STATISTICAL DESIGN AND ANALYSIS OF MASS SPECTROMETRY-BASED PROTEOMIC EXPERIMENTS WITH ISOBARIC LABELING

A Thesis Proposal

Submitted to the Faculty

of

Northeastern University

by

Ting Huang

In Partial Fulfillment of the

Requirements for the Comprehensive Examination

October 2020

Northeastern University

Boston, Massachusetts

TABLE OF CONTENTS

Page

ABSTRACT

Huang, Ting PhD, Northeastern University, October 2020. Statistical Design and Analysis of Mass Spectrometry-Based Proteomic Experiments with isobaric labeling. Major Professor: Olga Vitek.

Tandem mass tag (TMT) is a multiplexing technology widely-used in proteomic research. It enables relative quantification of proteins from multiple biological samples in a single mass spectrometry run with high efficiency and high throughput. However, experiments often require more biological replicates or conditions than can be accommodated by a single run, and involve multiple TMT mixtures and multiple runs. Such larger-scale experiments combine sources of biological and technical variation in patterns that are complex, unique to TMT-based workflows, and challenging for the downstream statistical analysis. These patterns cannot be adequately characterized by statistical methods designed for other technologies, such as label-free proteomics or transcriptomics. Therefore, there is a need for flexible statistical tools, which reflect diverse and complex designs of large-scale TMT experiments, and have good statistical performance.

We first develop a general statistical approach for relative protein quantification in mass spectrometry-based experiments with TMT labeling and group comparison design. It is applicable to experiments with multiple conditions, multiple biological replicate runs, multiple technical replicate runs, and unbalanced designs. It is based on a flexible family of linear mixed-effects models that handle complex patterns of technical artifacts and missing values. The approach is implemented in *MSstatsTMT*, a freely available open-source R/Bioconductor package compatible with data processing tools such as Proteome Discoverer, MaxQuant, OpenMS and SpectroMine. We propose to extend the existing approach for group comparison design to more com-

plex designs. In particular, we are interested in the repeated measures designs where protein abundance in a same subject is repeatedly measured for each condition in the TMT experiment.

1. PROBLEM STATEMENT AND CONTRIBUTIONS

1.1 Statement of the problem

1.1.1 Statement of the biotechnological problem

Mass spectrometry(MS)-based proteomics has emerged as powerful tool to study molecular and cellular biology and systems biology [1]. It can deal with large-scale characterization of protein composition and abundance in complex biological mixtures. In recent years, isobaric labeling of biological samples, combined with shotgun mass spectrometry (MS), is becoming a popular strategy for MS-based quantitative proteomics [2–4]. Two most commonly used isobaric labeling methods are Tandem Mass Tags (TMT [5]) and isobaric Tags for Relative and Absolute Quantitation (iTRAQ [6]). In these experiments, peptides from different samples are labeled with isobaric variants of a mass tag (also called channels) and combined to produce a single biological mixture. This multiplexed approach to quantification increases the sample throughput and decreases the experiment time. Importantly, it also reduces the between-run technical variation that is unavoidable during label-free sequential sample processing and data acquisition [7, 8]. The quantitative accuracy can be further improved by acquiring technical replicate MS runs from one mixture. However, in many investigations the number of biological samples and conditions exceeds the number of channels. These investigations require multiple MS runs with distinct biological replicates, and possibly also multiple technical MS runs. Investigations with such non-trivial designs may be unbalanced (i.e., they may have an unequal number of replicates from each condition in each mixture).

In addition to the complexities of the designs, larger-scale experiments introduce challenges in the downstream statistical analysis. The stochastic selection of peptide ions for fragmentation implies that the same ions are not consistently observed even between technical replicate MS runs. The problem is exacerbated when the experiment profiles multiple biological mixtures. Therefore, the resulting data have many missing intensities between MS runs [9, 10]. Moreover, multiple spectra of the same peptide ion, and multiple peptides of the same protein, may have varying quantitative patterns within and between MS runs, and require normalization [10]. The intensities can be further compromised by ion interference, due to co-isolating and co-fragmenting isobaric ions within the isolation window. This causes underestimation of changes in protein abundance [11, 12]. In some channels, the intensities can be missing if the analyte is below the limit of detection. Additional technical variation may come from sample preparation (e.g., protein extraction, digestion and isobaric labeling), run-to-run instrumental response fluctuation, etc [13]. A combination of complex experimental designs, missing values, biological and technical variation, and interferences complicates protein-level conclusions, and in particular the detection of differentially abundant proteins between conditions. Therefore, the MS proteomics community needs a flexible and versatile statistical analysis tool to address the challenges in mass spectrometry-based experiments with TMT labeling, which has a good performance.

1.1.2 Statement of the statistical and computational problem

The TMT experiments is becoming larger in scale and often quantify more biological replicates and technical replicates through multiple TMT mixtures and MS runs. Different experimental designs have different patterns of biological and technical variation, and different patterns of missing values. There is currently no integrated statistical tool, that is applicable to these diverse designs. Therefore, I propose to develop a statistical tool that can (1) perform protein quantification and normal-

ization in mass spectrometry-based experiments with TMT labeling, specifically for experiments with multiple conditions, multiple biological replicate runs and multiple technical replicate runs, (2) perform statistical modeling and inference 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, (3) enable the inter-operability of the method implementation with other popular data processing tools, (4) estimate the minimal sample size required to achieve a predefined statistical power, and (5) evaluate the method and the implementation in a variety of the datasets.

1.2 Statement of contributions

1.2.1 Preliminary work

For MS-based proteomic experiments with isobaric labeling, I have developed and implemented the following statistical workflow (described in Chapter 2):

- Remove the technical artifacts between channels with global median normalization.
- For each MS run, use robust estimation with Tukey's median polish to extricate existence of outliers and missing measurements.
- Remove the technical artifacts between MS runs by reference channel-based local protein-level normalization.
- Develop a modeling framework that is applicable to experiments with arbitrary group comparison designs. The design can include multiple conditions, multiple biological replicate runs, and multiple technical replicate runs. In addition, the design can be balanced or unbalanced. Apply Empirical Bayes moderation to deal with small sample size.

- Implement the workflow in the open-source R/Bioconductor packages MSstatsTMT [14]
- Publish the workflow in *Molecular & Cellular Proteomics* [15]

1.2.2 Future work

Based on the preliminary work, I propose to solve the following problems by the completion of my dissertation (described in Chapter 3):

- Extend the modeling framework developed by preliminary work to arbitrary complex designs, such as repeated measures design that includes paired design and time course design.
- Estimate the minimal sample size required in TMT experiments to achieve a predefined statistical power, which is used to design future experiments.
- Add the proposed models to the open-source R/Bioconductor packages MSstatsTMT
- Evaluate the performance of the proposed models in a variety of the datasets with diverse and complex design

2. MSSTATSTMT WITH GROUP COMPARISON DESIGN

I developed a general statistical approach for relative protein quantification in mass spectrometry-based experiments with TMT labeling. The approach is specifically designed for experiments with group comparison design, which can include multiple conditions, multiple biological replicate runs and multiple technical replicate runs. It is based on a flexible family of linear mixed-effects models that handle complex patterns of technical artifacts and missing values. The approach is implemented in *MSstatsTMT*, a freely available open-source R/Bioconductor package, and published in *Molecular & Cellular Proteomics* [15]. Below I present the details of the approach, as well as its evaluation on a controlled mixture and simulated datasets. More information about the approach and its evaluation results can be found in the manuscript.

2.1 Existing statistical methods for experiments with isobaric labeling

This section reviews the existing statistical analysis strategies for experiments with isobaric labeling, and partitions them into a series of common steps summarized in the rows of Figure 2.1. Specifically, *spectrum-level normalization* reduces artifacts of sample preparation or mass analysis at the level of reporter ion intensities. *Protein summarization* takes as input all reporter ion intensities (or their ratios) of a protein in a run, and aggregates them into a single estimate of protein abundance per channel per run. *Protein-level normalization* reduces the technological artifacts in the protein summaries. *Statistical modeling and inference* quantifies the sources of systematic and random variation for each protein, and tests proteins for differential abundance. Various workflows approach these steps in various ways. The steps can be applied locally (i.e., separately within a spectrum or a protein), or globally (i.e., simultane-

ously to all spectra or proteins in a run), and may or may not rely on a reference channel with constant protein abundance across the runs. Not every workflow uses every step.

Columns in Figure 2.1 summarize two representative workflows that we call *Ratio+Median+Limma* [16, 17] and *Sum+IRS+edgeR* [18]. We selected these workflows because they represent two main commonly used approaches (ratio-based and sum-based), have an open-source implementation, and are compatible with multiple data processing tools. Additionally, the table summarizes the statistical analysis workflow of *Proteome Discoverer* 2.2, based on its user guide book and the method by McAlister *et al.* [19]. Whenever possible, we expand the discussion to other approaches that focus on each particular step.

Spectrum-level normalization We loosely classify spectrum-level normalizations into two groups. The first group uses *local ratio-based normalization*, as exemplified by *Ratio+Median+Limma*. If a reference channel is available, for each spectrum the method subtracts from the log_2 intensities of the endogenous channels the log_2 intensity of the reference channel. In absence of a reference channel, the method assumes a constant protein abundance across the mixtures, and replaces the reference channel with the median of log_2 intensity in the spectrum [17]. The resulting log_2 ratios are centered around 0 (and the ratios on the original scale are centered around 1).

The second group of methods use *global spectrum-level normalization*. These methods do not calculate ratios, but assume constant total protein abundance across all the spectra, channels and runs. For example, variance stabilizing normalization (VSN) transforms the reporter ion intensities to roughly equalize their variance over the entire intensity range. Originally designed for transcriptomics, the method was specifically adapted to proteomic experiments with isobaric labeling [20–22]. *CONSTANd* [23] was explicitly designed for multiplexed proteomic experiments as an instance of constrained optimization. For each run, the method constructs a data

Method	**Ratio+Median+Limma** Adapted from Herbrich *et al.* [22] and Kammers *et al.* [23]	**Sum+IRS+edgeR** Adapted from Plubell *et al.* [21]	**Proteome Discoverer 2.2** User guide book [51] Adapted from McAlister *et al.* [7]	**MSstatsTMT**
Spectrum-level normalization	**Local ratio-based normalization**: log_2 transform the intensities; for each spectrum, calculate log_2 **ratio** – without reference channel: subtract the median of log_2 intensities of all channels in the spectrum – with reference channel: subtract log_2 intensity of the reference channel	None	None	**Global median normalization**: log_2 transform the intensities; equalize the median of the log_2 intensities across all spectra, channels and MS runs
Protein summarization	**Median summarization**: for each protein and each channel, estimate protein ratio as the **median** of all the log_2 ratios of all the spectra of the protein	**Sum summarization**: for each protein and each channel, estimate protein summary as the **sum** of all the spectrum intensities on the original (not log) scale	**Sum summarization**: for each protein and each channel, estimate protein summary as the **sum** of all the spectrum intensities on the original (not log) scale	**Tukey median polish**: for each protein and each run, impute missing values with Accelerated Failure Time model and estimate protein summary in each channel with **Tukey's median polish**
Protein-level normalization	**Global zero median normalization**: for each run and each channel, subtract the median of all the log_2 ratios across proteins, such that the median ratio of each channel is zero	– Remove proteins with missing summaries – **Global equal sum normalization**: sum the summaries of all proteins in each channel, and equalize the sums over all channels and runs – **Local IRS normalization with reference channel**: scale the normalized summaries in the reference channel in each run to equalize their geometric means across runs	For each run: – **Global equal sum normalization**: sum the summaries of all proteins in each channel, and equalize the sums over all channels and runs – **Local protein scaling**: scale the normalized summaries of each protein to have an average of 100.	**Local normalization with reference channel**: for each protein, equalize the log_2 protein summaries in the reference channel of each MS run to their median across all the runs
Statistical modeling and inference	linear model with **limma**. The linear model includes fixed run effect and condition effect .	Negative Binomial regression **edgeR** with library size correction; uses subsets of data with pairs of conditions	**one-way ANOVA**	**linear mixed-effects model** fit simpler model for proteins where parameters of full model are not estimable
Applicable experimental designs	– single mixture with single technical replicate MS run – multiple mixtures with single technical replicate MS run OR single mixture with multiple technical replicate MS runs	– treat every design as single mixture with single technical replicate MS run	– treat every design as single mixture with single technical replicate MS run	– single mixture with single technical replicate MS run – single mixture with multiple technical replicate MS runs – multiple mixtures with single technical replicate MS run – multiple mixtures with multiple technical replicate MS runs
Implementation	adapted from code in Kammers *et al.* [52] to handle multi-group designs	adapted from code in Wilmarth *et al.* [53]	proprietary, Proteome Discoverer 2.2	R/Bioconductor package *MSstatsTMT*

Figure 2.1.: **Representative workflows for differential analysis of mass spectrometry experiments with isobaric labeling.** All the four methods take as input the same PSM report from a data processing tool. Rows in the table classify the statistical analyses into a series of common steps. Columns in the table are representative workflows, adapted from the corresponding publications or open-source code. Local steps are applied within a spectrum or a protein. Global steps are applied to all spectra or proteins.

matrix where rows are spectra, columns are channels, and entries are intensities of the reporter ions in that run. The method estimates a normalized version of the data matrix with maximal similarity to the original matrix while satisfying two equality constraints. The first constraint ensures that the summation of each row of the normalized matrix is equal to 1. The second constraint ensures the summation of each column is equal to a value determined by the number of spectra and channels in a run.

Protein summarization Statistical methods for protein summarization can also be loosely classified into two groups. The first group assumes that all the spectra represent the protein abundance equally well. *Ratio+Median+Limma* employs *Median summarization*, which for each protein and each channel estimates protein-level log_2 as the median of the log_2 ratios of the spectra [16,17]. In contrast, *Sum+IRS+edgeR* [18] and McAlister *et al.* [19] (adapted by *Proteome Discoverer* 2.2) use *Sum summarization*, which sums the reporter ion intensities of all the spectra on the original (i.e., not log-transformed) scale. Other methods include *Tukey's median polish*, which considers all the channels in an MS run simultaneously, takes as input the log_2 reporter ion intensities of the protein across all the channels, and iteratively fits a two-way robust additive model [24].

The second group of protein summarization methods assigns different weights to different spectra of the protein, and estimates protein abundance with a weighted average of the reporter ion intensities or ratios of its spectra. Summarization methods in this group differ in how they estimate the weights. For example, the R package *isobar* [25] calculates ratios of intensities between pairs of channels in a spectrum, estimates the noise variance of the ratios, and uses the inverse of the variances as the ratios' weights. The output of the procedure is not a summary of protein abundance per channel, but a ratio of protein abundances between pairs of channels or pairs of conditions. Method *iPQF* [26] in the R package MSnbase [24] estimates the weights

of the spectra based on multiple spectral characteristics, such as peptide mass and charge.

Protein-level normalization Many recent methods apply normalization to protein summaries, and many require at least one reference channel in each MS run. *Sum+IRS+edgeR* employs two normalization procedures. The first is a *global equal sum normalization*, which sums the summaries of all proteins in each channel on the original scale, and equalizes the sums over all channels and runs. The second is a *local Internal Reference Scaling (IRS)* normalization [18], which normalizes each protein separately. For each protein, IRS normalization first calculates a geometric mean of the normalized protein summaries in the reference channel across the runs. Next, the method calculates a scale factor, i.e., a ratio of the protein summary in the reference channel of each run to the geometric mean above. Finally, the protein summary in every channel is multiplied by the scale factor of its run. An alternative approach in [16] calculates the ratio of a protein summary in a channel to the protein summary in the reference channel of the run.

The second group of methods does not use a reference channel at this stage. For example, *Ratio+Median+Limma* implements a *global zero median normalization* [16, 17]. For each run and channel, it subtracts the median of all the protein-level log_2 ratios, to set to zero the median over all the protein summaries in the channel. McAlister *et al.* [19] use a *global equal sum normalization* that equalizes the sum of the protein summaries across the channels and runs. *Proteome Discoverer* 2.2 supplements the global equal sum normalization with *protein scaling*, which scales the normalized summaries across each protein to generate the protein ratios with a total or average of 100.

Statistical modeling and testing for differential abundance Most statistical methods for detecting differentially abundant proteins are applied after protein summarization and normalization. The simplest approach, implemented in Proteome Discoverer 2.2, fits a *one-way Analysis of Variance (ANOVA)* to all the protein sum-

maries from all the runs [19]. Alternatives use statistical methods originally designed for transcriptomics, such as R/Bioconductor packages *limma* [27] and *edgeR* [28].

More complex statistical modeling is required for experiments with multiple MS runs and missing values. *Ratio+Median+Limma* [17] extends *limma* to explicitly account for multiple MS runs. The method takes as input the log_2 protein ratios produced by normalization and summarization, and fits a two-way additive linear model with a fixed group effect and a fixed MS run effect, which does not distinguish between biological and technical replicate MS runs. It then uses the Empirical Bayes procedure in *limma* to combine the estimates of random variation across all the proteins in a moderated t-statistic [29]. While the original implementation of *limma* did not allow proteins with missing values, more recent *limma* 3.44 includes proteins with missing values into analyses. D'Angelo *et al.* [30] expanded the use of *limma* by imputing missing values within an MS run, and excluding peptide ions that were completely missing in at least one MS run. In experiments with multiple MS runs, this exclusion significantly reduced the number of proteins that can be tested for differential abundance. *Sum+IRS+edgeR* [18] uses edgeR, originally designed for transcriptomic experiments. The model is primarily appropriate for experiments that generate data in form of discrete counts since it assumes a negative binomial distribution. The implementation is limited to two conditions (or subsets of the dataset with pairs of conditions), requires an additional normalization with respect to the total protein abundances in a sample (called library size), and removes proteins with any missing values.

Several statistical methods take as input reporter ion intensities before protein summarization. Paulo *et al.* [31] first summarize the spectra at the peptide level, and use the summaries as input to an additive linear model that includes a group effect and a peptide effect, but ignores a run effect. Oberg *et al.* [9] take as input MS/MS spectra, and fit a linear mixed-effects model that decomposes the variation in the reporter ion intensities into contributions from multiple sources, including multiple MS

runs. The model has many parameters, is limited to balanced designs, and requires computationally-intensive procedures such as stage-wise or iterative regression [32].

2.2 Proposed statistical methodology

2.2.1 Input to MSstatsTMT and notation

Figure 2.2 outlines a representative design of a proteomic experiment with isobaric labeling, and the input to *MSstatsTMT* for one protein. The experiment has $m = 1, \ldots, M$ biological *Mixtures*. Each mixture contains samples from distinct biological subjects, labeled with isobaric tags (e.g., TMT 10- or 11-plex). Each mixture is profiled in $t = 1, \ldots, T$ *Technical replicate* mass spectrometry (MS) runs. Therefore, the experiment has a total of $M \times T$ MS *Runs*. In practice, *MSstatsTMT* can be applied to any number and type of technical replicates. For example, technical replicates can be separately digested and randomly labeled in order to reflect the variation due to digestion and labeling. Biological replicates from different conditions can be assigned to different channels in each MS run.

This manuscript focuses on a group comparison design, i.e., a design with $c = 1, \ldots, C$ *Conditions* (such as treatments, or disease types), where each condition is represented by different subjects. Each MS run consists of $b = 1, \ldots, B$ *Biological replicates (BioRep)* from each of the C conditions. Thus, each MS run has $B \times C$ distinct biological replicates. For simplicity, below we refer to each column in Figure 2.2 as a *Channel*. In the example of Figure 2.2, the experiment has $MTCB$ channels.

In each MS run, the protein is represented by $f = 1, \ldots, F$ *Features*. The features are MS2 or MS3 spectra identified by a search engine such as Mascot or Sequest. In each run and each channel, each feature is quantified by a log_2-transformed intensity of the reporter ion (defined as the height of the reporter ion peak, or any other

measurement) by a data processing tool such as Proteome Discover, MaxQuant, or SpectroMine and denoted X_{mtcbf}. The log_2 transformation is important, because measurements on the log scale conform more closely to the Normal distribution [33] and better satisfy the statistical modeling assumptions.

Figure 2.2 represents a balanced design, i.e., a design where all the conditions in a mixture have the same number of biological replicates. In practice, the experiment design can be unbalanced and can contain a different number of biological replicates within a mixture and a condition, and a different number of technical replicates per mixture. *MSstatsTMT* applies to these situations.

The data structure in Figure 2.2 can also be unbalanced due to missing feature intensities. Occasionally, a reporter ion channel can be missing within a feature. More frequently, missing features arise when peptide ions are inconsistently identified between the MS runs and especially between the mixtures. Some of the observed intensities can be compromised by interferences and thus become outliers.

2.2.2 Protein summarization and normalization in MSstatsTMT

Each step of *MSstatsTMT* is summarized in Figure 2.1. The combined outcome of protein summarization and normalization is illustrated for one example protein in Figure 2.3.

Global median normalization between channels This step simultaneously considers all the features identified in the experiment. Similarly to *isobar* [25] and to normalizations used in label-free quantification [34], *MSstatsTMT* assumes that the total abundance of the analytes is equal across all the channels and runs. Therefore, *MSstatsTMT* applies a global equal median normalization between channels to account for differences in labeling efficiency and other technical artifacts. *MSstatsTMT* equalizes the median of the reporter ion intensities across all the channels and MS runs.

	Mixture 1																Mixture *M*													
	Technical Replicate Run 1							Technical Replicate Run *T*								Technical Replicate Run 1							Technical Replicate Run *T*							
	Condition 1			Condition *C*				Condition 1			Condition *C*					Condition 1			Condition *C*				Condition 1			Condition *C*				
	BioRep 1 (127C)	...	BioRep *B* (129N)	...	BioRep 1 (128C)	...	BioRep *B* (130N)	...	BioRep 1 (127C)	...	BioRep *B* (129N)	...	BioRep 1 (128C)	...	BioRep *B* (130N)	...	BioRep 1 (128C)	...	BioRep *B* (130N)	...	BioRep 1 (127C)	...	BioRep *B* (129N)	...	BioRep 1 (128C)	...	BioRep *B* (130N)	...	BioRep 1 (127C)	...	BioRep *B* (129N)
Feature 1	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X
Feature 2	X	...	X	...	X	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA
Feature 3	X	...	X	...	X	...	X	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA
Feature 4	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA
Feature 5	NA	...	NA	...	NA	...	NA	...	X	...	X	...	X	...	X	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA
Feature 6	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	X	...	X	...	X	...	X
Feature 7	NA	...	NA	...	NA	...	NA	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	NA	...	NA	...	NA	...	NA
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Feature *F*	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	NA	...	X	...	X	...	X	...	X	...	X	...	X	...	X	...	X

Figure 2.2.: **Representative design of a proteomic experiment with isobaric labeling, for one protein** The experiment has M mixtures, T technical replicates MS Runs per biological mixture, C conditions and S biological replicates per condition and mixture, resulting in $M \times T \times C \times B$ observations per feature. The protein has F features. Subjects in a mixture are randomly quantified with isobaric channels (e.g., 127C and 129N). In the language of experimental design, an MS run is a whole plot (in blue), each combination of conditions and biological replicates is a subplot (in orange), and a feature is a sub-subplot (in purple). The symbol X in each cell denotes the log_2 reporter ion intensity of the observed feature and NA denotes missing feature intensity. When a feature is not identified in one MS run, the values of all the corresponding cells are NA. For example, Feature 3 is only identified in Technical Replicate Run 1 of Mixture 1.

Protein summarization This step and all the subsequent steps of *MSstatsTMT* consider one protein and one MS run at a time. It focuses on the sub-subplot aspect of the experimental design, and summarizes the log_2 intensities of the features in each channel and MS run while accounting for missing and outlying feature intensities. This summarization is identical to the summarization used for label-free experiments in MSstats [35]. Specifically, it fits the observed intensities of a protein to a two-way model.

$$X_{mtcbf} = \mu_{mt} + Feature_{f(mt)} + Channel_{b(mtc)} + \epsilon_{mtcbf}, \qquad (2.1)$$
$$\sum_{f} Feature_{f(mt)} = 0, \quad \sum_{cb} Channel_{b(mtc)} = 0$$

Assuming that the missing feature intensities primarily arise from low-abundant analytes, *MSstatsTMT* extends the model above with the Accelerated Time Failure assumption [36], and imputes the missing feature intensities within each MS run. To impute an intensity of a feature *MSstatsTMT* requires at least one non-missing channel for the same feature in that run, and at least one non-missing feature from the same protein in the same channel in that run. If the entire feature was not quantified in a run, it is left missing. If the entire protein was not quantified in a channel, all the intensities from that protein in that channel are left missing. Next, to eliminate the undue influence of outliers, *MSstatsTMT* re-estimates the parameters of the additive model from the observed and the imputed values with the Tukey's median polish [37]. Finally, *MSstatsTMT* summarizes the protein abundance Y_{mtcb} in a channel and in a run containing biological replicate b of condition c profiled by technical MS run t of mixture m as

$$Y_{mtcb} = \hat{\mu}_{mt} + \widehat{Channel}_{b(mtc)} \qquad (2.2)$$

The values Y_{mtcb} are the sub-subplot level summaries in this design.

Local protein-level normalization with reference channel This second normalization takes the protein summaries in Eq. (2.2) as input. Since different features

of a protein are typically identified in different MS runs, and since the features differ in ionization efficiency and other biochemical properties, the protein summaries are not comparable between runs. To account for this, *MSstatsTMT* relies on the presence of at least one reference channel. The reference channel lacks biological variation and reflects technological artifacts (such as different labeling and ionization efficiency). For each protein, *MSstatsTMT* equalizes the protein summaries in the reference channel of each MS run to the median of the reference channels between the runs. It then applies the corresponding shifts to the protein-level summaries in the remaining channels of each run. If the design includes multiple reference channels per MS run, *MSstatsTMT* starts by averaging the protein summaries of the reference channels.

The local protein normalization by *MSstatsTMT* is similar in spirit to that of IRS normalization, but is different in that it is applied to the log-scaled protein summaries. It is equivalent to calculating log-ratios between the endogenous and the reference channels, and rescaling the log-ratios to a common median value. The local protein normalization by *MSstatsTMT* is similar to the approach by Kammers *et al.* [17] in *Ratio+Median+Limma* in balanced designs. However, the results of *MSstatsTMT* and *Ratio+Median+Limma* differ substantially when MS runs contain different number of replicates from each condition. Figure 2.3 illustrates that in unbalanced designs, normalization without a reference channel can remove the true biological signal. Normalization with respect to a reference channel avoids this artifact.

2.2.3 Statistical modeling and inference in MSstatsTMT

The normalized protein-level summaries are used as input to statistical modeling. For experimental designs with multiple biological replicates, multiple mixtures, and

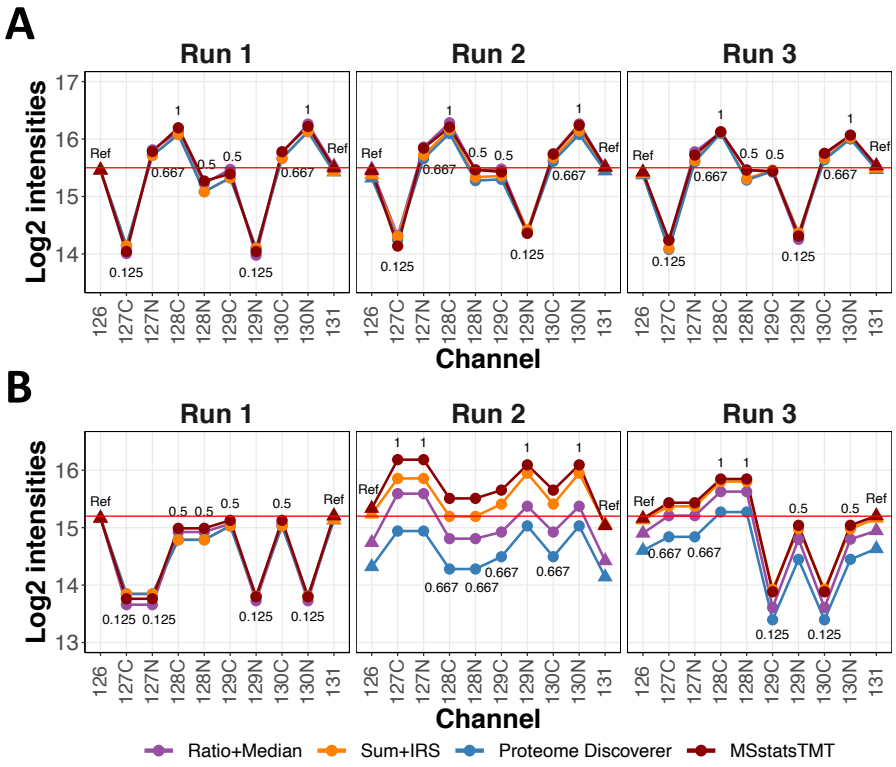

Figure 2.3.: **Spectrum-level normalization, protein summarization and protein-level normalization in representative workflows in Figure 2.1 in a hypothetical experiment with two reference channels and three runs** The individual steps are detailed in Section 2.1. Each panel is a MS run. X-axis: TMT channels. Y-axis: log_2 intensity. Colored dots are log_2 protein intensities summarized and normalized by each workflow, labeled with the true abundance in the respective channel and mixture. Triangle dots indicate reference channels. To make the scale of the protein summaries comparable between workflows, the normalized protein intensities reported by *Sum+IRS+edgeR* and *Proteome Discoverer* were log_2 transformed, and equalized to the protein summaries in Channel 126 of Run 1. Red horizontal line indicates the median of the protein summaries in the reference channels across the runs, as estimated by *MSstatsTMT*. (A) Balanced design, where each run has an equal number of replicates from each condition. All the workflows equalized the reference channels between the runs, and reported similar normalized protein summaries. (B) Unbalanced design, where each run has a different number of replicates from each condition. Normalizations by *Ratio+Median+Limma* and *Proteome Discoverer* failed to eliminate undue variation between reference channels, and compressed the differences in protein summaries across conditions and runs. Normalizations with respect to the reference channels by by *Sum+IRS+edgeR* and *MSstatsTMT* avoided this artifact.

multiple technical replicates (such as in Figure 2.4), *MSstatsTMT* fits the following model:

$$Y_{mtcb} = \mu + Mixture_m + TechRep(Mixture)_{t(m)} + Condition_c + Subject_{mcb} + \varepsilon_{mtcb} \quad (2.3)$$

$$\text{where } Mixture_m \overset{iid}{\sim} N(0, \sigma_M^2),\ TechRep(Mixture)_{t(m)} \overset{iid}{\sim} N(0, \sigma_T^2),\ \sum_{c=1}^{C} Condition_c = 0,$$

$$Subject_{mcb} \overset{iid}{\sim} N(0, \sigma_S^2),\ \varepsilon_{mtcb} \overset{iid}{\sim} N(0, \sigma^2)$$

The term *Subject* represents biological replicates, with the convention that each biological replicate has a unique identifier across mixtures and conditions. $Mixture$ and $TechRep(Mixture)$ distinguish technical variation between mixtures, and between replicate mass spectrometry runs of a same mixture. ε represents the technical variation that is not explained by $Mixture$ and $TechRep(Mixture)$. As the result of detailed modeling of systematic sources of variation, random errors at the level of protein summaries ε_{mtcb} can be assumed independent and non-systematic. When the experimental design does not include replicates for all the sources of variation in Figure 2.2 some terms in Eq. (2.3) are not estimable. In this case *MSstatsTMT* fits simpler models as described in [15].

Parameters of the model are estimated using restricted maximum likelihood. Since the number of biological replicates in each condition is often small, *MSstatsTMT* adopts Empirical Bayes moderation of the standard errors, as proposed in the R package *limma* for analysis of gene expression microarrays [38]. Model-based tests for differentially abundant proteins between pairs of conditions is carried by comparing the terms *Condition* (see [15] for technical details). Finally, *MSstatsTMT* adjusts the p-values of the tests to account for multiple comparisons between the proteins by the method of Benjamini-Hochberg FDR [39].

2.3 Implementation of MSstatsTMT

We implemented this workflow for general group comparison designs in the open-source R/Bioconductor package *MSstatsTMT* [14]. *MSstatsTMT* includes converters

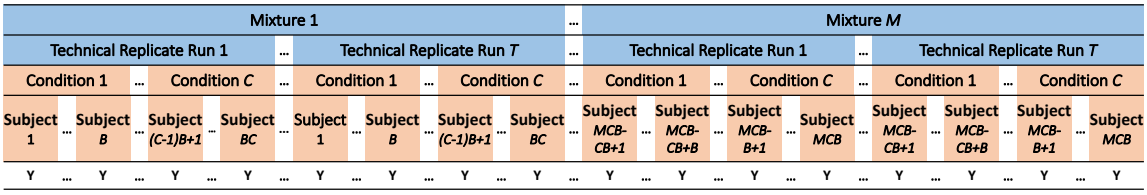

Figure 2.4.: **One protein data with group comparison design** The experiment has M mixtures, T technical replicates MS Runs per biological mixture, C conditions and MCB subjects, resulting in $MTCB$ observations.

from Proteome Discoverer, MaxQuant, OpenMS and SpectroMine. In addition to formatting the data, the converters construct spectral features as follows. The converters remove spectra with an excessive number of missing reporter ion intensities, or peptide identifications shared by multiple proteins. If multiple spectra have the same peptide ion identification, the converters only retain a single "best" spectrum with the minimal number of missing values, highest intensity, or lowest interference score. If the experiment contains fractions, and a peptide ion is present in multiple fractions, the peptide ion is only kept in the fraction where it has the highest mean intensity. If the peptide ion has the same highest mean intensity in multiple fractions, it is only kept in the fraction where it has highest maximal intensity. Proteins with more than one summary value in more than one condition are retained for the downstream statistical analysis.

The missing value imputation and protein summarization steps in *MSstatsTMT* rely on functionalities in the R package MSstats [35]. Statistical modeling, inference and hypothesis testing relies on the functionalities in the R packages *lme4* [40] and *lmerTest* [45]. The Empirical Bayes moderation relies on the functionalities in the R package *limma* [27]. Analyses of all the datasets in this manuscript were completed in under one hour on a MacBook Pro with Intel Core i5 and 8 GB memory.

2.4 Evaluation

2.4.1 Experimental datasets

This section summarizes the datasets in this chapter.

SpikeIn-5mix-MS3: controlled mixtures

The controlled mixtures were used to evaluate *MSstatsTMT* in situations with known ground truth. However, they lack biological variation, and therefore imperfectly represent real-life investigations.

Experimental design The controlled mixtures aimed to evaluate the ability of *MSstatsTMT* to deal with non-trivial designs with multiple TMT mixtures and multiple technical replicates. 500, 333, 250, and 62.5 fmol peptides from 48 UPS1 proteins were spiked-into 50 μg SILAC HeLa peptides in duplicate. This produced a dilution series corresponding to 1, 0.667, 0.5, and 0.125 times of the highest UPS1 peptide amount (500 fmol). In addition, a reference sample was generated by pooling all four diluted UPS1 peptide samples (286.5 fmol) and combined with 50 μg of SILAC HeLa in duplicate. These ten replicates were labeled with TMT 10-plex reagents, mixed and analyzed by LC-MS/MS. The procedure was repeated five times, to generate a total of five such controlled mixtures. Each mixture was profiled in three mass spectrometry runs, producing 15 MS runs from 5 TMT mixtures in total. The overall experimental design is shown in Figure 2.5.

Data acquisition and processing Raw data for SpikeIn-5mix-MS3 were acquired using SPS [19]. The data were processed with Proteome Discoverer 2.2.0.388 and Mascot Server 2.6.1. Statistical analyses with Proteome Discoverer were done within the software, using proteins marked as "Master" in the protein report. For all the other statistical analyses, reports from Proteome Discoverer 2.2 containing peptide-

spectrum matches (PSM) and reporter ion quantifications was exported to R. The PSM reports contained 5,903 proteins for the MS3 dataset. Since protein groups containing both spiked-in and background proteins complicated the calculation of the ground truth fold change, 1000 protein groups with multiple proteins were filtered out. For the same reason, we also removed 19 spiked-in UPS1 proteins sharing sequence with endogenous SILAC-HeLa proteins. The final dataset consisted of 4,812 proteins (including 21 UPS proteins).

Pairwise comparisons We evaluated the statistical approaches by their ability to detect changes in the abundance of UPS1 proteins between pairs of conditions. Each condition was labeled with the concentration of the UPS1 proteins, i.e., 1, 0.667, 0.5, and 0.125. The pairwise comparisons were labeled as the ratios of the concentrations of the UPS1 proteins, i.e., 0.667/0.5, 1/0.667, 1/0.5, 0.667/0.125, 0.5/0.125, and 1/0.125. Therefore, the true fold changes of the UPS1 proteins in these comparisons were 1.33, 1.5, 2, 5.328, 4, and 8.

Simulated datasets derived from SpikeIn-5mix-MS3

To evaluate *MSstatsTMT* in situations with both biological variation and known ground truth, we created two synthetic datasets by adding biological variation to SpikeIn-5mix-MS3.

SpikeIn-5mix-3TechRep-MS3-Sim We simulated a dataset with the same design as SpikeIn-5mix-MS3. The dataset consisted of five mixtures, each profiled with three technical replicate MS runs. The dataset was simulated by, first, summarizing all the spectra of a protein in SpikeIn-5mix-MS3 with *MSstatsTMT* as described below, and then adding to the protein summaries in each mixture a simulated random biological variation. Specifically, denote Y_{mtcb} the protein abundance in mixture m and technical MS run t, in the channel containing biological replicate b of condition c. The simulated protein abundance Z_{mtcb} was generated as

TMT10plex reagent		126	127N	127C	128N	128C	129N	129C	130N	130C	131
Mixture 1	Run 1	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref
	Run 2	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref
	Run 3	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref
Mixture 2	Run 4	Ref	0.5	1	0.667	0.125	1	0.667	0.125	0.5	Ref
	Run 5	Ref	0.5	1	0.667	0.125	1	0.667	0.125	0.5	Ref
	Run 6	Ref	0.5	1	0.667	0.125	1	0.667	0.125	0.5	Ref
Mixture 3	Run 7	Ref	0.125	0.667	1	0.5	0.5	0.125	0.667	1	Ref
	Run 8	Ref	0.125	0.667	1	0.5	0.5	0.125	0.667	1	Ref
	Run 9	Ref	0.125	0.667	1	0.5	0.5	0.125	0.667	1	Ref
Mixture 4	Run 10	Ref	1	0.5	0.125	0.667	0.667	1	0.5	0.125	Ref
	Run 11	Ref	1	0.5	0.125	0.667	0.667	1	0.5	0.125	Ref
	Run 12	Ref	1	0.5	0.125	0.667	0.667	1	0.5	0.125	Ref
Mixture 5	Run 13	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref
	Run 14	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref
	Run 15	Ref	0.667	0.125	0.5	1	0.125	0.5	1	0.667	Ref

Figure 2.5.: **Design of SpikeIn-5mix-MS3** Each row is a mixture, profiled in three technical replicate MS runs. Each column is a channel. Colors show conditions that represent the concentration of UPS1 proteins. The last column represents the reference channel and is used for local protein normalization. Each entry is the TMT10-plex label of a sample. Mixture 1 and 5 are identical since they have exactly the same design.

$Z_{mtcb} = Y_{mtcb} + \varepsilon_{mcb}$, where $\varepsilon_{mcb} \overset{iid}{\sim} Normal(0, \sigma_S^2)$. The same random term was added to all the technical replicates of a subject. We generated datasets with five values $\sigma_S = \{0.05, 0.1, 0.15, 0.2, 0.4\}$ motivated by the biological investigations in this manuscript.

SpikeIn-15mix-MS3-Sim We simulated another dataset with the same number of runs (15 MS runs total), but now including a larger number of biological replicates and no technical replicates. The simulated random biological variation was added to the protein summaries in SpikeIn-5mix-MS3 as described above.

2.4.2 Evaluation strategy

We evaluated the performance of *MSstatsTMT* v1.6.2 while comparing all pairs of conditions in the datasets in Experimental procedures, using workflows in Figure 2.1. All workflows except Proteome Discoverer took as input features produced by the *MSstatsTMT* converter. For *Ratio+Median+Limma*, we selected spectrum-level normalization without using reference channel as recommended by Kammers *et al.* [17]. For Proteome Discoverer 2.2, the input and the results of statistical analysis were as reported by the software.

We defined a *testable protein* a protein with enough data to perform a test for differential abundance with a particular workflow, and used the number of testable proteins in each workflow as a criterion for evaluation. Evaluations on the controlled mixtures and on the simulation experiments considered the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), where the total TP+FP+TN+FT equals to the number of testable proteins specific to each comparison and each workflow. We also considered the empirical false discovery rate (eFDR= FP / TP + FP), the sensitivity (TP / TP + FN) and the specificity (TN / TN + FP) of detecting differentially abundant proteins among the testable proteins at the FDR=0.05 cutoff. We further considered area under the ROC curve (AUC), which

represents sensitivity versus 1-specificity at various FDR-adjusted p-value cutoffs, calculated using R package *pROC* [46]. Finally, we compared the estimated fold changes to the true fold changes.

2.4.3 Evaluation on controlled mixtures

We first evaluated *MSstatsTMT* on controlled mixtures that contained ground truth, but lacked biological variation. Since the controlled mixtures had a balanced design, all normalization and summarization methods produced relatively similar results (similarly to the illustration in Figure 2.3), and differences in performance were primarily due to statistical modeling and inference.

MSstatsTMT best balanced the number of true and false positive differentially abundant proteins Figure 2.6 summarizes the performance of the representative workflows in Figure 2.1 on SpikeIn-5mix-MS3. For investigations without biological variation, *MSstatsTMT* fit the simple model that is similar (but not identical) to the models fit by *limma* and one-way ANOVA (implemented in *Proteome Discoverer*). Thus, these workflows had a similar number of testable proteins. *EdgeR* fit a different model, which assumed that reporter ion intensities were count data following a Negative Binomial distribution. The inappropriate assumption, combined with subsetting the dataset for each pair of conditions, negatively affected the number of testable proteins.

limma produced the largest number of both true and false positive differentially abundant proteins, and the largest eFDR. This was due to a combination of the treatment of missing values and of the Empirical Bayes step, which under-estimated the biological variation. Since the dataset had a relatively small number of true differentially abundant proteins (21) as compared to the background proteins (4791), all the workflows produced similar sensitivity, specificity and AUC (calculated with respect to their individual number of testable proteins). The performance of one-way

ANOVA was close (but slightly worse) than *MSstatsTMT*. Overall, *MSstatsTMT* best balanced the number of true positives and the eFDR.

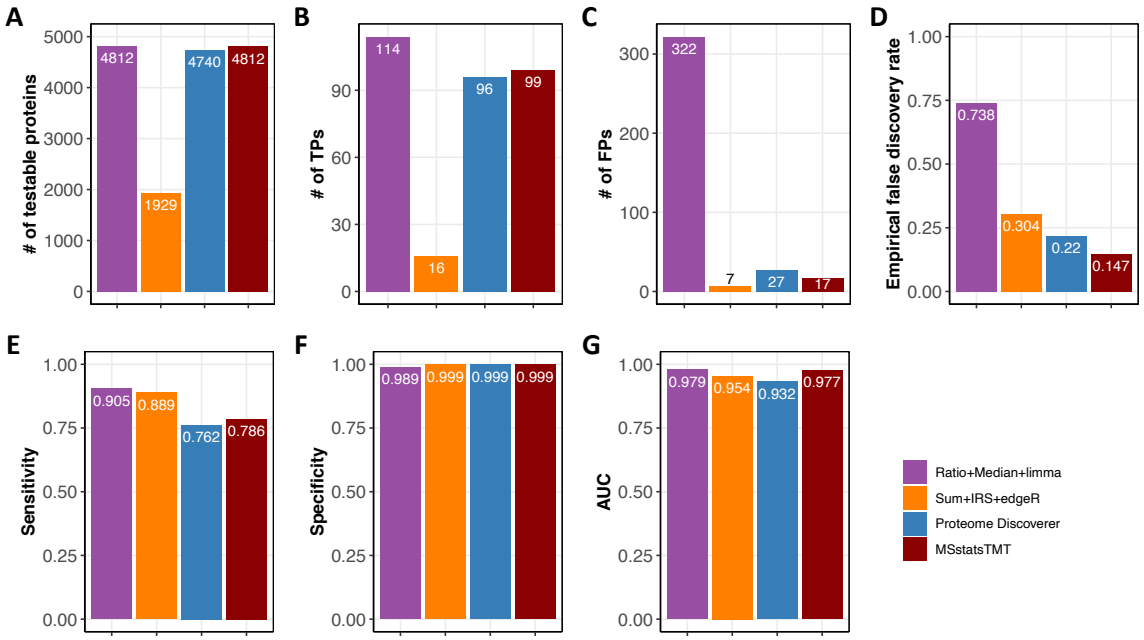

Figure 2.6.: **Detection of differentially abundant proteins in all pairs of conditions in SpikeIn-5mix-MS3 (FDR cutoff of 0.05)** Colors represent statistical modeling and inference methods in Figure 2.1. (A) Number of testable proteins. (B) Number of true positive differentially abundant proteins. (C) Number of false positive differentially abundant proteins. (D) Empirical false discovery rate. (E) Sensitivity of correctly detecting the spiked-in proteins. (F) Specificity of correctly detecting the background proteins. (G) Area under ROC curve (AUC).

2.4.4 Evaluation on controlled mixtures with simulated source of biological variation

To evaluate *MSstatsTMT* in experiments in presence of biological variation, we simulated various amounts of biological variation added to SpikeIn-5mix-MS3 as described in **Experimental Datasets**. As before, due to the balanced nature of the

designs, differences in performance primarily come from statistical modeling and inference.

MSstatsTMT accurately characterized biological variation in investigations with both biological and technical replicates The simulated dataset SpikeIn-5mix-3TechRep-MS3-Sim had five biological mixtures, three technical replicate MS runs per mixture, and a balanced design. Figure 2.7 summarizes the performance of the workflows. *MSstatsTMT* fit the model in Eq. (2.3), which distinguished these sources of variation. In contrast, the models in *limma*, *edgeR* and one-way ANOVA (implemented in *Proteome Discoverer*) did not have enough flexibility to distinguish biological and technical variation. They combined the variation from these two sources, which lead to over-estimation of the degrees of freedom and under-estimation of the standard error, and increased the false positive differentially abundant proteins.

MSstatsTMT best balanced true and false positive differentially abundant proteins in investigations with many biological replicates We further simulated an experiment SpikeIn-15mix-MS3-Sim with the same number of runs, but no technical replicates. Instead, it contained up to 15 distinct biological mixtures, corresponding to up to 30 biological replicates per condition.

Since the experiment did not include technical replicates, the models in *limma* and one-way ANOVA (implemented in *Proteome Discoverer*) were similar (but not identical) to the model in *MSstatsTMT*. Figure 2.8 and Figure 2.9 illustrate that increasing the number of biological replicates (and the number of mixtures) improved the sensitivity and the specificity of most workflows. This underscored the importance of biological replicates for achieving accurate results.

At the same time, additional mixtures introduced more missing values, with up to 60% of proteins having at least one missing summary. Therefore, the difference in performance was due primarily to the treatment of missing values. In *limma*, the negative impact of treatment of missing values and of the Empirical Bayes step was

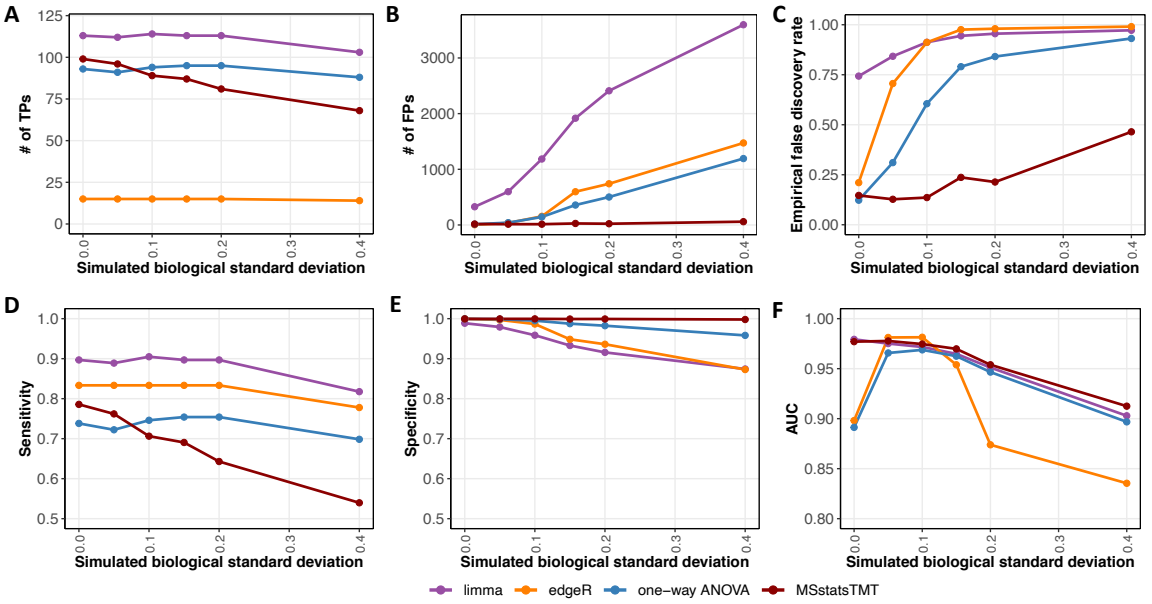

Figure 2.7.: **Detection of differentially abundant proteins in all pairs of conditions in SpikeIn-5mix-3TechRep-MS3-Sim (FDR cutoff of 0.05)** Colors represent statistical modeling and inference methods in Figure 2.1. X-axis: simulated biological standard deviation (standard deviation = 0 corresponds to the original controlled mixtures SpikeIn-5mix-MS3). (A) True positive differentially abundant proteins. (B) False positive differentially abundant proteins. (C) Empirical false discovery rate. (D) Sensitivity of detecting the spiked-in proteins. (E) Specificity of detecting the background proteins. (F) Area under ROC curve (AUC).

exacerbated, and resulted in a large number of false positive differentially abundant proteins. One-way ANOVA was less sensitive than *MSstatsTMT*, however the discrepancy became smaller with the increase of sample size. *edgeR* filtered out proteins with missing summaries, and therefore reported the smallest number of testable proteins, true positive and false positive differentially abundant proteins.

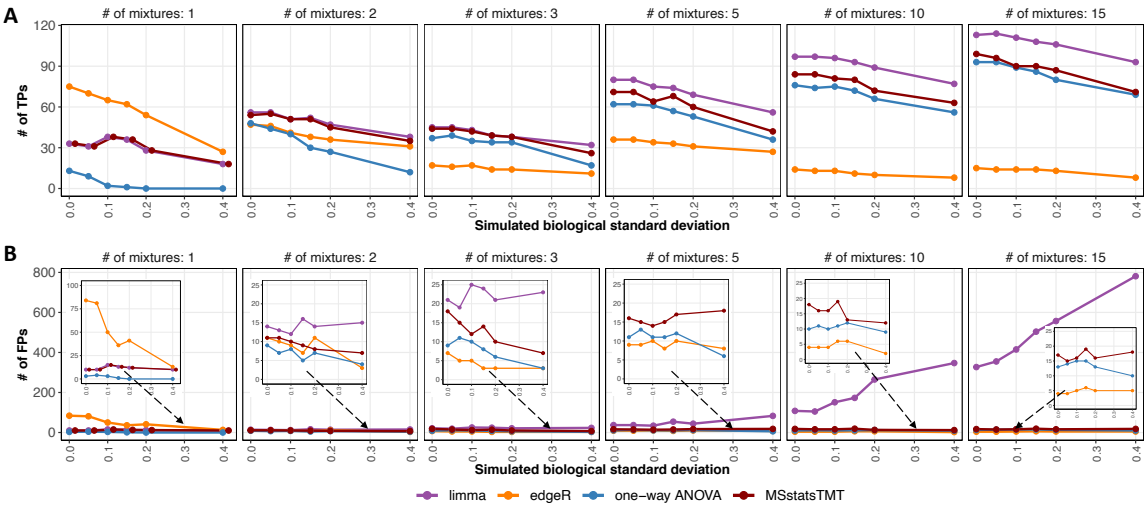

Figure 2.8.: **Detection of differentially abundant proteins in all pairs of conditions in SpikeIn-15mix-MS3-Sim (FDR cutoff of 0.05).** SpikeIn-15mix-MS3-Sim experiment simulated 15 biological mixtures and no technical replicates. Panels in the figure represent randomly selected subsets of 1, 2, 3, 5, 10 mixtures, and 15 mixtures. Colors represent statistical modeling and inference methods in Figure 2.1. X-axis: simulated biological standard deviation (standard deviation = 0 corresponds to the original controlled mixtures SpikeIn-5mix-MS3). (A) Number of true positive differentially abundant proteins. (B) Number of false positive differentially abundant proteins.

2.5 Discussion

This chapter proposes a statistical workflow for detecting differentially abundant proteins in mass spectrometry-based proteomic experiments with TMT labeling and group comparison designs. The workflow is implemented as an open-source

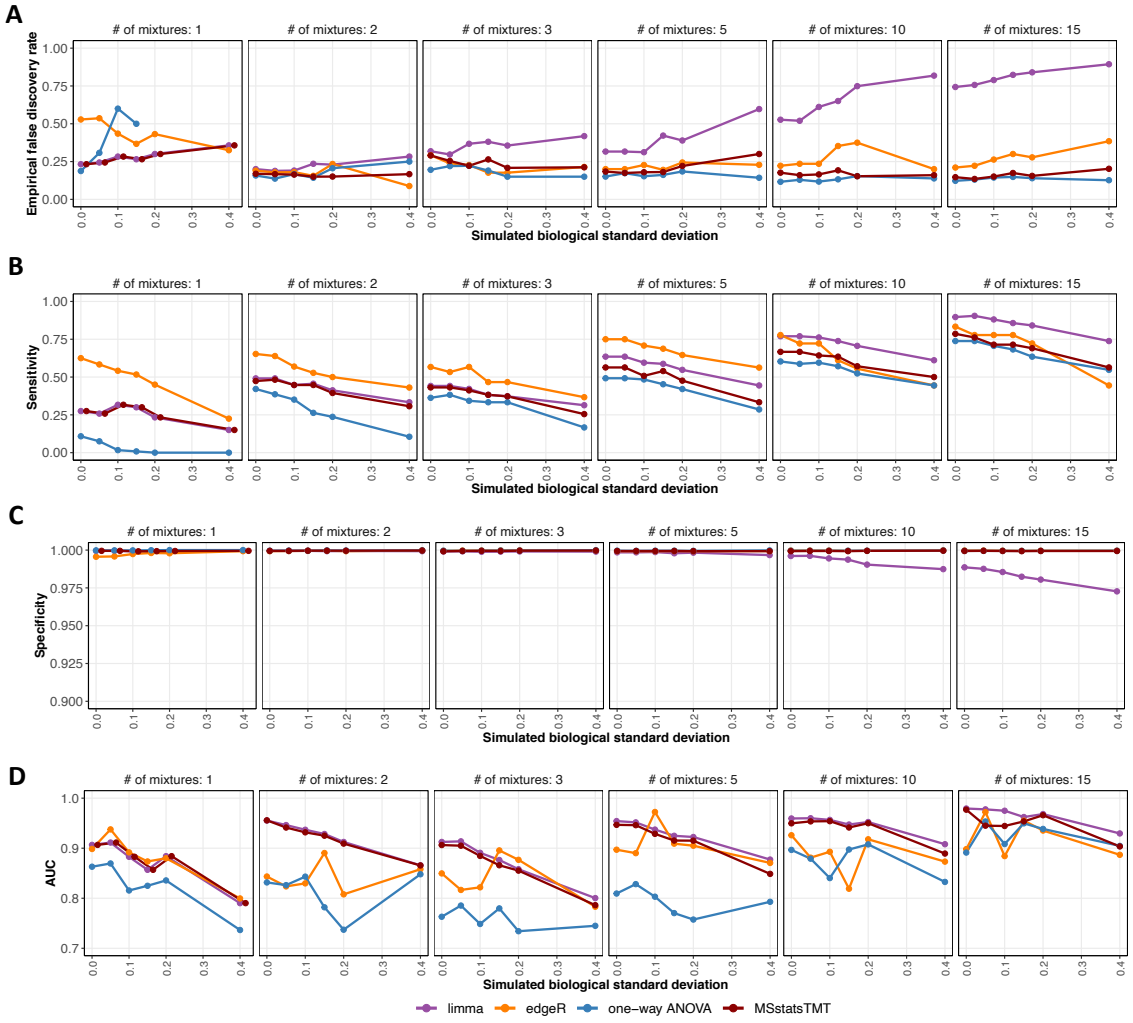

Figure 2.9.: **Detection of differentially abundant proteins in all pairs of conditions in SpikeIn-15mix-MS3-Sim (FDR cutoff of 0.05).** SpikeIn-15mix-MS3-Sim experiment simulated 15 biological mixtures and no technical replicates. Panels in the figure represent randomly selected subsets of 1, 2, 3, 5, 10 mixtures, and 15 mixtures. Colors represent statistical modeling and inference methods in Figure 2.1. X-axis: simulated biological standard deviation (standard deviation = 0 corresponds to the original controlled mixtures SpikeIn-5mix-MS3). (A) Empirical false discovery rate. (B) Sensitivity of correctly detecting the spiked-in proteins. (C) Specificity of correctly detecting the background proteins. (D) Area under ROC curve (AUC).

R/Bioconductor package, which takes as input exports from data processing tools such as Proteome Discoverer, MaxQuant, OpenMS, or SpectroMine.

Our evaluations indicate that performance of *MSstatsTMT*, as well as of all the other workflows, depends on the intrinsic characteristics of the investigation. For the protein-level normalization step, the use of a reference channel has little impact on the detection of differentially abundant proteins in balanced designs. However, in unbalanced designs, normalization without a reference channel can eliminate true fold changes, and reference channel-based normalization is preferred. The presence of a reference channel also improves the probability of selecting a peptide ion for fragmentation and the accuracy of the peptide identification. Therefore, *MSstatsTMT* encourages the users to add a reference channel to their designs.

Similarly, for the statistical modeling and inference step, most workflows performed similarly well in simple designs. For the controlled mixtures with no biological variation, high signal-to-noise ratio, and balanced designs, *MSstatsTMT* performed similarly to *Ratio+Median+Limma* and *Proteome Discoverer*. The situation changes when the investigation incorporates larger biological variation, more biological replicates and mixtures, and combinations of biological and technical replicates. Such complex designs require more consideration regarding statistical modeling. Since *limma*, *edgeR* and one-way ANOVA do not distinguish between the biological and the technical variance, and have limitations in handling missing values. This leads to inaccuracies in the inference, and loss of performance. In contrast, *MSstatsTMT* selects an appropriate model for each protein, and reflects both the experimental design and the pattern of missing protein summaries. This increases the sensitivity of detecting differentially abundant proteins while controlling false positive rate.

3. FUTURE WORK: MSSTATSTMT WITH COMPLEX DESIGNS

The work presented so far focuses on the TMT experiments with a group comparison design. We now propose the next steps for extending the models to deal with complex designs. Specifically, we are interested in the experiments with repeated measures designs. Repeated measures design is different from the group comparison design in the previous chapter. It takes repeated measurements of the protein abundance on each subject and utilizes the same subject (e.g., patient, cell line) for each condition in the study. Therefore, the subject serves as an additional block and the replicates within a block may be viewed as several conditions (paired design) or a single condition/treatment that is evaluated at different points in time (time-course design). Thus, the proposed model needs to take the subject block into account and be applicable to diverse cases of repeated measures designs.

3.1 Proposed model

In TMT experiments with time course design, the sub-subplot structure is the same as in a group comparison design but the whole-sub plot is different from a group comparison design. Figure 3.1 shows that time course design repeatedly measures protein abundance on a same subject across conditions/time points. For time course

designs with multiple biological replicates, multiple mixtures, and multiple technical replicates (such as in Figure 2.4), we propose to fit the following model:

$$Y_{mtcb} = \mu + Mixture_m + TechRep(Mixture)_{t(m)} + Condition_c + \tag{3.1}$$
$$Subject(Mixture)_{b(m)} + Condition \times Subject(Mixture)_{cb(m)} + \varepsilon_{mtcb}$$
$$\text{where } Mixture_m \overset{iid}{\sim} N(0, \sigma^2_M),\ TechRep(Mixture)_{t(m)} \overset{iid}{\sim} N(0, \sigma^2_T),\ \sum_{c=1}^{C} Condition_c = 0,$$
$$Subject(Mixture)_{b(m)} \overset{iid}{\sim} N(0, \sigma^2_S),\ Condition \times Subject(Mixture)_{cb(m)} \overset{iid}{\sim} N(0, \sigma^2_{CS}),$$
$$\varepsilon_{mtcb} \overset{iid}{\sim} N(0, \sigma^2)$$

$Mixture$ and $TechRep(Mixture)$ distinguish technical variation between mixtures, and between replicate mass spectrometry runs of a same mixture. $Subject$ represents biological replicates and is nested under $Mixture$. Since all the conditions under one mixture are designed to use same subjects, $Condition \times Subject(Mixture)_{cb(m)}$ estimates the protein abundance change across conditions for a same subject. ε represents the technical variation that is not explained by $Mixture$ and $TechRep(Mixture)$. Based on the model in Eq. (3.2), we will develop the ANOVA decomposition and pairwise comparisons of conditions.

Mixture 1	...	Mixture *M*
Technical Replicate Run 1 ... Technical Replicate Run *T*	...	Technical Replicate Run 1 ... Technical Replicate Run *T*
Condition 1 ... Condition *C* ... Condition 1 ... Condition *C*	...	Condition 1 ... Condition *C* ... Condition 1 ... Condition *C*
Subject 1 ... Subject *B* ... Subject 1 ... Subject *B* ... Subject 1 ... Subject *B* ... Subject 1 ... Subject *B*	...	Subject *(M-1)B+1* ... Subject *MB* ... Subject *(M-1)B+1* ... Subject *MB* ... Subject *(M-1)B+1* ... Subject *MB* ... Subject *(M-1)B+1* ... Subject *MB*
Y ... Y ... Y ... Y ... Y ... Y ... Y ... Y	...	Y ... Y ... Y ... Y ... Y ... Y ... Y ... Y

Figure 3.1.: **One protein data with time course design** The experiment has M mixtures, T technical replicates MS Runs per biological mixture, C conditions/time points and MB subjects per condition, resulting in $MTCB$ observations. The protein abundances in these MB subjects are repeatedly measured for each condition.

One special case of time course designs contains as many conditions as the channels (such as in Figure 3.2). In other words, each mixture in the experiment has one subject per condition and we propose the following reduced model

$$Y_{mtcb} = \mu + Mixture_m + TechRep(Mixture)_{t(m)} + Condition_c + \varepsilon_{mtc} \tag{3.2}$$

$$\text{where } Mixture_m \overset{iid}{\sim} N(0, \sigma_M^2),\ TechRep(Mixture)_{t(m)} \overset{iid}{\sim} N(0, \sigma_T^2),\ \sum_{c=1}^{C} Condition_c = 0$$

$$\varepsilon_{mtc} \overset{iid}{\sim} N(0, \sigma^2)$$

The term ε combines the biological variation and the technical variation that is not explained by $Mixture$ and $TechRep(Mixture)$.

Mixture 1								Mixture *M*							
Technical Replicate Run 1				Technical Replicate Run *T*				Technical Replicate Run 1				Technical Replicate Run *T*			
Condition 1	…	Condition *C*	…	Condition 1	…	Condition *C*	…	Condition 1	…	Condition *C*	…	Condition 1	…	Condition *C*	
Subject 1	…	Subject 1	…	Subject 1	…	Subject 1	…	Subject *M*	…	Subject *M*	…	Subject *M*	…	Subject *M*	
Y	…	Y	…	Y	…	Y	…	Y	…	Y	…	Y	…	Y	

Figure 3.2.: **One protein data with time course design** The experiment has M mixtures, T technical replicates MS Runs per biological mixture, C conditions/Channels and one subject per condition per mixture, resulting in MTC observations. The protein abundances in these B subjects are repeatedly measured for each condition in each mixture.

Paired design is often used when the experiments have disease tissue and healthy tissue from a same patient. It can be viewed as another special case of time course designs where there are only two conditions or time points. Thus, the model in Eq. (3.2) can be directly applied to the experiments with paired design.

All the designs above are assume to be balanced, i.e., each mixture has the same composition of conditions and subjects. However, a balanced design is sometimes difficult to achieve. For example, it is not always able to take both the disease tissue and healthy tissue from same patient and a subset of subjects are not paired. For such case, the corresponding linear models have to be adjusted.

3.2 Implementation

We propose to implement this workflow for general repeated measures designs in the open-source R/Bioconductor package *MSstatsTMT* [14]. Statistical modeling, inference and hypothesis testing relies on the functionalities in the R packages *lme4* [40] and *lmerTest* [45].

3.3 Evaluation

We plan to evaluate the performance of the proposed model on the following datasets:

- CPTAC data with paired design [47]: 111 unique tumor samples with 102 paired normal adjacent tissues (NATs) samples were distributed among 25 10-plex TMT mixtures. Each TMT mixture uses the first 9 channels to label 9 individual samples and uses the last channel for normalization. The first 8 channels of each mixture contained 4 tumor/NAT pairs where each pair of patient samples is adjacent to each other.

- SARS-CoV-2 data with time course design [48]: Cells were infected with SARS-CoV-2 and profiled for proteome at different times after infection. Control samples and virus samples after 2, 6,10 and 24 hours, as well as one reference sample, were labeled with 11-plex TMT. This experiment was repeated three times (3 control and 3 virus cell lines).

- SWIP data with spatial design [49]: 7 subcellular fractions from brain tissue were isolated from one wild-type(WT) and one mutant(MUT) mouse. These samples, as well as two pooled quality control (QC) samples, were labeled with 16-plex TMT. This experiment was repeated three times (3 WT and 3 MUT mice).

We will use the same evaluation metrics as that for a group comparison design, including the number of testable proteins and differentially abundant proteins.

3.4 Plan for completion of the research

Timeline	Task	Progress
	MSstatsTMT with group comparison design	Complete
	Sample size estimation for classification	Complete
2020.12	Collect TMT datasets with complex design	Ongoing
2021.01	Statistical modeling and inference for paired design	
2021.02-2021.03	Statistical modeling and inference for time course design	
2021.04	Statistical modeling and inference for complex design	
2021.05	Sample size estimation for TMT experiments	
2021.06-2021.07	Thesis writing and defense	

REFERENCES

REFERENCES

[1] Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, 2016.

[2] Johannes Griss, Goran Vinterhalter, and Veit Schwammle. IsoProt: a complete and reproducible workflow to analyze iTRAQ/TMT experiments. *Journal of Proteome Research*, 18(4):1751–1759, 2019.

[3] Ana Martinez-Val, Fernando Garcia, Pilar Ximenez-Embun, Nuria Ibarz, Eduardo Zarzuela, Isabel Ruppen, Shabaz Mohammed, and Javier Munoz. On the statistical significance of compressed ratios in isobaric labeling: a cross-platform comparison. *Journal of Proteome Research*, 15(9):3029–3038, 2016.

[4] Navin Rauniyar and John R Yates III. Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of Proteome Research*, 13(12):5293–5309, 2014.

[5] Andrew Thompson, Jürgen Schäfer, Karsten Kuhn, Stefan Kienle, Josef Schwarz, Günter Schmidt, Thomas Neumann, and Christian Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904, 2003.

[6] Philip L Ross, Yulin N Huang, Jason N Marchese, Brian Williamson, Kenneth Parker, Stephen Hattan, Nikita Khainovski, Sasi Pillai, Subhakar Dey, Scott Daniels, Subhasish Purkayastha, Peter Juhasz, Stephen Martin, Michael Bartlet-Jones, Feng He, Allan Jacobson, and Darryl J Pappin. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12):1154–1169, 2004.

[7] Jeremy D O'Connell, Joao A Paulo, Jonathon J O'Brien, and Steven P Gygi. Proteome-wide evaluation of two common protein quantification methods. *Journal of Proteome Research*, 17(5):1934–1942, 2018.

[8] Jan Muntel, Joanna Kirkpatrick, Roland Bruderer, Ting Huang, Olga Vitek, Alessandro Ori, and Lukas Reiter. Comparison of protein quantification in a complex background by DIA and TMT workflows with fixed instrument time. *Journal of Proteome Research*, 2019.

[9] Ann L Oberg, Douglas W Mahoney, Jeanette E Eckel-Passow, Christopher J Malone, Russell D Wolfinger, Elizabeth G Hill, Leslie T Cooper, Oyere K Onuma, Craig Spiro, Terry M Therneau, , and H Robert Bergen III. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *Journal of Proteome Research*, 7(1):225–233, 2008.

[10] Alejandro Brenes, Jens L Hukelmann, Dalila Bensaddek, and Angus I Lamond. Multi-batch TMT reveals false positives, batch effects and missing values. *Molecular & Cellular Proteomics*, pages mcp–RA119, 2019.

[11] Lily Ting, Ramin Rad, Steven P Gygi, and Wilhelm Haas. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature Methods*, 8(11):937, 2011.

[12] Mikhail M Savitski, Toby Mathieson, Nico Zinn, Gavain Sweetman, Carola Doce, Isabelle Becher, Fiona Pachl, Bernhard Kuster, and Marcus Bantscheff. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *Journal of Proteome Research*, 12(8):3586–3598, 2013.

[13] Paul D Piehowski, Vladislav A Petyuk, Daniel J Orton, Fang Xie, Ronald J Moore, Manuel Ramirez-Restrepo, Anzhelika Engel, Andrew P Lieberman, Roger L Albin, David G Camp, Richard D Smith, and Amanda J Myers. Sources of technical variability in quantitative lc–ms proteomics: human brain tissue sample analysis. *Journal of proteome research*, 12(5):2128–2137, 2013.

[14] Ting Huang, Meena Choi, Sicheng Hao, and Olga Vitek. MSstatsTMT: Protein Significance Analysis in shotgun mass spectrometry-based proteomic experiments with tandem mass tag (TMT) labeling. *R package Bioconductor*, v1.6.2, 2020.

[15] Ting Huang, Meena Choi, Manuel Tzouros, Sabrina Golling, Nikhil Janak Pandya, Balazs Banfai, Tom Dunkley, and Olga Vitek. Msstatstmt: Statistical detection of differentially abundant proteins in experiments with isobaric labeling and multiple mixtures. *Molecular & Cellular Proteomics*, 19(10):1706–1723, 2020.

[16] Shelley M Herbrich, Robert N Cole, Keith P West Jr, Kerry Schulze, James D Yager, John D Groopman, Parul Christian, Lee Wu, Robert N O'Meally, Damon H May, Martin W McIntosh, and Ingo Ruczinski. Statistical inference from multiple iTRAQ experiments without using common reference standards. *Journal of Proteome Research*, 12(2):594–604, 2013.

[17] Kai Kammers, Robert N Cole, Calvin Tiengwe, and Ingo Ruczinski. Detecting significant changes in protein abundance. *EuPA Open Proteomics*, 7:11–19, 2015.

[18] Deanna L Plubell, Phillip A Wilmarth, Yuqi Zhao, Alexandra M Fenton, Jessica Minnier, Ashok P Reddy, John Klimek, Xia Yang, Larry L David, and Nathalie Pamir. Extended multiplexing of tandem mass tags (TMT) labeling reveals age and high fat diet specific proteome changes in mouse epididymal adipose tissue. *Molecular & Cellular Proteomics*, 16(5):873–890, 2017.

[19] Graeme C McAlister, David P Nusinow, Mark P Jedrychowski, Martin Wuhr, Edward L Huttlin, Brian K Erickson, Ramin Rad, Wilhelm Haas, and Steven P Gygi. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical Chemistry*, 86(14):7150–7158, 2014.

[20] Magnus Ø Arntzen, Christian J Koehler, Harald Barsnes, Frode S Berven, Achim Treumann, and Bernd Thiede. IsobariQ: software for isobaric quantitative proteomics using IPTL, iTRAQ, and TMT. *Journal of Proteome Research*, 10(2):913–920, 2011.

[21] Bo Wen, Ruo Zhou, Qiang Feng, Quanhui Wang, Jun Wang, and Siqi Liu. IQuant: an automated pipeline for quantitative proteomics based upon isobaric tags. *Proteomics*, 14(20):2280–2285, 2014.

[22] Natasha A Karp, Wolfgang Huber, Pawel G Sadowski, Philip D Charles, Svenja V Hester, and Kathryn S Lilley. Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics*, 9(9):1885–1897, 2010.

[23] Evelyne Maes, Wahyu Wijaya Hadiwikarta, Inge Mertens, Geert Baggerman, Jef Hooyberghs, and Dirk Valkenborg. CONSTANd: A normalization method for isobaric labeled spectra by constrained optimization. *Molecular & Cellular Proteomics*, 15(8):2779–2790, 2016.

[24] Laurent Gatto and Kathryn S Lilley. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, 28(2):288–289, 2011.

[25] Florian P. Breitwieser, Andre Muller, Loic Dayon, Thomas Kocher, Alexandre Hainard, Peter Pichler, Ursula Schmidt-Erfurth, Giulio Superti-Furga, Jean-Charles Sanchez, Karl Mechtler, Keiryn L. Bennett, and Jacques Colinge. General statistical modeling of data from protein relative expression isobaric tags. *Journal of Proteome Research*, 10(6):2758–2766, 2011. PMID: 21526793.

[26] Martina Fischer and Bernhard Y Renard. iPQF: a new peptide-to-protein summarization method using peptide spectra characteristics to improve protein quantification. *Bioinformatics*, 32(7):1040–1047, 2016.

[27] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[28] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 2012.

[29] Lianbo Yu, Parul Gulati, Soledad Fernandez, Michael Pennell, Lawrence Kirschner, and David Jarjoura. Fully moderated T-statistic for small sample size gene expression arrays. *Statistical Applications in Genetics and Molecular Biology*, 10(1):42, 2011.

[30] Gina D'Angelo, Raghothama Chaerkady, Wen Yu, Deniz Baycin Hizal, Sonja Hess, Wei Zhao, Kristen Lekstrom, Xiang Guo, Wendy I White, Lorin Roskos, Michael A Bowen, and Harry Yang. Statistical models for the analysis of isobaric tags multiplexed quantitative proteomics. *Journal of Proteome Research*, 16(9):3124–3136, 2017.

[31] Joao A Paulo, Jeremy D O'Connell, Robert A Everley, Jonathon O'Brien, Micah A Gygi, and Steven P Gygi. Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 S. cerevisiae proteins across 10 carbon sources. *Journal of Proteomics*, 148:85–93, 2016.

[32] Elizabeth G Hill, John H Schwacke, Susana Comte-Walters, Elizabeth H Slate, Ann L Oberg, Jeanette E Eckel-Passow, Terry M Therneau, and Kevin L Schey. A statistical model for iTRAQ data analysis. *Journal of Proteome Research*, 7(8):3091–3101, 2008.

[33] J Martin Bland and Douglas G Altman. Statistics notes: the use of transformation when comparing two means. *British Medical Journal*, 312(7039):1153, 1996.

[34] Stephen J Callister, Richard C Barry, Joshua N Adkins, Ethan T Johnson, Weijun Qian, Bobbie-Jo M Webb-Robertson, Richard D Smith, and Mary S Lipton. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of Proteome Research*, 5(2):277–286, 2006.

[35] Meena Choi, Cyril Galitzine, Tsai Tsung-Heng Huang, Ting, and Olga Vitek. MSstats: Protein Significance Analysis in DDA, SRM and DIA for Label-free or Label-based Proteomics Experiments. *R package Bioconductor*, v3.20.0, 2014.

[36] John D Kalbfleisch and Ross L Prentice. *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons, 2011.

[37] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[38] Gordon Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 3, 2004.

[39] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[40] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[41] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[42] David A Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[43] Ting Wang and Edgar C Merkle. merderiv: Derivative computations for linear mixed effects models with application to robust standard errors. *Journal of Statistical Software, Code Snippets*, 87(1):1–16, 2018.

[44] Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

[45] Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 2017.

[46] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1):77, 2011.

[47] Michael A Gillette, Shankha Satpathy, Song Cao, Saravana M Dhanasekaran, Suhas V Vasaikar, Karsten Krug, Francesca Petralia, Yize Li, Wen-Wei Liang, Boris Reva, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, 182(1):200–225, 2020.

[48] Denisa Bojkova, Kevin Klann, Benjamin Koch, Marek Widera, David Krause, Sandra Ciesek, Jindrich Cinatl, and Christian Münch. Proteomics of sars-cov-2-infected host cells reveals therapy targets. *Nature*, 583(7816):469–472, 2020.

[49] Jamie L Courtland, Tyler W Bradshaw, Greg Waitt, Erik J Soderblom, Tricia Ho, Anna Rajab, Ricardo Vancini, Il Hwan Kim, and Scott Soderling. Genetic disruption of washc4 drives endo-lysosomal dysfunction and cognitive-movement impairments in mice and humans. *BioRxiv*, 2020.