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

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. Systems biology takes a quantitative approach to investigate the interactions, dynamics, and emergent properties of biomolecular networks. Biological systems are imagined as complex, multi-layer networks, with the ensemble of proteins, known as the proteome, playing a critical role within this interactive structure. Proteins have various functions, such as providing structural integrity, catalyzing chemical reactions, and regulating cellular functions (Karahalil 2016).

The central dogma of molecular biology offers insights not only at the level of individual terms but also provides a hierarchical understanding of biology. Hierarchical organization is the underlying principle in all topics of biology. Proteins exhibit hierarchical structures that contribute to their functionality. At the primary structure level, proteins are composed of smaller building blocks called amino acids. Combining them results in the protein secondary structure such as alpha helix and beta sheet. These secondary structures further assemble to create higher-order tertiary structures, which represent specific structural domains within the protein. Finally, the quaternary structure describes the functional state of the protein as a whole at a given time. Proteomics is a comprehensive approach that investigates the complete protein composition of a specimen, aiming to understand the intricate biological network it represents. For instance if pathogens interact with cells of an organism a cascade of biochemical reactions takes place. These reactions influence the outcome of the cellular proteome in regards of localization, abundance, post-translational modifications (Beltran et al. 2017).

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 such as fluorescence-activated cell sorting (FACS), magnetic-activated cell sorting (MACS), and buoyancy-activated cell sorting (BACS) were employed. The rise and validation of cell sorting techniques 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. This is achieved by a physical procedure done with a device which is made of three major components: the ion source, an analyzer, and a detector. The ion sources charges molecules and accelerate them through a magnetic or electric field. The analyzer separates particles according to their mass to charge ratio and the detector senses charged particles and amplifies their signal (Parker et al. 2010). However before acquiring data, sample preparation and pre-processing is required. Sample preparation involves cell sorting and protein isolation. After sample preparation the first step involves tryptic digestion. It is a widely used technique in shotgun proteomics and involves the enzymatic cleavage of proteins into smaller peptides using the proteolytic enzyme trypsin. The process known as proteolysis takes place in multiple parts of multi cellular organisms and was firstly detected in the small intestine digesting proteins (Wang et al. 2008).

To observe multiple samples in one run, labeling is needed to identify each sample within a batch. The two techniques for labeling peptides are metabolic labeling and isobaric labling. The data analyzed by ProteoScanR primarily utilized TMT tags.

Before ionization peptides need to be further separated according to their chemical properties, size or species by liquid chromatography (LC), because the number of peptide processed simultaneously by the MS device is limited.

To analyze a biological sample consisting of peptides in solution the liquid needs to be vaporized into gas phase. Electrospray ionization (ESI) pushes the analyte through a capillary while applying an electric current to the liquid, vaporizing the sample to a charged aerosol. Subsequently after the ionization peptides are accelerated and separated according to their mass to charge ratio (m/z) through the magnetic or electric field generated by the mass analyzer. Before the signal of the separated ions can be obtained an optional subsequent step involves trapping the ions an electric and/or magnetic field in order to detect ions based on their m/z consecutively.

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. The data obtained from mass spectrometry has three dimensions: m/z ratio, intensity, and retention time. To separate peaks from each other, algorithms are used to identify local minima in the data. The centroid of each peak is determined by fitting a Gaussian peak shape, which can be interpreted as locating the peaks of each m/z spectrum as a function of retention time. The centroid of a peak corresponds to an isotope.

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

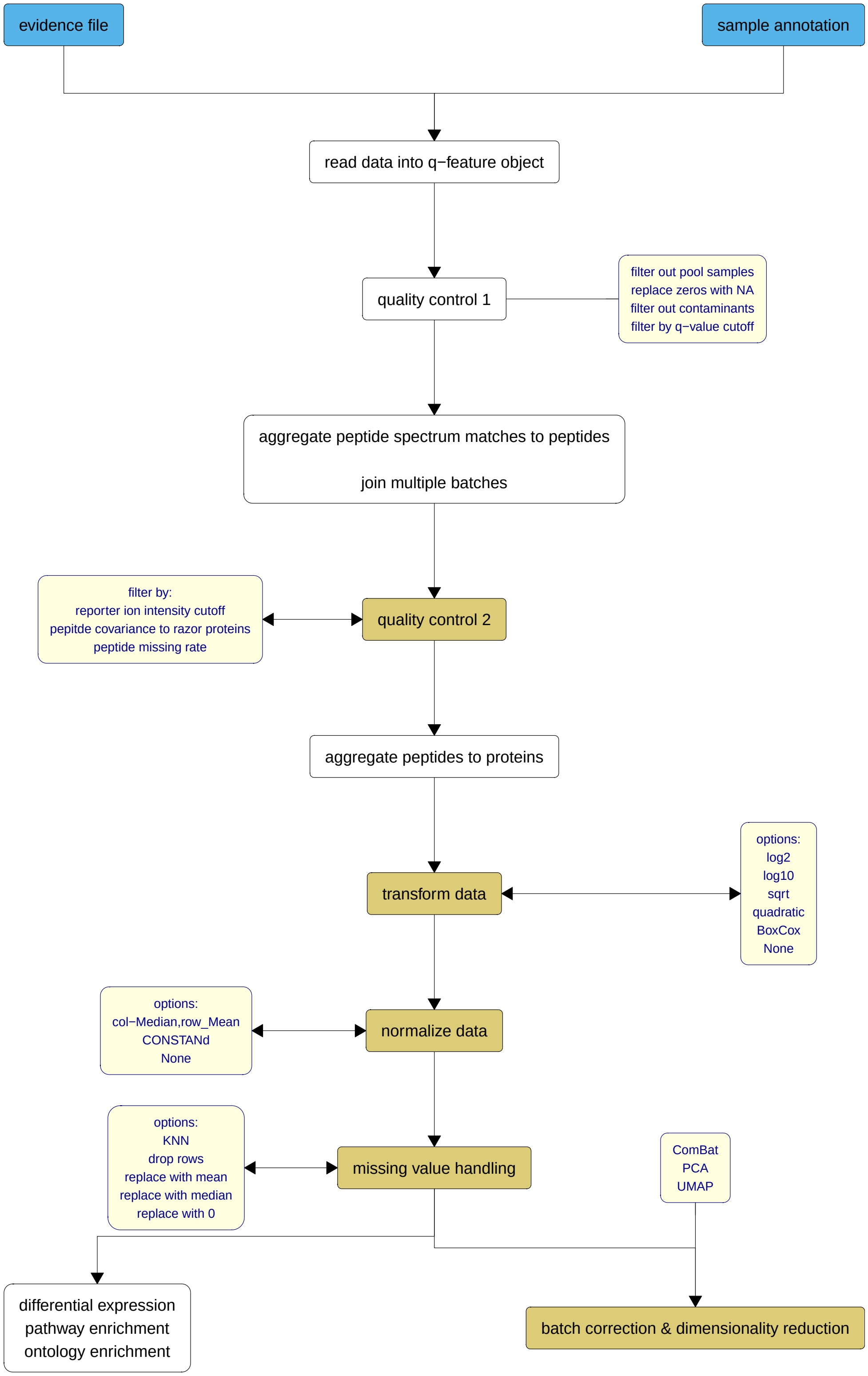
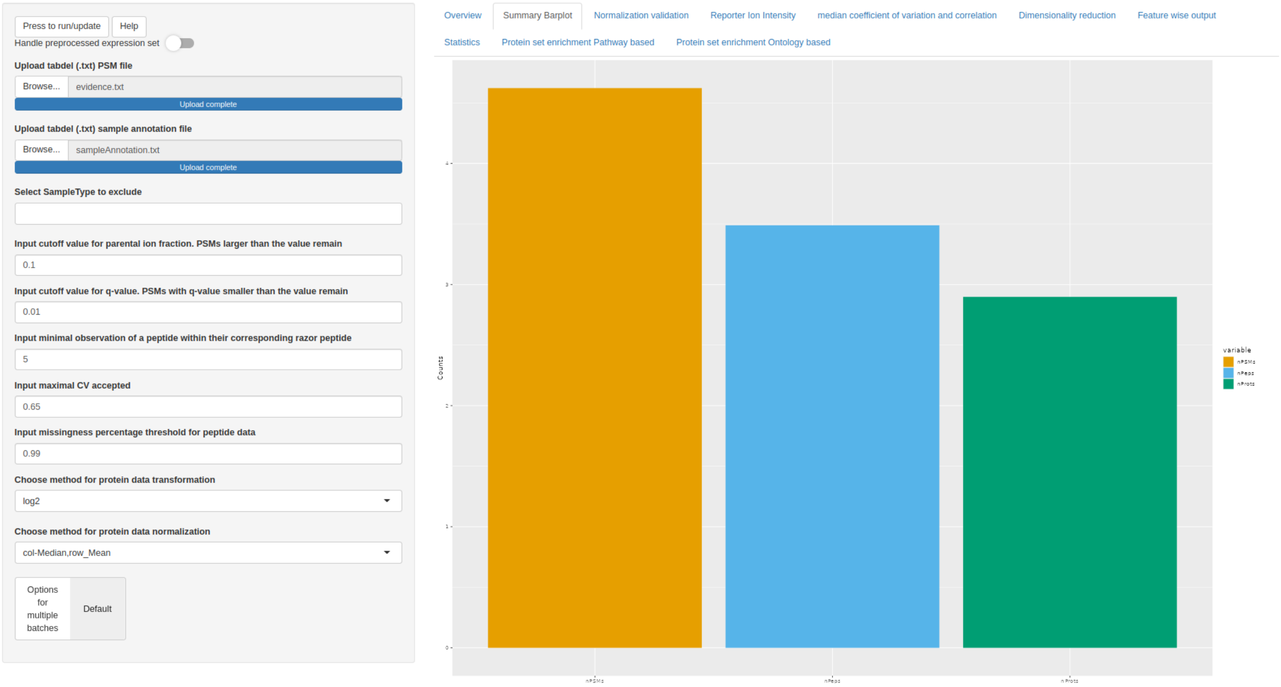


Figure 1: Flowchart presenting the processing by ProteoScanR. Gold: techniques which involve explanatory data analysis. Yellow: Adjustable features. Blue: Inputs

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 custom 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 with 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 the biological context. 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. The online database offers pre-built gene set collections for previous research questions in multiple fields of human biology.

3 Results



Handle preprocessed expression set ○

Upload tabdel (.txt) PSM file

Browse... evidence.txt

Upload complete

Upload tabdel (.txt) sample annotation file

Browse... sampleAnnotation.txt

Upload complete

Select SampleType to exclude

Input cutoff value for parental ion fraction. PSMs larger than the value remain

0.1

Input cutoff value for q-value. PSMs with q-value smaller than the value remain

0.01

Input minimal observation of a peptide within their corresponding razor peptide

5

Input maximal CV accepted

0.65

Input missingness percentage threshold for peptide data

0.99

Choose method for protein data transformation

log2

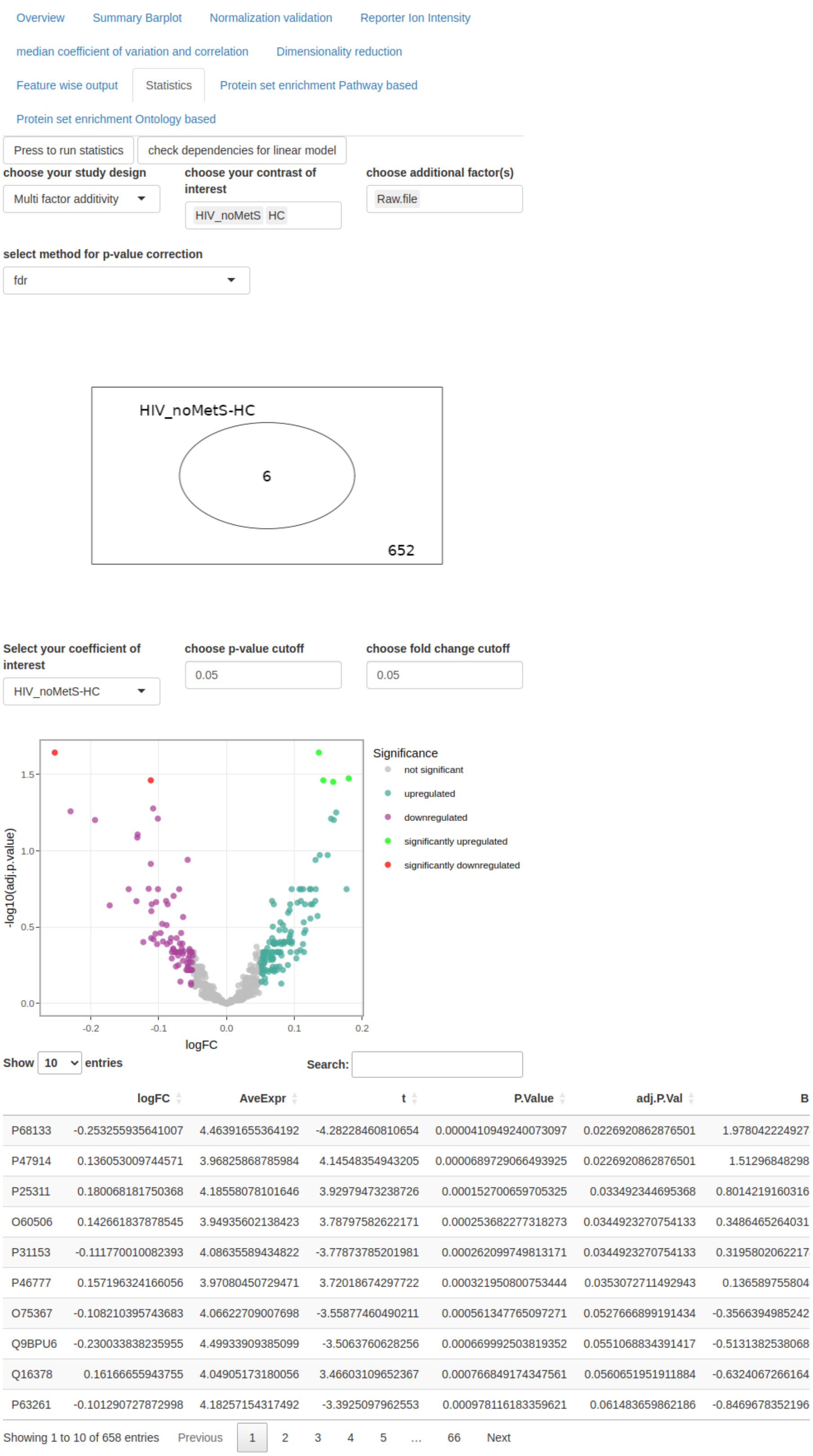
Choose method for protein data normalization

col-Median,row_Mean

Options for multiple batches: Default

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 preprocessing 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.


Overview
Summary Barplot
Normalization validation
Reporter Ion Intensity
median coefficient of variation and correlation
Dimensionality reduction
Feature wise output
Statistics
Protein set enrichment Pathway based
Protein set enrichment Ontology based
Press to run statistics
check dependencies for linear model

choose your study design: Multi factor additivity

choose your contrast of interest: HIV_noMetS HC

choose additional factor(s): Raw.file

select method for p-value correction: fdr

Select your coefficient of interest: HIV_noMetS-HC

choose p-value cutoff: 0.05

choose fold change cutoff: 0.05

Show 10 entries

Search:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
P68133	-0.253255935641007	4.46391655364192	-4.28228460810654	0.0000410949240073097	0.0226920862876501	1.978042224927
P47914	0.136053009744571	3.96825868785984	4.14548354943205	0.0000689729066493925	0.0226920862876501	1.51296848298
P25311	0.180068181750368	4.18558078101646	3.92979473238726	0.000152700659705325	0.033492344695368	0.8014219160316
O60506	0.142661837878545	3.94935602138423	3.78797582622171	0.000253682277318273	0.0344923270754133	0.3486465264031
P31153	-0.111770010082393	4.08635589434822	-3.77873785201981	0.000262099749813171	0.0344923270754133	0.3195802062217
P46777	0.157196324166056	3.97080450729471	3.72018674297722	0.000321950800753444	0.0353072711492943	0.136589755804
O75367	-0.108210395743683	4.06622709007698	-3.55877460490211	0.000561347765097271	0.0527666899191434	-0.3566394985242
Q9BPU6	-0.230033838235955	4.49933909385099	-3.5063760628256	0.000669992503819352	0.0551068834391417	-0.5131382538068
Q16378	0.16166655943755	4.04905173180056	3.46603109652367	0.000766849174347561	0.0560651951911884	-0.6324067266164
P63261	-0.101290727872998	4.18257154317492	-3.3925097962553	0.000978116183359621	0.061483659862186	-0.8469678352196

Showing 1 to 10 of 658 entries

Figure 3: Statistics module

4 Bibliography

Beltran, P. M. J., Federspiel, J. D., Sheng, X. & Cristea, I. M. (2017), 'Proteomics and integrative omic approaches for understanding host–pathogen interactions and infectious diseases', *Molecular Systems Biology* **13**, 922.

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.

Parker, C. E., Warren, M. R. & Mocanu, V. (2010), *Mass Spectrometry for Proteomics*.

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.

Wang, Y., Luo, W. & Reiser, G. (2008), ‘Trypsin and trypsin-like proteases in the brain: Proteolysis and cellular functions’, *Cellular and Molecular Life Sciences* **65**, 237–252.

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.