Freie Universität Berlin

Computational Methods and Statistical Models for improving Quantification Accuracy of High-throughput Omics Data

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
eingereicht beim Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Martina Fischer

Berlin
April 2018

Betreuer:	PD Dr. Y. Bernhard Renard
Erstgutachter:	PD Dr. Y. Bernhard Renard
Zweitgutachter:	Assoc.-Prof. Dr. Josselin Noirel
Drittgutachter:	Assoc.-Prof. Dr. Laurent Gatto
Tag der Disputation:	19. Oktober 2018

Abstract

The advances of high-throughput technologies in genomics and proteomics have revolutionized the biological research field. The increased resolution capabilities have strengthened the focus on quantification analysis and the massively parallel nature of the instruments ultimately enables quantification at a genome- and proteome-wide scale. The field has experienced an explosion of applications, which is accompanied by many computational challenges and a strong demand for novel quantification analysis tools.

Quantitative workflows are complex, involving many steps from sample preparation to data acquisition, processing, and the inference of quantitative estimates. Multiple sources within this process cause different bias in quantification. The goal to retrieve best estimates of the true underlying quantities from a sample remains a difficult task. It leaves a constant need for new method development to reduce systematic errors and biases in the data.

In this thesis, new statistical and computational strategies are presented to improve the quantification accuracy of data types from high-throughput omics applications. The aim is to correct biases and minimize the overall variance; therefore understanding the potential error sources and data characteristics is essential. One focus in this work is to identify biases and solutions which are common to different omics workflows and data types. A second aim is to assess the statistical confidence of resulting quantitative measures. A general lack persists in high-throughput quantification on how to measure and report reliability of quantitative estimates, especially in quantitative proteomics research. A lot of knowledge on statistical methodology for large-scale data analysis has been acquired in the microarray era. Generally, independent of underlying technologies, final quantitative values often exhibit similar properties from a statistical point of view. Hence, a strong potential lies in revealing parallels between different omics fields and in the transfer of established statistical concepts. In addition, however, it is equally important to precisely integrate and account for specific data characteristics and technique-induced biases. Overall, quantitative analyses are highly heterogeneous and one-fit-all methods are not appropriate.

The contribution of this thesis comprises three major projects which address different biological objectives and different data types based on three quantitative high-throughput techniques. New approaches concerning data pre-processing, quantification inference and resolution, and quantitative comparisons, are introduced.

In the first project, affinity purification is coupled with mass spectrometry (AP-MS) aiming to identify protein-protein-interactions. Here, quantitative counts of proteins obtained from pull-down experiments are compared with counts from negative controls in order to separate true interactions from false-positive hits. Current methods for AP-MS analysis mainly rely on scoring systems to rank potential interaction proteins. However, uncertainty on where to set the cutoff score remains for candidate selection and also no estimation on the expected number of false positives is given. Statistical pre-and postprocessing is an underrepresented topic in AP-MS analysis. A thorough statistical framework is introduced, which can embed any scoring method and enables to replace scores by statistical p-values using a permutation principle. In addition, a two-stage poisson model adapted from RNA-Seq to AP-MS data is proposed as an alternative method for assessing interactions. For pre-processing, different normalization methods and statistical filtering, adjusted to AP-MS data, are investigated. Several experiments demonstrate how the number of true interactions can be significantly increased while controlling a false detection rate.

The second project concerns the accurate inference of protein quantities. In mass spectrometry, measurements are assessed at the peptide spectrum level. Although all peptide spectra assigned to the same protein are assumed to share similar intensity values, in fact, a substantial heterogeneity exists due to random and systematic biases. Clever summarization strategies are needed. Current methods rely exclusively on peptide quantitative information. However, this work hypothesizes that a wealth of other peptide features are available that reflect spectra reliability. Several features are correlated with the observed variance heterogeneity and their relation to quantification accuracy in the spectra is investigated. As a result, a new peptide-to-protein summarization method is presented, referred to as iPQF (isobaric Protein Quantification based on Features), which integrates peptide features with quantitative values for protein quantification. As a novelty, peptide spectra are weighted according to their feature reliability. Extensive evaluation of iPQF in comparison to nine other summarization methods proves the added value of feature information to enhance protein ratio accuracy.

NGS-based quantification equivalently relies on shotgun measurements and requires summarization strategies. The third project focuses on accurate inference of quantities on strain level in NGS-based metagenomics data. Specific challenges arise on strain level due to the presence of highly similar reference sequences, which underlie a strong quantification bias due to shared read mappings. There is increasing demand for analyzing microbial communities at higher resolution, but only few tools provide quantitative profiling beyond species level. In this work, DiTASiC (Differential Taxa Abundance including Similarity Correction) is presented as a novel tool for abundance estimation and also for differential analysis applicable down to exact genome level in metagenomics samples. A new generalized linear model framework is introduced for the resolution of shared read counts, additionally including an error term to assess abundance estimation uncertainties. In a new statistical approach, the abundance variances are integrated to infer abundance distributions for differential testing sensitive to strain level. Performance evaluations on latest benchmark studies show highly accurate abundance estimations down to sub-strain level and improved detection of differentially abundant taxa.

Altogether, these three contributions improve the current repertoire of computational methods in high-throughput quantification of omics data. This work intends to raise awareness for the complexity of quantification analysis. On one side, it highlights the comprehensive usage and transfer of established statistical concepts across different omics techniques. Equally, it aims to emphasize the importance to specifically address underlying data characteristics and the need to offer individualized strategies in order to achieve high quantification accuracy.

Acknowledgements

First of all, I want to thank my supervisor Bernhard Renard for all his advice and support during my PhD years. I highly appreciated that he always made an effort to find time for meetings when any problems occurred. I am thankful for his openness for also controversial discussions, valuable exchanges of ideas and also letting me pursue new but unplanned directions such as my iPQF project.

Furthermore, I am grateful to Josselin Noirel and Laurent Gatto for willingly accepting to review my thesis. In addition, I want to thank all my co-authors and collaborators over the past years: Samuel Wagner, Susann Zilkenat and Roman Gerlach for introducing me to the AP-MS world and their enthusiasm to run experiments for us. I also want to express special gratitude to Laurent Gatto for his interest in my work and all his effort in integrating my iPQF code into his MSnbase package. I was very pleased that my work was given the opportunity to live on in such an established package. Many thanks also go to Benjamin Strauch for great team work in our DiTASiC project, it was a pleasure to combine the best of our analytical and coding skills to create a wonderful new tool.

Special credit also goes to Axel Benner and Manuela Zucknick from my previous position at DKFZ for laying the foundation of my statistical knowledge and teaching me many skills for data analysis. Further, I like to thank different people within the Robert Koch Institute: Jörg Döllinger for fruitful discussions, Woijtek for selfless help with the computational infrastructure and Stina for being the most lovely master student to supervise in the past. And of course, one of the biggest thanks is devoted to my colleagues of the MF1 bioinformatics group for creating such a relaxing, helping and fun atmosphere. I particularly thank Thilo for proofreading my thesis and our joint work in the MiMB book chapter, for all his advice and his great motivating skills. Moreover, I´m sending many thanks to my fellow PhD students Kathrin, Christine, Tobias and Vitor, Martin and Franzi in the past, for many inspiring scientific and non-scientific chats and their always helping hands in case of technical struggles. Deep gratitude also goes to Mathias as my hero in code trouble shooting and for making seemingly hopeless external tools run. I also want to thank Robert and Carlus for the last good time in our sunny office.

Last, but not least, I want to express my deep gratitude to my parents and to the wonderful friends I have - who stand with me in good and bad times and who strongly believe in my skills when I tend to doubt. My parents who always support whatever zig-zagging paths I chose and who still take care of my german comma rule setting or formatting issues till this thesis. My biomathematics family who sticks together for almost 10 years, encouraging each other in all ways, with special thanks to Janina and Anna for their strong support. Thanks to my long-term flatmates and all my Berlin friends for making my life so colourful. My climbing friends for all our adventures and pushing one’s mind at heights as ”there is no choice but to grow..”

Contents

1. Introduction **1**

1.1. Quantification of high-throughput omics data 1

1.2. Quantification biases: Challenges and common concepts between different fields 9

1.3. Differential quantification 14

1.4. Thesis objectives 18

2. APMS-WAPP: pre- and postprocessing of AP-MS data **23**

2.1. Study of protein interactions by AP-MS 23

2.2. Statistical framework for AP-MS data 26

2.3. Experimental setup 31

2.4. Results 33

2.5. Discussion of results 39

3. iPQF: a new peptide-to-protein summarization method **42**

3.1. Peptide-to-protein summarization 42

3.2. Features of peptide spectra 44

3.3. iPQF algorithm 45

3.4. Experimental setup 47

3.5. Results 50

3.6. Discussion of results 54

4. DiTASiC: quantification on strain level in metagenomics data **57**

4.1. Metagenomics profiling 57

4.2. DiTASiC workflow 60

4.3. Experimental setup 65

4.4. Results 66

4.5. Discussion of results 71

5. Summary and outlook **74**

5.1. Outlook 76

A. Appendix **81**

B. Appendix **109**

C. Appendix **133**

Bibliography **156**

1. Introduction

1.1. Quantification of high-throughput omics data

The identification and quantification of components in a sample is fundamental in virtually all branches of biological research. The questions which components are present in a sample, how much of them, and how the quantities change between different samples, are main objectives in data studies. Most state of the art technologies allow both qualitative as well as quantitative measurements, and often identifications are required as prerequisite for subsequent quantitative assessments. However, pure qualitative analyses are not sufficient to capture the complexity of biological systems, and quantitative investigations are crucial to achieve deeper understanding of the underlying processes.

Genomics and proteomics are two major research fields which thoroughly study all levels from gene to phenotype level. Quantitative genomics gives insight into the functional elements of the genome. On cell level, gene regulation and signaling pathways are investigated, while overall transcriptome dynamics during development or in a disease state are studied on a global scale. Therein, differential gene expression research has attained extreme popularity. Quantitative proteomics addresses the gene products present in a cell state and enables a direct picture of protein directed processes, revealing protein dynamics and function, protein interactions as well as their subcellular distributions [1, 2].

This chapter is devoted to the quantification process of high-throughput genomics and proteomics data. The corresponding technologies are introduced and aspects of the quantification process from the view of both fields are presented.

Established workhorses for proteomics analyses have been for a long period the two-dimensional gel electrophoresis (2-DE) and the western blot. However, the methods have been shown to meet limitations concerning low resolution and quantification range, frequently only displaying the most abundant proteins [3]. The popular western-blot method tends to be more qualitative and is commonly only used for semi-quantitative analysis. In the genomics field, hybridization based approaches, such as custom-made or commercially high-density oligo microarrays, are well-established quantification methods, high-throughput and inexpensive [4]. Yet, these arrays require prior knowledge of the target sequences and have a limited detection range due to both high background levels and signal saturation.

In the last decade, new developments of high-throughput technologies have emerged and have revolutionized the biological research field. The fields of genomics and proteomics have experienced an enormous increase in sensitivity, speed, accuracy and overall resolution by the advances of novel instrumentation. The increased capabilities on high resolution have significantly shifted the focus towards quantitative analyses which has become central in studies today. Especially the massively parallel sequencing techniques have empowered to address quantitative questions at a genome- and proteome-wide scale.

For proteomics, mass spectrometry (MS) is the most powerful approach present with its main platforms available being Orbitraps, QTOF instruments and triple-quadrupole instruments. All are characterized by highly increased sensitivity and data acquisition speed. The new platforms enable measuring hundreds to thousands of proteins (depth of 5000-10000) from a given proteome of any biological system within a single experiment [5].

The breakthrough of next-generation DNA-RNA sequencing technologies, referred to as NGS, has further enlarged the reach of proteomic research to practically any species. The new sequencing era enables to survey entire genomes, transcriptomes and epigenomes within a single sequencing run by deep-sequencing techniques [6]. It directly assesses the DNA or RNA sequence and provides single base resolution of the target sequence. Thereby, no prior sequence knowledge is required and an unbiased detection of also novel sequences in the sample is possible. Quantification has changed from signal intensity-based approaches in arrays to the form of discrete digital counting in NGS technologies. As a result, high sensitivity and a large dynamic range of abundance levels can be achieved as no physical saturation or upper limit exists. Another advancement is the large generation of sequence data at decreasing costs, with prospects of up to one billion short sequences per instrument run. A variety of platforms are on the market, to name Illumina, Pacific Biosciences, Oxford Nanopore technologies, Thermo Fisher Scientific and Roche, which vary in sequencing technology and specialize either on low or high output (from 100 Mb to 35 Gb) with correspondingly long or short sequence fragments (from 50nt to 120kb) [7].

Given high resolution in combination with high parallelization enables conducting analyses at all biological levels. Target objects range from small peptides in MS applications, gene fragments or splice isoforms in NGS, to the study of all transcripts or respectively all proteins present in a cell or any tissue sample. In addition, processing large numbers of samples simultaneously enables the investigation of samples with different conditions as well as the examination of whole communities comprising different organisms.

The past years have seen an explosion of MS and NGS based applications in all areas of the life sciences [8–11]. In this work, three applications will be addressed in more detail. Two popular areas in MS-based quantitative proteomics are protein interaction studies, in which affinity approaches are coupled with mass spectrometry, and second, abundance studies of whole proteomes using techniques based on stable isotopes. The new advances of NGS to measure genetic material directly from environmental samples and to assess hundreds of different genomes within a sample gave rise to the field of metagenomics [12]. One of its objectives is to unravel composition and change of microbial communities, discover novel species and to understand their interactions with the host and environment.

A broad range of technical and experimental variations regarding mass spectrometry and next-generation-sequencing protocols have been devised and are adjusted to the various applications. The fast progression of the technical field is further accompanied by a rapid development of computational methodology. The rate and volume of the new data generation creates a strong demand for data storage strategies, data processing and efficient analysis methods. Analysis methods need to address the complexity of the data sets as well as to account for the individual characteristics and shortcomings of the instruments. To sum up, it can be said that the era of high-throughput technologies have brought enormous statistical and computational challenges to the field of bioinformatics.

A large variety of computational methods has been developed in the last decade and most effort was initially concentrated on identification tools. The identification of components in a sample is a prerequisite for the process of quantification. NGS and MS techniques share the concept to identify resulting fragments by matching reads or spectra, respectively, against a reference database [13, 14]. Here, already a few unique fragment mappings are sufficient to identify the full target sequence and to confirm its presence in the sample.
In contrast, quantitative analyses are more sensitive, here a failed detection of single fragments or a false assignment already causes a bias in quantification. For accurate quantification of targets in a sample, a precise resolution and a highly accurate back-tracing of obtained fragments to their sequence of origin is crucial. Quantitative workflows are in need of several processing steps, on experimental as well as on computational analysis side, to reduce quantitative biases. Generally, quantification processes are strongly dependent on the given technology, as each is characterized by specific systematic and non-systematic error profiles impacting quantification (see section 1.2). Overall, as underlying techniques and resulting data sets are very heterogeneous, there is no easy or one-fits-all method in quantification research [15–17]. A strong demand for the development of specified quantification approaches persists, which more precisely account for characteristics and potential biases in the data to improve accuracy in quantification.

Quantitative data can have different forms. It is commonly classified into absolute and relative quantification of compounds in one sample or into quantitative comparisons between two or more samples. Absolute quantification aims to determine the actual amounts or concentrations of the compounds considered in a sample, measured e.g. in ng, ml^{-1} or discrete numbers of copies. Whereas, in relative quantification, the quantity of compounds is described in relation to the amount of all other compounds present in the sample by setting the total abundance to 100%. In quantitative comparison studies, the relative abundance change of a compound between different conditions or samples is assessed and commonly fold-changes are reported. Naturally, absolute quantification comprises relative quantification, as relative ratios can always be inferred from known absolute amounts. However, deriving precise absolute quantities by high-throughput techniques is very difficult due to several experimental impacts, while some biases can be countervailed by relative quantification (refer to section 1.1.2 and 1.2).

1.1.1. Next-Generation-Sequencing

Genome sequencing started with the automated Sanger sequencing method, which has dominated the field for almost two decades and is referred to as *the first generation*. However, it is considered rather low-throughput and was not effectually designed for quantification purpose. Instead hybridization based arrays dictated the quantitative genomics field [18]. With next-generation-sequencing technologies the previous limitations are overcome, providing high resolution and capabilities for accurate quantification analyses [6, 10].

The new sequencing process comprises a number of steps [7], which can be broadly grouped into 1) template preparation, 2) sequencing, 3) imaging, and 4) data analysis. The combination of specific protocol parts characterizes different sequencing technologies and adaptions to different data objects. In a first step, the DNA to be sequenced is fragmented into a library of smaller fragments and synthetic adapters are added to each end. These adapters are universal and specific to each platform and are used for fragment amplification in the next step. The reaction takes place in situ and clusters of amplified DNA for each fragment are formed either on beads or on a glass microfluidic channel. All fragment clusters are then sequenced in a massively parallel fashion, thus, millions of reactions take place simultaneously in one run. Key feature of the sequencing technique is to utilize fluorescence to achieve a base-by-base view of the target sequence. For this, most sequencers follow the principle of sequencing-by-synthesis (SBS), in which single bases are detected during incorporation into growing DNA strands. Each fragment sequence is consecutively resynthesized and with each incorporation step a unique emission signal of the corresponding base is released. The series of signals determines the resulting sequence *read*, which derives from a single DNA fragment. The length of the read depends on the number of sequencing cycles. In the setup of paired-end sequencing, the instrument sequences from both ends of a fragment and generates two reads of opposite direction.

The total number of reads for a gene, transcript or whole genome can be directly correlated to its abundance level, as each read is directly related to a single library fragment in a one-to-one relationship. The main challenge in quantification of next-generation-sequencing data concerns how to convert the output of reads to reasonable abundance estimates. Three steps need to be pursued: (1) identification of the origin of the reads by applying assignment approaches, (2) counting the number of classified reads per target sequence, and (3) potential corrections of the initial count estimates to infer accurate quantitative estimates.

Read assignment can be conducted in different ways and a large number of tools are available [19]. Methods can be divided into reference-based and reference independent approaches. In case the sample comprises known and previously sequenced genomes, referred to as reference genomes, mapping the reads against a reference database using sequence homology is a common approach [20]. This can be realized either by full alignment approaches, which yield the exact matching base positions of a read to the corresponding reference sequence, or by applying so called pseudo-alignment approaches [21]. The latter is based on fast k-mer hashing strategies to assign reads to reference sequences without executing exact base alignments, which significantly accelerates read assignments in large data sets. For quantification purposes, pseudo-alignment approaches are often considered sufficient as inferring the reference sequence of origin is necessary, while the exact mapping position is not. Reference independent approaches use unsupervised clustering to group similar reads or conduct an assembly of reads into genes or draft genomes [22]. These methods do not rely on homologies to known sequences, thus enable to detect and quantify novel genomes as well.

In summary, an appropriate read mapping approach has to be chosen according to the application and the target of interest. It is very important to note, that the selected read assignment approach significantly determines the read count measure. The read count refers to the initial abundance estimate for target transcripts or taxa, and thus has a crucial impact on overall quantification accuracy.

The second fundamental point to consider is that each data type comes with specific characteristics and individual challenges for quantification. For instance, the analysis of RNA expression levels by an RNA-seq experiment faces the complexity of the transcriptome and the fact that not all transcripts are known. In the mapping step, the reference genome only serves as a proxy for the transcriptome. Challenges arise from the exon intron structure as well as from different transcript isoforms due to alternative splicing events. One approach is to infer gene expression by counting the number of reads mapped to the exon sequence regions and normalize by the total sequence length of unique exons [23]. In the setting of metagenomics, one encounters vast sizes of data sets, a generation of millions of reads, which are derived from different genomic origins [24]. Here, users are also confronted with the challenge that a large majority of microbes are still unidentified and largely not represented in reference databases. Additionally, in order to resolve a complex mix of microbial genomes present in a sample, reads need to be mapped to thousands of potential reference sequences simultaneously. This gave rise to new sequence homology approaches using short k-mers, which are more runtime and memory efficient, to assign taxonomies [21]. Moreover, many different aspects of a microbial community can be investigated, ranging from the gene to the functional level, as well as considering taxonomic levels at different resolution. Further, different abundance parameters exist, differentiating between absolute and relative abundance and copy numbers of a taxon or gene [22]. Therein, absolute abundances are difficult to infer from only sequence data and need to be combined with density techniques such as flow cytometry [25] or quantitative PCR [26].

The main focus in this work is on metagenomics data aiming to describe the composition of a community by relative amounts of taxa present in a sample (see Chapter 4). In a standard approach, the proportion of classified reads that map to a reference sequence is used to define taxa relative abundance; however, challenges arise for higher resolution levels and correction strategies for initial read counts are needed.

Most notably, quantification in genomics has gone through a significant change, from former signal intensity based approaches to discrete digital counting of sequence reads. This enables new levels of resolution and improved sensitivity in quantification, as the number of sequencing reads can be unlimitedly increased or decreased according to sample processing. The digital count nature of the sequence data facilitates abundance comparisons between samples and whole populations, as the need to correct for differing background signal levels is not given. Thereby, most important for the computational analysis field is that the wealth of discrete statistical models, designed for quantitative comparisons, becomes applicable to comparative genomics.

1.1.2. Protein quantification techniques

A large variety of different quantitative proteomic protocols exists. They share the following steps of a typical proteomic bottom-up experiment: (1) sample preparation, (2) protein digestion, (3) peptide separation, (4) ionization, (5) mass spectrometry, and (6) data analysis [27].
First, proteins of a sample are purified and subsequently digested into peptides by using proteases such as trypsin, which cleave proteins at their arginine and lysine residues. The peptide mixture is injected to a high performance liquid chromatography (HPLC) column and peptides elute according to their hydrophobicity. At the end of the column, peptides are ionized by a process called electrospray ionization and are further transferred into the mass spectrometer for analysis. As a result, the instrument generates mass spectra. At first, MS1 spectra report signal intensities of peptide ions at a mass-to-charge (m/z) scale. Tandem MS instruments then proceed to select the high intensity peptide ions, also referred to as 'precursor' ions, for further fragmentation. This results in MS2 spectra exhibiting a series of different ion types, among them b- or y-ions of a specific peptide with mass differences indicating the single amino acids present. The peptide sequence can be correspondingly inferred from the MS2 spectrum by applying database search strategies.

Quantification measures can be retrieved either from the MS1 or MS2 spectra level. Generally, we can distinguish between direct quantification via signal intensity based approaches and indirect inference using spectrum count approaches.

Each peptide signal in the MS1 spectrum effectively consists of a cluster of intensity peaks. Peak picking algorithms are applied to detect the peak patterns of each peptide. The area under the peaks is considered to be proportional to the peptide abundance and can be calculated by integrating all assigned peak intensities [28]. The corresponding protein abundance can be subsequently inferred by averaging the total intensities of all spectra matched to the protein. The latter step is referred to as peptide-to-protein summarization and is addressed in more detail in this thesis (see Chapter 3).
Spectrum count approaches are based on the assumption that with increasing protein amount, the number of MS2 spectra identifying a protein will increase proportionally. The simplest protein abundance metric refers to the number of distinct peptides identified. However, using the number of actual peptide-to-spectrum

matches (PSMs) is recommended as a more reasonable abundance measure. Further, so called protein abundance indices (PAIs) also take into account that larger proteins yield more measurable peptides compared to smaller ones [3]. The basic PAI normalizes the number of identified peptides by the number of theoretically observable peptides [29]. A more robust version, the exponentially modified PAI is defined by $emPAI = 10^{PAI} - 1$ and is a popular measure in many software suites [30]. In other variations of the PAI, the spectrum counts are normalized for protein length (SAF index) [31], and additionally normalized for the sum of all protein abundances in the sample (NSAF), or normalized for molecular weight of the protein [32]. Generally, the spectrum count methods rely heavily on the quality of the peptide identification step. Incorrect identifications have a direct impact on the quantification quality and also the number of PSMs per protein affects the accuracy. Low PSM numbers make meaningful quantitative comparisons more difficult. The described quantification measures are commonly used in the label-free approach. As the term label-free implies, quantities are purely extracted from the MS scans without introducing any form of labelling in the experiment. Label-free approaches allow absolute and relative quantification in single samples as well as non-limited sample processing and enable all combinations of quantitative comparisons between samples.

Overall, the field of quantitative proteomics can be classified into the two main families of label-free and label-based methods [28]. An application using label-free quantification is presented in Chapter 2 and Chapter 3 is based on data derived from labelling techniques.

Label-based methods enable direct comparison of two or more proteome states within the same analysis and are by design ideal for relative quantification assessments. Key of the labeling concept is to use stable isotope substitutions to introduce a mass difference between labeled and unlabeled peptides or between differently labeled peptides. The peptides maintain an identical behavior in an MS experiment as the isotope label only induces a shift at the m/z scale. As a consequence, quantitative comparisons of peptides from different samples can directly take place within an MS spectrum. The family of labeling strategies is further subclassified according to the label incorporation process and whether quantitative information is retrieved from the MS1 or MS2 spectrum level. Labeling methods based on MS1 level are further subdivided into *in vivo* (metabolic) and *in vitro* (chemical and enzymatic) approaches [28]. One prominent *in vivo* approach is stable isotope labeling of amino acids in cell culture (SILAC) [33]. Here, proteins are labeled inside living cells in culture during growth and replication. In contrast, in *in vitro* strategies, labeling takes place after the cell lysis and chemical or enzymatic steps are applied to couple the label to the peptide. Different chemical labeling options exist, such as the O^{18} method [34] or the ICAT method labeling reduced cysteines [35]. Both methods result in a 4 Da or respectively 8 Da mass shift between a *light* (unlabeled)and *heavy* (labeled) peptide in an MS1 spectrum allowing differential quantification assessment of two samples.

The second category of isotope labeling methods applies different isobaric mass tags to the peptides of each sample after protein digestion. It enables a multiplexed analysis, measuring diverse samples simultaneously within one experiment run. Here, quantification takes place in the MS2 spectrum. The quantitative information is concealed in so called reporter ions which are released after fragmentation. The labels are designed to contain different parts, one reactive tag which can bind to the peptide, and commonly a spacer group and a reporter group. The key idea is that the masses of the spacer and reporter group are chosen to balance each other to ensure exactly the same mass for all labels. Consequently, the same peptide, labeled differently for different samples, will be represented by only one intensity peak in the MS1 spectrum. However, labels are precisely fragmented in the subsequent MS2 step. The intensity signals of the reporter tags will provide the quantitative information at the known reporter masses, while fragment ions of higher m/z values serve for peptide identification. Among the most popular methods is iTRAQ, according to *isobaric mass tags for absolute and relative quantification* [36], and the TMT method referred to as *tandem mass tag* [37]. iTRAQ labels allow multiplexing up to eight samples, while TMT isobaric tags are available till 6-plex forms. The expected output of an *n-plex* are *(n-1)* peptide ratios per MS2 spectrum as one label is usually defined as reference.

Altogether, all presented proteomic quantification techniques allow relative and absolute quantification and hold specific advantages and disadvantages correspondingly [3,17,28]. Label-free and label-based approaches significantly differ when considering relative quantification of proteomes from different conditions. Label-free methods essentially relate to the comparison of peptide abundances extracted from different MS runs, while label-based methods provide direct quantitative comparisons of proteomes within the same run. With it, labeling approaches clearly benefit from the fact that the same experimental conditions are assured for the differently labelled samples and experimental biases can cancel each other as a consequence. Further, the mass tags of the labels clearly define where the quantitative information is found. In contrast, in label-free approaches, a main requirement is to first identify and associate the same peptide from different experimental runs, quantify, and pre-process it separately, before a quantitative comparison can be conducted. Thereby, different biases can arise in each sample process independently. Additionally, the method can be challenged by the fact that not every peptide is necessarily selected for fragmentation in each run or drops below the background signal level. Thus, accuracy of relative quantification in label-free methods is highly dependent on reproducible sample processing and peptide separation. High-resolution instruments are advantageous to enhance collective peptide detection in the different runs. However, the method holds the advantage to be simple as no additional labeling preparation step is needed and allows an unlimited number of samples to be compared. A high dynamic range and resolution of abundance fold changes up to 60:1 is promised in label-free approaches in comparison to 20:1 stated for label-based ones [30,38].

Assessment of the absolute quantity of a protein poses an even greater challenge, as the MS response for each peptide is influenced by many different factors (see also next section) [17]. Among others, not all peptides can be captured by MS technologies, e.g. some fall beyond the analyzable mass range or cannot be retained on the chromatographic column. But strategies are available to infer absolute quantities. For example, adding an isotope label to a reference peptide of known absolute quantity in the iTRAQ method is one strategy. In label-free methods, taking the sum of all peptide intensities observed for a protein and divided by the number of theoretically expected peptides, is used as an approximation for absolute protein concentration [39]. Equivalently popular is the 'best-flyer' approach, based on the assumption that few unique proteotypic peptides with strong signal are present for each protein. The average of the three most intense signals given for a protein is computed [40].

In summary, a broad set of high-throughput techniques are available for quantification in genomic and proteomic samples. A wealth of different experimental protocols has been introduced to specifically target various applications. The previous sections have demonstrated many different forms on how to quantify transcripts and peptides, for genome- and proteome-based analyses, respectively. Naturally, all these methods come with individual characteristics, strength and weaknesses, however, also share common aspects in the quantification process. The next section specifically focuses on arising quantification biases and the associated computational challenges.

1.2. Quantification biases: Challenges and common concepts between different fields

The objective of accurate quantification in proteomics as well as genomics means a multi-disciplinary challenge. It requires a joint effort of the fields, chemistry, physics, biology, computer science and statistics. Within the process, sample preparation and data analysis are the steps which can be influenced most by the individual scientist conducting the experiment. The importance of data processing is often underestimated and, with regard to the fast technical advances, the development of statistical data analysis methods is lagging behind. The objectives of this thesis are therefore dedicated to provide approaches of data processing to solve the computational challenges to improve quantification accuracy in high-throughput omics applications.

Particularly in large-scale data scenarios, data sets from genomics and proteomics are often noisy or incomplete. Quantitative workflows in either MS-based proteomics or NGS-based genomics are quite complex, involving sample preparation, data acquisition, processing of raw data, and final inference of quantitative estimates. Multiple sources within this process cause different bias in quantification. It is necessary to understand the potential error sources and characteristics of each resulting data type. One important conclusion drawn in the past years in the research field of omics data quantification is that no 'one-fits-all' method exists and it is also not reasonable to define one [17].

In this section, different biases influencing quantification workflows are discussed. Therein, one aim is to particularly highlight biases which are common to different quantification techniques. The importance to correct for systematic errors as well as individual biases is emphasized. Corresponding bioinformatics approaches are presented and especially concepts which can be transferred between fields are investigated.

Several factors impact precision and accuracy of quantification on the technical side, which are specific for either NGS or MS data acquisition. Here, a summary is given to depict the large number of experimental influences in the two technologies. Pre-processing of raw data is the first important step to control technical biases and to ensure a certain level of data quality.

In proteomics-based quantification, a precise extraction of the peptide signal intensity peaks from the raw spectra is the first essential step. This step can be challenged by the presence of double peaks due to peptide isotope overlaps, surrounding artefactual spectral peaks and the difficulty to detect and separate low signal peaks from noise [28]. Good performance of a peak picker algorithm is crucial as errors are carried forward through the entire analysis with substantial impact on the final quantification result. Particularly in isobaric labelling approaches, co-elution of peptides is a well-known effect to distort quantification [41–43]. Other ions additional to the peptide precursor ion of interest are co-isolated within the same selection window for further MS/MS analysis. As a consequence, generated reporter ions, which provide the quantitative information upon fragmentation in the MS2 spectrum, are indistinguishable. In most cases, co-isolated ions are not biologically meaningful and hold constant abundances. Hence, this causes a ratio compression towards a fold change of 1 for an up- or downregulated peptide of interest. This effect can be severe, especially for low abundant peptides, and can result in heavily biased quantitative peptide ratios. Different solutions to reduce peptide interference have been proposed [44–46]. One computational approach is to conduct the analysis using only peptides with most extreme fold changes [47]. A promising technical strategy is the application of triple-stage mass spectrometry (MS3) by adding a third fragmentation step and corresponding studies by Ting *et al.* [48] demonstrated significant reduction of the interference effect. Quantification accuracy is further influenced by experimental factors such as varying ionization, fragmentation efficiency, or by limited labelling efficiency [49]. An important computational step concerns the step of baseline correction in which the background intensity is estimated and peaks are adjusted by subtracting the background correspondingly. This step has potentially strong impact on final quantities and requires careful execution [28]. In contrast, count-based quantification is free of complex intensity processing steps and is primarily dependent on the number of generated spectra and their accurate identification. Here, experimental factors are scan speed, peak width, retention time stability, and protein length itself which together determine the number of detectable peptides. In summary, diverse pre-processing methods need to be applied dependent on the protocol and techniques used. A variety of computational tools are available for processing and correction of raw quantitative MS data. Some tools are also specifically designed to adjust for technical biases known in iTRAQ, SILAC or label-free approaches [50–55].

In NGS based quantification, equivalently, experimental protocols and computa-

tional processing affect the difference between true abundances and the actual observed amount of reads. Some biases are inherent due to sample storage, DNA extraction, and the library preparation process [6]. In particular, errors introduced during DNA fragmentation and within the PCR cycles can cause a non-uniform representation of sequencing reads. It has been shown that sampling is frequently distorted for genomic fragments which exhibit low or high GC content [56]. Also the generation of duplicate reads can arise from biases in the PCR step. The resulting non-uniform read coverage can have a significant impact on final abundance estimates and requires correction methods [22]. The sequencing process itself is prone to different sequencing read errors dependent on the technology used. Common are base substitution errors or insertions and deletions due to homopolymer and carry-forward errors [57]. Diverse error correction algorithms are available and all are based on the assumption that errors occur infrequent and random, while the majority of reads will call a base at a specific position correctly [58–62]. Erroneous sequence reads potentially challenge the subsequent read mapping. However due to advanced mapping tools, they have been reported to have overall small effects on final abundance estimates [63]. Overall, bioinformatics processing methods are applied to raw read sequence data to establish a high quality data level for subsequent read mapping. Methods typically comprise trimming of reads with low quality bases at the ends, and different filtering steps to remove overall low-quality reads, sequence adaptors, contaminants as well as unwanted host sequences, or potential duplicate reads [22].
However, filtering steps always present a compromise between data quality enhancement and the risk of removing potentially important information. Wrong filtering can also cause biases: in case of duplicate reads, for example, reads might be PCR artefacts or might stem from abundant and deeply sequenced organisms in a metagenomics sample. Likewise in proteomics experiments, redundant spectra, referring to several MS/MS events received for one peptide, can either result from a highly abundant protein or arise due to a selection bias. A distinction is generally difficult and filtering needs to be highly sensitive in both applications.

Further, in both technologies, the use of counts as abundance measure is strongly dependent on high quality identifications of spectra or reads, respectively. Identification tools commonly report read assignments or PSMs along with confidence scores. Again, filtering plays a crucial role to identify all reliable identifications and simultaneously discard false-positive hits. However, the challenge remains of defining an optimal filter threshold to also capture low supported but abundant proteins or genomic sequences . Further, selection of a threshold also depends on the objective; filtering strictly for high-scoring PSMs, for example, can be beneficial to detect significantly smaller abundance changes as reported by Cooper *et al.* [64].

Generally, quantitative measurements are assessed only at the peptide spectrum and genomic read level, and inference strategies are needed to receive actual quantities of proteins, transcripts or taxa of interest. As described, the most straightforward aggregation method is to simply count reads or peptide spectra. Considering intensity-based measures, all peptide spectra originating from the same protein should ideally possess similar intensity values. In reality, however, significant signal heterogeneity is observed at the peptide spectrum level. This heterogeneity arises as a consequence of previously described random and systematic biases or due to biological influences such as modifications or present isoforms [65,66]. Hence, peptide intensities of a protein are not necessarily of equal quality. Combining peptide measures to one final protein abundance estimate, or selecting the 'best' peptide spectra to be used for protein quantification, poses a challenging bioinformatics task. For solving this problem, several peptide-to-protein summarization methods have been proposed [17] and a new solution is presented in this work (refer to Chapter 3). Frequently, standard statistical approaches such as the mean or the median are used to retrieve an average intensity measure of multiple spectra assigned to a protein. More sophisticated approaches integrate the fact that low intensity peptides suffer from larger variances due to decreased signal-to-noise ratio compared to high intensity peaks. Corresponding methods for weighting peptides according to their absolute intensity, as well as for low intensity peptide filtering, and for variance stabilization have evolved [66–69]. Further, tailored noise models accounting for specific error structures and underlying ratio distributions are available [47,70]. Overall, the quantitative aggregation of the shotgun measurements is an important and often neglected step in the quantification process, which has an immediate impact on resulting quantification accuracy.

Moreover, the summarization process is often challenged by the problem that certain reads and spectra map equally well to different genome or protein reference sequences and their origin cannot be inferred explicitly. These are defined as shared reads or shared spectra respectively. The ambiguity issue arises for MS as well as NGS data as a consequence of present protein, transcript or genome sequences sharing sequence similarities. Thereby, the read length is also an influential factor as short reads complicate homology detections, while long reads help to reduce the multi-matching problem when high sequence quality is given. Same is observed according to low and high precision in MS/MS spectra resolution. Generally, standard mapping tools do not resolve shared read assignments. Using pure mapping counts may result in biased quantification, specifically in abundance overestimation or positive abundance calls for similar but absent sequences. Computational correction algorithms are required for the resolution of shared counts (refer to the contribution of this thesis in Chapter 4). In a simple heuristic proteomics approach, shared peptides are assigned to the most detected protein within the possible set [2](or shared counts are proportionally distributed according to the number of unique spectra identified for each individual peptide of a protein [71]. An equivalent strategy is applied to NGS data in the metagenomics field, multiply mapped reads are heuristically assigned to reference genomes according to uniquely mapped read proportions [22,72]. More sophisticated models [73,74] evolved which integrate the genome sequence similarities to achieve more precise count resolution. In the RNA-Seq field, kallisto with its EM algorithm became popular to correct ambigious read counts for transcript abundance estimation [75].

In intensity-based quantification, however, shared peptides can cause severe quantification errors and are more difficult to resolve. Shared peptides show significantly increased intensity signals in the MS1 spectrum in comparison to uniquely mapped peptides of the same protein. The MS1 signal of a shared peptide effectively reflects the quantitative signal sum of all precursor proteins sharing the peptide [28, 49] . Likewise in isobaric labelling experiments, the extracted ratio of a shared peptide in the MS2 spectrum refers to a weighted average ratio with weights according to the absolute quantities of proteins involved [76]. Hence, a strong bias is introduced if combining unique and shared peptide intensities in the summarization step. A control step before peptide aggregation is crucial and commonly shared peptide signals, which often appear as outlier measures, are discarded [77]. Alternatively, instead of individual protein quantification, the abundance of protein groups sharing many peptides can be assessed [78].

Another issue arises with peptide spectra or genomic reads which cannot be mapped and assigned to any reference sequence. This means that they fall through the mapping step and are omitted for quantification correspondingly. Reasons for unmapped reads in a metagenomics scenario can be either novel taxa, for which no sequenced genome is available yet, or concern taxa which are only represented by certain strains in the reference database. Thus, in the latter case, not all sequence variations of a taxon are covered to enable accurate mapping. Generally, unmapped reads provoke a significant abundance bias as the relative abundance of known and identified genes or taxa are correspondingly overestimated [22]. Especially quantitative comparisons between samples, which hold different amounts of unmapped reads, risk strongly biased results. Notably in metagenomics studies, even in well-studied environments such as the human gut 43% of prokaryotic species are assumed not to be covered by reference genomes. It becomes even more problematic in soil or seawater samples in which above 90% of the microbes are expected to be unknown [79]. One approach is to estimate relative taxa abundances proportional to the total number of sequenced reads or define and estimate a separate ‘unclassified’ category [74]. However, unmapped reads can also occur because of contamination, host DNA or sequencing errors. Thus detection and removal of the latter is crucial while distinguishing them from novel taxa findings. The same points and challenges also apply to non-identifiable spectra which may stem from novel proteins, or from different isoforms, as well as from contaminants in MS experiments.

A standard data processing step, common to any quantitative experiment, is the detection and filtering of outliers. This concerns measurements observed outside the range of the majority of data points, atypically low or high values suspected to arise due to different reasons. However, important biological information can be hidden in outliers, for example the modification of a peptide. Further, single identifications of peptides or genes referred to as one-hit wonders also deserve attention as they may indicate novel findings [49]. The challenge lies in distinguishing between error, novelty and extremely large data variability. A popular method applied to continuous measurements is to compute z-scores and classify outliers according to exceeding 2-3 fold standard deviations [80].

A last important point concerns missing values which bring a large unknown factor into quantification assessment. If no reads or spectra are reported for a peptide, gene, taxon or protein, at the same time no evidence is given that those are actually absent. As described before, loss can occur in various sample processing steps due to biochemical or bioinformatics flaws. Corresponding reasons can be, for example, signals that fall beyond the detection range, signals being concealed by noise, suppression due to GC content or false positive identifications. Missing values naturally imply an abundance underestimation, while also causing relative abundance overestimation of detected components in composition studies. Major difficulties arise in proteomic labelling experiments with different conditions, here, a missed state evokes an infinity ratio for the corresponding peptide. Ignoring missing values is the simplest and often applied method. But numerous methods for substituting missing values, so called imputation strategies, are available according to three missing mechanisms: values are missing completely at random (MCAR), missing at random (MAR) but conditional dependent on known values, or missing not at random (MNAR), e.g. falling beyond detection limits [81–84]. Common statistical methods, to name k nearest neighbours, bayesian and maximum likelihood methods are applied to random missing values, while more data-adapted methods are required for MNAR. Overall, the choice of a suitable imputation strategy is highly dependent on the data set and its characteristics.

This section demonstrates different sources which cause errors and biases in quantification workflows, examining steps from sample preparation to quantification calculation. It reveals how different errors affect quantitative estimates in different ways and why specified solutions are required. Generally, reduction of biases can be either approached by technical and experimental advances or by applying computational correction strategies.
A general rule is that accuracy and precision of quantification is always increased with an increasing number of measurements. More data yields more overall robustness, higher identification reliability, less mis-assignments and results in reduced variance of abundance measures. Hence, higher read coverage in NGS experiments as well as increased numbers of PSMs naturally improve gene, protein and taxa quantification. At the same time, individual modifications, sequence errors, false -positive identifications and any outliers have less impact within a large number of measurements, which are assumed to largely reflect the true underlying abundance.

Overall, all potential quantification errors presented in this section contribute to the uncertainty of a final abundance estimate and combined give rise to the total variance per estimate. This variance will be referred to as *abundance variance* in this thesis. In the next section, the focus is on the comparison of abundance estimates obtained from different samples and the view will be extended to further global variances.

1.3. Differential quantification

Accurate quantification in a sample forms the base for successful differential quantification analysis when comparing two or more samples. Correct detection of significant abundance change of proteins, genes or taxa between samples and a precise inference of log-fold changes is primarily determined by the accuracy of estimations given in each sample. Further, it has to be noted that errors and corresponding biases occurring during the quantification process of a sample are not reproducible. Each experiment replication is influenced by already slightly varying factors such as temperature, instrument calibration or even a different person conducting the experiment [17]. As a result, variances of abundance estimations vary for each sample.

When comparing samples, the number of influences and the complexity of variances increases for the assessment of relative abundances. In addition to the variance emerging from uncertainties in the abundance estimation process within each sample, one needs to account for the variance arising between samples. Here we need to distinguish between variances as a result of technical, experimental or biological reasons [85]. We refer to technical replicates as an experimental repetition of two identical sample probes and the observed variance is reflecting the error variation in the quantification process, whereas the variance from experimental replicates comprises influences from different experimental sets. However, samples taken from different tissue or under different external conditions bear variances due to biological origin. In a complex study with several samples, all these variances are mixed up and contribute together to an uncertainty in relative abundance estimation. The challenge for bioinformatics methods is to decouple these variances and choose appropriate thresholds to capture significant abundance changes of interest. Aim of most differential studies is to identify biological induced abundance changes and separate them from abundance shifts evolved solely due to experimental and technical aspects.

In order to minimize experimental biases for sample comparisons ensuring the same experimental conditions is key. This is a clear advantage of isobaric labelling in proteomics in which multiple samples can be processed together and run under the exact same conditions, which allows certain errors to cancel each other [49]. In contrast, in label-free approaches, samples are processed independently and arising errors are not balanced naturally [3, 17].

Another challenge in quantitative comparison studies concerns the fact that initial sample loads are often varying between samples. This results in the generation of different total amounts of reads or spectra respectively, which in turn is reflected by shifted signal intensity scales or overall shifted read numbers in the different samples. Normalization of the samples is crucial to account for this variation and make samples comparable for differential analysis [86, 87] . An extensive development of normalization methods for genomics data has taken place during the microarray era. The main assumption followed in gene expression analysis is that the majority of genes do not alter in expression between different samples [88]. Hence, most normalization methods aim to scale the majority of abundance measurements to the same level in all samples. This is traditionally implemented by aligning certain data metrics or entire data distributions between samples. Well known is the sumtotal normalization at which each measure of a sample is divided by the total sum of measures of the sample [87, 89]. Other scaling methods align samples according to the median, mean or the 75^{th} percentile [88]. A popular and strong approach is the quantile normalization method, which forces all sample measures to the same quantiles, resulting in equalized sample distributions [86]. These normalization concepts are used in many quantitative fields and have been transferred from microarray to RNA-Seq as well as to metagenomics and to diverse applications in proteomics. They all do share expectations that measurements reflect certain biological stability across samples. However, it is crucial to carefully investigate for each data type whether the assumptions are met and are meaningful with regard to the question of interest. For example, no abundance consistency is expected across samples of pull-down experiments and corresponding negative controls (see also Chapter 2). Inappropriate data normalization can impede detection of differential abundant candidates and strongly bias data sets. A chosen normalization method should ideally capture data characteristics as best as possible. It should be adapted to continuous or discrete measures, handle sparsity or be robust to outliers. Further, in very heterogeneous data cases the use of housekeeping proteins or specific spike-ins can be beneficial [90].

Generally, at this stage of data post-processing, independent of the individual underlying experimental techniques, when dealing directly with continuous or discrete abundance measures, common concepts and principles from statistics and computational approaches come together. In the end, we are facing discrete count measures as a quantitative output from RNA-Seq, from NGS-based metagenomics experiments, and in all MS-proteomic experiments relying on spectral counts. Equivalently, continuous intensity measures are in the center of microarray analyses and of proteomic studies, which utilize peak signal information. Finally, the statistical questions of interest mainly revolve around comparisons of quantities from either different samples or labels, aiming to identify significant differentially abundant candidates. Hence, despite all disparities within genomic and proteomic techniques, quantitative analyses of taxa in a metagenomics study or of transcripts in an RNA-Seq study or of proteins in a shotgun proteomics experiment can become analogous problems from a statistical point of view.

Thus, Gaussian or Beta distributions are frequently used to model gene expressions of microarrays or peptide MS signal intensities, while Poisson and Binomial distributions, appropriate for modelling discrete count data, are popular in RNA-Seq, metagenomics and corresponding MS studies [91]. Standard statistical tests for assessing quantitative differences, such as the well-established t-test, Welch-test, Gauß-test or Wilcoxon-test and others, however require adaptions to the new high-dimensional data set specificities. Commonly the amount of quantified genes or proteins within samples greatly exceeds the total number of samples considered. Hence a *many-features-few-replicates* problem arises. Various statistical methods evolved or were adjusted to analyze microarray data, one of the most prominent methods being LIMMA [92]. The innovation of LIMMA is to makes use of the highly parallel measurements and to borrow strength between gene-wise models. The introduced empirical Bayes approach, using a moderated t-statistic, enables variance stabilization across gene measures, even when the number of samples is small. This concept of LIMMA was extensively transferred to methods for high-throughput data from all kind of technologies [93]. Various new tools have emerged keeping the core idea of LIMMA and integrating data specific characteristics. It is important to note that methods for microarray data are not directly applicable to NGS data. The latter consists of discrete counts of reads and often sample variances occur to be greater than the sample means, referring to an issue called overdispersion. As a consequence, models based on negative binomial distributions, suitable to account for overdispersion, have been developed for NGS-based analysis [89, 91]. Therein, the tool edgeR [94] was introduced as natural follow-up of LIMMA by the Smyth lab and Anders and Huber proposed their tool DESeq as further improvement to edgeR in the RNA-Seq field [95, 96]. And further, the most recently published tool sleuth [97] by the Pachter lab again incorporates the variance stabilization concept of LIMMA. Because of many similarities between transcriptomics and metagenomics data, many methods originally developed for RNA-Seq are equivalently applied in comparative metagenomics studies [22, 98]. One of the main differences that transcripts are significantly shorter than genome sequences has rather computational implications. But, one important aspect to consider in metagenomics studies is that metagenomes are frequently undersampled, which causes potential sparsity in the count data. This can bias differential abundance testing as zero counts cannot be equalized with absence. A group of zero-inflated models have evolved to explicitly account for undersampling and are combined, for example, with a Gaussian mixture distribution in metagenomeSeq [99], with a log-normal distribution in RAIDA [100] or with a beta-regression in ZIBSeq [101].

Differential testing methods from genomics and transcriptomics are also slowly seeping into the younger comparative quantitative proteomics field [102]. Different studies have confirmed similar statistical properties of transcript and protein abundance values. Pavelka *et al.* even attested microarray and NSAF values to follow the same global error model [103]. Further, particularly proteomic studies based on spectral counts benefit from developments in the RNA-Seq field. Correspondingly, methods ranging from modified Students t-test to DESeq are applied to shotgun proteomics data in the literature [104]. However, overall, differential quantitative methodology in proteomics is still lagging behind, leaving a great demand for investigations on method transfers and the need for tools that precisely capture spectra data complexities.

A last statistical correction step equally concerns all genomic and proteomic high-throughput data types in a differential abundance study. The fact of having large numbers of measured genes and proteins in a sample and a corresponding large number of differential tests conducted comes with the risk for a certain amount of false-positive candidates and the need to control the type-1 error rate. Application of multiple testing procedures is crucial to retain a prescribed false-discovery rate (FDR) or a more conservative family-wise-error rate (FWER) to ensure a final reliable differential candidate list. Well-established methods are the Holm-Bonferroni correction [105] or the less stringent Benjamini-Hochberg method [106] which can be directly applied to the resulting p-values from any test.

In summary, many statistical methods can be shared between comparative genomics and proteomics analyses. After technique specific data acquisition and processing, final quantitative values exhibit many similarities, and statistical problems in differential analysis become very much the same. There is a strong potential in re-using and adapting given methods and transferring statistical concepts between the different fields, which is by far not exploited extensively.

1.4. Thesis objectives

Accurate quantification in high-throughput data is a highly challenging task. The previous sections show the overall complexity of quantification workflows and reveal multiple sources that cause diverse biases in quantification. Although various efforts have been undertaken to enhance individual steps in the process, the challenge to retrieve true underlying quantities in a sample is not solved. An urgent demand persists to provide adapted solutions that minimize the variance of abundance estimates. Existing parallels and common biases between different techniques but also specific technique-induced biases are depicted. One-fit-all methods are considered to be not appropriate. New methods are strongly desired that inherit established statistical concepts and are also customized to data types of different technologies. Furthermore, a general lack persists in the quantitative field on how to measure and report the reliability of quantification results. Statistical assessments are often under-represented in analysis tools, especially in quantitative proteomics research. Aim of this thesis is to introduce new statistical analysis strategies and enhance data pre-processing in certain data types to achieve improved quantification accuracy.

This thesis comprises three major projects focusing on different data types stemming from three quantitative high-throughput techniques. Two projects consider MS-based proteomics data, one project is based on affinity-purification MS experiments and the other on isobaric labelling techniques. The third project addresses NGS-based metagenomics data. All projects pursue to answer different biological questions, however also share common problems in quantitative analysis. The thesis focuses on accounting for the individual data characteristics and challenges and on developing adapted methods for enhanced quantification accuracy. This work aims to highlight how established strategies from other fields can be transferred and utilized, while at the same time data-specific features need to be integrated. Overall, the three projects cover the main topics of pre-processing of quantitative measures, quantification inference and resolution, and comparison of quantities.

In the first project, affinity purification (AP) is coupled with mass spectrometry aiming to identify protein-protein-interactions [107–109]. In the purification step, a tagged bait protein is applied to capture potential interaction partners (preys), which are subsequently sequenced in a label-free MS approach. Simultaneously, negative controls are conducted to identify contaminant proteins. Protein abundances are inferred by spectral counts. As a result of the AP-MS study, a list of possible interaction proteins are obtained, accompanied by their count measures for bait and control samples. The challenge in the analysis of AP-MS data is to reliably separate truly interacting proteins from false-positives by contrasting the quantitative information in bait versus control samples. A variety of methods exists, which use scoring schemes to describe the likelihood of true interactions, the most popular one is the SAINT software [110]. However, often the question remains which cutoff score to choose to find highly reliable interaction candidates and how many false-positives are expected in a final list. An assessment by statistical measures is frequently missing in the analysis of AP-MS data. Generally, the field of AP-MS analysis shows little method exchange with other quantitative disciplines. However, on a more abstract level, the identification of enriched truly interacting proteins can be related to a standard differential abundance analysis between two conditions. Further, as the quantitative comparison is based on discrete counts, parallels between AP-MS and RNA-Seq analysis can be revealed. Thus, the suitability of models developed for RNA-Seq is investigated and how AP-MS data can benefit from it. Yet a main difference concerns the two-sided focus in RNA-Seq analysis in comparison with a one-sided interest in AP-MS data. Here only an increase of counts in bait samples indicates true interaction. In this work, a two-stage Poisson model (TSPM) [111]), developed for RNA-Seq data, is consequently adapted to the features of AP-MS data. It is introduced as a new scoring method to identify interaction candidates in an AP-MS study. Further, a permutation framework embedding scoring methods like SAINT is proposed. This allows assessing the actual significance of scores by testing whether they could have been derived by chance and replaces scores by proper statistical p-values. The permutation principle applied is a well-established method in statistics [112–114]. It can be further coupled with the algorithm of Westfall and Young, which yields p-values that control the family-wise-error rate (FWER) [115]. This finally enables the user to select cutoffs in candidate lists according to a defined significance level. Moreover, data pre-processing plays a big role in microarray and NGS analysis, however is rarely considered in AP-MS workflows. Several normalization methods and also a common statistical filtering strategy are adapted to the characteristics of AP-MS data in this work. Their benefit and impact on detecting true interactions is examined. As a result, the detection of truly interaction protein candidates is shown to be significantly improved by the application of pre- and postprocessing methods. In this project, a new statistical thorough framework is provided for candidate evaluation in AP-MS protein interaction analysis.

The second project is also involved with MS-based proteomics data, but focusing on quantification based on signal intensity measures instead of spectral counts. While spectral counting refers to a very robust and easy approach for protein abundance inference, more in-depth information is hidden in spectra intensities and PSMs. MS-based measurements are generally assessed at the peptide spectrum level and all peptide measures assigned to the same protein need to be summarized to infer the final protein abundance. Generally, peptide intensities stemming from the same protein are assumed to hold similar values. However, in fact, a substantial hetero-

geneity of peptide measures is observed per protein due to the impact of random and systematic biases [65,66]. Hence, not all spectra measures are of equal quality. Clever peptide-to-protein summarization methods are crucial to account for this variation to accurately infer the true underlying protein abundance. This project concentrates on this specific and very important pre-processing step in quantification, which is often addressed poorly.
Investigations and method development was conducted on isobaric labelling data, primarily iTRAQ and TMT studies [47,68,70,74]. However, the presented concepts are generally transferrable to SILAC and label-free approaches. As presented in the previous section 1.2, most summarization methods purely rely on given peptide quantitative information. The only feature which has been studied and acknowledged to indicate reliability of peptide measures is the absolute signal intensity [66,116,117]. Peptides reported with low absolute ion intensities have been shown to be more prone to noise and error, while high-intensity peptides improve protein quantification accuracy. However, this work hypothesizes that a wealth of additional peptide features does exist with the potential to hold more valuable information concerning the quality of peptide measures. It is the aim to answer the question whether the observed peptide signal heterogeneity can be described and explained by underlying peptide spectra characteristics. Several peptide features are studied, such as charge state, sequence length, peptide mass, identification score, modification state, absolute ion intensities, and distances within redundantly measured spectra derived from the same precursor. A systematic investigation is conducted on how individual features correlate with the observed variance heterogeneity and how they impact quantification accuracy. We believe that particularly the combination of features and their opposed strength enables to assess the quality of spectra measurements. A feature-based weighting of peptide spectra is developed, so that individual spectra contribute to the protein quantification according to their feature reliability. In summary, this work provides a novel peptide-to-protein summarization method, referred to as *iPQF*, which integrates peptide features with quantitative measures. It demonstrates the added value of spectra feature information to improve the accuracy of final protein quantification.

The challenge of correct summarization and quantification inference based on shotgun measurements plays a crucial role in MS and NGS-based quantifications alike. In place of peptide spectra, genomic reads need to be assigned and summarized to infer final transcript or taxa abundances. Equivalently to the concept of spectral counting, counting the number of mapped reads per reference sequence is most straightforward; however, the genomic origin of reads cannot always be determined precisely and may cause biased count abundances.
This issue of shared read mappings becomes particularly crucial in metagenomics settings when aiming for higher taxonomic resolution. Considering strain and substrain levels relates to the study of highly similar genomes sequences and a significant amount of reads will map equally well to different genome sequences. In this third project, we address NGS-based metagenomics data with a focus on strain level quantification and its special challenges. Most tools only provide resolutions

down to species level and many developments have concentrated more on identification and speed advances [118]. One of the few tools explicitly accounting for genome similarities and the shared count problem are GRAMMy [74] and GASiC [73], which are still alignment-based. Key idea of this project is to utilize the new and fast k-mer based pseudo-alignment strategies in the field and combine it with an improved model for read ambiguity resolution. The aim is to provide accurate quantitative resolution on strain- and sub-strain level to enable the analysis of samples comprising even many similar strain clusters. This refers to an urgently needed step in metagenomics profiling. Further, while many differential abundance methods exist, none are adapted to strain level specificities.

Here, a new generalized linear model (GLM) framework for shared read count resolution is introduced, which is designed to capture more precisely the data structure of mapping count data given for taxa. Further, as no method can assure perfect ambiguity resolution, it is important to acknowledge the uncertainty in the abundance estimates. The proposed model integrates this variance in form of an additional error term. The integration of the abundance variance is of particular importance for differential abundance analysis in the presence of similar reference genomes. It plays a crucial role in the detection of small but significant fold-changes between metagenomes holding many strain clusters. As described in the previous section, many statistical models are available in the genomics field for differential analysis, however focusing almost exclusively on biological and technical between-sample variances. This project specifically concentrates on modelling and integrating the variance that arises within the abundance estimation. A novel statistical approach is introduced, in which the taxa abundance estimates along with standard errors are used to derive abundance distributions. Hence, instead of comparing point estimates, the divergence of two abundance distributions indicates the differential abundance change of a taxon. It refers to an empirical approach without imposing prior assumptions on overall abundance change between samples. Overall, a comprehensive approach, to which we refer as *DiTASiC* (Differential Taxa Abundance including Similarity Correction) is provided for abundance estimation and differential testing sensitive to strain level, along with statistical measures for evaluation.

1.4.1. Thesis Outline

This thesis presents new computational and statistical strategies to improve the quantification accuracy of data types from high-throughput omics applications. The three main contributions are described in detail in the following chapters 2, 3, and 4. In chapter 5, the impact of the three contributions is summarized and a future outlook is given. All contributions were developed under the supervision of Bernhard Renard, who is Co-author in each project.

In Chapter 2, a comprehensive and novel statistical framework, referred to as *APMS-WAPP*, for the analysis of AP-MS data to identify protein-protein interaction candidates is presented. Experimental data for evaluation and method development was provided by Susann Zilkenat and Samuel Wagner (both from the Institute of Microbiology and Infection Medicine (IMIT) in Tübingen), and by Roman Gerlach (Robert Koch Institute). Samuel Wagner also contributed to drafting the result section of the *Salmonella* study in the manuscript. The chapter is based on the publication:

> *Pre-and post-processing workflow for affinity purification mass spectrometry data.* **M. Fischer**, S. Zilkenat, R. G. Gerlach, S. Wagner, B. Y. Renard. *Journal of Proteome Research* (2014), 13(5), 2239-2249.

Chapter 3 is dedicated to the step of peptide-to-protein summarization. It addresses the challenge to infer protein quantities from heterogeneous peptide spectra measurements and integrates feature information as a novelty. The introduced method *iPQF* was published in:

> *iPQF: a new peptide-to-protein summarization method using peptide spectra characteristics to improve protein quantification.* **M. Fischer**, B. Y. Renard. *Bioinformatics* (2015), 32(7), 1040-1047.

Chapter 4 approaches higher taxonomic resolution in NGS-based metagenomics data. *DiTASiC* is presented as a new approach enabling accurate taxa abundance estimation and differential testing sensitive to strain and sub-strain level. This project was developed with Benjamin Strauch, who contributed to software development and performance evaluations. The chapter is based on the publication:

> *Abundance estimation and differential testing on strain level in metagenomics data.* **M. Fischer**, B. Strauch, B. Y. Renard. *Bioinformatics* (2017), 33(14), i124-i132.

2. APMS-WAPP: pre- and postprocessing of AP-MS data

The reliable detection of protein−protein interactions by affinity purification mass spectrometry (AP-MS) is crucial for the understanding of biological processes. Quantitative information can be used to separate truly interacting proteins from false-positives by contrasting counts of proteins binding to specific baits with counts of negative controls. Several approaches have been proposed for computing scores for potential interaction proteins, for example, the commonly used SAINT software. However, it remains a subjective decision where to set the cutoff score for candidate selection; furthermore, no precise control for the expected number of falsepositives is provided. In related fields, successful data analysis strongly relies on statistical pre- and post-processing steps, which, so far, have played only a minor role in AP-MS data analysis. We introduce a complete workflow, embedding either the scoring method SAINT or alternatively a two-stage Poisson model into a pre- and post-processing framework. To this end, we investigate different normalization methods and apply a statistical filter adjusted to AP-MS data. Furthermore, we propose permutation and adjustment procedures, which allow the replacement of scores by statistical p values. The performance of the workflow is assessed on simulations as well as on a study focusing on interactions with the T3SS in *Salmonella Typhimurium.* Preprocessing methods significantly increase the number of detected truly interacting proteins, while a constant false-discovery rate is maintained. The developed R-package *APMS-WAPP* is freely available.

2.1. Study of protein interactions by AP-MS

The reliable identification of protein−protein interactions plays a key role in numerous biological questions, for instance, in the search for components forming a protein complex or for inferring the function of a protein by its known interaction partners.

Affinity purification combined with mass spectrometry analysis (AP-MS) has emerged as a popular technique to study protein interactions [107–109]. A protein of interest, the baitprotein, is purified in the AP step with potential interaction partners binding to it. This is followed by a digestion of the extracted protein mixture into peptides, which are separated by liquid chromatography (LC) and subsequently sequenced by mass spectrometry. On the basis of the acquired MS/MSspectra, peptides are identified by database search strategies and proteins are subsequently inferred. The abundance of the proteins in label-free MS experiments can be assessed by determining continuous MS intensities or by MS/MS spectral counting.

The outcome of a label-free AP-MS study is a list of possible interaction partners (preys) of the bait protein supported by quantitative information.
However, these raw AP-MS data sets bear a large number of false-positive interactions (here referred to as *contaminants*). Contaminants frequently occur as proteins that bind nonspecifically, for example, to affinity matrix, antibody, or tag [107]. It is crucial to have negative controls, in which the affinity purification is repeated in the same setting but without the bait protein because these may indicate contaminant proteins. Hence, the main challenge in the analysis of AP-MS data lies in the reliable separation of true interaction proteins from contaminants.
Different computational approaches have been proposed to address this task in label-free AP-MS data using the stronger quantitative evidence of true interactions in bait than in control purifications. These methods range from heuristic filtering methods to empirical and probabilistic scoring approaches [108, 119, 120].

Eliminating all proteins that were detected in the negative controls constitutes the most rigid treatment, whereas filtering for proteins with a ratio of spectral counts in the bait versus control experiments exceeding a certain threshold is often a more suitable alternative. Furthermore, different frequency filters have been introduced, judging the reproducibility in replicates as well as the abundance of a protein in different baits within a large-scale study. In a serial dilution approach of bait and control samples, quantification profiles are used to characterize true and false interaction proteins [121]. Widespread is the application of a statistical t-test to compute the significance for an interaction based on spectral counts, which can be misleading because the underlying discrete counts do not follow a normal distribution as required for the test. Minimal fold change requirements as well as positions in volcano plots have been added as additional criteria to the p value [122]. In the empirical method CompPass, [123] a D-score is calculated corresponding to an adjusted spectral count based on the uniqueness and the abundance of the prey proteins as well as the reproducibility of interactions. Truly interacting proteins are defined by a D-score threshold estimated by simulations. In another empirical approach, the relative protein abundance is estimated by normalized spectral abundance factors (NSAFs) [124]. A prey is regarded as a contaminant if the ratio of its NSAF values between bait and control samples lies below an empirically selected threshold. A sophisticated, probabilistic approach, SAINT [110], was developed on spectral count data and later extended to include MS-intensity data [125]. Scoring the interaction of a bait-prey pair by SAINT is based on a Bayesian model, estimating the distribution of true and false interactions including different features.

The main challenge in the analysis of AP-MS data is the reliable generation of a cutoff score for candidate selection and the estimation of the expected number of false-positives. This is similar to existing standards in other areas of proteomic research, for example, for the identification of peptides and proteins [126, 127]. Some scoring methods give instructions on how to at least approximate a false-discovery rate (FDR); however, the accuracy of this estimation cannot be guaranteed and has not been assessed so far due to the lack of appropriate benchmark data sets [108]. Finally, the aim is to provide wet-lab scientists with a highly reliable list of protein interaction candidates, which gives valuable advice on how many and which vali-

dation experiments are worth conducting.
In related fields, such as the analysis of microarray and nextgeneration sequencing data, it has been shown that successful data analysis relies on the impact of statistical preprocessing steps [86, 87]. However, normalization or statistical filtering is not considered or has played only a minor role in AP-MS data analysis so far.

In this contribution, we investigate the impact of pre- and post-processing steps on AP-MS data analysis. Considering different normalization methods adapted from microarray and RNA-seq analysis and applying a statistical filter adjusted to APMS data, we show how the detection of truly interacting proteins can significantly be improved by preprocessing of the data. For postprocessing, we propose a permutation methodology with the application of the Westfall and Young algorithm [115] to replace ad hoc interaction scores with a more proper and interpretable statistical measure. This allows setting the cutoff in the list of potential interaction proteins according to a desired significance level or, respectively, to the expected number of false-positive interactions one is willing to accept in the final output list. We focus on single-bait data sets that are designed to specifically identify the true interaction partners of a protein of interest, a common objective in particular in bacterial experiments [128]. The experiment should include replicates to ensure a certain level of confidence, and additional negative controls are required. The label-free quantification method considered is spectral counting.

We use SAINT as a current de facto gold standard for these experiments and introduce a complete pipeline for the analysis of AP-MS data with pre- and post-processing steps framing the scoring method SAINT.In addition, we investigate if the analysis of AP-MS data can benefit from existing techniques established for the analysis of RNA-seq data. Studies of differential expression between two conditions on RNA-seq data also result in discrete count data and exhibit many parallels to the identification of interaction proteins in AP-MS data, yet a major difference is that all methods proposed for RNA-seq data are two-sided tests. They simultaneously consider up- and down-regulation in expression, while in AP-MS data the focus is entirely on one-sided strategies identifying significantly higher values in bait samples than in controls. We focus on a two-stage Poisson model (TSPM) [111] and adapt it to AP-MS data.

We introduce two alternative workflows for the analysis of AP-MS data with pre- and post-processing procedures but replace SAINT with the adapted TSPM approach for evaluating the interactions. TSPM can be combined with two different postprocessing procedures, the procedure of Westfall and Young and the method of Benjamini−Hochberg [106], providing different controls of false-positive interactions. The performance of the three proposed workflows is assessed on simulated data as well as on experimental data focusing on interactions with export apparatus components of the type-III secretion system on pathogenicity island 1 of *Salmonella Typhimurium.* Thereby, we comprehensively study and discuss the added value of each single component within the workflow. As a result, we show the impact of pre- and post-processing methods and how the detected number of truly interacting proteins can significantly be increased while maintaining a constant false detection rate.

2.2. Statistical framework for AP-MS data

As shown in Figure 2.1, the proposed workflow for the analysis of AP-MS data consists of three main parts: preprocessing, scoring, and p-value assessment for each protein. Preprocessing comprises normalization and filtering of the data. Normalization is necessary to make samples comparable by removing systematic biases. Here we compare five different normalization methods. Next, a filtering step allows early elimination of obvious contaminants from further analysis. To determine the interaction potential of a protein and to provide an initial ranking of interaction candidates, we can apply either SAINT or TSPM. SAINT delivers a score, and TSPM provides a test statistic from an LRT for each individual protein. To evaluate the significance of the scores, a permutation procedure is performed to assess whether it could have been derived by chance. The permutation procedure builds the empirical distribution of scores, and subsequent application of the Westfall and Young algorithm allows the replacement of scores by p values that can be interpreted in a statistical way. In the case of TSPM, p values can alternatively be derived from the known underlying χ^2 distribution. Further adjustment methods account for the total number of candidate proteins. Finally, the approach enables the estimation of the portion of false-positives in a list of interaction candidates by a family-wise error rate (FWER) or an FDR.

2.2.1. Data requirements

The workflow focuses on single-bait experiments with the goal of identifying all detectable interaction partners with high confidence. For scoring and postprocessing, a minimum of three replicate bait samples is a prerequisite to account for variation in the experiment and to ensure reliable results. Preprocessing is also applicable to two replicates. The method works for single-bait replicates generated by independent sample preparation, purification, and MS runs as well as on technical replicates. Furthermore, negative control experiments are essential for the detection of contaminant proteins; also, here at least three replicates are advisable. The CRAPome database [129] offers new possibilities to integrate negative controls in the case that the same purification condition is met in the repository. Protein abundances are assessed by spectral counting.

2.2.2. Preprocessing: Normalization

The purpose of normalization is to remove systematic biases from the data and to enable a comparison between the samples [86]. Biases in LC−MS/MS data can evolve from different sources [130]: varying sample-processing conditions can lead to different amounts of probe material in the samples; furthermore, instrument calibration, LC columns, or changes in temperature during the experiment may influence measured protein abundances. When investigating differential protein abundances, the detection of a change could be due to technical and experimental aspects. Hence, normalization of the data is crucial to remove these biases to enable the detection of existing biological changes caused by truly interacting proteins.

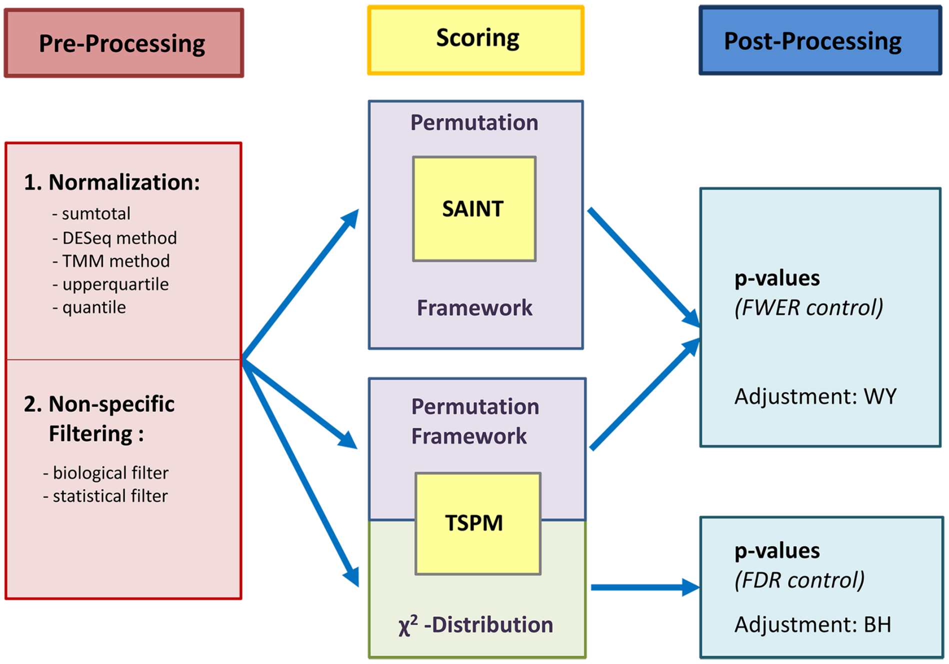


Figure 2.1.: Workflow overview showing the three main parts for the analysis of AP-MS data: preprocessing, scoring, and postprocessing of interaction proteins. Preprocessing comprises normalization and filtering of the data; here we compare five different normalization methods. Purpose of the filter is the early elimination of obvious contaminants. In the next step, scoring of the interaction proteins is conducted by either SAINT or TSPM. A permutation framework in combination with the subsequent WY-adjustment replaces the scores by p values. In the case of the TSPM model, p values can alternatively be derived from a distribution coupled to a BH adjustment. Finally, a control of false-positive interaction proteins is provided.

Normalization methods have been a major focus of research in genomics, and main ideas are transferable. The assumption made in most of these normalization methods is that the majority of genes are not differentially expressed between different conditions [86, 88], and thus normalization is performed across all samples to align these genes. This concept can be transferred to proteome analysis when the abundance of most proteins remains unchanged; however, this assumption does not hold for AP-MS data. In an AP-MS setting, proteins are measured only if they purify either with the bait or in the control. Thus, proteins showing the same quantities in both samples are likely contaminants and will not be common. Hence, the idea of scaling the abundance of the majority of proteins to the same level in all samples is not appropriate. However, we do expect a uniform expression of the protein abundances within the replicate samples for either control or bait, which justifies the separate application of normalization procedures to remove technical and experimental biases.

A second issue is that generally fewer identified proteins are expected in control than in bait samples. Thus, a sample-wise normalization procedure can boost lower abundant proteins in the controls relative to higher abundant proteins in the baits.

We balance this effect by rescaling normalized counts by the median count level in baits and controls, respectively.

We adapted, implemented, and applied different normalization methods commonly known from the analysis of microarray and RNA-seq data [87] to AP-MS data. Up to date, the most prevalent and, to our knowledge, the only commonly available normalization method for AP-MS data is the sumtotal method [87, 89]. Here each protein count within a sample is divided by the total number of counts being measured within that sample. Difficulties can occur if a sample contains an outlier in terms of an extremely high count; as a consequence, all counts in the corresponding sample are decreased by the normalization. A more robust version with regard to outliers is the scaling method referred to as upperquartile [88]. Instead of the sum of counts, the 75th percentile (the upper quartile) of all counts within each sample is calculated and serves as a denominator. An even stronger approach to align count distributions in terms of quantiles is the quantile normalization, which has become popular in microarray analysis [86]. The goal of this method is to adjust the distributions of protein counts across the samples by forcing them to the same quantiles.

Furthermore, two normalization methods that evolved with RNA-seq data, thus being designed for discrete count data, are applied: the DESeq approach by Anders and Huber [95] and the TMM method by Robinson and Oshlack [131]. In DESeq, a ratio is calculated for each protein by dividing the counts of a protein in a given sample by the geometric mean of counts for that protein across all samples. Finally, each count is corrected by dividing it by the median of all ratios determined in the corresponding sample. TMM requires the selection of a test and a reference sample to compute scaling factors. Each count in the chosen test sample is then divided by a weighted mean of log ratios between test and reference sample.

All normalization methods align replicates and remove systematic biases. However, the choice of the most appropriate normalization method may depend on data characteristics. (See also Appendix A)

2.2.3. Preprocessing: Filtering

Another crucial preprocessing step concerns the filtering of the data [132] to eliminate obvious contaminants from further analysis. The overall aim of filtering is to enrich the data for truly interacting proteins while simultaneously reducing the number of potentially interacting proteins. This is especially important for the subsequent testing procedure, as reducing the number of proteins to be tested decreases the multiple testing problem. The proposed method comprises two different filters, a biologically motivated filter and a statistical variance filter.

The role of the biological filter is to account for proteins showing higher counts in the control samples than in the bait samples. Exhibiting a stronger binding affinity to the matrices than to the bait protein is pointing to a clear contaminant. These proteins are identified by contrasting their median count in the controls to the median in the bait samples and are removed from the data set.

The second filter is motivated by the idea that a truly interacting protein should show an increase in counts in the bait samples compared with the controls. A protein is assumed to be a contaminant in case it shows similar counts across all samples. This is indicated by a low variance of the counts. It is important that the filtering involves all samples independently of the underlying class labels (bait and control); this approach is termed nonspecific filtering. As a filtering criterion, the overall variance of the counts is computed for each protein, and the fraction of proteins with the lowest overall variance is removed. A more robust criterion is to derive the interquantile range (IQR) of the protein counts corresponding to the difference between the 25 and 75% quantiles. The computation of the IQR is useful if outlying counts are expected and the sample size is large enough ($\geq$8).

The subsequent challenge is to define the cutoff for filtering, specifying the group of proteins having a variance or IQR below the cutoff and thus being considered as contaminants. One possibility is to set the cutoff according to a quantile. In the case that no prior knowledge is available for defining a quantile cutoff, a common approach is to determine the shortest interval containing 50% of the data in the variance distribution, assuming that the majority of proteins holds a small variance. The mean of the calculated interval can be used as cutoff [133]. (See Appendix A for more details on the biological filter and the cutoff choice.)

2.2.4. Scoring

SAINT

SAINT (significance analysis of interactome) [110, 125, 134–136] was developed for scoring protein−protein interactions in label-free quantitative AP-MS data. A Bayesian model calculates the posterior probability of observing a true interaction based on the count of a specific prey protein. By further averaging over replicate samples, a confidence score for each protein−protein interaction is obtained. A score in close proximity to one represents a true interaction, while a score decreasing to zero refers to a likely contaminant.

TSPM model

We adapt a two-stage Poisson model (TSPM) [111] to AP-MS data, which was originally developed for the analysis of RNAseq data and apply it for scoring protein−protein interactions. TSPM considers generalized linear models under the assumption that the observed counts for each protein are derived from a Poisson distribution. A reduced model is fitted under the null hypothesis that the counts for each protein have no discriminative character between bait and control samples, and hence all samples are treated the same, ignoring their labels. An alternative full model is fitted assuming that there is an association between the counts of a protein and the corresponding sample labels. This means that different count profiles are expected in the bait and control samples.

In the following step, a likelihood ratio test (LRT) is applied to compare the two models. In case the LRT leads to the rejection of the null model in favor of the alternative model, the considered protein is likely to be a true interaction protein.

The model fitting and the subsequent LRT are carried out proteinwise, and thus a LRT statistic is obtained for each individual interaction candidate.
The procedure was adapted to a one-sided test, as we are only interested in cases where counts in the bait samples exceed the ones measured in the controls. A second adaptation concerns overdispersion in the data. (See also Appendix A)

2.2.5. Postprocessing: Permutation framework and adjustment procedures

Independently of the scoring method, we aim at replacing the score by a true significance level. This allows revealing how valuable a score is or whether it could have been derived by chance. In the case that the underlying distribution of the scores is known, a statistical p value can easily be inferred. The distribution of SAINT scores is not known, and p values cannot be calculated directly. Using other scoring schemes, distributions might also be uncertain due to a small sample size, which is a common issue in AP-MS experiments. Therefore, we propose the application of a permutation procedure, which builds an empirical distribution to assess statistical p values for the given scores.

The permutation principle is a well-established method [89, 113, 114, 137], originally introduced by Fisher [112]. First, the original score is calculated for each protein by the scoring method. In the following step, the sample labels are permuted to simulate the effect of having a known distribution of false results. This means that a former control replicate is now labeled as a bait replicate while a bait replicate turns into a control, and thus a permuted data set is created. A subsequent score is calculated for each protein of the permuted data set. All possible permutations between bait and control labels are conducted, and each time the scores are computed. The number of possible permutations corresponds to the binomial coefficient $\binom{n+m}{n}$, with n and m being the number of replicates of bait or control, respectively. For instance, 69 permutation scores are obtained for each protein in a four versus four setting of bait and control replicates. The resulting empirical distribution of scores for a considered protein corresponds to an estimation of the underlying count distribution. If the original score of the protein exceeds its permutation scores, this indicates that it is better than random chance. In contrast, a protein receives a very weak support if the exchange of control and bait labels leads to a better score than the original one.

The standard approach to estimate a p value for a protein is by calculating the fraction of its permutation scores that are at least as extreme as the score obtained from the original data set. Here a major problem arises in the case of a small number of replicates. Considering three replicates per group corresponds to 20 possible permutations, which leads to a smallest attainable p value of 0.05. In AP-MS settings, the number of replicates tends to be small and requires an integrative procedure. A powerful method in *small sample number, large feature number* situations is the algorithm introduced by Westfall and Young [115]. It integrates the overall ranking of the original scores and accounts for the number of exceeding permutation scores by regarding each protein individually as well as in the context of other candidate proteins. In addition, the total number of proteins is incorporated in a stepwise manner. The entire procedure results in p values controlling the FWER. Thereby, the algorithm of Westfall and Young constitutes a less conservative method compared with other FWER controlling methods, for example, Bonferroni [138], Holm [105], or Hochberg [139]. Because it computes an FWER, the selection of proteins below a threshold of 0.05 refers to the expectation that no falsepositives are contained in the corresponding list of interaction proteins with a probability of 95%. SAINT as well as TSPM can be combined with the Westfall and Young procedure, and their results become comparable based on the FWER.

For TSPM, a second procedure is applicable to calculate p values without the need for permutation sampling, as the underlying distribution is already known. Each protein receives an LRT statistic, which converges to an asymptotic χ^2 distribution [111]. Thus, p values for each protein can directly be inferred from the χ^2 distribution. An additional adjustment of the p values is necessary to account for the number of proteins tested: Here the method of Benjamini−Hochberg [106](which is less conservative than the Westfall and Young method) can be applied to control the FDR. Selecting candidate proteins below a threshold of 0.05 provides a list of interaction proteins while restricting the expected number of false-positives in the list to 5%. This constitutes a different concept to control false-positive interactions in a final list of candidate proteins.

Implementation

The introduced framework is implemented in the package apmsWAPP for R [140] (available from version 2.14), and the TSPM-based workflows are also available in the OpenMS framework [141] and can be downloaded from `https://sourceforge.net/projects/apmswapp/`. Application of the three different workflows in R is based on two main commands, enabling researchers with little knowledge of R to use it, and the OpenMS framework provides a graphical user interface.

2.3. Experimental setup

We conducted two experiments to evaluate the proposed workflows: (i) a simulation study to evaluate the impact of the individual workflow components and to test the reliability of the FDR or FWER, respectively, and (ii) a real data study in *Salmonella Typhimurium.*

2.3.1. Simulation setup

The simulation study is designed to allow evaluating the performance with a well-defined and easily verifiable ground truth. Typical challenges of an AP-MS experiment are simulated including contaminants, low overall number of count signals, or low difference of counts between bait and control. Overall, the simulation comprises eight samples, four repeats of a bait experiment, and four control replicates.

A total set of 500 proteins is simulated, consisting of 400 contaminant proteins and 100 truly interacting proteins. The interacting proteins and contaminants are further separated into different protein classes (see Appendix Fig. A2 and A3a): We include classes of truly interacting proteins that do not have any counts in the control experiments and classes that appear in the control samples but have a stronger presence in the bait experiments. The contaminants are defined by four different classes. To simulate the effects of experimental noise, we rely on the common assumption that spectral counts follow a Poisson distribution [142], which also constitutes the basis for the SAINT model. For all contaminants, counts are simulated from a single Poisson distribution for bait and control samples. In contrast, counts for truly interacting proteins are derived from two different Poisson distributions representing the control and the bait experiment. Thereby, the difference in the Poisson distributions depends on the respective class of truly interacting proteins. (Refer to Appendix Fig. A2) To simulate biases of real AP-MS data, counts of two randomly chosen samples are up- and down-scaled by changing the parameter of the Poisson distribution by a factor of two. Furthermore, two single outliers are added, corresponding to proteins that possess an extremely high count in one of the samples, to challenge the proposed methods.

We conducted a total set of 50 simulations, sampling the different protein classes, to assess the variability and robustness of the results. One additional simulation set was created, in which the counts are sampled from negative binomial distributions to evaluate the performance on a different distribution, and two further simulations assess the robustness for larger sample sizes.

2.3.2. Salmonella data study

The experiment focused on interactions of the export apparatus component SpaS of the type-III secretion system on pathogenicity island 1 of *Salmonella Typhimurium*, comprising three replicate bait and control samples.

The simulation and the real data set were analyzed by applying all combinations of pre- and post-processing and scoring methods. The data were normalized by one of the five proposed normalization methods (sumtotal, DESeq, TMM, upperquartile, quantile) and analyzed with and without filtering. In case filtering is performed, the biological filter and the statistical filter are applied, setting the parameter of the latter to an IQR with a cutoff of 0.3 for the simulated data and to the parameters' overall variance with cutoff of 0.2 for the real data set. A more conservative filtering is appropriate for the real data set, as it contains a smaller number of potentially interacting proteins. Overall, the following three workflows were applied: the first workflow containing SAINT coupled to the permutation-based approach by Westfall and Young (SAINTWY), the alternative workflow integrating TSPM in combination with Westfall and Young (TSPM-WY), or the Benjamini− Hochberg adjustment (TSPM-BH). R-code to reproduce the simulation data as well as all method calls is provided as Supporting Information of the publication.

2.4. Results

2.4.1. Simulation results

In the following section, we investigate the performance of all individual workflow components on the simulation data. We aim to evaluate the impact of the methods on the results and reveal advantages and disadvantages depending on data characteristics. Because the simulation data serves as a ground truth, we can reliably compare the different methods by evaluating (i) how many of the 100 truly interacting proteins in the data are recovered below a multiplicity-adjusted p value of 0.05 and (ii) whether the methods allow controlling the number of false-positives. Note that the methods SAINT-WY and TSPM-WY control the FWER, while TSPM-BH restricts the FDR. The significance level is set to 0.05, holding the FWER or, respectively, the FDR at 5%, and the corresponding results need to be considered separately. We report the median of the number of truly interacting proteins detected by 50 simulations and provide the corresponding 95% confidence interval. Further results using a significance threshold of 0.1 can be found in Appendix Fig. A5 and a more detailed evaluation on detecting the different protein classes, which form the base of the simulation data, are presented in Appendix A. Additional results for the negative binomial simulation study are shown in Appendix Table A1, and robustness by increasing sample size is analyzed in Appendix Fig. A19 and A20.

Preprocessing Impact on the Count Distribution

Normalization and filtering influence the count distribution of the control and bait samples. The effect of normalization is clearly visible in boxplots of counts across the samples before and after normalization (as shown in Figure 2.2). We observe the expected stabilization of count distributions within replicate bait samples and within replicate controls (see Appendix Fig. A6): the quantile normalization forces all count distributions to have the same shape. TMM, DESeq, and upperquartile show a similar tendency, but are less strict. The sumtotal normalization reveals its difficulties with outliers, in terms of extremely high counts, which lead to the repression of the first bait replicate in this example (see Figure 2.2). These minor differences can have major effects on the downstream analysis, as can be seen in the following section.

Considering a count distributions of the different protein classes which were introduced in the simulation data (refer to Appendix Fig. A3a), Figure 2.2 demonstrates that a precise separation of bait and control distributions is obtained by the quantile normalization exemplary for one protein class, which is characterized by a high number of counts across all samples, but a stronger presence in the bait samples. The other classes are visualized in Appendix Fig. A3b .

Filtering of the data strongly reduces the number of interaction candidates by removing a significant number of contaminants, approximately 70% in this case.

Thereby, a complete removal of single-hit contaminants (defined by a very low count in only one sample) is obtained, while the number of truly interacting pro-

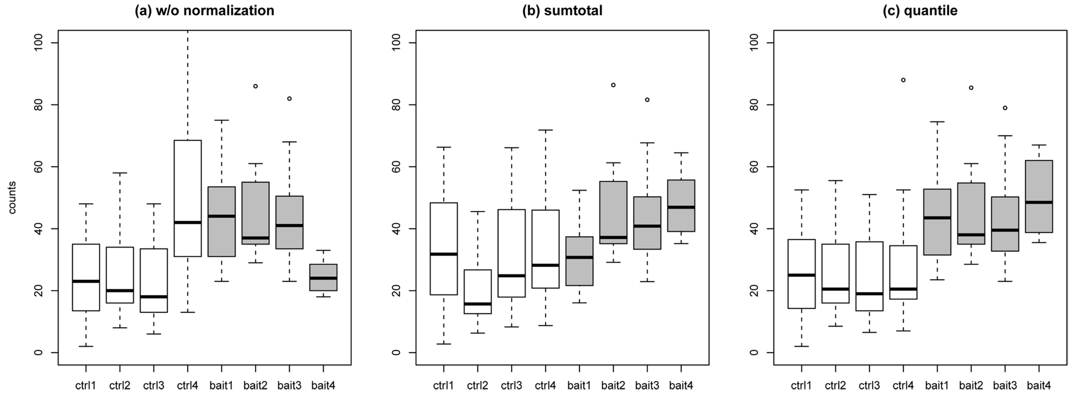


Figure 2.2.: Count distribution of bait and control samples (excerpt) shown for a class of truly interacting proteins holding high counts across all samples but a stronger presence in the baits (in gray): (a) without normalization, (b) with sumtotal normalization, and (c) with quantile normalization of one exemplary selected data set. Boxplots indicate that baits and controls are not clearly separated in cases (a) and (b) but receive a more precise separation by the quantile normalization in case (c), with the medians of all baits being above the upper quartiles of all controls.

teins is almost completely maintained (Refer to Appendix Fig. A7). In particular, between one and five truly interacting proteins are lost due to the filtering depending on the normalization method used (see Appendix Fig. A8), which is acceptable as the benefit of filtering is still larger than its decreasing effect.

Workflow based on SAINT combined with Westfall and Young

Without any preprocessing, SAINT-WY detects on average 47 out of the 100 true interactors. Normalization and preprocessing are crucial (as shown in Figure 2.3) and allow a detection rate of up to 76% for the quantile normalization in combination with filtering, while the sumtotal normalization exhibits the weakest performance. The narrow 95% confidence band points to reliable estimations. The median curve of contaminants, which are found in the corresponding list of proteins assessed below an adjusted p value of 0.05, is close to zero, proving the reliability of the FWER. (refer to Figure 2.3)

Furthermore, we investigate the impact of preprocessing of the data on the SAINT scores itself and observe an increase in the scores for the truly interacting proteins (refer to Appendix Fig. A9). Looking closer at the relationship of SAINT scores and p values obtained for the proteins in one selected set, truly interacting proteins with an adjusted p value below 0.05 show scores in a range from 0.51 to 1.0 (see Appendix Fig. A10). Hence, SAINT-WY also constitutes a robust criterion for generating a cutoff score, while the false-positive rate is controlled, corresponding to a SAINT cutoff score of 0.51 in this example.

Workflow based on TSPM combined with Westfall and Young

Independently of the normalization method used, TSPM-WY exhibits difficulties to detect any of the 100 truly interacting proteins if no filtering of the data is conducted (see Appendix Fig. A14). The filtering step is essential here and enables the median detection of 45 truly interacting proteins, as shown in Figure 2.3, and in combination with normalization a median detection value of 85% and above is attained by the TMM, the upperquartile, and the quantile normalization. In contrast, the median curve of contaminants that were found below an adjusted p value of 0.05 strictly remains zero, proving the correctness of the FWER (see also Appendix Fig. A15 and A16).

The reason TSPM-WY shows a very weak performance without the filtering step is due to single, outlying, highintensity counts that are present in one of the control samples. The affected protein receives an expected small test statistic (score) by TSPM, however, a high test statistic in the permutation sets. Because Westfall and Young is a sensitive method, integrating the information of all proteins, many truly interacting proteins receive a high adjusted p value due to this outlier. The filtering step leads to the removal of the outlier because the biological filter eliminates proteins in the case the median count of the controls exceeds the median count of the bait samples.

Workflow based on TSPM combined with Benjamini-Hochberg

TSPM-BH is per se the less conservative method and already enables a median identification of 57.5 truly interacting proteins without any preprocessing of the data (refer to Figure 2.3). Further normalization and filtering allows a median detection rate of 96% and above of the truly interacting proteins for the TMM, upperquartile, and quantile normalization. In particular, filtering has a significant impact when analyzing the data without normalization or using the sumtotal normalization with an increase of 15% in the number of true interactors. As this approach controls the FDR, a large number of true interactions are expected to be identified; however, more false-positives might also be included. The median contaminant curve reflects this issue, showing one or two contaminants in the final list in 50% of the cases but holding the FDR at the required 5%.

Comparison of the results by SAINT and TSPM

The two workflows SAINT-WY and TSPM-WY both control the FWER, and comparing their results by solely regarding the average number of detected truly interacting proteins shows that more true interactors are identified by TSPM at the same FWER when using normalization and filtering, while SAINT similarly outperforms TSPM when no filtering and normalization are used. It is noteworthy that the 95% confidence band for TSPM-WY exhibits greater variation compared with the more stable estimations obtained for the SAINT workflows. However, a clear separation between the confidence bands of the two methods is observed, indicating that even with this variation TSPM-WY is preferable (see Figure 2.3). Moreover,

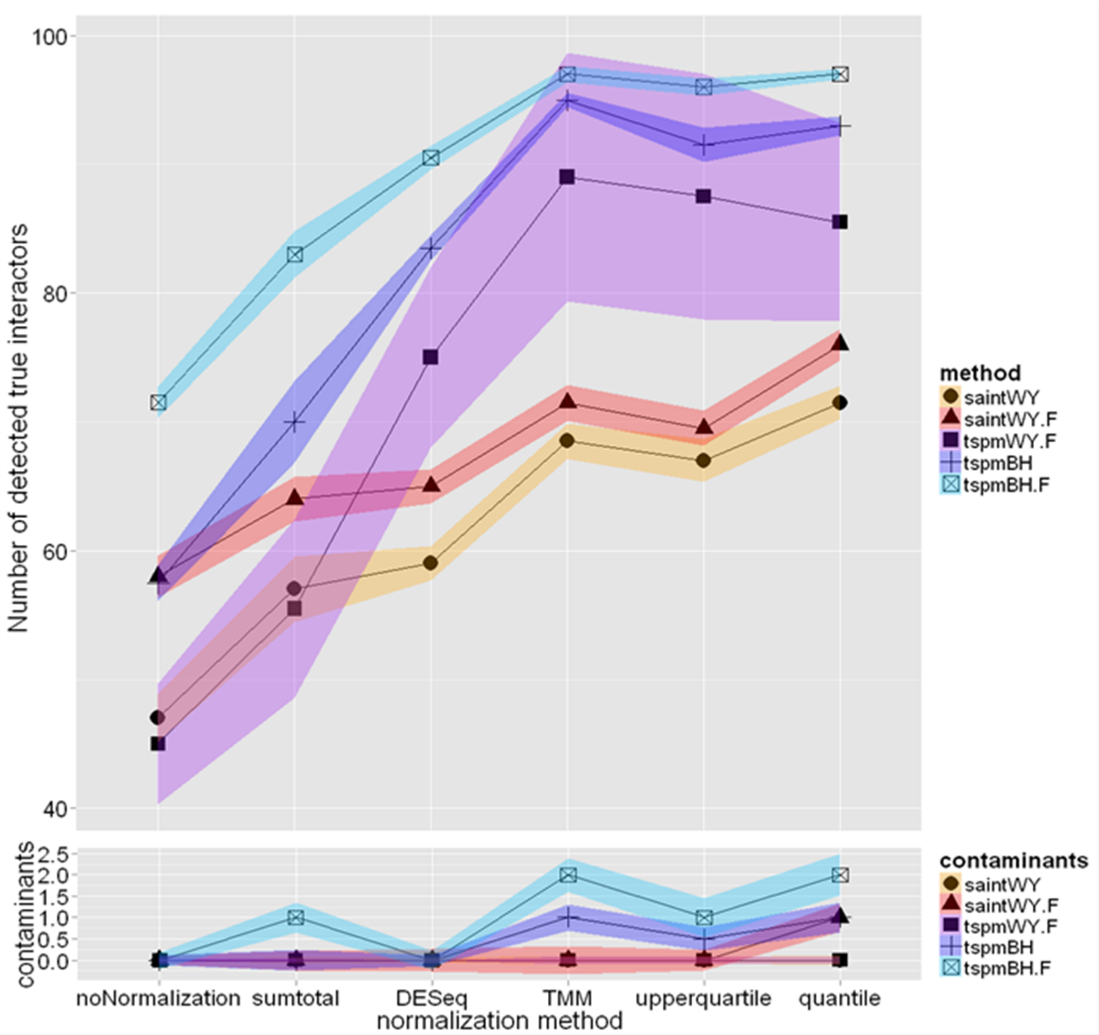


Figure 2.3.: Number of identified truly interacting proteins below a threshold of 0.05 by the different workflows. Median values of 50 simulations and corresponding 95without filtering for (i) SAINT-WY and (ii) TSPM-BH and with filtering for (iii) SAINT-WY (saintWY.F), (iv) TSPM-WY (tspmWY.F), and (v) TSPM-BH (tspmBH.F), according to the normalization method applied (reported on the x axis). A maximum number of 100 true interactors can be obtained based on the ground truth. Median values and 95% confidence bands are presented for the identified false-positives (contaminants) correspondingly.

considering the performance in identifying the true interactors of the individual protein classes, forming the base of the simulation data, reveals diverse strength of the two approaches (see Appendix Fig. A11).

The results indicate that TSPM-WY predominantly values a strong presence in the bait samples independent of low or high counts in the controls. Concerning the issue of outliers in the data, SAINT-WY is not as sensitive as TSPM-WY; detection rates of 40% and above are obtained for classes holding low counts in the controls as long as normalization of the data is conducted independent of the filtering. Without preprocessing of the data, SAINT-WY favors proteins showing a large difference in the counts between bait and control based on small counts in the controls. The less conservative method of TSPM-BH more strongly values small counts in the controls and allows detecting smaller differences between bait and control.

2.4.2. Results for the Salmonella data

In this section, we consider the results obtained from the analysis of the real data study, investigating interactions of the export apparatus component SpaS of the type-III secretion system on pathogenicity island 1 of *Salmonella Typhimurium.* The type-III secretion system is a well-studied macromolecular machinery composed of at least 15 interacting structural proteins [143] (see Appendix Fig. A17). The type-III secretion holo-complex can be subdivided into three parts: needle complex, export apparatus, and cytosolic components. The needle complex builds the structural core of the system and consists of the base proteins PrgH, PrgK, and InvG as well as the filament proteins PrgI [144] and PrgJ [145]. The substrate translocation mediating export apparatus is composed of the polytopic inner membrane proteins InvA, SpaP, SpaQ, SpaR, and SpaS [128]. The cytosolic components InvC, OrgA, OrgB, SpaO, and InvI are thought to prepare substrates for subsequent secretion. The cytosolic components are rather loosely associated with the rest of the complex and are easily lost during purification.

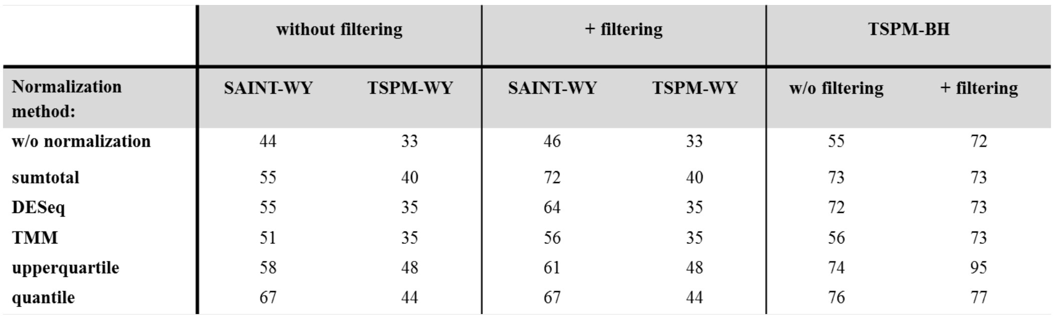
In our analysis, we evaluate whether the expected and wellknown components of the system are identified by the different methods. Furthermore, the experiment predominantly aims to discover potential new interaction candidates being involved. Because this study constitutes rather a screening approach, we decided to choose a less conservative significance level of 0.1. We evaluate the interacting proteins found by the different preprocessing methods and the three workflows SAINT-WY, TSPM-WY, and TSPM-BH.

A common set of 29 interaction candidates is detected by all methods independently of the pre- and post-processing. (See the Supporting Information.) Among those, many known components of the *S. Typhimurium* SPI-1 type-III secretion system needle complex and export apparatus [128] are found: the base components PrgH, PrgK, and InvG, the needle filament protein PrgI, and the export apparatus components InvA, SpaP, and SpaS (the bait).

Depending on the normalization method used, application of the filtering as well as on the chosen scoring method (SAINT or TSPM) and postprocessing (WY and BH), additional proteins can be detected. Considering purely the number of detected proteins obtained by the different methods (see Table 2.1 and Appendix Fig. A18) confirms the trend we observed in the simulation study. SAINT-WY yields more interaction candidates than TSPMWY. TSPM-BH results in more candidates, which is to be clearly expected because the procedure of Westfall and Young is more conservative than Benjamini−Hochberg at the same significance level.

However, the number of detected candidates does not reflect which method per-

Table 2.1.: Number of Identified Interaction Candidates below a Threshold of 0.1 in the *Salmonella* Data Study Investigating Interactions with the Type-III Secretion System (Application of the two FWER-controlled workflows SAINT-WY and TSPM-WY and the FDR-based workflow TSPM-BH for additional screening (i) without normalization (w/o norm.), (ii) with five different normalization methods, (iii) without filtering, and (iv) with filtering of the data.)



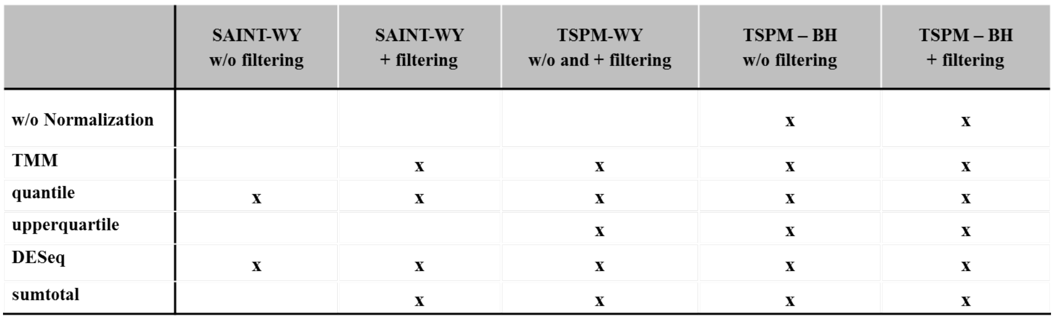
	without filtering		+ filtering		TSPM-BH	
Normalization method:	**SAINT-WY**	**TSPM-WY**	**SAINT-WY**	**TSPM-WY**	**w/o filtering**	**+ filtering**
w/o normalization	44	33	46	33	55	72
sumtotal	55	40	72	40	73	73
DESeq	55	35	64	35	72	73
TMM	51	35	56	35	56	73
upperquartile	58	48	61	48	74	95
quantile	67	44	67	44	76	77

forms best here. In the next step, we evaluate whether additional biological reasonable candidates are identified. First of all, two additional known T3SS needle complex proteins are found, namely, the inner rod protein PrgJ [145] and the export apparatus protein SpaQ [128]. With this, all known components of the cell-envelope-associated T3SS holocomplex were identified except SpaR, which evaded detection by mass spectrometry due to its extremely hydrophobic nature. A very promising and not immediately apparent interaction candidate is Ribonuclease R (*UniProtID: E1WF54*). It has been shown that the S1 RNA-binding domain of this protein can positively regulate the functioning of the T3SS in *Yersinia pestis* and *Yersinia pseudotuberculosis* [146]. It was also shown that RNase R plays a role in the regulation of type-III secreted effector proteins in *Shigella* spp. and enteroinvasive *Escherichia coli* (EIEC) [147]. However, a mechanism for the action of RNase R on T3SS or a direct interaction of this protein with the T3SS needle complex has not yet been presented.

It has been hypothesized that mRNA signals contribute to the targeting of substrates to the translocation machinery of *Yersinia's* T3SSs [148, 149]. The identification of RNA polymerase (E1WEJ7, E1WEJ8, E1WIJ2), ribosomal components (12 in total), and degradosome components (E1W7L4, E1WF54, E1WDY1) indeed suggests a close proximity of transcription and translation components and the T3SS, and this may promote the idea of mRNA targeting.

Our analysis shows that normalizationOur analysis shows that normalization plays a crucial role in the detection of Ribonuclease R (E1WF54) – it is only reported by SAINT-WY if the quantile or DESeq normalization is applied (as shown in Table 2.2). In case the filtering step is added, the candidate can further be detected by the TMM and sumtotal normalization methods in SAINT-WY. TSPM-WY enables the detection of the protein, as long as any of the normalization methods is executed. The less stringent method of TSPM-BH provides the determination of the protein independent of the normalization method or the filtering used. In contrast, the corresponding score calculated by the original SAINT

Table 2.2.: Detection of the Interaction Candidate Ribonuclease R (UniProt ID: E1WF54) below a Threshold of 0.1 (Denoted by x) by the methods SAINT-WY, TSPM-WY, and TSPM-BH with or without Filtering, Respectively, and Dependent on the Normalization Method Applied



	SAINT-WY w/o filtering	SAINT-WY + filtering	TSPM-WY w/o and + filtering	TSPM – BH w/o filtering	TSPM – BH + filtering
w/o Normalization				x	x
TMM		x	x	x	x
quantile	x	x	x	x	x
upperquartile			x	x	x
DESeq	x	x	x	x	x
sumtotal		x	x	x	x

without filtering is 0.572.

There are a number of further potentially relevant interaction candidates in which the preprocessing and scoring method decides whether the corresponding protein is found or remains undetected(see Appendix Table A2). For instance, normalization and filtering have an impact on the detection of the proteins L5 (E1WIK5), L15 (E1WIJ8), RpoA (E1WIJ2), L16 (E1WIL0), S11 (O54296), L17 (E1WIJ1), and HflC (E1WF51). Filtering, in general, enables a greater or even complete independence of the choice of normalization method. SAINT and TSPM rely on different features for scoring protein candidates, consequently leading to a diverse assessment as well as preference for some proteins. SAINT is solely responsible for the determination of protein S12 (E1WIM5) and YajC (E1W8R7), while TSPM is the only scoring method to detect HflK (E1WF50), FtsH (E1WI79), and HtpX (E1WG81), in most cases with the support of all normalization methods independent of filtering. Interaction candidate HtpX and the FtsH holo-complex consisting of FtsH, HflC, and HflK are the major proteases responsible for the turnover of integral inner membrane proteins in bacteria [150]. It is conceivable that these proteases are also involved in the quality control of the T3SS needle complex.

2.5. Discussion of results

In this contribution, we introduce a complete workflow for the analysis of AP-MS data, embedding a scoring method for interaction proteins into a pre- and post-processing framework. Preprocessing of data plays an important role in the analysis of genomic data; however, normalization or statistical filtering has so far not been considered in the analysis of AP-MS data.

To date, to our knowledge, sumtotal normalization is the only normalization method commonly applied in AP-MS data analysis. We implemented and investigated the performance of four additional normalization methods from microarray and RNA-seq analysis and adapted it to the features of AP-MS data. We account for the difference in control and bait experiments and solve the issue of fewer identified proteins in controls by a median rescaling approach. Our simulation study demonstrates the significant impact of normalization methods on the detection of truly interacting proteins − an increase of 20− 40% in the number of true interactors can be obtained. Different promising interaction candidates are also found in the Salmonella study due to normalization. However, normalization methods can vary in performance depending on data characteristics. (Refer to the section 'Evaluation of Results and Discussion of Merits in Appendix A)

As a second preprocessing step, we introduced a biological and statistical filtering of the data to remove obvious contaminants in an early stage and to reduce the multiple testing problem correspondingly. In the case of large and noisy data sets, filtering enables a more sensitive detection of true interactions, as our simulation study demonstrates. In contrast, the Salmonella data set is small and received high-quality measurements; hence filtering of the data is less crucial and results in only minor improvements. Furthermore, the framework enables an extension of the filtering to include additional prefilters, so that future studies can, for example, benefit from contaminant lists provided by the CRAPome database.

After preprocessing of the data, we investigated the performance of two different scoring methods − SAINT and TSPM − to evaluate the interaction potential of a protein. SAINT is a well-established method, and our simulation study confirms its overall good performance. As an alternative scoring scheme, we introduce TSPM, which we adapted to AP-MS data. TSPM is based on quantitative measures and labeling of bait and control only. The simulation study proves its efficiency in successfully separating truly interacting proteins from contaminant proteins. The identification of promising interaction candidates in the *Salmonella data* further supports the choice of TSPM as a new scoring scheme for AP-MS data.

We observe diverse features of the two proposed scoring methods by investigating different protein classes in the simulation study, which may result in a diverse assessment of proteins, as shown in the *Salmonella* data study. However, it is not our aim to favor one of the two methods, SAINT and TSPM; we showed strength and pitfalls of both methods. We note that the choice of normalization and filtering is far more impactful than the choice of the scoring scheme. In general, the overall idea of the proposed workflows is also applicable to other scoring schemes.

For postprocessing, we aimed at replacing scores by p values, which allow the estimation of false-positive interactions in a final list of candidate proteins. We proposed a permutation approach combined with the integrative procedure of Westfall and Young to calculate p values that are controlled by the FWER. Considering the simulation results, SAINT scores of the selected proteins range from 0.5 to 1.0. This indicates how difficult it can be to set thresholds and that many truly interacting proteins may be missed by subjectively set thresholds. Thus, the proposed approach constitutes a robust criterion for generating a cutoff score in a list of interaction proteins produced by SAINT or any other scoring scheme.

This also addresses the stated need for appropriate benchmark data sets to validate and compare the performance of different methods for the analysis of AP-MS data as well as to assess the accuracy of their error estimation procedures.3 The proposed approach can be transferred to any scoring scheme, thus providing a basis for comparison studies. Our findings indicate that the error estimation proposed by SAINT, averaging the complement of the scores for the selected interaction proteins, is more conservative (see Appendix Table A3).

Another issue in AP-MS studies is small sample size, and thus it is advisable to use integrative procedures. The algorithm of Westfall and Young is powerful in these settings and constitutes a less conservative method compared with other FWER controlling methods. Moreover, the conservative nature of the method ensures the generation of a highly reliable list of interaction proteins.

The method of TSPM can be combined with two different postprocessing concepts. We can apply the permutation-based procedure of Westfall and Young to the TSPM test statistics to allow comparisons to SAINT. Alternatively, p values can be directly calculated from a χ^2 distribution combined with a less conservative adjustment such as the Benjamini−Hochberg method. Hence, TSPM enables us to use a less conservative approach for detecting true interactions in AP-MS data by controlling false-positives by an FDR.

Our approach currently relies on the presence of negative controls, while an alternative strategy in AP-MS experiments is to use different bait experiments. Our framework is not directly applicable to this setting. When permuting negative controls with baits, it is ensured that the controls just contain noise, while baits always carry information. In future research, we will evaluate whether an iterative procedure starting with the strongest signals may allow the inclusion of different baits.

To summarize and give a guideline for the potential user analyzing AP-MS data (see also Appendix Fig. A21 for guideline overview): Normalization of the data is crucial; the quantile normalization is based on good experience in other fields and has also proven its successful application in our study. Filtering of the data is meaningful, but a low cutoff should be chosen in case no additional biological knowledge is available. The choice of the three proposed workflows SAINT-WY, TSPM-WY, or TSPMBH depends on the intention of the experiment and the significance threshold needs to be adapted correspondingly. In case a highly reliable group of true interaction proteins should be identified − rather accepting to lose some true interactors than to include false-positive candidates − the best choice is SAINT-WY or, given there are no outliers in the data set, TSPM-WY. If the experiment constitutes a screening approach to find new candidates, accepting a certain and controlled amount of false-positive hits, TSPM-BH should be used. Using the different methods simultaneously is also an option and increases the reliability of candidates, which are independently supported; however, the results need to be carefully integrated and interpreted according to the different FWER/FDR concepts.

3. iPQF: a new peptide-to-protein summarization method

Isobaric labelling techniques such as iTRAQ and TMT are popular methods for relative protein abundance estimation in proteomic studies. However, measurements are assessed at the peptide spectrum level and exhibit substantial heterogeneity per protein. Hence, clever summarization strategies are required to infer protein ratios. So far, current methods rely exclusively on quantitative values, while additional information on peptides is available, yet it is not considered in these methods.
iPQF (isobaric Protein Quantification based on Features) is presented as a novel peptide-to-protein summarization method, which integrates peptide spectra characteristics as well as quantitative values for protein ratio estimation. Diverse features characterizing spectra reliability are investigated and significant correlations to ratio accuracy in spectra are revealed. As a result, a feature- based weighting of peptide spectra is developed. The iPQF algorithm is available within the established R/Bioconductor package *MSnbase* (version $\geq$ 1.17.8).
A performance evaluation of iPQF in comparison to nine different protein ratio inference methods is conducted on five published MS2 and MS3 data sets with predefined ground truth. This work demonstrates the benefit of using peptide feature information to improve protein ratio estimation. Compared to purely quantitative approaches, our proposed strategy achieves increased accuracy by addressing peptide spectra reliability.

3.1. Peptide-to-protein summarization

Mass spectrometry based proteomics has evolved as the method of choice for identification and quantification of proteins [1], and major advances were achieved in the development of new quantification techniques. Isobaric labelling techniques such as iTRAQ and TMT have gained much popularity, allowing for simultaneous absolute and relative protein quantification in different samples within a single run [33, 35, 36]. This enables the investigation of changes in protein abundance across various conditions, which is crucial for the study of regulation processes, diagnostics research, and biomarker studies. Thereby, accuracy in protein ratio estimates plays an essential role. However, accuracy problems in iTRAQ and TMT data have been demonstrated by different studies [66, 68, 85, 117, 151] and reliable protein ratio estimation remains a challenging task.
Several steps are involved in the quantification process. First, peptides are identified and quantified by iTRAQ or TMT reporter ions in the MS/MS spectra. Factors contributing directly to the variability of peptide quantitative estimates

include: efficiency of protein digestion and labeling, co-eluting peptides, reporter ion peak detection, intensity assessment, label interference, and a limited dynamic range of the instrument [28, 152]. A frequently reported bias is the underestimation of ratios and its compression towards one, which is supposed to arise from co-eluting peptides [41,42,153]. Several approaches address these issues by proposing specific sample preparations [152], intensity calculation methods and correction strategies [28, 51, 52]. Further MS3 data acquisition is considered as a new promising strategy to reduce and potentially eliminate the peptide interference effect [48]. The next major step in this process is the inference from peptides to proteins. Measurements of label intensities are assessed at the spectra level and subsequently a summarization strategy is needed to estimate the corresponding protein ratios. Generally, all peptide spectra assigned to a protein are assumed to share the same expression profile. Indeed substantial variance heterogeneity is observed due to random and systematic biases [65, 66]. The question arises how a peptide-to-protein summarization method can appropriately address this existing variance heterogeneity. Different studies demonstrated that the coefficient of variance is dependent on the absolute signal intensity, suffering from higher variation in low-intensity than in high-intensity data [66, 68, 116, 117]. Therefore different summarization methods were developed to account for these intensitydependent effects by filtering for low intensity peptides [67], weighting peptides according to their absolute intensities [68, 69] or by applying a variance stabilization method [66] [154]. Other approaches examine the error structure and the underlying ratio distributions and develop noise models accordingly [47, 70]. Further, standard statistical concepts, such as averaging by mean or median, are still one of the most commonly used methods to find protein ratio estimates from a range of peptide quantities. Multiple tools and comprehensive iTRAQ quantification pipelines either offer or are exclusively based on simple median or weighted mean calculations for protein ratio inference [52, 69, 155]. Additionally, strategies for filtering outlying peptide ratios are frequently proposed, including methods like Grubb´s and Dixon's test [156,157]. A different category of approaches requires the integration of replicate samples or spike-in proteins to enable an assessment of the internal experimental variation [68]. All these summarization methods have in common that they only focus on quantitative peptide information in order to infer protein quantities. So far, the main feature, which is extensively studied and related to the reliability of peptide quantities, is the absolute intensity signal. However, there are several additional characteristics of peptides available, which are known to have an impact on the overall reliability of a specific peptide and its measurements.

3.1.1. Objectives

In this work, we identify and investigate the impact of diverse peptide spectra features such as charge state, sequence length, identification score, mass, and a distance metric within uniquely and redundantly measured spectra. We examine how these features correlate with the variance heterogeneity and to which extent they are related to quantification accuracy in spectra. Our aim is to find a combination of feature criteria that allows inferring ratio reliability by using the complementary strength of the features. As a result, we developed *iPQF* which integrates the information of peptide spectra characteristics with given quantitative values. We show the added value of peptide spectra feature information to improve protein ratio estimation.

The proposed algorithm can be combined with any purely quantitative approach. In addition, a fundamental intention was not to disregard any information, but rather to keep peptide spectrum matches and down weight unreliable spectra according to the features instead of losing information by filtering. Further, no internal replicates or specific sample setup in the design of iTRAQ and TMT experiments is required which may restrict applicability.

Finally, we evaluate the performance of our approach on five different published iTRAQ and TMT data sets providing a ground truth of known peptide and protein quantities. Thereby, we consider three MS2 data sets with minimal amount of biases, one MS2 data set showing a high peptide interference effect as well as one MS3 data set. A comparison study with nine commonly used peptide- to-protein summarization methods is conducted. To our knowledge, this is also the most comprehensive comparison study of summarization methods.

3.2. Features of peptide spectra

Considering the relative quantification values of peptide spectra being assigned to the same protein, a substantial heterogeneity is observed (shown in Appendix Fig. B1). The objective of this work is to investigate whether the observed peptide variation can be explained by underlying peptide spectra characteristics. Thereby, we aim to relate diverse features of spectra to the quality of their quantitative information. As a result, the reliability of given peptide spectra can be inferred and protein quantification can be improved by accounting for it.

In order to study the impact of features on the quantification accuracy, we assess the deviance of ratios from the spectra to a given ground truth by calculating the Euclidian distance across all iTRAQ/TMT labels, subsequently referred to as *quantification error*. Next, a correlation study is conducted by calculating Spearman´s correlation coefficient between feature values and the peptide spectra quantification error.

We examine the impact of the following peptide features: identification score, sequence length, charge state, mass, absolute ion intensity, modification state, and a distance metric within uniquely and redundantly measured spectra as explained below. The shared status of a peptide is not considered and corresponding spectra are discarded, as the negative impact of an incorrectly assigned peptide may be larger than the potential gain of an additional peptide for protein ratio estimation. Here, we define a group of *redundant* spectra as several MS/MS events for one peptide, while *unique* spectra are referred to peptides quantified exactly by one MS/MS event. For redundant peptide spectra of a protein, which are subject to the same conditions in the MS experiment, an even higher ratio similarity across channels is expected than among different sequence fragments of a protein. Hence, a peptide spectrum exhibiting ratios diverging from all other ratios in the redundant spectra group is suspected to be less reliable. For each protein we form different groups according to its different redundant spectra and one group pooling all uniquely measured spectra. The idea is that not only the number of spectra per protein matters, but also the degree of ratio similarity within these groups. For each peptide spectrum we compute the mean Euclidian distance of its ratios to the ratios of all other spectra belonging to the same group.

The identification score indicates the correctness of the peptide spectrum match. A low score implies less reliable peptide identification and consequently an uncertainty in the peptide to protein assignment, potentially resulting in an incorrect peptide ratio for the protein ratio calculation.

The impact of absolute ion intensity was already intensively studied and is well known as a key indicator for the reliability of ratio estimates. It has been shown that the accuracy of peptide ratio estimates depends strongly on the involved absolute intensities [66, 116, 117, 158]. Low intensities are expected to be subject to noise and ratios exhibit large variations, while ratio estimates converge to the true value as intensity increases. Here, we calculate the mean absolute intensity across all labels for each spectrum.

Peptide modifications in iTRAQ experiments occur mainly due to enzymatic or sample preparation related reactions. A slightly increased false positive protein identification rate was reported by allowing more modifications to be present [159]. Further, varying peptide expression behavior in a protein and shifted ratios were observed due to modifications. In our investigation, we distinguish between modified and unmodified peptide spectra without further distinguishing specific types of modification.

The features charge state, mass, and sequence length are inter-related and have direct or indirect impact on peptide identification. Higher charge states give rise to a variety of possible fragments carrying diverse amounts of charges. The peptide search space needs to be expanded accordingly and the risk of false-positive identifications is increased as a consequence. Further, long sequence peptides tend to show a bias to higher identification scores compared to short sequences dependent on the identification tool. The importance of these features and their crucial role has also been shown in other work [160–162].

3.3. iPQF algorithm

The proposed algorithm iPQF (isobaric Protein Quantification based on Features) is a peptide-to-protein summarization method. For each peptide spectrum, it requires peptide identification, reporter ion intensities, and assignment to the respective protein. Next, a summarization strategy is required to combine given peptide spectra quantities to estimate protein quantities.

iPQF presents a novel approach by using information of spectra features to evaluate peptide spectra ratios. Spectra receive weights and contribute to the protein quantification according to their reliability.

The algorithm is conducted protein-wise, which means individual protein quantifi-

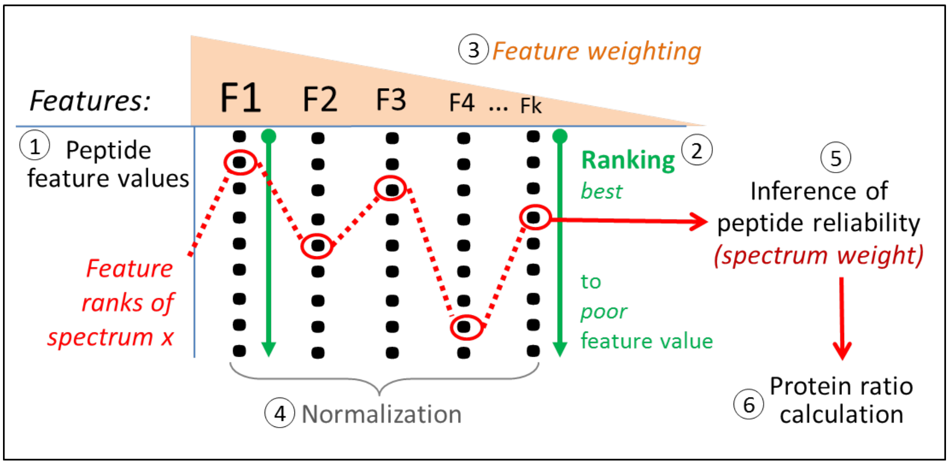


Figure 3.1.: Outline of the iPQF summarization method for protein ratio inference using a feature-based weighting of peptide spectra. Six steps are conducted for each protein to estimate the reliability of its underlying spectra ratios.

cations are not influenced by other protein quantifications in the data set. However, the number of identified peptide spectra per protein is important; we recommend a minimum of three spectra for protein quantification.

The algorithm consists of six steps, which are calculated for each protein (Figure 3.1) (see also example process in Appendix B):

(1) **Feature assessment**: Feature values are computed for each of the seven different features for all peptide spectra belonging to the specific protein.

(2) **Spectra ranking per feature**: Peptide values obtained for each feature are ranked from most to least reliable feature value based on knowledge of associated low and high quantification errors which was acquired in our correlation study. Hence, if a peptide spectrum receives a high rank for a specific feature, this means its reported quantification is considered more reliable by this feature compared to a spectrum showing a lower rank.

(3) **Feature weighting**: For each peptide spectrum we obtain several ranks, one for each presented feature, and each rank individually states the quantification reliability of the spectrum. Yet the explanation power of the features is different, and the impact of the diverse features is weighted according to strong and weak correlations observed with quantification errors (see results in section 3.4). We propose a default weighting order of features based on consensus observations in the different data sets and prove its robustness (for a more detailed explanation and the robustness analysis refer to Appendix B)

(4) **Normalization of ranks**: We normalize the ranks of each feature by the overall number of spectra to ensure the ranks to be within the range of zero and one.

(5) **Inference of overall peptide spectra reliability**: The feature ranks obtained for each spectrum are combined to receive an overall reliability measure called peptide spectrum weight. We do so by calculating a classic average rank per spectrum and normalize it by the weighted sum of all features. As a result, peptide spectra receiving weights close to one represent reliable ratios to enable the inference of true protein ratios, while peptide spectra weights decreasing to zero refer to a reduced confidence in its given quantification values.

(6) **Protein ratio calculation**: A weighted mean approach using squared peptide spectrum weights is conducted to estimate the underlying protein ratio.

Further, iPQF protein estimates can be additionally coupled to pure quantitative strategies using a mean approach, referred to as *combined iPQF* approach here. Generally, we recommend applying the algorithm based on relative spectra intensities in order to estimate protein ratios instead of using absolute intensities. The variance in absolute intensities can be large, while relative intensities are more robust.

3.3.1. Implementation

The introduced iPQF algorithm is implemented in R (version $\geq$ 3.1.3), and was integrated into the existing R/Bioconductor package *MSnbase* (version $\geq$ 1.17.8) [163], which offers a variety of processing functions for iTRAQ data (see MSnbase vignette). Further, the algorithm is designed for optional combination with any summarization method, which focuses exclusively on quantitative values, to combine strengths of both approaches.

3.4. Experimental setup

3.4.1. Dataset description

We evaluate peptide quantification data from five different published MS2 and MS3 data sets based on iTRAQ and TMT experiments, which have predefined protein fold-changes. Thereby we consider three MS2 data sets with smaller fold changes and minimal interference effect as well as one MS2 data set affected by high peptide interference events. Overall, the data sets hold diverse data characteristics concerning the data set size, the number of identified spectra per protein, the expected ratios, and the range of peptide feature values, thus covering different possible protein peptide scenarios.

(1) Data set (MS2) from Hultin-Rosenberg *et al.*: Peptides from a lung cancer cell line A549 were labeled with iTRAQ 8-plex tags according to a 2:2:1:1:2:2:1:1 fold change. Here, the data set showing most identifications in the publication was chosen, which is based on a 400 μg loaded peptide amount, prefractionated by IPG-IEF and analyzed on a LTQ Orbitrap Velos (Thermo Scientific).

Peptide spectra identification and protein inference was performed using Proteome Discoverer 1.1 with Mascot 2.2 (Matrix Science), and identified peptides below a 1FDR level were quantified. Further, peptide intensities were isotope impurity corrected.

(2) Data set (MS2) from Breitwieser *et al.*: A 4-plex iTRAQ experiment was designed with human plasma proteins holding constant ratios of 1:1:1:1 and two spiked-in proteins, a rat ceruplasmin being mixed in 1:2:5:10 ratio concentrations and a mouse ceruplasmin with 10:5:2:1 ratios. MS analysis was conducted on a hybrid LTQ Orbitrap XL (Thermo Scientific) coupled to a HPLC nanoflow system (Agilent 1200). Peptide spectra were searched and quantified using Mascot 2.3 and Phenyx 2.6.1 and only concordant peptide identifications were kept. Protein inference was set to hold an FDR level of 1%.

(3) Data set (MS2) from Zhou *et al.*: Replicate samples from mouse cell lysates were created with equal concentrations, labeled with iTRAQ 8-plex reagents (expected ratios 1:1:1:1:1:1:1:1) and measured by a TripleTOF 5600 (Absciex). The ProteinPilot software was used for peptide spectra identification and quantification, holding the protein FDR below 1%.

(4-5) Data set with MS3 and MS2 spectra from Ting *et al.*: A 6-plex TMT experiment was designed with a two-proteome mixture model containing human cell lines and yeast Lys-C digests to study the peptide interference effect. Yeast peptides were mixed according to 10:4:2.5:10:4:2.5 ratios and human peptides with equal amounts (1:1:1) were added to the first three labels. The MS2 data set presents compressed yeast ratios in the first three labels due to human peptide interference, while in the MS3 data set the interference effect is almost eliminated. Samples were measured on a LTQ Orbitrap Velos. The focus here is on the yeast peptide and protein identification and quantification which was performed by Sequest with a protein FDR of 1.5%.

All data sets were filtered for shared peptides, contaminants, and for spectra showing missing or zero intensities in one of the iTRAQ/TMT labels. An additional filtering was applied in case of MS3 data set (4) due to extreme outliers in the data set (see also filtering by Ting *et al.*), using a less restrictive approach than in the original publication and discarding only spectra deviating more than tenfold from the expected ratios which are biologically not reasonable (Appendix Fig. B2d). Peptide spectra intensities were normalized according to the median intensity present in each label for data set (2) and (3). No normalization was applied in the case of data set (1) and (4)-(5), as this would contrast with the foldchange setting defined for all peptides. Further, protein identifications based on the support of only one or two peptide spectra are not considered for quantification and evaluation here. As a result of the preprocessing, 624 proteins based on 5,885 peptide spectra are considered in data set (1), 145 proteins with 13,758 spectra in data set (2), 2,811 proteins with 217,822 spectra in data set (3), and 781 proteins with 8934 spectra in MS3 data set (4) (processing and analysis of the corresponding data set

(5) with MS2 spectra representing the impact of peptide interference can be found in the Appendix B).

The aim of this work is the computation of accurate protein ratios based on relative peptide intensities; here, we focus mainly on the relative intensity level of proteins and peptide spectra and not on the absolute intensities. Hence, ratios are calculated for all data sets. For data set (2) and (3), a ratio of a spectrum is defined by dividing its absolute intensity of one iTRAQ label by its summed intensities of all labels. This is a robust approach, as it satisfies greater label independence in the ratio calculation and peptide ratios are not exclusively based on one specific label. In case of data set (1), we followed the ratio computation in the corresponding publication, in which intensities were divided by the mean intensity of iTRAQ label 113 and 114. For data set (4)-(5) we relied on the provided ratios.

3.4.2. Method comparison

All introduced data sets come with predefined ratios for spectra and proteins, thus allowing the performance evaluation of diverse peptide-to-protein summarization methods. In order to compare the different summarization methods and to assess their accuracy in estimating protein ratios, we consider the *protein estimation error*. The error is defined as the squared differences of the protein ratio estimates to the ground truth with subsequent summation across labels.

We investigate and compare up to nine commonly used peptideto- protein summarization methods with our proposed iPQF approach. Protein ratios are estimated based on given peptide spectra ratios for each label individually in all presented methods:

- *Median*: The estimated protein ratio corresponds to the median of peptide ratios being assigned to the protein.
- *Mean*: The mean is used instead of the median.
- *Mean (Top5, Top3)*: A group of five or three spectra showing the highest absolute intensities are selected respectively and the mean is applied [164,165].
- *Tukey´s Median Polish*: An additive model is iteratively fitted to the ratios until the sum of absolute residuals falls below a significantly small threshold. The sum of the resulting overall median and label effect, given by the model, is used to estimate the protein ratio. [163, 166]
- *Sum of intensities*: The absolute peptide spectra intensities of one protein are summed for each label. Protein ratios are calculated on the basis of the intensity sums. [158]
- *Total Least Squares*: The objective to find the protein ratio is to fit a straight line between peptide spectra ratios of two different labels. Different from linear regression, here orthogonal distances are minimized between ratios and an optimal line. [158, 167]

- *isobar*: A noise model is built which estimates the underlying noise variance dependent on absolute spectra intensities. The inverse of the noise variance serves as a weighting factor for individual peptide spectra. Protein ratios are subsequently computed by a weighted average approach. [47]

Additionally, a comparison to protein quantification results of Mascot and ProteinPilot, which were provided in the supplements of the corresponding publications, is included for data sets (1) and (3), respectively.

3.5. Results

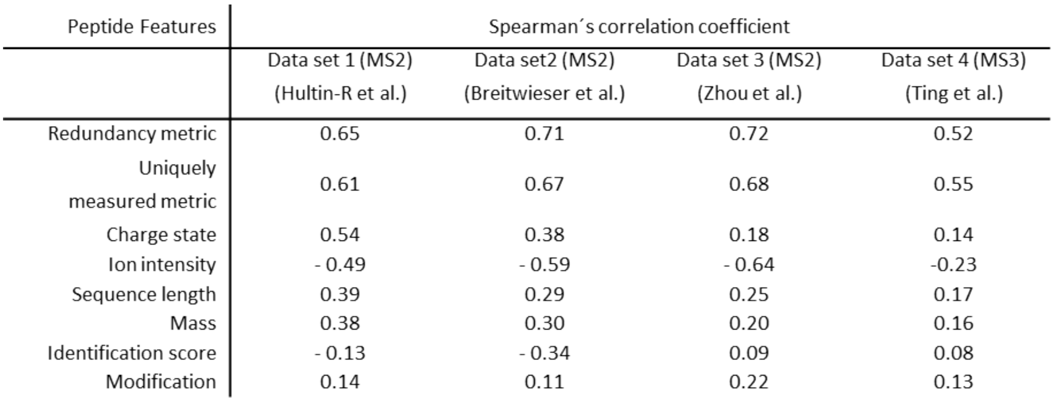
First, we demonstrate the correlation between peptide spectra features and quantification accuracy. Second, we evaluate the performance of the iPQF algorithm in comparison to nine summarization methods. Results are provided for the different MS2 data sets as well as for the additional MS3 data set.

3.5.1. Peptide feature correlation study

The distributions of peptide ratios measured by the different labels are shown in Appendix Fig. B2a-e. Ratio values are spread around the ground truth values of the corresponding data set. Considering the quantification error per spectrum, defined by the Euclidian distance of the measured ratios to the expected ratios, a right skewed distribution is observed in all data sets. The two spike-in proteins of data set (2) (Breitwieser *et.al*) each exhibit a group of strongly diverging peptide spectra ratios from the ground truth, which causes a second peak in the quantification error distribution (Appendix Fig. B2b).

The correlation of quantification errors to peptide spectra features is analyzed to study the feature impact on ratio accuracy. The corresponding Spearman´s correlation coefficients are reported in Table 3.1. All correlation coefficients are assessed to be statistically significant by using Spearman's rho statistic to estimate a rankbased measure of association. Overall, correlations observed are strikingly

Table 3.1.: Correlation study of peptide spectra features to relative quantification error



Peptide Features	Spearman´s correlation coefficient			
	Data set 1 (MS2) (Hultin-R et al.)	Data set2 (MS2) (Breitwieser et al.)	Data set 3 (MS2) (Zhou et al.)	Data set 4 (MS3) (Ting et al.)
Redundancy metric	0.65	0.71	0.72	0.52
Uniquely measured metric	0.61	0.67	0.68	0.55
Charge state	0.54	0.38	0.18	0.14
Ion intensity	- 0.49	- 0.59	- 0.64	-0.23
Sequence length	0.39	0.29	0.25	0.17
Mass	0.38	0.30	0.20	0.16
Identification score	- 0.13	- 0.34	0.09	0.08
Modification	0.14	0.11	0.22	0.13

consistent across the three MS2 data sets, despite different sample complexity, experimental setups, different instrumentation and different analysis software used. Further, even with an additional isolation and fragmentation step resulting in an MS3 scan, the same correlation trend with slightly decreased correlation coefficients is observed.

The most meaningful feature reflecting ratio accuracy is the proposed similarity metric within redundantly and uniquely measured spectra groups, holding positive correlations between 0.52 and 0.72. Hence, a small mean Euclidian distance of a specific peptide spectrum to spectra belonging to the same redundant or unique group, respectively, implies a small quantification error. However, the error increases with the peptide spectrum diverging from its group (Fig. 3.2A, Appendix Fig. B3-4). Further, ratio accuracy is decreasing with increasing charge of a peptide spectrum, especially apparent in the most common range between a charge state of two and four (Fig. 3.2B, Appendix Fig. B5). The increase of noise and ratio variation in low absolute ion intensity data has been shown before and is also confirmed in this study (Appendix Fig. B6). A consistently increasing ratio error is observed with increasing sequence length from mainly 5 to 30 amino acids, illustrated by a positive correlation between 0.17 and 0.39 (Fig. 3.2C, Appendix Fig. B7). The high inter-relation between length and mass of a peptide is also clearly reflected by similar correlation coefficients to the quantification error, further supporting both features as indicators of ratio reliability (Appendix Fig. B8). Correlation of the identification score varies between the data sets due to the different scoring systems, data set (1) and (2) are based on Mascot, while data set (3) relies on ProteinPilot, and data set (4)-(5) on Sequest. Generally higher scores correspond to smaller ratio errors; however, it is interesting to observe that error variation increases at the same time (see further details in Appendix Fig. B9). For the group of modified spectra the ratio error appears to be increased in all data sets compared to non-modified spectra (Appendix Fig. B10).

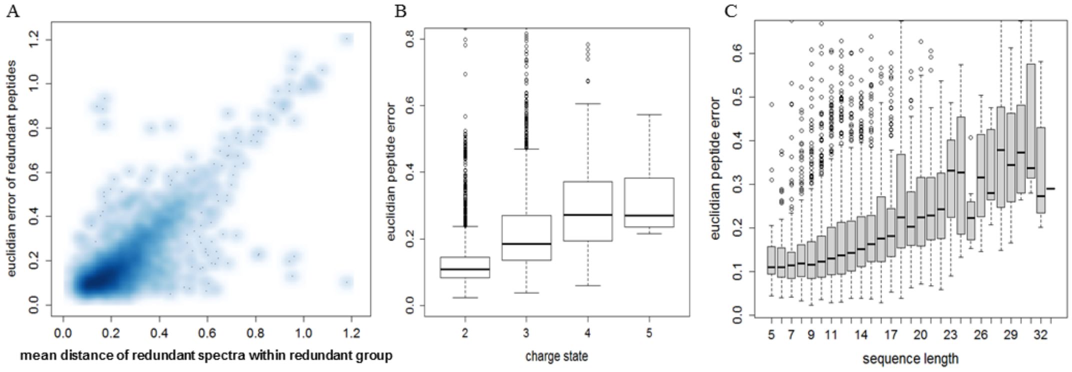


Figure 3.2.: Correlation of spectra features to quantification error, shown for three selected features of data set (1) (Hultin.-Rosenberg et.al.). The impact of the features (A) redundancy metric, (B) charge state, and (C) sequence length on spectra ratio accuracy is displayed. A significant trend is observed in all cases.

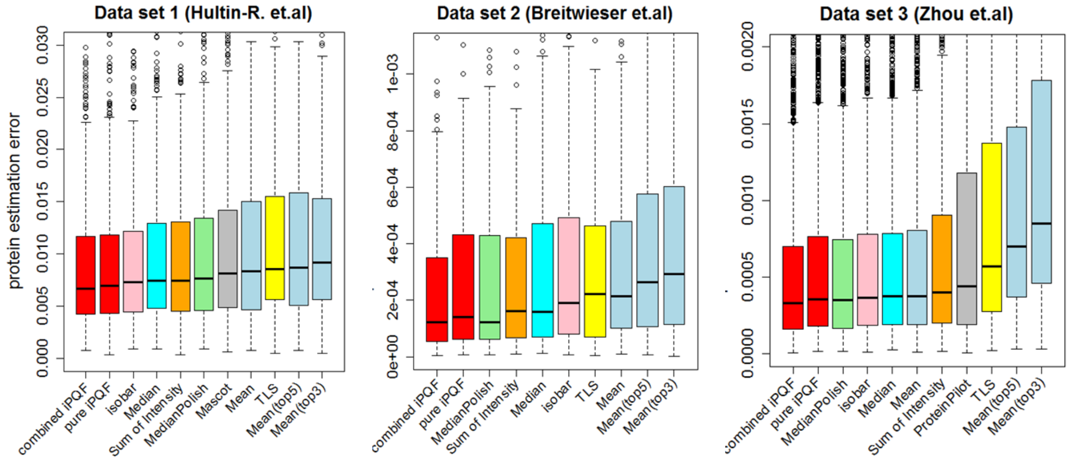


Figure 3.3.: Performance evaluation of iPQF approaches and nine summarization strategies shown for three different MS2 data sets (data set (1): 624 proteins with 5885 spectra; data set (2): 145 proteins with 13758 spectra; data set (3): 2811 protein with 217822 spectra). Boxplots display the protein estimation error of each method applied (note that methods are ordered according to error size). Improved and robust protein ratio accuracy is observed for the iPQF approaches in all three data sets.

The visualization of peptide feature-error-correlations displays a homogenous trend for all data sets, notably for MS2 as well as MS3 data (refer to Appendix Fig. B11). The impact of peptide interference events on feature-error-correlations is shown by means of the data set (5) with MS2 spectra (see Appendix Table B1 and Fig. B12). Additionally, the two spike-in rat and mouse proteins of data set (2) are shown separately in Appendix Fig. B13. In particular, short peptide sequences are assigned to the rat protein and the observed outlier peptide group consists exclusively of redundantly measured spectra showing low absolute intensities.
Further, a study of inter-correlations between features reveals a strong and expected relation structure among features such as length, mass, charge state, and score (Appendix Fig. B14). However, despite significant correlations of individual features, the combination of features is crucial to eliminate pitfalls of single features and make use of opposed strength. The proposed iPQF approach combines the information from all different features to obtain overall ratio reliability for each spectrum.

3.5.2. Evaluation of protein summarization methods

For evaluation of peptide-to-protein summarization methods, we rely on diverse data sets, in particular concerning the number of peptide spectra per protein (Appendix Fig. B15). Data set (1) and (4) consists of a large number of proteins being supported by predominantly three to ten or respectively twenty peptide spectra, while data set (2) comprises only 145 proteins based on a range of three to over hundred spectra. Data set (3) is an overall large data set holding a median of 26 spectra per protein and diverse cases of several hundred spectra.

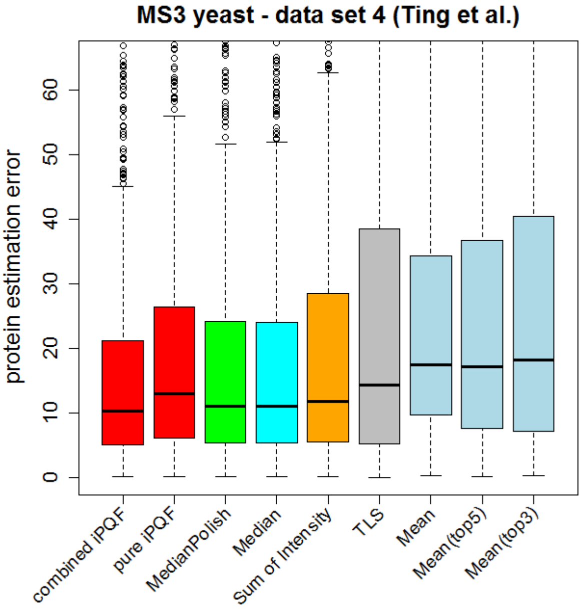


Figure 3.4.: Performance of iPQF approaches and seven other summarization strategies in a MS3 data setting (Note that isobar could not be run on data set 4). A reduced protein estimation error is attained for the combined iPQF strategy, confirming the benefit of features also in MS3 data (781 proteins with 8934 spectra).

We present a performance evaluation of iPQF and nine additional protein ratio inference methods, which are primarily based on quantitative peptide information only. The accuracy of each method is described by the *protein estimation error*, which is assessed for each protein of a data set, and a statistical summary is displayed in form of boxplots. Method comparisons are shown for three MS2 data sets in Figure 3.3. We present two forms of iPQF results, the *pure* form of iPQF using spectra feature information only and a *combined* form in which iPQF is coupled to one of the quantitative approaches, here shown for iPQF combined with the MedianPolish approach. The combined form illustrates the added value of feature information to quantitative approaches.

Overall, the pure iPQF approach shows better performance in data set (1) and comparable performance to the other summarization methods in data set (2) and (3), proving the importance of feature information. The combined iPQF approach exhibits the best protein ratio accuracy of all methods in each of the three different data sets. Further, iPQF approaches prove robustness, while other methods vary in performance dependent on the data set applied.

In particular, feature information is of high value in data set (1), which is dominated by small peptide spectra numbers per protein corresponding to sparsely available quantitative information. Thus, iPQF approaches perform best using all additional knowledge to weight spectra, while the diverse mean-based approaches struggle most due to high variation within protein profiles based on few spectra. The more robust and sophisticated approaches show an intermediate performance.

As spectra numbers vary more in data set (2), the pure iPQF approach becomes comparable to the other methods; however, the combined iPQF improves over all methods by throughout lower quantiles including a significantly reduced upperquartile of the estimation error. The large spectra numbers in data set (3) result in similar quantiles of estimation errors of most methods, even the mean approach performs equivalent to the more robust median and all other sophisticated approaches. In contrast, mean (top5/ top3) methods restrict themselves to few peptide spectra with high absolute intensity and have a significant performance loss. Also in this data set, the combined iPQF achieves improved ratio accuracy, shown by consistently lower quantiles. Generally, the commercial and commonly used tools Mascot and ProteinPilot do not show competitive performance, here.

Evaluation of iPQF in MS3 data is presented in Figure 3.4 and also confirms superior performance of the combined iPQF approach, while pure iPQF shows comparable results to other approaches. Generally high protein estimation errors are observed due to many outlying ratios in the data set which significantly impact the performance of the mean based approaches. A performance comparison of the methods on data set (5) being affected by peptide interference also supports the integration of feature information (Appendix Fig. B16).

Further, we evaluate accuracy details of the methods by considering specific deviation ranges from the ground truth ratios and assess the amount of protein ratios which could be estimated within this deviation range. A superior sensitivity can be observed for the iPQF approaches (Appendix Fig. B17). Additionally, the AUC measure (area under the curve) is provided for all methods, showing the highest AUC for the combined iPQF (Appendix Table B2).

3.6. Discussion of results

Inference of protein ratios based on heterogeneous peptide spectra measurements remains a crucial issue, which receives little consideration in most quantification pipelines. In this work, we present a new summarization strategy *iPQF*, which integrates spectra characteristics with quantitative values for protein ratio estimation. We investigate different peptide spectra features and reveal significant correlations between features and quantification accuracy. As a result, we are able to show the added value of feature information to achieve improved protein ratio accuracy.

Peptide spectra features contain valuable information in addition to pure quantitative information. Since no individual feature shows near-perfect correlation to quantification error, the combination of features can be crucial to compensate for failures of individual features and to make use of their diverse strengths. Overall, it is unlikely that a peptide spectrum is mischaracterized by a large set of features at the same time.

In particular, proteins with a high diversity of underlying feature values profit from the approach, while feature uniformity naturally reduces the impact by giving similar weights to spectra. This is primarily relevant for proteins holding a small to medium number of peptide spectrum matches exhibiting ratio variation. Here particularly, benefit of the iPQF approach is shown. In cases of large numbers of peptide spectra the feature impact is decreased and approaches using the mean already perform considerably well.

Another prerequisite for successful protein ratio estimation is that peptide ratio measurements are spread around the true protein ratio value. The best protein quantification method still remains dependent on given peptide quantities, and cannot work if peptide values coherently and systematically diverge from the ground truth. Feature and quantification error correlation will also not necessarily be sufficiently strong in these divergent cases as the error is strongly biased.

A major issue in iTRAQ and TMT data sets is the peptide interference effect which causes the underestimation of ratios and its compression towards one. MS3 data acquisition has proven to significantly reduce the interference effect. Evaluation of iPQF approaches also confirms the applicability in MS3 data settings and still shows a robust performance under interference impact compared to other methods.

The flexible design of the algorithm enables further extensions. One option is to join results of a purely quantitative method with estimations obtained by iPQF to benefit from both strategies. Here, we provide a combination of iPQF with MedianPolish and show significant improvements over both individual methods in our results. The advantage of a joined approach is that in case of few peptide spectra per protein additional feature knowledge can compensate for the sparse information in the quantitative setup, while more sophisticated summarization strategies can be applied with rich quantitative information available. Further, a different option is to exclusively employ the spectra feature-based reliability measure provided by iPQF and integrate it in existing summarization approaches. Beyond this, new and relevant features of interest can be easily added to the implemented feature framework.

Generally, the idea of a feature-based spectra weighting is transferable beyond iTRAQ data. While our studies only focus on feature- error-correlations in iTRAQ and TMT data, similar findings are expected for SILAC as well as label-free data. Algorithmic steps of iPQF are technically applicable to quantitative proteomic methods requiring peptide summarization, but careful evaluation in the context of the specific experiment is necessary.

Moreover our proposed approach is independent of using replicate samples or spike-in proteins, independent of the instrument, and the selected multiplex. Further, in contrast to modelling approaches mostly requiring larger numbers of peptide spectra, iPQF is also applicable in small settings. Also no assumption concerning underlying ratio distributions or specific data criteria is required. Hence, we also chose replicate independent summarization methods and use corresponding settings in tools, such as isobar [47], for evaluation comparison.

In addition, a fundamental intention was to keep peptide spectra by applying a feature based weighting instead of losing information by filtering. Filtering of low-intensity spectra or outlier ratios is commonly performed; however this significantly reduces the protein coverage as few peptide readings per protein typically dominate the data sets [66]. Further, defining a cutoff for outlier filtering is a critical issue as important information is potentially discarded.

Overall, we provide a broad performance comparison of nine different protein ratio inference methods on five published data sets with predefined ground truth. To the best of our knowledge, an overall benchmark study of current methods assessing diverse biases and impact on protein ratio accuracy in iTRAQ/TMT data is missing. Here, we also provide a basis for future comparison of summarization methods.

To summarize, the goal of the protein quantification process is the inference of protein quantities based on peptide quantities. However, peptide ratios assigned to a protein exhibit substantial heterogeneity and require clever summarization strategies. In this work *iPQF* is presented, which integrates peptide spectra characteristics as well as quantitative values for protein ratio estimation. The novelty of the algorithm is to weight spectra according to their feature reliability. Comprehensive evaluation of iPQF in comparison to other summarization methods yields a superior and robust performance. As a result, the benefit of feature information to achieve improved protein ratio accuracy is shown.

4. DiTASiC: quantification on strain level in metagenomics data

Current metagenomics approaches allow analyzing the composition of microbial communities at high resolution. Important changes to the composition are known to even occur on strain level and to go hand in hand with changes in disease or ecological state. However, specific challenges arise for strain level analysis due to highly similar genome sequences present. Only a limited number of tools approach taxa abundance estimation beyond species level and there is a strong need for dedicated tools for strain resolution and differential abundance testing.

DiTASiC (Differential Taxa Abundance including Similarity Correction) is presented as a novel approach for quantification and differential assessment of individual taxa in metagenomics samples. A generalized linear model is introduced for the resolution of shared read counts which cause a significant bias on strain level. Further, we capture abundance estimation uncertainties, which play a crucial role in differential abundance analysis. A novel statistical framework is built, which integrates the abundance variance and infers abundance distributions for differential testing sensitive to strain level.

As a result, we obtain highly accurate abundance estimates down to sub-strain level and enable fine-grained resolution of strain clusters. We demonstrate the relevance of read ambiguity resolution and integration of abundance uncertainties for differential analysis. Accurate detections of even small changes are achieved and false-positives are significantly reduced. Superior performance is shown on latest benchmark sets of various complexities and in comparison to existing methods. DiTASiC code is freely available from `https://rki_bioinformatics.gitlab.io/ditasic`.

4.1. Metagenomics profiling

Rapid advances in NGS technologies have revolutionized the field of metagenomics [12,24]. Metagenomics enables the study of complex communities in environmental or human samples by direct analysis of whole shotgun metagenomes, without prior need for cultivation. Among others, two major goals in metagenomics profiling studies are pursued. One is to unravel the taxonomic composition of the community in a given sample, the second concerns the abundance change of taxa between different metagenomes [168].

Especially, differences occurring on strain level in microbiomes can be of high relevance for disease and health state [169]. Investigations on strain level have been proven to be crucial for the understanding of evolutionary processes, adaption, pathogenicity, drug resistance, and transmission [170–173]. However, although importance of resolution on strain level is acknowledged, there are still only a limited number of tools focusing on accurate profiling beyond species level [118]. Altogether, in this context, three main concepts are relevant: strain identification, abundance estimation, and differential abundance assessment. Our objective in this work is to address all these steps by specifically focusing on strain level resolution and its arising challenges. In particular for differential abundance evaluation on the strain level, there is a need for novel tools. Here, we use the term strain level referring to the highest possible resolution available and always work on the exact genome level.

Many concepts have been pioneered for taxa identification and quantification, apart from assembly and binning methods, diverse metagenomics profiling tools have specialized on this task [118, 174]. In practice, these concern the assignment of the sequenced reads to taxa and corresponding inference of taxa abundance. Read assignment can be conducted either by the full alignment of reads to genome sequences or by using pseudo-alignment approaches [21]. The latter is sufficient for many metagenomics quantification studies due to the fact that only the assignment of reads is required and not exact alignments. Another variant is to rely on marker genes instead of complete genome sequences [175–177], however, a general drawback is the requirement of high sequencing coverage contrasting typical metagenomics scenarios of many low abundant taxa [178]. One of the first and popular reference-based tools for read assignment in metagenomics was MEGAN [179], which assigns the reads to the lowest common ancestor in the taxonomic tree at which a unique alignment is achieved. However, this approach limits MEGAN to the identification and quantification of only higher taxonomic levels. A main characteristic on strain level is the presence of highly similar reference sequences, causing many reads to match to multiple genomes equally. A further common practice is to assign multiply mapped reads heuristically to reference genomes according to uniquely mapped read proportions [22, 72]. Yet, this can easily result in biased abundance estimates due to reference sequence similarities as observed for example by Liu *et al.* (2017). GRAMMy [74] and GASiC [73] were the first tools to include reference genome similarities in a model for the resolution of ambiguously mapped reads. Since being based on read alignments, these methods can encounter computational limits in large sample sizes. A new era evolved by utilizing fast k-mer approaches, significantly accelerating read assignments, with Kraken being a popular representative [21], but showing reduced resolution power on strain level [180]. As a consequence, the importance of combining fast mapping approaches with methods for read ambiguity resolution was recognized. This was likewise applied in the field of RNA-Seq, resulting in the development of kallisto [75], which promises to also support metagenomics abundance analysis [180]. kallisto consists of two parts, a new fast pseudo-aligner based on k-mer hashing and an expectation-maximization (EM) algorithm on equivalence classes, which carries out the statistical resolution of read ambiguities.

4.1.1. Project objectives

In this work, we present DiTASiC which relies on pseudo-alignments for mapping and is built on a novel generalized linear model (GLM) framework for read ambiguity resolution. Hereby, we significantly improve on our previous development in this field, GASiC. Our new model framework is developed to adapt more precisely to the characteristics of absolute mapping count data observed for taxa. Moreover, our method improves on existing pure abundance profiling strategies by including additional error terms in the model and capturing abundance estimation uncertainties.

The integration of variance of abundance estimates plays a crucial role for the differential abundance analysis. This variance reflects the uncertainty in the resolution of read mapping ambiguities in the presence of similar reference sequences. Hence, it is of particular importance on strain level to integrate this variance to enable accurate detections of differential or non-differential abundance of a taxon in co-existence of similar strains, most notably in the case of smaller changes.

Most approaches developed for identification of differential abundance in the field of comparative metagenomics focus exclusively on experimental sources of variance, namely on sample variance relevant within technical and biological replicates. A large variety of tools is available [98]; amongst others, software packages implementing diverse parametric and non-parametric statistical standard tests [181–185]. Another group comprises zero-inflated models either combined with Gaussian mixture distribution [99], log-normal distribution [100], or beta-regression [101], concentrating on the potential sparsity in count data. Further, popular methods from RNA-Seq analysis such as edgeR [131], DESeq2 [96] and voom [186] are also commonly applied in comparative metagenomics. Without doubt, the integration of experimental variance is of high necessity when comparing groups of samples. However, here, we want to emphasize and raise awareness for variance in abundance estimates and its impact on differential abundance analysis on strain level.

Further it should be noted that many methods treat the differential assessment of taxa and genes equivalently. However, assumptions such as the majority of features will show non-differential abundance, which has widely been proven for gene expression, are not necessarily valid for taxa abundance in a sample. Antibiotics treatment and other life influential factors have shown rapid changes of microbial compositions in human samples [187] and similar scenarios are found in ecological environments [188]. Thus, commonly used assumptions cannot be easily transferred to composition change.

In summary, we present DiTASiC, which addresses abundance estimation as well as differential abundance of taxa specifically focusing on strain level. A new GLM framework is proposed for resolution of read mapping ambiguities and allows inference of highly accurate taxa abundance estimates. Second, a statistical framework, which integrates abundance estimate uncertainties, is built for differential abundance testing. Here, no prior assumptions on overall composition change are required. A resulting list of tested taxa is reported with estimated abundances, fold-changes and p-values to infer significance. The performance of DiTASiC is evaluated on different metagenomics data sets from four different data sources and in comparison to existing tools.

4.2. DiTASiC workflow

DiTASiC is designed as a comprehensive approach for abundance estimation and differential abundance assessment of individual taxa. Thereby, the main focus is on distinguishing on the strain level with highly similar sequences and its corresponding challenges. The steps of the DiTASiC workflow are illustrated in Figure 4.1; it consists of three main parts: mapping, abundance estimation, and differential abundance assessment.

In the first two parts we built on some of the core ideas of our previously published tool GASiC [73], while strongly improving on abundance quantification and introducing new methodology to address the critical aspects of variance of abundance estimates and differential abundance.

In a metagenomics sample measured by NGS technologies we face millions to billions of reads which are derived from diverse taxa. DiTASiC relies on a pre-filtering of species by fast profiling tools such as Kraken [21], CLARK [189], or Kaiju [190], or by using Mash [191], a genome distance calculator, to reduce the number of potential reference genomes and keep the main focus on species expected in the data. Here, we specifically aim at revealing the picture on the highest available strain levels. In the first **mapping step**, all reads are assigned to the given references as a first attempt to decipher their potential origin. The number of hits per reference genome is counted. We refer to it as *mapping abundance* of a taxon. In the next step of **abundance estimation**, a new generalized linear model is introduced for the resolution of *shared* read counts, which are crucial on strain level. As a result, more accurate abundance estimates are obtained for the different strains along with standard errors for abundance uncertainty. In the last part, the focus is on the comparison of whole metagenomics samples and the assessment of **differential abundance** of taxa. Thereby, we concentrate on a method to integrate the variance of abundance estimates. Abundances are transformed into distributions, divergence of distributions is used to infer differential events and corresponding p-values are calculated. The details of the three DiTASiC parts are explained in the following sections.

The following notation is applied: different metagenomics samples are denoted as $D = \{ D_k, \text{k=1,\ldots,K} \}$, each containing $N = \{ N_k, \text{k=1,\ldots,K} \}$ total input reads. A set of taxa $S = \{ S_i, \text{i=1,\ldots,M} \}$ with known reference sequences is considered. Thereby, S_i is synonymously used for both the taxa itself as well as its exact reference genome. Mapping and abundance estimation are addressed for each data set separately, while the last step of differential abundance estimation is defined on a pair of samples from D.

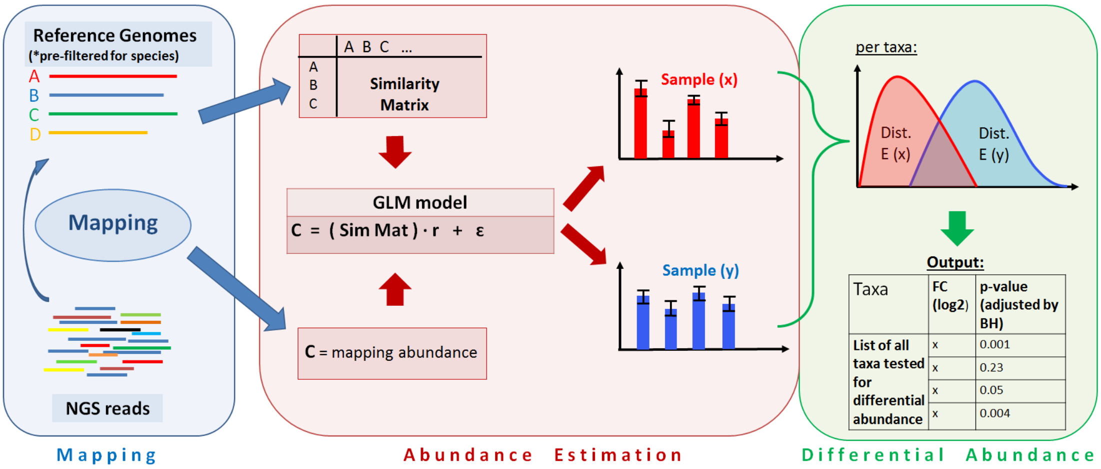


Figure 4.1.: Workflow of DiTASiC. It consists of three main parts: (1) mapping, (2) taxa abundance estimation, and (3) differential abundance assessment. (1) We rely on prior pre-filtering of species by external profiling tools such as Kraken or Mash. Reads are mapped to the given reference genome sequences and the number of matching reads per reference are counted (mapping abundance). A similarity matrix reflecting the genome similarities is constructed. (2) Subsequently, a generalized linear model (GLM) is built for resolution of read count ambiguities, resulting in corrected abundance estimates along with standard errors. (3) For the comparison of metagenomes, abundances are formulated as distributions and their divergence reflects differential events. A final list of tested taxa with fold change and adjusted p-values is reported.

4.2.1. Mapping step

To identify their origin, the assignment of reads is conducted by a *competitive* mapping approach, which means all selected reference genome sequences S are simultaneously offered to all reads of a sample D $\in \mathcal{D}$ for mapping. Particularly on strain level, reference sequences exhibit high sequence similarities, thus some reads are expected to match to different genome sequences equally well. These reads are defined as *shared reads* and we account for all their multiple hits. However, the exact matching position in a reference genome S_i is not of importance and several position hits of one read on the same reference S_i are counted as one. For the mapping itself, a pseudo-alignment approach provided as part of the kallisto implementation [75] is applied. As no exact alignments are required for our purpose, a pseudo-aligner is sufficient and proves to be much faster and accurate using a fast kmer-based approach. Here, we gain significant improvements over our previously published tool GASiC, which relied on individual reference alignments by Bowtie 2 [192].

Altogether, we extract and count the number of read hits each reference genome receives and refer to it as *mapping abundance* c_i of taxon S_i. In case the data set $\mathcal{D}$ consists of mainly dissimilar references and is dominated by clearly unique mappings, the observed mapping abundances c_i may already closely reflect the underlying true abundances of the taxa. However, if many similar references are present, which is a common scenario on strain level, a large bias is present due to multiple hits of shared reads. The sum of the mapping abundances of all taxa then drastically exceeds the number of input reads.

4.2.2. Abundance estimation

Following the idea introduced in GASiC, we rely on a simulation-based representation of reference genome similarities to resolve the effect of shared reads. A similarity matrix is constructed, which encodes the proportion of reads which are expected to be shared among all pairwise combinations of reference sequences considered. Reads are simulated using Mason [193] based on each reference sequence, and are subsequently mapped to all references following the same competitive mapping setup as applied to the reads of D in the step before. The key element is to imitate sequencing, read, and mapping characteristics as good as possible to reproduce the source of ambiguities. Parameters such as read length and mismatch probability are crucial for the simulation of reads, and are inferred from the raw reads of D. The square matrix A= (a_{ij}), i,j= 1,. . . ,M ,is computed column-wise for each reference, with a_{ij} referring to the count of reads simulated from reference j which map to reference i. Next, the matrix is normalized column-wise by the read count a_{jj}, the number of simulated reads which are assigned back to their reference of origin. Thus, the matrix A = (a_{ij} / a_{jj}), i,j= 1,. . . ,M , holds values between zero and one.

Replacing the classic linear model of GASiC, we formulate a new generalized linear model (GLM) with the vector of absolute mapping abundances c and similarity matrix A to correct for the shared read biases. Aiming to recover the true, but unknown, abundances r of the taxa:

$$\mathbf{c} = \mathbf{A} \cdot \mathbf{r} + \varepsilon$$

with $\mathbf{A} = (a_{ij})$, i,j= 1, . . . ,M, $\mathbf{c} = (c_1, c_2, ..., c_M)^T$, $\mathbf{r} = (r_1, r_2, ..., r_M)^T$ with non-negativity constraint $\mathbf{r} \geq 0$, and error term ε. The observed mapping count c_i of taxon i corresponds to a summed mixture of the underlying true abundance r_i of taxon i and a proportion of shared read counts r_j due to the other references:

$$c_i = r_i + \sum_{i \neq j}^{M} a_{ij} \cdot r_j + \varepsilon_i$$

with taxon i and taxa $j = \{1...M\} \neq i$.

The GLM is defined by an identity link function as a linear relation of components holds to explain the observed mapping counts. However, in this setting of discrete counts the error ε is defined to follow a poisson distribution. We expect and observed no overdispersion in the abundance estimates within a sample after ambiguity correction by the model. This is in contrast to measurements of replicate samples, which may display overdispersion and motivate a negative-binomial assumption [95]. The GLM is internally solved by an *iteratively reweighted least* *squares (IRLS)* to find the maximum likelihood estimates referring to the *true abundance estimates* r_i for each taxon i. Along with the abundance estimates, standard errors are computed which report the range of accuracy and reliability of the abundance estimates. Further, p-values are given for each taxa estimate as a measure of significance.

In case of high uncertainty about the presence of a crucial amount of taxa within the selected set of references, the application of an implemented filtering is possible. Thereby, p-values above a set threshold, commonly a value of 0.05, and estimates below a minimum number of assigned reads are used as indicators for false-positive estimates. The filtering step helps to numerically stabilize the equation system in case of many absent taxa and a re-optimization step is subsequently conducted.

4.2.3. Differential abundance

In this section, the focus is on comparing metagenomics samples. The objective is to identify which taxa significantly change their abundance from one metagenome sample to another as well as which hold a constant abundance. For the differential abundance assessment of similar strains the integration of the variance of their abundance estimates is crucial. Hence, in place of directly comparing abundance point estimates of taxa between samples, we make use of the estimates as well as their standard errors.

First, the comparison of different samples requires accounting for potentially different numbers of total input reads N. The number of input reads has a significant impact on the computed abundances r and standard error estimates. A linear dependence is clearly noticeable (see Appendix Fig.C1) and is in agreement with theoretical derivations of the GLM framework. The abundance count estimate r scales linear with the number of reads whereas the standard error scales quadratic. This means the accuracy of abundance estimates improves with increased number of input reads as expected. Altogether, a normalization factor is required and a factor of N_x/N_y is correspondingly applied to samples D_x and D_y to achieve a comparable base between samples.

In the next step, we integrate abundance estimates and corresponding standard errors to infer an abundance distribution for each taxon in each sample. Here, it is assumed that the unknown true abundance count of a taxon underlies a poisson distribution. The potential bias due to falsely assigned reads to taxa, after correction for read ambiguities by the GLM model, is not expected to exceed the variance of a poisson distribution. But, an analytical approach is not feasible here, as the exact distribution is described in practice by a mixture of poisson distributions. However, an empirical approach can be pursued, which is realized by a two-step sampling process: In the first step, we define intervals with abundance estimates $r_i + /-$*their standard errors* as boundaries for each listed taxon. We use a scale unit of one standard error, as this reflects the uncertainty interval which is expected to contain the abundance estimate. Subsequently, potential abundance point estimates are uniformly sampled from this interval. Concurrently each of these sampled values refers to a λ value of a poisson distribution. In the second step, for each taxon and each potential λ of it, 500 values in a default setup are drawn from the corresponding defined poisson distribution with parameter λ. This creates one empirical distribution based on a specific λ for the taxon. Pooling all empirical distributions, created by all the different λ which are assigned to the taxon, results in an overall empirical distribution comprising 50,000 poisson draws by default setup. We refer to it as *empirical abundance distribution* of a taxon.

In order to assess whether taxa show differential abundance between two samples, their abundance distributions need to be compared. As we rely on empirical distributions here, no analytical form of standard differential testing is applicable. Yet, we can transfer the assessment of differential abundance to the question to which extent the corresponding abundance distributions overlap. Clearly separated distributions refer to a significant abundance change, while an increasing overlap points to smaller or no significant difference. Measuring the separation of the distributions is implemented by randomly drawing pairs of values from either distribution. The difference within each pair is computed and yields an *overall distribution of differences* as a result. Thereby, the location of the zero value related to the distribution of differences is meaningful. A zero value moving towards the center of the distribution reflects a higher previous overlap and corresponds to a less significant abundance change. An empirical p-value is correspondingly inferred by determining the quantile of the zero value within the distribution.

In case a taxon is only detected within one sample, while absent in the other, the single abundance distribution of the taxon is tested against a user-defined threshold corresponding to a minimum read count. The latter test yields the significance of taxa presence in this one sample.

Generally, p-values are calculated individually for all taxa considered in the samples of comparison, either to assess differential abundance of taxa present in both samples or to infer new appearance of taxa in only one sample. Thus, p-values need to be adjusted for multiplicity, which is performed by the method of Benjamini-Hochberg [106]. A final report is provided listing all taxa tested for differential abundance along with normalized abundance estimates for each sample, log2 fold changes, and adjusted p-values.

4.2.4. Implementation

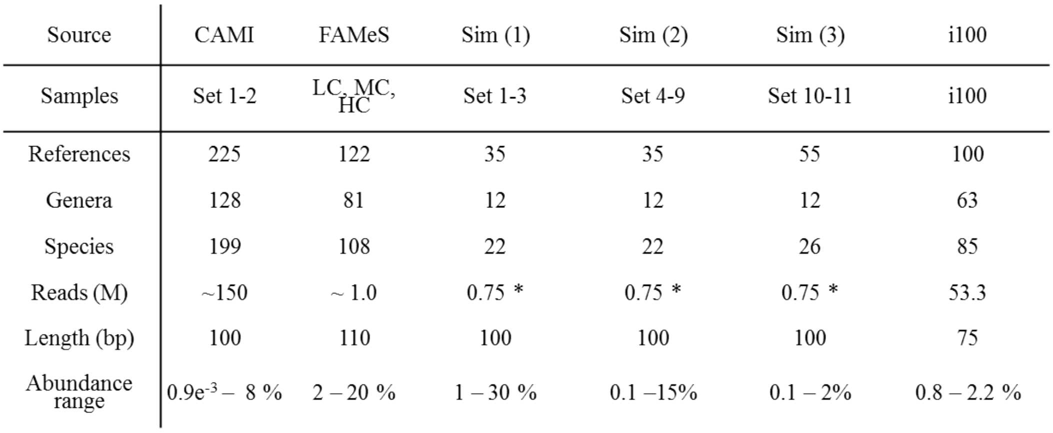
DiTASiC is implemented in Python3 and R (version $\geq$ 3.3.1), and is available from `https://rki_bioinformatics.gitlab.io/ditasic`. Further, a linked webpage and user manual provides easy guidance through the three main commands. DiTASiC is based on a flexible design and allows the integration of mapping algorithms and read simulators of choice. Our implementation uses the current state of the art pseudo-alignment algorithm provided within the kallisto framework [75], which can be individually called by the command *kallisto pseudo*. As a prerequisite, an overall index is built on selected reference sequences. Using the generated tsv and ec file formats, we extract the mapping counts of the contigs and merge them according to genomes. This allows circumventing the use of large SAM files. Further, read simulators need to be optimally adapted to capture the read characteristics. Here, the Mason simulator [193] serves as default.

4.3. Experimental setup

We tested DiTASiC and existing approaches on a variety of data sets from four different sources (Table 4.1), challenging the tools by number of taxa, total number of input reads, read characteristics, abundance complexity, and degree of reference similarities. A comprehensive simulation setup is established to enable abun-dance estimation as well as differential evaluation on an exact ground truth at which taxa proportions are known. In total, we consider eleven different simulation sets characterized by many strain clusters (Appendix Fig. C2-3), and distinguish between three groups: Group (1) serves to evaluate the abundance performance with different proportions of absent taxa, group (2) defined by all 35 taxa ensures an unbiased differential abundance evaluation in pairwise comparisons, and group (3) focuses on the resolution of large and highly similar strain clusters as well as on the impact of missing strains. Further, we relied on the Illumina based FAMeS data set of Pignatelli and Moya (2011) [194], evolved from the original set by Mavromatis et al. (2007) [195], which covers low (LC), medium (MC) and high complexity (HC) metagenomics profiles (Appendix Fig. C4). Additionally, we tested the popular Illumina 100 data sample [196], which serves as benchmark set in the latest relevant studies [180, 197]. Last, we used two benchmark data sets of medium complexity from a current comparative metagenomics challenge, CAMI [118]. We further extended the CAMI sets by simulated spike-in data, adding 30 new strains of genera already present in the original set and 20 million reads per sample, to create an additional ground truth for differential assessment.

Further details on the data sets and parameter settings are found in the Appendix C. In all presented data sets, ground truth of relative abundances of taxa

Table 4.1.: Characteristics of the four data sources: CAMI, FAMeS, Illumina 100 data (i100), and the simulation setups (Sim (1), (2), (3)). Each reference set is defined by the union of references of the underlying samples. All read profiles follow Illumina characteristics (* reads are simulated by Mason).



Source	CAMI	FAMeS	Sim (1)	Sim (2)	Sim (3)	i100
Samples	Set 1-2	LC, MC, HC	Set 1-3	Set 4-9	Set 10-11	i100
References	225	122	35	35	55	100
Genera	128	81	12	12	12	63
Species	199	108	22	22	26	85
Reads (M)	~150	~ 1.0	0.75 *	0.75 *	0.75 *	53.3
Length (bp)	100	110	100	100	100	75
Abundance range	$0.9e^{-3}$ – 8 %	2 – 20 %	1 – 30 %	0.1 –15%	0.1 – 2%	0.8 – 2.2 %

is available. Comparing the samples, a ground truth to classify differentially or non-differentially abundant taxa is given for the simulation and CAMI study, while fold-change accuracy can be assessed in all data sources.

4.4. Results

In the following sections, we demonstrate the performance of DiTASiC on the presented data sets in comparison to existing tools. We separately investigate three aspects: (i) abundance estimation, (ii) absent and missing taxa, and (iii) differential abundance. Evaluations focus on the accuracy of estimates of relative taxa abundance as well as fold change, and on sensitivity and specificity concerning detection of differentially abundant taxa.

4.4.1. Abundance estimation

In this first part, we address the quantification of taxa in a given metagenomics sample, aiming for the highest taxonomic level. We highlight the strength of our proposed GLM model for the resolution of shared read counts and subsequent inference of corrected abundance estimates for taxa considered.

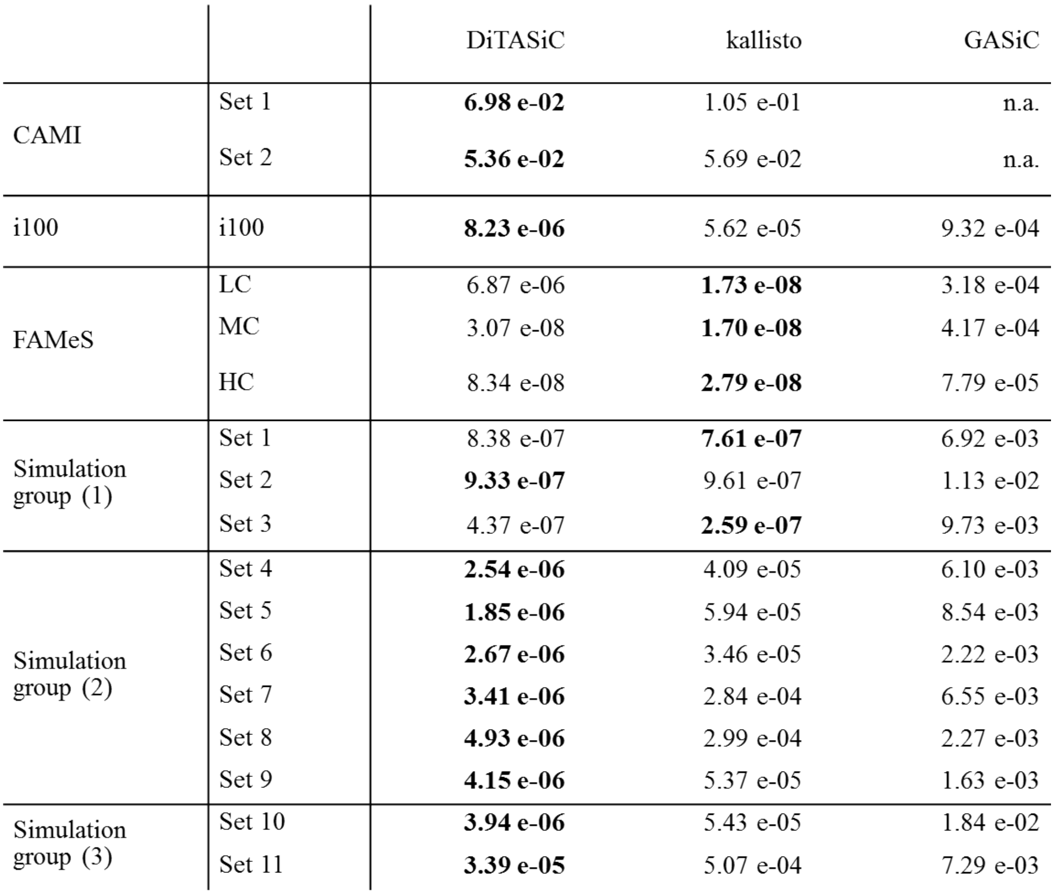
We compare to our previously published tool GASiC [73], which relies on individual reference alignments and a non-negative LASSO modelling approach for abundance estimation, and present significant improvements. Further we test against the most recently published tool for RNA-Seq analysis, kallisto [75,180], which has also been shown to perform superior to other existing tools in the application to metagenomics. We also evaluate on the same benchmark data to allow further comparison of tools (see Appendix C). While we compare against the full version of kallisto, it is important to note, that we use and integrate the pseudo-aligner of kallisto for mapping purpose, but not kallisto's quantification and modelling framework. Yet our main focus in this work is the modelling and resolution of arising read ambiguities due to highly similar genome sequences considered. Hence, the comparison of DiTASiC to kallisto in this section refers to a comparison of our GLM model to the statistical EM framework of kallisto.

All tools are applied to each sample individually, in total we consider and evaluate 17 different samples from four data sources. The output of all three tools are absolute read counts assigned to each taxa in the data set considered. Normalization is applied by dividing all absolute taxa counts by the total number of input reads of the corresponding sample. We receive an estimation of a quantitative taxa composition of a sample as a result.

All data sets described here provide a ground truth of taxa abundance proportions, enabling us to assess the difference between truth and estimate. As an error measure we apply the *SSE* (Sum of Squared Errors) to evaluate the accuracy of the given estimates, the SSE also penalizes abundance estimates obtained for absent taxa.

The resulting error measures of abundance estimation by DiTASiC, GASiC, and kallisto, according to all different data sets are reported in Table 4.2. Overall, Di-

Table 4.2.: Accuracy of taxa abundance estimates by DiTASiC, kallisto and GASiC. Accuracy is defined by the SSE (sum of squared error) between estimates and available ground truth. A significant error reduction is shown for DiTASiC compared to GASiC and a comparable performance is observed for kallisto (highest accuracy is depicted in bold print). GASiC was not run on CAMI data due to computational limitations.



		DiTASiC	kallisto	GASiC
CAMI	Set 1	**6.98 e-02**	1.05 e-01	n.a.
	Set 2	**5.36 e-02**	5.69 e-02	n.a.
i100	i100	**8.23 e-06**	5.62 e-05	9.32 e-04
FAMeS	LC	6.87 e-06	**1.73 e-08**	3.18 e-04
	MC	3.07 e-08	**1.70 e-08**	4.17 e-04
	HC	8.34 e-08	**2.79 e-08**	7.79 e-05
Simulation group (1)	Set 1	8.38 e-07	**7.61 e-07**	6.92 e-03
	Set 2	**9.33 e-07**	9.61 e-07	1.13 e-02
	Set 3	4.37 e-07	**2.59 e-07**	9.73 e-03
Simulation group (2)	Set 4	**2.54 e-06**	4.09 e-05	6.10 e-03
	Set 5	**1.85 e-06**	5.94 e-05	8.54 e-03
	Set 6	**2.67 e-06**	3.46 e-05	2.22 e-03
	Set 7	**3.41 e-06**	2.84 e-04	6.55 e-03
	Set 8	**4.93 e-06**	2.99 e-04	2.27 e-03
	Set 9	**4.15 e-06**	5.37 e-05	1.63 e-03
Simulation group (3)	Set 10	**3.94 e-06**	5.43 e-05	1.84 e-02
	Set 11	**3.39 e-05**	5.07 e-04	7.29 e-03

TASiC strongly reduces the error on all data sets compared to GASiC by several orders of magnitude. Further, DiTASiC shows either comparable and in many cases improved performance to kallisto. Generally, reported error values are dependent on data size and prevailing genome similarities. However, the presented values refer to a remarkably high accuracy of abundance estimates overall. Smallest divergences of estimates from the ground truth are found for the FAMeS data sets (Appendix Fig. C5). This is expected due to less pronounced reference similarities within the data and moderate median abundance proportions, meaning less challenge for the resolution models. The CAMI data do pose a much greater challenge, considering 255 taxa for quantification with several strain clusters and some extremely small relative abundance values. Yet, highly accurate taxa estimates, apart from few small outliers, are obtained by DiTASiC; notably also for very low relative abundances below 0.01% (see also Appendix Fig. C6). CAMI data was not analyzed with GASiC due to computational limitations. The commonly used i100 data set is characterized by shorter reads derived from different bacterial strain clusters. DiTASiC achieves an improved accuracy in comparison to kallisto, and also to further tools when compared to the values reported in a recent benchmark study of different abundance profiling tools on the i100 set [180] (see Appendix C and Fig. C7). The simulation data serves as a challenge with a high number of similar strains and a smaller number of reads available for assignment. In comparison, samples in CAMI hold 150 times more reads with only seven times more taxa. The results show that DiTASiC performs superior in all sets of simulation group (2), where all taxa are present, while errors are proportionally higher in sets of group (1), where taxa are absent. Group (1) is primarily defined by the absence of distant strains or entire strain clusters; the EM algorithm of kallisto proves to be slightly more accurate in these scenarios. However, sets in simulation group (3) are characterized by the absence of strains from highly similar clusters and by the presence of very large clusters of high sequence similarities. Here, DiTASiC demonstrates to be more powerful (Appendix Fig. C8). Notably, we observe an increased error of abundance estimates in kallisto predominantly for highly similar strain sequences. In contrast, DiTASiC reveals its particular strength in the resolution of these strain clusters, it demonstrates to precisely distinguish abundances down to sub-strains with sequence similarities above 95%. Different examples are found for the CAMI, i100 and simulation data, considering diverse *Escherichia coli* cluster, *Corynebacterium* and *Staphylococcus aureus* cluster (Appendix Fig. C9). Here, an accurate cluster resolution is obtained by DiTASiC, and common errors such as abundance interchange or equalization of similar sub-strains are avoided.

Appendix Figure C10 visualizes the taxa abundance estimates of the different tools in comparison to the observed mapping abundances, exemplary for three simulation sets of different complexity. It clearly demonstrates how the mapping abundance, biased due to read ambiguities, mainly overestimates the ground truth and further assigns abundance counts to absent taxa. GASiC shows some significant over- and underestimations, while the accuracy of DiTASiC and kallisto is consistently high. Further, a study of two replicate sets, defined by read sets simulated with the same abundance profile, proves robustness and precise reproducibility of results by DiTASiC as well as kallisto, with significant improvement over GASiC (Appendix Fig. C11).

4.4.2. Absent and missing taxa

We recommend prior pre-filtering of references to focus on reference genomes of species expected in the data. Still, frequently we consider more references than taxa actually present in the data and an inclusion of all potentially abundant strains is advised.

Hence, in the simulation group (1) and (3) and the FAMeS data, which hold different proportions of absent taxa, we tested the detection performance of DiTASiC. The internal filtering is conducted to infer potential false-positive taxa in the given sets. In the simulation group (1) the abundant taxa proportions of 28%, 40% and 45%, respectively, are exactly detected with neither false-positive nor false-negative calls. In the FAMeS data, proportion of absent taxa based on the reference set corresponds to 8%, 9%, and 8% in the three samples. DiTASiC achieves sensitivity and specificity of 100% for the MC and HC data. In the LC set, a false-negative is caused by missing one abundant taxon, resulting in a decreased sensitivity of 99.1%. (Appendix Table C1). In simulation group (3), set 10 serves to study the impact of absent strains from highly similar clusters and indicates unbiased abundance estimation of strains of the affected clusters by DiTASiC (refer to Appendix Fig. C7). In another study, reads derived from 55 taxa are contrasted to a reduced reference set of 35 taxa to investigate the impact of missing taxa in a selected reference set. First, we observe that 11% of the reads are not aligned; second, it is shown that abundance estimates of some taxa are overestimated by DiTASiC. However, a closer look reveals that it concerns closely related strains which show an increased abundance due to miss-ing strains within their cluster. The results propose that no overall abundance bias is caused (Appendix Fig. C12).

4.4.3. Differential taxa abundance

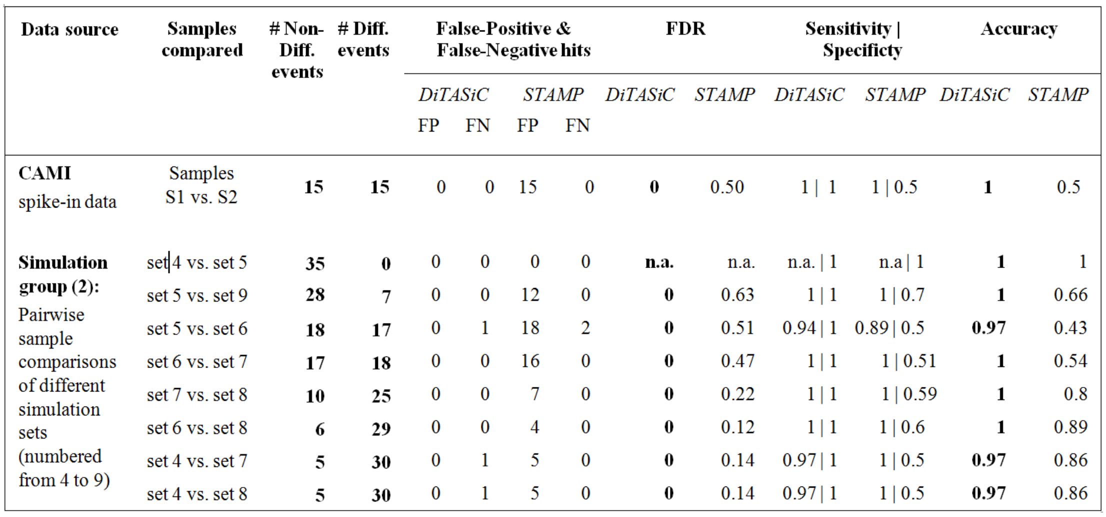
Here, we evaluate pairwise comparisons of metagenomics samples, aiming to reveal the change of taxa compositions at the highest taxonomic level. We demonstrate how the entire process of read ambiguity resolution and incorporating the uncertainty of abundance estimates has a crucial impact on differential assessment on strain level. As a result, a more accurate detection of differential events is achieved, particularly in case of small changes. False-positives are significantly reduced.

In order to evaluate independent of technical and biological variance factors, we do not consider replicate samples and comparisons here. This way we can test our specifically addressed differential method and prove the validity and impact of the abundance variance without bias. We compare our approach to STAMP [182,183], which is available for pairwise comparisons to exemplary demonstrate the importance of the issues of read ambiguities and abundance estimation uncertainties. The mapping abundances of the taxa serve as input for STAMP. STAMP is a software package providing several statistical tests for differential taxonomic and functional assessment and a user-friendly graphical interface. The recommended option of a G-test with Yates continuity correction followed by a Benjamini-Hochberg adjustment is selected.

Different metagenome comparisons are conducted within the presented data sources. Evaluations focus on correct detections of differentially abundant taxa and on accuracy of taxa fold changes. For the simulation data and the CAMI spike-in data, ground truth is available for specific classification into differential and non-differential taxa, results are described by measures of sensitivity, specificity, and accuracy as combined measure of correct detections. For the FAMeS and the original CAMI data, no classification is provided, here, the accuracy of fold changes is evaluated by using the *SSE* instead.

Different pairwise comparisons of the simulation data cover various scenarios of non-differential and differential events. A p-value cutoff of 0.05, adjusted for multiplicity, is used to define differentially abundant taxa. Evaluation results for the

Table 4.3.: Evaluation of differential taxa abundance by DiTASiC and STAMP based on sample comparisons within the simulation data and the CAMI data set. A p-value cutoff of 0.05 is used to define differentially abundant taxa. In most scenarios, DiTASiC achieves exact detections, holding a false-discovery rate (FDR) of zero and accuracy above 97% overall. A reduced accuracy performance by STAMP, using mapping abundances, confirms the significant impact of read ambiguities and abundance estimate uncertainties. *in case no differential events, FDR and sensitivity cannot be computed (n.a.)



Data source	**Samples compared**	**# Non-Diff. events**	**# Diff. events**	**False-Positive & False-Negative hits**				**FDR**		**Sensitivity | Specificty**		**Accuracy**	
				DiTASiC		*STAMP*		*DiTASiC*	*STAMP*	*DiTASiC*	*STAMP*	*DiTASiC*	*STAMP*
				FP	FN	FP	FN						
CAMI spike-in data	Samples S1 vs. S2	**15**	**15**	0	0	15	0	**0**	0.50	1 | 1	1 | 0.5	**1**	0.5
Simulation group (2): Pairwise sample comparisons of different simulation sets (numbered from 4 to 9)	set 4 vs. set 5	**35**	**0**	0	0	0	0	**n.a.**	n.a.	n.a. | 1	n.a | 1	**1**	1
	set 5 vs. set 9	**28**	**7**	0	0	12	0	**0**	0.63	1 | 1	1 | 0.7	**1**	0.66
	set 5 vs. set 6	**18**	**17**	0	1	18	2	**0**	0.51	0.94 | 1	0.89 | 0.5	**0.97**	0.43
	set 6 vs. set 7	**17**	**18**	0	0	16	0	**0**	0.47	1 | 1	1 | 0.51	**1**	0.54
	set 7 vs. set 8	**10**	**25**	0	0	7	0	**0**	0.22	1 | 1	1 | 0.59	**1**	0.8
	set 6 vs. set 8	**6**	**29**	0	0	4	0	**0**	0.12	1 | 1	1 | 0.6	**1**	0.89
	set 4 vs. set 7	**5**	**30**	0	1	5	0	**0**	0.14	0.97 | 1	1 | 0.5	**0.97**	0.86
	set 4 vs. set 8	**5**	**30**	0	1	5	0	**0**	0.14	0.97 | 1	1 | 0.5	**0.97**	0.86

simulation data are presented in Table 4.3.

For all scenarios, DiTASiC reports no false-positive hits, holding a false-discovery rate (FDR) of zero and a resulting specificity of one. Further, in five out of eight comparisons also a sensitivity of 100% is achieved. In the other three cases, the detection of one known differentially abundant taxon fails resulting in one false-negative detection and corresponding sensitivities of 97%. Here, it concerns the differential detection of the sub-strain *E.coli K12 MG1655*, which holds accurate abundance estimates but fairly large standard errors, arising due to uncertainties because of high sequence similarity of 98% with another *E.coli* sub-strain *DH10B*. The known relative abundance decrease by 1% is very small and hereby falls in the abundance variance range, while an increase by 3% for sub-strain *DH10B* could be detected as well as differences below 1% for the other *E.coli* strains in the cluster. In general contrast are the results obtained for STAMP, showing a strong tendency of identifying non-differential taxa as differentially expressed, causing high numbers of false-positives. As abundance estimates underlie some variation, additionally biased due to read ambiguities, these results confirm how the inclusion of standard errors is crucial to identify taxa with consistent abundances. The FDR of STAMP ranges from 12 % to 63% and the overall accuracy from 46% to 86%.

A similar situation is observed for the CAMI spike-in data. DiTASiC correctly detects all 15 differential and 15 non-differential taxa. However, all 30 taxa are found to be differentially abundant by STAMP, resulting in an accuracy of only 50%. Considering the entire CAMI data set, fold changes, spanning from 0.0009 to 1024, are proven to be highly accurate for DiTASiC with an *SSE* 19 times smaller com-

pared to the STAMP output. Further, the assigned p-values by DiTASiC clearly separate the spiked-in non-differential and differential taxa (Appendix Fig. C13). All other taxa of the data set, holding fold change values greater than zero, also receive very small p-values stating differential abundance, but cannot be further confirmed.
Pairwise metagenome comparisons within the FAMeS data also exhibit high fold change accuracies, as consequence to the former highly accurate abundance estimates. Corresponding *SSE* values are two magnitudes smaller compared to the ones computed by STAMP (see Appendix Table C2).

4.5. Discussion of results

This work demonstrates the challenges concerning strain level resolution in metagenomics data and the need for dedicated methods for quantification and differential abundance testing. DiTASiC addresses these challenges and provides novel approaches.
The inference of taxa abundances by directly counting mapped reads is not suitable on strain level. Although read mappers have significantly improved in speed and mapping accuracy, they cannot resolve shared read assignments and thereby cannot directly output correct abundances. Our results show the bias introduced by the pseudoaligner of kallisto (without its well working EM-based quantification framework): the abundances of most taxa are overestimated and many actually absent taxa are assigned positive abundances. This effect is due to shared read counts, caused by highly similar reference sequences of strains in a metagenomics sample. DiTASiC is based on a new GLM framework, adapted to characteristics of taxa data for the resolution of shared read counts. As a result, it provides highly accurate abundance estimates for taxa in different metagenomics samples. Thereby, DiTASiC proves excellent performance independent of abundance profile complexities and also shows reduced errors in comparison to existing tools on a recent benchmark study on the i100 data [180]. It enables accuracy in a large range of relative abundances from 0.001% to 30% present in the various data sets. Further, while generally the read coverage in a metagenomics sample is a critical factor for abundance estimation, the degree of reference similarities of present taxa means a greater challenge. Thus, on the FAMeS data set with 122 taxa, but many dissimilar species, all tools achieve overall higher abundance accuracy compared to the simulation sets with only 35 taxa holding almost the same number of input reads, but different challenging strain clusters. However, the GLM model of DiTASiC proves specific strength in highly accurate abundance resolution within strain clusters, as is shown for various examples in the i100, CAMI and simulation studies. In particular, it demonstrates to precisely distinguish abundances down to sub-strains which share sequence similarities above 95%. Whereas this is more challenging for kallisto, which was similarly reported in a benchmark study by McLoughlin, 2016. An important point is that the similarity matrix used in DiTASiC is not necessarily symmetric. Hence, the simulated proportion of reads shared from reference i with reference j can differ from the proportion reference j shares with reference

i. We observe these dissimilarities in the matrix e.g. for the *E.coli* clusters and hypothesize that this may explain the good performance of DiTASiC, as it allows capturing sub-strain sequences, which may be shorter, but highly similar to other longer strain sequences.

The framework of DiTASiC is also robust with increasing sequencing error, as the internal matrix simulations account for the error profiles found in the raw reads. However, as a consequence, misaligned reads in addition to shared reads will cause abundance bias, which poses another resolution challenge. Further, missing or unknown taxa in reference sets may introduce quantification bias. However, one of our studies indicates that closely related strains compensate for missing ones and not affected strain cluster remain stable. Overall, DiTASiC shows certain robustness on imperfect reference sets with either missing or false-positive taxa included. Nevertheless, explicitly accounting for non-mapped reads and their missed abundance proportion could be included in future work.

All in all, the accuracy of the abundance estimation has an immediate impact on the accuracy that can be achieved in the differential abundance analysis of the taxa. This is clearly observable in the comparisons of the FAMeS data sets, which result in highly accurate fold change estimates in consequence of the accurate abundance estimates that were obtained.

However, for differential abundance testing, in order to distinguish differentially and non-differentially abundant taxa, the uncertainty of the abundance estimates plays a crucial role. Especially on strain level, this variance reflects uncertainties in the underlying read ambiguity resolution in the presence of highly similar reference sequences. DiTASiC introduces a new statistical framework, which integrates the abundance variance and forms abundance distributions for differential testing sensitive to strain level.

Generally in comparative metagenomics, it is difficult to predict how a community of taxa in a sample will change, as there is a variety of influential factors involved. A study by [187], demonstrates how human actions can cause next-day abundance change in the microbiome. Hence, putting assumptions on data for composition change is complex. Further, although taxa abundance data and gene expression data share discrete count data characteristics, assumptions commonly made for gene expression for differential analysis cannot be easily transferred. One of the most common assumptions is that the majority of features will not be differentially changed. This is reasonable for genes in a cell as no global change of expression of all genes is biologically expected. In metagenomics studies though, antibiotics treatment has shown to cause rapid change of microbial compositions in human samples [198]. Further, gene expression data in RNA-Seq studies are often characterized by overdispersion and correspondingly modelled by negative binomial distributions.

Different popular RNA-Seq tools as well as standard statistical test are frequently applied to metagenomics gene data for differential analysis, however, have been shown to not capture the data well in all cases [98]. Similar problems are observed when considering differential taxa abundance. In a study of plaque samples, DE-Seq and edgeR were also shown to not fit the data properly [99]. Hence overall, it

is important to distinguish gene and taxa level and critically assess corresponding assumptions. Furthermore, defining assumptions to capture all diverse structures of metagenomics data might pose an almost impossible challenge. Here, we propose an independent statistical framework for differential testing of all individual taxa in the set, without putting any assumptions on overall composition change.
We evaluated our approach on diverse scenarios, covering sets with only non-differential events to sets with overall change, and can indicate overall correct detections. Further, the method is not dependent on the presence of a taxon in both samples of comparison, it also serves as test on taxa emergence or extinction.
In contrast, STAMP yields many false-positives, which reflects the importance of read ambiguity resolution and integration of abundance uncertainties for strain level analysis. In cases of extremely similar strain sequences, however, large standard errors for the estimates can occur, as shown for the two *E.coli* sub-strains, and can consequently cause a lower limit for the detection of very small fold-changes in DiTASiC.
Generally, DiTASiC is neither limited to bacteria nor any taxonomic level. Also its concept is applicable to any ambiguity resolution in which the similarities causing the ambiguities can be described. Further, variance of sample replicates pose another crucial variance source, integration could be achieved by not sampling from the mixture of poisson distributions of one experiment, but across all replicates. DiTASiC is independent of specific databases or any additional data information, it simply relies on the raw reads and on a (pre-filtered) species reference set in fasta format, the latter can also contain assemblies or fragmented sequences.

To summarize, this contribution focuses on the resolution on strain level in metagenomics data concerning taxa quantification and differential abundance assessment. We point out the challenges arising on strain level due to the presence of highly similar reference sequences. We present DiTASiC, which provides a new GLM framework for the resolution of shared read counts and introduce a statistical framework, which integrates abundance variances, for differential testing sensitive to strain level. As a result, highly accurate abundance estimates down to sub-strain level as well as detections of differentially abundant taxa are obtained. Evaluations are conducted on different data sources and in comparison to existing methods.

5. Summary and outlook

Enormous advances in high-throughput technologies have promoted quantification studies in genomics and proteomics research in the last decade. Quantification of genes, proteins or taxa in mixed samples is fundamental to gain deeper understanding of biological processes. The massively parallel nature and enhanced resolution capabilities of MS and NGS instruments have opened a large range of applications. Quantification workflows are highly diverse, following different techniques and protocols that comprise many steps from sample preparation to data acquisition [17,28]. Multiple error sources are hidden within the process which cause biases in quantitative measures. Processing of raw measurements and analyses tailored to the biological question of interest are one of the most important steps and can ruin a whole experiment if not conducted accurately. Accuracy in quantification is vital for any downstream analysis and has an immediate impact on differential abundance studies.

With the fast advances of modern NGS and MS devices, the development of computational methods for quantitative data processing is still lagging behind. The inference of accurate quantitative estimates from a sample remains a complex challenge. A strong demand for new approaches to address and correct quantitative biases constantly persists. This thesis presents new statistical strategies to improve the quantification accuracy of data types from high-throughput applications. The work points out parallels between different omics fields and stresses the benefit of transferring established statistical concepts, while equivalently emphasizing the importance of integrating specific error and data characteristics to improve quantification results.

In Chapter 2, a comprehensive and novel statistical framework for the analysis of AP-MS data based on protein count measures to identify protein-protein interaction candidates is presented. Chapter 3 is dedicated to the fundamental preprocessing step of inferring accurate protein quantities from heterogeneous peptide spectra measurements, an often underestimated task referred to as peptide-to-protein summarization. Chapter 4 approaches higher taxonomic resolution in NGS-based metagenomics data. It addresses accurate abundance estimation and differential testing sensitive to strain and sub-strain level.

Overall, the three projects provide novel approaches to solve current challenges in the high-throughput quantitative analysis field. The work demonstrates the applicability and transfer of well-known pre-processing methods such as normalization and statistical filtering, originating from genomics data analysis, to the quantitative proteomics field and confirms its potential to enhance results in proteomics studies alike. Equivalently, statistical models designed for differential analyses in microarray studies, and further adjusted to NGS data, are also shown to be adapt-able to MS data. The integration of statistical concepts and a thorough evaluation of quantitative estimates by statistical measures is strengthened in all three analysis methods. Moreover, the work stresses, that in addition to pure quantitative information, other feature information arising with the experiment can be highly valuable to appraise measurement reliability. It emphasizes that the quality of quantitative measures is frequently reflected by attached features and that this information should be fully considered for quantification assessment. Generally, errors in the quantification process cause variance in quantitative measures and, despite the application of correction methods, uncertainties in final abundance estimates always remain. The last project focuses on the variance arising within the abundance estimation process and highlights the importance to integrate this variance in quantification models. The abundance variance is particularly relevant for differential quantitative comparisons. Yet, in most methods, only sample variances of technical or biological replicates are considered.

With the first project referred to as APMS-WAPP, a comprehensive statistical framework of pre- and post-processing is introduced for detecting protein-protein interactions in AP-MS data. Based on various simulation data sets and an experimental study of the pathogenicity island 1 of *Salmonella Typhimurium*, the significant impact of pre-processing methods on enhancing interaction candidate lists is demonstrated. Application of normalization methods adjusted to AP-MS characteristics yield an increase of 20-40% in the number of true interaction candidates in simulated benchmark sets. Further, an additional filtering step reduces the multiple testing problem and increases detection sensitivity. As a novel and alternative scoring scheme to evaluate candidates, the TSPM model originating from RNA-Seq analyses is transferred and proves its efficiency in separating true interactors from contaminants. For post-processing, a permutation approach combined with FWER or FDR controlling procedures enables a specified cutoff choice according to the intention of the experiment. As a result, three workflow options are provided in our R-package to generate reliable lists of interaction candidates for wet-lab scientists.

The second project addresses the problem of heterogeneous peptide measurements and the need for sophisticated strategies to infer accurate protein abundance estimates. iPQF is presented as novel peptide-to protein summarization method which uses spectrum feature information in conjunction with quantitative values for improving protein abundance estimation. Significant correlations between spectra features and quantification accuracy of a spectrum are revealed in this work. Diverse features are investigated and their combined strength is proven to be powerful to assess spectra reliability. The introduced summarization method is the first to consider feature information and to control the contribution of peptides to protein quantification according to their spectra feature reliability. This approach is shown to be particularly strong in scenarios of small to medium numbers of spectra per protein with large measurement variation, which refers to a common case in many proteomics data sets. A comprehensive evaluation of iPQF to nine protein abundance inference methods on five different data sets demonstrates a robust and superior performance. It proves the benefit of feature integration for improved protein quantification.

In NGS based studies, quantification is determined by the summarization and resolution of reads assigned to genome reference sequences. Specific challenges arise when aiming for higher taxonomic quantitative resolution in metagenomics data due to the presence of highly similar reference sequences. DiTASiC addresses these challenges and provides novel approaches for taxa quantification and differential abundance testing beyond species level. A new GLM framework is introduced to resolve the significant shared read count biases and a new statistical framework enables sophisticated differential testing by integrating estimation uncertainties and forming abundance distributions. DiTASiC provides highly accurate taxa abundance estimations in data sets of diverse complexity, with taxa proportions ranging from 0.001% to 30% in a sample. The tool proves particular strength in accurate resolution of clusters holding many similar strains, even precisely distinguishing abundances of sub-strains with sequence similarities above 95%. Its superior performance on strain and sub-strain level to state-of-the art tools is highlighted on latest benchmark sets.

5.1. Outlook

Although new approaches are presented in this thesis, which enhance the current repertoire of computational methods to achieve higher quantification accuracy in three high-throughput applications, still various challenges remain. Naturally, different shortcomings are found in the existing methods. In this section, individual improvements and extensions for the presented methods are considered and a global perspective on desired advances for high-throughput quantification analysis is given.

The introduced AP-MS analysis framework is applied to protein data and relies exclusively on spectral count information. While spectral counting refers to a robust quantitative measure, additional quantitative information such as the number of different peptides and the actual sequence coverage per protein can be valuable to refine the quantitative picture of an interaction candidate. It is worthwhile to distinguish whether the number of spectral counts of a protein is only based on one detected peptide sequence or is composed of different peptides supporting the protein detection. A higher peptide count and higher protein coverage in bait samples certainly points to a more reliable protein interaction candidate, especially if a reduced peptide number per protein is observed in negative controls. Equivalently, the average MS1 intensity of the peptides could serve as an additional indicator. Peptide number and average MS1 intensity could be used as a weighting factor applied to interaction sores for a more fine-grained candidate ranking. This weighting should take place after application of the permutation system. Further, the presence of shared peptides needs be more closely investigated in the pre-processing. Particularly candidates with no unique peptide hits are more likely to be false-positive identifications and should be flagged correspondingly.

In general, proteins do not act individually but form protein complexes to fulfill essential biological functions. The identification of protein complexes is another objective in AP-MS analysis and is not considered in this work so far. A bait protein is commonly involved in multiple complexes. Consequently, prey proteins captured in one purification may originate from different complexes and do not share a direct interaction necessarily. Hence, a protein complex cannot be inferred straightforward from one pull-down experiment and strategies are needed to evaluate an existing linkage of two proteins to the same complex. Two main types of models, the spoke and matrix model have evolved to detect co-complex interactions between two proteins based on their counts obtained from different bait pull-down experiments [199]. The former model focuses only on bait-prey connections, while the latter also integrates prey-prey interactions and builds a matrix of all bait-prey counts. Considering protein counts from several negative controls can be referred to a pure false-positive prey-prey matrix. Hence, the permutation strategy of APMS-WAPP could be adapted in a way that protein complex scores obtained from the bait-prey count matrix are contrasted to scores from a matrix based on negative controls. Moreover, in a large multiple bait study, true interactions of bait-prey pairs are expected to re-occur in reverse pull-downs and generally a consistent interaction network is assumed. Protein interactions forming a complex should receive higher scores from a consistent and original bait-prey count matrix than from a distorted matrix. Thus, instead of using negative controls, a distorted matrix could be created by randomly shuffling counts within the matrix or by exchanging small bi-clusters of counts between bait-prey groups to ensure similar data distributions. This could be a new permutation concept for protein complex evaluation and a potential extension of the APMS-WAPP applicability.

As negative controls in AP-MS studies are largely independent of the bait proteins of interest, findings from previous control samples can already help to identify contaminants in new studies. Meanwhile, a contaminant repository for affinity purification mass spectrometry data, referred to as the CRAPome, has been established by Mellacheruvu *et al.* [129]. It is recommended to link the CRAPome database to the pre-processing of APMS-WAPP for pre-filtering for known contaminants as a next implementation step. Further, assessing the abundance ratio that known contaminants exhibit between bait and control samples provides useful information about the background signal strength and further evaluates the computed FDR threshold.

Furthermore, the developers of SAINT have introduced SAINT-MS1 as a follow-up tool [125]. SAINT-MS1 still focuses on the protein level, but now enables using continuous MS1 intensities as quantitative measure instead of spectral counts. Hence, in this context, another new idea is to conduct the entire analysis on peptide level instead of protein level. This provides more values and information per protein and could be particularly beneficial if only few replicates are available. Using the MS1 intensities of the peptides can potentially enhance the separation of true protein interactions from contaminants. For example, in case of equal numbers of spectral counts in bait and control samples, the underlying peptide signal intensities can refine the picture by either showing a clear offset or confirming equal signal values between the samples. An alternative suggestion is to couple the presented summarization method iPQF with SAINT-MS1 in a pre-processing step, to achieve more sensitive protein abundance estimates as input for the APMS framework.

With the development of the iPFQ algorithm, significant correlations of peptide spectra features to quantification accuracy have been demonstrated. However, so far, the present work focuses solely on characteristics of the peptides, considering charge state, length, modification state and other features at the PSM level. Future investigations should be devoted to more technical features representing aspects of the experimental workflow, the signal generation and the data acquisition process. Another problem concerns outlying data values, which can arise due to technical factors and go beyond a biologically meaningful range. Although such extreme spectra values receive a smallest possible weight by the algorithm, they still contribute to the overall protein quantification to some extent. The actual variance of the peptide intensities is not integrated by the algorithm when transforming the feature values into discrete ranks. Hence, a spectrum receiving the smallest weight can be a substantial outlier or can as well represent a spectrum with a maximum but overall small divergence to all other spectra values. A recommendation for future developments of iPQF is to integrate a pre-quality filter for testing the variance of assigned spectrum values. Either this variance can be used to apply a non-linear ranking scheme or to simply discard spectra with an x-fold variance that clearly exceeds an expected biological value range.

Assessing the quality of raw measurements and directly integrating properties of the measures into subsequent data algorithms ensures more data specific analysis. A crucial step in the presented metagenomics profiling tool DiTASiC is to imitate the mapping process of the reads to the reference sequences as good as possible to reproduce and resolve the source of ambiguities. The choice of a read simulator that closely matches the profile of the raw read data is fundamental in this approach. While few pre-processing scripts are provided along with DiTASiC to determine read characteristics for finding an optimal parameter setting for the read simulation, an improved parameter inference directly from the raw read profile is desired. Read simulators often follow fixed internal settings according to specific sequencing platforms. But using empirical read-length distribution and error models directly retrieved from the given read sample would provide a more precise emulation. Further, to this end, DiTASiC is only implemented in single read mode. However, including paired read information would be beneficial. Corresponding implementation in DiTASiC is easy, only a new extraction script from tsv and ec files based on a paired-read mapping is required. Paired-read information helps to improve the mapping accuracy to the taxa and can consequently increase the final abundance accuracy.

In metagenomics analysis, one also frequently faces the problem that a certain number of reads cannot be mapped to any reference sequence given. Reasons for unmapped reads of a sample can be either novel taxa, for which no reference sequence is known yet, or genome sequences which have been missed to be included, or reads derived due to contamination or sequencing errors. DiTASiC proves certain robustness on imperfect data sets. In studies, the tool has shown that closely related strains compensate for missed strains by receiving increased abundance estimates, while distant strains remain unaffected. However, the number of reads, which cannot be mapped to any reference and are not captured by closely related strains, signify missed abundance proportions. A precise assessment of actual missed abundance proportions becomes especially crucial when comparing metagenomics communities with differing amounts of unmapped reads. Relative abundances of detected taxa may be affected with different abundance overestimation and may not be comparable between samples. One suggestion is to include an *unmapped reference category* in the DiTASiC model, incorporating the number of unmapped reads as an additional equation to solve within the overall model system. The added row and column in the similarity matrix refers to a unit vector, as unmapped reads naturally exclude the shared read status.

DiTASiC was developed with the purpose to enable profiling and accurate abundance resolution down to strain and sub-strain level in metagenomics samples. The scope of DiTASiC can however be extended to various other applications dealing with the distinction of highly similar genome sequences. The performance of DiTASiC was investigated in a test case concerned with the identification of candidates involved in a horizontal gene transfer event and revealed promising results. A sample containing reads of both the donor and acceptor genome was analyzed aiming to identify the correct candidates out of a set of 500 potential and highly similar species candidates. The true donor and acceptor received the most significant p-values in the abundance estimation by DiTASiC. Follow up investigations in this direction would be highly interesting to learn more about DiTASiCs applicability on other objectives.

In summary, all three projects conducted in this thesis aim to minimize quantification variation to improve the accuracy of quantification. Here, new stand-alone strategies addressing certain quantification biases are presented, with focus on statistical data pre-processing, peptide heterogeneity and integration of estimation uncertainties. Overall, high-throughput quantification workflows are highly diverse and multiple sources, from sample preparation to data acquisition and final inference, cause different errors and variation in quantitative measures. Hence ideally, all these individual biases from the beginning to the end of the process should be captured, corrected, and all arising variances should be taken into account to assess an overall quantification reliability of final quantitative estimates. Although many individual and elaborate software solutions for specific data correction steps exist, there is still no comprehensive collection which also allows easy tool concatenations. Some commercial software suites have emerged, which try to offer a full service from raw data processing to differential quantitative assessment. However, even comprehensive tools can mostly only cover a narrow range of experimental setups. Thereby underlying data assumptions, sensitive parameters and limitations of the methods are rarely emphasized in their descriptions. A general dilemma persists between broad and easy usability of tools and the necessity for data adapted strategies. Particularly users of non-computational background are confronted with a difficult challenge to find an optimal quantification workflow according to their

given data set.
A vision for quantification analysis would be to unify all available data correction tools of a technology, for instance organized by a platform hosting analysis modules for the different steps of the quantification process. Further, prior data set assessments are essential to retrieve information on value distribution, sparsity, outliers or other data features in order to choose appropriate processing methods. In addition, data assumptions underlying the tools need to be more clearly reported.

A point which is also poorly addressed in data processing concerns the best application order of correction algorithms. For instance, whether to apply imputation strategies for missing values before or after peptide aggregation or whether normalization on peptide or protein level provides more robustness. This comes with a general demand for reliable benchmark data and testing setups to enable reasonable evaluation and comparison of tools. A lack of realistic benchmark sets with known ground truth for evaluation persists in the proteomics and metagenomics field.

As different biases occur along the quantification process and have shown to affect quantities differently, a comprehensive evaluation of the individual impact of each bias on final quantitative estimates is needed. Investigating which biases potentially overlie and to which extent each bias contributes to the final variance is an important matter and not answered yet. Further, an interesting question is whether the effect of processing induced biases in a sample poses an actual threat when studying biological variation between samples. This concerns another ongoing challenge on how to integrate all the different variances arising in a quantification process. It involves all experimental induced variances as well as abundance estimation variance in individual samples and also global variances with technical, experimental and biological replicates, while only the latter is mainly considered in methods so far. Particularly for differential abundance analysis, a decoupling of these variances becomes crucial to identify true effects caused by biological factors. A first proposal for integrating abundance variance with replicate variance in a statistical model has been recently presented by Pimentel *et al.* [97]. Further research efforts devoted to this matter are urgently needed for the high-throughput quantification field.

Additionally, future advances in sensitivity and resolution accuracy in NGS and MS devices are awaited to overcome certain issues and to decrease technical variances.

A. Appendix

Fig. A1:

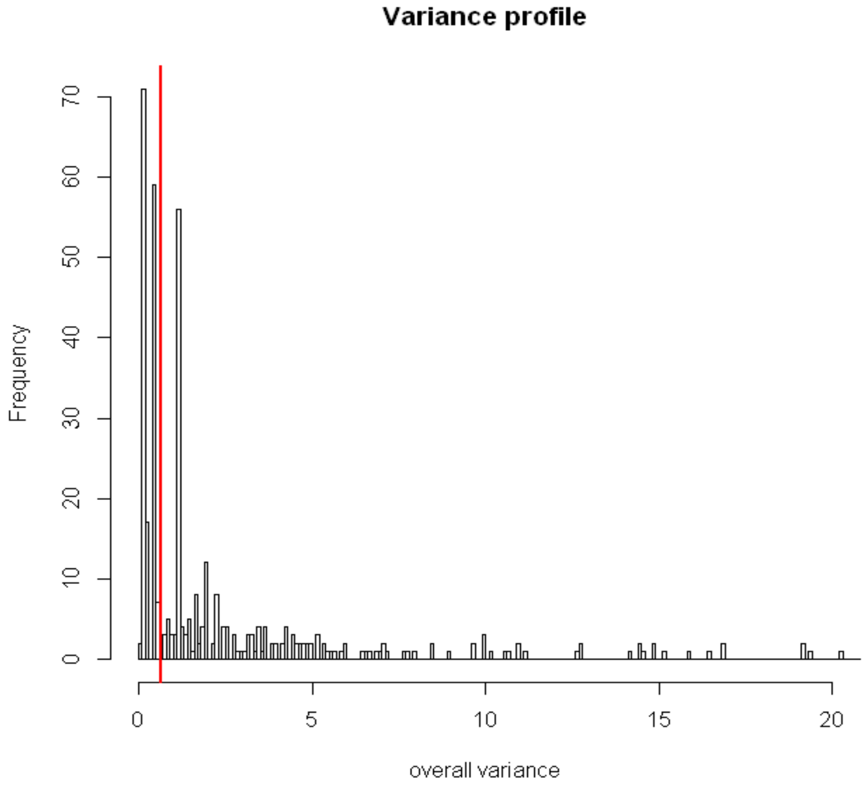


Fig. A1: Variance distribution of the proteins. The variance of the counts across all samples (bait and control) is calculated for each protein. The majority of proteins which exhibit no or only minor changes in counts between bait and control samples appear as a first peak close to zero in the variance profile. A cutoff for filtering is proposed in red.

Fig. A2:

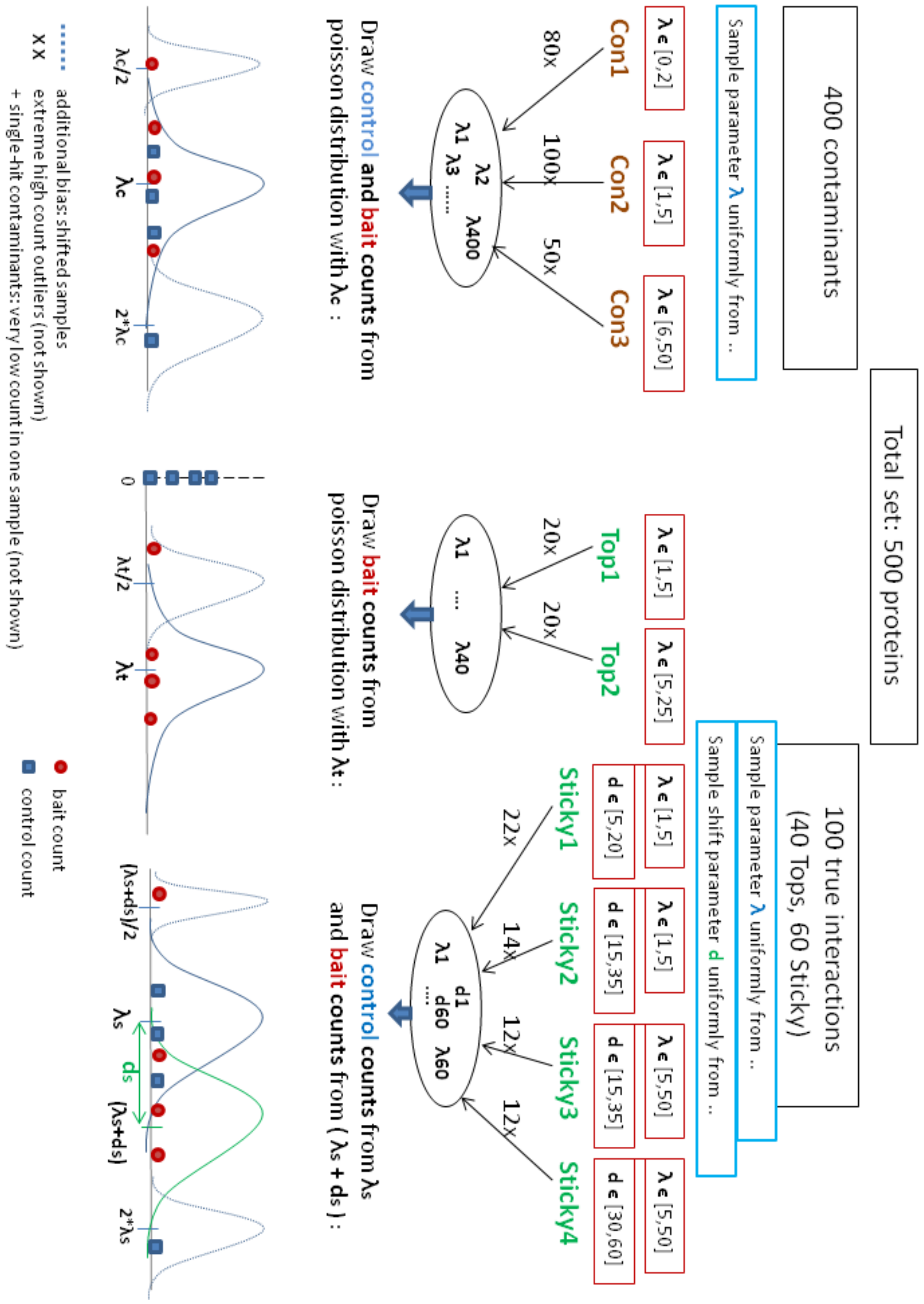
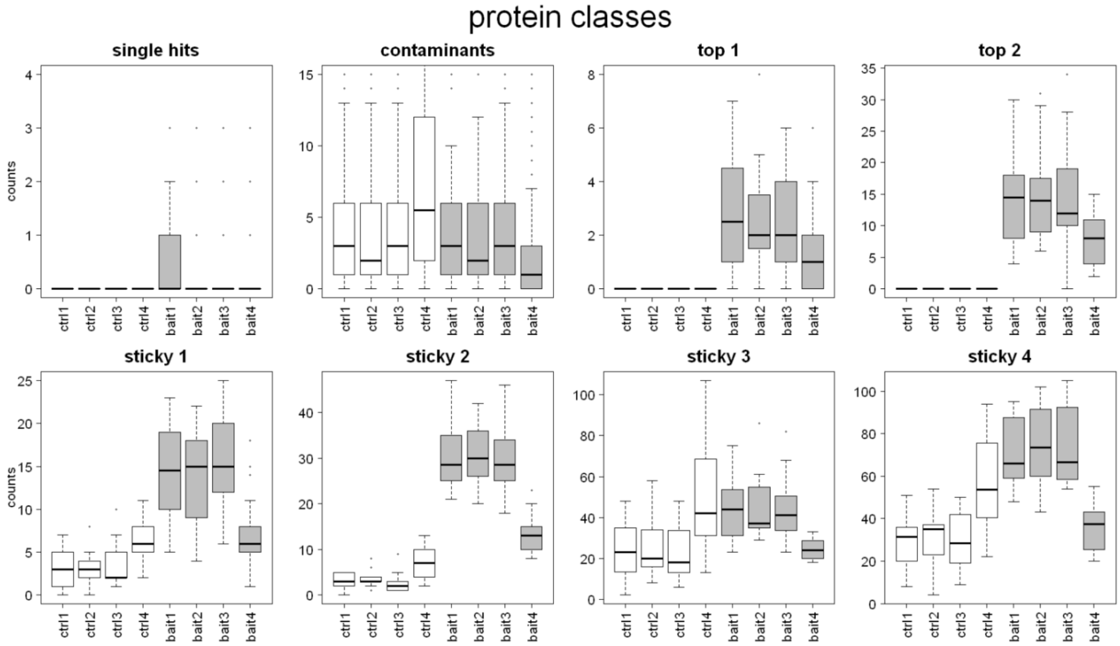

Total set: 500 proteins
400 contaminants
100 true interactions
(40 Tops, 60 Sticky)
Sample parameter λ uniformly from ..
Sample shift parameter d uniformly from ..
λ ϵ [0,2]
λ ϵ [1,5]
λ ϵ [6,50]
λ ϵ [1,5]
λ ϵ [5,25]
λ ϵ [1,5]
d ϵ [5,20]
λ ϵ [1,5]
d ϵ [15,35]
λ ϵ [5,50]
d ϵ [15,35]
λ ϵ [5,50]
d ϵ [30,60]
Con1
Con2
Con3
Top1
Top2
Sticky1
Sticky2
Sticky3
Sticky4
80x
100x
50x
20x
20x
22x
14x
12x
12x
λ1 λ2 λ3 ...... λ400
λ1 .... λ40
λ1 d1 d60 .... λ60
Draw control and bait counts from poisson distribution with λc :
Draw bait counts from poisson distribution with λt :
Draw control counts from λs and bait counts from (λs + ds) :
λc/2
λc
2*λc
0
λt/2
λt
(λs+ds)/2
λs
ds
(λs+ds)
2*λs
additional bias: shifted samples
x x extreme high count outliers (not shown)
+ single-hit contaminants: very low count in one sample (not shown)
bait count
control count

Fig. A2: Flow chart showing the construction of a simulation data set, consisting of 400 contaminants (170 single hit contaminants not shown) and 100 true interactions. The parameter λ and *d* are sampled uniformly from closed intervals of real numbers, e.g. λ ϵ [1,5] = {x ϵ R |1 ≤ λ ≤5}.

Fig. A3:
(a)



(b)

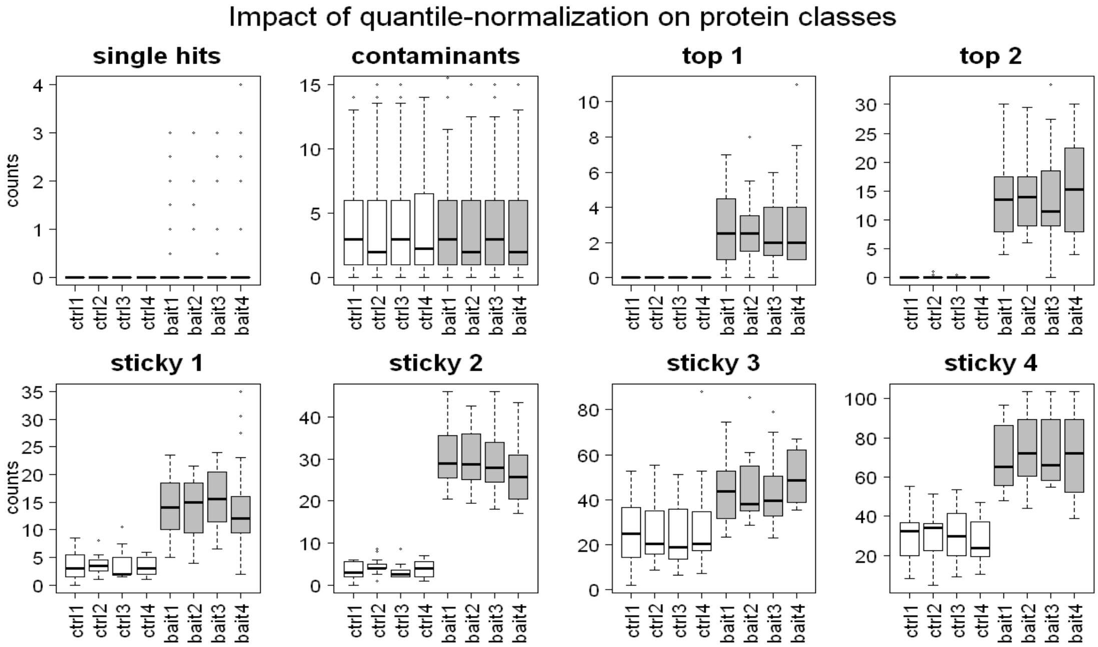


Fig. A3: **Count distribution of bait and control samples shown for the different protein classes exemplary for one simulation data set (a) without pre-processing and (b) with quantile normalization of the data, but without filtering**. The different protein classes are defined as: (i) *single-hit contaminants*: random appearance of a low count in one of the bait

samples, (ii) *contaminants*: proteins showing a similar expression across all samples (3 different classes of contaminants are pooled here), (iii) *top1*: no counts in the controls and low number of counts in the baits, (iv) *top2*: no counts in the controls and high counts in the baits, (v) *sticky1*: holding low counts across all samples with a weak dominance in the baits , (vi) *sticky2*: holding low counts across all samples with a strong dominance in the baits, (vii) *sticky3*: holding high counts across all samples with a weak dominance in the baits, and (viii) *sticky4*: holding high counts across all samples with a strong dominance in the baits.

Fig. A3b shows that normalization results in a clear separation of the count distributions between bait and control in case of the truly interacting proteins (top1-2, sticky1-4) compared to Fig. A3a.

Fig. A4:

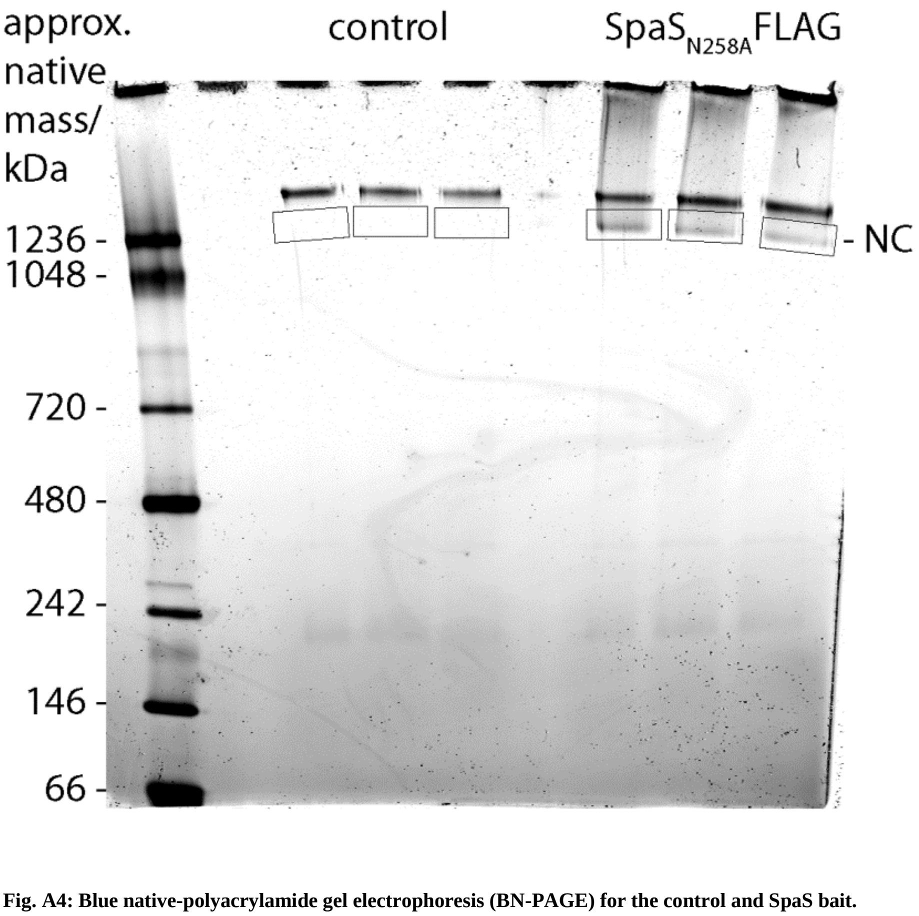


Fig. A4: Blue native-polyacrylamide gel electrophoresis (BN-PAGE) for the control and SpaS bait.

Fig. A5:

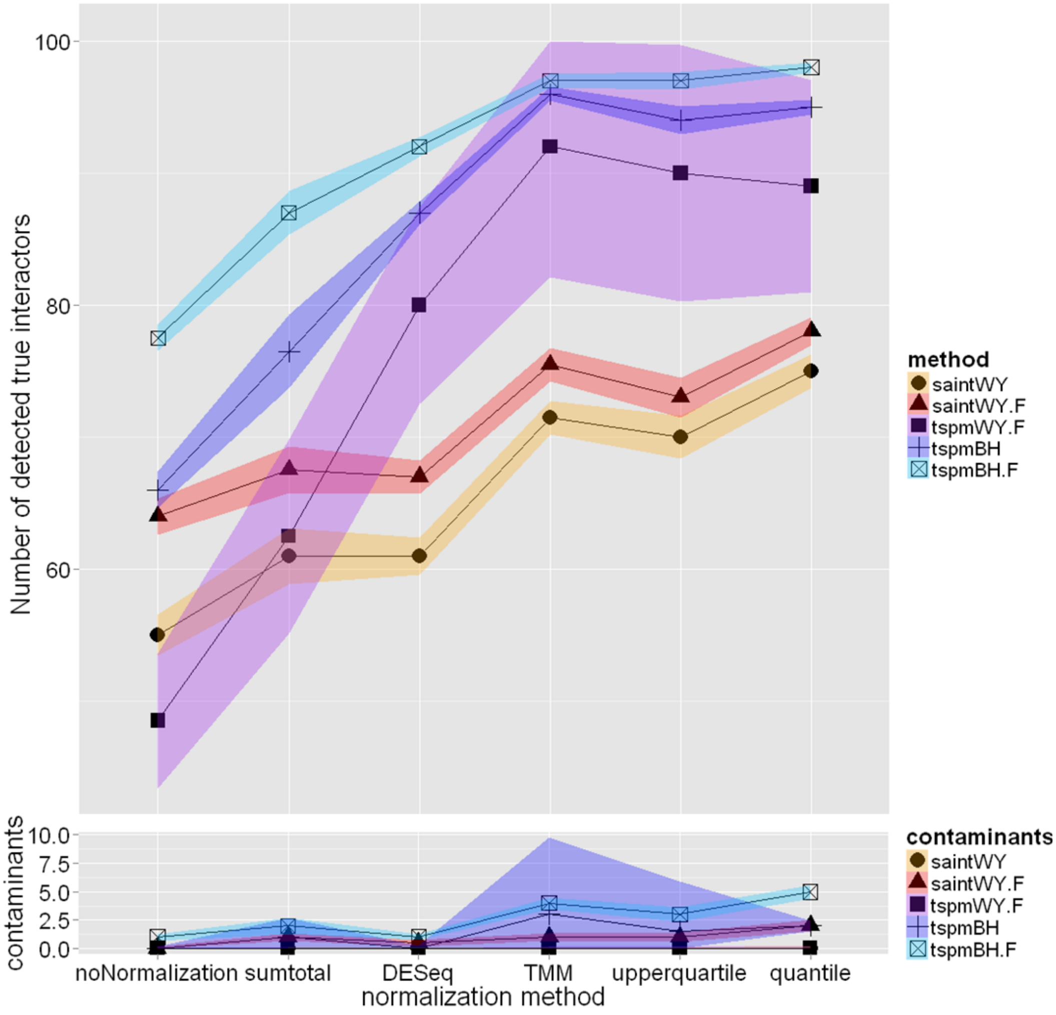


Fig. A5: **Number of identified truly interacting proteins below a threshold of 0.1 by the different workflows.** Median values of 50 simulations and corresponding 95% confidence bands are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-WY* (saintWY.F), (iv) *TSPM-WY* (tspmWY.F), and (v) *TSPM-BH* (tspmBH.F), according to the normalization method applied (reported on the x-axis). A maximum number of 100 true interactors can be obtained based on the ground truth. Median values and 95% confidence bands are presented for the identified false-positives (contaminants) correspondingly. In comparison to Figure 2.3 in the main text, more truly interacting candidates are detected and at the same time more contaminants are included in the final list, especially in case of the FDR based workflows, which is clearly expected due to the higher threshold of 0.1.

Fig. A6:

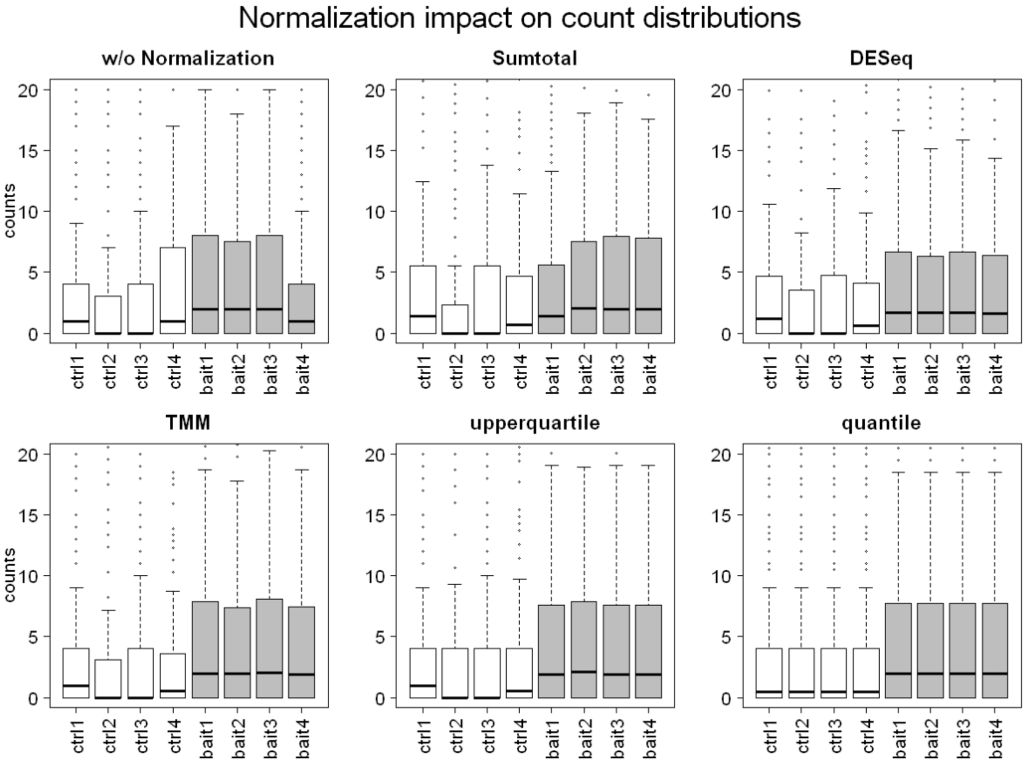


Fig. A6: **Count distribution of bait and control samples** of one selected simulation data set containing all interaction candidates (i) without normalization, (ii) with sumtotal normalization, (iii) with normalization by DESeq, (iv) with normalization by TMM, (v) with upperquartile normalization, and (vi) with quantile normalization. Normalization of the data results in the expected stabilization of count distributions within replicate bait samples and within replicate controls**.**

Fig. A7:

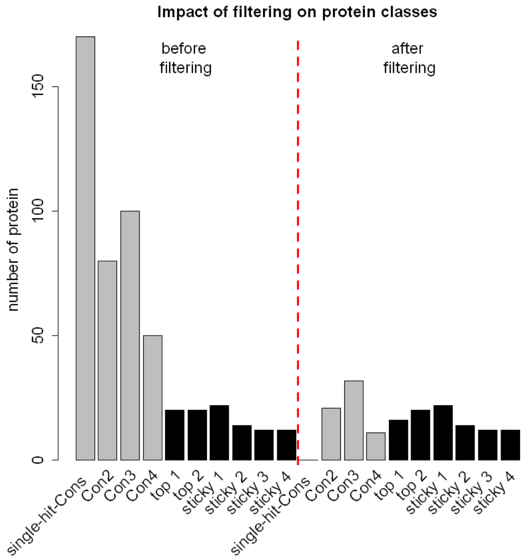


Fig. A7: **Number of proteins in the protein classes before and after the filtering step** (based on one representative simulated raw data set without normalization). As a result of the filtering step, single-hit contaminants (defined by a low count in only one sample) are completely removed and the remaining contaminant classes (Con 2-4) are significantly decreased, corresponding to approximately 70% in this case. However, the number of truly interacting proteins (top 1-2, sticky 1-4) in the data set is almost completely maintained, only 4% of the truly interacting proteins are lost due to the filtering (see also Fig. A8).

Fig. A8:

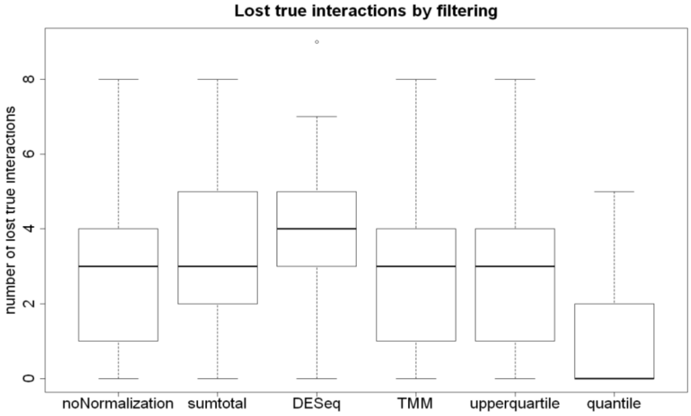


Fig. A8: **Distribution of truly interacting proteins lost due to the filtering step in 50 simulations dependent on the normalization method.** The lowest number of proteins is lost applying the quantile normalization. Overall, the median corresponds to three interactions lost by filtering, which is acceptable as the benefit of filtering is still larger than its decreasing effect.

Fig. A9:

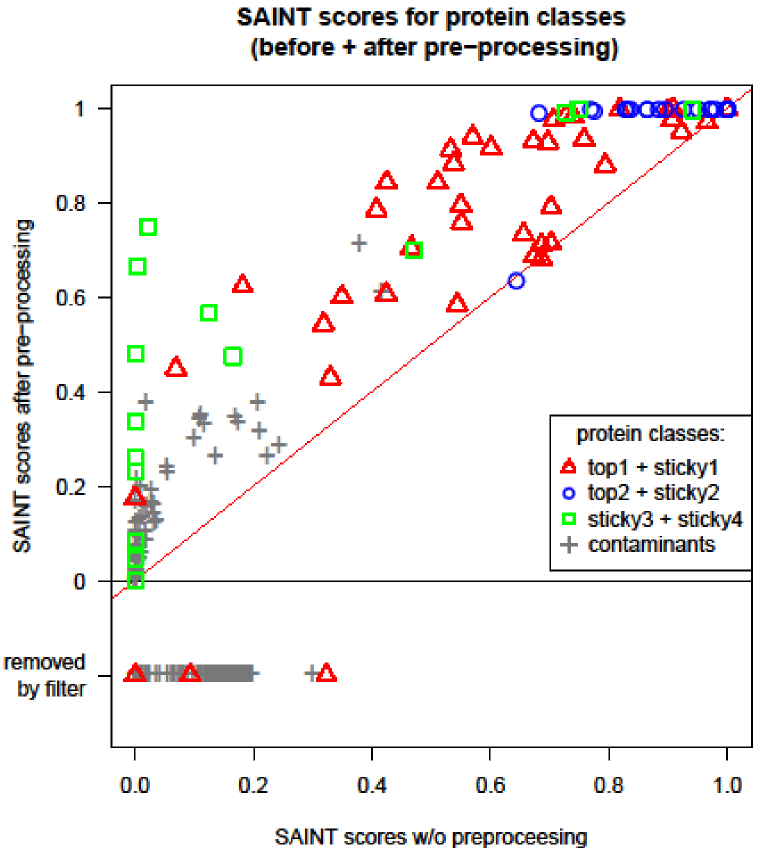


Fig. A9: **SAINT scores before and after pre-processing** (TMM normalization and filtering) of one selected data set, for different protein classes: true interactors with low counts in the controls and (i) a weak presence in the baits (*top1+sticky1*) or (ii) a strong presence in the baits (*top2+sticky2*), (iii) true interactors with high counts across all samples, but a superior presence in the baits (*sticky3+4*), and (iv) *contaminant* proteins. Dots above the diagonal represent proteins which receive an increased SAINT score after pre-processing. Proteins removed due to the filtering step are shown below the x-axis (also refer to Fig. A6).

Before pre-processing, especially the classes *sticky3-4* receive very low scores predominantly exactly at zero. As a consequence, proteins holding a score of zero, must always obtain a p-value of exactly one because no permutation score can be smaller than an original score of zero, thus there is no chance of identifying these truly interacting proteins without pre-processing. Preprocessing of the data raises the scores obtained by SAINT up to 0.75 for the classes *sticky3-4* and scores are also improved for the classes of *top1* and *sticky1*.

Fig. A10:

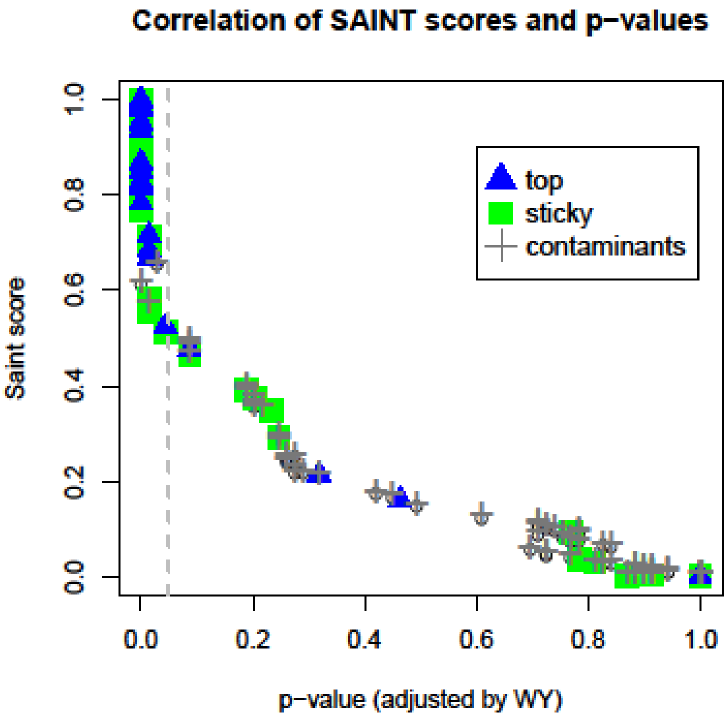


Fig. A10: **Correlation of SAINT scores and p-values** shown for the three classes of proteins: (i) easily detectable true interactors (*top*), (ii) challenging class of true interactors (*sticky*), and (iii) *contaminant* proteins of one selected simulation data set. P-values are calculated by *SAINT-WY* with pre-processing (quantile normalization + filtering). True interactors with an adjusted p-value<0.05 (vertical dashed line) correspond to scores ranging from 0.5 to 1.0.

Fig. A11:

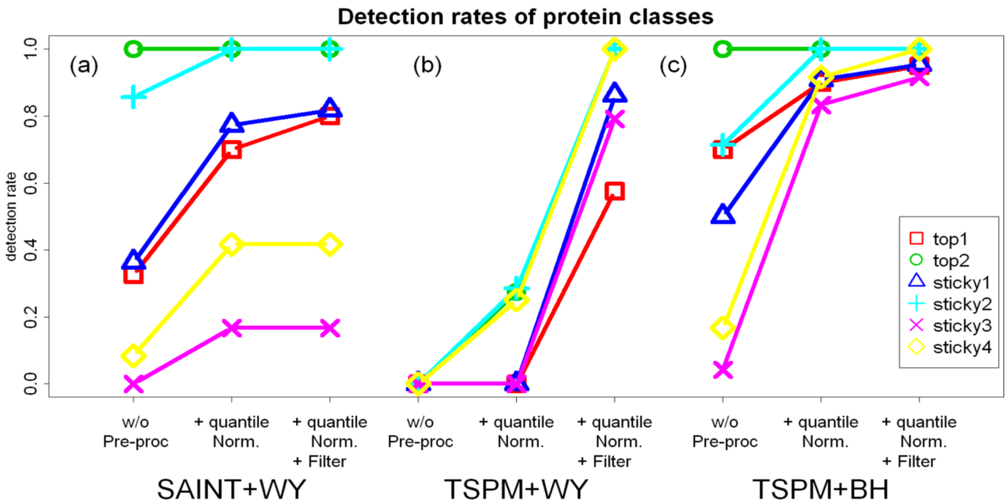


Fig. A11: Median detection rates of the six individual classes of truly interacting proteins applying the quantile normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without preprocessing, (ii) with quantile normalization, (iii) with quantile normalization and filtering of the data. Different classes of true interactors are *top1-2* having no counts in the controls and weak or strong presence in the baits respectively, *sticky1-2* holding low counts and *sticky3-4* high counts in the controls with weak or strong presence in the baits respectively.

Fig. A12:

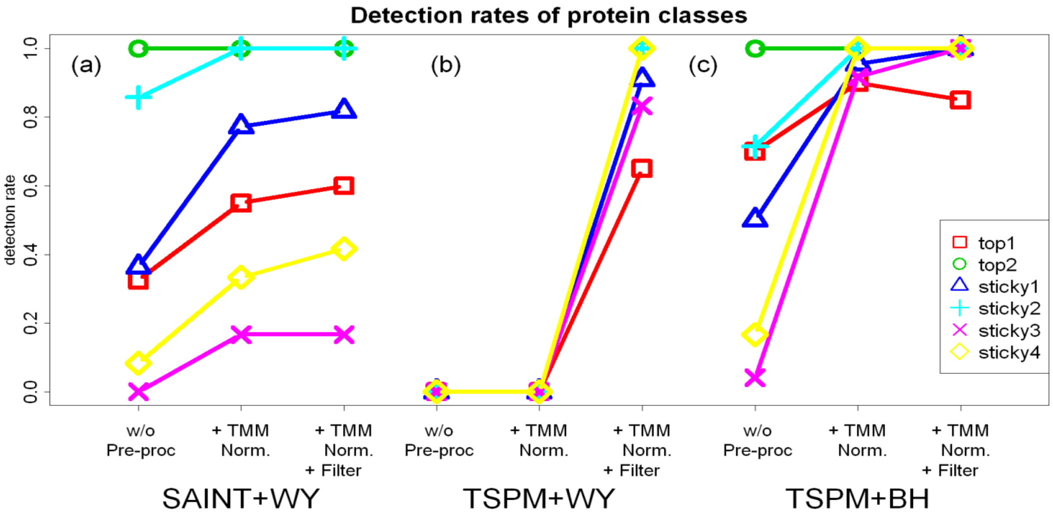


Fig. A12: Median detection rates of the six individual classes of truly interacting proteins applying the TMM normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without pre-processing, (ii) with TMM normalization, (iii) with TMM normalization and filtering of the data. Different classes of true interactors are top1-2 having no counts in the controls and weak or strong presence in the baits respectively, sticky1-2 holding low counts and sticky3-4 high counts in the controls with weak or strong presence in the baits respectively.

Fig. A13:

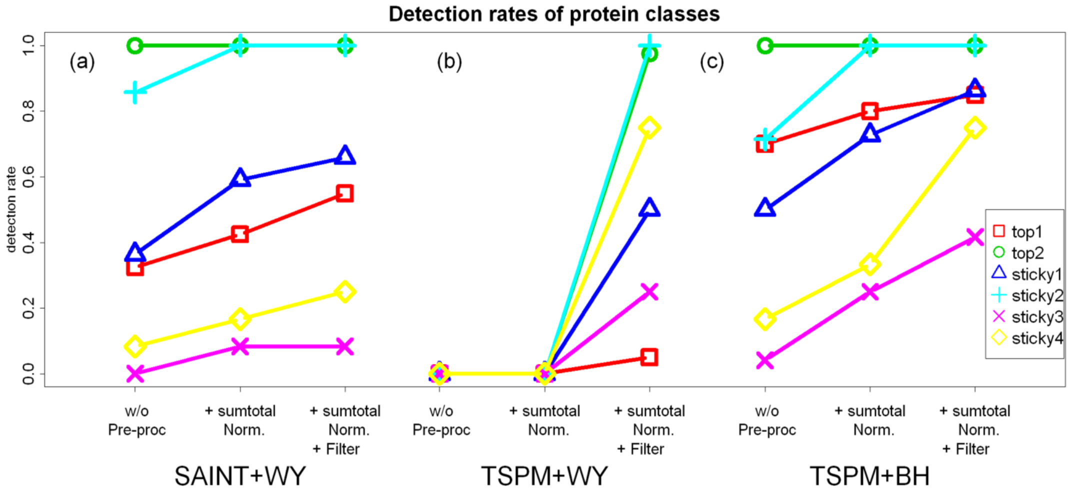


Fig. A13: Median detection rates of the six individual classes of truly interacting proteins applying the sumtotal normalization. Median detection rates of 50 simulations are calculated for the three different methods (a) *SAINT-WY*, (b) *TSPM-WY* and (c) *TSPM-BH*, in each case (i) without pre-processing, (ii) with sumtotal normalization, (iii) with sumtotal normalization and filtering of the data. Different classes of true interactors are top1-2 having no counts in the controls and weak or strong presence in the baits respectively, sticky1-2 holding low counts and sticky3-4 high counts in the controls with weak or strong presence in the baits respectively. *TSPM+WY* reveals difficulties in identifying protein classes defined by a small difference in the bait samples.

Fig. A14:

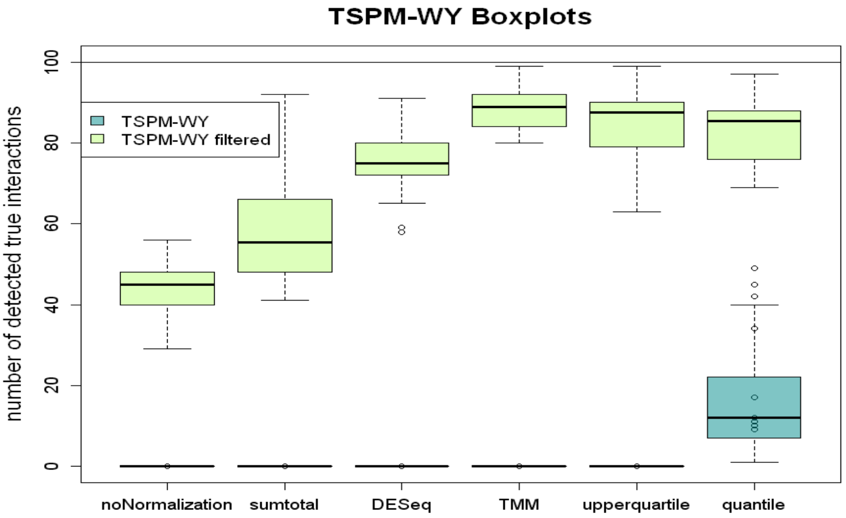


Fig. A14: Boxplots showing the number of truly interacting proteins identified by the workflow *TSPM-WY* below an adjusted p-value of 0.05 based on 50 simulations, dependent on the normalization method applied and with or without filtering of the data. The workflow exhibits difficulties to detect any of the true interactions without the filtering due to an outlier in the data. Filtering and normalization significantly increase the number of detections.

Fig. A15:

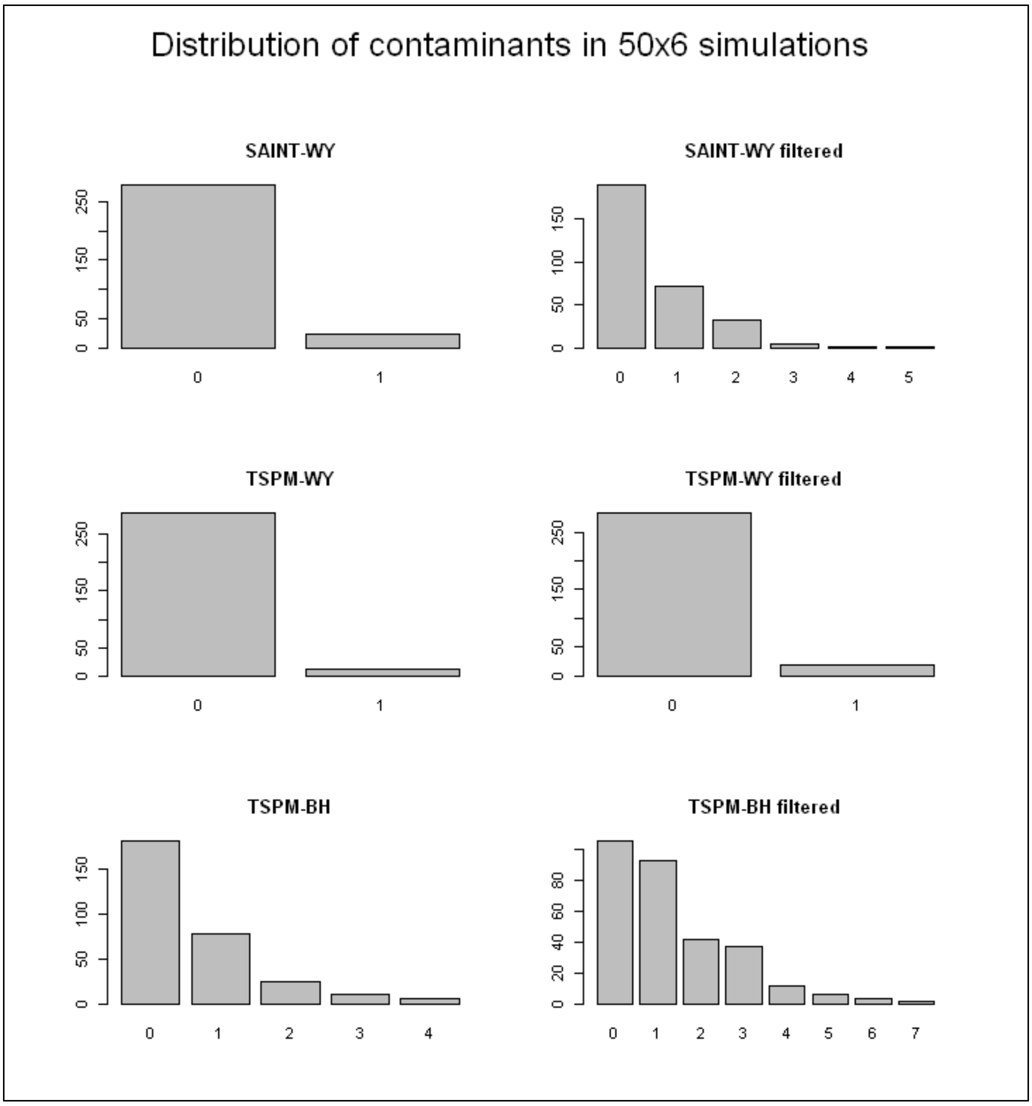
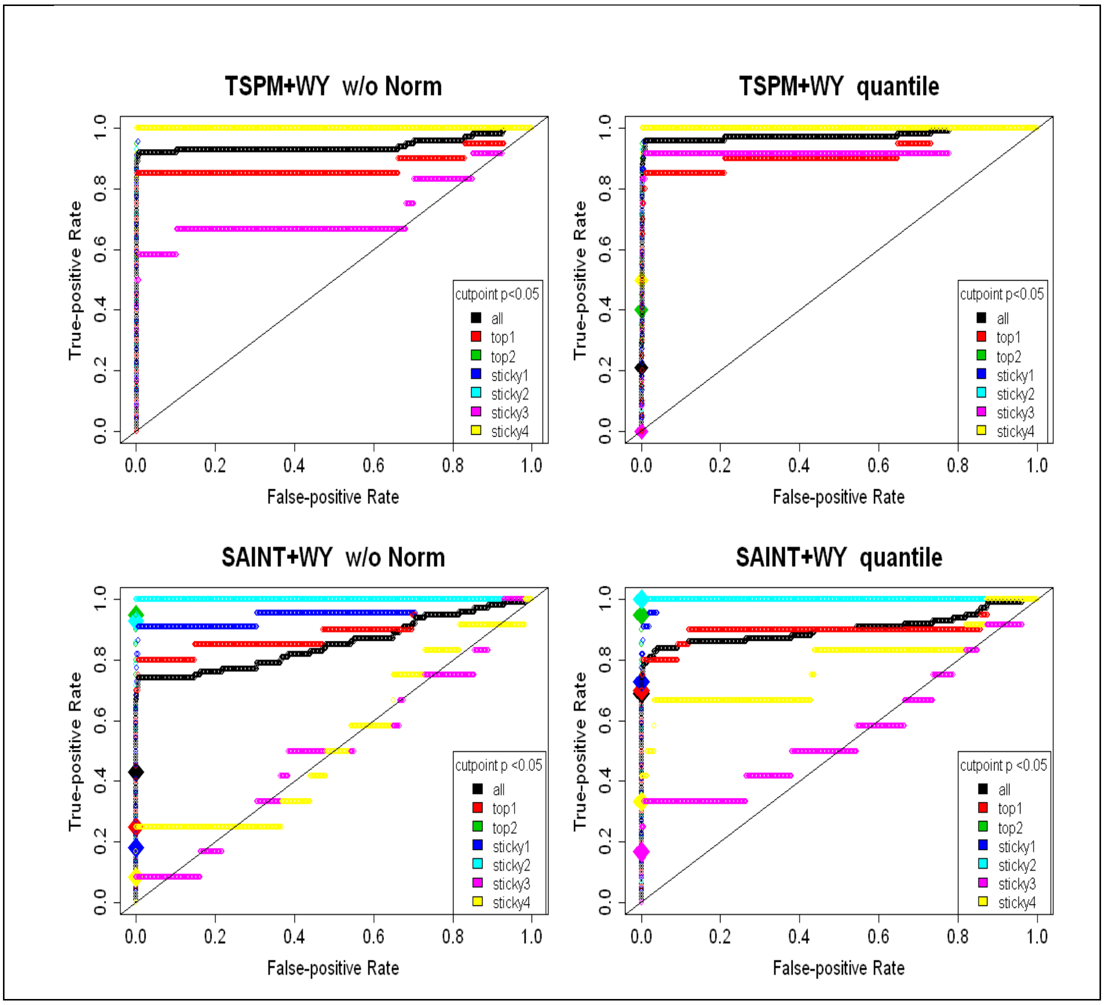


Fig. A15: Overview on the number of contaminants detected below a threshold of 0.05 in each of the 50 simulations and normalization methods applied, for the different workflows (i) *SAINT-WY*, (ii) *TSPM-WY*, and (iii) *TSPM-BH*, without and with filtering of the data**.** Thereby, *SAINT-WY*, *TSPM-WY* and *TSPM-WY filtered* show at most one contaminant in a final list of candidates in all the simulation runs. Between zero and seven contaminants are found for *TSPM-BH* without and with filtering, which is to be expected for the FDR control.

Fig. A16:

(a)


TSPM+WY w/o Norm
TSPM+WY quantile
SAINT+WY w/o Norm
SAINT+WY quantile
True-positive Rate
False-positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
cutpoint p<0.05
cutpoint p <0.05
■ all
■ top1
■ top2
■ sticky1
■ sticky2
■ sticky3
■ sticky4

(b)

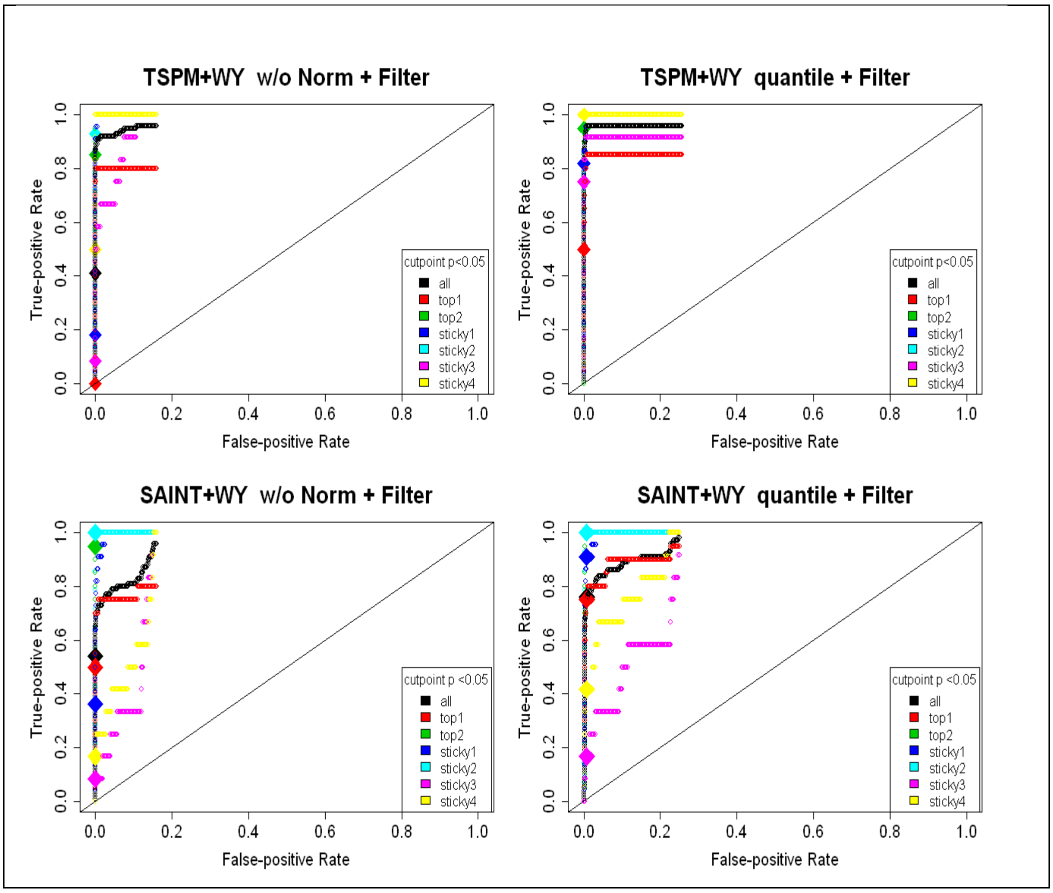


Fig. A16: ROC curves presenting the correlation of false positives and true positives for the different protein classes for the method *SAINT-WY* and *TSPM-WY* based on one exemplary simulation data set (i) without normalization, (ii) with quantile normalization, (iii) without filtering, and (iv) with filtering of the data. The overall ROC curve of all true interactions is shown in black. The relationship of true positives to false positives obtained by choosing all proteins below a threshold of 0.05 is marked by a rhomb.

Fig. A17:

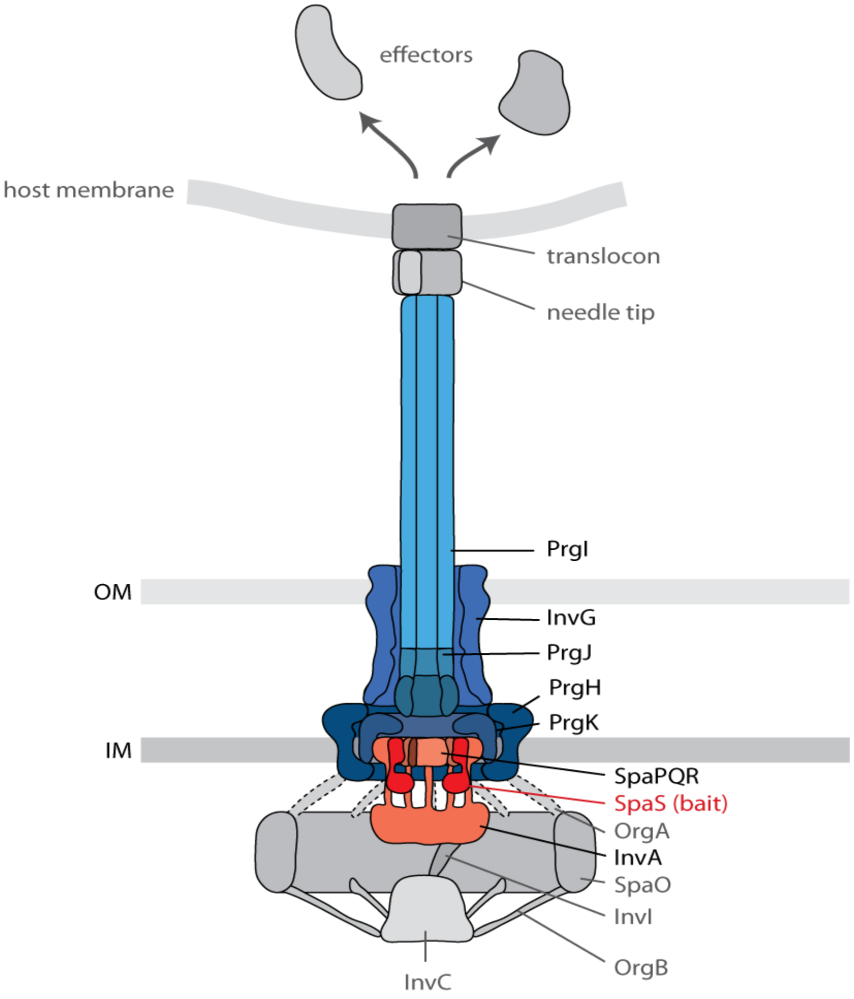


Fig. A17: Schematic representation of the macromolecular machinery of the type III secretion system encoded by pathogenicity island 1 of *Salmonella* Typhimurium in the presence of host cells: Proteins of the needle complex (blue) and the export apparatus (red) are clearly expected from the experiment. The cytosolic components (gray) are rather loosely associated with the rest of the complex and are easily lost during purification.

Fig. A18:

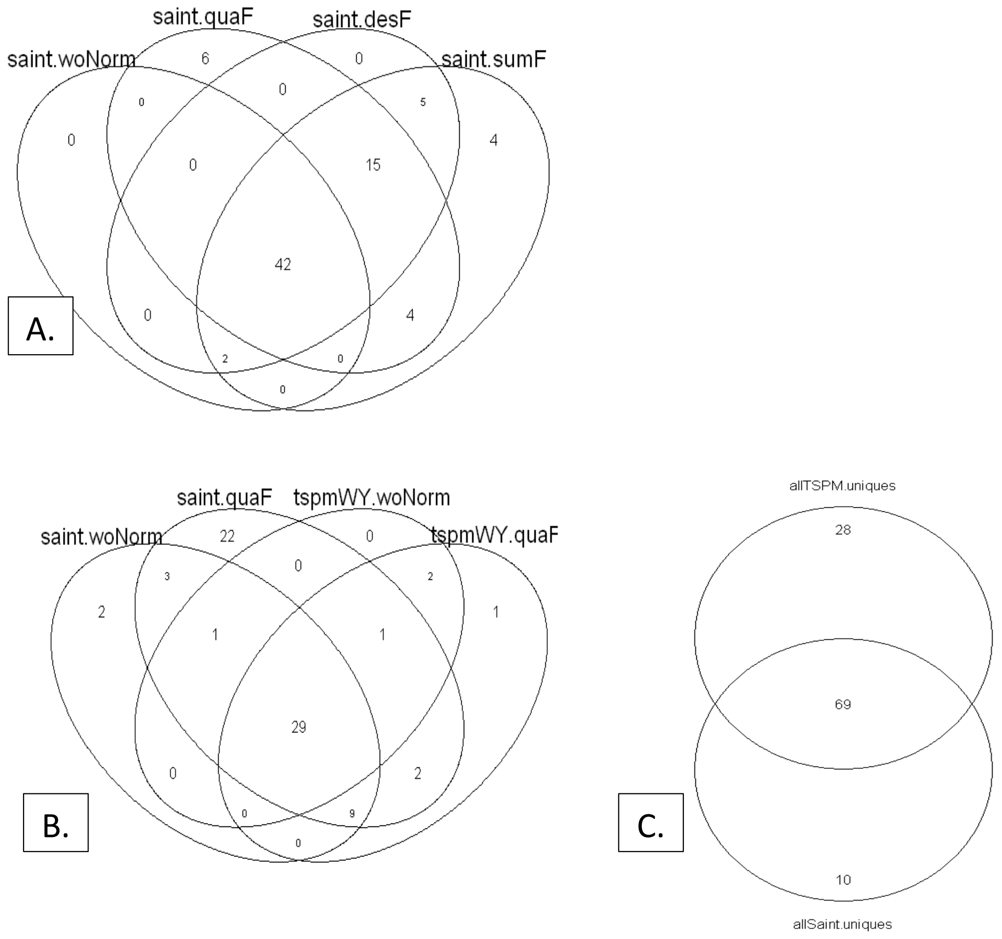


Fig. A18:

Venn diagram according to Table 2.1 (main manuscript), showing the intersection of candidates detected by A) *SAINT WY* without any preprocessing *(saint.woNorm)*, with quantile, DESeq and sumtotal normalization including filtering *(saint.quaF, saint.desF, saint.sumF)*; B) *SAINT-WY* and *TSPM-WY*, each without preprocessing *(saint.woNorm, tspmWY.woNorm)* or with quantile normalization and filtering respectively *(saint.quaF, tspmWY.quaF)*; C) The intersection is assessed for all proteins pooled by the different preprocessing methods for SAINT (79 proteins *allSAINT.uniques*) and all proteins pooled by the different preprocessing methods for TSPM (97 proteins *allTSPM.uniques*). The smallest intersect of all methods, independent of the preprocessing method used, are 29 candidates (subset of the 69 proteins).

Fig. A**19:**

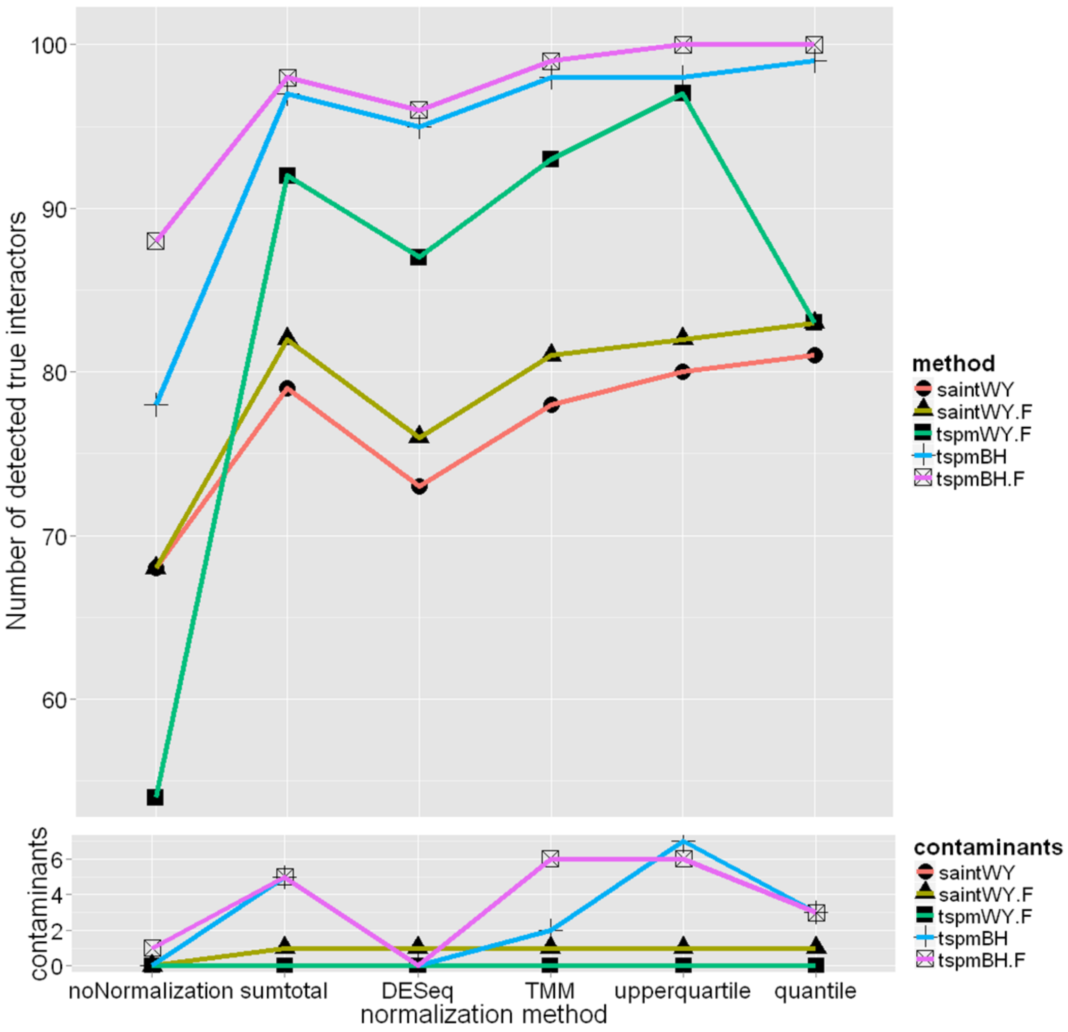


Fig. A19: **Robustness study on sample size using 5 replicate bait and control samples**. Number of identified truly interacting proteins below a threshold of 0.05 by the different workflows are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-WY* (saintWY.F), (iv) *TSPM-WY* (tspmWY.F), and (v) *TSPM-BH* (tspmBH.F), according to the normalization method applied (reported on the x-axis). A maximum number of 100 true interactors can be obtained based on the ground truth.

We observe the same trend as in Figure 2.3 in the main text – only the sumtotal normalization shows an improved performance, as the additional replicates compensate for the introduced outliers and biases in the data.

Fig. A**20:**

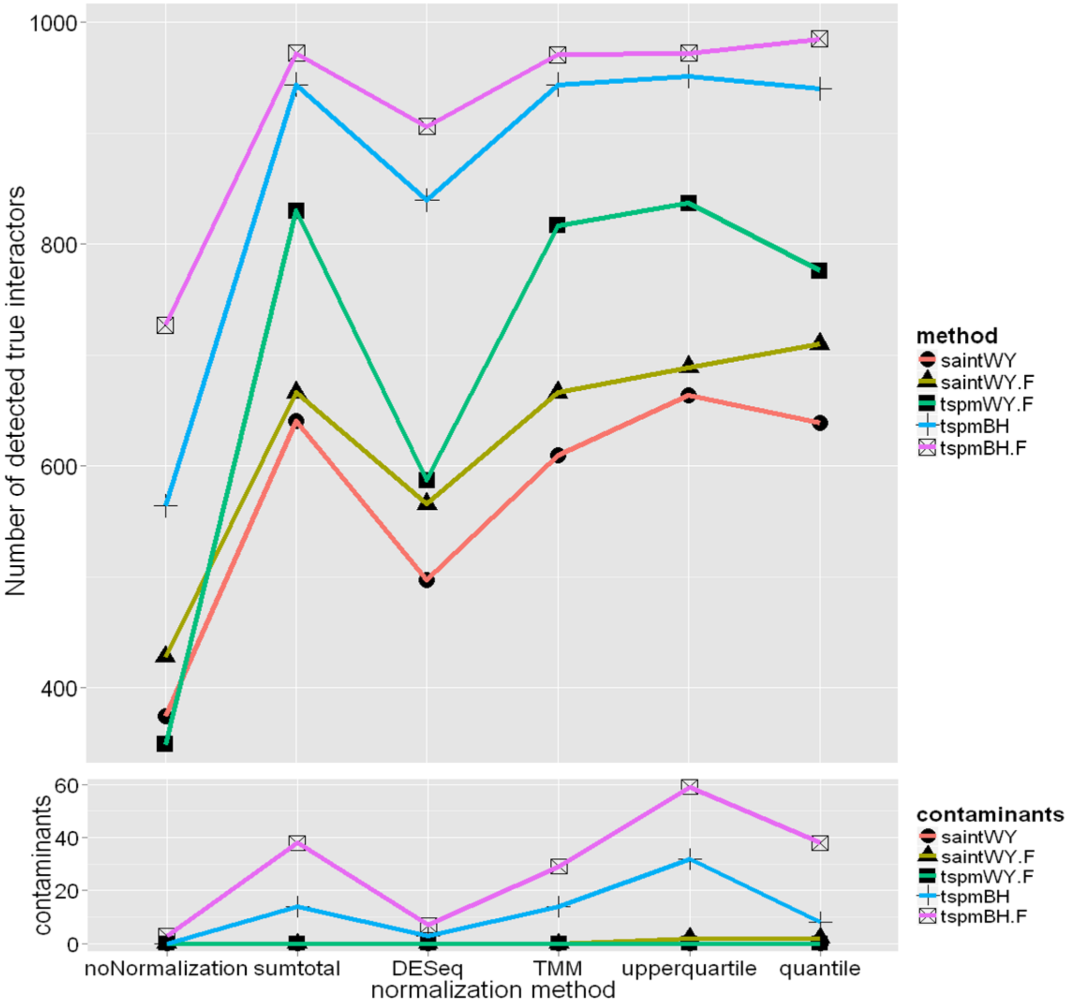


Fig. A20: **Robustness study on sample size using a total set of 5000 proteins**. Number of identified truly interacting proteins below a threshold of 0.05 by the different workflows are shown without filtering for (i) *SAINT-WY* and (ii) *TSPM-BH* , and with filtering for (iii) *SAINT-WY* (saintWY.F), (iv) *TSPM-WY* (tspmWY.F), and (v) *TSPM-BH* (tspmBH.F), according to the normalization method applied (reported on the x-axis). A maximum number of 1000 true interactors can be obtained based on the ground truth.

We observe the same trend as in Figure 2.3 in the main text – as already seen in Fig. A19 the sumtotal normalization shows an improved performance, as the increased number of candidate proteins compensates for the introduced outliers in the data.

Fig. A21:

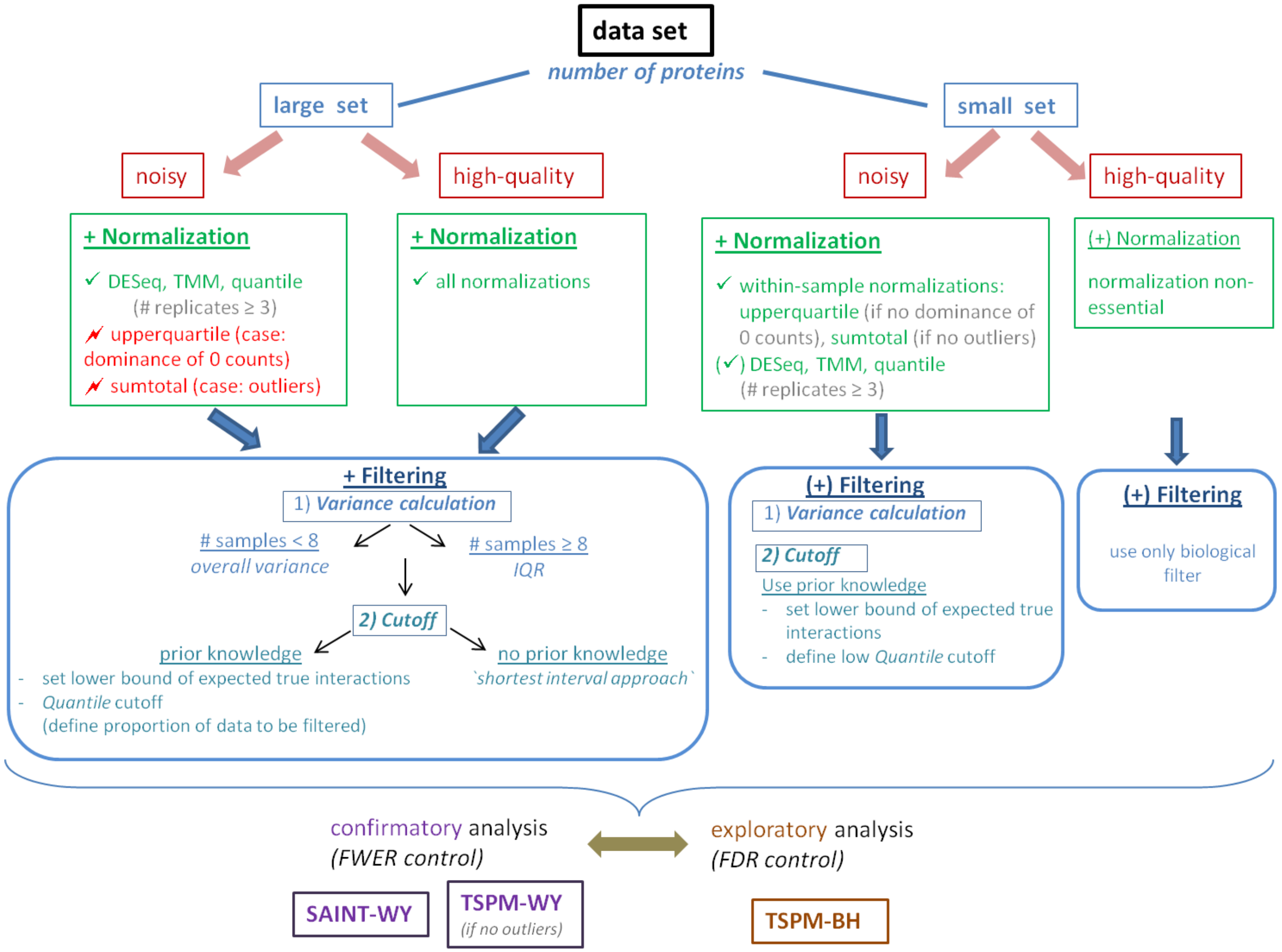


Fig. A21: **Workflow guidelines dependent on data characteristics.** Here '+' denotes essential steps and '(+)' denotes less essential steps.

Table A1:

without filtering			+ filtering		*TSPM-BH*	
Normalization method:	*SAINT-WY*	*TSPM-WY*	*SAINT-WY*	*TSPM-WY*	w/o filtering	+ filtering
w/o Norm.	41	0	48	36	43	66 (1)
sumtotal	41	0	46	39	50	69 (1)
DESeq	46	0	50	45	62	79 (3)
TMM	55	0	61	71	79 (4)	89 (3)
upperquartile	54	0	58	71	76 (3)	81 (3)
quantile	55	4	58 (1)	60	72 (3)	84 (4)

Table A1: Number of identified truly interacting proteins below a threshold of 0.05, based on 100 true interactors in the negative binomial simulation study.
Application of the two FWER controlled workflows *SAINT-WY* and *TSPM-WY,* and the FDR based workflow *TSPM-BH* (i) without normalization (*w/o Norm*.), (ii) with five different normalization methods, (iii) without filtering, and (iv) with filtering of the data. Numbers of contaminants below a threshold of 0.05 are shown in brackets.The same trend is observed here as in the results of the simulation study based on poisson distributions: Normalization as well as filtering increase the number of detected true interactions. *TSPM-WY* exhibits the same difficulties in detecting true interactions without filtering due to outliers in the data, but performs competitive to SAINT with the filtering. However, all methods here identify less true interactions compared to the simulation study based on poisson distributions, as the latter data setup fits presumably better to the scoring models.

Table A2:

	SAINT-WY w/o filtering	*SAINT-WY* + filtering	*TSPM – WY* w/o and with filtering	*TSPM – BH* w/o filtering	*TSPM – BH* + filtering
HflK (*Uniprot: E1WF50*)					
w/o Norm.			**x**	**x**	**x**
TMM			**x**	**x**	**x**
quantile			**x**	**x**	**x**
upperquartile			**x**	**x**	**x**
DESeq			**x**	**x**	**x**
sumtotal			**x**	**x**	**x**
FtsH (*Uniprot: E1WI79*)					
w/o Norm.				**x**	**x**
TMM				**x**	**x**
quantile			**x**	**x**	**x**
upperquartile			**x**	**x**	**x**
DESeq				**x**	**x**
sumtotal			**x**	**x**	**x**
HtpX (*Uniprot: E1WG81*)					
w/o Norm.					
TMM					
quantile				**x**	**x**
upperquartile					**x**

DESeq					
sumtotal					
L17 (*Uniprot: E1WIJ1*)					
w/o Norm.	x	x			
TMM	x	x			
quantile	x	x	x	x	x
upperquartile	x	x			x
DESeq	x	x			
sumtotal	x	x			
S12 (*Uniprot: E1WIM5*)					
w/o Norm.	x	x			
TMM	x	x			
quantile	x	x			
upperquartile	x	x			
DESeq	x	x			
sumtotal	x	x			
L5 (*Uniprot: E1WIK5*)					
w/o Norm.					x
TMM	x	x			x
quantile	x	x		x	x
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
L15 (*Uniprot: E1WIJ8*)					
w/o Norm.					x
TMM	x	x			x
quantile	x	x		x	x
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
YajC (*Uniprot: E1W8R7*)					
w/o Norm.		x			
TMM	x	x			
quantile	x	x			
upperquartile	x	x			
DESeq	x	x			
sumtotal	x	x			
S11 (*Uniprot: O54296*)					
w/o Norm.				x	x
TMM				x	x
quantile	x	x		x	x
upperquartile				x	x
DESeq	x	x		x	x
sumtotal		x		x	x
RpoA (*Uniprot: E1WIJ2*)					
w/o Norm.					x
TMM					x
quantile	x	x		x	x
upperquartile				x	x
DESeq				x	x
sumtotal		x		x	x
L16 (*Uniprot: E1WIL0*)					
w/o Norm.					x
TMM					x

quantile	x	x		x	x
upperquartile				x	x
DESeq		x		x	x
sumtotal		x		x	x
HflC (*Uniprot: E1WF51*)					
w/o Norm.		x			x
TMM		x			x
quantile					
upperquartile	x	x		x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x
PrgJ (*Uniprot: E1WAB7*)					
w/o Norm.	x	x		x	x
TMM	x	x		x	x
quantile	x	x	x	x	x
upperquartile	x	x	x	x	x
DESeq	x	x		x	x
sumtotal	x	x	x	x	x
SpaQ (*Uniprot: E1WAD3*)					
w/o Norm.	x	x		x	x
TMM	x	x		x	x
quantile	x	x	x	x	x
upperquartile	x	x	x	x	x
DESeq	x	x		x	x
sumtotal	x	x		x	x

Table A2: Detection of interaction candidates in the *Salmonella* study below an adjusted p-value of 0.1 (denoted by **x**) dependent on the methods applied: (1) *SAINT-WY*, (2) *TSPM-WY*, and (3) *TSPM-BH* in combination without and with the five proposed normalization methods and with or without the filtering step.

Table A3:

without filtering		**+ filtering**
Normalization method:	FDR by *SAINT* < 0.05	FDR by *SAINT* < 0.05
w/o Norm.	38	40
sumtotal	53	57
DESeq	51	55
TMM	60	64
upperquartile	58	61
quantile	61	65

Table A3: **Number of identified truly interacting proteins in the simulation data according to the approximated FDR proposed by SAINT below a threshold of 0.05**. Numbers are assessed for the simulation data (i) without normalization (w/o Norm.), (ii) applying the five different normalization methods, (iii) without filtering, and (iv) with filtering of the data. No contaminant proteins are found within the identified proteins. A comparison to Figure 2.3 in the main text reveals that the FDR by SAINT is more conservative than the FWER by Westfall&Young as less truly interacting proteins are detected here.

METHOD Section: *additional Details*

Overdispersion in TSPM

TSPM generally relies on poisson distributed data; however, overdispersion can occur in case a protein shows a larger variation in its counts across the samples than theoretically expected. Overdispersion is a common issue in experiments in which samples originate from different biological conditions, thus counts reflect technical as well as biological variance. However, the setup for AP-MS data is different since - based on a defined pool of proteins - interaction partners are searched with and without a bait protein. Thus, overdispersion is not expected in single-bait experiments as counts reflect rather a technical variance, theoretically following a poisson distribution. However, in the event overdispersion occurs, TSPM is able to treat these proteins separately. It uses a random effects model and an adjusted score test in a first step to identify overdispersed candidates, which are followed by a quasi-likelihood approach respectively. The original TSPM implementation relies on the presence of several proteins displaying overdispersion, while our adaption of TSPM to AP-MS data consequently allows the case of having no overdispersed proteins in the data.

Scoring method: SAINT

The underlying assumption of SAINT is that each observed protein count is derived from a mixture distribution. Thereby, spectral counts of a protein are assumed to follow either a poisson distribution representing true interactions or a poisson distribution with a different mean count in case of a false interaction. A Bayesian modeling approach is used to estimate these count distributions for true and false interactions in order to infer from the given count whether the considered prey and bait protein share a true interaction. Thereby, various features of the proteins are integrated, such as the protein length, the total number of spectra in a sample as well as any present interactions involving the prey or the bait in the overall experiment. The distribution of false interactions is modeled based on the negative controls. SAINT also allows using a sufficiently large number of independent bait purifications instead of controls, provided that they are not closely related.

Filtering Step

Further details on the biological filter and its impact:

From a biological point of view, there is no sense of keeping candidates showing higher counts in the controls than in the baits, as they are clearly no true interaction proteins. Concerning statistical testing, it is favorable to reduce the number of tests to meaningful candidates, as the multiple testing problem increases with each test. If the proportion of noise candidates is too large, the multiple testing corrections will impede the identification of true interaction candidates.

Further, the permutation principle guarantees the appearance of candidates holding higher counts in the controls than in the baits in the data set. By substituting bait and control samples in the permutation step, a former true interactor will turn into an outlier of this kind at some point of the permutation process. Hence, the generated permutation sets will always contain these outlier proteins in the overall data set and a balanced overall distribution exists.At the same time, if the ‘original’ outlier proteins remain in the data, they can receive a very strong impact on other proteins in case their corresponding counts in the controls are very high. The potential high scores, the ‘original’ outliers may obtain in the permutation sets, disturb the very sensitive procedure of Westfall & Young. A possible resulting effect is shown for the *TSPM-WY* workflow, which fails to detect any of the true interactions without filtering of the data due to an extreme outlier. In case TSPM is used in combination with the Benjamini-Hochberg adjustment removal of these outliers is not affecting an overall null distribution, as calculations are conducted protein-wise.

Guidelines for the cutoff choice:

One main challenge in the filtering step is to define a reasonable cutoff. In general, the decision is based on the quality of the data as well as on the number of truly interaction proteins one expects in the studied system. In case the data set is large and a certain amount of noise is expected, filtering becomes more important. Here, a cutoff according to a quantile can be set in order to filter 20%, 30% or any selected proportion of the data which is clearly expected to be noise (for example: A quantile cutoff of 0.2 filters 20% of the proteins showing the smallest variance). In contrast, if the data set is very small and the measurements are assumed to be of high quality, a low cutoff should be chosen or even no filtering should be conducted. In general, the cutoff decision is always coupled to the intention of the experiment and constitutes a critical tradeoff between new detections and loss of potential candidates due to filtering.It is strongly recommended to use available biological knowledge concerning the minimal number of expected true interactions; a parameter in the filtering step can be set accordingly and defines a fixed lower bound. In case that no prior knowledge is available for defining a quantile cutoff, a common approach is to determine the shortest interval containing 50% of the data in the variance distribution, assuming the majority of proteins holds a small variance. The mean of the calculated interval can be used as cutoff (default of the variance filter in *apmsWAPP*). For users who are willing to investigate their data in more depth, we recommend to view the overall variance or IQR distribution of the proteins. The majority of proteins which exhibit no or only minor changes in counts between bait and control samples appear as a first peak close to zero in the variance profile (see Fig. A1: example with proposed cutoff in red).

Preserving the type-1-error-control:

In general, the choice of the filtering method needs to be in agreement with the following test procedure since the risk of obtaining overly optimistic results and a loss of the type-1-error control persists otherwise. The proposed combination of an overall variance filter with the permutation-based Westfall & Young method is expected to increase the power, while maintaining the control of the type-1-error as long as the filtering is conducted before the permutation.(*Bourgon et al. PNAS 2010*)

Implementation of the framework

The introduced framework is implemented in the package *apmsWAPP* for R (version 2.14 and above) and is available as a workflow in the OpenMS framework. Both can be downloaded from https://sourceforge.net/projects/apmswapp/.

Application of the three different workflows in R is based on two main commands, enabling researchers with little knowledge of R to use it. The different pre- and postprocessing options can easily be set in the main command. Application of the workflow based on SAINT requires a LINUX environment; the R-package was tested with SAINT version 2.3.4. Data input formats correspond to the input formats used by SAINT – a bait-file, an interaction-file and a prey-protein-file in the form of three tab-delimited files

For Open MS, we provide a KNIME based workflow which integrates the AP-MS pre- and post-processing steps along with identifications based on MS/MS search. Details regarding installation and system requirements are provided in the README file.

SIMULATION RESULTS: Study on different protein classes

A total set of 500 proteins is simulated, consisting of 400 contaminant proteins and 100 truly interacting proteins. The interacting proteins and contaminants are further separated into different *protein classes* (see Fig. A2 and 3a). We include truly *top interacting proteins* that do not have any counts in the control experiments and show either (i) a low number of counts (*top1*) or (ii) a high number of counts (*top2*) across the bait experiments. Further, we have a more challenging class of truly interacting proteins which appear in the control samples, but have a stronger presence in the bait experiments (*sticky proteins*). We distinguish four different classes (*sticky 1-4*) with overall low or high number of counts and weak or strong presence in the bait samples. Moreover, four different classes of contaminants are introduced expressing various count levels.

Workflow based on SAINT and Westfall & Young

We investigate the performance of *SAINT-WY* in detecting the six different classes of truly interacting proteins which were introduced in the simulation data. These classes comprise different challenges concerning overall low and high counts in the samples with weak and strong presence in the bait replicates. As described in the manuscript, on average 47% of the truly interacting proteins were identified by *SAINT-WY* without preprocessing of the data, most of these proteins originate from the classes *top2* and *sticky2* as shown in Fig A11a, which share the characteristic of having very small counts in the control samples but a strong presence in the bait samples. The median detection rate of the remaining four classes is below 35%. These are more challenging to detect as they are either defined by a smaller increase of counts in the bait samples compared to the controls or show an overall number of high counts. Normalization of the data has an enormous impact on the detection rate of these individual protein classes. Application of the quantile normalization increases in particular the median detection rate of the classes *top1*, *sticky1* and *sticky4* by 30-40%. Further filtering only results in small improvements for the quantile normalization (see Fig. A11a), but shows greater impact on some protein classes in combination with other normalization methods (see Fig. A12-13).

Workflow based on TSPM and Westfall & Young

Here, we evaluate how the detection of the six individual classes of truly interacting proteins is affected by normalization and filtering when applying *TSPM-WY*. Fig. A11b visualizes the crucial application of the filtering step and reveals that approximately 30% of proteins of the classes *top2*, *sticky2* and *sticky4* are detected by the quantile normalization without filtering. These three classes are defined by a large difference of counts between bait and control. Additional filtering raises the median detection rate of true interactors to 80% and above in five of the protein classes (see Fig. A11b). The results vary dependent on the normalization method used: Application of the TMM normalization in combination with filtering yields even better results, while the sumtotal normalization shows difficulties in the identification of protein classes defined by a weaker presence in the bait samples (see Fig. A12-13).

Workflow based on TSPM and Benjamini-Hochberg

Finally, we evaluate the performance of *TSPM-BH* in detecting the six individual protein classes. As described in the manuscript, *TSPM-BH* identifies 75.5% of the truly interacting proteins without preprocessing; Fig. A11c reveals that these predominantly belong to the four protein classes which share the characteristic of low counts in the controls. The substantial impact of the quantile normalization is visualized; median detection rates for all protein classes are raised to 80% and above. Additional filtering further improves the detection rate of all classes to 90% and above.

ADDITIONAL EVALUATION OF RESULTS and DISCUSSION OF MERITS

Normalization methods can vary in performance depending on data characteristics: The sumtotal normalization needs to be used carefully as it is sensitive to single outliers in terms of high counts – they largely contribute to the total count of a sample and consequently result in the repression of all proteins in the sample. The quantile normalization alters the count distributions the most, but has proven a very good performance in microarray analysis and our results confirm an overall excellent performance in the analysis of AP-MS data. It is able to identify more truly interacting proteins in most analyses at the same false-positive rate than the other normalization methods. The methods upperquartile, TMM and DESeq are less strict in aligning the count distributions, showing an overall good performance with the TMM being superior. In case many zero counts dominate the data set, the upperquartile method is not appropriate as no normalization is conducted if the 75th percentile is zero.

As a second preprocessing step, we introduced a biological and statistical filtering of the data in order to remove obvious contaminants at an early stage and to reduce the multiple testing problem correspondingly. In the case of large and noisy data sets, in which a certain amount of noise is expected, filtering of the data becomes more important and enables a more sensitive detection of true interactions as our simulation study demonstrates. In contrast, the *Salmonella* data set is small and received high-quality measurements by the LTQ Orbitrap Elite mass spectrometer, hence filtering of the data is not crucial in this case and results in only minor improvements. Further, removal of outliers by the filtering can be essential, as we observed for the scoring method TSPM in combination with Westfall & Young. The main challenge in the application of the filtering is to define a reasonable cutoff – truly interacting proteins might be removed if the cutoff is set too high, while only a minor effect is obtained in case it is set too low. It is recommended to use available biological knowledge concerning the minimal number of expected true interactions; a parameter in the filtering step can be set accordingly (also refer to '*Guidelines for the cutoff choice'*).

After preprocessing of the data, we investigated the performance of two different scoring methods – SAINT and TSPM – to evaluate the interaction potential of a protein. We observe diverse features of the two proposed scoring methods, which may result in the preference of different proteins. In the *Salmonella* data study, some interaction candidates are exclusively detected by SAINT or TSPM respectively, showing a weak preference for TSPM. An additional investigation of different protein classes in the simulation study reveals that SAINT (coupled with WY) preferentially detects proteins with small counts in the controls, showing a large difference to counts in the baits. TSPM (coupled with BH) more strongly values small counts in the controls and is also more sensitive in detecting smaller differences between bait and control. In general, TSPM puts more weight on single high counts occurring in a bait sample than SAINT does. This may also become a pitfall in case of the permutation procedure, if an outlier (an extremely high count) in the controls turns into a bait sample by permutation. We observe this issue in *TSPM-WY*, which requires filtering, while *TSPM-BH* and *SAINT-WY* are not affected. A clear advantage of TSPM lies in the substantial reduction of runtime, corresponding to several minutes applying the permutation procedure with TSPM compared to a few hours with SAINT. The choice of either SAINT or TSPM (we showed strength and pitfalls of both methods) should depend on data characteristics and the experimental setup. We note that the choice of normalization and filtering is far more impactful than the choice of the scoring scheme.

For postprocessing, we aimed at replacing scores by p-values that can be interpreted in a statistical way and which allow the estimation of false positive interactions in a final list of candidate proteins. If the distribution of scores, given by any scoring scheme, is unknown, as for SAINT, p-values cannot directly be inferred. Therefore, we proposed a permutation procedure to estimate the empirical distribution and apply the integrative procedure of Westfall & Young to calculate p-values. This results in the generation of multiplicity adjusted p-values for all proteins which are controlled by the family-wise-error rate (FWER). Hence, selected proteins below a threshold of 0.05 refer to a list of true interaction candidates in which no false positives are expected with a probability of 95%. Considering the simulation results, SAINT scores of the selected proteins range from 0.5 to 1.0. This indicates how difficult it can be to set thresholds and that many truly interacting proteins may be missed by subjectively set thresholds. Thus, the proposed approach constitutes a robust criterion for generating a cutoff score in a list of interaction proteins produced by SAINT or any other scoring scheme.

The method of TSPM can be combined with two different postprocessing concepts. On the one hand, we can apply the permutation based procedure of Westfall & Young to the TSPM test statistics in order to allow comparisons to SAINT. On the other hand, p-values can be directly calculated from a χ^2-distribution without the need for permutation sampling. Thus we can choose a less conservative adjustment method for the latter to adjust the raw p-values, such as the Benjamini-Hochberg method. This approach can result in the detection of more potential interaction proteins below a threshold of 0.05, however more false positives might be included, but the expected number of false positives is limited to 5%. Hence, TSPM enables us to use a less conservative approach for detecting true interactions in AP-MS data by controlling false positives by a FDR.

B. Appendix

Fig. B1

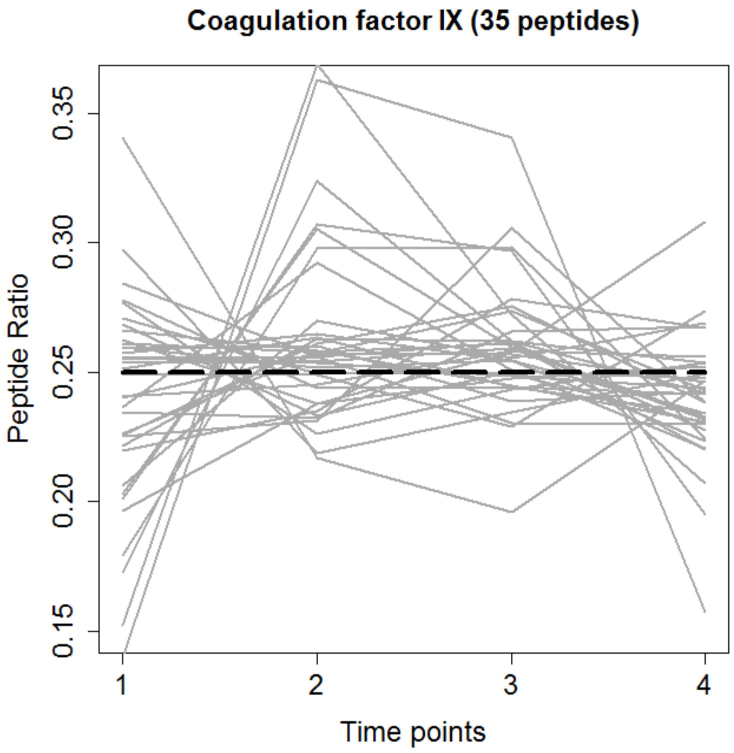
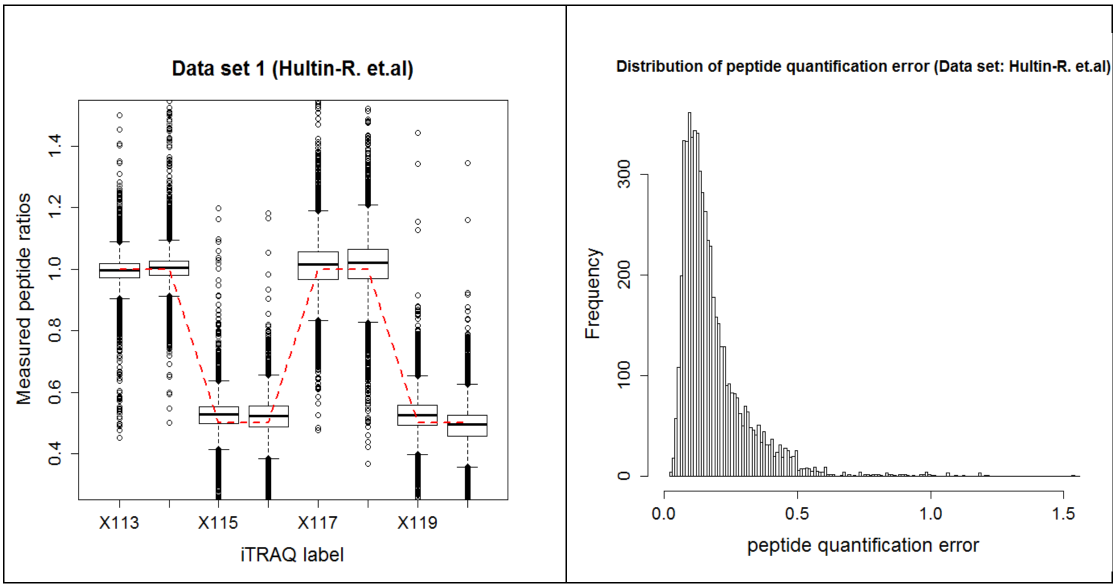


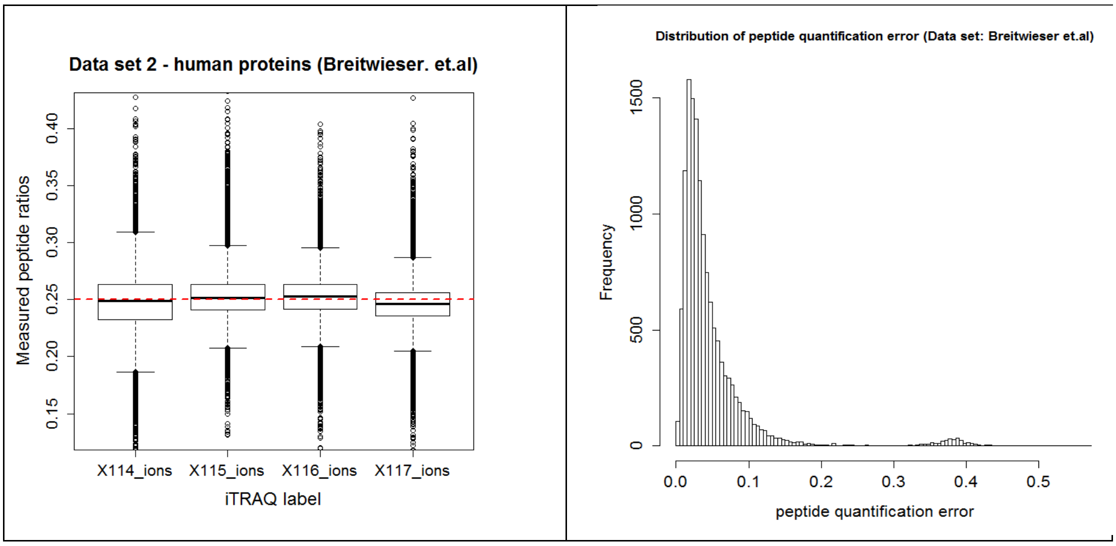
Fig. B 1: Peptide heterogeneity of an exemplary chosen protein, human coagulation factor IX (accession P00740), with 35 assigned peptide spectra. Every line corresponds to ratios of one peptide spectra. The expected ratios are marked by a dashed black line.

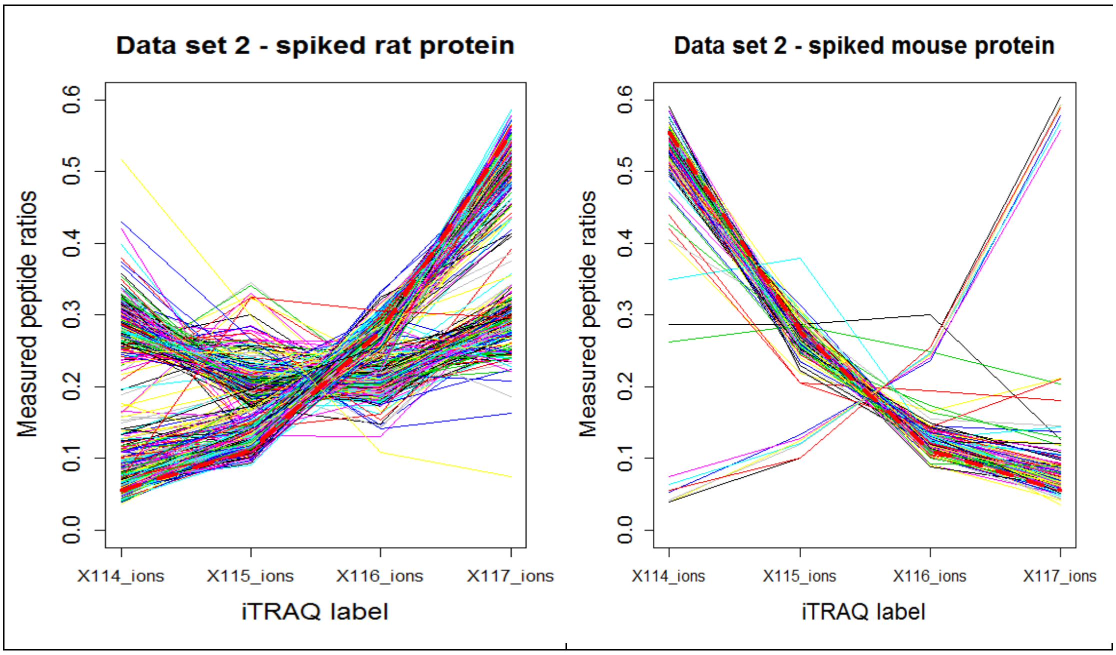
Fig. B2

a)

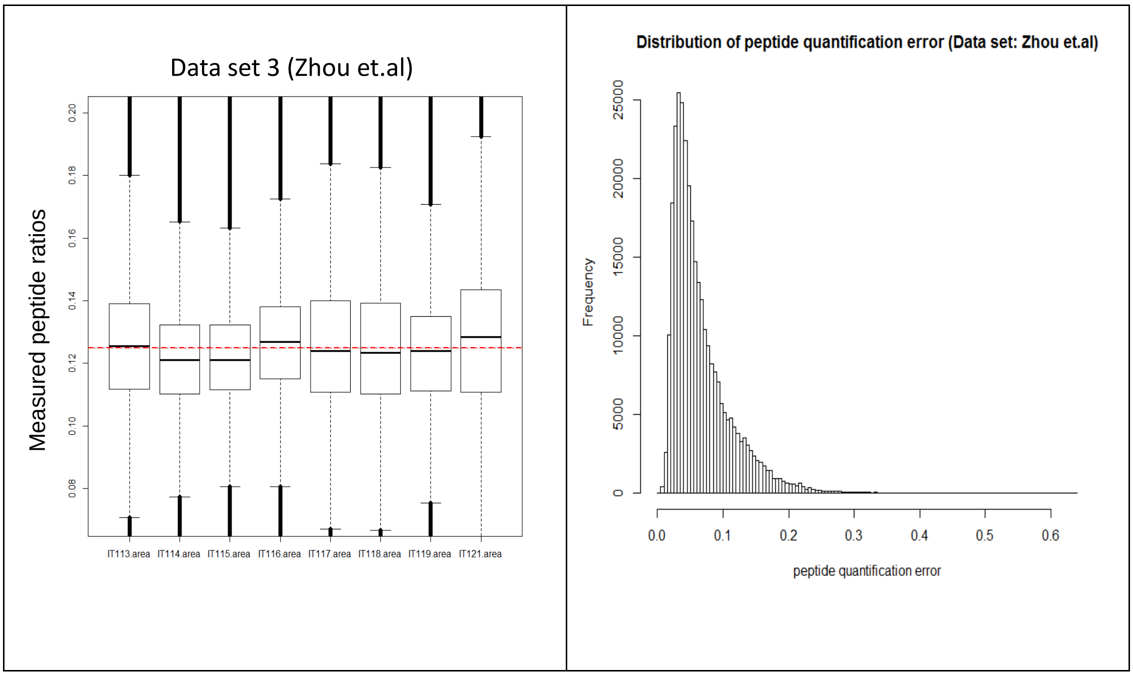

Data set 1 (Hultin-R. et.al)
Measured peptide ratios
0.4
0.6
0.8
1.0
1.2
1.4
X113
X115
X117
X119
iTRAQ label
Distribution of peptide quantification error (Data set: Hultin-R. et.al)
Frequency
0
100
200
300
0.0
0.5
1.0
1.5
peptide quantification error

b)

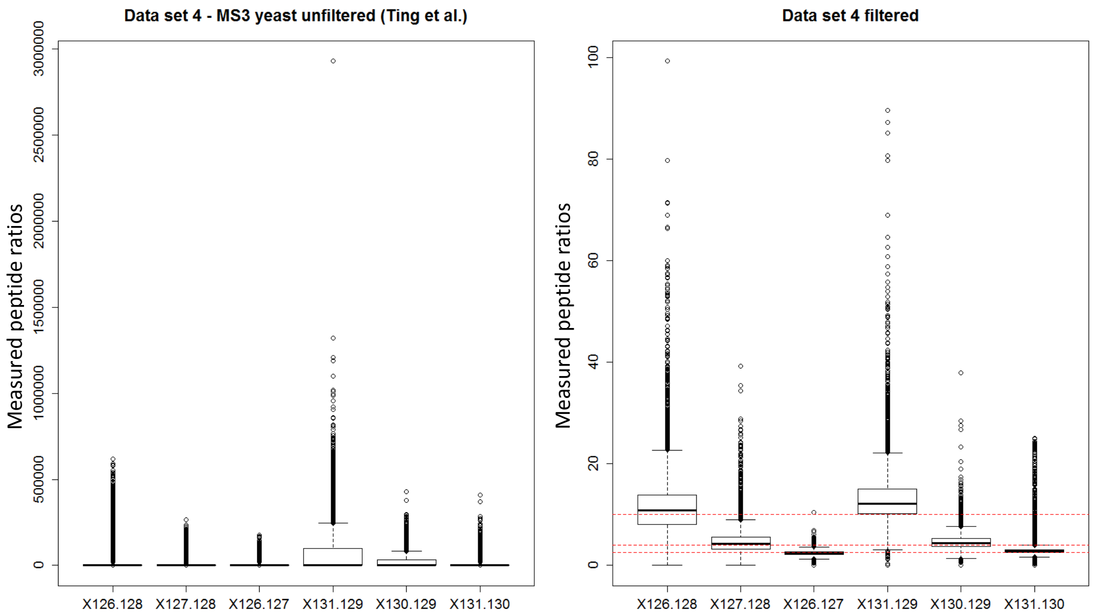

Data set 2 - human proteins (Breitwieser. et.al)
Measured peptide ratios
0.15
0.20
0.25
0.30
0.35
0.40
X114_ions
X115_ions
X116_ions
X117_ions
iTRAQ label
Distribution of peptide quantification error (Data set: Breitwieser et.al)
Frequency
0
500
1000
1500
0.0
0.1
0.2
0.3
0.4
0.5
peptide quantification error

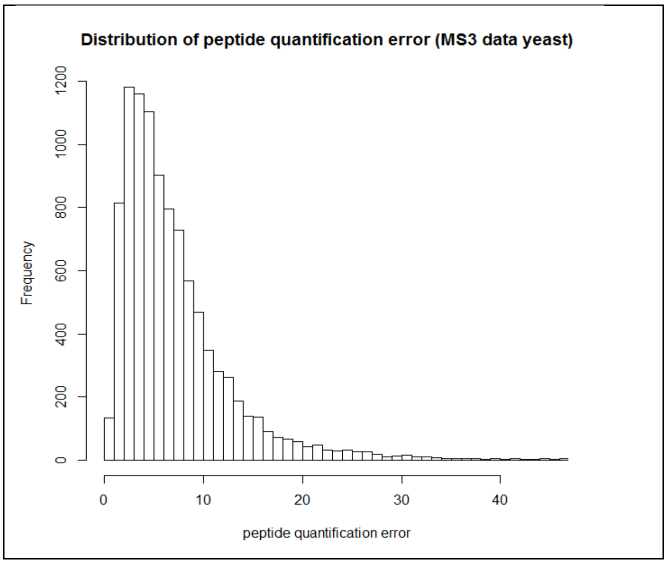

Data set 2 - spiked rat protein
Measured peptide ratios
0.0
0.1
0.2
0.3
0.4
0.5
0.6
X114_ions
X115_ions
X116_ions
X117_ions
iTRAQ label
Data set 2 - spiked mouse protein
Measured peptide ratios
0.0
0.1
0.2
0.3
0.4
0.5
0.6
X114_ions
X115_ions
X116_ions
X117_ions
iTRAQ label

c)


Data set 3 (Zhou et.al)
Measured peptide ratios
0.08
0.10
0.12
0.14
0.16
0.18
0.20
IT113.area
IT114.area
IT115.area
IT116.area
IT117.area
IT118.area
IT119.area
IT121.area
Distribution of peptide quantification error (Data set: Zhou et.al)
Frequency
0
5000
10000
15000
20000
25000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
peptide quantification error

d)


Data set 4 - MS3 yeast unfiltered (Ting et al.)
Measured peptide ratios
0
500000
1000000
1500000
2000000
2500000
3000000
X126.128
X127.128
X126.127
X131.129
X130.129
X131.130
Data set 4 filtered
Measured peptide ratios
0
20
40
60
80
100
X126.128
X127.128
X126.127
X131.129
X130.129
X131.130


Distribution of peptide quantification error (MS3 data yeast)
Frequency
0
200
400
600
800
1000
1200
0
10
20
30
40
peptide quantification error

e)

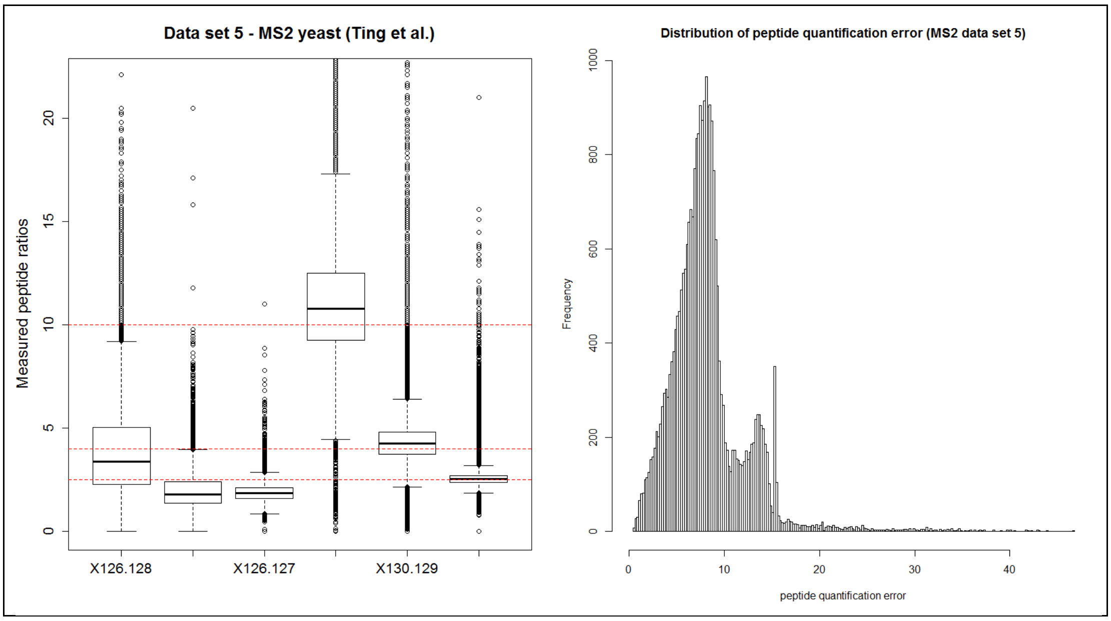


Fig. B 2 a-e: Distributions of measured peptide spectra ratios per label (left) and distribution of the corresponding peptide spectra quantification error (right), shown for all five data sets (a-e). Boxplots represent a summary of the measured peptide ratios per label and the dashed red line refers to the expected ratios (ground truth) for the corresponding data set. Data set 1 and data set 3 refer to 8-plex iTRAQ experiments, data set 2 is based on a 4-plex iTRAQ experiment, and data set 4-5 in form of MS2 and MS3 data present 6-plex TMT experiments.

(A) In data set 1, all peptide spectra are assumed to follow the same ground truth of 1:1:0.5:0.5:1:1:0.5:0.5. **(B)** Data set 2 consists of 143 human proteins with predefined constant ratios and two spike-in proteins from mouse and rat with expected decreasing and increasing ratios respectively; peptide ratios from the human proteins are displayed by boxplots, while peptide ratios from the two spike-in proteins are plotted separately showing the individual peptide ratio trend. In case of the spike-in rat protein a group of strongly diverging peptide ratios from the ground truth is observed. **(C)** In data set 3, all peptide spectra are expected to follow a predefined constant ground truth. **(D)** In the MS3 data set 4 all yeast peptide spectra are mixed according to ratios of 10:4:2.5:10:4:2.5. Original ratio values of the data set exhibit extreme outlier ratios. Hence the application of a filtering step is reasonable as fold changes greater than 100 are not expected due to biological source. In the filtering step spectra deviating more than ten-fold from the expected ground truth are discarded. **(E)** In data set 5 containing the MS2 spectra, the peptide interference effect in the first three labels can be clearly observed (an excerpt of the y-axis is shown for better visualization). Ratios of the yeast peptides are significantly compressed by the co-elution of human peptides and result in deviation from the ground truth as already reported in the publication by Ting et al. A filtering step removing extreme outlying ratios greater than 200 was also applied here and as a result the data set comprises 1482 proteins based on 26915 peptide spectra.

Overall in (a)-(d), the observed ratio values spread around the ground truth values of the corresponding data sets, the dashed line always lies within the box.

Considering the peptide spectra quantification error for each data set (right figure), defined by the Euclidian distance of the measured ratios to the expected ratios across all iTRAQ/TMT labels, a right skewed distribution is observed in (a) –(d). In case of data set 2, a second peak in the quantification error distribution is apparent due to strongly divergent spectra found in the spike-in proteins. Data set 3 indicates more accurate peptide ratio measurements in comparison to data set 1 shown by smaller peptide quantification errors; data set 2 is a 4-plex experiment and therefore not directly comparable to data set 1 and 3. Data set 4 and 5 show high quantification errors due to high ratio variation and the distribution of the MS2 data set is not right-skewed due to the bias of co-eluting peptides.

Correlation of Peptide Spectra Features to Quantification Error

Fig. B3

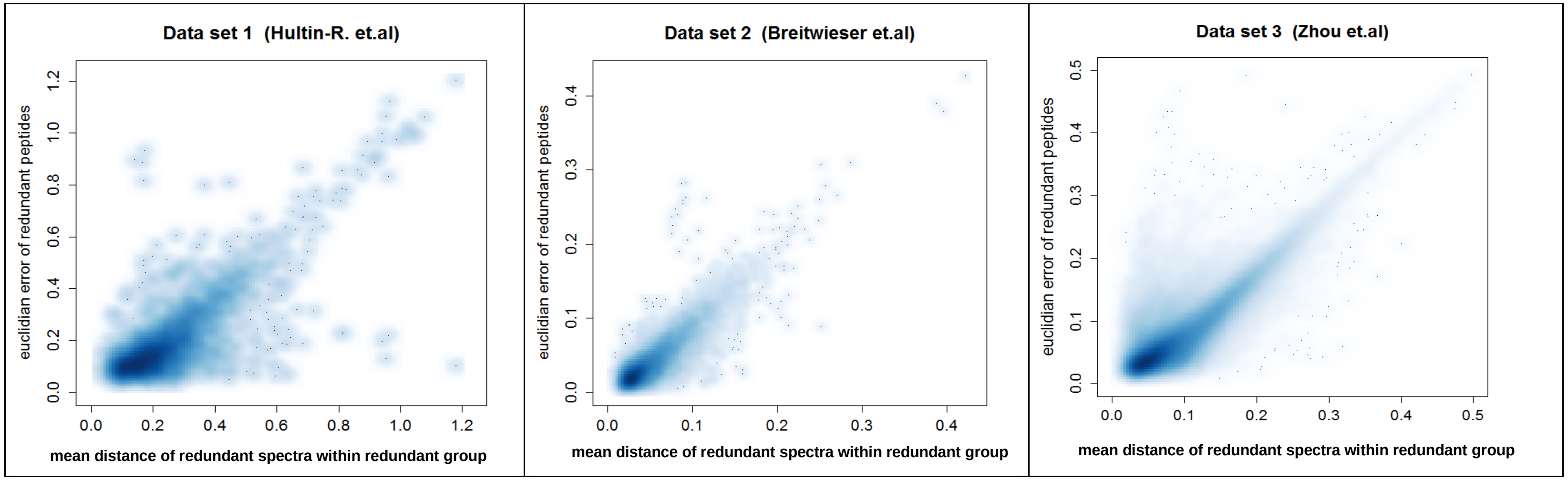


Fig. B 3: **Distance Metric of redundant peptide spectra correlated with quantification error.** The ratio similarity of redundant peptide spectra is specified by the mean distance of a peptide spectrum to all other peptide spectra belonging to the same redundant spectra group. All three data sets clearly reveal that the quantification error increases with a redundant peptide spectrum quantitatively diverging from its group of redundant spectra**.** Different distance and error scales are observed here dependent on the overall data accuracy of the data sets. Total numbers of redundant peptide spectra considered are 4161, 13311 and 211268 peptide for data set 1, 2 and 3 respectively.

Fig. B4

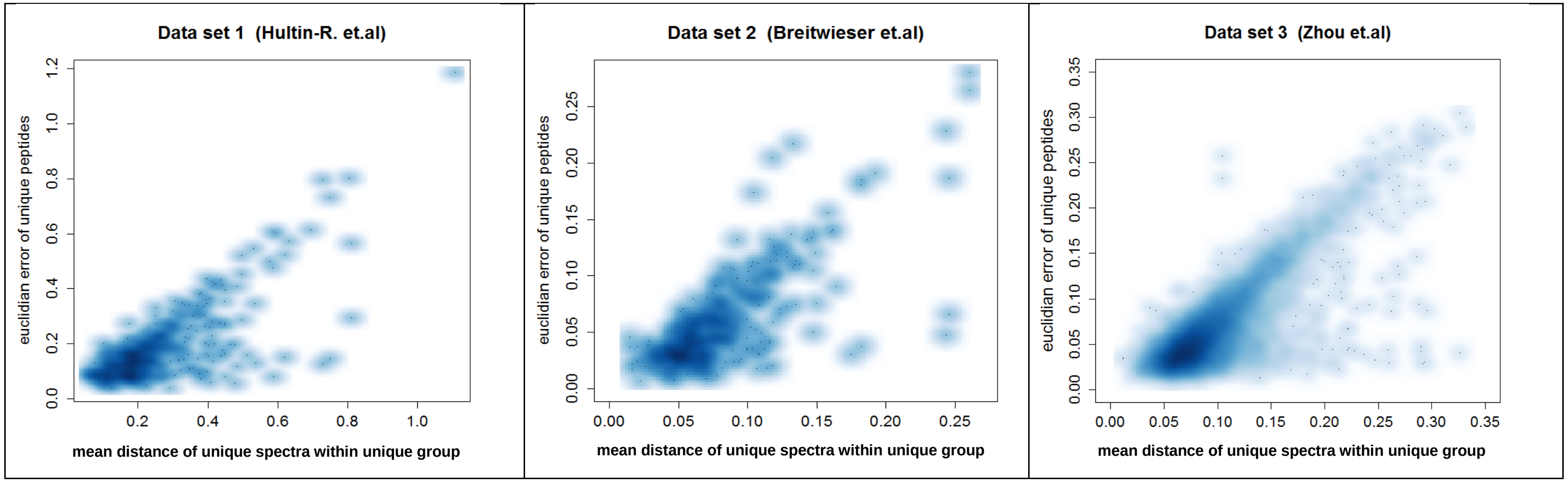


Fig. B 4: **Distance Metric of uniquely measured peptide spectra correlated with quantification error.** *Unique* spectra are referred to peptides quantified exactly by one MS/MS event. All unique spectra measured for a specific protein are pooled into one group per protein. The same similarity metric is applied to the unique group as for the redundant group in order to detect uniquely measured diverging spectra. Equivalent correlations are observed corresponding to an increasing quantification error with increased distance of a uniquely measured peptide spectrum to other spectra of the group. The number of uniquely measured spectra in the data sets is smaller referring to 628, 413 and 6554 spectra for data set 1, 2 and 3 respectively.

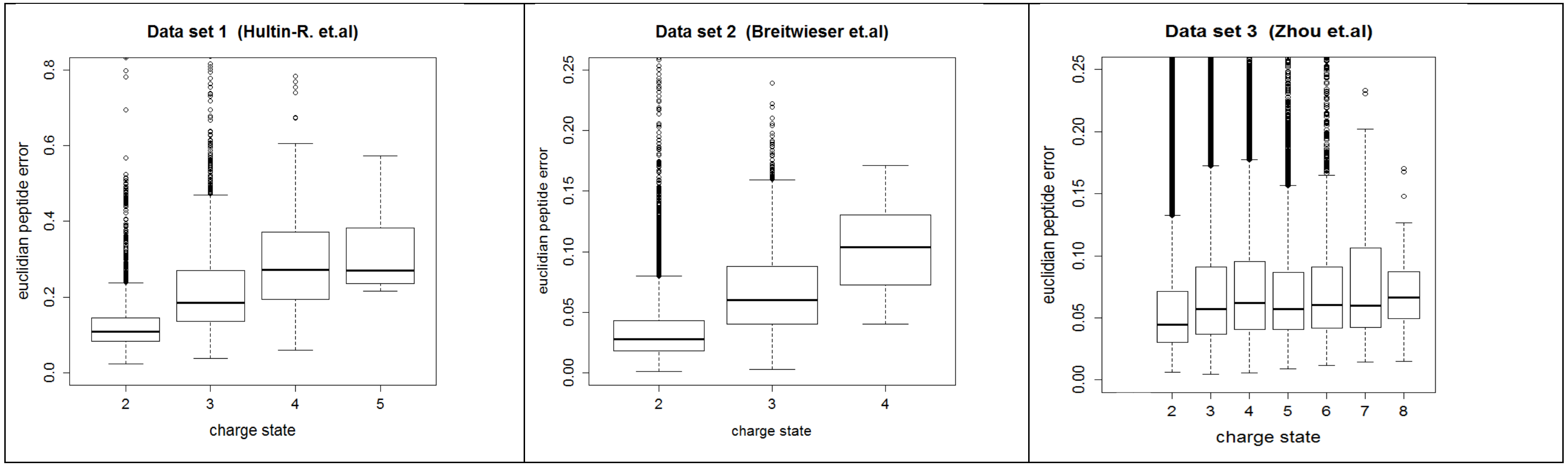


Fig. B 5: **Charge state of a peptide correlated with quantification error**. Peptide spectra with higher charge state are accompanied with increased quantification error. This is particularly evident for the most common charge states of 2, 3 and 4. Higher charge states shown for dataset 3 refer to less than 1% of the spectra, thus are less representative.

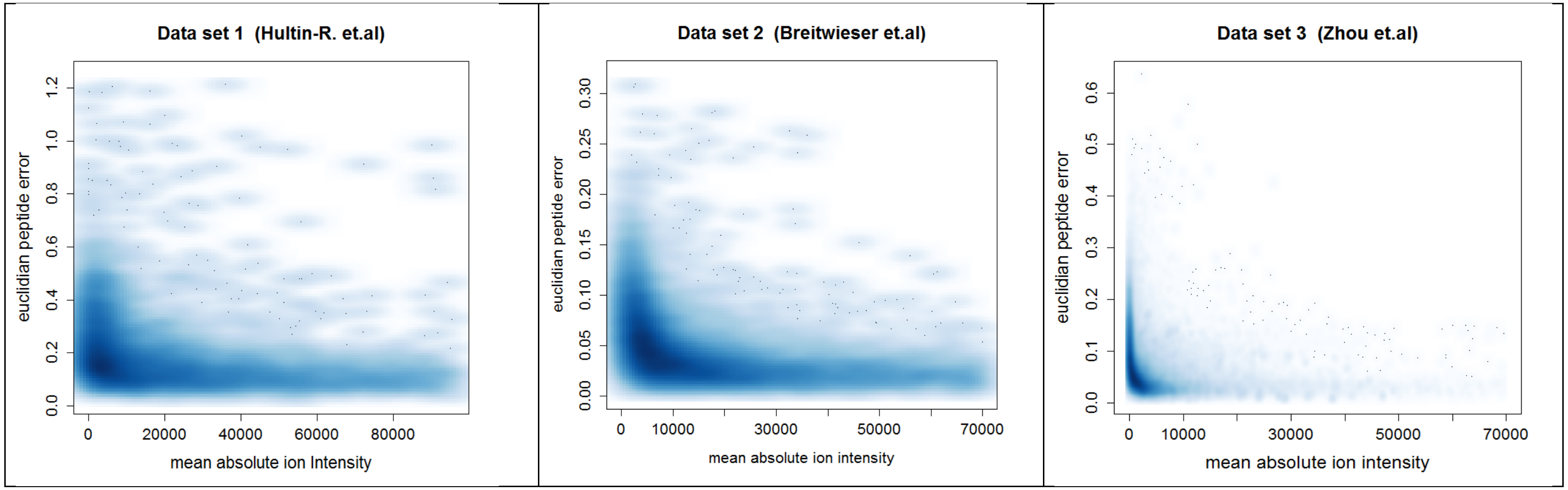


Fig. B 6: **Mean absolute ion intensity of a peptide correlated with quantification error**. This study also confirms the increase of noise and error in low absolute ion intensity data and improved quantification accuracy in high intensity data. An excerpt of the intensity range is shown from zero to approximately ~80000 for visualization reasons; the trend is sustained for higher intensities. The error scales of the different data sets also reflect the varying noise impact on low intensity data, being strongest in data set 1.

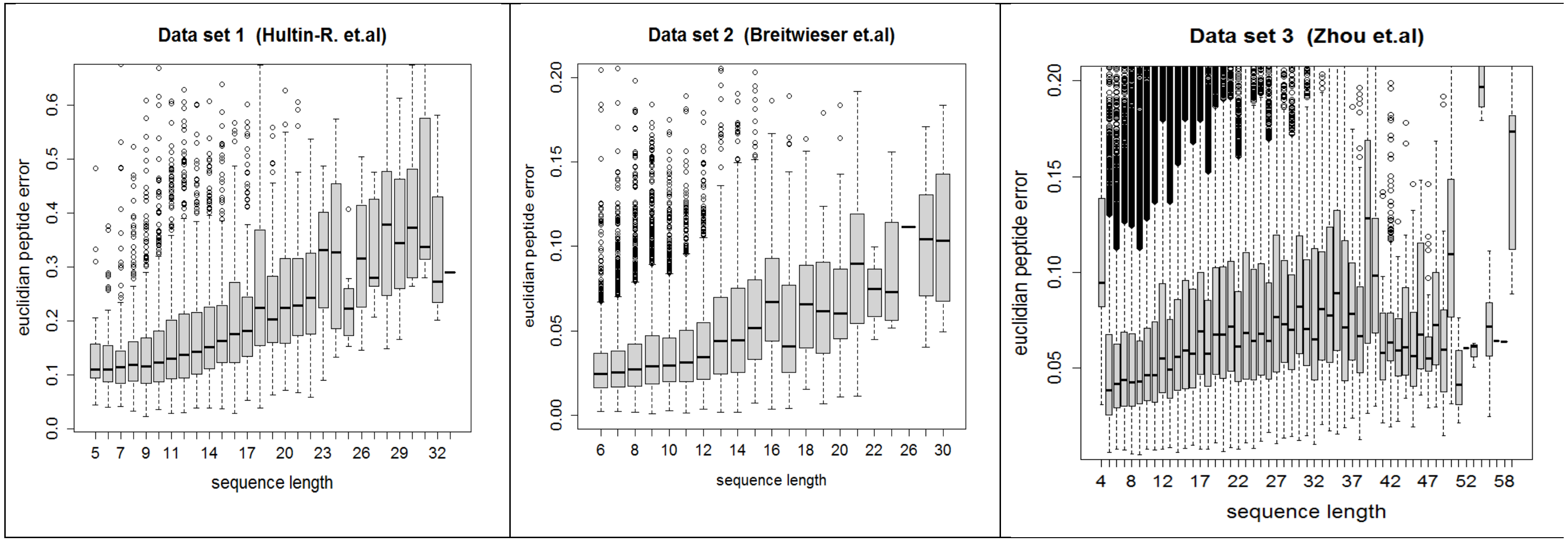


Fig. B 7: **Peptide sequence length correlated with quantification error.** Increased sequence length is shown to come along with higher quantification error. Sequence length present in the three data sets mainly spans from 5 to approximately 30 amino acids. Data set 3 exhibits only a few larger peptide sequences in addition.

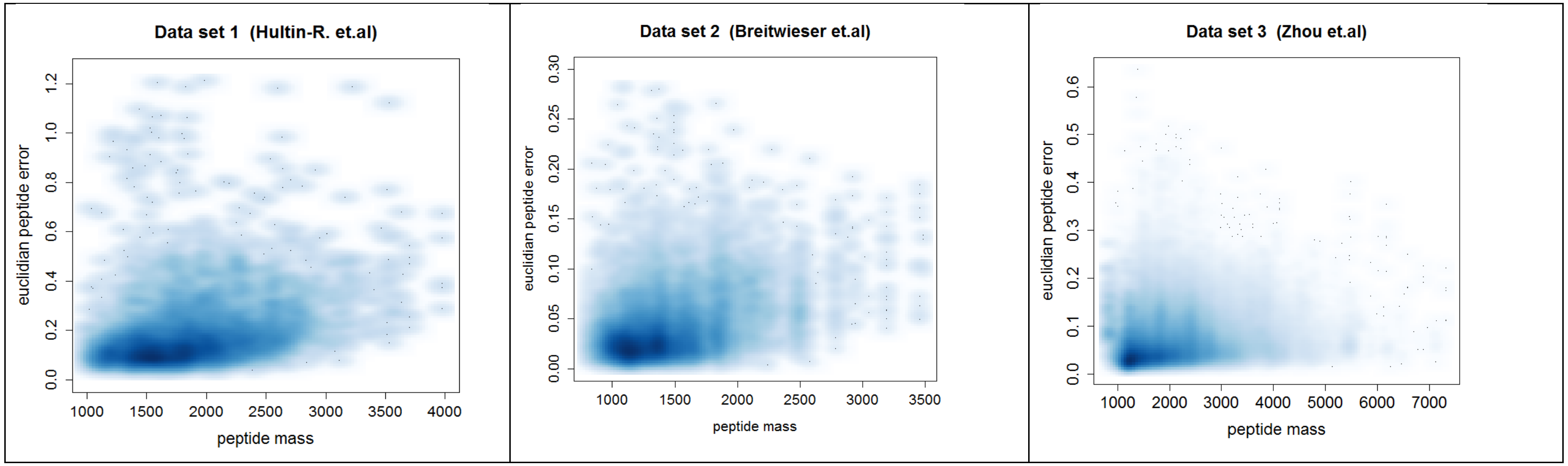


Fig. B 8: **Peptide mass correlated to quantification error.** The observed trend is equivalent to the one observed for the sequence length due to the strong inter-correlation of these features (see Fig. B14). Quantification error increases with peptide mass increasing for all three data sets.

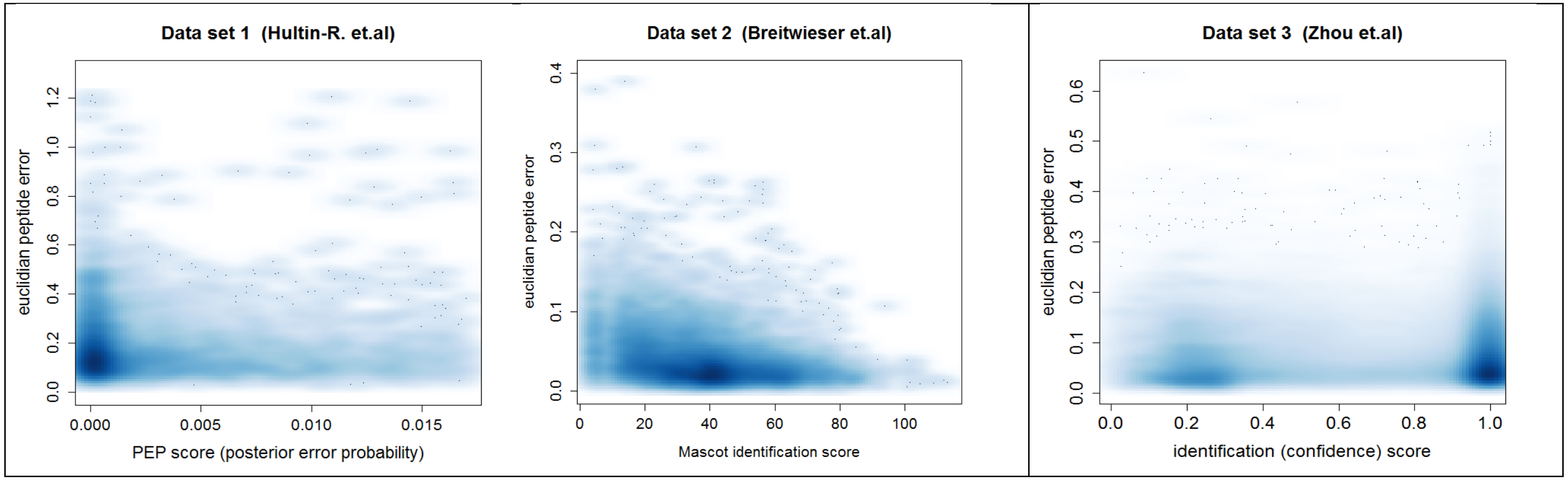


Fig. B 9: **Identification score and quantification error.** Correlation of the identification score varies between the three data sets due to the different scoring systems. In data set 1, the majority of spectra receive a small PEP score referring to highly reliable identifications, however a large error variation is observed at the same time. In data set 2 and 3 higher scores generally correspond to increased identification reliability, thus expecting small quantification errors. In case of data set 2 the negative correlation can be observed, yet the majority of peptide spectra hold a Mascot score around 40. In case of data set 3, the majority of spectra exhibit a score of 1 accompanied by a large error variation again and interestingly a second cluster is found around a score of 0.2.

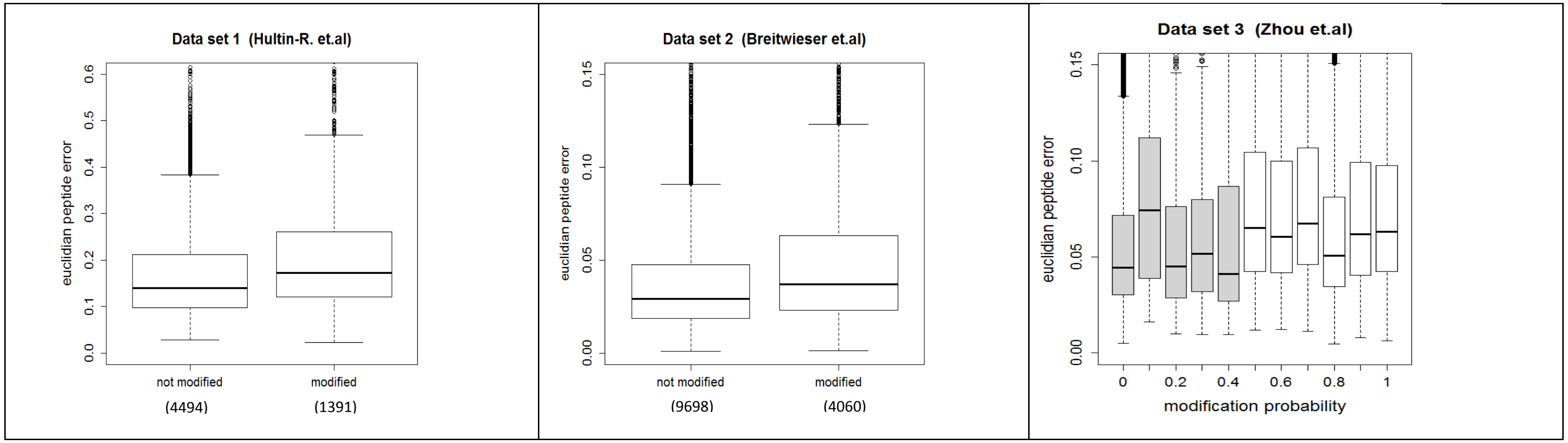


Fig. B 10: **Modification and quantification error.** An increased quantification error is observed in the smaller group of modified peptide spectra compared to the group of non-modified peptide spectra in data set 1 and 2. For data set 3 a modification probability is provided, displaying a significant shift of quantification errors for modification probabilities below 0.5 and above 0.5.

MS3 Data set (Ting et al.) - Peptide Feature – Error – Correlations

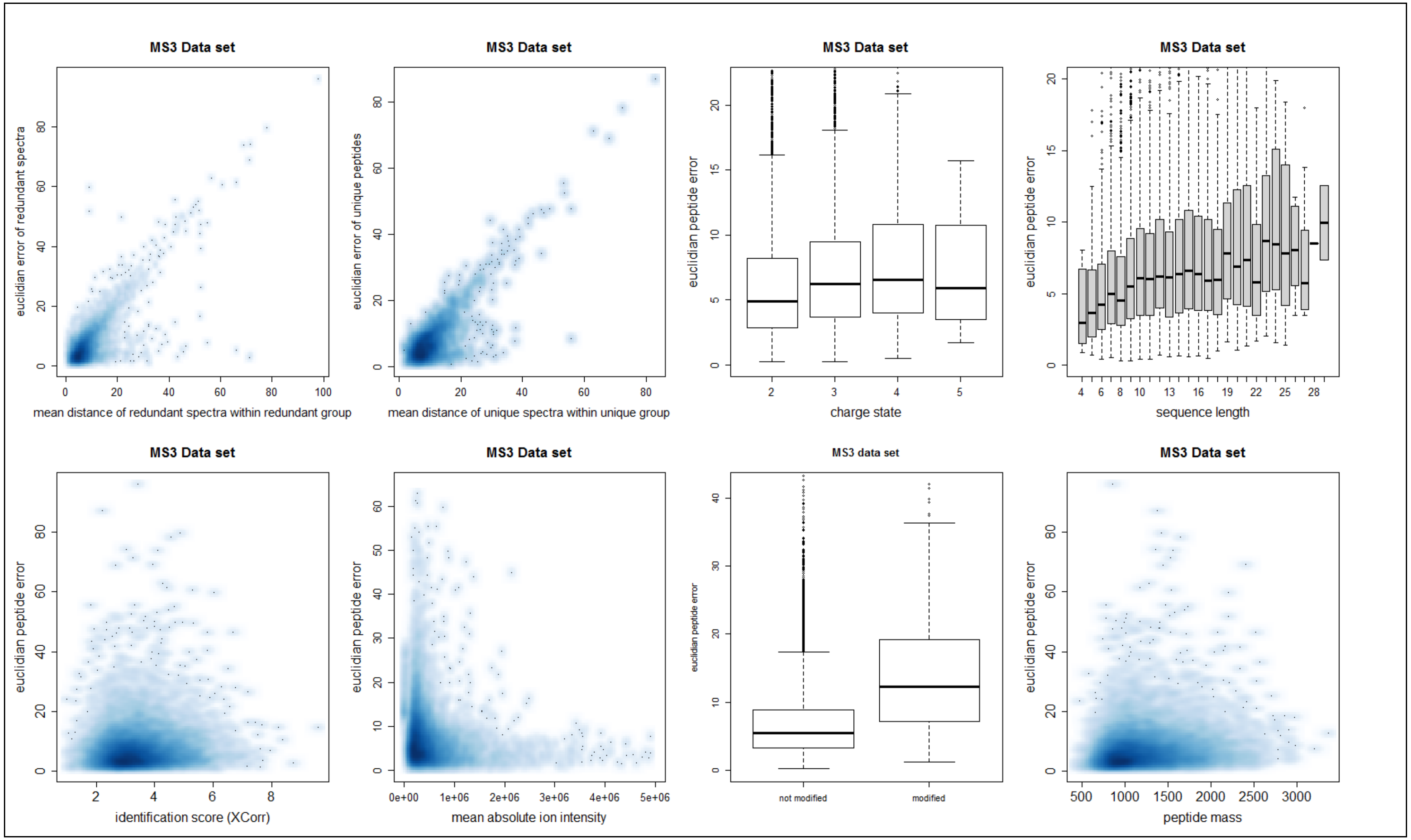


Fig. B 11: **Peptide feature -error-correlations on the basis of an MS3 data set.** Similar correlation trends as shown for the MS2 data sets are also observed for all different features using the technique of triple-stage mass spectrometry (MS3). This confirms the potential applicability of the iPQF algorithm in MS2 as well as MS3 data.

MS2 Data set (Ting et al.) - **Impact of peptide interference** on Peptide Feature – Error – Correlations

Peptide Features	All channels (126, 127, 128, 129, 130, 131)	Channels without interference (129, 130, 131)	Channels with interference (126, 127, 128)
Redundancy metric	0.355	**0.563**	0.059
Uniquely measured metric	0.416	**0.495**	0.092
Charge state	0.334	**0.405**	0.156
Ion intensity	-0.424	**-0.58**	-0.066
Sequence length	0.211	**0.384**	-0.013
Mass	0.193	**0.374**	-0.029
Identification score	-0.22	-0.07	**-0.27**
Modification	0.239	**0.253**	0.088

Table B 1: Correlation coefficients of peptide spectra features to relative quantification error observed for the MS2 data set by Ting et al. in different channels. Peptide interference is expected in the first three channels (126, 127, 128) and no interference is assumed in the last three channels (129, 130, 131) according to the design of the experiment (see also Fig. B 2e for ratio compression effect due to interference). As a result peptide -feature- error correlations are observed to be strongest in channels without interference as the quantification errors are not biased due to compressed ratios. Coefficients calculated on the basis of all channels show slightly decreased values, however are still in concordance with the reported correlation trend of the other data sets investigated (see Table 3.1 in manuscript). Channels underlying peptide interference cause high quantification errors and hence distort feature-error-correlations.

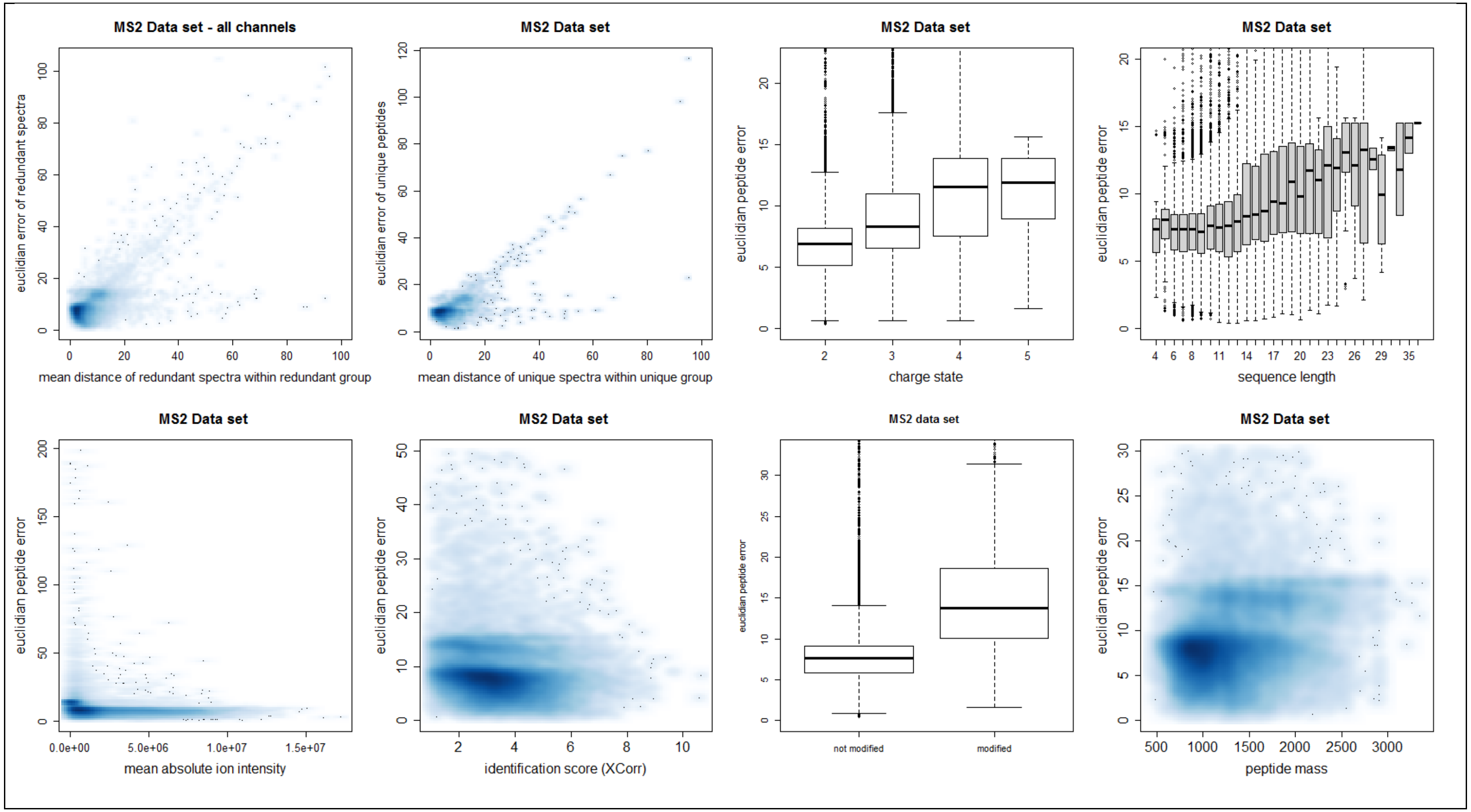


Fig. B 12: **Peptide feature -error-correlations on the basis of all channels of the MS2 data set by Ting et al.** An excerpt of the y-axis is chosen for the different plots for improved visualization of the correlations; outlying ratios causing high quantification errors are not shown. Overall, similar correlation trends, as shown for all other data sets, are observed. Although peptide interference in the first three channels causes reduced correlation, the overall correlation considering all channels still reflects significant associations between features and quantification errors.

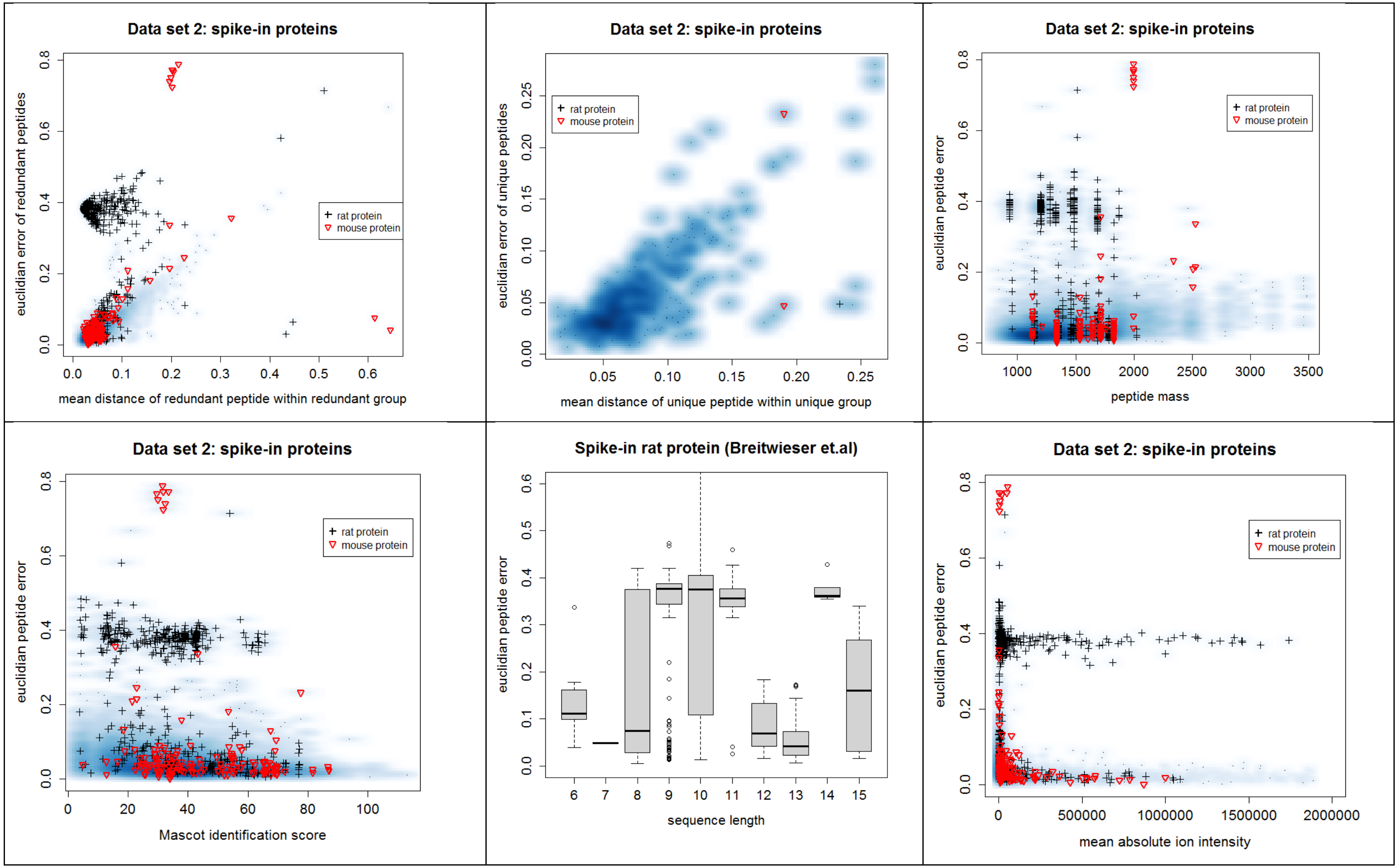


Fig. B 13: **Spike-in proteins of data set 2 (Breitwieser et.al) - Correlation of peptide features to quantification error** is presented for peptide spectra belonging to the rat and mouse spike-in proteins. Blue dots represent the underlying human data for comparison. As described in Fig B 2B), the rat protein has a highly divergent spectra group causing high quantification errors, reflected in the plots (black cross). In particular, short peptide sequences are assigned to the rat protein and the observed outlier peptide spectra group consists exclusively of redundantly measured spectra with high distance similarity. Further rat peptide spectrum matches hold predominantly low absolute intensities and low Mascot scores. The observed spectra outlier group could refer to a specific contaminant peptide with a high selection rate.

Inter-Correlation Study of Peptide Features (Data set 1)

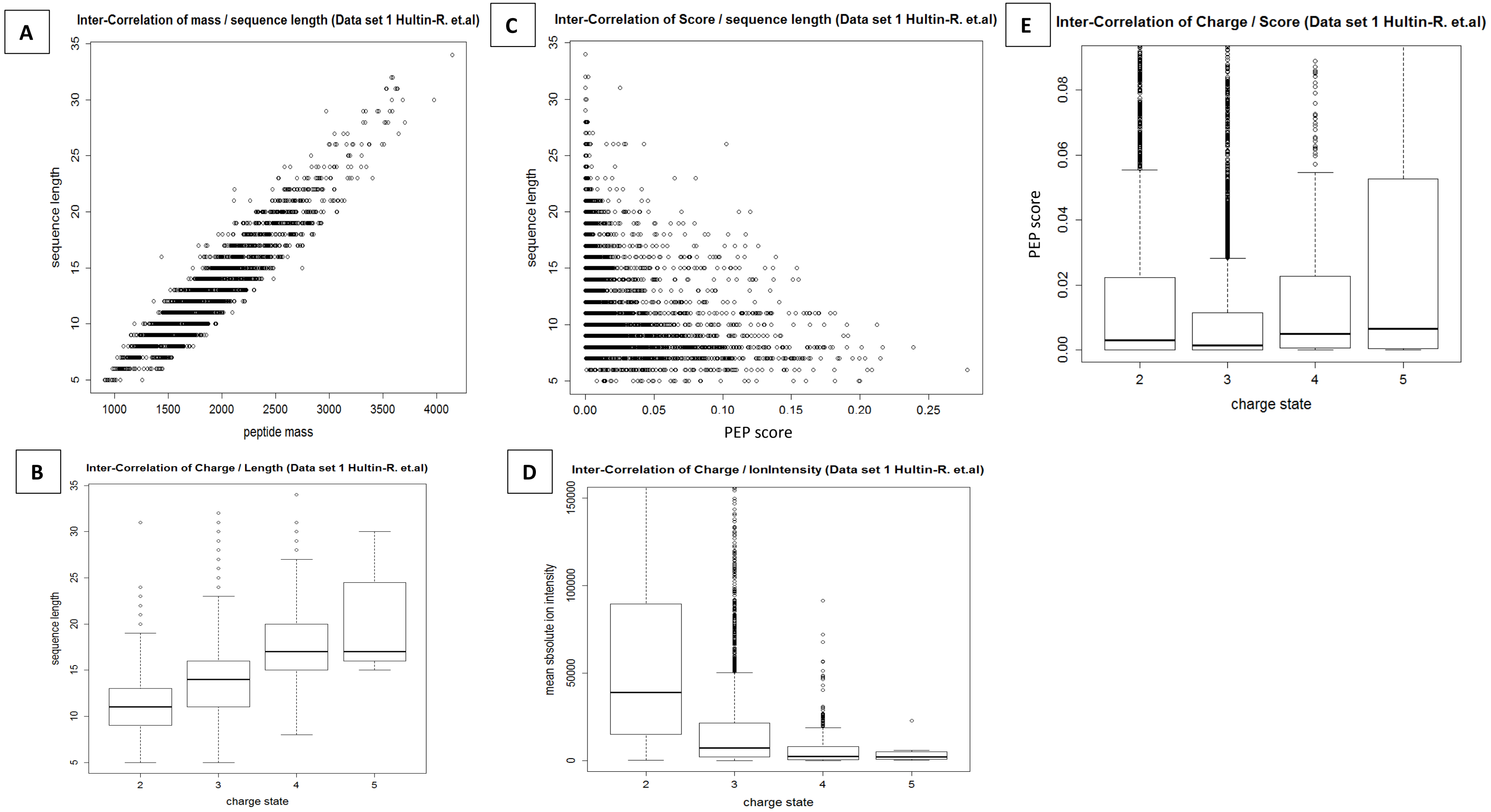
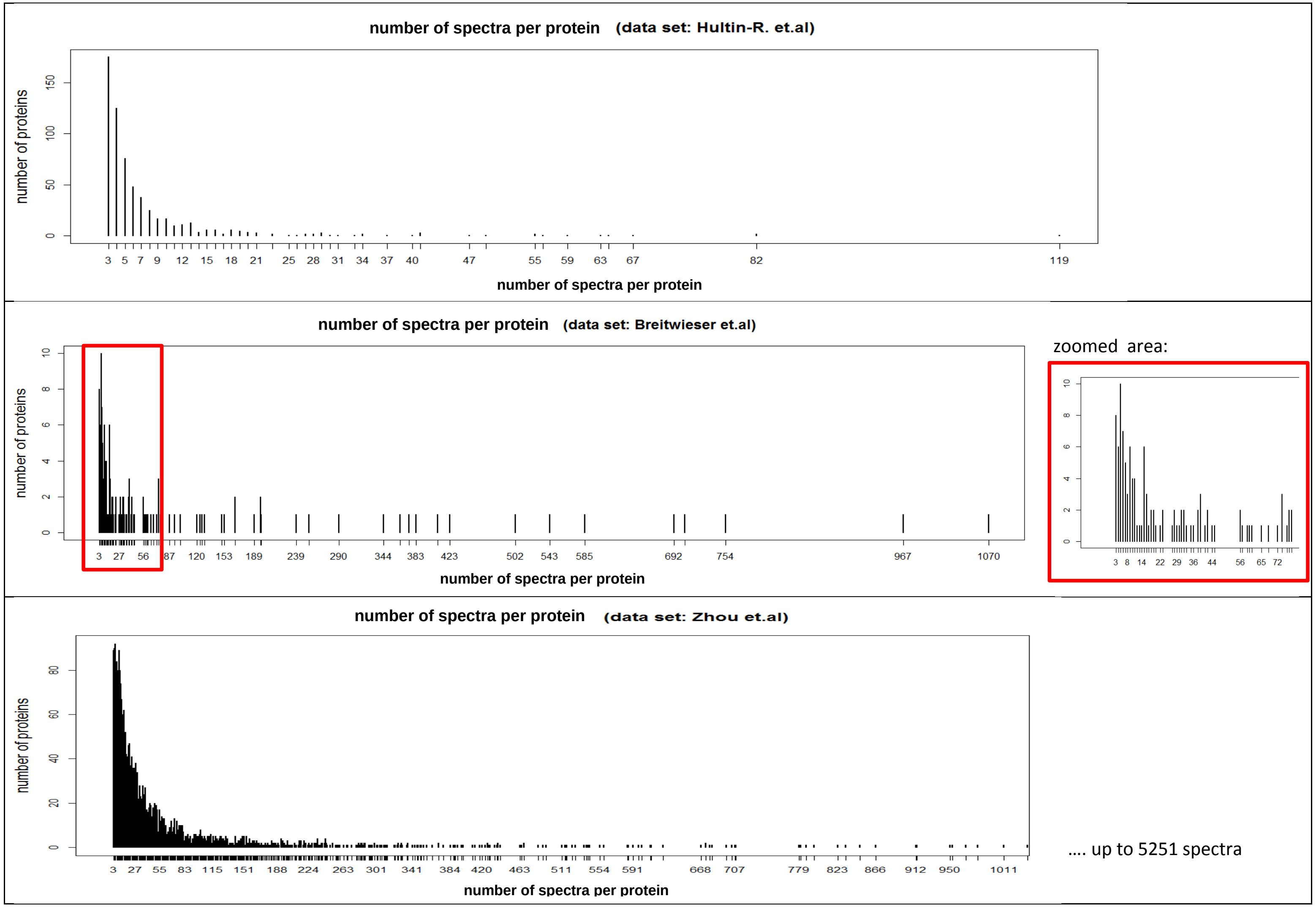


Fig. B 14: **Inter-correlation study of peptide features (shown for data set 1 (Hultin-R. et.al).** A) A strong correlation is displayed between peptide mass and sequence length as expected. B) Correspondingly, the number of possible charges is increasing with the sequence length increasing. C) As already proposed in other studies, we also show that long sequence peptides tend to have higher identification scores, here referring to a small PEP score in case of data set 1. D) Further, we observe low absolute ion intensities with increased charge state. E) Interestingly the range of identification scores assigned to peptide spectra with two, three or four charges stays similar, though we have seen an increased quantification error with increased charge state in Fig. B 5. C) and E) are examples that the combination of features, here score, charge state, and length, is crucial to improve protein ratio accuracy.


number of spectra per protein (data set: Hultin-R. et.al)
number of proteins
0
50
100
150
3 5 7 9 12 15 18 21 25 28 31 34 37 40 47 55 59 63 67 82 119
number of spectra per protein
number of spectra per protein (data set: Breitwieser et.al)
number of proteins
0
2
4
6
8
10
3 27 56 87 120 153 189 239 290 344 383 423 502 543 585 692 754 967 1070
number of spectra per protein
zoomed area:
3 8 14 22 29 36 44 56 65 72
number of spectra per protein (data set: Zhou et.al)
number of proteins
0
20
40
60
80
3 27 55 83 115 151 188 224 263 301 341 384 420 463 511 554 591 668 707 779 823 866 912 950 1011
number of spectra per protein
.... up to 5251 spectra

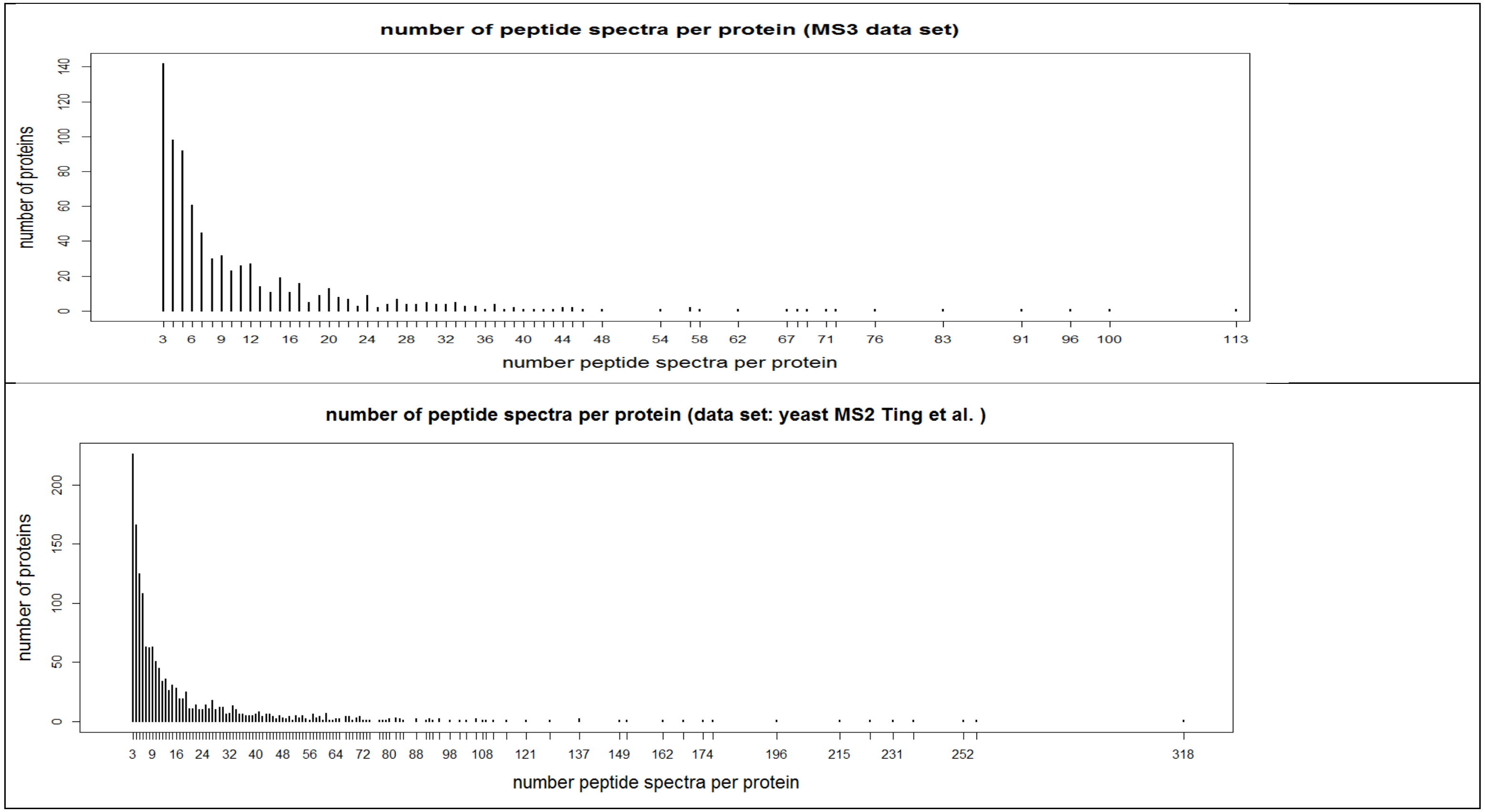


Fig. B 15: **Distribution showing the number of peptide spectra per protein for each of the five data sets**. Data set 1 consists of 624 proteins and is dominated by three to ten peptide spectra per protein. Data set2 comprises only 145 proteins, which are mainly supported by three to eighty peptide spectra. Data set 3 is an overall large data set with 2811 proteins, showing many cases of several hundred peptide spectra per protein (an excerpt of the distribution is shown here for visualization reason, the maximum lies at 5251 peptide spectra). Data set 4 (MS3) consists of 781 proteins with predominantly three to twenty spectra per protein and the corresponding data set 5 (MS2) consists of 1482 proteins holding predominantly three to forty spectra. Proteins based on less than three peptide spectra were not considered for iPQF quantification and evaluation, hence are not shown here.

Performance evaluation of iPQF approaches on data influenced by co-eluting peptides

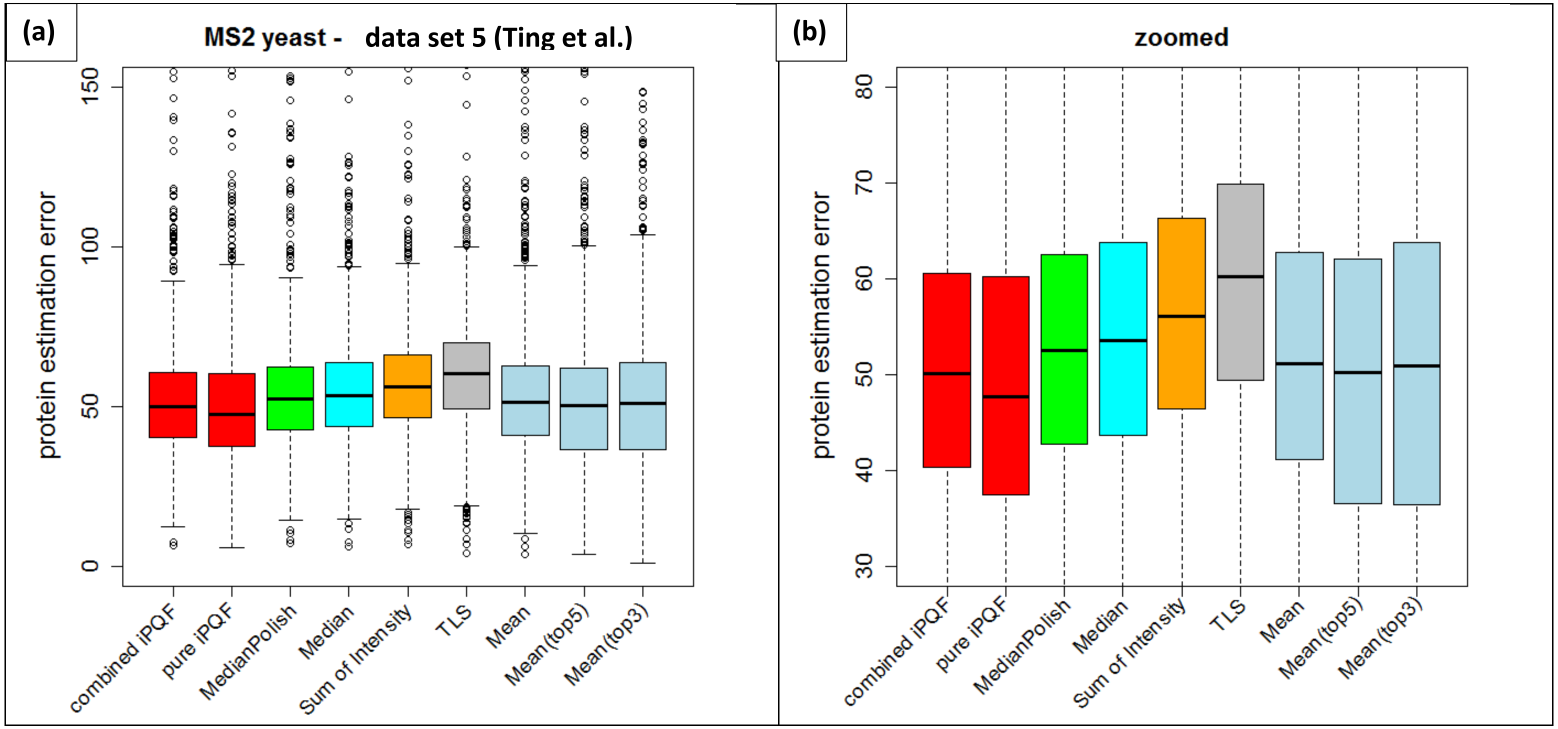


Fig. B 16: **Performance evaluation of iPQF approaches and seven summarization strategies on a data set affected by peptide interference. (a) Boxplots displaying the distribution of the protein estimation error, (b) Zoomed y-axis for detailed comparison.**

The MS2 yeast data set by Ting et al. exhibits substantial peptide interference in the first three channels (126, 127, 128) and no interference in the last three channels (129, 130, 131) according to the design of the experiment. Hence, the overall protein estimation error is high due to strong divergence of peptide ratios in the first three channels. However in comparison to other summarization methods, iPQF approaches show improved performance and confirm the benefit of integrating feature information. Mean based approaches are strongly influenced by extreme outlier ratios and benefit from this here. Note that isobar method is missing here as it could not be run on data set 5.

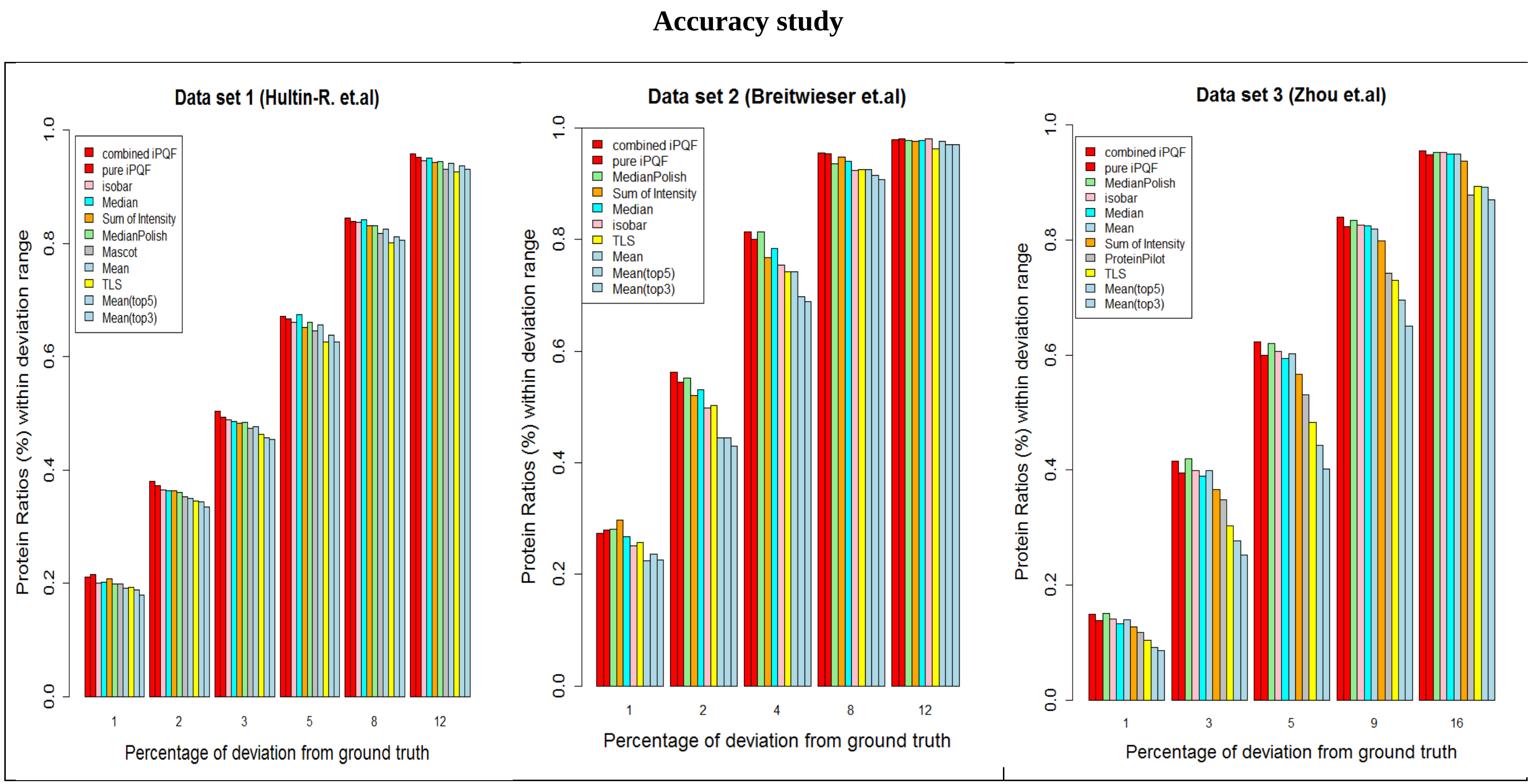


Fig. B 17: **Accuracy study of the different summarization methods, shown for each of three MS2 data sets.** For selected percentages of deviation from the ground truth, we assess the percentage of protein ratios which are estimated within the corresponding deviation range. This is shown for the different summarization methods applied. The combined iPQF shows slightly more ratio estimates within small ranges of deviation than other summarization methods. The pure iPQF even achieves more estimates in the low deviation range in data set 1. In particular, iPQF approaches prove performance robustness across the three different data sets, while other methods vary in performance. With the percentage of deviation increasing, all methods tend to perform similar and acquire most protein ratios.

Table B 2

	Data set 1 *(Hultin-R. et al.)*	Data set 2 *(Breitwieser et.al)*	Data set 3 *(Zhou et.al)*
combined iPQF	**0.6899**	**0.7327**	**0.7192**
pure iPQF	**0.6850**	**0.7246**	0.7036
MedianPolish	0.6761	0.7245	**0.7166**
isobar	0.6796	0.6867	0.7075
Median	0.6836	0.7130	0.7012
Sum of Intensities	0.6763	0.7104	0.6801
Total Least Squares (TLS)	0.6521	0.6807	0.6157
ProteinPilot.P	-	-	0.6345
Mascot	0.6646	-	-
Mean	0.6704	0.6695	0.7028
Mean (Top5)	0.6583	0.6508	0.5895
Mean (Top3)	0.6494	0.6401	0.5547

Table B 2: AUC measure (Area under the curve) of accuracy for the different summarization methods, assessed for each of the three MS2 data sets. For a series of deviation percentages from the ground truth, the percentage of protein ratios, which were estimated within a specific deviation range, is assessed by each method. The considered percentage of deviation is defined as series from zero to a cutoff, at which 95% of the protein ratios are covered by the first summarization method. This cutoff refers to a deviation percentage of 11.5 in data set 1 (Hultin-R. et.al), 7.8 in data set 2 (Breitwieser et.al) and 15.3 in data set 3 (Zhou et.al). AUC values are normalized by the maximum area value (cutoff * 0.95). The *combined iPQF* shows the highest AUC values referring to a higher precision in protein ratio estimation, covering 95% of the data within the smallest deviation range.

iPQF example workflow

Example Protein: MISP - Mitotic spindle positioning protein (IPI:IPI00217121.1), 7 peptide spectra assigned (from data set Hultin-Rosenberg *et al.*)

(1) Feature Assessment

Spectra ID	Features	Distance metric (unique, redundant spectra	Chargestate	Ion intensity	Sequence length	Identification score	Modifica-tion	Mass
6		0.1395713	2	100545.703	9	0.00956463	0	1230.683
7		0.1793798	2	16370.6901	9	0.01013137	0	1230.683
8		0.2223448	3	939.6888	9	0.09520685	0	1231.691
1185		0.1388411	2	300702.2851	8	0.05208204	0	1293.716
1186		0.1388411	3	8600.0059	8	0.0197555	0	1294.723
2817		0.2155298	3	49997.7403	9	0.00176861	0	1368.774
2818		0.2155298	2	627215.9769	9	0.04728342	0	1367.768

(2) Spectra ranks per feature (based on knowledge acquired in feature-error-correlation study: high rank ~ more reliable spectra)

Spectra ID	Features	Distance metric (unique, redundant spectra	Chargestate	Ion intensity	Sequence length	Identification score	Modifica-tion	Mass
6		3	2.5	3	5	6	4	2
7		4	2.5	5	5	5	4	1
8		7	6	7	5	1	4	3
1185		1.5	2.5	2	1.5	2	4	4
1186		1.5	6	6	1.5	4	4	5
2817		5.5	6	4	5	7	4	7
2818		5.5	2.5	1	5	3	4	6

(3) Normalization of ranks: Ranks of each feature are divided by the number of spectra to be within the range of 0 and1.

Spectra ID	Features	Distance metric (unique, redundant spectra	Charge state	Ion intensity	Sequence length	Identification score	Modification	Mass
6		0.4285714	0.3571429	0.4285714	0.7142857	0.8571429	0.5714286	0.2857143
7		0.5714286	0.3571429	0.7142857	0.7142857	0.7142857	0.5714286	0.1428571
8		1	0.8571429	1	0.7142857	0.1428571	0.5714286	0.4285714
1185		0.2142857	0.3571429	0.2857143	0.2142857	0.2857143	0.5714286	0.5714286
1186		0.2142857	0.8571429	0.8571429	0.2142857	0.5714286	0.5714286	0.7142857
2817		0.7857143	0.8571429	0.5714286	0.7142857	1	0.5714286	1
2818		0.7857143	0.3571429	0.1428571	0.7142857	0.4285714	0.5714286	0.8571429

(4) Feature Weighting (impact of the different features: ranks are multiplied by feature weight)

Feature Order (weight = squared order value)		7	6	5	4	3	2	1
Spectra ID	Features	Distance metric (unique, redundant spectra	Charge state	Ion intensity	Sequence length	Identification score	Modification	Mass
6		21	12.85714	10.714286	11.428571	7.714286	2.285714	0.2857143
7		28	12.85714	17.857143	11.428571	6.428571	2.285714	0.1428571
8		49	30.85714	25	11.428571	1.285714	2.285714	0.4285714
1185		10.5	12.85714	7.142857	3.428571	2.571429	2.285714	0.5714286
1186		10.5	30.85714	21.428571	3.428571	5.142857	2.285714	0.7142857
2817		38.5	30.85714	0.5714286	11.428571	9	2.285714	1
2818		38.5	12.85714	0.1428571	11.428571	3.857143	2.285714	0.8571429

(5) Inference of overall spectra reliability

Computation of an average rank per spectrum and normalization by the weighted sum of all features.

Peptide spectra weights:

Spectra ID	6	7	8	1185	1186	2817	2818
Weight	0.18426990	0.12618588	0.01317993	0.34349229	0.14612551	0.03613495	0.15061155

(6) Protein ratio calculation

A weighted mean approach using peptide spectrum weights is conducted to estimate the underlying protein ratio

[Matrix with relative intensities of peptide spectra (8 plex)]

```
ID      X113       X114      X115       X116       X117       X118       X119       X121
6     0.9922024  1.0077976  0.5051960  0.5550352  0.9948927  1.0465061  0.5142161  0.5249595
7     1.0166331  0.9833669  0.5696561  0.5095101  1.0295853  1.0594107  0.5374796  0.5245989
8     0.9589733  1.0410267  0.4046786  0.5723454  0.9669290  1.0189080  0.3920415  0.4598631
1185  0.9917052  1.0082948  0.5653267  0.5633381  1.1026590  1.0625592  0.5520376  0.4468080
1186  1.0194056  0.9805944  0.5564336  0.5981815  1.0654046  0.9505913  0.5109424  0.4179024
2817  0.9632542  1.0367458  0.5135213  0.4295675  0.9611111  1.0579834  0.4294693  0.3642797
2818  1.0183185  0.9816815  0.5449597  0.5032795  1.0212967  0.9699722  0.5249286  0.4804103
```

Output: estimated protein ratio

```
   X113       X114       X115       X116       X117       X118       X119       X121
1.0015389  0.9984611  0.5464364  0.5463467  1.0489784  1.0281571  0.5266055  0.4690521
```

Feature Weighting

The feature-error-correlation plots and the correlation coefficients as a measurement show different explanation power of the presented features. We propose a default order of the features from the most meaningful to the one with least impact on ratio accuracy and assign weights to the features. Weights of the features correspond to their squared order values.

A comparison analysis of two different feature orders is conducted to prove the general robustness of the feature weighting in the iPQF algorithm.

Features	Default feature order	Changed feature order
Distance metric (Redundant-Uniquely measured spectra)	7	7
Charge state	6	5
Ion intensity	5	6
Sequence length	4	4
Mass	1	3
Identification score	3	1
Modification	2	2

The feature order used by default in iPQF is compared to a slightly varied feature order. The ‘changed feature order’ is computed based on the mean correlation coefficients of the five different data sets considered. We investigate the evaluation results of the ‘pure iPQF’ approach using the default and the changed weighting of features and present it in Fig. B 15. Protein evaluation values are visualized by boxplots for each of the different data sets. As a result no significant shift in protein evaluation values between the default and the changed feature weighting is observed and a general robustness can be confirmed. Overall, it can be noted that the given basic order is of importance to distinguish between most and least strong features, however, change of close positions do not significantly impact results (here e.g. exchange of identification score and modification). Further in the case of highly correlated features, such as length and mass, feature information becomes redundant. Here it is preferable to select one for a higher weight and assign a lower weight to the other. Assigning similar weights to highly correlated features tends to introduce a bias as this doubles the impact of one feature category.

We provide the user with a default ranking of features; however it is optional to input a user-defined ranking as well.

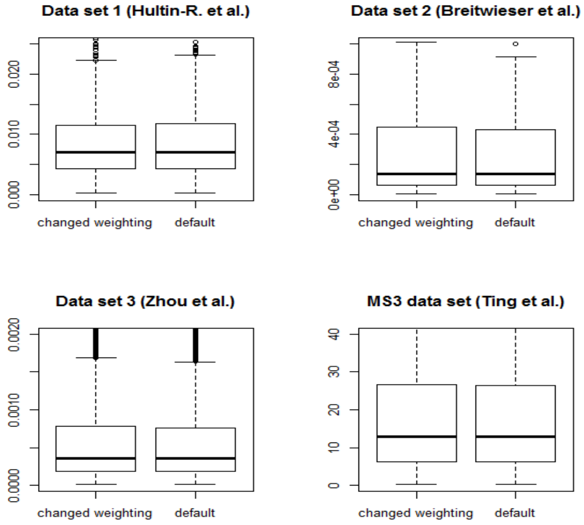


Fig. B 15:
Evaluation of ‘pure iPQF’ approach according to different feature weighting. Boxplots of protein evaluation values are shown for the three different MS2 data sets and the MS3 data set. No significant shift between the default and the changed feature weighting is observed, thus a general robustness is given.

C. Appendix

Fig. C 1:

(a)

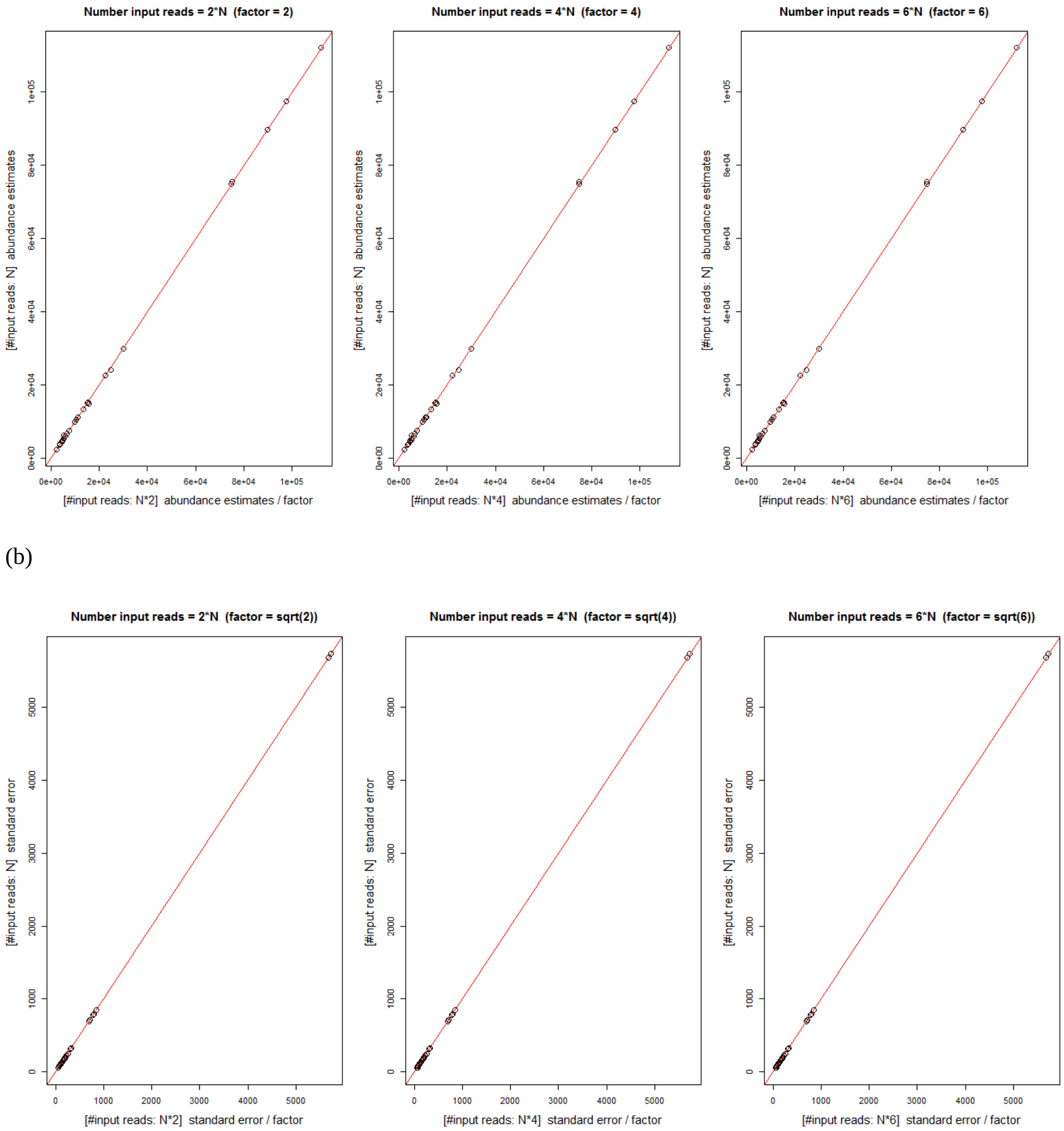
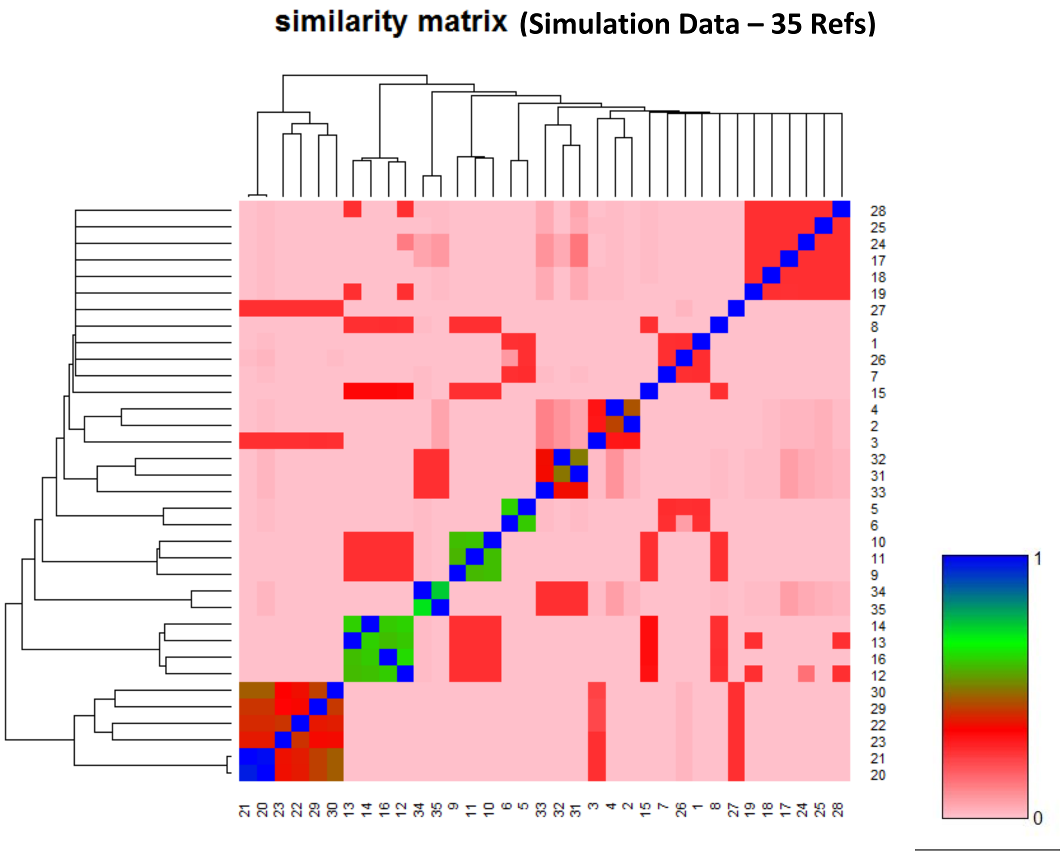


Fig. C 1: **Impact of total number of input reads on (a) abundance estimates and (b) standard errors.** We conducted a study applying different total numbers of input reads (exemplary for the ‘original’ simulation set 4): increasing the original number of input reads N (N =750.000) by the factor of 2, 4, and 6, corresponding to total amounts of 1.5, 3, and 4.5 million input reads for the set. We conducted comparisons of abundance estimates and standard errors computed on the ‘sets with increased read number’ against the results obtained by the ‘original’ set. It can be observed that the abundance estimates scale linear with the number of reads, whereas the standard errors scale quadratic.

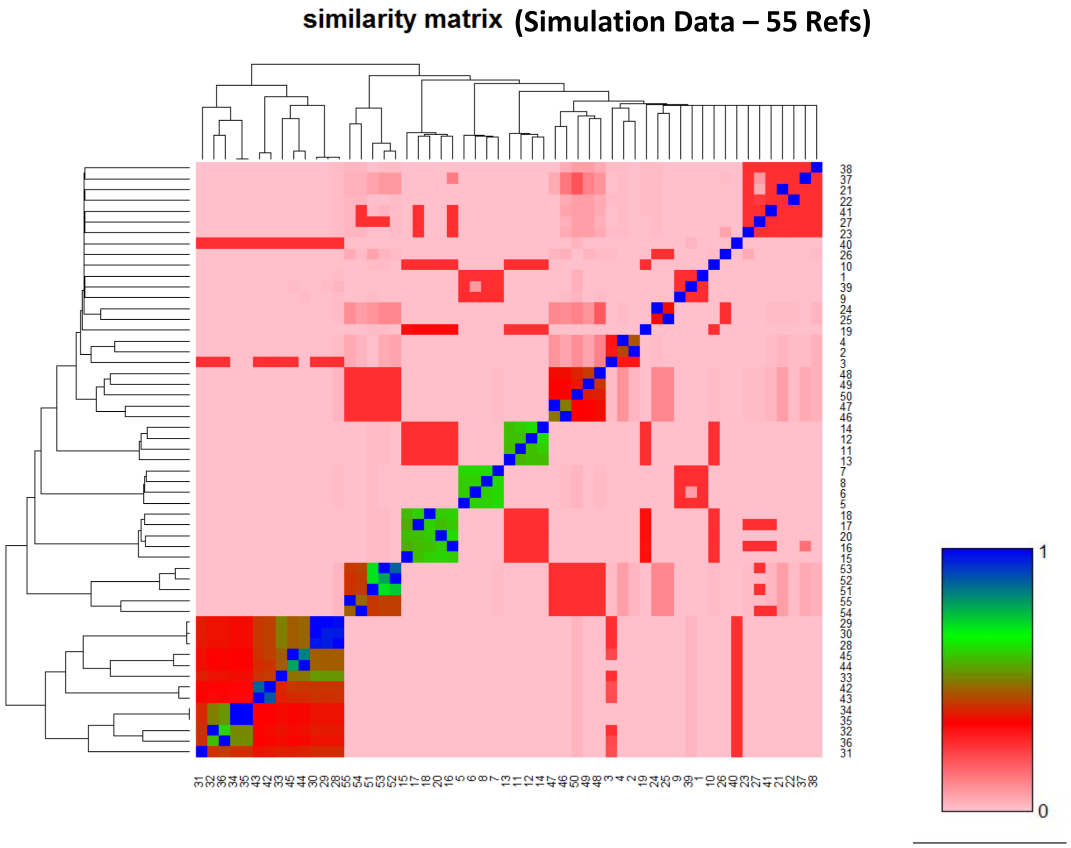
Fig. C 2:



1: Alistipes_finegoldii_DSM_17242
2: Bacillus_anthracis_Sterne
3: Bacillus_cereus_ATCC_10987
4: Bacillus_cereus_E33L
5: Bacteroides_fragilis_638R
6: Bacteroides_fragilis_NCTC_9343
7: Bacteroides_thetaiotaomicron_VPI_5482
8: Bifidobacterium_adolescentis_ATCC_15703
9: Bifidobacterium_bifidum_BGN4
10: Bifidobacterium_bifidum_PRL2010
11: Bifidobacterium_bifidum_S17
12: Bifidobacterium_longum_BBMN68
13: Bifidobacterium_longum_DJO10A
14: Bifidobacterium_longum_infantis_157F
15: Bifidobacterium_longum_infantis_ATCC_15697
16: Bifidobacterium_longum_JCM_1217
17: Clostridium_phytofermentans_ISDg
18: Clostridium_saccharolyticum_WM1
19: Clostridium_SY8519
20: Escherichia_coli_K_12_substr__DH10B
21: Escherichia_coli_K_12_substr__MG1655
22: Escherichia_coli_O7_K1_CE10
23: Escherichia_coli_S88
24: Eubacterium_eligens_ATCC_27750
25: Eubacterium_rectale_ATCC_33656
26: Odoribacter_splanchnicus_DSM_20712
27: Pantoea_ananatis_PA13
28: Roseburia_hominis_A2_183
29: Shigella_dysenteriae_Sd197
30: Shigella_flexneri_2a_301
31: Streptococcus_salivarius_57_I
32: Streptococcus_salivarius_CCHSS3
33: Streptococcus_salivarius_JIM8777
34: Streptococcus_suis_D9
35: Streptococcus_suis_ST3

Fig. C 2: **Similarity matrix of the simulation data sets comprising 35 reference genomes** (see list of taxa accession numbers in the subsequent section ‘Data Set description’). The heatmap visualizes all pairwise reference sequence similarities ranging from 0 to 100% similarity (visualized from pink to dark blue). The diagonal of the matrix refers to the proportion of simulated reads mapping back to their reference of origin. Different clusters of strains exhibiting high reference sequence similarities can be observed. Notably is the first big cluster of diverse *Escherichia coli* strains (bottom left), comprising two sub-strains which share 98% sequence similarity, two more distant *E.coli* strains, and further two *Shigella* strains known to be closely related to *E.coli*. The second big cluster comprises four different strains of *Bifidobacterium longum*, followed by a cluster of *Bifidobacterium bifidum*, which expresses moderate similarities to the former. Further, various smaller clusters of strains are present.

Fig. C 3:



similarity matrix (Simulation Data – 55 Refs)

1: Alistipes_finegoldii_DSM_17242
2: Bacillus_anthracis_Sterne
3: Bacillus_cereus_ATCC_10987
4: Bacillus_cereus_E33L
5: Bacteroides_fragilis_638R
6: Bacteroides_fragilis_NCTC_9343
7: Bacteroides_fragilis_strain_BOB25
8: Bacteroides_fragilis_YCH46
9: Bacteroides_thetaiotaomicron_VPI_5482
10: Bifidobacterium_adolescentis_ATCC_15703
11: Bifidobacterium_bifidum_BGN4
12: Bifidobacterium_bifidum_PRL2010
13: Bifidobacterium_bifidum_S17
14: Bifidobacterium_bifidum_ATCC_29521
15: Bifidobacterium_longum_subsp_longum_44B
16: Bifidobacterium_longum_BBMN68
17: Bifidobacterium_longum_DJO10A
18: Bifidobacterium_longum_infantis_157F
19: Bifidobacterium_longum_infantis_ATCC_15697
20: Bifidobacterium_longum_JCM_1217
21: Clostridium_phytofermentans_ISDg
22: Clostridium_saccharolyticum_WM1
23: Clostridium_SY8519
24: Clostridium_botulinum_A3_str_Loch_Maree
25: Clostridium_botulinum_B1_str_Okra
26: Clostridium_botulinum_B_str_Eklund_17B
27: Clostridium_cf_saccharolyticum_K10
28: Escherichia_coli_K_12_substr__DH10B
29: Escherichia_coli_K_12_substr__MG1655
30: Escherichia_coli_str_K_12_substr_MC4100
31: Escherichia_coli_O7_K1_CE10
32: Escherichia_coli_S88
33: Escherichia_coli_O104_H4_str_2011C_3493
34: Escherichia_coli_O127_H6_str._E2348_69_substr._CVDNalr_genomic
35: Escherichia_coli_O127_H6_str._E2348_69_substr._UMD753_genomic
36: Escherichia_coli_O83_H1_str_NRG_857C
37: Eubacterium_eligens_ATCC_27750
38: Eubacterium_rectale_ATCC_33656
39: Odoribacter_splanchnicus_DSM_20712
40: Pantoea_ananatis_PA13
41: Roseburia_hominis_A2_183
42: Shigella_dysenteriae_Sd197
43: Shigella_dysenteriae_1617
44: Shigella_flexneri_5_str_8401
45: Shigella_flexneri_2a_301
46: Streptococcus_salivarius_57_I
47: Streptococcus_salivarius_CCHSS3
48: Streptococcus_salivarius_JIM8777
49: Streptococcus_salivarius_strain_HSISS4
50: Streptococcus_salivarius_strain_NCTC_8618
51: Streptococcus_suis_D9
52: Streptococcus_suis_ST3
53: Streptococcus_suis_05HAS68
54: Streptococcus_suis_JS14
55: Streptococcus_suis_T15

Fig. C 3: **Similarity matrix of the simulation data sets comprising 55 reference genomes, exhibiting large similar strain clusters** (see list of taxa accession numbers in the subsequent section ‘Data Set description’). The heatmap visualizes all pairwise reference sequence similarities ranging from 0 to 100% similarity (visualized from pink to dark blue). Additional strain and sub-strain sequences were added to the simulation set of 35 references to challenge the tools: a big cluster of overall 13 taxa of *Escherichia coli* strains containing three different sub-strain clusters with sequence similarities above 95%, mixed with diverse distant *E.coli* strains and closely related S*higella* strains. Further, cluster of *Bifidobacterium longum, Bifidobacterium bifidum, Bacteroides fragilis* as wells as two different *Streptococcus* species cluster were largely extended to test the resolution performance of the tools within large and highly similar strain clusters.

Fig. C 4:

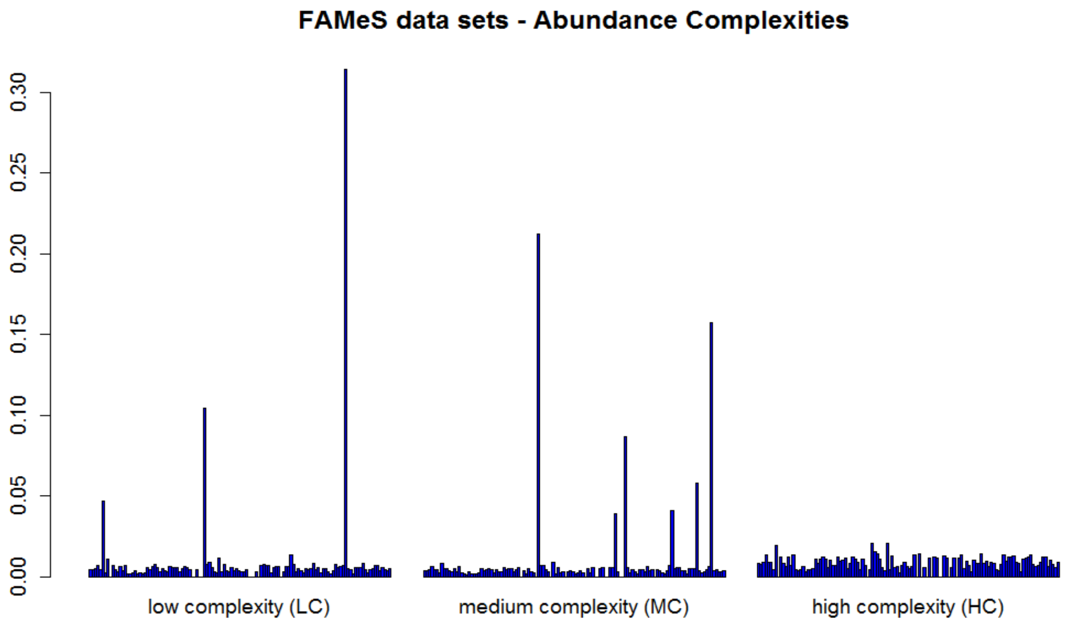


Fig. C 4: The **FAMeS data** comprises three different samples with abundance profiles according to low (*LC*), medium (*MC*) and high complexity (*HC*), a common classification in metagenomics. Thereby, a low complexity sample may represent a bioreactor community with one dominant among low abundant genomes, while medium complexity refers to a moderately complex community with few dominating taxa. High complexity samples are frequently characterized by no dominating taxa present or also by very long tails of low abundant taxa.

Fig. C 5:

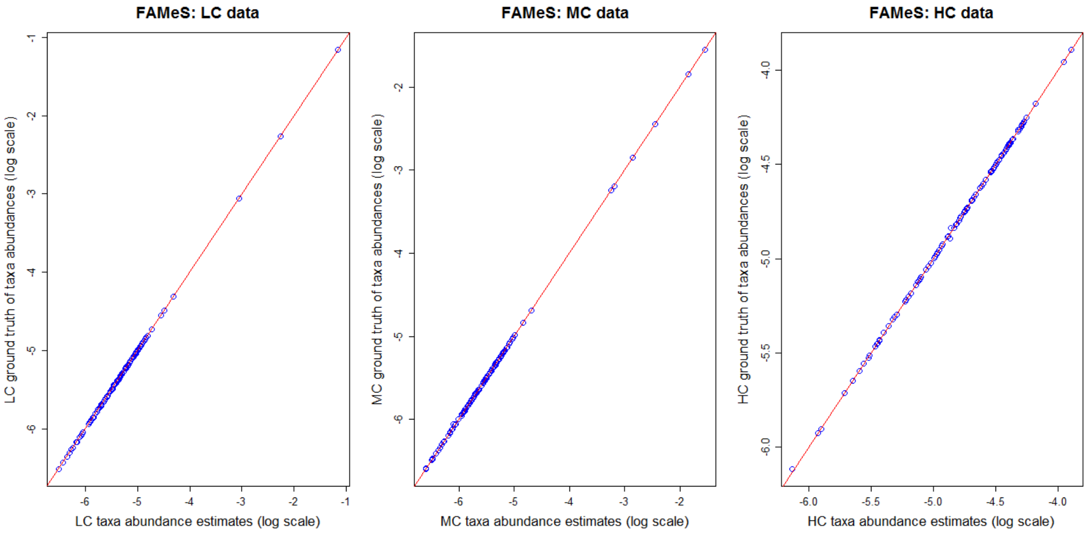


Fig. C 5: **Accuracy of abundance estimates by DiTASiC for the FAMeS data sets.** For all three samples of LC, MC, and HC, abundance estimates exhibit only tiny divergences from the ground truth. High accuracy is depicted by the points found on the diagonal. Hence, highly accurate abundance estimates of the considered 122 taxa are achieved across all three different abundance complexity profiles.

Fig. C 6:

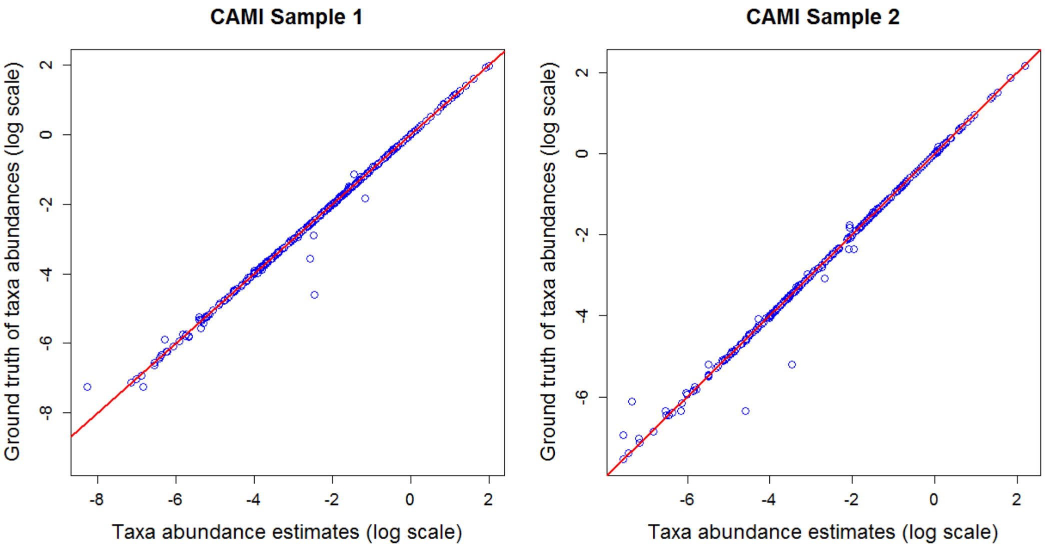


Fig. C 6: **Accuracy of abundance estimates by DiTASiC for samples of the CAMI benchmark data set.** For both samples, abundance estimates of the 255 taxa show high accuracy apart from very few outliers. High accuracy is depicted by points found on the diagonal. Notably, accurate estimates are also achieved for very low relative abundances below 0.01%.

Fig. C 7:

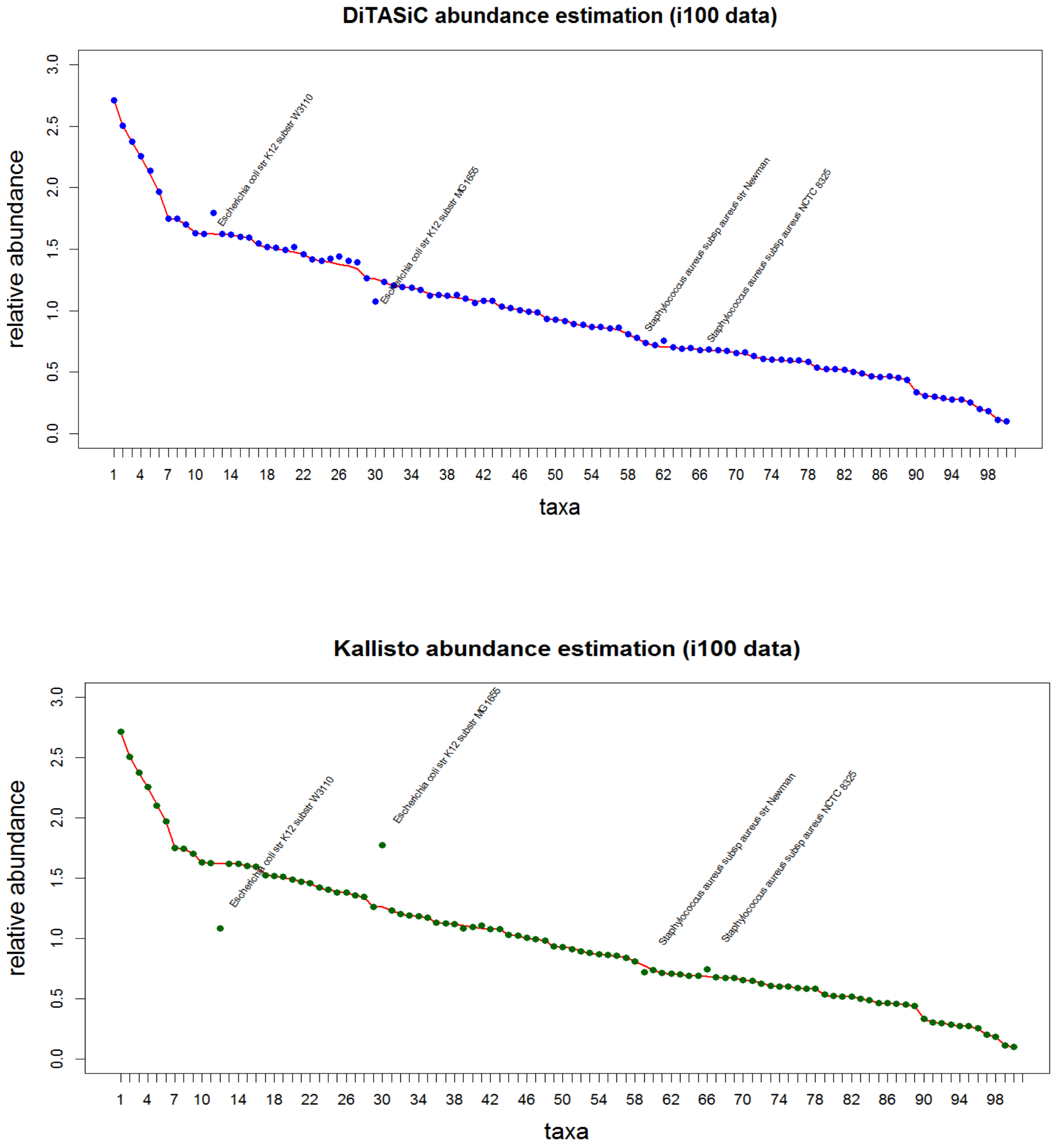
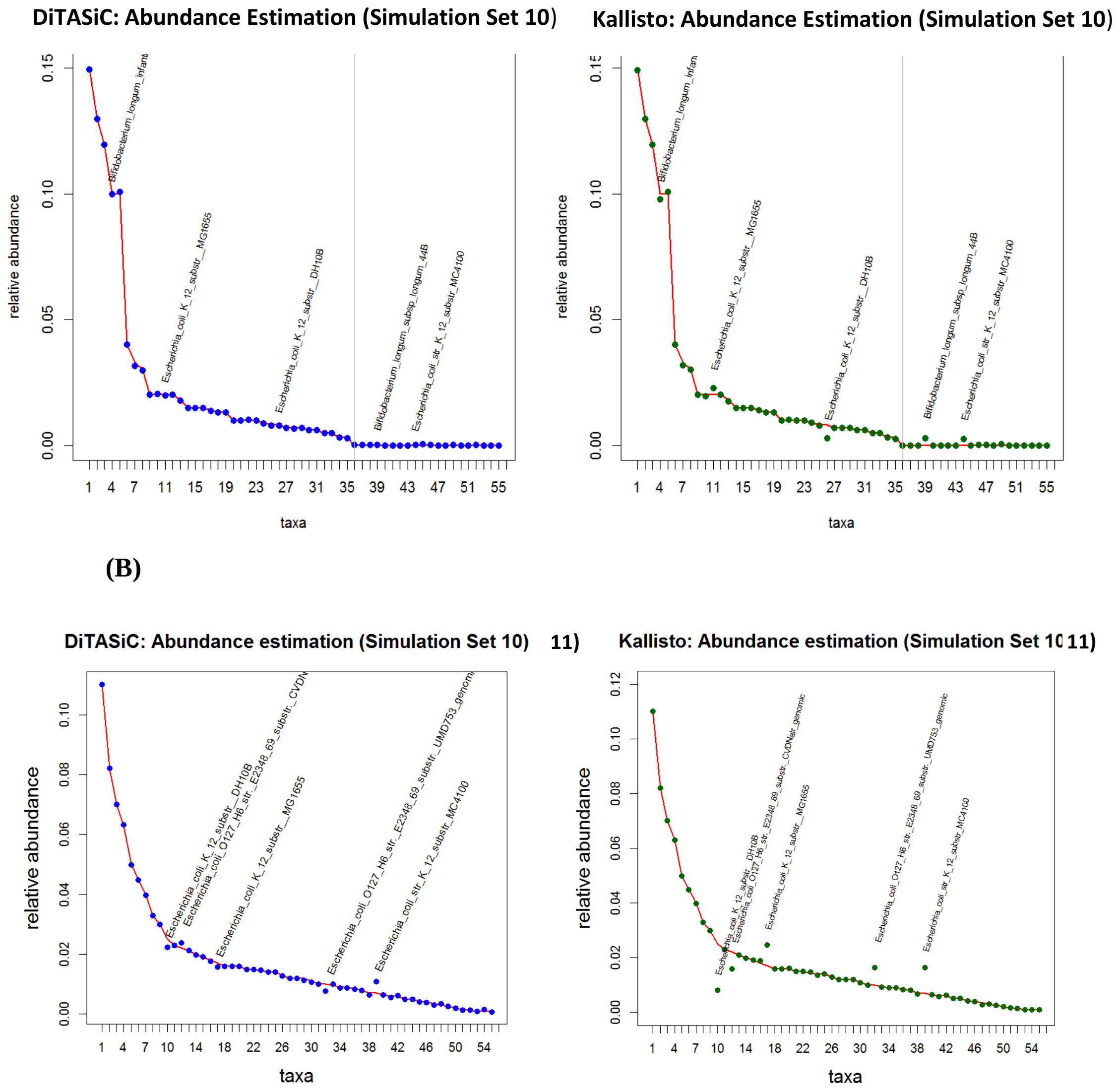


Fig. C 7: **Accuracy of abundance estimates by DiTASiC and kallisto for the Illumina 100 benchmark data set (i100)** *(Mende et al., 2012)*. The red line refers to the ground truth values and the points show the abundance estimates obtained by the corresponding tool. Overall, a high accuracy of abundance estimates is achieved for the 100 taxa by both tools across the entire abundance range. A bias in abundance estimation is observed for some strains of high sequence similarity, namely for the *Escherichia coli* sub-strains and for two *Staphylococcus aureus* strains. A more accurate abundance resolution of these strain clusters is obtained by DiTASiC in comparison to kallisto (also refer to Fig. C 9C). Further, results of DiTASiC can be related to a recent benchmark study of different abundance profiling tools tested on the *i100* data set by Schaffer *et al*. (2017): see end of Appendix C.

Fig. C 8:

(A)



(B)

Fig. C 8: **Accuracy of abundance estimates by DiTASiC and kallisto for (A) data set 10 and (B) data set 11 of simulation group (3).** The red line refers to the ground truth values and the points show the abundance estimates obtained by the corresponding tools. (A) Set 10 serves to study the impact of absent strains from highly similar clusters (gray vertical line in the plot to mark the section of absent strains). Overall, highly accurate abundance estimates are obtained by DiTASiC. Hence an un-biased estimation of strains of the clusters affected by absent strains is achieved. kallisto exhibits difficulties with some strains of high sequence similarity, here concerning the *Escherichia coli K12* sub-strain cluster and the *Bifidobacterium longum* strain group, causing a bias of abundance estimations and calling two of the absent strains abundant. (B) Set 11 focuses on the resolution of large and highly similar strain clusters, having all 55 taxa abundant in the data set (refer also to the matrix of reference similarities in Fig. C3). Overall accurate abundance estimations are obtained and also an accurate resolution within the diverse strain clusters is achieved by DiTASiC. The large *E.coli* cluster causes some abundance biases for both tools, especially for the sub-strain sequences of sequence similarities above 95%. Here, DiTASiC proves more accurate estimations and an overall better resolution within the considered cluster (see also Fig. C 9A).

Fig. C 9 (A)

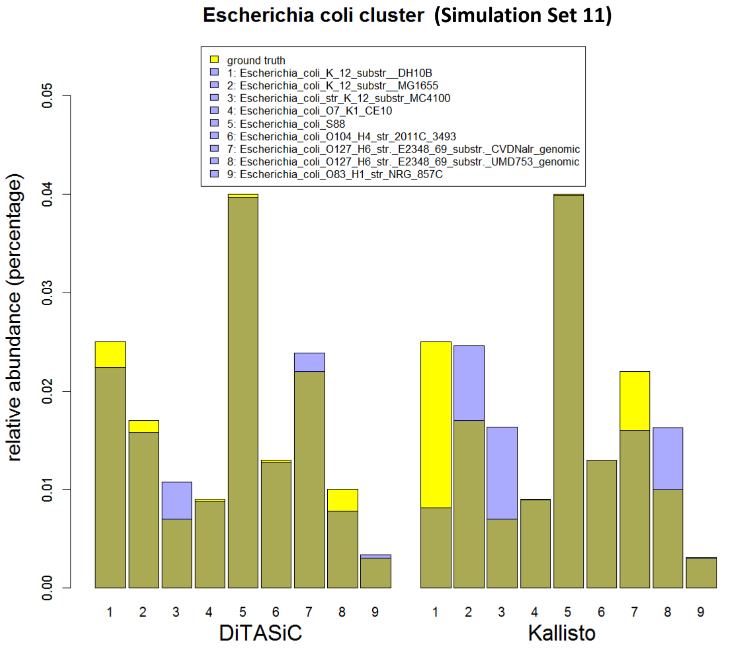


Fig. C 9 (B):

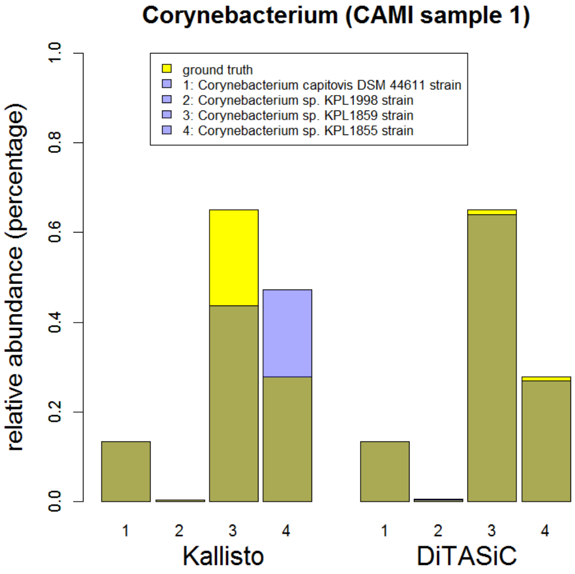
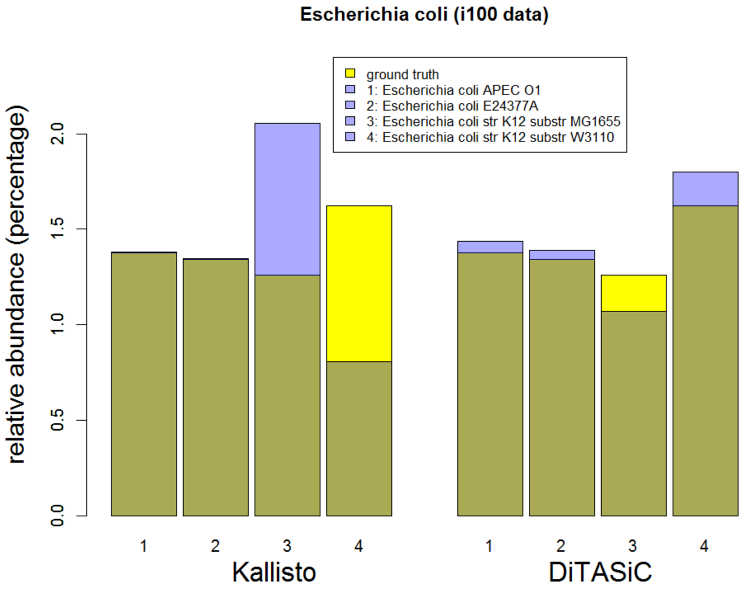


Fig. C 9 (C)


Escherichia coli (i100 data)
ground truth
1: Escherichia coli APEC O1
2: Escherichia coli E24377A
3: Escherichia coli str K12 substr MG1655
4: Escherichia coli str K12 substr W3110
relative abundance (percentage)
0.0
0.5
1.0
1.5
2.0
1
2
3
4
Kallisto
1
2
3
4
DiTASiC

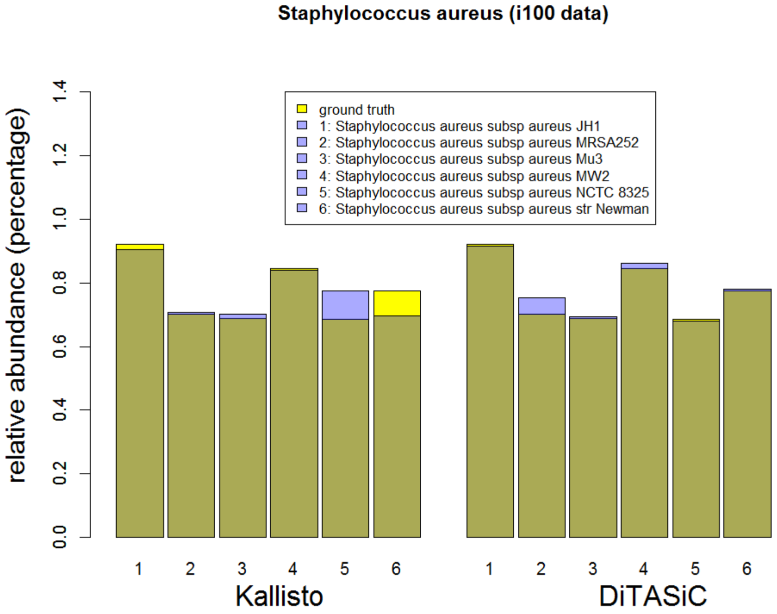

Staphylococcus aureus (i100 data)
ground truth
1: Staphylococcus aureus subsp aureus JH1
2: Staphylococcus aureus subsp aureus MRSA252
3: Staphylococcus aureus subsp aureus Mu3
4: Staphylococcus aureus subsp aureus MW2
5: Staphylococcus aureus subsp aureus NCTC 8325
6: Staphylococcus aureus subsp aureus str Newman
relative abundance (percentage)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1
2
3
4
5
6
Kallisto
1
2
3
4
5
6
DiTASiC

Fig. C 9 (D)

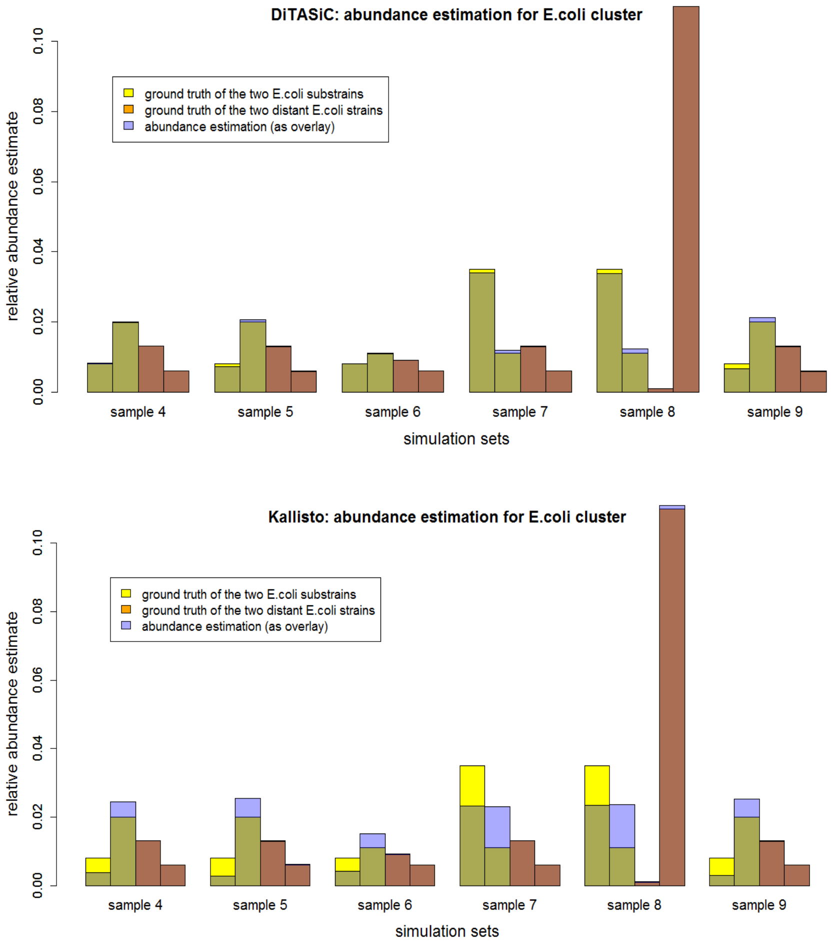
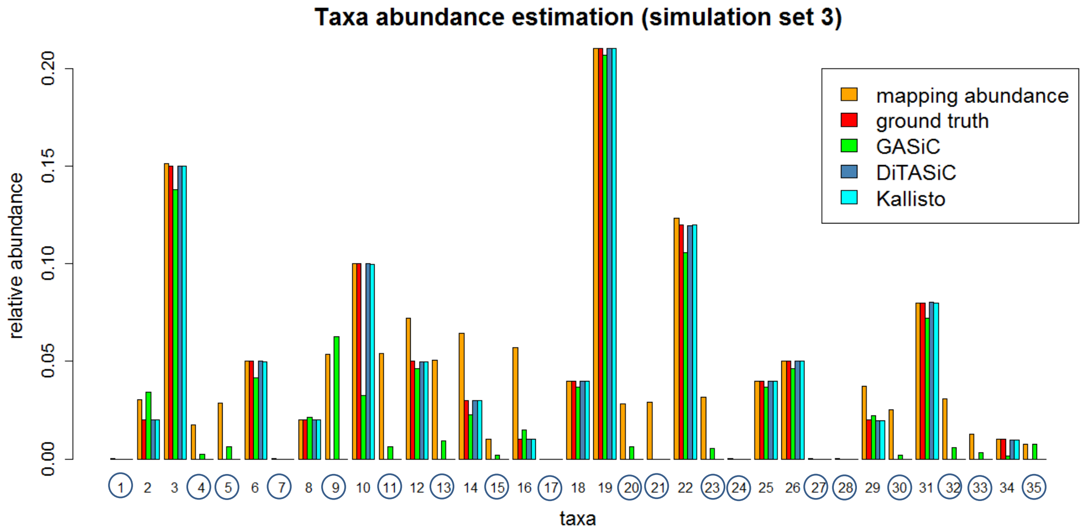


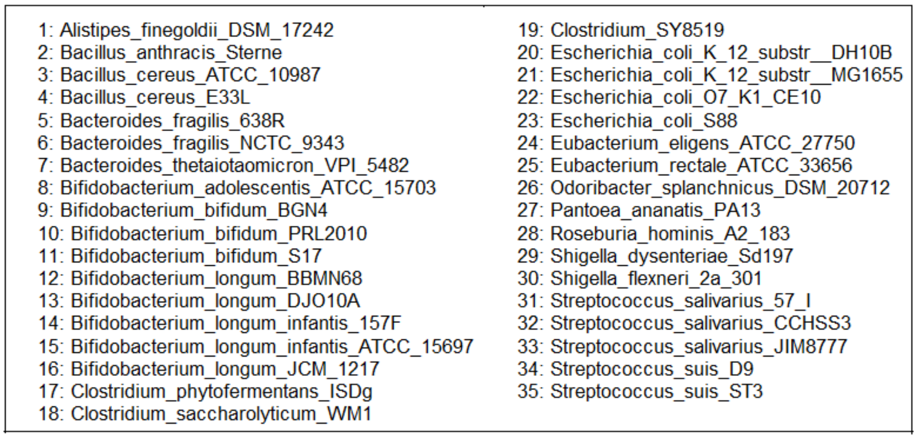
Fig. C 9: Accuracy of abundance resolution within strain clusters by DiTASiC and kallisto shown for different examples in the CAMI, i100 and simulation data (A-D). The ground truth abundance is displayed in yellow and the estimated abundances by the corresponding tools are overlayed with purple colour. **(A)** The largest cluster in simulation set 11 comprises nine different *Escherichia coli* strains. Challenging is the resolution of the present sub-strains which share sequence similarities above 98%. DiTASiC enables a more accurate abundance resolution in comparison to kallisto. **(B)** An example of a *Corynebacterium* cluster in the CAMI set reveals a perfect resolution by DiTASiC, again, two of the strains are characterized by high sequence similarity. **(C)** An increased error in abundance estimation in the i100 data was shown in Fig S7 for the *Escherichia coli* sub-strains and for two *Staphylococcus aureus* strains. A more accurate abundance resolution of these strain clusters is obtained by DiTASiC. **(D)** Here, we consider the six different simulations sets of group (2) focusing on the abundance estimates obtained for the 4 strains of the *E.coli* cluster (*E. coli K-12 substr. DH10B* , *E. coli K-12 substr. MG1655*, *E.coli O7:K1 str. CE10*, *E.coli S88)* (visualized only for group (2), as in group (1) not all strains of the *E.coli* cluster are abundant). The *E.coli* cluster consists of two sub-strains, which share 98% sequence similarity, and two more distant strains. DiTASiC enables a highly accurate abundance resolution of the entire strain cluster, as is shown by an almost perfect abundance estimation overlay in the plot across all samples. kallisto exhibits problems in the resolution of the two sub-strains, which is shown by a consistent abundance underestimation of *E. coli K-12 substr. DH10B* and abundance overestimation of *E. coli K-12 substr. MG1655*, while the two distant strains receive accurate estimations. Overall, it can be observed that a common error in the resolution of a strain cluster is an abundance interchange or equalization of abundances of similar sub-strains. In the resolution by DiTASiC these errors are shown to be avoided.

Fig. C 10:

(a)



Taxa, shown in the plot according to the numbers:


1: Alistipes_finegoldii_DSM_17242
2: Bacillus_anthracis_Sterne
3: Bacillus_cereus_ATCC_10987
4: Bacillus_cereus_E33L
5: Bacteroides_fragilis_638R
6: Bacteroides_fragilis_NCTC_9343
7: Bacteroides_thetaiotaomicron_VPI_5482
8: Bifidobacterium_adolescentis_ATCC_15703
9: Bifidobacterium_bifidum_BGN4
10: Bifidobacterium_bifidum_PRL2010
11: Bifidobacterium_bifidum_S17
12: Bifidobacterium_longum_BBMN68
13: Bifidobacterium_longum_DJO10A
14: Bifidobacterium_longum_infantis_157F
15: Bifidobacterium_longum_infantis_ATCC_15697
16: Bifidobacterium_longum_JCM_1217
17: Clostridium_phytofermentans_ISDg
18: Clostridium_saccharolyticum_WM1
19: Clostridium_SY8519
20: Escherichia_coli_K_12_substr__DH10B
21: Escherichia_coli_K_12_substr__MG1655
22: Escherichia_coli_O7_K1_CE10
23: Escherichia_coli_S88
24: Eubacterium_eligens_ATCC_27750
25: Eubacterium_rectale_ATCC_33656
26: Odoribacter_splanchnicus_DSM_20712
27: Pantoea_ananatis_PA13
28: Roseburia_hominis_A2_183
29: Shigella_dysenteriae_Sd197
30: Shigella_flexneri_2a_301
31: Streptococcus_salivarius_57_I
32: Streptococcus_salivarius_CCHSS3
33: Streptococcus_salivarius_JIM8777
34: Streptococcus_suis_D9
35: Streptococcus_suis_ST3

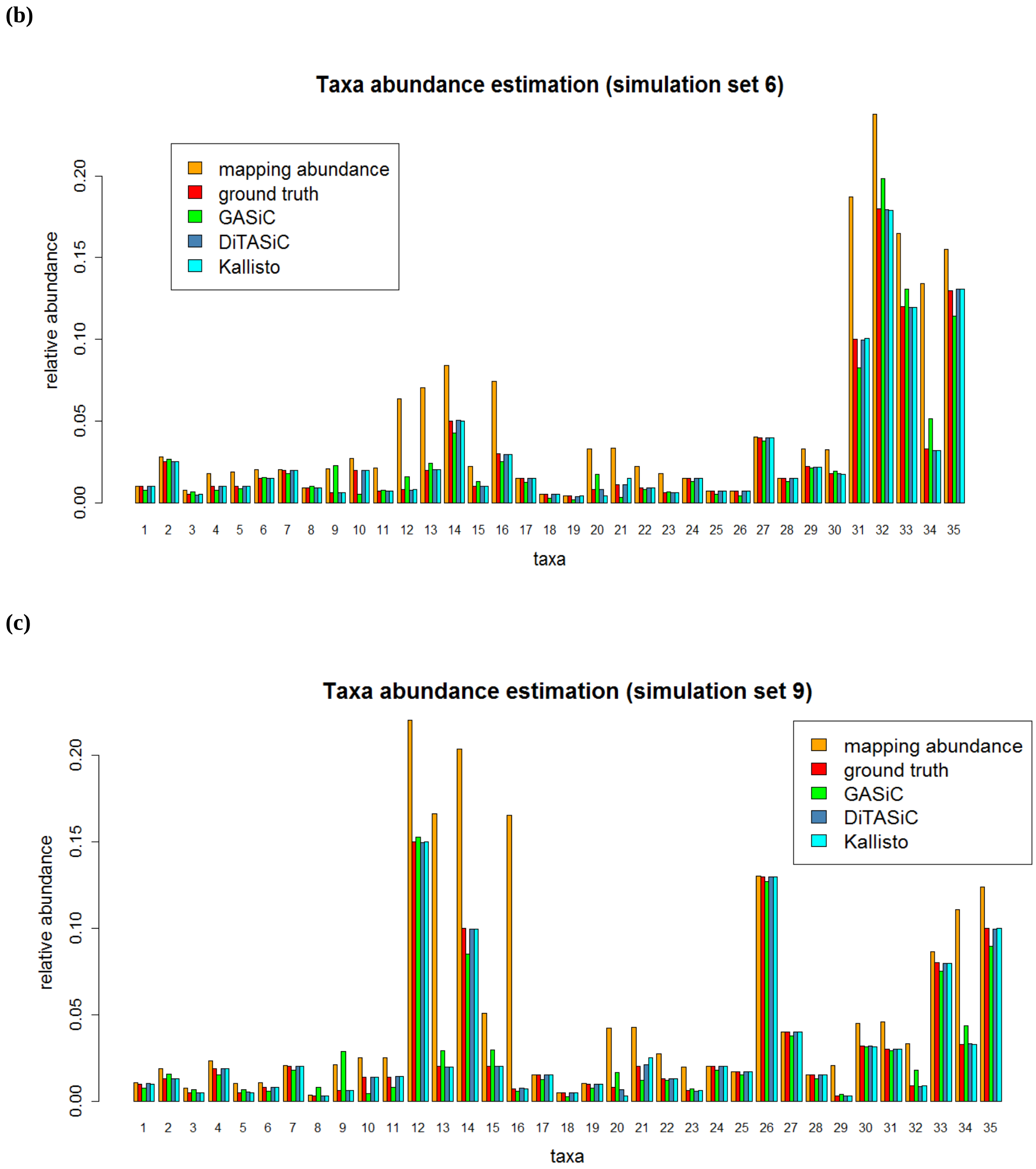
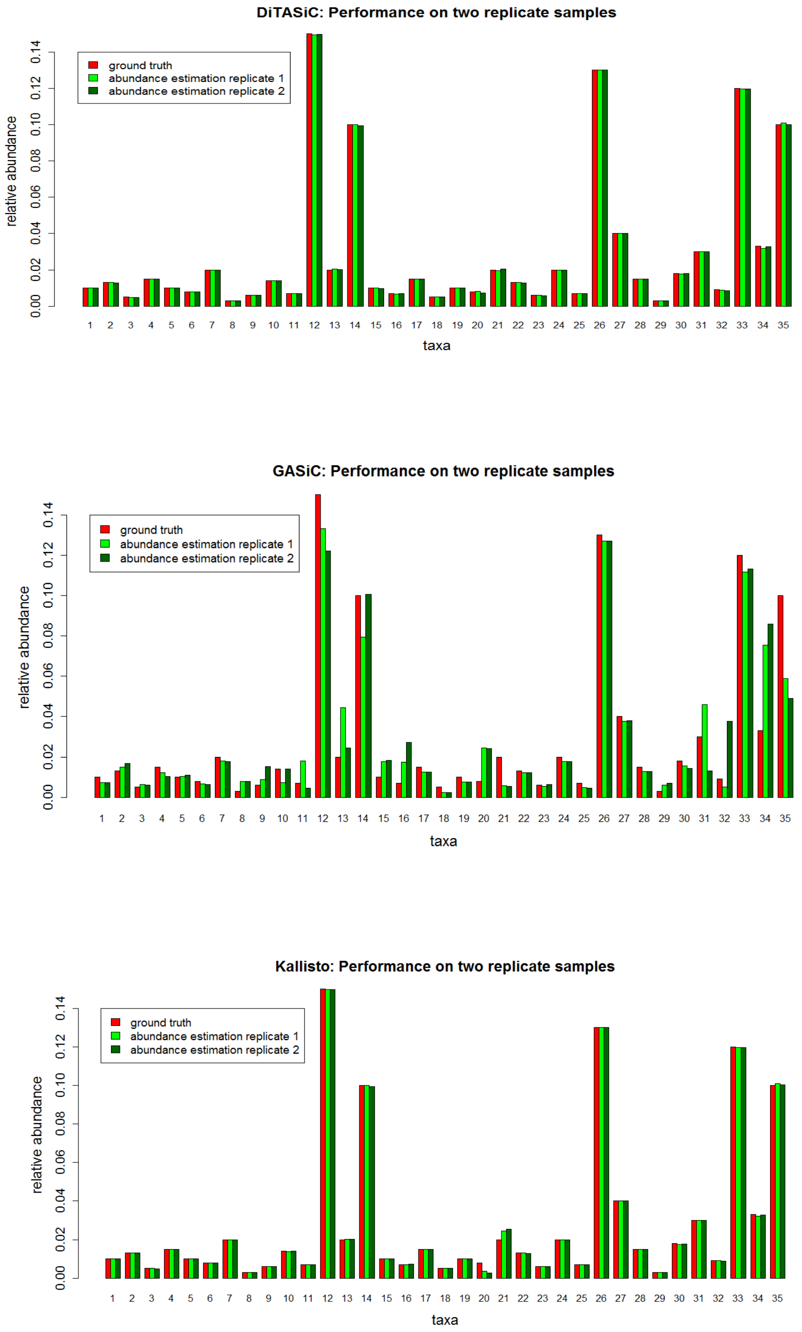


Fig. C 10: **Taxa abundance estimates exemplary for three simulation data set of various abundance profiles (a-c), presenting the different tools DiTASiC, GASiC and kallisto in comparison to the ground truth and observed mapping abundances**. Mapping abundances are biased due to read ambiguities which causes overestimation or assignment of reads to absent taxa (absent taxa are marked with a circle around the taxa number). DiTASiC as well as kallisto exhibit highly accurate estimations and a clear improvement over GASiC.

Fig. C 11:



(a)

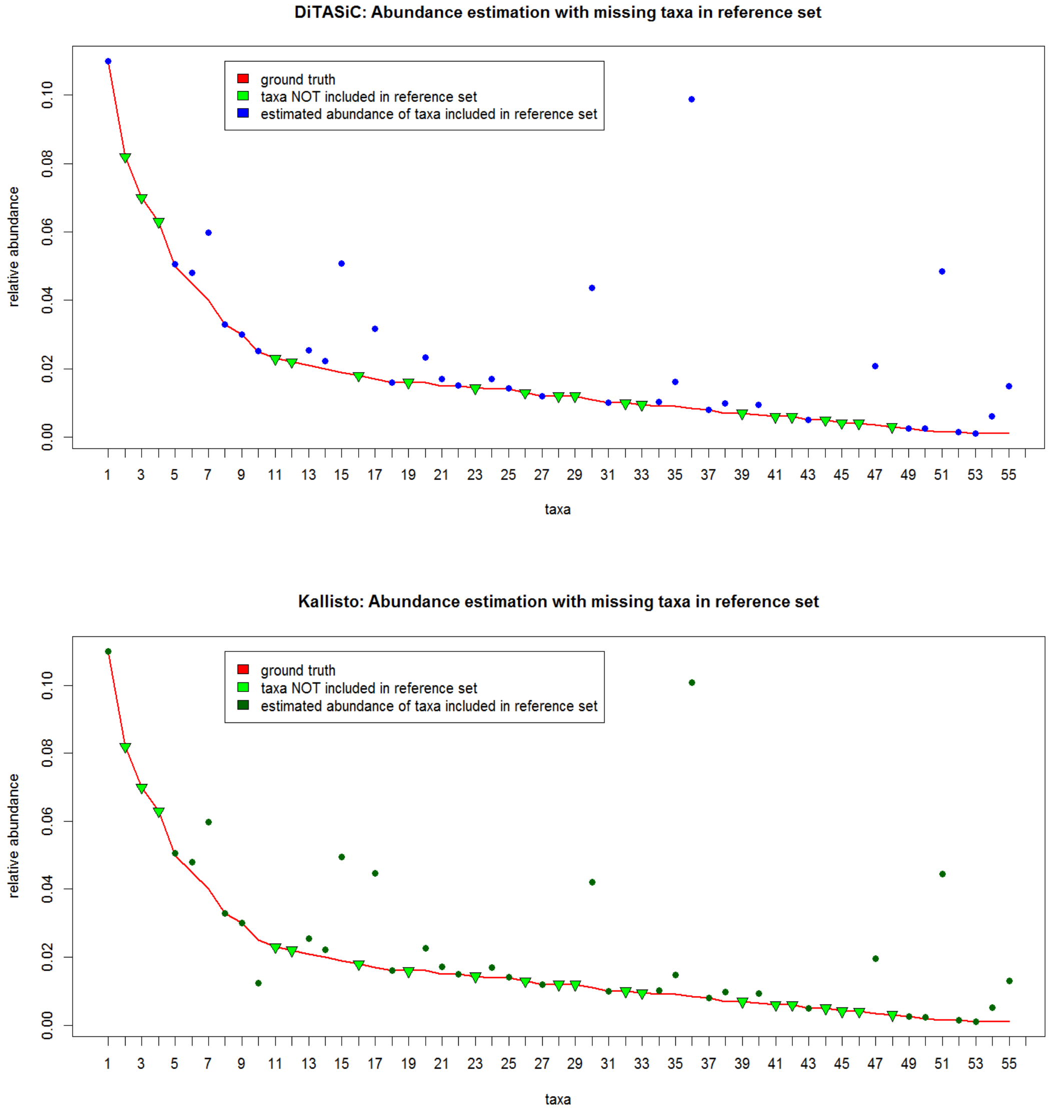
(b)

(c)

Fig. C 11: **Robustness evaluation of (a) DiTASiC, (b) GASiC, and (c) kallisto, on two replicate samples** from the simulation data (data set 4 and data set 5, respectively; taxa are numbered according to the list given in Fig. C 7). DiTASiC and kallisto show an overall robust performance in abundance estimation of all 35 taxa in the replicates, and a significant improvement compared to GASiC.

Fig. C 12:

(A)


DiTASiC: Abundance estimation with missing taxa in reference set
ground truth
taxa NOT included in reference set
estimated abundance of taxa included in reference set
relative abundance
0.00
0.02
0.04
0.06
0.08
0.10
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55
taxa
Kallisto: Abundance estimation with missing taxa in reference set
ground truth
taxa NOT included in reference set
estimated abundance of taxa included in reference set
relative abundance
0.00
0.02
0.04
0.06
0.08
0.10
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55
taxa

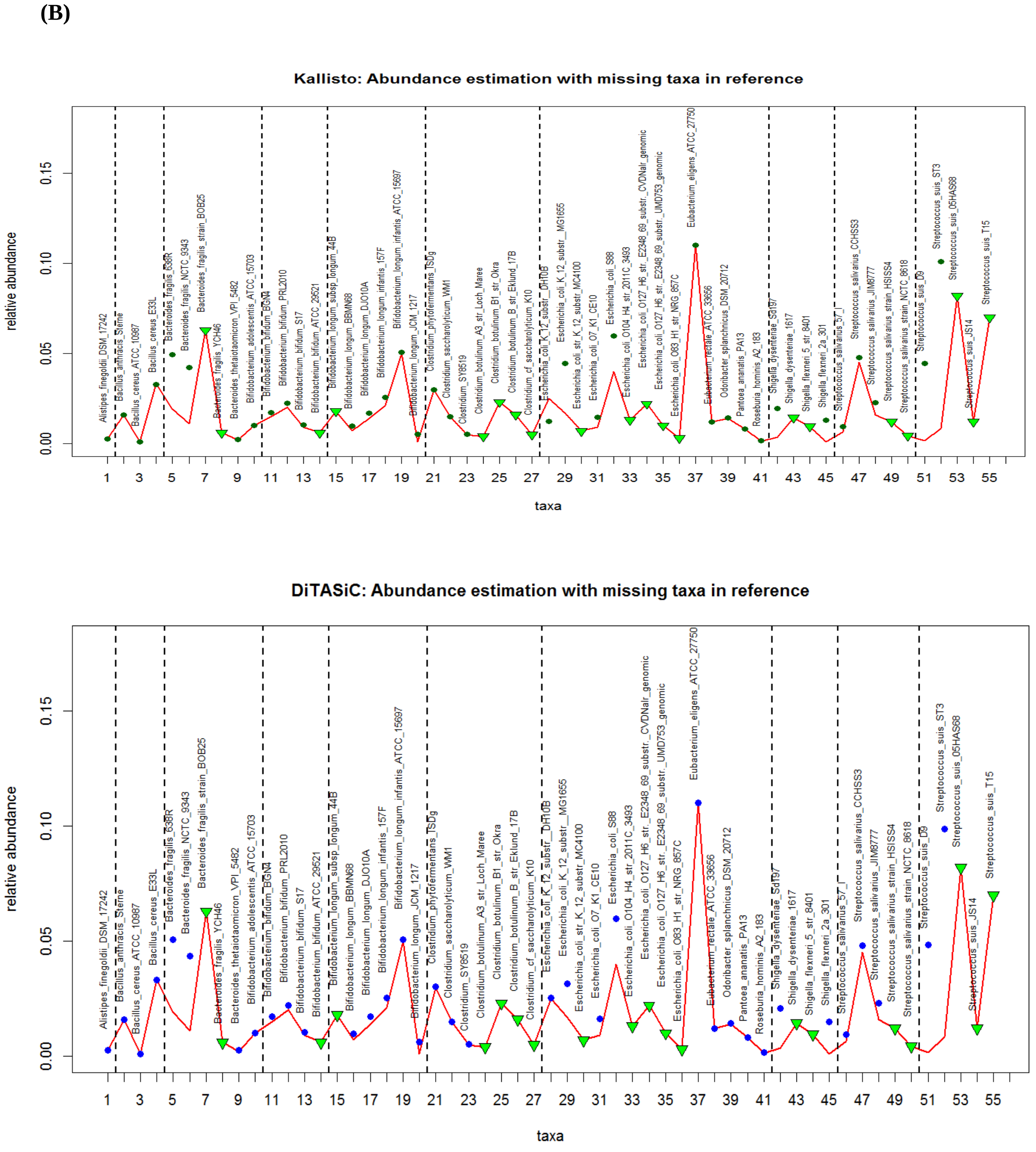
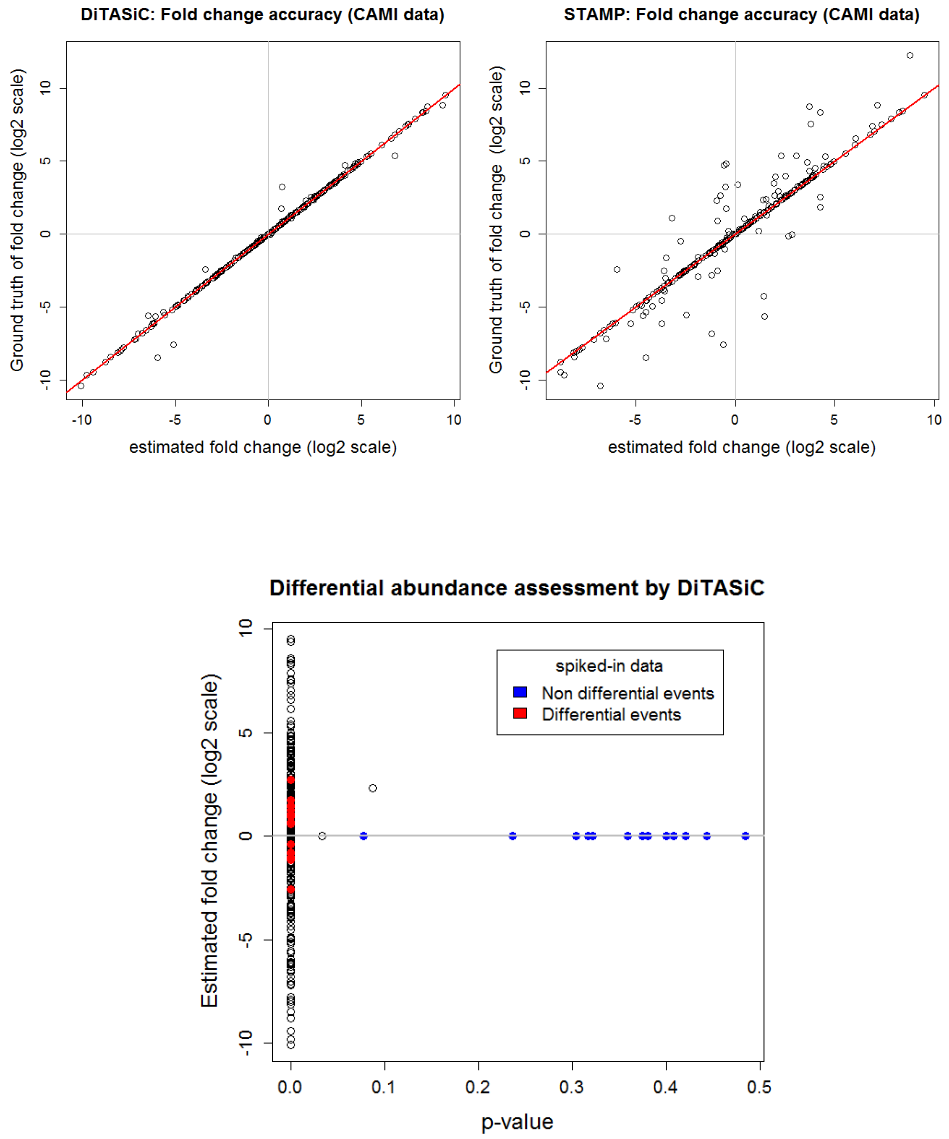


Fig. C 12: Impact of missing taxa in a reference set on the abundance estimation. The red line refers to the ground truth values, points refer to the abundance estimates obtained by the corresponding tool, while triangles mark absent taxa. Vertical lines are drawn to define sections of strain clusters. In this study**,** reads derived from 55 taxa are contrasted to a reduced reference set of 35 taxa to investigate the impact of missing taxa in a selected reference set. One consequence is that 11% of reads are not aligned and therefore are eliminated from the subsequent model calculations; second, the abundances estimated for some taxa are overestimated by the tools. However, a closer look reveals that it always concerns closely related strains which show an increased abundance due to missing strains within their cluster. However, no overall abundance bias is observed. Noticeable, while DiTASiC only exhibits abundance overestimations, kallisto also shows underestimation and overestimations within one cluster to compensate for missing taxa.

Fig. C 13:

(a)



(b)

Fig. C 13: **(a) Fold change accuracy achieved by DiTASiC in comparison to fold change accuracy obtained by STAMP in the CAMI data, and (b) differential abundance assessment using p-values by DiTASiC.** (a) Fold change estimates are proven to be highly accurate for DiTASiC with an *SSE* 19 times smaller compared to the STAMP output. This is depicted in the plots by fold change estimates found on the diagonal for DiTASiC, while many estimates are divergent from the diagonal in the plot by STAMP. (b) Computed p-values by the statistical framework in DiTASiC prove to clearly separate the spiked-in non-differential and differential taxa. Other taxa of the data set, holding fold change values greater than zero, also receive very small p-values stating differential abundance, but cannot be further confirmed here.

Table C1

	Simulation Data			FAMeS Data (Pignatelli et al.)		
	Sim 1	Sim 2	Sim 3	LC	MC	HC
# absent taxa (TN)	**25**	**21**	**19**	**10**	**12**	**10**
# false-positives (FP)	**0**	**0**	**0**	**0**	**0**	**0**
# present taxa (TP)	**10**	**14**	**16**	**112**	**110**	**112**
# false-negatives (FN)	**0**	**0**	**0**	**1**	**0**	**0**
Sensitivity	**1**	**1**	**1**	**0.991**	**1**	**1**
Specificity	**1**	**1**	**1**	**1**	**1**	**1**

Table C 1: **Detection of absent taxa**. We tested the detection performance with different proportions of absent taxa included, namely in the simulation group (1) and the FAMeS data sets. In the simulation data of group (1) only 28%, 40% and 45% taxa out of the 35 provided references are abundant in the data. Absolute numbers of absent and present taxa of each data set are reported in this table as well as absolute numbers of false-positive or false-negative detections. DiTASiC achieves exact detections, resulting in a sensitivity and specificity of 100%. The proportion of absent taxa in the FAMeS data refers to 8%, 9% and 8% based on the reference set of 122 taxa overall. A sensitivity and specificity of 100% is again reported for DiTASiC for the MC and HC data. In the LC set a reduced sensitivity is caused by one missed abundant taxon.

Table C2

Sample comparison	**DiTASiC**	**STAMP**
FAMeS: LC vs. MC	0.0047	0.5089
FAMeS: LC vs. HC	0.0013	0.4992
FAMeS: MC vs. HC	0.0051	0.0986
CAMI : S1 vs S2	25.07	476.91

Table C 2: ***SSE* values of fold change accuracy obtained by DiTASiC in comparison to STAMP in different sample comparisons.** *SSE* values of DiTASiC are significantly smaller compared to the ones computed by STAMP, indicating the importance of read ambiguity resolution and integration of abundance estimate uncertainties for differential abundance analysis.

Data Set Description

1) Simulation Data:

Nine data sets comprise 35 reference genomes from bacterial strains downloaded from NCBI, two additional data sets (set 10,11) were extended by further strain and sub-strain sequences (total 55 reference genomes) to create a high strain cluster density.

Each data set consists of 750,000 reads of 100bp length simulated by Mason (Holtgrewe, 2010), following Illumina read characteristics with default parameter settings. Reads are simulated according to the following abundance profiles.

Mason parameters:

- Total number of simulated reads: 750000
- Read length: 100 bp
- Replicate study Sim 4/5: Default.seed = 2048 (Sim 4), seed = 22 (Sim 5)

Taxa abundance list (1):

Taxa Name	**GenBank accession number**	**Ground Truth: relative taxa abundance**							
		Group (1)			**Group (2)**				
		Sim 1	**Sim 2**	**Sim 3**	**Sim 4 ~5**	**Sim 6**	**Sim 7**	**Sim 8**	**Sim 9**
Alistipes finegoldii DSM 17242	GCF_000265365.1	0.01	0.01	0	0.01	0.01	0.005	0.001	0.01
Bacillus anthracis str. Sterne	GCF_000008165.1	0.02	0.04	0.02	0.013	0.025	0.025	0.002	0.013
Bacillus cereus ATCC 10987	GCA_000008005.1	0.3	0.2	0.15	0.005	0.005	0.009	0.003	0.005
Bacillus cereus E33L	GCA_000011625.1	0.2	0.25	0	0.015	0.01	0.01	0.008	0.019
Bacteroides fragilis 638R	GCA_000210835.1	0	0	0	0.01	0.01	0.15	0.004	0.005
Bacteroides fragilis NCTC 9343	GCA_000025985.1	0	0.01	0.05	0.008	0.015	0.015	0.13	0.008
Bacteroides thetaiotaomicron VPI-5482	GCA_000011065.1	0	0	0	0.02	0.02	0.01	0.17	0.02
Bifidobacterium adolescentis ATCC 15703	GCA_000010425.1	0	0.02	0.02	0.003	0.009	0.009	0.009	0.003
Bifidobacterium bifidum BGN4	GCA_000265095.1	0	0	0	0.006	0.006	0.002	0.002	0.006
Bifidobacterium bifidum PRL2010	GCA_000165905.1	0.21	0.18	0.1	0.014	0.02	0.011	0.011	0.014
Bifidobacterium bifidum S17	GCA_000164965.1	0	0.01	0	0.007	0.007	0.03	0.025	0.014
Bifidobacterium longum BBMN68	GCA_000166315.1	0.14	0.14	0.05	0.15	0.008	0.008	0.008	0.15
Bifidobacterium longum DJO10A	GCA_000008945.1	0	0	0	0.02	0.02	0.003	0.02	0.02
Bifidobacteriumlongum infantis 157F	GCA_000196575.1	0	0	0.03	0.1	0.05	0.05	0.005	0.1
Bifidobacterium longum infantis ATCC 15697	GCA_000020425.1	0	0	0	0.01	0.01	0.014	0.006	0.02
Bifidobacterium longum JCM 1217	GCA_000196555.1	0	0.01	0.01	0.007	0.03	0.03	0.007	0.007
Clostridium phytofermentans ISDg	GCA_000018685.1	0	0	0	0.015	0.015	0.015	0.015	0.015
Clostridium saccharolyticum WM1	GCA_000144625.1	0.08	0	0.04	0.005	0.005	0.08	0.08	0.005
Clostridium SY8519	GCA_000270305.1	0.02	0.02	0.21	0.01	0.004	0.004	0.008	0.01
Escherichia coli K-12 substr. DH10B	GCA_000019425.1	0	0	0	0.008	0.008	0.035	0.035	0.008
Escherichia coli K-12 substr. MG1655	GCA_000005845.1	0	0	0	0.02	0.011	0.011	0.011	0.02
Escherichia coli O7:K1 str. CE10	GCA_000227625.1	0	0	0.12	0.013	0.009	0.013	0.001	0.013
Escherichia coli S88	GCA_000026285.1	0	0	0	0.006	0.006	0.006	0.11	0.006
Eubacterium eligens ATCC 27750	GCA_000146185.1	0	0	0	0.02	0.015	0.015	0.01	0.02
Eubacterium rectale ATCC 33656	GCA_000020605.1	0.01	0	0.04	0.007	0.007	0.012	0.012	0.017
Odoribacter splanchnicus DSM 20712	GCA_000190535.1	0.01	0.03	0.05	0.13	0.007	0.007	0.007	0.13
Pantoea ananatis PA13	GCA_000233595.1	0	0	0	0.04	0.04	0.04	0.016	0.04
Roseburia hominis A2-183	GCA_000225345.1	0	0	0	0.015	0.015	0.026	0.003	0.015
Shigella dysenteriae Sd197	GCA_000012005.1	0	0	0.02	0.003	0.022	0.022	0.13	0.003
Shigella flexneri 2a str. 301	GCA_000006925.2	0	0	0	0.018	0.018	0.002	0.017	0.032
Streptococcus salivarius 57.I	GCA_000305335.1	0	0.03	0.08	0.03	0.1	0.005	0.018	0.03
Streptococcus salivarius CCHSS3	GCA_000253335.1	0	0	0	0.009	0.18	0.18	0.009	0.009
Streptococcus salivarius JIM8777	GCA_000253315.1	0	0	0	0.12	0.12	0.12	0.002	0.08
Streptococcus suis D9	GCA_000231885.1	0	0.05	0.01	0.033	0.033	0.022	0.005	0.033
Streptococcus suis ST3	GCA_000204625.1	0	0	0	0.1	0.13	0.004	0.1	0.1

Taxa abundance list (2):

Taxa Name	**GenBank accession number**	**Ground Truth: relative taxa abundance Group (3): Sim 10**	**Ground Truth: relative taxa abundance Group (3): Sim 11**
Alistipes_finegoldii_DSM_17242	GCF_000265365.1	0.01	0.0025
Bacillus_anthracis_Sterne	GCF_000008165.1	0.013	0.016
Bacillus_cereus_ATCC_10987	GCA_000008005.1	0.005	0.001
Bacillus_cereus_E33L	GCA_000011625.1	0.015	0.033
Bacteroides_fragilis_638R	GCA_000210835.1	0.01	0.019
Bacteroides_fragilis_NCTC_9343	GCA_000025985.1	0.008	0.011
Bacteroides_fragilis_strain_BOB25	GCA_000965785.1	0	0.063
Bacteroides_fragilis_YCH46	GCA_000009925.1	0	0.006
Bacteroides_thetaiotaomicron_VPI_5482	GCA_000011065.1	0.008	0.002
Bifidobacterium_adolescentis_ATCC_15703	GCA_000010425.1	0.02	0.01
Bifidobacterium_bifidum_BGN4	GCA_000265095.1	0.003	0.015
Bifidobacterium_bifidum_PRL2010	GCA_000165905.1	0.006	0.02
Bifidobacterium_bifidum_S17	GCA_000164965.1	0.014	0.009
Bifidobacterium_bifidum_ATCC_29521	GCA_001025135.1	0	0.006
Bifidobacterium_longum_subsp_longum_44B	GCA_000261265.1	0	0.018
Bifidobacterium_longum_BBMN68	GCA_000166315.1	0.15	0.007
Bifidobacterium_longum_DJO10A	GCA_000008945.1	0.02	0.014
Bifidobacterium_longum_infantis_157F	GCA_000196575.1	0.1	0.021
Bifidobacterium_longum_infantis_ATCC_15697	GCA_000020425.1	0.01	0.05
Bifidobacterium_longum_JCM_1217	GCA_000196555.1	0.007	0.001
Clostridium_phytofermentans_ISDg	GCA_000018685.1	0.015	0.03
Clostridium_saccharolyticum_WM1	GCA_000144625.1	0.005	0.015
Clostridium_SY8519	GCA_000270305.1	0.01	0.005
Clostridium_botulinum_A3_str_Loch_Maree	GCA_000019545.1	0	0.004
Clostridium_botulinum_B1_str_Okra	GCA_000019305.1	0	0.023
Clostridium_botulinum_B_str_Eklund_17B	GCA_000307125.1	0	0.016
Clostridium_cf_saccharolyticum_K10	GCA_000210535.1	0	0.005
Escherichia_coli_K_12_substr__DH10B	GCA_000019425.1	0.008	0.025
Escherichia_coli_K_12_substr__MG1655	GCA_000005845.1	0.02	0.017
Escherichia_coli_str_K_12_substr_MC4100	GCA_000499485.1	0	0.007
Escherichia_coli_O7_K1_CE10	GCA_000227625.1	0.013	0.009
Escherichia_coli_S88	GCA_000026285.1	0.006	0.04
Escherichia_coli_O104_H4_str_2011C_3493	GCA_000299455.1	0	0.013
Escherichia_coli_O127_H6_str._E2348_69_substr._CVDNalr_genomic	GCA_000442065.2	0	0.022
Escherichia_coli_O127_H6_str._E2348_69_substr._UMD753_genomic	GCA_000442085.2	0	0.01
Escherichia_coli_O83_H1_str_NRG_857C	GCA_000183345.1	0	0.003
Eubacterium_eligens_ATCC_27750	GCA_000146185.1	0.02	0.11
Eubacterium_rectale_ATCC_33656	GCA_000020605.1	0.007	0.012
Odoribacter_splanchnicus_DSM_20712	GCA_000190535.1	0.13	0.014
Pantoea_ananatis_PA13	GCA_000233595.1	0.04	0.008
Roseburia_hominis_A2_183	GCA_000225345.1	0.015	0.0014
Shigella_dysenteriae_Sd197	GCA_000012005.1	0.003	0.0035
Shigella_dysenteriae_1617	GCA_000497505.1	0	0.0144
Shigella_flexneri_5_str_8401	GCA_000013585.1	0	0.0095
Shigella_flexneri_2a_301	GCA_000006925.2	0.018	0.001
Streptococcus_salivarius_57_I	GCA_000305335.1	0.03	0.0065
Streptococcus_salivarius_CCHSS3	GCA_000253335.1	0.009	0.045
Streptococcus_salivarius_JIM8777	GCA_000253315.1	0.12	0.016
Streptococcus_salivarius_strain_HSISS4	GCA_000448685.2	0	0.012
Streptococcus_salivarius_strain_NCTC_8618	GCA_000785515.1	0	0.0042
Streptococcus_suis_D9	GCA_000231885.1	0.033	0.0015
Streptococcus_suis_ST3	GCA_000204625.1	0.1	0.0085
Streptococcus_suis_05HAS68	GCA_000168355.3	0	0.082
Streptococcus_suis_JS14	GCA_000186405.1	0	0.012
Streptococcus_suis_T15	GCA_000494895.1	0	0.07

2) CAMI Data set:

Within the CAMI challenge (https://data.cami-challenge.org) (Sczyrba *et al.*, 2017), we selected a benchmark data set of medium complexity, which is provided for testing tools, with a ground truth of taxa proportions being available (‘*2. Toy Test Dataset Medium_Complexity’***)**. It comprises two 15 gb samples each holding about 150 million paired-end reads of 100 bp length based on HiSeq sequencing. A total of 225 bacterial and archaea genomes are present in both samples. Different clusters of strains with high sequence similarities are present within the 128 genera and 199 species. The relative abundances of the taxa range from 0.00009% to 8% in a medium complexity environment with median values of 0.1% and 0.08% for the samples, respectively. Comparison of the two samples yields taxa fold changes with a large span from 0.0009 to 1024. However, no ground truth is given for differential abundance classification and only fold change accuracy can be evaluated.

Therefore we extend the data set by simulating spike-in data: we selected 30 new strains from genera already present in the original set. A total of 20 million reads per sample are simulated from the new references based on a defined abundance given for each sample. Simulations are conducted using Mason (Holtgrewe, 2010), with error profiles matching the original reads, and are subsequently merged with the original set. Abundances of the added taxa are defined such that a ground truth of 15 differential and 15 non-differential events is created for additional differential assessment.

Mason parameters:

- Total number of simulated reads: N.sim = 20,000,000
- Read length: 100 bp
- Seed: 22
- Mismatch probability (begin): 0.005
- Mismatch probability (avrg): 0.01
- Mismatch probability (end): 0.03

* Mismatch probabilities are assessed by a pre-processing script which conducts a quick read-subset mapping for an approximate mismatch inference (refer to DiTASiC manual)

Merge ‘simulated set’ with ‘original set’:

Total number of reads (original CAMI set): N.org = 149,136,946
Factor = N.org / (N.org ~ N.sim) = 0.882
→ Relative abundance values (ground truth) of original CAMI reads are normalized by *Factor*
→ Relative abundance values (ground truth) for simulated reads created to sum up to (1-*Factor*) = 0.118

Taxa abundance list of the simulated 30 taxa (spiked into original CAMI set):

GenBank accession number	Ground Truth: relative taxa abundance for sample 1 ~ 2			
	Set 1	Set 1 – normalized values for Mason Simulation	Set 2	Set 2 – normalized values for Mason Simulation
GCA_900094705.1	0.005	0.04237288	0.005	0.04237288
GCF_000020965.1	0.01	0.08474576	0.005	0.04237288
GCF_000222305.1	0.0072	0.061016947	0.0072	0.061016947
GCF_000333455.1	0.003	0.025423728	0.0045	0.038135592
GCF_000385945.1	0.0015	0.012711864	0.0015	0.012711864
GCF_000428765.1	0.004	0.033898304	0.0023	0.019491525
GCF_000429685.1	0.013	0.110169488	0.013	0.110169488
GCF_000463735.1	0.003	0.025423728	0.0017	0.014406779
GCF_000470655.1	0.0055	0.046610168	0.0055	0.046610168
GCF_000471625.1	0.002	0.016949152	0.0048	0.040677965
GCF_000585495.1	0.0082	0.069491523	0.0082	0.069491523
GCF_000716525.1	0.001	0.008474576	0.0028	0.023728813
GCF_000817975.1	0.004	0.033898304	0.004	0.033898304
GCF_001298525.1	0.0033	0.027966101	0.0063	0.053389829
GCF_001402715.1	0.002	0.016949152	0.002	0.016949152
GCF_001418395.1	0.0066	0.055932202	0.003	0.025423728
GCF_001418715.1	0.004	0.033898304	0.004	0.033898304
GCF_001484195.1	0.0024	0.020338982	0.005	0.04237288
GCF_001485005.1	0.0015	0.012711864	0.0015	0.012711864

GCF_001514055.1	0.012	0.101694912	0.002	0.016949152
GCF_001514495.1	0.0044	0.037288134	0.0044	0.037288134
GCF_001544695.1	0.001	0.008474576	0.0027	0.022881355
GCF_001546055.1	0.003	0.025423728	0.003	0.025423728
GCF_001591345.1	0.0025	0.02118644	0.004	0.033898304
GCF_001591385.1	0.0013	0.011016949	0.0013	0.011016949
GCF_001592205.1	0.002	0.016949152	0.0015	0.012711864
GCF_001606025.1	0.0017	0.014406779	0.0017	0.014406779
GCF_001636425.1	0.0007	0.005932203	0.0023	0.019491525
GCF_001720585.1	0.001	0.008474576	0.0066	0.055932202
GCF_900044055.2	0.0012	0.010169523	0.0012	0.010169523
Sum:	0.118	1	0.118	1

3) Illumina 100 (*i100*) data set by Mende et al. (2012):

We applied the *i100* benchmark data set provided in the publication by Mende *et al.*, consisting of a total of 53.33 million single reads (~26.6 million paired reads) of 75 bp length following Illumina read characteristics. The reads are derived from 100 unique bacterial genomes and were originally simulated by the iMESSi metagenomics simulator.

Reads:
According to the publication, we retrieved the paired read sample ‘illumina_100species.1.fq’ and ‘‘illumina_100species.2.fq’ from the link: http://www.bork.embl.de/~mende/simulated_data/

Reference sequences:
We refer to Table2 *(Genomes Used in the Medium Complexity Metagenome and Estimated Coverage (100 genomes))* of the Supplementary Material of Mende *et al*. As stated in their description, the dataset includes all chromosomes of the genomes as well as all plasmids. Chromosome and additional plasmids sequences were retrieved according to the provided accessions for the i100 data available from http://www.bork.embl.de/~mende/simulated_data/bacterial_data.txt.

Ground Truth of Abundance Proportions:
We refer to a slightly corrected version of the i100 ground truth table provided by Schaeffer et al. (2017), named ‘i100_truth.csv’ available from https://github.com/pachterlab/metakallisto. The table follows the format s*pecies*, *abundance*, *counts*, and *genome size*. Thereby, ‘*counts*’ corresponds to the column ‘*Est_proportion of total sequence’* of the table by Mende et al. with minor corrections. The given ‘*counts*’ are used as ground truth (named *GT.counts*).

DiTASiC calculation
parameters used for the matrix calculation (default settings), defined parameters:
- Read length: 75 bp
- Mismatch probability (begin): 0.007
- Mismatch probability (avrg): 0.013
- Mismatch probability (end): 0.036

* Mismatch probabilities are assessed by a pre-processing script which conducts a quick read-subset mapping for an approximate mismatch inference (refer to DiTASiC manual)
Note: DiTASiC uses the reads as single end reads

kallisto calculation
kallisto quant command, only parameter: - l 75 (length)
Note: kallisto is run in paired end read mode

Evaluation

Parameter outputs	**kallisto** (*paired mode*)	**DiTASiC** (*single mode*)
n (number of taxa (exact genome level))	100	100
T (number of reads processed; see also in .json output file of kallisto)	26667004	53334008
A (number of aligned reads)	26202326	46516552
µ (true absolute counts, ground truth GT)	'*GT.counts*' (see description above, sum(GT.counts) = T)	[*GT.counts* / sum (*GT.counts*)] * T (scaled for the number of single reads)
t (absolute count estimate)	kallisto count estimates	DiTASiC count estimates

$$AVGRE = \frac{1}{n}\sum_{i}^{n}\frac{\left|t_i * \frac{T}{A} - \mu_i\right|}{\mu_i} \qquad RRMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}\left(\frac{\left|t_i * \frac{T}{A} - \mu_i\right|}{\mu_i}\right)^2}$$

$$SSE = \sum_{i=1}^{n}\left(\frac{t_i}{A} - \frac{\mu_i}{T}\right)^2$$

Evaluation values computed:

	Exact Genome level		
	AVGRE	RRMSE	SSE
DiTASiC	**0.86**	**2.19**	**8.23 e-06**
kallisto (reproduced)	1.09 *	5.38 *	5.62 e-05

Compare to Table 1 provided in the publication by Schaeffer *et al.:*

	Exact Genome level	
	AVGRE	RRMSE
kallisto	**0.97 ***	**5.42 ***
Bracken	-	-
CLARK	-	-
GASiC	7.21	19.31
eXpress	2.57	11.92

- CLARK and Bracken results are reported by Schaeffer *et al.* to be missing as "they do not output strain level counts."

* Evaluation values of kallisto reproduced in our computed i100 study and evaluation values given by Schaeffer *et al.* are shown to be very similar. The minor value differences observed might be explained by minor changes in reference sequences of new or older versions available in NCBI. (NCBI download of this study: 03/15/2017)

Bibliography

[1] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *Science (New York, N.Y.)*, vol. 312, pp. 212–217, Apr. 2006.

[2] J. Armengaud, "Microbiology and proteomics, getting the best of both worlds!," *Environmental Microbiology*, vol. 15, pp. 12–23, Jan. 2013.

[3] S.-E. Ong and M. Mann, "Mass spectrometry-based proteomics turns quantitative," *Nature Chemical Biology*, vol. 1, pp. 252–262, Oct. 2005.

[4] R. Bumgarner, "DNA microarrays: Types, Applications and their future," *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, vol. 0 22, pp. Unit–22.1., Jan. 2013.

[5] S. Eliuk and A. Makarov, "Evolution of Orbitrap Mass Spectrometry Instrumentation," *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, vol. 8, pp. 61–80, 2015.

[6] M. L. Metzker, "Sequencing technologies — the next generation," *Nature Reviews Genetics*, vol. 11, pp. 31–46, Jan. 2010.

[7] E. R. Mardis, "Next-Generation Sequencing Platforms," *Annual Review of Analytical Chemistry*, vol. 6, no. 1, pp. 287–303, 2013.

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

[9] T. C. Walther and M. Mann, "Mass spectrometry-based proteomics in cell biology," *The Journal of Cell Biology*, vol. 190, pp. 491–500, Aug. 2010.

[10] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, "The next-generation sequencing revolution and its impact on genomics," *Cell*, vol. 155, pp. 27–38, Sept. 2013.

[11] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, "Advanced sequencing technologies: methods and goals," *Nature Reviews Genetics*, vol. 5, no. 5, pp. 335–344, 2004.

[12] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS computational biology*, vol. 6, p. e1000667, Feb. 2010.

[13] L. McHugh and J. W. Arthur, "Computational Methods for Protein Identification from Mass Spectrometry Data," *PLoS Computational Biology*, vol. 4, Feb. 2008.

[14] S. Thankaswamy-Kosalai, P. Sen, and I. Nookaew, “Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics,” *Genomics*, vol. 109, pp. 186–191, July 2017.

[15] L. Gatto, K. D. Hansen, M. R. Hoopmann, H. Hermjakob, O. Kohlbacher, and A. Beyer, “Testing and Validation of Computational Methods for Mass Spectrometry,” *Journal of Proteome Research*, Nov. 2015.

[16] P. Mallick and B. Kuster, “Proteomics: a pragmatic perspective,” *Nature Biotechnology*, vol. 28, pp. 695–709, July 2010.

[17] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, “Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present,” *Analytical and Bioanalytical Chemistry*, vol. 404, pp. 939–965, Sept. 2012.

[18] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray,” *Science*, vol. 270, pp. 467–470, Oct. 1995.

[19] A. Hatem, D. Bozdağ, A. E. Toland, and m. V. Çatalyürek, “Benchmarking short sequence mapping tools,” *BMC Bioinformatics*, vol. 14, p. 184, June 2013.

[20] B. Langmead, “Aligning short sequencing reads with Bowtie,” *Current protocols in bioinformatics*, vol. 11, p. 7, Dec. 2010.

[21] D. E. Wood and S. L. Salzberg, “Kraken: ultrafast metagenomic sequence classification using exact alignments,” *Genome Biology*, vol. 15, p. R46, Mar. 2014.

[22] S. Nayfach, B. Rodriguez-Mueller, N. Garud, and K. S. Pollard, “An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography,” *Genome Research*, vol. 26, pp. 1612–1625, Nov. 2016.

[23] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, pp. 57–63, Jan. 2009.

[24] A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos, “Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies,” *Bioinformatics and Biology Insights*, vol. 9, pp. 75–88, 2015.

[25] P. Hingamp, N. Grimsley, S. G. Acinas, C. Clerissi, L. Subirana, J. Poulain, I. Ferrera, H. Sarmento, E. Villar, G. Lima-Mendez, K. Faust, S. Sunagawa, J.-M. Claverie, H. Moreau, Y. Desdevises, P. Bork, J. Raes, C. de Vargas, E. Karsenti, S. Kandels-Lewis, O. Jaillon, F. Not, S. Pesant, P. Wincker, and H. Ogata, “Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes,” *The ISME journal*, vol. 7, pp. 1678–1695, Sept. 2013.

[26] C. M. Liu, M. Aziz, S. Kachur, P.-R. Hsueh, Y.-T. Huang, P. Keim, and L. B. Price, "BactQuant: an enhanced broad-coverage bacterial quantitative real-time PCR assay," *BMC microbiology*, vol. 12, p. 56, Apr. 2012.

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

[28] M. Vaudel, A. Sickmann, and L. Martens, "Peptide and protein quantification: A map of the minefield," *PROTEOMICS*, vol. 10, pp. 650–670, Feb. 2010.

[29] J. Rappsilber, U. Ryder, A. I. Lamond, and M. Mann, "Large-scale proteomic analysis of the human spliceosome," *Genome Research*, vol. 12, pp. 1231–1245, Aug. 2002.

[30] Y. Ishihama, Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber, and M. Mann, "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein," *Molecular & cellular proteomics: MCP*, vol. 4, pp. 1265–1272, Sept. 2005.

[31] D. W. Powell, C. M. Weaver, J. L. Jennings, K. J. McAfee, Y. He, P. A. Weil, and A. J. Link, "Cluster analysis of mass spectrometry data reveals a novel component of SAGA," *Molecular and Cellular Biology*, vol. 24, pp. 7249–7259, Aug. 2004.

[32] F. Blondeau, B. Ritter, P. D. Allaire, S. Wasiak, M. Girard, N. K. Hussain, A. Angers, V. Legendre-Guillemin, L. Roy, D. Boismenu, R. E. Kearney, A. W. Bell, J. J. M. Bergeron, and P. S. McPherson, "Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 3833–3838, Mar. 2004.

[33] 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," *Mol Cell Proteomics*, vol. 1, pp. 376–86, May 2002.

[34] M. Miyagi and K. C. S. Rao, "Proteolytic 18o-labeling strategies for quantitative proteomics," *Mass Spectrometry Reviews*, vol. 26, pp. 121–136, Feb. 2007.

[35] 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. 994–999, Oct. 1999.

[36] 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,” *Mol Cell Proteomics*, vol. 3, pp. 1154–69, Dec. 2004.

[37] A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed, and C. Hamon, “Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS,” *Anal Chem*, vol. 75, pp. 1895–904, Apr. 2003.

[38] S. L. Sanders, J. Jennings, A. Canutescu, A. J. Link, and P. A. Weil, “Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry,” *Molecular and Cellular Biology*, vol. 22, pp. 4723–4738, July 2002.

[39] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, “Global quantification of mammalian gene expression control,” *Nature*, vol. 473, pp. 337–342, May 2011.

[40] C. Ludwig, M. Claassen, A. Schmidt, and R. Aebersold, “Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry,” *Molecular & cellular proteomics: MCP*, vol. 11, p. M111.013987, Mar. 2012.

[41] A. Sandberg, R. M. M. Branca, J. Lehtiö, and J. Forshed, “Quantitative accuracy in mass spectrometry based proteomics of complex samples: The impact of labeling and precursor interference,” *Journal of Proteomics*, vol. 96, pp. 133–144, Jan. 2014.

[42] M. Bantscheff, M. Boesche, D. Eberhard, T. Matthieson, G. Sweetman, and B. Kuster, “Robust and Sensitive iTRAQ Quantification on an LTQ Orbitrap Mass Spectrometer,” *Molecular & Cellular Proteomics : MCP*, vol. 7, pp. 1702–1713, Sept. 2008.

[43] S. Y. Ow, M. Salim, J. Noirel, C. Evans, and P. C. Wright, “Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation,” *Proteomics*, vol. 11, pp. 2341–2346, June 2011.

[44] C. D. Wenger, M. V. Lee, A. S. Hebert, G. C. McAlister, D. H. Phanstiel, M. S. Westphall, and J. J. Coon, “Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging,” *Nature Methods*, vol. 8, pp. 933–935, Nov. 2011.

[45] M. M. Savitski, T. Mathieson, N. Zinn, G. Sweetman, C. Doce, I. Becher, F. Pachl, B. Kuster, and M. Bantscheff, “Measuring and Managing Ratio Compression for Accurate iTRAQ/TMT Quantification,” *Journal of Proteome Research*, vol. 12, pp. 3586–3598, Aug. 2013.

[46] M. M. Savitski, G. Sweetman, M. Askenazi, J. A. Marto, M. Lang, N. Zinn, and M. Bantscheff, "Delayed fragmentation and optimized isolation width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers," *Analytical Chemistry*, vol. 83, pp. 8959–8967, Dec. 2011.

[47] F. P. Breitwieser, A. Muller, L. Dayon, T. Kocher, A. Hainard, P. Pichler, U. Schmidt-Erfurth, G. Superti-Furga, J. C. Sanchez, K. Mechtler, K. L. Bennett, and J. Colinge, "General statistical modeling of data from protein relative expression isobaric tags," *J Proteome Res*, vol. 10, pp. 2758–66, June 2011.

[48] L. Ting, R. Rad, S. P. Gygi, and W. Haas, "MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics," *Nature Methods*, vol. 8, pp. 937–940, Nov. 2011.

[49] N. Rauniyar and J. R. Yates, "Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics," *Journal of Proteome Research*, vol. 13, no. 12, pp. 5293–5309, 2014.

[50] M. E. Monroe, N. Tolić, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith, "VIPER: an advanced software package to support high-throughput LC-MS peptide identification," *Bioinformatics*, vol. 23, pp. 2021–2023, Aug. 2007.

[51] I. P. Shadforth, T. P. Dunkley, K. S. Lilley, and C. Bessant, "i-Tracker: For quantitative proteomics using iTRAQ$^{\mathrm{TM}}$," *BMC Genomics*, vol. 6, p. 145, Oct. 2005.

[52] A. M. Boehm, S. Pütz, D. Altenhöfer, A. Sickmann, and M. Falk, "Precise protein quantification based on peptide quantification using iTRAQ," *BMC bioinformatics*, vol. 8, p. 214, 2007.

[53] W. X. Schulze and M. Mann, "A Novel Proteomic Screen for Peptide-Protein Interactions," *Journal of Biological Chemistry*, vol. 279, pp. 10756–10764, Dec. 2004.

[54] D. K. Han, J. Eng, H. Zhou, and R. Aebersold, "Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry," *Nature Biotechnology*, vol. 19, pp. 946–951, Oct. 2001.

[55] L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.-Y. Brusniak, O. Vitek, R. Aebersold, and M. Müller, "SuperHirn – a novel tool for high resolution LC-MS-based peptide/protein profiling," *PROTEOMICS*, vol. 7, pp. 3470–3480, Oct. 2007.

[56] Y. Benjamini and T. P. Speed, "Summarizing and correcting the GC content bias in high-throughput sequencing," *Nucleic Acids Research*, vol. 40, p. e72, May 2012.

[57] X. Yang, S. P. Chockalingam, and S. Aluru, “A survey of error-correction methods for next-generation sequencing,” *Briefings in Bioinformatics*, vol. 14, pp. 56–66, Jan. 2013.

[58] L. Ilie, F. Fazayeli, and S. Ilie, “HiTEC: accurate error correction in high-throughput sequencing data,” *Bioinformatics (Oxford, England)*, vol. 27, pp. 295–302, Feb. 2011.

[59] W.-C. Kao, A. H. Chan, and Y. S. Song, “ECHO: a reference-free short-read error correction algorithm,” *Genome Research*, vol. 21, pp. 1181–1192, July 2011.

[60] D. R. Kelley, M. C. Schatz, and S. L. Salzberg, “Quake: quality-aware detection and correction of sequencing errors,” *Genome Biology*, vol. 11, no. 11, p. R116, 2010.

[61] L. Salmela, “Correction of sequencing errors in a mixed set of reads,” *Bioinformatics (Oxford, England)*, vol. 26, pp. 1284–1290, May 2010.

[62] J. Schröder, H. Schröder, S. J. Puglisi, R. Sinha, and B. Schmidt, “SHREC: a short-read error correction method,” *Bioinformatics (Oxford, England)*, vol. 25, pp. 2157–2163, Sept. 2009.

[63] D. M. O’Sullivan, T. Laver, S. Temisak, N. Redshaw, K. A. Harris, C. A. Foy, D. J. Studholme, and J. F. Huggett, “Assessing the accuracy of quantitative molecular microbial profiling,” *International Journal of Molecular Sciences*, vol. 15, pp. 21476–21491, Nov. 2014.

[64] B. Cooper, J. Feng, and W. M. Garrett, “Relative, label-free protein quantitation: spectral counting error statistics from nine replicate MudPIT samples,” *Journal of the American Society for Mass Spectrometry*, vol. 21, pp. 1534–1546, Sept. 2010.

[65] C. Bauer, F. Kleinjung, D. Rutishauser, C. Panse, A. Chadt, T. Dreja, H. Al-Hasani, K. Reinert, R. Schlapbach, and J. Schuchhardt, “PPINGUIN: Peptide Profiling Guided Identification of Proteins improves quantitation of iTRAQ ratios,” *BMC Bioinformatics*, vol. 13, p. 34, Feb. 2012.

[66] N. A. Karp, W. Huber, P. G. Sadowski, P. D. Charles, S. V. Hester, and K. S. Lilley, “Addressing Accuracy and Precision Issues in iTRAQ Quantitation,” *Molecular & Cellular Proteomics*, vol. 9, pp. 1885–1897, Jan. 2010.

[67] J. Hu, J. Qian, O. Borisov, S. Pan, Y. Li, T. Liu, L. Deng, K. Wannemacher, M. Kurnellas, C. Patterson, S. Elkabes, and H. Li, “Optimized proteomic analysis of a mouse model of cerebellar dysfunction using amine-specific isobaric tags,” *Proteomics*, vol. 6, pp. 4321–34, Aug. 2006.

[68] L. Hultin-Rosenberg, J. Forshed, R. M. M. Branca, J. Lehtiö, and H. J. Johansson, “Defining, comparing, and improving iTRAQ quantification in

mass spectrometry proteomics data," *Molecular & cellular proteomics: MCP*, vol. 12, pp. 2021–2031, July 2013.

[69] W.-T. Lin, W.-N. Hung, Y.-H. Yian, K.-P. Wu, C.-L. Han, Y.-R. Chen, Y.-J. Chen, T.-Y. Sung, and W.-L. Hsu, "Multi-Q: A Fully Automated Tool for Multiplexed Protein Quantitation," *Journal of Proteome Research*, vol. 5, no. 9, pp. 2328–2338, 2006.

[70] C. Zhou, M. J. Walker, A. J. Williamson, A. Pierce, C. Berzuini, C. Dive, and A. D. Whetton, "A hierarchical statistical modeling approach to analyze proteomic isobaric tag for relative and absolute quantitation data," *Bioinformatics*, vol. 30, pp. 549–58, Feb. 2014.

[71] Y. Zhang, Z. Wen, M. P. Washburn, and L. Florens, "Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins," *Analytical Chemistry*, vol. 82, pp. 2272–2281, Mar. 2010.

[72] Y. Liu, F. Ripp, R. Koeppel, H. Schmidt, S. Lukas Hellmann, M. Weber, C. F. Krombholz, B. Schmidt, and T. Hankeln, "AFS: identification and quantification of species composition by metagenomic sequencing," *Bioinformatics*, 2017.

[73] M. S. Lindner and B. Y. Renard, "Metagenomic abundance estimation and diagnostic testing on species level," *Nucleic Acids Research*, vol. 41, p. e10, Jan. 2013.

[74] L. C. Xia, J. A. Cram, T. Chen, J. A. Fuhrman, and F. Sun, "Accurate Genome Relative Abundance Estimation Based on Shotgun Metagenomic Reads," *PLOS ONE*, vol. 6, p. e27992, Dec. 2011.

[75] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology*, vol. 34, pp. 525–527, May 2016.

[76] A. I. Nesvizhskii and R. Aebersold, "Interpretation of Shotgun Proteomic Data The Protein Inference Problem," *Molecular & Cellular Proteomics*, vol. 4, pp. 1419–1440, Jan. 2005.

[77] K. Helsens, E. Timmerman, J. Vandekerckhove, K. Gevaert, and L. Martens, "Peptizer, a Tool for Assessing False Positive Peptide Identifications and Manually Validating Selected Results," *Molecular & Cellular Proteomics*, vol. 7, pp. 2364–2372, Jan. 2008.

[78] S. Jin, D. S. Daly, D. L. Springer, and J. H. Miller, "The Effects of Shared Peptides on Protein Quantitation in Label-Free Proteomics by LC/MS/MS," *Journal of Proteome Research*, vol. 7, pp. 164–169, Jan. 2008.

[79] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M.

Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, C. Bowler, C. d. Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork, "Structure and function of the global ocean microbiome," *Science*, vol. 348, p. 1261359, May 2015.

[80] F. E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, vol. 11, pp. 1–21, Feb. 1969.

[81] S. P, M, E. J, and G. M, "A Comparison of Six Methods for Missing Data Imputation," *Journal of Biometrics & Biostatistics*, vol. 6, May 2015.

[82] 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. 520–525, June 2001.

[83] 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. 2088–2096, Nov. 2003.

[84] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies," *Journal of Proteome Research*, vol. 15, pp. 1116–1125, Apr. 2016.

[85] C. S. Gan, P. K. Chong, T. K. Pham, and P. C. Wright, "Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ)," *Journal of Proteome Research*, vol. 6, pp. 821–827, Feb. 2007.

[86] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, pp. 185–193, Jan. 2003.

[87] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. L. Gall, B. Schaëffer, S. L. Crom, M. Guedj, and F. Jaffrézic, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Briefings in Bioinformatics*, Sept. 2012.

[88] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, Feb. 2010.

[89] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani, “Normalization, testing, and false discovery rate estimation for RNA-sequencing data,” *Biostatistics*, vol. 13, pp. 523–538, Jan. 2012.

[90] D. Chelius and P. V. Bondarenko, “Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry,” *Journal of Proteome Research*, vol. 1, pp. 317–323, Aug. 2002.

[91] Z. Fang, J. Martin, and Z. Wang, “Statistical methods for identifying differentially expressed genes in RNA-Seq experiments,” *Cell & Bioscience*, vol. 2, p. 26, July 2012.

[92] G. K. Smyth, “Linear models and empirical bayes methods for assessing differential expression in microarray experiments,” *Statistical Applications in Genetics and Molecular Biology*, vol. 3, p. Article3, 2004.

[93] 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.

[94] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, pp. 139–140, Jan. 2010.

[95] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 3, p. R106, 2010.

[96] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[97] H. Pimentel, N. L. Bray, S. Puente, P. Melsted, and L. Pachter, “Differential analysis of RNA-seq incorporating quantification uncertainty,” *Nature Methods*, vol. 14, pp. 687–690, July 2017.

[98] V. Jonsson, T. Österlund, O. Nerman, and E. Kristiansson, “Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics,” *BMC genomics*, vol. 17, p. 78, Jan. 2016.

[99] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop, “Differential abundance analysis for microbial marker-gene surveys,” *Nature Methods*, vol. 10, pp. 1200–1202, Dec. 2013.

[100] M. B. Sohn, R. Du, and L. An, “A robust approach for identifying differentially abundant features in metagenomic samples,” *Bioinformatics*, vol. 31, pp. 2269–2275, July 2015.

[101] X. Peng, G. Li, and Z. Liu, “Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data,” *Journal of Computational Biology*, vol. 23, no. 2, pp. 102–110, 2015.

[102] J. Wang, L. Li, T. Chen, J. Ma, Y. Zhu, J. Zhuang, and C. Chang, "In-depth method assessments of differentially expressed protein detection for shotgun proteomics data with missing values," *Scientific Reports*, vol. 7, p. 3367, June 2017.

[103] 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. 631–644, Apr. 2008.

[104] 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*, vol. 129, pp. 83–92, Nov. 2015.

[105] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, no. 6, pp. 65–70, 1979.

[106] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. Series B (Methodological), pp. 289–300, 1995.

[107] W. H. Dunham, M. Mullin, and A.-C. Gingras, "Affinity purification coupled to mass spectrometry: Basic principles and strategies," *PROTEOMICS*, vol. 12, no. 10, pp. 1576–1590, 2012.

[108] A. I. Nesvizhskii, "Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments," *PROTEOMICS*, vol. 12, no. 10, pp. 1639–1655, 2012.

[109] Y. V. Miteva, H. G. Budayeva, and I. M. Cristea, "Proteomics-Based Methods for Discovery, Quantification, and Validation of Protein–Protein Interactions," *Analytical Chemistry*, vol. 85, pp. 749–768, Jan. 2013.

[110] H. Choi, B. Larsen, Z.-Y. Lin, A. Breitkreutz, D. Mellacheruvu, D. Fermin, Z. S. Qin, M. Tyers, A.-C. Gingras, and A. I. Nesvizhskii, "SAINT: probabilistic scoring of affinity purification-mass spectrometry data," *Nature Methods*, vol. 8, no. 1, pp. 70–73, 2011.

[111] P. L. Auer and R. W. Doerge, "A two-stage Poisson model for testing RNA-Seq data."," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, pp. 1–26, 2011.

[112] R. Fisher, *The Design of Experiments.* Edinburgh: Oliver and Boyd, 1935.

[113] Y. Lai, "Conservative adjustment of permutation p-values when the number of permutations is limited," *Int. J. Bioinformatics Res. Appl.*, vol. 3, no. 4, pp. 536–546, 2007.

[114] H. Yang and G. Churchill, "Estimating p-values in small microarray experiments," *Bioinformatics (Oxford, England)*, vol. 23, pp. 38–43, Jan. 2007.

[115] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for P-value adjustment.* John Wiley Sons, Jan. 1993.

[116] C. Hundertmark, R. Fischer, T. Reinl, S. May, F. Klawonn, and L. Jänsch, "MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1004–1011, Apr. 2009.

[117] D. W. Mahoney, T. M. Therneau, C. J. Heppelmann, L. Higgins, L. M. Benson, R. M. Zenka, P. Jagtap, G. L. Nelsestuen, H. R. Bergen, and A. L. Oberg, "Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides," *Journal of Proteome Research*, vol. 10, pp. 4325–4333, Sept. 2011.

[118] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Droege, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jorgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, L. H. Hansen, S. J. Sorensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. D. Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Goeker, N. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy, "Critical Assessment of Metagenome Interpretation − a benchmark of computational metagenomics software," *bioRxiv*, p. 099127, Jan. 2017.

[119] M. Pardo and J. S. Choudhary, "Assignment of Protein Interactions from Affinity Purification/Mass Spectrometry Data," *Journal of Proteome Research*, vol. 11, no. 3, pp. 1462–1474, 2012.

[120] I. M. Armean, K. S. Lilley, and M. W. B. Trotter, "Popular Computational Methods to Assess Multiprotein Complexes Derived From Label-Free Affinity Purification and Mass Spectrometry (AP-MS) Experiments," *Molecular & Cellular Proteomics*, vol. 12, pp. 1–13, Jan. 2013.

[121] O. Rinner, L. N. Mueller, M. Hubálek, M. Müller, M. Gstaiger, and R. Aebersold, "An integrated mass spectrometric and computational framework for the analysis of protein interaction networks," *Nature Biotechnology*, vol. 25, no. 3, pp. 345–352, 2007.

[122] N. C. Hubner and M. Mann, "Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC)," *Methods*, vol. 53, pp. 453–459, Apr. 2011.

[123] M. E. Sowa, E. J. Bennett, S. P. Gygi, and J. W. Harper, "Defining the Human Deubiquitinating Enzyme Interaction Landscape," *Cell*, vol. 138, pp. 389–403, July 2009.

[124] M. E. Sardiu, Y. Cai, J. Jin, S. K. Swanson, R. C. Conaway, J. W. Conaway, L. Florens, and M. P. Washburn, “Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics,” *Proceedings of the National Academy of Sciences*, vol. 105, pp. 1454–1459, May 2008.

[125] H. Choi, T. Glatter, M. Gstaiger, and A. I. Nesvizhskii, “SAINT-MS1: Protein–Protein Interaction Scoring Using Label-free Intensity Data in Affinity Purification-Mass Spectrometry Experiments,” *Journal of Proteome Research*, vol. 11, pp. 2619–2624, Apr. 2012.

[126] R. A. Bradshaw, A. L. Burlingame, S. Carr, and R. Aebersold, “Reporting protein identification data: the next generation of guidelines,” *Molecular & cellular proteomics: MCP*, vol. 5, pp. 787–788, May 2006.

[127] B. Y. Renard, W. Timm, M. Kirchner, J. A. J. Steen, F. A. Hamprecht, and H. Steen, “Estimating the confidence of peptide identifications without decoy databases,” *Analytical chemistry*, vol. 82, pp. 4314–4318, June 2010.

[128] S. Wagner, L. Königsmaier, M. Lara-Tejero, M. Lefebre, T. C. Marlovits, and J. E. Galán, “Organization and coordinated assembly of the type III secretion export apparatus,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 17745–17750, Oct. 2010.

[129] D. Mellacheruvu, Z. Wright, A. L. Couzens, J.-P. Lambert, N. A. St-Denis, T. Li, Y. V. Miteva, S. Hauri, M. E. Sardiu, T. Y. Low, V. A. Halim, R. D. Bagshaw, N. C. Hubner, A. Al-Hakim, A. Bouchard, D. Faubert, D. Fermin, W. H. Dunham, M. Goudreault, Z.-Y. Lin, B. G. Badillo, T. Pawson, D. Durocher, B. Coulombe, R. Aebersold, G. Superti-Furga, J. Colinge, A. J. R. Heck, H. Choi, M. Gstaiger, S. Mohammed, I. M. Cristea, K. L. Bennett, M. P. Washburn, B. Raught, R. M. Ewing, A.-C. Gingras, and A. I. Nesvizhskii, “The CRAPome: a contaminant repository for affinity purification-mass spectrometry data,” *Nature methods*, vol. 10, pp. 730–736, Aug. 2013.

[130] 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, p. S5, Nov. 2012.

[131] M. D. Robinson and Alicia Oshlack, “A scaling normalization method for differential expression analysis of RNA-seq data,” *Genome Biology*, vol. 11, no. 3, p. R25, 2010.

[132] R. Bourgon, R. Gentleman, and W. Huber, “Independent filtering increases detection power for high-throughput experiments,” *Proceedings of the National Academy of Sciences*, vol. 107, pp. 9546–9551, May 2010.

[133] R. Gentleman, V. Carey, W. Huber, and F. Hahne, “genefilter: methods for filtering genes from microarray experiments.”

[134] H. Choi, S. Kim, A.-C. Gingras, and A. I. Nesvizhskii, "Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data," *Molecular systems biology*, vol. 6, p. 385, June 2010.

[135] A. Breitkreutz, H. Choi, J. R. Sharom, L. Boucher, V. Neduva, B. Larsen, Z.-Y. Lin, B.-J. Breitkreutz, C. Stark, G. Liu, J. Ahn, D. Dewar-Darch, T. Reguly, X. Tang, R. Almeida, Z. S. Qin, T. Pawson, A.-C. Gingras, A. I. Nesvizhskii, and M. Tyers, "A Global Protein Kinase and Phosphatase Interaction Network in Yeast," *Science*, vol. 328, pp. 1043–1046, May 2010.

[136] D. V. Skarra, M. Goudreault, H. Choi, M. Mullin, A. I. Nesvizhskii, A.-C. Gingras, and R. E. Honkanen, "Label-free quantitative proteomics and SAINT analysis enable interactome mapping for the human Ser/Thr protein phosphatase 5," *Proteomics*, vol. 11, pp. 1508–1516, Apr. 2011.

[137] 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, no. 6, pp. 1212–1222, 2010.

[138] O. J. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.

[139] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, pp. 800–802, 1988.

[140] R. D. C. Team, *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, 2012.

[141] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher, "OpenMS - an open-source software framework for mass spectrometry," *BMC bioinformatics*, vol. 9, p. 163, 2008.

[142] 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. 2373–2385, Dec. 2008.

[143] D. Büttner, "Protein export according to schedule: architecture, assembly, and regulation of type III secretion systems from plant- and animal-pathogenic bacteria," *Microbiology and molecular biology reviews: MMBR*, vol. 76, pp. 262–310, June 2012.

[144] T. Kubori, A. Sukhan, S. I. Aizawa, and J. E. Galán, "Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, pp. 10225–10230, Aug. 2000.

[145] T. C. Marlovits, T. Kubori, M. Lara-Tejero, D. Thomas, V. M. Unger, and J. E. Galán, “Assembly of the inner rod determines needle length in the type III secretion injectisome,” *Nature*, vol. 441, pp. 637–640, June 2006.

[146] J. A. Rosenzweig, G. Weltman, G. V. Plano, and K. Schesser, “Modulation of yersinia type three secretion system by the S1 domain of polynucleotide phosphorylase,” *The Journal of biological chemistry*, vol. 280, pp. 156–163, Jan. 2005.

[147] T. Tobe, C. Sasakawa, N. Okada, Y. Honma, and M. Yoshikawa, “vacB, a novel chromosomal gene required for expression of virulence genes on the large plasmid of Shigella flexneri,” *Journal of bacteriology*, vol. 174, pp. 6359–6367, Oct. 1992.

[148] D. M. Anderson and O. Schneewind, “A mRNA signal for the type III secretion of Yop proteins by Yersinia enterocolitica,” *Science (New York, N.Y.)*, vol. 278, pp. 1140–1143, Nov. 1997.

[149] B. Blaylock, J. A. Sorg, and O. Schneewind, “Yersinia enterocolitica type III secretion of YopR requires a structure in its mRNA,” *Molecular microbiology*, vol. 70, pp. 1210–1222, Dec. 2008.

[150] K. Ito and Y. Akiyama, “Cellular functions, mechanism of action, and regulation of FtsH protease,” *Annual review of microbiology*, vol. 59, pp. 211–231, 2005.

[151] M. Kirchner, B. Y. Renard, U. Kothe, D. J. Pappin, F. A. Hamprecht, H. Steen, and J. A. Steen, “Computational protein profile similarity screening for quantitative mass spectrometry experiments,” *Bioinformatics*, vol. 26, pp. 77–83, Jan. 2010.

[152] J. M. Burkhart, M. Vaudel, R. P. Zahedi, L. Martens, and A. Sickmann, “iTRAQ protein quantification: a quality-controlled workflow,” *Proteomics*, vol. 11, pp. 1125–1134, Mar. 2011.

[153] S. Y. Ow, M. Salim, J. Noirel, C. Evans, I. Rehman, and P. C. Wright, “iTRAQ Underestimation in Simple and Complex Mixtures: “The Good, the Bad and the Ugly”,” *Journal of Proteome Research*, vol. 8, pp. 5347–5355, Nov. 2009.

[154] 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*, vol. 18, no. suppl_1, pp. S96–S104, 2002.

[155] G. Onsongo, M. D. Stone, S. K. Van Riper, J. Chilton, B. Wu, L. Higgins, T. C. Lund, J. V. Carlis, and T. J. Griffin, “LTQ-iQuant: A freely available software pipeline for automated and accurate protein quantification of isobaric tagged peptide data from LTQ instruments,” *Proteomics*, vol. 10, pp. 3533–8, Oct. 2010.

[156] L. Choe, M. D'Ascenzo, N. R. Relkin, D. Pappin, P. Ross, B. Williamson, S. Guertin, P. Pribil, and K. H. Lee, "8-Plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease," *PROTEOMICS*, vol. 7, no. 20, pp. 3651–3660, 2007.

[157] X.-j. Li, H. Zhang, J. A. Ranish, and R. Aebersold, "Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry," *Analytical chemistry*, vol. 75, no. 23, pp. 6648–6657, 2003.

[158] B. Carrillo, C. Yanofsky, S. Laboissiere, R. Nadon, and R. E. Kearney, "Methods for combining peptide intensities to estimate relative protein abundance," *Bioinformatics*, vol. 26, pp. 98–103, Jan. 2010.

[159] M. J. Tenga and I. M. Lazar, "Impact of Peptide Modifications on iTRAQ Quantitation Accuracy," *Analytical chemistry*, vol. 83, pp. 701–707, Feb. 2011.

[160] V. A. Fusaro, D. R. Mani, J. P. Mesirov, and S. A. Carr, "Prediction of high-responding peptides for targeted protein assays by mass spectrometry," *Nature Biotechnology*, vol. 27, pp. 190–198, Feb. 2009.

[161] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search," *Analytical chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002.

[162] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and M. J. MacCoss, "Semi-supervised learning for peptide identification from shotgun proteomics datasets," *Nature Methods*, vol. 4, pp. 923–925, Nov. 2007.

[163] L. Gatto and K. S. Lilley, "MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation," *Bioinformatics (Oxford, England)*, vol. 28, pp. 288–289, Jan. 2012.

[164] J. C. Silva, M. V. Gorenstein, G.-Z. Li, J. P. Vissers, and S. J. Geromanos, "Absolute quantification of proteins by lcmse a virtue of parallel ms acquisition," *Molecular & Cellular Proteomics*, vol. 5, no. 1, pp. 144–156, 2006.

[165] B. C. Searle, "Scaffold: a bioinformatic tool for validating ms/ms-based proteomic studies," *Proteomics*, vol. 10, no. 6, pp. 1265–1269, 2010.

[166] F. Mosteller and J. W. Tukey, "Data analysis and regression: a second course in statistics.," *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

[167] S. Van Huffel and J. Vandewalle, *The total least squares problem: computational aspects and analysis*, vol. 9. Siam, 1991.

[168] G. Neelakanta and H. Sultana, “The use of metagenomic approaches to analyze changes in microbial communities,” *Microbiology Insights*, vol. 6, pp. 37–48, 2013.

[169] T. Nawy, “MICROBIOLOGY: The strain in metagenomics,” *Nature Methods*, vol. 12, p. 1005, Nov. 2015.

[170] B. J. Shapiro, J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabó, M. F. Polz, and E. J. Alm, “Population genomics of early events in the ecological differentiation of bacteria,” *Science (New York, N.Y.)*, vol. 336, pp. 48–51, Apr. 2012.

[171] M. J. Rosen, M. Davison, D. Bhaya, and D. S. Fisher, “Microbial diversity. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche,” *Science (New York, N.Y.)*, vol. 348, pp. 1019–1023, May 2015.

[172] T. D. Lieberman, K. B. Flett, I. Yelin, T. R. Martin, A. J. McAdam, G. P. Priebe, and R. Kishony, “Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures,” *Nature Genetics*, vol. 46, pp. 82–87, Jan. 2014.

[173] E. S. Snitkin, A. M. Zelazny, C. I. Montero, F. Stock, L. Mijares, NISC Comparative Sequence Program, P. R. Murray, and J. A. Segre, “Genome-wide recombination drives diversification of epidemic strains of Acinetobacter baumannii,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 13758–13763, Aug. 2011.

[174] S. Lindgreen, K. L. Adair, and P. P. Gardner, “An evaluation of the accuracy and speed of metagenome analysis tools,” *Scientific Reports*, vol. 6, p. 19233, Jan. 2016.

[175] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, “Metagenomic microbial community profiling using unique clade-specific marker genes,” *Nature Methods*, vol. 9, pp. 811–814, June 2012.

[176] M. Scholz, D. V. Ward, E. Pasolli, T. Tolio, M. Zolfo, F. Asnicar, D. T. Truong, A. Tett, A. L. Morrow, and N. Segata, “Strain-level microbial epidemiology and population genomics from shotgun metagenomics,” *Nature Methods*, vol. 13, no. 5, pp. 435–438, 2016.

[177] C. Luo, R. Knight, H. Siljander, M. Knip, R. J. Xavier, and D. Gevers, “ConStrains identifies microbial strains in metagenomic datasets,” *Nature Biotechnology*, vol. 33, pp. 1045–1052, Oct. 2015.

[178] H. Li, “Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis,” *Annual Review of Statistics and Its Application*, vol. 2, no. 1, pp. 73–94, 2015.

[179] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, pp. 377–386, Mar. 2007.

[180] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter, "Pseudoalignment for metagenomic read assignment," *Bioinformatics*, Feb. 2017.

[181] J. R. White, N. Nagarajan, and M. Pop, "Statistical methods for detecting differentially abundant features in clinical metagenomic samples," *PLoS computational biology*, vol. 5, p. e1000352, Apr. 2009.

[182] D. H. Parks, G. W. Tyson, P. Hugenholtz, and R. G. Beiko, "STAMP: statistical analysis of taxonomic and functional profiles," *Bioinformatics (Oxford, England)*, vol. 30, pp. 3123–3124, Nov. 2014.

[183] D. H. Parks and R. G. Beiko, "Identifying biologically relevant differences between metagenomic communities," *Bioinformatics (Oxford, England)*, vol. 26, pp. 715–721, Mar. 2010.

[184] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower, "Metagenomic biomarker discovery and explanation," *Genome Biology*, vol. 12, p. R60, June 2011.

[185] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed, "Gut metagenome in European women with normal, impaired and diabetic glucose control," *Nature*, vol. 498, pp. 99–103, June 2013.

[186] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome Biology*, vol. 15, p. R29, Feb. 2014.

[187] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, "Host lifestyle affects human microbiota on daily timescales," *Genome Biology*, vol. 15, p. R89, 2014.

[188] S. M. Gibbons and J. A. Gilbert, "Microbial diversity–exploration of natural ecosystems and microbiomes," *Current Opinion in Genetics & Development*, vol. 35, pp. 66–72, Dec. 2015.

[189] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers," *BMC genomics*, vol. 16, p. 236, Mar. 2015.

[190] P. Menzel, K. L. Ng, and A. Krogh, "Fast and sensitive taxonomic classification for metagenomics with Kaiju," *Nature Communications*, vol. 7, p. 11257, Apr. 2016.

[191] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy, "Mash: fast genome and metagenome distance estimation using MinHash," *Genome Biology*, vol. 17, p. 132, 2016.

[192] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, Mar. 2012.

[193] M. Holtgrewe, "Mason – A Read Simulator for Second Generation Sequencing Data," *Technical Report FU Berlin*, Oct. 2010.

[194] M. Pignatelli and A. Moya, "Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data," *PLoS ONE*, vol. 6, no. 5, p. e19984, 2011.

[195] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides, "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nature Methods*, vol. 4, pp. 495–500, June 2007.

[196] D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork, "Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data," *PLOS ONE*, vol. 7, p. e31386, Feb. 2012.

[197] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, "Bracken: estimating species abundance in metagenomics data," *PeerJ Computer Science*, vol. 3, p. e104, Jan. 2017.

[198] L. Dethlefsen and D. A. Relman, "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108 Suppl 1, pp. 4554–4561, Mar. 2011.

[199] B. Teng, C. Zhao, X. Liu, and Z. He, "Network inference from AP-MS data: computational challenges and solutions," *Briefings in Bioinformatics*, vol. 16, pp. 658–674, July 2015.

Zusammenfassung

Die Fortschritte der Hochdurchsatz-Technologien in der Genomik und Proteomik haben das biologische Forschungsfeld revolutioniert. Die verbesserte technische Auflösung hat den Fokus auf quantitative Analysen verstärkt und die hohe Parallelisierbarkeit in den Instrumenten ermöglicht nun die Quantifizierung gesamter Genome und Proteome. Das Forschungsfeld bietet eine Vielfalt an Anwendungen, welche mit vielen bioinformatischen Herausforderungen verbunden sind und mit einem großen Bedarf an neuen Quantifizierungs-Analyse Programmen einhergeht.Quantitative Analysen sind komplex. Sie umfassen zahlreiche Schritte, angefangen von der Probenaufbereitung, der Daten-Erfassung, das Daten-Prozessieren, bis zur finalen quantitativen Schätzung. Mehrere Schritte innerhalb des Prozesses können eine Verzerrung der quantitativen Werte verursachen. Die korrekten Mengen in einer biologischen Probe zu erfassen, bleibt eine schwierige Aufgabe und fordert ständig neue Methodenentwicklungen, um systematische Fehler und Verzerrungen in den Daten zu reduzieren.

In dieser Doktorarbeit werden neue bioinformatische Strategien zur Verbesserung der Quantifizierung von Daten aus Hochdurchsatz Anwendungen vorgestellt. Ziel der Arbeit ist es, systematische Fehler zu korrigieren und die Varianz von quantitativen Schätzungen zu minimieren. Dabei ist die Erfassung der potenziellen Fehlerquellen und der vorliegenden Datencharakteristiken entscheidend. Teil der Arbeit ist es gemeinsame Fehler und Lösungen verschiedener omik Analysen und Datentypen zu identifizieren. Ein weiteres Ziel ist es die Genauigkeit der quantitativen Schätzungen statistisch zu erfassen. Bei der Quantifizierung von Hochdurchsatz-Daten mangelt es grundlegend daran wie man die Güte quantitativer Schätzungen misst und angibt, dies gilt vor allem in der quantitativen Proteomforschung. Viele statistische Methoden für umfangreiche Datenanalysen wurden vor allem in der Microarray Zeit entwickelt. Grundlegend gilt, dass unabhängig von zugrundeliegenden Technologien, resultierende quantitative Werte aus statistischer Perspektive oft gleiche Eigenschaften haben. Es liegt ein großes Potenzial darin Parallelen zwischen den verschiedenen omik Feldern zu erfassen und etablierte statistische Methoden zu übertragen. Gleichermaßen wichtig ist es, jedoch, auch spezifische Datencharakteristiken und technisch bedingte Fehler zu erkennen und zu integrieren. Zusammengefasst: quantitative Analysen sind extrem heterogen und die Suche nach einer alles erfüllenden Methodik wäre nicht passend.

Die vorliegende Doktorarbeit umfasst drei Hauptprojekte, welche drei verschiedene biologische Fragestellungen und Datentypen von drei quantitativen Hochdurchsatz Techniken behandelt. Es werden neue Ansätze vorgestellt zur Prä-Prozessierung von Daten, zur quantitativen Inferenz und Auflösung von verzerrten Messungen, sowie Methoden für quantitative Vergleichsstudien.

Ziel des ersten Projektes ist die Identifikation von Protein-Protein Interaktionen unter Anwendung einer Affinitätschromatographie in Kombination mit einem Massenspektrometer (AP-MS). Dabei werden quantitative Mengen eines Proteins aus einem Pull-down Experiment mit Proteinmengen von negativen Kontrollexperimenten verglichen, mit dem Ziel echte Interaktionen von falsch-positiven Detektionen zu trennen. Gegenwärtige Methoden für AP-MS Analysen nutzen meist ein Punkteverfahren zum Ranking von potenziellen Interaktionsproteinen. Es gibt dabei jedoch keine Angabe wie der Cutoff zur Auswahl von Interaktions-Kandidaten am besten zu setzen ist und auch eine Einschätzung zur Anzahl von falsch-positiven Identifizierungen fehlt. Statistische Daten Prä- und Post-Prozessierung ist ein selten behandeltes Thema in AP-MS Analysen. In dieser Arbeit wird ein umfassend statistisches Rahmenwerk vorgestellt, welches um jedes Punkteverfahren gelegt werden kann und durch Anwendung eines Permutationsprinzips das Ersetzen von Punkten durch statistische P-Werte ermöglicht. Zusätzlich wird ein Zwei-Stufen Poisson Modell, welches von RNA-Seq Daten zu AP-MS Daten angepasst wird als alternative Methode zur Erfassung von Interaktionen vorgeschlagen. Für die Prä-Prozessierung werden verschiedene Normalisierungsmethoden und ein statistischer Filterprozess mit entsprechender Anpassung für AP-MS

Daten betrachtet. Verschiedene Experimente veranschaulichen wie die Detektion von wahren Interaktionen signifikant gesteigert werden kann, während gleichzeitig die Falsch-Detektions Rate kontrolliert wird.

Das zweite Projekt beschäftigt sich mit der genauen Schätzung von Proteinmengen. Bei einem Massenspektrometer werden die Messungen auf Peptid Spektrum Ebene durchgeführt. Obgleich man erwarten würde, dass all Peptid-Spektren die einem Protein zugeordnet werden ähnliche Intensitäts-Werte aufweisen, existiert in der Tat eine starke Werte Heterogenität. Diese Heterogenität entsteht aufgrund von zufallsbedingten und systematischen Fehlern. Intelligente Strategien zur Inferenz der zugrundeliegenden Proteinmengen sind gefragt. Aktuelle Methoden basieren fast ausschließlich auf quantitativen Informationen. Diese Arbeit vertritt die Hypothese, dass eine Fülle von weiteren Peptid Merkmalen verfügbar ist, die die Zuverlässigkeit von Spektren-Werten wiedergeben. Verschiedene Merkmale werden hier mit der beobachteten Varianz Heterogenität korreliert und ihr Zusammenhang mit der Wertegenauigkeit in Spektren erforscht. Als Ergebnis wird eine neue Peptide-Protein Inferenz Methode vorgestellt, welche als iPQF (isobaric Protein Quantification based on Features) bezeichnet wird. Die Methode integriert Peptide Merkmale zusammen mit quantitativen Werten für die Proteinquantifizierung. Die Wertung von Peptid Spektren entsprechend ihrer Merkmale ist neu. Eine umfangreiche Evaluierung von iPQF im Vergleich zu neun anderen Inferenz-Methoden belegt den Zugewinn der Merkmalsnutzung, um die Protein Quantifizierung zu verbessern.

NGS basierte Quantifizierung beruht ebenfalls auf der Messung vieler Sequenzfragmente und erfordert Methoden zur Zusammenfassung. Das dritte Projekt beschäftigt sich mit präziser Quantifizierung von Organismen aus metagenomischen Proben. Besondere Herausforderungen stellen sich bei Analysen von Sub-Spezies Ebenen aufgrund der Präsenz vieler sehr ähnlicher Referenzgenome. Diese führen aufgrund von Mehrfachzuordnungen von Reads zu einer starken Verzerrung in der Quantifizierung. Generell herrscht ein großes Interesse an feinerer Auflösung von mikrobiellen Proben, aber nur wenige Methoden erlauben eine tiefere quantitative Erfassung als die der Spezies-Ebene. In dieser Arbeit wird DiTASiC (Differential Taxa Abundance including Similarity Correction) als ein neues Tool zur Quantifizierung von Organismen und differentiellen Analyse in metagenomischen Proben vorgestellt, welches auf exaktem Genomlevel anwendbar ist. Ein neues generalisiertes lineares Modell zur Auflösung der Mehrfachzählungen der Reads wird eingeführt, welches zusätzlich einen Fehlerterm zur Abschätzung der Quantifizierung enthält. In einem neuen statistischen Ansatz wird die Quantifizierungs-Varianz integriert und quantitative Verteilungen abgeleitet, welches wichtig für ein differentielles Testen auf Sub-Spezies Ebene ist. Untersuchungen auf den neusten Testdaten zeigen präzise quantitative Schätzungen bis zu Sub-sub-Spezies Ebenen und eine verbesserte Detektion differentiell vorkommender Organismen.

Zusammengefasst tragen alle drei Projekte zur Verbesserung des gegenwärtigen Repertoires von bioinformatischen Methoden in der Hochdurchsatz-Quantifizierung von omik Daten bei. Die Arbeit verweist auf die Komplexität von Quantifizierungs-Analysen. Zum einen betont sie die umfassende Nutzbarkeit und den Transfer von etablierten statistischen Konzepten zwischen verschiedenen omiks Feldern. Gleichzeitig wird auf die Wichtigkeit verwiesen, zugrundeliegende Daten-Charakteristiken zu adressieren und auf die Notwendigkeit individuelle Strategien zu entwickeln, um eine hohe Quantifizierungs-Genauigkeit zu erreichen.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.

Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Martina Fischer, Berlin, 10.04.2018