UNIVERSITY OF TURKU

DATA ANALYSIS TOOLS FOR MASS SPECTROMETRY PROTEOMICS

Tomi Suomi

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 3 — TECHNICA – INFORMATICA — TURKU 2021

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral Programme in Mathematics and Computer Sciences

Supervised by

Prof. em. Olli Nevalainen
Faculty of Science and Engineering, University of Turku, Finland

Prof. Jukka Heikkonen
Faculty of Science and Engineering, University of Turku, Finland

Reviewed by

Prof. Sampsa Hautaniemi
Faculty of Medicine, University of Helsinki, Finland

Prof. Lukas Käll
KTH Royal Institute of Technology, Sweden

Opponent

Associate Prof. Fredrik Levander
Faculty of Engineering, Lund University, Sweden

ISBN 978-951-29-8541-8 (PRINT)
ISBN 978-951-29-8542-5 (PDF)
ISSN 2736-9390 (PRINT)
ISSN 2736-9684 (ONLINE)
Grano, Helsinki, Finland, 2021

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
SUOMI, TOMI: Data analysis tools for mass spectrometry proteomics
Doctoral dissertation, 113 pp.
Doctoral Programme in Mathematics and Computer Sciences
February 2021

ABSTRACT

Proteins are large biomolecules which consist of amino acid chains. They differ from one another in their amino acid sequences, which are mainly dictated by the nucleotide sequence of their corresponding genes. Proteins fold into specific three-dimensional structures that determine their activity. Because many of the proteins act as catalytes in biochemical reactions, they are considered as the executive molecules in the cells and therefore their research is fundamental in biotechnology and medicine.

Currently the most common method to investigate the activity, interactions, and functions of proteins on a large scale, is high-throughput mass spectrometry (MS). The mass spectrometers are used for measuring the molecule masses, or more specifically, their mass-to-charge ratios. Typically the proteins are digested into peptides and their masses are measured by mass spectrometry. The masses are matched against known sequences to acquire peptide identifications, and subsequently, the proteins from which the peptides were originated are quantified. The data that are gathered from these experiments contain a lot of noise, leading to loss of relevant information and even to wrong conclusions. The noise can be related, for example, to differences in the sample preparation or to technical limitations of the analysis equipment. In addition, assumptions regarding the data might be wrong or the chosen statistical methods might not be suitable. Taken together, these can lead to irreproducible results. Developing algorithms and computational tools to overcome the underlying issues is of most importance. Thus, this work aims to develop new computational tools to address these problems.

In this PhD Thesis, the performance of existing label-free proteomics methods are evaluated and new statistical data analysis methods are proposed. The tested methods include several widely used normalization methods, which are thoroughly evaluated using multiple *gold standard* datasets. Various statistical methods for differential expression analysis are also evaluated. Furthermore, new methods to calculate differential expression statistic are developed and their superior performance compared to the existing methods is shown using a wide set of metrics. The tools are published as open source software packages.

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietojenkäsittelytiede
SUOMI, TOMI: Data analysis tools for mass spectrometry proteomics
Väitöskirja, 113 s.
Matemaattis-tietotekninen tohtoriohjelma
Helmikuu 2021

TIIVISTELMÄ

Proteiinit ovat aminohappoketjuista muodostuvia isoja biomolekyylejä. Ne eroavat toisistaan aminohappojen järjestyksen osalta, mikä pääosin määräytyy proteiineja koodaavien geenien perusteella. Lisäksi proteiinit laskostuvat kolmiulotteisiksi rakenteiksi, jotka osaltaan määrittelevät niiden toimintaa. Koska proteiinit toimivat katalyytteinä biokemiallisissa reaktioissa, niillä katsotaan olevan keskeinen rooli soluissa ja siksi myös niiden tutkimusta pidetään tärkeänä.

Tällä hetkellä yleisin menetelmä laajamittaiseen proteiinien aktiivisuuden, interaktioiden sekä funktioiden tutkimiseen on suurikapasiteettinen massaspektrometria (MS). Massaspektrometreja käytetään mittaamaan molekyylien massoja – tai tarkemmin massan ja varauksen suhdetta. Tyypillisesti proteiinit hajotetaan peptideiksi massojen mittausta varten. Massaspektrometrillä havaittuja massoja verrataan tunnetuista proteiinisekvensseistä koottua tietokantaa vasten, jotta peptidit voidaan tunnistaa. Peptidien myötä myös proteiinit on mahdollista päätellä ja kvantitoida. Kokeissa kerätty data sisältää normaalisti runsaasti kohinaa, joka saattaa johtaa olennaisen tiedon hukkumiseen ja jopa pahimmillaan johtaa vääriin johtopäätöksiin. Tämä kohina voi johtua esimerkiksi näytteen käsittelystä johtuvista eroista tai mittalaitteiden teknisistä rajoitteista. Lisäksi olettamukset datan luonteesta saattavat olla virheellisiä tai käytetään datalle soveltumattomia tilastollisia malleja. Pahimmillaan tämä johtaa tilanteisiin, joissa tutkimuksen tuloksia ei pystytä toistamaan. Erilaisten laskennallisten työkalujen sekä algoritmien kehittäminen näiden ongelmien ehkäisemiseksi onkin ensiarvoisen tärkeää tutkimusten luotettavuuden kannalta. Tässä työssä keskitytäänkin sovelluksiin, joilla pyritään ratkaisemaan tällä osa-alueella ilmeneviä ongelmia.

Tutkimuksessa vertaillaan yleisesti käytössä olevia kvantitatiivisen proteomiikan ohjelmistoja ja yleisimpiä datan normalisointimenetelmiä, sekä kehitetään uusia datan analysointityökaluja. Menetelmien keskinäiset vertailut suoritetaan useiden sellaisten standardiaineistojen kanssa, joiden todellinen sisältö tiedetään. Tutkimuksessa vertaillaan lisäksi joukko tilastollisia menetelmiä näytteiden välisten erojen havaitsemiseen sekä kehitetään kokonaan uusia tehokkaita menetelmiä ja osoitetaan niiden parempi suorituskyky suhteessa aikaisempiin menetelmiin. Kaikki tutkimuksessa kehitetyt työkalut on julkaistu avoimen lähdekoodin sovelluksina.

Acknowledgements

First of all, I would like to express my gratitude to both of my supervisors, Prof. Olli Nevalainen and Prof. Jukka Heikkonen for their support and patience, insightful comments, and encouragement to complete my work. I also wish to express my special thanks to Prof. Laura Elo and all my co-workers at the Medical Bioinformatics Centre for their enthusiasm and valuable assistance throughout the years.

I want to specifically acknowledge my co-authors Tommi Välikangas, Fatemeh Seyednasrollah, Maria Jaakkola, Thomas Faux, Olli Kannaste, Garry Corthals, Jussi Salmi, and Esa Uusipaikka. The research could not have been successfully conducted without their active participation. I would also like to acknowledge Susumu Imanishi, Thaman Chand, Anni Vehmas, Anne Rokka, Arttu Heinonen and Tiina Pakula for providing and preparing part of the data and participating in fruitful discussions. My thanks also to Robert Moulder, Damien Coté, and Pirjo Nuutila for their insights.

Lastly, I would like to express my thanks to my parents, my family, and my friends who have given me constant support.

- Tomi Suomi

Table of Contents

Abbreviations

AQUA	Absolute quantification of proteins and their modification states
AUC	Area under the curve
BPCA	Bayesian principal component analysis
CPTAC	Clinical proteomic tumor analysis consortium
cRAP	Common repository of adventitious proteins
CV	Coefficient of variation
DDA	Data-dependent acquisition
DE	Differential expression
DIA	Data-independent acquisition
ESI	Electrospray ionization
FDR	False discovery rate
HOMD	Human oral microbiome database
HPC	High-performance computing
HPLC	High-pressure liquid chromatography
HPP	The human proteome project
ICAT	Isotope-coded affinity tag
IGC	Integrated gene catalog of human gut microbiome
iTRAQ	Isobaric tag for relative and absolute quantitation
k-NN	k-nearest neighbors
LC	Liquid chromatography
LLS	Local least squares

LOESS	Locally estimated scatterplot smoothing
LogFC	Logarithmic fold change
MALDI	Matrix-assisted laser desorption ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MS2	Tandem mass spectrometry
MSE	Mean squared errors
m/z	Mass to charge ratio
NSAF	Normalized spectral abundance factor
pAUC	Partial area under the curve
PCA	Pricipal component analysis
PCV	Pooled coefficient of variation
PECA	Peptide-level expression change averaging
PEV	Pooled estimate of variance
PLGEM	The power law global error model
PMAD	Pooled median absolute deviation
PRIDE	Proteomics identification database
RLR	Robust linear regression
ROC	Receiver operating characteristic
ROPECA	Reproducibility-optimized peptide-level expression change averaging
ROTS	Reproducibility-optimized test statistic
SD	Standard deviation
SGSD	Shotgun standard data
SILAC	Stable isotope labeling by amino acids in cell culture
SRM	Selected reaction monitoring

SVD	Singular value decomposition
SWATH-MS	Sequential window acquisition of all theoretical mass spectra
T2D	Type 2 diabetes mellitus
UPS	Universal proteomics standard
VSN	Variace stabilization normalization

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. Brief Bioinform. 2018 Jan 1;19(1):1-11. doi: 10.1093/bib/bbw095. PMID: 27694351; PMCID: PMC5862339.

II Kannaste O*, **Suomi T***, Salmi J, Uusipaikka E, Nevalainen O, Corthals GL. Cross-correlation of spectral count ranking to validate quantitative proteome measurements. J Proteome Res. 2014 Apr 4;13(4):1957-68. doi: 10.1021/pr401096z. Epub 2014 Mar 26. PMID: 24611565. (*shared first author)

III **Suomi T**, Seyednasrollah F, Jaakkola MK, Faux T, Elo LL. ROTS: An R package for reproducibility-optimized statistical testing. PLoS Comput Biol. 2017 May 25;13(5):e1005562. doi: 10.1371/journal.pcbi.1005562. PMID: 28542205; PMCID: PMC5470739.

IV **Suomi T**, Corthals GL, Nevalainen OS, Elo LL. Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins. J Proteome Res. 2015 Nov 6;14(11):4564-70. doi: 10.1021/acs.jproteome.5b00363. Epub 2015 Sep 29. PMID: 26380941.

V **Suomi T**, Elo LL. Enhanced differential expression statistics for data-independent acquisition proteomics. Sci Rep. 2017 Jul 19;7(1):5869. doi: 10.1038/s41598-017-05949-y. PMID: 28724900; PMCID: PMC5517573.

Publications not included to Thesis

The following publications are related to the topic of the Thesis, but are not included as part of it. The first two involve benchmarking of tools and methods for mass spectrometry proteomics data analysis, the next two development of algorithms for phosphoproteomics data analysis, and the last two are related to data-independent acquisition proteomics and metaproteomics.

1. Pursiheimo A, Vehmas AP, Afzal S, **Suomi T**, Chand T, Strauss L, Poutanen M, Rokka A, Corthals GL, Elo LL. Optimization of Statistical Methods Impact on Quantitative Proteomics Data. J Proteome Res. 2015 Oct 2;14(10):4118-26. doi: 10.1021/acs.jproteome.5b00183. Epub 2015 Sep 8. PMID: 26321463.

2. Välikangas T, **Suomi T**, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform. 2018 Nov 27;19(6):1344-1355. doi: 10.1093/bib/bbx054. PMID: 28575146; PMCID: PMC6291797.

3. Saraei S, **Suomi T**, Kauko O, Elo LL. Phosphonormalizer: an R package for normalization of MS-based label-free phosphoproteomics. Bioinformatics. 2018 Feb 15;34(4):693-694. doi: 10.1093/bioinformatics/btx573. PMID: 28968644.

4. Suni V, **Suomi T**, Tsubosaka T, Imanishi SY, Elo LL, Corthals GL. SimPhospho: a software tool enabling confident phosphosite assignment. Bioinformatics. 2018 Aug 1;34(15):2690-2692. doi: 10.1093/bioinformatics/bty151. PMID: 29596608; PMCID: PMC6061695.

5. Pietilä S, **Suomi T**, Aakko J, Elo LL. A Data Analysis Protocol for Quantitative Data-Independent Acquisition Proteomics. Methods Mol Biol. 2019;1871:455-465. doi: 10.1007/978-1-4939-8814-3_27. PMID: 30276755.

6. Aakko J, Pietilä S, **Suomi T**, Mahmoudian M, Toivonen R, Kouvonen P, Rokka A, Hänninen A, Elo LL. Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota - Implementation and Computational Analysis. J Proteome Res. 2020 Jan 3;19(1):432-436. doi: 10.1021/acs.jproteome.9b00606. Epub 2019 Dec 4. PMID: 31755272.

1 Introduction

Proteins are large molecules made of amino acid residue chains. They are generally considered as the executive molecules in the cells because many of them act as catalytes in biochemical reactions [1]. For example, proteins can act as carriers which move molecules around or they can be structural proteins for cellular components. The large-scale study of proteins is called *proteomics* [2]. Compared to genomics, proteomics is quite often considered as a more challenging field of research [3]. This is because the genome of an organism is commonly considered as being somewhat constant while the proteome (*i.e.* the set of expressed proteins) rapidly changes between individual cells and over time [3].

In principle, proteomics aims to understand the function of all proteins that are found from an organism [4]. It is an intriguing possibility to systematically identify and analyse all the proteins that are expressed by a cell or tissue. For example, the human proteome project (HPP) aims at mapping the entire human proteome [5]. It has been estimated that there are around 20 000 protein-coding genes in human and with currently available techniques, 90 % of these proteins are now mapped with credible evidence [6].

Proteomics can be done in multiple ways, but conventionally it is divided into two main paradigms: *top-down* and *bottom-up* proteomics [4]. In the top-down proteomics, the analysis is performed using intact proteins, and in the bottom-up paradigm, the proteins are first cleaved into smaller parts before they are analysed. These smaller sections of proteins are called peptides and they can be used for identification and quantification of the proteins [3].

Currently the most common method to investigate activity, interactions, and functions of proteins on large scale is high-throughput mass spectrometry (MS) [7; 8]. Quantitative MS methods are used to analyse protein expressions as a function of cellular state [7]. They have been used, for instance, to gain information about the molecular composition, regulation, and pathways in various types of samples [9]. They are also often used in search of clinical biomarkers that are indicators of biological or pathogenic processes or responses to therapeutic interventions [10]. In drug discovery, mass spectrometry can be used to help designing compounds that interfere with protein functions [11]. Overall, accumulating knowledge of molecular processes has a critical role on our understanding, diagnosis, and treatment of various diseases.

Although mass spectrometry can be used to detect and quantify thousands of proteins in a sample, there remain multiple challenges in the analysis of MS data. One major challenge is that the results from MS analyses are susceptible to biases [12]. These can originate for example from instrument calibration, sample preparation, or temperature changes [13]. The exact reason is usually unknown so it cannot be compensated by adjusting the experimental settings [12; 13]. Therefore, normalization methods are used to make the acquired data comparable [12]. However, it is typically not obvious which normalization method should be used. To this end, in **Publication I**, a set of widely used normalization methods were benchmarked in the context of mass spectrometry proteomics.

Another important aspect in MS data analysis is peptide quantification. It can be done either at the precursor ion or at the fragment ion level [4]. At precursor level, the intensities of the precursor ions correspond to the peptide abundance [14]. However, the peptides must still be simultaneously fragmented and identified. It has been suggested that the number of identified peptides (*i.e. spectral counting*) could be used directly as measure of peptide quantification, but it is known to have poor signal-to-noise ratio [15; 16]. As this would allow a much simplified workflow, the suitability of spectral counts for differential expression analysis was assessed in **Publication II**.

Detection of differentially expressed proteins is one of the most common tasks in many MS proteomics studies. To find statistical differences between sample groups, differential expression analysis is performed on the quantified protein data [17]. For this purpose, the standard method has been the Student's *t*-test, although it might not always be an optimal solution [18; 19]. Similarly as with normalization, the selection of the most suitable method is not straightforward. To enable data-driven selection of the statistic, a reproducibility-optimized test statistic was implemented and published as an R package in **Publication III**. It adjusts a modified *t*-statistic based on the underlying data and it has been successfully applied to various omics studies, including proteomics [20].

A special feature of *bottom-up* proteomics is that the measurements are made at the peptide-level, but in most cases they are summarized to protein-level for further analysis [21]. While this commonly used approach was used also in Publication III, it has its limitations, as peptides from the same protein can behave differently. To take this into account, a differential expression analysis method leveraging all the peptide-level data was introduced and published as an R package in **Publication IV**. Finally, the emerging data-independent acquisition proteomics has made it feasible to combine the reproducibility optimization procedure with the peptide-level data. Such approach was developed and published in **Publication V**.

In conclusion, this work evaluates common steps of mass spectrometry proteomics data analysis and provides suggestions for normalization and differential expression analysis. Many of the insights that were gained throughout this work are

applicable also more generally to omics data beyond mass spectrometry proteomics. For instance, the data-driven approach for differential expression analysis is widely applicable to other omics data. The developed software tools that were published as open source allow researchers to apply the new methods in their research.

1.1 Aims of the Thesis

The overall aim of this Thesis is to improve existing and develop new computational tools to address research problems in the field of biotechnology. The data acquired from biological samples contain a lot of noise, which leads to loss of relevant information. A part of this noise is related to differences in the sample preparation, whereas another part of it is related to technical limitations of the analysis equipment. Because in many cases there are very limited chances to prevent the occurrence of noise, developing algorithms and computational tools to overcome the underlying issues is of most importance. For example, the results that are based on label-free mass-spectrometry proteomics are commonly used for further validation studies, which can be expensive and time consuming. Therefore, it is important to produce as accurate candidate lists as possible from this *discovery* phase using effective computational tools. The resulting findings have the potential to, for example, increase understanding of complex biological systems including disease processes, to be used as predictive or diagnostic biomarkers, or to aid in drug development.

To address these needs, the specific aims of this Thesis work are as follows:

1. Investigate and evaluate existing computational methods for processing label-free mass-spectrometry based proteomics data.

2. Improve existing tools and develop new statistical approaches for differential expression analysis of mass spectrometry proteomics.

3. Implement and publish the new methods as easy to use software packages.

Publications I, and **II** cover the investigation of existing methodologies and **Publications III**, **IV** and **V** describe new statistical approaches that are made available as software packages written in R.

1.2 Structure of the Thesis

The basics of mass-spectrometry proteomics are shortly explained in Chapter 2. Software tools to perform quantification, methods for data normalization, and various approaches for differential expression analysis are described in Chapter 3. The datasets used in the present work are described in Chapter 4 and the main results of the related

publications are given in Chapter 5. Reprints of these articles are given at the end of the Thesis.

2 Mass spectrometry in proteomics

Mass spectrometers are used to quantify in high-throughput complex protein mixtures that have been generated in biological studies (**Figure 1**). They measure the masses of molecules, or more specifically, their mass to charge (m/z) ratios and produce mass spectra of the samples. In principle, a mass spectrometer consists of three main parts: an ion source, a magnetic field to accelerate the particles, and a detector. Electromagnetic fields are used to move the molecules, which means that the molecules must be first ionized [22].

A widely used method to obtain an overall protein profile of a sample, is the so-called label-free *shotgun* proteomics [8]. While the present work focuses on the label-free proteomics, there are labeled techniques, where the samples are labeled for example using stable isotopes [23]. The following sections describe the typical steps of a mass spectrometry experiment.

2.1 Protein digestion and peptide separation

An important part of the protein identification process is the protein digestion (*i.e.* the cleaving of proteins into peptides). The enzymes that perform digestion are called proteases. For mass spectrometry experiments, proteins are commonly cleaved enzymatically by trypsin [7].

For peptide separation, a commonly used technique is chromatography. It includes a family of techniques which separate mixtures into their individual components. In mass spectrometry proteomics, integrated liquid-chromatography (LC-MS) systems are often preferred for complex samples [7]. This is in comparison to

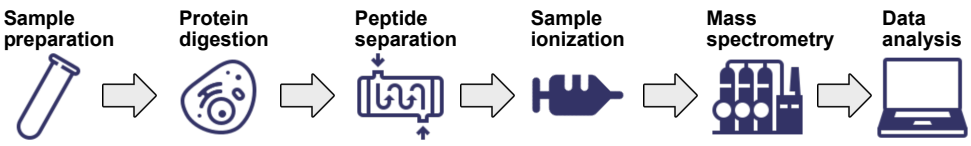


Figure 1. Workflow of a typical mass spectrometry experiment. Proteins are first digested into peptides, then separated using chromatography and further ionized for mass spectrometry. After mass spectrometry, the peptides are identified computationally and the corresponding protein abundances are estimated. Finally, the data can be analyzed using various computational tools. This study focuses on the method development.

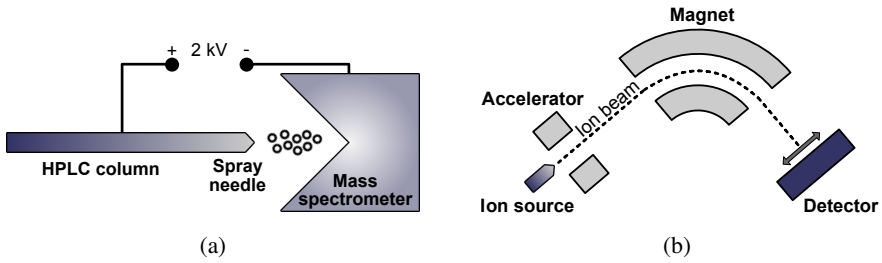


Figure 2. Principles of mass spectrometry. (a) Inionization of separated peptides by the electrospray ionization. (b) Measurement of mass-to-charge ratios of the charged ions.

MALDI based systems, where a laser is used to ionize dry crystalline samples, while liquid based separation techniques allow direct coupling to a mass spectrometer via electrospray ionization (ESI) [7]. Because of the high-pressure pumps that are used, the process is referred to as high-pressure liquid chromatography (HPLC).

2.2 Ionization and mass spectrometry

In mass spectrometry proteomics, the electrospray ionization [24] is primarily used. In principle, a liquid flow carries the sample material to the ionization source (**Figure 2a**). The liquid is sprayed from a needle with several kV of electrical potential between the needle and mass spectrometer [8].

Peptide ions enter mass spectrometer and are guided and manipulated by electric fields in a vacuum system. In the simplest form, they are first accelerated and then their flight path is altered [8], which can be measured by a particle detector (**Figure 2b**). There are different variations of these instruments, but each of them generate mass spectrums that report the signal intensities of the ions at mass-to-charge (m/z) scale [22].

2.3 Data acquisition

In shotgun proteomics the MS instrument is commonly operated in the so-called *data-dependent acquisition* (DDA) mode, where the machine selects and isolates intense precursor ions and fragments them to produce a secondary spectra (*i.e.* tandem mass spectra, MS/MS) [4]. To get peptide identifications, the acquired spectra are matched to a database of peptide sequences. Shotgun proteomics suffers from low reproducibility because of under-sampling and because MS/MS spectra are quite often taken outside the elution peak (*i.e.* when most of the peptide to be identified would be passing out of the chromatography column). Therefore, only a proportion of peptides that are detected are identified reliably in all samples [25]. The so-called *targeted mass spectrometry* that includes for example selected reaction monitoring

(SRM) [26], complements shotgun proteomics. In targeted SRM analysis only such molecular ions that match the mass of the targeted peptide are selected and subsequently fragmented. By counting the precursor-fragment ion pairs (*i.e.* SRM transitions), the targeted peptides are quantified [26; 27].

A more recent technology, the *data-independent acquisition* (DIA) essentially combines the volume of protein identifications in shotgun proteomics and the reproducibility of selected reaction monitoring. Multiple different DIA methods have been published, such as MS^E [28], XDIA [29], SWATH [30], MSX [31] or UDMS^E [32], each having their unique properties. Regardless, they all collect MS/MS scans *independently* without precursor information during the acquisition process. Because of the dissociation between precursor and its fragments, processing of the acquired data is more challenging compared to the more standard data-dependent acquisition. While the research has been focused on preprocessing and protein quantification [33; 34; 35; 36], much less attention has been paid on the statistical analysis of DIA data. Only few tools claim to take into account the properties of DIA data, such as MSstats [37] and mapDIA [38].

2.4 Peptide identification

When the peptides are fragmented and analyzed by mass spectrometry, their exact masses are acquired. The masses are then compared against theoretical masses produced computationally (*in silico*) from a reference database. The aim is to find such sequences from a given database that explain the experimental data. Protein sequences in a reference database are digested *in silico* by scanning the sequence for possible cleavage sites and calculating the exact masses for the peptides [7; 39; 40]. In practice, an algorithm performs the comparison between the theoretical and acquired peaks. One of the first such algorithms was SEQUEST [41], which calculates cross-correlation similarity between mass-to-charge ratios of the observed tandem mass spectrum fragment ions and the ratios predicted from the reference database. There are several open source tools and commercial programs available that are designed for this purpose, including for example Mascot [42], OMSSA [43], TANDEM [44], Andromeda [45], Comet [46], and MS-GF+ [47].

An up-to-date reference of protein sequences for a selected organism needs to be obtained before performing the search. One of the most commonly used database is UniProt [48] which contains both manually curated protein entries of SwissProt [49] and automatically added unreviewed protein entries of TrEMBL [50] for a number of organisms. There are also other similar databases *e.g.* RefSeq [51] or Ensembl [52]. For specific use cases, there is a collection of reference databases available, *e.g.* UniPept [53] to find unique tryptic peptides in metaproteomics, IGC [54] for human gut metaproteome, or HOMD [55] for human oral microbiome. Similarly, there are databases like cRAP for common contaminants in proteomics experiments, but these

can also come as parts of the search engines. Finally, instead of using a database, *proteogenomics* can provide a very specific reference that is generated directly from DNA- or RNA-sequencing data of the same set of samples [56]. In general, it is critical to have the expected sequences in the search space and narrowing down the possible candidates helps in achieving reliable identifications.

Peptide identification without relying on a database of protein sequences at all is possible using *de novo* sequencing methods, *e.g.* PEAKS [57], pNovo [58], or Novor [59]. These approaches allow the identification of peptides that are not in any database or are otherwise unknown. However, deriving a sequence from a fragment mass spectrum relies on the spectral quality, mass accuracy, and resolution of the instrument [60].

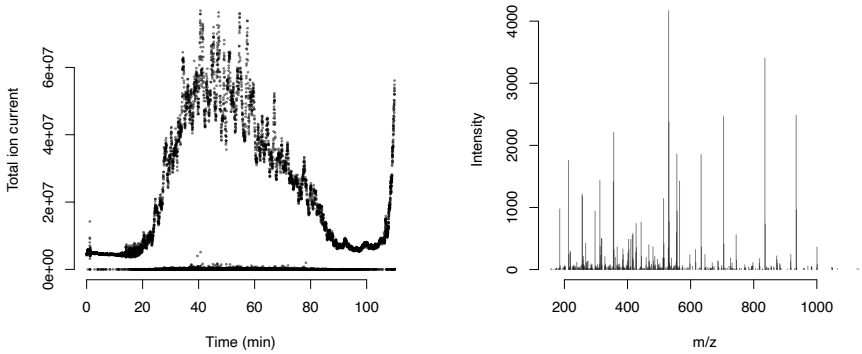
One measurement for technical progress of mass spectrometers is the number of proteins that can be identified in a study [7], although a dependence between the experiment length (*i.e.* chromatography gradient length) and the number of identified proteins has been observed [61]. There are often thousands of identified proteins in an experiment and the numbers are slowly increasing as the technology progresses. Despite the increasing numbers, around 75 % of the acquired spectra still remain unidentified [62], which is mainly because of poor signal-to-noise ratio, incomplete reference databases, and post-translational modifications that are not expected [63].

While mass spectrometry proteomics can be considered as a technology that can probe the majority of the proteins in a sample [7], there is also a major gap between the concentrations of the most abundant molecules and those existing as trace amounts. Current mass spectrometry can detect differences of around five orders of magnitude while protein concentrations can span 12 orders of magnitude [63]. There are ways to overcome this issue, for example by depleting most abundant proteins with antibodies before analysis to help in the detection of low abundance proteins [64]. However, analysis of complex mixtures remains uncomprehensive and if a particular peptide is not identified from a sample, it does not indicate that it was not originally present [7].

2.5 Quantification

Quantitative analysis in mass spectrometry proteomics can be done by labeling samples chemically (ICAT, iTRAQ) [65; 66] or metabolically in a cell culture (SILAC) [23]. These methods are commonly used for relative quantification, but absolute quantification can be achieved by *spiking* the sample with synthetic peptides to which the unlabelled peptides are compared (AQUA) [67].

Due to the cost and extra work of isotope labelling methods, various label-free methods have emerged. The quantification can be done either at the precursor ion (MS1) or fragment ion (MS2) level. On MS1-level the intensities for precursor ions are measured over time (**Figure 3a**). The intensity at a particular time corresponds to



(a) (b)

Figure 3. Example visualizations from raw data of an MS experiment. (a) Total ion current over time. (b) Tandem mass (MS/MS) spectrum from a peptide.

the peptide abundance [14], but the corresponding peptides must simultaneously be identified on MS2-level (**Figure 3b**). Alternatively, the number of identified peptide spectra can directly be used in the so-called *spectral counting* method. It is obviously more straightforward to implement, but is known to have poor signal-to-noise ratio [15; 16]. One of the suggested improvements to spectral counting has been using cumulative fragment signal intensities per protein [68; 69].

3 Data analysis

After acquiring the estimated protein expression levels from the quantification software (commonly provided in a matrix format), the data analysis can be done. This usually includes *quality control* of the samples, data *normalization* to make the samples comparable, sometimes complemented by *imputation* of missing values, and finally *differential expression analysis* to estimate which proteins have different expression levels between the compared sample groups (**Figure 4**). These steps are described in more detail in the next subsections.

3.1 Quality control

The first step of computational analysis involves quality control of the acquired data. This includes checking the distribution of expressions, sample correlations, overall clustering of the samples, and principal component analysis (PCA). These are done in order to pinpoint any deviating samples, which might affect the outcome and should possibly be removed from the analysis.

The expression value distributions from each sample are commonly visualized by box plots (**Figure 5a**) or violin plots. Box plots visualize numerical data by their quartiles and they have extending lines (*whiskers*) that indicate variability outside the upper and lower quartiles of the data. Median value is shown as a band in the middle of the diagram. Violin plots are similar but they also show the probability density of the data.

Correlations between samples describe the similarity between the samples on a general level, when all values are taken into consideration. Commonly used options are the *Pearson's correlation coefficient* for linear relationships between the values and *Spearman's rank correlation* to assesses ranking between the values. With mul-

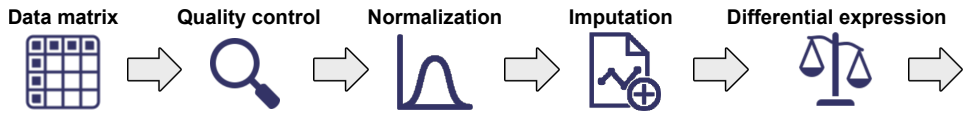


Figure 4. Workflow of data analysis in mass spectrometry proteomics experiments. This study involves all steps through evaluating different methods and their combinations. New tools that are published deal with the differential expression analysis.

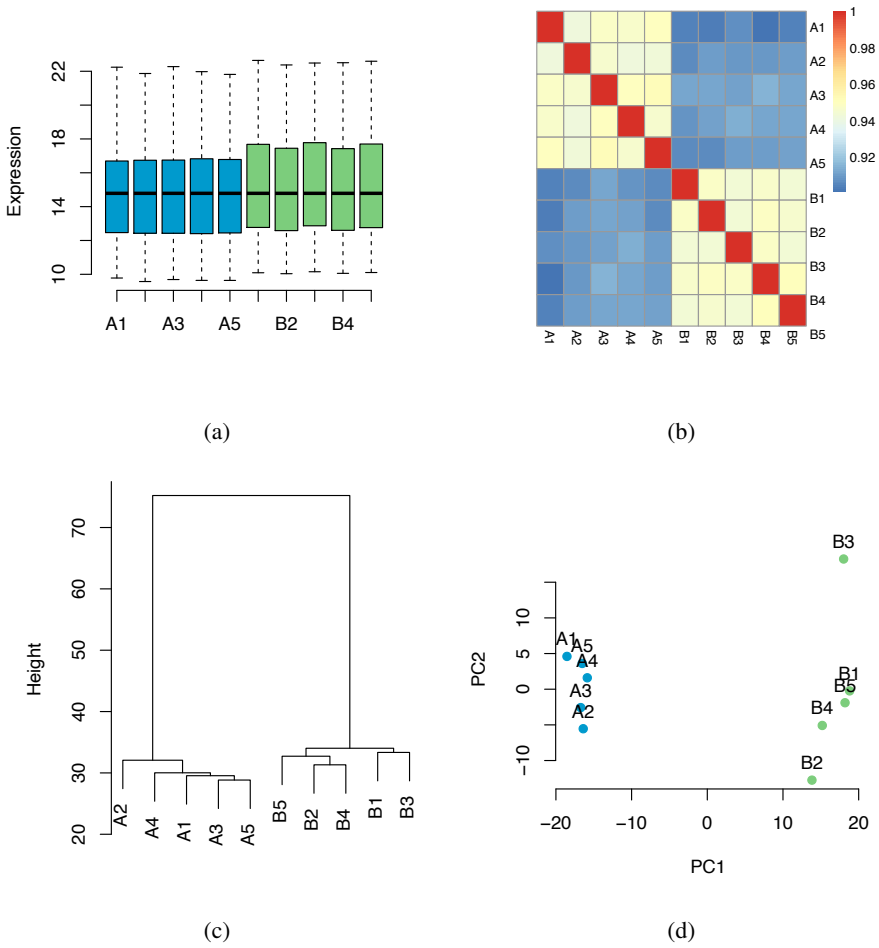


Figure 5. Examples of quality control visualizations for simulated high-throughput mass spectrometry proteomics data with five replicates in two groups. (a) Box plot. (b) Correlation heat map. (c) Similarity dendgrogram using Euclidean metrics. (d) Principal component analysis.

tiple samples, the correlation is often visualized as a heat map (**Figure 5b**).

In hierarchical clustering the samples are grouped together according to their similarity metric, which is often either the sample correlation or Euclidean distance. The similarity is visualised as a *dendrogram*; a graph where the most similar samples end up nearest to one another (**Figure 5c**). There are multiple options in agglomerative clustering that can be used to select in which order the nodes become connected together. Additionally, hierarchical clustering is often used together with heat maps to re-order the image.

Principal component analysis (PCA) is also commonly employed to visualize the data (**Figure 5d**). It is a procedure that uses transformation to convert a set of values to principal components, which are linear combinations of the initial variables. The principal components are supposed to account as much of the variability in the data as possible, in descending order [70]. This means that often only the very first components are needed to represent the data.

3.2 Normalization

Mass spectrometry based proteomics has gone through very rapid development in recent years. Its current aims include identification and quantification of proteins as accurately as possible [71]. Mass spectrometry workflows can detect thousands of proteins and their modifications [72], but the results from the MS analyses are still susceptible to biases [12]. These biases are caused, for example, by differences in sample preparation, changes in sample temperature, instrument calibration, or depend on the measured protein abundances [13]. The exact reason of a bias is usually unknown which means that it cannot be compensated by adjusting the experimental settings [12; 13].

Normalization aims to make the acquired data of the samples more comparable [12]. Many of the normalization methods have originally been developed for the microarray technology [13], but several of them can be used similarly for proteomics data as well. Because in mass spectrometry the analysis of multiple samples one by one takes usually long time and the performance of the instrumentation changes, it has also been suggested that linear regression normalization that accounts for run order of the samples is useful [73].

There are extensive reviews inspecting various normalization methods [74; 75; 76; 77]. For example, Bolstad et al. [74] and Choe et al. [75] found no major differences between the methods, while Callister et al. [78] claimed linear regression normalization as the best. In these studies the different normalization methods were evaluated by variance across the replicates, looking at the absolute distance between a LOESS curve and the x axis of an MA plot, comparing observed fold changes against known changes, or by applying a *t*-test between groups of samples and evaluating the results. Many of these approaches are also used in the present research.

Large and systematic comparisons of the normalization methods in context of proteomics have been lacking. Evaluation using multiple data sets and differential expression analysis has not been systematically performed yet. In addition, the accuracy of logarithmic fold change (LogFC) estimates (*i.e.* accuracy of difference in expression level between groups) after normalization has not been systematically investigated before on proteomics data. To address this, a thorough comparison of popular normalization methods and their variants was performed in the present study (**Publication I**). To thoroughly benchmark the normalization methods, three label-free proteomics data sets were used. They include known concentrations of specific spike-in proteins so they are suitable for testing because the differences between sample groups are known, and therefore one can evaluate the ability of the methods to find the true signal in the data. In addition, a mouse study was included to reflect a typical research setting.

Inherently, the various normalization approaches perform differently depending on the data, making the selection of the most suitable method hard. To help with that, different tools have been proposed, including *SPANS* [79] and *Normalyzer* [12]. They include a large set of different normalization methods together with frequently used evaluation metrics to aid in the selection. These include the median coefficient of variation (CV) [78], median standard deviation (SD) [73], pooled estimate of variance (PEV) [12; 78; 73], pooled coefficient of variation (PCV) [12], and pooled median absolute deviation (PMAD) [12].

The next subsections outline the ideas of the most popular normalization techniques which were evaluated.

3.2.1 Quantile normalization

Quantile normalization [80] is a crude technique to make statistical properties of two distributions similar. In the quantile normalization procedure, a reference distribution is used to change the properties of another distribution of equal length. The process is based on ranking the values of both distributions, then replacing the original values with the values of the reference distribution having the same rank. Thus, the highest ranked value in the data gets the value of the highest ranked value in the reference distribution, the next highest value the next etc. Overall, the distribution becomes a permutation of the reference distribution according to ranks. This approach has been common in microarray studies.

3.2.2 Median normalization

Median normalization assumes that the samples, and more specifically their intensity values, are separated by a constant. To counteract for this effect, median normalization simply scales all the samples to have identical median values. Here, the median

normalization was implemented using Normalyzer [12].

3.2.3 Linear regression normalization

The underlying assumption with linear regression normalization is that the bias in the data is linearly dependent on the protein intensities (*i.e.* as the intensity increases, so does the bias) [78]. We explored the robust linear regression (RLR) and its variants RlrMA and RlrMACyc. The first method uses the median values of each protein as its reference to which all the samples are normalized against. The RlrMA is similar, but the data are MA transformed (*i.e.* converted to fold change vs average -axes) before normalization. In the RlrMACyc, there is no common reference (such as median) anymore, but instead the MA transformations and the normalizations are done pairwise between all pairs of samples in cyclic fashion. The standard RLR normalization was implemented using Normalyzer [12] and the variants using the R package MASS [81].

3.2.4 Local regression normalization

The local regression normalization technique extends linear regression normalization. Unlike a simple linear regression, it assumes that the relationship between the intensity and possible bias is not linear [78]. Here, two variants of locally estimated scatterplot smoothing (LOESS) were explored. The first method utilizes MA transformed data where mean of all samples is used as a reference. The second variant (LoessCyc) selects only two samples at a time for MA transformation and normalization, but the process is iterated multiple times over all the samples. Both variants were implemented with the limma R-package [82].

3.2.5 Variance stabilization normalization

The underlying assumption with variance stabilization normalization (VSN) is that the variance in the data is dependent on the overall mean (*i.e.* proteins with higher intensity have higher variance). Thus, this approach aims to make the sample variances independent from their corresponding mean intensities. This is achieved by parametric transformations and maximum likelihood estimation [83]. Here, the method was implemented using the Bioconductor R-package Vsn [83].

3.3 Imputation

Mass spectrometry based proteomics commonly suffers from missing values. Essentially, the missing values are a result of peptide peaks that are not recorded in a sample during MS measurement [13]. Such missing values also propagate to protein-

level when the values are eventually summarized. Overall, the missing values are roughly divided to two main categories. They are either abundance-dependent (*i.e.* stemming from instrument limitations) or the values are missing completely at random.

Besides proteomics, imputation has gained a lot of attention in single-cell RNA-sequencing, where the acquired data is even more sparse. An evaluation of 18 imputation methods [84] suggested that scVI [85], DCA [86], and MAGIC [87] perform well.

While benchmarking the various normalization methods and differential expression statistics in this study, we have also investigated the effects of imputation in the context of mass spectrometry proteomics data [88]. The next sections outline the various imputation methods that were used.

3.3.1 Zero imputation

First, a simplistic method of zero imputation was tested, where all missing values were replaced with zeros. This is more of a technical operation that allows the statistical methods to calculate differential expression statistic in situations with too many missing values.

3.3.2 Background imputation

Background imputation is another simple approach that uses the lowest intensity value of the whole data set to represent the background. This method simulates a situation where the lowest detected value is the technical limitation of the instrument and it is used as such to represent all the missing values.

3.3.3 Censored imputation

In censored imputation, only such proteins that have more than one missing value per sample group are considered as missing because of technical limitations. The lowest intensity in the whole data is used for imputation, similarly as with the background imputation. In cases where only a single value is missing, it is considered *missing completely at random*, so no imputation is performed.

3.3.4 K-nearest neighbor imputation

The k-nearest neighbor (k-NN) imputation works by finding the k most similar proteins from other samples than the one being imputed for. This is commonly measured by Euclidean metrics. The average of the k proteins is used to estimate the missing value. The model can also be weighted so that more similar proteins affect the es-

timation more [89]. A number of k between 10 and 20 has been suggested to be appropriate [89].

3.3.5 Bayesian principal component analysis imputation

The Bayesian principal component analysis (BPCA) imputation combines principal component regression and Bayesian estimation to calculate the expected values for the missing data [90]. If a standard principal component analysis is performed on the data, the resulting visualizations should resemble each other before and after the imputation procedure. Our implementation was based on the pcaMethods R package [91].

3.3.6 Local least squares imputation

The the local least squares (LLS) imputation uses the standard least squares regression to estimate the missing values. A set of k most similar proteins from other samples are used for the regression mode. In most cases a k value of 150 is considered enough [92; 93]. Our LLS imputation was based on the pcaMethods R package [91].

3.3.7 Singular value decomposition imputation

Singular value decomposition (SVD) obtains mutually orthogonal expression patterns that allows approximation of all values in the data via linear combinations [94]. These patterns are identical to principal components and called as eigenproteins. The missing values are estimated by regressing the protein against the k most significant eigenproteins, then using the coefficients to reconstruct the original values [94]. It has been observed that k often needs to be only 20 % of all eigenproteins [89], which was also used in this study. Because singular value decomposition works only with complete matrices, the missing values are first replaced for example by mean values and the imputation procedure is repeated until convergence [94].

3.4 Differential expression analysis

Differential expression analysis is performed on the data to find statistical differences between sample groups. Such analysis is often performed on different omics data, where the detected features are for example gene or protein expressions. One of the most common method has been the Student's *t*-test, but is not always an optimal solution [18; 19]. While many alternative methods are available [95; 96; 82], it seems that different statistics work well in different datasets [97; 98; 99; 100]. New statistical approaches are constantly being developed but unfortunately there is no

consensus on how to select an appropriate method *a priori*. Therefore, currently the aim is to select a method that is shown to perform consistently well in similar data as the one being analyzed.

3.4.1 Spectral counting methods

Counting the number of identified peptides (*i.e. spectral counting*) has been used as an alternative way to quantify proteins instead of using MS1 intensity. This is considered easier approach and numerous methods have already been suggested for differential expression analysis of spectral count data [16; 101; 102; 103; 104; 105; 106; 107; 108]. Some of these have been developed specifically to handle spectral count data, while others are adaptations of previously published microarray data analysis tools.

Sometimes the spectral counts are additionally normalized by protein lengths and divided by the sum of all counts in a sample. This produces normalized spectral abundance factors (NSAF) [109] that are suggested to have similar properties as microarray data [102]. This allows to directly use many of the methods designed for microarray data. Other approaches consider directly the properties of spectral counts. For example, QSpec [16] uses Poisson distribution to model the counts, while others rely on beta-binomial distribution [103].

Commonly the methods are evaluated using simulated or spike-in data [16] or relying on specific knowledge about the biology [101]. In **Publication II**, we also considered the overall similarity of the spectral counting methods when the same input data was given.

3.4.2 ROTS

The main idea of reproducibility-optimized test statistic (ROTS) is to robustly select a dataset specific statistic. This is achieved by performing a range of modified t-statistics and selecting the one that produces highest overlap of top features over group-preserving permuted datasets. The method has been applied in various contexts, including microarrays [110], mass spectrometry proteomics [20], bulk RNA-seq [111], as well as single-cell RNA-seq [100]. While the performance of the method has found to be good, it is also practical to select a statistic that optimizes the top ranked features. After all, it is often the case that only a few of the top candidates are selected for further validation.

In **Publication III** we introduced an R package to perform ROTS analysis and demonstrated its usage using three different case studies. At the same time, multiple visualization options (**Figure 6**) were included to the package for convenience. Finally, the method has been published in Bioconductor where it has remained around the top 20 % of most downloaded packages since.

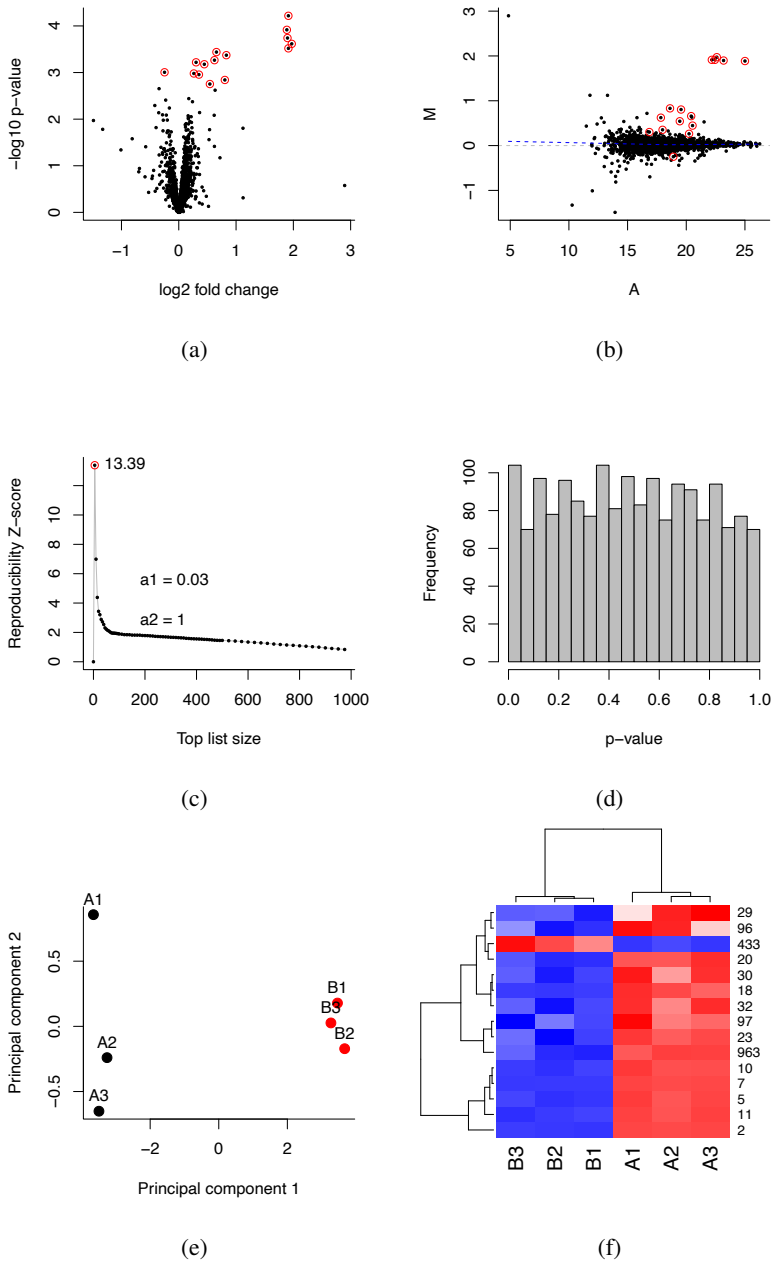


Figure 6. Visualizations provided by the ROTS R-package include (a) volcano plot, (b) MA plot, (c) reproducibility plot, (d) histogram of *p*-values. (e) PCA plot, and (f) heatmap of differentially expressed features.

3.4.3 PECA

In *bottom-up* proteomics the measurements are made at the peptide-level, but in most cases proteins are the desired units to quantify and, for example, measure differential expression. Thus, intensities are eventually combined to protein-level using for example mean [112], sum [113], or linear models (*e.g.* [114; 115; 116]). Overall, estimation of protein abundances and differential expression analysis are challenging because even peptides that originate from the same protein can behave differently. However, some peptide-centric methods have been proposed [117].

In **Publication IV** we propose a peptide-level expression change averaging method (PECA) to determine differential protein expression using directly the peptide-level measurements. By combining multiple statistics from the same protein (*i.e.* peptides), we are able to improve the differential expression analysis in MS-based proteomics studies. Briefly, the differential expression statistic is calculated for each measured peptide and the *p*-value of median peptide-level statistic is used for scoring. Because under the null hypothesis *p*-values follow uniform distribution $U(0, 1)$ and the order statistics from it follow a beta distribution, the overall significance of the i^{th} peptide (here median) can be estimated via beta distribution's probability density function [118]. This method is completely different from tools like InfernoRDN, which only provide different roll-up methods to summarize peptide abundances before statistical testing. [119].

The method tailored for mass spectrometry proteomics was implemented as part of an R package and published in the Bioconductor repository.

3.4.4 ROPECA

In **Publication V**, we develop the idea even further, and introduce a new reproducibility-optimized peptide change averaging method (ROPECA) to perform differential expression analysis on mass spectrometry proteomics data. This method is made possible by the emerging data-independent acquisition (DIA) proteomics technology, which in a way combines both shotgun proteomics and targeted proteomics [30; 31; 32].

In data-independent acquisition the MS/MS scans are gathered systematically throughout the process, but it happens without clear association between the precursor and its fragments. This makes the identification of the peptides more difficult, and has made the processing much more challenging. Methods and other software for data-independent acquisition data have only recently been published; most of the progress has been in pre-processing of the data and quantification of the proteins [120; 35; 36]. For statistical analysis of the data obtained using DIA, there has not been many improvements. While there are few exceptions, including MSstats [37] and mapDIA [38], the data analysis methods are quite limited and there is still room

for improvement.

The new method uses all peptide measurements in estimating the overall protein-level changes, much like the previously developed PECA approach. However, it also considers the reproducibility of the top ranked features in a way ROTS does. Combining two of the best performing approaches is made possible by DIA proteomics, where systematic data collection produces much less missing values on the peptide-level. While the method can also be used with traditional shotgun proteomics data, its performance might suffer from missing values.

The method has been included as part of R package PECA, which has been published in the Bioconductor repository.

4 Datasets

Advances in high-throughput technologies such as mass spectrometry proteomics, has led to a massive accumulation of data. Today, most of the mass spectrometry proteomics datasets are publicly available from repositories such as PeptideAtlas [121] or PRIDE [122]. This opens up the possibility of mining and reusing the data [123]. The following sections describe the datasets used in this Thesis.

4.1 Technical datasets

The UPS1 data we use for benchmarking, was originally published as part of our comparison study [20], where a number of statistical methods were assessed in the context of mass spectrometry proteomics. The data includes 48 proteins of the universal proteomics standard set (UPS1) that are spiked into a yeast proteome in different amounts. Concentrations of 2, 4, 10, 25 and 50fmol/μL were created in three technical replicates. The data is publicly available from the PRIDE Archive (id PXD002099) and it has been used in **Publications I**, **II**, and **IV**.

The CPTAC data [124; 125; 126] also contains 48 universal proteomics standard set (UPS1) proteins that are spiked into a yeast proteome in varying amounts. It has spike-in concentrations of 0.25, 0.74, 2.2, 6.7 and 20fmol/μL with three technical replicates. Originally, the data was sent to multiple laboratories for independent analysis to assess the current state-of-the-art protocols. All data are available from the CPTAC-portal, from where we processed the mass spectrometry data acquired at one of the anonymized test sites (Orbitrap instrument located at site 86). The data has been used in **Publications I** and **III**.

The shotgun standard data (SGSD) of Bruderer et al. [25] has 12 spike-in proteins that have been added to a human cell line background for benchmarking purposes. In total, it has eight sample groups with different spike-in concentrations, each with three technical replicates. One of the eight sample groups has a large relative spike-in concentration compared to the background. Because we have previously observed that this might lead to issues at least with DDA data [20], we decided to exclude the sample group from our analysis. The data are publicly available from PeptideAtlas (id PASS00589) and has been used in **Publication I**. The same benchmark data, but acquired in data-independent acquisition mode and processed using Spectronaut [127], was used in **Publication V**.

4.2 Biological datasets

The mouse data of Vehmas et al. [128] contains liver samples of five transgenic mice and seven wild-type mice. It is publicly available from the the PRIDE Archive (id PXD002025) and it was used in **Publication I** to assess the similarity of normalization methods in biological context.

To test the similarity of spectral counting methods in **Publication II**, three different biological datasets were used. A rat data containing brain tissues of three epileptogenic and three control rats, wound healing data from six pigs, and finally, a yeast data of Pavelka et al. [102] containing eight samples from different phases of cell growth.

A hybrid proteome data from Kuharev et al. [129], originally used for testing different quantification tools, is used here to asses differential expression analysis. The data is a mixture of three different organisms (human, yeast and *E. coli*) and it has been prepared in such a way that there are two different sample groups, each with five technical replicates. The proportion of human proteome is kept constant while yeast and *E. coli* change between the groups. Roughly 35 % of the total proteins are differentially expressed, which should reflect actual biological samples more than spike-in data [129]. The data are publicly available from the the PRIDE Archive (id PXD001240) and it was used in **Publication V**.

A longitudinal human twin study by Liu et al. [130] contains data from 72 monozygotic and 44 dizygotic twins from two different time points. It has plasma proteins quantified in data-independent acquisition mode using the SWATH technique [131]. The data is publicly available from the the PRIDE Archive (id PXD001064) and is used in **Publication V** to evaluate our new method using biological data.

5 Results

5.1 Benchmark of existing methods

The overall proteomics workflow includes multiple steps starting from peptide identification and their quantification, both having a number of competing algorithms and methods. To this end, we have evaluated a number common of software workflows that perform these tasks [88]. Briefly, we benchmarked the freely available MaxQuant, OpenMS, Proteios, and the commercially available Peaks and Progenesis softwares. We found Progenesis to perform well both in terms of differential expression analysis results and the amount of missing values [88]. The commonly used MaxQuant software performed best in estimating the fold changes and the differential expression analysis, but only after filtering the data [88]. Overall, the missing values produced by the different workflows reduced their performance, but we found that imputation or filtering could be used to mitigate the issue [88].

Subsequent steps of the proteomics workflow include normalization of the data and statistical analysis. The following sections summarize the comparisons of existing normalization methods, and differential expression statistics (**Publications I** and **II**).

5.1.1 Comparison of normalization methods

In addition to the different software, the effects of various normalization methods were explored. We found that normalization decreased variation between technical replicates with all the tested data (**Figure 7**). Especially the variace stabilization normalization (Vsn) decreased pooled median absolute deviation (PMAD) more than other methods (Wilcoxon signed rank test $p < 0.05$), except EigenMS in the CPTAC data. Other measures were the pooled coefficient of variation (PCV) and the pooled estimate of variance (PEV), both leading to the same conclusion. Similarity of technical replicates using Pearson correlation was also highest with Vsn in all tested spike-in datasets (**Figure 8**).

Besides the differential expression analysis, the accuracy of the fold change estimates of the protein intensities were evaluated. There were differences in the accuracy of the fold changes by the different software workflows, but MaxQuant consistently produced low mean squared errors (MSE) between observed and the expected changes. A detailed investigation was performed on the different normal-

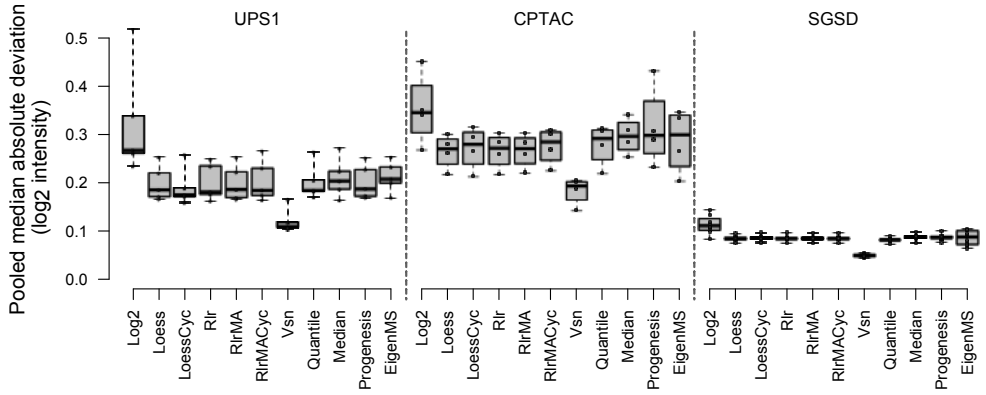


Figure 7. The effect of normalization methods on intragroup variation (pooled median absolute deviation, PMAD) between technical replicates in UPS1, CPTAC and the SGSD data

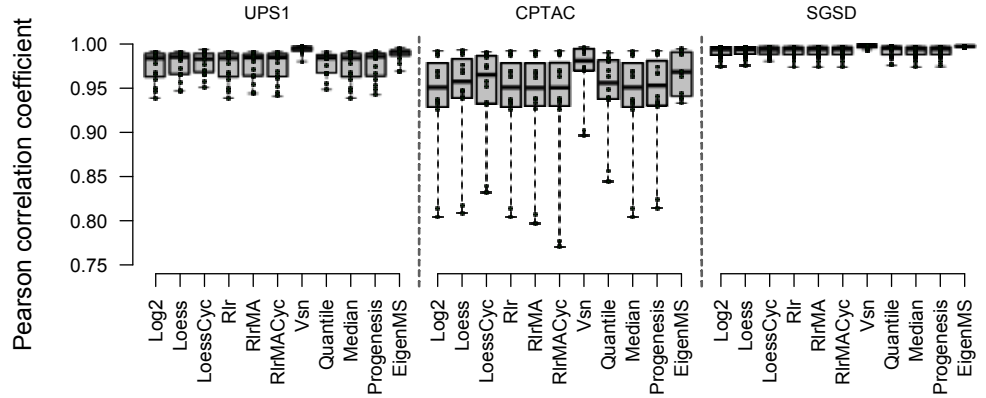


Figure 8. The effect of normalization methods on intragroup variation (Pearson correlation coefficient) between technical replicates in UPS1, CPTAC and the SGSD data

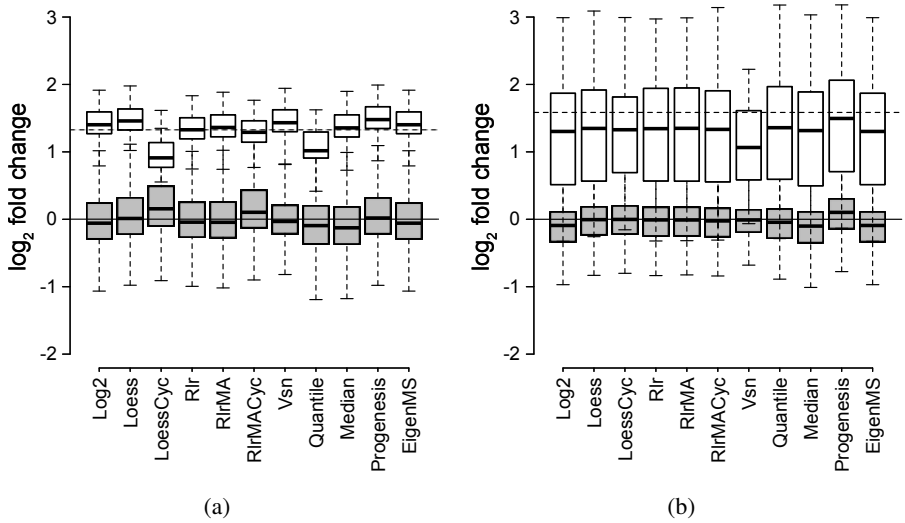


Figure 9. The $\log_2$ fold changes of the spike-in proteins (white box plots) and the background proteins (grey box plots) in (a) 10 fmol vs 25 fmol comparison of UPS1 data and (b) 0.74 fmol vs 2.2 fmol comparison of CPTAC data. The real fold changes of the spike-in proteins are represented by dashed lines.

ization methods (**Figure 9**). The fold changes of the spike-in proteins were typically underestimated for both normalized and $\log_2$-transformed data. The variance stabilization normalization (Vsn) sometimes resulted in smaller fold changes compared to other methods, *e.g.* in 0.74 fmol vs 2.2 fmol comparison of the CPTAC data (Wilcoxon signed rank test $p < 0.01$). In the UPS1 data, the fold changes of the spike-in proteins were generally closer to the theoretical changes after normalization regardless of the method, than in the un-normalized $\log_2$-transformed data.

5.1.2 Comparison of spectral counting methods

We evaluated the performance of multiple spectral counting methods and their variants (PLGEM, PepC, QSpec, Beta Binomial) in detecting differential expression. Proteins were ranked based on the statistical significance of their differential expression. Sums of absolute rank differences for all proteins were calculated for all pairwise comparisons and for randomly permuted ranks to determine statistical significance of the differences. Overall, the observed differences were smaller than what would be expected by random change ($p = 0.05$). In addition, the spectral counting methods were assessed by calculating overall correlations of protein ranks. In the UPS spike-in data, the correlation between PLGEM SC and Beta Binomial, but also between PLGEM SC and PepC, was exceptionally poor.

In addition to calculating the overall correlations between the spectral counting

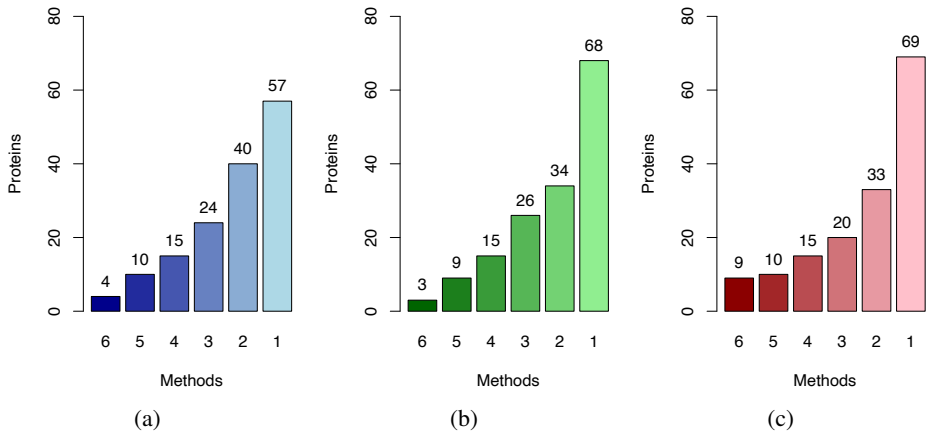


Figure 10. Number of proteins reported by at least *n* methods in their lists of top 25 differentially expressed proteins in (a) rat, (b) pig and (c) yeast data. Number of unique proteins reported by the different methods in their top 25 lists were 57 in rat data, 68 in pig data, and 69 in yeast data. Similarly, only four, three, and nine proteins were reported as differentially expressed by all six methods in rat, pig, and yeast data, respectively.

methods, we also considered only the top ranking proteins. For this purpose, combined top lists of the different analysis methods were constructed, and the number of methods reporting such proteins in their top lists were calculated (**Figure 10**). In all data sets, most proteins reaching top 25 by the differential expression statistics, were assigned by only one method. Furthermore, only a handful of proteins were reported by all the tested methods.

Finally, the overall performance of the spectral counting methods to detect differentially expressed proteins was assessed using receiver operating characteristic (ROC) curves (**Figure 11**). The un-normalized version of QSpec deviated most from others in the 25 fmol vs 50 fmol comparison highlighting the need for normalization. The overall best performance in terms of sensitivity and specificity (*i.e.* highest area under the ROC curve) was achieved by PLGEM NSAF.

5.2 New methods

The following sections summarize the proposed methods for differential expression analysis and their comparisons to existing tools (**Publications III**, **IV**, and **V**).

5.2.1 Peptide-level expression change averaging

Performance of the proposed peptide-level expression change averaging method (PECA) was evaluated using receiver operating characteristic (ROC) curves. **Figure 12** shows the curves and their area under the curve (AUC) values for 2 vs 4 fmol and the 25 vs

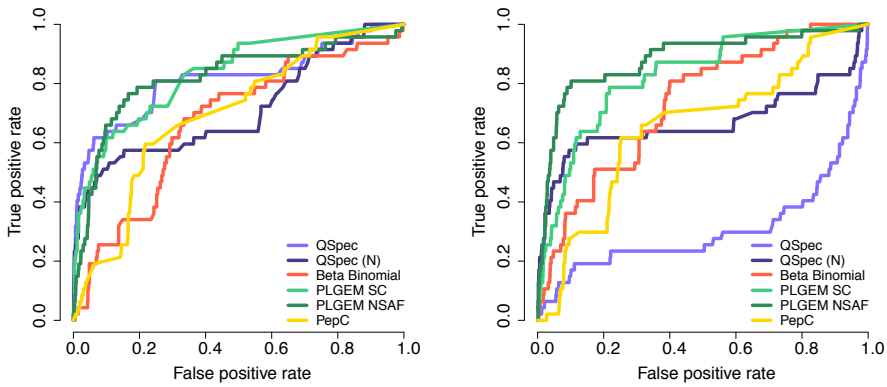


Figure 11. Receiver operating characteristic (ROC) curves for spectral counting methods in UPS spike-in data. The comparisons are from (a) 2 fmol vs 4 fmol and (b) 25 fmol vs 50 fmol setting.

50 fmol comparisons with different options for differential expression and summary statistic. Same statistical tests were used with both the pre-summed protein values and the corresponding peptide-level values. PECA with modified *t*-test and median as the aggregation method produced the best results. The overall performance is also significantly better than the others in all the tested comparisons (DeLong's test $p < 0.01$).

Differentially expressed UPS proteins detected by PECA ($p < 0.05$) from the 2 fmol vs 4 fmol comparison are shown in **Figure 13**. The summarized protein-level values (black) have higher intensity in the 4 fmol sample, but because similar changes can be detected in the yeast proteins (*i.e.* false positives), many of them would not be detected as differentially expressed. The peptide-level values (grey) show a systematic change while the yeast proteins have on average equal number of up and down-regulated peptides. This means that by utilizing the peptide-level measurements, the UPS proteins can more easily be detected as differentially expressed. In total, with the proposed peptide-level approach, there are 38 proteins detected as differentially expressed ($p < 0.05$), out of which 16 are UPS proteins. Similarly, when using a protein-level approach, there are a total of 55 proteins detected as differentially expressed out of which only one is a UPS protein. This clearly shows the improved ability of the PECA method to detect correct signal from noisy data.

The proposed PECA method was also benchmarked against similar methods using the un-summarized peptide-level data for differential expression analysis. **Figure 14** shows the results for comparison on the 2 vs 4 fmol and on the 25 vs 50 fmol datasets. Because of internal filtering mechanisms of these methods, only such proteins that were common between the results were used in the benchmarking (947

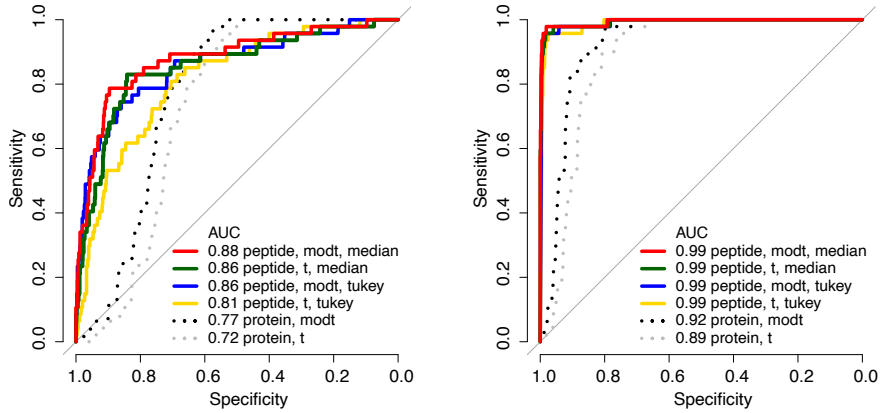


Figure 12. Receiver operating characteristic (ROC) curves and their area under the curve (AUC) for peptide-based PECA and the corresponding protein-level analysis in (a) 2 vs 4 fmol and (b) 25 vs 50 fmol comparisons of the USP spike-in data.

out of 1387 proteins). With *InfernoRDN*, three different rollup methods were tested. In all comparisons the PECA method outperformed the other tested methods.

5.2.2 Reproducibility-optimized peptide-level expression change averaging

The ROPECA method was benchmarked using DIA data, which contains a constant background proteome and 12 spike-in proteins in eight different concentrations. Performance of ROPECA, MSstats, and mapDIA, together with a commonly used protein-level *t*-test, were compared using ROC curves (**Figure 15a**). The results were merged from all possible pairwise comparisons from the tested sample groups. Because in some cases the overall differences were small and we wanted to focus on the most significant findings, we calculated only partial area under the curves (pAUC) for specificity above 0.9. Overall, ROPECA detected more true positives and less false positives than other methods. When comparing to previous PECA and ROTS methods, the new approach had the ability of PECA to detect true positives accurately from peptide-level data, while also the ability of ROTS to report less false positives.

The ROPECA method was also tested using a hybrid proteome data, where human, yeast, and *E. coli* are mixed together to create benchmarking samples. A data set with a larger proportion of up or down-regulated proteins could be considered more accurately to reflect biological changes than benchmarks with just a few spike-in proteins. Performance of the methods was investigated using ROC curves

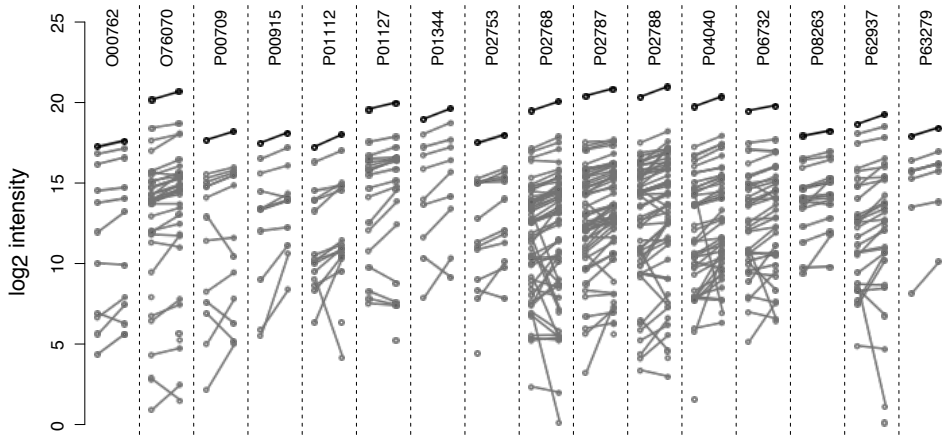


Figure 13. Visualization of peptide intensities (grey) and their corresponding protein-level values (black) of differentially expressed UPS proteins by PECA in the 2 vs 4 fmol comparison of UPS spike-in data. For each protein, the 2 fmol abundance is on the left and 4 fmol on the right.

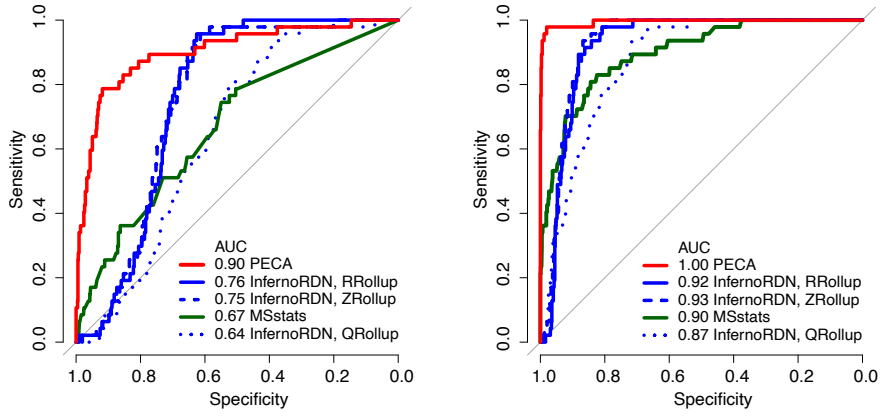


Figure 14. Receiver operating characteristic (ROC) curves of peptide-centric methods using (a) 2 vs 4 fmol and (b) 25 vs 50 fmol setting of the USP spike-in data.

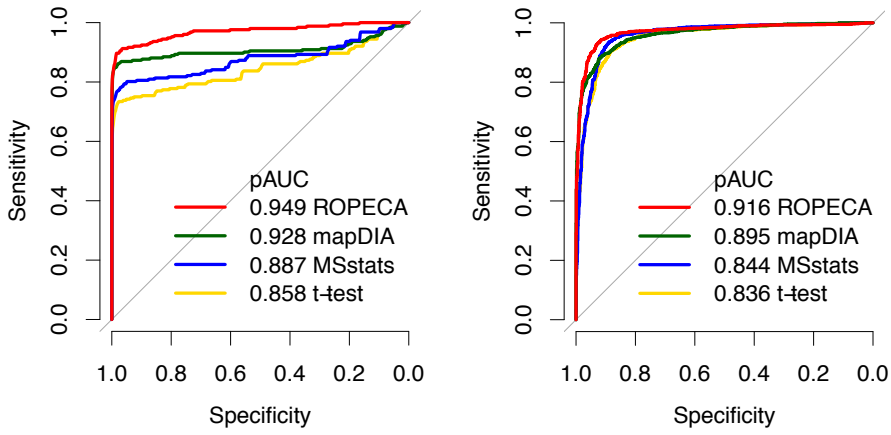


Figure 15. Receiver operating characteristic (ROC) curves for (a) the DIA profiling standard data and (b) the hybrid proteome benchmark data.

(**Figure 15b**). ROPECA produced significantly higher pAUC (0.916) than the other methods (bootstrap test $p < 0.01$). When comparing to previously introduced PECA and ROTS methods, ROPECA outperformed them in terms of true and false positives

In addition to the benchmark datasets, a publicly available twin study data set was used in order to test the methods in a clinical setting. From all the available data ($n = 116$), 14 individuals diagnosed with type 2 diabetes mellitus (T2D) were selected as the test group and the rest of the population as the control group. Differential expression analysis was performed between the selected groups and to test the reproducibility of the detections, the results (FDR < 0.05) from the full dataset we compared against randomly sampled subsets. **Figure 16** shows the overlaps of the ROPECA method and the commonly used *t*-test from 100 randomly sampled subsets. With the different subsets ROPECA produced more common detections than standard *t*-test (Wilcoxon signed rank test $p <0.01$).

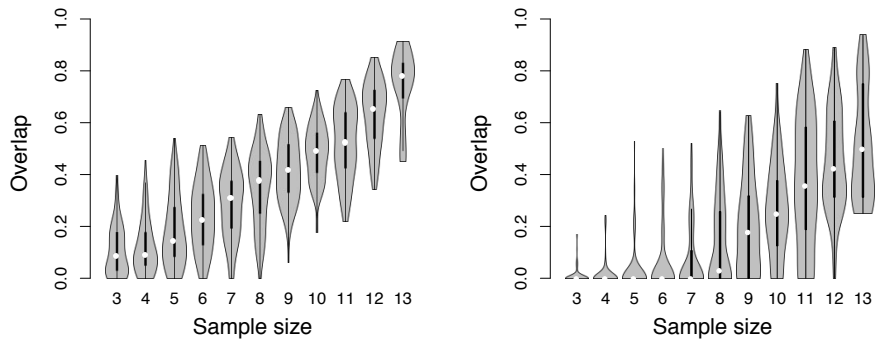


Figure 16. Violin plots showing the reproducibility in clinical twin study. Overlap of differentially expressed proteins (FDR < 0.05) between full data and 100 sampled subsets using (a) ROPECA and (b) *t*-test with varying number of samples.

6 Discussion

In this Thesis, I have tackled the different aspects of MS proteomics data analysis and demonstrated how different choices made during the analysis may affect the final results. I have compared popular normalization and spectral counting methods and found that there are clear differences in their performance. I have also proposed new peptide-level methods for differential expression analysis and shown that the new methods perform well. After all, it is desirable to find the best protein candidates for further validation, even if the data is noisy. Therefore, understanding both the limitations and the possibilities of the underlying technology is critical.

Normalization makes the samples comparable for statistical testing and forms the foundation for differential expression analysis. In the benchmark study of **Publication I**, the variance stabilization normalization (VSN) consistently performed well in reducing variation between replicates and resulted in high AUCs with the differential expression analysis, which eventually ranked it as the top performing normalization method. This observation is in line with other studies as well [12; 73].

Previous studies have suggested the linear or local regression normalization to perform well in reducing intragroup variation [12; 73; 78], and the same was observed also in the present study. However, it turned out that their performance depends on the data. The local regression normalization performed well in the UPS1 dataset while the linear regression normalization performed well in the CPTAC and SGSD datasets, which indicates that there is a different kind of bias in the datasets.

The UPS1 dataset has much larger variation of total protein abundances than the other datasets. This could stem from a number of reasons, including different laboratory protocols and instrumentation. It should be noted, however, that such sample to sample variation is typical in *real* experimental settings too. Therefore, the normalization methods should not be too sensitive to the underlying characteristics of the data. More importantly, even if there is a high quality dataset with fairly equal total intensities between the samples to begin with, it cannot be deduced whether a simple logarithmic transformation without any normalization is sufficient. An interesting future possibility could be to algorithmically select the most suitable normalization method based on a given data.

More generally, it can be argued that ideally one should not make any assumptions on the bias, unless the method is used for some specific purpose. One such case would be the normalization of phosphoproteomics samples, where the samples un-

dergo an enrichment step before quantification. To tackle such a specific scenario, we have published a tool that normalizes enriched samples using non-enriched controls as a reference [132].

The overall importance of normalization is highlighted by the fact that our benchmarking efforts were also featured on *News in Proteomics Research*, which is arguably the most popular blog in the field [1].

In addition to normalization, we have also benchmarked the different workflows for protein quantification and found that the commercial Progenesis software performed systematically well [88]. However, after filtering and LLS imputation, the open source MaxQuant had the overall best performance [88]. We also showed that MaxQuant produced the most accurate estimates of fold changes for the spike-in proteins (true positives), while other tools estimated accurately only the background proteins (true negatives) of the benchmark data [88]. This supports the general use of MaxQuant for protein quantification. Regarding differential expression analysis, we have previously also shown that the reproducibility-optimized statistical testing performs generally better than other tested methods with proteomics data [20]. Therefore, MaxQuant together with VSN and ROTS forms the basic recommended combination for the analysis of proteomics data.

It is obvious that different software parameters or different algorithms can affect the final outcome of the study. This is especially true for modular workflows, for which different algorithms can be separately selected for each processing step. Arguably, an experienced user could be able to optimise the performance of these customisable software workflows by selecting a proper combination of algorithms together with their parameters based on prior expert knowledge and the data at hand. It would be tempting to systematically benchmark all possible combinations of the different processing steps, including normalizations and statistical methods, using extensive set of benchmarking datasets. However, the number of possible combinations quickly becomes unfeasible to test, even using high-performance computing (HPC) clusters.

Spectral counting is an intriguing and much simpler method for quantifying proteins than the traditional MS1 based approach. For differential expression analysis, it requires methods that are tailored for such data. I studied the utility of spectral counting methods in **Publication II**. Each algorithm produced a significance metric for each protein and the final rankings were used to evaluate the level of agreement between the results. While the similarity of the results was not as good as desired, a number of proteins were differentially expressed with high confidence by multiple methods. While the potential of the spectral counting methods has been recognized [15], their inadequate performance has also been documented [133]. In contrast to our results suggesting that PLGEM performed the best, others have found QSpec to

[1] http://proteomicsnews.blogspot.com/2019/07/an-incredibly-comprehensive-evaluation.html

perform better [133]. Thus, no general conclusions could be made regarding the relative performance of the methods. Overall, our study supported the idea that spectral counting in general is not as reliable as the more traditional MS1 based quantification. A reasonable approach to spectral counting workflows could be a *voting* type approach, where results of multiple methods are summarised to generate a final consensus list of differentially expressed proteins.

All spectral counting methods reported a number of background proteins of the test data to be significantly differentially expressed between sample groups (*i.e.* false positives). Perhaps increasing the amount of spike-in proteins inherently affects the proportion of all other material in the sample. This likely affects especially the spectral counting approach, where the peptides are competing for detection by LC-MS/MS system.

Differential expression analysis is an important part of a quantitative proteomics experiment. It provides candidate proteins for further downstream analysis or targets for validation. This is particularly important with the commonly used label-free approach, which only allows relative quantification between samples. In **Publication III**, I implemented the reproducibility-optimized test statistic as an R package ROTS. Besides showing that it performs well, I have put emphasis that it is easy to use and well maintained to promote its wide use. These factors have likely played a role in keeping the software consistently in top 20 % of the most downloaded R packages in Bioconductor. It has also led to being included in comparative studies by third parties in high-impact journals [134], where it has performed well.

Finally, I showed in **Publications IV** and **V** that differential expression analysis using peptide-level data is a viable method for proteomics. Both of the proposed methods (PECA and ROPECA) worked better than the other tested methods. Both methods allow to utilize the full potential of MS data by considering directly the peptide measurements. Others have also shown the benefits of peptide-based models [117]. This is different from most state-of-the-art approaches where the peptide-level signal is still commonly summarized in various ways before performing statistical testing [135]. I published both methods as an easy-to-use R package. Perhaps this is one of the reasons why our effort was featured on the official blog of ThermoFischer Scientific, one of the manufacturers of mass spectrometry instruments [2].

It was observed that sometimes a single peptide can have a strong opposite fold change compared to the other peptides of the same protein. This is a major issue with the protein-level summarization approach, because such peptides would effectively level out the summarized signal. This suggests that all the available data should be utilized for statistical testing. Interestingly, our benchmarks showed the biggest improvements in the most difficult test cases, where the differences between the tested

[2] https://www.thermofisher.com/blog/proteomics/differential-protein-quantitation-at-the-peptide-level/

groups were smallest. On the other hand, it is likely that with high enough concentration levels, the peptide-level approach no longer brings any benefits compared to protein-level analysis. However, the tested concentrations and fold changes fall within a range that is commonly encountered in real samples. More importantly, the smaller differences can be detected reliably, the better.

The improved peptide-level expression change averaging method for DIA proteomics data was evaluated not only by the different spike-in benchmark data, but also using a clinical twin study to asses the biomedical relevance of the findings. Many of the reported proteins were supported by the literature, but not detected as differentially expressed by conventional methods. This essentially means that applying novel methods to exiting data has the potential to reveal completely new information from the accumulating publicly available data.

Overall, the data-independent acquisition technique shows great promise as new methods and algorithms are constantly being published. One interesting avenue is the application of the technique in the context of metaproteomics, where peptide identification is extremely challenging because of trace amounts and overlapping peptide sequences between the different species. To this end, we have developed a software workflow for DIA metaproteomics [136] that includes the new differential expression analysis tools, and applied it successfully in the context of complex human gut microbiota [137].

To conclude, mass spectrometry proteomics offers great opportunities for biological and biomedical studies. However, interpretation of the data requires careful attention and cross-disciplinary expertise. In this Thesis, contributions to the field were made by systematically investigating many of the challenges in the data analysis in order to help interpretation of the data.

7 Summary of publications

I: A systematic evaluation of normalization methods in quantitative label-free proteomics

In this work, a set of widely used normalization methods were tested. They represent the various strategies that are commonly used to normalize high-throughput omics data. Several datasets that had spike-in proteins representing the ground truth were used to evaluate the methods. They were evaluated by their ability to reduce variation between the replicates, accuracy of fold changes between sample groups, and more importantly, the accuracy of differential expression analysis by a statistical test. In this study, the variance stabilization normalization (VSN) was found to perform overall best, outperforming the other methods especially when examining differential expression analysis results.

II: Cross-Correlation of Spectral Count Ranking to Validate Quantitative Proteome Measurements

In this work, we studied various differential expression methods targeted especially for spectral counts of label-free mass spectrometry proteomics. The methods were evaluated using a dataset containing different concentrations of spike-in proteins representing the ground truth. It was found out that power law global error model had superior performance compared to the other tested methods in ranking the spike-in proteins as differentially expressed between sample groups. In addition, we studied the overall similarities of the methods by using three biological datasets. Generally, the methods had considerable differences when looking at the overall ranking of all proteins in the sample.

III: ROTS: An R package for reproducibility-optimized statistical testing

This work focuses on the reproducibility-optimized test statistic (ROTS) that works by adjusting a *t*-statistic based on the underlying data. Over the years the method has been applied to various omics studies with great success. Here, a publicly available R package was implemented and published in Bioconductor repository. Convenient features were included to the package and its performance was illustrated using three

case studies, including not only proteomics but also bulk and single cell RNA-seq data.

IV: Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins

In mass spectrometry proteomics peptides are identified and quantified before summarizing them to protein level. In this work, we introduced a differential expression analysis method (PECA) that utilizes all the peptide-level measurements when estimating the final protein-level statistic. Similarly as before, we used a controlled spike-in experiment that has additional proteins with known changes between the sample groups. It was shown that this kind of approach produces more accurate differential expression estimates than other methods. More importantly, the benefits of using such approach became more clear if the differences between the sample groups were small, which is often the case in real experiments. The method itself was published as an R package in Bioconductor.

V: Enhanced differential expression statistics for data-independent acquisition proteomics

In this work, we made a new reproducibility optimized method (ROPECA) that utilizes the underlying peptide-level measurements of mass spectrometry proteomics data. It has the benefits of both previously introduced methods ROTS and PECA. This was made possible by using the newly emerging data-independent acquisition (DIA) mass spectrometry data, where there is much less missing peptide measurements. The new method was benchmarked against other similar tools using a spike-in gold standard data with known concentration changes and a hybrid proteome data, where proteomes of multiple species are mixed together to produce known changes between sample groups. It was shown that our new method performed better than any of the previous approaches. Besides artificially generated data, the improved accuracy of the new method was explored by applying it to a clinical twin study.

List of References

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, P. Walte, J. Wilson, and T. Hunt, *Molecular Biology of The Cell*. Garland Science, 6th ed., 2008.

[2] M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams, "Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it.," *Biotechnology & Genetic Engineering Reviews*, vol. 13, pp. 19–50, 1996.

[3] H. Zhang, T. Liu, Z. Zhang, S. H. Payne, B. Zhang, J. E. McDermott, J.-Y. Zhou, V. A. Petyuk, L. Chen, D. Ray, S. Sun, F. Yang, L. Chen, J. Wang, P. Shah, S. W. Cha, P. Aiyetan, S. Woo, Y. Tian, M. A. Gritsenko, T. R. Clauss, C. Choi, M. E. Monroe, S. Thomas, S. Nie, C. Wu, R. J. Moore, K.-H. Yu, D. L. Tabb, D. Fenyö, V. Bafna, Y. Wang, H. Rodriguez, E. S. Boja, T. Hiltke, R. C. Rivers, L. Sokoll, H. Zhu, I.-M. Shih, L. Cope, A. Pandey, B. Zhang, M. P. Snyder, D. A. Levine, R. D. Smith, D. W. Chan, K. D. Rodland, and CPTAC Investigators, "Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer.," *Cell*, vol. 166, pp. 755–765, jul 2016.

[4] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen, *Computational Methods for Mass Spectrometry Proteomics*. Chichester, UK: John Wiley & Sons, Ltd, nov 2007.

[5] P. Legrain, R. Aebersold, A. Archakov, A. Bairoch, K. Bala, L. Beretta, J. Bergeron, C. H. Borchers, G. L. Corthals, C. E. Costello, E. W. Deutsch, B. Domon, W. Hancock, F. He, D. Hochstrasser, G. Marko-Varga, G. H. Salekdeh, S. Sechi, M. Snyder, S. Srivastava, M. Uhlén, C. H. Wu, T. Yamamoto, Y.-K. Paik, and G. S. Omenn, "The human proteome project: current state and future direction.," *Molecular & Cellular Proteomics : MCP*, vol. 10, p. M111.009993, jul 2011.

[6] S. Adhikari, E. C. Nice, E. W. Deutsch, L. Lane, G. S. Omenn, S. R. Pennington, Y.-K. Paik, C. M. Overall, F. J. Corrales, I. M. Cristea, J. E. Van Eyk, M. Uhlén, C. Lindskog, D. W. Chan, A. Bairoch, J. C. Waddington, J. L. Justice, J. LaBaer, H. Rodriguez, F. He, M. Kostrzewa, P. Ping, R. L. Gundry, P. Stewart, S. Srivastava, S. Srivastava, F. C. S. Nogueira, G. B. Domont, Y. Vandenbrouck, M. P. Y. Lam, S. Wennersten, J. A. Vizcaino, M. Wilkins, J. M. Schwenk, E. Lundberg, N. Bandeira, G. Marko-Varga, S. T. Weintraub, C. Pineau, U. Kusebauch, R. L. Moritz, S. B. Ahn, M. Palmblad, M. P. Snyder, R. Aebersold, and M. S. Baker, "A high-stringency blueprint of the human proteome," *Nature Communications*, vol. 11, p. 5301, dec 2020.

[7] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics.," *Nature*, vol. 422, pp. 198–207, mar 2003.

[8] H. Steen and M. Mann, "The ABC's (and XYZ's) of peptide sequencing.," *Nature Reviews Molecular Cell Biology*, vol. 5, pp. 699–711, 2004.

[9] B. F. Cravatt, G. M. Simon, and J. R. Yates, "The biological impact of mass-spectrometry-based proteomics.," *Nature*, vol. 450, pp. 991–1000, dec 2007.

[10] D. A. Megger, T. Bracht, H. E. Meyer, and B. Sitek, "Label-free quantification in clinical proteomics.," *Biochimica et Biophysica Acta*, vol. 1834, pp. 1581–90, aug 2013.

[11] M. Schirle, M. Bantscheff, and B. Kuster, "Mass spectrometry-based proteomics in preclinical drug discovery.," *Chemistry & Biology*, vol. 19, pp. 72–84, jan 2012.

[12] A. Chawade, E. Alexandersson, and F. Levander, "Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets," *Journal of Proteome Research*, vol. 13, pp. 3114–3120, 2014.

[13] Y. V. Karpievitch, A. R. Dabney, and R. D. Smith, "Normalization and missing value imputation for label-free LC-MS analysis.," *BMC Bioinformatics*, vol. 13 Suppl 1, no. Suppl 16, p. S5, 2012.

[14] X. Wang, W. Zhu, K. Pradhan, C. Ji, Y. Ma, O. J. Semmes, J. Glimm, and J. Mitchell, "Feature extraction in the analysis of proteomic mass spectra.," *Proteomics*, vol. 6, pp. 2095–100, apr 2006.

[15] D. H. Lundgren, S.-I. Hwang, L. Wu, and D. K. Han, "Role of spectral counting in quantitative proteomics.," *Expert Review of Proteomics*, vol. 7, no. 1, pp. 39–53, 2010.

[16] H. Choi, D. Fermin, and A. I. Nesvizhskii, "Significance analysis of spectral count data in label-free shotgun proteomics.," *Molecular & Cellular Proteomics : MCP*, vol. 7, pp. 2373–85, dec 2008.

[17] M. Krzywinski and N. Altman, "Points of significance: Designing comparative experiments," *Nature Methods*, vol. 11, pp. 215–216, feb 2014.

[18] S. Mukherjee and S. J. Roberts, "A theoretical analysis of gene selection.," *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*, pp. 131–41, 2004.

[19] L.-X. Qin, K. F. Kerr, and Contributing Members of the Toxicogenomics Research Consortium, "Empirical evaluation of data transformations and ranking statistics for microarray analysis.," *Nucleic Acids Research*, vol. 32, pp. 5471–9, oct 2004.

[20] A. Pursiheimo, A. P. Vehmas, S. Afzal, T. Suomi, T. Chand, L. Strauss, M. Poutanen, A. Rokka, G. L. Corthals, and L. L. Elo, "Optimization of Statistical Methods Impact on Quantitative Proteomics Data," *Journal of Proteome Research*, vol. 14, pp. 4118–4126, oct 2015.

[21] S. R. Eddy, "What is a hidden Markov model?," *Nature Biotechnology*, vol. 22, no. 10, pp. 1315–1316, 2004.

[22] J. P. Savaryn, T. K. Toby, and N. L. Kelleher, "A researcher's guide to mass spectrometry-based proteomics," *Proteomics*, pp. 1–9, aug 2016.

[23] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann, "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.," *Molecular & Cellular Proteomics : MCP*, vol. 1, pp. 376–86, may 2002.

[24] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules.," *Science*, vol. 246, pp. 64–71, oct 1989.

[25] R. Bruderer, O. M. Bernhardt, T. Gandhi, S. M. Miladinović, L.-Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, and L. Reiter, "Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues," *Molecular & Cellular Proteomics : MCP*, vol. 14, pp. 1400–1410, may 2015.

[26] V. Lange, P. Picotti, B. Domon, and R. Aebersold, "Selected reaction monitoring for quantitative proteomics: a tutorial.," *Molecular Systems Biology*, vol. 4, no. 222, p. 222, 2008.

[27] P. Picotti and R. Aebersold, "Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.," *Nature Methods*, vol. 9, pp. 555–66, may 2012.

[28] J. C. Silva, M. V. Gorenstein, G.-Z. Li, J. P. C. Vissers, and S. J. Geromanos, "Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition.," *Molecular & Cellular Proteomics : MCP*, vol. 5, pp. 144–56, jan 2006.

[29] P. C. Carvalho, X. Han, T. Xu, D. Cociorva, M. d. G. Carvalho, V. C. Barbosa, and J. R. Yates, "XDIA: improving on the label-free data-independent analysis," *Bioinformatics (Oxford, England)*, vol. 26, pp. 847–848, mar 2010.

[30] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a

new concept for consistent and accurate proteome analysis.," *Molecular & Cellular Proteomics : MCP*, vol. 11, p. O111.016717, jun 2012.

[31] J. D. Egertson, A. Kuehn, G. E. Merrihew, N. W. Bateman, B. X. MacLean, Y. S. Ting, J. D. Canterbury, D. M. Marsh, M. Kellmann, V. Zabrouskov, C. C. Wu, and M. J. MacCoss, "Multiplexed MS/MS for improved data-independent acquisition," *Nature Methods*, vol. 10, no. 8, pp. 744–746, 2013.

[32] U. Distler, J. Kuharev, P. Navarro, Y. Levin, H. Schild, and S. Tenzer, "Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics.," *Nature Methods*, vol. 11, pp. 167–70, feb 2014.

[33] C. R. Weisbrod, J. K. Eng, M. R. Hoopmann, T. Baker, and J. E. Bruce, "Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification.," *Journal of Proteome Research*, vol. 11, pp. 1621–32, mar 2012.

[34] H. Pak, F. Nikitin, F. Gluck, F. Lisacek, A. Scherl, and M. Muller, "Clustering and filtering tandem mass spectra acquired in data-independent mode.," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 1862–71, dec 2013.

[35] J. Teleman, H. L. Rost, G. Rosenberger, U. Schmitt, L. Malmstrom, J. Malmstrom, and F. Levander, "DIANA–algorithmic improvements for analysis of data-independent acquisition MS data," *Bioinformatics (Oxford, England)*, vol. 31, no. 4, pp. 555–562, 2015.

[36] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii, "DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics," *Nature Methods*, vol. 12, pp. 258–264, jan 2015.

[37] M. Choi, C.-Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek, "MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 2524–6, sep 2014.

[38] G. Teo, S. Kim, C.-C. Tsou, B. Collins, A.-C. Gingras, A. I. Nesvizhskii, and H. Choi, "mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry," *Journal of Proteomics*, 2015.

[39] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, "Review of peak detection algorithms in liquid-chromatography-mass spectrometry.," *Current Genomics*, vol. 10, pp. 388–401, sep 2009.

[40] H. Lam, "Building and searching tandem mass spectral libraries for peptide identification.," *Molecular & Cellular Proteomics : MCP*, vol. 10, p. R111.008565, dec 2011.

[41] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 976–89, nov 1994.

[42] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data.," *Electrophoresis*, vol. 20, pp. 3551–67, dec 1999.

[43] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm.," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–64, 2004.

[44] R. Craig and R. C. Beavis, "TANDEM: matching proteins with tandem mass spectra.," *Bioinformatics (Oxford, England)*, vol. 20, pp. 1466–7, jun 2004.

[45] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: a peptide search engine integrated into the MaxQuant environment.," *Journal of Proteome Research*, vol. 10, no. 4, pp. 1794–1805, 2011.

[46] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: an open-source MS/MS sequence database search tool.," *Proteomics*, vol. 13, pp. 22–4, jan 2013.

[47] S. Kim and P. A. Pevzner, "MS-GF+ makes progress towards a universal database search tool for proteomics.," *Nature Communications*, vol. 5, p. 5277, oct 2014.

[48] R. Apweiler, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Research*, vol. 32, pp. 115D–119, jan 2004.

[49] E. Gasteiger, E. Jung, and A. Bairoch, "SWISS-PROT: connecting biomolecular knowledge via a protein database.," *Current Issues in Molecular Biology*, vol. 3, pp. 47–55, jul 2001.

[50] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998.," *Nucleic Acids Research*, vol. 26, pp. 38–42, jan 1998.

[51] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.," *Nucleic Acids Research*, vol. 44, pp. D733–45, jan 2016.

[52] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, P. Larsson, I. Longden, W. McLaren, B. Overduin, B. Pritchard, H. S. Riat, D. Rios, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sobral, G. Spudich, Y. A. Tang, S. Trevanion, J. Vandrovcova, A. J. Vilella, S. White, S. P. Wilder, A. Zadissa, J. Zamora, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Vogel, and S. M. J. Searle, "Ensembl 2011," *Nucleic Acids Research*, vol. 39, pp. D800–D806, jan 2011.

[53] B. Mesuere, B. Devreese, G. Debyser, M. Aerts, P. Vandamme, and P. Dawyndt, "Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples," *Journal of Proteome Research*, vol. 11, no. 12, pp. 5773–5780, 2012.

[54] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, and MetaHIT Consortium, "An integrated catalog of reference genes in the human gut microbiome.," *Nature Biotechnology*, vol. 32, pp. 834–41, aug 2014.

[55] I. F. Escapa, T. Chen, Y. Huang, P. Gajare, F. E. Dewhirst, and K. P. Lemon, "New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract," *mSystems*, vol. 3, dec 2018.

[56] A. I. Nesvizhskii, "Proteogenomics: concepts, applications and computational strategies," *Nature Methods*, vol. 11, pp. 1114–1125, nov 2014.

[57] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.," *Rapid Communications in Mass Spectrometry : RCM*, vol. 17, no. 20, pp. 2337–42, 2003.

[58] H. Chi, R.-X. Sun, B. Yang, C.-Q. Song, L.-H. Wang, C. Liu, Y. Fu, Z.-F. Yuan, H.-P. Wang, S.-M. He, and M.-Q. Dong, "pNovo: de novo peptide sequencing and identification using HCD spectra.," *Journal of Proteome Research*, vol. 9, pp. 2713–24, may 2010.

[59] B. Ma, "Novor: real-time peptide de novo sequencing software.," *Journal of the American Society for Mass Spectrometry*, vol. 26, pp. 1885–94, nov 2015.

[60] C. Hughes, B. Ma, and G. A. Lajoie, "De Novo Sequencing Methods in Proteomics," in *Proteome Bioinformatics. Methods in Molecular Biology.*, pp. 105–121, Humana Press, 2010.

[61] R. A. Zubarev, "The challenge of the proteome dynamic range and its implications for in-depth proteomics.," *Proteomics*, vol. 13, pp. 723–6, mar 2013.

[62] J. Griss, Y. Perez-Riverol, S. Lewis, D. L. Tabb, J. A. Dianes, N. Del-Toro, M. Rurik, M. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, and J. A. Vizcaíno, "Recognizing millions of consis-

tently unidentified spectra across hundreds of shotgun proteomics datasets," *Nature Methods*, vol. 13, pp. 651–656, aug 2016.

[63] W. Timp and G. Timp, "Beyond mass spectrometry, the next step in proteomics.," *Science Advances*, vol. 6, no. 2, p. eaax8978, 2020.

[64] C. Tu, P. A. Rudnick, M. Y. Martinez, K. L. Cheek, S. E. Stein, R. J. C. Slebos, and D. C. Liebler, "Depletion of abundant plasma proteins and limitations of plasma proteomics.," *Journal of Proteome Research*, vol. 9, pp. 4982–91, oct 2010.

[65] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold, "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.," *Nature Biotechnology*, vol. 17, pp. 994–9, oct 1999.

[66] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin, "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.," *Molecular & Cellular Proteomics : MCP*, vol. 3, pp. 1154–69, dec 2004.

[67] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi, "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 6940–5, jun 2003.

[68] N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer, "Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis.," *Nature Biotechnology*, vol. 28, pp. 83–9, jan 2010.

[69] N. Colaert, C. Van Huele, S. Degroeve, A. Staes, J. Vandekerckhove, K. Gevaert, and L. Martens, "Combining quantitative proteomics data processing workflows for greater sensitivity.," *Nature Methods*, vol. 8, no. 6, pp. 481–483, 2011.

[70] J. E. Jackson, *A Use's Guide to Principal Components*. Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons, Inc., mar 1991.

[71] D. A. Megger, T. Bracht, H. E. Meyer, and B. Sitek, "Label-free quantification in clinical proteomics.," *Biochimica et Biophysica Acta*, vol. 1834, pp. 1581–90, aug 2013.

[72] F. Meissner and M. Mann, "Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology.," *Nature Immunology*, vol. 15, pp. 112–7, feb 2014.

[73] K. Kultima, A. Nilsson, B. Scholz, U. L. Rossbach, M. Fälth, and P. E. Andrén, "Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides.," *Molecular & Cellular Proteomics : MCP*, vol. 8, no. 10, pp. 2285–2295, 2009.

[74] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 185–93, jan 2003.

[75] S. E. Choe, M. Boutros, A. M. Michelson, G. M. Church, and M. S. Halfon, "Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.," *Genome Biology*, vol. 6, no. 2, p. R16, 2005.

[76] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.," *Nucleic Acids Research*, vol. 30, p. e15, feb 2002.

[77] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong, "Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.," *Journal of Cellular Biochemistry*, vol. Suppl 37, pp. 120–5, 2001.

[78] S. J. Callister, R. C. Barry, J. N. Adkins, E. T. Johnson, W. J. Qian, B. J. M. Webb-Robertson, R. D. Smith, and M. S. Lipton, "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics," *Journal of Proteome Research*, vol. 5, no. 2, pp. 277–286, 2006.

[79] B.-J. M. Webb-Robertson, M. M. Matzke, J. M. Jacobs, J. G. Pounds, and K. M. Waters, "A statistical selection strategy for normalization procedures in LC-MS proteomics experi-

ments through dataset-dependent ranking of normalization scaling factors," *Proteomics*, vol. 11, pp. 4736–4741, dec 2011.

[80] D. Amaratunga and J. Cabrera, "Analysis of data from viral DNA microchips," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1161–1170, 2001.

[81] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition*. Springer Texts in Statistics, New York: Springer-Verlag, 2004.

[82] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies.," *Nucleic Acids Research*, vol. 43, p. e47, apr 2015.

[83] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression.," *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, pp. S96–104, jul 2002.

[84] W. Hou, Z. Ji, H. Ji, and S. C. Hicks, "A systematic evaluation of single-cell RNA-sequencing imputation methods," *Genome Biology*, vol. 21, p. 218, dec 2020.

[85] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics.," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[86] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[87] D. van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion.," *Cell*, vol. 174, no. 3, pp. 716–729.e27, 2018.

[88] T. Välikangas, T. Suomi, and L. L. Elo, "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation.," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1344–1355, 2018.

[89] J. Tuikkala, L. Elo, O. S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in microarray data with gene ontology.," *Bioinformatics (Oxford, England)*, vol. 22, pp. 566–72, mar 2006.

[90] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 2088–96, nov 2003.

[91] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, "pcaMethods–a bioconductor package providing PCA methods for incomplete data.," *Bioinformatics (Oxford, England)*, vol. 23, pp. 1164–7, may 2007.

[92] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 187–98, jan 2005.

[93] Q. Xiang, X. Dai, Y. Deng, C. He, J. Wang, J. Feng, and Z. Dai, "Missing value imputation for microarray gene expression data using histone acetylation information.," *BMC Bioinformatics*, vol. 9, p. 252, may 2008.

[94] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays.," *Bioinformatics (Oxford, England)*, vol. 17, pp. 520–5, jun 2001.

[95] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 5116–21, apr 2001.

[96] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.," *FEBS letters*, vol. 573, pp. 83–92, aug 2004.

[97] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data.," *BMC Bioinformatics*, vol. 14, p. 91, mar 2013.

[98] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.," *Genome Biology*, vol. 14, no. 9, p. R95, 2013.

[99] F. Seyednasrollah, A. Laiho, and L. L. Elo, "Comparison of software packages for detecting differential expression in RNA-seq studies.," *Briefings in Bioinformatics*, vol. 16, pp. 59–70, jan 2015.

[100] M. K. Jaakkola, F. Seyednasrollah, A. Mehmood, and L. L. Elo, "Comparison of methods to detect differentially expressed genes between single-cell populations.," *Briefings in Bioinformatics*, vol. 18, pp. 735–743, jul 2017.

[101] X. Fu, S. A. Gharib, P. S. Green, M. L. Aitken, D. A. Frazer, D. R. Park, T. Vaisar, and J. W. Heinecke, "Spectral index for assessment of differential protein expression in shotgun proteomics.," *Journal of Proteome Research*, vol. 7, pp. 845–54, mar 2008.

[102] N. Pavelka, M. L. Fournier, S. K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens, and M. P. Washburn, "Statistical similarities between transcriptomics and quantitative shotgun proteomics data.," *Molecular & Cellular Proteomics : MCP*, vol. 7, pp. 631–44, apr 2008.

[103] T. V. Pham, S. R. Piersma, M. Warmoes, and C. R. Jimenez, "On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics.," *Bioinformatics*, vol. 26, pp. 363–9, feb 2010.

[104] P. C. Carvalho, J. S. G. Fischer, E. I. Chen, J. R. Yates, and V. C. Barbosa, "PatternLab for proteomics: a tool for differential shotgun proteomics.," *BMC Bioinformatics*, vol. 9, p. 316, jul 2008.

[105] M. Li, W. Gray, H. Zhang, C. H. Chung, D. Billheimer, W. G. Yarbrough, D. C. Liebler, Y. Shyr, and R. J. C. Slebos, "Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling," *Journal of Proteome Research*, vol. 9, no. 8, pp. 4295–4305, 2010.

[106] N. L. Heinecke, B. S. Pratt, T. Vaisar, and L. Becker, "PepC: proteomics software for identifying differentially expressed proteins based on spectral counting.," *Bioinformatics (Oxford, England)*, vol. 26, pp. 1574–5, jun 2010.

[107] K. M. Little, J. K. Lee, and K. Ley, "ReSASC: a resampling-based algorithm to determine differential protein expression from spectral count data.," *Proteomics*, vol. 10, pp. 1212–22, mar 2010.

[108] J. G. Booth, K. E. Eilertson, P. D. B. Olinares, and H. Yu, "A bayesian mixture model for comparative spectral count data in shotgun proteomics.," *Molecular & Cellular Proteomics : MCP*, vol. 10, p. M110.007203, aug 2011.

[109] B. Zybailov, A. L. Mosley, M. E. Sardiu, M. K. Coleman, L. Florens, and M. P. Washburn, "Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae.," *Journal of Proteome Research*, vol. 5, pp. 2339–47, sep 2006.

[110] L. L. Elo, S. Filén, R. Lahesmaa, and T. Aittokallio, "Reproducibility-optimized test statistic for ranking genes in microarray studies.," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 423–31, 2008.

[111] F. Seyednasrollah, K. Rantanen, P. Jaakkola, and L. L. Elo, "ROTS: reproducible RNA-seq biomarker detector-prognostic markers for clear cell renal cell cancer.," *Nucleic Acids Research*, vol. 44, p. e1, jan 2016.

[112] F.-y. Cheng, K. Blackburn, Y.-m. Lin, M. B. Goshe, and J. D. Williamson, "Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses.," *Journal of Proteome Research*, vol. 8, pp. 82–93, jan 2009.

[113] K. Ning, D. Fermin, and A. I. Nesvizhskii, "Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data," *Journal of Proteome Research*, vol. 11, no. 4, pp. 2261–2271, 2012.

[114] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W.-J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney, "A statistical framework for protein quantitation in bottom-up MS-based proteomics.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 2028–34, aug 2009.

[115] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, and O. Vitek, "Protein quantification in label-free LC-MS experiments.," *Journal of Proteome Research*, vol. 8, pp. 5275–84, nov 2009.

[116] Y. V. Bukhman, M. Dharsee, R. Ewing, P. Chu, T. Topaloglou, T. Le Bihan, T. Goh, H. Duewel, I. I. Stewart, J. R. Wisniewski, and N. F. Ng, "Design and analysis of quantitative differential proteomics investigations using LC-MS technology.," *Journal of Bioinformatics and Computational Biology*, vol. 6, pp. 107–23, feb 2008.

[117] L. J. E. Goeminne, A. Argentini, L. Martens, and L. Clement, "Summarization vs. Peptide-Based Models in Label-free Quantitative Proteomics: Performance, Pitfalls and Data Analysis Guidelines," *Journal of Proteome Research*, p. 150401063858005, 2015.

[118] A. K. Gupta and N. Saralees, *Handbook of Beta Distribution and Its Applications*. CRC Press, 2004.

[119] A. D. Polpitiya, W.-J. Qian, N. Jaitly, V. a. Petyuk, J. N. Adkins, D. G. Camp, G. a. Anderson, and R. D. Smith, "DAnTE: a statistical tool for quantitative analysis of -omics data.," *Bioinformatics (Oxford, England)*, vol. 24, pp. 1556–8, jul 2008.

[120] A. Bilbao, E. Varesio, J. Luban, C. Strambio-De-Castillia, G. Hopfgartner, M. Müller, and F. Lisacek, "Processing strategies and software solutions for data-independent acquisition in mass spectrometry.," *Proteomics*, vol. 15, pp. 964–80, mar 2015.

[121] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, and R. Aebersold, "The PeptideAtlas project.," *Nucleic Acids Research*, vol. 34, pp. D655–8, jan 2006.

[122] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma, and J. A. Vizcaíno, "The PRIDE database and related tools and resources in 2019: improving support for quantification data.," *Nucleic Acids Research*, vol. 47, no. D1, pp. D442–D450, 2019.

[123] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, and J. A. Vizcaíno, "Making proteomics data accessible and reusable: current state of proteomics databases and repositories.," *Proteomics*, vol. 15, pp. 930–49, mar 2015.

[124] P. A. Rudnick, K. R. Clauser, L. E. Kilpatrick, D. V. Tchekhovskoi, P. Neta, N. Blonder, D. D. Billheimer, R. K. Blackman, D. M. Bunk, H. L. Cardasis, A.-J. L. Ham, J. D. Jaffe, C. R. Kinsinger, M. Mesri, T. A. Neubert, B. Schilling, D. L. Tabb, T. J. Tegeler, L. Vega-Montoto, A. M. Variyath, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. A. Carr, S. J. Fisher, B. W. Gibson, A. G. Paulovich, F. E. Regnier, H. Rodriguez, C. Spiegelman, P. Tempst, D. C. Liebler, and S. E. Stein, "Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses.," *Molecular & Cellular Proteomics : MCP*, vol. 9, pp. 225–41, feb 2010.

[125] A. G. Paulovich, D. Billheimer, A.-J. L. Ham, L. Vega-Montoto, P. A. Rudnick, D. L. Tabb, P. Wang, R. K. Blackman, D. M. Bunk, H. L. Cardasis, K. R. Clauser, C. R. Kinsinger, B. Schilling, T. J. Tegeler, A. M. Variyath, M. Wang, J. R. Whiteaker, L. J. Zimmerman, D. Fenyo, S. A. Carr, S. J. Fisher, B. W. Gibson, M. Mesri, T. A. Neubert, F. E. Regnier, H. Rodriguez, C. Spiegelman, S. E. Stein, P. Tempst, and D. C. Liebler, "Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance.," *Molecular & Cellular Proteomics : MCP*, vol. 9, pp. 242–54, feb 2010.

[126] D. L. Tabb, L. Vega-Montoto, P. a. Rudnick, A. M. Variyath, A.-J. L. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. a. Carr, K. R. Clauser, J. D. Jaffe, K. a. Kowalski, T. a. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman, "Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry.," *Journal of Proteome Research*, vol. 9, pp. 761–76, feb 2010.

[127] O. M. Bernhardt, N. Selevsek, L. C. Gillet, O. Rinner, P. Picotti, R. Aebersold, and L. Reiter,

"Spectronaut: A fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data," *Biognosys. ch*, 2012.

[128] A. P. Vehmas, M. Adam, T. D. Laajala, G. Kastenmüller, C. Prehn, J. Rozman, C. Ohlsson, H. Fuchs, M. Hrabě de Angelis, V. Gailus-Durner, L. L. Elo, T. Aittokallio, J. Adamski, G. Corthals, M. Poutanen, and L. Strauss, "Liver lipid metabolism is altered by increased circulating estrogen to androgen ratio in male mouse.," *Journal of Proteomics*, vol. 133, pp. 66–75, feb 2016.

[129] J. Kuharev, P. Navarro, U. Distler, O. Jahn, and S. Tenzer, "In-depth evaluation of software tools for data-independent acquisition based label-free quantification.," *Proteomics*, vol. 15, pp. 3140–51, sep 2015.

[130] Y. Liu, A. Buil, B. C. Collins, L. C. J. Gillet, L. C. Blum, L.-Y. Cheng, O. Vitek, J. Mouritsen, G. Lachance, T. D. Spector, E. T. Dermitzakis, and R. Aebersold, "Quantitative variability of 342 plasma proteins in a human twin population.," *Molecular Systems Biology*, vol. 11, no. 1, p. 786, 2015.

[131] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmström, L. Malmström, and R. Aebersold, "OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data," *Nature Biotechnology*, vol. 32, pp. 219–223, mar 2014.

[132] S. Saraei, T. Suomi, O. Kauko, and L. L. Elo, "Phosphonormalizer : an R package for normalization of MS-based label-free phosphoproteomics," *Bioinformatics*, vol. 34, no. September 2017, pp. 693–694, 2018.

[133] S. R. Langley and M. Mayr, "Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics," *Journal of Proteomics*, 2015.

[134] C. Soneson and M. D. Robinson, "Bias, robustness and scalability in single-cell differential expression analysis," *Nature Methods*, feb 2018.

[135] M. M. Matzke, J. N. Brown, M. a. Gritsenko, T. O. Metz, J. G. Pounds, K. D. Rodland, A. K. Shukla, R. D. Smith, K. M. Waters, J. E. McDermott, and B.-J. Webb-Robertson, "A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments.," *Proteomics*, vol. 13, pp. 493–503, feb 2013.

[136] S. Pietilä, T. Suomi, J. Aakko, and L. L. Elo, "A Data Analysis Protocol for Quantitative Data-Independent Acquisition Proteomics.," *Methods in Molecular Biology*, vol. 1871, pp. 455–465, 2019.

[137] J. Aakko, S. Pietilä, T. Suomi, M. Mahmoudian, R. Toivonen, P. Kouvonen, A. Rokka, A. Hänninen, and L. L. Elo, "Data-independent acquisition mass spectrometry in metaproteomics of gut microbiota - implementation and computational analysis.," *Journal of Proteome Research*, nov 2019.