2-2016

A flexible and versatile framework for statistical design and analysis of quantitative mass spectrometry-based proteomic experiments

Meena Choi
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Meena Choi

Entitled

A FLEXIBLE AND VERSATILE FRAMEWORK FOR STATISTICAL DESIGN AND ANALYSIS OF QUANTITATIVE MASS SPECTROMETRY-BASED PROTEOMIC EXPERIMENTS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Olga Vitek
Co-chair

Hyonho Chun
Co-chair

Bruce Craig

Yu Michael Zhu

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Hyonho Chun

Approved by: Hao Zhang 2/26/2016

Head of the Departmental Graduate Program Date

A FLEXIBLE AND VERSATILE FRAMEWORK FOR STATISTICAL DESIGN AND ANALYSIS OF QUANTITATIVE MASS SPECTROMETRY-BASED PROTEOMIC EXPERIMENTS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Meena Choi

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2016

Purdue University

West Lafayette, Indiana

ACKNOWLEDGMENTS

I owe Dr. Olga Vitek a real debt of gratitude. Dr. Vitek gave me the first chance to research in the field of biology. Dr. Vitek introduced me to MS-based proteomics, gave me the chance to meet and work with some of the best biologists around the world, and provided me the opportunity to be an instructor in many short courses. Dr. Vitek taught me how to approach research problems with a scientific view and how to communicate with collaborators. After five and half years, I have built invaluable experience and built a network with the best experimentalists around the world; it became my treasure. Also, Dr. Vitek advised and encouraged me professionally and personally in moments of difficulty. Dr. Vitek patiently waited for me to overcome adversity and reach the next level. Her tremendous support helped me achieve the doctoral journey.

I also want to specially thank Dr. Hyonho Chun for her great support during my PhD program. Dr. Chun has given me priceless advice on how to build a career and how to balance my work and life. Dr. Chun also supported me through difficult personal moments.

Special thanks to my committee members, Dr. Bruce Chaig and Dr. Michael Zhu for their very helpful input and suggestions. Their guidance has served me well. Without the guidance of my committee members I would not have been able to complete this dissertation.

Special thank to Dr. Ruedi Aebersold at ETH Zürich for his generous support. Dr. Aebersold made it possible to visit Zürich four times. His supportive and previsional research advise led me to have interest in MS proteomics.

My sincere appreciation goes to all members in Vitek's lab at Purdue University, Dr. Tim Clough, Dr. Ching-Yun(Veavi) Chang, Dr. Danni Yu, Mike Cheng, Kyle Bemis, April Harry, Robert Ness and at Northeastern University, Dr. Tsung-Heng Tsai, Dr. Cyril Galitzine, Dr. Zeynep Filiz, Dr. Eralp Dogu, and Ting Huang for their friendship and assistance.

My sincere appreciation also goes to all my collaborators. Thanks to all the former and current members in Aebersold's lab, particularly my dear collaborators, Dr. Silvia Surinova and Dr. Ruth Hüttenhain and organizers of several courses for Proteomics, Dr. Christina Ludwig, Dr. Olga Schubert, and Dr. Ariel Bensimon for introducing me to proteomics experiments and providing invaluable discussions and collaborations. Thanks to Dr. Eduard Sabidó, Dr. Eva Borràs at the Proteomics Unit of Center for Genomic Regulation In Spain for their amazing projects, discussion, insights in proteomics, and the chance to be an instructor in a short course in Spain. Thank you, Dr. Bernd Wollescheid in ETH Zürich , Dr. Ferdinando Cerciello, Dr. Maria Pavlov and Nadine Sobotzki for their kind help with discussions about various projects and collaborations. All of these individuals, the best biologists in proteomics whom possess deep insight, gave me the vision necessary to apply statistics to this research area.

My big appreciation to Brendan MacLean, Yuval Boss, Nick Shulman in Dr. Mike MacCoss's lab at University of Washington. I am grateful for their tireless efforts in many collaborations. I appreciate the help with computational insights and for informing others about our tools within the proteomics community.

Many thanks to Dr. Erik Verschueren in the Krogan lab at UCSF School of Medicine for all his help, suggestions, discussions and feedback about my research and our tools.

Lastly, thank my parents, Jong-Yi Choi and Yeong-Ran Kim, for their endless love and support. Their love helped me to finish this rough and long journey. Thank you to my sister and best friend, Yoona Choi, and brother-in-law, Sergio Daniel Rey-

Silva, who have been with me and supported my studies in US. Thank you to my lovely nephew, Daniel Jeean Rey, who gives me the sweetest happiness and love.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

ABBREVIATIONS

MS	Mass Spectrometry
DDA	Data-Dependent Acquisition
SRM	Selected Reaction Monitoring for quantitative proteomics
DIA	Data-Independent Acquisition
SWATH	Sequential Window Acquisition of all Theoretical fragment-ion spectra
m/z	the ratio of mass to charge
MLE	Maximum Likelihood Estimator
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

ABSTRACT

Choi, Meena PhD, Purdue University, May 2016. A Flexible and Versatile Framework For Statistical Design and Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. Major Professor: Olga Vitek.

Quantitative mass spectrometry (MS)-based proteomics is an indispensable technology for biological and clinical research. As the proteomics field grows, MS-based proteomic workflows are becoming more complex and diverse. The accuracy and the throughput of the MS measurements and of the signal processing tools dramatically increased. However, many existing statistical tools and workflows have not followed the technological development. Therefore, there is a need for flexible statistical tools, which reflect diverse and complex workflows, are computationally efficient for large datasets, and maximize the reproducibility of the results.

We propose a family of linear mixed effects models, and a split-plot view of the experimental design, that represent measurements from quantitative mass spectrometry-based proteomics. The whole plot part of the design reflects the structure of the biological variation of the experiment, such as case-control design, paired design, or time-course design. The subplot part of the design reflects the structure of the technological variation, such as fragmentation patterns, labeling strategy, and presence of multiple peptides per protein. We propose an estimation procedure that separately estimates the parameters of the subplot and the whole plot parts of the design, to maximize the flexibility of the model, increase the speed of the analysis, and facilitate the interpretation.

The proposed modeling framework was validated using 10 controlled mixtures and 10 experimental datasets from targeted Selected Reaction Monitoring (SRM), Data-Dependent Acquisition (DDA or shotgun), and Data-Independent Acquisition (DIA

or SWATH-MS), where signals were extracted with mutiple signal processing tools. We implemented the proposed method in the software package MSstats, which checks the correctness of the user input, recognizes arbitrary complex experimental design, visualizes the data and performs statistical modeling and inference. It is interoperable with other existing computational tools such as Skyline.

1. PROBLEM STATEMENT AND CONTRIBUTIONS

1.1 Statement of the problem

1.1.1 Statement of the biotechnological problem

Mass spectrometry(MS)-based proteomics is an indispensable tool for molecular and cellular biology and systems biology. It deals with large-scale characterization of protein composition and abundance of complex biological mixtures. In recent years MS proteomics workflows have become more complex and diverse, in terms of (1) experimental design, which now often includes complex comparisons of experimental conditions or times, (2) experimental technology, which now offers alternative modes of spectral acquisition such as targeted Selected Reaction Monitoring (SRM) [1], Data-Dependent Acquisition (DDA or shotgun) [2, 3], and Data-Independent Acquisition (DIA or SWATH-MS) [4], and (3) goals of protein quantification, which determines relative quantification of changes in protein abundance between groups, and absolute quantification, which determines the concentration of proteins on an absolute scale that is important for a wide range of questions in systems biology and clinical research. In the last few years, technical measurements from MS technology have become more accurate, and better spectral processing tools have been developed. As a result, the nature of the data has changed dramatically. The experiments have larger number of biological replicates, more spectral features, and higher signal to noise ratios in the quantified features. Many of the existing statistical tools have not kept up with these recent developments. Therefore, the MS proteomics community needs a flexible and versatile statistical analysis tool to reflect the diverse types of mass spectrometry-based experiments, which has a good performance.

1.1.2 Statement of the statistical and computational problem

The different types of MS experiments now have different characteristics. As the proteomics field grows, the experiments become larger in scale, quantifying more biological replicates, more proteins, and more features per protein. Different experimental workflows have different patterns of biological and technical variation, and different patterns of missing values. There is currently no integrated statistical approach that is applicable across these diverse types of MS experiments, and for which the computation is scalable to large datasets. Even when new statistical methods are developed, the experimentalists with little statistical background often struggle to use these methods in practice.

Therefore, my goal is to develop a statistical tool that can (1) perform an appropriate analysis based on experimental design for each protein, including borderline cases where the data structure for individual proteins deviates from the overall structure of the experiment, (2) perform model-based inference efficiently and on a large scale, (3) enable the inter-operability of the method implementation with other popular signal processing and data analysis tools, (4) thoroughly evaluate the method and the implementation in a variety of the datasets, and (5) facilitate the use of the method by experimentalists who have little statistical knowledge.

1.2 Statement of contributions

1.2.1 Statistical methods

The main insight of the proposed method is the representation of quantitative mass spectrometry-based proteomic experiments as an instance of split plot design. The whole plot part of the design reflects the structure of the biological variation of the experiment, such as case-control design, paired design, or time-course design. The subplot part of the design reflects the structure of the technological variation,

such as fragmentation patterns, labeling strategy, and presence of multiple peptides per protein. I proposed to analyze the whole plot part and subplot part separately, as it maximizes the flexibility of the model, increases the speed of the analysis, and facilitates the interpretation.

For the subplot part of the design:

- To impute censored missing values with likelihood estimation by accelerated failure time model.
- To use robust estimation with Tukey's median polish to extricate existence of outliers and missing measurements.
- For experiments with a labeling strategy, I extended linear models to separately take into account (1) the deviation of the reference intensity in the run and (2) the deviation of the endogenous intensity in a run from the reference.

The summary from subplot analysis is used as the input for the whole plot part of the model. For the whole plot part of the design:

- I proposed a modeling framework that is applicable to experiments with arbitrary complex designs, such as group comparisons of groups, time course or paired-design.
- Extend the models to borderline cases, such as proteins with a single replicate in a condition with or without technical replicates

The modeling framework is now applicable to targeted SRM, DDA or shotgun, and DIA or SWATH-MS.

1.2.2 Open-source software and implementation

I integrated and extended the software packages, MSstats [5], implemented by [6, 7], and SRMstats, implemented by [8]. My specific contributions are :

- I expressed arbitrary experimental designs in an internal representation, which allows us to carry out model-based inference. The experimental design is recognized automatically based on the structure of the input data. The appropriate model for the experimental design is then used for inference.

- I developed efficient data structures to ensure scalability.

- I implemented tools for visualizing the data and the model-based conclusions. Data visualization for checking quality of data, diagnostics for evaluating the quality of model fit, and visualization of test results are more flexible, in particular, the display of pre-specified proteins, and customization of components of plots using ggplot2 functionalities.

- Infrastructure for checking the correctness of input. MSstats takes as input data in a tabular format, produced by any spectral processing tool across different types of MS experiments such as SuperHirn, MaxQuant, Progenesis, MultiQuant, OpenMS, OpenSWATH, Spectronaut. Before starting statistical analysis, MSstats can recognize faulty annotation in the required input format, incomplete data structure, or duplicated information.

- Progress reports generated to help troubleshoot potential problems with functionalities of MSstats and also to keep records about statistical analysis. It includes information on the R session, options selected by the user, checks of successful completion of intermediate analysis steps, and possible reasons for the errors if the analysis produces an error.

- Interoperability with other existing computational tools
  - MSstats is integrated as an external tool in Skyline, a widely used computational framework for processing signals in mass spectra (more than 5000 registered users). MSstats is now available as an external tool in Skylime. It supports automatic installation and point-and click execution with a graphical user interface (GUI) for easy use. [9]

- MSstats provides the functionality to convert the output from MaxQuant, another popular spectra processing tool for MS proteomics, into the required input format for MSstats.
- MSstats satisfies the interoperability requirements from Bioconductor and takes as input data in the MSnSet format, which is the general format for proteomics on Bioconductor. MSstats has been available on Bioconductor since October 2013.

MSstats has 8962 lines of R code. I have tested it with datasets with up to 3,097 proteins, 162,492 features, 24 MS runs, which is average size of DIA data. There are more than 3,000 downloads for MSstats R package from Bioconductor, each with unique IPs since October 2013, placing the package in the top 20% most downloaded Bioconductor packages. MSstats is the most downloaded external tool in the mass spectrometry signals processing software, Skyline, with 5,900 downloads since February 2013.

1.2.3 Evaluation and case studies

I evaluated the methods and the implementation using ten controlled experimental datasets as listed in Section 4.1.1. These published or unpublished datasets were tested across maximum four spectral processing tools. I also evaluated the methods and the implementation using a variety of biological or clinical case studies as listed in Section 4.1.2.

1.2.4 Interdisciplinary education

I developed educational materials that introduce the basic statistical methodology used in MSstats, as well as the use of package itself. The target audience is experi-

mentalists who have little familiarity with statistical methods. I have participated in 12 short courses in USA, Europe and Asia.

1.3 Related manuscripts

This dissertation proposes the statistical framework for relative protein quantification in mass spectrometry-based proteomics and implementation in the following chapters 3, 4, 5, 6.

Chapter 3 proposes a statistical framework with split-plot design approach for protein significance analysis in mass spectrometry based proteomics experiments and chapter 4 shows the evaluation of the proposed method using a variety of datasets, based on a manuscript in preparation.

Chapter 5 proposes the implementation by `R` package for statistical analysis of quantitative mass spectrometry-based, based on [5] and [9].

Chapter 6 shows the application of the proposed statistical framework and downstream analysis for biomarker study, based on [10], [11], [12], [13].

2. INTRODUCTION

2.1 Mass spectrometry-based proteomics

2.1.1 Quantitative proteomics

Proteomics based on high-resolution mass spectrometry(MS) has matured as a powerful tool for the analysis of protein abundance, modifications and interaction from the quantification of thousands of proteins in biological and clinical investigation [14]. As the MS-based proteomics field grows, diverse quantification strategies for large scale have been extensively developed during the present decade [15]. For relative quantification of proteins, generally there are two categories of measurements; spectral counting vs measurements of peak intensities [16]. Spectral counting approach relatively quantifies protein abundance by counting the number of peptide-spectrum matches for each protein from MS/MS search result. It can be robust measurement of protein abundance because an increase in protein abundance results in a increase in the number of its digested peptides [17] and have been used for discovery studies in label-free experiments. Quantification by peak intensity measures the peak area or peak heights in chromatography as intensity of detected peptide/fragment ions. As amount of peptides increase, the measured peak intensities also increase. It can be used for label-free, labeling and targeted approach described in next section. In addition, the absolute quantification of abundance of proteins is possible with anchor point peptides or proteins that are introduced at known concentrations [18, 19]. In this dissertation, I focus on relative quantification by chromatography-based peak intensities.

2.1.2 Acquisition of mass spectra and characteristics

For chromatography-based quantification of proteomics experiments, there are several acquisition methods of mass spectra.

Data-dependent acquisition (DDA)

In DDA, the proteins in biological samples are digested into the peptide mixture by enzyme. The peptides are separated and analyzed in lipid chromatography coupled with mass spectrometry(LC-MS) and are detected with the peak signal by elution time from a liquid chromatography column and the ratio of mass to charge(m/z). Then the detected peptide ions are fragmented, identified by sequence of peptide fragment ions and quantified by peak intensity. The peak intensity is related to the peptide abundance, and it can be used for relative quantification. [20] It is commonly called 'bottom-up' workflow and is used for identifying large number of proteins in small number of complex samples in discovery-based research. The disadvantage is lack of reproducibility for precursor selection between samples and experiments. Also, the measured peak intensities are noisy and many missing values at low abundance.

Selected reaction monitoring (SRM)

SRM is a targeted proteomics experiment. Unlike DDA, it targets the predefined peptides and focuses on detecting and quantifying these peptides. The proteins in samples are digested and loaded in LC-MS as DDA. Then the different instrument, triple quadruple mass spectrometer, is used to monitor the precursor/fragment ion signals of target peptides based on priori information such as m/z of precursor ion, retention time, optimized collision energy and unique fragment ions of the targeted peptides [3]. The peak intensities of precursor/fragment ion combinations of

a peptide, called transitions, are measured for relative quantification. This method increases sensitivity and specificity and has high reproducibility [21].

Data-independent acquisition (DIA)

DIA is the novel approach to take strength of DDA and SRM and overcome their limitations [22]. Within cycle LC time range, peptide ions are detected by searching as DDA, and they are fragmented and all of them are quantified in predefined m/z window as SRM. This procedure is repeated until covering the full m/z range [4, 23]. Therefore It can consistently and accurately quantify the large number of features per proteins and large number of proteins.

Label free or label-based workflow

There are several different types of labeling protocols by chemical tagging approach, such as iTRAQ labeling experiment [24, 25] that each condition is separately labeled with different isobaric tag, or by metabolic tagging, such as stable isotope labeling by amino acids in cell culture (SILAC) [26, 27] that cells or mice in different conditions are grown in the medium with different isotope state. We only focus on label-free experiments or label-based workflow with isotope labeled reference peptides. The label-based workflow with isotope labeled reference peptides spikes in the labeled synthetic peptide with the same sequence of the targeted peptides as reference. It allows to adjust technical MS run variation with reference peptides and improve the precision of quantification [28].

2.1.3 Signal processing tools

There are plenty of commercial or free software packages to process and analyze diverse types of MS proteomics experiments as above, in terns of peptide identification

and quantification. Each software can analyze the dataset from different acquisition methods and has different features of workflow.

Skyline [29] is one of most popular spectral processing tools, which can analyze SRM and Parallel Reaction Monitoring(PRM), DIA and targeted DDA quantitative methods. The prominent feature of this tool is to distinguish between peaks missing at random and peaks missing at low abundance, and report, respectively, 'NA' and 0.

MaxQuant [30] is another most popular spectral processing tool for proteomics experiments, which can analyze several labeling techniques as well as label-free quantification. It also supports the search engine, Andromeda [31]. It provides several intensities, such as original intensities and summarized intensities by MaxLFQ [32], iBAQ [33, 34]. It reports the intensities of any missing peaks as 'NA'. Perseus is the companion software of MaxQuant for downstream statistical analysis, especially for shotgun proteomics data analyses, including basic statistical analysis with visualization of data and statistical results.

Progenesis QI for proteomics for proteins (Nonlinear Dynamics/Waters) supports DDA and DIA for identification and relative quantification of proteins. It reports the intensity of any missing values as 0.

SuperHirn [35] is open source software tool for label-free quantification of DDA experiments.

Proteome Discoverer(Thermo scientific^{TM}) supports to identify proteins and peptides by several search algorithms (SEQUEST [36,37], Mascot [38], etc.) and quantify proteins for different isobaric mass tagging strategies.

openMS [39] offers the analyses for DDA, SILAC, iTRAQ, SRM and SWATH with their own identification and searching algorithm and provides quantified intensities.

OpenSWATH [40] is a free proteomics software to analyze DIA data including quality control by mProphet [41] and pyprophet [42]. It reports the intensity of any missing values as 0.

Spectronaut (Biognosys AG) is developed for the analysis of DIA and SWATH datasets including automatic quality control and interference correction.

ISOQuant [43] and DIA-Umpire [44] are also for analyses of DIA datasets.

Beyond the list of software packages, many other softwares have been developed for identifying and quantifying signals with different aspects. It is important that all signal processing tools across any type of acquisition methods produce a data matrix for quantified intensities , which can be used for downstream statistical analysis like protein significance analysis or classification.

2.2 Related work : MSstats

Our group has previously proposed linear mixed models for protein-level inference for each type of acquisition strategy [6, 8], and I implemented them and expanded statistical model as a flexible family of linear mixed models for protein-level inference (Figure 2.1, Figure 2.2) in open-source R-based software MSstats [5]. The family has the advantage of being applicable to a broad variety of experimental situations [7] such as different types of experimental design, with or without technical replicates, single feature or single subject in experiment.

The basic linear mixed model in MSstats was developed in 2009 [6]. At the time these were mostly label-free shotgun experiments with a small number of biological replicates, and a few noisy features per protein. Traditional methods had a hard time to find changes. Our group proposed to (1) condition the experiments on the selected biological replicates (i.e., use fixed effects), and (2) assume that the between-feature variation equals the between-subject variation. This allowed us to increase the statistical power, and detect subtle but consistent changes. The price was in

increasing false positives, but this was acceptable in screening experiments with a small sample size and a large noise.

Since then the nature of the data changed dramatically. The experiments now have larger sample size, more features, and much less noise. The assumptions (1) and (2) are now not appropriate. Especially, the statistical model, which can distinguish between-feature variation and between-subject variation, is needed. The assumptions (1) and (2) are eliminated by treating subjects as random (i.e., using random effects). In this way, the model separates the between-subject and the between-feature variation, and the tests are based on biological variation only. There are two problems with that. First, if the experiment has a small number of biological replicates, it lacks power and will not find any changes. Second, random effects are more difficult to fit in complex designs, and may have problems of convergence. For the first problem there is nothing to do, except repeatedly explaining to the users the issues of biological replication and power. In order to solve the second problem, the new statistical method is needed, which can reflect recent nature of proteomics data.

2.3 Related work : other statistical methods

There are many approaches for protein-level inference from intensity-based spectral data. We classify them loosely in two groups below.

Two-step approaches Some researchers advocate a simple two-step approach, where the intensities (or log-intensities) of all the features are first summarized in each run, and then subjected to statistical modeling. For example, Skyline takes the sum of all the intensities in a run on the original scale, while filtering out any run that have any missing values. Although this strategy avoids biases due to missing intensities, it also leads to loss of some potentially valuable information. Perseus uses summed feature intensities on the original scale after a complex feature-level normalization from MaxQuant [32]. Progenesis also outputs summarized features as

part of the analysis. The downstream statistical inference proceeds on the summarized (and possibly log-transformed) data using simple methods such as t-test, ANOVA [45], or permutation tests [32], or methods originally developed for gene expression microarrays, e.g. Linear Models for Microarray Data (LIMMA) [46, 47], Significance Analysis of Microarrays (SAM) [48,49], and Rank Product [50,51]. Other alternatives are *ad hoc* methods for summarized data, such as ROTS [52]. Although common, these two-step approaches are not motivated by methodological considerations. As the result, there is a great diversity of detailed decisions made by various methods and tools, all of which affect the final conclusions, but for which the methodological properties are unknown.

Linear models Other researchers advocate linear mixed effects models, which perform statistical inference directly from the quantified features, e.g. [45, 53]. The frequentist version of the linear model for DDA workflow is proposed in [54, 55]. The linear models have been extended to express the limited ability of mass spectrometers to detect low-abundant analytes, by explicitly modeling the underlying censoring mechanisms [56,57], or by combining linear models with the presence/absence analysis of the analytes [58,59]. Linear models can also account for outliers by downweighting poor quality peaks [60]. Moreover, Bayesian specifications of linear models enable a probabilistic treatment of missing values [61, 62], and a filtering strategy for outlier detection [63]. Although these approaches are extremely valuable and innovative, their complexity limits their practical adoption in general circumstances. They may not be easily extended to arbitrary complex experimental designs, or may not scale to an arbitrary large number of features, proteins and samples.

Even though empirical comparisons of performance of various methods increasingly appear, e.g. comparisons of peptide-level versus summarization-based methods for DDA analysis [64], comparison of peptide-level quantification by peptide selection [65], or comparison of testing methods [66] or of imputation methods [67], there is currently little conceptual understanding of differences and similarities of the properties shared by various approaches.

My research contributes a framework, which builds upon our previous linear mixed effects modeling, and unifies many of the approaches above. I show that this unified modeling allows us to take advantage of the best aspects of the previously proposed methods, and maximize the accuracy of detecting differentially abundant proteins and of estimation of fold changes across all signal processing tools and many published datasets. The implementation of this framework in MSstats enables the analysis of experiments with arbitrary complex designs, and shortens the analysis time.

	log(peak intensity)	=	Expected reference abundance	+	Deviation from the reference due to: feature	+	condition or time	+	between-condition interference	+	biol. replicate	+	between-subject interference	+	Random meas. error
General case	**Group comparison:** y_{ijkl}	=	μ_{1111}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+	$S(C)_k$	+			ε_{ijkl}
	Time course: y_{ijkl}	=	μ_{1111}	+	F_i	+	T_j	+	$(F \times T)_{ij}$	+	S_k	+	$(T \times S)_{jk}$	+	ε_{ijkl}
	Paired design: y_{ijkl}	=	μ_{1111}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+	S_k	+	$(C \times S)_{jk}$	+	ε_{ijkl}
Single feature with technical replicates	**Group comparison:** y_{1jkl}	=	μ_{1111}	+			C_j	+			$S(C)_k$	+			ε_{1jkl}
	Time course: y_{1jkl}	=	μ_{1111}	+			T_j	+			S_k	+	$(T \times S)_{jk}$	+	ε_{1jkl}
	Paired design: y_{1jkl}	=	μ_{1111}	+			C_j	+			S_k	+	$(C \times S)_{jk}$	+	ε_{1jkl}
Single feature, no technical replicates	**Group comparison:** y_{1jkl}	=	μ_{1111}	+			C_j	+							ε_{1jkl}
	Time course: y_{1jkl}	=	μ_{1111}	+			T_j	+			S_k	+			ε_{1jkl}
	Paired design: y_{1jkl}	=	μ_{1111}	+			C_j	+			S_k	+			ε_{1jkl}
Single subject with technical replicates	**Group comparison:** y_{ij1l}	=	μ_{1111}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+					ε_{ij1l}
	Time course: y_{ij1l}	=	μ_{1111}	+	F_i	+	T_j	+	$(F \times T)_{ij}$	+					ε_{ij1l}
	Paired design: y_{ij1l}	=	μ_{1111}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+					ε_{ij1l}
Single subject, no technical replicates	**Group comparison:** y_{ij11}	=	μ_{1111}	+	F_i	+	C_j	+							ε_{ij11}
	Time course: y_{ij11}	=	μ_{1111}	+	F_i	+	T_j	+							ε_{ij11}
	Paired design: y_{ij11}	=	μ_{1111}	+	F_i	+	C_j	+							ε_{ij11}

Figure 2.1.: Previously proposed linear mixed effect model for label-free experiments [5,6].

	log(peak intensity)	=	Expected reference abundance	+	Deviation from the reference due to: feature	+	condition or time	+	between-condition interference	+	biol. replicate	+	between-subject interference	+	run	+	between-run interference	+	Random meas. error
General case	**Group comparison:** y_{ijkl}	=	μ_{1001}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+	$S(C)_k$	+			R_l	+	$(F \times R)_{il}$	+	ε_{ijkl}
	Time course: y_{ijkl}	=	μ_{1001}	+	F_i	+	T_j	+	$(F \times T)_{ij}$	+	S_k	+	$(T \times S)_{jk}$	+	R_l	+	$(F \times R)_{il}$	+	ε_{ijkl}
	Paired design: y_{ijkl}	=	μ_{1001}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+	S_k	+	$(C \times S)_{jk}$	+	R_l	+	$(F \times R)_{il}$	+	ε_{ijkl}
Single feature with technical replicates	**Group comparison:** y_{1jkl}	=	μ_{1001}	+			C_j	+			$S(C)_k$	+			R_l	+			ε_{1jkl}
	Time course: y_{1jkl}	=	μ_{1001}	+			T_j	+			S_k	+	$(T \times S)_{jk}$	+	R_l	+			ε_{1jkl}
	Paired design: y_{1jkl}	=	μ_{1001}	+			C_j	+			S_k	+	$(C \times S)_{jk}$	+	R_l	+			ε_{1jkl}
Single feature, no technical replicates	**Group comparison:** y_{1jkl}	=	μ_{1001}	+			C_j	+							R_l	+			ε_{1jkl}
	Time course: y_{1jkl}	=	μ_{1001}	+			T_j	+			S_k	+			R_l	+			ε_{1jkl}
	Paired design: y_{1jkl}	=	μ_{1001}	+			C_j	+			S_k	+			R_l	+			ε_{1jkl}
Single subject with technical replicates	**Group comparison:** y_{ij1l}	=	μ_{1001}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+					R_l	+	$(F \times R)_{il}$	+	ε_{ij1l}
	Time course: y_{ij1l}	=	μ_{1001}	+	F_i	+	T_j	+	$(F \times T)_{ij}$	+					R_l	+	$(F \times R)_{il}$	+	ε_{ij1l}
	Paired design: y_{ij1l}	=	μ_{1001}	+	F_i	+	C_j	+	$(F \times C)_{ij}$	+					R_l	+	$(F \times R)_{il}$	+	ε_{ij1l}
Single subject, no technical replicates	**Group comparison:** y_{ij1l}	=	μ_{1001}	+	F_i	+	C_j	+							R_l	+			ε_{ij1l}
	Time course: y_{ij1l}	=	μ_{1001}	+	F_i	+	T_j	+							R_l	+			ε_{ij1l}
	Paired design: y_{ij1l}	=	μ_{1001}	+	F_i	+	C_j	+							R_l	+			ε_{ij1l}

Figure 2.2.: Previously proposed linear mixed effect model for SRM experiments with stable isotope labeled reference peptides [5, 8].

3. METHOD

3.1 Notation and goals of statistical analysis

Whole plot

Subplot	Condition$_1$						⋯				⋯	Condition$_I$						⋯			
	Subject$_1$			Subject$_2$			⋯	Subject$_J$			⋯	Subject$_{(I-1)J+1}$			Subject$_{(I-1)J+2}$			⋯	Subject$_{IJ}$		
	Run$_1$	Run$_2$	Run$_3$	Run$_4$	Run$_5$	Run$_6$	⋯	Run$_{JK-2}$	Run$_{JK-1}$	Run$_{JK}$	⋯	Run$_{(I-1)JK+1}$	Run$_{(I-1)JK+2}$	Run$_{(I-1)JK+3}$	Run$_{(I-1)JK+4}$	Run$_{(I-1)JK+5}$	Run$_{(I-1)JK+6}$	⋯	Run$_{IJK-2}$	Run$_{IJK-1}$	Run$_{IJK}$
Feature$_1$	y	y	y	y	y	y	⋯	y	y	y	⋯	y	y	y	y	y	y	⋯	y	y	y
Feature$_2$	y	y	y	y	y		⋯	y	y	y	⋯	y	y	y	y	y	y	⋯	y	y	y
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
Feature$_L$	y		y			y	⋯		y	y	⋯		y	y	y	y	y	⋯	y		y

Figure 3.1.: The general data structure of a quantitative proteomic experiment. A label-free experiment with a group comparison design with technical replicates. The whole plot aspect of the experimental design is shown in pink. The subplot is shown in yellow. μ denotes the overall mean signal in the experiment. y denotes the log intensity of the observed feature in each cell. Empty cells indicate missing values.

Figure 3.1 outlines measurements from one protein in a typical group comparison experiment, where data are acquired with a label-free workflow such as DDA or DIA. The experiment in the figure has $i = 1, \ldots, I$ *Conditions*. The conditions can be predefined groups, e.g. healthy and disease. Alternatively, the conditions can represent complex combinations of treatments, where a condition is a unique combination of all the levels of the treatments.

Each condition consists of $j = 1, \ldots, J$ *Subjects*, i.e. distinct biological replicates (e.g., patients, mice, etc). The subjects are the main experimental units, and in group comparison designs the subjects are *nested* within the conditions. Furthermore, each subject sample is profiled in $k = 1, \ldots, K$ mass spectrometry *Runs*. When a subject is represented by multiple runs, the runs are technical replicates. The experiment in

Figure 3.1 has $I \times J \times K$ runs. In practice the number of biological and technical replicates can vary across conditions and subjects.

In each run the protein is represented by $l = 1, \ldots, L$ spectral features. The features are peptide ions in DDA experiments, combinations of peptide ions and transitions in SRM experiments, and combinations of peptide ions and fragments in DIA experiments. For the purposes of this dissertation we do not distinguish transitions or fragments generated by a same or different peptide. For example, a protein with 2 peptide ions and 3 transitions per peptide has 6 features.

Each feature in each run is quantified by its *Intensity*, defined as peak area, peak hight at apex, or any other measure of the strength of the signal reported by a signal processing tool. Throughout the dissertation we denote that y_{ijkl} the peak intensities that are $\log_2$ transformed (and use 0 after the $\log_2$ transformation if an intensity is quantified as 0 on the original scale) and normalized, to account for the technological artifacts between the runs. In practice some peak intensities can be misidentified or mis-quantified. Some peaks can be missing at random. Some peaks can be missing according to a censoring mechanism, due to the inability of mass spectrometers to detect signals from low-abundant analytes. The quantification of missing peaks can differ between signal processing tools (e.g., as 0, 'NA', or the noise in the area where the peak is expected). We show below that the proposed approach can work with all these choices.

Although Figure 3.1 outlines an experiment with a group comparison design and a label-free workflow, other experiments (e.g., experiments with time course or paired designs, experiments with stable isotope labeled synthetic reference peptides, experiments with reference peptides from a metabolically labeled reference sample, or experiments with unbalanced numbers of replicates or missing values) can all be represented with a similar layout. For example, Figure 3.5 shows the layout of an experiment with heavy-labeled reference peptides. Although our discussion focuses on group comparisons, it is easily extended to other experiments with complex designs.

The quantitative proteomic experiments typically have three goals. The first goal is to compare pairs of conditions (or combinations of conditions), to detect proteins that change in abundance more systematically than as expected by random chance. The second goal is to estimate the magnitude of fold changes associated with these comparisons. The third goal is to summarize all the measurements that pertain to a protein, and obtain a single value of relative protein quantification per run, per subject, or per condition. The latter can be used as input to machine learning methods, e.g. unsupervised clustering or supervised classification. *The goal of statistical modeling and inference is to accurately represent all the systematic and random variation in the data, and provide the best model-based quantities that address these goals.*

3.2 Overview of split-plot design

The main contribution is to recognize that quantitative mass spectrometry experiments incorporate a restriction on randomization. In other words, the order of the quantified protein features is not randomized, and all the features of a protein are acquired simultaneously within a run. In statistical language, the layout in Figure 3.1 is an instance of a split-plot design [68], and the features are *sub-samples* of the run. This aspect of the design is highlighted in Figure 3.1 in yellow, and is called *subplot*. The order of conditions, subjects and runs is randomized, and their variation reflect the underlying biological and technical variation. This aspect of the design is highlighted as in Figure 3.1 in pink, and is called *whole plot*. In early days of quantitative proteomics the subplot variation was comparable to the whole plot variation, however in modern experiments the subplot variation is comparatively small.

The model for the experiment in Figure 3.1 is shown in Figure 3.2, and its extension to experiments with labeled reference peptides is in Figure 3.6. Although the full model is similar to the models in our previous work, the split-plot view of the experimental design changes the parameter estimation procedure. It is easy to show ([68] and Figure 3.3) that in the special case of a balanced split-plot experiment

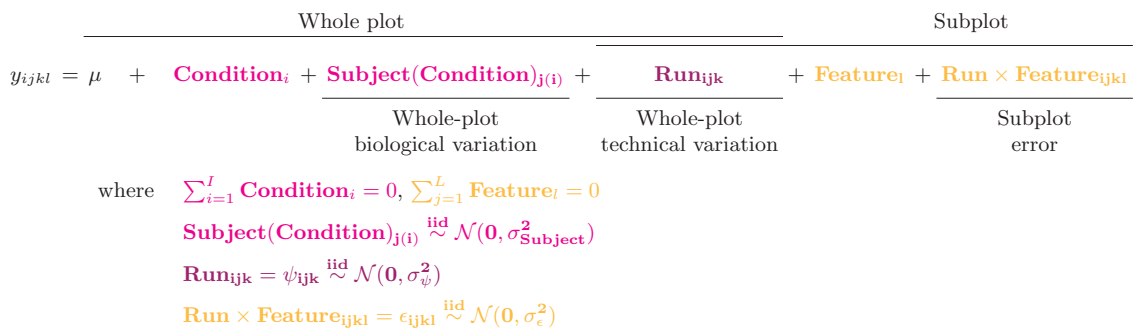


Figure 3.2.: Linear mixed effects model and assumptions for the experiments in Figure 3.1. The model has three variance components, that reflect the biological and the technological variation.

with no missing values and equal technical variance of all features, the goals of group comparison, fold change estimation, and summarization of protein abundance, the optimal procedure of statistical inference only relies on average log-feature intensities in each run. In other words, the goals can be achieved using a modified two-step statistical inference procedure, where summarization of log-intensities in the run is an intermediate step.

Model term	Sum of Squares	Degrees of freedom	E(MS): *Condition*, *Feature* : F, *Subject* : R
Condition	$JKL\sum(\bar{y}_{i\ldots} - \bar{y}_{\ldots})^2$	$I-1$	$\sigma^2_\epsilon + \mathrm{L}\sigma^2_\psi + \mathrm{KL}\sigma^2_{Subject} + \frac{JKL\sum Condition_i^2}{I-1}$
$Subject_{(Condition)}$	$KL\sum\sum(\bar{y}_{ij\cdot} - \bar{y}_{i\ldots})^2$	$I(J-1)$	$\sigma^2_\epsilon + \mathrm{L}\sigma^2_\psi + \mathrm{KL}\sigma^2_{Subject}$
Wholeplot error (Run)	$L\sum\sum\sum(\bar{y}_{ijk\cdot} - \bar{y}_{ij\cdot})^2$	$IJ(K-1)$	$\sigma^2_\epsilon + \mathrm{L}\sigma^2_\psi$
Feature	$IJK\sum(\bar{y}_{\ldots l} - \bar{y}_{\ldots})^2$	$L-1$	$\sigma^2_\epsilon + \frac{IJK\sum Feature_l{}^2}{L-1}$
Subplot error (R×F)	$\sum\sum\sum\sum(y_{ijkl} - \bar{y}_{ijk\cdot} - \bar{y}_{\ldots l} + \bar{y}_{\ldots})^2$	$(IJ-1)(K-1)(L-1)$	σ^2_ϵ
Total	SS_{Total}	$IJKL-1$	

Figure 3.3.: Analysis of variance table for split-plot of group comparison design with label free experiment is summarized in Figure 3.2. The expected mean squares for this split-plot design are with random subjects and condition and feature fixed, are shown in. σ_ϵ is subplot error. σ_ψ is whole plot error.

We use the analogy of the balanced design to propose a two-step estimation and inference procedure applicable to general quantitative proteomic experiments (Figure 3.4). The general formulation in terms of linear mixed effects models links this approach to linear mixed effects models previously proposed by our group and by other groups. The estimation procedure links this approach to the *ad hoc* two-step approaches described in the related work above. This methodological connection allows us to combine the best aspects of these approaches.

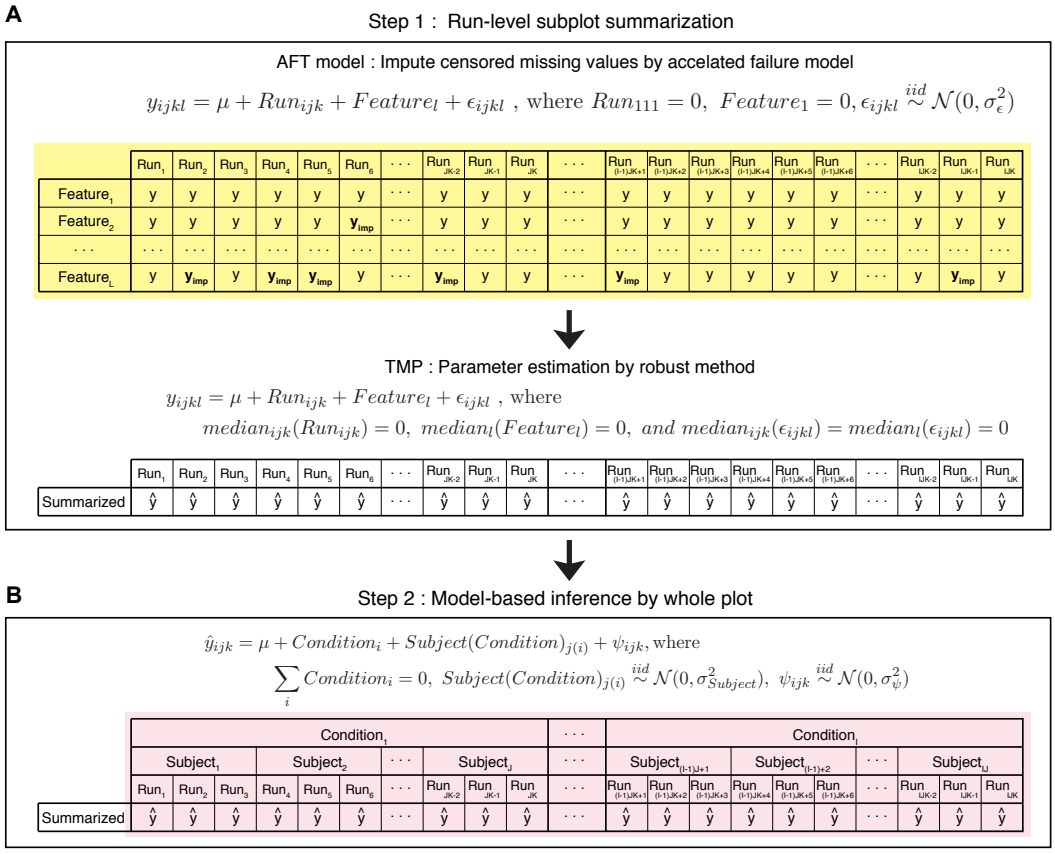


Figure 3.4.: Proposed two-step estimation and inference procedure. (A) Subplot. The step imputes missing log-feature intensities using a two-way linear model, which incorporates the accelerated failure time censoring mechanism, and summarizes the resulting data structure using Tukey's median polish. (B) Whole plot. The step models the output or run-level summarization with a linear mixed effects model that accurately represents the experimental design, and reports model-based conclusions.

3.3 Subplot summarization of feature intensities in a MS run

Figure 3.4 (A) shows that subplot is a two-way table, where rows are features, and columns are mass spectrometry runs. All the features are quantified simultaneously within a run, and therefore parameter estimation at the subplot level performs *run-level summarization*. A desirable side-effect of the subplot is that it is agnostic of the biological origin of the samples, e.g., of whether these are biological or technical replicates, or samples from a same group or from different groups. Since the structure of subplot is very simple, many approaches to run-level summarization (including those that can address complications such as outlying measurements and missing data) can be successfully applied regardless of the complexity of the experimental design. Below we describe three such approaches, and then discuss how these approaches are integrated in the proposed general and versatile workflow.

3.3.1 Approach 1: Two-way fixed effects Analysis of Variance (ANOVA)

It is described in Figure 3.4 (A), and is the simplest model for subplot. The parameters are estimated by minimizing the residual sum of squares, and run-level summaries are average log-intensities within each run. In experiments with no missing values and equal technical variance of all features the estimation is unbiased and has minimum variance. However, the approach is undermined by outliers, missing values, or unequal variance.

3.3.2 Approach 2: Accelerated failure time (AFT) model

It was brought to proteomics from statistical methods for survival analysis. It accounts for the censoring mechanism, i.e. for missing peak intensities that arise due to the limited ability of mass spectrometers to detect low abundant analytes. The AFT model in proteomics was originally proposed for group comparison designs,

after averaging peak intensities from all peptides into a single quantity per protein per run [57]. In contrast, here we propose to use the AFT model within subplot, before the run-level summarization.

In presence of censored observations, the observed log-intensities of peaks y_{ijkl} can be thought of as $y_{ijkl} = max(y_{ijkl}, c_{ijkl})$, were c_{ijkl} is the censoring threshold, i.e. the lowest quantifiable signal on the $\log_2$ scale. We define an indicator of whether the peak was detected and quantified as

$$\delta_{ijkl} = \begin{cases} 1 & \text{if } c_{ijkl} \leq y_{ijkl}, \text{ i.e., observed} \\ 0 & \text{if } c_{ijkl} > y_{ijkl}, \text{ i.e., censored} \end{cases} \tag{3.1}$$

The likelihood for the model in Figure 3.4 (A) is then the product of the likelihoods of the observed peak intensities (top line of Eq. (3.2)) and censored peak intensities (bottom line of Eq. (3.2))

$$\begin{aligned} L(Run_{ijkl}, Feature_{ijkl}|y_{ijkl}) &= \prod_{i,j,k,l} f\left(y_{ijkl} \mid \mu + Run_{ijk} + Feature_l,\ \sigma_\epsilon^2\right)^{\delta_{ijkl}} \\ &\times \prod_{i,j,k,l} F\left(y_{ijkl} \mid \mu + Run_{ijk} + Feature_l,\ \sigma_\epsilon^2\right)^{1-\delta_{ijkl}} \end{aligned} \tag{3.2}$$

Here f is the probability density function and F is the cumulative probability density function of the Normal distribution with the expected value $\mu + Run_{ijk} + Feature_l$ and variance σ_ϵ^2. The latter expresses the fact that the intensities of the censored peaks are unknown, but are below the threshold. The parameters Run_{ijk}, $Feature_l$ and σ_ϵ^2 are estimated by maximizing the likelihood.

The AFT model has the same mechanism of accounting for missing values with a censoring threshold c_{ijkl} as in Eq. (3.1), and the same maximum likelihood-based estimation Eq. (3.2). The AFT model for subplot describes the contributions of the systematic sources variation (in our case, features and runs) to the untransformed intensity of the peaks, x_{ijkl}, in a way that works directly with cumulative probability density functions. It assumes that

$$F(x_{ijkl}) = \Phi\left(2^{\mu+\sigma z_{ijkl}} \cdot 2^{Run_{ijk}+Feature_l}\right), \text{ where } Run_{111} = 0,\ Feature_1 = 0 \quad (3.3)$$

Φ is the cumulative density function of the Standard Normal distribution. If x_{ijkl} were to represent the time to a failure, then the parameters of the model would have had the effect of 'accelerating or contracting' the time with respect to the reference, hence the name of the model.

Denoting as $z_{ijkl} \sim \mathcal{N}(0,1)$ we can re-write the model as

$$\begin{aligned} x_{ijkl} &= 2^{\mu+\sigma z_{ijkl}} \cdot 2^{Run_{ijk}+Feature_l} \\ log_2\ x_{ijkl} &= \mu + \sigma z_{ijkl} + Run_{ijk} + Feature_l \\ y_{ijkl} &= \mu + Run_{ijk} + Feature_l + \sigma z_{ijkl} \end{aligned} \quad (3.4)$$

where y_{ijkl} is $\log_2$ transformation of x_{ijkl}.The similarity of Eq. (3.4) to linear regression is the source of the flexibility, and also of the popularity of this model in applications. The parameters μ, Run_{ijk}, $Feature_l$ and σ are estimated by maximum likelihood with censoring mechanism, as described in the main manuscript. In absence of missing values, the parameter estimates Run_{ijk} and $Feature_l$ are the same with those of two-way ANOVA. The AFT model can be easily extended to other distributions of the peak intensities in addition to the Normal.

Several aspects of parameter estimation in presence of censoring require attention. First, the censoring threshold is effectively one more parameter in the model. Here we assume that the threshold is feature-specific but constant across the runs, i.e. $c_{ijkl} = c_l$, and estimate it by the smallest observed log-intensity of each feature. Second, in mass spectrometric experiments missing values may also arise by random chance. Some signal processing tools such as Skyline distinguish between these two types of missing values, and in this case randomly missing peak intensities can be left missing. However, as the technology improves, the proportion of randomly missing values becomes small, and it is reasonable to assume that all the missing values are in fact censored. Third, although the proposed approach takes care of the censored

missing values, it can be negatively affected by misidentified or misquantified peaks, which manifest themselves through outlying intensity values. Finally, the AFT model can be used to predict the log-intensities of the missing peaks. This prediction is more accurate than the imputation of missing values with, say, minimal value in the run, because it takes into account the intensities of the feature and the run, and also the censoring threshold. We will take advantage of this aspect of the model in the proposed approach below.

3.3.3 Approach 3: Tukey's Median Polish (TMP)

TMP [69] is a robust parameter estimation method for a two-way fixed effects ANOVA. It has a long and successful history in bioinformatics, used, e.g., in affymetrix microarrays to summarize multiple probes in a probe set [70]. TMP iteratively subtracts medians of rows and columns from the observed log-intensities in Figure 3.4 (A) until there is no change. The values that remain in the table after these operations are the residuals of the model fit. The run-level summarization is obtained by summing the fitted overall and run effects in the run.

This median-based summarization down-weighs the outliers and the highly variable features, and the simple structure of subplot allows us to implement this approach in an arbitrary complex experimental design. However, the approach is negatively affected by the missing values, especially by informative missing values that are due to censoring. The effect is particularly strong when more than 50% of the intensities in a run are sensored.

3.3.4 Proposed workflow: integration of approaches 1, 2 and 3

We propose a subplot summarization that combines the advantages of the three approaches above. First, we fit the AFT model to the subplot, and predict the

censored log-intensities of peaks. Second, we summarize the resulting subplot data structure, which now includes both the observed and the predicted log-intensities, using Tukey Median Polish. The summaries are then analyzed in the whole plot using a linear model that accurately reflects the experimental design (see Section 3.4).

The proposed framework is easily extended to experiments with labeled reference peptides. Example data structure and statistical model for these experiments are described in Section 3.3.5.

3.3.5 Extension : experiment with labeled reference peptides

The proposed summarization approach for subplot can be extended for the experiments with labeling strategy. SRM experiments with isotope labeled reference peptides is the case. Figure 3.5 demonstrates example data structure for simplified group comparison experiment with triplicate. In the notation of the figure, $i = 1, ..., I$ is the index of a *Condition*, $j = 1, ..., J$ is the index of a biological replicate, called *Subject* in each group, $k = 1, ..., K$ is the index of a mass spectrometry run. When a subject for certain condition is represented by multiple runs, the runs are technical replicates. The experiment has $I \times J \times K$ mass spectrometry *Runs*. $l = 1, ..., L$ is the index of a *Feature*, $m = 0$ or 1 is the index of a *Label*, 0 denotes labeled reference features and 1 denotes endogenous features. To compare with data structure of label-free experiment, there are additional L rows for labeled reference features.

The linear mixed effects model in Figure 3.2 is easily extended to reflect for the experiments with labeled reference peptides as Figure 3.6. Whole plot is the same as in label-free experiment. But, subplot model is different including *Label* and $Run \times Label$.

We specify the model in Figure 3.6, and use the approach in Figure 3.4 to estimate model parameters. The subplot summarization step should have two additional variables, *Label* and $Run \times Label$, in the model in order to consider labeled reference

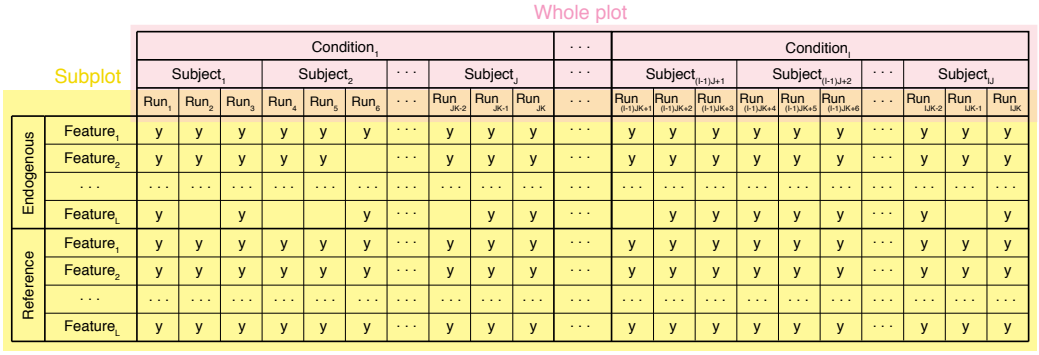


Figure 3.5.: The general data structure of the experiment with labeled reference peptides. A group comparison design with technical replicates. The whole plot aspect of the experimental design is shown in pink. The subplot is shown in yellow. y denotes the log intensity of the observed feature in each cell. Empty cells indicate missing values.

$$y_{ijklm} = \mu + \underbrace{\mathbf{Condition}_i + \mathbf{Subject(Condition)}_{\mathbf{j(i)}}}_{\text{Whole-plot biological variation}} + \underbrace{\mathbf{Run}_{\mathbf{ijk}}}_{\text{Whole-plot technical variation}} + \underbrace{\mathbf{Label}_{\mathbf{m}} + \mathbf{Run} \times \mathbf{Label}_{ijkm} + \mathbf{Feature}_{\mathbf{l}} + \epsilon_{ijklm}}_{\text{Subplot error}}$$

Whole plot | Subplot

where $\sum_{i=1}^{I} \mathbf{Condition}_i = 0$, $\sum_{l=1}^{L} \mathbf{Feature}_l = 0$, $\sum_{m=0}^{1} \mathbf{Label}_m = 0$

$\mathbf{Subject(Condition)}_{\mathbf{j(i)}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}_{\mathbf{Subject}})$

$\mathbf{Run}_{\mathbf{ijk}} = \psi_{\mathbf{ijk}} \overset{\mathbf{iid}}{\sim} \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}_{\psi})$

$\epsilon_{ijklm} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$

Figure 3.6.: Linear mixed effects model for the split plot design in a label-based experiment with group comparison design. The model has three variance components, that reflect the biological and the technological variation.

feature intensities.

Imputation step for missing values should be considered, even though SRM experiments with isotope labeled reference have very few missing values. For label-based

experiment, we consider only endogenous intensities for imputation because we can assume that missing values for reference intensities are random. Because reference intensities are expected to be consistent across runs, random missing assumption for missing reference intensities is reasonable. Imputation for missing values in endogenous intensities will be performed by the same model for label-free experiments in Section 3.3.2.

For run-level summarization step, Tukey's median polish is used for estimating parameter in Eq. (3.5). First, $Run_{ijk}+Label_m+Run\times Label_{ijkm}$ substitutes for Run_{ink} in Eq.(4) as in Eq. (3.5), which considers the data matrix including $(i \times j \times k) * 2$ columns and l rows.

$$
\begin{aligned}
y_{ijklm} &= \mu + Run_{ijk} + Label_m + Run \times Label_{ijkm} + Feature_l + \epsilon_{ijklm}, \text{ where} \\
& median_{ijkm}(Run_{ink} + Label_m + Run \times Label_{ijkm}) = 0, \\
& median_l(Feature_l) = 0, \\
& median_{ijkm}(\epsilon_{ijklm}) = median_l(\epsilon_{ijklm}) = 0
\end{aligned} \tag{3.5}
$$

Then we can get summarized intensities per run and label from estimation of parameters as in Eq. (3.6).

$$
\begin{aligned}
\widehat{y_{ijk.1}} &= \widehat{Label}_1 + \widehat{Run_{ijk}} + \widehat{Run \times Label_{ijk1}} \text{ for endogenous intensities} \\
\widehat{y_{ijk.0}} &= \widehat{Label}_0 + \widehat{Run_{ijk}} + \widehat{Run \times Label_{ijk0}} \text{ for reference intensities}
\end{aligned} \tag{3.6}
$$

Next, summarized reference intensity is used to adjust summarized endogenous intensity for MS run variations, because we can assume the equal summarized reference intensity across MS runs. Final summarized endogenous intensity is adjusted as in Eq. (3.7), which shifts the summarized intensities in a run by a constant to equalize the median of summarized reference intensities across runs.

$$\widehat{y_{ijk.1,adjusted}} = \widehat{y_{ijk.1}} - (\widehat{y_{ijk.0}} - median_{ijk.0}(Label_0 + Run_{ink} + Run \times Label_{ijk.0})) \tag{3.7}$$

These run-level summarized endogenous intensities are used for the same whole-plot inference with label-free experiments for finding differential abundance proteins.

3.4 Whole plot: representation of experimental design and model-based inference.

After the log-intensities of the peaks are summarized in each run, we model the whole plot using a linear mixed effects model that appropriately reflects the experimental design. Figure 3.4 (B) illustrates one such model for a group comparison design. This model is identical to a particular model in the previous version (v2) of MSstats, which specifies expanded scope of biological replication, and which only has one input feature. The positive side-effect of this framework is that the between-feature variance does not need to be estimated, and therefore the parameter estimation is simpler and faster. Figure 3.7 provides another example of such model for the time course design.

	Condition$_1$ or Time$_1$						...				...	Condition$_I$ or Time$_I$						...			
	Subject$_1$			Subject$_2$			...	Subject$_J$			...	Subject$_1$			Subject$_2$			...	Subject$_J$		
	Run$_1$	Run$_2$	Run$_3$	Run$_4$	Run$_5$	Run$_6$	...	Run$_{JK-2}$	Run$_{JK-1}$	Run$_{JK}$	...	Run$_{(I-1)JK+1}$	Run$_{(I-1)JK+2}$	Run$_{(I-1)JK+3}$	Run$_{(I-1)JK+4}$	Run$_{(I-1)JK+5}$	Run$_{(I-1)JK+6}$	...	Run$_{IJK-2}$	Run$_{IJK-1}$	Run$_{IJK}$
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$

Figure 3.7.: The general data structure for time-course or paired design with technical replicates.

The difference between group comparison design and time course design is subject notation. The subjects for group comparison are nested in condition. However, the subjects for time course design are crossed in different time points. For example, if there are I Conditions and J subjects for each condition for group comparison design, there are $I \times J$ unique subjects are in experiment. If there are I time points

and J subjects for each condition for time course design, there are total J subjects in experiment and these subjects are measured multiple time points. Therefore, the notation for subject in Figure 3.7 is different from Figure 3.1 and Figure 3.4(B). Then, variables for linear mixed model of whole plot are also different as Eq. (3.8).

$$\begin{aligned} \hat{y}_{ijk} &= \mu + Time_i + Subject_j + Time_i \times Suject_j + \epsilon_{ijk}, \text{ where} \\ &\sum_i Time_i = 0,\ Subject_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{Subject}),\ \epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\epsilon}) \end{aligned} \tag{3.8}$$

If there is no technical replicate, there is no interaction term in the model as below. Paired design has the same data structure with time course design, shared subjects across different conditions. The same model as Eq. (3.8) with replacing $Time$ with $Condition$ can be used.

4. EXPERIMENTAL RESULTS

4.1 Experimental data

We demonstrate the performance of the method using ten controlled mixtures of known composition and ten biological investigations, acquired with label-free DDA, DIA, and SRM with isotope labeled reference peptides. The controlled mixtures are spike-in mixtures or dilutions of a same parent mixture. They cover a broad range of known fold changes between condition, and are well suited for evaluating the sensitivity and specificity of detecting differentially abundant proteins, and evaluating the accuracy of estimation of fold changes between conditions. However, the spike-in datasets typically have a small number of technical replicates and no biological replicates. The processed data are not hand-curated, and have many missing values. Therefore, they may not be fully representative of the real biological investigation. In contrast, biological and clinical investigations have biological replicates (and, in our case, no technical replicates), and smaller fold changes. For some datasets the output of signal processing is manually curated. The datasets have few missing values. Each dataset was processed with up to four signal processing tools to demonstrate that the proposed statistical method performs well regardless of the signal processing. The datasets are listed below.

4.1.1 Controlled spike-in mixtures

Controlled mixture - label-free DDA and SRM with labeled reference peptides

Thirty commercial proteins were prepared at 1.5 pmolμL in three different subsets of 10 proteins each. Proteins from these subsets were spiked to a 15-μg *Escherichia coli* background in different amounts and five different mixtures were prepared in triplicates. The final amount of each protein in each mixture was either 100, 200 or 400 fmol/μg of *E. coli* background for the subsets marked as 1, 2 and 4, respectively, as Table 4.1. Samples were digested using the sequential in-solution Lys-C/trypsin digestion protocol.

The peptides mixtures were analyzed by online nanoflow liquid chromatography tandem mass spectrometry (nanoLCMS/MS) using an EASY-nLC system (Proxeon Biosystems, Odense, Denmark) connected to the LTQ Orbitrap Velos instrument (Thermo Fisher Scientific, Bremen, Germany) through a nanoelectrospray ion source. The instrument was operated in data-dependent acquisition mode, with full MS scans used over a mass range of m/z 250-2,000 with detection in the Orbitrap (1 microscan, resolution of 60,000). In each cycle of data-dependent acquisition analysis, following each survey scan, the twenty most intense ions with multiple charged ions above a threshold ion count of 5,000 were selected for fragmentation at normalized collision energy of 35%. Fragment ion spectra produced via collision-induced dissociation (CID) were acquired in the ion trap, with AGC set to 5e4, an isolation window of 2.0 m/z, an activation time of 0.1 ms, and a maximum injection time of 100 ms. All data were acquired with Xcalibur software v2.2.

MS/MS spectra were searched using Proteome Discovery software suite (v1.3.0.339) and Mascot (v2.3.01) as search engine. Acquired data were searched against an in-house generated database containing all the spiked-in proteins (Table 1), and the *E. coli* Swissprot protein database (July 2012 version) plus the most common protein

contaminants. Precursor ion mass tolerance was set to 7 ppm at the MS1 level and to 0.5 Da at the fragment ion level. Up to three missed cleavages for trypsin were allowed. Oxidation of methionine and protein N-terminal acetylation were considered as variable modifications, whereas carbamidomethylation on cysteines was set as a fixed modification. False discovery rate (FDR) in peptide identification was evaluated by using a decoy database and it was set to a maximum of 1%. Peptide areas were extracted with the Precursor Area Ion Detector module of Proteome Discoverer with a mass tolerance of 2 ppm. Only the spiked-in proteins were selected for further quantitative analysis and all calculations and comparisons were processed.

	Mixture				
Protein	1	2	3	4	5
Protein subset1 (7 proteins)	1	2	4	2	1
Protein subset2 (7 proteins)	4	1	2	1	2
Protein subset3 (7 proteins)	2	4	1	2	1

Table 4.1.: Composition of the different controlled mixtures for controlled spike-in dataset. The values are relative amounts among the different protein subsets. Subset 1 includes 10 proteins (P02701, P00711, Q29443, Q29550,P0CG53, P68082, P00432,P02754, P24627, P80025), Subset 2 includes 10 proteins (P00915, P02787, P02663, P01008, P00921, P05307, P61769, P02662, P01012, P02666), and Subset 3 includes 10 proteins (Q3SX14, P00563, P02769, Q58D62, P00698, P00004, P00442, P01133, P02753)

The 2015 study of the Proteome Informatics Research Group (iPRG) of the Association of the Biomedical Resource Facilities (ABRF) [71]

There are total six proteins spiked in each sample with yeast proteins background. Four proteins were spiked with four different concentrations by Latin Square design in four biological samples with triplicate. Another two proteins were spiked with another four different concentrations in the same four biological samples with triplicate. Detailed concentrations are in Table 4.2. It is a DDA dataset. Skyline was used

to analysis and to calculate MS1 peak intensities. The processed data from Skyline was distributed from iPRG. Peptides were searched by Comet, OMSSA and MSGF+. Reported three monoisotope peaks per peptide were summed for peptide-level in a MS run. The same raw files also were reanalyzed by MaxQuant with defaults and Progenesis with searching result by Comet. All pairwise fold changes between concentrations (65/55, 55/15, 65/15, 15/2, 55/2 65/2, total six fold changes) among four proteins by Latin Square design were used for sensitivity evaluation. Specificity was calculated using constant background yeast proteins.

	Concentnration(fmol)			
Protein	Sample1	Sample2	Sample3	Sample4
P44015	**65**	**55**	**15**	**2**
P44752	**55**	**15**	**2**	**65**
P44374	**15**	**2**	**65**	**55**
P44983	**2**	**65**	**55**	**15**
P44683	**11**	**0.6**	**10**	**500**
P55249	**10**	**500**	**11**	**0.6**
Background	+200ng yeast digest			

Table 4.2.: Protein IDs and concentrations of spike-in proteins for 2015 iPRG study

Dynamic Range Benchmark [32]

A mixture of 48 human UPS proteins were spiked in *E. coli* lysate with different ratios between two conditions. There were six different concentration ratios between conditions. Detailed proteins and ratios are in Table 4.3. Each condition was analyzed in 4 replicates. Data was acquired by DDA. Raw files were processed by three spectral processing tools. I used the original peptide-level intensities, not performed by MaxLFQ of MaxQuant output. Second dataset was processed by Progenesis QI with searching results by Comet. We will evaluate the ability of statistical methods to detect 10000, 1000, 100, 10, 0.1 true fold changes between conditions among 40

proteins, and true constant signals with 8 human UPS proteins and around 2200 *E.coli* background proteins.

Protein	Relative concentration UPS1	UPS2
P00441ups, P01375ups, P02741ups, P02788ups P05413ups, P08758ups, P10145ups, P10636-8ups	10000	1
P06396ups, O00762ups, P01112ups, P01579ups P09211ups, P51965ups, P99999ups, P02787ups	1000	1
O76070ups, P01127ups, P01344ups, P08263ups P10599ups, P55957ups, P61769ups, P01008ups	100	1
P00709ups ,P02753ups, P06732ups,P12081ups P16083ups, P61626ups, P63279ups, Q15843ups	10	1
P00167ups, P01133ups, P02144ups, P04040ups P15559ups, P62937ups, P63165ups, Q06830ups	1	1
P68871ups,P02768ups, P00915ups, P00918ups P01031ups, P41159ups, P62988ups, P69905ups	1	10
Background	*E.coli* digest	

Table 4.3.: Protein IDs and concentrations of spike-in proteins for dynamic range benchmark dataset

Label-based SRM controlled spiked-in dataset [8]

Total twelve proteins were spiked into six mixtures with the same background. Six proteins were spiked with six different concentrations, 8 to 512 folds, in each mixture according to the Latin Square design as positive control. Remaining six proteins were mixed at constant concentrations across mixtures as negative controls. Each two proteins among 6 proteins had different concentrations. Detailed concentrations per protein are in Table 4.4. Each mixture had two replicates. This is from SRM with isotope labeled reference peptides. MultiQuant was used for quantifying the peaks. We will evaluate the ability of statistical methods methods to detect true 8, 32, 128, 256, 512 fold changes, starting from maximum concentrations as baseline among six proteins and true constant fold changes with six negative control proteins.

Protein	Mixture 1	2	3	4	5	6
YBR132C	Max	Max/512	Max/256	Max/128	Max/32	Max/8
YBR144C	Max/8	Max	Max/512	Max/256	Max/128	Max/32
YBR147W	Max/32	Max/8	Max	Max/512	Max/256	Max/128
YBR184W	Max/128	Max/32	Max/8	Max	Max/512	Max/256
YBR203W	Max/256	Max/128	Max/32	Max/8	Max	Max/512
YBR184W	Max/512	Max/256	Max/128	Max/32	Max/8	Max
YBR168W	Max	Max	Max	Max	Max	Max
YBR186W	Max	Max	Max	Max	Max	Max
YBR204C	Max/32	Max/32	Max/32	Max/32	Max/32	Max/32
YBR228W	Max/32	Max/32	Max/32	Max/32	Max/32	Max/32
YBR250W	Max/256	Max/256	Max/256	Max/256	Max/256	Max/256
YBR270C	Max/256	Max/256	Max/256	Max/256	Max/256	Max/256

Table 4.4.: Experimental design of label-based SRM controlled spike-in experiment. 'Max' denotes the maximal concentration in each mixture and was 50 fmol.

CPTAC study III from site 52 [72]

CPTAC dataset for SRM with labeled reference peptides has seven proteins that were spiked in human plasma with three biological replicates at nine levels of concentrations as in Table 4.5. The dataset for study III from site 52 was used for evaluation. Peaks were detected and quantified with MultiQuant. We evaluated the sensitivity for detecting six different true fold changes in protein abundance between D-513 fmol/μL and six largest concentrations (1500/513, 2760/513, 4980/513, 9060.513, 16500/513, 30000/513).

Mouse liver mitochondrial protein lysate [73]

100ng and 33ng mitochondrial lysates from mouse liver cell were processed in duplicates and three injection replicates, total twelve MS runs. Samples were acquired in DIA. Total 25 proteins were quantified. All proteins were expected to change

Condition	Concentration (fmol/μL)
D	513
E	1500
F	2760
G	4980
H	9060
I	16500
J	30000

Table 4.5.: Concentrations of spike-in proteins in each condition for CPTAC SRM dataset

between conditions (100ng vs. 33ng) with 3 fold changes. Dataset was processed by Skyline. We evaluated the ability of statistical methods to detect three fold changes.

Profiling standard sample set [74]

Twelve non-human proteins, grouped into three master mixtures, were spiked into HEK-293. Each mixture was diluted in eight different concentrations with triplicate as in Table 4.6. Therefore, there were 24 MS runs in total. The DIA dataset is available in the supplementary of [74], which was processed with Spectronaut 5. The raw data also were re-processed in Skyline with centroided with 15 ppm mass accuracy and 5 minute retention time tolerance. We used only quantifiable peptides with q-value less than 0.01 from mProphet in Skyline. The features with q-value greater than 0.01 were replaced with censored missing values. Among 3 mixtures, we used two master mixers, Mix1 and Mix2 (5 proteins per mixer), which were diluted with small concentrations, for evaluation. We evaluated the statistical methods in terms of the sensitivity for different fold changes from lowest concentration baseline (1.1, 1.21, 1.33, 10, 11.01, 12.11 13.33 for Mix1 and 1.59, 2.52, 4, 50, 79.37, 125.99, 200 for Mix2) and the specificity for constant background proteins across samples.

Protein		Sample 1	2	3	4	5	6	7	8
Mix1	concentration(fmol/μL)	1.5	1.65	1.815	1.995	15	16.515	18.165	19.995
	relative concentration	**1**	**1.1**	**1.21**	**1.33**	**10**	**11.01**	**12.11**	**13.33**
Mix2	concentration(fmol/μL)	100	62.995	39.685	25	2	1.26	0.795	0.5
	relative concentration	**200**	**125.99**	**79.37**	**50**	**4**	**2.52**	**1.59**	**1**
Mix3	concentration(fmol/μL)	0.05	0.2	0.8	3.2	12.8	51.2	204.8	819.2
	relative concentration	**1**	**4**	**16**	**64**	**256**	**1024**	**4096**	**16384**
Background	concentration(fmol/μL)	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	relative concentration	1	1	1	1	1	1	1	1

Table 4.6.: Sample series of DIA Profiling standard sample set. Mix 1 includes 5 proteins (P02754, P80025, P00921, P00366, P02662), Mix2 contains 5 proteins (P61823, P02789, P12799, P02676, P02672), and Mix3 has 2 proteins (P02666, P68082).

Gold standard data [40]

The dataset included 16 proteins in ten dilution steps into three different backgrounds(water, whole-cell protein extracts from human or yeast). Detailed concentrations are available in Table 4.7. We used the datasets with human cell protein extracts background. The DIA datasets from supplementary of [40], which was processed with OpenSWATH, were used with filtering out peaks with greater than 0.01 of m-score. We also processed the same raw files in Skyline with 40,000 rp and +/- 5 min and then filter out peaks by q-value greater than 0.01 from mProphet. The peaks with q-value greater than 0.01 were replaced with censored missing values. We evaluated the ability of statistical methods to detect nine different fold changes from maximum concentration as baseline (2, 4, 8,16, 32, 64, 128, 256, 512 true fold changes).

4.1.2 Biological and clinical investigations.

Cardiovascular disease study [6]

This study was for investigation for cardiovascular disease between control and four disease stages (0, 1, 2, 3, 4 in `Condition`). 246 samples from control and disease

Condition (Dilution)	Concentration (amol/μL)
1x	30000
2x	15000
4x	7500
8x	3750
16x	1875
32x	937.5
64x	468.75
128x	234.38
256x	117.19
512x	58.6

Table 4.7.: Concentrations of spike-in proteins for dilution steps in gold standard data

patients were analyzed with single injection by label-free DDA as described in [75]. There are 77 identified proteins. The dataset was processed by Monarch, `http://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapid=20704167`). Unusually, this DDA dataset had no missing values because the procedure reported the background signal if a feature in a run was not detected.

A study of subjects with ovarian cancer [76]

Original published raw data, SRM with isotope labeled reference peptides, has total 83 patients plasma samples. Skyline succeeded to analyze 81 patients samples. The dataset including 66 ovarian cancer (OC) patients and 15 patients with benign ovarian tumors was used to evaluate. Each patient sample measured once without technical replicate. Total 36 proteins were used to evaluate the ability of statistical method to detect differential abundance proteins between OC and benign groups.

The training and the validation sets from a study of subjects with malignant pleural mesothelioma [10]

To identify candidate biomarkers for MPM in serum, the experiment targeted 51candidate peptides with SRM with isotope labeled reference peptides. There were two datasets. First was the training set including total 75 subjects: 25 MPM, 25 healthy donors(HD), 25 non-small cell lung cancer (NSCLC). Second, the validation set consisted of total 75 subjects: 23 MPM, 26 HD, 26 NSCLC. Each sample was injected once without technical replicate. All samples were processed by Skyline.

The training and the validation sets from a study for subjects with colorectal cancer [11]

70 proteins were targeted for plasma samples with SRM with isotope labeled reference peptides in order to identify candidate protein biomarker for non-invasive detection of CRC. The training cohort included 100 subjects in control group and 100 subjects with CRC. The validation cohort had 67 subjects in controls, and 202 subject with different clinical stages of CRC. Each sample for subject was measured in a single injection without technical replicate. The training cohort was analyzed with Skyline. The validation cohort was processed with MultiQuant 1.2.

Time course investigation of central carbon metabolism of *S. cerevisiae* [77]

45 proteins in the glycolysis/gluconeogenesis/TCA cycle/glyoxylate cycle network were targeted in the experiment. Three biological replicates were measured at ten time points (T1-T10). It covered dynamic growth phases of *S. cerevisiae*, in a glucose-rich

medium (T1-T4), diauxic shift (T5-T6), post-diauxic phase (T7-T9), and stationary phase (T10). Each transition was quantified automatically using MultiQuant with no missing values.

Human liver micro tissue study with APAP treatment [74]

This study identified differential abundance proteins by four different concentrations of APAP in human liver micro tissues from DIA dataset. There are 5 conditions (one control, S1, and four concentrations of APAP, S3, S4, S7, S9) with three biological replicates per condition. Total 2788 proteins were measured. Increasing concentrations of APAP was expected to detect more number of proteins as changed. The dataset from supplementary of [74] was used, which was analyzed with Spectronaut 5 with default settings.

Saccharomyces cerevisiae proteome quantification [78]

S. cerevisiae cell cultures in biological triplicates were sampled at six time points (0 min (T0), 15 min(T1), 30 min (T2), 60 min (T3), 90 min (T4), 120 min (T5)) after osmotic stress. SWATH data was extracted by the Spectronaut. To validate the fold changes in proteins of SWATH, 100 proteins were quantified with SRM. Among them, measurements from 90 proteins are available in the supplementary of [78]. The data with SRM was processed using Skyline.

4.2 Evaluation strategy

4.2.1 Methods used in evaluation

We compared the performance of the proposed framework to three other summarization methods.

TMP To evaluate the importance of modeling missing values via a censoring mechanism, we considered the split-plot approach with Tukey median polish summarization, without imputing missing values (Section 3.3.3).

Linear model To evaluate the importance of both modeling missing values via a censoring mechanism and robust estimation, we considered the split-plot approach with linear model summarization (Section 3.3.1).

log(sum) Log(sum) is frequently used, and is implemented in signal processing tools such as Skyline and MaxQuant. The summarization consists of summing the peak intensities in the run on the original scale, and applying the $\log_2$ transformation to the sum. This summarization effectively gives higher weights to peaks with higher intensities, and sets the intensities of missing values to 0.

In order to produce comparable results, the summarized intensities in a run were analyzed with the same family of linear mixed effects model for all the approaches above. The specific models varied between the datasets, to appropriately reflect the experimental design as described in [7]. For example, the model for group comparison designs in presence of biological and technical replicates is given in Figure 3.4(B). For all the datasets we used the whole plot model to compare protein abundances between all possible pairs of conditions, while controlling the False Discovery Rate separately for each comparison at the significance level, 0.05.

4.2.2 Criteria for evaluation

We evaluated the performance of the methods with respect to the goals of testing proteins for differential abundance between conditions, and estimating the associated log-fold change. For the controlled mixtures the true changes in abundance are known. Therefore, the ability of the methods to detect differentially abundant proteins was evaluated in terms of sensitivity, specificity and positive predictive value (PPV), defined in Table 4.8 and Eq. (4.1). The ability of the methods to accurately estimate the log-fold changes was evaluated in terms of Mean Squared Error (MSE) defined in Eq. (4.2).

Number of proteins		Decision	
		Differentially abundant	Not differentially abundant
Truth	Differentially abundant	True positive (TP)	False negative (FN)
	Not differentially abundant	False positive (FP)	True negative (TN)

Table 4.8.: Outcomes of testing proteins for differential abundance between conditions in a controlled mixture.

$$\text{Sensitivity} = \frac{TP}{TP+FN},\ \text{Specificity} = \frac{TN}{TN+FP},\ \text{and}\ PPV = \frac{TP}{TP+FP} \quad (4.1)$$

$$MSE = \frac{\sum_{proteins}\ \sum_{condition\ pairs} \{\text{estimated } \log_2(\text{fold change}) - \text{true } \log_2(\text{fold change})\}^2}{\#\text{ proteins} \times \#\text{ condition pairs}} \quad (4.2)$$

For the biological and clinical investigations the true number of differentially abundant proteins and the true log-fold changes are unknown. Therefore, we simply reported the number of detected proteins that change abundance between conditions.

4.2.3 Summarization of criteria across statistical methods and signal processing tools

Signal processing tools make many choices regarding identification of peaks and reporting their intensity, all of which impact both testing for differential abundance and estimation of log-fold changes. Therefore, we processed the experimental datasets with up to four signal processing tools. The DDA datasets were processed with MaxQuant, Progenesis, Skyline and ProteomeDiscover or SuperHirn. The DIA datasets were processed with Skyline, Spectranaut and OpenSWATH. The SRM datasets were processed with Skyline and with MultiQuant.

This manuscript does not aim at comparing the performance of various signal processing tools. Instead, our goal is to verify that the proposed downstream data analysis framework performs well regardless of the signal processing. Therefore, for the controlled mixtures we reported relative performance of detecting differential abundance of the four approaches (Proposed, TMP, Linear model, log(sum)), separately for each signal processing tool, and separately for each true fold change. This was calculated as

$$\begin{aligned} \text{relative sensitivity} &= \frac{\text{sensitivity}}{\max \text{sensitivity}\{\text{Proposed, TMP, Linear model, log(sum)}\}} \\ \text{relative specificity} &= \frac{\text{specificity}}{\max \text{specificity}\{\text{Proposed, TMP, Linear model, log(sum)}\}} \end{aligned} \tag{4.3}$$

Larger values indicate better performance. The best performant approach has relative sensitivity and specificity of 1. Similarly, we reported relative performance of estimation of log-fold changes of the four approaches (Proposed, TMP, Linear model, log(sum)), separately for each signal processing tool, and separately for each true fold change. This was calculated as

$$\text{relative } MSE_{approach} = \frac{MSE_{approach}}{\max MSE\{\text{Proposed, TMP, Linear model, log(sum)}\}} \tag{4.4}$$

Smaller values indicate better performance. The best performant approach has relative sensitivity and specificity of 1.

4.3 Evaluation results

4.3.1 Robust parameter estimation for summarization accounts outliers.

First we evaluated robust parameter estimation by TMP for outlier issue. If there is no problem for quality of data, which means no outliers, no missing values, small variability across features, four summarization methods generate similar performances. Figure 4.1 illustrates the example proteins with different pattern of outliers using the dataset, iPRG 2015 DDA, as described in Section 4.1.1 and processed by Skyline. Figure 4.1(A) shows the profile plot with all individual measurements for each peptide ion and each run and summarized intensities by different summarization methods in one example protein, TIM9. It has no outlier and no missing value. Also summarized intensities across different methods are similar except by log(sum). Even though summarized intensities by log(sum) method are higher than other, it is parallel across runs and has similar variability with other methods. Estimated fold changes and even adjust p-values are similar across summarization methods. However, different outlier pattern affect differently to the summarization methods. Linear model-based summarization is easily affected by outliers in the peptides or features with both high and low intensity. Figure 4.1(B) presents the example protein with outliers in low intensities, in Run=3 for Condition 1. In this case we can see that summarized intensities for run with lower outlier with linear model (yellow dot in Run=3) is lower than other summarized intensities with other methods. It makes that estimated fold change with linear model is worst among summarization methods. Also, in Figure 4.1(C), the protein, which has outliers in high intensities, has worse estimated fold change with linear model than with TMP. log(sum) approach is insensitive by outliers in low intensity peptides, but it is still affected by outliers in high intensity peptides. That is because summing original intensities gives more weight to higher original intensities, which means that high intensities contribute more to summed intensities. Therefore it could be sensitive to variation for high

intensities. Figure 4.1(C) shows the example with outliers in the highest peptide at condition 3 and with more variability in high intensity. Even though significance testing results with FDR cutoff=0.05 are the same across summarization methods, estimated fold change with log(sum) is the worst. On the other hand, TMP method is least influenced by outliers. Figure 4.1(D) shows the example protein, which has some zero intensities at condition2 and 3, which are censored missing values, and also outliers in the peptide ions with low intensities at condition4. In this case, log(sum) method can not detect the change between condition2 and 3 with quite smaller estimated fold change than true fold change. Linear model-based summarization has worst estimated fold change, even can not detect the change between condition2 and 3. Therefore TMP method performs better than log(sum) and linear model-based method. With TMP approach, Imputation for missing values before TMP helps the performance than TMP alone in terms of estimated fold change.

4.3.2 Model-based imputation before robust estimation accounts missing values.

Even though TMP is least sensitive for outlier, TMP method can be affected by missing values. TMP is median-based estimation with observed measurement. Therefore less number of observed measurement can be change TMP summarization. Then imputation for missing values before TMP can improve the performance. In order to show how imputation affects the performance, Figure 4.2 demonstrates two example proteins from the a label-free controlled mixture dataset in Section 4.1.1, which have some outliers and missing values in low intensities. The protein in Figure 4.2(A) has the best performance with the proposed method in terms of sensitivity with the same specificity across methods. The comparison Mix2-Mix1 is detected as significant difference only with the proposed method. Mix1 has many missing values in low intensities. Therefore after imputing missing values, TMP summarization works better than without imputation or other methods. Compared with the protein

in Figure 4.2(A), the example in Figure 4.2(B) has more number of peptide than Figure 4.2(A). Then it can be less affected by summarization methods. However, it still has outliers and many missing values in low intensities. In this example protein, Impute+TMP has better specificity with the same sensitivity than TMP. The table in Figure 4.2(B) shows that the comparison Mix1-Mix4 is detected as significant difference incorrectly in TMP. The proposed method also has better estimated fold change than log(sum) with the same sensitivity and specificity. In addition, we summarized how much the proposed method improves MSE compared to log(sum) across all 28 proteins in Figure 4.2(C). As the percentage of missing value increases, the proposed method improves MSE more than log(sum). Therefore, it is important how to summarize in subplot level If the datasets have many missing values or unequal variance between features, for example, DDA and DIA case. SRM datasets are expected to have similar performance across different methods because they have relatively few missing values and less variability between features.

4.3.3 Datasets and signal processing matter for performance

We showed some individual protein examples above sections in order to how the different methods summarized differently some special cases. Next, we evaluated the performance of methods with 4 DDA datasets, 3 SRM datasets, 3 DIA/SWATH datasets in order to see the performance across spectral acquisition experiments. For 4 DDA experiments, we processed by up to four signal processing tools among Skyline, MaxQuant, Progenesis, superHirn, Proteome Discoverer per DDA dataset. Figure 4.3, Figure 4.4, Figure 4.5, and Figure 4.6 visually summarized all performance with less than 100 true fold change across true fold changes, datasets, spectral acquisitions, and spectral processing tools. Overall, the performance varies between the datasets and is confounded by datasets and spectral tools. DDA experiment has more variability in quality of data, such as outliers and missing values. Therefore, it has major difference between methods. DIA is intermediate between SRM and DDA, which doesn't have

many outliers, but still has missing values in low intensities. SRM datasets with isotope labeled reference have mostly very clear, good quality of data, compared with other type of experiments. Therefore, most of summarization methods works very well similarly.

First, the abilities to detect proteins with known fold changes for different summarization methods are shown by the relative sensitivity and relative specificity as described in Eq. (4.3). In Figure 4.3(A), for DDA experiments, sensitivities with the proposed method and TMP are better than with log(sum) or linear model-based summarization and specificities are little better across datasets and spectral tools. For DIA datasets in Figure 4.4(A) sensitivities for more than fold change=4 are similar across 3 datasets and spectral processing tools. But, mostly the proposed method and log(sum) for fold changes less than 4 have similarly better results than others in terms of sensitivity across datasets and spectral processing tools. For SRM datasets in Figure 4.4(B), all different summarization methods work very well in terms of sensitivity. However, the proposed method works better than other methods in terms of MSE and specificity.

In some cases such as DDA of iPRG, 2015 processed by Progenesis, log(sum) have better sensitivity than the proposed method. But, also has worse specificity. In these cases, PPV tells more details in Table 4.9. In general, PPV for the proposed method are better than for log(sum). It means that log(sum) overfits. Therefore, it detects more number of false positive proteins as differentially abundant proteins, even though finding more number of true positive proteins as differentially abundant proteins. The previous version of MSstats, which use the full linear mixed effect model without considering split-plot approach, has the same issue. Even if It could have better sensitivities for some datasets or processing tools, it commonly has much more number of false positive proteins and generates worse specificity and worse PPV.

Second, accuracy of fold change estimation among different summarization methods can be evaluated with relative MSEs. As same as sensitivity, MSEs for DDA

DDA	Spike-in dataset														
	MaxQuant					Progenesis					Skyline				
	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2
TP	217	218	156	215	222	209	209	195	207	200	217	217	195	199	217
FP	16	15	3	12	24	8	10	4	5	11	11	11	8	7	24
total P	234	243	159	227	246	217	219	199	212	211	228	228	203	206	241
PPV	0.93	0.94	0.98	0.95	0.90	0.96	0.95	0.98	0.98	0.95	0.95	0.95	0.96	0.97	0.90
	Proteome discover														
	Proposed	TMP	Linear	log(sum)	v2										
TP	203	210	126	202	207										
FP	3	3	1	3	12										
total P	206	213	127	205	219										
PPV	0.99	0.99	0.99	0.99	0.95										

DDA	iPRG study, 2015														
	MaxQuant					Progenesis					Skyline				
	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2
TP	26	23	19	25	27	20	20	24	24	33	22	21	18	13	11
FP	5	4	5	28	160	125	97	144	199	898	16	16	50	98	390
total P	31	27	24	53	187	145	117	168	223	931	38	37	68	111	401
PPV	0.84	0.85	0.79	0.47	0.14	0.14	0.17	0.14	0.11	0.04	0.58	0.57	0.26	0.12	0.03

DDA	Dynamic range benchmark - Cox, 2014														
	MaxQuant					Progenesis					Skyline				
	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2
TP	31	26	22	29	32	26	20	13	31	39	39	39	35	36	38
FP	7	8	5	13	36	3	2	2	7	186	9	10	2	14	50
total P	38	34	27	42	68	29	22	15	38	225	48	49	37	50	88
PPV	0.82	0.76	0.81	0.69	0.47	0.90	0.91	0.87	0.82	0.17	0.81	0.80	0.95	0.72	0.43

DIA	Profiling standard sample set - Bruderer, 2015									
	Spectronaut					Skyline				
	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2
TP	68	63	53	70	65	60	51	48	62	55
FP	889	869	566	1045	1012	303	402	55	297	1692
total P	957	932	619	1115	1077	363	453	103	359	1747
PPV	0.07	0.07	0.09	0.06	0.06	0.17	0.11	0.47	0.17	0.03

SRM	Controlled spiked-in data - Chang, 2012					Spike-in dataset				
	MultiQuant					MultiQuant				
	Proposed	TMP	Linear	log(sum)	v2	Proposed	TMP	Linear	log(sum)	v2
TP	4	0	2	2	25	190	191	192	192	192
FP	0	0	0	0	2	2	2	12	8	18
total P	4	0	2	2	27	192	193	204	200	210
PPV	1	0	1	1	0.93	0.99	0.99	0.94	0.96	0.91

Table 4.9.: PPV across run-level summarization methods, datasets and spectral processing tools. Three DDA datasets, one DIA dataset, and two SRM datasets, which have both differentially abundant proteins and constant proteins between comparisons, are presented. 'v2' represents the previous version, v2, of MSstats.

experiments have more variability among methods and known fold changes in Figure 4.5, even though MSEs for the proposed method in most DIA and SRM experiments perform better than others in Figure 4.6(A) and (B) .

Beyond relative comparisons between four methods, we compared two paired summarization methods, the proposed method vs log(sum) and the proposed method vs TMP, by absolute difference between original sensitivities or MSEs. The comparison between the proposed method vs log(sum) shows how much the proposed method

performs better than log(sum) in Figure 4.7(A) and (B) by acquisition experiments by the percentage of missing values. We see that the proposed method works better than log(sum). In DDA, there are more improvements than DIA and SRM. In DIA, cases with more missing values have more improvements in MSE. Even SRM has improvement in terms of MSE. The comparison between the proposed method vs TMP shows how much imputation improves performance than TMP in Figure 4.7(C) and (D). DIA and SRM, which have less outliers issue, have better performance for Imputation than TMP only for large percentage of missing intensities. Because DIA have more missing value than SRM, the improvement for DIA is bigger than SRM. We can tell that imputation is helpful in case of lower fold changes which has more missing values, compared with TMP alone, that means missing value is main issue for DIA analysis. Even though SRM has similar performance across summarization method, imputation help to improve MSEs for lower true fold changes, which has more missing values. In DDA experiment, there are some improvements for sensitivities, but worse MSEs than TMP. That is because outliers could affect imputation and then bias could be generated and outliers and variation across peptides were bigger difficulty than missing values for DDA.

4.3.4 Performance of detecting differentially abundant proteins in biological investigations

Usually controlled mixture dataset has small number of replicates with small number of biological replicates or without any biological replicates. However, actual experiments for research have more technical and biological replicates in complex design of experiment. To check the difference between summarization methods in real experiments, we tested ten published biological investigation datasets with various experiments, across acquisition methods, different diseases, also time-course design of experiment.

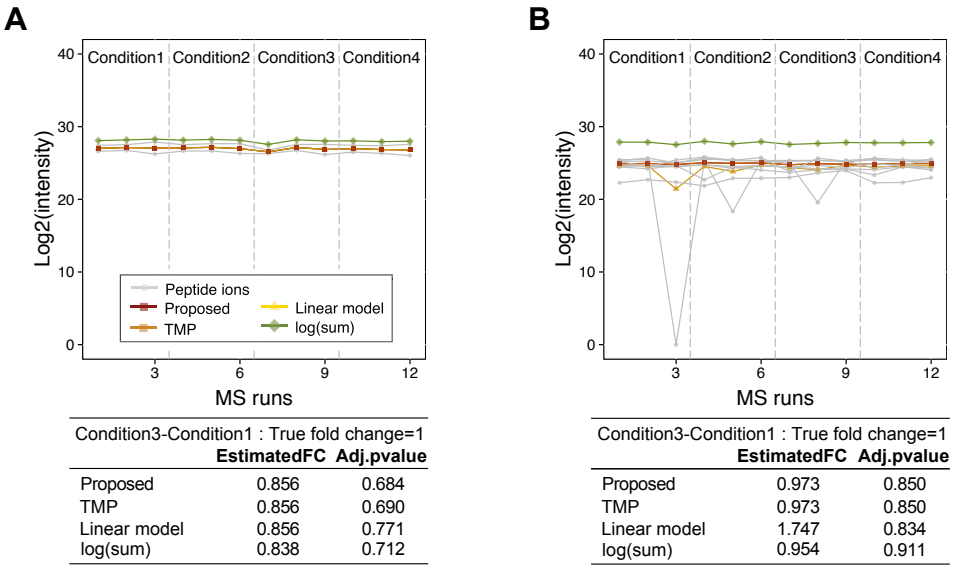
Figure 4.8 shows the number of significant proteins in each comparison across different summarization methods, because we do not know true changes between comparison and can not calculate statistical performance. Original difference between numbers of significant proteins in most of comparisons especially for SRM is very small, such as only one or two proteins are different in list of detected significant proteins. The difference between the methods is less apparent, because of the larger sample size and more careful hand curation, large sample size, more targeted. Mostly log(sum) has largest number of significant proteins for many comparisons. It is similar with the result of controlled mixture datasets in terms of total number of detected proteins as differentially abundant, even though log(sum) is less sensitive for true positive proteins as Table 4.9. It means that there is the possibility that log(sum) can detect more false positive proteins by overfitting for biological studies.

4.4 Discussion

Even though the proposed method takes advantage of both imputation and TMP in the summarization stage, there are some potential disadvantages. TMP helps to remove outlier effects. On the other hand, we could throw away potentially useful information from outliers. The imputation strategy can possibly add bias with imputed values in the summarization level. Also, the quality of the data including outliers can affect the imputation; because the threshold for censored missing values is decided among all observed intensities. Therefore, DDA datasets, which usually have issues with outliers and miss values, do not have much improve with imputation.

In addition, the proposed approach still cannot impute the missing value if there is no measurement at all in a particular run or in a particular condition. The proteins with completely missing intensities in a particular condition can be interesting candidates for some types of experiments, such as PTM.

The performance varies between the datasets and is confounded by datasets and spectral tools. Some datasets are easier than others. Some have aspects that affect the performance of individual methods in specific ways. This emphasizes several points. First, it is important to evaluate the proposed method over multiple datasets and multiple spectral processing tools. The newly suggested statistical models in the journal are tested usually with one or two datasets by one specific processing tool. But, the results can be the evaluation for particular datasets and processing tools. Second, it is also important to process the data correctly with an understanding of spectral processing tools, prior to statistical analysis for protein differential abundance.



Condition3-Condition1 : True fold change=1

	EstimatedFC	Adj.pvalue
Proposed	0.856	0.684
TMP	0.856	0.690
Linear model	0.856	0.771
log(sum)	0.838	0.712

Condition3-Condition1 : True fold change=1

	EstimatedFC	Adj.pvalue
Proposed	0.973	0.850
TMP	0.973	0.850
Linear model	1.747	0.834
log(sum)	0.954	0.911

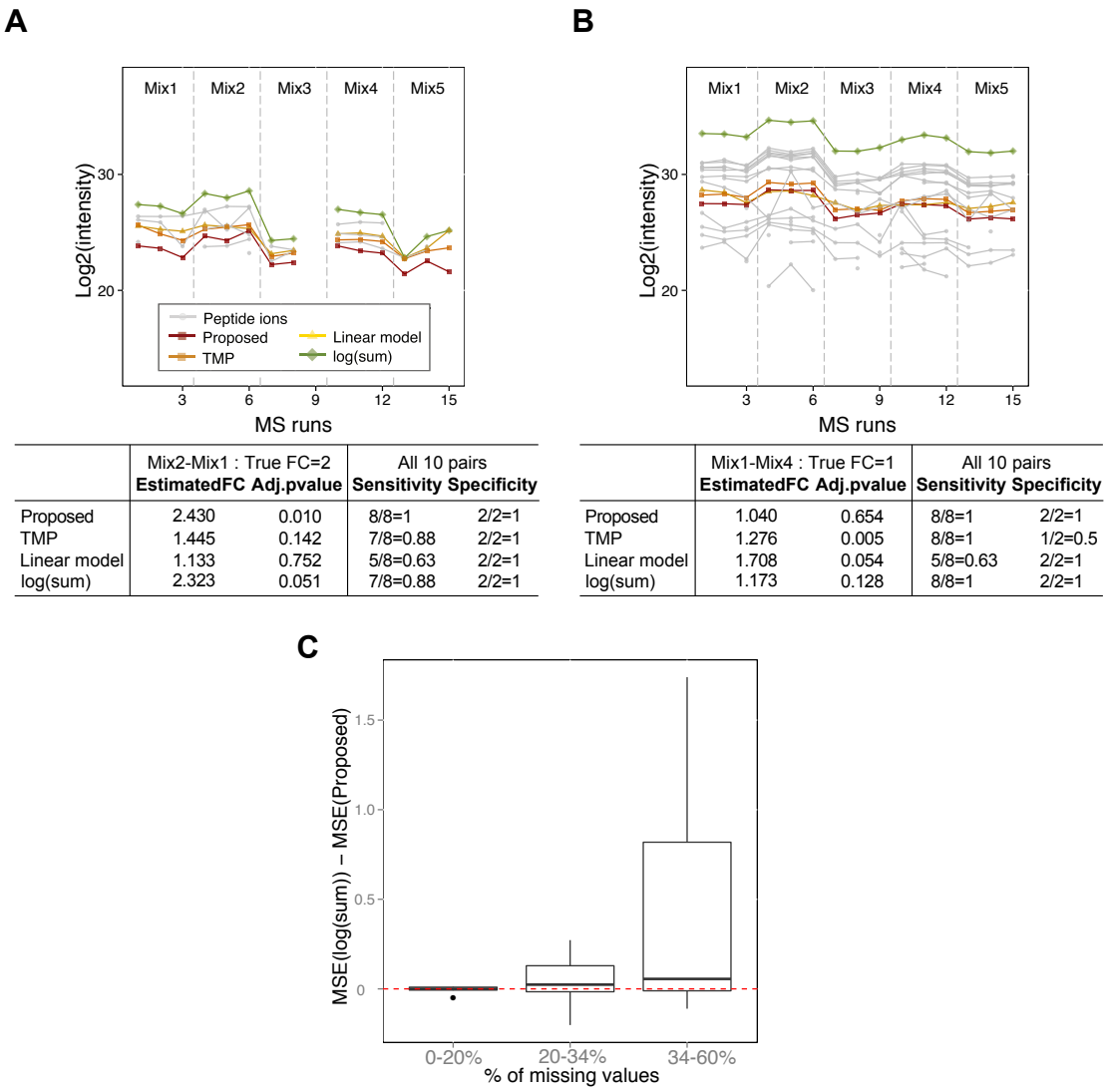
Condition3-Condition1 : True fold change=1

	EstimatedFC	Adj.pvalue
Proposed	0.881	0.915
TMP	0.881	0.916
Linear model	0.685	0.834
log(sum)	0.383	0.794

Condition2-Condition3 : True fold change=7.5

	EstimatedFC	Adj.pvalue
Proposed	7.688	0.001
TMP	5.990	0.001
Linear model	31.03	0.008
log(sum)	1.701	0.895

Figure 4.1.: Profile plots with processed feature-level intensities and run-level summarized intensities and testing results for example proteins with or without outliers from iPRG 2015 DDA dataset (Section 4.1.1). Legend for profile plot in the box of 2(A) : Gray dots show log 2 transformed and normalized feature-level intensities individually and line means each peptide ions. Colored dots and lines show run-level summarized intensities by different summarization methods. Tables for each protein show estimated fold change(FC) and adjusted p-value(Adj.pvalue) across different run-level summarization methods in rows. (A) No outlier: Protein TIM9, (B) Outliers in low intensity: Protein INV2, (C) Outliers in high intensity: Protein SIR3, (D) Outliers in low and high intensity: Protein ISCB



	Mix2-Mix1 : True FC=2		All 10 pairs	
	EstimatedFC	**Adj.pvalue**	**Sensitivity**	**Specificity**
Proposed	2.430	0.010	8/8=1	2/2=1
TMP	1.445	0.142	7/8=0.88	2/2=1
Linear model	1.133	0.752	5/8=0.63	2/2=1
log(sum)	2.323	0.051	7/8=0.88	2/2=1

	Mix1-Mix4 : True FC=1		All 10 pairs	
	EstimatedFC	**Adj.pvalue**	**Sensitivity**	**Specificity**
Proposed	1.040	0.654	8/8=1	2/2=1
TMP	1.276	0.005	8/8=1	1/2=0.5
Linear model	1.708	0.054	5/8=0.63	2/2=1
log(sum)	1.173	0.128	8/8=1	2/2=1

Figure 4.2.: Profile plots with processed feature-level intensities and run-level summarized intensities and testing results for example proteins including missing values. Controlled spike-in DDA in Section 4.1.1 is used. Legend for profile plot is the same as Figure 4.1. (A) Protein, P02753, has 61% missing values. (B) Protein, P00563, has 31% missing values. (C) Improvement in fold change estimation of the proposed method as compared to log(sum). Y-axis is the difference between MSE for log(sum) and MSE for Imputation+TMP among 10 possible pairs for each protein. Positive values in y-axis means that Imputation+TMP method performs better for fold change estimation than log(sum) method. X-axis is the group of the percentage of missing values among 28 proteins. The percentage of missing values is calculated by dividing the number of NA measurements by the required number of measurements (the number of MS runs $\times$ the number of peptides)

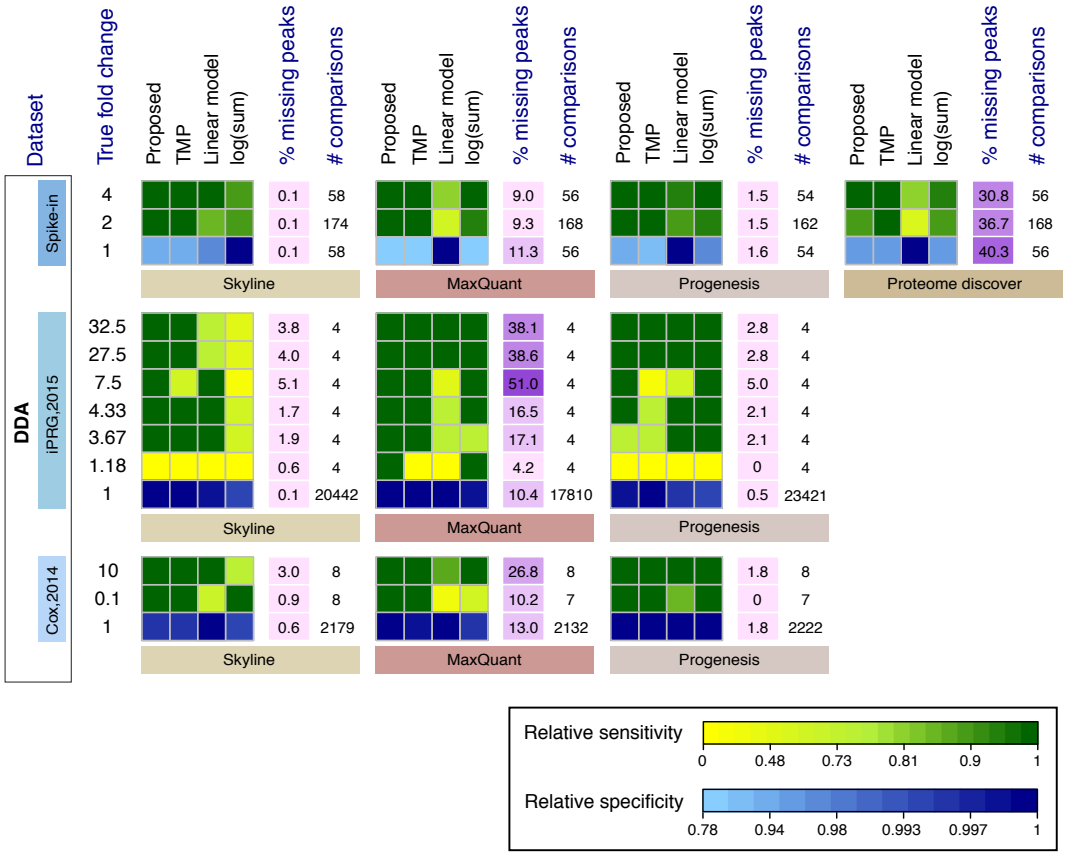


Figure 4.3.: Relative sensitivity or specificity of run-level summarization methods, evaluated on 3 controlled mixtures of DDA. (A) Relative sensitivity and specificity of pairwise comparisons, for the true fold change below 100. Each panel quantified with different signal processing tools. Colors indicate relative sensitivity (true fold changes different from 1) and relative specificity (true fold changes equal to 1), calculated by standardizing each sensitivity and specificity by the maximum value in each row, separately by the panel. Darker green or blue indicate better performance as shown in color key.

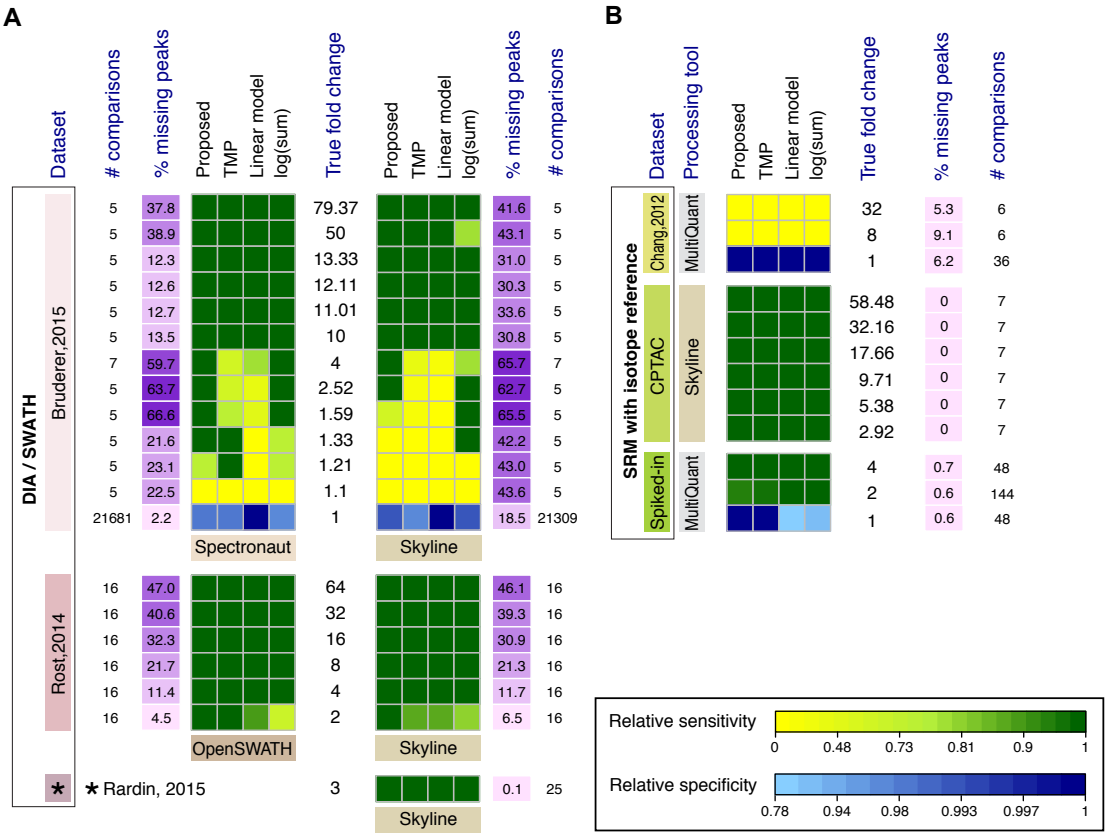


Figure 4.4.: Relative sensitivity or specificity of run-level summarization methods, evaluated on controlled mixtures of DIA and SRM. (A) DIA datasets, relative sensitivity, specificity of pairwise comparisons, for the true fold change below 100. Left panel for peaks intensities quantified with the original signal processing tools, Spectronaut or OpenSWATH. Right panel for peaks intensities quantified with Skyline. Colors are as in Figure 4.3. (B) SRM datasets, relative sensitivity, specificity of pairwise comparisons, for the true fold change below 100. Colors are as in (A).

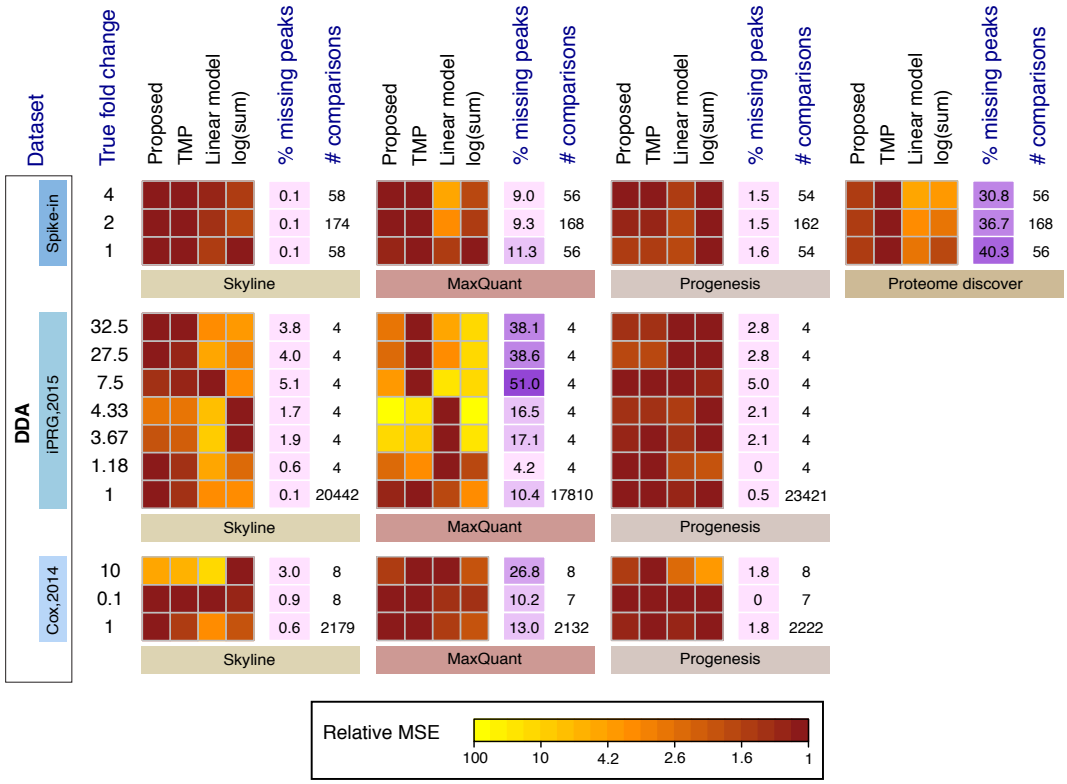


Figure 4.5.: Accuracy of fold change estimation of run-level summarization methods among pairwise comparisons, for the true fold change below 100, evaluated on 3 controlled mixtures of DDA. Each panel quantified with different signal processing tools. Colors indicate relative MSEs. Darker reds indicate better performance as shown in color key.

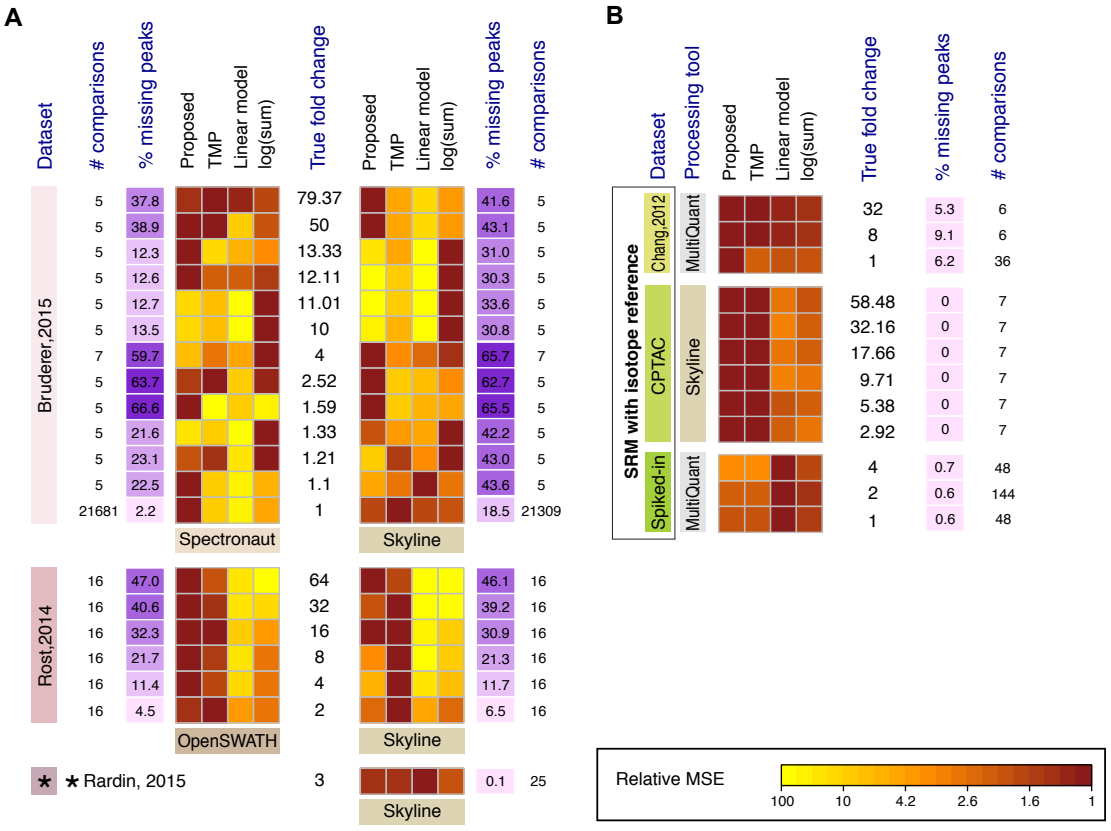


Figure 4.6.: Accuracy of fold change estimation of run-level summarization methods, evaluated on 3 controlled mixtures of DIA and 3 of SRM datasets. (A) DIA datasets. Relative MSEs of pairwise comparisons, for the true fold change below 100. Each panel quantified with different signal processing tools. Left panel for peaks intensities quantified with the original signal processing tools, Spectronaut or OpenSWATH. Right panel for peaks intensities quantified with Skyline. (B) SRM datasets, relative MSE of pairwise comparisons, for the true fold change below 100. Colors indicate relative MSEs. Darker reds indicate better performance as shown in color key.

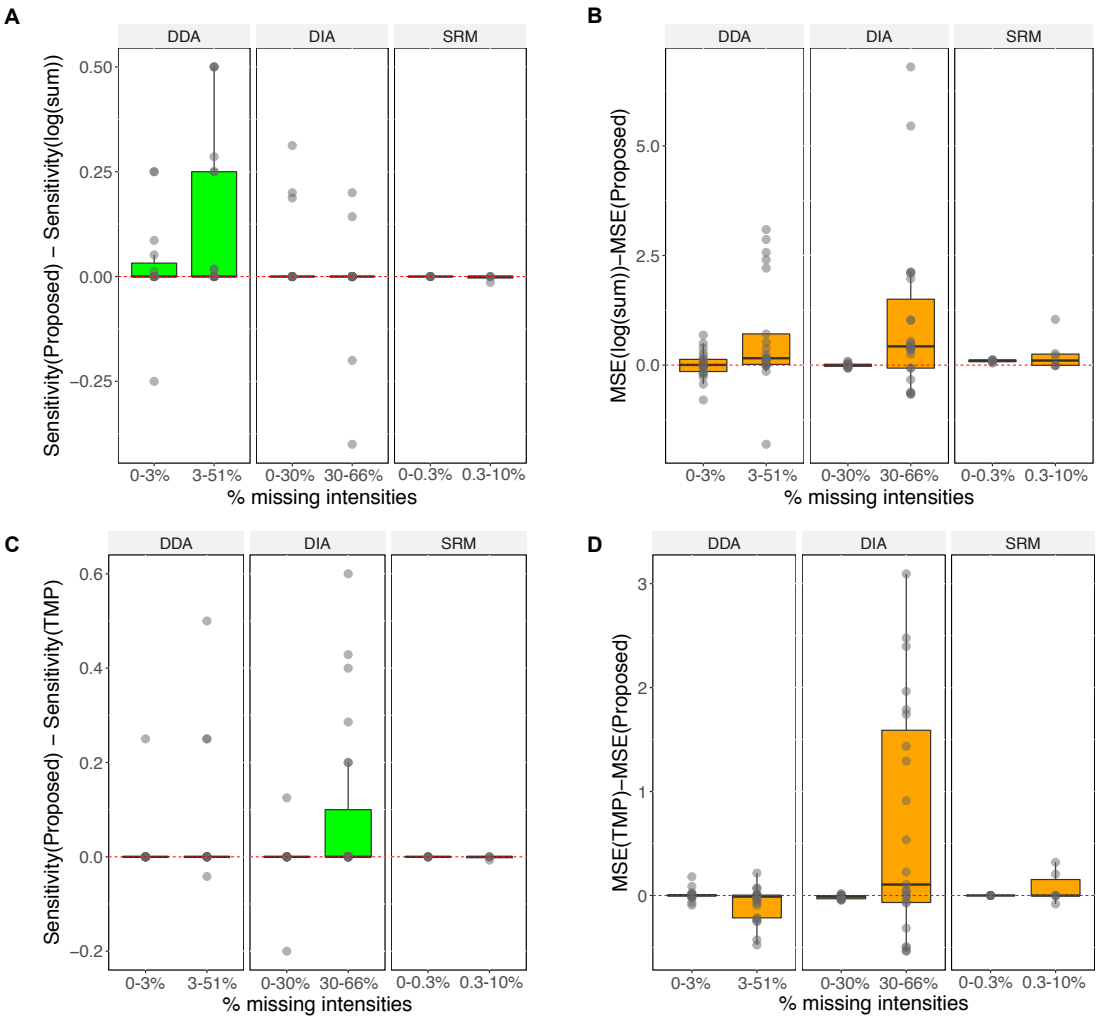


Figure 4.7.: Comparison between performance of run-level summarization methods across datasets and spectral processing tools. (A) Improvement in sensitivity of the proposed method as compared to log(sum), combined for all the datasets and spectral processing tools and separated by the percentage of missing values and by dataset type. (B) Improvement in fold change estimation of the proposed method as compared to log(sum), combined for all the datasets and spectral processing tools and separated by the percentage of missing values and by dataset type. (C) Improvement in sensitivity the proposed method as compared to TMP. (D) Improvement in fold change estimation of the proposed method as compared to TMP.

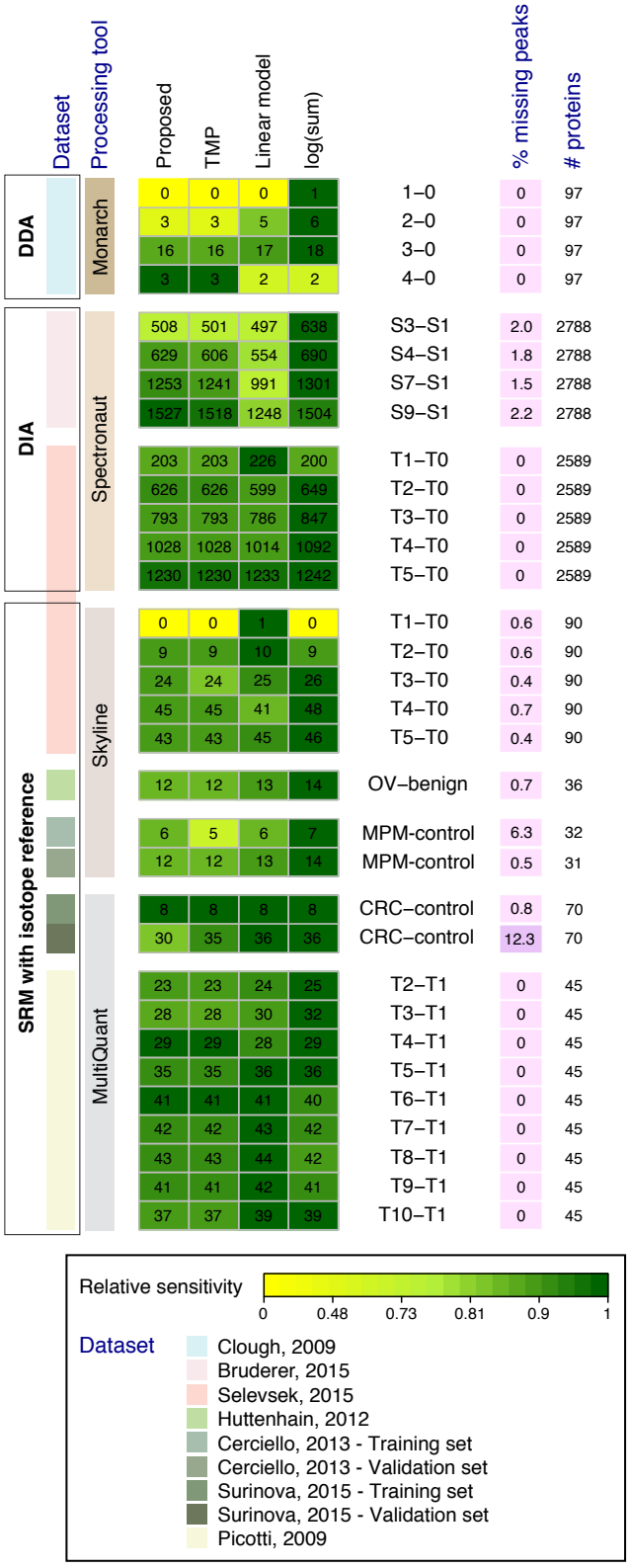


Figure 4.8.: Sensitivity of run-level summarization methods in biological and clinical investigations. Colors are as in Figure 5. The numbers in cells are the number of differentially abundant proteins for each comparison. Since these datasets have no ground truth, the relative specificity of the four statistical approaches is unknown.

5. OPEN-SOURCE SOFTWARE AND IMPLEMENTATION

5.1 R package, MSstats, statistical tool for quantitative MS proteomics

MSstats is an open-source R-based package for statistical relative quantification of peptides and proteins in mass spectrometry-based proteomic experiments that integrates the suggested methodology from our group across several mass spectrometric workflows and data acquisition strategies, contains functionalities for model-based analyses and enable interoperability with the existing popular spectral processing tools and computational tools. It takes as input identified and quantified spectral peaks, and outputs a list of differentially abundant peptides or proteins, or summaries of peptide or protein relative abundance.

For special cases of some experimental workflows, the underlying statistical methodology was previously implemented in R-based packages MSstats 1.0 [7] and SRMstats [8,79]. MSstats 2.0 supersedes MSstats 1.0 and SRMstats, in that it implements all the analysis steps that are available in these packages. In addition, it extends the methodology and the implementation across three acquisition methods (SRM, DDA and DIA) and labeling strategy ((label-free, and workflow using labeled reference proteins or peptides). MSstats 3.0 improves the statistical method as proposed above chapter and execution time and also facilitates the interoperability with existing computational tools. I describe the most recent version, MSstats 3.0, in this section.

5.1.1 Applicability

MSstats is applicable to multiple types of sample preparation, including label-free workflows, workflows that use stable isotope labeled reference proteins and peptides,

and workflows that use fractionation. It is applicable to targeted SRM, DDA or shotgun, DIA or SWATH-MS. It is applicable to experiments that make arbitrary complex comparisons of experimental conditions or times.

MSstats performs statistical analysis steps, that follow peak identification and quantitation. Therefore, input to MSstats is the output of other spectral processing software tools (such as Skyline or MultiQuant) that read raw spectral files and identify and quantify spectral peaks. The output from any spectral processing tool, which satisfies the required information, can be applicable for MSstats.

5.1.2 Statistical functionalities

MSstats performs three analysis steps as in Figure 5.1. The first step, *data processing and visualization*, performs log transformation of intensities and normalizes the intensities of the peaks, and summarizes the protein abundance for subplot level. Then it generates workflow-specific and customizable numeric summaries for data visualization and quality control.

The second step, *whole plot inference*, automatically detects the experimental design (e.g. group comparison, paired design or time course, presence of labeled reference peptides or proteins) from the data. It then reflects the experimental design, and fits an appropriate linear mixed model by means of `lm` and `lmer` functionalities in R. The model is used to detect differentially abundant proteins or peptides.

The third step, *statistical experimental design*, views the dataset being analyzed as a pilot study of a future experiment, utilizes the variance components of the current datasets, and calculates the minimal number of replicates necessary in the future experiment to achieve a pre-specified statistical power.

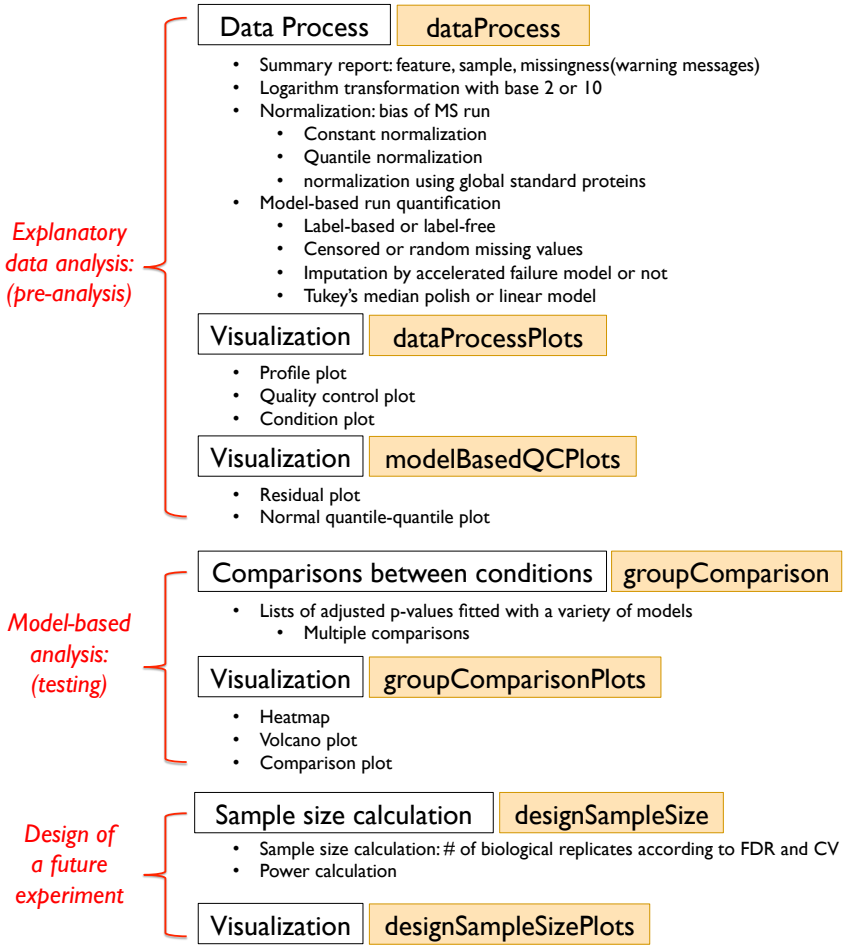


Figure 5.1.: Overview of the functionalities and of the associated functions in MSstats. Colored boxes indicate the actual function names.

5.1.3 Suggested statistical analysis workflow for MS experiments

Required input

MSstats performs statistical analysis steps, that follow peak identification and quantitation. Therefore, input to MSstats is the output of other software tools (such

as Skyline or MultiQuant) that read raw spectral files and identify and quantify spectral peaks. MSstats requires 10 columns input including the following variables: 'ProteinName', 'PeptideSequence(or PeptideModifiedSequence)', 'PrecursorCharge', 'FragmentIon', 'ProductCharge', 'IsotopeLabelType', 'Condition', 'BioReplicate', 'Run', 'Intensity'. There are several convenient ways to make the required input from popular spectral processing tools. (1) Skyline supports MSstats input report format, which automatically extracts all required input columns for MSstats and additional columns, 'Truncated' and 'StandardType', which can help to control quality of data. (2) MSstats provides the function, `MQtoMSstatsFormat`, to convert the output of MaxQuant to required input for MSstats. (3) SWATH data from OpenSWATH software can be reformatted by R package, SWATH2stats, in Bioconductor, which was developed to support the conversion the output from OpenSWATH to input for MSstats after MSstats release.

Pre-processing data and quality control of MS runs

After reading the input for analysis, data processing steps follow as below.

- Check the correctness of input such as correct input format, names of columns in data structure, correct options.
- Detect duplicate rows, which is multiple rows for a same feature in a same run which are generated by signal processing tools and warn the user should decide which rows should be used.
- Detect incomplete rows in the input. MSstats requires that the input contains a separate row for every feature in every run. If MSstats detects incomplete rows, it will output the list of problematic features. With the decision of the option for incomplete rows from user, the data processing can be stopped with list of incomplete rows or the incomplete rows can be filled with adding intensity=NA.

- Perform logarithm transform with base 2 (default) or 10 of the intensities.

- Do normalization to remove systematic bias between mass spectrometry runs. The normalization is applied after the logarithm transform. There are several options for normalization: (1) constant normalization shifts all the intensities in a run by a constant, to equalize the median of reference intensities across runs. (2) quantile normalization [80] applies a non-linear transformation to all the intensities in a run, to equalize the distribution of reference intensities across runs. (3) normalization with standard proteins, which are expected with equal amount across MS runs, is applied to endogenous intensities. First, the normalization equalizes endogenous intensities of global standard proteins across runs. Second, it applies the same between-run shifts to the remaining endogenous proteins in the experiment. For SRM experiments with stable isotope labeled reference peptides, the normalization is typically based on labeled reference peptides of all the proteins. If all the transitions in a biological or technical replicate are split into multiple methods (and are recorded in multiple files), this structure of the data is detected automatically by MSstats, by reading the values of the column 'Run'. In this case the normalization is performed separately for each method.

- Produce an output summarizing the experimental design including the warning message with the list of problematic features, subjects, conditions and their labels.

- Summarize feature-level intensities for protein-level per run by the proposed statistical approach as in Section 3.3.4

The output is the reformatted data that includes feature-level data for the data visualization and run-level summarized data for downstream model-based inference.

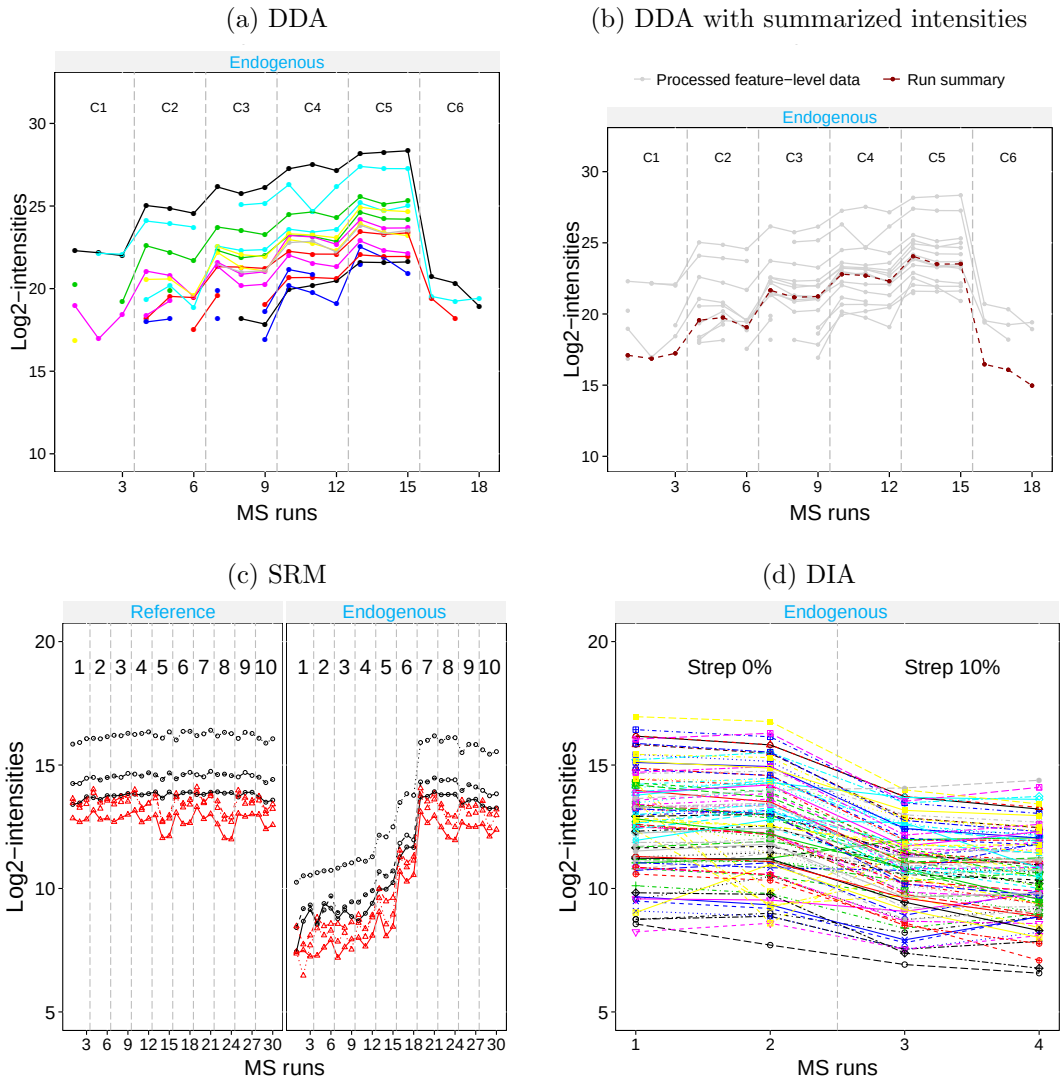
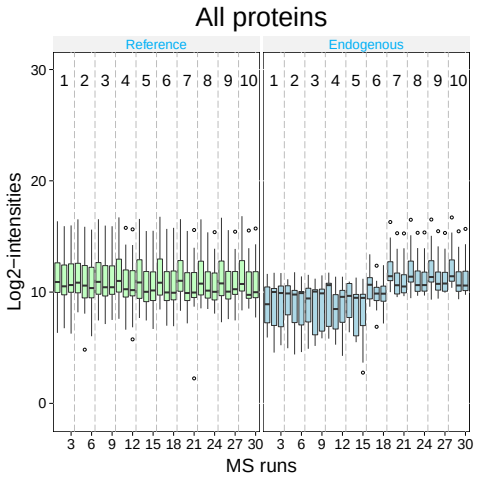
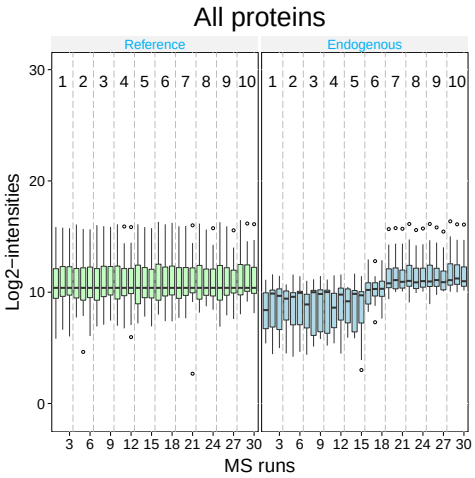


Figure 5.2.: Visualization of one representative protein in a DDA, an SRM and a DIA experiment. Colors represent peptides, and multiple line types of a same color represent the fragments of the peptide. Vertical lines separate times or conditions. (a) Protein Alcohol dehydrogenase-Yeast spiked into a complex background in 6 concentrations from DDA Spike-in dataset by Latin Square design [35] (b) With summarized intensities after subplot summarization for the same data in (a). Red dashed line shows summarized intensities. (c) Protein ACH1, at 10 times points after a stress. from a time course of *S. Cerevisiae* in Section 4.1.2 (d) Protein FabG of Streptococcus, with 0% and 10% human plasma added. a group comparison of *S. Pyogenes* from [40]. All three datasets are available in MSstats as example datasets.

(a) No normalization



(b) Constant normalization



(c) Quantile normalization

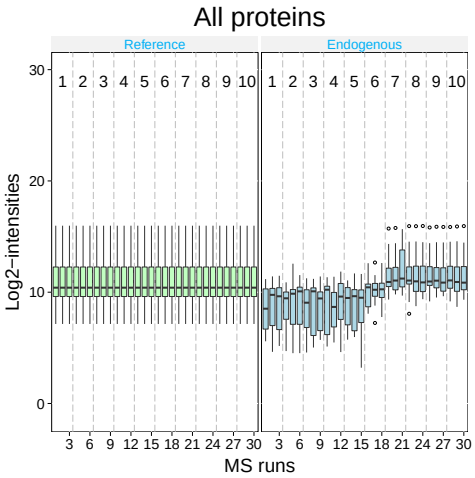


Figure 5.3.: Quality control (QC) plots for all the proteins combined. A time course study of *S. Cerevisiae* in Section 4.1.2 is used. X-axis: MS run. Y-axis: log-intensities of transitions. Reference/endogenous signals are in the left/right panel. (a) Before normalization. (b) After constant normalization. (c) After quantile normalization. The plots visualize potential artifacts in mass spectrometry runs.

The function, called `ProcessPlots`, takes as input the quantitative data from the function dataProcess, and generates three types of plots for data visualization and quality control.

Profile plot as Figure 5.2 helps identify potential sources of variation (both variation of interest and nuisance variation) for each protein. Such plots should be done after the normalization.

QC plot Figure 5.3 visualizes potential systematic biases between mass spectrometry runs. After constant normalization, the median intensities of reference transitions across all proteins should be equal between runs (Figure 5.3(b)). After quantile normalization, the distribution of reference intensities across all proteins should be equal between runs (Figure 5.3(c)).

Condition plot visualizes potential systematic differences in protein intensities between conditions with error bars for confidence interval with 0.95 significant level or standard deviation for each condition for descriptive purpose only.

Model-based whole-plot inference

The function `groupComparison` needs (1) the output of function `dataProcess`, which includes subplot summarization, and automatically recognizes the design of experiments based on the structure of the input data. Then It requires the users (2) to state the conditions that they would like to compare. The statistical model in Section 3.4 will be used to evaluate each protein for evidence of differential abundance between these conditions, while taking into account the experimental design, and the available sources of variation. It then reports log fold change estimation, standard error of the log fold change, test statistic of the Student test, degree of freedom, raw p-values and p-values adjusted for multiple testing across the entire protein set by Benjamini and Hochberg [81].

Visualization for inference result

The function `groupComparisonPlots` takes as input the results of testing in function `groupComparison` above and visualizes them as below.

Volcano plots (Figure 5.4(a)) visualize the outcome of one comparison between conditions for all the proteins, and combines the information on statistical and practical significance. The y-axis displays the FDR-adjusted p-values on the negative log10 scale, and represents statistical significance. The horizontal dashed line represents the FDR cutoff, commonly 0.05. The x-axis is the model-based estimate of log-fold change, and represents practical significance.

Heatmaps (Figure 5.4(b)) illustrate the patterns of up- and down-regulation of proteins in several comparisons. Columns in the heatmaps are comparisons of conditions, and rows are proteins. The heatmaps display signed FDR-adjusted p-values of the tests, colored in red/blue for significantly up-/down-regulated proteins, while taking into account the specified FDR cutooff and the additional optional fold change cutoff.

Comparison plots illustrate model-based estimates of log-fold changes, and the associated uncertainty, in several comparisons of conditions for one protein. X-axis is the comparison of interest. Y-axis is the log fold change. It presents the model-based estimates of log-fold change, and the error bars for the model-based 95% confidence intervals.

Sample size calculation for a future experiment

This last analysis step views the dataset as a pilot study of a future experiment, utilizes its variance components, and calculates the minimal number of replicates required in a future experiment to achieve the desired statistical power. The calculation is performed by the function `designSampleSize`, which takes as input the output of

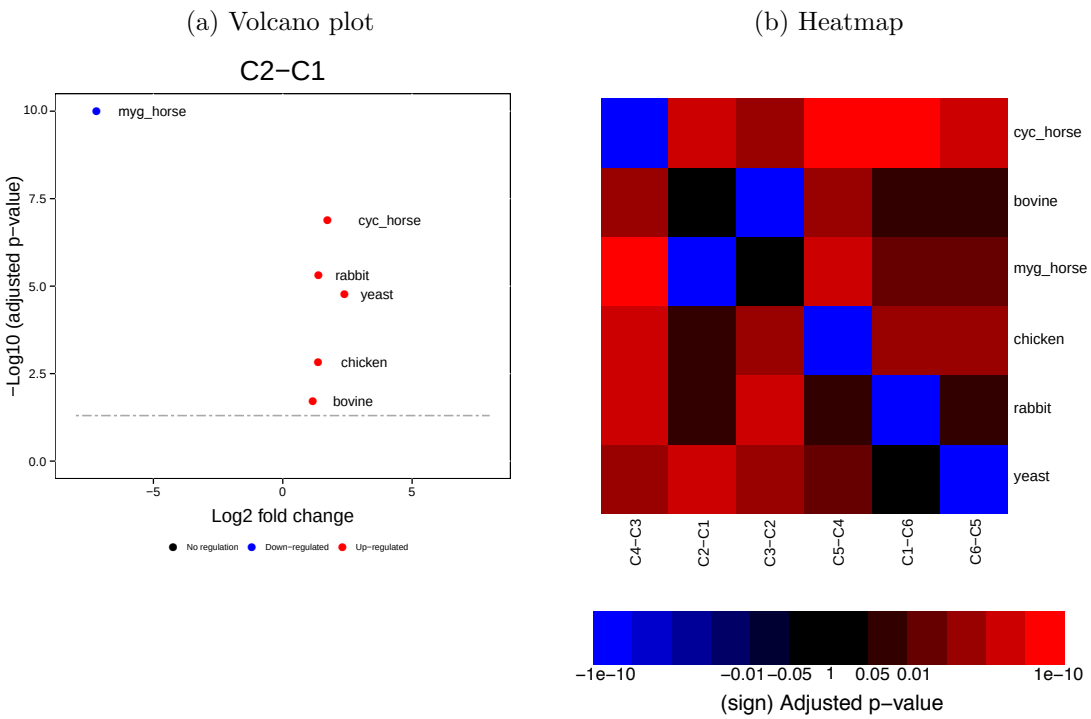


Figure 5.4.: Visualization for testing result with DDA Spike-in dataset by Latin Square design [35] (a) Volcano plot for the comparison, C2-C1. The dashed line represent the FDR cutoff=0.05. (b) Heatmap for results of testing proteins for differential abundance in six pairwise comparisons of conditions. As color key shows below heatmap, brighter colors indicate stronger evidence in favor of differential abundance. Black color represents proteins are not significantly differentially abundant.

function `groupComparison`, which contains the fitted model and variance components in whole plot level. Sample size calculation assumes same experimental design (i.e. group comparison, time course or paired design) as in the current dataset, and uses the model fit to estimate the median variance components across all the proteins. Finally, sample size calculation assumes that a large proportion of proteins (specifically, 99%) will not change in abundance in the future experiment. This assumption also provides conservative results. Using the estimated variance components, the function relates the number of biological replicates per condition, average statistical power across all the proteins (power), minimal fold change that we would like to detect, and the False Discovery Rate (FDR). Either minimum number of biological repli-

cates per condition or average statistical power is the output with specifying all other quantities.

Execution time

For the experiment with simple design of experiments including small number of proteins, features and subjects, executing time is short with few seconds. But, as the design of experiments gets complex and the number of samples and feature increases, MSstats version 3.0 or later saves more time than previous version (v2) with full linear model. Here are the examples about running time for three large datasets. I analyzed the datasets with 1.8GHz inter core i5.

DDA datasets generally have several thousand proteins, even though the number of samples is small. As one example of DDA datasets, Dynamic benchmark DDA data in Section 4.1.1, processed by MaxQuant, has 2173 proteins with 22365 unique combination of peptides and charge states. The percentage of missing values is 14%. It took 5 minutes for the proposed method in MSstats version 3, 15 minutes for MSstats version 2.3.7.

DIA datasets usually have large number of proteins as DDA. But, DIA has more number of features. Profiling standard sample DIA data in Section 4.1.1, processed by Skyline, contains 3097 proteins with 162492 unique features, 19% missing values. It took 140 minutes for the proposed method of v 3 and 550 minutes for MSstats v 2.3.7.

SRM datasets have relatively small number of proteins and features. However, if the number of subjects are large, running time between statistical methods are different. The validation set of CRC biomarker study with SRM in Section 4.1.2 has large number of subjects, 269 subjects in total, even though the number of features is small, 276 unique features in 70 proteins. 12% of measurements are missing. It took 1.5 minutes for the proposed method of v 3, but 16 minutes for v 2.3.7.

5.2 Availability of MSstats

5.2.1 MSstats website : msstats.org

MSstats is available under the Artistic-2.0 license at `msstats.org`. The most recent version is available at `msstats.org` and related source codes are available at MSstats github (`https://github.com/MeenaChoi/MSstats`). The versioning of the main package is updated several times a year, to fix the reported bugs and to synchronise with the Bioconductor release. The published data sets and example R scripts are available in this website to help user as the guide.

5.2.2 Bioconductor

MSstats satisfies all requirements of Bioconductor including interoperability, maintenance and documentation and is available in Bioconductor (`http:\\www.bioconductor.org`) from version 2.0 and later (now 3.3) since October 2013. 1734 downloads from unique IPs, which is in top 20% of downloads in Bioconductor.

5.2.3 External tool in Skyline

MSstats as an external tool in Skyline [9] is designed as a link between researchers with and without statistical background. Proteomic practitioners (the primary audience of the package) have a limited familiarity with R, and in the past this has hindered a broad adoption of R-based implementations. MSstats from version 2.0 and later is available at `http://proteome.gs.washington.edu/software/Skyline/tools.html` as a popular Graphical user interface (GUI) tool for quantitative proteomics with 1100 registered users. The external tool support within Skyline manages MSstats installation, point-and-click execution, parameter collection in Windows forms and output display. Skyline manages the annotations of the experimental de-

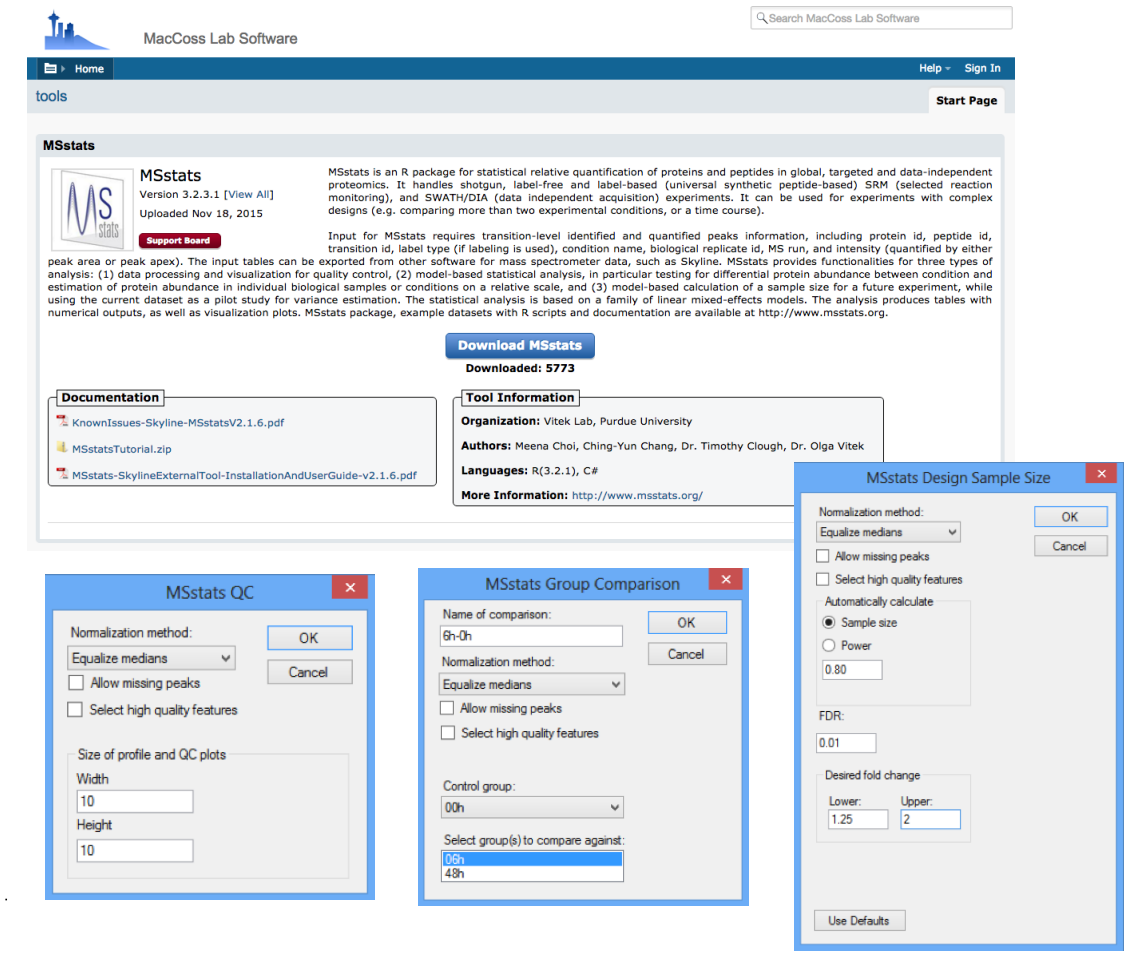


Figure 5.5.: Screen captures of MSstats external tool webpage and GUIs for three main functionalities, (1) MSstats QC : preprocessing data and run-level summarization with suggested statistical model (2) MSstats Group Comparison : whole plot inference for interested group comparison, and (3) MSstats Design Sample Size : sample size calculation.

sign and the processing of raw data. It outputs a custom report that is fed as a single stream input into MSstats. This design buffers proteomics users from the details of the R implementation, while enabling rigorous statistical modeling. MSstats also benefits from inclusion in Skyline community resources such as message boards, support in tutorials and examples of publicly available datasets.

6. APPLICATION OF MSSTATS

Several joint projects with my collaborators confirm the statistical framework proposed in the chapters above for broad biological and clinical investigation with mass spectrometry-based proteomics experiments.

We addresses the statistical framework for biomarker study in MS proteomics experiments, including (1) design of experiments including separation of training set and independent validation datasets, normalization between datasets, (2) use of MSstats for finding significantly abundant proteins, statistical quantification for each subject as in Chapter 5, (3) selection of candidate proteins by statistical variable selection method. Selected protein markers by suggested statistical strategy are matched with candidates from many literatures and expectation. Other biological studies, such as antibody-based immunochemistry assay or signatures from genomic datasets, confirm the proposed statistical framework.

Here is the list of peer-reviewed publications.

- Cerciello *et.al. Identification of a seven glycopeptide signature for malignant pleural mesothelioma in human serum by selected reaction monitoring* [10] is the protein biomarker investigation for malignant pleural mesothelioma in human serum by SRM.

- Borràs *et.al. Protein-based classifier to predict conversion from clinically isolated syndrome to multiple sclerosis* [13] studies the protein biomarkers to predict conversion from clinically isolated syndrome to Multiple Sclerosis.

- Surinova *et.al. Prediction of colorectal cancer diagnosis based on circulating plasma proteins* [11] is the non-invasive diagnostic biomarker studies for colorectal cancer.

- Surinova *et.al. Non-invasive prognostic protein biomarker signatures associated with colorectal cancer* [12] is the non-invasive prognostic biomarker studies for colorectal cancer.

I describe the proposed statistical procedure for biomarker study as downstream analysis with one of examples above, the biomarker investigation for colorectal cancer [11] in this chapter.

6.1 Design of experiments for biomarker study

The workflow for the development of predictive biomarkers consists of three phases. Phase 1is to discover biomarker candidates between tumor and normal tissue epithelia by discovery-driven mass spectrometric (MS) profiling of the glycoproteome. Phase 2 is the screening stage in patient plasma by targeted MS via selected reaction monitoring (SRM). The first cohort was used both the discovery and the screening phase with different MS technologies. The subset of candidates among the list of candidates from Phase 1 is reproducible in terms of consistently quantified and detected proteins as significantly differential protein. Phase 3 is the biomarker development stage, which select biomarker signatures and evaluate their performance. Two large-scale independent clinical cohorts, training and validation cohorts were involved in Phase 3. The training cohort consisted of patients with colorectal cancer (CRC) and subjects representing a control population at risk, in order to find biomarker signatures and develop prediction model. The validation cohort contained the two groups, CRC group including approximately equal number of patients from 4 clinical stages and control group including clinically healthy blood donors and subjects with vari-

ous non-malignant gastrointestinal tract (GIT) conditions such as adenoma, benign condition, diverticular disease, dysplastic polyps, and Crohns disease.

6.2 Relative quantification and statistical significance analysis

Label-free quantification for discovery stage Label-free quantification was processed by OpenMS 1.7. Quantitation with peptide sequences were were log2-transformed, and a scale-normalization procedure [82] was performed. Protein significance analysis between CRC and control was performed with MSstats (v1.0) with restricted scope of conclusions for biological replication.

Label-based quantification for screening and clinical cohorts Automatic SRM peak integration was performed by MultiQuant 1.2 for the screening and validation cohorts and by Skyline for the training cohort.

Two steps of normalization was used to logarithm base 2-transformed peak areas, separately for each cohort. The first normalization used internal stable isotope labeled standard reference peptides for each targeted endogenous peptide. in order to remove systematic variations in the signal during MS runnings. The second normalization was based on internal standard bovine proteins, which we can assume the same amounts across runs. It help to explain potential artifacts during sample preparation before data acquisition.

Statistical analysis for differentially abundant proteins between CRC and control was performed by MSstats v2.3.5 with expanded scope of conclusions for technical replication and with restricted scope of conclusions for biological replication. Model-based estimation of sample-level quantification for individual protein was calculated by MSstats and used predictive analysis in next step.

6.3 Predictive analysis

Before prediction analysis, normalization between cohorts and consideration of missing quantification should be required in order to make relatively quantified intensities comparable between cohorts. The normalization between cohorts was performed by making the median normalization log2-relative quantification of the training cohorts and validation cohort same. Imputation for missing relative quantifications was also considered, because NA is not accepted for suggested predictive model and zero is biased value, which affect parameter estimation and performance. Missing summarized intensities were imputed with a minimum summarized quantification observed for that protein in each cohort, assuming that missing intensities were under limit of detection.

Statistical analysis workflow of prediction analysis for protein biomarker signatures had several steps as below. Step 1 discovered predictive signatures using multivariate logistic regression with 10-fold cross validation with 100 CRC patients and 100 control subjects in the training set. For each fold, first, protein significance analysis between CRC and controls with feature-level intensities was carried out for each protein, using the nine-tenths of the subjects with FDR < 0.05 and fold change > 1.1. Significant proteins changing in abundance in CRC are the candidates for predictive modeling. Then subject-level summarized relative abundance of these candidate proteins were used as input to logistic regression. Stepwise selection by minimizing the Akaike information criterion (AIC) reduced the list of candidates within each fold. Among the ten folds, the candidate signatures, which were selected more than five times, are ceruloplasmin (CP), leucine-rich alpha-2-glycoprotein (LRG1), serpin peptidase inhibitor, clade A (SERPINA3), serum paraoxonase/ arylesterase 1 (PON1), and tissue inhibitor of metalloproteinases 1 (TIMP1) and they are final biomarkers for CRC predictive model. The parameters of the multivariate logistic regression model were estimated by these five final proteins in whole training set.

To assess the reproducibility of the selected biomarkers, the prediction analysis was repeated an additional three times on differently partitioned subjects with 10-fold cross-validation and also employed 8-fold cross validation. Overall, the results confirmed that the selected biomarkers by the original analysis are robust to the specific choice of the parameters and of the folds.

Step 2 characterized the selected biomarkers on the full training cohort by providing the original scale of fold changes and standard errors, which log2 transformed scale were estimated using a linear mixed effect model in MSstats and transformed to the original scale using Delta method [83] in Eq. (6.1).

$$\begin{aligned} \widehat{FC}_{original} &= 2^{\widehat{FC}_{log2scale}} \\ \widehat{SE}_{original} &= \widehat{SE}_{log2scale} \times ln(2) \times 2^{\widehat{FC}_{log2scale}} \end{aligned} \tag{6.1}$$

, where $\widehat{FC}$ is the estimated fold change, and $\widehat{SE}$ is the estimated standard error.

Two visualizations were shown for characterization of selected biomarkers. The plot with proportion of subjects with CRC for subgroups, which were partitioned by relative protein abundance separately by each protein, showed that an increase in protein abundance for four out of the five proteins was associated with CRC and one out of the five proteins had the opposite trend. Another plot illustrated model-based probabilities of CRC as a function of estimated log2-abundance of CP, while fixing the estimated abundances of the other proteins to their quantiles.

In step 3, the performance of the final predictive model was evaluated in the independently acquired validation dataset. The threshold was determined based on the best accuracy in the training set for detection of the disease and the control. The performance for detection of the disease and the control is assessed by AUC of ROC, specificity, sensitivity, and accuracy with the determined threshold. In addition, the predictive ability of the protein biomarker signatures to distinguish CRC with some characteristics from controls, such as clinical stages or tumor size was tested.

Furthermore, the selected diagnostic signatures have been previously linked to colorectal cancer and are assessed independently in other many literatures. Also their ability to predict colorectal cancer are confirmed with other diagnostic test, ELISA. Moreover, it compares the performance with selected proteins list with other diagnostic candidates, as CEA which is popularly used and shows the better performance of the suggested statistical analysis in terms of accuracy.

In summary, we proposed the experimental and statistical framework for biomarker investigation in MS proteomics from the discovery, screening to validation and provided other biological evidence to support the suggested statistical framework.

7. SUMMARY AND FUTURE WORK

This dissertation shows that a two-step split-plot approach improves current statistical methods for protein significance analysis across different acquisition methods for relative quantification of MS proteomics. In addition, the proposed statistical framework is fully implemented in the open source R package, MSstats with ready-made visualization. Through multiple channels with easy-to-use GUIs, such as the external tool in spectral processing tools, the number of MSstats users has been increasing. Beyond protein significance analysis, I present the biomarker investigation framework as the downstream analysis with several cancer biomarker studies and also show its validity.

The proposed approach leaves room for a number of directions for future research as shown below.

7.1 Publication of ExperimentData package in Bioconductor

I will make all datasets that were used in this dissertation publicly available as an ExperimentData package in Bioconductor. It can help other researchers to evaluate statistical methods.

7.2 Decision on the threshold for censored missing values

The datasets with more noise features tend to have worse performance in the imputation method. The suspected reason is that noisy measurements affect the decision on threshold for censored missing value at the imputation stage. A better

strategy for the decision on the threshold for censored missing values is needed for the data including many outliers.

7.3 Adjustment of degree of freedom for model-based inference

After run-level summarization in subplot, the degrees of freedom for inference per protein are the same regardless of the number of features or missing values, except the proteins which have any MS run with complete missing. A better adjustment to degree of freedom for different number of features or missing measurement is needed to be considered.

7.4 Investigation into quality control of spectral processing tools

This dissertation shows relative performance separately in each signal processing tool in order not to compare spectral processing tools, but to focuses on comparing the summarization methods given a signal processing tool. As we discussed above, the performance varies between the datasets and also spectral processing tools. I suspect that depends on how to handle missing values and quality of data by the spectral processing tools. I will look into various decisions made by each tool, including a filtering strategy for quality control, and how these decisions affect the downstream statistics and suggest the best decision.

7.5 Potential extension for PTM

Relative quantification of post-translation modification (PTM) is one of the biggest questions in mass spectrometry-based proteomics. PTM provides a more precise mechanism and key functional roles for cellular function. However, there are some challenges for PTM analysis, such as many missing values due to a very low abundance for PTM quantification, and the need for statistical methods that can improve sen-

sitivity and specificity for PTM data. I will extend the methods implemented in the proposed statistical modeling and analysis framework, and implemented in MSstats, to a peptide-level statistical analysis workflow for PTM data. The method will analyze post-translated peptides and non-post-translated peptides separately, integrate the results into a single conclusion that distinguishes differential post-translational modifications from differential protein abundance, and adjust p-values for multiple testing that considers the correlation between peptides from the same protein.

REFERENCES

REFERENCES

[1] S. Pan, R. Aebersold, R. Chen, J. Rush, D. R. Goodlett, M. W. McIntosh, J. Zhang, and T. A. Brentnall. Mass spectrometry based targeted protein quantification: methods and applications. *J. Proteome Res.*, 8(2):787–797, February 2009.

[2] T. C. Walther and M. Mann. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.*, 190(4):491–500, August 2010.

[3] B. Domon and R. Aebersold. Options and considerations when selecting a quantitative proteomics strategy. *Nat. Biotechnol.*, 28(7):710–721, July 2010.

[4] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by Data-independent Acquisition: A new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics*, 11(6):O111.016717–O111.016717, June 2012.

[5] M. Choi, C. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, 30:2524–2536, 2014.

[6] T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, and O. Vitek. Protein quantification in label-free LC-MS experiments. *J. Proteome Res.*, 8(11):5275–5284, 2009.

[7] T. Clough, S. Thaminy, S. Ragg, and O. Aebersold, R.and Vitek. Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics*, 13(Suppl 16):S6, November 2012.

[8] C.-Y. Chang, P. Picotti, R. Huttenhain, V. Heinzelmann-Schwarz, M. Jovanovic, R. Aebersold, and O. Vitek. Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics*, 11(4):M111.014662, April 2012.

[9] D. Broudy, T. Killeen, M. Choi, N. Shulman, D. Mani, S. Abbatiello, R. Ahmad, A. Sahu, B. Schilling, K. Tamura, Y. Boss, V. Sharma, B. Gibson, S. Carr, O. Vitek, M. MacCoss, and B. MacLean. A framework for installable external tools in Skyline. *Bioinformatics*, 30:2521–2523, 2014.

[10] F. Cerciello, M. Choi, A. Nicastri, D. Bausch-Fluck, A. Ziegler, O. Vitek, E. Felley-Bosco, R. Stahel, R.and Aebersold, and B. Wollscheid. Identification of a seven glycopeptide signature for malignant pleural mesothelioma in human serum by selected reaction monitoring. *Clin. Proteomics*, 10:16, November 2013.

[11] S. Surinova, M. Choi, S. Tao, P. J. Schüffler, C.-Y. Chang, T. Clough, K. Vysloŭzil, M. Khoylou, J. Srovnal, Y. Liu, M. Matondo, R. Hüttenhain, H. Weisser, J. M. Buhmann, M. Hajdúch, H. Brenner, O. Vitek, and R. Aebersold. Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol. Med.*, 7(9):1166–1178, September 2015.

[12] S. Surinova, L. Radova, M. Choi, J. Srovnal, H. Brenner, O. Vitek, M. Hajduch, and R. Aebersold. Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. *EMBO Mol. Med.*, 7(9):1153–1165, September 2015.

[13] E. Borràs, E. Cant, M. Choi, L.M. Villar, J.C. Álvarez Cermeño, C. Chiva, X. Montalban, O. Vitek, M. Comabella, and E. Sabid. Protein-based classifier to predict conversion from clinically Isolated syndrome to multiple sclerosis. *Mol. Cell. Proteomics*, page doi:10.1074/mcp.M115.053256, 2015.

[14] F. Meissner and M. Mann. Quantitative shotgun proteomics:considerations for a high-quality workflow in immunology. *Nat. Immunol.*, 15(2):112–117, February 2014.

[15] W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics*, 4:1487–1502, September 2005.

[16] W. Zhu, J. W. Smith, and C.-M. Huang. Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. and Biotechnol.*, 2010(840518):1–6, 2010.

[17] H. Choi, D. Fermin, and A. I. Nesvizhskii. Significance Analysis of Spectral Count Data in Label-free Shotgun Proteomics. *Mol. Cell. Proteomics*, pages 2373–2385, November 2008.

[18] D. S. Kirkpatrick, S. A. Gerber, and S. P. Gygi. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods*, 35(3):265–273, March 2005.

[19] C. Ludwig, M. Claassen, A. Schmidt, and R. Aebersold. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry. *Mol. Cell. Proteomics*, 11(3):M111.013987–M111.013987, March 2012.

[20] O. Vitek. Getting started in computational mass spectrometry-based proteomics. *PLoS Comput. Biol.*, 5(5):e1000366, May 2009.

[21] J. Stahl-Zeng, V. Lange, R. Ossola, K. Eckhardt, W. Krek, R. Tebersold, and B. Domon. High sensitivity detection of plasma proteins by multiple reaction monitoring of N-Glycosiyes. *Mol. Cell. Proteomics*, 6:1809–1817, September 2007.

[22] A. Doerr. DIA mass spectrometry. *Nat. Methods*, 12(1):35–35, January 2015.

[23] J. D Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates III. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods*, 1(1):39–45, September 2004.

[24] P. L. Ross, Y. H. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics*, 3:1154–1169, September 2004.

[25] S. Gygi, B. Rist, S. A. Gerber, Fr. Turecek, M. H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature*, 17:994–999, September 1999.

[26] Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.*, 7(12):952–958, November 2006.

[27] Shao-En Ong, Blagoy Blagoev, Irina Kratchmarova, Dan Bach Kristensen, Hanno Steen, Akhilesh Pandey, and Matthias Mann. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics*, 1:376–386, June 2002.

[28] L. Käll and O. Vitek. Computational mass spectrometry–based proteomics. *PLoS Comput. Biol.*, 7:e1002277, December 2011.

[29] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26–27(7):966–968, March 2010.

[30] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, November 2008.

[31] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann. Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, 10(4):1794–1805, April 2011.

[32] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics*, 13(9):2513–2526, September 2014.

[33] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[34] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. PäC̈abo, and M. Mann. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, 7:1–8, November 2011.

[35] L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.-Y. Brusniak, O. Vitek, R. Aebersold, and M. Müller. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, 7(19):3470–3480, October 2007.

[36] J. R. Yates III, J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, 67:1426–1436, June 1995.

[37] J. K. Eng, A. L. McCormack, and J. R. Yates III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–989, May 1994.

[38] D. N. Perkins, D. J. Pappin, D. M. Creasy, , and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551–3567, 1999.

[39] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher. OpenMS - An open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9(163):1–11, 2008.

[40] H. L. Röst, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinović, O. T. Schubert, W. Wolski, Ben C. C., J. Malmström, L. Malmström, and R. Aebersold. OpenSWATH enables automated, targeted analysis of data- independent acquisition MS data. *Nat. Biotechnol.*, 32(3):219–223, March 2014.

[41] L. Reiter, O. Rinner, P. Picotti, R. Hüttenhain, M. Beck, M.-Y. Brusniak, M. O. Hengartner, and R. Aebersold. mProphet: automated data processing and statistical validation for large-scale SRM experiments. *Nat. Methods*, 8(5):430–435, March 2011.

[42] J. Teleman, H. L. Röse, G. Roseberger, U. Schmitt, L. Malmström, J. Malmström, and F. Levander. DIANA-algorithmic improvements for analysis of data-independent acquisition MS data. *Bioinformatics*, 31:555–562, October 2014.

[43] U. Distler, J. Kuharev, P. Navarro, Y. Levin, H. Schild, and S. Tenzer. Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods*, 11(2):167–170, December 2013.

[44] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods*, 12(3):258–264, January 2015.

[45] A. V. Nefedov, M. J. Gilski, and R. G. Sadygov. Bioinformatics tools for mass spectrometry-based high-throughput quantitative proteomics platforms. *Curr. Proteomics*, 8(2):125137, 2011.

[46] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3(1):1–25, 2004.

[47] L Ting, M. J. Cowley, S. L. Hoon, M. Guilhaus, M. J. Raftery, and R. Cavicchioli. Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Mol. Cell. Proteomics*, 8:2227–2242, September 2009.

[48] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.*, 98(9):5116–5121, April 2001.

[49] B. AP Roxas and Q. Li. Significance analysis of microarray for relative quantitation of LC/MS data in proteomics. *BMC Bioinformatics*, 9(1):187, 2008.

[50] P. Breitling, R.and Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 573(1):83–92, August 2004.

[51] Y. Wang, T.-H. Ahn, Z. Li, and C. Pan. Sipros/ProRata: a versatile informatics system for quantitative community proteomics . *Bioinformatics*, 29:2064–2065, August 2013.

[52] L. L. Elo, S. Filen, R. Lahesmaa, and T. Aittokallio. Reproducibility-optimized test statistic for ranking genes in microarray studies. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 5(3):423–431, 2008.

[53] Y. V. Bukhman, M. Dharsee, R. Ewing, P. Chu, T. Topaloglou, T. Le Bihan, T. goh, H. duewel, I. I. Stewart, J. R. wisniewski, and ng N. F. Design and analysis of quantitative differential proteomics investigations using LC-MS technology. *J. Bioinform. Comput. Biol.*, 6(1):107–123, March 2008.

[54] A. D. Polpitiya, W. J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. G. Camp, G. A. Anderson, and Smith R. D. DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics*, 24(13):1556–1558, June 2008.

[55] L. L. Elo, J. Hiissa, J. Tuimala, A. Kallio, E. Korpelainen, and T. Aittokallio. Optimized detection of differential expression in global profiling experiments: case studies in clinical transcriptomic and quantitative proteomic datasets. *Brief. Bioinform.*, 10(5):547–555, August 2009.

[56] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W.-J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, 25(16):2028–2034, August 2009.

[57] C. D. Tekwe, R. J. Carroll, and A. R. Dabney. Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics*, 28(15):1998–2003, July 2012.

[58] B.-J. M. Webb-Robertson, L. A. McCue, K. M. Waters, M. M. Matzke, J. M. Jacobs, T. O. Metz, S. M. Varnum, and J. G. Pounds. Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data. *J. Proteome Res.*, 9(11):5748–5756, November 2010.

[59] X. Wang, G. A. Anderson, R. D. Smith, and A. R. Dabney. A hybrid approach to protein differential expression in mass spectrometry-based proteomics. *Bioinformatics*, 28(12):1586–1591, June 2012.

[60] J.-P. Lambert, G. Ivosev, A. L. Couzens, B. Larsen, M. Taipale, Z.-Y. Lin, Q. Zhong, S. Lindquist, M. Vidal, R. Aebersold, T. Pawson, R. Bonner, S. Tate, and A.-C. Gingras. Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. *Nat. Methods*, 10(12):1239–1245, October 2013.

[61] H. Choi, S. Kim, D. Fermin, C.-C. Tsou, and A. I. Nesvizhskii. QPROT: Statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. *J. Proteomics*, 129(C):121–126, November 2015.

[62] F. Koopmans, L. N. Cornelisse, T. Heskes, and T. M. H. Dijkstra. Empirical bayesian random censoring threshold model improves detection of differentially abundant proteins. *J. Proteome Res.*, 13(9):3871–3880, September 2014.

[63] G. Teo, S. Kim, C.-C. Tsou, Ben C., A.-C. Gingras, A. I. Nesvizhskii, and H. Choi. mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry. *J. Proteomics*, 129(C):108–120, November 2015.

[64] L. J. E. Goeminne, A. Argentini, L. Martens, and L. Clement. Summarization vs. peptide-based models in label-free quantitative proteomics: performance, pitfalls and data analysis guidelines. *J. Proteome Res.*, 14(6):2457–2465, June 2015.

[65] R. E. Higgs, J. P. Butler, B. Han, and M. D. Knierman. Quantitative proteomics via high resolution MS quantification: capabilities and limitations. *Int. J. Proteomics*, 2013(5):674282, 2013.

[66] A. Pursiheimo, A. P. Vehmas, S. Afzal, T. Suomi, T. Chand, L. Strauss, M. Poutanen, A. Rokka, G. L. Corthals, and L. L. Elo. Optimization of statistical methods impact on quantitative proteomics data. *J. Proteome Res.*, 14(10):4118–4126, October 2015.

[67] B.-J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, and K. M. Waters. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.*, 14(5):1993–2001, May 2015.

[68] D. C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons, Inc., New Jersey, US, 8th edition, 2013.

[69] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

[70] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, March 2003.

[71] The Association of Bimolecular Resource Facilities. *IPRG-2015 Study - Differential abundance analysis in label-free quantitative proteomics*, 2015. Available at `https://www.abrf.org/research-group/proteome-informatics-research-group-iprg`, accessed November 28, 2015.

[72] T. A. Addona, S. E. Abbatiello, B. Schilling, S. J. Skates, D. R. Mani, D. M. Bunk, C. H. Spiegelman, L. J. Zimmerman, A.-J. L. Ham, H. Keshishian, S. C. Hall, S. Allen, R. K. Blackman, C. H. Borchers, C. Buck, H. L. Cardasis, M. P. Cusack, N. G. Dodder, B. W. Gibson, J. M. Held, T. Hiltke, A. Jackson, E. B. Johansen, C. R. Kinsinger, J. Li, M. Mesri, T. A. Neubert, R. K. Niles, T. C. Pulsipher, D. Ransohoff, H. Rodriguez, P. A. Rudnick, D. Smith, D. L. Tabb, T. J. Tegeler, A. M. Variyath, L. J. Vega-Montoto, Å. Wahlander, S. Waldemarson, M. Wang, J. R. Whiteaker, L. Zhao, N. L. Anderson, S. J. Fisher, D. C. Liebler, A. G. Paulovich, F. E. Regnier, and S. A. Tempst, P.and Carr. Multi-site assessment of the precision and reproducibility of multiple reaction monitoring–based measurements of proteins in plasma. *Nat. Biotechnol.*, 27(7):633–641, June 2009.

[73] M. J. Rardin, B. Schilling, L.-Y. Cheng, B. MacLean, D. J. Sorensen, A. K. Sahu, M. J. MacCoss, O. Vitek, and B. W. Gibson. MS1 peptide ion intensity chromatograms in MS2 (SWATH) data independent acquisitions. Improving post acquisition analysis of proteomic experiments. *Mol. Cell. Proteomics*, 14(9):2405–2419, August 2015.

[74] R. Bruderer, O. M. Bernhardt, T. Gandhi, S. M Miladinović, L.-Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, and L. Reiter. Extending the limits of quantitative proteome profiling with Data-Independent Acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol. Cell. Proteomics*, 14(5):1400–1410, May 2015.

[75] R. E. Higgs, M. D. Knierman, V. Gelfanova, J. P. Butler, and J. E. Hale. Label-free LC-MS method for the identification of biomarkers. *Methods Mol. Biol.*, 428:209–230, August 2008.

[76] R. Hüttenhain, M. Soste, N. Selevsek, H. Röst, A. Sethi, C. Carapito, T. Farrah, EW. Deutsch, U. Kusebauch, E. Moritz, R. L. Nimeus-Malmström, O. Rinner, and R. Aebersold. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. *Sci. Tansl. Med.*, 4(142):142ra94, July 2012.

[77] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*, 138(4):795–806, August 2009.

[78] N. Selevsek, C.-Y. Chang, L. C. Gillet, P. Navarro, O. M. Bernhardt, L. Reiter, L.-Y. Cheng, O. Vitek, and R. Aebersold. Reproducible and consistent quantification of the *Saccharomyces cerevisiae* proteome by SWATH-MS. *Mol. Cell. Proteomics*, 14(3):739–749, March 2015.

[79] S. Surinova, H. Ruth, C.-Y. Chang, L. Espona, O. Vitek, and R. Aebersold. Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. *Nat. Protoc.*, 8(8):1602–1619, July 2013.

[80] D. Amaratunga and J. Cabrera. Analysis of data from viral DNA microchips. *J. Amer. Statist. Assoc.*, 96(456):1161–1170, December 2001.

[81] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Statist. Soc. B*, 57(1):289–300, 1955.

[82] G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, December 2003.

[83] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., New Jersey, US, 3rd edition, 2012.

VITA

VITA

Meena Choi was born in Wonju, South Korea. She received a B.S. in Biology from Korea Advanced Institute of Science and Technology, South Korea in 2002 and a B.S. in Management from Purdue University in 2008. She received a M.S in Applied Statistics from Purdue University in 2011.