DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Evaluating Measurement Error in Mass Spectrometry-Based Proteomics

Citation

Lim, Matt Y. 2020. Evaluating Measurement Error in Mass Spectrometry-Based Proteomics. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37365524

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Evaluating measurement error in mass spectrometry-based proteomics

A dissertation presented

by

Matthew Yue Cheng Lim

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Virology

Harvard University

Cambridge, Massachusetts

December 2019

© 2019 Matthew Yue Cheng Lim

All rights Reserved

Dissertation Advisor: Steven P. Gygi Matthew Yue Cheng Lim

Evaluating measurement error in mass spectrometry-based proteomics.

Abstract

Over the last few decades, liquid chromatography with tandem mass spectrometry (LC-MS/MS) has become a pillar in the field of proteomics. As the popularity of and access to LC-MS/MS techniques grows, a concerted effort has been dedicated to increase its quantitative capabilities. Generally, the quantitative mass spectrometry-based proteomics field is divided into two camps: 1) users of isotope labels and 2) those who prefer a label-free approach. While labelling techniques, such as Tandem Mass Tags (TMT), avoid the stochasticity of single sample analyses by multiplexing, isolation of unwanted ions can lead to interference thereby negatively affecting measurement accuracy. Conversely, while label-free methods are not susceptible to peptide interference, stochasticity results in a large increase of missing values, limiting the quantitative usefulness of these data sets.

The work of this dissertation has been focused on understanding how measurement error occurs across various LC-MS/MS experiments from instrumentation to downstream data analysis. We first analyzed how the latest mass spectrometry technologies affect peptide interference, one of the major causes of measurement error in TMT LC-MS/MS experiments. In order to obtain large highly quantitative data sets, our results show that a balance must be maintained between interference, signal-to-noise, and identification rates. Secondly, we developed a two-sample, two-proteome experiment to evaluate the commonly-used Match-Between-Runs algorithm (MBR) – a software solution used by the label-free community in an

attempt to circumvent the missing values problem inherent to label-free LC-MS/MS. We find that while MBR will incorrectly transfer identifications at a large rate, quantitative algorithms bundled with the software will not quantity most incorrect transfers. While this audit enforces quantitative accuracy, it effectively negates any gains produced by MBR. Lastly, we developed a TMT-based approach to measure global phosphorylation occupancy but found that mass spectrometry measurements were inadequate at reliably measuring the small changes associated with this type of experiment leading to estimations of negative occupancies. As such, a Bayesian statistical tool was developed to better model the data. By utilizing the measurement error and generating credible intervals, we report an elegant strategy for presenting phosphorylation occupancy data that better encompasses the estimation, as well as related uncertainty.

Table of Contents

List of Figures

List of Tables

Acknowledgements

When I first moved to Boston to start graduate school, I do not think I could have imagined what the next five and a half years would have in store for me. If it takes a village to raise a child, then it takes a city like Boston to raise a graduate student. The journey to complete this dissertation would have been impossible alone. I am forever grateful for the time, energy, wisdom, and knowledge that my colleagues, friends, and family have invested in me over the last five years.

Perhaps the largest elephant in the room in a graduate student's journey is their lab – I am incredibly grateful for the opportunity to have found a temporary home in the laboratory of Dr. Steven P. Gygi. When I joined his group at the start of 2016, I could tell there was something special about the group; Dr. Gygi's pursuit of accuracy and precision and his vast knowledge of mass spectrometry are second to none. His continuous pushing to make me a better presenter, writer, and scientist were invaluable and whatever accomplishments I have achieved throughout my tenure are a credit to his mentorship.

But what made the group peerless was the caliber of the post-doctoral scholars with which Dr. Gygi surrounded himself. The countless nuggets of knowledge I was able to glean from conversations were incredibly humbling. Needless to say, my work would have been impossible without input and guidance from each of them. Countless episodes of priceless instruction from Edward Huttlin, David Nusinow, Julian Mintseris, Qing Yu, Devin Schweppe, and Brian Erickson, among others, have allowed me to understand enough about mass spectrometry and bioinformatics to realize I know hardly enough.

Additionally, I would like to extend a special thanks to Jonathon O'Brien, a former lab member, who helped immensely with the publication of my first first-author academic paper. I

am also incredibly grateful for the friendship and support of the group's wet-lab gurus, Ekaterina Stepanova (Katya) and Tian Zhang. From helping me get lab supplies to giving me rides home when it was late and raining, it was a privilege to call them lab mates and friends.

However, within the Gygi group, there was one member who went far above the call of duty to help a fledgling graduate student grow – João A. Paulo. From day one, João was always looking out for me. I cannot emphasize how much the man has done for the lab and for me. It should be noted that every single mass spectrometry experiment I have conducted in the Gygi lab was analyzed on an instrument João maintained and every written work was edited by him (whether he likes that attribution or not). The countless hours João has spent mentoring me and showing me tough love are the cornerstone of the academic side of my graduate school career.

I would also like to acknowledge Maria Bollinger from the Virology Program at Harvard Medical School. Understanding the bureaucratic hoops one must jump through in graduate school is probably worth a degree in itself. Without Maria's help to navigate the behind the scenes paperwork, I probably would still be missing a dozen forms. Her friendship and wisdom always reminded me that there was more to life than the marble halls of academia.

Outside the walls of Harvard, I am forever grateful for my friends, church, and family. It is a credit to their support and love they freely gave me during the hardest times of graduate school that I have survived the ordeal. Specifically, Lee-Shing Chang and Wilbur Hu – two of my closest friends in Boston that I met through the church City on a Hill – have lent me their ears to complain and vent when things were not working. Their words of wisdom, countless prayers, and incessant encouragement were extremely precious to me and I am eternally grateful for their friendship.

Moving to Boston alone was a bit of an experience and I was only able to survive a year of it before asking my wife, Andrea Wong Lim, and my childhood friend, Stefan Takamura, to make the trek out to New England. Thankfully, they acquiesced to my request and moved in with me, subsidizing my rent. Between the cold winters and high cost of living, I am grateful they did not run back to sunny Southern California at the first chance they got. As this chapter in my life closes, I can look back fondly at the many zany adventures we shared that helped me keep my sanity throughout my graduate school journey.

I would be remiss if I did not dedicate a portion of my acknowledgements to my wife specifically. Andrea Noelle Wong Lim has been an amazing blessing to me and I do not think I would have been able to finish this dissertation without her love and support. She always reminds me that the most important things in life are the people around us. Her presence is a never-ending reminder that true success is not tied to my career choices but what kind of person I choose to be.

Lastly, I would like to thank my parents John and Caroline Lim. Their sacrifices throughout my childhood are why I was able to make it to graduate school in the first place. I do not think this dissertation would have existed without their undying love and support.

Soli Deo gloria

Chapter 1: Introduction

Proteomics is an area of research dedicated to the large-scale investigation of proteins, from their sequencing and identification to interaction networks within cells or tissues[1,2]. While the earliest concepts of modern proteomics can be traced back to the 1970s, it was not until the mid-1990s that the field began taking its current shape[1,3–5]. Previously, two-dimensional gel electrophoresis (2DE) coupled with a sequencing technique like Edman-degradation was the norm; however, during this era of growth, tandem mass spectrometry (MS/MS) began to rise in popularity. Advances in mass spectrometry and database searching paved the way for MS/MS to become the premiere technology for protein sequencing and identification[5–7]. Additionally applications of stable-isotope dilution theory began to find its way into mass spectrometry-based proteomics giving the traditionally discovery-based technique a more reliable quantitative spin[8–10]. The rise of MS/MS provided reliable, high-throughput protein identification with specific quantitation that rivaled, and in some cases surpassed, antibody-based methods like the enzyme-linked immunosorbent assay (ELISA) and Western blotting.

It should be noted that MS/MS in the field of proteomics can be primarily divided into two distinct subsets. One form of mass spectrometry-based proteomics is top-down proteomics, the analysis of intact proteins[11,12]. This mass spectrometry method bypasses the need for peptide-level quantitation as each precursor correlates to an individual protein. However, ionization of intact proteins is difficult and the subsequent identification from MS/MS analysis can be complicated due to high charge states and increasingly complex spectra[13]. Bottom-up (a.k.a. shotgun) proteomics focuses on the analysis of peptides from proteolytic digestion of a protein sample. The field of shotgun proteomics is well developed in comparison to its sibling, top-down, as tryptic peptides (peptides resulting from digestion with the enzyme trypsin) are highly soluble and readily ionizable making them amenable to liquid chromatography MS/MS (LC-

MS/MS)[13]. Further mentions of MS/MS and LC-MS/MS will refer to the bottom-up method unless otherwise specified.

A Brief History of Quantitative Proteomics

The identification of proteins is the core component of any of proteomics experiment; yet qualitative analysis alone is insufficient to fully describe biological phenomena. To that end, numerous methods employing various techniques have been utilized to quantify relative and absolute protein changes since the early days of proteomics (Table 1-1). Early quantitative methods involved the comparison of spot or band intensities from 2DE or Western blotting experiments but are plagued with issues of accuracy, precision, and throughput[14,15]. While 2DE analysis provides depth through the sheer number of proteins observed on a gel, 2DE data lacked specificity and was easily confounded by co-migrating proteins. Moreover, without *a priori* knowledge or subsequent identification of isolated spots, the biological significance derived from 2DE experiments alone remains limited due to a decoupling of quantitation and identification[16]. To combat some of these drawbacks, antibody-based methods, like the Western blotting and ELISA, were developed. These antibody-based methods are highly specific and quantitative but involve labor-intensive workflows that limit the number of comparisons the experiment can perform[15,17]. Additionally, by nature of using antibodies, these methods are limited by the affinity and specificity of the primary and secondary antibodies used in the experiment affecting limit of detection and depth. The targeted approach of antibody-based methods provided a solution to the co-migration problem observed in 2DE at a direct cost to depth while still requiring knowledge of the proteins of interest.

MS/MS-based methods by comparison, provide a solution to the specificity problem by combining both the sequencing and quantitation elements of an experiment. Early MS/MS-based

Table 1-1: Brief comparison of common quantitation methods used in proteomics.

Method	Primary Technique	Specificity	Depth	Accuracy	Precision	Throughput
2D-electrophoresis	Gel Staining	None	Medium/High	Low	Low	Medium/Low
Western blot	Antibody based	High	Low	Medium/Low	Medium/Low	Low
ELISA	Antibody based	High	Low	Medium/High	Medium/High	Low
Spectral Counting	MS/MS	High	High	Medium/Low	Low	High
Label-free MS1-based*	MS/MS	High	High	Medium/Low	Medium/Low	High
SILAC	MS/MS	High	High	High	High	Low
Absolute Quant	MS/MS	High	Medium/Low	High	High	Low
Isobaric tagging	MS/MS	High	High	High	High	Medium/High

Dashed line segregates MSMS-based methods from earlier methods
*Label-free MS1-based quantitation methods include those methods utilizing Area Under the Curve and Intensity-based analysis

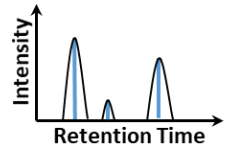
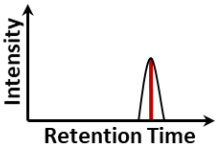
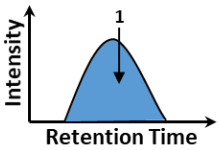
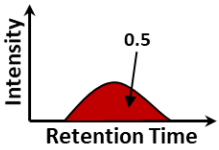
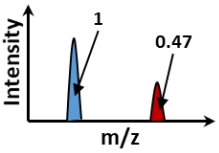
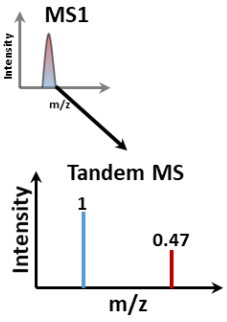
quantitation methods leaned heavily on ideas from the field of stable-isotope dilution theory resulting in methods that were both highly accurate and precise[9,10,18,19]. By leveraging stable isotopes like ^{13}C, ^{15}N, and ^{2}H biological samples could be labelled pre-harvest through metabolic labelling (e.g. Stable Isotope Labeling by/with Amino acids in Cell culture or Mammals – SILAC or SILAM), post-harvest through chemical labelling (e.g. Tandem Mass Tags - TMT), or compared to synthetic peptides containing heavy isotopes[10,20–23]. While highly quantitative, these methods introduced a labelling step that could increase the complexity and cost of the experiment. As such, many research groups have attempted to find solutions that would circumvent the need to utilize stable isotopes. Through observations of empirical data, correlations were drawn between the number of peptide-spectra matches (PSMs) and the relative abundance of a given protein[24]. Other researchers noticed that there was a relationship between the peak areas and relative abundance of a peptide thus creating the idea of label-free quantitation[25–28]. These two methods simplified the sample preparation process, thus increasing the number of MS/MS analyses performed; but neither method could consistently account for the stochasticity inherently observed in MS/MS experiments, thus reducing the accuracy and precision when compared to methods employing stable isotope[29,30].

Comparing Quantitative MS/MS Methods

As mass spectrometry instrumentation and technology has improved over the previous decades, so too have the quantitation methods. Despite various acquisition techniques, at the core there are only four unique families of quantitation methods based on what measurements are used and how the quantitation is performed. Briefly, these methods are spectral counting, label-free intensity-based, isotope-based quantitation, and isobaric tag-based quantitation (Figure 1-1).

Figure 1-1: Primary means of quantitation in an LC-MS/MS proteomics experiment. Fundamentally, only 4 distinct quantitation methods exist depending on what the numeric value represents and how many comparisons are made simultaneously. For example, spectral counting and label-free intensity-based quantitation methods both quantify only one condition per LC-MS/MS run. However, the numeric value in the former represents the number of PSMs attributed to a peptide while in the latter it represents the ions present in an LC-MS/MS scan.

Figure 1-1 (continued)

Quantitation Method	First LC-MS/MS Experiment	Second LC-MS/MS Experiment	Quantitation
Spectral Counting	 PSMs found in Condition 1 analysis for given precursor	 PSMs found in Condition 2 analysis for given precursor	Compare count of PSMs for a protein across LC-MS/MS experiments e.g. 3:1 ratio
Label-free Intensity-based Quantitation (e.g. AUC, intensity)	 AUC in Condition 1 analysis for given precursor	 AUC in Condition 2 analysis for given precursor	Compare AUC of precursor masses across LC-MS/MS experiments e.g. 2:1 ratio
Labelled Intensity-based Quantitation (e.g. SILAC, AQUA)	 AUC in Condition 1 analysis for given precursor AUC in Condition 2 analysis for given precursor	No second run	Compare area under the curve for precursor masses within the same LC-MS/MS experiment e.g. 2.1:1 ratio
Isobaric Label Quantitation (e.g. TMT, iTRAQ)	 S:N value for reporter ion representing Condition 1 S:N value for reporter ion representing Condition 2	No second run	Compare S:N ratios for reporter ions in tandem MS analysis of precursor from the same MS1 within the same LC-MS/MS experiment e.g. 2.1:1 ratio

Spectral counting quantitation method

One of the simplest mass spectrometry-based quantitation methods is that of spectral counting which was based on researchers' observations that the number of PSMs attributed to a given peptide was often related to the relative abundance of its protein across multiple samples analyzed by LC-MS/MS[24,29]. Spectral counting is a label-free quantitation method that attains quantitative information for a peptide by its sampling rate in the mass spectrometry experiment (i.e. how many PSMs were found for each peptide). Protein level quantitation is further inferred by the extent of coverage (i.e. how much of a protein was identified by the peptides in the experiment). This information can be fed into a statistical model to quantify peptides and proteins in a complex mixture[2,24,31].

By focusing on the frequency of identification as the primary measurement for quantification, spectral counting buffers against issues of variable observed intensity but is still susceptible to stochasticity problems (Figure 1-2). However, as a simple method based on empirical observation it has proven to be effective at accurate enough to measure relative abundances of proteins in a yeast sample and even mapping the human interactome[2,24].

Label-free intensity-based quantitation methods

Another common family of quantitation methods is label-free intensity based quantitation. These methods utilize intensity information (e.g. S:N, AUC, peak intensity) to generate quantitative values for peptides observed in an MS/MS analysis. Quantitative information can be obtained from data-dependent acquisition (DDA) MS1 scans or from MS2 scans from selected-reaction-monitoring (SRM or MRM) and parallel-reaction monitoring (PRM) acquisition methods. However, as no distinguishing features (e.g. isotope labels) are

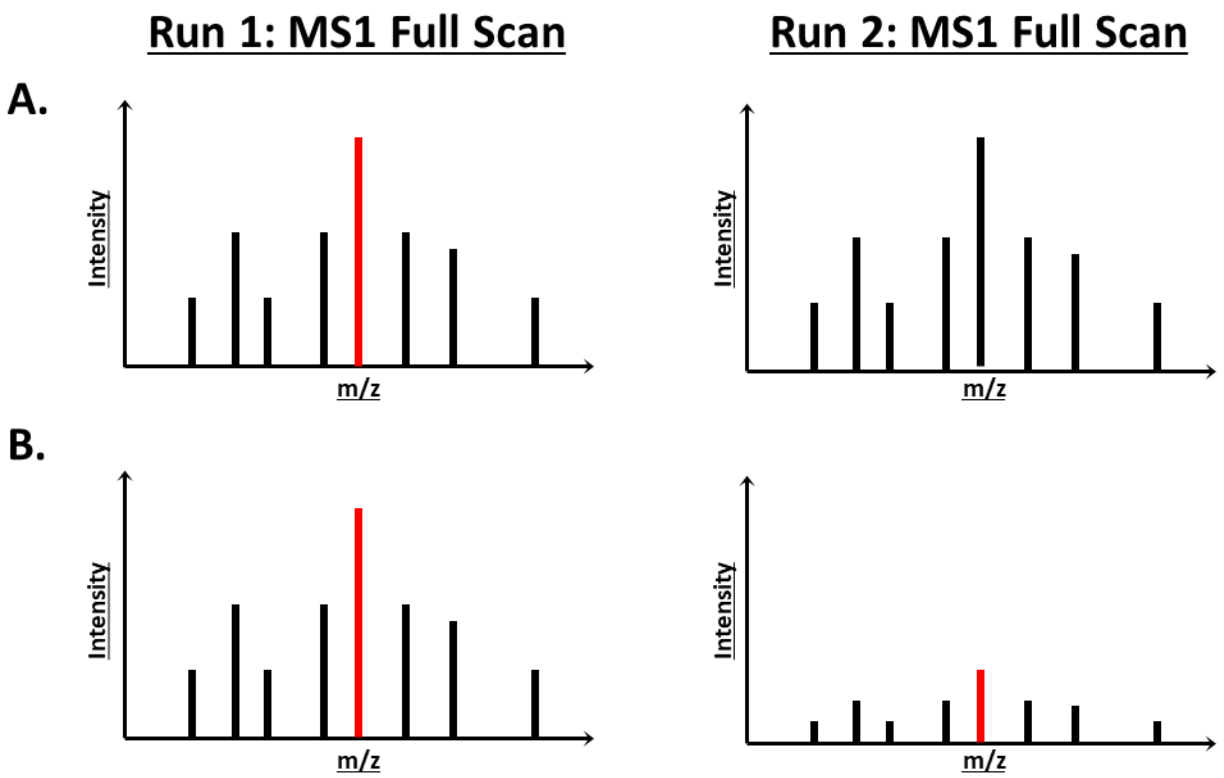


Figure 1-2: Two common pitfalls with label-free based quantitation methods. **A)** Stochasticity, or the random identification (or lack thereof) of any ion species with the mass spectrometer. This problem can result in the an ion species be selected (red) for identification and quantitation in one LC-MS/MS run while being ignored in a second run. The quantitation of the selected ion species becomes impossible as lack of evidence does not imply absence. **B)** Variable observed intensity is common between independent LC-MS/MS runs as ionization efficiency and detection are a function of numerous variables, some known and some unknown. For example, temperature, humidity, and air pressure can affect how well a peptide will ionize; slight variations can result in observed intensity differences. As such direct quantitation of the selected species (red) is impossible across LC-MS/MS runs and some form of normalization must be performed. No tandem MS occurs during MS1; however, during label-free quantitation, quantitative information is provided by data from the MS1 scan.

used, only one condition can be analyzed per MS/MS run. This poses two significant problems involving variable observed intensity and the stochastic identification of peptide (Figure 1-2).

Variable observed intensity means that comparisons across independent LC-MS/MS runs may not be accurate. This is because the observed intensity in an independent LC-MS/MS analysis is a function of many external variables, such as temperature, that affect ionization both directly and indirectly[32,33]. As such, direct comparisons between raw collected data have limited value. Labelled approaches circumvent this problem by treating one label as an internal standard to which a quantitative comparison can be made; however, no reliable internal standard is provided in a label-free experiment[29]. To deal with the variable observed intensities of independent LC-MS/MS experiments, normalization schemes have been employed by various groups for all label-free intensity-based quantitation methods[27,28,34].

Stochastic identification of peptides is, perhaps, the leading cause of missing values in label-free analysis. This effect appears to be intensity dependent as peptides with the largest intensity values are usually observed in multiple runs while low abundant peptides may be identified sporadically across multiple MS/MS analyses[30]. As such, there have been several attempts to address this issue ranging from imputation to identification transfer[27,28,30]. Fundamentally, imputation techniques can be problematic as the assignment of an intensity value to a missing value implies knowledge of something unknown. Fixed standard deviations and vastly incorrect imputed values are among the common problems when imputing a missing value. Alternatively, researchers have attempted to transfer identifications from one MS/MS run to another utilizing chromatographic data[27,28,35]. This technique relies heavily on the quality of liquid chromatography (LC) and is susceptible to variations in elution order and the consistency of LC setup (e.g. column, gradient, and pumps)[36]. In short, this method aligns multiple LC-

MS/MS runs, allowing shifts of up to 20 minutes, and then scans each run for unidentified features in the MS1. If an identified feature is found matching an unidentified feature's precursor mass within a user defined elution window, that identification is transferred to the unidentified feature. While powerful, this method can generate false matches which are impossible to detect in a standard experiment[35–37]. However, one of the most popular ways to handle stochastic identification of peptides is to utilize the label-free quantitation (LFQ) algorithm within the MaxQuant software suite. This method ignores the issue of stochasticity by requiring a peptide to have been identified in the compared LC-MS/MS analyses[27]. As such, peptides identified in only one LC-MS/MS run are not suitable for quantitation and are ignored.

Ultimately, label-free intensity-based quantitation methods are powerful as they provide a simple way to perform quantitative proteomics on many LC-MS/MS analyses without having to rely on isotope labels. This can reduce errors from sample handling and reaction efficiencies, as well as the costs for reagents. These methods are not without their flaws and there are research groups dedicated to advancing the technologies and methodologies surrounding label-free intensity-based quantitation.

Isotope-based quantitation

One of the earliest methods to quantify changes observed in a mass spectrometer was to utilize stable-isotope dilution theory[8]. These isotope-based methods allowed for the simultaneous quantitation of two samples utilizing intensity information from an MS1 scan and identification information from a subsequent tandem MS scan. Quantitation is performed by comparing the signal-to-noise (S:N), peak intensity, or the integrated area under the intensity curve (AUC) values between the peak containing ions labelled with heavy isotopes and the peak containing ions without heavy labels – respectively called the heavy and light peaks. The

simultaneous analysis of two samples minimized the effect of run-to-run variation often seen as stochasticity or variations in observed intensity by allowing one sample, or channel, to serve as an internal standard (Figure 1-2)[1,9,10,30]. These isotope labelling methods leverage the quality of isobaric elution to ensure accurate comparisons are made; that is, peptides containing a heavy label elute simultaneously with their light labelled counterparts. Of these isotope-based quantitation methods, there are three major sub categories: metabolic labelling methods, chemical labelling methods, and spike-in methods[29].

Metabolic labelling methods such as SILAC are some of the most commonly used quantitative mass spectrometry techniques and are lauded for their accuracy and precision[10]. However, as these methods necessitate the incorporation of heavy amino acids throughout the entire proteome. It can be time consuming, costly, or even impossible to generate labelled samples for MS/MS analysis. Furthermore, the number of labels in traditional metabolic labelling experiments is generally limited to $^{13}C_6$-arginine and $^{13}C_6$-lysine meaning only 3 conditions can be analyzed simultaneously at maximum[29].

This limitation can be overcome by novel approaches such as NeuCode (neutron-encoded) lysine which leverages the high-resolution capabilities of modern Fourier-transform mass spectrometers, instruments that detect and measure ion currents in the time domain (i.e. frequencies) before converting these values to mass measurements through the use of Fourier transforms, to detect small mass defects in isotopologues[38–41]. Although expensive, NeuCode lysine can extend a SILAC experiment from a 3-plex to a 9-plex through the use of commercially available reagents. Furthermore, combining this metabolic labelling technique with a chemical labelling strategy known as mTRAQ (the non-isobaric form of Isobaric Tags for Relative and Absolute Quantitation, iTRAQ) allows for the creation of an 18-plex[42,43].

Chemically labelling cell lysate or digest with a heavy and light chemical tag is an alternative solution to metabolic labelling. These methods include ICAT, dimethyl labelling, and the incorporation of ^{18}O during protein digestion as well the non-isobaric forms of isobaric tags like mTRAQ and mTMT[9,42,44–46]. Functionally, these methods target reactive regions of a peptide such that a heavy and light sample can be created post-harvest. By breaking the dependency of metabolic incorporation, these methods can be more universal; however, labelling and incorporation efficiencies become more of a concern as not every peptide will receive the modification. As with the metabolic labelling strategy, these methods are often limited to a small number of comparisons.

Lastly, another isotope-based quantitation is to spike-in known amounts of a synthetic peptide that was made with heavy amino acids or a cell lysates metabolically labelled with heavy amino acids. Unlike the other methods mentioned here, the use of heavy spike-ins is not primarily meant for comparisons across conditions but to generate an absolute measurement by calculating a relative abundance to a known amount. Absolute quantitation (AQUA) or proteins is an example of a spike-in technique which has been shown to quantify low abundance phosphorylation events with ease[21,47].

While these methods are accurate, they are limited to the number of simultaneous comparisons that can be made, limiting the number of replicates or conditions that can be analyzed in a single MS/MS run. Furthermore, the use of chemical labels such as ICAT and dimethyl labelling has been supplanted by isobaric tag-based quantitation methods, which will be discussed later. However, due to the high accuracy and sensitivity of SILAC and absolute quantitation methods, these methods are still readily employed in a variety of current proteomic experiments.

Isobaric tag-based quantitation methods

Like labeled intensity-based quantitation, isobaric-tag based quantitation methods rely on the use of stable isotopes to facilitate multiple simultaneous comparisons of samples or conditions in a single LC-MS/MS analysis. Currently, there are two main types of isobaric-tags commercially available, TMT and iTRAQ. These two types of tags are chemically distinct but utilize the same principals to quantify peptides. Isobaric tag-based quantitation relies on the use of chemical tags that are covalently added to a sample after proteolytic digest[23,48]. Presently, iTRAQ is commercially available in 2-, 4-, or 8-plex from Sigma-Aldrich/SCIEX while TMT is available as a 2-, 6-, 10-, or 11-plex from ThermoFisher Scientific. A 16-plex version of TMT, TMTpro, was recently developed by ThermoFisher with the reporter ion based on a proline structure. However, due to its novelty, it has yet to reach mainstream use. Other isobaric tags such as the DiLeu tag and Combinatorial Isobaric Mass Tags have been produced in academic labs but have not been commercially adopted[49–51]. It is likely that the ability to multiplex will increase beyond 16 in the near future as the research of isobaric tag-based technologies continues.

An isobaric tag can be divided into three distinct regions, a reporter-ion fragment, a mass balancer, and a reactive group (most typically an amine-reactive group such as an N-hydroxysuccinimide ester). Stable isotopes, such as ^{13}C and ^{15}N, can be distributed across the chemical bond that separates the reporter-ion fragment from the mass balancer resulting in each tag having a different reporter-ion mass while the sum of the reporter-ion fragment and the mass balancer masses remains unchanged. As such, each sample that will be part of the multiplexed experiment will receive a tag with a unique reporter-ion mass before being mixed with the other samples (Figure 1-3). Since the overall mass and chemical properties of the tags are identical,

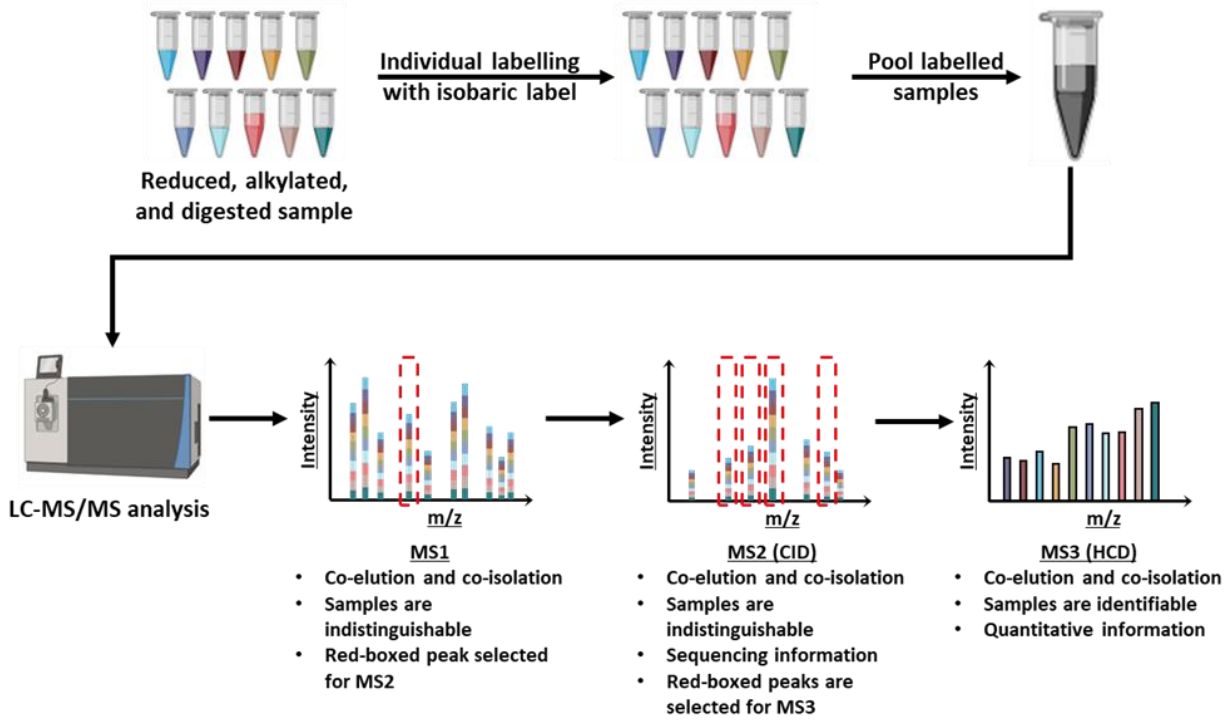


Figure 1-3: Example generic workflow of an isobaric tag-based LC-MS/MS experiment. Depicted workflow starts from samples that were reduced, alkylated, and proteolytically digested. Each sample receives a unique isobaric tag before being combined into one tube. The pooled sample is then analyzed by LC-MS/MS. Peptides from all samples co-elute across the LC gradient and are indistinguishable at the MS1 and MS2 (if collision-induced dissociation, CID, is used). In an MS3-based experiment, as shown here, quantitation occurs after HCD fragmentation. Only after HCD fragmentation can the relative contribution of each sample be identified for a given peptide ion.

peptides will co-elute and co-isolate in the mass spectrometer resulting in the simultaneous LC-MS/MS analysis of any given peptide across all samples in the multiplexed experiment. However, quantitation is not performed directly on peptide ions but on the reporter-ion fragments that are fragmented off the peptide during higher-energy collisional dissociation (HCD). The relative abundances of intensity values of reporter ion fragments are analyzed in a low-mass region (126-131 m/z for TMT and 113-121 m/z for iTRAQ) and are correlated to the relative abundances of the peptide within each sample.

Because HCD fragmentation can be used to sequence and identify peptides, it is possible to perform quantitation and identification within a single HCD MS2 scan[52,53]. This approach is popular due to its ease of use and large number of identifications. However, MS2 level quantitation of reporter ions in isobaric tag-based experiments is prone to interference from co-eluting and co-isolating peptides (Figure 1-4A). As such, fragment ions that do not belong to the peptide of interest can contribute to the reporter ion signal intensities leading to ratio compression – the trend where relative abundances approach a 1:1 ratio[54]. Ratio compression directly affects the assays ability to accurately measure fold changes while misleadingly increasing precision; as more ratios trend towards 1:1 the coefficient of variation between measurements decreases.

To address the concerns about quantitative accuracy of isobaric tag-based methods, several research groups developed methods involving gas-phase purification techniques to reduce the complexity of MS spectra and improve the overall accuracy and precision of isobaric tag-based quantitation[54,55]. While methods that manipulated the charge-state of ions in the gas-phase through proton-transfer reactions (PTR) proved to be quite effective at reducing interference, the ability to conduct an additional MS scan (MS3) proved to be a simple, easy, and

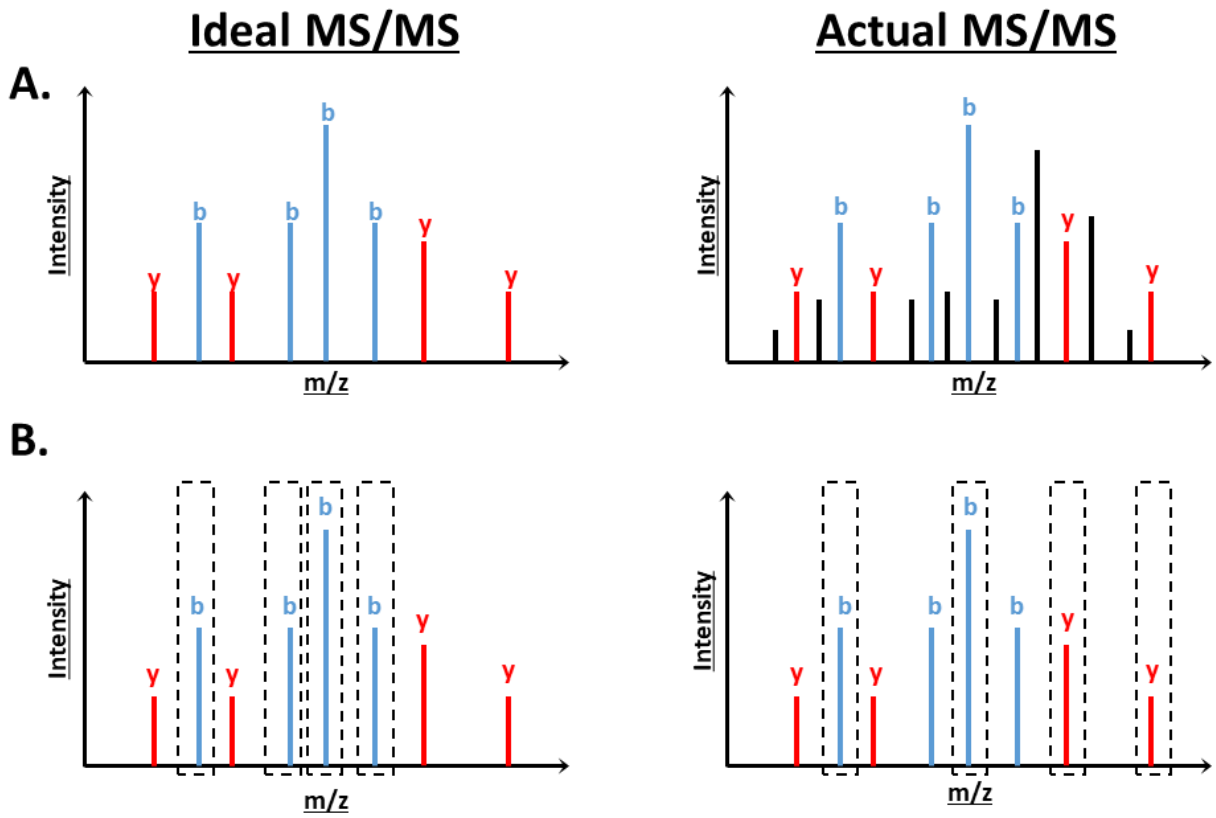


Figure 1-4: Two common pitfalls with isobaric tag-based quantitation methods. **A)** Interference from co-eluting and co-isolated peaks. Ideally, the only fragment ions present in the MS/MS spectra would be the *b*- and *y-type* ions (blue and red respectively) from the selected precursor in the MS1. However, the isolation of a precursor in the MS1 is rarely completely pure and unwanted ions are selected for due to reasons including co-elution and co-isolation. Fragment ions of these unwanted selections can be found in the MS/MS (black) and can contribute reporter ions during quantitation resulting in ratio compression. **B)** Selection of fragment ions for SPS-MS3 that do not contain an isobaric tag (ions selected for quantitation boxed in dashed lines). Because amine reactive isobaric tags the most common, the N-termini of peptides are guaranteed to contribute a reporter ion. This means that b-ions are the most desirable for quantitation. While some tryptic peptides can end lysine, not all will. As such selecting y-ions can result in insufficient reporter ion signal for quantitation.

affordable way for users to improve the accuracy and precision of their isobaric tag-based experiments without the need of extra reagents.

Both the MS3 and PTR methods reduced interference increasing the accuracy of isobaric tag quantitation methods at the cost of sensitivity. By limiting the number of ions to increase purity these methods suffered from reduced signal intensity resulting in diminished sensitivity. This problem was initially addressed for the MS3 method by selecting multiple peaks from the MS2 spectra, also known as synchronous precursor selection (SPS), for follow up MS3 analysis; the combination of SPS and MS3 level analysis (SPS-MS3) was enabled using an isolation waveform with multiple frequency notches[56]. By increasing the number of MS2 precursors selected for MS3 analysis, the technique increased the likelihood of selecting precursors that were labelled with an isobaric tag. The dramatic improvement in sensitivity, accuracy, and precision the SPS-MS3 method provided led it to become the de facto protocol for those concerned about accuracy and precision above all else[57].

While both *b*- and *y-type* ions are useful for sequencing a peptide, only b-ions are guaranteed to contain an isobaric label. This is because most isobaric labels use amine reactive chemistry resulting in the labelling of the N-terminus and all lysine residues present in a peptide. Furthermore, as trypsin is the proteolytic workhorse for mass spectrometry experiments only 50% of peptides will contain a lysine at their C-terminus (assuming no missed cleavages and equal abundance of lysine and arginine)[58]. This means that about 50% of the time, the selection of a y-ion during SPS-MS3 would result in no additional reporter-ion signal during the MS3 analysis – assuming the absence of interfering ions (Figure 1-4B). As such, the SPS-MS3 technique was still lacking as there is a chance of selecting a y-ion fragment in the MS2. Currently, there exists two techniques that have been shown to directly address this issue: the

targeted method known as TOMAHAQ and real-time database searching[59,60]. Both technologies, at their core, rely on on-the-fly decision making to increase the likelihood of selecting daughter ions in the MS2 that will contain reporter ions when fragmented for MS3 analysis. However, TOMAHAQ and real-time database searching are bleeding-edge methods are only starting to see mainstream usage.

The importance of peptide level quantitation

Fundamentally, LC-MS/MS approaches in proteomics measure ionized peptides in a sample, not proteins. Because of this distinction, protein-inference (i.e. the organization of peptides into a list or probable proteins) can become quite challenging, especially when various proteoforms are involved. As such, various methods from the application of Occam's razor to Bayesian modelling to determine how peptides should be attributed to proteins[61]. While all methods can sufficiently characterize the proteome, noticeable levels of variation exist between the peptide-to-protein assignments. Therefore, as protein identification precedes quantification in mass spectrometry experiments, a corollary is that quantification of protein abundances from peptide level data can be equally complicated. In most LC-MS/MS proteomic experiments, proteins are quantified by a summary statistic of the peptide abundance data. Statistical analysis is then performed on these summary values. As such, these methods often ignore information like measurement uncertainty and proteoform differences as they attempt to quantify global proteomic changes.

Peptide level quantitation incorporates uncertainty

A simple way to quantify protein measurements is to take the mean, median, or weighted average of all peptides or, in the case of isobaric tag-based methods, reporter-ion intensity-based values attributed to a protein[27,62–65]. Protein quantitation is then performed using these summary

statistics as surrogates for relative protein abundances. As a result, the uncertainty behind these measurements are abstracted away from users. Other methods like CompMS and SCAMPI rely on complex mathematical models to better account for the physical manner by which the LC-MS/MS data is acquired[64,66]. For the price of computational intensity, these methods can provide gains in accuracy and precision – although these improvements are often marginal or useful only in fringe situations such as the accurate quantitation of small fold changes.

While these complex methods may minimally affect accuracy and precision, an additional benefit is that they provide error estimates of the protein abundance measurements which can allow users to better understand the accuracy and precision of their data. Furthermore, by performing statistical analysis at the peptide level, these algorithms can leverage replicate information from multiply sequenced peptides and proteins with multiply sequenced unique peptides thus increasing confidence in the final protein abundance estimate. Simpler methods that utilize peptide level data have been employed in the label-free community showing a desire for a middle ground approach[27].

Peptide level quantitation is often necessary to handle multiple proteoforms

One important piece of information lost by summarizing peptide information into a unifying protein value pertains to proteoforms. The term "proteoform" includes splice variants, isoforms, and post-translational modifications like phosphorylation, acetylation, and ubiquitination[67]. Mass spectrometry is an excellent tool to identify these different proteoforms but quantitation becomes problematic as peptide assignments can often belong to several proteoforms[61]. While some isoforms and post-translational modifications may not have functional roles within a biological system, some can have real biological significance.

For example, the protein Dok-1 exists in human cells in a full-length and a truncated form with each form localizing to different locations, the cytosol and perinuclear region respectively[68]. However, all 15 tryptic peptides found in the truncated form of Dok-1 exist in the full-length isoform while another 8 peptides, 7 wholly within the full length and 1 spanning the truncation, can be found only within the full-length protein (Figure 1-5). In this instance, to quantitatively understand expression levels of both the full-length and truncated form of Dok-1, peptide level analysis must be employed. This is especially true if the absolute abundance levels of the full-length protein dwarf the truncated form.

Another example of a proteoform that must be quantified at the peptide level is phosphorylation as a population can exist as a mixture of phosphorylated and unphosphorylated forms. Phosphorylation occurs mainly on serine, threonine, and tyrosine residues and is an important post-translational modification used in a variety of cellular processes, such as signaling[69]. However, as phosphorylation occurs at the site level, the two populations of protein sequences are indistinguishable except for the phosphorylation event. As such, most peptides are shared between phosphorylated and unphosphorylated protein species; however, unlike the isoform problem, quantifying the modified proteoform involves isolating and analyzing the phosphorylated peptide in an LC-MS/MS experiment and no peptide reassignments need to occur. Furthermore, since the function of a phosphorylation event is often intimately tied to its localization to a specific site, biology dictates that phosphoproteomic analysis focus on the site and peptide level rather than the overall protein expression. This is further emphasized as a single protein can have multiple phosphorylation sites corresponding to several different functions.

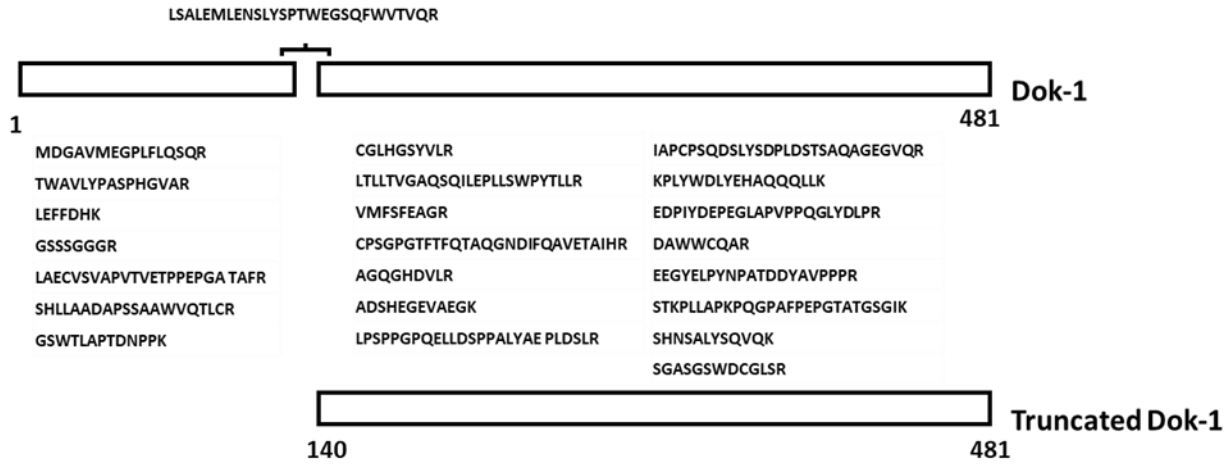


Figure 1-5: Schematic of the protein Dok-1 and its truncated isoform. Seven unique tryptic peptides are found solely within the N-termini portion of Dok-1 while a single unique tryptic peptide spans the truncation. Fifteen unique tryptic peptides are shared between Dok-1 and its truncated isoform. Protein level quantitation of both isoforms will inadvertently be influenced by the other. Some methods, such as TOPn, will ignore the truncation isoform as it consists only of shared peptides while other methodologies will attempt to attribute quantitation values to both Dok-1 and the truncated isoform. Peptide-level analysis can be used to either deconvolute the intensity values by leveraging the peptides unique to the full-length Dok-1 or to indicate error and ambiguity in the measurement. In silico digestion of Dok-1 was performed assuming no missed cleavages.

The quantitative characterization of phosphopeptides and phosphorylation sites is often enough to elucidate biological phenomena; however, sometimes understanding the peptide level changes with respect to the overall protein is important[70–72]. This phenomenon, known as phosphorylation occupancy or stoichiometry, is a relationship between a phosphorylated peptide and non-phosphorylated peptide abundance and can have important implications in biology such as in the case of the protein separase[21,73]. In these instances, understanding the relative change in abundance of a phosphorylation site is insufficient to grasp the whole biological picture as multiple occupancy states can explain a relative change (Figure 1-6). While low-throughput or targeted LC-MS/MS strategies are often employed to calculate phosphorylation stoichiometry, large-scale analysis is quite challenging as the core of the work revolves around the peptide-level analysis of rare peptide species.

Direct methods to measure phosphorylation stoichiometry rely on measuring both the phosphorylated and unphosphorylated peptide species with respect to the protein abundance[72]. This method of measurement is performed over the course of several LC-MS/MS experiments as a phosphopeptide enriched sample must be analyzed in addition to a non-enriched sample as phosphoproteins are rarely detected without enrichment[69,70]. Alternatively, phosphorylation stoichiometry can be measured indirectly in one experiment by phosphatase treating half the sample and utilizing an isotope labelling strategy[71]. While the first type of method is susceptible to the same issues of stochasticity as a label-free experiment, both types of methods are prone to generating impossible stoichiometry values (i.e. greater than 100% and less than 0%). Better instrumentation and peptide detection technology to and improved peptide-level quantitation can help increase the reliability of this inherently peptide-level measurement of a biological phenomenon.

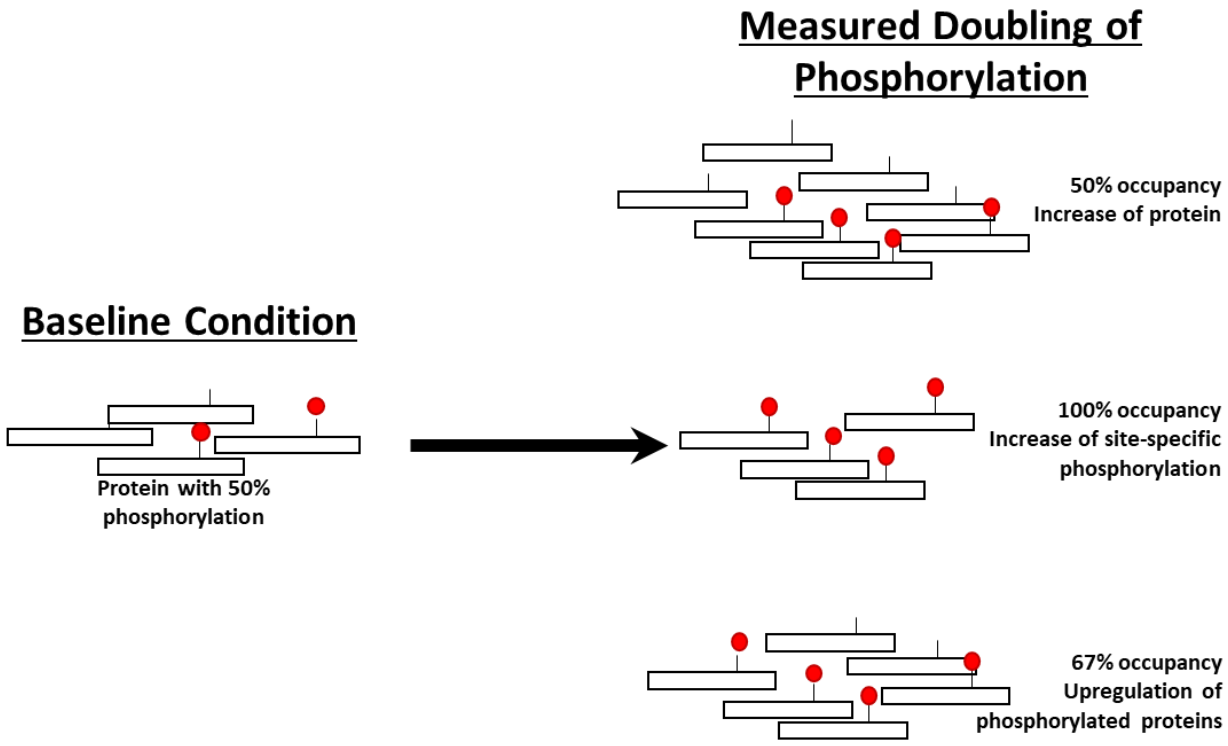


Figure 1-6: Doubling of phosphopeptide abundance can result from any number of different biological states. In all cases a doubling of phosphorylation is measured. However, a doubling could imply upregulation of protein with no change to its occupancy level. Alternatively, this doubling could be attributed to the increase of phosphorylation stoichiometry. Yet another alternative could be the selective upregulation of phosphorylated proteins by stabilization or selective degradation of the unphosphorylated form. Simply measuring the relative change is insufficient to grasp the complete biology in these instances.

References

1. Graves, P. R. & Haystead, T. A. J. Molecular Biologist's Guide to Proteomics. *Microbiol. Mol. Biol. Rev. Microbiol. Mol. Biol. Rev.* **66**, 39–63 (2002).

2. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).

3. Anderson, N. G. & Anderson, N. L. Twenty years of two-dimensional electrophoresis: Past, present and future. *Electrophoresis* **17**, 443–453 (1996).

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

5. Gygi, S. P. & Aebersold, R. Mass spectrometry and proteomics. *Curr. Opin. Chem. Biol.* **4**, 489–494 (2000).

6. Eng, J. K., Mccormack, A. L. & Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

7. Sadygov, R. G., Cociorva, D. & Yates, J. R. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat. Methods* **1**, 195–202 (2004).

8. de Leenheer, A. P. & Thienpont, L. M. Applications of isotope dilution-mass spectrometry in clinical chemistry, pharmacokinetics, and toxicology. *Mass Spectrom. Rev.* **11**, 249–307 (1992).

9. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).

10. Ong, S. E. *et al.* 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 (2002).

11. Toby, T. K., Fornelli, L. & Kelleher, N. L. Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annu. Rev. Anal. Chem.* **9**, 499–519 (2016).

12. Chen, B., Brown, K. A., Lin, Z. & Ge, Y. Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.* **90**, 110–127 (2018).

13. Chait, B. T. Mass Spectrometry: Bottom-Up or Top-Down? *Science (80-.).* **314**, 65–66 (2006).

14. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-

dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 9390–9395 (2000).

15. Ghosh, R., Gilda, J. E. & Gomes, A. V. The necessity of and strategies for improving confidence in the accuracy of western blots. *Expert Rev. Proteomics* **11**, 549–560 (2014).

16. Fullaondo, A., Vicario, A., Aguirre, A., Barrena, I. & Salazar, A. Quantitative analysis of two-dimensional gel electrophoresis protein patterns: A method for studying genetic relationships among Globodera pallida populations. *Heredity (Edinb).* **87**, 266–272 (2001).

17. Sakamoto, S. *et al.* Enzyme-linked immunosorbent assay for the quantitative/qualitative analysis of plant secondary metabolites. *J. Nat. Med.* **72**, 32–42 (2018).

18. Paša-Tolic, L. *et al.* High throughput proteome-wide precision measurements of protein expression using mass spectrometry [10]. *J. Am. Chem. Soc.* **121**, 7949–7950 (1999).

19. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 6591–6596 (1999).

20. Ong, S. E. & Mann, M. Mass Spectrometry–Based Proteomics Turns Quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).

21. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6940–5 (2003).

22. Thompson, A. *et al.* Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).

23. Ross, P. L. *et al.* Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).

24. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).

25. Bondarenko, P. V., Chelius, D. & Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography - Tandem mass spectrometry. *Anal. Chem.* **74**, 4741–4749 (2002).

26. Chelius, D. & Bondarenko, P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* **1**, 317–323 (2002).

27. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

28. Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11**, 1–11 (2012).

29. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).

30. O'Connell, J. D., Paulo, J. A., O'Brien, J. J. & Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* **17**, 1934–1942 (2018).

31. Sowa, M. E., Bennett, E. J., Gygi, S. P. & Harper, J. W. Defining the Human Deubiquitinating Enzyme Interaction Landscape. *Cell* **138**, 389–403 (2009).

32. Banerjee, S. & Mazumdar, S. Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *Int. J. Anal. Chem.* **2012**, 1–40 (2012).

33. Blackler, A. R., Speers, A. E. & Wu, C. C. Chromatographic benefits of elevated temperature for the proteomic analysis of membrane proteins. *Proteomics* **8**, 3956–3964 (2008).

34. MacLean, B. *et al.* Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

35. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, 1–16 (2019).

36. Bielow, C., Mastrobuoni, G. & Kempa, S. Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **15**, 777–787 (2016).

37. Lim, M. Y., Paulo, J. A. & Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J. Proteome Res.* **18**, 4020–4026 (2019).

38. Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass Spectrom.* **47**, 226–236 (2012).

39. Hebert, A. S. *et al.* Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **10**, 332–334 (2013).

40. Overmyer, K. A. *et al.* Multiplexed proteome analysis with neutron-encoded stable isotope labeling in cells and mice. *Nat. Protoc.* **13**, 293–306 (2018).

41. Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **10**, 1–19 (2011).

42. DeSouza, L. V., Romaschin, A. D., Colgan, T. J. & Siu, K. W. M. Absolute quantification of potential cancer markers in clinical tissue homogenates using multiple reaction monitoring on a hybrid triple quadrupole/linear ion trap tandem mass spectrometer. *Anal. Chem.* **81**, 3462–3470 (2009).

43. Merrill, A. E. *et al.* NeuCode labels for relative protein quantification. *Mol. Cell. Proteomics* **13**, 2503–2512 (2014).

44. Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S. & Heck, A. J. R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **4**, 484–494 (2009).

45. Yao, X., Freas, A., Ramirez, J., Demirev, P. A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: Model studies with two serotypes of adenovirus. *Anal. Chem.* **73**, 2836–2842 (2001).

46. Paulo, J. A. & Gygi, S. P. mTMT: An Alternative, Nonisobaric, Tandem Mass Tag Allowing for Precursor-Based Quantification. *Anal. Chem.* **91**, 12167–12172 (2019).

47. Kettenbach, A. N., Rush, J. & Gerber, S. A. Absolute quantification of protein and post-translational modification abundance with stable isotope-labeled synthetic peptides. *Nat. Protoc.* **6**, 175–186 (2011).

48. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).

49. Xiang, F., Ye, H., Chen, R., Fu, Q. & Li, N. N,N-Dimethyl leucines as novel Isobaric tandem mass tags for quantitative proteomics and peptidomics. *Anal. Chem.* **82**, 2817–2825 (2010).

50. Frost, D. C., Greer, T. & Li, L. High-resolution enabled 12-plex DiLeu isobaric tags for quantitative proteomics. *Anal. Chem.* **87**, 1646–1654 (2015).

51. Braun, C. R. *et al.* Generation of Multiple Reporter Ions from a Single Isobaric Reagent Increases Multiplexing Capacity for Quantitative Proteomics. *Anal. Chem.* **87**, 9855–9863 (2015).

52. Jedrychowski, M. P. *et al.* Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol. Cell. Proteomics* **10**, 1–9 (2011).

53. Zecha, J. *et al.* TMT labeling for the masses: A robust and cost-efficient, in-solution labeling approach. *Mol. Cell. Proteomics* **18**, 1468–1478 (2019).
54. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

55. Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8**, 933–935 (2011).

56. Mcalister, G. C. *et al.* MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes Graeme C. McAlister, 1 David P. Nusinow, 1. *Anal. Chem.* **86**, 7150–7158 (2014).

57. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

58. Kozlowski, L. P. Proteome-pI: Proteome isoelectric point database. *Nucleic Acids Res.* **45**, D1112–D1116 (2017).

59. Erickson, B. K. *et al.* A Strategy to Combine Sample Multiplexing with Targeted Proteomics Assays for High-Throughput Protein Signature Characterization. *Mol. Cell* **65**, 361–370 (2017).

60. Erickson, B. K. *et al.* Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).

61. Huang, T., Wang, J., Yu, W. & He, Z. Protein inference: A review. *Brief. Bioinform.* **13**, 586–614 (2012).

62. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: The protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).

63. Boehm, A. M., Pütz, S., Altenhöfer, D., Sickmann, A. & Falk, M. Precise protein quantification based on peptide quantification using iTRAQ^{TM}. *BMC Bioinformatics* **8**, (2007).

64. Gerster, S. *et al.* Statistical approach to protein quantification. *Mol. Cell. Proteomics* **13**, 666–677 (2014).

65. Silva, J. C., Gorenstein, M. V., Li, G. Z., Vissers, J. P. C. & Geromanos, S. J. Absolute quantification of proteins by LCMSE: A virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **5**, 144–156 (2006).

66. O'Brien, J. J. *et al.* Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags. *J. Proteome Res.* **17**, 590–599 (2018).

67. Smith, L. M. & Kelleher, N. L. Proteoform: A single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).

68. Kobayashi, R., Patenia, R., Ashizawa, S. & Vykoukal, J. Targeted mass spectrometric analysis of N-terminally truncated isoforms generated via alternative translation initiation. *FEBS Lett.* **583**, 2441–2445 (2009).

69. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* **26**, 676–687 (2015).

70. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).

71. Wu, R. *et al.* A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Methods* **8**, 677–683 (2011).

72. Olsen, J. V. *et al.* Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci. Signal.* **3**, ra3–ra3 (2010).

73. Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715–726 (2001).

Chapter 2: Systematic analysis of the effects of ion selection parameters on peptide purity in hrMS2 and SPS-MS3 analyses

Abstract

Quantitative mass spectrometry using isobaric tagging, such as Tandem Mass Tags (TMT) and Isobaric Tags for Relative and Absolute Quantitation (iTRAQ), is a technique that is commonly used. However, the intricacies surrounding interference, and thus peptide purity, caused by co-eluting and co-fragmenting peptides is not well understood. To understand better how instrument performance impacts interference and purity, a yeast and mouse two-proteome model was developed. These two species were selected because of their highly different proteomes. The 5Da mass difference between the reporter ions generated from the TMT-126 and TMT-131 reagents was utilized to minimize the effect of isotopic impurities, thus ensuring observed ratio abnormalities are primarily from interference. Additionally, TMT reagents TMT-130N and TMT-130C were used as a duplex to assess whether interference was increased when reporter ions were only separated by a 6mDa shift. The labelled samples were combined 1:1 such that the mass of yeast and mouse peptides were equal. Samples were analysed by high resolution MS2 and SPS-MS3 on an Orbitrap Fusion Lumos and interference was defined as the composition of the total TMT signal observed in a peptide-spectrum-match (PSM) attributed to the wrong species while purity was defined as 1 – (Peptide Interference). Primarily, we find that selecting the most stringent quadrupole isolation widths prior to reporter ion MS/MS analysis will result in increased purity at the cost of signal-to-noise. Additionally, we find that decreasing the number of ions for Synchronous Precursor Selection-MS3 (SPS-MS3) analysis will yield similar results for MS3 based quantitation. Furthermore, we mapped interference to the attributes of identified peptides and found that longer, and thus heavier, peptides with higher charge states tend to be observed with lower interference. A stronger relationship was observed between

Sequest Xcorr values and interference with higher values resulting in lower observed interference.

Introduction

Major technological improvements in mass spectrometry detector technologies such as the FT-ICR and Orbitrap have allowed mass spectrometry to remain a powerful and flexible resource in proteomics[1–3]. By increasing the resolving power and mass accuracy, these technologies have allowed researchers to perform quantitative proteomics experiments with both accuracy and confidence. When combined with liquid chromatography and tandem mass spectrometry (LC-MS/MS), this technology allows for the unbiased high-throughput quantitative analysis and sequencing of complex samples containing a mixture of various proteins digested into peptides[4,5]. Complex protein mixtures are separated by online or offline liquid chromatography (LC) instruments by physical properties such as size, charge, and hydrophobicity before injection into a mass spectrometer for tandem mass spectrometry (MS/MS) analysis. The decreased complexity of LC separated samples allows researchers to dig deeper into these mixtures to find low abundant proteins whose signal would have been obscured by more abundant proteins.

A further benefit of the improved technology is the ability to analyse multiplexed samples beyond the capabilities of stable isotope labelling in cell culture (SILAC). While SILAC has been readily available since 2002 and considered one of the most accurate forms of multiplexing, it is limited to samples obtained from cell culture and is not practically useful beyond comparing three conditions[6–9]. To address the limitations of SILAC, isobaric tagging methods were developed, such as iTRAQ and tandem mass tags (TMT), which can easily and affordably multiplex samples beyond three conditions. In the case of TMT, commercially

available reagents allow for up to 11 conditions (16 with the new proline based TMTpro reagent) to be analysed simultaneously and do not require samples to originate from cell culture[10–12]. However, these isobaric tags are not without their limitations, one of them being a reduction in quantitative accuracy due to peptide interference (i.e., co-isolation of co-eluting precursors for MS/MS analysis)[13–16].

While certain methods are more resilient against peptide interference, the reduction in accuracy from peptide interference stems from how TMT reagents are quantitatively analysed[17,18]. Peptides are chemically labelled with the TMT reagents at the N-terminus and or the amine group on lysine side chains; each sample to be multiplexed receives a different TMT reagent. During peptide fragmentation for LC-MS/MS analysis, the TMT reagents are fragmented into the reporter ion and the mass balancer region – the latter of which stays covalently bound to the peptide fragment. The relative abundance of the different TMT reporter ions reflects the relative abundance of the peptide in each sample. Because reporter ions are quantified instead of direct quantitation of peptide precursors and subsequent transitions, co-eluting and co-isolating precursors can contribute reporter ion signals which adversely affect accurate quantitation. Specifically this phenomenon tends to result in ratio compression, which falsely implies ratios are closer to 1:1 than in actuality[15,16].

To reduce interference, our lab has previously reported an MS3 method that utilizes CID MS2 as a filter before HCD MS3 fragmentation and reporter ion quantitation[16]. This method was enhanced further by using multiple frequency notch waveforms to isolate multiple ions in an MS2 spectrum for MS3 analysis, a method termed synchronous precursor selection MS3 (SPS-MS3)[19]. When using these methods instead of traditional high resolution MS2 (hrMS2) methods, the amount of peptide interference was drastically reduced. Coupling these methods with data

filtering methods such as spectrum filtering through linear discriminate analysis has facilitated the acquisition of data sets with minimal interference[20].

Although our current methods using MS3 have improved the quality of data sets, obtaining data sets completely devoid of peptide interference remains elusive. In this study, we seek to quantify the effects of notch selection on peptide interference using current MS3 methods. To this end, we employed a two-proteome model using a fractionated 1:1 mixture of yeast whole cell lysate and mouse whole brain lysate with each proteome tagged with a different TMT labels[21]. This model allows us to easily track and quantify interference when performing single proteome searches as the yeast and mouse proteome have minimal sequence overlap. The samples were used to test whether varying the number of notches selected during SPS-MS3 would reduce interference as well as the effects of increasing gradient length. Our findings suggest that shorter acquisition times and fewer notches result in reduced peptide interference, thus increasing peptide purity.

Method

Two-proteome model

A sample of wild type *Saccharomyces cerevisiae* was lysed with 8M urea buffered to pH 8.5 in 100mM EPPS to form yeast whole cell lysate as described previously[16]. Protease and phosphatase inhibitors were added to the buffer to prevent protein degradation. After a BCA assay, 1mg of yeast whole cell lysate was precipitated by chloroform-methanol precipitation and resuspended in 100mM EPPS at pH 8.5 and digested with 10µg of Lys-C overnight at room temperature in a shaker. 10µg of Trypsin was then added to the sample and digested for 6 hours in a 37C shaker. A whole brain from a wild type C57BL/6 mouse was obtained and lysed in 8M urea buffered to pH 8.0 in 100mM EPPS with a tissue drill as described previously[22]. The buffer

also contained protease and phosphatase inhibitors. After application of the tissue drill, the sample was further syringed lysed in a 21-gauge syringe tip. From here the sample was treated identically to the yeast whole cell lysate (chloroform-methanol precipitation and enzymatic digestion).

Both samples were desalted via Sep-Pak purification before separate treatment with TMT labelling reagents. TMT reagents were utilized in pairs: a 5Da difference in reporter ion mass (TMT-126 and TMT-131) and a 6mDa difference (TMT-130N and TMT TMT-130C). Four aliquots of mouse lysate and four of yeast were labelled with the four TMT reagents specified above. Samples were then cleaned again, separately, via Sep-Pak. A quantitative colorimetric peptide assay was performed to determine the peptide concentration of each sample. The two samples were then mixed at a 1:1 ratio which was confirmed by a ratio check. A duplex combination strategy was employed such that four TMT-duplexes were generated such that each species would be observed with each label over the course of the study. For single-shot analysis on the Orbitrap Fusion Lumos, combined samples were dried down, de-salted via Stop-and-go Extraction tips (STAGE tips), and resuspended in loading buffer comprised of 5% acetonitrile and 5% formic acid before LC-MS/MS analysis[23].

Basic-pH reverse phase fractionation

The combined sample containing the TMT-131 labelled mouse and TMT-126 labelled yeast lysates was then subjected to basic-pH reverse phase chromatography on an Agilent 300Extend C18 column (3.5µm particles, 4.6 mm ID, 250 mm length) attached to an Agilent 1260 Infinity LC with degasser and a single wavelength detector set at 220nm. After the combined peptide sample was loaded onto the column, the sample was eluted with a 50-minute linear gradient ranging from 8% to 40% acetonitrile in 10mM ammonium bicarbonate buffered

to pH 8. The flow rate was set to 0.6 mL/min and the sample fractionated into a 96 well plate which was then combined in a checkerboard pattern into 24 fractions, A 1-12 and B 1-12. Fractions B4, B5, and B6 were selected for analysis on the Orbitrap Fusion Lumos and samples A4, A6, and A7 were analysed on an Orbitrap Fusion instrument.

Liquid Chromatography and Mass Spectrometry

Samples were analysed on an Orbitrap Fusion Lumos with a Proxeon EASY-nLC 1000 before the source (Thermo Fisher Scientific, San Jose). The instrument was operated in data-dependent mode for all SPS-MS3 methods. For each analysis, 1μg of peptides was separated by liquid chromatography at a flow rate of ~350nL/min on a microcapillary column with a 100 μm inner diameter packed with ~35cm of Accucore C-18 resin (2.6 μm, Thermo Fisher Scientific). Separation was performed in-line with the mass spectrometer with a gradient of 6 to 26% acetonitrile in 0.125% formic acid. Gradient length was varied between 45-, 90-, and 180-minutes for unfractionated analyses and 60-, 120-, and 180-minutes for fractionated analyses.

Quadrupole isolation widths for analyses of unfractionated samples were varied between 0.4, 0.7, 1.2, and 2.0 Th. For all fractionated analyses, the isolation window for all quadrupole related isolations was set to the lowest rated setting of 0.4 Th. Additionally, the initial MS1 survey scan was collected by an Orbitrap detector (resolution: 120,000; mass range: 350-1,400 *m/z*; automatic gain control (AGC): $5*10^5$, maximum injection time: 100ms. All precursors selected from the MS1 scan for MS2 analysis was performed using a Top10 method where the 10 most intense precursors from the MS1 spectrum were selected for analysis. For hrMS2 analysis, the selected precursors were isolated by the quadrupole and subjected to HCD with normalized collision energy of 35. These fragment ions were analysed in the Orbitrap (resolution: 60,000; AGC: $1.5*10^4$; maximum injection time: 120ms). For MS3 analysis on

unfractionated samples, the selected precursors from the MS1 scan were subjected to a CID-MS2 analysis with a normalized collision energy of 35% after quadrupole isolation. During CID-MS2 the ions were analysed in the instrument's ion trap. From here we used a TopN method (N being either 1, 2, 5, or 10) where the N most intense fragment ions from the CID-MS2 were isolated by SPS for further fragmentation. The fragment ions selected from the CID-MS2 analysis were subjected to HCD-MS3 analysis with a normalized collision energy of 55. Fractionated samples were analysed using a 10-notch SPS-MS3 method. MS3 analysis was performed in the Orbitrap (resolution: 60,000; scan range: 100-1000 *m/z*; AGC: $1.0*10^5$; maximum injection time: 120ms).

Quantitative Data Analysis and Data Presentation

Spectra from all mass spectrometry experiments were processed using a Sequest-based in-house software pipeline. Briefly, spectra were converted into mzXML format and searched using against three databases: a yeast only database containing entries from SGD (*Saccharomyces* Genome Database; last modified January 13, 2015), a mouse only data base containing entries from Uniprot (Universal Protein Resource; ProteomeI D: UP000000589; last modified August, 21, 2019), and a concatenated database combining both the yeast and mouse databases. All databases included a list of common contaminants. For the database search, each database was concatenated with a list composed of reversed protein sequences derived from all proteins in the appropriate database. Static modifications for TMT tags on the peptide's N-terminal and lysine residues (+229.163 Da) and carbamidomethylation of cysteine residues (+57.021 Da) as well as a variable modification for the oxidation of methionine (+15.995 Da) were added to the search parameters. Additionally, for searches performed on all MS3 acquired data, the peptide mass tolerance was set to 50ppm and the fragment ion tolerance was set to 0.9 Da; MS2 acquired data was searched with the same peptide mass tolerance but with a fragment

ion tolerance of 0.03 Da. These tolerance values were selected to maximize sensitivity during Sequest searches and subsequent linear discriminant analysis.

All peptide-spectrum matches (PSMs) returned from the search were adjusted to a 1% false discovery rate (FDR) by the target-decoy method and linear discriminant analysis per previous protocols[24]. The linear discriminant analysis considered the following parameters: XCorr, ΔCn, missed cleavages, peptide length, charge state, and precursor mass accuracy. Once PSMs were filtered to a 1% FDR they were collapsed to reduce the protein-level FDR to 1%. Protein assembly was performed using principles of parsimony to limit the number of proteins to the smallest set that can account for all observed peptides. TMT reporter ion quantitation was performed by utilizing the signal-to-noise ratio (SNR) data for the TMT-126 and TMT-131 channels or TMT-130N and TMT-130C channels defined by the closest matching centroid to the expected mass of the reporter ion. SNR values that were below 1 were replaced with a value of 1 as, by the definition of noise, it is impossible to distinguish signal from noise below the noise level.

Peptide interference for a PSM was calculated as the amount of the total reporter ion SNR attributed to the incorrect species. For example, interference in a yeast PSM would be calculated using the following equation:

$$\text{Peptide Interference} = \frac{\text{TMT SNR}_{\text{Mouse Channel}}}{\text{TMT SNR}_{\text{Yeast Channel}} + \text{TMT SNR}_{\text{Mouse Channel}}}$$

For PSMs attributed to the mouse species, the numerator would be replaced with the component of the total signal coming from the TMT channel assigned to the yeast species. Peptide purity is defined as: $1 - (\text{Peptide Interference})$.

To calculate statistical differences, a surrogate measurement was used in place of Peptide Interference:

$$\frac{\text{TMT SNR}_{\text{Incorrect Species}}}{\text{TMT SNR}_{\text{Correct Species}}}$$

This ratio of TMT SNR was log-2 transformed to avoid boundary conditions surrounding the definition of Peptide Interference (i.e. peptide interference must be between 0 and 1). It should be noted that the following transformation can be performed on the ratio to obtain the original definition of Peptide Interference:

$$1 - \frac{1}{1 + \frac{\text{TMT SNR}_{\text{Incorrect Species}}}{\text{TMT SNR}_{\text{Correct Species}}}}$$

Analysis of variance (ANOVA) was performed utilizing the log-2 transformed ratios and Tukey's Honestly Significant Difference post-hoc tests were performed to identify conditions that were statistically different.

Results

Two-proteome model allows for analysis of peptide interference during SPS-MS3 analysis

We utilized a two-proteome model using a 1:1 mixture of yeast whole cell lysate from *S. cerevisiae* and mouse whole brain lysate labelled with TMT reagents from either the 126 and 131 pair or the 130N and 130C pair to quantify peptide interference (Figure 2-1). These unfractionated mixtures were analysed using a Thermo Orbitrap Fusion Lumos instrument to understand better how instrumentation affects peptide interference. As such, several instrument parameters, including whether peptide quantitation occurred by the high-resolution MS2 (hrMS2) or SPS-MS3 methods, were evaluated (Table 2-1). A total of 76 independent LC-MS/MS runs were analysed, 52 utilizing an SPS-MS3 method and 24 using a hrMS2 method.

Increased chromatographic gradient lengths correlates to increased peptide interference

As longer LC gradients can sometimes allow for better separation of chromatographic features, we believed that varying the gradient length would result in an increase in peptide and

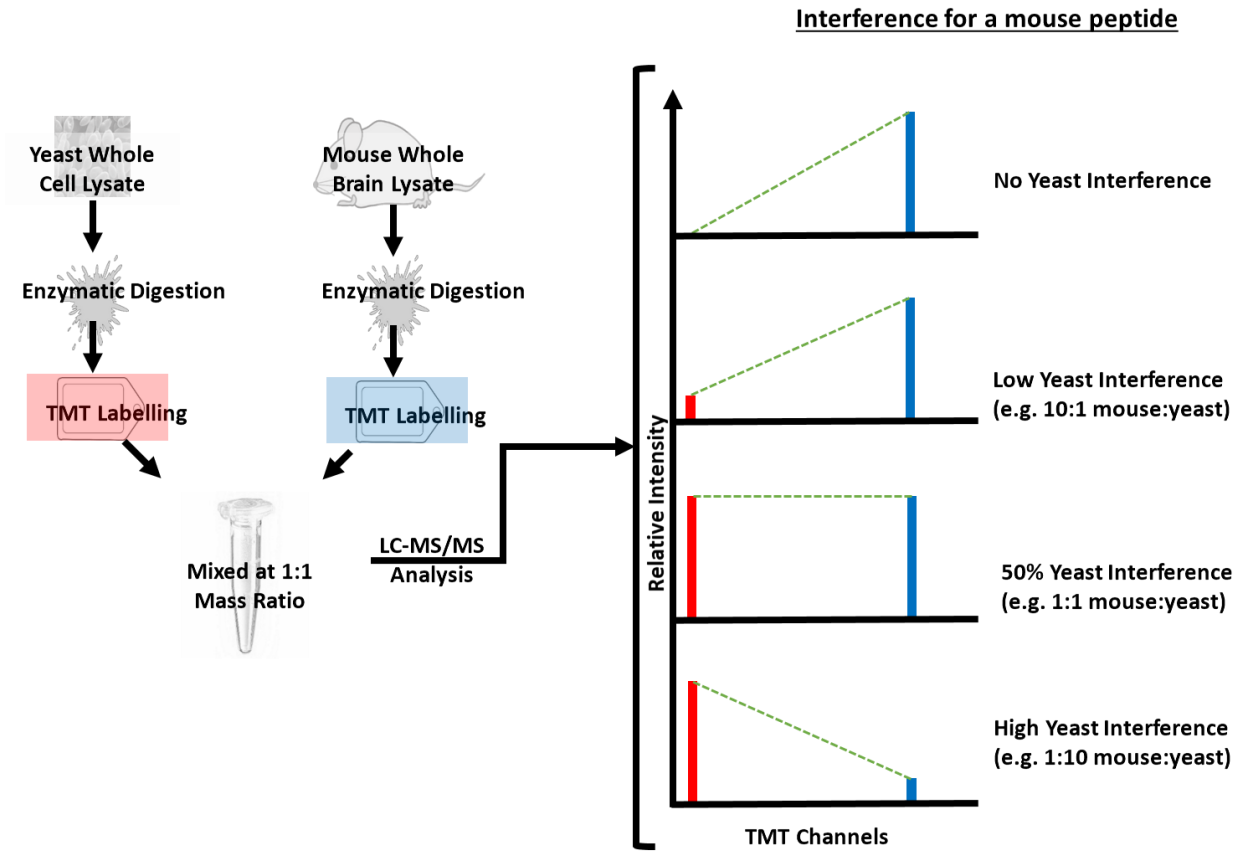


Figure 2-1: Graphical representation of the experimental workflow described in the method section for unfractionated 1:1 mixtures of mouse and yeast peptides. Different levels of interference for a hypothetical mouse peptide depict how interference would be measured using reporter ions (e.g. mouse receiving TMT-131 and yeast receiving TMT-126). To calculate interference for a yeast peptide from the same sample an identical approach using the appropriate TMT reporter ion channels would be utilized. This experimental approach is valid for all duplex combinations of TMT reporter ion channels used in this experiment. Additionally, for fractionated data, the workflow is identical except a fractionation step is performed before LC-MS/MS analysis.

Table 2-1: Table of various experimental conditions tested.

	hrMS2 Conditions	SPS-MS3 Conditions
LC gradient length (mins)	45, 90, & 180	45, 90, & 180
IW1 (Th)*	0.4, 0.7, 1.2, & 2.0	0.4, 0.7, 1.2, & 2.0
IW2 (Th)**	Not Applicable	0.4, 0.7, 1.2, & 2.0
Number of SPS ions selected	Not Applicable	1, 2, 5, & 10
Total number of LC-MS/MS Runs	**24**	**52**

Red text indicates experimental conditions used as baseline.
*IW1 is the quadrupole isolation width specified for precursor selection prior to an MS2
**IW2 is the quadrupole isolation width specified for precursor selection prior to an MS3

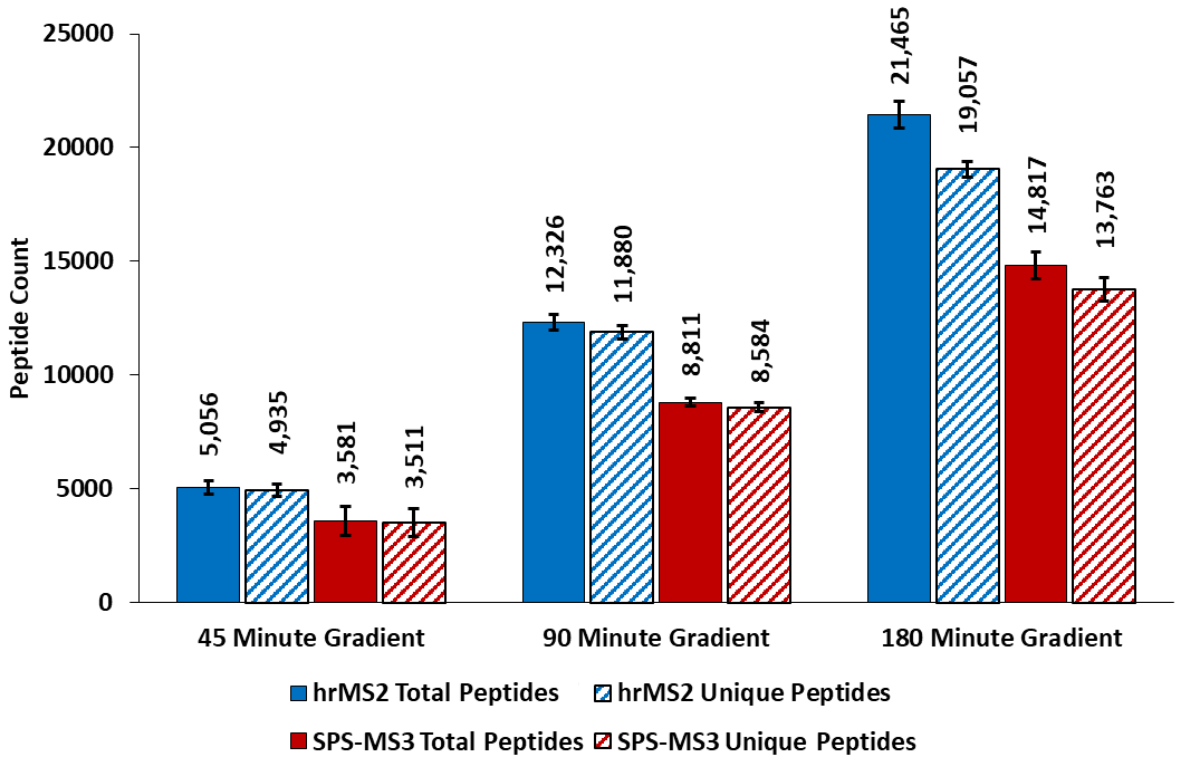
protein identifications[5]. To test this, we utilized both hrMS2 and SPS-MS3 methods while surveying gradient lengths of 45, 90, and 180 minutes. Isolation widths for the MS1 to MS2 transition (IW1) were kept at 0.7 Th while the isolation width for the MS2 to MS3 transition (IW2) was kept at 1.2 Th. A 10-notch SPS-MS3 method was used for all MS3 methods.

This belief proved to be accurate with both total and unique peptide identifications doubling, on average, as gradient lengths doubled (Figure 2-2). Additionally, we observed that, while identifications increased with gradient length regardless of acquisition method, hrMS2 methods would consistently identify an average of 42% more total peptides corresponding to an average increase of 39% for unique peptides than comparable SPS-MS3 methods utilizing an identical LC gradient (p-value < 0.001). The increase in identification rates was slightly diminished at the protein level with hrMS2 methods identifying an average of 34% more proteins than SPS-MS3 (p-value < 0.001). This is due to the increased workload of the mass spectrometer when conducting high-resolution data dependent MS3 scans.

However, to understand further how the increase in LC gradient length was affecting quantitation, we analysed the relationship between MS acquisition method and gradient length with peptide interference and TMT Summed SNR. The results from these comparisons showed a general trend that increasing gradient length resulted in a decrease in the peptide TMT Summed SNR observed (p-values < 0.001) with SPS-MS3 methods generally resulting in higher levels of SNR (p-values < 0.001) (Figures 2-3A). We believe this negative relationship between TMT Summed SNR and gradient length is due to large increase in peptide identifications. With longer LC gradients, the mass spectrometer can target lower abundant peptides that were previously obscured by highly abundant peaks. These low abundant peaks will, as a by-product of being low abundant, generate TMT Summed SNR values that are lower than the previously identified and

Figure 2-2: Summary of peptide and protein identifications from unfractionated LC-MS/MS analysis experiments, black error bars depict one standard deviation. Identifications from hrMS2 methods are shown in blue while identifications from MS3 methods are shown in red. **A)** Total (solid bars) and unique (hashed bars) peptides. **B)** Total proteins. 45-minute LC gradient lengths were used as part of the baseline method for the analysis of unfractionated samples in this study. As such, n = 4 for all 90- and 180-minute gradient experiments, while n = 16 for 45-minute hrMS2 experiments and, due to the numerous additional parameters tested, n = 44 for 45-minute MS3 methods.

Figure 2-2 (continued)



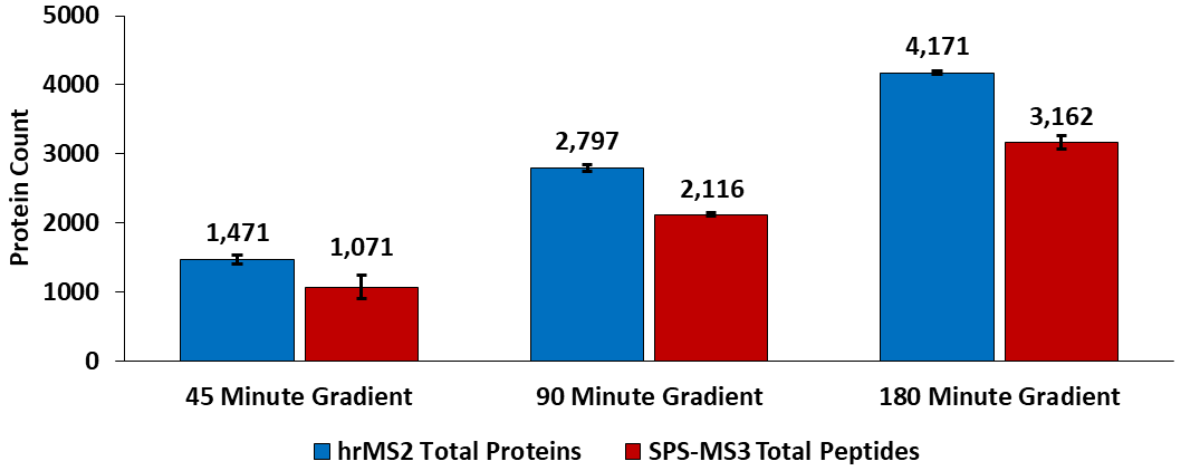
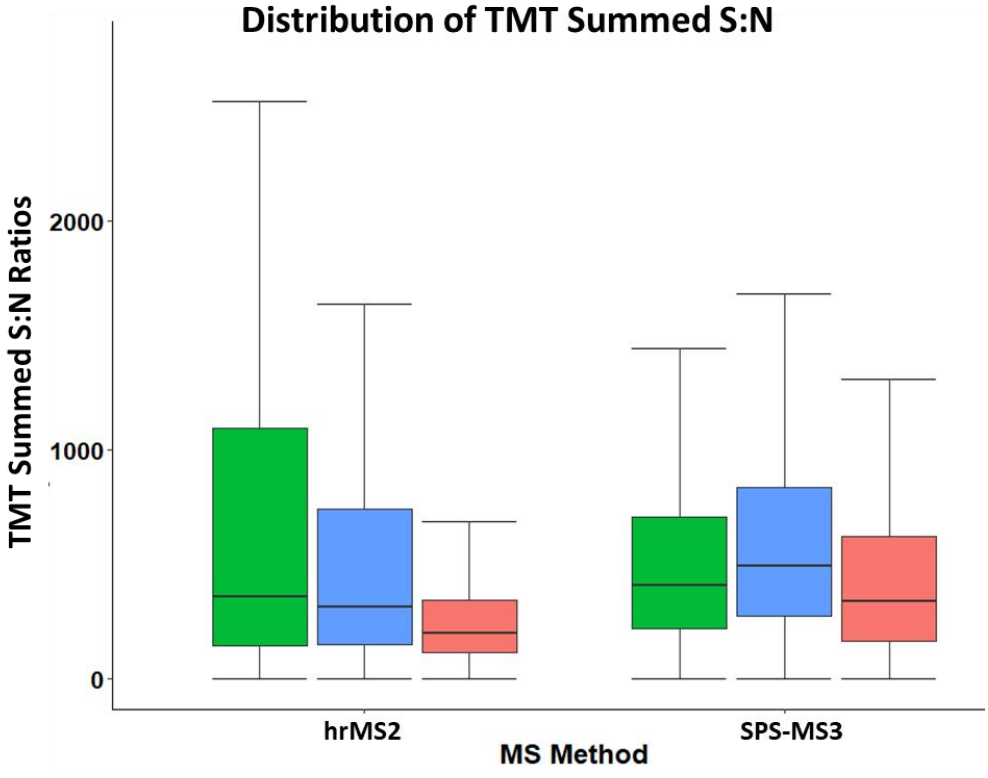
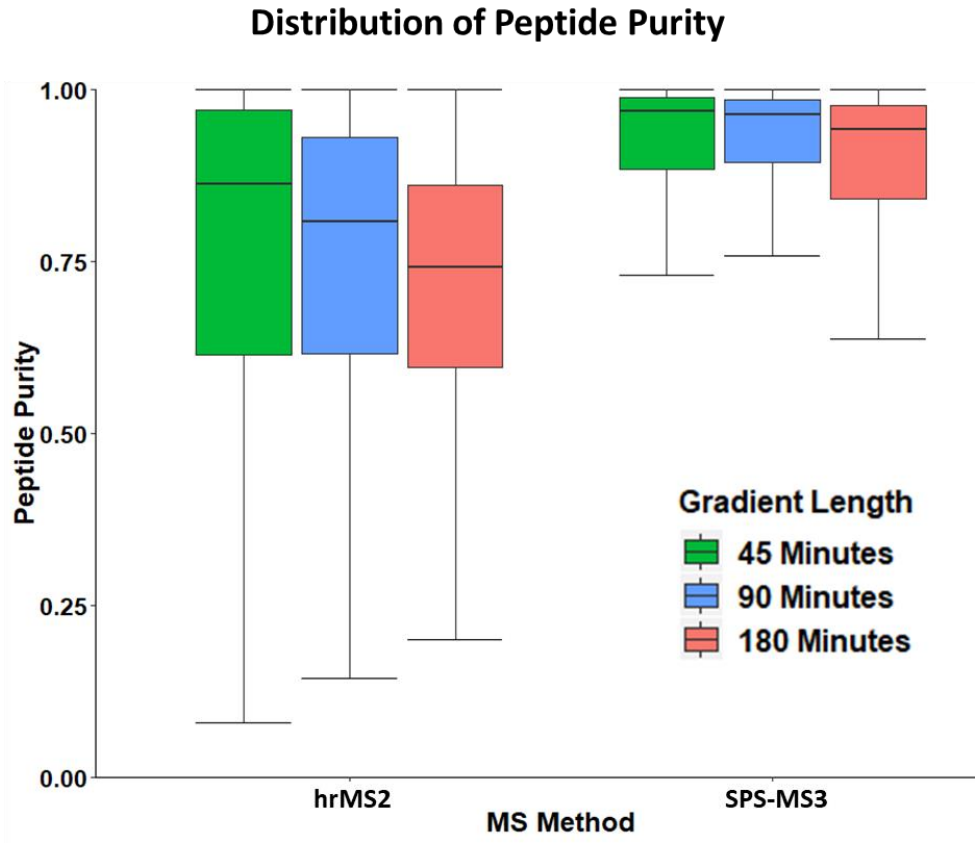


Figure 2-3: Analysis of the effect of varying LC gradient lengths on peptide purity. **A)** Boxplots showing the distribution of TMT Summed SNR for observed peptides over various LC gradient lengths and various MS acquisition methods. **B)** Boxplots of peptide purity over various LC gradient lengths and MS acquisition methods. Notably, MS3 methods resulted in less inter- and intra-method (i.e. gradient length) variation compared to their hrMS2 counterparts. Additionally, all MS3 methods resulted in higher average and median peptide purities implying lower peptide interference is generated with this method.

Figure 2-3 (continued)





quantified abundant features thus reducing the average. Furthermore, the SPS-MS3 method is likely to generate stronger TMT Summed SNR values due to its quantitation by MS3 rather than MS2, note that default IW2 settings are higher, and thus reporter ion intensities, than IW1.

Analysis of peptide purity and interfere showed that SPS-MS3 increased peptide purity over the hrMS2 method (p-value < 0.001) (Figure 2-2B). Additionally, in the hrMS2 data, increasing gradient lengths resulted in a significant decrease in purity and thus an increase in peptide interference likely due to the emergence of low abundant peaks in the LC elution profile that cannot be baseline resolved due to increased peak widths from the extended gradient length (p-value < 0.001). This trend was only partially observed in the SPS-MS3 data as no significant difference was observed between the use of a 45- or a 90-minute LC gradient. However, significant p-values were obtained when comparing 45- to 180-minute gradient lengths and 90- to 180- minute gradient lengths.

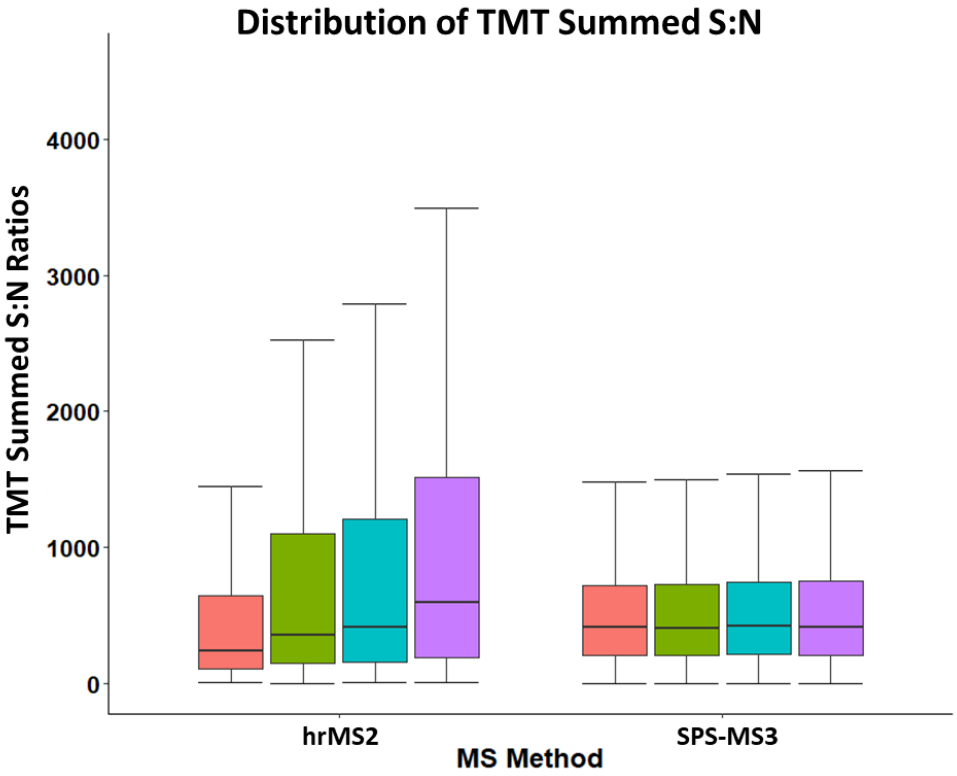
Larger quadrupole Isolation Widths Increase Peptide Interference

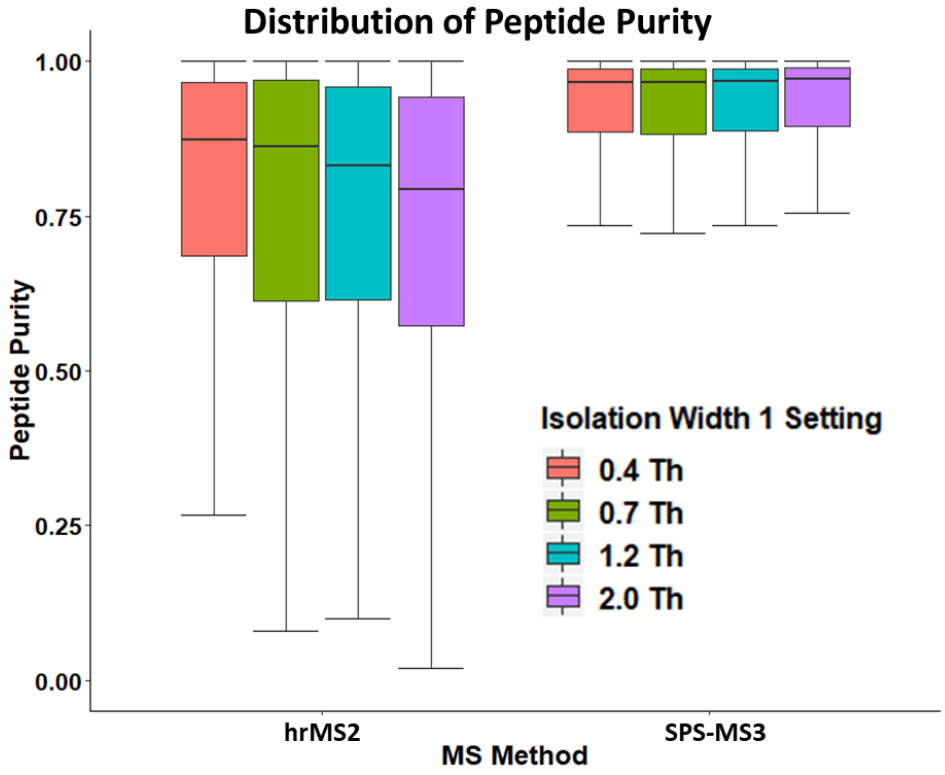
TMT quantitation requires the targeted fragmentation and analysis of a precursor ion. As such we hypothesized that modulating the isolation width used to filter ions before HCD fragmentation would affect the levels of peptide interference. For hrMS2 TMT analysis, IW1 controls ion selection prior to HCD fragmentation while IW2 controls the selection in the case of all MS3 analyses. To test the effects of various IW1 settings, we kept the LC gradient length constant at 45 minutes for our hrMS2 and 10-notch SPS-MS3 methods while varying IW1 between 0.4, 0.7, 1.2, and 2.0 Th. IW2 was held constant at 1.2 Th during the IW1 trials (Table 2-1).

As expected, increasing widths of IW1 resulted in a direct increase in peptide TMT SNR when performing hrMS2 analysis (Figures 2-4A). This phenomenon, however, was not true for SPS-MS3 analysis as changing IW1 values as TMT quantitation for SPS-MS3 is not reliant on

Figure 2-4: Analysis of the effect of varying IW1 settings on peptide purity. **A)** Boxplots showing the distribution of TMT Summed SNR for observed peptides over various IW1 settings and various MS acquisition methods. **B)** Boxplots of peptide purity over various IW1 settings and MS acquisition methods. Notably, TMT Summed SNR and peptide purity of MS3 methods did not vary with changing IW1 settings as IW1 does not directly affect quantitation in an MS3 method. Of note, SPS-MS3 methods had higher peptide purity than any of the hrMS2 methods suggesting that hrMS2 cannot reach SPS-MS3 levels of purity even with improved quadrupole isolation specificity.

Figure 2-4 (continued)





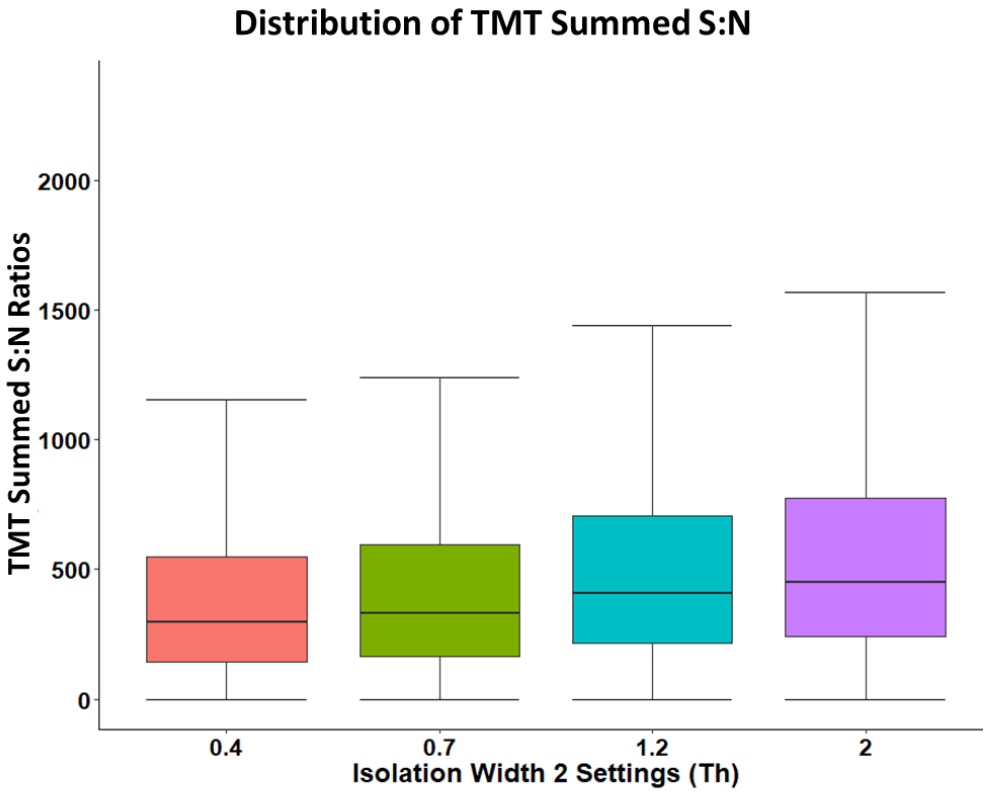
ions filtered with IW1. All comparisons of TMT Summed SNR from hrMS2 analyses with various IW1 settings resulted in significant p-values (p-values < 0.001). A 52% increase in the average TMT Summed SNR was observed between the IW1 lowest setting, 0.4 Th, and the highest tested setting, 2.0 Th. Increasing the baseline IW1 setting of 0.7 Th to the baseline IW2 setting of 1.2 Th resulted in a 7% increase of the TMT Summed SNR average.

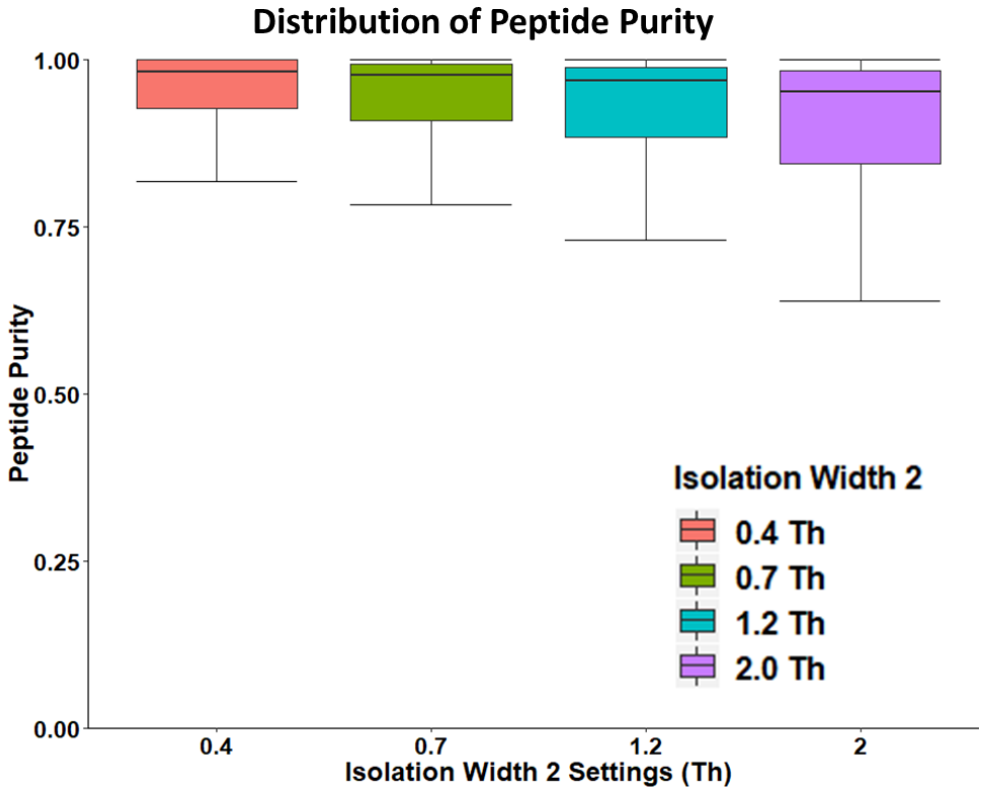
Peptide purity, however, negatively correlated with increasing IW1 settings for hrMS2 analyses with higher IW1 settings resulting in lower peptide purity (Figure 2-4B). Significantly different decreases in peptide purity were observed in all pairwise comparisons of increasing IW1 settings for hrMS2 analyses with an average peptide purity of 83% observed for IW1 settings of 0.4 Th and 77% for IW1 settings of 2.0 Th (p-values < 0.001). No significant difference was observed for any IW1 changes when performing an SPS-MS3 analysis (ANOVA p-value = 0.0582).

As mentioned previously, in MS3-based TMT analyses, IW1 only controls selection for ions used for peptide identification. Because identification and quantitation are decoupled during an MS3 analysis, this means that IW1 does not play a direct role in the latter. Therefore, to test how quadrupole isolation widths affect M3-based TMT analyses, IW1 was left constant at 0.7 Th while IW2 was varied between 0.4, 0.7, 1.2, and 2.0 Th (Table 2-1). Analogous to IW1, increasing values of IW2 resulted in statistically significant increases in TMT Summed SNR for observed peptides (p-values < 0.001) (Figure 2-5A). Additionally, all increasing IW2 comparisons resulted in decreases of peptide purity (p-values < 0.001) (Figure 2-5B). However, only a 3% absolute decrease in the average peptide purity was observed (96% to 93%) when increasing IW2 from 0.4 Th to 2.0 Th, highlighting the powerful interference reducing capabilities the MS3-based methods over hrMS2 analyses.

Figure 2-5: Analysis of the effect of varying IW2 settings on peptide purity. **A)** Boxplots showing the distribution of TMT Summed SNR for observed peptides over various IW2 settings for SPS-MS3 methods. **B)** Boxplots of peptide purity over various IW1 settings for SPS-MS3 methods. Increasing IW2 settings resulted in decreased peptide purity and increased TMT Summed SNR, similar to the relationships observed for varying IW1 in hrMS2 analyses.

Figure 2-5 (continued)





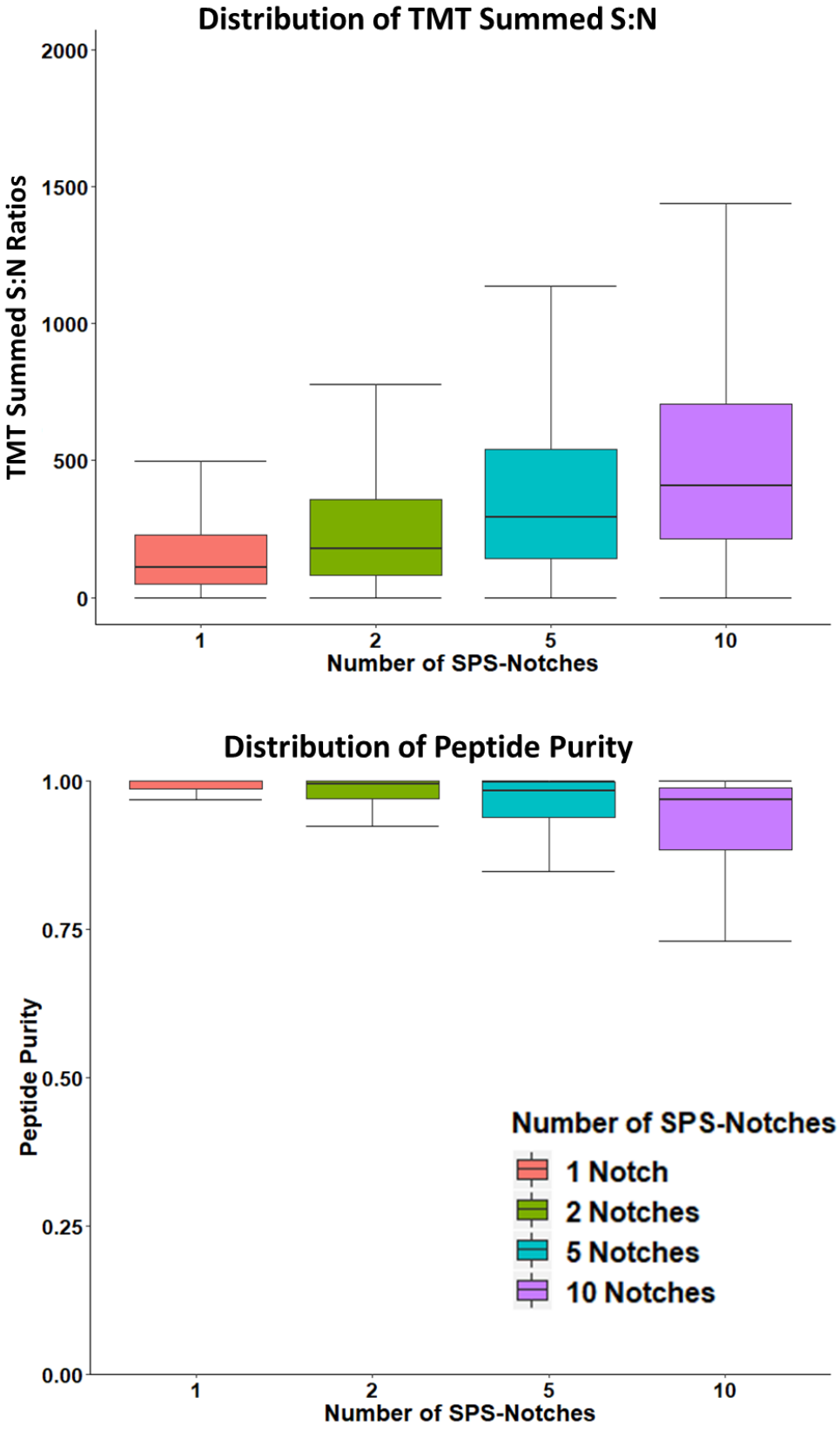
In general, the conserved inverted relationship between TMT Summed SNR and peptide interference is likely due to larger isolation widths allowing more MS features surrounding a target precursor to pass quadrupole selection. Additionally, the distribution of peptide purities for all SPS-MS3 methods displayed less variation than the distributions for hrMS2 during the IW1 experiments, in-line with previous findings that gas-phase separation methods can effectively reduce interference (Figures 2-4B and 2-5B)[16,21].

Changing the number of ions selected during SPS-MS3 analysis can affect peptide interference

So far, our data has shown that a 45-minute 10-notch SPS-MS3 method provides the least amount of peptide interference. However, the 10-notch method was designed to address sensitivity issues with the initial MS3 method[19]. We sought to understand the impact of increasing the number of SPS-ions selected on peptide interference by varying the number of ions selected between 1, 2, 5, and 10 (Table 2-1). A direct relationship between the number of ions selected and TMT Summed SNR was observed with increasing number of ion selections resulting in statistically significant increases in TMT Summed SNR (p-values < 0.001) (Figure 2-6A). Interestingly, while a general trend of increasing the number of SPS-ions resulted decreasing peptide purity (ANOVA p-value < 0.001), one pairwise comparison was noticeably not statistically significant (Figure 2-6B). However, when increasing the number of SPS-ions selected from 1 to 10, the maximum difference in the average peptide purities observed was 97% (1 ion selected) to 95% (10 ions selected). The most notable difference in the observed peptide purities when increasing the number of SPS-ions selected was the variance, not the mean. This is likely due to the increasing chance of selecting an incorrect ion when the number of SPS-ions selected increases.

Figure 2-6: Analysis of the effect of varying number of SPS-ions selected on peptide purity. **A)** Boxplots showing the distribution of TMT Summed SNR for observed peptides over various number of SPS-ions selected. **B)** Boxplots of peptide purity over various numbers of SPS-ions selected. Increasing the number of SPS-ions selected resulted in decreased peptide purity and increased TMT Summed SNR, similar to the relationship observed for varying IW2 in MS3 analyses.

Figure 2-6 (continued)



Characterizing the properties of observed peptides and their relationship to peptide interference

We next investigated whether type of peptide exists that was specifically enriched when conducting any of the various LC-MS/MS methods tested previously. Peptides were binned by peptide length, charge state, precursor m/z, and Sequest Xcorr value (Figure 2-7). No substantial enrichment was observed between LC-MS/MS of the same acquisition type (i.e. hrMS2 vs MS3-based). Additionally, no relationship between any of the binned categories and peptide interference was observed in any variation of the hrMS2 methods. However, when analysing the MS3-based methods, we observed that longer, higher charged, and heavier peptides tended to have lower observed peptide interference. Additionally, peptides that received a Sequest Xcorr value above a 6 had noticeably reduced interference levels compared to other peptides.

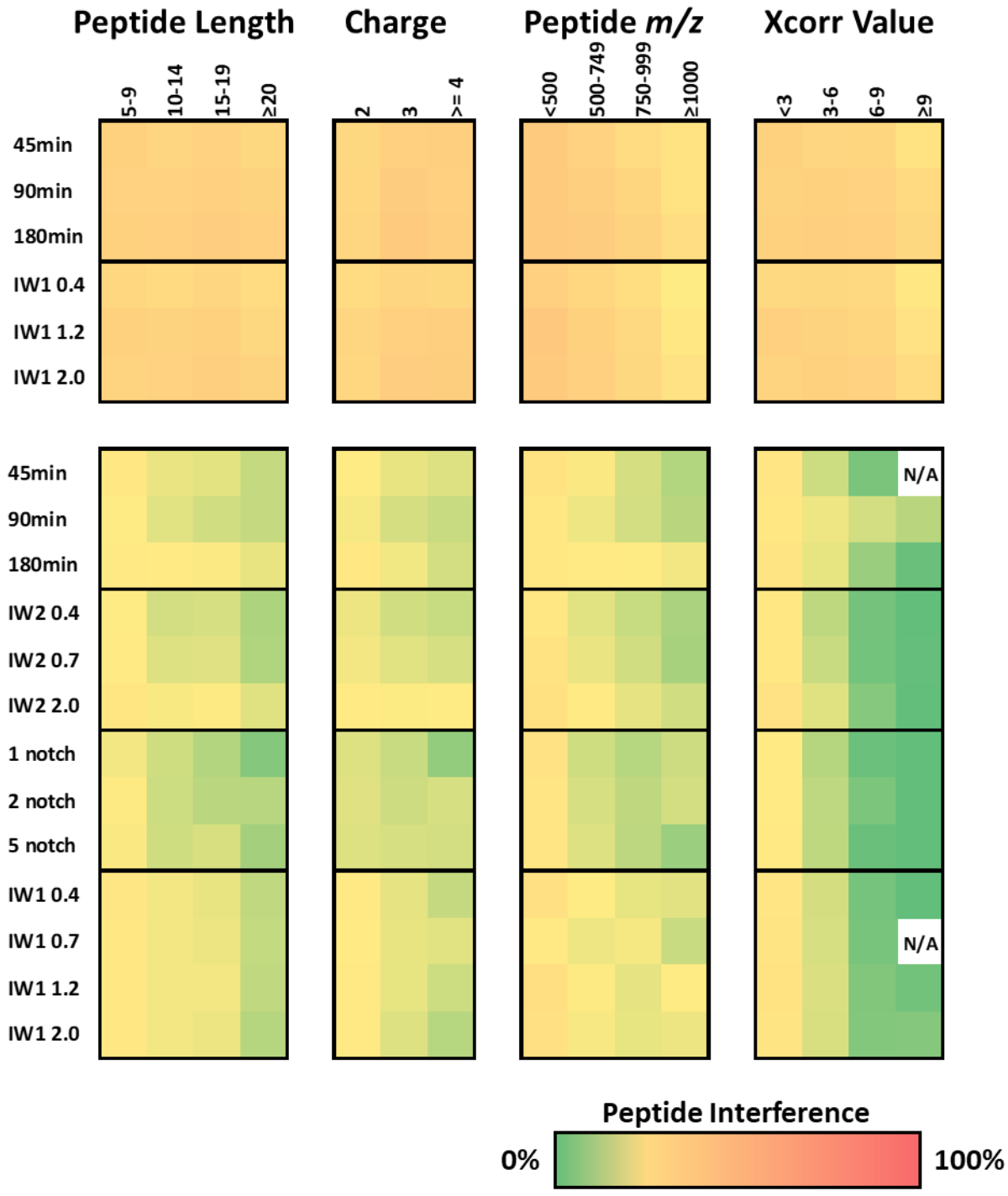
The effects of fractionation on Fusion Lumos analyses

The 1:1 mixture of yeast and mouse lysates was then fractionated by basic-pH reverse phase chromatography on an HPLC instrument into 24 fractions. Three fractions were selected and analysed by various LC-MS/MS methods such that each method was tested three times. We tested 10-notch SPS-MS3 methods with a varying LC gradients of 60-, 120-, and 180-minute gradient lengths, to assess our interference model using LC gradient lengths more accustomed to a standard protocol (Figure 2-8).

Peptide interference for data sets generated by these 10-notch SPS-MS3 methods displayed a decrease in peptide purity averages ranging between 96% and 98% (p-values < 0.001). As such, we concluded that while data acquisition length is correlated with the peptide interference such that peptide interference is increased when data acquisition time is increased. This affect is not as pronounced when using a 10-notch SPS-MS3 method on fractionated samples.

Figure 2-7: Heat map showing the relationship between peptide length, charge state, m/z, and Xcorr with peptide interference. There is an immediate distinction between hrMS^2 and SPS-MS^3 methods as the former has interference between 20-40% while the latter only has interference between 0-20%. Furthermore, we noticed that peptide length and peptide charge correlated inversely with peptide interference. This implied that *m/z* may be inversely correlated as well. To that extent, we noticed that peptides with *m/z* below 500 had more interference (~25%) while peptides with *m/z* above 1000 had less interference (~15%). As Xcorr can be used as a measure of quality for a mass spectrum (cross correlation, Xcorr, scores how well the real spectrum matches with a theoretical spectrum) we wanted to see whether Xcorr could also be a proxy for peptide interference. Xcorr values larger than 6 correlated with almost no interference. However, this population with extremely high Xcorr values only corresponded with a small fraction of the whole data set while the vast majority of peptides had Xcorr values below 6.

Figure 2-7 (continued)



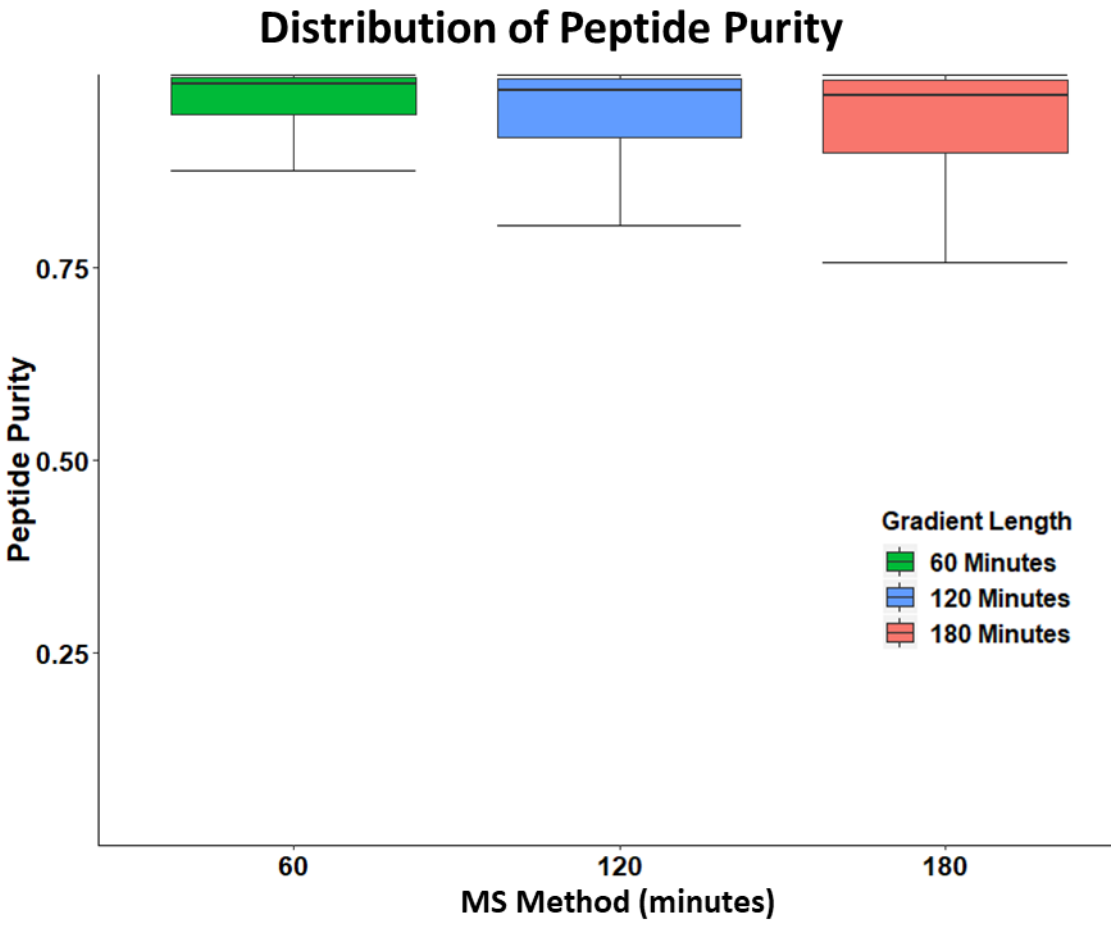


Figure 2-8: Analysis of the effect of varying LC gradient length on peptide purity for fractionated samples analyzed by 10-notch SPS-MS3 methods. Low variation between the observed median and mean of peptide purities was observed between the various LC gradient lengths.

Tracking the CID MS2 and HCD MS3 spectra of peptides in a protein reveals that incorrect ions are selected occasionally leading to increased peptide interference

SPS-MS3 quantitation relies on the selection of precursors identified from a low resolution MS2. We investigated ions the instrument selected during CID MS2 that would be used for a subsequent HCD MS3, to test whether peptide interference in SPS-MS3 can be attributed to incorrect notch selection. For consistency, we used data generated from 60-minute 10-notch SPS-MS3 methods. This resulted in four of the top ten yeast proteins with the greatest number of unique peptides identified containing at least one HCD MS3 spectrum with ≥ 50% peptide interference. One of these proteins was PCK1, also known as phosphoenolpyruvate carboxykinase 1, which is an essential protein regulating gluconeogenesis[25].

In our data set, we observed PCK1 peptides 33 times over the course of three experimental runs that translated into the identification of 13 unique peptides. Of these 12 unique peptides, up to 6 were identified in any given run. Interestingly, while 7 PSMs of the 33 observations of PCK1 peptides failed quality control due to low isolation specificity of TMT Summed SNR, only one spectrum had ≥ 50% peptide interference (Table 2-2). This peptide, MNATVGSTSEVEQK, was observed once in each run. In the instance when the spectrum yielded ≥ 50% peptide interference, there were noticeable abnormalities in the notches taken (Table 2-3). Specifically, six selected SPS ions were not associated to the peptide assigned after Sequest searching. These six ions contributed the majority of TMT signal likely resulting in the high interference. As such, ion selection during the CID MS2 portion of the SPS-MS3 method is important at determining the amount of peptide interference in a given spectrum.

Table 2-2: PSMs for yeast protein PCK1 can fail the quality control check for different reasons including interference.

Raw File	Scan Number	Peptide	126 S:N	131 S:N	Sum S:N	Isolation Specificity*	Interference**
a02169	8279	MNATVGSTSEVEQK	202.51	377.45	579.96	0.65	**0.65**
a02149	5986	SHVVDYDDSSITENTR	74.78	35.62	110.40	**0.45**	0.32
a02169	26059	LTPEQVMYHFISGYTSK	14.39	4.48	**18.87**	**0.00**	0.24
a02159	27291	VNGVPAELLNPAK	11.31	3.05	**14.35**	**0.41**	0.21
a02149	911	STIEINFK	4.96	0.00	**4.96**	0.84	0.00
a02149	1204	STIEINFK	11.43	0.00	**11.43**	1.00	0.00
a02159	26149	LTPEQVMYHFISGYTSK	25.27	0.00	**25.27**	0.93	0.00
a02169	2626	TTLSADPHR	91.48	0.00	**91.48**	0.99	0.00

Reasons a PSM failed the quality control check are shown in bolded red type. An isolation specificity cut-off of 0.5 and a Sum S:N cut-off of 100 were used as a quality control check for this analysis.

*Scaled between 0 and 1.

**Scaled between 0 and 1. A loose cut-off of 0.5 was set for interference.

Table 2-3: Incorrect SPS-Ion selection for PCK1 leads to high interference*

SPS Ion	Notch m/z	Intensity	Reason
1	487.6328	26819.3	y2 Ion (+1 with loss of ammonia)
2	504.4098	25364.7	y2 Ion (+1)
3	522.0707	15101	b9 Ion (+2 with loss of water)
4	531.1243	11981.8	Unknown
5	566.4671	136301.1	Unknown
6	586.6226	12442.4	b10 Ion (+2 with loss of water)
7	595.6882	35376.7	Unknown
8	654.2347	14523.1	Unknown
9	719.0631	16142.5	Unknown
10	746.4006	15537.2	Unknown

Red text indicates SPS ions that could not be identified from peptide precursor assigned by Sequest
*Data from scan 8279, raw file a02169.

Discussion

Based on our results and previous published studies, we conclude that SPS-MS3 drastically reduces peptide interference compared to hrMS2 analyses while minimizing the loss of peptide identifications due to the decreased sensitivity of the original "single-notch" MS3 method. However, here we find that, despite the improvements to accuracy and precision provided by the SPS-MS3 method, the number of total peptide identifications remains an average of 40% lower than that of traditional MS2 methods. Furthermore, while lengthening data acquisition times and chromatography gradients may improve the number of peptide identifications this can result in an increase of peptide interference affecting the accurate quantitation of the identified peptides.

Additionally, with the current improvements in mass spectrometry technology that improve sensitivity and isolation width precision, the number of SPS-ions required to address the MS3 sensitivity problems while maintaining low peptide interference are reduced. This has an additional benefit of reducing the likelihood of selecting an incorrect ion from the CID MS2 spectra that are used for subsequent HCD MS3, thus further reducing peptide interference. However, simply reducing the number of notches can only have a limited affect at reducing peptide interference in the SPS-MS3 method. A single-notch MS3 method (a non-SPS method), is the limit for low peptide interference. Specifically, a single-notch method runs the risk of an incorrect ion or a y-ion of a non-lysine peptide being selected for SPS-MS3 which can either result in 100% interference or no signal for quantitation during TMT-based experiments. This ultimately leads to a drastically reduced number of quantifiable peptides.

Since the current SPS-MS3 methods with the latest instrumentation produce data sets with extremely low interference, the future of TMT based multiplex proteomics requires

advancements in two areas: increased peptide identifications for MS3 methods and the ability to select the correct ions from CID MS2 spectra always. While the former goal will increase MS3's viability over MS2, the latter could potentially reduce peptide interference beyond the 1-notch limit. Active research into real-time database searching and online field asymmetric ion mobility spectrometry (FAIMS) has shown promising results at increasing peptide and protein identifications while reducing interference[26–29]. These new technologies allow for on-the-fly decision making for SPS ion selection and an additional gas phase purification step, respectively. Such breakthroughs would uncover low abundant proteins and rare posttranslational modifications to accurate quantitative proteomics studies.

References

1. Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annu. Rev. Biomed. Eng.* **11**, 49–79 (2009).

2. Marshall, A. G. Fourier transform mass spectrometry. *Spectrosc. Biomed. Sci.* **10**, 87–105 (2011).

3. Sleno, L. The use of mass defect in modern mass spectrometry. *J. Mass Spectrom.* **47**, 226–236 (2012).

4. Bondarenko, P. V., Chelius, D. & Shaler, T. A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography - Tandem mass spectrometry. *Anal. Chem.* **74**, 4741–4749 (2002).

5. Hsieh, E. J., Bereman, M. S., Durand, S., Valaskovic, G. A. & MacCoss, M. J. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J. Am. Soc. Mass Spectrom.* **24**, 148–153 (2013).

6. Ong, S. E. *et al.* 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 (2002).

7. Mann, M. Fifteen Years of Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). in *Methods and Protocols* **1188**, 1–7 (Springer Science+Business Media, 2014).

8. Rose, C. M. *et al.* Neutron encoded labeling for peptide identification. *Anal. Chem.* **85**, 5129–5137 (2013).

9. Overmyer, K. A. *et al.* Multiplexed proteome analysis with neutron-encoded stable isotope labeling in cells and mice. *Nat. Protoc.* **13**, 293–306 (2018).

10. Thompson, A. *et al.* Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).

11. Ross, P. L. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).

12. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

13. Zhang, G. *et al.* Protein quantitation using mass spectrometry. in *Computational Biology* (ed. Fenyö, D.) **673**, 211–222 (Humana Press, 2010).

14. Sandberg, A. S., Branca, R. M. M., Lehtiö, J. & Forshed, J. Quantitative accuracy in mass spectrometry based proteomics of complex samples: The impact of labeling and precursor interference. *J. Proteomics* **96**, 133–144 (2014).

15. Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8**, 933–935 (2011).

16. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

17. Ronsein, G. E. *et al.* Parallel reaction monitoring (PRM) and selected reaction monitoring (SRM) exhibit comparable linearity, dynamic range and precision for targeted quantitative HDL proteomics. *J. Proteomics* **113**, 388–399 (2015).

18. Vidova, V. & Spacil, Z. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal. Chim. Acta* **964**, 7–23 (2017).

19. Mcalister, G. C. *et al.* MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes Graeme C. McAlister, 1 David P. Nusinow, 1. *Anal. Chem.* **86**, 7150–7158 (2014).

20. Du, X. *et al.* Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. *J. Proteome Res.* **7**, 2195–2203 (2008).

21. Wenger, C. D. *et al.* Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nat. Methods* **8**, 933–935 (2011).

22. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).

23. Ishihama, Y., Rappsilber, J. & Mann, M. Modular Stop and Go Extraction Tips with Stacked Disks for Parallel and Multidimensional Peptide Fractionation in Proteomics. *J. Proteome Res.* **5**, 988–994 (2006).

24. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

25. Latorre-Muro, P. *et al.* Dynamic Acetylation of Phosphoenolpyruvate Carboxykinase Toggles Enzyme Activity between Gluconeogenic and Anaplerotic Reactions. *Mol. Cell* **71**, 718-732.e9 (2018).

26. Pfammatter, S., Bonneil, E. & Thibault, P. Improvement of Quantitative Measurements in Multiplex Proteomics Using High-Field Asymmetric Waveform Spectrometry. *J. Proteome Res.* **15**, 4653–4665 (2016).

27. Schweppe, D. K. *et al.* Characterization and optimization of multiplexed quantitative

analyses using high-field asymmetric-waveform ion mobility mass spectrometry. *Anal. Chem.* **91**, 4010–4016 (2019).

28. Erickson, B. K. *et al.* Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *J. Proteome Res.* **18**, 1299–1306 (2019).

29. Schweppe, D. K. *et al.* Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *bioRxiv* 668533 (2019). doi:10.1101/668533

Chapter 3: Evaluating False Transfer Rates from the Match-Between-Runs Algorithm with a Two-Proteome Model

Attributions: The following chapter was previously published on September 23, 2019 in the *Journal of Proteome Research.* Reproduced with permission from "Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model" Copyright 2019 American Chemical Society.

Matthew Y. Lim performed experiments, conducted data analysis, and wrote the manuscript.

Joao A. Paulo provided samples and conducted mass spectrometry instrumentation

Steven P. Gygi provided the initial concept for the project.

Abstract

Stochasticity between independent LC-MS/MS runs is a challenging problem in the field of proteomics resulting in significant missing values (i.e., abundance measurements) among observed peptides. To address this issue, several approaches have been developed including computational methods such as MaxQuant's Match-Between-Runs (MBR) algorithm. Often dozens of runs are all considered at once by MBR, transferring identifications from any one run to any of the others. To evaluate the error associated with these transfer events, we created a two-sample/two-proteome approach. In this way, samples containing no yeast lysate (n = 20) were assessed for false identification transfers from samples containing yeast (n = 20). While MBR increased the total number of spectral identifications by ~40%, we also found that 44% of all identified yeast proteins had identifications transferred to at least one sample without yeast. However, of these only 2.7% remained in the final dataset after applying the MaxQuant LFQ algorithm. We conclude that false transfers by MBR are plentiful, but few are retained in the final dataset.

Introduction

Stochasticity is an important issue for quantitative multi-run data-dependent acquisition (DDA) LC-MS/MS experiments as lack of observable evidence does not prove absence[1]. Attempts to address this issue have been both chemical and computational[2–4]. Chemically, various labelling methods have been developed to analyze multiple MS experiments simultaneously and address missing values[5]. Isobaric labelling techniques like Tandem Mass Tags (TMT) and iTRAQ provide peptide level modification and simplify downstream analysis while increasing sample throughput[5–7]. These chemical labels, however, are limited by the number of samples that can be analyzed simultaneously, currently 11 for TMT and 8 for iTRAQ.

Attempts to expand the multiplexing capabilities of these compounds often requires computational manipulation and experimental rearrangement (e.g. use of a bridge channel) or the complete redesign of the chemical compound[8,9].

Label-free quantitation inherently does not require chemical or metabolic labelling, but is susceptible to stochasticity[2,10]. As such, attempting to quantify across independent, label-free LC-MS/MS analyses can result in inaccuracies and missing data. To combat this, several computational methods have been applied[3,10]. However, most of these strategies employ imputation to fill in the missing values.

An alternative method to statistical imputation is to perform an identification transfer by leveraging chromatographic and mass-to-charge information. The most popular variation of this technique is the Match-Between-Runs (MBR) algorithm, which is included within the MaxQuant software suite[11,12]. Briefly, the MBR algorithm assesses each identified peak in an MS1 spectrum from an LC-MS/MS run and compares its retention time to unidentified peaks in another. An identification is transferred if an unidentified peak with the same properties (e.g. m/z and charge state) is found within a specified retention time window. As retention time is critical for the algorithm to function, the MBR algorithm first realigns compared chromatograms (by default, up to 20 min deviations) before attempting to transfer identifications. Thereby, identifications through peptide-spectrum matches (PSMs) from one run can be transferred to peaks having no tandem MS information in another run. However, this strength also presents the primary difficulty in assessing the accuracy of MBR.

With no tandem MS information, the authenticity of an identification transfer is not guaranteed. The need to validate identification transfers becomes more important as research groups begin to utilize the MBR algorithm for experiments with many runs. For example, a

recent publication comparing the proteomes of 29 different healthy human tissues utilized MBR with a total of over 1,800 MS/MS RAW files analyzed during the study[13]. Previously, work to assess the quality of identification transfers by MBR has utilized alignment of "ID-pairs" to investigate false transfer rate[14]. Simply put, peptide identifications in MBR compared LC-MS/MS runs should align closely after MBR chromatogram recalibration; transfers in regions where "ID-pairs" do not align well are assumed to be incorrect. By this method, MBR was found to have between 2% (LC-MS/MS runs analyzed on the same day) and 74% (LC-MS/MS a month apart) false transfers regardless of realignment and recalibration of sample chromatograms[14].

Here, we present a novel method utilizing a human-only Sample (H) compared with a human+10% yeast spike-in Sample (HY) to assess the false transfer rate via the MBR algorithm. To eliminate issues of carry over and to reduce the previously identified effects of column consistency on the MBR analysis, we first analyzed Sample H 20 consecutive times immediately followed by 20 consecutive analyses of the Sample HY. By leveraging Sample H as ground truth due to its single proteome composition, we are able to measure how often a false transfer occurs by counting the number of yeast protein identifications found in Sample H after MBR transfers compared to a standard analysis.

Through this study, we find that while the use of MBR greatly improves the missing values problem, false transfers from the MBR algorithm do occur at a measurable rate. When observing identifications across the experiment, we find that 44% of yeast proteins in the two-proteome sample were incorrectly transferred at least once to the human-only sample. Additionally, most of these incorrect transfers result in one-hit-wonder identifications in line with the belief that false transfers are spurious. However, by processing the MBR data with LFQ

enabled in MaxQuant, these spurious transfers were frequently assigned zero or near zero quantification values, thereby preventing incorrect quantitation.

Materials and Methods

Human Cell Culture

HCT116 cells were obtained from ATCC and cultured as described previously[15]. Briefly, cells were cultured in DMEM supplemented with 10% Fetal Calf Serum and 5% Penicillin/Streptomycin. Cells were kept at 37°C with 5% CO_2 until harvest. Harvesting occurred after cells reached 80% confluency by visual inspection. After ice-cold PBS wash, cells were lysed on-plate with 1 mL of an 8M urea lysis solution containing 200mM EPPS, pH 8.5, and protease inhibitors. Lysate was homogenized by trituration through a 21-gauge needle followed by sedimentation by centrifugation at 21,000 x g for 15 mins. Clarified lysate was flash frozen and stored at -80°C.

Yeast Samples

Saccharomyces cerevisiae was acquired from Dharmacon and cultured in standard yeast-peptone-dextrose (YPD) media as described previously[16,17]. Briefly, when the culture reached mid-log phase (measured by optical density of 0.6/mL), the culture was pelleted by centrifugation and resuspended in an 8M urea lysis solution containing 200mM EPPS, pH 8.5, and complete-mini, EDTA-free protease inhibitors (Roche). The resuspension was lysed via bead beating in microcentrifuge tubes. Lysate was clarified by centrifugation, flash frozen, and stored at -80°C.

Sample Preparations

Human cell lysate and yeast cell lysate were prepared for label-free LC-MS/MS analysis following protocols previously described[2]. Briefly, a protein level bicinchoninic acid protein

assay (Pierce) was performed on the clarified lysates to determine protein concentration. Lysates were then reduced at room temperature in the dark with incubation of 5mM tris(2-carboxyethly)-phosphine (TCEP), followed by alkylation with 10mM iodoacetamide to covalently block reactive cysteine groups. The reaction was quenched with the addition of 15mM dithiotreitol. Blocked lysates were then chloroform-methanol precipitated.

Precipitated proteins were resuspended in 200mM EPPS pH 8.5 and placed in an orbital shaker at room temperature for overnight digestion with Lys-C at a 1:100 protease:protein ratio (Wako). Sequencing grade trypsin (Promega) was added to the Lys-C digest and incubated for 6hrs on an orbital shaker at 37°C. A Quantitative Colorometric Peptide assay (Pierce) was then performed to measure the concentration of digested peptides present.

At this point, the human sample was split into 2 aliquots of 30µg each. Three µg of yeast peptides were added to one of the human samples, while an equal volume of HPLC grade water was added to the other sample. Both samples were independently de-salted using stop-and-go-extraction tips containing a C18 solid phase[18]. De-salted samples were then dried in vacuum centrifuge and resuspended in a 5% acetonitrile, 5% formic acid buffer for LC-MS/MS analysis.

Liquid Chromatography and Tandem Mass Spectrometry (LC-MS/MS)

Both unfractionated samples were each analyzed 20 times consecutively on an Orbitrap Fusion Lumos mass spectrometer operated in positive-mode with a Proxeon EASY-nLC 1200 liquid chromatograph (Thermo Fisher Scientific) as described previously[2]. Peptide fractionation was performed on a 100 µm inner diameter microcapillary column packed with 35cm of Accucore C18 resin (2.6 µm, 150 Å, Thermo Fisher Scientific). Approximately 1µg of peptide was loaded onto the column for LC-MS/MS analysis.

Separation occurred across a 90min gradient from 4% to 35% acetonitrile in 0.125% formic acid. The flow rate was set to 525nL/min over the gradient. To prevent carry over, the 20 analyses of the human only sample (H) were queued first followed by the 20 analyses of the human+10% yeast spike-in sample (HY).

A data dependent Top Speed (3 second) method was used to collect spectra: high resolution MS1 spectra (Orbitrap resolution: 120,000; mass range: 350-1,400Th; and automatic gain control (AGC) target: $4x10^5$; maximum injection time 50ms) and high resolution MS2 spectra (Quadrupole isolation window: 1.6Th; Orbitrap resolution: 7,500; HCD energy: 30%; AGC target: $5x10^4$; maximum injection time: 22ms). Dynamic exclusion was enabled with a duration time of 120 s.

Data Analysis

Spectra collected from our 40 LC-MS/MS analyses were analyzed with the MaxQuant software package (Version 1.6.3.4)[2,11]. Spectra were searched against a concatenated human (Uniprot ID: UP000005640, Downloaded November 8, 2018) and yeast (Uniprot ID: UP000002311, Downloaded November 8, 2018) database. Database reversal for false discovery rate determination using the target-decoy method was performed by MaxQuant[19]. To prevent bias, MaxQuant default parameters were used. All 40 LC-MS/MS runs were analyzed simultaneously and given a unique Experiment tag in the "Raw data" upload section of MaxQuant. LFQ was enabled with default settings.

LC-MS/MS runs were analyzed once without and once with MBR enabled. Proteins containing even a single shared peptide between human and yeast databases were removed to avoid artificially increasing the amount of yeast identifications due to common peptides between species (n = 77). To account for leucine/isoleucine isomers, a regular expression search was

conducted for peptides containing those amino acids, allowing them to match sequences containing either form of the isomers from both databases. MBR was enabled through the "Identification" subtab in the "Global Parameters" tab of MaxQuant. The default settings for MBR were used (0.7min match window and 20min alignment time).

Extracted Ion Chromatograms (XICs) of peptides were viewed using the Skyline client[20]. A peptide list was imported into the software and the "msms.txt" file generated from the MaxQuant software suit was supplied along with the raw files used in this study. Default Skyline settings were used. For Skyline analysis, raw files were aligned and calibrated within Skyline. All graphs generated by Skyline for a given XIC were synchronized such that the x- and y-axes were identical.

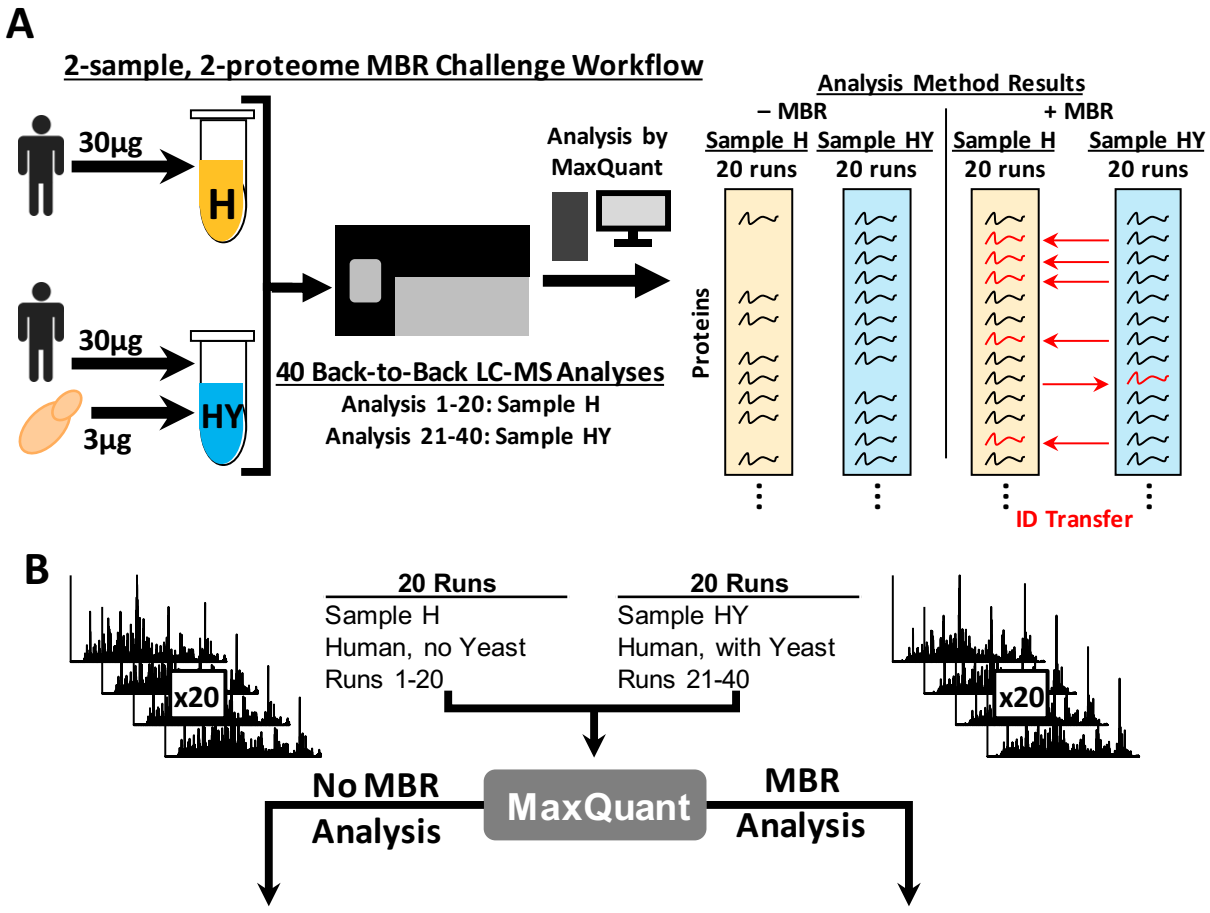
Results

Assessing Experimental Stochasticity

We devised a two-sample, two-proteome system consisting of a pure sample and a mixed sample to measure how frequently the MBR algorithm incorrectly transfers an identification. The pure sample, Sample H, consisted of a human-only peptide mixture from a digested HCT116 cellular lysate. Sample HY, the mixed sample, was an identical aliquot of Sample H with an additional spike-in of yeast (S. cerevisiae) at 10% of the human lysate mass (Figure 3-1A). Forty, 90-min, back-to-back LC-MS/MS analyses were performed such that the first 20 analyses were conducted on Sample H (single-proteome sample) and the last 20 analyses were on Sample HY (two-proteome sample) to avoid sample carryover. As such, Sample H serves as ground truth as it is devoid of yeast proteins. The raw data were analyzed twice by MaxQuant, once with MBR off and again with MBR on to assess how many yeast proteins would be identified by matching in the single-proteome sample (Figure 3-1B).

Figure 3-1: Experiment setup and overview. **A)** Diagram of experimental workflow. Sample H (yellow) and HY (blue) were generated from the same human lysate. Sample HY was only differentiated by a spike-in of yeast whole cell lysate at 10% of the human lysate mass. Both samples were analyzed 20 times each by LC-MS/MS back-to-back on the same column with analysis of Sample H performed first to prevent carry over. MaxQuant was used to process the LC-MS/MS data with and without MBR. **B)** Diagram summarizing the results of the 40 MS/MS runs by MaxQuant analysis type.

Figure 3-1 (continued)



Summary – No MBR

	Species	Total IDs
Peptides[a] (By MS/MS)	Total	938,148
	Human	895,396
	Yeast	42,752
Proteins[b]	Total	4,614
	Human	4,122
	Yeast	492

Summary – With MBR

	Species	Total IDs	% ↑
Peptides[a] (By MS/MS or by matching)	Total	1,342,413	43
	Human	1,281,350	43
	Yeast	61,063	43
Proteins[b]	Total	4,614	0
	Human	4,122	0
	Yeast	492	0

[a] Peptides that overlapped the human and yeast databases were marked as human. Peptides were blasted against the human database and allowed to match either leucine or isoleucine isomers.

[b] Proteins containing peptides that overlapped the human and yeast databases were removed from analysis to remove confounding data. Protein groups which included human proteins were marked as human.

To summarize the complete data set and assess stochasticity, we examined the number of total and unique peptide identifications over all MS/MS analyses performed during this study (Figure 3-1B and Supplemental Table 1). Without MBR enabled, 938,148 peptides were identified over the 40 MS/MS analyses split between 895,396 human identifications compared to 42,752 yeast identifications. This translated into 4,122 human and 492 yeast proteins. Note that any yeast protein which shared one or more peptides with a human protein was removed. By enabling MBR, peptide identifications were increased by 43% equally split between both species. Furthermore, no protein identifications were added as MBR does not increase total protein identifications within the dataset.

Match Between Runs Aids in the Completion of Data Sets

To evaluate the effect of MBR on multiple MS/MS experiments, we first looked to establish a baseline identification rate by assessing the number of unique peptides and proteins identified, by species, in each of the 40 MS/MS analyses (Figure 3-1B, Table 3-1, Figure 3-2). Without MBR, only 57% and 56% of all observed human peptides were detected in each MS/MS analysis of Sample H and HY, respectively (Figure 3-2). A similar per run identification rate of 59% was observed for yeast peptides in the 20 MS/MS analyses of Sample HY. By enabling MBR, the peptide level identification rate for human peptides improved to an average of 81% across all 40 analyses, while the average identification rate of yeast peptides in the 20 analyses of Sample HY increased to 83%.

A more holistic analysis was conducted at the protein level (Table 3-1). On average, 3,738 of the 4,614 proteins (81%), across both species, were identified in each MS/MS run. By species, an average of 85% of the 4,122 human proteins were identified per run. This trend was

Table 3-1: The effect of Match-Between-Runs (MBR) on protein identifications per run

			No MBR		With MBR		
Sample	**n**	**Species**	**Protein IDs**	**% Total[a]**	**Protein IDs**	**% Total[a]**	**Fold ↑[b]**
All Runs	40	Total[c]	3,738 ± 190	81 ± 4	4,290 ± 195	93 ± 4.2	1.15 ± 0.01
		Human	3,521 ± 47	85 ± 1	4,023 ± 18	98 ± 0.4	1.14 ± 0.01
		Yeast	217 ± 212	44 ± 43	267 ± 209	54 ± 42	4.49 ± 3.61
H	20	Total[c]	3,554 ± 40	77 ± 1	4,097 ± 10	89 ± 0.2	1.15 ± 0.01
(Human only)		Human	3,546 ± 40	86 ± 1	4,037 ± 9	98 ± 0.2	1.14 ± 0.01
		Yeast	8 ± 2	1.6 ± 0.4	60 ± 5	12 ± 0.9	7.87 ± 1.64
HY	20	Total[c]	3,922 ± 43	85 ± 1	4,482 ± 14	97 ± 0.3	1.14 ± 0.01
(Human with Yeast)		Human	3,496 ± 40	85 ± 1	4,009 ± 14	97 ± 0.3	1.15 ± 0.01
		Yeast	426 ± 6	87 ± 1	473 ± 3	93 ± 0.7	1.11 ± 0.01

[a]Percent total is calculated with respect to protein identification counts in Figure 1

[b]Fold increase is calculated per run as the ratio of protein identifications with MBR to that without MBR

[c]Total protein is the sum of human and yeast protein (4,122 + 492)

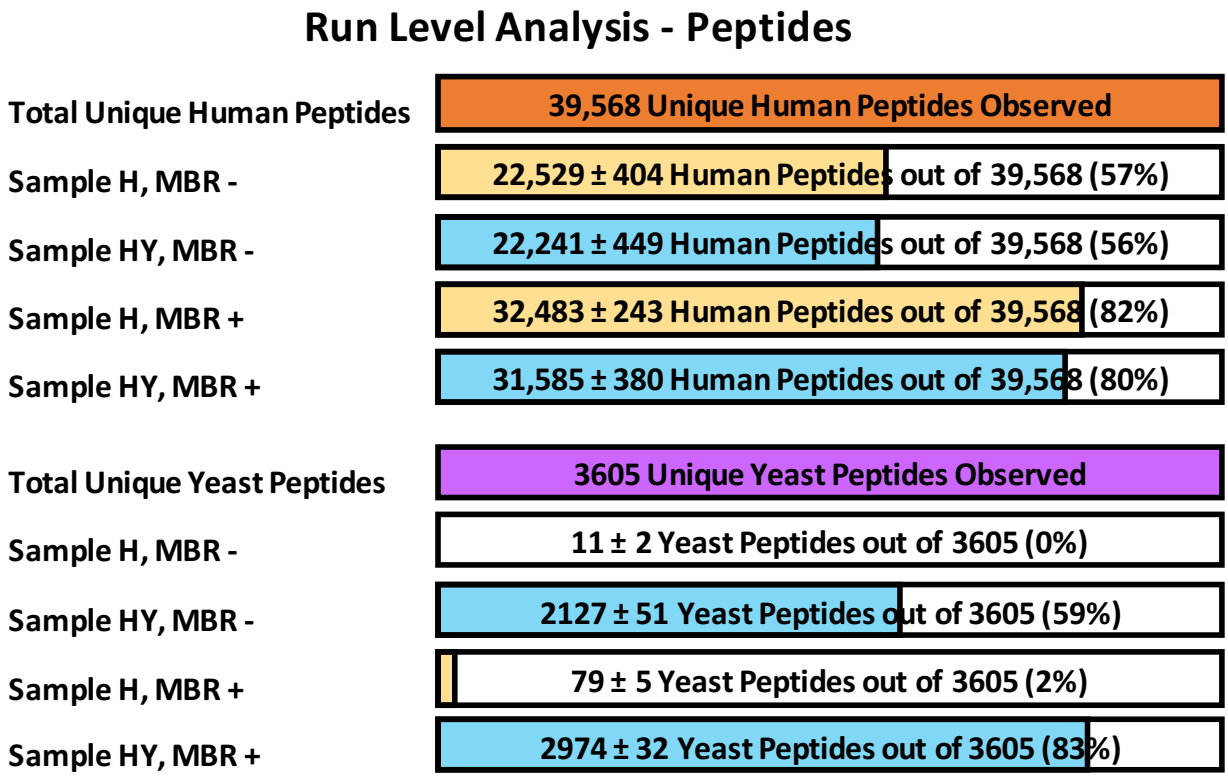


Figure 3-2: Run level analysis of MBR false transfers. False transfers are shown as a percentage of total identifications at the peptide level. Total unique proteins and peptides reported are the combined unique protein list between all 40 LC-MS/MS runs performed in this study from the human (orange bar) and yeast (purple bar) proteomes. Yellow bars indicate percentage identified in analyses of Sample H while blue bars represent percentage identified in Sample HY.

preserved when analyzing Sample H and HY separately. An equivalent average identification rate was observed in Sample HY for yeast proteins (87% of the 494 were identified) while only 1.6% of the yeast proteins were identified in Sample H.

Assessing false transfer rates using a two-proteome model

With the baseline case established, we repeated the analysis with MBR enabled to measure the increase in average identification rate (Table 3-1). Globally, an average of 4,290 proteins (or 93%) were identified per run with 98% of all human proteins being identified per run (up from 85% without MBR) showing an increase in completeness of the dataset. Between samples H and HY, this identification rate was preserved for the human proteins – 98% and 97% respectively.

To measure the false transfer rate, we next looked at the yeast proteins transferred between samples H and HY. For Sample HY, the same completeness trend observed for human proteins was found for yeast proteins with an average of 93% of the 492 yeast proteins identified in each analysis. However, the number of yeast proteins identified on average in Sample H increased to 60 proteins out of 492, or 12%. This is an average 7.87-fold increase of yeast protein identifications in a human-only sample.

Furthermore, we utilized the Skyline suite to assess the differences between the XICs of yeast peptide identifications transferred to Sample H by the MBR algorithm. By comparing the yeast peptide XICs from the 20 LC-MS/MS analyses of Sample H to the 20 from Sample HY we were able to observe that while some incorrectly transferred peptides had little to no signal in the calibrated retention time window in the 20 LC-MS/MS, others contained various incorrect peaks with the appropriate m/z and charge state. For example, yeast peptide IIDDDVPTILQGAK

Figure 3-3: Extracted Ion Chromatograms for peptide IIDDDVPTILQGAK across 40 runs. **A)** Sample H and **B)** Sample HY. XICs of transferred peptide identifications are boxed in red while XICs of peptide identifications made by MS/MS are boxed in blue. Peptide IIDDDVPTILQGAK belongs to the yeast protein HS104 which was identified in Sample H with MBR but not quantified with LFQ. The MaxQuant msms.txt file with MBR and LFQ enabled was supplied to Skyline to generate XICs for this peptide (±5 PPM). Dashed lines represent the bounds of integration for a peak selected automatically by Skyline. Black arrows show the peak Skyline selected for integration. Monoisotopic precursor and the M+1 and M+2 isotopes are shown in blue, purple, and brown respectively in each Skyline plot.

Figure 3-3 (continued)

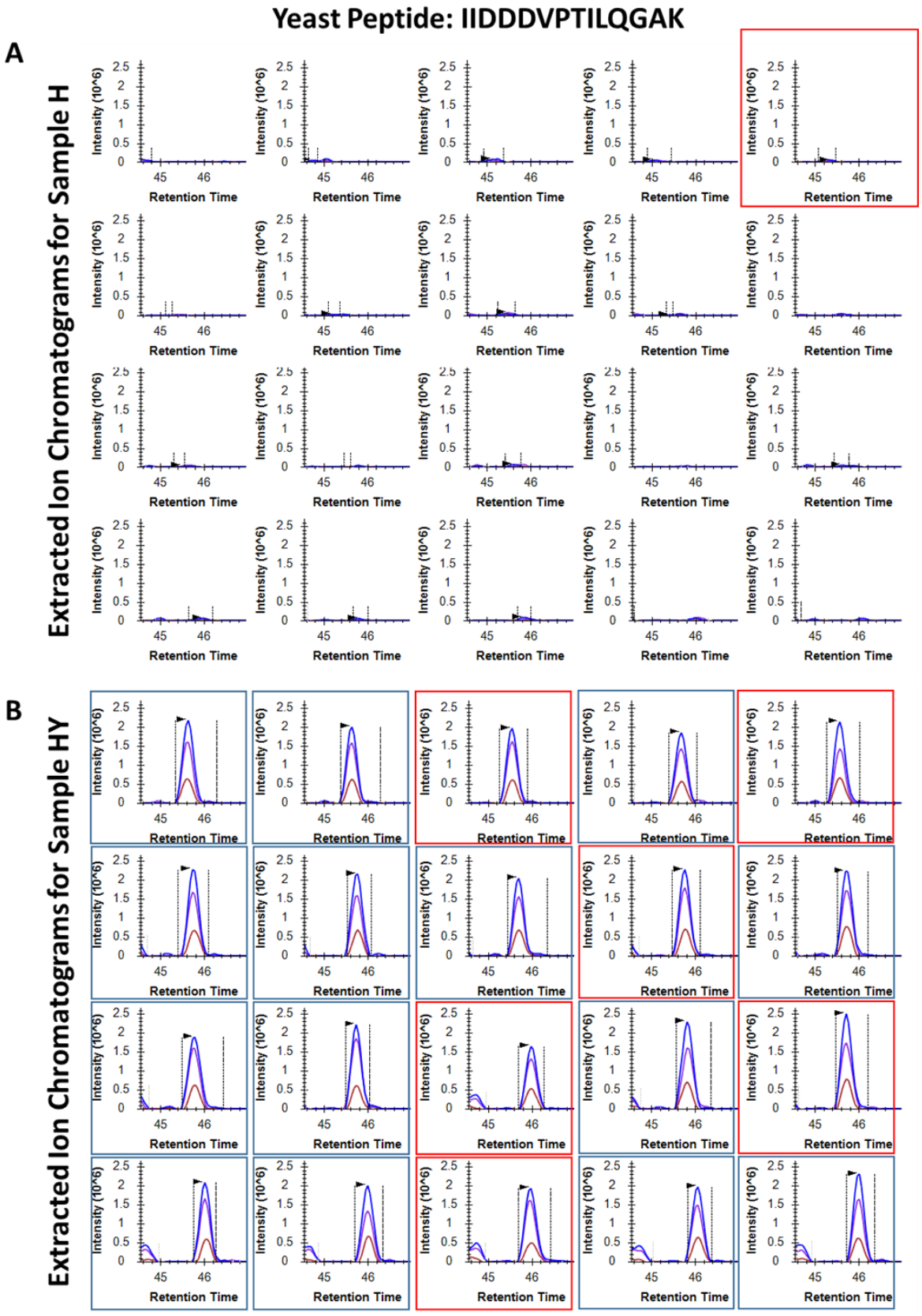
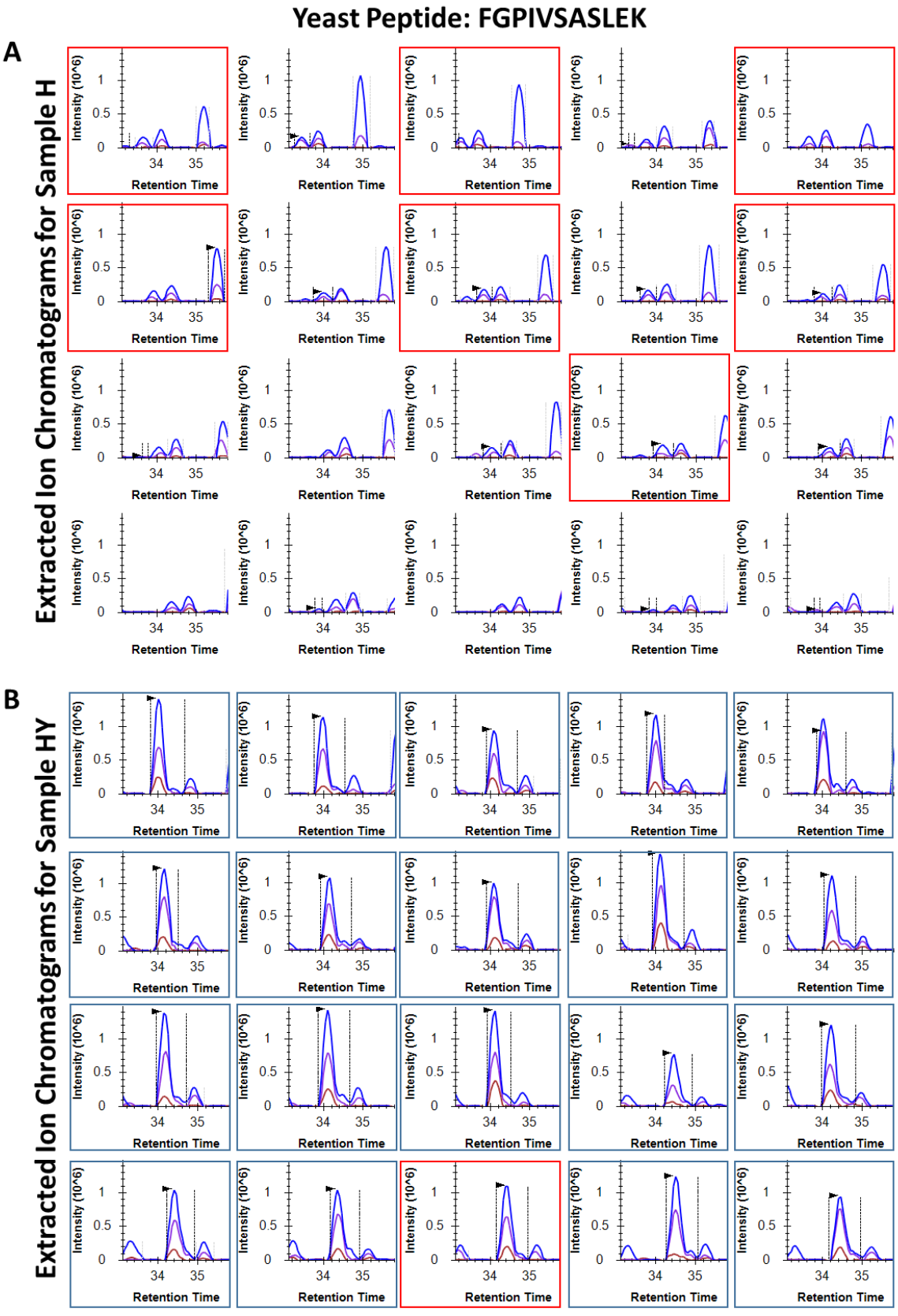


Figure 3-4: Extracted Ion Chromatograms for yeast peptide FGPIVSASLEK across 40 runs. **A)** Sample H and **B)** Sample HY. XICs of transferred identifications are boxed in red. Peptide FGPIVSASLEK belongs to the yeast protein PABP which was identified in Sample H with MBR and quantified with LFQ in 6 of the 20 Sample H replicates. Even though Sample H contains no yeast proteins, peaks sharing the same precursor M/z value can be found near the calibrate retention time for FGPIVSASLEK. This highlights the potential difficulty the MBR algorithm faces when attempting to assign an identification transfer. The MaxQuant msms.txt file with MBR and LFQ enabled was supplied to Skyline to generate XICs for this peptide (±5 PPM). Dashed lines represent the bounds of integration for a peak selected automatically by Skyline. Black arrows show the peak Skyline selected for integration. Monoisotopic precursor and the M+1 and M+2 isotopes are shown in blue, purple, and brown respectively in each Skyline plot.

Figure 3-4 (continued)



contained no signal in Sample H while peptide FGPIVSASLEK contained several peaks within the identified retention time window (Figures 3-3 and 3-4).

Requiring Identifications in Multiple Runs Reduces Spurious Identifications

Next, we assessed how often any given yeast peptide or protein was identified in the 20 LC-MS/MS runs of Sample H (Figure 3-5A). Without MBR, only 57 yeast peptides corresponding to 37 proteins were identified in the 20 analyses. When requiring identification in at least 5 runs, this number reduced to 10 proteins. However, analysis with MBR resulted in the identification of 215 yeast proteins in the 20 samples containing only human proteins. Furthermore, when requiring identification in at least 5 runs, 76 yeast proteins were still identified. However, of the 215 yeast proteins identified in at least one MBR enabled analysis of Sample H, 167 transferred identifications were subsequently reported as having 0 intensity by the LFQ algorithm, 36 proteins had been previously identified without MBR, and 12 yeast proteins were transferred and quantified by LFQ as having a non-0 intensity (Figure 3-5B). As such, while 44% of all yeast proteins were identified in at least one run with MBR enabled, only 2.4% received an intensity value greater than 0 from the downstream LFQ algorithm. This is due to the default setting of the LFQ algorithm requiring a minimum of 2 peptides per protein for quantitation.

Incorrect identifications by Match-Between-Runs are quantified near-0 when LFQ is implemented

To evaluate the effect MBR has on label-free quantitation, we further assessed how the MaxQuant LFQ algorithm handled the 12 yeast protein identifications receiving LFQ intensity values greater than 0 when transferred to Sample H. These proteins were not identified by MS/MS during any of the Sample H analyses, but each identification was transferred, on

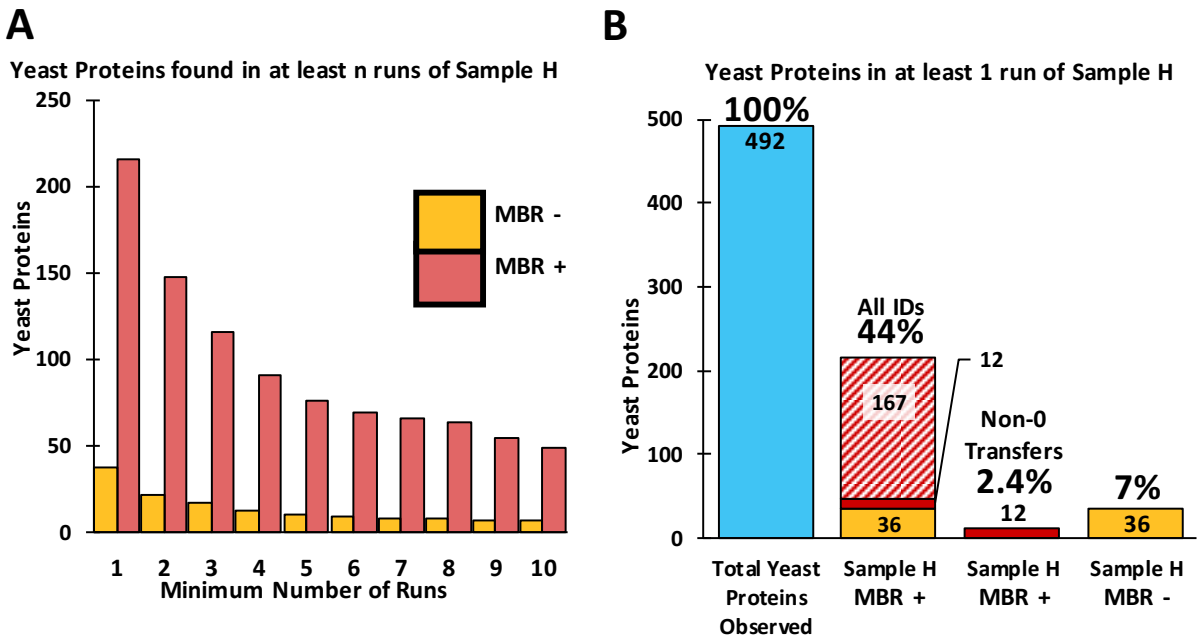


Figure 3-5: Analysis of yeast protein identifications occurring in sample H (no yeast). **A)** Bar plot showing the number of yeast proteins found in *n* or more runs. **B)** Stacked bar plots showing the type of yeast protein identifications occurring in at least one run from sample H. Blue bar shows the number of yeast proteins observed across the 40 MS/MS runs. Yellow bars represent yeast proteins always identified irrespective of MBR. Diagonally dashed bar depicts the number of identifications transferred by MBR but then removed by the downstream LFQ algorithm. Solid red bars show transfers remaining after MBR and downstream LFQ analysis.

average, to 12 analyses of Sample H (Table 3-2). Conversely, these 12 yeast proteins were identified by MS/MS in all 20 analyses performed on Sample HY.

Despite many Sample H analyses receiving an identification of the 12 proteins by matching, only an average of 2 Sample H analyses per yeast protein contained a transfer that would be quantified as non-0 during LFQ analysis. When analyzing the average LFQ ratio of these quantifiable transfers, an average H:HY ratio was 0.18 – indicating that on average, these non-0 assignments at the LFQ level resulted in a quantitation ratio near 1:5. The largest average ratio was 1:2 and observed in the protein PURA, transcriptional activator protein Pur-alpha (Table 3-2). Meanwhile elongation factor 3A, EF3A, and the alpha-tubulin folding protein, had the smallest ratio of 1:100. However, when including transfers that were quantified with an intensity of 0 by LFQ, the average H:HY LFQ ratio of the 12 proteins is near 2:100 indicating a negligible identification.

Discussion

Match between runs is popular approach which partially addresses the problem of stochastic identifications in LC-MS/MS by leveraging chromatographic data. No standardized method, however, is available to assess the false transfer rate when utilizing the MBR algorithm in MaxQuant. Here we present a novel method that utilizes single-proteome and dual-proteome samples to quantitatively measure false transfers. While our data set is biased by primarily focusing on assessing the presence or absence of low abundant proteins, it does mimic the exclusivity of rare proteins when comparing between treatments or tissues.

Table 3-2: Overview of 12 yeast proteins remaining after MBR and LFQ analysis in Sample H

	Sample H			Sample HY	Average LFQ Ratio H:HY	
Protein Name	**MS/MS ID**	**Matching ID**	**LFQ Quantified***	**MS/MS ID**	**LFQ Quantified***	**All Transfers for Protein**
PABP	0	19	6	20	0.21	0.064
HXKA	0	18	2	20	0.09	0.009
RL6B	0	19	1	20	0.18	0.009
FAS1	0	11	1	20	0.03	0.002
DHE4	0	15	3	20	0.32	0.048
G6PI	0	11	1	20	0.08	0.004
ALF	0	4	1	20	0.01	0.000
EF3A	0	5	1	20	0.01	0.000
PFKA2	0	11	1	20	0.14	0.007
FAS2	0	14	2	20	0.27	0.027
UGPA1	0	8	2	20	0.36	0.036
PURA	0	4	1	20	0.50	0.025
Average	0	12	2	20	0.18	0.019

*LFQ Quantified is defined as a protein receiving an intensity greater than 0 during LFQ analysis.

Our findings suggest that, on average, an ~8-fold increase in incorrect identifications can occur at the protein level when allowing the algorithm to perform matches between 40 independent MS/MS runs – 20 of a single proteome sample and 20 of a dual-proteome sample. These spurious identifications are often "one-hit-wonders" and are a result of non-systematic transfers of peptide identifications.

Despite these shortcomings, MBR is not without its merits. We found that in identical samples, MBR increased the number of peptides identified by an average of 43%. This increase was reduced to 15% at the protein level due to the assignment of multiple peptides to a single protein. As such, 98% of all detected human proteins were observed in all 40 MS/MS runs, up from the initial 86% identification rate without MBR. The result of enabling MBR in identical samples is near-complete identification of all observed proteins in the data set, which alleviates the missing-value problem. Furthermore, enabling LFQ in the MaxQuant algorithm resolved most spurious matches due to false transfer. Although a small subset remained after LFQ analysis, on average these represent a 1:50 ratio between the human-only (H) and human + yeast sample (HY). Until further developments to the MBR algorithm to measure, quantify, and control the false transfer rates are added, it is recommended that utilizing MBR in conjunction with some form of post-processing software that can address false transfers (i.e., LFQ).

Supporting Information

Formatted tables of MaxQuant outputs are provided as an XLSX file. RAW files and data for LC-MS/MS experiments have been uploaded to the ProteomeXchange Consortium via the PRIDE partner repository[21]. The dataset identifier is PXD014415.

References

1. Stead, D. A. *et al.* Information quality in proteomics. *Brief. Bioinform.* **9**, 174–188 (2008).

2. O'Connell, J. D., Paulo, J. A., O'Brien, J. J. & Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* **17**, 1934–1942 (2018).

3. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).

4. O'Brien, J. J. *et al.* The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann. Appl. Stat.* **12**, 2075–2095 (2018).

5. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).

6. Thompson, A. *et al.* Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).

7. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* **84**, 7469–7478 (2012).

8. Dephoure, N. & Gygi, S. P. Hyperplexing : A Method for Higher-Order Multiplexed Quantitative Proteomics Provides a Map of the Dynamic Response to Rapamycin in Yeast. *Sci. Signal.* **5**, 1–9 (2012).

9. Braun, C. R. *et al.* Generation of Multiple Reporter Ions from a Single Isobaric Reagent Increases Multiplexing Capacity for Quantitative Proteomics. *Anal. Chem.* **87**, 9855–9863 (2015).

10. Wiberg, H. K. *et al.* Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J. Proteome Res.* **14**, 1993–2001 (2015).

11. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

12. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

13. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, 1–16 (2019).

14. Bielow, C., Mastrobuoni, G. & Kempa, S. Proteomics Quality Control: Quality Control Software for MaxQuant Results. *J. Proteome Res.* **15**, 777–787 (2016).

15. Lim, M. Y., O'Brien, J., Paulo, J. A. & Gygi, S. P. Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling. *J. Proteome Res.* **16**, 4217–4226 (2017).

16. Paulo, J. A., O'Connell, J. D. & Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J. Am. Soc. Mass Spectrom.* **27**, 1620–1625 (2016).

17. Paulo, J. A. & Steven, P. A comprehensive proteomic and phosphoproteomic analysis of yeast deletion mutants of 14-3-3 orthologs and associated effects of rapamycin. *Proteomics* **15**, 474–486 (2015).

18. Ishihama, Y., Rappsilber, J. & Mann, M. Modular Stop and Go Extraction Tips with Stacked Disks for Parallel and Multidimensional Peptide Fractionation in Proteomics. *J. Proteome Res.* **5**, 988–994 (2006).

19. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

20. MacLean, B. *et al.* Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

21. Perez-riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019 : improving support for quantification data. *Nucleic Acids Res.* **47**, 442–450 (2019).

Chapter 4: Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling

Attributions: The following chapter was previously published on October 6, 2017 in the *Journal of Proteome Research.* Reproduced with permission from "Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling"
Copyright 2017 American Chemical Society.

Matthew Y. Lim performed experiments, conducted data analysis, and wrote the manuscript.

Jonathon O'Brien designed the statistical framework.

Joao A. Paulo provided samples and conducted mass spectrometry instrumentation.

Steven P. Gygi provided the initial concept for the project.

Abstract

Phosphorylation stoichiometry, or occupancy, is one element of phosphoproteomics that can add useful biological context[1]. We previously developed a method to assess phosphorylation stoichiometry on a proteome-wide scale[2]. The stoichiometry calculation relies on identifying and measuring the levels of each nonphosphorylated counterpart peptide with and without phosphatase treatment. The method, however, is problematic in that low stoichiometry phosphopeptides can return negative stoichiometry values if measurement error is larger than the percent stoichiometry. Here, we have improved the stoichiometry method through the use of isobaric labeling with 10-plex TMT reagents. In this way, 5 phosphatase treated and 5 untreated samples are compared simultaneously so that each stoichiometry is represented by 5 ratio measurements with no missing values. We applied the method to determine basal stoichiometries of HCT116 cells growing in culture. With this method, we analyzed 5 biological replicates simultaneously with no need for phosphopeptide enrichment. Additionally, we developed a Bayesian model to estimate phosphorylation stoichiometry as a parameter confined to an interval between 0 and 1 implemented as an R/Stan script. Consequently, both point and interval estimates are consistent with the plausible range of values for stoichiometry. Finally, we report absolute stoichiometry measurements with credible intervals for 6,772 phosphopeptides containing at least a single phosphorylation site.

Introduction

Phosphorylation is one of the most common post-translation modifications found in cells. By chemically attaching a phosphate group to amino acid residues such as serine, threonine, and tyrosine, cells can change a protein's function, localization, or degradation in addition to other important cellular activities including signal transduction[3,4]. Because of its many cellular

functions, a variety of experiments have been designed to probe different elements of phosphorylation. Quantitative experiments such as Western blotting and phosphopeptide mass spectrometry analysis are often implemented to measure phosphorylation dynamics[3]. While generally useful, these methods are often limited to identifying fold changes which may not provide sufficient information to fully understand the underlying biological mechanism.

One facet of quantitative phosphorylation proteomics that can have potential biological insight is phosphorylation stoichiometry, or occupancy[1]. A measured fold change of 2 for a phosphopeptide's levels can be caused a by a multitude of different cellular processes: a doubling in a protein production, a doubling in a phosphorylation occupancy, a decrease in protein degradation of the nonphosphorylated version, or any number of other cellular events. Additionally, a 2 fold increase in relative phosphorylation levels can mean anything from an increase of 2% to 4% overall occupancy to an increase of 50% to 100% occupancy. Such stark differences in the absolute amount of phosphorylation occupancy could suggest that different cellular processes are activated in response to stimuli at different phosphorylation stoichiometries[2,5–8].

Traditionally, phosphorylation stoichiometry has been measured using low throughput methods such as quantitative western blotting. In 2001, we used AQUA peptides as absolute internal standards to measure the absolute amounts of both the phosphorylated and nonphosphorylated forms of site Ser-1126 on the protein separase[1]. We showed that this site was held at very high stoichiometry until the anaphase-metaphase transition, whereupon it became dephosphorylated, releasing its protease activity to finish mitosis. In recent years, we and others have developed high-throughput whole proteome techniques to assess phosphorylation stoichiometry *en masse*[2–5,9]. All current global proteome methods suffer from the same

unavoidable drawback – their inability to distinguish phosphorylation stoichiometry of individual sites in multiply phosphorylated peptides. As such, it can only be claimed that for a multiply phosphorylated peptide, this is the maximum possible stoichiometry considering all sites. One such method utilizes stable isotope labelling with amino acids in cell culture (SILAC) to measure three distinct ratios to generate a phosphorylation stoichiometry measurement[5]. However, SILAC can only be utilized where heavy amino acids can be doped into cell culture. Additionally, the SILAC method for assessing phosphorylation stoichiometry can only detect stoichiometry for sites that undergo a change in stoichiometry based on two conditions[5].

We have previously published a method utilizing phosphatase treatment to assess phosphorylation stoichiometry[2]. A sample is divided into two aliquots, chemically labelled with a unique label, and one is treated with phosphatase. Phosphorylation stoichiometry can be assessed by analyzing the increase in signal of the non-phosphorylated form of phosphopeptides after phosphatase treatment[2,8]. Calculated stoichiometries are then assigned to phosphopeptides by matching the stoichiometries of these non-phosphorylated peptides to their known phosphorylated form from a phosphopeptide database or a previously generated phosphopeptide library. This indirect measurement circumvents issues of phosphorylation enrichment efficiencies as well as ionization efficiency for phosphorylated peptides, and potential digestion problems related to analyzing phosphorylated peptides[2,6,8]. Importantly, no comparison to a second condition is required allowing for the basal phosphorylation stoichiometry of a cell to be assessed. Others have adapted our method further for iTRAQ labelling or kinase treatment to improve this phosphatase-based method[6,7].

This method can, however, report negative stoichiometries. For example, if the true occupancy level is 2% but the measurement error is 5%, it is possible to calculate negative

values. To our knowledge, no group has successfully addressed the negative stoichiometries resulting from measurement error. Furthermore, previous attempts at analyzing phosphorylation stoichiometry relied on sample standard deviations to calculate confidence intervals for each stoichiometry measurement[2,5–7]. These intervals frequently include stoichiometry values below 0% or above 100%, which are not possible. Fortunately, these issues can be resolved by carefully defining a statistical model with appropriate distributions and ranges.

TMT reagents are a conduit for sample multiplexing in quantitative proteomics[10–13]. TMT chemically modifies the N-terminus and all free lysine residues of a peptide and is commercially available as a 2-, 6-, 8-, and 10-plex[10,12,14]. Each label is divided into two regions, a reporter ion region and a mass balance. All labels have the same nominal mass, but differ in the placement of heavy ^{13}C and ^{15}N atoms, distributed between the reporter ion and mass balance regions[10]. TMT labelled peptides are, thus, indistinguishable during chromatographic separations and even via MS1 analysis[10,12,14,15]. However, during peptide fragmentation in a mass spectrometer, the balance remains attached to the peptide while the reporter region falls off as a reporter ion. Each label, or channel, has a unique reporter ion mass. Quantitation is performed by assessing the relative ratios of reporter ions[10,12]. A multi-notch MS3 method can be used to collect accurate reporter ion ratios, greatly reducing or removing completely interference caused by co-eluting and co-fragmenting peptides[15,16].

Here we have extended the TMT workflow to include stoichiometry analysis. We determined absolute stoichiometry from five biological replicates of asynchronously growing HCT116 cells under basal conditions. We used statistical modeling to address negative stoichiometries in our dataset. We treated stoichiometry as an estimable parameter rather than a

directly calculated statistic. Finally, we provide occupancy measurements for 6,772 unique phosphopeptides containing at least one phosphorylation site in HCT116 cells.

Materials and Methods

Cell Culture

HCT116 cells were cultured in DMEM (Gibco) supplemented with 10% (v/v) fetal bovine serum (Hyclone) and 50µL/mL penicillin and 50µL/mL streptomycin (Gibco) in a 15cm dish as described previously[13,17]. Cells were incubated at 37°C at 5% CO2 until approximately 80% confluent. Cells were then washed with ice cold phosphate buffered saline (Gibco) and lysed on plate with 1mL of an 8M urea lysis buffer containing a protease and phosphatase inhibitor cocktail (Roche). Lysate was collected and stored at -80°C until sample preparation for mass spectrometry.

Sample Preparation

HCT116 lysate was homogenized by passing the lysate through a 21-gauge needle followed by sedimentation by centrifugation at 21,000 x g for 15mins[13]. The supernatant was transferred to a new tube and protein concentration was determine by a bicinchoninic acid (BCA) assay (ThermoFisher Scientific). The proteins were then reduced and alkylated to block reactive cysteine groups and chloroform-methanol precipitated. Proteins were resuspended in 200mM EPPS pH 8.5 and digested with Lys-C (Wako) overnight at room temperature and subsequently digested with sequencing grade trypsin (Promega) for 6hrs at 37°C. Digests were then de-salted using C18 solid-phase extraction (SPE) (Sep-Pak, Waters) and dried down in a vacuum centrifuge.

Phosphatase Experiment to Generate Stoichiometry

We adapted our previous phosphatase method[2] to make use of TMT. Briefly, five dried down de-salted digests were resuspended in 100mM EPPS pH 8.5 and separated into two equivalent 50μg aliquots. Each digest corresponded to a biological replicate. Each aliquot was labeled with a TMT10 reagent for 90mins at room temperature and then quenched with hydroxylamine. The quenched reaction was flash frozen and dried down in a vacuum centrifuge and then resuspended in CutSmart Buffer (New England Biolabs) and one labelled aliquot from each replicate was treated with 200 units of Calf Intestinal Phosphatase (New England Biolabs) while the other aliquot from the replicate was treated with water. All aliquots were incubated at 37°C for 3 hours and then acidified with formic acid to a final concentration of 1%. All aliquots were then combined at a 1:1:1:1:1:1:1:1:1:1 ratio[11]. The pooled sample was then subjected to C18 SPE (Sep-Pak, Waters) and then dried down in a vacuum centrifuge before resuspension in 10mM ammonium bicarbonate and 5% acetonitrile for off-line basic pH reversed-phase (BPRP) fractionation.

Phosphopeptide Enrichment Experiment

A separate phosphopeptide enrichment experiment was performed on HCT116 cell lysates to generate a phosphopeptide library as previously described[18]. Briefly, 10mg of protein from HCT116 lysates was digested and subjected to enrichment with immobilized metal affinity chromatography with Fe^{3+} (Fe-IMAC). The phosphopeptide enriched digest was then labelled with a TMT10 reagent as described above. The sample was then dried down in a vacuum centrifuge, resuspended in 1% formic acid and subjected to C18 solid phase extraction (SPE) (Sep-Pak, Waters). The de-salted phosphopeptide enrichment was dried down in a vacuum centrifuge before resuspension in 10mM ammonium bicarbonate and 5% acetonitrile for off-line basic pH reversed-phase (BPRP) fractionation.

BPRP fractionation

Off-line BPRP HPLC was performed on an Agilent 1100 pump with a degasser and photodiode array detector[11]. A gradient of 13%-37% acetonitrile in 10mM ammonium bicarbonate was used over 50min. The pooled TMT-labelled sample and the phosphopeptide enriched sample were each separated into 96 fractions by the instrument. For each fractionation experiment, fractions were collected in a 96-well plate and combined into 24 fractions as previously described[11]. The 24 fractions were acidified to 1% formic acid and dried down in a vacuum centrifuge. Dried down fractions were resuspended in 5% acetonitrile and 5% formic acid for LC-MS/MS analysis.

Liquid Chromatography and Tandem Mass Spectrometry (LC-MS/MS)

Data for all LC-MS/MS experiments were collected on an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA) with LC separation performed on an attached Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). LC-MS/MS method was modified from a previous study[11]. A 100μm inner diameter microcapillary column packed with 35cm of Accucore C18 resin (2.6μm, 150Å, ThermoFisher) was used to separate peptides. Approximately 2μg of peptide were loaded onto the column for analysis.

A 150min gradient of 6% to 25% acetonitrile in 0.125% formic acid was used at a flow rate of ~450nL/min to separate peptides from the pooled TMT-labelled samples: MS1 spectra (Orbitrap resolution 120,000; mass range: 350-1,400 m/z; automatic gain control (AGC) target: $5x10^5$; maximum injection time: 100ms). We then used a Top10 method to select precursors for further downstream analysis. MS2 spectrum were collected after collision-induced dissociation (CID) (AGC target: $2x10^4$; Normalized collision Energy (NCE): 35%; maximum injection time:

120ms; and isolation window of 0.7Th). MS2 analysis was performed in the ion trap. We performed an MS3 analysis for each MS2 scan acquired by isolating multiple MS2 fragment ions that were used as precursors for the MS3 analysis with a multi-notch isolation waveform. We detected the MS3 analysis in the Orbitrap (resolution 50,000) after high energy collision induced dissociation (HCD) (NCE: 65% with instrument parameters: AGC target: $2.5x10^5$; maximum injection time: 150ms; and isolation window of 1.3Th).

For the phosphopeptide enriched sample, a high-resolution MS2 method was utilized for analysis as there was no quantitation to perform. Peptides were again separated by a 150min gradient. MS1 spectra were obtained in the Orbitrap (resolution 120,000; mass range: 350-1400 m/z; AGC target: $5x10^5$; maximum injection time: 100ms). We selected precursors for MS2 analysis using a TopSpeed method of 3sec. MS2 analysis occurred in the Orbitrap as well (HCD fragmentation; NCE: 38%; AGC target: $1x10^5$; maximum injection time: 150ms; isolation window: 1.6Th).

Data Analysis

Spectra acquired from LC-MS/MS experiments for the TMT-pooled phosphatase experiments were processed using a Sequest-based software pipeline[11,19]. First a modified version of ReAdW.exe converted spectra to the mzXML format. These files were then searched against a database which contained the human proteome (Uniprot Database ID: 9606, downloaded February 4, 2014) concatenated to a database of all protein sequences reversed[20]. A precursor ion tolerance of 50ppm and a product ion tolerance of 0.9Da were used as search parameters. Static modifications for TMT tags (+229.163Da) on lysine residues and the peptide's N termini and carbamidomethylation (+57.021 Da) on cysteine residues were used in conjunction with a variable modification for oxidation (+15.995 Da) on methionine.

Peptide-spectrum matches (PSMs) were then filtered using linear discriminant analysis to a false discovery rate (FDR) of 1% as described previously[21]. XCorr, ΔCn, missed cleavages, peptide length, charge state, and precursor mass accuracy were used as parameters for the LDA. The false discovery rate was estimated by using the target-decoy method. Peptides were identified and collapsed using principles of parsimony to a final protein-level FDR of 1%.

For quantitation, we extracted the signal-to-noise (S:N) ratio of the closest matching centroid to the expected mass of the TMT reporter ion for each TMT channel from MS3 scans triggered by MS2 scans. MS3 spectra were filtered for a minimum TMT reporter ion sum S:N of 200 and an isolation specificity of at least 0.5.

Data from the phosphopeptide enrichment were processed similarly except an additional variable modification of phosphate (+79.966) on serine, threonine, and tyrosine residues was included as a Sequest search parameter. Additionally, because the analysis was a high-resolution MS2 scan, product ion tolerance was tightened to 0.03 Da. Site localization was performed using Ascore[22]. No quantitation was performed. The generated localized phosphopeptide list was filtered to remove any duplicate phosphopeptides to create a unique-matchable list.

Filtered PSMs from the phosphatase experiment were then matched to the unphosphorylated form of peptides from the unique-matchable phosphopeptide list. A TMT-based reporter ion quantitation method was then performed on these matched PSMs utilizing the S:N ratios for each reporter ion channel from the phosphatase experiment. To calculate stoichiometry we compared S:N ratios for reporter ion channels corresponding to the same biological replicate. This was done with three different computational approaches, which we will refer to as the standard stoichiometry, 0% lower limit, and Bayesian method. We defined the standard stoichiometry calculation as:

$$\text{Stoichiometry} = \frac{\text{TMT S: N}_{phosphatase\ treated} - \text{TMT S: N}_{untreated}}{\text{TMT S: N}_{phosphatase\ treated}} * 100\%$$

For our 0% lower limit method, the calculation of stoichiometry was identical except that any negative stoichiometry calculated was replaced with 0%. Arithmetic means and sample standard deviations were calculated for both methods across the five biological replicates for each peptide.

An in-house Bayesian modelling program in R/Stan treated stoichiometry as an estimable parameter rather than a statistic. Briefly, to prevent negative estimation of stoichiometry and to generate credible intervals that contain only physically possible numbers (i.e. stoichiometry estimations constrained between 0-100%) we chose to model stoichiometry as a beta distribution – a distribution naturally constrained to the unit interval. Additionally, instead of calculating stoichiometry as a statistic directly from the raw data, we calculated the fraction of S:N contributed by the untreated channel and used this statistic to make inferences about the phosphorylation:

$$\text{S: N Contribution}_{untreated} = \frac{\text{TMT S: N}_{untreated}}{\text{TMT S: N}_{phosphatase\ treated} + \text{TMT S: N}_{untreated}}$$

This statistic is calculated for each pair of TMT channels corresponding to a biological replicate. All calculated values are then fed into our in-house software which then fits the following Bayesian model,

$$y_{ij} \sim Beta(\phi_i, \lambda)$$

$$\phi_i \sim Pareto(0.1, 1.5)$$

$$\lambda = logit^{-1}(\mu_i + \beta_j)$$

$$\mu_i \sim halfNormal(0, 5)$$

$$\beta_j \sim Normal(0, 5)$$

$$S_i = 1 - \frac{logit^{-1}(\mu_i)}{1 - logit^{-1}(\mu_i)}$$

where $i = 1, \dots, n_p$ indexes the n_p phosphorylation sites. $j = 1, \dots, n_t$ indexes the n_t tubes/replicates. y_{ij} represents the observed untreated signal-to-noise contribution, which was defined above, and the Beta distribution here is defined in terms of mean parameters, ϕ_i, and a precision parameter, λ. μ_i represents the true contribution of untreated signal to the i'th site and the β_j's represents tube effects (pipetting error). Finally, S_i represents the true phosphorylation of the i'th site. This is the main parameter of interest. Notice that it is the use of a half-Normal distribution for μ_i that forces stoichiometry between 0 and 1. All prior distributions were selected to be weakly informative.

Bayesian methods gave us the flexibility to pick distributions and domains that place stoichiometry within the correct interval. It is not clear how this would be achieved with frequentist methods. Our Bayesian method is not deterministic and requires simulations to describe the posterior distributions of our parameters. In the domain specific programming language Stan, Markov chain Monte Carlo simulations using Hamiltonian Dynamics, also known as a Hamilton Monte Carlo, achieve this goal. After executing a pre-defined number of simulated draws, 2000 is the default in Stan, we discard the first half (since convergence may not have been achieved) and use the latter to describe the distributions of interest. Here we aim to determine the probability distribution of each stoichiometry, given the observed data. We summarize this distribution with the posterior mean and percentiles that correspond with 80% and 95% credible intervals for each peptide. Additionally, the posterior mean of λ provides a measurement of how much overall variation is seen in the data.

Convergence can be assessed by looking at traceplots which show the values of a parameter after each iteration. In our experiment, we always observed convergence within the first few hundred iterations.

Results

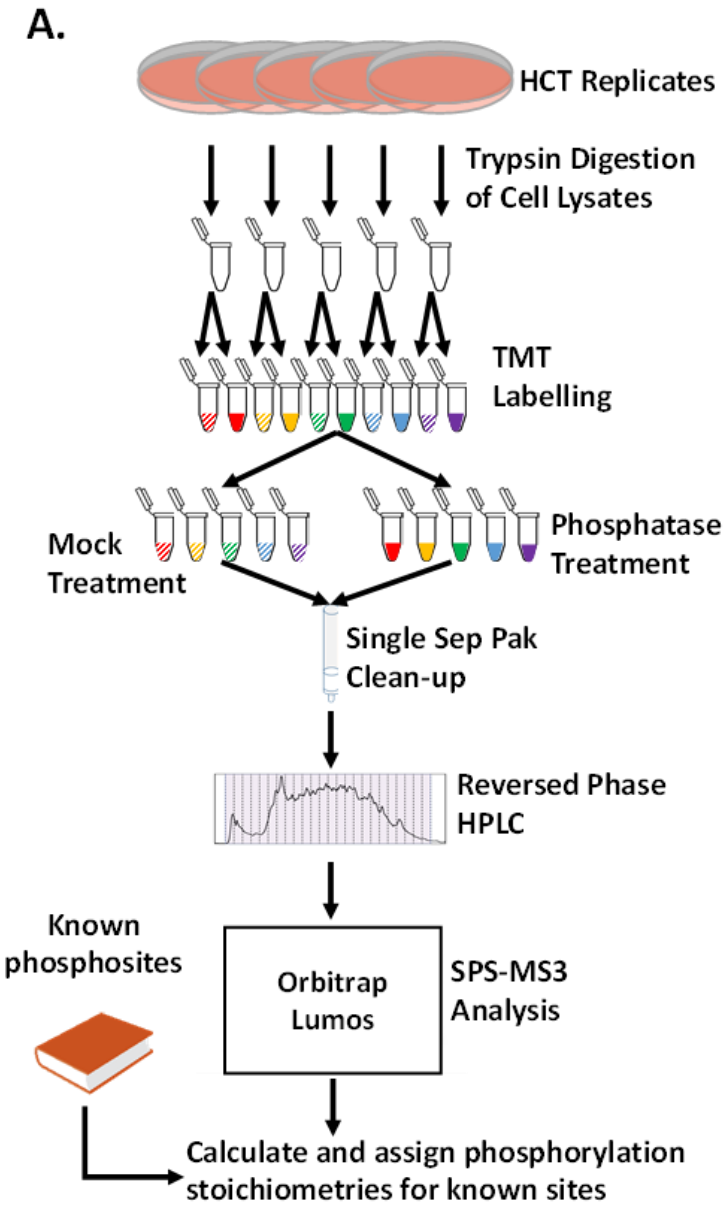
Experiment Workflow

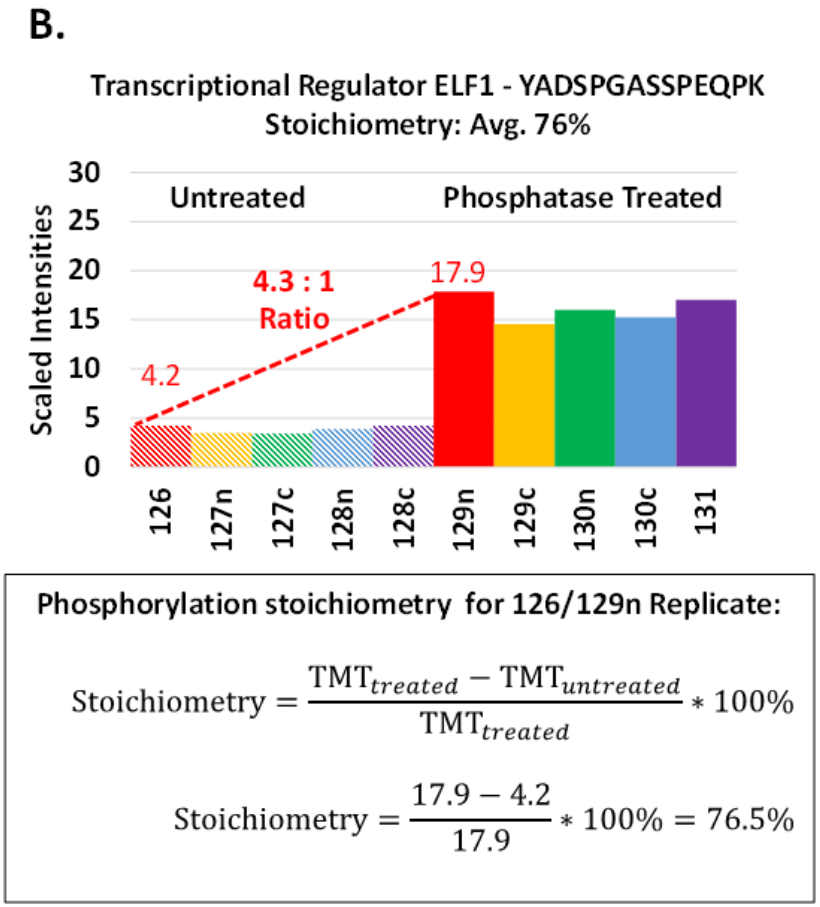
The TMT10-plex workflow for determining phosphorylation occupancy is shown in Figure 1. We chose to implement the workflow using 5 biological replicates of HCT116 cells (Figure 4-1A). To minimize variability, samples were only subjected to individual de-salting columns once, after digestion (before splitting). TMT10 reagent usage was optimized by dividing each replicate into 2 aliquots such that each aliquot received a unique TMT tag. After TMT labelling, all aliquots were dried down in a vacuum centrifuge before reconstitution in phosphatase buffer. All aliquots were recombined for a single de-salting step before off-line BPRP fractionation prior to mass spectrometer analysis. For each biological replicate, phosphorylation stoichiometry was calculated for each peptide whose phosphorylated version could be found in a known library (Figure 4-1A,B). TMT10 enabled us to analyze all 5 biological replicates simultaneously, which was not possible previously.

Generation of a phosphopeptide library found in HCT-116 cells

The first iteration of our phosphatase-based method used a database of known phosphorylation sites found in the literature[2]. Instead of using a literature-based database, we created our own by performing a Fe-IMAC enrichment on confluent HCT-116 cells (Figure 4-2A). We enriched phosphopeptides from 10mg of protein and then separated the enriched sample into 24 fractions by off-line BPRP HPLC. Each fraction was subjected to high-resolution MS2 analysis using HCD fragmentation. We identified over 42,000 unique phosphopeptides that were

Figure 4-1: Outline of TMT-based phosphostoichiometry experiment. **A)** Workflow for phosphorylation stoichiometry experiment. Briefly, reduced and alkylated cell lysate from 5 biological replicates of HCT-116 cells were separately digested with trypsin, and each sample was split into 2 aliquots for TMT-10 labelling. One labelled aliquot from each sample was subjected to phosphatase treatment while its sister aliquot underwent a mock treatment. All 10 aliquots were combined for Sep-Pak clean-up and subjected to reversed phase HPLC and then analyzed by SPS-MS3 on a Thermo Orbitrap Fusion Lumos. Stoichiometries were then calculated for each peptide and assigned to phosphopeptides from a previous independent phosphopeptide identification experiment. **B)** Sample calculation of how stoichiometry is calculated for an observed peptide from our experiment. The stoichiometry for each sample is calculated. In the example shown, the stoichiometry is calculated for the red sample. An equivalent formula is to use the ratio of treated to untreated to calculate the stoichiometry: $1 - \frac{1}{T:U}$.





Phosphorylation stoichiometry for 126/129n Replicate:

$$\text{Stoichiometry} = \frac{\text{TMT}_{treated} - \text{TMT}_{untreated}}{\text{TMT}_{treated}} * 100\%$$

$$\text{Stoichiometry} = \frac{17.9 - 4.2}{17.9} * 100\% = 76.5\%$$

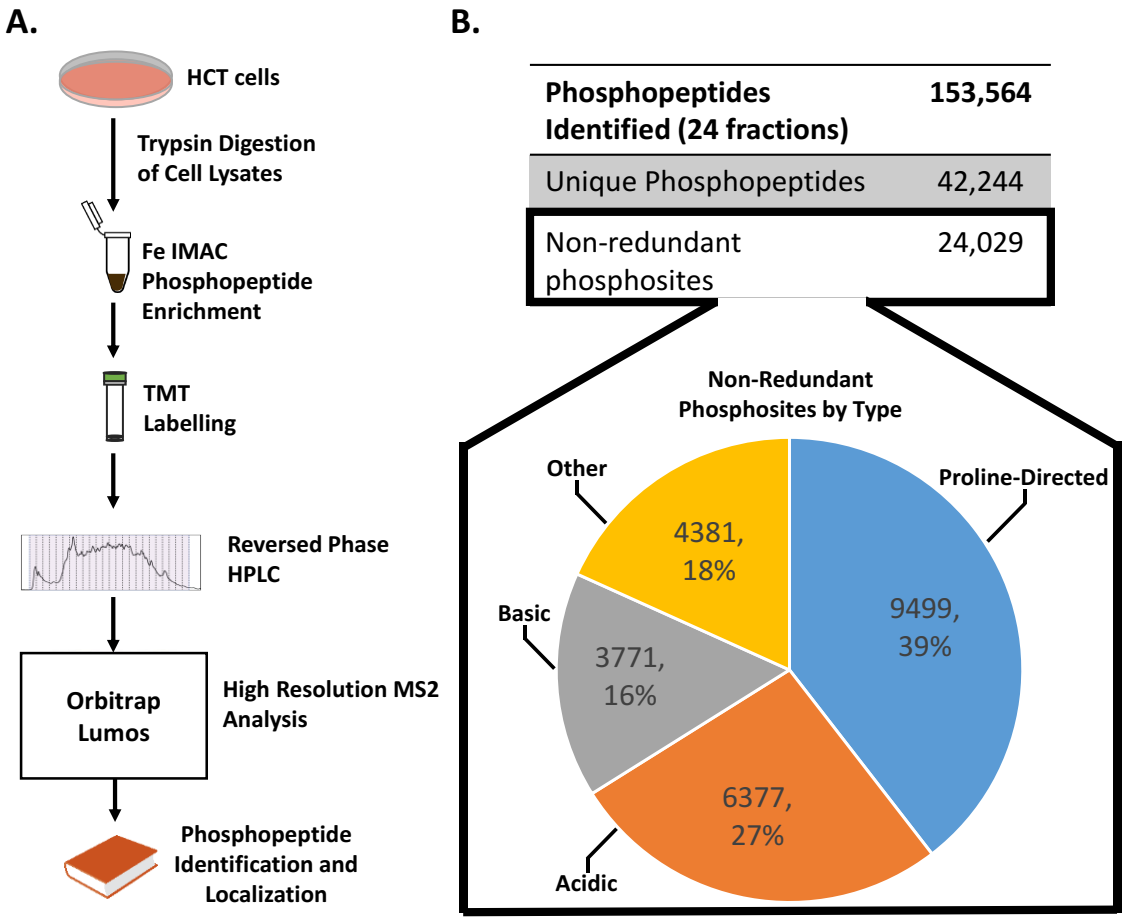


Figure 4-2: Outline of phosphosite library generation experiment. **A)** Workflow for independent phosphopeptide identification experiment. Fe-IMAC enrichment was performed on the digested cell lysate from HCT-116 cells. The phosphopeptide enriched digest was then TMT-labelled to account for chemical changes caused by TMT-labelling and subjected to fractionation by reverse-phase HPLC. Fractions were analyzed by high resolution MS2 analysis. Resulting phosphopeptide identifications were localized to sites using a modified A-score to generate the known phosphorylation sites library used in Fig 1a. **B)** Summary of phosphopeptides identified during this experiment. Pie chart breaks down the phosphopeptides by type: Acidic, Basic, Proline-Directed, and Other. Sites were assigned a type based on a previously described algorithm.

localized to 24,028 sites categorized by type (acidic, basic, proline-directed, other) based on our lab's previous algorithm (Figure 4-2B, Supplemental Table 2)[19]. This dataset was then utilized as the known peptide library (Figure 4-1A). 40% of observed phosphorylation sites were of the proline-directed type, 26% acidic, 19% basic, and 16% did not fall any of the listed categories (Figure 4-2B). After assigning stoichiometry to the matched sites, we observed that sites with an acidic motif were found at higher average stoichiometries (Figure 4-3)[2].

Phosphatase experiment observes 25% from generated phosphopeptide database

We analyzed all 24 fractions of our TMT10 labelled phosphatase-based stoichiometry experiment on an Orbitrap Fusion Lumos instrument. Over 124,000 total peptides were identified, corresponding to 8,351 proteins (Figure 4-4). For consistent quality, we then filtered our data set for peptides with precursor isolation specificity of at least 0.5 and a sum S:N ratio of 200 across the 10 TMT reporter ion channels. This resulted in 72,074 unique peptides being passed for quantification (Figure 4-4). After matching our identified peptides to their phosphorylated forms in our phosphopeptide library, we assigned 6,772 unique peptides a phosphorylation stoichiometry value (Figure 4-4, Supplemental Table 3). The stoichiometries for these peptides were then calculated in the standard method, 0% lower limit method, and our Bayesian modelling method.

Calculating stoichiometries directly from raw data can result in negative values

We first proceeded to calculate stoichiometries for our phosphopeptides using the standard method (Figure 4-1B). We looked at six examples of the phosphorylation calculation that were found in targeted studies according to previous literature (Figure 4-5A)[23]. While the averages of the five replicates were all physically possible (between 0% and 100% stoichiometry), we noticed that the individual stoichiometry measurements for each replicate

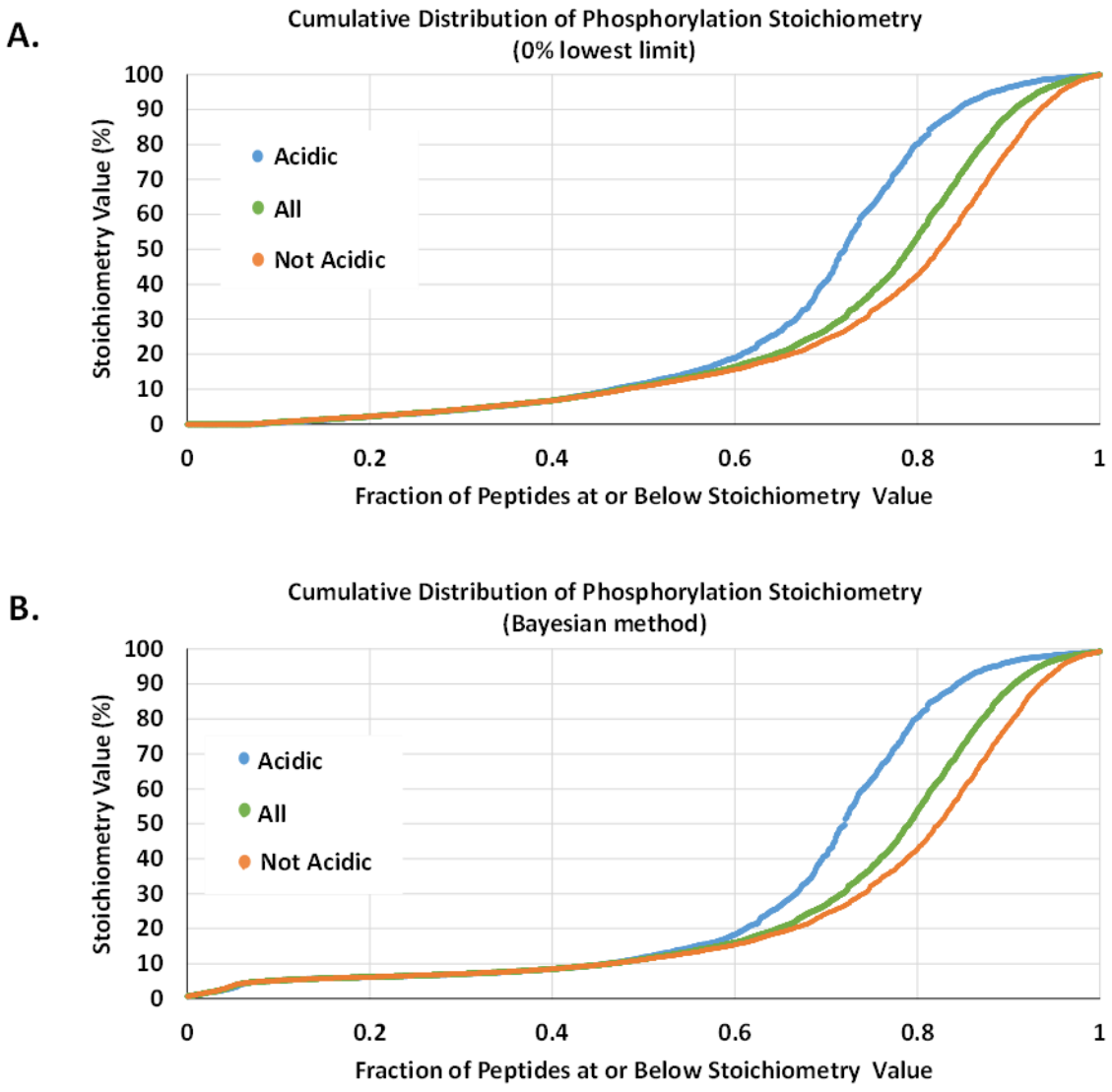


Figure 4-3: Cumulative distribution plots of peptide stoichiometry based on assigned type. Our plots recapitulate the original trends identified in our earlier findings[2]. **A)** Cumulative distribution plot when the 0% lower limit method is used. **B)** Cumulative distribution plot when stoichiometry is estimated using our Bayesian modelling approach. Minor changes are observed in the trend when utilizing the Bayesian modelling approach.

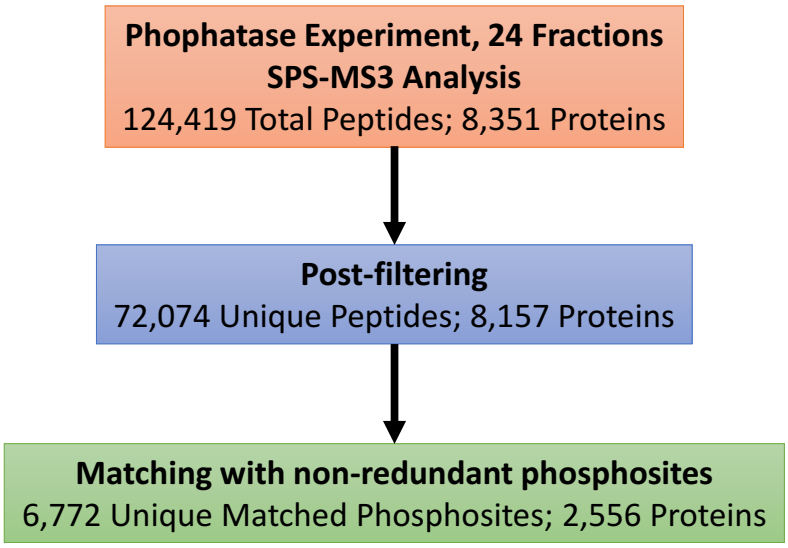
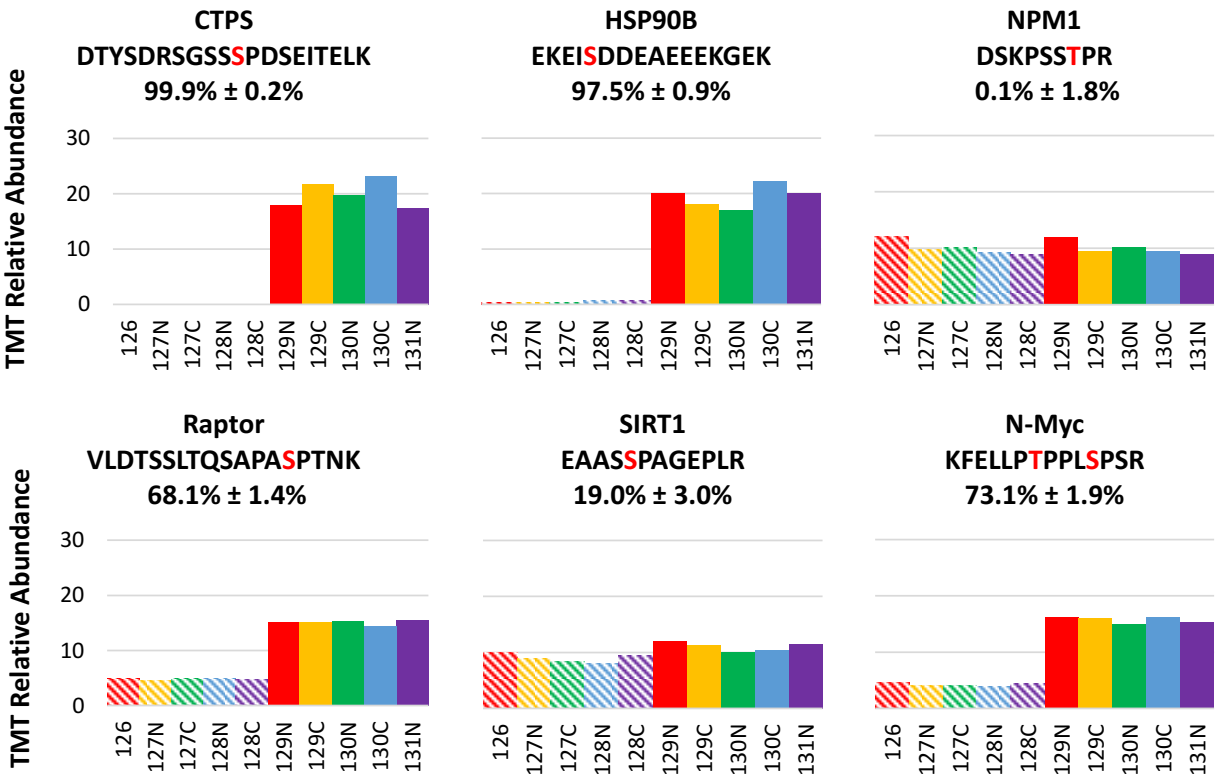


Figure 4-4: Summary of phosphorylation stoichiometry experiment results. 124,419 total peptides corresponding to 8,351 proteins were identified. 6,772 unique peptides (2,556 proteins) were matched to phosphorylation sites identified in our phosphopeptide enrichment experiment.

Figure 4-5: Analysis of stoichiometries for selected peptides. **A)** Example TMT-data for peptides known to harbor phosphorylation sites. Stoichiometries were calculated for each sample (red, yellow, green, blue, and purple), the average and standard deviation are reported. Solid colors represent channels where the aliquot was treated with phosphatase while the striped colors represent channels where the aliquot was mock treated. **B)** Table displaying the individual sample phosphorylation stoichiometries calculated for each peptide in a). Red characters represent the expected phosphorylation site. All sites chosen were identified as regulatory phosphorylation events through targeted studies based on the phosphositeplus.org database[23].



B.

Protein	Peptide	Stoichiometry in Biological Replicate				
		1	2	3	4	5
CTPS	DTYSDRSGSSSPDSEITELK	100.0%	99.6%	100.0%	100.0%	100.0%
HSP90B	EKEISDDEAEEEKGEK	98.1%	98.1%	97.8%	96.1%	97.5%
NPM1	DSKPSSTPR	-0.5%	-1.8%	2.1%	1.9%	-1.3%
Raptor	VLDTSSLTQSAPASPTNK	66.9%	69.3%	68.0%	66.4%	69.8%
SIRT1	EAASSPAGEPLR	17.1%	20.6%	16.7%	23.5%	19.0%
N-Myc	KFELLPTPPLSPSR	71.2%	74.4%	73.0%	75.7%	71.3%

could be calculated as negative values (Figure 4-5A,B). An example is nucleophosmin, NPM1, which had a positive average stoichiometry near 0%, but had individual replicates that were assigned negative stoichiometries using our standard method of stoichiometry calculation (Figure 4-5A,B)[23]. We then attempted to address these negative stoichiometry issues by either setting the lower limit of stoichiometry to 0% and by developing our Bayesian model.

The boundary conditions of the stoichiometry measurement affect its distribution

Previously, peptide phosphorylation stoichiometry was treated as statistic and calculated directly from the raw data[2,6,7]. As such, we initially calculated this stoichiometry statistic and plotted a histogram of the results. The resulting distribution was centered near 0% resulting in substantial negative stoichiometries being calculated (Fig 4-6A). Additionally, a second population of stoichiometries near 100% was observed. Both observations are in line with previous data from our lab[2].

To address the issue of negative stoichiometries, we then calculated stoichiometry but only allowed the lowest value to be 0%, as reported previously[2]. This resulted in all negative stoichiometries being set to 0%. The resulting histogram showed little to no change in bins containing average stoichiometries of 30% or more but showed an increase in the bin height of the bins containing 6% or less average stoichiometries (Fig 4-6B). While this solved the issue of negative stoichiometries, it created a new problem of artificially reducing our error estimates, as discussed later.

As an alternative to limiting the lowest stoichiometry to 0%, we created a Bayesian model that would treat phosphorylation stoichiometry as an unobserved parameter defined on the interval 0 to 1. In doing so, we can utilize all of the observed measurements to estimate stoichiometry and inform our error and precision. We ran our statistical model on the dataset and

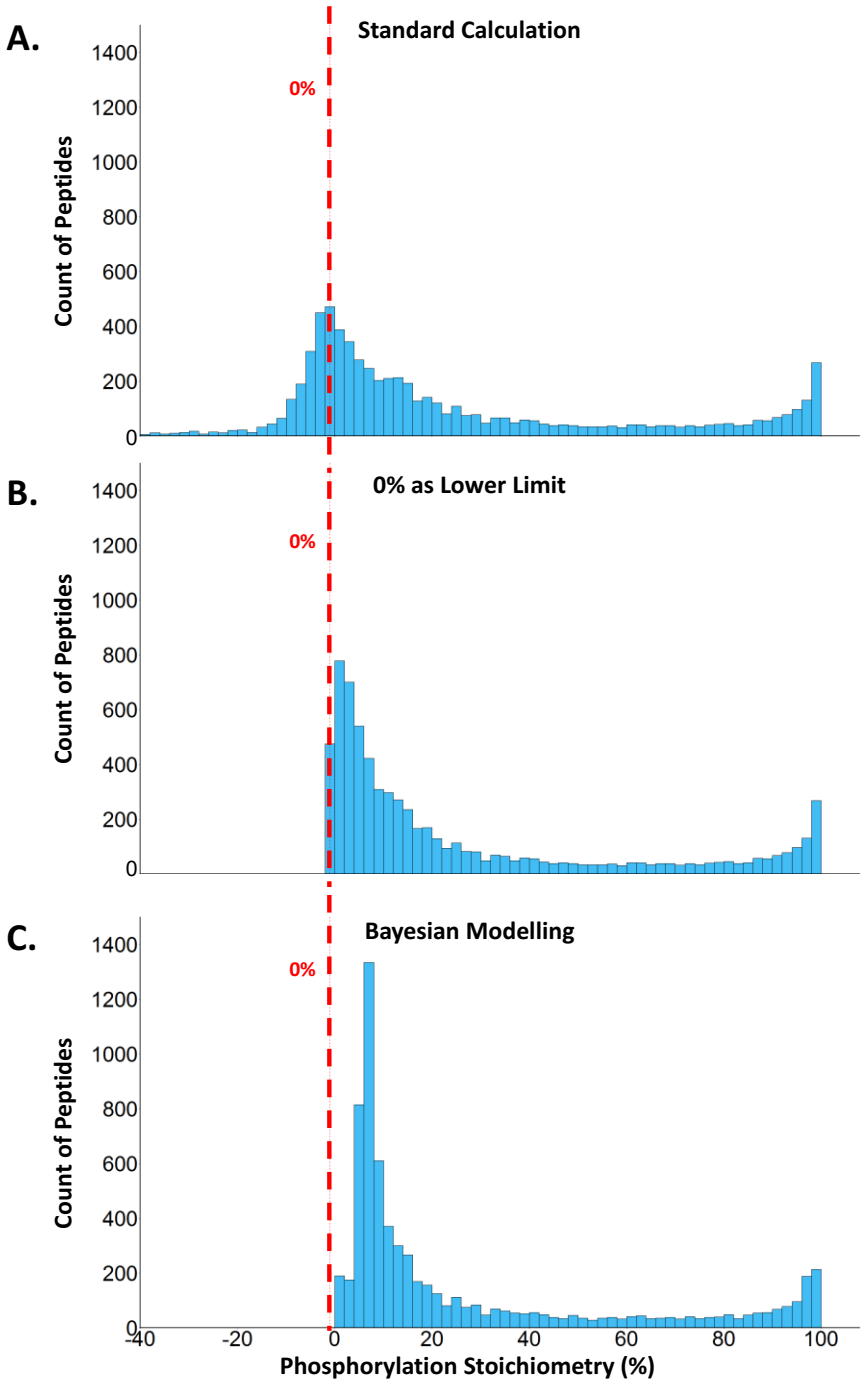


Figure 4-6: Histograms of the phosphorylation stoichiometry for each estimation method. **A)** Histogram when no correction is performed. **B)** Histogram where each negative stoichiometry is replaced with 0. **C)** Histogram when stoichiometry is estimated using the Bayesian modelling approach. The red dashed line represents 0%.

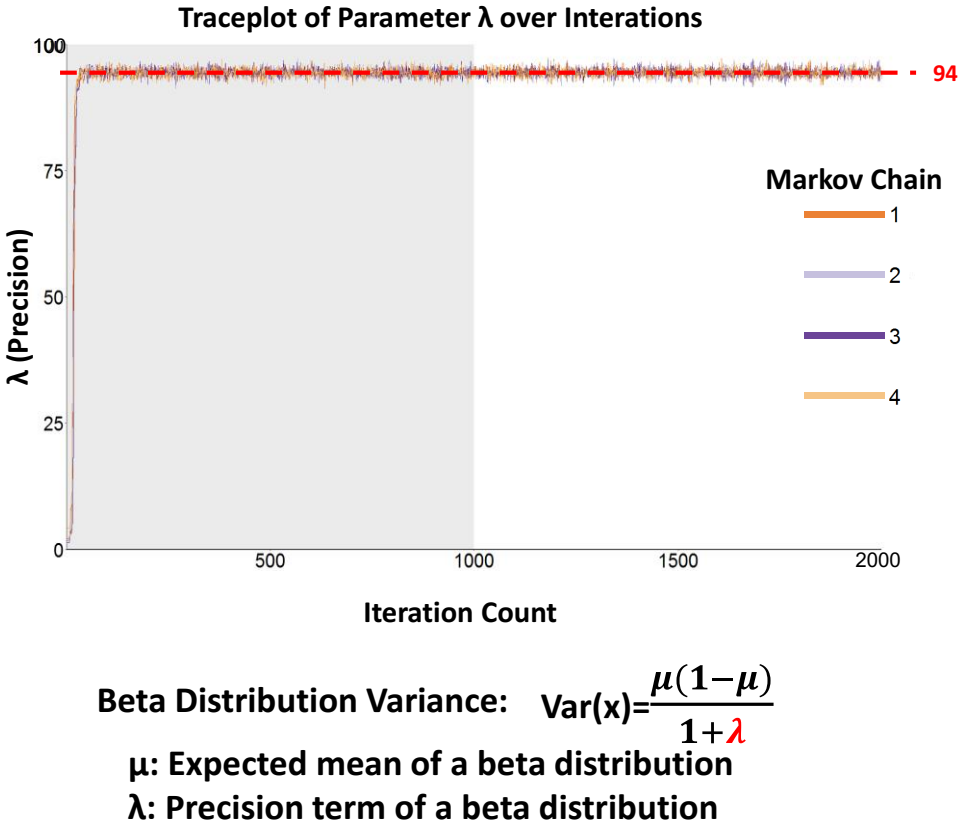


Figure 4-7: Traceplot depicting value of Lambda during each iteration during of the Monte-Carlo Markov Chain simulation. The shaded region of the plot represents iterations used in the warm-up phase of the simulation. Lambda converges rapidly during the warmup to its final value of 92. As seen in the equation for the variance of a Beta distribution, the precision parameter λ is inversely proportional to the variance. However variance is also a function of location.

observed that it converged rapidly with a precision value of 94 (Figure 4-7). The precision value is inversely related to the variance of a beta distribution given a specific expected mean. As such, increasing precision results in decreasing variance. Plotting the distribution of the stoichiometries as a histogram highlighted large increases in the bins containing average expected stoichiometries between 6-10% (Figure 4-6C).

To assess how the Bayesian modelling was affecting the stoichiometries obtained by traditional methods, we compared the differences between the standard method for calculating stoichiometry and the 0% lower limit method with the Bayesian model. We found that a majority of stoichiometry values did not change dramatically (Figure 4-8). Additionally, we observed that most changes to the stoichiometry when going from average measurements to expected means from the Bayesian model resulted in a 5-10% increase. These data agree with the change in the distribution of the histograms, implying that our statistical method preferentially affects the calculations yielding negative or low stoichiometries (Figure 4-6 and Figure 4-3).

The increase in the observed stoichiometry value when utilizing the Bayesian modelling method suggests that measuring a 0% stoichiometry is extremely difficult with the current instrumentation and that perhaps the lower end of our reliable estimation of stoichiometries is approximately 5-10%. This was further confirmed when we assessed how phosphorylation site motifs affect stoichiometry by utilizing the stoichiometries obtained from our 0% lower limit and Bayesian modelling method. When using the 0% lower limit method, peptides assigned a phosphorylation stoichiometry containing an acidic motif peptide were more likely to be observed at a higher stoichiometry, with ~20% of all peptides kept at 0% (Figure 4-3A). This is in line with our previous findings in yeast whole cell lysate[2]. When utilizing the Bayesian modeling method, this trend is preserved; however, we noticed that peptides estimated to be at

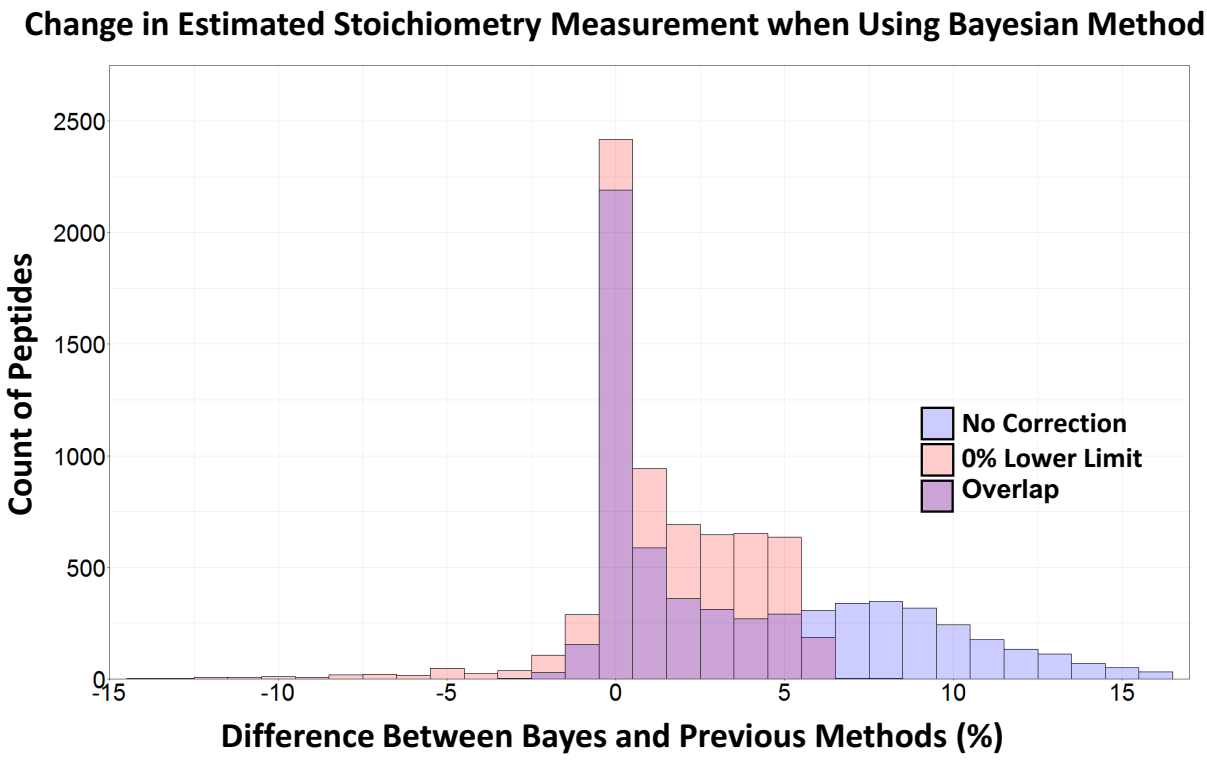


Figure 4-8: Distribution of differences between the Bayesian models' estimation of stoichiometry and other methods of calculating stoichiometry. Most peptides experience little change in their stoichiometry; however, some peptides demonstrate an increase in stoichiometry when using the Bayesian modelling.

0% by the 0% lower limit method were pushed off the x-axis in the Bayesian modelling method (Figure 4-3A,B). This seemingly implied that the cell maintains a low level of phosphorylation for peptides thought to be kept at 0% stoichiometry. However, both the cumulative distribution plots from Supplemental Figure 1 and the histograms from Figure 5 only visualize the point estimate of each stoichiometry distribution for each peptide. Large variance could render these point estimations worthless and necessitate the investigation of the error intervals surrounding each stoichiometry point estimator. As such, it cannot be inferred that a majority of the proteome is kept at 5-10% stoichiometry without first looking at the error intervals.

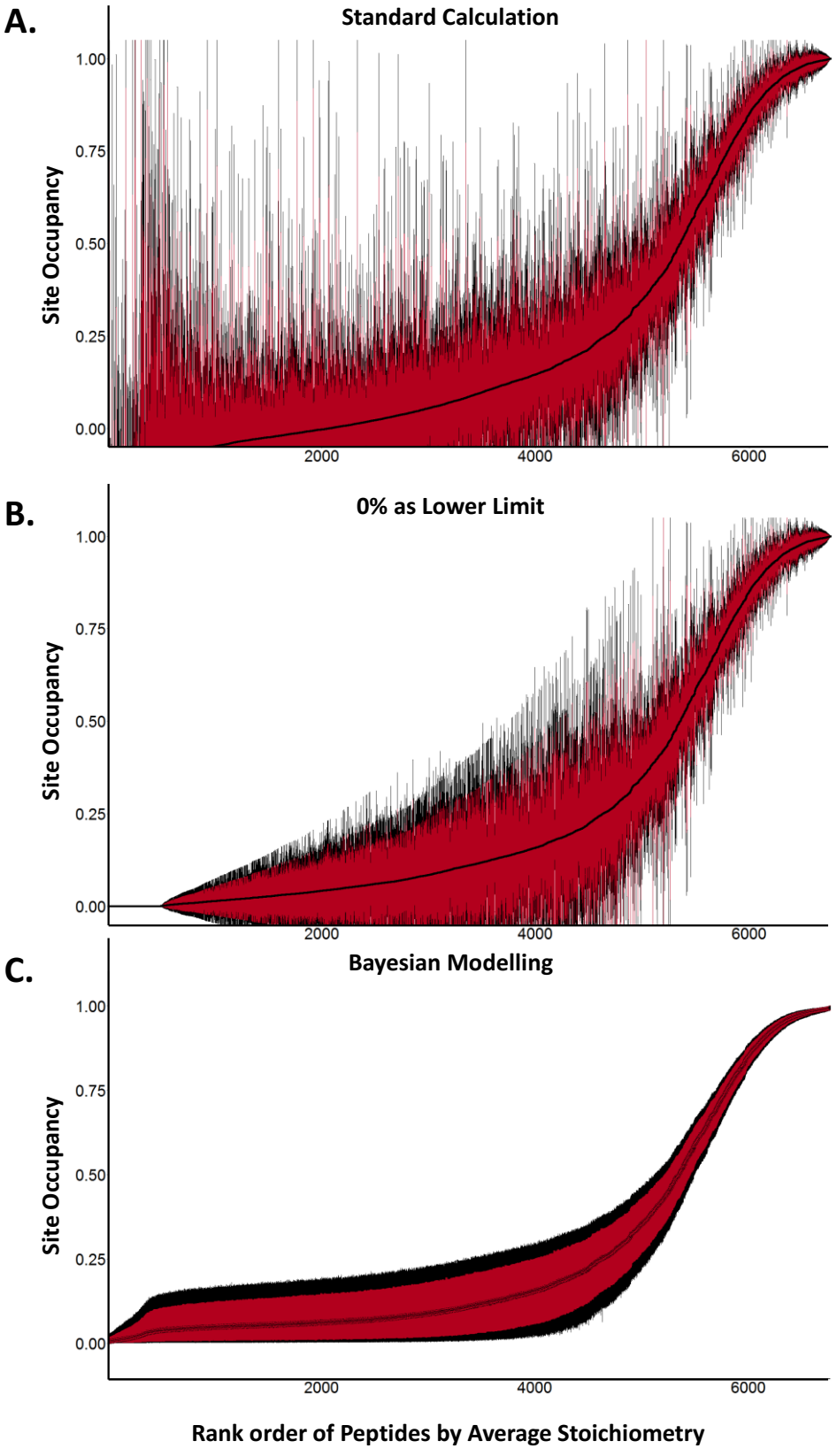
Proper modelling prevents error intervals from containing senseless results

To assess the variation of the stoichiometry distributions by the different estimation methods, we plotted the rank ordered peptide phosphorylation averages with their 80% and 95% confidence intervals. When calculating the stoichiometry value using the standard method, we observed a noticeable number of the average phosphorylation stoichiometries that fell below 0%; furthermore, a majority of peptides had confidence intervals that included negative values or values exceeding 100% (Figure 4-9A). Additionally, we noted the abundance of relatively large confidence intervals throughout the dataset.

We then assessed how setting the lower limit of stoichiometry to 0% would affect this plot. About 12% of the data were incorrectly reported as having no variance while about half of the peptides displayed a trend of increasing interval size as the peptide's stoichiometry average increased (Fig 6B). This linked relationship between increasing average and standard deviation coupled with the region of no variance caused us to question the validity of this method. By artificially clipping the negative stoichiometries we calculated to 0%, measurements of variability were artificially reduced, with greater reductions occurring the closer the

Figure 4-9: Caterpillar plots depicting phosphostoichiometries estimated with different analysis methods. Peptides were rank ordered (lower values first) by their estimated stoichiometry. 80% confidence intervals (red bars) and 95% confidence intervals (black bars) were drawn around each point. The y-axis represents the phosphorylation stoichiometry as a fraction instead of a percent. Resulting caterpillar plots are shown for each method. **A)** Standard method with no corrections performed. **B)** All negative stoichiometry calculations were replaced with 0. **C)** Stoichiometry values were estimated using our Bayesian model.

Figure 4-9 (continued)



stoichiometry average was to 0%. Furthermore, we still had problems with nonsensical error intervals containing values outside of the 0 to 1 range.

When utilizing our Bayesian model to estimate stoichiometry, the program additionally generates credible intervals around the expected stoichiometry value. We performed the same plotting method as above, which shows that all peptides have credible intervals corresponding to physical reality (Figure 4-9C). We observed additionally a vertical shift at the low end of the graph indicating that most peptides previously thought to be at 0% phosphorylation stoichiometry now had stoichiometry point estimators slightly higher (Figure 4-9C, Figure 4-6C, and Figure 4-8). Overall, while the general shape and trend of the plots remain unchanged, the error intervals improved dramatically when utilizing the Bayesian model. This is further highlighted by the observation that approximately 2,000 peptides with confidence intervals containing only physically possible results when utilizing the standard method and approximately 3,000 with the 0% lower limit method (Figure 4-10). Additionally, the credible intervals using the Bayesian method suggest that, for a majority of peptides, even though the point estimator suggests 5% stoichiometry the true stoichiometry lies anywhere between 0% and 20% stoichiometry.

Discussion

In this study, 5 biological replicates of HCT116 cells were analyzed to gain insight into the basal level of phosphorylation stoichiometry of this colorectal cancer cell line. Prior to determining stoichiometry we collected a reference database of 24,028 phosphorylation events under basal conditions which served as a library of sites to attempt stoichiometry assessment. Our occupancy analysis was performed using TMT labeling which increased the sample

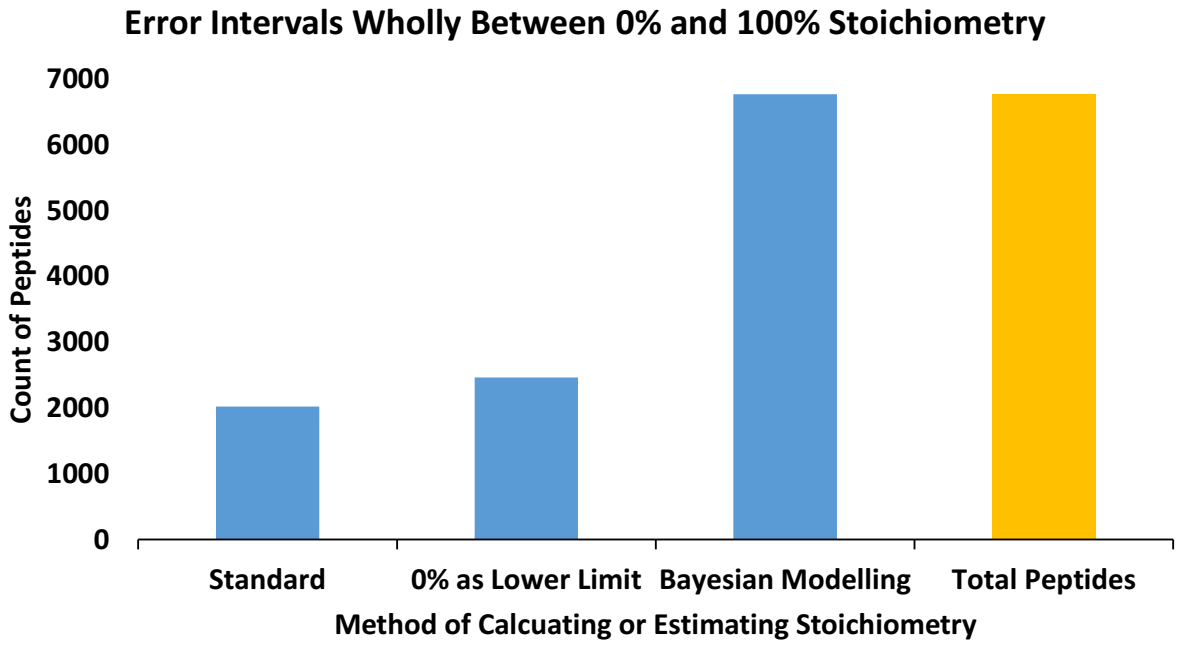


Figure 4-10: Bar charts depicting number of peptides contained within 95% wholly within physical limits. Chart counts peptides with 95% confidence intervals (or credible intervals in the case of the Bayesian Modelling) containing stoichiometry values between 0% and 100% for each method of Stoichiometry calculation or estimation. Having a confidence interval or credible interval that contains values greater than 100% or less than 0% implies that the true phosphorylation stoichiometry could lie outside the physical bounds of 0-100%.

multiplexing capacity to allow simultaneous analysis of all 5 biological replicates. In addition, by utilizing TMT, there were no missing values in that all 5 measurements were determined for all peptides in the dataset. In total, we assigned 6,772 unique peptides, from our generated reference library, stoichiometry values.

As stoichiometry is defined as the fractional occupancy, its values should, ideally, reside within the unit interval [0, 1]. Despite quantifying peptides across five replicates, by following the standard method of calculating phosphorylation stoichiometry values, we initially obtained some negative stoichiometry values[2]. This occurred stochastically when sites were present at low stoichiometries such that the error in the 5 measurements was greater than the % occupancy. Additionally, our initial attempts at calculating stoichiometry resulted in confidence intervals containing values greater than 1 suggesting over 100% occupancy. Both phenomena are physically impossible.

Previous iterations of this phosphatase method estimated stoichiometry from a stoichiometry statistic calculated from the raw data rather than treating stoichiometry solely as an estimable parameter[2,6–8]. As the raw data are the S:N values collected from the instrument ranging from 1 to positive infinity, nothing constrains a stoichiometry statistic calculated with the formula in Figure 1b to the unit interval. If we treat stoichiometry as a constrained parameter we wish to estimate, rather than a statistic calculated from the raw data, we can utilize novel approaches to estimate the true stoichiometry of a peptide utilizing alternative statistics that leverage the raw data's properties.

Furthermore, negative stoichiometries traditionally have been dealt with by replacing the negative stoichiometries with 0% or discarding those measurements[2,6,7]. However, as mentioned above, the raw data can be transformed into a meaningful statistic from which a stoichiometry

parameter can be estimated. This alternative statistic is the proportion of the sum S:N of the paired TMT channels corresponding to a replicate contributed by the untreated channel. Furthermore, the statistic, which is based on the proportionality of the data, can easily be converted into the traditional stoichiometry measurement thus allowing us to use this statistic to estimate phosphorylation stoichiometry as a parameter.

We implemented our Bayesian modelling by developing an R/Stan script included in the supplemental materials. The software samples mean peptide effects on stoichiometry, sample handling effects, and overall experimental precision. We chose to utilize an overall experimental precision due to the low sample size when analyzing precision per peptide. This measurement of precision provides a quantitative measure of the global experimental variance while still providing individual peptide variance as the variance of a beta distribution is governed by the mean and the precision term. This can be preferable to using a per-peptide error as we only acquired five measurements per peptide, one for each biological replicate, resulting in unstable estimations of error. The tradeoff is that an overall experimental error gives a coarse overview of the error may not accurately represent each peptide. A further benefit is that a single experimental precision provides a quick and quantitative factor to compare multiple experiments.

Based on the amount of uncertainty surrounding many of the stoichiometry point estimators, we found it was more effective to bin point estimators based into low (0-25%), medium (25-70%), and high (70-100%) categories. Similar to the Wu et al paper, we found that acidic residues are phosphorylated at a higher stoichiometry than sites with other phosphorylation motifs (Figure 4-3)[2]. This specific phosphorylation of acidic motifs is likely due to the high activity of Casein kinase II which targets the motif SxxE/D[24].

Conclusion

We simultaneously compared the basal phosphorylation stoichiometry of 5 biological replicates of HCT116 using a TMT based workflow eliminating previous problems involving missing data. We then presented a novel statistical method to address negative stoichiometries from using the phosphatase based phosphorylation stoichiometry experiment. While the credible intervals were larger than we had hoped, the global phosphorylation can be binned into low, medium, and high phosphorylation stoichiometry categories which allow for a quick first-pass assessment of the phosphorylation state of the cell. Further study into improving measurement precision by utilizing targeted approaches or real time search may further narrow these bins. Overall, our study provides a methodical way to make sense of complex phosphorylation occupancy experiments and a quantitative read out for experimental error.

Supporting Information

The following files are available free of charge at the ACS website http://pubs.acs.org:
Sup_Info_Phosphatase_Lim_et_al.pdf:

Sup_Table1_localized_phosphopeptides_HCT116.xlsx: Table of all phosphopeptides localized to phosphorylation sites identified during the phosphopeptide library experiment
Sup_Table2_peptides_assigned_stoiciometries.xlsx: Relative abundances and stoichiometries of peptides identified and quantified during the phosphatase experiment

The code to run the Bayesian modelling (an R-script that initializes the Stan code) is available upon request (JPR does not support these file formats for upload).

References

1. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6940–5 (2003).

2. Wu, R. *et al.* A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nat. Methods* 8, 677–683 (2011).

3. Humphrey, S. J., James, D. E. & Mann, M. Protein Phosphorylation: A Major Switch Mechanism for Metabolic Regulation. *Trends Endocrinol. Metab.* 26, 676–687 (2015).

4. Newman, R. H., Zhang, J. & Zhu, H. Toward a systems-level view of dynamic phosphorylation networks. *Front. Genet.* 5, 1–22 (2014).

5. Olsen, J. V. *et al.* Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci. Signal.* 3, ra3-ra3 (2010).

6. Glibert, P. *et al.* Phospho-iTRAQ: Assessing Isobaric Labels for the Large-Scale Study Of Phosphopeptide Stoichiometry. *J. Proteome Res.* 14, 839–849 (2015).

7. Tsai, C.-F. *et al.* Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. *Nat. Commun.* 6, 6622 (2015).

8. Domanski, D., Murphy, L. C. & Borchers, C. H. Assay development for the determination of phosphorylation stoichiometry using multiple reaction monitoring methods with and without phosphatase treatment: application to breast cancer signaling pathways. *Anal Chem* 82, 5610–5620 (2010).

9. Horinouchi, T., Terada, K., Higashi, T. & Miwa, S. Using Phos-Tag in Western Blotting Analysis to Evaluate Protein Phosphorylation. in *Kidney Research: Experimental Protocols, Methods in Molecular Biology* 1397, 267–277 (Springer Science+Business Media, 2009).

10. McAlister, G. C. *et al.* Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.* 84, 7469–7478 (2012).

11. Paulo, J. A., O'Connell, J. D. & Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J. Am. Soc. Mass Spectrom.* 27, 1620–1625 (2016).

12. Thompson, A. *et al.* Tandem mass tags: a novel quantificaiton strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75, 1895–1904 (2003).

13. Paulo, J. A., Mancias, J. D. & Gygi, S. P. Proteome-Wide Protein Expression Profiling Across Five Pancreatic Cell Lines. *Pancreas* 1 (2017). doi:10.1097/MPA.0000000000000800

14. Ross, P. L. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* 3, 1154–1169 (2004).

15. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* 8, 937–940 (2011).

16. Mcalister, G. C. *et al.* MultiNotch MS3 Enables Accurate, Sensitive, and Multiplexed Detection of Differential Expression across Cancer Cell Line Proteomes Graeme C. McAlister, 1 David P. Nusinow, 1. *Anal. Chem.* 86, 7150–7158 (2014).

17. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 162, 425–440 (2015).

18. Villén, J. & Gygi, S. P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* 3, 1630–1638 (2008).

19. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 143, 1174–1189 (2010).

20. Wasmuth, E. V & Lima, C. D. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, 1–12 (2016).

21. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 4, 207–214 (2007).

22. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* 24, 1285–1292 (2006).

23. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43, D512–D520 (2015).

24. Ahmed, K. & Issinger, O. *Protein Kinase CK2 Cellular Function in Normal and Disease States*. (2015). doi:10.1007/978-3-319-14544-0

Chapter 5: Discussion

Discussion

While genomics provides a blueprint for a cell, proteomics provides an understanding of how the encoded cellular gears and machineries allow for functionality. To that end, mass spectrometry is a powerful tool with its ability to identify and quantify thousands of proteins and dozens of associated post translational modifications. However, with the increasing focus on quantitation, researchers hoping to leverage mass spectrometry for quantitative proteomics need to better understand measurement error in their experiments, regardless of whether labelled or label-free methods are employed. As shown, this body of work summarizes some of the most commonly occurring measurement errors in the field: interference caused by isolation of co-eluting precursors, incorrect measurements due to incorrect identifications, and instrument imprecision when measuring small relative changes. The findings suggest that, in simpler terms, better instrumentation, proper identification, and larger fold changes reduce measurement error.

In Chapter 2 of this dissertation, the effects of incremental improvements to mass spectrometry instrumentation on peptide interference during quantitative labelled methods utilizing Tandem Mass Tags (TMT) were assessed. Data showed that improvements to the instrumentation led to direct reduction in interference thereby resulting in better quantitation. Generally, this finding suggests that as mass spectrometry hardware improves, tighter isolation widths with sharper cutoffs will reduce the effects of co-isolation of co-eluting peaks. To offset the reduction of signal intensity traded for this increased precision of precursor isolation, "brighter sources" will allow for increased ion transmission. However, the most important finding of this chapter is that interference reduction from hardware improvements in reporter ion-based quantitation during high resolution tandem mass spectrometry (hrMS2) methods is not as drastic as employing a gas phase filtering methods, such as synchronous precursor selection

MS3 (SPS-MS3). Results from Chapter 2 show that shrinking isolation widths alone are insufficient to reduce interference at the level when compared to utilizing the SPS-MS3 method. Additionally, the reduction in interference from the selection of fewer precursors during SPS-MS3 leads to two implications: 1) selection of the correct precursors and 2) improvement in the quality of this selection by reduced isolation widths will noticeably reduce interference. These findings lend credence, from a measurement error perspective, to the push for intelligent data acquisition, such as real-time database searching.

Fundamentally, accurate quantitation is tied to appropriate identification: quantitation is only meaningful on identified peptides and proteins. As a corollary, misidentification leads to inappropriate quantitation. This issue is only broached when comparison between multiple LC-MS/MS runs is necessary. While labelled approaches, such as TMT, handle this issue through multiplexing, label-free experiments must choose whether to transfer identifications between runs or remove data due to incompleteness. The focus of Chapter 3 is on the former. The data show that incorrect identifications are rampant when the common Match-Between-Runs (MBR) algorithm facilitates the identification transfers. Currently, the accepted solution to this problem is to, effectively, undo these transfers rendering use of the MBR algorithm an exercise in futility.

As MBR is a commonly utilized tool, understanding this potentially major issue is critical. Transparent calculations of false discovery rates and detailed explorations into statistical underpinnings behind identification transfers are required for this technique to be a credible solution in quantitative mass spectrometry. Future directions for the MBR field can be divided into improving the current algorithm and further understanding stochasticity. Chapter 3 presented a solution to estimate the false transfer rate utilizing a two-proteome model – however, more work is required to devise a universal solution for calculating false transfer rate that would be

analogous to the target-decoy method. Furthermore, the idea of MBR raises the question: What makes an identification transferable at the MS1 level? Leveraging retention time and m/z, while incredibly useful, seems to be an incomplete solution when other characteristics like charge state are seemingly ignored. Given that MBR occurs post hoc of data collection means that any future algorithm designed should seek to employ all the information acquired to gain a better understanding of what a potential identification transfer looks like.

Quantitative measurements from mass spectrometry experiments can be utilized in downstream calculations to gain more biological insight. Phosphorylation stoichiometry is one such experiment where quantitative measurements from a mass spectrometry experiment are utilized further in a separate calculation. In Chapter 4, phosphorylation stoichiometry was calculated indirectly by utilizing the relative abundance of unphosphorylated peptides between phosphatase-treated and untreated samples. Previously limited to a single comparison due to limitations surrounding the reductive dimethylation technique, the work in Chapter 4 showed that expanding the number of comparisons to five using modern TMT technology is possible. While each comparison can be different, by employing the four additional comparisons for replicates, a better understanding of measurement accuracy is achieved. Based on the data from the five replicates, boundary effects were a noticeable problem with the lower boundary of 0% occupancy proving to be extremely difficult to measure. By leveraging replicate information and developing a new statistical tool to estimate stoichiometry, four important conclusions were drawn: first, low occupancy events are incredibly difficult to measure with this method; second, the mass spectrometer struggles to measure an exact 1:1 ratio (i.e. 0% stoichiometry); third, reporting some measurement of variance when presenting mass spectrometry data is important due to scan-to-scan and run-to-run variations; fourth, the majority of phosphorylation events

observed by this method were found to have low stoichiometry. Together these conclusions suggest two directions of future work: 1) method development to reduce error and 2) biological applications.

While the occurrence of non-physical phosphorylation stoichiometry measurements resulting from the indirect phosphatase method was addressed by the work in Chapter 4, the resulting data highlighted another source of error needing to be addressed; namely, that the resulting stoichiometry measurements are peptide-based rather than site-based. Missed cleavages and common variable modifications, like methionine oxidation, are allowed in the database search performed on the collected LC-MS/MS data. By virtue of the indirect phosphatase method, stoichiometry is calculated by measuring the abundance of unphosphorylated peptide which is inherently devoid of localization data. As such, by this method. a phosphorylation site can be measured in multiple peptide forms. For example, if a peptide is observed in two forms (e.g., with and without oxidized methionine) each peptide would generate a different stoichiometry measurement. A similar, but potentially more complicated, situation would involve two peptides that span the same phosphorylation site except one contains a missed cleavage. Importantly, in both examples, each of the mentioned peptide forms need not be observed with the same stoichiometry.

As such, the next major improvement to the presented model and statistical framework would be to account for multiple peptides being attributed to the same phosphorylation site. Data from multiple LC-MS/MS scans (i.e. multiple peptides) would need to be combined to calculate an accurate stoichiometry for a site or set of sites. However, this is not trivial due to the nature of reporter ion-based quantitative data; each scan provides intensity-based relative abundance

measurements for each TMT channel, but no reliably accurate direct method to compare these measurements across scans exists.

Combining stoichiometry measurements from multiple peptides is complicated further in the case of missed cleavages: each peptide could contain a different set of phosphorylation sites with the abundances of each form (fully cleaved and containing missed cleavages). Take the hypothetical case where three peptides, a missed cleavage containing peptide and its two fully cleaved siblings, are observed in the mass spectrometer. If each fully cleaved peptide spans a known phosphorylation site three stoichiometry measurements will be generated over the two sites – one measurement from each unique peptide sequence (missed cleavage containing peptides count as a unique peptide). Additionally, each stoichiometry measurement is not fully independent as the phosphorylation event may be influencing the missed cleavage rate. In this imaginary situation, it is likely beneficial to combine all three measurements. Future research into handling this phenomenon of single phosphorylation sites assigned to multiple unique peptide sequences will be required if the indirect phosphatase method for measuring global phosphorylation stoichiometry is to see more widespread adoption.

Although challenges persist with this method to calculate stoichiometry, the current improvements presented in this dissertation highlight the promise that this method has for various biological applications. Interestingly, a reason to study phosphorylation stoichiometry is that standard analyses of phosphorylation events look only at relative fold changes. Though powerful, relying solely on fold changes places the finer nuances of phosphorylation changes in a black box. Ironically, global phosphorylation analysis at the whole cell lysate level is still an abstraction of the true phosphorylation landscape within the cell. Based on the data from Chapter 4, most observed phosphorylation sites in an average asynchronous HCT cell population are kept

at a low stoichiometry (<30%) with high variation. This phenomenon leads to three hypotheses: 1) that all observed low stoichiometries are true but difficult to measure; 2) the combination of low observed stoichiometry and high variance can be explained (at least partially) by these sites only undergoing phosphorylation during a specific cellular state or process; and 3) that the combination of low mean and high variance is influenced by vastly different stoichiometries observed in different sub-cellular localizations. While the first hypothesis can be tested by targeted stoichiometry experiments, the second and the third hypotheses seek to addresss a philosophical question akin to the one that spurred on the initial pursuit of phosphorylation; "what different states result in 50% stoichiometry?" is parallel to "what different states result in a 2-fold change?" Understanding phosphorylation stoichiometry in the context of phosphorylation events in various cellular pathways or states and sub-cellular microenvironments can potentially lead to interesting biological discoveries.

Differences in phosphorylation stoichiometry due to cell state or pathway specific phosphorylation events is one possible way bulk global phosphorylation stoichiometry simplifies cellular phosphorylation. Historically, phosphorylation stoichiometry was most prominently highlighted for its role as an alternative mechanism to regulate the protein separase independent of securin. The phosphorylation of separase was found to be important due to consistent subtle decrease in stoichiometry during the cell cycle. This example highlights the potential of phosphorylation stoichiometry analysis to discover biological functions for many identified but uncharacterized phosphorylation sites overlaying the analysis on top of a pathway context. While previous methods were low throughput, incorporating the presented improved phosphatase-based stoichiometry analysis in future work investigating phosphorylation events during different phases of the cell cycle or during signal transduction is an easy way to deepen the insight gained

from such studies. Furthermore, synchronizing cells by arresting them during specific stages in a pathway can reduce the overall cell-to-cell variation reducing the uncertainty of the stoichiometry measurements by homogenizing the cell population. These experiments involving tighter cell state synchronization will help elucidate how much of the variation observed in the data presented in Chapter 4 was due to the analysis being performed on bulk HCT cells in addition to any biological phenomena discovered.

However, even understanding how phosphorylation stoichiometry changes in different cell states is insufficient. Local concentrations of proteins and other molecules can be orders of magnitude higher in certain sub-cellular contexts compared to their overall cellular concentration. These larger local concentrations can drive local chemical reactions and biological pathways. As such, it is easy to hypothesize that this is true for phosphorylation events as well. For example, understanding that a site's phosphorylation stoichiometry is at 100% in the nucleus, but 0% in the cytosol, given a specific cellular state could highly inform a site's role in regulating the proteins function. Additionally, while the bulk data from Chapter 4 displays an almost digital-like response from phosphorylation (i.e., phosphorylation is either on or off), understanding the sub-cellular stoichiometries of phosphorylation sites during different cellular states and pathways is essential to elucidate fully whether phosphorylation is a more digital-like or analog-like signal.

While the data presented here generates many distinct questions for follow-up, ultimately, the work performed in this dissertation highlights how advancements in hardware and software affect measurement error in mass spectrometry-based proteomics. At its core, measurement error stems from inconsistencies with physical data acquisition. For reporter ion quantitation, incorrect and impure precursor selections drive interference and error. In label-free

experiments, dealing with run-to-run variation and stochastic identification hampers accurate quantitation. Novel technologies and techniques, such as real-time database searching, data independent acquisition, and field asymmetric ion mobility separation, can improve both reporter ion and label-free quantitation. These improvements are at the data acquisition level and, thus, lessen the necessity of software solutions to correct for data acquisition insufficiencies. However, the nature of reality is that imperfections will always exist. As such, measurement variation must not be ignored. Hardware improvements and innovations that directly affect data acquisition may reduce this variation, but nothing will fully eliminate it. To that end, quantitation of mass spectrometry data should begin to account for scan-to-scan variation and data reporting should include confidence intervals (or an equivalent estimate of variation) for each protein estimate. Measurement error will always exist, the best course of action for the field is to account for it and honestly report it.

Appendix A – RatioCheckR – an interactive tool to calculate load control normalization via TMT reporter ion ratios

Attributions:

Matthew Y Lim conducted the experiments, designed the software tool, and wrote the manuscript.

João A Paulo provided samples and conceived the initial idea for the project.

Abstract

Quantitative proteomics, typically involves comparisons between samples. These comparisons rely on loading controls to establish a frame of reference. Protein quantitation assays like the bicinchoninic acid assay (BCA) and their peptide counter parts are often used mass spectrometry workflows to estimate the total amount of protein within each sample to use as loading control normalization factor. However, these biochemical assays can be inaccurate due to complications including buffer compatibility and small linear ranges. Here we present a software tool, RatioCheckR that can be used to quickly determine accurate loading control normalization factors for isobaric tag-based mass spectrometry experiments based on information acquired from short high-resolution tandem mass spectrometry analyses. These normalization factors can correct concentration estimates from traditional biochemical assays allowing for more accurate 1:1 mixing of isobaric tag labelled samples. Our results show that up to 50% deviation from a 1:1 ratio is observed when utilizing uncorrected biochemical assay estimates to inform sample mixing. By correcting this data with ratio check information, the maximum deviation is reduced to 9%, which indicates a closer 1:1 ratio.

Introduction

Comparisons (e.g., between samples, conditions, and/or standards) are the cornerstone of quantitative proteomics. Through measuring relative abundances, researchers can understand changes in context and, with the appropriate conditions, even convert these relative measurements into absolute values[1–4]. However, in order for comparisons to be valid, conditions for equal sample load is essential.

Loading controls establishes a basis for comparisons by making the assumption that equal amounts of common, unchanging materials in the sample are present[5–9]. Essentially, the loading

control shows specific changes with respect to overall sample changes allowing for cross sample normalization to a specific protein. While simple, this assumption is incredibly important and must hold true for any comparisons to be viewed as valid. Common loadings controls, such as Actin and GAPDH (proteins that are generally considered common and unchanging), have been used in Western blotting experiments to show that an observed change is occurring without affecting unrelated systems[10–12]. However, the use of specific proteins for loading controls, is not without its flaws. Specifically, problems arise when the protein is not consistent between all samples due to a given perturbation[7,8,12].

As an alternative, total protein amount or sample mixtures can be used as a loading control[6–9,12]. Utilization of total protein slightly changes the original loading control assumption: comparisons are made with the assumption that overall protein expression is not affected. As such, normalization is performed to total protein amount rather than a specific protein. In Western blotting experiments, where overloading and heavy reliance on normalization is a common issue, this method has been shown to be a viable solution to better estimate protein concenterations[6–8]. It should be noted that while total protein can serve as a good loading control, issues arise when total protein expression is dramatically uncoupled from cell doubling (e.g. during immune cell activation)[13].

Quantitative liquid chromatography with tandem mass spectrometry (LC-MS/MS) experiments do not require a separate blot or stain to ensure equal loadings of samples[6,8,9]. Instead, data collected during sample preparation through the use of protein or peptide abundance assays as these abundance values are useful for reaction optimizations. For example, the bicinchoninic acid assay (BCA) is often used to determine total protein mass to optimize digestion conditions while the quantitative colorimetric peptide (QCP) assay can be used to

measure peptide amounts post-digestion, a factor important to optimize chemical labelling with reagents such as Tandem Mass Tags (TMT)[14].

While protein and peptide abundance measurements upstream of LC-MS/MS analysis provides an adequate estimation for normalization, data acquired from the LC-MS/MS analysis can also be used as a loading control[9,14,15]. The number of peptide spectral matches (PSMs) or a form of the total ion signal observed are common values used for normalization in these types of experiments can be enough for label-free type experiments. However, due to the compositional nature of experiment utilizing isobaric tag reagents like TMT, additional precautions must be taken to ensure load control normalization is valid[14,16]. Namely, the ratio of each TMT reporter ion channel with any other channel must be close to 1:1 to ensure equal sampling of all reporter ion channels during any trap-based analysis (e.g. Orbitrap or linear ion trap analysis).

To that end, our research group has adopted a TMT protocol that involves the LC-MS/MS analysis of a small mixture of our TMT-labelled samples (utilizing either BCA or QCP assay information) prior to the combination of the bulk of our samples[14]. This step, which we term a ratio check, is a direct sampling of how the instrument will analyze the combined sample and is no longer a surrogate or proxy. As such the information provided from a ratio check can assess how well the BCA or QCP assay estimated a 1:1 relationship among all TMT channels. Correcting these values establishes a closer 1:1 mixture of all TMT-labelled samples.

In this study, we show how the precision of BCA and QCP assay can be insufficient in estimating a 1:1 ratio for TMT labelled samples and how ratio check corrected samples exhibit mixings truer to 1:1. Additionally, we provide a step-by-step protocol on how to perform and utilize a ratio check. Lastly, we introduce a software tool, RatioCheckR that can accept data from

common mass spectrometry data processing tools, such as MaxQuant and Proteome Discoverer, to assist researchers who are keen on incorporating a ratio check step in their current workflows.

Methods

Human tissue culture

SH-SY5Y and PANC1 were obtained from ATCC and grown as previously described [17,18]. Briefly, cells were cultured in Dulbecco's Modified Eagle Media supplemented with 10% (v/v) Fetal Bovine Serum and a 5% (v/v) penicillin/streptomycin cocktail. After washing with cold phosphate buffered saline (PBS), cells were lysed on plate. Lysis was performed on ice when cells were confluent using an 1mL of an 8M Urea lysis solution with a miniComplete protease inhibitor tablet (Roche) buffered to pH 8.5 with 200mM EPPS. Cell lysates were homogenized by trituration through a 21-gauge syringe to shear DNA. After this, lysates were pelleted by centrifugation at 21,000 x g for 15 minutes at room temperature. Clarified lysate was then processed for LC-MS/MS analysis.

LC-MS/MS sample preparation

Cell lysates were processed for LC-MS/MS analysis as previously described[14]. Protein abundances for each lysate was determined by a commercially available BCA kit (Pierce). Absorbance was measured at 562nm. Lysates were reduced by 5mM of tris(2-carboxyethly)-phosphine (TCEP) and reactive cysteines were alkylated in the dark with 10mM iodoacetamide. To prevent over alkylation, the reaction was stopped by adding 15mM of dithiothreitol to the solution before precipitation by the chloroform-methanol method[14].

Precipitates were then resuspended in 200mM EPPS buffered to pH 8.5 for digestion with Lys-C (Wako) at a 1:100 protease:protein ratio. Lys-C digestion was performed overnight with vigorous shaking at room temperature. Sequencing grade trypsin (Promega) was then added and

the sample was transferred to an orbital shaker in a 37°C warm room for additional digestion for 6 hours. Peptides from digested proteins were measured by the commercially available quantitative colorimetric peptide (QCP) assay (Pierce). Absorbance was measured at 480nm. Samples were then labeled with TMT-10 reagents for 90 minutes and then quenched with hydroxylamine. 2µg of each TMT-labelled sample was combined by utilizing protein or peptide concentrations from BCA or QCP assay measurements, respectively. Combined samples were then subjected to LC-MS/MS analysis as a ratio check. Ratio check corrected BCA and QCP assay numbers were then utilized to generate two more combined samples which were submitted to LC-MS/MS analysis as ratio checks.

Liquid chromatography and tandem mass spectrometry

All LC-MS/Ms analyses were performed on a Q-Exactive mass spectrometer (Thermo Fisher Scientific, San Jose, CA) with LC separation performed on a Famos 920 autosampler (LC Packings) attached to a quaternary Accela 600 pump (Thermo Fisher). 35cm of Accucore C18 resin (2.6µm, 150Å, ThermoFisher) was packed in a 100 µm inner diameter microcapillary column to separate peptides. Approximately 2µg of labelled material was loaded onto the column.

A 120 minute gradient ranging from 6% to 25% acetonitrile in 0.125% formic acid was used as the mobile phase for the liquid chromatography. Flow rate was set to ~450nL/min. MS1 spectra were collected as centroids with the instrument running in positive mode: Orbitrap resolution – 70,000; mass range – 300 to 1,500 m/z; Automatic Gain Control (AGC) target – 3 x 10^6; Maximum ion injection time – 250 ms. Precursors for MS2 analysis were selected using a Top 10 method. MS2 spectra were collected after high energy collision induced dissociation (HCD, normalized collision energy: 32%). High resolution MS2 spectra were collected as

centroids: Orbitrap resolution – 35,000; Fixed first mass – 110.0 m/z; AGC target – 1 x 10^6; Maximum ion injection time – 100 ms; Isolation width – 1.6 Th.

Data Analysis

All spectra collected during LC-MS/MS analyses were searched using an in-house Sequest-based software suite[14,19,20]. Raw spectra data were converted into mzXML format and searched against a human proteome database (Uniprot Database ID: 9606, downloaded February 4, 2014) that was concatenated with common contaminants and a database containing all protein sequences reversed. Ion tolerance for precursors and product ions was set to 50ppm and 0.03Da, respectively. Methionine oxidation (+15.99491) was added as variable modification while cysteine alkylation (+57.02146) and TMT labelling (+229.16293) on lysine and the peptide N-terminus were added fixed modifications.

False discovery rate (FDR) was determined by the target-decoy method and PSMs were filtered to a 1% FDR at the peptide level using linear discriminant analysis as described previously[21,22]. Protein were then identified and collapsed to 1% FDR level by the principles of parsimony. Signal-to-noise ratios for TMT reporter ions were extracted from high resolution MS2 scans. MS2 spectra were filtered such that the signal-to-noise ratio of all TMT reporter ions was greater than 200 and the scan's isolation specificity, also known as the scan's isolation purity, was at least 0.5[23]. Protein level abundances were then calculated as the sum of peptide reporter ion signal-to-noise values. Normalization across TMT channels was then performed utilizing RatioCheckR. These normalization factors were then used to adjust BCA and QCP assay protein concentration estimates.

Results

Data for total protein amount normalization can be acquired during various experimental steps

Most LC-MS/MS experiments require that researchers determine a sample's protein and/or peptide abundances prior to analysis on a mass spectrometer. To demonstrate how estimates from traditional biochemical assays can deviate from 1:1 estimations, we designed an experiment to compare an equal mixture of the neuronal-like cell line SH-SY5Y and the pancreatic cell line PANC1 (Figure A-1A). Quantitative information about sample abundances were taken at three distinct phases of the experiment: I) protein abundances by BCA, II) peptide abundances by QCP assay, and III) reporter ion relative abundance by ratio check.

Acquisition of quantitative information by each method has benefits and caveats (Figure A-1B). Traditional biochemical assays can often be quick and are readily available in most laboratories. However, when using these assays to quantify samples for LC-MS/MS analysis important caveats must be considered, including compatibility with lysis and digestion buffers as well as the relatively large amount of starting material (μg-scale) required. While ratio checks circumvent these issues, the method is not without its caveats, most notably the consumption of instrument time and sample.

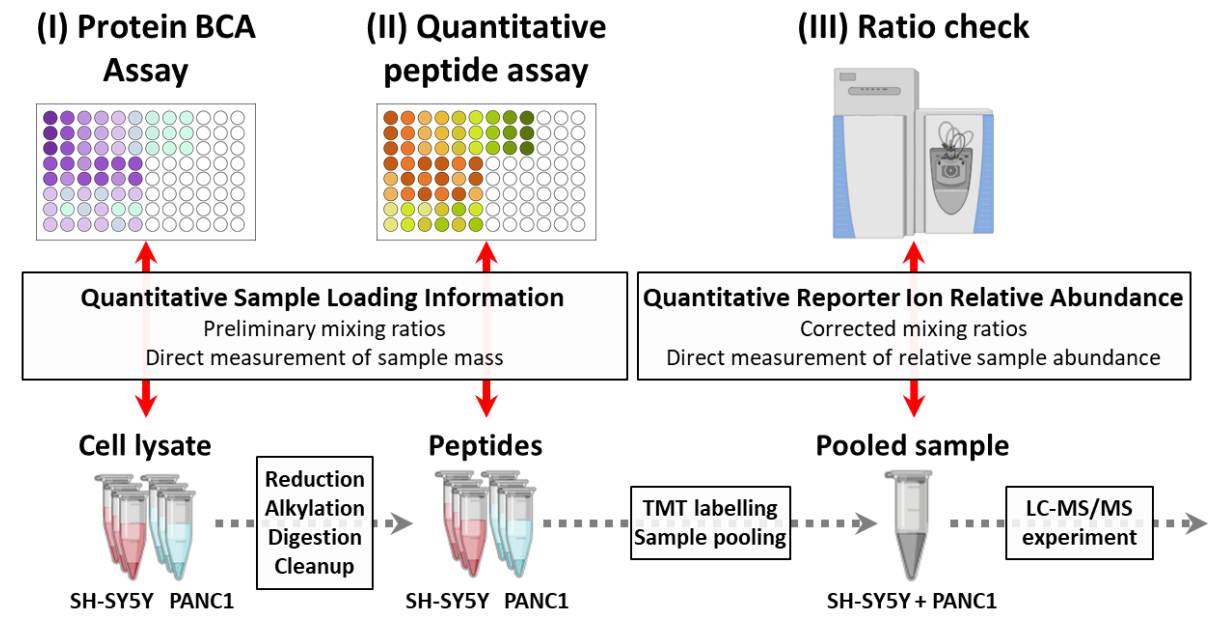
Ubiquitously expressed proteins can still deviate from a 1:1 ratio

As a method of loading control is to utilize a single protein for normalization, we examined whether a ubiquitously expressed protein would show a 1:1 ratio in our LC-MS/MS analyses. The protein GNB1L1, also known as RACK1, is a cytosolic protein that is a core member of the 40S ribosomal subunit[24,25]. Due to its biological function, it is an ideal candidate as a single protein loading control.

LC-MS/MS analysis of the RACK1 peptide YWLCAATGPSIK showed noticeable deviations of the expected 1:1 ratio (Figure A-2). Samples mixed with quantitative data acquired from the BCA or QCP assay showed the largest deviation of reporter ion ratios

Figure A-1: Overview of the quantitative assays used to generate loading controls in LC-MS/MS experiments. **A)** Experimental workflow for a typical TMT experiment. Roman numerals indicate quantitative steps that can be used to inform how samples should be pooled in a multiplex experiment. Additionally, protein and peptide concentrations from BCA and quantitative peptide assays can be used to determine the amount of digestion enzymes and labelling reagents to use. **B)** A high-level analysis of the benefits and caveats of quantitation methods used to determine sample mixing for multiplexed experiments.

A.



B.

	I. BCA/Protein Assays	II. QCP/Peptide Assays	III. Ratio Checks
Benefits	• Quick and simple • Provides protein concentration for optimum enzymatic digestion	• Accounts for digestion efficiency • Accounts for errors in precipitation/sample cleanup recovery	• All benefits of QCP/Peptide assays • Low amount of sample material required (<μg) • Accounts for flyability and detectability of sample • Provides quality control statistics (e.g. missed cleavage rates, PPM, etc)
Caveats	• Buffer compatibility • Sample must be within linear range • Requires relatively large amount (μg) of sample material	• Buffer compatibility • Sample must be within linear range • Requires relatively large amount (μg) of sample material	• Requires instrument time • Sample must be digested and labelled

(Figure A-2A,B). These ratios were closer to 1:1 in ratio check-corrected data (Figure A-2C,D). These findings suggest that single protein normalization should be avoided due to potential variations and that ratio check-corrected data should be used to inform the mixing of TMT-labelled samples. However, in experiments where single protein normalization is unavoidable (e.g, immunoprecipitation where normalization to a single bait protein is required), researchers should be aware of this limitation and take appropriate action[15].

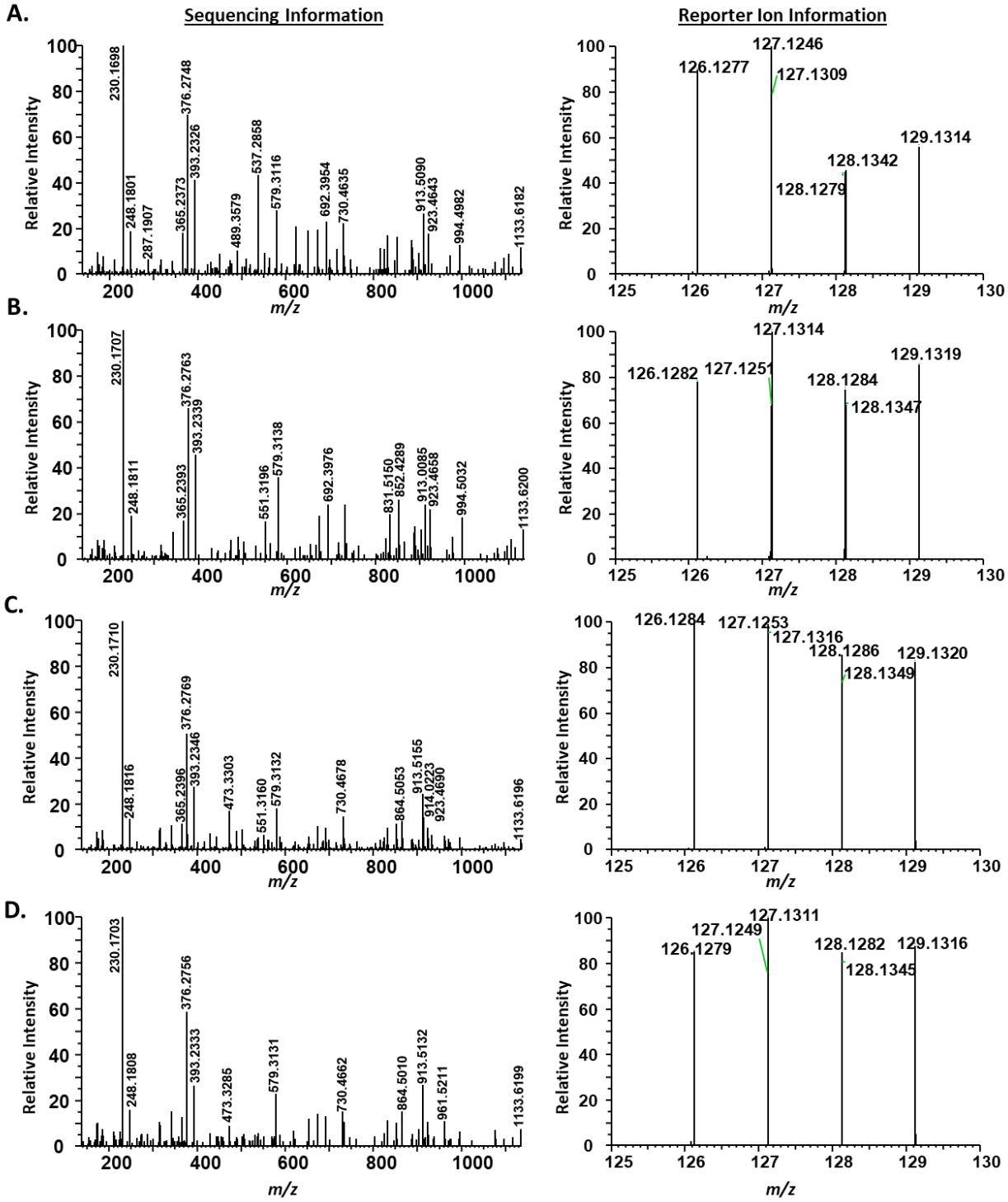
Ratio check corrected estimations are truer to a 1:1 ratio

Protein abundances for both cell lines were measured by BCA while peptide abundances were measured by QCP assay. For both protein and peptide biochemical assays, three technical replicates were performed for three dilutions (1:2, 1:10, and 1:50) for each cell line (Figure A-3). Replicates of the appropriate dilution showed good grouping along the standard curve in both assays. However, only one 1:50 replicate exhibited absorbance above background at 480nm during the QCP assay, highlighting the importance of remaining within the linear range of the biochemical assay (Figure A-2B).

When utilizing BCA and QCP assays to generate 1:1 mixtures of TMT-labelled samples noticeable deviations from the 1:1 ratio were observed (Figure A-4A,B). In some cases, a 2:1 ratio, or 50% deviation of the most abundant sample, was observed when utilizing protein quantitation from BCA (Figure A-4A). QCP assay values generated values closer to 1:1 with the largest deviation from the most abundant sample being 22%. Normalization values were calculated from these deviations and samples were remixed.

The ratio check-corrected normalized values displayed tighter groupings closer to the target 1:1 ratio (Figure A-4C,D). Ratio check corrected BCA normalized ratios displayed the most precision with the largest deviation being 6% compared to the largest deviation of ratio check

Figure A-2: Spectra for the peptide YWLCAATGPSIK from the protein GNB2L1 from samples mixed using. **A,B)** BCA protein loading information and **C,D)** QCP peptide loading information. Region of the MS2 spectra used for sequencing is displayed on the left while the region used for reporter ion quantification is displayed on the right. Samples mixed using uncorrected values are **A,C** while ratio check corrected samples are **B,D**.


A.
Sequencing Information
Reporter Ion Information
Relative Intensity
m/z
230.1698
376.2748
248.1801
393.2326
537.2858
365.2373
579.3116
692.3954
730.4635
913.5090
923.4643
994.4982
1133.6182
287.1907
489.3579
127.1246
126.1277
127.1309
129.1314
128.1342
128.1279
B.
230.1707
376.2763
393.2339
248.1811
365.2393
579.3138
551.3196
692.3976
831.5150
852.4289
913.0085
923.4658
994.5032
1133.6200
127.1314
129.1319
126.1282
127.1251
128.1284
128.1347
C.
230.1710
376.2769
393.2346
248.1816
365.2396
473.3303
551.3160
579.3132
730.4678
864.5053
913.5155
914.0223
923.4690
1133.6196
126.1284
127.1253
127.1316
128.1286
129.1320
128.1349
D.
230.1703
376.2756
393.2333
248.1808
473.3285
579.3131
730.4662
864.5010
913.5132
961.5211
1133.6199
127.1311
127.1249
126.1279
128.1282
129.1316
128.1345

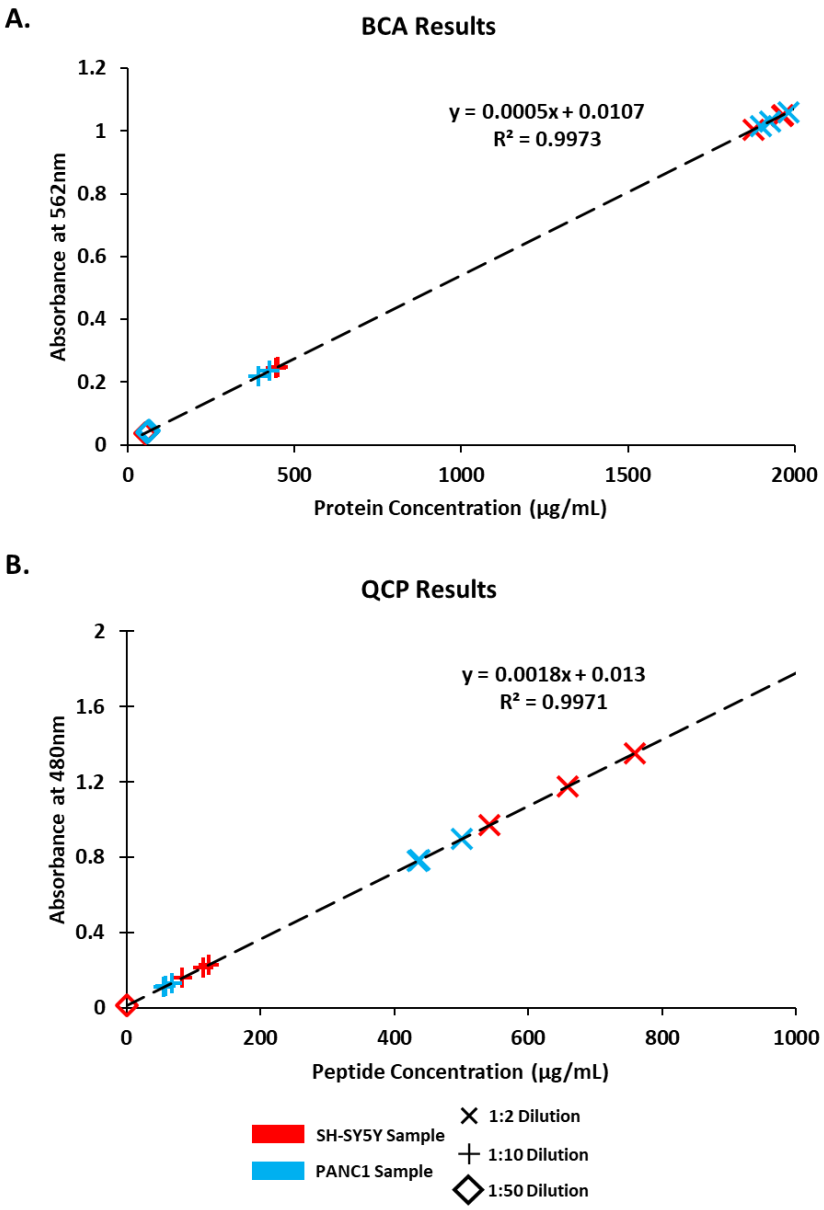


Figure A-3: Plots of protein and peptide assay data. Measured by **A)** BCA or **B)** QCP, respectively. Three dilutions were used for each sample and 3 replicates for each dilution were performed. Note that only 1 replicate from the 1:50 dilution is shown in the QCP assay results as all other replicates obtained 480nm absorbance values that were too low to quantify over background, thereby stressing the importance of remaining in the linear range. Dashed line represents linear regression line fit to protein or peptide standards supplied with the BCA and QCP assay kits, respectively.

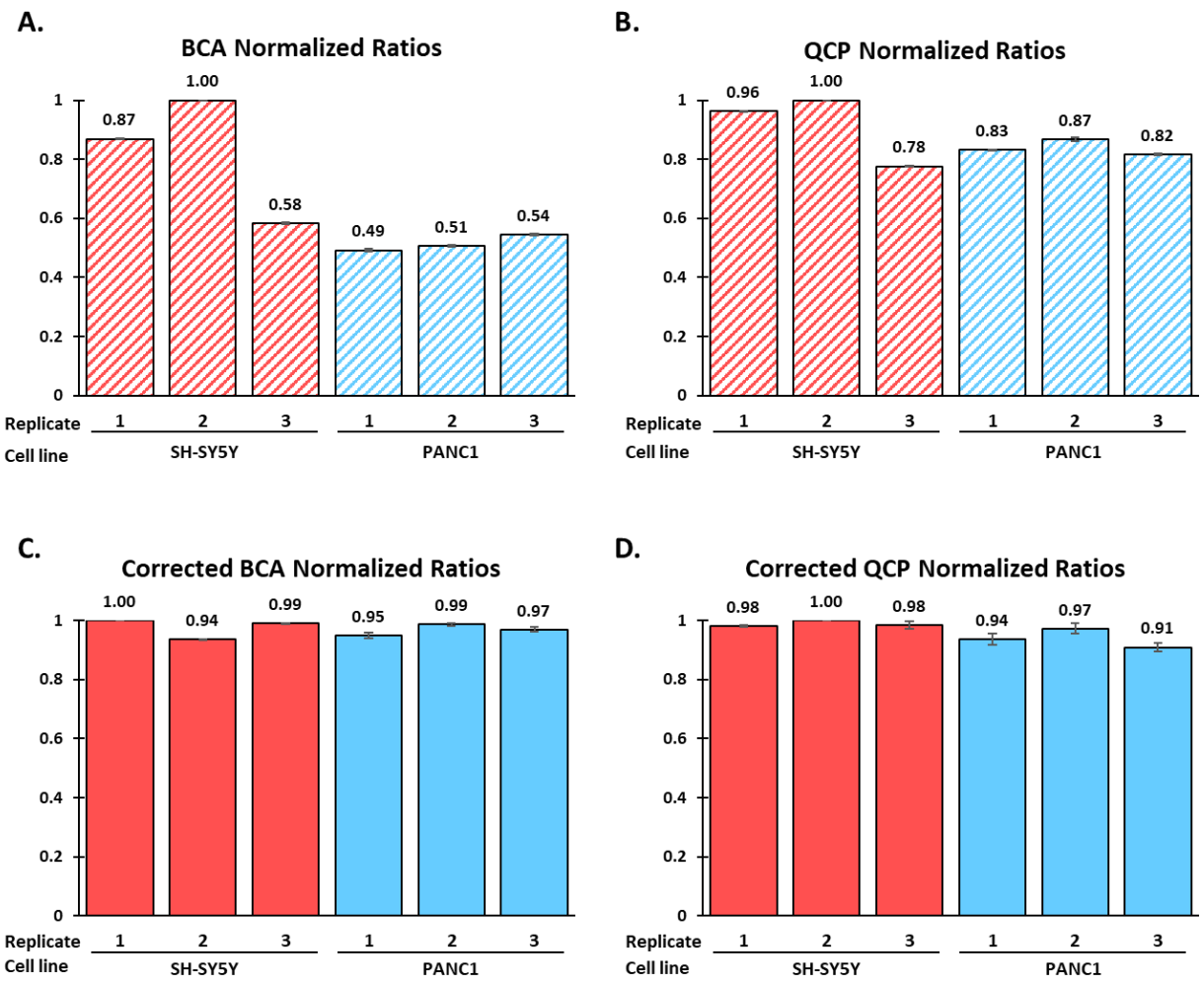


Figure A-4: Bar charts depicting the normalized TMT reporter ion relative abundances for samples. Mixed using **A)** BCA or **B)** QCP protein or peptide abundance measurements, respectively, to control for loading. Sample mixing was corrected using information from normalized TMT reporter ions and reanalyzed by LC-MS/MS. **C)** Corrected BCA and **D)** Corrected QCP.

corrected QCP normalized ratio being 9%. However, all ratios were noticeably closer to 1:1 thus reducing estimation errors due to the compositional nature of TMT data.

Discussion

Establishing the basis for fair comparisons is critical component of any quantitative experiment. Here we show how utilization of BCA and QCP assays to acquire protein and peptide estimations for sample mixing in TMT-based experiments may not always yield a true 1:1 ratio. Although these biochemical tools are frequently used, they are prone to complications including buffer compatibility and a limited linear range. As a result, a 1:1 ratio estimated by these assays can be as deviant as a 2:1 ratio creating potential downstream normalization problems. While these deviations may not affect label-free analyses as dramatically, the compositional nature of isobaric tag-based proteomic experiments, such as TMT, prefers that the total amount of material for all samples remain as true to a 1:1 ratio as possible[16].

Additionally, as data acquisition is performed using the mass spectrometer, estimates by BCA and QCP assays do not factor in properties such as ionization efficiency and detectability in the instrument. By performing a ratio check, a short high resolution MS2 analysis, mixing ratios closer to 1:1 can be achieved. Here we show that, in our hands, the maximum deviation from a 1:1 ratio for after ratio check correction was 9%, or a ratio of 1:1.10. Supplementing either the BCA or the QCP assay in a standard workflow with a ratio check will not dramatically increase the workload and can provide increased quantitative accuracy. Furthermore, the software tool used to calculate the normalization factors to correct the BCA or QCP assay estimates has been provided so that other researchers can easily integrate this analysis step.

References

1. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 6940–5 (2003).

2. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).

3. Speicher, D. W. Proteome bioinformatics. *Curr. Protoc. Protein Sci.* **1549**, (2010).

4. MacCoss, M. J. Computational analysis of shotgun proteomics data. *Curr. Opin. Chem. Biol.* **9**, 88–94 (2005).

5. Ghosh, R., Gilda, J. E. & Gomes, A. V. The necessity of and strategies for improving confidence in the accuracy of western blots. *Expert Rev. Proteomics* **11**, 549–560 (2014).

6. Welinder, C. & Ekblad, L. Coomassie staining as loading control in Western blot analysis. *J. Proteome Res.* **10**, 1416–1419 (2011).

7. SUZUKI, O., KOURA, M., NOGUCHI, Y., UCHIO-YAMADA, K. & MATSUDA, J. Use of Sample Mixtures for Standard Curve Creation in Quantitative Western Blots. *Exp. Anim.* **60**, 193–196 (2011).

8. Aldridge, G. M., Podrebarac, D. M., Greenough, W. T. & Weiler, I. J. The use of total protein stains as loading controls: An alternative to high-abundance single-protein controls in semi-quantitative immunoblotting. *J. Neurosci. Methods* **172**, 250–254 (2008).

9. Wiśniewski, J. R. & Mann, M. A proteomics approach to the protein normalization problem: Selection of unvarying proteins for MS-based proteomics and western blotting. *J. Proteome Res.* **15**, 2321–2326 (2016).

10. Dang, W. & Sun, L. Determination of internal controls for quantitative real time RT-PCR analysis of the effect of Edwardsiella tarda infection on gene expression in turbot (Scophthalmus maximus). *Fish Shellfish Immunol.* **30**, 720–728 (2011).

11. Barber, R. D., Harmer, D. W., Coleman, R. A. & Clark, B. J. GAPDH as a housekeeping gene: Analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* **21**, 389–395 (2005).

12. Dittmer, A. & Dittmer, J. β-Actin is not a reliable loading control in Western blot analysis. *Electrophoresis* **27**, 2844–2845 (2006).

13. Ron-Harel, N. *et al.* Defective respiration and one-carbon metabolism contribute to impaired naïve T cell activation in aged mice. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 13347–13352 (2018).

14. Navarrete-Perea, J., Yu, Q., Gygi, S. P. & Paulo, J. A. Streamlined Tandem Mass Tag (SL-TMT) Protocol: An Efficient Strategy for Quantitative (Phospho)proteome Profiling Using Tandem Mass Tag-Synchronous Precursor Selection-MS3. *J. Proteome Res.* **17**, 2226–2236 (2018).

15. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).

16. O'Brien, J. J. *et al.* Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags. *J. Proteome Res.* **17**, 590–599 (2018).

17. Paulo, J. A., Mancias, J. D. & Gygi, S. P. Proteome-Wide Protein Expression Profiling Across Five Pancreatic Cell Lines. *Pancreas* 1 (2017). doi:10.1097/MPA.0000000000000800

18. Stepanova, E., Gygi, S. P. & Paulo, J. A. Filter-Based Protein Digestion (FPD): A Detergent-Free and Scaffold-Based Strategy for TMT Workflows. *J. Proteome Res.* **17**, 1227–1234 (2018).

19. Paulo, J. A., O'Connell, J. D. & Gygi, S. P. A Triple Knockout (TKO) Proteomics Standard for Diagnosing Ion Interference in Isobaric Labeling Experiments. *J. Am. Soc. Mass Spectrom.* **27**, 1620–1625 (2016).

20. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**, 1174–1189 (2010).

21. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

22. Du, X. *et al.* Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. *J. Proteome Res.* **7**, 2195–2203 (2008).

23. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

24. Sengupta, J. *et al.* Identification of the versatile scaffold protein RACK1 on the eukaryotic ribosome by cryo-EM. *Nat. Struct. Mol. Biol.* **11**, 957–962 (2004).

25. Gerbasi, V. R., Weaver, C. M., Hill, S., Friedman, D. B. & Link, A. J. Yeast Asc1p and Mammalian RACK1 Are Functionally Orthologous Core 40S Ribosomal Proteins That Repress Gene Expression. *Mol. Cell. Biol.* **24**, 8276–8287 (2004).

Supplemental Table 1

Supplemental Table 1 is a multi-sheet Microsoft Excel document that contains all peptides and proteins identified during LC-MS/MS experiments conducted in Chapter 3 – "Evaluation False Transfer Rates from the Match-Between-Runs Algorithm with a Two-Proteome Model." The table can be downloaded from the journal website at the following link: https://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.9b00492/suppl_file/pr9b00492_si_001.xlsx Alternatively, the table can be accessed through ETDs @ Harvard.

Supplemental Table 2

Supplemental Table 2 is a Microsoft Excel document that contains an annotated list of localized phosphorylation events during phosphopeptide enrichment LC-MS/MS experiments conducted in Chapter 4 – "Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling." The table can be downloaded from the journal website at the following link:
https://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.7b00571/suppl_file/pr7b00571_si_002.xlsx
Alternatively, the table can be accessed through ETDs @ Harvard.

Supplemental Table 3

Supplemental Table 3 is a multi-sheet Microsoft Excel document that contains all peptides quantified during LC-MS/MS phosphatase treated TMT experiments conducted in Chapter 4 – "Improved Method for Determining Absolute Phosphorylation Stoichiometry Using Bayesian Statistics and Isobaric Labeling" that were previously identified to contain a phosphorylation event (See Supplemental Table 2 for list of phosphorylation events identified in the experiment after phosphopeptide enrichment). The table can be downloaded from the journal website at the following link:

https://pubs.acs.org/doi/suppl/10.1021/acs.jproteome.7b00571/suppl_file/pr7b00571_si_003.xlsx

Alternatively, the table can be accessed through ETDs @ Harvard