From the Department of **M**edical **B**iochemistry and **B**iophysics
Karolinska **I**nstitutet, Stockholm, **Sweden**

ADVANCING BIOINFORMATICS METHODS FOR IN-DEPTH PROTEOME ANALYSIS BASED ON HIGH-RESOLUTION MASS SPECTROMETRY

ZHANG, BO
(张 博)

Stockholm 2017

Published by Karolinska Institutet
Printed by E-Print AB

ISBN 987-91-7676-617-0

ADVANCING BIOINFORMATICS METHODS FOR IN-DEPTH PROTEOME ANALYSIS BASED ON HIGH-RESOLUTION MASS SPECTROMETRY

THESIS FOR DOCTORAL DEGREE (PH.D.)

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i **Stora Seminarierummet (A3:311), Scheeles väg 2, Campus Solna**

May 12th 2017, Friday, 14:00

By

Bo Zhang

Principal Supervisor:
Roman A. Zubarev, Professor
Karolinska Institutet
Department of Medical Biochemistry & Biophysics
Division of Physiological Chemistry I

Co-supervisors:
Erik L.L. Sonnhammer, Professor
Stockholm University
Department of Biochemistry and Biophysics
Science for Life Laboratory

Lukas Käll, Associate Professor
KTH - Royal Institute of Technology
School of Biotechnology
Science for Life Laboratory

Opponent:
Bernhard Küster, Professor
Technical University of Munich
School of Life Sciences Weihenstephan
Chair of Proteomics and Bioanalytics

Examination Board:
Jonas Bergquist, Professor
Uppsala University
Department of Chemistry - BMC
Division of Analytical Chemistry

Fredrik Levander, Associate Professor
Lund University
Department of Immunotechnology

Qiang Pan-Hammarström, Professor
Karolinska Institutet
Department of Laboratory Medicine
Division of Clinical Immunology

To my parents, who named me "doctor".

献给我的父母，感其取名之恩。

"Persevere in carving, any metal or stone can be engraved."—Xun Kuang

「锲而不舍 金石可镂」— 荀子《劝学》

Abstract

Mass spectrometry-based shotgun proteomics has become one of the essential techniques for comprehensive studies of living systems. Due to the inherent complexity of proteomes and the data, bioinformatics plays a critical role to translate mass spectra into biological information and knowledge. Adapting to the increased availability of high-resolution mass analyzers, computational strategies for processing shotgun proteomics data should have some adjustments to utilize the advantages of modern instruments. This thesis presents five constituent papers to illustrate the methodological advancements for analyzing shotgun proteomics data that are generated from high-resolution mass spectrometry. Paper-I describes the DeMix workflow for protein identification, in which we broke down an old paradigm of tandem mass spectrometry by converting the unwanted co-fragmentation events into an advantage of natural multiplexing. DeMix simplifies the data processing procedure and significantly improves protein identification outcomes. Paper-III describes a label-free extension of the DeMix workflow, termed DeMix-Q, which makes use of the quantitative features of extracted ion chromatograms (XICs) for reliably propagating peptide identifications across LC-MS/MS experiments. DeMix-Q improves the reproducibility of peptide quantification by addressing the missing value problem that is caused by the data-dependent acquisition of MS/MS. Based on the results, the concept of quantification-centered proteomics has been proposed. In the practice of quantification-centered proteomics, a flexible proteome summarizing approach termed Diffacto is described in Paper-V, which utilizes the information about covariation of peptides' abundances to improve the relative quantification of proteins. Diffacto offers automatic quality control to remove inconsistent and unreliable quantitative data on peptides. The combination of a weighted summarizing method and an efficient FDR estimation provides significant enhancement of data utility for large-scale comparative proteomics. In Paper-II, an improved pI estimation method has been introduced to the novel device for sample fractionation based on isoelectric focusing technique. In Paper-IV and V, the applications of peptide *de novo* sequencing have been demonstrated for analyzing complex proteomes in the absence of reference databases.

List of Scientific Publications

Papers included in this thesis:

I. Zhang, B., Pirmoradian, M., Chernobrovkin, A., and Zubarev, R. A. (2014) DeMix Workflow for Efficient Identification of Cofragmented Peptides in High Resolution Data-dependent Tandem Mass Spectrometry. *Mol. Cell. Proteomics* 13, 11–17

II. Pirmoradian, M., Zhang, B., Chingin, K., Astorga-Wells, J., and Zubarev, R. A. (2014) Membrane-assisted isoelectric focusing device as a micropreparative fractionator for two-dimensional shotgun proteomics. *Anal. Chem.* 86, 5728–5732

III. Zhang, B., Käll, L., and Zubarev, R. A. (2016) DeMix-Q: Quantification-centered Data Processing Workflow. *Mol. Cell. Proteomics* 15, 1467–1478

IV. Lundström, S. L., Zhang, B., Rutishauser, D., Aarsland, D., and Zubarev, R. A. (2017) SpotLight Proteomics: uncovering the hidden blood proteome improves diagnostic power of proteomics. *Sci. Reports.* 7, 41929

V. Zhang, B., Pirmoradian, M., Zubarev, R. A., and Käll, L. (2017) Covariation of Peptide Abundances Accurately Reflects Protein Concentration Differences. *Mol. Cell. Proteomics*, doi: 10.1074/mcp.O117.067728

Other publications:

i. Pirmoradian, M., Budamgunta, H., Chingin, K., Zhang, B., Astorga-Wells, J., and Zubarev, R. A. (2013) Rapid and deep human proteome analysis by single-dimension shotgun proteomics. *Mol. Cell. Proteomics* 12, 3330–8

ii. Barnidge, D. R., Lundström, S. L., Zhang, B., Dasari, S., Murray, D. L., and Zubarev, R. A. (2015) Subset of Kappa and Lambda Germline Sequences Result in Light Chains with a Higher Molecular Mass Phenotype. *J. Proteome Res.* 14, 5283–5290

iii. Thiagarajan, D., Frostegård, A. G., Singh, S., Rahman, M., Liu, A., Vikström, M., Leander, K., Gigante, B., Hellenius, M.-L., Zhang, B., Zubarev, R. A., de Faire, U., Lundström, S. L., and Frostegård, J. (2016) Human IgM Antibodies to Malondialdehyde Conjugated with Albumin Are Negatively Associated with Cardiovascular Disease Among 60-Year-Olds. *J. Am. Heart Assoc.* 5 (12), e004415

CONTENTS

Abbreviations

AA	amino acid residue
AMT	accurate-mass-and-time
AVONA	analysis of variance
BLAST	Basic Local Alignment Search Tool
BSA	bovine serum albumin
CDR	complementary determining region
CID	collision-induced dissociation
cIEF	capillary isoelectric focusing
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CV	coefficient of variation
Da	dalton, the unified atomic mass unit
DDA	data-dependent acquisition
DIA	data-independent acquisition
ECD	electron-capture dissociation
ESI	electrospray ionization
ETD	electron transfer dissociation
FARMS	Factor Analysis for Robust Microarray Summarization
FDR	false discovery rate
FQR	false quantification rate
FT	Fourier transform
FWHM	full width at half maximum
HCD	higher-energy collisional dissociation
HiRIEF	high-resolution isoelectric focusing
iPRG	the ABRF Proteome Informatics Research Group
iTRAQ	isobaric tags for relative and absolute quantitation
LC	liquid chromatography
LFQ	label-free quantification
m/z	mass-to-charge ratio
MC	Monte Carlo
MJ-cIEF	multijunction capillary isoelectric focusing
MPIB	Max Planck Institute of Biochemistry
MS	mass spectrometry
MS^1	survey spectrum, the first stage of tandem mass spectrometry
MS/MS	tandem mass spectrometry
PCR	polymerase chain reaction
pI	isoelectric point
PIP	peptide identity propagation
PMF	peptide mass fingerprint
ppm	part per million
PQPQ	protein quantification by peptide quality control
PSM	peptide-spectrum match
PTM	post-translational modification
S/N	signal-to-noise ratio
SpC	spectral counting
TMT	Tandem Mass Tag
XIC	extracted ion chromatogram

Chapter One: Background

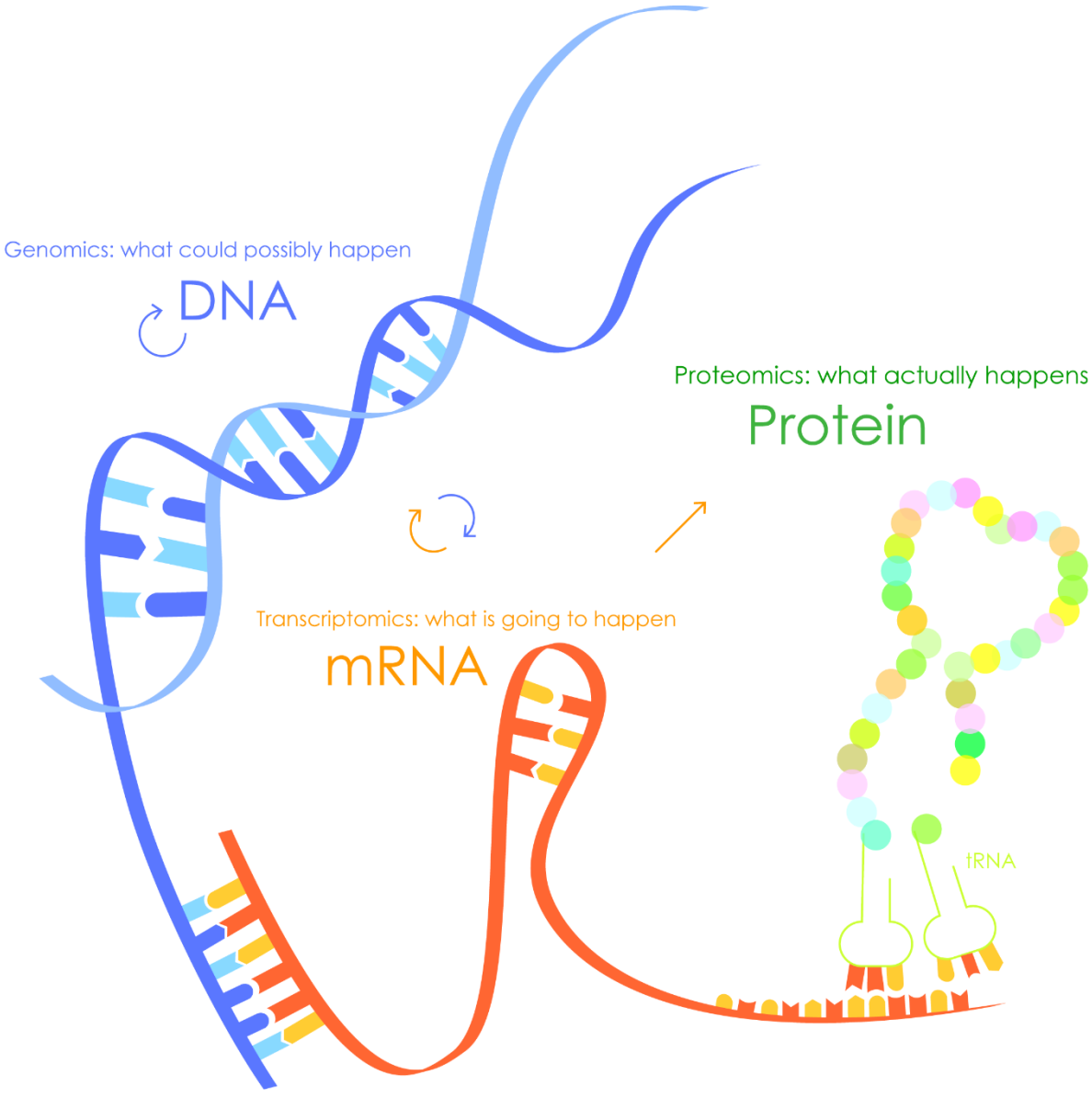

Figure 1.1 | The central dogma of molecular biology.

1.1 Introduction

Comprehensively understanding molecular and cellular functions is the key to systems biology (Kitano, 2002). Moving towards the ultimate goal of systems biology, revolutionary advancements in technology, such as the massive parallel sequencing and high-resolution mass spectrometry, have changed almost every aspect of biomedical research. In light of the central dogma of molecular biology (Crick, 1970), we started to bravely explore the vast territory of systems biology, from DNA through RNA to proteins and beyond.

Proteomics is the system-scale study of proteins– the functional entities involved in almost all the biochemical processes– that is tightly related to systems biology (Weston and Hood, 2004). Compared to genomics and transcriptomics, respectively the studies of DNA and mRNA, proteomics did not benefit much from the massive sequencing technology. The reason is obvious, proteins (chains of amino acids) are fundamentally different from DNA and RNA (chains of nucleotides). Unlike in the processes of transcription (DNA to mRNA) and reverse-transcription (mRNA back to DNA), the information-flow is "irreversibly" encoded by nucleotide triplets (codons) in the process of translation (mRNA to protein). The combination of four nucleobases– A, C, G, and U (or T for DNA)– gives as many as 64 possible triplets, which causes the degeneracy of genetic code that exhibits redundancy for encoding the 20 amino acids. Besides, there is no evidence so far indicating an equivalent of the polymerase chain reaction (PCR) for reproducing proteins after the synthesis.

Many believe that the protein abundances can be determined, to some extent, by the expression level of mRNAs (Gygi et al., 1999, Liu et al., 2016). This claim could be supported by an apparent correlation between the profiles of transcriptome and the proteome, if given a steady cellular state and the factors reflecting protein translation and degradation (Wilhelm et al., 2014, Liu et al., 2016). However, proteomes are the snapshots of living systems. Direct study of the performers of biological functions should give a higher accuracy of elucidating cellular mechanisms, compared to a study via the correlating entities (Wang et al., 2017). A complex system, e.g., a mammalian cell, often requires many copies of proteins to maintain the biological functions. Fortunately, to proteins in the scope of a cell, the relative scale of "many" is several magnitudes larger than that to the genes and transcripts. According to previous knowledge (Schwanhausser et al., 2011), the median protein abundance in a

mammalian cell is 50,000 copies, in contrast to the abundances of mRNAs (median 17 copies) and genes (one or two copies). Therefore, when the protein samples are collected, the most difficult job of amplifying the analytes has already been done. However, the analytical challenge is to capture a clear snapshot of the proteome.

Mass spectrometry-based proteomics

Lacking the equivalent of PCR, the technique of "*sequencing by synthesis*" is not feasible for analyzing proteins, and the traditional protein sequencing method– *Edman degradation*– could be extremely slow and costly for a system-wide analysis. Therefore, the contemporary strategy for proteomics is "*divide and conquer*", which strives, by all means, to purify, isolate, break down, and subsequently measure the analytes.

Multiple sources of information about the properties of proteins could be used to characterize the protein contents in the samples. Among the physiochemical properties of proteins, such as isoelectric point (pI), hydrophobicity, and antibody affinity, **mass** (with the symbol **m**) is perhaps the only intrinsic property that can be precisely measured. Importantly, regardless of the structure, the mass of a protein molecule can be determined by summing up the masses of all the elements of all its chemical components– amino acids, terminal groups, and modifications.

Mass spectrometry (MS) is the analytic technique for measuring the masses (m), or more precisely speaking, the mass-to-charge ratios of the charged particles (Thomson, 1913). The principle of MS can be described by the Lorentz force law and the Newton's second law of motion:

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})$$

$$\boldsymbol{F} = \mathrm{m} \cdot \mathbf{a}$$

Equating the two expressions gives:

$$\left(\frac{m}{q}\right) \cdot \mathbf{a} = \boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}$$

This equation describes the dynamics of charged particles in an electromagnetic field in a vacuum (***E*****:** electric field, $\boldsymbol{v} \times \boldsymbol{B}$**:** vector cross product of the velocity and the magnetic field), where the change of velocity (**a**) determines the mass-to-charge ratio (m/q) of the charged particle.

The standard unit of the physical quantity of m/q is kilograms per coulomb (kg/C). To simplify the interpretation of data in mass spectra, the common notation for the mass-to-charge ratio is *m/z*. In this notation, mass (m) is related to the unified atomic mass unit, dalton (Da), which is defined as one twelfth of the mass of a neutral carbon-12 (~$1.6605402 \times 10^{-27}$ kg); and the charge (z) is related to the elementary charge (e), which is the electric charge of a proton (~$1.60217662 \times 10^{-19}$ C). By this definition, the atomic unit of *m/z* is dimensionless, which is the ratio between the mass number and the charge number.

Although proteomics relies on the different technology than genomics and transcriptomics, the field is not without revolutionary advancements. (Aebersold and Mann, 2003, Aebersold and Mann, 2016). In fact, a comparably fast development of high-resolution mass spectrometers, with the increased accessibility, has already transformed proteomics. For instance, the Orbitrap mass analyzer (Makarov, 2000) has enabled high-resolution Fourier transform MS on a benchtop, which has been well adopted as a powerful technique in various protein-related biomedical research (Michalski et al., 2011b, Zubarev and Makarov, 2013). Thanks to the Orbitrap, deep profiling of complex proteomes became not only feasible (Thakur et al., 2011, Nagaraj et al., 2012) but also time-efficient (Richards et al., 2015), As a result, the comprehensive maps of the human proteome were eventually drafted (Kim et al., 2014, Wilhelm et al., 2014), more than a decade after the availability of the human genome sequence (Lander et al., 2001, Venter et al., 2001).

Separation prior to MS

Complementary techniques are almost always applied in conjunction with MS, to provide extra dimensions for physical separation of proteins. Electrophoresis and liquid chromatography (LC) are the most commonly applied techniques.

Electrophoresis separates proteins by moving the charged molecules in an electric field through a matrix or solution. The electrophoretic mobility depends on the net charge and the molecular weight of the analyte. Because proteins are zwitterionic molecules that may carry both positive and negative charges, each protein has an isoelectric point (pI). Therefore, isoelectric focusing (IEF), as one type of electrophoresis, provides a pH gradient to variate the net charges of analytes. When the pH of surrounding environment equals the pI of a particular protein molecule, the net charge of the zwitterionic molecule becomes zero (neutral), and consequently, the protein becomes stationary (focused) in the electric field.

LC, or more specifically, the reversed-phase LC separates molecules based on their hydrophobic characters, by dissolving and eluting samples from a non-polar stationary phase using a polar mobile phase that has a concentration gradient that changes the molecules' affinities to the stationary phase. A high-performance LC uses a pump to force the sample solved in the mobile phase through the stationary phase. The instrumental setting of reversed-phase (ultra)high-performance liquid chromatography coupled with mass spectrometry, (U)HPLC-MS for short, has become the most common technology used for in-depth proteomic analysis.

Tandem layouts and combinations of these techniques can provide a multidimensional separation for approaching the rather comprehensive analysis of complex proteomes (Zhang et al., 2010). For example, the MudPIT approach (Link et al., 1999) combines two types of LC (ion-exchange and reversed-phase, IEX/RP-LC) and then connects to MS. For the first time, it provided a "comprehensive" analysis of over 100 proteins in a single experiment. With modern instruments, compared to its earlier implementations, such a two-dimensional LC-MS system has become a hundred times more powerful that enabled profiling of over 10,000 proteins (Geiger et al., 2012). More recently, the HiRIEF approach, which links the state-of-the-art IEF, LC, and MS technologies, has demonstrated the deep proteome profiling of mammalian cells by quantitatively analyzing over 13,000 human proteins in a single experiment (Branca et al., 2014).

Tandem mass spectrometry

Although the mass of a protein could be precisely measured by MS, it is far from a unique identity of the protein and is not sufficient to infer the corresponding sequence of the protein. In a protein sequence, the number of possible combinations of 20 amino acids is enormous, which could form as many as 20^n different sequences depending on the length (n). Consequently, based on a single value of *m/z*, it is rather difficult to deduce the elementary composition of large proteins and is almost impossible to know the order of the amino acids. However, the problem can also be solved in the manner of "*divide and conquer*" by breaking down the ion of the analyte to preferably two pieces and subsequently measuring the *m/z* of its fragments.

The instrumental setting for this method is called tandem mass spectrometry (MS/MS or MS^n) (McLafferty, 1981). MS/MS generates hierarchical mass information on the precursor-fragments relations. MS/MS has four basic steps: isolation, activation, fragmentation, and

detection. The first mass analyzer (e.g., a quadrupole) isolates the target precursors from other ions that have different *m/z*. After the isolation, precursor ions are stored in a chamber (e.g., an ion trap) where certain energy will be given to activate the precursors. Activation will eventually break down the amino acid backbones. For fragmenting the peptide ions, the most commonly applied method of activation is collision with neutral gas molecules (CID/HCD) (Mcluckey, 1992, Wells and McLuckey, 2005); and a less common method is transferring electrons to the positively charged molecules (ECD/ETD) (Zubarev et al., 1998, Syka et al., 2004). Breakages on the amino acid chains will generate fragment ions (m_f^+) and/or neutral particles (m_n) that, in combination, compose the masses from the precursors (m_p^+):

$$m_p^+ = m_f^+ + m_n$$

Depending on the locations of backbone breakage, the fragment ions could form a ladder of masses that tells the composition and the order of amino acids, which is the sequence of the precursor. Detection of ions can be done in a high-resolution mass analyzer, (e.g., Orbitrap) for a high mass accuracy at the level of ppm. However, the mechanisms of fragmentation are not yet fully understood, and the backbone breakages are only empirically predictable (Degroeve and Martens, 2013). When the fragmentation does not follow the theoretical or empirical rules, it could be sometimes difficult to interpret the observed spectra (Palzs and Suhal, 2005, Zubarev et al., 2008).

Shotgun proteomics

Of the approximately 20,000 protein coding genes embedded in the human genome, the canonical human protein sequences have a median length of over 400 amino acids, and a median mass of over 45 kDa (UniProt, 2015). Ideally, proteins should be analyzed in their native forms that carry the full information about their sequences (Catherman et al., 2014). However, proteins often carry a wide variety of posttranslational modifications (PTMs). The combination of PTMs, isoforms, and mutations will result in an exponentially increased number of the forms of proteins (Cox and Mann, 2011), and many of them will likely have distinctive masses. Due to technical limitations, a "top-down" approach for direct measurement of all these "proteoforms" is rather difficult (Kelleher, 2004). Therefore, shotgun proteomics, an analog to the shotgun sequencing of genomes, has become the formidable technique in large-scale studies of complex proteomes, which is also known as the "bottom-up" approach.

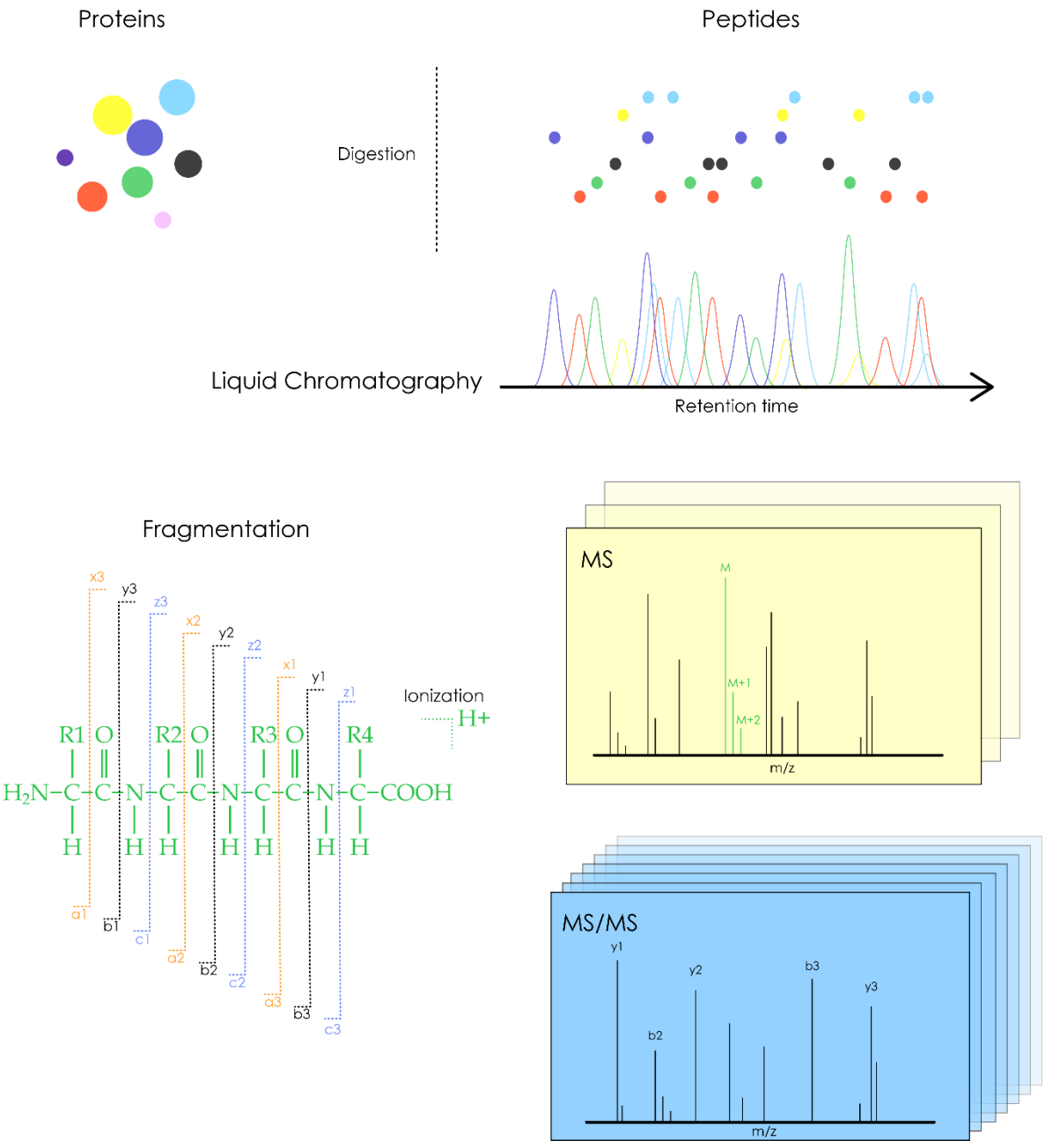

Figure 1.2 | Shotgun proteomics and LC-MS/MS. Purified proteins are digested into peptides using proteolytic enzymes (typically trypsin). Proteolytic peptides are separated by liquid chromatography (LC) according to hydrophobicity. Peptides eluted from LC are ionized (protonated) and injected into a mass spectrometer. The *m/z* values of peptide precursor ions are recorded in the survey MS spectra. After isolation, activation, and fragmentation, the *m/z* values of fragment ions are recorded in MS/MS spectra. The nomenclature for labeling the fragment ions was suggested by (Roepstorff and Fohlman, 1984) and modified by (Biemann, 1988). Collisional dissociation (CID and HCD) mainly produces b/y species of fragment ions.

Shotgun proteomics involves proteolytic digestion to cleave the intact proteins into proteolytic peptides. The digestion makes the analytes less heterogeneous and more friendly to MS. Trypsin exclusively cleaves the peptide bonds at the C-termini to lysine (Lys/K) or arginine (Arg/R) residue (Olsen et al., 2004), which is the protease of choice for a proteome-scale digestion. Each of these two amino acids has an approximately 5% occupancy in human proteins. By tryptic digestion, the resultant proteolytic peptides have an average length around ten amino acids. Compared to that of intact proteins, the relative size of tryptic peptides is more amenable to MS.

In addition, a full-tryptic digestion will yield peptides (most of which) containing only one lysine or arginine residue located at the C-termini of the sequences. According to the theoretical "mobile proton" model of fragmentation, protonated tryptic peptides are friendly to the collisional fragmentation (CID and HCD), because they are less likely to "sequester" the attached protons by the basic side chains (Palzs and Suhal, 2005, Swaney et al., 2010).

Data acquisition

Proteolytic digestion generates tens of peptides per protein, which inevitably causes new challenges for an MS-based analysis. First of all, MS is currently not fast enough to analyze all peptides in a reasonable amount of time. For instance, human cells may typically express about 10,000 genes and translate them into proteins. As the median length of human proteins is about 400 AA, one could expect to have on average 40 proteolytic peptides per protein, assuming a full digestion using trypsin. In this regard, even without considering sequence variations and PTMs, one would have at least 400,000 peptides to analyze in a single experiment. Given three hours for a single LC-MS/MS experiment, one needs to have an at least 40 Hz identification rate to get a full coverage of the expressed proteome. Currently, an Orbitrap mass analyzer has a maximum 18 Hz of acquiring high-resolution (15,000 FWHM) spectra; which empirically offers ca. 10 Hz identification rate in best scenarios with 50 to 60% success rate of MS/MS identification.

Secondly, the dynamic range of proteins in mammalian cells spans over seven orders of magnitude (Schwanhausser et al., 2011), but a modern MS could barely reach five orders of magnitude, leaving a "dark corner" of the proteome consisting of mostly the least abundant proteins (Zubarev, 2013). The instrumental requirement for a comprehensive proteome-wide analysis is considerably high, which could be analogous to using one scale to measure the

weights of both a mosquito (~5 mg) and a human adult (~6o kg); or using one ruler to measure both the diameter of a human hair (~80 μm) and the height of the Eiffel Tower (300 m).

In general, proteins with higher abundances give peptide ions that generate stronger signals and are easier to analyze. Thus, a strategy called data-dependent acquisition (DDA) for MS/MS is commonly applied in shotgun proteomics experiments. By the DDA approach, in every cycle of MS analysis, the most abundant precursor ions will be sequentially isolated for MS/MS. However, the DDA approach has a critical drawback– it is not entirely reproducible. That is to say, repeating the experiment with the same sample and the same instrument settings, one would obtain a somewhat different set of peptides derived from the MS/MS spectra (Tabb et al. 2010). The under-sampling problem leaves a serious concern about the reproducibility in proteomics experiments (Domon & Aebersold 2010). In order to avoid the stochasticity introduced by DDA, the targeted data acquisition could be applied, which specifies when to acquire an MS/MS spectrum of a specific precursor ion. The latter approach relies on the prior knowledge and allows for a limited list of targets, which makes it poorly suitable for the proteome-wide discovery. Alternatively, the data-independent acquisition (DIA) has been proposed to sequentially acquire MS/MS spectra for a wider *m/z* range (compared to DDA and targeted approaches) of precursors in an unbiased manner (Gillet et al., 2012, Egertson et al., 2013).

The choice of data acquisition method depends on the purpose of a study, and the balance between the discovery and validation, as well as between sensitivity and reproducibility (Leitner and Aebersold, 2013).

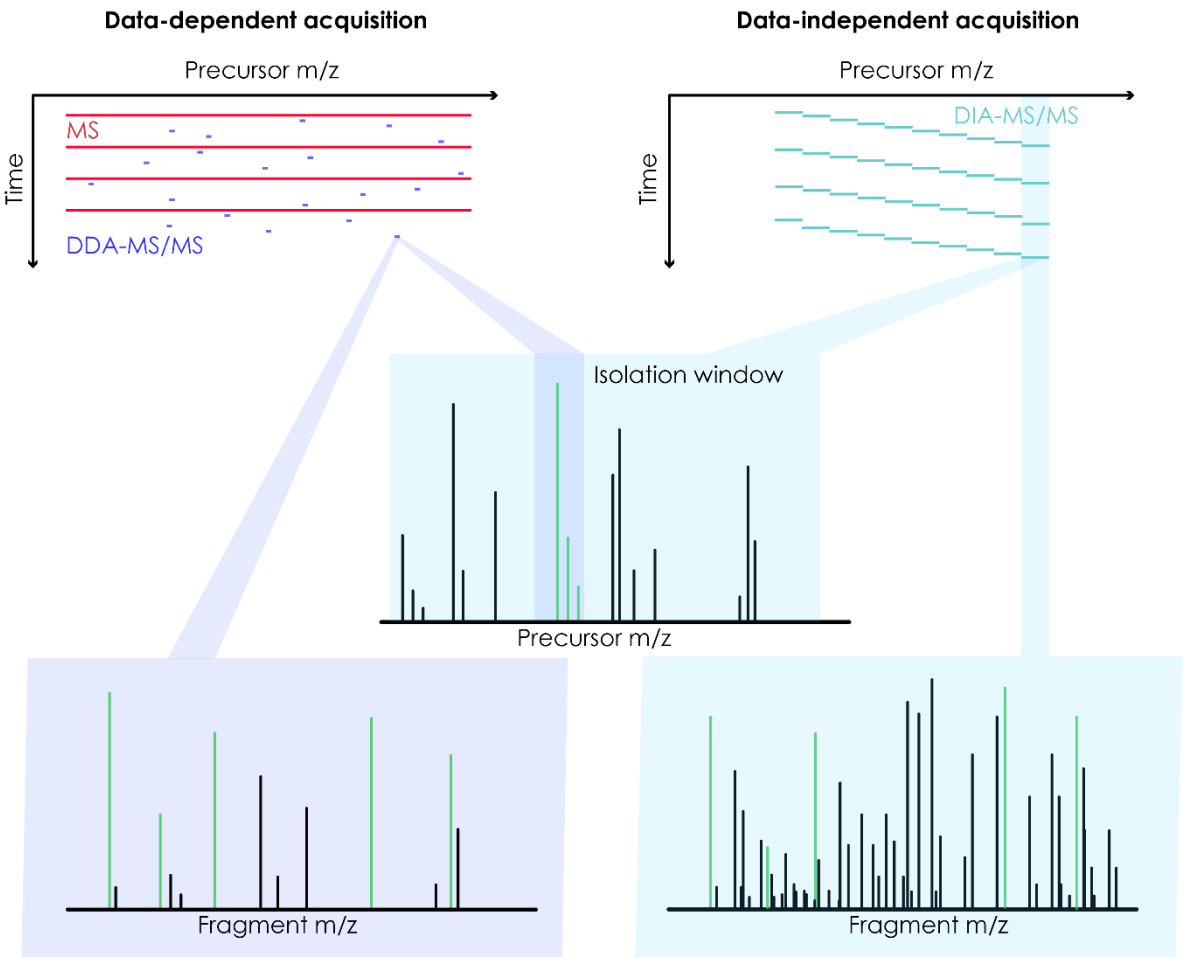

Figure 1.3 | Data-dependent vs. data-independent acquisition of MS/MS spectra. DDA targets the most intensive precursors that are observed in the survey MS. Narrow isolation windows (usually, $m/z < 4.0$) are often used to separate targeted precursor ions for MS/MS. To prevent repetitively analyzing the same targets within a short time, the mechanism of dynamic exclusion applied in DDA introduces a stochastic tendency in the MS/MS data. In contrast, DIA does not target specific precursors and does not necessarily require survey MS for acquiring precursor information. Wide isolation windows (typically, $m/z > 20$) are sequentially used for acquiring MS/MS spectra in an unbiased manner. Consequently, DIA-derived MS/MS spectra are often more complex and noisy than that of DDA, due to co-fragmentation of the mixture of precursor ions.

1.2 Bioinformatics

MS-based shotgun proteomics could be categorized as a high-throughput, data-driven research field (Cox and Mann, 2007, Cox and Mann, 2011), generating huge amounts of data that require highly efficient software for downstream analysis. Proteomics, in this regard, is also an interdisciplinary field, in which, bioinformatics plays a critical role in making sense of the data, and ultimately transforming mass spectra into the knowledge related to biological systems.

Compared to the rapid instrumental developments, the data analysis paradigms and computational methods for shotgun proteomics have not changed much since its early ages. Especially, the dramatically improved resolution and mass accuracy have not been fully utilized (Olsen et al., 2005, Zubarev and Mann, 2007). Apparently, developments in bioinformatics for proteomics are lagging behind other fast-growing 'omics fields.

Here, I will briefly introduce the common bioinformatics methods for proteomics, and address the current methodological limits in analyzing high-resolution and high-throughput data, which can be improved by the studies included in this thesis.

Identification

In shotgun proteomics, the analytes of an LC-MS/MS experiment are ionized proteolytic peptides. Unlike DNA sequencing, the typical readout from a mass spectrometer is not the processed information about the sequence itself, but the much less intuitive *m/z* values. Due to the inherent complexity, it could be one of the most diverse and challenging tasks in shotgun proteomics to associate a set of *m/z* values to the identity of a certain peptide (Marcotte, 2007).

In order to know the mass (m) of a given peptide, one needs to determine its charge state (z). Fortunately, as proteins and peptides are carbon-based macromolecules, multiple isotopic peaks may be observed due to the natural abundances of stable isotopes of chemical elements, especially the relatively abundant (~1.1%) carbon-13. High-resolution MS (e.g., FWHM > 15,000) can well distinguish the *m/z* difference between peptide ions ($m \ll 100$ kDa)

with different compositions of stable isotopes, for example, between carbon-12 and carbon-13 ($\Delta m = {}^{13}C - {}^{12}C = 1.003355$ Da ≈ 1 Da). The charge state could be determined by the *m/z* difference between two isotopic peaks, $z = {\Delta m}/{z}^{-1}$. For example, a difference of ${\Delta m}/{z} = 0.5$ indicates a doubly charged precursor, while a ${\Delta m}/{z} = {1}/{3}$ difference means triply charged ions, and so forth.

In addition, it is important to determine the monoisotopic mass, which is the mass of the analyte molecule where each of the chemical elements is from the most abundant isotopes, such as ^{1}H, ^{12}C, ^{14}N, ^{16}O, as well as ^{32}S. Because the 1 Da mass difference between the stable isotopes of carbon is slightly different from (~ 6.3 mDa heavier than) that between the isotopes of nitrogen, the "fine structures" composed of peaks of various (heavier) isotopes are often convoluted into a broader peak, whose position is defined less accurately than that of the monoisotopic peak. Resolving isotopic fine structures requires ultrahigh resolution (e.g., FWHM 500,000 to 1,000,000) that is available exclusively through FTMS. Importantly, the monoisotopic mass is additive, while the most abundant isotopic mass is not.

As mentioned before, the peptide mass alone might not be a unique identity of the sequence, but combining extra-dimensional information (RT or pI) with the mass may create peptide mass fingerprints (PMFs) that greatly improves the specificity (Smith et al., 2002, Pasa-Tolic et al., 2004). More detailed and reliable information about the peptide sequence is stored in MS/MS spectra. SEQUEST (Eng et al., 1994) and Mascot (Perkins et al., 1999) are the two traditional tools that are widely used in proteomics for identification of peptides from MS/MS spectra. Both are in the category of MS/MS database search engine that matches the peptide fragment fingerprints (PFFs) to the theoretical fragments ***in silico*** generated from a database of protein sequences. The certainty of a peptide-spectral matching (PSM) is estimated from the probability of randomly matching the MS/MS spectra to the theoretical peaks (*p*-values), or the expected number of random matchings in the given database (E-values). However, both search engines were developed quite a while before the prevalence of high-resolution instruments. Since then, many new algorithms have been proposed to improve the accuracy and reliability of peptide identification (Käll et al., 2007, Cox et al., 2011b, Kim and Pevzner, 2014).

Instead of matching experimental peaks to the theoretical peaks in a database search, spectra that have been previously identified can also be used as a reference library for matching (Lam and Aebersold, 2011, Yen et al., 2011). This approach is often applied for analyzing DIA data (Schubert et al. 2015). Compared to database searching, this method offers higher specificity but is still limited by the empirical knowledge stored in the spectral library.

The drawback of using a protein database or a spectral library is the little tolerance to any sequence variation or posttranslational modification (PTM). PTMs and mutations will accordingly introduce mass shifts to the masses of peptide ions, as well as the related fragments. In fact, a large proportion of unidentified spectra might probably be peptides that carry sequence variants and PTMs (Savitski et al., 2006b), and the border between mutations and PTMs is blurry. For example, a methylated aspartic acid (Asp) is identical to a glutamic acid (Glu) from the *m/z* point of view. Wrong conclusions can be drawn from using an inappropriate database (Knudsen and Chalkley, 2011). Therefore, identification and localization of PTMs and mutations require either a modification to the theoretical masses or a searching strategy that tolerates mass errors (Mann and Wilm, 1994, Chick et al., 2015).

Table 1 | Mass increments (Δm) of canonical amino acid residues.

Amino acid	Code	Monoisotopic mass	Amino acid	Code	Monoisotopic mass
Glycine	G	57.02146	Glutamine	Q (GA/AG)	128.05858
Alanine	A	71.03711			
Serine	S	87.03203	Lysine	K	128.09496
Proline	P	97.05276	Glutamate	E	129.04259
Valine	V	99.06841	Methionine	M	131.04049
Threonine	T	101.04768	Histidine	H	137.05891
Cysteine	C	103.00919	Phenylalanine	F	147.06841
Leucine/Isoleucine	L/I	113.08406	Arginine	R	156.10111
Asparagine	N (GG)	114.04293	Tyrosine	Y	163.06333
Aspartate	D	115.02694	Tryptophan	W	186.07931

The mass is identical between leucine (L) and isoleucine (I).
The mass of asparagine (N) equals the mass two glycine residues (GG), and the mass of glycine-alanine (GA) is indistinguishable from that of glutamine (Q).
Glutamine (Q) is less than 36.4 mDa (0.03%) lighter than lysine (K).
Carbamidomethylation is a deliberate PTM introduced to cysteine residues during sample preparation, which gives a mass shift of +57.02146 Da to the mass of cysteine.
Oxidation is also a common PTM that introduces a +15.99491 Da mass shift.

Nevertheless, it is possible to wholly or partially construct the peptide sequences directly from MS/MS spectra. Tools for peptide *de novo* sequencing (Taylor and Johnson, 1997, Frank et al., 2007, Ma and Johnson, 2012) have evolved from the early approach of sequence tags (Mann and Wilm, 1994). The *de novo* approaches directly infer the sequence of a peptide from the mass increments (Table 1) in the ladder of fragment peaks, without any prior knowledge of the protein sequences. Peptide *de novo* sequencing requires almost full backbone

coverage from the fragments and works best with short and lower-charged peptides. However, amino acids having the same masses (Table 1) can hardly be distinguished, and the introduction of sequence variances and PTMs can make the sequencing process even more difficult and less accurate. In order to improve the accuracy of peptide *de novo* sequencing, high-resolution spectra with complementary fragmentation techniques (e.g., HCD+ETD) are highly recommended (Nielsen et al., 2005, Zubarev et al., 2008, Chi et al., 2013). In practice, this approach is mainly used when a reference protein database is not available, for example, sequencing of antibodies (Bandeira et al., 2008).

Quantification

The changes of protein abundances can reflect the systematic responses to perturbations. In various contexts, differentially expressed proteins could reveal the key components of biological pathways and interaction networks, thus could be used as biomarkers for disease diagnose, or potential therapeutic targets for treatments. Accordingly, many proteomics studies have a common design, which is quantitative profiling of the protein contents of in different systems (e.g., perturbed vs. unperturbed) followed by comparative analysis (Ong and Mann, 2005).

Fortunately, modern mass spectrometers provide a decent quantitative accuracy and a relatively wide dynamic range that is suited for the proteome-wide analysis (Zubarev, 2013). Most of the time, protein quantification is done on a relative scale by comparing protein abundances between the samples. It is, however, possible to estimate the absolute protein abundance (e.g., copy number per cell) by comparing measured abundances to internal standards with know absolute concentrations (Gerber et al., 2003, Silva et al., 2006, Wiśniewski et al., 2014).

Quantitative analyses have two main approaches: stable isotopic labeling and the alternative, label-free quantification (Bantscheff et al., 2007, Cappadona et al., 2012). The labeling techniques, such as iTRAQ (Ross et al., 2004), TMT (Thompson et al., 2003), and SILAC (Mann, 2006), provide the power of multiplexing to MS analysis (Rauniyar and Yates, 2014). By isotopic labeling, as many as ten (or even more) samples can be quantified simultaneously in the same LC-MS/MS experiment (McAlister et al., 2012, Werner et al., 2014). However, the increased multiplexing ability comes at a price. As mentioned before, it requires ultrahigh resolving power (> 70,000 FWHM at 200 *m/z*) to distinguish the densely-

coded mass tags (Hebert et al., 2013, Rose et al., 2013, Merrill et al., 2014). By the labeling approach, the quantitative information is often obtained by comparing the intensities of the same ionic species, but with different isotopic compositions. Inevitably, the process of quantification is built on top of the identification results that attribute the peptides' quantities to the source proteins; thus, errors in identification will cause imprecise quantification.

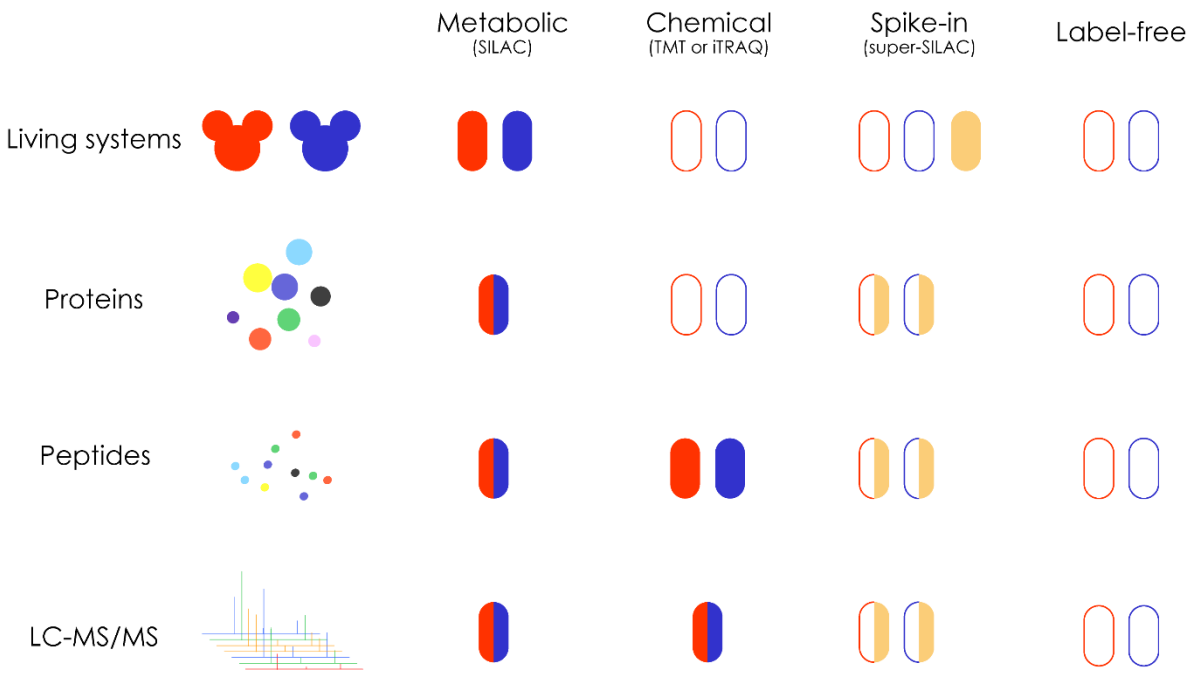

Figure 1.4 | Quantification techniques for shotgun proteomics. Adapted from (Bantscheff et al., 2007, Bantscheff et al., 2012). Solid capsules represent samples labeled by stable isotopes (e.g., ^{13}C and ^{15}N), while empty capsules represent unlabeled samples. Metabolic labeling methods (e.g., SILAC) incorporate heavy isotopes into living systems. Chemical labeling introduces post-translational modifications to specific amino acid residues or terminal groups using reagents with stable isotopes. Popular chemical labeling techniques (e.g., TMT and iTRAQ) are usually applied after protein digestion. Spike-in approaches combine a quantitative standard (labeled) into each protein sample. For example, in super-SILAC, an isotope-labeled cell mixture can serve as the internal standard (Geiger et al., 2010). Isotope labeling (metabolic and chemical) enables multiplexing, i.e., quantifying more than one sample in each LC-MS/MS experiment. Label-free quantification measures each sample individually.

LFQ, on the other hand, is appealing for having less experimental steps and is not limited by the size of studies. However, it requires for each sample an individual LC-MS/MS experiment (Hein et al., 2012). The very basic method of LFQ is spectral counting (SpC), which relies on counting the number of identified MS/MS spectra to approximate the relative logarithm of the abundance of the proteins (Liu et al., 2004, Ishihama et al., 2005, Griffin et al., 2010). However, SpC is often deemed as semi-quantitative and has poor performance for lower abundant proteins with few identified peptides (Tabb et al., 2015). The more accurate solution is based on extracted-ion chromatograms (XICs) that can be applied to both precursors and fragments depending on the data acquisition mode (DDA or DIA). By measuring the time-dependent eluting profile of the ions, one can generate a map of chromatographic features that contains the quantitative information about the peptides, even without being associated with the corresponding sequences.

Larger variances might be introduced into data analysis while combining multiple experiments. Compared to the labeling approaches, LFQ has a greater challenge of controlling the run-to-run variances. Especially, as mentioned before, the identification of peptides can be more or less different from run to run, due to the stochastic tendency in DDA. If the quantification is performed on top of the identification result, many peptides will have missing abundances in some of the experiments. The problem with missing values could be seen as one of the biggest obstacles in LFQ. For this reason, many sophisticated algorithms have been proposed for LFQ to address the issue with significant variances (Clough et al., 2012, Lyutvinskiy et al., 2013, Cox et al., 2014). Moreover, computational efficiency might also be concerned in a large-scale implementation (Khan et al., 2009). Comprehensive reviews of LFQ algorithms can be found in recent studies (Matzke et al., 2013, Sandin et al., 2014).

Proteome-wide summarization

Shotgun proteomics involves proteolytic digestion, which breaks proteins down to peptides. In this scenario, an extra but necessary procedure is to reconstruct proteins based on the identified peptides, which is called "protein inference". Due to sequence homology, the links between some of the peptides and the original protein molecules might no longer be unique after the digestion. A practical approach to protein inference is to apply the principle of parsimony and assembles the entire set of identified peptides into the least possible number of distinguishable protein groups (Nesvizhskii et al., 2007, Ma et al., 2009). Alternatively,

proteins could also be inferred based on likelihoods (Serang and Noble, 2012) or quantitative patterns (Forshed et al., 2011, Lukasse and America, 2014).

Compared to the already difficult problem of protein inference, quantitative estimation of protein concentrations is more complicated, as it is based on multiple measurements of individual peptides. A common assumption is that the peptides' abundances are proportional to the concentration of the source proteins (Walther and Mann, 2010). Accordingly, summing up peptide abundances is the approach that is well adopted in shotgun proteomics (Schwanhausser et al., 2011, Wilhelm et al., 2014). The question is, do all the peptides from the same source protein response equally to the concentration change? The answer might likely to be negative because multiple factors are involved in the quantitative measurements of peptides, many of which can violate the assumption of proportionality. Such factors include efficiency of enzymatic cleavage, peptide ionization efficiency, charge distribution, mass range of the instrument, sequence variance and homology, and so forth (Bantscheff et al., 2012). However, the biggest obstacle is still related to missing values, which are frequently encountered, especially in LFQ. By aligning the chromatographic features, identities of some peptide may be propagated across LC-MS runs even in the absence of an MS/MS evidence (Thakur et al., 2011, Bateman et al., 2013, Weisser et al., 2013). Other procedures including imputation are often needed for enabling proteome-wide summarization of protein abundances and their comparisons (Karpievitch et al., 2012).

Integrative workflows

The field of proteomics is diverse, not only in the technological choices but also the analytical methods. Practically, no "one-size-fits-all" strategy could be applied universally in proteomic studies (Mallick and Kuster, 2010). Consequently, data analysis workflows often need to be tailored for specific types of applications and instruments.

Common procedures in shotgun proteomics processing include MS data format conversion, mass peak picking, spectral quality control, chromatographic feature detection, retention time alignment, mass calibration, MS/MS database search, FDR estimation, protein inference, protein quantification, spectral annotation, visualization, and so forth. Since many of these procedures could be standardized, many integrated workflows, such as the commercial ProteomeDiscoverer package and the freeware MaxQuant (Cox and Mann, 2008), have gained popularity for providing bundles of pre-optimized tools for efficient processing

of typical proteomic datasets. However, the source codes of these programs are unavailable to general users, which reduces the flexibility of data analysis. In contrast, the OpenMS Proteomics Pipeline (Kohlbacher et al., 2007) is an open-source platform that provides greater flexibility for building customized analytical pipelines, especially for analyzing non-typical datasets.

Besides the main processes of protein identification and quantification, the pre-processing and post-processing steps could also affect the analytical performance (Wenger et al., 2011). Important pre-processing steps include precursor mass calibration (Cox et al., 2011a), deconvolution of charges and isotopes, spectral cleaning, as well as demultiplexing of chimeric MS/MS spectra (Egertson et al., 2013). One post-processing procedure, which has become standard in proteomic analysis, is estimating the FDR of identification based on the theory of target-decoy competition (Elias and Gygi, 2007). For a given scoring threshold, the proportion of spectra matching to decoy (reversed or shuffled) sequences can reflect the fraction of false discoveries in the identification list. Base on the theory, advanced algorithms (e.g., Percolator) can be applied to calculate the posterior-error-probabilities (PEPs), giving a more accurate approximation of the FDR (Käll et al., 2007, Käll et al., 2008). Different methods may have various scoring schemes for peptide-spectral matching. Thus, it is also important to generate a consensus result when peptide-spectral matches are divergent (Shteynberg et al., 2013). A protein group having more than one unique peptides could be considered more reliable than the "one-hit wonders" (Ong and Mann, 2005, Huang et al., 2012). Sophisticated protein inference approaches are especially useful when the number of identifications is large and when the FDR estimation at the protein level is no longer accurate. However, FDR estimation at the protein level is always harder and less accurate than for the peptide-spectral matches (Nesvizhskii et al., 2003, Serang et al., 2010, Savitski et al., 2015).

1.3 Comparative proteomics

Currently, the most common design in proteomics studies is the comparative analysis of two or more sample groups. In this context, the relative quantification approach plays a significant role in investigating the similarities and discrepancies at the protein level between the distinct functional states of the living systems. Changes of proteins' concentrations may indicate the

cellular responses to perturbations in a biological system. Covariation of the abundance changes may provide information about protein interaction and signal transduction pathways. Key nodes in the protein networks may serve as biomarkers or therapeutic targets for complex diseases. The challenge in comparative analysis is to discriminate signals from noises (Matzke et al., 2013). Therefore, advanced statistical methods and large sample sizes are often needed (Serang and Käll, 2015). For increasing sample size, highly multiplexed techniques, such as neutron-encoded labeling (McAlister et al., 2012, Hebert et al., 2013, Savitski et al., 2014), can be applied; LFQ, on the other hand, offers a simpler solution for further expanding proteomic studies to larger sample sizes, if only the issue with reproducibility could be well-addressed (Tabb et al., 2016).

1.4 Aims

The general aim of this thesis is to provide a set of advanced bioinformatics methods for extracting biologically relevant information from proteomics datasets. The main focus is on improving the data utility for high-resolution tandem mass spectrometry that could facilitate biomedical research.

The specific aims in the constituent papers are:

- to establish a computational workflow for efficient processing of high-resolution mass spectrometry data and improving peptides identification results (Paper-I),
- to illustrate a novel method for predicting the isoelectric point of peptides and proteins based on a new cIEF device (Paper-II),
- to establish a novel workflow for improving the reproducibility of XIC-based peptide quantification, and to introduce a quantification-centered analytical paradigm for enhancing data utility in proteomics (Paper-III),
- to propose a novel analytical approach for studying human antibodies in blood as disease biomarkers (Paper-IV),
- to describe an advanced algorithm for protein quantification, and to propose a flexible and reliable data analysis strategy for designing and conducting large-scale comparative proteomics studies (Paper-V).

Chapter Two: Present Investigations

2.1 Methodological considerations

LC-MS/MS

Proteomics experiments performed for this thesis were carried out with high-performance reversed-phase LC systems made by Thermo Fisher Scientific: EASY-Spray C18 columns (50 cm in Papers I and V, 15 cm in Papers II and IV), or in-house packed 10 cm C18 column (Paper-IV). Positive mode electrospray ionization (ESI) was used for coupling LC to MS. Fourier-transform mass spectrometry (FTMS) were performed with Orbitrap mass analyzers made by Thermo Fisher Scientific: Velos (Paper-II), Q-Exactive (Paper-I), Q-Exactive Plus (Papers IV and V) and Fusion (Paper-IV). In Paper-II, an additional dimension of sample separation was applied prior to LC-MS/MS, which used the cIEF technique that will be described later.

The data-dependent acquisition (DDA) was applied. In Paper-I, four different isolation windows (±1.0, ±2.0, ±3.0, and ±4.0 *m/z*) were used for investigating the specificity of precursor selection and the effect of multiplexing. Basically, a wider isolation window gives less specificity to the precursor ions and generates complexed spectra due to co-fragmentation of precursors. Further widening the isolation windows will decrease the specificity of the precursor-fragments relations in MS/MS, and will yield noisy chimeric spectra that are similar to DIA-MS/MS. However, processing DIA-derived data requires a different strategy to deconvolute each of the MS/MS spectra and reconstruct the information on the precursor-fragments relations (Egertson et al. 2013; Tsou et al. 2015).

In addition, a segmented DDA strategy (Vincent et al., 2013) was applied in Paper V, by limiting the range of precursor *m/z* in low, mid and high three non-overlapping mass ranges for triggering DDA. Such strategy intentionally reduced data redundancy of MS/MS for

investigating the quantification quality after implementing peptide identity propagation (PIP) across LC-MS/MS runs.

Higher-energy collisional-induced dissociation (HCD) was the default technique applied for fragmentation. Electron-transfer dissociation (ETD) was also implemented in Paper-IV, in order to generate complementary fragment ions in the HCD-ETD spectral pairs for reliable peptide *de novo* sequencing.

MJ-cIEF device and pI

In Paper-II, an additional dimension of sample separation was applied using the newly developed isoelectric focusing device: pI-Trap (multiple-junction capillary isoelectric focusing fractionator, MJ-cIEF). The pI is defined as the point of pH value at which a given peptide acquires a net charge of zero. The net charge of peptides could be calculated based on the equations derived from the Henderson–Hasselbalch equation, which was described in details in the reference of Protein Modification Screening Tool (ProMoST) (Halligan, 2009).

The negative charge (C_-) is determined by the pH of the environment and the pKa value of each of the n negative groups.

$$C_- = \sum_{i=1}^{n} \frac{-1}{1 + 10^{pKa(i) - pH}}$$

Similarly, the positive charge (C_+) is determined by the pH of the environment and the pKa value of each of the m positive groups.

$$C_+ = \sum_{i=1}^{m} \frac{1}{1 + 10^{pH - pKa(i)}}$$

When setting the net charge (sum of the positive and negative charges) to zero:

$$C_- + C_+ = 0$$

the corresponding pH (i.e., the estimated pI) could be determined by the pKa values of the chargeable groups in the chain of amino acids. The pKa values of chargeable groups were obtained from the website of ProMoST [1] (Halligan, 2009).

[1] http://proteomics.mcw.edu/promost.html

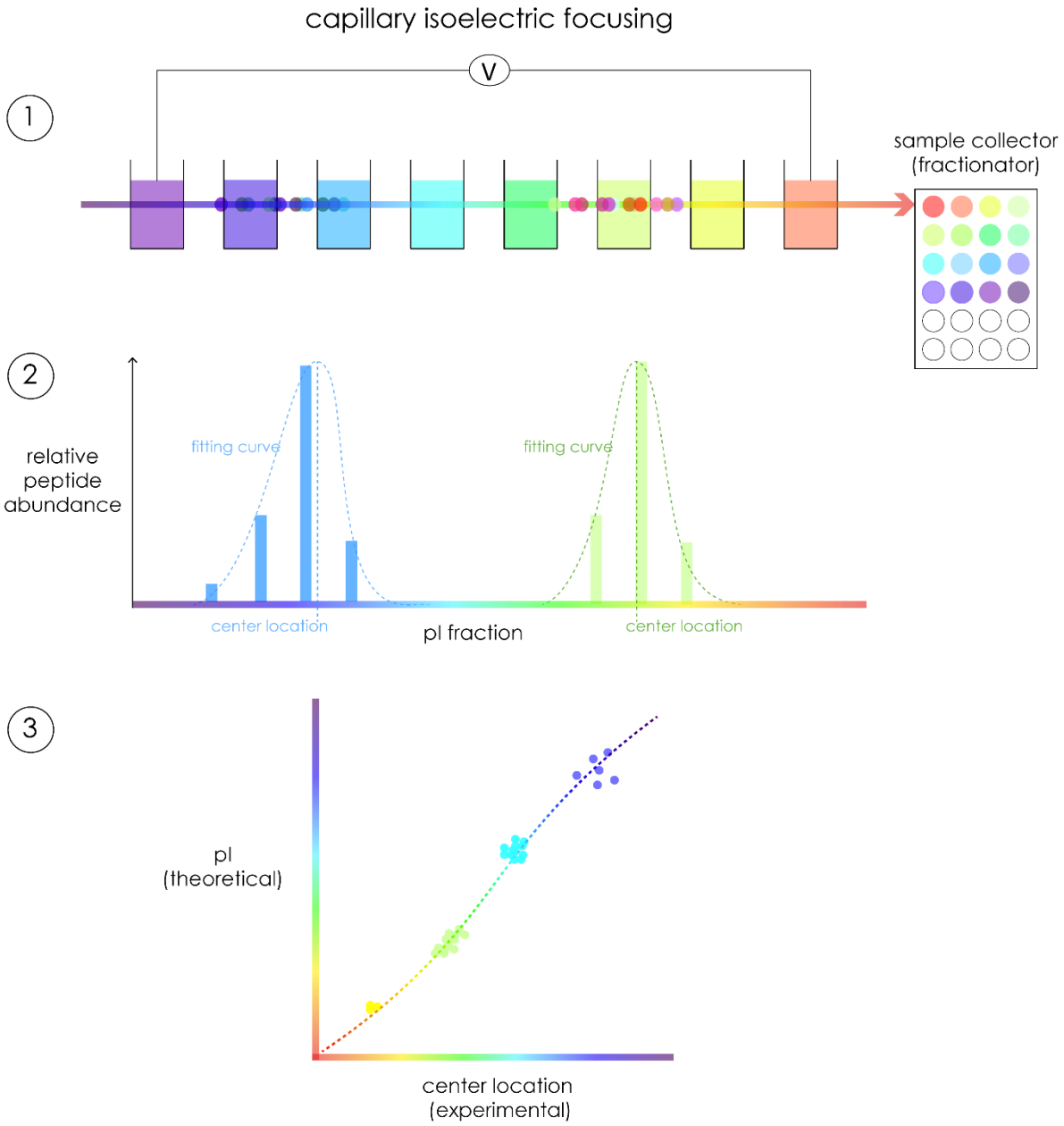

Figure 2.1 | Calibration of pI gradient for the MJ-cIEF sample fractionator (pI-Trap). Peptides are injected into the cIEF device, focused in the capillary column based on their pI properties, and then collected in different fractions depending on the time of flowing through the column. Abundant peptides observed in multiple fractions are used as markers for connecting the fractional location of the capillary column to the theoretical pI values. Center locations of the marker peptides in the cIEF column are estimated by fitting skewed Gaussian curves to the relative peptide abundances observed in different fractions. Theoretical pI values are predicted by computationally approximating the optimal pH that yields the minimal net charge of the peptide molecules. (Colors indicate the pH or pI: blue basic, red: acidic).

Chargeable groups include side chains of acidic (negative) residuals: glutamic acid (E), aspartic acid (D), tyrosine (Y), and basic (positive) residuals histidine (H), lysine (K), and arginine (R). However, the Cysteine (C) side chain with the fixed modification (carbamidomethylation) was not considered as a chargeable group. Terminal groups (N-terminus and C-terminus) are also chargeable but were considered individually for the 20 terminal amino acids. A Python script has been developed for calculating the optimal pH value by approximating the minimum net charge for a given sequence of amino acids.

The information on pI of the peptides was associated with the fractional location of the cIEF device. Peptides collected from more than three fractions were chosen as the self-calibrating markers. Skewed Gaussian curves were fitted to the observed peptide abundances, producing the estimated center locations of the marker peptides. Theoretical pI values and the center locations of the marker peptides were connected by five-degree polynomial curve-fitting, which transformed the fractionator to a non-linear pI gradient (Figure 2.1).

External datasets

Data generated from different laboratories are especially useful for testing the compatibility of the data processing approaches. In Paper-I, the reference dataset produced by researchers from University of Texas Southwestern (Guo et al., 2014) was used to validate that the new peptide identifications are not artifacts. Paper-III and V used the data of the ABRF-iPRG2015 study (Choi et al., 2017), to assess the reproducibility of peptide quantification. The iPRG study contains 12 single-shot LC-MS/MS (Orbitrap) experiments that measured four samples with a common background proteome and six marker proteins spiked in by different concentrations. In Paper-V, to demonstrate the consistency of the protein quantification results, two datasets were obtained from supplementary materials of two clinical studies of breast cancer: 1) The CPTAC breast cancer dataset (Mertins et al., 2016) contains 77 breast cancer samples (quality control passed) analyzed by iTRAQ experiments; and 2) The MPIB dataset acquired from the study conducted at the Max Planck Institute of Biochemistry, Germany (Tyanova et al. 2016), which contains 40 breast cancer samples compared to a spike-in standard (super-SILAC) of isotopic-labeled mixture of breast cancer cell cultures.

Identification

Search engines. Mascot (Perkins et al., 1999) was used in Paper-I to benchmark the improvements from other advanced methods. In Paper-I and Paper-III, the "universal" search engine MS-GF+ (Granholm et al., 2014, Kim and Pevzner, 2014) was implemented, to achieve the deepest proteome coverage. In Papers-I, II, III, and V, the Andromeda search engine (Cox et al., 2011b) was used, which is integrated into the popular MaxQuant package (Cox and Mann, 2008, Michalski et al., 2011a, Tyanova et al., 2016). With high-resolution MS/MS and the target-decoy strategy (Elias and Gygi, 2007), matching MS/MS spectra to the protein database can be simplified to counting of observed fragment ions with little tolerance to mass errors (< 20 ppm). Applying the fragment counting strategy, the Morpheus search engine (Wenger and Coon, 2013) was used in Papers I, IV, and Paper-V. On top of Morpheus-derived PSM results, an advanced scoring strategy was proposed in Paper-I with improved usage of the mass accuracy (Zubarev and Mann, 2007) and the information on complementary fragment pairs (i.e., b/y ion-pairs).

Precursor information. Reliable peptide identification based on MS/MS requires accurate *m/z* of the precursor ions measured in the survey spectrum (MS^1). In Papers-I, II, III, and V, the precursor mass information was firstly calibrated by identifications from the first-pass database search. Systematic mass deviations were automatically corrected by offsetting the average mass error for each of the precursors, then search the MS/MS spectra again with the calibrated precursor information. The concept used in this approach is called "software lock-mass" (Cox et al., 2011a).

DDA-derived MS/MS usually records one precursor *m/z* per spectrum, which implicates only one peptide identification from the spectrum. However, in a rapid analysis of a complex proteomic sample (e.g., human cells), the distribution of precursor ions in a given mass range is considerably dense (Michalski et al., 2011a). Consequently, only in rare cases, the precursor ions collected within the isolation window could be pure. In this regard, most of MS/MS spectra are chimeric, "naturally" containing extra pieces of information about co-fragmenting precursors. Associating these additional precursors with the MS/MS spectra, a set of MS/MS "clones" were generated. Such a strategy for naturally multiplexing of MS/MS spectra was applied in Papers-I, III, and V.

Peptide *de novo* sequencing. Peptide sequences could be infered directly from mass increments of fragment ions in MS/MS. As an alternative to a database search, peptide *de novo* sequencing is especially useful when the protein contents could not be referenced to a predefined database. For example, in Paper-IV, the sequence of human antibodies, especially

sequences from the complementary determining region (CDR), are of considerable uncertainty, due to the enormous possibility ($\sim 10^{15}$ in theory) of recombination and hypermutations (Schroeder and Cavacini, 2010). It is not feasible to provide a comprehensive reference database for matching all possible CDR sequences. In Paper-IV, pNovo+ (Chi et al., 2013) was applied to produce reliable full-length peptide sequences directly from high-resolution HCD/ETD spectral pairs (Savitski et al., 2005, Savitski et al., 2006a). In Paper-V, Novor (Ma, 2015) was used to generate *de novo* peptides from HCD spectra. In the absence of the complementary information from ETD, the latter approach (Paper-V) is arguably less accurate than the former one (Paper-IV). However, both methods applied stringent criteria for quality control, to achieve a reasonable accuracy of connecting the *de novo* sequences to their source proteins. The filtration was made by analysis of sequence similarity using protein BLAST (Altschul et al., 1990), with strict limits of sequence identity between *de novo* peptides and the protein sequences in the databases.

Peptide identity propagation (PIP). To counteract the stochasticity caused by DDA and increase the reproducibility of peptide identification, PIP was applied in Paper-III, IV, and V. This option is implemented in MaxQuant as a feature of "match-between-runs", and also in OpenMS (Kohlbacher et al., 2007, Weisser et al., 2013) as FeatureLinkerUnlabeledQT. PIP relies on reproducible LC chromatogram, calibrated retention time, as well as accurate measurements of the precursors' *m/z*. Sequence identities of the chromatographic features were propagated across multiple LC-MS/MS experiments by the concept similar to the accurate-mass-and-time (AMT) tag (Smith et al., 2002) and the peptide mass fingerprint (PMF) (Moruz et al., 2013). An alternative to aligning chromatographic features that are detected by clustering isotopic peaks, the more sensitive (but less accurate) PIP could be done based on XIC, which is implemented by Skyline (Schilling et al., 2012) and OpenMS (EICExtractor). In Paper-III, both feature-based and XIC-based PIP approaches were combined, and quantitative estimations of FDR were applied.

Protein inference. After associating MS/MS spectra with peptide sequences, protein inference was done by applying the principle of parsimony (Ma et al., 2009), which reports the minimum number of proteins covering the entire list of identified peptide sequences. The uniqueness of peptide-protein association was determined by whether a peptide could be attributed to more than one proteins. However, a protein group may contain more than one protein sequence (for example, isoforms) that share the same set of unique peptides. When the number of unique peptides is relatively small (often less than three), or the similarities between multiple protein sequences are relatively high (e.g., multiple isoforms per gene), the reported protein groups may contain multiple entries of the database.

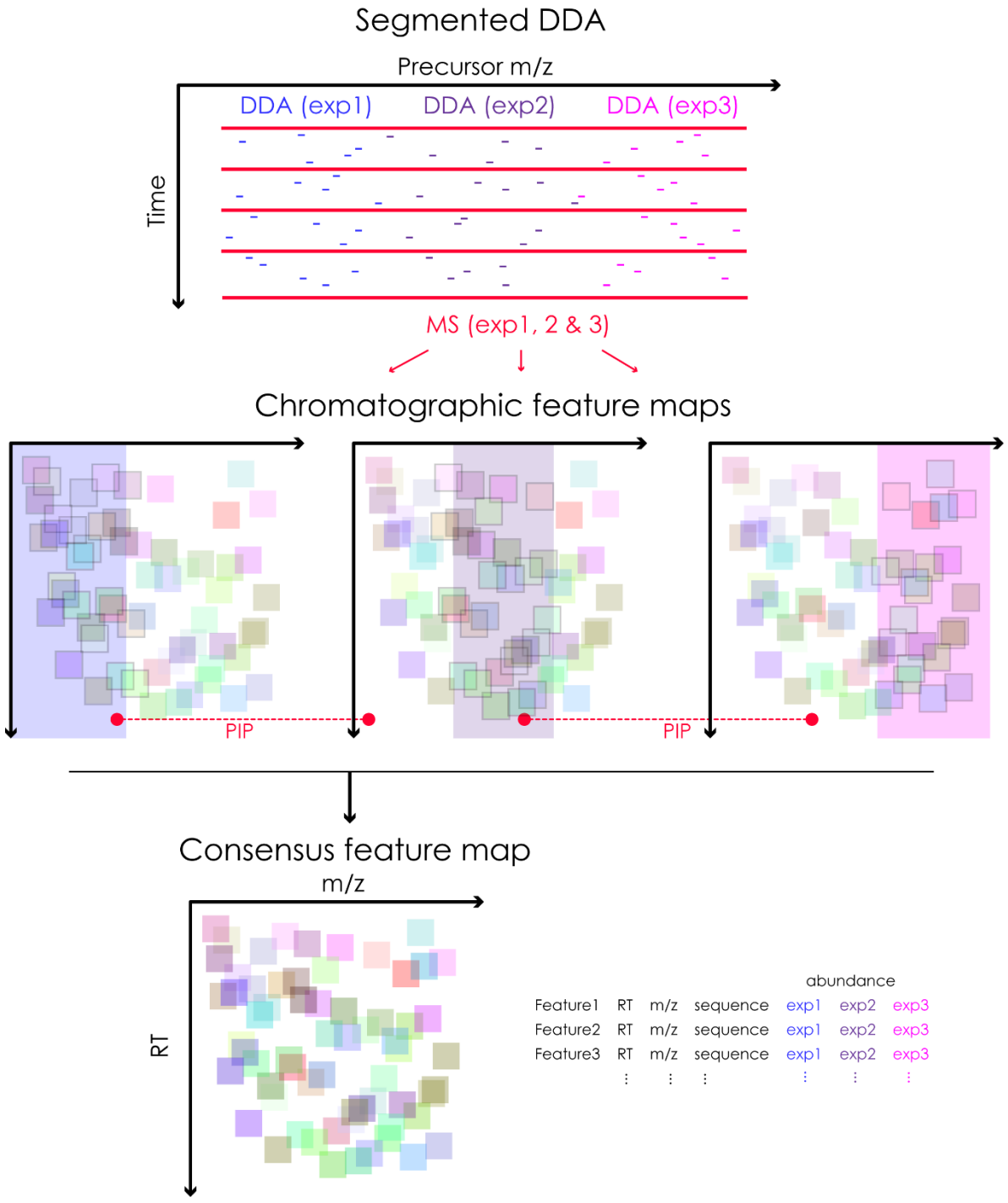

Figure 2.2 | Segmented DDA and peptide identity propagation. In Paper-V, the full-range survey MS spectra were applied in each experiment and were assembled into chromatographic feature maps. However, MS/MS was applied to analyze precursors in three different *m/z* ranges (low, mid, and high). MS/MS-derived identifications results were assigned to the feature maps and propagated between experiments. A consensus feature map was generated by aligning and combining all the individual feature maps.

The choice of a protein database affects not only the search space for peptide identification but also the complexity of protein inference when applying the parsimony principle. In Paper-I, the UniProt "Human Complete Proteome" (composed of canonical and additional sequences) was used. In Paper-II the Ensembl database derived from the *Saccharomyces cerevisiae* (Yeast) genome was used. In Paper-III, the UniProt (SwissProt) database of yeast proteins was provided in the data of iPRG-2015-study. In Paper-IV, the UniProt (SwissProt) database of human proteins was used for the first-pass MS/MS identification for excluding known sequences, and also used for the BLAST search for sequence similarities with *de novo* peptides. In the final step, the SwissProt database was concatenated with the *de novo* peptide sequences for implementing traditional MS/MS identification. In Paper-V, as the 20 mixture contains three proteome components (human, yeast, and BSA), a concatenated UniProt "reference proteome" database was generated to cover the proteins that are likely to be observed in the samples. Common contaminants such as BSA, trypsin, and various keratins (from hair and skin) may hinder the identification process if the corresponding sequences are not included in the database. For this reason, an additional database that contains such proteins sequences was searched together with the main database, but peptides linked to potential contaminations were excluded from subsequent analysis.

Quantification

Label-free quantification (LFQ) is the primary method that has been investigated in this thesis and has been performed at different levels. In Paper-I, quantification was only carried out at the level of chromatographic features, for illustrating the abundance distribution of the newly identified co-fragmenting precursors. In Papers II, III, and partially in Paper-IV, the quantification was done at the peptide-level by summing up abundances of features attributed to the same peptide sequences. In Paper-V (and partially in Paper-IV), the quantification was investigated at the protein-level. Four traditional protein summarizing approaches (MaxLFQ, PQPQ, Top-3, and Median) were compared with the newly proposed Diffacto method. MaxLFQ summarizes relative protein abundances by linear regressions based on pairwise peptide log-ratios (Cox et al., 2014); PQPQ clusters peptides by measuring the linear correlations of log-abundances among multiple runs (Forshed et al., 2011, Zhu et al., 2014), the abundance of the largest cluster of peptides was summed up; Top-3 sums up the

abundances of the most intensive three peptides (Silva et al., 2006); the Median approach uses the mid-points of all peptide abundances. In the Diffacto approach, relative protein abundances were summarized by weighted geometric means of peptides' log-ratios. Unlike the compared methods that generate per-experiment protein summarization, the abundances of proteins were given by sample groups in the Diffacto results.

Missing values are obstacles in protein quantification, especially, LFQ. In Paper-III, the problem of missing values for peptide quantification has been investigated. By performing peptide identity propagation (PIP), the identities of the chromatographic features were inferred, and the quantitative information was directly extracted from the features. However, feature-based PIP only alleviated but did not fully solve the missing value problem. The more sensitive XIC-based PIP was applied to further recover the remaining missing values. Quality control is important in this process. A target-decoy approach was the applied for estimating the reliability of XIC. For each precursor ion, two target peaks (monoisotopic and the first ^{13}C isotopic, i.e., M and M+1) and two decoy peaks (by shifting precursor's *m/z* and RT) were extracted from the MS^1 spectra, and were scored based on the mass and RT deviations, as well as the CV between replicate experiments. A score threshold corresponding to 5% FDR was applied to assess the quality of XIC. This quality threshold was also applied in Paper-V.

Nevertheless, for the protein-level summarization in Paper-V, the missing value problem was further addressed by the proposed Diffacto algorithm. In the calculation of weighted geometric means, only the valid (non-missing) measurements were used. If only a significantly large proportion of peptide quantities went missing (for a specific sample group), the remaining missing values were imputed as bellow the detection limit.

Factor analysis was applied in Paper-V to measure the covariation of peptide abundances from multiple measurements by LC-MS/MS. A Bayesian factor analysis algorithm, FARMS (Hochreiter et al., 2006) that was originally developed for summarizing gene expression microarray data, was adapted for handling proteomics datasets. Similar to the PQPQ approach that measures the correlation of peptide abundances, factor analysis estimates the consistency of peptide signals by comparing to the estimated underlying factor reflecting the concentration changes of the protein. Peptides with incoherent trends of quantities were deemed as unreliable and were down-weighted or excluded from the protein summarization. The factor analysis was an unsupervised process, which estimated for each protein the proportion of information content (S/N) (Talloen et al., 2007) in the quantitative measurements of peptides, without using the information about the identities of samples.

False quantification rate (FQR) was investigated in Paper-V. For each protein quantified from the 20-mixture data, the ranks of 190 pairwise ratios were correlated, between the reference concentrations and the quantification results, using Spearman's rank correlation. Proteins with negative correlation (below the threshold r = 0) were considered as false quantifications.

Statistics

Peptide and protein identification was filtered at less than 1% FDR by the target-decoy approach (Elias and Gygi, 2007), with the assumption that MS/MS spectra could match to artificial sequences by chance, which reflect the proportion of random matches to the target sequences at a given threshold. In Paper-III and Paper-V, the peptide identity propagation (PIP) were controlled at FDR < 5%, using a similar target-decoy approach where the decoy features were created by shifting the *m/z* and RT values of target features.

For quantification of differentially expressed proteins in comparative analysis, the statistical significance was given by FDR < 5%. In Paper-IV, FDR estimated by Bonferroni correction of *p*-values derived from pairwise t-tests. In Paper-V, FDR was calculated as *q*-values derived from *p*-values of either AVONA or Monte Carlo random permutation tests (Sandve et al., 2011), with a conservative estimation of the proportion of true null hypotheses (Pounds and Cheng, 2006).

2.2 Results and Discussions

Natural multiplexing of MS/MS data

In Paper-I, we investigated the identification efficiency of high-resolution MS/MS spectra from a set of single-dimensional shotgun proteomics experiments. We found that DDA-derived MS/MS data frequently contain chimeric spectra, which is against the paradigm of "one MS/MS spectrum – one peptide identification" in traditional database search algorithms. Naturally, chimeric spectra contain the information about other precursors that were not initially targeted by DDA but were co-eluted from the LC and co-fragmented in the MS/MS. We observed a correlation between the width of precursor isolation window in DDA and the rate of acquiring chimeric MS/MS spectra, as well as the negative impact of chimeric spectra on the peptide identification. Estimated by the proportion of MS/MS spectra that contain both lysine (*m/z* 147.11280) and arginine (*m/z* 175.11895) peaks as the C-terminal (y1+) ions, we found that a majority of over 98% MS/MS spectra were chimeric.

By associating precursors' chromatographic features with the isolation windows of MS/MS data acquisition, we generated clonal MS/MS spectra with the same fragment mass peaks but newly assigned precursor information. We matched these spectra via database search, using a simplified scoring method that did not penalize the existence of unrelated peaks. As a result, we obtained significantly more PSMs and peptide identifications at the 1% FDR level.

We demonstrated the power of natural multiplexing from traditional DDA-derived MS/MS data, by applying MS/MS spectral cloning that broke down the old paradigm and yielded more than one peptide ID per MS/MS spectrum (Figure 2.3). An integrated data analysis workflow, DeMix, was developed on top of the OpenMS pipelines. DeMix transformed the previously unwanted events, i.e., co-fragmentation of peptides, into an advantage of multiplexing. We achieved an overall identification rate of 1.25 PSM per MS/MS spectrum, which exceeded the instrumental limit of data acquisition rate and yielded nine peptides per second identification efficiency.

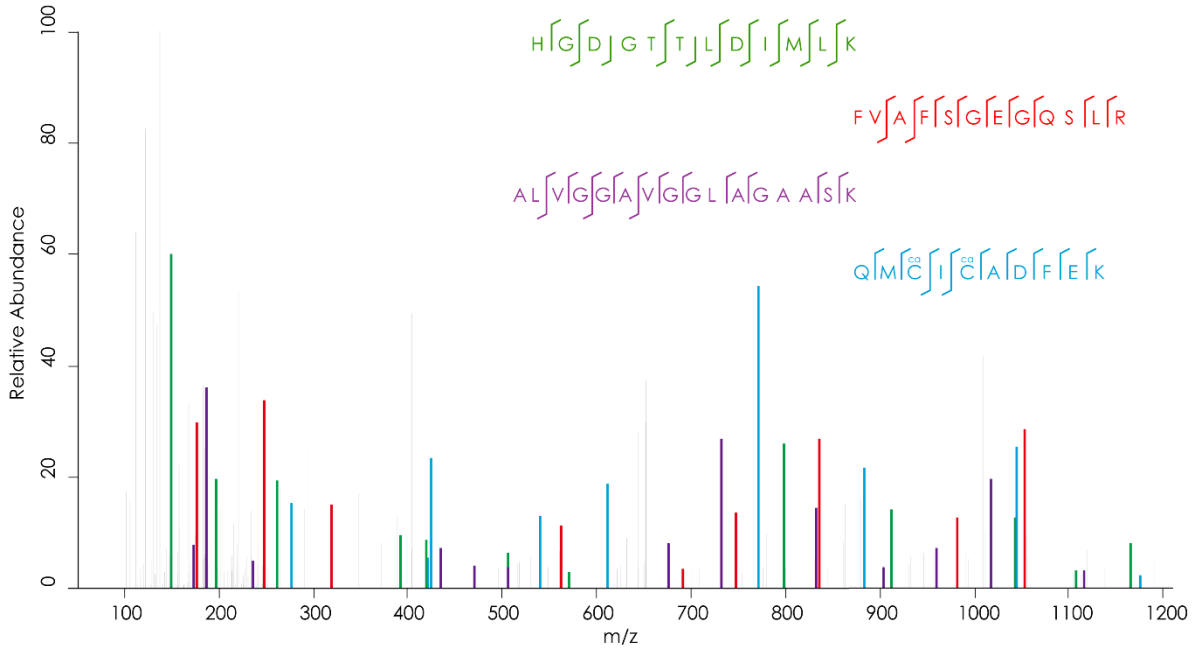

Figure 2.3 | Multiplexing MS/MS (an example). The spectrum was associated with four peptide precursors and was identified from the DeMix workflow. [HGDGTTLDIMLK]2+, *m/z* 650.8311; [ALVGGAVGGLAGAASK]2+, *m/z* 649.8724; [FVAFSGEGQSLR]2+, *m/z* 649.3308; and the original precursor, [QMCICADFEK]2+, *m/z* 651.2697.

We showed that the identified co-fragmenting peptides mainly came from the middle or low abundance range (Figure 2.4), which increased the dynamic range of data analysis and improved the reproducibility of peptide identification between experiments.

In principle, widening the precursor isolation window will associate more chromatographic features to the MS/MS spectra, increasing the capability of multiplexing. However, we observed a decreasing trend of identification efficacy for spectra acquired from isolation windows that are wider than 4.0 *m/z*. Possible explanations might be that the chimeric spectra become too noisy, or the ions of the dominate fragments suppress the signals of other ions and decrease the overall S/N of the spectra. This observation may raise a concern about the effectiveness of analyzing lower abundant peptides via DIA approaches that use much wider (>20 *m/z*) isolation windows, although the burden of identification is different between DDA and DIA data.

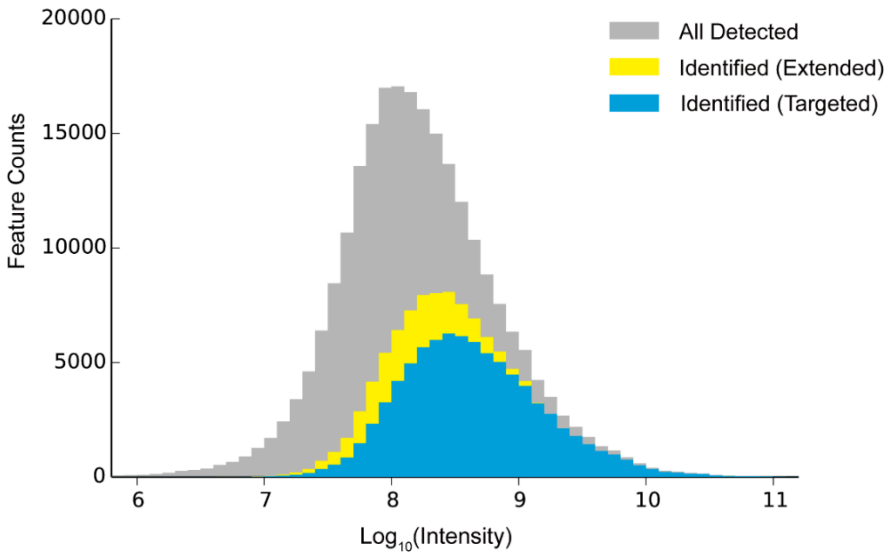

Figure 2.4| Abundance distributions of chromatographic features. Gray: all peptide-like features detected in LC-MS; blue: identified peptides targeted in conventional DDA strategy; yellow: peptides additionally identified by feature-based deconvolution.

Quantification-centered proteomics

It is frequently claimed that DDA has a stochastic tendency that causes the problem with missing values in quantification. However, since DDA is a strategy for acquiring MS/MS data and serves the purpose of identification, the survey spectra (MS^1) that are acquired before MS/MS should remain unbiased to all precursors. Therefore, the trouble with DDA-induced missing values should not be a fundamental problem for the quantification process that is performed at the level of MS^1.

One of the key findings in Paper-I was that the chromatographic features assembled from MS^1 spectra could provide reliable information about not only the accurate mass of precursors but also their quantities (i.e., intensities). In light of the DeMix workflow, we looked for a novel strategy that focuses on the quantitative aspect of proteomics data. Thus, in Paper-III, we developed an extension to the workflow, named DeMix-Q, that does the job of quantification while avoiding some pitfalls in the identification processes.

We demonstrated, in Paper-III, that peptides' identities that are inferred from MS/MS could be reliably propagated across multiple LC-MS/MS experiments, by aligning individual maps of chromatographic features that create a consensus feature map (see Figure 2.2). Although such strategy was already implemented in both MaxQuant and OpenMS, the performance was not satisfactory. We still observed significant proportions (ca. 15%) of missing values in the quantification outputs, when analyzing the data of iPRG-2015 study that repetitively measured, technically, the same proteomic contents. On the contrary, missing values composed only less than 2% of the quantitative data (derived from the Skyline workflow) provided by the iPRG-2015 study. However, we found, in the Skyline data, the CV of the quantities between replicated experiments was significantly larger than that of the MaxQuant and OpenMS results. The reason was, Skyline extracted XICs without quality control; while MaxQuant and OpenMS, on the other hand, generated chromatographic features that are "peptide-like", based on the isotopic distributions of precursor peaks and the shapes of peptides' eluting profiles.

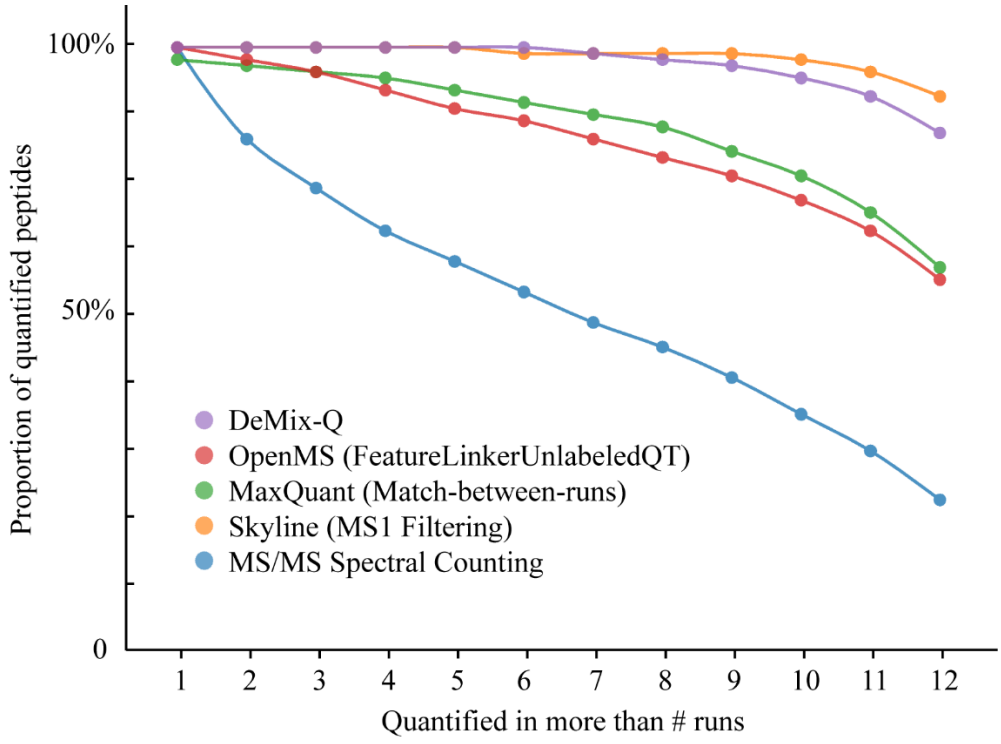

Figure 2.5 | Comparison of five quantification approaches regarding missing values. The fraction of peptides quantified in all runs drops as a function of sample size but with different rates for different quantification methods.

In the DeMix-Q workflow, we integrated the reliable feature-based PIP with the sensitive XIC-based PIP, and also applied FDR estimation in the latter approach for quality control. In addition, DeMix-Q implemented a global normalization of abundances, which corrects for each LC-MS/MS experiment the time-dependent median abundance shifts of features by comparing to the global medians. After normalization, the systematic errors caused by the fluctuation of peptide ionization were removed from the peptide quantities. As a result, DeMix-Q achieved sensitive and reproducible peptide quantification with less than 3% missing values when measuring 26,753 peptides in the 12 replicated LC-MS/MS experiments. Remarkably, the median CV of peptide quantification in 12 replicated runs decreased from 22.7% (Skyline) to 11.6% (DeMix-Q).

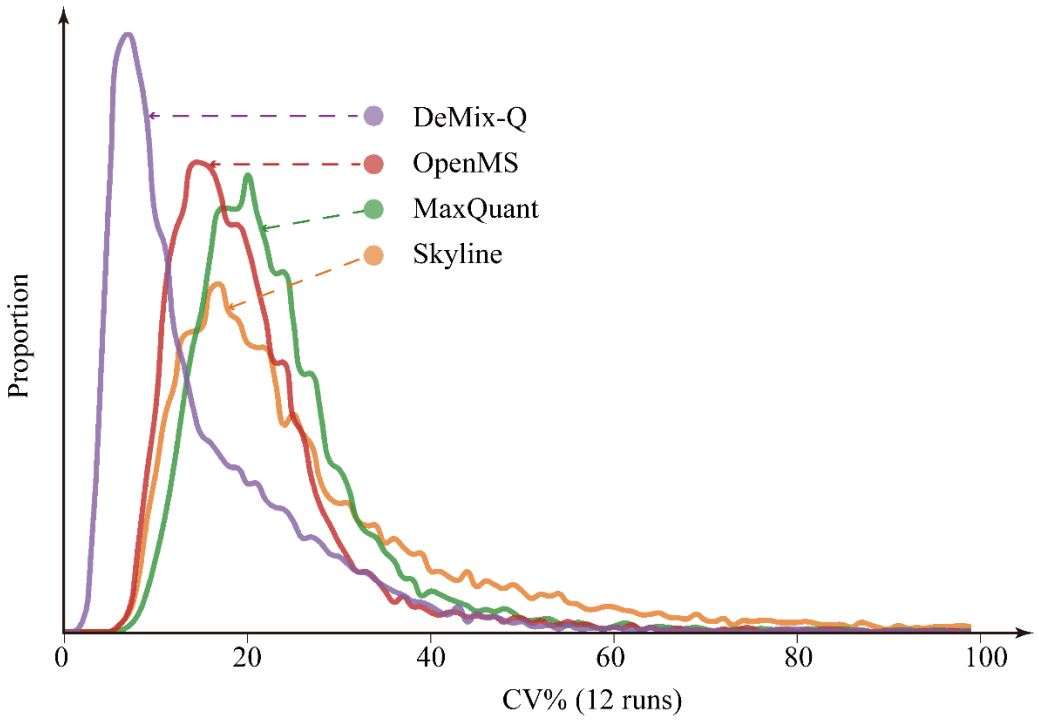

Figure 2.6 | Comparison of CV distributions. DeMix-Q provided lower median CV than other methods while quantifying the largest number of peptides across all runs. Median CVs: DeMix-Q: 11.6%, OpenMS: 18.6%, MaxQuant: 21.9%, Skyline: 22.7%.

As demonstrated in Paper-I and Paper-III, chromatographic features generated from MS^1 spectra are, in fact, reproducible and quantitative measurements of peptides ions. MS/MS-derived identifications were later associated with these quantitative measurements. Therefore, the relation between the identification and quantification may be well reversed.

The peptide identification result may likely compose a subset, rather than a superset, of the quantified chromatographic features.

We showed that more than a hundred thousand of peptide-like features could be detected in single LC-MS/MS experiment, even for a less complexed (yeast) proteome. Typical DDA data could merely identify one-third of these features and are biased against the lower abundant species. In that sense, with an ever-increasing size of proteomics studies, DDA-derived MS/MS spectra become more and more redundant, especially, for the abundant peptides and proteins. Thus, we made a hypothesis: repetitively identifying the most abundant peptides by MS/MS is unnecessary, then demonstrated that reproducible quantification could be achieved in the absence of redundant MS/MS data. With all the findings in Paper-I and Paper-III, we proposed the quantification-centered strategy for proteome-wide data analysis.

In Paper-V, we generated a set of LFQ data that consists of 63 experiments to measure the mixtures of three proteome components in 20 different concentrations. The DDA strategy applied for this dataset was intentionally segmented to three mass ranges in the triplicated measurements (see Figure 2.2). Consequently, for each of the experiment, around two-thirds of the peptide identifications had to be propagated from other experiments by aligning the chromatographic feature maps. Nonetheless, the segmented DDA did not prevent us from obtaining reproducible quantification results. The fraction of missing values in peptide quantification was only 12%, which was mainly due to the limit of detection. The median CV of replicated measurements was 12%, which was in line with our previous findings of LFQ.

When summarizing peptides' quantities to the abundances of proteins, the difficulty is often underestimated in traditional identification-based approaches. However, the concept for the quantitative analysis is similar between shotgun proteomics and the old-style gene expression microarrays. With the increased proteome coverage and the improved reproducibility of LC-MS/MS experiments, the data structures between MS-based shotgun proteomics and microarray-based transcriptomics become more and more similar. Therefore, we brought a widely-used summarizing method from transcriptomics to proteomics, which applies a Bayesian factor analysis algorithm to measure the covariation of peptides' signals in response to protein concentration changes. The factor analysis enabled the quality assessment for quantitative analysis. By this approach, peptides with incoherent trends of abundance changes could be detected from the covariation metrics, and the overall signal-to-noise ratio (S/N) could be estimated for each group of peptides that are tentatively attributed to the same source protein. Factor loadings, reflecting the correlation between the peptide abundances underlying factor of the concentration change, provided weights of individual peptides and

removal of unreliable quantities; S/N values indicated the overall consistency of peptides' abundances and were used to categorize the protein quantification results as informative or non-informative. The distribution of S/N exhibited a strong correlation with the number constituent peptides, showing the vital importance of having multiple quantitative measurements per each protein (Figure 1 in Paper-V).

With the two-level quality controls based on the factor analysis, we demonstrated the improved quantification accuracy with reduced errors in summarizing the LFQ dataset. Compared to the traditional approaches, the false quantification rate dropped from as much as 14% to 1.6% (Figure 2.7). In addition, the weighted summarizing approach addressed the problem with missing values that further improved quantification accuracy.

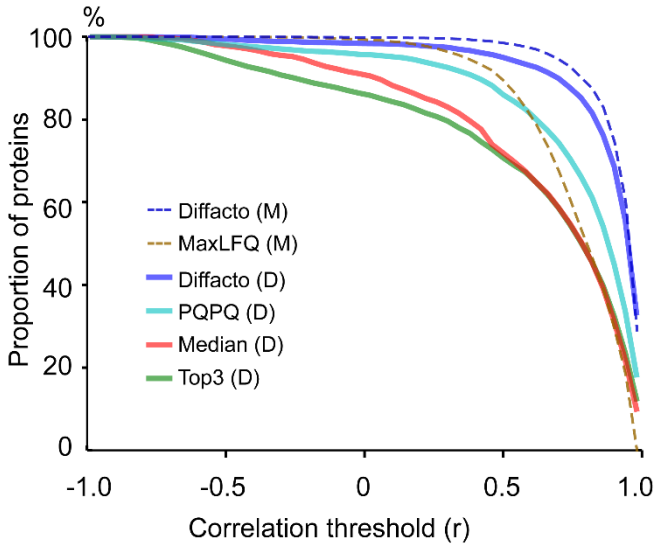

Figure 2.7 | Evaluation of the precision of protein quantification results. Dashed lines: quantification based on MaxQuant (M) peptide abundances. Solid lines: quantification based on DeMix-Q (D) peptide abundances. Abundances of informative proteins summarized by different techniques were correlated to the actual protein concentrations. The proportions of quantified proteins (y-axis) at the correlation threshold (r = 0) were used to estimate false quantify cation rates: 14.3%, 9.6%, 4.3%, 1.6%, 0.68%, and 0.13%, respectively in Top3 (D), Median (D), PQPQ (D), Diffacto (D), MaxLFQ (M), and Diffacto (M) results.

We named the summarizing method Diffacto, for relative quantification of differentially expressed proteins based on factor analysis. We applied Diffacto to analyze two sets of data generated from clinical breast cancer studies and revealed the persistent proteomic signatures

of three subtypes of breast cancer. We improved the consistency of the quantification results between the two independent studies.

In addition to the flexible quantification approach, a reliable FDR estimation method was also applied, based on sequential Monte Carlo random permutation tests (Sandve et al., 2011). Such non-parametric statistical approach provided conservative FDR control for proteomics studies with complex designs.

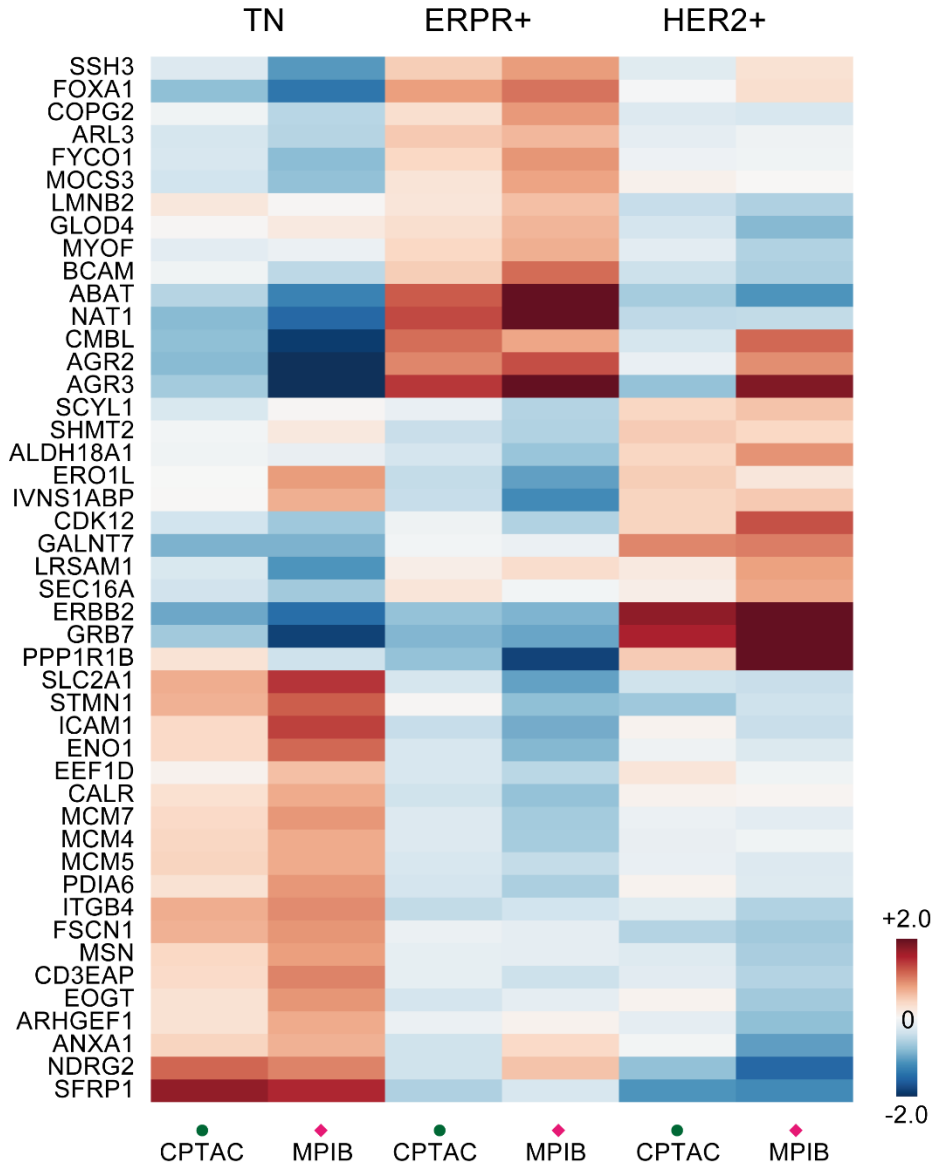

Figure 2.8 | Expression patterns of the differential proteins. Protein fold-changes estimated by Diffacto (weighted geometric means) showed good agreements not only in the directions of regulation, but also in the magnitudes of changes between the results from CPTAC and MIPB data. Such protein expression patterns clearly clustered into three groups that represent the most persistent proteomic signatures of the three subtypes of breast cancer.

Discovering biomarkers in the hidden proteome

In Paper-II, we developed a novel MJ-cIEF device (named pI-Trap) for sample fractionation. By establishing a stable pH gradient and an electric field, the device focuses protein or peptide molecules in different regions of the capillary column, according to the pI properties. Sample fractionation was done by sequentially collecting the samples flowed through the cIEF column (see Figure 2.1 and Figure 1 of Paper-II).

To calibrate the pI gradient of the sample fractionator, we associated the center locations of the eluting profiles of peptides with theoretical pI values. The tight regression curve indicated the high precision of the pI estimation. By instrumental optimizations, we decreased the standard deviation of the pI prediction from 0.44 to 0.21 for the whole range of peptides.

The MJ-cIEF device was initially designed for providing an orthogonal dimension of sample fractionation prior to LC and MS. With the pI fractionation one can decrease the dynamic range in complex proteomes, such as the human blood proteome (dynamic range larger than 10 orders of magnitude). In addition, the searching space in the identification process could be limited by the pI ranges, which could increase the specificity of PSMs. Compared to the gel-based HiRIEF approach that is limited to a narrow (e.g., acidic) range of peptides (Branca et al., 2014), our method covered a much wider range of peptides including neutral and basic ones, although did not show a comparable resolution.

As discussed in the paper, our approach can be used to analyze PTMs that modify the chargeable groups of proteins. The PTMs that introduce pI changes could be due to protein damage. Accordingly, the changes of pI will shift the focusing positions of these molecules, moving them out of the shadow of other proteins that have extreme abundances (e.g., serum albumins). Therefore, pI shifting PTMs could be used as potential biomarkers of diseases related to protein damages, such as the Alzheimer's disease.

In Paper-IV, we tried another approach to decrease the dynamic range in the blood proteome by enriching polyclonal antibodies (i.e., IgGs). However, sequences of native human antibodies are of vast heterogeneity, due to the recombination, hypermutation, and PTMs. Consequently, it is practically not feasible to create a comprehensive reference database for identification of antibody-derived sequences. Therefore, we made an attempt to analyze blood antibodies using peptide *de novo* sequencing, which did not depend on the *a priori* knowledge about the protein sequences.

We found that *de novo* sequences revealed a hidden proteome that has almost the same size and similar properties (e.g., abundances) of the typical and searchable proteome. The

expansion of IgG-derived sequences rendered a subdomain of "IgGome" in the blood proteome. Utilizing the hidden proteome, we gained a high predictive power of differentiating blood samples from the patients with Alzheimer's disease and the patients with Dementia with Lewy Bodies.

In Paper-V, we demonstrated another implementation of *de novo* sequencing. We made a Diffacto analysis based on the iPRG-2015 data, where the chromatographic features were associated with *de novo* peptides and were subsequently attributed to the abstract sources of proteins by BLAST search against the universal SwissProt database. Even with such an identification procedure, we still obtained the high specificity in detecting all the spike-in proteins, as well as the accurate estimation of the relative protein abundances. Our results showed the usefulness of the quantification-centered data analysis, which provided a "spotlight" for conducting proteomics studies in the absence of "reference" databases.

Open source projects

Source codes and pipelines of the workflows developed for this thesis are freely available via GitHub.com.

DeMix: a workflow for identification of co-fragmenting peptides.

https://github.com/userbz/DeMix

DeMix-Q: a quantification-centered data analysis workflow.

https://github.com/userbz/DeMix-Q

Diffacto: finding differentially expressed protein by factor analysis.

https://github.com/statisticalbiotechnology/diffacto

Chapter Three: Concluding Remarks

In this thesis, I present the analytical methods developed during the four years of my Ph.D. study. The DeMix workflow started from improving the identification aspect of shotgun proteomics data analysis, which changes the paradigm of processing MS/MS data and uses the rich information from chimeric spectra to identify more than one peptide per spectrum. The expansion of peptide and protein identification from the co-fragmenting peptides increased the proteome coverage and dynamic range of data analysis.

In light of the successful implementation of DeMix, the quantitative extension of the workflow (DeMix-Q) addressed the reproducibility problem in label-free quantification. A new procedure for peptide identity propagation (PIP) has been introduced to rescue the missing peptide measurements. Based on the results of DeMix-Q, the idea of quantification-centered proteomics has been proposed for processing high-resolution shotgun proteomics data in large-scale.

In the practice of the quantification-centered proteomics, the Diffacto approach for proteome summarization was presented. By applying factor analysis, Diffacto seeks to differentiate between signals and noises from the covariation of peptides' abundances measured in multiple experiments. The covariation structure accurately reflects the protein concentration differences, which also provides the basis for a reliable quality control for quantifying differentially expressed proteins between biological conditions. Therefore, the false quantification rate has been reduced. In addition, with the released burden of identification, peptide *de novo* sequencing has been implemented to characterize complex proteomes, which is highly flexible in terms of protein identification and is no longer strictly limited to the reference sequence database.

In my opinion, the rapid development of high-resolution mass spectrometry will continue, and mass spectrometers will become faster, more sensitive, accurate, and cost-effective (Eliuk and Makarov, 2015). Accordingly, the scale of quantitative proteomics studies will increase significantly, providing the basis for quantification-centered proteomics. MS-

based shotgun proteomics will soon become a data-intensive research field, where more advanced bioinformatics methods can be implemented.

One the problem that has frequently been mentioned is the reproducibility issue caused by DDA. However, since the problem can be well-addressed by the proposed methods in this thesis, the boundary between DDA and DIA for MS/MS data acquisition might likely blur. It is possible to implement both DDA with DIA within the same experiment, and subsequently analyze the data using a universal processing workflow (Tsou et al., 2015). The Diffacto approach for proteome-wise summarization is, in principle, applicable to data of DIA data and targeted proteomics, which could be an interesting extension to the current study.

High-resolution MS offers unprecedented specificity for associating complex mass spectra with the identities of the analytes and simplifies the identification process. Also, the combination of complementary fragmentation techniques has become practical, such as EThcD that combines ETD with HCD within one fragmentation process (Frese et al., 2012). Such techniques may double the information contents in each MS/MS spectrum, and further increase the specificity of peptide-spectral matching. Using the complementary information in new algorithms may improve the accuracy of both peptide database search and *de novo* sequencing.

As shotgun proteomics studies expand in depth and sizes, the quantitative data structure resembles that of microarray-based transcriptomics. Therefore, many problems we encountered in proteomics, such as the demonstrated summarization problem, might already have practical solutions in transcriptomics. Compared to microarrays, modern mass spectrometers provide higher dynamic range and better quantitative accuracy, which should make transcriptomics-oriented algorithms easy to be implemented in shotgun proteomics.

While proteomics is becoming a data-driven research field, techniques of data science (e.g., machine learning) can be applied to extend our understandings of the complex properties and elusive mechanisms in proteomics, such as retention time, isoelectric point, enzymic cleavage, isotopic distribution, ionization efficiency, charge distribution, and gas-phase reactions. Integrating high-dimensional information into the analytical procedures will enhance the utility of proteomics data.

Acknowledgements

"No man is an island, entire of itself." Without love and support, on this lonely planet Earth, I would not survive the long march towards the truth. I need to thank a lot of people for making the five years of my life meaningful and enjoyable. Looking over this journey, it becomes a beautiful picture of everyone smiling to me in the glorious Swedish sunshine. But first of all, I owe the lovely country, the friendly city, and marvelous Karolinska Institutet a tremendous appreciation. Tack så mycket **Sverige**!

I would express my sincere gratitude towards Professor **Roman A. Zubarev**, my supervisor and mentor of science. His generosity allowed me to join the lab; his openness let me follow my way of doing research; and his enthusiasm for science affected me deeply, inspired me to think out of the box and try all the crazy ideas. Because of him, I love science more than ever. Спасибо, Роман Александрович!

My appreciation also goes to my co-supervisors. Professor **Lukas Käll** provided me great support and guidance in practicing bioinformatics. From all the discussions with him, I learned the way of statistical thinking, and I really enjoyed working with him. I would thank Professor **Erik L.L. Sonnhammer.** Although I have not been working closely with him, it is still my honor to have the support from such an outstanding computational biologist. Also, I would thank **Craig E. Wheelock** for being my mentor and always showing his support to me.

I appreciate the committee of my halftime review, Professors **Janne Lehtiö, Mikhail V. Gorshkov**, and **Lars Arvestad.** The three experts in proteomics and bioinformatics gave me valuable feedbacks and great suggestions that broadened my view of science and helped me proceeding my research.

I am grateful to all current and former members of Zubarev Lab and PK/KI core facility. Firstly, the senior Ph.D. students: **Yang Hongqian (Angie)**, **Mohammad Pirmoradian**, **Nataliya (Tarasova) Östberg,** and **Xie Xueshu**, who showed me by example how to survive the journey and enjoy it.

The senior researchers: **A. Jimmy Ytterberg**, **Akos Vegvari**, **Aleksandr Manoilov**, **Alexey Chernobrovkin**, **Juan Astorga-wells**, **Luciano Di Stefano**, **Massimiliano (Max) Gaetani**, **Susanna Lundström**, **Sergey Rodin**, as well as **Atim Enyenihi**, **Alexandra Bernadotte**, **Consuelo Marín-Vicente**, **David M. Good**, **Dorothea Rutishauser**, **Ernesto Gonzalez de Valdivia**, **Hao Piliang, Konstantin (Kostya) Chingin**, **Marta Guerrero-Valero**, **Shiva Kalantari**, **Ülkü Güler,** and **Yaroslav (Yar) Lyutvinskiy**, thanks for doing great science in the lab and providing enlightening comments and discussions.

Special thanks to **Carina Palmberg** and **Marie Ståhlberg** for keeping instruments and everything in the lab in its perfect status, and also for those great suggestions of Swedish restaurants. I would express my great gratitude towards **Victoria Balabanova**, for providing generous aid to me by dealing with all the administrative matters in a super-efficient manner. Also, thank **Gizella Bengtsson** for taking over the jobs during her leave.

Thanks to the junior fellows: **Amir Ata Saei**, **Pierre Sabatier, Tina Heyder**, **Wang Jijing (Janet)**, **Andrea Fossati, Harleen Dhot, Harshavardhan Budamgunta, Liban Abikar, Neil Rumachik, Niels M. Leijten, Marjan Abbasi, Muna Muse, Oleksii Rebrov,** and **Vincent Gemard,** for joining the journey with me towards the dreams of Ph.D.

I appreciate the friendship between the Zubarev Lab and the **Wheelock Lab**, especially, **Antonio (Toni) Checa, Cristina Gomez, Johan Kolmert, Shama Naz, David Balgoma, Daniel Sar, Stuart Snowden,** and **Ting Hsiu-Chi.** My gratitude also goes to other colleagues at MBB, **Ella Cederlund**, **Guo Jing, Peng Xiaoxiao, Ren Xiaoyuan, Tang Xiao, Tomas Bergman**, and **Xu Jianqiang**. It was always a pleasure to have the little chatting about everyday life.

Thanks to **Alessandra Nanni**, Prof. **Sten Linnarsson,** and Prof. **Elias Arnér,** for guiding the doctoral education; **Chad Tunell,** for helping with IT stuffs; the **Lunch-seminar** committee, for orgainzing free sandwiches and good talks; **Olof Rådmark, Susie Björkholm, Åke Rökaeus,** and **Jan-Olov Höög,** for organizing teaching activities at the Department of MBB.

Thanks to **Chi Hao**, **Hassan Foroughi, Linus Östberg, Liu Yansheng, Matthew The,** and **Wen Bo**, for the discussions about bioinformatics, proteomics, and life.

Great appreciations go to the organizers, sponsors and regular participants of the **PathProt** forum (Oeiras, Portugal): **Pedro L. Fernandes, Alexander** & **Olga Kel, Carlos Cordeiro, Marta Sousa Silva, António Ferreira,** and **Mikhail M. Savitski.**

I appreciate **Thermo** Scientific (Bremen), **Alexander Makarov**, **Christoph Henrich**, and **Carmen Paschke**, for showing me the awesomeness behind the Orbitrap.

I attended the **MaxQuant** summer school, twice (München and Oxford), and learned from Prof. **Matthias Mann**, Prof. **Jürgen Cox**, and **Marco Hein**.

Thanks to the organizers and participants of the Advanced Scientific Programming in Python (**ASPP**) summer school at the University of Reading, I enjoyed the amazing event.

I learned from Professors **Joshua J. Coon, Stephen Gygi**, and **Michael MacCoss** at the **ASMS** conferences (Baltimore and St. Louis). I appreciate the travel grant from the **HUPO** Congress (Taipei). I am grateful to the cabin crews of **Cathay Pacific** who saved my passport at Taipei Taoyuan airport.

I am indebted to **Yang Honqian, Xie Xueshu** & **Alan Shaw, Xu Shaohua** & **Zhang Lu**, **Ye Xiaofei**, **Wang Jijing, Wei Guifeng, Chen Maoshan** & **Wang Tingting, Kostya, Mohammad & Mahya, Nataliya & Linus, Susanna, Juan, Alexey, Toni, Amir,** and **Pirrere,** who befriended me and involved me with all the exciting events.

I would also express my gratitude to **Xu Peiyuan, Yu Nan, Liu Yilin, Isabel Li,** and **Alicja Kwiatkowski,** who offered warm companionships.

I thank my parents who give me life, strength, and wisdom, their love allowed me to move forward without fear. I am grateful to my parents-in-law for their understanding and blessings. I appreciate the good wishes from my grandma, aunties, sister and brothers, as well as my childhood friends. （感谢妈妈爸爸赋予我生命、力量和智慧，给与我至高无上的爱，令我无所顾忌，勇往直前。感谢岳父岳母的关心、理解和祝福。感谢奶奶、姑姑、妹妹、哥哥们的牵挂。谢谢童心未改的伙伴们：彭瑞、周乾康、冯乾、吴诚昊、郑泽南、吕泌志）

Finally, I would express my deepest gratitude to my dearest, **Miao Yinghan**, my lovely gorgeous brilliant wife. Now I realize how brave you were and how much sacrifice you had made to leave the tropical summer and follow me to this frozen land. You are my sunshine. Thank you, for everything!

♥

References

Aebersold, R. & Mann, M. 2003. Mass spectrometry-based proteomics. *Nature,* 422**,** 198-207. doi:10.1038/nature01511

Aebersold, R. & Mann, M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature,* 537**,** 347-355. doi:10.1038/nature19949

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology,* 215**,** 403-10. doi:10.1016/S0022-2836(05)80360-2

Bandeira, N., Pham, V., Pevzner, P., Arnott, D. & Lill, J. R. 2008. Automated de novo protein sequencing of monoclonal antibodies. *Nature biotechnology,* 26**,** 1336-1338. doi:10.1038/nbt1208-1336

Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. 2012. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem,* 404**,** 939-65. doi:10.1007/s00216-012-6203-4

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. 2007. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry,* 389**,** 1017-31. doi:10.1007/s00216-007-1486-6

Bateman, N. W., Goulding, S. P., Shulman, N. J., Gadok, A. K., Szumlinski, K. K., Maccoss, M. J. & Wu, C. C. 2013. Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA). *Molecular & Cellular Proteomics,* 13**,** 329-338. doi:10.1074/mcp.M112.026500

Biemann, K. 1988. Contributions of mass spectrometry to peptide and protein structure. *Biomed Environ Mass Spectrom,* 16**,** 99-111.

Branca, R. M. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff, Å., Forshed, J., Käll, L. & Lehtiö, J. 2014. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature Methods,* 11**,** 59-62. doi:10.1038/nmeth.2732

Cappadona, S., Baker, P. R., Cutillas, P. R., Heck, A. J. R. & Van Breukelen, B. 2012. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids,* 43**,** 1087-1108. doi:10.1007/s00726-012-1289-8

Catherman, A. D., Skinner, O. S. & Kelleher, N. L. 2014. Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun,* 445**,** 683-93. doi:10.1016/j.bbrc.2014.02.041

Chi, H., Chen, H., He, K., Wu, L., Yang, B., Sun, R. X., Liu, J., Zeng, W. F., Song, C. Q., He, S. M. & Dong, M. Q. 2013. PNovo+: De novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *Journal of proteome research,* 12**,** 615-625. doi:10.1021/pr3006843

Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L. & Gygi, S. P. 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol,* 33**,** 743-9. doi:10.1038/nbt.3267

Choi, M., Eren-Dogu, Z. F., Colangelo, C., Cottrell, J., Hoopmann, M. R., Kapp, E. A., Kim, S., Lam, H., Neubert, T. A., Palmblad, M., Phinney, B. S., Weintraub, S. T., Maclean, B. & Vitek, O. 2017. ABRF Proteome Informatics Research Group (iPRG) 2015 Study: Detection of Differentially Abundant Proteins in Label-Free Quantitative LC-MS/MS Experiments. *J Proteome Res,* 16**,** 945-957. doi:10.1021/acs.jproteome.6b00881

Clough, T., Thaminy, S., Ragg, S., Aebersold, R. & Vitek, O. 2012. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC bioinformatics,* 13 Suppl 16**,** S6. doi:10.1186/1471-2105-13-S16-S6

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. & Mann, M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics,* 13**,** 2513-26. doi:10.1074/mcp.M113.031591

Cox, J. & Mann, M. 2007. Is proteomics the new genomics? *Cell,* 130**,** 395-8. doi:10.1016/j.cell.2007.07.032

Cox, J. & Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology,* 26**,** 1367-1372. doi:10.1038/nbt.1511

Cox, J. & Mann, M. 2011. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annual Review of Biochemistry,* 80**,** 273-299. doi:10.1146/annurev-biochem-061308-093216

Cox, J., Michalski, A. & Mann, M. 2011a. Software lock mass by two-dimensional minimization of peptide mass errors. *Journal of the American Society for Mass Spectrometry,* 22**,** 1373-1380. doi:10.1007/s13361-011-0142-8

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. 2011b. Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of proteome research,* 10**,** 1794-1805. doi:10.1021/pr101065j

Crick, F. 1970. Central dogma of molecular biology. *Nature,* 227**,** 561-3.

Degroeve, S. & Martens, L. 2013. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics,* 29**,** 3199-203. doi:10.1093/bioinformatics/btt544

Egertson, J. D., Kuehn, A., Merrihew, G. E., Bateman, N. W., Maclean, B. X., Ting, Y. S., Canterbury, J. D., Marsh, D. M., Kellmann, M., Zabrouskov, V., Wu, C. C. & Maccoss, M. J. 2013. Multiplexed MS/MS for improved data-independent acquisition. *Nature Methods,* 10**,** 744-746. doi:10.1038/nmeth.2528

Elias, J. E. & Gygi, S. P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods,* 4**,** 207-14. doi:10.1038/nmeth1019

Eliuk, S. & Makarov, A. 2015. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annual Review of Analytical Chemistry,* 8**,** 61-80. doi:10.1146/annurev-anchem-071114-040325

Eng, J. K., Mccormack, A. L. & Yates, J. R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom,* 5**,** 976-89. doi:10.1016/1044-0305(94)80016-2

Forshed, J., Johansson, H. J., Pernemalm, M., Branca, R. M., Sandberg, A. & Lehtio, J. 2011. Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ). *Mol Cell Proteomics,* 10**,** M111 010264. doi:10.1074/mcp.M111.010264

Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A. & Pevzner, P. A. 2007. De novo peptide sequencing and identification with precision mass spectrometry. *Journal of proteome research,* 6**,** 114-23. doi:10.1021/pr060271u

Frese, C. K., Altelaar, A. F., Van Den Toorn, H., Nolting, D., Griep-Raming, J., Heck, A. J. & Mohammed, S. 2012. Toward full peptide sequence coverage by dual fragmentation combining electron-transfer and higher-energy collision dissociation tandem mass spectrometry. *Anal Chem,* 84**,** 9668-73. doi:10.1021/ac3025366

Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. & Mann, M. 2010. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods,* 7**,** 383-5. doi:10.1038/nmeth.1446

Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. 2012. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics,* 11**,** M111.014050-M111.014050. doi:10.1074/mcp.M111.014050

Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci U S A,* 100**,** 6940-5. doi:10.1073/pnas.0832254100

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., Bonner, R. & Aebersold, R. 2012. 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,* 11**,** O111.016717-O111.016717. doi:10.1074/mcp.O111.016717

Granholm, V., Kim, S., Navarro, J. C. F., Sjölund, E., Smith, R. D. & Käll, L. 2014. Fast and accurate database searches with MS-GF+percolator. *Journal of proteome research,* 13**,** 890-897. doi:10.1021/pr400937n

Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., Koziol, J. A. & Schnitzer, J. E. 2010. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol,* 28**,** 83-9. doi:10.1038/nbt.1592

Guo, X., Trudgian, D. C., Lemoff, A., Yadavalli, S. & Mirzaei, H. 2014. Confetti: A Multi-protease Map of the HeLa Proteome for Comprehensive Proteomics. *Molecular & cellular proteomics : MCP,* 13**,** 1573-1584. doi:10.1074/mcp.M113.035170

Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. 1999. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology,* 19**,** 1720-1730.

Halligan, B. D. 2009. ProMoST: a tool for calculating the pI and molecular mass of phosphorylated and modified proteins on two-dimensional gels. *Methods Mol Biol,* 527**,** 283-98, ix. doi:10.1007/978-1-60327-834-8_21

Hebert, A. S., Merrill, A. E., Bailey, D. J., Still, A. J., Westphall, M. S., Strieter, E. R., Pagliarini, D. J. & Coon, J. J. 2013. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature Methods,* 10**,** 332-4. doi:10.1038/nmeth.2378

Hein, M. Y., Sharma, K., Cox, J. & Mann, M. 2012. Proteomic analysis of cellular systems. *Handbook of Systems Biology***,** 3-25.

Hochreiter, S., Clevert, D. A. & Obermayer, K. 2006. A new summarization method for Affymetrix probe level data. *Bioinformatics,* 22**,** 943-9. doi:10.1093/bioinformatics/btl033

Huang, T., Wang, J., Yu, W. & He, Z. 2012. Protein inference: A review. *Briefings in Bioinformatics,* 13**,** 586-614. doi:10.1093/bib/bbs004

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J. & Mann, M. 2005. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics,* 4**,** 1265-72. doi:10.1074/mcp.M500061-MCP200

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & Maccoss, M. J. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods,* 4**,** 923-925. doi:10.1038/nmeth1113

Käll, L., Storey, J. D. & Noble, W. S. 2008. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinformatics,* 24**,** 42-48. doi:10.1093/bioinformatics/btn294

Karpievitch, Y. V., Dabney, A. R. & Smith, R. D. 2012. Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics,* 13 Suppl 1**,** S5-S5. doi:10.1186/1471-2105-13-S16-S5

Kelleher, N. L. 2004. Top-down proteomics. *Anal Chem,* 76**,** 197A-203A.

Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M. & Kruglyak, L. 2009. Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 15544-15548. doi:10.1073/pnas.0904100106

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., et al. 2014. A draft map of the human proteome. *Nature,* 509**,** 575-81. doi:10.1038/nature13302

Kim, S. & Pevzner, P. A. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun,* 5**,** 5277-5277. doi:10.1038/ncomms6277

Kitano, H. 2002. Systems biology: a brief overview. *Science (New York, N.Y.),* 295**,** 1662-1664. doi:10.1126/science.1069492

Knudsen, G. M. & Chalkley, R. J. 2011. The effect of using an inappropriate protein database for proteomic data analysis. *PloS one,* 6**,** e20873-e20873. doi:10.1371/journal.pone.0020873

Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., Pfeifer, N., Schulz-Trieglaff, O. & Sturm, M. 2007. TOPP--the OpenMS proteomics pipeline. *Bioinformatics,* 23**,** e191-7. doi:10.1093/bioinformatics/btl299

Lam, H. & Aebersold, R. 2011. Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics. *Methods (San Diego, Calif.),* 54**,** 424-31. doi:10.1016/j.ymeth.2011.01.007

Lander, E. S., Consortium, I. H. G. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921. doi:10.1038/35057062

Leitner, A. & Aebersold, R. 2013. SnapShot: mass spectrometry for protein and proteome analyses. *Cell,* 154**,** 252-252 e1. doi:10.1016/j.cell.2013.06.025

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. & Yates, J. R., 3rd 1999. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol,* 17**,** 676-82. doi:10.1038/10890

Liu, H., Sadygov, R. G. & Yates, J. R., 3rd 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem,* 76**,** 4193-201. doi:10.1021/ac0498563

Liu, Y., Beyer, A. & Aebersold, R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell,* 165**,** 535-50. doi:10.1016/j.cell.2016.03.014

Lukasse, P. N. J. & America, A. H. P. 2014. Protein Inference Using Peptide Quantification Patterns. *Journal of proteome research,* 13**,** 3191-3199. doi:10.1021/pr401072g

Lyutvinskiy, Y., Yang, H., Rutishauser, D. & Zubarev, R. A. 2013. In silico instrumental response correction improves precision of label-free proteomics and accuracy of proteomics-based predictive models. *Mol Cell Proteomics,* 12**,** 2324-31. doi:10.1074/mcp.O112.023804

Ma, B. 2015. Novor: real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry,* 26**,** 1885-94.

Ma, B. & Johnson, R. 2012. De Novo Sequencing and Homology Searching. *Molecular & Cellular Proteomics,* 11**,** O111.014902-O111.014902. doi:10.1074/mcp.O111.014902

Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., Schilling, B., Drake, P. M., Gibson, B. W. & Tabb, D. L. 2009. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of proteome research,* 8**,** 3872-3881. doi:10.1021/pr900360j

Makarov, A. 2000. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem,* 72**,** 1156-62.

Mallick, P. & Kuster, B. 2010. Proteomics: a pragmatic perspective. *Nature biotechnology,* 28**,** 695-709. doi:10.1038/nbt.1658

Mann, M. 2006. Functional and quantitative proteomics using SILAC. *Nature Reviews Molecular Cell Biology,* 7**,** 952-958. doi:10.1038/nrm2067

Mann, M. & Wilm, M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem,* 66**,** 4390-4399. doi:10.1021/ac00096a002

Marcotte, E. M. 2007. How do shotgun proteomics algorithms identify proteins? *Nature biotechnology,* 25**,** 755-7. doi:10.1038/nbt0707-755

Matzke, M. M., Brown, J. N., Gritsenko, M. A., Metz, T. O., Pounds, J. G., Rodland, K. D., Shukla, A. K., Smith, R. D., Waters, K. M., Mcdermott, J. E. & Webb-Robertson, B. J. 2013. A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics,* 13**,** 493-503. doi:10.1002/pmic.201200269

Mcalister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D. & Gygi, S. P. 2012. Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem,* 84**,** 7469-7478. doi:10.1021/ac301572t

Mclafferty, F. W. 1981. Tandem mass spectrometry. *Science,* 214**,** 280-7.

Mcluckey, S. A. 1992. Principles of Collisional Activation in Analytical Mass-Spectrometry. *Journal of the American Society for Mass Spectrometry,* 3**,** 599-614. doi:Doi 10.1016/1044-0305(92)85001-Z

Merrill, A. E., Hebert, A. S., Macgilvray, M. E., Rose, C. M., Bailey, D. J., Bradley, J. C., Wood, W. W., El Masri, M., Westphall, M. S., Gasch, A. P. & Coon, J. J. 2014. NeuCode labels for relative protein quantification. *Molecular & cellular proteomics : MCP,* 13**,** 2503-12. doi:10.1074/mcp.M114.040287

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., Huang, K. L., et al. 2016. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature,* 534**,** 55-62. doi:10.1038/nature18003

Michalski, A., Cox, J. & Mann, M. 2011a. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res,* 10**,** 1785-93. doi:10.1021/pr101060v

Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M. & Horning, S. 2011b. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics,* 10**,** M111 011015. doi:10.1074/mcp.M111.011015

Moruz, L., Hoopmann, M. R., Rosenlund, M., Granholm, V., Moritz, R. L., Käll, L. & Ka, L. 2013. Mass fingerprinting of complex mixtures: Protein inference from high-resolution peptide masses and predicted retention times. *Journal of proteome research,* 12**,** 5730-5741. doi:10.1021/pr400705q

Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O. & Mann, M. 2012. System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Molecular & Cellular Proteomics,* 11**,** M111.013722-M111.013722. doi:10.1074/mcp.M111.013722

Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. 2003. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. *Anal Chem,* 75**,** 4646-4658. doi:10.1021/ac0341261

Nesvizhskii, A. I., Vitek, O. & Aebersold, R. 2007. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods,* 4**,** 787-797. doi:10.1038/nmeth1088

Nielsen, M. L., Savitski, M. M. & Zubarev, R. A. 2005. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Molecular & cellular proteomics : MCP,* 4**,** 835-45. doi:10.1074/mcp.T400022-MCP200

Olsen, J. V., De Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S. & Mann, M. 2005. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular & cellular proteomics : MCP,* 4**,** 2010-2021. doi:10.1074/mcp.T500030-MCP200

Olsen, J. V., Ong, S.-E. & Mann, M. 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics : MCP,* 3**,** 608-614. doi:10.1074/mcp.T400003-MCP200

Ong, S. E. & Mann, M. 2005. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol,* 1**,** 252-62. doi:10.1038/nchembio736

Palzs, B. & Suhal, S. 2005. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews,* 24**,** 508-548. doi:10.1002/mas.20024

Pasa-Tolic, L., Masselon, C., Barry, R. C., Shen, Y. F. & Smith, R. D. 2004. Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques,* 37**,** 621-+.

Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20**,** 3551-67.

Pounds, S. & Cheng, C. 2006. Robust estimation of the false discovery rate. *Bioinformatics,* 22**,** 1979-87. doi:10.1093/bioinformatics/btl328

Rauniyar, N. & Yates, J. R. 2014. Isobaric Labeling-Based Relative Quantification in Shotgun Proteomics. *Journal of proteome research,* 13**,** 5293-5309. doi:10.1021/pr500880b

Richards, A. L., Hebert, A. S., Ulbrich, A., Bailey, D. J., Coughlin, E. E., Westphall, M. S. & Coon, J. J. 2015. One-hour proteome analysis in yeast. *Nature Protocols,* 10**,** 701-714. doi:10.1038/nprot.2015.040

Roepstorff, P. & Fohlman, J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom,* 11**,** 601. doi:10.1002/bms.1200111109

Rose, C. M., Merrill, A. E., Bailey, D. J., Hebert, A. S., Westphall, M. S. & Coon, J. J. 2013. Neutron encoded labeling for peptide identification. *Anal Chem,* 85**,** 5129-37. doi:10.1021/ac400476w

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. & Pappin, D. J. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics,* 3**,** 1154-69. doi:10.1074/mcp.M400129-MCP200

Sandin, M., Teleman, J., Malmström, J. & Levander, F. 2014. Data processing methods and quality control strategies for label-free LC-MS protein quantification. *Biochimica et biophysica acta,* 1844**,** 29-41. doi:10.1016/j.bbapap.2013.03.026

Sandve, G. K., Ferkingstad, E. & Nygard, S. 2011. Sequential Monte Carlo multiple testing. *Bioinformatics,* 27**,** 3235-41. doi:10.1093/bioinformatics/btr568

Savitski, M. M., Kjeldsen, F., Nielsen, M. L. & Zubarev, R. A. 2006a. Complementary Sequence Preferences of Electron-Capture Dissociation and Vibrational Excitation in Fragmentation of Polypeptide Polycations. *Angewandte Chemie International Edition,* 45**,** 5301-5303. doi:10.1002/anie.200601240

Savitski, M. M., Nielsen, M. L., Kjeldsen, F. & Zubarev, R. A. 2005. Proteomics-Grade de Novo Sequencing Approach. *Journal of proteome research,* 4**,** 2348-2354. doi:10.1021/pr050288x

Savitski, M. M., Nielsen, M. L. & Zubarev, R. A. 2006b. ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of

modifications, and fingerprinting complex protein mixtures. *Molecular & cellular proteomics : MCP,* 5**,** 935-948. doi:10.1074/mcp.T500034-MCP200

Savitski, M. M., Reinhard, F. B., Franken, H., Werner, T., Savitski, M. F., Eberhard, D., Martinez Molina, D., Jafari, R., Dovega, R. B., Klaeger, S., Kuster, B., Nordlund, P., Bantscheff, M. & Drewes, G. 2014. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science,* 346**,** 1255784. doi:10.1126/science.1255784

Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. 2015. A Scalable Approach for Protein False Discovery Rate Estimation in Large Proteomic Data Sets. *Mol Cell Proteomics,* 14**,** 2394-404. doi:10.1074/mcp.M114.046995

Schilling, B., Rardin, M. J., Maclean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., Sorensen, D. J., Bereman, M. S., Jing, E., Wu, C. C., Verdin, E., Kahn, C. R., Maccoss, M. J. & Gibson, B. W. 2012. Platform-independent and Label-free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline: APPLICATION TO PROTEIN ACETYLATION AND PHOSPHORYLATION. *Molecular & Cellular Proteomics,* 11**,** 202-214. doi:10.1074/mcp.M112.017707

Schroeder, H. W., Jr. & Cavacini, L. 2010. Structure and function of immunoglobulins. *J Allergy Clin Immunol,* 125**,** S41-52. doi:10.1016/j.jaci.2009.09.046

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. 2011. Global quantification of mammalian gene expression control. *Nature,* 473**,** 337-42. doi:10.1038/nature10098

Serang, O. & Käll, L. 2015. Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less. *J Proteome Res,* 14**,** 4099-103. doi:10.1021/acs.jproteome.5b00568

Serang, O., Maccoss, M. J. & Noble, W. S. 2010. Efficient Marginalization to Compute Protein Posterior Probabilities from Shotgun Mass Spectrometry Data research articles. *Journal of proteome research,* 9**,** 5346-5357. doi:10.1021/pr100594k

Serang, O. & Noble, W. 2012. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface,* 5**,** 3-20.

Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L. & Deutsch, E. W. 2013. Combining Results of Multiple Search Engines in Proteomics. *Molecular & Cellular Proteomics,* 12**,** 2383-2393. doi:10.1074/mcp.R113.027797

Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. & Geromanos, S. J. 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics,* 5**,** 144-56. doi:10.1074/mcp.M500230-MCP200

Smith, R. D., Anderson, G. A., Lipton, M. S., Masselon, C., Pasa-Tolic, L., Shen, Y. & Udseth, H. R. 2002. The use of accurate mass tags for high-throughput microbial proteomics. *Omics : a journal of integrative biology,* 6**,** 61-90. doi:10.1089/15362310252780843

Swaney, D. L., Wenger, C. D. & Coon, J. J. 2010. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res,* 9**,** 1323-9. doi:10.1021/pr900863u

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. 2004. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 9528-33. doi:10.1073/pnas.0402700101

Tabb, D. L., Wang, X., Carr, S. A., Clauser, K. R., Mertins, P., Chambers, M. C., Holman, J. D., Wang, J., Zhang, B., Zimmerman, L. J., Chen, X., Gunawardena, H. P., Davies, S. R., Ellis, M. J., Li, S., Townsend, R. R., Boja, E. S., Ketchum, K. A., Kinsinger, C. R., Mesri, M., et al. 2016. Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *J Proteome Res,* 15**,** 691-706. doi:10.1021/acs.jproteome.5b00859

Tabb, D. L., Wang, X., Carr, S. A., Clauser, K. R., Mertins, P., Chambers, M. C., Holman, J. D., Wang, J., Zhang, B., Zimmerman, L. J., Chen, X., Gunawardena, H. P., Davies, S. R., Ellis, M. J. C., Li, S.,

Townsend, R. R., Boja, E. S., Ketchum, K. A., Kinsinger, C. R., Mesri, M., et al. 2015. Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *Journal of proteome research*. doi:10.1021/acs.jproteome.5b00859

Talloen, W., Clevert, D. A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S. & Gohlmann, H. W. 2007. I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics,* 23**,** 2897-902. doi:10.1093/bioinformatics/btm478

Taylor, J. A. & Johnson, R. S. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom,* 11**,** 1067-75.

Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J. & Mann, M. 2011. Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & cellular proteomics : MCP,* 10**,** M110.003699-M110.003699. doi:10.1074/mcp.M110.003699

Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K. & Hamon, C. 2003. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem,* 75**,** 1895-904.

Thomson, J. J. 1913. *Rays of positive electricity and their application to chemical analyses*, Longmans, Green and Company.

Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C. & Nesvizhskii, A. I. 2015. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods,* 12**,** 258-64, 7 p following 264. doi:10.1038/nmeth.3255

Tyanova, S., Temu, T. & Cox, J. 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc,* 11**,** 2301-2319. doi:10.1038/nprot.2016.136

Uniprot, C. 2015. UniProt: a hub for protein information. *Nucleic Acids Res,* 43**,** D204-12. doi:10.1093/nar/gku989

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., et al. 2001. The sequence of the human genome. *Science,* 291**,** 1304-51. doi:10.1126/science.1058040

Vincent, C. E., Potts, G. K., Ulbrich, A., Westphall, M. S., Atwood, J. A., 3rd, Coon, J. J. & Weatherly, D. B. 2013. Segmentation of precursor mass range using "tiling" approach increases peptide identifications for MS1-based label-free quantification. *Anal Chem,* 85**,** 2825-32. doi:10.1021/ac303352n

Walther, T. C. & Mann, M. 2010. Mass spectrometry-based proteomics in cell biology. *J Cell Biol,* 190**,** 491-500. doi:10.1083/jcb.201004052

Wang, J., Ma, Z., Carr, S. A., Mertins, P., Zhang, H., Zhang, Z., Chan, D. W., Ellis, M. J. C., Townsend, R. R., Smith, R. D., Mcdermott, J. E., Chen, X., Paulovich, A. G., Boja, E. S., Mesri, M., Kinsinger, C. R., Rodriguez, H., Rodland, K. D., Liebler, D. C. & Zhang, B. 2017. Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction. *Molecular & Cellular Proteomics,* 16**,** 121-134. doi:10.1074/mcp.M116.060301

Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., Sturm, M., Kenar, E., Kohlbacher, O., Aebersold, R. & Malmström, L. 2013. An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *Journal of proteome research*. doi:10.1021/pr300992u

Wells, J. M. & Mcluckey, S. A. 2005. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol,* 402**,** 148-85. doi:10.1016/S0076-6879(05)02005-7

Wenger, C. D. & Coon, J. J. 2013. A Proteomics Search Algorithm Speci fi cally Designed for High-Resolution Tandem Mass Spectra.

Wenger, C. D., Phanstiel, D. H., Lee, M. V., Bailey, D. J. & Coon, J. J. 2011. COMPASS: a suite of pre- and post-search proteomics software tools for OMSSA. *Proteomics,* 11**,** 1064-74. doi:10.1002/pmic.201000616

Werner, T., Sweetman, G., Savitski, M. F., Mathieson, T., Bantscheff, M. & Savitski, M. M. 2014. Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal Chem,* 86**,** 3594-601. doi:10.1021/ac500140s

Weston, A. D. & Hood, L. 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res,* 3**,** 179-96.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature,* 509**,** 582-7. doi:10.1038/nature13319

Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. 2014. A ' proteomic ruler ' for protein copy number and concentration estimation without spike - in standards. doi:10.1074/mcp.M113.037309

Yen, C.-Y., Houel, S., Ahn, N. G. & Old, W. M. 2011. Spectrum-to-spectrum searching using a proteome-wide spectral library. *Molecular & cellular proteomics : MCP,* 10**,** M111.007666-M111.007666. doi:10.1074/mcp.M111.007666

Zhang, X., Fang, A., Riley, C. P., Wang, M., Regnier, F. E. & Buck, C. 2010. Multi-dimensional liquid chromatography in proteomics--a review. *Anal Chim Acta,* 664**,** 101-13. doi:10.1016/j.aca.2010.02.001

Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R. M., Orre, L. M. & Lehtio, J. 2014. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol Cell Proteomics,* 13**,** 1552-62. doi:10.1074/mcp.M113.031203

Zubarev, R. & Mann, M. 2007. On the proper use of mass accuracy in proteomics. *Molecular & cellular proteomics : MCP,* 6**,** 377-381. doi:10.1074/mcp.M600380-MCP200

Zubarev, R. A. 2013. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics,* 13**,** 723-6. doi:10.1002/pmic.201200451

Zubarev, R. A., Kelleher, N. L. & Mclafferty, F. W. 1998. Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society,* 120**,** 3265-3266. doi:10.1021/ja973478k

Zubarev, R. A. & Makarov, A. 2013. Orbitrap mass spectrometry. *Anal Chem,* 85**,** 5288-96. doi:10.1021/ac4001223

Zubarev, R. A., Zubarev, A. R. & Savitski, M. M. 2008. Electron Capture/Transfer versus Collisionally Activated/Induced Dissociations: Solo or Duet? *Journal of the American Society for Mass Spectrometry,* 19**,** 753-761. doi:10.1016/j.jasms.2008.03.007