

With the growing prominence of single-cell techniques across various omics fields, there is a pressing need to develop a standardized pipeline for proteomics data in the realm of systems biology. Unlike DNA sequencing and RNA sequencing, proteomics analysis via mass spectrometry incurs high costs in terms of labor and equipment. Additionally, commercially available software solutions often come with hefty price tags and limited transparency regarding the underlying methods employed. However, MaxQuant presents a promising alternative, complemented by the flexibility using the R programming language. Introducing ProteoScanR, a state-of-the-art proteomics pipeline integrated into the user-friendly Proteomics Workbench interface. Guided by SCoPE2 , the development process of ProteoScanR was thoroughly tested and validated using both bulk and single-cell mass spectrometry data sets. The pipeline implementation in the Proteomics Workbench leverages the power of an interactive environment built in R Shiny, empowering users to discover valuable insights for their specific data set. Within the interactive environment, users have the flexibility to customize cutoffs and thresholds for quality control, as well as employ various approaches for data transformation, normalization, missing value imputation, and batch correction. ProteoScanR and the Proteomics Workbench serves as a valuable tool in guiding the identification of expressed proteins in cells under study. Subsequently, the pathway enrichment analysis provides additional biological contexts for a comprehensive understanding of their functional implications. The master thesis project serves as a foundation for future advancements in the field of single-cell proteomics. Moreover, the codebase has been designed with robustness and scalability in mind, ensuring ease of maintenance and future expansion of the application. The source code is accessible and can be located at the following URL: <https://github.com/Lukas67/ProteoScanR>

The central dogma underpins genetic information’s transformation into functional proteins . Proteins have various functions, such as providing structural integrity, catalyzing chemical reactions, and regulating cellular functions . Omics fields like genomics, transcriptomics, proteomics, and metabolomics provide an holistic view of biomolecules. These proteins are comprehensively studied through proteomics to unveil intricate biological networks.

Higher organisms are composed of specialized cells organized into tissues, such as skin, muscle, and blood. Each tissue consists of cells with specific functions, resulting in variations in protein expression. Bulk proteomics is a technique used to analyze the protein composition of a sample, which in case of a tissue contains various cell types. Taking tissue samples can lead to an averaging effect across the entire cellular ensemble, making it difficult to discern specific cell types. To overcome this limitation, cell sorting techniques were employed and enabled targeted single-cell proteomics (SCP) . Single-cell proteomics reflects the protein ensemble of a specific cell type at a particular time, providing a focused perspective on the studied field compared to bulk methods .

Mass spectrometry enables qualitative and quantitative analysis of the entire repertoire of a biological sample. Mass spectrometers measure the mass to charge ratio (m/z) of charged particles. Given the high resolution of MS data, algorithms are employed to convert the raw signal into an interpretable form. Software packages like MaxQuant are commonly used to process the data, providing it for further analysis and statistical testing.

With advancements in mass spectrometry technology and computational methods, proteomic analysis has emerged as a powerful tool for investigating complex protein samples. However, analyzing proteomic data poses challenges in data processing, statistical analysis, and interpretation. This thesis aims to address these challenges by exploring computational methods for downstream analysis of proteomics data. It is important to acknowledge that downstream analysis in proteomics lacks a standardized approach, and the selection of computational methods depends on the specific dataset and research objectives. ProteoScanR comprises a series of steps to streamline the data in an interactive environment, called Proteomics Workbench. This user-friendly interface will enable researchers, including those with limited computational expertise, to navigate and comprehend the data effectively.

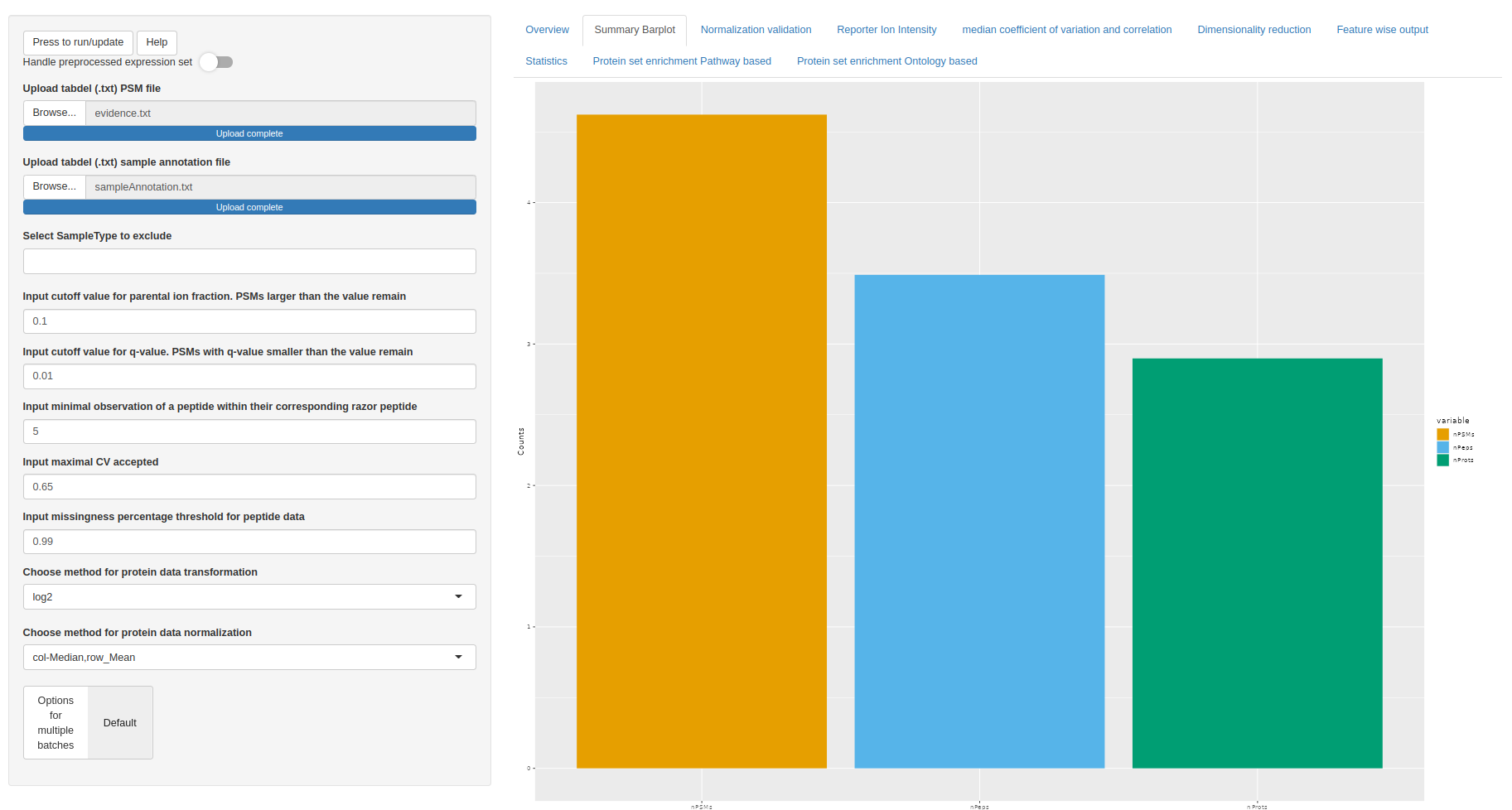


Figure 1: Screenshot of proteomics workbench. Summary barplot showing the number of peptide spectrum matches, peptides and proteins

The ProteoScanR pipeline is employed in the proteomics workbench. Users find settings for the pre-processing on the left hand side of the interface (see figure 1). The main panel shows the data in different aspects and helps the user finding a good fit for the methods applied to their individual dataset.

Beforehand normalization was tested with applying logarithmic transformation to the base 2 and performing column median row mean normalization (not shown in figure). The mutual information (MI) within the healthy control (HC) group’s sample type exhibited a reduction towards the lower edge of the interquartile range (IQR).

After performing a logarithmic transformation to the base 2 and quantile normalization, the mutual information (MI) within the healthy control (HC) group’s sample type exhibits a minor reduction in the interquartile range (IQR). However, it should be noted that there is an increase in the number of outliers compared to the raw data. When switching to the quantile normalization the loss of mutual information between sample types decreased. Therefore the advised normalization technique for this particular data set would be the quantile normalization compared to the default method .