MEDICAL UNIVERSITY OF VIENNA

Computational Approaches for Quantifying Proteins and Posttranslational Modifications from Labeled Mass Spectrometry Data

Doctoral thesis at the Medical University of Vienna
for obtaining the academic degree

Doctor of Philosophy

Submitted by

DI (FH) Florian Paul Breitwieser

Supervisor:
Jacques Colinge, PhD
Research Center for Molecular Medicine
of the Austrian Academy of Sciences
Bioinformatics Dept.

Vienna, 06/2014

Declaration

This thesis is a presentation of my original research work, which was done under the guidance of J. Colinge at the Research Center of Molecular Medicine, Vienna, Austria. Parts of the work stem from collaborative efforts and wouldn't have been possible without the contribution of others. The following text lists their contributions in detail.

This thesis is cumulative and presents as main results two publications which were previously published in a peer reviewed journal. The background chapter contains parts of a previously published book chapter. Furthermore, as detailed below, several further publications are described as additional outcomes in the result chapter.

Full citations of the first-author publications:

Book chapter — Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91

Paper I — Florian P. Breitwieser, André Müller, Loïc Dayon, Thomas Köcher, Alexandre Hainard, Peter Pichler, Ursula Schmidt-Erfurth, Giulio Superti-Furga, Jean-Charles Sanchez, Karl Mechtler, Keiryn L. Bennett, and Jacques Colinge (June 2011). "General statistical modeling of data from protein relative expression isobaric tags." eng. In: *J Proteome Res* 10.6, pp. 2758–2766

Paper II — Florian P. Breitwieser and Jacques Colinge (Sept. 2013). "Isobar(PTM): a software tool for the quantitative analysis of post-translationally modified proteins." eng. In: *J Proteomics* 90, pp. 77–84

Chapter 2 includes figures and ideas from the book chapter published by Breitwieser and Colinge (2012a). The manuscript of the book chapter was devised and written by the author of the thesis. All figures were generated by the author. J. Colinge contributed to the ideas and proof-reading of the text.

The main results (chapter 4) previously appeared in peer-reviewed journals:

Section 4.1 has been published by Breitwieser et al. (2011). The original project idea came from J. Colinge. The author of the thesis developed the concept, main ideas and implementation of the software package. The statistical concepts were developed jointly with J. Colinge. The author has written an initial version of the manuscript. J. Colinge devised the method for p-value combination for replicate analysis, provided crucial statistical input, and wrote the final versions

of the manuscript. A. Müller designed the test dataset and performed the mass spectrometric analysis of the spike-in datasets in the laboratory of K. L. Bennett and G. Superti-Furga. L. Dayon and A. Hainard generated and provided the human ceruspinal fluid data set in the laboratory of J.-C. Sanchez. T. Köcher and P. Pichler generated and provided the mouse ventricles data set in the laboratory of K. Mechtler. U. Schmidt-Erfuth provided the background samples for the test data set analysis. All authors contributed to the discussion of the results.

Section 4.2 has been published by Breitwieser and Colinge (2013). The author of this thesis designed and implemented the computational framework, wrote the initial versions of the manuscript and generated all figures. J. Colinge finalized the manuscript and provided critical input.

Section 4.3 describes additional outcomes of the thesis. Section 4.3.1 details the results obtained in biological and mass spectrometry research projects, in which the author of the thesis contributed in the analysis and presentation of the results; developing, applying and extending the methodology presented in sections 4.1 and 4.2. Section 4.3.2 presents an unpublished extension of the statistical methods, which improves on certain parts. The author devised the method and wrote the full text. J. Colinge provided input to the methods and writing. Full citations of the application papers:

Application I — Eric B. Haura, André Müller, Florian P. Breitwieser, Jiannong Li, Florian Grebien, Jacques Colinge, and Keiryn L. Bennett (Jan. 2011). "Using iTRAQ combined with tandem affinity purification to enhance low-abundance proteins associated with somatically mutated EGFR core complexes in lung cancer." eng. In: *J Proteome Res* 10.1, pp. 182–190

Application II — Keiryn L. Bennett, Marion Funk, Marion Tschernutter, Florian P. Breitwieser, Melanie Planyavsky, Ceereena Ubaida Mohien, André Müller, Zlatko Trajanoski, Jacques Colinge, Giulio Superti-Furga, and Ursula Schmidt-Erfurth (Feb. 2011). "Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ®-labelled specimens." eng. In: *J Proteomics* 74.2, pp. 151–166

Application III — André C. Müller, Florian P. Breitwieser, Heinz Fischer, Christopher Schuster, Oliver Brandt, Jacques Colinge, Giulio Superti-Furga, Georg Stingl, Adelheid Elbe-Bürger, and Keiryn L. Bennett (July 2012). "A comparative proteomic study of human skin suction blister fluid from healthy individuals using immunodepletion and iTRAQ labeling." eng. In: *J Proteome Res* 11.7, pp. 3715–3727

Application IV Georg E. Winter, Uwe Rix, Scott M. Carlson, Karoline V. Gleixner, Florian Grebien, Manuela Gridling, André C. Müller, Florian P. Breitwieser, Martin Bilban, Jacques Colinge, Peter Valent, Keiryn L. Bennett, Forest M. White, and Giulio Superti-Furga (Nov. 2012). "Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML.". eng. In: *Nat Chem Biol* 8.11, pp. 905–912

Application V Andreas Pollreisz, Marion Funk, Florian P. Breitwieser, Katja Parapatics, Stefan Sacu, Michael Georgopoulos, Roman Dunavoelgyi, Gerhard J. Zlabinger, Jacques Colinge, Keiryn L. Bennett, and Ursula Schmidt-Erfurth (Mar. 2013). "Quantitative proteomics of aqueous and vitreous fluid from patients with idiopathic epiretinal membranes." eng. In: *Exp Eye Res* 108, pp. 48–58

All further chapters of the thesis were written solely by the author. J. Colinge provided input to the writing.

Reprint permissions for the material integrated in this thesis which has been previously published were obtained from the respective publishers: Springer Science + Business Media granted the reuse and modification of the figs. 2.1, 2.4, 2.6 and 2.10 and tables 2.1 and 2.2 in chapter 2, originally published by Breitwieser and Colinge (2012a). The American Chemical Society granted the reprint of the full article in section 4.1 (originally published by Breitwieser et al., 2011), fig. 4.1 (originally published by Haura et al., 2011), and fig. 4.4a (originally published by Müller et al., 2012). Elsevier granted the reprint of the full article in section 4.2 (originally published by Breitwieser and Colinge, 2013), fig. 4.2 (originally published by Keiryn L. Bennett et al., 2011), and figs. 4.3a and 4.3b (originally published by Pollreisz et al., 2013).

Contents

Abstract

Proteomic technologies are a fundamentally important tools of biological and medical research. Modern mass spectrometric equipment enables researchers to identify thousands of proteins in biological samples in a matter of hours. For the quantitative comparison of protein content, isotope-coded mass labels are employed which mark the proteins of the respective samples. Especially popular are the isobaric methods iTRAQ and TMT which make the simultaneous quantitative comparison of up to 10 samples possible.

The data of these experiments are complex, high-dimensional, and noisy. Even though suitable statistical modeling and efficient software tools are pivotal for the success of the data analysis, few comprehensive and open bioinformatical analysis frameworks exist for quantitative proteomics. In this thesis, thus a software package and statistical framework are developed, which enable and facilitate the analysis of isobarically labeled mass spectrometric data.

The first part of the thesis describes statistical models for isobarically tagged data, which enable better inference by capturing technical variability in a intensity-dependent noise function, and biological variability with a heavy tailed distribution. The performance characteristics of this method were tested on especially prepared test datasets with spiked proteins at known ratios and unchanging background proteins. By resampling of the data, it could be demonstrated that the method both controls the rate of false positives and provides a good sensitivity in selecting true positives. Using additional biological datasets it further could be shown that the method works well with data from different types of mass spectrometers and setups.

The methods were implemented in a novel R package called `isobar`, which is part of the Bioconductor project. Along with the statistical framework, `isobar` implements methods for a complete quantitative workflow from mass spectrometric peaklists to protein quantification and analysis reports in PDF and XLS formats. Protein grouping is implemented within the package. The analysis can also be automated and the package thus integrated into existing pipelines. `isobar` was designed according to Bioconductor design principles and is implemented in the S4 class system.

The aforementioned methods and software were developed for the quantification of protein differences. Besides the protein expression change, differential post-translational modifications (PTMs) are prime modulator of protein function. PTMs are of great importance in meany research questions and can be identified and quantified with mass spectrometry. In the second part of the thesis, we thus extend the statistical models and R package for the quantitative PTM analysis. This includes the integration of modules for the localization of the PTM in the peptide sequence,

the correction of the modified peptide ratio with the protein ratio, and the creation of extended analysis reports with specific details for identified PTMs.

The methods and the software were applied and extended in several further publications. Furthermore, the `isobar` package is downloaded over 100 times per month from Bioconductor. In conclusion, this thesis contributes to the advancement of quantitative protein research with the development of novel bioinformatical software and methods.

German abstract

Proteomische Technologien sind fundamental wichtige Werkzeuge der biologischen und medizinischen Forschung. Mit modernen Massenspektrometern können Forscher innerhalb von wenigen Stunden tausende Proteine in biologischen Proben detektieren und quantifizieren. Für die Protein-Quantifizierung werden isotop-kodierte Labels verwendet, mit denen die Proteine der jeweiligen Proben markiert werden können. Besonders populär sind die isobaren Methoden iTRAQ und TMT, mit welchen bis zu 10 Proben in einem Experiment verglichen werden können.

Die Daten von diesen Experimenten haben eine komplexe Struktur, sind hoch-dimensional und verrauscht. Obwohl die passende statistische Modellierung und effiziente Software essentiell für den Erfolg der Datenanalyse sind, gibt auf dem Bereich der quantitativen Proteomik wenig umfassende und offene bioinformatische Analyseframeworks. In dieser Arbeit wird deshalb ein bioinformatisches Softwarepaket und statistisches Rahmenwerk entwickelt, welche die Analyse von quantitativen proteomischen Daten ermöglichen und erleichtern.

Der erste Teil dieser Arbeit beschreibt statistische Modelle welche eine bessere Inferenz für quantitative proteomische Experimente ermöglichen; durch die Modellierung der technischen Variabilität mit einer intensitätsabhängigen Varianzfunktion und der biologischen Variabilität mittels einer endlastigen Verteilung. Die Leistungsfähigkeit dieser Methode wurde an speziell erzeugten Test-Datensätzen getested, welche gleichbleibende Hintergrundproteine und eingemischten Proteine in bekannten Konzentrationen beinhaltet. Mittels Resampling konnte demonstriert werden, das die Methode sowohl die Rate der falsch-positiv selektierten Proteine kontrolliert, als auch eine gute Performanz im selektieren echt positiver Proteine hat. An weiteren biologischen Datensätzen wurde weiters gezeigt, dass die Methode mit unterschiedlichen Massenspektrometern und Setups funktioniert.

Die Modelle wurden in einem neuartigen R-Softwarepaket namens `isobar` implementiert, welches Teil des Bioconductor-Projekt ist. Zusammen mit dem statistischen Rahmenwerk implementiert `isobar` Methoden für einen kompletten Workflow von massenspektrometrischen Peaklisten zur Proteinquantifizierung und Analyseergebnissen im PDF und XLS Format. Protein-Gruppierung wird innerhalb des Paket implementiert. Eine Analyse kann automatisiert und in vorhandede Analyse-Pipelines integriert werden. `isobar` ist nach den Bioconductor Design-Prinzipien konzipiert und in dem objektorientierten S4 Klassensystem implementiert.

Die oben genannte Methoden und Software wurden für die Quantifizierung von Protein-Unterschiede entwickelt. Neben der unterschiedlichen Expression von Proteinen, sind post-translationale Modifikationen (PTM) zentrale Modulatoren der Proteinfunktion. PTMs sind von

großer Bedeutung in vielen Forschungsfragen, und können ebenfalls mit Massenspektrometrie identifiziert und quantifiziert werden. Im zweiten Teil der Arbeit werden deswegen die statistischen Modelle und das R-Paket für die quantitative PTM Analyse erweitert. Dies inkludiert die Integration von Modulen zur Lokalisierung der Modifikation in der Peptidsequenz, die Anpassung des PTM-Ratios mit Protein-Ratios, und das Erstellen von erweiterten Analyseberichten mit spezifischen Details zu identifizierten PTMs.

Die Methoden und die Software wurden in mehreren Publikationen angewendent und erweitert. Das `isobar`-Paket wird weiters über einhundert mal pro Monat über Bioconductor installiert. Abschließend kann gesagt werden, dass diese Arbeit mit neuer bioinformatischer Software und Methoden zur Weiterentwicklung der Proteinforschung mit iTRAQ und TMT beiträgt.

Acknowledgments

I would like to thank the people who made it possible for me to do the research for this thesis.

First, I want to express my gratitude to my adviser Jacques Colinge, who provided me with his guidance and advice throughout the process of writing this thesis. Thank you for your trust in giving me the opportunity to pursue this PhD project. It has been a challenging, substantial and important period of my life, which provided me with ample motivation to pursue science.

I also thank my thesis committee members Giulio Superti-Furga and Arndt van Haeseler for their advice.

During my time at CeMM I got involved in many interesting projects which manifested in various publications. Thank you to all my collaborators! I want to mention especially the labs of Keiryn Bennett and Giulio Superti-Furga. Further I want to thank the labs of Karl Mechtler at the IMP Vienna, and Jean-Charlez Sanchez at the University of Geneva, for valuable discussions and data for the first publication.

My colleagues in the Bioinformatics group provided me both company and assistance. Thank you for sharing this time of my life. I want to thank especially Alexey Stukalov, who infused my with some of his perfectionism.

Finally, I want to thank more generally all the people at CeMM. The atmosphere was always vibrant and stimulating for research. The diverse areas of research and the active aura of all of the people involved was one of my main initial motivations to go into science. Thank you to all people who have been part, and whom I was able to meet over the years.

1. Introduction

In the last decade, technological advances revolutionized the way in which research in molecular biology is conducted. Novel developments in high-throughput technologies enabled researchers to conduct biological investigations at previously unprecedented depths (Metzker, 2010; Mallick and Kuster, 2010). The genomic part of biological systems can be explored using DNA microarrays and next-generation sequencing; by sequencing the whole genome, mapping the DNA methylation state, or analyzing differential expression of gene transcripts (Cirulli and Goldstein, 2010; Bock, 2012; Oshlack, Robinson, and Young, 2010). The gene transcripts, or transcriptome, are often used as a proxy to look at the protein space, the proteome. However, the picture that the transcriptome can paint of the proteome is bleak and imprecise (Vogel and Marcotte, 2012). For one, protein abundances correlate only roughly with the transcript abundances. Furthermore, and arguably more importantly, the transcriptome can provide only an estimate of protein abundances. The proteome, however, has many more dimensions - such as protein localization, interactions and modifications - which are of great importance, and are controlled mainly by post-translational processes (Altelaar, Munoz, and Heck, 2013).

The driving technology behind large-scale protein investigations is mass spectrometry (MS, see Jürgen Cox and Mann (2011) and section 2.1). To identify proteins in a sample, usually the so-called shotgun approach is employed. In shotgun, or bottom-up proteomics, proteins are first digested to peptides, which are easier to analyze by MS. The peptides are separated, introduced into the MS and surveyed in the first MS dimension. Sequentially certain peptide ions are isolated, and analyzed by fragmenting its backbone and measuring the fragmentation spectra (second MS dimension) (Yaoyang Zhang et al., 2013). Based on the resulting data, peptides are identified and subsequently used to infer proteins (Colinge and Keiryn L Bennett, 2007; Claassen, 2012). MS technology is evolving at a rapid pace, and every year or two, the time needed to identify the constituents of a protein sample almost halves (Nagaraj et al., 2012; Hebert et al., 2013b). Using specific enrichment, researchers can also focus on post-translation protein modifications of proteins (PTMs), such as phosphorylation and acetylation (Engholm-Keller and Larsen, 2013; Olsen and Mann, 2013). Furthermore, the array of possibilities for MS-based proteomics include the identification of proteins which bind to small molecules, drugs or nucleic acid sequences (Winter et al., 2012; Bantscheff et al., 2011; Pichlmair et al., 2011) and the unraveling of the constituents of protein complexes (Varjosalo et al., 2013).

The quantification of the difference in protein abundance across different samples, as well as the difference in the prevalence of PTMs, is of great interest in biological research. Several analytical strategies have been developed to use MS to get precise estimates of the differences (see Bantscheff et al. (2012) and section 2.2). Isotope-coded labeling of samples is an efficient

approach to relative protein quantification and allows the simultaneous analysis of the labeled samples (i. e. multiplexing). The mass spectrometry results then can be used to both identify the protein constituents of the samples and quantify their differential abundances (Gouw, Krijgsveld, and Heck, 2010; Christoforou and Lilley, 2012). Isobaric tags are a special kind of isotope labels that enable a high number of multiplexing of up to 10 samples. In isobarically tagged data, the quantitative information is encoded in the fragmentation spectrum, i. e. the second mass spectrometric dimension rather than the first. Isobaric tags, such as the commercialized iTRAQ and TMT kits, can be used with any sample and MS strategy, and have been successfully employed in many experiments (e. g. Winter et al., 2012; Borgdorff et al., 2013; Müller et al., 2012). The methods in this thesis are developed for and tested on isobarically tagged data.

The datasets generated by quantitative proteomics methods are large and complex. The best way of their analysis is the topic of extensive research (see Lin et al. (2006), Schwacke et al. (2009), and Karp et al. (2010, e. g.) and section 2.3). The structure of quantitative (shotgun) proteomics data is intricate for several reasons: (1) The inference of protein presence in the sample is not straight-forward, as peptides might match to multiple proteins (Colinge and Keiryn L Bennett, 2007). (2) A variable number of spectra is available per peptide, leading to differences in accuracy and precision of estimators (Bantscheff et al., 2012). (3) The quantitative information displays a heterogeneity of variance (heteroskedacity), which leads to lower precision of ratios calculated from low-intense peaks (Karp et al., 2010). (4) The accuracy of measured ratios is influenced by isotope impurities and co-eluting peptides, which can distort the quantitative information (Christoforou and Lilley, 2012). (5) In the analysis of PTM data, several additional levels of uncertainty are introduced, with often uncertain localization of the modification site, due to the necessity of calculating fold-changes on single peptides, and the correction of peptide with protein abundance changes (see Chalkley and Clauser (2012) and Wu et al. (2011) and section 2.5).

All these aspects require careful examination in the statistical modeling of the data.

Bioinformatics software plays a pivotal part in the analysis of quantitative proteomics data (Cappadona et al., 2012; Bantscheff et al., 2012). Proteomics uses various instrumental platforms and software search engines for protein identification, and thus different file formats (Jones et al., 2012). The inference of proteins through protein grouping is often performed in separate tools, as well as the localization of PTM sites on modified peptides (Nesvizhskii, Vitek, and Aebersold, 2007; Chalkley and Clauser, 2012). Standalone programs for protein quantification may often provide well-designed graphical user interface, but their application for non-standard investigations may be limited. Furthermore, software with a GUI often do not provide a command line interface which allows its use in an analysis pipeline. Software pipelines are essential for labs, such as proteomics core facilities, which create lots of high-throughput data sets (Cappadona et al., 2012). The analysis results should present relevant information on the ratio (such as

variance, statistical significance, effect size, and confidence intervals), the protein (protein name and description), and the modification site - if applicable. Ideally, it should be possible for bioinformaticians to delve deeper into the data while the wet-lab scientists can directly identify important proteins and use the results for downstream analysis.

The challenges posed by quantitative proteomics data on statistical methods, as well as bioinformatical software, are discussed in more detail in the next chapter (chapter 2). This thesis aims at developing a software tool and better statistical methods for the analysis of isobarically tagged quantitative proteomics data (see chapter 3). The results of this work are presented in chapter 4; specifically section 4.1 develops the statistical and software framework for protein quantification, which is extended in section 4.2 for the analysis of PTM data, and applied in several publications described in section 4.3. Finally, Chapter 5 concludes the thesis and frames the results in the wider context of the field.

2. Background on quantitative proteomics

This chapter introduces the main concepts of mass spectrometry and its use in protein identification and quantification (sections 2.1 and 2.2). Mass spectrometry-based proteomics is a diverse field, in which many types of instruments and experimental workflows are employed. An understanding of the concepts is important for the development of data analysis methods and software. The current understanding of the structure of data from isobarically labeled proteomics experiments is reviewed in section 2.3, as well as necessary preprocessing and data analysis steps. The challenges in statistical modeling of the data are presented, and are used to motivate the developments investigations of this thesis. Section 2.4 then introduces the particular challenges posed by PTM-centric quantitative proteomics data, which provides a further motivation for the development of an appropriate analysis tool. Finally, section 2.5 shortly reviews the currently available software tools for the analysis of isobarically tagged data.

2.1. Shotgun proteomics and mass spectrometry

"*Mass spectrometry is the art of measuring atoms and molecules to determine their molecular weight. Such mass or weight information is sometimes sufficient, frequently necessary, and always useful in determining the identity of a species.*"

— John B. Fenn, Nobel Laureate 2002

The principle of mass spectrometry is that ions (i. e. atoms with a net positive or negative charge) are affected by electric or magnetic fields based on the ratio of their mass to their charge (Hoffmann and Stroobant, 2007). In 1897, the physicist Joseph J. Thomson used this principle to prove the presence of particles in a cathode ray (i. e. stream of electrons), and determine the $\frac{m}{z}$ of electrons. He thus could prove the existence of sub-atomic particles - at least once the charge of electrons was determined as -1 by Millikan (1913). In the beginning, the separated ions were detected on photographic plates, though already early on coupled an electron detector to measure the separated ion stream (Joseph J Thomson, 1912) and thus a mass spectrum. Aston (1920) used the instruments to discover the existence of stable isotopes, and Nier and Gulbransen (1939) determined the isotope ratio of ^{13}C in nature. Already 1941, isotope tracers were employed to study chemistry and biology. In the second world war, mass analyzers were used to enrich 235U and characterize uranium. Many novel types of mass analyzers were developed (Mattauch, 1936; Stephens, 1946; Paul and Steinwedel, 1953; Comisarow and Marshall, 1974b; Stafford Jr et al., 1984; Makarov, 2000) and provided more sensitivity, speed, and precision.

Already in 1959, mass spectrometry was first applied to sequence peptides (Biemann, Gapp, and Seibl, 1959). The real breakthrough for protein analysis, however, was in the 1980s. Hunt et al. (1981) first described the use of multiple levels of mass analysis (tandem mass spectrometry, MS/MS, Futrell and Miller, 1966) to determine peptide sequences. Fenn et al. (1989) and Tanaka et al. (1987) developed soft ionization techniques that keep peptides intact in the process of transforming them into the gaseous ionized state, in which they can be analyzed by MS. J. K. Eng, McCormack, and J. R. Yates 3. (1994) developed first algorithms to identify peptides based on the mass spectra. Hybrid instruments for tandem mass spectrometry were continually developed (Cornish and Cotter, 1993; Morris et al., 1997).

"Shotgun" proteomics evolved as the standard approach for large-scale protein identification (Yaoyang Zhang et al., 2013). It enables the identification of thousands of proteins in a single mass spectrometric experiment. The name lends itself from the shotgun genome sequencing approach, which tries to reconstruct the whole genome from (overlapping) reads of small DNA fragments (Metzker, 2010). Analogously, shotgun proteomics is an approach to infer proteins by the study of protein fragments, which are more amenable to mass spectrometric analysis. The protein fragments (peptides) are typically generated by enzymatic digestion.

The further parts of this section detail the components of a shotgun proteomics experiment (see fig. 2.1), namely protein isolation, separation and digestion techniques; the mass spectrometry instrumentation which is used; and the data processing and protein identification from the raw mass spectrometric output.

2.1.1. Protein isolation, separation and digestion

The isolation of proteins - or a specific subset of proteins - from the sample is the first step in the analysis (see fig. 2.1). The subject of the analysis can be for example the whole lysates of cells or tissues, or specific parts thereof. Affinity purification can be used to enrich for specific protein complexes or protein interactors (e. g. Giambruno et al., 2013; Varjosalo et al., 2013). The achieve a greater depth in the analysis of the sample, the proteins can be fractionated using gel electrophoresis or other chromatographic techniques (Dowell et al., 2008). Then, the proteins are digested by a proteolytic enzyme. For most applications, trypsin is employed, which cleaves the amino acid sequence at the carboxyl side of arginines and lysines residues (Olsen, Ong, and Mann, 2004). The peptides should not be too small (below six amino acids) or too large (above 30). Most peptides generated by trypsin fall in this range. The mixture of peptides can be fragmented "online" or "offline" again to increase the analysis depth. Online separation refers to a setup with a liquid chromatography (LC) columAnalogousirectly coupled to the mass spectrometer. The eluate is ionized and analyzed at the time it elutes from the column. This efficient setup, termed LC-MS, is the workhorse of most high-throughput protein identification

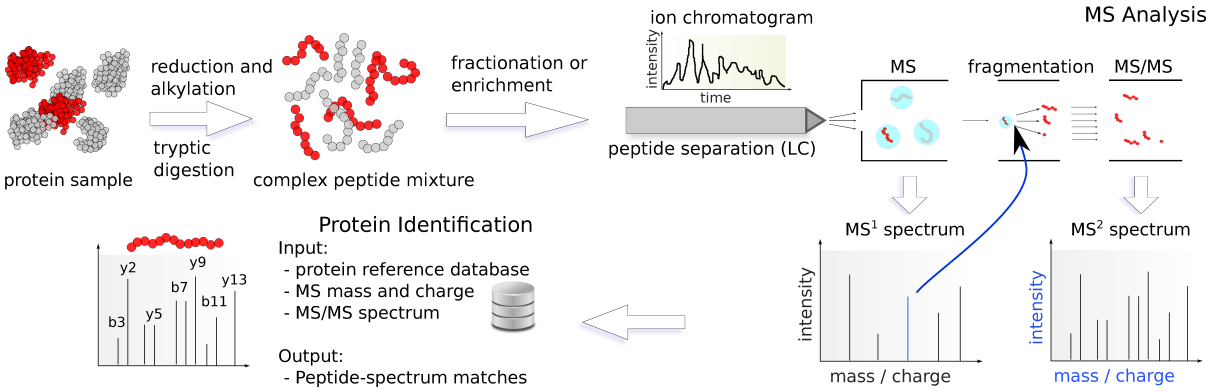


Figure 2.1: Proteomics workflow for large-scale protein identification ("shotgun" proteomics). (1) Proteins are isolated, denatured, digested, and fractionated. (2) The peptide fractions are separated on a liquid chromatography column over time (30 minutes to 2 hours), ionized and injected into the mass spectrometer. At periodic intervals, an MS^1 spectrum is gathered, which measures the mass-to-charge ratios and intensities of the ions. Certain ions (precursors) are selected (usually 3 to 10 after each MS^1 survey scan), and then sequentially isolated and fragmented. The MS^2 spectra capture the fragment mass-to-charge ratios and intensities. (3) The MS^2 peaklists and precursor mass are matched against theoretical spectra of proteins in a reference database.

experiments. Online separation is typically performed with reversed-phase high pressure liquid chromatography (J. R. Yates, Ruse, and Nakorchevsky, 2009). The columns, which have very small diameters ($75\,\mu$m to $100\,\mu$m) and require high pressure, allow using very small amounts of starting materials and provide high selectivity and sensitivity in the analysis. It is important to note that the resolving power of the physical separation techniques is as important as the ion mass-to-charge ratio resolving power of the mass spectrometer.

Peptides are more easily separated and analyzed by mass spectrometry, and thus shotgun proteomics has evolved as the standard technique for large-scale protein identification. However, the peptide-based approach has drawbacks. First, for some proteins or protein regions no detectable peptides may be generated. A prominent example are the tails of histone proteins, which are important in transcriptional control and contain many lysines and arginines. The peptides generated via tryptic digestion are usually too small to be detected (Witze et al., 2007). A further drawback of peptide-based proteomics is the indirect identification of proteins. Peptides are in many cases not uniquely matching to one protein, especially when considering splice variants. Protein inference from shotgun data is still a topic of research (Claassen, 2012). As will be discussed later in section 2.3.5, shared peptides are also important in protein quantification. A final issue appears when shotgun proteomics data is used for the investigation of PTMs. Many proteins have multiple sites that can be differentially modified (Witze et al., 2007). When using shotgun data, the view is usually limited to a small stretch of the sequence, and thus any

combinatorial code is lost.

The study of intact proteins in the mass spectrometer is called *top-down* proteomics (Siuti and Kelleher, 2007), and does not suffer from the issues mentioned for the peptide-based approach. Top-down mass spectrometry retains the sequence and modification variants present in the sample. It has been successful for targeted studies of single or few proteins, thus providing a complementary approach for targeted analysis (Armirotti and Damonte, 2010). However, it requires specific attunement for the studied proteins, and thus is not applicable for large-scale untargeted protein identification. The *middle-down* approach is an alternative that is often used for the identification of PTMs. It has been especially used in the study of histone tails, where it is believed that different modifications patterns encode the function. As the name suggests, middle-down falls between the above-mentioned approaches. It is peptide-based but employs proteases that cleave rare amino acids and thus generate very long peptides in the range of 5 kDa to 6 kDa (Sidoli, Cheng, and Ole N. Jensen, 2012).

Studies into post-translational modification usually require a specific enrichment step for peptides or proteins that bear the modification of interest (see 2.4 and Zhao and Ole N. Jensen, 2009). The enrichment depends on the properties of the modified residues: For acetylated lysines and phosphorylated tyrosines, antibodies are employed (S. C. Kim et al., 2006; Rikova et al., 2007), phosphorylated serines and threonines; for phosphorylated serines and threonines, immobilized metal affinity chromatography (IMAC, Porath et al., 1975) with different chelating metals (Pinkse et al., 2004; Feng et al., 2007) is used; for glycopeptides the affinity to lectins is exploited (Morelle et al., 2006).

2.1.2. Protein mass spectrometry instrumentation

Mass spectrometry is an analytical technique for identifying the *mass-to-charge* ratio of charged ions[1]. The process can be divided into three steps: ion generation by an ion source, ion separation by a mass analyzer, and ion detection in a mass detector (see fig. 2.2.A). Mass spectra are generated and registered by a computer system.

The standard setup for proteomics includes two levels of mass analysis (tandem mass spectrometry or MS/MS, see fig. 2.2.B). Both the mass spectra of intact peptide ions are measured, as well as the fragments of specified peptide ions. This requires an intermediary mass filter to select precursors based on their mass-to-charge ratio, and a collision cell to generate fragments.

[1] It is worth to keep in mind that a mass spectrometer (despite its name) cannot measure the molecular weight of objects directly, but the *ratio* of the mass of analytes to the charge they carry. Neutral analytes, thus, cannot be measured. The usual notation of this quantity is $\frac{m}{z}$ (Todd, 1991), where m is the mass in Dalton, and z is the number of charges.

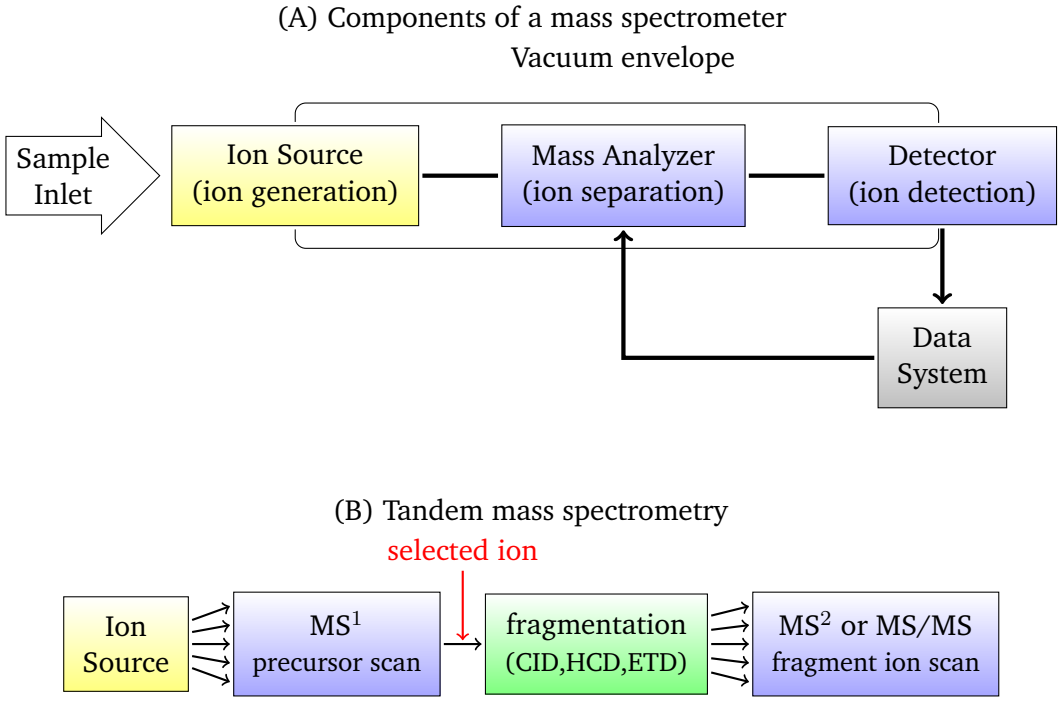


Figure 2.2: (A) General mass spectrometer schematic. The sample is introduced into the instrumental vacuum envelope as molecules are ionized (ion source). The ions are electrostatically propelled to the mass analyzers, where they are separated based on their mass-to-charge ratios. A detector then measures the current that is carried by the ions. (B) Tandem mass spectrometry schematic. To gain insight into specific ions, two levels of mass measurements are performed. First the $\frac{m}{z}$ of intact molecules are measured (precursor scan). Then specific ions (i. e. a specific $\frac{m}{z}$) are selected, isolated, and fragmented. The fragment ions are recorded (fragment ion scan) and, for protein identification, matched against the theoretical fragmentation patterns of peptides.

Ion sources

Ionization sources transform analytes from solid or liquid form into gas phase ions that are amenable to MS analysis (J. R. Yates, Ruse, and Nakorchevsky, 2009). Usually it is pivotal that the analytes are kept intact in the process - which is difficult for large organic molecules such as peptides or proteins. Electrospray ionization (ESI, Fenn et al., 1989) and matrix assisted laser desorption/ionization (MALDI, Tanaka et al., 1987) have developed as the standard ionization methods for proteomics.

MALDI requires the analytes to be mixed with a matrix compound and co-crystallized on a plate. Directed laser pulses cause the desorption of the matrix and ionization of the analyte. After the laser pulse, the ions cool down for a short time (in the range of 150 nanoseconds) before being accelerated into the mass analyzer by an extracting pulse. A very interesting application is MALDI imaging, which enables the investigation of the spatial distribution of proteins in thin tissue sections (Cornett et al., 2007; Casadonte and Caprioli, 2011).

ESI is a continuous ionization source (in contrast with the pulsed acquisition with MALDI) for liquid samples (M. Wilm, 2011). Thus, it can be coupled *online* to liquid chromatography, ionizing the sample as it elutes from the column. The eluate goes through a thin metallic needle with high voltage, which creates a fine spray of highly charged drops. As the solvent evaporates, the droplet decreases in size and its surface tension increases. The repulsion of the ions on the surface becomes so great that the ions are released and directed through an opening into the vacuum of the mass spectrometer. Nano-electrospray ionization (nanoESI) is a variant of ESI with a very small spray needle and low flow through rate, which reduces the required amount of sample material (M. S. Wilm and Mann, 1994; M. Wilm, 2011). Liquid-chromatography coupled mass spectrometry (LC-MS) using ESI is the workhorse for most large-scale proteomics investigations (Niessen, 2010).

Mass analyzers and mass filters

Mass analyzers enable the measurement of mass-to-charge ratios ($\frac{m}{z}$) of gas phase ions by separating them in time or space (Glish and Vachet, 2003). The charged ions are affected by electrical or magnetic fields and thus separated.

The main measures of the performance of mass analyzers are resolution, accuracy, mass range, and scan rate. *Resolution* is the resolving power of the mass spectrometer between adjacent peaks at certain m/z values. It is defined as the ratio $\frac{(m/z)}{\Delta(m/z)}$, where m/z is the observed value and $\Delta(m/z)$ the smallest difference for two ions that can be separated (Murray et al., 2013). High resolution has the advantage of separating isotopes and near-isobaric ions into separate peaks and delivers more precise mass measurements. Typically, high resolution mass analyzer

are highly accurate, but this is no necessity. *Accuracy* defines how close the observed value is to the true value. It is usually denoted in parts per million (ppm); e. g. measurements of a mass analyzer with an accuracy of 10 ppm are expected to be within ± 0.1 at $\frac{m}{z}$ 100, within $pm1$ at $\frac{m}{z}$ 1000, and within ± 10 at $\frac{m}{z}$ 10000 of the true value. Notably, the mass accuracy is usually not constant over the $\frac{m}{z}$ range (A. Makarov et al., 2006).

Quadrupole. The linear quadrupole consists of an array of four metal rods, two being connected to a direct current (dc), and the others to an alternating current oscillating at specified radio-frequencies (rf) (Paul and Steinwedel, 1953). The amplitude of the rf and dc voltage determine which $\frac{m}{z}$ retain stable trajectories when passing through to the detector ("mass-selective stability"). To scan the spectrum, these parameters are sequentially incremented (at a constant rf/dc ratio). In the process, all ion species that are not scanned are lost. Quadrupoles have limited mass ranges - the upper limit varies from 300 to 4000 $\frac{m}{z}$. Quadrupoles are relatively cheap and small, and are used in many mass spectrometers as mass filters.

Quadrupole and linear ion trap. The quadrupole ion trap (QIT) is a further development by Paul and Steinwedel (1953). It is closely related to the quadrupole mass analyzer, but features a third dimension in the electric field that enables trapping of the ions. They were initially operated, like quadrupoles, in a mass-selective stability mode, and not widely used. Only the development of the inverse "mass-selective instability" mode by Stafford Jr et al. (1984) led to its widespread adoption. In mass-selective instability mode, the ions are stored during the mass analysis and sequentially ejected to a mass analyzer. Linear versions of the ion trap offer increased ion storage capabilities and faster scan time compared to the three-dimensional Paul traps (Douglas, Frank, and Mao, 2005) and were commercialized in the LTQ Orbitrap by Thermo Scientific (A. Makarov et al., 2006).

Time-of-flight. Time-of-flight (TOF) mass spectrometer separate ions based on their velocity, and were first described by Stephens (1946). In a constant electrostatic field, ions attain velocities proportionally to their mass-to-charge ratio (Mamyrin, 2001). Measuring their time of arrival at the detector after a stretch of field-free region can be used to calculate their mass:

$$t = \sqrt{\frac{2md_1}{eE}} + d_2\sqrt{\frac{m}{2eV_0}}$$

where m is the mass of the particle, e its charge, E the electrostatic field, d_1 is the length of accelerating region, d_2 the length of field-free region, and V_0 the accelerating potential (typically 2-10 kV). The distance d_2 is usually between 0.5 and several meters. The mass resolution is a function of the flight time (with constant precision of time measurements) and thus can be

increased by a longer field-free region, or lower accelerating potential. The resolving power of TOF mass spectrometers can be increased with a reflectron - an electrostatic ion mirror that doubles the path length at a similar size total size of the instrument. Reflectrons also reduce the spread of flight times of ions due to initial differences in kinetic energies (Alikhanov, 1957; Mamyrin et al., 1973). TOF works very well with pulsed ionization such as MALDI. TOF analyzers have the advantage of theoretically unlimited mass range.

Fourier-transform mass analyzer. Fourier transform mass spectrometers (FTMS, Comisarow and Marshall, 1974a) measure the coherent motion of ions that orbit with frequencies proportional to their $\frac{m}{z}$. The frequencies are recorded and transformed into mass-to-charge ratios by Fourier transform. The longer the orbiting ions are measured, the more precise is the signal - usual times are $.5$ to 1 seconds. FTMS are very accurate and can measure the whole spectrum at once. Resolving power is inversely proportional to the $\sqrt{m/z}$, i. e. lower for higher $\frac{m}{z}$ values.

Orbitrap mass analyzer. Orbitrap mass analyzers separate ions orbiting around and along a central electrode (at a voltage of 3.5 to 5 kV) (Roman A. Zubarev and A. Makarov, 2013). The image current is picked up by sensors on outer electrodes, and the $\frac{m}{z}$ can be calculated using Fourier transformation similar to FTMS. Orbitraps have good analytical performance in terms of resolutions, speed, mass accuracy and dynamic range, while being relatively small. Currently, standard and high-field versions of the Orbitrap analyzer exist (Roman A. Zubarev and A. Makarov, 2013). Orbitrap mass analyzer require direct injections. As such, they can be directly linked to a MALDI ionization source, but require a trapping module such as an ion trap to be interfaced with ESI (Perry, Cooks, and Noll, 2008). In the last years, Thermo Scientific's mass spectrometers that feature Orbitrap mass analyzer - such as the Orbitrap Velos, QExactive, and Orbitrap Fusion - have been the most widely used analyzers for high-throughput proteomics.

Mass detector

The detector generates a usable signal from the ions that pass through the mass analyzer (Hoffmann and Stroobant, 2007). A difference can be made between detectors that consume the ions, where ion impacts results in electrical current directly (such as Faraday cups and electron multipliers), and those that recognize ions flying by, inducing electrical current (used in FTMS and Oribtrap mass spectrometers) (Koppenaal et al., 2005).

The simplest form of a detector are Faraday cups and have been employed already in the first mass spectrometers by Joseph J Thomson (1912). A striking ion causes secondary ions to be emitted, inducing an electron current towards the detector. Faraday cups are not very sensitive,

but highly accurate and provide a direct relation between the current that is measured and the number of ions.

Electron multipliers are much more sensitive detectors used in most modern mass spectrometer (but those based on image-current). Their basic building block is either a series of dynodes (each similar to a Faraday cup), or a continuous structure, with increasing potential (Farnsworth, 1934). When ions hit the surface of the first dynode, secondary electrons are emitted and attracted to the next dynode. The second dynode, which has a higher potential than the first, is hit and produces more secondary electrons. This leads to a cascade and amplifies the signal in the order of 10^6. As electrons are depleted from the wall of the electron multiplier, these mass detectors need some time to recover.

Image-current based detectors recognize charged ions that pass by and thus induce a current. The basic principle is that ions with the same $\frac{m}{z}$ exhibit coherent oscillations, either along the central electrode (in Orbitraps) or between two electrodes (in FTMS). With multiple ion species present, the measured signal is a sum of sine waves. Using Fourier transformation, the signal can be decomposed to the frequencies of the oscillations. The frequencies are proportional to m/z in FTMS and $\sqrt{m/z}$ in Orbitraps. The resolving power of these detectors is proportional to the measurement time (Roman A. Zubarev and A. Makarov, 2013).

Tandem mass spectrometry (MS/MS)

Tandem mass spectrometry uses multiple levels of mass analysis to characterize the analytes (McLafferty, 1981). In the first level, denoted MS^1, the mass analysis provides a *survey spectrum* of ions that are present. Certain ions, termed the *precursor ions*, are targeted and successively analyzed in the next MS level. For that, targeted ions are separated from the others in mixtures and fragmented by collision. The fragment ions - also termed *product ions* - are measured (MS^2 spectrum), providing a fingerprint of the target molecule. This process (of selection, fragmentation, and mass analysis) can be repeated further on specific fragments (MS^n). MS^2 is the standard level for protein identification, but for more selectivity in identification and quantification, higher levels can be employed Ting et al. (2011), Ahn et al. (2007), and Witze et al. (2007).

The acquisition mode defines which ions are targeted and get selected as precursors. Classically, shotgun proteomics uses data-dependent acquisition (DDA), in which the computer system of the mass spectrometer decides on certain precursors after each MS^1 acquisition - usually the n most intense $\frac{m}{z}$ are taken, with n depending on the acquisition speed. Normally, the precursor $\frac{m}{z}$ that were selected are excluded for a certain time from another MS^2 acquisition (Yaoyang Zhang et al., 2013).

Data-dependent acquisition is to a certain degree stochastic. In complex mixtures it is not possible to fragment and identify all precursors (Michalski, Juergen Cox, and Mann, 2011). This leads to poor reproducibility across runs (Tabb et al., 2010) in terms of identified peptides (to a certain degree less on the level of proteins). In replicate runs, this can be reduced to some extent by using putting the $\frac{m}{z}$ of peptides of interest on "inclusion lists", whereby they will be selected as precursors when present (Mikhail M. Savitski et al., 2010).

Precursors are selected within a small mass window, e. g. ± 2 $\frac{m}{z}$. This window is based on the precision of the mass filter (see section 2.1.2). It is beneficial to have a large enough window to include high intense molecule isotopes. Due to the complexity of peptide mixtures, however, it is not uncommon that other peptides with matching $\frac{m}{z}$ are present. Those peptide species are co-selected and present challenges in the identification of spectra, as many peaks are not assignable to one peptide sequence (Houel et al., 2010). Furthermore, co-eluting peptides are a hurdle for accurate MS^2-based quantification (see section 2.3.2).

Data-independent acquisition modes (DIA) intentionally create chimeric spectra from peptide ions over a large $\frac{m}{z}$ range (Geiger, Juergen Cox, and Mann, 2010; Egertson et al., 2013). DIA methods survey the whole $\frac{m}{z}$ range one broad $\frac{m}{z}$ window at a time. SWATH-MS for example, which has been developed by the Aebersold group, uses windows of 25 $\frac{m}{z}$ (Gillet et al., 2012; Collins et al., 2013). In each window, the peptide ions are isolated, fragmented, and the fragments are measured. The benefit of DIA method is that it fully captures the data. Methods to extract single peptide fragment spectra, or search for peptide fragments in the whole spectra, are in development. The focus of this work, however, is on the classic DDA method, in which spectra are mapped to single peptide sequences. Naturally, DIA methods cannot be combined with MS^2-based quantification techniques. DIA has a lot of potential for the future, but does not provide the throughput of DDA, yet.

Selected Reaction Monitoring (SRM). When the objective is to have assays for known entities (and not the discovery of novel ones), targeted proteomics via selected reaction monitoring (SRM) provides a highly reproducible and sensitive tool. Classically, triple quadrupole mass spectrometers are used (Domon and Aebersold, 2010). Parallel reaction monitoring has been developed recently, using Orbitrap instruments, which provide higher resolution and accuracy (Gallien et al., 2012). Together with spiked peptide standards (AQUA, Gerber et al., 2003), SRM can be used to determine absolute quantities of proteins of interest even in complex backgrounds.

Dissociation methods Several methods for fragmentation of the peptide ions exist. The most common ones are collision-induced dissociation (CID), higher energy collisional dissociation (HCD), and electron-transfer dissociation (ETD). The methods are performed in different

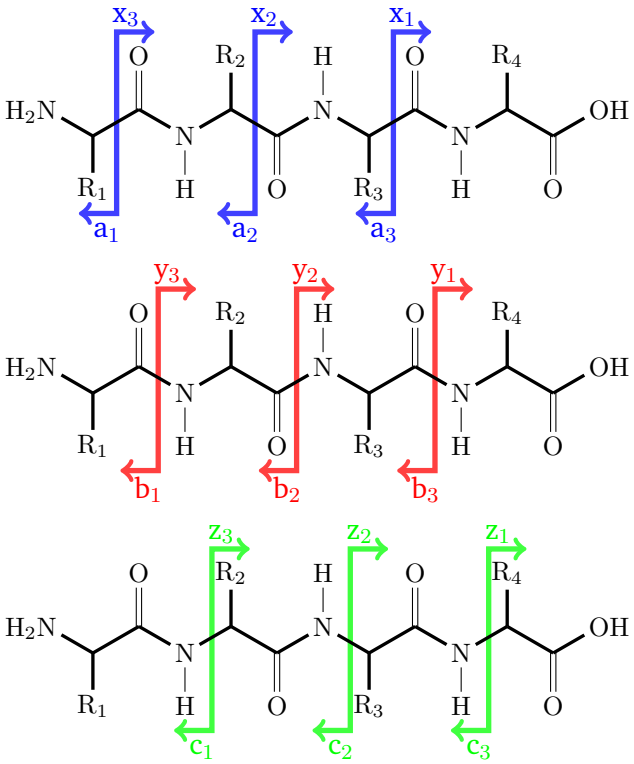


Figure 2.3: Nomenclature for peptide fragments resulting from peptide backbone cleavages. The breaks at the peptide bond can occur at three different positions (blue, red, green). The product ions are named with the letters **a, b, c** when including the peptide N-terminus, and **x, y, z** when including the peptide C-terminus. The subscript denotes the number of amino acid residues present in the fragment.

instruments, or different part of the same instrument. Depending on the method, the amino acid sequence breaks predominantly at particular parts of the peptide bond. A nomenclature for naming the peptide fragments has been developed (Murray et al., 2013), see fig. 2.3. For breaks at the peptide backbone C-C bond, the resulting fragments are called a and x ions, denoting the fragments containing the peptide N-terminus and peptide C-terminus, respectively. For breaks at the peptide backbone C-N bond, the respective fragments are called b and y ions. Finally, for breaks at the peptide backbone N-C bind, the fragments are called c and z. Due to the different nature of how the breaks are induced, different fragmentation methods produce different fragments. CID and HCD mostly produce b and y ions, and ETD results predominantly in c and z ions.

CID (or CAD, collision-assisted dissociation) leads to the dissociation of peptide ions by the collision with inert gas, such as helium, or argon. In the collision, some kinetic energy is converted into internal energy, resulting in breaks of the ions at the peptide bonds (Sleno and Volmer, 2004). CID is typically performed in ion traps, which provides high sensitivity and fast

analysis. As the peptide ions break, they fall out of phase and hit the detector. However, ion traps have the inherent drawback of a low mass cutoff, i. e. no data is available for the low $\frac{m}{z}$ region (Y.-H. Yang et al., 2009). While the cutoff poses no big issue for peptide identification, reporter ion-based quantification with isobaric tags is not possible, as the reporter ions reside in the low-$\frac{m}{z}$ region of the spectrum (see 2.2.4).

HCD occurs in quad or octopoles with gas molecules. HCD collision cells are available as add-ons to orbitrap systems (Olsen et al., 2007). On the orbitrap, the peptide ions go through the C-trap into the octopole collision cell, where the fragmentation occurs. The fragments are brought back to the C-trap and injected into the orbitrap, where the mass analysis occurs. The orbitrap generates high-resolution high-accuracy MS^2 spectra and also retains low $\frac{m}{z}$ ions. It thus works well with MS^2-based (isobaric) quantification methods, and additionally immonium ions can be detected. In initial implementations, it was less sensitive than CID, mainly due to longer cycle time (Jedrychowski et al., 2011) - even though it produced better identification scores than CID due to better accuracy. This drawback inspired the development of hybrid methods to generate a full CID spectrum for identification and a "short" HCD spectra just for quantification for every precursor (Köcher et al., 2009). However, with the novel versions of mass spectrometers in the Orbitrap family (Elite, QExactive, Fusion), the cycle time of HCD is much faster.

ETD is a chemical fragmentation method (Syka et al., 2004). Charged anions transfer an electron to the precursor, inducing peptide bond breaks. In contrast with CID and HCD, where the low-energy bonds have a higher probability of breaking, this process gives equal fragmentation probability at all bonds. ETD can reveal more sequence information - especially on long sequences - and provide a better fragmentation spectra for peptides with volatile modifications such as phosphorylation (Chi et al., 2007). ECD (electron capture dissociation) is a similar fragmentation method where electrons are directly introduced to the trapped ions (R. A. Zubarev et al., 2000).

2.1.3. Data processing and protein search engines

Computational methods and statistics are pivotal in the identification of proteins from mass spectrometry data (Colinge and Keiryn L Bennett, 2007; W. S. Noble and MacCoss, 2012; Nesvizhskii, Vitek, and Aebersold, 2007). The first steps in data processing are calibration, transformation, and peak picking of the data. For Orbitrap and ion cyclotron resonance mass spectrometers, frequencies are Fourier-transformed to the *m/z* domain (Roman A. Zubarev and A. Makarov, 2013). After the transformation to the *m/z* domain, peak picking resolves isotopic clusters and reduces the mass spectra to peaks that are potential molecules (Eckel-Passow et al., 2009).

The identification of peptides in shotgun proteomics data is based on the $\frac{m}{z}$ measured in MS^1 and the corresponding MS^2 fragment spectrum (Colinge and Keiryn L Bennett, 2007). To determine the charge of peptide ions, the presence of isotope patterns is exploited. The masses of peptides with equal amino acid composition are not all the same, but rather distributed with spacing of multiples of (about) one Da. This pattern is due to naturally occurring isotopic variants, which constitute a certain percentages of the atoms. Knowing the natural rate of occurrence of atomic isotopes, it is possible to calculate the expected isotopic pattern of amino acid sequences. The pattern can be used to deduce the peptide charge state - and thus the peptide mass - as the distance between the peaks is $1/z$.

The three main approaches for protein identification from MS data are *de novo* identification of proteins, database search, and spectrum library search (Nesvizhskii, 2010). In all approaches, the search space for peptides is first restricted to a window around the precursor mass. *De novo* algorithms match the spectrum against all possible amino acid sequences (Dancik et al., 1999). Database search engines are the most common solution. In this method, the peptide search space is constrained by considering only peptides that derive from proteins in a protein sequence database. Theoretical fragmentation spectra are generated and correlated with the experimental spectra (Jimmy K. Eng et al., 2011). Spectral library searches, on the other hand, match the experimental spectrum against a library spectrum, which were experimentally generated and are already matched to an amino acid sequence (Nesvizhskii, 2010).

The most common database search engines are the commercial Mascot (Perkins et al., 1999) and Sequest (J. K. Eng, McCormack, and J. R. Yates 3., 1994), and the free X!Tandem (Craig and Beavis, 2004; Bjornson et al., 2008) and OMMSA (Geer et al., 2004). Several further search engines have been developed. Phenyx is a search engine that is based on the OLAV family of likelihood scoring schemes and demonstrated better performance than Mascot (Colinge et al., 2003). MS-GF+ is a novel search engine that uses generating functions for scoring for the different fragmentation methods (S. Kim et al., 2010).

The protein search engines return peptide-spectrum matches, accompanied by p values or another type of score. The most common approach to select trusted hits does not use the scores or p values directly, but calculates a threshold based on a procedure to estimate the false-discovery rate (Elias and Steven P. Gygi, 2007): The sequences in the target database, e. g. all human protein sequences and its splice variants, are reversed or scrambled to generate a decoy database. Searches against target and decoy databases (or a combined database) are performed. At a specific score threshold, the number of peptide-spectrum matches in the decoy database is considered an estimate on the number of false positives in the target database. Score thresholds are thus adjusted to meet a specified false-discovery rate - either at the level of spectra, peptides, or proteins (Colinge and Keiryn L Bennett, 2007). The false discovery rate on the level of proteins is naturally much higher than on the level of spectra, as wrong peptide-spectrum

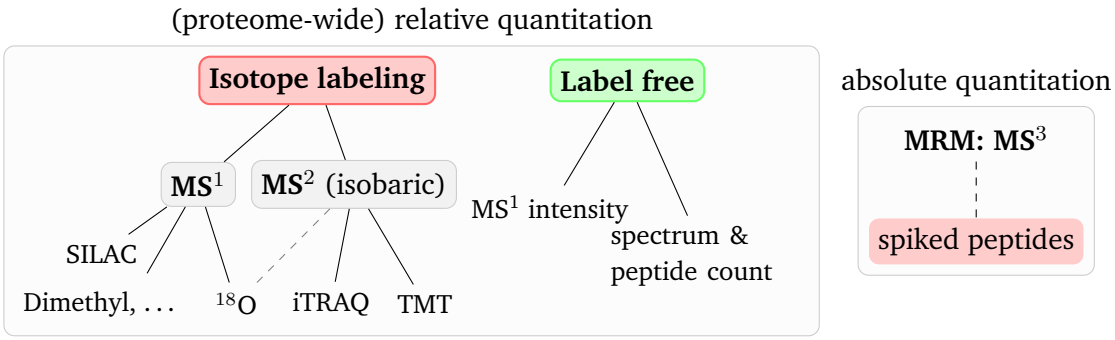


Figure 2.4: Overview of quantitative proteomics techniques. (Figure and text adapted from Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91)

matches are more likely to match to a novel protein (Claassen, 2012).

Shotgun proteomics does not give protein identifications directly - the peptide matches have to be combined to infer proteins (Yaoyang Zhang et al., 2013). As the connections between the peptides and proteins are lost during digestion, the protein inference is not trivial (Claassen, 2012). Many proteins share parts of their sequence and thus many peptides do not match uniquely to one protein. As the issue of shared peptides is of importance in the quantification of proteins, it is discussed in more detail in section 2.3.5.

Interesting developments are made for real-time protein identification search engines that run directly on the mass spectrometer in parallel to the mass analysis (Altelaar, Munoz, and Heck, 2013). Real-time protein identification enables targeting specific peptides or proteins, and real-time decision making to improve detection rates, quantification accuracy, and localization of post-translational modification sites (D. J. Bailey et al., 2012).

2.2. Protein quantification by mass spectrometry

The quantification of protein changes across samples is an important part in many proteomic analysis. Quantitative methods can be divided by their scope: proteome-wide or targeted, relative or absolute (see fig. 2.4). Targeted quantification necessitates the development of specific assays for proteins of interest. By spiking heavy isotope-coded versions of endogenous peptide at known concentrations, absolute abundance measures for the proteins may be derived. For research applications, typically unbiased proteome-wide quantification is desired. This comes with the drawbacks of less precision and only relative quantification, i. e. fold-changes of protein abundances may be derived, but not the protein abundances themselves.

Another distinction are labeling and label-free methods. Labeling approaches use isotope-coded tags to label proteins, which are then pooled and co-analyzed. Quantitative information then can be extracted from the MS^1 or MS^2 spectra. Label-free approaches are based on the comparison of separate runs.

This section reviews protein quantification methods. The focus is on MS^2-based labeling, which can be used for any sample material and enables the simultaneous analysis of up to 10 samples.

2.2.1. Targeted and absolute quantification

Targeted mass spectrometry techniques aim at accurate identification and quantification of a specified set of proteins. Selected reaction monitoring (SRM, see fig. 2.4) is a precise and reproducible method that utilizes specific peptide assays with three mass-spectrometric dimension (MS^3, Picotti and Aebersold, 2012). SRM assays have improved the limits of detection and reproducibility compared to *disocvery* proteomics experiments (Domon and Aebersold, 2010; Picotti and Aebersold, 2012). Relative quantification methods (as described in the following sections) can be used in conjunction with SRM. Furthermore, it is especially suitable for absolute quantification methods of a set of proteins of interest. For absolute quantification, labeled or synthetic peptides that are heavier than their endemic counterparts are spiked to the sample at known concentrations. The ratio of the intensities of the peak of the endemic and spiked peptides can then be used to calculate the abundance of the endemic peptides (Brönstrup, 2004).

Targeted methods are a powerful complement to untargeted methods, especially when precision, reproducibility, and a low limit of detection are essential. The drawbacks are that the number of observable proteins in one run is limited to about 50 to 100, and for each protein an assay has to be designed (Picotti and Aebersold, 2012). However, Picotti et al. (2013) recently generated a complete assay map for yeast using synthetic peptides. FUTURE

2.2.2. Relative label-free quantification

Label-free methods compare measurements of separate MS experiments without the use of stable isotope labels. Label-free methods are thus analogous to single-color arrays in the microarray world. Two approaches for label-free protein quantification can be distinguished: counting-based and intensity-based methods. The former uses the number of spectra or peptide matched to each protein as expression index, while the later relies MS^1 intensities (see fig. 2.4).

Counting approaches have the assumption that peptides are sampled proportionally to protein abundances. “Spectra count” or “peptides count”, i. e. the number of unique spectra or peptide that match to the protein sequence, are readily available statistics and easy to understand. The

"protein abundance index" PAI (Rappsilber et al., 2002) is a derivation of the peptide count. It incorporates the number of observable peptides for each protein in its formula. Ishihama et al. (2005) proposed an exponential form of the PAI (emPAI) and demonstrated linear correlation of its value to protein abundance. Zybailov, Florens, and Washburn (2007) showed that a spectral counting measure, also taking into account protein length, is more accurate. The "normalized spectral abundance factor" additionally normalizes run-to-run variability, and has been extended to further to distribute counts for shared peptides among its proteins (distributed NSAF, dNSAF, Ying Zhang et al., 2010). Both Ying Zhang et al. (2010) and Ishihama et al. (2005) demonstrated linear correlation of their respective measure with absolute protein abundance in test datasets. In that way, these dNSAF and emPAI may be seen as absolute quantification methods. For real-world samples, however, that correlation is very limited due to extensive separation, high sample complexity, and limited sampling of each peptide. In general, the spectrum count based methods have a higher dynamic range than peptide counting (Bantscheff et al., 2012). But both low sensitivity in detecting differentially abundance in proteome-wide experiments. In proteome-wide experiments, many proteins are present at the detection limit, and are seen with a low number of spectra. The comparison of the abundance of such proteins is very inaccurate due to sampling stochastics.

Feature-intensity-based label-free approaches use the MS signal intensities (Nahnsen et al., 2013). The most important steps are signal processing (including baseline and noise filtering, centroiding and charge estimation), feature finding (which tries to find all signals caused by specific peptides), and alignment of the feature maps of different samples (Nahnsen et al., 2013). It is essential that the feature maps of the different samples are aligned well. A very good reproducibility of the liquid chromatography is thus a prerequisite. Recent implementations of feature-intensity based label-free quantification are available in OpenMS (Weisser et al., 2013), MaxQuant (Jürgen Cox and Mann, 2008) and Skyline (Schilling et al., 2012).

Intensity-based approaches generally have a higher sensitivity than counting based approaches (Nahnsen et al., 2013). With label-free methods, theoretically unlimited numbers of samples can be compared. However, label-free methods have disadvantages compared to labeling methods which are described in the next paragraphs. First, more instrument time is needed, as the samples cannot be multiplexed. Furthermore the additional variability in the chromatography - which has to be matched across samples - and feature detection in intensity-based approaches lead to poorer reproducibility and accuracy.

2.2.3. Relative quantification with stable isotope labels

Labeling approaches involve the incorporation of mass labels with different isotope compositions (Gouw, Krijgsveld, and Heck, 2010). The labels can be either incorporated by the metabolism of

the studied organism *in vivo*, or are chemically attached to the extracted proteins or peptides *in vitro* (see table 2.1). Differentially labeled samples can be pooled and co-analyzed: As stable isotopes are exhibit nearly identical chemical behavior, the labeled peptides co-separate and co-elute on the chromatography into the mass spectrometer. The mass spectrometer then can separate the differentially labeled species based on their $\frac{m}{z}$ in MS^1 and represents them as pairs (or triplets) of peaks which are spaced with a certain mass difference. For MS^2-based isobaric approaches, the differentially labeled species share the same peak in MS^1 and produced sample-specific reporter ion peaks in the fragment spectra.

The following paragraphs review the various strategies for labeled quantitative proteomics. Table 2.2 summarizes their advantages and disadvantages.

Metabolic *in vivo* labeling is achieved by growing an organism in media which contains only heavy isotopes of certain essential molecules. The heavy molecules are metabolically integrated into the organism and built into the proteins. Cell cultures, prokaryotic organisms, plants and mammals have been successfully labeled - the later by the use of specific "heavy" diets (Gouw, Krijgsveld, and Heck, 2010). The stable isotopes are commonly in nitrogen or carbon sources Oda et al. (1999) and Washburn et al. (2002), or in heavy isotope variants of essential amino acids Ong et al. (2002). In metabolic labeling, the incorporation is done at the earliest possible moment in sample preparation. This means, the differentially labeled samples can be mixed at the earliest possible time point, minimizing variation due to differential sample handling. Only living organisms may be labeled metabolically - tissue samples or body fluids, for example, are out of reach (Gouw, Krijgsveld, and Heck, 2010). To circumvent this issue, Geiger et al. (2011) proposed the use of labeled cell cultures to use as a standard against human tissue, for example. A big disadvantage with this approach is the high number of differences in protein abundances between cell line and tissue.

Differential *in vitro* labeling can be applied to any sample after protein extraction. ^{18}O labeling occurs proteolytically by performing protein digestion in heavy $H_2^{18}O$ water (Yao et al. (2001), see Miyagi and Rao (2007) for a review). Chemical labeling with isotope-coded mass tags can occur on reactive amino acid side chains (Ong and Mann, 2005). S. P. Gygi et al. (1999) developed the isotope-coded affinity tag (ICAT) with a biotin group that is used for purification. ICAT targets the sulfhydryl-group of cysteines. Dimethyl labeling (J.-L. Hsu et al., 2003), tandem mass tags (TMT, Thompson et al., 2003), isobaric tags for relative and absolute quantification (iTRAQ, Ross et al., 2004), isotope-coded protein label (ICPL, Schmidt, Kellermann, and Lottspeich, 2005), and others are modifying primary amines (at N-terminus and lysine residue side chain). The di-methylation is a very cheap, fast, and specific (Kovanich et al., 2012), however might suffer

Table 2.1: Overview of quantitative proteomics techniques. (Table adapted from Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91)

Labeling method	#[1]	Incorporation	Reference
Labels producing mass difference observable in MS[1]			
Stable isotope labeling in cell culture (SILAC)	3	metabolical	Ong et al. (2002)
^{15}N labeling	2	metabolical	Oda et al. (1999)
^{13}C labeling	2	metabolical	
Isotope-coded affinity tag (ICAT)	2	chemical	S. P. Gygi et al. (1999)
Isotope-coded protein label (ICPL)	2	chemical	Schmidt, Kellermann, and Lottspeich (2005)
^{18}O labeling	2	enzymatical	Mirgorodskaya et al. (2000)
Dimethyl Labeling	3	chemical	J.-L. Hsu et al. (2003)
Neutron encoding (NeuCode) SILAC	6[2]	metabolical	Hebert et al. (2013a)
Isobaric labels producing reporter ions in MS[2] fragment spectrum			
Isobaric tags for relative and absolute quantitation (iTRAQ)	8	chemical	Ross et al. (2004)
Tandem mass tags (TMT)	6	chemical	Thompson et al. (2003)
TMT exploiting mass defects	10[3]	chemical	Werner et al. (2012)
Deuterium isobaric amine-reactive tags (DiART)	6	chemical	J. Zhang, Wang, and S. Li (2010)
N,N-dimethyl leucines (DiLeu)	4	chemical	Xiang et al. (2010)
Isobaric peptide termini label (IPTL)	2	chemical	Koehler et al. (2011)
Caltech isobaric tags (CIT)	2[4]	chemical	Sohn et al. (2012)

[1] Number of labels [2] NeuCode is not yet commercially available [3] TMT 18-plex should be available soon (Werner et al., 2012) [4] CIT has modular structure of tags with click chemistry - a high number of tags is theoretically possible

Table 2.2: Advantages and disadvantages of incorporating stable isotope labels by metabolic or chemical processes, and MS^1 *versus* MS^2 (isobaric) based quantification. (Table adapted from Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91)

	Incorporation	
	Advantage	Disadvantage
metabolical	Incorporation at organism level - lowest experimental variation.	*Per se* not applicable to tissue or body fluids - samples have to be grown in defined media.
chemical	Applicable to any sample. Fast.	Higher variation due to labeling and proteolysis efficiencies.
	Quantitation	
	Advantage	Disadvantage
MS^1-based	Straightforward label design. Quantification based on MS^1 peak envelope (more data points).	Signal congestion by increased MS^1 complexity. Limited in the number of samples.
MS^2-based isobaric tags	Multiplexing of up to 10 samples. No increase in MS^1 complexity.	Quantitation based on few MS^2 spectra. MS must be able to to analyze low $\frac{m}{z}$ region. Coelution hampers accuracy.

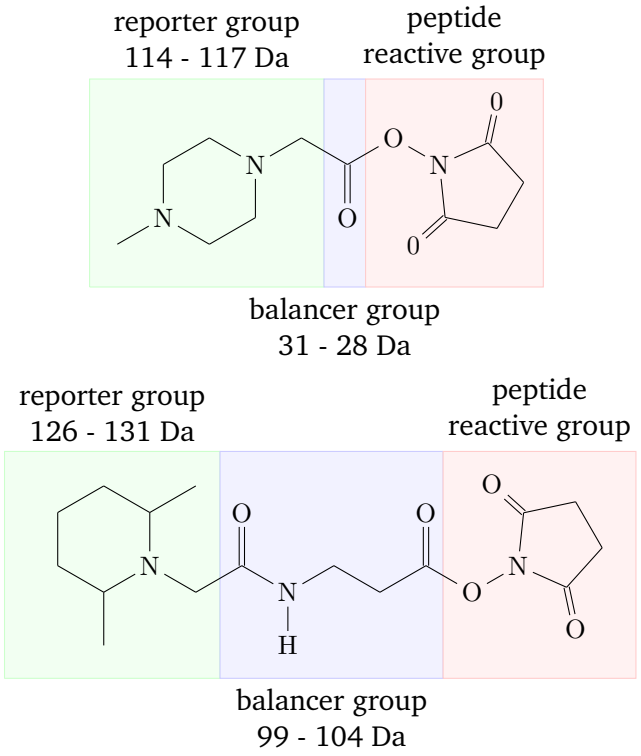


Figure 2.5: Structure of iTRAQ and TMT labels. (a) iTRAQ fourplex tags feature a mass reporter group with 114 to 117 Dalton, balanced by a group with 28-31 Dalton, giving a total of 145 Dalton. (b) TMT sixplex have reporter group masses of 126-131 Dalton, balancer group of 99-104 Dalton, giving a total of 230.

from partial tagging or chromatographic separation.

2.2.4. MS^2-based quantification using isobaric tags

Normally, mass labels differentially add weight to the peptide mass, separating the peptide peaks of different samples by defined mass deltas (usually 4 - 6 Dalton) in MS^1 (see fig. 2.6). Isobaric tags are special: Built up of isotope-coded reporter groups and balancer groups, they are nearly isobaric, leading to co-selection for fragmentation of differentially labeled peptides (Christoforou and Lilley, 2012). Upon fragmentation, the reporter groups break off, and create unique peaks in the MS^2 mass spectrum (see fig. 2.5). The benefit of isobaric tags is that they can be highly multiplexed. MS^1 mass labels increase the complexity in MS^1 because differentially labeled peptide precursors appear at distinct $\frac{m}{z}$ values. The higher MS^1 complexity causes the ion current to be divided between the labeled species and thus decreases the sensitivity and sampling depth.

Isobaric tags do not increase the complexity of the MS^1 spectrum, as the reporter groups only dissociate upon fragmentation (Christoforou and Lilley, 2012). Different tagging kits with up to 10 labels are commercially available. High multiplexing capability opens the venue for more experimental designs, such as time series, and enables the use of more replicates in one experiment. Additionally, the total analysis time (and thus the cost) is reduced. Commercially available tags are 2, 6, and 10-plex tandem-mass tags (TMT) from Thermo Scientific (Thompson et al., 2003), and 4 and 8-plex iTRAQ from Applied Biosystems (Ross et al., 2004) (see table 2.3). Most of the tags in the kits have approximately 1 Dalton difference (i. e. by coding a ^{13}C at the position of a ^{12}C in the lesser tags), but not the tags in TMT 10-plex, where the recognizable differences are based on mass-defects. Previous studies found a decrease in identifications from iTRAQ fourplex to TMT sixplex to iTRAQ 8-plex (Pichler et al., 2010). The identification rate with TMT 10-plex should be on par with TMT sixplex since the basic tag structure is the same. Sohn et al. (2012) introduced modular tags that can be synthesized with variable number of isotopes and thus have multiplexing capabilities.

The reporter groups of isobaric tags have masses in the range of $110\,\mathrm{Da}$ to $130\,\mathrm{Da}$. For their identification, it thus is required to use a fragmentation mode and mass analyzer which cover this region (Bantscheff et al., 2012). The hitherto most widely used method of generating peptide fragment spectra, CID in the ion trap of a mass spectrometer is therefore not applicable: Ion traps imposes a low mass cutoff, also called the “one-third rule”. The radiofrequency amplitude in the ion trap is set according to the precursor ion $\frac{m}{z}$. Ions which are below of 25-30% of the precursor $\frac{m}{z}$ cannot trapped, and thus not detected (see section 2.1.2 and Y.-H. Yang et al., 2009). For a while iTRAQ and TMT experiments have therefore been done on TOF instruments with a much better mass range. Nowadays, the most common method is HCD collision combined with orbitrap mass analysis (Yi Zhang et al., 2009). HCD acquisition was initially, in comparison

Table 2.3: iTRAQ and TMT tagging kits, their total tag mass, and reporter ion masses after HCD. All kits are available with amine-reactive tags. TMT 2-plex and sixplex kits are additionally available with carbonyl-reactive and sulfhydryl-reactive tags.

tag	tag mass	reporter ion masses [Da]
iTRAQ fourplex	145.1 Da	114.1112, 115.1083, 116.1116, 117.1150
iTRAQ 8-plex	301 Da	113.107 87, 114.111 23, 115.108 26, 116.111 62, 117.114 97, 118.112 01, 119.1153, 121.1220
TMT 2-plex	230 Da	126.127 725, 127.131 079
TMT sixplex	230 Da	126.127 725, 127.124 760, 128.134 433, 129.131 468, 130.141 141, 131.138 176
TMT 10-plex	230 Da	126.127 726 1, 127.131 080 9, 127.124 761, 128.134 435 7, 128.128 115 8, 129.137 790 5, 129.131 470 6, 130.141 145 3, 130.134 825 4, 131.138 180 2

with CID, very slow. Köcher et al. (2009) proposed the sequential acquisition of a short HCD scan and a standard CID scan for each precursor $\frac{m}{z}$, The HCD scan would be used solely for the quantitative information from reporter ions, while the identification can be inferred from the CID scan. As the acquisition speed of mass spectrometers greatly improved over the last years, the sole use of HCD scans is most common. However, in various phosphorylation experiments, we observed that the combination of full HCD scans and full CID scans, which are both matched against the protein database, results usually in about 30% more peptide-spectra matches. We thus propose this combined use of CID and HCD for the deep investigation into a sample. The software tool presented in chapter 4 can transparently integrate and merge the identifications.

2.2.5. Current developments in isotope labeling

Until recently, all isotope-coded mass labels were spaced by multiples of roughly 1 Dalton, the mass of a neutron. With new high-precision instruments being able to resolve more peaks, however, novel tags are being introduced that have mass differences in the milli-Dalton (mDa) range.

The basis for these tags are mass defects that enable to differentiate between isotopologues (Sleno, 2012). In short, the weight of an atom is not equal to the sum of its unbound parts: The binding energy consumes mass. The difference in bound to unbound mass Δm equals to $\Delta m = \frac{\Delta E}{c^2}$, where ΔE is the difference in (binding) energy, and c the speed of light (Einstein, 1905). For example, the sum of the parts of a ^{12}C atom (consisting of six protons, six neutrons and six electrons) is equal to $12.098\,943$ u. The actual mass of the molecule, however, is 12 u. Further on, there is a recognizable difference between the mass delta of isotopes. For example, ^{13}C-^{14}N and ^{12}C ^{15}N versions of a isotope-coded tag have a mass difference of about 6 mDa.

With new high-precision instruments, such small mass difference can be resolved. McAlister et al. (2012) first reported novel TMT tags exploiting mass defects (see section 2.2.4 and fig. 2.5). The mass difference of 6 mDa (milli-Dalton) requires instruments with a resolution of 50,000. Werner et al. (2012) first employed 8-plex TMT tags, and also 10-plex TMT mass tags are currently on the market. Everley et al. (2013) described two novel sets of sixplex TMT tags that can be combined with the standard sixplex set, to have effectively 18-plex. These tags are not on the market yet.

Hebert et al. (2013a) introduced the concept of mass defects to MS^1-based labels. They report novel SILAC tags termed NeuCode SILAC. For a heavy lysine molecule with $+8$ Da, 39 isotopologues exist when considering the heavy isotopes ^{13}C, ^{2}H, ^{15}N and ^{18}O. The 39 isotopologues are spaced roughly 1 mDa apart, which is too small to resolve with current techniques. With high-resolution Fourier-transform instruments, most peaks with differences of 12 mDa can be resolved. With this resolving power, four of the 39 isotopologues could be distinguished. Hebert et al. (2013a) provided a proof of concept with with 36 mDa spaced lysine $+8$ Da molecules. Other SILAC amino acids, which are at $+4$ Da and $+12$ Da could be used neutron encoded in the same way.

A benefit of the close spacing of the different tags is that they are co-selected for MS^2 fragmentation - in standard MS^1-labeling, the ion current is divided between the labeled species (see section 2.2.4). The MS^2 spectrum does change, however: MS^2 fragments appear as doublets, similar to IPTL (Koehler et al., 2011) spectra. The paired MS^2 fragment ions could be used to determine, which peaks are real peptide fragments and not noise. C-terminal product ions could be identified and used for de-novo mapping of the fragment spectra (Richards et al., 2013).

Combining MS^1 and MS^2 tags, the number of samples that can be analyzed simultaneously is multiplied. Dephoure and Steven P. Gygi (2012) combined triplex metabolic with sixplex isobaric labeling summing up to a total of 18 samples that were analyzed simultaneously. A year late, the number of possible samples was tripled to a total of 54 by using 18-plex isobaric tags with triplex metabolic tags (Everley et al., 2013).

2.3. Data processing and algorithms for quantitation

This section provides an overview of the methods to go from the raw mass spectrometric measurements to the inference of differential protein abundance. Figure 2.6 list some steps: peak or reporter ion extraction, preprocessing and normalization, merging with protein identification, protein ratio computation, and statistical modeling and inference. We contrast the specificities of MS^1 and MS^2 labeling techniques, which are predominantly apparent only in the first steps of the analysis. However, the main focus of this thesis is development of methods for MS^2-based

isobaric quantification methods, thus also the combined part is partial in its focus on this type of tags.

2.3.1. Specificities of MS^1 labeling

In MS^1-based labeling, quantitation is performed on the peptide features of corresponding peptide pairs or triplets. Typically, ion chromatograms are extracted and integrated over their elution time, similar to label-free intensity based approaches (see section 2.2.2). The extracted ion chromatograms (XIC) of corresponding peaks can be used for quantitation.

The detection of peptide features is a crucial part in data extraction. Overlapping isotope patterns can be caused by incomplete labeling, or a small mass differences between labeled and unlabeled species. Peptide features then cannot be clearly separated.

For many MS^1 labeling methods, the observed mass difference between labeled and unlabeled species depends on the number of times a tag has been integrated into the sequence. SILAC-labeled peptides have a mass difference in the range of $4\,\text{Da}$ to $10\,\text{Da}$, are usually nearly completely labeled, and are then well separated in the intensity profile (Ong et al., 2002). ^{18}O, when completely labeled, introduces a mass shift of $4\,\text{Da}$, but is prone to incomplete labeling (Gouw, Krijgsveld, and Heck, 2010). The different dimethyl label add $28\,\text{Da}$, $32\,\text{Da}$, and $36\,\text{Da}$, thus being spaced $4\,\text{Da}$ apart (Kovanich et al., 2012). For ^{15}N labeling, the spacing depends on the number of nitrogens in the peptide sequence.

The isotope patterns, especially of large and multiply charged peptides, can overlap. The separation of overlapping peaks is challenging, and several methods have been proposed for the different methods (Matthiesen, 2007; Matthiesen and Carvalho, 2010).

2.3.2. Specificities of MS^2 labeling

Extraction of reporter intensities

The quantitative information of MS^2-based methods resides in the product ion spectrum (Christoforou and Lilley, 2012). The most commonly used methods, iTRAQ and TMT, produce reporter ions in the low-$\frac{m}{z}$ region of the MS^2 spectrum, where the interference from fragment and immonium ions is low (Bantscheff et al., 2007b). Knowing the $\frac{m}{z}$ of the reporter ions of the different tags, the extraction is in principle straight-forward. Fragment spectra, however, can be collected in either profile or centroid mode. The profile mode shows the original continuous series of signals and thus needs integration (or centroiding) of the peaks beforehand. Most software tools rely on centroid data. Notably the R package `MSnbase` (Gatto and Lilley, 2012) provides the possibility of handling profile data, with various methods to integrate or interpolate

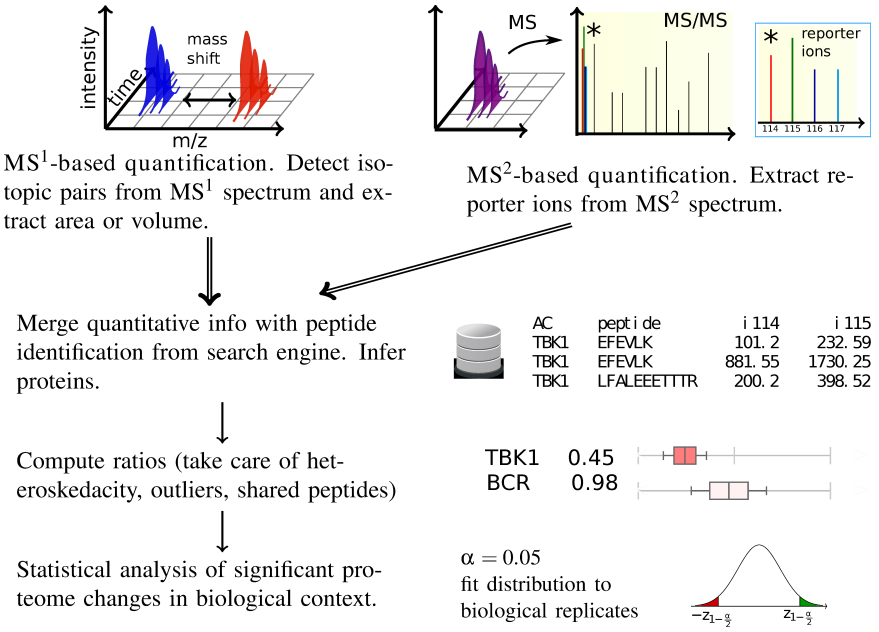


Figure 2.6: Quantification by stable isotope labeling. MS^1-based labeling induces mass shifts of all labeled peptides. In combination with the unlabeled sample, all peptides appear as peak pairs which are separated by a specific mass delta. The peptide features are extracted and used to estimate their differential abundances (top left). MS^2-based isobaric tags do not induce mass shifts in MS^1, as the tags contain balancer groups such that each total tag mass is roughly equal. The presence of the differentially labeled peptides is only revealed upon fragmentation, when the reporter and balancer groups of the isobaric tags dissociate, and significant reporter ion peaks appear in the MS^2 spectrum (top right). The quantitative information is extracted, preprocessed, and merged with identification information from the protein search engine. Proteins are then inferred and ratios calculated, before the statistical significance of the changes are assessed, and downstream analysis performed. (Figure and text adapted from Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91)

over the peak. The advantage of the profile mode is more data to infer signals, however the raw data has a significantly larger file size than centroided peak lists.

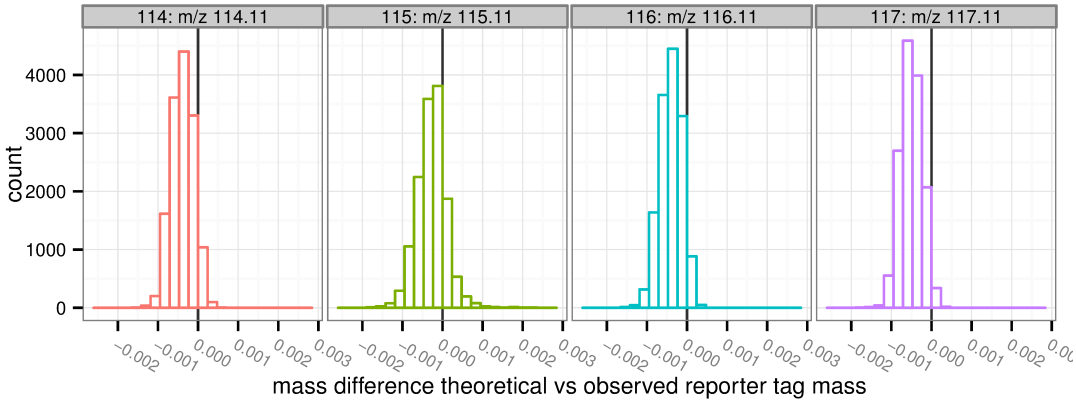


Figure 2.7: Histogram of reporter mass precision. Histograms show the difference of extracted iTRAQ fourplex reporter ion $\frac{m}{z}$ to their true mass. Generated using R package `isobar` on data from Phanstiel et al. (2011), reanalyzed in Breitwieser and Colinge (2013), figure unpublished.

The histograms of reporter ion masses provide a graphical way to check for major interferences in the data, potential problems in certain channels, and whether the fragment precision window was set reasonably (see fig. 2.7). The software tool developed in chapter 4 can generate automated quality control reports with reporter mass precision plot - which to our knowledge no other iTRAQ / TMT analysis tool does.

Isotope impurity correction

It is not possible to enrich isotopes to 100%. Consequently, all isotope labels are, to a certain extend, impure. For isobaric tags, this means that a certain percentage of reporter groups differ by one or more Dalton from the nominal mass, which causes signal transfer from one to another channel. For iTRAQ and TMT, the manufacturers measure the impurities for each batch and supply correction matrices with the relative frequencies of signal transfer for each kit (Christoforou and Lilley, 2012).

The isotope impurity matrix, which lists the percentage of tags differing by $\pm$ one and two Da, can be transformed to a mathematically more suitable form. For a kit with n tags, a $n \times n$ matrix A is created, where the entry in the cell $a_{i,j}$ corresponds to the amount of signal tag i brings to channel j (see table 2.4). To correct for isotope impurities and get the (unobserved) actual intensities of the molecules, the equation can be phrased as the following system of linear equations:

$$Ax = b\ , \tag{2.3.1}$$

where A is the $n \times n$ isotope impurity correction matrix, $b \in \mathbb{R}^n$ is the vector of observed intensities, and $x \in \mathbb{R}^n$ the (unobserved) real intensities. This can be rewritten for x as the multiplication of the matrix inverse A^{-1} with b:

$$x = A^{-1}b \tag{2.3.2}$$

When corrected intensities in x are negative, they are usually set to 0 (Lin et al., 2006; Arntzen et al., 2011), or flagged and ignored (Perkins et al., 1999). Bielow (2012) proposed parameter estimation by non-negative least squares with the added constraint that x contains no negative values.

Table 2.4: (A) Structure of an isotope impurity correction matrix. (B) Example of a iTRAQ fourplex impurity matrix with vendor-supplied values.

(A)

	A	B	C	D
A	A	A+1	A+2	A+3
B	B -1	B	B+1	B+2
C	C-2	C-1	0	C+1
D	D-3	D-2	D-1	0

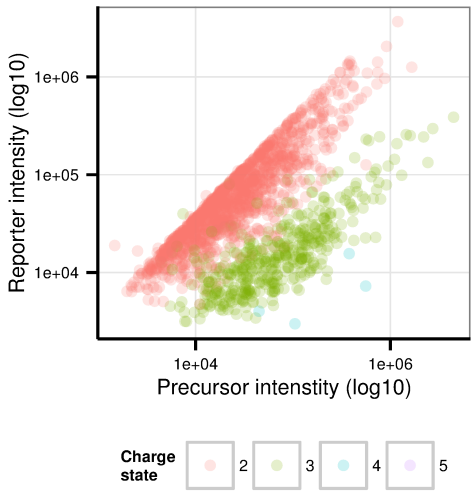
(B)

	114	115	116	117
114	0.93	0.06	0.00	0.00
115	0.02	0.92	0.06	0.00
116	0.00	0.03	0.92	0.04
117	0.00	0.00	0.04	0.92

Fold change bias due to interference by coeluting peptides

MS^2 based quantification methods have issues with accuracy: In complex samples, the ratios get compressed, or, even worse, distorted. The commonly accepted cause are co-eluting peptides with masses within the precursor mass selection window that are co-selected and fragmented (Ow et al., 2009; Christoforou and Lilley, 2012). The fragment spectrum thus is the product of the precursor, and all interfering species. For identification, these chimeric spectra are a complicating factor but no big issue as long as there is no single very prominent co-eluting peptide. High-intense peaks that are not explained by the precursor can reduce the search engine score for the sequence to spectrum match (Houel et al., 2010). In general, however, the fragments of interfering species scatter across the mass range with lower intensities than the precursor fragments, and thus do not cause big problems.

For MS^2 quantification, however, the situation is different: The reporter $\frac{m}{z}$ is fixed across all peptides, and therefore, all interfering species add their share to the same peaks. While the individual interfering signals may be small, it adds up. A further complicating factor is that the precursor abundance in MS^1 does not relate directly to reporter fragment abundances in MS^2, especially across charge state (see fig. 2.8). The distortion can be substantial (Ow et al., 2009; Ting et al., 2011; Wenger et al., 2011).



charge	0%	25%	50%	75%	100%
2	0.22	1.69	2.52	3.41	12.78
3	0.03	0.14	0.21	0.37	5.36
4	0.01	0.02	0.03	0.05	0.09

Figure 2.8: (A) Scatterplot of precursor intensities versus the summed reporter intensities per charge state. Each point represents a peptide with a specific modification state and charge (median of intensities were taken for multiple spectrum matches). (B) Quantiles of reporter intensities divided by precursor intensities. Data for both taken from test dataset 1 in Breitwieser et al. (2011).

Better separation reduces the sample complexity and thereby coelution (see also section 2.1.1). Extensive offline fractionation in multiple dimensions, better chromatography, and just longer gradients have been proposed (Ow et al., 2011; de Jong and Griffin, 2012; Lau et al., 2011). The trade-off for most of the improvements in sample separation is an increase in instrument time. Furthermore, many improvements are just incremental (Christoforou and Lilley, 2012).

Naturally, narrowing the precursor isolation window should decrease negative effects by coelution. However, no major improvements could be achieved with this strategy, as it also impacts the sensitivity. Novel MS acquisition modes can be more successful. Wenger et al. (2011) proposed a gas-phase purification of precursor ions leading to charge state reduction. Ting et al. (2011) use a second step of mass filtering on fragment ions to reduce interference: prominent fragment $\frac{m}{z}$ are selected and fragmented again to reveal their reporters. It is unlikely that coeluting ions also have similar fragment masses; therefore the issue of interference is mostly eliminated. In the second step of selection, however, it is required that the fragment ion includes the tagging group. Tryptic peptides can be labeled on the N-terminus and on lysines, C-terminal fragments thus may have no reporter groups. To circumvent this issue, Ting et al. (2011) proposed the use of the enzyme Lys-C whose proteotypic peptides cam have the isobaric tags on both terminii.

Post-acquisition methods use the MS^1 ion chromatogram to assess whether and how much interfering peptide species were present at the point before the acquisition of each MS^2 spectrum. Spectra with too much interference are usually discarded from subsequent analysis. Mikhail M.

Savitski et al. (2010) proposed the signal-to-interference measure (s2i) and released scripts to calculate the measure on Orbitrap raw data files using `multiplierz` (Parikh et al., 2009). Mertins et al. (2011) reported a slight increase in accuracy filtering by the precursor isolation purity measure of the commercial software SpectrumMill. Stringent filters for the interference do create more accurate quantification values, however at the expense of a substantial loss in quantified proteins. Mikhail M Savitski et al. (2013) thus proposed a novel method that corrects the fold-changes and leads to more accurate estimates without requiring a very stringent thresholds.

2.3.3. Data normalization

Normalization attempts to compensate for unwanted systematic differences between the channels that are caused by disparities in sample processing, labeling and digestion (Ting et al., 2009).

The structure of these systematic biases depends on the measurement technology, and is most apparent in the comparison of technical replicates. Bland-Altman plots (or Tukey's mean-difference plot) are a powerful graphical tool to inspect the correlation and agreement of values (Zaki et al., 2012). In the microarray field these plots are known as MA plots, which plot the log-ratios (M) against the average intensities (A). Microarrays show a non-linear dependency of the mean on the average, calling for a non-linear normalization function (Quackenbush, 2002). Several approaches have been developed for microarray normalization: Forcing the whole distribution to be similar by quantile normalization (Bolstad et al., 2003), fitting curves such as splines (Workman et al., 2002) or by local regression (Y. H. Yang et al., 2002), or adjusting based on additive/multiplicative models (Huber et al., 2002; Durbin et al., 2002).

In labeled proteomics data, the ratio mean normally does not depend on the intensity, i. e. there is only linear bias (Ann L. Oberg and Douglas W. Mahoney, 2012). However, as Ann L. Oberg and Douglas W. Mahoney (2012) mention, that this should be evaluated on each experiment. A known exception are instruments with a limited dynamic range, where the reporter intensities are saturated (Lin et al., 2006; Karp et al., 2010).

Figure 2.10 shows an adaption of the Bland-Altmann plot for technical iTRAQ replicates (ratio-intensity plot), with a local regression fit line that is roughly linear, thus showing no mean-intensity dependency. However, it is apparent that the variance changes as function of the mean. This heterogeneity of variance is discussed in the next subsection.

Systematic shifts of the ratios away from the often expected mean of zero do occur (Ann L. Oberg and Douglas W. Mahoney, 2012). Under the assumption that most proteins are not differentially abundant, this shift can be countered by applying scaling factors derived from the channel median or average signals, or total signal intensity (Arntzen et al., 2011). Figure 2.10 shows

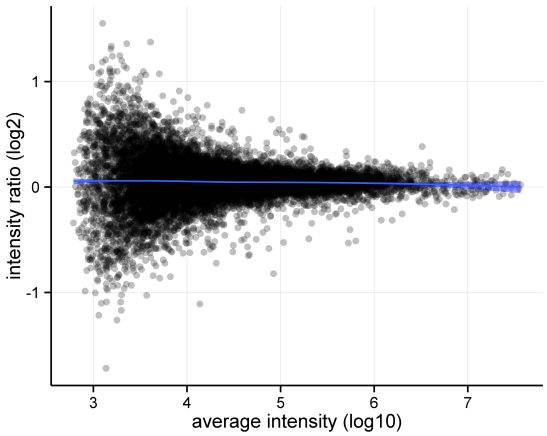


Figure 2.9: Ratio versus average intensity of reporter intensities of isobarically tagged technical replicate data. The blue line, which is nearly linear and horizontal at zero, shows the fit of a local polynomial regression line. While no intensity to mean ratio relationship is apparent, the variance shows a strong heterogeneity as function of the mean intensity.

spectrum-level data before and after normalization. Normalization leads to a centering of the ratio distribution. Using the sum of the intensities can be problematic when a small fraction of proteins account for a large proportion of the spectra. Then, even small expression changes in these highly expressed proteins can skew the normalization. These measures can also be computed based solely on house-keeping genes that are considered to be stably expressed. It is important to keep in mind, though, that the expression of proteins that are considered as house-keeping can vary significantly (Ferguson et al., 2005).

The software developed in chapter 4 allows various options for the normalization, including median, average or summed intensities. Furthermore, it allows the use of house-keeping or spiked proteins (see A).

Proteomics experiments, in which it cannot be assumed that most proteins show no change in abundance, are difficult to normalize. For example, when proteins from different affinity purifications are compared, it cannot be expected that their mean change is zero. A good strategy for affinity purification can be a normalization by the abundance of the tagged protein. In any case, the results of normalization procedures should always be carefully assessed: Big normalization factors point to problems with the data.

Some software tools normalize on the ratios of peptides or proteins rather than the channel intensities. IsobariQ supports normalization by the median ratio (Arntzen et al., 2011). Multi-Q fits a normal distribution on peptide ratios, and uses the inverse of the mode for normalization (Lin et al., 2006).

Normalization by Analysis of Variance (ANOVA)

Normalization by a scaling factor accounts only for the tag effect in a single experiment. For more elaborate experimental designs, a statistical model that captures the different sources of variation is required. Hill et al. (2008) propose ANOVA models, in which observed reporter intensities are explained by the effects of the experiment, specimen (biological replicate), isobaric tag (channel), peptide, and potentially more effect. For normalization, the protein effect can be excluded. The logarithm of the intensity can be modeled by additive effects (following the example and notation of Ann L. Oberg and Douglas W. Mahoney, 2012):

$$\log(y_{i,j,k,p,m}) = expt_i + spec_{i,j} + tag_j + pep_{k,p,m} + \epsilon_{i,j,k,p,m}$$

where $y_{i,j,k,p,m}$ represents the observed intensities in experiment i, tag j, specimen k, peptide p, and protein m. The random sources of variation are thus experiment ($expt_i$), tag (tag_j), and specimen ($spec_{i,j}$). $\epsilon_{i,j,k,p,m}$ is residual unexplained variation, which is assumed to be distributed independent and identically according to a normal distribution. The peptide effect $pep_{k,p,m}$ is modeled to account for specific peptide effects on the observed intensities, such as ionization efficiency and abundance.

The normalized intensities $y\_norm_{i,j,k,p,m}$ can then be attained by subtraction of experiment and tag specific effects:

$$y\_norm_{i,j,k,p,m} = \log(y_{i,j,k,p,m}) - [\hat{expt}_i + \hat{spec}_{i,j} + \hat{tag}_j],$$

2.3.4. Heterogeneity of variance.

It has been observed that the reporter ion ratios exhibit heterogeneity of variance (also known as heteroskedacity) as function of the intensity (Bantscheff et al., 2008; Hundertmark et al., 2009; Karp et al., 2010). The ratios of lower-intense peaks exhibit larger variation and vice versa, as seen in figs. 2.9 and 2.10.

A variety of approaches have been used to counter the effect of this issue; by the exclusion of low-intense reporters (Ow et al., 2009; Haura et al., 2011), taking the ratio of summed intensities (Keller et al., 2005; Carrillo et al., 2010), ratio estimation by two-dimensional linear regression (Bantscheff et al., 2007a), or weighted averaging when summarizing to the peptide or protein ratio from multiple spectra ratios, reducing the influence from low intense values (Hu et al., 2006; Onsongo et al., 2010). Two statistical approaches have been developed and applied for quantitative proteomics data: Variance stabilizing transformation (VST, Karp et al., 2010)

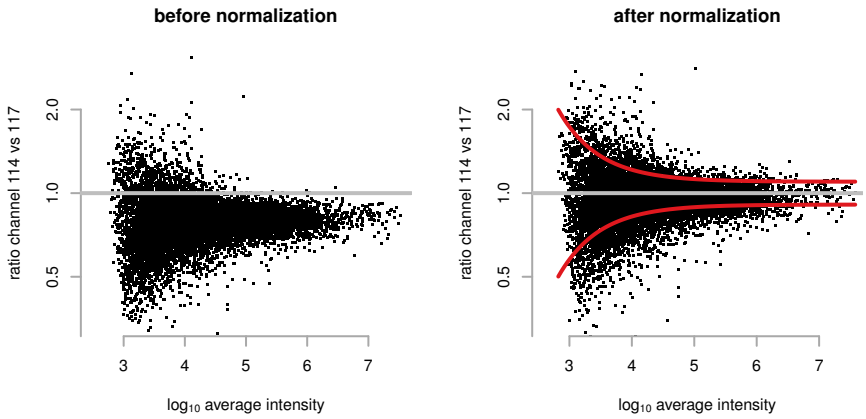


Figure 2.10: Ratio versus logged intensity plots of 1:1 isobarically tagged data before and after applying normalization shows the linear shift of the data. The heterogeneity of variance is modeled in the second plot by a noise model (red). (Figure and text taken from Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91)

and error modeling with variance functions (Hundertmark et al., 2009; Yi Zhang et al., 2010; Breitwieser et al., 2011).

Variance functions model the relationship between variance and mean, and have been applied in different areas of biomedical statistical data analysis (Davidian and Carroll, 1987), including microarrays Weng et al. (2006). Hundertmark et al. (2009) and Yi Zhang et al. (2010) demonstrated their usefulness for quantitative proteomics and phospho-proteomics isobarically labeled data. Hundertmark et al. (2009) fit the noise model parameters using a maximum-likelihood. Mandel et al. (2013) present a mixture model approach for estimating the variance function and a coherent statistical treatise of the subject.

Variance stabilizing transformations use the generalized logarithm transformation and stabilize the variance of microarray data up to the first order (Durbin et al., 2002; Huber et al., 2002) attempting to remove the variance-mean relationship. A benefit of VST in microarray data analysis is that - in contrast to logarithm - it is defined for negative values. A large part of the microarray data can be negative after background correction. In isobarically labeled data, however, the fraction of signals that becomes negative by isotope impurity correction is usually negligible or can be avoided altogether (see section 2.3.2 and Bielow, 2012).

The variance stabilizing transformation function can be derived by the delta method: Using a first-order Taylor series expansion, we can find a transformation $h(Y)$ with approximately

constant variance. Suppose a random variable Y with the following mean and variance:

$$\mathrm{E}[Y] = \mu \qquad\qquad \mathrm{V}[Y] = \sigma^2 = \Omega(\mu)$$

Then let:

$$h(y) = \int \frac{1}{\sqrt{\Omega(\mu)}}\,\mathrm{d}\mu \quad \Rightarrow \quad \mathrm{V}[f(Y)] \approx constant$$

This results from a first-order Taylor series expansion:

$$\begin{aligned} & h(y) \approx h(\mu) + (Y-\mu)h'(\mu) \\ \Rightarrow \quad & [h(y) - h(\mu)]^2 \approx (Y-\mu)^2[h'(\mu)]^2 \\ \Rightarrow \quad & \mathrm{V}[h(Y)] \approx \mathrm{V}(Y)[h'(\mu)^2] = \Omega(\mu)[h'(\mu)]^2 \\ \Rightarrow \quad & \mathrm{V}[h(Y)] \approx 1 \end{aligned}$$

Huber et al. (2002) assumed a quadratic error model for the dependence of the variance on the mean, which results in a inverse hyperbolic transformation function $h(y)$:

$$\begin{aligned} & v(u+k) = (c_1 u_k + c_2)^2 + c_3 \\ \Rightarrow \quad & h(y) = c_1^{-1} arsinh(\frac{c_2}{\sqrt{c_3}} + \frac{c_1}{\sqrt{c_3}} y), \text{with} \\ & arsinh(x) = log(x + \sqrt{x^2+1}) \end{aligned}$$

The transformation is approximately logarithmic at large values and linear at $y = 0$. Karp et al. (2010) demonstrated the applicability of variance stabilizing normalizations to quantitative proteomics data. Douglas W Mahoney et al. (2011) remark that VST standardizes the variance across all proteins, which is not required in downstream analysis - only equal variance within the protein is required. Furthermore, the interpretation of VST transformed data is difficult (Douglas W Mahoney et al., 2011).

In chapter 4 we present a framework for capturing variance with a noise model. The noise model is learned once per instrumental setup on technical replicate data, and can then be applied to future datasets. As the dependence on a technical replicate experiment for the estimation of model parameters can be a barrier when analyzing public or previously generated data, we further propose a method to fit the function just on the variability within proteins, which can be applied to biological experiments.

2.3.5. Shared peptides in protein quantification

Protein inference is a necessary part in shotgun proteomics data analysis, and many approaches have been suggested (see section 2.1.3 and Claassen (2012), Y. F. Li and Radivojac (2012), and Serang and W. Noble (2012)). The general goal in protein inference is the compilation of a list of the protein constituents in the sample, which contains few false positives and best explains the mass spectrometric evidence. Due to extensive sequence similarity in proteins - especially in splice isoforms and protein families - this is no straight-forward task (Nilsen and Graveley, 2010).

For protein quantification it is required to decide which peptides are combined to calculate its ratio. It is common that scientific publications state that only specific peptides were used for quantification. After the grouping proteins, however, two levels of specificity exist (see fig. 2.11). This section attempts to clarify the ambiguity, and motivate the approach of the software tool which is developed in chapter 4.

The list of peptide-spectrum matches is seen as static. It is assumed it contains only confident identifications. The usual protein grouping approach takes two steps (Colinge and Keiryn L Bennett, 2007):

First, some proteins might be indistinguishable based on the mass spectrometric evidence (proteins A and B in fig. 2.11). It cannot be said whether protein A is present, or protein B, or both. If truly only specific peptides are used, neither protein A nor B should be reported or quantified. Some approaches filter the database entries beforehand to minimize the appearance of shared peptides and indistinguishable proteins (Claassen, 2012). For example, only "canonical" protein sequences might be retained in the database, ignoring splice variants. In protein identification, the cost of this approach is that splice isoforms are not resolved even when there would be specific peptides. In protein quantification, furthermore the canonical protein uses the quantification values of the isoforms, which are potentially differentially regulated. Thus, the cost of this approach is higher as it can lead inadvertently to mingled quantification data.

We propose in thesis an approach (similiar to the one proposed in the ProteinProphet algorithm (Nesvizhskii and Aebersold, 2005)) that circumvents the above-mentioned issues: Indistinguishable proteins are collapsed into a single entry, and all the collapsed proteins are shown and reported. After the collapsing of indistinguishable proteins, the remainder of the peptide-protein identifications can be grouped. After the grouping, peptides are either specific to one protein (termed "reporter specific", rs), shared by proteins with a subset of identifications ("group specific", gs), or shared across protein sets, that have specific peptides. Most quantitative software tools are not precise in defining which set of peptides is used, and whether the peptides are reporter or group specific. The reason for this ambiguity is that when only those proteins, which have "reporter specific" peptides, are considered present, the "group specific" peptides become

"reporter specific". For example, in fig. 2.11, if protein C is not considered, the peptides 3, 4 and 5 would be specific to the first (collapsed) protein.

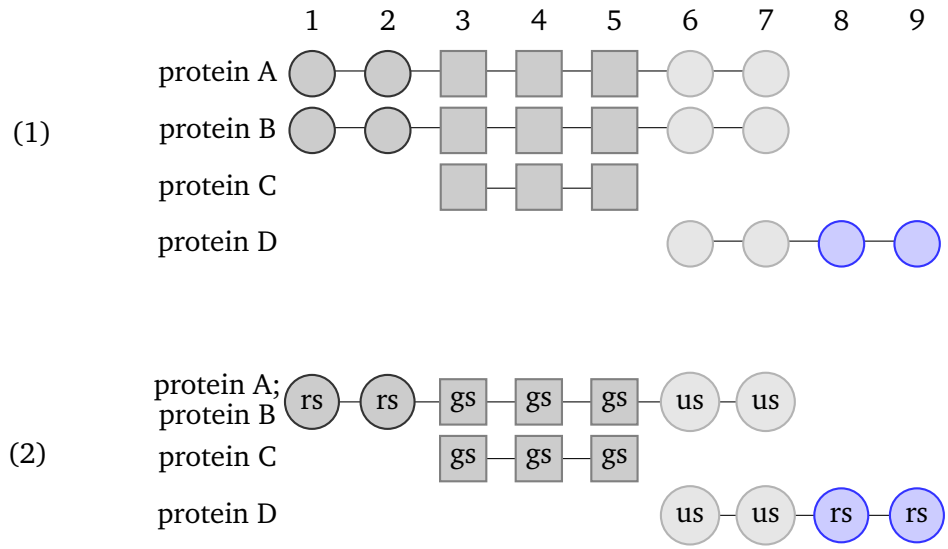


Figure 2.11: Example protein grouping. Nine peptides are identified, mapping to four different proteins in the database. rs: reporter-specific; gs: group-specific; us: unspecific

Quantification software should provide clear information, which peptides are used in the calculation of protein ratios, and preferably also give the choice to include or exclude group specific peptides. The software developed in chapter 4 performs the protein grouping based on the peptide-spectrum matches, and thus enables the control over which peptides are used for quantification. Furthermore, it could be used as basis to implement more advanced statistical protein grouping models (Gerster et al., 2013; Blein-Nicolas et al., 2012; Dost et al., 2012).

2.3.6. Calculation of protein ratios

A varying number of spectra are available per protein. The more spectra and peptide are available per protein, the more accurate and precisely the ratios can be estimated (Bantscheff et al., 2012). The first step is naturally the definition of which peptides are used for quantification of a protein (see section 2.3.5). Typically only peptides which map specifically to one of the proteins, which are considered present, are taken.

The logarithmic scale is the natural choice when working with fold-changes, as it transforms the multiplicative terms into additive terms. On the log-scale, the sum and arithmetic mean of the ratios are sensible, and thus the law of large numbers and the central limit theorem can apply.

When assuming independent identically distributed (iid) data, the sample mean is the estimator of the population mean, which has the maximum likelihood. Correspondingly, when peptide

ratios are considered to be iid with a mean equal to the true protein ratio, the sample mean of the peptide ratios is the best estimator of the true protein ratio. However, when the peptide ratios have differing variances, the maximum likelihood estimator of the mean is a inverse-variance weighted average.

Various approaches are used to summarize to protein ratios. Libra (Keller et al., 2005) uses the mean peptide ratio after removing outliers which are more than 2 standard deviations from the sample mean. MaxQuant (Jürgen Cox and Mann, 2008) uses the median peptide ratio. ASAPRatio (X.-J. Li et al., 2003) uses Dixon's test to remove outliers before calculating a peak area-weighted mean of the peptide ratios. Multi-Q (Lin et al., 2006) removes outliers outside 3 standard deviations of the mean, and calculate an intensity-weighted mean. VEMS (Rodríguez-Suárez et al., 2010) uses the median of peptide ratios after correcting for intensity biases. Similarly to Lin et al. (2006), Rodríguez-Suárez et al. (2010) observed saturation effects, and thus mean-intensity biases, which means less accuracy for higher-intense ratios.

Intensity- or area-weighted means can give an improvement over standard arithmetic means, as the precision increases with the peak intensities. However, this assumes a linear relationship between precision and intensity, which is not the case (Hundertmark et al., 2009; Karp et al., 2010). When it is possible to model the variance, inverse-variance weighted means are the better estimators. Karp et al. (2010), on the other hand, transform the data such that the variance is stabilized at a uniform value, and therewith the arithmetic mean is a good choice. Karp et al. (2010) use a 20% trimmed mean, which is robust against outliers. IsobariQ Arntzen et al. (2011), which supports the use of VST as proposed by Karp et al. (2010), calculates the protein ratio as median of the peptide ratios.

In section 4.1 we model the noise function and demonstrate the gains in precision using inverse-variance weighted means.

2.3.7. Statistical inference of protein regulation

This section covers some topics related to the statistical analysis of proteomics data; namely the issue of fold-change thresholds, statistical modeling, experimental design and multiple hypothesis testing.

Fold-change cutoffs and statistical inference

Many proteomics research papers use a fold-change threshold to select interesting proteins. Proteins with a fold change value above, say, 2 are considered regulated and interesting for further analysis. The actual value of the threshold is somewhat arbitrary, and differs between publications. For example, Polisetty et al. (2012) use the fold-change cutoff of 2 to select proteins

for subsequent analysis. Arrigoni et al. (2013) and Tolin et al. (2013), on the other hand, use fold-change cutoffs of 1.5 to select differing proteins.

To circumvent the arbitrariness of the actual fold-change threshold, several publications use a z-score or robust z-score (Jürgen Cox and Mann, 2008; P. P. Hsu et al., 2011; Arntzen et al., 2011). This approach assumes a normal distribution of the protein ratios. The z statistic is calculated as

$$z_i = \frac{r_i - \mu}{\sigma},$$

where r_i is the ratio of protein i, μ and σ are the mean and standard deviation of the protein ratios. Since the number of proteins n is usually large ($n > 100$), the use of the sample mean and standard deviation for μ and σ are justified. The p value for differential abundance of the ratio r_i is then calculated as the probability of observing a value, which is as or more extreme than z_i, assuming z_i comes from a standard normal distribution.

The actual implementations of the z score vary slightly. P. P. Hsu et al. (2011) use a robust z-score defined as the number of median absolute deviations from the mean. Jürgen Cox and Mann (2008) and Arntzen et al. (2011), on the other hand, use a robust version of the z-score, with different calculations of the standard deviations on either side, such that skewed distributions could be captured. If multiple samples are analyzed, it is usually required that the cutoff is met in all samples (e. g. P. P. Hsu et al., 2011).

There are two main issues with the z-score approach:

- The technical variability of the protein ratios is not considered. Each protein ratio may be summarized from one or 100 spectra ratios, and each spectrum ratio comes with its own precision due to the intensity effect (Hundertmark et al., 2009). To counter this effect Jürgen Cox and Mann (2008) calculate a version of the z-score, for which the estimates of population mean and standard deviation are based on intensity bins of the proteins. Each bin contains at least 300 proteins, and it is expected that lower-intensity bins have higher variability. While the binning does improve the inference, it is a very coarse-grained instrument to control the technical variability.
- It is assumed that the distribution of protein ratios is normal. However, as we will show in chapter 4, the distribution is heavy-tailed due to the differences in the variance of protein ratios.

In section 4.1, we thus present a statistical method which improves on the z-score approach. Instead of the assumption of a normal distribution, we use a heavy-tailed model of the variability. Furthermore, a second level is introduced to capture the technical variability of the protein ratio. This level assures that the protein change is seen with enough signal to be considered no random event.

Multiple hypothesis testing

The issue of multiple hypothesis testing naturally occurs in modern large-scale high-throughput experiments. For thousands of genes or proteins the same question is asked: Is this analyte regulated or not? For each individual test, the error rate α can be controlled, which signifies the expected proportion of false positives when the test would be repeated many times. When m proteins are tested simultaneously against the null hypothesis using a specific α, the expected number of false positives thus will be $F = m \times \alpha$ (Dudoit, Shaffer, and Boldrick, 2003). Thus the probability of having one or more false positives (called the family wise error rate) increases with the number of hypothesis tests. The family-wise error rate can be controlled globally by testing each hypothesis at a level $\frac{\alpha}{m}$ - the Bonferroni correction (Dudoit, Shaffer, and Boldrick, 2003). However, this procedure is very conservative, especially for large m. The false discovery rate, the expected proportion of false positives in all rejected hypothesis, is more reasonable to control in this setting: Ordering p values from smallest to largest, let k be the largest index i for which $p_{(i)} \leq \frac{i}{m}\alpha$ is chosen, and all hypotheses H_i for $i = 1, 2, \ldots, k$ are rejected (Benjamini and Hochberg, 1995). Benjamini and Yekutieli (2001) showed that the false-discovery rate is appropriate in many situations with dependence, and provided a method for general dependence. Storey and Robert Tibshirani (2003) introduced the q-value, which gives for each test the expected false discovery rate when rejecting it. This is achieved by calculating the false discovery rate in windows along the p values and fitting a spline.

The R statistical environment provides in its standard package `stats` several implementations of p value adjustment procedures via the function `p.adjust`. The software package, which is described in chapter 4, exploits this functionality to provide the choice over the various multiple hypothesis correction methods.

Experimental design and statistical power

Hypothesis testing always is a balance between false positives and false negatives - selecting proteins wrongly as “significant” when they are in fact not (type I error) versus not selecting proteins as “significant” even though they should be (type II error). The relative value of these errors depends on whether a wasted effort in downstream validation, or a lost opportunity of identifying candidates is more important.

To reduce the number of false negatives, statistical power is essential, and this comes with sample size. Levin (2011) calculated with simulation experiments that, even with a low combined technical and biological variation of 25%, four biological replicates per sample group are required to measure a fold change of 1.5 reliably. In most proteomics studies, however, the sample size is smaller, and thus the statistical power. This is also due to time (and therefore cost) constraints:

Eckel-Passow et al. (2009) calculated that a mass-spectrometry experiment which tested 50 disease and control pairs, about half a year of uninterrupted instrument time would be required. Quantitative proteomics is, for this reason, often used for hypothesis generation, where the sample pairs are considered as representative of the biological population (Karp et al., 2010).

Mass-spectrometry resources may be allocated on biological replicates, fractionation, or technical replicates. Technical replicates and fractionation increase the proteome coverage: In standard experiments, each technical replicate may add up 25% newly identified proteins. The majority of variability between experiments, though, is normally due to biological variability, and biological replicates give the greatest amount of information (Douglas W Mahoney et al., 2011).

The fundamental principle of experimental design, as introduced by Fisher (1935), are replication, randomization, and blocking. Avoiding systematic errors in the conclusions (*bias*) and being *efficient*. Replication reduces random variation thus helps to decide whether observed effects are due to chance or not and identifying true differences reliably. Randomization aims at eliminating bias from sources of confounding variation, e. g. by randomized selection of individuals and treatments. Blocking reduces variances from known sources. Douglas W Mahoney et al. (2011) propose using both the tags and the MS experiment as blocking factors by balancing groups over MS runs and tags. Ann L Oberg and Vitek (2009) discuss experimental design for quantitative proteomics in more detail. This methods in this thesis, however, are aimed at pairwise comparisons of samples, which is the most prevalent design in quantitative proteomics.

2.4. Quantifying changes of post-translational modifications

Through post-translational modifications (PTM), cells control proteins in their function, localization, and half life (Altelaar, Munoz, and Heck, 2013). Mass spectrometry can help to identify protein modification sites on a large-scale (Olsen and Mann, 2013). Choudhary et al. (2009), for example, made the first large-scale analysis of lysine acetylation and identified 3,600 sites. They thus demonstrated that acetylation is regulating not just gene expression (i. e. histones) but many protein pathways. Phanstiel et al. (2011) investigated the proteome and phosphoproteome differences of human ES and iPS cells, showing subtle but reproducible differences. Minguez et al. (2012) used mass spectrometry to decipher the network of functionally associated post-translational modification sites.

The modifications bind to specific amino acid residues or protein N-terminii and induce specific mass shifts: for example, phosphate groups (PO_4^{3-}) are bound to serine, threonine and tyrosine,

and induce a mass shift of 79.966 Da; actyl groups (CH_3CO) bind to lysines and protein N-terminii, and add 42.011 Da to the peptide mass; and methyl groups (CH_3) can bind to arginines and lysines once, twice, or trice (for lysine), and each modification adds 14.016 Da.

Phosphorylation is one of the most important and most studied post-translational modification (Olsen and Mann, 2013). Even though it is estimated that $\frac{1}{3}$ to $\frac{1}{2}$ of proteins are phosphorylated (depending on cellular state), the identification requires special enrichment (see section 2.1.1) and MS setups. Phosphorylation is volatile, and phosphorylated peptides have poor ionization efficiencies (Engholm-Keller and Larsen, 2013). Neutral losses of the phosphate group can lead to non-informative fragment spectra, but also site-determining diagnostic ions. Further fragmentation of the neutral-loss fragment (MS^3) has been proposed, as well as using the more gentle ETD and ECD fragmentation (Olsen and Mann, 2013). HCD, through high accuracy and the feasibility of immonium ion measurements, also performs well (Olsen et al., 2007). Combing ETD and HCD has been reported to improve identification and localization (Frese et al., 2013).

2.4.1. Variable modifications increase search space

Proteins are identified by matching observed fragment spectra against theoretical fragment spectra from theoretical proteotypic peptides from a protein database (section 2.1.3). When searching for modified peptides, the masses of the potentially modified residues are added to the search. For example, when trying to identify phosphorylated peptides, each serine, threonine, and tyrosine residue may bear a modification or not ("variable modification"). The search space of peptides, which may be mapped against spectra, gets exponentially larger when more variable modifications are added (Cappadona et al., 2012). A peptide with 3 such residues has $2^3 = 8$ possible phosphorylation states, a peptide with 5 $2^5 = 32$. To counter the increase in search space for modifications, the search engine scores have to be increased to keep a defined FDR.

2.4.2. Localization of modification sites is ambiguous

Protein search engine usually return the highest-scoring peptide match for each spectrum. The higher the score, the better the agreement between the theoretical and experimental peptide fragment spectrum. What the score does not tell, however, is the confidence in the localization of the modification moiety, when multiple sites are possible. Several algorithms have been proposed to provide a score or p value for the localization (Chalkley and Clauser, 2012).

The most straight-forward way to score the localization are difference scores. These approaches simply subtract the score of the next-best modification configuration. Mikhail M. Savitski et al. (2011a) proposed the Mascot Delta Score for determining localization confidence, and proved that it is comparable or better than certain probability scores (see below). The advantage of

the difference scores are there general applicability to different search engines (Vaudel et al., 2013).

Probability scoring algorithms, on the other hand, re-score the spectra against the matched peptide considering different configurations of the modification (Chalkley and Clauser, 2012). In the assignment of a certain configuration of the modification, *site-determining ions* are naturally of great importance. Beausoleil et al. (2006) developed Ascore for Sequest, and subsequent algorithms were developed based on similar probabilistic ideas for other search engines (Olsen et al., 2006) or fragmentation methods (C. M. Bailey et al., 2009), mostly showing improvements over the existing methods. Two notable exclusion with a different scoring scheme are PhosphoScore (Ruttenberg et al., 2008), which uses a graph theoretic approach, and Saeed et al. (2012), which uses a dynamic programming algorithm (unfortunately they are tied to one specific search engine (Sequest) and cannot be used else). Taus et al. (2011) developed PhosphoRS based on the Ascore probability score and enhanced the scoring, determining peak depths automatically. PhosphoRS is integrated in ProteomeDiscoverer, a commercial software by Thermo Scientific. An stand-alone version is also available, which accepts XML input and is thus search engine agnostic.

While some search engines begin to integrate these scores, the researcher is usually required to call external programs to be able to remove peptides, which have uncertain modification positions.

2.4.3. Quantification of modified peptides

Once the above-mentioned issues are resolved, the spectrum measurements can be summarized to a peptide ratio - similar to protein ratio calculation (section 2.3.6). However, for the quantification of modified peptides, much fewer data is available. Per definition it is only from one peptide, and often only single or few spectra could be matched. While the principles explained in section 2.3.6 remain unchanged, the proper use of the data is even more important.

After the ratios are calculated, usual question, which is posed at the data, is: Which sites are differentially modified? While it may seem that the data can provide an answer, the modified peptide ratio does not reflect this quantity. The modified peptide ratio reflects not just the modification state change, but also the protein abundance change (Wu et al., 2011). Therefore, if the former quantity is desired, the later has to be known, as well. Because the investigation into the modification states usually presumes its specific enrichment, the protein abundance cannot be taken from the same dataset. Instead, a separate protein-level experiment of the same samples has to be conducted (Wu et al., 2011). Then, the modified peptide ratio can be corrected and reveal the desired quantity.

In section 4.2, we re-validate the statistical models, which we developed for the protein-level analysis, for the PTM level. Using the noise model, we can provide precise weighted estimates of the modified peptide ratio. Furthermore, we present a transparent way to integrate separate protein-level experiments, and current the peptide ratio. We take into account the increase in variability of the modification ratio estimator. In the final spreadsheet reports, all three quantities are juxtaposed.

2.4.4. Databases for post-translational modifications

Protein knowledge databases such as the Universal Protein Resource (UniProt) are essential for biological research. Proteins get annotated with name, description, functional domains, placed within molcular ontologies, and much more (). It is standard practice that a certain subset of this information - at least the gene name alongside the accession code - is included in reporter on proteomics experiment. In PTM experiments, however, the resulting reports often lack annotation on the specific sites which were identified. An obvious reason is that the PTM data in public repositories is less standardized and less comprehensive, and contains more noise.

However a couple of PTM databases are available, which try to be comprehensive and provide curated datasets. The neXtProt project (Lane et al., 2012) tries to create a knowledge platform for human proteins, including their post-translational modifications, abundance, subcellular localization and interactions. neXtProt provides a REST API to access this information. PhosphoSitePlus Hornbeck et al. (2012) is a resource which contains experimentally determined PTMs from several organisms. Furthermore, there are Phospho.ELM (Diella et al., 2004; Dinkel et al., 2011) and PHOSIDA (Gnad, Gunawardena, and Mann, 2011; Gnad et al., 2007).

To our knowledge, no open tool for protein quantification harvests public PTM databases and integrates the knowledge in the quantification report. It is important for researchers to know whether a observed modification site has been previously experimentally observed or not. The PTM analysis pipeline presented in this thesis (section 4.2) generates reports which integrate PTM site knowledge available in neXtProt and PhosphoSitePlus.

2.5. Software tools for the analysis of isobarically labeled data

A number of software tools have been developed for the analysis of quantitative proteomics data. Table 2.5 lists a few recently published software tools for the analysis of isobarically tagged data.

The *isobar* software is the original contribution of this thesis, and described in more detail in chapter 4. It is implemented in R and Perl, part of the Biocondutor project, and is continually

Table 2.5: Software tools developed for the quantification of isobarically tagged data.

Software Tools (last update)	**Package**				**Quantification**			**Reports**			**PTM Analysis**		
	OS[1]	CP[2]	GUI[3]	MS[4]	Var[5]	Stat[6]	Rep[7]	PL[8]	QC[9]	A[10]	Loc[11]	Cor[12]	DB[13]
isobar v1.9 (2014)	y	y	n	y	y (nm)	y	y	n	y	y	y	y	y
Msnbase v1.11 (2014)	y	y	n	y	y (vsn)	n	n	y	n	n	n	n	n
IsobariQ v2.0a (2012)	y	y	y	n	y (vsn)	y	y	n	y	y	n	n	n
Multi-Q v1.6.5 (2010)	n	n	y	y	y	n	n	n	n	y	n	n	n
Libra v4.6.3 (2013)	y	y	y	y	n	n	n	n	n	y	n	n	n
VEMS v5 (2011)	n	n	y	y	n	y	y	n	n	y	n	n	n
iQuantitator v1.0 (2009)	y	n	y	y	y	y	y	y	n	y	n	n	n

[1] Open Source: Is the package source code available? [2] Cross Platform: Is the package usable on Linux, OS X, and Windows? [3] Graphical User Interface: Does the software has a GUI? [4] Multiple Search Engines: Is the package designed to be used only with the results from one search engine, or not? [5] Is the variance heterogeneity modeled? [6] Are statistical models for the chosing differentialy regulated proteins employed? [7] Can replicates be analyzed and summarized? [8] PipeLine: Can the processing be automated, such that the software is used in a pipeline? [9] Are quality control reports or graphics generated? [10] Are analysis reports generated? [11] Can software for the validation of PTM localizations integrated? [12] Can protein expression data be integrated to correct the modified peptide ratio? [13] Is PTM database information integrated?

developed, and open source under the L-GPL version 2 license. *isobar* was originally presented in Breitwieser et al. (2011) and Breitwieser and Colinge (2013).

MSnbase is a continually developed R/Bioconductor package which implements many features for the processing, visualization, and quantitation of mass spectrometry data (Gatto and Lilley, 2012). In comparison to *isobar*, it has a stronger base in the initial steps of data integration and supports the *mzR* package for the import of mzML and RAW (M. C. Chambers et al., 2012). *MSnbase* provides S4 class representations for qualitative and quantitative MS data, which are likely to evolve to Biocondutor's standard classes for proteomics.

iQuantitator is a further R package, which is however not part of Bioconductor (Schwacke et al., 2009). Their statistical models, which are based on the Baysian framework, allow the sharing of information across levels and can integrate replicates naturally. The models are based on ANOVA models Hill et al. (2008), restated for the Bayesian framework following Gelman (2005). The parameter inference is done using Markov Chain Monte Carlo Gibbs sampling which is written in C. While several steps have been made to lessen the computational burden, beyond a few experiments, the requirements of runtime and memory become impracticable for this implementation (Bielow, 2012). The package is intented for statisticians who can adapt the model definitions. *iQuantitator* creates rich and hyperlinked PDF analysis reports via the use of Sweave, which integrates LaTeXand R code, and TikZ for visualization. The source code is available from the web-site of the publisher of the publication, however there are no newer versions available.

IsobariQ (Arntzen et al., 2011) is a stand-alone Windows-only software tool for use with results

from the Mascot search engine. It provides a graphical user interface with various options for visualization and quantification, and can inter grate the R package `vsn` for normalization. Uniquely among free packages, it supports IPTL, which relies on paired fragment ions instead of reporter ions. Significant proteins are determined by a z-test. The latest stable version of IsobariQ is from 2012.

Multi-Q is a closed-source program, which is developed in .NET, and supports MASCOT, SEQUEST and X!Tandem search results. Profile-mode MS^2 spectra are smoothed by a 3-point moving average, then reporter ions are extracted with user-defined accuracy. Cutoffs for both lower and higher limits of intensity can be applied to counter high variability in the low, and detector saturation in the high intense region. Yu et al. (2007) released a web-server version of Multi-Q which is based on the Trans Proteomics Pipeline. The latest version of Multi-Q is from 2010, and might not work with new versions of Mascot (Bielow, 2012).

Virtual Expert Mass Spectrometrist (VEMS) is a search engine and proteomics platform, to which quantification with isobaric labels has been added (Rodríguez-Suárez et al., 2010). Apart from its own search engine, MASCOT input is supported. VEMS allows importing RAW data, and integrates intensities and subtracts the baseline similar to Multi-Q.

Libra is the quantification module of the Trans Proteomics Pipeline (TPP) (Keller et al., 2005). TPP is open source, and available for Linux, OS-X and Windows. Libra uses simple algorithms to remove outliers and quantify protein ratios. Statistical guidelines for selecting significant proteins are not provided. Through TPP, it supports many formats for input; a stand-alone version is however not available.

3. Aims of this thesis

Quantitative proteomics is a pivotal tool for biological research (Bantscheff et al., 2012). Isobaric tags, such as iTRAQ and TMT, are one of the most popular choices for labeled quantitative proteomics, and allow multiplexing of up to 10 samples (Christoforou and Lilley, 2012). As outlined in the previous chapter, adapted statistical models are crucial for the analysis of the data. Furthermore, flexible bioinformatical software is essential to handle the analysis, and present the researcher with the right information (Cappadona et al., 2012). Despite of the crucial importance of data analysis methods and tool, however, researchers have few choices in this regard. Furthermore, certain areas such as shared peptides, the distribution of protein ratios, and PTM analysis received little attention.

The main goal of this thesis was thus to develop a software tool, which can adapt to various needs of the researcher and provides statistical guidance for the selection of differentially abundant proteins. More specifically, this included the development of statistical models for isobarically tagged data, which capture technical and biological variability, and can serve to select interesting proteins; the development of a software tool for protein quantification based on the R programming language, which is versatile and extensible; and finally the development of specific modules for the handling and analysis of PTM-centric data.

The next chapter presents the results. Section 4.1 builds the foundation with an investigation into the structure of specifically generated test datasets, the derivation of models for handling technical variability, and models for selecting significantly different proteins. Furthermore, a novel software package is presented, which implements the statistical models, as well as methods for the data import from various formats, protein grouping, automatization, and report generation in Excel and PDF format. Section 4.2 builds on the protein quantification framework of section 4.1 and adds methods for the analysis of PTM-centric data, such as the calculation of localization scores for PTM sites, the correction of peptide abundance changes with protein abundance changes from a separate experiment, and the integration of public PTM databases in output reports. The performance of the methods described in sections 4.1 and 4.2 is evaluated on test as well as biological datasets from various MS platforms.

The methods and the software, which are presented in the in the results chapters, were applied and further extended in several publications. The relevant data analysis parts of these publications presented in section 4.3. Furthermore, section 4.3 describes an unpublished extension of the statistical methods, using an hierarchical empirical Bayesian model to improve the variance estimates.

4. Results

4.1. General statistical modeling of data from protein relative expression isobaric tags

4.1.1. Prologue

Chapters 1 to 3 motivated the development of software and statistical methods for the analysis of isobarically tagged quantitative proteomics data. In the attached publication (originally published by Breitwieser et al. (2011)), we investigate several aspects of the modeling of quantitative proteomics data, including:

1. Structure of technical variability in reporter ion ratios.
2. Protein grouping and shared peptides.
3. Summarizing of repeated measures for proteins.
4. Biological variability distribution of proteins.
5. Selection of significantly different proteins.
6. Identification of differentially regulated splice variants.

We propose methods for modeling the technical noise, protein quantification and inference of differential abundance. The performance characteristics of the method are evaluated on specifically generated test-datasets with spiked proteins. The broader applicability of the method is further demonstrated on samples that were generated at other institutions with different mass spectrometry platforms and isobaric tagging kits.

The methods described in the publication were implemented into the novel software package `isobar`. The package is implemented in object-oriented R code, and enables the import of commonly used file formats, statistical analysis, and report generation of iTRAQ and TMT data.

The main supporting information associated with this publication is contained in section 4.1.3. Further supporting materials and the package user manual are available online at `http://pubs.acs.org/doi/abs/10.1021/pr1012784` and `http://www.ms-isobar.org`, resp., and as attachment to this thesis.

4.1.2. Manuscript

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

General Statistical Modeling of Data from Protein Relative Expression Isobaric Tags

Florian P. Breitwieser,[†] André Müller,[†] Loïc Dayon,[‡] Thomas Köcher,[¶] Alexandre Hainard,[‡] Peter Pichler,[§] Ursula Schmidt-Erfurth,[∥] Giulio Superti-Furga,[†] Jean-Charles Sanchez,[‡] Karl Mechtler,[¶] Keiryn L. Bennett,[†] and Jacques Colinge*[,†]

[†]CeMM, Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

[‡]Biomedical Proteomics Group, Department of Structural Biology and Bioinformatics, Faculty of Medicine, University of Geneva, Geneva, Switzerland

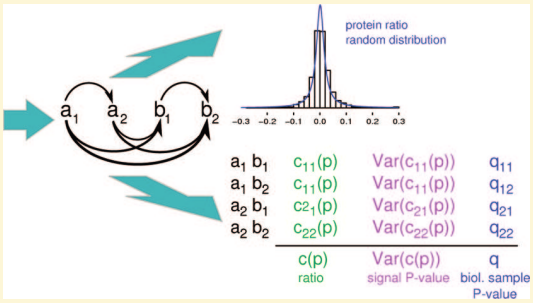
[¶]Institute of Molecular Pathology, Vienna, Austria

[§]CD Laboratory for Proteome Analysis, University of Vienna, 1030 Vienna, Austria

[∥]Department of Ophtalmology, Medical University of Vienna, Vienna, Austria

Supporting Information

ABSTRACT: Quantitative comparison of the protein content of biological samples is a fundamental tool of research. The TMT and iTRAQ isobaric labeling technologies allow the comparison of 2, 4, 6, or 8 samples in one mass spectrometric analysis. Sound statistical models that scale with the most advanced mass spectrometry (MS) instruments are essential for their efficient use. Through the application of robust statistical methods, we developed models that capture variability from individual spectra to biological samples. Classical experimental designs with a distinct sample in each channel as well as the use of replicates in multiple channels are integrated into a single statistical framework. We have prepared complex test samples including controlled ratios ranging from 100:1 to 1:100 to characterize the performance of our method. We demonstrate its application to actual biological data sets originating from three different laboratories and MS platforms. Finally, test data and an R package, named isobar, which can read Mascot, Phenyx, and mzIdentML files, are made available. The isobar package can also be used as an independent software that requires very little or no R programming skills.



KEYWORDS: bioinformatics, statistics, iTRAQ, TMT, quantitative proteomics

INTRODUCTION

Proteomic technologies provide access to the protein content of biological samples[1,2] and are important tools for current medical, biological, and systems biology research. Several highly efficient approaches also using MS exist to measure quantitative information related to proteins[3−5] that can be combined with PTM analysis.

In this work, we consider methods allowing the measurement of proteome-wide protein relative expression.[5] In general, protein digestion by an enzyme, e.g., trypsin, and tandem mass spectrometry (MS/MS) are required to identify the resultant peptides.[6] The samples for comparison are prepared such that the peptides from each of them are labeled in order to distinguish them after sample pooling and shared MS analysis. Several methods have been designed along this principle, e.g., ICPL,[7] ICAT,[8] SILAC,[9] COFRADIC,[10] $^{16}O/^{18}O$,[11] iTRAQ,[12] and TMT[13] to cite the most common ones. iTRAQ is especially convenient as (1) it can be multiplexed (up to 4 samples can be analyzed simultaneously), and (2) quantitative information resides in each single MS/MS spectrum (not necessary to combine spectra). Multiplexing is achieved through the use of isobaric tags (equal mass) to label the peptides. These tags fragment during MS/MS, thus yielding reporter peaks with distinct m/z ratios,[12] e.g., 114, 115, 116, and 117 Da. Direct comparison of the reporter peak intensities, or channel intensities, provides an estimate of relative expression. TMT (2- or 6-plex) works according to the same principle, and there exists an 8-plex version of iTRAQ; the theory we develop here applies to all of them. In this work, we are interested in the prevalent experimental settings where biological samples are compared in a single experiment (with or without replicates). Experimental design that is composed of multiple iTRAQ/TMT experiments is out of the scope of this work and has been studied by others.[14−16]

Regarding statistical analysis, iTRAQ/TMT data have similarities with gene microarray data, though they also have clear specificities. One notable difference is the variability of available information due to the variable number of measured spectra.

Received: December 23, 2010
Published: April 28, 2011

Consequently, the estimation of protein ratios comes with variable accuracy. One major breakthrough in iTRAQ data analysis was the introduction of signal intensity noise models.[17−19] We present a coherent approach that extends and improves the applicability of this concept by also modeling biological sample variability and we validate our results extensively on complex and realistic test samples comprised of albumin- and IgG-depleted human plasma background and spiked ceruloplasmins (CERU). The influence of the number of available spectra to estimate protein ratios is reported as well.

Many reported methods for iTRAQ and TMT data analysis do not provide statistical guidance to select regulated proteins, and *ad hoc* thresholds must be applied to expression fold changes. Recent developments that employed statistics[14,15,17−19] are compared with our approach. We further demonstrate the application of the proposed method to several biological data sets obtained from different MS platforms (ThermoFisher Scientific ESI-LTQ Orbitrap TMT 6-plex[20] and iTRAQ 4-plex,[21] Applied Biosystems MALDI-TOF/TOF TMT 6-plex[20]). We finally show how the presented statistical framework can provide, to our best knowledge, the first practical tool for assessing the expression of proteins with no specific peptides such as isoforms. Only theoretical work[22] addressed this question so far.

Our layered modeling enables straightforward exploitation of quantitative proteomic data and is implemented in a R package named *isobar*.

METHODS

Test Samples

Mouse and rat lyophilized CERU were dissolved and digested (trypsin) to prepare 10 fmol/μL stock solutions. These were mixed in a reciprocal fashion for each 4-plex iTRAQ channel (114:115:116:117); Set 1 = 1:2:5:10 (rat) and 10:5:2:1 (mouse), and Set 2 = 1:10:50:100 (rat) and 100:50:10:1 (mouse). A complex background peptide mixture was generated by depleting albumin and IgG (ProteoPrep) from 160 μL of human plasma. Reduced and alkylated depleted plasma was separated by 1D-SDS-PAGE, and following visualization of the proteins by colloidal coomassie staining, several regions were excised and the proteins in the gel digested in situ with trypsin.[23] The resultant background peptide mixture was extracted from the gel slices and purified, and four fractions were combined with 6 pmol of digested human CERU and mixed with Set 1 and Set 2 to obtain the final test samples (TS1) and (TS2). In TS1, 1:2:5:10 relates to 6.1:12.2:30.5:61 fmol CERU peptides, whereas in TS2, 1:10:50:100 relates to 0.7:6.8:34.2:68.3 fmol CERU peptides.

Peptides were separated at pH 10 on a Gemini-NX column (Phenomenex, Torrance, CA). Forty fractions were collected and subsequently analyzed with a hybrid LTQ-Orbitrap XL mass spectrometer (ThermoFisher Scientific, Waltham, MA) coupled to an Agilent 1200 HPLC nanoflow system via a nanoelectrospray ion source (Proxeon, Odense, Denmark). Analyses were performed in a data-dependent acquisition mode using a top 3 higher-energy collision-induced dissociation (HCD) method. MS data were searched against human Swiss-Prot[24] 2010.09, with mouse and rat ceruloplasmins appended, using Mascot[25] 2.3 and Phenyx[26] 2.6.1, imposing a protein group false discovery rate (FDR) <1%. Namely, we selected proteins with at least 2 distinct peptides above a score T_1 or a single peptide above T_2, and additional peptides for such validated proteins with score better than T_3 were also accepted. For Mascot we used $T_1 = 16$, $T_2 = 40$, $T_3 = 10$, whereas for Phenyx we used $T_1 = 5.5$, $T_2 = 9.5$, $T_3 = 3.5$ (all with P-value $<10^{-3}$); after this selection, proteins are grouped on the basis of shared peptides[6] and protein group reporters only considered. Conflicts between Mascot and Phenyx peptide identifications were discarded. The whole procedure, including protein grouping, was repeated against a reversed database[27] to assess the protein group FDR. Individual peptide identifications were found to have <0.1% false positive (FP) rate.

Biological Samples

To illustrate some of its important features, we demonstrated the application of isobar on two large data sets we published recently. The first one (TAC) is an LTQ Orbitrap data set (2D LC−MS/MS) with CID MS/MS complemented by a narrow HCD scan to measure iTRAQ 4-plex channels[21] only. The analyzed samples originated from a mouse model of cardiac stress obtained by transverse aortic constriction: left ventricles of control (sample class A, I_{114}, I_{115}) and operated (class B, I_{116}, I_{117}) animals were compared using technical replicates, i.e., two channels per sample class.

The second data set[20] (CSF) was composed of the analysis of human cerebrospinal fluid from patients suffering from sleeping sickness with both an LTQ Orbitrap and a MALDI-TOF/TOF platforms and labeling with TMT 6-plex. The two classes of samples were (A) hemolymphatic stage (first stage) and (B) meningoencephalitic (second stage) patients. Channels I_{126}, I_{128}, I_{130} measured each a pool of 3 first stage patients and I_{127}, I_{129}, I_{131} pools of 3 s stage patients.

Software

Parsers for Mascot and Phenyx were implemented in Perl and data analysis in R as a Bioconductor package. The parsers export results in a simple tabular format and support a protein validation strategy with 3 thresholds, as described above. A PSI mzIdentML[28] parser was also developed purely in R. The data import in R can deal with an arbitrary number of parser outputs in case multiple search engines are used.

RESULTS AND DISCUSSION

We present our results establishing a statistical framework adapted to the analysis of data sets where each iTRAQ or TMT channel is used for a different sample (no replicate), which we then extend to cover the situation with replicates. For conciseness, most of the text is written for iTRAQ 4-plex, but it applies generally, and to further simplify, many cases are discussed considering channels I_{114} and I_{115} as a generic example. Detailed mathematical derivations and additional figures are presented as Supporting Information (SI) and a detailed example of isobar use is also provided as Supporting Information. As we shall see, the logarithmic scale is a natural scale to analyze expression ratios, and therefore, unless otherwise specified, ratios are log ratios.

Noise Model

We denote $P = \{p_1; \ldots; p_n\}$ the n proteins identified in one iTRAQ experiment and $S_i = \{s_{i,1}; \ldots; s_{i,mi}\}$ the m_i spectra identifying protein i. Reporter intensities are written $\{I_{114,i,j}; I_{115,i,j}; I_{116,i,j}; I_{117,i,j}\}$ or $\{I_{114}; I_{115}; I_{116}; I_{117}\}$ when the context is clear. Isotopic impurities in iTRAQ reagents cause the transfer of signal from one channel to the others. Isobar can correct the intensities according to the reagent batch impurity rates,[29−31] and we performed this correction with all the samples discussed. Ratio normalization was achieved imposing equal median intensity in each channel.

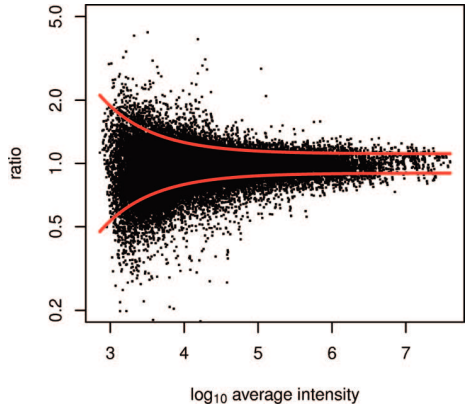


Figure 1. Log ratios (y-axis) of 1:1 samples versus average signal intensity (x-axis) display more variability for low signals (heteroscedasticity). A noise model (red line) is estimated to capture this trend and integrate it into higher level statistical models.

In samples TS1 and TS2, all of the proteins except the ceruloplasmins have a ratio of 1:1:1:1, and hence we should observe identical reporter peak intensities theoretically. In practice, this is not the case since intensity measurements contain noise, and as shown by Hundertmark et al.,[17] this noise is essentially multiplicative and varies with signal intensity (heteroscedasticity). In the log-scale, the noise becomes additive and can be analyzed conveniently. We model signal intensity dependent noise variance with the noise model

$$f(x) = a + r\,e^{-\lambda x}, \qquad (1)$$

with x the signal log intensity. We observe that over a large part of the intensity range, in small intensity windows due to the heteroscedasticity, the noise is normally distributed with mean equal to zero (data not shown). That is, for a signal log intensity x, the noise in one iTRAQ channel is modeled by a normal distribution $N(0,f(x))$ and parameters a,r,λ are learned from a 1:1 experiment, Figure 1. Parameter a stands for the minimum noise level, r is the amplitude of the signal-dependent component, and λ is the rate of its decrease. This approach has been already presented,[17,18] and the details are omitted here. Nonetheless, we made several improvements to apply it to current large data sets (TS1 contains 14991 spectra, 1648 peptides, and 157 proteins; TS2 15457 spectra, 1825 peptides, and 180 proteins), including non-1:1 samples, and to obtain more robust noise models by averaging noise functions trained on all pairs of iTRAQ channels (see Supporting Information). In general, the noise model is very stable for a given MS platform and hence it is only necessary to learn it once.

Computation of a Protein Ratio

A protein ratio $c(p_i)$ is obtained by combining the ratios measured from the MS/MS spectra of its peptide spectra $c(s_{i,j}) = \log(I_{115,i,j}/I_{114,i,j})$. The design of the test data sets, with human, mouse, and rat CERU shared peptides at different ratios, was ideal to investigate the selection of the peptides integrated in the protein ratio computation. We found that only specific (not shared) peptides can be used[32,33] (Figure S2), and even spectra from specific peptides must be filtered by eliminating outliers (Figure S3), most likely because of coeluting material.[30] We write S'_i as the subset of spectra for protein i after selection.

We then examined whether a peptide-specific bias exists in the spectra passing the above selection, but we did not observe any trend stronger than a sampling effect (Figure S4). Although one could argue that peptides that are easy to detect should give more intense signals and hence contain less noise, we explain the absence of such an effect by the fact that peptides are fragmented at various time points during their elution curve. Peptide dependence is thus ignored as other researchers have done already.[19]

Several options are available to summarize the multiple spectrum ratios into a single protein ratio, and we considered usual estimators such as the median and the average (both trimmed here because of the outliers elimination) and three different weighted averages (intensity-based Multi-Q,[34] or weighted by either the standard deviation $(f(x))^{1/2}$ or the variance $f(x)$. A modified *boosted* median evaluated on half the data consisting of the most extreme ratios was also added to address commonly observed underestimated iTRAQ ratios,[30,35] which are also assumed to be caused by coeluting material. As comparison criteria, accuracy and stability (limited variance) of the estimations were considered as well as the number of spectra available influence. We found no major differences for ratios smaller than 1:10 (or 10:1), and as expected, the larger the number of available spectra, the more accurate the estimations. For larger ratios, differences in variance and accuracy become visible and the boosted median slightly outperforms the other candidates in accuracy. Nonetheless, the weighted average

$$c(p_i) = \sum_{j \in S'_i} \alpha_{i,j} c(s_{i,j}), \qquad (2)$$

with $\alpha_{i,j} \propto (\mathrm{Var}(c(s_{i,j})))^{-1} = (f(\log(I_{114,i,j})) + f(\log(I_{115,i,j})))^{-1}$ and $\sum_j \alpha_{i,j} = 1$, has the smallest variance and is sufficiently accurate for correct biological interpretation. Based on this, we selected i as $c(p_i)$ estimator.

Protein ratio $c(p_i)$ must be complemented by an estimation of its variance to determine how reliable it is. From the individual spectrum variance, determined from eq 1, we can classically compute $c(p_i)$ variance $V_{\mathrm{estim},i}$ (details in Supporting Information). In practice, $V_{\mathrm{estim},i}$ can become very small when individual spectrum ratios are spread but in large numbers, and though mathematically correct, $V_{\mathrm{estim},i}$ can thus be misleading. A common solution[36] to this problem is to also take into account the sample variance $V_{\mathrm{spectrum},i}$ that captures measurement dispersion (see Supporting Information) and finally set

$$\mathrm{Var}(c(p_i)) = \max\{V_{\mathrm{estim},i}, V_{\mathrm{spectrum},i}\}. \qquad (3)$$

Table 1 shows that direct use of $V_{\mathrm{estim},i}$ yields too many FPs, whereas eq 3 does not (Table 1 "isobar"). There are also fewer cases where a few low intensity spectra can cause excessively small $V_{\mathrm{spectrum},i}$, thus further motivating the application of eq 3.

Biological Sample Variability Modeling

Estimation of the protein ratio variance $\mathrm{Var}(c(p_i))$ is not sufficient since it only reflects the accuracy of the ratio estimation and does not indicate whether it is significant. To determine significance, it is important to relate this ratio to biological sample variability, i.e., an additional *biological sample* model layer based on the distribution of random protein ratios. We have found that the correct theoretical model for random protein ratios is a Cauchy distribution[37] (Figure 2 and Figure S5). Cauchy is a heavy-tailed distribution that reflects inherent variability in iTRAQ/TMT data. These can become very precise with accurate instruments but, nonetheless, will always contain a few ratios far

Table 1. False Positive Rates Estimated from 1:1 TS1[a]

sp[b]	isobar	isobar $V_{estim,i}$	*t* test	fc[c]
1	0.02	0.19	0.00	0.08
2	0.01	0.09	0.07	0.05
3	0.01	0.05	0.06	0.02
5	0.0	0.02	0.21	0.01
10	0.00	0.01	0.29	0.00
15	0.00	0.00	0.37	0.00

[a] Data were re-sampled (50 times) to estimate the FP rates for any number of available spectra. Methods "isobar", "isobar $V_{estim,i}$", and "*t* test" were imposed with a maximum 5% FP rate, and fold change analysis was imposed a minimum 1.5-fold change. Isobar with $V_{estim,i}$ alone (isobar boost) instead of eq 3 requires more than 3 spectra, and fold change analysis more than 2. The t-test is too sensitive to apparent differences supported by multiple spectra. [b] Number of available spectra. [c] Fold change method.

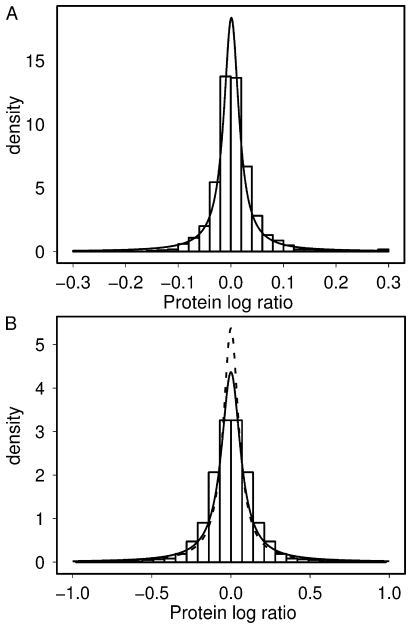


Figure 2. Protein ratio random distribution. (A) The distribution of technical replicate ratios on a LTQ Orbitrap is sharp (small ratios are significant) and fitted accurately by a Cauchy distribution (solid line). (B) Considering patient samples of the same class (CSF data set measured on LTQ Orbitrap), we also observe a Cauchy distribution (solid line) but more spread since data contain more variability. The CSF sample was also analyzed on a MALDI-TOF/TOF platform, and the corresponding Cauchy model (dashed line) is close to the LTQ Orbitrap model. This shows that isobar replicate analysis captures biological sample variability precisely. See Figure S5 in Supporting Information for a comparison with a Gaussian model.

from the correct value (even after filtering for nonspecific peptides and outliers). Remarkably, this is further supported by the observation that the Cauchy distribution is accurate for all the protein ratio estimators we assessed (average, median, etc.) and on all MS platforms (Figure S5). Cauchy belongs to the larger family of A-stable distributions,[38] which also comprises the Gaussian, but we could not obtain significantly improved fit considering the whole family.

In an experiment where each channel corresponds to a different sample, the protein ratios observed when comparing two channels are a mixture of ratios from nonregulated and regulated proteins, and therefore biological sample variability should be learned separately. If the reference is variability between technical replicates, then it is possible to directly learn random protein ratios variability from a 1:1 experiment in addition to the noise model. This distribution is reproducible, and hence can be learned once only for a given MS platform. On the contrary, if the reference is variability between biological replicates or samples of the same class, e.g., patients, random protein ratio variability must be learned by comparing replicates with each other in a preliminary experiment, Figure 2B. When no preliminary experimental data are available, one can simply select the *x*% most extreme ratios or fit a mixture model to the data at hand. Namely, we know that random protein ratios distribute like a Cauchy, and we can assume a generic distribution for the regulated proteins such as a normal. By fitting a mixture α Normal + $(1 - \alpha)$ Cauchy, we can then use the Cauchy component as an estimate of the random part; see Supporting Information. We shall see hereafter that experimental design including replicates in multiple channels, gives direct access to the random protein distribution.

Deciding for Protein Regulation

So far, we have introduced three layers of modeling: (1) the noise model for spectra; (2) the protein ratio estimator and its variance that inform on ratio accuracy in terms of signal quality (how many spectra, how intense); and (3) the biological sample variability model that identifies significantly regulated proteins. It is a natural choice to select proteins measured with a *sufficient signal* and *regulated*. Therefore, we compute a signal *P*-value (layer 2) and a biological sample *P*-value (layer 3) and require that both are better than a chosen level of risk, e.g., 5%.

Performance Evaluation and Comparison with Other Methods

The test data sets allow careful determination of isobar performance: we can exploit spiked mouse and rat CERU to measure TP rates and the numerous background plasma 1:1 proteins to measure FP rates. In addition, we can resample the data to obtain performance estimates with respect to the number of available spectra. Starting with FPs, see Table 1 "isobar", we obtain estimates in reasonable agreement with the imposed 5% that reduce with larger numbers of spectra. Concerning TPs, it turns out that large ratios are very easy to identify and therefore we only use TS1 here (TS2 ratios are larger). We report one easy rat CERU 5:10 = 2 ratio, for reference, followed by the case at the lowest concentration, rat CERU 1:2 = 2. Since noise is multiplicative, we can multiply iTRAQ channels intensities to create two lower ratios 1:1.3 and 1:1.5 in a realistic manner (the multiplication is a simple shift in the log scale, and it does not affect the noise structure). Results presented in Table 2 "isobar" show that performance scales very nicely with the ratio magnitude and the number of spectra available. Furthermore, relating the observed performance to actual protein concentrations, we note that the concentrations used in Table 2 are in the low fmol range and estimating the relative abundance of rat CERU with respect to the total amount of proteins in TS1 (by means of peptide MS intensities) we find ~2%. Therefore, the high abundant rat CERU ratio 5:10 involves material at $2 \times 5/(10 + 5 + 2 + 1) \simeq 1\%$ and the lower abundant 1:2 ratio at $2 \times 1/(10 + 5 + 2 + 1) \simeq 0.2\%$.

In Tables 1 and 2, we remark that isobar used with $V_{estim,i}$ causes too many FPs for proteins identified with small a number

Table 2. True Positive Rates Estimated from TS1 Data (Rat CERU) Resampled 500 Times[a]

ratio	sp	expt[b]	isobar	isobar boost	fc
2[c]	1	1.91	0.82	(0.97)	(0.93)
	2	1.94	0.96	(1.00)	(0.97)
	3	1.94	0.99	(1.00)	0.99
	5	1.95	1.00	1.00	1.00
	10	1.95	1.00	1.00	1.00
	15	1.96	1.00	1.00	1.00
2[d]	1	1.72	0.46	(0.88)	(0.72)
	2	1.76	0.59	(0.95)	(0.77)
	3	1.78	0.75	(0.97)	0.82
	5	1.79	0.93	0.99	0.88
	10	1.80	0.97	1.00	0.96
	15	1.81	1.00	1.00	0.98
1.5	1	1.36	0.30	(0.76)	(0.46)
	2	1.41	0.35	(0.79)	(0.41)
	3	1.39	0.33	(0.74)	0.28
	5	1.45	0.57	0.81	0.36
	10	1.49	0.54	0.90	0.27
	15	1.48	0.56	0.94	0.20
1.3	1	1.15	0.12	(0.65)	(0.28)
	2	1.20	0.15	(0.50)	(0.13)
	3	1.23	0.24	(0.49)	0.11
	5	1.29	0.34	0.56	0.07
	10	1.29	0.26	0.59	0.01
	15	1.30	0.23	0.58	0.00

[a] FP thresholds (5% and 1.5) as in Table 1. TP rates obtained where the FP rates were ≥5% (Table 1) are in brackets. We note isobar high sensitivity and accuracy on the first, more abundant, ratio 2 and the more progressive performance with more difficult data. All nonreported TS1 ratios have better TP rates than the low abundant ratio 2 reported here. [b] Experimental ratio given by isobar in the linear scale. [c] 5:10 more abundant material. [d] Low abundant.

of spectra. This could be used to boost isobar performance as soon as more spectra are available and, according to TS1 data, the limit is obviously 5 spectra, though it might be difficult to determine for an unknown sample. We can hence only recommend careful application of this boosted procedure.

Many different approaches exist for interpreting iTRAQ/TMT data. These methods have various scopes, from providing peptide and protein ratios to selecting differentially expressed peptides and proteins. Here we discuss conceptual differences with some other methods aiming at computing protein ratios and identifying differential expression and assess the performance of a few simple methods.

Even when eliminating outliers, methods that rely on the computation of a ratio through a median, an average, or similar techniques can be inaccurate due to high noise levels in low signal spectra. The commonly used application of a fold change analysis necessitates a stringent threshold that must be specified *a priori* to limit FPs. In Table 1, it is shown that imposing a minimum fold change of 1.5 yields excessive FPs when less than 3 spectra are available, and in comparison with isobar, sensitivity is similar or worse for difficult data (Table 2). See also the Applications subsection below.

Comparing isobar with methods founded on statistics, we first note that authors interested in peptide selection[17,18] obtained

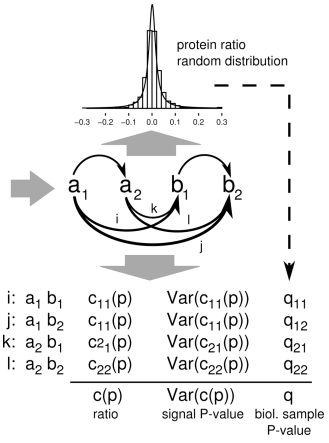


Figure 3. Analysis of experiments comparing two classes of samples (A and B), with replicates (a_i and b_i), starts with the estimation of the protein ratio random distribution (upward arrow) using samples of the same class. The real comparisons (downward arrow) involves pairwise comparisons of all of the pairs of distinct classes and an integration step to obtain the final protein ratio $c(p)$ and its *P*-values. Biological sample *P*-values are estimated on the basis of the random ratio distribution as indicated by the dashed arrow on the right.

good results determining a peptide ratio confidence interval through the noise function and requiring the exclusion of 0 from this interval. Unfortunately, as already mentioned, this does not extend to protein selection since large numbers of spectra yield small $V_{\mathrm{estim},i}$ values and many FPs are created (data not shown). Our solution to this problem was the introduction of the sample level modeling. Recently, variance stabilizing transformation (TSV)[19] was adapted to iTRAQ data. TSV considers a slightly more general noise (mixture of additive and multiplicative) and uses it to transform the spectrum ratios such that less confident ratios are reduced (closer to 1:1). A trimmed average is then applied by Karp et al. to estimate the protein ratios, which is similar to eq 2: less confident data have a decreased influence the protein ratio. We do computations in the original scale, whereas Karp et al. do them in a transformed scale and they do not provide statistical modeling of biological sample variability, whereas we do. Another common approach to select proteins is a *t* test[39] on log-transformed channel intensities but we found it dangerous when more than a few spectra are available because small differences appear significant (Table 1 "*t* test").

Use of Replicates

It is customary to apply iTRAQ or TMT to compare two classes of biological samples introducing replicates. The isobar models extend naturally to exploit additional information. Assuming two sample classes A and B, we can learn the protein ratio random distribution considering all the intraclass pairs, i.e., (a_1;a_2), (a_1;a_3), ..., (b_1;b_2), ... Such pairwise comparisons are done using the methods above, all the results are pooled, and a Cauchy model is fitted, Figure 2B. True sample comparisons, i.e. class A versus B, is achieved by combining the analysis of all the interclass pairs, i.e., (a_1;b_1), (a_1;b_2), ..., (a_2;b_1), (a_2;b_2), ..., with a new integration step to determine single signal and biological sample *P*-values, Figure 3.

Integration of the ratios is accomplished as follows. We use the weighted average of the protein ratios originating from each sample pair (a_k;b_l) with weights inversely proportional to the variances. The variance of the final protein ratio estimation is

Table 3. True and False Positive Rates for the Analysis with Technical Replicates[a]

sp	TP 2	TP 1.5	TP boost 1.5	TP 1.3	TP boost 1.3	FP	FP boost
1	0.97	0.32	(0.43)	0.13	(0.30)	0.00	0.05
2	0.99	0.61	(0.74)	0.30	(0.48)	0.00	0.06
3	1.00	0.76	0.83	0.53	0.64	0.01	0.03
5	0.99	0.93	0.90	0.69	0.79	0.01	0.02
10	1.00	0.99	1.00	0.93	0.95	0.00	0.00
15	1.00	1.00	1.00	1.00	1.00	0.00	0.00

[a] Compare with Table 2 to observe the huge improvement with respect to an experimental design without replicates. This improvement is smaller with easier data (not shown). Again, to replace eq 3 by $V_{estim,i}$ might "boost" isobar performance but is risky with few spectra.

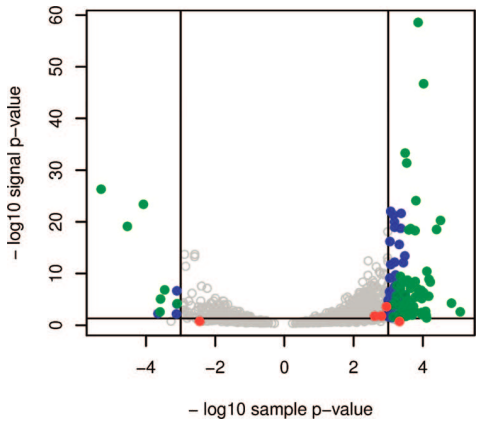


Figure 4. Volcano plot of Köcher et al. data[21] relating protein expression change (x-axis) and signal strength (y-axis). Top left and right quadrants are used to select clear (signal) and strong (ratio) changes. Proteins selected by the original method (reproducible fold change >1.5) only are in red (5 proteins), by isobar and the original method in green (76), and by isobar only in blue (54); nonselected proteins are in gray (893). All of the important proteins discussed in Köcher et al. were found to be significant by isobar, and we almost doubled the overall sensitivity (+70%). The original selection and isobar selection were done at 5% FP rates (see Table 1). Isobar biological sample maximum P-value was set to 0.1% because technical replicates were used; signal maximum P-value was set to 5%.

obtained from the variances as computed in each pair, details in Supporting Information. The integrated analysis of protein ratio significance with respect to biological sample variability also requires a new step of modeling. For a given protein p_i and from each pair of samples $(a_k;b_l)$ we obtain a biological sample P-value $q_{k,l,i}$. We multiply these P-values to obtain a product $q_i = \Pi_{k,l}q_{k,l,i}$ and we prove in Supporting Information that, assuming random protein ratios, q_i follows a theoretical distribution whose cumulative distribution function is

$$R_N(q) = q\sum_{i=0}^{N-1}(-1)^i\frac{\ln^i q}{i!}, \qquad (4)$$

with N the number of pairs used ($N = 4$ for 2 versus 2 samples, $N = 9$ for 3 versus 3). The integrated sample P-value for protein p_i is given by $R_N(q_i)$. To estimate performance, there is no replicate in TS1, but as explained above, we can precisely simulate data with repeats and create ratios 1:1:2:2, 1:1:1.5:1.5, and 1:1:1.3:1.3.

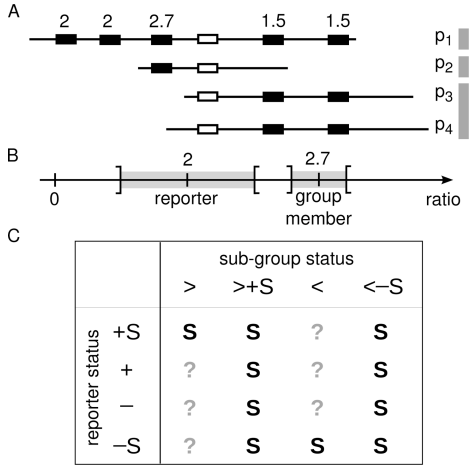


Figure 5. Principle of protein group analysis. (A) A protein group and its subgroups. The white peptide cannot be assigned to a subgroup specifically, and hence it cannot be used for group structure analysis. (B) When the group reporter ratio and a subgroup specific peptide ratio, e.g., p_2, are different enough, taking into account their estimator variance indicated here as a confidence interval in gray, we can conclude for expression of the subgroup. (C) Combinations of observations on the reporter and the subgroup ratios and their implication of subgroup ratio significance. For the reporter, "+" and "−" indicate the sign of the ratio (log-scale) and "S" significance; for the subgroup, ">" and "<" indicate larger/smaller ratios, "+"" and "−" the sign, and "S" significance of the subgroup without referring to the reporter significance. The table cells give "S" for significance or "?" for unknown.

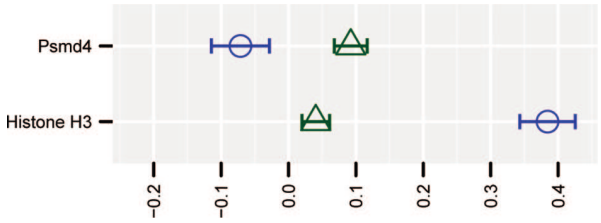


Figure 6. Two examples of protein groups from TAC data set that we could analyze (11 in total, see Supporting Information). Group reporters are in green and group members in blue; dashed lines indicate significant protein ratios. Psmd4 is 26S proteasome non-ATPase regulatory subunit 4, isoforms Rpn10A, Rpn10C, Rpn10E (coreporters) versus Rpn10A, Rpn10D (group members, one subgroup); Histone H3.1 and H3.2 versus H3.3.

TP estimation shows the expected gain in sensitivity and FP rates rigorously estimated on the many non-CERU 1:1:1:1 proteins are low, Table 3.

Practical experience shows that when comparing highly variable samples such as patient samples, certain pairs $(a_k;b_l)$ may yield some weaker ratios with a contradictory direction though the general trend is clear. Additionally, some pairs may yield no ratio because one iTRAQ channel has no signal. The replicate integration framework proposes here is very flexible with respect to these difficulties: opposite direction ratios can be given an appropriate weight to moderate the overall ratio and changing N in eq 4 adapts to the number of available pairs as a χ^2-test does with the number of degrees of freedom. The isobar library code addresses all these aspects transparently (mathematical details in Supporting Information), and we note that when $N = 1$, the

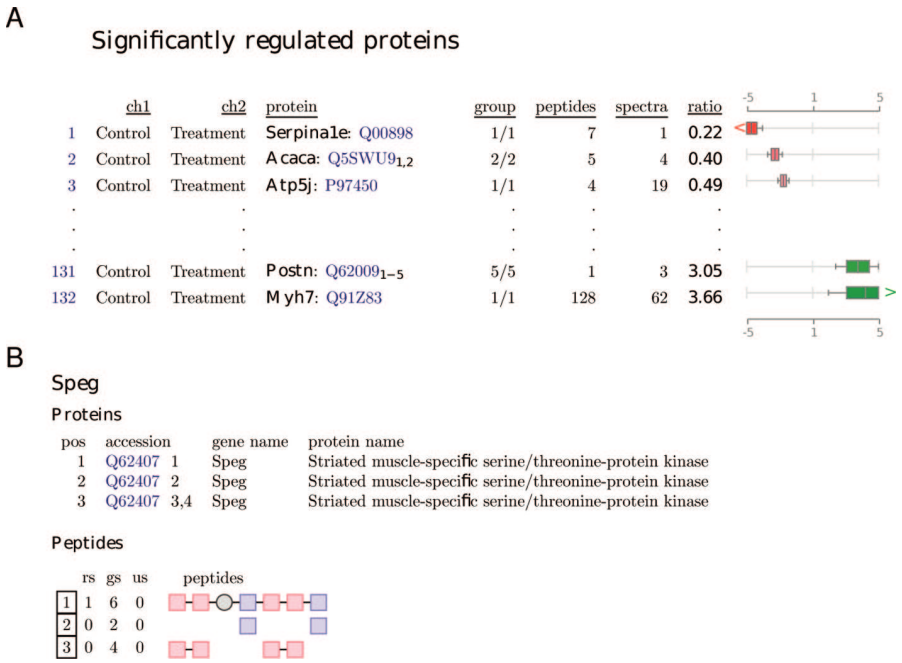


Figure 7. Sample of isobar PDF user report. Regulated proteins are listed first (A), followed by a complete list of all the proteins found in the sample (not shown) and complemented by a protein group structure representation (B), which color-codes subgroups that can potentially be quantified. The PDF document contains links to facilitate navigation as well as hyperlinks to a protein database Web site.

replicate layer does nothing and can hence be elegantly regarded as an extension of the "no replicate" model.

As an alternative to isobar, ANOVA analysis, applied to log intensities of iTRAQ channels, was proposed to treat repeated experiments[15,40] but it obviously covers the single experiment with replicate design. Compared to the layered model presented here, ANOVA also provides selection guidance via *P*-values but, however, is likely to suffer from data heteroscedasticity, a well-known ANOVA limitation. Furthermore, differences in protein ratio accuracy due to variable number of spectra cannot be considered in the ANOVA framework.

Applications

In the original analysis if the TAC data set,[21] we determined the protein relative abundance ratios using a simple average (geometric average in the linear scale) for the pairs of channels: I_{116} vs I_{114} and I_{117} vs I_{114}. Specific peptides were considered only, we then required that the two ratios did not diverge by more than 10% and used I_{115} vs I_{114} as a visual control. The original analysis hence used the replicates as controls but did not integrate them into a single computation and ratios were obtained ignoring signal intensities. Reprocessing the data with isobar almost doubled the number of significantly regulated proteins at a comparable FP rate, i.e., 5%, see Figure 4. The CSF data set was analyzed on two MS platforms, and we show that although the noise models for the LTQ Orbitrap and MALDI TOF/TOF are quite different, isobar determines very similar biological sample variability at the level of protein ratios (Figure 2B), which is a desirable property. Isobar also increased the number of CSF selected proteins compared to the original analysis (data not shown).

Unraveling the Structure of Protein Groups

It is well-known, on the basis of specific peptides, that MS analysis only reveals the presence of so-called *protein groups*, defined as sets of proteins identified by the same set of peptides.[6] The protein that contains all the peptides is the group reporter (there can be several group reporters) and if it has at least one specific peptide then its presence in the sample is certain (we ignore FP identification problems here). In general, the actual expression of the other proteins in the group is impossible to determine.

When quantitative information is provided, there is a potential to elucidate the structure of part of the protein groups.[41,42] The fundamental concept is to exploit differences in ratios from shared peptides compared to those specific to the group reporter, Figure 5A. In a seminal paper, Dost et al.[22] introduced regression methods to try to estimate the relative expression ratios of protein group members. The difficulty is that this problem is *ill-posed*, which in other words means that there is no unique solution. The reason is simple: when shared peptide ratios are different, the information concerning the relative abundance of the two proteins is missing.

We introduce here a novel approach that solves a restricted problem but with reliable results. Given a group reporter ratio, we only try to predict whether group member ratios are significantly larger or smaller by isobar statistical models. On the basis of such predictions, we can identify in some cases group members that influence the ratios of peptides shared with the group reporter distinctly and thus conclude that the group members are present in the sample. Note that the shared peptides are not included in the group reporter ratio calculation since they are not specific.

To make an example, let us assume that proteins p_1, p_2, p_3, p_4 are identified by the same set of peptides and p_1 is the group reporter; see Figure 5A. Subgroups can be identified considering the shared peptides (gray bars, right-hand side of Figure 5A). When peptides are specific to a subgroup and together yield a ratio different from the group reporter peptide ratio, there is a potential to assess expression of the subgroup. For instance, in Figure 5A, the reporter protein will have a ratio 2 and the

dx.doi.org/10.1021/pr1012784 |*J. Proteome Res.* 2011, 10, 2758–2766

peptides specific to the subgroup made of p_3 and p_4 will together give a ratio 1.5, thus indicating they change the expression of the peptides they share with the reporter. Referring to Figure 5B, we can compare the shared peptides common ratio, computed as a protein ratio with estimated variance, with the group reporter ratio and require that x% CIs do not overlap. A related test was used by Hundertmark et al.[17] to predict modified regulated peptides.

Applying this test to TS1 1:1 protein groups with $x = 80\%$, we estimate a FP rate less than 1.5% (50 resamplings). The TP rate depends on relative abundance of reporters and subgroups and differences in ratios; see Supporting Information for some estimates. TAC data set contains 118 protein groups and 10 cases where a subgroup is expressed can be detected; see Figure 6 and Supporting Information. Beyond the technical challenge there is potential scientific value in such additional findings since, for instance, histone modification seems to play a role in cardiac hypertrophy[43] and the ability to distinguish H3.3 from H3.1 and H3.2 might be advantageous for cardiac studies.[44] In Supporting Information we briefly discuss three other cases where association with cardiac pathologies exists and isoforms have relevant distinct roles (Fh, FHOD3 and Camk2d).

In addition to expression prediction, it is possible to detect *differential* expression in peptides shared with subgroups. For instance, a positive significant ratio for the reporter combined with a significantly larger subgroup positive ratio implies significant regulation of the subgroup; see Figure 5C for all of the combinations.

Isobar Output Formats

Beyond its functionality enabling data analysis from within the R environment, isobar has two user report formats. These PDF or spreadsheet documents can be produced either from R commands or through a scripting procedure that hides most R programming details. Figure 7 illustrates parts of the PDF output, and complete examples are provided as Supporting Information.

CONCLUSIONS

High-throughput quantitative technology in proteomics certainly constitutes a clear revolution empowering researchers.[3−5] To efficiently exploit such large and complex data sets is thus very important, and in this context, we believe that the application of statistically sound methods have the potential to pertinently summarize and present data. We therefore developed and implemented an approach that precisely models technical and biological sources of variability and obtained high sensitivity maintaining selectivity at low protein concentration. The proposed approach also naturally extends to experimental designs including biological or technical replicates in multiple iTRAQ/TMT channels.

Comparison with classical solutions such as average ratio estimations (possibly trimmed) and fold change analysis[21] or a t test[39] shows that only isobar can control false positives while maintaining high sensitivity whatever the number of available spectra. In particular and as a consequence of the statistical modeling approach, the false positive rate can be specified *a priori*. More advanced methods published recently do not provide P-values to select proteins and cannot integrate replicates[19] or might suffer more from data heteroscedasticity.[15,40]

Furthermore, the results we obtained show that modeling biological sample variability is advantageous to select regulated proteins. Therefore, we strongly recommend that iTRAQ or TMT experiments are conducted such that a preliminary step to measure the random protein ratio distribution is performed or replicates are integrated. Isobar supports both cases appropriately and can identify truly relevant protein ratios, i.e., a few or many depending on actual sample differences, instead of relying on an empirical selection of the x% most extreme ratios, which is commonly done in absence of a model. The application of isobar to two biological data sets illustrated its potential convincingly. The sensitivity of a classical fold change analysis was doubled while maintaining a similar false positive rate (TAC sample[21]). Additionally, isobar robustly estimated biological sample variability when the same sample is analyzed on alternative MS platforms (CSF sample,[20] LTQ Orbitrap versus MALDI-TOF/TOF).

As one further benefit obtained through the application of a pertinent statistical framework, isobar can analyze the structure of some protein groups, when enough spectra are available, with the ability to predict expression, or even significant differential expression, of group members such as isoforms. Results obtained on TAC data indicates that additional biological insight might be gained via this procedure.

To conclude, isobar is available as an open source R package able to process Mascot, Phenyx, or PSI mzIdentML files, and its utilization is possible with very limited programming skills via an automatic procedure.

ASSOCIATED CONTENT

Supporting Information

The R package named isobar (submitted to Bioconductor) is made available under the LGPL license from our Web site (http://bioinformatics.cemm.oeaw.ac.at) that comprises test data sets 1 and 2, the Perl parsers, and R code necessary to process mzIdentML, Mascot, and Phenyx output files. We also provide one detailed example R code showing analysis of data and sample output files (PDF and spreadsheet formats). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jcolinge@cemm.oeaw.ac.at.

ACKNOWLEDGMENT

Part of this work was supported by the Austrian Ministry of Sciences GEN-AU grants APP-III (KLB) and BIN-III (JC).

REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207.

(2) Domon, B.; Aebersold, R Mass spectrometry and protein analysis. *Science* **2006**, *312*, 212–7.

(3) Pan, S.; Aebersold, R.; Chen, R.; Rush, J.; Goodlett, D. R.; McIntosh, M. W.; Zhang, J.; Brentnall, T. A. Mass spectrometry based targeted protein quantification: methods and applications. *J. Proteome Res.* **2009**, *8*, 787–97.

(4) Patel, V. J.; Thalassinos, K.; Slade, S. E.; Connolly, J. B.; Crombie, A.; Murrell, J. C.; Scrivens, J. H. A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J. Proteome Res.* **2009**, *8*, 3752–9.

(5) Julka, S.; Regnier, F. Quantification in proteomics through stable isotope coding: a review. *J. Proteome Res.* **2004**, *3*, 350–63.
(6) Colinge, J.; Bennett, K. L. Introduction to computational proteomics. *PLoS Comput. Biol.* **2007**, *3*, e114.
(7) Schmidt, A.; Kellermann, J.; Lottspeich, F. A novel strategy for quantitative proteomics using isotope-coded protein labels. *Proteomics* **2005**, *5*, 4–15.
(8) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999**, *17*, 994–9.
(9) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376–86.
(10) Gevaert, K.; Goethals, M.; Martens, L.; Van Damme, J.; Staes, A.; Thomas, G. R.; Vandekerckhove, J. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **2003**, *21*, 566–9.
(11) Wang, Y. K.; Ma, Z.; Quinn, D. F.; Fu, E. W. Inverse 18O labeling mass spectrometry for the rapid identification of marker/target proteins. *Anal. Chem.* **2001**, *73*, 3742–50.
(12) Ross, P. L.;Multiplexed protein quantitation in Saccharomyces cerevisiae using aminereactive isobaric tagging reagents. *Mol. Cell. Proteomics* **2004**, *3*, 1154–69.
(13) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **2003**, *75*, 1895–904.
(14) Keshamouni, V. G.; Michailidis, G.; Grasso, C. S.; Anthwal, S.; Strahler, J. R.; Walker, A.; Arenberg, D. A.; Reddy, R. C.; Akulapalli, S.; Thannickal, V. J.; Standiford, T. J.; Andrews, P. C.; Omenn, G. S. Differential protein expression profiling by iTRAQ-2DLCMS/ MS of lung cancer cells undergoing epithelial-mesenchymal transition reveals a migratory/ invasive phenotype. *J. Proteome Res.* **2006**, *5*, 1143–54.
(15) Oberg, A. L.; Mahoney, D. W.; Eckel-Passow, J. E.; Malone, C. J.; Wolfinger, R. D.; Hill, E. G.; Cooper, L. T.; Onuma, O. K.; Spiro, C.; Therneau, T. M.; Bergen, r., H. R. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.* **2008**, *7*, 225–33.
(16) Wang, P.; Tang, H.; Zhang, H.; Whiteaker, J.; Paulovich, A. G.; Mcintosh, M. Normalization regarding non-random missing values in high-throughputmass spectrometry data. *Pac. Symp. Biocomput.* **2006**, 315–26.
(17) Hundertmark, C.; Fischer, R.; Reinl, T.; May, S.; Klawonn, F.; Jansch, L. MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics. *Bioinformatics* **2009**, *25*, 1004–11.
(18) Zhang, Y.; Askenazi, M.; Jiang, J.; Luckey, C. J.; Griffin, J. D.; Marto, J. A. A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia. *Mol. Cell Proteomics* **2010**, *9*, 780–790.
(19) Karp, N. A.; Huber, W.; Sadowski, P. G.; Charles, P. D.; Hester, S. V.; Lilley, K. S. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell Proteomics* **2010**, *9*, 1885–1897.
(20) Tiberti, N.; Hainard, A.; Lejon, V.; Robin, X.; Ngoyi, D. M.; Turck, N.; Matovu, E.; Enyaru, J.; Ndung'u, J. M.; Scherl, A.; Dayon, L.; Sanchez, J. C. Discovery and verification of osteopontin and Beta-2-microglobulin as promising markers for staging human African trypanosomiasis. *Mol. Cell Proteomics* **2010**, *9*, 2783–2795.
(21) Koecher, T.; Pichler, P.; Schutzbier, M.; Stingl, C.; Kaul, A.; Teucher, N.; Hasenfuss, G.; Penninger, J. M.; Mechtler, K. High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all. *J. Proteome Res.* **2009**, *8*, 4743–4752.
(22) Dost, B.; Bandeira, N.; Li, X.; Shen, Z.; Brigg, S.; Bafna, V. Shared peptides in mass spectrometry based protein quantitation. In *RECOMB Conference Proceedings*; Springer: New York, 2009.
(23) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68*, 850–8.
(24) Wu, C. H.;The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **2006**, *34*, D187–91.
(25) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551–3567.
(26) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: towards highthroughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3*, 1454–63.
(27) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **2004**, *22*, 214–9.
(28) Eisenacher, M. mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. *Methods Mol. Biol.* **2011**, *696*, 161–177.
(29) Boehm, A. M.; Putz, S.; Altenhofer, D.; Sickmann, A.; Falk, M. Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinf.* **2007**, *8*, 214.
(30) Ow, S. Y.; Salim, M.; Noirel, J.; Evans, C.; Rehman, I.; Wright, P. C. iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly. *J. Proteome Res.* **2009**, *8*, 5347–55.
(31) Shadforth, I. P.; Dunkley, T. P.; Lilley, K. S.; Bessant, C. i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics* **2005**, *6*, 145.
(32) Jin, S.; Daly, D. S.; Springer, D. L.; Miller, J. H. The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. *J. Proteome Res.* **2008**, *7*, 164–9.
(33) Usaite, R.; Wohlschlegel, J.; Venable, J. D.; Park, S. K.; Nielsen, J.; Olsson, L.; Yates Iii, J. R. Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. *J. Proteome Res.* **2008**, *7*, 266–75.
(34) Lin, W. T.; Hung, W. N.; Yian, Y. H.; Wu, K. P.; Han, C. L.; Chen, Y. R.; Chen, Y. J.; Sung, T. Y.; Hsu, W. L. Multi-Q: a fully automated tool for multiplexed protein quantitation. *J. Proteome Res.* **2006**, *5*, 2328–38.
(35) Bantscheff, M.; Boesche, M.; Eberhard, D.; Matthieson, T.; Sweetman, G.; Kuster, B. Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **2008**, *7*, 1702–13.
(36) Weng, L.; Dai, H.; Zhan, Y.; He, Y.; Stepaniants, S. B.; Bassett, D. E. Rosetta error model for gene expression analysis. *Bioinformatics* **2006**, *22*, 1111–1121.
(37) Evans, M.; Hastings, N.; Peacock, B. *Statistical Distributions*; Wiley-Interscience: New York, 2000.
(38) Levy, P. *Calcul des Probabilités*; Gauthier-Villars,: Paris, 1925.
(39) Rodriguez-Suarez, E.; Gubb, E.; Alzueta, I. F.; Falcon-Perez, J. M.; Amorim, A.; Elortza, F.; Matthiesen, R. Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage. *Proteomics* **2010**, *10*, 1545–1556.
(40) Schwacke, J. H.; Hill, E. G.; Krug, E. L.; Comte-Walters, S.; Schey, K. L. iQuantitator: a tool for protein expression inference using iTRAQ. *BMC Bioinf.* **2009**, *10*, 342.
(41) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4*, 1419–40.
(42) Choe, L.; D'Ascenzo, M.; Relkin, N. R.; Pappin, D.; Ross, P.; Williamson, B.; Guertin, S.; Pribil, P.; Lee, K. H. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* **2007**, *7*, 3651–3660.
(43) Kong, Y.; Tannous, P.; Lu, G.; Berenji, K.; Rothermel, B. A.; Olson, E. N.; Hill, J. A. Suppression of class I and II histone deacetylases blunts pressure-overload cardiac hypertrophy. *Circulation* **2006**, *113*, 2579–2588.
(44) Goldberg, A. D.;Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* **2010**, *140*, 678–691.

4.1.3. Supporting information

Noise model

We denote $P = \{p_1; \ldots; p_n\}$ the n proteins identified in one experiment and $S_i = \{s_{i,1}; \ldots; s_{i,m_i}\}$ the m_i spectra identifying protein i. Reporter intensities are written $\{I_{114,i,j}; I_{115,i,j}; I_{116,i,j}; I_{117,i,j}\}$ or $\{I_{114}; I_{115}; I_{116}; I_{117}\}$ when the context is clear. We also write S the set of all the spectra s.

From the $1:1$ spectra we can learn a so called *noise model* or *variance function* to capture noise typical magnitude in function of signal intensity. We re-introduce this method in a slightly different and more direct way compared to Hundertmark et al. (2009). Considering 2 channels only, say $\{I_{114}; I_{115}\}$, we write the log-ratio of a spectrum $C = \log(I_{115}/I_{114}) = X_2 - X_1$, where $X_1 = \log(I_{114})$ and $X_2 = log(I_{115})$. Denoting $\mathrm{E}(X_i) = \mu_i$ the true signal intensities (without noise), we have

$$C = X_2 - X_1 = \mu_2 - \mu_1 + \varepsilon_2 - \varepsilon_1, \tag{S1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ represents the noise. In a $1:1$ ratio, $\mu_1 = \mu_2$ and, assuming independence of the measurements X_1 and X_2, we find $\mathrm{Var}(C) = 2 \times \mathrm{Var}(\epsilon)$. From a $1:1$ experiment, we can thus learn signal noise ϵ variance directly from observed ratio variance: $\mathrm{Var}(\epsilon) = \mathrm{Var}(C)/2$ and a common choice to model signal intensity dependence is

$$\mathrm{Var}(\epsilon(x)) = f(x) = a + re^{-\lambda x}, \tag{S2}$$

with x the channel log-intensity and a, r, and λ parameters of the noise model.

To train the parameters a, r, and λ, either a genetic algorithm (Hundertmark et al., 2009) or a modified Levenberg-Marquardt iteration have been used (Yi Zhang et al., 2010), both tested on rather small data sets. Following a straightforward maximum likelihood approach, we write

$$c(s_{i,j}) = \log(I_{115,i,j}/I_{114,i,j})$$

and, using the noise distribution $N(0, f(x))$, we obtain a likelihood function

$$L(a, r, \lambda) = \prod_{i \in \{1;\ldots;n\}} \prod_{j \in \{1;\ldots;m_i\}} \frac{e^{-\frac{s_{i,j}^2}{2\sigma_{i,j}^2}}}{\sqrt{2\pi}\sigma_{i,j}},$$

where $\sigma_{i,j}^2$ is the noise variance estimated by the noise model:

$$\sigma_{i,j}^2 = f\left(\frac{\log(I_{115,i,j}) + \log(I_{114,i,j})}{2}\right).$$

Now maximizing $L(a, r, \lambda)$, or more conveniently $\log(L(a, r, \lambda))$, we find the parameter estimates $(\hat{a}, \hat{r}, \hat{\lambda})$. In isobar, this is achieved using the R function `nlminb` and we could easily process data sets comprising > 100000 spectra within a few seconds.

Several noise models $f_i(x) = a_i + r_i e^{-\lambda_i x}, i \in 1; \ldots; n$ are obtained from distinct pairs of iTRAQ or TMT channels. They can be averaged to obtain a more robust common estimate $f(x)$. Writing

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

and imposing that $f(x)$ retains the same form, i.e. $f(x) = a + re^{-\lambda x}$, we find

$$a = \frac{1}{n} \sum_{i=1}^{n} a_i,$$

by letting $x \to \infty$, and then

$$r = \frac{1}{n} \sum_{i=1}^{n} r_i,$$

by setting $x = 0$, and finally

$$\lambda = -\log\left(\frac{1}{n} \sum_{i=1}^{n} r_i e^{-\lambda_i}\right) + \log(r),$$

by setting $x = 1$.

We further extend the applicability of noise models by considering the situations where a $1:1$ experiment is not available. In this case, spectrum ratios do not necessarily equal 0 (we use the log-scale) but, provided enough spectra are available to estimate a protein ratio reliably, we can subtract it from all the spectrum ratios of the protein and observe their variability around 0 again. Pooling such corrected spectrum ratios for all proteins we obtain enough data to fit the noise model. Namely, let $n_{minspectra}$ be the minimum number of spectra required to estimate the protein ratio (for instance 7 in small data sets and more to limit computation time in large data sets), and

$$A_{\text{minspectra}} = \{i \in \{1; \ldots; n\}; m_i \geq n_{\text{minspectra}}\}$$

the subset of proteins having enough spectra. We obtain a set of corrected spectrum ratios

$$R = \{\log(I_{115,i,j}/I_{114,i,j}) - c(p_i); i \in A_{\text{minspectra}}, j \in \{1; \ldots; m_i\}\}.$$

The set R can then be used as a substitute for $1:1$ spectra. Figure S1 shows a comparison of noise models trained on $1:1$ and non $1:1$ data.

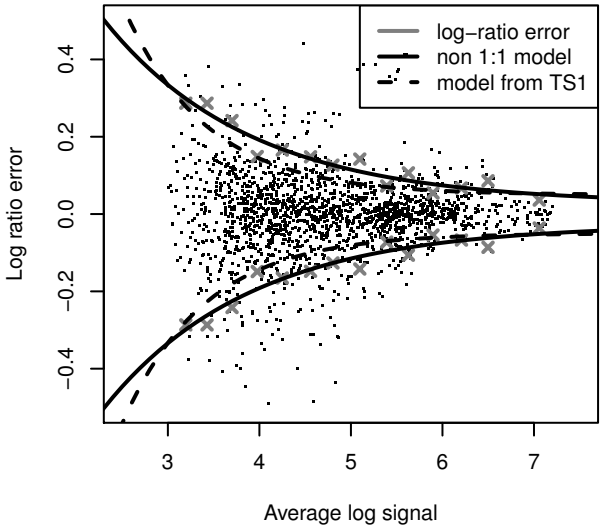


Figure S1: Comparison of noise model fit on the basis of $1:1$ data versus non $1:1$ data. We used a non $1:1$ data set that was measured on a LTQ Orbitrap instrument (ThermoFinnigan) for which we have a 1:1 sample available (TS1). We observe from the figure the similarity between the 2 noise models. The dots represent the corrected spectrum ratios R, or ratio errors, and we see that they distribute the same as direct ratio errors of a $1:1$ data set, e. g. compare with Figure 1 in the paper.

Protein ratio

We first illustrate various affects observed at a peptide level that motivated our method of selecting specific peptides (fig. S2), applying outlier elimination (fig. S3) and ignoring peptide dependence (fig. S4).

The variance of weighted sum estimators $x = \sum_j \beta_j x_j$ can be computed directly (measures assumed independent) by

$$\mathrm{Var}(x) = \frac{1}{\sum_j \beta_j}. \tag{S3}$$

In the case of the protein ratio, its estimator variance $V_{estim,i}$ is calculated setting $x = c(p_i)$, $x_j = c(s_{i,j})$, and

$$\beta_j = \frac{1}{\mathrm{Var}(c(s_{i,j}))} = \frac{1}{f(\log(I_{114,i,j})) + f(\log(I_{115,i,j}))}, \tag{S4}$$

with $f()$ the noise model.

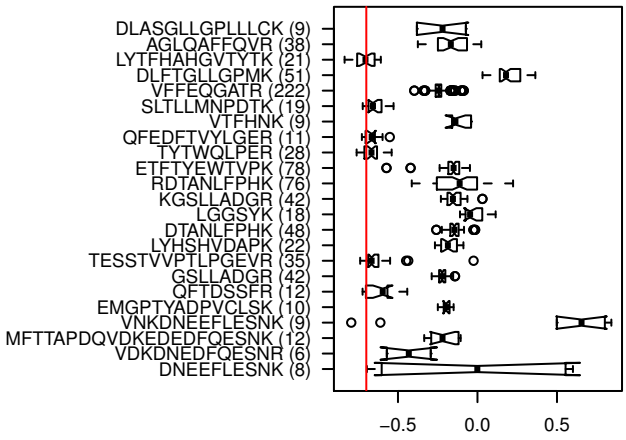


Figure S2: In TS1 data we consider mouse CERU peptides channel 116 versus 114 with a theoretical ratio of $\log(0.2)$ (red line). Each peptide is represented by a boxplot to indicate the range of its spectrum ratios. All the peptides not centered around the red lines are shared with rat and human CERUs, but SGAGREDSACLPWAYYSTVDR, RAEDEHLGLLGPPLHANVGDK, VNKDNEEFLESNK, which are outliers (see fig. S3).

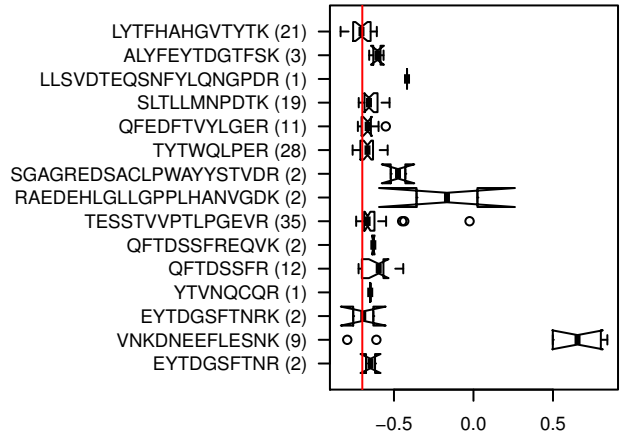


Figure S3: In TS1 data we again consider mouse CERU peptides channel 116 versus 114 and observe that even specific peptide can yield a few spectra with a ratio far from the correct value (peptides SGAGREDSACLPWAYYSTVDR, RAEDEHLGLLGPPLHANVGDK, AND VNKDNEEFLESNK in this example). This is most likely due to co-eluting material and requires detection and elimination of outlier ratios before computing the protein ratio.

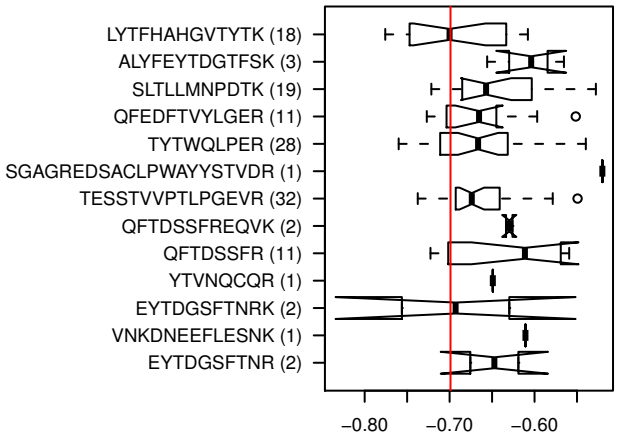


Figure S4: The spectrum ratios of specific mouse CERU peptides after outlier eliminations are represented as boxplots (TS1 data, channels 116 versus 114). No significant peptide dependence is observed.

The computation of the sample variance of a sample $\{x_j\}$ with weights $\{\beta_j\}$ is given by

$$\frac{\sum_j \beta_j}{(\sum_j \beta_j)^2 - \sum_j \beta_j^2} \sum_j \beta_j (x_j - x)^2 \tag{S5}$$

The protein ratio sample variance $V_{\mathrm{spectrum},i}$ is calculated setting $x = c(p_i)$, $x_j = c(s_{i,j})$, and β_j as in Eq. (S4). Moreover, sample variance estimation is unreliable with 2 spectra only and impossible with 1. In these cases we use the following heuristics $V'_{\mathrm{spectrum},i}$:

- 1 spectrum available: $V'_{\mathrm{spectrum},i} = (V_{\mathrm{estim},i})^{0.75}$;
- 2 spectra available: $V'_{\mathrm{spectrum},i} = \max\left\{V_{\mathrm{spectrum},i}; (V_{\mathrm{estim},i})^{0.75}\right\}$.

Biological sample variability modeling

The random protein ratio distribution is accurately modeled by a Cauchy distribution. We show various MS platform data sets with a normal and a Cauchy models for comparision. The Cauchy model cannot be distinguished from the data by a Kolmogorov-Smirnov test, whereas the normal model is significantly different. For instance, in fig. S5 panel B, Cauchy distribution P-value=0.15 and the normal distribution P-value<1.7E-13.

We then illustrate the application of the mixture Normal-Cauchy distribution to estimate the random protein ratio distribution on the CSF (MALDI-TOF/TOF) data set (see fig. S6). Comparing the first two patient pools (one of each class) we obtain protein rations combining regulated and non-regulated proteins.

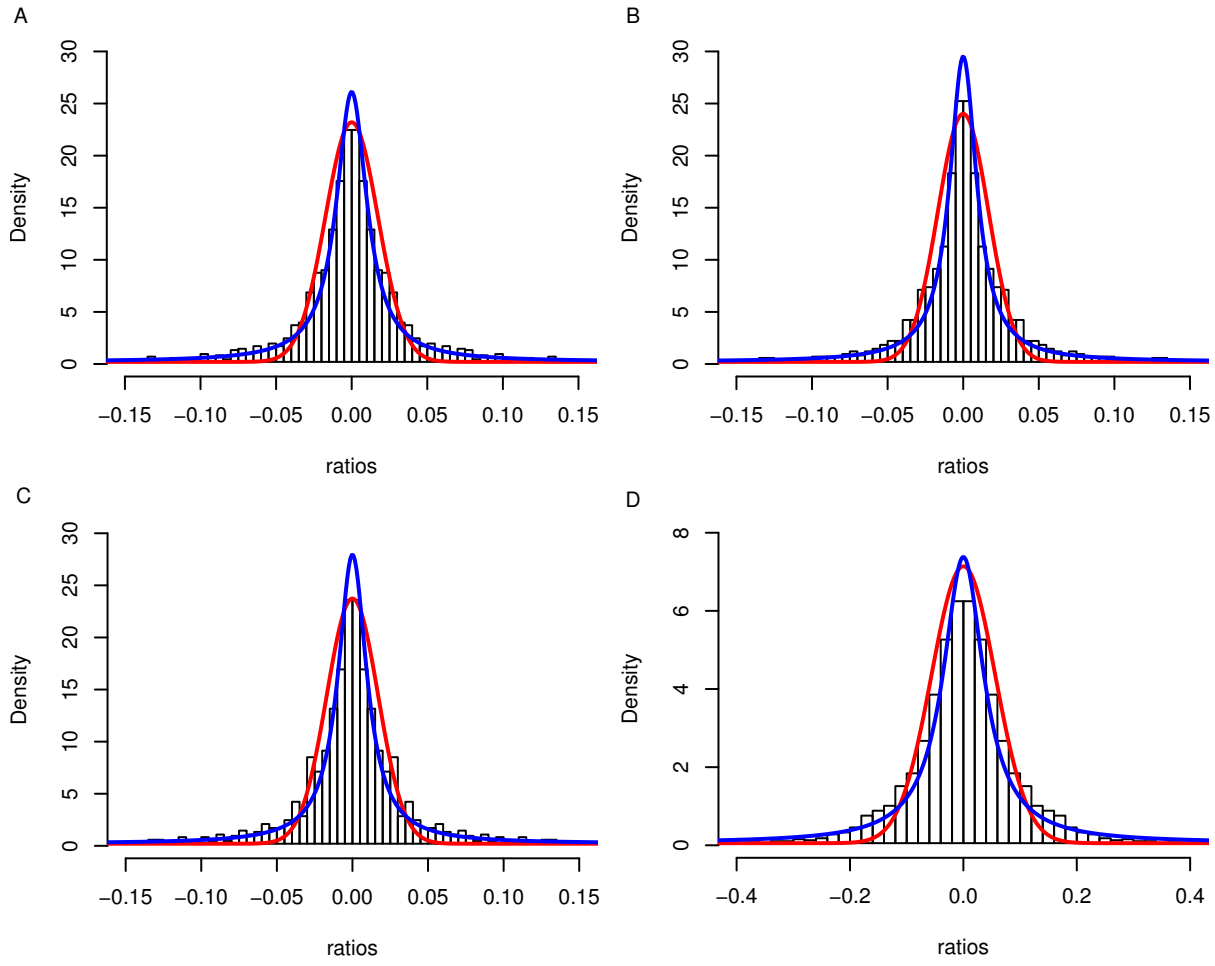


Figure S5: Cauchy versus normal distributions to fit random protein ratios. Cauchy (blue) captures accurately the spread distribution with a sharp peak centered around 0, whereas the normal curve (red) drops too rapidly and gives an overoptimistic model of the ratios. A-C are ratios from TS1 (LTQ Orbitrap HCD) and D are ratios from a MALDI data set (Choe et al., 2007), all submitted to specific peptide and outlier filtering. A. Protein ratios estimated by an average. B. Weighted average as recommended (paper Eq. (2)). C. Multi-Q (Lin et al., 2006) estimator. D. Weighted average, MALDI data set.

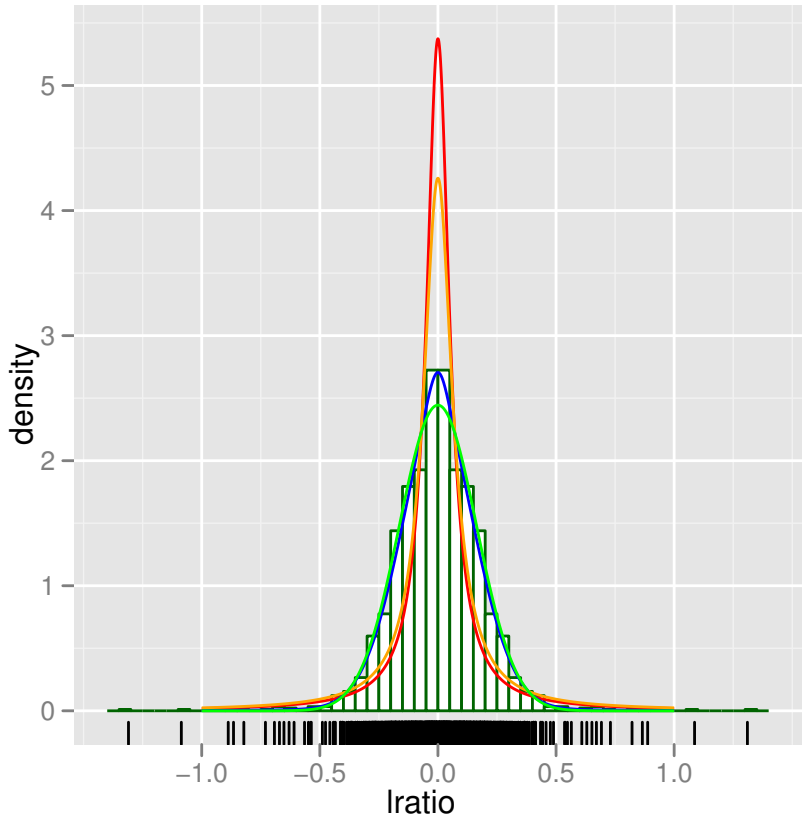


Figure S6: Application of the Normal-Cauchy mixture to estimate the Cauchy component (orange) modeling non-regulated protein ratios. We note the good match with the true model (red) obtained comparing samples of the same class. The mixture model is in blue and the Normal component in green.

Replicates

Given a protein p_i, for each pair of sample $(a_k; b_l)$ we have a protein log-ratio $c_{k,l}(p_i)$ with associated variance estimation $v_{k,l,i}$. Application of weighted averages gives

$$c(p_i) = \sum_{k,l} \alpha_{k,l} c_{k,l}(p_i), \text{ with } \alpha_{k,l} \propto \frac{1}{v_{k,k,i}} \tag{S6}$$

and, naturally, $\sum_{k,l} \alpha_{k,l} = 1$. Variance of $c(p_i)$ is obtained from eq. (S3) with $\beta_j = v_{k,l,i}$.

To determine the null-distribution of the product of biological sample P-values $q_i = \prod_{k,l} q_{k,l,i}$ we proceed as follows. Under the null hypothesis, each P-value $q_{k,l,i}$ is uniformly distributed over the interval $[0; 1]$. Therefore, we have to determine the distribution of the product of N such random variables:

Proposition S1 1 *Let $Q_1 \sim U(0; 1), i = 1, 2, \ldots, N$, be N uniformly distributed random variables. Their product $Q = \prod_{i=1}^{N} Q_i$ follows a distribution given by the probability density function*

$$r_N(q) = \frac{\ln^{N-1} q}{(N-1)!}$$

and the cumulative distribution function

$$R_N(q) = q \sum_{i=0}^{N-1} (-1)^i \frac{\ln^i q}{i!}.$$

Proof. We prove the result by induction on N, starting with $N = 2$.

$$\begin{aligned} R_2(q) &= \Pr(Q_1 Q_2 < q) \\ &= \Pr(Q_1 Q_2 < q \mid Q_1 Q_2 > 0) \underbrace{\Pr(Q_1 Q_2 > 0)}_{=1} + \Pr(Q_1 Q_2 < q \mid Q_1 Q_2 = 0) \underbrace{\Pr(Q_1 Q_2 = 0)}_{=0} \\ &= \Pr(Q_1 Q_2 < q \mid Q_1 Q_2 > 0) \end{aligned}$$

For notation convenience we define the set

$$T_N(q) = \left\{ x \in]0; 1]^N; \prod_{i=1}^{N} x_i < q \right\}.$$

Now, we find

$$\begin{aligned} R_2(q) &= \iint_{r_2(q)} \mathrm{d}(q_1, q_2) \\ &= \int_{q_1=0}^{1} \int_{q_2=0}^{\min\left(\frac{q}{q_1},1\right)} \mathrm{d}q_2 \; \mathrm{d}q_1 = \int_{q_1=0}^{1} \min\left(\frac{q}{q_1}, 1\right) \mathrm{d}q_1 \\ &= \int_{q_1=0}^{q} \mathrm{d}q_1 + q \int_{q_1-q}^{1} \frac{1}{q_1} \; \mathrm{d}q_1 = q(1 - \ln q). \end{aligned}$$

For $N > 2$ we have

$$\begin{aligned} R_N(q) &= \int \cdots \int_{T_N(q)} \mathrm{d}(q_1, \cdots, q_N) \\ &= \int_{q_1=0}^{1} \left[\int \cdots \int_{T_{N-1}\left(\min\left(\frac{q}{q_1},1\right)\right)} \right] \mathrm{d}q_1 = \int_{q_1=0}^{1} R_{N-1}\left(\min\left(\frac{q}{q_1}, 1\right)\right) \mathrm{d}q_1 \\ &= \int_{q_1=0}^{q} R_{N-1}(1) \; \mathrm{d}q_1 + \int_{q_1=q}^{1} R_{N-1}\left(\frac{q}{q_1} \; \mathrm{d}q_1\right) \\ &= \int_{q_1=0}^{q} \mathrm{d}q_1 + \int_{q_1=q}^{1} \left(\frac{q}{q_1} \sum_{i=0}^{N-2} (-1)^i \frac{\ln^i\left(\frac{q}{q_1}\right)}{i!} \right) \mathrm{d}q_1 \\ &= q + \sum_{i=0}^{N-2} \frac{(-1)^i}{i!} \int_{q_1=q}^{1} \frac{q}{q_1} \ln^i \frac{q}{q_1} \; \mathrm{d}q_1 \end{aligned}$$

Using the primitive $\int \frac{a}{x} \ln^n(\frac{a}{x}) \; \mathrm{d}x = -\frac{a}{n+1} \ln^{n+1}(\frac{a}{x})$, we further simplify the integral term

$$\begin{aligned} R_N(q) &= q + \sum_{i=0}^{N-2} \frac{(-1)^i}{i!} \left[(-q) \frac{\ln^{i+1} \frac{q}{q_1}}{i+1} \right]_q^1 \\ &= q + q \sum_{i=0}^{N-2} \frac{(-1)^{i+1}}{(i+1)!} \ln^{i+1} q \\ &= q \left(1 + \sum_{i=1}^{N-1} \frac{(-1)^i}{i!} \ln^i q \right) \\ &= q \left(\sum_{i=1}^{N-1} \frac{(-1)^i}{i!} \ln^i q \right) \end{aligned}$$

When biological samples contain enough variability and some pairs $(a_k; b_l)$ yield a protein ratio with a sign (log-scale) that is opposite compare to the majority for a protein p_i, then it makes sense that such ratios should penalize the final P-value. To set the maximum number of accepted sign discrepancies is a parameter of the software and here we only explain how the model adapts naturally. If the majority of ratios $c_{k,l}(p_i)$ is positive, the P-values $q_{k,l,i}$ are obtained from

the relevant Cauchy distribution by $1 - F_{\text{Cauchy}}(c_{k,l}(p_i))$. On the other hand, if the majority is negative, then the P-values are obtained by $F_{\text{Cauchy}}(c_{k,l}(p_i))$.

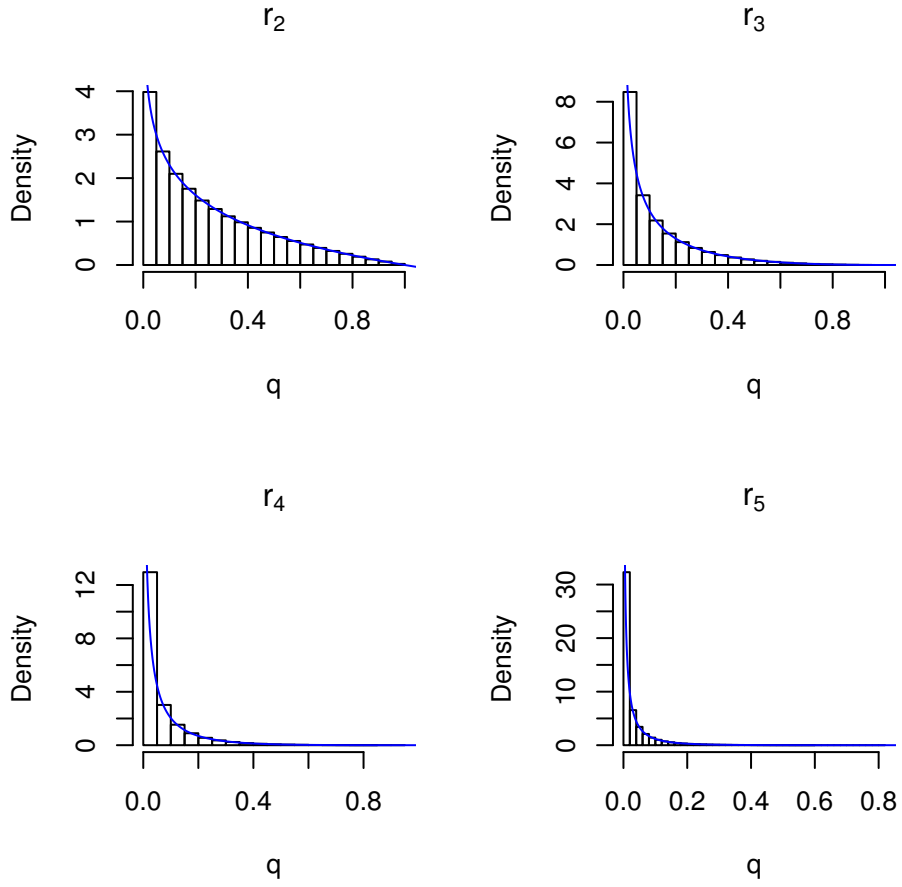


Figure S7: Illustration of the density $r_N(q)$ for $N = 2, 3, 4, 5$. Histograms were obtained by simulations using products of random samples drawn from a uniform distribution and the solid blue line is the theoretical probability density function.

Missing observations can either be dealt with by reducing N appropriately (more permissive) or by assigning them a neutral P-value of 0.5 and maintaining the original N (more stringent). Isobar allows for both choices. The distribution $f_{\text{replicates},N}(q)$ scales with N requiring smaller products q_i for significance when N is larger, fig. S7.

Protein group structure

The application of the method to TS1 $1:1$ proteins, re-sampling 50 times, allows us to precisely estimate the FP rates depending on the number of available spectra, table S1. We can also obtain some estimations of the TP rates but as the latter depend in the relative as well as the absolute concentrations of group reporters and group members, there is no single TP rate. The

values we obtain doing so are indicative only and we can use rat CERU as a group reporter and peptides shared with mouse CERU as a group member protein. We consider two cases: rat $1:2$ and mouse $10:5$ (channels 114 vs. 115), and rat $5:10$ and mouse $2:1$ (channels 116 vs. 117). The first one is much easier to detect as the group member is at higher concentration and hence can influence the shared peptide ratios largely. TP rates are in table S1 as well. Protein

Table S1: False positive rates ($1:1$ TS1 data re-sampled 50 times) and true positive rates for two cases (TS1 rat vs. mouse CERU, re-sampled 500 times).

#sp[a]	FP	TP $114/115$	TP $116/117$
1	0.000	0.59	0.02
2	0.000	0.75	0.02
3	0.012	0.89	0.32
5	0.006	0.98	0.63
10	0.000	1.00	0.63
15	0.000	1.00	0.88

[a]number of available spectra

group structure analysis of the TAC data sets revealed 10 out of 118 protein groups where it was possible to predict expression of a group member. In 4 cases, we could also conclude for the significant differential expression of either the group reporter or the group member, fig. S8.

In the main text we indicate the ability to detect expression of Histone H3.3 independently of H3.1 and H3.2 might be advantageous in cardiac studies. We mention here another three protein groups we could "separate" that might be relevant as well. First, Camk2d is involved in cardiac hypertrophy (W. Zhang et al., 2010) and heart failure (Toko et al., 2010) and its splice variants have special functions in the heart (Xu et al., 2005; Singh et al., 2005). Second, we are able to distinguish the mitochondrial from the cytoplasmic isoform of Formarate hydratase (X.-H. Liu et al., 2004) (the mitochondrial isoforms might be important, as pressure overload by aortic banding is likely to go along with increased energy demand). Third, FHOD3 (FH1/FH2 domain-containing protein 3) is a gene involved in myofibril maintenance (Iskratsch et al., 2010) and isoforms 1 and 4 are cardiac specific, whereas the other isoforms are not specific to the heart.

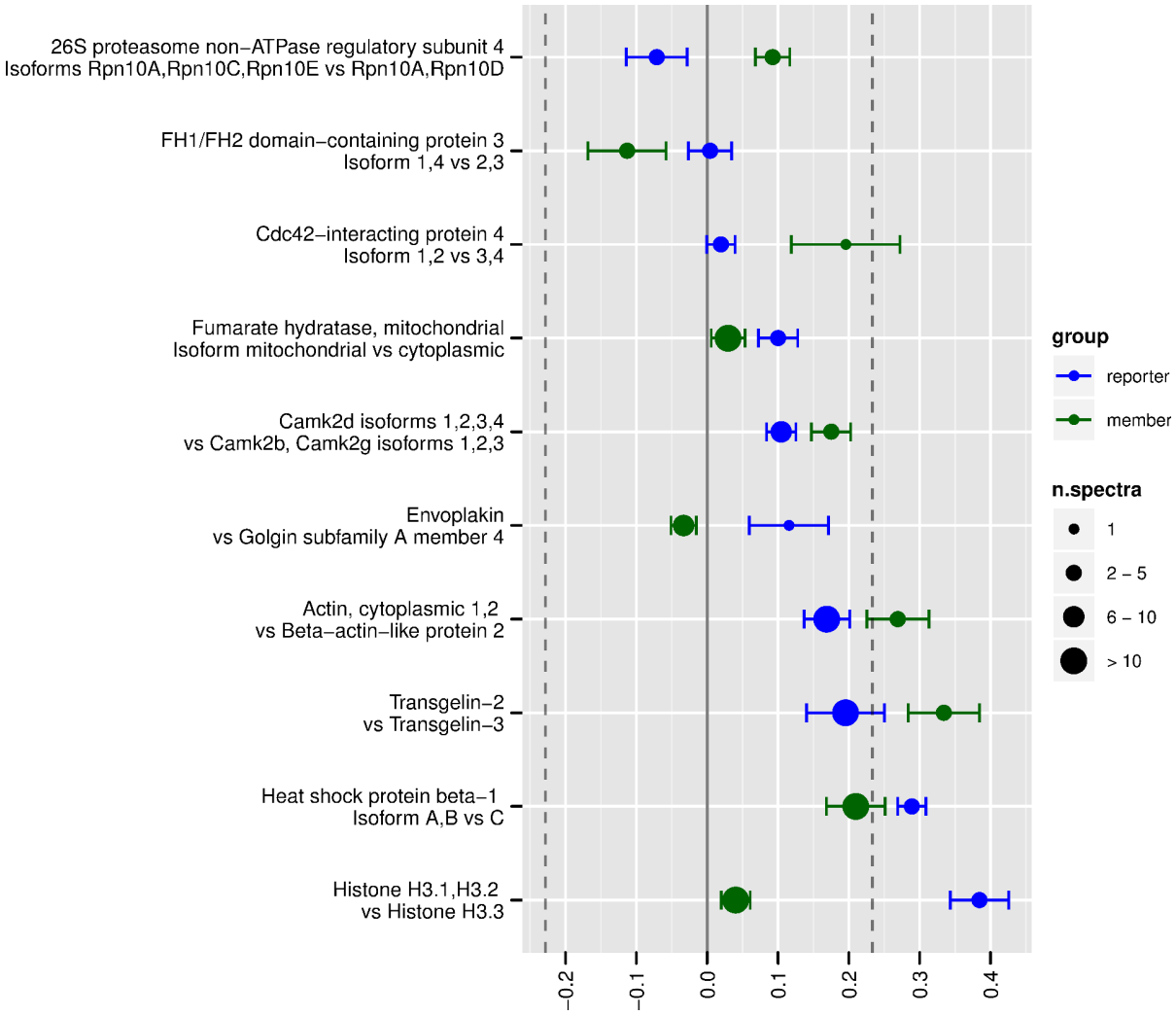


Figure S8: The 10 protein groups from the TAC data set with a predicted group member as expressed. The dashed lines represent significant protein ratios and we see for 4 out of the 10 groups we could analyze, it is possible to even predict differential expression.

4.2. isobar^{PTM}: A software tool for the quantitative analysis of posttranslationally modified proteins

4.2.1. Prologue

The first part of this chapter presented statistical models for protein quantification and their implementation. Beyond protein abundance regulation, PTMs provide another important mechanism of protein control. In many pathways, PTMs are better indicators of the biological status than the protein abundance.

The following publication addresses several of the aspects peculiar to the analysis of quantitative PTM data (see section 2.4). The specific challenges of PTM data include:

1. The reliance on fewer data points for the estimation of ratios, as they are summarized at the level of modified peptides, and thus exhibit higher (technical) variability of the ratio.
2. The uncertain localization of modification groups, as returned by MS search engines such as Mascot, requires additional software for their validation.
3. Observed changes at the level of modified peptides are the product of changes of the modification state *and* the protein abundance. Thus, complementary datasets on protein expression differences can be essential to separate these effects.
4. The function of all the different PTM sites is not as well-known as the function of the proteins. However, the integration of public PTM databases can help to identify novel and known modification sites in the experiment.

We extended the software to integrate these aspects and thus enable easier analysis of PTM datasets. The statistical models were re-evaluated and extended for the peptide level analysis. General improvements on the models include the support of a generalized T-distribution for the capturing of biological variability. To test the whole pipeline, a large scale public dataset on phosphorylation and protein differences in embryonic stem cell and induced pluripotent stem cells (Phanstiel et al., 2011) was re-analyzed.

The supporting information to this publication is in section 4.2.3. The user manual for the PTM functionality in `isobar` is available at `http://www.ms-isobar.org/isobar-ptm`, and as attachment to this thesis.

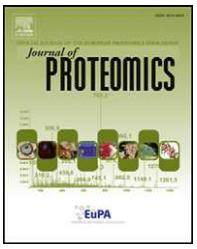
4.2.2. Manuscript

JOURNAL OF PROTEOMICS 90 (2013) 77–84

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Isobar^{PTM}: A software tool for the quantitative analysis of post-translationally modified proteins☆

*Florian P. Breitwieser, Jacques Colinge**

CeMM — Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH-BT. 25.3, 1090 Vienna, Austria

ARTICLE INFO

Available online 5 March 2013

Keywords:
Bioinformatics
Computational proteomics
Quantitative proteomics
iTRAQ
TMT
Statistics

ABSTRACT

The establishment of extremely powerful proteomics platforms able to map thousands of modification sites, e.g. phosphorylations or acetylations, over entire proteomes calls for equally powerful software tools to effectively extract useful and reliable information from such complex datasets. We present a new quantitative PTM analysis platform aimed at processing iTRAQ or Tandem Mass Tags (TMT) labeled peptides. It covers a broad range of needs associated with proper PTM ratio analysis such as PTM localization validation, robust ratio computation and statistical assessment, and navigable user report generation. Isobar^{PTM} is made available as an R Bioconductor package and it can be run from the command line by non R specialists.

Biological significance
"IsobarPTM is a new software tool facilitating the quantitative analysis of protein modification regulation streamlining important issues related to PTM localization and statistical modeling. Users are provided with a navigable spreadsheet report, which also annotate already public modification sites."
This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

1. Introduction

The dynamic execution of the genetic program encoded in the genome is controlled by a multitude of regulatory mechanisms such as transcription factors, alternative splicing, silencing by non coding RNAs, and epigenetic marks. The large repertoire of gene products generated by the translation/transcription machinery is further submitted to another level of modulation provided by PTMs. These modifications increase the diversity of biomolecules available to cells to adapt to environmental changes or to assemble in specialized tissues.

A large number of PTMs have been described (591 entries in the RESID [1] database vers. 70.01) which modify the properties of proteins for diverse purposes and whose deregulated control can cause multiple disorders. A classical and very important example is the phosphorylation of threonine, tyrosine, or serine that is used to activate proteins upon specific stimuli and to realize signaling cascades [2]. Dysfunctions in such signaling can cause cell proliferation and cancer. More generally, PTMs participate in signal integration within the cell, protein degradation, binding, etc. Commonly studied PTMs are catalyzed by enzymes such as kinases, phosphatases, or acetyltransferases. It has been also shown that distinct PTMs can have a cross-talk, e.g. to establish substitution strategies when one is deficient [3].

Given the importance of PTM regulation in a broad range of biological processes, the analysis of their differences across biological samples is of prime interest in proteomics and is best achieved with quantitative techniques. The measure of PTMs by MS is generally challenging [4,5] since most modifications are lost upon ionization or fractionation resulting in low MS signals

☆ This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

* *Corresponding author*. Tel.: +43 14016070020; fax: +43 140160970000.
E-mail address: jcolinge@cemm.oeaw.ac.at (J. Colinge).

1874-3919/$ – see front matter
http://dx.doi.org/10.1016/j.jprot.2013.02.022

and it might be necessary to operate chromatography and MS equipments in particular conditions. A number of analytical protocols – often relying on chromatographic enrichment for the PTM of interest – have been established successfully, e.g. in the case of phosphorylation [6], ubiquitinylation [7], or acetylation [8].

In this work, we present isobar^{PTM} a new software tool aimed at analyzing the MS/MS spectra of modified peptides resulting from isobarically labeled samples using the Tandem Mass Tags [9] (TMT) or iTRAQ [10] reagents. Isobar^{PTM} is a peptide level extension of the isobar statistical and software framework which we introduced for the analysis of protein ratios [11]. The analysis of modified peptides does not only require determining peptide ratios instead of protein ratios but actually necessitates additional data processing steps. These include the validation of the modification sites on the peptides, the integration of publicly known PTMs, and the relation of modified peptide ratios with the corresponding protein ratios to eliminate apparent PTM regulation caused by the sole protein regulation. As it was the case previously, this new PTM extension is released as free open source software implemented in R and available as part of the isobar Bioconductor package. It provides a complete workflow for handling quantitative PTM data from their validation to user report generation. Currently, Mascot [12], Phenyx [13], Rockerbox [14], comma separated, and PSI mzIdentML identification formats are supported. Isobar is available from the Bioconductor web site (http://www.bioconductor.org).

2. Materials and methods

Programming was done in the R statistical programming language [15] and all the features described in this paper were implemented in the isobar package [11]. The novel PTM functionality is accessible via user report generation options and new specific functions of isobar.

The access to public PTMs from neXtProt [16] is performed via REST-compatible searches (URL http://www.nextprot.org/rest/). The results are retrieved in JSON format and parsed into the ptm.info data frame of the isobar package.

Integration of the PhosphoRS [17] phosphorylation localization tool was realized by using the free stand-alone command line version of PhosphoRS. PhosphoRS does not feature a graphical user interface but requires XML input instead. Isobar^{PTM} integrates generic readers and writers for such a situation and thus provides a seamless interface to PhoshpoRS and other similar external tools.

Validation of statistical models at the peptide level was achieved using data from isobar original publication [11] to assess true and false positive rates of peptide selection as well as the adequacy of the statistical distributions underlying isobar statistics. We further validated the ratio null distribution

2.1. *Application sample data*

We downloaded Phanstiel et al. raw MS data [18] from Tranche. Peak picking and processing was performed using ProteoWizard [19] and the resulting peak lists were searched with Mascot 2.3.0 against the UniProtKB/SwissProt human database [20] appended with sequences of common contaminants (sheep keratin and bovine serum albumin). Fixed modifications were set to cysteine Carbamidomethylation, iTRAQ 4-plex at the peptide N-terminus and lysine side chains. Methionine oxidation was set as variable modification. The phospho dataset was searched with phosphorylation on serine, threonine, and tyrosine residues as variable modifications and mass tolerance was set according to the original publication [18], i.e. precursors 4.5 Da and fragments 0.01 Da. In-house developed scripts were used to filter peptide-spectrum matches to a 1% false discovery rate (FDR) at the protein group and peptide level utilizing reversed database searches. Accordingly, proteins with 2 unique peptides above an ion score threshold of 16, or with a single peptide above a threshold of 40 were selected as unambiguous identifications. Additional peptides for these validated proteins with ion score >12 were also accepted. Only those peptides with a PhopshoRS [17] probability >0.9 were considered for quantitation. The quantification was performed with default isobar settings. From the peak lists, fragments with reporter tag mass +/– 0.005 m/z were extracted and corrected for isotopic impurities. iTRAQ channels were normalized to an equal median intensity. The higher-energy c-trap dissociation (HCD) noise model supplied with the isobar package was used.

3. Results and discussion

In our previous work [11] that established the isobar statistical framework we carefully integrated important elements for selecting significant ratios. Briefly, we eliminated outlier ratios from individual spectra obviously distorted by co-eluting peptides and modeled the technical as well as the biological variability. This allowed for a simple and safe selection of protein ratios that were reliably measured and with sufficient magnitude compared to the sample natural variability. This previous work also included generalized statistical models to take advantage of replicates with a single iTRAQ or TMT experiment, and, in general, put great emphasis on the value of statistically sound methods to obtain robust and competitive methods. Here, we describe isobar^{PTM}, the extension of isobar for the analysis of modified peptide ratios.

Clearly, to bring the whole analysis to the peptide level requires computing peptide ratios instead of protein ratios. That is, all the spectra assigned to a specific peptide/PTM combination (distinct copies of the same peptide can display different patterns of PTMs) are combined in a single weighted ratio calculation taking into account signal intensities and technical variability as previously described for the protein level [11]. Beyond the change in the analysis level, several additional issues that are specifically related to PTM analysis arise and must be properly addressed (Fig. 1). In this section, we present and discuss these various issues followed by two general improvements relevant to PTM quantitation and a comparison with other tools.

3.1. *Validation of PTM site localization*

The localization of PTM sites on modified peptides identified by MS can be ambiguous and, accordingly, only reliably localized PTMs should enter the quantitative analysis. This problem

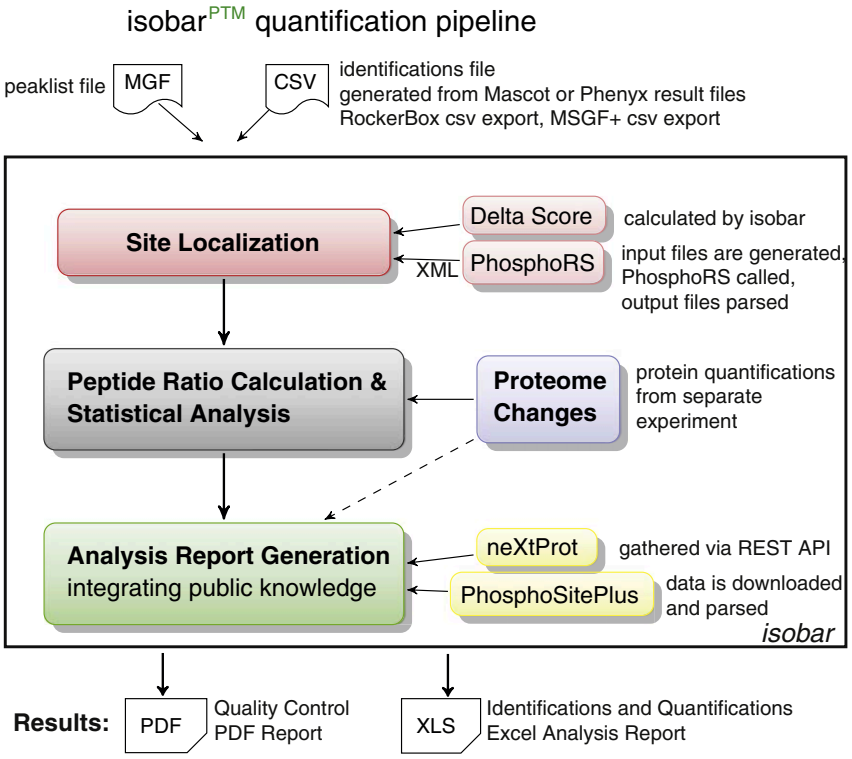


Fig. 1 – Workflow for generating quantitative PTM analysis reports. Peptide–spectrum matches with uncertain localizations of the modifications are removed using a difference or probability score (red box). Reliable matches are used to calculate ratios of modified peptides. Protein ratios from a separate experiment (blue box) can be used to correct modified peptide ratios (solid line) or integrated in the analysis report, and are displayed next to the modified peptide ratios (dotted line). The analysis report (green box) in Excel format integrates previously published knowledge on identified sites, harvesting neXtProt and PhosphoSitePlus. A PDF report containing quality control figures is generated.

mostly occurs when several amino acids of a peptide can carry a certain PTM. For instance the peptide AAGSWHSILSK can be phosphorylated at 3 positions (serines) and if it is singly phosphorylated there are 3 possible localizations. Protein identification search engines provide scores for peptide–spectrum matches that can identify the correct localization provided the peptide fragment coverage is sufficient. In practice, nonetheless, the score alone is not reliable enough [21]. To generally address this issue we integrated a universal method of validating PTM localizations, i.e. the Mascot Delta Score [22]. Although this technique was introduced for phosphorylations and is based on Mascot peptide ion scores, it is in reality of general applicability. It compares the difference between the best- and second best-scoring peptide–spectrum matches for a given peptide and PTM, with distinct modification sites, e.g. AAGS(phos)WHSILSK versus AAGSWHS(phos)ILSK to refer to the above example. The peptide identification score difference informs on the amount of information in the fragmentation spectrum to support one localization versus another one. It provides a measure of confidence in the localization and its analysis was performed by its authors. Since it only relies on score differences it is applicable to any PTM under the condition that the search engine provides multiple peptide/PTM matches for each spectrum and not only the best-scoring one. This is the case of Mascot and many other programs such as Phenyx.

Given the importance of identifying phosphorylated peptides, more advanced procedures of reliable localization have been proposed for this specific case [17,23–27]. To offer the possibility to implement or use external specialized and different PTM localization functions we introduced a generic mechanism of spectrum annotation in isobar^{PTM}, which we exploited to integrate PhosphoRS [17] for phosphorylation localization as an alternative to the Mascot Delta Score approach.

3.2. *Summarizing and quantifying at the level of the modified peptides*

As explained above the computation of modified peptide ratios necessitates introducing another level of organization in the data such that all the spectra – with safe PTM localizations – can be combined for one specific peptide sequence and PTM pattern. We validated that the statistical models introduced for the protein level [11] are still valid at the peptide level by repeating the analysis we conducted for protein ratios [11]. In particular, we assessed that (1) a heavy tailed distribution is appropriate to model peptide ratio null distributions (Supplementary Figs. S1–S3); (2) regulated peptide selection false positive rates are accurately estimated by the statistical models (Supplementary Table S1). We further estimated the true positive rate for different peptide ratios and underlying protein abundance

(Supplementary Table S2). These results, which resemble protein ratio results strongly, are not surprising since isobar protein and peptide ratios are computed identically. As a matter of fact, we do not distinguish between different peptides when we compute protein ratios [11] meaning that a ratio is always a weighted sum in our calculations (sum because we work in the log-scale and weighted by a variance estimate of each spectrum ratio [11]). We concluded this validation by showing that modified peptide ratios also follow a heavy tailed distribution (Supplementary Fig. S4).

The accurate modeling of modified peptide ratios is not necessarily sufficient to obtain biologically relevant results. The observed ratio of a modified peptide is the integrated change of the modification state and the underlying protein abundances and, when quantifying modification state changes, the change in protein abundance – if measured – should not be ignored. Wu et al., comparing the phosphoproteomes of FUS3 or STE7 yeast knockout strains against wild type [28], discussed this problem in great detail and found that 25% of the apparently regulated phosphopeptides disappeared after protein ratio correction. Having access to a high coverage of the proteome in yeast, they were able to calibrate over 96% of the phosphopeptide ratios. In our experience, working with human samples, the overlap between the proteins detected with both unmodified peptides, to estimate protein abundance change, and modified peptides simultaneously resides in the 60–90% range depending on the sample. Note that a PTM enrichment procedure preceding MS, as it is commonly done for phosphopeptide mapping, might require measuring the protein ratios from a separate set of samples. In isobar^{PTM}, we enabled the optional correction of modified peptide ratios when the protein ratio is available, in which case the peptide ratio is divided by the protein ratio. Namely, if Rn is the observed modified peptide ratio and Rp the observed protein ratio, then Rm, the corrected peptide ratio (i.e. its modification state change), is Rm = Rn − Rp (ratios in the log-scale). An adjustment to the estimated variance of Rm is also determined to comply with our general procedure of selecting significantly regulated peptides; the formulas are provided as Supplementary Information.

To exemplify ratio corrections on a human sample, we decided to reanalyze the iTRAQ 4-plex dataset published by Phanstiel et al. [18], who compared embryonic stem cell (ESC) lines with induced pluripotent stem cell (iPSC) lines and a fibroblast cell line. Using the ESC H1 as a reference, in line with the authors, we found that the strongest difference in phosphorylation is observed when comparing with the fibroblast cell line NFF (Fig. 2A), whereas the differences comparing H1 with another ESC line H9 and an iPSC line DF19.7 were very modest (ESCs are similar to iPSC [18]). Turning to the question of correcting phosphorylation site ratios with protein ratios, we found protein ratios for 77% of the phosphopeptides we identified. Applying the same fold-change threshold of 2 as Phanstiel et al., 48% of corrected phosphopeptide ratios were no longer significant after correction with a matching protein ratio, a massive change in the overall sample picture (Fig. 2B & C). Specific examples of four phosphorylated peptides are shown in Supplementary Fig. S5, including cases where the corrected ratio is augmented, reduced, and reversed compared to the original ratio.

Analyzing the enrichment of specific GO terms in differentially expressed and phosphorylated proteins using DAVID (http://david.abcc.ncifcrf.gov), we could recapitulate the findings of Phanstiel et al. Proteins higher in ESCs compared to NFF were enriched in cell cycle-related processes (e. g. chromosomal organization), those higher in NFF were enriched in cytoskeletal processes.

3.3. *Generation of user reports and integration with published PTM data*

The isobar package creates reports for quality control (Fig. 3) and quantification analysis and this feature has been extended to cover modified peptides. Reporting results at the peptide level dramatically augments the size of the data to return to the user and the PDF report we generate for the protein level is no longer appropriate. We hence extended and made fully navigable the already existing spreadsheet user report to also accommodate the peptide level (Fig. 3). It now provides links from quantified peptides to identified spectrum matches, enabling checking of the raw data, etc. Identification information includes search engine scores, modification site localization scores, and extracted isobaric report masses and intensities.

Public databases collect thousands of protein modification sites reported in the literature. To present an overview of existing knowledge about experimentally identified modification sites, we query PTM information-containing databases during user report generation. The neXtProt database [16] is our main source, which we reach via their on-line API (Materials and methods). An alternative source we also support is PhosphoSitePlus [29] that provides a second comprehensive resource of experimentally observed PTMs, primarily phosphorylations although ubiquitinylations and acetylations are covered as well. Isobar integrates PhosphoSitePlus data, automatically downloading the most recent of their monthly updated datasets at the time of report generation, parsing and mapping the data to the experimentally identified proteins. The isobar^{PTM} PTM annotation framework allows users to include supplementary PTM annotation resources if needed.

3.4. *Further improvements*

Having described all the necessary new functionalities implemented to support the analysis of quantitative PTM data, we briefly mention two improvements of isobar that are of general interest and thus impact modified peptide data processing as well.

Firstly, combinations of CID with HCD or electron transfer dissociation (ETD) fragmentation methods are commonly used in iTRAQ or TMT protocols to achieve more identifications on the basis of a fast method (CID), while more accurate quantification is obtained on the basis of the slower but more precise method (HCD or ETD) limited to a narrow mass range covering the iTRAQ or TMT channels [30]. In such a case, isobar can merge identification runs (e.g. from CID) and quantification runs (e.g. from HCD spectra) while reading the MS data, and even combine identifications obtained from quantification runs when they include regular fragment information as well. For instance, CID and HCD can provide complementary peptide identifications [31], which in our laboratory equipped with an

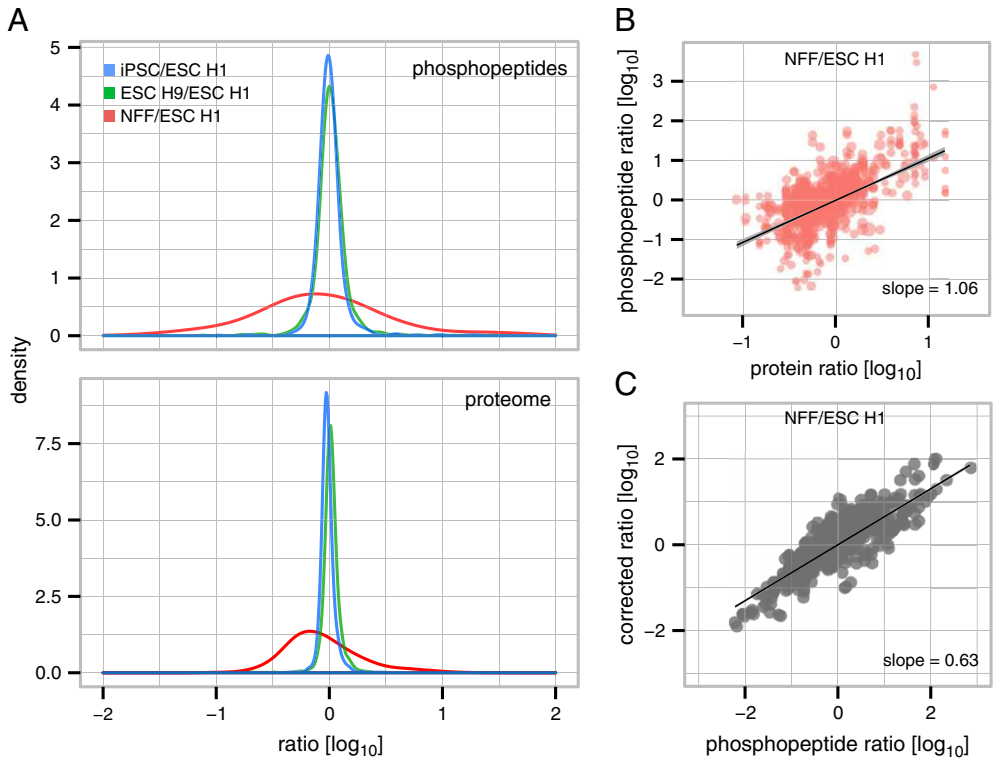


Fig. 2 – Analysis of Phanstiel et al. data. Ratios are relative to the 114 channel corresponding to H1 embryonic stem cells. (A) We observe the larger spread of ratios both in the phosphoproteome (top) and the proteome (bottom) when comparing to NFF fibroblast cells (red, channel 115) compared to H9 embryonic stem cells (green, channel 116) and DF19.7 induced pluripotent stem cells (blue, channel 117). (B) Protein ratios versus phosphopeptide ratios. We note the positive correlation indicating that a significant part of the phosphopeptide ratios originate from the protein regulation and not the phosphorylation site regulation. (C) Original versus corrected phosphopeptide ratios. The slope 0.63 < 1 confirms the general reduction of the ratios after correction.

LTQ-Orbitrap Velos (ThermoFisher Scientific, Waltham, MA) each account for 20–30% of the peptide–spectrum matches in the analysis of phosphopeptide enriched fractions.

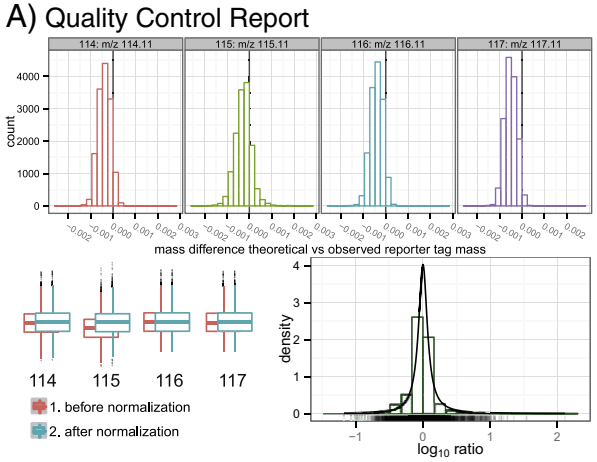
Secondly, we could find a more accurate model of heavy tailed distribution than the Cauchy. We have observed that generalized Student's t distribution better models the tails and thus improve the sensitivity of isobar (Supplementary Figs. S1–S4, S6). This distribution belongs to the generalized logistic distribution family that is a very general model of heavy tailed distribution parameterized by five parameters, which is too much for practical applications where data can be sparse. The generalized Student's t distribution has three parameters as compared to Cauchy which has only two, and it is a widely used model for heavy tailed distributions. Cauchy remains isobar default to ensure maximum robustness with smaller datasets (less than 1000 ratios, Supplementary Table 3).

3.5. *Use without programming*

The presented tool can be used with minimal configuration and no direct interaction with R: a plain text property file specifies basic parameters such as the isobaric tagging kit used (iTRAQ/TMT, 2-, 4-, 6-, or 8-plex), peak list and identification file names, and how the report and quantification should be produced (see Fig. 3). An R script, which can be called from the command line, runs the analysis with the provided parameters and generates the results. Many further options can be specified to customize the analysis and report — examples are provided with the package to guide beginners.

3.6. *Comparison with existing tools*

In Table 1 we present a feature comparison of software used in recent publications for the quantitation of isobarically tagged PTM experiments. The Coon group has developed the COMPASS [32] proteomics analysis suite for OMSSA, used recently for the quantitation of stem cell proteomes and phosphoproteomes [18]. The Marto group introduced Multiplierz [33] that provides an excellent basis for extensible workflows and data access and has been used for example for the quantitation of the mTOR regulated phosphoproteome [34]. Thermo Scientific's commercial Proteome Discoverer enables to construct a workflow from identification to quantitation. As it can be appreciated from the table, isobar's distinguishing features are its statistical fundament for quantitation and significance analysis, the high level integration of public PTM data for report generation, and the configurability and extensibility with bioinformatics packages for R/Bioconductor.



B) Report Properties

general	type	"iTRAQ4plexSpectra", …	isobaric tagging kit
	isotope.impurities	matrix	isotope impurity matrix
	peaklist	file name(s)	MGF files
	identifications	file name(s)	identification CSV files
ptm	modif	"PHOS", "ACET", …	modification to track in the report
	ptm.info.f	function name	*e.g.* "getPtmInfoFromNextProt"
	correct.ratios.with	data.frame	proteome ratio table
quantification	noise.model	NoiseModel object	technical variability model
	class.labels	vector	classes of tag channels
	normalize	boolean	normalize intensities?
	summarize	boolean	summarize ratios of the same class?
	ratios.opts	list, see '?peptideRatios'	additional properties for quantitation
report	write.qc.report	boolean	generate quality control report?
	write.xls.report	boolean	generate spreadsheet analysis report?
	xls.report.format	"long" or "wide"	layout of spreadsheet
	xls.report.columns	"p.value.ratio", …	columns visible in spreadsheet

C) Analysis Report

	Sequence	Phosphorylation Position	ACs	ID	Description	Gene	Spectra	Channels	Ratio	Significa	Ratio Minus Sd	Ratio Plus Sd	P Value Rat	P Value Sample
1	SSPNPFVGS(p)PPK	S401*;S380*	P98082-[1,3]	DAB2_HUMAN	Disabled homolog 2	DAB2	1	116 / 114	6.48	1	4.13	10.17	0.0000	0.0492
5	KAEPS(p)EVDMNSPK	S65	Q9NR30-1	DDX21_HUMAN	Nucleolar RNA helicase 2	DDX21	5	115 / 114	0.06	1	0.03	0.15	0.0004	0.0338
6	NEEPS(p)EEELDAPKPK	S121*	Q9NR30-1	DDX21_HUMAN	Nucleolar RNA helicase 2	DDX21	6	115 / 114	0.09	1	0.03	0.29	0.0224	0.0377
2	EES(p)EEEEDEDDEEEEEEK	S32*	P35659	DEK_HUMAN	Protein DEK	DEK	2	115 / 114	0.02	1	0.01	0.06	0.0003	0.0231
2	S(p)LLVEGK	S51*	P35659	DEK_HUMAN	Protein DEK	DEK	2	115 / 114	0.11	1	0.07	0.17	0.0000	0.0419
2	KPATPAEDDEDDDLDLFGS(p)D	S162*;S528*	P2		on factor 1-delta	EEF1D	2	115 / 114	7.04	1	5.43	9.13	0.0000	0.0471
5	AS(p)STSTPEPTR	S485	Q		enabled homolog	ENAH	5	115 / 114	0.01	1	0.00	0.11	0.0348	0.0186
3	LWT(p)PLK	T143	Q9NYF3	FA53C_HUMAN	Protein FAM53C	FAM53C	3	115 / 114	0.06	1	0.05	0.08	0.0000	0.0331
3	ATEDGEEDEVS(p)AGEK	S1435*	Q9BXW9-2	FACD2_HUMAN	Fanconi anemia group D2 pro	FANCD2	3	115 / 114	0.07	1	0.03	0.17	0.0015	0.0349
28	RAPS(p)VANVGSHC(c)DLSLK	S2152*;S2144*	P21333-[1,2]	FLNA_HUMAN	Filamin-A	FLNA	28	115 / 114	7.01	1	3.50	14.05	0.0025	0.0472
4	AGGSAALSPS(p)K	S33*	Q92522	H1X_HUMAN	Histone H1x	H1FX	4	115 / 114	0.07	1	0.05	0.10	0.0000	0.0346
2	S(p)APAPK	S7*	O60814,P06899	H2B1B_HUMAN, H	Histone H2B type 1-B, Histon	H2BFS, HIST	2	115 / 114	0.10	1	0.06	0.14	0.0000	0.0394
2	LEDVGS(p)DEEDDS(p)GKDK	S255*&S261*	P08238	HS90B_HUMAN	Heat shock protein HSP 90-be	HSP90AB1	2	115 / 114	0.12	1	0.08	0.19	0.0000	0.0441
1	C(c)TPAC(c)LS(p)FGPK	S40	P34932	HSP74_HUMAN	Heat shock 70 kDa protein 4	HSPA4	1	115 / 114	42.38	1	28.64	62.72	0.0000	0.0247

Fig. 3 – Isobar^{PTM} quantification reports. (A) Quality control report showing reporter tag mass precision, reporter tag intensities before and after normalization, and a histogram of peptide ratios along with the fit Cauchy biological variability ratio distribution [11]. (B) Report generation is controlled by a properties file. Columns: property name, possible values, and explanation. (C) Spreadsheet user report. It includes modified peptide sequence with the positions of the modifications in the protein sequence (separated by semicolons if in multiple identical peptides or by ampersands if multiple occurrences in the same peptide). A star identifies positions previously reported in the literature, tooltips display information on the latter PTMs (here from neXtProt). The report has multiple tabs for identifications and contains multiple links to navigate them, e.g. from a modified peptide as featured in the figure to all the spectra supporting its identification.

Table 1 – Comparison with similar software packages.

	Isobar^{PTM}	Proteome Discoverer	COMPASS	multiplierz
Availability	Open source	Commercial	Open source	Open source
iTRAQ and TMT Quant	Yes	Yes	Yes	Yes
Statistical Framework	Yes, technical and biological variability	no	no	Technical variability modeled [a]
PTM Localization	Yes [b]	Yes [c]	No	Yes [a]
Annotation of PTM sites	Yes [d]	No	No	No
Correction with Protein Ratios	Yes	Yes	Yes	Yes
Restrictions	No graphical user interface	Closed source	For usage with OMSSA only	Scripting skills required

[a] Scripts for robust error model and Mascot Delta Score available on the multiplierz homepage http://blais.dfci.harvard.edu/index.php?id=106.
[b] PhosphoRS and Mascot Delta Score.
[c] PhosphoRS.
[d] NextProt and PhosphoSitePlus.

4. Conclusion

To measure and understand PTMs in disease and biological processes is an important objective of current research in proteomics. Such experiments remain challenging but the technology has made such tremendous progresses that in-depth and proteome-scale mappings of specific PTMs can be realized with unprecedented accuracy. As a consequence, data analysis faces difficulties that are common to most omics fields: the access to reliable and highly automated methods of processing and selecting relevant data conditions the extent to which discoveries can be accomplished. With this consideration in mind, we started to develop a combined statistical and software framework – isobar – that we originally targeted towards protein expression studies [11]. The work presented here implements a second step aimed at including the peptide PTM regulation level within the scope of the analyses supported by this platform. We named this specific branch of the project isobar^{PTM}.

The approach we have followed remains in line with the original concepts that guided isobar design: the establishment of robust and accurate statistical models provides the most appropriate basal layer to construct a successful software platform. In isobar^{PTM} we greatly benefited from the initial effort to the point where no real additional statistical modeling was necessary, just validations and small adaptations. The models developed for the proteins turned out to be adequate for the peptides as well and we could concentrate on establishing the new software functionalities. Doing so, we also benefited from the general improvements and bug-fixes we kept introducing in the isobar libraries that has been applied to a multitude of projects by ourselves [35] and others [36] meanwhile.

Practically, successful and high quality analysis of PTM data on a large-scale preventing the manual inspection of each and every interesting spectrum implies the execution of several tasks that are generally not all accessible to the average proteomics laboratory in the best conditions. With isobar^{PTM} we have streamlined the fundamental steps of extracting and combining identification and MS data, including when hybrid fragmentation strategies e.g. CID-HCD are adopted, performing an automatic validation of the localization of the modification sites and removing dubious cases, and applying state of the art statistical modeling to compute ratios and assess their significance (Fig. 1). Furthermore, convenient user reports are produced which include a navigable sophisticated spreadsheet that represents a convenient paradigm for reporting large sets of results as generated by peptide level studies.

Finally, we believe that bioinformatics tools should be as interoperable as possible and the development of open source R Bioconductor packages represents an effective way of implementing this goal. In particular, follow up functional analyses such as GO term or pathway enrichments are made straightforward thanks to many existing Bioconductor packages. Developing within the R platform allows other bioinformaticians to use isobar at all possible levels, from calling high-level functions down to completely redesigned analyses capitalizing on the low-level functions. For non-bioinformaticians and for usage within an automated pipeline, we make the complete analysis with report generation accessible on the command line requiring simple configuration via text files only. In the future of the isobar project, we will give significant attention to the development of a graphical user interface.

Isobar and isobar^{PTM} can be downloaded from http://www.ms-isobar.org or from the Bioconductor web site.

Acknowledgments

We thank all our CeMM colleagues and in particular André Müller, Uwe Rix, Alexey Stukalov, and Keiryn Bennett for useful feedback and advices. JC is supported by an Austrian Science Fund (FWF) grant No P 24321-B21.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.02.022.

REFERENCES

[1] Garavelli JS. The RESID database of protein modifications as a resource and annotation tool. Proteomics 2004;4:1527–33.
[2] Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, Pedrioli PG, et al. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci Signal 2010;3:rs4.
[3] van Noort V, Seebacher J, Bader S, Mohammed S, Vonkova I, Betts MJ, et al. Cross-talk between phosphorylation and lysine acetylation in a genome-reduced bacterium. Mol Syst Biol 2012;8:571.
[4] Mallick P, Kuster B. Proteomics: a pragmatic perspective. Nat Biotechnol 2010;28:695–709.
[5] Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol 2003;21:255–61.
[6] Bodenmiller B, Aebersold R. Quantitative analysis of protein phosphorylation on a system-wide scale by mass spectrometry-based proteomics. Methods Enzymol 2010;470: 317–34.
[7] Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell 2011;44:325–40.
[8] Henriksen P, Wagner SA, Weinert BT, Sharma S, Bacinskaja G, Rehman M, et al. Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in *Saccharomyces cerevisiae*. Mol Cell Proteomics 2012;11:1510–22.
[9] Thompson A, Schafer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem 2003;75:1895–904.
[10] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3:1154–69.
[11] Breitwieser FP, Muller A, Dayon L, Kocher T, Hainard A, Pichler P, et al. General statistical modeling of data from protein relative expression isobaric tags. J Proteome Res 2011;10:2758–66.
[12] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999;20:3551–67.
[13] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. Proteomics 2003;3:1454–63.
[14] van den Toorn HW, Munoz J, Mohammed S, Raijmakers R, Heck AJ, van Breukelen B. RockerBox: analysis and filtering of massive proteomics search results. J Proteome Res 2011;10: 1420–4.
[15] R_Core_Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012.
[16] Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, et al. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 2012;40:D76–83.
[17] Taus T, Kocher T, Pichler P, Paschke C, Schmidt A, Henrich C, et al. Universal and confident phosphorylation site localization using phosphoRS. J Proteome Res 2011;10:5354–62.
[18] Phanstiel DH, Brumbaugh J, Wenger CD, Tian S, Probasco MD, Bailey DJ, et al. Proteomic and phosphoproteomic comparison of human ES and iPS cells. Nat Methods 2011;8:821–7.
[19] Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24:2534–6.
[20] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.
[21] Chalkley RJ, Clauser KR. Modification site localization scoring: strategies and performance. Mol Cell Proteomics 2012;11: 3–14.
[22] Savitski MM, Lemeer S, Boesche M, Lang M, Mathieson T, Bantscheff M, et al. Confident phosphorylation site localization using the Mascot Delta Score.Mol Cell Proteomics 2011;10 [M110 003830].
[23] Bailey CM, Sweet SM, Cunningham DL, Zeller M, Heath JK, Cooper HJ. SLoMo: automated site localization of modifications from ETD/ECD mass spectra. J Proteome Res 2009;8:1965–71.
[24] Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol 2006;24:1285–92.
[25] Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 2006;127:635–48.
[26] Ruttenberg BE, Pisitkun T, Knepper MA, Hoffert JD. PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. J Proteome Res 2008;7:3054–9.
[27] Swaney DL, Wenger CD, Thomson JA, Coon JJ. Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. Proc Natl Acad Sci U S A 2009;106:995–1000.
[28] Wu R, Dephoure N, Haas W, Huttlin EL, Zhai B, Sowa ME, et al. Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol Cell Proteomics 2011;10(8):M111 009654.
[29] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 2012;40:D261–70.
[30] Kocher T, Pichler P, Schutzbier M, Stingl C, Kaul A, Teucher N, et al. High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all. J Proteome Res 2009;8:4743–52.
[31] Frese CK, Altelaar AF, Hennrich ML, Nolting D, Zeller M, Griep-Raming J, et al. Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos. J Proteome Res 2011;10:2377–88.
[32] Wenger CD, Phanstiel DH, Lee MV, Bailey DJ, Coon JJ. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. Proteomics 2011;11:1064–74.
[33] Parikh JR, Askenazi M, Ficarro SB, Cashorali T, Webber JT, Blank NC, et al. multiplierz: an extensible API based desktop environment for proteomics data analysis. BMC Bioinformatics 2009;10:364.
[34] Hsu PP, Kang SA, Rameseder J, Zhang Y, Ottina KA, Lim D, et al. The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling. Science 2011;332:1317–22.
[35] Winter GE, Rix U, Carlson SM, Gleixner KV, Grebien F, Gridling M, et al. Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML. Nat Chem Biol 2012;8:905–12.
[36] Gluck F, Hoogland C, Antinori P, Robin X, Nikitin F, Zufferey A, et al. EasyProt — an easy-to-use graphical platform for proteomics data analysis. J Proteomics 2012;79C:146–60.

4.2.3. Supporting information

Null distribution

In the original paper describing the isobar statistical models (Breitwieser et al., 2011) we showed that unregulated protein ratios follow a heavy tailed distribution. Here, exploiting three datasets already described in this paper to cover multiple MS and labeling techniques (Orbitrap iTRAQ and TMT, MALDI-TOF/TOF TMT), we collected peptide ratios from unregulated proteins in these datasets and observed that they also follow a heavy-tailed distribution (figs. S1 to S3).

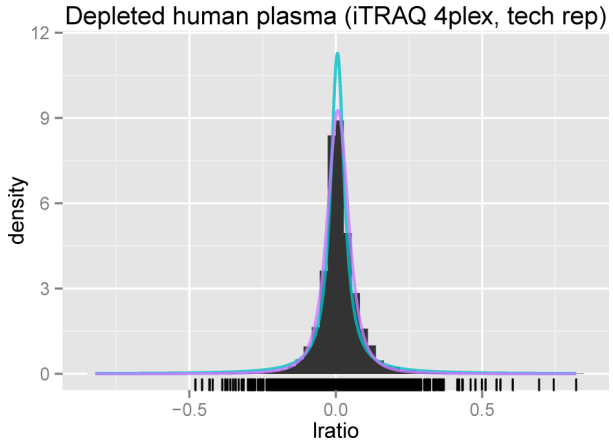


Figure S1: Peptide ratios from TS1 dataset (Breitwieser and Colinge, 2013), LTQ-Orbitrap, iTRAQ 4-plex labeling. Cauchy (blue) and generalized Student's T (pink) models.

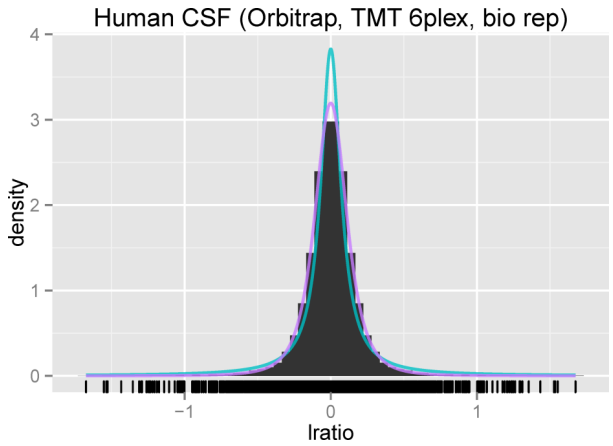


Figure S2: Peptide ratios from MALDI-TOF/TOF dataset (Breitwieser et al., 2011), TMT 6-plex labeling. Cauchy (blue) and generalized Student's T (pink) models.

We then performed the same analysis using the data of Phanstiel et al. (2011) collecting phosphorylated peptide ratios between replicates, i. e. modified peptide ratios for unregulated

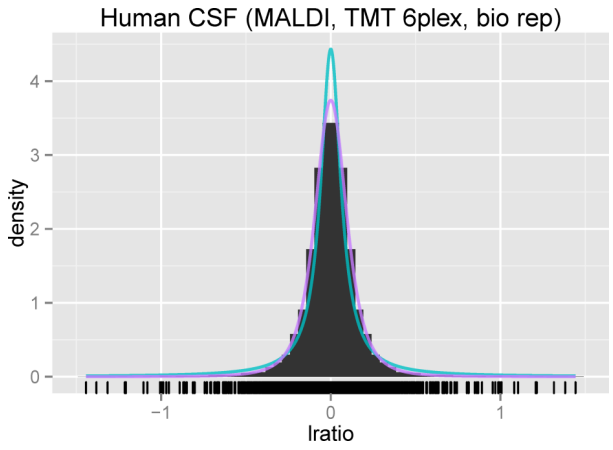


Figure S3: Peptide ratios from LTQ-Orbitrap, CSF fluid dataset (Breitwieser and Colinge, 2013), TMT 6-plex labeling. Cauchy (blue) and generalized Student's T (pink) models.

PTMs. The result is identical (fig. S4) thus validating isobar null distribution at the modified peptide level.

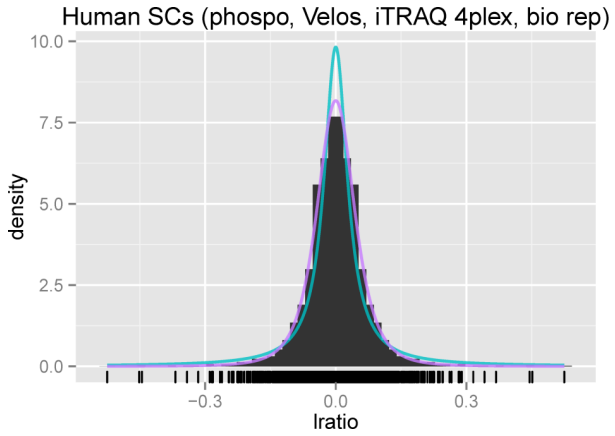


Figure S4: Unregulated phosphorylated peptide ratios (Phanstiel et al., 2011). Cauchy (blue) and generalized Student's T (pink) models.

Validation of the selection model and performance evaluation

Again exploiting further the depleted plasma test sample generated for the original isobar paper, we conducted a novel false/true positive rate evaluation but at the peptide level. Selecting peptide ratios at 5% false positives according to the statistical model actually delivered false positive rates close and below this limit when comparing biological replicates of the same sample where no ratio should be selected ideally (Suppl. Table S1).

Table S1: False positive rates observed when a 5% threshold was imposed on the selection.

Num spectra	Isobar Cauchy[a]	Isobar general. T[b]	T-test	Fold change[c]
1	0.01	0.01	0.00	0.08
2	0.01	0.01	0.06	0.04
3	0.01	0.03	0.09	0.03
5	0.01	0.02	0.27	0.02
10	0.00	0.02	0.45	0.01
15	0.00	0.01	0.50	0.00
20	0.00	0.01	0.54	0.00

[a] Isobar significant ratio selection procedure with Cauchy null distribution for modeling biological variability.
[b] Cauchy replaced by the more accurate generalized Student's T.
[c] Fold change of 1.5 considered as a significant ratio.

True positive rate estimations are dependent on the peptide abundance, the number of spectra available, and the actual ratio magnitude. From the test sample we selected a large number of peptides at different concentrations, known ratios, and we randomly selected different numbers of spectra (when more were available) following the procedure we applied to characterize protein detection performance in Breitwieser and Colinge (2013). Results are similar to protein level performance (table S2) and also show that isobar selection, which was able to control false positives successfully (table S1), is not always the most sensitive but more sensitive methods are typically the ones yielding far unacceptable false positive rates. We hence conclude that the selection method is appropriately ported to the peptide level.

Table S2: True positive rates.

Num spectra	Isobar Cauchy	Isobar general. T	T-test	Fold change
Peptide ratio 1.3				
1	0.13	0.16	0.00	0.41
2	0.11	0.14	0.07	0.24
3	0.14	0.22	0.21	0.19
5	0.13	0.34	0.37	0.17
10	0.13	0.39	0.62	0.16
15	0.17	0.47	0.75	0.20
20	0.20	0.48	0.73	0.23
Peptide ratio 1.5				
1	0.27	0.27	0.00	0.51
2	0.22	0.24	0.07	0.43
3	0.31	0.39	0.24	0.40
5	0.28	0.46	0.44	0.34
10	0.40	0.61	0.68	0.50
15	0.47	0.72	0.82	0.61
20	0.54	0.68	0.80	0.64

Table S2: (continued)

Peptide ratio 2 (abundant)				
1	0.66	0.65	0.00	0.87
2	0.75	0.76	0.33	0.95
3	0.88	0.91	0.78	0.94
5	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00
15	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00
Peptide ratio 2 (low)				
1	0.42	0.41	0.00	0.68
2	0.47	0.47	0.19	0.75
3	0.59	0.62	0.51	0.77
5	0.69	0.79	0.78	0.80
10	0.94	0.97	0.98	0.97
15	0.96	0.99	1.00	0.99
20	0.98	1.00	1.00	1.00

Modified peptide ratio correction

As discussed in the paper, Wu et al. (2011) showed that correcting modified peptide ratios by the abundance change ratio of the corresponding proteins much improves their accuracy. When both ratios are available, i. e. modified peptide and protein ratios, we perform this correction and compute an upper bound on the variance of the corrected ratio, which is used in the statistical test taking care of the modified peptide selection.

Let R_n be the observed log-ratio of a modified peptide and Rp the log-ratio of the corresponding protein. We estimate the true modified peptide ratio R_m by the following formula:

$$R_m = R_n - R_p \tag{4.2.1}$$

The variance of R_m is given by

$$\mathrm{Var}(R_m) = \mathrm{Var}(R_n) + \mathrm{Var}(R_p) + 2\,\mathrm{Cov}(R_n, R_p) \tag{4.2.2}$$

The covariance $\mathrm{Cov}(R_n, R_p)$ of the peptide and protein ratios, however, is unknown. Omitting the covariance term means assuming independence between R_n and R_p but this is wrong in general since an increase in R_p causes an increase of R_n in the iTRAQ or TMT measurements. For the same reason, a positive correlation can be assumed generally. We use Pearson's correlation coefficient ρ formula to modify equation (4.2.2) and obtain an upper bound of $\mathrm{Var}(Rm)$, hence

yielding conservative ratio selections. Namely, we have

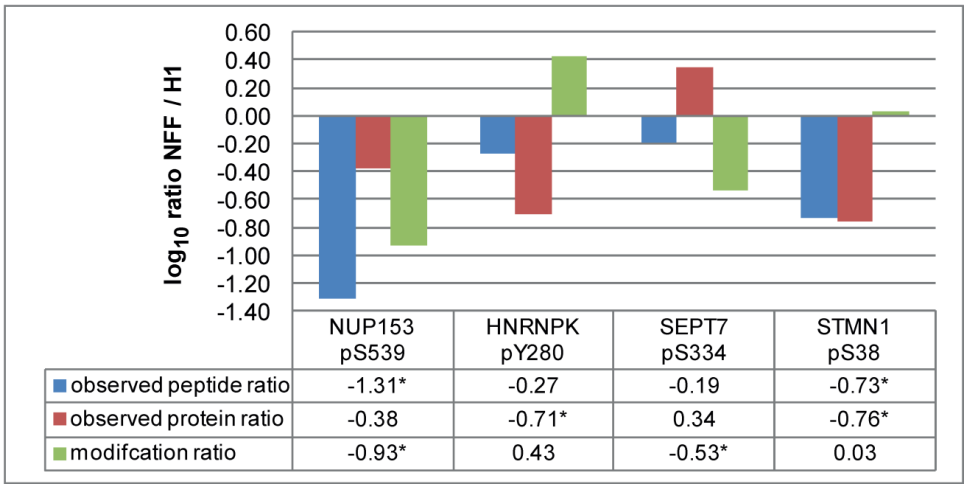
$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{s(X)s(Y)},$$
$$\mathrm{Cov}(X,Y) = \rho(X,Y)s(X)s(Y),$$

where s is the standard deviation. With $\rho = \rho(R_n, R_p)$ we further obtain

$$\mathrm{Cov}(R_n, R_p) = \rho \times s(R_n)s(R_p)$$
$$\mathrm{Var}(R_m) = \mathrm{Var}(R_n) + \mathrm{Var}(R_p) + 2\rho \times s(R_n)s(R_p).$$

ρ is not known for the pair (R_n, R_p), but it is assumed positive and $\mathrm{Var}(R_m)$ is thus bounded by $(\rho = 1)$:

$$\mathrm{Var}(R_m) = \mathrm{Var}(R_n) + \mathrm{Var}(R_p) + 2s(R_n)s(R_p).$$



	NUP153 pS539	HNRNPK pY280	SEPT7 pS334	STMN1 pS38
observed peptide ratio	-1.31*	-0.27	-0.19	-0.73*
observed protein ratio	-0.38	-0.71*	0.34	-0.76*
modifcation ratio	-0.93*	0.43	-0.53*	0.03

Figure S5: Examples of corrected modified peptide ratios. Observed modified peptide ratios (blue) are corrected according to observed protein ratios (red) to obtain a corrected modified peptide ratio (green).

An improved heavy tailed distribution model

In figs. S1 to S4 we have shown that unregulated peptide ratios followed a heavy tailed distribution that was well modeled by a Cauchy, which is isobar default null distribution for such ratios. In recent work we found the generalized Student's T distribution to provide a more precise model of the distribution tails. It is visible in figs. S1 to S4 (pink curves) and this is also valid for protein ratios (fig. S6).

To better model the tails of the null provides a more sensitive selection of peptides (or proteins) as can be nicely observed in table S2 without causing false positives beyond the pre-imposed error rate (table S2).

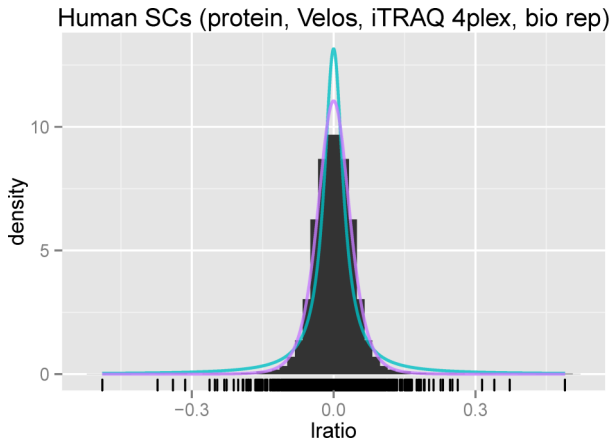


Figure S6: Unregulated protein ratios (Phanstiel et al., 2011). Cauchy (blue) and generalized Student's T (pink) models.

4.3. Additional outcomes and applications

This section presents additional outcomes of the thesis that build on and extend the results presented in sections 4.1 and 4.2. This thesis has been developed in an integrated environment with mass spectrometrists and biologists. While the test data sets presented in section 4.1 were crucial for further development and testing of the methods, the analysis of several actual data sets (Haura et al., 2011; Keiryn L. Bennett et al., 2011; Müller et al., 2012; Winter et al., 2012) lead to conception and, later, extension of the ideas and software tool.

4.3.1. Use and enhancement of the isobar package in publications

The first two publications Haura et al. (2011) and Keiryn L. Bennett et al. (2011) were analyzed before we published our method in Breitwieser et al. (2011). We used ad-hoc thresholds for intensity values, and the median for summarizing to protein ratios. However, in Keiryn L. Bennett et al. (2011), we already showed the heavy-tailedness of the protein ration distribution.

In Müller et al. (2012), we used the published `isobar` package, and extended it to calculate the counting methods dNSAF and emPAI. In Winter et al. (2012), we further extended `isobar` to use the strategy of precursor purity filtering developed by Mikhail M. Savitski et al. (2011b).

Haura et al. (2011) - Using iTRAQ combined with tandem affinity purification to enhance low-abundance proteins associated with somatically mutated EGFR core complexes in lung cancer. In Haura et al. (2011) we reported the applicability of isobaric tagging to improve the identification ratio of low-abundant proteins. We tagged and transduced mutated EGFR proteins into two mutant lung cancer cell lines (HCC827 and PC9, see fig. 4.1). Using affinity purification of the tagged proteins, we enriched binding proteins, such as the EGFR core complex members. The analysis was performed with two biological replicates and two technical replicates with iTRAQ 4-plex, as well as standard label-free shotgun proteomics. Using an inclusion list approach and higher energy dissociation, we could identify several complex members. Interestingly, we could show the presence of a certain protein, UBS3B, in the core complex of both cell lines. Using shotgun proteomics, UBS3B was seen only in one of the two cell lines. We argued that the combined analysis using iTRAQ pushes the protein above the detection limit for both samples.

For the analysis, we developed Perl scripts and investigated the structure of the technical variability. To counter the noise, a sliding window approach was used to calculate and subtract noise levels relative to retention time. Furthermore, to counter high variability of low intense reporters, a intensity threshold of 2000 was used, and low-intense ions were excluded from quantification. The ratios were summarized to protein level using the median spectra ratios of all

protein-specific peptides. Biological replicates were not combined, and no statistical assessment of the significance of changes across the cell lines was performed.

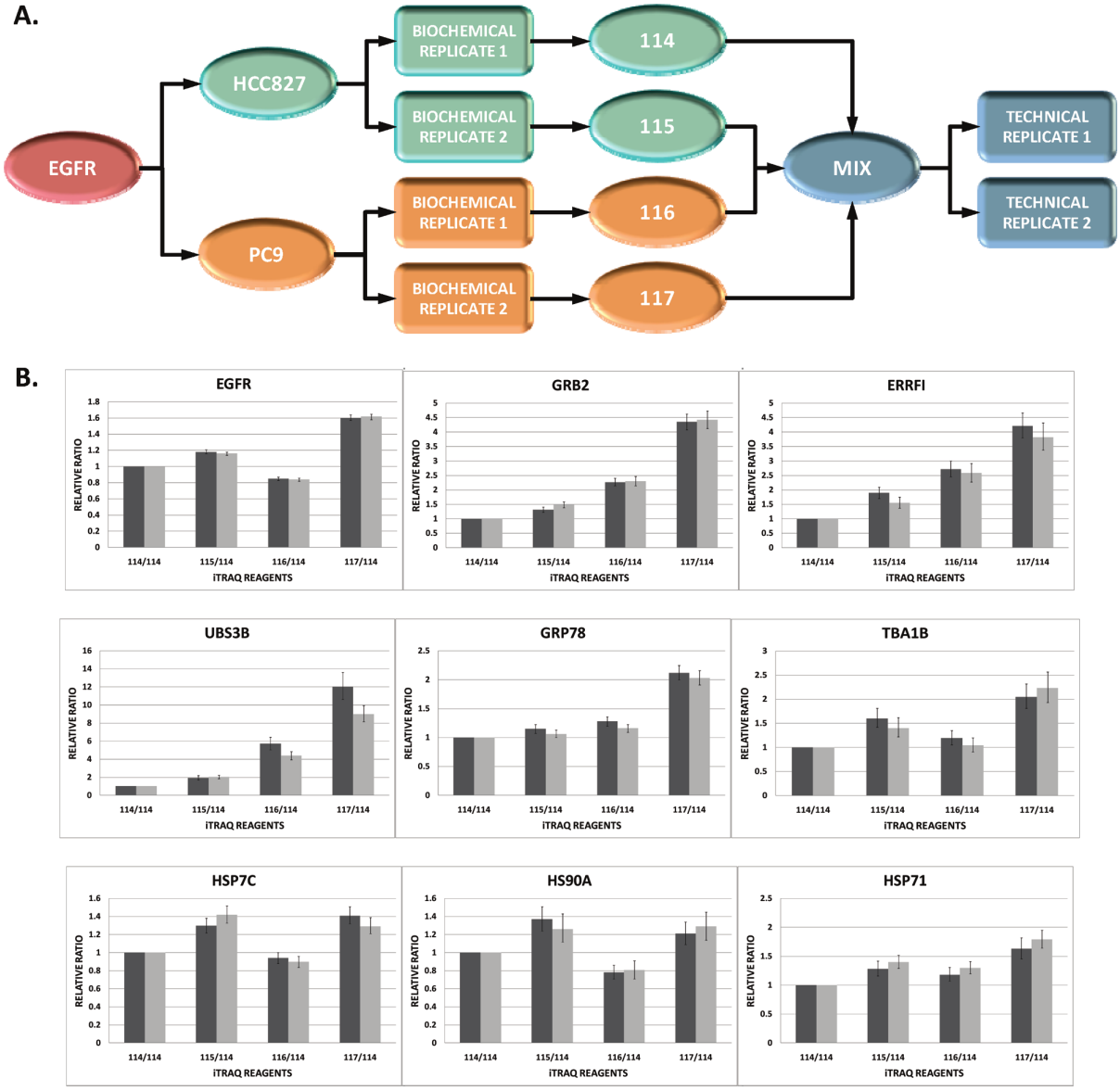


Figure 4.1: (A) Schematic overview of the iTRAQ labeling strategy: Mutant tagged EGFR was transduced in two lung cancer cell lines; biological replicates were iTRAQ labeled; mixed; and analyzed in technical replicates. (B) Relative quantitation levels for proteins interacting with EGFR. Figure from Haura et al. (2011, figure 2).

Keiryn L. Bennett et al. (2011) - Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ-labelled specimens. In Keiryn L. Bennett et al. (2011), we set the ground for analysis

of the human eye fluid. We compared separation and quantification strategies - one-dimensional gel LC-MS and two-dimensional LC-MS with and without iTRAQ labeling - to identify and quantify proteins in human aqueous humor, the fluid in the chamber in front of the eye lens. From ten patients with cataract we could extract on average 49 μg of total protein. Two patients each were pooled, and 4 pools labeled with iTRAQ. The calculation of the ratios and technical variability handling was done as in Haura et al. (2011).

We used the same methods for summarizing protein ratios as in Haura et al. (2011). Additionally, we observed that the distribution of protein ratios across the humor of different patient pools has heavy-tails, which are better explained by a Cauchy distribution than the Gaussian (see fig. 4.2). The fitted distribution is used to assess the range within which 95% of the protein ratios are situated.

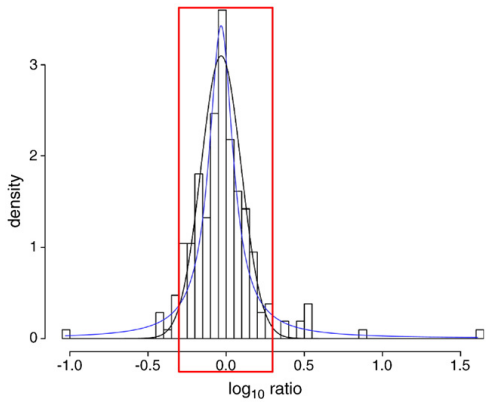
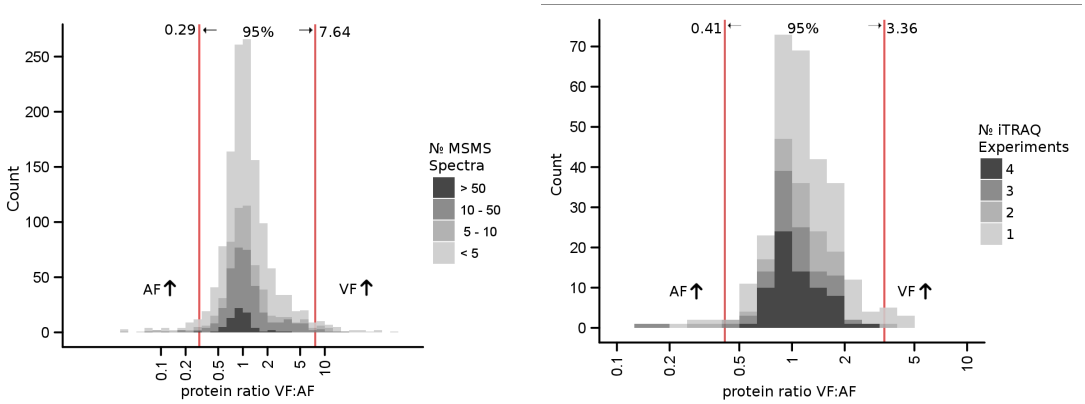


Figure 4.2: Fold change differences in aqueous humor proteins of cataract patients in two pools. The lines shows the fit of a Gaussian (black) and Cauchy (blue) distribution. The Cauchy was used to assess 95% cutoffs for differential regulation (red box). Figure from Keiryn L. Bennett et al. (2011, figure 6)

.

Pollreisz et al. (2013) - Quantitative proteomics of aqueous and vitreous fluid from patients with idiopathic epiretinal membranes. Pollreisz et al. (2013) demonstrates the application of the knowledge gathered for the analysis of eye fluids in Keiryn L. Bennett et al. (2011). We characterized the proteomes of aqueous and vitreous humor in human eyes from patients with idiopathic ediretinal membranes (iERM). Fluids from 24 patients undergoing surgery for removal of iERM were collected. Samples from eight patients were analyzed in four iTRAQ experiment, grouping the aqueous (AF) and vitreous fluid (VF) of two patients, each. The protein fold change in VF:AF was relatively big - 95% are between 0.29 and 7.64 (see fig. 4.3a). Figure 4.3b shows the combined protein ratios of up to 4 iTRAQ experiments (and eight patients).



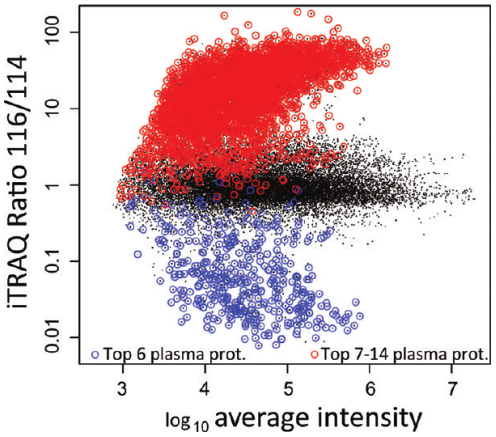
(a) Intra-individual protein ratios of eight patients.

(b) Combined protein ratios from 4 experiments and eight patients.

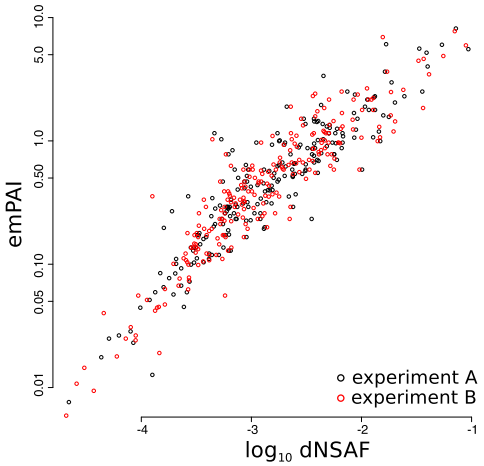
Figure 4.3: Histograms of VF:AF protein ratios (adapted from figure 4, Pollreisz et al. (2013))

Müller et al. (2012) - A comparative proteomic study of human skin suction blister fluid from healthy individuals using immunodepletion and iTRAQ labeling. In Müller et al. (2012) we compared human blister fluid proteome of healthy individuals obtained by skin suction. The study aimed at helping understanding skin-related diseases through the proteome, developing methods for blister fluid analysis. First, we assessed depletion strategies using two commercial spin columns, with antibodies against the top 6 or top 14 abundant proteins. We found that while the "top 6" method is better at filtering its share of proteins, "top 14" has more spectra depleted in total (see fig. 4.4a). We further compared iTRAQ 8-plex (eight individual patients tagged) and 4-plex (two patients pooled per tag) tagging kits. The 8-plex kit promises double the sample throughput, however we identified, in concordance with previous observations, a much lower number of observations with this kit: The number of proteins identified with 8-plex is a third of 4-plex. Interestingly, the pooling of two patients already reduced the range of the 95% interval of protein ratios by a factor of $\frac{1}{5}$. We further hypothesized that there might be a correlation between the abundance of proteins and the variability of their ratios. We implemented the "distributed normalized spectral count" (dNSAF, Ying Zhang et al. (2010)) and "empirical protein abundance index" (emPAI, Ishihama et al. (2005)) measures for estimation of the protein amount based on the number of spectra and peptides, respectively. The measures show good correlation (see fig. 4.4b), and we proceeded with dNSAF in the publication. We found no correlation between coefficient of variation and the protein abundance. We concluded with the discussion of skin-related proteins and their variability between patients.

Winter et al. (2012) - Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML. The BCR-ABL fusion oncoprotein, which results from a reciprocal translocation between chromosomes 9 and 22, is the driving cause of chronic myeloid leukemia (CML). Potent



(a) Juxtaposition of depletion strategies. "Top 6" better depletes the 6 common proteins (blue dots), but the proteins unique to "top 14" (red dots) are seen with more spectra (figure 3 in publication)



(b) Scatterplot of dNSAF and emPAI in two experiments demonstrates correlation of the label-free measures (figure unpublished)

Figure 4.4: Data from Müller et al. (2012)

kinase inhibitor targeting BCR-ABL and downstream pathways, such as imatinib, nilotinib and dasatinib have been developed to treat CML. A mutation of BCR-ABL at position 315 (T315I) renders them ineffective. In Winter et al. (2012) we investigated possible synergies of known small molecule inhibitors against BCR-ABL^{T315I} CML. After assessing dose-response for eight clinical kinase inhibitors individually, we tested them pair-wise and found pronounced synergy between danusertib and bosutinib specifically for killing BCR-ABL^{T315I}-transformed cells. To elucidate the cause we first investigated the proteins binding the eight kinase inhibitors using iTRAQ 8-plex labeling.We could quantify the affinity ratios of 43 kinases *versus the average*, identifying known specificities such as Tec and dasatinib, and stronger affinity of either danusertib or bosutinib against members of the Mapk signaling pathways (fig. 4.5). We could prove a significant enrichment on this pathway using the targets of danusertib and bosutinib individually and combined.

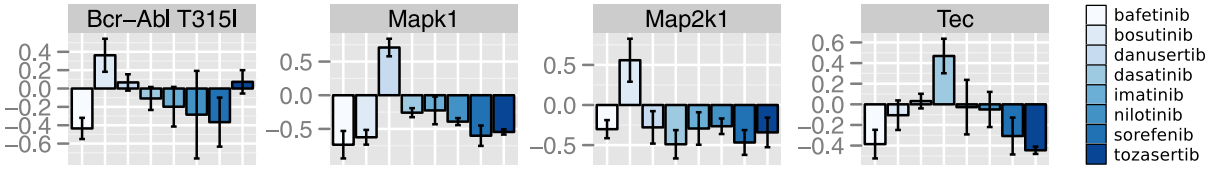


Figure 4.5: Protein affinities. $\log_{10}$ protein ratios versus average of eight kinase inhibitors (data from Winter et al. (2012), figure unpublished).

We further analyzed the global transcriptome changes upon individual and combined drug treatment and discovered many changes induced by the combination treatment, with an enrichment

of genes containing a c-Myc motif. Additional iTRAQ experiments demonstrated differential phosphorylation of the Mapk pathway, leading to its inhibition. In conclusion, our investigation demonstrated the effectiveness of a combination of danusertib and bosutinib against BCR-ABL^{T315I} CML due to their combined affect on the downstream Mapk pathway.

In the process of the analysis, we implemented a strategy to integrate precursor purity measures: iTRAQ signals are known to suffer from coeluting material, causing ratio compression (see also background section 2.3.2). Requiring a minimum signal-to-noise ratio can thus slightly improve the quantification accuracy. Mikhail M. Savitski et al. (2011b) provide scripts to calculate signal-to-interference. We observed that identified spectra have a higher precursor purity than unidentified (see fig. 4.6), and used a cutoff of 0.5 to improve quantification accuracy while limiting the effect on the number of protein identifications.

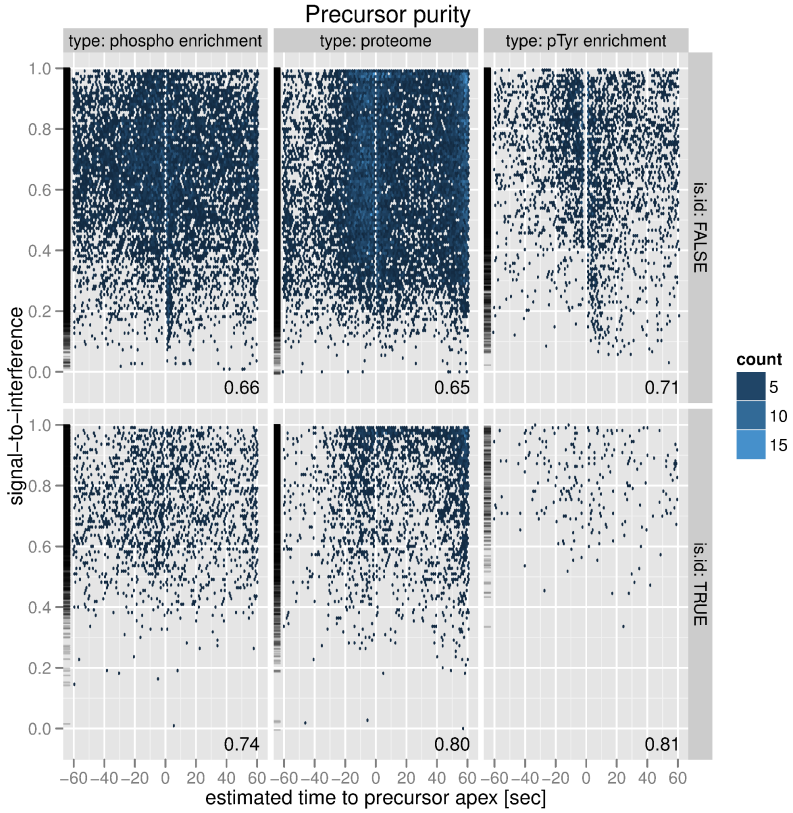


Figure 4.6: Precursor purity as measured by "signal-to-interference" (s2i) measure (Mikhail M. Savitski et al., 2011b) versus time to peak apex. Upper panel shows measures of spectra that are not matched to a peptide sequence, lower panel shows matched spectra. Columns shows phosphorylation enrichment experiment (mainly peptides with phosphorylated serines and threonines), global proteome experiment, and specific tyrosine phosphorylation experiment. For identified spectra (lower panel), the average signal-to-interference is much higher (data from Winter et al. (2012), figure unpublished).

4.3.2. Hierarchical modeling of protein ratios

The large-scale nature of data from modern biomolecular assays - such as microarrays, next-generation sequencing, and mass spectrometry - lends itself to the exploitation of not only the information obtained for each gene or protein, but also its structure across genes and proteins. The variance modeling presented in section 4.1 captures the precision of individual measurements utilizing the variance structure in technical replicates. We used these intensity-dependent estimates of the data precision to summarize the spectrum-level data to peptide- or protein-level ratios, weighing more precise data points higher. The variance of the peptide- or protein-level ratio is calculated as the maximum of the estimator variance, and the weighted variance (see eq. 3 in section 4.1). Using the maximum limits the number of false positives when few spectra are available (see Table 1 in section 4.1). The reason for the higher numbers of false positives, when few data points are available, are fluctuations in the sample variance: Sometimes, by chance, data points are near to each other. As the sample variance is used in the inference, this can lead to inflated type I errors.

It has long been recognized in the field of microarray data analysis that better inference on the individual genes can be achieved by using the observed information across all genes (Efron, 2008). This concept of "borrowing strength" (and precision) from the entirety for the estimation of the individual has been first demonstrated by Stein (1956), and later generalized by James and Stein (1961). James and Stein (1961) showed that when the means of multiple populations are to be estimated, the smallest total error will be made, when each mean is regressed towards the mean across all populations - even if the populations are not related. This approach has developed into "Empirical Bayes", in which the prior distributions are estimated based on the data (Efron, 2010). Thus, the mean and/or the variance are regressed towards the total mean, using the Bayes formula. Limma (Smyth, 2004), Cyber-T (Baldi and Long, 2001), and Significance Analysis for Microarrays (Tusher, R. Tibshirani, and Chu, 2001) employ such strategies.

The focus of the algorithms in the above-mentioned microarray analysis packages is the moderation of the biological variability of genes. In contrast, this section attempts to improve on the estimation of technical variability of the ratios from mass spectrometry data. One difference between microarray and mass spectrometry data is that the number of replicate spots per probe is usually the same for microarrays. Affymetrix GeneChip arrays, for example, have 11-20 pairs of spots for each 25-mer probe (Irizarry et al., 2003a), while the Agilent Microarray Platform typically have 10 replicate spots of the 60-mer probes (Zahurak et al., 2007). Normalization and summarization methods such as RMA (Irizarry et al., 2003b; Irizarry et al., 2003a) and MAS (Hubbell, W.-M. Liu, and Mei, 2002) combine the information to one reading per gene per array. After variance stabilization (or another type of normalization), each gene should have roughly equal variances.

In contrast, proteins are identified with one, some, or many spectra in mass spectrometry experiments. Thus, the uncertainty on the technical level is higher. We anticipated that therefore the moderation of the technical variance estimates would give an improvement for the inference of statistically different ratios.

The following paragraphs develop a simple hierachical Bayesian model, assuming a normal likelihood with unknown mean and variance. We take our prior belief regarding the distribution of probable variance values from the data. For each individual protein, its observed values are combined with the prior belief, which results in posterior distributions for the parameters. We first develop the Bayesian model without regard to the intensity-dependent variability of the spectra ratios themselves. These calculations lead to an analytical solution of the posterior, which has the same distributional family as the prior, with updated parameters. The model is first developed generally assuming the same weight for all spectrum ratios, and is later adapted to integrate the estimates from the noise model from section 4.1.

The results show that the posterior means are regressed towards the prior mean inversely dependent on the number of spectra observed. For proteins with many spectra, the data dominates the prior, and the regression is minimal. However, occasional extremes in mean or variability, which can appear by chance when few spectra are available, are absorbed. The marginal posterior on the mean follows a t-distribution. As the contribution of the prior belief distribution can be seen as pseudo-observations, the t-test is correspondingly also available when only one spectra is available. We demonstrate that this empirical Bayes estimation - especially the modified weighted version, which integrates intensity-based variance estimates - leads to greatly improved estimators with a higher number of true positive protein selections on the test dataset, while the false positive rate is well controlled.

Model specification

We consider the spectrum ratios (our data) are distributed normally, which is in line with the previous results. Now, however, we explicitly model the prior distributions of the unknown protein mean and variance. Note that at this point, we assume that the variance of the ratios is the same, independent of the signal intensity. Only at a later point we re-introduce the variance estimates of the spectra ratios, which we get from the noise model.

So the likelihood of the data, given the parameters, is normal. The natural conjugate prior distribution for an unknown mean and variance with a normal likelihood is a Normal-Inverse Gamma distribution (Baldi and Long, 2001). For the ease of modeling, it is common to parametrize the Normal distribution with precision instead of variance, where the precision is the inverse of the variance. We thus use a Normal-Gamma prior for the unknown mean and precision. Notably, the Normal-Gamma model is in line with the observation of heavy-tails of

the distribution of protein means: The marginal distribution of the mean of a Normal-Gamma is a Student's T distribution (Bishop, 2006; Gelman et al., 2003). As previously, all ratios are considered to be log-transformed.

Likelihood We consider one protein at a time for the likelihood function. We assume that the protein's spectrum-level ratios $\boldsymbol{X} = X_1, \ldots, X_n$, are independent and identically normally distributed, with the unknown parameters mean μ and precision τ. The likelihood function of the data given the parameters μ and τ is

$$\begin{aligned}\Pr(\boldsymbol{X} \mid \mu, \tau) &= \prod_{i=1}^{n} \mathcal{N}(X_i \mid \mu, \tau^{-1}) \\ &= \prod_{i=1}^{n} \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(X_i - \mu)^2\right) \\ &= \frac{\tau^{n/2}}{(2\pi)^{n/2}} \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right)\end{aligned}$$

Model prior We define a Gamma prior distribution on the unknown precision τ, and a Normal prior distribution on the mean μ, whose precision depends on τ:

$$\Pr(\tau) = \text{Gamma}(\tau \mid \alpha_0, \beta_0) \quad \text{where } \text{Gamma}(\tau \mid \alpha_0, \beta_0) \stackrel{\text{def}}{=} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau^{\alpha_0 - 1} \exp(\beta_0 \tau)$$

$$\Pr(\mu \mid \tau) = \mathcal{N}(\mu \mid \mu_0, (\kappa_0 \tau)^{-1}) \quad \text{where } \mathcal{N}(\mu \mid \mu_0, (\kappa_0\tau)^{-1}) \stackrel{\text{def}}{=} \sqrt{\frac{\kappa_0 \tau}{2\pi}} \exp\left(-\frac{\kappa_0 \tau}{2}(\mu - \mu_0)^2\right)$$

α_0 and β_0 are the shape and rate parameters of the Gamma distribution on τ, and μ is distributed normally with mean μ_0 and a precision $\kappa_0\tau$. The joint prior distribution is the product of the prior on μ given τ, and the prior on τ

$$\Pr(\mu, \tau) = \Pr(\mu \mid \tau) \times \Pr(\tau), \tag{4.3.1}$$

which leads to the Normal-Gamma distribution which is defined as follows:

$$\text{NG}(\mu, \tau \mid \mu_0, \kappa_0, \alpha_0, \beta_0) \stackrel{\text{def}}{=} \mathcal{N}(\mu \mid \mu_0, (\kappa_0\tau)^{-1}) \times \text{Gamma}(\tau \mid \alpha_0, \beta_0) \tag{4.3.2}$$

$$= \sqrt{\frac{\kappa_0}{2\pi}} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \tau^{\alpha_0 - \frac{1}{2}} \exp\left(-\frac{\tau}{2}(\kappa_0(\mu - \mu_0)^2 + 2\beta_0)\right) \tag{4.3.3}$$

The shape parameter α_0 and rate parameter β_0 for the Gamma prior are estimated based on the distribution of protein sample variances. The μ_0 is typically zero, as we consider ratios in the log-space, and the null hypothesis states that there is no change. However, in the case the data distribution is shifted from zero, it is advisable to estimate μ_0 based on the sample mean

or mode. κ_0 is the only parameter which has to be set disregarding the data, and it controls the weight given to the prior compared to the data. The default is $\kappa_0 = 2$, which says that the prior belief about μ is considered as much as two observations of the data (Gelman et al., 2003, pg. 81). We believe that this is a sensible choice, however it is possible to set other values. Figure 4.8 provides a comparison of the posterior means with different values of κ_0.

Model posterior and posterior marginals According to the Bayes' theorem, the posterior is proportional to the prior times the likelihood. Thus the posterior is proportional to the joint distribution of normal likelihood and the Normal-Gamma prior:

$$\Pr(\mu, \tau \mid \boldsymbol{X}) \propto \Pr(\boldsymbol{X} \mid \mu, \tau) \times \Pr(\mu \mid \tau) \times \Pr(\tau), \tag{4.3.4}$$

which follows a Normal-Gamma distribution with updated parameters (for the derivation see e. g. Kadane (2011)):

$$\Pr(\mu, \tau \mid \boldsymbol{X}) = \mathrm{NG}(\mu, \tau \mid \mu_n, \kappa_n, \alpha_n, \beta_n), \tag{4.3.5}$$

where

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{X}}{\kappa_0 + n} \tag{4.3.6}$$

$$\kappa_n = \kappa_0 + n \tag{4.3.7}$$

$$\alpha_n = \alpha_0 + \frac{n}{2} \tag{4.3.8}$$

$$\beta_n = \beta_0 + \tfrac{1}{2} SS + \frac{n\kappa_0}{2(n + \kappa_0)} \left(\mu_0 - \bar{X}\right)^2, \tag{4.3.9}$$

and $SS = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ is the sum of squared differences from the sample mean. For inference, the posterior marginals of μ and τ are of interest. The marginal posteriors are obtained by integrating the posterior distribution over the other parameter.

$$\Pr(\tau \mid \boldsymbol{X}) = \int_{-\infty}^{\infty} \Pr(\mu, \tau \mid \boldsymbol{X}) \, \mathrm{d}\mu = Gamma(\tau \mid \alpha_n, \beta_n) \tag{4.3.10}$$

$$\Pr(\mu \mid \boldsymbol{X}) = \int_{0}^{\infty} \Pr(\mu, \tau \mid \boldsymbol{X}) \, \mathrm{d}\tau = T_{2\alpha_n}(\mu \mid \mu_n, \beta_n / (\alpha_n \kappa_n)) \tag{4.3.11}$$

Integrating variance estimates from noise model In section 4.1 we developed a variance function that captures the relationship of signal intensity and variability. We demonstrated that the weighted mean - where the weights are the inverse of the estimated variance - are more accurate and precise.

The model thus far ignores the intensity-dependent variability. Here we modify the posterior parameters which use the data:

$$\mu_n^* = \frac{\kappa_0 \mu_0 + n\bar{X}^*}{\kappa_0 + n} \tag{4.3.12}$$

$$\beta_n^* = \beta_0 + \tfrac{1}{2} WSS + \frac{n\kappa_0}{2(n+\kappa_0)} \left(\mu_0 - \bar{X}\right)^2, \tag{4.3.13}$$

where

$$\bar{X}^* = \frac{1}{\sum_{i=1}^n W_i} \sum_{i=1}^n W_i X_i \tag{4.3.14}$$

$$WSS = \frac{n}{\sum_{i=1}^n w_i} \sum_{i=1}^n \left(X_i - \bar{X}^*\right)^2 \tag{4.3.15}$$

$$\boldsymbol{W} = W_1, \ldots, W_n = \frac{1}{\mathrm{Var}(X_1)}, \ldots, \frac{1}{\mathrm{Var}(X_n)} \tag{4.3.16}$$

and $\mathrm{Var}(X_i)$ is the variance of X_i as estimated by the noise model. Thus, the posterior uses the weighted sample mean $\bar{X}^*$ and weighted sum of squares. We compare the weighted and unweighted form of the posteriors for $alpha$ and β in the results, which demonstrates its superiority. The weighted version thus is proposed as default method.

Inference In line with the model developed in section 4.1, we use two cutoffs to assess the statistical significance at a level of α. First, the protein ratio has to be extreme enough in light of the background variability that is observed in biological replicates. For this, a Cauchy or generalized T-distribution (see section 4.2) are fitted on biological replicates. The probability of observing a ratio that extreme in biological replicates has to be less than α. Second, the null hypothesis, which states that there is 'no (technically) detectable change' of the protein at hand, has to be rejected, too. Here, the marginal posterior probability distribution of μ, $\Pr(\mu \mid \boldsymbol{X})$, which is t-distributed with $2\alpha_n$ degrees of freedom (see eq. (4.3.11)). If $\Pr(|\mu| > \mu_0 \mid \boldsymbol{X}) < \alpha/2$, thus if the standardized protein ratio is greater than the appropriate t quantile, the null hypothesis of no change is rejected:

$$\left| \frac{\mu_n}{\beta_n/(\alpha_n \kappa_n)} \right| > t_{2\alpha_n}(1 - \alpha/2), \tag{4.3.17}$$

where $|x|$ is the absolute value of x, and $t_{2\alpha_n}(1-\alpha/2)$ is the $1-\alpha/2$ quantile of a standard T-distribution with $2\alpha_n$ degrees of freedom.

Results and Discussion

Fit of Gamma distribution to sample variance We tested the fit of the Gamma distribution to the distribution of sample variances in the three samples described in Breitwieser et al. (2011). The variances are calculated as described in Breitwieser et al. (2011). The Gamma provides a good fit for all tested samples. The Kolmogorov-Smirnoff test fails to reject the null hypothesis that the data stems from the fitted distributions (p-values and more details are the fig. 4.7).

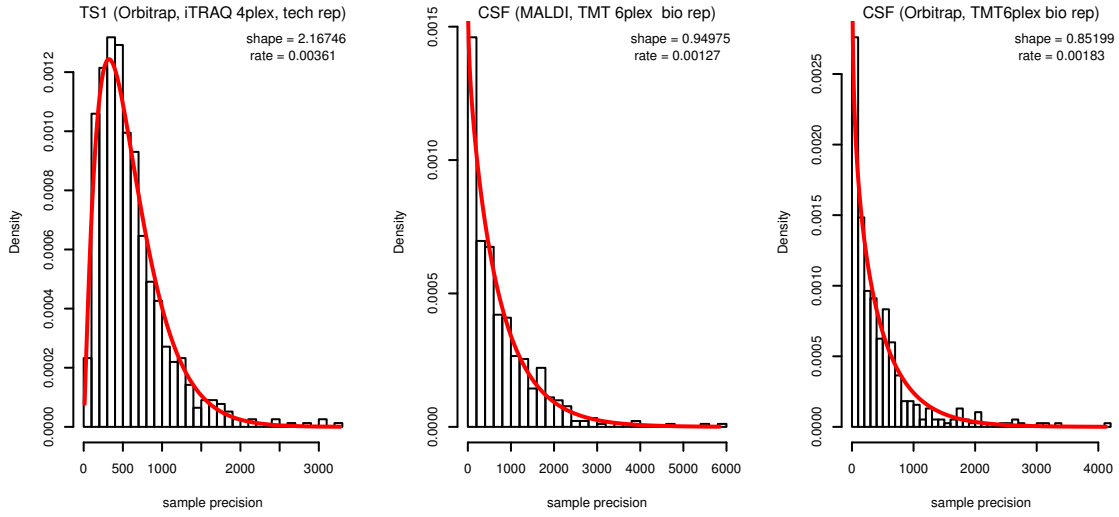


Figure 4.7: Fit of Gamma distribution to sample precision in three datasets. From left to right: Test dataset 1 from Breitwieser et al. (2011), analyzed with iTRAQ 4plex on Orbitrap (Kolmogorov-Smirnoff goodness of fit (KS) p-value: 0.422); Human cerospinal fluid of different individuals (CSF) analyzed with TMT 6-plex on MALDI-TOF/TOF (Tiberti et al. (2010), KS p-value: 0.160); CSF analyzed with TMT 6-plex on Orbitrap (KS p-value: 0.097).

Posterior parameters of the test dataset The posterior parameter distributions combine the prior knowledge and the data. μ_n can be interpreted as the average from κ_0 prior observations with mean μ_0 and n observations with mean $\bar{X}$ (eq. (4.3.6)). κ_0 is the only parameter which has to be set by the user. Figure 4.8 presents the effects on the posterior mean based on three different values of κ_0 in the test dataset (TS1). It can be seen that the posterior mean is regressed towards zero, with a magnitude relative to the number of spectra of the protein. Furthermore, increasing numbers of κ_0 increase the strength of the regression. We choose a default value for $\kappa_0 = 2$.

Figure 4.9 shows the sample and posterior parameter estimates of mean and variance in TS1 with $\kappa_0 = 2$. The variance estimates are shrunken towards the prior mean, too. However, certain

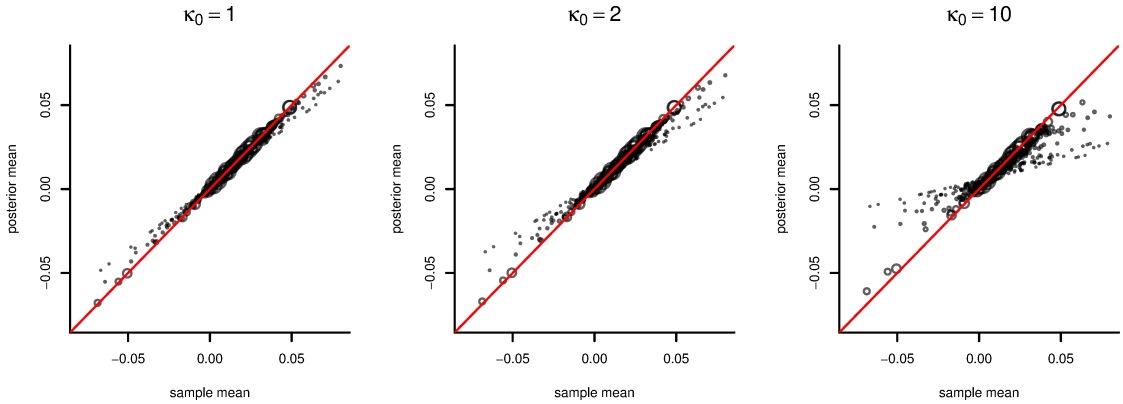


Figure 4.8: Scatterplot of sample versus posterior mean for test dataset 1 at $\kappa_0 = 1$, 2, and 10. Each point is one protein ratio, and the point size is relative to the square-root of its number of spectra. Prior variance parameters were fitted on the same dataset with κ_0 set to 2. Only ratios in the range of $\pm \log10(1.2)$ is shown for display purposes.

number of outliers are apparent, for which the effect is opposing. These result from proteins, which have a sample mean $\bar{X}$ at some distance from μ_0, as well as a high number of spectra n (see term for posterior sum of squares β_n in eq. (4.3.9)). In general, we can observe a regression towards the mean for both parameter estimates, which is as expected.

False positives and true positives We tested the performance of the hierachical model on the test dataset described in section 4.1. The parameters for the Gamma prior were $\alpha = 0.852$ and $\beta = 0.0018$, as estimated on the CSF dataset. We estimated the false positive rates on the background proteins. From each background protein between one and 20 spectra were randomly sampled. We tested the original isobar algorithm using the Cauchy and T distribution (columns 'Isobar Cauchy' and 'Isobar general. T', resp., see sections 4.1 and 4.2), a t-test, for which the degrees of freedom is equal to the number of spectra minus 1, and fold-change test, with a fold-change threshold at 1.5. Furthermore, the herein described empirical Bayes method was tested with weighted parameters (see 'Integrating variance estimates from the noise model') as well as unweighted parameters. For all the tested methods a significance threshold of $\alpha = 0.05$ was set (not applicable to the fold-change method). For each protein and number of spectra, the data was resampled 500 times. Table 4.1 demonstrates that the isobar methods control the false positive rate at the imposed significance level, while the t-test shows exceedingly high number of false positives with many spectra.

The estimation of the true positive-rates was conducted in the same way as above, however, the spiked proteins with known ratios were used instead of the background proteins. For each number of spectra, the data was resampled 5000 times. Table 4.2 demonstrates that both the

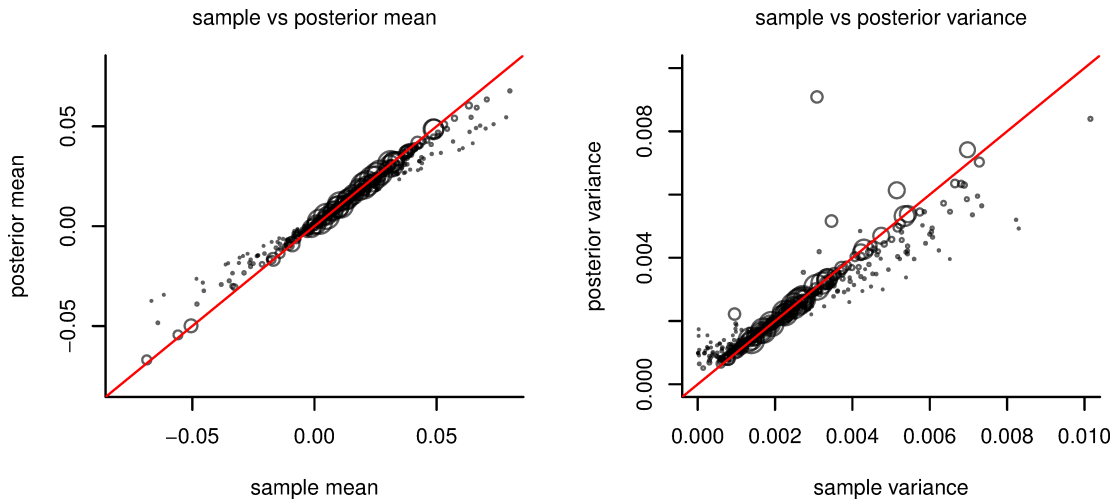


Figure 4.9: Scatterplot of sample versus posterior mean and sample versus posterior precision for test dataset 1. Each point is one protein ratio, and the point size is relative to the square-root of its number of spectra. Prior variance parameters were fitted on the same dataset with κ_0 set to 2. Only data for ratios in the range of $\pm \log10(1.2)$ is shown for display purposes.

weighted and unweighted versions of the 'eBayes' method are better than the originally proposed isobar method as well as the other methods. While at low number of spectra, the results are similar, at higher numbers the 'eBayes' methods manages to select up to twice the number of proteins with low fold changes.

Comparing the weighted eBayes method ('isobar eBayes') versus the unweighted ('isobar eBayes nw') shows that the former outperforms the later, and thus the value of the noise model.

Num spectra	Isobar Cauchy	Isobar general. T	T-test	Fold Change	Isobar eBayes	Isobar eBayes nw
1.00	0.02	0.02	0.00	0.07	0.03	0.03
2.00	0.01	0.01	0.06	0.05	0.03	0.04
3.00	0.02	0.03	0.06	0.03	0.02	0.04
5.00	0.02	0.05	0.17	0.01	0.01	0.02
10.00	0.00	0.00	0.17	0.00	0.00	0.01
15.00	0.00	0.00	0.21	0.00	0.00	0.00
20.00	0.00	0.00	0.27	0.00	0.00	0.00

Table 4.1: False positive rates observed on the test dataset with a 5% false positive threshold (resampled 500 times for each number of spectra).

Num spectra	isobar T original				T-test				Fold Change			
	r1.3	r1.5	r2	r2a	r1.3	r1.5	r2	r2a	r1.3	r1.5	r2	r2a
1.00	0.13	0.28	0.46	0.80	0.32	0.52	0.73	0.94	0.00	0.00	0.00	0.00
2.00	0.14	0.30	0.56	0.93	0.23	0.42	0.77	0.97	0.06	0.09	0.14	0.42
3.00	0.21	0.37	0.69	0.99	0.16	0.35	0.82	0.99	0.10	0.17	0.35	0.87
5.00	0.34	0.49	0.85	1.00	0.10	0.36	0.88	1.00	0.29	0.41	0.78	1.00
10.00	0.28	0.49	0.94	1.00	0.04	0.30	0.95	1.00	0.39	0.62	0.99	1.00
15.00	0.23	0.47	0.98	1.00	0.02	0.25	0.97	1.00	0.45	0.76	1.00	1.00
20.00	0.18	0.44	0.99	1.00	0.01	0.21	0.99	1.00	0.49	0.86	1.00	1.00

Num spectra	isobar eBayes				isobar eBayes nw			
	r1.3	r1.5	r2	r2a	r1.3	r1.5	r2	r2a
1.00	0.12	0.18	0.44	0.62	0.12	0.17	0.43	0.62
2.00	0.17	0.35	0.70	0.97	0.16	0.31	0.67	0.95
3.00	0.23	0.51	0.87	1.00	0.18	0.40	0.82	1.00
5.00	0.31	0.69	0.96	1.00	0.25	0.54	0.94	1.00
10.00	0.49	0.88	1.00	1.00	0.29	0.66	1.00	1.00
15.00	0.59	0.95	1.00	1.00	0.29	0.72	1.00	1.00
20.00	0.64	0.98	1.00	1.00	0.28	0.77	1.00	1.00

Table 4.2: True positive rates observed on the test dataset with a 5% false positive threshold at known ratios (data resampled 5000 times for each number of spectra). The expected ratios were 1.3, 1.5, and two times 2 with low ("r2") and high abundance ("r2a") of the spike-in material.

Conclusion

The technical variability of quantitative proteomics data is high, not only due to the intensity-dependent variability (which we presented in section 4.1), but also due to the varying number of spectra per protein. With a small number of spectra, the variance estimate for protein ratios imprecise in spite of the estimates of the variance, which we have for the individual spectra ratios. We thus developed a hierachical model for quantitative proteomics data which can moderate the protein means and variance, and help in the inference of significant changes.

The hierachical model was first developed assuming a common variance of the individual spectrum-level ratios of a protein in the data likelihood function. With this assumption, the Normal-Gamma prior is conjugate, which enables to derive an analytical solution for the posterior parameter distribution (Kadane, 2011). However, as mentioned above, the variability of spectrum ratios is varying based on signal intensity. We rationalized that the combination of the hierachical model with the estimates from the variance function would bring an improvement over using either alone. To integrate the noise model estimates of the variance, we modified the formulas of the posterior parameters to use the weighted mean and weighted sum of squares. As these are the only places in which the sample data influences the posterior parameters, we reasoned that this should give the desired results.

We tested the method on a test dataset as well as biological datasets. We could demonstrate that (a) the Gamma model for the precision fits the data well; (b) the model shrinks the estimates towards the prior mean, dependent on the number of spectra; (c) the empirical Bayes approach gives a better true positive rate than the previous method; (d) the modified weighted empirical Bayes method outperforms the unweighted one.

Notably, Schwacke et al. (2009) present a full Bayesian modeling for proteomics data down to the spectrum level intensities. Such a models require a powerful sampler, which enables to examine the parameter space. Schwacke et al. (2009) thus implement a Markov-Chain Monte Carlo Gibbs sampler. While these models can provide the most complete picture of the data, they can be too time or resource consuming for the analysis of actual datasets (Bielow, 2012). The model that is presented here is less comprising, but does not need sampling to infer the posterior parameters.

In conclusion, we have shown that the proposed hierachical modeling approach provides a significant improvement over the previous solution, especially when combined with the weighted estimates from the noise model. We belief that the combined model with the augmented sensitivity can help identifying additional biologically relevant, regulated proteins.

5. Discussion

This thesis presented novel computational and statistical approaches for the analysis of quantitative proteomics data. Figure 5.1 shows the various areas of the contributions, which are summarized in the following sections.

Isobaric labels are vital tools in the field of quantitative proteomics. We investigated the structure of technical and biological variability in isobarically labeled data, and devised statistical models for its analysis (section 4.1). These models were implemented in a novel R software package, which facilitates a complete workflow from mass spectrometry peak lists to the generation of reports of quantative protein differences (introduced in section 4.1). The package and methods were initially designed for protein-level analysis, but subsequently extended for the analysis on the level of post-translational protein modifications (section 4.2). The applicability of the methods has been demonstrated in further publications, in which the package was used and extended in the analysis of biological data sets (section 4.3).

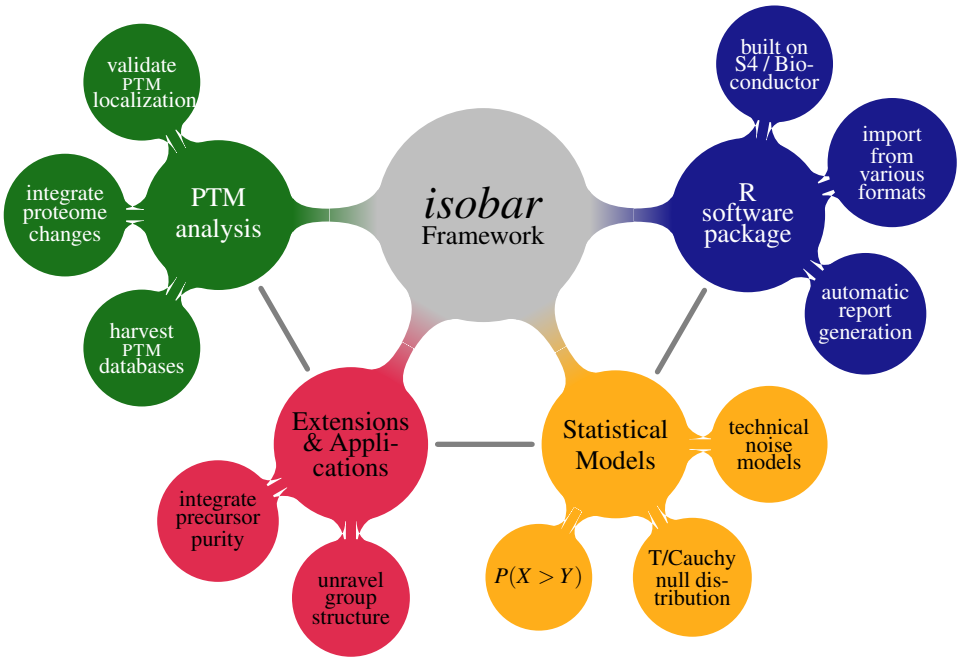


Figure 5.1: Main components of the framework for protein and PTM quantification presented in this thesis. The R software package `isobar` implements an analysis pipeline based on novel statistical models for the quantification of isobarically tagged data. Specific modules for the quantification of PTM data are integrated, as well as extensions for protein group structure elucidation, hybrid HCD-CID experiments, and label-free quantification.

5.1. Bioinformatics package for quantitative proteomics

The combining theme of the work presented in this thesis is the `isobar` software package (Breitwieser et al., 2011; Breitwieser and Colinge, 2013; Breitwieser and Colinge, n.d.). It was introduced to provide a computational framework for the quantification of proteomics experiments (Cappadona et al., 2012), and an implementation of the statistical methods described in section 4.1. Two main goals for the package were to (a) serve as an interface to investigate quantitative proteomics data in R, through the implementation of respective data representation classes, methods and functions, and (b) provide functionality to generate, with minimal efforts, quality-control and analysis reports, which can serve as results to be handed on, and used for down-stream analysis.

`isobar` is an open source R package (provided under the L-GPL version 2 license) which implements methods for a complete analysis pipeline from identification results to quantification reports. It can be automatized and run in a scripted environment, or used interactively for data exploration. It has been integrated into the Easyprot, a proteomics analysis software which is developed and employed at the University of Geneva (Gluck et al., 2013). `isobar` can be downloaded from its central development repository at `http://github.com/fbreitwieser/isobar` and, as part of the Bioconductor project, via the project's installer.

The main classes and methods of the package are implemented in the S4, R's object oriented class system (J. Chambers, 2008). The S4 containers for qualitative and quantitative information extend from base Bioconductor classes. One further package for iTRAQ/TMT quantification is available in Bioconductor: The `MSnbase` package, described in Gatto and Lilley (2012). `MSnbase` (Gatto, 2013; Gatto and Lilley, 2012) provides a similar set of base classes and functionality for capturing the quantitative information. Most notably, it also includes the import of RAW files using the `mzR` package (M. C. Chambers et al., 2012), the quantification based on profile-mode spectra, and handling MS^E data independent acquisition data through `synapter` (Nicholas J Bond et al., 2013; Gatto, Nick J Bond, and Shliaha, 2014). While each package has unique features (protein grouping, PTM modules, statistical analysis and report generation for `isobar`), it would be advantageous to consolidate the packages to provide better interoperability. `isobar` features conversion methods to `MSnbase` classes, but information is lost in the process. Ideally, for a future version of `isobar`, its code is refactored to base the data representation classes on `MSnbase`.

As described in the package vignette (see appendix A), the package supports the import from various file formats (appendix A.2), performs standard tasks such as isotope impurity correction and normalization (sections 2.3.2 and 4.1 and appendix A.3.2), and integration of and thresholding by of precursor purity measures (see sections 2.3.2 and 4.3). Furthermore, the package provides the feature of combining the results of multiple search engines, which is to our

knowledge currently not supported by other packages in the R environment. Also, experiments with multiple levels of analysis, such as CID/HCD can be imported (appendix A.2.3). Protein grouping is inferred on the fly, and it considers indistinguishable proteins and peptide specificity. Based on the grouping, `isobar` gives the option to quantify using peptides which are specific to the protein, or consider also peptides which are shared with splice variants which have no specific peptides themselves (Breitwieser and Colinge, n.d.). We demonstrated the possibility to further use the information to infer quantitative differences - and thus the presence - of proteins which are only seen with shared peptides (section 4.1). Furthermore, the grouping objects could be used to implement a better strategy for protein quantification exploiting the quantitative information of shared peptides, as proposed by Blein-Nicolas et al. (2012), Dost et al. (2012), and Gerster et al. (2013), and explicitly model shared peptides to improve quantification. Dost et al. (2012) showed that the shared peptides can also be used to quantify the relative abundances of different proteins across a family - a very interesting finding, as up to 50% of peptides identified in a typical experiments may be shared, when considering splice variants.

To allow the automated use of the software in a pipeline, as well as by inexperienced users, we developed methods and scripts such that, from a small text definition file, the package performs all necessary data analysis steps. PDF and Excel reports are generated, which provide quality control information and quantification results. To our knowledge, `isobar` is the sole R package which provides quality control reports for the quantification of isobarically tagged data. The PDF report - generated using Sweave (Leisch, 2002), LaTeX(Knuth, 1979) and TikZ (Tantau, 2013) - displays the results in tabular form, and features a box-and-whiskers type plot to represent and juxtapose the extend and precision of protein ratios (see figure 7 in section 4.1). Furthermore, the protein grouping is graphical presented. The Excel format, on the other hand, allows more flexibility for the user in displaying, filtering and selecting the data (see figure 3 in section 4.2). It contains the relevant spectrum-level information - such as the individual reporter intensities and search engines scores - and integrates PTM or protein information from external databases.

One design principle in the implementation of the `isobar` was to be agnostic of the specific methods used. For example, the `vsn` package, which implements variance stabilizing transformations (Huber et al., 2002; Karp et al., 2010), can be used, as well as z-score measures as protein fold-change cutoffs. The noise model may be fitted with any variance function, not necessarily exponential as we proposed. By using the `distr` package, the biological ratio distribution may be of any implementation of the `Distribution` class, such as Cauchy, T, or Gaussian.

5.2. Modeling of isobarically tagged proteomics data

To investigate technical variability, ratio estimation and distribution, and shared peptides, we designed two test datasets. The test dataset had a semi-complex background from depleted

human plasma and we spiked proteins at known concentrations. Compared to test-datasets with only spiked proteins and no background Boehm et al. (2007), Rodríguez-Suárez et al. (2010), and Hill et al. (2008, e. g.), the use of background proteins allows to estimate - at least to some extent - the effects of coelution and ratio distortion. We selected spike proteins which were mixed in concentrations of $1:2:5:10$ and $1:10:50:100$ into the background. The samples were differentially labeled with iTRAQ 4-plex reagents, and analyzed on an Orbitrap Velos Mass spectrometer (2D shotgun approach with forty offline fractions, HCD fragmentation, see section 4.1). The spiked proteins - ceruplasmins of the species human, rat and mouse - were selected because of their sequence similarity, which lead to shared tryptic peptides. We designed the test dataset to be able to explore, additionally to the standard properties of isobaric ratios, the quantitative properties of shared peptides. To our knowledge, this is the first isobarically tagged test dataset with this goal in mind.

In the analysis, we additionally used biological datasets from collaborators in Vienna and Geneva.

5.2.1. Technical noise model

In the first part of the investigation, we assessed the technical variability of reporter ion measurements using the test data set background, which was at $1:1$ concentration. In accordance with previous reports, we observed a strong dependence of the deviation of reporter intensity ratios from the true value on the signal intensities (Bantscheff et al., 2008; Hundertmark et al., 2009; Karp et al., 2010). Homogeneity of variance is an assumption in many statistical tests, such as the t-test and ANOVA.

Variance stabilizing transformations and noise models have been used for dealing with heteroskedacity in microarray data analysis, and were proposed for isobarically tagged data by Hundertmark et al. (2009), Yi Zhang et al. (2010), and Karp et al. (2010). We implemented the strategy of noise models, which capture the signal intensity dependent variance as exponential function of the log-transformed intensity. While Hundertmark et al. (2009) used a small number of synthetic peptides for estimating the noise model, we propose and employ, in line with Yi Zhang et al. (2010), the use of full tryptic digests of labeled technical replicates of cell lysates, which provide thousands of points for the estimation. Yi Zhang et al. (2010) demonstrate that correcting for ion injection time can provide a better fit without a constant term in the error model function.

The standard model we proposed uses an exponential function and includes a constant term (Breitwieser et al., 2011). The methods were designed to be used with processed peak lists, where the information on ion injection time is usually not available. However, within the software package (discussed in the next section), classes for noise models without intercept,

which are more appropriate for ion injection time corrected intensities, are provided (Yi Zhang et al., 2010).

Mandel et al. (2013) criticize the use of maximum-likelihood approach for parameter estimation by Hundertmark et al. (2009), as their model includes the individual ratios as nuisance parameters. This leads to a biased estimator, as shown by Neyman and Scott (1948). We employ maximum likelihood estimation, too, but by using technical replicates, we can avoid this issue, as the true ratios are known (zero, on the logarithmic scale). The function parameters are fitted using the R function `nlminb`, which does constrained quasi-newton optimization, and it fits $100,000$s of spectra in few seconds.

We further extended the methods to allow the fit of a noise model also from non one-to-one data. We normalize the protein ratios and were able to fit MALDI-TOF/TOF and LTQ-Orbitrap data from setups, for which no technical replicate data was available (Breitwieser et al., 2011). As the protein ratios are normalized and not used as nuisance parameters, the above-mentioned issue should not affect the parameter estimation for the noise model. On the test dataset, the estimated noise models based on non one-to-one data showed no strong difference to the one estimated on the background proteins (see fig. S1).

In conclusion of this point, we demonstrated in coherence with Hundertmark et al. (2009) and Yi Zhang et al. (2010), that noise models can capture signal intensity-dependent technical variability of isobarically tagged data, provide confidence intervals for measured ratios, and be learned once on technical replicates and then be used for further experiments. Furthermore, we provided implementations for the fitting of noise models, also on non $1:1$ data. The noise functions are encapsulated within S4 class representations, and thus can easily be exchanged.

5.2.2. Protein ratio calculation

Using the variance function value, we are able to compute averages, inversely weighted by the variance estimates, as estimators of the protein ratio mean. We demonstrated that, based on the test dataset, the weighted average provided the best compromise in accuracy and precision compared to other estimators. We do not model the peptide level, but summarize directly from spectrum to protein ratios, as we did not observe peptide-specific effects. Furthermore, we did not observe ratio suppression due to limited dynamic range, which has been described by Lin et al. (2006) and Rodríguez-Suárez et al. (2010) on data acquired on Q-TOF machines. To a certain extent, the use of weighted average and outlier removal may counter the effect. However, if there is a strong effect due to limited dynamic range in a data set the use of Multi-Q (Lin et al., 2006) or VEMS (Rodríguez-Suárez et al., 2010) method may give better results than our methods. By default, we exclude outliers (according to the "boxplot method" (Tukey, 1977)),

since we observed individual ratio deviations which might be due to coeluting peptides. As with many of the described choices, this option can be overridden by the user.

The default peptide set used for quantification includes just "reporter-specific" peptides (see section 2.3). However, it is possible to use group specific peptides, too. It is select-able, if only group-specific peptides from splice variants should be used, or all possible group-specific peptides.

5.2.3. Heavy-tailed protein ratio distribution and biological background variability

The models developed in this thesis are designed for pairwise comparisons, which is the simplest and most wide-spread experimental design of quantitative proteomics experiments (Karp et al., 2010). The reason for the choice of this simple design is the time and cost involved in proteomics experiments. The focus of the experiments is, thus, often on the hypothesis generation using few samples (which are then seen as representative of the population), with the biological validation done afterwards.

The selection of interesting proteins in such experiments is commonly performed using a z-score or robust z-score approach (for example in MaxQuant Jürgen Cox and Mann (2008) and IsobariQ (Arntzen et al., 2011)). These approaches standardize the protein ratios with the global (population) mean ratio μ and standard deviation σ^2 (see section 2.3.7). The standardized score is compared to standard normal distribution quantiles to calculate the probability of observing a ratio that extreme by chance. Jürgen Cox and Mann (2008, Supplementary Information) remarked that the normality is a reasonable assumption, since, in the limit of a large number of peptides per protein, the distribution of the (peptide) averages converges to a Normal according to the central limit theorem (CLT).

The CLT states that the average of many samples, which are drawn independently from some distribution with mean μ and variance σ^2, follows in the limit a normal distribution with mean μ and variance σ^2. In actual data, however, the assumption of a common variance will fail. Each protein mean has a different variance due to the effects of differences in signal intensity and number of spectra/peptides, and thus a more appropriate limiting model might be one of samples with the same mean, but different variances. It can be shown that the limiting distribution for this model is a generalized Student's T distribution (with parameters degrees of freedom, location, and scale) (Bishop, 2006, pg. 103).

We reported in sections 4.1 and 4.2 an improved fit of the Cauchy and generalized T distributions on replicate ratios. In the light of the aforementioned result, this does not surprise. Thus, in calculating a p value for a protein ratio using the fit of a T distribution, we will provide a more

precise measure of its probability to be a random outlier (notably, the Cauchy is the special case of a T distribution with one degree of freedom). The z-score will generate a too small p value for larger ratios due to the light tails of the normal distribution.

5.2.4. Technical ratio significance and hierarchical modeling

As mentioned above, the method presented in this thesis improves over the z-score by using the fit of a T or Cauchy distribution on the data, instead of assuming a normal. A further improvement is the consideration of technical variability of protein ratios, for which differences arise because of the different number of spectra and the precision of the spectrum-level ratios (see section 5.2.1). We demonstrated the heterogeneity of variance in spectrum ratios. The standard z-score, however, does not consider differences of different proteins. To incorporate intensity-dependent differences in variability, Jürgen Cox and Mann (2008) propose a modified version of the z-score, where the proteins are binned based on their intensity. However, this can capture the differences in technical variability only roughly. The proposed bin size is 300, which is more than the number of proteins observed in many pulldown experiments (Fernbach et al., 2009).

The method presented in this thesis calculates a ratio p value for a second level of inference, which eliminates outlying protein ratios which have a low precision. We assume the individual protein means are normally distributed, and exclude those whose precision is not sufficient to state with a specified level of confidence that the protein is different from zero.

Notably, when variance-stabilizing transformations are used, the spectrum-level ratios have roughly equal variance after the transformation (Karp et al., 2010; Schwacke et al., 2009). However, the standard errors of the sample means will still be different, as it incorporate the square root of the number of spectra.

The hierarchical model presented in section 4.3.2 is an unpublished extension to the previously developed methods. We could demonstrate a great improvement in selecting true positive proteins using the posterior marginal distribution of the protein means for inference. The model is simple and similar to Baldi and Long (2001), however, the modeled data are ratios and not the intensities. Furthermore the scope of the models presented in this thesis is just a single experiment to improve the estimation of technical variability.

5.2.5. Assessment of true/false positive rates

As summarized in the last two sections, to call protein abundance changes significant, it is required (1) that the signals are strong enough , and (2) the change is extreme enough in the light of the biological biological variability.

We determined the performance of our methodology in terms of specificity and sensitivity by resampling from the test data set. In MS proteomics data, interesting proteins are frequently identified just with few spectra. Therefore, we assessed the performance separately for defined number of spectra available $(1, 2, \ldots, 15)$ using resampling methodology on our test dataset. The true positive rate (sensitivity) was estimated on spiked proteins (with known ratios) at different fold changes. The false positive rate (specificity) was calculated on the background proteins.

We demonstrated that the combined method can provide better control of the false positive rate across the different number of spectra than those methods which were compared. Furthermore, the sensitivity ranks well compared to the others.

5.3. Quantification pipeline for post-translational modifications

Section 4.2 extended the software package and methodologies, which we have developed for protein-centric quantification, for datasets focusing on post-translational modifications of proteins. While many tools developed for quantitative proteomics are also applicable for the quantification on the peptide level (Allmer, 2012), few software packages were explicitly designed and tested for the quantification of PTMs. We had three properties in mind: (1) validation of modification localization; (2) correction of modification ratios with protein ratios; (3) integration of public PTM databases in a navigable spreadsheet user report.

First the statistical models and software package, which were introduced for the protein level analysis in section 4.1, were adapted for the analysis of PTM data. We reassessed the fit of the model on peptide-level data and the performance in the selection of significant proteins at varying number of spectra. The peptide-level results were consistent with the protein-level, and controlled the false positive rate at the imposed level while providing good sensitivity at the varying number of spectra.

The validation of PTM site localization is an important step in the PTM data analysis (Chalkley and Clauser, 2012). Often separate tools are employed to assess the reliability of site-localizations after the peptide-spectrum matching by the search engine. We developed a general interface to `isobar` to filter identifications based on scores of site localization tools. The Mascot Delta score (Mikhail M. Savitski et al., 2011a), which is also applicable to other search engines (Chalkley and Clauser, 2012), was directly implemented into the tool. For PhosphoRS, a probabilistic site localization tool (Taus et al., 2011), wrapper functions that call the tool and process the output into a suitable form are provided.

Wu et al. (2011) reported that the observed changes on the abundance of modified peptides integrates both the abundance change on the protein level as well as the abundance change on the modification state level. In many research applications it is desired to separate the

effects, which, however, requires an additional experiment in which the changes in protein levels are measured without the PTM enrichment step. We implemented a framework in `isobar` to use the quantification results of a protein-level experiment to adjust the peptide-level ratios. Furthermore we reported that the resulting variance of the adjusted ratio is the sum of the variances of the peptide-level and protein-level ratios plus a further covariance term. The covariance is unknown, but the term is bounded by assuming either independence or perfect correlation of the peptide- and protein-level ratios. We reasoned that a sensible choice for setting the covariance uses the correlation coefficient of the data (times the standard deviations of the respective ratios). On average, this should give the best estimation of the covariance. Having an updated ratio mean and variance, we can calculate the significance of the ratio as previously. To our knowledge, no other software tool enables the direct integration of the protein and peptide abundance.

The analysis reports in Excel formats then report all information; on the modified peptide ratio, the protein ratio, and the adjusted modification state ratio. Furthermore, the public databases NeXtProt and PhosphoSitePlus databases (Lane et al., 2012; Hornbeck et al., 2012) are harvested and the specific knowledge for experimentally observed modification sites is reported as well.

In general, we know of no other quantification tool which provides the mentioned breadth for PTM analysis, and we think this can provide an useful analysis workflow for researchers.

5.4. Directions of future research and development

5.4.1. Graphical User Interface (GUI)

One of the main features of the software package is report generation. The generation of the reports can be initiated within a R session, or by the definition of a text definition file, and a command line script call. Nearly every aspect of the quantification can be modified within the text definition file. In total, `isobar` allows the parametrization of 60 distinct properties, as of version 1.9. Even though normally only a small subset is used, and documentation is provided, the large number of parameters can present an obstacle to new users.

A graphical user interface (GUI) for the report generation could help to bring focus on the important set of parameters, and provide online help for its use. The properties may be grouped, put into submenus or tabs. Furthermore, a GUI can confine the input to valid or reasonable values. The whole process of report-generation can be self-explanatory and guided. Figure 5.2 shows a user interface prototype. Certainly, users would also benefit from a web version of `isobar` by not having to install R, the package and its dependencies themselves.

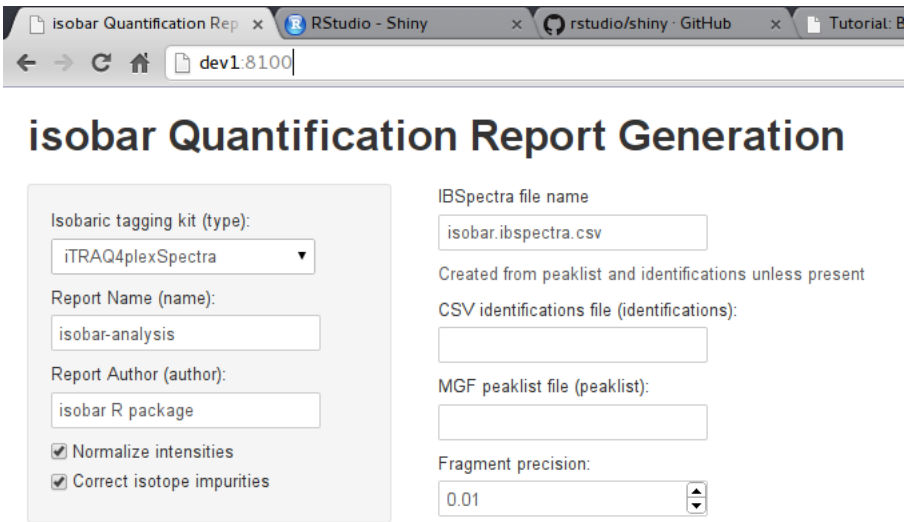


Figure 5.2: Prototype for web site for automated report generation using the Shiny server package (http://www.rstudio.com/shiny).

5.4.2. Support of MS1 and MS3-based quantification

The methods developed in this thesis have been specifically designed for and tested on isobarically tagged MS^2 quantification methods. However, the structure of technical variability should be similar for other mass spectrometry-based quantification techniques. Furthermore, the biological variability is expected to be heavy-tailed, too.

Quantification of isotopically-labeled samples can be done in MS^1, MS^2, or in extension, MS^n. Differences in pre-processing of MS^1-labeled data are explained in section 2.3.1. MS^3-based quantification, as proposed by Ting et al. (2011), is used for isobarically labeled samples as a method to remove the ratio compression effect introduced by co-eluting peptides. In a preliminary analysis of MS^3 data, the integration in the software package proved straight-forward using the methods initially developed for combining paired CID and HCD runs. MS^2-based quantification using isobaric peptide termini labeling (Koehler et al., 2011; Koehler et al., 2013) and pseudo-isobaric dimethyl labeling (Zhou et al., 2013; Bamberger et al., 2014) produce paired peptide fragment ions, instead of reporter ions, for quantification. This approach also does not suffer from co-elution in the same way as the iTRAQ and TMT reporter-ion based methods.

The `isobar` package, which was conceived as software for quantitation of (reporter ion-based) isobarically tagged data, was already extended with methods to allow label-free quantitation based on peptide or spectra count, specifically emPAI and dNSAF (Ishihama et al., 2005; Ying Zhang et al., 2010). `isobar` could be developed further to provide a platform for the aforementioned quantification techniques; labeled and label-free, MS^1, MS^2, and MS^n based.

5.4.3. Statistical Inference

The statistical methods presented in this thesis improve over z-score or fold-change methods for selecting interesting proteins, which enable to select proteins in pairwise comparisons. Such methods are widely used in quantitative proteomics for hypothesis generation and selection of interesting proteins (Karp et al., 2010; Jürgen Cox and Mann, 2008; Arntzen et al., 2011; P. P. Hsu et al., 2011). These approaches consider the sample representative of the population and are useful for generating hypothesis. However, when the experimental design includes multiple samples and a more complicated setup is used, more advanced statistical methods are required.

With the advances in instrumentation (Hebert et al., 2013b) and higher multiplexing capabilities of the labeling techniques (Hebert et al., 2013a; Werner et al., 2012), the use of better experimental designs for proteomics studies will become more standard. In the microarray field, the standardization of the arrays and relatively low cost of experiments have lead to clear guidelines on the use of replicates to provide sufficient statistical power (Shi et al., 2008) (in proteomics, sample sizes cannot be easily assessed, as each protein ratio comes with its own variability). Bioinformatics software packages like `limma` provide a well-established models and statistics for differential microarray expression analysis (Smyth, 2004).

`limma` has also been used in several publications for the analysis of proteomics data (Ting et al., 2009; Castello et al., 2012; Schwämmle, León, and Ole Nørregaard Jensen, 2013). With the ever-higher throughput of the mass spectrometers (Hebert et al., 2013b) as well as increased multiplexing options in labeling approaches (Hebert et al., 2013a; Werner et al., 2012), the use of more involved experimental designs is becoming more standard. We believe that, for the time being, the empirical Bayes models of `limma` can be useful for the analysis of such data sets. The protein (or modified peptide) log-transformed ratios and weights can be used as input. The calculation of protein ratios and their weights (i. e. inverse variance) still should be performed with full consideration of the peculiarities of the quantitative proteomics data. As this can be provided by `isobar`, the combination of `isobar` and `limma` can provide a working solution for more involved quantitative proteomics experiments.

5.5. Final conclusions

Proteomics is an extraordinary tool of biological research (Altelaar, Munoz, and Heck, 2013). Fueled by ever more powerful mass spectrometry, researchers can go deeper into the proteome than ever before (Hebert et al., 2013b; Olsen and Mann, 2013). To assess protein abundance changes across multiple samples, isotope labeling techniques are employed. The resulting

datasets are large and complex, and powerful bioinformatical software, and sound statistical models, are essential for their analysis (Mueller et al., 2008).

This thesis thus presented a novel software tool for the analysis of quantitative proteomics data, which integrates statistical models geared for isobaric tag quantification, as well as methods for the quantification of post-translational protein modifications (Breitwieser and Colinge, 2013; Breitwieser et al., 2011).

The `isobar` package integrates the steps of data processing, statistical analysis, and report generation for quantitative proteomics. The statistical methods underpinning the package capture technical and biological variability and allow defining p value thresholds for significant regulation. We exploit the technical variance structure to summarize protein ration using weighted average. The protein ratio null distribution, which represents the biological background variability, is modeled with a heavy tailed-modeled. Combining the noise model and protein ratio distribution for selection, we demonstrated better performance in terms of sensitivity and specificity for selecting significant proteins compared to fold change or t-test analysis.

Protein function and fate is controlled on various levels. Of prime importance are post-translational protein modifications, which can change the physico-chemical properties of the protein. The unprecedented proteome-scale mapping possibilities in the investigation and quantification of PTMs presents several challenges for the data analysis (Allmer, 2012). While tools developed for protein quantification can also be used or adapted for peptide quantification, the analysis is complicated by the several steps which have to be integrated for PTM analysis.

We hence developed and implemented modules for a PTM quantification workflow, based on our afore-mentioned methods. The statistical models were re-validated on the peptide level, and required little adaption. Several usual steps involved in PTM quantification were integrated and implemented: performing automatic validation of modification site localization using PhosphoRS or the delta score (Taus et al., 2011; Mikhail M. Savitski et al., 2011a); correction of peptide ratios with protein abundance changes (Wu et al., 2011); integration of public PTM site databases (Hornbeck et al., 2012; Lane et al., 2012).

The software was developed as part of the Bioconductor project. Download statistics from the Bioconductor website [1] demonstrate continued interest in the project which has been continuously downloaded about or over 100 times per month since it was added (see fig. 5.3). Furthermore, the project is in use in the quantitative proteomics pipeline at CeMM as well as at the University of Geneva.

In conclusion, this thesis has developed statistical models and bioinformatical software for the analysis of protein and PTM quantification data (Breitwieser et al., 2011; Breitwieser and Colinge, 2013), whose use and applicability were demonstrated in several publications (Borgdorff et al.,

[1] http://www.bioconductor.org

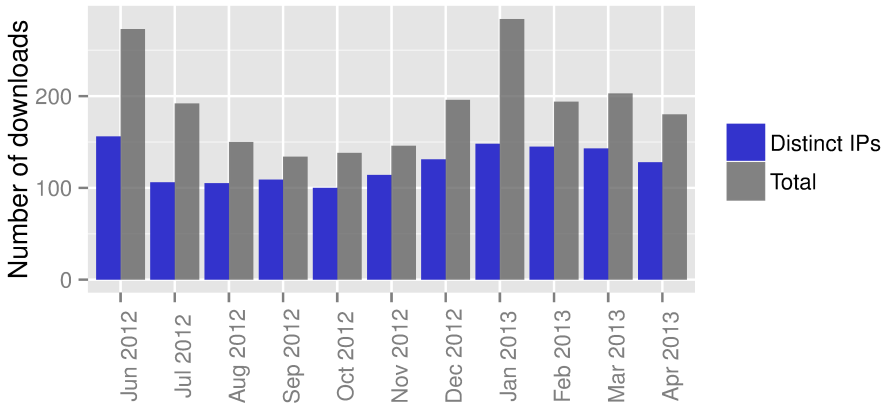


Figure 5.3: Number of downloads of the isobar R package via Bioconductor. Source: http://bioconductor.org/packages/stats/bioc/isobar.html

2013; Pollreisz et al., 2013; Rudashevskaya et al., 2013; Müller et al., 2012; Winter et al., 2012). The methods facilitate quantative proteomics analysis and are applicable a broad range of instruments and experiments. Due to its open nature and design, the software package can be easily extended or adapted. The support of novel tags such TMT 10-plex (McAlister et al., 2012) was achieved by adding appropriate class definitions. Using reporter ions from MS^3 (Ting et al., 2011), or correcting intensities for precursor impurities, using the method reported by Mikhail M Savitski et al. (2013), could be added without great effort.

6. Bibliography

Ahn, Natalie G., John B. Shabb, William M. Old, and Katheryn A. Resing (Jan. 2007). “Achieving in-depth proteomics profiling by mass spectrometry.” In: *ACS Chem Biol* 2.1, pp. 39–52 (cit. on p. 12).

Alikhanov, S. G. (1957). “A new impulse technique for ion mass measurements”. In: *Soviet Phys. JETP* 4 (cit. on p. 11).

Allmer, Jens (Jan. 2012). “Existing bioinformatics tools for the quantitation of post-translational modifications.” In: *Amino Acids* 42.1, pp. 129–138 (cit. on pp. 109, 113).

Altelaar, A F Maarten, Javier Munoz, and Albert J R. Heck (Jan. 2013). “Next-generation proteomics: towards an integrative view of proteome dynamics.” In: *Nat Rev Genet* 14.1, pp. 35–48 (cit. on pp. 1, 17, 42, 112).

Armirotti, Andrea and Gianluca Damonte (Oct. 2010). “Achievements and perspectives of top-down proteomics.” In: *Proteomics* 10.20, pp. 3566–3576 (cit. on p. 7).

Arntzen, Magnus O., Christian J. Koehler, Harald Barsnes, Frode S. Berven, Achim Treumann, and Bernd Thiede (Feb. 2011). “IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT.” In: *J Proteome Res* 10.2, pp. 913–920 (cit. on pp. 30, 32, 33, 39, 40, 46, 107, 112).

Arrigoni, Giorgio, Serena Tolin, Roberto Moscatiello, Antonio Masi, Lorella Navazio, and Andrea Squartini (Nov. 2013). “Calcium-dependent regulation of genes for plant nodulation in Rhizobium leguminosarum detected by iTRAQ quantitative proteomic analysis.” In: *J Proteome Res* 12.11, pp. 5323–5330 (cit. on p. 40).

Aston, Francis W. (1920). “The constitution of atmospheric neon”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 39.232, pp. 449–455 (cit. on p. 4).

Bailey, Christopher M., Steve M M. Sweet, Debbie L. Cunningham, Martin Zeller, John K. Heath, and Helen J. Cooper (Apr. 2009). “SLoMo: automated site localization of modifications from ETD/ECD mass spectra.” In: *J Proteome Res* 8.4, pp. 1965–1971 (cit. on p. 44).

Bailey, Derek J., Christopher M. Rose, Graeme C. McAlister, Justin Brumbaugh, Pengzhi Yu, Craig D. Wenger, Michael S. Westphall, James A. Thomson, and Joshua J. Coon (May 2012). “Instant spectral assignment for advanced decision tree-driven mass spectrometry.” In: *Proc Natl Acad Sci U S A* 109.22, pp. 8411–8416 (cit. on p. 17).

Baldi, Pierre and Anthony D Long (2001). “A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes”. In: *Bioinformatics* 17.6, pp. 509–519 (cit. on pp. 92, 93, 108).

Bamberger, Casimir, Sandra Pankow, Sung Kyu Robin Park, and John R Yates (2014). “Interference-free Proteome Quantification with MS/MS-based Isobaric Isotopologue Detection”. In: *Journal of Proteome Research* (cit. on p. 111).

Bantscheff, Marcus, Markus Boesche, Dirk Eberhard, Toby Matthieson, Gavain Sweetman, and Bernhard Kuster (Sept. 2008). "Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer." In: *Mol Cell Proteomics* 7.9, pp. 1702–1713 (cit. on pp. 34, 105).

Bantscheff, Marcus, Dirk Eberhard, Yann Abraham, Sonja Bastuck, Markus Boesche, Scott Hobson, Toby Mathieson, Jessica Perrin, Manfred Raida, Christina Rau, Valérie Reader, Gavain Sweetman, Andreas Bauer, Tewis Bouwmeester, Carsten Hopf, Ulrich Kruse, Gitte Neubauer, Nigel Ramsden, Jens Rick, Bernhard Kuster, and Gerard Drewes (Sept. 2007a). "Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors." In: *Nat Biotechnol* 25.9, pp. 1035–1044 (cit. on p. 34).

Bantscheff, Marcus, Carsten Hopf, Mikhail M. Savitski, Antje Dittmann, Paola Grandi, Anne-Marie Michon, Judith Schlegl, Yann Abraham, Isabelle Becher, Giovanna Bergamini, Markus Boesche, Manja Delling, Birgit Dümpelfeld, Dirk Eberhard, Carola Huthmacher, Toby Mathieson, Daniel Poeckel, Valérie Reader, Katja Strunk, Gavain Sweetman, Ulrich Kruse, Gitte Neubauer, Nigel G. Ramsden, and Gerard Drewes (Mar. 2011). "Chemoproteomics profiling of HDAC inhibitors reveals selective targeting of HDAC complexes." In: *Nat Biotechnol* 29.3, pp. 255–265 (cit. on p. 1).

Bantscheff, Marcus, Simone Lemeer, Mikhail M. Savitski, and Bernhard Kuster (Sept. 2012). "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present." In: *Anal Bioanal Chem* 404.4, pp. 939–965 (cit. on pp. 1, 2, 19, 24, 38, 48).

Bantscheff, Marcus, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster (Oct. 2007b). "Quantitative mass spectrometry in proteomics: a critical review." In: *Anal Bioanal Chem* 389.4, pp. 1017–1031 (cit. on p. 27).

Beausoleil, Sean A., Judit Villén, Scott A. Gerber, John Rush, and Steven P. Gygi (Oct. 2006). "A probability-based approach for high-throughput protein phosphorylation analysis and site localization." In: *Nat Biotechnol* 24.10, pp. 1285–1292 (cit. on p. 44).

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300 (cit. on p. 41).

Benjamini, Yoav and Daniel Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency". In: *Annals of Statistics*, pp. 1165–1188 (cit. on p. 41).

Bennett, Keiryn L., Marion Funk, Marion Tschernutter, Florian P. Breitwieser, Melanie Planyavsky, Ceereena Ubaida Mohien, André Müller, Zlatko Trajanoski, Jacques Colinge, Giulio Superti-Furga, and Ursula Schmidt-Erfurth (Feb. 2011). "Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ®-labelled specimens." In: *J Proteomics* 74.2, pp. 151–166 (cit. on pp. iii, iv, 86–88, 168).

Bielow, Chris (2012). "Quantification and Simulation of Liquid Chromatography-Mass Spectrometry Data". PhD thesis. Freie Universität Berlin (cit. on pp. 30, 35, 46, 47, 101).

Biemann, K., G. Gapp, and J. Seibl (1959). “Application of mass spectrometry to structure problems. I. Amino acid sequence in peptides”. In: *Journal of the American Chemical Society* 81.9, pp. 2274–2275 (cit. on p. 5).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. (cit. on pp. 94, 107).

Bjornson, Robert D., Nicholas J. Carriero, Christopher Colangelo, Mark Shifman, Kei-Hoi Cheung, Perry L. Miller, and Kenneth Williams (Jan. 2008). “X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers.” In: *J Proteome Res* 7.1, pp. 293–299 (cit. on p. 16).

Blein-Nicolas, Mélisande, Hao Xu, Dominique de Vienne, Christophe Giraud, Sylvie Huet, and Michel Zivy (Sept. 2012). “Including shared peptides for estimating protein abundances: a significant improvement for quantitative proteomics.” In: *Proteomics* 12.18, pp. 2797–2801 (cit. on pp. 38, 104).

Bock, Christoph (Oct. 2012). “Analysing and interpreting DNA methylation data.” In: *Nat Rev Genet* 13.10, pp. 705–719 (cit. on p. 1).

Boehm, Andreas M., Stephanie Pütz, Daniela Altenhöfer, Albert Sickmann, and Michael Falk (2007). “Precise protein quantification based on peptide quantification using iTRAQ.” In: *BMC Bioinformatics* 8, p. 214 (cit. on p. 105).

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (Jan. 2003). “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.” In: *Bioinformatics* 19.2, pp. 185–193 (cit. on p. 32).

Bond, Nicholas J, Pavel V Shliaha, Kathryn S Lilley, and Laurent Gatto (2013). “Improving qualitative and quantitative performance for MSe-based label-free proteomics”. In: *Journal of Proteome Research* 12.6, pp. 2340–2353 (cit. on p. 103).

Borgdorff, V., U. Rix, G. E. Winter, M. Gridling, A. C. Müller, F. P. Breitwieser, C. Wagner, J. Colinge, K. L. Bennett, G. Superti-Furga, and S. N. Wagner (June 2013). “A chemical biology approach identifies AMPK as a modulator of melanoma oncogene MITF.” In: *Oncogene* (cit. on pp. 2, 113, 167).

Boucheron, Nicole, Roland Tschismarov, Lisa Goeschl, Mirjam A. Moser, Sabine Lagger, Shinya Sakaguchi, Mircea Winter, Florian Lenz, Dijana Vitko, Florian P. Breitwieser, Lena Müller, Hammad Hassan, Keiryn L. Bennett, Jacques Colinge, Wolfgang Schreiner, Takeshi Egawa, Ichiro Taniuchi, Patrick Matthias, Christian Seiser, and Wilfried Ellmeier (Mar. 2014). “CD4(+) T cell lineage integrity is controlled by the histone deacetylases HDAC1 and HDAC2.” In: *Nat Immunol* (cit. on p. 167).

Breitwieser, Florian P. (2012). *Statistical Modeling of Post-translational Protein Regulation Dynamics*. Presented at the Young Investigators Day. Faculty of Computational Life Sciences, University of Vienna (cit. on p. 169).

Breitwieser, Florian P. (Feb. 2013). *isobar: Quantitative Analysis of Protein and PTM iTRAQ/TMT data*. Presented at IMBA Impromptu Seminar (invited talk). Vienna, Austria (cit. on p. 169).

Breitwieser, Florian P. and Jacques Colinge (n.d.). *isobar:Analysis and quantitation of isobarically tagged MSMS proteomics data* (cit. on pp. 103, 104).

– (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91 (cit. on pp. ii, iv, 17, 21, 22, 28, 35, 169).

– (2012b). *isobar: Quantifying changes of the proteome and its post-translational modifications*. Presented at the 9th Siena Meeting: From Genome to Proteome. Siena, Italy (cit. on p. 169).

– (2012c). isobar *R package for the analysis of quantitative proteomics data*. Presented at the 12th Annual Bioinformatics Open Source Conference. Vienna, Austria (cit. on p. 169).

– (Sept. 2013). "Isobar(PTM): a software tool for the quantitative analysis of post-translationally modified proteins." In: *J Proteomics* 90, pp. 77–84 (cit. on pp. ii–iv, 29, 46, 80–82, 103, 113, 167).

Breitwieser, Florian P., André Müller, Loïc Dayon, Thomas Köcher, Alexandre Hainard, Peter Pichler, Ursula Schmidt-Erfurth, Giulio Superti-Furga, Jean-Charles Sanchez, Karl Mechtler, Keiryn L. Bennett, and Jacques Colinge (June 2011). "General statistical modeling of data from protein relative expression isobaric tags." In: *J Proteome Res* 10.6, pp. 2758–2766 (cit. on pp. ii, iv, 31, 35, 46, 49, 80, 86, 97, 103, 105, 106, 113, 168).

Breitwieser, Florian P., André Müller, Giulio Superti-Furga, Keiryn L. Bennett, and Jacques Colinge (2010). *Statistical Models for Quantitative Proteomics using Isobaric Tags*. Presented at the 4th Central and Eastern European Proteomics Conference. Vienna, Austria (cit. on p. 170).

Brönstrup, Mark (Dec. 2004). "Absolute quantification strategies in proteomics based on mass spectrometry." In: *Expert Rev Proteomics* 1.4, pp. 503–512 (cit. on p. 18).

Burkard, Thomas R., Melanie Planyavsky, Ines Kaupe, Florian P. Breitwieser, Tilmann Bürckstümmer, Keiryn L. Bennett, Giulio Superti-Furga, and Jacques Colinge (2011). "Initial characterization of the human central proteome." In: *BMC Syst Biol* 5, p. 17 (cit. on p. 168).

Burkard, Thomas R., Uwe Rix, Florian P. Breitwieser, Giulio Superti-Furga, and Jacques Colinge (2010). "A computational approach to analyze the mechanism of action of the kinase inhibitor bafetinib." In: *PLoS Comput Biol* 6.11, e1001001 (cit. on p. 169).

Cappadona, Salvatore, Peter R. Baker, Pedro R. Cutillas, Albert J R. Heck, and Bas van Breukelen (Sept. 2012). "Current challenges in software solutions for mass spectrometry-based quantitative proteomics." In: *Amino Acids* 43.3, pp. 1087–1108 (cit. on pp. 2, 43, 48, 103).

Carrillo, Brian, Corey Yanofsky, Sylvie Laboissiere, Robert Nadon, and Robert E Kearney (Jan. 2010). "Methods for combining peptide intensities to estimate relative protein abundance." In: *Bioinformatics* 26.1, pp. 98–103 (cit. on p. 34).

Casadonte, Rita and Richard M. Caprioli (Nov. 2011). “Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry.” In: *Nat Protoc* 6.11, pp. 1695–1709 (cit. on p. 9).

Castello, Alfredo, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M. Beckmann, Claudia Strein, Norman E. Davey, David T. Humphreys, Thomas Preiss, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze (June 2012). “Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.” In: *Cell* 149.6, pp. 1393–1406 (cit. on p. 112).

Chalkley, Robert J. and Karl R. Clauser (Feb. 2012). “Modification site localization scoring: Strategies and performance.” In: *Mol Cell Proteomics* (cit. on pp. 2, 43, 44, 109).

Chambers, John (2008). *Software for data analysis: programming with R*. Springer (cit. on p. 103).

Chambers, Matthew C., Brendan Maclean, Robert Burke, Dario Amodei, Daniel L. Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A. Baker, Mi-Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L. Seymour, Lydia M. Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W. Deutsch, Robert L. Moritz, Jonathan E. Katz, David B. Agus, Michael MacCoss, David L. Tabb, and Parag Mallick (Oct. 2012). “A cross-platform toolkit for mass spectrometry and proteomics.” In: *Nat Biotechnol* 30.10, pp. 918–920 (cit. on pp. 46, 103).

Chi, An, Curtis Huttenhower, Lewis Y. Geer, Joshua J. Coon, John E P. Syka, Dina L. Bai, Jeffrey Shabanowitz, Daniel J. Burke, Olga G. Troyanskaya, and Donald F. Hunt (Feb. 2007). “Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry.” In: *Proc Natl Acad Sci U S A* 104.7, pp. 2193–2198 (cit. on p. 15).

Choe, Leila, Mark D’Ascenzo, Norman R. Relkin, Darryl Pappin, Philip Ross, Brian Williamson, Steven Guertin, Patrick Pribil, and Kelvin H. Lee (Oct. 2007). “8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer’s disease.” In: *Proteomics* 7.20, pp. 3651–3660 (cit. on p. 64).

Choudhary, Chunaram, Chanchal Kumar, Florian Gnad, Michael L. Nielsen, Michael Rehman, Tobias C. Walther, Jesper V. Olsen, and Matthias Mann (Aug. 2009). “Lysine acetylation targets protein complexes and co-regulates major cellular functions.” In: *Science* 325.5942, pp. 834–840 (cit. on p. 42).

Christoforou, Andy L. and Kathryn S. Lilley (Sept. 2012). “Isobaric tagging approaches in quantitative proteomics: the ups and downs.” In: *Anal Bioanal Chem* 404.4, pp. 1029–1037 (cit. on pp. 2, 24, 27, 29–31, 48).

Cirulli, Elizabeth T. and David B. Goldstein (June 2010). “Uncovering the roles of rare variants in common disease through whole-genome sequencing.” In: *Nat Rev Genet* 11.6, pp. 415–425 (cit. on p. 1).

Claassen, Manfred (Nov. 2012). “Inference and validation of protein identifications.” In: *Mol Cell Proteomics* 11.11, pp. 1097–1104 (cit. on pp. 1, 6, 17, 37).

Colinge, Jacques and Keiryn L Bennett (July 2007). “Introduction to computational proteomics.” In: *PLoS Comput Biol* 3.7, e114 (cit. on pp. 1, 2, 15, 16, 37).

Colinge, Jacques, Adrián César-Razquin, Kilian Huber, Florian P. Breitwieser, Peter Májek, and Giulio Superti-Furga (Apr. 2014). “Building and exploring an integrated human kinase network: Global organization and medical entry points.” In: *J Proteomics* (cit. on p. 167).

Colinge, Jacques, Alexandre Masselot, Marc Giron, Thierry Dessingy, and Jérôme Magnin (Aug. 2003). “OLAV: towards high-throughput tandem mass spectrometry data identification.” In: *Proteomics* 3.8, pp. 1454–1463 (cit. on p. 16).

Collins, Ben C., Ludovic C. Gillet, George Rosenberger, Hannes L. Röst, Anton Vichalkovski, Matthias Gstaiger, and Ruedi Aebersold (Dec. 2013). “Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system.” In: *Nat Methods* 10.12, pp. 1246–1253 (cit. on p. 13).

Comisarow, Melvin B and Alan G Marshall (1974a). “Fourier transform ion cyclotron resonance spectroscopy”. In: *Chemical Physics Letters* 25.2, pp. 282–283 (cit. on p. 11).

– (1974b). “Selective-phase ion cyclotron resonance spectroscopy”. In: *Canadian Journal of Chemistry* 52.10, pp. 1997–1999 (cit. on p. 4).

Cornett, Dale S., Michelle L. Reyzer, Pierre Chaurand, and Richard M. Caprioli (Oct. 2007). “MALDI imaging mass spectrometry: molecular snapshots of biochemical systems.” In: *Nat Methods* 4.10, pp. 828–833 (cit. on p. 9).

Cornish, Timothy J and Robert J Cotter (1993). “Tandem time-of-flight mass spectrometer”. In: *Analytical Chemistry* 65.8, pp. 1043–1047 (cit. on p. 5).

Cox, Jürgen and Matthias Mann (Dec. 2008). “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.” In: *Nat Biotechnol* 26.12, pp. 1367–1372 (cit. on pp. 19, 39, 40, 107, 108, 112).

– (June 2011). “Quantitative, high-resolution proteomics for data-driven systems biology.” In: *Annu Rev Biochem* 80, pp. 273–299 (cit. on p. 1).

Craig, Robertson and Ronald C Beavis (2004). “TANDEM: matching proteins with tandem mass spectra”. In: *Bioinformatics* 20.9, pp. 1466–1467 (cit. on p. 16).

Dancik, V., T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner (1999). “De novo peptide sequencing via tandem mass spectrometry.” In: *J Comput Biol* 6.3-4, pp. 327–342 (cit. on p. 16).

Davidian, Marie and Raymond J Carroll (1987). “Variance function estimation”. In: *Journal of the American Statistical Association* 82.400, pp. 1079–1091 (cit. on p. 35).

de Jong, Ebbing P. and Timothy J. Griffin (Oct. 2012). “Online nanoscale ERLIC-MS outperforms RPLC-MS for shotgun proteomics in complex mixtures.” In: *J Proteome Res* 11.10, pp. 5059–5064 (cit. on p. 31).

Dephoure, Noah and Steven P. Gygi (Mar. 2012). “Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast.” In: *Sci Signal* 5.217, rs2 (cit. on p. 26).

Diella, Francesca, Scott Cameron, Christine Gemünd, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Pontén, Nikolaj Blom, and Toby J. Gibson (June 2004). “Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins.” In: *BMC Bioinformatics* 5, p. 79 (cit. on p. 45).

Dinkel, Holger, Claudia Chica, Allegra Via, Cathryn M. Gould, Lars J. Jensen, Toby J. Gibson, and Francesca Diella (Jan. 2011). “Phospho.ELM: a database of phosphorylation sites–update 2011.” In: *Nucleic Acids Res* 39.Database issue, pp. D261–D267 (cit. on p. 45).

Domon, Bruno and Ruedi Aebersold (July 2010). “Options and considerations when selecting a quantitative proteomics strategy.” In: *Nat Biotechnol* 28.7, pp. 710–721 (cit. on pp. 13, 18).

Dost, Banu, Nuno Bandeira, Xiangqian Li, Zhouxin Shen, Steven P. Briggs, and Vineet Bafna (Apr. 2012). “Accurate mass spectrometry based protein quantification via shared peptides.” In: *J Comput Biol* 19.4, pp. 337–348 (cit. on pp. 38, 104).

Douglas, Donald J., Aaron J. Frank, and Dunmin Mao (2005). “Linear ion traps in mass spectrometry”. In: *Mass Spectrometry Reviews* 24.1, pp. 1–29 (cit. on p. 10).

Dowell, James A., Dustin C. Frost, Jiang Zhang, and Lingjun Li (Sept. 2008). “Comparison of two-dimensional fractionation techniques for shotgun proteomics.” In: *Anal Chem* 80.17, pp. 6715–6723 (cit. on p. 5).

Dudoit, S., J.P. Shaffer, and J.C. Boldrick (2003). “Multiple hypothesis testing in microarray experiments”. In: *Statistical Science*, pp. 71–103 (cit. on p. 41).

Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke (2002). “A variance-stabilizing transformation for gene-expression microarray data.” In: *Bioinformatics* 18 Suppl 1, S105–S110 (cit. on pp. 32, 35).

Eckel-Passow, J. E., A. L. Oberg, T. M. Therneau, and HR Bergen 3rd (July 2009). “An insight into high-resolution mass-spectrometry data.” In: *Biostatistics* 10.3, pp. 481–500 (cit. on pp. 15, 42).

Efron, Bradley (2008). “Microarrays, empirical Bayes and the two-groups model”. In: *Statistical Science*, pp. 1–22 (cit. on p. 92).

– (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press (cit. on p. 92).

Egertson, Jarrett D., Andreas Kuehn, Gennifer E. Merrihew, Nicholas W. Bateman, Brendan X. MacLean, Ying S. Ting, Jesse D. Canterbury, Donald M. Marsh, Markus Kellmann, Vlad

Zabrouskov, Christine C. Wu, and Michael J. MacCoss (Aug. 2013). “Multiplexed MS/MS for improved data-independent acquisition.” In: *Nat Methods* 10.8, pp. 744–746 (cit. on p. 13).

Einstein, Albert (1905). “Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?” In: *Annalen der Physik* 323.13, pp. 639–641 (cit. on p. 25).

Elias, Joshua E. and Steven P. Gygi (Mar. 2007). “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.” In: *Nat Methods* 4.3, pp. 207–214 (cit. on p. 16).

Eng, J. K., A. L. McCormack, and John R. Yates 3rd (1994). “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database”. In: *Journal of the American Society for Mass Spectrometry* 5.11, pp. 976–989 (cit. on pp. 5, 16).

Eng, Jimmy K., Brian C. Searle, Karl R. Clauser, and David L. Tabb (Nov. 2011). “A face in the crowd: recognizing peptides through database search.” In: *Mol Cell Proteomics* 10.11, R111.009522 (cit. on p. 16).

Engholm-Keller, Kasper and Martin R. Larsen (Mar. 2013). “Technologies and challenges in large-scale phosphoproteomics.” In: *Proteomics* 13.6, pp. 910–931 (cit. on pp. 1, 43).

Everley, Robert A., Ryan C. Kunz, Fiona E. McAllister, and Steven P. Gygi (June 2013). “Increasing throughput in targeted proteomics assays: 54-plex quantitation in a single mass spectrometry run.” In: *Anal Chem* 85.11, pp. 5340–5346 (cit. on p. 26).

Farnsworth, Phllo T (Aug. 1934). *Electron multiplier*. US Patent 1,969,399 (cit. on p. 12).

Feng, Shun, Mingliang Ye, Houjiang Zhou, Xiaogang Jiang, Xingning Jiang, Hanfa Zou, and Bolin Gong (Sept. 2007). “Immobilized zirconium ion affinity chromatography for specific enrichment of phosphopeptides in phosphoproteome analysis.” In: *Mol Cell Proteomics* 6.9, pp. 1656–1665 (cit. on p. 7).

Fenn, John B., Matthias Mann, Chin Kai Meng, Shek Fu Wong, and Craig M. Whitehouse (1989). “Electrospray ionization for mass spectrometry of large biomolecules”. In: *Science* 246.4926, pp. 64–71 (cit. on pp. 5, 9).

Ferguson, Roisean E., Helen P. Carroll, Adrian Harris, Eamonn R. Maher, Peter J. Selby, and Rosamonde E. Banks (Feb. 2005). “Housekeeping proteins: a preliminary study illustrating some limitations as useful references in protein expression studies.” In: *Proteomics* 5.2, pp. 566–571 (cit. on p. 33).

Fernbach, Nora V., Melanie Planyavsky, André Müller, Florian P. Breitwieser, Jacques Colinge, Uwe Rix, and Keiryn L. Bennett (Oct. 2009). “Acid elution and one-dimensional shotgun analysis on an Orbitrap mass spectrometer: an application to drug affinity chromatography.” In: *J Proteome Res* 8.10, pp. 4753–4765 (cit. on pp. 108, 169).

Fisher, Ronald Aylmer (1935). *The Design of Experiments*. Edinburgh and London: Oliver & Boyd. (cit. on p. 42).

Frese, Christian K., Houjiang Zhou, Thomas Taus, A F Maarten Altelaar, Karl Mechtler, Albert J R. Heck, and Shabaz Mohammed (Mar. 2013). “Unambiguous phosphosite localization

using electron-transfer/higher-energy collision dissociation (EThcD).” In: *J Proteome Res* 12.3, pp. 1520–1525 (cit. on p. 43).

Futrell, Jean H and CD Miller (1966). “Tandem Mass Spectrometer for Study of Ion-Molecule Reactions”. In: *Review of Scientific Instruments* 37.11, pp. 1521–1526 (cit. on p. 5).

Gallien, Sebastien, Elodie Duriez, Catharina Crone, Markus Kellmann, Thomas Moehring, and Bruno Domon (Dec. 2012). “Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer.” In: *Mol Cell Proteomics* 11.12, pp. 1709–1723 (cit. on p. 13).

Gatto, Laurent (2013). *MSnbase: Base Functions and Classes for MS-based Proteomics*. R package version 1.9.4 (cit. on p. 103).

Gatto, Laurent, Nick J Bond, and Pavel V Shliaha (2014). “Package "synapter"”. In: (cit. on p. 103).

Gatto, Laurent and Kathryn S. Lilley (Jan. 2012). “MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.” In: *Bioinformatics* 28.2, pp. 288–289 (cit. on pp. 27, 46, 103).

Geer, Lewis Y., Sanford P. Markey, Jeffrey A. Kowalak, Lukas Wagner, Ming Xu, Dawn M. Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H. Bryant (2004). “Open mass spectrometry search algorithm.” In: *J Proteome Res* 3.5, pp. 958–964 (cit. on p. 16).

Geiger, Tamar, Juergen Cox, and Matthias Mann (Oct. 2010). “Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation.” In: *Mol Cell Proteomics* 9.10, pp. 2252–2261 (cit. on p. 13).

Geiger, Tamar, Jacek R. Wisniewski, Juergen Cox, Sara Zanivan, Marcus Kruger, Yasushi Ishihama, and Matthias Mann (Feb. 2011). “Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics.” In: *Nat Protoc* 6.2, pp. 147–157 (cit. on p. 20).

Gelman, Andrew (2005). “Analysis of variance - why it is more important than ever”. In: *The Annals of Statistics* 33.1, pp. 1–53 (cit. on p. 46).

Gelman, Andrew, John B. Carlin, Stern Hal S., and Donald B. Rubin (2003). *Bayesian Data Analysis, 2nd Edition*. Chapman and Hall/CRC (cit. on pp. 94, 95).

Gerber, Scott A, John Rush, Olaf Stemman, Marc W Kirschner, and Steven P Gygi (June 2003). “Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.” In: *Proc Natl Acad Sci U S A* 100.12, pp. 6940–6945 (cit. on p. 13).

Gerster, Sarah, Taejoon Kwon, Christina Ludwig, Mariette Matondo, Christine Vogel, Edward Marcotte, Ruedi Aebersold, and Peter Bühlmann (Nov. 2013). “Statistical approach to protein quantification.” In: *Mol Cell Proteomics* (cit. on pp. 38, 104).

Giambruno, Roberto, Florian Grebien, Alexey Stukalov, Christian Knoll, Melanie Planyavsky, Elena L. Rudashevskaya, Jacques Colinge, Giulio Superti-Furga, and Keiryn L. Bennett (Sept. 2013). “Affinity purification strategies for proteomic analysis of transcription factor complexes.” In: *J Proteome Res* 12.9, pp. 4018–4027 (cit. on p. 5).

Gillet, Ludovic C., Pedro Navarro, Stephen Tate, Hannes Röst, Nathalie Selevsek, Lukas Reiter, Ron Bonner, and Ruedi Aebersold (June 2012). "Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis." In: *Mol Cell Proteomics* 11.6, O111.016717 (cit. on p. 13).

Glish, Gary L. and Richard W. Vachet (Feb. 2003). "The basics of mass spectrometry in the twenty-first century." In: *Nat Rev Drug Discov* 2.2, pp. 140–150 (cit. on p. 9).

Gluck, Florent, Christine Hoogland, Paola Antinori, Xavier Robin, Frederic Nikitin, Anne Zufferey, Carla Pasquarello, Vanessa Fétaud, Loïc Dayon, Markus Müller, Frederique Lisacek, Laurent Geiser, Denis Hochstrasser, Jean-Charles Sanchez, and Alexander Scherl (Feb. 2013). "EasyProt–an easy-to-use graphical platform for proteomics data analysis." In: *J Proteomics* 79, pp. 146–160 (cit. on p. 103).

Gnad, Florian, Jeremy Gunawardena, and Matthias Mann (Jan. 2011). "PHOSIDA 2011: the posttranslational modification database." In: *Nucleic Acids Res* 39.Database issue, pp. D253–D260 (cit. on p. 45).

Gnad, Florian, Shubin Ren, Juergen Cox, Jesper V. Olsen, Boris Macek, Mario Oroshi, and Matthias Mann (2007). "PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites." In: *Genome Biol* 8.11, R250 (cit. on p. 45).

Gouw, Joost W., Jeroen Krijgsveld, and Albert J R. Heck (Jan. 2010). "Quantitative proteomics by metabolic labeling of model organisms." In: *Mol Cell Proteomics* 9.1, pp. 11–24 (cit. on pp. 2, 19, 20, 27).

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold (Oct. 1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." In: *Nat Biotechnol* 17.10, pp. 994–999 (cit. on pp. 20, 21).

Haura, Eric B., André Müller, Florian P. Breitwieser, Jiannong Li, Florian Grebien, Jacques Colinge, and Keiryn L. Bennett (Jan. 2011). "Using iTRAQ combined with tandem affinity purification to enhance low-abundance proteins associated with somatically mutated EGFR core complexes in lung cancer." In: *J Proteome Res* 10.1, pp. 182–190 (cit. on pp. iii, iv, 34, 86–88).

Hebert, Alexander S., Anna E. Merrill, Derek J. Bailey, Amelia J. Still, Michael S. Westphall, Eric R. Strieter, David J. Pagliarini, and Joshua J. Coon (Apr. 2013a). "Neutron-encoded mass signatures for multiplexed proteome quantification." In: *Nat Methods* 10.4, pp. 332–334 (cit. on pp. 21, 26, 112).

Hebert, Alexander S., Alicia L. Richards, Derek J. Bailey, Arne Ulbrich, Emma E. Coughlin, Michael S. Westphall, and Joshua J. Coon (Oct. 2013b). "The One Hour Yeast Proteome." In: *Mol Cell Proteomics* (cit. on pp. 1, 112).

Hill, Elizabeth G., John H. Schwacke, Susana Comte-Walters, Elizabeth H. Slate, Ann L. Oberg, Jeanette E. Eckel-Passow, Terry M. Therneau, and Kevin L. Schey (Aug. 2008). "A statistical

model for iTRAQ data analysis." In: *J Proteome Res* 7.8, pp. 3091–3101 (cit. on pp. 34, 46, 105).

Hoffmann, Edmond and Vincent Stroobant (2007). *Mass Spectrometry. Principles and Applications (Third Edition)*. Wiley Online Library (cit. on pp. 4, 11).

Hornbeck, Peter V., Jon M. Kornhauser, Sasha Tkachev, Bin Zhang, Elzbieta Skrzypek, Beth Murray, Vaughan Latham, and Michael Sullivan (Jan. 2012). "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse." In: *Nucleic Acids Res* 40.Database issue, pp. D261–D270 (cit. on pp. 45, 110, 113).

Houel, Stephane, Robert Abernathy, Kutralanathan Renganathan, Karen Meyer-Arendt, Natalie G. Ahn, and William M. Old (Aug. 2010). "Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies." In: *J Proteome Res* 9.8, pp. 4152–4160 (cit. on pp. 13, 30).

Hsu, Jue-Liang, Sheng-Yu Huang, Nan-Haw Chow, and Shu-Hui Chen (Dec. 2003). "Stable-isotope dimethyl labeling for quantitative proteomics." In: *Anal Chem* 75.24, pp. 6843–6852 (cit. on pp. 20, 21).

Hsu, Peggy P., Seong A. Kang, Jonathan Rameseder, Yi Zhang, Kathleen A. Ottina, Daniel Lim, Timothy R. Peterson, Yongmun Choi, Nathanael S. Gray, Michael B. Yaffe, Jarrod A. Marto, and David M.n Sabatini (June 2011). "The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling." In: *Science* 332.6035, pp. 1317–1322 (cit. on pp. 40, 112).

Hu, Jun, Jin Qian, Oleg Borisov, Sanqiang Pan, Yan Li, Tong Liu, Longwen Deng, Kenneth Wannemacher, Michael Kurnellas, Christa Patterson, Stella Elkabes, and Hong Li (Aug. 2006). "Optimized proteomic analysis of a mouse model of cerebellar dysfunction using amine-specific isobaric tags." In: *Proteomics* 6.15, pp. 4321–4334 (cit. on p. 34).

Hubbell, Earl, Wei-Min Liu, and Rui Mei (Dec. 2002). "Robust estimators for expression analysis." In: *Bioinformatics* 18.12, pp. 1585–1592 (cit. on p. 92).

Huber, Wolfgang, Anja von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." In: *Bioinformatics* 18 Suppl 1, S96–104 (cit. on pp. 32, 35, 36, 104).

Hundertmark, C., R. Fischer, T. Reinl, S. May, F. Klawonn, and L. Jänsch (Apr. 2009). "MS-specific noise model reveals the potential of iTRAQ in quantitative proteomics." In: *Bioinformatics* 25.8, pp. 1004–1011 (cit. on pp. 34, 35, 39, 40, 59, 105, 106).

Hunt, D. F., A. M. Buko, J. M. Ballard, J. Shabanowitz, and A. B. Giordani (Sept. 1981). "Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer." In: *Biomed Mass Spectrom* 8.9, pp. 397–408 (cit. on p. 5).

Irizarry, Rafael A., Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed (Feb. 2003a). “Summaries of Affymetrix GeneChip probe level data.” In: *Nucleic Acids Res* 31.4, e15 (cit. on p. 92).

Irizarry, Rafael A., Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed (Apr. 2003b). “Exploration, normalization, and summaries of high density oligonucleotide array probe level data.” In: *Biostatistics* 4.2, pp. 249–264 (cit. on p. 92).

Ishihama, Yasushi, Yoshiya Oda, Tsuyoshi Tabata, Toshitaka Sato, Takeshi Nagasu, Juri Rappsilber, and Matthias Mann (Sept. 2005). “Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.” In: *Mol Cell Proteomics* 4.9, pp. 1265–1272 (cit. on pp. 19, 89, 111).

Iskratsch, Thomas, Stephan Lange, Joseph Dwyer, Ay Lin Kho, Cris dos Remedios, and Elisabeth Ehler (Dec. 2010). “Formin follows function: a muscle-specific isoform of FHOD3 is regulated by CK2 phosphorylation and promotes myofibril maintenance.” In: *J Cell Biol* 191.6, pp. 1159–1172 (cit. on p. 69).

James, William and Charles Stein (1961). “Estimation with quadratic loss”. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1961, pp. 361–379 (cit. on p. 92).

Jedrychowski, Mark P., Edward L. Huttlin, Wilhelm Haas, Mathew E. Sowa, Ramin Rad, and Steven P. Gygi (Dec. 2011). “Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics.” In: *Mol Cell Proteomics* 10.12, p. M111.009910 (cit. on p. 15).

Jones, Andrew R., Martin Eisenacher, Gerhard Mayer, Oliver Kohlbacher, Jennifer Siepen, Simon J. Hubbard, Julian N. Selley, Brian C. Searle, James Shofstahl, Sean L. Seymour, Randall Julian, Pierre-Alain Binz, Eric W. Deutsch, Henning Hermjakob, Florian Reisinger, Johannes Griss, Juan Antonio Vizcaíno, Matthew Chambers, Angel Pizarro, and David Creasy (July 2012). “The mzIdentML data standard for mass spectrometry-based proteomics results.” In: *Mol Cell Proteomics* 11.7, p. M111.014381 (cit. on p. 2).

Kadane, Joseph B (2011). *Principles of uncertainty*. CRC Press (cit. on pp. 95, 100).

Karp, Natasha A., Wolfgang Huber, Pawel G. Sadowski, Philip D. Charles, Svenja V. Hester, and Kathryn S. Lilley (Sept. 2010). “Addressing accuracy and precision issues in iTRAQ quantitation.” In: *Mol Cell Proteomics* 9.9, pp. 1885–1897 (cit. on pp. 2, 32, 34, 36, 39, 42, 104, 105, 107, 108, 112).

Keller, Andrew, Jimmy Eng, Ning Zhang, Xiao-jun Li, and Ruedi Aebersold (2005). “A uniform proteomics MS/MS analysis platform utilizing open XML file formats.” In: *Mol Syst Biol* 1, p. 2005.0017 (cit. on pp. 34, 39, 47).

Kim, Sangtae, Nikolai Mischerikow, Nuno Bandeira, J Daniel Navarro, Louis Wich, Shabaz Mohammed, Albert J R. Heck, and Pavel A. Pevzner (Dec. 2010). “The generating function of

CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search." In: *Mol Cell Proteomics* 9.12, pp. 2840–2852 (cit. on p. 16).

Kim, Sung Chan, Robert Sprung, Yue Chen, Yingda Xu, Haydn Ball, Jimin Pei, Tzuling Cheng, Yoonjung Kho, Hao Xiao, Lin Xiao, Nick V. Grishin, Michael White, Xiang-Jiao Yang, and Yingming Zhao (Aug. 2006). "Substrate and functional diversity of lysine acetylation revealed by a proteomics survey." In: *Mol Cell* 23.4, pp. 607–618 (cit. on p. 7).

Knuth, Donald Ervin (1979). *TEX and METAFONT: New directions in typesetting*. American Mathematical Society (cit. on p. 104).

Köcher, Thomas, Peter Pichler, Michael Schutzbier, Christoph Stingl, Axel Kaul, Nils Teucher, Gerd Hasenfuss, Josef M. Penninger, and Karl Mechtler (Oct. 2009). "High precision quantitative proteomics using iTRAQ on an LTQ Orbitrap: a new mass spectrometric method combining the benefits of all." In: *J Proteome Res* 8.10, pp. 4743–4752 (cit. on pp. 15, 25).

Koehler, Christian J., Magnus Ø. Arntzen, Gustavo Antonio de Souza, and Bernd Thiede (Feb. 2013). "An approach for triplex-isobaric peptide termini labeling (triplex-IPTL)." In: *Anal Chem* 85.4, pp. 2478–2485 (cit. on p. 111).

Koehler, Christian J., Magnus O. Arntzen, Margarita Strozynski, Achim Treumann, and Bernd Thiede (June 2011). "Isobaric peptide termini labeling utilizing site-specific N-terminal succinylation." In: *Anal Chem* 83.12, pp. 4775–4781 (cit. on pp. 21, 26, 111).

Koppenaal, David W., Charles J. Barinaga, M Bonner Denton, Roger P. Sperline, Gary M. Hieftje, Gregory D. Schilling, Francisco J. Andrade, and James H Barnes 4th (Nov. 2005). "MS Detectors". In: *Anal Chem* 77.21, 418A–427A (cit. on p. 11).

Kovanich, Duangnapa, Salvatore Cappadona, Reinout Raijmakers, Shabaz Mohammed, Arjen Scholten, and Albert J R. Heck (Sept. 2012). "Applications of stable isotope dimethyl labeling in quantitative proteomics." In: *Anal Bioanal Chem* 404.4, pp. 991–1009 (cit. on pp. 20, 27).

Lane, Lydie, Ghislaine Argoud-Puy, Aurore Britan, Isabelle Cusin, Paula D. Duek, Olivier Evalet, Alain Gateau, Pascale Gaudet, Anne Gleizes, Alexandre Masselot, Catherine Zwahlen, and Amos Bairoch (Jan. 2012). "neXtProt: a knowledge platform for human proteins." In: *Nucleic Acids Res* 40.Database issue, pp. D76–D83 (cit. on pp. 45, 110, 113).

Lau, Edward, Maggie P Y. Lam, S. O. Siu, Ricky P W. Kong, Wai Lung Chan, Zhongjun Zhou, Jirong Huang, Clive Lo, and Ivan K. Chu (May 2011). "Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications." In: *Mol Biosyst* 7.5, pp. 1399–1408 (cit. on p. 31).

Leisch, Friedrich (2002). "Sweave: Dynamic generation of statistical reports using literate data analysis". In: *Compstat*. Springer, pp. 575–580 (cit. on p. 104).

Levin, Yishai (June 2011). "The role of statistical power analysis in quantitative proteomics." In: *Proteomics* 11.12, pp. 2565–2567 (cit. on p. 41).

Li, Xiao-Jun, Hui Zhang, Jeffrey A. Ranish, and Ruedi Aebersold (Dec. 2003). "Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry." In: *Anal Chem* 75.23, pp. 6648–6657 (cit. on p. 39).

Li, Yong Fuga and Predrag Radivojac (2012). "Computational approaches to protein inference in shotgun proteomics." In: *BMC Bioinformatics* 13 Suppl 16, S4 (cit. on p. 37).

Lin, Wen-Ting, Wei-Neng Hung, Yi-Hwa Yian, Kun-Pin Wu, Chia-Li Han, Yet-Ran Chen, Yu-Ju Chen, Ting-Yi Sung, and Wen-Lian Hsu (Sept. 2006). "Multi-Q: a fully automated tool for multiplexed protein quantitation." In: *J Proteome Res* 5.9, pp. 2328–2338 (cit. on pp. 2, 30, 32, 33, 39, 64, 106).

Liu, Xiao-Hua, Ling-Jia Qian, Jing-Bo Gong, Jing Shen, Xue-Min Zhang, and Xiao-Hong Qian (Oct. 2004). "Proteomic analysis of mitochondrial proteins in cardiomyocytes from chronic stressed rat." In: *Proteomics* 4.10, pp. 3167–3176 (cit. on p. 69).

Mahoney, Douglas W, Terry M Therneau, Carrie J Heppelmann, Leeann Higgins, Linda M Benson, Roman M Zenka, Pratik Jagtap, Gary L Nelsestuen, H. Robert Bergen, and Ann L Oberg (Aug. 2011). "Relative Quantification: Characterization of Bias, Variability and Fold Changes in Mass Spectrometry Data from iTRAQ-Labeled Peptides." In: *J Proteome Res* (cit. on pp. 36, 42).

Makarov (Mar. 2000). "Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis". In: *Anal Chem* 72.6, pp. 1156–1162 (cit. on p. 4).

Makarov, Alexander, Eduard Denisov, Oliver Lange, and Stevan Horning (July 2006). "Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer." In: *J Am Soc Mass Spectrom* 17.7, pp. 977–982 (cit. on p. 10).

Mallick, Parag and Bernhard Kuster (July 2010). "Proteomics: a pragmatic perspective." In: *Nat Biotechnol* 28.7, pp. 695–709 (cit. on p. 1).

Mamyrin, BA (2001). "Time-of-flight mass spectrometry (concepts, achievements, and prospects)". In: *International Journal of Mass Spectrometry* 206.3, pp. 251–266 (cit. on p. 10).

Mamyrin, BA, VI Karataev, DV Shmikk, and VA Zagulin (1973). "The massreflectron, a new non-magnetic time-of-flight mass spectrometer with high resolution". In: *Sov. Phys. JETP* 64, pp. 82–89 (cit. on p. 11).

Mandel, Micha, Manor Askenazi, Yi Zhang, and Jarrod A Marto (2013). "Variance function estimation in quantitative mass spectrometry with application to iTRAQ labeling". In: *The Annals of Applied Statistics* 7.1, pp. 1–24 (cit. on pp. 35, 106).

Mattauch, Josef (1936). "A double-focusing mass spectrograph and the masses of N 15 and O 18". In: *Physical Review* 50.7, p. 617 (cit. on p. 4).

Matthiesen, Rune (Aug. 2007). "Methods, algorithms and tools in computational proteomics: a practical point of view." In: *Proteomics* 7.16, pp. 2815–2832 (cit. on p. 27).

Matthiesen, Rune and Ana Sofia Carvalho (2010). "Methods and algorithms for relative quantitative proteomics by mass spectrometry." In: *Methods Mol Biol* 593, pp. 187–204 (cit. on p. 27).

Maurer, Margarita, André C. Müller, Katja Parapatics, Winfried F. Pickl, Christine Wagner, Elena L. Rudashevskaya, Florian P. Breitwieser, Jacques Colinge, Kanika Garg, Johannes Griss, Keiryn L. Bennett, and Stephan N. Wagner (June 2014). "Comprehensive comparative and semiquantitative proteome of a very low number of native and matched epstein-barr-virus-transformed B lymphocytes infiltrating human melanoma." In: *J Proteome Res* 13.6, pp. 2830–2845 (cit. on p. 167).

McAlister, Graeme C., Edward L. Huttlin, Wilhelm Haas, Lily Ting, Mark P. Jedrychowski, John C. Rogers, Karsten Kuhn, Ian Pike, Robert A. Grothe, Justin D. Blethrow, and Steven P. Gygi (Sept. 2012). "Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses." In: *Anal Chem* 84.17, pp. 7469–7478 (cit. on pp. 26, 114).

McLafferty, F. W. (Oct. 1981). "Tandem mass spectrometry." In: *Science* 214.4518, pp. 280–287 (cit. on p. 12).

Mertins, Philipp, Namrata D. Udeshi, Karl R. Clauser, D. R. Mani, Jinal Patel, Shao-En Ong, Jacob D. Jaffe, and Steven A. Carr (Dec. 2011). "iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics." In: *Mol Cell Proteomics* (cit. on p. 32).

Metzker, Michael L. (Jan. 2010). "Sequencing technologies - the next generation." In: *Nat Rev Genet* 11.1, pp. 31–46 (cit. on pp. 1, 5).

Michalski, Annette, Juergen Cox, and Matthias Mann (Apr. 2011). "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS." In: *J Proteome Res* 10.4, pp. 1785–1793 (cit. on p. 13).

Millikan, Robert Andrews (1913). "On the elementary electrical charge and the Avogadro constant". In: *Physical Review* 2.2, pp. 109–143 (cit. on p. 4).

Minguez, Pablo, Luca Parca, Francesca Diella, Daniel R. Mende, Runjun Kumar, Manuela Helmer-Citterich, Anne-Claude Gavin, Vera van Noort, and Peer Bork (2012). "Deciphering a global network of functionally associated post-translational modifications". In: *Mol Syst Biol* 8, p. 599 (cit. on p. 42).

Mirgorodskaya, O. A., Y. P. Kozmin, M. I. Titov, R. Körner, C. P. Sönksen, and P. Roepstorff (2000). "Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards." In: *Rapid Commun Mass Spectrom* 14.14, pp. 1226–1232 (cit. on p. 21).

Miyagi, Masaru and K C Sekhar Rao (2007). "Proteolytic 18O-labeling strategies for quantitative proteomics." In: *Mass Spectrom Rev* 26.1, pp. 121–136 (cit. on p. 20).

Morelle, Willy, Kévin Canis, Frédéric Chirat, Valegh Faid, and Jean-Claude Michalski (July 2006). "The use of mass spectrometry for the proteomic analysis of glycosylation." In: *Proteomics* 6.14, pp. 3993–4015 (cit. on p. 7).

Morris, Howard R, Thanai Paxton, Maria Panico, Roy McDowell, and Anne Dell (1997). "A novel geometry mass spectrometer, the Q-TOF, for low-femtomole/attomole-range biopolymer sequencing". In: *Journal of Protein Chemistry* 16.5, pp. 469–479 (cit. on p. 5).

Mueller, Lukas N, Mi-Youn Brusniak, D. R. Mani, and Ruedi Aebersold (Jan. 2008). "An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data." In: *J Proteome Res* 7.1, pp. 51–61 (cit. on p. 113).

Müller, André C., Florian P. Breitwieser, Heinz Fischer, Christopher Schuster, Oliver Brandt, Jacques Colinge, Giulio Superti-Furga, Georg Stingl, Adelheid Elbe-Bürger, and Keiryn L. Bennett (July 2012). "A comparative proteomic study of human skin suction blister fluid from healthy individuals using immunodepletion and iTRAQ labeling." In: *J Proteome Res* 11.7, pp. 3715–3727 (cit. on pp. iii, iv, 2, 86, 89, 90, 114, 168).

Murray, Kermit K., Robert K. Boyd, Marcos N. Eberlin, G John Langley, Liang Li, and Yasuhide Naito (2013). "Definitions of terms relating to mass spectrometry (IUPAC Recommendations 2013)". In: *Pure and Applied Chemistry* None, None–None (cit. on pp. 9, 14).

Nagaraj, Nagarjuna, Nils Alexander Kulak, Juergen Cox, Nadin Neuhauser, Korbinian Mayr, Ole Hoerning, Ole Vorm, and Matthias Mann (Mar. 2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap." In: *Mol Cell Proteomics* 11.3, p. M111.013722 (cit. on p. 1).

Nahnsen, Sven, Chris Bielow, Knut Reinert, and Oliver Kohlbacher (Mar. 2013). "Tools for label-free peptide quantification." In: *Mol Cell Proteomics* 12.3, pp. 549–556 (cit. on p. 19).

Nesvizhskii, Alexey I. (Oct. 2010). "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." In: *J Proteomics* 73.11, pp. 2092–2123 (cit. on p. 16).

Nesvizhskii, Alexey I. and Ruedi Aebersold (Oct. 2005). "Interpretation of shotgun proteomic data: the protein inference problem." In: *Mol Cell Proteomics* 4.10, pp. 1419–1440 (cit. on p. 37).

Nesvizhskii, Alexey I., Olga Vitek, and Ruedi Aebersold (Oct. 2007). "Analysis and validation of proteomic data generated by tandem mass spectrometry." In: *Nat Methods* 4.10, pp. 787–797 (cit. on pp. 2, 15).

Neyman, Jerzy and Elizabeth L Scott (1948). "Consistent estimates based on partially consistent observations". In: *Econometrica: Journal of the Econometric Society*, pp. 1–32 (cit. on p. 106).

Nier, Alfred O and Earl A Gulbransen (1939). "Variations in the relative abundance of the carbon isotopes". In: *Journal of the American Chemical Society* 61.3, pp. 697–698 (cit. on p. 4).

Niessen, Wilfried MA (2010). *Liquid chromatography-mass spectrometry*. CRC Press (cit. on p. 9).

Nilsen, Timothy W. and Brenton R. Graveley (Jan. 2010). "Expansion of the eukaryotic proteome by alternative splicing." In: *Nature* 463.7280, pp. 457–463 (cit. on p. 37).

Noble, William Stafford and Michael J. MacCoss (Jan. 2012). "Computational and statistical analysis of protein mass spectrometry data." In: *PLoS Comput Biol* 8.1, e1002296 (cit. on p. 15).

Oberg, Ann L. and Douglas W. Mahoney (2012). “Statistical methods for quantitative mass spectrometry proteomic experiments with labeling.” In: *BMC Bioinformatics* 13 Suppl 16, S7 (cit. on pp. 32, 34).

Oberg, Ann L and Olga Vitek (May 2009). “Statistical design of quantitative mass spectrometry-based proteomic experiments.” In: *J Proteome Res* 8.5, pp. 2144–2156 (cit. on p. 42).

Oda, Y., K. Huang, F. R. Cross, D. Cowburn, and B. T. Chait (June 1999). “Accurate quantitation of protein expression and site-specific phosphorylation.” In: *Proc Natl Acad Sci U S A* 96.12, pp. 6591–6596 (cit. on pp. 20, 21).

Olsen, Jesper V., Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann (Nov. 2006). “Global, in vivo, and site-specific phosphorylation dynamics in signaling networks.” In: *Cell* 127.3, pp. 635–648 (cit. on p. 44).

Olsen, Jesper V., Boris Macek, Oliver Lange, Alexander Makarov, Stevan Horning, and Matthias Mann (Sept. 2007). “Higher-energy C-trap dissociation for peptide modification analysis.” In: *Nat Methods* 4.9, pp. 709–712 (cit. on pp. 15, 43).

Olsen, Jesper V. and Matthias Mann (Nov. 2013). “Status of large-scale analysis of post-translational modifications by mass spectrometry.” In: *Mol Cell Proteomics* (cit. on pp. 1, 42, 43, 112).

Olsen, Jesper V., Shao-En Ong, and Matthias Mann (June 2004). “Trypsin cleaves exclusively C-terminal to arginine and lysine residues.” In: *Mol Cell Proteomics* 3.6, pp. 608–614 (cit. on p. 5).

Ong, Shao-En, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann (May 2002). “Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.” In: *Mol Cell Proteomics* 1.5, pp. 376–386 (cit. on pp. 20, 21, 27).

Ong, Shao-En and Matthias Mann (Oct. 2005). “Mass spectrometry-based proteomics turns quantitative.” In: *Nat Chem Biol* 1.5, pp. 252–262 (cit. on p. 20).

Onsongo, Getiria, Matthew D. Stone, Susan K. Van Riper, John Chilton, Baolin Wu, Leeann Higgins, Troy C. Lund, John V. Carlis, and Timothy J. Griffin (Oct. 2010). “LTQ-iQuant: A freely available software pipeline for automated and accurate protein quantification of isobaric tagged peptide data from LTQ instruments.” In: *Proteomics* 10.19, pp. 3533–3538 (cit. on p. 34).

Oshlack, Alicia, Mark D. Robinson, and Matthew D. Young (2010). “From RNA-seq reads to differential expression results.” In: *Genome Biol* 11.12, p. 220 (cit. on p. 1).

Ow, Saw Yen, Malinda Salim, Josselin Noirel, Caroline Evans, Ishtiaq Rehman, and Phillip C. Wright (Nov. 2009). “iTRAQ underestimation in simple and complex mixtures: "the good, the bad and the ugly".” In: *J Proteome Res* 8.11, pp. 5347–5355 (cit. on pp. 30, 34).

Ow, Saw Yen, Malinda Salim, Josselin Noirel, Caroline Evans, and Phillip. C. Wright (2011). "Minimising iTRAQ ratio compression through understanding LC-MS elution dependence and high-resolution HILIC fractionation". In: *PROTEOMICS* 11.11, pp. 2341–2346 (cit. on p. 31).

Parikh, Jignesh R., Manor Askenazi, Scott B. Ficarro, Tanya Cashorali, James T. Webber, Nathaniel C. Blank, Yi Zhang, and Jarrod A. Marto (2009). "multiplierz: an extensible API based desktop environment for proteomics data analysis." In: *BMC Bioinformatics* 10, p. 364 (cit. on p. 32).

Paul, Wolfgang and Helmut Steinwedel (1953). "Ein neues Massenspektrometer ohne Magnetfeld". In: *Zeitschrift Naturforschung Teil A* 8, p. 448 (cit. on pp. 4, 10).

Perkins, D. N., D. J. Pappin, D. M. Creasy, and J. S. Cottrell (Dec. 1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." In: *Electrophoresis* 20.18, pp. 3551–3567 (cit. on pp. 16, 30).

Perry, Richard H., R Graham Cooks, and Robert J. Noll (2008). "Orbitrap mass spectrometry: instrumentation, ion motion and applications." In: *Mass Spectrom Rev* 27.6, pp. 661–699 (cit. on p. 11).

Phanstiel, Douglas H., Justin Brumbaugh, Craig D. Wenger, Shulan Tian, Mitchell D. Probasco, Derek J. Bailey, Danielle L. Swaney, Mark A. Tervo, Jennifer M. Bolin, Victor Ruotti, Ron Stewart, James A. Thomson, and Joshua J. Coon (2011). "Proteomic and phosphoproteomic comparison of human ES and iPS cells." In: *Nat Methods* 8.10, pp. 821–827 (cit. on pp. 29, 42, 71, 80, 81, 85).

Pichler, Peter, Thomas Köcher, Johann Holzmann, Michael Mazanek, Thomas Taus, Gustav Ammerer, and Karl Mechtler (Aug. 2010). "Peptide labeling with isobaric tags yields higher identification rates using iTRAQ 4-plex compared to TMT 6-plex and iTRAQ 8-plex on LTQ Orbitrap." In: *Anal Chem* 82.15, pp. 6549–6558 (cit. on p. 24).

Pichlmair, Andreas, Caroline Lassnig, Carol-Ann Eberle, Maria W. Górna, Christoph L. Baumann, Thomas R. Burkard, Tilmann Bürckstümmer, Adrijana Stefanovic, Sigurd Krieger, Keiryn L. Bennett, Thomas Rülicke, Friedemann Weber, Jacques Colinge, Mathias Müller, and Giulio Superti-Furga (2011). "IFIT1 is an antiviral protein that recognizes 5'-triphosphate RNA." In: *Nat Immunol* 12.7, pp. 624–630 (cit. on p. 1).

Picotti, Paola and Ruedi Aebersold (June 2012). "Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions." In: *Nat Methods* 9.6, pp. 555–566 (cit. on p. 18).

Picotti, Paola, Mathieu Clément-Ziza, Henry Lam, David S. Campbell, Alexander Schmidt, Eric W. Deutsch, Hannes Röst, Zhi Sun, Oliver Rinner, Lukas Reiter, Qin Shen, Jacob J. Michaelson, Andreas Frei, Simon Alberti, Ulrike Kusebauch, Bernd Wollscheid, Robert L. Moritz, Andreas Beyer, and Ruedi Aebersold (Feb. 2013). "A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis." In: *Nature* 494.7436, pp. 266–270 (cit. on p. 18).

Pinkse, Martijn W H., Pauliina M. Uitto, Martijn J. Hilhorst, Bert Ooms, and Albert J R. Heck (July 2004). "Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns." In: *Anal Chem* 76.14, pp. 3935–3943 (cit. on p. 7).

Polisetty, Ravindra Varma, Poonam Gautam, Rakesh Sharma, H. C. Harsha, Sudha C. Nair, Manoj Kumar Gupta, Megha S. Uppin, Sundaram Challa, Aneel Kumar Puligopu, Praveen Ankathi, Aniruddh K. Purohit, Giriraj R. Chandak, Akhilesh Pandey, and Ravi Sirdeshmukh (June 2012). "LC-MS/MS analysis of differentially expressed glioblastoma membrane proteome reveals altered calcium signaling and other protein groups of regulatory functions." In: *Mol Cell Proteomics* 11.6, p. M111.013565 (cit. on p. 39).

Pollreisz, Andreas, Marion Funk, Florian P. Breitwieser, Katja Parapatics, Stefan Sacu, Michael Georgopoulos, Roman Dunavoelgyi, Gerhard J. Zlabinger, Jacques Colinge, Keiryn L. Bennett, and Ursula Schmidt-Erfurth (Mar. 2013). "Quantitative proteomics of aqueous and vitreous fluid from patients with idiopathic epiretinal membranes." In: *Exp Eye Res* 108, pp. 48–58 (cit. on pp. iv, 88, 89, 114, 168).

Porath, J., J. Carlsson, I. Olsson, and G. Belfrage (Dec. 1975). "Metal chelate affinity chromatography, a new approach to protein fractionation." In: *Nature* 258.5536, pp. 598–599 (cit. on p. 7).

Quackenbush, John (Dec. 2002). "Microarray data normalization and transformation." In: *Nat Genet* 32 Suppl, pp. 496–501 (cit. on p. 32).

Rappsilber, Juri, Ursula Ryder, Angus I. Lamond, and Matthias Mann (Aug. 2002). "Large-scale proteomic analysis of the human spliceosome." In: *Genome Res* 12.8, pp. 1231–1245 (cit. on p. 19).

Richards, Alicia L., Catherine E. Vincent, Adrian Guthals, Christopher M. Rose, Michael S. Westphall, Nuno Bandeira, and Joshua J. Coon (Sept. 2013). "Neutron-encoded signatures enable automated product ion annotation from tandem mass spectra." In: *Mol Cell Proteomics* (cit. on p. 26).

Rikova, Klarisa, Ailan Guo, Qingfu Zeng, Anthony Possemato, Jian Yu, Herbert Haack, Julie Nardone, Kimberly Lee, Cynthia Reeves, Yu Li, Yerong Hu, Zhiping Tan, Matthew Stokes, Laura Sullivan, Jeffrey Mitchell, Randy Wetzel, Joan Macneill, Jian Min Ren, Jin Yuan, Corey E. Bakalarski, Judit Villen, Jon M. Kornhauser, Bradley Smith, Daiqiang Li, Xinmin Zhou, Steven P. Gygi, Ting-Lei Gu, Roberto D. Polakiewicz, John Rush, and Michael J. Comb (Dec. 2007). "Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer." In: *Cell* 131.6, pp. 1190–1203 (cit. on p. 7).

Rix, U., L. L. Remsing Rix, A. S. Terker, N. V. Fernbach, O. Hantschel, M. Planyavsky, F. P. Breitwieser, H. Herrmann, J. Colinge, K. L. Bennett, M. Augustin, J. H. Till, M. C. Heinrich, P. Valent, and G. Superti-Furga (Jan. 2010). "A comprehensive target selectivity survey of the

BCR-ABL kinase inhibitor INNO-406 by kinase profiling and chemical proteomics in chronic myeloid leukemia cells." In: *Leukemia* 24.1, pp. 44–50 (cit. on p. 169).

Rodríguez-Suárez, Eva, Ewa Gubb, Itziar Frades Alzueta, Juan Manuel Falcón-Pérez, António Amorim, Felix Elortza, and Rune Matthiesen (Apr. 2010). "Virtual expert mass spectrometrist: iTRAQ tool for database-dependent search, quantitation and result storage." In: *Proteomics* 10.8, pp. 1545–1556 (cit. on pp. 39, 47, 105, 106).

Ross, Philip L., Yulin N. Huang, Jason N. Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, Subhasish Purkayastha, Peter Juhasz, Stephen Martin, Michael Bartlet-Jones, Feng He, Allan Jacobson, and Darryl J. Pappin (Dec. 2004). "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents." In: *Mol Cell Proteomics* 3.12, pp. 1154–1169 (cit. on pp. 20, 21, 24).

Rudashevskaya, Elena L., Florian P. Breitwieser, Marie L. Huber, Jacques Colinge, André C. Müller, and Keiryn L. Bennett (Feb. 2013). "Multiple and sequential data acquisition method: an improved method for fragmentation and detection of cross-linked peptides on a hybrid linear trap quadrupole Orbitrap Velos mass spectrometer." In: *Anal Chem* 85.3, pp. 1454–1461 (cit. on pp. 114, 167).

Ruttenberg, Brian E., Trairak Pisitkun, Mark A. Knepper, and Jason D. Hoffert (July 2008). "PhosphoScore: an open-source phosphorylation site assignment tool for MSn data." In: *J Proteome Res* 7.7, pp. 3054–3059 (cit. on p. 44).

Saeed, Fahad, Trairak Pisitkun, Jason D. Hoffert, Guanghui Wang, Marjan Gucek, and Mark A. Knepper (Oct. 2012). "An Efficient Dynamic Programming Algorithm for Phosphorylation Site Assignment of Large-Scale Mass Spectrometry Data." In: *Proceedings (IEEE Int Conf Bioinformatics Biomed)*, pp. 618–625 (cit. on p. 44).

Savitski, Mikhail M., Frank Fischer, Toby Mathieson, Gavain Sweetman, Manja Lang, and Marcus Bantscheff (Oct. 2010). "Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays." In: *J Am Soc Mass Spectrom* 21.10, pp. 1668–1679 (cit. on pp. 13, 31).

Savitski, Mikhail M., Simone Lemeer, Markus Boesche, Manja Lang, Toby Mathieson, Marcus Bantscheff, and Bernhard Kuster (Feb. 2011a). "Confident phosphorylation site localization using the Mascot Delta Score." In: *Mol Cell Proteomics* 10.2, p. M110.003830 (cit. on pp. 43, 109, 113).

Savitski, Mikhail M, Toby Mathieson, Nico Zinn, Gavain Sweetman, Carola Doce, Isabelle Becher, Fiona Pachl, Bernhard Kuster, and Marcus Bantscheff (2013). "Measuring and managing ratio compression for accurate iTRAQ/TMT quantification". In: *Journal of Proteome Research* 12.8, pp. 3586–3598 (cit. on pp. 32, 114).

Savitski, Mikhail M., Gavain Sweetman, Manor Askenazi, Jarrod A. Marto, Manja Lang, Nico Zinn, and Marcus Bantscheff (Dec. 2011b). "Delayed fragmentation and optimized isolation

width settings for improvement of protein identification and accuracy of isobaric mass tag quantification on Orbitrap-type mass spectrometers." In: *Anal Chem* 83.23, pp. 8959–8967 (cit. on pp. 86, 91).

Schilling, Birgit, Matthew J. Rardin, Brendan X. MacLean, Anna M. Zawadzka, Barbara E. Frewen, Michael P. Cusack, Dylan J. Sorensen, Michael S. Bereman, Enxuan Jing, Christine C. Wu, Eric Verdin, C Ronald Kahn, Michael J. Maccoss, and Bradford W. Gibson (May 2012). "Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation." In: *Mol Cell Proteomics* 11.5, pp. 202–214 (cit. on p. 19).

Schmidt, Alexander, Josef Kellermann, and Friedrich Lottspeich (Jan. 2005). "A novel strategy for quantitative proteomics using isotope-coded protein labels." In: *Proteomics* 5.1, pp. 4–15 (cit. on pp. 20, 21).

Schwacke, John H., Elizabeth G. Hill, Edward L. Krug, Susana Comte-Walters, and Kevin L. Schey (2009). "iQuantitator: a tool for protein expression inference using iTRAQ." In: *BMC Bioinformatics* 10, p. 342 (cit. on pp. 2, 46, 101, 108).

Schwämmle, Veit, Ileana Rodríguez León, and Ole Nørregaard Jensen (Sept. 2013). "Assessment and improvement of statistical tools for comparative proteomics analysis of sparse data sets with few experimental replicates." In: *J Proteome Res* 12.9, pp. 3874–3883 (cit. on p. 112).

Serang, Oliver and William Noble (2012). "A review of statistical methods for protein identification using tandem mass spectrometry." In: *Stat Interface* 5.1, pp. 3–20 (cit. on p. 37).

Shi, Leming, Wendell D. Jones, Roderick V. Jensen, Stephen C. Harris, Roger G. Perkins, Federico M. Goodsaid, Lei Guo, Lisa J. Croner, Cecilie Boysen, Hong Fang, Feng Qian, Shashi Amur, Wenjun Bao, Catalin C. Barbacioru, Vincent Bertholet, Xiaoxi Megan Cao, Tzu-Ming Chu, Patrick J. Collins, Xiao-Hui Fan, Felix W. Frueh, James C. Fuscoe, Xu Guo, Jing Han, Damir Herman, Huixiao Hong, Ernest S. Kawasaki, Quan-Zhen Li, Yuling Luo, Yunqing Ma, Nan Mei, Ron L. Peterson, Raj K. Puri, Richard Shippy, Zhenqiang Su, Yongming Andrew Sun, Hongmei Sun, Brett Thorn, Yaron Turpaz, Charles Wang, Sue Jane Wang, Janet A. Warrington, James C. Willey, Jie Wu, Qian Xie, Liang Zhang, Lu Zhang, Sheng Zhong, Russell D. Wolfinger, and Weida Tong (2008). "The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies." In: *BMC Bioinformatics* 9 Suppl 9, S10 (cit. on p. 112).

Sidoli, Simone, Lei Cheng, and Ole N. Jensen (June 2012). "Proteomics in chromatin biology and epigenetics: Elucidation of post-translational modifications of histone proteins by mass spectrometry." In: *J Proteomics* 75.12, pp. 3419–3433 (cit. on p. 7).

Singh, Puneet, John J. Leddy, George J. Chatzis, Maysoon Salih, and Balwant S. Tuana (Feb. 2005). "Alternative splicing generates a CaM kinase IIbeta isoform in myocardium that targets the sarcoplasmic reticulum through a putative alphaKAP and regulates GAPDH." In: *Mol Cell Biochem* 270.1-2, pp. 215–221 (cit. on p. 69).

Siuti, Nertila and Neil L. Kelleher (Oct. 2007). “Decoding protein modifications using top-down mass spectrometry.” In: *Nat Methods* 4.10, pp. 817–821 (cit. on p. 7).

Sleno, Lekha (Feb. 2012). “The use of mass defect in modern mass spectrometry.” In: *J Mass Spectrom* 47.2, pp. 226–236 (cit. on p. 25).

Sleno, Lekha and Dietrich A. Volmer (Oct. 2004). “Ion activation methods for tandem mass spectrometry.” In: *J Mass Spectrom* 39.10, pp. 1091–1112 (cit. on p. 14).

Smyth, Gordon K. (2004). “Linear models and empirical bayes methods for assessing differential expression in microarray experiments.” In: *Stat Appl Genet Mol Biol* 3, Article3 (cit. on pp. 92, 112).

Sohn, Chang Ho, J Eugene Lee, Michael J. Sweredoski, Robert L J. Graham, Geoffrey T. Smith, Sonja Hess, Gregg Czerwieniec, Joseph A. Loo, Raymond J. Deshaies, and J. L. Beauchamp (Feb. 2012). “Click chemistry facilitates formation of reporter ions and simplified synthesis of amine-reactive multiplexed isobaric tags for protein quantification.” In: *J Am Chem Soc* 134.5, pp. 2672–2680 (cit. on pp. 21, 24).

Stafford Jr, G. C., P. E. Kelley, J. E. P. Syka, W. E. Reynolds, and J. F. J. Todd (1984). “Recent improvements in and analytical applications of advanced ion trap technology”. In: *International Journal of Mass Spectrometry and Ion Processes* 60.1, pp. 85–98 (cit. on pp. 4, 10).

Stein, Charles (1956). “Inadmissibility of the usual estimator for the mean of a multivariate normal distribution”. In: *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*. Vol. 1. 399, pp. 197–206 (cit. on p. 92).

Stephens, WE (1946). “A pulsed mass spectrometer with time dispersion”. In: *Bull. Am. Phys. Soc* 21.691, p. 22 (cit. on pp. 4, 10).

Storey, John D. and Robert Tibshirani (Aug. 2003). “Statistical significance for genomewide studies.” In: *Proc Natl Acad Sci U S A* 100.16, pp. 9440–9445 (cit. on p. 41).

Syka, John E P., Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt (June 2004). “Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.” In: *Proc Natl Acad Sci U S A* 101.26, pp. 9528–9533 (cit. on p. 15).

Tabb, David L., Lorenzo Vega-Montoto, Paul A. Rudnick, Asokan Mulayath Variyath, Amy-Joan L. Ham, David M. Bunk, Lisa E. Kilpatrick, Dean D. Billheimer, Ronald K. Blackman, Helene L. Cardasis, Steven A. Carr, Karl R. Clauser, Jacob D. Jaffe, Kevin A. Kowalski, Thomas A. Neubert, Fred E. Regnier, Birgit Schilling, Tony J. Tegeler, Mu Wang, Pei Wang, Jeffrey R. Whiteaker, Lisa J. Zimmerman, Susan J. Fisher, Bradford W. Gibson, Christopher R. Kinsinger, Mehdi Mesri, Henry Rodriguez, Stephen E. Stein, Paul Tempst, Amanda G. Paulovich, Daniel C. Liebler, and Cliff Spiegelman (Feb. 2010). “Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry.” In: *J Proteome Res* 9.2, pp. 761–776 (cit. on p. 13).

Tanaka, Koichi, Yutaka Ido, S Akita, Y Yoshida, and T Yoshida (1987). “Detection of high mass molecules by laser desorption time-of-flight mass spectrometry”. In: *Proceedings of the Second Japan-China Joint Symposium on Mass Spectrometry*. Vol. 185 (cit. on pp. 5, 9).

Tantau, Till (2013). “Graph drawing in TikZ”. In: *Graph Drawing (Lecture Notes in Computer Science Volume 7704)*. Springer, pp. 517–528 (cit. on p. 104).

Taus, Thomas, Thomas Köcher, Peter Pichler, Carmen Paschke, Andreas Schmidt, Christoph Henrich, and Karl Mechtler (Nov. 2011). “Universal and Confident Phosphorylation Site Localization Using phosphoRS.” In: *J Proteome Res* (cit. on pp. 44, 109, 113).

Thompson, Andrew, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, R. Johnstone, A. Karim A Mohammed, and Christian Hamon (Apr. 2003). “Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.” In: *Anal Chem* 75.8, pp. 1895–1904 (cit. on pp. 20, 21, 24).

Thomson, Joseph J. (1897). “Cathode Rays”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 44.269, pp. 293–316 (cit. on p. 4).

Thomson, Joseph J (1912). “Further experiments on positive rays”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 24.140, pp. 209–253 (cit. on pp. 4, 11).

Tiberti, Natalia, Alexandre Hainard, Veerle Lejon, Xavier Robin, Dieudonné Mumba Ngoyi, Natacha Turck, Enock Matovu, John Enyaru, Joseph Mathu Ndung’u, Alexander Scherl, Loïc Dayon, and Jean-Charles Sanchez (Dec. 2010). “Discovery and verification of osteopontin and Beta-2-microglobulin as promising markers for staging human African trypanosomiasis.” In: *Mol Cell Proteomics* 9.12, pp. 2783–2795 (cit. on p. 97).

Ting, Lily, Mark J Cowley, Seah Lay Hoon, Michael Guilhaus, Mark J Raftery, and Ricardo Cavicchioli (Oct. 2009). “Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling.” In: *Mol Cell Proteomics* 8.10, pp. 2227–2242 (cit. on pp. 32, 112).

Ting, Lily, Ramin Rad, Steven P. Gygi, and Wilhelm Haas (Nov. 2011). “MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics.” In: *Nat Methods* 8.11, pp. 937–940 (cit. on pp. 12, 30, 31, 111, 114).

Todd, JOHN FJ (1991). “Recommendations for nomenclature and symbolism for mass spectroscopy”. In: *Pure and Applied Chemistry* 63, pp. 1541–1566 (cit. on p. 7).

Toko, Haruhiro, Hidehisa Takahashi, Yosuke Kayama, Toru Oka, Tohru Minamino, Sho Okada, Sachio Morimoto, Dong-Yun Zhan, Fumio Terasaki, Mark E. Anderson, Masashi Inoue, Atsushi Yao, Ryozo Nagai, Yasushi Kitaura, Toshiyuki Sasaguri, and Issei Komuro (Aug. 2010). “Ca2+/calmodulin-dependent kinase IIdelta causes heart failure by accumulation of p53 in dilated cardiomyopathy.” In: *Circulation* 122.9, pp. 891–899 (cit. on p. 69).

Tolin, Serena, Giorgio Arrigoni, Roberto Moscatiello, Antonio Masi, Lorella Navazio, Gaurav Sablok, and Andrea Squartini (June 2013). “Quantitative analysis of the naringenin-inducible

proteome in Rhizobium leguminosarum by isobaric tagging and mass spectrometry." In: *Proteomics* 13.12-13, pp. 1961–1972 (cit. on p. 40).

Tukey, John (1977). *Exploratory Data Analysis*. Addison-Wesley (cit. on p. 106).

Tusher, V. G., R. Tibshirani, and G. Chu (Apr. 2001). "Significance analysis of microarrays applied to the ionizing radiation response." In: *Proc Natl Acad Sci U S A* 98.9, pp. 5116–5121 (cit. on p. 92).

Ubaida Mohien, Ceereena, Jürgen Hartler, Florian Breitwieser, Uwe Rix, Lily Remsing Rix, Georg E. Winter, Gerhard G. Thallinger, Keiryn L. Bennett, Giulio Superti-Furga, Zlatko Trajanoski, and Jacques Colinge (July 2010). "MASPECTRAS 2: An integration and analysis platform for proteomic data." In: *Proteomics* 10.14, pp. 2719–2722 (cit. on p. 168).

Varjosalo, Markku, Roberto Sacco, Alexey Stukalov, Audrey van Drogen, Melanie Planyavsky, Simon Hauri, Ruedi Aebersold, Keiryn L. Bennett, Jacques Colinge, Matthias Gstaiger, and Giulio Superti-Furga (Apr. 2013). "Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS." In: *Nat Methods* 10.4, pp. 307–314 (cit. on pp. 1, 5).

Vaudel, Marc, Daniela Breiter, Florian Beck, Jörg Rahnenführer, Lennart Martens, and René P. Zahedi (Mar. 2013). "D-score: a search engine independent MD-score." In: *Proteomics* 13.6, pp. 1036–1041 (cit. on p. 44).

Vogel, Christine and Edward M. Marcotte (Apr. 2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses." In: *Nat Rev Genet* 13.4, pp. 227–232 (cit. on p. 1).

Washburn, Michael P., Ryan Ulaszek, Cosmin Deciu, David M. Schieltz, and John R. Yates 3rd (Apr. 2002). "Analysis of quantitative proteomic data generated via multidimensional protein identification technology." In: *Anal Chem* 74.7, pp. 1650–1657 (cit. on p. 20).

Weisser, Hendrik, Sven Nahnsen, Jonas Grossmann, Lars Nilse, Andreas Quandt, Hendrik Brauer, Marc Sturm, Erhan Kenar, Oliver Kohlbacher, Ruedi Aebersold, and Lars Malmström (Feb. 2013). "An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics." In: *J Proteome Res* (cit. on p. 19).

Weng, Lee, Hongyue Dai, Yihui Zhan, Yudong He, Sergey B. Stepaniants, and Douglas E. Bassett (May 2006). "Rosetta error model for gene expression analysis." In: *Bioinformatics* 22.9, pp. 1111–1121 (cit. on p. 35).

Wenger, Craig D., M Violet Lee, Alexander S. Hebert, Graeme C. McAlister, Douglas H. Phanstiel, Michael S. Westphall, and Joshua J. Coon (Nov. 2011). "Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging." In: *Nat Methods* 8.11, pp. 933–935 (cit. on pp. 30, 31).

Werner, Thilo, Isabelle Becher, Gavain Sweetman, Carola Doce, Mikhail M. Savitski, and Marcus Bantscheff (Aug. 2012). "High-resolution enabled TMT 8-plexing." In: *Anal Chem* 84.16, pp. 7188–7194 (cit. on pp. 21, 26, 112).

Wilm, Matthias (July 2011). “Principles of electrospray ionization.” In: *Molecular and Cellular Proteomics* 10.7, p. M111.009407 (cit. on p. 9).

Wilm, Matthias S and Matthias Mann (1994). “Electrospray and Taylor-Cone theory, Dole’s beam of macromolecules at last?” In: *International Journal of Mass Spectrometry and Ion Processes* 136.2, pp. 167–180 (cit. on p. 9).

Winter, Georg E., Uwe Rix, Scott M. Carlson, Karoline V. Gleixner, Florian Grebien, Manuela Gridling, André C. Müller, Florian P. Breitwieser, Martin Bilban, Jacques Colinge, Peter Valent, Keiryn L. Bennett, Forest M. White, and Giulio Superti-Furga (Nov. 2012). “Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML.” In: *Nat Chem Biol* 8.11, pp. 905–912 (cit. on pp. iv, 1, 2, 86, 89–91, 114, 168).

Witze, Eric S., William M. Old, Katheryn A. Resing, and Natalie G. Ahn (Oct. 2007). “Mapping protein post-translational modifications with mass spectrometry.” In: *Nat Methods* 4.10, pp. 798–806 (cit. on pp. 6, 12).

Workman, Christopher, Lars Juhl Jensen, Hanne Jarmer, Randy Berka, Laurent Gautier, Henrik Bjørn Nielser, Hans-Henrik Saxild, Claus Nielsen, Søren Brunak, and Steen Knudsen (Aug. 2002). “A new non-linear normalization method for reducing variability in DNA microarray experiments.” In: *Genome Biol* 3.9, research0048 (cit. on p. 32).

Wu, Ronghu, Noah Dephoure, Wilhelm Haas, Edward L Huttlin, Bo Zhai, Mathew E Sowa, and Steven P Gygi (Aug. 2011). “Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes.” In: *Mol Cell Proteomics* 10.8, p. M111.009654 (cit. on pp. 2, 44, 83, 109, 113).

Xiang, Feng, Hui Ye, Ruibing Chen, Qiang Fu, and Lingjun Li (Apr. 2010). “N,N-dimethyl leucines as novel isobaric tandem mass tags for quantitative proteomics and peptidomics.” In: *Anal Chem* 82.7, pp. 2817–2825 (cit. on p. 21).

Xu, Xiangdong, Dongmei Yang, Jian-Hua Ding, Wang Wang, Pao-Hsien Chu, Nancy D. Dalton, Huan-You Wang, John R Bermingham Jr, Zhen Ye, Forrest Liu, Michael G. Rosenfeld, James L. Manley, John Ross Jr, Ju Chen, Rui-Ping Xiao, Heping Cheng, and Xiang-Dong Fu (Jan. 2005). “ASF/SF2-regulated CaMKIIdelta alternative splicing temporally reprograms excitation-contraction coupling in cardiac muscle.” In: *Cell* 120.1, pp. 59–72 (cit. on p. 69).

Yang, Yee Hwa, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terence P. Speed (Feb. 2002). “Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.” In: *Nucleic Acids Res* 30.4, e15 (cit. on p. 32).

Yang, Yung-Hun, Kwangwon Lee, Kyoung-Soon Jang, Yun-Gon Kim, Sung-Hee Park, Chang-Soo Lee, and Byung-Gee Kim (Apr. 2009). “Low mass cutoff evasion with q(z) value optimization in ion trap.” In: *Anal Biochem* 387.1, pp. 133–135 (cit. on pp. 15, 24).

Yao, X., A. Freas, J. Ramirez, P. A. Demirev, and C. Fenselau (July 2001). “Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus.” In: *Anal Chem* 73.13, pp. 2836–2842 (cit. on p. 20).

Yates, John R., Cristian I. Ruse, and Aleksey Nakorchevsky (2009). “Proteomics by mass spectrometry: approaches, advances, and applications.” In: *Annu Rev Biomed Eng* 11, pp. 49–79 (cit. on pp. 6, 9).

Yu, Chuan-Yih, Yin-Hao Tsui, Yi-Hwa Yian, Ting-Yi Sung, and Wen-Lian Hsu (July 2007). “The Multi-Q web server for multiplexed protein quantitation.” In: *Nucleic Acids Res* 35.Web Server issue, W707–W712 (cit. on p. 47).

Zahurak, Marianna, Giovanni Parmigiani, Wayne Yu, Robert B. Scharpf, David Berman, Edward Schaeffer, Shabana Shabbeer, and Leslie Cope (2007). “Pre-processing Agilent microarray data.” In: *BMC Bioinformatics* 8, p. 142 (cit. on p. 92).

Zaki, Rafdzah, Awang Bulgiba, Roshidi Ismail, and Noor Azina Ismail (2012). “Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review.” In: *PLoS One* 7.5, e37908 (cit. on p. 32).

Zhang, Junxiang, Yan Wang, and Shuwei Li (Sept. 2010). “Deuterium isobaric amine-reactive tags for quantitative proteomics.” In: *Anal Chem* 82.18, pp. 7588–7595 (cit. on p. 21).

Zhang, Wei, Feng Qi, Dong-Qin Chen, Wen-Yan Xiao, Jing Wang, and Wei-Zhong Zhu (Aug. 2010). “Ca2+/calmodulin-dependent protein kinase IIdelta orchestrates G-protein-coupled receptor and electric field stimulation-induced cardiomyocyte hypertrophy.” In: *Clin Exp Pharmacol Physiol* 37.8, pp. 795–802 (cit. on p. 69).

Zhang, Yaoyang, Bryan R. Fonslow, Bing Shan, Moon-Chang Baek, and John R. Yates 3rd (Apr. 2013). “Protein Analysis by Shotgun/Bottom-up Proteomics.” In: *Chem Rev* 113.4, pp. 2343–2394 (cit. on pp. 1, 5, 12, 17).

Zhang, Yi, Manor Askenazi, Jingrui Jiang, C. John Luckey, James D Griffin, and Jarrod A Marto (May 2010). “A robust error model for iTRAQ quantification reveals divergent signaling between oncogenic FLT3 mutants in acute myeloid leukemia.” In: *Mol Cell Proteomics* 9.5, pp. 780–790 (cit. on pp. 35, 59, 105, 106).

Zhang, Yi, Scott B. Ficarro, Shaojuan Li, and Jarrod A. Marto (Aug. 2009). “Optimized Orbitrap HCD for quantitative analysis of phosphopeptides.” In: *J Am Soc Mass Spectrom* 20.8, pp. 1425–1434 (cit. on p. 24).

Zhang, Ying, Zhihui Wen, Michael P. Washburn, and Laurence Florens (Mar. 2010). “Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins.” In: *Anal Chem* 82.6, pp. 2272–2281 (cit. on pp. 19, 89, 111).

Zhao, Yingming and Ole N. Jensen (Oct. 2009). “Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques.” In: *Proteomics* 9.20, pp. 4632–4641 (cit. on p. 7).

Zhou, Yuan, Yichu Shan, Qi Wu, Shen Zhang, Lihua Zhang, and Yukui Zhang (Nov. 2013). “Mass defect-based pseudo-isobaric dimethyl labeling for proteome quantification.” In: *Anal Chem* 85.22, pp. 10658–10663 (cit. on p. 111).

Zubarev, R. A., D. M. Horn, E. K. Fridriksson, N. L. Kelleher, N. A. Kruger, M. A. Lewis, B. K. Carpenter, and F. W. McLafferty (Feb. 2000). “Electron capture dissociation for structural characterization of multiply charged protein cations.” In: *Anal Chem* 72.3, pp. 563–573 (cit. on p. 15).

Zubarev, Roman A. and Alexander Makarov (May 2013). “Orbitrap Mass Spectrometry.” In: *Anal Chem* (cit. on pp. 11, 12, 15).

Zybailov, Boris L., Laurence Florens, and Michael P. Washburn (May 2007). “Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors.” In: *Mol Biosyst* 3.5, pp. 354–360 (cit. on p. 19).

7. Abbreviations

Da	Dalton
$\frac{m}{z}$	mass-to-charge ratio
ppm	Parts per million
AP	Affinity prurification
CID	Collision induced dissociation
DDA	Data-dependent acquisition
ECD	Electron-capture dissociation
ETD	Electron-transfer dissociation
HCD	Higher energy collisional dissociation, or higher energy collisional C-trap dissociation
HPLC	High performance LC, or high pressure LC
IMAC	Immobilized metal affinity
iTRAQ	Isobaric tags for relative and absolute quantitation
LC	Liquid chromatography
MS	Mass spectrometry
MS^1	First mass spectrometry dimension (i.e. survey scan)
MS^2	Second mass spectrometry dimension (i.e. product ion scan)
MS^n	n^{th} mass spectrometry dimension
PTM	Post-translational modification
Q-TOF	Quadrupole-Time-of-flight
SILAC	Stable isotope labeling by amino acids in cell culture
SRM	Selected reaction monitoring
TMT	Tandem mass tags
WLS	Weighted least squares
XIC	Extracted ion chromatogram

A. Vignette for *isobar* R package

A.1. Introduction

The `isobar` package is designed as an extensible and interactive environment for data analysis and exploration of data produced by Mass Spectrometry analysis of proteins and peptides labelled with isobaric tags, such as iTRAQ and TMT. `isobar` implements the theory presented in Breitwieser et al., Journal of Proteome Research 2011.

`isobar` allows analyzing iTRAQ 4plex and 8plex, and TMT 2plex and 6plex experiments representing them as `IBSpectra` objects. The respective classes are `iTRAQ4plexSpectra`, `iTRAQ8plexSpectra`, `TMT2plexSpectra`, `TMT6plexSpectra` and `TMT10plexSpectra`.

The first thing you need to do is load the package.

```
> library(isobar) ## load the isobar package
```

A.2. Loading data

`isobar` can read identifications and quantifications from tab-separated and MGF files. Perl scripts are supplied to generate a tab-separated version from the vendor formats of Mascot and Phenyx, see appendix A.5.3. The format is simple and described in appendix A.5.1. Experimental support for the mzIdentML format within R is also available - please contact the mantainer in case of problems.

ID.CSV tab-separated file containing peptide-spectra matches and spectrum meta-information such as retention time, m/z and charge. Generated by parser scripts.

MGF contains peak lists from which quantitative information on reporter tags are extracted. Must be centroided.

IBSPECTRA.CSV tab-separated file containing the same columns as ID.CSV plus *quantitative information* extracted from MGF file - that means the reporter tag masses and intensities as additional columns.

`readIBSpectra` is the primary function to generate a `IBSpectra` object. The first argument is one of iTRAQ4plexSpectra, iTRAQ8plexSpectra, TMT2plexSpectra, TMT6plexSpectra and TMT10plexSpectra denotes the tag type and therefore class.

```
>   ## generating IBSpectra object from ID.CSV and MGF
>   ib <- readIBSpectra("iTRAQ4plexSpectra",list.files(pattern=".id.csv"),
+           list.files(pattern=".mgf"))
>   ## write in tabular IBSPECTRA.CSV format to file
>   write.table(as.data.frame(ib),sep="\t",row.names=F,
+           file="myexperiment.ibspectra.csv")
>   ## generate from saved IBSPECTRA.CSV - MGF does not have to be supplied
>   ib.2 <- readIBSpectra("iTRAQ4plexSpectra","myexperiment.ibspectra.csv")
```

In case the MGF file is very big, it can be advanteguous to generate a smaller version containing only meta- and quantitative information before import in R. On Linux, the tool grep is readily available.

```
egrep '^[A-Z]|^1[12][0-9]\.' BIG.mgf > SMALL.mgf
```

A.2.1. ibspiked test samples

The examples presented are based on the dataset ibspiked_set1 which has been designed to test isobar's functionality and searched against the Swissprot human database with Mascot and Phenyx. ibspiked_set1 is an iTRAQ 4-plex data set comprised of a complex background (albumin- and IgG-depleted human plasma) and spiked proteins. MS analysis was performed in ThermoFisher Scientific LTQ Orbitrap HCD instrument with 2D shotgun peptide separation (see original paper for more details). The samples used for each iTRAQ channel are as follows:

- Depleted human plasma background (>150 protein detected);
- Spiked-in proteins with the following ratios
  - CERU_HUMAN (P00450) at concentrations $1:1:1:1$;
  - CERU_RAT (P13635) at concentrations $1:2:5:10$;
  - CERU_MOUSE (Q61147) at concentrations $10:5:2:1$.

A second data set with ratios 1:10:50:100 is available as ibspiked_set2 from http://bininformatics.cemm.oeaw.ac.at/isobar.

The Ceruplasmins have been selected as the share peptides. Hereafter, we load the data pacakage and the ceru protein IDs are identified via the protein.g function, which provides a mean to retrieve data from ProteinGroup objects. ProteinGroup is a slot of IBSpectra objects and contains informations on proteins and their grouping. See A.2.2.

```
> data(ibspiked_set1)
> ceru.human <- protein.g(proteinGroup(ibspiked_set1),"CERU_HUMAN")
> ceru.rat <- protein.g(proteinGroup(ibspiked_set1),"CERU_RAT")
> ceru.mouse <- protein.g(proteinGroup(ibspiked_set1),"CERU_MOUSE")
> ceru.proteins <- c(ceru.human,ceru.rat,ceru.mouse)
```

A.2.2. Protein information and grouping in `ProteinGroup`

When an `ibspectra.csv` is read, protein are grouped to identify proteins which have unique peptides. By default, only peptides with unique peptides are grouped.

The algorithm to infer protein groups works as follows:

1. Group proteins together which have been seen with exactly the same peptides (`indistinguishableProte` - these are the `protein.g` identifiers.
2. Create protein groups (`proteinGroupTable`):
   a) Define proteins with specific peptides as reporters (`reporterProteins`)
   b) Get proteins which are contained [1] by reporterProteins and group them below.
3. Create protein groups for proteins without specific peptides as above.

A.2.3. Loading data from CID/HCD (or CID/MS3, etc) experiments

A combined CID/HCD approach, in which for each precursor two fragmentation spectra are acquired, has proven useful to increase the number of identified and quantified peptide-spectrum matches. Usually, the reporter intensity information is taken from the HCD spectrum, and the peptide is identified based on the fragment ions in the CID spectrum.

To import these experiments, a comma-separated *mapping file* is needed, which contains the association from `identification` to `quantification` spectrum title.

Example mapping file (`mapping.csv`):

```
"hcd","cid"
"spectrum 1","spectrum 2"
"spectrum 3","spectrum 4"
```

[1] That means these proteins have a subset of the peptides of the reporter

By calling `readIBSpectra` with `mapping.file="mapping.csv"`, the spectra titles in the `id.file` are matched to those in the `peaklist.file`. If the column names for the quantification and identification spectrum are not `hcd` and `cid`, resp., they can be set with the argument `mapping=c(identification.spectrum="column name 1",quantification.spectrum="column name 2")`.

Example mapping file `mapping2.csv`**:**

```
"quant-spectrum-ms3","id-spectrum-ms2"
"spectrum 1","spectrum 2"
"spectrum 3","spectrum 4"
```

```
> readIBSpectra(...,
+               mapping.file="mapping2.csv",
+               mapping=c(identification.spectrum="id-spectrum-ms2",
+                         quantification.spectrum="quant-spectrum-ms3")
+               )
```

The argument `mapping.file` can take multiple files as argument (which are read and concatenated), or a `data.frame`.

A.2.4. Loading data from CID/HCD experiments with full HCD spectrum

If a full HCD spectrum was acquired, and both the HCD and CID spectrum are searched against a protein database, the argument `id.file.domap` to `readIBSpectra` can be used to merge both CID and HCD identifications:

```
> readIBSpectra("TMT6plexSpectra",
+               id.file="cid.identifications.csv",
+               peaklist.file="hcd.peaklist.mgf",
+               id.file.domap="hcd.identifications.csv",
+               mapping.file="mapping.csv",
+               ...
+               )
```

Here, the CID (`cid.identifications.csv`) and HCD identifications (`hcd.identifications.csv`) are combined and mapped according to `mapping.csv`. Diverging identifications are discarded.

A.2.5. MSnbase integration

MSnbase by Laurent Gatto provides data manipulation and processing methods for MS-based proteomics data. It provides import, representation and analysis of raw MS data stored in mzXML, mzML and mzData using the mzR package and centroided and un-centroided MGF peak lists. It allows using and preprocessing raw data whereas isobar requires centroided peak lists. In the future, the isobar class IBSpectra might be based on or replaced by MSnbase's class MSnSet. For now, methods for coercion are implemented:

```
>   as(ibspectra,"MSnSet")
>   as(msnset,"IBSpectra")
```

A.3. Data analysis

A.3.1. Reporter mass precision

The distribution of observed masses from the reporter tags can be used to visualize the precision of the MS setup on the fragment level and used to set the correct window for isolation.

The expected masses of the reporter tags are in the slot reporterTagMasses of the implementations of the IBSpectra class. The experimental masses are in the matrix mass of AssayData; they can also be accessed by the method reporterMasses(x).

```
> sprintf("%.4f",reporterTagMasses(ibspiked_set1))  ## expected masses

[1] "114.1112" "115.1083" "116.1116" "117.1150"

> mass <- assayData(ibspiked_set1)[["mass"]] ## observerd masses
> apply(mass,2,function(x) sprintf("%.4f",quantile(x,na.rm=TRUE,probs=c(0.025,0.975))))

     114        115        116        117
[1,] "114.1110" "115.1081" "116.1115" "117.1148"
[2,] "114.1116" "115.1087" "116.1120" "117.1153"
```

reporterMassPrecision provides a plot of the distribution.

```
>   print(reporterMassPrecision(ibspiked_set1))
```

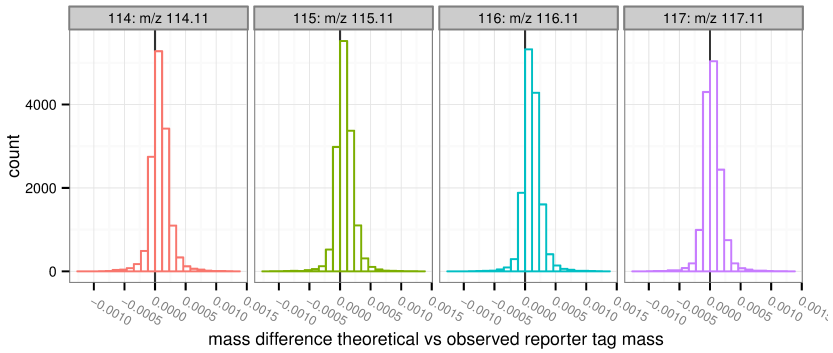


Figure 1.1: Reporter mass precision plot.

A.3.2. Normalization and isotope impurity correction

Isotope impurity correction factors are supplied by labelling reagent manufacturers. Default values that can be modified by the user are available in `isobar` and corrections are obtained by simple linear algebra.

Due to differences between samples it is advisable to normalize data before further processing. By default, `normalize` corrects by a factor such that the median intensities in all reporter channels are equal.

See figure 1.2.

```
> ib.old <- ibspiked_set1
> ibspiked_set1 <- correctIsotopeImpurities(ibspiked_set1)
> ibspiked_set1 <- normalize(ibspiked_set1)

> par(mfrow=c(1,2))
> maplot(ib.old,channel1="114",channel2="117",ylim=c(0.5,2),
+        main="before normalization")
> abline(h=1,col="red",lwd=2)
> maplot(ibspiked_set1,channel1="114",channel2="117",ylim=c(0.5,2),
+        main="after normalization")
> abline(h=1,col="red",lwd=2)
```

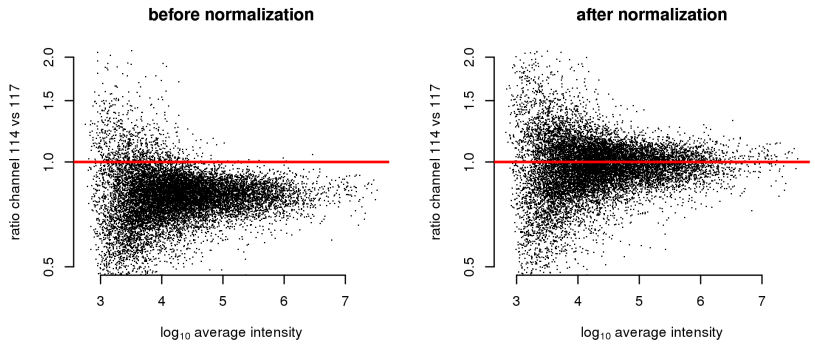


Figure 1.2: Ratio versus intensity plots ('MA plots') before and after applying normalization.

A.3.3. Fitting a noise model

A noise model is a approximation of the expected technical variation based on signal intensity. It is stable for a certain experimental setup and thus can be learned once. Noise is observed directly when comparing identical samples in multiple channels (1:1 iTRAQ/TMT sample) and we can use `ibspiked_set1` background proteins as a 1:1 sample. Therefore we exclude the ceruplasmins before fitting a noise model using `NoiseModel`. See figure 1.3.

```
> ib.background <- subsetIBSpectra(ibspiked_set1,protein=ceru.proteins,direction="exclude")
> noise.model <- NoiseModel(ib.background)

[1]  0.03423425 12.14500685  1.43708103
```

Though only recommended when sufficient data are available, a method exist for the estimation of a noise model without a 1:1 dataset. It takes longer time as it first computes all the protein ratios to shift spectrum ratios to 1:1. To examplify this procedure, we only take rat and mouse CERU proteins from `ibspiked_set1`, see figure 1.3. The resultant noise model is a rough approximation only because of the very limitted data, see Breitwieser et al. Supporting Information, submitted, for a real example.

```
> ib.ceru <- subsetIBSpectra(ibspiked_set1,protein=ceru.proteins,
+                            direction="include",
+                            specificity="reporter-specific")
> nm.ceru <- NoiseModel(ib.ceru,one.to.one=FALSE,pool=TRUE)

3 proteins with more than 10 spectra, taking top 50.
[1] 0.0000000001 0.4473733696 0.2057470615
```

```
> maplot(ib.background,noise.model=c(noise.model,nm.ceru),
+        channel1="114",channel2="115",ylim=c(0.2,5),
+        main="95% CI noise model")
```

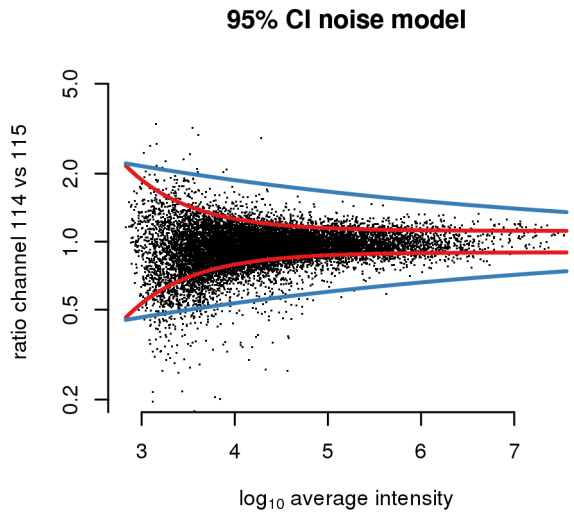


Figure 1.3: Red lines denote the 95 % confidence interval as estimated by the noise model on background proteins. The blue line is estimated as non 1:1 noise model based on only spectra of CERU proteins.

A.3.4. Protein and peptide ratio calculation

`estimateRatio` calculates the relative abundance of a peptide or protein in one tag compared to another. It calculates a weighted average (after outlier removal) of the spectrum ratios. The weights are the inverse of the spectrum ratio variances. It requires a `IBSpectra` and `NoiseModel` object and definitions of channel1, channel2, and the protein or peptide. The result is channel2/channel1.

```
> ## Calculate ratio based on all spectra of peptides specific
> ##  to CERU_HUMAN, CERU_RAT or CERU_MOUSE. Returns a named
> ##  numeric vector.
> 10^estimateRatio(ibspiked_set1,noise.model,
+                  channel1="114",channel2="115",
+                  protein=ceru.proteins)['lratio']

   lratio
0.9276031
```

```
> ## If argument 'combine=FALSE', estimateRatio returns a data.frame
> ##  with one row per protein
> 10^estimateRatio(ibspiked_set1,noise.model,
+                  channel1="114",channel2="115",
+                  protein=ceru.proteins,combine=FALSE)[,'lratio']

[1] 1.0446652 1.8324549 0.5074106

> ## spiked material channel 115 vs 114:
> ##                  CERU_HUMAN (P00450): 1:1
> ##                  CERU_RAT   (P13635): 2:1  = 2
> ##                  CERU_MOUSE (Q61147): 5:10 = 0.5
>
> ## Peptides shared between rat and mouse
> pep.shared <- peptides(proteinGroup(ibspiked_set1),
+                        c(ceru.rat,ceru.mouse),set="intersect",
+                        columns=c('peptide','n.shared.groups'))
> ## remove those which are shared with other proteins
> pep.shared <- pep.shared$peptide[pep.shared$n.shared.groups==2]
> ## calculate ratio: it is between the rat and mouse ratios
> 10^estimateRatio(ibspiked_set1,noise.model,
+                  channel1="114",channel2="115",
+                  peptide=pep.shared)['lratio']

   lratio
0.6304827
```

When examining the global differences and differences in between classes, `proteinRatios` can be used. It is also suitable to inspect sample variability. The argument `cl` can be used to define class labels. If `combn.method='interclass'` or `intraclass` and `summarize=TRUE`, `proteinRatios` return a single summarized ratio across and within classes, resp..

```
> protein.ratios <- proteinRatios(ibspiked_set1,noise.model,cl=c("1","0","0","0"))
> ## defined class 114 and 115 as class 'T', 116 and 117 as class 'C'
> classLabels(ibspiked_set1) <- c("T","T","C","C")
> proteinRatios(ibspiked_set1,noise.model,protein=ceru.proteins,
+               cl=classLabels(ibspiked_set1),combn.method="interclass",
+               summarize=T)[,c("ac","lratio","variance")]

      ac      lratio     variance
1 P00450  0.00678429 0.0006185627
```

```
2 P13635  0.60024322 0.0512591412
3 Q61147 -0.56460030 0.0466327614
```

A.3.5. Protein ratio distribution and selection

Protein ratio distributions can be calculated ideally on biological replicated. To examine differentially expressed proteins, both sample variability information (random protein ratios) as a *fold-change* constraint, and ratio *precision* can be used. For a experimental setup with biolgical replicates in the same experiment (but different channels), the distribution of biological variability can be learned by computing the ratios between the replicates. With no replicates available, one has the choice to (a) model the actual protein ratios and just select the most extreme ratios; (b) learn the distribution from a previous experiment; or (c) assume a standard Cauchy distribution with location 0 and scale 0.1, 0.05, and 0.025, which correspond with $\alpha = 0.05$ roughly to fold changes of 4, 2, and 1.5.

A Cauchy distribution fits accurately this type of random protein ratio distribution: Cauchy is displayed in red, Gaussian in blue. In the case of `ibspiked_set1`, the many 1:1 proteins provide us with adequate data to learn the random protein ratio distribution, however only of the *technical* variation.

```
> #protein.ratios <- proteinRatios(ibspiked_set1,noise.model)
> protein.ratiodistr.wn <- fitWeightedNorm(protein.ratios[,'lratio'],
+                                          weights=1/protein.ratios[,'variance'])
> protein.ratiodistr.cauchy <- fitCauchy(protein.ratios[,"lratio"])

> library(distr) # required library
> limits=seq(from=-0.5,to=0.5,by=0.001)
> curve.wn <- data.frame(x=limits,y=d(protein.ratiodistr.wn)(limits))
> curve.cauchy<-data.frame(x=limits,y=d(protein.ratiodistr.cauchy)(limits))
> g <- ggplot(data.frame(protein.ratios),aes(x=lratio)) +
+   geom_histogram(colour = "darkgreen", fill = "white",aes(y=..density..),
+                  binwidth=0.02) + geom_rug() +
+   geom_line(data=curve.wn,aes(x=x,y=y),colour="blue") +
+   geom_line(data=curve.cauchy,aes(x=x,y=y),colour="red")
> print(g)
```

Now, when supplying a `ratiodistr` parameter to `estimateRatio` and `proteinRatios`, sample and signal p-values are calculated, what we illustrate in the code below

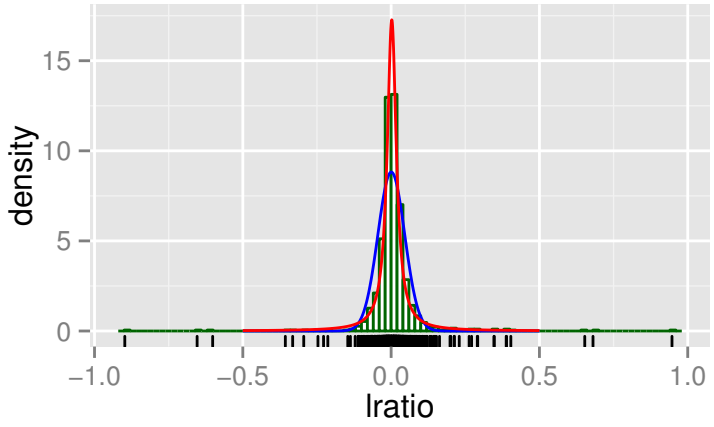


Figure 1.4: Histogram of all protein ratios in `ibspiked_set1`. A fit with a Gaussian and Cauchy probability density function is shown in blue and red, respectively.

```
> rat.list <-
+   estimateRatio(ibspiked_set1,noise.model=noise.model,channel1="114",channel2="115",
+                 protein=reporterProteins(proteinGroup(ibspiked_set1)),combine=F,
+                 ratiodistr=protein.ratiodistr.cauchy)
> rat.list[rat.list[,"is.significant"]==1,]
```

```
           lratio     variance n.spectra n.na1 n.na2  p.value.rat
P13635  0.2630333 0.0070344050       240     0     0 9.364772e-04
Q61147 -0.2946405 0.0009601071       139     0     0 4.822918e-22
       p.value.sample is.significant
P13635     0.02245261              1
Q61147     0.01973357              1
```

A.3.6. Detection of proteins with no specific peptides

It is well known that MS analysis only reveals the presence of so-called protein groups, defined as sets of proteins identified by the same set of peptides. The protein that contains all the peptides is the group reporter (there are possibly several group reporters) and if it has one specific peptide at least then its presence in the sample is certain. The status of the other proteins in the group is in general impossible to determine. When quantitative information is available, there is a potential to elucidate the structure of part of the protein groups.

In the example below, a subset `IBSpectra` object is created, containing only peptides shared between CERU_RAT and CERU_MOUSE, and those specific to CERU_RAT.

```
> ## peptides shared between CERU_RAT and CERU_MOUSE have been computed before
> pep.shared

 [1] "AGLQAFFQVR"      "DNEEFLESNK"      "DTANLFPHK"       "EMGPTYADPVCLSK"
 [5] "ETFTYEWTVPK"     "GSLLADGR"        "KGSLLADGR"       "LYHSHVDAPK"
 [9] "NMATRPYSLHAHGVK" "RDTANLFPHK"      "VFFEQGATR"

> ## peptides specific to CERU_RAT
> pep.rat <- peptides(proteinGroup(ibspiked_set1),protein=ceru.rat,
+                     specificity="reporter-specific")
> ## create an IBSpectra object with only CERU_RAT and shared peptides
> ib.subset <- subsetIBSpectra(ibspiked_set1,
+                              peptide=c(pep.rat,pep.shared),direction="include")
> ## calculate shared ratios
> sr <- shared.ratios(ib.subset,noise.model,
+                     channel1="114",channel2="117",
+                     ratiodistr=protein.ratiodistr.cauchy)
> sr

       reporter.protein protein2   ratio1 ratio1.var n.spectra.1        ratio2
lratio           P13635   Q61147 0.946961 0.01468257         241 -6.172894e-06
        ratio2.var n.spectra.2
lratio 0.001755512         275

>

>   ## plot significantly different protein groups where 90% CI does not overlap
>   ## CERU_MOUSE and CERU_RAT is detected, as expected.
>   shared.ratios.sign(sr,z.shared=1.282)

    reporter.protein protein2 n.spectra.1 n.spectra.2          proteins
1.1           P13635   Q61147         241         275 P13635 \nvs Q61147
1.2           P13635   Q61147         241         275 P13635 \nvs Q61147
          g         ratio         var n.spectra id
1.1 reporter  9.469610e-01 0.014682575      > 10  1
1.2   member -6.172894e-06 0.001755512      > 10  1
```

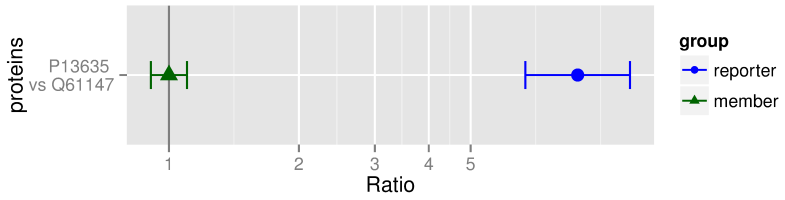


Figure 1.5: Peptides of spiked ceruplasmins have significantly different ratios between groups. Group *reporter* consists of peptides specific to `CERU_RAT (P13635)`, group *member* are peptides shared between `CERU_RAT` and `CERU_MOUSE (Q61147)`.

A.4. Report generation

`isobar` provides a rich interface for creating Excel and PDF reports for further analysis and quality control. The main entry function is `create.reports`. Alternatively the Rscript `create_reports.R` can be used. It is located in the `report` folder of the `isobar` installation, and reads the properties from a file in the working directory.

The posible values are defined in the `report/properties.R` file in the isobar installation. To generate a report with standard properties the following code should do the trick:

```
> create.reports(type="iTRAQ4plexSpectra",
+                identifications="my.id.csv",peaklist="my.mgf")
```

The properties can also be defined in a `properties.R` file which is located in the working directory. The properties are set in the following order:

- 'global' properties in `ISOBAR-DIRECTORY/report/properties.R`[2]
- 'local' properies in `WORKING-DIRECTORY/properties.R`
- command line arguments to `create_reports.R` or `create.reports` function

Appendix A.5.2 provides a syntax-highlighted version of the properties file supplied with isobar, which sets the default parameters and provides some help in the comments. The number of paramters which can be set may seem a lot at first, however most times only a few are needed.

For successful completion, LaTeX- for the PDF reports - and Perl - for the Excel reports - need to be installed.

[2] located in `system.file('report','properties.R',package='isobar')`

A.4.1. Files used for report generation

```
> ## execute to find the path and file location in your installation.
> system.file("report",package="isobar") ## path
> list.files(system.file("report",package="isobar")) ## files
```

create_reports.R R script which can be used to create QC and PDF reports It initializes the environment, reads properties and calls `Sweave` on QC and DA Sweave files. Additionally it generates a Excel data analysis report by calling `tab2xls.pl`.

isobar-qc.Rnw Sweave file with quality control plots.

isobar-analysis.Rnw Sweave file for generating a data analysis report with the list of all protein ratios and list of significantly different proteins.

properties.R Default configuration for `create_reports.R`. It is parsed as R code.

report-utils.R Helper R functions used in Sweave documents.

report-utils.tex Helper LaTeX functions used in Sweave documents.

A.5. Appendix

A.5.1. File formats

ID CSV file format

The Perl parsers create ID CSV files - identification information for all matched spectra without quantitative information. You can create your own parser, the resulting file should be tab-delimited and contain the following columns. Only bold columns are obligatory. The information is redundant - that means if a peptide may stem from two different proteins the information of the identification is repeated.

accession	Protein AC
peptide	Peptide sequence
modif	Peptide modification string
charge	Charge state
theo.mass	Theoretical peptide mass
exp.mass	Experimentally observed mass
parent.intens	Parent intensity
retention.time	Retention time
spectrum	Spectrum identifier
search.engine	Protein search engine and score

IBSpectra CSV file format

IBSpectra file format has the same columns as the ID CSV format and additionally columns containing the quantitation information, namely `X`*tagname*`_mass` and `X`*tagname*`_ions`, for mass and intensity of each tag *tagname*. Below an example of the further columns for an `iTRAQ 4plex` IBSpectra.

X114_mass	reporter ion mass
X115_mass	reporter ion mass
X116_mass	reporter ion mass
X117_mass	reporter ion mass
X114_ions	reporter ion intensity
X115_ions	reporter ion intensity
X116_ions	reporter ion intensity
X117_ions	reporter ion intensity

A.5.2. `properties.R` for report generation

```
##
## Isobar properties.R file
##     for automatic report generation
##
## It is standard R code and parsed using sys.source

#######################################################################
## General properties

## Report type: Either 'protein' or 'peptide'
# report.level="peptide"
```

```
report.level="protein"
#attr(report.level,"allowed.values") <- c("protein","peptide")

## Isobaric tagging type. Use one of the following:
# type='iTRAQ4plexSpectra'
# type='iTRAQ8plexSpectra'
# type='TMT2plexSpectra'
# type='TMT6plexSpectra'
type=NULL
#attr(type,"allowed.values") <- IBSpectraTypes()

isotope.impurities=NULL
correct.isotope.impurities=TRUE

## Name of project, by default the name of working directory
## Will be title and author of the analysis reports.
name=basename(getwd())
author=paste0("isobar␣R␣package␣v",packageDescription("isobar")$Version)

## specifes the IBSpectra file or object
##   - can be a data.frame (e.g. ibspectra=as.data.frame(ibspiked_set1) )
##   - if it is a character string, it is assumed to be a file
##     - if it ends on .rda, then it is assumed to be a R data object
##     - if it does not exists, then it is may generated based on
##       the peaklist and identifications properties
ibspectra=paste(name,"ibspectra.csv",sep=".")

## When replicates or 'samples belonging together' are analyzed, a
## ProteinGroup object based on all data should be constructed
## beforehand. This then acts as a template and a subset is used.
protein.group.template=NULL

## Via database or internet connection, informations on proteins (such
## as gene names and length) can be gathered. protein.info.f defines
## the function which takes a ProteinGroup object as argument
protein.info.f=getProteinInfoFromTheInternet

## Where should cached files be saved? Will be created if it does not
## exist
# cachedir="."
cachedir="cache"
## Regenerate cache files? By default, chache files are used.
regen=FALSE

## An ibspectra object can be generated from peaklists and
## identifications.

## peaklist files for quantitation, by default all mgf file in
## directory
peaklist=list.files(pattern="*\\.mgf$")
## id files, by default all id.csv files in directory
identifications=list.files(pattern="*\\.id.csv$")
## mapping files, for data quantified and identified with different but
## correspoding spectra. For example corresponding HCD-CID files.

## masses and intensities which are outside of the 'true' tag mass
## +/- fragment.precision/2 are discarded
```

```
fragment.precision=0.01
## filter mass outliers
fragment.outlier.prob=0.001

## Additional arguments of readIBSpectra can be set here
##  decode.titles should be set to TRUE for Mascot search results
##  as Mascot encodes the spectrum title (e.g. space -> %20)
readIBSpectra.args = list(
    mapping.file=NULL,
    decode.titles=FALSE
)

######################################################################
## Quantification properties

normalize=TRUE
# if defined, normalize.factors will be used for normalization
normalize.factors=NULL
normalize.channels=NULL
normalize.use.protein=NULL
normalize.exclude.protein=NULL
normalize.function=median
normalize.na.rm=FALSE

peptide.specificity=REPORTERSPECIFIC

use.na=FALSE

## the parameter noise.model can be either a NoiseModel object or a file name
data(noise.model.hcd)
noise.model=noise.model.hcd
## If it is a file name, a noise model is estimated as non one-to-one
## and saved into the file. otherwise, the noise model is loaded from
## the file
# noise.model="noise.model.rda"

## Define channels for creation of a noise model, ideally a set of
##  channels which are technical replicates.
noise.model.channels=NULL

## If noise.model.is.technicalreplicates is FALSE, the intensities
##   are normalized for protein means, creating artifical technical
##   replicates. For this procedure, only proteins with more than
##   noise.model.minspectra are considered.
noise.model.is.technicalreplicates=FALSE
noise.model.minspectra=50

## Class definitions of the isobaric tag channels.
## A character vector with the same length as channels
##   (e.g. 4 for iTRAQ 4plex, 6 for TMT 6plex)
## Example for iTRAQ 4plex:
# class.labels=as.character(c(1,0,0,0))
# class.labels=c("Treatment","Treatment","Control","Control")
## Also names are possible - these serves as description in the report
##  and less space is used in the rows
# class.labels=c("Treatment"="T","Treatment"="T","Control"="C","Control"="C")
class.labels=NULL
```

```
## The following definitions define which ratios are calculated.

## summarize ratios with equal class labels, set to TRUE when replicates are used
summarize=FALSE

## combn.method defines which ratios are calculated - versus a channel or a class,
##   all the ratios within or across classes, or all possible combinatioins.
## When summarize=TRUE is set, use "interclass", "versus.class", or "intraclass"
# combn.method="global"
# combn.method="versus.class"
# combn.method="intraclass"
# combn.method="interclass"
combn.method="versus.channel"
vs.class=NULL

cmbn=NULL

## Arguments given to 'proteinRatios' function. See ?proteinRatios
ratios.opts = list(
    sign.level.sample=0.05,
    sign.level.rat=0.05,
    groupspecific.if.same.ac=TRUE)

quant.w.grouppeptides=c()

min.detect=NULL

preselected=c()


### Biological Variability Ratio Distribution options
## ratiodistr can be set to a file or a 'Distribution object. ' If
##  NULL, or the specified file is not existent, the biological
##  variability of ratios is estimated on the sample at hand and
##  written to cachedir/ratiodistr.rda or the specified file.
ratiodistr=NULL

## Ideally, when the biological variability is estimated for the
## sample at hand, a biological replicate is present (/ie/ same class
## defined in class labels).  Classes can also be assigned just for
## estimation of the ratio distribution, /eg/ to choose biologically
## very similar samples as pseudo replicates.
ratiodistr.class.labels=NULL

## Function for fitting. Available: fitCauchy, fitTlsd
ratiodistr.fitting.f=fitCauchy

## Use symetrical ratios - i.e. for every ratio r add a ratio -r
##   prior to fitting of a distribution
ratiodistr.symmetry=TRUE

## If defined, use z-score instead of ratio distribution
# zscore.threshold=2.5
zscore.threshold=NULL

#########################################################################
```

```
## PTM properties

## PhosphoSitePlus dataset which can be used to annotate known
## modification sites. Download site:
## http://www.phosphosite.org/staticDownloads.do
phosphosite.dataset <- NULL

## Modification to track. Use 'PHOS' for phosphorylation.
# ptm <- c('ACET','METH','UBI','SUMO', 'PHOS')
ptm <- NULL

## file name of rda or data.frame with known modification sites
## gathered with ptm.info.f.  defaults to 'cachedir/ptm.info.rda'
ptm.info <- NULL

## Function to get PTM modification sites from public datasets
# ptm.info.f <- getPtmInfoFromNextprot
# ptm.info.f <- function(...)
#    getPtmInfoFromPhosphoSitePlus(...,modification="PHOS")
# ptm.info.f <- function(...)
#    getPtmInfoFromPhosphoSitePlus(...,modification=ptm)
ptm.info.f <- getPtmInfoFromNextprot

## A protein quantification data.frame (generated with
##  'proteinRatios').  The ratio and variance are used to correct the
##  observed modified peptide ratios Needs to have the experimental
##  setup as the modified peptide experiment
correct.peptide.ratios.with <- NULL

## The correlation between peptide and protein ratios defines the
## covariance

##  Var(ratio m) = Var(ratio mp) + Var(ratio p)
##                                   + 2 * Cov(ratio mp, ratio p),
##  Cov(ratio mp, ratio p) = 2 * cor * Sd(ratio mp) * Sd(ratio p),
##  with m = modifcation, mp = modified peptide, p = protein
peptide.protein.correlation <- 0

## quantification table whose columns are attached to the XLS
## quantification table
compare.to.quant <- NULL

#########################################################################
## Report properties

write.qc.report=TRUE
write.report=TRUE
write.xls.report=TRUE

## Use name for report, ie NAME.quant.xlsx instead of
## isobar-analysis.xlsx
use.name.for.report=TRUE

## PDF Analysis report sections: Significant proteins and protein
## details
show.significant.proteins=FALSE
show.protein.details=TRUE
```

```
### QC REPORT OPTIONS ###
#qc.maplot.pairs=FALSE # plot one MA plot per tag (versus all others)
qc.maplot.pairs=TRUE # plot MA plot of each tag versus each tag

### XLS REPORT OPTIONS ###
## Spreadsheet format: Either 'xlsx' or 'xls'
# spreadsheet.format="xlsx"
spreadsheet.format="xlsx"

## XLS report format 'wide' or 'long '.

## 'wide' format outputs ratios in separate columns of the same record
## (i.e. one line per protein)
## 'long' format outputs ratios in separate records (i.e. one line per
## ratio)
# xls.report.format="wide"
xls.report.format="long"

## XLS report columns in quantification tab
##  possible values: ratio, is.significant, CI95.lower, CI95.upper,
##                   ratio.minus.sd, ratio.plus.sd,
##                   p.value.ratio, p.value.sample, n.na1, n.na2,
##                   log10.ratio, log10.variance,
##                   log2.ratio, log2.variance
##  only for summarize=TRUE: n.pos, n.neg
xls.report.columns <- c("ratio","is.significant","ratio.minus.sd",
                        "ratio.plus.sd","p.value.ratio","p.value.sample",
                        "log10.ratio","log10.variance")

########################################################################
## Etc

sum.intensities=FALSE

datbase="Uniprot"

scratch=list()

##
#  compile LaTeX reports into PDF files
compile=TRUE

# zip final report files into archive
zip=FALSE

# warning level (see 'warn' in ?options)
warning.level=1
```

A.5.3. Dependencies

LaTeX and PGF/TikZ

LaTeX is a high-quality typesetting system; it includes features designed for the production of technical and scientific documentation. It is available as free software[3]. PGF is a TeX macro package for generating graphics It comes with a user-friedly syntax layer called TikZ[4].

LaTeX is used for creating PDF analysis reports, with the PGF package creating the graphics. Go to http://www.latex-project.org to get information on how to download and install a LaTeX system and packages.

Perl

Perl is a high-level, general-purpose, interpreted, dynamic programming language. Perl is required for two tasks:

- Conversion of Pidres XML and Mascot DAT files to ID CSV format;
- Creation of Microsoft Excel format data analysis report.

Go to http://www.perl.org to download and get help on the installation of Perl on your Operating System. For file format conversion, perl module `Statistics::Lite` is required. For Excel export `Spreadsheet::WriteExcel`. All Perl scripts are in the subdirectory `pl` of the isobar package installation.

```
> ## execute to find the path and file location in your installation.
> system.file("pl",package="isobar") ## path
> list.files(system.file("pl",package="isobar")) ## files
```

`mascotParser2.pl` and `pidresParser2.pl` convert from respective protein search outputfiles to a XML file format, which can be converted into a CSV file readable by *isobar* by using `psx2tab2.pl`.

`mascotParser2.pl` coverts from Mascot format, and requires the file `modifconv.csv` as a definition of modification names. `pidresParser2.pl` converts from Phenyx output and requires the file `parsersConfig.xml`. `tab2xls.pl` converts csv file to different sheets of an Excel spreadsheet.

[3] http://www.latex-project.org
[4] http://sourceforge.net/projects/pgf

```
> ## execute on your system
> system(paste("perl",system.file("pl","mascotParser2.pl",package="isobar"),
+              "--help"))
> print(paste("perl",system.file("pl","pidresParser2.pl",package="isobar"),
+              "--help"))
```

B. Curriculum Vitae

Name: Florian Paul Breitwieser

Address: Hofferplatz 4, 1160 Wien, AUSTRIA

E-mail: florian.bw@gmail.com

EDUCATION

CeMM - Research Center for Molecular Medicine / Medical University of Vienna

Doctoral Student at Bioinformatics Dept., 2009 - 2014

Dissertation: *'Computational Approaches for Quantifying Proteins and Posttranslational Modifications from Labeled Mass Spectrometry Data'*

Advisor: Jacques Colinge

Upper Austrian University of Applied Sciences

DI (FH) in Bioinformatics, 2007, passed with high distinction

Diploma thesis: *'Genetic Variation in Protein Biomarkers'* (conducted at the University of New South Wales, Sydney)

Supervisors: Marc Wilkins and Karin Pröll

WORK EXPERIENCE

since 10/2009

Predoc at CeMM - Research Center for Molecular Medicine, Vienna, Austria.

Main project on protein and PTM quantitation of iTRAQ/TMT datasets. Involved in several expression, modification, and interaction proteomics experiments.

09/2008-09/2009

Bioinformatician at CeMM - Research Center for Molecular Medicine, Vienna, Austria.

Developed tools, algorithms and databases for proteomics and protein interactomics. Maintained the mass spectrometry analysis pipeline.

09/2007-05/2008

Developer at Centro International de Agricultura Tropical (CIAT), Cali, Colombia.

Developed of a web platform for project management using Java JCR and a repository for sharing of scientific documents.

09/2006-06/2007

Undergraduate thesis work at University of New South Wales, Australia. Researched the variability and heritability of gene expression (in 12 families / 330 individuals, microarray data) of proteins which have been reported as biomarkers for cancer.

10/2005-06/2006

Project manager of study project for biotech startup Anagnostics. Planned and organized the development of a tool for generating scripts controlling a novel analyzer using cylindrical DNA or protein microarrays.

Awards and Honors

2012 Full AUPA grant for FEBS summer school on high performance proteomics

2012 FWF Grant for Doctoral Thesis Project

2011 Student Travel Fellowship for BOSC conference

2007 First class honors at graduation at Upper Austrian University of Applied Sciences

2004 Scholarship for “Best bioinformatics student” in year 2003/2004

C. List of Publications and Presentations

Peer-reviewed Articles

2014 Jacques Colinge, Adrián César-Razquin, Kilian Huber, Florian P. Breitwieser, Peter Májek, and Giulio Superti-Furga (Apr. 2014). "Building and exploring an integrated human kinase network: Global organization and medical entry points." eng. In: *J Proteomics*

2014 Nicole Boucheron, Roland Tschismarov, Lisa Goeschl, Mirjam A. Moser, Sabine Lagger, Shinya Sakaguchi, Mircea Winter, Florian Lenz, Dijana Vitko, Florian P. Breitwieser, Lena Müller, Hammad Hassan, Keiryn L. Bennett, Jacques Colinge, Wolfgang Schreiner, Takeshi Egawa, Ichiro Taniuchi, Patrick Matthias, Christian Seiser, and Wilfried Ellmeier (Mar. 2014). "CD4(+) T cell lineage integrity is controlled by the histone deacetylases HDAC1 and HDAC2." eng. In: *Nat Immunol*

2014 Margarita Maurer, André C. Müller, Katja Parapatics, Winfried F. Pickl, Christine Wagner, Elena L. Rudashevskaya, Florian P. Breitwieser, Jacques Colinge, Kanika Garg, Johannes Griss, Keiryn L. Bennett, and Stephan N. Wagner (June 2014). "Comprehensive comparative and semiquantitative proteome of a very low number of native and matched epstein-barr-virus-transformed B lymphocytes infiltrating human melanoma." eng. In: *J Proteome Res* 13.6, pp. 2830–2845

2013 Florian P. Breitwieser and Jacques Colinge (Sept. 2013). "Isobar(PTM): a software tool for the quantitative analysis of post-translationally modified proteins." eng. In: *J Proteomics* 90, pp. 77–84

2013 V. Borgdorff, U. Rix, G. E. Winter, M. Gridling, A. C. Müller, F. P. Breitwieser, C. Wagner, J. Colinge, K. L. Bennett, G. Superti-Furga, and S. N. Wagner (June 2013). "A chemical biology approach identifies AMPK as a modulator of melanoma oncogene MITF.". eng. In: *Oncogene*

2013 Elena L. Rudashevskaya, Florian P. Breitwieser, Marie L. Huber, Jacques Colinge, André C. Müller, and Keiryn L. Bennett (Feb. 2013). "Multiple and sequential data acquisition method: an improved method for fragmentation and detection of cross-linked peptides on a hybrid linear trap quadrupole Orbitrap Velos mass spectrometer." eng. In: *Anal Chem* 85.3, pp. 1454–1461

2013 Andreas Pollreisz, Marion Funk, Florian P. Breitwieser, Katja Parapatics, Stefan Sacu, Michael Georgopoulos, Roman Dunavoelgyi, Gerhard J. Zlabinger, Jacques Colinge, Keiryn L. Bennett, and Ursula Schmidt-Erfurth (Mar. 2013). "Quantitative proteomics of aqueous and vitreous fluid from patients with idiopathic epiretinal membranes." eng. In: *Exp Eye Res* 108, pp. 48–58

2012 André C. Müller, Florian P. Breitwieser, Heinz Fischer, Christopher Schuster, Oliver Brandt, Jacques Colinge, Giulio Superti-Furga, Georg Stingl, Adelheid Elbe-Bürger, and Keiryn L. Bennett (July 2012). "A comparative proteomic study of human skin suction blister fluid from healthy individuals using immunodepletion and iTRAQ labeling." eng. In: *J Proteome Res* 11.7, pp. 3715–3727

2012 Georg E. Winter, Uwe Rix, Scott M. Carlson, Karoline V. Gleixner, Florian Grebien, Manuela Gridling, André C. Müller, Florian P. Breitwieser, Martin Bilban, Jacques Colinge, Peter Valent, Keiryn L. Bennett, Forest M. White, and Giulio Superti-Furga (Nov. 2012). "Systems-pharmacology dissection of a drug synergy in imatinib-resistant CML.". eng. In: *Nat Chem Biol* 8.11, pp. 905–912

2011 Florian P. Breitwieser, André Müller, Loïc Dayon, Thomas Köcher, Alexandre Hainard, Peter Pichler, Ursula Schmidt-Erfurth, Giulio Superti-Furga, Jean-Charles Sanchez, Karl Mechtler, Keiryn L. Bennett, and Jacques Colinge (June 2011). "General statistical modeling of data from protein relative expression isobaric tags." eng. In: *J Proteome Res* 10.6, pp. 2758–2766

2011 Keiryn L. Bennett, Marion Funk, Marion Tschernutter, Florian P. Breitwieser, Melanie Planyavsky, Ceereena Ubaida Mohien, André Müller, Zlatko Trajanoski, Jacques Colinge, Giulio Superti-Furga, and Ursula Schmidt-Erfurth (Feb. 2011). "Proteomic analysis of human cataract aqueous humour: Comparison of one-dimensional gel LCMS with two-dimensional LCMS of unlabelled and iTRAQ®-labelled specimens." eng. In: *J Proteomics* 74.2, pp. 151–166

2011 Thomas R. Burkard, Melanie Planyavsky, Ines Kaupe, Florian P. Breitwieser, Tilmann Bürckstümmer, Keiryn L. Bennett, Giulio Superti-Furga, and Jacques Colinge (2011). "Initial characterization of the human central proteome." eng. In: *BMC Syst Biol* 5, p. 17

2010 Ceereena Ubaida Mohien, Jürgen Hartler, Florian Breitwieser, Uwe Rix, Lily Remsing Rix, Georg E. Winter, Gerhard G. Thallinger, Keiryn L. Bennett, Giulio Superti-Furga, Zlatko Trajanoski, and Jacques Colinge (July 2010). "MASPECTRAS 2: An integration and analysis platform for proteomic data." eng. In: *Proteomics* 10.14, pp. 2719–2722

2010 U. Rix, L. L. Remsing Rix, A. S. Terker, N. V. Fernbach, O. Hantschel, M. Planyavsky, F. P. Breitwieser, H. Herrmann, J. Colinge, K. L. Bennett, M. Augustin, J. H. Till, M. C. Heinrich, P. Valent, and G. Superti-Furga (Jan. 2010). "A comprehensive target selectivity survey of the BCR-ABL kinase inhibitor INNO-406 by kinase profiling and chemical proteomics in chronic myeloid leukemia cells." eng. In: *Leukemia* 24.1, pp. 44–50

2010 Thomas R. Burkard, Uwe Rix, Florian P. Breitwieser, Giulio Superti-Furga, and Jacques Colinge (2010). "A computational approach to analyze the mechanism of action of the kinase inhibitor bafetinib." eng. In: *PLoS Comput Biol* 6.11, e1001001

2009 Nora V. Fernbach, Melanie Planyavsky, André Müller, Florian P. Breitwieser, Jacques Colinge, Uwe Rix, and Keiryn L. Bennett (Oct. 2009). "Acid elution and one-dimensional shotgun analysis on an Orbitrap mass spectrometer: an application to drug affinity chromatography." eng. In: *J Proteome Res* 8.10, pp. 4753–4765

Book Chapters

2012 Florian P. Breitwieser and Jacques Colinge (2012a). "Analysis of Labeled Quantitative Mass Spectrometry Proteomics Data". In: *Computational Medicine*. Ed. by Zlatko Trajanoski. Springer Vienna, pp. 79–91

Conference Presentations **(All talks were given by the author)**

2013 Florian P. Breitwieser (Feb. 2013). *isobar: Quantitative Analysis of Protein and PTM iTRAQ/TMT data*. Presented at IMBA Impromptu Seminar (invited talk). Vienna, Austria

2012 Florian P. Breitwieser and Jacques Colinge (2012b). *isobar: Quantifying changes of the proteome and its post-translational modifications*. Presented at the 9th Siena Meeting: From Genome to Proteome. Siena, Italy

2012 Florian P. Breitwieser (2012). *Statistical Modeling of Post-translational Protein Regulation Dynamics*. Presented at the Young Investigators Day. Faculty of Computational Life Sciences, University of Vienna

2011 Florian P. Breitwieser and Jacques Colinge (2012c). isobar *R package for the analysis of quantitative proteomics data*. Presented at the 12th Annual Bioinformatics Open Source Conference. Vienna, Austria

2010 Florian P. Breitwieser, André Müller, Giulio Superti-Furga, Keiryn L. Bennett, and Jacques Colinge (2010). *Statistical Models for Quantitative Proteomics using Isobaric Tags*. Presented at the 4th Central and Eastern European Proteomics Conference. Vienna, Austria