Context-based analysis of mass spectrometry proteomics data

Constantin Ammar

München 2020

Context-based analysis of mass spectrometry proteomics data

Constantin Ammar

Dissertation
an der Fakultät für Mathematik, Informatik
und Statistik
der Ludwig–Maximilians–Universität
München

eingereicht von
Constantin Ammar
aus München

München, den 11.08.2020

1. Gutachter: Prof. Dr. Ralf Zimmer
2. Gutachter: Prof. Dr. Oliver Kohlbacher
Tag der mündlichen Prüfung: 09.11.2020

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Ammar, Constantin

Name, Vorname

München, 11.08.2020

Ort, Datum

Constantin Ammar

Unterschrift Doktorand/in

Formular 3.2

Contents

1 Introduction **1**
- 1.1 Bottom-up mass spectrometry-based proteomics . . . 2
- 1.2 Essential steps in MS proteomics data analysis . . . 4
- 1.3 Finding contexts . . . 7

2 Detecting peptides in MS proteomics data **9**
- 2.1 Abstract . . . 11
- 2.2 Introduction . . . 11
- 2.3 Experimental section . . . 13
  - 2.3.1 Proteome analysis using a Q-ToF MS . . . 13
  - 2.3.2 Data analysis of data-dependent LC-MS/MS experiments . . . 14
  - 2.3.3 Library generation settings for the in-house data set . . . 15
  - 2.3.4 Selection of processed fragmentation spectra . . . 15
  - 2.3.5 Import of raw fragmentation spectra . . . 15
  - 2.3.6 Assessment of the similarity of fragmentation spectra . . . 16
  - 2.3.7 Choosing a global similarity threshold . . . 16
  - 2.3.8 Centroid clustering and CIPs . . . 17
  - 2.3.9 Spectral coverage . . . 17
  - 2.3.10 Comparison to custom-made spectral libraries . . . 17
  - 2.3.11 Comparison with SpectraST . . . 18
  - 2.3.12 Benchmarking via cross validation . . . 19
  - 2.3.13 Processing of targeted LC-MS/MS runs for CE and isolation window study . . . 19
  - 2.3.14 Benchmarking spectral library performance on DIA data . . . 19
- 2.4 Results . . . 20
  - 2.4.1 Spectral variability is widespread over experimental conditions . . . 20
  - 2.4.2 Usage of MCIPs yields almost complete spectral coverage . . . 23
  - 2.4.3 MCIP library performs comparable to and enhances custom-made libraries . . . 24
  - 2.4.4 Direct comparison with SpectraST shows significantly increased sensitivity . . . 26
  - 2.4.5 MCIPs increase sensitivity without affecting specificity . . . 26
  - 2.4.6 Analysis of a same-sample DDA and SWATH data Set . . . 27
  - 2.4.7 Comparing the MCIP approach and SpectraST on a SWATH data set 28
- 2.5 Discussion . . . 28

3 Detecting regulated proteins in MS-proteomics data 33
3.1 Abstract . . . 35
3.2 Introduction . . . 36
3.3 Methods . . . 38
3.3.1 MS-EmpiRe compared to state-of-the-art approaches . . . 38
3.3.2 Normalization . . . 38
3.3.3 Empirical error distributions . . . 40
3.3.4 Merging scores over replicates . . . 41
3.3.5 Correcting for outlier measurements . . . 42
3.3.6 Correcting for outlier peptides . . . 44
3.3.7 Combining the peptide scores . . . 44
3.3.8 Re-processing of the proteome wide benchmarking dataset . . . 44
3.3.9 Processing of the different proteomics studies . . . 45
3.3.10 Filtering of the benchmarking dataset . . . 45
3.3.11 In silico benchmarking . . . 46
3.4 Results and Discussion . . . 47
3.4.1 Fold change based normalization reveals structure of the benchmarking dataset . . . 47
3.4.2 Assessment of empirical error distributions underlines the importance of context dependent fold changes . . . 48
3.4.3 MS-EmpiRe shows up to 121% sensitivity increase in an experimental benchmarking set . . . 49
3.4.4 MS-EmpiRe identifies up to 1,200 additional significant proteins in quantitative MS datasets . . . 50
3.4.5 Set-based comparison reveals strong differences between significant proteins for each method . . . 52
3.4.6 A detail view on the quantitative data validates the proteins called by MS-EmpiRe . . . 52
3.4.7 In silico benchmarking shows high sensitivity and conservative FDR estimation . . . 55
3.5 Conclusion . . . 57
3.6 Data availability . . . 58

4 Detecting differential alternative splicing in MS proteomics data 59
4.1 Abstract . . . 61
4.2 Introduction . . . 61
4.3 Results . . . 65
4.3.1 Benchmarking . . . 65
4.3.2 Analysis of a clinical dataset . . . 65
4.4 Discussion & Conclusion . . . 71
4.5 Methods . . . 72
4.5.1 Simulating non spliced proteins . . . 72
4.5.2 Benchmarking on the technical datasets . . . 72
4.5.3 Benchmarking on the *E. coli* datasets . . . 72
4.5.4 Preprocessing the CPTAC data set . . . 73

4.5.5 Data normalization . . . 73
4.5.6 Mapping peptides to isoforms . . . 73
4.5.7 Calculation of FCFCs . . . 74
4.5.8 Generation of empirical FCFC error distributions . . . 74
4.5.9 Combination of FCFCs . . . 75
4.5.10 GO enrichment . . . 75
4.5.11 Data availability . . . 75

5 Detecting relevant proteins for *E. coli* carbon starvation in MS proteomics data 77
5.1 Abstract . . . 79
5.2 Introduction . . . 79
5.3 Results . . . 80
5.3.1 A trade-off between growth rate and death rate across six different growth perturbations . . . 80
5.3.2 A protective survival sector in the proteome . . . 81
5.3.3 Proteomics analysis pipeline . . . 83
5.3.4 Proteomics analysis shows significant enrichment for stress protection and cell envelope . . . 83
5.3.5 Stress response, oxidative damage and catabolism are not limiting survival . . . 86
5.3.6 Survival is sensitive to perturbations of the cell envelope . . . 86
5.3.7 Survival is limited by the mechanical stability of the cell envelope . . . 89
5.3.8 Time-lapse microscopy reveals cell envelope failures in starvation . . . 89
5.4 Discussion and Conclusion . . . 89
5.5 Methods . . . 91
5.5.1 Proteomics data processing . . . 91
5.5.2 Differential expression analysis of proteomics data . . . 91
5.5.3 Scoring the direction of the proteomic response . . . 91
5.5.4 Z-value based ranking of the proteome perturbations . . . 92
5.5.5 Combination of the ranked proteome perturbations . . . 92
5.5.6 Absolute quantification of proteins . . . 92
5.5.7 GO enrichment analyses . . . 93

6 Conclusion and Outlook 95

A Supplement - Detecting differential alternative splicing in MS proteomics data 99
A.1 Supplemental figures . . . 99
A.2 Supplemental text . . . 103

B Supplement - Detecting relevant proteins for *E. coli* carbon starvation in MS proteomics data 105
B.1 Experimental protocols . . . 105
B.1.1 Strains . . . 105

B.1.2 Culture medium 105
B.1.3 Culture conditions 105
B.1.4 Viability measurements 106
B.1.5 Stress conditioning 106
B.1.6 Anaerobic culturing 106
B.1.7 Time-lapse microscopy 107
B.2 Supplemental figures 107

Bibliography **111**

Acknowledgements **129**

List of Figures

1.1 Experimental MS proteomics workflow . . . 3
1.2 MS proteomics data processing . . . 5

2.1 MCIP analysis workflow . . . 14
2.2 MCIP cluster example . . . 18
2.3 MCIP spectral similarity . . . 23
2.4 MCIP DDA comparison . . . 26
2.5 MCIP target decoy . . . 27
2.6 MCIP SWATH comparison . . . 29

3.1 Single linkage clustering for signal normalization . . . 40
3.2 Schematic of the MS-EmpiRe workflow . . . 43
3.3 Experimental setup and fold change based metrics . . . 48
3.4 Assessment of the differential detection performance . . . 51
3.5 Application of MS-EmpiRe, MaxLFQ+t-test and MSqRob to three different quantitative LC-MS/MS datasets . . . 54
3.6 In silico benchmarking of MS-EmpiRe, MaxLFQ+t-test and MSqRob . . . 56

4.1 MS-EmpiReS workflow . . . 62
4.2 Benchmarking and application to a clinical proteomics dataset . . . 67
4.3 Visualization of DAS events for three top scoring genes with important regulatory functions . . . 68

5.1 Physiological relation between growth and death . . . 81
5.2 Data analysis pipeline . . . 82
5.3 Comparison of MS proteomics data from different sources . . . 84
5.4 Correlations of Z-values across CARLS conditions . . . 85
5.5 Effect of pre-stressing on proteome and survival kinetics . . . 87
5.6 Cell envelope integrity is essential for survival . . . 88
5.7 Timelapse microscopy shows envelope failure . . . 90

A.1 Examples of E. coli proteins with inconsistent peptides . . . 100
A.2 Examples of spliced genes with no visible regulation . . . 101
A.3 Pairwise comparison of all samples in the CPTAC dataset . . . 102

B.1 Example growth and death curves of growth perturbations shown in Fig. 1 . . . 109

B.2 LacZ OE shows proteome-wide downregulation 109
B.3 Anaerobic starvation 110

List of Tables

2.1 Spectral dataset overview 13

3.1 The two state-of-the-art differential quantification methods MaxLFQ (with t-test) and MSqRob compared against MS-EmpiRe. 39

4.1 20 of the top ranked DAS genes in the CPTAC data set 67

Abstract

The identification and characterization of proteins in a biological systems can be achieved with current mass spectrometry-based proteomics. A major challenge is the bioinformatics analysis and interpretation of the huge data sets generated. This thesis argues that crucial tasks can be significantly improved via context-based methods, i.e. algorithms exploiting statistical context information or related measurements.

Modern biology is to a large extent facilitated by highly parallelized technologies that are able to capture certain classes of molecules on a system-wide scale. These technologies rely heavily on computational methods both for the processing of measured signals as well as for their biological interpretation. Among the main classes of molecules are proteins, which determine most of the state and function of biological systems. The current method of choice for the system-wide investigation of proteins is based on complex *mass spectrometry* (MS) instrumentation coupled to liquid chromatography. Modern MS proteomics already allows comprehensive and quantitative proteome measurements within minutes. Thus, large clinical cohorts can be measured and even proteomic profiling applications in clinical diagnostics and personalized medicine are emerging. The wealth of proteomics data generated and the different technological setups pose considerable computational challenges to ensure robust and sensitive utilization of the data. In particular, the key challenges in computational proteomics addressed in this thesis are (i) the identification of peptides from raw signals, (ii) the quantification and (iii) identification of differential proteins and isoforms from peptide signals and (iv) the generation of biological hypotheses from the quantified signal measurements and derived statistical features from multiple experiments. A key idea in all methods introduced in this thesis is to utilize contextual information, either by screening many similar data types ((i), (iv)) or generating signal-to-noise contexts from replicate measurements ((ii), (iii)).
In the project on peptide identification (i), we focus on the approach of spectral library searching in the context of novel "Data-Independent-Acquisition" (DIA) proteomics. One of the working hypotheses for DIA algorithms is that the intensity patterns (spectra) leading to identifications of the same peptide are very similar to each other. We perform a systematic evaluation of these similarities over different repositories and find significant differences. We cluster peptide spectra and propose "Multiple Characteristic Intensity Patterns" (MCIP) to represent each peptide. This approach strongly increases the sensitivity.
For differential quantification (ii), we introduce the concept of "Mass Spectrometry analysis using Empirical and Replicate based statistics" (MS-EmpiRe). We introduce empirical signal-to-noise distributions generated from replicate measurements, giving a tailored con-

text for each peptide. Our approach is purely based on peptide fold change information, which strongly reduces the noise. We achieve more than 100% sensitivity increases in benchmarking datasets with strict control of the specificity. In state-of-the-art experimental data sets we see more than 1000 additional proteins in a single comparison and confirm that the detected regulated proteins are also convincing upon detailed inspection of the peptide signals.

To detect differential splicing in proteomics data (iii), we introduce the concept of quantification-based splicing detection to MS proteomics. We show that expanding the concepts of MS-EmpiRe to "fold changes of fold change" setups enables us to find differential abundance changes of protein splice isoforms (MS-EmpiReS). We extensively benchmark MS-EmpiReS and demonstrate its application on a clinical study on colon cancer (≈100 patients), where we substantially increase the detected differentially spliced variants in cancer as compared to state-of-the-art approaches. Moreover, we identify functionally relevant differential splice events on the protein level.

In an experimental collaboration with the Basan Lab at Harvard University we address the biological question (iv), how the proteomic state influences the starvation kinetics of *E. coli* bacteria. We identify from physiological measurements, which regulatory responses are of interest and introduce an approach to combine proteomics data from over 100 MS proteomics measurements from three different experimental repositories. We find that envelope proteins and in particular membrane anchors play a crucial role in the starvation of *E. coli*, a finding of interest for the broad microbiology community that we validated experimentally.

Zusammenfassung

Die Identifikation und Charakterisierung von Proteinen in einem biologischen System kann mit moderner Massenspektrometrie-basierter Proteomik auf systemweiter Skala durchgeführt werden. Eine Herausforderung hierbei ist die bioinformatische Analyse und Interpretation der entstehenden umfangreichen Datensätze. Die vorliegende Dissertation zeigt, dass entscheidende Analyseschritte durch kontextbasierte Methoden, d.h. Algorithmen, die statistische Kontextinformation oder verwandte Messungen nutzen, deutlich verbessert werden können.

Moderne biologische Forschung wird zu einem substantiellen Teil von hochgradig parallelisierten Technologien ermöglicht, die in der Lage sind, bestimmte Klassen von Molekülen auf einer systemweiten Skala zu erfassen. Diese Technologien sind sowohl für die Verarbeitung der gemessenen Signale, als auch für deren biologische Interpretation stark auf computergestützte Methoden angewiesen. Eine der wichtigsten Molekülklassen sind die Proteine, die Zustand und Funktion biologischer Systeme weitgehend bestimmen. Die derzeit führende Methode für die systemweite Untersuchung von Proteinen nutzt komplexe *Massenspektrometrie* (MS) -basierte Instrumente, die mit Flüssigchromatographie gekoppelt sind. Moderne MS Proteomik Technologie ermöglicht derzeit bereits umfassende und quantitative Messungen eines Proteoms innerhalb von Minuten. So können große klinische Kohorten vermessen werden und auch Anwendungen in der personalisierten Medizin und Diagnostik werden hierdurch möglich. Die Fülle der erzeugten Proteomikdaten aus teilweise sehr unterschiedlichen technologischen Konfigurationen stellt eine erhebliche rechnerische Herausforderung dar. Die sinnvolle und verlässliche Interpretation dieser Daten ist Aufgabe der computergestützten Proteomik und die vorliegende Arbeit behandelt essentielle Herausforderungen dieses Forschungsfeldes. Untersucht wird in dieser Arbeit: (i) die Identifizierung von Peptiden aus Rohsignalen, (ii) die Quantifizierung und (iii) die Identifizierung von differentiellen Proteinen und Isoformen aus Peptidsignalen und (iv) die Generierung biologischer Hypothesen aus den quantifizierten Signalmessungen. Eine zentrale Idee bei allen vorgestellten Methoden ist die Nutzung von Kontextinformationen, entweder durch Screening vieler ähnlicher Datentypen ((i), (iv)) oder durch Generierung von Signal-Rausch-Kontexten aus Replikatmessungen ((ii), (iii)).
In dem Projekt zur Peptid Identifizierung (i) konzentrieren wir uns auf die Suche mit Spektral-Bibliotheken im Kontext eines neuen Proteomik-Verfahrens, das als
"Data-Independent-Acquisition" (DIA) bezeichnet wird. Eine der Arbeitshypothesen für DIA-Algorithmen ist, dass die Intensitätsmuster (Spektren), die zur Identifizierung desselben Peptids führen, einander sehr ähnlich sind. Wir führen eine systematische Auswertung

solcher Ähnlichkeiten über verschiedene spektrale Datenbanken hinweg durch und finden teilweise signifikante Unterschiede. Wir clustern die Spektren und schlagen vor, jedes Peptid durch multiple "Characteristic Intensity Patterns" (MCIP) zu repräsentieren. Dieser Ansatz führt zu einer deutlichen Erhöhung der Sensitivität, mit der Peptide identifiziert werden.
Für die differentielle Quantifizierung (ii) führen wir das Konzept der "Massenspektrometrie-Analyse unter Verwendung empirischer und replikatbasierter Statistik" (MS-EmpiRe) ein. Wir generieren empirische Signal-zu-Rausch-Verteilungen aus Replikatmessungen, die einen maßgeschneiderten quantitativen Kontext für jedes Peptid liefern. Unser Ansatz wertet ausschliesslich Abundanzänderungen zwischen Peptiden gleicher Sequenz aus, was zu einer deutlichen Verringerung des Rauschens führt. Wir erreichen mehr als 100% Sensitivitätssteigerungen in benchmarking Datensätzen mit guter Kontrolle der Spezifität. In einzelnen modernen experimentellen Datensätzen sehen wir mehr als 1000 zusätzliche Proteine. Wir validieren diese zusätzlich gefundenen Proteine durch detaillierte Visualisierung der zugrunde liegenden Peptidsignale. Zur Erkennung von differentiellem Spleißen in Proteomikdaten (iii) führen wir das Konzept der quantitativen Spleißerkennung in die MS-Proteomik ein. Wir zeigen, dass die Erweiterung der Konzepte von MS-EmpiRe auf "doppelt-differentielle" Setups es ermöglicht, statistisch signifikante Spleißereignisse in Proteomik-Datensätzen (MS-EmpiReS) zu finden. Wir führen umfangreiche Benchmarkings von MS-EmpiReS durch und wenden unsere Pipeline auf eine klinische Studie zu Dickdarmkrebs (ca. 100 Patienten) an, wo wir funktionell relevante, differentielle Spleißereignisse auf Proteinebene identifizieren.
In einer experimentellen Kollaboration mit dem Labor von Prof. Markus Basan an der Harvard University befassen wir uns mit der biologischen Frage (iv) wie der Proteomzustand die Kinetik hungernder *E. coli* Bakterien beeinflusst. Wir identifizieren charakteristische regulatorische Muster aus physiologischen Messungen und testen anschließend in über 100 MS-Proteomik Messungen aus drei verschiedenen experimentellen Datenbanken, welche Proteine diesem Muster folgen. Dadurch konnte Hüllproteinen und insbesondere Membran-Ankern eine entscheidende Rolle beim Hungern von *E. coli* zugeordnet werden. Wir zeigen mehrere biologische Experimente, die diesen Befund validieren.

Chapter 1

Introduction

"It is by avoiding the rapid decay into the inert state of 'equilibrium' that an organism appears so enigmatic; so much so, that from the earliest times of human thought some special non-physical or supernatural force [..] was claimed to be operative in the organism, and in some quarters is still claimed." [1]

Erwin Schrödinger - 'What is life?'

Standard biology textbooks define living organisms as having the properties of evolutionary adaptation, internal homoeostasis, compartmentalization and organization, metabolism, growth, response to stimuli and reproduction [2]. On the cellular level, the latter six of these seven 'enigmatic' properties, as physicist Erwin Schrödinger put it, are directly mediated by proteins. It is hence no overstatement to claim that the foundations of life are built of proteins.
Consequently, studying and characterizing proteins is essential for a wide variety of questions in many areas of biology. Over the past three decades, technologies have emerged that allow identification and (relative) quantification of thousands of proteins (proteomics) in biological samples. The leading approach for this characterization is based on mass spectrometry (MS). The field has seen a transformation from very basic 'proof of principle' research to the current state, where instruments are commercially available and accessible in laboratories and core facilities around the world. In accordance with these developments, computational and data processing aspects take up an ever increasing fraction of proteomics research. The innovations in technology lead to an unprecedented wealth of available data and computational challenges range from basic identifications of peptides to concrete interpretation of biological processes.
In this thesis, I address computational challenges along this whole range. The following introduction aims at defining core principles of MS proteomics and corresponding computational problems. It will also give a brief overview on how contextual information is used in this thesis to improve the state-of-the-art in computational proteomics research.

1.1 Bottom-up mass spectrometry-based proteomics

Mass spectrometry-based proteomics experiments consist of three main steps: a) Sample preparation, b) Chromatographic separation and ionization and c) Mass spectrometry measurement. Sample preparation (Figure 1.1a) starts with lysing intact cells and subsequently isolating the proteins in the sample. Often, the proteins are digested via proteases such as Trypsin, which preferably cleaves proteins C-terminal of the amino acids Arginine and Lysine [3]. This results in smaller peptides, which, compared to intact proteins, are much easier to handle chemically and in the mass spectrometer. To achieve higher coverage, proteins can be pre-fractionated, for example by *gel electrophoresis*, *strong cation exchange chromatography*, *peptide isoelectric focusing* and *sodium dodecyl sulfate–polyacrylamide gel electrophoresis* (SDS-PAGE) [4, 5]. Each fraction can then be measured in an individual measurement.
The digestion step defines the technique of *bottom-up* proteomics [6], a term stemming from the fact that proteins have to be reconstructed from the smaller peptides. The opposing technique is *top-down* proteomics [7], where intact proteins are directly submitted to the mass spectrometer, retaining valuable information about the intact protein. However, due to the benefits of protease digestion for sample handling, sensitivity and throughput, the vast majority of biological studies is carried out using bottom up proteomics. In the standard setup, peptides are submitted to liquid chromatography (LC) after digestion (Figure 1.1b). In principle, pressurized liquid drags the peptides through a column coated with hydrophobic material. The higher the hydrophobicity of the peptide, the stronger it aligns at the coating. The increased contact increases the *retention time* a peptide needs to cross the column, which is often adjusted to range between 3h and 30min [8]. The LC step hence prevents the whole sample to be submitted to the MS at once (thereby reducing data complexity) and provides valuable retention time information. As usually there are (at least) hundreds of thousands of peptide species in the column [9, 10] a large number of different peptide species reaches the end of the column at any given moment. Peptides that reach the end of the column are channeled through a needle and ionized, using *electrospray ionization* [11]. In principle, a (usually negative) voltage is applied at the entry of the mass spectrometer. This results in positively charged droplets that form in a cascade-like process, as positive charges in the liquid align on the droplet surface, which energetically favors an increase in surface by the formation of smaller droplets up to the single ion. For many peptide species, a substantial fraction of its molecules is ionized this way and becomes 'visible' to the mass spectrometer.

An example mass spectrometry setup is depicted in 1.1c, which contains two mass analyzers and a collision cell. Mass analyzers are electrostatic or electrodynamic components that are able to determine the *mass over charge* (m/z) values of charged particles. Popular types of analyzers are the *linear ion trap* (LIT), *orbitrap* (OT), *quadrupol* (Q) and *time of flight* (TOF) [6]. The OT is an electrostatic analyzer, where ions rotate around a central axis. The rotation frequency and knowledge about the applied electric field allows to determine m/z. For the TOF analyzer, accelerated ions are sent on an elliptical trajectory before landing on a detection plate. This allows to determine the 'time of flight' and consequently m/z. The LIT is a coupled electrodynamic and electrostatic analyzer, where ions can be trapped axially by electric potentials and radially by radio frequency fields. The quadrupol analyzer employs oscillating electric fields. Collision cells usually consist of LITs filled with a colli-

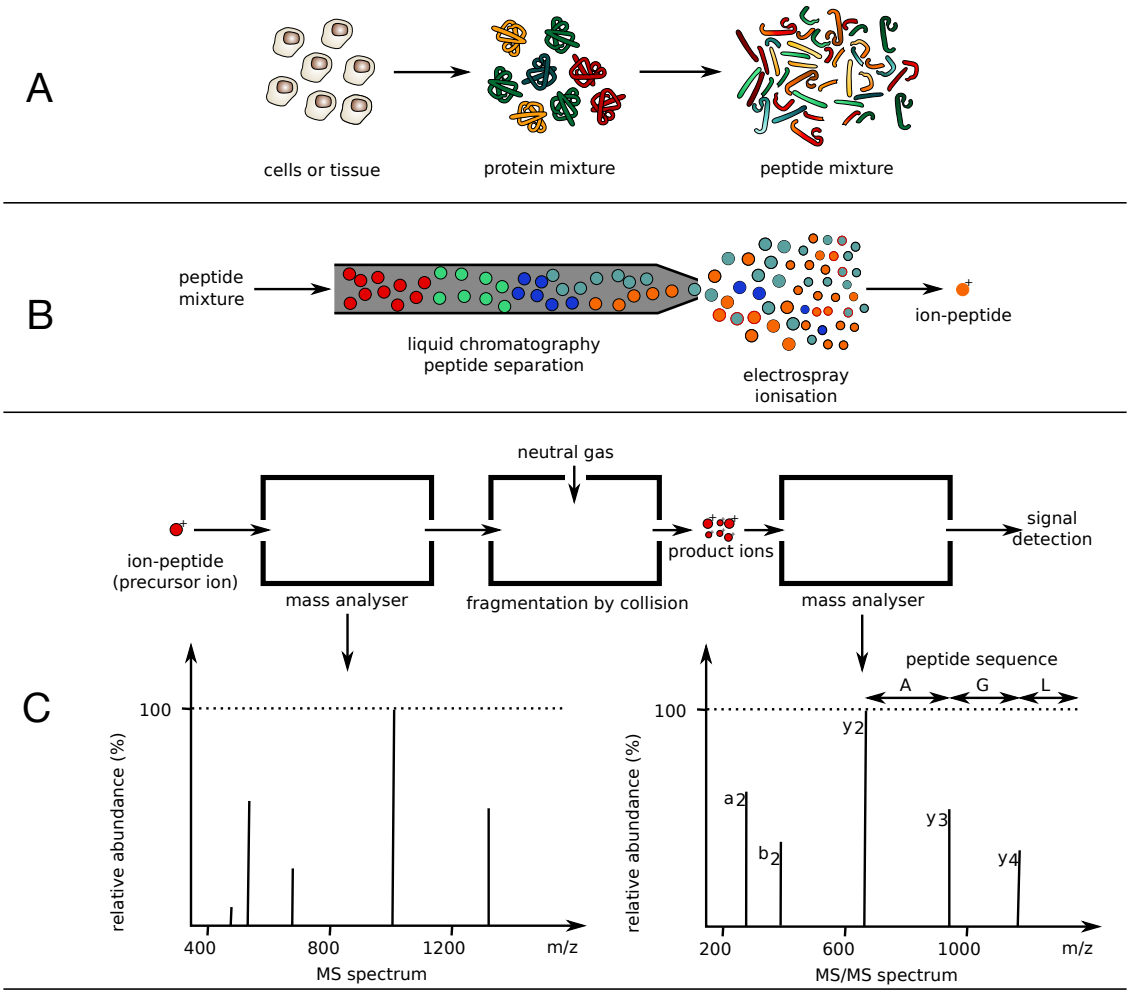


Figure 1.1: Basic steps of proteomics data acquisition. a) For sample preparation, the biological sample is lysed, proteins are extracted and digested into peptides. b) The peptides are then submitted to liquid chromatography where they are separated by hydrophobicity. Peptides that reach the end of the column are ionized by electrospray ionization. c) Ionized peptides are submitted to an MS setup. In a first step the mass and intensities of the intact peptide ions are determined by a mass analyzer. Selected peptide ions are then fragmented and the fragment ions are again measured by a mass analyzer. Image adapted from [12].

sional gas such as helium. The ions are accelarated, collide with neutral gas and fragment [6]. Importantly, the mass analyzers are able to extract ion intensities that scale linearly with ion counts within the *dynamic range*. This allows to precisely determine relative changes. However, as it is not guaranteed that all molecules get ionized and detected, it is generally not possible to determine exact molecule numbers from intensity values [13]. Additionally, masses can be derived from the m/z values with various computational approaches [14]. In the first mass analyzer in Figure 1.1c, the m/z values and corresponding intensities of all peptide ions are acquired, resulting in an *MS spectrum* [4]. Peptide ions within a selected m/z range are then submitted to the collision chamber and fragmented at their amide bonds. Depicted in Figure 1.1 is a standard *data dependent acquisition* (DDA) setup. The DDA process aims at selecting only one peptide ion for fragmentation. This results in pre- and suffix fragment ions of the original peptide sequence. The fragment ions are then submitted to a second mass analyzer and m/z values and intensities are determined. The resulting spectrum is often called *MS/MS spectrum* [4]. The pre- and suffix fragment ions can be used to identify the peptide sequence [4]. A more recent method of data acquisition, termed *data-independent acquisition* (DIA) [15] acquires MS/MS scans at fixed time intervals over broader mass ranges. Often an MS/MS spectrum is generated by a mixture of different peptides. This results in more complex spectra, but generally retains more information about the sample due to the consistent scanning.

1.2 Essential steps in MS proteomics data analysis

The data acquisition steps described in the previous section constitute the basic structure of MS proteomics data. In principle, the data acquired in an MS proteomics run is a collection of MS and MS/MS spectra and their respective acquisition times. From this data, the measured peptides and proteins have to be derived and quantified. In Figure 1.2, four essential steps of proteomics data analysis are displayed: peptide identification, peptide quantification, protein identification and downstream analyses. Figure 1.2a depicts peptide identification from a DDA MS/MS spectrum with a *sequence database* and a *spectral library*. For sequence-based identification, a database of possible peptides is derived from a database of protein sequences. Comparing the mass of the intact *precursor ion* with the sequence database allows to narrow down the number of candidate peptides. The m/z values of the fragment ions (i.e. the pre- and suffix ions of the precursor) in the MS/MS spectrum are then compared with the expected m/z values of all candidate peptides and a score is calculated. For sequence database searching, a wide variety of computational tools exist. Popular examples are SEQUEST [17], Mascot [18], Andromeda [19], MSGF+ [20], X!Tandem [21], MyriMatch [22], OMMSA [23], MSFragger [24] and pFind [25]. An alternative approach to sequence based searching is *spectral library* searching [26], where the MS/MS spectrum is scored against previously measured and identified spectra. In this approach, a central focus is on the intensity information of the fragment ions. Popular tools for spectral library searching are SpectraST [27], BiblioSpec [28] and X! Hunter [29]. In the context of DIA data analysis, spectral library searching has recently gained renewed importance and will be examined in more detail in chapter 2 of this thesis.
To quantify the measured peptides, several approaches exist. The key idea is always to

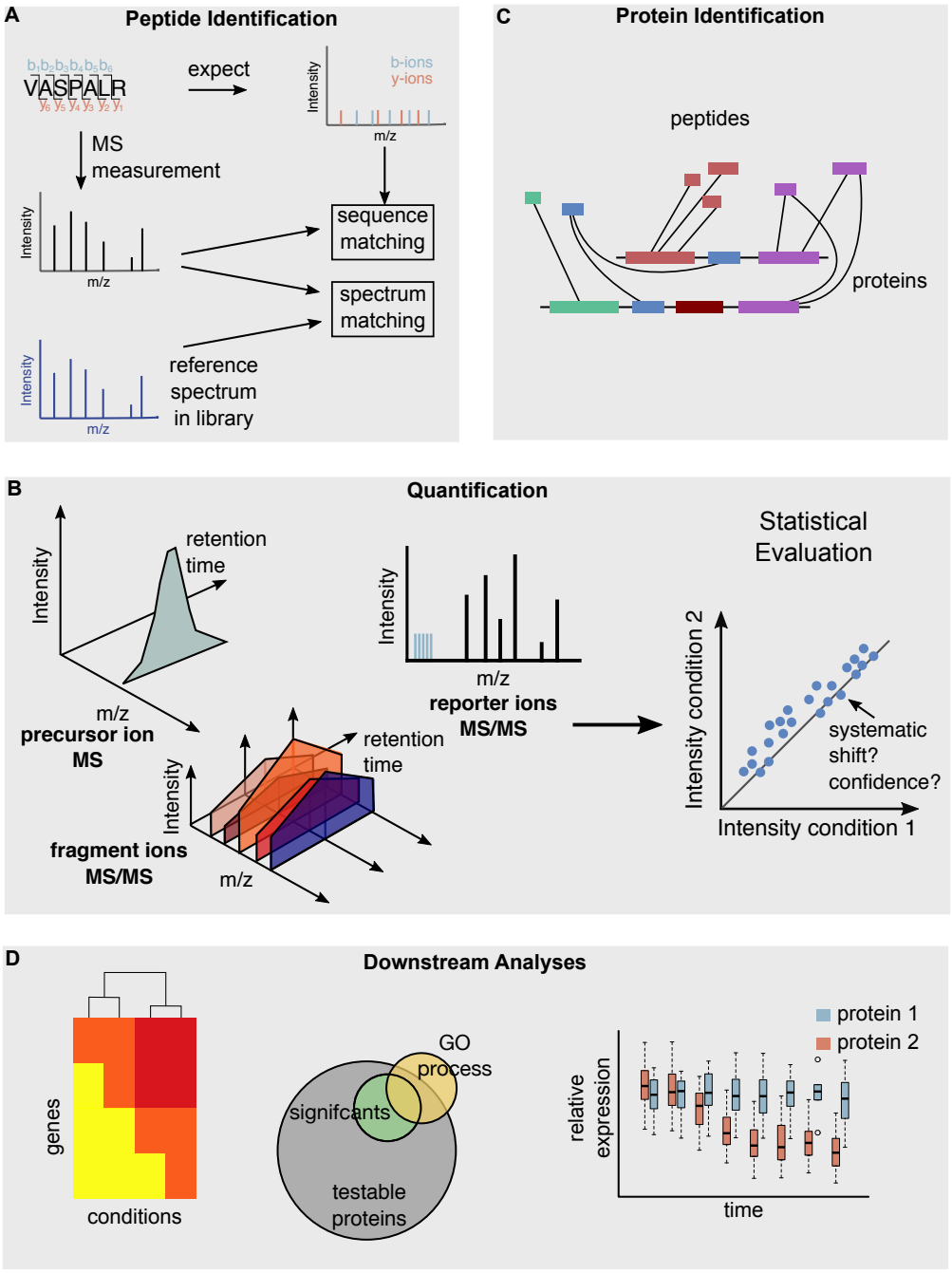


Figure 1.2: Essential steps in computational proteomics. a) Peptide identification. Peptide spectra can be identified by either comparing the sequence derived from a spectrum with a database or by comparing against previously identified spectra. b) Quantification. Intensity values for a peptide ion can be derived differently, depending on the quantification technique. Signals often have to be integrated over time. When comparing a protein between biological conditions, often many intensities exist for each condition. A statistical assessment is necessary to distinguish between systematic changes and noise. c) Protein identification. Smaller peptides have to be mapped to the protein sequence. Mechanisms such as alternative splicing can impede a unique mapping of peptides to a distinct protein. Optimal protein matches have to be derived computationally. d) Downstream analyses such as visualizations, enrichment analyses or combined scorings are needed to extract biological information from the quantified protein values. Image inspired by [16].

extract intensity values that represent the abundance of the peptide ions as accurately as possible. On the left in Figure 1.2b, we see the chromatographic profile of an intact peptide ion in the MS spectrum. This profile comes from the fact that peptides of the same sequence are distributed slightly over the LC column and reach the end of the column at different times. The profile hence reflects the distribution of the peptide in the LC column and extracting the area under the profile curve is a good approximation for the intensity of the peptide ion. This type of quantification is used for DDA-based *label-free quantification* (LFQ) [30] and for stable isotope-based labeling, such as *Stable Isotope Labeling by Amino acids in Cell culture* (SILAC) [31]. For LFQ, the intensities of the same peptide from different runs are compared. For SILAC, metabolically labelled heavy isotope peptides can be measured in the same run as native peptides. Their intensities can then be compared within the same run. In addition to the precursor quantification approaches, there are also approaches to extract peptide intensities from the MS/MS spectra. One method is based on isobaric labels such as *tandem mass tags* (TMTs) [32] or *Isobaric tags for relative and absolute quantitation* (iTRAQ) [33]. For isobaric labelling, peptides from different conditions are labelled with different isobaric labels and pooled together. On the MS level, the conditions cannot be distinguished as the labels are isobaric. After fragmentation however, different reporter ions detach from each label and can be quantified. For DIA data, MS/MS spectra are acquired repeatedly at a fixed frequency. This allows to extract chromatographic intensity profiles of the fragment ions of a peptide. The computational challenges associated with quantification range from the correct extraction of the intensity profiles to proper normalization and to the appropriate statistical evaluation of quantitative differences [30, 34]. Chapter 3 of this thesis will focus on the latter point.

A further computational challenge in MS proteomics is protein identification. For protein identification (Figure 1.2c), the identified peptide sequences are mapped back to the protein sequences available in the database. If a peptide maps uniquely to a protein, then the mapping is trivial. If a peptide maps to different proteins the mapping becomes more difficult and appropriate computational solutions have to be found. In particular *alternative splicing*, where multiple proteins with overlapping common parts are produced for a single gene, can lead to ambiguities in the peptide mapping. Such ambiguities are often addressed by applying the principle of parsimony and choosing the smallest set of proteins that explains all available peptides [35, 36]. A recent approach also applies more sophisticated bayesian modeling [37]. In chapter 4 of this thesis, we propose a new approach to test abundance changes between common and exclusive peptides in order to detect alternative splicing. This way, we not only use sequence information, but also quantitative information to identify splicing. This allows for a more in-depth examination of the regulatory aspects of alternative splicing and distinction between possible protein candidates.

The previous three points aim at providing well definied and quantified proteins. It should be noted that also several larger computational frameworks exist that aim at providing a comprehensive toolset for MS proteomics analyses. Popular and freely accessible pipelines are OpenMS [38] (DDA & DIA), MaxQuant [39] (DDA) and Skyline [40] (DIA and targeted data acquisition). After identification and quantification of proteins, an important computational aspect is also the downstream analysis that aims at the evaluation and biological interpretation of the data. These analyses usually differ for each biological problem and often tailored solutions are required. The bioinformatics tasks range from clusterings to statistical

evaluations and data visualization. In chapter 5 of this thesis, we focus on such a biological analysis in the context of *Escherichia coli* starvation.
It should be noted that I have introduced the core elements and steps of MS proteomics with the help of representative examples. The introduction should convey enough knowledge and intuition to apprehend the contents of this thesis. However, every step along the MS proteomics pipeline can be seen as a research field in itself and the interested reader is referred to the specialized literature for more details [41, 42, 43, 6].

1.3 Finding contexts

The topics of this thesis range along the computational steps displayed in Figure 1.2. In all projects, we utilize context information to facilitate computational improvements or biological insights.
In the spectral library project presented in chapter 2, the context is generated by collecting data from different spectral repositories. For each peptide, we then have spectra coming from many different MS proteomics runs. When creating a spectral library, we try to retain the relevant information of this experimental context using a clustering-based approach. This allows us to substantially increase the sensitivity and applicability of spectral library searching. For the differential quantification project introduced in chapter 3, we generate empirical statistical contexts, which estimate the quantitative variation inherent in the data. We use replicate measurements to generate empirical noise distributions and design these distributions in a way that allows to immediately put each quantitative feature (i.e. peptide) into a statistical context. This approach yields high statistical power for the detection of regulated proteins and performs favourably compared to state-of-the-art approaches. For the differential alternative splicing project introduced in chapter 4, we extend the statistical contexts of chapter 3 to 'double differential' setups. Double differential means that not the change of one object between conditions is evaluated, but the difference in the changes of two objects. To create the distributions, we again rely on replicate measurements, but asses the differences between pairwise changes. This approach allows us to quantitatively detect differential alternative splicing in proteomics data. For the *E. coli* proteomics project introduced in chapter 5, we rely on both, experimental and statistical context information. By obtaining all relevant datasets available for our biological problem, we create an experimental context. Consequently, we have a multitude of biological conditions for each protein. Our aim is then to calculate a representative value which reflects the overall response of a protein over the experimental context. To calculate the representative value we use our differential quantification tool. We hence extend the context-based differential changes to context-based multi-experiment changes. This allows us to obtain new insights into the molecular details of *E. coli* starvation.

Chapter 2

Detecting peptides in MS proteomics data

Motivation

As discussed in the introduction, a standard MS proteomics data file is basically a collection of mass spectra over time. These spectra largely depict the masses and intensities of peptide ions and their fragments. One of the first and most basic steps in the evaluation of MS proteomics data is therefore the assignment of mass spectra to their corresponding peptides (*peptide identification*). For standard proteomics setups, the so called *database searching* has established itself as the method of choice [16]. An alternative, sensitive method for identifying peptides is *spectral library searching* [26]. Spectral library searching suffers from the drawback that it performs best when the spectral library is *custom made*, meaning that the spectral library is generated under very similar experimental conditions as the newly measured query run [44]. This limits the applicability and has resulted in considerably smaller distribution of spectral library searching as compared to database searching. Spectral library searching, which was originally developed in the early 2000's is currently experiencing a renaissance in the context of new *data-independent acquisition* (DIA) proteomics approaches. DIA approaches benefit strongly from (and were initially completely dependent on) spectral library searching [34]. In the light of these developments, improving spectral library searching becomes increasingly important. This is the motivation for the following chapter, where we present a computational contribution to spectral library searching. We perform a systematic investigation of the experimental context a spectrum has been acquired in. We generate these experimental contexts from publicly available *spectral repositories* that contain large numbers of identified mass spectra. We propose a clustering-based adaptation to current spectral library searching by using *Multiple Characteristic Intensity Patterns* (MCIPs). We see that the MCIP approach in conjunction with spectral repositories mitigates the necessity for *custom made* spectral libraries and shows increased performance on DIA data.

Publication

The content of this chapter was published in *Journal of Proteome Research* ([45]). Reprinted (adapted) with permission from C. Ammar, E. Berchtold, G. Csaba, A. Schmidt, A. Imhof, and R. Zimmer, "Multi- reference spectral library yields almost complete coverage of heterogeneous LC-MS/MS data sets," *Journal of proteome research*, vol. 18, no. 4, pp. 1553–1566, 2019.
Accessible under: http://pubs.acs.org/articlesonrequest/AOR-eDu4gHVUDPW8VCP5TSSR
Supplemental materials can be found online with the publication. Here, the reformatted manuscript is presented with minor modifications.

Author contributions

I headed the project, implemented and performed bioinformatics analyses and wrote the manuscript with suggestions from all authors. I jointly designed the MCIP approach together with Gergely Csaba, who contributed to the analyses and provided helper classes. Evi Berchtold helped with the bioinformatics analyses. Andreas Schmidt performed the mass spectrometry experiments. Axel Imhof supervised the mass spectrometry analyses. Ralf Zimmer supervised method development, bioinformatics analyses and the writing of the manuscript. Technical descriptions of MS proteomics measurements were provided by Andreas Schmidt and Axel Imhof.

2.1 Abstract

Spectral libraries play a central role in the analysis of data-independent-acquisition (DIA) proteomics experiments. A main assumption in current spectral library tools is that a single characteristic intensity pattern (CIP) suffices to describe the fragmentation of a peptide in a particular charge state (peptide charge pair). However, we find that this is often not the case. We carry out a systematic evaluation of spectral variability over public repositories and in-house data sets. We show that spectral variability is widespread and partly occurs under fixed experimental conditions. Using clustering of preprocessed spectra, we derive a limited number of multiple characteristic intensity patterns (MCIPs) for each peptide charge pair, which allow almost complete coverage of our heterogeneous data set without affecting the false discovery rate. We show that a MCIP library derived from public repositories performs in most cases similar to a "custom-made" spectral library, which has been acquired under identical experimental conditions as the query spectra. We apply the MCIP approach to a DIA data set and observe a significant increase in peptide recognition. We propose the MCIP approach as an easy-to-implement addition to current spectral library search engines and as a new way to utilize the data stored in spectral repositories.

2.2 Introduction

Data-dependent acquisition (DDA) approaches are still the standard of proteomics data acquisition. In DDA, selected precursor ions are isolated in a small mass window and subsequently submitted for fragmentation and MS measurement [4, 46] giving MS2 spectra. The most widely applied DDA approach is also called shotgun proteomics, whereby in each duty cycle fragmentation spectra of the N most intense precursor ions are acquired (Top N). The corresponding MS2 data are commonly analyzed by scoring the mass to charge (m/ z) values of the most intense fragment peaks against a theoretical prediction of m/z values of fragment ions derived from sequence databases [17, 18, 21]. The theoretical m/z values of fragment ions are discriminative as, in most cases, each peak in the MS2 fragmentation spectrum stems from the same precursor ion. Additionally, the m/z value of the submitted precursor ion is known, which narrows down the number of possible matches in the sequence database. Peptide precursors not selected for fragmentation are excluded from the result since sequence confirmation is missing [47]. As precursor ion selection can be described as semirandom [48], DDA approaches are also problematic for quantification, as a peptide measured in a first sample might not be identified in a second sample, even though it is abundant. Selected reaction monitoring (SRM, alternatively multiple reaction monitoring (MRM) or parallel reaction monitoring (PRM)) approaches [49, 50] address the problem of reproducibility by a fixed preselection of peptide precursor ions. This approach allows very sensitive and accurate quantitation of a small number of proteins in each LC-MS run; however, the overall coverage of the proteome is low due to the preselection. Higher coverage can only be achieved by measuring the sample with multiple precursor lists. Data-independent acquisition (DIA) approaches try to overcome these limitations by omitting the preselection of precursor ions [51, 52, 53]. To reduce spectral complexity, many applications scan MS/MS spectra of medium-sized isolation windows (5-50m/z) over a wide m/z

range [54, 15, 34, 55, 56, 57]. In general, the possibilities for spectral searches via sequence databases are challenging for DIA data [58, 59] due to the ambiguity of m/z values in complex peptide mixtures. Thus, many commonly used approaches rely on spectral libraries, also considering fragment ion intensities [15, 34]. These libraries are obtained from DDA proteomics experiments by generating a characteristic intensity pattern (CIP) of m/z and ion intensity pairs (m,i) from confidently identified MS2 spectra for each peptide in a distinct charge state (peptide charge pair) [27, 28, 29, 60]. A library pattern must be constructed such that it is sufficiently specific (implying few false positives) while maintaining high sensitivity (few false negatives). A library of CIPs is then compared to the measured fragmentation spectrum using a similarity measure. Most current approaches for the construction of library patterns employ the scoring measure dot product [61, 62, 63, 64] or the related spectral contrast angle [65, 66, 67] as scoring measure. To our knowledge, all tools try to approximate one unique CIP from the available measured fragmentation spectra. Prior to their use in DIA approaches, spectral libraries have been employed to speed up and increase confidence in peptide recognition [27] and therefore, large spectral repositories [68, 69, 70, 71] have been compiled. In current DIA applications like OpenSWATH [34] chromatography-based scores such as retention time are used to find MS2 fragmentation spectra, which are then matched with a library CIP. Hence, having an accurate spectral library and a highly reproducible and calibrated LC system are key factors determining the quality of a DIA experiment. In the context of these developments, improved spectral libraries gain renewed importance. In this study, we present a systematic analysis of fragmentation spectra identified with high confidence by generating and evaluating a model spectral library. We integrate data from the databases ProteomeTools [72] (further referred to as Kuster Set), Pan Human Library [73] (further referred to as Aebersold Set), as well as from our own lab (further referred to as Imhof Set). The Kuster Set contains 8 different combinations of fragmentation type, fragmentation energy, and readout, all acquired on an Orbitrap Fusion Lumos mass spectrometer. The Aebersold Set had fixed fragmentation settings and was acquired on an AB SCIEX TripleTOF 5600+ system from different human tissues and cell lines. The Imhof Set had fixed fragmentation settings and was acquired on an AB SCIEX TripleTOF 6600 from different organisms and cell lines (see also Table 2.1 for an overview). We only use peptides, which have been measured and identified at least 20 times across several experiments, yielding ≥ 20 replicate fragmentation spectra for each peptide charge pair. Hence, for each peptide, we obtain an empirical estimate how similar the fragmentation behaviors of the individual spectra are (i.e., it can be derived whether certain ways of fragmentation happen more often than others).

We first demonstrate that a surprisingly large fraction of MS2 spectra corresponding to the same peptide charge pair is strongly heterogeneous across experimental conditions. This heterogeneity represents a large drawback of using public repositories for spectral library searching, which are mostly obtained under different experimental conditions than the query spectra they are used on. A common practice in many proteomics laboratories is hence the generation of custom-made spectral libraries, especially in the context of DIA experiments [74]. This means it is necessary to generate a spectral library from DDA runs of the desired sample under as similar experimental conditions as possible. One obvious problem is the experimental and computational effort that has to go into creating a custom- made library. Additionally, the set of peptides contained in a custom-made library is usually orders of

magnitude smaller than the peptides available in online repositories. On the basis of our findings, we propose the multiple characteristic intensity pattern (MCIP) approach, which is similar to the SpectraST approach by Lam et al. [60] but differs with respect to the following points: (i) SpectraST uses semiraw (.mzXML) fragmentation spectra for the generation of spectral libraries, without further preprocessing [60]. We conduct our library generation on MaxQuant [39] preprocessed peptide identifications without modifications and consider only b- and y-ions (with molecular losses). (ii) As we use preprocessed spectra, we can either apply a ranking prior to the clustering or use an unranked approach. In both cases, we apply a systematic clustering until all spectra are contained in a cluster and retain all clusters involved. (iii) We determine one CIP from each cluster. This can yield more than one CIP per peptide charge pair. We compare a spectral library generated with the MCIP approach from a repository with a custom-made spectral library and show comparable performance for most data sets. An overview over the major steps taken in this study is given in Figure 2.1. The MCIP method outperforms the current single CIP approach employed in spectral library searching. We suggest this easy to implement "one-size-fits-all" method as a new way to utilize the data available in spectral archives.

id	instrument	readout	fragmentation	background matrix	lab	collison energy
L_HCD_O_20	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	20%
L_HCD_O_23	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	23%
L_HCD_O_25	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	25%
L_HCD_O_28	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	28%
L_HCD_O_30	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	30%
L_HCD_O_35	Orbitrap Fusion Lumos	Orbitrap	HCD	synthetic peptides	Kuster	35%
L_HCD_I_28	Orbitrap Fusion Lumos	ion trap	HCD	synthetic peptides	Kuster	28%
L_CID_I_35	Orbitrap Fusion Lumos	ion trap	CID	synthetic peptides	Kuster	35%
Q_CID_AEBERSOLD	AB SCIEX TripleTOF 5600+	TOF analyzer	CID	human tissue + cell lines	Aebersold	rolling
Q_CID_IMHOF	AB SCIEX TripleTOF 6600+	TOF analyzer	CID	drosophila tissue + cell lines	Imhof	rolling

Table 2.1: Overview over the data sets used in this study and the corresponding experimental parameters

2.3 Experimental section

2.3.1 Proteome analysis using a Q-ToF MS

An Ultimate 3000 HPLC system (Thermo Fisher Scientific) was used. Tryptic peptides were desalted on a trapping column (5 x 0.3 mm inner diameter; packed with C18 PepMap100, 5μm particle size, 100 1 Å pore diameter, Thermo-Fisher Scientific) to perform nanoreversed phase separation. 0.1% formic acid (FA) was initially used. The flow of the loading pump was set to 25 μL/min and washing was performed for 10 min under isocratic conditions. An analytical column (150 x 0.075 mm inner diameter; packed with C18RP Reposil-Pur AQ, 2.4 μm particle size, 100 pore diameter, Dr. Maisch) was used for separation with a linear gradient from 4% to 40% B in 170 min and a gradient flow of 270 nL/min. To separate the sample, the solvent A 0.1% FA in water and B 80% acetonitrile (ACN), 0.1% FA in water were used. Using a nano-ESI source, the 6600 TOF mass spectrometer was directly coupled

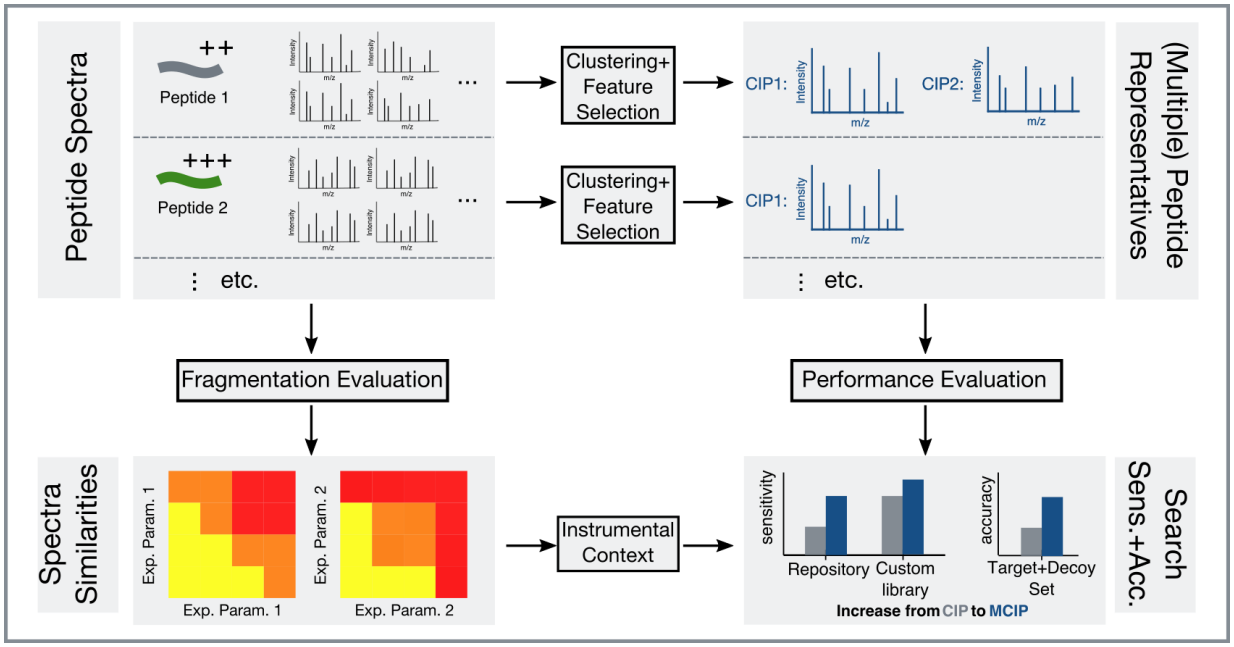


Figure 2.1: MCIP analysis workflow. Preprocessed peptide spectra are collected from many data sets and MS runs. Similarities of the spectra are compared over different public repositories and in-house data sets. Multiple characteristic intensity patterns (MCIPs) are generated from the spectra. Search performance (sensitivity, accuracy, etc.) is evaluated in different cross-validation settings, also considering the different experimental contexts.

to the HPLC (both AB Sciex). DDA settings were chosen with 225ms survey scan and mass range 300 to 1800 m/z. Up to 40 MS/MS scans were allowed (100-1800 m/z). Exclusion time of fragmented precursors was set to range between 10 and 50 s, depending on the experiment (see Supplemental Table 2). Rolling collision energy setting was enabled, which performs fragmentation at optimized collision energy for the peptide charge pairs. Precursor charge states from +2 to +5 were specifically detected. SWATH runs were generated with the same HPLC settings and 40 mass windows (Supplemental Table 2).

2.3.2 Data analysis of data-dependent LC-MS/MS experiments

The Aebersold Set and the Imhof Set were analyzed with MaxQuant (version 1.5.1.2 and higher) using the Andromeda search engine [19] with a FASTA protein database specific to the sample (see Supplemental Table 1). The following settings were used: fixed modification carbamido- methylation of cysteine, variable modifications oxidation of methionine, and acetylation at the protein N-terminus ; for precursors Δmass = 30 ppm in the first search and in the second search 6 ppm, for fragment ions the Δmass = 60 ppm, enzyme trypsin with specific cleavage and max two missed cleavages. The minimum peptide sequence length was set to 7 and for modified peptides the minimum required scored was set to 40. For modified peptides the score was set to 40. The false discovery rate (FDR) for a peptide spectrum match was set to 1%. MaxQuant preprocessing included mass centroiding of peaks and corresponding intensity adaption, de-isotoping, and detection of cofrag-

mented peptides [39]. The results were returned as msms.txt files containing the relevant spectral information of fragment ion intensities, retention times, fragment masses, as well as charge and modification states of the identified peptide. The MS proteomics data of the Imhof Set, including MaxQuant results, "have been deposited to the ProteomeXchange Consortium ($http://proteomecentral.proteomexchange.org$) via the PRIDE partner repository" [75], with the data set identifiers PXD005060, PXD005063, PXD005100, PXD005111, PXD006245, and PXD006691. For the Kuster Set, the MaxQuant files were directly downloaded from the PRIDE repository PXD004732. The raw data for the Aebersold Set was downloaded from the PRIDE repository PXD000953. More details on the data sets used are displayed in Supplemental Table 1.

2.3.3 Library generation settings for the in-house data set

For the Imhof Set, the spectral library was generated from DDA data only, with the explicit runs marked in Supplemental Table 1. For the instrument, the standard configurations as recommended by Sciex were applied to all setups with the vast majority of parameters fixed between all runs. Different settings were only applied to the parameters "Exclude for" (range 10s-50s), "Mass tolerance" (15 ppm-50 ppm), "Switch After" (30 spectra -40 spectra), and "With intensity greater than" (100–150). Rolling collision energy was set in all cases. The specific parameters for each input sample are listed in Supplemental Table 2.

2.3.4 Selection of processed fragmentation spectra

Peptides were separated by charge into peptide charge pairs because differences in the charge state significantly alter the fragmentation pattern (see Supplemental Figure S1). Only peptide charge pairs, which had at least 20 replicate spectra (see Supplemental Figure S2), were included to enable the statistical analysis of repeated fragmentation of chemically identical peptides. In our main analysis, we restricted our MCIP approach to only b- and y-ions in charge states up to 2+ with different molecular losses (examples: b3, y4-NH3, y6(2+), b5(2+)-H2O). Modified peptides were excluded.

2.3.5 Import of raw fragmentation spectra

To quantify the impact of using all peaks without filtering, an additional analysis with raw spectra was carried out. To assess the influence of the preprocessing method, two different methods of preprocessing the data were applied. In the first approach, the raw spectra were imported from the MaxQuant ".apl" files contained in the "andromeda" folder in the MaxQuant output folder. We parsed these files and extracted a list of m/z values with corresponding intensities, without b- and y-ion annotation for each spectrum. The spectra were assigned to their respective MaxQuant identification via the spectrum index. In the second approach, the raw ".wiff" files were processed into the ".mzXML" format with the MSConvert tool [76] without any additional filters (yielding profile data), parsed, and assigned to the respective MaxQuant identification via the spectrum index. The influence of raw spectral scoring can be seen in Supplemental Figure S3 with an overall lower performance compared to the MaxQuant approach.

2.3.6 Assessment of the similarity of fragmentation spectra

The similarity among spectra of the same peptide charge pair (replicate spectra) can be used as a measure to characterize the fragmentation behavior of peptide charge pairs. As spectra are vectors of (m/z, intensity) pairs, they can differ in the m/z values (different peaks) or their intensities or both. To assess similarity between replicate spectra, all replicate spectra (at least 20, see previous section) available for a peptide charge pair were compared pairwise to each other. Each fragmentation spectrum was represented as a normalized replicate fragmentation vector (NRFV) $I = (i_1, i_2, ..., i_n)$, with i, j denoting the intensities in the pattern and the index j of the vector implicitly denoting the different fragmentation ions (m/z values). To get vectors of equal length, each fragmentation ion with intensity > 0 in any of the replicate spectra was included in every vector. Imputed 0 values were used if a corresponding ion was not observed. For raw spectra (Supplemental Figure S3) best bipartite matching was used. Only vectors with at least four nonzero values (n > 4) were used. Each vector was normalized to length |I| = 1 (unit vector). After determining which intensities were included in the NRFVs, the spectral similarities between all NRFVs of a peptide charge pair were assessed in a pairwise fashion. For each pair of vectors X and Y of NRFVs, the dot score was calculated using the dot product similarity measure DP defined as

$$DP(X, Y) = \sum_{k=1}^{n} x_k y_k \tag{2.1}$$

with x_k and y_k denoting the kth element of X and Y, respectively. A pair of fragmentation spectra was called similar if the dot score of their two corresponding NRFVs was larger than a predefined similarity threshold (see below).

2.3.7 Choosing a global similarity threshold

A global similarity threshold of dot score 0.6 was adapted from the SpectraST search engine [60] and was subsequently tested using the sampling approach discussed below. This was done to check whether this threshold would give overall discriminative results. Each spectrum in the data set was represented as a NRFV and assigned 1000 differently shuffled decoy vectors. Each NRFV was then dot scored against each decoy vector, which resulted in a distribution of 1000 shuffled dot scores for each NRFV. From each distribution of shuffled dot scores, a local discriminative dot score was extracted, such that less than 5% of the shuffled dot scores were above this threshold (in other words, the 95% quantile was extracted). Thus, the use of this dot score would result in 5% acceptance of decoy spectra for a particular NRFV. All locally discriminative dot scores were collected. From the distribution of locally discriminative dot scores, again the 95% quantile was extracted (see Supplemental Figure S4). This 95% quantile was 0.62 in this study, which agreed well with the global similarity threshold of 0.6. The approach of extracting two quantiles was taken because the distribution of shuffled dot scores varied distinctly for different spectra. Hence, taking only one quantile of the distribution of all shuffled dot scores of all spectra combined would result in some spectra (the spectra with generally large shuffled dot scores) being ambiguous. Still, a dot score cutoff of 0.6 might be comparably low considering current high-resolution data.

2.3.8 Centroid clustering and CIPs

A central goal of this study is to find a minimal set of characteristic intensity patterns (CIPs), able to characterize all observed fragmentation spectra of a peptide charge pair. In order to derive these, a centroid clustering approach was employed to determine clusters of similar NRFVs. For each NRFV, the neighborhood (all fragmentation spectra with a similarity score greater than the chosen similarity threshold) was determined. The medoid NRFV, corresponding to the spectrum with the best signal-to-noise ratio (defined via the average intensity of the second to sixth highest peak divided through the median of the remaining peaks), was defined as a CIP, analogous to the SpectraST approach [60]. Additionally, also NRFVs with the largest number of neighbors were defined as CIPs. If not all NRFVs were neighbors to this CIP, it becomes a cluster with all its neighbors and the procedure was repeated on the remaining NRFVs. A visualization of the MCIP clustering procedure is given in Figure 2.2. Depending on the number of CIPs resulting from this procedure, each peptide charge pair was assigned either a single CIP (all spectra of a peptide charge pair assembled in a single cluster) or multiple CIPs (MCIPs). The CIPs were referred to by size of their respective cluster: CIP_1 corresponds to the largest cluster and CIP_i to the ith largest cluster.

2.3.9 Spectral coverage

The spectral coverage was introduced as a measure for the sensitivity of the approach. A spectral library was constructed with the entries for each peptide charge pair consisting either of a single CIP of the largest cluster or of MCIPs $\{CIP_1, ..., CIP_n\}$ of the n largest clusters. The single CIP or each element of the MCIPs $\{CIP_1, ..., CIP_n\}$ was then compared to all NRFVs of the peptide charge pair using the dot score. If the dot score was above the similarity threshold for any of the CIPs, the respective spectrum was marked as covered. The spectral coverage denotes the fraction of replicate spectra covered.

2.3.10 Comparison to custom-made spectral libraries

To compare the performance of a custom-made library with a MCIP library, we implemented a test set and three training sets. For each experimental setup S, we selected all peptide charge pairs with at least 10 spectra in setup S (and at least 10 spectra in other setups). Five spectra belonging to S were randomly assigned to the test set. The remaining spectra of S were assigned to the first training set, termed the custom training set. All spectra that did not belong to S were assigned to the MCIP training set. The union of custom training set and MCIP training set was termed MCIP custom training set. Hence, the custom training set corresponded to the scenario of a custom-made spectral library, the MCIP set corresponded to the scenario of having a heterogeneous spectral repository and the custom MCIP set corresponded to the scenario of integrating a repository library with a MCIP library. Only the main CIP was determined from the custom training set, and MCIPs (and also one CIP as a control) were determined from the MCIP training sets. The dot scores of the respective CIPs/MCIPs with the test set were computed.

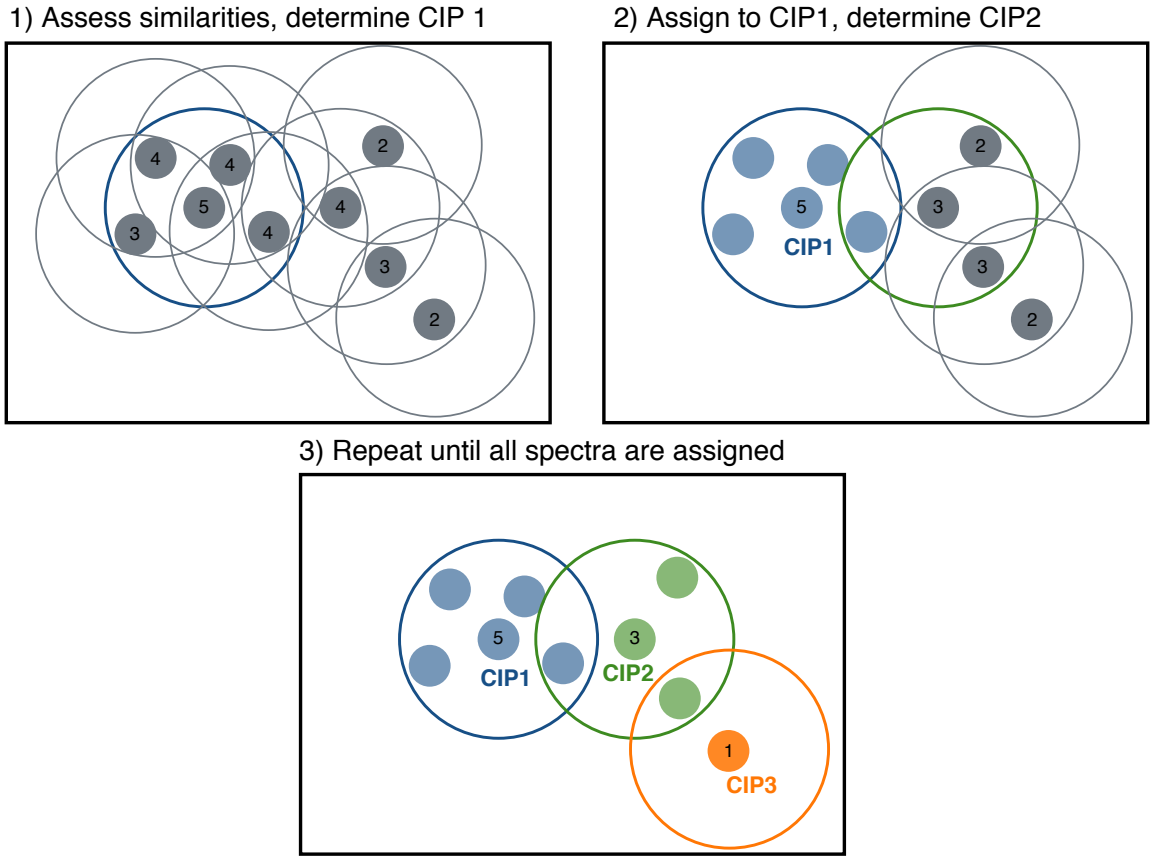


Figure 2.2: Example of the MCIP method applied to a set of input spectra using the maximum neighbor approach. Each point represents a fragmentation spectrum, and the distance of the points to each other represents the spectrum similarity. Large circles represent the similarity thresholds. Numbers on the points indicate the number of neighbors below the similarity threshold. Spectrum with the maximum number of neighbors (i.e., the medoid) is chosen as CIP_1. All neighbors are assigned to CIP_1, spectra outside the similarity threshold are clustered again, and CIP_2 is found. Clustering is repeated until all spectra are assigned to a CIP.

2.3.11 Comparison with SpectraST

A comparison of the spectral coverage with the popular SpectraST search engine [60] was carried out. For this, input files in ".pep.XML" format suitable for SpecraST were created from the MaxQuant spectrum identifications. Hence, for each training set belonging to a specific training and test set combination, a set of ".pep.XML" files was generated that contained only the spectra of the specific training set. SpectraST library spectra were then generated from these ".pep.XML" files. This ensures that the comparison between the MCIP approach and the SpectraST approach is carried out with exactly the same underlying data. To generate the SpectraST library spectra, .pep.XML output files were submitted to SpectraST in library create mode using the default configurations. The resulting raw library was processed to a consensus library using the corresponding SpectraST option. The consensus library mode was chosen because it has been shown to give the highest number of positive

identifications [27]. The consensus library was then quality filtered using the highest quality level (option -cL5) in SpectraST. The raw spectra from the Kuster Set were converted into ".mzXML" format with the tool MSConvert [76], and the ".mzXML" files were subsequently searched with SpectraST.

2.3.12 Benchmarking via cross validation

To conduct performance testing, a cross-validation approach was used. The replicate spectra of each peptide charge pair were split into two fractions. The first fraction consisted of 20% of the spectra, and each spectrum was assigned a decoy spectrum P decoy which contained m/z-shuffled intensities of the original spectrum. Shuffling was carried out using unbiased random permutations of the m/z values. As only identified m/z values were used, no further constraint was applied to the permutation. By shuffling the spectra, the total intensity and the m/z values were preserved while the spectrum was changed completely. A 1:1 mixed test set containing original and decoy spectra was then generated. The second fraction consisted of the remaining 80% of spectra. On this fraction, CIP(s) were created as described in the previous sections. The CIP(s) were then similarity scored against the test set using the dot score. A similarity score below the similarity threshold for an original spectrum P_{orig} was marked as false negative, a score above the threshold with a decoy spectrum P_{decoy} was marked as a false positive. The m/z-shuffling approach is similar to the method employed by Lam et al. [77], where counting of decoy matches is used library wide to estimate the FDR. Each set of replicate spectra was individually checked via 5-fold cross validation in this study. This allowed estimating the relative fractions of false positives and false negatives per peptide charge pair, rather than library wide.

2.3.13 Processing of targeted LC-MS/MS runs for CE and isolation window study

The targeted data acquisition setup mentioned in the discussion and supplemental Figures S13/S14, was not accessible to standard DDA processing via MaxQuant. The ".wiff" files were converted to ".mzXML" using MSConvert [74], and the ".mzXML" files were then processed using an in-house scoring method, termed ReScore. ReScore is a re-implementation of the scoring described in the publication of the MaxQuant search engine Andromeda [19]. The scorings are exactly re-implemented as described in the publication. However, as not all in-depth details of the processing were accessible, the absolute values are different. The scores were compared to Andromeda using DDA runs that were carried out along with the targeted LC-MS/MS runs on the same standardized HeLa Pierce lysate (PXD006691). The scores show strong correlation with the Andromeda scoring, and the vast majority of Andromeda scores is higher than the corresponding ReScore (Supplemental Figure S5). Hence, a certain ReScore cutoff can be used as a reliable cutoff for the Andromeda score.

2.3.14 Benchmarking spectral library performance on DIA data

To assess the spectral library performance on DIA data, a combination of the OpenSwath DIA search engine [34] and the corresponding spectral search engine SpectraST was used. A

SWATH datafile acquired in the scope of a benchmarking study of Navarro et al. [78] was downloaded from the PRIDE repository PXD002952 (file id I150211) and processed into the mzXML format with MSConvert. Additionally, the corresponding OpenSwath identifications were directly downloaded from the PRIDE repository (1% protein-level FDR and 1% peptide-level FDR). Spectral libraries were generated with SpectraST and with the MCIP method as described in the section above. Precursor tolerance for SpectraST was adapted to the SWATH window width. Noncanonical peaks were excluded (-s_UAS 0.0), and only the most intense peaks (-s_LNP 10) were chosen, as recommended by Schubert et al. [74] For the MCIP method, fragment ions were identified at 15 ppm accuracy, including molecular losses. Dot scores were extracted scoring the highest intensity spectrum in the OpenSwath peak group against the library spectrum. As SWATH data has different noise levels than DDA data, decoy distributions were generated for SpectraST as well as the MCIP approach. The decoy distributions were obtained by taking the dot scores of the library spectra with SWATH spectra that were 40 min away from the peak group retention time or in a differing m/z range. This ensures that the library spectra are scored against mass spectra not containing the library peptide. More than 100 samplings were carried out per peptide. To compare the significance of peptide hits with respect to the noise levels, an empirical p value (how often a dot score was higher equal to or higher to the library dot score) was calculated for each peptide for the MCIP and for the SpectraST approach. The resulting p-value distribution was corrected for multiple testing via the Benjamini-Hochberg method [79].

2.4 Results

2.4.1 Spectral variability is widespread over experimental conditions

We considered the data sets listed in Table 2.1 containing a total of 10 different experimental settings. We chose a subset of experiments for the Kuster Set with large peptide overlap with either the Aebersold or the Imhof Set, resulting in a heterogeneous set of experimental conditions (see Supplemental Figure S6). To obtain a more detailed understanding of spectral variability, we sorted all replicate spectra corresponding to their respective experimental condition. We then combined all possible pairs of experimental conditions, resulting in 45 pairs (one example pair: Orbitrap Fusion Lumos in HCD mode at CE 25 vs Sciex Q-ToF 5600+ using CID and optimized rolling collision energy). We assessed the dot scores between the experimental conditions in a pairwise manner. The median values of the resulting dot score distributions are displayed in Supplemental Figure S7 and show a clear clustering after experimental settings. To visualize dissimilar clustering, we plotted the lower 10% quantiles of the pairwise dot score distribution in Figure 2.3a. We observe a large spread in the distributions of dot scores with visible dependence on the experimental settings. The calculated dot scores are most stable for Orbitrap data generated by HCD fragmentation with collision energies (CEs) from 20 to 30. The Kuster CID@CE35 setup with low-resolution ion trap readout differs most from the remaining setups. The Q-ToF data sets cluster together with the highest CE Orbitrap dataset. We see relatively low dot scores within identical experimental settings (diagonal of the heatmap) for the Q-ToF data sets and for

high collision energies as well as for low-resolution readout in the Kuster Sets. This underlines that even under fixed experimental conditions fragmentation can vary. Some examples are shown in Supplemental Figure S8, and more interpretation of this phenomenon is given in the Discussion. To investigate the influence of CE on spectral similarity, we considered the distributions of pairwise dot scores between the Orbitrap set at CE 20 and the Orbitrap sets at higher CE (Figure 2.3b, left). We see a clear influence of the CE difference on the pairwise distributions. We extracted the 10% quantile and the median from these distributions and plotted them against the CE difference (Figure 2.3b, right). We see almost no influence of small CE changes. For CE changes between 5 and 8 we see a clear drop in similarity and between 10 and 15 an even stronger drop, which indicates complex processes underlying peptide fragmentation. Ion trap readout generally shifts the dot scores toward lower values (see Supplemental Figure S9). To elucidate the drastic effects of the observed variability on peptide identification, we carried out a clustering on the 45 combinations of experimental conditions (Figure 2.3c). Each combination in the heatmap shows the fraction of peptides that have only one cluster. This corresponds to no spectra clustering outside the main cluster and hence no spectra being missed in a spectral library search. We see that depending on the experimental combination, as little as 35% of peptides fulfill this condition, with most combinations ranging from 60% to 90%. From these observations we conclude two main points: (i) The machine setup can play a crucial role for spectral recognitions and (ii) even fixed experimental settings can lead to spectra being missed.

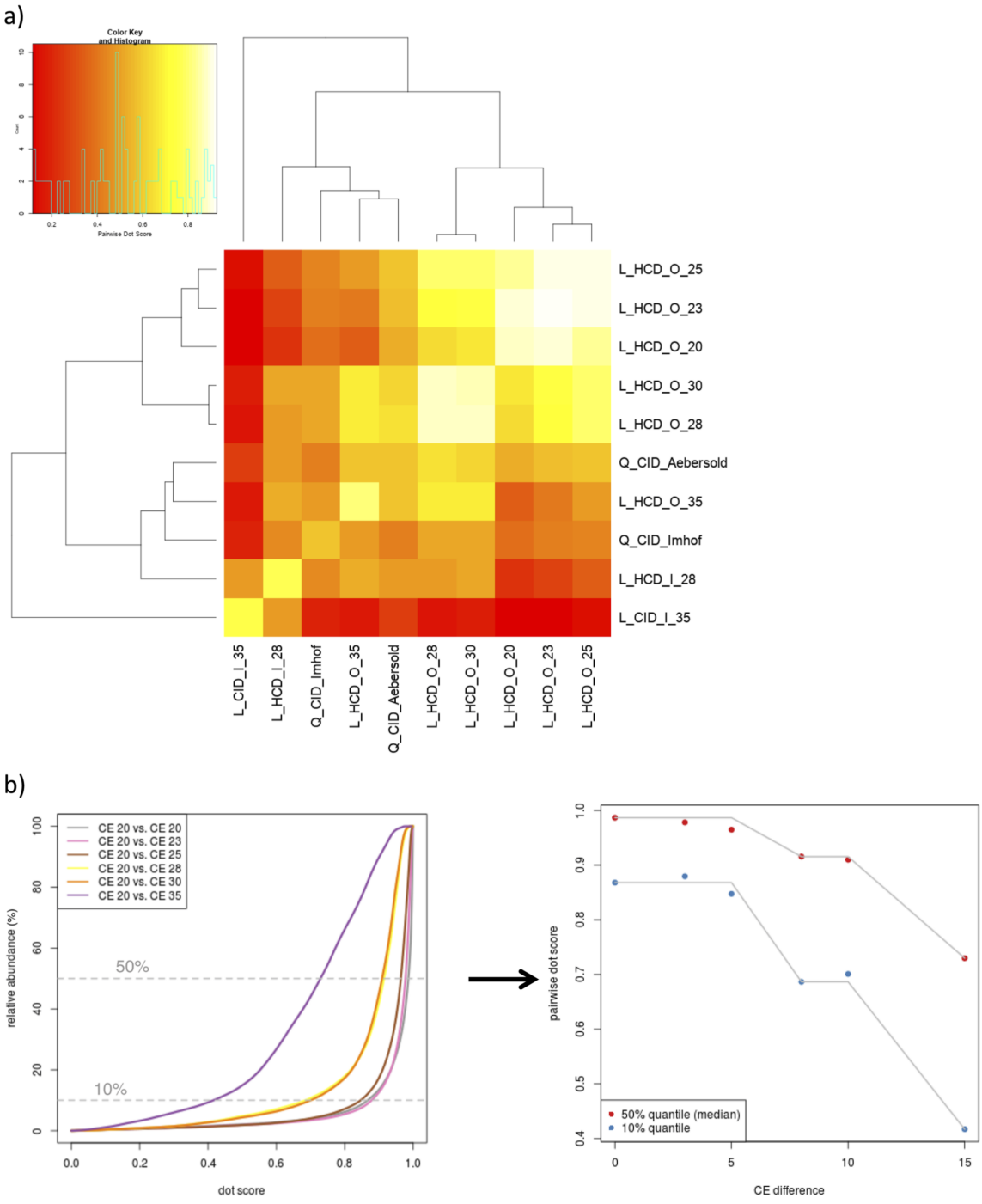


Figure 2.3: (continued)

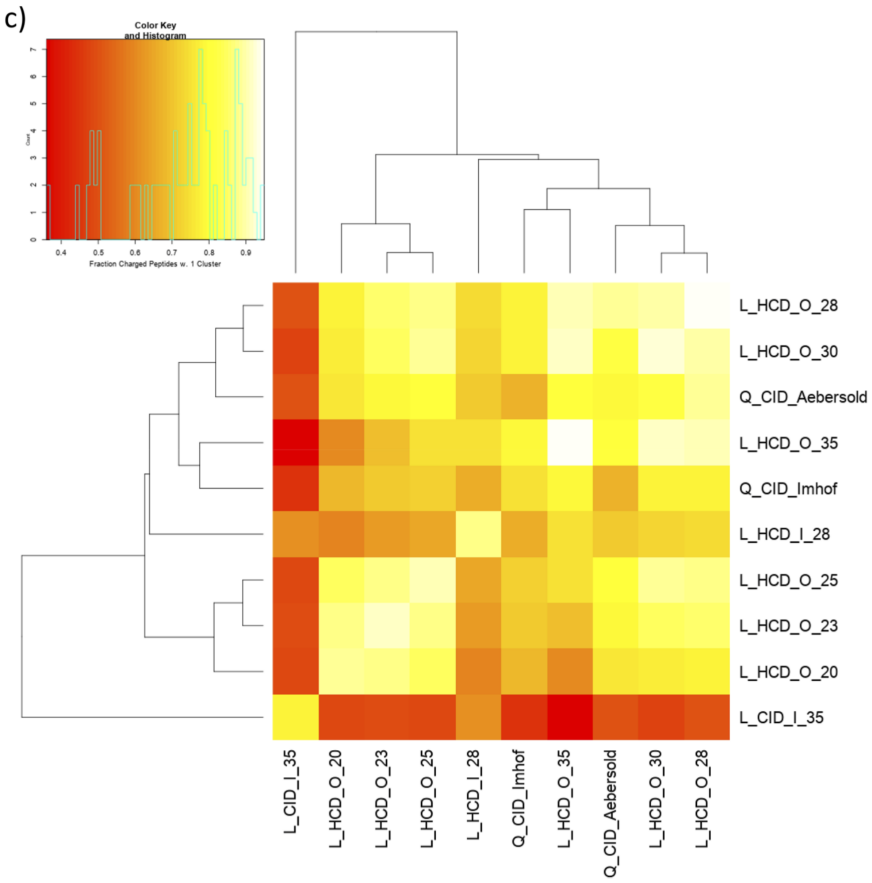


Figure 2.3: Spectral similarity between experimental conditions. (a) Ten percent quantiles of the distributions of dot scores between the different experimental conditions. (b) Distribution of pairwise dot scores for the Orbitrap HCD conditions with the collision energy (CE) 20 setting (left). Median and 10% quantile are drawn in as dashed lines and plotted against the CE difference (right). Two clear drops in similarity are visible. (c) Fraction of peptides with no missing spectra after a spectral library search, determined between the different experimental conditions. Depending on the condition pair, the fraction can go down to around 35%.

2.4.2 Usage of MCIPs yields almost complete spectral coverage

To achieve a global impression of the performance, we assessed the spectral coverage (see Expermental Section) for all peptide charge pairs. The spectra were compared either with one CIP (in accordance with the current library approaches) or with MCIPs. Figure 2.4a shows a striking improvement upon successive integration of more and more CIPs (red up to orange line) until near-complete coverage is reached. The largest gain is visible upon integration of the second CIP (blue line), which corresponds to the second largest cluster. This improvement clearly shows that significantly higher peptide recognition is possible by simply including two representative spectra for a peptide charge pair instead of just one.

2.4.3 MCIP library performs comparable to and enhances custom-made libraries

Custom-made spectral libraries can be seen as the gold standard for creating a high-performing spectral library [80]. These libraries, however, come with drawbacks compared to spectral repositories, mainly due to the effort in creating the library or the limited number of peptides in the library. To evaluate the performance of our libraries, we generated a set of test spectra for each experimental condition. For each set of test spectra, we generated spectral libraries from three different sets of training spectra: (1) The custom set only contained spectra measured under the same experimental condition as the test set. (2) The MCIP set contained spectra measured under all available conditions except the condition of the test set. (3) The MCIP custom set contained spectra measured under all available conditions. The MCIP set was chosen in this way to recreate the scenario of having a repository library (the MCIP set) and own data measured under a different experimental setting (the test set). We then determined either a single CIP or MCIPs from the different sets of training spectra and compared them with the test spectra (see also methods section). This allowed us to directly assess the effect of extending the current single CIP approach to an MCIP approach. We first examined the fraction of missed spectra, which denotes spectra that would not be detected in a spectral library search, when using a (rather low) similarity threshold of 0.6 (Figure 2.4b). Using MCIPs always performs better than the current single CIP approach. As a general trend, we see that differently clustering spectra are more common in experimental setups where either the spectral resolution is low (ion trap) or fragmentation energy settings are high. The most challenging setup Kuster CID@CE35 with ion trap readout leads to around one-third of spectra being missed when using a single CIP approach. Integrating multiple CIPs reduces the missed fraction by a factor of 2. Using a custom library further reduces the missed rate to around 5%. Using the MCIP custom training set yields an overall missed rate of 3%. For the other experimental conditions, the MCIP approach gives a similar performance as the custom library approach. In around one-half of the cases the custom library approach is slightly better; in the other half the MCIP approach performs slightly better. The MCIP approach in combination with the custom approach always increases accuracy. The results described also hold for maximum neighbor clustering (see Supplemental Figure S10) and are stable for different training sets. An example of a spectrum being detected by a CIP outside the main cluster is given in Figure 2.4c. To give an intuitive visualization of the similarities between the spectra, the fragment ion intensities are connected via lines. We see that CIP1, which was acquired at HCD@CE23, has a significantly less prominent b6-ion, which significantly alters the shape of the fragmentation profiles for the higher energy CID@CE35 and HCD@CE28 spectra. The annotated raw spectra corresponding to Figure 2.4c are displayed in Supplemental Figure S11.

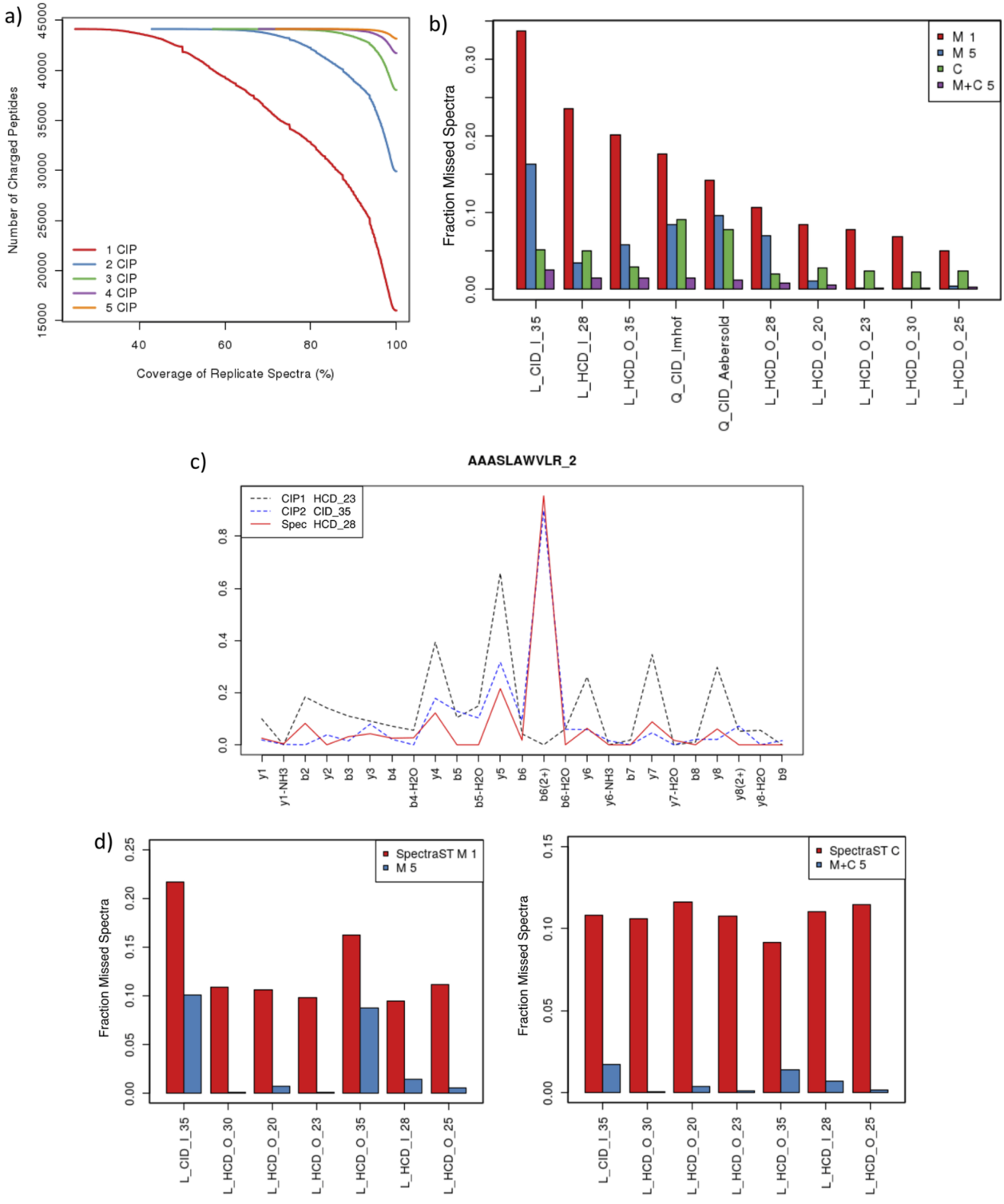


Figure 2.4: (caption on next page)

Figure 2.4: Comparison of the MCIP approach with current approaches. (a) Spectral coverage of the whole data set for different numbers of characteristic intensity patterns (CIPs) integrated in the spectral library. Number of peptide charge pairs (y axis) is displayed, for which the spectral coverage is larger than or equal to the value denoted on the x axis. The single CIP approach (red line) leaves a large fraction of spectra uncovered. Integrating one more CIP (blue line) into the library gives a strong increase in coverage with successively smaller increases upon integration of more CIPs until almost complete coverage is reached for up to 5 CIPs per peptide charge pair. (b) Fraction of spectra with dot score < 0.6 to the CIP/MCIPs (missed spectra). Each group of histograms displays one experimental condition the method is tested on. Different clustering approaches are indicated in the legend: M1 equals the MCIP approach with a single CIP (state-of-the-art approach). M5 equals the MCIP approach with a maximum of 5 CIPs included. C equals the custom library approach, and M+C 5 equals the MCIP approach with custom spectra included. (c) Example of a query spectrum dissimilar to CIP1 (dot score 0.28) but similar to CIP2 (dot score 0.96). Raw spectra are displayed in Supplemental Figure S11. (d) Missed spectra of MCIP and SpectraST using identical training and test sets. Number of spectra missed is significantly lower for the MCIP approach for both the noncustom (left) and the custom (right) approach.

2.4.4 Direct comparison with SpectraST shows significantly increased sensitivity

To further assess the performance of our MCIP approach, we carried out a comparison with the SpectraST [60] spectral search engine, which is among the most popular in the field [26]. We again determined the fraction of missed spectra at a similarity threshold of 0.6. We generated the SpectraST library on the identical spectra as our own library (see also Expermental Section). We see that the MCIP approach outperforms the single CIP approach of SpectraST in terms of sensitivity in the custom setup as well as in the noncustom setup (Figure 2.4d).

2.4.5 MCIPs increase sensitivity without affecting specificity

As has been shown in the previous sections, MCIPs are able to cover all replicate spectra for many peptide charge pairs and thereby improve sensitivity in spectral searches. However, this might come at the cost of reduced specificity (i.e., increase in false positives). Here we investigate, whether using MCIPs affects the number of false positives and the overall accuracy. We tested this by first generating CIPs on 80% of the replicate spectra and then scoring these CIPs against a mixture of the remaining replicate spectra and shuffled decoy spectra. This allowed distinguishing true positives (match of CIP with replicate spectrum) from false positives (match of CIP with decoy spectrum). The procedure is described in more detail in the Expermental Section. Figure 2.5a) shows that the overall accuracy for MCIPs (blue line) increases significantly in comparison to a single CIP (red line). For >99% of peptide charge pairs, the minimum accuracy increases by around 10% when integrating all CIPs available for each peptide charge pair in the spectral library (blue line). In Figure 2.5b,

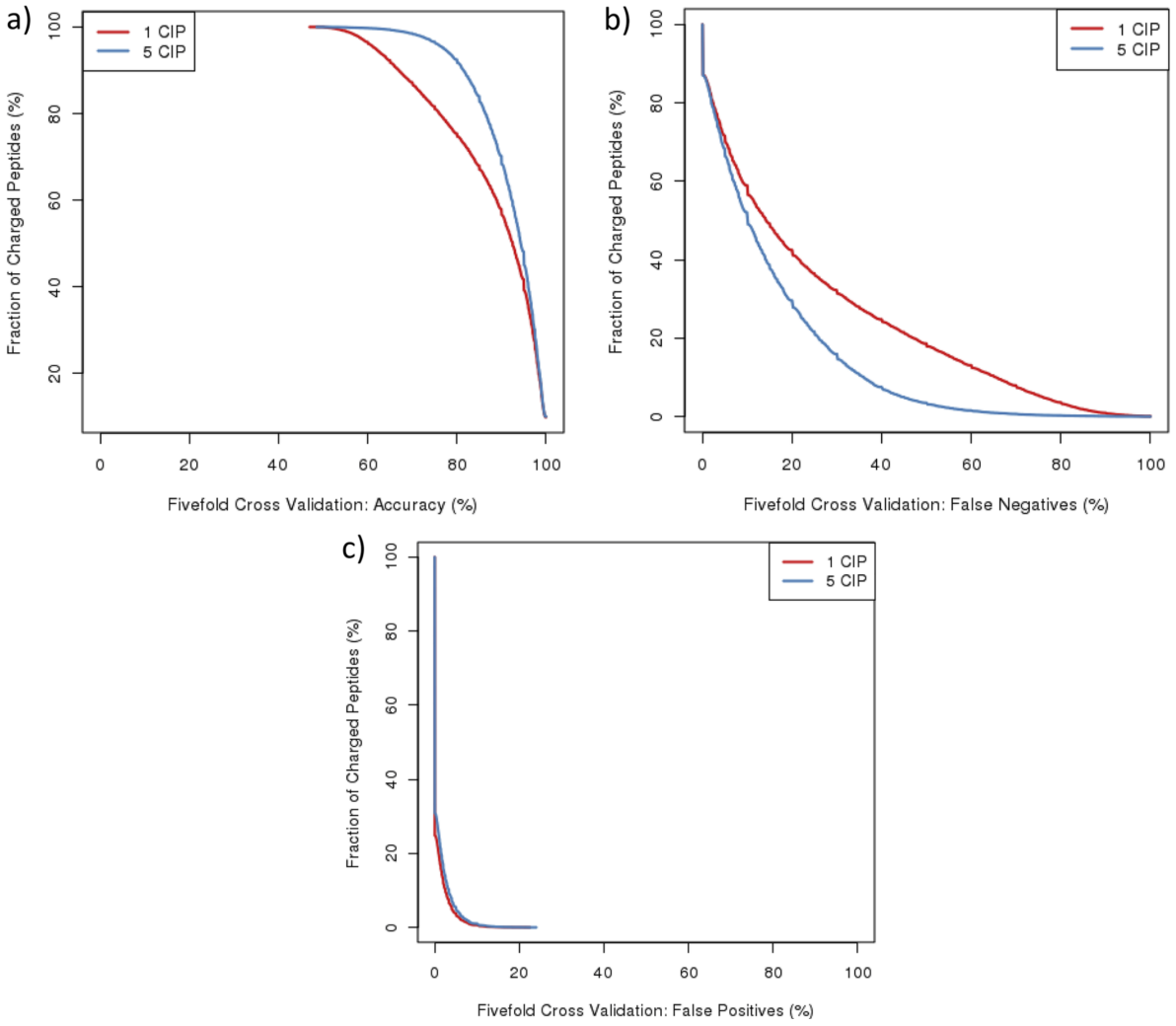


Figure 2.5: Assessment of accuracy using permuted decoy spectra, as described in the Experimental Section, cumulative plot. (a) Comparison of the overall identification accuracy between the single characteristic intensity pattern (CIP) approach (red line) and multiple CIPs (MCIPs) (blue line). Significant improvement upon integration of MCIPs is visible. (b) Effect of MCIP integration on false negatives: false negatives rate is strongly decreased as now also the differently fragmenting ions are integrated. (c) Effect of MCIP integration on the false positives rate, which is only marginally increased upon integration of MCIPs.

we see (as expected from the spectral coverage results) a strong decrease in false negatives upon integration of MCIPs. At the same time, we see that the false positive rate displayed in Figure 2.5c is very mildly affected. The results of integrating all MCIPs instead of 5 MCIPs are virtually identical.

2.4.6 Analysis of a same-sample DDA and SWATH data Set

One of the current applications of spectral libraries is the analysis of SWATH data [34]. In this setup, CIPs are matched with more complex MS2 spectra. Larger spectral libraries can

increase performance [81] but also show more stringent cutoffs after multiple testing correction [82]. As described in the Expermental Section, we used a data set where the same sample had been identified using DDA and SWATH. We then utilized this setup to derive a spectral library of peptides, which was expected to be in the SWATH data set. We then searched the library patterns against the DDA run as well as the SWATH run with the fraction of nonidentified spectra ("errors") plotted against the similarity threshold (Figure 2.6a). For the DDA run, the results are analogous to the results already presented, with a significant improvement of identification upon integration of MCIPs (red and blue line, respectively). For the SWATH run, we observe lower baseline identification with approximately 20% of the patterns not being identified at all, likely due to the higher noise in the SWATH patterns. Nevertheless, also for the SWATH data set we observe very similar effects when comparing the single CIP approach (green) with the MCIP approach (magenta) with an ≈30% increase in identification accuracy at reasonable similarity scores (e.g., 29% for a dot score of 0.6).

2.4.7 Comparing the MCIP approach and SpectraST on a SWATH data set

As described in the Expermental Section, we carried out a comparison of the MCIP approach and SpectraST on a public SWATH benchmarking data set of Navarro et al. [78]. The instrument setup TripleTOF 6600 with 64 variable windows was chosen as it showed the highest performance in the study of Navarro et al. The SWATH peak groups were identified with OpenSwath [34] at peptide and protein level FDR of 1%. A spectral library was created on identical input data for the MCIP approach and for SpectraST. Both libraries were scored against the SWATH data set, and the dot scores with the identified peak groups were extracted. As is visible in Figure 2.6b, all dot scores are lower as compared to the less noisy DDA data. The dot scores generated with the MCIP approach show a clear shift toward higher dot scores. As we saw that the overall dot score distributions were different on SWATH data, we again generated sets of decoy distributions, as described in the Expermiental Section. We see that the decoy distributions of SpectraST and the MCIP approach are both shifted toward very low values, with lower values for the MCIP approach (see Supplemental Figure S12). Using the decoy distributions, we estimated the significance of peptide hits as described in the Expermental Section. In Figure 2.6c we see that the combination of higher dot scores and lower noise levels in the MCIP approach strongly increases the number of significant hits compared to SpectraST.

2.5 Discussion

In our study we introduced a simple and efficient strategy to deal with the heterogeneity of peptide fragmentation. We see that instrument settings can have a huge influence on the peptide fragmentation behaviour, especially for high-energy and low- resolution spectra. We have shown that exclusion of dissimilar peptide spectra is overcautious and results in the negligence of many potential hits. We observe that even under fixed experimental conditions spectra can vary from each other. This effect is strongly enhanced by low-resolution readout. Additionally, very high collision energy changes also have an effect on differing pep-

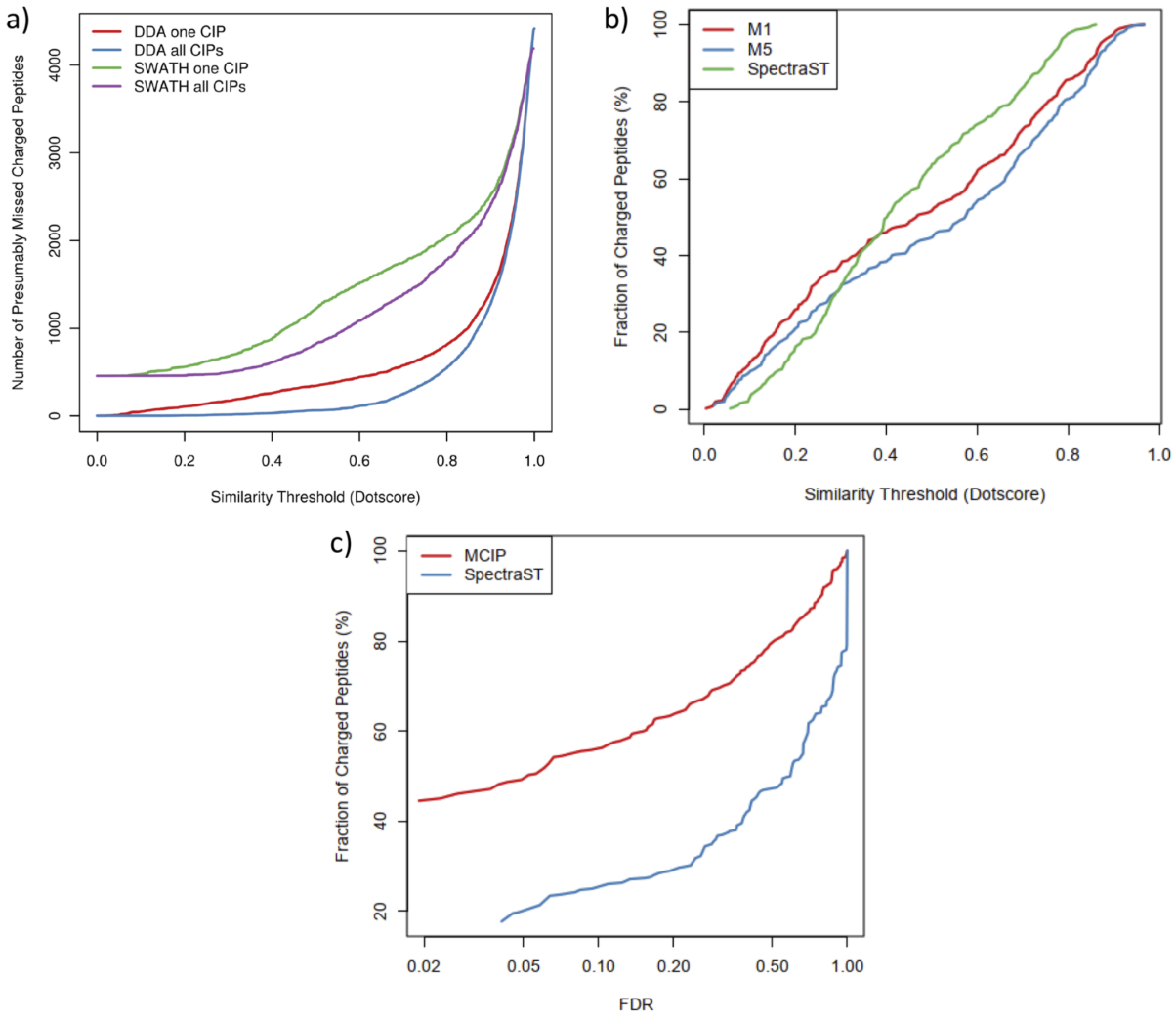


Figure 2.6: Application of the MCIP approach on SWATH data. (a) Repeated analysis of the same sample with DDA and SWATH. For reasonable similarity thresholds, a significant decrease in unidentified peptides can be seen to SWATH data when integrating MCIPs (violet line) in comparison to a single CIP (green line). Analogous behaviour is seen for the DDA approach, with a significantly smaller number of missed peptides in both cases (blue line MCIPs, red line single CIP). (b) Dot score distributions when scoring a publicly available SWATH data set with the MCIP approach and with SpectraST. Shift toward higher dot scores using the MCIP approach is visible. (c) Ad-hoc significance estimation using an empirical estimate for the background noise. Higher significance levels are observed for the MCIP approach compared to SpectraST.

tide fragmentation. Unfortunately, it is beyond the scope of our study to fully explain the differences in peptide fragmentation under fixed conditions. However, we carried out some initial screens using targeted LC-MS/MS runs where we varied the applied collision energy within the same run. This was done to test whether a wrong charge state assignment from the machine could account for the effects. Our results show that dot scores are robust over

a range from −3 to +3 V in most of the cases, which covers differences in collision energy settings caused by wrong precursor charge state assignment (Supplemental Figure S13). We complemented these runs with experiments on the same lysate, where we tried to investigate the influence of the background matrix (co-eluting peptides/ions in the same isolation window). For this purpose, the precursor isolation width was varied between 1 and 5 Da. For broader isolation windows, we observed a systematic enrichment in differently fragmenting spectra (Supplemental Figure S13) and an increase in spectral dissimilarity within the same experimental run (Supplemental Figure S14). For low abundant peptides, it has already been shown that ion interferences with the background matrix can alter the fragmentation spectrum [83], which we also see in corresponding analyses in Supplemental Figures S14 and S15. Recent studies also show this effect with SWATH-MS data [84]. As a low- resolution readout should also strongly increase the effect of interferences, we speculate that the background matrix might be responsible for the observed differences in the fragment spectra. Our reductionist approach of relying on MaxQuant preprocessed spectra comes at the cost of possibly neglecting important spectral information. The dot score values determined from this approach will be different from the dot score values derived from raw spectra, as the representative vectors are shorter and vector length influences the outcome. Nevertheless, heuristic measures to shorten the vector are applied in common library generation tools [27, 29] and have been shown to only mildly affect the overall sensitivity. Additionally, we explicitly tested for accuracy, which is displayed in Figure 2.5 of this study. It should be noted that MaxQuant spectra do not carry fragment ion annotations for fragment ions in charge states larger than 2. To check whether this affected the outcome of the scoring, we repeated our measurements only on peptides in charge state 2 (which should hence not produce fragment ions with charge larger than 2) with no qualitative differences in the outcome. As the examined databases mostly contained unmodified peptides, we carried out our investigation only on unmodified peptides. To get an impression of the influence of modifications on the fragmentation behavior we repeated parts of our analysis on a subset of modified peptide charge pairs (Supplemental Figure S16). In this analysis we compared spectra with the same modification against each other. We see no clear differences from the unmodified peptide charge pairs and would hence expect no systematic influence of modifications on the overall fragmentation behavior. On the basis of our findings, we conclude that even though considerable efforts are being undertaken to extend the amount of available experimental setups in spectral repositories (as, for example, in the scope of the ProteomeTools [72] project), this might only be part of the solution. Due to the large variety of machine setups available, a public library is unlikely to be a perfect fit for the desired setup (including instrument model, fragmentation mode, collision energy settings, fragment ion readout). Additionally, when a well-fitting spectral library is found, the user has to constrain to the parameters of the library, which comes at the cost of flexibility in tuning the machine setup. However, even when this is fulfilled, the user is not able to utilize the full amount of spectral data available online, as only the peptides available in the specified setup can be used. As we have shown, using MCIPs frees the user of these constraints and hence improves the usability of spectral resources. With the advent of quantitative DIA methods like SWATH, the phenomenon of MCIPs becomes important in the context of quantification. If MCIPs are not taken into account, in a significant fraction of cases, fold changes might be miscalculated because peptides that are actually there will be missed because the fragmentation spectrum

is different. The increase in spectral recognition of our SWATH data set upon integration of MCIPs (Figure 2.6) is a first indication that SWATH benefits from our approach.

Chapter 3

Detecting regulated proteins in MS-proteomics data

Motivation

In chapter 2, we have introduced a new computational approach to identify peptides in MS proteomics data. However, simply identifying peptides (and subsequently deriving proteins) in a sample is often not sufficiently instructive, as coverage of the proteome is in general not complete [10, 9]. Hence, if a protein is not detected, this is certainly no proof of absence. While in some cases it can already be biologically interesting to detect certain proteins (for example splice isoforms [85]), biological insight is more often generated from studying the regulation of the proteome: when a biological system is perturbed or its boundary conditions change, this often manifests in a proteomic response, meaning that some proteins increase or decrease their abundance [43]. Studying these responding proteins often allows inference of underlying biological processes or the regulation of the biological system. Many current proteomics experiments are hence *quantitative* and provide peptide abundance estimates that allow to detect such changes. As high noise levels and systematic biases are inherent in quantitative MS proteomics data, the protein change cannot simply be 'read off' of the abundance estimates [13]. Rather, detecting changing proteins from these abundance estimates is a computational and statistical challenge, which is commonly termed *differential quantification*. One priority is to reduce the noise in the data, another priority is to accurately estimate the noise and to maximize the statistical power.
In the following chapter, we present a novel approach to differential quantification. A main difference between our method and other state-of-the-art methods is that we put a strong focus on precisely estimating the noise inherent in the measurements. This allows us to immediately embed every sub-measurement into a 'noise context' which allows us to estimate the statistical significance. We present an approach to retain these sub-estimates throughout the whole analysis pipeline, which increases the statistical power. Our computational approach substantially increases the performance as compared to other state-of-the-art approaches, giving more than 100% sensitivity increases on a benchmarking dataset and more than 1000 additional significant proteins in biological datasets.

Publication

This research was was originally published in Molecular and Cellular Proteomics. C. Ammar, M. Gruber, G. Csaba, and R. Zimmer. *MS-EmpiRe utilizes peptide-level noise distributions for ultra sensitive detection of differentially expressed proteins.* Mol Cell Proteomics. 2019; Vol 18, no. 9: pp. 1880–1892: ©the American Society for Biochemistry and Molecular Biology or the authors.
Accessible under: https://www.mcponline.org/content/18/9/1880
Here, the reformatted manuscript is presented with minor modifications. Supplemental materials can be found online with the publication
I also presented this work as a promoted speaker at the 2018 World Congress of the Human Proteome Organization (HUPO).

Author contributions

I carried out initial analyses on proteomics quantification and peptide noise, which initiated the project. The model was jointly designed by Gergely Csaba, Markus Gruber and me. Gergely Csaba implemented the prototype for empirical error distribution scoring, which Markus Gruber translated into R. I designed the experimental benchmarking parts of the study and carried out the analyses together with Markus Gruber and Gergely Csaba. Gergely Csaba and Markus Gruber designed the simulations. Gergely Csaba, Markus Gruber and I analysed the simulations. Markus Gruber and I wrote the methods section of the manuscript and I wrote the remaining manuscript with suggestions from Markus Gruber, Ralf Zimmer and Gergely Csaba. Ralf Zimmer supervised method development, bioinformatics analyses and the writing of the manuscript.

3.1 Abstract

Mass spectrometry based proteomics is the method of choice for quantifying genome-wide differential changes of protein expression in a wide range of biological and biomedical applications. Protein expression changes need to be reliably derived from a large number of measured peptide intensities and their corresponding peptide fold changes. These peptide fold changes vary considerably for a given protein.
Numerous instrumental setups aim to reduce this variability, while current computational methods only implicitly account for this problem. We introduce a new method, MS-EmpiRe, which explicitly accounts for the noise underlying peptide fold changes. We derive dataset-specific, intensity-dependent empirical error fold change distributions, which are used for individual weighing of peptide fold changes to detect differentially expressed proteins (DEPs). In a recently published proteome-wide benchmarking dataset, MS-EmpiRe doubles the number of correctly identified DEPs at an estimated FDR cutoff in comparison to state-of-the-art tools. We additionally confirm the superior performance of MS-EmpiRe on simulated data. MS-EmpiRe requires only peptide intensities mapped to proteins and, thus, can be applied to any common quantitative proteomics setup. We apply our method to diverse MS datasets and observe consistent increases in sensitivity with more than 1,000 additional significant proteins in deep datasets, including a clinical study over multiple patients. At the same time, we observe that even the proteins classified as most insignificant by other methods but significant by MS-EmpiRe show very clear regulation on the peptide intensity level. MS-EmpiRe provides rapid processing (< 2 min for 6 LC-MS/MS runs (3h gradients)) and is publicly available under github.com/zimmerlab/MS-EmpiRe with a manual including examples.

3.2 Introduction

A major fraction of current Mass Spectrometry (MS) based proteomics experiments is quantitative in nature and aims at the detection and quantification of differentially expressed proteins (DEPs) between biological conditions [86]. As MS measurements are subject to substantial noise, researchers have to rely on statistical tests which detect changing proteins at a given false discovery rate (FDR). The de-facto value of a quantitative proteomics experiment could hence be defined by the overall sensitivity (i.e. the fraction of all changing proteins, which is actually detected by a statistical test) at a reasonable FDR. Huge instrumental efforts are being undertaken to increase the overall sensitivity [87, 88, 15, 80, 8] nevertheless, protein quantification remains a challenging task. In general, protein level intensities have to be inferred from peptide level intensities. This is complicated by the fact, that two peptides of the same protein - even though they are equally abundant in the sample - can be orders of magnitude different from each other in their measured intensities, for example due to differing ionization efficiencies [89] of the peptides. Additionally, ions with similar mass can interfere with the quantified peptides and distort the signal [86]. As many more low intensity than high intensity signals are present in a sample, interference of low intensity signals is common. A further challenge is due to *missing values* in the data. This denotes peptides that are only identified in some of the samples and missing in the other samples. Several setups are available for quantitative proteomics. In label free quantification (LFQ), each sample is measured in an individual liquid chromatography tandem mass spectrometry (LS-MS/MS) run and the peptide intensities are compared between the runs. In the most widely used setup, peptide intensities are derived from the full (MS1) scans [90]. The sets of peptides identified in each run are often not identical and, therefore, lead to missing values. This problem can be addressed by matching the MS1 peaks in the neighboring runs, but this solves the problem only to a limited extent.
A quantification approach that is less computationally challenging is chemical labeling via tandem mass tags (TMT) [91]. For TMT, up to 11 samples are isobarically labeled on the peptide level and mixed before submission to LC-MS/MS. The labels have reporter ions of distinct masses, which are detected in the fragmentation spectrum. Depending on the machine type, the fragmentation spectrum for reporter ion quantification can be a "classical" MS2 spectrum, or an intensity reduced MS3 spectrum, which is generated by further fragmentation of MS2 fragments [88].
In general, the challenges of protein inference, differing ionization, noisy peptide data and missing values are expressed to a certain degree in all quantitative MS setups and computational approaches have to deal with them appropriately.
A common approach for differential expression analysis is to derive protein level fold changes from the peptide fold changes and to apply statistical tests such as the t-test to assign a significance to it. This approach is for example implemented in the Perseus pipeline [92]. Peptide level models have been proposed [93, 94] and have recently been shown to offer superior performance compared to protein level approaches [95, 96, 97]. A recent implementation is given in the MSqRob package [96]. The majority of peptide level models are based on linear regression, which can be problematic for data with strong distortion, outliers, or small peptide numbers. We propose an orthogonal approach, consisting of the direct assessment of peptide level noise, which we term **M**ass **S**pectrometry analysis using **Empi**rical

and **Re**plicate based statistics (MS-EmpiRe). We introduce empirically generated, intensity dependent error fold change distributions and utilize this for between-sample normalization and to derive differential expression probabilities for each peptide. We then show that these probabilities can be combined to the protein level via a modified Stouffer's method [98]. The data for MS-EmpiRe can be measured with a variety of quantitative proteomics setups, as we need only peptide intensities grouped to proteins as input. We test the performance of the method on a recently published proteome wide benchmarking dataset of O'Connel, Gygi and coworkers [99], containing LFQ as well as TMT-MS3 data. With MS-EmpiRe we observe up to 121% more sensitivity in comparison to the approach reported in O'Connell et al. We additionally compare our approach with the peptide level tool MSqRob and see similar performance increases. On simulated differential expression changes, we see similar performance results as on the benchmarking set and demonstrate MS-EmpiRe's superior classification abilities.
MS-EmpiRe is available as a R package on GitHub (github.com/zimmerlab/MS-EmpiRe).

3.3 Methods

3.3.1 MS-EmpiRe compared to state-of-the-art approaches

We compare our method with two current methods, MaxLFQ with Perseus [90, 92] and MSqRob [95, 96], which have different strategies of solving the challenges associated with MS based protein quantification (Tab. 3.1). For the comparison, we focus on the principle issues of differential quantification: Normalization between different experimental samples, the estimation of the protein fold change, the statistical test applied, derivation of the corresponding test statistic (which represents the protein level change), estimation of the variance parameter(s) of the statistical test and outlier correction. In MaxLFQ, normalization is carried out by minimizing the sum of peptide level fold changes with run specific normalization factors. The sum is taken over all runs, also between conditions, with the underlying assumption that most of the proteins do not change. Several statistical tests can be applied to MaxLFQ data, where the t-test shows best performance [90]. The test statistic used in the t-test is the difference between the mean protein level LFQ intensities per condition. The LFQ intensities are pseudo-intensities derived from the median of the peptide level fold changes. The variance of the t-test is derived from the variance of the LFQ intensities between replicates. Outliers are implicitly taken care of by taking the median of the peptide level fold changes as a reference for the LFQ intensity.
The MSqRob method relies on a linear model, similar to the limma method for microarray data [100]. The linear model describes each peptide intensity as a composition of 4 different effects: The effect of the technical replicate, the effect of the biological replicate, the effect of the peptide specific ionization and the effect of protein level regulation. The normalization step is hence included in the linear model estimation. MSqRob uses the t-test and the test statistic is derived from the protein level regulation effect. Ridge regression and an empirical Bayes approach are used to stabilize the intensity estimates and the protein level variance, respectively. To reduce the effects of outliers, M-Estimation with Huber weights is used, which shrinks the effect of high-residual observations [96].
MS-EmpiRe carries out normalization based on peptide level fold change distributions. Between-replicate fold changes are used, which ensures that the fold change distribution should be centered around zero. The statistical test applied is a slightly modified version of Stouffer's method. This way, individual peptide level probabilities for regulation can contribute to the test statistic. The peptide level probabilities are derived from the peptide fold changes in the context of empirical, dataset-specific and intensity dependent background distributions. The variance estimation is hence carried out via these distributions. This allows a refined weighing of the influence of peptide noise on a given fold change. Outlier detection is carried out by explicitly modeling the influence of outlier signals and downscaling of outlier peptides, as described in more detail below.

3.3.2 Normalization

Mass spectrometry data suffer from sample specific effects, i.e. systematic perturbations which affect whole samples. For instance, the total amount of protein which is processed per run has a significant effect on the signal measured per peptide. Therefore, raw signals

	MS-EmpiRe	MaxLFQ	MSqRob
Normalization	Median/Mode of peptide fold change distributions (only replicate samples)	Minimization of peptide fold changes (all samples)	Estimation of sample specific contribution via linear model
Protein fold change estimation	Maximum likelihood estimation from peptide fold change distribution	Fold change of MaxLFQ protein intensities	Estimation of condition specific contributions via linear model
Statistical test	Modified Stouffer's method	t-test	Moderated t-test
Test statistic	Sum of Z-transformed, background normalized peptide fold changes	Difference of MaxLFQ protein intensities	Estimation of condition specific contributions via linear model
Variance estimation	Empirical and Replicate based peptide distributions	Protein-level sample variance	Empirical Bayes Estimation
Outlier detection	Signal scoring based on outlier distribution, downscaling of outlier peptides	Median of peptide fold changes	M-Estimation with Huber weights

Table 3.1: The two state-of-the-art differential quantification methods MaxLFQ (with t-test) and MSqRob compared against MS-EmpiRe.

originating from two different samples are rarely comparable. To correct for sample specific effects, all signals of a sample are typically multiplied by a *scaling factor*. In the context of RNA sequencing data, which are subject to sample specific effects as well, procedures to detect appropriate scaling factors are introduced e.g. by DESeq and edgeR [101, 102, 103]. While DESeq finds scaling factors by comparison of every sample to a virtual sample, edgeR computes all pairwise scaling factors. Both methods use the median of many gene level fold changes as an estimate for the scaling factor. MaxLFQ [90] is a normalization procedure for mass spectrometry data. Instead of relying on the median, MaxLFQ solves a system of linear equations to identify scaling factors such that the change of peptide signals between any two samples (and fractions) is minimized.

All previously mentioned normalization procedures rely on the assumption that most of the signals do not change between any two samples, even when samples from different experimental conditions are compared. We use the same assumption, but only for samples from the same experimental condition (i.e. replicate samples, which should indeed measure the same peptide values) and use a different factor to normalize between conditions.

Normalization within a condition in MS-EmpiRe is done by single linkage clustering as described in Fig. 3.1. Each cluster contains either one or multiple samples. We start with as many clusters as we have replicates and successively merge the two most similar clusters until we end up with one cluster that contains all samples. Similarity between any two clusters is defined as follows: For two clusters, we compute a *fold change distribution*. We build every possible sample pair between the two clusters and compute the fold change for every peptide which was detected in both samples of a pair. The variance of this distribution is used to determine the similarity between clusters while the median is used as an estimate for a systematic signal shift. To merge two clusters c_1 and c_2 we scale all signals of samples in c_2 by the median. This step yields a new cluster that contains all samples from c_1 and c_2 and in which all samples are shifted onto each other. Single linkage clustering is applied to each condition separately.Samples from two different conditions are then shifted in a similar way, the difference is the selection of the shift parameter. Since we can no longer assume that peptides do not change, except for experimental noise, we propose to use the most probable fold change from the distribution instead of the median. This choice is similar to the idea of centralization proposed by Zien et al. [104]. Instead of enforcing a minimal change between all peptides, this shift only targets the majority of peptides and is still in accordance with the assumption that most proteins do not change. The shift parameter can also be defined

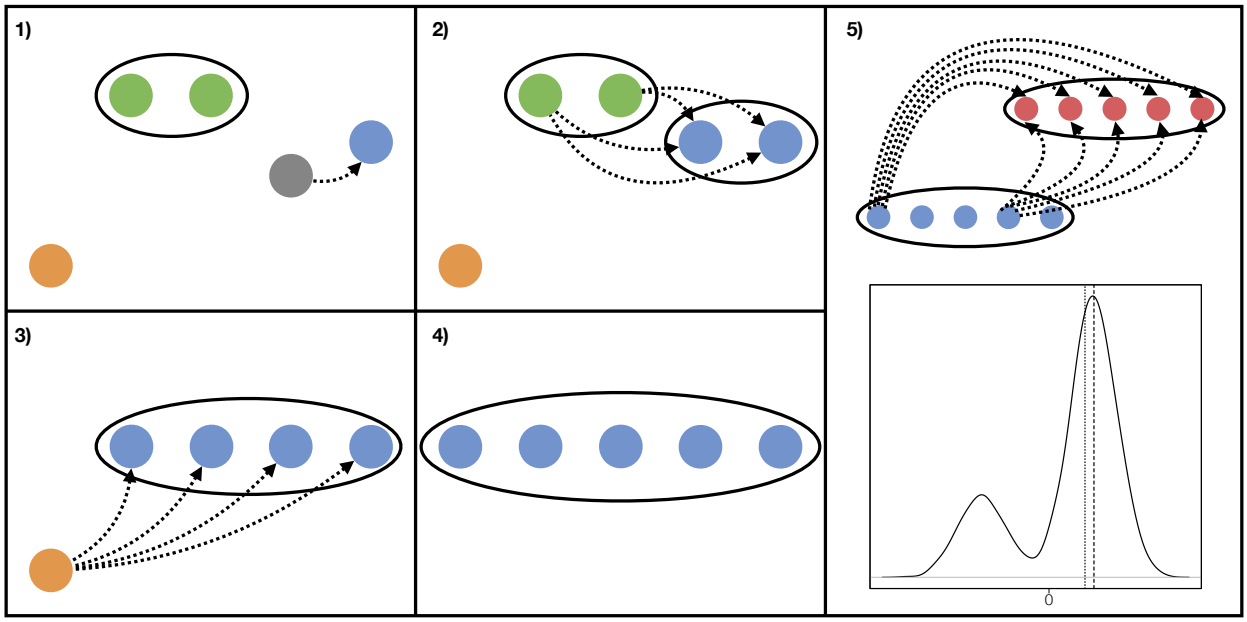


Figure 3.1: Single linkage clustering for signal normalization. **1)** We start with one cluster containing two samples (green) and three clusters of size one. We identify the two nearest clusters (grey and blue) and merged them to one new cluster by shifting the signals of the grey sample according to the median $\log_2$ fold change to the blue sample. **2)** We merge the green and blue cluster. Since they both consist of multiple samples, we determine the shift parameter by computing signal fold changes between any possible inter-cluster sample pair. **3)** The last cluster merge step **4)** The algorithm results in one cluster containing all samples. **5)** The two final clusters of different conditions are merged (blue and red). The resulting distribution of all inter-cluster fold changes is shown below.

by the user using prior knowledge for certain subsets of proteins for example. After the normalization, signals for the same peptide between any two samples are comparable.

3.3.3 Empirical error distributions

Our goal is to detect DEPs between different conditions. However, only peptide level measurements are available from current standard MS experiments and protein level changes have to be inferred. We argue that each peptide level change should be assessed in context of the noise associated with the measurement. MS-EmpiRe is therefore centered around replicate based *empirical error distributions* (Fig. 3.2.1 and 3.2.2). The empirical error distribution is fully based on the data and derived as follows: We compute the $\log_2$ fold change of every peptide signal between any two replicate pairs in each condition. As the $\log_2$ fold change between replicate samples should be zero, each deviation from zero can be seen as an error. This results in one large collection of errors, approximately $C \times \frac{N(N-1)}{2} \times P$ for C conditions with N replicates each and P detected peptides. Since we observed that the variance of peptide measurements depends on signal strength (Fig. 3.3e) we decided to split the complete distribution into intensity dependent sub-distributions. Each of the resulting

sub-distributions contains just a subset of all peptide fold changes. For the construction we sort the peptides ascending to their mean signal strength. We slide a window over the sorted list of peptides to determine the relevant subset for each distribution. The window size and how far it is shifted in each step are parameters that can be controlled by the user. Adjusting it can increase the resolution of the sub-distributions at the cost of computational time. The default window size is 10% of the total number of peptide measurements with a maximum of 1200. The size of the shift is set to $\frac{1}{8}$ of the actual window size. Note that each peptide may appear in multiple sub distributions if the shift is smaller than the window size. To assign each peptide to only one sub-distribution, we save the mean signal of the first and last element of each distribution. We then calculate the distance of the mean signal to the start and end of each distribution for every peptide. Each peptide is then assigned to a distribution such that the minimum of those two distances is maximized, i.e.

$$\max_{sub}(\min(|sub_{start} - s|, |sub_{end} - s|)) \tag{3.1}$$

After this step we have a collection of empirical error distributions that describe the observed measurement errors in relation to signals. Any observed peptide fold change can now be put into context of the background noise. This allows us to determine the probability of the peptide fold change under the corresponding empirical error distribution. We denote this probability as the *empirical p-value* (see also Supplemental Figs. 1&2).

3.3.4 Merging scores over replicates

We can now determine the empirical p-value for every peptide between any two samples. What we rather want, however, is the same information for whole proteins between two conditions including replicate data. This means we have to express the empirical p-value in terms of a score that we can combine over replicates as well as peptides. Furthermore the score should be able to distinguish between negative and positive fold changes. This way we can identify groups of peptides that consistently show the same direction of change between multiple replicate pairs. One score fulfilling these criteria is the Z-value, i.e. a score that follows a standard normal distribution. We can transform an observed fold change into the corresponding Z-value as follows:

$$Z_{fc} = \phi^{-1}(p_{emp}) \tag{3.2}$$

where ϕ^{-1} is the inverse of the cumulative distribution function of the standard normal distribution and p_{emp} is the empirical p-value. This is analogous to Stouffer's method [98] for combined probability tests.

This means we can transform any empirical error distribution to a standard normal distribution (Fig. 3.2.4). In the following sections we will show how those Z-values can be transformed to joint probabilities over replicate data as well as multiple peptides.

To distinguish between background noise and signals, usually not only 2 samples, but N vs M replicate measurements from two different conditions are compared. Those yield up to $N \times M$ scores per peptide which are merged to make a protein-level statement between the two conditions. Under the null hypothesis of no change, each of the $N \times M$ Z-values follows a standard normal distribution. Under this assumption, we can simply compute the sum of the

$N \times M$ standard normally distributed Z-values which follows a normal distribution as well. Looking at the sum is reasonable for the following reasons: It should become extreme only if we have multiple measurements that consistently deviate in the same direction. Too few too weak deviations are canceled out by the non deviating measurements. The same is true for strong deviations in different directions. The mean of the resulting normal distribution is zero as it is the sum over the individual means. In general, the variance of a random variable that is the sum over multiple random variables is the sum over the full covariance matrix of the variables, i.e.

$$var\left(\sum_{i=0}^{N\times M} X_i\right) = \sum_{i=0}^{N\times M}\sum_{j=0}^{N\times M} cov\left(X_i, X_j\right) \tag{3.3}$$

The means and variances are known for each of the variables since they follow a standard normal distribution. We are also able to compute the covariances for dependent variables. This is necessary because some of the possible sample comparisons are not independent, in particular any two sample pairs that share either the first or second sample. It can be shown that the covariance of two Z-value random variables that share one of the samples is 0.5 (Supplemental Material, Section 1 and Supplemental Fig. 3). For each peptide, we can now assess the unexpectedness under the previously derived background distribution over all sample pairs.

3.3.5 Correcting for outlier measurements

One problem about the sum described in the previous section, is that it is susceptible to single outlier measurements. A single extreme Z-value can be sufficient to make the resulting sum significant. This is because of the null hypothesis that *each* of the sample pair comparisons must not be differential. We therefore introduce a correction to estimate the probability, that a single outlier shifts the distribution towards higher values (Fig.3.2.5). For this correction, we estimate the Z-value of the peptide when it is not regulated (Z_{normed}) and substract it from the original Z-value (Z_{orig}). Z_{normed} is estimated as follows: We compute all possible fold changes of the peptide between two conditions (replicate 1 vs. replicate 1, replicate 1 vs. replicate 2, etc.). This results in a (very small) fold change distribution. Analogous to section 3.3.2, we use the median of this distribution as a scaling factor and shift all signals of the second condition by the median. This minimizes the difference of signals between the two conditions and simulates a non-regulated peptide. We again compute the summed Z-value for those shifted peptides, i.e Z_{normed}. If the peptide measurements were differentially regulated previous to the shift, Z_{normed} would be less extreme than Z_{orig}. If the shift does not change the signal, Z_{orig} and Z_{normed} are more or less the same. We can hence introduce a new value $Z_{corrected} = Z_{orig} - Z_{normed}$, which denotes the difference between a regulated and a non-regulated shift. We now want to use the distribution of $Z_{corrected}$ to estimate, how unlikely an observed $Z_{corrected}$ value is. The higher $abs(Z_{corrected})$ is, the more extreme the original measurement was. However there exists no closed form for the distribution of $Z_{corrected}$. We therefore sample such a distribution by simulating a set of non-differential measurements. For each simulated measurement, we compute $Z_{corrected}$. Similar to Eq. 3.2, we look up the cumulative probability of a measured $Z_{corrected}$ in the simulated distribution

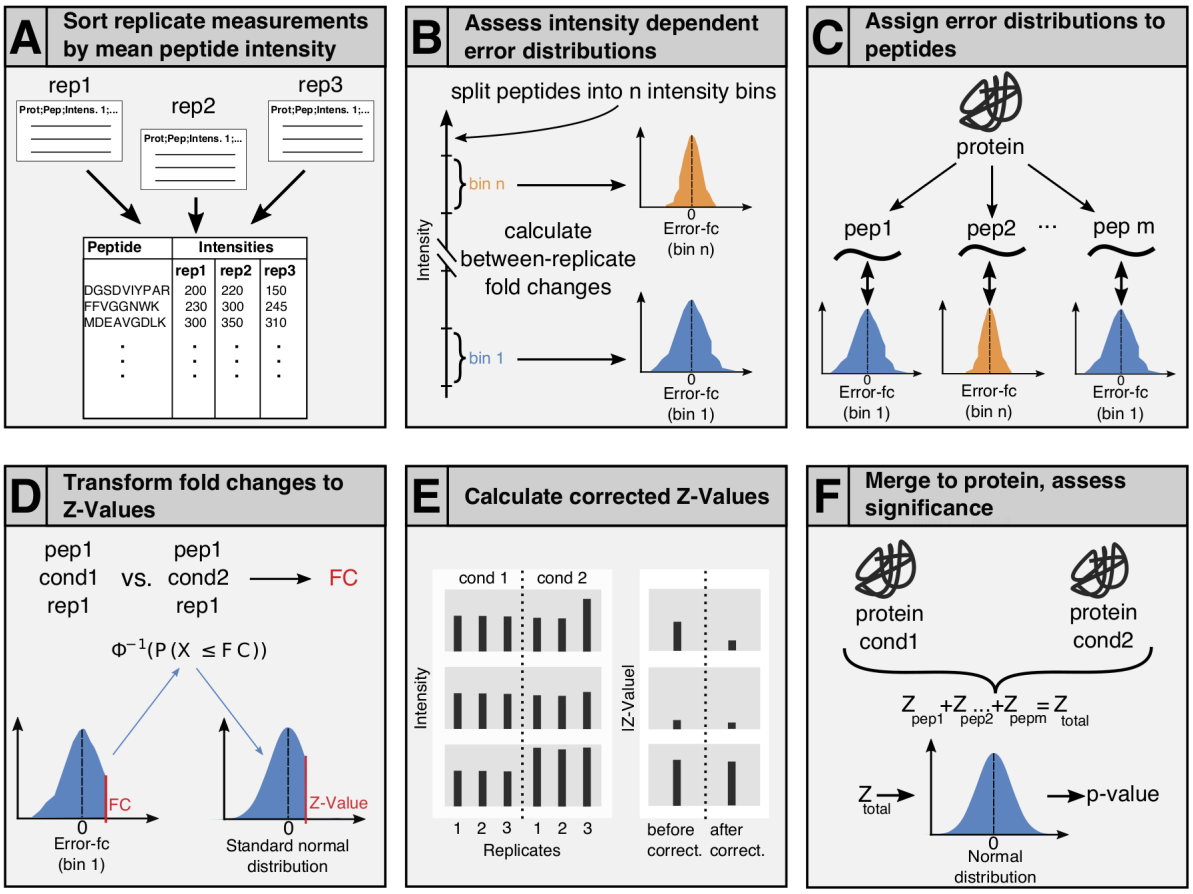


Figure 3.2: Schematic of the MS-EmpiRe workflow. 1) All identified peptides from a proteomics run are sorted by their mean intensities. 2) The peptides are the split into subgroups based on their intensity. For each subgroup, the *error fold changes* of the individual peptides are calculated. An error fold change simply denotes the log_2 fold change of a peptide between two replicate conditions. All error fold changes within a subgroup form an *empirical error distribution*. Distributions corresponding to lower intensity peptides show a larger variance than for high intensity peptides. 3) When a protein is tested for differential expression, each peptide gets assigned an empirical error distribution. Peptides of similar intensities can get the same distribution assigned. 4) For each peptide fold change, the probability that this fold change happened by chance (e.g. the p-value) is assessed from the empirical distribution. This means that the same fold change will get a much lower p-value when the distribution is wide as compared to when it is narrow. To make this value manageable, the p-value is then transformed to a Z-value, by transferring the mass of the empirical probability distribution to a standard normal distribution. 5) The Z-values for each peptide are corrected for outliers. For this, the probability is estimated that a high Z-value on the peptide level has happened by chance due to individual outliers. 6) The corrected Z-values can directly be summed to the protein level, and the corresponding protein-level p-value can be obtained as well as the FDR after multiple testing correction.

of $Z_{corrected}$ and transform its empirical p-value to a Z-value. Whether this correction is needed may depend on the data. The null distribution of the corrected score is a standard normal distribution.

3.3.6 Correcting for outlier peptides

The previous section shows how to correct for outlier measurements of single intensities. Peptides that show a much more extreme change over many samples than the remaining peptides of the same protein have to be accounted for separately. To detect outlier peptides, we compute fold changes for every peptide and sample pair of the same protein. If the median of a single peptide is more extreme than the 25%/75% quantile of the whole protein distribution and protein and peptide are shifted in the same direction, it is marked as an outlier. To compute Z_{normed} for an outlier peptide, we shift the signals by the 25%/75% quantile of the protein distribution instead of the median of the peptide distribution. This modified shift results in a less extreme $Z_{corrected}$ for the peptide.

3.3.7 Combining the peptide scores

What we have so far is a Z-value that expresses how likely it is for a peptide fold changes over all possible replicate comparisons to occur by chance. Each of those Z-values follows a standard normal distribution.

Similar to the first step of merging the peptide scores over all replicate pairs, we can join those scores for all peptides from the same protein (Fig. 3.2.6). In contrast to measurements for the same peptide from different sample pairs, peptide measurements can be regarded as independent measurements for the same protein. This means that under the null hypothesis that every peptide score is a standard normal distributed variable, the sum of such peptide scores is distributed

$$\sum_{i=0}^{P} Z_i \sim \mathcal{N}(0, P) \tag{3.4}$$

with P being the number of different peptides mapped to a certain protein. Using this sum of peptide scores we can now express the probability of a protein under the null hypothesis of no change while taking into account all replicate measurements. To correct for multiple testing we finally apply the Benjamini-Hochberg false discovery adjustment.

3.3.8 Re-processing of the proteome wide benchmarking dataset

We downloaded the raw data of the study of O'Connell et al. [99] from the PRIDE repository PXD007683 and processed the TMT as well as the LFQ dataset with MaxQuant [39] version 1.6.0.16 with standard settings and the respective quantification set (11 plex TMT-MS3 and LFQ). The LFQ data consisted of 11 single-shot runs and for the TMT data 10 runs corresponding to 10 fractions were available. The mapping of the raw files is available in Supplemental Tab. 1. Each set contained three conditions: 10% yeast spike-in, 5% yeast spike-in, 3.3% yeast spike-in. For 10% yeast spike-in, three replicates were measured, for

5% and 3.3%, four replicates were measured. The datasets were searched against a combined yeast (7,904 entries) and human (20,317 entries) database downloaded from Uniprot (04/2018, reviewed). Specific digestion was set for trypsin with two missed cleavages allowed. Carbamidomethylation of Cysteine was set as a fixed modification and Oxidation of Methionine and N-terminus Acetlyation were set as variable modifications. 20ppm mass tolerance were set for precursor ions und 0.5m/z were set for fragment ions. Results were filtered to 1% FDR at the peptide-spectrum-match (PSM) and protein level. For the LFQ data, the "match between runs" option was set (default configuration). We compared the number of identified proteins with the results reported from O'Connell et al. and observed only slight differences, possibly due to different databases and MaxQuant versions (-2% for TMT and -1% for LFQ for our protein). Hence the numbers are a bit different but should not influence the overall outcome. We found around 400 proteins after filtering in the MaxLFQ with Perseus setup, which agrees well with O'Connell et al. As we used MaxQuant instead of SEQUEST [17] for the TMT analysis, we had a noticeably lower identification rate (around 100 proteins less compared to the SEQUEST results presented in the main text of O'Connell et al.). This is similar to the MaxQuant results for TMT reported in the supplement of O'Connell et al.

3.3.9 Processing of the different proteomics studies

We tested and compared MS-EmpiRe on three different proteomics studies of Sharma et al. [105], Ramond et al. [106] and Ping et al. [107]. We downloaded the MaxQuant processed data of Sharma et al. and Ramond et al. directly from the corresponding PRIDE repositories. For the Ping et al. data, we downloaded the raw files from the PRIDE repository and processed the TMT data analogous to the method described above.

3.3.10 Filtering of the benchmarking dataset

Between the different tools, we noticed large differences in the number of proteins that are actually submitted to statistical testing. MaxLFQ with Perseus showed the most conservative filtering, while MSqRob was most permissive. The decision whether or not to accept proteins with only a single quantified peptide value had the most impact on the filtering. With only one peptide per protein, a misidentified peptide can immediately lead to a false classification. As in MS-EmpiRe a peptide needs to be consistently quantified over multiple replicates to gain significance, the probability for such an event decreases and we hence decided to use a less conservative filtering of only one peptide per protein. We also compared the one-peptide with the two-peptide approach and observed no significant effects on the FDR (see Supplemental Fig. 5). This underlines the fact that MS-EmpiRe is designed appropriately to deal with sparse peptide evidence caused by many missing values. For filtering of MS-EmpiRe the following peptides/proteins were excluded: reverse peptides and contaminants, peptides mapping to yeast as well as to human and proteins quantified in only one replicate. As yeast and human are on very distant branches of the evolutionary tree there are many changes in the protein (and thus peptide) sequences even for homologous proteins. By excluding peptides mapping to yeast as well as to human proteins, we ensured a clear mapping of every peptide to either yeast or human, without having to exclude many peptides. This

way less than 0.5% of all peptides are excluded from the analysis, even though a large fraction of yeast proteins is homologous (LFQ: peptides used: 58,805 reverse/contaminants: 328 (0.56%) organism-unique: 58,240 (99.04%) organism-ambiguous: 237 (0.40%), TMT: peptides used: 61,267 reverse/contaminants: 312 (0.51%) organism-unique: 60,707 (99.09%) organism-ambiguous: 248 (0.40%)).

3.3.11 In silico benchmarking

The HeLa background proteins from the study of O'Connell et al. were normalized via MS-EmpiRe and each sample was considered a replicate measurment. This resulted in 11 quasi-replicate runs, out of which 6 were randomly chosen. The 6 replicate measurements were split into two sets with 3 replicates each. One of the sets was chosen for in silico expression changes. For the selected set, a subset of the proteome was chosen and was artificially ̈regulated ̈. For each protein in the subset, an expression change factor was drawn from a distribution. The peptide level changes for the protein change were then sampled around this factor. The changed and the unchanged subset were then compared as two separate experiments with MS-EmpiRe. As in the benchmark it was known which proteins were regulated and the differential quantification performance (sensitivity, specificity etc.) could be assessed.

3.4 Results and Discussion

3.4.1 Fold change based normalization reveals structure of the benchmarking dataset

We used the benchmarking dataset from the study of O'Connell et al. [99], where yeast is spiked into human cell lysate at different concentrations (10%, 5% and 3.3%) (Fig. 3.3a). Hence, when comparing the abundance of yeast proteins e.g. in the 10% sample with the 5% sample, one expects a fold change of 2 for each yeast protein. The two other combinations 5% vs. 3.3% and 10% vs. 3.3% give a fold change of 1.5 and 3, respectively. When applying a differential quantification algorithm, the changing proteins are known (e.g. the yeast proteins) and, thus, measures like specificity and sensitivity can be assessed. The samples were measured twice, once using a TMT-MS3 approach and once using label free quantification (LFQ). To visualize the normalization procedure employed by MS-EmpiRe, we used the LFQ dataset. Between every sample pair, we calculated the log_2 fold change for each individual peptide. This resulted in a distribution of fold changes for each sample pair, which was either a between replicate, or a between sample distribution. For the between replicate distribution we would only expect deviations from zero due to measurement errors or biological variation. Therefore, we call this distribution the *empirical error distribution*. For the between sample distribution, we would only expect systematic deviations from zero for the regulated proteins. In Fig. 3.3b, the between sample distributions are displayed before normalization. For clarity, human proteins (which should not change at all) and yeast proteins (which have systematic changes applied to them) are displayed separately. We can already see some trends in the distributions, which underlines the fact that the fold change based view is an intuitive measure for quantitative datasets. When applying subsequent between-replicate and between-sample normalization, as described in the methods section, we obtain the visibly clustered distributions displayed in Fig. 3.3c. The human peptides (around 90% of peptides) are not shifted and distributed around 0. The yeast proteins are aligned around the shift that was experimentally applied to them (i.e. the log_2 transformation of 1.5, 2 or 3). If too much or too little yeast had been applied to one of the samples, this would reflect in a stronger deviation in a subset of the replicate distributions. This is not the case in our dataset and we see – with slight deviations- an alignment around the desired value (dashed lines). In a real life example, we would not expect systematic changes around one fold change in one direction, but larger spread deviations in both directions. The example, however, visualizes that a fold change based approach on quantitative data sets is an effective procedure to normalize datasets without altering the structure of the underlying data. In general, the distributions reveal a ubiquitous problem in MS based proteomics data. The data is so noisy, that a lot of the measured yeast peptides do not even show regulation. This is most striking for the 1.5 set, where around one quarter of the peptides show no regulation, or even regulation into the wrong direction. Hence there is no way to classify these peptides correctly by themselves. Since usually, multiple replicates and peptides exist for a protein, the quantification of a protein can be seen as multiple drawings from such a distribution. This underlines that peptide fold changes should always be analyzed in the context of the dataset specific noise.

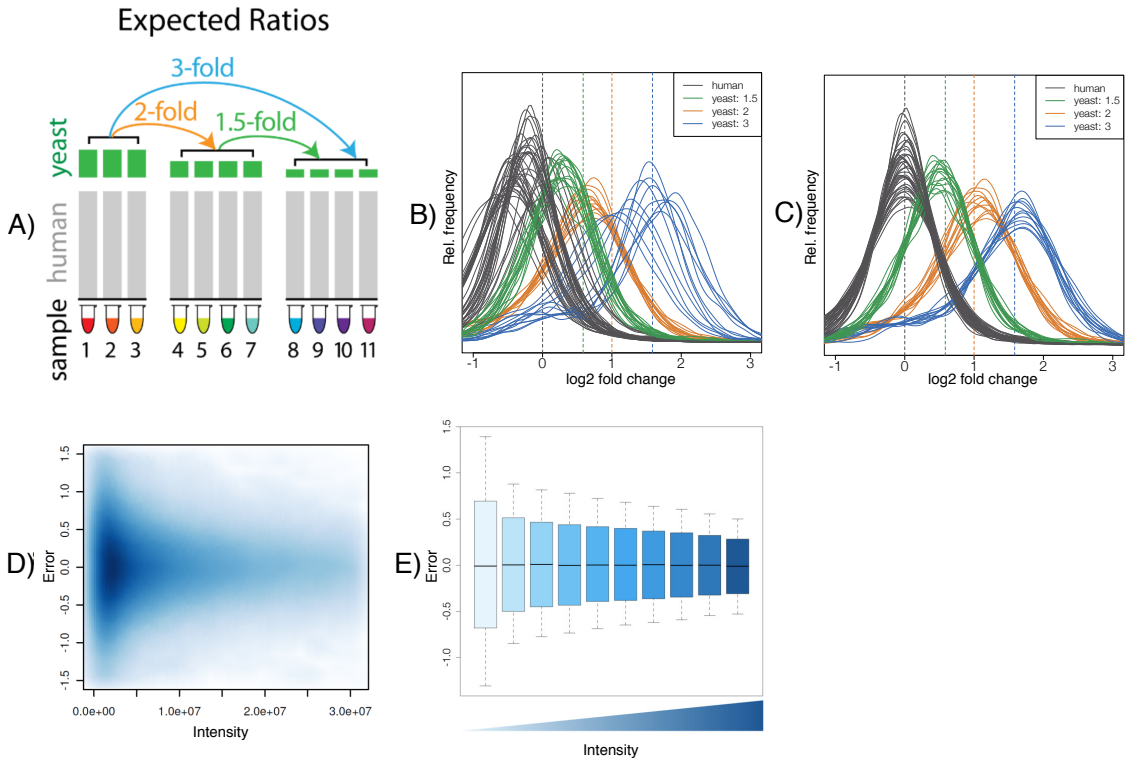


Figure 3.3: Experimental setup and fold change based metrics. a) Benchmarking setup for quantitative assessment of fold changes taken from O'Connell et al. [99]. Different amounts of yeast lysate are spiked into human cell lysate. The three groups contain 10%, 5% and 3.3% yeast lysate, respectively. b) Peptide level fold change distribution between all conditions, before normalization. c) The distribution after fold change based normalization. e) Intensity dependent peptide fold changes between two replicates of the LFQ data (error fold changes), displayed as smoothed density scatter plot. f) Error fold changes for 10 intensity regions displayed as box plots. Each box contains the same number of data points. The quantiles correspond to the fractions 0.05, 0.15, 0.50, 0.85, 0.95.

3.4.2 Assessment of empirical error distributions underlines the importance of context dependent fold changes

As already described in the methods section, one of the key features of the MS-EmpiRe algorithm is the quantitative assessment (and subsequent utilization) of fold change consistency. For this we make use of the fact that the log fold change between replicate measurements should be 0, which allows us to derive empirical error distributions containing the fold changes of peptides between replicate measurements. These distributions have already been discussed in the previous section and the human peptides displayed in Fig. 3.3c are a good example. However, when looking at the empirical error distributions over a whole dataset, as in the previous section, we neglect the well-known fact that low-intensity peptides are subject to significantly more variation than high intensity peptides. An intuitive way to visualize this is by plotting the error fold changes against the mean intensities of the peptides, as displayed in Fig. 3.3d. In this density scatter plot, we see that the majority of peptides

is actually of low intensity and in the low intensity region a large spread of the fold changes is visible. Furthermore, we see global dependence of the error fold changes on the intensity and high intensity peptides are much less prone to deviate strongly from zero. Nevertheless, outliers exist for all intensities, which underlines the need for a more quantitative assessment. This is depicted in Fig. 3.3e, where empirical error distributions are given for distinct intensity bins. The original fold change distribution is split into 10 boxes and each box contains equal amounts of data points. We see that the lowest intensity box is particularly noisy, with around 40% of the data above a fold change of 1.5 (log_2 fold change of around 0.6). Hence, if a low intensity peptide has a 50% intensity increase from one condition to the other, the likelihood that it is an actually irrelevant change is at around 40%. When we consider the highest intensity box, this value drops from 40% to around 5%. Identical fold changes hence have a very different meaning, depending on the context. The MS-EmpiRe approach transforms an observed fold change in the context of its corresponding empirical error distribution and, therefore, quantitatively accounts for this phenomenon. Especially, as every datasets carries its own noise, and e.g. TMT-MS3 data shows significantly reduced noise (see Supplemental Fig. 4) the empirical assessment for each dataset is crucial.

3.4.3 MS-EmpiRe shows up to 121% sensitivity increase in an experimental benchmarking set

The key question addressed in the setup of O'Connell et al. [99] is, how many proteins can be detected as differentially expressed in a proteome wide benchmarking setup. Analogous to their paper, we assessed how many of the experimentally shifted yeast proteins we were able to detect via MaxLFQ coupled to the Perseus pipeline. We then compared this approach with our MS-EmpiRe approach and the more recent tool MSqRob. Perseus was executed analogous to the settings given in O'Connell et al. (reverse and contaminant filtering, at least two replicate measurements per protein and two sided homeoscedatic t-test with Benjamini-Hochberg correction) and MSqRob was executed with default settings. An FDR of 5% was set for all approaches. In Fig. 3.4a and 3.4b, we show the results of the benchmark for the more challenging LFQ setup. The number of peptides available for testing differs markedly, depending on how conservative the corresponding tool is in filtering peptides for quantification (see also methods section). MS-EmpiRe clearly outperforms MaxLFQ+t-test in terms of sensitivity, with up to 120% more DEP detections in the fold change 1.5 setup. When comparing MS-EmpiRe with MSqRob, it seems that MSqRob is slightly more sensitive. However, for MSqRob the observed FDR (i.e. the number of human proteins detected) is between 9% to 15% instead of the required 5% for MSqRob. MaxLFQ+t-test and MS-EmpiRe also violate the FDR in the fold change 2 setup, but only by 1 and 2 percentage points, respectively. To make the sensitivity analysis more comparable, we show an "FDR corrected" bar for MSqRob, where we set the FDR cutoff of MSqRob to a more stringent value (see methods section), such that the actual FDR is also at around 5%. In this setup MS-EmpiRe outperforms MSqRob in terms of sensitivity in all cases and detects more than twice the number of proteins of MSqRob in the most challenging fold change 1.5 setup. As MS-EmpiRe and MaxLFQ+t-test both violated the FDR for the fold change 2 dataset, we looked at the corresponding data in more detail. We saw that many of the misclassified

human proteins in the dataset were particularly tough cases, where many peptides over many replicate conditions show consistent up- or down regulation (see Supplemental table 2). Of course, cases like this are included in the FDR estimation. However, the condition seems to have more cases of consistent protein up-regulation than expected. We considered "tuning" the FDR estimator to be more conservative, but took into account that the FDR violation was only mild and that neither the other benchmarking conditions, nor the simulations showed further FDR violations. As slight regulation or systematic distortions might always occur under experimental settings, we decided to leave the model unchanged. The setup shown in Fig. 3.4a contains the input sets after the individual filtering applied by each method. This corresponds to a real-life application of the methods, but also reduces the comparability of the classification capabilities. We hence compared each method on the same set of peptides, which consisted of the intersection of all input peptides. This led to a significant decrease in the number of proteins detected. Interestingly, the drastic reduction of input peptides strongly increased the number of detected proteins for MSqRob for the 1.5 set. This implies, that MSqRob is prone to give an over-optimistic scoring to proteins with sparse peptide evidence and hence more stringent filtering might be appropriate. In the fold change 1.5 set, MSqRob also does not violate the FDR constraint. In the two other sets, the peptide filtering does not seem to suffice to control the FDR and also for MS-EmpiRe the fold change 2 set still violates the FDR slightly. Nevertheless, MS-EmpiRe is the most sensitive method over all sets. When comparing the methods on the less challenging TMT data set in Fig. 3.4c, we see overall high sensitivity, which is also discussed in O'Connell et al. Also here, we see a substantial increase in sensitivity of around 20% compared to the t-test, when considering the 1.5 fold change set.

3.4.4 MS-EmpiRe identifies up to 1,200 additional significant proteins in quantitative MS datasets

To test the performance of MS-EmpiRe in different experimental settings, we applied our method to public MS datasets from three different studies. The first study by Sharma et al. [105] was a deep (≈12,000 proteins) LFQ proteomics study of neuronal cell development. For the second dataset, we chose the LFQ study of Ramond et al. [106] that was also tested in the MSqRob study [95]. The data is from a single knockout experiment in *Franciscella* plants (≈1,000 proteins). The last dataset by Ping et al. [107] was a deep (≈10,000 proteins) TMT-MS3 study of Alzheimer's disease (AD), Parkinson's disease (PD) and co-morbid (ADPD) patients. We applied MS-EmpiRe, MaxLFQ+t-test and MSqRob to the LFQ datasets and MS-EmpiRe as well as the t-test to the TMT dataset. For the Sharma et al. and the Ping et al. datasets, many conditions had been measured and we randomly picked subsets for further investigation. For the Sharma et al. dataset, we chose three stages of *in vitro* neuron development and for the Ping et al. dataset, we chose control vs. disease in the anterior cingulate gyrus.

The numbers of differentially called proteins (DCPs) identified by each method and study are displayed in Fig. 3.5a. We see that the number of DCPs differs strongly between the studies, ranging from thousands of DCPs in the study of Sharma et al. to only tens of DCPs in the study of Ramond et al. We see that MS-EmpiRe is the most sensitive method in all

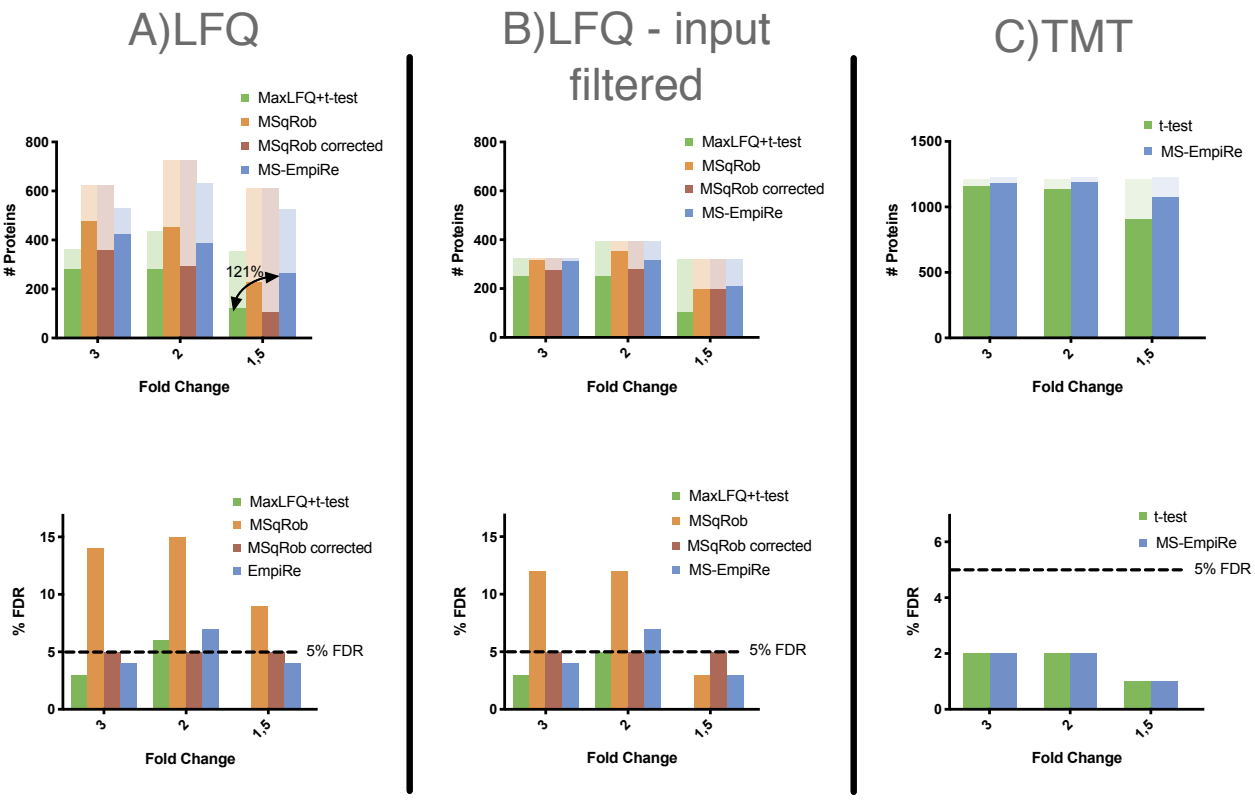


Figure 3.4: Assessment of the differential detection performance on the benchmarking setup of O'Connell et al. a) Number of proteins detected in the LFQ data by the MaxLFQ+t-test setup, MSqRob and MS-EmpiRe. The light shades show the numbers of yeast proteins accessible for testing in the setups, which differ for every method. As MSqRob shows high FDR rates (bottom plot), an FDR-corrected bar is introduced for MSqRob. MaxLFQ+t-test shows low sensitivity at good error rate control similar to MSqRob. MaxLFQ+t-test is very conservative for the fold change 1.5 setup, with no false positives (no bar visible). MS-EmpiRe increases the detection substantially in all cases with good error rate control. This is most pronounced in the most challenging fold change 1.5 setup. b) Number of proteins detected when using an intersected input set. Due to this conservative approach the numbers and differences are lower in general, nevertheless MS-EmpiRe is the best performing method. c) Comparison of MaxLFQ+t-test and MS-EmpiRe on a TMT dataset. MSqRob was excluded as it currently does not support TMT data. The overall performance is better due to higher depth from sample fractionation, lower noise and fewer missing values. Quantification on the protein level hence already works well, still MS-Empire shows a significant sensitivity increase of around 19% for fold change 1.5.

cases. Especially in the deep datasets, we found a highly increased number of DCPs. In the Sharma et al. dataset, up to 1,200 additional DCPs are identified by MS-EmpiRe. In the clinical data of Ping et al., up to 1,000 additional DCPs can be identified when using MS-EmpiRe.

3.4.5 Set-based comparison reveals strong differences between significant proteins for each method

In the previous section, we have shown that the absolute numbers of DCPs are quite different for the individual methods. To check how consistent the results of the methods are, we analysed the overlaps of the sets of DCPs. To assess these overlaps, we restricted on a set of proteins, where regulation should be very clear: we considered all proteins, where the $log2$ fold change estimate was larger than 1 and the FDR was below 0.05 in at least one method. In Fig. 3.5b we see the intersection sets of the individual methods for the Sharma et al. data, comparing *days in vitro* (DIV) 5 with 15. The set of proteins detected by all methods is the largest, but still consists of only 30% of all DCPs. MS-EmpiRe has large overlaps with MSqRob and with MaxLFQ+t-test, while the exclusive overlap between MaxLFQ+t-test and MSqRob is very small. This provides further confidence in many hits of MS-EmpiRe and indicates that MS-EmpiRe is able to close the gap between MaxLFQ+t-test and MSqRob. The remaining sets are hits detected by only one method. Here, MS-EmpiRe has the largest set. The one-method sets have to be treated with special care (high chance for false positives) and they will be investigated in more detail the next section. The intersection sets for the other studies and conditions are displayed in Supplemental Fig. 6. For most other conditions, the detection rate of MaxLFQ+t-test/TMT+t-test is very low and hence the overlap between all three methods is also very small. There is again a large overlap between MSqRob and MS-EmpiRe and large one-method sets. For the Ramond et al. datasets, seven of the eight proteins passing the threshold are in the combined set of all three methods. This indicates, that MS-EmpiRe is not over-sensitive in datasets with little regulation happening (few DEPs).

3.4.6 A detail view on the quantitative data validates the proteins called by MS-EmpiRe

In contrast to benchmarking datasets, for r̈eal-lifeq̈uantitative MS-datasets, a ground truth is not available. We cannot easily decide, whether a DCP is a DEP (i.e. actually regulated). What is possible, however, is to visualize all the quantitative data available for a protein to allow manual inspection in detail. As the quantitative MS data is on the peptide level, we visualize the peptide intensities using *peptide fold change* plots. For this, we assess the fold changes between all replicates of a peptide in condition1 and in condition2 and represent them as a box plot. The fold change plot for a given protein has as many boxes as peptides measured and each box contains as many fold changes as there are replicate pairs between the conditions. In Fig. 3.5c we show for each method the protein with largest FDR difference to the two other methods. This means the selected protein has a significant (small) FDR in one method while both other methods assign it a very insignificant (high) FDR. For MS-EmpiRe, we see highly consistent fold changes in one direction, indeed indicating a DEP. In contrast to the proteins detected by MaxLFQ+t-test and MSqRob, a distinct change of the protein is visible, adding further confidence into the MS-EmpiRe results.
For a comprehensive check, we provide visualizations of all proteins detected in the intersection datasets on the website https://www.bio.ifi.lmu.de/files/gruber/empire/, allowing in detail inspection of the DCPs. For the clinical dataset of Ping et al., we see that the

DCPs show small fold changes in general but very consistent regulation on the peptide level (Fig. 3.5d). This indicates that the precise quantification of TMT-MS3 together with the MS-EmpiRe approach is a powerful combination for the detection of biomarkers or clinically relevant proteins.
On the web pages we also provide further plots on the datasets: comparative volcano plots, comparative FC-scatter plots and plots of the peptide intensities as well as the MaxLFQ intensities for each protein.

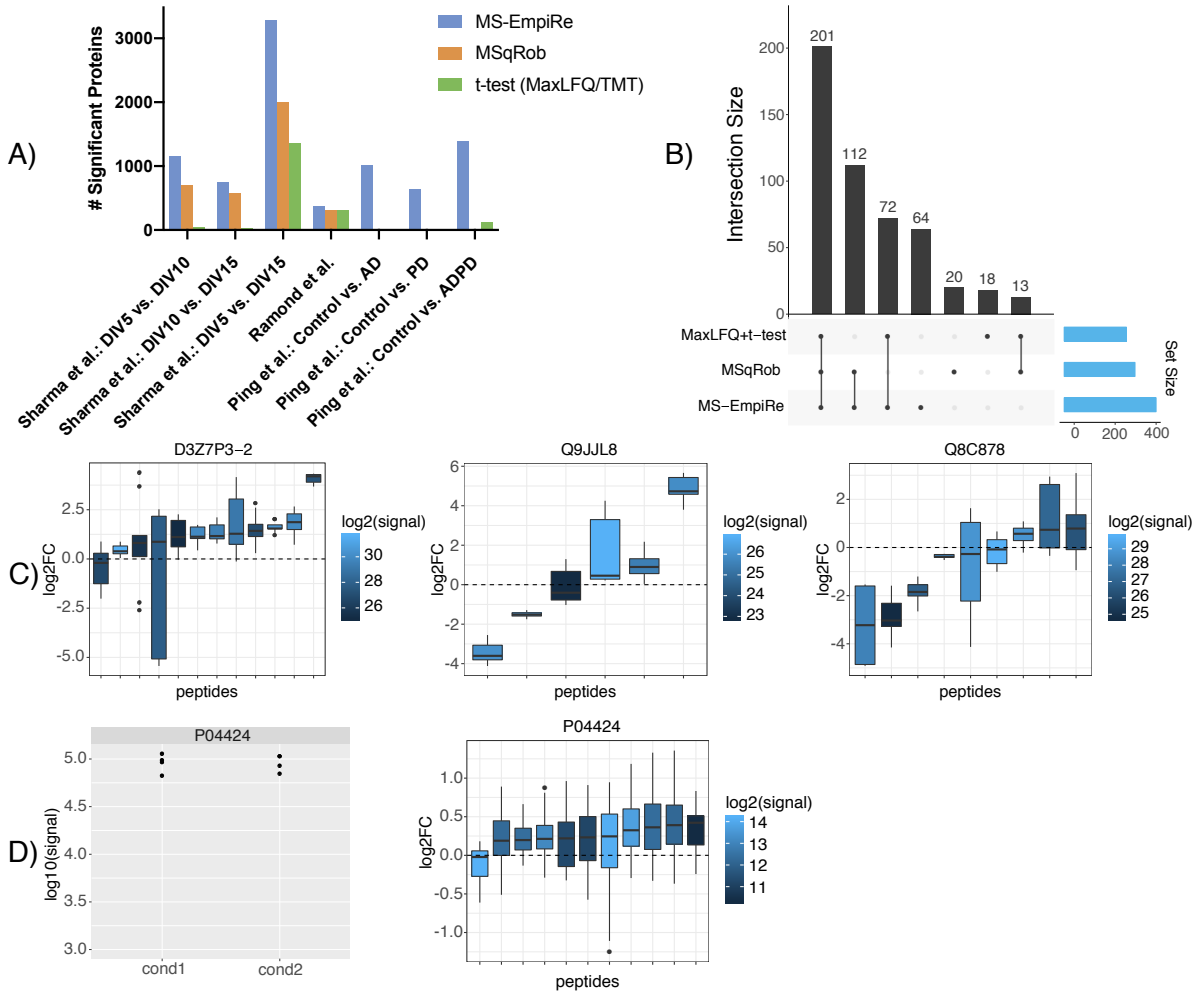


Figure 3.5: Application of MS-EmpiRe, MaxLFQ+t-test and MSqRob to three different quantitative LC-MS/MS datasets. a) Number of DCPs detected in the three different datasets by the three different methods. Each bar represents the number of proteins found in the set determined by the dots below. MS-EmpiRe is the most sensitive approach. b) Overlaps of protein hits on a subset of very clear hits for DIV5 vs. DIV15. MS-EmpiRe shows large overlaps with MaxLFQ+t-test and MSqRob, while the overlap between MaxLFQ+t-test and MSqRob is small. c) Investigation of the proteins called by only one method. Peptide fold change plots for the proteins with the largest FDR difference to the other two methods. A consistent shift of the boxes above or below $log2$ fold change 0 (black dashed line) indicates regulation. Left (MS-EmpiRe): protein D37ZPP3-2 ($FDR_{emp} < 0.01$, $FDR_{msqr} = 0.45$, $FDR_{mlfq} = 0.83$). Almost all peptides imply clear up regulation. Middle (MSqRob): protein Q8C878 ($FDR_{emp} = 0.95$, $FDR_{msqr} < 0.01$, $FDR_{mlfq} = 0.83$). We see varying up- and downregulation. Right (MaxLFQ+t-test): protein Q9JJL8 ($FDR_{emp} = 0.74$, $FDR_{msqr} = 0.61$, $FDR_{mlfq} < 0.01$). We see varying up- and downregulation. d) MaxQuant protein intensities vs. fold change plot for protein P04424 in the clinical dataset ($FDR_{emp} < 0.01$, $FDR_{ttest} = 0.99$). MS-EmpiRe is able to clearly resolve the small but systemtatic fold changes. Many more validation plots for all methods tested can be found under https://www.bio.ifi.lmu.de/files/gruber/empire/.

3.4.7 In silico benchmarking shows high sensitivity and conservative FDR estimation

The importance of experimental benchmarking setups for quantitative proteomics cannot be overstated. Without reference standards, it is impossible to estimate the performance of experimental and computational methods. Unfortunately, performing an experimental benchmarking is cumbersome as it requires very precise mixing and sample handling. Additionally, only constant fold changes can be applied to a given setup, which does not reflect an actual regulative scenario. To complement the experimental benchmarking, we generated an in silico benchmarking set, as described in the methods section. In short, we used the human background proteins measured in O'Connell et al. as replicate measurements and we divided six replicate measurements into two groups. We then applied *in silico* intensity changes on the protein and peptide level to one of the groups and compared the two groups in a differential quantification context. As we know which proteins are "artificially regulated", we can assess measures like sensitivity and specificity analogous to the experimental benchmarking setup discussed in the previous sections. We simulated two setups: one similar to the one in O'Connell et al., where we always applied the same fold change (with some noise) to a sub-fraction of the proteome, including a 10% fraction. Additionally, we simulated a more realistic scenario, where the *in silico* expression changes were not always the same, but were drawn from a distribution. We designed the distribution to be bimodal such that up- and down regulation was possible. The results for sensitivity and precision (i.e. specificity) for LFQ data are depicted in Fig. 3.6. The boxes result from changing different fractions of the proteome (individual simulations where 5%, 10%, … ,40% of the proteome are changed). Surprisingly, we noticed that the fraction of proteome changing significantly influences the sensitivity of the applied statistical test, especially for the MaxLFQ+t-test setup. For example, when 30% of the proteome is changing with a fold change of 1.5, this is better detected by a statistical test as when only 5% of the proteome is changing with a fold change of 1.5. The reason for this is apparently a loss of significance after multiple testing correction. As multiple testing correction can be seen as a shifting of the p values into the direction of a uniform distribution, stronger deviations from the uniform distribution (e.g. many regulated proteins) are less strongly affected. In Supplemental Fig. 7 we see, that the protein level scoring underlying the t-test does not allow a very distinct discrimination between regulated and non-regulated proteins as compared to MS-EmpiRe, which explains the losses in sensitivity with MaxLFQ+t-test. The clearer distinction between regulated and non-regulated proteins by the peptide level tools is also reflected in the fact, that the peptide level tools MS-EmpiRe and MSqRob show less dependence on the proteome fraction in terms of sensitivity. In general, the results of the in silico simulations in Fig. 3.6 show a similar picture as compared to the experimental benchmarking setup. MS-EmpiRe and MaxLFQ+t-test show conservative error estimation, which however comes at the cost of drastically reduced sensitivity for MaxLFQ+t-test. MS-EmpiRe is the most sensitive tool and MSqRob detects only slightly less proteins. However, MSqRob shows problems in terms of error rate control, especially for setups with strong fold changes. This might be due to an over-optimistic error estimation of MSqRob due to the many clear classification cases. Comparing the fixed setup with the setup where we generate dynamic noise, we see that the overall identification rate decreases, whereas the general trends for all three methods

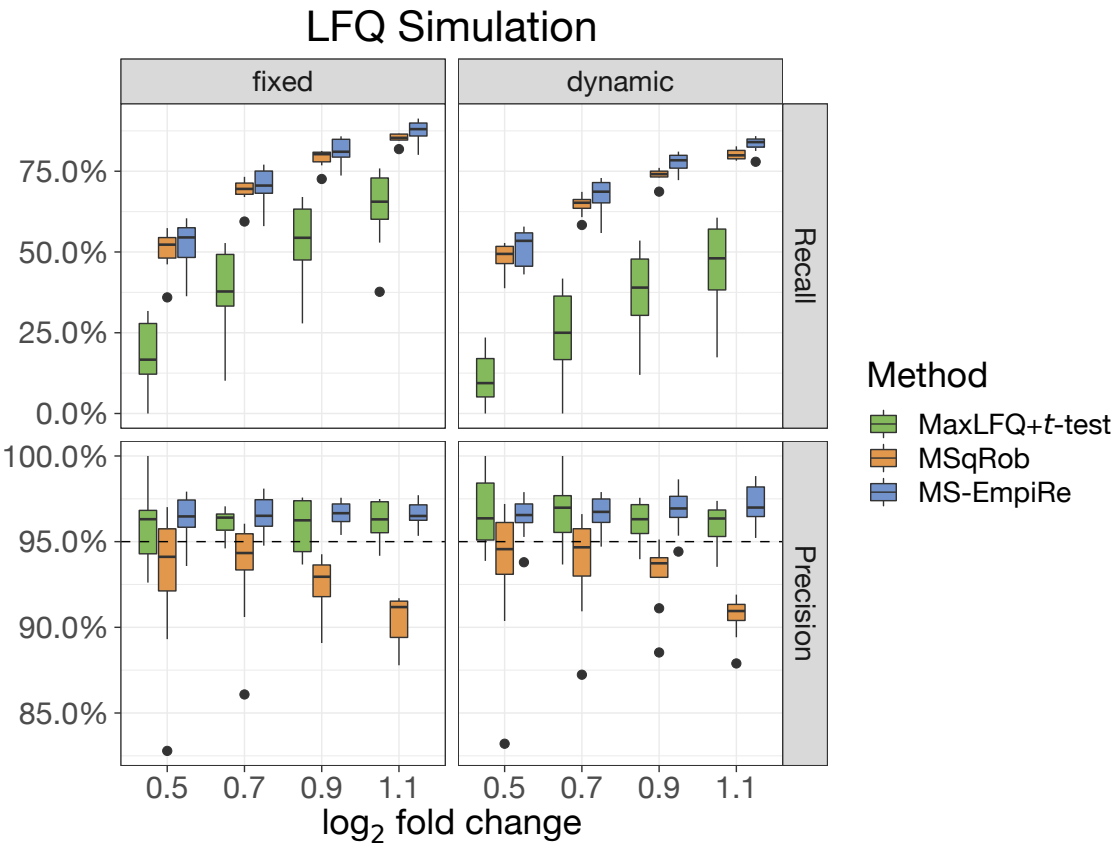


Figure 3.6: In silico benchmarking of MS-EmpiRe, MaxLFQ+t-test and MSqRob. The x tics represent the median fold change by which the data is shifted. The boxes contain the sensitivity/specificity results, when different fractions of the proteome are changed. In particular each box contains eight values corresponding to 5% up to 40% of the proteome changing in 5% steps. Sensitivity and specificity for different fold changes upon constant (left) as well as dynamic (right) proteome changes are shown. A clear dependency on the fraction of the proteome changing is visible. As in the benchmarking set, MaxLFQ+t-test shows low sensitivity with good error rate control, MSqRob shows high sensitivity but often violates the error estimation and MS-EmpiRe shows high sensitivity with good error rate control.

are very similar. Error rate estimation does not decrease and hence all methods show the desired response towards high noise in the data.

3.5 Conclusion

Current Mass Spectrometry proteomics publications often report the number of quantified proteins for a given proteomics setup. This number however, can be misleading, as the number of quantified proteins does not necessarily reflect the number of proteins that can actually be detected in a differential quantification experiment [99]. This is especially the case for more noise-prone proteomics setups, such as LFQ. Given the popularity of such setups, increasing the depth of a proteomics experiment at the level of differential detection becomes an ever more important issue. Considering recent studies [95, 96], it is likely that the future development of differential quantification will increasingly use peptide level information. Peptide level tools like MSqRob show significantly improved sensitivity, though we have shown in this study that FDR control is still difficult in a proteome-wide setup. The popular MaxLFQ+t-test pipeline also implements the most conservative approach in the experimental benchmarking setup. With MS-EmpiRe, we introduce a new peptide level tool that shows high sensitivity at accurate error rate estimation. While FDR estimation for MS-EmpiRe is almost as accurate as for MaxLFQ+t-test in the experimental setup, it even outperforms MaxLFQ+t-test in the in silico simulation. We have shown that MS-EmpiRe gives up to two fold increase in sensitivity for small fold changes (1.5 fold change), which can be highly relevant for biological applications. Even though a fold change of only 1.5 is challenging for a proteomics setup, such a change may already reflect a drastic alteration in a biological system. The key to the sensitivity of MS-EmpiRe is the direct modeling of errors on the peptide fold change level. This gives an immediate statistical weight to individual peptide fold changes, which are then transferred to the protein level via basic statistics and without additional optimizations or parameters. This simple approach relies on the only assumption of consistency between replicates. Therefore, our method heavily relies on good replicate measurements. Even though MS-EmpiRe is able to work with only two replicate measurements per condition, ideally three or more replicate samples should be available, to obtain accurate error estimates. For MS proteomics data, where robust workflows exist and the creation of replicate measurements is a standard, we believe that this requirement matches well with current experimental practices. From our perspective, the consistency of replicates is a minimalistic and reasonable assumption that can be made for proper processing of proteomics data. A possible deviation from replicate consistency might occur, when uncontrolled factors in an biological experiment change between replicates. In our setup, this might lead to an underestimation of DEPs. However, replicate-inconsistent setups are highly critical and should be handled with care. Based on our analysis, we conclude that MS-EmpiRe is currently the most sensitive tool for differential protein detection. MS-EmpiRe requires as inputs only peptide intensities and protein identifications and, thus, applicable to virtually any modern proteomics measurement. MS-EmpiRe is an easy-to-use option for proteomics researchers and helps to improve the quality and biological insight gained from MS proteomics studies.

3.6 Data availability

All proteomics datasets used in this study are publicly available through the ProteomeXchange Consortium (http://www.proteomexchange.org/) via the corresponding PRIDE partner repositories: O'Connell et al. [99]: PXD007683. Sharma et al. [105]:PXD001250. Ramond et al. [106]:PXD001584. Ping et al. [107]: PXD007160. All relevant quantititative protein and peptide intensity tables are available via:

https://www.bio.ifi.lmu.de/software/msempire/index.html.

Chapter 4

Detecting differential alternative splicing in MS proteomics data

Motivation

In chapter 3, we have introduced a novel method for differential quantification, i.e. the detection of regulated proteins between conditions. Differential quantification is one of the most important concepts in quantitative proteomics, mainly because it gives an overview over which genes respond to a given biological perturbation. Such a gene-level overview can be very instructive, however it does not represent the full complexity of the proteomic response. In reality, a single gene can often produce a wide variety of proteins, so called *proteoforms* [108], which vary in their sequence as well as their chemical modifications. MS proteomics currently has very limited capabilities in detecting proteoforms, because proteins are digested into smaller peptides and the information about the full (modified) protein sequence is lost and the sequence coverage is still comparably low [10, 9]. For these reasons, the complexity of the proteome is largely unexplored. On the sequence level, the main source of diversity is given by the mechanism of *alternative splicing*. Alternative splicing occurs in eukaryotes, where the coding parts of a gene are distributed over multiple exons. During transcription, the exons can be spliced together in different ways, potentially resulting in alternative protein products [109]. On the transcript level, the phenomenon of splicing is ubiquitous and widely detected in RNA-sequencing data. On the protein level, however, there is much less evidence for splicing and the impact of alternative splicing on the proteome is subject to controversial debate [110, 111]. The main reason for this debate are the technological limitations of MS proteomics listed above, which prevent a comprehensive study of splicing. There are, however, also limitations on the computational side: For example, there is a multitude of computational methods to detect the regulation of splicing based on quantitative information (*differential alternative splicing*) in transcriptomics data [112]. For proteomics data, however, there are currently no comparable methods available to quantitatively detect differential alternative splicing. In the following chapter, we want to fill this gap by introducing such a method. We extend the idea of creating background contexts from replicate measurements in MS-EmpiRe to so called 'double differential' setups (i.e. changes of changes). This allow us to asses the changes between different gene regions and therefore the quantitative detection

of differential alternative splicing. We benchmark our method on several datasets and show that we can obtain relevant splice events in a proteomic colon cancer cohort. Studying the regulation of splicing can facilitate insights on biological responses based on splicing and therefore also make a contribution to the underlying question of proteome complexity.

Publication

The contents of this chapter have been sent out for peer review. I also presented this work at the 2020 reboot of the annual conference of the *American Society of Mass Spectrometry* (ASMS), one of the largest conferences in the field.

Author contributions

I headed the project, implemented the package, performed bioinformatics analyses and wrote the manuscript. I jointly discussed and designed the method together with Gergely Csaba and Markus Gruber. Gergely Csaba provided helper classes. Ralf Zimmer and Gergely Csaba gave comments on the manuscript. Ralf Zimmer supervised method development, bioinformatics analyses and the writing of the manuscript.

4.1 Abstract

The regulation of alternative splicing is a complex process and can result in alternative isoforms or the same isoform(s) at different abundance in different conditions. Differential alternative splicing between conditions, especially on the protein level, helps to estimate the impact of splicing. In mass spectrometry-based proteomics data, distinguishing isoform-specific peptides are rarely measured and detection and quantification of isoforms is difficult. We introduce MS-EmpiReS, the first quantification-based computational approach for differential alternative splicing detection in proteomics data. Our approach detects both, isoforms-specific peptides and systematic abundance fold changes between different regions of a gene. We apply MS-EmpiReS to differential proteomics measurements between normal and diseased tissues from a larger clinical colon cancer cohort. MS-EmpiReS could exploit a 100-fold increase in the number of testable peptides and, thereby, detected a large number of cancer-relevant alternative splicing candidates, indicating a potential use of proteomic splice signatures in disease contexts.

4.2 Introduction

The exon-based gene structure of eukaryotic organisms enables the production of multiple proteins from a single gene via alternative splicing (AS). The impact of AS on the proteome is subject to controversial debate [111, 110, 113, 114, 85, 115]. In particular, it often remains unclear whether and to what extent the alternative transcript observed in very deep sequencing data will actually result in relevant amounts of alternative protein products. Alternative splicing on the protein level is usually detected in mass spectrometry (MS) data by utilizing the sequence information of identified peptides. For example, junction peptides can be identified which could span over a spliced-out exon. If an additional peptide within the exon is identified, this is a clear indication of splicing. Also intron retentions and alternative start sites could be identified via the respective peptides. Such *sequence-based* approaches [113, 85, 116] are the most basic form of AS detection, usually aiming at the assessment of the general prevalence of splicing, for example in the human proteome. Detection of AS could mean several things: (i) Detecting different isoforms in different samples to validate that these isoforms are actually translated to proteins, (ii) finding different isoforms between conditions establishing major differences, or (iii) finding different isoforms in one condition maybe with a shift in the relative abundance of these isoforms between conditions. Consequently, the latter case is the most difficult to identify, but would include the other cases as well. In order to cover all three cases and to obtain detailed insights into the regulation of splicing, it is necessary not only to assess, whether a protein is expressed or not but also to detect quantitative differences in the expression of the isoforms. The ambiguity of peptides mapping to a multitude of protein isoforms is well known and often addressed by applying principles of parsimony, but also more quantitative approaches have been introduced [117, 118].

The basic question we address is, whether splicing patterns change between conditions. In Figure 4.1A, we display different cases of splicing regulation, which can include switching of isoforms between conditions, expression of an additional isoform in one condition as well as

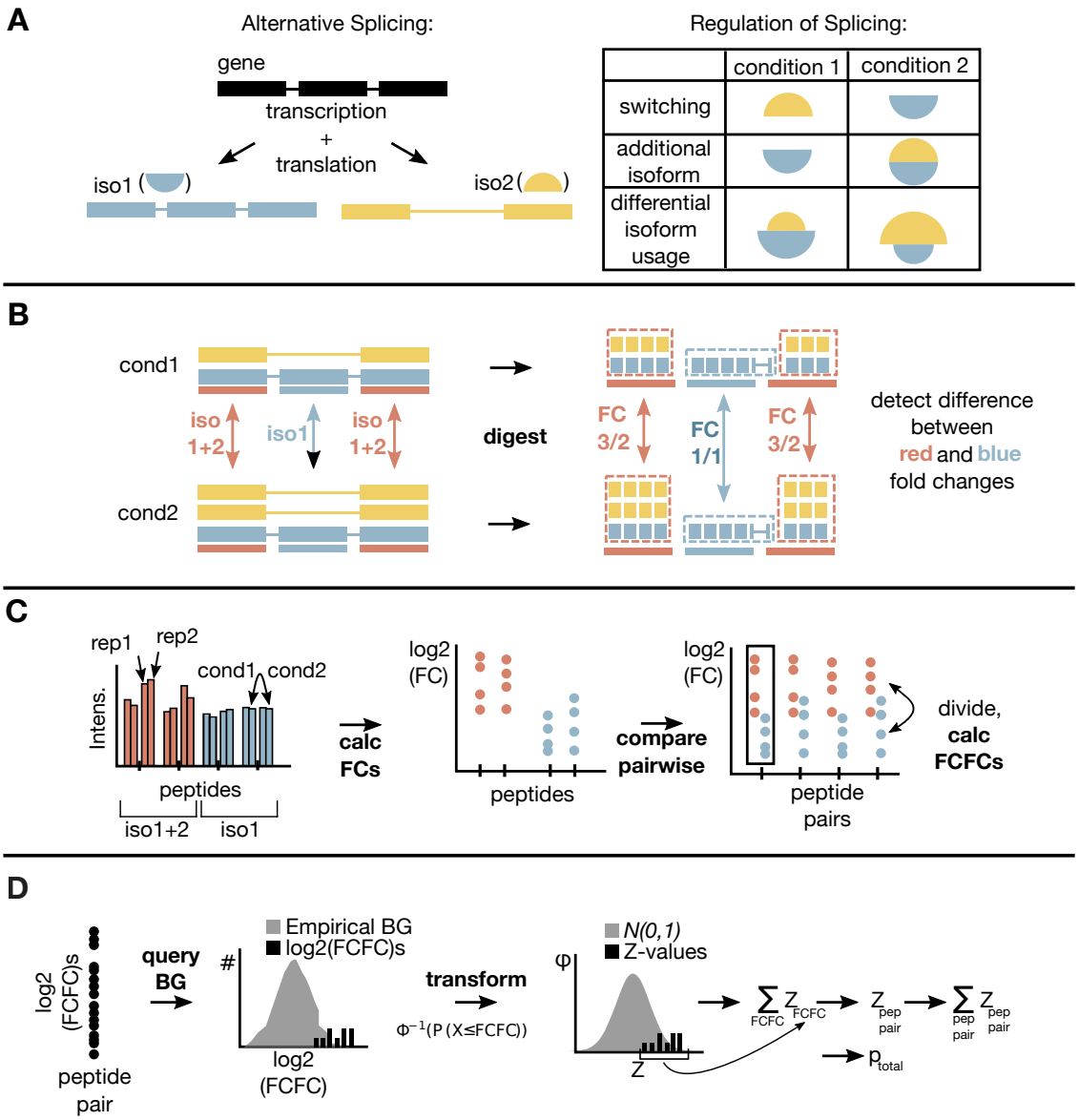


Figure 4.1: MS-EmpiReS workflow. A) Exemplary alternative splicing event (exon skipping) and its effect on the protein products and different forms of splicing regulation. Differential isoform usage, which is only detectable via quantitative differences of abundances, includes the other two cases, which could speculatively be predicted via detection alone. B) Principal idea of quantitative splicing detection for differential isoform usage. The yellow isoform doubles in condition 2. After digestion, the peptides (small squares) either map to only iso1 (marked blue) or both iso1+2 (marked red). The fold changes of the red peptides (3/2) are different to the fold changes of the blue peptides (1). This difference can be detected. The fold change for the yellow isoform, which has no exclusive region and, thus, no unique peptides in our example, is 2. C) Peptide level comparison between red and blue regions, with an example of two red and two blue peptides with two replicate samples each. Low noise peptide fold changes between conditions are assessed as a first step. Red and blue peptides are then compared in a pairwise manner and "fold changes of fold changes" (FCFCs) are assessed. D) The FCFCs are used to query an empirical error distribution of FCFCs derived from replicate measurements, where no systematic change is expected. The observed FCFCs can be expressed as Z-values (with direction of the change) and combined (summed) for the respective isoforms to obtain estimates for the change of isoform changes between the conditions.

differential abundance changes between two expressed isoforms in both conditions. In the following, we will refer to the latter two events as *Differential Alternative Splicing* (DAS). Due to the relatively low sensitivity of MS proteomics setups, it is in general not possible to distinguish between not expressed and not detected, which would be necessary to identify isoform switching. Additionally, absolute quantification of molecule copy numbers would be required, which is not feasible in a precise manner in MS proteomics [13]. Therefore, we rely on relative quantification, which enables precise fold change estimations (0.1-2 fold errors) via so called "local" peptide fold changes [119, 120]. In this study, we utilize the fact that peptide fold changes can be calculated in a substantial fraction of DAS events. This complements current studies on splicing regulation, which employ different types of sequence-based approaches [85, 121, 122, 123] and often rely on transcriptomic DAS detection [124, 125, 126, 127, 128]. We introduce a new computational method based on our recently published differential quantification method for "Mass Spectrometry analysis using Empirical and Replicate bases statistics" (MS-EmpiRe) [120]. MS-EmpiRe utilizes empirical between-replicate distributions to assign probabilities to individual peptide fold changes. In our new extended algorithm we present a framework to compare peptide fold changes against each other in the context of splicing (MS-EmpiRe**S**). This enables us to score whether peptides mapping to one region of the protein have significantly different fold changes than peptides mapping to another region. In combination with isoform mappings from the Ensembl [129] database we assemble these regional local fold changes to fold change differences between isoforms and thereby identify candidates for all types of DAS. The basic principle is displayed in Figure 4.1B with an example of two expressed isoforms in two conditions (case (iii)). Isoform 1 doubles from condition 1 to condition 2, while isoform 2 does not change between the conditions. After enzymatic digestion, the peptides either map only to isoform 2 (blue), or map to both isoform 1 and isoform 2 (red). It should be noted that in general peptides mapping only to isoform 1 or peptides mapping to an additional isoform also exist. These scenarios can always be reduced to a similar case as displayed here (see methods) and are hence omitted in the Figure for clarity. The fold changes of the red peptides should center around 3/2, because there are two copies in condition 1 and three copies in condition 2. The fold changes of the blue peptides should center around 1, because there is no change in isoform 2 between conditions. With MS-EmpiReS, we statistically evaluate the fold change differences between such groups of peptides, as described below. We additionally complement this quantification-based approach with a sequence based approach to utilize the full information available in the dataset. The MS-EmpiReS approach hence differs from current approaches due to the quantification-based identification. Current evaluations of protein-level DAS are based on detecting sequence-based splicing in MS proteomics measurements. Detected events are subsequently quantified, which leads to a drastic loss in sensitivity.
The bioinformatics pipeline of MS-EmpiReS is visualized in more detail in Figure 4.1C with the example case of two peptides in each group and two replicate measurements for each peptide. The fold changes between conditions are determined for every peptide and all peptide pairs between the two groups are formed. For every peptide pair, four against four peptide fold changes are compared and the fold changes are divided, resulting in 16 "fold changes of fold changes" blue/red (FCFCs), which are log2 transformed. The absolute value of the FCFC indicates how dissimilar the change between conditions is. Positive or negative FCFCs reflect that the blue group changes stronger than the red group or vice-a-versa,

respectively. The FCFCs are compared to an empirical error distribution, describing the FCFCs of non-changing peptide pairs (see methods for more details). From the empirical error distribution and a given FCFC, a normally distributed Z-value can be derived. The Z-values can be combined using a modified Stouffer [130] approach to calculate an overall score, which tests the null hypothesis: no difference in the change of the two peptide groups [120]. We denote the multiple testing corrected score as p_{adj}. Dependencies of the variables have to be taken into account at several points of the calculation (see methods).
Peptides are mapped to protein isoforms based on the Ensembl genome annotation. In order to detect quantitative differences between isoforms, the FCFCs of peptide pairs are used, where the peptides stem from different isoforms. The Ensembl annotation contains the current state of the art of known isoforms comprising all relevant splice events (exon skippings, alternative donor/acceptor sites, intron retentions, etc.). As some genes have a large number of annotated isoforms but not all of them are expressed as proteins in the condition under study, *equivalence classes* are determined which group peptides unique to a specific (set of) isoforms (see methods). Thereby, FCFCs are compared for peptide pairs relevant for the isoforms of interest. Again, FCFCs are accumulated over all these peptide pairs in order to estimate the significance of DAS.
MS-EmpiReS is available as a java package under
https://www.bio.ifi.lmu.de/software/msempire_s/index.html and provides DAS detection from standard quantitative proteomics measurements including on-demand visualizations.

4.3 Results

4.3.1 Benchmarking

We benchmarked MS-EmpiReS in several ways, always using negative controls. We use differential proteomics datasets with proteins changing between two conditions, but no splicing of the proteins. For each protein, we randomly distributed the corresponding peptides into two groups, representing two virtual isoforms (i.e. "red" and "blue" peptides) and tested them against each other with MS-EmpiReS. Each group had to contain at least two peptides. In the case that MS-EmpiReS classified a comparison as significant, this indicated a false positive hit. These (realistic) benchmark sets account only for false positives, as this is the most important factor to control. For the first benchmark, we simulated proteins with randomly sampled intensities, representing non-spliced proteins (see methods for details). As there is no systematic shift in the data, the p-values from the tests should be uniformly distributed, if model and implementation are adequate. We confirmed this to be the case, also when one isoform has few peptides/replicates and the other isoform has many peptides/replicates (Figure 4.2A). We then tested several experimental datasets (Figure 4.2B). First, we tested MS-EmpiReS on technical datasets for both label free quantification (LFQ) and Tandem Mass Tag (TMT) data. In the technical datasets, 6 replicate measurements of human cell lysate were split into two groups and tested against each other, to simulate a differential setup. We see that testing thousands of proteins as described above does not give a single significant hit (having $< 1\%$ p_{adj} in MS-EmpiReS, see methods) (Figure 4.2B, left). This indicates that MS-EmpiReS handles the technical variation and biases within a proteomics setup well without giving false positives. For the last and most challenging benchmark, we compared quantitative *E. coli* proteomics datasets (3 conditions for TMT and 27 conditions for LFQ data, see methods for details). *E. coli* are chosen because as prokaryotes they have no splicing mechanism and because they show extensive proteome remodelling between conditions (differential regulation for ≈30% of the proteome on average), thereby introducing both technical and biological biases, thus, a very challenging benchmark. After testing, we indeed observed 0.4% false positive hits. Inspection of these false positive hits showed that some peptides have very systematic and replicate-consistent shifts (Supplemental Figure 1), possible due to post-translational modifications, which can affect the fold change [131]. As with our model, we explicitly test for systematic shifts between groups of peptides, rare combinations of such peptide shifts can lead to very strong significance scores. However, as the rate of such events was low at 0.4% we deemed our model sufficient for confident splicing identification.

4.3.2 Analysis of a clinical dataset

We applied MS-EmpiReS to a clinical proteomics cohort of about 100 colon cancer patients, measured by the Clinical Proteome tumour Analysis Consortium (CPTAC) with TMTs [132]. In the study, two samples have been extracted from each patient, one with cancerous tissue and one with healthy adjacent tissue (see methods for details). We first performed sequence-based splicing detection with MS-EmpiReS. For this, we searched for peptide sequences that have conflicting genomic coordinates and hence cannot exist on the same protein due to

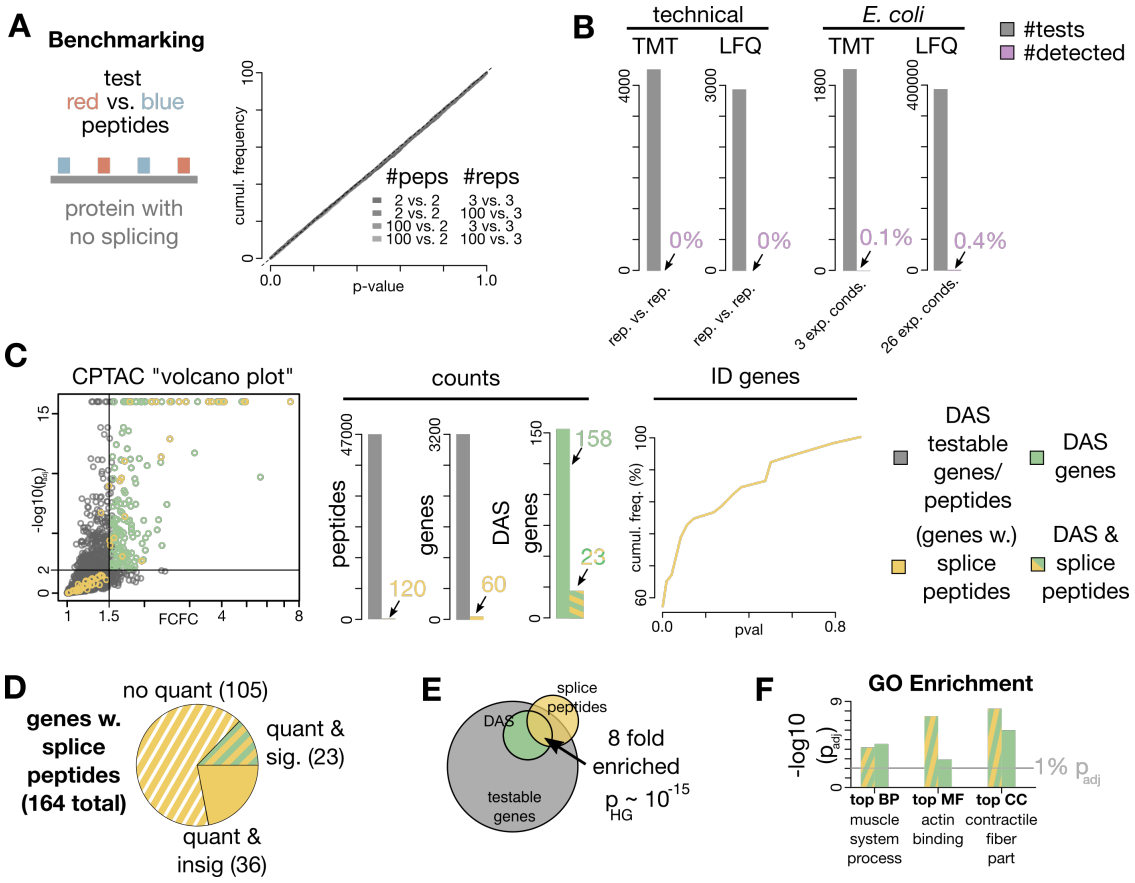


Figure 4.2: Benchmarking and application to a clinical proteomics dataset. A) For benchmarking, simulated peptides of non-spliced proteins are randomly assigned into two groups and tested against each other. Simulating peptides by drawing random intensities gives perfectly uniform p-values as required, even for drastically different peptide and replicate numbers. B) Negative controls on several experimental (LFQ and TMT) datasets using technical replicates and the non-splicing organism *E. coli* (every hit is hence a false positive). MS-EmpiReS is applied to random subsets of peptides and significant hits (adjusted p-value (p_{adj}) < 0.01) are displayed in violet. A very small fraction of significant changes of peptides between conditions is detected by MS-EmpiReS, possibly due to chemical modifications or systematic biases. C) Application of MS-EmpiReS on the CPTAC colon cancer data set ($\approx$ 100 patients). The results of isoform changes can be displayed as a volcano plot with absolute, non logged, FCFCs. Genes containing isoform pairs with p_{adj}<1% and FCFC>1.5 are classified as DAS (green). The yellow color indicates genes for which distinct junction peptides for two different isoforms exist and both isoforms were sufficiently quantified. Still, a large number of these yellow genes is in the insignificant area of the volcano plot. This means that for some genes different isoforms exist which show no clear change between cancerous and healthy tissue. Counts corresponding to the volcano plot are displayed on the right. The number of testable proteins and peptides (grey) is increased by 1-2 orders of magnitude, respectively. The number of DAS genes identified with MS-EmpiReS is more than six fold the number of DAS genes detected via junction peptides alone. On the right, the cumulative distribution of p-values of genes with junction peptides is displayed. We see that more than 40% of genes with junction peptides have insignificant p-values >0.01. (Caption continued on next page.)

Figure 4.2: (continued) D) Overview over sequence-based splicing. We see that the majority of genes with splicing evidence is actually not accessible for quantitative assessment (differential isoform usage). Only 23 of 164 genes with junction peptides can be classified as DAS. This underlines the strong difference between AS and DAS, with DAS having condition-specific regulation. E) The number of genes with detected junction peptides is strongly enriched within the DAS genes quantitatively classified by MS-EmpiReS. D) Top scoring GO "biological process" (BP), "molecular function" (MF) and "cellular component" (CC) are the same for both approaches.

Gene	Name	Splicegene Lau et al.	Literat. Cancer Splicing
ACTN1	actinin alpha 1	**y**	**y**
ACTN4	actinin alpha 4	**y**	**y**
CALD1	caldesmon 1	**y**	**y**
CAPZB	capping actin protein of muscle Z-line subunit beta	**n**	**y**
CFL1	cofilin 1	**n**	**y**
CHID1	chitinase domain containing 1	**y**	**n**
COL6A3	collagen type VI alpha 3 chain	**y**	**y**
EPB41L2	erythrocyte membrane protein band 4.1 like 2	**y**	**y**
H2AFY	macroH2A.1 histone	**y**	**y**
LRRFIP1	LRR binding FLII interacting protein 1	**y**	**y**
MAP3K20	mitogen-activated protein kinase kinase kinase 20	**n**	**y**
PDLIM5	PDZ and LIM domain 5	**y**	**y**
PDLIM7	PDZ and LIM domain 7	**y**	**y**
PKM	pyruvate kinase M1/2	**y**	**y**
RPS7	ribosomal protein S7	**n**	**y**
SPTAN1	spectrin alpha, non-erythrocytic 1	**y**	**y**
TNC	tenascin C	**y**	**y**
TPM1	tropomyosin 1	**y**	**y**
TPM2	tropomyosin 2	**y**	**y**
TPM4	tropomyosin 4	**n**	**y**

Table 4.1: 20 of the top ranked DAS genes in the CPTAC data set. The "Splicegene Lau et al." column indicates whether the gene is listed as alternatively spliced in the recently published database on protein splicing by Lau et al. The "Literature Cancer Splicing" column indicates whether there are explicit mentions of the gene as being alternatively spliced in the context of cancer. Bold genes are shown in detail in Figure 4.3.

splicing. At least one of the peptides had to be a junction peptide spanning an exon junction (see methods for details). We identified 164 genes with such splice peptides. We then filtered out peptides with less than 5 measured replicates (patients) in any of the conditions (cancer and normal), reducing the number of testable peptides from ≈166.000 to ≈138.000. We subsequently applied the new quantification-based approach of MS-EmpiReS. Results of MS-EmpiReS on the dataset are displayed as a volcano plot with FCFCs on the x-axis (Figure 4.2C). Each protein with at least two peptides in at least two equivalence classes is accessible to DAS testing, which resulted in around 3200 testable genes in the CPTAC dataset. Compared to the sequence-based approach, around 50 times the number of genes are available for testing. This results in a six-fold increase in the number of significant genes and enables a first quantitative proteome-wide screening for DAS. With MS-EmpiReS we aim to distinguish regulated from non-regulated splice events, which is not possible via the purely sequence-based approach. To investigate the differences between these two approaches, we examine the genes that are detected as spliced using the sequence-based approach. We see that around 40% of these genes have no significant p-value even before multiple testing cor-

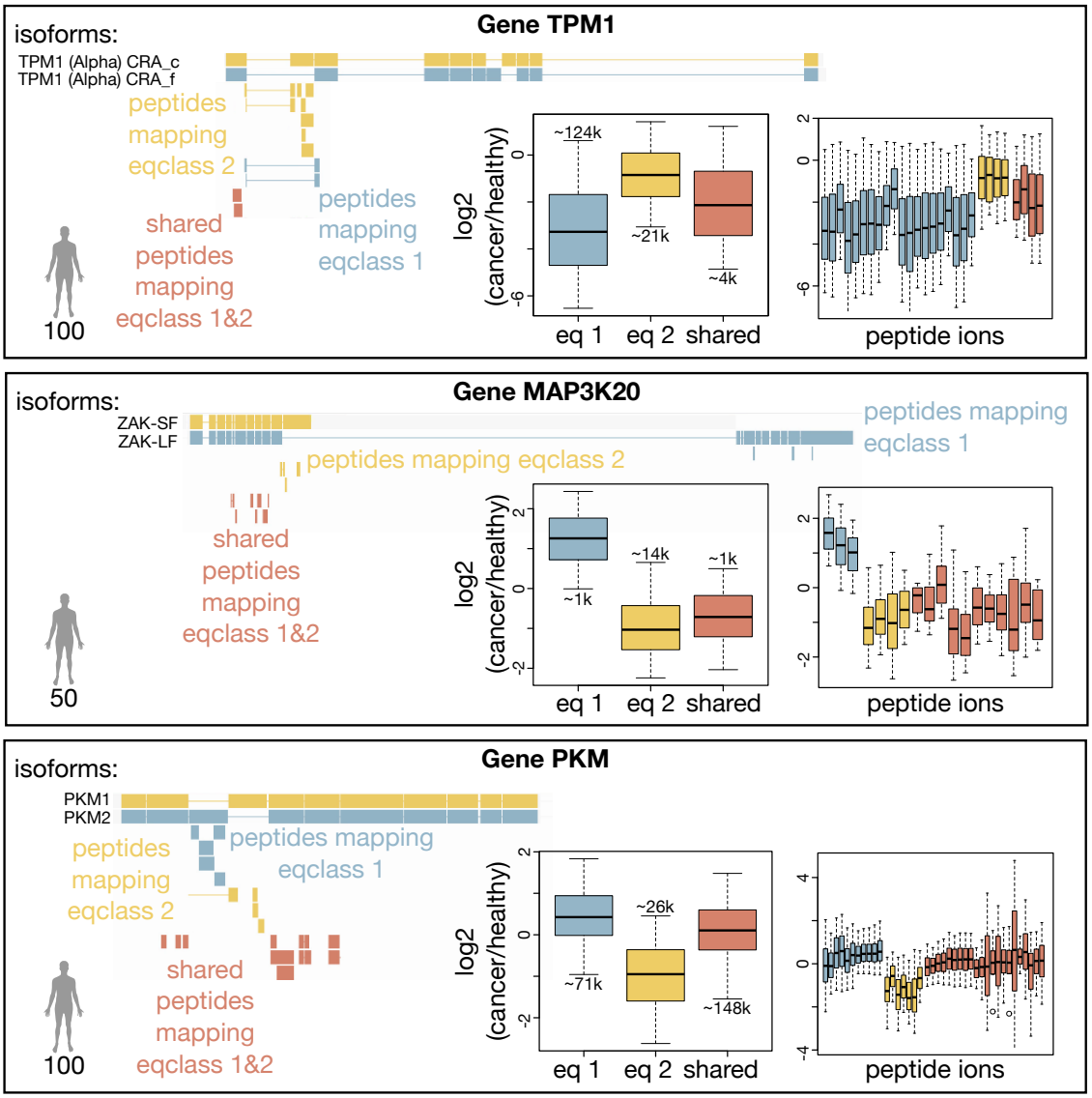


Figure 4.3: Visualization of DAS events for three top scoring genes with important regulatory functions. Two transcript representatives of the equivalence classes (isoforms) are displayed in yellow and blue with quantified peptides aligned below (see Figure 1). The box plots summarize peptide fold changes between cancer and normal for each equivalence class (blue and yellow isoforms), and shared (red). The boxplots on the right show the respective fold changes per peptide. We observe both, clear and significant differences between the two isoforms (ECs) and a plausible (mixture) change for the isoforms in between. The number of patients quantified in both equivalence classes is indicated in the bottom left corner. The examples underline that MS-EmpiReS is able to detect and quantify splice events for functionally relevant proteins, enabling a direct description and interpretation of quantitative splicing changes in cancer vs. normal tissue of patients.

rection (Figure 4.2C) and visualization reveals many cases which appear to be non-regulated (Supplemental Figure 2). This underlines that the approach of detecting isoform-specific peptides without additional quantitative assessment is not sufficient to obtain comprehensive information about the regulation of splicing. We see that a major hurdle for quantitative evaluation is consistent quantification: less than half of the genes with splice peptides are not properly quantified (Figure 4.2D), either because there are not enough replicate measurements or because there are not enough other peptide ions to ensure proper quantification (see Supplemental Table 1). This reveals a large potential for increases in sensitivity, for example with more targeted data acquisition approaches [133, 134]. To check, whether our quantitative approach (DAS genes marked green in Figure 4.2C) is consistent with the junction peptide approach with additional quantification (genes marked yellow/green in Figure 4.2C), we performed two further checks: We first assessed the enrichment of genes with junction peptides within the DAS genes (Figure 4.2E), which was very strong. This indicates that our quantification-based approach is a suitable way to detect actual splice events. We also performed GO enrichment of the DAS genes and the DAS genes with additional junction peptides. The top scoring results for "GO biological process", "GO molecular function" and "GO cellular component" are the same for both approaches (Figure 4.2F).
Twenty of the top ranked DAS genes are listed in Table 4.1 (see Supplemental Table 3 for the full list). We performed two additional checks on this list of genes. The first check was to look up each gene in a recently published database on splicing in the human proteome, which was generated using large scale profiling of MS proteomics data of human tissues with a junction peptide-based approach [85]. For each gene, we indicated whether it was detected as alternatively spliced in the database. For the second check, we searched the literature for explicit mentions of the gene as being alternatively spliced in a cancer context and indicated if we found such mentions (see Supplemental Table 4 for references). We could validate all genes in at least one of these checks. Detailed visualizations of DAS events are given in Figure 4.3 (see Supplemental MS-EmpiReS output files for all visualizations). The first example is Tropomyosin 1, a gene that regulates muscle contraction in association with the Troponin complex. It is known to be a tumour suppressor gene with splice events impacting colony formation and regulatory activity [135]. We see downregulation of both equivalence classes, with strong downregulation of the equivalence class including the CRA_a isoform and mild downregulation of the CRA_m isoform equivalence class. Even though the regulation goes into the same direction, MS-EmpiReS clearly resolves the splice event ($p_{adj} < 10^{-15}$), potentially indicating a higher relevance of CRA_a to the suppression of colon cancer. The second example is the gene MAP3K20, also known as ZAK kinase, which is a MAPKKK family signal transduction molecule and activates cancer-related signaling pathways such as NF-κB, Wnt/β-catenin, and AP1. The two equivalence classes map to the ZAK long form (ZAK-LF) and the ZAK short form (ZAK-SF) which differ strongly from each other. The ZAK-LF has been shown to induce tumour growth in immunodeficient mice [136]. In accordance with this finding, we see a switching event in the colon cancer patients, with the tumour associated isoform being upregulated and the ZAK-SF being downregulated, indicating a splicing induced signalling switch on the protein level. As we have peptides mapping to equivalence class 1, equivalence class 2 and shared peptides between both classes, we can roughly estimate the ratios between the isoforms as shown in the Supplemental Text. This estimation indicates that the ZAK-SF is almost two orders of magnitude more abundant

than the cancer up-regulated ZAK-LF, potentially indicating a higher impact of the ZAK-LF on the phenotype

The third example is the pyruvate kinase M 1/2 gene, which mediates the last step of glycolysis, namely the dephosphorylation of phosphoenolpyruvate to pyruvate. It is hence an essential metabolic gene and has been widely studied in the context of cancer [137]. For example, it has been shown that switching of the PKM2 isoform to PKM1 reverses the Warburg effect in cancer cells [138]. In concordance with this finding, we see a slight upregulation of the PKM2 associated peptides but also a stronger downregulation of the PKM1 associated peptides in the patient data.

4.4 Discussion & Conclusion

In this study we have introduced a novel bioinformatics framework to detect Differential Alternative Splicing (DAS) in MS proteomics data based on quantitative information. Our framework is able to detect splicing induced differences in the relative abundance of two isoforms between conditions via FCFCs. We thereby extend the current sequence-based approach to splicing detection in proteomics data. To our knowledge, a quantitative approach for this purpose has until now not been available for proteomics data. This approach covers all types of splicing and adds additional evidence by considering many more of the measured peptides. As proteins represent the functional players in the cell, we believe that studying regulation of splicing on the protein level is crucial for obtaining further insights into the biological implications of alternative splicing, making MS-EmpiReS a valuable computational tool for the splicing community. To maximize flexibility, we specifically designed MS-EmpiReS to be a proteomics and not a proteogenomics tool (i.e. dependent on additional transcriptomics data). In principle, MS-EmpiReS can analyze any quantitative proteomics experiment where two or more conditions are compared and at least two replicates are measured in each condition. The only input needed is quantified peptides and a condition mapping. By screening the complete Ensembl annotation, we get a comprehensive picture of possible splice events. We currently provide analysis possibilities for human and mouse data, but in principle, other organisms can be easily added. MS-EmpiReS only detects a splice event if there are at least two non-overlapping peptides quantified for each isoform, the peptides on each isoforms have similar fold changes and therefore a statistically significant shift between the isoforms has been detected. This reduces the chance for false positives, as we have shown on the conservative *E. coli* benchmarking dataset. Nevertheless, false positives are possible and we encourage manual inspection of the splice events via the visualizations provided by MS-EmpiReS. We applied our method to a larger CPTAC dataset containing data from ≈100 patients. This clinical proteomics dataset carries a high level of noise and also of biological variation. Nevertheless, we were able to recover clear proteomic splice events of important regulator genes, which we validate by the sequence-based approach, independent databases and literature knowledge. MS-EmpiReS identifies and displays the respective evidence from peptide to isoform level and, thereby, supports the interpretation of cancer-related splicing events and isoform abundances. In summary, MS-EmpiReS thus is a useful and easy to use addition to standard computational proteomics pipelines.MS-EmpiRe is a conservatively benchmarked, widely applicable tool that can reliably identify all types of Differential Alternative Splicing also in disease relevant candidates in challenging personalized medicine contexts.

4.5 Methods

4.5.1 Simulating non spliced proteins

In the first step of the simulation, differential peptides, i.e. peptides quantified in two conditions were simulated. For each simulated peptide, r_1 and r_2 intensity values were drawn from a log normal distribution, with r_1 and r_2 being the numbers of replicate measurements in each condition. A log normal distribution was deemed adequate as it is generally used as an approximation to describe protein/peptide abundances [13]. Two groups of peptides were tested against each other for each simulated protein, meaning that for each simulated protein, $g_1 + g_2$ peptides were drawn, with g_1 and g_2 the numbers of peptides in each group. 3,000 proteins were simulated for each test, with the parameters r_1, r_2, g_1, g_2 specified in Figure 4.2A.

4.5.2 Benchmarking on the technical datasets

The technical dataset was downloaded as specified in the data availability section. The data was acquired in the context of a study on differential quantification [139], where human cell lysate with differing amounts of yeast spike in were measured with LFQ and TMT-MS3. This dataset had been used in the MS-EmpiRe study [120], where it had been processed as follows: The data was searched and quantified with the MaxQuant [14] software v. 1.6.0.16. Standard settings and additional LFQ or TMT quantification were set. The database used was combined from yeast (7,904 entries) and human (20,317 entries), which were individually downloaded from Uniprot [140] (April 2018, reviewed) was used. For the DAS benchmark, the human proteins were selected and for each quantification method, 6 replicate runs were compared as 3 vs. 3 replicates.

4.5.3 Benchmarking on the *E. coli* datasets

The *E. coli* datasets were downloaded from their respective PRIDE repositories (see data availability section). For the LFQ dataset [141], the .raw files were downloaded and searched with MaxQuant v. 1.5.7.4 against the reviewed Uniprot *E. coli* K-12 database (03/2019), using standard settings with additional "Label Free Quantification" and "Match between runs" set. For the TMT dataset (acquired via an SPS-MS3 workflow), MaxQuant search files were directly downloaded. As we collected the data from different studies, the details of the preprocessings (for example the MaxQuant versions) differed slightly. Our model is however not dependent on such preprocessings and should not be affected by this. Peptide intensities were extracted from the "peptides.txt" files for both TMT and LFQ data. Only peptides with unique mapping to a protein were considered. The conditions were compared in a pairwise manner resulting in 352 condition pairs for the LFQ data (see Supplemental Table 3) and 3 condition pairs for the TMT data. For each condition pair, we iterated through all proteins. For each protein, peptides were randomly distributed in two groups of equal size. The groups were then tested against each other for DAS, resulting in one p-value for each protein in the condition pair. Multiple testing correction was carried out using the Benjamini-Hochberg [142] procedure on the p-values for each condition pair and proteins

were classified as significant with an p_{adj}<1% and an FCFC>1.5. Additional consistency constraints were applied as described below.

4.5.4 Preprocessing the CPTAC data set

MSGF+ [143] search files and MASIC [144] selected ion chromatograms (SICs) were downloaded from the CPTAC data portal (see data availability section). Peptides were filtered to a MSGF+ q- value <0.005, a MASIC InterferenceScore >0.9 and a PeakSignalToNoiseRatio>35. In the case of multiple identifications of the same peptide, the identification with the strongest signal was chosen. Fractions of the same TMT multiplex were merged and two normalization steps were performed. In a first step, TMT channels of the same multiplex were normalized to correct for differing sample amounts in the channels, using the MS-EmpiRe within-replicate normalization. In a second step, the TMT-10 131 channels with a technical spike-in were used for peptide specific normalization between TMT multiplexes. For each peptide, fold changes between the reference channels were obtained and one normalization factor per multiplex was estimated using Levenberg-Marquardt optimization. The dataset was acquired with MS2 level quantification, thereby possibly giving rise to the phenomenon of ratio compression [145] (i.e. underestimation of absolute fold change strength) due to co-fragmenting precursor ions. However, several factors mitigated this problem: From the data acquisition side, small isolation windows (0.7 Da) were used, which reduced the probability of co-fragmentation. Additionally, the data was highly fractionated (96 fractions of which 12 were chosen per multiplex) and the ≈200 patient samples were distributed on a total of 22 multiplexes. This should reduce the chance that the same peptide measured in multiple multiplexes has the same type of co-isolation. Rather, co-isolation would result in additional noise between replicates, which can be accurately modeled. On the computational side, we used conservative interference score cutoffs and most importantly, our model is able to handle inconsistent fold changes, as shown in the *E. coli* benchmark. As displayed in Supplemental Figure 3, our normalization results in low noise levels and separation between healthy and disease samples already on the peptide intensity level, indicating good quantification.

4.5.5 Data normalization

Data were normalized similar to MS-EmpiRe [120], which uses the concept of centralization [146]. Briefly, samples were shifted by a constant factor to cancel out systematic biases, e.g. due to differing sample amounts. The factors were derived from peptide fold change distributions using the median for replicate normalization and the mode for normalization between conditions.

4.5.6 Mapping peptides to isoforms

Peptides were mapped to genes and protein isoforms using the Ensembl homo sapiens GrCh37.75 genome annotation. Peptides mapping to more than one gene were eliminated from the analysis. In order to detect sequence-based splicing, conflicting peptide pairs were searched, meaning pairs of peptides that cannot exist on the same isoform. One peptide was required to span an exon junction and the other peptide was required to be conflicting

with this junction, meaning that it starts or ends inside the junction. For the quantitative approach of MS-EmpiReS, an *equivalence class* was obtained for each peptide. Similar to the BANDITS [147] package for transcriptomic DAS detection, we define an equivalence class as the set of all isoforms a peptide maps on. This means that peptides mapping to the same equivalence class should have equal abundances and be independent of DAS (under the assumption that the annotation is comprehensive). Peptides mapping to different equivalence classes could potentially stem from different isoforms or different mixtures of isoforms and can show abundance changes after DAS. Grouping peptides by equivalence classes hence enables very "clean" testing for DAS. In the case that conflicting splice peptides were detected for a gene, we grouped the remaining peptides differently and distributed all remaining peptides of the gene around the splice peptides. For this, we iterated through all other peptides of the gene and tested for each peptide if its equivalence class overlaps (i.e. has shared isoforms) with one splice peptide and does not overlap with the other splice peptide. If this criterion was fulfilled, we grouped it to the respective splice peptide. An equivalence class was included in the further analysis if there were at least two distinct peptide sequences quantified, or if a conflicting splice peptide was quantified in another equivalence class. To reduce the influence of technical noise in the latter case, we also required at least two peptide ions to be quantified (not necessarily with differing sequences). In total, of ≈138000 quantified peptides, ≈64000 mapped to different equivalence classes, ≈47000 of which passed the filtering criteria. For gene-level testing, as displayed in Figure 4.2, the most significant pair of equivalence classes was chosen for each gene and multiple testing correction was carried out subsequently via the Benjamini-Hochberg [142] procedure, the corrected score is denoted as p_{adj}.

4.5.7 Calculation of FCFCs

After isoform mapping, the equivalence classes were tested pairwise against each other. For this, all peptide pairs between two equivalence classes were obtained. For a peptide pair p_1 and p_2 the FCFCs were calculated. To define the FCFC, we first define the peptide intensity $I(c, r, p)$, which is dependent on condition c, replicate r and peptide p. A peptide fold change is subsequently defined as

$$FC(p, c_1, r_1, c_2, r_2) = log2((I(p, c_2, r_2)/I(p, c_1, r_1)) \tag{4.1}$$

with conditions c_1,c_2 and respective replicates r_1,r_2. The FCFC is then defined as

$$FCFC(p_1, c_1, r_1, p_2, c_2, r_2) = FC(p_2, c_1, r_1, c_2, r_2) - FC(p_1, c_1, r_1, c_2, r_2) \tag{4.2}$$

with peptides p_1 and p_2. We see that the only factor changing between first and second term is the peptide. In this definition, peptides p_1 and p_2 are only compared between identical replicates.

4.5.8 Generation of empirical FCFC error distributions

To put FCFCs into a statistical context, we generated empirical FCFC error distributions from replicate measurements. As a first step, we generated empirical FC error distributions,

analogous to MS-EmpiRe [120]. For this, log2 FCs between peptides of equal sequence and charge were obtained between replicate measurements, generating an empirical error distribution of FCs. As replicate measurements carry the sum of biological and technical variation, this empirical FC error distribution should exactly reflect this variation and hence be a good estimate of the noise underlying the experiment. Analogous to MS-EmpiRe, we also separated the empirical FC error distribution into sub distributions depending on the intensities of the peptides measured. As a general trend, peptides with lower intensity are subject to higher variation than peptides with higher intensity and it thus makes sense to have multiple empirical FC error distributions, covering different intensity ranges. To increase runtime and memory efficiency, we binned the empirical FC error distribution into log2 FC intervals of 0.01. To generate the empirical FCFC error distributions, we took two empirical FC error distributions and created the difference distribution, as described in [148]. Technically this is simply achieved by comparing all possible pairs of bins. For each pair of bins, we calculated the log2 FCFC by subtracting the log2 FCs of the pair and obtained the corresponding frequency by multiplying the frequencies of the pair. Empirical FCFC error distributions were generated for all necessary pairs of intensity ranges.

4.5.9 Combination of FCFCs

A pair of equivalence classes with n and m peptides generates $n \cdot m$ peptide pairs and each peptide pair generates a maximum of $r_1 \cdot r_2$ FCFCs, with r_1 and r_2 being the respective number of replicates in each condition. Our aim is to obtain the overall null probability for DAS of the equivalence class pair from this (possibly very large) set of FCFCs. Analogous to the MS-EmpiRe paper [120], we transform the FCFCs into normally distributed random variables using a modified Stouffer approach and combine these variables by summation. We see that many of the $FCFC(p_1, c_1, r_1, p_2, c_2, r_2)$ are dependent on each other. For example, the pair $FCFC(p_1, c_1, 1, p_2, c_2, 2)$ and $FCFC(p_1, c_1, 1, p_2, c_2, 3)$ has a shared replicate. When we consider the FCFCs to be random variables, which we combine, these dependencies affect the variance of the combined distribution and have to be taken into account. A main goal is hence to appropriately estimate the variance of the combined random variables, which can be achieved via summation over the full covariance matrix. The basic concepts for this estimation have been introduced in the MS-EmpiRe paper [120] and a generalized package, which has been used in this work, is described in detail by Berchtold et al. [148].

4.5.10 GO enrichment

GO enrichment was based on the Gene Ontology .obo database, using the "is_a" relation. Enrichment was calculated via overrepresentation analysis using a Hypergeometric test. For the the CPTAC data the Ensembl gene mapping was used and the sets indicated in Figure 4.2C were enriched (see Supplemental Table 5).

4.5.11 Data availability

The *E. coli* proteomics data and the technical benchmarking data "was downloaded from http://proteomecentral.proteomexchange.org via the corresponding PRIDE partner reposi-

tories" [75] (LFQ E. coli: PXD000498, TMT E. coli: PXD008339, technical benchmarking: PXD007683). "The colon cancer data used in this publication were generated by the Clinical Proteomic tumour Analysis Consortium" (NCI/NIH) [149]. The "PSM" data and relevant mappings were downloaded via
https://cptac-data-portal.georgetown.edu/cptac/s/S045. The CPTAC input file for MS-EmpiReS used in this analysis is provided as Supplemental Table 6. The MS-EmpiReS outputs are provided as Supplemental File 1.

Chapter 5

Detecting relevant proteins for *E. coli* carbon starvation in MS proteomics data

Motivation

The previous chapters introduced tools and algorithms to be used by data analysts. Proper computational tools are essential, they do not guarantee, however, that proteomics technology can be fully utilized for its designated purpose: facilitating insights into specific biological questions. A major hurdle for this is the knowledge gap that often prevails between researchers designing and performing biological experiments and researchers analyzing biological data. The arguably best approach to overcome this problem is close interdisciplinary collaboration, bridging the gap with regular and detailed exchange of knowledge. Ideally, there is a solid foundation on both sides, facilitating clear communication. Coming from a biophysics background and having personally worked on a specific biological question, namely the quantitative characterization of *Escherichia coli* (*E. coli*) population dynamics, I was in a unique position to carry out such a collaboration. My previous work led to basic insights and a phenomenological description of *E. coli* starvation [150].
The following chapter describes a continuation of this previous work: we utilize the bacterial death rate that we described in [150] as a quantitative phenotype to define regulatory constraints that proteins have to fulfil in order to be important for starvation survival. We then utilize the data stored in different proteomics repositories to define a regulatory context for each protein. We introduce an adapted Stouffer [130] approach that allows us to globally rank proteins coming from very different repositories and conditions. This allowed us to uncover an essential role of the cell envelope for carbon starvation. To our knowledge, this finding is novel and the connection is not known to the experts in the field.
The project was carried out in collaboration with the Systems Biology lab of Prof. Markus Basan at Harvard Medical School. During this collaboration, I visited the Basan lab to analyze data and develop computational approaches in direct exchange with the biological expert, Dr. Severin Schink. Being directly in the lab enabled short turnaround times from experiment to data analysis and vice versa.

Publication

The contents of this chapter have not been published yet. A manuscript that will strongly overlap with the chapter below is in preparation.

Author contributions

The project was jointly initiated by Severin Schink and me. I proposed and designed the bioinformatics and proteomics analysis parts of the study and contributed to the design and execution of the growth-to-death tradeoff experiments, which facilitated the data analysis. Severin Schink headed design and execution of the biological experiments, with contributions from Markus Basan. I implemented the data analysis pipeline and performed bioinformatics analyses. Severin Schink gave input and feedback on the data analyses. Markus Basan supervised biological experiments and evaluations. Ralf Zimmer supervised the computational analyses. The manuscript was jointly written by Severin Schink and me, with me focussing on the computational aspects and Severin Schink focussing on the biological aspects. Ralf Zimmer and Markus Basan gave comments.

5.1 Abstract

Bacteria like *Escherichia coli* can reorganize their physiology to survive long periods of nutrient-limitation. What part of this reorganization causes the improved survival is a difficult question, because the change in physiology includes a global reorganization of the proteome, structure and organization of the cell that overshadows any subtle changes. In this work, we aim to identify survival-relevant reorganization. We find that a fundamental trade-off between fast growth and long survival is set by the proteome allocation of the bacteria. We utilize this trade-off by statistically scoring several orthogonal proteome perturbations measured with mass spectrometry-based proteomics, resulting in a comprehensive ranking of over 2.000 *E. coli* proteins. Our combined ranking allows us to narrow down the set of proteins that correlate with starvation survival. We find a significant fraction of survival genes to be located in the periplasm and outer membrane. We confirm that the envelope of *E. coli* is indeed a weak spot during starvation using antibiotics and genetic perturbations and verify that improving the mechanical stability of the outer membrane leads to better survival. Our results uncover a new protective feature of the cell envelope that goes well beyond the interpretation of the outer membrane and envelope solely being a barrier that prevents abiotic substances to reach the cytoplasm.

5.2 Introduction

Nutrient limitation is a defining part of the life cycle of microorganisms. In the absence of external nutrients, the only energy sources for heterotrophic organisms is either internal storage [151, 152, 153] or nutrients retrieved by recycling dead biomass [154, 155, 156]. This finite energy source can be temporarily used to maintain the cell, but will eventually deplete, exposing the cells to a slow deterioration process. The survival kinetics during starvation, i.e. how many cells will be still alive after a certain time, is determined by the consumption rate of these nutrients, called the maintenance rate [155]. While the concept of a maintenance rate is well-established since mid-last century [157], several fundamental questions remain unclear decades later. What kind of maintenance is the cell doing, what determines how much maintenance a cell needs, and why are cells dying when this maintenance cannot be met? The question of maintenance is particularly puzzling, because sporulating organisms can minimize the number of active processes - without losing their reproductive ability. The maintenance rate of non-sporulating bacteria is similarly not an engraved biophysical constant. Instead, *E. coli*, for example, can change its maintenance rate and death rate depending on environmental conditions during the growth phase [158]. In this work, we are using the adaptation of *E. coli* in different growth environments to identify which changes of the cellular composition cause changes in the survival kinetics. We use six different types of growth environments to show a global trade-off between proteome composition and starvation survival. By statistically scoring changes in abundance of individual proteins, measured using mass spectrometry-based proteomics, we narrow down the set of proteins that correlate with improved survival across all six conditions. From the abundance correlations, we identify several candidate processes, including the cell envelope. Using genetic and systemic perturbations, we show that tampering with the mechanical stability of the cell envelope leads to faster death.

This includes the peptidoglycan layer, the outer membrane and the anchoring proteins that connects both. Strengthening the cell envelope, by increasing its stiffness, on the other hand, leads to longer survival, indicating that *E. coli*'s maintenance is required to prevent pressure induced cell envelope failure.

5.3 Results

5.3.1 A trade-off between growth rate and death rate across six different growth perturbations

To study whether death rate depends on the proteome composition, we culture *E. coli* K-12 in different growth conditions that are known to affect the proteome composition and bring them into identical starvation conditions by washing and resuspending them into carbon free medium, see methods for details. As a reference condition, we use wild-type *E. coli* grown in minimal medium supplemented with glucose, which yields moderately fast growth at a rate of 0.98/h and death rate of 0.57/d, see Fig. B.1. In addition, we study six different perturbations, listed and shown in Fig. 5.1.

As first perturbation, we omit the washing step and let *E. coli* adapt on acetate excreted during fermentive growth [159]. After one day in this stationary phase we wash the culture to remove all nutrients remaining from growth. The result is a slower death rate (white symbol placed at growth rate 0, Fig. 5.1). As a second perturbation, we limit carbon uptake during growth, either by growing the culture on different carbon substrates (blue triangles), by downregulation of a carbon transporter (blue circles) or by growth in a carbon limited chemostat (blue squares). Individual points are different strengths of the respective perturbations. In all three cases of catabolic limitation, *E. coli* grows slower and subsequently dies slower. Thirdly, we downregulated glutamate synthesis (green circles), thereby inducing an anabolic limitation [160, 161]. As a result, *E. coli* grew slower and died slower, similar to catabolic limitation, but with a less pronounced effect. Growing *E. coli* in rich medium, either medium supplemented with glucose and casamino acids (red square) or lysogenic broth (LB, red circle), resulted in faster growth and faster death. These four growth perturbations, stationary phase, catabolic limitation, anabolic limitation and rich medium, follow a common linear trend of death rate increasing with growth rate. To test if the growth-death relation is due to proteome constraints, we perform a second set of perturbations that targets specifically the proteome allocation. The addition of sublethal doses of chloramphenicol (yellow symbols) leads to an increased expression of ribosomal proteins, which reduces the expression of the remaining proteome [162, 160]. This ribosomal limitation leads to slower growth and faster death, see Fig. 1. Expressing an irrelevant protein (LacZ, grey symbols) leads to a downregulation of virtually all other proteins, Fig. B.2, and results in slower growth and faster death, similar to ribosomal limitation. These two 'proteome stress' perturbations follow a trend orthogonal to 'nutrient quality', see Fig. 1, thereby breaking correlation between growth rate and death rate. The fact that the proteome composition has a strong effect on death rates points to the existence of a 'survival sector' within the proteome, a set of proteins that determines the how long *E. coli* can survive during starvation. Next, we will identify this sector and interpret the function of processes and pathways within it.

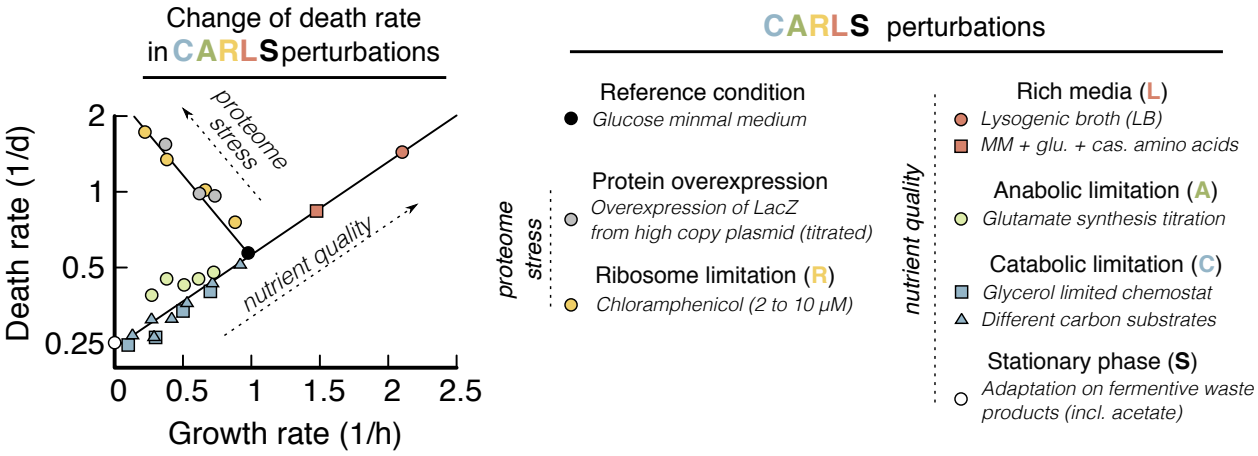


Figure 5.1: Physiological relation between growth and death. Bacteria were grown under different growth perturbations, together abbreviated CARLS (Catabolic limitation (C), Anabolic limitation (A), Ribosome limitation (R), Rich media (L), Stationary Phase (S) and Protein Overexpression (OE)) and transferred to a common starvation condition by washing and resuspension in carbon free minimal medium. The color indicates the type of perturbation and each point represents a different strength of perturbation. Growth and death curves for the individual conditions are displayed in Fig. B.1. Clear orthogonal responses are visible for the perturbations affecting the nutrient quality (C, A, L, S – positive correlation) and proteome stress (R, OE – negative correlation). Data of the glycerol limited chemostat and different carbon sources are taken from [158].

5.3.2 A protective survival sector in the proteome

The survival sector could either be protective or harmful. However, we see an increase in death rate upon downregulation of virtually all proteins. This downregulation is achieved by the over-expression of an irrelevant protein (LacZ), Fig. B.2. The grey points in Fig. 5.1 show the increase in death rate after LacZ overexpression. We conclude that the 'survival sector' must play a protective role. This means that the abundance of proteins within the 'survival sector' should correlate with increased survival across all growth perturbations in Fig. 5.1. We use this as our guiding strategy to identify the 'survival sector', and search the proteins whose abundance correlates with survival across all perturbations of Fig. 5.1.

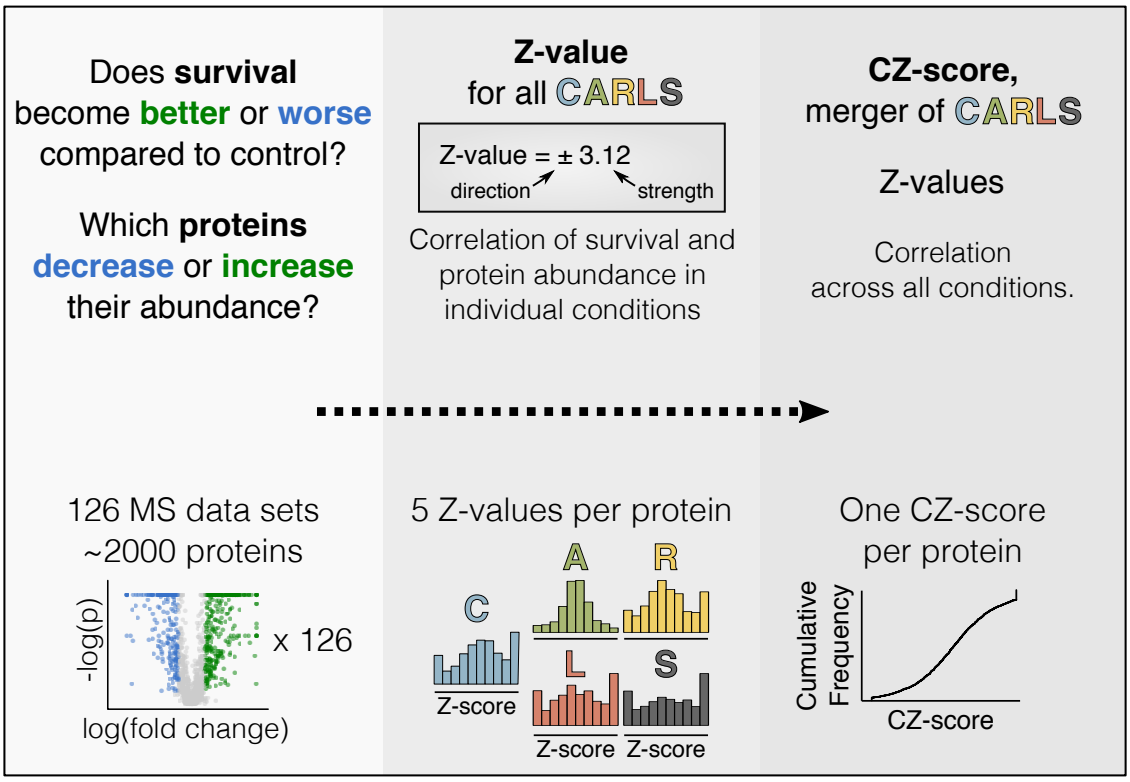


Figure 5.2: Data analysis pipeline. For each CARLS condition we evaluate (1) if survival gets better or worse and (2) which proteins significantly increase or decrease their expression, relative to the glucose minimal medium reference condition (left). Each perturbation includes multiple MS runs corresponding to different perturbation strengths, totaling 126 MS runs across all five CARLS perturbations. The correlation of individual proteins with survival across different MS-runs is merged into a Z-value for each perturbation (center). These five Z-values per proteins are merged into a single 'Combination of Z-values' score (abbr. 'CZ-score') that measures correlation of protein abundance with survival across all experiments (right).

5.3.3 Proteomics analysis pipeline

A data analysis pipeline, outlined in Fig. 5.2, identifies the relevant proteins by correlating changes of protein abundance with changes of death rates. We focused on the perturbations catabolic limitation (C), anabolic limitation (A), ribosome limitation (R) and rich medium (L) and stationary phase (S), which we abbreviate 'CARLS'. We collected a total of 126 LC-MS/MS runs from three different repositories: Schmidt et al. [163], Hui et al. [160] and Houser et al. [164], see methods for details. Analyzing this large and heterogeneous dataset is a considerable challenge, because the three different data sets were measured in different labs with different machines and differing quantification techniques. First, we analyzed each perturbation individually (e.g. different carbon sources for C) and scored each protein depending on how well changes of the survival correlated with changes of protein abundance (left of Fig. 5.2). Next, we merged the scores of all conditions within a single perturbation into a Z-value that quantifies both the strength and the direction of the correlation (middle of Fig. 5.2). Finally, all Z-values were merged into a single score called 'Combination of Z-values' (CZ-score, see methods for details), that allowed us to combine scores across data sets and growth perturbations (right of Fig. 5.2). This pipeline drastically reduces data complexity from 126 MS data sets, down to a single CZ-score for each protein. Individual Z-values and merged CZ-scores for each protein are listed in Table S1. Z-value distributions from different data sets, e.g. from stationary phase measured by Houser et al. and Schmidt et al., or catabolic limitation by Schmidt et al. and Hui et al. show strong correlation, despite both proteomics setups and culture conditions varying, validating that our approach can extract consistent information (Fig. 5.3). While our phenomenological analysis of perturbations in Fig. 1 only showed two orthogonal changes in death rate, pair-wise comparison of the proteome change in the CARLS perturbations reveals that the proteome responses in the five perturbations are substantially different, Fig. 5.4A, with major parts of the proteome correlating and anti-correlating in each comparison. Thus, no two conditions lead to the same proteome remodeling and each condition is providing distinct information to our analysis.

5.3.4 Proteomics analysis shows significant enrichment for stress protection and cell envelope

We used the single protein CZ-scores to test for significant enrichment of gene ontology (GO) processes and cellular compartments using a Kolmogorov-Smirnov test. Figure 5.4B shows all redundancy-reduced GO biological processes and GO cellular components with FDR<0.01. 'Cell envelope', and 'oxidative damage', 'response to stress' and 'catabolic process' show strong enrichment, and a positive correlation with survival. These pathways will now be tested for their causal contribution to the growth/death relationships. The full list of GO processes and cellular components is shown in Table S2.

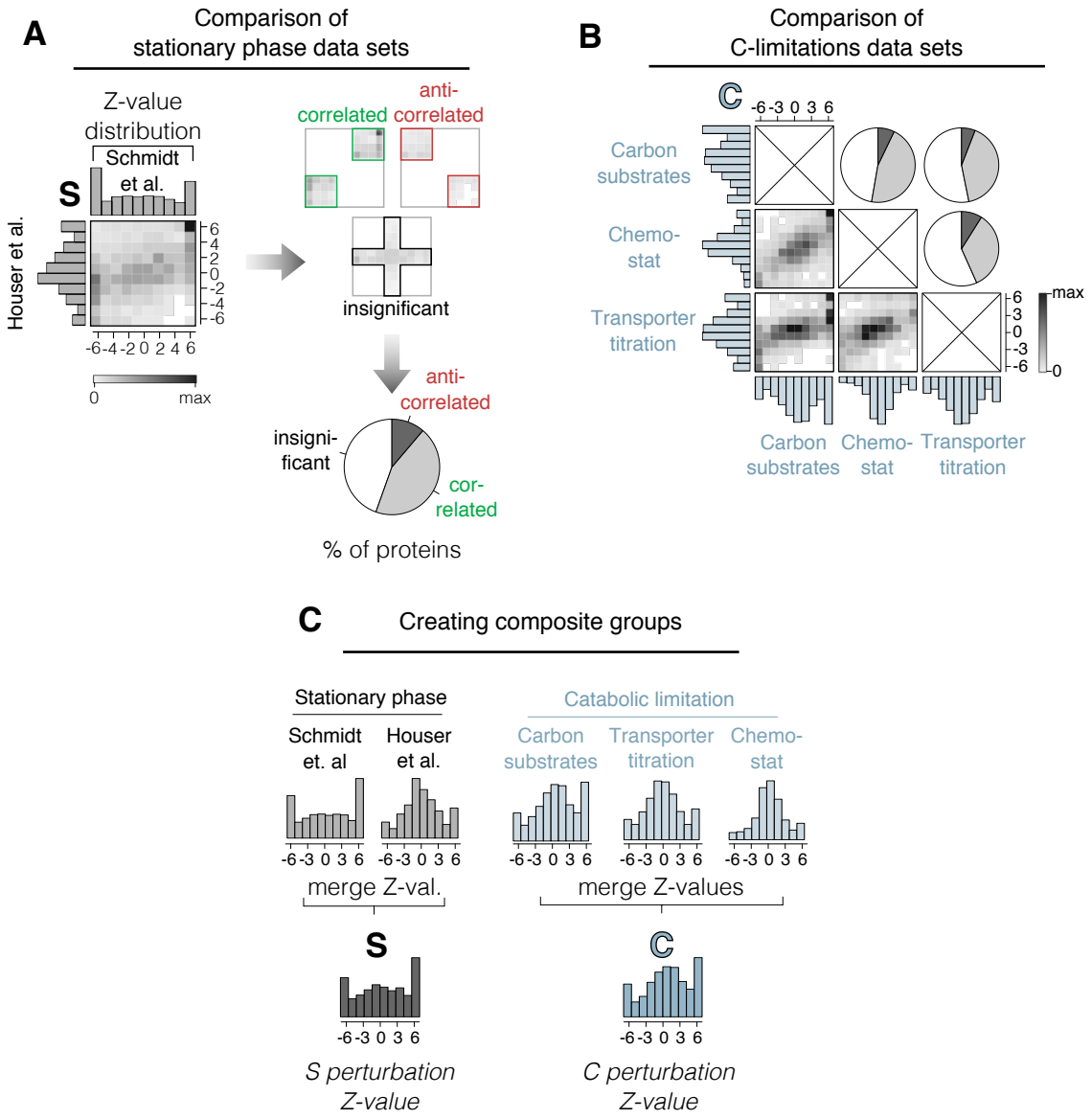
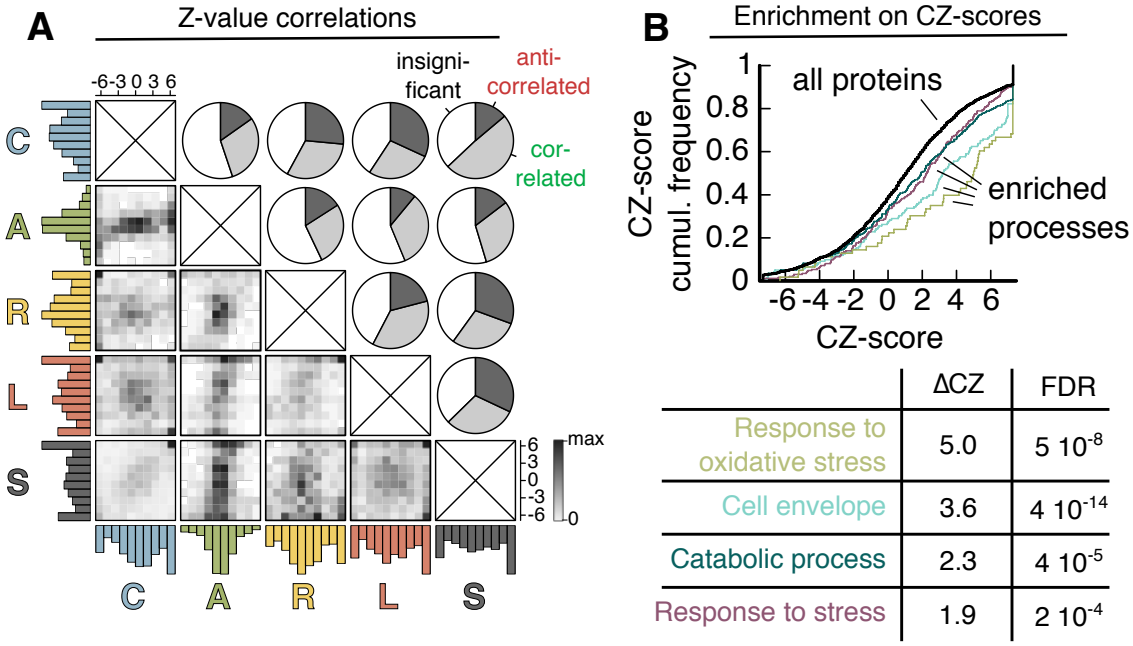


Figure 5.3: Comparison of MS proteomics data from different sources. (A) Comparison of Z-value distributions from two different data sets, by Houser et al and Schmidt et al, both measured after one day in stationary phase after growth on glucose. On left: Z-value distributions of individual experiments are shown on left and top of correlogram. Matrix inside the correlogram depicts frequency distribution of individual proteins that are measured in both data sets and which have the respective Z-values. On right: Proteins scored in top right and bottom left corner are counted 'correlated', proteins in the top left and bottom right are counted as 'uncorrelated'. Cut-off for significance is chosen at $Z = 1.28$, corresponding to a p-value of 0.1. Proteins with at least one Z-value less than 1.28 are counted as uncorrelated. (B) Analysis of correlograms and quantification in pie charts of three data sets of different catabolic limitation analog to panel A. 'Carbon substrates' and 'chemostat' taken from Schmidt et al., 'Transporter titration' taken from Hui et al. Different types of catabolic limitation show high correlation between data sets. (C) Data sets from panel A and B, respectively, are merged with the CZ procedure to form a single Z-value distribution for each growth perturbation.



	ΔCZ	FDR
Response to oxidative stress	5.0	$5\ 10^{-8}$
Cell envelope	3.6	$4\ 10^{-14}$
Catabolic process	2.3	$4\ 10^{-5}$
Response to stress	1.9	$2\ 10^{-4}$

Figure 5.4: Correlations of Z-values across CARLS conditions. Density plots of pairs of Z-values (bottom left) and quantification of correlation (top right) show that no two growth conditions lead to the same proteome remodeling, despite only two trends observed in Fig. 1. In the quantification charts, proteins scoring higher than $|Z| = 1.28$ (p-value 0.1) in both directions are denoted correlated or anticorrelated, depending on the signs of both values. Proteins scoring less are denoted 'insignificant'. The highest correlation in proteome remodeling are seen for C-S (carbon limitation and stationary phase), the highest anti-correlation for C-L (carbon limitation and rich medium) and R-S (ribosome limitation and stationary phase). Each perturbation adds unique information to the data analysis pipeline. (B) Cumulative CZ-score frequency of complete set of CZ-scores (black – 'all proteins') selected enriched processes (colors). Median change of CZ-scores (ΔCZ) and false discovery rate (FDR) of selected processes shown in corresponding colors below plot. See Table S2 for complete list of processes and cellular compartments.

5.3.5 Stress response, oxidative damage and catabolism are not limiting survival

To test whether the 'stress response' causally determines the survival kinetics, we pre-stressed bacteria with 50 mM NaCl, pH6 or 40°C and measure the effect on the survival kinetics. For each of three pre-stress conditions, we find substantial upregulation of proteins, see Fig. 5.5A-C, including those from the respective response ('response to osmotic stress', 'response to acidic pH' & 'response to heat'), as well as the general 'response to stress'. Despite this upregulation of the cellular stress response, we observed no significant change of death rate for either pre-stress condition, see Fig. 5.5D. Thus, the stress response is not limiting starvation survival. In [155, 158] we have investigated the impact of catabolism on carbon starvation and found it not to be limiting. Oxidative damage by reactive oxygen species (ROS) can be a major threat to bacteria and have been previously proposed as survival limiting in starvation [165]. ROS are formed as a natural byproduct of oxygen metabolism in aerobic conditions. To test if ROS are survival limiting, we starved *E. coli* in anaerobic conditions, after growth in either aerobic or anaerobic conditions. If ROS were limiting, the culture will survive better compared to an aerobic control. In contrast, in anaerobic conditions, irrespective of the previous condition, death rate even accelerated, Fig. B.3, presumably due to the lack of oxidative respiration for biomass recycling.

5.3.6 Survival is sensitive to perturbations of the cell envelope

The 'cell envelope' includes the periplasm and outer membrane, sketched in Fig. 5.6A. It provides mechanical stability, imports nutrients and is a selective chemical barrier [166] that keeps abiotic chemicals out. It consists of three distinct regions, the peptidoglycan layer (cell wall), the outer membrane and a space in-between (periplasm). Outer membrane and cell wall are physically linked by anchors proteins (blue). Nutrients and ions are imported via porins (red) and transporters (yellow) across the outer and inner membranes, respectively. Numerous other proteins (grey) are responsible for correct synthesis and maintenance of peptidoglycan, outer membrane ('OM') and outer membrane proteins ('OMP'). The abundance of proteins in the cell envelope depends on growth state of the cell, see methods for quantification details. In slow growth, about 26% of the protein mass is associated with the cell envelope, compared to 13% in rich medium (LB), see Fig. 5.6B. The biggest share of protein mass is due to Lpp (up to 8.5%) and OmpA (up to 3.3%), two proteins that anchor the outer membrane to the cell wall. These two anchors show a strong dependence on growth rate, increasing their abundance from 5.1% on LB to 11.8% in stationary phase. Smaller shares of the protein mass are associated with outer membrane porins (3.8% to 4.7%), mostly due to the high abundance of OmpF and OmpC, and with a large number of proteins associated with nutrient and ion uptake (2.0% to 1.0%). Proteins responsible for correct assembly and maintenance of the cell envelope (grey shadings) make up less than 1% of the protein mass of *E. coli*.

To test if the survival kinetics are sensitive to perturbations of the cell envelope, we tested key knock-outs of non-essential cell envelope proteins, see Fig. 5.6C. We found key genes throughout the envelope that are essential for survival, including two of the most abundant proteins of *E. coli*, the membrane anchoring proteins (Lpp, OmpA). Porins (OmpF, OmpC

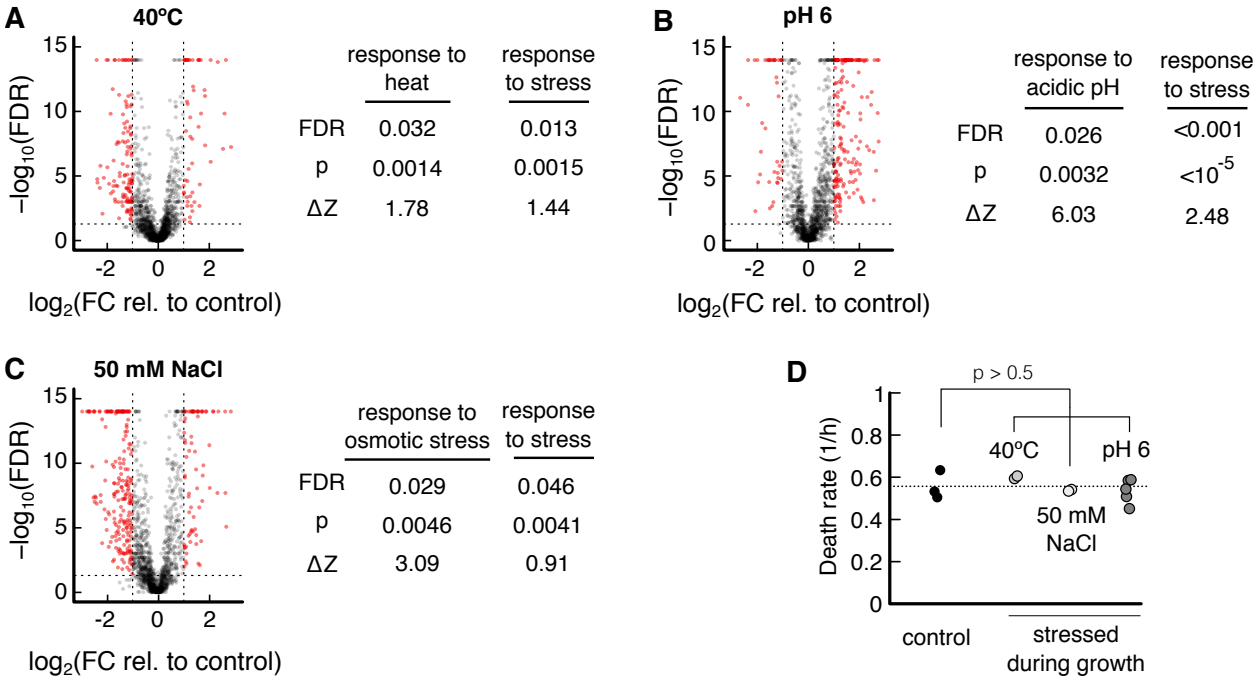


Figure 5.5: Effect of pre-stressing on proteome and survival kinetics. (A-C) On left side of each panel, volcano plots of individual proteins, showing the probability of being a response versus the logarithm of the fold change. Proteins with fold change higher than 2 and false discovery rate smaller than 0.05, are colored in red. On the right side of the panel, for each pre-stress (40°C, pH 6 and 50 mM NaCl), the corresponding stress response and the general 'response to stress' are tested for significant upregulation using Kolmogorov-Smirnov tests. In each pre-stress, both the specific and the general stress response are significantly upregulated, FDR < 0.05. (D) After stressing during growth, bacteria are transferred to pre-warmed, carbon-free minimal medium without stress. Death rates of neither pre-stressing condition lead to a significant change in death rate.

and double-KO OmpF & OmpC) and transporters also take up a substantial fraction of the proteome, Fig. 5.6B, but show no increase of death rate, Fig. 5.6C, indicating that nutrient uptake is not limited by these proteins.

Several of the less abundant proteins show a high increase of death rate upon knock-out, too. Outer membrane protein (OMP) chaperones (skp) and OMP assembly factors (bamE), regulators of peptidoglycan hydrolases (prc & nlpI – regulators of MepS, yraP & nlpD – regulators of AmiC), regulators of cell envelope homeostasis (cpxA - kinase of the cpx regulon, rseA – anti-sigma factor of RpoE) and outer membrane lipid asymmetry maintenance (mlaA, mlaC). Many of these genes have in common that they lead to a destabilization of the envelope. This implies that the overall integrity of the envelope is important for survival.

In line with this result, we find that the antibiotic polymixin B, which targets the envelope via the outer membrane drastically reduces the number of viable cells, while the antibiotic tetracycline, which targets translation does not show any effect, see Fig. 5.6D.

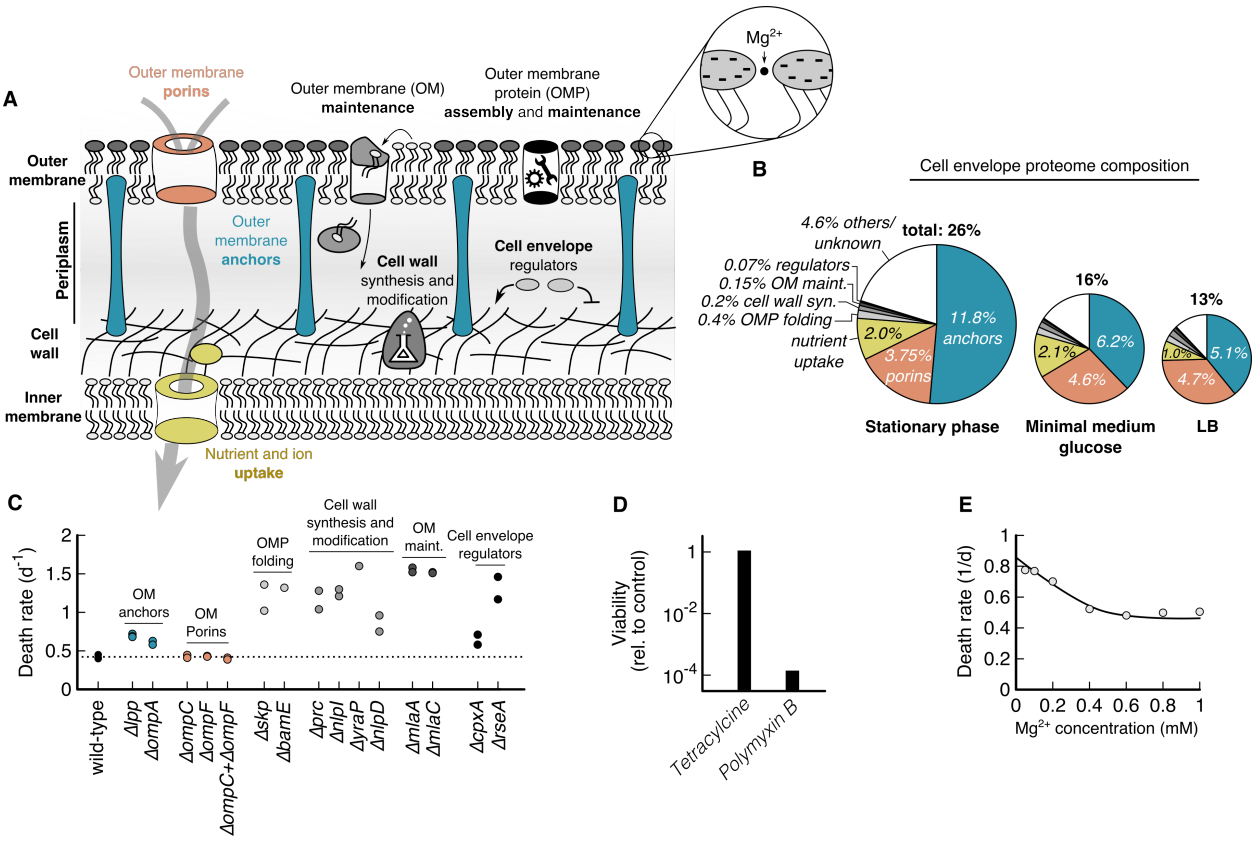


Figure 5.6: Cell envelope integrity is essential for survival. (A) Cartoon of the cell envelope, consisting of the peptidoglycan layer (cell wall), the outer membrane, and the space in-between. Proteins of high abundance are colored: Anchors (blue), outer membrane porins (red) and nutrient and ion uptake (yellow). Other proteins for assembly and maintenance of the cell envelope are colored in shades of grey. Note the lipid asymmetry in the outer membrane with highly negatively charged Lipid A, shielded by Mg2+ ions, on the outer leaflet, and phospholipids on the inner leaflet. (B) Cell envelope proteome composition in three different growth conditions. Size of the pie is proportional to the total abundance of cell envelope proteins, defined as fraction of the total protein mass. Outer membrane anchors OmpA and Lpp (blue) show a strong increase with decreasing growth rate from 5.1% to 11.8% of the total protein mass. Porins (red), proteins required for nutrient or ion uptake (yellow) have considerable abundance, but show no consistent increase with decreasing growth rate. (C) Death rates of knock-outs of key cell envelope genes. Knock-outs of outer membrane anchors (blue) show an increase of death rate. Outer membrane porins (red) do not increase death rate, indicating that nutrient uptake is not limited by porins. Several knock-outs of cell envelope assembly and maintenance lead to strongly increased death rates. (D) Death rates decrease with increasing concentration of Mg2+ in the medium. Regular minimal medium used in this work contains 0.41 mM Mg2+, see methods for details. Bacteria require high concentrations of divalent ions to shield negatively charged Lipid A [167, 168].

5.3.7 Survival is limited by the mechanical stability of the cell envelope

To distinguish whether mechanical stability is 'essential, but not limiting' or 'essential and limiting' survival, we aimed to increase the stability of the cell envelope. The stiffness of the outer membrane depends on the concentration of divalent Mg2+, which plays a crucial role in shielding the highly negatively charged Lipid A, see inset on top right of Fig. 5.6A. We found that increasing the Mg2+ to 1 mM decreases the death rate by a factor of two, Fig. 5.6E. Treating the culture with 50 mM EDTA, a chelating agent of divalent ions, death rate increased by a factor of two, see Fig. 5.6E. These results show that the mechanical stability of *E. coli* is 'essential and limiting' survival.

5.3.8 Time-lapse microscopy reveals cell envelope failures in starvation

We hypothesize that the mechanical stability of the envelope is needed to protect *E. coli* from internal turgor pressure. Such turgor pressure is actively regulated by bacteria, which pump ions across the inner membrane. This active transport requires dissipation of energy, which is a scarce resource in starvation. To investigate ion dysregulation and its connection of cell death on single cell level, we monitored starving *E. coli* stained with a membrane potential reporter $DiBAC_4(3)$ in an inert glass chamber on a time-lapse microscope. During starvation, bacteria are in plasmolysis, with cytoplasm contracted and detached from the outer membrane, see Fig. 5.7 (starvation). We find that bacteria spontaneously lose their membrane potential, followed by an expansion of the cytoplasm . In some cells a bleb formsx while on others it does not. In all cases, we observe a slow bleeding of cytoplasm, reported by a cytoplasmic fluorescent protein (mKate), indicating mechanical failure of the cell envelope.

5.4 Discussion and Conclusion

In this chapter we have presented an investigation on relevant proteins for *E. coli* carbon starvation. We have found that differing physiological perturbations during growth have differing impact on the death rate of *E. coli* bacteria. We then searched for proteins in MS proteomics data that were regulated into the opposite direction as the death rate under the perturbations. We developed the CZ score to evaluate how well a protein responded into the direction of interest over all measured perturbations. The underlying assumption was, that there is an overall 'survival sector' in the proteome and increased expression of this sector results in better starvation survival. It should be highlighted that this is not necessarily the case. The increased survival could be the 'by-product' of each individual perturbation and due to differing and complex proteome configurations. In previous studies, it has however been shown that specific regulatory sectors do exist in the *E. coli* proteome, with clearly defined functions, such as Catabolism or Anabolism [162, 160]. Moreover, we see strongly enriched proteome sectors in the CZ score. Subsequent to the data analyses, we perform physiological analyses. This reveals that most of the enriched sectors are not limiting for starvation. For example, high levels of oxidative damage are deadly for cells

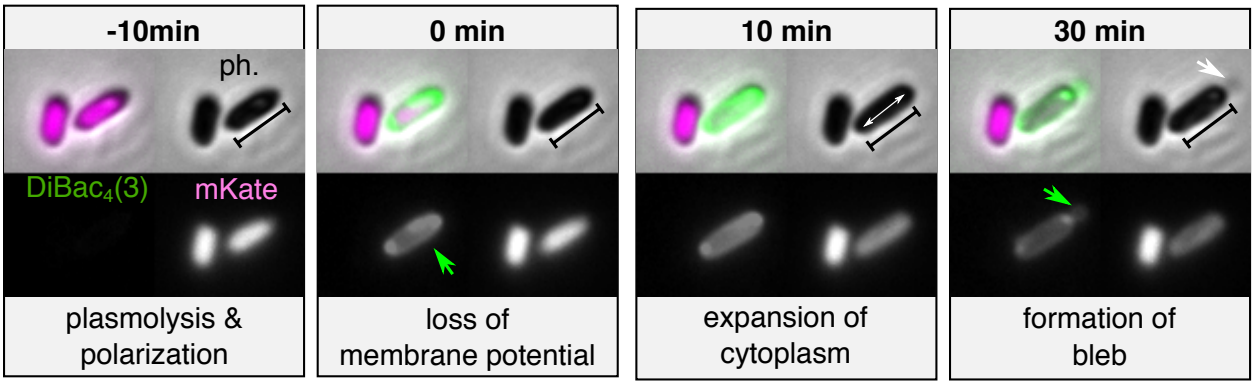


Figure 5.7: Timelapse microscopy of starving *E. coli* indicates pressure-related death event. After loss of polarization, we see expansion of the cytoplasm, in some cases followed by bleb formation. The green $DiBAC_4(3)$ dye quantifies depolarization of the membrane, while the purple mKate is a fluorescent cytoplasmic protein.

in general and decrease starvation survival, while anaerobic (i.e. 'non-oxidative') starvation of bacteria does not decrease the death rate. A strongly enriched proteome sector that survives the physiological tests is the cell envelope. We see that a substantial fraction of the proteome is allocated into this sector and that weakening the envelope increases death rate, while stabilising the envelope increases survival. Additionally, we observe membrane rupture events during starvation. These combined results indicate that we discovered an essential role of the cell envelope for starvation survival, using a combined data-driven and physiological approach.

5.5 Methods

5.5.1 Proteomics data processing

The MS proteomics data was downloaded from the corresponding PRIDE partner repositories (Schmidt et al.: PXD000498, Hui et al.: PXD001467, Houser et al.: PXD002140). For the Schmidt and Houser datasets, the .raw files were downloaded, for the Hui dataset, peptide-intensity mappings were downloaded. The Schmidt and Hui datasets were searched with MaxQuant [39] v. 1.5.7.4 using standard settings and 'Label Free Quantification' (LFQ) and 'Match between runs' enabled. The Schmidt and the Houser datasets were searched against the reviewed Uniprot *E. coli* K-12 database (03/2019).

5.5.2 Differential expression analysis of proteomics data

All datasets were processed with the MS-EmpiRe [169] algorithm for differential quantification, which assigned fold changes and significance scores to each protein in a condition pair. Each condition contained three replicates. For the Schmidt and Houser datasets, the MaxQuant 'peptides.txt' and 'proteinGroups.txt' files were used input. The Hui dataset was pre-processed as follows: As in the experimental setup of Hui et al. the data was quantified relative to a 15N-labelled spike-in, a direct assessment of LFQ values was not optimal. Similar to the approach of Geiger et al. [53], we first assessed the fold changes of each peptide relative to its heavy labeled spike-in. To preserve the intensity information, we re-scaled each spike-in fold change by the median intensity of the spike-in peptide over all conditions. This resulted in pseudo-intensities, which we further processed in a 'LFQ-like' manner using MS-EmpiRe. Our dataset consisted of the perturbations 'Transporter Titration', 'Ribosome Limitation' and 'Anabolic Limitation' in the Hui set, 'Carbon Substrates', 'Chemostat', 'Rich Media', 'Stationary Phase', 'Osmotic Shock', 'Heat Stress' and 'PH Stress' in the Schmidt set and 'Stationary Phase' in the Houser set. Each dataset had a glucose reference condition and the 'Stationary Phase', 'Osmotic Shock', 'Heat Stress' and 'PH Stress' perturbations were compared to the corresponding glucose reference using differential expression analysis with MS-EmpiRe.

5.5.3 Scoring the direction of the proteomic response

As displayed in Fig. 5.1 and discussed in the main text, each growth condition has a corresponding death rate. Additionally, when one condition has a lower death rate than the other, we expect the lower death rate to be caused by increased expression of proteins. In our ranking, we hence wanted to identify proteins that show increased expression for decreased death rate. Each condition consisted of several sub-conditions (e.g. different levels of transporter titration in the C data set). The sub-conditions were sorted from lower to higher death rate. They were then compared in an increasing manner (e.g. C1 vs C2, C2 vs C3, C1 vs C3). This way, positive log2 fold changes always correlate with better survival. For a given protein, we are mainly interested whether it consistently follows the direction of the response. The response is consistent when the fold change of every comparison is positive. For the evaluation, we hence focussed on the sign and the significance of each

comparison. To implement this, we used signed Z-values to calculate a final score for the overall response.

5.5.4 Z-value based ranking of the proteome perturbations

In the MS-EmpiRe algorithm, peptide fold changes are assessed in the context of *empirical background distributions* which estimate the noise of a peptide fold change. As discussed in the MS-EmpiRe paper [169], it is helpful to transform such fold changes into standard normally distributed Z-values, via an adaptation of the Stouffer [130] method. The Z-values carry the information about the direction of the change, the background distribution and the strength of the change (i.e. direction and significance). To combine the proteomics responses, we obtained the Z-value for each protein from MS-EmpiRe. Multiple Z-values for a given condition can simply be summed. The summed Z-values can then be transformed again to standard normal by estimating the variance of the summed distribution, see section 3.3.4 of this thesis for more details.

5.5.5 Combination of the ranked proteome perturbations

In a first iteration, we applied the procedure described above to all perturbations to assess the overall $N(0,1)$ Z-value of each perturbation (i.e. for the 'C,A,R,L,S'). This resulted in a list of scored proteins in each perturbation, where a high Z-value means that the protein correlates well with starvation survival. Proteins with scores in less than three perturbations were excluded from the analysis. Depending on the number of samples and the quality of the data, the distributions of the scored perturbations can span different ranges. As we are interested in proteins that score generally well over all perturbations, we re-scaled the individual distributions of scored perturbations with the factor $rs = 3.09/Z_{max}$, such that the maximum absolute Z-value was 3.09 (corresponds to a p-value of 0.001). Re-scaling was only applied to lower significance (i.e. if rs<1). This prevented that a single perturbation dominated the overall score as follows: If a perturbation consists of many samples with strong changes between them, the Z-values become more extreme than if a perturbation consists of few samples with more subtle changes. The final score was then determined by again assessing normalized N(0,1) Z-Values via equations 1 and 2. Using linear combinations, we ensured, that the overall contributions of C,A,R,L,S perturbtaions had equal weight. Due to the rescaling before, the final score should not directly be transformed back into an overall p-value, but is a useful measure for an overall ranking. The overall score is robust against missing values, as perturbations that are not available also do not contribute to the variance v of the summed distribution $\tilde{\phi} = N(0, \sqrt{v})$. Due to this effect, fewer perturbations that are very clear can also result in a high score.

5.5.6 Absolute quantification of proteins

For absolute quantification, we used protein synthesis rates derived by Li et al. [170] from ribosomal sequencing data of a MG1655 glucose reference condition. Synthesis rates were used as proxies for copy numbers and multiplied by the respective molecular weight to obtain mass estimates. Further conditions were compared relatively to the reference with

MS-EmpiRe and the mass estimates were scaled by the respective fold changes. To determine the mass fraction of a gene set, the genes of the set were summed and divided by summed mass of all genes.

5.5.7 GO enrichment analyses

The Gene Ontology (GO) was downloaded from http://geneontology.org (03/2019) together with the *E. coli* 'ecocyc.gaf' annotation. The relations 'is_a' and 'part_of' were used for the construction of the gene sets. The analysis was carried out using the Kolmogorov-Smirnov test with signed scores. Multiple testing correction was carried out via the Benjamini-Hochberg procedure [79].

Chapter 6

Conclusion and Outlook

The topics of this thesis cover essential aspects of MS proteomics data analysis. They have to be understood in their respective technological and computational context. The MCIP approach introduced in chapter 2, for example, covers the topic of peptide identification. We show that we can increase the number of identified peptides in an MS proteomics run with our approach. However, despite all current efforts, a large fraction (usually around 50%) of spectra acquired in an MS proteomics run are currently unidentified. Major reasons for missed identifications are chemical modifications of the peptide, uncharacteristically digested peptides, sequence mutations and chimeric spectra (i.e. spectra stemming from multiple precursors) [25]. Modified peptides and uncharacteristically digested peptides can be increasingly identified with 'open' searching approaches, where a wide variety of chemical modifications and sequences is screened for each peptide spectrum [25, 24]. Sequence mutations can be addressed by either integrating genomics data in the search [171] or by de-novo peptide identification approaches. Chimeric spectra can be deconvoluted for increased identification, which is especially important for DIA spectra [58, 59]. In the field of spectral library searching, remarkably accurate predictions of fragmentation spectra from deep neural networks have been shown [172, 173, 174, 175]. Novel proteomics instruments increasingly measure the *ion mobility* of peptide ions, which adds an additional independent dimension that can be utilized for peptide identification [176]. Despite these possibilities, the so called 'dark proteome' still makes up a substantial fraction of proteomics data. A recent combination of latent space embedding and open spectral library searching of a large scale proteomics repository shows that more than 30% of spectra still cannot be assigned to a peptide sequence [177]. Increases in sensitivity are likely necessary to further elucidate the dark proteome. The MCIP approach introduced in this thesis is one effort in this direction and can be seen as one of many steps in the collective effort to reach comprehensive peptide identification.

The MS-EmpiRe model introduced in chapter 3 improves the statistical detection of regulated proteins in quantitative proteomics data. We show substantial increases in the detection of regulated proteins (i.e. the sensitivity). Such sensitivity increases are however limited. Ultimately, the statistics are constrained by the precision of the measurements. Apart from the overall number of detected proteins, two main factors limit accurate quantification in proteomics data: missing values and technical noise. In our analysis, sensitivity

depends strongly on these factors and consequently, many developments in MS proteomics technology aim at reducing the missing values and technical noise. Superior properties in both of these aspects are the main reason for the success of DIA methods. For DDA data, recent computational approaches show a strong decrease in the number of missing values by integrated matching of chromatographic features in neighbouring proteomics runs [178]. Novel data acquisition approaches provide an improved basis for this by increased detection of features [179]. Isobaric labelling approaches have also seen futher developments in both the chemical design of the compounds that allow higher precision [180] and higher multiplexing [181] and the data acquisition strategies that allow for higher coverage and higher accuracy [182]. A future challenge for differential quantification will also be the appropriate handling of clinical proteomics data. Data acquisition methods are getting increasingly fast and increasingly automated. A key point will be scalability of computational methods and their ability to appropriately handle datasets with thousands of patient samples.

As discussed in chapter 4, the detection of alternative splicing events in proteomics data is highly challenging. Recent approaches, including the approach presented in this thesis, focus on the improvement of computational methods for increased detection of alternative splicing. In our approach, we add a statistical framework to evaluate differential alternative splicing to the proteomics pipeline. Improving quantification as described in the section above is hence a key factor for improving the study of splicing regulation on the protein level. Also increasing the number of quantified peptides, for example with novel data acquisition approaches, is highly important. The emerging field of Proteogenomics deals with the integration of proteomics data with other data types such as genomics, transcriptomics or translatomics data. Such context-based approaches could benefit from novel more targeted data acquisition strategies [183]. Peptides of interest can be determined from the context of the other data and then be targeted in the MS run. One technology that is theoretically predestined for the detection of alternative splicing events is top-down proteomics [7], where the intact protein is measured by the MS and no further mapping to isoforms is necessary. A key challenge is to achieve the necessary throughput, sensitivity and cost efficiency compared to bottom up proteomics.

In chapter 5 we have presented a collaborative study to better understand carbon starvation of *E. coli*. On the computational side, the challenge was to find an appropriate quantitative description of the complex experimental setup. In principle, quite a few frameworks for downstream processing and analysis of omics data exist [184, 185, 186, 187, 188, 189]. However, for more complicated setups it is still challenging to always find an appropriate solution. In regard to the biological aspects of our study, it should be noted that for our analysis, we study the regulation of proteins to understand the response of the overall system, which implies a significant reduction in complexity. For the analysis of the systemic response, we have correlated groups of genes with underlying and annotated biological processes. As opposed to this reductionist approach is the integrative approach, pursued by a subbranch of systems biology, where complete regulatory models of the system are approximated. Currently these models are limited by data quality. Even if data acquisition were perfect and all molecular components of the systems could be tracked with highest spatial and time resolution, it is unclear to which extent such models could predict the behaviour

of a biological system. An important point in our study is that we validate our analyses by performing additional physiological and imaging experiments. This feedback between theory and experiment is imperative in modern biology and will only become more important in the future.

In conclusion, this thesis underlines the importance of computational methods in MS proteomics. We have demonstrated substantial improvements along several steps in the computational proteomics pipeline and have shown an interesting application of biological data analysis. MS proteomics has already greatly contributed to biology and it will be exciting to see, how far it will further develop in the future. Current studies [190, 191, 192, 193, 171] show the great promise for MS proteomics in data-driven fields such as personalized medicine, which will offer great challenges and great opportunities for computational proteomics.

Appendix A

Supplement - Detecting differential alternative splicing in MS proteomics data

Note: Supplemental Tables and other outputs of substantial size have been deposited online and are accessible under: https://www.bio.ifi.lmu.de/files/ammar/DISSERTATION/index.html

A.1 Supplemental figures

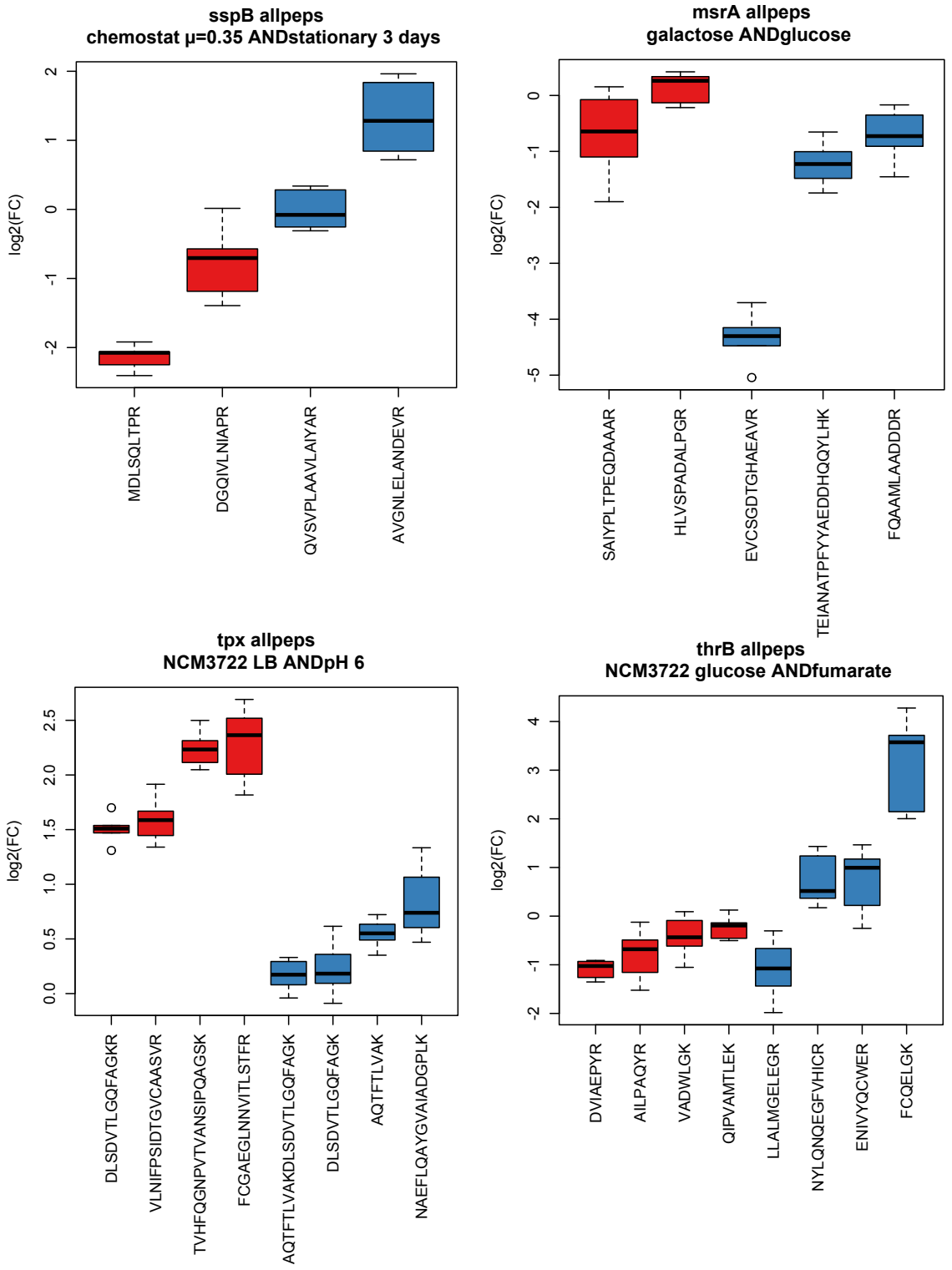


Figure A.1: Examples of *E. coli* proteins with inconsistent peptides. We see strong and systematic differences in the fold changes of the individual peptides. These systematic shifts could be due to post-translational modifications or systematic biases in the data. Peptides were randomly assigned into the red and blue groups and tested with MS-EmpiReS.

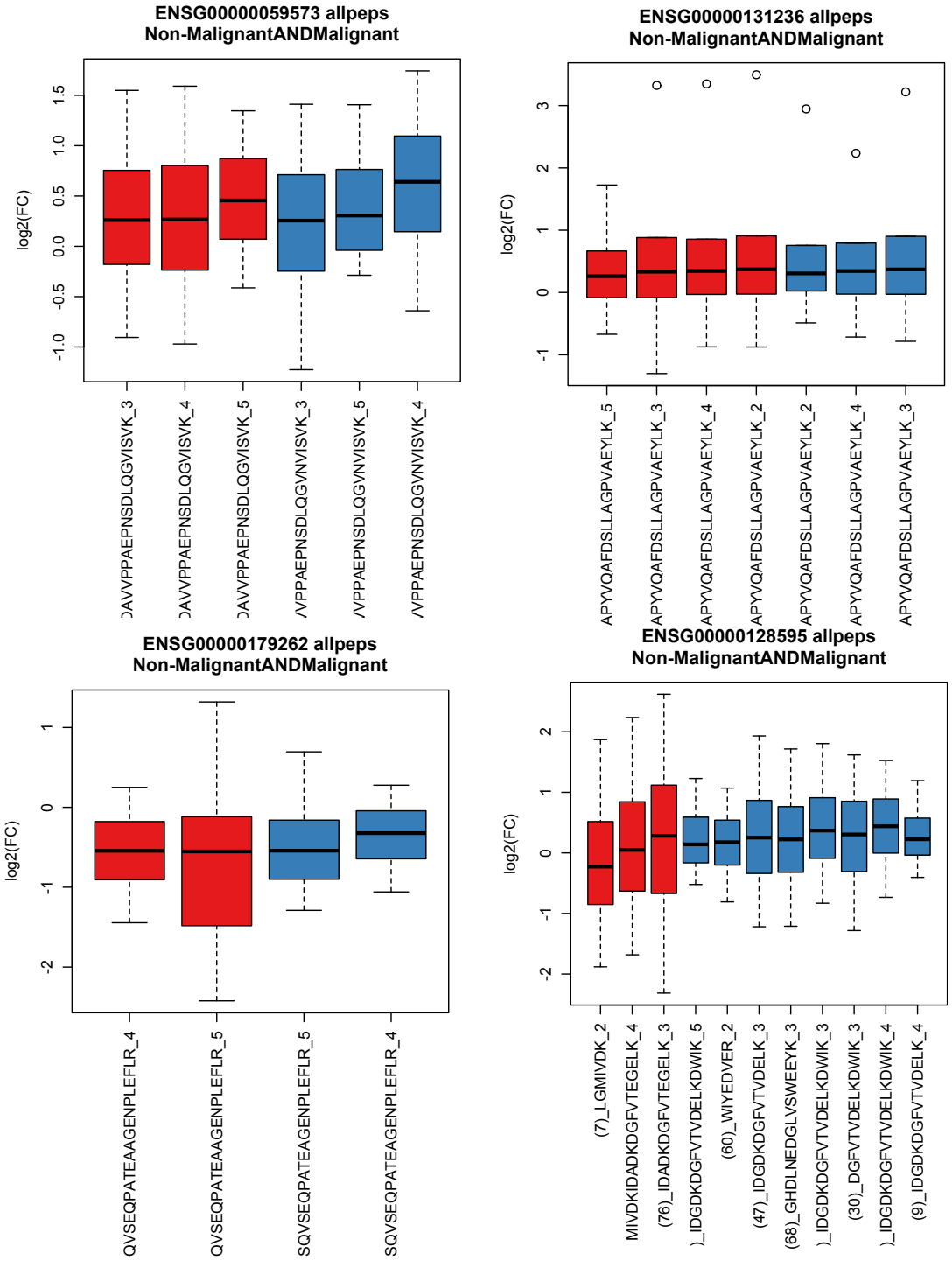


Figure A.2: Examples of spliced genes with no visible regulation. The genes have splice conflicts on the sequence level, however the ratios of both isoforms (red and blue) show no substantial change relative to each other between conditions.

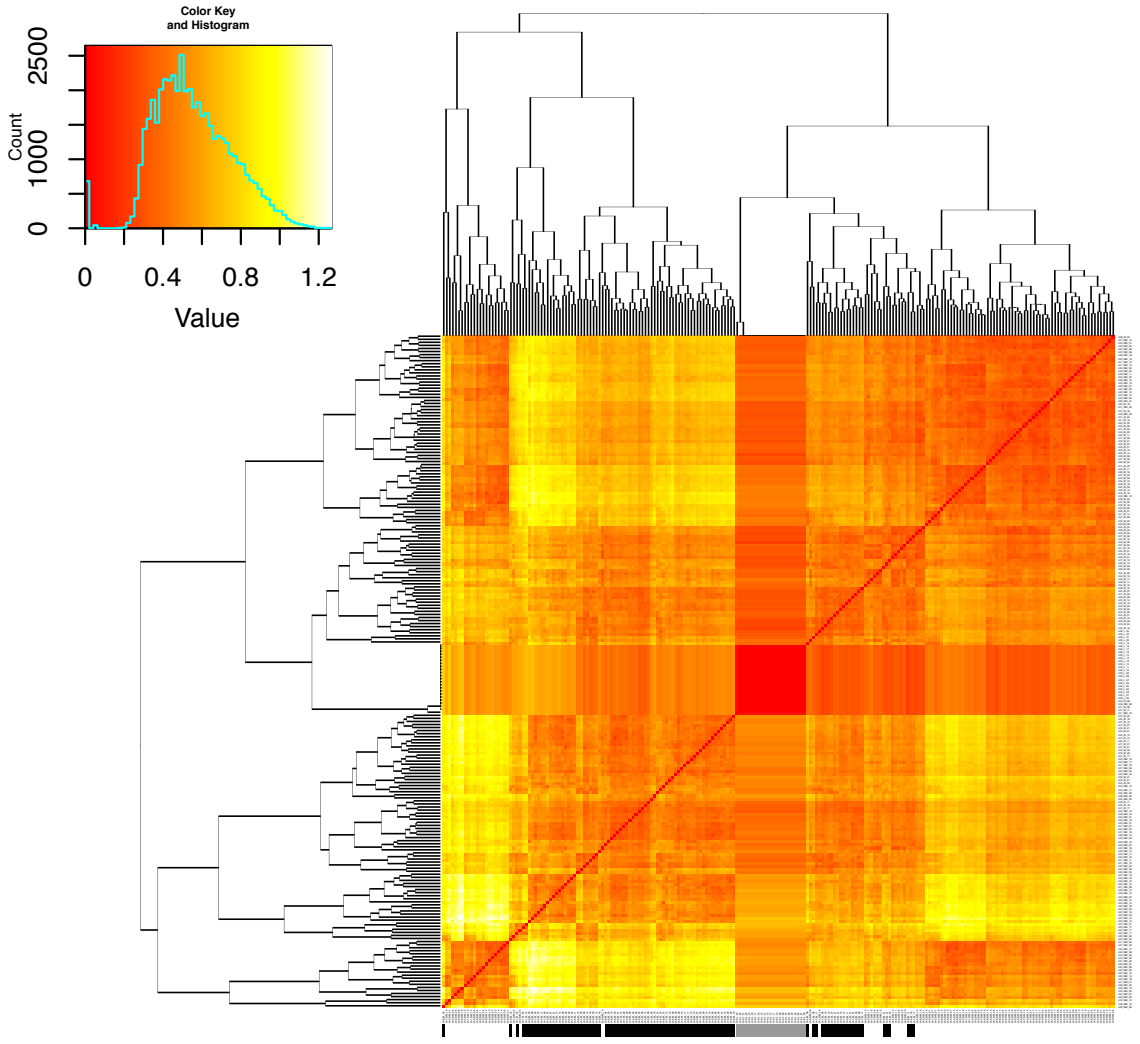


Figure A.3: Pairwise comparison of all samples in the CPTAC dataset after normalization. The color in the heat map indicates the standard deviation of the peptide fold change distribution between the samples. Many of the cancer and healthy samples cluster together already. Cancer samples are indicated with a black bar, linker channel measurements are indicated in grey.

A.2 Supplemental text

In the case that we have three particular groups of sufficiently quantified peptides, we can estimate the ratios between the equivalence classes: One group has to map to equivalence class 1, one group has to map to equivalence class 2 and one group has to map to both equivalence classes. To show this, we write the copy numbers of peptides in equivalence class 1 as A_1, A_2, with the indices 1, 2 denoting the conditions 1 and 2. Analogous, we write B_1, B_2 for equivalence class 2. With this notation, we can write one equation for each group of peptides:

$$I)\ \frac{A_2}{A_1} = f_A, \tag{A.1}$$

$$II)\ \frac{B_2}{B_1} = f_B, \tag{A.2}$$

$$III)\ \frac{A_2 + B_2}{A_1 + B_1} = f_{A+B}, \tag{A.3}$$

with f_A, f_B, f_{A+B} being the respective fold changes between conditions that we can estimate from the proteomics measurements. Re-aligning the equations gives

$$I)\ -A_1 f_A + A_2 = 0, \tag{A.4}$$

$$II)\ -B_1 f_B + B_2 = 0, \tag{A.5}$$

$$III)\ -A_1 f_{A+B} + A_2 - B_1 f_{A+B} + B_2 = 0. \tag{A.6}$$

Calculating III) -I) - II) and re-aligning gives

$$\frac{A_1}{B_1} = \frac{f_{A+B} - f_B}{f_A - f_{A+B}}, \tag{A.7}$$

which is an estimate of the ratio between equivalence classes 1 and 2 in condition 1. It should be noted that this calculation does not model the noise which is a substantial factor in the data and represents a very idealized picture. The results of the calcuation should be seen as a rough estimation of the order of magnitude of the isoform ratio.

Appendix B

Supplement - Detecting relevant proteins for *E. coli* carbon starvation in MS proteomics data

Note: Supplemental Tables and other outputs of substantial size have been deposited online and are accessible under: https://www.bio.ifi.lmu.de/files/ammar/DISSERTATION/index.html

B.1 Experimental protocols

B.1.1 Strains

All strains used in this study are derived from wild type *E. coli* K-12 strain NCM3722 [194]. Strains NQ381 and NQ399 used for 'catabolic limitation' are reported in You et al. [161], NQ393, used for 'anabolic limitation' was reported in Hui et al. All knockouts were transferred from the Keio collection [195] to NCM3722 via P1 transduction to yield strain.

B.1.2 Culture medium

The culture medium N-C- minimal medium [196], contains 1g K2SO4, 17.7 g K2HPO4, 4.7 g KH2PO4, 0.1 g MgSO4 7H2O and 2.5 g NaCl per liter. The medium was supplemented with 20 mM NH4Cl, as nitrogen source, and varying carbon sources. The 'reference glucose condition' contained 0.2% glucose. All chemicals were purchased from Sigma Aldrich, St. Louis, Mo, USA.

B.1.3 Culture conditions

Prior to each experiment, bacteria were streaked out from -80°C glycerol stock on an LB agar plate supplemented with antibiotics if necessary. Bacteria were cultured in three steps. First, a seed culture was grown in lysogenic broth (LB) from a single colony. Second, the seed culture was diluted in N-C- - minimal medium supplemented with 20 mM NH4Cl and a carbon source and grown overnight for at least 5 doublings to exponential phase. The next morning, the overnight culture was diluted into fresh, pre-warmed N-C– minimal medium

supplemented with 20 mM NH4Cl and a carbon source and grown for another 5 to 10 doublings. At an optical density of 0.5 or below, the culture was washed by centrifugation (3 min at 3000 g) and resuspension into fresh, carbon-free, pre-warmed N-C- minimal medium supplemented with 20 mM NH4Cl. This washing step removes excreted fermentative byproducts such as Acetate. For growth conditions known to fully respire carbon, e.g. wild-type NCM3722 grown on glycerol [159], this washing step was omitted. For small culture volumes (5 to 7 ml), 20 mm x 150 mm glass test tubes (Fisher Scientific, Hampton, NH, USA) with disposable, polypropylene Kim-Kap closures (Kimble Chase, Vineland, NJ, USA) were used. For larger volumes, baffled Erlenmeyer flasks (Chemglass, Vineland, NJ, USA) were used.

B.1.4 Viability measurements

For viability measurements, cultures were diluted in untreated, sterile 96 well plates (Cell-treat, Pepperell, MA, USA) in three to four steps using a multichannel pipette (Sartorius, Göttingen, Germany) to a target cell density of about 4000 CFU/ml. 100 µl of the diluted culture was spread on LB agar plates supplemented with 25 g/ml of 2,3,5-triphenyltetrazolium chloride to stain colonies bright red using Rattler Plating Beads (Zymo Research, Irvine, CA, USA), and incubated for 12 to 24 hours. Images of agar plates were taken with a Canon EOS Rebel T3i (Tokyo, Japan) mounted over an LED light box 'Lightpad A920' (Artograph, Delano, MN, USA), and analyzed using a custom script in Cell Profiler (ref). Colony forming units per volume (CFU/ml) were calculated by multiplying the number of colonies per agar plate by the dilution factor.

B.1.5 Stress conditioning

For pre-stressing, wild-type *E. coli* NCM3722 was grown in glucose minimal medium, either in a water bath at 40°C ('heat stress'), in medium supplemented with 50 mM NaCl ('osmotic stress') or in N-C- medium adjusted to pH 6 using KOH ('pH stress'). At an optical density OD600 of about 0.5, cultures were washed and transferred to pre-warmed, carbon-free N-C- supplemented with 20 mM NH4Cl, and the decay of viability was recorded for about 10 days.

B.1.6 Anaerobic culturing

For anaerobic growth and starvation, cultures were grown in 0.05% glucose minimal medium in an vinyl anaerobic chamber (COY Lab Products, Grass Lake, Mi, USA), in Erlenmeyer flasks (Chemglass, Vineland, NJ, USA) on a magnetic stirrer (IKA RO10, Staufen, Germany), and not washed after the end of growth. For aerobic growth and anaerobic starvation, cultures were grown in 0.05% glucose minimal medium in an air incubator. At an optical density of about 0.5, cultures were centrifuged, supernatants were discarded, and pellets were introduced to the anaerobic chamber. In the anaerobic chamber, pellets were resuspended in pre-warmed, carbon-free minimal medium. All media were degassed prior to being introduced to the anaerobic chamber, and left with open lid to be equilibrated for one week.

B.1.7 Time-lapse microscopy

Microscopy was performed on a widefield inverted Nikon Ti2 fluorescence microscope (Nikon, Tokyo, Japan) equipped with incubation chamber (Okolab, Pozzuoli, Italy) kept at 37ºC, Hamamatsu Flash 4.0 sCMOS camera (Hamamatsu Photonics, Hamamatsu City, Japan), Lumencor Spectra-X light engine (Lumencor, Beaverton, Or, USA) and a 100x, 1.4 NA phase objective (Nikon, Tokyo, Japan). Culture chambers were built by assembling two cover slides (name) with adhesive Secure Seal of 120 µm thickness (Grace Bio-Lab, Sigma-Aldrich, St. Louis, Mo, USA). The result is a flat, hollow chamber with inert glass on top and bottom. A starved culture, diluted to OD 0.075 was loaded through laser-cut holes in the top cover slide and spun down for 3 minutes at 2200 g in a Centrifuge 5430 (Eppendorf, Hamburg, Germany), and holes were sealed. To detect depolarization, the dye $DiBAC_4(3)$ was used, which exhibits increased flourescence when it binds to intracellular proteins and membrane. Increased depolarization leads to increased influx and higher flourescence.

B.2 Supplemental figures

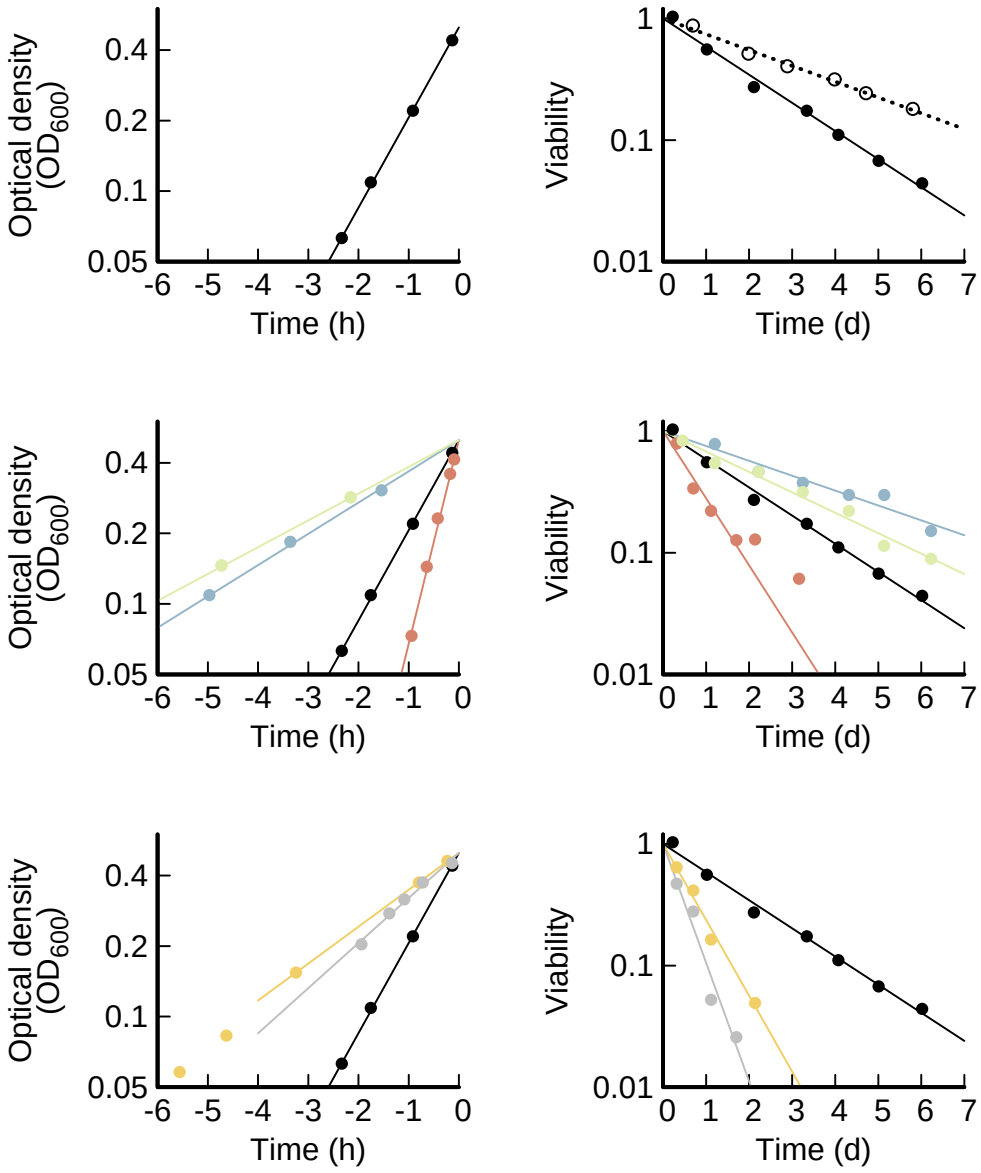


Figure B.1: (Caption next page.)

Figure B.1: (Previous page.) Example growth and death curves of growth perturbations shown in Fig. 1. All cultures were grown until OD 0.5 or less, (left panels), followed by resuspension into fresh, pre-warmed, carbon-free medium (right panels). Data points before -6 h, below OD 0.05, after 7 days or below viability of 0.01 are not shown. Generally, all data points between OD 0.05 and OD 0.5 were used to fit growth rates, and all data points between 109 CFU/ml and 107 CFU/ml were used to measure death rates. (A) Growth on glucose minimal medium. (B) A culture grown on glucose and washed during growth dies exponentially (black – reference condition). If the culture adapted one day in stationary phase on excreted acetate, before being washed and resuspended in carbon-free medium, the death rate will decrease (white symbols – stationary phase). (C) Change of nutrient quality. Comparison of growth in catabolic limitation via titration of LacY (blue), anabolic limitation via titration of glutamate synthesis (green) and LB, a rich medium, (red) with the reference condition (black). (D) Death rate of cultures grown on different nutrient qualities show that growth limitation leads to slower death (blue and green), while rich medium leads to faster death (red). Note that on LB, the decay of viability appears to be non-exponential. In this case we fit only the initial part of the decay. (E) Proteome stress. Comparison of growth when cultures are either limited by ribosome inhibiting 3 µM Chloramphenicol (yellow) or by expression of large quantities of a LacZ, an irrelevant protein (grey). (F) Proteome stress leads to very fast death compared to the reference condition. Neither Chloramphenicol, nor the inducer of LacZ expression are present during starvation. A summary of all growth and death rates is shown in Table S1.

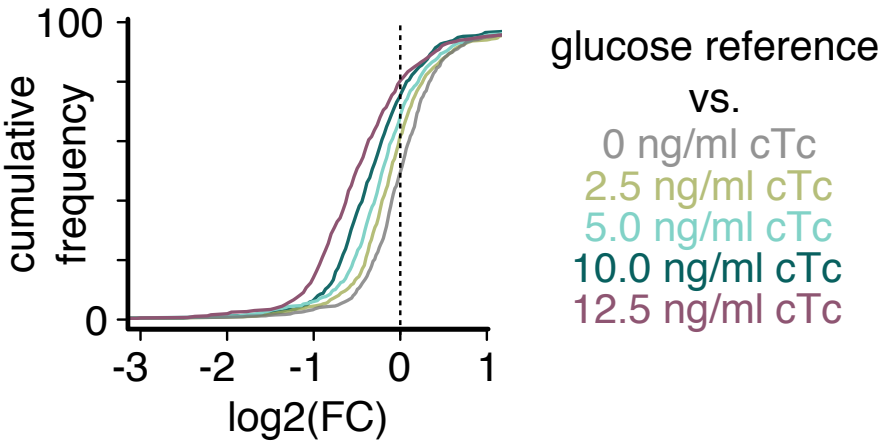


Figure B.2: Investigation of proteomic influence on starvation survival. Cumulative distribution of fold changes for LacZ overexpression relative to a glucose reference for different strengths of perturbation by chlortetracicline (cTc) induciton [197] (2.5ng/ml - 12.5ng/ml). The grey line corresponds to the uninduced strain relative to the reference and can be seen as an estimation of the margin of error. Upon perturbation, we see a systematic down regulation, proportional to the inducer strength, without any visible up regulation. For the highest perturbation, more than 80% of the proteome has a negative fold change and no positive fold change is visible above the estimated error.

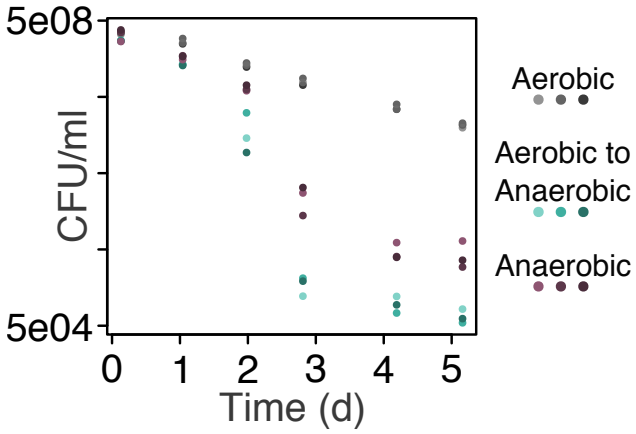


Figure B.3: Comparison of death curves under anaerobic conditions. Death curves under standard aerobic conditions, after aerobic growth and transferral into anaerobic conditions and after both anaerobic growth and starvation are compared. We see a clear increase in death for both anaerobic conditions as compared to the standard aerobic condition.

Bibliography

[1] E. Schrödinger, *What is Life? - The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.

[2] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and J. B. Reece, *Campbell biology*. Pearson, 2017.

[3] J. V. Olsen, S.-E. Ong, and M. Mann, "Trypsin cleaves exclusively c-terminal to arginine and lysine residues," *Molecular & Cellular Proteomics*, vol. 3, no. 6, pp. 608–614, 2004.

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

[5] E. Mostovenko, C. Hassan, J. Rattke, A. M. Deelder, P. A. van Veelen, and M. Palmblad, "Comparison of peptide and protein fractionation methods in proteomics," *EuPA Open Proteomics*, vol. 1, pp. 30–37, 2013.

[6] Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek, and J. R. Yates III, "Protein analysis by shotgun/bottom-up proteomics," *Chemical reviews*, vol. 113, no. 4, pp. 2343–2394, 2013.

[7] N. L. Kelleher, "Peer reviewed: Top-down proteomics," *Analytical Chemistry*, vol. 76, no. 11, pp. 196 A–203 A, 2004.

[8] F. Meier, P. E. Geyer, S. Virreira Winter, J. Cox, and M. Mann, "BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes," *Nature Methods*, p. 1, 2018.

[9] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, *et al.*, "A draft map of the human proteome," *Nature*, vol. 509, no. 7502, pp. 575–581, 2014.

[10] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, *et al.*, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, no. 7502, pp. 582–587, 2014.

[11] M. Yamashita and J. B. Fenn, "Electrospray ion source. another variation on the free-jet theme," *The Journal of Physical Chemistry*, vol. 88, no. 20, pp. 4451–4459, 1984.

[12] E. Barillot, L. Calzone, P. Hupe, J.-P. Vert, and A. Zinovyev, *Computational systems biology of cancer*. CRC Press, 2012.

[13] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, "Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present," *Anal. Bioanal. Chem.*, vol. 404, pp. 939–965, Sept. 2012.

[14] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nat. Biotechnol.*, vol. 26, pp. 1367–1372, Dec. 2008.

[15] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis," *Molecular & Cellular Proteomics*, vol. 11, no. 6, 2012.

[16] M. Y. Hein, K. Sharma, J. Cox, and M. Mann, "Proteomic analysis of cellular systems," in *Handbook of systems biology: concepts and insights*, pp. 3–25, Academic Press, 2013.

[17] J. K. Eng, A. L. McCormack, and J. R. Yates, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *Journal of the american society for mass spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[18] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *ELECTROPHORESIS: An International Journal*, vol. 20, no. 18, pp. 3551–3567, 1999.

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

[20] S. Kim and P. A. Pevzner, "Ms-gf+ makes progress towards a universal database search tool for proteomics," *Nature communications*, vol. 5, p. 5277, 2014.

[21] R. Craig and R. C. Beavis, "Tandem: matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004.

[22] D. L. Tabb, C. G. Fernando, and M. C. Chambers, "Myrimatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis," *Journal of proteome research*, vol. 6, no. 2, pp. 654–661, 2007.

[23] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of proteome research*, vol. 3, no. 5, pp. 958–964, 2004.

[24] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii, "Msfragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics," *Nature methods*, vol. 14, no. 5, p. 513, 2017.

[25] H. Chi, C. Liu, H. Yang, W.-F. Zeng, L. Wu, W.-J. Zhou, R.-M. Wang, X.-N. Niu, Y.-H. Ding, Y. Zhang, *et al.*, "Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine," *Nature biotechnology*, vol. 36, no. 11, pp. 1059–1061, 2018.

[26] J. Griss, "Spectral library searching in proteomics," *Proteomics*, vol. 16, no. 5, pp. 729–740, 2016.

[27] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold, "Building consensus spectral libraries for peptide identification in proteomics," *Nature methods*, vol. 5, no. 10, pp. 873–875, 2008.

[28] B. E. Frewen, G. E. Merrihew, C. C. Wu, W. S. Noble, and M. J. MacCoss, "Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries," *Analytical chemistry*, vol. 78, no. 16, pp. 5678–5684, 2006.

[29] R. Craig, J. Cortens, D. Fenyo, and R. C. Beavis, "Using annotated peptide mass spectrum libraries for protein identification," *Journal of proteome research*, vol. 5, no. 8, pp. 1843–1849, 2006.

[30] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq," *Molecular & cellular proteomics*, vol. 13, no. 9, pp. 2513–2526, 2014.

[31] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann, "Stable isotope labeling by amino acids in cell culture, silac, as a simple and accurate approach to expression proteomics," *Molecular & cellular proteomics*, vol. 1, no. 5, pp. 376–386, 2002.

[32] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms," *Analytical chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003.

[33] S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid, "Protein labeling by itraq: a new tool for quantitative mass spectrometry in proteome research," *Proteomics*, vol. 7, no. 3, pp. 340–350, 2007.

[34] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmström, L. Malmström, *et al.*, "Openswath enables automated, targeted analysis of data-independent acquisition ms data," *Nature biotechnology*, vol. 32, no. 3, p. 219, 2014.

[35] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data: the protein inference problem," *Molecular & cellular proteomics*, vol. 4, no. 10, pp. 1419–1440, 2005.

[36] S. Tyanova, T. Temu, and J. Cox, "The maxquant computational platform for mass spectrometry-based shotgun proteomics," *Nature protocols*, vol. 11, no. 12, p. 2301, 2016.

[37] J. Pfeuffer, T. Sachsenberg, T. M. Dijkstra, O. Serang, K. Reinert, and O. Kohlbacher, "Epifany: A method for efficient high-confidence protein inference," *Journal of proteome research*, vol. 19, no. 3, pp. 1060–1072, 2020.

[38] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, *et al.*, "Openms: a flexible open-source software platform for mass spectrometry data analysis," *Nature methods*, vol. 13, no. 9, pp. 741–748, 2016.

[39] J. Cox and M. Mann, "Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification," *Nature biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008.

[40] L. K. Pino, B. C. Searle, J. G. Bollinger, B. Nunn, B. MacLean, and M. J. MacCoss, "The skyline ecosystem: Informatics for quantitative mass spectrometry proteomics," *Mass spectrometry reviews*, vol. 39, no. 3, pp. 229–244, 2020.

[41] J. H. Gross, *Mass spectrometry: a textbook*. Springer Science & Business Media, 2006.

[42] G. Zhang, R. S. Annan, S. A. Carr, and T. A. Neubert, "Overview of peptide and protein analysis by mass spectrometry," *Current protocols in protein science*, vol. 62, no. 1, pp. 16–1, 2010.

[43] R. Aebersold and M. Mann, "Mass-spectrometric exploration of proteome structure and function," *Nature*, vol. 537, no. 7620, pp. 347–355, 2016.

[44] R. Bruderer, O. M. Bernhardt, T. Gandhi, Y. Xuan, J. Sondermann, M. Schmidt, D. Gomez-Varela, and L. Reiter, "Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results," *Molecular & Cellular Proteomics*, vol. 16, no. 12, pp. 2296–2309, 2017.

[45] C. Ammar, E. Berchtold, G. Csaba, A. Schmidt, A. Imhof, and R. Zimmer, "Multi-reference spectral library yields almost complete coverage of heterogeneous lc-ms/ms data sets," *Journal of proteome research*, vol. 18, no. 4, pp. 1553–1566, 2019.

[46] B. Domon and R. Aebersold, "Mass spectrometry and protein analysis," *Science*, vol. 312, pp. 212–7, Apr 2006.

[47] A. Michalski, J. Cox, and M. Mann, "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc- ms/ms," *Journal of proteome research*, vol. 10, no. 4, pp. 1785–1793, 2011.

[48] R. G. Sadygov, D. Cociorva, and J. R. Yates, "Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book," *Nature methods*, vol. 1, no. 3, pp. 195–202, 2004.

[49] A. Schmidt, N. Gehlenborg, B. Bodenmiller, L. N. Mueller, D. Campbell, M. Mueller, R. Aebersold, and B. Domon, "An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures," *Molecular & Cellular Proteomics*, vol. 7, no. 11, pp. 2138–2150, 2008.

[50] P. Picotti and R. Aebersold, "Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions," *Nature methods*, vol. 9, no. 6, p. 555, 2012.

[51] S. Purvine, J.-T. Eppel*, E. C. Yi, and D. R. Goodlett, "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer," *Proteomics*, vol. 3, no. 6, pp. 847–850, 2003.

[52] R. S. Plumb, K. A. Johnson, P. Rainville, B. W. Smith, I. D. Wilson, J. M. Castro-Perez, and J. K. Nicholson, "Uplc/mse; a new approach for generating molecular fragment information for biomarker structure elucidation," *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, vol. 20, no. 13, pp. 1989–1994, 2006.

[53] T. Geiger, J. Cox, and M. Mann, "Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation," *Molecular & Cellular Proteomics*, vol. 9, no. 10, pp. 2252–2261, 2010.

[54] J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nature methods*, vol. 1, no. 1, pp. 39–45, 2004.

[55] A. Panchaud, A. Scherl, S. A. Shaffer, P. D. von Haller, H. D. Kulasekara, S. I. Miller, and D. R. Goodlett, "Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean," *Analytical chemistry*, vol. 81, no. 15, pp. 6481–6488, 2009.

[56] P. C. Carvalho, X. Han, T. Xu, D. Cociorva, M. d. G. Carvalho, V. C. Barbosa, and J. R. Yates III, "Xdia: improving on the label-free data-independent analysis," *Bioinformatics*, vol. 26, no. 6, pp. 847–848, 2010.

[57] A. Panchaud, S. Jung, S. A. Shaffer, J. D. Aitchison, and D. R. Goodlett, "Faster, quantitative, and accurate precursor acquisition independent from ion count," *Analytical chemistry*, vol. 83, no. 6, pp. 2250–2257, 2011.

[58] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii, "Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics," *Nature methods*, vol. 12, no. 3, p. 258, 2015.

[59] Y. Li, C.-Q. Zhong, X. Xu, S. Cai, X. Wu, Y. Zhang, J. Chen, J. Shi, S. Lin, and J. Han, "Group-dia: analyzing multiple data-independent acquisition mass spectrometry data files," *Nature methods*, vol. 12, no. 12, pp. 1105–1106, 2015.

[60] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein, and R. Aebersold, "Development and validation of a spectral library searching method for peptide identification from ms/ms," *Proteomics*, vol. 7, no. 5, pp. 655–667, 2007.

[61] S. E. Stein and D. R. Scott, "Optimization and testing of mass spectral library search algorithms for compound identification," *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 9, pp. 859–866, 1994.

[62] I. Beer, E. Barnea, T. Ziv, and A. Admon, "Improving large-scale proteomics by clustering of mass spectrometry data," *Proteomics*, vol. 4, no. 4, pp. 950–960, 2004.

[63] C. A. Sherwood, A. Eastham, L. W. Lee, J. Risler, O. Vitek, and D. B. Martin, "Correlation between y-type ions observed in ion trap and triple quadrupole mass spectrometers," *Journal of proteome research*, vol. 8, no. 9, pp. 4243–4251, 2009.

[64] C.-Y. Yen, S. Houel, N. G. Ahn, and W. M. Old, "Spectrum-to-spectrum searching using a proteome-wide spectral library," *Molecular & Cellular Proteomics*, vol. 10, no. 7, 2011.

[65] K. X. Wan, I. Vidavsky, and M. L. Gross, "Comparing similar spectra: from similarity index to spectral contrast angle," *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 1, pp. 85–88, 2002.

[66] J. Wang, J. Pérez-Santiago, J. E. Katz, P. Mallick, and N. Bandeira, "Peptide identification from mixture tandem mass spectra," *Molecular & Cellular Proteomics*, vol. 9, no. 7, pp. 1476–1485, 2010.

[67] D. L. Tabb, M. J. MacCoss, C. C. Wu, S. D. Anderson, and J. R. Yates, "Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility," *Analytical chemistry*, vol. 75, no. 10, pp. 2470–2477, 2003.

[68] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich, and R. Aebersold, "The peptideatlas project," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D655–D658, 2006.

[69] P. Jones, R. G. Côté, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob, and R. Apweiler, "Pride: a public repository of protein and peptide identifications for the proteomics community," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D659–D663, 2006.

[70] M. Riffle and J. K. Eng, "Proteomics data repositories," *Proteomics*, vol. 9, no. 20, pp. 4653–4663, 2009.

[71] J. A. Vizcaíno, R. G. Côté, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, *et al.*, "The proteomics identifications (pride) database and associated tools: status in 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D1063–D1069, 2012.

[72] D. P. Zolg, M. Wilhelm, K. Schnatbaum, J. Zerweck, T. Knaute, B. Delanghe, D. J. Bailey, S. Gessulat, H.-C. Ehrlich, M. Weininger, *et al.*, "Building proteometools based on a complete synthetic human proteome," *Nature methods*, vol. 14, no. 3, pp. 259–262, 2017.

[73] G. Rosenberger, C. C. Koh, T. Guo, H. L. Röst, P. Kouvonen, B. C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, *et al.*, "A repository of assays to quantify 10,000 human proteins by swath-ms," *Scientific data*, vol. 1, p. 140031, 2014.

[74] O. T. Schubert, L. C. Gillet, B. C. Collins, P. Navarro, G. Rosenberger, W. E. Wolski, H. Lam, D. Amodei, P. Mallick, B. MacLean, *et al.*, "Building high-quality assay libraries for targeted analysis of swath ms data," *Nature protocols*, vol. 10, no. 3, p. 426, 2015.

[75] J. A. Vizcaíno, A. Csordas, N. Del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q.-W. Xu, R. Wang, and H. Hermjakob, "2016 update of the PRIDE database and its related tools," *Nucleic Acids Res.*, vol. 44, p. 11033, Dec. 2016.

[76] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, *et al.*, "A cross-platform toolkit for mass spectrometry and proteomics," *Nature biotechnology*, vol. 30, no. 10, pp. 918–920, 2012.

[77] H. Lam, E. W. Deutsch, and R. Aebersold, "Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics," *Journal of proteome research*, vol. 9, no. 1, pp. 605–610, 2010.

[78] P. Navarro, J. Kuharev, L. C. Gillet, O. M. Bernhardt, B. MacLean, H. L. Röst, S. A. Tate, C.-C. Tsou, L. Reiter, U. Distler, *et al.*, "A multicenter study benchmarks software tools for label-free proteome quantification," *Nature biotechnology*, vol. 34, no. 11, p. 1130, 2016.

[79] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[80] R. Bruderer, O. M. Bernhardt, T. Gandhi, Y. Xuan, J. Sondermann, M. Schmidt, D. Gomez-Varela, and L. Reiter, "Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results," *Molecular & Cellular Proteomics*, vol. 41, no. 0, p. mcp.RA117.000314, 2017.

[81] J. X. Wu, X. Song, D. Pascovici, T. Zaw, N. Care, C. Krisp, and M. P. Molloy, "Swath mass spectrometry performance using extended peptide ms/ms assay libraries," *Molecular & Cellular Proteomics*, vol. 15, no. 7, pp. 2501–2514, 2016.

[82] G. Rosenberger, I. Bludau, U. Schmitt, M. Heusel, C. L. Hunter, Y. Liu, M. J. MacCoss, B. X. MacLean, A. I. Nesvizhskii, P. G. Pedrioli, *et al.*, "Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses," *Nature methods*, vol. 14, no. 9, pp. 921–927, 2017.

[83] S. Gallien, E. Duriez, K. Demeure, and B. Domon, "Selectivity of lc-ms/ms analysis: implication for proteomics experiments," *Journal of proteomics*, vol. 81, pp. 148–158, 2013.

[84] R. Peckner, S. A. Myers, A. S. V. Jacome, J. D. Egertson, J. G. Abelin, M. J. MacCoss, S. A. Carr, and J. D. Jaffe, "Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics," *Nature methods*, vol. 15, no. 5, p. 371, 2018.

[85] E. Lau, Y. Han, D. R. Williams, C. T. Thomas, R. Shrestha, J. C. Wu, and M. P. Y. Lam, "Splice-Junction-Based mapping of alternative isoforms in the human proteome," *Cell Rep.*, vol. 29, pp. 3751–3765.e5, Dec. 2019.

[86] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, "Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present," *Analytical and Bioanalytical Chemistry*, vol. 404, no. 4, pp. 939–965, 2012.

[87] J. V. Olsen, "Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap," *Molecular & Cellular Proteomics*, vol. 4, no. 12, pp. 2010–2021, 2005.

[88] L. Ting, R. Rad, S. P. Gygi, and W. Haas, "MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics," *Nature Methods*, vol. 8, no. 11, pp. 937–940, 2011.

[89] D. A. Abaye, F. S. Pullen, and B. V. Nielsen, "Peptide polarity and the position of arginine as sources of selectivity during positive electrospray ionisation mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 25, no. 23, pp. 3597–3608, 2011.

[90] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ," *Molecular & Cellular Proteomics*, vol. 13, no. 9, pp. 2513–2526, 2014.

[91] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS," *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003.

[92] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox, "The Perseus computational platform for comprehensive analysis of (prote)omics data," *Nature Methods*, vol. 13, no. 9, pp. 731–740, 2016.

[93] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, and O. Vitek, "Protein quantification in label-free LC-MS experiments," *Journal of Proteome Research*, vol. 8, no. 11, pp. 5275–5284, 2009.

[94] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W. J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney, "A statistical framework for protein quantitation in bottom-up MS-based proteomics," *Bioinformatics*, vol. 25, no. 16, pp. 2028–2034, 2009.

[95] L. J. Goeminne, A. Argentini, L. Martens, and L. Clement, "Summarization vs peptide-based models in label-free quantitative proteomics: Performance, pitfalls, and data analysis guidelines," *Journal of Proteome Research*, vol. 14, no. 6, pp. 2457–2465, 2015.

[96] L. J. E. Goeminne, K. Gevaert, and L. Clement, "Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics," *Molecular & Cellular Proteomics*, vol. 15, no. 2, pp. 657–668, 2016.

[97] H. W. Koh, H. L. Swa, D. Fermin, S. G. Ler, J. Gunaratne, and H. Choi, "EBprot: Statistical analysis of labeling-based quantitative proteomics data," *Proteomics*, vol. 15, no. 15, pp. 2580–2591, 2015.

[98] R. Rosenthal, "Combining results of independent studies," *Psychological bulletin*, vol. 85, no. 1, p. 185, 1978.

[99] J. D. O'Connell, J. A. Paulo, J. J. O'Brien, and S. P. Gygi, "Proteome-wide evaluation of two common protein quantification methods," *Journal of proteome research*, vol. 17, no. 5, pp. 1934–1942, 2018.

[100] G. K. Smyth, "Limma: linear models for microarray data," *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, 2005.

[101] S. Anders and W. Huber, "Differential expression analysis for sequence count data.," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[102] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2009.

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

[104] A. Zien, T. Aigner, R. Zimmer, and T. Lengauer, "Centralization: A new method for the normalization of gene expression data," *Bioinformatics*, vol. 17, pp. S323–S331, 2001.

[105] K. Sharma, S. Schmitt, C. G. Bergner, S. Tyanova, N. Kannaiyan, N. Manrique-Hoyos, K. Kongi, L. Cantuti, U.-K. Hanisch, M.-A. Philips, *et al.*, "Cell type–and brain region–resolved mouse brain proteome," *Nature neuroscience*, vol. 18, no. 12, p. 1819, 2015.

[106] E. Ramond, G. Gesbert, I. C. Guerrera, C. Chhuon, M. Dupuis, M. Rigard, T. Henry, M. Barel, and A. Charbit, "Importance of host cell arginine uptake in francisella phagosomal escape and ribosomal protein amounts," *Molecular & Cellular Proteomics*, vol. 14, no. 4, pp. 870–881, 2015.

[107] L. Ping, D. M. Duong, L. Yin, M. Gearing, J. J. Lah, A. I. Levey, and N. T. Seyfried, "Global quantitative analysis of the human brain proteome in Alzheimer's and Parkinson's disease," *Scientific data*, vol. 5, p. 180036, 2018.

[108] L. M. Smith, N. L. Kelleher, M. Linial, D. Goodlett, P. Langridge-Smith, Y. A. Goo, G. Safford, L. Bonilla, G. Kruppa, R. Zubarev, *et al.*, "Proteoform: a single term describing protein complexity," *Nature methods*, vol. 10, no. 3, pp. 186–187, 2013.

[109] B. Modrek and C. Lee, "A genomic view of alternative splicing," *Nature genetics*, vol. 30, no. 1, pp. 13–19, 2002.

[110] R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe, "The ribosome-engaged landscape of alternative splicing," *Nature Structural & Molecular Biology*, vol. 23, no. 12, pp. 1117–1123, 2016.

[111] M. L. Tress, F. Abascal, and A. Valencia, "Alternative splicing may not be the key to proteome complexity," *Trends Biochem. Sci.*, vol. 42, pp. 98–110, Feb. 2017.

[112] G. A. Merino, A. Conesa, and E. A. Fernández, "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human rna-seq studies," *Briefings in bioinformatics*, vol. 20, no. 2, pp. 471–481, 2019.

[113] X. Wang, S. G. Codreanu, B. Wen, K. Li, M. C. Chambers, D. C. Liebler, and B. Zhang, "Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity," *Mol. Cell. Proteomics*, vol. 17, pp. 422–430, Mar. 2018.

[114] F. Birzele, G. Csaba, and R. Zimmer, "Alternative splicing and protein structure evolution," *Nucleic Acids Res.*, vol. 36, pp. 550–558, Feb. 2008.

[115] M. Reixachs-Solé, J. Ruiz-Orera, M. M. Albà, and E. Eyras, "Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome," *Nat. Commun.*, vol. 11, p. 1768, Apr. 2020.

[116] F. Abascal, I. Ezkurdia, J. Rodriguez-Rivas, J. M. Rodriguez, A. del Pozo, J. Vázquez, A. Valencia, and M. L. Tress, "Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level," *PLoS Comput. Biol.*, vol. 11, p. e1004325, June 2015.

[117] Y.-Y. Chen, S. Dasari, Z.-Q. Ma, L. J. Vega-Montoto, M. Li, and D. L. Tabb, "Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines," *Analytical and bioanalytical chemistry*, vol. 404, no. 4, pp. 1115–1125, 2012.

[118] S. Jin, D. S. Daly, D. L. Springer, and J. H. Miller, "The effects of shared peptides on protein quantitation in label-free proteomics by lc/ms/ms," *Journal of proteome research*, vol. 7, no. 01, pp. 164–169, 2008.

[119] F. Erhard and R. Zimmer, "Count ratio model reveals bias affecting NGS fold changes," *Nucleic Acids Res.*, vol. 43, p. e136, Nov. 2015.

[120] C. Ammar, M. Gruber, G. Csaba, and R. Zimmer, "MS-EmpiRe utilizes peptide-level noise distributions for ultra-sensitive detection of differentially expressed proteins," *Mol. Cell. Proteomics*, vol. 18, pp. 1880–1892, Sept. 2019.

[121] Y. Liu, M. Gonzàlez-Porta, S. Santos, A. Brazma, J. C. Marioni, R. Aebersold, A. R. Venkitaraman, and V. O. Wickramasinghe, "Impact of alternative splicing on the human proteome," *Cell Rep.*, vol. 20, pp. 1229–1241, Aug. 2017.

[122] M. A. Komor, T. V. Pham, A. C. Hiemstra, S. R. Piersma, A. S. Bolijn, T. Schelfhorst, P. M. Delis-van Diemen, M. Tijssen, R. P. Sebra, M. Ashby, G. A. Meijer, C. R. Jimenez, and R. J. A. Fijneman, "Identification of differentially expressed splice variants by the proteogenomic pipeline splicify," *Mol. Cell. Proteomics*, vol. 16, pp. 1850–1863, Oct. 2017.

[123] A. Kahles, K.-V. Lehmann, N. C. Toussaint, M. Hüser, S. G. Stark, T. Sachsenberg, O. Stegle, O. Kohlbacher, C. Sander, Cancer Genome Atlas Research Network, and G. Rätsch, "Comprehensive analysis of alternative splicing across tumors from 8,705 patients," *Cancer Cell*, vol. 34, pp. 211–224.e6, Aug. 2018.

[124] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Res.*, vol. 22, pp. 2008–2017, Oct. 2012.

[125] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, Jan. 2010.

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

[127] J. Wang, Y. Pan, S. Shen, L. Lin, and Y. Xing, "rMATS-DVR: rMATS discovery of differential variants in RNA," *Bioinformatics*, vol. 33, pp. 2216–2217, July 2017.

[128] G. P. Alamancos, A. Pagès, J. L. Trincado, N. Bellora, and E. Eyras, "Leveraging transcript quantification for fast computation of alternative splicing profiles," *RNA*, vol. 21, pp. 1521–1531, Sept. 2015.

[129] N. C. Institute and National Cancer Institute, "Ensembl," *Definitions*, 2020.

[130] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr, *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.* Princeton Univ. Press, 1949.

[131] D. Malioutov, T. Chen, E. Airoldi, J. Jaffe, B. Budnik, and N. Slavov, "Quantifying homologous proteins and proteoforms," *Mol. Cell. Proteomics*, vol. 18, pp. 162–168, Jan. 2019.

[132] S. Vasaikar, C. Huang, X. Wang, V. A. Petyuk, S. R. Savage, B. Wen, Y. Dou, Y. Zhang, Z. Shi, O. A. Arshad, M. A. Gritsenko, L. J. Zimmerman, J. E. McDermott, T. R. Clauss, R. J. Moore, R. Zhao, M. E. Monroe, Y.-T. Wang, M. C. Chambers, R. J. C. Slebos, K. S. Lau, Q. Mo, L. Ding, M. Ellis, M. Thiagarajan, C. R. Kinsinger, H. Rodriguez, R. D. Smith, K. D. Rodland, D. C. Liebler, T. Liu, B. Zhang, and Clinical Proteomic Tumor Analysis Consortium, "Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities," *Cell*, vol. 177, pp. 1035–1049.e19, May 2019.

[133] C. Wichmann, F. Meier, S. Virreira Winter, A.-D. Brunner, J. Cox, and M. Mann, "MaxQuant.Live enables global targeting of more than 25,000 peptides," *Mol. Cell. Proteomics*, vol. 18, pp. 982–994, May 2019.

[134] D. K. Schweppe, J. K. Eng, Q. Yu, D. Bailey, R. Rad, J. Navarrete-Perea, E. L. Huttlin, B. K. Erickson, J. A. Paulo, and S. P. Gygi, "Full-Featured, Real-Time database searching platform enables fast and accurate multiplexed quantitative proteomics," *J. Proteome Res.*, vol. 19, pp. 2026–2034, May 2020.

[135] J. Hu, A. L. Ho, L. Yuan, B. Hu, S. Hua, S. S. Hwang, J. Zhang, T. Hu, H. Zheng, B. Gan, G. Wu, Y. A. Wang, L. Chin, and R. A. DePinho, "From the cover: Neutralization of terminal differentiation in gliomagenesis," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 14520–14527, Sept. 2013.

[136] A. S. Adler, M. L. McCleland, S. Yee, M. Yaylaoglu, S. Hussain, E. Cosino, G. Quinones, Z. Modrusan, S. Seshagiri, E. Torres, V. S. Chopra, B. Haley, Z. Zhang, E. M. Blackwood, M. Singh, M. Junttila, J.-P. Stephan, J. Liu, G. Pau, E. R. Fearon, Z. Jiang, and R. Firestein, "An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth," *Genes Dev.*, vol. 28, pp. 1068–1084, May 2014.

[137] T. L. Dayton, T. Jacks, and M. G. Vander Heiden, "PKM2, cancer metabolism, and the road ahead," *EMBO Rep.*, vol. 17, pp. 1721–1730, Dec. 2016.

[138] H. R. Christofk, M. G. Vander Heiden, M. H. Harris, A. Ramanathan, R. E. Gerszten, R. Wei, M. D. Fleming, S. L. Schreiber, and L. C. Cantley, "The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth," *Nature*, vol. 452, no. 7184, pp. 230–233, 2008.

[139] J. D. O'Connell, J. A. Paulo, J. J. O'Brien, and S. P. Gygi, "Proteome-Wide evaluation of two common protein quantification methods," *Journal of Proteome Research*, vol. 17, no. 5, pp. 1934–1942, 2018.

[140] T. U. Consortium and The UniProt Consortium, "The universal protein resource (UniProt) 2009," *Nucleic Acids Research*, vol. 37, no. Database, pp. D169–D174, 2009.

[141] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann, "The quantitative and condition-dependent escherichia coli proteome," *Nat. Biotechnol.*, vol. 34, pp. 104–110, Jan. 2016.

[142] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[143] S. Kim and P. A. Pevzner, "MS-GF makes progress towards a universal database search tool for proteomics," *Nature Communications*, vol. 5, no. 1, 2014.

[144] M. E. Monroe, J. L. Shaw, D. S. Daly, J. N. Adkins, and R. D. Smith, "MASIC: A software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC–MS(/MS) features," *Computational Biology and Chemistry*, vol. 32, no. 3, pp. 215–217, 2008.

[145] M. M. Savitski, T. Mathieson, N. Zinn, G. Sweetman, C. Doce, I. Becher, F. Pachl, B. Kuster, and M. Bantscheff, "Measuring and managing ratio compression for accurate iTRAQ/TMT quantification," *J. Proteome Res.*, vol. 12, pp. 3586–3598, Aug. 2013.

[146] A. Zien, T. Aigner, R. Zimmer, and T. Lengauer, "Centralization: a new method for the normalization of gene expression data," *Bioinformatics*, vol. 17, no. suppl_1, pp. S323–S331, 2001.

[147] S. Tiberi and M. D. Robinson, "Bandits: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty," *Genome Biology*, vol. 21, no. 1, pp. 1–13, 2020.

[148] E. Berchtold, G. Csaba, A. Hadziahmetovic, M. Gruber, C. Ammar, and R. Zimmer, "Empires: Differential analysis of gene expression and alternative splicing," *bioRxiv*, 2020.

[149] M. J. Ellis, M. Gillette, S. A. Carr, A. G. Paulovich, R. D. Smith, K. K. Rodland, R. R. Townsend, C. Kinsinger, M. Mesri, H. Rodriguez, D. C. Liebler, and Clinical Proteomic Tumor Analysis Consortium (CPTAC), "Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium," *Cancer Discov.*, vol. 3, pp. 1108–1112, Oct. 2013.

[150] S. J. Schink, E. Biselli, C. Ammar, and U. Gerland, "Death rate of e. coli during starvation is set by maintenance cost and biomass recycling," *Cell systems*, vol. 9, no. 1, pp. 64–73, 2019.

[151] T. Fung and N. Kwong, "Residual Glycogen Metabolism in Escherichia coli is Specific to the Limiting Macronutrient and Varies During Stationary Phase," vol. 17, p. 5, 2013.

[152] R. Hengge-Aronis and D. Fischer, "Identification and molecular analysis of glgS, a novel growth-phase-regulated and rpoS-dependent gene involved in glycogen synthesis in Escherichia coli," *Molecular Microbiology*, vol. 6, no. 14, pp. 1877–1886, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2958.1992.tb01360.x.

[153] R. E. Strange, "Bacterial "Glycogen" and Survival," *Nature*, vol. 220, pp. 606–607, Nov. 1968. Number: 5167 Publisher: Nature Publishing Group.

[154] S. E. Finkel, "Long-term survival during stationary phase: evolution and the GASP phenotype," *Nature Reviews Microbiology*, vol. 4, pp. 113–120, Feb. 2006.

[155] S. J. Schink, E. Biselli, C. Ammar, and U. Gerland, "Death Rate of E. coli during Starvation Is Set by Maintenance Cost and Biomass Recycling," *Cell Systems*, vol. 9, pp. 64–73.e3, July 2019.

[156] M. M. Zambrano and R. Kolter, "GASPing for Life in Stationary Phase," *Cell*, vol. 86, pp. 181–184, July 1996. Publisher: Elsevier.

[157] S. J. Pirt and C. N. Hinshelwood, "The maintenance energy of bacteria in growing cultures," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 163, pp. 224–231, Oct. 1965. Publisher: Royal Society.

[158] E. Biselli, S. Schink, and U. Gerland, "Slower growth of e. coli leads to longer survival in carbon starvation due to a decrease of the maintenance rate.," *Molecular Systems Biology (in press)*, 2020.

[159] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson, and T. Hwa, "Overflow metabolism in Escherichia coli results from efficient proteome allocation," *Nature*, vol. 528, pp. 99–104, Dec. 2015. Number: 7580 Publisher: Nature Publishing Group.

[160] S. Hui, J. M. Silverman, S. S. Chen, D. W. Erickson, M. Basan, J. Wang, T. Hwa, and J. R. Williamson, "Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria," *Molecular Systems Biology*, vol. 11, p. 784, Feb. 2015. Publisher: John Wiley & Sons, Ltd.

[161] C. You, H. Okano, S. Hui, Z. Zhang, M. Kim, C. W. Gunderson, Y.-P. Wang, P. Lenz, D. Yan, and T. Hwa, "Coordination of bacterial proteome with metabolism by cAMP signalling," *Nature*, vol. 500, pp. 301–306, Aug. 2013.

[162] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, "Interdependence of Cell Growth and Gene Expression: Origins and Consequences," *Science*, vol. 330, pp. 1099–1102, Nov. 2010.

[163] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann, "The quantitative and condition-dependent Escherichia coli proteome," *Nature Biotechnology*, vol. 34, pp. 104–110, Jan. 2016.

[164] J. R. Houser, C. Barnhart, D. R. Boutz, S. M. Carroll, A. Dasgupta, J. K. Michener, B. D. Needham, O. Papoulas, V. Sridhara, D. K. Sydykova, C. J. Marx, M. S. Trent, J. E. Barrick, E. M. Marcotte, and C. O. Wilke, "Controlled Measurement and Comparative Analysis of Cellular Components in E. coli Reveals Broad Regulatory Changes in Response to Glucose Starvation," *PLoS Computational Biology*, vol. 11, Aug. 2015.

[165] S. Dukan and T. Nyström, "Oxidative stress defense and deterioration of growth-arrestedescherichia coli cells," *Journal of Biological Chemistry*, vol. 274, no. 37, pp. 26027–26032, 1999.

[166] H. I. Zgurskaya, C. A. Lopez, and S. Gnanakaran, "Permeability barrier of gram-negative cell envelopes and approaches to bypass it," *ACS infectious diseases*, vol. 1, no. 11, pp. 512–522, 2015.

[167] E. Schneck, T. Schubert, O. V. Konovalov, B. E. Quinn, T. Gutsmann, K. Brandenburg, R. G. Oliveira, D. A. Pink, and M. Tanaka, "Quantitative determination of ion distributions in bacterial lipopolysaccharide membranes by grazing-incidence x-ray fluorescence," *Proceedings of the National Academy of Sciences*, vol. 107, no. 20, pp. 9147–9151, 2010.

[168] L. A. Clifton, M. W. Skoda, A. P. Le Brun, F. Ciesielski, I. Kuzmenko, S. A. Holt, and J. H. Lakey, "Effect of divalent cation removal on the structure of gram-negative bacterial outer membrane models," *Langmuir*, vol. 31, no. 1, pp. 404–412, 2015.

[169] C. Ammar, M. Gruber, G. Csaba, and R. Zimmer, "Ms-empire utilizes peptide-level noise distributions for ultra-sensitive detection of differentially expressed proteins," *Molecular & Cellular Proteomics*, vol. 18, no. 9, pp. 1880–1892, 2019.

[170] G.-W. Li, D. Burkhardt, C. Gross, and J. S. Weissman, "Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources," *Cell*, vol. 157, pp. 624–635, Apr. 2014.

[171] S. Vasaikar, C. Huang, X. Wang, V. A. Petyuk, S. R. Savage, B. Wen, Y. Dou, Y. Zhang, Z. Shi, O. A. Arshad, *et al.*, "Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities," *Cell*, vol. 177, no. 4, pp. 1035–1049, 2019.

[172] S. Degroeve and L. Martens, "Ms2pip: a tool for ms/ms peak intensity prediction," *Bioinformatics*, vol. 29, no. 24, pp. 3199–3203, 2013.

[173] X.-X. Zhou, W.-F. Zeng, H. Chi, C. Luo, C. Liu, J. Zhan, S.-M. He, and Z. Zhang, "pdeep: predicting ms/ms spectra of peptides with deep learning," *Analytical chemistry*, vol. 89, no. 23, pp. 12690–12697, 2017.

[174] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghe, A. Huhmer, *et al.*, "Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning," *Nature methods*, vol. 16, no. 6, pp. 509–518, 2019.

[175] S. Tiwary, R. Levy, P. Gutenbrunner, F. S. Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic, and J. Cox, "High-quality ms/ms spectrum prediction for data-dependent and data-independent acquisition data analysis," *Nature methods*, vol. 16, no. 6, pp. 519–525, 2019.

[176] F. Meier, A.-D. Brunner, S. Koch, H. Koch, M. Lubeck, M. Krause, N. Goedecke, J. Decker, T. Kosinski, M. A. Park, *et al.*, "Online parallel accumulation–serial fragmentation (pasef) with a novel trapped ion mobility mass spectrometer," *Molecular & Cellular Proteomics*, vol. 17, no. 12, pp. 2534–2545, 2018.

[177] W. Bittremieux, D. May, J. A. Bilmes, and W. S. Noble, "Deep neural network embedding for efficient repository-scale analysis of hundreds of millions of mass spectra," in *MOD am 09:30, Proceedings of the 68th ASMS Conference on Mass Spectrometry and Allied Topics, Online Reboot, June 1 - Aug 31*, 2020.

[178] X. Shen, S. Shen, J. Li, Q. Hu, L. Nie, C. Tu, X. Wang, D. J. Poulsen, B. C. Orsburn, J. Wang, *et al.*, "Ionstar enables high-precision, low-missing-data proteomics quantification in large biological cohorts," *Proceedings of the National Academy of Sciences*, vol. 115, no. 21, pp. E4767–E4776, 2018.

[179] F. Meier, P. E. Geyer, S. Virreira Winter, J. Cox, and M. Mann, "Boxcar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes," *Nature methods*, vol. 15, pp. 440–448, 2018.

[180] S. V. Winter, F. Meier, C. Wichmann, J. Cox, M. Mann, and F. Meissner, "Easi-tag enables accurate multiplexed and interference-free ms2-based proteome quantification," *Nature methods*, vol. 15, no. 7, pp. 527–530, 2018.

[181] A. Thompson, N. Wölmer, S. Koncarevic, S. Selzer, G. Böhm, H. Legner, P. Schmid, S. Kienle, P. Penning, C. Höhle, *et al.*, "Tmtpro: Design, synthesis, and initial evaluation of a proline-based isobaric 16-plex tandem mass tag reagent set," *Analytical Chemistry*, vol. 91, no. 24, pp. 15941–15950, 2019.

[182] B. K. Erickson, J. Mintseris, D. K. Schweppe, J. Navarrete-Perea, A. R. Erickson, D. P. Nusinow, J. A. Paulo, and S. P. Gygi, "Active instrument engagement combined with a real-time database search for improved performance of sample multiplexing workflows," *Journal of proteome research*, vol. 18, no. 3, pp. 1299–1306, 2019.

[183] C. Wichmann, F. Meier, S. V. Winter, A.-D. Brunner, J. Cox, and M. Mann, "Maxquant. live enables global targeting of more than 25,000 peptides," *Molecular & Cellular Proteomics*, vol. 18, no. 5, pp. 982–994, 2019.

[184] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox, "The perseus computational platform for comprehensive analysis of (prote) omics data," *Nature methods*, vol. 13, no. 9, pp. 731–740, 2016.

[185] L.-C. Chang, H.-M. Lin, E. Sibille, and G. C. Tseng, "Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline," *BMC bioinformatics*, vol. 14, no. 1, p. 368, 2013.

[186] P. Moulos and P. Hatzis, "Systematic integration of rna-seq statistical algorithms for accurate detection of differential gene expression patterns," *Nucleic acids research*, vol. 43, no. 4, pp. e25–e25, 2015.

[187] X. Wang, D. D. Kang, K. Shen, C. Song, S. Lu, L.-C. Chang, S. G. Liao, Z. Huo, S. Tang, Y. Ding, *et al.*, "An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection," *Bioinformatics*, vol. 28, no. 19, pp. 2534–2536, 2012.

[188] J. Xia, E. E. Gill, and R. E. Hancock, "Networkanalyst for statistical, visual and network-based meta-analysis of gene expression data," *Nature protocols*, vol. 10, no. 6, pp. 823–844, 2015.

[189] F. Hong, R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser, and J. Chory, "Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis," *Bioinformatics*, vol. 22, no. 22, pp. 2825–2827, 2006.

[190] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, and M. Mann, "Plasma proteome profiling to assess human health and disease," *Cell systems*, vol. 2, no. 3, pp. 185–195, 2016.

[191] M. A. Gillette, S. Satpathy, S. Cao, S. M. Dhanasekaran, S. V. Vasaikar, K. Krug, F. Petralia, Y. Li, W.-W. Liang, B. Reva, *et al.*, "Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma," *Cell*, vol. 182, no. 1, pp. 200–225, 2020.

[192] J. M. Bader, P. E. Geyer, J. B. Müller, M. T. Strauss, M. Koch, F. Leypoldt, P. Koertvelyessy, D. Bittner, C. G. Schipke, E. I. Incesoy, *et al.*, "Proteome profiling in cerebrospinal fluid reveals novel biomarkers of alzheimer's disease," *Molecular systems biology*, vol. 16, no. 6, p. e9356, 2020.

[193] T. Scheidt, O. Alka, H. Gonczarowska-Jorge, W. Gruber, F. Rathje, M. Dell'Aica, M. Rurik, O. Kohlbacher, R. P. Zahedi, F. Aberger, *et al.*, "Phosphoproteomics of short-term hedgehog signaling in human medulloblastoma cells," *Cell Communication and Signaling*, vol. 18, no. 1, pp. 1–18, 2020.

[194] E. Soupene, W. C. Van Heeswijk, J. Plumbridge, V. Stewart, D. Bertenthal, H. Lee, G. Prasad, O. Paliy, P. Charernnoppakul, and S. Kustu, "Physiological studies of escherichia coli strain mg1655: growth defects and apparent cross-regulation of gene expression," *Journal of Bacteriology*, vol. 185, no. 18, pp. 5611–5626, 2003.

[195] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, vol. 2, Jan. 2006.

[196] L. N. Csonka, T. P. Ikeda, S. A. Fletcher, and S. Kustu, "The accumulation of glutamate is necessary for optimal growth of salmonella typhimurium in media of high osmolality but not induction of the prou operon.," *Journal of Bacteriology*, vol. 176, no. 20, pp. 6324–6333, 1994.

[197] M. Basan, S. Hui, and J. R. Williamson, "ArcA overexpression induces fermentation and results in enhanced growth rates of E. coli," *Scientific Reports*, vol. 7, p. 11866, Dec. 2017.

Acknowledgements

I feel very thankful that I have had great mentors in my professional life and a great family, great friends, and a great girlfriend in my personal life. They all have brought me here.

In particular, I want to thank:

My mentors, colleagues and collaboration partners:

My supervisor Ralf Zimmer for giving me the chance to pursue a Ph.D. in his lab, for always being available for any question and for never sparing an effort in order to help me.
Gergely Csaba, for the great times in the lab and on the road and for the tremendous amount of things I learned from you.
Markus Joppich for being a reliable comrade and office-mate over the years.
Markus Gruber for the great work together on MS-EmpiRe and for being a kind and helpful colleague.
Evi Berchtold for her help getting me started and for being patient with me.
Michael Kluge and Alexander Grün for being great sports and colleagues.
Volker Heun for the jokes and Caroline Friedel and Franziska Schneider for the entertaining conversations and Frank Steiner for the technical support.
Severin Schink for the amazing times working and playing together in Boston and during my master thesis. I was very lucky to meet you.
Markus Basan for welcoming me in his lab, for the entertaining and brilliant discussions over lunch and for always giving support and advice.
Werner Mewes for helping and advising me wherever he could and for doing the missionary work for Bioinformatics.
Axel Imhof and Andreas Schmidt for the collaboration on the MCIP paper and for sharing their expertise.
Ulrich Gerland, for paving the way towards my Ph.D. by being an enthusiastic and encouraging master thesis advisor.
Katia Parodi for guiding my first steps in science, for trusting me, challenging me, and giving me great feedback. To Kathrin Frey for her great advice and supervision.
Maximilian Strauss, for kindly helping me take the next step.
The QBM graduate school and staff for support and funding.

Prof. Oliver Kohlbacher, for taking the time and effort to review my thesis.

My family and friends:

My mother Sibylle Schardey, who gave up so much for us, for her guidance through life.
My beloved and dauntingly intelligent little brother Leander, for always being by my side and sharing his life with me.
My longest companion Aliyah, the kindest person I know, for doing the extraordinary as a matter of course.
My little sister Naila, for keeping an eye on me.
My father Yousif, for his support and advice.
Oma Lily, for being my greatest supporter.
My uncles Hinrich and Martin, for always encouraging me and setting the bar high.
My friend Peter Wollförster, for the enthusiastic discussions and interesting experiments.
My tallest little brother Maxi, for being a steady companion in the exploration of life.
My most loyal friend Alex, for the countless adventures and the unceasing support in any situation.
My longest friend Johannes, for always doing the right thing.
My hilariously creative friend Max, for making my life more colorful.
My great friend Christoph, for making me feel like home whenever we meet.
The honorable Jonas, for being a hilarious friend and great roommate.
Jaani, Fred and Marc, für den Kult.
Annick, for welcoming me into her home in Boston.
Ludwig, Duffy and Simeon for the great times in New York.
Lastly, I want to thank Verena for understanding me, never ceasing to encourage me, and for seeing what's good. I could never have hoped for more.