged into a unified PSM dataset using R. These files had already been filtered by SearchGUI to include only confident matches (non-decoy, rank 1). The combined dataset was converted into a tibble (idtbl) for downstream manipulation.

An additional filtering step was initially applied to remove ambiguous PSMs (spectrum IDs matching multiple peptides). However, this step removed nearly all data and was therefore excluded, as ambiguity resolution had likely been handled during the initial SearchGUI processing.

EncyclopeDIA output:

EncyclopeDIA provides the raw match-level data between the DIA spectra and the library in the features.txt files – which is basically already akin to the PSM object we have seen in the course. There are some differences because of the way EncyclopeDIA searches the library and produces the output, and because our data comes from a DIA experiment:

* The decoy hits are provided in a separate file.
* The spectra are not automatically assigned a rank (because of the way that the matching is done when using a DIA). To circumvallate this, we manually calculate a rank by grouping the results by scan and ranking them according to the HyperScore value.
* The identified peptides and proteins are already corrected for FDR < 0.01 (EncyclopeDIA outputs a separate result file where the q-score (or FDR) can be checked.

The combined feature file for all samples was then filtered for rank = 1 (in other words, selecting the peptide with higher score, we used HyperScore forthis), and the decoy hits were removed from the data frame.

1. **Identification Summary**

|  |  |  |
| --- | --- | --- |
|  | **Searchgui** | **EncyclopeDIA** |
| **Total PSMs (after filtering)** | 1,505,627 | 105,600 |
| **Unique peptide sequences** | 18,687 | 48,672 |
| **Unique protein accessions** | 11,860 | 12803 |
| **Unique razor proteins** |  | 11790 |

These results indicate a high-confidence and deep coverage of the proteome in the analysed dataset.

1. **Comparison with previous results**

The authors provide a summary of their analysis in an .mztab file.