Methods for harmonizing and calibrating quantitative mass spectrometry experiments

Lindsay Kristina Pino

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Michael J. MacCoss, Chair

William Stafford Noble, Chair

Andrew N. Hoofnagle

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Methods for harmonizing and calibrating quantitative mass spectrometry experiments

Lindsay Kristina Pino

Co-Chairs of the Supervisory Committee:
Professor Michael J. MacCoss
Genome Sciences

Professor William Stafford Noble
Genome Sciences

The field of mass spectrometry proteomics has made great technological progress, and these techniques are now being used to address essential questions in basic biology and are increasingly being used on samples of clinical significance. In particular, the development of data independent acquisition mass spectrometry (DIA-MS) has made it possible to measure tens of thousands of peptides from a protein digest in a 1-2 hour time scale. As generating larger and larger proteomic data sets becomes easier and easier, questions about normalizing batch effects and assessing data quality have arisen in the mass spectrometry community. In the following chapters, I describe three projects that aimed to address various challenges associated with scaling up quantitative mass spectrometry experiments.

I first introduce the need for reference materials for mass spectrometry proteomics. In Chapter 2, I describe a single-point external calibration strategy to calibrate signal intensity measurements to a common reference material, which places MS measurements on the same scale and harmonizes signal intensities between instruments, acquisition methods, and sites. In Chapter 3, I extend the reference material calibration approach to multi-point calibration and demonstrate the consequences of linearity in quantitative analyses. We apply this approach to yeast lysate, human cerebrospinal fluid, and formalin-fixed paraffin-embedded samples. In Chapter 4, I apply the methods

developed in the previous two chapters to investigate the yeast proteome response to genetic and environmental modulators of replicative lifespan. I show that the protein-level signatures associated with replicative lifespan extension suggest a higher-level response beyond protein abundances. Lastly, I present closing remarks and future directions in Chapter 5.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without the support of many people. First, thank you to my advisors, Mike MacCoss and Bill Noble, for your guidance and encouragement not only on these thesis projects but my overall career development. To Mike, thank you for continuously pushing me past my self-imposed limitations and being the nucleus for many of my closest professional relationships; to Bill, thank you for believing in my potential and thank you for your patience while I came to realize it. Thank you to Andy Hoofnagle for your specific guidance on and scientific contributions to two of these thesis chapters. Also, I would like to thank the entirety of my thesis committee, Maitreya Dunham, Matt Kaeberlein, and Shao-En Ong, for your many valuable contributions to not only this work but my future directions. I hope to continue learning from you all.

The experimental and computational work in this thesis reflects my fortunate circumstance of being jointly trained in two labs. I'd like to thank Alex Hu, Andy Lin, Will Fondrie, and Wout Bittremieux of the Noble Lab for their assistance with programming and computational proteomics. In the MacCoss lab, I'd especially like to thank Han-Yin Yang for being a constant source of advice and camaraderie; and I'd also like to thank Brian Searle for becoming like a third thesis adviser to me both in scientific contributions and in professional development. The friendships I've made during my graduate career have also supported this work, and so I'm thankful to Damien Wilburn for many science talks over beers and especially to Hannah Pliner for the miles of hiking, backpacking, and dog park-walking.

I would not have considered an advanced degree without the encouragement of my past supervisors, Sue Abbatiello and Steve Carr.

Finally, thank you to my family for building me into the person I am today. Thank you to my

siblings, Andrew and Katie, for many years of teaching me the importance of a good role model, and for being such good role models yourselves. Thank you to my parents, John and Elaine, for prioritizing our education even when it came at personal expense, and for your patience and unconditional love while I figured out my place in the world. Last, thank you to Alex for supporting my decision to go back to school by dropping everything, moving thousands of miles away to be here for me physically and emotionally through timepoint experiments, instrument failures, and procrastinated deadlines.

DEDICATION

For all those family, friends, and colleagues who believed in me.

Chapter 1: INTRODUCTION

This chapter is adapted from the following works:

Pino LK, Searle BC, Bollinger JG, Nunn B, MacLean BX, MacCoss MJ. (2017) The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. Mass Spectrometry Reviews.

Pino LK, Just S, MacCoss MJ, Searle BC. (2019) Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *(in preparation)*

1.1 THE ROLE OF PROTEINS IN BIOMEDICAL RESEARCH

Since the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2001, 2004; Venter et al., 2001), a wealth of functional genomic techniques have emerged as the focus of research shifts to assigning function and understanding the regulation of each of the identified gene products. The focus of these efforts is to better understand how the information stored in a genome encodes all the complexity necessary to sustain a complex multicellular organism (Lander, 2011). Nothwithstanding impressive gains in these technologies, interpretation of their results is limited without corresponding data on proteins, the primary functional macromolecules encoded by the genome. This limitation is highlighted by the observation that measurements performed at the nucleic acid level (e.g. transcriptomic studies using microarray or RNA-Seq methods) tend to correlate very poorly with those performed at the protein level (Greenbaum et al., 2003; Schrimpf et al., 2009), especially in cases when experimental noise is not considered (Csárdi et al., 2015). A combination of factors likely contribute to the poor protein-transcript correlation, including the variable lifetime of each protein dictated by its respective synthesis and degradation rates; the existence of multiple different forms of each transcript product due to post-translational modifications; and finally, the temporal/spatial regulation imparted by protein complexes and the highly compartmentalized nature of cellular processes. Accordingly, the direct analysis of proteins, albeit more technically challenging, is absolutely crucial to a complete

understanding of gene regulation and systems biology.

1.2 Analysis of proteins by mass spectrometry

To meet these ends, tandem mass spectrometry (MS/MS) has emerged as the dominant analytical platform for the direct characterization of the protein fraction from complex biological matrices (Ong and Mann, 2005). A majority of mass spectrometry-based proteomic workflows have utilized a *bottom-up* approach in which proteins are extracted from a sample via lysis (cell culture) or homogenization (tissue), linearized, and digested with an endoprotease like trypsin (Figure 1.1a). The resulting peptide mixture is commonly separated via nano-flow reverse-phase liquid chromatography (LC), which separates peptides by hydrophobicity. As peptides are separated, they are ionized and emitted directly into a mass spectrometer for measurement as a continuous stream of ions.

When a peptide ion (*precursor*) enters the mass spectrometer, it generally encounters three stages: precursor selection, fragmentation, and detection. Although mass spectrometer hardware configurations perform these steps differently – ion detection in particular – these three stages summarize most of the methods used for proteomics. During precursor selection, the mass spectrometer is tuned to only allow a specific mass-to-charge (*m/z*) value to pass through the instrument. All other precursors entering the mass spectrometer at the same time are filtered out of the ion stream. Precursors with the selected *m/z* are then passed to the collision cell, where they are bombarded with gas molecules, causing the backbone (amide bonds) between amino acids to break (Figure 1.1b). Typically, the energy required to break the backbone only allows for a single fragmentation event for a given precursor molecule, resulting in two fragment ions from the original intact precursor. However, because there are multiple precursor molecules being fragmented, across the population of precursor molecules there occurs a distribution of fragmentation events. Theoretically, any fragmentation event could occur at any time, resulting in an equal abundance of all possible fragment ions; in practice, some fragmentation events are more favorable than others, resulting in those fragment ions being more abundant.

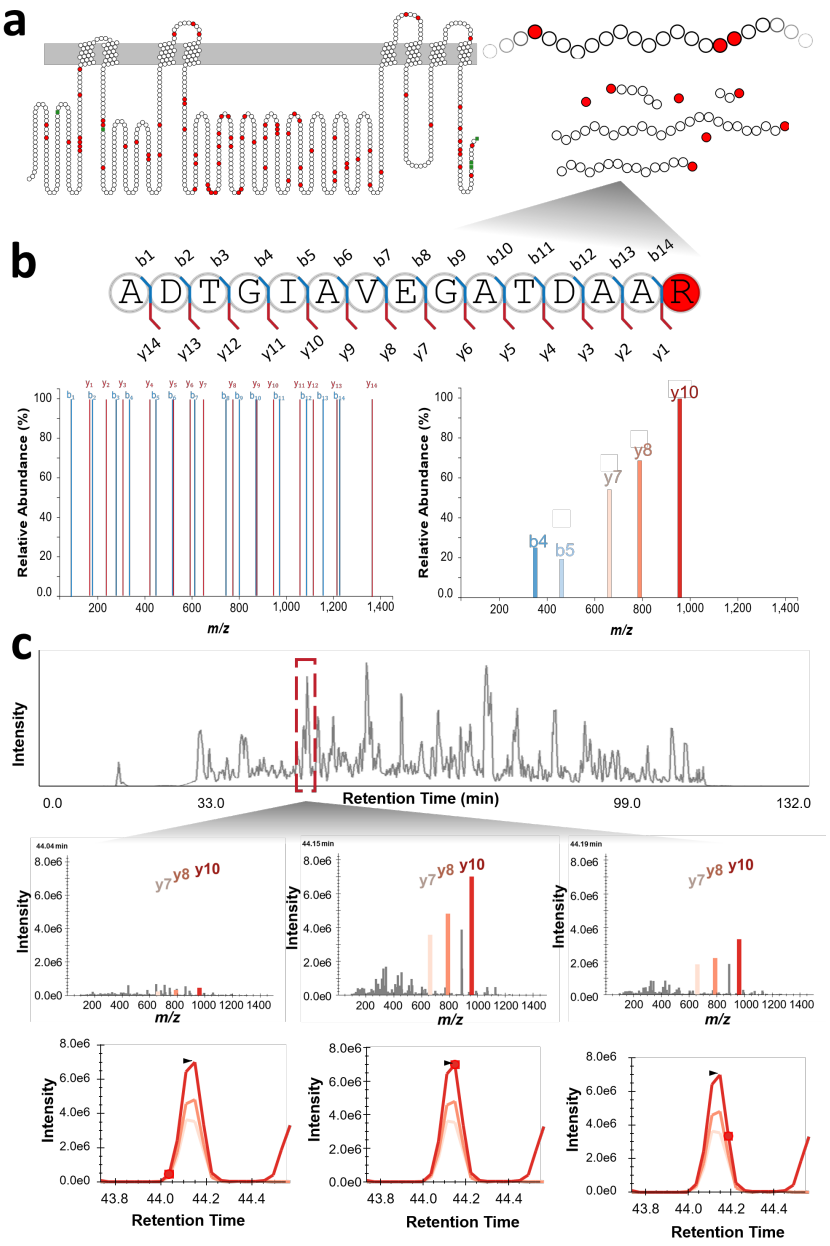


Figure 1.1: Bottom-up proteomics from protein sample to chromatogram. (*Legend continued on the following page*)

Figure 1.1: (continued) (a) Proteins exist in high-order structures under biological contexts, visualized here as yeast protein Pma1, a membrane-bound protein composed of 918 amino acids depicted as circles with the lysine and arginine residues colored in red. The first step in a bottom-up proteomics workflow is to solubilize proteins, linearize them with detergents or high salt solutions, and then digest the protein sequence into peptides using trypsin, cleaving the protein sequence at lysines and arginines. (b) Digested peptides are ionized and then fragmented by mass spectrometer. Fragmentation events are theoretically possible at any amide bond between residues, with each fragmentation event creating a *b* and *y* ion piece of the original peptide. In practice, only a portion of possible fragmentation events are observed. (c) The process of fragmentation and measurement is performed continuously while a protein mixture is separated by liquid chromatography, such that fragment measurements of the same precursor are made repeatedly while the peptide is eluting. Fragment spectra from the same peptide can be reoriented along the retention time axis to create fragment ion chromatograms, the basis of most quantitative proteomics experiments.

In the last stage, ions are measured by a detector and reported as intensities. When precursor ions are measured, the resulting data are called *mass spectra* (MS, MS1); when fragment ion measurements are made, they are called *tandem mass spectra* (MS/MS, MS2). Both types of spectra contain the same data: *m/z* values on the horizontal axis and intensities on the vertical. The detection stage is the most variable in quantitative mass spectrometry, with the three most common techniques at the detection stage being quadrupole (triple quadrupole instruments, QqQ), ion trap (in particular the Orbitrap), and time-of-flight (TOF). In this work, we incorporate all three of these instrument configurations (see Chapter 2); later in this chapter, we discuss the differences in assay development based on instrumentation limitations.

For quantitative proteomics applications, the ultimate goal is to construct *fragment ion chromatograms* (Figure 1.1c). As precursors are separated and are analyzed by the mass spectrometer, they produce peaks, gradually increasing in abundance as they elute off the LC column and then decreasing again. Throughout their elution profile, the mass spectrometer is collecting precursors, fragmenting them, and measuring their fragmentation spectra. At the beginning of the precursor elution, these fragment intensities are low; as the precursor reaches the apex of its elution, the fragment intensities also increase. By interpolating the fragment ion intensities over the retention time, a fragment ion chromatogram is constructed, displaying the change in a precursor's fragment ion intensities over time. In MS2-based quantitative proteomics, the precursor's abundance

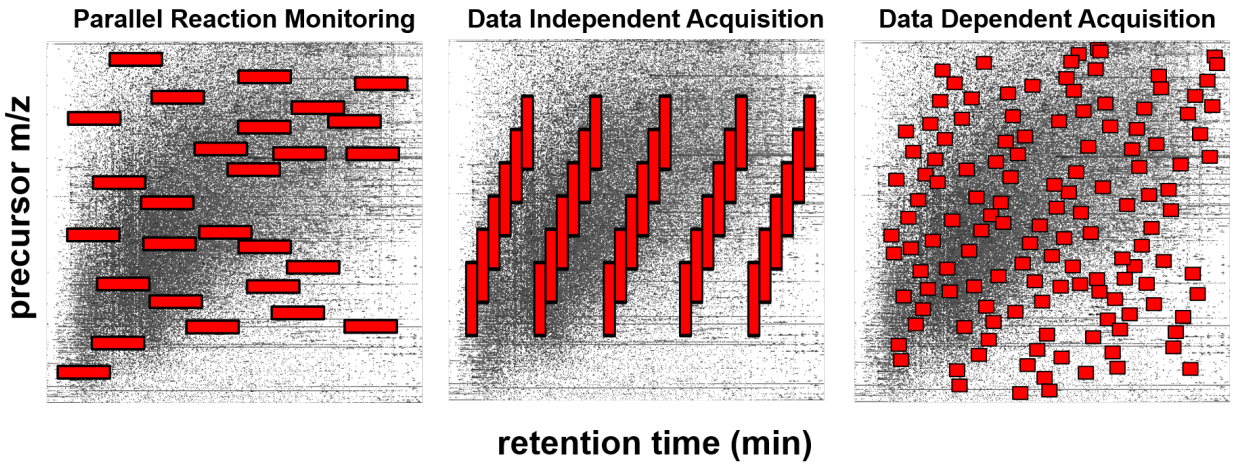


Figure 1.2: Mass spectrometry acquisition strategies common in proteomics. Mass spectrometry data is three dimensional: here, precursor mass-to-charge (*m/z*) on vertical, retention time on horizontal, and intensity as a greyscale color. The same data is shown three times with each of three types of programmed acquisition in red boxes. Left, parallel reaction monitoring (PRM), which acquires predetermined precursor *m/z* for predetermined retention time windows; middle, data independent acquisition (DIA), which cycles through predetermined precursor *m/z* ranges over the duration of retention time; and right, data dependent acquisition (DDA), which stochastically acquires the most intense precursor *m/z* values at discrete retention times.

is calculated as the sum of the integrated areas of its fragment ion's chromatograms.

Both absolute and relative quantitative measurements, reviewed in detail elsewhere (Ong and Mann, 2005), are possible using commonly applied MS acquisition methods (Figure 1.2). Targeted acquisition methods, including selected reaction monitoring (SRM) (Picotti and Aebersold, 2012), also known as multiple reaction monitoring (MRM) (Zhang et al., 2011), and parallel reaction monitoring (PRM) (Peterson et al., 2012), quantify peptides from a preprogrammed list of precursor-fragment pairs and scheduled isolation windows based on previously-determined chromatography elution times. Data-independent acquisition (DIA) (Venable et al., 2004) such as Sequential Window Acquisition of all Theoretical Fragment ion spectra (SWATH) (Gillet et al., 2012) forgo preprogrammed precursor-fragment pairs, widening the isolation windows to activate all ions in a pre-specified mass-to-charge (*m/z*) range. (A detailed review of DIA methodology can be found elsewhere (Chapman et al., 2013; Bilbao et al., 2015a), including peptide-centric

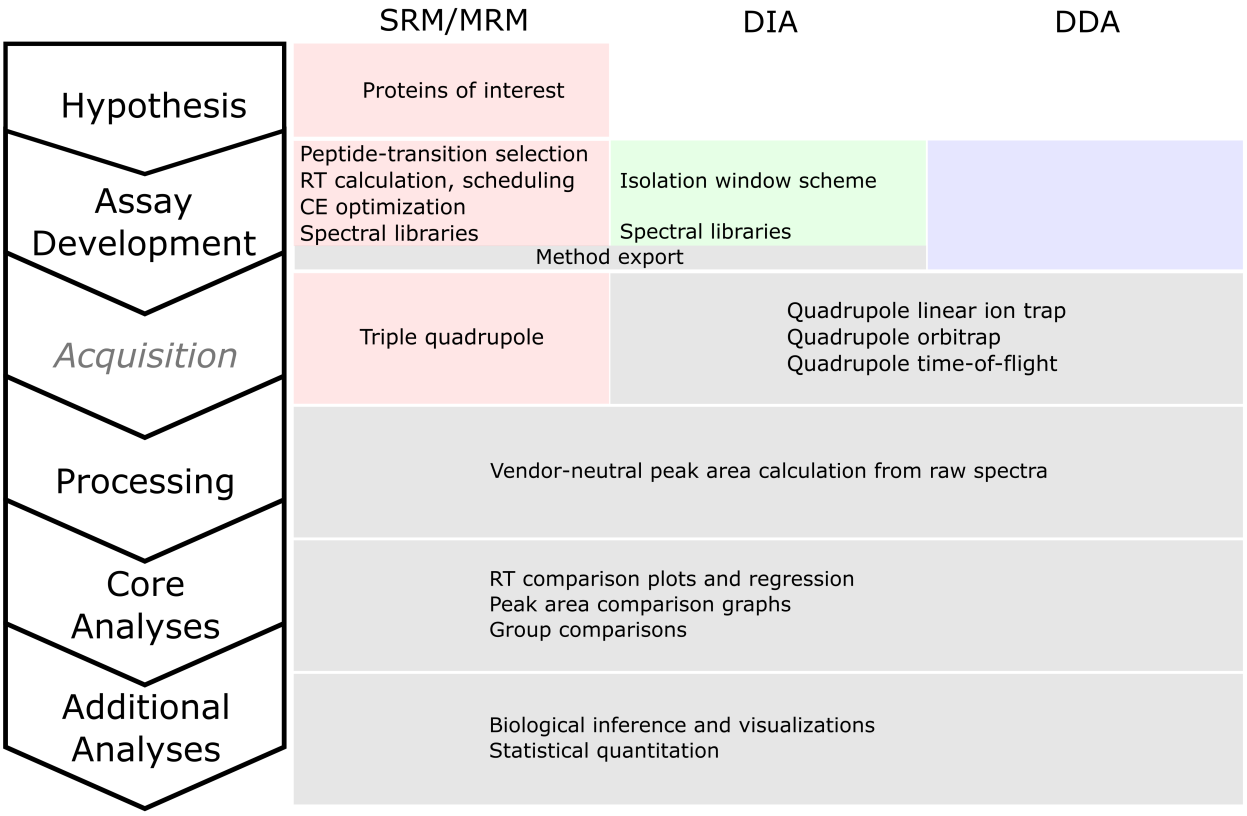


Figure 1.3: Generalized workflow for quantitative MS assay development. Six main steps are outlined, beginning with the development of a hypothesis and continuing through additional analyses, with examples of the associated Skyline ecosystem features.

approaches to DIA (Ting et al., 2015).) It is also possible, through MS1 filtering informatics techniques (Schilling et al., 2012), to use data dependent acquisition (DDA) for quantitative analysis as opposed to conventional detection analysis.

The type of acquisition influences the selectivity, reproducibility, repeatability, limit of detection, dynamic range, and data density of the assay (Domon and Aebersold, 2010). Additionally, acquisition type places specific requirements on assay development and influences the computational strategy for analyzing data. A variety of individual informatics tools have been developed to aid in assay development and to process the data collected with various acquisition types, reviewed elsewhere (Colangelo et al., 2013; Cham Mead et al., 2010).

1.2.1 Targeted mass spectrometry proteomics

The requirements for developing an effective quantitative MS proteomics assay are specific to the type of experiment and the peptide targets being assayed. For all experiments, prior to MS acquisition, it is obligatory to create a program for the instrument that defines the instrument parameters and defines how the data is to be collected by the instrument. In addition, depending on the acquisition mode of the instrument (i.e., SRM/MRM, PRM, DIA and DDA), multiple decisions must be made to optimize the acquisition of the data (Figure 1.3). For example, the experiments with the most intensive assay development, scheduled SRM/MRM and PRM type experiments, necessitate selection of target peptides and their fragment ions (SRM only) prior to acquisition, validation of transitions (that is, a precursor-fragment ion pair) by MS/MS spectra, potentially optimization of individual parameters (such as collision energy - CE) and determination of retention times (RT) for optimal MS instrument scheduling. On the other hand, for DIA experiments, the only required step pre-acquisition is calculating isolation window schemes. In this section, we describe the steps required for assay development, noting which steps are necessary for which experiment types.

Peptide and transition selection for targeted experiments.

Many proteomics hypotheses are rooted in biological observations, and so selecting proteins of interest and peptides that are exclusively representative of those proteins is often the first experimental design step in targeted bottom-up proteomic experiments, such as SRM/MRM and PRM. Selection of peptides for targeted assays is a complex process, involving consideration of (1) specific peptides or amino acid modifications of interest, (2) biological influences on the protein of interest, (3) chemical influences on peptide suitability for MS experiments, and (4) for SRM/MRM experiments, the selection of fragment ions for quantitation.

Specific peptides or amino acid modifications of interest. In the first case, specific amino acid modifications, especially post-translational modifications at the protein level, may dictate a peptide sequence of interest. This is especially seen in targeted phosophoproteomics assays, where

the phosphosite of interest has previously been determined by prior experiments (Schilling et al., 2012; Sherrod et al., 2012; Abelin et al., 2016). In these cases, it may be easiest to manually enter the peptide sequences of interest. Skyline accepts peptides added directly to the document as lists in the Targets window. Peptides added as lists may have modifications and even charge states specified in the added sequence text. They may also be modified manually within Skyline one at a time, or in bulk by changing the Skyline modification settings.

Biological influences on the protein of interest. For situations where the peptide sequence is not defined by the experiment, Skyline accepts lists of proteins, either entered manually, copy-pasted, or as a FASTA file import. After proteins are added to the document, Skyline digests the proteins *in silico* to generate a list of peptides. The result of Skyline's in silico digestion depends on the particular endoprotease specified in the settings of the Skyline document. The most common endoproteases used in bottom-up proteomics are Lys-C, which hydrolyzes specifically at the carboxyl side of lysine; chymotrypsin, which cleaves peptide bonds formed by aromatic residues (e.g. tyrosine, phenylalanine, and tryptophan); GluC, which cleaves peptide bonds C-terminal to glutamic acid residues; and trypsin, which cleaves the carboxyl side of lysine or arginine residues. Other Skyline Peptide Settings that affect results of peptide list generation are common biochemical sample preparation concerns such as missed cleavages, oxidized methionine, and peptide amino acid length (Anderson and Hunter, 2005; Lange et al., 2008; Prakash et al., 2009). After endoprotease(s) are selected and biochemical considerations are defined in the Peptide Settings in the Skyline document, researchers can add proteins of interest to the Skyline target list and Skyline automatically performs *in silico* digestion on the proteins and the resulting peptides are displayed organized by protein of origin.

A point of consideration for proteomics research with clinical applications is the selection of peptides that may have naturally occurring amino acid variations due to individual subjects' genetic backgrounds. Single nucleotide polymorphisms (SNPs) in the genome may give rise to amino acid changes in the final proteoform, which may alter a peptide sequence. To help guide users collecting data on clinical samples that may include SNP-related variation, Skyline provides users with access

to the informatics tool Population Variation (Fujimoto et al., 2013). Population Variation reveals all human sequence variation within a set of user-specified peptides or proteins by identifying the minor allele frequency of peptide targets. The tool then filters SNP data records from dbSNP by criteria directly relevant to proteomics experiments, storing entries with minor allele frequency > 0.01, a non-null protein accession, and a protein-influencing mutation (missense, stop-gain, frameshift). The refined list is stored as a SQLite database and can be accessed through a Skyline plug-in. Running the Population Variation Skyline plug-in outputs a table listing the isoforms and peptide variants for all proteins included in the Skyline document. Researchers can use this output to consider variant peptide targets to ensure that the assay accurately measures.

Chemical considerations of selected peptides. Next, the hypothesis-based, biologically considered peptides must be validated for chemical considerations, namely MS signal robustness. Peptides from the same protein of interest have a range of MS signal response, with some peptides reliably responding strongly and others responding weakly or variably to MS conditions (Kuster et al., 2005). These widely ranging responses are dictated by sequence-specific physiochemical properties (e.g., length of the amino acid sequence, charge, presence of various amino acids, and hydrophobicity) and can be empirically determined using prior knowledge from MS experiments (Stergachis et al., 2011) or by using predictive algorithms.

Empirical determination of high-responding peptides requires performing preliminary MS experiments with the potential targets, often synthesized or purchased, in the intended sample matrix (Stergachis et al., 2011). The mass spectrometrist then evaluates the potential target peptide and transition pairs for signal response and chemical noise interference. Skyline facilitates this empirical evaluation with simple transition deletion and addition tools, including ability to Undo these operations, allowing researchers to easily create or modify transition lists for targeted assay development. Besides empirical determination, however, it is also possible to query past MS experiments to evaluate peptide signal response, making use of Skyline-supported online repositories like PeptideAtlas (Desiere, 2006), Human Proteinpedia (Mathivanan et al., 2008), GPM Proteomics Database Craig and Beavis (2004), and PRIDE (Jones et al., 2007). A caveat to using

repositories, as opposed to an assay-specific preliminary experiment, is that peptide response is not the same across instruments and acquisition methods.

In addition to empirical determination, predictive algorithms provide an alternative or complementary method to select the target peptides most likely to be high-responding for a set of proteins (Mallick et al., 2006; Fusaro et al., 2009; Eyers et al., 2011; Muntel et al., 2014). For researchers interested in using predictive algorithms for SRM/MRM and PRM peptide selection, Skyline has implemented the publically available, open-source PREGO algorithm (Searle et al., 2015) as a plug-in. PREGO predicts high responding peptides using an artificial neural network on DIA experimental data. The artificial neural network was trained using 11 minimally redundant, maximally relevant physiochemical properties that describe peptide size, structure, and hydrophobicity. PREGO outperforms previous predictive algorithms, correctly predicting more high-responding peptides than other algorithms. This performance improvement is believed to stem from a more representative training set. As mentioned above, peptide signal response differs between instruments and acquisition types. PREGO, being trained on a DIA dataset, may perform better because peptide signals in DIA datasets better represent peptide signals in SRM datasets. An important note is that these predictive algorithms mentioned above do not predict transition signal response, only peptide response.

The final number of peptides required for a quantitative assay depend on the analytical rigor of the experiment, the details of the project, and the purpose. A broad rule of thumb is that the quantification of a protein requires at the quantification of at least one peptide unique to the protein's sequence. For basic research purposes, two peptides per protein are used to ensure that the peptide is detected, but validation experiments are rarely performed to confirm analyte reproducibility, dynamic range, and stability. For clinical applications, a protein can be measured by just one peptide; however, the peptide has undergone extensive validation testing to fully characterize the peptide as a measurand. A more thorough discussion of measurement validation and their implications on assay development is described elsewhere (Carr et al., 2014).

Selection of transitions for SRM/MRM experiments. By definition of the method, all transitions for a precursor are measured in a PRM experiment, and therefore PRM experiments do not require selection of fragments prior to acquisition. Rather, transitions are curated after acquisition to improve the precision of the measurement. In comparison, SRM/MRM experiments target only the transitions preprogrammed for acquisition. Selection of optimal transitions is critical for quantitative experiments, as poorly designed assays will suffer unreliable, inaccurate, or nonspecific quantitation (Ludwig et al., 2011).

It is common to choose y-type ion fragments, due to high ion abundance compared to the alternative, b-type ion fragments (Holstein , Sherwood). Similar to peptide selection, transition selection must be evaluated for chemical considerations, namely transition MS signal response and transition selectivity. Transition signal response may be assessed empirically through preliminary MS experiments to evaluate potential transitions in the appropriate experimental sample matrix and under the experimental instrument conditions. The mass spectrometrist must manually confirm that the transitions are high-responding and free of interference, and remove any transitions that do not meet those criteria. Alternatively, predictive algorithms for thermodynamic peptide fragmentation (Zhang, 2004, 2005) may provide computationally-assisted transition selection, and computational tools have been designed to aid in SRM method development (Rost et al., 2012), though none have been integrated with Skyline yet.

Current standard practice (Carr et al., 2014) monitors three or more transitions per peptide to make a reliable quantitation. However, if the transition has been evaluated as high-responding and free of interference, it is possible to perform quantitative analysis on one transition, using the other monitored transitions for confirming the identity of the peptide precursor.

Retention time determination for scheduled MS experiments.

Most quantitative mass spectrometry experiments hyphenate reversed-phase high performance liquid chromatography (RP HPLC) to separate and simplify complex proteomic samples. Coupling LC to MS adds a time dimension to the data, as peptides elute off the solid-state column at a par-

ticular time in the chromatographic gradient. As with other modes of reversed phase chromatography, LC-MS peptide RT is dependent on several experimental factors, such as the physiochemical properties of the target peptide itself; background matrix of the sample; column-specific details including stationary phase material, bed length, and temperature; and the chromatography details including gradient percentage and delivery speed (Moseley et al., 1991). In the case of liquid chromatography coupled SRM/MRM and PRM experiments (LC-SRM/MRM, LC-PRM), the number of peptide precursor-fragment transitions to be measured may exceed the speed at which the instrument can measure them and still maintain a cycle time appropriate for quantification (2-3 seconds per cycle maximum). In these cases, *scheduling* methods enable measurements of tens to hundreds of individual peptides, by allowing only a subset of the targeted peptides to be measured in any given cycle. The acquisition schedule for these methods includes precursor m/z, transition m/z, and the RT, or time window during which the precursor peptide elutes off the LC column.

Skyline's ecosystem incorporates several complementary tools to predict peptide RT. The first, SSRCalc (Krokhin, 2006; Spicer et al., 2007), is based on calculated hydrophobicity, as determined from the peptide amino acid sequence, to predict a peptide RT. This approach is particularly useful when empirical RT is unknown for a peptide. Alternatively, when peptide RT has been previously observed, a standard set of reference peptides can be used to calibrate RT prediction for any number of target peptides of interest on new columns or chromatography methods. In this approach, termed indexed retention time (iRT) (Escher et al., 2012), the reference peptides act as anchor points across a range of hydrophobicities, allowing the HPLC run-time to be calibrated and the assay-specific peptides to be aligned to the observed iRT reference peptide anchors. The iRT method is particularly useful in interlaboratory and large-scale experiments, projects which typically necessitate use of multiple LC systems and columns. For these projects, the iRT workflow integrated into Skyline provides a simple method to transfer chromatography empirical knowledge from one system to another, or to easily transition to a new column when the previous is replaced.

After predicting peptide RT through either method, or simply by using prior measurements that have already been imported, Skyline can export an acquisition table including all relevant information for a scheduled LC-SRM/MRM or LC-PRM method, including start and end times for

peptide elution. The priority for these experiments is to capture the entirety of the chromatograhic peak as the peptide elutes from the column, but with as narrow a window as possible. The mass spectrometer is limited in the number of peptide precursors it can measure at any given time, as dictated by the speed of the instrument (duty cycle), and the number of transitions to measure at that time, as dictated by predicted RT and the width of the scheduling time window. In order to assay as many peptides as possible, it is necessary to adjust the scheduling windows to reflect the instrument's speed and the number of transitions eluting at each time point. Skyline facilitates this adjustment with a visualization option in the retention time pane that displays the number of transitions eluting over the chromatographic gradient under several potential scheduling window widths.

Instrument parameter optimization.

Determining the optimal set of MS instrument parameters for a targeted experiment is necessary in order to create an effective assay. One parameter of particular importance to targeted experiments is collision energy (CE). Optimized CE increases fragment ion intensity, which confers stronger, more reliable signal response (Sherwood et al., 2009). Computational estimation of optimal CE based on precursor *m/z* and a simple linear equation (Equation 1) is useful for both triple quadrupole (Picotti et al., 2009) and quadrupole time-of-flight instruments (Griffin et al., 1991; Prakash et al., 2009). An automated pipeline for optimizing CE specifically for quantitative assays is integrated in Skyline to achieve maximum fragment ion intensity (MacLean et al., 2010a) and therefore strongest, most reliable signal response for the peptides in the assay. Recent versions have added the ability to store optimized parameter values in a library for future re-use and easier sharing.

MS/MS spectral library creation.

Although not strictly required for assay development, inclusion of spectral libraries in quantitative proteomics aids in downstream data processing. In spectral library searching, spectra acquired

by tandem mass spectrometry (MS/MS) are compared with previously identified reference spectra (Craig et al., 2005). The benefits to library searching as opposed to database searching, in which spectra are compared with spectra predicted from amino acid sequences (Eng et al., 1994), is a more accurate comparison of fragment ion intensities and a more efficient spectra search.

The Skyline ecosystem includes a suite of software tools, Bibliospec (Frewen et al., 2006), for creating and searching MS/MS peptide spectrum libraries. The Bibliospec 2.0 software package is composed of two informatics tools: BlibBuild and BlibFilter. All Skyline installations include these tools, and Skyline itself provides user interface for creating spectral libraries. The first step in building a spectral library is creating a full redundant library of peptide MS/MS spectra matched with known peptide identifications, which is performed computationally by BlibBuild and written to sqlite3 database file. To obtain peptide identifications for this step, an assortment of available database search programs are supported by BiblioSpec 2.0 (Table 2). Second, BlibFilter refines the redundant library to choose just one representative spectrum for each peptide, preserving the original retention times of the redundant spectra, and then writes a new non-redundant sqlite3 database containing this information. BlibFilter choses the one representative spectrum by measuring the similarity between all pairs of redundant spectra for a given peptide, and selecting the spectrum with the highest average similarity score.

The Skyline GUI also supports MS/MS spectral library creation. To do so, it takes the best scoring PSM from a variety of supported search engines (Table 2) as a reference spectrum, picking the most intense in the event of a tie. In addition to creation of spectral libraries, Skyline supports several sources of reference libraries, including Peptide Atlas (Desiere, 2006), the National Institute of Standards and Technology (NIST), and the Global Proteome Machine (GPM) (Craig et al., 2004). Most Skyline users will choose to use their spectral libraries, once created, for targeted method creation and data extraction.

Final method export and refinement.

Once a Skyline document is built with the settings and optimizations described above, the final assay is exported either as a native method for triple quadrupole instruments or as scheduled isolation lists for certain Q-TOF and the Thermo Q-Exactive instruments. After acquiring mass spectrometry data, the acquisition files are imported into the Skyline document for method refinement such as peptide and transition validation. The cycle of export, acquisition, and refinement is repeated until the assay is considered effective, at which point final acquisition and quantitative analysis begins.

1.2.2 Data independent acquisition mass spectrometry

DIA workflows attempt to acquire the same precursor isolation windows repeatedly across the elution profile of a peptide. Unlike parallel reaction monitoring (PRM) (Peterson et al., 2012), which targets specific peptides, DIA targets wide, evenly spaced precursor isolation windows that are tiled across a *m/z* range of interest. Originally Venable et al (Venable et al., 2004) envisioned DIA as a method to detect peptides without requiring a precursor signal. Consequently, the original methods focused on acquiring MS/MS with narrow precursor windows (10 m/z) at 3 Hz with an approximate 35 second cycle time. This approach allowed them to generally acquire at least one MS/MS spectrum within the elution profile of each peak but quantitation could only be performed on precursor ions in interspersed MS spectra. Modern Orbitrap and ToF instruments can collect MS/MS at 10 Hz or faster and allow for PRM-like quantitation using MS/MS spectra.

Despite the wide appeal of DIA for quantitative proteomics, one drawback is that many approaches commonly require generating comprehensive DDA-based spectrum libraries (Deutsch et al., 2018) before interpreting any DIA data (Egertson et al., 2015; Ludwig et al., 2018; Reubsaet et al., 2019). While this approach produces high-performance libraries with instrument-specific fragmentation and retention times, it does so at the expense of time, sample, and significant effort offline fractionating that sample (Bruderer et al., 2017). These trade-offs are additionally expensive when considering the fact that DDA-based spectrum libraries are typically considered to not be reusable for other experiments. However, several approaches have been developed to detect pep-

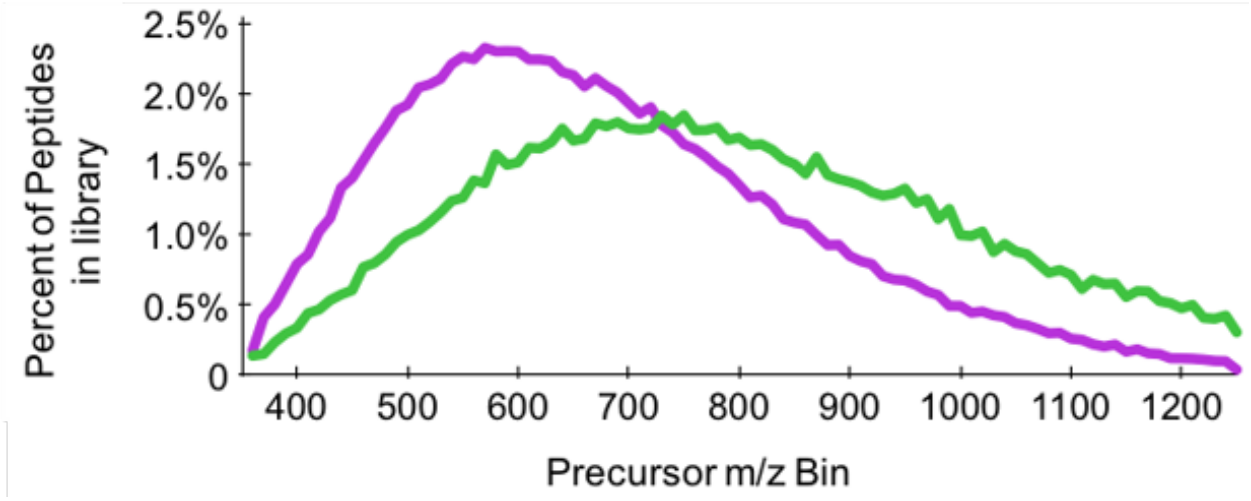


Figure 1.4: Appropriate peptide precursor ranges are dependent on the sample type. Using a fixed isolation width (10 m/z), the distribution of peptide precursor *m/z* in a whole-human proteome repository (the Pan-Human library, purple) and in a human phosphopeptide library (green) differ significantly, indicating the importance of setting precursor isolation width ranges appropriately for each experiment.

tides directly from DIA experiments (Tsou et al., 2015; Ting et al., 2017), and here we demonstrate how to use them to successfully acquire and analyze DIA experiments without spectrum libraries using a DIA-only workflow (Searle et al., 2018). In this work, we focus on analyzing DIA data with open-source software using Proteowizard (Chambers et al., 2012), Skyline (MacLean et al., 2010b) and EncyclopeDIA (Searle et al., 2018). It should be noted that commercial software for detecting peptides without spectrum libraries (e.g. Scaffold DIA or Spectronaut Pulsar) can also be used.

Balancing compromises within DIA methods is critical for successful experiments. The intent of DIA is to measure as much of the proteome as possible, while still maintaining quantitative rigor. These compromises manifest as a result of three competing goals, the desire to: a) maximize the total precursor range of targeted peptides, b) maximize the number of points measured across every chromatographic peak, and c) minimize the number of peptides simultaneously contributing signal in each window.

Precursor range of targeted peptides.

Although it is impractical to measure every possible tryptic peptide in a proteome, the total precursor range can be optimized to target the majority of peptides. For example, while some peptides produce more intense signals below 400 *m/z* or above 900 *m/z*, we find that 93% of peptides in the Pan-Human library (Rosenberger et al., 2014) can be observed within that 400-900 *m/z* range (Figure 1.4). Assuming a fixed cycle time, narrowing our focus to this range allows us to collect narrower precursor isolation windows, and thus lowering signal interference for any given peptide. However, this same range only encapsulates 77% of the phosphopeptides in a human phosphopeptide library (Lawrence et al., 2016) of similar scope, suggesting that it is important to match the precursor range to the proteome of interest if specific peptides or modifications are targeted.

Number of points across chromatographic peaks.

Since quantitative measurements are made at the fragment-level, it is imperative that there are sufficient fragment ion scans for each precursor isolation window to appropriately represent the peptide peak shape. Following the conventional practice of quantitative mass spectrometry, most DIA tools use trapezoidal rule-based integration to measure peak area. While generally robust, significant measurement errors can be caused simply by undersampling across the shape of the peak (Figure 1.5a). Based on an model sampling at fixed intervals across Gaussian distributions (Figure 1.5b), we estimate that restricting measurements to a minimum of nine points sufficiently limits bias caused by trapezoidal integration to below an average of 1% (Figure 1.5c). In general we recommend attempting to achieve a minimum average of 10 points to describe a peak to ensure that faster eluting peptides at the beginning and end of the chromatographic gradient are adequately measured.

Several data acquisition variables can be adjusted to achieve this requirement: total precursor isolation range, scan rate, and peptide elution width. Average peak elution widths are typically dependent on the liquid chromatography setup, and can be determined by looking at past runs (DIA or DDA) using a fixed gradient. The necessary cycle time is the ratio of the average peak

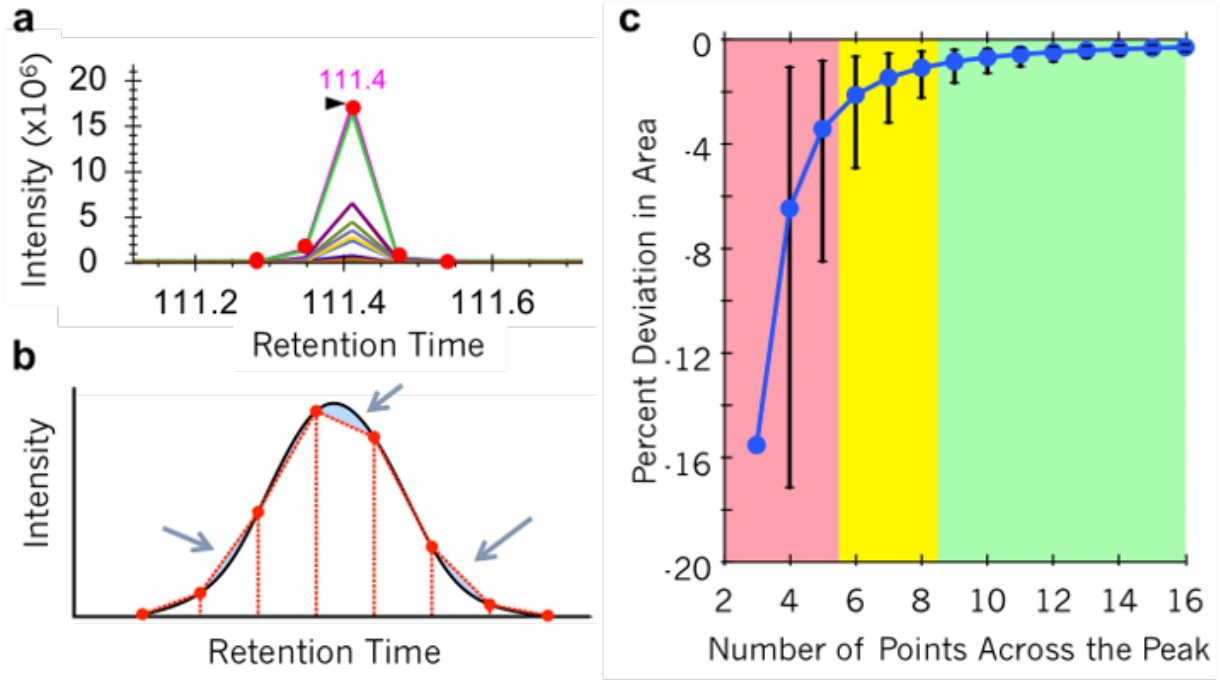


Figure 1.5: Peptide quantification by integrated peak area is sensitive to the number of points sampled across the peak. (a) The trapezoidal quantitation can produce poor measurements with only five points (or fewer). (b) Error (shaded in blue) in trapezoidal quantitation (red dashed lines) typically cancel ut when measuring a Gaussian peak (black solid line) with eight to nine points across the peak. (c) The average percent deviation with 95% error bars in actual/calculated area at different number of points across a Gaussian peak.

width and the minimum number of points needed to describe a peak (typically 10):

$$cycle\,time\,[sec] = \frac{average\,peak\,width\,[sec]}{minimum\,points\,across\,the\,peak} \tag{1.1}$$

The optimal scan rate is typically instrument specific. Combined with the estimated cycle time, it is possible to determine the relationship between total precursor range and the fixed precursor isolation width (or average width if using variable width windows):

$$precursor\,isolation\,width\,[m/z] = \frac{total\,precursor\,range\,[m/z]}{cycle\,time\,[sec] \times scan\,spee\,[hz]} \tag{1.2}$$

This calculation assumes equal transmission of ions across the entire precursor isolation window, and no ions outside that window. It should be noted that no Q1 quadrupole produces a true square-wave transmission, and that some researchers increase the precursor isolation width to add small margins on either side that account for loss of sensitivity at each window edge. We find that this is typically not necessary for instruments that employ a hyperbolic segmented Q1 quadrupole (e.g. for Thermo instruments, the QE+, QE-HF, QE-HFX, and Fusion Lumos), but may help with other Q1 designs.

Cost of interfering peptides.

At first it might appear logical to increase the total precursor range to be as large as possible to measure the most number of peptides. While we previously observed that most tryptic peptides could be detected within the total precursor range of 400-900 m/z, there are some peptides for which the most intense charged ion falls outside that range, and others that are rarely (if ever) observed in that range. However, as the total precursor range increases, the precursor isolation width also increases, and with that the number of interfering peptides. We find that at some point, interference caused by wider precursor isolation widths outweighs any benefit gained from sampling more peptides.

To demonstrate this, we analyzed the same tryptic peptide sample generated from HeLa lysate using five different acquisition schemes that were scaled to keep cycle time constant (Figure 1.6). We found that while the Pan-Human library (Rosenberger et al., 2014) contains peptides from

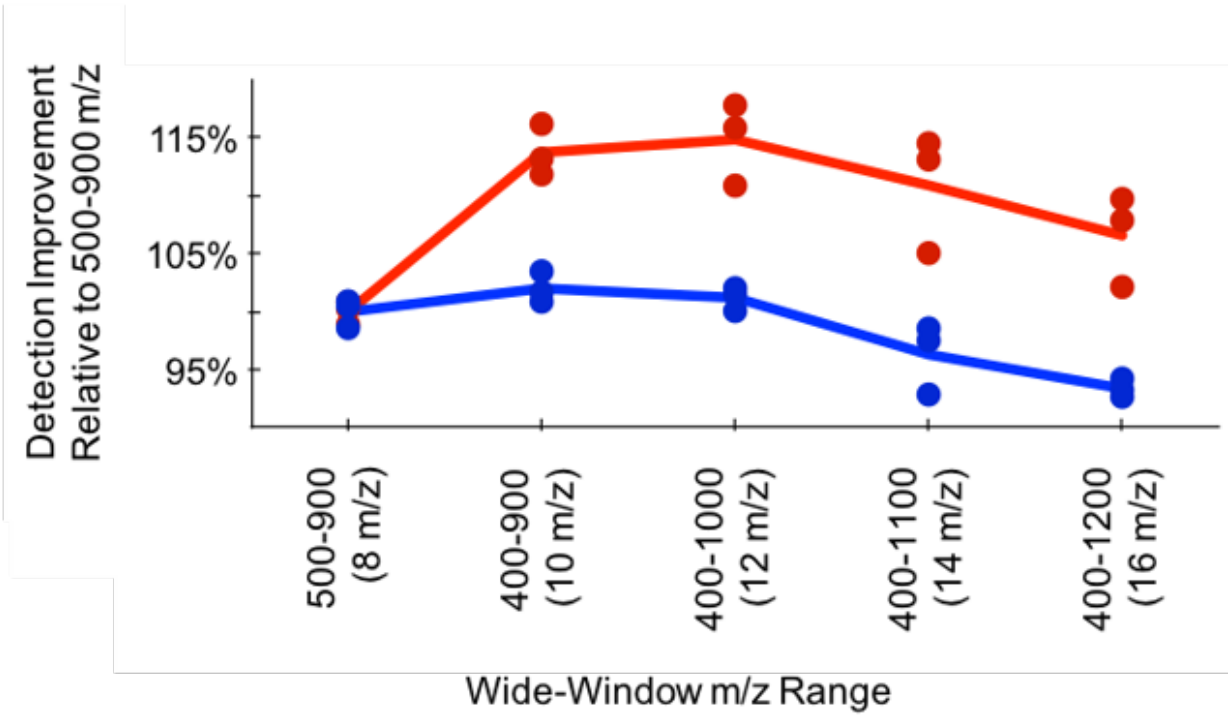


Figure 1.6: Optimal precursor range for surveying the unmodified human proteome trends similarly regardless of library size. While the percentage of new HeLa peptide detections increases with wider total precursor ranges, the number of detections in individual wide-window experiments is highest for both Pan-Human library searches (blue) and sample-specific library searches (red) at 400-1,000 *m/z* due to the tradeoff of increased overall *m/z* range with decreased window selectivity for any given window.

400-1200 m/z, increasing the total precursor range from 500-900 *m/z* did not actually provide any meaningful increase in sensitivity when searching it, and indeed we found that the number of detected peptides dropped significantly beyond 400-1000 m/z. However, when searching a sample-specific library, where retention times and fragmentation patterns are tuned specifically for the instrumentation and chromatographic setup used in this specific experiment, we found that we detected more peptides at wider windows, presumably due to increasing the precision for matching peptide signatures.

We note that some selectivity can be gained back by using variable-width windows (Zhang et al., 2015), which adjust the window-width based on the expected number of precursors in those windows. Overlapping precursor isolation windows (Amodei et al., 2019) or using a similar strategy that employs a scanning quadrupole (Moseley et al., 2018) can also improve selectivity after computationally deconvolving the overlapping isolation regions. Overlapping-window deconvolution is now a built-in feature in the freely available, open-source software tool, Proteowizard (Chambers et al., 2012). As such, we typically recommend the 400-1000 *m/z* precursor range with overlapping windows using Orbitrap mass spectrometers as a good tradeoff between breadth and selectivity for most unenriched, tryptically digested proteomics experiments.

An alternative strategy to gain selectivity is to use gas-phase fractionation (GPF) (Spahr et al., 2001) coupled with DIA. GPF involves no additional sample preparation; instead, GPF injects the same sample multiple times, each time focusing the precursor isolation range on a different fraction of the overall precursor range of interest. This approach, which we discuss in further detail later, has been shown to significantly improve sensitivity in modern DIA workflows (Ting et al., 2017), but comes at the expense of requiring additional injections and sample.

Leveraging Gas-Phase Fractionated DIA to generate chromatogram libraries.

In order to maximize the number of peptide and protein detections in DIA experiments, it is common to search the DIA data against fractionated DDA spectral libraries. The reasoning is that whole proteomes are too complex to trigger MS/MS scans on all possible peptide ions in a con-

ventional DDA method, and so to reveal as much spectral information as possible, researchers first fractionate the sample and perform DDA on each fraction. However, relying on DDA-based spectral libraries to make peptide detections from DIA data assumes that DDA MS/MS spectra are reasonable representations of DIA MS/MS spectra. For several reasons, this is fundamentally false. First, DDA MS/MS are triggered by the presence of a Top-N intense MS1 ion, while DIA MS/MS are triggered systematically regardless of precursor ion intensities. Second, DDA fragmentation is performed with a charge-state optimized collision energy, while DIA fragmentation uses a fixed specified collision energy. Third, DDA precursor selection and fragmentation should theoretically contain only one precursor species, while DIA fragments a specified range of precursor m/z. Finally, DDA spectral libraries from off-line fractionated samples do not result in MS/MS spectra that reflect possible cofragmentation interferences that would occur in the original unfractionated sample because the matrix has been simplified. Spectral libraries from DDA data also suffer from challenges in mapping fragmentation patterns and retention times across instruments and platforms.

To address these challenges with offline-fractionated DDA-based spectral libraries, we propose a better representation of DIA data is a library itself built from DIA data, particularly DIA data collected of the exact experimental sample on the exact instrument platform. These experiment representative DIA-based libraries are called chromatogram libraries. Similar to a spectral library, a chromatogram library contains information about peptides such as fragmentation pattern and retention time, however the chromatogram library is made from DIA data of the specific sample matrix rather than DDA data, and therefore also includes valuable information such as known interferences within a specific sample matrix. Another important difference between spectral libraries and the chromatogram library is precise, experiment-specific retention time information. While it is easy to calculate *m/z* for precursors and fragments, it is not easy to predict RT from peptide sequence alone. Having indexed retention time (iRT) values is more helpful than no RT information (typically 3-5 minute prediction window), but it is even more valuable to have accurate, empirical RT on the same column, in the same samples (30-60 second *prediction* window).

However, the main goal of a library is to “dive deeper” into the sample, detecting peptides of

low abundance even in a complex matrix. Conventional library-based approaches perform deep-dives by physically simplifying the sample using basic reverse-phase (bRP) liquid chromatography. While bRP fractionation does not chemically affect the peptides themselves, bRP fractionation does affect the spectra and retention time for a given peptide. This is because bRP simplifies the sample matrix, and simplifies it differently for each fraction, meaning that the elution time for a peptide in a bRP fraction does not match its elution time in an unfractionated sample. Also, because the peptides are taken out of the context of the original sample, features in the MS2 spectra will not reflect the same features present in the unfractionated sample.

Rather than simplifying the sample in the liquid phase with bRP, fractionating it in the gas phase produces spectra and retention time that reflect the peptide in the complex matrix. Because the entire original samples is subjected to the same chromatographic gradient before each GPF injection, the peptide elution times is not changed. Additionally, peptides with similar retention times and similar *m/z* properties will be fragmented together, producing the chimera spectra expected in the unfractionated samples.

Preparing the chromatogram library sample requires skimming a small aliquot from a representative set of the experimental samples and pooling them together. By combining samples into a representative pool, any peptides present in all the samples will be equally concentrated in the pool. On the other hand, any peptides present in only a few samples or present in low abundance in a subset of samples will be diluted, which may mean that the peptide isn't detected in the chromatogram library. While this may be, it is important to remember that peptides of very low abundance in a pooled but narrow isolation window GP fractionated sample will not likely be detected in a wide isolation window, unfractionated sample.

Although a single pooled sample is appropriate for the majority of basic research purposes, there are two scenarios where acquiring multiple chromatogram libraries may be best practice. These scenarios involve experiments in which the samples have different matrix complexities, as matrix complexity will affect the retention time of the samples. Specifically, complex experimental designs with 3+ sample groups may benefit from a multiple chromatogram library strategy, in which each sample group is pooled for a sample type-specific chromatogram library.

A second scenario requiring multiple chromatogram libraries involves experiments spanning multiple columns, whether due to planned instrument maintenance or unplanned column changes due to column clogs. It is important to note that a library must be collected on each column used in an experiment. This requirement can be satisfied either by pooling the representative sample with enough volume to acquire chromatogram libraries on the same sample on multiple columns, or the experiment acquisition queue can be designed such that the chromatogram library for each column reflects the sample replicate blocks acquired on that column.

Generating an inclusion list.

Above, we discuss considerations in choosing a precursor range to survey. After a precursor range is decided, the inclusion list can be generated. There are several general inclusion window strategies for DIA which we will refer to as "normal", variable, and overlapping. The goal of all windowing schemes is to transmit as many precursor ions as possible (i.e. lowest IIT) with the shortest cycle time (i.e. most points-across-the-peak). As a more detailed review of various inclusion list approaches can be found elsewhere (Ludwig et al., 2018), here we will focus on general best practices for window width and placement.

In generating an inclusion list, window width and placement should account for precursor mass defect. The precursor mass defect is an observation that, across all possible *m/z* values, there are certain *m/z* values at which no peptide precusors exist (*forbidden zones*) (Egertson et al., 2013) (Figure 1.7). Using the knowledge of forbidden zones and the observation that quadrupole transmission is imperfect at the edges of isolation windows, an inclusion list with windows bordered by forbidden zones maximizes the transmission of precursor ions in the window range. By placing a window edge at one of these forbidden zones where no precursor can possibly exist, the edge effects of quad transmission are less pronounced. Practically, this entails shifting the width and edges of the windows off of the chosen integer value (for example, a 24 *m/z* window from 400 *m/z* to 424 *m/z*) so that the window boundaries – where quadrupole ion transmission suffers – will fall in forbidden zones where no precursor can exist (400.23 *m/z* to 424.44 *m/z*). There are sev-

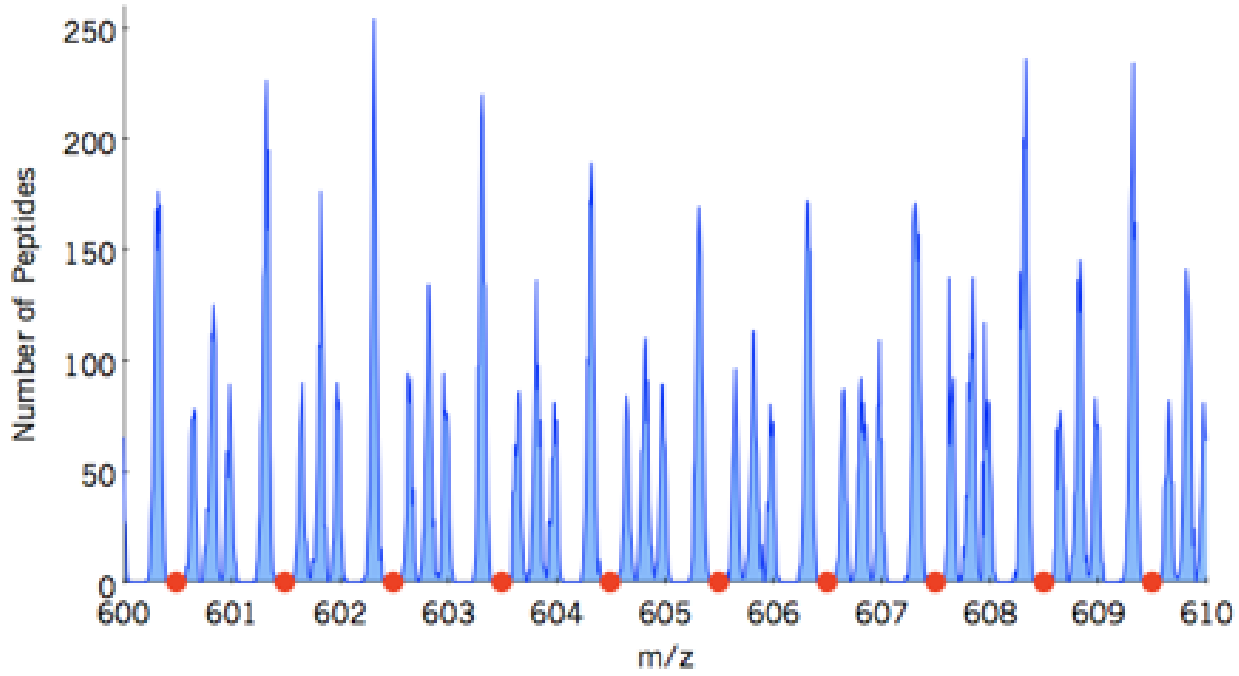


Figure 1.7: Placement of isolation window edges should account for precursor mass defect. The precursor mass defect is visualized for HeLa precursors between 600-610 *m/z* (blue), emphasizing certain *m/z* values that are impossible for a precursor to have (”forbidden zones”) (red points).

eral software options for generating inclusion lists using these principles, including Skyline's Edit Isolation Scheme feature (MacLean et al., 2010b) and EncyclopeDIA's Window Scheme Wizard (Searle et al., 2018).

Sample queueing.

Capturing accurate, on-column retention times is a crucial step in performing DIA without spectral libraries. This is made additionally difficult because gas phase fractions typically do not include the same peptides, so it is challenging to computationally align retention times between gas phase fractions. Therefore, it is important with GPF strategies that retention times across each fraction are as stable as possible.

Retention time is sensitive to changes in column age, so it is best to first condition the column to a sample with similar matrix complexity as the experimental samples. In practice, column conditioning simply means running several samples after equilibrating a new column. As the column ages and is exposed to the sample matrix, retention times should stabilize. Tracking retention time stabilization can be done by choosing a handful of endogenous peptides or spiking a synthetic set of peptides into the conditioning sample and tracking the retention times and peak shapes over each injection.

Because the pooled sample is an empirical average of the unpooled, single-shot quantitative samples, running at least half of the single-shot quantitative samples before running the pooled library sample gives the most stable retention times. It's recommended that the chromatogram library pool should be run in the middle of the experimental sample queue, sandwiched on either side by the unpooled, single-shot quantitative samples (Figure 1.8).

1.3 Informatics for quantitative mass spectrometry proteomics.

The first step in many data analysis workflows is to convert the native, vendor-specific file formats into a portable file format like mzXML (Pedrioli et al., 2004) or mzML. The end result of most workflows is a quantitative matrix of calculated peak areas, or area under the curve (AUC), for each

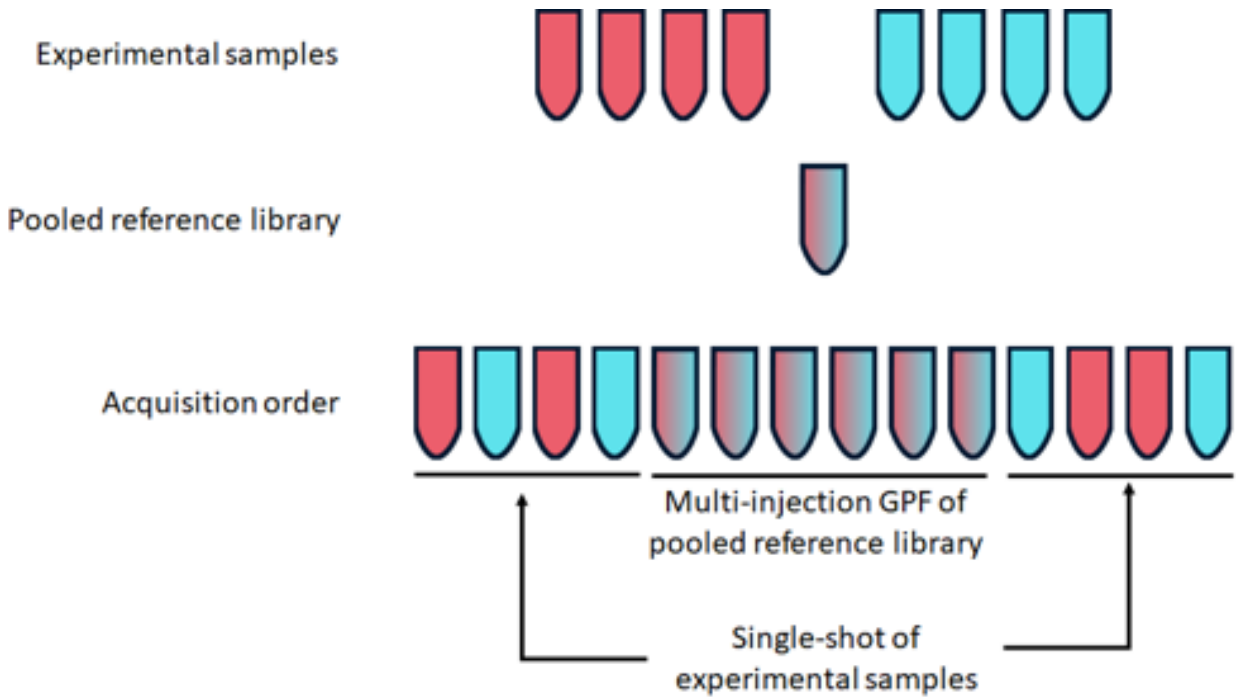


Figure 1.8: Example of an experimental sample queue for the chromatogram library approach. The pooled consensus library sample should be sandwiched between the wide-window, single-shot experimental samples in order to most accurately capture retention time.

peptide ion (modified peptide plus charge state) for each sample in the experiment. Many freely available informatics tools struggle with community adoption, due to issues with limited end user design, and lack a complete pipeline spanning method development through data analysis for an experiment.

Properties such as easy access, large dataset management, integration with other commonly used tools, intuitive data visualization, timely issue resolution, documentation, support, as well as facilitated sharing of data files and the methods used to collect them (Codrea et al., 2007) are important aspects that influence software adoption. With these needs in mind, the freely-available and open-source Skyline ecosystem was developed with a user-friendly interface, comprehensive file compatibility, vendor-neutral data processing, intuitive visualization, and reasonable computational requirements (MacLean et al., 2010b). The original objective of the Skyline project was to create a single informatics tool to generate MS methods and to analyze the data collected for chromatography-based quantitative MS experiments. In addition to these core functions, Skyline now invites the community to share their own informatics tools through an external tool store (Broudy et al., 2014) for software tools that support point-and-click installation and can be run from the Skyline Tools menu. Furthermore, the introduction of additional software to the Skyline ecosystem such as Chorus for sharing raw MS files (`http://chorusproject.org`) and Panorama for sharing Skyline processed experimental results (Sharma et al., 2014), has helped facilitate large-scale MS datasets and inter-laboratory collaborations.

1.3.1 Quantitative data processing for targeted proteomics.

Chromatogram extraction. Mass spectrometry data contains three dimensions: m/z, retention time, and intensity. In the first step of data processing, Skyline extracts the retention time and intensity information for a given *m/z* (Figure 1.9, Step 1). For PRM or DIA experiments, this information is calculated from the measured spectra as extracted-ion chromatograms (XIC), and for SRM/MRM experiments, the measured chromatograms are themselves imported. No file conversion is necessary prior to this step; raw files from the instrument are directly imported. It should be

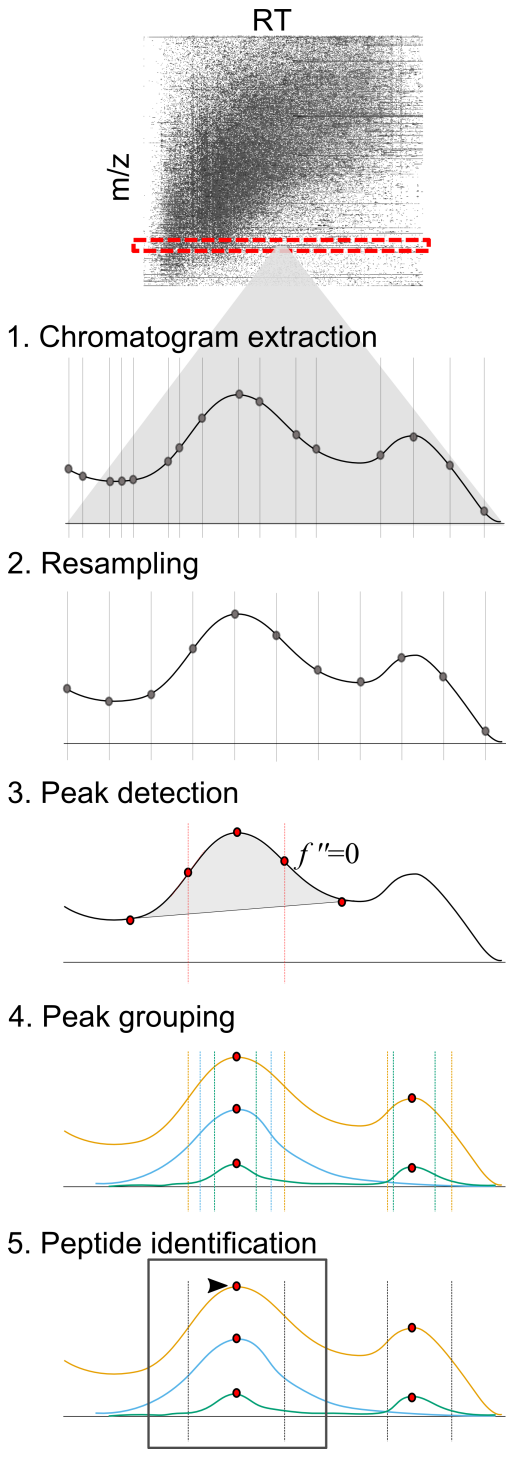


Figure 1.9: Data processing pipeline in Skyline. Skyline derives information from native, vendor-specific file formats or from portable files, producing peak area calculations, and visualizations of the data.

noted, however, that several settings in Skyline affect the chromatogram extraction process, such as retention time window width and parameters for instrument resolving power for profile spectra or mass accuracy for centroided spectra, therefore researchers should be sure that the Skyline document is prepared with the appropriate instrument and experimental details before importing data. These settings can be exported and imported from other Skyline documents, aiding repeatability in data processing and ensuring the proper instrument and experimental details are preserved across laboratory sites and experiments.

Resampling. For all tandem mass spectrometry data acquisition types, the time intervals between MS2 scans are irregular. For example, in an SRM/MRM experiment, the rate of MS2 scans depends on the number of transitions scheduled for collection at a given time and the dwell time for each. For its purposes, Skyline requires all chromatogram time, intensity points for a peptide to be placed on a uniform scale with a consistent interval. Even for DIA, this requires some adjustment of MS1 with MS2 scans and ions for multiple charge states or isotope labeling. To place these points, a linear interpolation of each raw chromatogram is performed. Skyline calculates an interval that captures as much information about the peak as possible (Figure 1.9, Step 3). Intervals placed too wide distort the shape of the peak, while intervals too narrow are costly in storage and processing time. The end product of resampling is an interval width that works best for the dataset, avoiding as much distortion as possible.

Peak detection. The resampled data is then searched for areas that represent peaks. Peak detection is performed by the Chromatogram Retention time Alignment and Warping for Differential Analysis of Data (CRAWDAD) Peaks algorithm (Finney et al., 2008). CRAWDAD finds the maxima and minima by points were the first derivative is equal to zero, then takes the second derivative in the retention time dimension, noting the point at which the second derivative is equal to zero in order to find inflection points. This set of points (local maxima, local minima, and inflection points) define a detected peak. In the absence of spectral library retention time information for peptide spectrum matches (IDs) within the files being analyzed (usually for DDA, PRM or DIA

- with initial processing by tools like DIA-Umpire (Tsou et al., 2015)), Skyline takes only the 20 most intense peaks for each transition from CRAWDAD. When ID times are present, Skyline also includes all CRAWDAD detected peaks containing IDs, or aligned IDs in runs which do not contain any IDs for the target being analyzed. This results in an initial set of raw peak detections for each individual chromatogram with boundaries set at the inflection points and peak areas in interval units.

Peak grouping. Next Skyline creates peak groups for each targeted modified peptide or molecular structure, combining the raw peaks for its chromatograms and grouping them by retention time overlap. Peak grouping is based on elution profile similarity (Figure 1.9, Step 4), with apex RT, start RT, and end RT drawn from the local maxima and inflection points from the previous step. It should be noted that different charge states and isotopes (heavy labeled peptides, medium labeled peptides, endogenous or light peptides) are each considered together. After grouping, the individual peak boundaries are replaced with a single boundary for each entire peak group. This boundary may be adjusted outward from the original 2D inflection point boundary, using Savitzky-Golay smoothing and combined information of all chromatograms contributing to the peak group. Peak statistics are also recalculated to reflect the new agreed-upon boundary values and interval unit areas are multiplied by the number of seconds in the chosen interval to yield an ion count estimate (ions / second * seconds = ions).

Peptide identification. During the peptide identification step, commonly called *peak picking* the top 10 results from peak grouping are evaluated for probability that they represent the peptide. For each of the 10 considered peak groups, a number of peak group features are calculated. These features, derived both from the CRAWDAD calculate statistics and raw chromatogram data, are weighted with particular coefficients, and summed to give a final score to the peak group. The seven scores and corresponding coefficients in Skyline's default peak picking model are log intensity (1.0), coelution count (1.0), identified count (20.0), library intensity correlation (3.0), shape score (4.0), weighted co-elution (-0.05), and retention time delta from prediction (-0.7). The peak group

with the highest score is identified (*picked*) as the peak for that peptide.

Many of these scoring features used in the Skyline default peak picking strategy are similar to those used in the mProphet method (Reiter et al., 2011). Researchers also have the ability to use other peak picking algorithms, such as the mProphet model itself, after initial data import by using a Re-integrate command to generate and apply these models, using decoys and semi-supervised machine learning. As evident from the exceptionally high weight given to the identification count feature, if external tools for peptide identification are used to identify a time of peptide elution within the data, Skyline will give very high priority to finding a peak at that time, using retention time alignment between runs to propagate ID times between runs.

Peak area calculation. In Skyline, the peak area, or area under the curve (AUC), refers to the total integrated area within the peak boundaries, minus the background area (in intensity for seconds of time units - or ion count where intensity is ions per second). Background area is defined as the total integrated area of the minimum of background height and intensity at each point, where background height is the minimum intensity of the two points where the chromatogram crosses the integration boundaries, which is assumed to be the level of intensity contributed not by the transitions themselves but from chemical noise (background) in the measurement. The background area is subtracted from the total integrated area within the peak boundaries to return the final reported peak area. Although Skyline allows display of chromatograms with various smoothing options (2D, 1D, Savitzky-Golay) applied, it uses the interpolated points displayed in the unsmoothed graphs to calculate peak area. Total area values sum the AUC values of individual chromatograms, rather than performing a separate AUC calculation on a summed chromatogram.

1.3.2 The chromatogram library approach for analyzing DIA-MS.

While analyzing DIA experiments to detect peptides from DDA-based libraries (Weisbrod et al., 2012; Röst et al., 2014; Bruderer et al., 2015) is now commonplace, early on DIA datasets were analyzed using ordinary database search engines (Venable et al., 2004) such as Sequest (Eng et al., 1994). Two major classes of tools have emerged to detect peptides directly from DIA experiments

by taking advantage of the repetitive MS/MS measurements in DIA. Spectrum-centric analysis tools for DIA, such as DIA-Umpire (Tsou et al., 2015), attempt to demultiplex several peptide signals from the same MS/MS spectra by time aligning elution peaks for both fragment and precursor ions. Fragment ions that co-vary across retention time are likely to come from the same peptide, and matching precursor ions indicate the potential masses for that peptide. These time-aligned ions are converted into demultiplexed "pseudo" spectra that usually represent a single peptide and can be interpreted with any database searching engine. A powerful benefit for this approach is that it can leverage a wealth of downstream MS/MS software since the pseudo spectra effectively resemble DDA MS/MS. In contrast, peptide-centric analysis (Ting et al., 2015) looks for specific peptides across all spectra in a precursor isolation window. PECAN (Ting et al., 2017) queries DIA data using peptide sequences and their predicted fragmentation models. While the performance of DIA-Umpire and PECAN are comparable under normal conditions, PECAN has been shown to perform better when analyzing GPF-DIA experiments due to DIA-Umpire's reliance on observing precursor ion peaks.

A detailed tutorial walking through each step of data analysis with EncyclopeDIA is provided in full (Appendix A); here we focus on the broader concepts of analyzing DIA data without spectrum libraries.

Verifying raw data quality.

The raw data quality of DIA data can be assessed either qualitatively like shotgun or quantitatively like targeted proteomics before any peptide detection or quantification is performed. The most basic assessment is noting file size across the runs. Quantitative single-shot samples should all be similarly sized files; likewise, each gas phase fraction should be roughly similar in file size. If any file is substantially lower in size, this may indicate a sample or acquisition issue that should be investigated further.

Next, data quality can be assessed by observing the total ion current (TIC) trace over the chromatographic gradient. Ideally, for complex whole proteome samples, the TIC profile makes a

right-angled plateau with no obvious spikes in the gradient.

Finally, before any detection or quantification is performed, the ion inject time (IIT) can be evaluated. Ideally, the IIT across a DIA experiment will not be affected by the precursor isolation windows. In a visualization software such as EncyclopeDIA's RAW File Browser, this should be depicted as the same average IIT across precursor isolation windows. Although IIT should ideally remain unchanged across retention time, it is common to observe that the average IIT across RT forms an upside-down U shape, where the maximum IIT is reached at the beginning and ends of a chromatographic gradient because less ions are eluting at those times. This is because more ions are eluting in the middle of the gradient, requiring less time to collect enough ions to trigger the scan. However at the beginning and ends of the gradient, there are fewer ions, and with less ions to fill the ion trap, the instrument spends more time accumulating ions and more likely triggers once reaching the maximum IIT.

Data file preparation.

Acquisition files from the mass spectrometer are first converted using MSConvert. For data acquired using the overlapping window scheme discussed above, files require a computational demultiplexing step before they can be processed by data analysis softwares. Demultiplexing the files separates the overlapping precursor isolation windows into their effective parts, for example the first few cycles in the overlap scheme described in Table 3 (d,e) are computationally demultiplexed into 12 *m/z* effective windows such that the converted output file contains isolation windows 400-412 m/z, 412-424 m/z, 424-436 m/z, etc. This step requires only one additional parameter flag during the MSconvert step.

Generating a chromatogram library.

A chromatogram library is generated from the gas-phase fractionated, narrow-window acquisitions of the pooled reference sample. While these acquisitions can be analyzed by searching against a spectral library, here we describe searching against a FASTA proteome. Specifically, peptides are

detected in the reference sample acquisitions by searching a ”target” proteome fasta in the context of a ”background” proteome fasta. In global proteome experiments, the target and background proteome will be the same. In enriched or targeted proteome experiments, the target proteome should contain only proteins the researcher is interested in, while the background proteome should be all possible proteins in the sample. For example, an experiment investigating global protein abundances in yeast should use the yeast reference proteome as both the target and background; for an experiment focused on changes in mitochondrial respiration in yeast, the reference proteome would be yeast mitochondrial proteins while the background is the reference yeast proteome.

Using PeCAn or Walnut, the gas phase fractions are searched against the target fasta and an equal number of decoys. The 1% FDR-thresholded detections in each gas phase fraction are combined into a single chromatogram library, which is again 1% FDR controlled. More details about the FDR estimation are described below in ”Searching for peptides”. The final chromatogram library elib file format stores chromatographic data about each detected peptide, including retention times, peak shape, fragment ion intensities, and known interferences, all specific to the reference sample and LC-MS platform.

Searching for peptides.

The EncyclopeDIA software uses chromatogram libraries to detect peptides in wide isolation window samples, leveraging the precise coordinates for m/z, time, and intensity stored in the sample- and instrumentation-specific chromatogram library. To control FDR, Encyclopedia generates a decoy for each target in the chromatogram library. The decoy sequences are generated by shuffling the target sequences and decoy fragmentation is borrowed from the fragmentation pattern observed in the library. EncyclopeDIA first assigns peaks by calculating a weighted dot product of the intensities in the acquired spectrum and library spectrum. Other feature scores are calculated after peaks are assigned, including retention time accuracy. Before scoring retention time accuracy, EncyclopeDIA uses the chromatogram library retention times to align the wide-window detections. All the detections in the wide-window acquisitions are aligned to the chromatogram library, which

is why it is so important to capture accurate retention times when acquiring the narrow-window, gas phase fractionated reference sample.

This set of target and decoy PSMs, without target-decoy competition, is input to Percolator (Käll et al., 2007). The following options are provided to Percolator: "–results-peptides" and "–decoy-results-peptides" tells Percolator to output peptide-level results to files rather than standard output; "–subset-max-train 200000" tells it to train on at most 200,00 examples, and "–post-processing-mix-max" means that it is not using target-decoy competition but is instead using the mix-max procedure.

Using chromatogram libraries to quantify peptides.

After peptide detection, peptides are quantified by integrating and summing their fragment ion chromatograms. In wide-window DIA data, because a range of precursor ions' fragments are analyzed in the same MS/MS scan, interference in fragment ions is common. When interferences are included in integrated peak areas, the resulting peptide quantification is inaccurate. Selecting fragment ions to use for peptide quantification based on DDA data does not reflect the best, interference-free fragments seen in DIA data. Therefore, for the same logic behind using DIA data and chromatogram libraries for optimal peptide detection, using DIA data and chromatogram libraries is optimal for peptide quantification.

EncyclopeDIA enables peptide quantification from chromatogram libraries, including automated interference detection to select the optimal fragment ions for quantification. It first calculates a global interference score for transitions across all wide-window samples in the experiment and only uses the set of best scoring, interference-free transitions to integrate and sum for peptide quantification. Using this method for automated transition refinement, we see that as more interference-free fragments are required for quantification, the reproducibility of peptide quantification improves. We find that the peak area quantifications for peptides with just one interference-free transition are extremely variable, with a median coefficient of variation (CV) greater than 50% (Figure 1.10). When we increase the number of interference-free transitions to a required three,

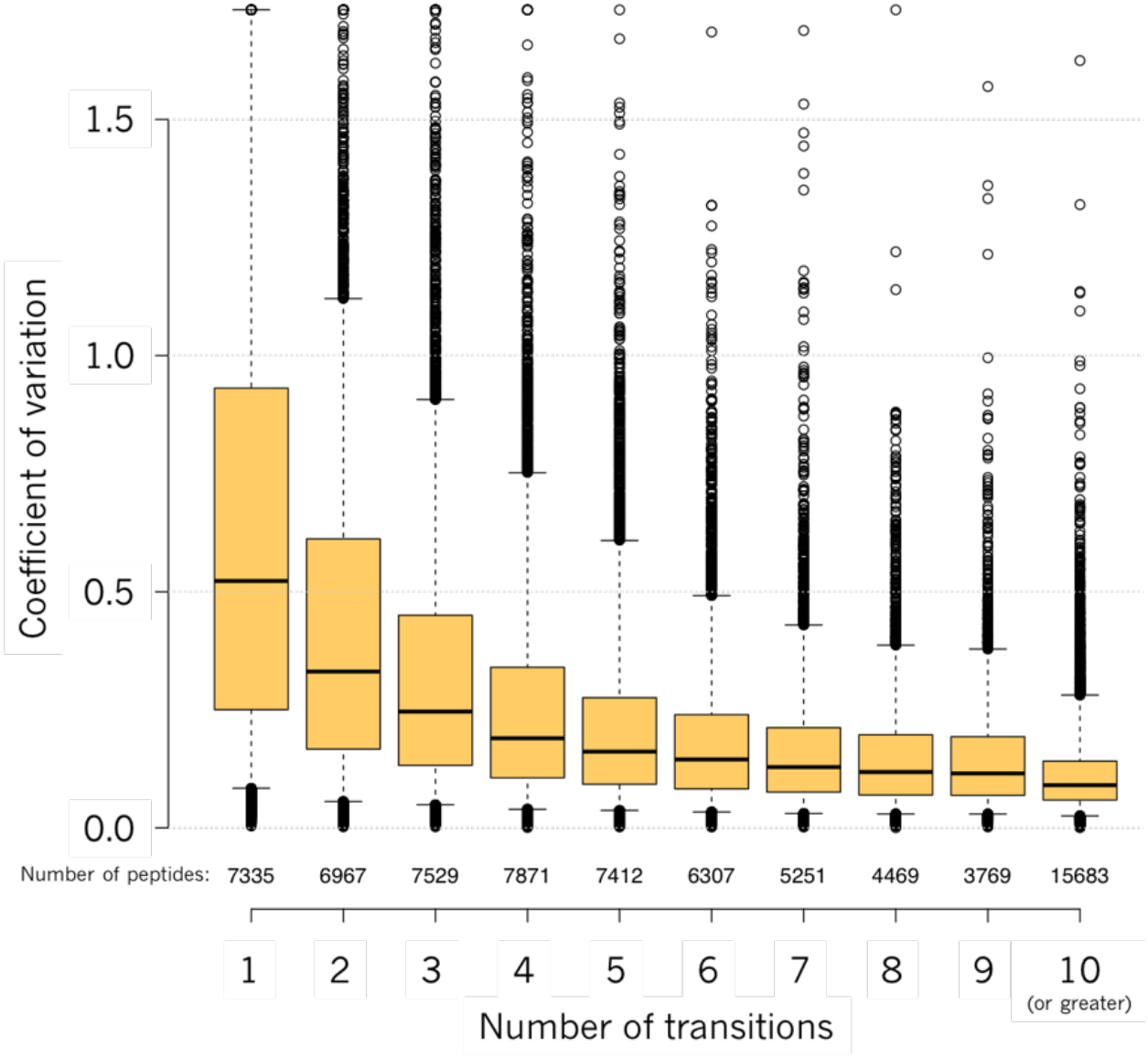


Figure 1.10: Peptide quantification is more reproducible with more transitions. Boxplots showing the coefficient of variation for HeLa peptides with a given N number of transitions after automated transition refinement. The median CV for peptides with three transitions is 25%, while 75% of peptides with five or more transitions have CVs below 20%. Boxes indicate medians and interquartile ranges, and whiskers indicate 5% and 95% values.

the peak area CV improves to 25%. Requiring five or more interference-free transitions improves the peak area CVs to below 20%.

Current limitations of DIA-MS

Many peptides generate the same fragment ions, either due to sequence similarity, or post translational modifications (PTMs). With DDA, while it is rare that peptides share both fragment ions and the same precursor mass, certain circumstances, such as PTM-site localization require additional software and statistics for reliable peptide detection. With DIA this problem is both better and worse. One major advantage of DIA is that because of the regular fragment ion scanning, is possible to retention time-align fragment ions to temporally separate similar peptide species (Searle et al., 2018) (Rosenberger et al. 2017; Meyer et al. 2017). However, unlike DDA, DIA lacks precursor selectivity and similar peptide sequences can fall in the same precursor isolation window: for example triply charged HSASQDGQDTIR (438.87 m/z) HSASQEGQDTIR (443.54 m/z) are both from Human Filaggrin and fall within the same 12 *m/z* precursor isolation window (436.45 *m/z* to 448.45 m/z). Here shared fragment ions may indicate the presence of either one or both of these peptides, and retention time or higher precursor isolation are the only methods to differentiate those outcomes. While the narrow windows in GPF-DIA can somewhat mitigate the effect of shared fragment ions, care must be taken when reporting peptides with similar sequences to avoid double counting the same fragment ion evidence.

1.3.3 Statistical analyses for quantitative proteomics.

Whether targeted approaches or data independent approaches are used to generate the quantitative matrix of peptide peak areas, the subsequent statistical analyses generally fall into several categories: differential abundance testing to find significantly changing peptides or proteins, biological inference and visualizations, and finally data sharing for quantitative proteomics.

Differential abundance testing.

The external tool MSstats (Choi et al., 2014) considers these data properties to calculate the relative quantification of proteins and peptides. MSstats begins with data processing and visualization of the identified and quantified spectral peaks. It then performs statistical modeling and inference using linear mixed models, customized to the method of sample generation and MS acquisition. Finally, researchers can specify a particular statistical power for their experiment, and MSstats determines the minimal number of replicates required to achieve that statistical power by considering the dataset as a pilot experiment.

Other external tools are designed for use with specific acquisition methods. For DDA analyses, an MS1 filtering approach through the external tool MS1Probe (Schilling et al., 2012) enables high throughput statistical quantification of peptide analytes. The external tool QuaSAR (Mani et al., 2012) produces figures of merit (limit of detection, LOD; limit of quantitation, LOQ) for statistical characterization of stable isotope dilution MRM-MS assays (SID-MRM-MS) generated with heavy labeled stable-isotope peptide standards. Within the QuaSAR external tool, AuDIT (Abbatiello et al., 2009) performs automated filtering of transition validation, improving sensitivity and specificity for peptide quantitation by SID-MRM-MS. For label-free quantitative DIA analyses, Skyline exported custom reports can be used to optimize fragment selection and detect interferences using the nonoutlier fragment ion (NOFI) ranking algorithm (Bilbao et al., 2015b).

In addition to the tools described above, Skyline also enables the export of results for analysis in other software suites. The MPPReport tool, for example, creates a results file designed for import into Agilent's Mass Profiler Professional multivariate statistics software package. Researchers can create their own custom reports with a wide range of values to view, edit, and export. Exported custom reports enable researchers to perform their own statistical analyses in Excel, R, Matlab, Java, C++, and other languages, and formats of custom reports can be saved as templates to share and re-use in future analyses.

Informatics for biological inference of quantitative MS proteomics data.

The ultimate goal of many MS proteomics experiments is deriving biological information. Towards this end, researchers have developed several tools to facilitate the visualization and biological importance of peptide and protein measurements. The external tool Protter (Omasits et al., 2013) combines known annotations of protein structure and function with experimental MS data to give researchers an interactive visualization of protein topology. Protter is especially powerful for visualization of membrane protein topology.

1.3.4 Data sharing in the quantitative proteomics community

Skyline, being designed for the mass spectrometry proteomics community, is ideal for interlaboratory collaborations and experimental results comparisons in a vendor-neutral manner. With these types of collaborations in mind, the Skyline ecosystem grew to include storage and sharing applications.

Panorama and CHORUS projects for raw and Skyline file storage and sharing.

Panorama (Sharma et al., 2014), a web-based application for storing, sharing, analyzing and reusing targeted Skyline assays, allows laboratories to communicate the details for replicating or reproducing targeted Skyline experiments. To this end, during the development of Panorama, data integrity, security, and scalability were stressed. Storing Skyline documents in Panorama does not confer any loss of information and data can be made public or kept private at the discretion of the researcher.

It is possible to automate entire informatics pipelines, from acquisition to Panorama publishing, using the command-line version of Skyline, called SkylineRunner. An exemplary case of informatics automation is AutoQC, a completely automated pipeline designed to monitor system suitability in bottom-up proteomics (Bereman et al., 2016). As a mass spectrometer runs, AutoQC imports quality control acquisitions into Skyline, extracts multiple identification-free metrics, and uploads the data to a Panorama Skyline document repository. Users can view system suitability metrics in

the web-based interface, including Levey-Jennings and Pareto plots.

In addition to the Panorama module, the CHORUS platform was developed to provide storage, analysis, and sharing function for raw mass spectrometry files with a simple user interface. When raw data is placed into CHORUS, it is uploaded to the Amazon Web Services (AWS) cloud and translated into a distributed data structure. By utilizing AWS cloud computing and the unique distributed file format, accessing DIA data remotely from CHORUS is faster than from the local hard drive. When researchers wish to request data from the cloud, Skyline requests the extracted ion chromatograms, CHORUS generates the chromatograms, and then returns a Skyline cache. In addition to this scalable data access and remote extraction of chromatograms, CHORUS also provides a browser-based vendor-neutral spectrum and chromatogram viewer, integrated protein database searching and quantitative analysis tools. CHORUS is intended to facilitate community-driven mass spectrometry proteomics, and is therefore a not-for-profit public/private partnership.

1.4 ORGANIZATION OF THIS DISSERTATION

In the following chapters, I describe three projects that aimed to address various challenges of scaling quantitative mass spectrometry experiments.

Chapter 2 describes a mass spectrometry proteomics-specific implementation of single point calibration with a reference material. While single point calibration is common in many other analytical chemistry fields, its application in mass spectrometry proteomics has fallen out of use in favor of expensive, experiment-specific internal standards. In Chapter 2, I present the theoretical and practical use of an external reference material to harmonize quantitative data across experiments, instrument platforms, and laboratories.

Chapter 3 extends the single point calibration approach to multi-point calibration and assessing analytical figures of merit for mass spectrometry proteomics at any scale. I demonstrate a novel experimental and computational framework we call *matched matrix calibration curves*, in which we dilute the reference material with an equally complex matrix to build a dilution series calibration curve. Then, a computational model developed to accommodate this new type of calibration curve

data uses a bilinear model of the data and bootstrapping to calculate a *limit of detection (LOD)* and *limit of quantitation (LOQ)* for every peptide detected in the reference material. In this chapter, I also illustrate the implications of reporting peptide measurements below the LOQ.

Chapter 4 applies the methods development in the previous two chapters to investigate the yeast proteome response to genetic and environmental modulators of replicative lifespan. I show that the protein-level signatures associated with replicative lifespan extension suggest a higher-level response beyond protein abundances. Lastly, I present closing remarks and future directions in Chapter 5.

Chapter 2: CALIBRATION USING A SINGLE-POINT EXTERNAL REFERENCE MATERIAL HARMONIZES QUANTITATIVE MASS SPECTROMETRY PROTEOMICS DATA BETWEEN PLATFORMS AND LABORATORIES

Chapter 2 is adapted with minimal modification from:

Pino LK, Searle BC, Huang EL, Noble WS, Hoofnagle AN, MacCoss MJ. (2018) Calibration using a single-point external reference material harmonizes quantitative mass spectrometry proteomics data between platforms and laboratories. Analytical Chemistry. *90* (21), 13112-13117.

2.1 ABSTRACT

Mass spectrometry (MS) measurements are not inherently calibrated. Researchers use various calibration methods to assign meaning to arbitrary signal intensities and improve precision. Internal calibration (IC) methods use internal standards (IS) such as synthesized or recombinant proteins or peptides to calibrate MS measurements by comparing endogenous analyte signal to the signal from known IS concentrations spiked into the same sample. However, recent work suggests that using IS as IC introduces quantitative biases that affect comparison across studies due to the inability of IS to capture all sources of variation present throughout an MS workflow. Here we describe a single-point external calibration (EC) strategy to calibrate signal intensity measurements to a common reference material, placing MS measurements on the same scale and harmonizing signal intensities between instruments, acquisition methods, and sites. We demonstrate data harmonization between laboratories and methodologies using this generalizable approach.

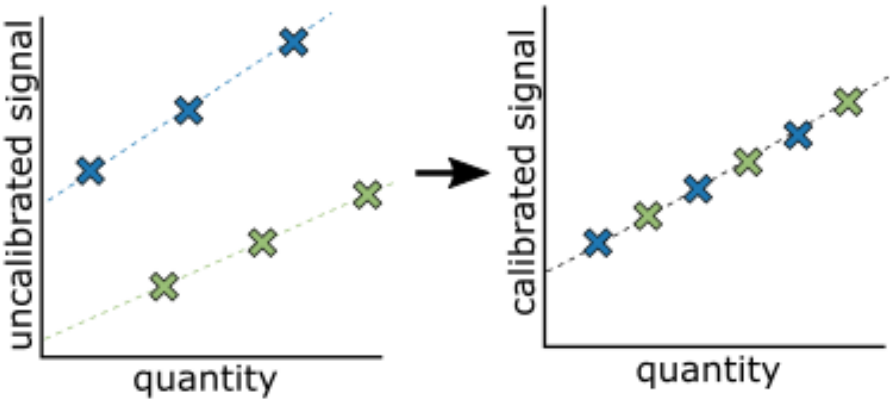


Figure 2.1: Calibration aims to place measurements from two different batches onto the same scale. Two hypothetical sample batches will not have comparable measurements without signal calibration. Whether batches refer to different physical sample preparations, two LC-MS platforms, or two laboratories, the signal must first be placed on the same scale.

2.2 Introduction

To convert signal from any analytical measurement into a more meaningful value, the signal is calibrated by scaling it relative to a reference standard. The goal of calibration is to put all measurements on the same scale, regardless of methodology, operator, instrumentation, or location (Fig 2.1). The bottom-up liquid chromatography-mass spectrometry (LC-MS) field has approached the calibration of protein abundance in two ways: either through internal or external calibration.

Internal standards for MS can be unpaired or paired. Unpaired (also referred to as surrogate) standards typically consist of an exogenous protein or peptide spiked into the experimental sample itself, reviewed in greater detail elsewhere (Domon and Aebersold, 2010), while paired standards typically take the form of isotopically-labeled peptides synthesized with heavy (15N, 13C, 18O) amino acids of the same sequence as the target analyte peptide. Although isotopically-labeled

synthetic peptides can serve as reasonable internal standards, this method suffers from several limitations. First, such peptides are not good calibrants because they do not necessarily reflect the level of the undigested protein (Shuford et al., 2017) and because methods for determining the amount of synthetic peptide in the standard often suffer from poor accuracy and precision (Hoofnagle et al., 2015). Second, this approach requires the enormous cost of synthesizing standards for every target in the experiment. Finally, a recent paper demonstrated lack of harmonized protein quantification when using stable isotope labeled peptides as internal calibrators (Shuford et al., 2017). An alternative paired internal reference approach is *winged peptides*, where the measured peptide is flanked by some series of amino acids, such that the peptide standard is digested out of the wings. However, wings do not accurately capture the digestion conditions of the native protein sequence (Shuford et al., 2017). Beyond winged peptides, researchers also attempt to use intact proteins as calibrants, but the inability to confirm that the standard protein has the same characteristics as the native protein (such as folding, PTMs, etc) prevents this approach from being an ideal calibrant. In addition to protein-level internal standards, a final alternative approach, super-SILAC (Geiger et al., 2010), pools experimental samples into a single master representative sample. The super-SILAC mixes can be used as internal standards, where the same master super-SILAC mix could be spiked into samples across experiments and laboratories as a calibrant. Because a super-SILAC mix includes all proteins in their endogenous states and respective matrices, this approach to signal calibration would address many of the abovementioned limitations. Although the super-SILAC approach is promising, it has not been demonstrated in the years following its proposal. Additionally, the SILAC method is only applicable to cell culturing experiments and is therefore limiting in scope. Because these internal standard approaches all suffer from known limitations, we propose to calibrate protein measurements relative to a common external reference material, which preserves all matrix and digestion properties of the protein measured.

In contrast to calibration by internal standard reference materials, external standard reference materials are separate samples whose acquisitions are interspersed among the experimental sample acquisitions. The external standard reference material is as representative matrix reflective of the experimental matrix; for example, an experiment measuring analytes in human cell lysates would

use a pool of human cell culture, or in plasma would use a pool of plasma, or in yeast would use a bulk culture of yeast. The reference material is prepared alongside experimental samples in each sample processing batch, capturing all the conditions that the experimental samples experience from protein extraction, to digestion kinetics, to instrument variation. Using this type of external calibration approach is common in clinical chemistry, where using a reference material such as normal human plasma for external calibration of patient samples improves precision and harmonization of measurements (Agger et al., 2010; Grant and Hoofnagle, 2014; Hoofnagle et al., 2008; Netzel et al., 2016). Despite these successful implementations of calibration by external references in clinical MS experiments, the broader MS community, despite advances in label-free quantification (Cox et al., 2014), has not yet broadly adopted such an approach.

Here, we describe a generalized approach for calibration by external reference to correct for sample preparation batch variance and instrument-to-instrument variance in not only selected reaction monitoring/multiple reaction monitoring (SRM/MRM) experiments, but any LC-MS experiment. With this external reference approach, the most robust calibrators and reference materials will be stable over time – just as with all other reference materials. We demonstrate this approach in yeast, employing the BY4741 strain as the external reference. The BY4741 strain is particularly useful as a reference material because the copies per cell for many proteins have been estimated (Ghaemmaghami et al., 2003), enabling not only harmonization of the MS signal but also conversion of the signal into a biologically meaningful quantity.

2.3 METHODS

Sample preparation The data regenerated in this work used yeast strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) (Dharmacon) cultured in YEPD to mid log phase, then treated with NaCl to a final concentration of 0.4M NaCl. Cell pellets were harvested and lysed individually with 8M urea buffer solution and bead beating (7 cycles of 4 minutes beating with 1 min rest on ice). Cell lysates were reduced, alkylated, digested for 16 hours, and desalted with a mixed-mode (MCX) method.

Selected reaction monitoring mass spectrometry (SRM-MS) Data were acquired using selected reaction monitoring (SRM) on a Proxeon EasyLC coupled to a Thermo Altis triple quadrupole mass spectrometer. Peptides were separated by reverse phase liquid chromatography using pulled tip columns created from 75 μm inner diameter fused silica capillary (New Objectives, Woburn, MA) in-house using a laser pulling device and packed with 3 μm ReproSil-Pur C18 beads (Dr. Maisch GmbH, Ammerbuch, Germany) to 30 cm. Trap columns were created from 150 μm inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 3 cm. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 80% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using a 90-minute gradient from 5 to 40% B in 25 minutes, 40 to 60% B in 5 minutes, followed by a 15-minute wash and then 15 minutes equilibration back to initial conditions. Total analytical run time was 45 minutes. Thermo RAW files were imported into Skyline (MacLean et al., 2010b) (Skyline-daily version 4.1.1.18151) for processing and Total Area Fragment results were exported using a Custom Report.

Data independent acquisition mass spectrometry (DIA-MS) Data were acquired using data-independent acquisition (DIA) on a Waters NanoAcquity UPLC coupled to a Thermo Q-Exactive HF orbitrap mass spectrometer. Peptides were separated by reverse phase liquid chromatography using pulled tip columns created from 75 μm inner diameter fused silica capillary (New Objectives, Woburn, MA) in-house using a laser pulling device and packed with 3 μm ReproSil-Pur C18 beads (Dr. Maisch GmbH, Ammerbuch, Germany) to 30 cm. Trap columns were created from 150 μum inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 3 cm. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 98% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using a 90-minute separating gradient starting at 5 and increasing to 35% B, followed by a 40-minute wash and equilibration (total 130 minute method). DIA methods followed the chromatogram library workflow, described in greater detail elsewhere (Searle et al., 2018). Briefly, the untreated (reference) samples and osmotic shocked peptide samples were pooled 1:0.33:0.33:0.33

to create a library sample, and a Thermo Q-Exactive HF was configured to acquire six gas phase fractions, each with 4 m/z DIA spectra using an overlapping window pattern from narrow mass ranges. For quantitative samples, the Thermo Q-Exactive HF was configured to acquire 25x 24 m/z DIA spectra using an overlapping window pattern from 388.43 to 1012.70 m/z. The specific windowing schemes for both the chromatogram library construction and quantitative experiments are described in Supplemental Table 1. All DIA spectra were programmed with a normalized collision energy of 27 and an assumed charge state of +2.

Thermo RAW files were converted to .mzML format using the ProteoWizard package (version 3.0.10106), where they were centroided using vendor provided file reading libraries. Converted acquisition files were processed using EncyclopeDIA (version 0.7.0) configured with default settings (10 ppm precursor and fragment tolerances, considering only Y ions, and trypsin digestion was assumed). EncyclopeDIA features were submitted to Percolator (version 3.1) for validation at 1% FDR.

Data Analysis All raw data is publicly available on Panorama Public (`https://panoramaweb.org/singlepointcal.url`, individual file descriptions provided in Supplemental Table 3, ProteomeXchange ID PXD011297) along with Skyline documents for the SRM and DIA experiments performed in this work. Additionally, the processed quantitative data from this work is available in Supplemental Table 2. A Skyline-based tutorial for applying the method described in this work is provided along with open source code in the form of an annotated Python notebook (Appendix B.2).

2.4 Calibration to an external reference sample

The process of calibrating to an external reference material is straightforward (Figure 2.2). The percent change of an experimental sample (E) relative to the reference material (C) is calculated from the peak area (A) of a given peptide as

$$R_{A_E - A_C} = \frac{A_E - A_C}{A_C} \times 100 \tag{2.1}$$

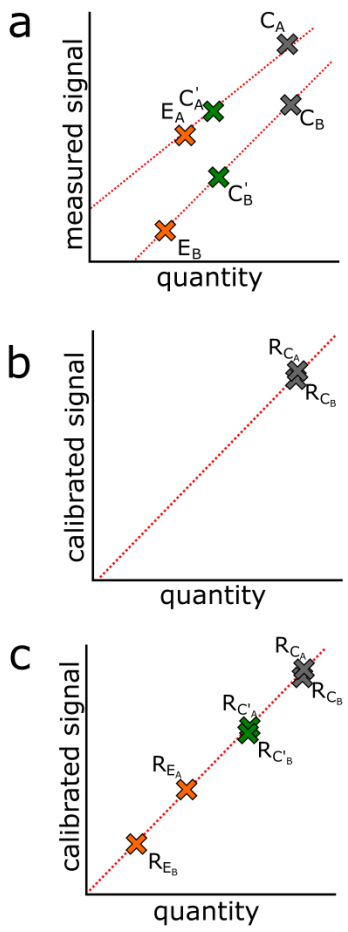


Figure 2.2: Schema for calibration by global reference and by working reference. (a) Signal from samples collected at one site (experimental, E_A; local reference, C'_A; global reference, C_A) are on a different scale compared to those collected at another site (experimental, E_B; local reference, C'_B; global reference, C_B). (b) To harmonize the signals, we set a common scale relative to the global reference material C_A and C_B). While the signal may not be measured on the same absolute scale (see panel a), as long as these reference materials are the same sample they should represent the same quantity. (c) Signals measured for batch A and batch B are calibrated by reporting their signal relative to the reference material signal. In cases necessitating a local working reference material (C'), experimental samples can be calibrated to their respective local working reference, then secondarily calibrated to the global reference.

where the relative change ($R_{A_E-A_C}$) is analogous to the delta notation used in isotopic composition chemistry (Craig, 1961a,b) but expressed as a percentage (%) instead of per mille (‰). To illustrate, consider a peptide with the same abundance (peak area) in the experimental and reference material. In this case, the R value is 0%:

$$R_{A_E-A_C} = \frac{1-1}{1} \times 100 = 0\%$$

In an alternate example, where the peptide is 2x more abundant in the experimental sample than the reference sample, the R value reflects that the abundance of the peptide in the experimental sample is twice (100% change) the abundance of the peptide in the reference material:

$$R_{A_E-A_C} = \frac{2-1}{1} \times 100 = 100\%$$

We note that R values are in the form of a percent relative to the reference but can also be converted to more meaningful units when those values are known in the reference material. Assuming that mass spectrometry response and analyte abundance are linear, if we quantify the analyte through any other method besides mass spectrometry, we can equate the unitage of the new method to the mass spectrometry signal. Converting the relative mass spectrometry signal from a percentage of the reference material to a relevant unitage such as concentrations (e.g. fmol/μl, μg/mL) makes interpretation of the measured values easier across different scientific fields and also enables transfer of the measurements between different lots or batches of reference material. To illustrate this point using our example in yeast, Ghaemmaghami *et al* quantified the molecules-per-cell abundance of nearly all proteins expressed in the yeast strain BY4741 under laboratory standard conditions (YEPD media, 37C incubation, mid-log growth phase) using a TAP-tag and quantitative Western blot approach. In expressing our measured mass spectrometry signals relative to the same reference material (BY4741 grown in laboratory standard conditions), we can associate the R of a reference material signal from a given protein to itself with a multiplier M from Ghaemmaghami *et al*.

$$R_{A_C-A_C} = \frac{1}{1} \times M_C = M_C$$

To demonstrate, consider the example above where the abundance of a peptide in the experimental sample is twice (100% change) the abundance of a peptide in the reference material, and

assume the peptide is unique to its protein of origin. Using the Ghaemmaghami *et al* molecules-per-cell multiplier for that protein in the BY4741 reference material, the equation becomes

$$R_{A_E - A_C} = \frac{2}{1} \times M_C = 2M_C$$

where the 100% increase is now converted to the units held by M. We imagine a reference material may have a multiplier based on any quantitative assay, including enzyme-linked immunosorbent assay (ELISA), GFP-tagged fluorescence, or protein-specific colorimetric assays, if that assay and the MS assay is performed on the same reference material. However, we note that using a multiplier is not required for single-point calibration by an external reference material as we describe here, because the purpose of the calibration is to place experimental measures relative to a reference material which is reported in the R value.

In the above scenarios, the reference material is the same for all experimental samples. However, we can imagine situations where this is not practical, for example, when experimental samples must be batched or where experimental samples are acquired longitudinally. In these situations, we introduce a local working reference material (C'). Here, three steps are required: 1) the peak area of a given peptide measured in an experimental sample in one batch is calibrated to its respective working reference material; 2) the experimental samples in another batch are calibrated to their respective working reference material; and 3) the working reference materials in turn are calibrated to each other through the global master reference (C) (Thienpont et al., 2002). In this scenario, the peak area of a given peptide measured in one experimental sample can be expressed as

$$R_{A_E - A_C} = R_{A_E - A_{C'}} \times R_{A_{C'} - A_C} \times 100 \tag{2.2}$$

in which the peak area for a given peptide in the experimental sample relative to the master sample is a value equivalent to the multiplication of the experimental standard relative to the working standard and the working standard relative to the master standard. To demonstrate, assume an experiment in which the abundance of measurand in the local working reference ($A_{C'}$) is 3-fold greater than that in the global master reference (A_C), e.g. $R_{A_{C'} - A_C} = 3$, and assume that the abundance of measurand in the experimental sample (A_E) is 3-fold greater than that in the local working

reference, e.g. $R_{A_E - A'_C} = 3$. Using these values and the equation above,

$$R_{A_E - A_C} = 3 \times 3 \times 100 = 900\%$$

we find that the experimental measurand, relative to the global master reference, is 900% more abundant.

2.5 Application of single point calibration to harmonize data across MS experiments, platforms, and laboratories.

To demonstrate the proposed single-point calibration approach, we reproduced a portion of an osmotic shock experiment described by Selevsek *et al* (Selevsek et al., 2015), in which cultures of *S. cerevisiae* strain BY4741 were grown unperturbed in YEPD media or shocked with 0.4M NaCl. We evaluated the MS signal of proteins under osmotic shock with and without calibration to the reference material (unperturbed BY4741). First, we compared the effect of calibration on measurements made from identical biological samples prepared on different days by the same operator at the same site using the same instrument and acquisition method (Figure 2.3a). Because these two samples were highly comparable (same operator, same site, same instrument), we should not expect to see dramatically uncorrelated values, and indeed the raw signal shows improved agreement between days without calibration to the reference material. Applying calibration to the reference material in this case does not improve the agreement between the two samples but does assign biologically meaningful units (protein copies per cell) to the measurements (Figure 2.3b).

Based on the precursors detected in these DIA datasets, we developed a targeted SRM method on a Thermo Altis triple quadrupole. We picked targets that spanned a range of signal response on the QEHF. We measured the two sample processing replicates using this scheduled SRM method and observe the same high agreement in both uncalibrated signals and calibrated signals that we see in measuring the two sample processing replicates by DIA ((Figure 2.4a, (Figure 2.4b))). Because these measurements were made in the same laboratory using the same method, the same chromatography column, the same instrument and the samples were acquired consecutively within 11 hours of each other to minimize instrument performance variability, we might expect that the

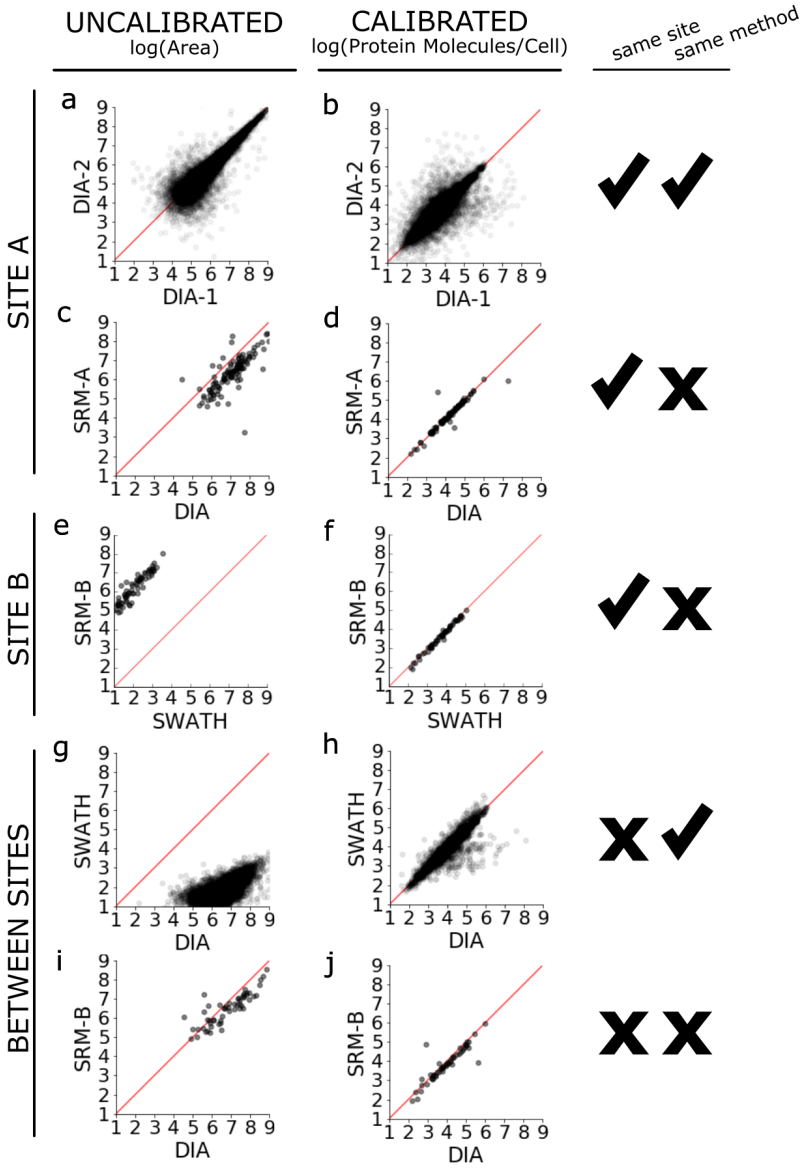


Figure 2.3: External reference calibration harmonizes quantification across MS methods and laboratories. (a,c,e,g,i) Uncalibrated peak areas (log10) of shared precursors from between paired data sets are plotted across sites (Site A, this work; Site B, Selevsek *et al*), and methodologies (DIA/SWATH and SRM). The bias of trends across MS methods reflects systematic differences in data acquisition and instrument platforms, as all data was bioinformatically processed using the same Skyline-based method. (b,d,f,h,j) Application of single-point external reference calibration and the biological unit multiplier (Ghaemmaghami et al., 2003) harmonizes the majority of quantitative values and converts area ratios to meaningful units (protein molecules per cell).

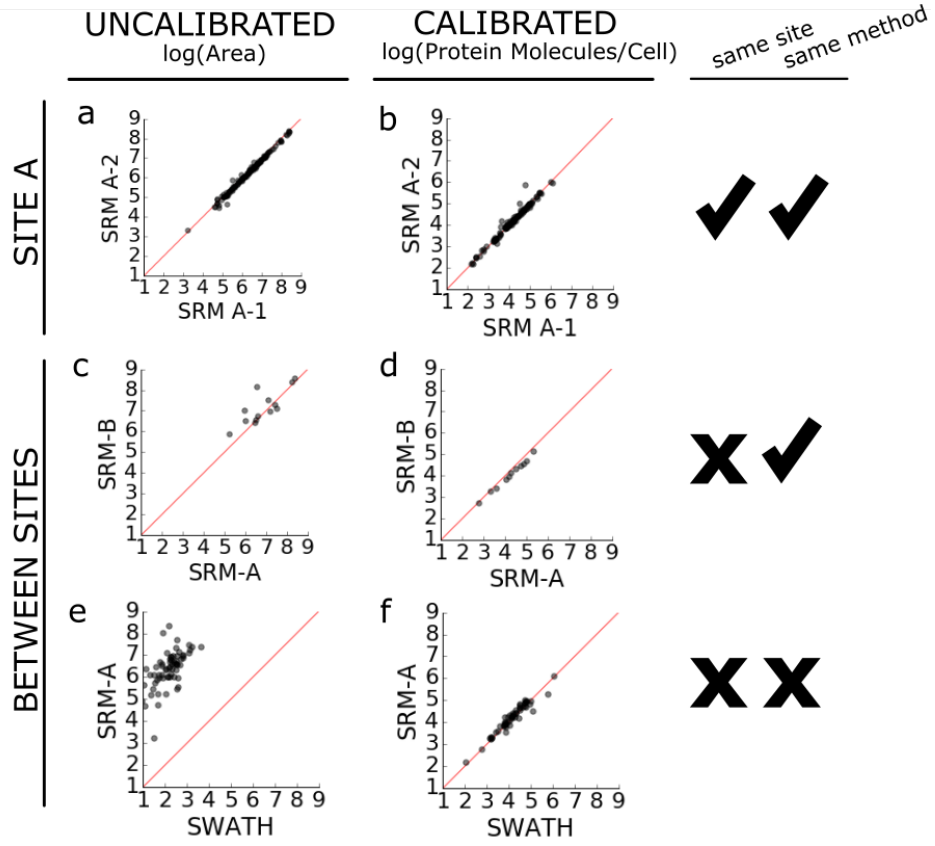


Figure 2.4: Additional examples of external reference calibration harmonizing quantification measurements between methods and laboratory sites. As in Figure 2.3, the inequality of quantifications before calibration reflect differences in the instruments and in laboratories, as all data was bioinformatically processed using the same Skyline-based method. Application of single-point external reference calibration and the biological unit multiplier9 harmonizes the values, bringing them to the line of equality. (a, b) The sample preparation replicates 1 and 2 were measured by SRM at site A (SRM A-1 and SRM A-2) and the values show high correlation due to being measured on the same instrument platform at the same laboratory. (c, d) The SRM values for peptides measured by both site A (SRM-A) and by Selevsek *et al* (site B; SRM-B) are compared, and show greater harmonization after single-point calibration. (e, f) The complementary comparison to Figure 2.3i, Figure 2.3j shows the comparison of the site B SWATH measurements with the site A SRM measurements.

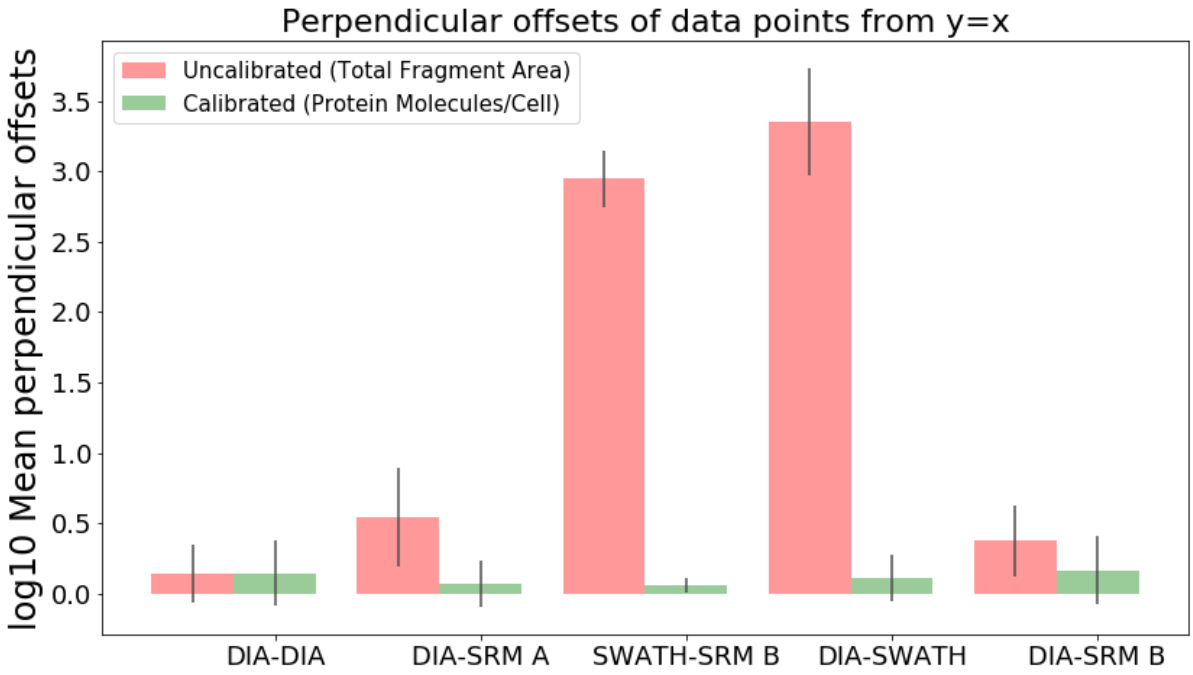


Figure 2.5: Calibration moves signal responses closer to the desired line of equality. For each pair of data sets compared in Figure 2, the perpendicular distance from each of the points to the line of equality was calculated. Points closer to the line of equality have smaller perpendicular offsets, points farther from the line of equality have larger perpendicular offsets. The mean perpendicular offset for uncalibrated points (red) is compared to the mean perpendicular offsets for calibrated points (green) for each of the comparisons, along with the standard deviation (error bars). In most cases, calibrated values have smaller perpendicular offsets than uncalibrated values, indicating the calibration places data points on the same scale.

signals would be highly correlated even without calibration of each batch.

Next, we compared the signal from our DIA method on a Thermo Q Exactive-HF to the signal from the SRM method on a Thermo Altis triple quadrupole. The uncalibrated Orbitrap and triple quadrupole signals roughly follow the same trend, as might be expected from identical samples collected on two LC-MS systems, but they also show a distinct bias away from the $y = x$ line of equality because the raw signal from the Orbitrap is higher than from the triple quadrupole (Figure 2.3c). After calibrating the osmotic shock signal to the reference material, we see the measurements falling tightly along the line of equality, indicating improved harmonization of the

signals (Figure 2.3d). The amount of improvement is quantified by calculating the perpendicular offset, which is the distance of a point to the line. The mean perpendicular offset drops from 0.5 to nearly zero by applying single-point calibration to the reference material (Figure 2.5). There are three notable outliers falling to the right of the $y = x$ line. Closer inspection reveals that these points are from low abundance peptides, where the signal is made irreproducible by interference (data not shown).

We assessed the agreement of quantitative measurements made on different instruments using different acquisition strategies (e.g. the Selevsek *et al* SWATH-MS experiments versus Selevsek *et al* SRM-MS experiments). Of note, although the 100 precursors targeted for SRM-MS were selected from detections and transitions derived from the SWATH-MS data, only 69 precursors had finite calibrated values to compare between the two. We found that although both platforms reported linear trends, the magnitude of the platforms' signals correlated poorly (Figure 2.3e). Applying single-point calibration improved this agreement (Figure 2.3f), harmonizing the difference in signal intensities between the two platforms.

We then compared the agreement of quantitative measurements made from samples prepared by different operators on different instruments using the same acquisition style but different implementations (e.g. the Selevsek *et al* SWATH-MS experiments performed on a Sciex 5600 tripleTOF versus our DIA-MS experiments performed on a Thermo Q Exactive HF). Although we refer to these two methods as SWATH and DIA, the methodological details of the two approaches are very similar (see Experimental Section and Supplemental Table 1). We found that 9932 shared precursors were measured with non-zero values between the two methods. The uncalibrated measurements of these 9932 precursors correlate poorly to each other and do not follow a $y = x$ line of equality (Figure 2.3g). However, applying calibration using the reference materials improves agreement of the measurements from a mean perpendicular offset of 3.3 uncalibrated to an offset of 0.1 offset across sample preparations, operators, and instruments in these two studies ((Figure 2.3h), Figure 2.5). We calculated the Pearson product-moment correlation coefficient between the uncalibrated data and between the calibrated data. The uncalibrated DIA v SWATH correlation coefficient is 0.63, while the calibrated DIA v SWATH correlation coefficient is 0.92. For

context, the correlation coefficient between the uncalibrated DIA-1 v DIA-2 data is 0.92, while the calibrated DIA-1 v DIA-2 correlation coefficient is 0.87. The improved correlation coefficient between DIA v SWATH data suggests that single-point calibration normalizes for experimental variations which may not affect all peptides in a systematic manner.

In addition to the global DIA and SWATH comparison, we compared targeted SRM methods between the two laboratory sites. Because we built our SRM method from our DIA detections, and Site B built their SRM method from their SWATH detections, many of the precursors were not shared. Of the 11 precursors shared in the two SRM methods, we see a dramatic improvement in data harmonization by applying single-point calibration (Figure 2.4c, Figure 2.4d).

Finally, we compared the quantitative measurements made at the different sites using different acquisition strategies on different instruments (e.g. the Selevesek *et al* SRM-MS experiments versus our DIA-MS work experiments). Of the 100 Selevsek *et al* SRM targeted peptides, 40 were also detected and measured by our DIA work. Similar to the poor agreement between different acquisition strategies on different instruments at the same site, we expected to see poor agreement when we looked between different sites (Figure 2.3i). We find that calibrating the measurements improved agreement slightly, and greatly improved the accuracy of the model (Figure 2.3j). The complementary comparison between the Selevsek *et al* targeted SRM experiment and our global DIA experiment was also performed (e.g. Selevsek *et al* global SWATH experiment compared to our SRM experiment) with similar improvements to data harmonization (Figure 2.4e, (Figure 2.4f).

2.6 CONCLUSIONS

In summary, our analyses demonstrate that calibrating to an external reference material improves the harmonization of quantitative LC-MS proteomics data. The single point calibration method, illustrated here in yeast, is generalizable to any proteomics experiment and is universally applicable across acquisition methods. To extrapolate from the various examples we show here, this approach is especially useful for longitudinal studies where samples are collected over extended time frames, consortium projects spanning multiple laboratories, and large scale projects employing multiple in-

struments. We note that while removing internal standards from an experiment increases variance in instrument response, employing an external reference approach does not preclude the use of internal standards. Neither approach is perfect, and in the most ideal metrological scenario, the external reference approach illustrated here could be used together with internal standards. These approaches to ensure accurate and precise measurements come at the experimental cost of an additional sample acquisition (in the case of external reference calibration) or additional transitions monitored by the method (in the case of internal standards).

Following other analytical fields such as isotope ratio mass spectrometry (Craig, 1961b), the proposed external reference material is a homogenous pool of unprocessed material. We propose that one aliquot of this unprocessed material could be measured by another assay, and those measured values used as a multiplier commutable to all the other aliquots. We emphasize that the reference material is a predefined standard appropriate for the experimental system. Here, for yeast, we chose a reference material (pellets of BY4741 strain yeast grown under the same conditions as Ghaemmaghami *et al*) with a useful unitage (protein copies-per-cell) established by TAP- and GFP-tagging methods described in the same work by Ghaemmaghami *et al*.

For all experiments, single-point calibration by external reference improves data harmonization the most when exact physical samples serve as global reference materials, suggesting that laboratories should preserve aliquots of their local working reference materials for future calibration to global reference materials such as the NIST yeast standard or commercially available pooled biofluid products like plasma or CSF (Grant and Hoofnagle, 2014). Even in the absence of an exact global reference, we harmonized LC-MS data by using a thoroughly described reference material and following well described procedures to approximate the previous local reference. Going forward, we propose that LC-MS experimental design should include the selection of an appropriate reference material to support data harmonization.

Chapter 3: MATCHED MATRIX CALIBRATION CURVES FOR ASSESSING ANALYTICAL FIGURES OF MERIT IN QUANTITATIVE PROTEOMICS

This chapter is adapted with minimal modification from:

Pino LK, Searle BC, Yang HY, Hoofnagle AN, Noble WS, MacCoss MJ. (2019) Matrix-matched calibration curves for assessing analytical figures of merit in quantitative proteomics. *bioRxiv* `https://doi.org/10.1101/719179`

3.1 ABSTRACT

Mass spectrometry has become an increasingly powerful tool for the quantification of protein abundances in complex samples. Advances in sample preparation and development of the data independent acquisition (DIA) mass spectrometry approaches have increased the number of peptides and proteins measured between samples. However, methods to assess quantitative figures of merit (i.e. limit of quantification, LOQ) are not easily extended to multiplex assays with hundreds or thousands of analytes. Here we present a general framework for assessing the quantitative accuracy of peptide measurements by mass spectrometry. In a study of the yeast proteome, only 52% of detected proteins (41% of detected peptides) have a peptide that is above the limit of quantification in an unfractionated reference. A similar trend was observed in human cerebrospinal fluid, suggesting that this observation is not sample specific. Our results demonstrate that increasing the number of detected peptides and proteins does not necessarily result in an increase in the number of quantitative peptides or proteins. Our framework provides a method for assessing confident quantitative proteomics figures of merit at scale.

3.2 INTRODUCTION

Mass spectrometry based proteomics has made great progress and is being used to address essential questions in basic biology and of biomedical significance. Of particular interest, the development of data independent acquisition mass spectrometry (DIA-MS) has made it possible to measure tens of thousands of peptides in a protein digest in 1-2 hours of instrument time. The sampling of tandem mass spectra in DIA-MS is unbiased (Ting et al., 2017) and systematic (Collins et al., 2017), in principle making it an appealing compromise between a narrowly focused targeted data acquisition strategy (noa, 2013) and an irregularly sampled discovery method. Although fully targeted proteomics assays often include validation experiments to assess whether the change in measured signal is reflective of the actual change in peptide abundance, proteomics assays measuring thousands of analytes in an unbiased fashion rarely assess which peptide measurements are truly quantitative.

A measurement is quantitative when the change in measured signal reflects a change in the quantity of the analyte (Nic et al., 2009). Specifically in mass spectrometry proteomics, for a method to be considered quantitative the relationship between the measured signal and the peptide quantity must be assessed. This assessment uses a *calibration curve*, where the analyte is diluted systematically to demonstrate that the measured signal is precise and above the *lower limit of quantitation* (LLOQ), the quantity below which a change in signal no longer reflects a change in quantity. Because liquid chromatography-tandem mass spectrometry is subject to matrix effects, calibration curves must be constructed in a relevant sample matrix. For endogenous compounds like peptides that are present in the sample matrix, assessment is frequently performed with *reverse calibration curves*, where a heavy isotope-labeled synthetic version of the analyte is diluted in the sample matrix (Abbatiello et al., 2015; Lynch, 2016). Although a signal measured below the LLOQ may still be used to assess a difference between two conditions, when compared to a signal above the LLOQ, the magnitude of the difference in signal is not reflective of the true difference in analyte quantity. In some papers, this phenomenon has been referred to as ratio compression (Venable et al., 2004). Thus, unless the relationship between the quantity and signal for each

analyte is documented, mass spectrometry measurements should be considered only differential rather than quantitative. In targeted proteomics studies, reverse calibration curves of increasing concentrations of stable isotope-labeled internal standard peptides can be used to approximate the LLOQ and precision of unlabeled peptide responses. However large-scale studies on the order of 1,000's to 10,000's of peptides like most DIA/SWATH-MS experiments do not evaluate peptide response. Calibration curves for up to 30 stable isotope-labeled internal standard peptides have been collected using DIA/SWATH-MS methods (Galitzine et al., 2018), but it is cost-prohibitive to synthesize stable isotope-labeled peptides for the number of targets detected in DIA. In this work, we propose a framework for discriminating between peptides that are only detectable and those which are both detectable and quantitative in a mass spectrometry experiment. We introduce an alternative to reverse calibration curves called *matrix-matched calibration curves*.

3.3 METHODS

3.3.1 Sample preparation and mass spectrometry data acquisition.

Yeast culture and sample preparation. Yeast strains BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) and S288C (MATα) (Dharmacon) were cultured in YEPD and ^{15}N minimal media, respectively, for matched matrix calibration curve experiments. Cultures of 50 mL were grown to mid-log phase, harvested, and lysed individually with 8M urea buffer solution and bead beating (7 cycles of 4 minutes beating with 1 min rest on ice). Cell lysates were reduced with 5mM DTT, alkylated with 15mM IAA, and digested for 16 hours with 1:50 trypsin to protein. The peptide digests were desalted with a mixed-mode (MCX) method, dried down via speedvac overnight, and brought up with synthetic iRT peptide standards (Pierce Peptide Retention Time Calibration Mixture) to 1μg/μl total proteome using calculations from a bicinchoninic acid (BCA) assay (Pierce BCA Protein Assay Kit) performed on the lysate.

Cerebrospinal fluid sample preparation. Pooled human cerebrospinal fluid (CSF) from healthy donors was purchased from Golden West Biologicals. CSF was denatured with 0.2% PPS Silent

Surfactant, reduced with 5mM DTT, alkylated with 15mM IAA, and digested for 16 hours with 1:25 trypsin to protein. The peptide digest were desalted with a mixed-mode (MCX) method, the desalted peptides split into two aliquots, and each aliquot dried down via speedvac overnight. 24 hours prior to MS acquisition, one dried aliquot was resuspended in 0.05μg/μL trypsin in ^{18}O-enriched water (purchased from Cambridge Isotope Laboratories, Inc.) following a standard ^{18}O-labeling protocol (Petritis et al., 2009), the other was resuspended in 0.05μg/μL trypsin in conventional molecular-grade water. The digest incubated overnight then was quenched with 5mM DTT, cooled to room temperature, and acidified with formic acid.

Formalin-Fixed Paraffin-Embedded (FFPE) sample preparation. Pooled human plasma (75μg/μl; Na-Citrate, Cat 7303806, Unit 23-45456A) were diluted with DPBS (Life technologies, 14190-144) to make a plasma dilution series with 13 different concentrations. The 30μl of human plasma or blank samples was well mixed with 80μl homogenized chicken liver in an open-ended syringe (company name and size). Each concentration mixture was quickly mixed with 200μl 20% formalin and followed by 90μl 1% agarose. The syringe was then sealed and left on the bench overnight at room temperature to allow protein-liver mixture form a gel-like structure. Each resulting product was then pushed out from syringe gently and placed into a tissue cassette for standard paraffin embedding procedure.

Six of 10μm-thick tissue slides were obtained from each protein-chicken liver block, and then deparaffined. Proteins on the deparaffinized tissue slides were re-solubilized in 60μl 0.1% RapiGest buffer by undergoing high heat and sonication cycles. Reconstituted protein mixture was reduced, alkylated and digested with 5 μl trypsin overnight. The protein digests were stored in -80 C until the day of analysis.

Liquid chromatography mass spectrometry. Peptides were separated by liquid chromatography before analysis by mass spectrometry, either with a Waters NanoAcquity UPLC for yeast and human CSF DIA experiments or a Thermo easy-nanoLC for FFPE tissue block SRM experiments. On all systems, peptides were separated by reverse phase liquid chromatography using pulled tip

columns created from 75 μm inner diameter fused silica capillary (New Objectives, Woburn, MA) in-house using a laser pulling device and packed with 3 μm ReproSil-Pur C18 beads (Dr. Maisch GmbH, Ammerbuch, Germany) to 30 cm. Trap columns were created from 150 um inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 3 cm.

^{14}N BY4741 yeast proteome separation on Waters NanoAcquity UPLC. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 98% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using a 90-minute gradient from 5 to 35% B, followed by a 40 minute wash and equilibration (35 to 60% B for 10 minutes, 60 to 95% B for 5 minute, 95% B for 5 minutes, 95 to 2% B for 1 minute, 2% B for 19 minute).

^{16}O human CSF proteome separation on Waters NanoAcquity UPLC. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 98% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using 60-minute gradient from 5 to 35% B, followed by a 40 minute wash and equilibration (35 to 60% B for 10 minutes, 60 to 95% B for 5 minute, 95% B for 5 minutes, 95 to 2% B for 1 minute, 2% B for 19 minute).

FPE tissue block proteome separation on Thermo easy-nanoLC. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 98% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using a 30-minute gradient from 0 to 40% B, followed by a 18 minute wash and equilibration (40 to 60% B for 5 minutes, 60% B for 5 minutes, 60 to 100% B for 1 minute, 100% B for 5 minutes, 100 to 0% B for 1 minute, 0% B for 1 minute).

Data independent acquisition mass spectrometry (DIA-MS). Yeast curve data were acquired using data-independent acquisition (DIA) method on a Thermo Q-Exactive HF Orbitrap mass spectrometer. Human CSF curve data were acquired using an equivalent DIA method on a Thermo Lumos mass spectrometer. Both DIA methods followed the chromatogram library workflow, described in greater detail elsewhere (Searle et al., 2018). Briefly, to create the chromatogram library, the mass spectrometer was configured to acquire six gas phase fractions of the undiluted reference proteome for each curve (e.g. ^{14}N BY4741 yeast proteome, ^{16}O human CSF).

Thermo Q-Exactive HF Orbitrap method details. Mass range of 388.43190-1,012.70480 *m/z* was monitored in the yeast experiments. The chromatogram library, gas-phase fractionated "narrow window" Thermo QEHF method details were as follows: 4 *m/z* overlapped windows (effectively 2 *m/z* isolation), 30k resolution, 55 maximum ion inject time, 1e6 AGC. The quantitative, single-shot "wide window" Thermo QEHF method details were as follows: 24 *m/z* overlapped windows (effectively 6 *m/z* isolation), 30k resolution, 55 maximum ion inject time, 1e6 AGC. All DIA spectra were programmed with a normalized collision energy of 27 and an assumed charge state of +2.

Thermo Lumos method details. Mass range of 394.4319 - 1,006.704807 *m/z* was monitored in the CSF experiments. The chromatogram library, gas-phase fractionated narrow window Thermo Lumos method details were as follows: 4 *m/z* overlapped windows (effectively 2 *m/z* isolation), 30k resolution, 60 maximum ion inject time, 4e5 AGC. The quantitative, single-shot wide window Thermo Lumos method details were as follows: 12 *m/z* overlapped windows (effectively 6 *m/z* isolation), 15k resolution, 20 maximum ion inject tiem, 4e5 AGC. All DIA spectra were programmed with a normalized collision energy of 33%.

Thermo RAW files were converted to .mzML format using the ProteoWizard package (version 3.0.10106), where they were peak picked using vendor libraries. Converted acquisition files were processed using EncyclopeDIA (version 0.8.0) configured with default settings (10 ppm precursor, fragment, and library tolerances, considering both B and Y ions, and trypsin digestion was assumed). EncyclopeDIA was configured to use Percolator (version 3.1).

Selected reaction monitoring mass spectrometry (SRM-MS). FFPE tissue block curve data were acquired using a targeted SRM-MS method on a Thermo TSQ Quantiva triple quadrupole mass spectrometer. The target list was developed based on clinical relevance to amyloidosis. Instrument details were as follows: dwell time 2ms, Q1 resolution set to 0.7 FWHM, Q3 resolution set to 0.7 FWHM, CID gas set to 1.5 mTorr.

Data availability. The RAW files, converted MZML files, Encyclopedia elib files, and Skyline documents have been deposited in ProteomeXChange Consortium (**?**) via the Panorama (Sharma et al., 2014) partner repository with the identifiers PXD014815 (ProteomeXchange) and `https://panoramaweb.org/matrix-matched_calcurves.url` (Panorama).

3.3.2 Constructing a serial dilution standard curve using the matched matrix calibration curve approach.

The curves used in this work followed Clinical and Laboratory Standards Institute (CLSI) recommendations (Lynch, 2016). Specifically, the CLSI recommends calibration curves for LC-MS assays are composed of at minimum a blank (a sample containing matrix only) and six to eight calibration standards, with the calibration standards commonly spaced logarithmically across several orders of magnitude. Our yeast dilution series is composed of 13 calibration points and a blank consisting of the matched matrix alone (Figure 3.1, Table 3.1). It is also recommended that calibration curves not be composed of one continuous serial dilution, because this can propagate pipetting errors throughout the curve. We therefore constructed these yeast calibration curves as a set of five serial dilutions, with each of points A, B, C, D, and E mixed individually from reference and matched matrix materials, then subsequent points are dilutions of those original five (F is a dilution of B, G is a dilution of C, H is a dilution of D, I is a dilution of E; then J is a dilution of F, K is a dilution of G, L is a dilution of H, and M is a dilution of I). If pipetting error occurred in one of the dilutions, it would appear as an outlying point in the final calibration curve.

For the cerebrospinal fluid curves, we followed the same fractional dilution scheme as above, but did not include points K, L, and M due to limited availability of the matched matrix material (^{18}O enriched CSF).

For the FFPE tissue block proof of concept, we created concentration points of human plasma by diluting healthy donor pooled plasma into PBS, then mixing an equal volume of each plasma dilution with liver homogenate using an open-end 2ml syringe (Table 3.2). Each concentration point-spiked liver homogenate sample was then formalin fixed and paraffin embedded into individ-

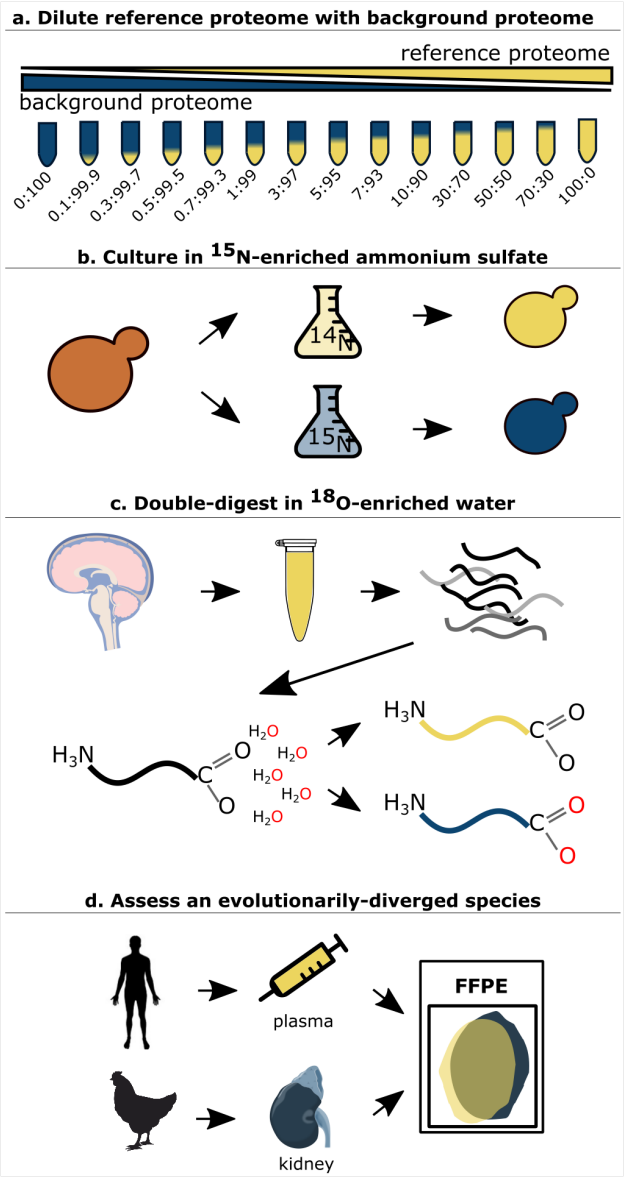


Figure 3.1: Three methods for constructing matched matrix calibration curves from a reference proteome. (a) All matched matrix calibration curves are constructed from a reference proteome and a background proteome. The reference proteome is diluted into a background proteome spanning several orders of magnitude in ratio. (b) Cell lysates can be matrix-matched by culturing the cells in a stable isotope media, such as ^{15}N or SILAC, in order to shift *m/z* of the background matrix. (c) Biofluids such as cerebrospinal fluid or plasma can me matrix-matched by incorporating ^{18}O into peptides via trypsin incubation in ^{18}O-enriched water. (d) Tissues or other biological samples requiring sensitive preparation protocols can be matrix-matched using an y-diverged species, ensuring that the matrix is similarly complex but contains no homologous analytes.

Point	Yeast reference (fractional dilution)	Yeast matched matrix (fractional dilution)
A	1	0
B	0.7	0.3
C	0.5	0.5
D	0.3	0.7
E	0.1	0.9
F	0.07	0.93
G	0.05	0.95
H	0.03	0.97
I	0.01	0.99
J	0.007	0.993
K	0.005	0.995
L	0.003	0.997
M	0.001	0.999
N	0	1

Table 3.1: Dilution series for the yeast matched matrix calibration curves. Using a fractional dilution scheme, the reference material is diluted with the matched matrix to create a dilution series. In this 14-point design, the calibration standards span three orders of magnitude of the reference material and include a blank with only the matched matrix.

Point	Plasma (fractional dilution)	PBS (fractional dilution)
A	1	0
B	0.5	0.5
C	0.25	0.75
E	0.1	0.9
F	0.05	0.95
G	0.025	0.975
I	0.01	0.99
J	0.005	0.995
K	0.0025	0.9975
L	0.00125	0.99875
M	0.001	0.999
O	0	1

Table 3.2: Dilution series for the FFPE tissue block matched matrix calibration curves. Using a fractional dilution scheme, healthy donor plasma (reference) is diluted with PBS to create a dilution series. An equal volume of each calibration point was then mixed with a homogenate of chicken liver and prepared as individual FFPE tissue blocks.

ual tissue blocks. Tissue blocks were scraped and prepared for analysis by mass spectrometry as described above.

3.3.3 A piecewise linear model to fit sparse, label-free LC-MS calibration curves.

We developed a model to fit the data produced by the matched matrix calibration curve method. The model is an extension of the work described previously by Galitizine et al (Galitzine et al., 2018). Below, we briefly summarize the main steps of the model then discuss each step in detail.

Algorithm 1 Model for determining LOD and LOQ from matched matrix calibration curves

Input: x curve points, y measured signals.

1: Fit piecewise regression (parameters b_n, b_s, m_s)
2: Find intersection of piecewise components ($P_x = \frac{b_n - b_s}{m_s}$)
3: Calculate standard deviation of noise segment (σ_{y_n})
4: Calculate LOD ($LOD = \frac{b_n + \sigma_{y_n} - b_s}{m_s}$)
5: Uniformly discretize 100 bins of x_i from the range $LOD < x_{max}$
6: **for** i to N **do**
7: Resample $n = x$ data points from x, y with replacement
8: Fit piecewise regression to the resampled points
9: For each x_i predict y_i using the regression parameters
10: For each x_i calculate $CV_{y_i} = \frac{\sigma_{y_i}}{\mu_{y_i}}$
11: **end for**
12: Calculate LOQ ($LOQ = min(x_i)$ for which $CV_{y_i} \leq 0.2$)

First, the model assumes two segments are present in the calibration curve: a noise segment where the measured signal y_n (reported as intensity, peak area, estimated concentration, etc) does not exceed background noise and a signal segment where the measured signal y_s is within the linear range for the analyte. Formally, we express this model in Equation 3.1 as

$$f(x) = \begin{cases} y_n = b_n & x < LOD \\ y_s = m_s x + b_s & x > LOD \end{cases} \tag{3.1}$$

where x is the experimentally constructed analyte dilution values given by concentration, copies-per-cell, fractional dilution, etc. We use weighted least squares to minimize the function (lmfit package version x.x) using as weights the inverse square root of the curve points, and we constrain the parameters (b_n, b_s, m_s) following (Equation 3.2)

$$\begin{aligned} m_s &\geq 0 \\ b_n &\geq b_s \\ b_n &\geq 0 \end{aligned} \tag{3.2}$$

With these constraints, we enforce that the signal segment must have a positive, nonzero slope, and we enforce that the intersection of the noise and the signal segments must be positive.

To determine the standard deviation associated with the noise segment, we calculate the empirical standard deviation of all y_n values associated with the noise segment. The y_n values are those where the corresponding x_n values are less than the intersection P_x of the noise and linear segments.

$$P_x = \frac{b_n - b_s}{m_s} \tag{3.3}$$

Thus, we compute the empirical standard deviation σ_{y_n} in y_n for all points for which $x \leq P_x$.

Next we determine the figures of merit: limit of detection (LOD) and limit of quantitation (LOQ). We define the limit of detection (LOD) as the x for which the corresponding signal y is one standard deviation (σ) above the noise segment,

$$LOD = \frac{b_n + \sigma_{y_n} - b_s}{m_s} \tag{3.4}$$

The limit of quantitation (LOQ; also referred to as the Lower Limit of the Measuring Interval (LLMI)) is defined by the Clinical and Laboratory Standards Institute (Lynch, 2016) as “the lowest measurand concentration at which all defined performance characteristics of the measurement

procedure are met." The performance characteristics we choose to define are the lowest analyte concentration which (1) is above the LOD and (2) achieves a coefficient of variation (CV) less than a threshold τ selected by a researcher (default is a 20% CV, $\tau = 0.2$). To determine the value x which meets these two criteria, we first uniformly discretize the range of x above the LOD into 100 bins (x_i), for which we will calculate 100 predicted y_i by bootstrapping. Then we calculate the standard deviation and mean in the 100 predicted y_i for each x_i.

For bootstrapping, we resample the entire dataset with replacement N times (default $N = 100$). Each of the N resampled data sets is fit to the piecewise regression model described in 3.1. We use the piecewise regression parameters to calculate the predicted response for a series of curve points spanning the range of curve points in the empirical data. The mean and standard deviation of the bootstrapped y_i values are used to calculate a bootstrapped coefficient of variance (CV_{y_i}) for each of the curve points in the series. Last, the LOQ is calculated as the lowest value in the curve point series above the LOD which passes at or below the CV_{y_i} threshold specified by the researcher (default $CV_{y_i} = 0.2$). The user has the option to set more or less conservative thresholds. For instance, the CV threshold recommended by Clinical and Laboratory Standards Institute guidelines is 10:1 signal:noise which equates to a $CV_{y_i} = 0.1$ threshold (Lynch, 2016).

The code is accessible on Bitbucket (`https://bitbucket.org/lkpino/matrix-matched_calcurves`).

3.4 Results and Discussion

Our goal was to construct calibration curves and determine the LLOQ for every detectable peptide in a given complex protein mixture of interest using one dilution series and without predetermining targets. We propose *matrix-matched calibration curves*, in which a complex protein sample of interest (a *reference material* (Pino et al., 2018)) is diluted with a *matrix-matched material*. A matrix-matched material may be any sample of equivalent biochemical complexity, but should not share any endogenous analytes with the reference material. For example, a matrix-matched material could be a stable-isotope labeled a reference material that preserve the matrix complexity but

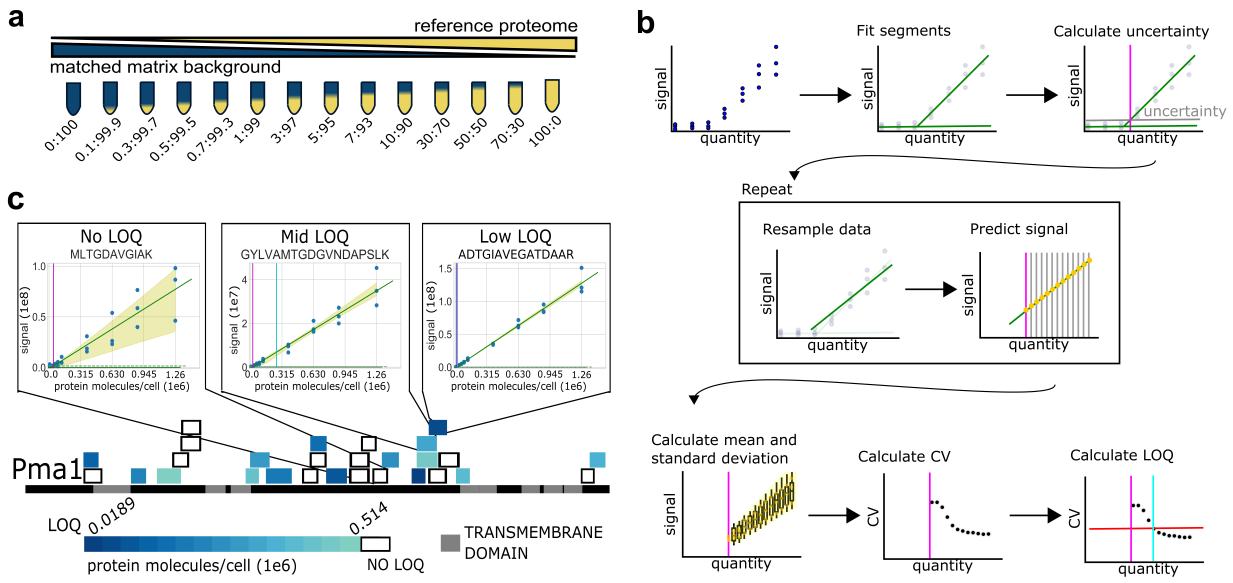


Figure 3.2: Constructing reference material calibration curves using a matched-matrix diluent. (a) A reference material is diluted into a matrix-matched material of similar matrix complexity but with no shared endogenous analytes, for example by stable isotope labeling the matrix or using a diverged species. The curve is made from dilutions spanning several orders of magnitude plus a *blank* with only the matrix-matched proteome. (b) The model for assessing the lower limit of quantification (LLOQ) using the sparse matrix-matched calibration curve data. We assess the LLOQ (cyan line) as the first point that is statistically different from the background (pink line) and has a CV $\leq 20\%$ using bootstrapping (red line). (c) The sequence of plasma membrane ATPase (Pma1) is represented as a black line. The transmembrane domains along the sequence are depicted in grey. Each peptide detected by DIA-MS is represented by a colored box placed along the sequence. The color of the box ranks the peptide LLOQs. Three of the peptide calibration curves are shown above the sequence. Yellow shading indicates two standard deviations above and below the median for the bootstrapped data.

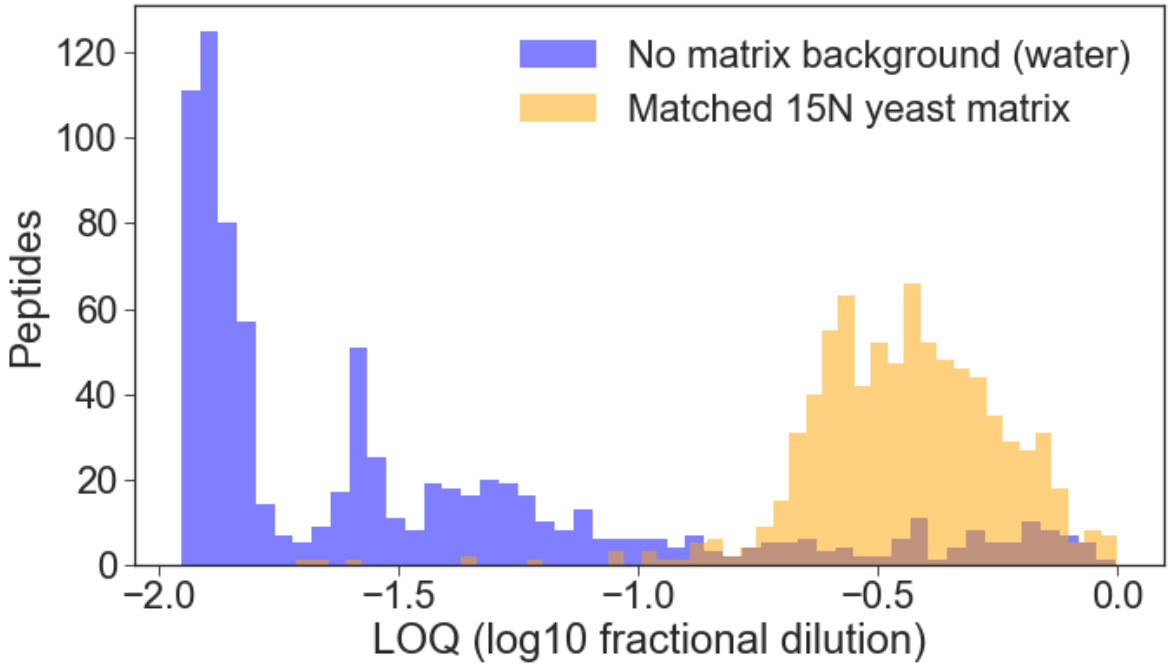


Figure 3.3: Reference materials must be diluted with a similarly complex material to preserve matrix properties. The yeast reference material was diluted in a matched matrix with 15N-shifted yeast (orange) or diluted in water with no matrix replacement (blue). Our curve-fitting model was then run on each data set to calculate the LOQs. The reported peptide LOQs are significantly more sensitive when the matrix complexity is not replaced, showing the importance of retaining matrix properties when building the calibration curve.

shift the peptide masses or using an equivalent biosample from an evolutionarily-diverged species (Figures 3.1, 3.3). Each point in the dilution series has the same total protein concentration, composed of some ratio of the reference and matrix-matched material (Figure 3.2a) spanning several orders of magnitude (Table 3.1). A strength of this approach is that every peptide (or other type of analyte) in the reference material is diluted through the curve, meaning that calibration curves are constructed for all peptides detected in the reference material. To fit calibration curves to this novel data, we developed a computational model (Figure 3.2b) which extends the work described previously by Galitizine *et al.* (Galitzine et al., 2018) to accommodate the sparseness of matrix-matched calibration curve data and to determine the LLOQ for each detected analyte. Briefly, the model first fits a piece-wise linear regression to the noise and the signal segments of the curve data, then bootstraps the observed data, refits the piece-wise regression to the bootstrapped data to predict signal over the range of quantities measured. Finally we calculate the coefficient of variance (CV) of the predicted signal and define the LLOQ as the minimum quantity at which the predicted signal passes a predetermined CV threshold (CV $\leq$ 20% for the results reported here) (see Methods).

We apply the matrix-matched calibration curve framework first in yeast, and find that it highlights the divide between detection and quantification especially at low protein abundances. In particular, highly abundant proteins often contain peptides that are detected at 1% FDR but are not quantifiable because the observed abundance in the reference material is below the LLOQ. Using the highly-abundant yeast proteome plasma membrane ATPase protein (Pma1) as an example, we detect 28 peptides at a 1% FDR threshold across the protein sequence (Figure 3.2c, Appendix C). Of the detected peptides, only half (15 peptides) are deemed quantitative, and the quantitative peptides display a range of LLOQs spanning more than 20x. A peptide with no LLOQ is a less accurate quantitative proxy for Pma1, while a peptide with a low LLOQ is a more accurate quantitative proxy for Pma1 and is more accurate over a wider linear range. The extreme range of peptide responsiveness emphasizes the necessity to carefully select which peptides should act as quantitative proxies for their protein of origin.

The yeast proteome has the advantage of an established reference quantity for each protein, allowing us to contextualize our results. Ghaemmaghami *et al.* affinity-tagged the protein coding

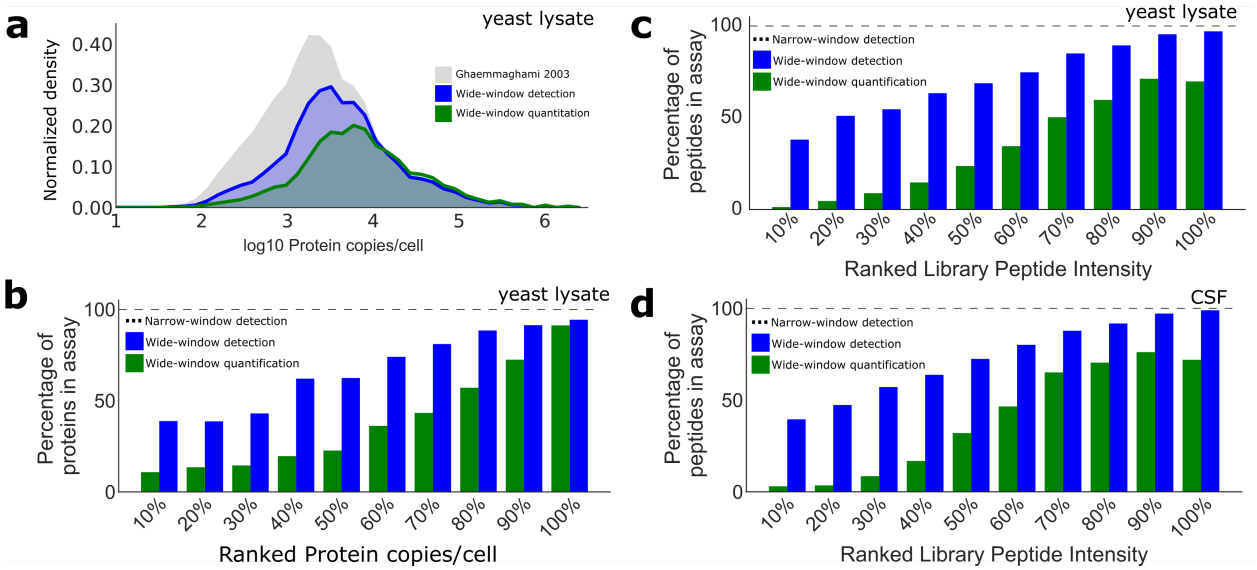


Figure 3.4: There is a difference between the detection of a peptide and the quantification of a peptide. The (a) number and (b) percentage of proteins detected in yeast at different orders of magnitude of abundance. Ghaemmaghami *et al.* comprehensively estimated protein copies per cell in yeast (black, 3,869 proteins) using epitope tagging (Ghaemmaghami et al., 2003). The wide-window DIA using a chromatogram-library approach (Searle et al., 2018) detects, at 1% protein-level FDR, 74% of these proteins (blue, 2,870 proteins). The number of proteins quantifiable by DIA-MS (proteins with at least one peptide with a defined LLOQ) encompasses 52% of the detected proteins, or 39% of the expressed proteins (green, 1,511 proteins). (c) Peptides detected in the yeast lysate narrow-window library are ranked by intensity, and the wide-window detected and quantitative peptides are shown for each decile. (d) Cerebrospinal fluid peptides detected in the narrow-window library (8,698 total peptides, 2,994 protein groups) are ranked by intensity, and the wide-window detected and quantitative (3,183 peptides; 1,303 protein groups) peptides are shown for each decile.

regions in yeast and reported the protein abundances in molecules-per-cell for 4,102 proteins, 3,869 of which could be quantified above 50 molecules/cell (Ghaemmaghami et al., 2003). Using data independent acquisition mass spectrometry (DIA-MS) (Searle et al., 2018), we detected peptides from 2,870 of the proteins they quantified in the reference yeast proteome (Figure 3.4a, b). Using matrix-matched calibration curves to assess the quantitative accuracy of the detected peptides, we found that half of the detected proteins had at least one quantitative peptide (1,511 proteins). The proteins with validated peptides are primarily high quantity proteins, particularly those above 10,000 molecules/cell. As the reported quantity (Ghaemmaghami et al., 2003) decreases, fewer detected proteins have at least one quantitative peptide (Figure 3.4a, b). We compared the peptides determined to be quantitative by matrix-matched calibration curves with the peptides determined to be quantitative by a more conventional synthetic peptide approach (Lawless et al., 2016). Overall, the proposed framework assessed 6x more candidate peptides and defined 4.7x more peptides as quantitative (Figure 3.5), demonstrating the higher throughput of the proposed framework compared to conventional approaches.

The matrix-matched calibration curve approach is generalizable beyond cell culture. To illustrate its flexibility, we adapted the framework to two human samples: cerebrospinal fluid (CSF) and formalin-fixed paraffin embedded (FFPE) tissue. For the CSF reference material, we chose a commercially-available pool of healthy donor CSF (Golden West Biologicals, Inc.) which we prepared following conventional protocols. For the CSF matrix-matched material, we performed a second enzymatic digest in the presence of ^{18}O-enriched water. This reaction preferentially exchanges one or both of the oxygens at the C-terminus of the peptide with ^{18}O, shifting the peptides by 2 or 4 mass units via incorporation of one or two ^{18}O atoms. Following the matrix-matched calibration curve framework, we found that 36% of peptides detected in the CSF reference material library (8,698 peptides; 2,994 protein groups) have a defined LLOQ (3,183 peptides; 1,303 protein groups) (Figure 3.4d). In both the yeast (Figure 3.4c) and CSF (Figure 3.4d) references, the most intense peptides in the reference are more likely to be detected and quantified. We also applied the matrix-matched calibration curve approach to an FFPE sample and acquired the data by another form of mass spectrometry (selected reaction monitoring). To construct the FFPE matrix-matched

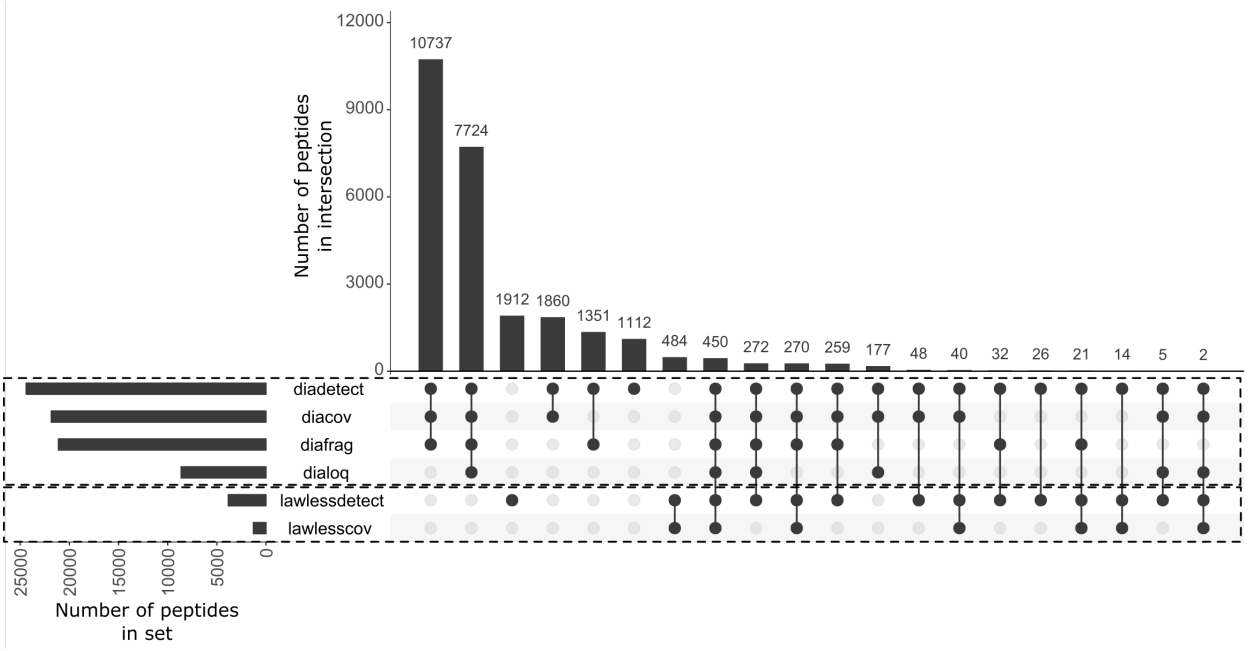


Figure 3.5: Matched matrix calibration curves can assess more candidate targets than conventional approaches without predetermining targets. The UpSet plot compares yeast peptides detected and deemed quantitative by DIA-MS and matched matrix calibration curves (this work, legend "dia*") with the peptides detected and validated by QConCat SRM-MS in Lawless et al 2016 (Lawless et al., 2016). Of all peptides detected in the wide-window DIA-MS at 1% FDR ("diadetect", 24,400 peptides), only 6,117 peptides (25%) display all three desirable quantitative traits ("diacov", peptides with $\leq$20% CV in the undiluted yeast sample; "diafrag", peptides with $\geq$3 interference-free fragment ions; "dialoq", peptides with a defined LOQ). Lawless et al 2016 assessed over 4,000 total peptides ("lawlessdetect", QConCat peptides tested by Lawless et al 2016), of which 1,281 peptides (50%) displayed the desired quantitative properties ("lawlesscov", QConCat peptides tested by Lawless 2016 with $\leq$ 20% CV). Overall, the proposed framework assessed 6x more candidate peptides and defined 4.7x more peptides as quantitative. The quantitative peptides in the proposed approach map to 1,629 proteins; the quantitative peptides by the QConCat approach map to 644 proteins. Both approaches include quantitative peptides for 520 proteins, and the QConCat approach includes peptides for 124 proteins not represented by the proposed approach, while the proposed approach includes 1,109 proteins not represented by the QConCat approach.

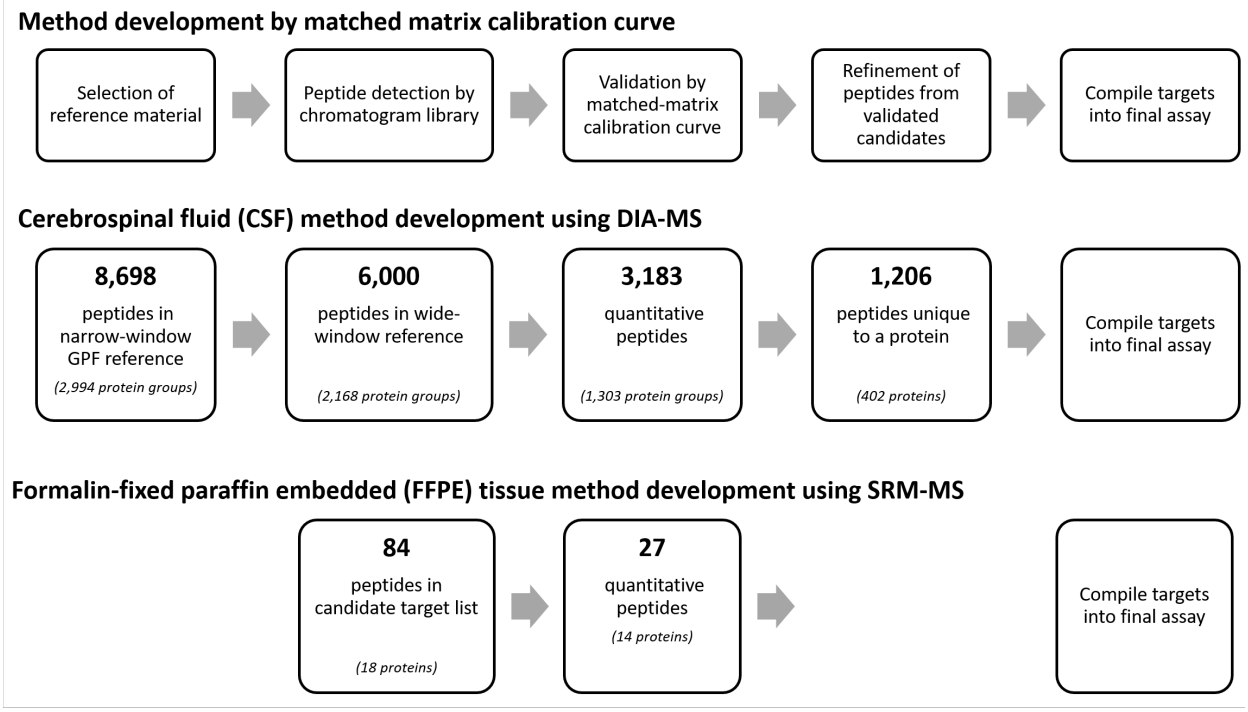


Figure 3.6: Matched matrix calibration curves can be used to rapidly develop targeted methods. Starting with all peptides detected in a gas-phase fractionated reference material as possible candidates, the first refinement step in this protocol discards any candidate target that cannot be detected in an unfractionationated, single-shot acquisition of the reference material. Next, the matched matrix calibration curve framework is used to assess quantitative figures of merit, discarding any candidates whose abundance in the reference material is below the analyte LOQ. For most targeted quantitative proteomics work, targets that are unique to a protein are considered better candidates. Even if experimenters are not starting from DIA data, the matched matrix calibration curve approach can be used to quickly eliminate poor quantitative targets from an assay.

calibration curve, we spiked human plasma into homogenized chicken liver as a reference and used the unspiked homogenized chicken liver for the background proteome. We targeted 84 peptides (18 proteins) and found that 27 of the targeted peptides were quantitative (14 proteins) (Figure 3.6). This demonstrates that the matched-matrix calibration curve approach is generalizable broadly across not only sample types but also mass spectrometry acquisition approaches.

3.5 CONCLUSION

A limitation of the approach is that the maximum possible peptide quantification is limited by the endogenous abundance of the peptide in the reference, which for low abundance peptides results in stunted linear range. Another consequence of the endogenous abundance limitation is that matrix-matched calibration curve data is extremely sparse compared to conventional calibration curves because low abundance reference peptides produce low signal which reduces to zero signal as the reference is diluted. Additionally, while the quantitative peptides reported here may serve as a starting point for future assay development, we emphasize that these LLOQs are specific to these exact conditions. Matrix-matched calibration curves, like all calibration curves, are only reflective of the peptide measured on a given platform. While most quantitative methods report precision, this does not assess whether a change in signal reflects the change in quantity. Therefore, the use of matrix-matched calibration curves should be performed for all proteomics experiments that require an assessment of which peptides reflect the change in quantity those that are just differential.

Chapter 4: MOLECULAR PHENOTYPING OF THE YEAST REPLICATIVE LIFE SPAN RESPONSE TO GENETIC AND ENVIRONMENTAL MODULATORS

4.1 ABSTRACT

Senescence and life span modulation are crucial biological processes for all cellular life, controlled through tightly regulated molecular mechanisms. Many of these mechanisms and their modulation are evolutionary conserved, including calorie restriction (also referred to as dietary restriction) and inhibition of the mTOR signaling pathway. While there are likely multiple cellular processes through which such life span modulators act, the full range of molecular phenotypes that cells display as they age under different contexts and genetic backgrounds remains to be characterized. In my dissertation research, I used state-of-the-art biochemical and proteomic tools to characterize the molecular phenotypes of life span extension in budding yeast (*Saccharomyces cerevisiae*). Yeast are an excellent model system for this application, as replicative life span (RLS) – the number of times a mother cell buds before reaching senescence – is a quantifiable measure of aging that can be collected for yeast on experimental timescales. Further, the yeast genome is well annotated, there are a wealth of established resources for this model system such as the Saccharomyces Genome Database and the Saccharomyces Genome Deletion Collection, and finally, as a eukaryote, they are a much better model for human health compared to prokaryotic systems.

Using our previously described mass spectrometry-based methods for quantifying the yeast proteome at scale, we were capable of creating quantitative proteomic signatures which we hoped to be useful as molecular phenotypes. Here, we applied these frameworks to construct molecular phenotypes of increased RLS in long-lived yeast mutants and to identify key proteins in life span extension through intervention methods.

4.2 INTRODUCTION

Aging in multicellular organisms is the accumulation of molecular, cellular, and finally tissue-specific alterations that ultimately culminates in the final phenotype (Tosato et al., 2007). While aging itself is not considered a disease, certain disease states are associated with aging and better understanding the complexities behind these diseases requires an understanding of the multifaceted process of aging itself. Several distinct but not incompatible schools of thought about aging exist, including the free radical theory of aging (Harman, 2003), the *inflamm-aging* theory of immunosenesence (Franceschi et al., 2000), and the mitochondrial damage theory (Cadenas and Davies, 2000). As there is no single gene or pathway that culminates in the phenotype of aging or longevity, it is necessary to use experimentally-unbiased research techniques to examine aging as a global process. To this end, we propose to use quantitative proteomics to serve as objective measurands, as proteins are the primary functional biomolecules of the cell.

Significance to aging and aging-related disease. Understanding complex aging-associated diseases is not possible without first understanding the molecular mechanisms that govern the basic biology of aging. As multicellular organisms age, distinct biochemical changes to their cells produce the phenotype referred to as aging. In humans, aging leads to increased risk of aging-associated disease such as neurodegeneration, cardiovascular disease, and cancer (Campisi, 2003). Several aging modulators, including specific genotypes and environmental perturbations, are evolutionarily conserved in single-celled eukaryotes and in multicellular organisms such as worms and mice (Smith et al., 2008). This shared functionality suggests that the underlying molecular mechanisms are likely also conserved. We used advanced proteomic techniques to construct the underlying protein-level molecular phenotypes of long-lived mutants of budding yeast (*Saccharomyces cerevisiae*), with a focus on defining how molecular phenotype changes in response to conserved genotype- and dose-dependent lifespan modulators.

Significance to systems and molecular biology. Modulators of lifespan such as calorie restriction (CR), and inhibition of mammalian target of rapamycin (mTOR) signaling have been found

efficacious in mice (McCay et al., 1935), flies (Mair, 2003), worms (Klass, 1977), and yeast (Lin et al., 2002). The mechanism(s) by which these modulators extend lifespan remain poorly understood, but several pathways are generally evolutionarily conserved, including mitochondrial respiration, autophagy, and the signaling pathways of protein kinase A, mTOR, and AMP-activated protein kinase (Wasko and Kaeberlein, 2013). The presence of multiple functional categories in aging phenotypes is to be expected from such a complex biological process, and thus employing a global, unbiased approach to capture all involved pathways will provide valuable insights. By systematically evaluating these phenotypes of aging in the tractable genome of the yeast model system, we may be able to determine key proteins shared between these paths and identify molecular differences.

These experiments are important both because they will elucidate the relationship between different aging modulators in yeast, and because they will improve confidence in quantitative protein signature profiling for future studies of yeast aging. In the subsequent chapter, we will use the molecular phenotypes discovered by this comparative analysis as a starting point for identifying genetic targets indicative of yeast longevity. We place special focus on genes and pathways that are conserved in multicellular organisms such as worms and mice.

Yeast replicative lifespan as a model for cellular aging. Yeast are an excellent model system for basic biology of aging studies because they enable genome-wide screens for molecular mechanisms associated with aging (Steinkraus et al., 2008). Yeast replicative lifespan (RLS) – the number of times a mother cell buds before reaching senescence – is a quantifiable, objective measure of aging that can be collected on experimental timescale (Mortimer and Johnston, 1959). Further, the yeast genome is well annotated, there are a wealth of established resources for this model system such as the Saccharomyces Genome Database (Cherry et al., 2011) and the Saccharomyces Genome Deletion Collection (Giaever et al., 2002), and finally, as a eukaryote, yeast is a more analogous model for human health than prokaryotic systems while retaining the experimental benefits of single-cellular models. Aging proteomics has also been successfully applied to yeast, revealing possible mechanisms of lifespan asymmetry in yeast mother / daughter cells (Yang et al.,

2015) and uncovering a novel mitochondrial unfolded protein response in prohibitin-deficient yeast (Schleit et al., 2013). We hypothesized our quantitative proteomics approach would replicate these previously reported processes, reveal novel protein members of established processes, and discover processes not yet associated with yeast longevity. In this chapter, we construct molecular phenotypes of increased RLS in long-lived yeast perturbations using the quantitative proteomics methods described previously and identify key proteins in life span extension through these interventions.

4.3 METHODS

Sample preparation

Calorie restriction Yeast strain BY4741 (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) (Dharmacon) was grown overnight in YEPD media (2% glucose). The overnight culture was used to inoculate flask cultures of six media conditions of YEP with glucose added to each flask spanning the range of calorie restriction (specifically 2% glucose, 1%, 0.5%, 0.05%, 0.005%, and 3% glycerol), in biological triplicate. Glucose concentrations were measured by enzymatic D-glucose assay (r-biopharm, Germany). Cultures were grown to OD600 0.2 before harvest. Cell pellets were mechanically lysed with beadbeating and 8M urea. Lysates were reduced with DTT, alkylated with iodoacetamide, and digested with trypsin.

Gene deletions Yeast single-gene deletion strains were chosen spanning a range of RLS modulation per McCormick 2015 (McCormick et al., 2015). Special consideration was made to choose strains not documented as petite/slow-growing, and strains whose deleted gene was documented with the most interactions per the SGD (Cherry et al., 2011). The final strains chosen were Δubp8, Δsgf73, Δeos1, Δidh2, Δtor1, Δadp1, and Δade17 (Table 4.1). Growth curves later showed Δeos1 to be slow growing in our hands, so this strain was dropped from further work.

Data independent acquisition mass spectrometry (DIA-MS) Data were acquired using data-independent acquisition (DIA) on a Waters NanoAcquity UPLC coupled to a Thermo Q-Exactive

Gene	Description	# genetic interactors	RLS inc	Slow growing
UBP8	Ubiquitin-specific protease component of the SAGA acetylation complex	462	46%	No
SGF73	Subunit of DUBm module of SAGA and SLIK	685	25%	No
EOS1	Protein involved in N-glycosylation	743	21%	No**
IDH2	Subunit of mitochondrial NAD(+)-dependent isocitrate dehydrogenase	302	13%	No
TOR1	subunit of TORC1, a complex that controls growth in response to nutrients by regulating translation, transcription, ribosome biogenesis, nutrient transport and autophagy	449	13%	No
ADP1	Putative ATP-dependent permease of the ABC transporter family	51	No extension	No
ADE17	Enzyme of 'de novo' purine biosynthesis	110	Shortening	No

Table 4.1: Descriptions of the yeast single-gene deletion strains used in this work. Each gene deleted from yeast listed with the gene description per SGD (Cherry et al., 2011), the RLS increase per McCormick 2015 (McCormick et al., 2015), and whether the strain is documented as slow growing per growth rates reported in Giaever 2002 (Giaever et al., 2002). **The Δeos1 strain, while not documented as slow growing, formed petite colonies in our hands, and was therefore excluded from the study.

HF Orbitrap mass spectrometer. Peptides were separated by reverse phase liquid chromatography using pulled tip columns created from 75 μm inner diameter fused silica capillary (New Objectives, Woburn, MA) in-house using a laser pulling device and packed with 3 μm ReproSil-Pur C18 beads (Dr. Maisch GmbH, Ammerbuch, Germany) to 30 cm. Trap columns were created from 150 μum inner diameter fused silica capillary fritted with Kasil on one end and packed with the same C18 beads to 3 cm. Solvent A was 0.1% formic acid in water (v/v), solvent B was 0.1% formic acid in 98% acetonitrile (v/v). For each injection, approximately 1 μg total protein was loaded and eluted using a 90-minute separating gradient starting at 5 and increasing to 35% B, followed by a 40-minute wash and equilibration (total 130 minute method). DIA methods followed the chromatogram library workflow, described in greater detail elsewhere (Searle et al., 2018). Briefly, the control (reference) sample and calorie restricted yeast peptide samples were pooled to create a library sample, and a Thermo Q-Exactive HF was configured to acquire six gas phase fractions, each with 4 m/z DIA spectra using an overlapping window pattern from narrow mass ranges. For quantitative samples, the Thermo Q-Exactive HF was configured to acquire 25x 24 m/z DIA spectra using an overlapping window pattern from 388.43 to 1012.70 m/z. All DIA spectra were programmed with a normalized collision energy of 27 and an assumed charge state of +2.

Thermo RAW files were converted to .mzML format using the ProteoWizard package (version 3.0.10106), where they were centroided using vendor provided file reading libraries. Converted acquisition files were processed using EncyclopeDIA (version 0.7.0) configured with default settings (10 ppm precursor and fragment tolerances, considering only Y ions, and trypsin digestion was assumed). EncyclopeDIA features were submitted to Percolator (version 3.1) for validation at 1% FDR.

Data analysis Glucose concentration calculations and plots were performed using R. Peptide quantification was performed using EncyclopeDIA-derived peak picking, boundary settings, and transition refinement; peak integration values were calculated using Skyline (MacLean et al., 2010b). Consensus clustering was performed in R using the ConsensusClusterPlus package (Wilkerson and Hayes, 2010). Dose-dependent abundance profiles were tested using the edgeR package

(Storey et al., 2005). Differential testing was performed in R using the MSstats package (Choi et al., 2014) (Appendix D.1).

4.4 Determining key proteins that induce life span extension in yeast upon intervention with calorie restriction.

Several metabolic intervention methods that extend RLS in yeast are also effective in multicellular organisms such as worms and mice (Wasko and Kaeberlein, 2013), suggesting broadly conserved mechanisms. Among these methods are calorie restriction and rapamycin treatment. Calorie restriction in yeast, induced by limiting glucose, extends RLS in genotype- and dose-dependent manners (Schleit et al., 2013); rapamycin, an inhibitor of the mTOR signaling pathway, influences nutrient sensing (Powers, 2006) and can be thought of as a mimetic for calorie restriction.

Our motivation was to find a common protein signature in yeast cultured under CR conditions that corresponds to the calorie restriction mechanism and therefore increases lifespan (Figure 4.1). Calorie restriction is documented in yeast to occur when the glucose concentration is lowered from 2% to between 0.05% and 0.5% (Kaeberlein et al., 2004). Further, under caloric restriction, yeast replicative lifespan is extended 30-40% (Lin et al., 2002). Therefore, our experimental design aimed to assess CR across a range of severities, spanning from abundant glucose through calorie restriction through fully depleted glucose, and including a glycerol control.

We might expect the proteome to respond in several ways. First, if the proteome responded in a dose-dependent manner, we would expect to see the proteins involved in CR to be differentially abundant at the different glucose concentrations. This may appear as a monotonic response (that is, proteins changing in abundance linearly with the amount of glucose available), or this may appear as a multimodal response (for example, with proteins differentially abundant only during CR conditions). To assess whether any proteins were differentially abundant in a dose-dependent manner, we used the EDGE software package to test if any proteins changed significantly compared to the mean abundance profile. However, no statistically significant trends were found in this data (Figure 4.2).

If not a dose-dependent response, we might expect that the yeast proteome would respond to

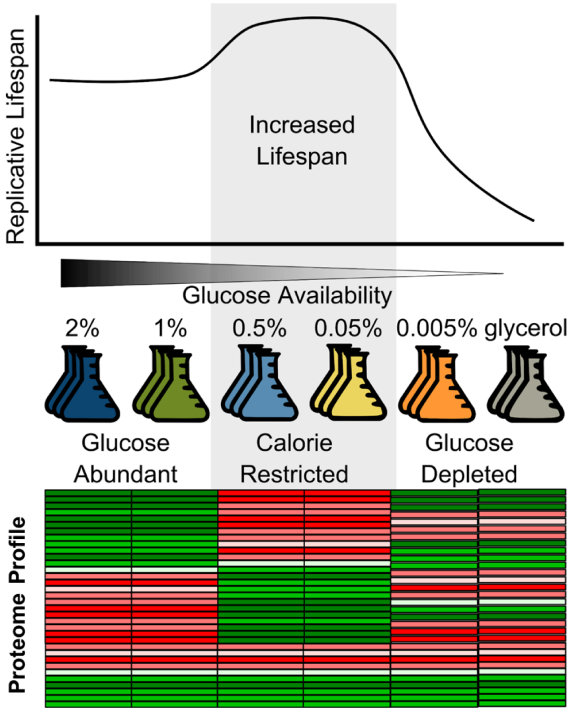


Figure 4.1: Yeast strain BY4741 was cultured under two glucose-abundant conditions (2% and 1% glucose in YEP), two glucose-restricted conditions (0.5% and 0.05% glucose in YEP), and two glucose-depleted conditions (0.005% glucose in YEP and glycerol in YEP).

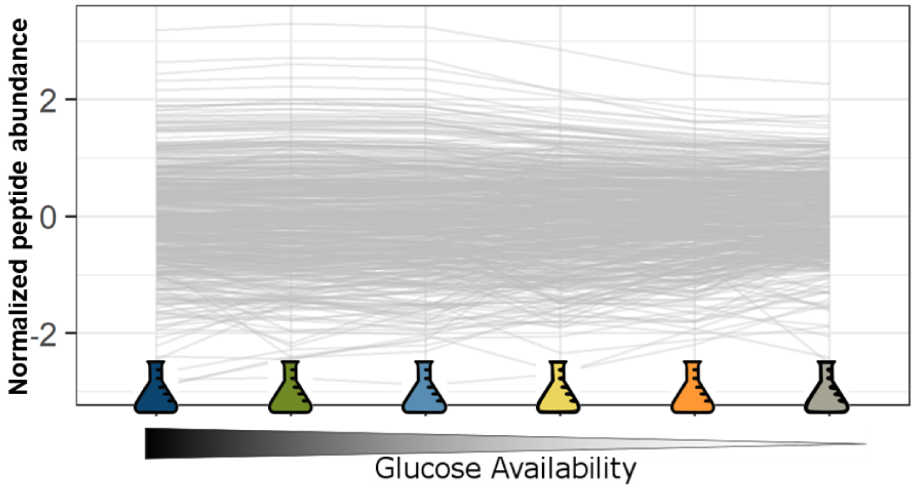


Figure 4.2: Peptide abundance shows no dose-dependent trends over glucose availability. Time course analysis (Storey et al., 2005) was applied to test for significant trends in peptide abundance over the continuous glucose availability.

Figure 4.3: Using initial or harvest glucose concentration labels cluster samples identically. Confusion matrices of k-means clustering with two values of k and two sample labels.

calorie restriction via a *switch* mechanism. In this scenario, proteins would only be differential in the CR condition, and not in either the glucose abundant or glucose depleted conditions. We tested this hypothesis using *consensus clustering* (Wilkerson and Hayes, 2010). In consensus clustering, a particular clustering technique is chosen (here, we used k-means clustering) and the data is clustered multiple times under different values of k, then the clusters are themselves clustered. If a sample is reproducibly clustered in the same group, it is more likely that the sample belongs in that group. When consensus clustering was performed on the peptide abundance profiles, clustering by k-means confirmed a k=2 (that is, two groups of samples among the three conditions and six doses). Consensus clustering even robustly preserves these two groups when k is increased to k=3 only moving one replicate of glycerol to an outgroup (Figure 4.3).

Unfortunately, the k=2 consensus cluster also holds when the sample labels are changed to the final glucose concentrations as measured at harvest, with one group represented by all samples with zero glucose at harvest and the second group including samples with any glucose remaining at harvest. Because these two clusters are perfectly preserved when class labels are changed to glucose concentration at the time of harvest, these clustering results are inconclusive.

During culture growth, we periodically sampled each replicate of each culture to measure the OD600 and the glucose concentration in the media, making measurements roughly once per doubling after an initial overnight growth. We harvested the cultures at the same OD600 in the interest of controlling biomass across the samples. Because the yeast grow more rapidly under glucose abundant conditions, this meant that the control 2% glucose condition was harvested earlier (chronologically) than the glucose deficient cultures. Unfortunately, this also meant the the glucose deficient conditions consumed or otherwise depleted nearly all of the available glucose in their media over their duration of incubation (Figure 4.4a). From the glucose concentration measurements, we see that the three conditions wit hthe most abundant starting glucose concentrations (2%, 1%, 0.5%) did not appreciably deplete their glucose, while the three most glucose deficient conditions (0.05%, 0.005%, glycerol) had consumed or otherwise depleted all available glucose before the cultures were harvested (Figure 4.4b).

To our knowledge, no other study of CR in yeast has also performed an enzymatic assay to

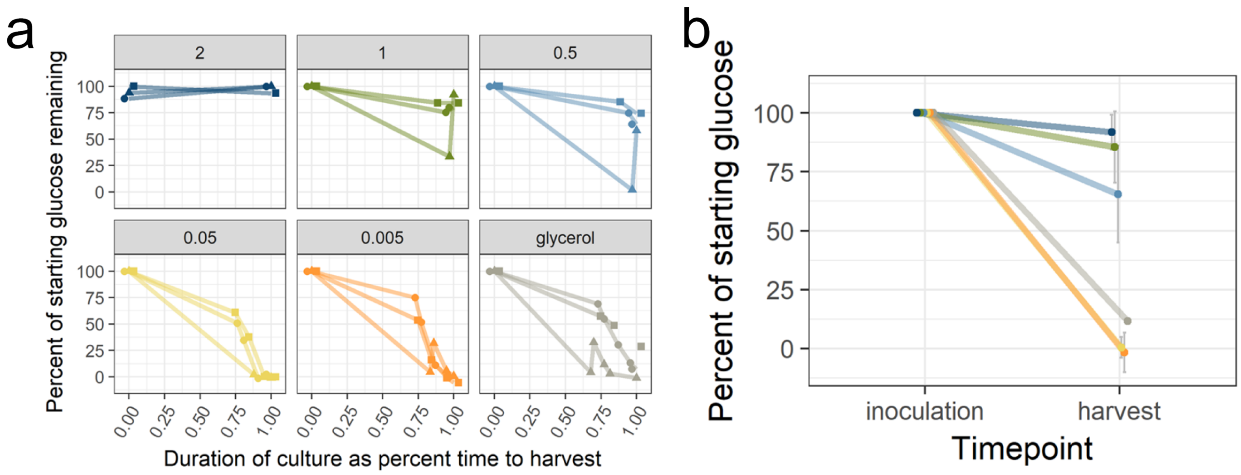


Figure 4.4: Cultures with low starting glucose concentrations rapidly consume their available glucose. (a) The concentration of glucose available in the growth media was tested periodically during culture growth for all conditions and all replicates. (b) Yeast grown in 2% glucose and 1% glucose did not consume all the available glucose in their media, having roughly the 100% of their initial glucose concentration available at the time of harvest. Cultures with lower initial glucose concentrations, especially the 0.05% and 0.005% glucose conditions,were depleted of all measurable glucose before the time of harvest.

measure the concentration of free glucose available to the yeast during what is believed to be calorie restriction. This may have important consequences on what we believe to be the mechanism of calorie restriction and appropriate calorie restriction dosage.

Despite the confounded labels, we analyzed the differential proteins between the two clusters. Using a t-test between the two clusters, we found a number of differentially abundant proteins, many of which are expected to be differential under glucose availability and even calorie restriction specifically. One of the most highly differential proteins after significance testing is Heat Shock Protein 12 (Hsp12) (Figure 4.5a). Hsp12 is documented in the Saccharomycese Genome Database as being glucose-repressed (`https://www.yeastgenome.org/locus/S000001880`), lending credibility to the observed differential abundance. Not only is the overall protein abundance highly differential between the low and high glucose groups, but all the peptides detected for Hsp12 trend consistently (Figure 4.5b) and the peptide chromatograms display ideal characteristics (Figure 4.5c). Hsp12 is also documented in SGD as being required for calorie-restriction

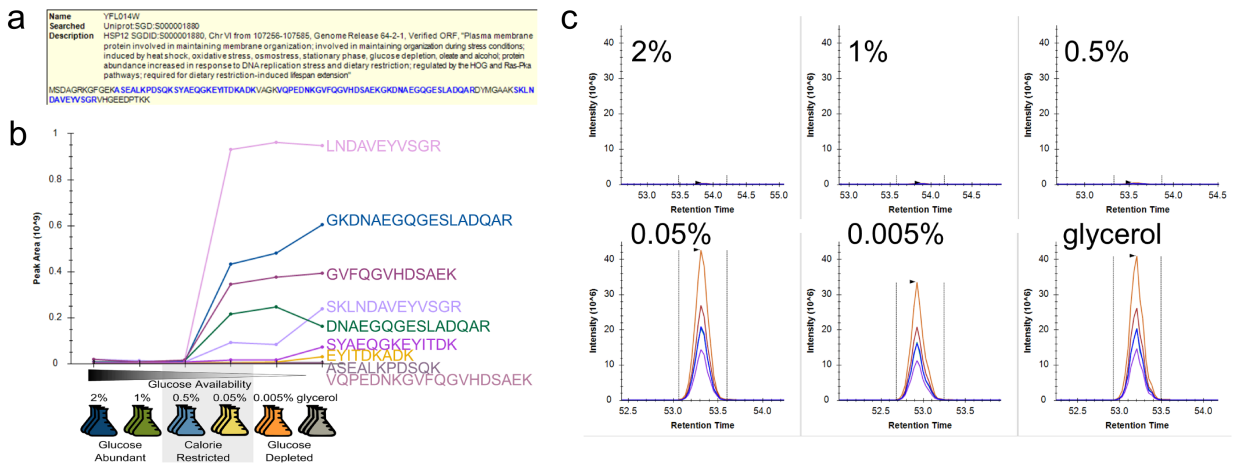


Figure 4.5: Increased expression of proteins required for lifespan extension clusters in the deficient glucose group.The plasma membrane protein Hsp12 (YFL014W) is known to increase in response to DNA replication stress (a hallmark of molecular aging) and in response to calorie restriction. Coverage of this protein is high and abundance is increased 7.5-fold in the group comprised of 0.05%, 0.005%, and glycerol samples compared to the group comprised of the 2%, 1%, and 0.5% glucose samples.

mediated lifespan extension, suggesting that the proteomic data here does capture the phenotype of lifespan extension.

4.5 Constructing quantitative molecular phenotypes of yeast longevity for life-span modulating genotypes.

As expected of a complex polygenic trait, several independent molecular pathways likely contribute to the ultimate aging phenotype. Existing work has identified multiple independent pathways that perturb cellular aging, including mitochondrial respiration. In this aim, I constructed molecular phenotypes for yeast mutants whose genotypes are associated with extended replicative lifespan. I hypothesized that molecular phenotypes would cluster into groups indicative of the complex functional processes underlying yeast longevity, yielding quantitative signatures of cellular aging. I performed differential protein abundance analysis with an ANOVA-based classic statistics approach, and finally clustered the proteome signatures into molecular phenotypes that

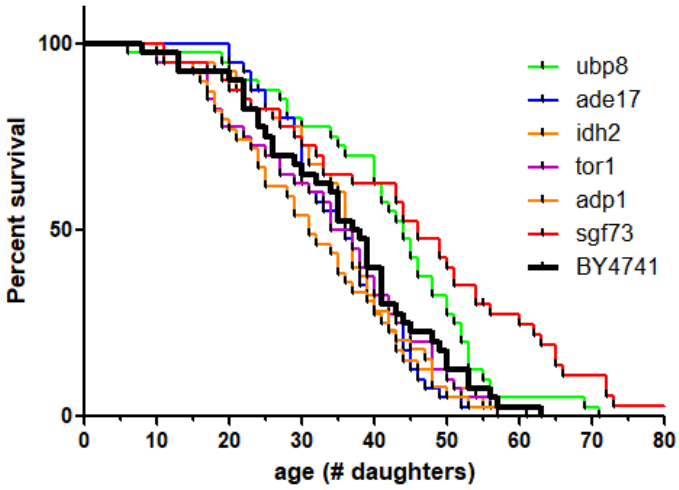


Figure 4.6: Survival curves for the yeast deletion strains used in this work. The results are complicated by the wildtype living very long (median lifespan was 37.5, normal wildtype should be closer to 27). Even with the very long-lived wildtype, Δsgf73 and Δudp8 appeared long. (n=40 cells started per genotype)

represent mechanisms of yeast longevity. I hypothesized that the resulting molecular phenotypes would replicate known processes such as mitochondrial respiration, reveal novel protein members of established processes, and discover processes not yet associated with yeast longevity.

The Δsgf73, Δubp8, Δidh2, and Δtor1 strains used in this work were selected to represent a range of replicative lifespan extension based on the results from McCormick 2015 (McCormick et al., 2015). We obtained the strains used in their work, cultured them under our conditions, and performed an RLS assay to confirm the lifespan for the exact strains used in this work's proteomic analyses.

Although strains Δsgf73 and Δubp8 were confirmed long-lived versus the control BY4741, the control strain itself lived longer than expected based on previous RLS assays (Figure 4.6). Because of this, the other two strains selected to represent RLS extension (Δidh2 and Δtor1) appear not to be long lived, at least these particular strains.

We performed differential protein abundance testing on the selected long-lived strains (Δsgf73,

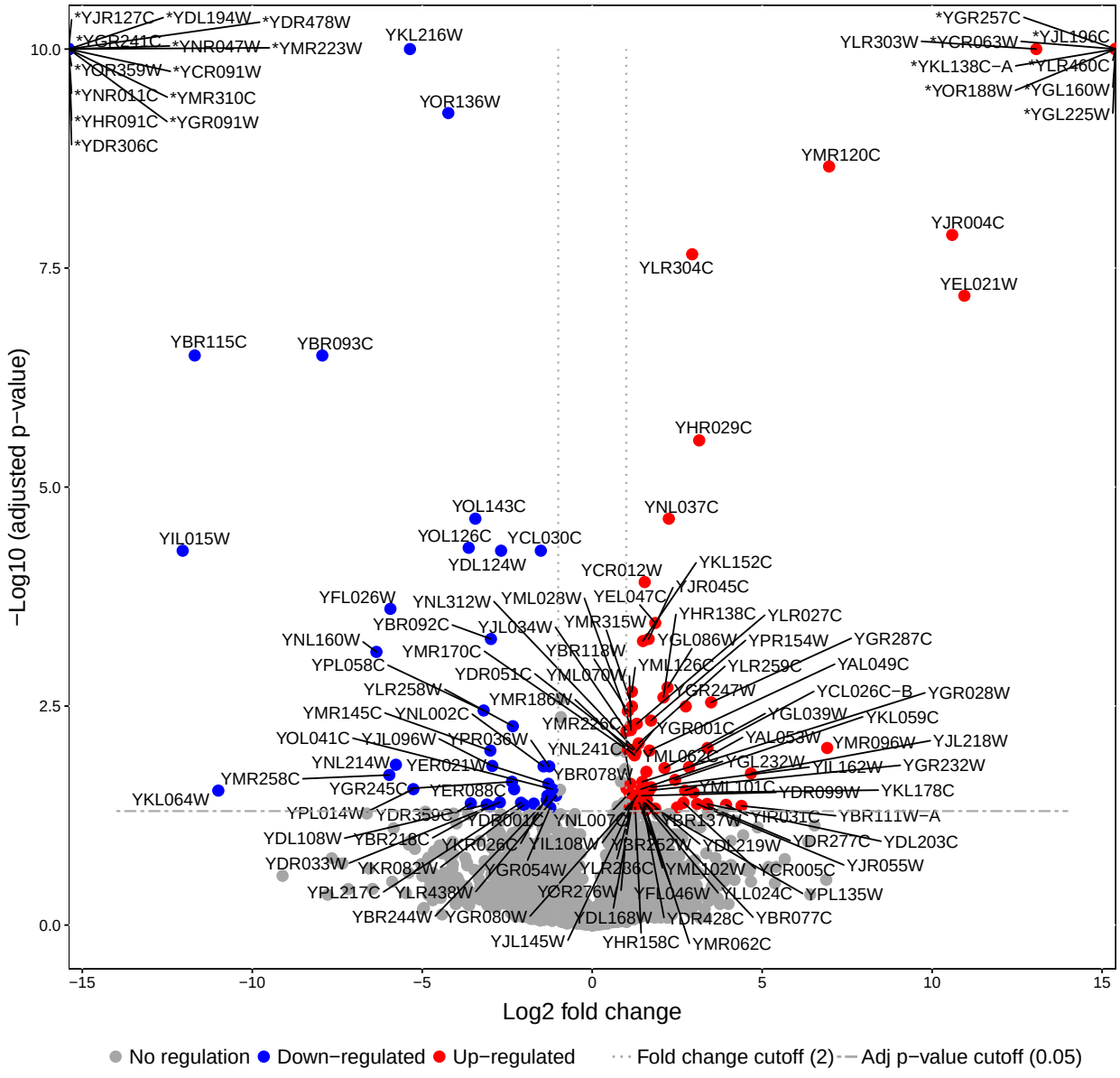


Figure 4.7: Volcano plot of protein abundances in the RLS-extending genotypes versus the control genotypes.

Δubp8, Δidh2, and Δtor1) versus the control strains (Δade17, Δadp1, BY4741 replicates cultured at the same time as the deletion strain replicates, and two BY4741 external reference cell pellets cultured at a different time but prepared alongside these strains). If a common signature of protein abundance was associated broadly across all RLS-extending genotypes, we might expect those proteins would be significant in a simple pairwise comparison. In total, we found 74 of 3553 detected proteins were significantly differential between the long-lived strains and the control strains (Figure 4.7, Appendix D.2, Appendix D.2). PANTHER GO-Slim Biological Process analysis comparing these 74 proteins against the background of all yeast genes found that tricarboxylic acid cycle and carbohydrate metabolic processes are overrepresented. PANTHER GO-Slim Molecular Function analysis additional finds that catalytic activity is overrepresented by these proteins.

To compare these high-level results between RLS-extension vs all other strains, we next performed pairwise comparisons between all single-gene deletions and the control BY4741 strain. The two negative control deletion strains, Δade17 (Figure 4.8) and Δadp1 (Figure 4.9) (Appendix D.2), show few differentially abundant proteins compared to the BY4741 experimental control. We might expect that these two strains would not have many differential proteins based on the number of known genetic interactors listed in SGD (Table 4.1).

The two strains chosen to represent a low RLS-extension but did not appear to be long-lived based on our in-house RLS assay, Δidh2 (Figure 4.10) and Δtor1 (Figure 4.11), both show slight more differentially abundant proteins compared to the BY4741 experimental control. Again, based on the number of known interactions for these two genes (Table 4.1), these results align with what we expect.

Last, the two strains both documented as extending RLS and proven in our hands to display RLS extension, Δsgf73 (Figure 4.12) and Δubp8 (Figure 4.13), show many differentially abundant proteins compared to the BY4741 experimental control. This also matched our expectations based on the number of unique interactions for these two genes (Table 4.1).

Of the 3553 proteins tested for differential abundance, 74 were determined significant when comparing the RLS genotypes versus the control genotypes. If we compare the results between the two group test with the results from the individual genotype tests, only 38 of those proteins

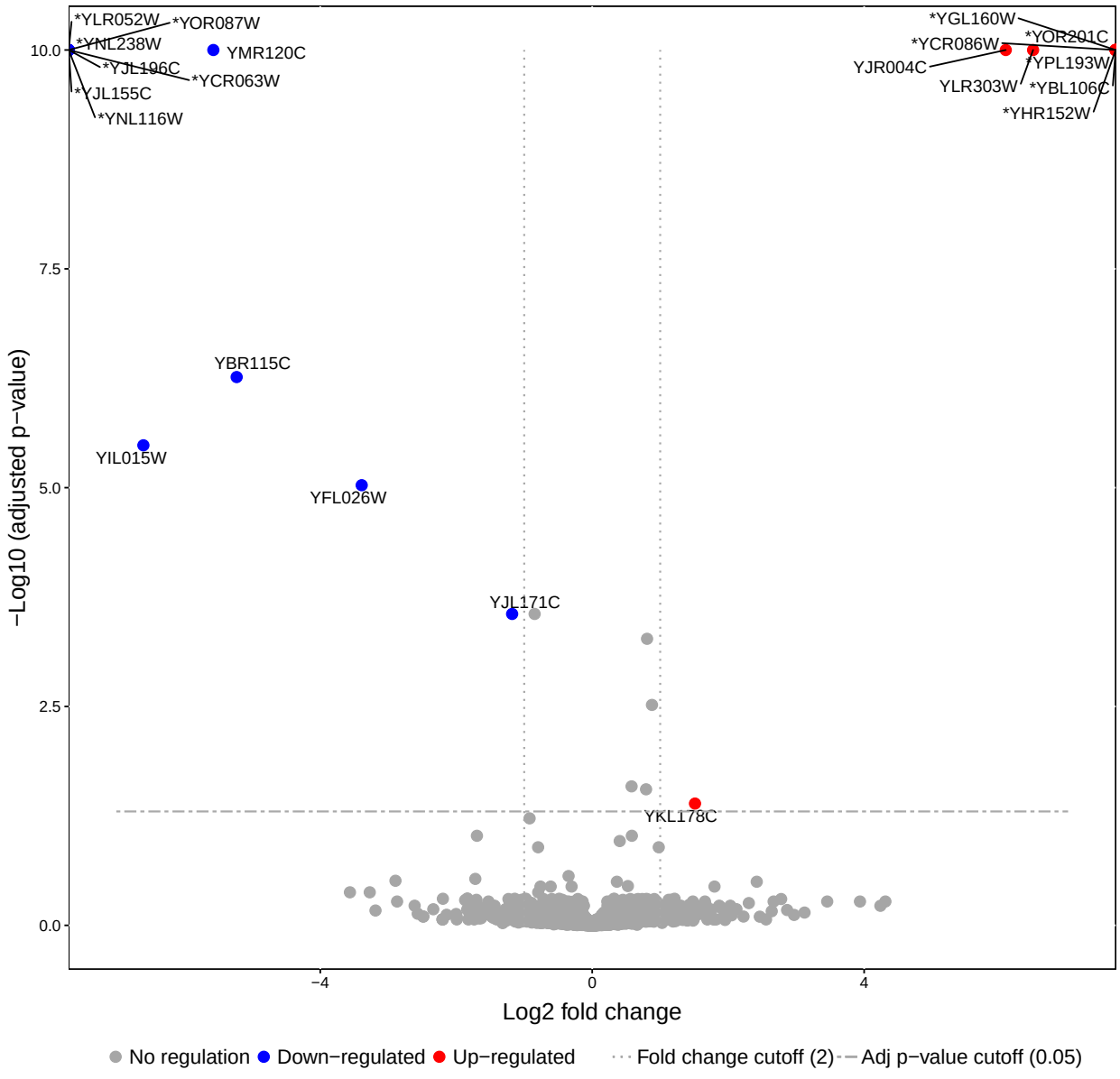


Figure 4.8: Volcano plot of protein abundances in Δade17 vs BY4741.

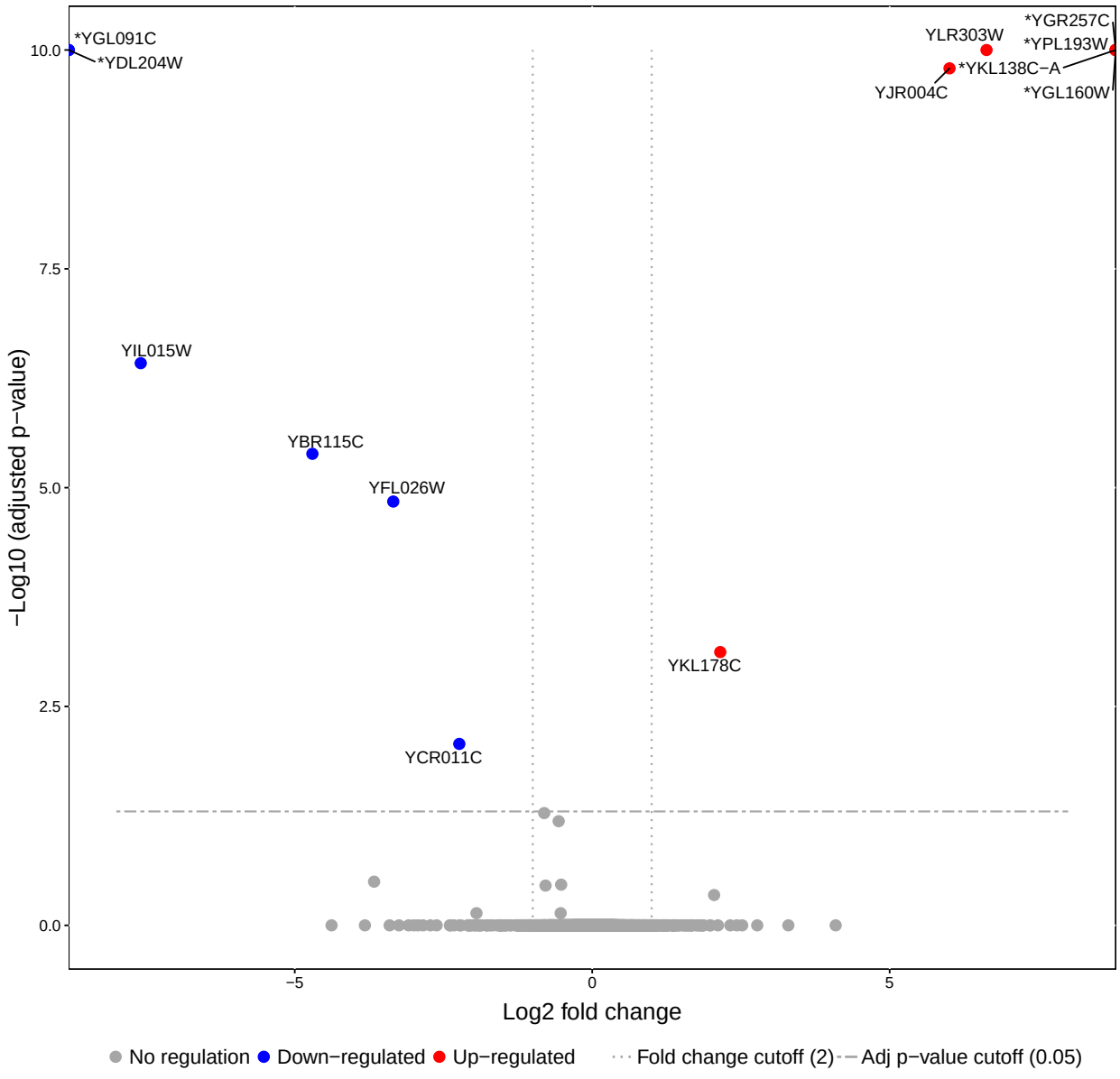


Figure 4.9: Volcano plot of protein abundances in Δadp1 vs BY4741.

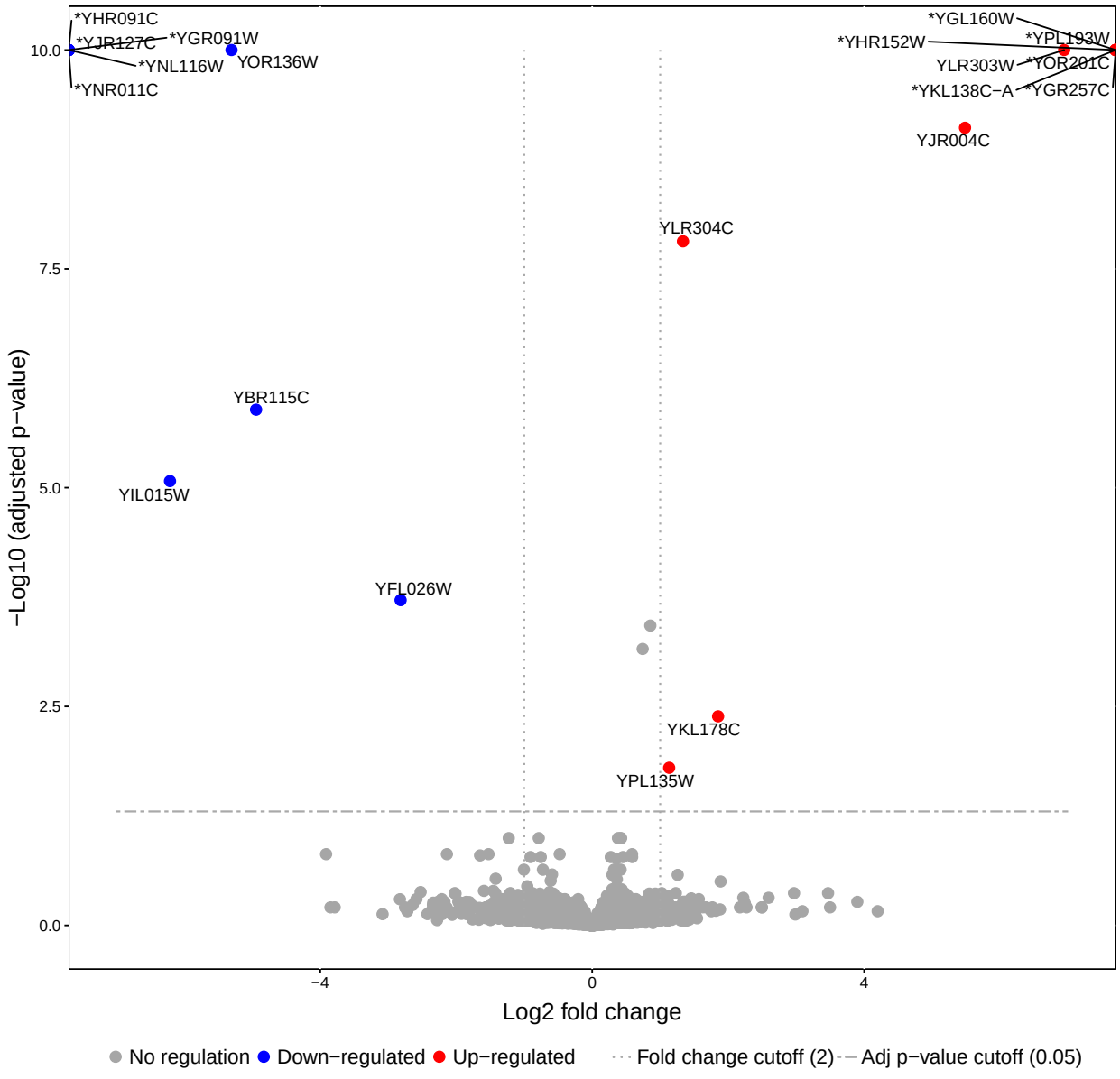


Figure 4.10: Volcano plot of protein abundances in Δidh2 vs BY4741.

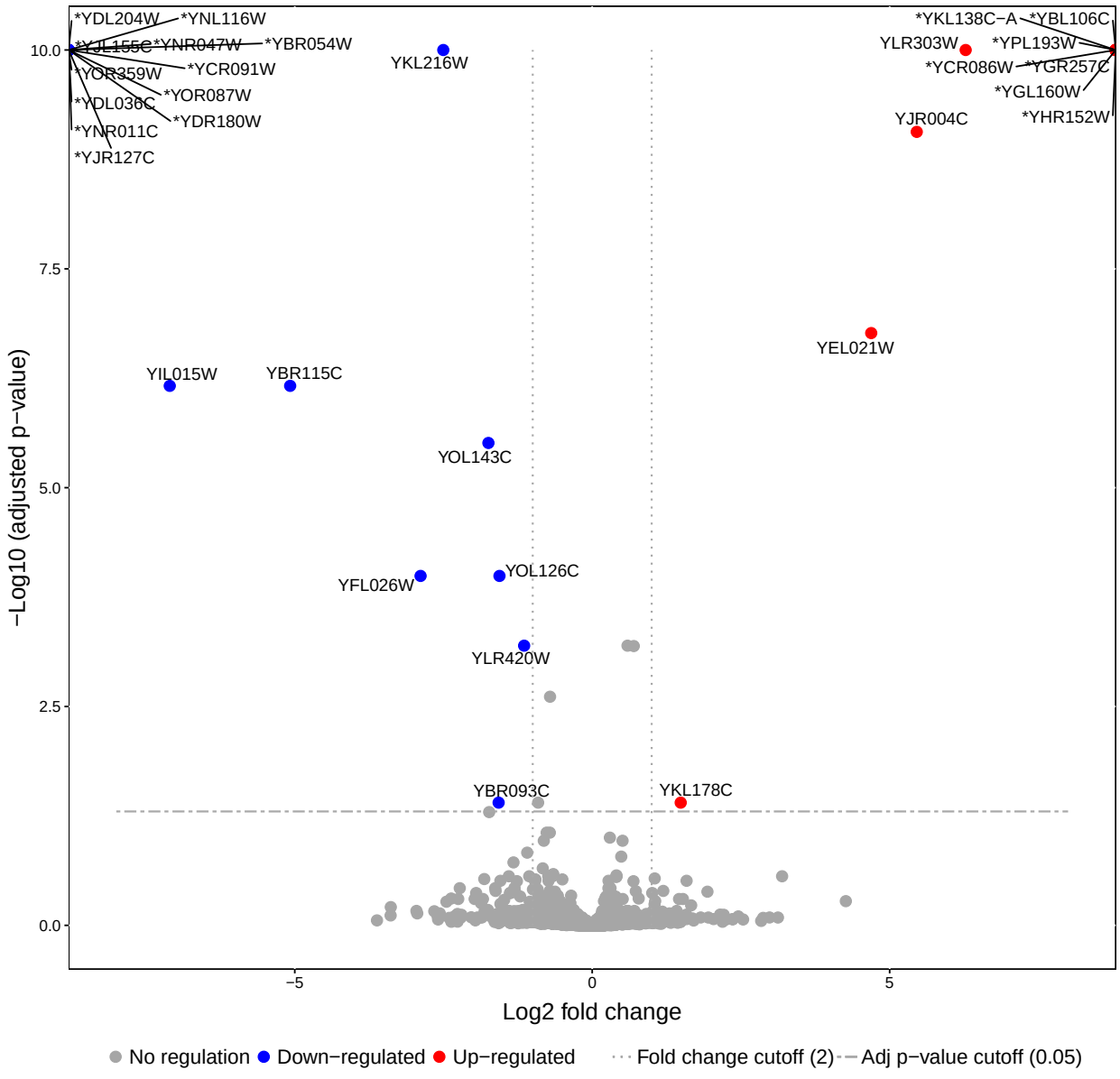


Figure 4.11: Volcano plot of protein abundances in Δtor1 vs BY4741.

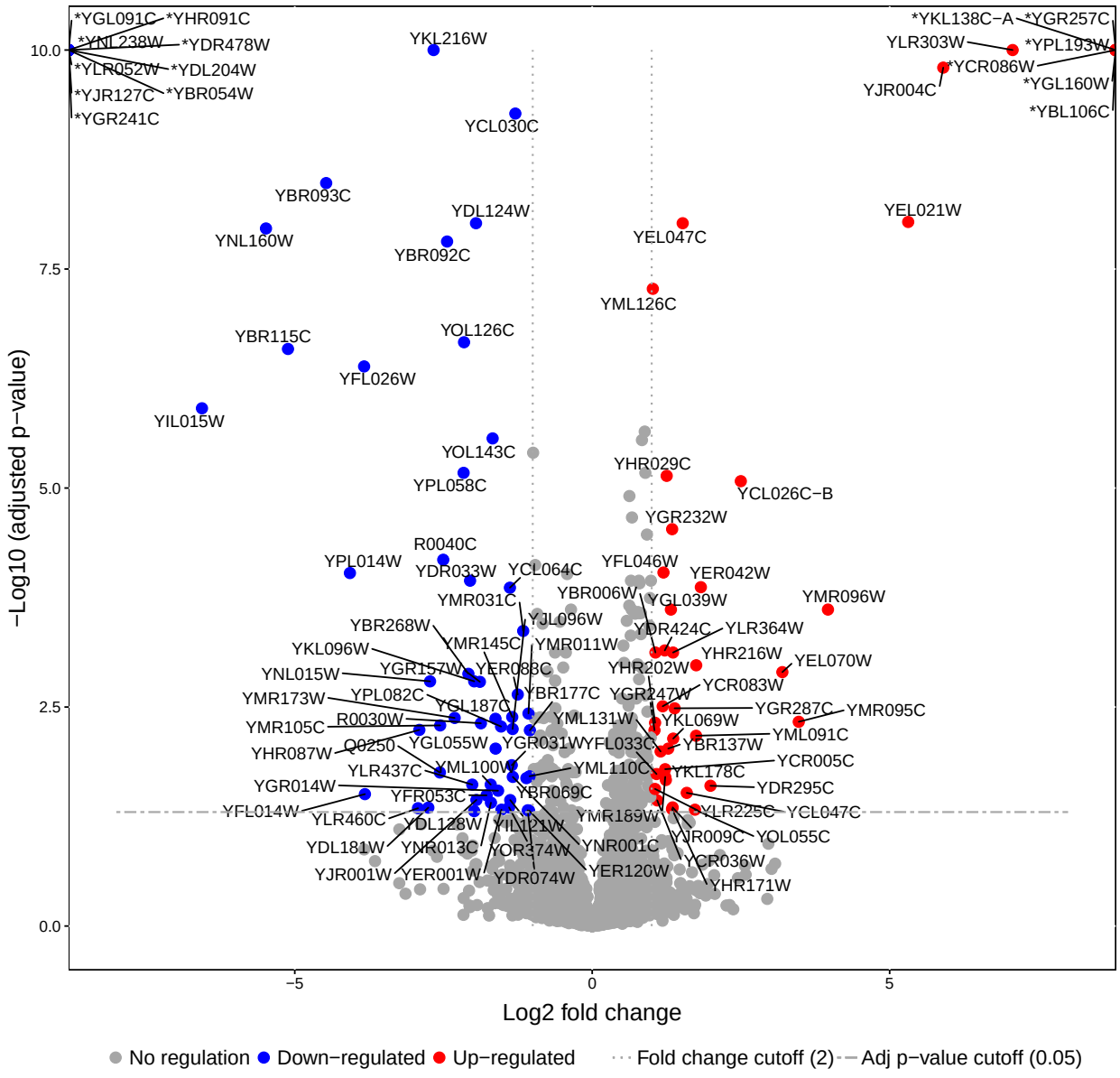


Figure 4.12: Volcano plot of protein abundances in Δsgf73 vs BY4741.

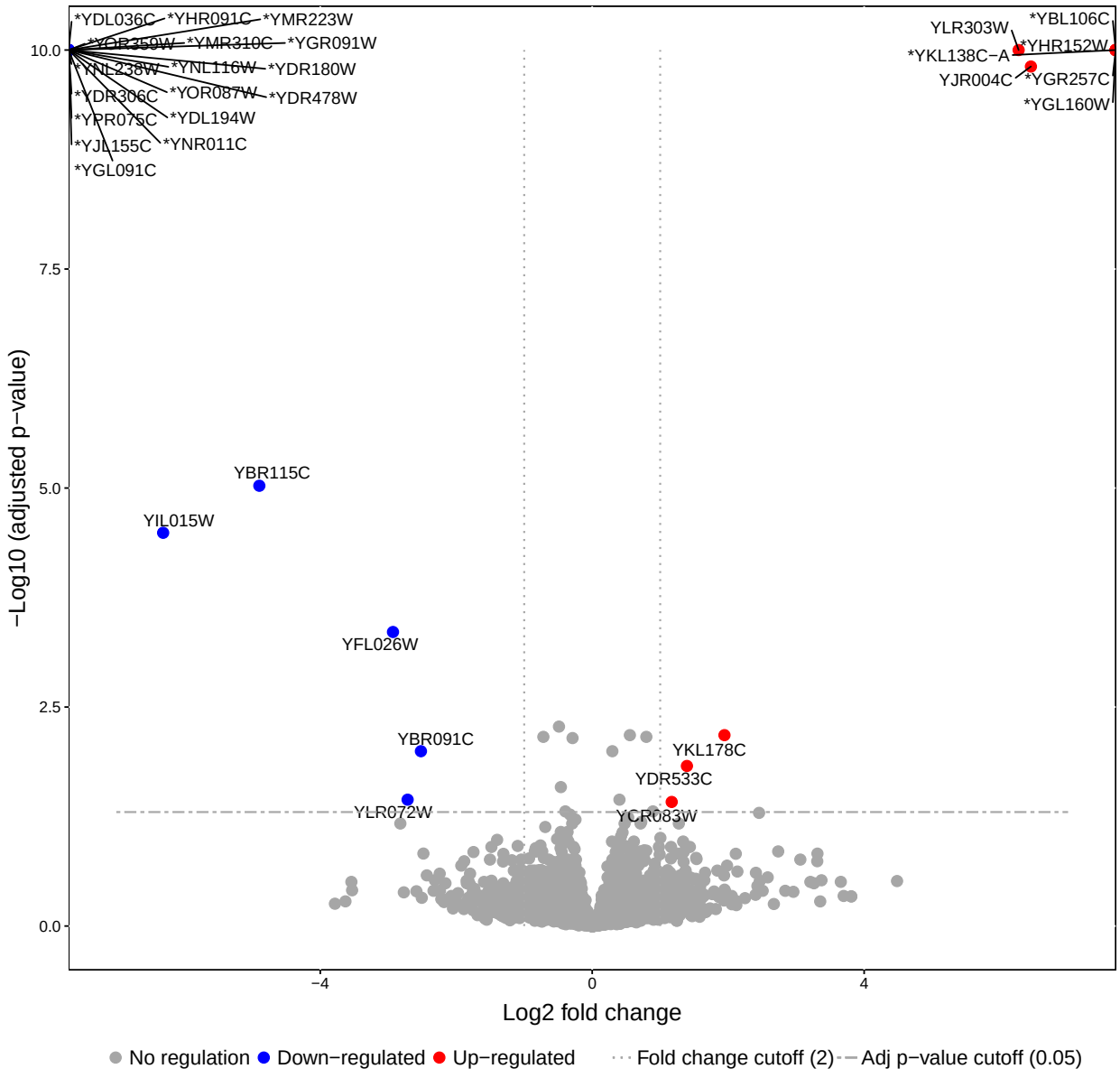


Figure 4.13: Volcano plot of protein abundances in Δubp8 vs BY4741.

were significant in at least one of the single genotype vs BY4741 tests (Appendix D.2). Further, of those 38 proteins, six were similarly differential in all genotypes and an additional two were differential in all but one genotype, suggesting that they may be involved with general response to the gene deletion process but are not likely involved in the extended RLS phenotype. Of the remaining 30 candidate proteins, three proteins are significant in three genotypes, five are significant in two genotypes, and the remaining are only significant in one genotype. This casual comparison suggests that there may not be a global signature of yeast RLS extension, at least using protein abundance data.

Because no clear protein signature is evident from comparing all four RLS-extending genotypes to the controls, we next split the genotypes into a high RLS extension group (Δsgf73 and Δubp8), a low RLS extension group (Δidh2 and Δtor1), a non-RLS extending group (Δade17 and Δadp1), and compared each of these three groups to the control BY4741. We additionally tested the external reference BY4741 cell pellet against the experimental control BY4741 as a sanity check, which found 14 significantly differential proteins disregarded from further consideration. We also found that three of the significantly differential proteins in this multigroup comparison were Idh2, Sgf73, and Ade17, which validates that the methods is working as intended, since these three proteins are three of the knocked out genes and therefore should indeed be absent from those strains.

When we compare the other differential proteins, we find that the low RLS group had 46 differential proteins; the high RLS group had 66. The groups shared 17 proteins differentially regulated at similar fold change which are affiliated with *ATP synthesis coupled electron/proton transport* per PANTHER overrepresentation testing, although this is a low number of proteins to use for overrepresentation testing. It's not surprising that Δidh2 may show differential proteins with this type of biological process annotation, as *idh2* is a subunit of a complex involved with the TCA cycle, but it is surprising to see that the extended RLS strains, Δsgf73 and Δubp8, also show an increase in those ATP synthesis-related proteins (Appendix D.2), since those two genes are part of the SAGA complex, which is typically associated with chromatin remodeling and not ATP synthesis. The differential proteins unique to the high-RLS genotypes did not have any significant overrepresentation; however, the differential proteins unique to the low-RLS genotypes was also

enriched for ATP synthesis coupled electron transport.

We can further determine which of these results is reasonable by looking back at the original mass spectrometry data. In particular, we can consider which of these significant proteins is represented by multiple peptides, which of those peptides is represented by multiple high-quality fragment ion chromatograms, and whether those peptides trend in the same abundance pattern across the RLS-extending genotypes. If these three conditions are met, the protein candidate should be considered more seriously. When we curate the 75 proteins differential in the RLS-extending groups, but not differential in the control and non-RLS extending groups, we quickly refine to 35 proteins with at least two unique peptides with high-quality fragment ion chromatograms. Refined proteins are nearly all lower abundance in the RLS extended groups, and are associated with cellular respiration, carboxylic acid metabolic processes, and nucleobase-containing small molecule metabolic processes.

4.6 Conclusions

The mechanisms of aging and life span modulation likely include multiple cellular processes; however, the range of molecular phenotypes that cells display as they age under different contexts and genetic backgrounds is not well understood. This work used state-of-the-art biochemical and proteomic methods to measure the molecular phenotypes of life span extension in yeast to reveal novel protein members of established processes and discover processes not yet associated with yeast longevity. Understandably these pathologies are complex and likely involve mechanisms beyond the differential expression of a specific protein or even a set of the same proteins. Our work is validated through several control measures that confirm the methods and data are appropriate, and we recapitulate known basic biology of aging, such as HSP12 association with calorie restriction. In addition, our differential proteomics reveals novel insights that generate hypotheses for further functional or spaciotemporal investigations. In particular, this chapter highlights two possible directions for future work: finer-tuned calorie restriction conditions to distinguish between the switch mechanism this work suggests and a potential calorie restriction level between 0.5% and

0.05% glucose concentration; and testing additional RLS-modulating genotypes to further stratify categories of RLS extension. With increasing quantitative proteomics projects being undertaken with at-scale methods such as the DIA-MS approaches used in this chapter, combining experiments together into larger data sets will likely afford more statistical power and increased ability to construct more sensitive and specific molecular phenotypes of complex pathologies.

Chapter 5: CLOSING REMARKS

Quantitative mass spectrometry experiments have historically been treated independently as individual assays developed and optimized for a specific laboratory and for a specific purpose (Carr et al., 2014). However, the growing adaption of DIA/SWATH methods now enable studies to build upon previous work and opens up the possibility of larger scale experiments across laboratories and time. With increasing experimental scales comes discussions about how best to integrate and combine these data sets.

5.1 SCALING QUANTITATIVE MASS SPECTROMETRY DATA RESPONSIBLY

One prospect that I find exciting about these large-scale endeavors is the re-utilization of data. Even within modest experimental efforts, the practical limitations of sample preparation – such as homogenization methods or the number of slots available in a centrifuge – preclude the complete randomization of more than just a couple dozen samples at a time, forcing experimentalists to split their samples into blocks that later reappear as batch effects during data analysis.

As mass spectrometry proteomics groups embark on more ambitious experiments like the ProCan and NCI60 projects (Guo et al., 2019), I anticipate it will become more and more crucial to include external reference materials during experimental design. Even though these projects have already begun, incorporating a working reference and global reference may harmonize these data sets.

Reutilization of data would also benefit from building data repositories that cater specifically to quantitative proteomics. While there are many raw data repositories, these resources typically focus on shotgun DDA reanalyses, so while they provide the means to search for peptides in the repository, they don't typically support the targeted quantitative proteomics workflows. Building a repository specifically designed to assist the reuse and reanalysis of quantitative proteomics data would require a standardized quantitative data format, experimental meta data descriptors, and

quality control/assurance approaches, but would enable researchers to build on prior work, rather than needlessly repeating past experiments.

5.2 Bridging proteomics data generation and data analysis

Before I began my graduate training, I recognized the rapid growth in quantitative proteomics data generation. I also realized that my classical training in biochemistry didn't include the necessary coding and statistical skills required to appropriately handle data at the scales it was being acquired, and so I resolved to gain these skills for myself in my predoctoral training. In doing so, I've positioned myself uniquely as a sort of liaison between the data generators and the data analysts in the mass spectrometry community.

I expect that, for future generations of mass spectrometrists, training will include more computational and statistical coursework. Until this new generation gains majority in the community, however, I expect that the most popular quantitative proteomics frameworks will be those that are not only well documented but also have easily accessible tutorials and workshops. Because most current mass spectrometrists aren't comfortable with programming or command line tools, I think moving the most common computational frameworks into user-friendly GUIs or Docker containers will be an easier bridge between generations. Until then, the ability to liaison between data generators and data analysts will be an important role for methods development and software adaptation.

5.3 Future Directions

5.3.1 External reference materials in mass spectrometry proteomics

Because not all analytes may be present in a given reference material under a single physiological condition, it will be important to build global reference materials that include all possible analytes. This may take the physical form of consensus pool mixtures made from a reference under multiple perturbations or disease states so that even condition-specific analytes are detectable. Then, working references with just a subset of all possible detected analytes would be used in individual

experiments, but calibrated periodically to the physical global reference.

Alternatively, computational solutions may be able to create *in silico* global references from imputing across multiple working references that represent a range of physiological conditions. This might take the form of something like current retention time alignment approaches, but instead of retention times, analyte abundances would be aligned across working references.

5.3.2 Detection and quantification as independent processes

As I demonstrated in Chapter 3, detection of a peptide doesn't imply that the peptide is quantitative. While the quantitative proteome appears to be a subset of the detected proteome, this may just be an artifact of the order of operations, which first relies on detection before quantitative assessment. In the future, it would be interesting to see this assumption challenged. That is, rather than detecting peptides first, it would be interesting to see a process that instead searches data for features that display quantitative properties, then perform detection on those selected features. Recent work has been done along these ideas for DDA analysis (The and Kall, 2019), but I think the scope of the idea could be expanded to DIA especially.

5.3.3 Molecular phenotypes beyond peptide abundances

Using quantitative proteomics even in presumably well-studied systems has revealed novel insights. For example, although genomics has already determined several gene-level subtypes of colon cancer, quantitative proteomics further clustered those subtypes based on their proteomic signatures (Zhang et al., 2014). Beyond applications in cancer, I foresee quantitative proteomics being used to further understand phenomena that cannot be explained by the genotype alone such as cellular differentiation, cardiovascular disease, and neurodegenerative disease.

Although some disease states and phenotypes are driven by differential protein abundances, I think even more molecular phenotypes are described by higher-level molecular interactions. I don't think that merely measuring proteome-wide abundances will answer all our questions about the status or mechanism of complex phenotypes. Even if quantitative proteomics alone is unable

to explain these phenomena, I think that these experiments will generate hypotheses for futher investigation. For example, although my work to model molecular phenotypes associated with yeast replicative life span in Chapter 4 was inconclusive on its own, the data itself is trustworthy and the conclusions do call into question previous work and new hypotheses to investigate.

This means that in the future quantitative proteomics may need to be paired with other technologies such as proximity-labeling approaches, post-translational modification detection, or even sequencing-based technologies like ATAC-seq which describe chromatin structure. Hypotheses generated by large scale experiments such as those used in this thesis may also need to be structurally or mechanistically validated using more traditional biochemical and genetic approaches.

5.4 Conclusion

Quantitative proteomics is poised to help bridge the gap between the genotypic information gleaned from DNA sequencing and the phenotypic presentations of biological structure and function. In this dissertation, I focused on data independent acquisition methods for quantitative proteomics due to the unprecedented scales of data that this approach enables, but many of the ideas and concepts explored here are generalizable to any analytical method. As larger and more ambitious quantitative proteomics experiments are undertaken, I anticipate that the analytical chemistry fundamentals demonstrated here will facilitate the robust and accurate molecular measurements required for biological advances.

BIBLIOGRAPHY

(2013). Method of the Year 2012. Nature Methods *10*, 1–1.

Abbatiello, S. E., Mani, D. R., Keshishian, H. and Carr, S. A. (2009). Automated Detection of Inaccurate and Imprecise Transitions in Peptide Quantification by Multiple Reaction Monitoring Mass Spectrometry. Clinical Chemistry *56*, 291–305.

Abbatiello, S. E., Schilling, B., Mani, D. R., Zimmerman, L. J., Hall, S. C., MacLean, B., Albertolle, M., Allen, S., Burgess, M., Cusack, M. P., Gosh, M., Hedrick, V., Held, J. M., Inerowicz, H. D., Jackson, A., Keshishian, H., Kinsinger, C. R., Lyssand, J., Makowski, L., Mesri, M., Rodriguez, H., Rudnick, P., Sadowski, P., Sedransk, N., Shaddox, K., Skates, S. J., Kuhn, E., Smith, D., Whiteaker, J. R., Whitwell, C., Zhang, S., Borchers, C. H., Fisher, S. J., Gibson, B. W., Liebler, D. C., MacCoss, M. J., Neubert, T. A., Paulovich, A. G., Regnier, F. E., Tempst, P. and Carr, S. A. (2015). Large-Scale Interlaboratory Study to Develop, Analytically Validate and Apply Highly Multiplexed, Quantitative Peptide Assays to Measure Cancer-Relevant Proteins in Plasma. Molecular & Cellular Proteomics *14*, 2357–2374.

Abelin, J. G., Patel, J., Lu, X., Feeney, C. M., Fagbami, L., Creech, A. L., Hu, R., Lam, D., Davison, D., Pino, L., Qiao, J. W., Kuhn, E., Officer, A., Li, J., Abbatiello, S., Subramanian, A., Sidman, R., Snyder, E., Carr, S. A. and Jaffe, J. D. (2016). Reduced-representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-scale Comparison of Drug-induced Phenotypes. Molecular & Cellular Proteomics *15*, 1622–1641.

Agger, S. A., Marney, L. C. and Hoofnagle, A. N. (2010). Simultaneous Quantification of Apolipoprotein A-I and Apolipoprotein B by Liquid-Chromatography-Multiple- Reaction-Monitoring Mass Spectrometry. Clinical Chemistry *56*, 1804–1813.

Amodei, D., Egertson, J., MacLean, B. X., Johnson, R., Merrihew, G. E., Keller, A., Marsh, D., Vitek, O., Mallick, P. and MacCoss, M. J. (2019). Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. Journal of The American Society for Mass Spectrometry *30*, 669–684.

Anderson, L. and Hunter, C. L. (2005). Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. Molecular & Cellular Proteomics *5*, 573–588.

Bereman, M. S., Beri, J., Sharma, V., Nathe, C., Eckels, J., MacLean, B. and MacCoss, M. J. (2016). An Automated Pipeline to Monitor System Performance in Liquid Chromatography–Tandem Mass Spectrometry Proteomic Experiments. Journal of Proteome Research *15*, 4763–4769.

Bilbao, A., Varesio, E., Luban, J., Strambio-De-Castillia, C., Hopfgartner, G., Müller, M. and Lisacek, F. (2015a). Processing strategies and software solutions for data-independent acquisition in mass spectrometry. PROTEOMICS *15*, 964–980.

Bilbao, A., Zhang, Y., Varesio, E., Luban, J., Strambio-De-Castillia, C., Lisacek, F. and Hopfgartner, G. (2015b). Ranking Fragment Ions Based on Outlier Detection for Improved Label-Free Quantification in Data-Independent Acquisition LC–MS/MS. Journal of Proteome Research *14*, 4581–4593.

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

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., Vitek, O., Rinner, O. and Reiter, L. (2015). Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. Molecular & Cellular Proteomics *14*, 1400–1410.

Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D. and Reiter, L. (2017). Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. Molecular & Cellular Proteomics *16*, 2296–2309.

Cadenas, E. and Davies, K. J. A. (2000). Mitochondrial free radical generation, oxidative stress, and aging11This article is dedicated to the memory of our dear friend, colleague, and mentor Lars Ernster (1920–1998), in gratitude for all he gave to us. Free Radical Biology and Medicine *29*, 222–230.

Campisi, J. (2003). Cellular senescence and apoptosis: how cellular responses might influence aging phenotypes. Experimental Gerontology *38*, 5–11.

Carr, S. A., Abbatiello, S. E., Ackermann, B. L., Borchers, C., Domon, B., Deutsch, E. W., Grant, R. P., Hoofnagle, A. N., Hüttenhain, R., Koomen, J. M., Liebler, D. C., Liu, T., MacLean, B., Mani, D. R., Mansfield, E., Neubert, H., Paulovich, A. G., Reiter, L., Vitek, O., Aebersold, R., Anderson, L., Bethem, R., Blonder, J., Boja, E., Botelho, J., Boyne, M., Bradshaw, R. A., Burlingame, A. L., Chan, D., Keshishian, H., Kuhn, E., Kinsinger, C., Lee, J. S. H., Lee, S.-W., Moritz, R., Oses-Prieto, J., Rifai, N., Ritchie, J., Rodriguez, H., Srinivas, P. R., Townsend, R. R., Eyk, J. V., Whiteley, G., Wiita, A. and Weintraub, S. (2014). Targeted Peptide Measurements in Biology and Medicine: Best Practices for Mass Spectrometry-based Assay Development Using a Fit-for-Purpose Approach. Molecular & Cellular Proteomics *13*, 907–917.

Cham Mead, J. A., Bianco, L. and Bessant, C. (2010). Free computational resources for designing selected reaction monitoring transitions. PROTEOMICS *10*, 1106–1126.

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L. and Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology *30*, 918–920.

Chapman, J. D., Goodlett, D. R. and Masselon, C. D. (2013). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. Mass Spectrometry Reviews *33*, 452–470.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S. and Wong, E. D. (2011). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research *40*, D700–D705.

Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B. and Vitek, O. (2014). MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics *30*, 2524–2526.

Codrea, M. C., Jiménez, C. R., Heringa, J. and Marchiori, E. (2007). Tools for computational processing of LC–MS datasets: A user\textquotesingles perspective. Computer Methods and Programs in Biomedicine *86*, 281–290.

Colangelo, C. M., Chung, L., Bruce, C. and Cheung, K.-H. (2013). Review of software tools for design and analysis of large scale MRM proteomic datasets. Methods *61*, 287–298.

Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R. L., Ohtsuki, S., Schlapbach, R., Selevsek, N., Thomas, S. N., Tzeng, S.-C., Zhang, H. and Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. Nature Communications *8*.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N. and Mann, M. (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. Molecular & Cellular Proteomics *13*, 2513–2526.

Craig, H. (1961a). Isotopic Variations in Meteoric Waters. Science *133*, 1702–1703.

Craig, H. (1961b). Standard for Reporting Concentrations of Deuterium and Oxygen-18 in Natural Waters. Science *133*, 1833–1834.

Craig, R. and Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinformatics *20*, 1466–1467.

Craig, R., Cortens, J. P. and Beavis, R. C. (2004). Open Source System for Analyzing, Validating, and Storing Protein Identification Data. Journal of Proteome Research *3*, 1234–1242.

Craig, R., Cortens, J. P. and Beavis, R. C. (2005). The use of proteotypic peptide libraries for protein identification. Rapid Communications in Mass Spectrometry *19*, 1844–1850.

Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. and Drummond, D. A. (2015). Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. PLOS Genetics *11*, e1005206.

Desiere, F. (2006). The PeptideAtlas project. Nucleic Acids Research *34*, D655–D658.

Deutsch, E. W., Perez-Riverol, Y., Chalkley, R. J., Wilhelm, M., Tate, S., Sachsenberg, T., Walzer, M., Käll, L., Delanghe, B., Böcker, S., Schymanski, E. L., Wilmes, P., Dorfer, V., Kuster, B., Volders, P.-J., Jehmlich, N., Vissers, J. P. C., Wolan, D. W., Wang, A. Y., Mendoza, L., Shofstahl, J., Dowsey, A. W., Griss, J., Salek, R. M., Neumann, S., Binz, P.-A., Lam, H., Vizcaíno, J. A., Bandeira, N. and Röst, H. (2018). Expanding the Use of Spectral Libraries in Proteomics. Journal of Proteome Research *17*, 4051–4060.

Domon, B. and Aebersold, R. (2010). Options and considerations when selecting a quantitative proteomics strategy. Nature Biotechnology *28*, 710–721.

Egertson, J. D., Kuehn, A., Merrihew, G. E., Bateman, N. W., MacLean, B. X., Ting, Y. S., Canterbury, J. D., Marsh, D. M., Kellmann, M., Zabrouskov, V., Wu, C. C. and MacCoss, M. J. (2013). Multiplexed MS/MS for improved data-independent acquisition. Nature Methods *10*, 744–746.

Egertson, J. D., MacLean, B., Johnson, R., Xuan, Y. and MacCoss, M. J. (2015). Multiplexed peptide analysis using data-independent acquisition and Skyline. Nature Protocols *10*, 887–903.

Eng, J. K., McCormack, A. L. and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry *5*, 976–989.

Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M. J. and Rinner, O. (2012). Using iRT, a normalized retention time for more targeted measurement of peptides. PROTEOMICS *12*, 1111–1121.

Eyers, C. E., Lawless, C., Wedge, D. C., Lau, K. W., Gaskell, S. J. and Hubbard, S. J. (2011). CONSeQuence: Prediction of Reference Peptides for Absolute Quantitative Proteomics Using Consensus Machine Learning Approaches. Molecular & Cellular Proteomics *10*, M110.003384.

Finney, G. L., Blackler, A. R., Hoopmann, M. R., Canterbury, J. D., Wu, C. C. and MacCoss, M. J. (2008). Label-Free Comparative Analysis of Proteomics Mixtures Using Chromatographic Alignment of High-Resolution $\upmu$LC-MS Data. Analytical Chemistry *80*, 961–971.

Franceschi, C., Bonafè, M., Valensin, S., Olivieri, F., De Luca, M., Ottaviani, E. and De Benedictis, G. (2000). Inflamm-aging. An evolutionary perspective on immunosenescence. Annals of the New York Academy of Sciences *908*, 244–254.

Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S. and MacCoss, M. J. (2006). Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. Analytical Chemistry *78*, 5678–5684.

Fujimoto, G. M., Monroe, M. E., Rodriguez, L., Wu, C., MacLean, B., Smith, R. D., MacCoss, M. J. and Payne, S. H. (2013). Accounting for Population Variation in Targeted Proteomics. Journal of Proteome Research *13*, 321–323.

Fusaro, V. A., Mani, D. R., Mesirov, J. P. and Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nature Biotechnology *27*, 190–198.

Galitzine, C., Egertson, J. D., Abbatiello, S., Henderson, C. M., Pino, L. K., MacCoss, M., Hoofnagle, A. N. and Vitek, O. (2018). Nonlinear Regression Improves Accuracy of Characterization of Multiplexed Mass Spectrometric Assays. Molecular & Cellular Proteomics *17*, 913–924.

Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R. and Mann, M. (2010). Super-SILAC mix for quantitative proteomics of human tumor tissue. Nature Methods *7*, 383–385.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K. and Weissman, J. S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P., Astromoff, A., Bakkoury, M. E., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kötter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W. and Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R. and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Molecular & Cellular Proteomics *11*, O111.016717.

Grant, R. P. and Hoofnagle, A. N. (2014). From Lost in Translation to Paradise Found: Enabling Protein Biomarker Method Transfer by Mass Spectrometry. Clinical Chemistry *60*, 941–944.

Greenbaum, D., Colangelo, C., Williams, K. and Gerstein, M. (2003). Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biology *4*, 117.

Griffin, P. R., Coffman, J. A., Hood, L. E. and Yates, J. R. (1991). Structural analysis of proteins by capillary HPLC electrospray tandem mass spectrometry. International Journal of Mass Spectrometry and Ion Processes *111*, 131–149.

Guo, T., Luna, A., Rajapakse, V. N., Koh, C. C., Wu, Z., Menden, M. P., Cheng, Y., Calzone, L., Martignetti, L., Ori, A., Iskar, M., Gillet, L., Zhong, Q., Varma, S., Schmitt, U., Qiu, P., Sun, Y., Zhu, Y., Wild, P. J., Garnett, M. J., Bork, P., Beck, M., Saez-Rodriguez, J., Reinhold, W. C., Sander, C., Pommier, Y. and Aebersold, R. (2019). Rapid proteotyping reveals cancer biology and drug response determinants in the NCI-60 cells. bioRxiv *na*.

Harman, D. (2003). The Free Radical Theory of Aging. Antioxidants & Redox Signaling *5*, 557–561.

Holstein (Sherwood), C. A., Gafken, P. R. and Martin, D. B. (2011). Collision Energy Optimization of b- and y-Ions for Multiple Reaction Monitoring Mass Spectrometry. Journal of Proteome Research *10*, 231–240.

Hoofnagle, A. N., Becker, J. O., Wener, M. H. and Heinecke, J. W. (2008). Quantification of Thyroglobulin, a Low-Abundance Serum Protein, by Immunoaffinity Peptide Enrichment and Tandem Mass Spectrometry. Clinical Chemistry *54*, 1796–1804.

Hoofnagle, A. N., Whiteaker, J. R., Carr, S. A., Kuhn, E., Liu, T., Massoni, S. A., Thomas, S. N., Townsend, R. R., Zimmerman, L. J., Boja, E., Chen, J., Crimmins, D. L., Davies, S. R., Gao, Y., Hiltke, T. R., Ketchum, K. A., Kinsinger, C. R., Mesri, M., Meyer, M. R., Qian, W.-J., Schoenherr, R. M., Scott, M. G., Shi, T., Whiteley, G. R., Wrobel, J. A., Wu, C., Ackermann, B. L., Aebersold, R., Barnidge, D. R., Bunk, D. M., Clarke, N., Fishman, J. B., Grant, R. P., Kusebauch, U., Kushnir, M. M., Lowenthal, M. S., Moritz, R. L., Neubert, H., Patterson, S. D., Rockwood, A. L., Rogers, J., Singh, R. J., Van Eyk, J., Wong, S. H., Zhang, S., Chan, D. W., Chen, X., Ellis, M. J., Liebler, D. C., Rodland, K. D., Rodriguez, H., Smith, R. D., Zhang, Z., Zhang, H. and Paulovich, A. G. (2015). Recommendations for the Generation, Quantification, Storage, and Handling of Peptides Used for Mass Spectrometry-Based Assays. Clinical Chemistry *62*.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jones, P., Cote, R. G., Cho, S. Y., Klie, S., Martens, L., Quinn, A. F., Thorneycroft, D. and Hermjakob, H. (2007). PRIDE: new developments and new datasets. Nucleic Acids Research *36*, D878–D883.

Kaeberlein, M., Kirkland, K. T., Fields, S. and Kennedy, B. K. (2004). Sir2-Independent Life Span Extension by Calorie Restriction in Yeast. PLoS Biology *2*, e296.

Klass, M. R. (1977). Aging in the nematode Caenorhabditis elegans: Major biological and environmental factors influencing life span. Mechanisms of Ageing and Development *6*, 413–429.

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. and MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods *4*, 923–925.

Krokhin, O. V. (2006). Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-A Pore Size C18 Sorbents. Analytical Chemistry *78*, 7785–7795.

Kuster, B., Schirle, M., Mallick, P. and Aebersold, R. (2005). Scoring proteomes with proteotypic peptide probes. Nature Reviews Molecular Cell Biology *6*, 577–583.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197.

Lange, V., Picotti, P., Domon, B. and Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. Molecular Systems Biology *4*.

Lawless, C., Holman, S. W., Brownridge, P., Lanthaler, K., Harman, V. M., Watkins, R., Hammond, D. E., Miller, R. L., Sims, P. F. G., Grant, C. M., Eyers, C. E., Beynon, R. J. and Hubbard, S. J. (2016). Direct and Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. Molecular & Cellular Proteomics *15*, 1309–1322.

Lawrence, R. T., Searle, B. C., Llovet, A. and Villen, J. (2016). Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nature Methods *13*, 431–434.

Lin, S.-J., Kaeberlein, M., Andalis, A. A., Sturtz, L. A., Defossez, P.-A., Culotta, V. C., Fink, G. R. and Guarente, L. (2002). Calorie restriction extends Saccharomyces cerevisiae lifespan by increasing respiration. Nature *418*, 344–348.

Ludwig, C., Claassen, M., Schmidt, A. and Aebersold, R. (2011). Estimation of Absolute Protein Quantities of Unlabeled Samples by Selected Reaction Monitoring Mass Spectrometry. Molecular & Cellular Proteomics *11*, M111.013987.

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C. and Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Molecular Systems Biology *14*, e8126.

Lynch, K. L. (2016). CLSI C62-A: A New Standard for Clinical Mass Spectrometry. Clinical Chemistry *62*, 24–29.

MacLean, B., Tomazela, D. M., Abbatiello, S. E., Zhang, S., Whiteaker, J. R., Paulovich, A. G., Carr, S. A. and MacCoss, M. J. (2010a). Effect of Collision Energy Optimization on the Measurement of Peptides by Selected Reaction Monitoring (SRM) Mass Spectrometry. Analytical Chemistry *82*, 10116–10124.

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

Mair, W. (2003). Demography of Dietary Restriction and Death in Drosophila. Science *301*, 1731–1733.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B. and Aebersold, R. (2006). Computational prediction of proteotypic peptides for quantitative proteomics. Nature Biotechnology *25*, 125–131.

Mani, D. R., Abbatiello, S. E. and Carr, S. A. (2012). Statistical characterization of multiple-reaction monitoring mass spectrometry (MRM-MS) assays for quantitative proteomics. BMC Bioinformatics *13*.

Mathivanan, S., Ahmed, M., Ahn, N. G., Alexandre, H., Amanchy, R., Andrews, P. C., Bader, J. S., Balgley, B. M., Bantscheff, M., Bennett, K. L., Björling, E., Blagoev, B., Bose, R., Brahmachari, S. K., Burlingame, A. S., Bustelo, X. R., Cagney, G., Cantin, G. T., Cardasis, H. L., Celis, J. E., Chaerkady, R., Chu, F., Cole, P. A., Costello, C. E., Cotter, R. J., Crockett, D., DeLany, J. P., Marzo, A. M. D., DeSouza, L. V., Deutsch, E. W., Dransfield, E., Drewes, G., Droit, A., Dunn, M. J., Elenitoba-Johnson, K., Ewing, R. M., Eyk, J. V., Faca, V., Falkner, J., Fang, X., Fenselau, C., Figeys, D., Gagné, P., Gelfi, C., Gevaert, K., Gimble, J. M., Gnad, F., Goel, R., Gromov, P., Hanash, S. M., Hancock, W. S., Harsha, H. C., Hart, G., Hays, F., He, F., Hebbar, P., Helsens, K., Hermeking, H., Hide, W., Hjernø, K., Hochstrasser, D. F., Hofmann, O., Horn, D. M., Hruban, R. H., Ibarrola, N., James, P., Jensen, O. N., Jensen, P. H., Jung, P., Kandasamy, K., Kheterpal, I., Kikuno, R. F., Korf, U., Körner, R., Kuster, B., Kwon, M.-S., Lee, H.-J., Lee, Y.-J., Lefevre, M., Lehvaslaiho, M., Lescuyer, P., Levander, F., Lim, M. S., Löbke, C., Loo,

J. A., Mann, M., Martens, L., Martinez-Heredia, J., McComb, M., McRedmond, J., Mehrle, A., Menon, R., Miller, C. A., Mischak, H., Mohan, S. S., Mohmood, R., Molina, H., Moran, M. F., Morgan, J. D., Moritz, R., Morzel, M., Muddiman, D. C., Nalli, A., Navarro, J. D., Neubert, T. A., Ohara, O., Oliva, R., Omenn, G. S., Oyama, M., Paik, Y.-K., Pennington, K., Pepperkok, R., Periaswamy, B., Petricoin, E. F., Poirier, G. G., Prasad, T. S. K., Purvine, S. O., Rahiman, B. A., Ramachandran, P., Ramachandra, Y. L., Rice, R. H., Rick, J., Ronnholm, R. H., Salonen, J., Sanchez, J.-C., Sayd, T., Seshi, B., Shankari, K., Sheng, S. J., Shetty, V., Shivakumar, K., Simpson, R. J., Sirdeshmukh, R., Siu, K. W. M., Smith, J. C., Smith, R. D., States, D. J., Sugano, S., Sullivan, M., Superti-Furga, G., Takatalo, M., Thongboonkerd, V., Trinidad, J. C., Uhlen, M., Vandekerckhove, J., Vasilescu, J., Veenstra, T. D., Vidal-Taboada, J.-M., Vihinen, M., Wait, R., Wang, X., Wiemann, S., Wu, B., Xu, T., Yates, J. R., Zhong, J., Zhou, M., Zhu, Y., Zurbig, P. and Pandey, A. (2008). Human Proteinpedia enables sharing of human protein data. Nature Biotechnology *26*, 164–167.

McCay, C. M., Crowell, M. F. and Maynard, L. A. (1935). The Effect of Retarded Growth Upon the Length of Life Span and Upon the Ultimate Body Size. The Journal of Nutrition *10*, 63–79.

McCormick, M. A., Delaney, J. R., Tsuchiya, M., Tsuchiyama, S., Shemorry, A., Sim, S., Chou, A. C.-Z., Ahmed, U., Carr, D., Murakami, C. J., Schleit, J., Sutphin, G. L., Wasko, B. M., Bennett, C. F., Wang, A. M., Olsen, B., Beyer, R. P., Bammler, T. K., Prunkard, D., Johnson, S. C., Pennypacker, J. K., An, E., Anies, A., Castanza, A. S., Choi, E., Dang, N., Enerio, S., Fletcher, M., Fox, L., Goswami, S., Higgins, S. A., Holmberg, M. A., Hu, D., Hui, J., Jelic, M., Jeong, K.-S., Johnston, E., Kerr, E. O., Kim, J., Kim, D., Kirkland, K., Klum, S., Kotireddy, S., Liao, E., Lim, M., Lin, M. S., Lo, W. C., Lockshon, D., Miller, H. A., Moller, R. M., Muller, B., Oakes, J., Pak, D. N., Peng, Z. J., Pham, K. M., Pollard, T. G., Pradeep, P., Pruett, D., Rai, D., Robison, B., Rodriguez, A. A., Ros, B., Sage, M., Singh, M. K., Smith, E. D., Snead, K., Solanky, A., Spector, B. L., Steffen, K. K., Tchao, B. N., Ting, M. K., Wende, H. V., Wang, D., Welton, K. L., Westman, E. A., Brem, R. B., Liu, X.-g., Suh, Y., Zhou, Z., Kaeberlein, M. and Kennedy, B. K. (2015). A Comprehensive Analysis of Replicative Lifespan in 4,698 Single-Gene Deletion Strains Uncovers Conserved Mechanisms of Aging. Cell Metabolism *22*, 895–906.

Mortimer, R. K. and Johnston, J. R. (1959). Life Span of Individual Yeast Cells. Nature *183*, 1751–1752.

Moseley, M. A., Deterding, L. J., Tomer, K. B. and Jorgenson, J. W. (1991). Nanoscale packed-capillary liquid chromatography coupled with mass spectrometry using a coaxial continuous-flow fast atom bombardment interface. Analytical Chemistry *63*, 1467–1473.

Moseley, M. A., Hughes, C. J., Juvvadi, P. R., Soderblom, E. J., Lennon, S., Perkins, S. R., Thompson, J. W., Steinbach, W. J., Geromanos, S. J., Wildgoose, J., Langridge, J. I., Richardson, K. and Vissers, J. P. C. (2018). Scanning Quadrupole Data-Independent Acquisition, Part A: Qualitative and Quantitative Characterization. Journal of Proteome Research *17*, 770–779.

Muntel, J., Boswell, S. A., Tang, S., Ahmed, S., Wapinski, I., Foley, G., Steen, H. and Springer, M. (2014). Abundance-based Classifier for the Prediction of Mass Spectrometric Peptide Detectability Upon Enrichment (PPA). Molecular & Cellular Proteomics *14*, 430–440.

Netzel, B. C., Grant, R. P., Hoofnagle, A. N., Rockwood, A. L., Shuford, C. M. and Grebe, S. K. G. (2016). First Steps toward Harmonization of LC-MS/MS Thyroglobulin Assays. Clinical Chemistry *62*, 297–299.

Nic, M., Jirat, J., Kosata, B., Jenkins, A. and McNaught, A. (2009). IUPAC Compendium of Chemical Terminology: Gold Book. 2.1.0 edition, IUPAC, Research Triagle Park, NC.

Omasits, U., Ahrens, C. H., Muller, S. and Wollscheid, B. (2013). Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics *30*, 884–886.

Ong, S.-E. and Mann, M. (2005). Mass spectrometry–based proteomics turns quantitative. Nature Chemical Biology *1*, 252–262.

Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., Cheung, K., Costello, C. E., Hermjakob, H., Huang, S., Julian, R. K., Kapp, E., McComb, M. E., Oliver, S. G., Omenn, G., Paton, N. W., Simpson, R., Smith, R., Taylor, C. F., Zhu, W. and Aebersold, R. (2004). A common open representation of mass spectrometry data and its application to proteomics research. Nature Biotechnology *22*, 1459–1466.

Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. and Coon, J. J. (2012). Parallel Reaction Monitoring for High Resolution and High Mass Accuracy Quantitative, Targeted Proteomics. Molecular & Cellular Proteomics *11*, 1475–1488.

Petritis, B. O., Qian, W.-J., Camp, D. G. and Smith, R. D. (2009). A Simple Procedure for Effective Quenching of Trypsin Activity and Prevention of 18O-Labeling Back-Exchange. Journal of Proteome Research *8*, 2157–2163.

Picotti, P. and Aebersold, R. (2012). Selected reaction monitoring–based proteomics: workflows, potential, pitfalls and future directions. Nature Methods *9*, 555–566.

Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. and Aebersold, R. (2009). Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics. Cell *138*, 795–806.

Pino, L. K., Searle, B. C., Huang, E. L., Noble, W. S., Hoofnagle, A. N. and MacCoss, M. J. (2018). Calibration Using a Single-Point External Reference Material Harmonizes Quantitative Mass Spectrometry Proteomics Data between Platforms and Laboratories. Analytical Chemistry *90*, 13112–13117.

Powers, R. W. (2006). Extension of chronological life span in yeast by decreased TOR pathway signaling. Genes & Development *20*, 174–184.

Prakash, A., Tomazela, D. M., Frewen, B., MacLean, B., Merrihew, G., Peterman, S. and MacCoss, M. J. (2009). Expediting the Development of Targeted SRM Assays: Using Data from Shotgun Proteomics to Automate Method Development. Journal of Proteome Research *8*, 2733–2739.

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

Reubsaet, L., Sweredoski, M. J. and Moradian, A. (2019). Data-Independent Acquisition for the Orbitrap Q Exactive HF: A Tutorial. Journal of Proteome Research *18*, 803–813.

Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., Faini, M., Schubert, O. T., Faridi, P., Ebhardt, H. A., Matondo, M., Lam, H., Bader, S. L., Campbell, D. S., Deutsch, E. W., Moritz, R. L., Tate, S. and Aebersold, R. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS. Scientific Data *1*.

Rost, H., Malmstrom, L. and Aebersold, R. (2012). A Computational Tool to Detect and Avoid Redundancy in Selected Reaction Monitoring. Molecular & Cellular Proteomics *11*, 540–549.

Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L. and Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nature Biotechnology *32*, 219–223.

Schilling, B., Rardin, M. J., MacLean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., Sorensen, D. J., Bereman, M. S., Jing, E., Wu, C. C., Verdin, E., Kahn, C. R., MacCoss, M. J. and Gibson, B. W. (2012). Platform-independent and Label-free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline. Molecular & Cellular Proteomics *11*, 202–214.

Schleit, J., Johnson, S. C., Bennett, C. F., Simko, M., Trongtham, N., Castanza, A., Hsieh, E. J., Moller, R. M., Wasko, B. M., Delaney, J. R., Sutphin, G. L., Carr, D., Murakami, C. J., Tocchi, A., Xian, B., Chen, W., Yu, T., Goswami, S., Higgins, S., Holmberg, M., Jeong, K.-S., Kim, J. R., Klum, S., Liao, E., Lin, M. S., Lo, W., Miller, H., Olsen, B., Peng, Z. J., Pollard, T., Pradeep, P., Pruett, D., Rai, D., Ros, V., Singh, M., Spector, B. L., Wende, H. V., An, E. H., Fletcher, M., Jelic, M., Rabinovitch, P. S., MacCoss, M. J., Han, J.-D. J., Kennedy, B. K. and Kaeberlein, M. (2013). Molecular mechanisms underlying genotype-dependent responses to dietary restriction. Aging Cell *12*, 1050–1061.

Schrimpf, S. P., Weiss, M., Reiter, L., Ahrens, C. H., Jovanovic, M., Malmström, J., Brunner, E., Mohanty, S., Lercher, M. J., Hunziker, P. E., Aebersold, R., Mering, C. v. and Hengartner, M. O. (2009). Comparative Functional Analysis of the Caenorhabditis elegans and Drosophila melanogaster Proteomes. PLoS Biology *7*, e1000048.

Searle, B. C., Egertson, J. D., Bollinger, J. G., Stergachis, A. B. and MacCoss, M. J. (2015). Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. Molecular & Cellular Proteomics *14*, 2331–2340.

Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., Villen, J. and MacCoss, M. J. (2018). Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nature Communications *9*.

Selevsek, N., Chang, C.-Y., Gillet, L. C., Navarro, P., Bernhardt, O. M., Reiter, L., Cheng, L.-Y., Vitek, O. and Aebersold, R. (2015). Reproducible and Consistent Quantification of the Saccharomyces cerevisiae Proteome by SWATH-mass spectrometry. Molecular & Cellular Proteomics *14*, 739–749.

Sharma, V., Eckels, J., Taylor, G. K., Shulman, N. J., Stergachis, A. B., Joyner, S. A., Yan, P., Whiteaker, J. R., Halusa, G. N., Schilling, B., Gibson, B. W., Colangelo, C. M., Paulovich, A. G., Carr, S. A., Jaffe, J. D., MacCoss, M. J. and MacLean, B. (2014). Panorama: A Targeted Proteomics Knowledge Base. Journal of Proteome Research *13*, 4205–4210.

Sherrod, S. D., Myers, M. V., Li, M., Myers, J. S., Carpenter, K. L., MacLean, B., MacCoss, M. J., Liebler, D. C. and Ham, A.-J. L. (2012). Label-Free Quantitation of Protein Modifications by Pseudo Selected Reaction Monitoring with Internal Reference Peptides. Journal of Proteome Research *11*, 3467–3479.

Sherwood, C. A., Eastham, A., Lee, L. W., Risler, J., Mirzaei, H., Falkner, J. A. and Martin, D. B. (2009). Rapid Optimization of MRM-MS Instrument Parameters by Subtle Alteration of Precursor and Productm/zTargets. Journal of Proteome Research *8*, 3746–3751.

Shuford, C. M., Walters, J. J., Holland, P. M., Sreenivasan, U., Askari, N., Ray, K. and Grant, R. P. (2017). Absolute Protein Quantification by Mass Spectrometry: Not as Simple as Advertised. Analytical Chemistry *89*, 7406–7415.

Smith, E. D., Tsuchiya, M., Fox, L. A., Dang, N., Hu, D., Kerr, E. O., Johnston, E. D., Tchao, B. N., Pak, D. N., Welton, K. L., Promislow, D. E. L., Thomas, J. H., Kaeberlein, M. and Kennedy, B. K. (2008). Quantitative evidence for conserved longevity pathways between divergent eukaryotic species. Genome Research *18*, 564–570.

Spahr, C. S., Davis, M. T., McGinley, M. D., Robinson, J. H., Bures, E. J., Beierle, J., Mort, J., Courchesne, P. L., Chen, K., Wahl, R. C., Yu, W., Luethy, R. and Patterson, S. D. (2001). Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I.Profiling an unfractionated tryptic digest. PROTEOMICS *1*, 93–107.

Spicer, V., Yamchuk, A., Cortens, J., Sousa, S., Ens, W., Standing, K. G., Wilkins, J. A. and Krokhin, O. V. (2007). Sequence-Specific Retention Calculator. A Family of Peptide Retention Time Prediction Algorithms in Reversed-Phase HPLC: Applicability to Various Chromatographic Conditions and Columns. Analytical Chemistry *79*, 8762–8768.

Steinkraus, K. A., Kaeberlein, M. and Kennedy, B. K. (2008). Replicative Aging in Yeast: The Means to the End. Annual Review of Cell and Developmental Biology *24*, 29–54.

Stergachis, A. B., MacLean, B., Lee, K., Stamatoyannopoulos, J. A. and MacCoss, M. J. (2011). Rapid empirical discovery of optimal peptides for targeted proteomics. Nature Methods *8*, 1041–1043.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments. Proceedings of the National Academy of Sciences *102*, 12837–12842.

The, M. and Kall, L. (2019). Focus on the spectra that matter by clustering of quantification data in shotgun proteomics. bioRxiv *na*.

Thienpont, L. M., Van Uytfanghe, K. and De Leenheer, A. P. (2002). Reference measurement systems in clinical chemistry. Clinica Chimica Acta *323*, 73–87.

Ting, Y. S., Egertson, J. D., Bollinger, J. G., Searle, B. C., Payne, S. H., Noble, W. S. and MacCoss, M. J. (2017). PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nature Methods *14*, 903–908.

Ting, Y. S., Egertson, J. D., Payne, S. H., Kim, S., MacLean, B., Käll, L., Aebersold, R., Smith, R. D., Noble, W. S. and MacCoss, M. J. (2015). Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Molecular & Cellular Proteomics *14*, 2301–2307.

Tosato, M., Zamboni, V., Ferrini, A. and Cesari, M. (2007). The aging process and potential interventions to extend life expectancy. Clinical Interventions in Aging *2*, 401–412.

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

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

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson,

D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Wasko, B. M. and Kaeberlein, M. (2013). Yeast replicative aging: a paradigm for defining conserved longevity interventions. FEMS Yeast Research *14*, 148–159.

Weisbrod, C. R., Eng, J. K., Hoopmann, M. R., Baker, T. and Bruce, J. E. (2012). Accurate Peptide Fragment Mass Analysis: Multiplexed Peptide Identification and Quantification. Journal of Proteome Research *11*, 1621–1632.

Wilkerson, M. D. and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Yang, J., McCormick, M. A., Zheng, J., Xie, Z., Tsuchiya, M., Tsuchiyama, S., El-Samad, H., Ouyang, Q., Kaeberlein, M., Kennedy, B. K. and Li, H. (2015). Systematic analysis of asymmetric partitioning of yeast proteome between mother and daughter cells reveals "aging factors" and mechanism of lifespan asymmetry. Proceedings of the National Academy of Sciences *112*, 11977–11982.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. C., Liebler, D. C. and the NCI CPTAC (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhang, H., Liu, Q., Zimmerman, L. J., Ham, A.-J. L., Slebos, R. J. C., Rahman, J., Kikuchi, T., Massion, P. P., Carbone, D. P., Billheimer, D. and Liebler, D. C. (2011). Methods for Peptide and Protein Quantitation by Liquid Chromatography-Multiple Reaction Monitoring Mass Spectrometry. Molecular & Cellular Proteomics *10*, M110.006593.

Zhang, Y., Bilbao, A., Bruderer, T., Luban, J., Strambio-De-Castillia, C., Lisacek, F., Hopfgartner, G. and Varesio, E. (2015). The Use of Variable Q1 Isolation Windows Improves Selectivity in LC–SWATH–MS Acquisition. Journal of Proteome Research *14*, 4359–4371.

Zhang, Z. (2004). Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. Analytical Chemistry *76*, 3908–3922.

Zhang, Z. (2005). Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides with Three or More Charges. Analytical Chemistry *77*, 6364–6373.

Appendix A: ENCYCLOPEDIA TUTORIALS FOR THE CHROMATOGRAM LIBRARY APPROACH TO DIA-MS DATA ANALYSIS

DIA DATA ANALYSIS WITH THE ENCYCLOPEDIA SOFTWARE SUITE

This tutorial is a practical guide for how to use the Encyclopedia software suite (Searle 2018, https://www.nature.com/articles/s41467-018-07454-w) for the chromatogram library approach to DIA-MS. We have a GUI-based workflow and also a command line workflow (thanks @atkeller). I've included options for visualizing the results of the Encyclopedia analysis in Skyline or in Encyclopedia itself ("Bri-line").

Citations

MSconvert (https://www.nature.com/articles/nbt.2377)

A cross-platform toolkit for mass spectrometry and proteomics. Chambers MC et al. *Nat Biotech* 30, 918-920 (2012). doi.org/10.1038/nbt.2377

Encyclopedia (https://www.nature.com/articles/s41467-018-07454-w)

Searle BC et al. *Nat Comm* 9, 5128 (2018). doi.org/10.1038/s41467-018-07454-w

You will need:

- MSConvert from Proteowizard: *Windows only!*
  - http://proteowizard.sourceforge.net/download.html
- EncyclopeDIA suite (*.jar file): *command line and cross-platform GUI*
  - https://bitbucket.org/searleb/encyclopedia/downloads/

Tl;dr three steps for DIA-MS analysis by chromatogram library

1. Convert .raw files to .mzML using MSConvert
2. Build library using Walnut or XCorDIA in EncyclopeDIA
3. Search wide-window data with library from step 2 using EncyclopeDIA

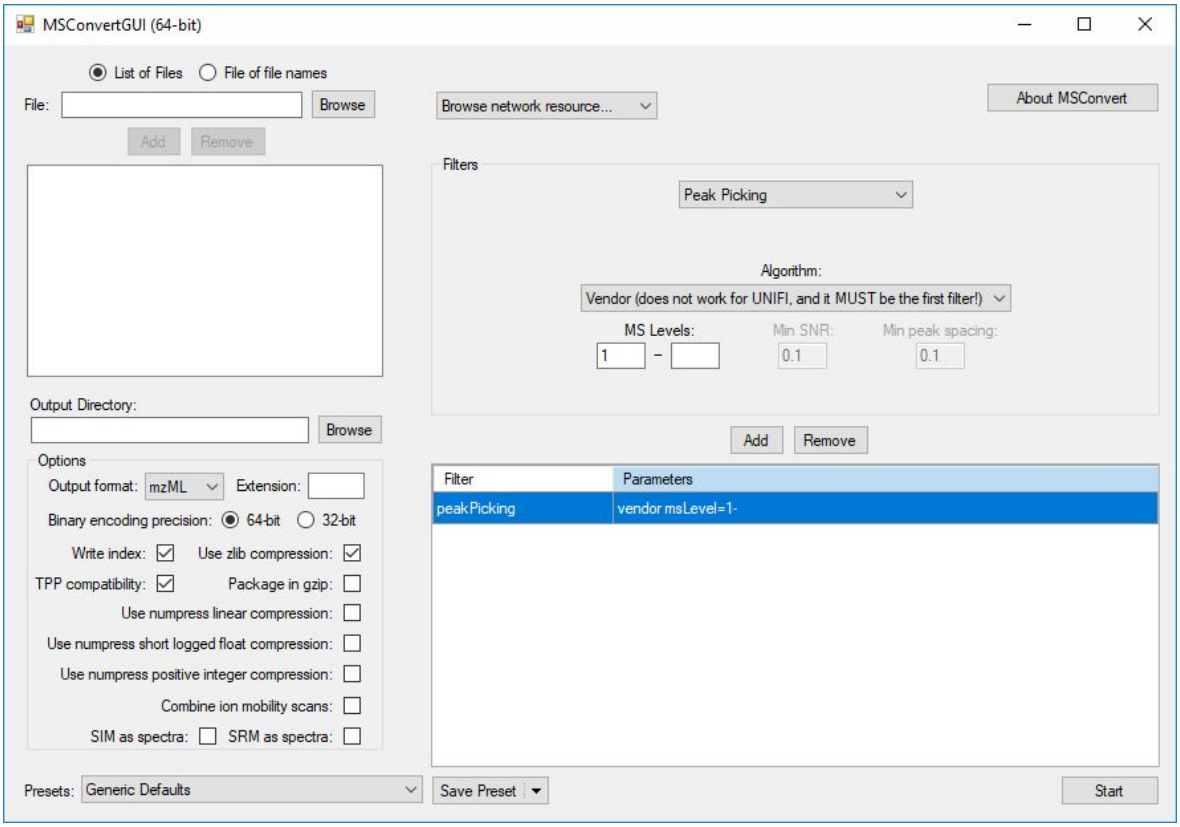
Appendix: Visualization options

A. Skyline
B. Bri-line
C. Viewing elib files DB Browser for SQLite

GUI-BASED WORKFLOW

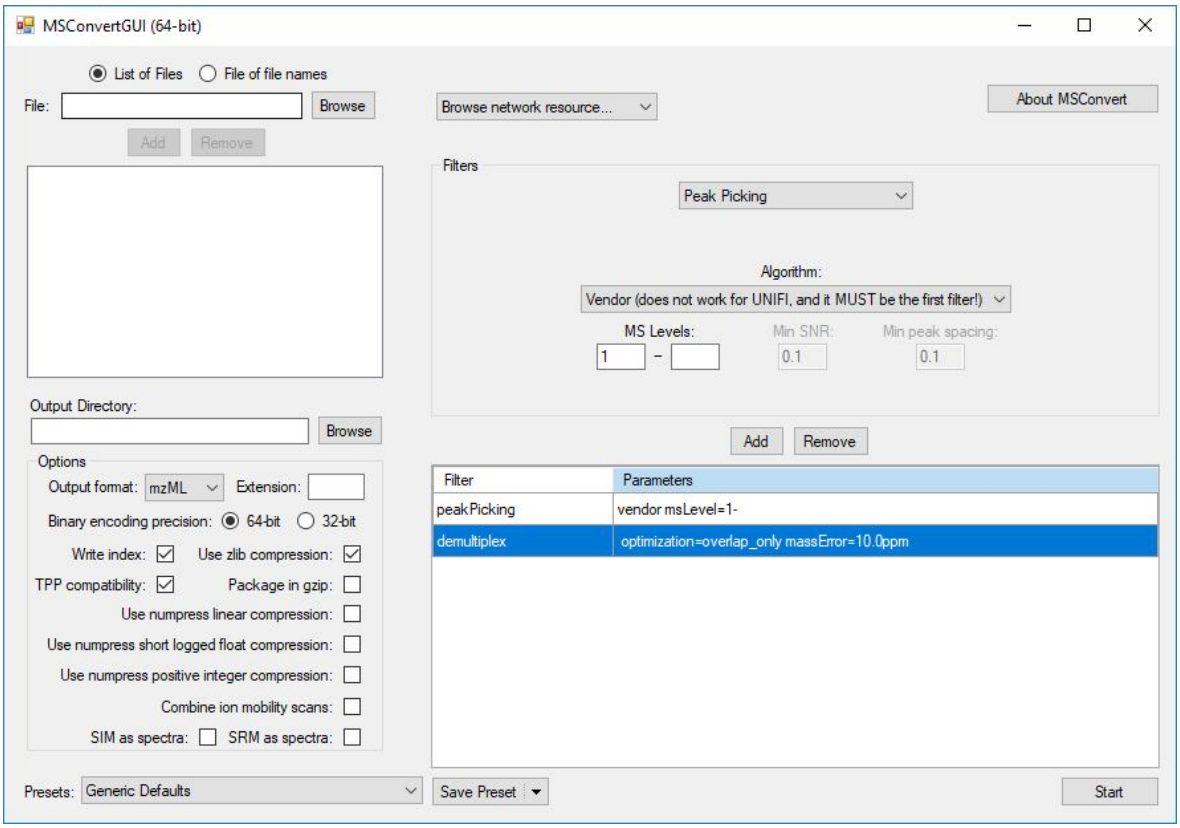
1. Convert .raw files to .mzML using MSConvert

For non-overlapping windows:



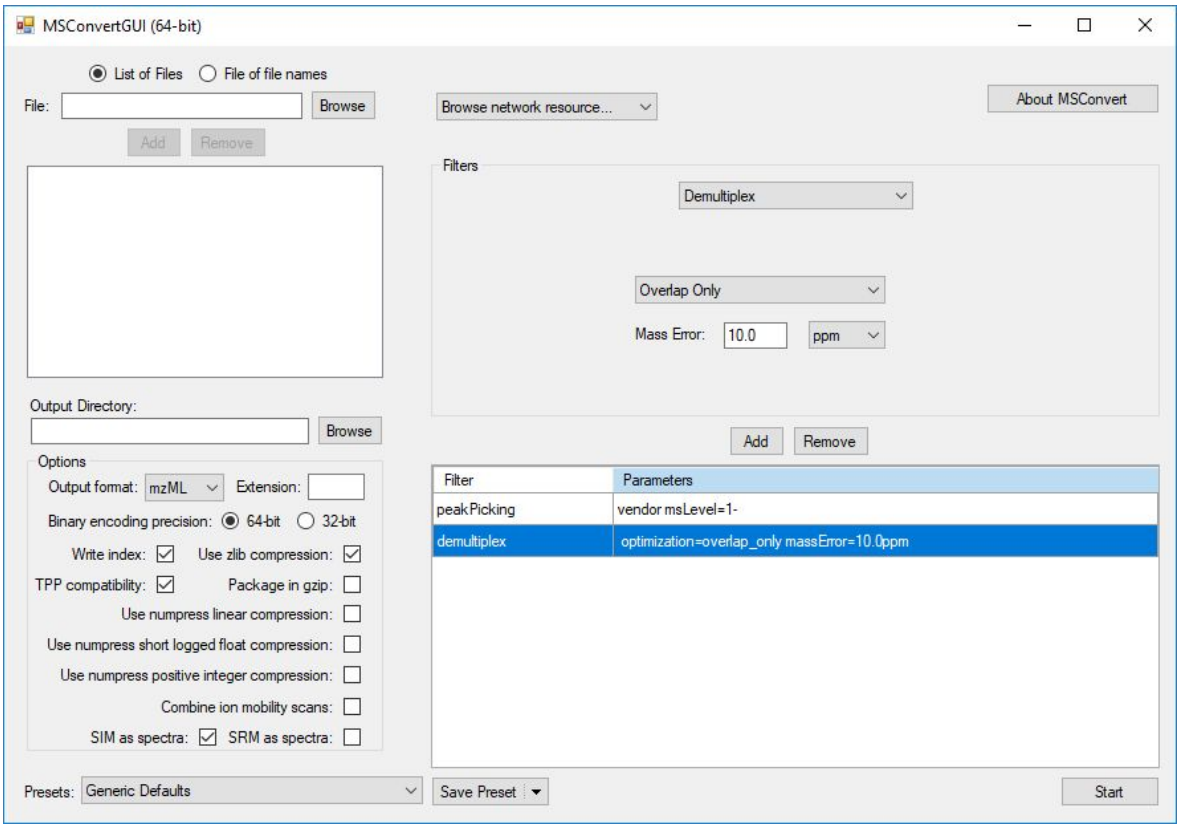
! NOTE: Make sure to have "peakPicking" as the first filter

For overlapping windows:



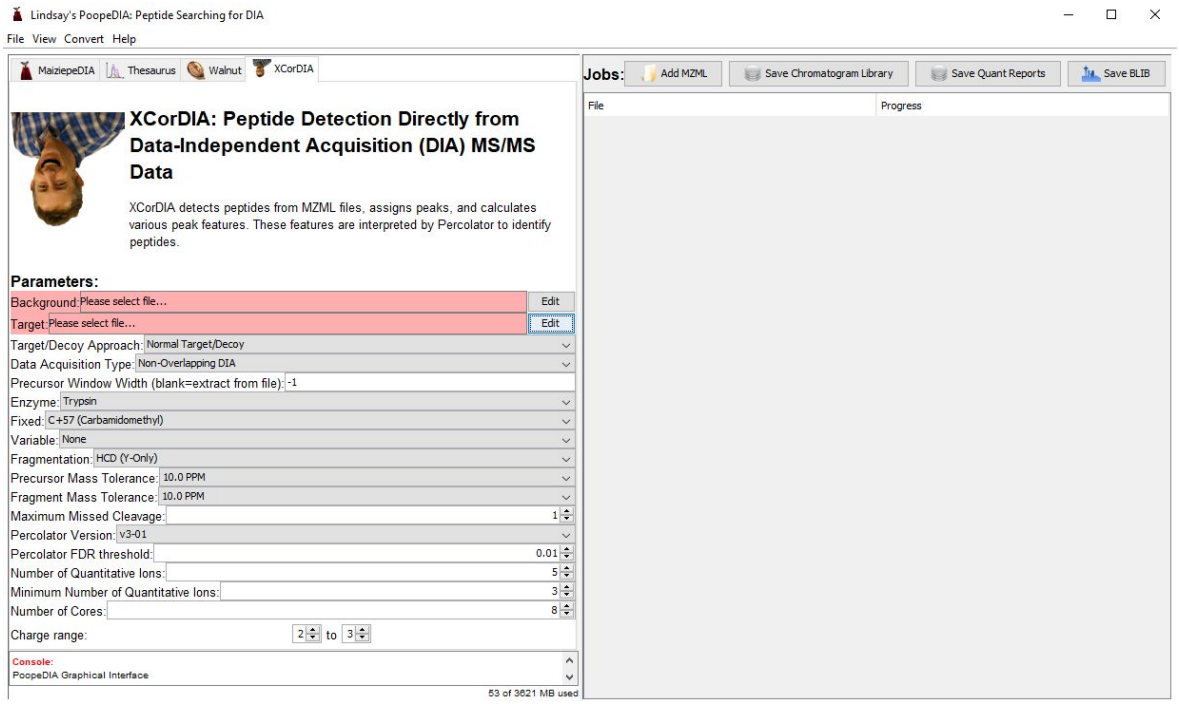
NOTE: data acquired on Lumos instruments may need another box checked (SIM as spectra) to convert precursor scans correctly. If so, your MSConvert should look like:

> *This step probably depends on whether you acquired an MS1 as a SIM scan or just as a mere MS1 scan in setting up the acquisition method on the Lumos. Since you can acquire a mere MS1 scan using the quad to filter (basically filters things outside of the scan range) I think it's basically the same as a SIM. It's kind of like how the method editor has a DIA method and a targeted MS2 method.*
>
> *-- Rich Johnson*

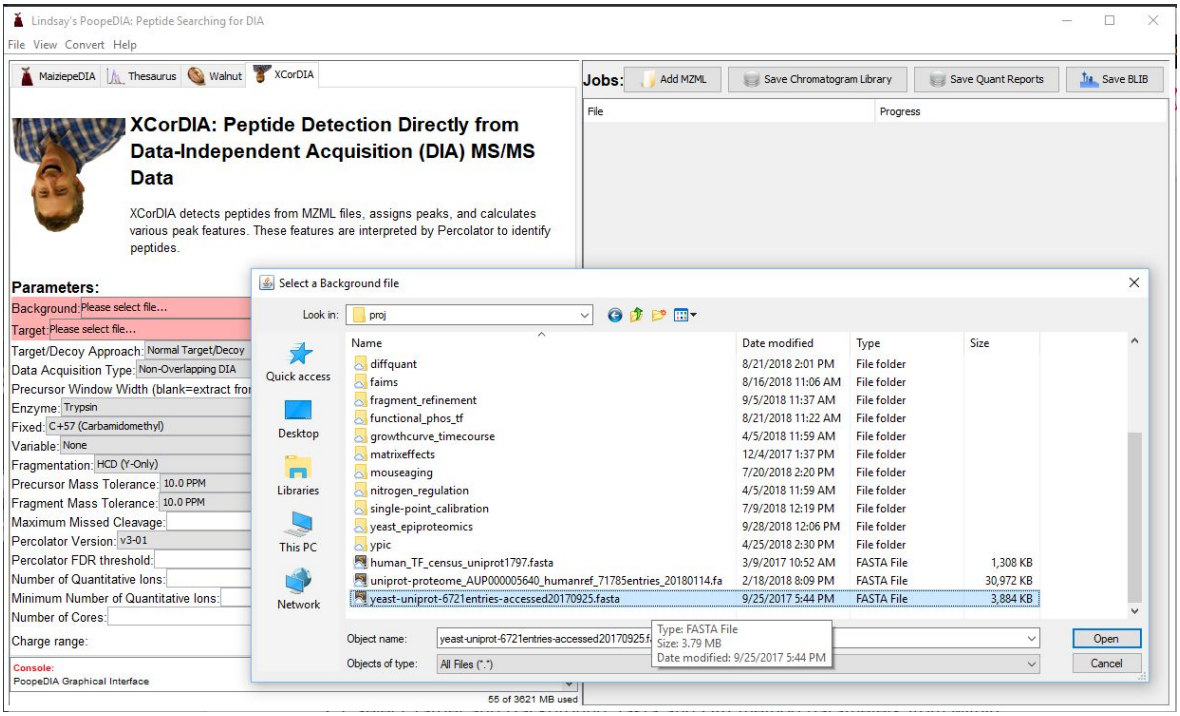


2. Build library using Walnut in EncyclopeDIA or XCorDIA in EncyclopeDIA

If you don't have access to the XCorDIA tab yet, you can use Walnut for this step!



2.1 On the left hand side right underneath The Upside Down Mike, where it says **Parameters**, find the "**Background**" field and click the corresponding "**Edit**" button to select the background fasta file. The background fasta file should basically be the reference fasta for your model system (*E.coli*, yeast, human, whatever). Here, I'll use an example experiment in yeast, so I downloaded the yeast reference from Uniprot, and navigated the file explorer to that downloaded fasta file.

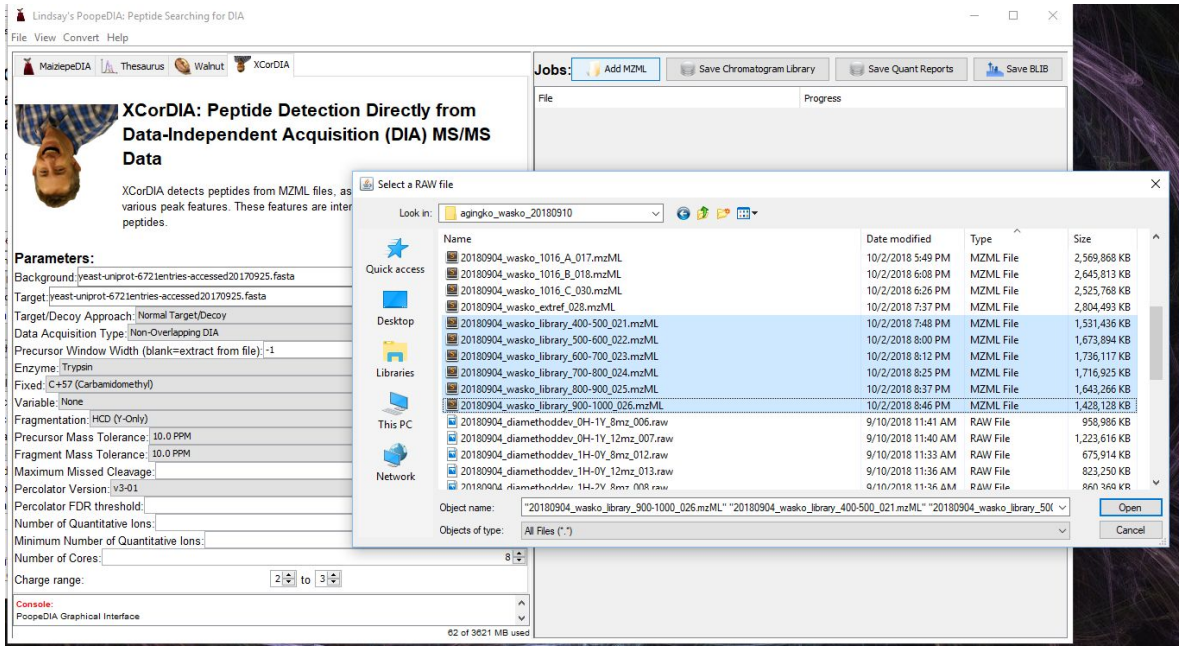


2.2 Again under "**Parameters:**", just underneath the "**Background**" field, which should now contain the filename of that organism fasta, click the "**Target**" field corresponding "**Edit**" button. Navigate to a fasta of your target search proteome. Here, I'm interested in the whole proteome (no specific subcellular fraction like mitochondria) so I'll select the same fasta that I used in the Background field.

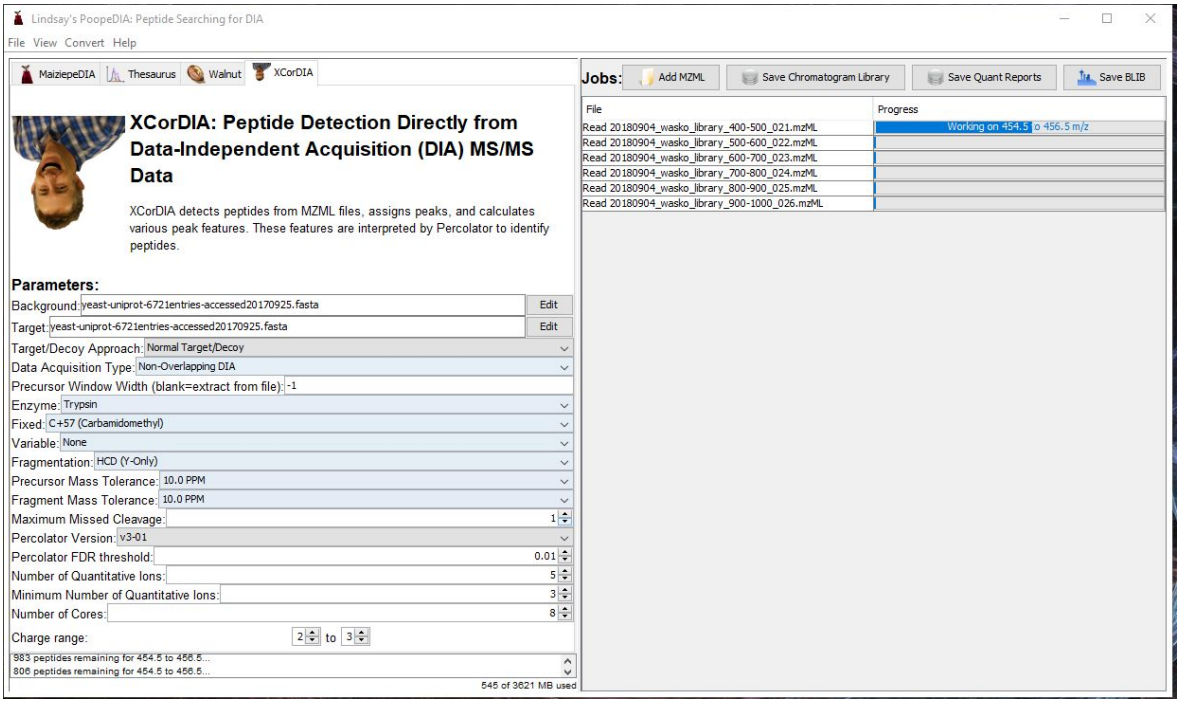
> ***!** More about the Target/Background fasta: For experiments looking at a "whole proteome" (lysates, for example), both the Target and the Background fasta are the same file (the yeast reference fasta, human reference fasta, etc). For experiments where subcellular fractionation was performed or where you're only interested in some subset of the proteome, use a "Background" fasta of the whole organism and a "Target" fasta just of the proteins you're interested in (for example, a mitochondrial isolation might use the human proteome for a Background fasta, and a MitoCarta fasta that only includes mitochondrial proteins).*
>
> ***!** Both Background and Target files are .fasta format. PeCAn users may recall processing a fasta to get a list of peptides for input, but Walnut includes the insilico digest step so you can just give Walnut the .fasta*

2.3 Set the remaining parameters if you have experiment-specific details that deviate from the defaults (for example, a different digestion enzyme than the default trypsin, or a different fragmentation type than HCD, etc)

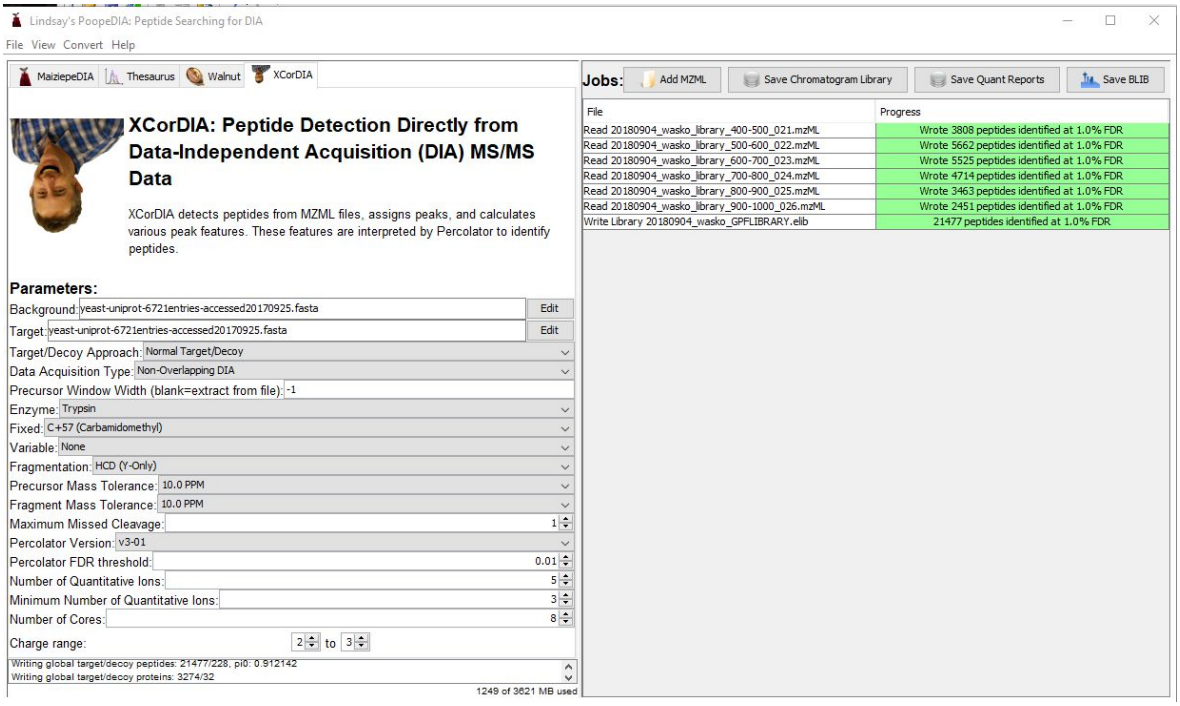
2.4 In the top right, where it says “**Jobs:**”, click “**Add MZML**”. Navigate the file explorer to your converted gas phase fractionated library files from step 1. Select all the gas phase fractionated library MZMLs and click “**Open**”.



The MZML files you selected should now appear under the “**Jobs**” buttons. You can monitor progress using the GUI.

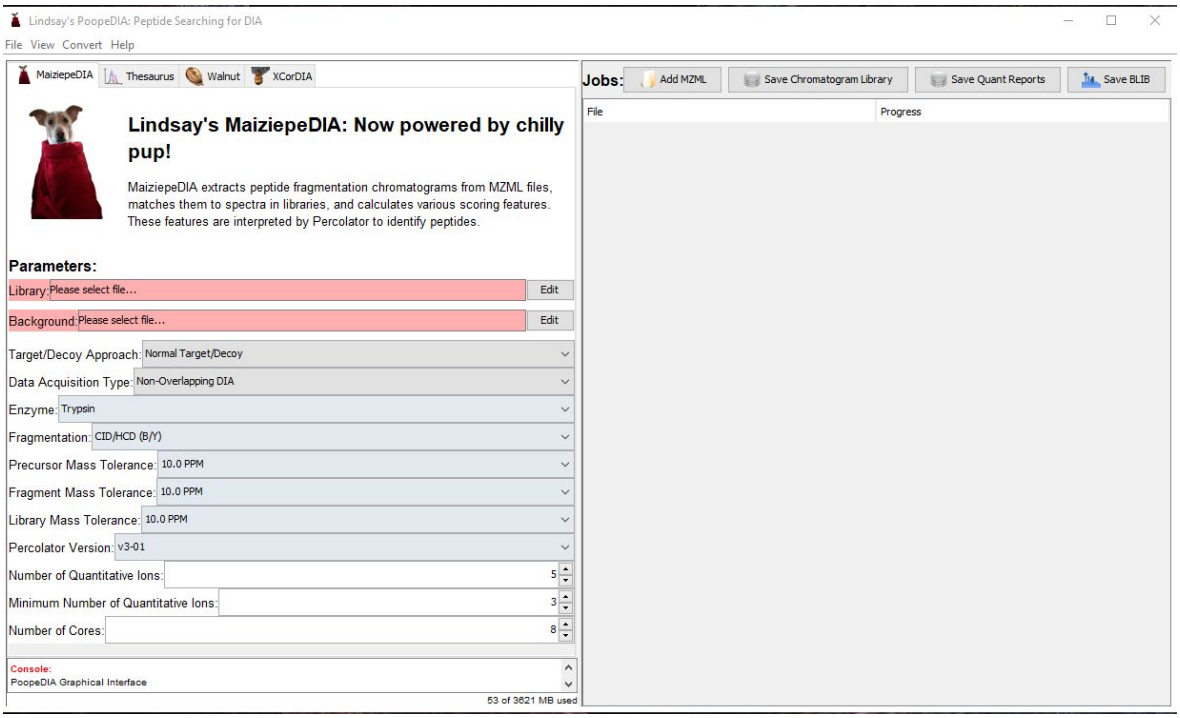


2.5 When the six gas phase fractionated files have finished running, click "**SAVE CHROMATOGRAM LIBRARY**" and give your library some descriptive filename.



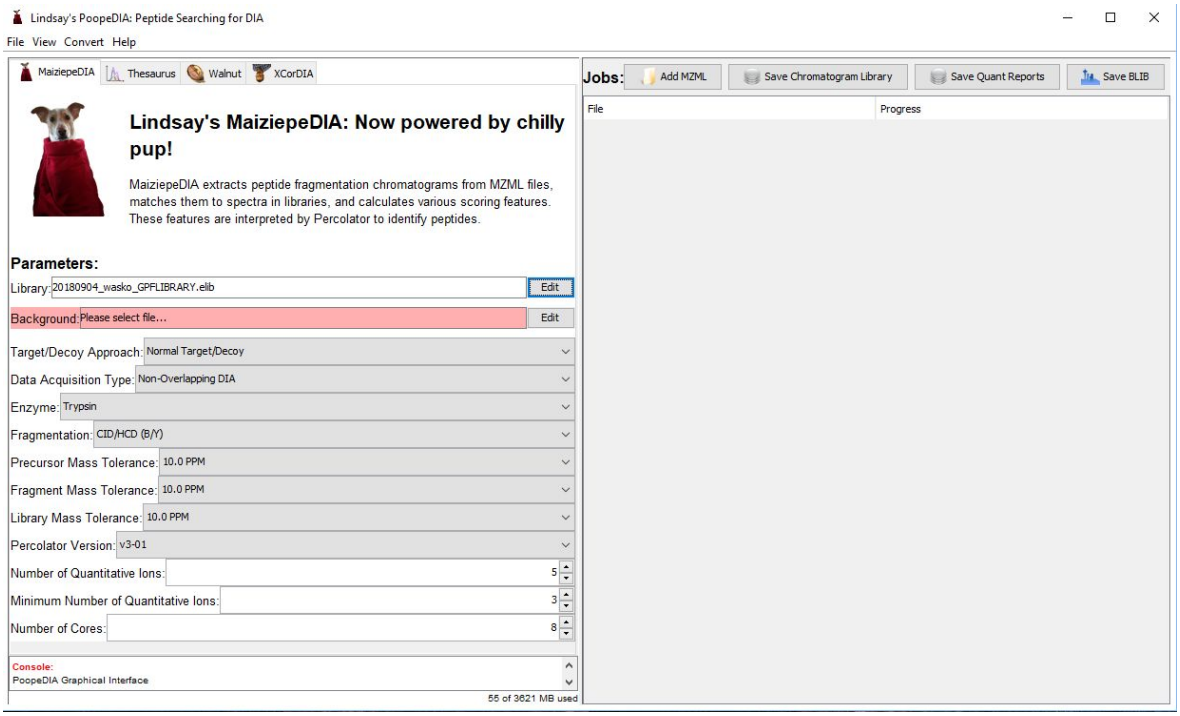
3. Search wide-window data with library from step 2 using EncyclopeDIA

3.1 Close and reopen the EncyclopeDIA GUI to clear EncyclopeDIA's cache/history

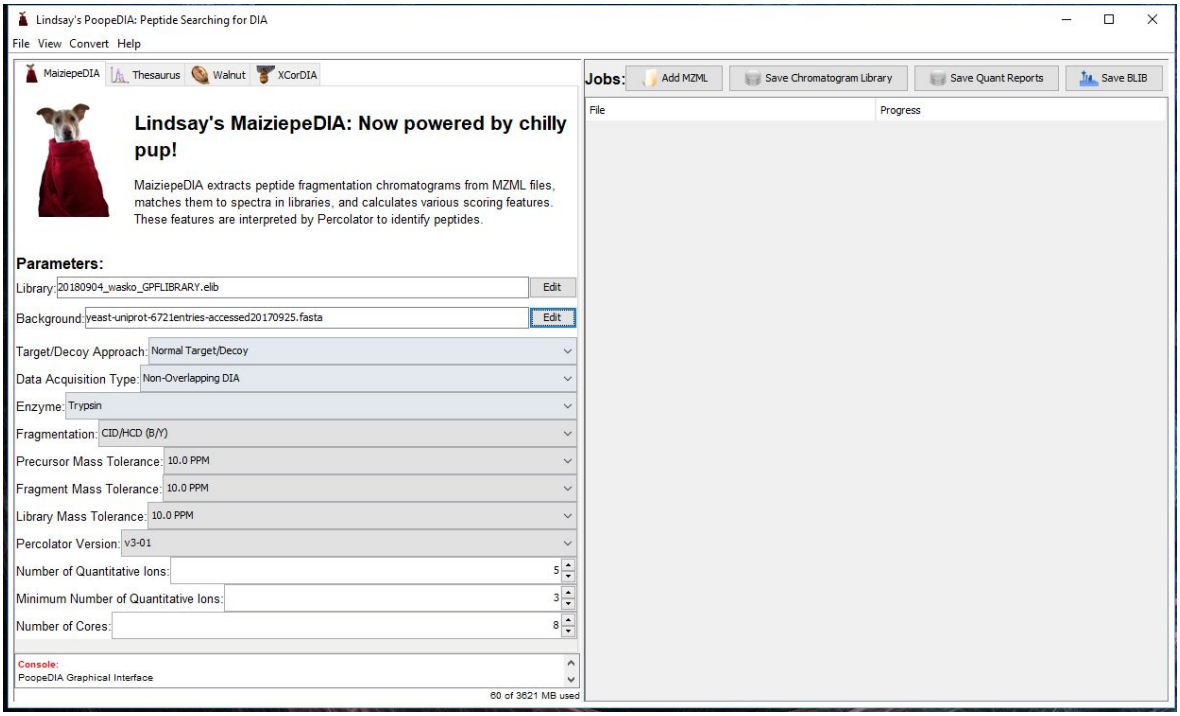


! Yours won't have this beauteous icon, it'll have some plebeian book.

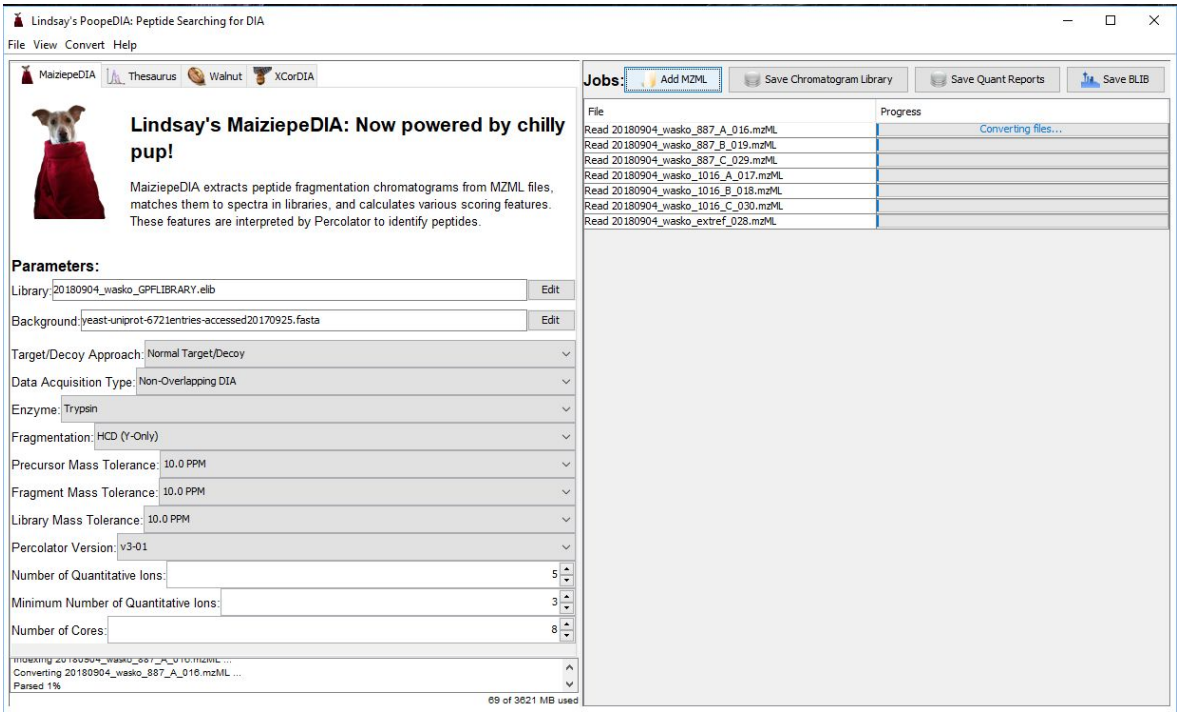
3.2 Within EncyclopeDIA GUI (not XCorDIA), on the left hand side under "Parameters:" across from the "Library" field, click the "Edit" button. Using the file explorer, select the .elib file you just saved in Step 2.2



3.3 Underneath the “Library” field, across from the “Background” field, click the corresponding “Edit” button and select the appropriate Background file (should be the same fasta you used above!)

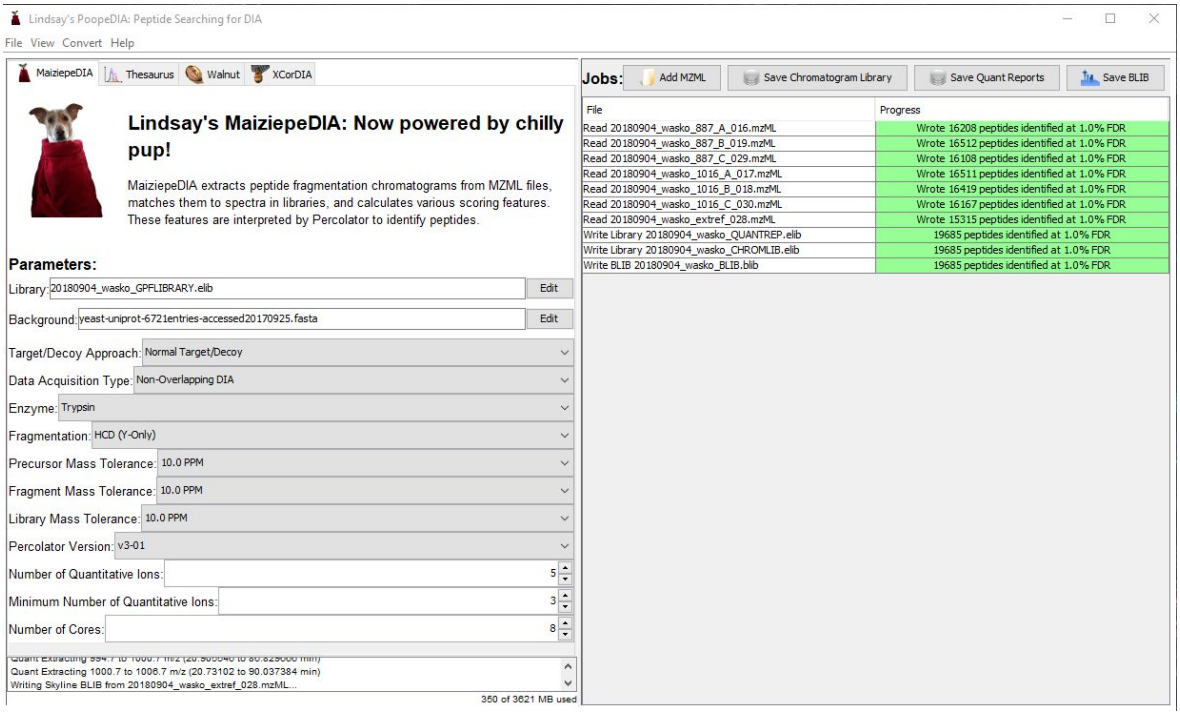


3.4 In the top right, next to “Jobs”, click the “Add MZML” button and select all of the wide-window .mzML files that were acquired using this narrow-window library file.



3.5 Click “**Save Quant Reports”** to perform a final experiment-wide FDR correction and export peptide quant, and protein quant.

- Select “Save Chromatogram Library” to build a file with bonus information like integration boundaries.
    - This is only applicable if you do not want to retention time-align across the MZML files (for example, if your MZMLs are fractionated in a way such that you don’t expect to sample the same peptides in each file)
- Select “Save Quant Reports” to get peptide/protein quantitation matrices in the form of a tsv.
    - Pick this if your MZMLs were the experimental samples you want to post-process.
    - If you are following this workflow as-is, this is what you should pick!
- Select “SAVE BLIB” to build a spectral library file that Skyline can use.
    - This option is effectively depreciated now that Skyline reads elib file formats



! Encyclopedia is determining a lot of information about the DIA experiment. Important details include peptide detections, fragment refinement, and peak

boundaries. We'll import that information into Skyline next so that we can visualize the results; however, there's also a visualizer built right into Encyclopedia. See Appendix for details.

! Encyclopedia outputs (blib, elib) can be used in Skyline for visualizing DIA-MS experiments. See Appendix for details.

COMMAND LINE WORKFLOW

1. Convert .raw files into .mzml using MSConvert

For non-overlapping windows:

```
msconvert.exe -v --zlib --64 --mzML --filter "peakPicking true
1-" *.raw
```

For overlapping windows:

```
msconvert.exe -v --zlib --64 --mzML --filter "peakPicking true
1-" --filter "demultiplex optimization=overlap_only" *.raw
```

NOTE: data acquired on Lumos instruments needs an extra flag (`--simAsSpectra`) to convert precursor scans correctly:

```
msconvert.exe --zlib --64 --mzML --filter "peakPicking true 1-"
--filter "demultiplex optimization=overlap_only" --simAsSpectra
*.raw
```

> *! Tip: If your file conversion is going slow, you probably aren't using a current version of MSconvert! Go back to the top and start again :)*

2. Build library using Walnut in EncyclopeDIA or XCorDIA in EncyclopeDIA (command line)

2a. OPTION 1: Walnut

For a full list of options and default values for your version of EncyclopeDIA:

```
java -jar encyclopedia.jar -walnut --help
```

The parameters you may want to change include the enzyme used to prepare the samples,

```
$ java -Xmx<GB_OF_MEM>G -jar encyclopedia.jar -walnut \
  -i <MZML_IN> \
  -f <BACKGROUND_FASTA> \
  -t <TARGET_FASTA> \
  -acquisition DIA \
```

```
  -enzyme <ENZYME> \
  -frag <FRAGMENTATION> \
  -ftol <FRAGMENT_TOLERANCE> \
  -ftolunits <FRAGMENT_TOLERANCE_UNITS> \
  -ptol <PRECURSOR_TOLERANCE> \
  -ptolunits <PRECURSOR_TOLERANCE_UNITS> \
  -minCharge <MIN_CHARGE> \
  -maxCharge <MAX_CHARGE>
```

In the MacCoss lab, with our typical DIA set up (trypsin digest, Orbitrap instrument, demultiplexing the RAW file overlapping windows with MSConvert), the command usually looks like this:

```
$ java -Xmx8G -jar encyclopedia.jar -walnut \
  -i DIA_narrow_run_400to500mz.mzML \
  -f human.fasta \
  -t human.fasta \
  -acquisition DIA \
  -enzyme trypsin \
  -frag YONLY \
  -ftol 10.0 \
  -ftolunits ppm \
  -ptol 10.0 \
  -ptolunits ppm \
  -minCharge 2 \
  -maxCharge 3
```

Running Walnut produces several results files, including:

```
<MZML_IN>.dia
<MZML_IN>.mzML.pecan.txt.log
<MZML_IN>.mzML.features.txt
<MZML_IN>.mzML.pecan.txt
<MZML_IN>.mzML.pecan.decoy.txt
```

2a. OPTION 2: XCORDIA

The command to run XCorDIA on the narrow-window acquisition files is the same as Walnut (above), but replace `-walnut` with `-xcordia.`

The output files from XCorDIA will differ slightly :

```
<MZML_IN>.dia
<MZML_IN>.mzML.elib
<MZML_IN>.mzML.xcordia.txt.log
<MZML_IN>.mzML.features.txt
<MZML_IN>.mzML.xcordia.txt
<MZML_IN>.mzML.xcordia.decoy.txt
```

2b. Merge Results into a Single Library File (.elib)

This command must be run within the same directory as the search command from 2.b.i. No matter if you used Walnut or XCorDIA to search the files! You have to be in the same directory because Encyclopedia will look for the Walnut/XCordia result files in the current directory in order to compile them into one elib, so if the result files aren't there, Encyclopedia has nothing to compile.

For a full list of options and default values for your version of EncyclopeDIA:

```
java -jar encyclopedia.jar -libexport --help
```

`OUTPUT_LIBRARY_NAME:` (filename) Any name of your choice. Entering **`narrow_merged`** here would result in a final output of `narrow_merged.elib` (or `narrow_merged.blib`).
`ALIGN_SPECTRA?:` (**`true`** or **`false`**) You will likely want to set this to **`false`** for this step as you're probably doing a narrow isolation gas phase fractionation. In general, if each of your mzML acquisitions collect an identical precursor range, this should be **`true`** -- otherwise it should be **`false`**.
`USE_BLIB_FLAG:` To export an elib, leave this blank. To export a .blib, set to: **`-blib`**

Typical run:

```
$ java -Xmx<GB_OF_MEM>G -jar encyclopedia.jar -libexport \
  -i <INPUT_DIRECTORY> \
  -o <OUTPUT_LIBRARY_NAME> \
  -a <ALIGN_SPECTRA?> \
  <USE_BLIB_FLAG> \
  -f <BACKGROUND_FASTA> \
  -t <TARGET_FASTA> \
  -ftol <FRAGMENT_TOLERANCE> \
  -ftolunits <FRAGMENT_TOLERANCE_UNITS>
```

Example:

```
$ java -Xmx8G -jar encyclopedia.jar -libexport \
  -i ./ \
  -o narrow_merged \
  -a false \
  -f human.fasta \
  -t human.fasta \
  -ftol 10.0 \
  -ftolunits ppm
```

3. Search wide-window data with library from step 2 using EncyclopeDIA (command line)

3.i. Search Each mzML

For a full list of options and default values for your version of EncyclopeDIA:

```
java -jar encyclopedia.jar --help
```

LIBRARY_ELIB_FILE: This should be the result from your narrow library search in step 2.b.ii, e.g. **path/to/your/narrow_merged.elib**

Typical run:

```
$ java -Xmx<GB_OF_MEM>G -jar encyclopedia.jar \
  -l <LIBRARY_ELIB_FILE> \
  -i <MZML_IN> \
  -f <BACKGROUND_FASTA> \
  -t <TARGET_FASTA> \
  -acquisition DIA \
  -enzyme <ENZYME> \
  -frag <FRAGMENTATION> \
  -ftol <FRAGMENT_TOLERANCE> \
  -ftolunits <FRAGMENT_TOLERANCE_UNITS> \
  -ptol <PRECURSOR_TOLERANCE> \
  -ptolunits <PRECURSOR_TOLERANCE_UNITS> \
  -minCharge <MIN_CHARGE> \
  -maxCharge <MAX_CHARGE>
```

Example:

TODO

Output files:

```
<MZML_IN>.dia
<MZML_IN>.mzML.elib
<MZML_IN>.mzML.encyclopedia.txt
<MZML_IN>.mzML.encyclopedia.txt.delta_rt.pdf
<MZML_IN>.mzML.encyclopedia.txt.log
<MZML_IN>.mzML.encyclopedia.txt.rt_fit.pdf
<MZML_IN>.mzML.encyclopedia.txt.rt_fit.txt
<MZML_IN>.mzML.features.txt
<MZML_IN>.mzML.first_round.txt
```

3.ii Merge Results into a Single Library File (.blib)

***!** Note: You most likely do not need to create a .blib anymore, because Skyline can read elib files.*

This command must be run within the same directory as the search command from 3.b.i. The parameters used should be similar to those used in step 2.b.ii, with the following exception:

`ALIGN_SPECTRA?:` (**`true`** or **`false`**) You will likely want to set this to **`true`** for this step as you're probably searching replicates with the same precursor mass range.

Typical run:

```
$ java -Xmx<GB_OF_MEM>G -jar encyclopedia.jar -libexport \
  -i <INPUT_DIRECTORY> \
  -o <OUTPUT_LIBRARY_NAME> \
  -a <ALIGN_SPECTRA?> \
  <USE_BLIB_FLAG> \
  -f <BACKGROUND_FASTA> \
  -t <TARGET_FASTA> \
  -ftol <FRAGMENT_TOLERANCE> \
  -ftolunits <FRAGMENT_TOLERANCE_UNITS>
```

Example:

```
$ java -Xmx8G -jar encyclopedia.jar -libexport \
  -i ./ \
```

```
        -o merged \
        -a true \
        -f human.fasta \
        -t human.fasta \
        -ftol 10.0 \
        -ftolunits ppm
```

DOCKER-POWERED COMMAND LINE WORKFLOW

Docker allows for running applications without having to worry about downloading or installing the software you'd like to run. Once docker is installed, it handles all of this automatically. See here for instructions on installing Docker: TODO

This workflow is identical to the COMMAND LINE WORKFLOW above, with the following modifications:

For msconvert, instead of using

```
msconvert.exe --help
```

You can use

```
docker run -it --rm \
  -v "C:\Users\your_username\path\to\your\data:/data" \
  chambm/pwiz-skyline-i-agree-to-the-vendor-licenses:3.0.19073-85be84641 \
  wine msconvert --help
```

And docker will download and install msconvert for you if needed and then run your command.

Similarly, for encyclopedia:

```
java -Xmx<GB_OF_MEM>G -jar encyclopedia.jar --help
```

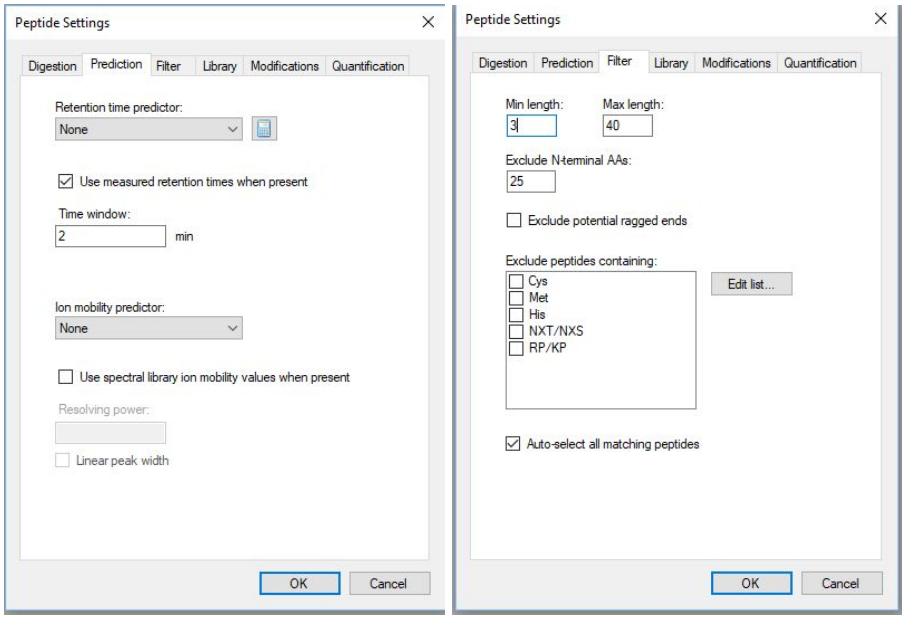
```

becomes

```
docker run -it --rm \
  -v “C:\Users\your_username\path\to\your\data:/data” \
  -e ‘JAVA_OPTS=-Xmx<GB_OF_MEM>G’ \
  atkeller/encyclopedia:v0.8.1_cv2 \
  encyclopedia --help
```

## Giving Docker Access to Your Data

Docker can’t take paths to your data directly, so you need to give it access using the -v flag.

If your data is under… TODO [Finish describing mounting volumes]
```

Appendix

Skyline.

! Tip: You can find a Skyline template with steps 4a-4h pre-cooked in this directory. You'll need Skyline-daily to use it, I think, but you can pick up at step 4i.

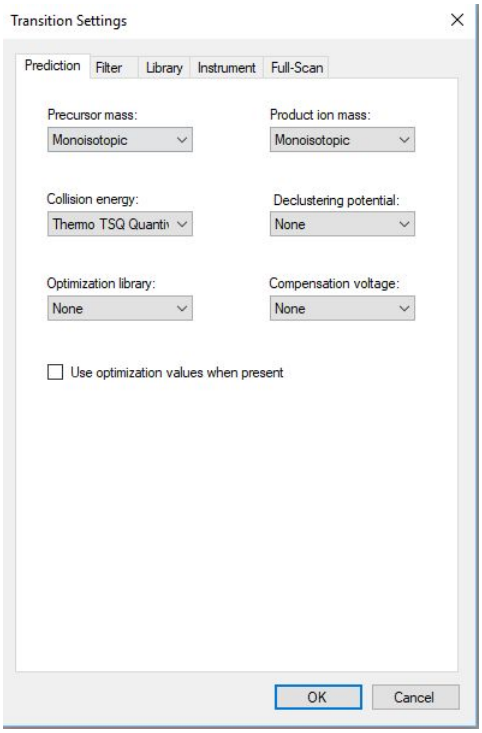
Settings > Peptide Settings.

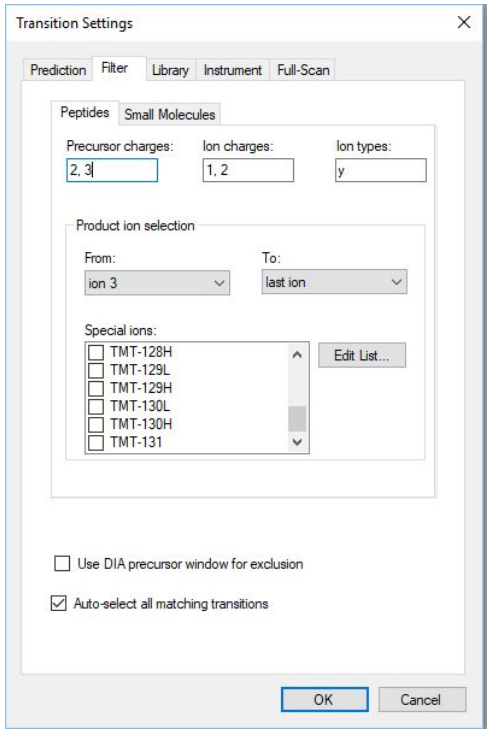


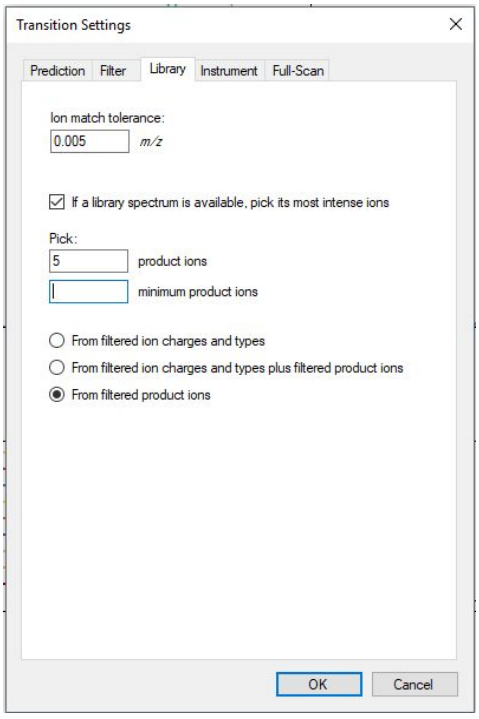
4a. Prediction: Check "Use measured retention times when present", Time window=2

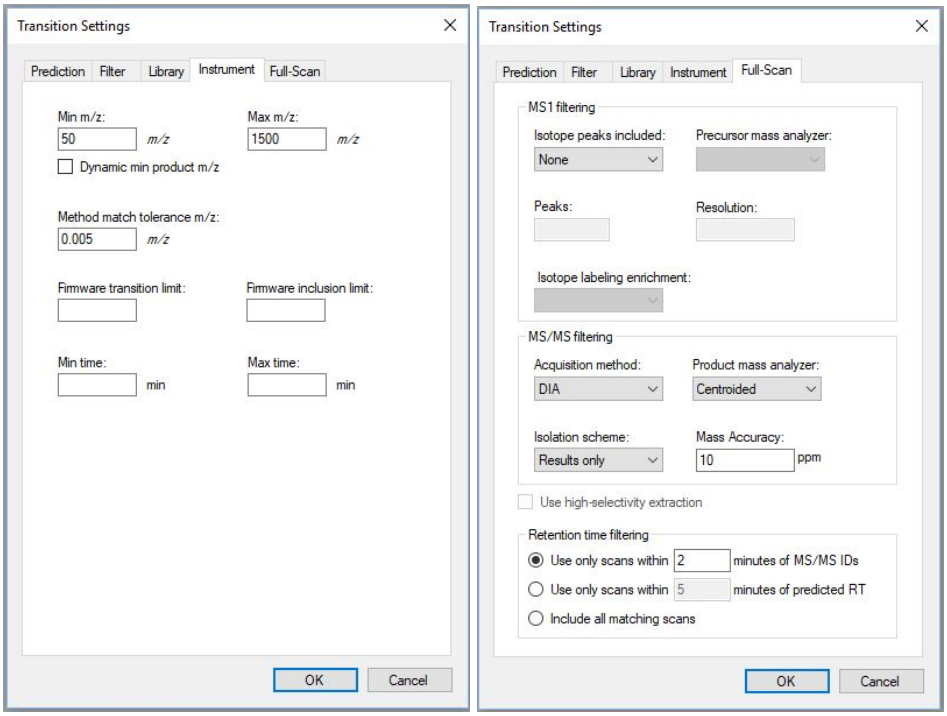
4b. Filter: Min length=3, Max length=25*, no excluded amino acids checked (*I sometimes increase the Max length to 40)

Settings > Transition Settings.


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Precursor mass:
Monoisotopic
Product ion mass:
Monoisotopic
Collision energy:
Thermo TSQ Quantiv
Declustering potential:
None
Optimization library:
None
Compensation voltage:
None
Use optimization values when present
OK
Cancel


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Peptides
Small Molecules
Precursor charges:
2, 3
Ion charges:
1, 2
Ion types:
y
Product ion selection
From:
ion 3
To:
last ion
Special ions:
TMT-128H
TMT-129L
TMT-129H
TMT-130L
TMT-130H
TMT-131
Edit List...
Use DIA precursor window for exclusion
Auto-select all matching transitions
OK
Cancel


Transition Settings
Prediction
Filter
Library
Instrument
Full-Scan
Ion match tolerance:
0.005
m/z
If a library spectrum is available, pick its most intense ions
Pick:
5
product ions
minimum product ions
From filtered ion charges and types
From filtered ion charges and types plus filtered product ions
From filtered product ions
OK
Cancel



4c. Prediction: Precursor/Product ion mass=”Monoisotopic”

4d. Filter: Ion charges=1,2 Ion types=”y”*, From=Ion 3, To=last ion, no special ions

*Set “ion types” to reflect how you searched the data in Encyclopedia!

4e. Library: Ion match tolerance=0.005 m/z, check “If a library spectrum is available, pick its most intense ions”, pick=5 product ions, select “From filtered product ions”

** If you’re intending to use quantifications from Skyline and skipping MSstats, you may want to consider setting the Pick: *3 minimum product ions*

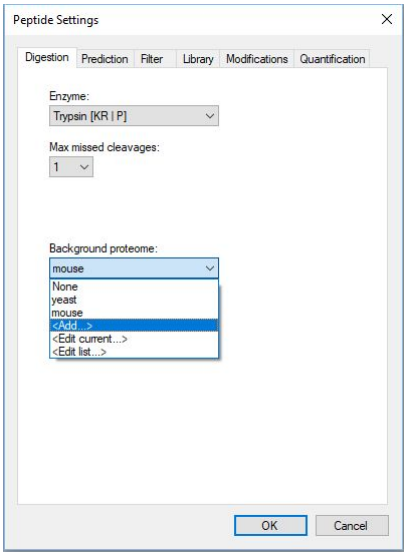
4f. Instrument: “Min m/z=50, Max m/z=1500, Method match tolerance m/z=0.005

4g. Full-Scan (MS1): “Isotope peaks included=None”

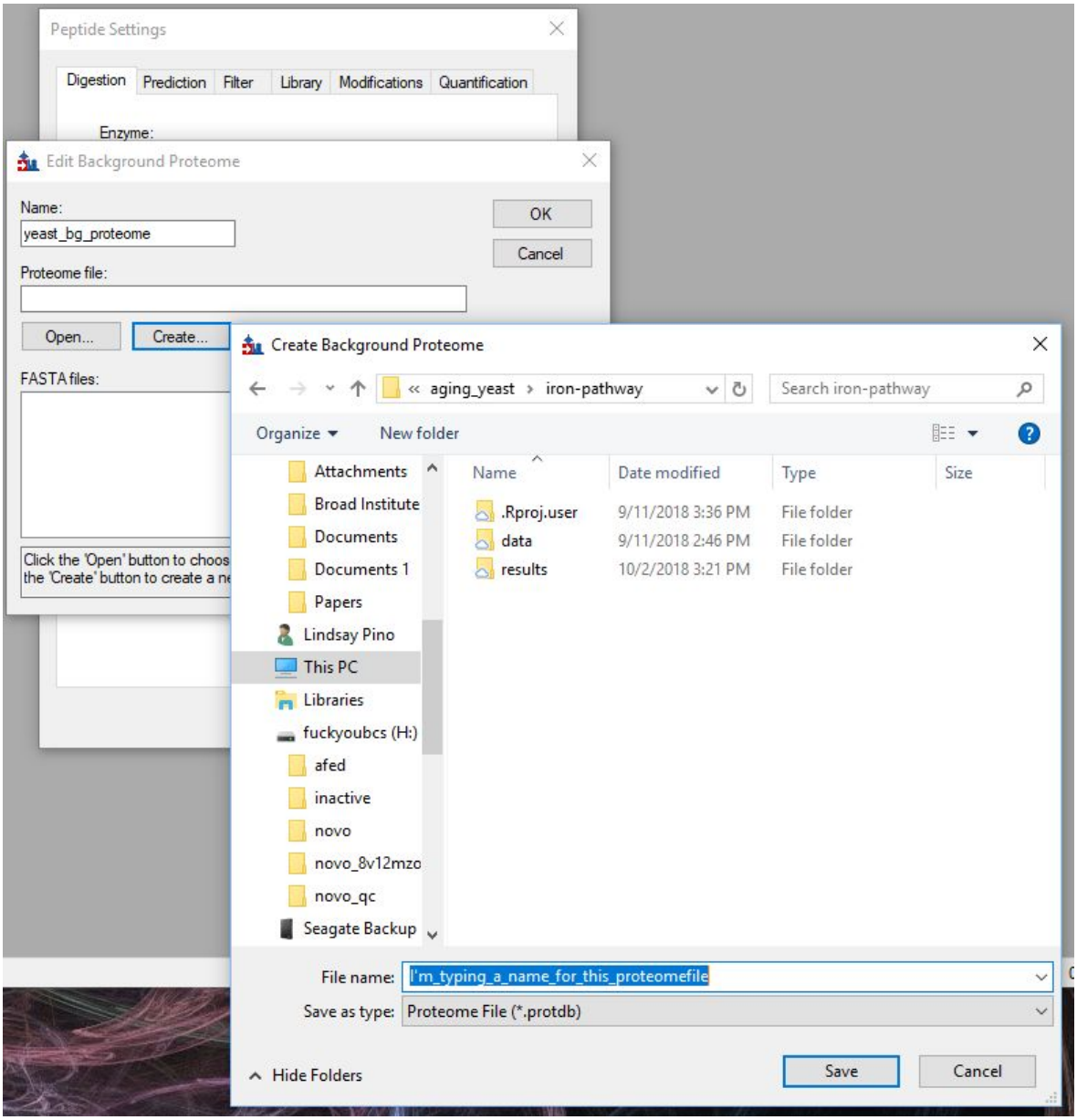
4h. Full-Scan (MS/MS): “Acquisition method=DIA, Product mass analyzer=Centroided, Isolation scheme=Results only, Mass Accuracy=10 ppm, check “Use only scans within “2” minutes of MS/MS IDs

Load the FASTA database and the ELIB library from EncyclopeDIA into Skyline.

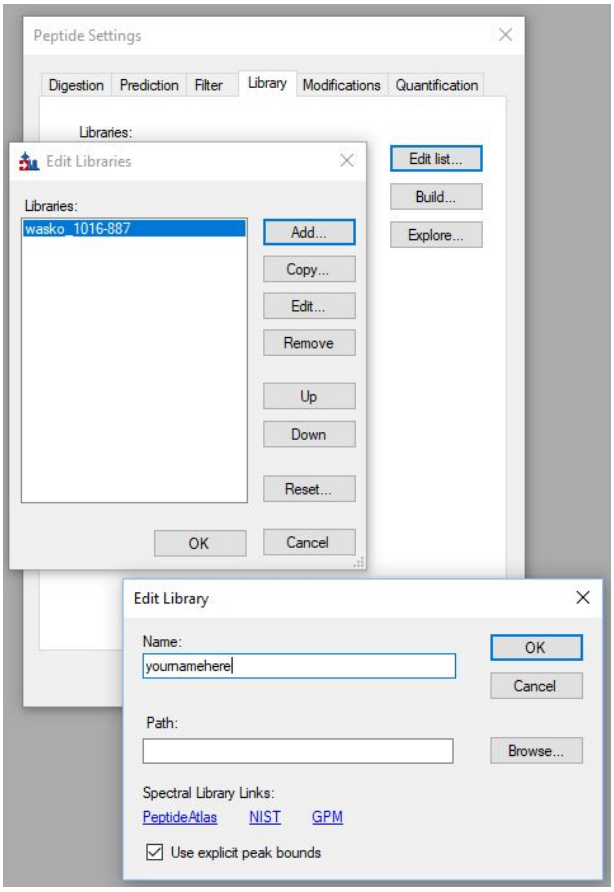
4i. Open Settings/Peptide Settings/Digestion. Select “<Add...>” for Background proteome



4j. Set Name=”[yournamehere]”, then click “Create…” and type in a descriptive name for the soon-to-be-created Proteome File and click “**Save**”. Select “Add File…” to add the background fasta you were just using in the Encyclopedia suite.



4k. Open Settings/Peptide Settings/Library. Select “Edit list…”, then “Add…”. Set the name to be “[yournamehere]” and select “Browse…” to select the .elib that you saved in EncyclopeDIA from the wide-window search (not the gas phase fractionated, narrow window library). “Ok” out to the Peptide Settings.



4l. Check the new “[yournamehere]” library and uncheck all other libraries. Click “Explore…” to view it

> ! If you get a pop up warning that ‘Peptide settings have been changed. Save changes?’ then just click “Yes”

4m. Check “Associate proteins” and click “Add All…”. Select “Add to all matching proteins” and “Include all peptides” and hit “OK”.

4n. If you have more than one library (*.elib) for your experiment, you must select each of those libraries from the dropdown menu in "Spectral Library Explorer" and repeat step 4m for each.

4o. Save the Skyline file!

4p. Import .mzML data into Skyline

> ***!*** *Tip: Use the .mzml files, not the .raw files! If you prefer using the .raw files for some reason, you will need to set the windowing scheme parameters in the "MS/MS filtering" box under the "Full Scan" tab in Skyline's "Transition Settings".*

TODO make a Skyline template file for the wide-window isolation window method.

Skyline: Export report for MSstats

To load the MSstats report format, install the MSstats external tool. Alternatively, you can quickly build a custom report using the little black binoculars to find whatever fields you're interested in exporting.

4q. File > Export > Report...

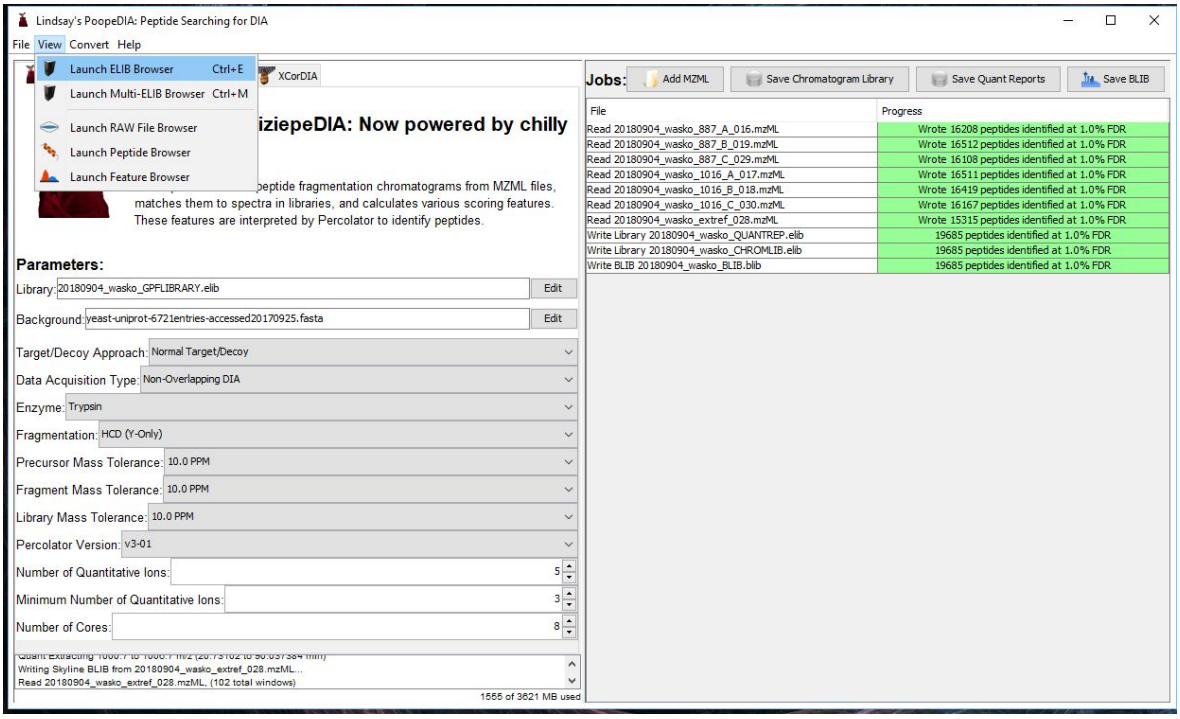
4r. Edit list... > Add...

4s. Make sure to name your report ("View Name:"), and check the fields you want the report to include. Binoculars are at the top left, next to the Redo arrow, and will find fields by name.

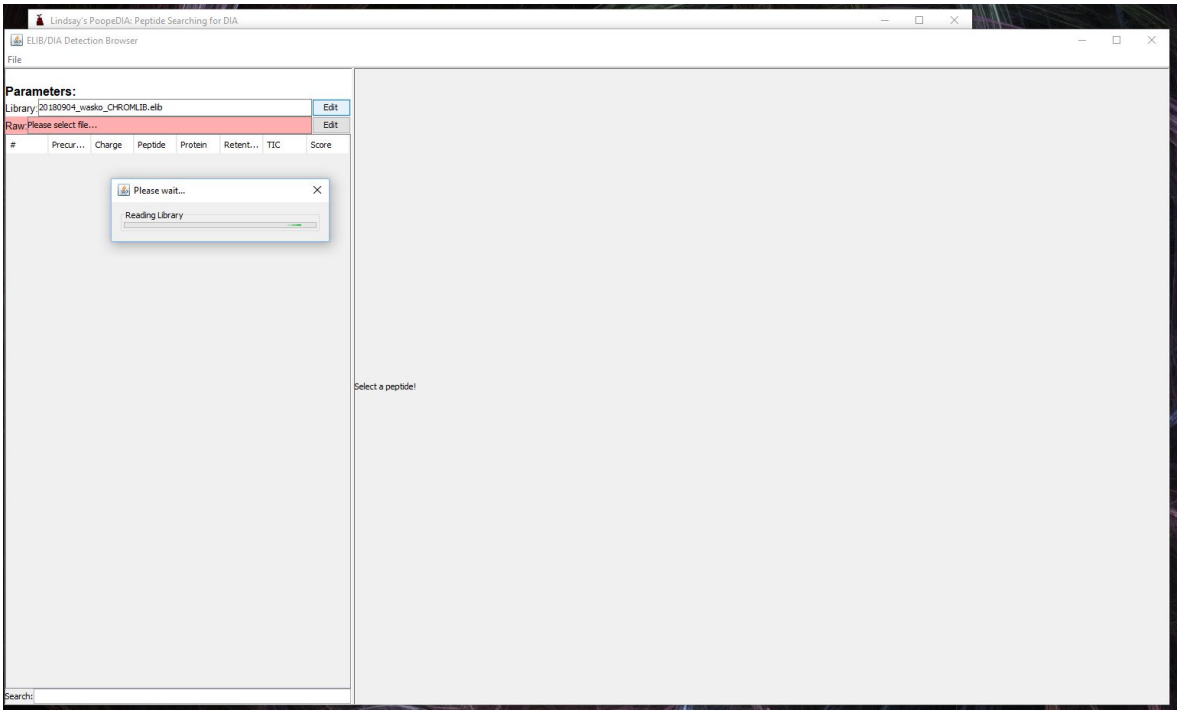
Bri-line.

Bri-Line elib Browser

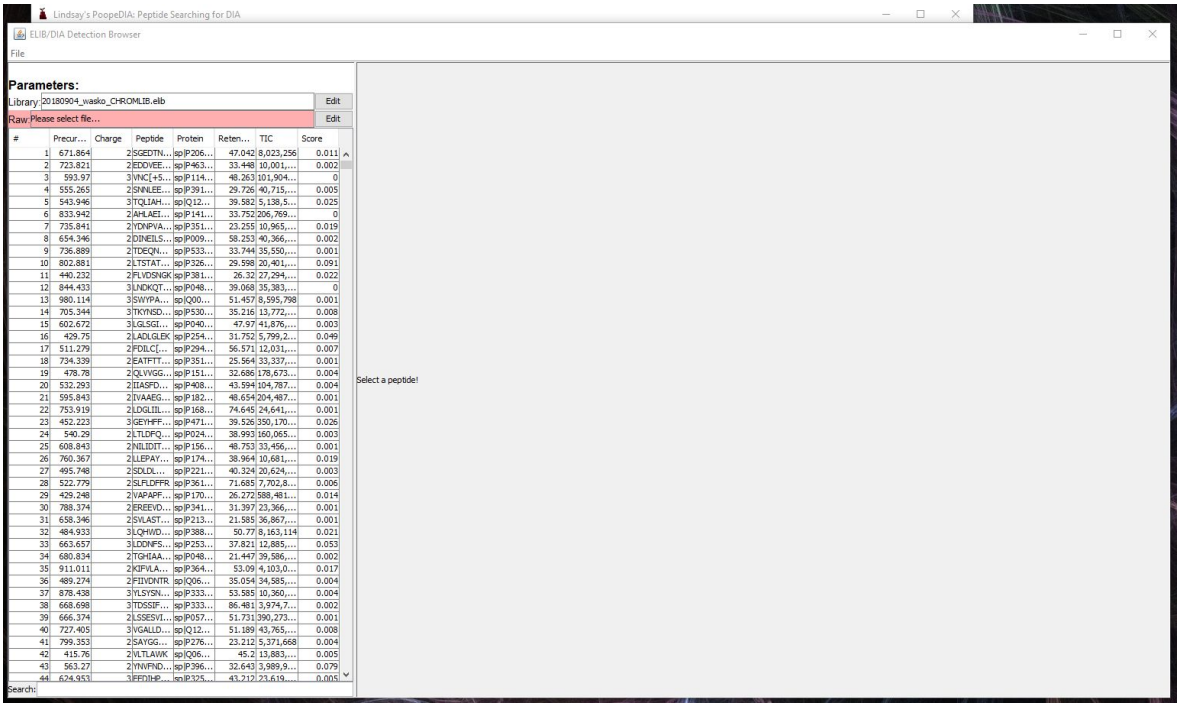
The top left under “View”, select the “Launch ELIB Browser” option



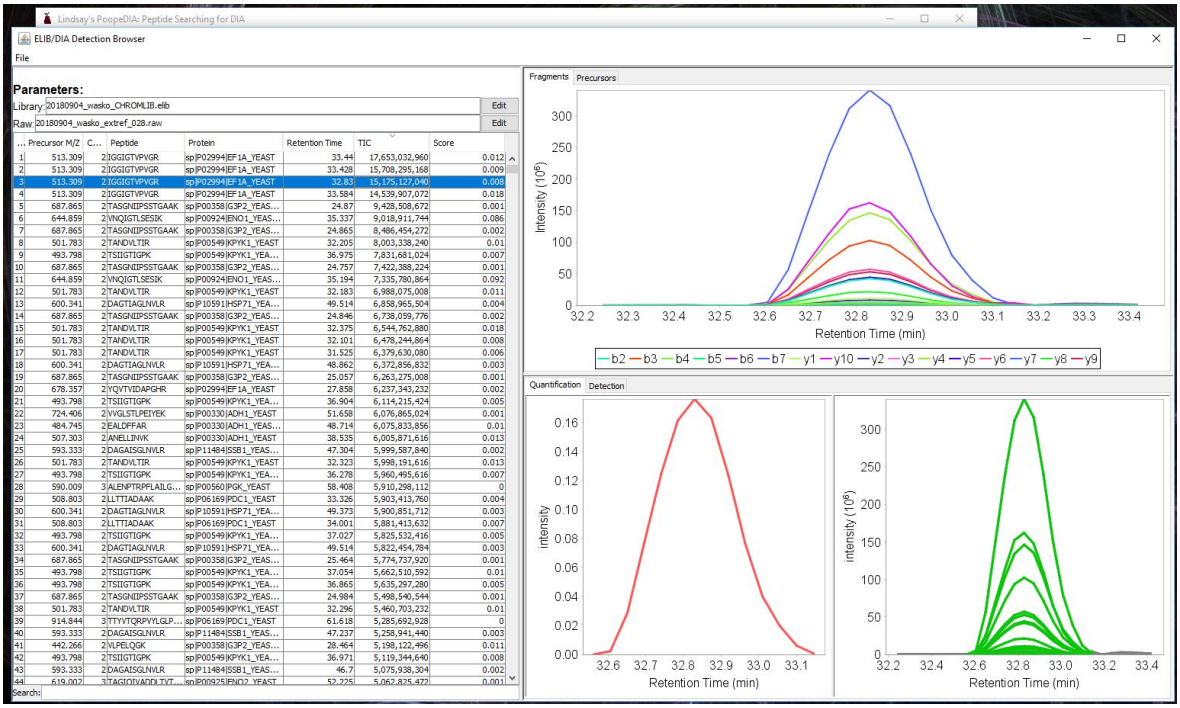
Next to the “Library” field, click the “Edit” button and navigate the explorer to the “Save Chromatogram Library” file you just saved.



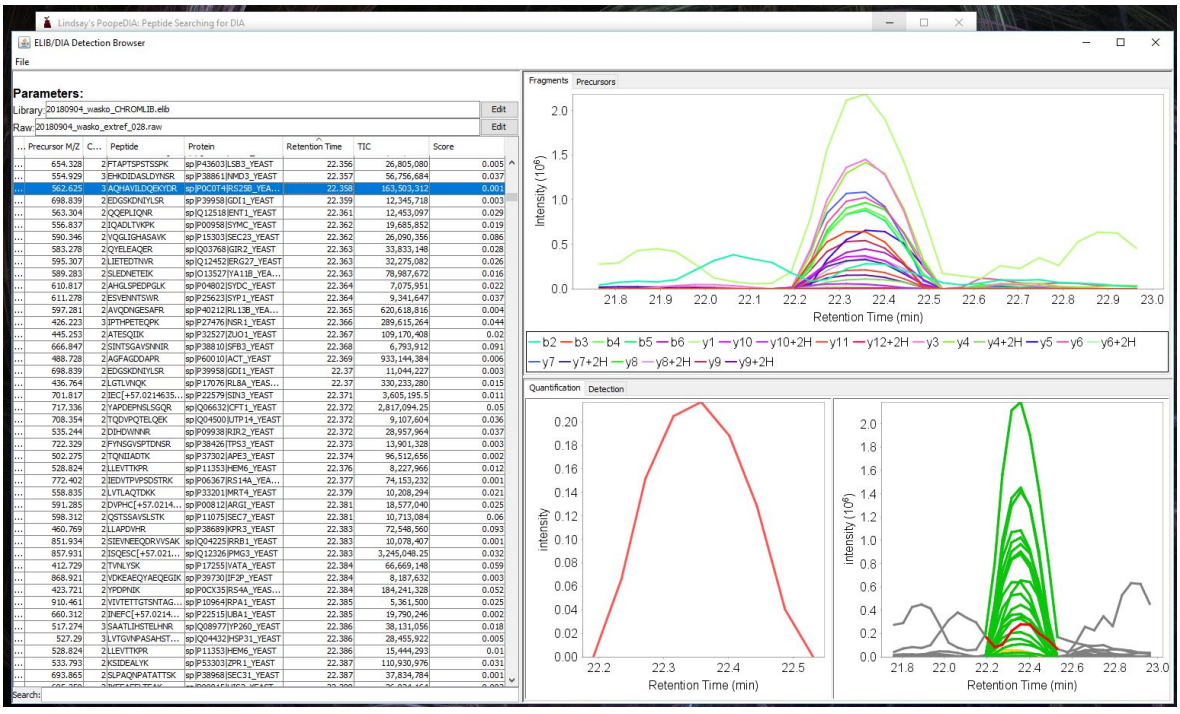
Once the library has loaded, you should have a populated target list like this:



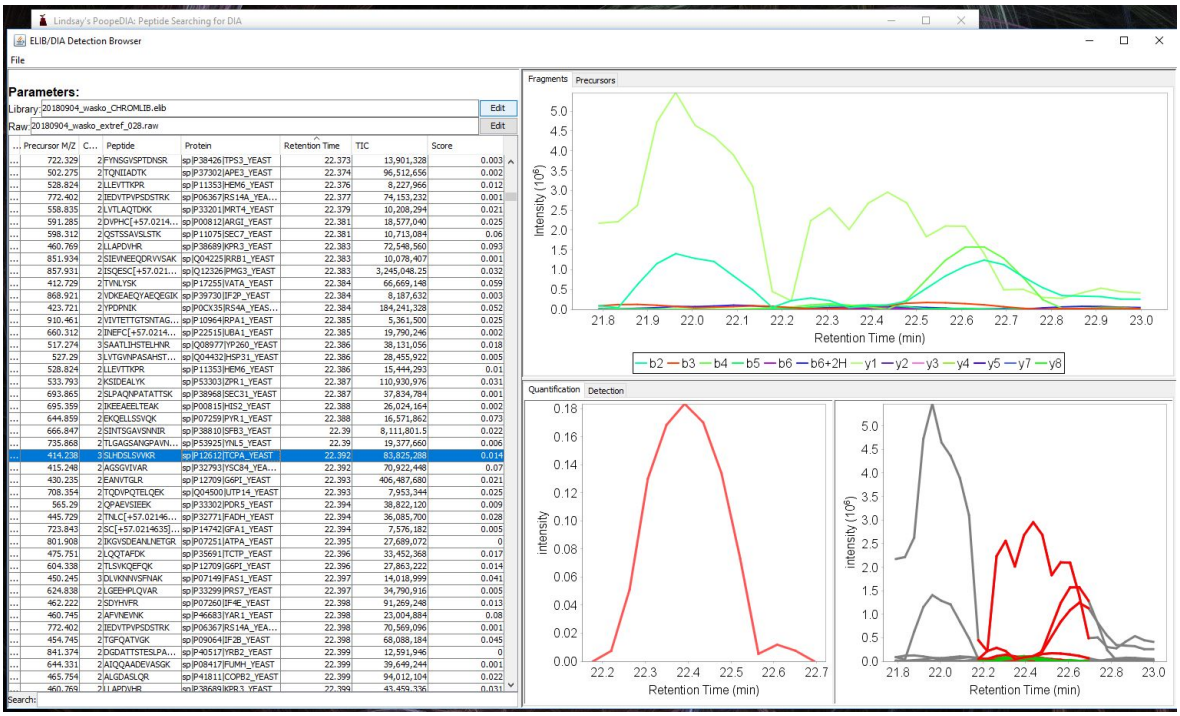
Next to the “Raw” field, click the “Edit” button to select one of the RAW files analyzed in that chromatogram library.



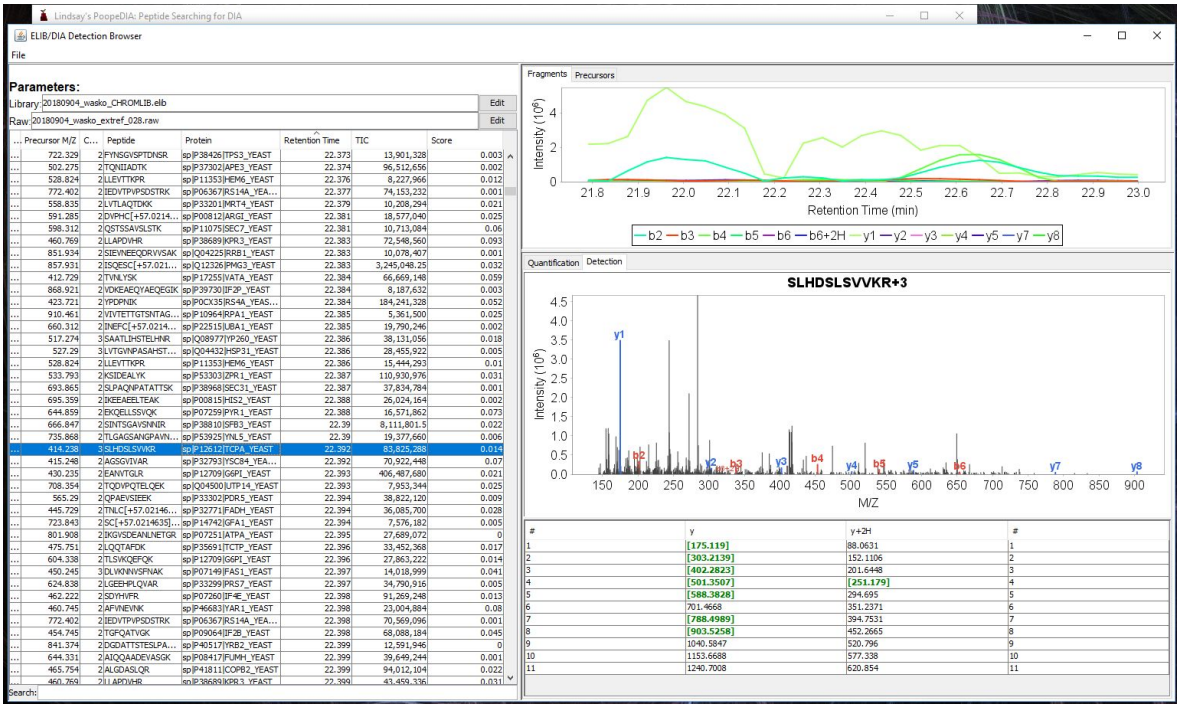
Above is an example of a “good” peptide, where Encylopedia found lots of interference-free fragments (all the fragment traces in the bottom right are green).



Above is an example illustrating Encyclopedia's fragment refinement. The bottom left pane shows a fragment ion chromatogram in red, which doesn't follow the average profile of the peak group.



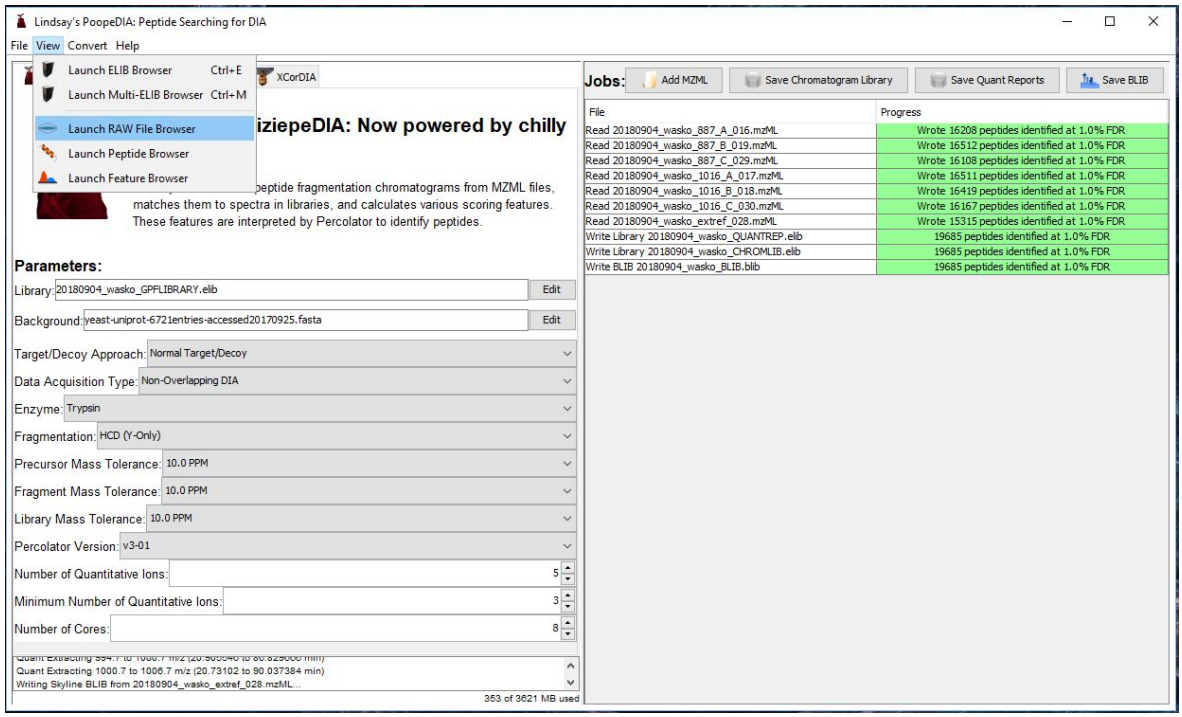
And finally, above is an example of peptide that was detected but doesn't look quantitative. There is no canonical chromatogram shape that looks good, but if we click on the "detection" tab in the middle of the three figures:



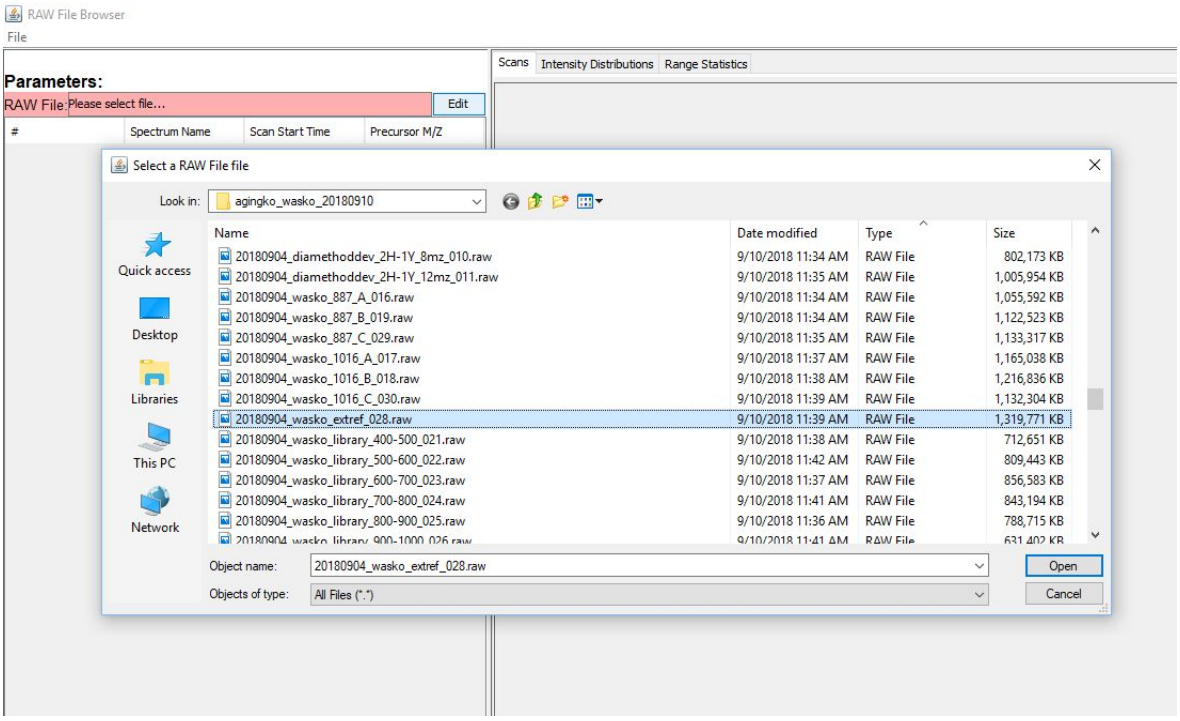
We can see that in the spectrum the detection was made, there are a reasonable number of fragments. This is just an example of how we can't quantify everything we detect!

Bri-Line RAW File Browser

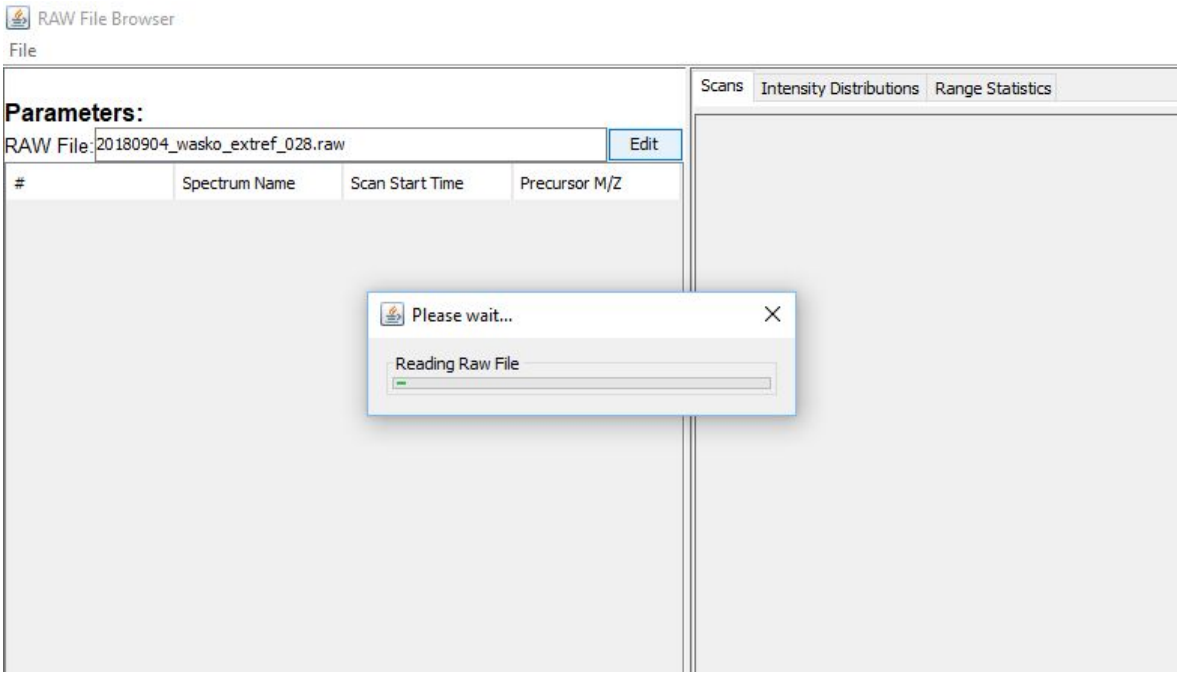
Top left under “View”, select “Launch RAW File Browser”



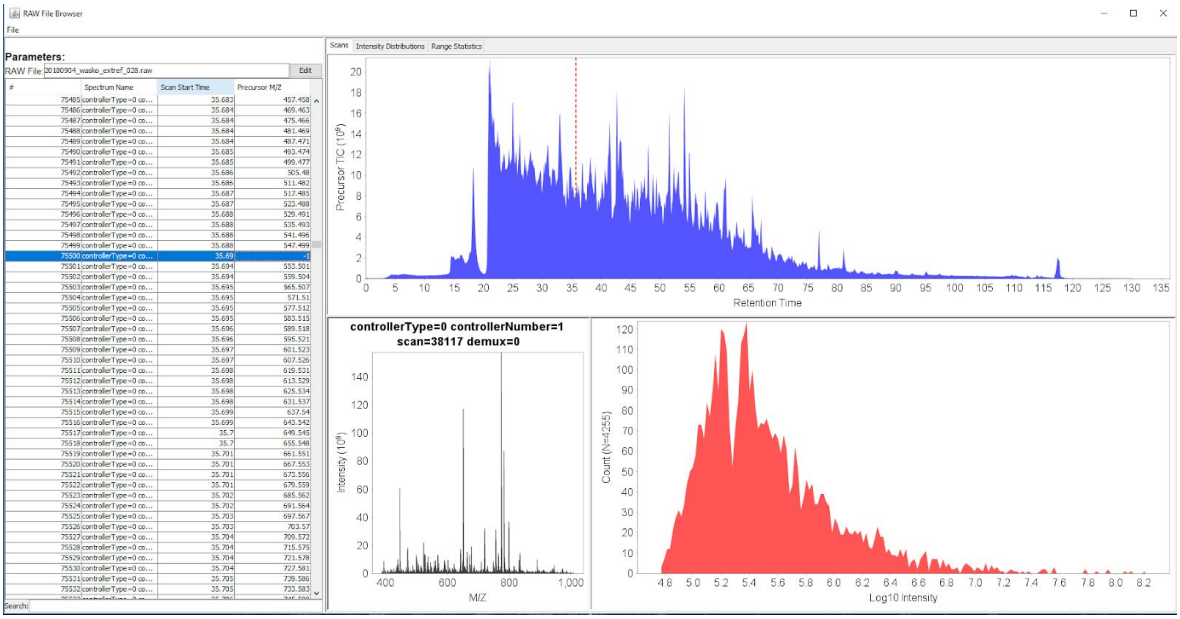
Select a “RAW” file that has been previously analyzed with the Encyclopedia suite (XCordia, Walnut, Encyclopedia, …)



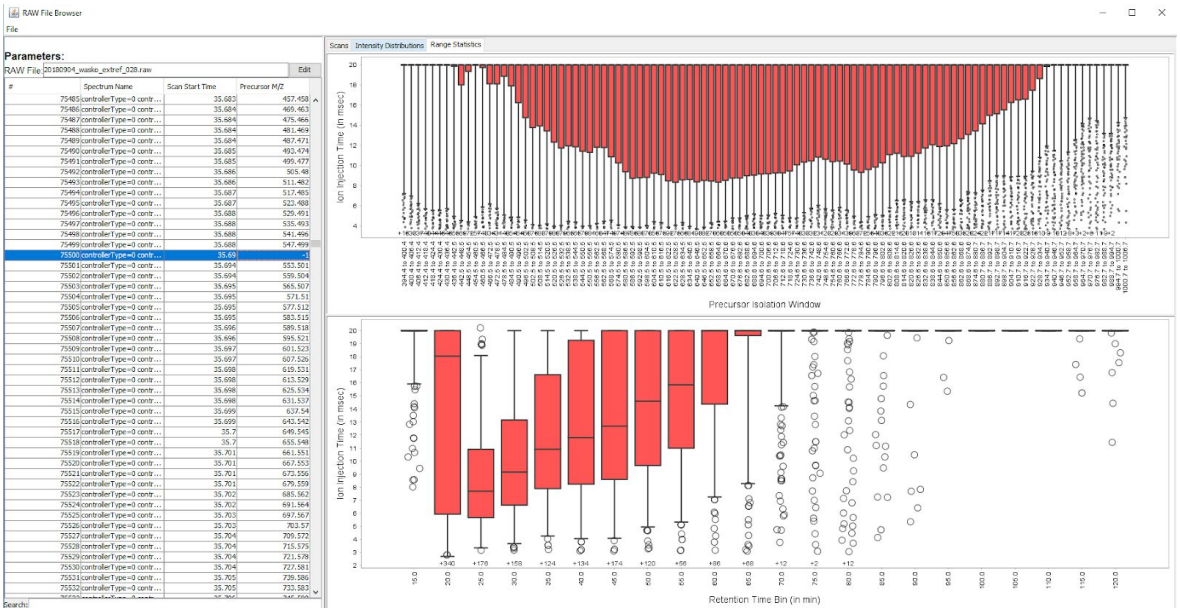
You should get a small pop up titled “Please wait…” with a “Reading Raw File” animation.



The target list on the left lists the Scan Number (#), SpectrumName, Scan Start Time, and Precursor M/Z.



On the top above the graphics, there are three tabs. The “Range Statistics” gives some valuable information about the DIA method like the Ion Inject Time across each precursor window and across retention time bin:

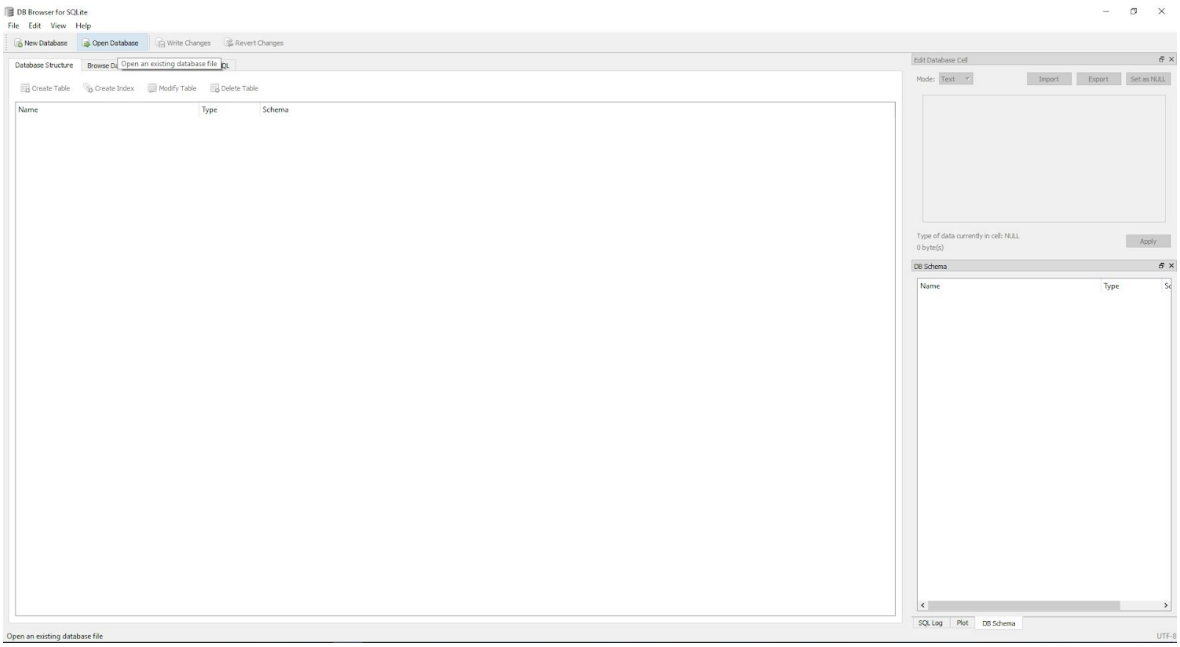


Viewing elib files with DB Browser for SQLite

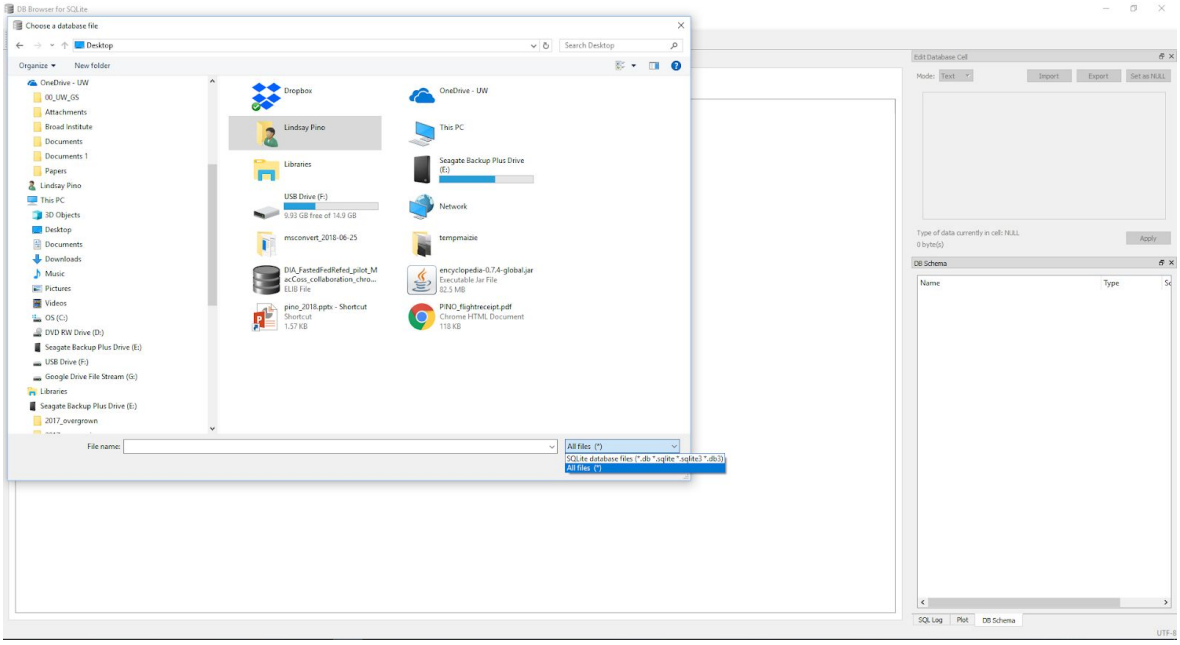
The *.elib files that Encyclopedia builds are SQL databases, which are kind of like multi-tab Excel files but fancier. To view these files and browse the data stored in the elib, follow the steps below:

Download DB Browser for SQLite here: https://sqlitebrowser.org/

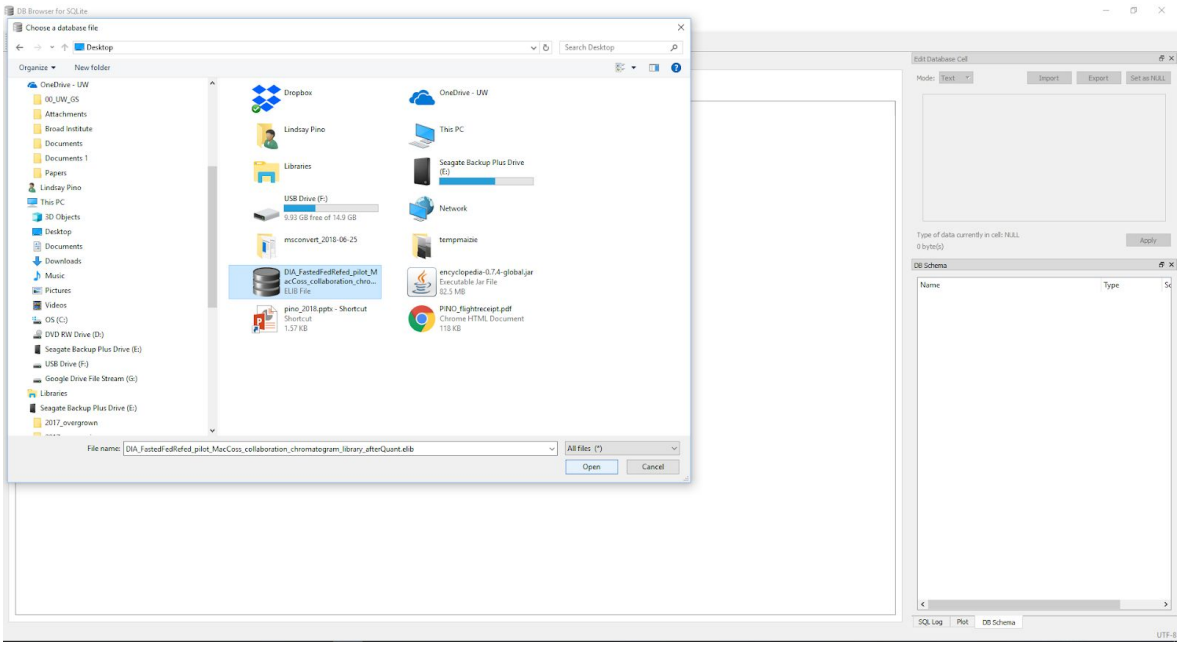
Open DB Browser for SQLite
Navigate to “Open Database” button on top left



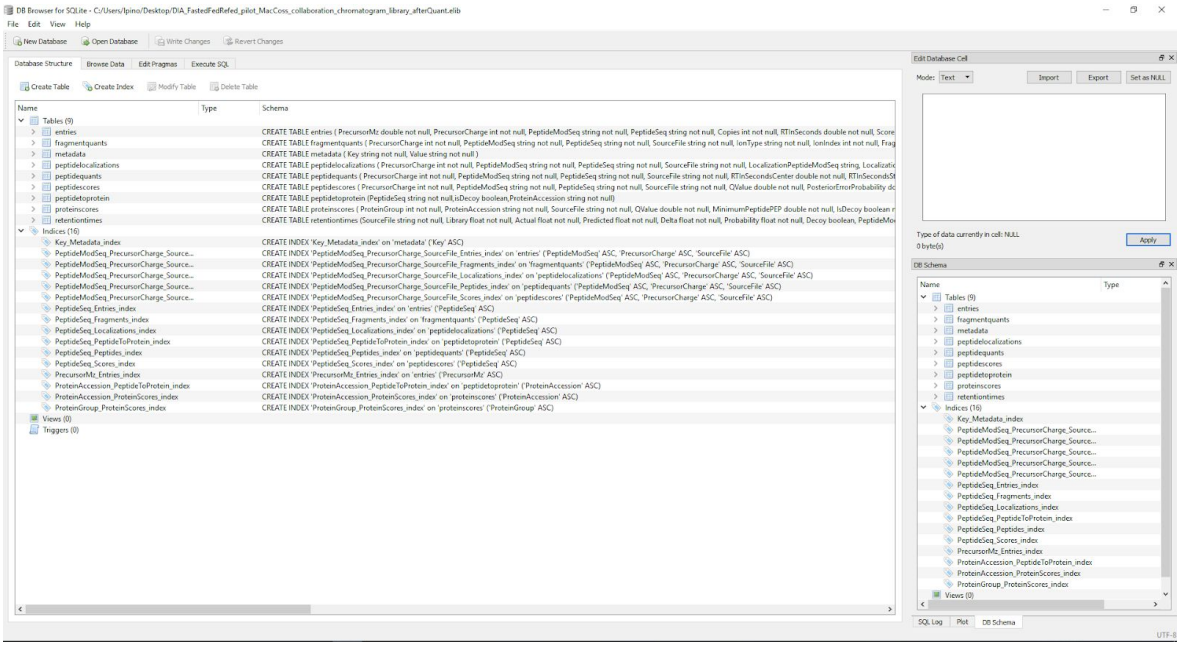
At the bottom of the “Choose a database file” pop-up next to the “File Name” field, select “All files” from the file type dropdown



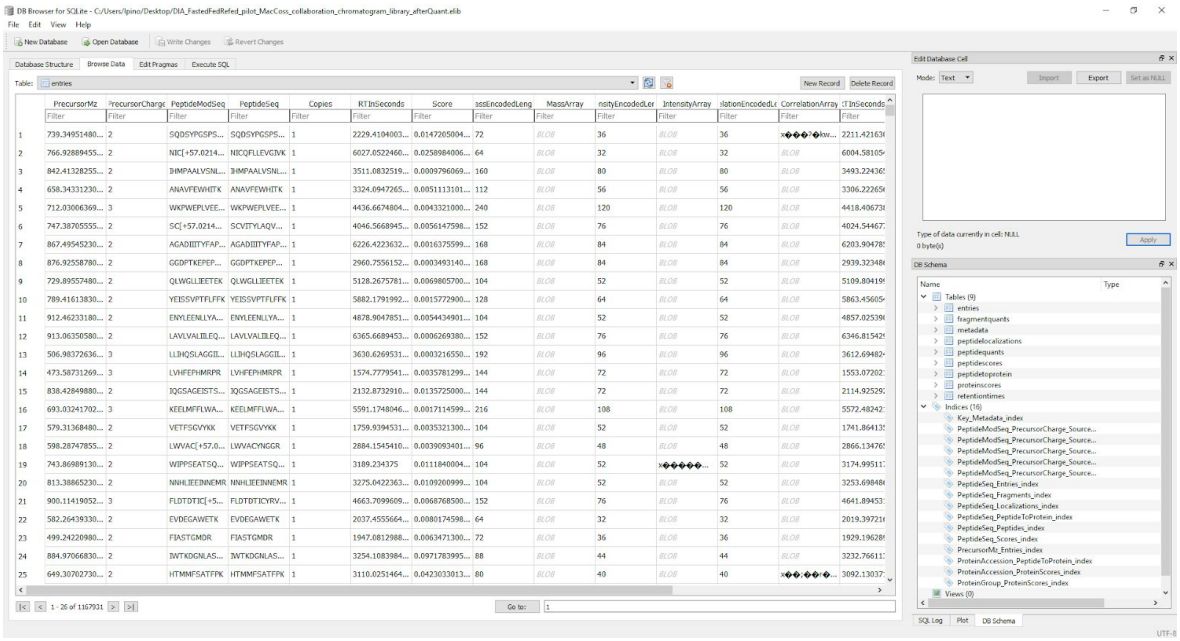
Navigate to the appropriate directory and select the elib file you want to view. Click .Open”



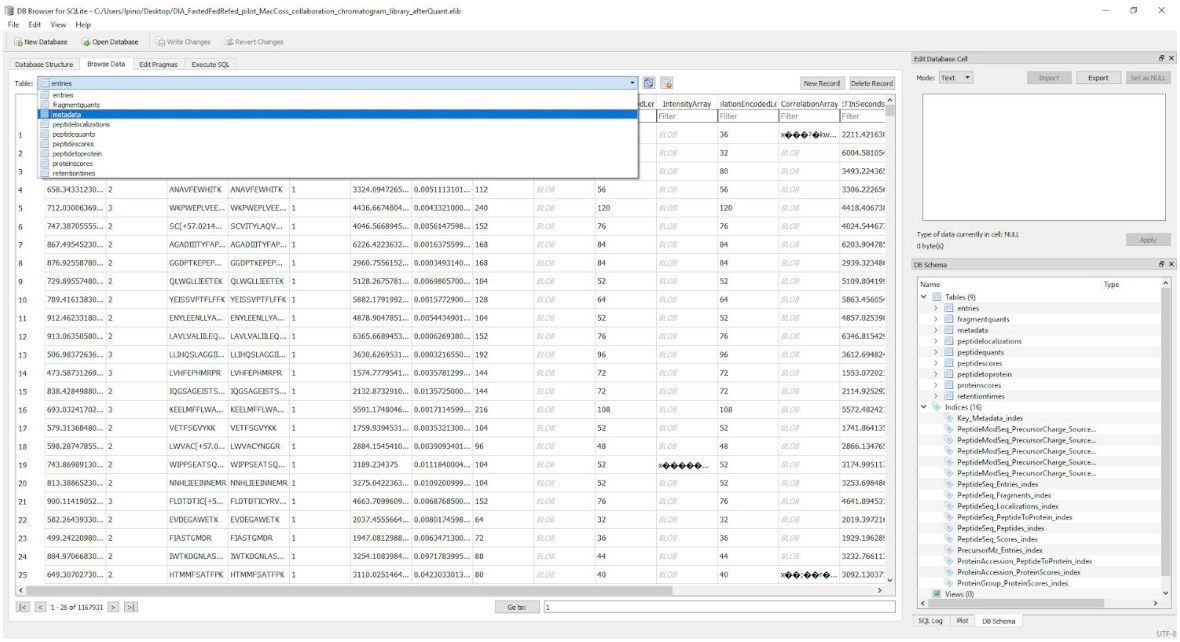
! It might take a minute to load. The window should look like this once it’s loaded:



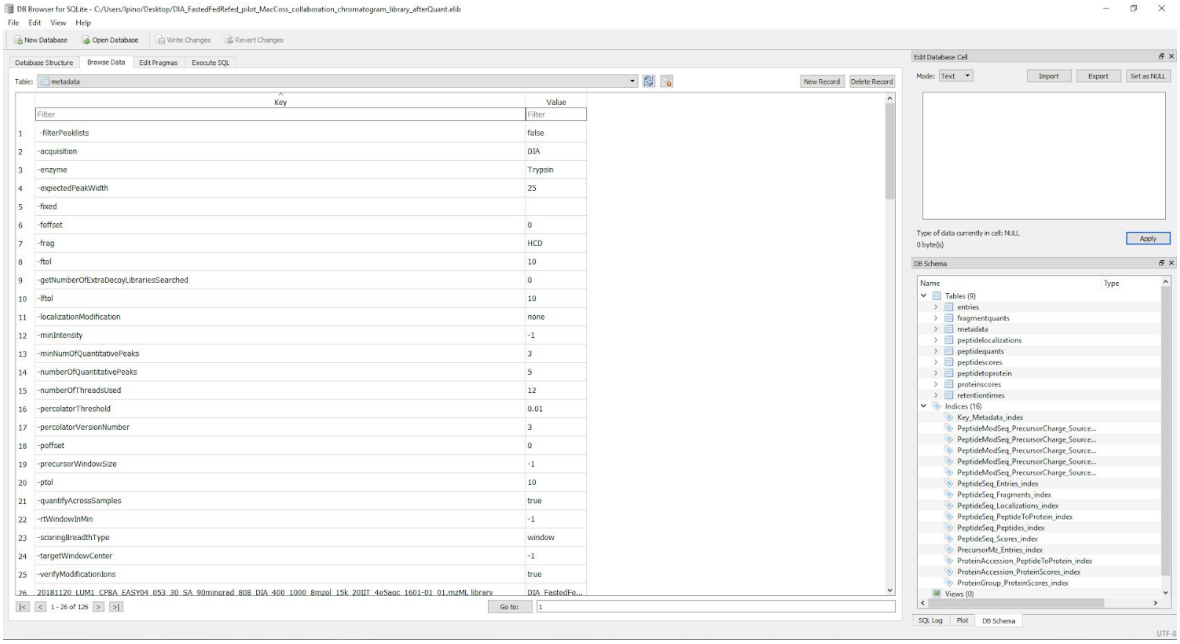
Click on “Browse Data” tab at top under the “Open Database” button.



Under the “Browse Data” tab, select the table you want to view from the “Table:” dropdown menu. For example, the meta data:



Choosing the “metadata” table will display information such as the parameter settings when Encyclopedia was run, the TIC for each raw file in the chromatogram library/quant report, etc



Appendix B: SINGLE POINT CALIBRATION TUTORIALS

B.1 SOURCE CODE

The source code to perform single point calibration and generate the figures in Chapter 2 is available at `https://bitbucket.org/lkpino/single-point_calibration/wiki/Home`.

B.2 MANUAL

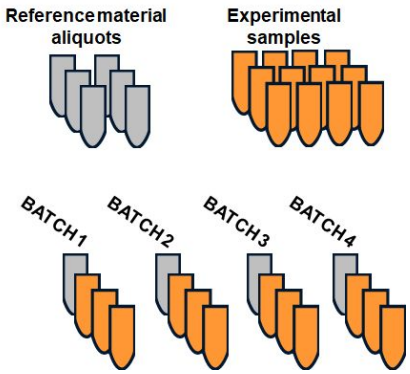
A walkthrough of how to harmonize quantitative mass spectrometry proteomics data using single-point calibration to a reference material.

This tutorial will detail the step-by-step process using single-point calibration to a reference material. The data I'm using is the same as what was presented at the 2018 Skyline User Group Meeting, where yeast was exposed to YEPD + high salt (osmotic shock, "sample") or a just YEPD solution (control, "reference"). The yeast in YEPD is our reference material in this experiment. In other words, we want to make measurements relative to this reference. Because this particular strain of yeast (BY4741) was used previously to measure protein molecules-per-cell by Ghaemmaghami *et al*, there is also a biological unit associated with each protein in the reference.

DESIGNING AN EXPERIMENT WITH A REFERENCE MATERIAL

Choosing a reference material. When setting up an experiment, researchers chose controls that represent the system being studied, often reflecting sources of intra- and intergroup variation so that comparisons to the experimental group will be robust. Similarly, choosing a reference material should represent the system being studied so that sources of variation in workflow are captured in parallel to the experimental samples. For example, an experiment in human plasma might use healthy individuals as controls, age- and gender-matched to experimental samples. An appropriate reference material in this situation might be a pool of healthy and experimental plasma samples, so that representative analytes and matrix effects are present. There is also commercially available pooled plasma, which might be another consideration for a reference material especially in situations where experimental samples are precious. Another consideration when choosing a reference material might be how well a material is characterized. For example, in yeast studies, the BY4741 strain has been thoroughly characterized by DNA sequencing, perturbation experiments, and quantitative protein quantification assays.

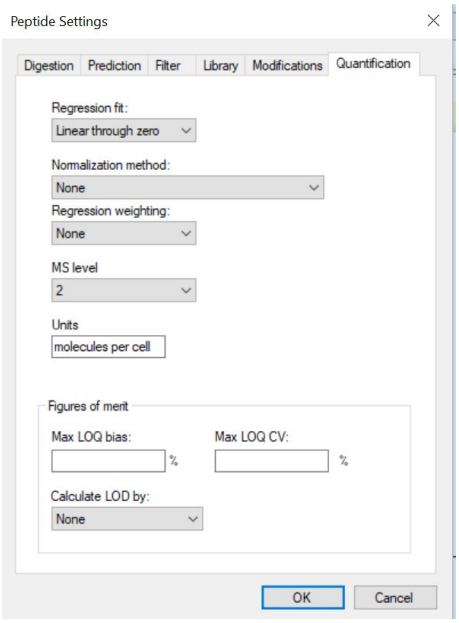
Preparing a reference material. There should be enough mass/volume of reference material to prepare at least one aliquot for each batch of samples, plus at least one extra aliquot so that this reference material can be calibrated to future materials. A reference material aliquot should be included in each sample batch, however the samples happen to be batched (e.g. sample preparation bottlenecks, longitudinal study time points, or instrument availability).



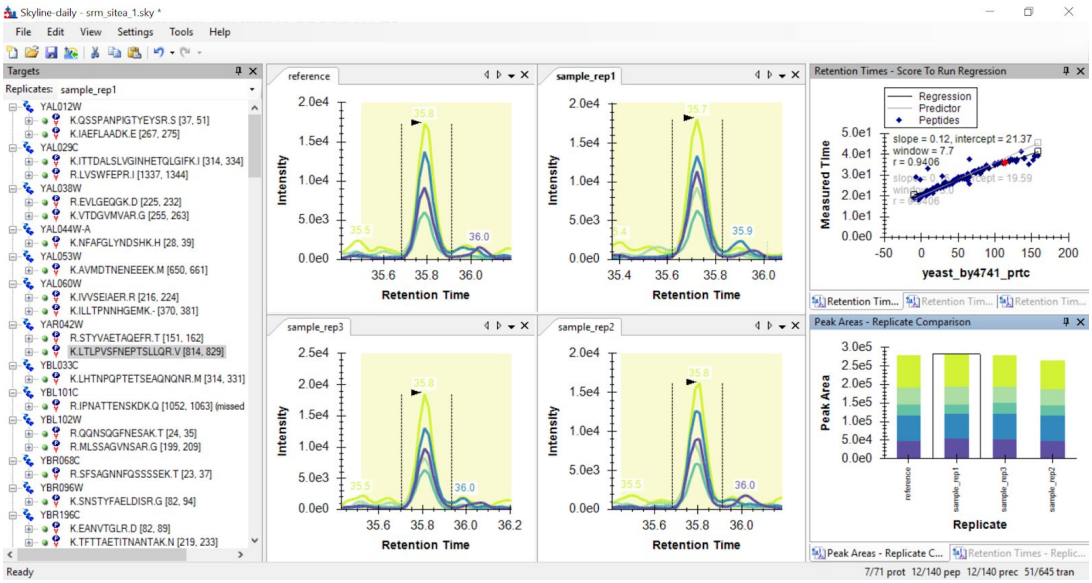
In this tutorial, we repeated an osmotic shock experiment performed by Selevsek *et al*, and calibrated the experiment samples to a common yeast reference material, BY4741 grown under standard laboratory conditions. We grew a bulk culture of BY4741 and harvested multiple aliquots of 50mL culture volumes to use for reference material. When an experimental sample batch was prepared, one of these reference material cell pellets was prepared in parallel and the LC-MS data acquired alongside the experimental samples. Next, we'll harmonize the sample preparation batches using Skyline to process the data.

SINGLE-POINT CALIBRATION TO A REFERENCE MATERIAL USING SKYLINE

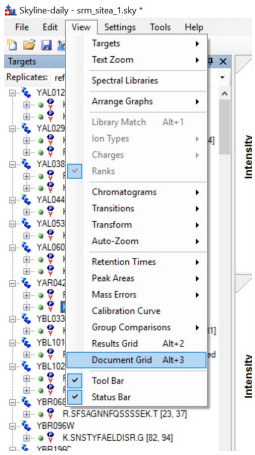
1. Set up the Skyline document.
    a. Format the experiment-specific **Peptide Settings**, **Transition Settings**, and target list. For single-point calibration, be sure to set up the **Quantification** tab under **Peptide Settings** as follows:
        - **Regression fit:** Linear through zero
        - **Normalization method:** None
        - **Regression weighting:** None



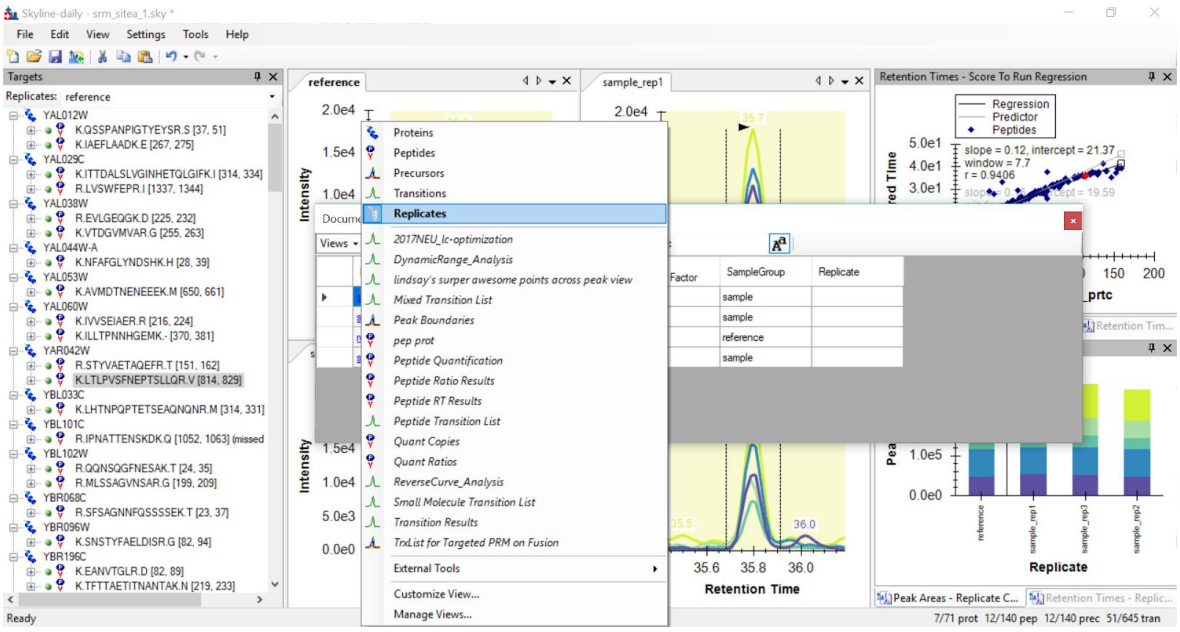
    b. Import the acquisition data.



c. View > Document Grid



d. In the Document Grid pane, click on “Views” and select “Replicates” from the drop-down menu. * Note: you won’t have all the italicized Document Grid views that are listed here, that’s okay



e. Fill out the Document Grid appropriately, according to your experimental design. In this batch, I have four raw files: one file for the reference material acquisition and a file for each of the osmotic shock sample replicates.

Sample Type

Set experimental samples to “Unknown”

Set reference material to “Standard”

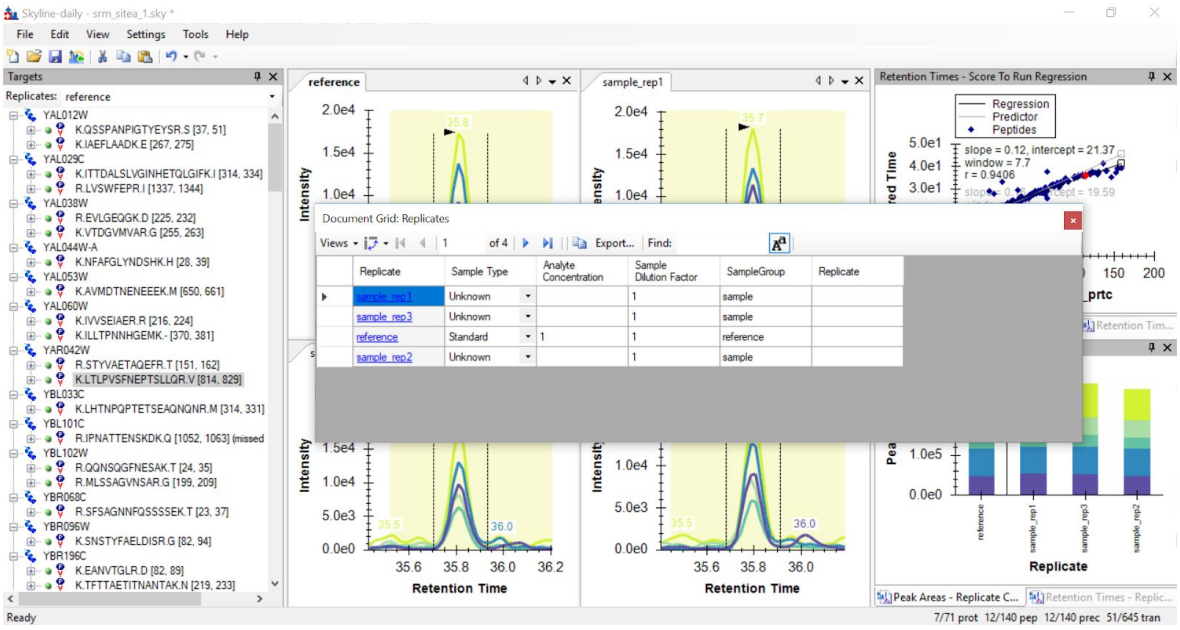
Analyte Concentration

Leave experimental samples blank

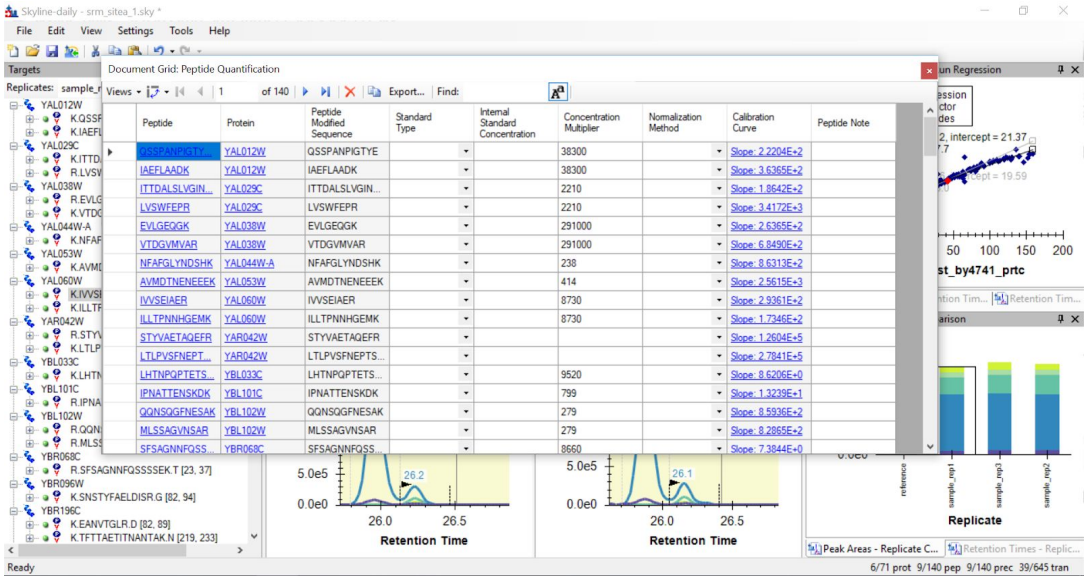
Set reference to “1”

Sample Dilution Factor

Set all files to “1”

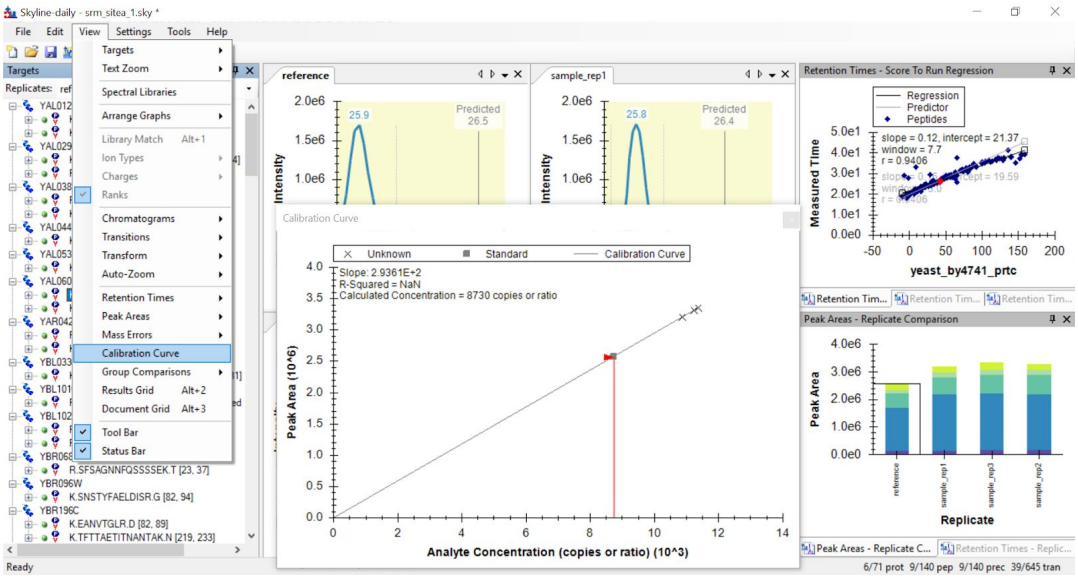


f. OPTIONAL: if you have meaningful units you want to attach to specific proteins or peptides, you can enter them into the Document Grid using View “Peptide Quantification”. Here is an example where I have entered the protein molecules-per-cell reported by Ghaemmaghami *et al*.

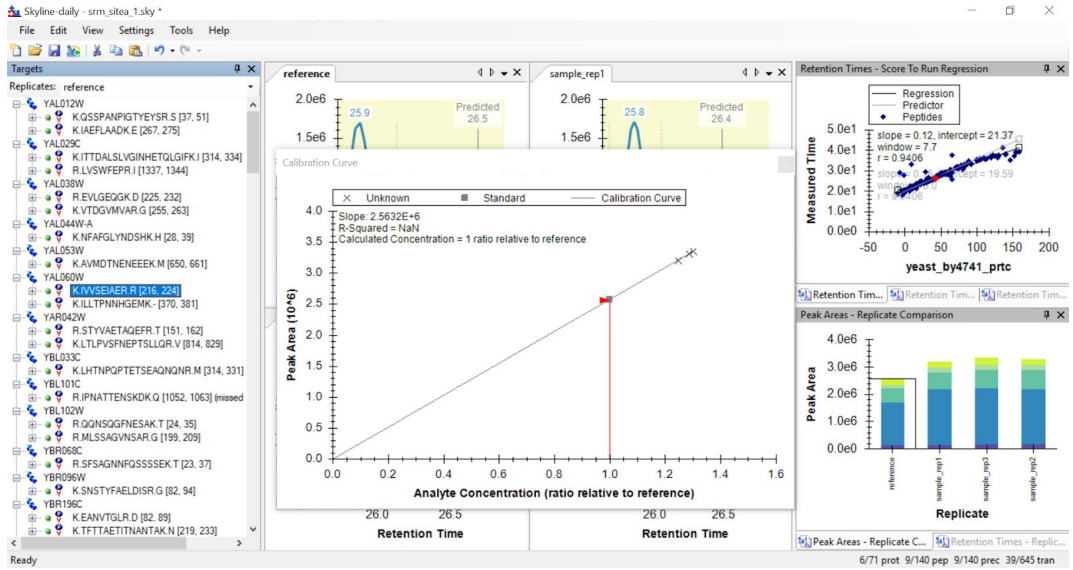


g. View > Calibration Curve

With a biological multiplier:

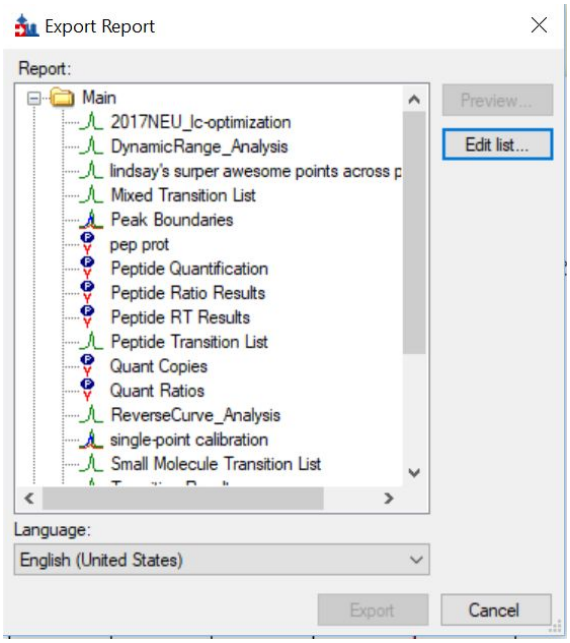


Without a biological multiplier:

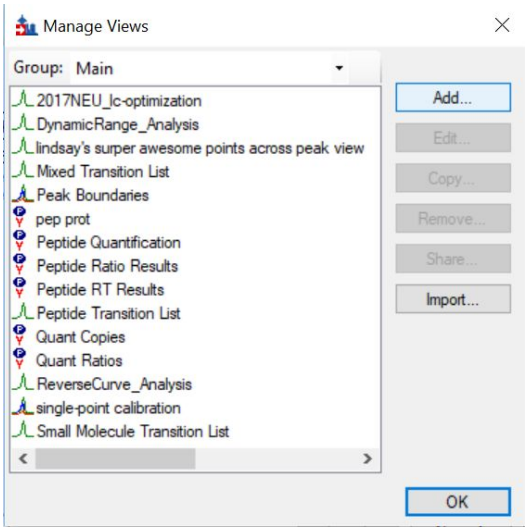


2. Repeat this process for each batch+reference you have. (Skyline does not yet support multiple reference samples in a single document, so we need to make a new document for each batch.)

3. Export the data from each of the batches' Skyline files
    a. File > Export > Report
    b. Select "Edit list..."

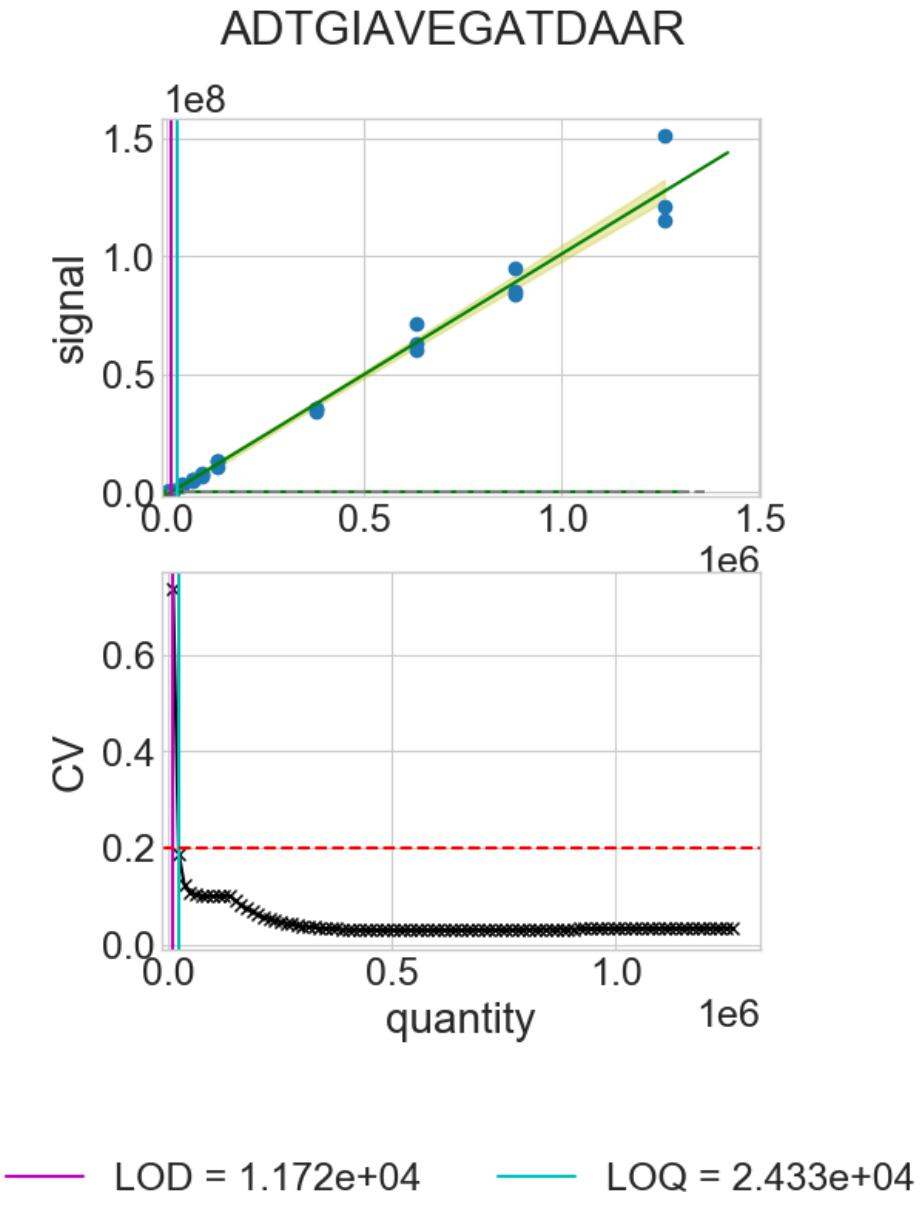


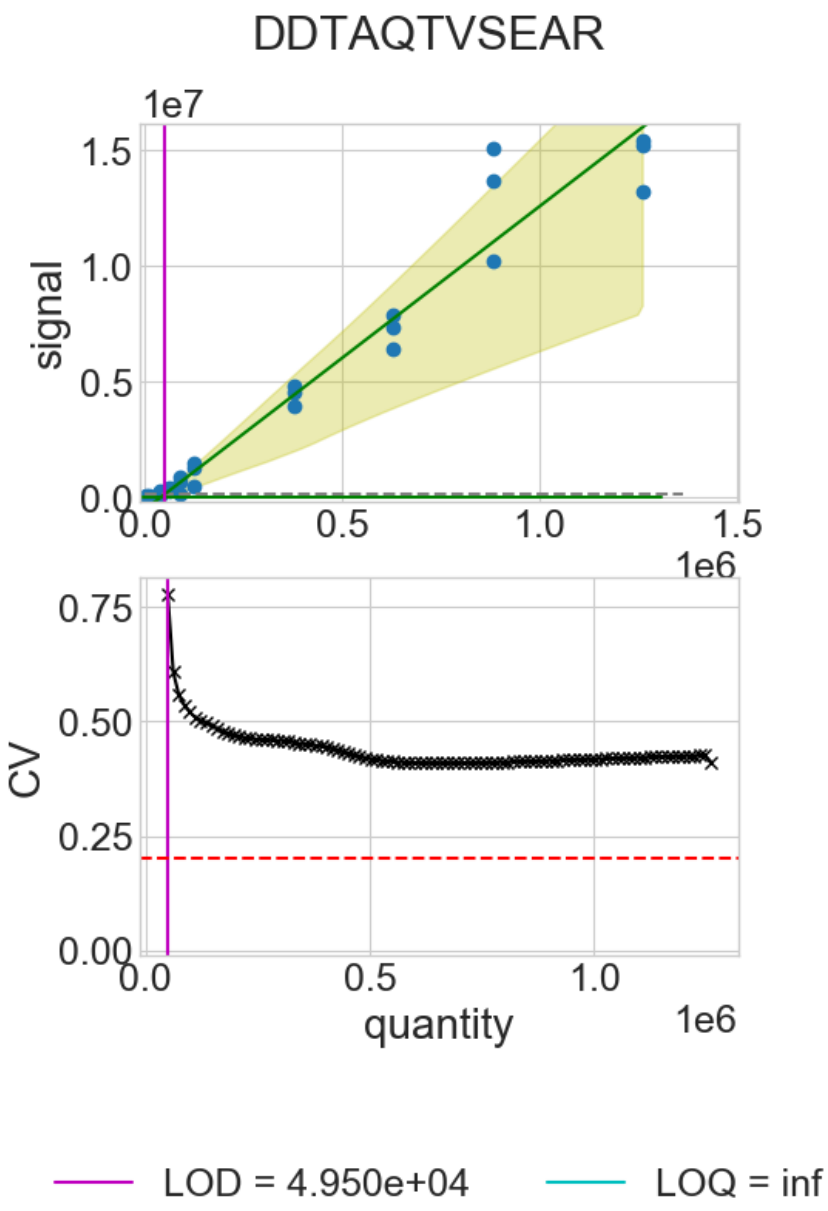
c. Select “Add...”

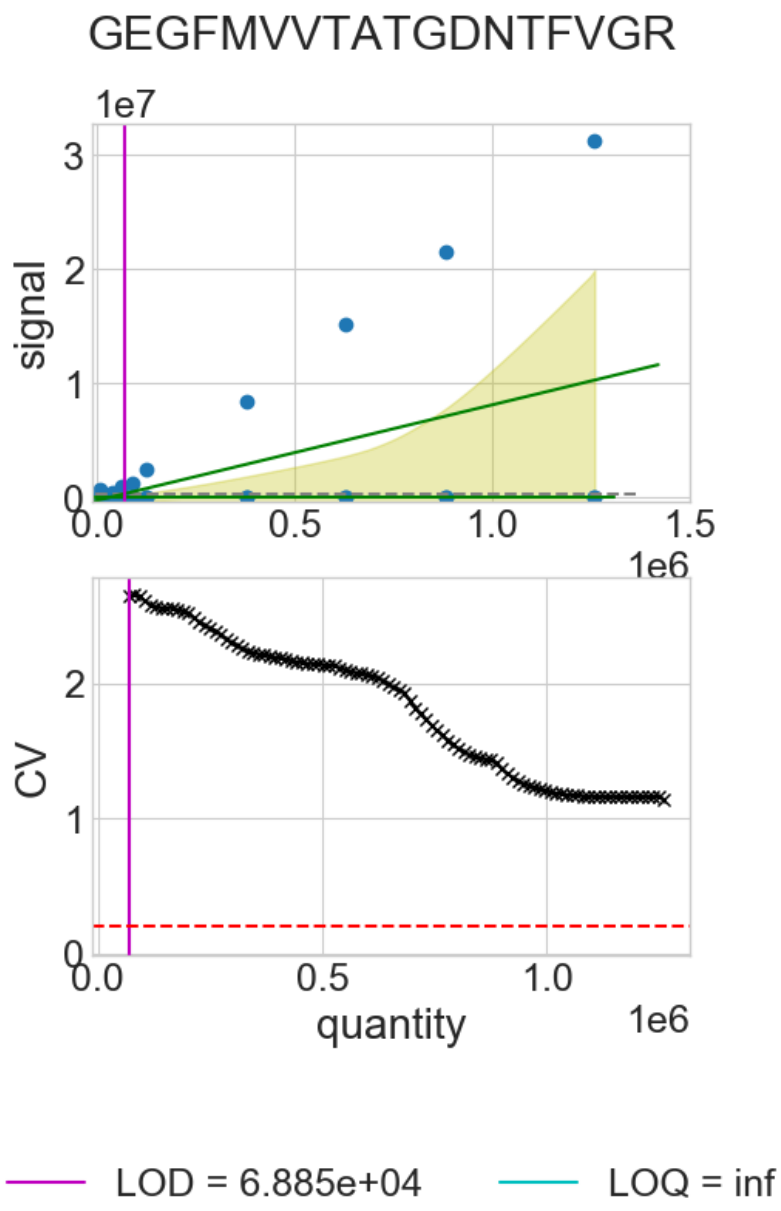


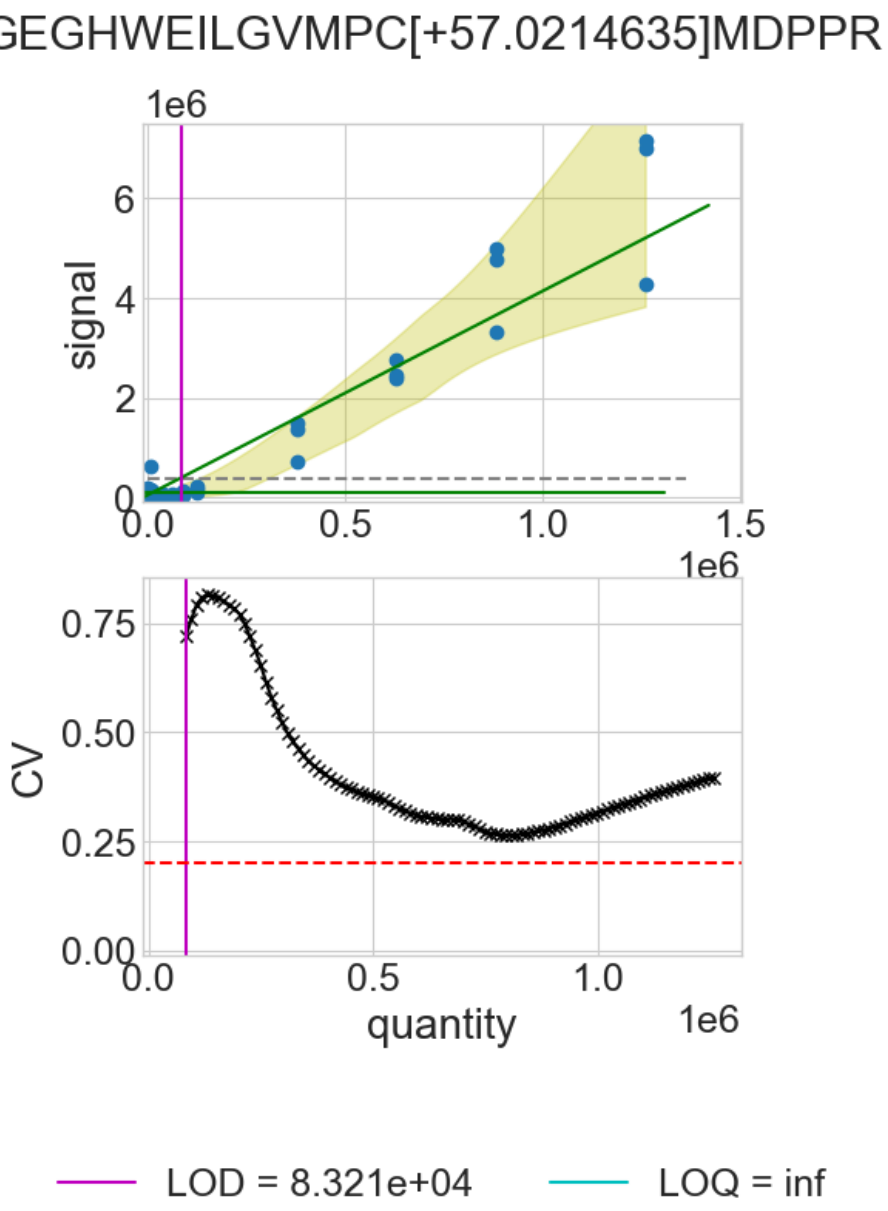
d. Choose the desired fields. Note: The “Calculated Concentration” field will export the Calibration Curve values that were computed in step 1.

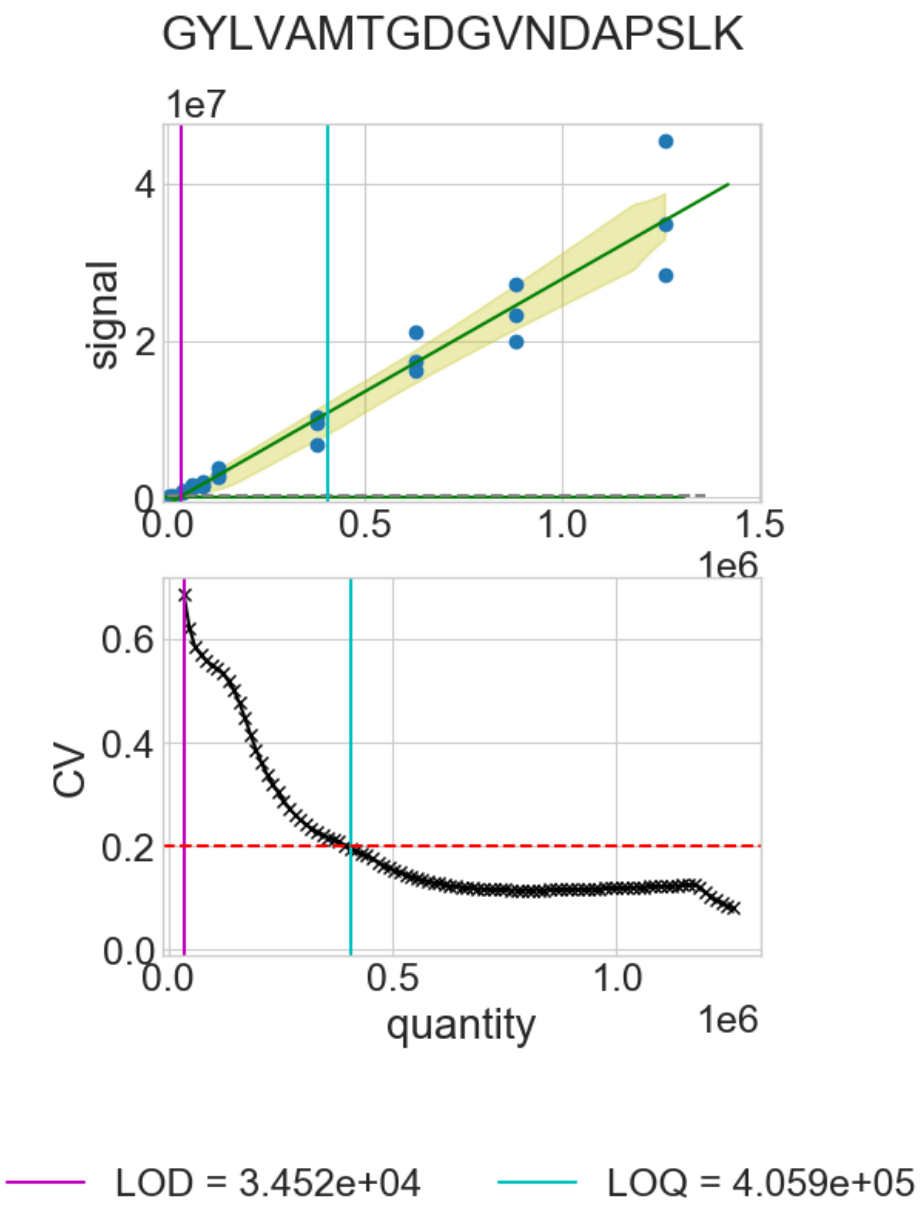
Appendix C: MATCHED MATRIX CALIBRATION CURVES FOR PROTEIN PMA1

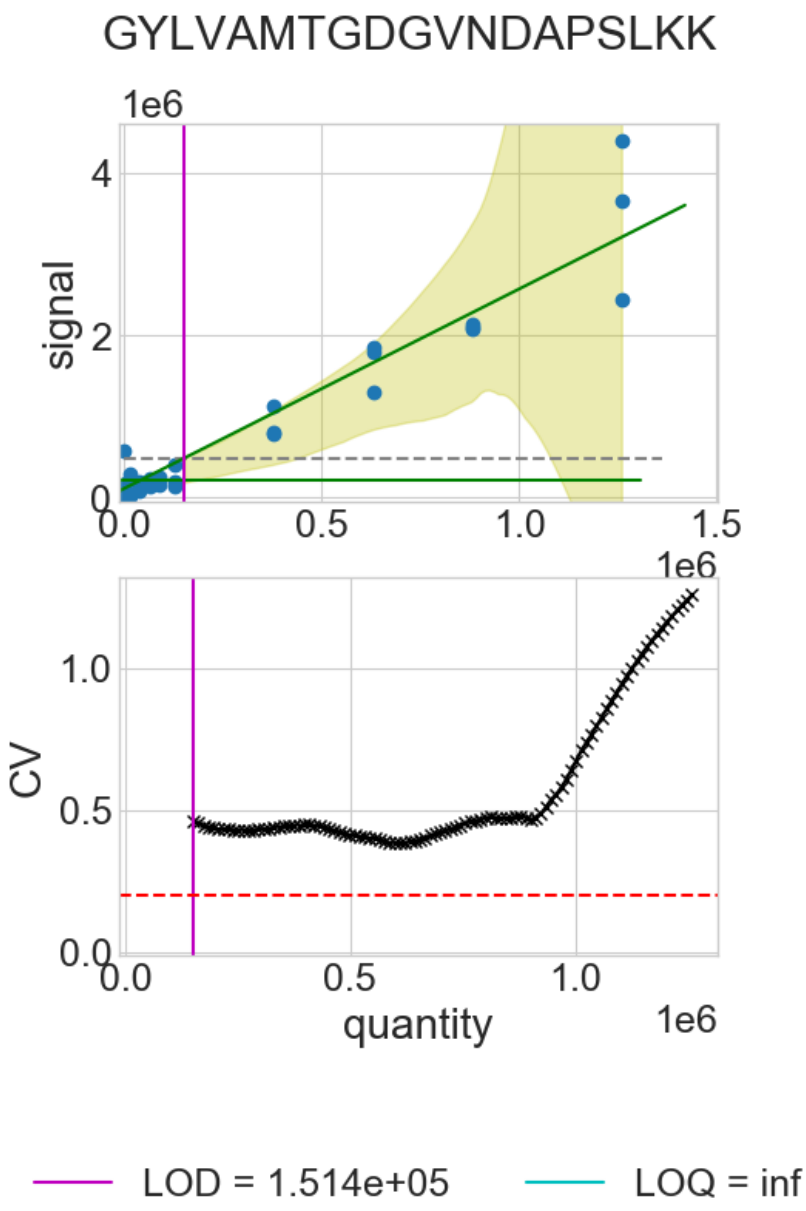


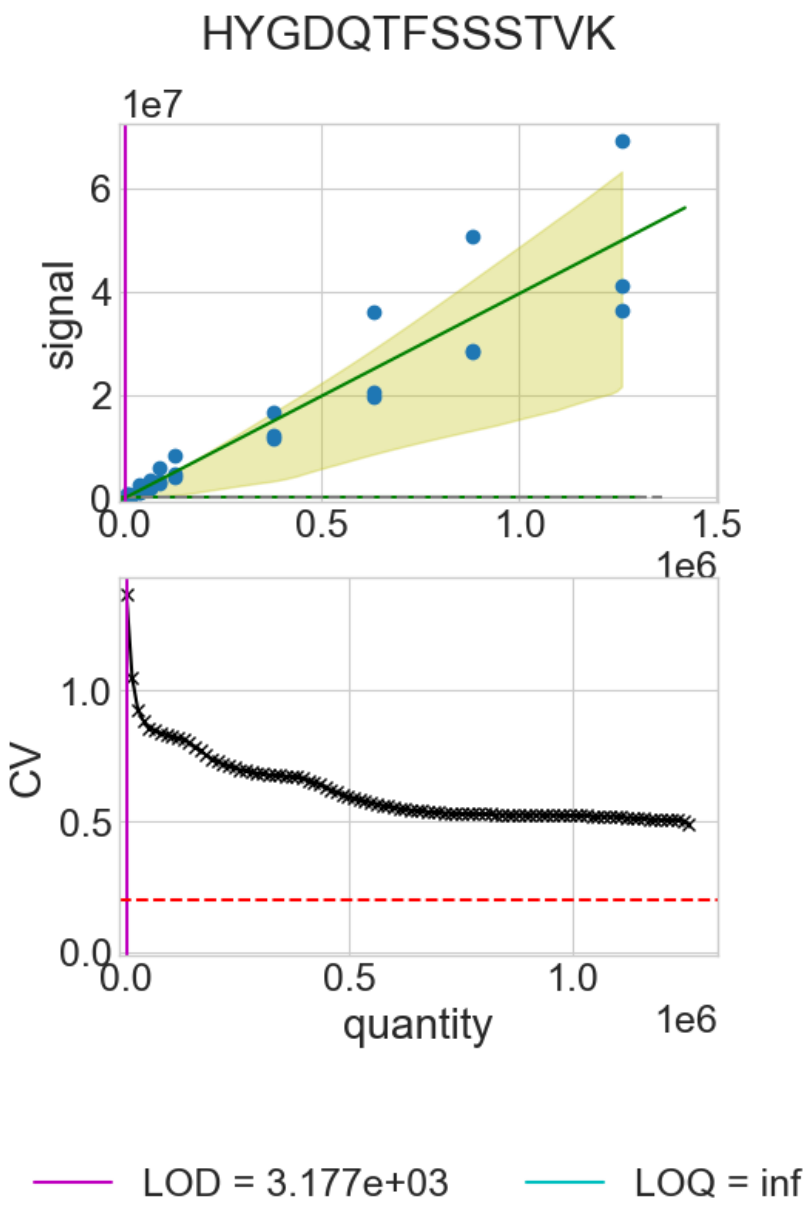

DDTAQTVSEAR
1e7
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
0.75
0.50
0.25
0.00
CV
0.0
0.5
1.0
quantity
1e6
LOD = 4.950e+04
LOQ = inf

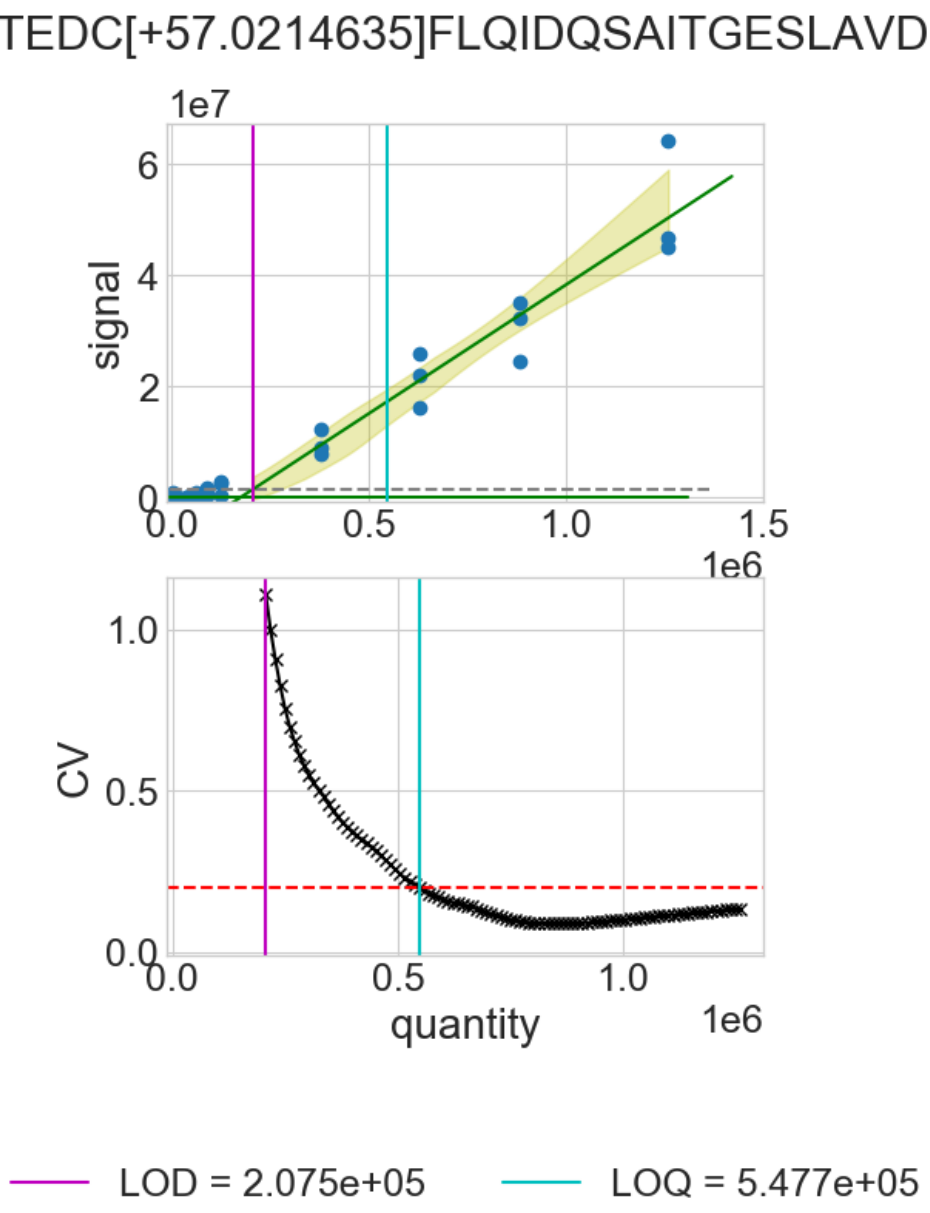

GEGFMVVTATGDNTFVGR
1e7
signal
3
2
1
0
0.0
0.5
1.0
1.5
1e6
CV
2
1
0
0.0
0.5
1.0
quantity
1e6
LOD = 6.885e+04
LOQ = inf

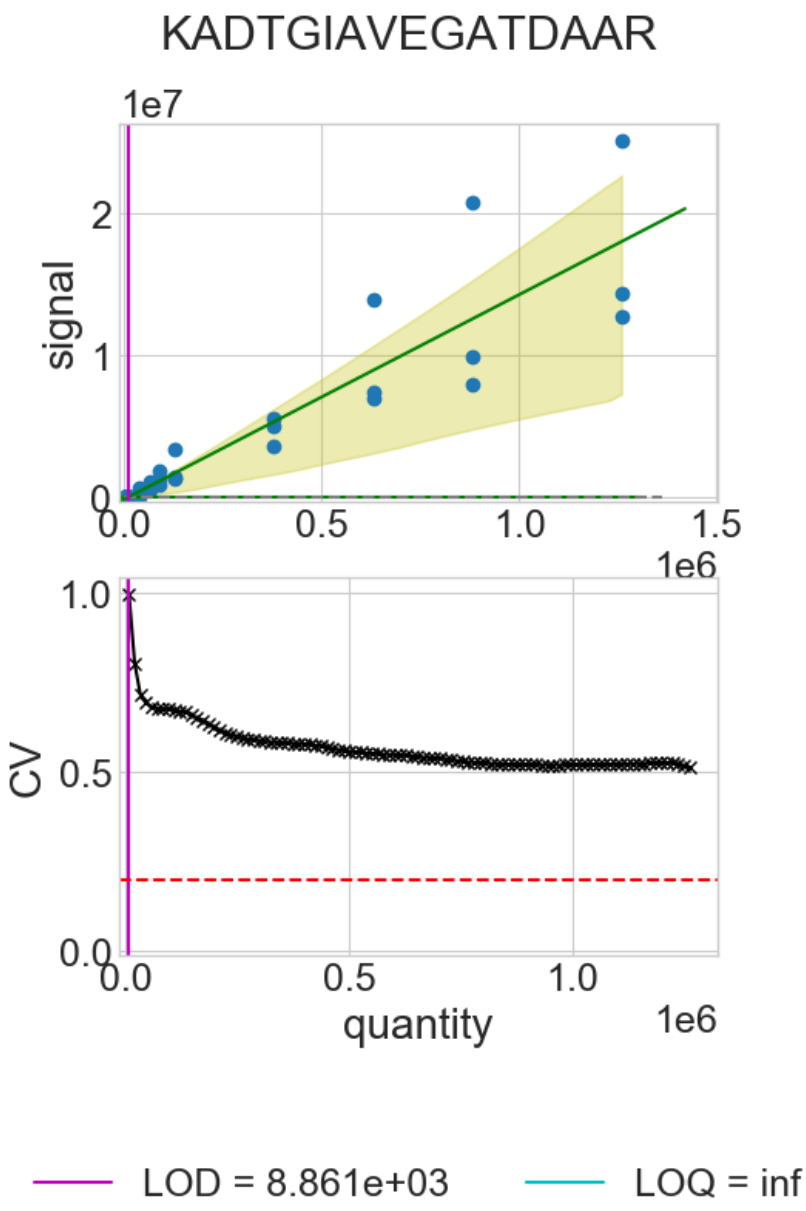

GEGHWEILGVMPC[+57.0214635]MDPPR
1e6
signal
0
2
4
6
0.0
0.5
1.0
1.5
1e6
CV
0.00
0.25
0.50
0.75
0.0
0.5
1.0
quantity
1e6
LOD = 8.321e+04
LOQ = inf

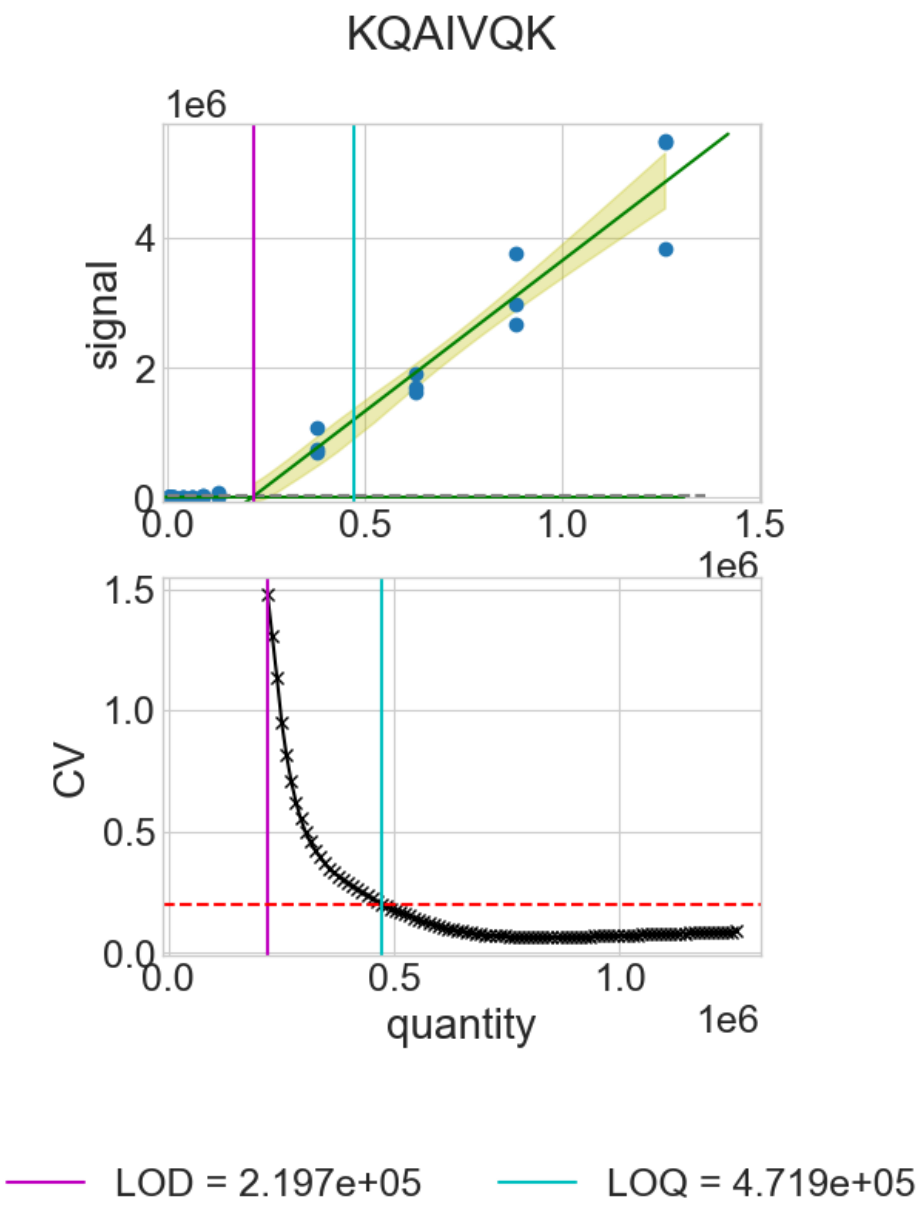

GYLVAMTGDGVNDAPSLK
1e7
signal
4
2
0
0.0
0.5
1.0
1.5
1e6
CV
0.6
0.4
0.2
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 3.452e+04
LOQ = 4.059e+05

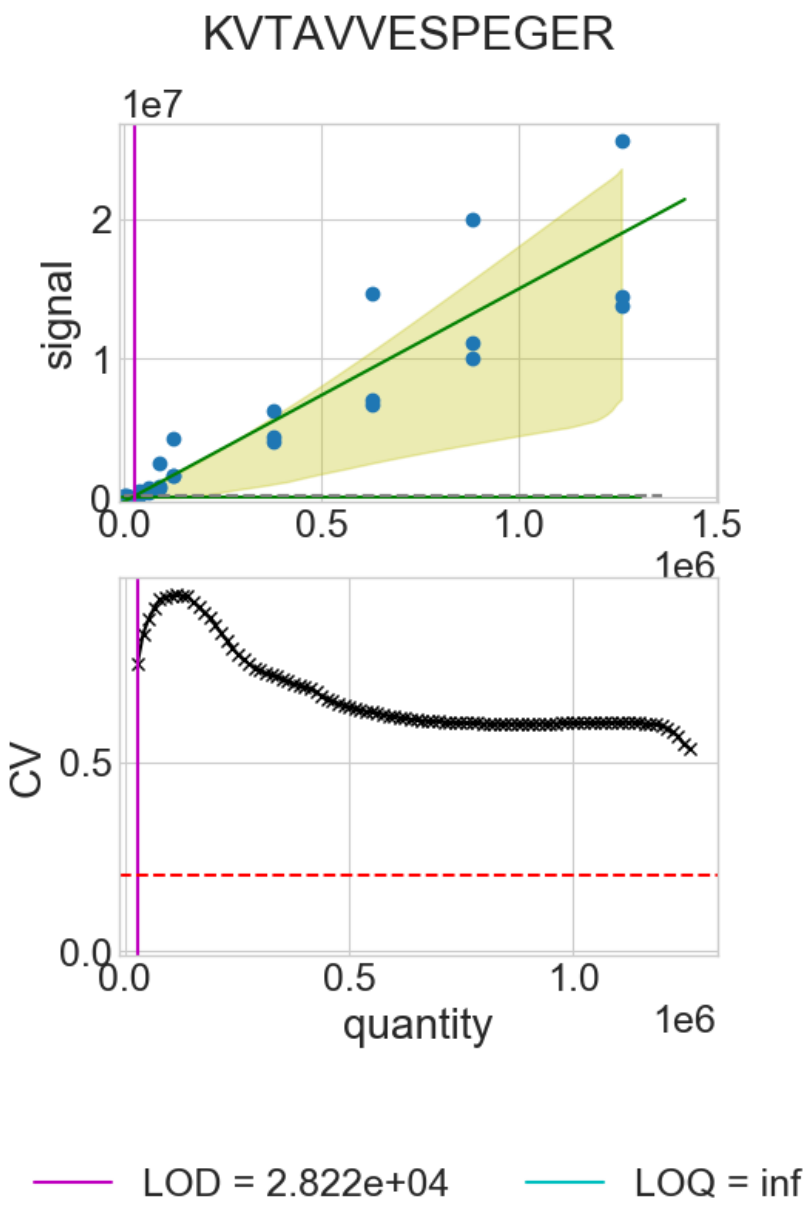

GYLVAMTGDGVNDAPSLKK
1e6
signal
0
2
4
0.0
0.5
1.0
1.5
1e6
CV
0.0
0.5
1.0
0.0
0.5
1.0
quantity
1e6
LOD = 1.514e+05
LOQ = inf

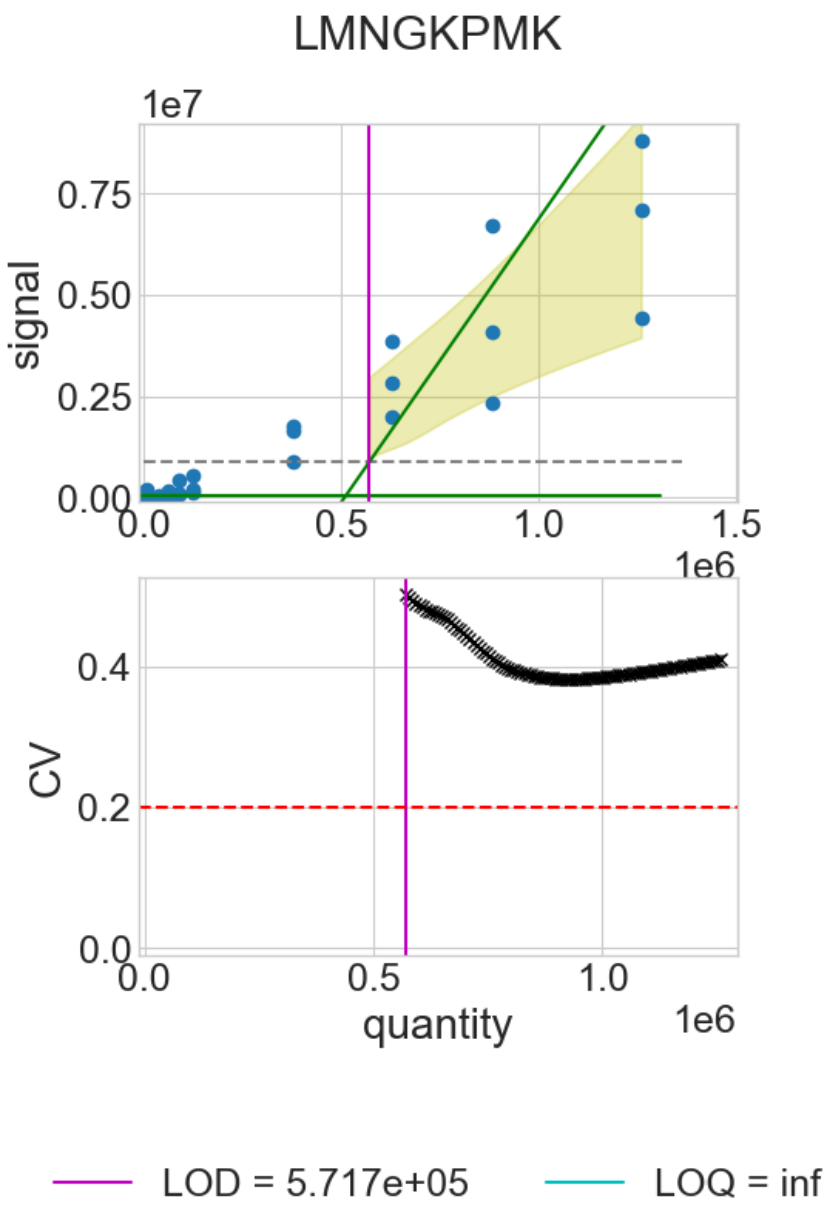

HYGDQTFSSSTVK
1e7
signal
6
4
2
0
0.0
0.5
1.0
1.5
1e6
CV
1.0
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 3.177e+03
LOQ = inf

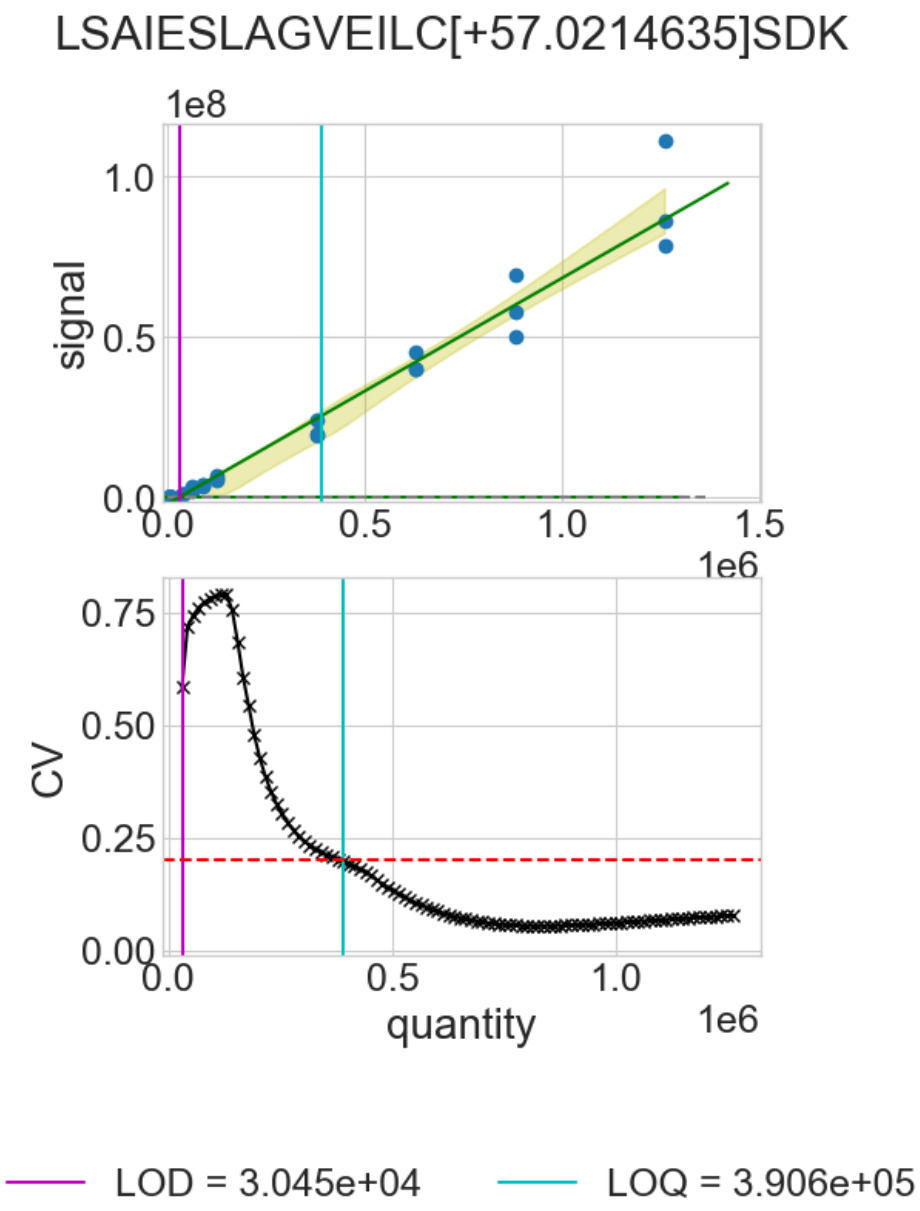

IVTEDC[+57.0214635]FLQIDQSAITGESLAVDK
1e7
signal
6
4
2
0
0.0
0.5
1.0
1.5
1e6
CV
1.0
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 2.075e+05
LOQ = 5.477e+05


KADTGIAVEGATDAAR
1e7
signal
2
1
0
0.0
0.5
1.0
1.5
1e6
CV
1.0
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 8.861e+03
LOQ = inf

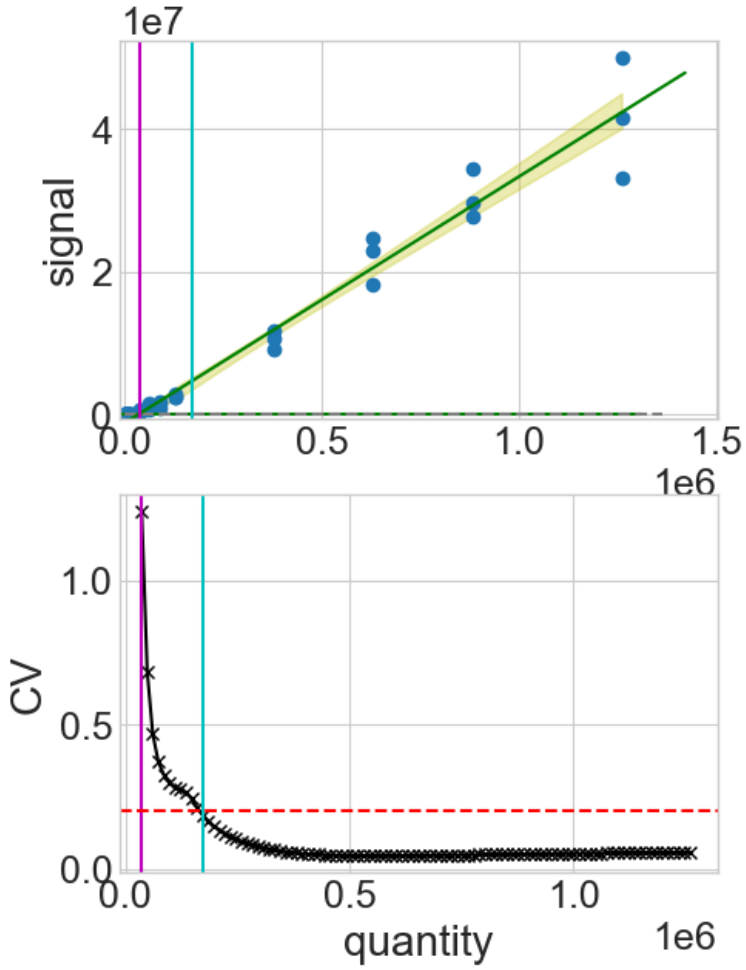

KQAIVQK
1e6
signal
4
2
0
0.0
0.5
1.0
1.5
1e6
CV
1.5
1.0
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 2.197e+05
LOQ = 4.719e+05


KVTAVVESPEGER
1e7
signal
2
1
0
0.0
0.5
1.0
1.5
1e6
CV
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 2.822e+04
LOQ = inf

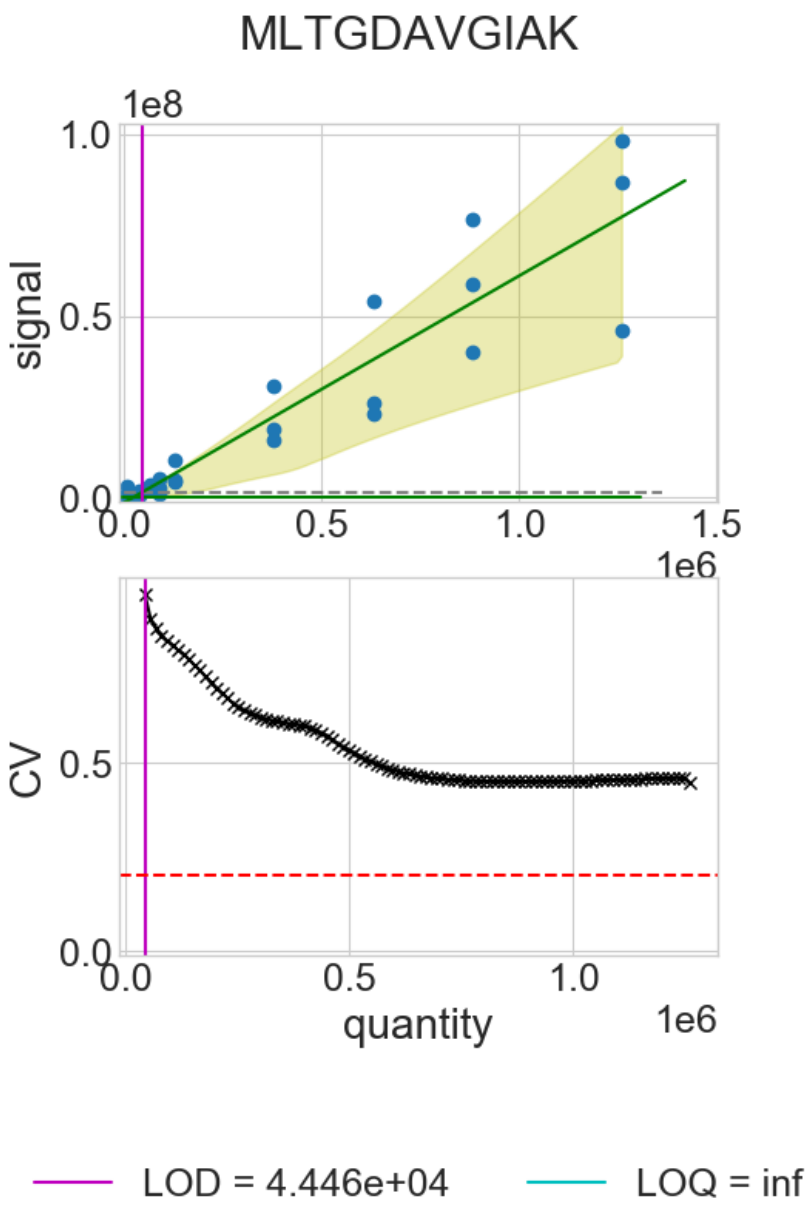

LMNGKPMK
1e7
0.75
0.50
0.25
0.00
signal
0.0
0.5
1.0
1.5
1e6
0.4
0.2
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 5.717e+05
LOQ = inf

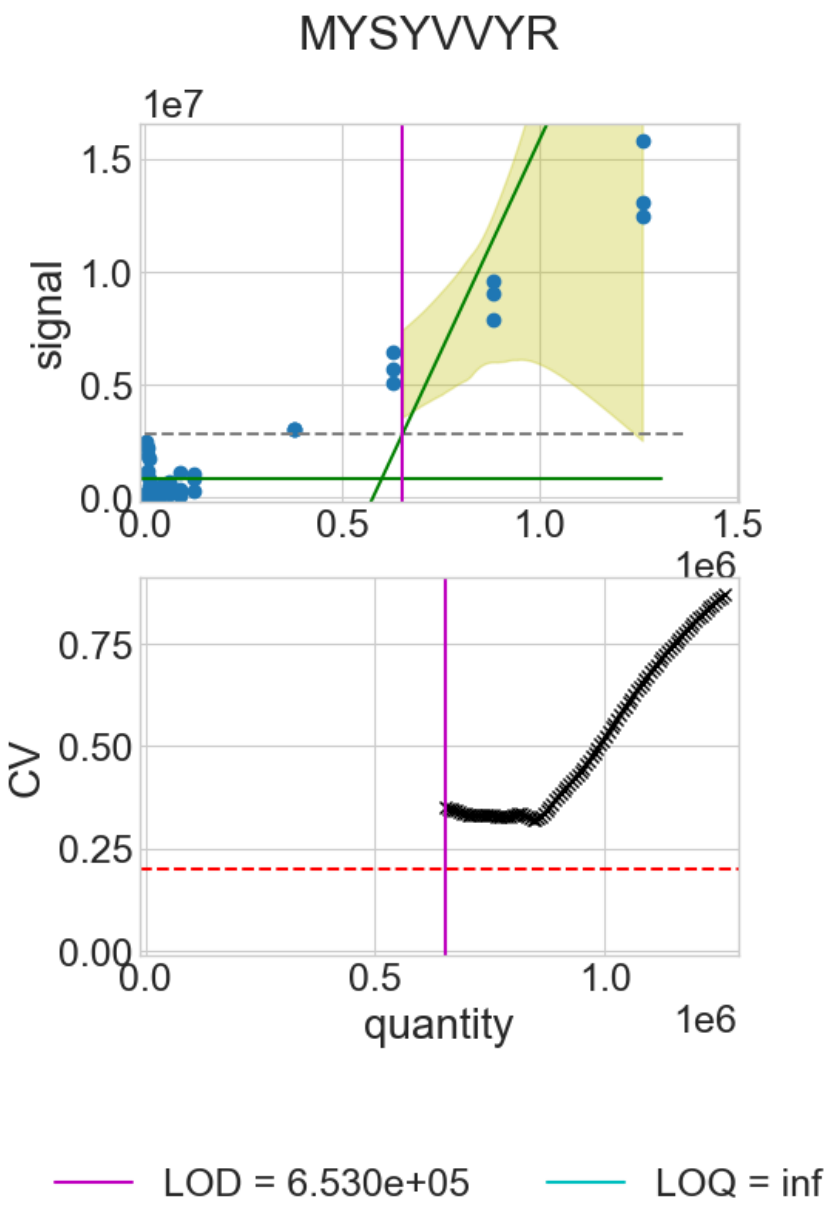

LSAIESLAGVEILC[+57.0214635]SDK
1e8
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
0.75
0.50
0.25
0.00
CV
0.0
0.5
1.0
quantity
1e6
LOD = 3.045e+04
LOQ = 3.906e+05

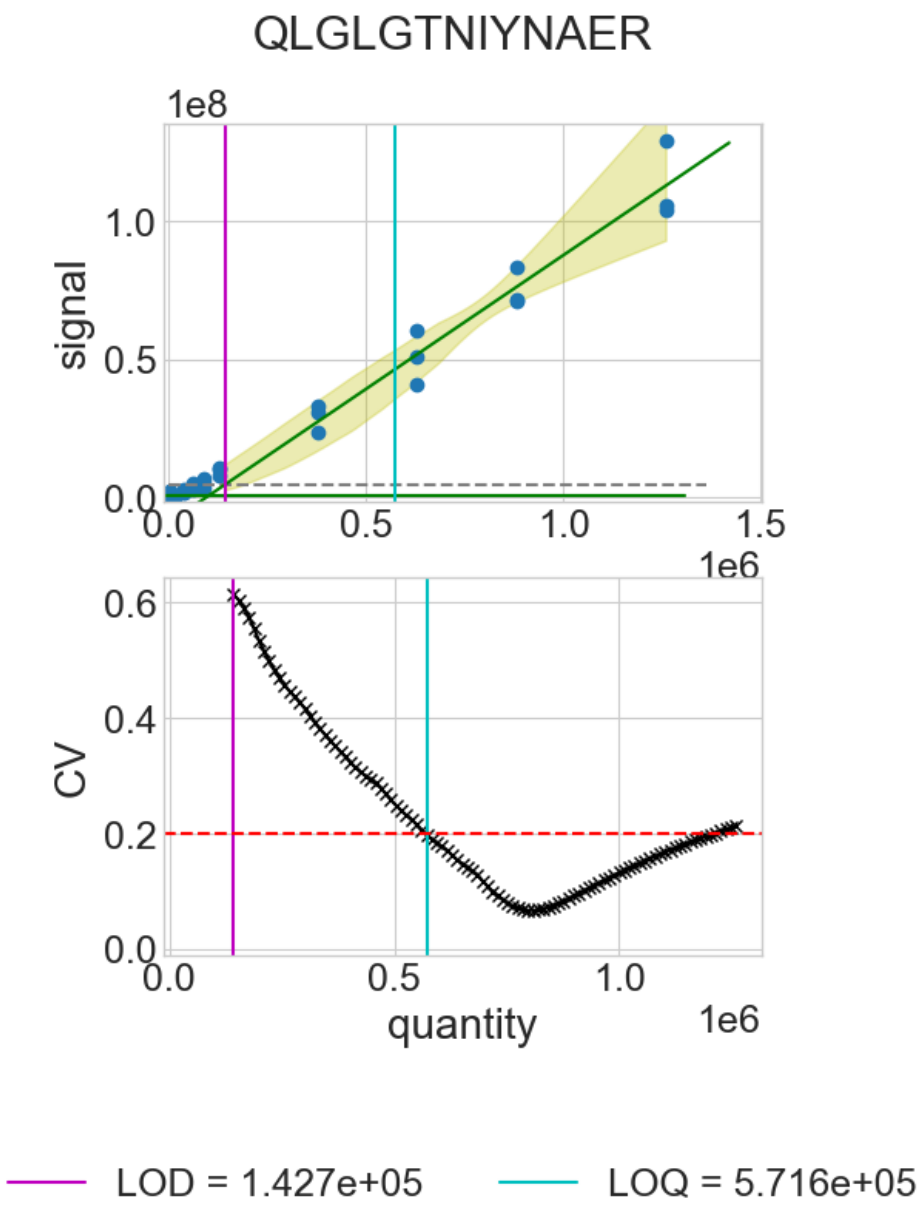
LSLHEPYTVEGVSPDDLMLTAC[+57.0214635]LAASR

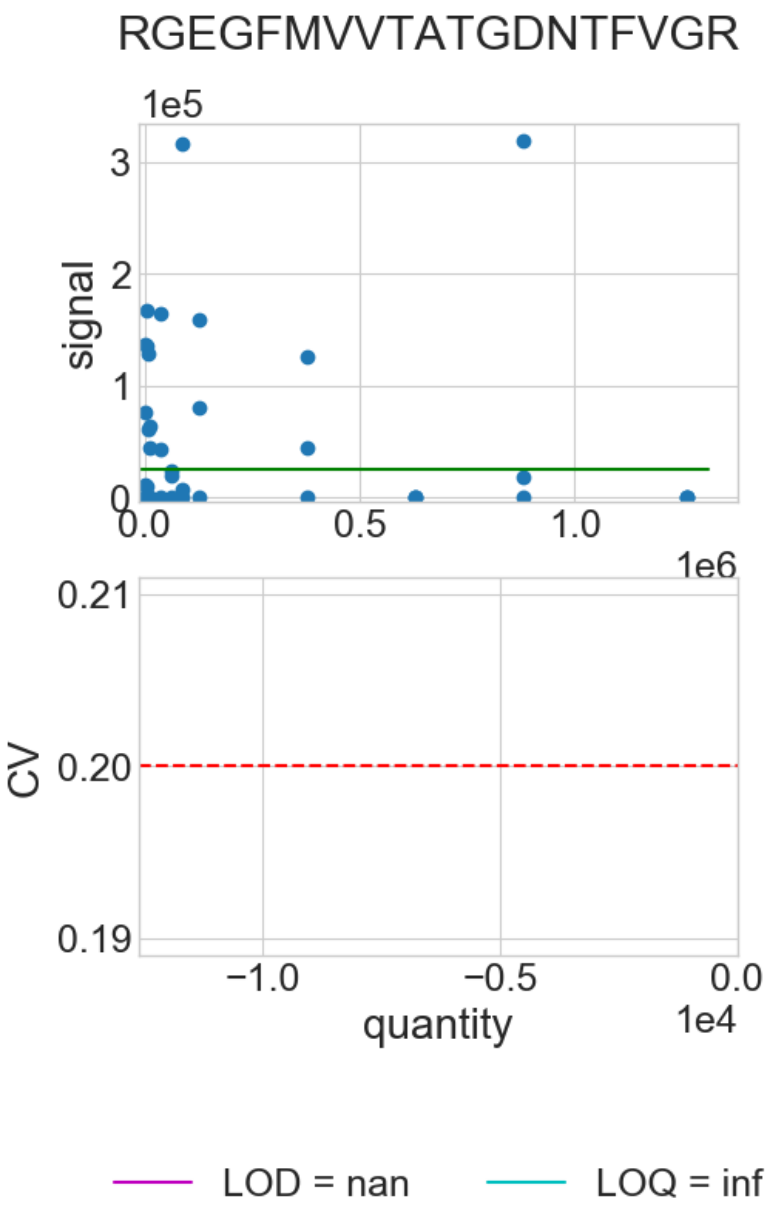


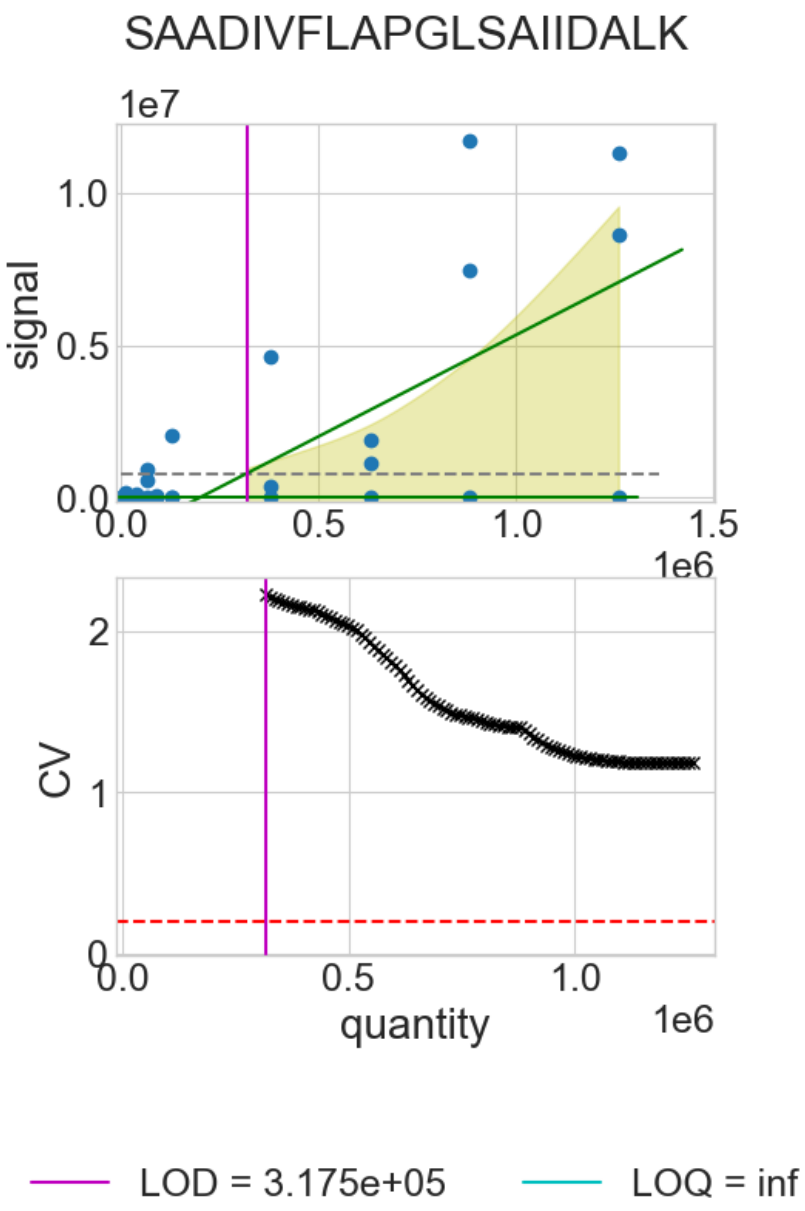
LOD = 3.633e+04 LOQ = 1.723e+05

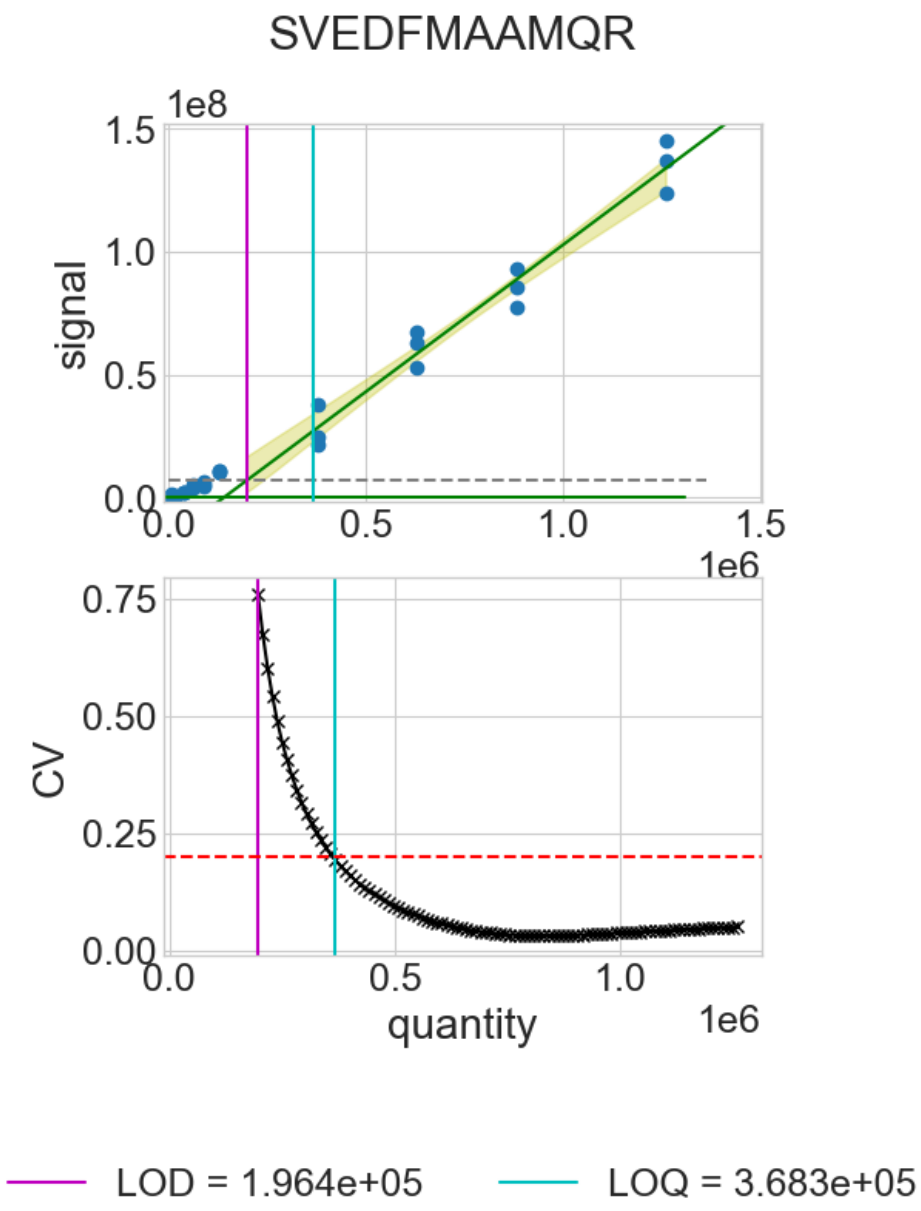

MLTGDAVGIAK
1e8
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
CV
0.5
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 4.446e+04
LOQ = inf

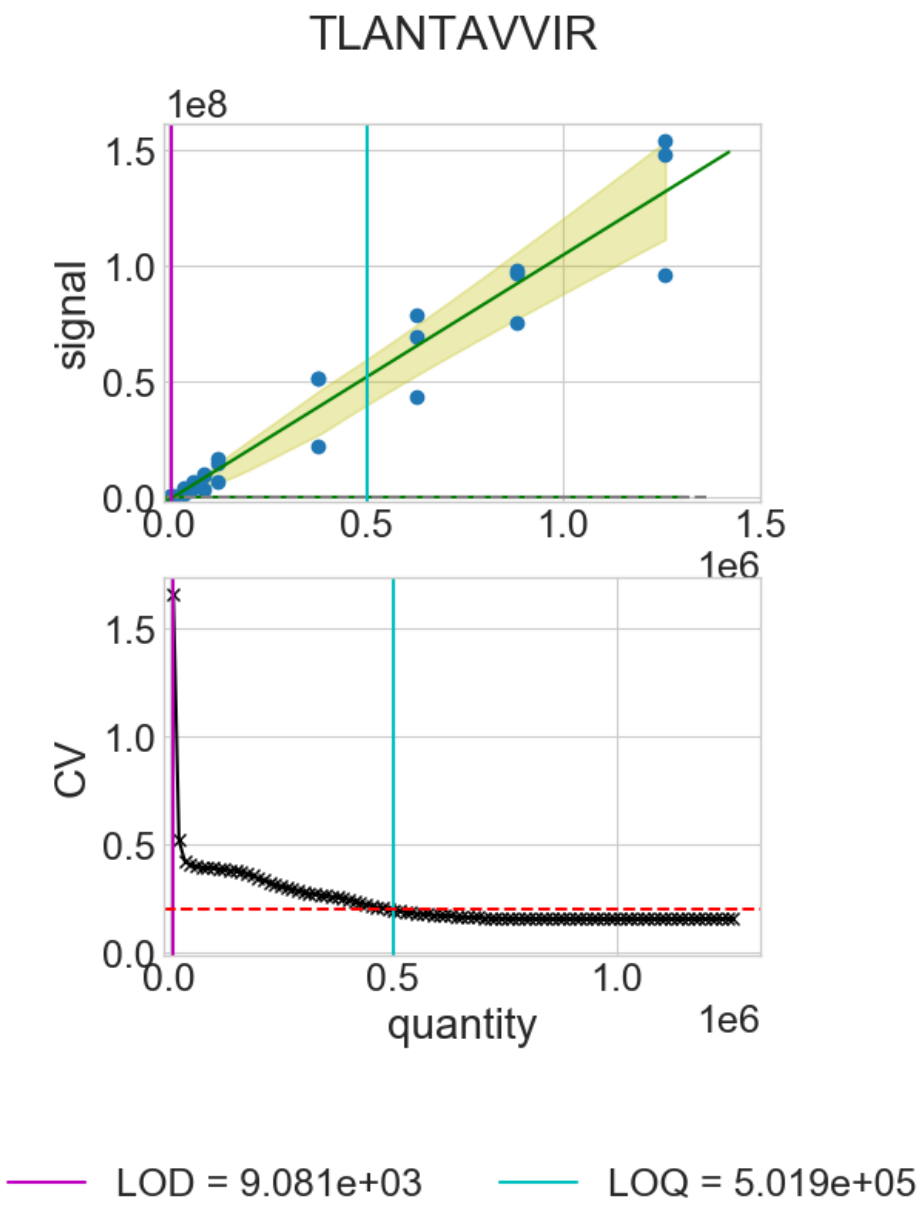

MYSYVVYR
1e7
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
0.75
0.50
0.25
0.00
CV
0.0
0.5
1.0
quantity
1e6
LOD = 6.530e+05
LOQ = inf

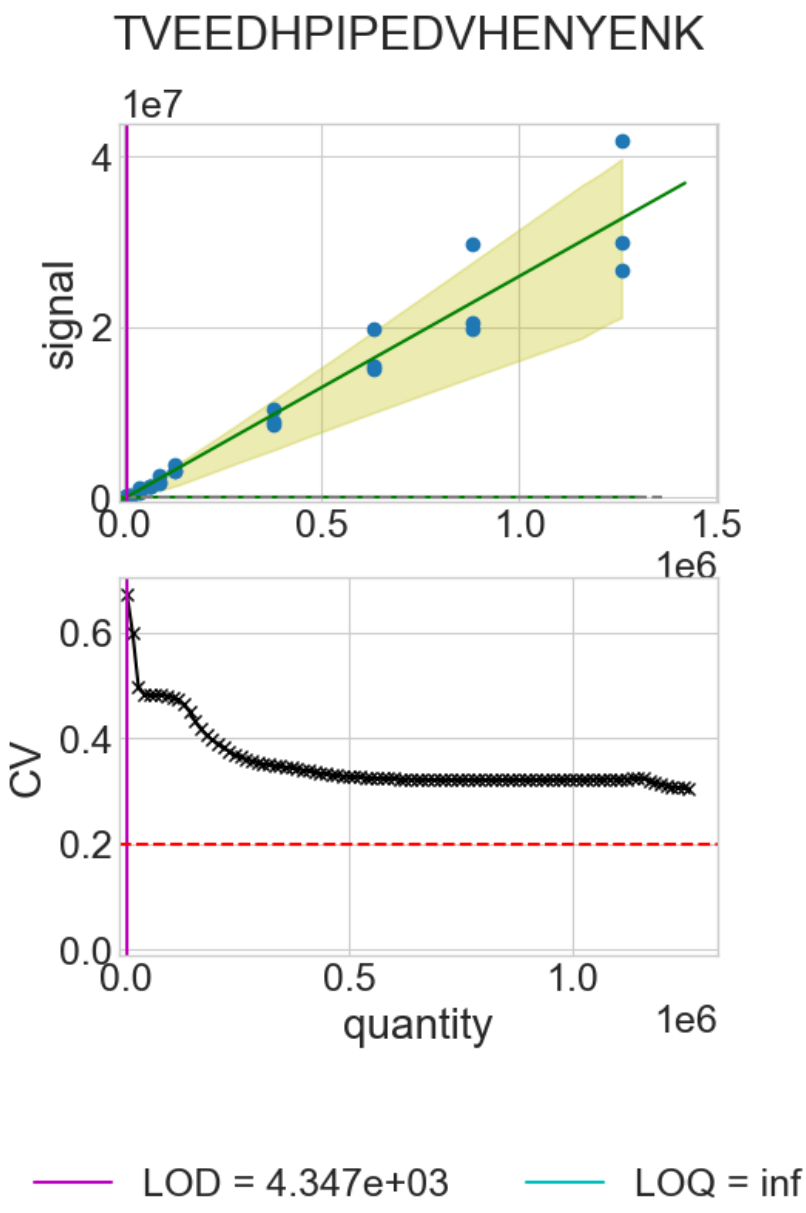

QLGLGTNIYNAER
1e8
signal
1.0
0.5
0.0
0.0
0.5
1.0
1.5
1e6
CV
0.6
0.4
0.2
0.0
0.0
0.5
1.0
quantity
1e6
LOD = 1.427e+05
LOQ = 5.716e+05

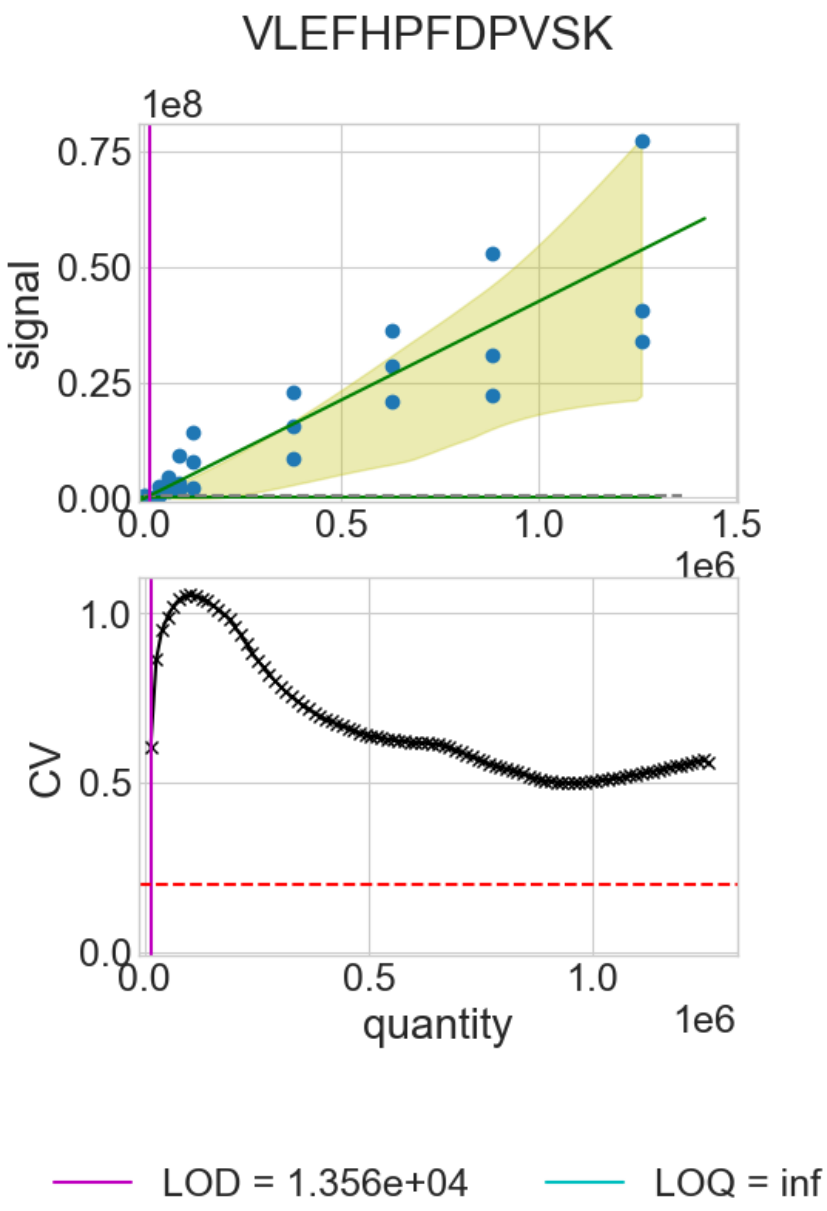

RGEGFMVVTATGDNTFVGR
1e5
3
2
1
0
signal
0.0
0.5
1.0
1e6
0.21
0.20
0.19
CV
−1.0
−0.5
0.0
quantity
1e4
LOD = nan
LOQ = inf

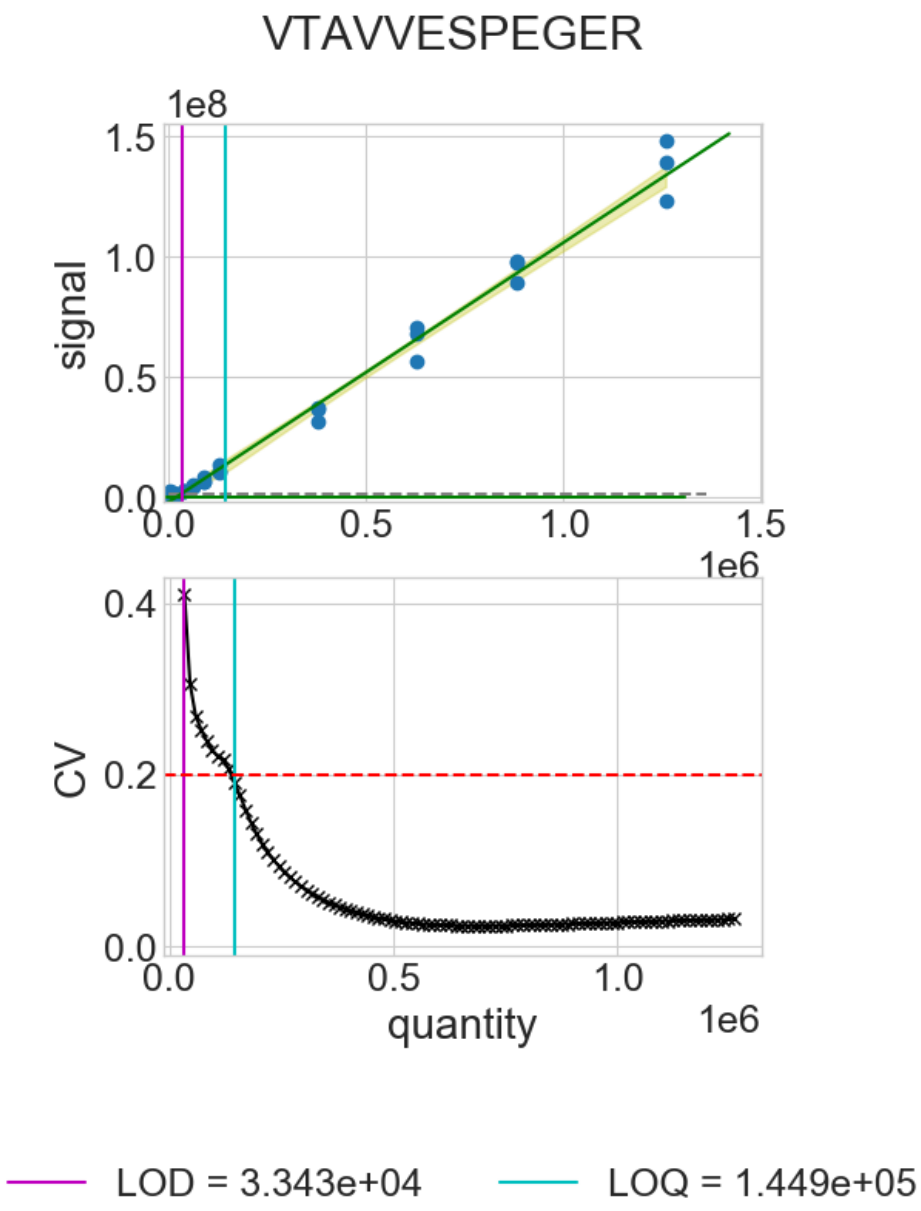

SAADIVFLAPGLSAIIDALK
1e7
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
2
1
0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 3.175e+05
LOQ = inf

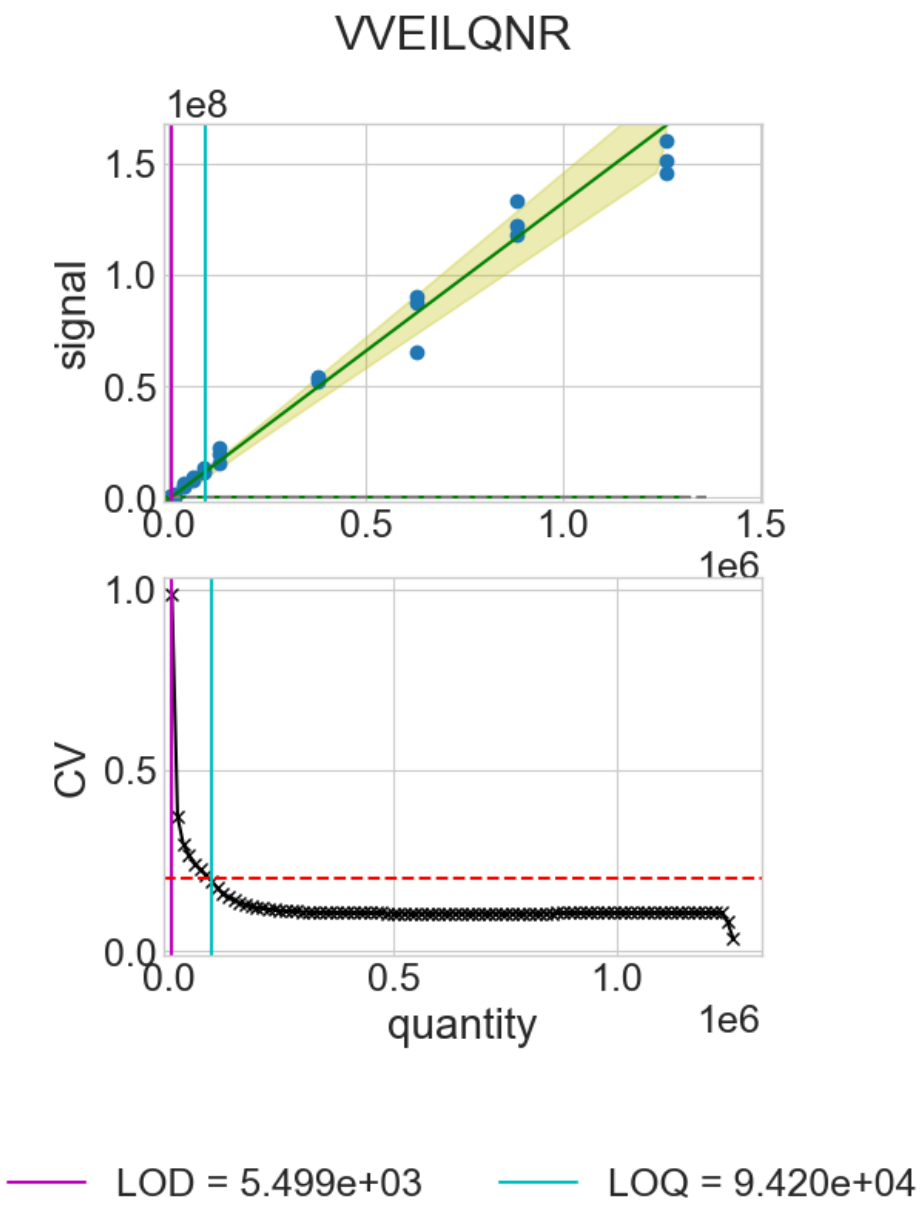

SVEDFMAAMQR
1e8
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
0.75
0.50
0.25
0.00
CV
0.0
0.5
1.0
quantity
1e6
LOD = 1.964e+05
LOQ = 3.683e+05

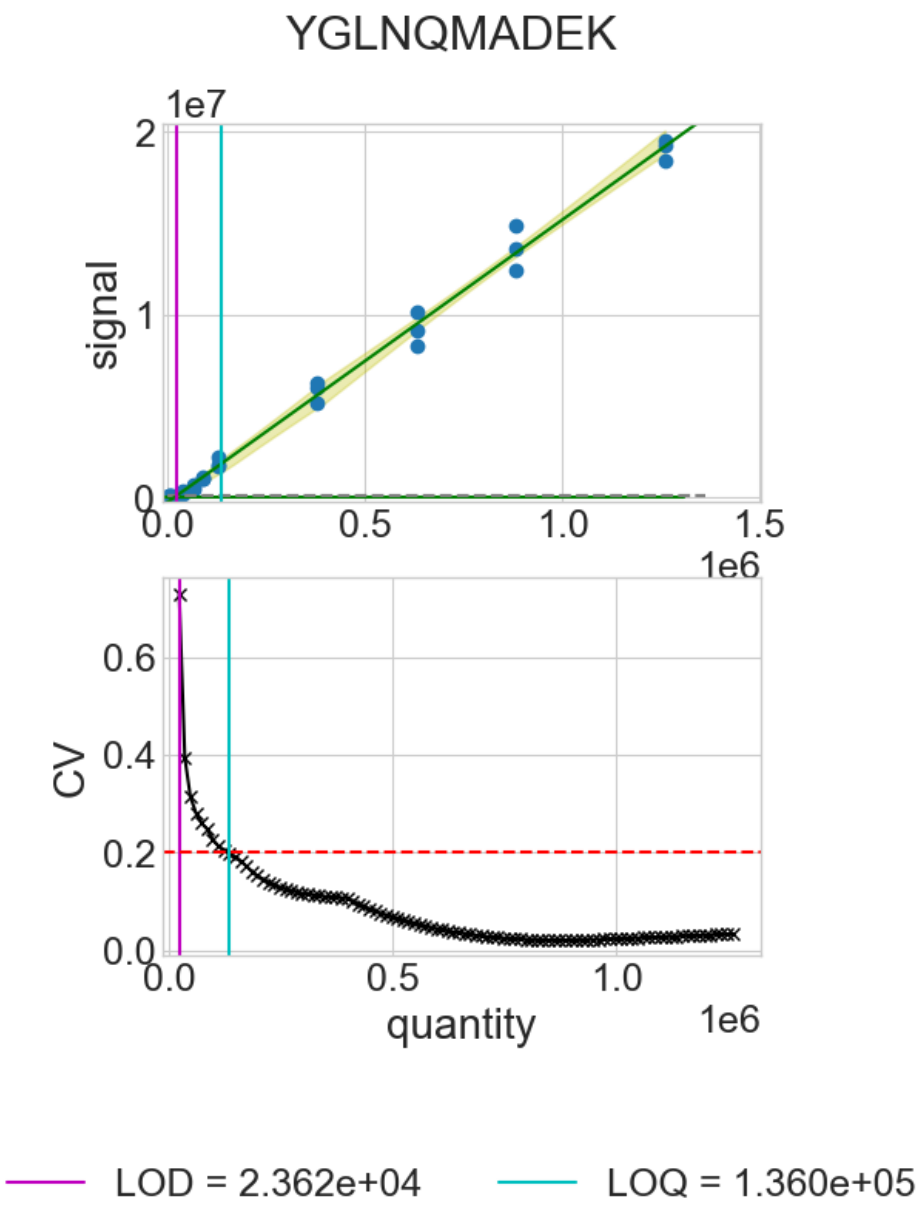

TLANTAVVIR
1e8
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
1.5
1.0
0.5
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 9.081e+03
LOQ = 5.019e+05

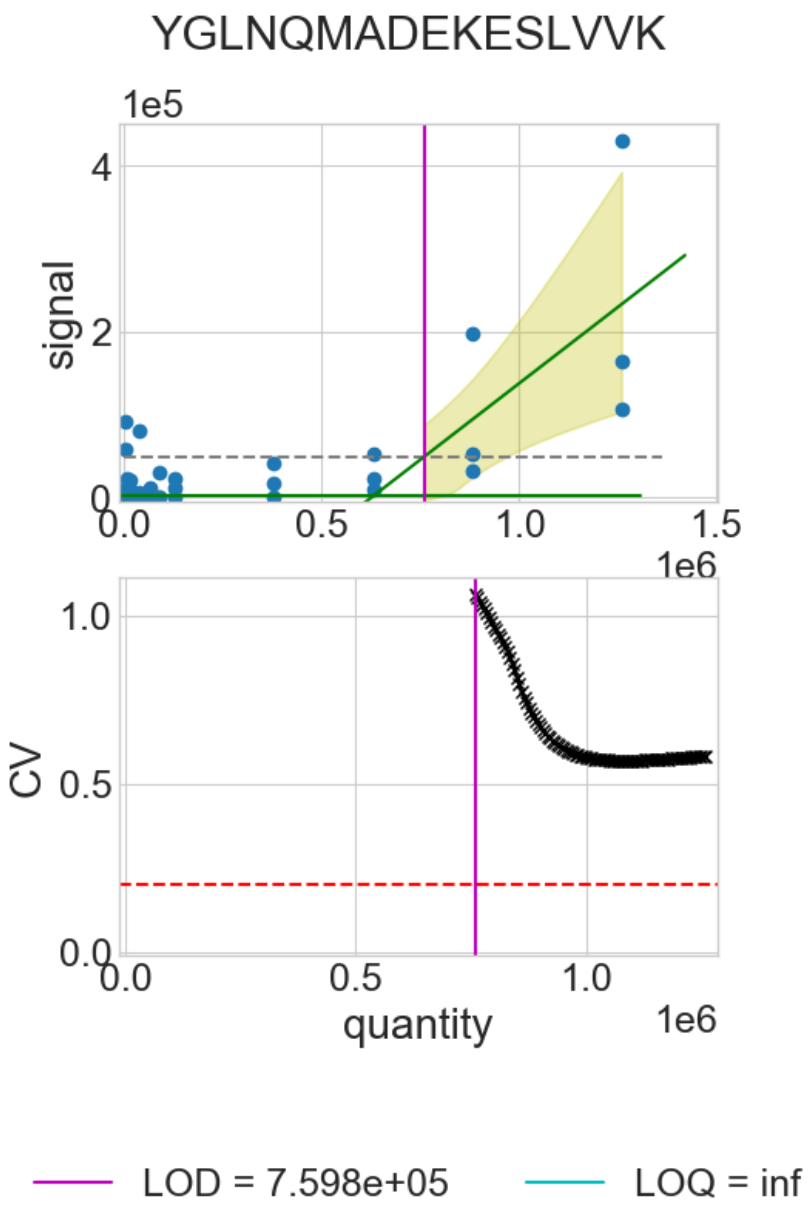

TVEEDHPIPEDVHENYENK
1e7
4
2
0
signal
0.0
0.5
1.0
1.5
1e6
0.6
0.4
0.2
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 4.347e+03
LOQ = inf


VLEFHPFDPVSK
1e8
0.75
0.50
0.25
0.00
signal
0.0
0.5
1.0
1.5
1e6
1.0
0.5
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 1.356e+04
LOQ = inf


VTAVVESPEGER
1e8
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
0.4
0.2
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 3.343e+04
LOQ = 1.449e+05


VVEILQNR
1e8
1.5
1.0
0.5
0.0
signal
0.0
0.5
1.0
1.5
1e6
1.0
0.5
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 5.499e+03
LOQ = 9.420e+04


YGLNQMADEK
1e7
2
1
0
signal
0.0
0.5
1.0
1.5
1e6
0.6
0.4
0.2
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 2.362e+04
LOQ = 1.360e+05


YGLNQMADEKESLVVK
1e5
4
2
0
signal
0.0
0.5
1.0
1.5
1e6
1.0
0.5
0.0
CV
0.0
0.5
1.0
quantity
1e6
LOD = 7.598e+05
LOQ = inf

Appendix D: STATISTICAL TESTING ANALYSES PERFORMED WITH MSSTATS

D.1 SOURCE CODE

Differential testing analyses in Chapter 4 were performed using peptide detections and integration information from EncyclopeDIA, quantitative peak areas calculated by Skyline, and protein-level summarization and statistical testing performed using MSstats (Choi et al., 2014). The following R markdown document includes all code used to perform the statistical testing in Chapter 4.

Aging Yeast K/O Differential Protein Analysis with MSstats

Pino LK

```
##
## 0. Prerequisites
##

# install dependency packages for MSstats from CRAN
install.packages(c("gplots","lme4","ggplot2","ggrepel","reshape","reshape2",
                   "data.table","Rcpp","survival","minpack.lm", "dplyr"))

# install dependency packages and MSstats from BioConductor
source("https://www.bioconductor.org/biocLite.R")
biocLite()
biocLite(c("limma","marray","preprocessCore","MSnbase", "MSstats"))

# load MSstats and dplyr
library(MSstats)
library(dplyr)


##
## 1. Preparing the data for MSstats input
##

# check the working directory (should be set to the project directory)
getwd()
setwd("G:/Shared drives/Lindsay Pino/proj/aging_yeast/gene-ko/results/2019summer")

## read in the Skyline export data
raw = read.csv('../../data/dia-ms/skylineexport_msstatsinput_20190612.csv')

## format the dataframe so that it plays nicely with MSstats
raw_msstats = SkylinetoMSstatsFormat(raw, filter_with_Qvalue = FALSE)


##
## 2. Data processing with the `dataProcess` function
##


# use the Tukey's median polish model for summarization
quant_tmp = dataProcess(raw = raw_msstats)


## 2.2 Visualization of data processing

# generate QC plots
#dataProcessPlots(data = quant_linear, type = "QCplot", address = 'Linear_')
```

```
# generate profile plots using the data summarized by linear mixed model
#dataProcessPlots(data = quant_linear, type = "Profileplot",
#                 width = 7, height = 7, address = "Linear_")

# generate the condition plots
#dataProcessPlots(data = quant_tmp, type = "conditionplot", address = "Linear_")


##
## 3. Group Comparison between all gene deletions and BY4741 control
##

# check unique conditions and check order of condition information
levels(quant_tmp$ProcessedData$GROUP_ORIGINAL)

# create a contrast matrices
# all pairwise comparisons against by4741
comparison1<-matrix(c(1,0,-1,0,0,0,0,0),nrow=1)
comparison2<-matrix(c(0,1,-1,0,0,0,0,0),nrow=1)
comparison3<-matrix(c(0,0,-1,1,0,0,0,0),nrow=1)
comparison4<-matrix(c(0,0,-1,0,1,0,0,0),nrow=1)
comparison5<-matrix(c(0,0,-1,0,0,1,0,0),nrow=1)
comparison6<-matrix(c(0,0,-1,0,0,0,1,0),nrow=1)
comparison7<-matrix(c(0,0,-1,0,0,0,0,1),nrow=1)

comparison<-rbind(comparison1,
                  comparison2,
                  comparison3,
                  comparison4,
                  comparison5,
                  comparison6,
                  comparison7)
row.names(comparison)<-c("ade17 vs by4741",
                         "adp1 vs by4741",
                         "extref vs by4741",
                         "idh2 vs by4741",
                         "sgf73 vs by4741",
                         "tor1 vs by4741",
                         "ubp8 vs by4741")

# perform the group comparison
gpcomp_tmp <- groupComparison(contrast.matrix = comparison, data = quant_tmp)
names(gpcomp_tmp$ComparisonResult)
head(gpcomp_tmp$ComparisonResult)
head(gpcomp_tmp$ModelQC)
head(gpcomp_tmp$fittedmodel)

# pull just the results out of the whole group comparison output
gpcomp_res <- gpcomp_tmp$ComparisonResult

# subset only significant proteins
list_sig <- gpcomp_res[gpcomp_res$adj.pvalue < 0.05
                       & abs(gpcomp_res$log2FC) > 2^1, ]
```

```
write.csv(list_sig, file="./msstats_results_sigpairwise.csv")
write.csv(gpcomp_res, file="./msstats_results_all.csv")
head(list_sig)
nrow(list_sig)


##
## 3.2 Visualization of differentially abundant proteins
##


# generate a volcano plot of the analyzed data
groupComparisonPlots(data = gpcomp_tmp$ComparisonResult,
                     type = 'VolcanoPlot',
                     sig = 0.05, FCcutoff = 2^1,
                     address = 'C:/Users/linds/Desktop/msstats_')

# generate a heatmap of the analyzed data
#groupComparisonPlots(data = gpcomp_tmp$ComparisonResult,
#                      type = 'Heatmap', sig = 0.05, FCcutoff = 2^1,
#                      address = 'C:/Users/linds/Desktop/heatmap_')


# generate a comparison plot
#groupComparisonPlots(data = gpcomp_tmp$ComparisonResult,
#                      type = 'ComparisonPlot',
#                      address = 'C:/Users/linds/Desktop/ComparisonPlot_')

##
## 3b. Comparison between RLS extending and nonextending
##

# all pairwise comparisons against by4741
comparison1b<-matrix(c(1,1,-1,0,0,0,0,0),nrow=1)
comparison2b<-matrix(c(0,0,-1,1,0,0,0,0),nrow=1)
comparison3b<-matrix(c(0,0,-1,0,1,0,1,0),nrow=1)
comparison4b<-matrix(c(0,0,-1,0,0,1,0,1),nrow=1)

comparisonb<-rbind(comparison1b, comparison2b, comparison3b, comparison4b)
row.names(comparisonb)<-c("ade17 and adp1 vs by4741",
                          "extref vs by4741",
                          "idh2 and tor1 vs by4741",
                          "sgf73 and ubp8 vs by4741")

# perform the group comparison
gpcomp_tmp_b <- groupComparison(contrast.matrix = comparisonb, data = quant_tmp)

# make the volcano plots
groupComparisonPlots(data = gpcomp_tmp_b$ComparisonResult,
                     type = 'VolcanoPlot',
                     sig = 0.05, FCcutoff = 2^1,
                     address = "C:/Users/linds/Desktop/msstats_b_")
```

```
##
## 3c. Comparison between RLS extending and everything else
##

# all pairwise comparisons against by4741
comparison1c<-matrix(c(-1,-1,-1,-1,1,1,1,1),nrow=1)

comparisonc<-rbind(comparison1c)
row.names(comparisonc)<-c("idh2 and tor1 and sgf73 and ubp8
                          vs ade17 and adp1 and by4741 and ext ref")

# perform the group comparison
gpcomp_tmp_c <- groupComparison(contrast.matrix = comparisonc, data = quant_tmp)

# make the volcano plots
groupComparisonPlots(data = gpcomp_tmp_c$ComparisonResult,
                     type = 'VolcanoPlot',
                     sig = 0.05, FCcutoff = 2^1,
                     address = "C:/Users/linds/Desktop/msstats_c_")

# pull just the results out of the whole group comparison output
gpcomp_c_res <- gpcomp_tmp_c$ComparisonResult

# subset only significant proteins
list_sig <- gpcomp_c_res[gpcomp_c_res$adj.pvalue < 0.05
                         & abs(gpcomp_c_res$log2FC) > 2^1, ]
write.csv(list_sig, file="C:/Users/linds/Desktop/msstats_results_c_sigpairwise.csv")
write.csv(gpcomp_c_res, file="C:/Users/linds/Desktop/msstats_results_c_all.csv")
head(list_sig)
nrow(list_sig)
```

D.2 SUMMARY OF SIGNIFICANTLY DIFFERENTIAL PROTEIN RESULTS

The first table includes the significant results of testing each genotype versus the control BY4741 strain. Differential proteins with an adjusted p-value of ≤ 0.5 and a log2 Fold Change (log2FC) ≥ 2 in any one of the single genotype pairwise comparisons are included. A blank cell for a given pairwise comparison indicates that the protein was not significantly differential in that test. Fold changes are reported as RLS-extending genotypes over control (e.g. a positive fold change value indicates increased abundance compared to the control, a negative fold change value indicates decreased abundance compared to the control). Column "RLS vs Control log2FC" indicates whether that protein was detected as significantly differential when all RLS-extending genotypes were compared together against all control genotypes. The second table includes significant results of testing all RLS-extending genotypes versus all control genotypes that were not found to be significant in the genotype . Differential proteins with an adjusted p-value of ≤ 0.5 and a log2 Fold Change (log2FC) ≥ 2 in any one of the single genotype pairwise comparisons are included. A blank cell for a given pairwise comparison indicates that the protein was not significantly differential in that test. Fold changes are reported as RLS-extending genotypes over control (e.g. a positive fold change value indicates increased abundance compared to the control, a negative fold change value indicates decreased abundance compared to the control). Column "RLS vs Control log2FC" indicates whether that protein was detected as significantly differential in the two-group comparison. The third table includes significant results of testing the four-group comparison: low RLS extending genotypes vs BY4741; high RLS extending genotypes vs BY4741; no RLS-extension genotypes vs BY4741; and a control external reference BY4741 vs the experimental control BY4741.

ORF	Symbol	Description	ade17 vs by4741 log2FC	adp1 vs by4741 log2FC	idh2 vs by4741 log2FC	tor1 vs by4741 log2FC	sgf73 vs by4741 log2FC	ubp8 vs by4741 log2FC	# Strains Significantly Differential	RLS vs Controls log2FC	extref vs by4741 log2FC
YGL160W	AIM14	Altered Inheritance rate of Mitochondria	Inf	Inf	Inf	Inf	Inf	Inf	6	Inf	Inf
YGR257C	MTM1	Manganese Trafficking factor for Mitochondrial SOD2		Inf	Inf	Inf	Inf	Inf	5	Inf	Inf
YKL138C-A	HSK3	Helper of ASK1		Inf	Inf	Inf	Inf	Inf	5	Inf	Inf
YCR063W	BUD31	BUD site selection	-Inf						1	Inf	
YJL196C	ELO1	ELOngation defective	-Inf						1	Inf	
YLR460C							-2.93		1	Inf	-Inf
YLR303W	MET17	METhionine requiring	6.48	6.63	6.94	6.28	7.07	6.27	6	13.06251533	
YEL021W	URA3	URAcil requiring				4.69	5.31		2	10.94798182	
YJR004C	SAG1	Sexual AGglutination	6.08	6.01	5.48	5.46	5.9	6.45	6	10.59044622	
YMR120C	ADE17	ADEnine	-5.57						1	6.969442601	
YMR096W	SNZ1	SNooZe					3.96		1	6.912125049	
YCL026C-B	HBN1	Homologous to Bacterial Nitroreductases					2.5		1	3.393016568	
YKL178C	STE3	STErile		2.15					1	2.983992525	
YDR033W	MRH1	Membrane protein Related to Hsp30p					-2.05		1	-2.717951282	
YBR092C	PHO3	PHOsphate metabolism					-2.44		1	-2.978975563	
YPL058C	PDR12	Pleiotropic Drug Resistance					-2.16		1	-3.193200423	
YOL126C	MDH2	Malate DeHydrogenase					-2.15		1	-3.633432429	
YOR136W	IDH2	Isocitrate DeHydrogenase			-5.31				1	-4.233102506	
YPL014W	CIP1	Cdk1 Interacting Protein					-4.07		1	-5.262152479	
YKL216W	URA1	URAcil requiring				-2.5	-2.67		2	-5.36173422	
YFL026W	STE2	STErile	-3.39	-3.34	-2.82	-2.89	-3.83	-2.93	6	-5.937186985	
YNL160W	YGP1	Yeast GlycoProtein					-5.48		1	-6.347576548	
YBR093C	PHO5	PHOsphate metabolism					-4.47		1	-7.939740531	
YBR115C	LYS2	LYSine requiring	-5.23	-4.7	-4.94	-5.08	-5.11	-4.9	6	-11.69299579	
YIL015W	BAR1	BARrier to the alpha factor response	-6.6	-7.59	-6.21	-7.1	-6.56	-6.31	6	-12.04708543	
YHR091C	MSR1	Mitochondrial tRNA Synthetase aRginine			-Inf		-Inf	-Inf	3	-Inf	
YJR127C	RSF2	ReSpiration Factor			-Inf	-Inf	-Inf		3	-Inf	
YNR011C	PRP2	Pre-mRNA Processing			-Inf	-Inf		-Inf	3	-Inf	
YDR478W	SNM1	Suppressor of Nuclear Mitochondrial endoribonuclease					-Inf	-Inf	2	-Inf	
YGR091W	PRP31	Pre-mRNA Processing			-Inf			-Inf	2	-Inf	
YOR359W	VTS1	VTi1-2 Suppressor				-Inf		-Inf	2	-Inf	
YGR241C	YAP1802	Yeast Assembly Polypeptide					-Inf		1	-Inf	
YDL194W	SNF3	Sucrose NonFermenting						-Inf	1	-Inf	
YDR306C								-Inf	1	-Inf	
YMR223W	UBP8	UBiquitin-specific processing Protease						-Inf	1	-Inf	
YMR310C								-Inf	1	-Inf	

YCR091W	KIN82	protein KINase				-Inf			1	-Inf			
YNR047W	FPK1	FliPase Kinase 1				-Inf			1	-Inf			
YPL193W	RSA1	RiboSome Assembly	Inf	Inf	Inf	Inf	Inf		5	NOT SIG			
YBL106C	SRO77	Suppressor of rho3	Inf			Inf	Inf	Inf	4	NOT SIG			
YHR152W	SPO12	SPOrulation	Inf		Inf	Inf		Inf	4	NOT SIG	Inf		
YNL116W	DMA2	Defective in Mitotic Arrest	-Inf		-Inf	-Inf		-Inf	4	NOT SIG			
YCR086W	CSM1	Chromosome Segregation in Meiosis	Inf			Inf	Inf		3	NOT SIG			
YNL238W	KEX2	Killer EXpression defective	-Inf				-Inf	-Inf	3	NOT SIG			
YJL155C	FBP26	Fructose BisPhosphatase	-Inf			-Inf		-Inf	3	NOT SIG			
YOR087W	YVC1	Yeast Vacuolar Conductance	-Inf			-Inf		-Inf	3	NOT SIG	-Inf		
YGL091C	NBP35	Nucleotide Binding Protein		-Inf			-Inf	-Inf	3	NOT SIG	-Inf		
YDL204W	RTN2	ReTiculoN-like		-Inf		-Inf	-Inf		3	NOT SIG	-Inf		
YOR201C	MRM1	Mitochondrial rRNA Methyltransferase	Inf		Inf				2	NOT SIG			
YLR052W	IES3	Ino Eighty Subunit	-Inf				-Inf		2	NOT SIG			
YBR054W	YRO2					-Inf	-Inf		2	NOT SIG	-Inf		
YDL036C	PUS9	PseudoUridine Synthase				-Inf		-Inf	2	NOT SIG	-Inf		
YDR180W	SCC2	Sister Chromatid Cohesion				-Inf		-Inf	2	NOT SIG	-Inf		
YCR011C	ADP1	ATP-Dependent Permease		-2.23					1	NOT SIG			
YFL014W	HSP12	Heat Shock Protein					-3.82		1	NOT SIG			
YHR087W	RTC3	Restriction of Telomere Capping					-2.91		1	NOT SIG			
YDL181W	INH1	INHibitor (of F1F0-ATPase)					-2.76		1	NOT SIG			
YNL015W	PBI2	Proteinase B Inhibitor					-2.73		1	NOT SIG			
Q0250	COX2	Cytochrome c OXidase					-2.56		1	NOT SIG			
YMR105C	PGM2	PhosphoGlucoMutase					-2.56		1	NOT SIG			
R0040C	REP2	REPlication					-2.5		1	NOT SIG			
YMR173W	DDR48	DNA Damage Responsive					-2.31		1	NOT SIG			
YBR268W	MRPL37	Mitochondrial Ribosomal Protein					-2.08		1	NOT SIG			
YLR437C	DIF1	Damage-regulated Import Facilitator					-2.02		1	NOT SIG			
YEL070W	DSF1	Deletion Suppressor of mptFive/pufFive mutation					3.2		1	NOT SIG			
YMR095C	SNO1	SNZ proximal Open reading frame					3.47		1	NOT SIG			
YLR072W	LAM6	Lipid transfer protein Anchored at Membrane contact site						-2.72	1	NOT SIG			
YBR091C	TIM12	Translocase of the Inner Membrane						-2.52	1	NOT SIG			
YPR075C	OPY2	Overproduction-induced Pheromone-resistant Yeast						-Inf	1	NOT SIG	-Inf		

Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	issue	MissingPercentage	ImputationPercentage
YGL225W	idh2 and tc	Inf	NA	NA	NA	NA	0	oneConditi	0.675675676	0
YOR188W	idh2 and tc	Inf	NA	NA	NA	NA	0	oneConditi	0.513513514	0
YJL218W	idh2 and tc	4.670821	1.116792	4.182356	24	0.000332	0.018520537	NA	0	0
YBR111W-A	idh2 and tc	4.392098	1.213966	3.617974	24	0.001375	0.04358991	NA	0	0
YIR031C	idh2 and tc	3.937797	1.083031	3.635904	24	0.001315	0.042382867	NA	0	0
YGR287C	idh2 and tc	3.498949	0.668159	5.236698	24	2.28E-05	0.002867435	NA	0	0
YDR277C	idh2 and tc	3.403135	0.943712	3.606117	24	0.001416	0.044030396	NA	0	0
YDL203C	idh2 and tc	3.379864	0.920814	3.670517	24	0.001206	0.04140425	NA	0	0
YHR029C	idh2 and tc	3.147576	0.365456	8.612732	24	8.34E-09	2.93E-06	NA	0	0
YJR055W	idh2 and tc	3.083978	0.825049	3.737934	21	0.001214	0.04140425	NA	0.081081081	0
YLR304C	idh2 and tc	2.940859	0.25855	11.37443	24	3.75E-11	2.20E-08	NA	0	0
YAL053W	idh2 and tc	2.848176	0.666744	4.271767	24	0.000265	0.015760386	NA	0	0
YGR247W	idh2 and tc	2.75301	0.532101	5.173854	24	2.68E-05	0.003189599	NA	0	0
YML101C	idh2 and tc	2.733788	0.699422	3.90864	24	0.000664	0.029145546	NA	0	0
YCR005C	idh2 and tc	2.685287	0.722871	3.714754	24	0.00108	0.040078674	NA	0	0
YBR137W	idh2 and tc	2.496216	0.696044	3.586292	24	0.001487	0.045451366	NA	0	0
YKL059C	idh2 and tc	2.438282	0.594439	4.101823	24	0.000407	0.022016234	NA	0	0
YNL037C	idh2 and tc	2.253731	0.296787	7.593767	24	7.84E-08	2.30E-05	NA	0	0
YGL086W	idh2 and tc	2.215927	0.4064	5.452584	24	1.33E-05	0.001943773	NA	0.004324324	0.004324324
YGL039W	idh2 and tc	2.126386	0.499605	4.256132	24	0.000275	0.016125546	NA	0.007475561	0.007475561
YHR138C	idh2 and tc	2.095618	0.394078	5.317768	24	1.86E-05	0.002517523	NA	0	0
YKR082W	idh2 and tc	-2.011792	0.552879	-3.638754	24	0.001305	0.042382867	NA	0	0
YPL217C	idh2 and tc	-2.093632	0.564241	-3.710526	24	0.001091	0.040078674	NA	0	0
YER088C	idh2 and tc	-2.295745	0.581561	-3.947559	24	0.000602	0.028130834	NA	0.001228501	0.001228501
YLR258W	idh2 and tc	-2.335125	0.478609	-4.87898	24	5.65E-05	0.005366404	NA	0	0
YGR245C	idh2 and tc	-2.359911	0.579866	-4.069751	24	0.000442	0.023126364	NA	0.014054054	0.014054054
YDL124W	idh2 and tc	-2.680577	0.374764	-7.152713	24	2.16E-07	5.33E-05	NA	0.006538797	0.006538797
YJL096W	idh2 and tc	-2.938633	0.682584	-4.305161	24	0.000243	0.015253991	NA	0	0
YMR145C	idh2 and tc	-2.995707	0.664402	-4.508879	24	0.000145	0.010176584	NA	0	0
YBR218C	idh2 and tc	-3.007831	0.83332	-3.609454	24	0.001404	0.044030396	NA	0.015135135	0.015135135
YDL108W	idh2 and tc	-3.102512	0.848133	-3.658051	24	0.001244	0.042034031	NA	0	0
YOL143C	idh2 and tc	-3.437613	0.45162	-7.61173	24	7.53E-08	2.30E-05	NA	0	0
YDR359C	idh2 and tc	-3.571656	0.967486	-3.691689	24	0.001144	0.040823671	NA	0	0
YNL214W	idh2 and tc	-5.775147	1.223631	-4.719678	16	0.000231	0.01478027	NA	0.216216216	0
YMR258C	idh2 and tc	-5.969484	1.389654	-4.295662	20	0.000352	0.019349957	NA	0.108108108	0
YKL064W	idh2 and tc	-10.9991	2.755647	-3.991476	21	0.000663	0.029145546	NA	0.189189189	0.108108108

ORF	Symbol	Name	ade17 and adp1 vs by4741 log2FC	idh2 and tor1 vs by4741 log2FC	sgf73 and ubp8 vs by4741 log2FC	extref vs by4741 log2FC
YNL160W	YGP1	Yeast GlycoProtein		-3.13	-6.31	
YDL181W	INH1	INHibitor (of F1F-ATPase)		-5.73	-4.45	
YMR105C	PGM2	PhosphoGlucoMutase		-3.15	-3.13	
YKL216W	URA1	URAcil requiring		-2.86	-2.86	
YML120C	NDI1	NADH Dehydrogenase Internal		-3.58	-2.37	
YNR001C	CIT1	CITrate synthase		-2.05	-2.31	
YGL187C	COX4	Cytochrome c OXidase		-2.33	-2.04	
YGR287C	IMA1	IsoMAltase		2.00	2.30	
YBR111W-A	SUS1	Sl gene Upstream of ySa1		3.36	3.14	
YJL218W				3.02	3.27	
YEL021W	URA3	URAcil requiring		4.99	5.85	
YMR120C	ADE17	ADEnine		5.57	6.26	
YGR091W	PRP31	Pre-mRNA Processing		-Inf	-Inf	
YHR091C	MSR1	Mitochondrial tRNA Synthetase aRginine		-Inf	-Inf	
YJR127C	RSF2	ReSpiration Factor		-Inf	-Inf	
YNR011C	PRP2	Pre-mRNA Processing		-Inf	-Inf	
YOR359W	VTS1	VTi1-2 Suppressor		-Inf	-Inf	
YBR093C	PHO5	PHOsphate metabolism			-6.00	
YMR258C	ROY1	Repressor Of Ypt52			-4.41	
R0040C	REP2	REPlication			-3.34	
YLR072W	LAM6	Lipid transfer protein Anchored at Membrane contact site			-3.25	
YNL015W	PBI2	Proteinase B Inhibitor			-3.21	
YDL128W	VCX1	VaCuolar H+/Ca2+ eXchanger			-3.17	
YPL014W	CIP1	Cdk1 Interacting Protein			-3.05	
YOR120W	GCY1	Galactose-inducible Crystallin-like Yeast protein			-2.71	
YBR092C	PHO3	PHOsphate metabolism			-2.69	
YOL126C	MDH2	Malate DeHydrogenase			-2.57	
YDL124W					-2.42	
YPL058C	PDR12	Pleiotropic Drug Resistance			-2.30	
YDR033W	MRH1	Membrane protein Related to Hsp3p			-2.08	
YLR364W	GRX8	GlutaRedoXin			2.21	
YCL047C	POF1	Promoter Of Filamentation			2.23	
YIL002W-A	CMI7	Cytosolic MIni protein of ~7 kDa			2.30	

YER042W	MXR1	peptide Methionine sulfoXide Reductase		2.40
YNL259C	ATX1	AnTioXidant		2.78
YEL070W	DSF1	Deletion Suppressor of mptFive/pufFive mutation		3.36
YCL026C-B	HBN1	Homologous to Bacterial Nitroreductases		3.58
YMR096W	SNZ1	SNooZe		3.95
YMR095C	SNO1	SNZ proximal Open reading frame		4.34
YBL106C	SRO77	Suppressor of rho3		Inf
YDL194W	SNF3	Sucrose NonFermenting		-Inf
YDR306C				-Inf
YDR478W	SNM1	Suppressor of Nuclear Mitochondrial endoribonuclease		-Inf
YGR241C	YAP182	Yeast Assembly Polypeptide		-Inf
YMR223W	UBP8	UBiquitin-specific processing Protease		-Inf
YMR310C				-Inf
YNL117W	MLS1	MaLate Synthase	-9.08	
YLR174W	IDP2	Isocitrate Dehydrogenase	-6.26	
YNR034W-A	EGO4	Exit from rapamycin-induced GrOwth arrest	-5.77	
YIL136W	OM45	Outer Membrane	-5.46	
YBR072W	HSP26	Heat Shock Protein	-5.19	
YOR136W	IDH2	Isocitrate DeHydrogenase	-4.92	
YMR250W	GAD1	GlutAmate Decarboxylase	-4.68	
YPR160W	GPH1	Glycogen PHosphorylase	-4.53	
YOR065W	CYT1	CYTochrome c1	-4.12	
YPR010C-A	MIN8	mitochondrial MINi protein of 8 kDa	-4.09	
YHR137W	ARO9	AROmatic amino acid requiring	-3.93	
Q0250	COX2	Cytochrome c OXidase	-3.88	
YDR529C	QCR7	ubiQuinol-cytochrome C oxidoReductase	-3.66	
YBL015W	ACH1	Acetyl CoA Hydrolase	-3.63	
YER103W	SSA4	Stress-Seventy subfamily A	-3.43	
YJL052W	TDH1	Triose-phosphate DeHydrogenase	-3.42	
YDR380W	ARO1	AROmatic amino acid requiring	-3.30	
YLR038C	COX12	Cytochrome c OXidase	-3.17	
YFR053C	HXK1	HeXoKinase	-2.81	
YOR374W	ALD4	ALdehyde Dehydrogenase	-2.65	
YEL024W	RIP1	Rieske Iron-sulfur Protein	-2.59	
YBR230C	OM14	Outer Membrane Protein of 14 kDa	-2.56	

YNL100W	MIC27	MItochondrial contact site and Cristae organizing system	-2.45
YBR268W	MRPL37	Mitochondrial Ribosomal Protein	-2.27
YDR322C-A	TIM11	Translocase of the Inner Mitochondrial membrane	-2.05
YGR244C	LSC2	Ligase of Succinyl-CoA	-2.05
YMR145C	NDE1	NADH Dehydrogenase	-2.03
YCR091W	KIN82	protein KINase	-Inf
YNR047W	FPK1	FliPase Kinase 1	-Inf

VITA

Lindsay K. Pino was born in Belleville, Illinois but moved frequently throughout her childhood as an Air Force brat, including South Korea, Florida, Virginia, and New Hampshire. She graduated from the Pennsylvania State University with a B.S. in biochemistry and molecular biology and a minor in microbiology. After graduating, she went to South Korea to teach and study abroad on a Fulbright scholarship. She repatriated in 2011 and rejoined the scientific community, working as a research associate at the Broad Institute in Boston, first in the Genomics Platform and then in the Proteomics Platform, where she discovered the scientific interests that drove her to pursue her doctorate. Outside of the lab, Lindsay likes to run, hike, and explore the Pacific Northwest with her shelter dog, Maizie; and stress-bakes an admirable macaron.