FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

Design and implementation of analysis pipeline for single cell type proteomics data

presenting ProteoScanR & Proteomics Workbench

Lukas Gamp

Main Supervisor: Dr. DI(FH) Gerhard Duernberger
External Supervisor: Assoc. Prof. Ujjwal Neogi, M.Sc. PhD
Second examiner: FH-Prof. Dr. Alexandra Graf

Abstract

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 (Cox & Mann 2008) 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 (Specht et al. 2021, Petelski et al. 2021, Vanderaa & Gatto 2021), 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

This project was performed at the systems virology lab at the Karolinska Institutet in Stockholm.

1 Introduction

The central dogma of molecular biology serves as the cornerstone of biological processes, providing a framework for understanding how genetic information is converted into functional proteins (Cobb 2017). Over the past six decades, researchers have explored various omics fields, such as genomics, transcriptomics, proteomics, and metabolomics, to comprehensively analyze biomolecules in diverse contexts. Omics research offers a holistic view of biological processes and regulatory mechanisms, and identify key players in complex biological networks using computational methods. Proteins have various functions, such as providing structural integrity, catalyzing chemical reactions, and regulating cellular functions (Karahalil 2016). Proteomics is a comprehensive approach that investigates the complete protein composition of a specimen, aiming to understand the intricate biological network it represents.

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) (Liou et al. 2015, Sutermaster & Darling 2019). 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 (Maes et al. 2020).

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 a charged particle from a larger molecule. Given the high resolution of MS data, algorithms are employed to convert the raw signal into an interpretable form. Software packages like MaxQuant (Cox & Mann 2008) 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.

2 Materials and methods

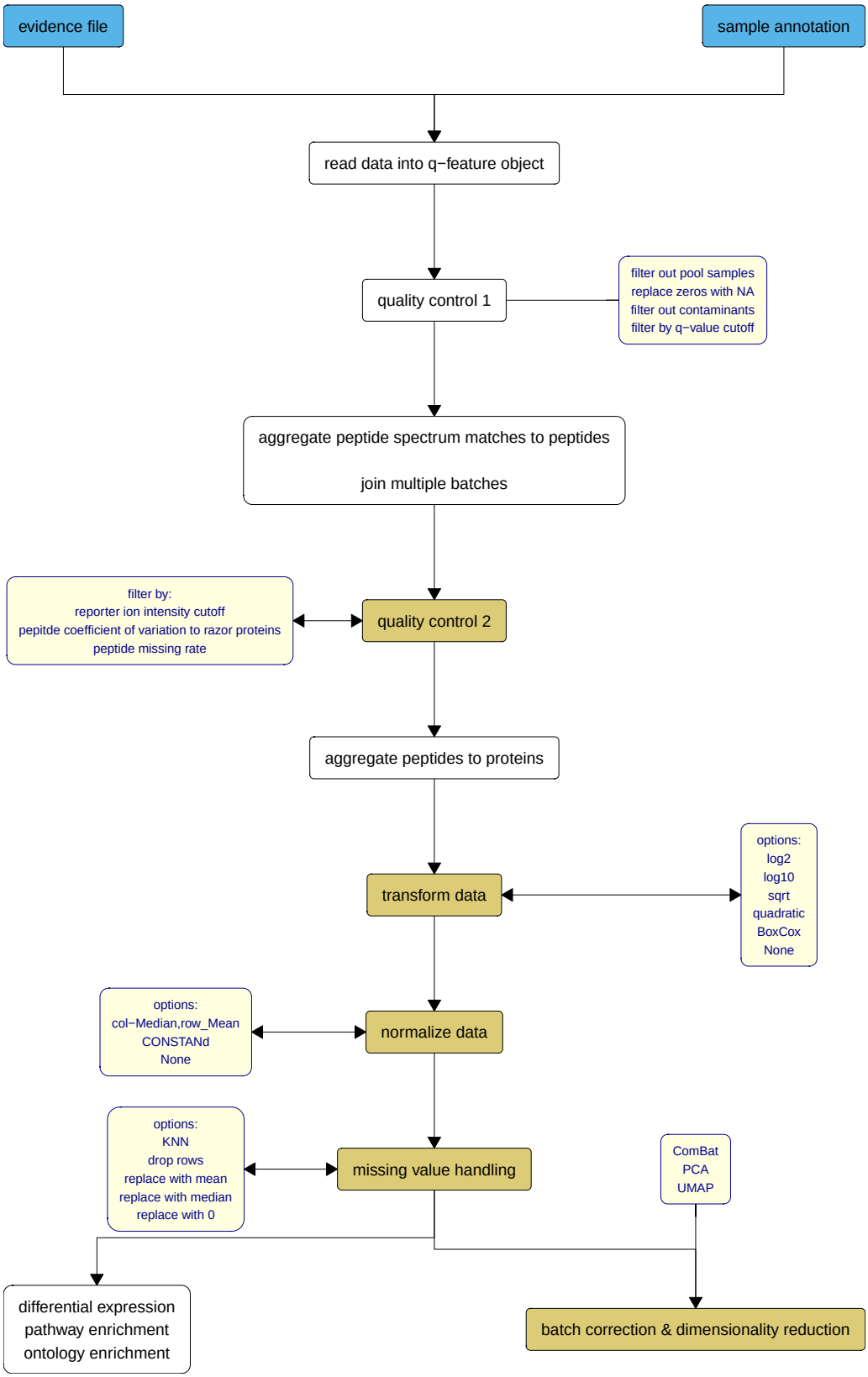

Figure 1: Flowchart presenting the processing by ProteoScanR. Gold: techniques which involve explanatory data analysis. Yellow: Adjustable features. Blue: Inputs

The flowchart (figure 1) describes the processes involved in the ProteoScanR pipeline. ProteoScanR initially filters the data by false discovery rate (FDR) converted to q-value (provided by MaxQuant (Cox et al. 2011)), according to precursor ion fraction (PIF) (Tannous et al. 2020, Specht et al. 2021) and subsequently aggregates peptide spectrum matches to peptides.

The quality control continues with calculating median reporter ion intensity (RI), median coefficient of variation (CV) and filtering according to cutoffs. Subsequently peptides with high missing rate are removed and peptides are aggregated to proteins.

Preparation for hypothesis testing involves data transformation and normalization. Missing values for proteins in individual channels are handled according to user selection with methods such as the K-nearest neighbor algorithm (=KNN) (Lan et al. 2013). Before statistical analysis, dimensionality reduction techniques can be applied and observed with or without correcting for batch effect by ComBat (Johnson et al. 2007). Furthermore data processing can be validated with an entropy based approach for the conservation of information.

Statistical testing involves differential expression analysis with the R package Limma (Phipson et al. 2016). Hypothesizes can be selected by the user and p-value correction can be employed with different approaches to meet conservative as well as liberal study designs.

After finding differentially expressed proteins, enrichment elucidates them in biological context by finding over represented proteins. The protein enrichment analysis can be employed in selected contexts such as pathways, cell-types or custom ontologies. The pathway enrichment analysis is performed using the R Bioconductor package clusterProfiler (Wu et al. 2021). The analysis utilizes the entire "Kyoto Encyclopedia of Genes and Genomes" (KEGG) database, which is accessed through a function that maps the UniProt IDs to corresponding pathways. For custom ontologies the protein enrichment analysis is done using the R bioconductor package piano (Väremo et al. 2013). In opposite to the pathway enrichment, piano maps proteins to a pre-selected gene collection set, which can be either downloaded from gsea-msigdb.org or custom made for the particular experiment and explains conditions, phenotype or other desired properties.

3 Results

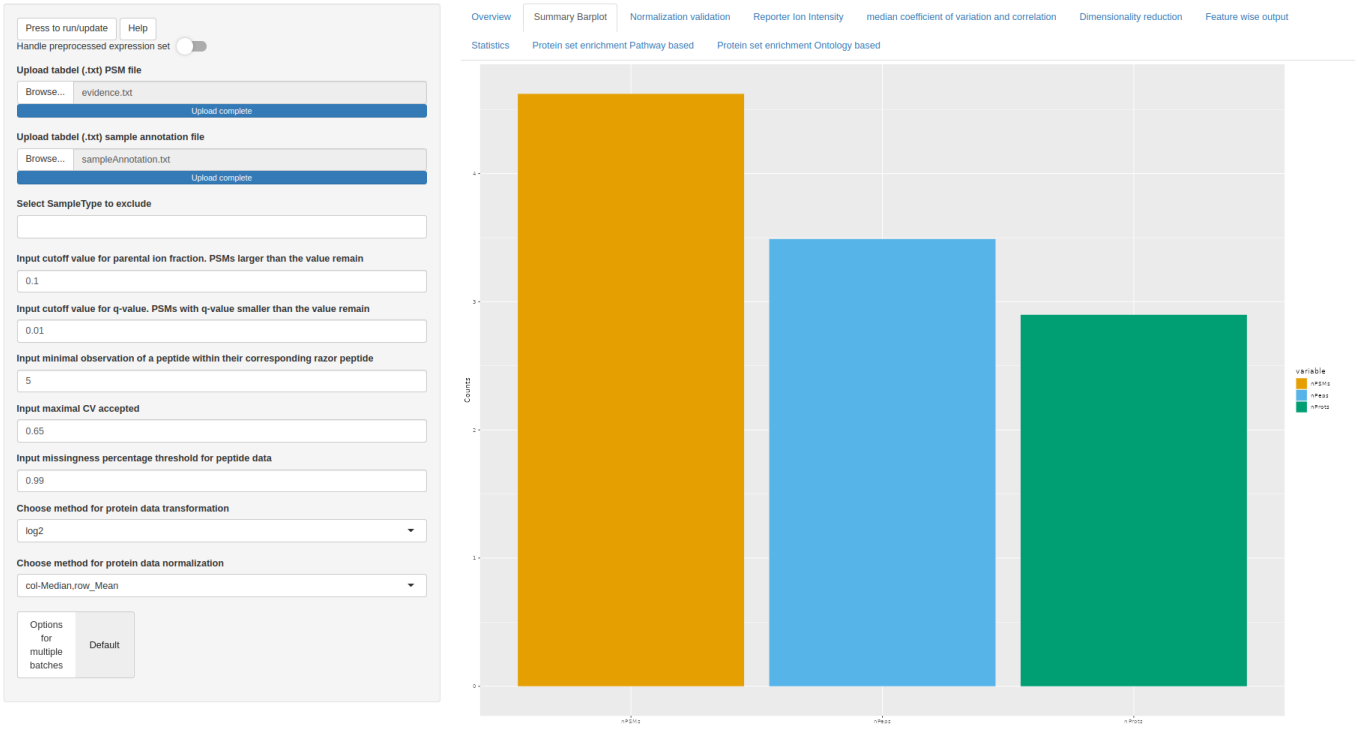

Figure 2: Barchart showing the number of peptide spectrum matches, peptides and proteins

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 2). 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.

As seen in figure 2 the aspects shown in the main panel include:

- Overview showing the individual processing steps of the analysis pipeline
- Barchart showing the number of peptide spectrum matches, peptides and proteins
- Mutual information intra- and inter- groupwise
- Reporter ion intensity with factor selection to highlight
- Median coefficient of variation for razor proteins and correlation heatmap for protein expression
- Dimensionality reduction analysis with principle component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP)
- Feature wise output for an individual protein over the course of the pipeline.
- Statistics module
- Pathway enrichment
- Protein set ontology enrichment

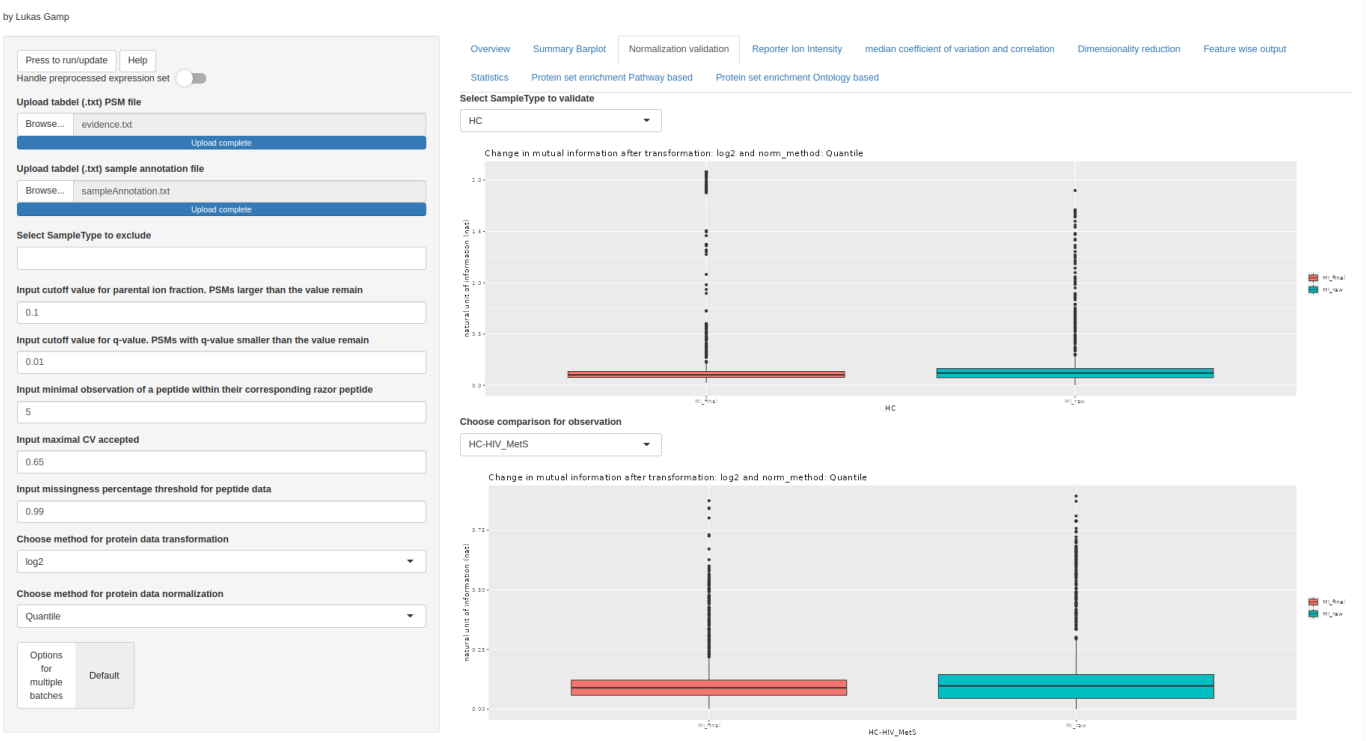

Figure 3: Mutual information with quantile normalization

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) (as seen in figure 3). 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 (as seen in figure 3). Therefore the advised normalization technique for this particular data set would be the quantile normalization compared to the default method.

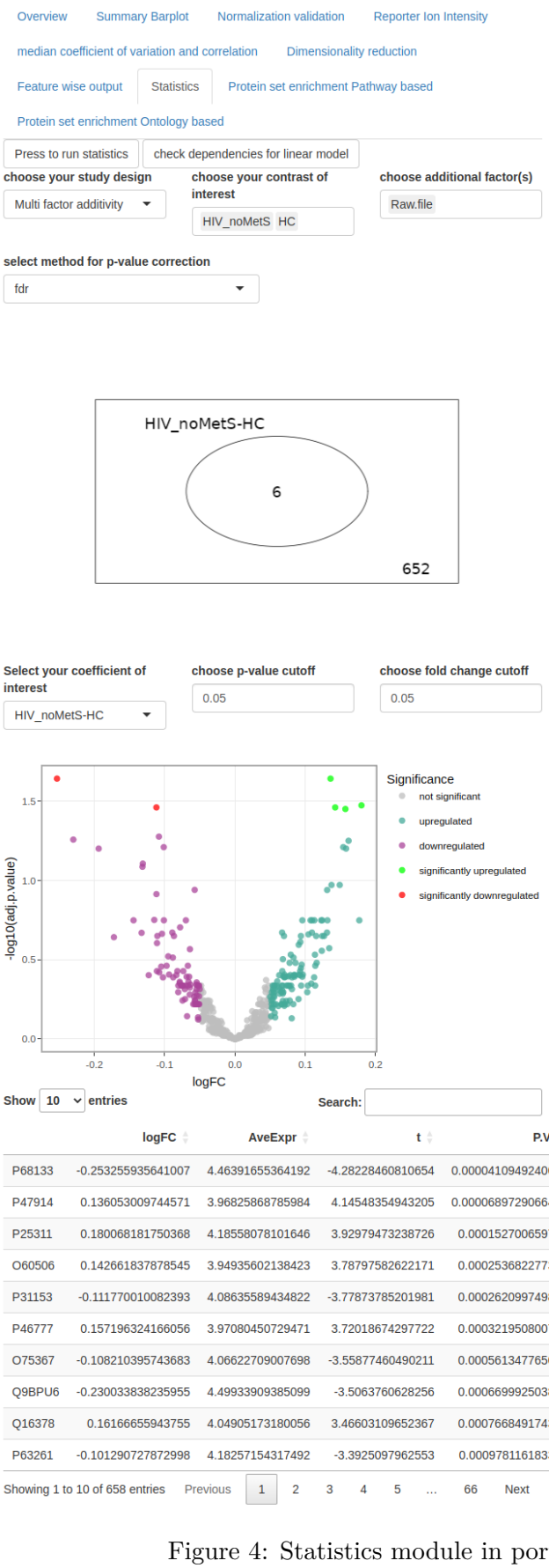

Figure 4: Statistics module in portrait mode

By comparing the log10-transformed and quantile-normalized expression values between the HIV group without metabolic syndrome and the healthy control (HC) group (see figure 4), six differentially expressed proteins were identified. The Venn diagram indicates significantly up/down regulated proteins. The Volcano represents -log10(adjusted p values) against log2 fold changes (logFC). These proteins include P68133 (Actin, alpha skeletal muscle), P47914 (60S ribosomal protein L29), P25311 (Zinc-alpha-2-glycoprotein), O60506 (Heterogeneous nuclear ribonucleoprotein Q), P31153 (S-adenosylmethionine synthase isoform type-2), and P46777 (60S ribosomal protein L5). The remaining 652 proteins have been identified as not significantly up/down regulated in the comparison analysis. These proteins did not exhibit statistically significant differences in expression between the compared groups or conditions.

4 Discussion and conclusion

The Proteomics Workbench interface and the ProteoScanR pipeline demonstrated how interactive engagement with the data not only enhances the experience of biologists but also improves the comprehension of the underlying significance of a biological dataset. By utilizing an entropy-based visualization approach, the conservation of information can be validated, allowing users to select appropriate methods and adjust thresholds, cutoffs, and techniques accordingly. With the selection of factors to indicate over multiple plots, biases can be identified on various levels. Additionally, the analysis can be examined on an individual protein basis, enabling the identification of suspicious results and closer validation of significant candidates. The statistics module provides users with a simple tool to assess and visualize the results, offering clarity to dense scatter plots through hover functions. Enrichment mapping links statistically significant proteins to either the KEGG pathway database or individual research data sets, depending on user selection. These final results place the data in a biological context, which can be explored interactively in graphical visualizations.

In conclusion, the ProteoScanR pipeline and Proteomics Workbench interface provide a user-friendly and efficient approach for proteomic data analysis, enabling biologists to address their research questions and validate computations. Future perspectives include deploying the application on a server for global accessibility, leveraging artificial intelligence techniques, and implementing clustering approaches to further enhance annotation and understanding of cellular heterogeneity and functional characteristics.

5 Bibliography

Cobb, M. (2017), '60 years ago, francis crick changed the logic of biology', *PLOS Biology* **15**, e2003243.

Cox, J. & Mann, M. (2008), 'Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification', *Nature Biotechnology* **26**, 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. (2011), 'Andromeda: A peptide search engine integrated into the maxquant environment', *Journal of Proteome Research* **10**, 1794–1805.

Johnson, W. E., Li, C. & Rabinovic, A. (2007), 'Adjusting batch effects in microarray expression data using empirical bayes methods', *Biostatistics* **8**, 118–127.

Karahalil, B. (2016), 'Overview of systems biology and omics technologies', *Current Medicinal Chemistry* **23**, 4221–4230.

Lan, L., Djuric, N., Guo, Y. & Vucetic, S. (2013), 'Ms-k nn: protein function prediction by integrating multiple data sources', *BMC Bioinformatics* **14**, S8.

Liou, Y.-R., Wang, Y.-H., Lee, C.-Y. & Li, P.-C. (2015), 'Buoyancy-activated cell sorting using targeted biotinylated albumin microbubbles', *PLOS ONE* **10**, e0125036.

Maes, E., Cools, N., Willems, H. & Baggerman, G. (2020), 'Facs-based proteomics enables profiling of proteins in rare cell populations', *International Journal of Molecular Sciences* **21**, 6557.

Petelski, A. A., Emmott, E., Leduc, A., Huffman, R. G., Specht, H., Perlman, D. H. & Slavov, N. (2021), 'Multiplexed single-cell proteomics using scope2', *Nature Protocols* **16**, 5398–5425.

Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. (2016), 'Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression', *The Annals of Applied Statistics* **10**.

Specht, H., Emmott, E., Petelski, A. A., Huffman, R. G., Perlman, D. H., Serra, M., Kharchenko, P., Koller, A. & Slavov, N. (2021), 'Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using scope2', *Genome Biology* **22**, 50.

Sutermaster, B. A. & Darling, E. M. (2019), 'Considerations for high-yield, high-throughput cell enrichment: fluorescence versus magnetic sorting', *Scientific Reports* **9**, 227.

Tannous, A., Boonen, M., Zheng, H., Zhao, C., Germain, C. J., Moore, D. F., Sleat, D. E., Jadot, M. & Lobel, P. (2020), 'Comparative analysis of quantitative mass spectrometric methods for subcellular proteomics', *Journal of Proteome Research* **19**, 1718–1730.

Vanderaa, C. & Gatto, L. (2021), 'Replication of single-cell proteomics data reveals important computational challenges', *Expert Review of Proteomics* **18**, 835–843.

Väremo, L., Nielsen, J. & Nookaew, I. (2013), 'Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods', *Nucleic Acids Research* **41**, 4378–4391.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X. & Yu, G. (2021), 'clusterprofiler 4.0: A universal enrichment tool for interpreting omics data', *The Innovation* **2**, 100141.

List of Figures