TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ENHANCED LABEL-FREE DISCOVERY PROTEOMICS THROUGH IMPROVED DATA ANALYSIS AND KNOWLEDGE ENRICHMENT

Tommi Välikangas

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 8 | TECHNICA – INFORMATICA | TURKU 2022

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ENHANCED LABEL-FREE DISCOVERY PROTEOMICS THROUGH IMPROVED DATA ANALYSIS AND KNOWLEDGE ENRICHMENT

Tommi Välikangas

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 8 | TECHNICA – INFORMATICA | TURKU 2022

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral Programme in Technology

Supervised by

Prof. Laura L. Elo
Turku Bioscience Centre
University of Turku and
Åbo Akademi University
Finland

Reviewed by

Asst. Prof. Veit Schwämmle
Dept. of Biochemistry and Mol. Biol.
University of Southern Denmark
Denmark

Prof. Matti Nykter
Fac. of Medicine and Health Tech.
University of Tampere
Finland

Opponent

Docent Markku Varjosalo
Institute of Biotechnology
University of Helsinki
Finland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

ISBN 978-951-29-8741-2 (PRINT)
ISBN 978-951-29-8742-9 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)
Painosalama, Turku, Finland 2022

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
TOMMI VÄLIKANGAS: Enhanced label-free discovery proteomics through improved data analysis and knowledge enrichment
Doctoral Dissertation, 217 pp.
Doctoral Programme in Technology
February 2022

ABSTRACT

Mass spectrometry (MS)-based proteomics has evolved into an important tool applied in fundamental biological research as well as biomedicine and medical research. The rapid developments of technology have required the establishment of data processing algorithms, protocols and workflows. The successful application of such software tools allows for the maturation of instrumental raw data into biological and medical knowledge. However, as the choice of algorithms is vast, the selection of suitable processing tools for various data types and research questions is not trivial. In this thesis, MS data processing related to the label-free technology is systematically considered. Essential questions, such as normalization, choice of preprocessing software, missing values and imputation, are reviewed in-depth. Considerations related to preprocessing of the raw data are complemented with exploration of methods for analyzing the processed data into practical knowledge. In particular, longitudinal differential expression is reviewed in detail, and a novel approach well-suited for noisy longitudinal high-througput data with missing values is suggested.

Knowledge enrichment through integrated functional enrichment and network analysis is introduced for intuitive and information-rich delivery of the results. Effective visualization of such integrated networks enables the fast screening of results for the most promising candidates (e.g. clusters of co-expressing proteins with disease-related functions) for further validation and research. Finally, conclusions related to the prepreprocessing of the raw data are combined with considerations regarding longitudinal differential expression and integrated knowledge enrichment into guidelines for a potential label-free discovery proteomics workflow. Such proposed data processing workflow with practical suggestions for each distinct step, can act as a basis for transforming the label-free raw MS data into applicable knowledge.

KEYWORDS: Mass spectrometry, labe-free, proteomics, normalization, missing values, imputation, longitudinal differential expression, knowledge enrichment

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietojenkäsittelytieteet
TOMMI VÄLIKANGAS: Enhanced label-free discovery proteomics through improved data analysis and knowledge enrichment
Väitöskirja, 217 s.
Teknologian tohtoriohjelma
Helmikuu 2022

TIIVISTELMÄ

Massaspektrometriaan (MS) pohjautuva proteomiikka on kehittynyt tehokkaaksi työkaluksi, jota hyödynnetään niin biologisessa kuin lääketieteellisessäkin tutkimuksessa. Alan nopea kehitys on synnyttänyt erikoistuneita algoritmeja, protokollia ja ohjelmistoja datan käsittelyä varten. Näiden ohjelmistotyökalujen oikeaoppinen käyttö lopulta mahdollistaa datan tehokkaan esikäsittelyn, analysoinnin ja jatkojalostuksen biologiseksi tai lääketieteelliseksi ymmärrykseksi. Mahdollisten vaihtoehtojen suuresta määrästä johtuen sopivan ohjelmistotyökalun valinta ei usein kuitenkaan ole yksiselitteistä ja ongelmatonta. Tässä väitöskirjassa tarkastellaan leimaamattomaan proteomiikkaan liittyviä laskennallisia työkaluja. Väitöskirjassa käydään läpi keskeisiä kysymyksiä datan normalisoinnista sopivan esikäsittelyohjelmiston valintaan ja puuttuvien arvojen käsittelyyn. Datan esikäsittelyn lisäksi tarkastellaan datan tilastollista jatkoanalysointia sekä erityisesti erilaisen ekspression havaitsemista pitkittäistutkimuksissa. Väitöskirjassa esitellään uusi, kohinaiselle ja puuttuvia arvoja sisältävälle suurikapasiteetti-pitkittäismittausdatalle soveltuva menetelmä erilaisen ekspression havaitsemiseksi.

Uuden tilastollisen menetelmän lisäksi väitöskirjassa tarkastellaan havaittujen tilastollisten löydösten rikastusta käytännön ymmärrykseksi integroitujen rikastumis- ja verkkoanalyysien kautta. Tällaisten funktionaalisten verkkojen tehokas visualisointi mahdollistaa keskeisten tulosten nopean tulkinnan ja kiinnostavimpien löydösten valinnan jatkotutkimuksia varten. Lopuksi datan esikäsittelyyn ja pitkittäistutkimusten tilastollisen jatkokäsittelyyn liittyvät johtopäätökset yhdistetään tiedollisen rikastamisen kanssa. Näihin pohdintoihin perustuen esitellään mahdollinen työnkulku leimaamattoman MS proteomiikka-datan käsittelylle raakadatasta hyödynnettäviksi löydöksiksi sekä edelleen käytännön biologiseksi ja lääketieteelliseksi ymmärrykseksi.

ASIASANAT: Massaspektrometria, proteomiikka, normalisointi, puuttuvat arvot, imputointi, erilainen ekspressio, pitkittäistutkimus, tiedon rikastaminen

Table of Contents

Abbreviations

AP-MS	Affinity-Purification Mass Spectrometry
AUC	Area Under the ROC-curve
BETR	Baysian Estimation of Temporal Regulation
BP	Biological Processes
BPCA	Bayesian Principal Component Analysis
CIP2A	Cancerous Inhibitor of Protein Phosphatase 2A
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CRAPome	Contaminant Repository for Affinity Purification-mass spectrometry data
CV	Coefficient of Variation
DAVID	Database for Annotation, Visualization and Integrated Discovery
DE	Differentially Expressed
FC	Fold Change
FDR	False Discovery Rate
FN	False Negatives
Fn	*Francisella tularensis* subspecies *novicida*
FP	False Positives
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
IBAQ	Intensity-Based Absolute Quantification
IgG	Immunoglobulin G
IP	ImmunoPrecipitates
IPA	Ingenuity Pathway Analysis
IQR	InterQuartile Range
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbours
LFQ	Label-Free Quantification
Limma	Linear models for microarray data
LLS	Local Least Squares
Lme	Linear mixed effects regression
logFC	Fold Change in $\log_2$ - transformed data

MAD	Median Absolute Deviation
MAR	Missing At Random
MaSigPro	Microarray Significant Profiles
MCAR	Missing Completely At Random
MCL	Markov Clustering
MNAR	Missing Not At Random
MS	Mass Spectrometry
MSE	Mean Squared Error
PANTHER	Protein Annotation Through Evolutionary Relationship
pAUC	partial Area Under the ROC-curve
PCA	Principal Component Analysis
PCV	Pooled Coefficient of Variation
PEV	Pooled Estimate of Variance
PID	Pathway Interaction Dabase
PMAD	Pooled Median Absolute Deviation
Pme	Polynomial mixed effects regression
POI	Proteins of Interest
PPI	Protein-Protein Interaction
PTM	Post-Translational Modification
ROC	Receiver Operating Characteristic
ROTS	Reproducibility Optimized Test Statistic
SGSDS	Shotgun Standard Set Data Set
SVA	Surrogate Variable Analysis
SVD	Singular Value Decomposition
Th0	T helper cell – activated, undifferentiated
Th17	T helper cell 17
Thp	T helper precursor cells
TN	True Negatives
UPS1	Universal Proteomics Standard Set 1
Vsn	Variance stabilization normalization

Definitions

True Positives (TP) Positive test result correctly identified as such.

False Positives (FP) Negative test result incorrectly identied as positive.

True Negatives (TN) Negative test result correctly identified as such.

False Negatives (FN) Positive test result incorrectly identified as negative.

Sensitivity TP/TP+FN

Specifity TN/TN+FP

ROC-curve In the ROC-curve analysis, sensitivity is plotted against specificity while varying the threshold for detection (e.g. significance value, differential expression statistic).

AUC Area Under the ROC-curve. Typically varies from 0.5 to 1, where 1 equals to perfect performance of the examined method in terms of sensitivity and specificity and 0.5 equals to performance achieved by random ranking of the detections. An AUC value of less than 0.5 corresponds to worse performance of the examined method than what would be expected by randomly ranking the detections and typically indicates a problem with labeling of the classes.

pAUC Partial area under the ROC-curve. Typically, the interest of the researcher is in the top findings (e.g. the (most) DE proteins), partial AUC then focuses on the most essential part of the ROC-curve. In the analysis performed in this thesis, pAUC refers to the area under the ROC-curve between specificity values 1 and 0.9. The pAUC values are rescaled to correspond to full AUC values (i.e. to have a maximal value of 1.0 and a non-discriminant value of 0.5) using the pROC-package [1].

Spike-in data

A dataset where specific proteins (e.g. UPS1 proteins) have been “spiked-in” in known concentrations for different sample groups and mixed with a stable background proteome. As this ground truth of the truly changing spike-in proteins (TP) and the stable background proteins (TN) is known, spike-in datasets enable the benchmarking of methods with the ROC-curve analysis.

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in Bioinformatics*. 2018; 19:1–11.

II Välikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Briefings in Bioinformatics*. 2018; 19:1344–1355.

III #Välikangas T, Suomi T, Chandler CE, Scott AJ, Tran BQ, Ernst RK, Goodlett DR, Elo LL. Enhanced longitudinal differential expression detection in proteomics with robust reproducibility optimization regression. bioRxiv 2021.

IV Tripathi SK*, Välikangas T*, Shetty A*, Khan MM, Moulder R, Bhosale SD, Komsi E, Salo V, Albuquerque RSD, Rasool O, Galande S, Elo LL, Lahesmaa R. Quantitative Proteomics Reveals the Dynamic Protein Landscape during Initiation of Human Th17 Cell Polarization. *iScience*. 2019; 11:334–355.

V Khan MM*, Välikangas T*, Khan MH*, Moulder R, Ullah U, Bhosale SD, Komsi E, Butt U, Qiao X, Westermarck J, Elo LL, Lahesmaa R. Protein interactome of the Cancerous Inhibitor of protein phosphatase 2A (CIP2A) in Th17 cells. *Current Research in Immunology*. 2020; 1:10–22.

The original publications have been reproduced with the permission of the copyright holders.

* Equal contribution as first authors.
# The article has been published as a preprint and has not yet been certified by peer review.

1 Introduction

Biological systems are complex collections of biologically active molecules. Proteins are one of the most abundant and diverse class of such functional biomolecules. Formed by long chains of amino acid residues, polypeptides; proteins are large macromolecules facilitating all cellular function [2]. The functionality of a protein is determined by the amino acid sequence of the polypeptide chain and the folding of the polypeptide chain into complex three-dimensional structures [2]. Changes in the abundance of proteins can significantly alter the state of biological systems. The analysis of proteins and protein abundances can thus be said to form the basis for understanding the functioning of the cell [3].

The set of all proteins expressed by a biological structure such as a genome, cell, tissue or an organism is referred to as a *proteome* and the science that studies proteomes is titled *proteomics*. Proteomics is a central tool used in basic biological research and has great relevance in life sciences more generally. *Clinical proteomics* is interested in the discovery of proteins, protein abundance changes and pathways related to specific cell stress states, such as disease emergence/progeression or toxicity [4,5]. Ultimately, key proteins involved in such stress states could be used as novel *biomarkers* and drug targets addressing these unwanted conditions.

For the reliable and reproducible study of protein expression from complex samples, robust identification and quantification methods are required. Protein mass spectrometry (MS) is an integral part of modern proteomics and forms the basis of reproducible proteins measurements. During the recent decades, the evolution of MS technologies has been rapid. Modern high quality MS workflows are able to identify and quantify tens of thousands of peptides and thousands of proteins, protein modifications and localizations in a single run [2]. The resolution, speed and cost efficiency of MS technologies are expected to even further increase in the future [2]. The successful application of a modern MS powered workflow requires many considerations from the experimental as well as the data analysis point of view. In this thesis, the focus is on considerations regarding proteomics data analysis, especially in questions related to the data analysis methodology of successful protein quantification.

In the first part of the thesis, a brief introduction to related MS technologies is given, after which the focus turns to more specific issues from the viewpoint of data analysis. Typical problems of normalization (**publications I** and **IV**), the choice of a suitable label-free proteomics software workflow (**publication II**), missing values and imputation (**publication II**) are addressed in more detail. After considering the specific data analysis challenges related to the reliable quantification of proteins, the focus of the thesis turns to data analysis methodology used to mine the processed data for knowledge. Specifically, the interest is in the robust detection of differential expression in longitudinal experimental settings (**publication III**). Following the discovery of the proteins of interest (POI), knowledge enrichment of the findings through combined enrichment and network analysis is explored (**publications IV** and **V**). Finally, the conclusions and implications from all the performed work included in the thesis are combined into a suggestion for a label-free proteomics discovery workflow.

2 Mass Spectrometry-Based Proteomics

Before the discovery of modern mass spectrometry, the identification of a proteins amino acid sequence was time consuming and required large amounts of sample material [6]. The sequence was determined by identifying chemically cleaved amino acids sequentially beginning from the N terminus with the Edman degradation method [6]. This method required substantial expertise and often an unambiguous sequence could not be assigned [6]. Since the development of mass spectrometry in 1990s, it soon replaced the Edman Degradation method [6]. In contrast to the Edman degradation, with MS, the sensitivity is increased, peptide fragmentation is much faster and the peptides or proteins do not need to be purified to homogeneity [6]. Furthermore, MS can be used to identify blocked or modified proteins.

While diverse sample types (e.g. cell cultures, tissue and fecal samples) can be analyzed with MS, the proteins need to be first extracted from the samples, purified, typically digested into peptides and separated [2,6,7]. The optimal way to extract proteins from a sample is varying and application dependent [2]. Although mass spectrometers can measure the mass of entire intact proteins, digesting the proteins into peptides is considered a more favorable solution in most settings [6,8,9]. The proteins are digested into peptides using enzymes that cleave the backbone of the proteins amino acid sequence at specific sites [6,7]. Most commonly this enzyme is a protease known as trypsin [2,6,7]. Trypsin cleaves the backbone of the protein very specificly on the carboxy (C) terminal side of the arginine and lysine amino acid residues. The cleaving by trypsin generally leaves the charge carrying amino acids at C termini side of the peptide and creates uniquely classifiable peptides in the suitable mass range for the MS analysis [1–3].

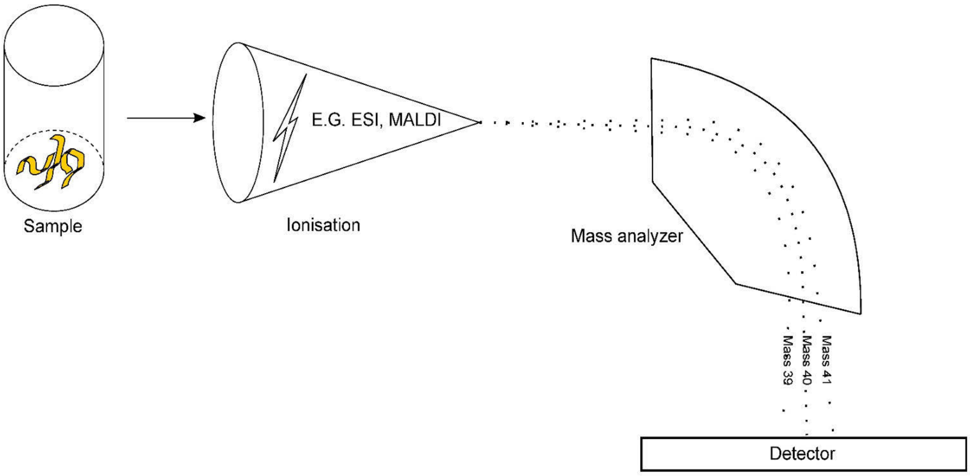


Figure 1. A simplified schematic diagram of a mass spectrometer.

The mass spectrometer consists of three fundamental components: an ion source, a mass analyzer and a detector [7] (**Figure 1**). Before entering the mass spectrometer to be ionized, the digested peptides need to be separated [6]. This typically happens by injecting the peptides into a microscale capillary high performance liquid chromatography (HPLC) or an ultrahigh pressure liquid chromatography (UHPLC) column coupled to the mass spectrometer [2,6]. If very complex samples are being processed, the proteins might have to be separated already prior to digestion into separately processed fractions by gel electrophoresis or other techniques [6]. The peptides flow through the column and through a needle point, get vaporized and ionized by a strong electric potential [6]. This type of setup, where the peptides are ionized from a liquid column is called electrospray ionization (ESI) or LC-MS [6,7]. Another type of approach, where the peptides are ionized from a dry crystalline matrix, is called the matrix assisted laser desorption/ionization (MALDI). While MALDI has been generally used to analyze relative simple peptide mixtures and ESI has been used in the analysis of more complex samples [7], recent studies have suggested that these two techniques might be complementary while neither is comprehensive in identifying all the peptides in a sample [10]. However, in this introduction the focus is in ESI or LC-MS systems.

Once the ionized peptides have entered the vacuum of the mass analyzer, they are controlled by strong electrical currents. The mass analyzer plays a central part in the MS technology. It separates the ionized peptides according to the mass to charge (m/z) ratios. The detector then registers the numbers of ions at each m/z value [6] (**Figure 2A-B**). Once all the m/z ratios and intensities of all the peptides are recorded, in a typical MS experiment individual peptides are further fragmented and the mass spectra of the resulting fragments are recorded in a second mass spectral scan [2,6] (**Figure 2C**). The first scan is generally called MS1 spectra or survey spectrum and the second scan as tandem MS or MS/MS or MS2 [2,6]. This type of MS experiment is generally referred to as tandem MS [6]. Typically, the MS2 spectra is obtained only for the most abundant peaks in the MS1 spectra [6]. The masses of the peptides and the fragments are used to identify the peptides while the intensity information is used for quantification [2]. Multiple different types of mass analyzers with different strategies exist [2,6,7]. Four basic types of mass analyzers can be distinguished: the ion trap, time of flight (TOF), quadrupole and Fourier transform ion cyclotron (FT-MS) [7]. However, multiple variations and combinations of these basic analyzer types have been developed. The basic types differ in how they determine the m/z ratios of the ionized peptides and in their resolution [6,7]. In the recent decade, especially the Orbitrap and TOF analyzers have greatly improved in their resolution and popularity [2].

The identity of the peptides can be inferred from the peptide and peptide fragment masses mainly in two ways: via a technique called *de novo* sequencing or through database searching [2,6,7]. In database searching, the experimentally acquired peptide fragmentation spectrum is typically matched to *in silico* digested theoretical spectra of a database using a specified search engine [6]. The matching theoretical peptide sequences are scored and the identity inferred based on the most probable solution [6,7]. There are several popular search engines such as Mascot [11], X!Tandem [12], MS-GF+ [13], Andromeda [14] and others for performing the peptide and protein database searching. The identification of proteins via database searching follows a similar strategy as peptide identification: based on the identified peptide sequences for a protein, the most probable proteins from which these sequences could arise from, are presented. Distinct species specific protein sequence databases, such as Swiss-Prot/TrEMBL [15], can be used for this type of protein identification. In *de novo* sequencing, the identity of the peptides is attempted to be inferred based on the tandem MS fragmentation spectrum alone [6]. However, such *de novo* sequencing is a very challenging task, which is why database searching is commonly applied for peptide and protein identification [2,6].

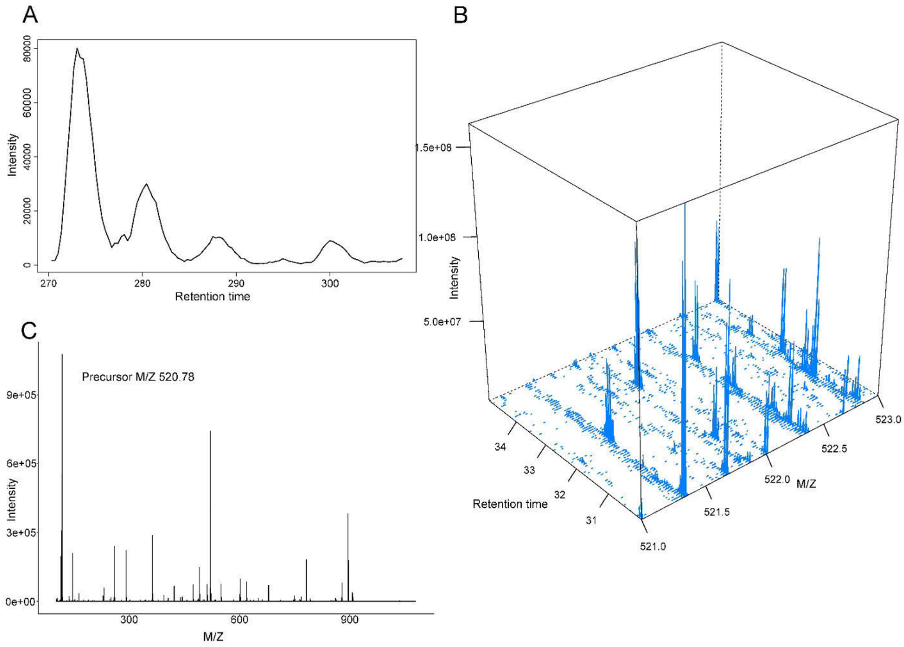


Figure 2. Representative chromatograms from a mass spectrometer. A) A representative MS1 level base peak chrotomatogram showing the most intense peaks over retention time across the whole range of masses. B) A 3D representation of raw MS1 level data showing the peak intensities at different m/z values over retention time. C) A representative peptide ion MS2 level fragmentation spectrum.

2.1 Label-free proteomics

Different sample types (e.g. cell cultures, fecal samples, body fluids) or properties of the investigated proteome may require distinct considerations for protein quantification via MS [2,16]. The quantification techniques used in MS experiments can be divided into two main categories: *label-based* quantification and *label-free* quantification [2,5,16].

In label-based MS, the samples are typically labeled with stable protein isotopes acting as internal standards or references, allowing for accurate quantification of protein abundances in the samples [5,16]. The labels can be incorporated into the samples using metabolic, chemical or enzymatic labeling resulting in various methodologies such as stable isotope labeling by amino acids in cell culture (SILAC), stable isotope labeling of mammals (SILAM), isotope-coded affinity tags (ICAT), isotope-coded protein labeling (ICPL), isobaric tags for relative and absolute quantification (iTRAQ), tandem mass tags (TMT) etc. [5,16]. The labeling based approaches offer accurate and robust quantification of protein abundances but require high sample concentrations, are expensive and can generally be performed for a limited number of samples per run [5,16,17]. In contrast to the labeling-based methods, label-free approaches are not dependent on the specific labeling of samples, can be simultaneously performed for a higher number of samples, typically require less sample preparation and are more cost effective [16,18]. Furthemore, label-free methods can be applied even when the metabolic labeling of samples is not possible [18].

The label-free approaches can be further divided into two main techniques: spectral counting and peak intensity methods. In spectral counting, protein quantification is performed by counting the number of identified MS2 fragment spectra matched to peptides derived from a specific protein [19]. However, the accuracy of the spectral count method has been questioned [19–21] and the peak intensity methods have been observed to give more robust quantifications at a larger dynamic range [19–22]. In peak intensity methods, peptide quantification is performed by measuring the area under the peak or the maximum intensity (height) in the MS1 or MS2 level data [16]. Protein quantification is then performed by "rolling up" or aggregating the peptide intensity measurements of a protein through various methods (such as summing, median, mean, top n). The intensity based quantification is motivated by the observation that the measured ion signal intensity linearly correlates with ion concentration [16,21]. Furthermore, the combined peak areas of the peptide ions for proteins have been observed to correlate with protein amount [16].

The quantitative label-free MS approaches can be applied in various experimental settings, such as the *shotgun* or the discovery approach, as well as the *directed* or *targeted* approaches [23]. The emphasis in the shotgun approach is the discovery analysis of the whole proteome [23]. In the targeted and directed approaches on the other hand, previous information is applied to select a set of proteins or peptides which are then explored in more detail [23]. Perhaps the most widely used application of targeted approaches is the selected reaction monitoring approach (SRM) [23]. In SRM, specific representative peptides relating to some proteins of interest are targeted. The characteristics of these peptides (such as the fragment ion spectrum) have been determined beforehand and can be used to select peptides for further analysis during the course of the experiment. Thus particular fragment ion(s) for a selected peptide precursor ion are monitored and the signal intensity associated with these specific pair of m/z values (referred to as the SRM transitions) are used for quantification [23,24]. The SRM transitions together with the retention time of the targeted peptide offer good selectivity for the SRM approach [23,24]. Developments in the MS technologies during the past decades have enabled the shotgun method to become a popular approach, able to identify and quantify thousands of proteins from complex samples in a single run [2,23].

Depending on the used MS instrument and settings of the instrument, the quantitative label-free shotgun experiments can be performed in *the Data Dependent Acquisition (DDA)* or *Data Independent Acquisition (DIA)* modes. Typically in DDA, for each scan cycle of the MS device, only 10-20 most abundant peptides at the MS1 level are sequentially selected for fragmentation and for the acquisition of the MS2 spectra [25]. Consequently, the selection of different subsets of peptides in different replicates and samples can cause moderate amounts of missing values in the DDA data [25]. On the other hand, in the DIA mode, all peptides within a selection window at a narrow m/z range are fragmented and the MS2 spectra acquired [25,26]. This acquisition is repeated in a stepwise manner for different selection windows in a wider mass-to-charge range (e.g. 400-1000 m/z range) [25]. However, the DIA mode typically requires more sample material for the generation of spectral libraries by multiple runs of the MS instrument in DDA mode [25,26], although also library-free DIA approaches have been developed [27]. Typically, DDA has been used in the knowledge-blind shotgun discovery studies while the strength of the DIA has been seen in developing targeted assays and more accurate and reproducible workflows [26]. Furthemore, while DIA may provide an inherent solution to the missing value problem presented by DDA, DIA requires more complicated data analysis and is still developing in terms of bioinformatics software and instrumentation [28]. In the following subchapters of the introduction of this thesis, data preprocessing methods related to the label-free peak intensity approach are discussed in more detail. The specific challenges and methods presented are discussed mainly from the point of view of DDA data, but most of the solutions suggested are likely applicable to DIA data as well. Some of the main strategies for quantification in MS based proteomics discussed in this chapter are summarized in **Figure 3**.

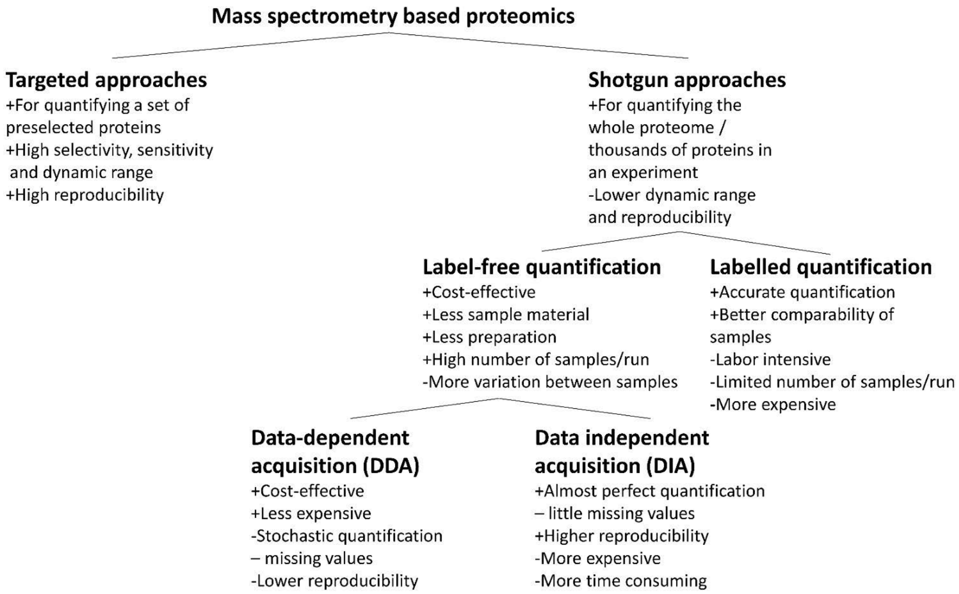

Mass spectrometry based proteomics
Targeted approaches
+For quantifying a set of preselected proteins
+High selectivity, sensitivity and dynamic range
+High reproducibility
Shotgun approaches
+For quantifying the whole proteome / thousands of proteins in an experiment
-Lower dynamic range and reproducibility
Label-free quantification
+Cost-effective
+Less sample material
+Less preparation
+High number of samples/run
-More variation between samples
Labelled quantification
+Accurate quantification
+Better comparability of samples
-Labor intensive
-Limited number of samples/run
-More expensive
Data-dependent acquisition (DDA)
+Cost-effective
+Less expensive
-Stochastic quantification – missing values
-Lower reproducibility
Data independent acquisition (DIA)
+Almost perfect quantification – little missing values
+Higher reproducibility
-More expensive
-More time consuming

Figure 3. Some of the central approaches for mass spectrometry based proteomics with their respective main advantages and possible challenges.

2.2 Normalization in proteomics

While MS has evolved rapidly during the past decades and modern MS instruments have greatly developed in terms of accuracy and robustness [2,5], the data produced by the instruments are still prone to systematic biases [29]. This kind of bias may result from small variations in the experimental conditions such as differences in sample handling or preparation, variations in liquid chromatography flow rates, changes in temperature during the course of the experiment, device calibration, etc [30,31]. More generally, the bias can be defined as unwanted non-biological variation occurring spontaneously during the course of the MS experiment [30–32] (**Figure 4A**). Furthermore, the specific cause of the bias is typically unknown and cannot solely be countered by adjusting the instrumentation or the experimental settings [29,31]. The observed bias can be dependent or independent of the measured protein or peptide abundances.

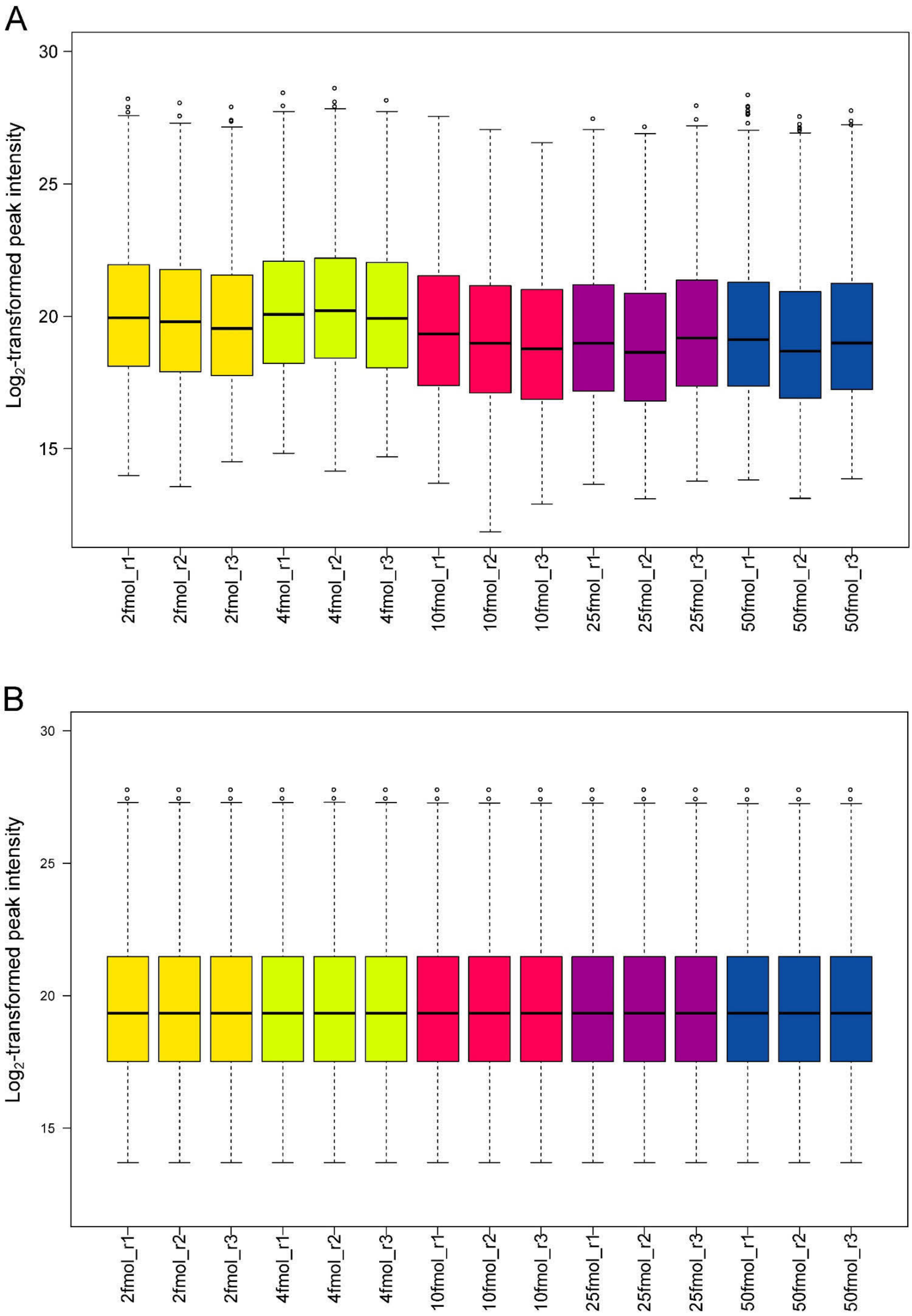


Figure 4. An example of the effects of normalization in proteomics. A) Sample distributions of the log$_2$-transformed peak intensities in an unnormalized label-free proteomics dataset. B) Quantile normalization performed on the same dataset. Samples are colored according to sample groups.

The process that tries to compensate for this bias, make the samples more comparable and the downstream data analysis more reliable, is called *normalization* [29–32] (**Figure 4B**). As a result of normalization, the data should be less biased and the true biological signal more prominent in the data. Normalization is not a process related solely to proteomics, rather most high-throughput methods producing large amounts of data through complex experimental settings and instruments, require normalization [29,31,32]. Normalization in proteomics thus builds on normalization in preceding, more established technologies, such as DNA microarrays [29,31]. Many popular normalization methods used in proteomics, such as quantile normalization [33], median normalization, variance stabilization normalization (vsn) [34] have all been developed or applied already in conjunction with DNA microarrays. Normalization in proteomics and the best practices for normalizing MS data have been investigated on several accounts [29–31,35–37]. Typically, normalization methods have been evaluated in their ability to decrease intragroup variability in technical and/or biological replicate sample groups [29,36,37]. However, prior to **publication I**, the effect of the normalization methods for detecting differential expression had not been systematically and comprehensively evaluated.

To summarize, missing or improper normalization can result in incomparable samples and biased or non-reliable findings in the differential expression analysis (e.g. incorrectly detecting variation arising from technical reasons as biological variation). In the first work included in this thesis (**publication I**), 11 popular normalization methods were evaluated in their ability to decrease intragroup variation and produce unbiased data from which the true biological signal can correctly be detected in the differential expression analysis. Furthermore, the effect of the different normalization procedures on the known magnitude of change of the true signal in the data was evaluated [32].

2.3 Label-free proteomics data processing software tools

As discussed in the previous chapter, normalization is an essential part of the MS powered label-free proteomics workflow. Normalization can be integrated as part of the workflow within a label-free proteomics software or it can be an external algorithm, applied after receiving the non-normalized protein abundances from the used processing software workflow. In addition to normalization, there are several typical steps common to a label-free proteomics software workflow.

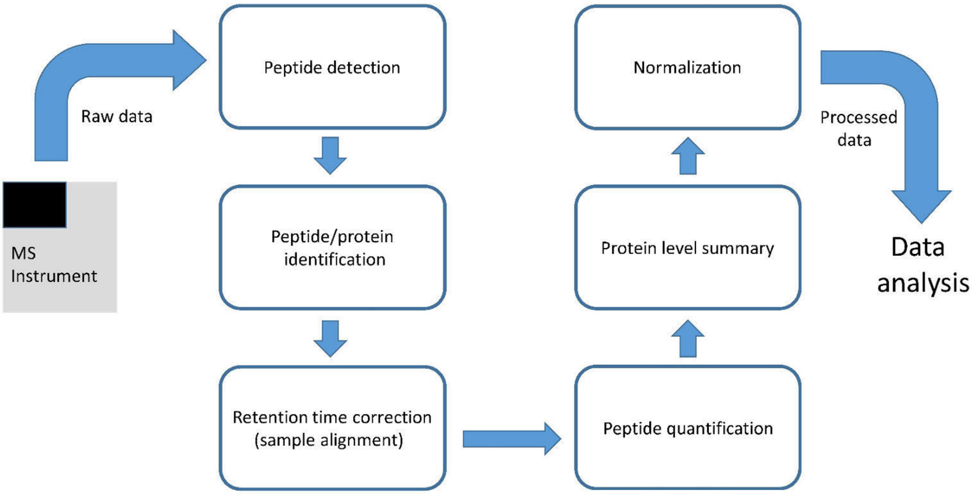


Figure 5. A simplified schematic of a possible label-free proteomic software data processing worklfow. Several alternatives to such simplified workflow and processing order exist. Retention time alignment may preceed or happen simultaneuously with peptide identification and peptide detection. Normalization may also take place before protein level summary.

Such common steps can include *feature detection*, *peptide identification*, *alignment* of the peptides in different samples and *aggregation* of peptide identification and quantification information into protein identities and relative quantities [38] (**Figure 5**).

Feature detection is the process, where the isotopic envelope for a single eluting compound is determined [38,39]. Typically, a single eluting peptide will produce a collection of peaks at the MS1 level, corresponding to different isotopes of the same peptide [38,39]. A single peptide feature thus has three dimensions: the retention time, m/z ratio and the measured intensity (**Figure 2B**). Feature detection algorithms aim to identify all three aspects related to a peptide as accurately as possible [38]. The resulting feature list for an isotope peak cluster can for example consist of a combination of monoisotopic m/z value (m/z value for the peak of the primary isotope), apex retention time (highest point of the feature), charge and intensity [38]. Quantification can be performed by using total intensities: the sum of all integrated isotope peaks along the retention time, or apex intensities: sum of only the isotope apex retention time intensities [38,39]. As the resolution of the MS instruments is ever improving, the more accurate total sum based quantification approaches have become popular choices [38]. However, considerable variation exists between different software in their approach to quantification and feature detection [38–43].

As mentioned earlier in chapter 2, identities for the detected features, or peptides, can be inferred by two main approaches: de novo sequencing or database searching. A database search engine is typically included or applied by a label-free software proteomics workflow [38]. For example, several search engines, such X!Tandem [12], MS-GF+[13] or OMSSA [44], can be applied within the OpenMS [41] and Proteios [45] software workflows, while the Andromeda [14] search engine and the PEAKS DB [46] search engine are incorporated into the MaxQuant [42] and Peaks [47] software environments, respectively. A suitable FASTA sequence database is generally provided for the applied database search engine(s) for the inference of protein identities. *De novo* sequencing is possible for example by using the PEAKS DB [46] included the Peaks studio software environment or open source tools such as pNovo 3 [48]. As database searching relies on the reference protein sequences provided in the FASTA sequence databases, variations in the amino acid sequence of the protein to be identified can induce problems for the identification [49,50]. Such variations in the amino acid sequence may arise from individual genetic variability and genetic mutations. As amino acid substitutions in the experimental sequence can impair its matching to *in silico* digested peptide spectra in the database, the identification of mutated or heavily modified peptides can be highly susceptible to errors [49,50]. However, many tools have been developed to overcome this problem, such as the SPIDER software [51] in the Peaks Studio software suite utilizing *de novo* sequence tags and cross-species database searching. Similarly, proteogenomics approaches [52,53] and associated tools have been developed [49,54], where RNA or DNA sequences from the same samples or a variety of samples under similar conditions (e.g. cancer) are used in inferring experiment specific peptide sequence databases with genetic variation considered.

Alignment is the process in which the discovered peptide identifications are transferred between different MS runs (samples) [38,43]. Furthermore, alignment corrects for retention time distortions between different samples by aligning the peptide maps of the different MS runs [38]. Alignment together with normalization allows for the reliable downstream comparison of samples and conditions. Alignment can be performed mainly at two different stages: before or after feature detection [38,43,55]. In *feature based alignment*, feature detection is performed before alignment whereas in *profile based alignment* the order is reversed [43,55]. Simplified, in profile based alignment approaches, the different runs are aligned based on their chromatographic profiles, where as in feature based alignment, the runs are aligned according to the identified features [38,55]. The less computationally costly feature based approach is used mainly for high resolution data, where the identification of the features is less error prone [38]. In addition, the alignment algorithms can be further divided into two main approaches: *reference run based alignment* and *reference free alignment* [38,43]. In the reference run based alignment, one representative run is selected either automatically by the software or by the user [38]. All the other runs are then aligned using the selected reference run. The Progenesis software for example applies a reference run based alignment approach. On the other hand, in the reference run free approach, the different runs are aligned without a chosen reference run using for example clustering approaches. For instance, MaxQuant [42] uses hierarchical clustering to align the runs [42,56].

As typically the interest of the experimental researcher is in protein quantities instead of peptide quantities, most of the label-free software provide options to “roll-up” the peptide quantifications into protein quantifications [56,57]. Several varying strategies for protein quantification exist. In MaxQuant [42] proteins can be quantified using the Intensity-Based Absolute Quantification (IBAQ) [57] or the MaxLFQ [56] methods. In Progenesis, relative protein quantification can be performed by using the sum of non-conflicting peptides for a protein (peptides unique to the quantifiable protein), by calculating the sum of all the peptides for a protein or by using the averaged intensity of top N peptides for a protein. Similarly, the ProteinQuantifier module in OpenMS [41] contain several options for calculating the protein intensities from peptide intensities.

Label-free proteomics software workflows can be divided into two main groups based on their modularity: modular and complete workflows [58]. Modular workflows, such as OpenMS [41] and Proteios [45], provide a software environment in which several algorithms for a given task (such as peak picking for feature detection, varying search engines for peptide and protein identification, etc.) can be plugged in to formulate a full label-free proteomics software workflow [41,59]. Such modular software environments are typically open source and provide a flexible software environment for developers to deliver their new tools for separate tasks as well as give the user control in defining the workflow as desired [41,59] (**Figure 6**). However, while being flexible, such modular workflows typically require a certain level of expertise from the researcher for proper application.

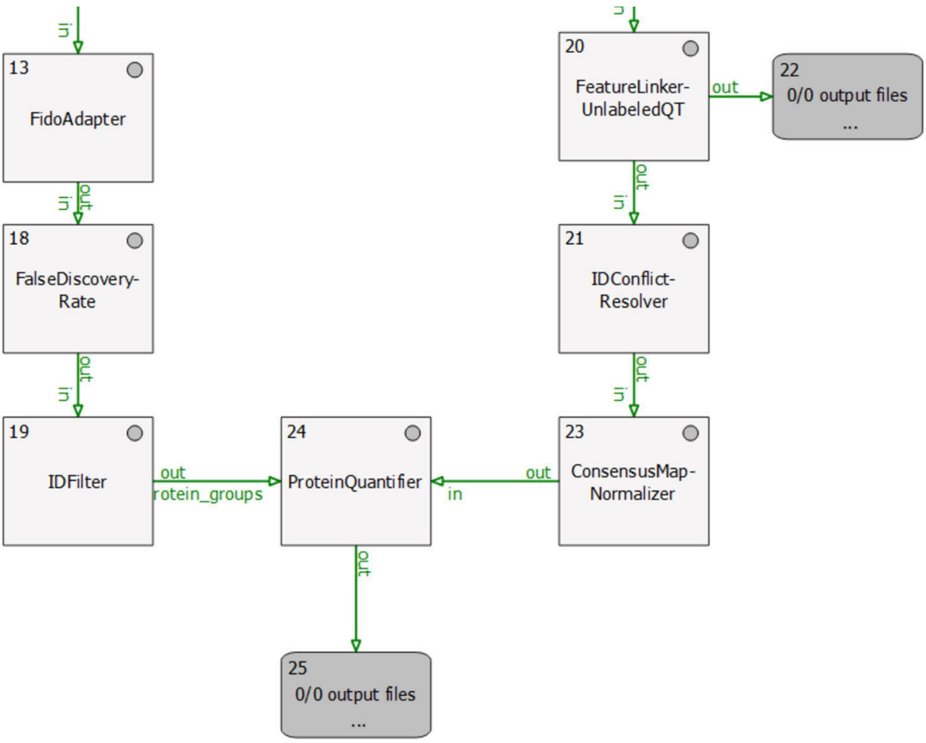


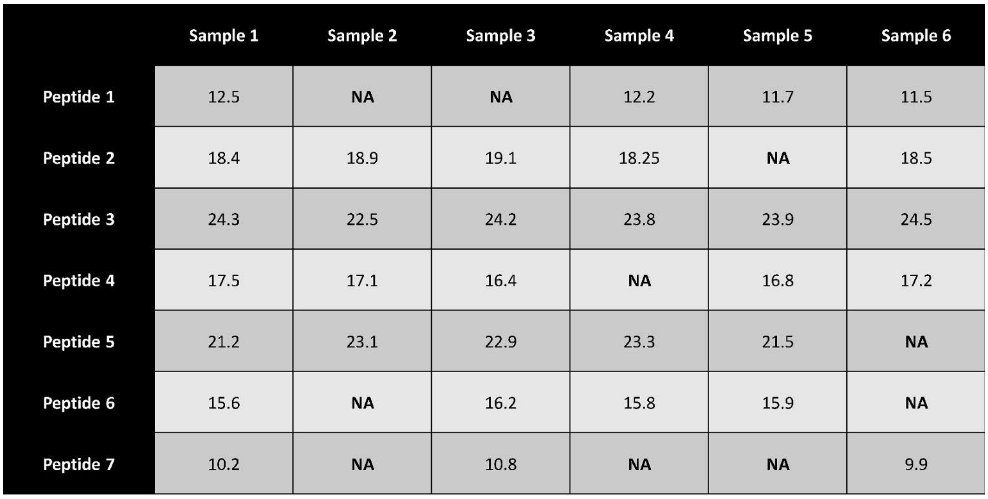
Figure 6. An example of part of a label-free MS discovery workflow with the OpenMS software.

As opposed to modular workflows, complete worklows typically have one or a few built-in specific algorithms for a given task [58]. Such complete workflows might be open source or non-profit, such as the popular MaxQuant [42] software, or commercial such as the Progenesis and Peaks [47] software solutions. These kinds of complete solutions are typically easy to use and require less prior knowledge about the workings of the specific algorithms used.

In the second work included in this thesis (**publication II**), five popular label-free proteomics software workflows were evaluated in their ability to correctly detect the known truly differentially expressed proteins and to correctly estimate the known fold changes [60]. Both commercial and non-profit, as well as complete and modular solutions were evaluated [60].

2.4 Missing values and imputation in label-free proteomics

Missing values are an inherent and a well known problem in MS data, especially in the label-free DDA data [25,30,31,61–63]. Typically, non-random missing values in MS data might occur when the instrument is not able to detect low abundant peptides or this weak signal cannot be distinguished from the background noise during data processing [61,63]. However, other more random processes, such as miscleavages, ionization competition, ion suppression, peptide misidentification, and retention time drifts, can cause missing peptide values in a given MS experiment [61]. Additionally, the data generation process of the DDA label-free proteomics data is known to result in a high proportion of missing values [25,31,61]. As described in chapter 2.1, typically 10-20 of the most abundant peptide precursor ions at the MS1 level at each scan cycle are selected for further fragmentation at the MS2 level in DDA mode [25]. This stochastic selection of peptide precursors leads to inconsistent detection of peptides and missing values in multiple MS runs, even if the same sample is measured multiple times [25]. The selection of different peptide subsets in different MS runs leads to a high number of low to medium abundant peptides not detected across all runs (**Figure 7**).



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Peptide 1	12.5	NA	NA	12.2	11.7	11.5
Peptide 2	18.4	18.9	19.1	18.25	NA	18.5
Peptide 3	24.3	22.5	24.2	23.8	23.9	24.5
Peptide 4	17.5	17.1	16.4	NA	16.8	17.2
Peptide 5	21.2	23.1	22.9	23.3	21.5	NA
Peptide 6	15.6	NA	16.2	15.8	15.9	NA
Peptide 7	10.2	NA	10.8	NA	NA	9.9

Figure 7. An example of a $\log_2$-transformed label-free proteomics peptide quantification matrix with missing values. Missing values are depicted as NA.

More generally, missing values have been classified according to their cause of origin as *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *Missing Not At Random* (MNAR) [61]. Typically with regards to MS data, it is assumed that most MAR observations are MCAR and MAR cases are not analyzed or treated separately [61]. Thus, missing values in MS data are mainly divided into two categories based on their origin: abundance dependent missing values (MNAR) and MCAR [31,61]. It has been observed, that while a relationship between peptide intensities and missing values exist (MNAR), depending on the dataset and technique used, as large a proportion as 50-70% of missing values might be MCAR [62].

Imputation of the missing values with different methods has been proposed as a solution to missing values in MS data [31,61,62]. A previous comparison of imputation in proteomics by [62], evaluated the imputation methods based on their accuracy when missing values were artificially generated. The performance of various imputation approaches with regards to missing values of different origins (MCAR and MNAR) have also been investigated [61]. Simple methods, such as imputation of half of the minimum observed values or minimum of the observed values, have been evaluated together with more sophisticated approaches [61,62]. The more complex methods have included local similarity approaches, such as k-nearest neighbours (KNN) or local least squares (LLS) [62], as well as global structure approaches such as the Probabilistic Principal Component Analysis (PPCA) [62] and Singular Value Decomposition (SVD) [61]. In general, the simple value imputations have been observed to be less accurate than the more sophisticated approaches [62]. However, if a majority of the missing values are MNAR, the simple minimum value imputations can perform well and even outperform the more complex methods [61].

Overall, missing values can distort the following statistical analysis in many ways. As many statistical approaches (e.g. PCA, many statistical tests) require complete data, the data needs to be either filtered or imputed to contain no missing values. The incorrect application of these approaches can result for example in biased results in the differential expression analysis (e.g. the detection of technical noise from imputation as true biological signal) or not detecting some true findings containing missing values at all. In the second included work in this thesis (**publication II**), the performance effects of seven imputation methods together with a filtering approach and a combined filtering imputation approach were also evaluated. The imputation and filtering approaches were evaluated in their effect on the performance of the different tested label-free-proteomics software workflows in the differential expression analysis.

2.5 Label-free proteomics for biomarker discovery

Once the MS-data has been properly preprocessed, the focus turns to what can be discovered from the data. There is a high demand for various biomarkers in the medical field [64–68]. A biomarker is more generally characterized as: ”a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or a response to an exposure or intervention” by the Food and Health Administration (FHA) and National Health Institute (NHI) Biomarker Working Group [69]. Depending on their intended usage, biomarkers can be further divided into main classes such as: diagnostic biomarkers, monitoring biomarkers, predictive biomarkers, prognostic biomarkers, etc. [70,71]. A good diagnostic biomarker is able to differentiate between the healthy and the diseased states early in the disease progression with good sensitivity and specificity [64,65,71]. A popular approach for discovering diagnostic or prognostic biomarker candidates is detecting *differential expression* or *differentially expressed* (DE) proteins between the conditions of interest (e.g. healthy and diseased).

The use of proteins as biomarkers is well established in clinical medicine and the concentrations of many plasma proteins for example, are routinely measured in clinics worldwide [64,65,72]. MS-based proteomics is routinely used in several applications, for example toxicological testing and therapeutic drug monitoring [64]. However, many clinical operations utilizing protein biomarkers apply various non-MS techniques, such as enzymatic or immunoassays [64,65]. During the recent decades, MS techniques have evolved rapidly [2,5] and gained popularity, especially in biomarker discovery studies [65]. MS-powered proteomics has the power to potentially analyze all, or at least thousands of, proteins in a sample, including their post-translational modifications (PTMs) [65,72]. Despite the rapid developments, challenges, such as the large dynamic ranges of proteins in a sample (e.g. blood plasma), still remain [65]. Quantitative label-free shotgun DDA proteomics has emerged as a popular choice for the initial discovery phase, due to its cost-effectiveness and lower sample preparation complexity [16,18]. A biomarker discovery study workflow utilizing MS-powered proteomics might involve a quantitative shotgun proteomics discovery phase followed by a validation of the discovered candidates with targeted techniques, such as immunoassays (e.g. enzyme-linked immunosorbent assays), western blotting, selected reaction monitoring (SRM)-, or multiple reaction monitoring (MRM)-MS [65]. For a successful discovery of a new applicable clinical biomarker, further validations utilizing independent disease related cohorts are also typically required [65,72].

Typically in a quantitative shotgun proteomics workflow, thousands of proteins are quantified in a limited number of samples, limiting the statistical power and the ability to detect differential expression between the conditions [72]. Specifically, poor experimental design and incomplete power analysis may lead to an insufficient number of samples [73,74]. In addition to increasing sample size (e.g. including more individuals), an alternative for increased statistical power are repeated measurements, or a longitudinal study design [74–76]. In addition to providing more statistical power, a longitudinal experimental design also delivers important information on the changes in the response/behavior of the individuals and proteins over time [74].

Following the discovery of the potential biomarker candidates (e.g. DE proteins), further analysis is typically performed to deepen the understanding of the possible biological mechanisms involded. For example, enriched pathways, biological processes, cellular locations and molecular functions within the candidate biomarkers can be explored [5]. The analysis of protein-protein interactions together with such functional enrichment information can enable the fast discovery of interesting proteins (e.g. clusters of co-expressing proteins with similar research question related functionalities). The successful refinement of data from the instrument into biological or medical knowledge and interpretations thus typically requires multiple successive and different computational analysis to be performed.

In the following chapters, computational methods for the discovery of the DE proteins in label-free proteomics data from longitudinal experiments and the further refinement of these findings into biological knowledge through the integration of network and functional enrichment analysis are considered in more detail.

2.6 Longitudinal differential expression in proteomics

Experiments with longitudinal study designs utilizing high-throughput techniques are not unprecedented. DNA microarray experiments with repeatedly measured samples over time, have been performed already more than two decades ago [77–80]. Similarly, timecourse experiments applying RNA-sequencing transcriptomics have been increasing in popularity during the past decade [81–84]. While longitudinal study designs in proteomics experiments have not so far been as popular as the traditional cross-sectional studies, also proteomics experiments with several time points have started to emerge [85–89].

Proteomic intensity data, such as the label-free peak intensity data discussed in the previous chapters, is continuous numerical data and is typically close to normally distributed after the logarithm transformation [31]. While no specialized longitudinal differential expression methods for proteomics data exist, multiple such methods have been developed in the context of DNA microarrays [81,90–92] and RNA sequencing [84]. Similar to proteomics data, DNA microarray data is continuous numerical intensity data distributed approximately normally after the logarithm transformation [93]. In addition to the longitudinal differential expression methods designed specifically for longitudinal high-throughput data, various more general frameworks can be applied in detecting the longitudinally DE proteins. Analysis of Variance (ANOVA) together with several different types of linear and non-linear regression models, with or without random effects, have been applied in analyzing longitudinal data more generally [88,94–99]. A popular approach for grouping the samples in longitudinal omics data is also *clustering*, where samples of the data are clustered based on similarity of their temporal profiles using various methods such as correlation analysis [100], hierarchical clustering [101] or unsupervised machine learning approaches [101].

While similar to other types of longitudinal high-throughput data, proteomics data has some special characteristics that create additional challenges for the applied methods. Some of these challenges, such as the presence of a high proportion of missing values and a high degree of experimental technical variation, has already been discussed in the previous chapters. However, as no optimal solution for these problems has yet been developed, the statistical framework applied in detecting the DE proteins should be tolerant against such challenges. The high degree of noise present in proteomics data renders the methods subjective to false positive and false negative detections, especially for the low intensity proteins.

In **publication III**, several specialized high-throughput longitudinal DE methods together with traditional modelling approaches and a novel method designed especially for longitudinal proteomics were evaluated. The methods were evaluated in their performance to correctly detect the truly longitudinally DE proteins using almost two thousand semi-simulated label-free proteomic spike-in datasets (for spike-in data, see definitions). Furthermore, the methods were tested for their tolerance against missing values, in their reproducibility, and their ability to produce biologically meaningful results.

2.7 Knowledge enrichment through integrated functional enrichment and network analysis

As discussed in the previous chapter, a typical approach for discovering the potential biomarker candidates is detecting differential expression between the experimental conditions of interest. Other approaches in discovering the potential biomarker candidates, or more generally the proteins of interest (POI), might include exploring the interaction partners for a known/suggested POI [102–105] or evaluating the genome- or proteome-wide effects of silencing a known/suggested POI [106–108]. Typically in most approaches, a resulting candidate list of potentially interesting proteins is acquired. After the reliable discovery of such POI, generally some further analysis are required to deepen the knowledge about the candidates and allow for possible interpretations into the underlying biological mechanisms [5]. For example, the discovered protein candidates can be functionally annotated and connected to known biological processes, molecular locations and pathways. As proteins are fundamentally interacting molecules forming complexes, signaling pathways and complicated interaction networks, the analysis of protein-protein interactions plays an essential and natural role in interpreting the results from proteomic experiments [109–112]. Existing and predicted interactions between the proteins can be investigated using protein-protein interaction (PPI) databases [109,112]. Interaction networks of the candidate proteins can reveal tight clusters of co-expressing proteins and possibly proteins with a central role in the interaction networks (i.e. hub proteins). Combining multiple levels of functional annotation together with protein-protein interactions can result in informative networks, essential for the meaningful interpretation of the results.

A popular approach in analyzing functional enrichment in high-throughput data is gene ontology (GO) [113,114]. In gene ontology, genes are annotated functionally into *terms* which are ontologically related to other terms [114]. The terms are hierarchically related to each other; a parent term (such as immune response) might have multiple more specific child terms (e.g. adaptive immune response, humoral immune response). There are three main classes of GO term annotations: biological processes (BP), molecular functions (MF) and cellular compartments (CC) [114]. The enrichment of GO terms in the feature candidate list can be performed through several specialized tools, such as AmiGO [115], the Database for Annotation, Visualization and Integrated Discovery (DAVID) [116,117], Protein Annotation Through Evolutionary Relationship (PANTHER) [118,119] and others.

For GO enrichment, the features in the result list (e.g. proteins) are first converted to gene annotations. Following, the enrichment of GO terms in features of the result list is explored as statistical overrepresentation against what would be randomly expected from a gene set of the size of the result list (e.g. with Fisher's exact test, binomial test [110]). In whole genome or transcriptome experiments, the used background genes defining the expected GO terms and the expected gene counts of GO terms, are typically the whole genome [115]. However, this might be problematic, as necessarily not all the genes can be expected to be expressed in a sample of a given type (e.g. cancerous tissue, blood serum, etc.) [120,121] or detected with equal reliability [121]. This is especially true for proteomics data, as only a subset of all the possible protein coding genes are expressed in a given tissue at a time [122]. A conservative choice for a background universe for the enrichment analysis in proteomics might then be all the proteins reliably detected in the experiment [110,120]. However, a unianonymous consensus concerning the selection of an appropriate background to be used for a statistical overrepresentation analysis has yet to be reached [121]. Another popular approach in exploring the enrichment of GO terms within the data, is gene set enrichment analysis (GSEA) [123]. As opposed to analyzing enrichment in a list of defined candidates (e.g. DE proteins), GSEA can be used to analyze the functional enrichment (including the GO terms) within the whole data, based on a selected score ranking the proteins (e.g. p-values, fold change). An advantage of GSEA is that no specific input list or background needs to be defined, rather the whole ranked data is used as an input list [123,124]. However, if the interest is specifcally in functional enrichment among a specific group of proteins, such as the DE proteins, GSEA might not be the most suitable available tool.

In addition to GO terms, pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [125], Reactome [126], Biocarta [127] and the Pathway Interaction Database (PID) [128], can be explored for additional biological insight. A pathway can be described as a set of chemical reactions leading to a detectable change in cell state [124]. These pathway databases include a high number of various interaction data arising from e.g. metabolism or signaling pathways, genetic interactions or drug development studies [124]. Similar to the GO terms, statistical overrepresentation or GSEA can be used to analyze enriched pathways in the input list or the data in general [110,124]. Although KEGG, Reactome, PID and Biocarta can be considered as some of the most popular pathways databases [109,110,124], altogether several hundreds of databases for biological pathways exist [129]. Another very popular complete commercial pathway and network analysis tool with a broad and constantly updated functional annotation information database is the Ingenuity Pathway Analysis (IPA, https://www.qiagenbioinformatics.com).

Although useful in providing biological insight and knowledge about the overall functionality within a group of proteins, ontological enrichment of related terms (such as GO) or pathways, do not provide information about the relative importance of specific proteins within the input list or data [120]. However, investigating the interactions, or the "local interactome", within the input list in addition to functional enrichment, can provide additional information into the relative importance of proteins in the result list [120]. Network databases such as STRING [112], BioCarta [127] or ReactomeFIViz [130] accommodate information about the known experimentally detected as as well as predicted interactions between proteins. The STRING database includes known and predicted interactions from more than 24 million proteins in 5090 organisms [112]. The gathered interactions from multiple sources are categorized according to their type (e.g. experimentally determined, co-occurrence, text mining, etc.) and scored according to their strength [112]. Given a list of input proteins, STRING composes an interaction network with the proteins of the input list as nodes of the network and the scored interactions between the proteins as edges of the network. Various sorts of processing (e.g. clustering, filtering based on interaction strength) can be performed within the on-line tool and the network can be downloaded for further processing [112].

In **publications IV** and **V**, various functional annotation and network analysis tools were used to perform integrative knowledge enrichment for the detected proteins of interest.

3 Aims of the Thesis

This thesis aims to discuss some of the major challenges related to the processing and analysis of label-free discovery proteomics data and suggests best practices and improvements where applicable. More specifically, the thesis addresses issues concerning:

- Normalization of the raw data (**publications I and IV**).
- Choice of the data processing software (**II**).
- Missing values and imputation (**II**).
- Differential expression detection in longitudinal experiments (**III**).
- Knowledge enrichment for the proteins of interest (**IV and V**).

Finally, on the basis of the performed research work, the thesis aims to provide suggestions and guidelines for a complete label-free proteomics data processing discovery workflow.

4 Materials and Methods

4.1 Datasets

The *UPS1* dataset. The technical benchmarking dataset of [131] contains forty eight (48) Universal Proteomics Standard Set 1 (UPS1) proteins spiked into an unchanging yeast proteome background digest. Five different concentrations were used for the spike-in proteins: 2, 4, 10, 25 and 50 fmol/µl. For every sample, three technical replicate runs were analyzed using a LTQ Orbitrap Velos MS. The UPS1 benchmarking dataset is freely available via the ProteomeXchange Consortium [132] partner repository, the Proteomics Identification Database (PRIDE) [133] with the identifier PXD002099. The UPS1 dataset was used in **publications I, II and III**.

The *UPS1B* dataset. The technical proteomic standard dataset of [134] is similar to the UPS1 dataset with forty eight (48) UPS1 proteins spiked into an unchanging yeast proteome, but with different concentrations. For the UPS1B dataset, nine different concentration groups for the spike-in proteins were utilized: 0.05, 0.125, 0.25, 0.5, 2.5, 5, 12.5, 25 and 50 fmol/µl. Again for each sample, three technical runs were analyzed usin a Orbitrap Velos MS. The UPS1B dataset is also openly available in the PRIDE [133] archive with the identifier PXD001819. The UPS1B dataset was used **publication II**.

The *CPTAC* dataset. The dataset of [135] Clinical Proteomic Tumor Analysis Consortium (CPTAC) Study 6 consists of forty eight (48) UPS1 proteins spiked into a steady yeast proteome digest in five different concentrations: 0.25, 0.74, 2.2, 6.7 and 20 fmol/µl. For each concentration sample, three technical replicate runs were analyzed using a Orbitrap MS (test instrument for this dataset was located at test site 86). The CPTAC dataset is freely available from the CPTAC-portal and was used for **publications I, II and III**.

The *SGSDS* dataset. The Shotgun Standard Set Data Set (SGSDS) of [136] includes 12 non-human proteins spiked into an unchanging human proteome [human embryonic kidney cell line proteins (HEK-293)]. Eight (8) different samples with established concentrations of the non-human spike-in proteins in three master mixes were generated for the dataset and three technical replicate runs for each sample were analyzed using a Q Exactive Orbitrap MS in both DDA and DIA modes. In **publications I, II and III,** the DDA mode dataset was utilized for benchmarking the methods. The SGSDS dataset can be freely accessed from PeptideAtlas [137] (username PASS00589, password WF6554orn).

***Mouse* data**. In addition to the spike-in datasets, a mouse dataset of [138] was used to evaluate the different normalization methods in **publication I**. The used mouse dataset includes liver samples from seven wild-type male mice together with samples from five male mice genetically modified to overexpress the aromatase enzyme cytochrome P450. Samples in the mouse dataset were analyzed with a LTQ Orbitrap Velos Pro MS. The mouse data is available from the ProteomeXchange repository [132] with the identifier PXD002025.

The *Francisella tularensis* subspecies *novicida* (*Fn*) dataset. The *Fn* dataset used in **publication III** for evaluating the performance of the longitudinal differential expression methods, consisted of a wild type strain and three null mutant strains of the acyltransferase enzymes LpxD1 (D1) [139], LpxD2 (D2) and LpxL (L) [140]. The modified acyltransferases are related to the production of important *Fn* membrane proteins and are essential components of the *Fn* lipolysaccharide pathway. Three biological replicates in each strain were measured in five temperatures: 18°C, 21°C, 25°C, 32°C and 37°C. Three technical replicate runs of each biological sample were analyzed using a LTQ Orbitrap Elite MS. The dataset consisted of 180 samples in total and has been stored in the PRIDE repository with the identifier PXD025439. The dataset will be released upon publication of the manuscript in a peer-reviewed journal. Details related to the dataset can be found in **publication III**.

The *Th17 proteome profiling* data. In **publication IV**, human peripheral blood mononuclear cells (PBMCs) were isolated from the umbilical cord blood of five healthy neonates. Naive CD4+ (Thp) cells were further purified from the isolate and either activated by T cell receptor (TCR) cross-linking with CD3 and CD28 antibodies (Th0 cells) or polarized with a cytokine cocktail in combination with TCR/CD28 cross-linking to commence Th17 cell differentiation. A Q Exactive HF MS was used to analyze three technical replicate runs of each biological sample at 24h and 72h after the onset of Th17 differentiation. Altogether 75 samples were analyzed. For details about the dataset, see **publication IV** [85]. In addition to **publication IV**, The Th17 proteome profiling data was used as a background reference proteome for the enrichment analysis in **publication V** and for an additional case study into normalization in this thesis. The dataset is availaible from the PRIDE [133] repository with the identifier PXD008973.

The *cancerous Inhibitor of protein phosphatase 2A (CIP2A) interactome*. In **publication V**, white blood cells were isolated from the umblical cord blood received from the Turku University Hospital. Naive CD4+ T cells were further purified from the isolate, activated with CD3 and CD28 antibodies (Th0 cells) and polarized with a cytokine cocktail to initiate differentiation into Th17 cells. After

72h the onset of Th17 polarization, CIP2A immunoprecipitation was performed using two separate antibodies recognizing distint regions of CIP2A and respective Immunoglobulin G (IgG) antibodies for control immunoprecipetates (IP). Each of the IPs were analyzed from two biological replicates using a Q Exactive HF MS. The IP-MS data consisted of eight samples (four CIP2A IPs, four IgG controls) altogether. For more details about the dataset, see **publication V** [102]. The data was deposited in the PRIDE [133] archive with the identifier PXD008983.

4.2 Normalization in proteomics

The evaluated methods in **publication I** consisted of a baseline transformation and 10 various normalization approaches [32]. The used baseline transformation was the **log$_2$-transformation**, commonly applied for high-throughput data prior to normalization and/or downstream data analysis [32]. The logarithm transformation is typically applied for various high-throughput data to make the data more normally distributed for statistical testing [31,37]. Furthermore, the logarithmic transformation allows the variances of the observed abundance measurements to be less dependent on the absolute magnitude of the measured abundances [34,141], a phenomenon typical for both DNA microarray and MS data. Similar deviation along the whole intensity range is a desirable quality in data, as it enables more equal detection of differences within the lowly and highly expressed features alike. In addition, the logarithm transformation transforms multiplicative relationships between measurements in the data to additive, allowing for simpler models to be used in normalizing the data [141,142]. A base two for the logarithm transformation is typically used for ease of interpretation; a log$_2$ fold change (FC) of one corresponds to a fold change of two, a log$_2$ fold change of two corresponds to a fold change of 4, etc. [31,142]. The log$_2$-transformed protein intensity of protein i in sample j can be represented as:

$$y'_{ij} = log_2(y_{ij})$$

where y_{ij} is the intensity for protein i in sample j, $i=1,...,n, j=1,...,p$, n is the number of proteins and p is the number of samples.

Another common transformation related to high-troughput techniques and used especially in conjunction with normalization, is the **MA transformation**. The MA transformation is based on the MA plot, which enables an easy comparison of abundance levels, patterns and possible biases between two samples [37,141]. It is generally assumed that most features are not differentially expressed in a given high-troughput dataset [34,141]. Under this assumption, for non-biased data or data properly normalized, the data points in an MA plot between the compared samples should be centered around the x axis (M=0) [37]. The M represents the difference between the two compared samples and the A is the average of the two samples. For example, the A value for protein *i* between samples *j* and *k* is:

$$A_{ijk} = \frac{y'_{ij} + y'_{ik}}{2}$$

while the M value is:

$$M_{ijk} = y'_{ij} - y'_{ik}$$

If the data points are not centered around the x axis throughout the whole length of the x axis (changing A values), we can assume that there is bias in the data and can try to correct this bias with normalization. For example, in **Figure 8**, it can be clearly observed that the samples in the unnormalized UPS1 spike-in data (**Figure 8A**) have different expression levels and a constant difference (M-values) of 1-2 for most of the proteins. If we were to make conclusions from the unnormalized data, we might detect a large number of DE proteins, which in the case of this spike-in dataset would be an erroneous conclusion, as the background proteins (colored black in **Figure 8**) are known to remain constant between the conditions. This bias is removed with normalization (**Figure 8B)** and the downstream conclusions are made more reliable.

4.2.1 Evaluated normalization approaches

The evaluated normalization approaches were roughly divided into two categories for this thesis: general approaches and specialized approaches. The general approaches are common approaches used for normalization in high-troughput data, whose form is not necessarily fixed and several variants of these methods can exist. The specialized methods on the other hand are specific algorithms or statistical frameworks designed specifically to normalize the data in a certain defined manner or to achieve a specific structure for the data.

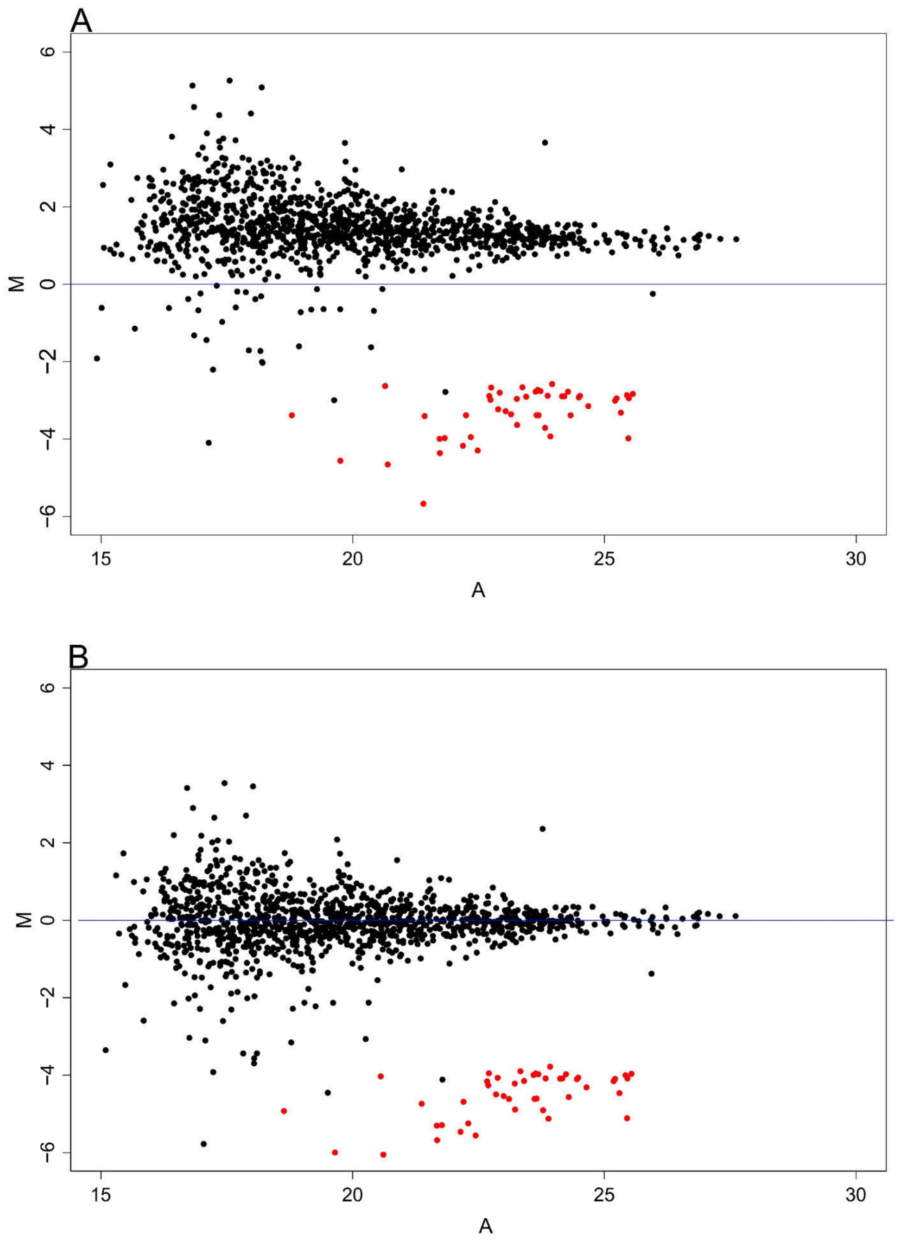


Figure 8. MA-plots. An MA-plot of the 2fmol and the 25fmol samples in A) the unnormalized log$_2$-transformed UPS1 data, and B) the normalized UPS1 data. The stable backround proteins are colored black while the truly changing spike-in proteins are colored red.

General approaches

Linear regression normalization (Rlr, RlrMA, RlrMACyc)

Sometimes, the unknown bias in the data can be assumed to be linearly dependent on the magnitude of the observed protein abundances. This type of linear dependency can be caused for example by the expansion of the measured protein intensities due to sample carry over on a LC column [37]. To normalize for this type of bias, a linear regression normalization can be performed. The general form of the linear regression normalization can be simply described as:

$$y_{ij}^{*} = \frac{(y'_{ij} - b_0)}{b_1}$$

where y_{ij}^{*} is the normalized abundance of protein i in sample j, b_0 and b_1 are the intercept and the slope from the fitted linear regression model, respectively.

The linear regression normalization can be performed in multiple ways. All the samples can be adjusted to one another in a cyclic manner or all the samples can be compared to a common reference sample, typically a mean or a median sample [29,33,36]. In **publication I**, variants of linear regression normalization approaches using a median reference sample (Rlr and RlrMA) as well as a cyclic approach (RlrMACyc) were evaluated. In the RlrMACyc approach, no predefined reference array was defined but instead the MA transformation and the normalization was always performed pairwise between two samples [32]. All samples were iterated in this way and the cycle was repeated three times, observed to be enough to reach convergence between iteration cycles [33,36]. In the RlrMA and the RlrMACyc approaches the data was MA transformed prior to normalization.

Local regression normalization (LoessF, LoessCyc)

In addition to being linear in nature, the bias in the data can also be non-linearly dependent on the observed protein abundances. A non-linear bias in the protein abundances can occur e.g. when the peptide abundances near the detection saturation limit of the instrument or by ion suppression effects [37]. One approach to account for a non-linear bias is *local regression* (loess) normalization [37,143]. In local regression, linear regression curves are fit locally to selected subpopulations or neighborhoods of proteins in the sample to predict for the normalized protein abundances. Similar to the linear regression approaches, variants of loess using a common average reference sample (LoessF) and a pairwise cyclic loess variant (LoessCyc) were explored. Loess was performed by using the loess normalization functions [144] included in the limma-package [91]. The data was MA transformed prior to the loess normalizations [32].

Median normalization (Median)

If the samples in the data are assumed to differ by a constant factor, the data can be normalized using a median normalization. Such a constant difference between the samples can occur for example due to different masses of peptides injected from different samples into the MS system [37]. With median normalization, the samples are scaled to have the same median, typically 0 or the median or mean of median intensities of the samples [29]. The median normalized intensity for protein i in sample j then becomes:

$$y_{ij}^{*} = (y_{ij}' - Med(y_j')) + c$$

where $Med(y_j')$ is the median intensity over all the protein intensities in sample j and c is the level where the new medians of each sample are adjusted to, such as the mean of median intensities:

$$c = \frac{\sum_{j=1}^{p} Med(y_j')}{p}$$

where p is the number of samples. Median normalization has been observed to be effective in practice when analyzing MS-data [145].

Specialized approaches

Quantile normalization (Quantile)

The quantile normalization is another evaluated normalization method developed first for DNA microarrays [33]. The quantile normalization is a non-parametric method and makes no assumptions about the specific nature of the bias, but instead aims to make the distributions of the samples in the data similar [33]. The quantile normalization for matrix M (*n* x *p)* consisting of *n* proteins and *p* samples can be summarized in a few steps:

1) Each sample *j* of the matrix M is ordered into ascending order based on the protein intensity values in the sample to get the sorted dataset matrix M_{sorted}.
2) The intensity values in row 1 in each sample *j* of the ordered dataset M_{sorted} are replaced by the mean intensity value over all the samples in row 1. The same procedure is repeated for all the rows of M_{sorted}.
3) The normalized protein intensities are obtained by reorganizing the rows of each sample *j* in M_{sorted} into the same order as in sample *j* in the original matrix M.

Variance stablization normalization (Vsn)

As mentioned earlier in this chapter, the variances of the observed protein abundances are typically dependent on the magnitude of the abundances in the untransformed data [34,141]. The variance stabilization normalization (Vsn) is a complete statistical method replacing the logarithm transformation. Vsn was first introduced for DNA microarrays by [34]. Vsn aims at making the samples of the data more comparable by transforming the data in a manner decreasing the dependence between the variances and the magnitude of abundances [34]. While the mean-variance dependence is accounted through a set of parametric transformations, the samples are simultaneously normalized to the same scale through a collection of linear mappings [34]. A detailed description of the method can be found in [34].

EigenMS Normalization (EigenMS)

Similar to the quantile normalization, EigenMS does not make any specific assumptions about the nature of the bias in the data [30]. The EigenMS normalization follows the approach of the surrogate variable analysis (SVA) [146], where the essential steps include [30]:

1) Composition of a model for protein expression to estimate the effects of the experimental factors using knowledge about the experimental design.
2) Singular value decomposition (SVD) on the model residuals to explore systematic trends remaining in the unexplained variation.
3) Use of the observed additional trends in the residuals as factors to be adjusted for in the downstream inferential model.

EigenMS adopts the approach of SVA, developed for microarrays [146], into the proteomics environment and further complements it with a rescaling algorithm. In the rescaling part, the systematic variability removed by the normalization method is replaced by a small amount of random variation. This rescaling is done in order to achieve approximately the correct number of degrees of freedom to be able to calculate valid significance values for the normalized data in the downstream analysis [30]. The EigenMS normalization aims in conserving the original group level differences in the data while removing the technical bias. Details about the method can be found in [30].

Progenesis normalization (Progenesis)

As Progenesis was used to process the data in this study, Progenesis normalization was also included in the comparison. The Progenesis normalization is similar to the median normalization; a global normalization factor is calculated to normalize between the samples. An automatically selected representative reference sample is used to calculate this scaling factor. More information about the normalization procedure provided by the software can be acquired from the Progenesis website (http://www.nonlinear.com/progenesis/qi-for-proteomics/v3.0/faq/how-normalisation-works.aspx).

4.2.2 Evaluation of normalization

Typically, the normalization methods in proteomics have been evaluated in their ability to decrease **intragroup variation** [29,36,37]. The examined sample groups might be composed of technical or biological replicates. Especially in the case of technical replicates, decreased intragroup variation indicates successful normalization, as unwanted technical variation is removed from the data. There are multiple ways to visualize the effect of normalization on data. Some common plot types to explore the need and effect of normalization include boxplots (**Figure 4**), MA plots (**Figure 8**), scatter plots (**Figure 9A-B**) and quantile-quantile plots (**Figure 9C-D**).

In **publication I**, the normalization methods were evaluated in their ability to decrease intragroup variation between technical and biological replicates. The UPS1, CPTAC and SGSDS spike-in datasets introduced in chapter 4.1, were used to evaluate the normalization methods ability to decrease unwanted intragroup variation between the technical replicates. In addition to the spike-in datasets, the experimental proteomics mouse dataset of [138] (chapter 4.1) was used to explore the ability of the normalization methods to decrease intragroup variation between biological replicates. As variability measures, the intragroup *pooled median absolute deviation* (PMAD), *intragroup pooled coefficient of variance* (PCV) and *intragroup pooled estimate of variance* (PEV) were used.

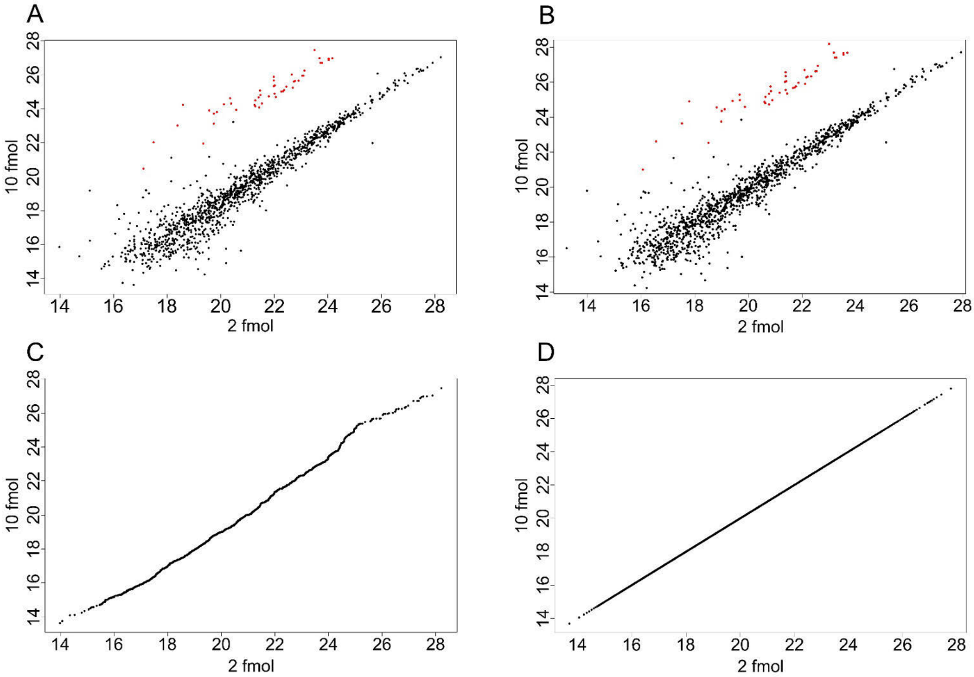


Figure 9. Scatter plots of the 2 fmol and the 10 fmol samples in the A) unnormalized UPS1 data, B) normalized UPS1 data. The background proteins are colored in black and the spike-in proteins in red. Quantile-quantile plots of the 2 fmol and 10 fmol samples in the C) unnormalized UPS1 data, D) normalized UPS1 data.

The PMAD for sample group *l* consisting of technical or biological replicates can be simply defined as:

$$PMAD_l = \frac{\sum_{i=1}^{n} MAD_{il}}{n}$$

where MAD_{il} is the *median absolute deviation* (MAD) for proteins *i=1,…,n* in sample group *l* and *n* is the number of proteins in the data. PCV for sample group *l* is defined as the mean *coefficient of variation* (CV) over all the proteins *i=1,…,n*:

$$PCV_l = \frac{\sum_{i=1}^{n} CV_{il}}{n}$$

where the CV for protein *i* in sample group *l* is:

$$CV_{il} = \frac{s(y_{il}^*)}{m(y_{il}^*)}$$

$s(y_{il}^{*})$ is the standard deviation and $m(y_{il}^{*})$ is the mean of the normalized protein abundances for protein *i* across the samples in sample group *l*. PEV is defined as the mean of variances in sample group *l* over all the proteins *i=1,...,n* :

$$PEV_l = \frac{\sum_{i=1}^{n}(s(y_{il}^{*})^2)}{n}$$

where $s(y_{il}^{*})^2$is the variance of protein *i* across the samples in sample group *l.*

As the aim of many proteomics experiments is to detect differentially expressed (DE) proteins between the experimental conditions, the effect of the normalization method on the correct definition of the DE proteins is typically of utmost interest for the researcher. While [36] explored the effect of a few normalization methods on the number of the DE proteins detected, no prior comprehensive evaluation of the effect of the normalization method to the validity of the detected DE proteins had been performed. Therefore, in addition to examining the normalization methods ability to decrease intragroup variation, we evaluted how well true **differential expression** could be detected from data normalized with the different methods in **publication I** [32]. As explained in the definitions chapter, the TPs (spike-in proteins) and the TNs (background proteins) are known in the spike-in datasets, enabling the benchmarking of the methods with a receving operator characteristic (ROC)-curve analysis (see the definitions chapter). The area under the ROC-curve (AUC) then describes how correctly true differential expression could be detected from data normalized with the different methods. Furthermore, it was evaluated how well the true known **fold change** (FC) between the examined samples could be estimated from the data normalized with the different methods. As the data was logarithm transformed prior to normalization and the results were examined on the $\log_2$ transformed scale, *logarithmic fold change* (logFC) was used instead of FC. The logFC of protein *i* between sample *j* and sample *k* was defined as:

$$logFC_{ikj} = y_{ik}^{*} - y_{ij}^{*}$$

Finally, as the data can be normalized in various ways depending on the samples examined, it was evaluated whether the **stage** in which the data is normalized had a major effect on the performance of the methods. It was explored whether performing the normalization globally on the whole data or normalizing only the examined samples separately in each examined comparison, had a major effect on the performance of the normalization methods.

4.2.3 An additional case study into normalization

The data in **publication IV** consisted of discovery work into the cellular proteome during Th17 cell polarization using a quantitative label-free DDA proteomics approach. The proteome profiles of CD4+ human T cells, CD3/CD28 activated T (Th0) cells, and Th17 cells were explored at 24h and 72h after the initiation of the polarization [85] (chapter 4.1).

As opposed to the work performed related to **publication I** and the spike-in datasets with relatively few DE proteins, there were considerable changes in protein expression in the Th17 proteome of **publication IV** during the polarization process (**Figure 10A**). Exploration of the non-normalized data of **publication IV** indicated a clear need for normalization (**Figure 10B-D**). While Progenesis was used to process the data for evaluating the normalization methods in **publication I**, MaxQuant [42] was used to analyze the Th17 quantitative proteomics data [87] for **publication IV**. Based on the experiences gained from **publications I and II**, it was known that different normalization methods varied in their performances but also different software workflows resulted in data from which the truly DE proteins could be detected with varying accuracies. Due to these reasons, *Vsn, LoessF and Median* normalizations, each noted to perform consistently in the evaluation work of **publication I** and represent different approaches into normalization, were explored together with the MaxQuant innate normalization method MaxLFQ [18] in order to choose the most suitable normalization approach for the data in **publication IV**.

The MaxLFQ algorithm [18] is a combined strategy for normalization and protein quantification implemented in the MaxQuant software [42]. MaxLFQ allows for the normalization of samples in large proteome wide experiments even when the samples are fractionated prior to the MS analysis [18]. The algorithm determines a normalization coefficient for each sample based on the extracted ion chromatograms (XIC) for the peptides. It is assumed that the majority of the peptides do not change between the samples and the normalization coefficients are optimized as such that the changes in XIC between the samples are minimized for the bulk of the peptides. In addition to normalizing the samples, MaxLFQ aims to estimate protein quantities as accurately as possible [18]. Protein quantities are calculated by determining pairwise protein ratios between all samples using only the common peptides present in both evaluated samples. A median ratio over all the common peptide ratios for a protein is used as a representative protein ratio between the examined samples. A least squares approach is then used to calculate the LFQ protein intensities from the matrix of pairwise protein ratios over all the samples. In addition to the LFQ approach, protein intensities in MaxQuant can be calculated using the intensity-based absolute quantification (IBAQ) method [57], where peptide intensities for a protein in the given sample are summed up and divided by the number of theoretical peptides for the given protein. Even though assuming that most proteins remain unchanged, the authors report good performance using the LFQ approach even when 30% of the proteome was changing [18]. To compare to the MaxLFQ normalized data, IBAQ data from MaxQuant was extracted and normalized with the previously mentioned selected best approaches from **publication I**.

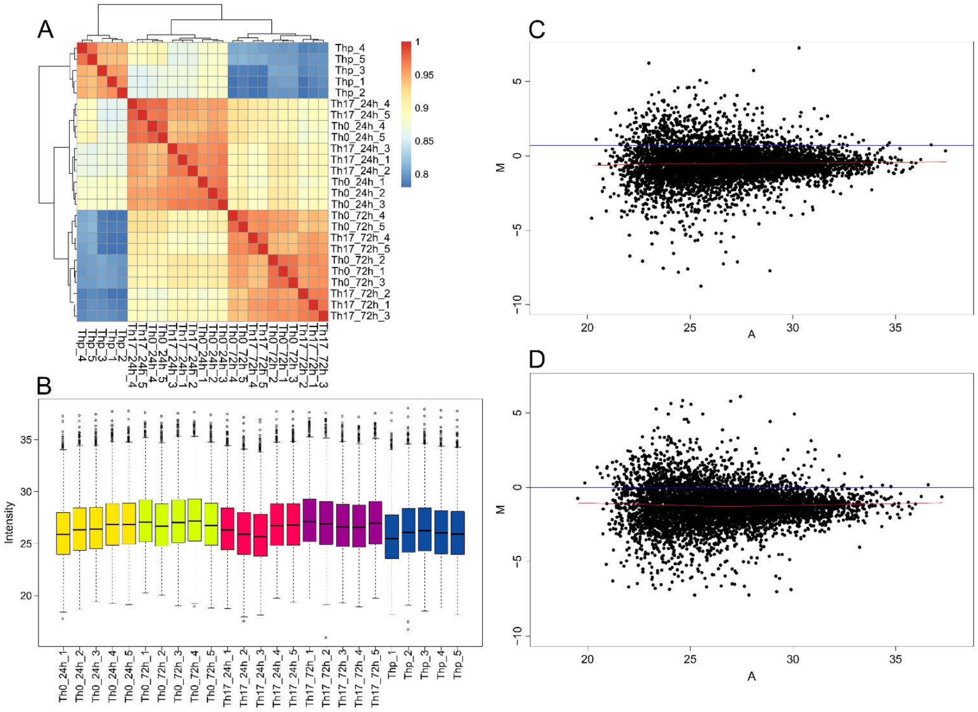


Figure 10. A) Correlation heatmap of the data in **publication IV**. Pearson correlation coefficients have been calculated pairwise between all samples. The samples are clustered using hierarchical clustering with complete linkage. B) Distributions of the samples as boxplots. Samples are colored according to sample groups in the data. MA-plots between the C) Th0 sample for replicate 1 at 24h and Th17 sample for replicate 1 at 24h and D) Th17 sample for replicate 3 at 24h and Th17 samples for replicate 5 at 72h. The red curve represents a loess fit between the M and A values. For all visualizations, the data has been $\log_2$-transformed and the technical replicates for each biological replicate have been averaged.

4.3 Label-free proteomics data processing software tools

Evaluation and comparison of the software workflows in **publication II** was performed using common settings as much as possible in the different evaluated software. Within each software workflow, the proper level of instrumentation was selected. The same spike-in datasets and FASTA files were used for each software [60]. The UPS1, CPTAC, SGSDS and UPS1B spike-in datasets (chapter 4.1) were used to evaluate the performance of the different software. Furthermore, the same search modification, digestion enzyme, peptide length, precursor and fragment mass tolerance settings were used for all the software workflows. Peptides unambiguously mapping to one protein (non-conflicting peptides) were used in calculating the relative protein-level abundances. Non-normalized protein intensities were extracted from each software workflow and normalized with the Vsn normalization previously detected to perform well with proteomics data in **publication I** [32].

4.3.1 Evaluated label-free software workflows

Commercial solutions

Peaks

Peaks Studio is a complete commercial proteomics software workflow. The Peaks Studio software package contains multiple specialized tools for various analysis purposes. PEAKS DB [46] is a database searching tool with incorporated de novo sequencing included in the Peaks Studio software. Peptides with unspecified single amino acid mutations can be searched using the SPIDER [51] search tool. In addition, PEAKS PTM can be used to search proteins and peptides with unspecified post translational modifications [147]. In **publication II**, Peaks was allowed to automatically identify the reference sample and align the runs.

Progenesis

Progenesis is a commercial proteomics quantification workflow. The Mascot [11] search engine within the Thermo Fisher provided software Proteome Discoverer was used to produce the peptide and protein identifications imported into Progenesis. Progenesis was allowed to perform automatic alignment of the MS runs.

Freeware and open source solutions

MaxQuant

MaxQuant [42] is a popular non-profit complete proteomics software workflow. MaxQuant uses its own built-in Andromeda [14] search engine to identify the peptides and the proteins. Quantification in MaxQuant is performed using either the IBAQ [148] method or the LFQ [18] as described in chapter 4.2.3. For this comparison, the IBAQ method was used to allow for the normalization of the data with the Vsn normalization. MaxQuant was allowed to automatically align the runs and transfer identifications between the runs.

OpenMS

OpenMS is an open source software environment for MS data processing algorithms [41]. OpenMS is fully modular; the user can combine various modules to construct a desired workflow. Workflows can be constructed for different types of MS-data with the use of distinct modules. Toppas [149] (Figure 6) is a graphical workflow editor which can be applied in building the OpenMS workflows instead of the command line user interface. Several related suitable algorithms, such as PeakPickerHiRes, BaselineFilter, FeatureFinderCentroided, MapAlignerPoseClustering, FeatureLinkerUnlabeledQT and ProteinQuantifier, were used to construct a full label-free proteomics workflow with identification and quantification. The X!tandem [12] and MS-GF+[13] search engines were used to identify the peptides and the proteins. The MapAlignerPoseClustering algorithm was allowed to automatically select a reference run and align the samples.

Proteios

Proteios [45,150] is another open source modular proteomics software environment. Proteios is designed to operate in a server environment and to be a repository in which all the metadata as well as the actual experimental data can be stored and managed in addition to performing various analyses on the data [45]. Analogous to OpenMS, distinct algorithms for different analysis tasks can be plugged into the Proteios environment [45]. X!Tandem [151] and MS-GF+ [13] were used to identify the peptides and the proteins similar to OpenMS. The Dinosaur [39] algorithm was used for feature detection and the innate Proteios alignment algorithm [43] to automatically align the runs.

4.3.2 Evaluation of the label-free software workflows

Similar to the evaluation of the normalization methods (chapter 4.2.2), ROC-curve analysis and the related partialAUC (pAUC) values (for the definition of pAUC, see definitions chapter) were used in evaluating the different softwares' ability to correctly detect the truly DE proteins. The ROC curves were drawn both over all the pairwise comparisons in each dataset and also separately for each pairwise comparison of sample groups in each dataset. The associated pAUC values were calculated for all the ROC curves.

The known true fold changes in the different comparisons of the spike-in datasets were used to evaluate each software's ability in estimating the fold changes. The mean squared error (MSE) between the estimated LogFCs and the true known LogFCs were calculated for each software workflow. The MSEs for the spike-in proteins and the non-changing background proteins were calculated separately for each pairwise comparison for each software workflow. The MSE for the spike-in or the background proteins in a given comparison for a software was calculated as:

$$MSE = \frac{\sum_{i=1}^{n}\left(LogFC_{estimated_i} - LogFC_{known_i}\right)^2}{n}$$

where *n* is the number of evaluated proteins.

4.4 Missing values and imputation in label-free proteomics

4.4.1 Evaluated imputation methods

The evaluated imputation methods in **publication II** were divided into four categories for this thesis based on their approach to imputation/missing values: single value approaches, local similarity approaches, global structure approaches and filtering approaches.

Single value approaches

Zero imputation (zero)

The missing values in the data were replaced with zeros. Typically, the Progenesis software does not produce missing values but sometimes produces zero values, indicating non-existent expression. The zero imputation implemented a similar approach, assuming values are missing due to non-existent expression.

Background imputation (back)

The missing values in the data were replaced with the minimum detected value. This approach is similar to the MinDet approach of [61] and assumes most of the missing values are abundance dependent MNARs.

Censored imputation (censored)

Censored imputation is a more sophisticated variant of the background imputation. If a protein in a sample group had only one missing value, it was consired as MCAR and no value was imputed for it. However, if more than one missing value for a protein in a sample group were detected, they were considered as MNAR and similar to the background imputation, a minimum detected value was imputed for them.

Local similarity approaches

Local least squares imputation (lls)

For each imputable protein *i*, the *k* most similar proteins were selected using the Pearson's correlation coefficient [152]. Least squares regression was then used to estimate the missing values as a linear combination of the non-missing values from the *k* most similar proteins. A value of 150 was used for *k*, observed to be suitable in most cases by [153–155].

K-nearest neighbor imputation (knn)

Similar to the lls imputation, *k* most similar proteins (nearest neighbours) for the imputable protein *i* were first selected [156]. Similarity was inferred using the Euclidean distance metric. A weighted average over the values of the *k* nearest neighbours in the imputable sample *j* was used to impute a value for protein *i* in sample *j*. Similarity of the *k* nearest neighbours and protein *i* was inferred using other samples than the imputable sample *j*. A value of 10 for *k* was used, previously observed to deliver good performance by [156].

Global structure approaches

Bayesian Principal Componet Analysis (bpca)

The Bayesian Principal Component Analysis (BPCA) is a global structure based imputation developed first for microarray data [157]. The BPCA algorithm consists of three parts combined together: principal component regression, Bayesian estimation and an expectation maximization like Variational Bayes (VB) algorithm. Estimates for the missing values and parameters for the model, such as scores for the principal components are repetitively updated using the VB algorithm until convergence. Orthogonality between the principal components is not enforced by the VB algorithm [152]. The principal components in BPCA are scaled differently than in standard PCA [152]. Such different scaling suppresses redundant components but can lead to unreliable estimates in the case of small numbers of observations.

Singular value decomposition imputation (svd)

Similar to PCA, the data is reduced to a set of mutually orthogonal expression patterns, the principal components, using singular value decomposition (SVD) [152,156]. Protein *i* with missing values was then regressed against the *k* most significant principal components or eigenproteins [156]. Sample *j* with a missing value for protein *i* was not used in the regression. The coefficients of the regression were used to determine a value for sample *j* in protein *i* as a linear combination of the *k* eigenproteins for sample *j*. A proportion of 20% of the eigenproteins have been previously observed to be a suitable value for *k* [156] and was also used as *k* for SVD in **publication II**.

Filtering approaches

Basic filtering (filtered)

A strict filtering was applied where all proteins having more than one missing value in any sample group were removed. Each sample group consisted of three technical replicates. Thus, a protein was required to have two valid non-missing values in each sample group to remain in the analysis.

Filtering + local least squares imputation (filtlls)

In this approach, filtering was first performed as in the filtered approach followed by imputation of the remaining missing values with the lls imputation.

4.4.2 Evaluation of the imputation approaches

The imputation and filtering approaches were evaluated together with the label-free software data processing software workflows (chapter 4.3) in **publication II**. Each imputation method was benchmarked with each of the five evaluated software in all pairwise comparisons of the UPS1, CPTAC, SGSDS and UPS1B datasets (chapter 4.1). ROC-curves over all the pairwise comparisons in each dataset for each imputation approach were drawn and the related pAUC values were recorded. The pAUC values of the imputation methods were ranked within each dataset and each software. An overall mean rank for each imputation method was calculated over all the examined datasets and software.

4.5 Longitudinal differential expression in proteomics

4.5.1 Evaluated longitudinal differential expression approaches

The evaluated methods in **publication III** were divided into four types for this thesis according to their origin: the baseline method, specialized longitudinal DE methods for high-throughput data, general regression modelling approaches and the new proposed solution.

Baseline method

Reproducibility Optimized Test Statistic (BaselineROTS)

ROTS [158] is a differential expression method observed to perform well on multiple platforms [32,131,159–161], especially in controlling type 1 errors [162]. ROTS maximizes the reproducibility of the top differentially expressed features with group preserving bootstraps [158]. ROTS was used as a baseline method for benchmarking the evaluated longitudinal methods. Differential expression between the conditions was examined at each timepoint and the minimum significance value over all the timepoints was recorded as the representative significance value for each protein.

Specialized longitudinal DE methods

Bayesian Estimation of Temporal Regulation (BETR)

Bayesian Estimation of Temporal Regulation is a longitudinal differential expression method developed for the analysis of timecourse DNA microarray data already a decade ago [90]. BETR accounts for the correlation in expression between the timepoints. Timepoints closer to each other are assumed to be more similar than those further apart. For each feature, two models are fitted. The first model assumes no differential expression between the examined conditions, while the second model considers timepoint correlated differential expression. The Bayes' rule is used to determine the probabilities of the features comings from either of these models.

Linear models for microarray data (Limma, LimmaSplines_L, LimmaSplines_H)

Limma is a well-established analysis toolkit developed for microarray and RNASeq data [91]. At the core of Limma are linear models, which can be utilized in several ways to assess differential expression. In **publication III**, Limma with two different strategies for detecting differential expression were applied. Similarly to [163], the first option (Limma) included designing the coefficient vector for the linear models in a way that each timepoint and condition combination was represented by a separate coefficient. Next, the examined contrasts of coefficients were defined to reflect the differences in expression between the conditions at each timepoint. Finally, it was investigated whether all the separate timepoint contrasts are zero, i.e. if differential expression at any timepoint between the conditions exists. The second option (LimmaSplines) involved fitting polynomial regression splines for each protein and exploring the condition related coefficients [91].

Microarray Significant Profiles (MaSigPro_L, MaSigPro_H)

Microarray Significant Profiles (MaSigPro) was first designed for the analysis of longitudinal DNA microarray experiments by [92]. Subsequently, MaSigPro has been further developed to incorporate the analysis of timecourse RNASeq data [164]. As proteomics data resembles more DNA microarray data than RNASeq data, the performance of the original MaSigPro in the detection of longitudinal differential expression in label-free proteomics data was explored. MaSigPro uses a two-step regression strategy where the significantly longitudinally differentially expressed features are first identified using generally defined polynomial regression models. Secondly, the relevant variables (e.g. time, condition, time*condition) are identified separately for each feature using a stepwise regression approach [92]. Dummy variables are used to encode the polynomial regression models. While the overall time-associated changes in expression over both (all) conditions is explored by MaSigPro by default [92], only the condition associated coefficients of the results were utilized, as the focus was on longitudinal differential expression between the conditions, not on overall longitudinal changes.

Timecourse

Similar to BETR, Timecourse applies a Bayesian framework in detecting longitudinal differential expression [165]. Timecourse uses the Maxwell–Boltzmann and/or the Hotellings t^2 -statistics through a multivariate empirical Bayes approach in ranking the features. Correlations in expression between time points and individual/replicate variances are considered. Furthermore, Timecourse estimates differential expression over all the features simultaneously, borrowing information across features to better estimate the variance-covariance matrices.

General regression modelling approaches

Linear mixed effects regression modelling (Lme)

Together with the specifically designed longitudinal DE methods, several regression based modelling approaches for detecting longitudinal differential expression were explored in **publication III**. Linear regression is a simple approach, where the expression is assumed to change in a linear fashion over time. Differential expression with linear regression can be investigated by examining the intercept and the slope at different levels of a condition related categorical factor included in the model.

Individual variation in the baseline and longitudinal expression can be taken into consideration by using random effects and the mixed modelling approach. Linear mixed modelling has been a popular approach in modelling longitudinal data [88,166,167]. In **publication III**, the linear mixed effects modelling approach was explored via two variants: 1) allowing an individual baseline for the replicates but using a common slope and 2) allowing an individual specific slope in addition to the individual baseline. The first linear mixed effects model variant with an individual baseline and a common slope for a protein was defined as:

$$y = \beta_0 + \beta_1 t + \gamma_o c + \gamma_1 c \cdot t + \ \delta_{0r} + \varepsilon$$

where β_0 is the intercept, β_1 is the linear time-related slope over both conditions and all the replicates, γ_o is the condition (c) related intercept, γ_1 is the condition related linear slope for the protein, $\delta_{0r} \sim N(0, \sigma_{0r}^2), r = 1, \dots, h$ is the random effect for the individual/replicate-specific baseline, h is the number of replicates and ε is the remaining error variation. The average condition related differences in longitudinal expression can be examined by investigating γ_o and differences in linear longitudinal expression patterns between the conditions by investigating γ_1. In the case of the second model variant, a random effect term $\delta_{1r}t, \delta_{1r} \sim N(0, \sigma_{1r}^2), r = 1, \dots, h$, describing the individual specific longitudinal linear experession (slope), was added to the model:

$$y = \beta_0 + \beta_1 t + \gamma_o c + \gamma_1 c \cdot t + \ \delta_{0r} + \ \delta_{1r} t + \varepsilon$$

For each protein with enough information to determine both model variants, a likelihood ratio test was performed to explore if the more complex second variant yielded a significantly better fit. If the fit for the second variant was not significantly better, the first model variant was used.

Polynomial mixed effects modelling (Pme_L, Pme_H)

A polynomial regression model enables the detection of more complex longitudinal trend differences between the conditions. Similar to the linear mixed effects modelling approach, two model variants for each protein were inspected in

publication III. The first variant utilized random effects only for the replicate-specific baseline, while the second variant allowed a random effect also for the linear term. Mixed models incorporating higher order random effects were typically unavailable due to insufficient information to define such models in the short time series data with only few replicates and were thus excluded from the analysis. The first polynomial mixed effects model variant was defined as:

$$y = \beta_0 + \sum_{i=1}^{d} \beta_i t^i + \gamma_o c + \sum_{i=1}^{d} \gamma_i c \cdot t^i + \delta_{0r} + \varepsilon$$

and the second variant as:

$$y = \beta_0 + \sum_{i=1}^{d} \beta_i t^i + \gamma_o c + \sum_{i=1}^{d} \gamma_i c \cdot t^i + \delta_{0r} + \delta_{1r} t + \varepsilon$$

where *d* is the degree of the polynomial. Similar to the linear approach, longitudinal differential expression can be examined by investigating γ_o and γ_i for the different polynomial degrees. For each protein, the second variant with random effects also for the linear term, was selected only if it resulted in a significantly better fit than the first variant. For all the polynomial regression models, orthogonal polynomials were used. As typically, the polynomials of different degree are highly correlated, orthogonal polynomials can decrease such collinearity and allow for a more independent exploration of coefficients of different polynomial degree within the same model [168,169].

Two levels of model complexity were explored for each longitudinal differential expression approach based on polynomial regression (LimmaSplines, MaSigPro and Pme). Less complex models with *t/2* degrees were explored together with more complex models of *t-1* degrees, where *t* was the number of time points. The less complex models are denoted with the extension _L in the results section while the more complex models are denoted with an _H extension, respectively. To detect any condition related longitudinal differential expression, a representative significance value for each protein was selected as the minimum over all the condition coefficient related significance values of the protein for MaSigPro, LimmaSplines and the regression modelling approaches.

The new proposed approach, Robust longitudinal Differential Expression (RolDE)

In high-throughput data, especially in proteomics data, there is typically a great deal of noise in addition to the true biological signal. Such noise variation renders the differential expression methods subjective to false detections and complicates the detection of the true signal. The new method, RolDE, is a composite method, consisting of three independent modules with different approaches to detecting differential expression. The combination of these diverse modules allows RolDE to robustly detect varying changes in longitudinal trends and expression levels in different data types and experimental settings.

The ***RegROTS*** module merges the power of regression modelling with the power of the established differential expression method ROTS [158,170]. A polynomial regression model for longitudinal protein expression is fitted separately for each replicate (individual) in each condition. Differential expression for a protein between two replicates, r_1 and r_2, in different conditions c_1 and c_2, is then examined by comparing all the coefficients of the corresponding degrees of the replicate-specific regression models:

$$\Delta\beta_{r_1c_1r_2c_2} = \left\{\beta_{0_{r_1c_1}} - \beta_{0_{r_2c_2}}, \beta_{1_{r_1c_1}} - \beta_{1_{r_2c_2}}, \dots, \beta_{d_{r_1c_1}} - \beta_{d_{r_2c_2}}\right\}$$

where $\Delta\beta_{r_1c_1r_2c_2}$ are all the coefficient differences between the replicate-specific models of r_1 in c_1 and r_2 in c_2 and d is the degree of the polynomial regression. If all coefficient differences are zero, no longitudinal differential expression between the two replicates in different conditions exist. For a thorough exploration of differential expression between the conditions, all possible combinations of replicates in different conditions are examined. The null hypothesis for a protein becomes:

$$G_1 = G_2 = , \dots , = G_{(d+1)} = 0$$

where G_1 are the coefficient differences between all the replicate-specific models in the different conditions related to the intercept and $G_2, \dots , G_{(d+1)}$ are all the coefficient differences related to different polynomials of the regression models. Using multigroup ROTS, the coefficient differences over all the proteins are investigated simultaneously. To preserve the proper degrees of freedom for statistical testing, multiple runs are typically required, so that each replicate in each condition is used only once in each run, when all possible comparisons between the replicates in the different conditions are performed. The different RegROTS runs are combined by using the rank product. The final score for protein in the RegROTS module then becomes:

$$S_{RegROTS} = \left(\prod_{i=1}^{q} R_i\right)^{\frac{1}{q}}$$

where R_i is the rank of the protein in run i and q is the number of runs.

In the ***DiffROTS*** module, the expression of the replicates in the different conditions are directly compared at all timepoints. Differential expression for a protein between two replicates, r_1 and r_2, in different conditions c_1 and c_2, is examined simply by:

$$\Delta y_{r_1 c_1 r_2 c_2} = \left\{ y_{1_{r_1 c_1}} - y_{1_{r_2 c_2}}, \dots, y_{t_{r_1 c_1}} - y_{t_{r_2 c_2}} \right\}$$

where $\Delta y_{c1r1c2r2}$ are the differences in the normalized protein expression between r_1 in c_1 and r_2 in c_2 at all timepoints, $y_{1_{r_1 c_1}}$ is the expression of the protein for r_1 in c_1 at timepoint 1 and t is the number of timepoints. If the expression level differences at all timepoints are zero, no differential expression between the examined replicates in different conditions exist. Similarly to the RegROTS module, differential expression is examined between all possible combinations of replicates in the different conditions. As with the RegROTS module, multigroup ROTS is used to examine differential expression over all the proteins simultaneously. The groups for multigroup ROTS now consist of all the expression level differences at different timepoints (e.g. G_1 for all the expression level differences for a protein at timepoint 1, G_2 at timepoint 2, etc.). Similar to RegROTS module, the possible comparisons between all the replicates in the different conditions are divided into multiple runs so that each replicate is used only once in each run and the different runs are combined via the rank product for a final score for the DiffROTS module:

$$S_{DiffROTS} = \left(\prod_{i=1}^{q} R_i \right)^{\frac{1}{q}}$$

The ***PolyReg*** module uses polynomial regression to explore differences in longitudinal expression between the conditions. The expression y for a protein in the PolyReg module is described as:

$$y = \beta_0 + \sum_{j=1}^{d} \beta_j t^j + \gamma_o c + \sum_{j=1}^{d} \gamma_j c \cdot t^j + \varepsilon$$

The average condition related differences in longitudinal expression can again be examined by exploring γ_o and differences in longitudinal expression patterns between the conditions by exploring $\gamma_j, j = 1, \dots, d$. For each protein, the results from the PolyReg module are summarized in a final score as the minimum over the significance values of the condition related coefficients:

$$S_{PolyReg} = \min_j p(\gamma_j).$$

where $j=0,1,\dots,d$. Alternative to a fixed effects only model, a polynomial mixed effects model can be used in the PolyReg module with either random effects only for the intercept or for both the intercept and the slope. Both the RegROTS module, as well as the PolyReg module, apply orthogonal polynomials to allow for a more independent investigation of the coefficients of different polynomial degrees within their regression models [168,169].

Finally, to conclusively detect any differential expression for a protein, the scores from the different modules are combined into a final RolDE score using the rank product (geometric mean) of the ranks of the scores from the different modules:

$$S_{RolDE} = \sqrt[3]{R(S_{RegROTS}) \cdot R(S_{DiffROTS}) \cdot R(S_{PolyReg})}$$

For details about the new proposed method, RolDE see **publication III**.

4.5.2 Evaluation of the longitudinal differential expression methods

The explored approaches for the detection of longitudinal differential expression were evaluated in multiple ways: 1) The proportion of proteins each method was able to provide a score for, 2) The performance of the methods in differentiating the known truly longitudinally DE proteins from the known non-DE proteins in the differential expression analysis, 3) The reproducibility of the findings, and 4) The biological relevance of the findings of each method.

Semi-simulated spike-in datasets

For the evaluation of the longitudinal differential expression methods, semi-simulated spike-in datasets with various longitudinal trends and trend differences were generated. In short, the mean values and standard deviations of proteins in sample groups of the UPS1, SGSDS and CPTAC spike-in datasets (chapter 4.1) were used to recreate semi-simulated sample groups with similar levels of protein expression and some variation. Thus, the same sample group could be generated multiple times with some random diversity. Longitudinal trends within a condition were generated by organizing the generated semi-simulated sample groups in the desired combinations (e.g. 2fmol, 4fmol, 10fmol, 25fmol 50fmol for a type of linear trend in the UPS1 dataset) (**Figure 11A**, **Figure 11B)**. As in the original datasets, only the concentrations of the spike-in proteins varied between the sample groups. The expression of the background proteins remained constant between sample groups, excluding experimental noise (**Figure 11C**). The number of sample groups within a condition was the same as in the original datasets, and in each sample group, the number of technical replicates was the same as in the original dataset. Thus, the number of timepoints in each semi-simulated dataset type (UPS1=5, SGSDS=8, CPTAC=5) was the same as the number of sample groups in the original dataset.

Furthermore, the pattern of missing values was directly replicated from the original dataset.

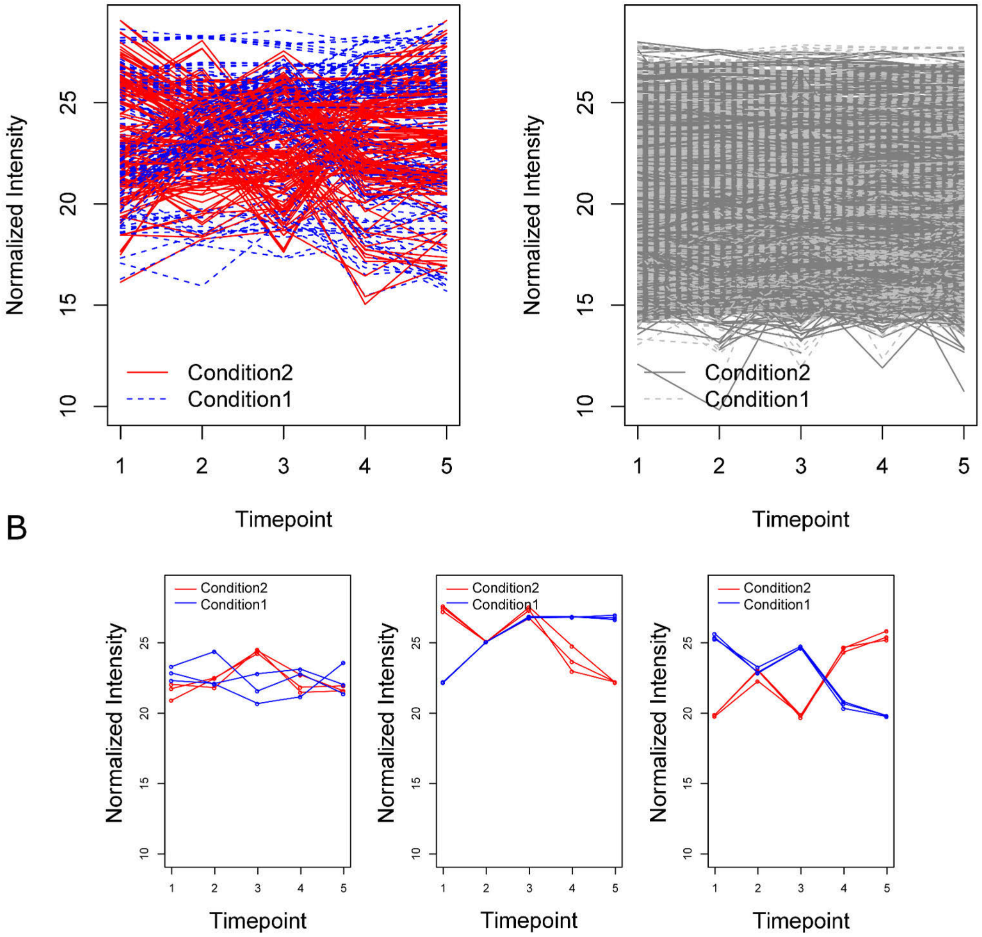


Figure 11. Examples of the generated longitudinal trends in the semi-simulated UPS1 dataset for evaluating the longitudinal differential expression methods in **publication III**. Longitudinal trends for A) all the spike-in proteins, B) selected spike-in proteins and C) all the background proteins in a representative semi-simulated UPS1 dataset. All the examples are from the same semi-simulated UPS1 dataset with mixed trends and trend differences for the the spike-in proteins. The different conditions in the dataset are colored as red and blue for the spike-in proteins. The true positive spike-in proteins in A and B show varying trends and longitudinal differential expression between the conditions. The true negative background proteins in C show a constant and unvarying expression over time and between the conditions.

For each semi-simulated dataset, two conditions with different trends – or same trends but different levels of expression – were combined (**Table 1**). Six basic trend categories were generated (Stable, Linear, LogLike, Poly2, Sigmoid and PolyHigher), with multiple variations of the category trend within each category (**Table 1**).

Table 1. The created longitudinal trends and combinations of trends for the UPS1 semi-simulated spike-in datasets. The longitudinal trends were created for the spike-in proteins by organizing the semi-simulated sample groups in the desired order. All possible combinatios of trend categories are shown in columns three and four.

CATEGORY	LONGITUDINAL TREND (SAMPLE FMOL)		ALL COMBINATIONS OF TRENDS	
STABLE	2, 2, 2, 2, 2		**Condition1**	**Condition2**
STABLE	4, 4, 4, 4, 4		Stable	Stable
STABLE	10, 10, 10, 10, 10		Stable	Linear
STABLE	50, 50, 50, 50, 50		Stable	LogLike
STABLE	25, 25, 25, 25, 25		Stable	Poly2
LINEAR	2, 4, 10, 25, 50		Stable	Sigmoid
LINEAR	50, 25, 25, 10, 4		Stable	PolyHigher
LINEAR	2, 4, 4, 10, 25		Linear	Linear
LINEAR	25, 25, 10, 4, 2		Linear	LogLike
LINEAR	4, 4, 10, 25, 50		Linear	Poly2
LOGLIKE	2, 10, 25, 25, 25		Linear	Sigmoid
LOGLIKE	50, 10, 4, 4, 4		Linear	PolyHigher
LOGLIKE	25, 25, 25, 10, 2		LogLike	LogLike
LOGLIKE	4, 4, 4, 10, 50		LogLike	Poly2
LOGLIKE	4, 10, 50, 50, 50		LogLike	Sigmoid
POLY2	2, 4, 10, 4, 2		LogLike	PolyHigher
POLY2	50, 25, 10, 25, 50		Poly2	Poly2
POLY2	2, 10, 10, 10, 2		Poly2	Sigmoid
POLY2	50, 10, 10, 10, 50		Poly2	PolyHigher
POLY2	25, 4, 4, 25, 50		Sigmoid	Sigmoid
SIGMOID	2, 4, 4, 25, 25		Sigmoid	PolyHgher
SIGMOID	50, 25, 25, 4, 4		PolyHigher	PolyHigher
SIGMOID	4, 4, 4, 10, 10			
SIGMOID	25, 25, 25, 10, 10			
SIGMOID	50, 50, 50, 25, 25			
POLYHIGHER	2, 10, 2, 25, 50			
POLYHIGHER	50, 10, 50, 4, 2			
POLYHIGHER	10, 50, 2, 25, 50			
POLYHIGHER	25, 4, 50, 10, 4			
POLYHIGHER	50, 2, 25, 2, 50			

To comprehensively evaluate the ability of the different methods to detect trend differences of various kinds, all unique combinations of the basic trend categories were used to generate trend differences between the conditions resulting in 21 trend difference combinations (**Table 1)**.

Two variants for each dataset were created to test the methods: *full* and *filtered*. In the full datasets, no filtering of missing values was performed. In the filtered datasets on the other hand, all proteins with missing values were filtered out. In addition to generating datasets where all the spike-in proteins had trend differences from the same combination (e.g. Linear vs. Sigmoid), mix datasets were generated. In the mix datasets 10 randomly selected trend difference combinations were created for the spike-in proteins (**Figure 11**). The semi-simulated mix datasets further assessed the method's abilities to effectively detect differential expression – even when multiple different patterns of longitudinal differential expression were present within a single dataset. The mix datasets were generated using the UPS1 data and its sample groups as a basis. Semi-simulated datasets with a high proportion of missing values to further push and stress test the methods were generated using the CPTAC data. Altogether, 1920 semi-simulated datasets with varying trend and/or expression level differences were generated.

Evaluation metrics

The proportion of proteins from the total number of proteins in a dataset each method was able to provide a valid score, was inspected over the full and the filtered semi-simulated datasets and the experimental *Fn* data (chapter 4.1). The performance of the methods in correctly detecting true longitudinal differential expression was evaluated with a ROC-curve analysis in the semi-simulated spike-in datasets. Partial AUC values (see definitions) from each semi-simulated dataset were recorded for each method.

The reproducibility of the methods was estimated using the experimental *Fn* data. The *Fn* dataset was divided into three replicate datasets according to the three technical replicates of each measurement. Overall reproducibility was estimated as similarity of the technical replicate result lists over all the possible pairwise comparison of strains using the Spearman's correlation coefficient. Reproducibility of the top results was further evaluated as the overlap of the top *k* results in the technical replicate result lists when *k* was varied. Median proportional overlap at each size of *k* over all the pairwise comparisons was considered.

Biological relevance of the findings of the different methods were assessed in the *Fn* dataset against the KEGG [125] Lipopolysaccharide Biosynthesis pathway (ftn00540) and the associated knockout pathway (ko00540) assumed to be affected by the generated modifications in the null mutant strains (D1, D2 and L). The modified asyltransferases in these strains were directly related to the production of lipid A and the Lipopolysacchararide (LPS) in *Fn*. Longitudinal differential expression between the wild type and each of the null mutant strains was examined pairwise from the full *Fn* dataset with no filtering of proteins performed and the technical replicates for a biological replicate averaged. Gene Set Enrichment Analysis (GSEA) [70] was used to explore how the defined pathway proteins were enriched in the result lists of the methods in the different comparisons of wild type and the null mutant strains. Normalized Enrichment Scores (NES) from the GSEA analysis within each comparison for each method were considered as measures of the methods ability to provide biologically meaningful results.

4.6 Knowledge enrichment through integrated functional enrichment and network analysis

4.6.1 Common data processing

The proteins of interest (POI) within **publication IV** and **publication V** were first determined (see the following chapters 4.6.2 and 4.6.3). Following the discovery of the POI, the enrichment of GO [114] terms within the POI was performed as statistical overrepresentation using DAVID [116]. As a background for the enrichment analysis, the whole detected proteome of Th17/Th0/Thp from **publication IV** was used for both studies (**publication IV** and **publication V**). The enriched biological processes (BP) were investigated using the GO FAT terms which filter out the very broad/general GO terms, and comprise only of the more specific lower-level terms [171]. Only statistically significantly overrepresented biological processes with a false discovery rate (FDR) of 0.05 were considered as enriched.

The known and predicted interactions among the DE proteins were queried from the STRING [112] database. All interactions, including both known and predicted interactions, were considered. Only high-quality interactions, with a combined interaction score of >=0.7, were included in the subsequent analysis. The resulting high-quality interaction network was downloaded and imported into Cytoscape for further visualization and processing [172]. Cytoscape is an open source software platform designed for visualizing complex networks and integrating these with any type of attribute data (e.g. annotations, expression data, state data) [172]. Within Cytoscape, different types of analysis, data modification and integration can be performed using the core algorithms included in the software. Furthermore, plenty of plug-ins for Cytoscape are available for more specialized tasks, such as clustering, GO enrichment, pulling STRING interactions, combining annotations, etc. Cytoscape offers flexible and rich ways to visualize data and networks, with many pre-installed and custom layouts and visualization styles [172]. The chosen styles and layouts can easily be further adjusted and modified according to user preferences.

Markov Clustering (MCL) in the Cytoscape plug-in *clusterMaker v2* [173] was used to determine clusters in the protein-protein interaction networks. The edges of the networks, the combined interaction scores assigned by STRING, were used as input for the algorithm. In MCL, a symmetric similarity matrix between the nodes in the network (proteins) is constructed by weighting node interactions (edges) according to their strength [174]. The cluster structure is then discovered by simulating the flow of the graph, and the probability of transitioning from one state (node) to another [174,175]. The probabilities are determined through iterative rounds of matrix multiplication and inflation [174]. The inflation step promotes the differences between areas with strong and weak flows in the network graph and the inflation parameter is used to control for the tightness of the clusters (i.e. granularity) [174,175]. The chosen inflation value highly influences the number of the resulting clusters [175]. The clustering algorithm converges to a partitioned graph, where clusters of high-flow regions are separated from low or no flow regions [175]. As the inflation parameter for the clustering of our PPI network, a value of 1.8 was used, as suggested by [175] for the MCL clustering of high-throughput data.

4.6.2 Quantitative proteomics reveals the dynamic protein landscape during initiation of human Th17 cell polarization

The proteins of interest in **publication IV** were determined as the differentially expressed proteins between Th17 cells and naive activated T cells (Th0) at two time points. Differential expression between the cell types was examined at 24h and 72h after the onset of polarization using ROTS [158,170].

After the common data processing described in the previous chapter, the POI between Th17 and Th0 cell types at 24h and 72h were visualized with their respective LogFCs and most frequent GO BP FAT terms for each cluster. To examine biological functional enrichment in the main cluster in more detail, cluster 1 was further clustered into subclusters using the Cytoscape plug-in ReactomeFIViz [130]. In addition, ReactomeFIViz was used to explore pathway enrichment in the detected subclusters. ReactomeFIViz integrates information from several pathway databases [130], such as Reactome [126], NCI PID [128], Panther [118] and Biocarta [127].

4.6.3 Protein interactome of the Cancerous Inhibitor of Protein phosphatase 2A (CIP2A) in Th17 cells

The proteins of interest in **publication V** were defined as the CIP2A interacting proteins determined by the Significance Analysis of INTeractome (SAINT) algorithm [176]. SAINT is a software tool assigning confidence scores for the detected interactions in affinity-purification mass spectrometry (AP-MS) data such as the data in **publication V** (chapter 4.1). Only interactions with a SAINT probability score of ≥ 0.95 coming from a distribution of true interactions were included in the further analysis. Furthermore, of the interacting proteins with a SAINT score ≥ 0.95, only proteins present in <60% of the datasets listed in the Contaminant Repository for Affinity Purification-mass spectrometry data (CRAPome) [177] were allowed to enter the subsequent analysis. The CRAPome lists common contaminants aggregated from multiple AP-MS studies.

Following the common data preprocessing (chapter 4.6.1), the CIP2A interacting proteins together with the LogFCs between the CIP2A immuno-precipitates and the IgG controls and the most frequent significantly enriched GO BP FAT term for each cluster, were visualized.

5 Results

5.1 Normalization in proteomics

5.1.1 Intragroup variation

Normalization in general decreased intragroup variation in **publication I**. Intragroup variation between technical replicates in the spike-in datasets, as well as between biological replicates in the experimental data, was lower in data after normalization by all the methods when compared to the non-normalized $\log_2$ transformed data (**Figure 12**). Of the explored normalization methods, the Vsn normalization was found to decrease intragroup variation the most (**Figure 12**). In addition to Vsn, also the EigenMS normalization was found to decrease intragroup variation more than the other examined normalized methods in the experimental mouse data (**Figure 12D**).

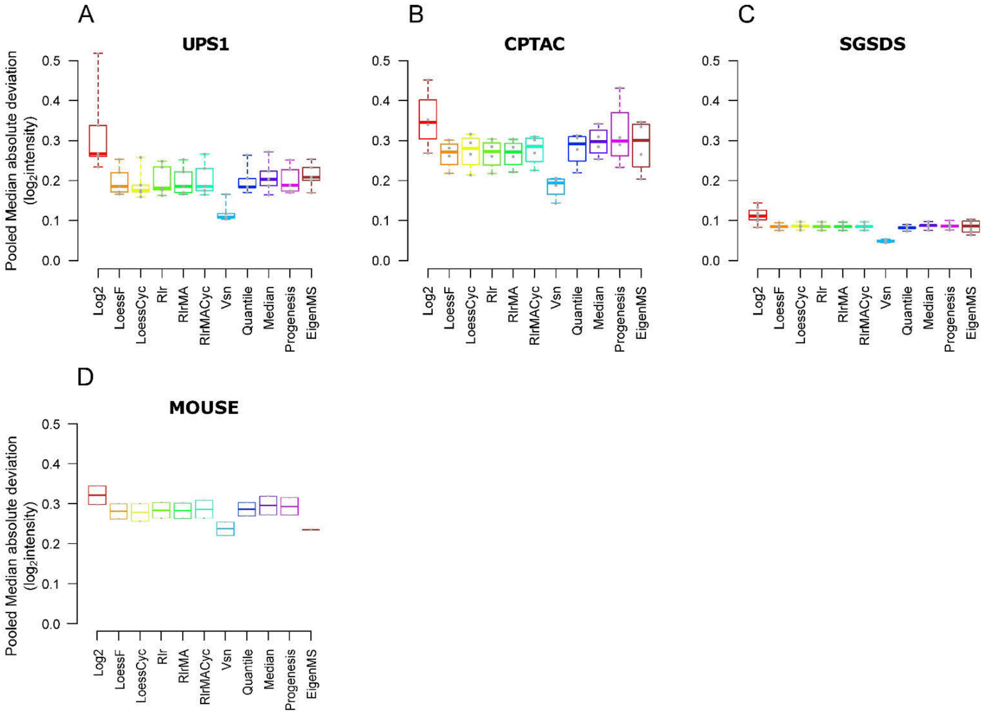


Figure 12. The pooled median absolute deviations (PMAD) between the technical replicates in data normalized with the different methods in the A) UPS1, B) CPTAC, C) SGSDS spike-in datasets and D) between the biological replicates in the experimental mouse dataset. Adopted with permission from **Publication I**: Figures 1 and 5.

5.1.2 Performance in the differential expression analysis

Overall, normalizing the data improved the consistency of correctly detecting the truly DE proteins (**Figure 13**). While no single normalization method was able to perform best in all examined comparisons and datasets, Vsn produced consistently good results. In many of the examined comparisons, the AUCs from Vsn normalized data were significantly better than the AUCs from data normalized with the other methods (**Figure 13**). However, several other normalization methods were also able to provide data from which the true DE proteins could be detected consistently and the AUC differences to Vsn were small. In addition to Vsn, LoessF, Rlr, RlrMA, Progenesis and Median normalization consistently performed well in the DE analysis. The Quantile normalization method performed mostly well in the examined comparisons but had clearly worse AUC values compared to the other methods in some comparisons. Interestingly, out of the cyclic methods, RlrMACyc performed rather consistently between the comparisons, while LoessCyc performed excellent in some comparisons but poorly on others (**Figure 13**).

Surprisingly, while EigenMS was effective in reducing intragroup variation, it produced AUC values in the differential expression comparable to the non-normalized data. This is interesting, since typically the normalization methods are evaluated in their ability to decrease intragroup variation [29,36,37] while the interest of the experiments are in the DE proteins. Based on these results, considering only the ability of a normalization method to reduce unwanted variation might not be sufficient.

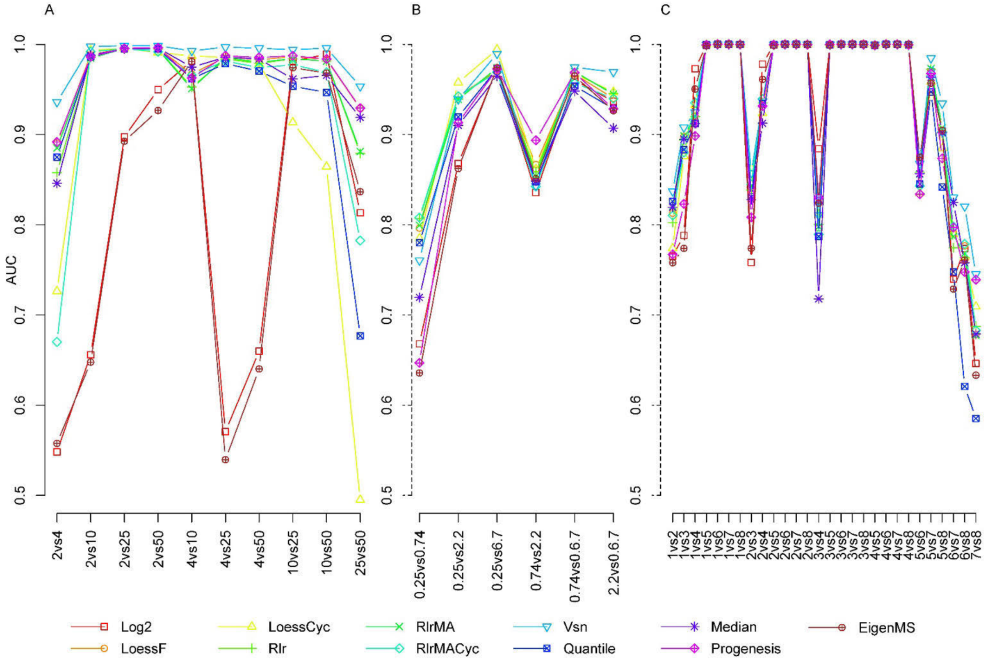


Figure 13. The areas under the ROC-curves (AUC) in all the pairwise comparisons of sample groups in the A) UPS1, B) CPTAC and C) SGSDS spike-in datasets. For the ROC-curve analysis, the spike-in proteins have been considered as true positives and the background proteins as true negatives. For the UPS1 and CPTAC datasets, the sample group numbers refer to the concentrations of the spike-in proteins (fmol) in the sample group. For the SGSDS data, the numbers refer to sample group identifiers. Adopted with permission from **Publication I**: Figure 2.

5.1.3 Logarithmic fold change

The LogFCs of the stable background proteins were more concentrated around 0 in Vsn-normalized data than in data normalized with the other methods (**Figure 14**). Furthermore, the logFCs of the spike-in proteins were systematically underestimated in the Vsn-normalized data when the spike-in proteins in either of the evaluated samples were spiked at low concentrations in the original data (**Figure 14C**).

However, when the concentration of the spike-in proteins was higher in both samples, the logFCs calculated from Vsn-normalized data coincided with logFCs calculated from data normalized with the other methods (**Figure 14B**). More generally, the logFCs of the spike-in proteins were often underestimated in all normalized and non-normalized datasets compared to expected theoretical logFCs based on the concentrations of the spike-in proteins (**Figure 14C**).

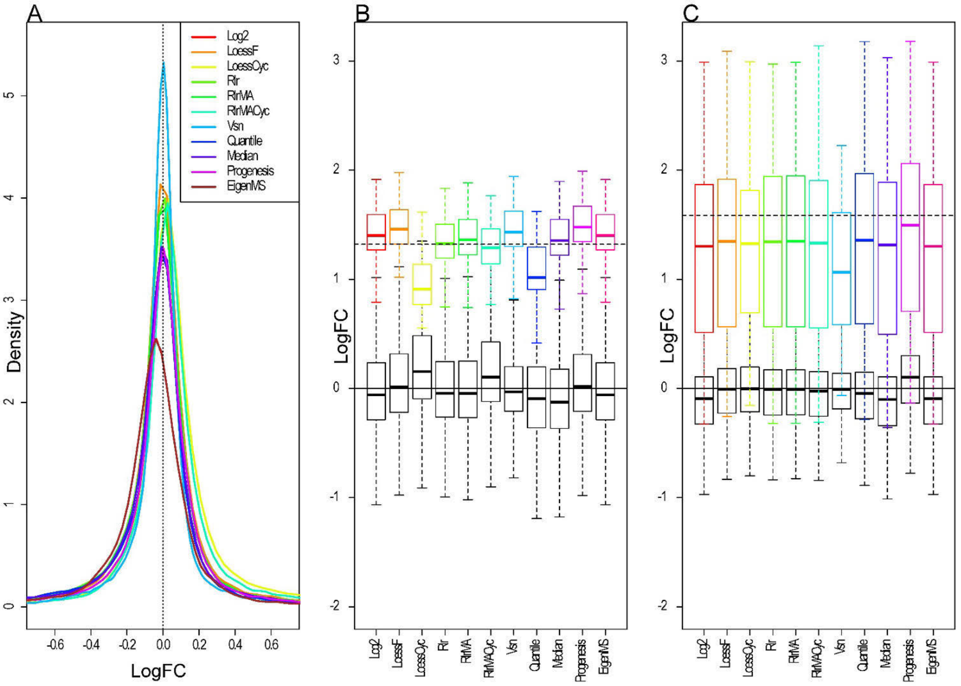


Figure 14. A) The densities of logarithmic fold changes (LogFC) of the background proteins over the UPS1, CPTAC and SGSDS spike-in datasets in data normalized with the different methods. The vertical dashed line represents the theoretical expected LogFC of the background proteins (0). B) The LogFCs of the proteins in the 10 fmol vs. 25 fmol sample group comparison from the UPS1 data normalized with the different methods. C) The LogFCs of the spike-in proteins in the 0.74 fmol vs. 2.2 fmol sample group comparison from the CPTAC data normalized with the different methods. In B and C, the colored boxplots represent the observed LogFCs of the spike-in proteins while the black boxes represent the LogFCs of the background proteins from data normalized with the different methods. The dashed horizontal line represents the expected theoretical LogFC between the spike-in proteins while the horizontal solid black line represents the theoretical expected LogFC for the background proteins. Adopted with permission from **Publication I**: Figure 4.

The arsinh function based parametric transformations performed by Vsn return the data on a scale that coincide with $\log_2$ transformed data on large intensity values but are typically larger on low intensities [34] (**Figure 15**). Such different scaling typically results in smaller logFC ratios. Therefore, even though differential expression between samples at low intensities might be better detected due to decreased variance, the logFC values at low intensities are systematically underestimated by Vsn.

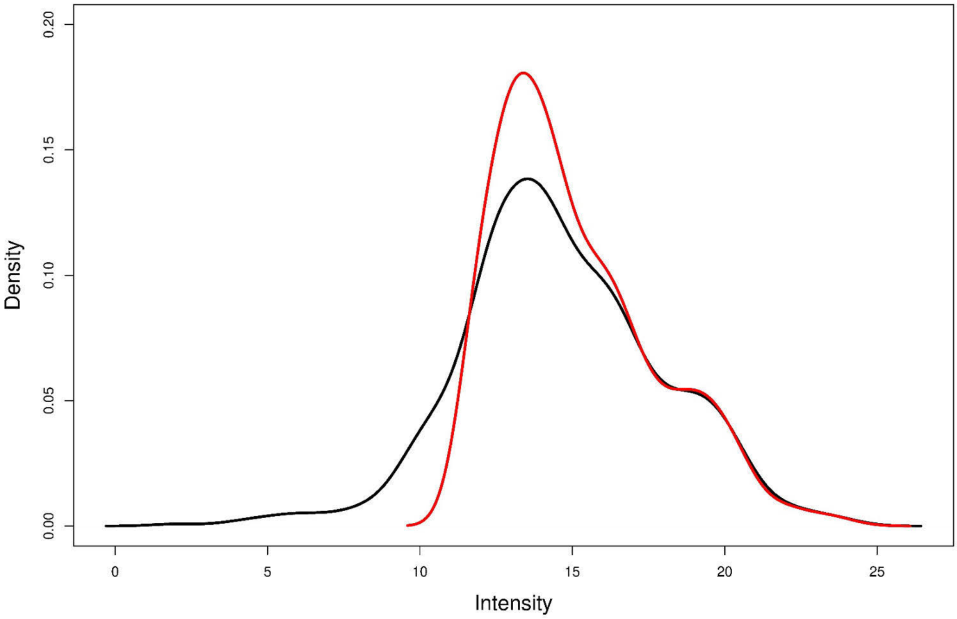


Figure 15. The density curves of intensity values over all the proteins in the 0.25 fmol sample in the CPTAC dataset. The black curve corresponds to $\log_2$-transformed data and the red curve to vsn-normalized data.

5.1.4 Effect of normalization stage

Apart from the cyclic methods, the stage in which the normalization was performed did not have a large effect in **publication I**. When the normalization was performed globally, simultaneously for all the samples in the data, the cyclic methods were not able to center the data on the x axis in the MA plots (**Figure 16**). When only the examined samples were normalized pairwise, the cyclic methods performed similarly to the non-cyclic methods (**Figure 16**).

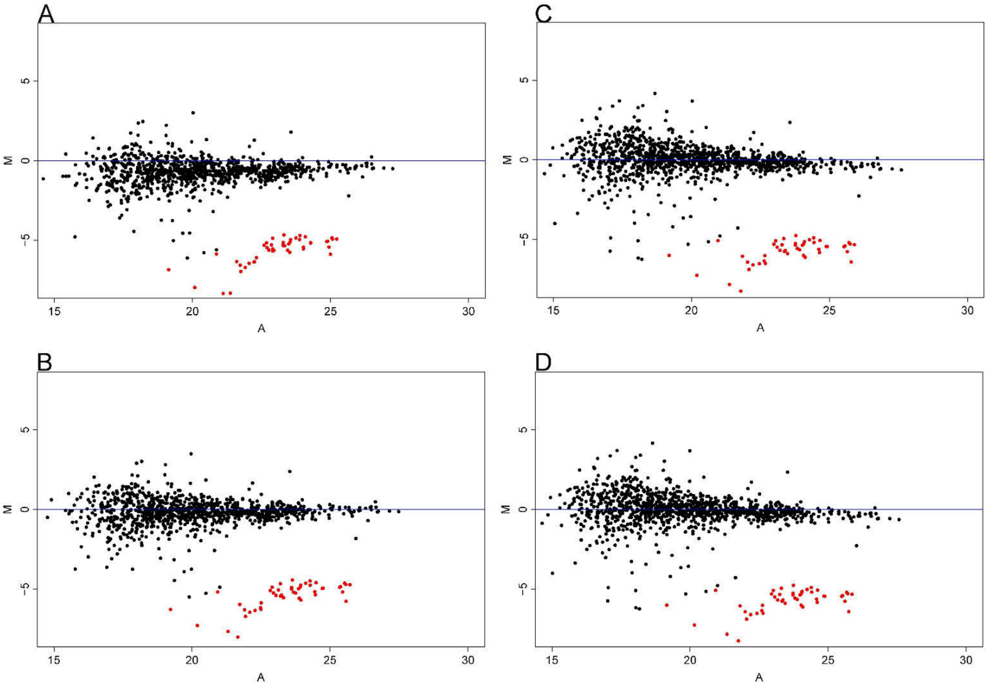


Figure 16. MA-Plots of the 2 fmol and 50 fmol samples in the A) globally cyclic loess normalized UPS1-data, B) pairwise normalized cyclic loess data, C) globally vsn normalized data and D) pairwise vsn normalized data. The non-changing background proteins are colored black while the truly differentially expressed spike-in proteins are colored as red.

5.1.5 An additional case study into normalization

In the Th17 proteome of **publication IV,** all examined normalization approaches aligned the sample distributions when compared to non-normalized data (**Figure 17A**). The effects of normalization were explored using best approaches from **publication I** together with the MaxQuant [42] innate normalization method MaxLFQ [18] to determine the most suitable normalization for this dataset with considerable changes in protein expression between the different timepoints.

The standard deviation was more constant along the whole intensity range in MaxLFQ normalized data when compared to IBAQ data normalized with the other methods or non-normalized IBAQ data (**Figure 17B**). Normalization of the IBAQ data with other methods than MaxLFQ seemed to have only a very slight effect on the mean-to-variance relationship when compared to non-normalized data. As discussed in chapter 4.2, a constant mean-to-variance relationship along the whole intensity range is a desirable quality in the data.

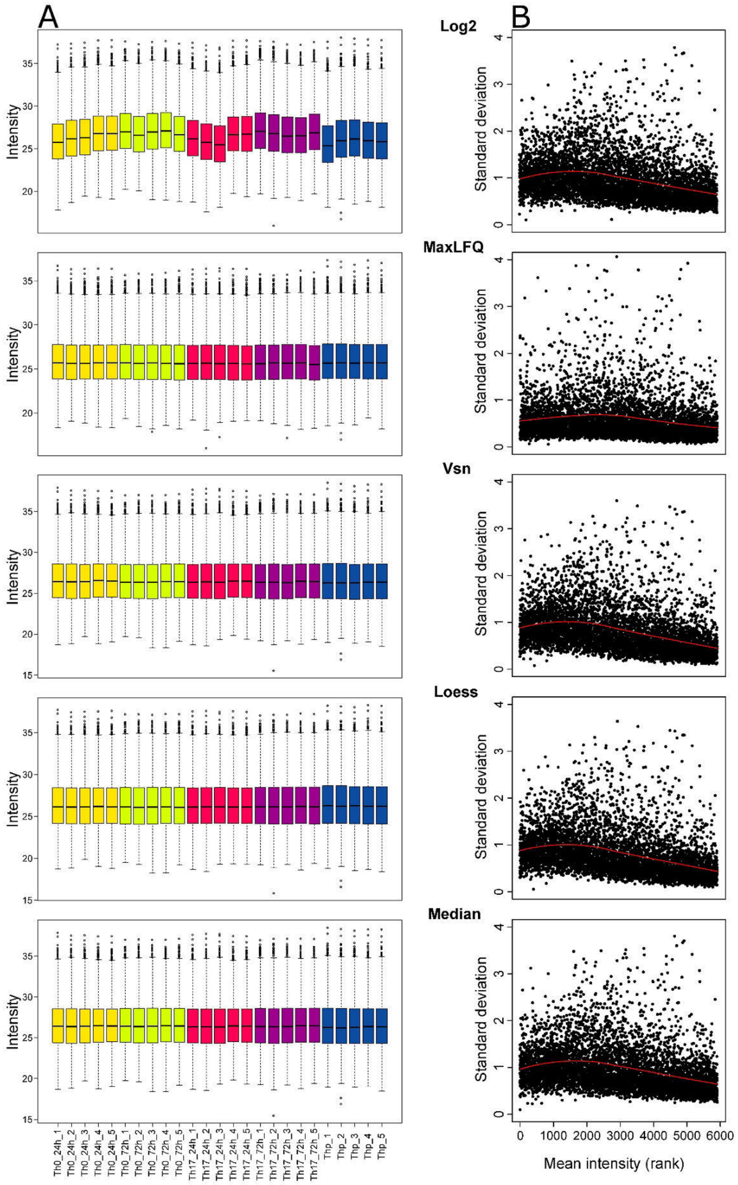


Figure 17. A) Distributions of samples, and B) mean intensity vs. standard deviation in the unnormalized log2-transformed data from **publication IV** and the same data normalized with different methods. For A, the samples are colored according to biological sample groups. For B, the proteins in each dataset are ordered ascending based on ranks of mean intensity of the proteins. The red line in B corresponds to a loess fit between standard deviation and mean intensity. Technical replicates of the data have been averaged.

Variance between technical replicates was decreased in data normalized by all the methods when compared to non-normalized data (**Figure 18A**). However, intragroup variation in LFQ data was considerably lower than in IBAQ data normalized with the other methods (**Figure 18A**). Similarly, variation between biological replicates was decreased in normalized data (**Figure 18B**). Again, intragroup variation between biological replicates was smallest in the MaxLFQ data followed by Vsn-, and Loess-normalized datas (**Figure 18B**).

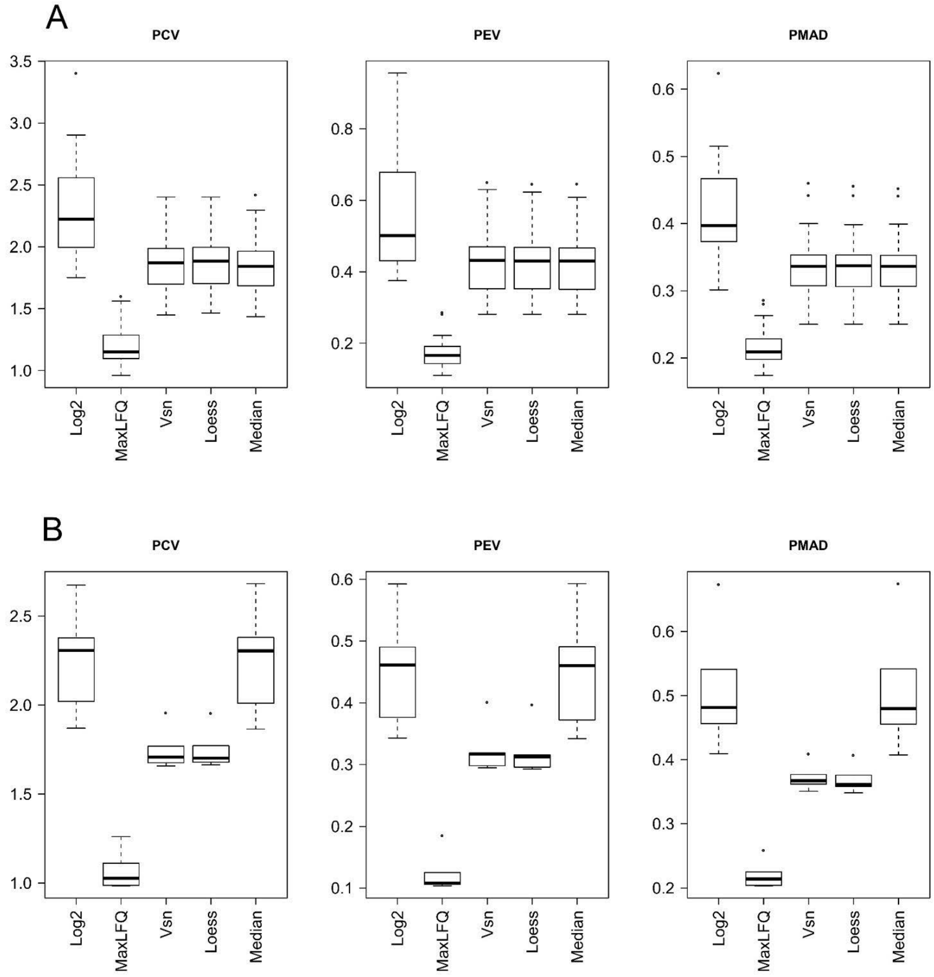


Figure 18. Intragroup variation between A) the technical replicates, and B) the biological replicates in the unnormalized log2-transformed data from **publication IV** and the same data normalized with different methods.PCV stands for Pooled Coefficient of Variation, PEV=Pooled estimate of Variance and PMAD=Pooled Median Absolute Deviation.

Furthermore, samples in the MaxLFQ data were overall more higly correlated with each other and clustered better according to the biological sample groups than in data normalized with the other methods or in the un-normalized data (**Figure 19**). In addition to being normalized differently, the MaxLFQ data was also quantified differently and thus contained a different proportion of missing values. However, no large differences in missing values or missing value patterns were observed between the two quantitation approaches. The proportion of missing values in the averaged IBAQ data was 4.4% and 6% in the MaxLFQ data.

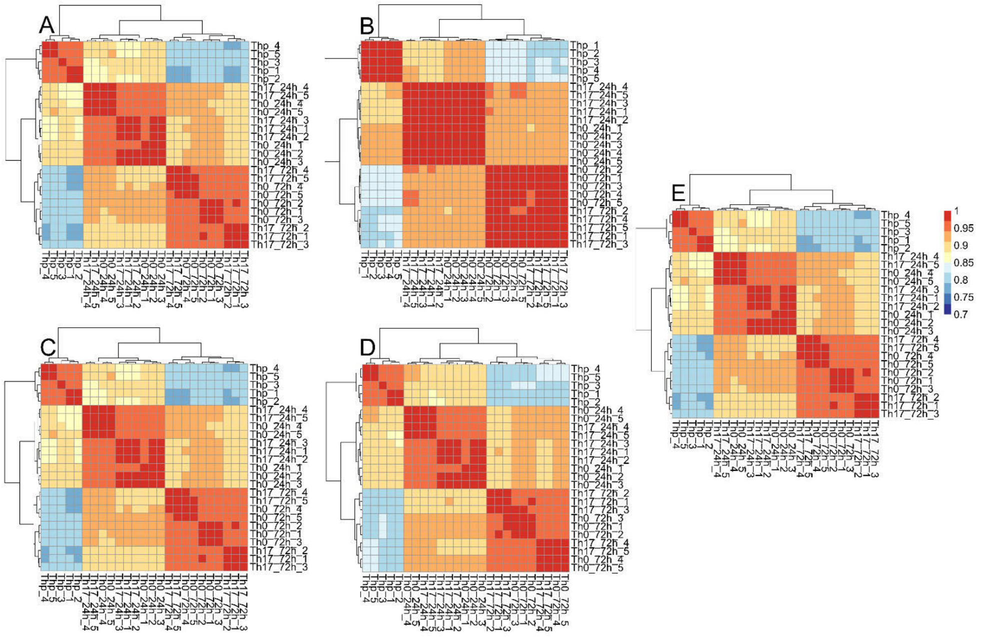


Figure 19. Pearson correlations coefficients between the samples in the averaged A) unnormalized log2-transformed, B) MaxLFQ normalized, C) vsn normalized, D) loess normalized and E) median normalized data from **publication IV**. The correlation data has been clustered using hierarchical clustering with Euclidean distances and the complete linkage method.

To conclude, while the previously observed well performing methods Vsn and Loess (**publication I**), were also observed to perform well in normalizing the Th17 proteome data, exploration of the data normalized with the different chosen approaches suggested MaxLFQ as the most suitable method for normalizing the data in **publication IV**.

5.2 Label-free proteomics data processing software tools

5.2.1 Protein identification and quantification

In **publication II**, different software workflows identified and quantified a varying number of proteins in each dataset (**Table 2**). Progenesis clearly quantified the lowest number of proteins in all datasets. However, in practice, there were no missing values in Progenesis data while all the other software produced varying proportions of missing values for each dataset.

Table 2. The number of quantified proteins, number of detected spike-in proteins and the proportion of missing values produced by the different label-free software workflows in the CPTAC and SGSDS datasets. Adopted with permission from **Publication II**: Table 1.

	PROGENESIS		**MAXQUANT**		**PROTEIOS**		**PEAKS**		**OPENMS**	
	CPTAC	SGSDS	CPTAC	SGSDS	CPTAC	SGSDS	CPTAC	SGSDS	CPTAC	SGSDS
PROTEINS QUANTIFIED	614	2168	1247	3487	1383	3554	1223	3161	1276	1401
SPIKE-IN PROTEINS DETECTED	32 (67%)	12 (100%)	41 (85%)	12 (100%)	42 (88%)	12 (100%)	42 (88%)	12 (100%)	41 (85%)	11 (92%)
PROPORTION OF MISSING VALUS IN THE DETECTED SPIKE-IN PROTEINS	0,0%	0,0%	29,4%	4,2%	20,8%	0,0%	21,7%	0,3%	23,6%	2,3%
PROPORTION OF MISSING VALUES IN THE DETECTED BACKGROUND PROTEINS	1,2%	0,0%	19,1%	3,4%	18,0%	6,4%	19,5%	3,2%	32,1%	8,8%
TOTAL PROPORTION OF MISSING VALUES	1,0%	0,0%	19,0%	3,0%	18,0%	6,0%	20,0%	3,0%	32,0%	9,0%

Most of the protein identifications between software workflows were shared with almost all of the proteins identified in the Progenesis workflow, identified also by other software workflows (**Figure 20**).

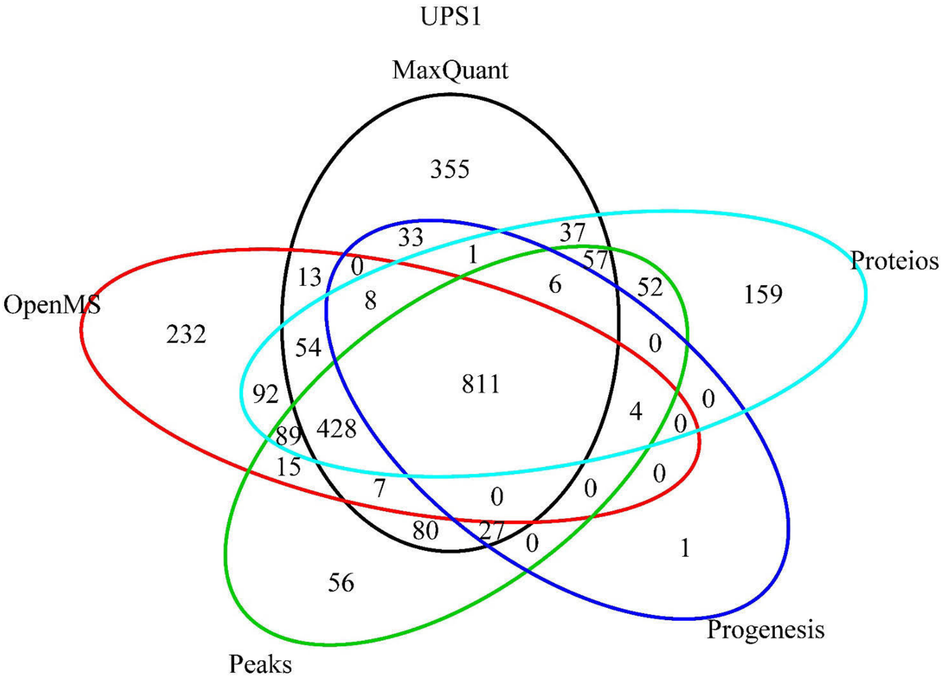


Figure 20. The number of shared and distinct protein identifications by each software workflow in the UPS1 dataset. Adopted with permission from **Publication II**: Figure 1.

5.2.2 Performance in the differential expression analysis

Missing value proportion in the data was the main determining factor differentiating the performance of the evaluated software in the differential expression analysis in **publication II**. In the presence of missing values, Progenesis clearly outperformed the other software (**Figure 21A, Figure21B, Table2**). The performance of MaxQuant on the other hand was most prominently hindered by the missing values in the CPTAC and UPS1B datasets. Proteios, Peaks and OpenMS all performed better in producing data from which the true DE proteins could be correctly detected in the differential exression analysis than MaxQuant. However, when no large proportion of missing values were present in the data, all software workflows performed well (**Figure 21C, Figure 21D**).

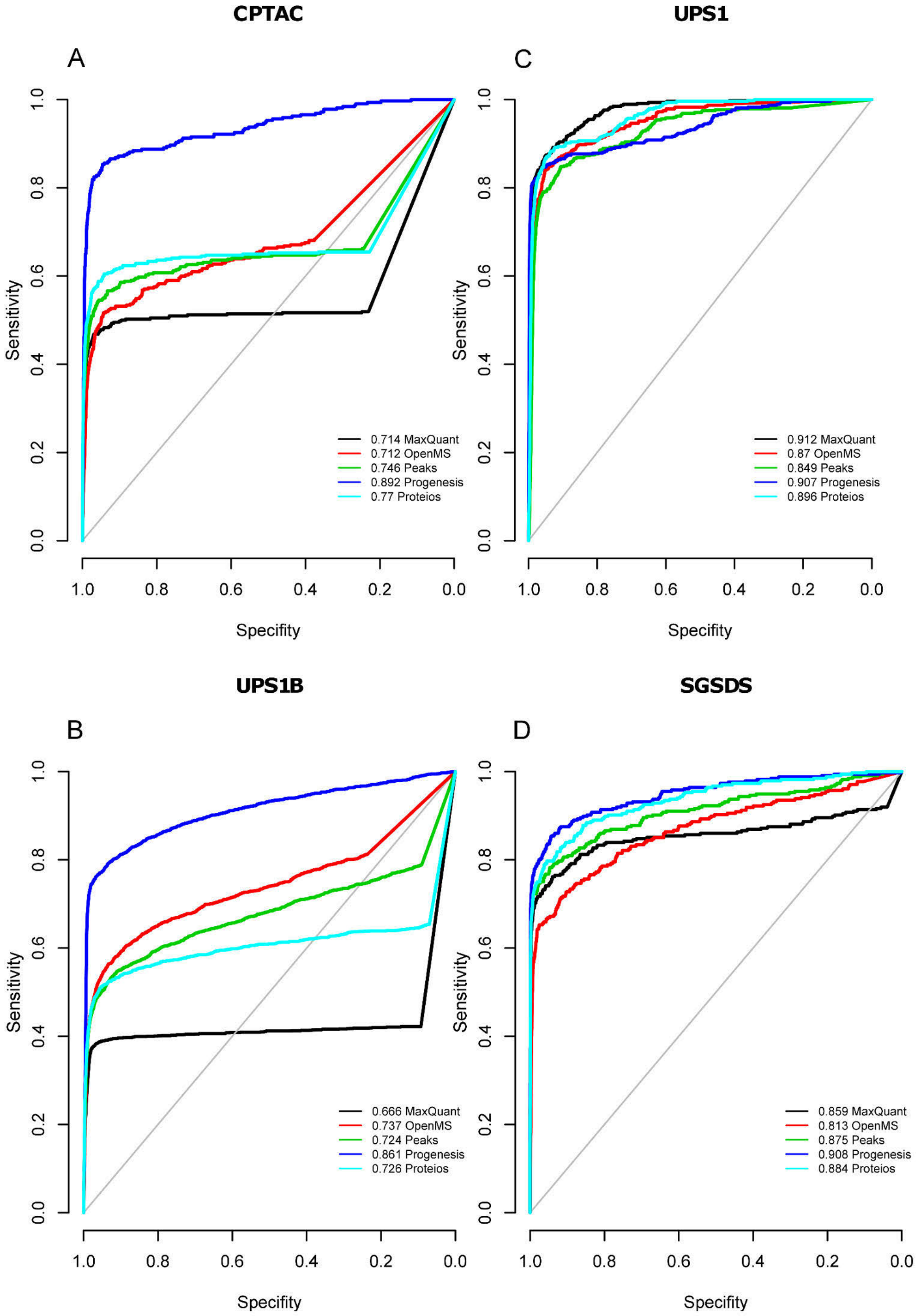


Figure 21. ROC-curves over all the pairwise comparisons of sample groups for the different software in the A) CPTAC data, B) UPS1B data, C) UPS1 data and D) SGSDS data. Adopted with permission from **Publication II**: Figure 2.

5.2.3 Evaluation of logarithmic fold changes by the software workflows

The LogFCs of the spike-in proteins were estimated most accurately in MaxQuant data (**Figure 22**). Most variation was detected in the estimates of OpenMS and Progenesis. The LogFCs of the background proteins were estimated accurately by all the software with Progenesis having slightly more accurate estimates than the other software.

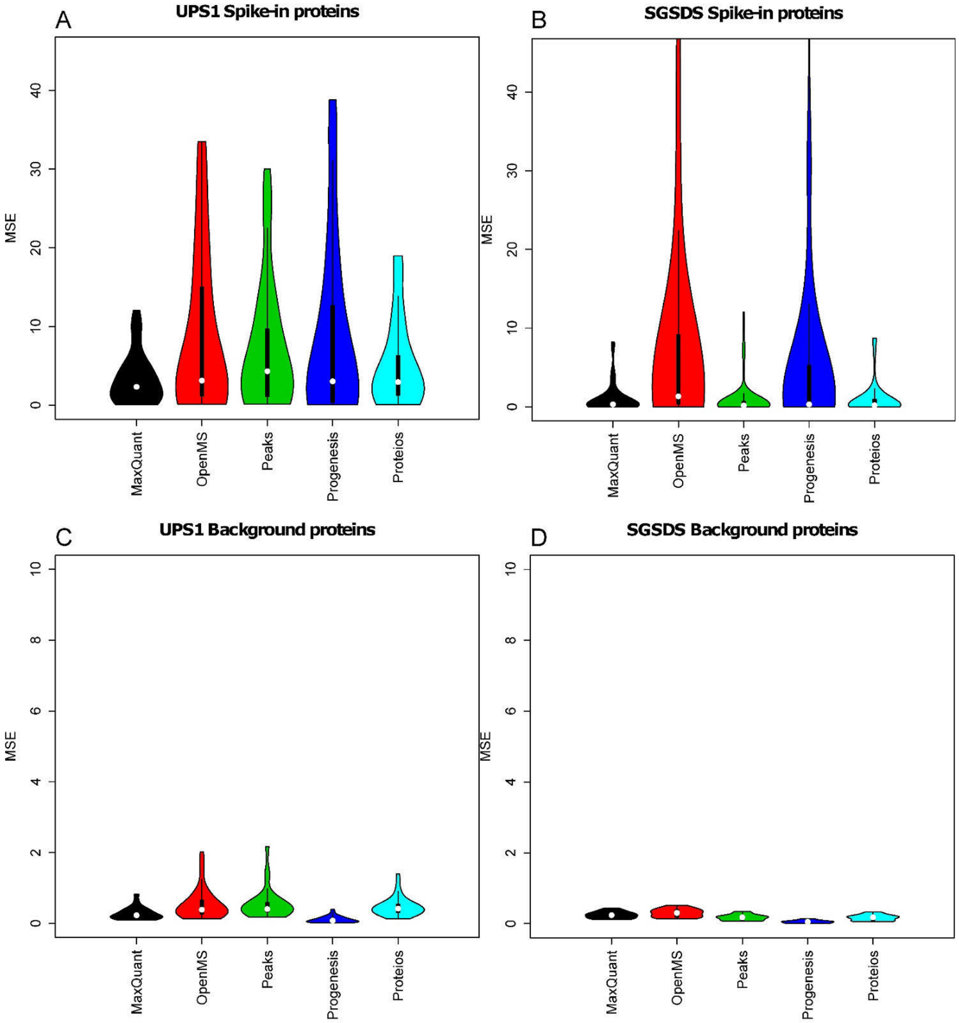


Figure 22. Mean squared erros (MSE) of the observed logarithmic fold changes (LogFC) from the theoretical expected LogFCs over all the pairwise comparisons of the A) spike-in proteins in the UPS1 data, B) spike-in proteins in the SGSDS data, C) background proteins in the UPS1 data and D) background proteins in the SGSDS data. Adopted with permission from **Publication II**: Figure 4.

5.3 Missing values and imputation in label-free proteomics

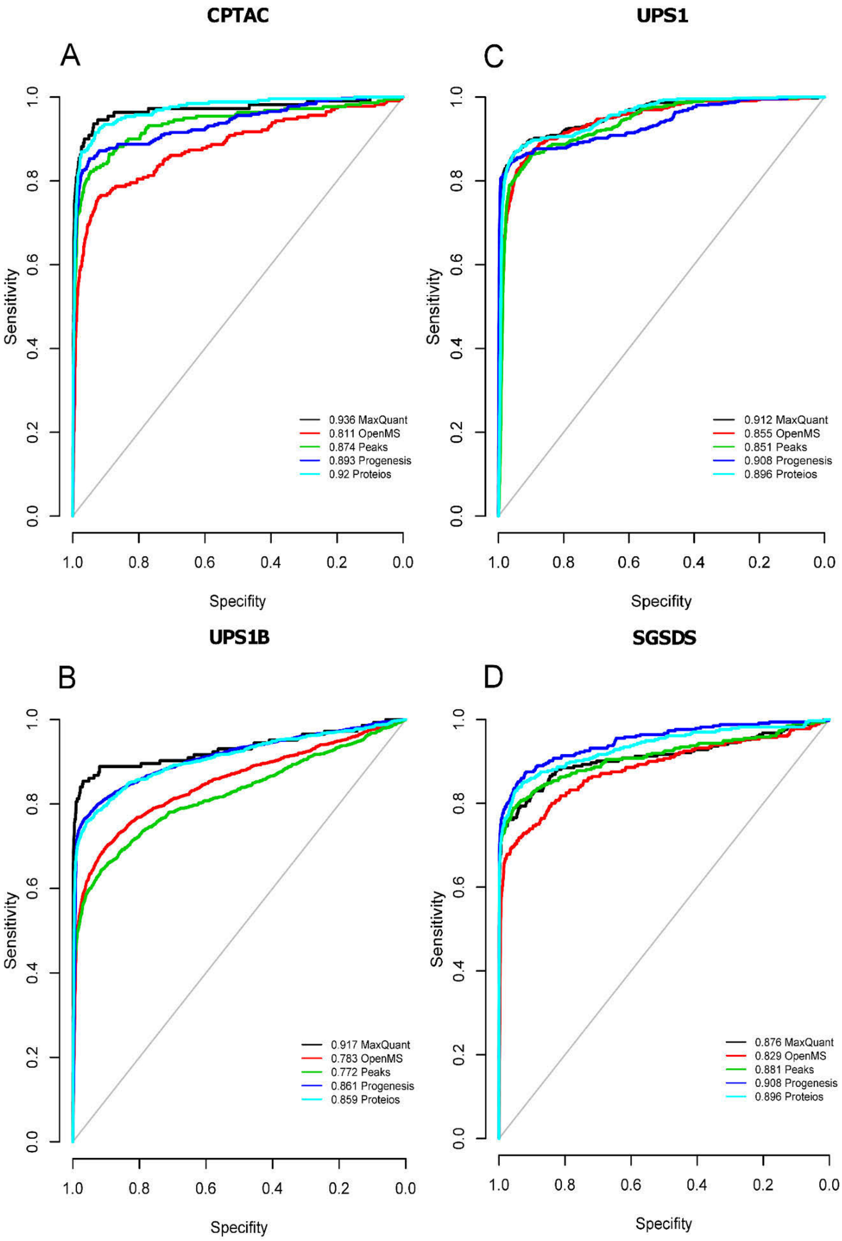


Figure 23. ROC-curves over all the pairwise comparisons of sample groups after the best performing filtlls (filtering + local least squares) imputation for the different software in the A) CPTAC data, B) UPS1B data, C) UPS1 data and D) SGSDS data. Adopted with permission from **Publication II**: Figure 5.

In general, filtering and/or imputation improved the correct detection of the spike-in proteins (TP) and the background proteins (TN), when the proportion of missing values in the datasets was high (**Figure 21A, Figure 21B, Figure 23A, Figure 23B**). Interestingly, the proportion of missing values in the background proteins did not seem to be so relevant for the performance effect of the imputation approaches (**Table 3**). Moreover, in the presence of a low proportion of missing values in the spike-in proteins (e.g. the UPS1 dataset), the application of imputation or filtering had a detoriating effect on performance more often (**Table 3**).

Table 3. Differences in pAUCs from the ROC-curves drawn over all the pairwise comparisons in data imputed with the different methods in all the evaluated software compared to corresponding unimputed data. Green color of the cell indicates improvement in pAUC after imputation while red colod indicates a decrease in pAUC after imputation. Mean ranks for each imputation method is calculated as the mean over all the ranks within each dataset and and each software.

Software	Dataset	back	bpca	censor	filtered	filtlls	knn	lls	svd	zero	Proportion of missing valus in the detected spike-in proteins	Proportion of missing values in the detected background proteins
MaxQuant	CPTAC	0.137	0.08	0.122	0.213	0.222	0.097	0.16	0.044	0.003	29.40 %	19.10 %
	SGSDS	0.029	0.028	0.014	0.014	0.017	0.027	0.036	0.031	-0.027	4.20 %	3.40 %
	UPS1	-0.191	0.002	-0.267	-0.003	0	-0.1	-0.01	-0.127	-0.306	0.00 %	14.50 %
	UPS1B	0.129	0.137	0.11	0.249	0.251	0.17	0.203	0.177	0.099	35.20 %	7.40 %
OpenMS	CPTAC	0.029	0.038	-0.026	0.089	0.099	0.022	0.072	0.001	-0.077	23.60 %	32.10 %
	SGSDS	-0.031	0.011	-0.046	0.014	0.016	0	0.002	-0.002	-0.068	2.30 %	8.80 %
	UPS1	-0.312	0.011	-0.34	-0.013	-0.015	-0.115	-0.042	-0.172	-0.287	1.00 %	29.80 %
	UPS1B	-0.049	0.027	-0.105	0.041	0.046	0.007	0.029	-0.011	-0.082	13.00 %	20.00 %
Peaks	CPTAC	-0.001	0.056	-0.05	0.133	0.128	0.065	0.094	0.071	-0.043	21.70 %	19.50 %
	SGSDS	-0.003	0.001	-0.016	0.003	0.006	0.004	0.01	0.006	-0.024	0.30 %	3.20 %
	UPS1	-0.221	-0.004	-0.295	0.001	0.002	-0.09	-0.041	-0.168	-0.284	1.20 %	16.40 %
	UPS1B	-0.019	0.045	-0.04	0.04	0.048	0.065	0.068	0.052	-0.057	21.40 %	5.90 %
Progenesis	CPTAC	0.001	0.002	0	0	0.001	0.001	0.002	0.001	-0.014	0.00 %	1.20 %
	SGSDS	0	0	0	0	0	0	0	0	0.003	0.00 %	0.00 %
	UPS1	0.001	0.002	0.001	0.001	0.001	0	0.001	0.001	-0.009	0.00 %	0.40 %
	UPS1B	-0.001	-0.001	-0.001	0	0	-0.001	-0.001	0	-0.005	0.00 %	0.10 %
Proteios	CPTAC	0.027	0.064	0.044	0.113	0.15	0.054	0.131	0.043	-0.106	20.80 %	18.00 %
	SGSDS	-0.014	-0.003	-0.043	0.002	0.012	0.006	-0.009	-0.012	-0.032	0.00 %	6.40 %
	UPS1	-0.283	-0.017	-0.312	-0.003	0	-0.21	-0.024	-0.143	-0.379	0.10 %	19.00 %
	UPS1B	0.05	0.095	0.029	0.133	0.133	0.093	0.119	0.043	0.009	19.90 %	8.10 %
Imputation Method mean ranks		**5.75±0.45**	**3.65±0.41**	**7.1±0.49**	**2.85±0.42**	**1.85±0.29**	**4.65±0.33**	**2.85±0.27**	**4.65±0.48**	**8.3±0.4**		

Filtering combined with the lls imputation was the most effective approach, increasing the performance of the software in the differential expression analysis the most (**Table 3**). Similarly, the pure filtering approach as well as the lls, bpca and knn imputations mostly improved the performance of the software in the differential expression analysis. Out of the pure imputations methods, the lls imputation method resulted in highest performance gains on average, while the bpca imputation improved the performance of the different software in the differential expression analysis most often (**Table 3**). Furthermore, the applicability of the different filtering and imputation methods was software dependent. While MaxQuant with the largest proportions of missing values in the spike-in proteins, benefitted most from imputation or filtering, Progenesis with no missing values remained largely unaffected. For the other software than MaxQuant, the effect of the simple imputations methods (zero, back, censored) was mainly deteriorating and resulted in reduced performance. MaxQuant benefitted from the back and censored imputations in all but the UPS1 dataset.

5.4 Longitudinal differential expression in proteomics

5.4.1 Performance in the differential expression analysis

Overall, in **publication III** the new proposed method RolDE performed best in detecting the truly longitudinally DE spike-in proteins from the semi-simulated spike-in label-free proteomic datasets (**Figure 24**). This was especially true in the presence of missing values in the data (**Figure 24B-D**). The specialized Bayesian longitudinal method Timecourse and all variants of Limma also displayed good overall performance. The higher order regression models outperformed the lower order models.

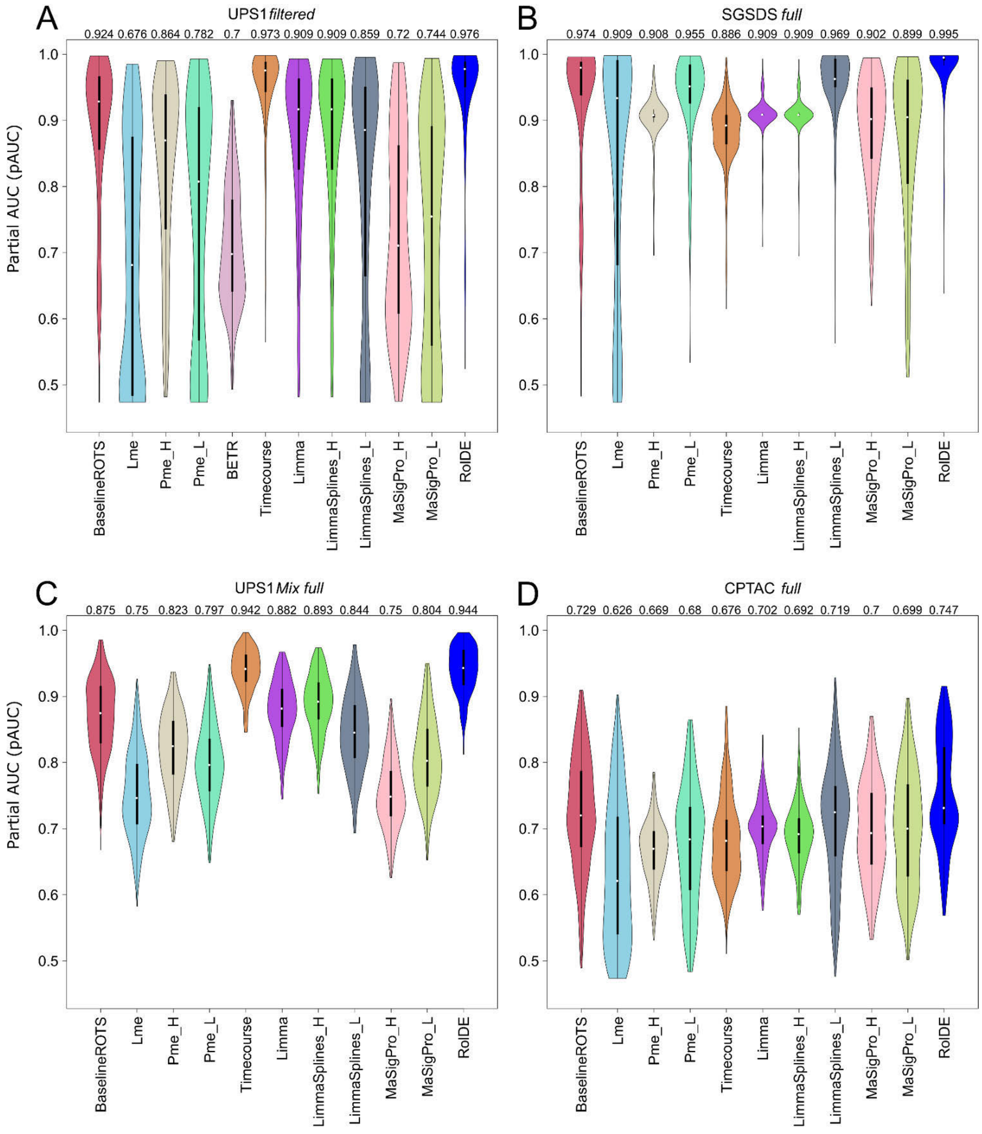


Figure 24. The partial AUCs (pAUC) of the ROC-curve analysis of longitudinal differential expression over all the semi-simulated datasets with varying longitudinal trend differences between two conditions in A) the UPS1 filtered datasets, B) SGSDS full datasets, C) UPS1 mix full datasets and D) CPTAC full datasets. In filtered datasets, all proteins with any missing values have been filtered out. In full datasets, no filtering of the datasets has been performed. 300 semi-simulated datasets have been tested for A, C and D. 210 semi-simulated datasets have been tested for B. Adopted with permission from **Publication III**: Figure 2.

Furthermore, the performance of RolDE was most balanced over the different categories, offering consistent performance in detecting trend differences of various types (**Figure 25**). In general, the polynomial complexity of the detected trend differences was concordant with the degree of the regression, as can be expected. Of the evaluated regression approaches, the full polynomial models (Pme_H) was able to detect the broadest spectrum of trend differences (**Figure 25**).

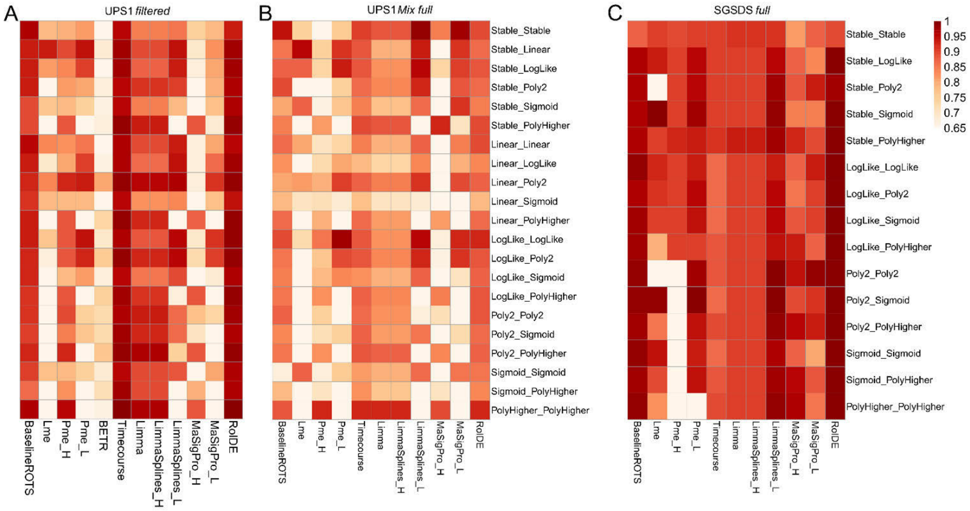


Figure 25. Interquartile (IQR) mean partial AUCs (pAUCS) of the ROC-curve analysis of longitudinal differential expression in all trend difference categories in the A) UPS1 filtered, B) UPS1 mix full and C) SGSDS full semi-simulated datasets. In filtered datasets, all proteins with any missing values have been filtered out. In full datasets, no filtering of the datasets has been performed. 300 semi-simulated datasets altogether have been tested for A, B. 210 semi-simulated datasets have been tested for C. Adopted with permission from **Publication III**: Supplementary Figure 2.

5.4.2 Ability to provide a valid ranking

Different methods were able to provide a ranking for a different proportion of proteins in the data (**Table 4**). In general, the linear regression based approach, Lme, together with RolDE were able to consistently deliver a valid ranking/score for the majority of proteins in each dataset (IQR mean value <10%). When missing values were present, the proportion of proteins Timecourse was able to provide a ranking for decreased markedly and BETR does not tolerate missing values.

Table 4. Interquartile (IQR) mean proportions of proteins each method was not able to determine a ranking / score for in the different dataset types. Adopted with permission from **Publication III**: Table 1.

	UPS1 FILTERED	SGSDS FILTERED	UPS1 MIX FILTERED	UPS1 FULL	SGSDS FULL	CPTAC FULL	UPS1 MIX FULL
BASELINEROTS	0.0 %	0.0 %	0.0 %	12.4 %	1.8 %	16.3 %	9.9 %
LME	0.0 %	0.0 %	0.0 %	7.7 %	0.5 %	7.4 %	7.2 %
PME_H	0.0 %	0.0 %	0.0 %	19.3 %	3.2 %	19.5 %	19.1 %
PME_L	0.0 %	0.0 %	0.0 %	11.9 %	1.2 %	11.4 %	11.2 %
BETR	0.0 %	0.0 %	0.0 %				
TIMECOURSE	0.0 %	0.0 %	0.0 %	23.4 %	6.9 %	27.8 %	24.4 %
LIMMA	0.0 %	0.0 %	0.0 %	19.3 %	3.2 %	19.2 %	19.0 %
LIMMASPLINES_H	0.0 %	0.0 %	0.0 %	19.3 %	3.2 %	19.2 %	19.0 %
LIMMASPLINES_L	0.0 %	0.0 %	0.0 %	11.9 %	1.2 %	11.1 %	11.1 %
MASIGPRO_H	20.3 %	22.8 %	23.2 %	30.6 %	24.1 %	40.6 %	32.9 %
MASIGPRO_L	33.6 %	35.7 %	46.0 %	41.1 %	36.5 %	52.2 %	51.8 %
ROLDE	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %

5.4.3 Reproducibility

Overall reproducibility was high and comparable in all versions of Limma (Limma, LimmaSplines_H, LimmaSplines_L) and RolDE (**Figure 26A**). The reproducibility of the top results was similarly high in RolDE and different variants of Limma (**Figure 26B**).

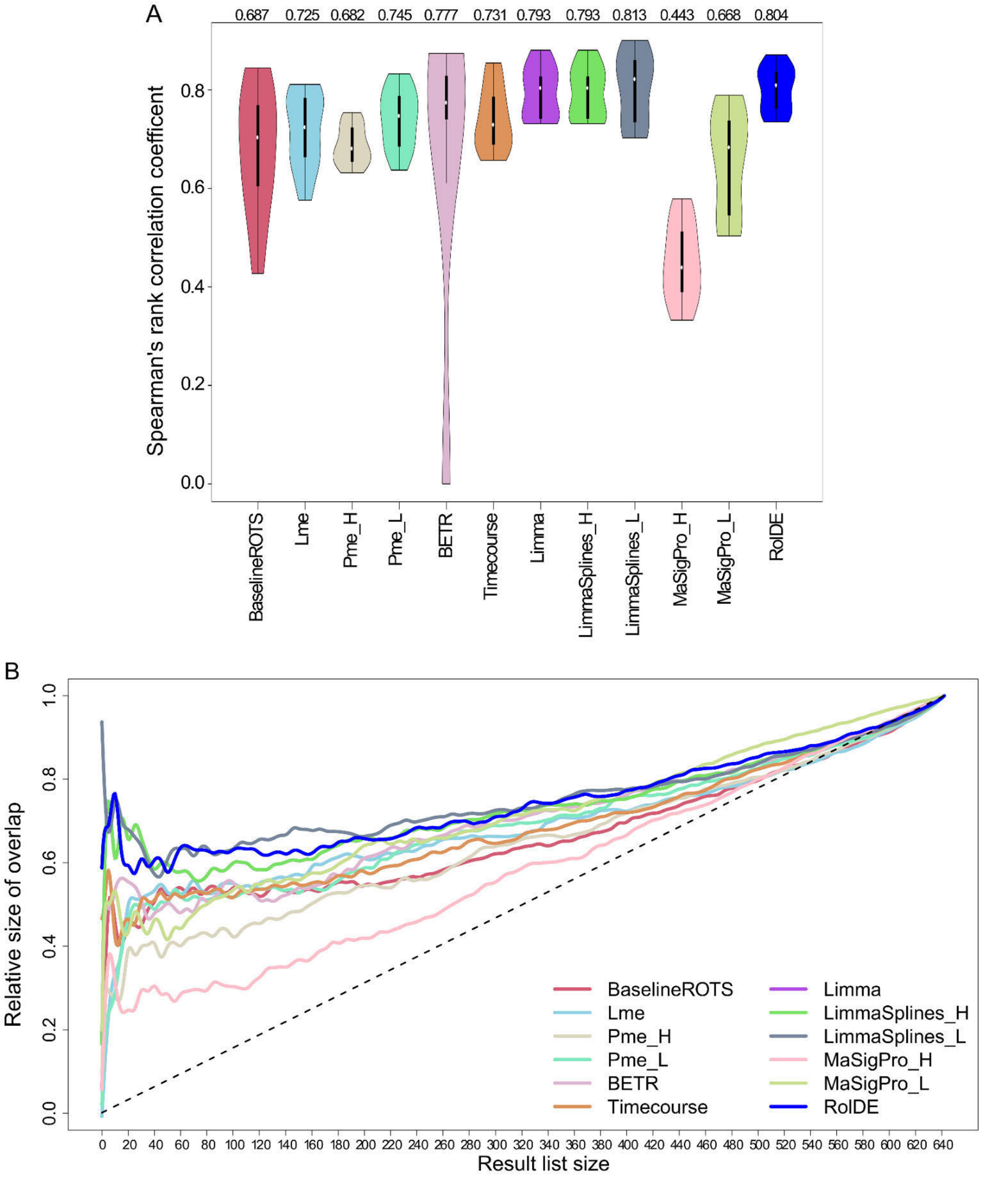


Figure 26. Reproduciblity in of the results of the evaluated methods. A) Spearman's rank correlation coefficients between technical replicate datasets over all the pairwise comparisons of longitudinal differential expression between different strains in the *Fn* dataset. B) Median proportional overlaps of the top *k* longitudinal differential expression findings between technical replicate dataset result lists over all the pairwise comparisons of strains in the *Fn* dataset. The included number of proteins, *k*, was varied from 1 to the length of the entire dataset. Adopted with permission from **Publication III**: Figure 4.

5.4.4 Biological relevance of the findings

All 16 proteins from the KEGG Lipopolysaccharide Biosynthesis pathway (ftn00540) were detected in our data, complemented with two proteins from the associated knockout pathway ko00540. These 18 proteins, assumed to be affected by the modified acyltransferases of the null mutants, were most often detected in the top results of RolDE, Pme_L and MaSigPro_L (**Table 5**).

Table 5. The normalized enrichment scores (NES) from the gene set enrichment analysis (GSEA) of the Lipopolysaccharide synthesis pathway and the associated knockout pathway proteins among the findings of the different methods in comparisons of the acyltransferase null mutant strains and the wild type. The NES of the methods were ranked within each comparison and a mean rank was calculated over all the comparisons for each method. Adopted with permission from **Publication III**: Figure 4.

	Gene Set Enrichment Analysis (GSEA) Normalized Enrichment Score (NES)			Method GSEA NES ranks within comparisons			
	WT vs. L	**WT vs. D2**	**WT vs. D1**	**WT vs. L**	**WT vs. D2**	**WTv vs. D1**	**Mean Rank**
BaselineROTS	0.879	1.008	1.009	12	12	12	12.0
Lme	1.090	1.226	1.199	11	4	7	7.3
Pme_H	1.354	1.019	1.117	4	11	10	8.3
Pme_L	1.475	1.140	1.496	1	9	2	4.0
BETR	1.154	1.185	1.533	8	8	1	5.7
Timecourse	1.095	1.216	1.480	10	6	3	6.3
Limma	1.228	1.226	1.160	6	3	8	5.7
LimmaSplines_H	1.221	1.218	1.151	7	5	9	7.0
LimmaSplines_L	1.152	1.115	1.401	9	10	4	7.7
MaSigPro_H	1.238	1.197	1.248	5	7	5	5.7
MaSigPro_L	1.367	1.531	1.054	3	1	11	5.0
RolDE	1.422	1.235	1.207	2	2	6	**3.3**

5.5 Knowledge enrichment through integrated functional enrichment and network analysis

5.5.1 Quantitative proteomics reveals the dynamic protein landscape during initiation of human Th17 cell polarization

The discovered Th17 proteome in **publication IV** clustered into one main and several smaller clusters according to the known interactions within the differentially regulated proteins between the Th17 and Th0 conditions. Two tight clusters (clusters 2 and 3) of interacting DE proteins similarly upregulated at 24h in the Th17 condition related to lipid biosynthesis and metabolism were discovered (**Figure 27A**). Fatty acids and lipid metabolism have been shown to be involved in driving Th17 differentiation [178–180]. Discovered interacting enzymes ACSL1, ACSL3 and ACSL4, all associated with lipid metabolism and commonly upregulated in the Th17 condition (**Figure 27A**), could provide valuable targets for future research related to Th17 mediated autoimmune diseases.

Over half of the interacting proteins in the main cluster, cluster 1, were associated with immune responses. Furthermore, pathways related to Th17 cell differentiation, Th1 and Th2 cell diffentiation were discovered as enriched in the subcluster 1 (**Figure 27B**). Coincidently, many proteins in subclusters 1 and 3 are known modulators of Th17 differentiation [85]. Antiviral pathways, including the interferon alpha/beta signaling pathway, were highly enriched in subcluster 2, consisting of interacting proteins mostly upregulated similarly in the Th17 condition at 72h. Interestingly, interferon-beta-1a (IFN-β-1a) has been also observed to inhibit Th17 cell differentiation [181]. Many of the proteins in subcluster 2 could thus provide targets for future research regarding their relationship with Th17 expression and differentiation. In summary, co-expressing clusters of interacting proteins with common interesting Th17 related functions were discovered.

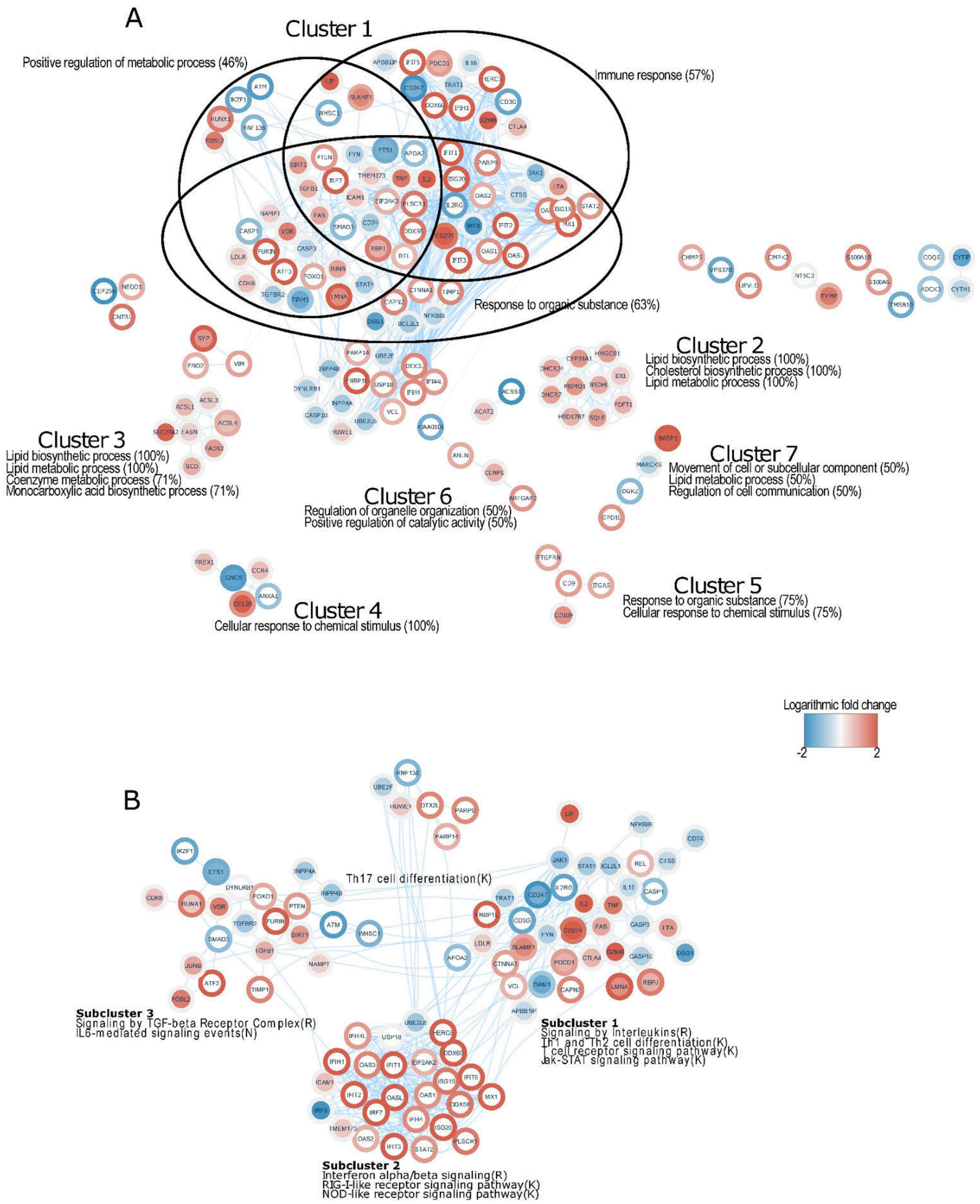

A
Cluster 1
Positive regulation of metabolic process (46%)
Immune response (57%)
Response to organic substance (63%)
Cluster 2
Lipid biosynthetic process (100%)
Cholesterol biosynthetic process (100%)
Lipid metabolic process (100%)
Cluster 3
Lipid biosynthetic process (100%)
Lipid metabolic process (100%)
Coenzyme metabolic process (71%)
Monocarboxylic acid biosynthetic process (71%)
Cluster 4
Cellular response to chemical stimulus (100%)
Cluster 5
Response to organic substance (75%)
Cellular response to chemical stimulus (75%)
Cluster 6
Regulation of organelle organization (50%)
Positive regulation of catalytic activity (50%)
Cluster 7
Movement of cell or subcellular component (50%)
Lipid metabolic process (50%)
Regulation of cell communication (50%)
Logarithmic fold change
-2
2
B
Th17 cell differentiation(K)
Subcluster 1
Signaling by Interleukins(R)
Th1 and Th2 cell differentiation(K)
T cell receptor signaling pathway(K)
Jak-STAT signaling pathway(K)
Subcluster 2
Interferon alpha/beta signaling(R)
RIG-I-like receptor signaling pathway(K)
NOD-like receptor signaling pathway(K)
Subcluster 3
Signaling by TGF-beta Receptor Complex(R)
IL6-mediated signaling events(N)

Figure 27. The protein-protein interaction (PPI) network within the DE Proteins between Th17 and Th0 cells. A) The PPI network from the DE proteins between Th17 and Th0 cells over both the 24h and 72h time points. B) Cluster 1 of the PPI network divided into sub-clusters and functionally annotated with identified enriched pathways. Letters in parentheses after pathway names denote sources for the annotations: R, Reactome; K, KEGG; N, NCI PID. The logarithmic fold changes between Th17 and Th0 depicted as continuous color mapping (at 24 h node inner color, at 72h node outer color). Adopted with permission from **Publication IV**: Figure 4.

5.5.2 Protein interactome of the Cancerous Inhibitor of protein phosphatase 2A (CIP2A) in Th17 cells

As a result of the performed MS analysis in **publication V**, a novel protein-protein interactome of Cancerous inhibitor of PP2A (CIP2A) was constructed. A large cluster of CIP2A interacting proteins with previous known or predicted interactions related to RNA metabolic processes and RNA splicing was discovered (**Figure 28**). RNA metabolic processes and RNA splicing as the main functions of the detected interactome was further confirmed by a separate statistical overrepresentation analysis performed in PANTHER [118]. RNA splicing has been observed to be one of the main CIP2A regulated processes [182]. CIP2A is an oncogene, first detected as an inhibitor of PP2A in cancerous cells [183]. Alternative protein isoforms in genes related to apoptosis typically have opposing, antagonistic functions in apoptosis regulation, suggesting an essential role for alternative splicing in the regulation of apoptosis [184]. Thus, in cancer progression, alternative RNA splicing can lead to the promotion of isoforms and pathways that promote cell proliferation and inhibit apoptosis [184,185].

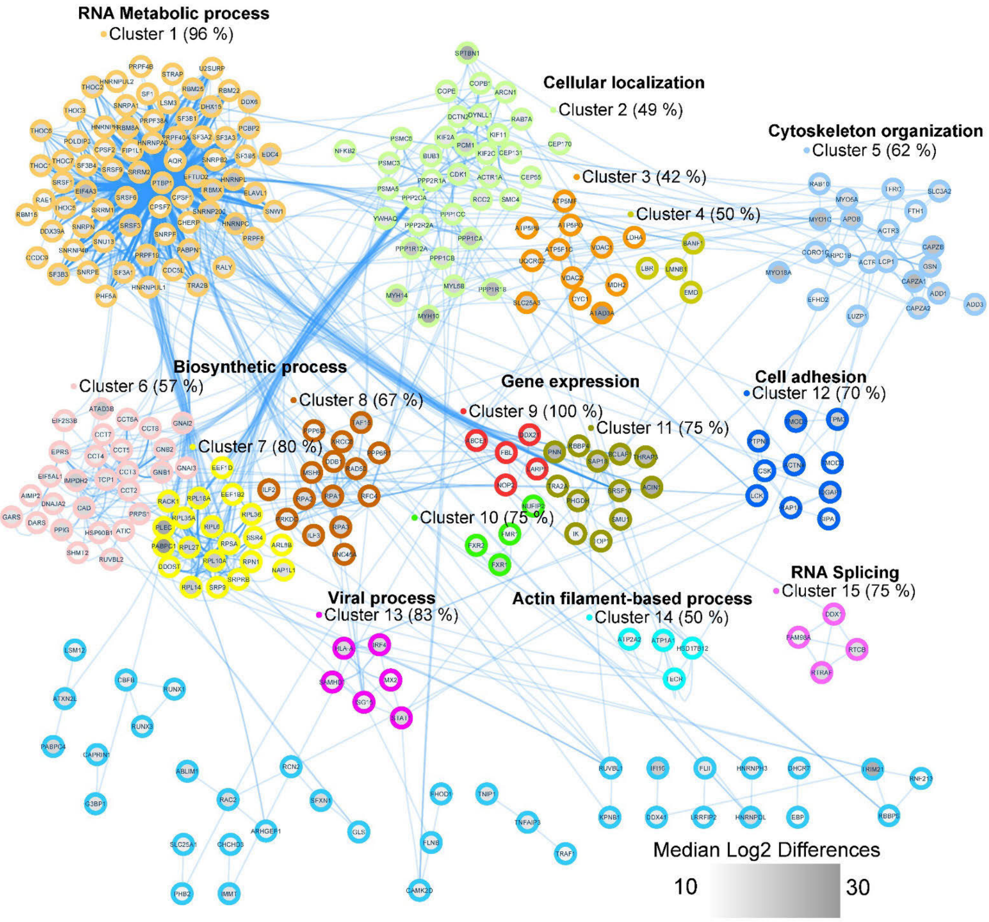


Figure 28. The protein-protein interaction (PPI) network among the proteins interacting with CIP2A. Identified clusters with over two members are indicated by different coloring schemes. For each cluster with over four members, a gene ontology (GO) term representing cluster function is displayed. Node inner color depicts median mass spectrometry intensity differences between the CIP2A immuno-precipitates and the IgG controls. Increasingly grey color represents stronger expression in the CIP2A immuno-precipitates compared to the IgG controls. The STRING database was used to query for the known PPI and the enrichment analysis was performed with DAVID against a Th17 proteome reference background. Adopted with permission from **Publication V**: Figure 4.

Viral processes were also detected as significantly enriched in the CIP2A interactome and the network analysis revealed a cluster of interacting proteins related to viral processes. Many of these proteins' functions are related to viral response and interferon signaling. Interferon regulatory factor 4 (IRF4) for example, has been shown to be involved in the dysregulation of PP2A and the production of interleukin-17 (IL-17), resulting in increased pro-inflammatory response of T cells related to the amplification of autoimmune diseases [186]. CIP2A is known to inhibit PP2A in cancer cells and the depletion of CIP2A in T cells has been shown to result in enhanced IL-17 production [106]. The effect of CIP2A on IRF4 could thus provide an interesting target for future studies and potential therapies.

6 Discussion

6.1 Normalization in proteomics

Normalization in general decreased intragroup variation in the data in **publication I** and improved the consistency of a statistical test in correctly detecting the true DE proteins (**Figures 11-12**). The vsn normalization was observed to consistently produce data from which the true DE proteins could be detected reliably. In previous studies of normalization in proteomics, [36] observed the linear regression normalization using a median reference array and the linear regression normalization complemented with the sample analysis order information to reduce intragroup variation the most. Of the four normalization methods evaluated, [37] also noted the linear regression normalization to perform better than the local regression normalization in removing intragroup variation. However, while no large differences between the examined methods were detected, [29], observed vsn and loess to perform best in reducing intragroup variation. Nonetheless, the exact linearity of the bias seems to vary between experiments, being more linearly dependent on the measured protein abundances in some experiments than in others. As the exact nature of this bias is typically not known, a simple linear regression normalization as a generally applicable method warrants for caution.

Many characteristics of the data, typically unknown beforehand, can affect the performance of a normalization method. Such characteristics may include: the total amount of proteins, the number (typically unknown) of true DE proteins, proportion of missing values, the underlying reason (typically unknown) for the observed missing values (technically censored, MAR/MNAR, missingness related to an experimental condition), the magnitude of differences (typically unknown) between the true DE proteins, amount of noise and bias (typically unknown), etc. Due to many of these underlying characteristics typically being unknown, a method performing adequately well in various different types of conditions and datasets might be more favorable than a method performing excellent in some specific comparisons/datasets while failing in others.

While suggestive, all differential expression performance testing in our normalization work was performed using spike-in datasets with relatively few proteins changing. Further evaluating the performance of the different normalization methods in datasets with a larger proportion of the proteins changing between the samples, would markedly increase the understanding needed for the selection of a proper normalization method in varying experimental settings. A series of spike-in and mixture datasets with a varying proportion of known proteins changing in various known ratios between the samples, could be explored to complement the current findings. Furthermore, the benchmarking of normalization methods in **publication I** was performed on protein level. Normalization can also be performed on peptide level, generally even with the same methods. As the units of interest are typically proteins, evaluating normalization at protein level was chosen to be explored in our current study. However, it would be equally interesting to explore normalization prior the peptide intensities are summarized to proteins and observe if the same methods would be applicable. However, as many of the label-free workflows, such as Progenesis used in the current study, use specialized methods to summarize the peptides within the workflow, the exported normalized peptides would then possibly have to be manually rolled-up to proteins.

As the data produced by the different label-free proteomics software from the same raw data are quite different (e.g. no missing values in Progenesis when compared to MaxQuant/Peaks for example), most likely even the software used has its implications for the choice of a suitable normalization method. The performed exploratory work and case study into normalization in conjunction with **publication IV** demonstrates that even though performing well, the previously observed best acting methods might not be the most suitable choices for all datasets and software. While Vsn and Loess performed well in reducing unwanted intragroup variation between the samples, the performed exploratory work suggested the innate MaxQuant normalization method, MaxLFQ, as the most suitable method for the normalization of data in **publication IV**.

While the comparison work performed in **publication I** was comprehensive, it was not complete, but rather directive and informative of favorable approaches for normalizing label free DDA proteomics data. As the spectrum of proteomics preprocessing software and possible normalization methods is vast and increasing, a single best method for all datasets and software workflows is unlikely to exist. Thus, in the case of a particular preprocessing software workflow and dataset, it can be reasonable to explore the performance of some of the known well performing approaches together with the provided approach by the software, if available. Moreover, the field is developing constantly and new normalization methods are introduced regularly. Including some of the recently developed methods in similar benchmarking experiments, would be highly informative.

Our evaluation of normalization was performed using **label-free DDA** proteomics data. While most of the conclusions of this study could be expected to be at least partly similar for label-free DIA data, for label-free spectral count data and for labeled proteomics data these conclusions might not be valid. If the nature of the data is very different (ratios, discrete counts) or if the structure of the data is very different (e.g. no similar mean-to-variance relationship, less noise), the requirements for a normalization method might be different.

6.2 Label-free proteomics data processing software workflows

While a considerably different number of proteins were identified and quantified by the different workflows, a substantial amount of the identifications were shared (**Figure 20**). These shared identifications will then be discovered regardless of the software used. Such common identifications, identified by multiple software, could be considered as more reliable. The final list of quantified proteins for each software workflow consisted of proteins identified with enough unique peptides, aligned and quantified reliably enough by the software. Following, even though the identification of the peptides and proteins by search engine plays a large role in the final list of quantified proteins by a software, it is not the only determinator. Also other properties of the software workflow (e.g. transferring identifications between the samples, aligning the samples) affect the final list of the quantified peptides and proteins by a software from the input raw data. The commercial Mascot [11] search engine applied in the Progenesis workflow through the Proteome Discoverer, is one of the most established search engines available and has been observed to deliver good results in the identification of histone modifications from the data [187]. However, in comparisons of the performance of search engines, no notable differences in performance have been observed between Mascot and open source alternatives such as X!Tandem [188,189]. Furthemore, while combining the results from multiple search engines as suggested by [190] is feasible in many software (e.g. OpenMS, Proteios), the final list of identifications also depends on how this combination is then performed.

In our comparison (**publication II**), Progenesis performed consistently well in the differential expression analysis and outperformed the other software when no imputation or filtering was performed. However, it delivered less accurate fold change estimates for the changing spike-in proteins than the other software while the fold changes for the non-changing proteins were estimated most accurately. The number of proteins quantified by Progenesis was markedly lower when compared to the other software, possibly indicating that Progenesis automatically filters out some of the unreliable measurements during its preprocessing of the data. However, even in union datasets, containing all proteins identified and quantified at least by one software, Progenesis outperformed the other software when the data was not imputed or filtered post software. In general, the number of missing values in the datasets generated by the software was the main factor determining the performance of the software in the differential expression analysis. Similar to our results, a previous evaluation of preprocessing software for label-free DDA data [58], observed Progenesis to perform best and to have the least amount of missing values and highest quantification accuracy at the peptide level when compared to MaxQuant and Proteios. In an another earlier comparison of two software [191], Progenesis and the Elucidator suite (discontinued), Elucidator was observed to estimate the fold change ratios better than Progenesis. Elucidator was also observed to correctly quantify the tested QC samples as more similar to each other than Progenesis.

The performance of each of the software workflows is dependent on many parameters. The evaluations performed in the included **publication II** were run with the default settings as much as possible, which might work in favor of some workflows. It is difficult to evaluate to which extent using the common default settings might possibly bias the results of the comparison in favor of some workflows as compared to others. As each software workflow includes multiple parameters (e.g. different ways of assigning unique protein identifications, alignment and matching window sizes, alignment settings, allowed post-translational modifications, etc.) or even alternative algorithms for specific purposes (e.g. alignment, normalization, summarizing peptides to proteins) it is unclear how much the performance of each software workflow could be tuned with optimal choices. Indeed, in **publication I**, the vsn normalization was observed to perform consistently well in Progenesis data when compared to other normalization methods. Based on this experience, the data from all software workflows was normalized using the vsn normalization in **publication II**. However, perhaps for some other software workflow, some other normalization method might have been more suitable, as the data produced from the different workflows have different attributes (e.g. number of missing values, variability of measures, etc.). Then, to exhaustively evaluate the performance of the different software workflows, all relevant parameter and algorithm combinations should be explored, but this soon becomes an infeasible task due to the large number of possible combinations. An expert user, familiar with a certain software workflow, might be able to considerably increase the performance of the chosen workflow by tuning the parameters. This is even more likely, when considering the modular workflows, where specific modules/algorithms for a certain task can be replaced. However, often in practice the software worfkflows are used without much further tuning the parameters of the selected workflow, corresponding to the experimental setting in **publication II**.

As mentioned earlier, the results from multiple database search engines can be combined for more reliable protein identification [190,192]. Similarly, ensemble methods integrating multiple models are routinely used in the field of machine learning for better predictive performance as compared to the separate approaches

[193]. Extending this principle to protein quantification would be a highly interesting prospect. Can the quantification results from the different software workflows be combined for higher reliability and more robust quantifications? However, as mentioned in the previous chapter, the data from different software typically accommodate different attributes (e.g. missing values, variability) and normalization should be carefully considered if combination of data from several softwares were considered to avoid serious batch effects. Furhermore, the required computational time for analyzing the data would markedly increase when using multiple software workflows, possibly limiting the usefulness of this approach for large datasets. Nevertheless, integrating results from multiple approaches remains an interesting topic for possible future research.

In addition to previous reseach work and experience, the available resources further determine the selection of a suitable software, as some of the processing software are commercial (e.g. Progenesis, Peaks) while others are non-commerical (e.g. MaxQuant [42], OpenMS [41]). Furthermore, as the comparisons in **publication II** were performed using label-free DDA data, the results might not be generalizable to other types of proteomics data. DIA data has significantly less missing values, which most likely will affect the performance of the software workflows. Additionally, entirely separate default workflows and modules for labeled data exist in most software, thus the performance of the software with labeled data should be compared separately.

6.3 Missing values and imputation in label-free proteomics

As discussed in the previous chapter and in the results section, the proportion of missing values in the data, especially in the true positives, was the largest determinator of performance in the differential expression analysis (**Figure 23, Table 3**). Missing values are a common phenomenon in proteomics data, especially in the label-free DDA proteomics data, and thus it is essential how they are processed. The proportions of missing values were observed to vary highly in data preprocessed with the different software workflows in **publication II**. Progenesis performs a type of imputation already together with its effective alignment algorithm and Progenesis data was observed to contain virtually no missing values. However, even among the other evaluated software, there were highly varying proportions of missing values in the data generated by them.

In the tested datasets and software, the more complex local similarity and global structure based methods, defined as MCAR/MAR methods by [61], outperformed the simple value based approaches. Lazar et al. [61] also concluded that without prior knowledge into the origin of the missing values, a MAR/MCAR approach should be preferred as they are more general in nature. Devoted MNAR methods, such as the back, censored, and zero imputation in our comparison, assume that the missing values result purely from abundance dependent left censoring and will perform poorly on other types of missing values [61]. However, if the missing values are known to be purely MNAR, the single value approaches can perform well and offer a simple approach to deal with missigness in the data [61]. In the tested spike-in datasets, missing values in the spike-in proteins had a larger impact on performance in the differential expression analysis than missing values in the background proteins. The spike-in proteins in the tested datasets are lowly expressed in some sample groups while highly expressed in others. As observed, simply substituting all the missing values in the spike-in proteins with proxies for low expression was not a fruitful strategy in our comparison, indicating that at least partly the missing values in the spike-in proteins are not resulting from left censoring. Similarly, [62] have observed the local similarity based approaches to outperform the simple single value imputations.

Filtering followed by imputation with the lls method was observed to improve the performance in the differential expression analysis the most. While filtering might be an effective approach in improving the detection of the truly DE proteins in the remaining filtered dataset, consideration should be applied, as some of the truly DE proteins can be also filtered out. In this comparison, performance in the filtered datasets was only assessed with regards to the remaining spike-in proteins in the dataset. Application of the bpca or lls imputations methods resulted mostly in improved performance or only slightly decreased performance in the differential expression analysis, while preserving all the proteins. The choice of a suitable imputation/filtering approach then depends also on whether the emphasis is in detecting as many true DE proteins as possible or in detecting less true DE proteins as reliably as possible. Without prior knowledge and depending on the emphasis of the researcher, the filtering, lls or bpca approaches could be considered as valid candidates in dealing with missing values. However, as was observed in the case of the UPS1 dataset in our comparison, sometimes the best performance is achieved when no imputation or filtering is performed. Thus, the decision to impute should depend on the proportion of missing values in the data but also on the ability of the downstream statistical analysis tool in dealing with missing values.

In **publication II**, the work related to missing values and imputation was performed on protein level. Lazar et al. [61] suggested that imputation should be performed already at the peptide level. Missing peptide values will contribute to protein values in various ways, depending on the chosen peptide-to-protein aggregation method. If the peptides are aggregated to proteins by simply summing the peptide abundances, missing peptide values will not contribute to protein values and this aggregation method would coincide with imputing zero for the missing peptide values. If protein level values are aggregated via averaging, missing peptide values will again not contribute to protein level values, which is the same as imputing an average of all the proteins peptide values for each missing peptide in a sample. Thus protein aggregation can be considered an imputation method itself and will impact the resulting pattern of missing values at protein level in various ways. A protein level missing value will only occur if all/enough peptides for a given protein are missing.

A similar investigation as performed in the included work with imputation performed already at the peptide level would be highly interesting. How would imputation at the peptide level affect the performance in the differential expression analysis? However, as many of the label software workflows evaluated in the current work use their own specialized peptide-to-protein aggregation methods, the extracted and imputed peptide level intensities would possibly need to be aggregated into proteins manually via other methods.

6.4 Longitudinal differential expression in proteomics

In the examined semi-simulated spike-in datasets in **publication III**, the new proposed method RolDE, displayed excellent performance, outperforming the other evaluated methods in each tested dataset type (UPS1, SGSDS, CPTAC, UPS1 mix) (**Figure 24**). Furthermore, the performance of RolDE was most balanced over categories of trend differences (**Figure 25**), indicating that RolDE can detect various kinds of longitudinal differential expression reliably. Specifically, RolDE clearly outperformed the other examined methods in the SGSDS and CPTAC full datasets. Missing values in general clearly decreased the performance of all the other methods except RolDE. In the CPTAC datasets, with a high proportion of missing values in the true positive spike-in proteins, all methods performed clearly worse than in other datasets with less missing values. As discussed in the previous chapters, missing values are prevalent in proteomics data, especially the popular label-free DDA technique, and the ability of the chosen differential expression method to withstand them is crucial for reliable results.

There is typically no prior knowledge of the types of longitudinal trend and expression level differences that can be found from the data. The ability of a method to consistently detect various trend differences as well as expression level differences between the conditions, is thus an essential quality for an exhaustive analysis of the data. In their comparison of longitudinal differential methods for RNASeq data [84], observed that many of the specific longitudinal DE methods performed worse than the traditional pairwise methods when the number of timepoints was low (<8). As currently relatively few timepoints are typical for longitudinal omics data, methods unable to perform reliably on such short timeseries data are limited in their usability. The examined semi-simulated and experimental datasets in this study contained 5 (UPS1, CPTAC, *Fn*), 8 (SGSDS) or 9 (T1D dataset of [88]) timepoints. Of the specifically designed longitudinal DE methods examined in our study, only RolDE and Timecourse consistently outperformed the pairwise baseline method ROTS (BaselineROTS). However, when missing values were present in the spike-in proteins as well as in the background proteins (SGSDS and CPTAC datasets), only RolDE was able to outperform the established pairwise differential expression method ROTS, known to perform well in cross-sectional data [32,131,159,161,170].

As the discovered protein biomarker candidates in a discovery proteomics study are commonly validated with other techniques [65], the top DE hits produced by the used differential expression method are of great importance. The overall reproducibility but especially the reproducibility of the top results are therefore of great interest to a researcher. Good top reproducibility indicates the competence of a method to reliably and robustly deliver the same findings in the presence of experimental random noise. Best reproducibility of the top results was observed with Limma and RolDE.

Generally, the higher order polynomial regression models outperformed the lower degree models. This is most likely due to the fact that the higher order models were able to detect trend differences in a broader spectrum of categories than the less complex models. Orthogonal polynomials were used with all the polynomial regression approaches. As the polynomials of different order can be highly correlated, using orthogonal polynomials can reduce such multicollinearity and allow for more independent inspection of coefficients of different polynomial degree [168,169].

Allowing for the overall time-associated changes and inspecting the overall significance values of the regression models instead of specific condition-related coefficients, might improve also the detection of the longitudinally DE proteins and increase the performance of most of the regression based methods (e.g. MaSigPro, Pme). This increase in performance would then result from the longitudinally DE proteins having an overall time-associated trend in their expression over the examined conditions in addition to being DE between the conditions. Thus, when only time-associated changes are also examined, the DE proteins can be detected as a consequence (not specifically because they are DE but because they have overall time-associated changes). However, if the interest is specifically in longitudinal differential expression and allowing also for the overall-time associated changes in ranking of the results, non-DE features with an on overall strong time-associated trend might also be detected, resulting in false positive detections. The prevalence of such overall non-DE time-associated changes are highly data specific and must be assessed in the experimental context where the method is used. Thus, the efficient applicability of methods such as MaSigPro, where by default the overall time-associated changes are included, require careful consideration and interpretation from the researcher.

Due to practical reasons, the timepoints in different conditions of longitudinal data are not always perfectly aligned. This is especially true for clinical data involving humans [86,88]. Furthermore, the number of timepoints between individuals within and between conditions might vary. The flexibility of a differential expression method to take many kinds of different experimental settings into account, reflect the general applicability of the method. RolDE can be used to detect longitudinal differential expression even when the number of timepoints differ between conditions and/or the individual timepoints and are not aligned. To conclude, the choice of a well-performing suitable longitudinal DE method results in the robust and reliable detections of interesting findings for futher analysis or validation.

6.5 Knowledge enrichment through integrated functional enrichment and network analysis

Through combining functional enrichment analysis, i.e the statistical overrepresentation of GO terms and pathway analysis, with protein-protein interaction networks, tightly connected clusters of interacting proteins with common functionalities were discovered (**Figure 27, Figure 28**). Moreover, many functionalities discovered for the detected clusters among the proteins of interest, are related to known important biological processes and disease progression. Thus, knowledge enrichment through combined functional and network analysis can bolster the biological and biomedical conclusions of the performed statistical analysis (i.e. DE analysis) and suggest potential biomarker candidates for further studies (e.g IRF4, ACSL4).

As discussed in chapter 2.7, the used background for the statistical overrepresentation analysis markedly effects the results of the analysis [110,120–122]. As both of the analysis in **publication IV** and **publication V**, were performed in the Th17 cell environment, the same background of all the detected proteins in the activated Th17 cells, CD3/CD28 activated non-differentiated Th0 cells and naive T helper precursor (Thp) cells was used in both studies. As the selected background defines the biological environment (and the possible enrichment terms / pathways) for the enrichment analysis, the selection of a suitable background is essential for a successful overrepresentation analysis.

With the choice of visualization style and effective use of color-coding, several informative attributes, such as fold change at different timepoints or difference in expression between the immunoprecipitates and controls, can be included in the network. Through the use of a proper visualization theme and refinement of the results, even more complex characteristics can be displayed, such as in [106]. In the network of [106], where the interactome of phosphorylated STAT3 was explored in two conditions, CIP2A-deficient and CIP2A-sufficient Th17 cells, node inner color represents the specificity of the interaction of the prey proteins with the bait protein (STAT3). Thus, the effect of CIP2A on the interaction of all the detected prey proteins with STAT3 can be viewed/explored straightforwardly through the visualized network.

To conclude, the combination of a properly selected functional enrichment analysis with a suitable network analysis and visualization, can provide valuable additional insight and knowledge about the biological mechanisms involved within the POI, essentially help in interpreting the findings and transforming data to knowledge.

7 Conclusions

As has been demonstrated in this thesis, proper preprocessing and analysis of the data is crucial for reliable findings in label-free proteomics. The use of suitable methodology for different stages of the workflow can greatly influence the interprations made from the data and add or decrease the effective value of the input raw data. Based on the performed research work, rough outlines for a possible label-free proteomics discovery workflow can be compiled in the following suggested main steps (**Figure 29**):

1. Preprocessing software

Considering the prior expertise and knowledge on label-free data processing software, available resources, algorithmic experience and established related research work (such as **publication II**), an informed decision can be made to select a suitable software for the preprocessing of label-free proteomics data. In **publication II**, the commercial Progenesis software performed overall best as is but was matched by the non-commercial MaxQuant after filtering of missing values or application of proper imputation.

2. Exploration

Initial exploration is an essential step, revealing important properties of the data and affecting choices for the later steps. Proportion of missing values, shape of the intensity distribution, intragroup variability, clustering of the samples according to technical/biological replicates, sample correlations, PCA, etc. can all shed light into characteristics of the data. Log$_2$-transformation of the data and averaging of the technical replicates should be considered where applicable.

3. Normalization

A crucial step for the the downstream analysis of the data. Ideally, several previously observed well-performing normalization methods (e.g. **publication I**) should be explored together with the normalization provided by the chosen software workflow (e.g. MaxLFQ), if available. Some previously well-performing normalization methods for the normalization of label-free proteomics data from **publication I** in this thesis and other works include: the variance stabilization normalization (vsn), local regression (loess), linear regression, linear run order regression and median normalization.

4. Exploration

Data normalized with different methods from step 3 should be thoroughly explored to determine the most suitable approach. MA-plots, boxplots, correlation heatmaps, clustering of the samples according to sample groups, meanSD-plots, intragroup variability (PCV, PEV, PMAD) and missing value proportions in normalized data are some tools in assessing the succesfulnes of normalization.

5. Imputation

In this thesis (**publication II**), the proportion of missing values in the data has been shown to markedly affect the performance of the statistical tests in correctly detecting the true DE proteins. The proportion of missing values in the preprocessed raw data is not equivalent between different software and considerations related to missing values are naturally related to the choice of the processing software in step 1. Filtering of proteins with lots of missing values and/or imputation of missing values are possible solutions for dealing with missing values in the data. In **publication II**, filtering or imputation with the local least squares (lls) and bayesian principle component analysis (bpca*)* methods generally improved the correct detection of the true DE proteins from the data. Furthermore, choices related to the downstream analysis tools in step 7 (e.g. statistical test), and their ability to withstand missing values in the data, greatly affects the need to filter or impute the data.

6. Exploration

If imputation or filtering of the data is performed, the effect of different imputation/filtering methods should be explored with similar metrics as in step 1 and step 4 for the selection of a suitable method.

7. Determination of the POI

Once the data has been suitably preprocessed, it can be further analyzed to determine the proteins of interest. In this thesis, the focus has been on longitudinally DE proteins between the conditions (**publication III**). It has been shown, that different statistical approaches greatly differ in their performance in detecting the true longitudinally DE proteins, especially in the presence of missing values. The new suggested method, RolDE, performed best in all data types, was especially tolerant to missing values, had good reproducibility and was among the top method in producing biologically meaningful results. For cross-sectional data, the Reprodudibility Optimized Test Statistic (ROTS) [158,170] has been shown to perform well [32,131,170].

8. Functional enrichment

Following the determination of the proteins of interest, the enrichment of biological functionalities related to the POI can be performed to increase understanding of the underlying biological mechanisims. For statistical overrepresentation analysis of terms and pathways among the POI, a suitable background should be used. Some popular tools for investigating statistical overrepresentation of ontology terms or pathways in a list of POIs include PANTHER [108], DAVID [107], KEGG [115], IPA [190] and Reactome [116]. Alternatively, for exploring the enrichment of terms and pathways in the whole data, GSEA can be performed using an appropriate score.

9. Network construction

To explore protein-protein interactions within the POI, network databases such as STRING or ReactomeFiViz can be queried.

10. Integrated knowledge enrichment and visualization

Finally, the processed data, POI, biological functional information and network information can be integrated and visualized for a clear and meaningful representation of the findings from the data (e.g. **publications IV** and **V**). Such information-rich representation can be applied in selecting the most potential and interesting targets for validation or further future research work.

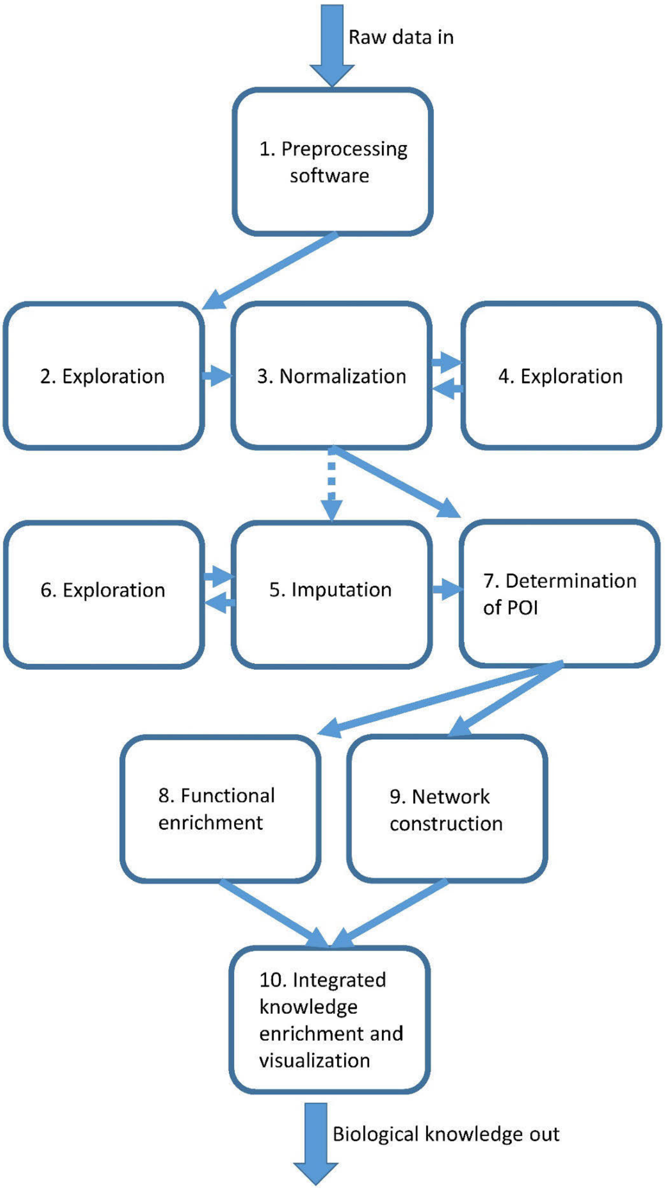


Figure 29. A potential label-free proteomics discovery data processing workflow. The dashed lines represent possible alternative paths.

Acknowledgements

First, I would like to express my gratitude to my supervisor and mentor into the world of bioinformatics: Professor Laura Elo, whose guidance and support has made the completion of this thesis possible. I am especially thankful for all the opportunities provided to participate in interesting projects, smaller and larger collaborations and research visits. I want to recognize Professor Matti Nykter and Professor Veit Schwämmle for reviewing my thesis and providing me valuable comments and suggestions. I am grateful for Dr. Markku Varjosalo for accepting the invitation and taking the time to be my opponent. I would like to thank the members of my PhD thesis advisory board committee – Professor David Goodlett and Dr. Kalle Rytkönen, thank you for your guidance and comments for my thesis plan. I would like to thank Dr. Tomi Suomi for working with me in most of the projects during these years and offering me his guidance and opinions. I feel privileged to have been able to work with an inspiring group of bioinformaticians, computational biologists and other experts and want to thank the whole group - past and present- of the Medical Bioinformatics Centre. I would like to thank Professor Riitta Lahesmaa and also the Molecular Systems Immunology group - past and present- for including me in many interesting and diverse collaboration projects and encouraging me on my way. I would like to especially recognize all my co-authors: Tomi Suomi, Laura Elo, Courtney Chandler, Alison Scott, Bao Tran, Robert Ernst, David Goodlett, Subhash Tripathi, Ankitha Shetty, Mohd Moin Khan, Robert Moulder, Santosh Bhosale, Elina Komsi, Verna Salo, Rafael Sales De Albuquerque, Omid Rasool, Sanjeev Galande, Meraj Hasan Khan, Ubaid Ullah, Umar Butt, Xi Qiao and Jukka Westermarck. Without their contributions, this research work would not have been possible. I would like to acknowledge the former Turku University Graduate School (UTUGS) Doctoral Programme in Mathematics and Computer Sciences Medicine (MATTI) and current Doctoral Programme in Technology for funding this thesis and for organizing doctoral training in Turku. I would like to give thanks to Mehrad Mahmoudian for advice after my Master’s studies and for facilitating me on a path towards this thesis. I want to thank my beloved wife Mari for her patience and for all the support during the years. Topi, thank you for existing and bringing so much joy to my life. Milla

Välikangas, Henna Välikangas and Jarkko Seppälä, thank you for proofreading the thesis. Lastly, I want to thank my parents, family and friends for all the support along the way.

22.12.2021
Tommi Välikangas

List of References

1. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011; 12:77
2. Meissner F, Mann M. Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. Nat Immunol 2014; 15:112–117
3. Altelaar AFM, Munoz J, Heck AJR. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 2013; 14:35–48
4. Tuli L, Tsai T-H, Varghese RS, et al. Using a spike-in experiment to evaluate analysis of LC-MS data. Proteome Sci. 2012; 10:13
5. Megger DA, Bracht T, Meyer HE, et al. Label-free quantification in clinical proteomics. Biochim. Biophys. Acta - Proteins Proteomics 2013; 1834:1581–1590
6. Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. Nat. Rev. Mol. Cell Biol. 2004; 5:699–711
7. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003; 422:198–207
8. Srzentić K, Fornelli L, Laskay ÜA, et al. Advantages of Extended Bottom-Up Proteomics Using Sap9 for Analysis of Monoclonal Antibodies. Anal. Chem. 2014; 86:9945–9953
9. Gregorich ZR, Chang Y-H, Ge Y. Proteomics in heart failure: top-down or bottom-up? Pflugers Arch. 2014; 466:1199–1209
10. Nadler WM, Waidelich D, Kerner A, et al. MALDI versus ESI: The Impact of the Ion Source on Peptide Identification. J. Proteome Res. 2017; 16:1207–1215
11. Perkins DN, Pappin DJ, Creasy DM, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999; 20:3551–3567
12. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004; 20:1466–1467
13. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 2014; 5:5277
14. Cox J, Neuhauser N, Michalski A, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. 2011; 10:1794–1805
15. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000; 28:45–48
16. Zhu W, Smith JW, Huang CM. Mass spectrometry-based label-free quantitative proteomics. J. Biomed. Biotechnol. 2010; 2010:
17. Ong S-E, Blagoev B, Kratchmarova I, et al. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Mol. Cell. Proteomics 2002; 1:376–386
18. Cox J, Hein MY, Luber CA, et al. Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Mol. Cell. Proteomics 2014; 13:2513–2526
19. Krey JF, Wilmarth PA, Shin J-B, et al. Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. J. Proteome Res. 2014; 13:1034–1044

20. Bantscheff M, Schirle M, Sweetman G, et al. Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. 2007; 389:1017–1031
21. Quantitative Proteomics. 2014;
22. Grossmann J, Roschitzki B, Panse C, et al. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. J. Proteomics 2010; 73:1740–1746
23. Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. Nat Biotech 2010; 28:710–721
24. Lange V, Picotti P, Domon B, et al. Selected reaction monitoring for quantitative proteomics: a tutorial. Mol. Syst. Biol. 2008; 4:222
25. Koopmans F, Ho JTC, Smit AB, et al. Comparative Analyses of Data Independent Acquisition Mass Spectrometric Approaches: DIA, WiSIM-DIA, and Untargeted DIA. Proteomics 2018; 18:
26. MacMullan MA, Dunn ZS, Graham N, et al. Quantitative Proteomics and Metabolomics Reveal Biomarkers of Disease as Potential Immunotherapy Targets and Indicators of Therapeutic Efficacy. Theranostics 2019; 9:7872–7888
27. Sinitcyn P, Hamzeiy H, Salinas Soto F, et al. MaxDIA enables library-based and library-free data-independent acquisition proteomics. Nat. Biotechnol. 2021;
28. Hu A, Noble WS, Wolf-Yadlin A. Technical advances in proteomics: new developments in data-independent acquisition. F1000Research 2016; 5:
29. Chawade A, Alexandersson E, Levander F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. J. Proteome Res. 2014; 13:3114–3120
30. Karpievitch Y V, Taverner T, Adkins JN, et al. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. Bioinformatics 2009; 25:2573–2580
31. Karpievitch Y V, Dabney AR, Smith RD. Normalization and missing value imputation for label-free LC-MS analysis. BMC Bioinformatics 2012; 13:S5–S5
32. Välikangas T, Suomi T, Elo LL. A systematic evaluation of normalization methods in quantitative label-free proteomics. Brief. Bioinform. 2018; 19:1–11
33. Bolstad BM, Irizarry RA, Åstrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma. 2003; 19:185–193
34. Huber W, von Heydebreck A, Sültmann H, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 2002; 18 Suppl 1:S96–S104
35. Karpievitch Y V, Nikolic SB, Wilson R, et al. Metabolomics Data Normalization with EigenMS. 2014; 1–10
36. Kultima K, Nilsson A, Scholz B, et al. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. Mol. Cell. Proteomics 2009; 8:2285–2295
37. Callister SJ, Barry RC, Adkins JN, et al. Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics. J. Proteome Res. 2006; 5:277–286
38. Sandin M, Teleman J, Malmström J, et al. Data processing methods and quality control strategies for label-free LC–MS protein quantification. Biochim. Biophys. Acta - Proteins Proteomics 2014; 1844:29–41
39. Teleman J, Chawade A, Sandin M, et al. Dinosaur: A Refined Open-Source Peptide MS Feature Detector. J. Proteome Res. 2016; 15:2143–2151
40. Sandin M, Krogh M, Hansson K, et al. Generic workflow for quality assessment of quantitative label-free LC-MS analysis. Proteomics 2011; 11:1114–1124
41. Sturm M, Bertsch A, Gröpl C, et al. OpenMS – An open-source software framework for mass spectrometry. BMC Bioinformatics 2008; 9:163
42. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotech 2008; 26:1367–1372

43. Sandin M, Ali A, Hansson K, et al. An Adaptive Alignment Algorithm for Quality-controlled Label-free LC-MS. Mol. Cell. Proteomics 2013; 12:1407–1420
44. Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. J. Proteome Res. 2004; 3:958–964
45. Hakkinen J, Vincic G, Mansson O, et al. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. J. Proteome Res. 2009; 8:3037–3043
46. Zhang J, Xin L, Shan B, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol. Cell. Proteomics 2012; 11:M111.010587
47. Ma B, Zhang K, Hendrie C, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. 2003; 17:2337–2342
48. Yang H, Chi H, Zeng W-F, et al. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. Bioinformatics 2019; 35:i183–i190
49. Choi S, Paek E. MutCombinator: identification of mutated peptides allowing combinatorial mutations using nucleotide-based graph search. Bioinformatics 2020; 36:i203–i209
50. Pevzner PA, Mulyukov Z, Dancik V, et al. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res. 2001; 11:290–299
51. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. J. Bioinform. Comput. Biol. 2005; 3:697–716
52. Zhang H, Liu T, Zhang Z, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell 2016; 166:755–765
53. Krug K, Jaehnig EJ, Satpathy S, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell 2020; 183:1436-1456.e31
54. Woo S, Cha SW, Na S, et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics 2014; 14:2719–2730
55. Vandenbogaert M, Li-Thiao-Té S, Kaltenbach H-M, et al. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. Proteomics 2008; 8:650–672
56. Cox J, Hein MY, Luber C a, et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. … 2014; 13:2513–2526
57. Schwanhüusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. Nature 2011; 473:337–342
58. Chawade A, Sandin M, Teleman J, et al. Data Processing Has Major Impact on the Outcome of Quantitative Label-Free LC-MS Analysis. J. Proteome Res. 2015; 14:676–687
59. Häkkinen J, Vincic G, Månsson O, et al. The Proteios Software Environment: An Extensible Multiuser Platform for Management and Analysis of Proteomics Data. J. Proteome Res. 2009; 8:3037–3043
60. Valikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief. Bioinform. 2018; 19:1344–1355
61. Lazar C, Gatto L, Ferro M, et al. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. J. Proteome Res. 2016; 15:1116–1125
62. Webb-Robertson B-JM, Wiberg HK, Matzke MM, et al. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. J. Proteome Res. 2015; 14:1993–2001
63. Wang P, Tang H, Zhang H, et al. Normalization Regarding NonRandom Missing Values in High-Throughput Mass Spectrometry Data. Pacific Symp. Biocomput. 2006; 315–326
64. Crutchfield CA, Thomas SN, Sokoll LJ, et al. Advances in mass spectrometry-based clinical biomarker discovery. Clin. Proteomics 2016; 13:1

65. Geyer PE, Holdt LM, Teupser D, et al. Revisiting biomarker discovery by plasma proteomics. Mol. Syst. Biol. 2017; 13:942
66. Serna G, Ruiz-Pace F, Cecchi F, et al. Targeted multiplex proteomics for molecular prescreening and biomarker discovery in metastatic colorectal cancer. Sci. Rep. 2019; 9:13568
67. Enroth S, Berggrund M, Lycke M, et al. High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. Commun. Biol. 2019; 2:221
68. Sajic T, Ciuffa R, Lemos V, et al. A new class of protein biomarkers based on subcellular distribution: application to a mouse liver cancer model. Sci. Rep. 2019; 9:6913
69. Group F-NBW, others. BEST (Biomarkers, EndpointS, and other Tools) resource. 2016;
70. Califf RM. Biomarker definitions and their applications. Exp. Biol. Med. (Maywood). 2018; 243:213–221
71. He T. Implementation of Proteomics in Clinical Trials. PROTEOMICS – Clin. Appl. 2019; 13:1800198
72. Cominetti O, Núñez Galindo A, Corthésy J, et al. Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry. J. Proteome Res. 2016; 15:389–399
73. Levin Y. The role of statistical power analysis in quantitative proteomics. Proteomics 2011; 11:2565–2567
74. Guo Y, Logan HL, Glueck DH, et al. Selecting a sample size for studies with repeated measures. BMC Med. Res. Methodol. 2013; 13:100
75. Lu N, Han Y, Chen T, et al. Power analysis for cross-sectional and longitudinal study designs. Shanghai Arch. psychiatry 2013; 25:259–262
76. Xu Z, Shen X, Pan W, et al. Longitudinal Analysis Is More Powerful than Cross-Sectional Analysis in Detecting Genetic Association with Neuroimaging Phenotypes. PLoS One 2014; 9:e102312
77. Cho RJ, Campbell MJ, Winzeler EA, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 1998; 2:65–73
78. Cho RJ, Huang M, Campbell MJ, et al. Transcriptional regulation and function during the human cell cycle. Nat. Genet. 2001; 27:48–54
79. Greenall A, Lei G, Swan DC, et al. A genome wide analysis of the response to uncapped telomeres in budding yeast reveals a novel role for the NAD+ biosynthetic gene BNA2 in chromosome end protection. Genome Biol. 2008; 9:R146–R146
80. Karlovich C, Duchateau-Nguyen G, Johnson A, et al. A longitudinal study of gene expression in healthy individuals. BMC Med. Genomics 2009; 2:33
81. Oh S, Song S, Grabowski G, et al. Time series expression analyses using RNA-seq: a statistical approach. Biomed Res. Int. 2013; 2013:203681
82. Aijo T, Butty V, Chen Z, et al. Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. Bioinformatics 2014; 30:i113-20
83. Tuomela S, Rautio S, Ahlfors H, et al. Comparative analysis of human and mouse transcriptomes of Th17 cell priming. Oncotarget 2016; 7:13416–13428
84. Spies D, Renz PF, Beyer TA, et al. Comparative analysis of differential gene expression tools for RNA sequencing time course data. Brief. Bioinform. 2017; 20:288–298
85. Tripathi SK, Valikangas T, Shetty A, et al. Quantitative Proteomics Reveals the Dynamic Protein Landscape during Initiation of Human Th17 Cell Polarization. iScience 2019; 11:334–355
86. Lietzén N, Cheng L, Moulder R, et al. Characterization and non-parametric modeling of the developing serum proteome during infancy and early childhood. Sci. Rep. 2018; 8:5883
87. Giddey AD, de Kock E, Nakedi KC, et al. A temporal proteome dynamics study reveals the molecular basis of induced phenotypic resistance in Mycobacterium smegmatis at sub-lethal rifampicin concentrations. Sci. Rep. 2017; 7:43858
88. Liu C-W, Bramer L, Webb-Robertson B-J, et al. Temporal expression profiling of plasma proteins reveals oxidative stress in early stages of Type 1 Diabetes progression. J. Proteomics 2018; 172:100–110

89. Khan MM, Chattagul S, Tran BQ, et al. Temporal proteomic profiling reveals changes that support Burkholderia biofilms. Pathog. Dis. 2019; 77:
90. Aryee MJ, Gutierrez-Pabello JA, Kramnik I, et al. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). BMC Bioinformatics 2009; 10:409
91. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47–e47
92. Conesa A, Nueda MJ, Ferrer A, et al. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. Bioinformatics 2006; 22:1096–1102
93. Transformation and Normalization BT - Analysis of Microarray Gene Expression Data. 2004; 67–84
94. Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu. Rev. Clin. Psychol. 2010; 6:79–107
95. Chu T-M, Weir B, Wolfinger R. A systematic statistical linear modeling approach to oligonucleotide array experiments. Math. Biosci. 2002; 176:35–51
96. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 2004; 3:Article3
97. Wolfinger RD, Gibson G, Wolfinger ED, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. 2001; 8:625–637
98. Cheng L, Ramchandran S, Vatanen T, et al. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat. Commun. 2019; 10:1798
99. Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. Genet. Res. 2001; 77:123–128
100. Haljasmägi L, Salumets A, Rumm AP, et al. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. Sci. Rep. 2020; 10:20533
101. Magni P, Ferrazzi F, Sacchi L, et al. TimeClust: a clustering tool for gene expression time series. Bioinformatics 2008; 24:430–432
102. Khan MM, Välikangas T, Khan MH, et al. Protein interactome of the Cancerous Inhibitor of protein phosphatase 2A (CIP2A) in Th17 cells. Curr. Res. Immunol. 2020; 1:10–22
103. Schweppe DK, Chavez JD, Lee CF, et al. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. Proc. Natl. Acad. Sci. 2017; 114:1732 LP – 1737
104. Scifo E, Szwajda A, Debski J, et al. Drafting the CLN3 protein interactome in SH-SY5Y human neuroblastoma cells: a label-free quantitative proteomics approach. J. Proteome Res. 2013; 12:2101–2115
105. Sakai Y, Shaw CA, Dawson BC, et al. Protein Interactome Reveals Converging Molecular Pathways Among Autism Disorders. Sci. Transl. Med. 2011; 3:86ra49--86ra49
106. Khan MM, Ullah U, Khan MH, et al. CIP2A Constrains Th17 Differentiation by Modulating STAT3 Signaling. iScience 2020; 23:100947
107. Côme C, Cvrljevic A, Khan MM, et al. CIP2A Promotes T-Cell Activation and Immune Response to Listeria monocytogenes Infection. PLoS One 2016; 11:e0152996–e0152996
108. Chesor M, Roytrakul S, Graidist P, et al. Proteomics analysis of siRNA-mediated silencing of Wilms' tumor 1 in the MDA-MB-468 breast cancer cell line. Oncol. Rep. 2014; 31:1754–1760
109. Wu X, Hasan M Al, Chen JY. Pathway and network analysis in proteomics. J. Theor. Biol. 2014; 362:44–52
110. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 2012; 8:e1002375–e1002375
111. Yugandhar K, Gupta S, Yu H. Inferring Protein-Protein Interaction Networks From Mass Spectrometry-Based Proteomic Approaches: A Mini-Review. Comput. Struct. Biotechnol. J. 2019; 17:805–811

112. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47:D607–D613
113. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. Nat Genet 2000; 25:
114. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47:D330–D338
115. Carbon S, Ireland A, Mungall CJ, et al. AmiGO: online access to ontology and annotation data. Bioinformatics 2009; 25:288–289
116. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2009; 4:44–57
117. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13
118. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017; 45:D183–D189
119. Mi H, Muruganujan A, Casagrande JT, et al. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 2013; 8:1551
120. Bessarabova M, Ishkin A, JeBailey L, et al. Knowledge-based analysis of proteomics data. BMC Bioinformatics 2012; 13 Suppl 1:S13–S13
121. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. Genome Biol. 2015; 16:186
122. Dezso Z, Nikolsky Y, Sviridov E, et al. A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biol. 2008; 6:49
123. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 2005; 102:15545–15550
124. Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. BMC Syst. Biol. 2014; 8 Suppl 2:S3–S3
125. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30
126. Matthews L, Gopinath G, Gillespie M, et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. 2009; 37:D619–D622
127. Nishimura D. BioCarta. Biotech Softw. Internet Rep. 2001; 2:117–120
128. Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009; 37:D674–D679
129. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res. 2006; 34:D504–D506
130. Wu G, Dawson E, Duong A, et al. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Research 2014; 3:146
131. Pursiheimo A, Vehmas AP, Afzal S, et al. Optimization of Statistical Methods Impact on Quantitative Proteomics Data. J. Proteome Res. 2015; 14:4118–4126
132. Deutsch EW, Csordas A, Sun Z, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 2017; 45:D1100–D1106
133. Perez-Riverol Y, Csordas A, Bai J, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019; 47:D442–D450
134. Ramus C, Hovasse A, Marcellin M, et al. Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods. Data Br. 2016; 6:286–294
135. Tabb DDL, Vega-Montoto L, Rudnick P a, et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J. Proteome Res. 2010; 9:761–76

136. Bruderer R, Bernhardt OM, Gandhi T, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen treated 3D liver microtissues. Mol. Cell. Proteomics 2015; mcp.M114.044305
137. Desiere F, Deutsch EW, King NL, et al. The PeptideAtlas project. Nucleic Acids Res. 2006; 34:D655-8
138. Vehmas AP, Adam M, Laajala TD, et al. Liver lipid metabolism is altered by increased circulating estrogen to androgen ratio in male mouse. J. Proteomics 2016; 133:66–75
139. Yun J, Wang X, Zhang L, et al. Effects of lipid A acyltransferases on the pathogenesis of F. novicida. Microb. Pathog. 2017; 109:313–318
140. McLendon MK, Schilling B, Hunt JR, et al. Identification of LpxL, a late acyltransferase of Francisella tularensis. Infect. Immun. 2007; 75:5518–5531
141. Eidhammer I, Barsnes H, Eide GE, et al. Experimental Normalization. Comput. Stat. Methods Protein Quantif. by Mass Spectrom. 2013; 96–109
142. Oberg AL, Mahoney DW. Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. BMC Bioinformatics 2012; 13 Suppl 1:S7–S7
143. Cleveland W, Devlin S. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. J. Am. Stat. Assoc. 1988; 83:596–610
144. Ballman K V, Grill DE, Oberg AL, et al. Faster cyclic loess: normalizing RNA arrays via linear models. Bioinforma. 2004; 20:2778–2786
145. Ting L, Cowley MJ, Hoon SL, et al. Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. Mol. Cell. Proteomics 2009; 8:2227–2242
146. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–1735
147. Han X, He L, Xin L, et al. PeaksPTM: Mass Spectrometry-Based Identification of Peptides with Unspecified Modifications. J. Proteome Res. 2011; 10:2930–2936
148. Schwanhäusser B, Busse D, Li N, et al. Global quantification of mammalian gene expression control. Nature 2011; 473:337–342
149. Junker J, Bielow C, Bertsch A, et al. TOPPAS: A Graphical Workflow Editor for the Analysis of High-Throughput Proteomics Data. J. Proteome Res. 2012; 11:3914–3920
150. Gärdén P, Alm R, Häkkinen J. PROTEIOS: an open source proteomics initiative. Bioinformatics 2005; 21:2085–2087
151. Venable JD, Dong M-Q, Wohlschlegel J, et al. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Meth 2004; 1:39–45
152. Stacklies W, Redestig H, Scholz M, et al. pcaMethods—a bioconductor package providing PCA methods for incomplete data. Bioinforma. 2007; 23:1164–1167
153. Xiang Q, Dai X, Deng Y, et al. Missing value imputation for microarray gene expression data using histone acetylation information. BMC Bioinformatics 2008; 9:252
154. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics 2005; 21:
155. Tuikkala J, Elo L, Nevalainen OS, et al. Improving missing value estimation in microarray data with gene ontology. Bioinformatics 2006; 22:
156. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17:
157. Oba S, Sato MA, Takemasa I, et al. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 2003; 19:
158. Elo, Laura, Filén S, Lahesmaa R, et al. Reproducibility-optimized test statistic for ranking genes in microarray studies. IEEE/ACM Trans. Comput. Biol. Bioinform. 2008; 5:423–31
159. Seyednasrollah F, Rantanen K, Jaakkola P, et al. ROTS: reproducible RNA-seq biomarker detector—prognostic markers for clear cell renal cell cancer. Nucleic Acids Res. 2015; 44:e1–e1
160. Suomi T, Elo LL. Enhanced differential expression statistics for data-independent acquisition proteomics. Sci. Rep. 2017; 7:5869

161. Jaakkola MK, Seyednasrollah F, Mehmood A, et al. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief. Bioinform. 2016; 18:735–743
162. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods 2018; 15:255
163. Gillespie CS, Lei G, Boys RJ, et al. Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays. BMC Res. Notes 2010; 3:81
164. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics 2014; 30:2598–2602
165. Tai YC, Speed TP. A multivariate empirical Bayes statistic for replicated microarray time course data. Ann. Stat. 2006; 34:2387–2412
166. Pusponegoro NH, Rachmawati RN, Notodiputro KA, et al. Linear Mixed Model for Analyzing Longitudinal Data: A Simulation Study of Children Growth Differences. Procedia Comput. Sci. 2017; 116:284–291
167. Molenberghs G, Verbeke G. A review on linear mixed models for longitudinal data, possibly subject to dropout. Stat. Modelling 2001; 1:235–269
168. Bradley RA, Srivastava SS. Correlation in Polynomial Regression. Am. Stat. 1979; 33:11–14
169. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. Introd. to Linear Regres. Anal. 5th Ed. 2012; 672
170. Suomi T, Seyednasrollah F, Jaakkola MK, et al. ROTS: An R package for reproducibility-optimized statistical testing. PLoS Comput. Biol. 2017; 13:
171. Jantzen SG, Sutherland BJG, Minkley DR, et al. GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets. BMC Res. Notes 2011; 4:267
172. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504
173. Morris JH, Apeltsin L, Newman AM, et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC Bioinformatics 2011; 12:436
174. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002; 30:1575–1584
175. Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 2006; 7:488
176. Choi H, Larsen B, Lin Z-Y, et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. Nat. Methods 2011; 8:70–73
177. Mellacheruvu D, Wright Z, Couzens AL, et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods 2013; 10:730–6
178. Berod L, Friedrich C, Nandan A, et al. De novo fatty acid synthesis controls the fate between regulatory T and T helper 17 cells. Nat. Med. 2014; 20:
179. Sun L, Fu J, Zhou Y. Metabolism controls the balance of Th17/T-regulatory cells. Front. Immunol. 2017; 8:
180. Young KE, Flaherty S, Woodman KM, et al. Fatty acid synthase regulates the pathogenicity of Th17 cells. J. Leukoc. Biol. 2017; 102:1229–1235
181. Ramgolam VS, Sha Y, Jin J, et al. IFN-beta inhibits human Th17 cell differentiation. J. Immunol. 2009; 183:5418–5427
182. Kauko O, Imanishi SY, Kulesskiy E, et al. Rules for PP2A-controlled phosphosignalling and drug responses. bioRxiv 2018;
183. Junttila MR, Puustinen P, Niemela M, et al. CIP2A inhibits PP2A in human malignancies. Cell 2007; 130:51–62
184. Schwerk C, Schulze-Osthoff K. Regulation of Apoptosis by Alternative Pre-mRNA Splicing. Mol. Cell 2005; 19:1–13
185. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. 2010; 24:2343–2364

186. Apostolidis SA, Rauen T, Hedrich CM, et al. Protein phosphatase 2A enables expression of interleukin 17 (IL-17) through chromatin remodeling. J. Biol. Chem. 2013; 288:26775–26784
187. Yuan Z-F, Lin S, Molden RC, et al. Evaluation of proteomic search engines for the analysis of histone modifications. J. Proteome Res. 2014; 13:4470–4478
188. Balgley BM, Laudeman T, Yang L, et al. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. Mol. Cell. Proteomics 2007; 6:1599–1608
189. Ivanov M V, Levitsky LI, Lobas AA, et al. Peptide identification in "shotgun" proteomics using tandem mass spectrometry: Comparison of search engine algorithms. J. Anal. Chem. 2015; 70:1614–1619
190. Shteynberg D, Nesvizhskii AI, Moritz RL, et al. Combining Results of Multiple Search Engines in Proteomics. Mol. Cell. Proteomics 2013; 12:2383–2393
191. Runxuan, Zhang; Barton, Alun; Brittenden, Julie; T.-J.Huang, Jeffrey; Crowther D. Evaluation for computational platforms of LC-MS based label-free quantitative proteomics: A global view. J. Proteomics Bioinform. 2010; 3:
192. Svecla M, Garrone G, Faré F, et al. DDASSQ: An open-source, multiple peptide sequencing strategy for label free quantification based on an OpenMS pipeline in the KNIME analytics platform. Proteomics 2021; 21:2000319
193. Rokach L. Ensemble Methods for Classifiers BT – Data Mining and Knowledge Discovery Handbook. 2005; 957–980

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2022

ISBN 978-951-29-8741-2 (PRINT)
ISBN 978-951-29-8742-9 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)