**Capseome Project Team 2**

1. **Introduction**

This report details the processing and analysis of proteomics data from the PRIDE dataset **PXD061542**, titled *"SOX9-dependent fibrosis drives renal function in nephronophthisis"* . The dataset originates from a study investigating the molecular mechanisms underlying nephronophthisis, a genetic disorder characterized by kidney fibrosis and cyst formation. Specifically, the research focused on the role of the transcription factor SOX9 in driving fibrotic processes that impair renal function.

The dataset comprises mass spectrometry-based proteomic analyses of FBXW7 knock-out IMCD-3 cell lines, aiming to identify proteomic changes associated with the loss of FBXW7. The samples were processed using an Orbitrap Exploris 480 mass spectrometer, and 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.

1. **PSM Object Creation and Preprocessing**

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.

**4. Identification Summary**

The final dataset consisted of:

* **Total PSMs**: 1,505,627
* **Unique peptide sequences**: 18,687
* **Unique protein accessions**: 11,860

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