Development of Data Independent Acquisition Methods to Systematically Analyze the Human Proteome

Brian C. Searle

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Michael J. MacCoss, Chair

Judit Villén

William S. Noble

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Development of Data Independent Acquisition Methods to Systematically Analyze the Human Proteome

Brian C. Searle

Chair of the Supervisory Committee:
Michael J. MacCoss
Genome Sciences

Data independent acquisition (DIA) mass spectrometry is a powerful technique that is improving the reproducibility and throughput of mass spectrometry-based proteomics studies. Here we explore several new approaches to leverage this technology. First we present an overview of modern data independent acquisition techniques, and demonstrate their internal detection rate and quantification consistency relative to targeted parallel reaction monitoring (PRM) and data dependent acquisition (DDA)

methods despite large variations in data sampling strategies. Second, we use DIA experiments to construct a prediction model that helps determine optimal peptide selection for targeted experiments. Third, we introduce a new experimental workflow that uses chromatogram libraries to enable sensitive peptide detection in quantitative experiments, while still maintaining the throughput necessary for large scale experiments. Finally, we discuss a new approach to statistically validate phosphopeptide positional isomers and explore their prevalence in studies of Human cells.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF APPENDIX FIGURES

TABLE OF TABLES

ACKNOWLEDGEMENTS

This work is the result of several collaborations. First, I would like to thank my coauthors on portions of this material: Jim Bollinger, Jarrett Egertson, Rob Lawrence, Mike MacCoss, Lindsay Pino, Andrew Stergachis, Sonia Ting, Judit Villén, and Han-Yin Yang. All of these people have helped this work in immeasurable ways. Specifically, this work would not have been possible without the mentorship of Mike MacCoss. Mike has been an inspiration to me, and my time in his lab has profoundly affected how I model the lab I hope to run in the future. I also appreciate the mentorship of Judit Villén and Josh Akey, who have coached me both scientifically and personally. I am grateful for the guidance of my other thesis committee members, Bill Noble, Shao-En Ong, and John Scott.

I thank my lab mates, but especially Lindsay Pino and Han-Yin Yang for willing to be a constant sounding board for my ideas, both good and bad. I would also like to thank Rob Lawrence for teaching me everything I know about bench work, and Jarrett Egertson, Rich Johnson, and Sonia Ting for teaching me everything I know about running a mass spectrometer.

I am indebted to my parents, Scott and Lai-Kum Searle, for their constant support throughout this experience.

Finally, this work was financially supported by the NIH, specifically F31 GM119273.

DEDICATION

This work is dedicated to my pal, Oliver.

1 CURRENT PERSPECTIVES ON DATA INDEPENDENT ACQUISITION

1.1 Introduction

Shotgun proteomics(1) is an increasingly powerful tool for studying human biology and disease(2, 3). With shotgun proteomics, enzymatically digested peptides are separated by high pressure liquid chromatography (LC) and subsequently analyzed with tandem mass spectrometry (MS/MS). Peptides are detected from peptide-specific precursor and fragment ions, and bioinformatics tools help assemble those detections into protein inferences. At any given retention time the intensities of those mass-to-charge (m/z) measurements generally scale with peptide abundance, and those intensities can be integrated over time to quantify peptides relative to a control measurement. The differences between MS/MS acquisition methods represent tradeoffs in assay speed, reproducibility, and depth of proteome coverage.

By far the most common approach to large-scale proteomic profiling capitalizes on data dependent acquisition (DDA), where mass spectrometers collect MS survey spectra that are interspersed with triggered MS/MS spectra for the top N most abundant precursor ions(4). These MS/MS spectra are collected on fragment ions produced by a narrow m/z window surrounding a precursor ion in an attempt to isolate a single peptide at a time. By focusing only on regions of the m/z space that contain peptide signals found in the MS survey spectra, DDA can produce deep proteome coverage(5, 6) often within a single acquisition(7). However, one disadvantage caused by selecting peptides for MS/MS

based on the precursor signal for each acquisition is that minor variations between samples and acquisition timing can lead to different peptides getting selected for MS/MS acquisition between different runs in an experiment. DDA dynamically excludes peptides from being measured multiple times across the peak elution. This approach increases the number of unique peptide detections but due to the lack of repeated measurements, peak area-based quantification can only be performed using MS1 survey spectra.

In contrast, targeted methods such as selected reaction monitoring(8) (SRM) and parallel reaction monitoring(9) (PRM) forgo profiling the entire proteome and instead specifically measuring representative peptides of interest. Rather than rely on a precursor signal, SRM and PRM methods systematically acquire MS/MS spectra irrespective of whether a precursor signal is or is not detected. Instrument cycle times are structured such that the same precursor isolation window is acquired several times across the elution of the peptide, enabling more selective quantitative measurements using MS/MS spectra of fragment ions instead of just MS survey spectra of precursor ions. However, with these methods sensitivity (with SRM) and selectivity (with PRM) is inversely proportional to the dwell/ion accumulation time, which limits the number of peptides that can be tracked. In general these measurements are scheduled to specific retention time windows to achieve higher throughput and multiplexed across multiple acquisitions such that hundreds or thousands of peptides can be measured in a single study(10).

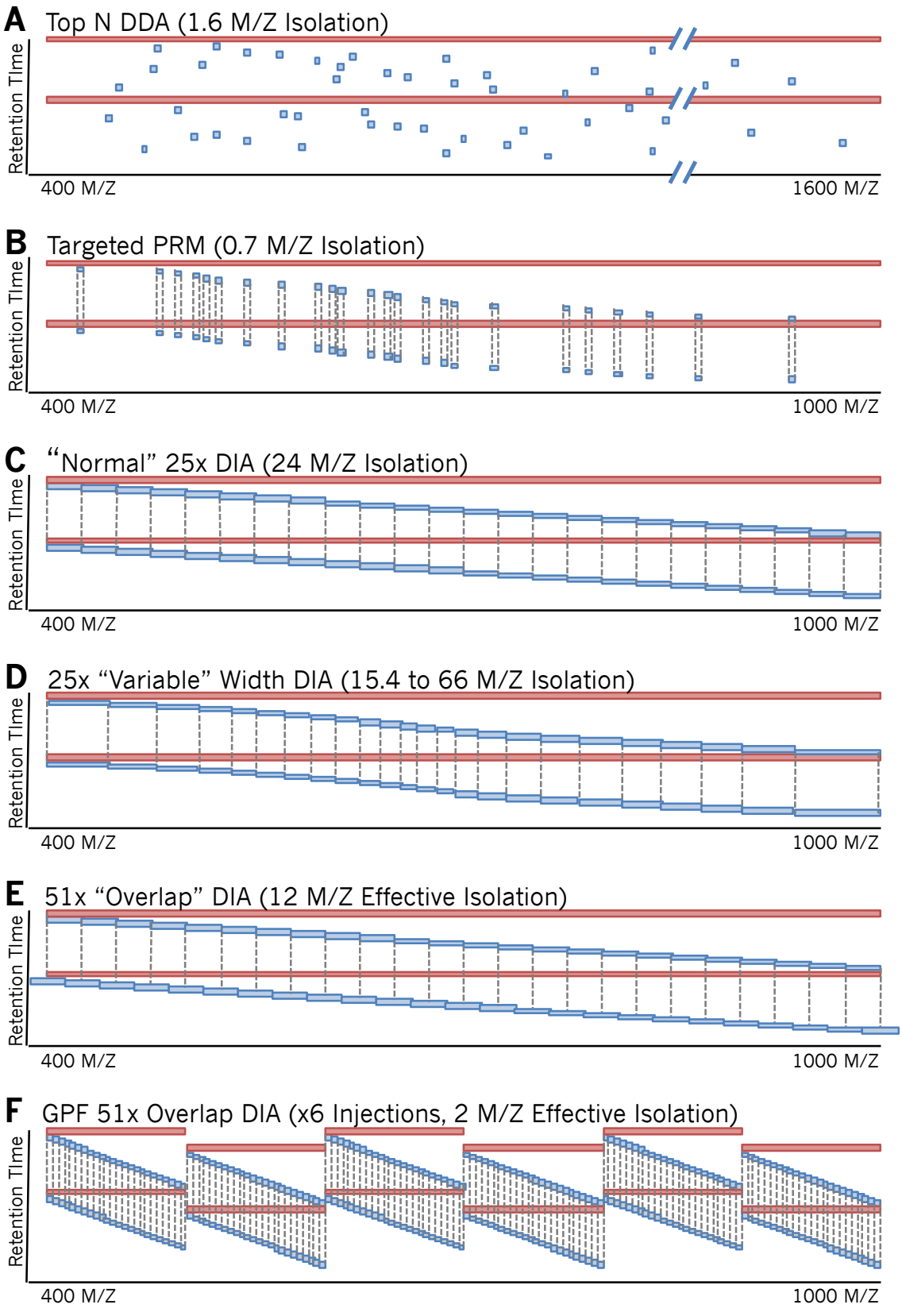

A
Top N DDA (1.6 M/Z Isolation)
Retention TIme
400 M/Z
1600 M/Z
B
Targeted PRM (0.7 M/Z Isolation)
Retention TIme
400 M/Z
1000 M/Z
C
“Normal” 25x DIA (24 M/Z Isolation)
Retention TIme
400 M/Z
1000 M/Z
D
25x “Variable” Width DIA (15.4 to 66 M/Z Isolation)
Retention TIme
400 M/Z
1000 M/Z
E
51x “Overlap” DIA (12 M/Z Effective Isolation)
Retention TIme
400 M/Z
1000 M/Z
F
GPF 51x Overlap DIA (x6 Injections, 2 M/Z Effective Isolation)
Retention TIme
400 M/Z
1000 M/Z

Figure 1-1: High resolution proteomics data acquisition methods.
(A) DDA methods collect MS scans (red bar) that span the entire m/z range of interest (e.g. 400 to 1600 m/z), followed by several MS/MS spectra (blue bars) with narrow precursor isolation selected from the most abundant peaks in the previous MS spectrum. (B) PRM methods collect repeated MS/MS spectra with narrow precursor isolation based on an inclusion list of target peptides only, regardless of precursor intensity. (C) Normal DIA tiles wide-window MS/MS spectra across the entire m/z range of interest at even intervals. (D) Variable-width DIA scales the MS/MS precursor isolation window widths to the expected number of peptides or precursors in each window. (E) Overlapped DIA acquires data similar to Normal DIA, but shifts every other cycle by half a window size, enabling computational deconvolution between cycles to cut the effective precursor isolation in half. (F) The gas phase fractionated DIA approach collects six runs for each biological sample, where the individual runs are tiled across different m/z ranges of interest but have precursor isolation widths equivalent to DDA or PRM.

Data independent acquisition(11) (DIA) is an alternative approach to MS/MS data acquisition that represents a compromise between PRM and DDA (Figure 1-1A-C). Like DDA, DIA attempts to measure signals from as many peptides in a proteome as possible, while still maintaining the systematic acquisition of MS/MS spectra found in PRM experiments. Originally Venable et al envisi`oned 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 instruments can collect

MS/MS at 10 Hz or faster and allow for PRM-like quantitation using MS/MS spectra by assuming some limitations.

For whole proteome analysis DIA is limited by compromises in several ways. First, in order to collect quantitative MS/MS for as many peptides as possible in a proteome, the MS/MS precursor isolation width is widened such that each spectrum often contains fragment ions from multiple coeluting peptides, and these signals must be deconvoluted to ensure accurate peptide detection and quantification. Second, the number of required measurements scales with the precursor m/z range, so some peptides must be monitored at suboptimal charge states. Third, unlike with DDA where fragmentation collision energies can be tuned specifically for an individual precursor's mass and charge, precursor isolation windows in DIA contain multiple peptides at varying charge states; consequently, collision energies generally are only optimized for peptides of one charge state. Despite these concessions, the reproducibility of DIA is a major advantage when compared to DDA for monitoring proteomic changes across tens or hundreds of samples. In addition, breadth of DIA measurements unlocks an important benefit over PRM acquisition: the possibility to iteratively re-interpret MS/MS experiments when new hypotheses are derived.

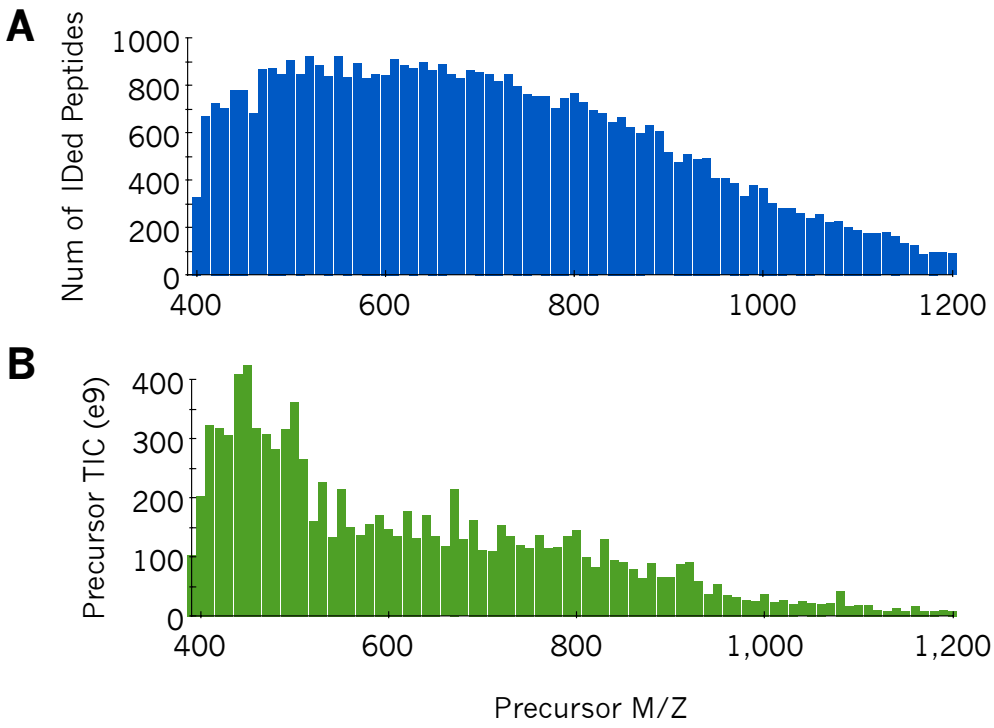


Figure 1-2: Histograms showing the number of detected peptides and the precursor integrated precursor intensity from 400 to 1200 m/z.

The number of peptides are shown in (**A**) while the integrated precursor intensity is shown in (**B**). In general a lower percentage of peptides are detected (and lower ion intensity measured) above 1000 m/z. Consequently this study focuses on the 400 to 1000 m/z range.

Several strategies have been developed to combat the most significant limitation: wide precursor isolation windows. Instead of using consistent window sizes, Zhang et al(12) proposed variable-width windows (Figure 1-1D) that scale depending on the number of expected peptides in each window. This strategy leverages the fact that the number of both detectable and identifiable precursors varies dramatically across the m/z space since most peptides fall in a relatively narrow mass and charge range (Figure 1-2). Egertson et al(13) developed a multiplexing approach (MSX) to randomly mix narrow precursor windows using multiple notched waveforms(14) in the precursor isolation

quadrupole (Q1) and computationally demultiplex fragment ions after the acquisition. While this approach can only practically be performed on select instruments, a variant of this approach uses overlapped windows (Figure 1-1E) that repeat every other cycle. The overlapping windows can be performed on most DIA-capable instrumentation and the results can be deconvoluted to reduce the precursor isolation window width by half (Figure 1-3). Finally, the PAcIFIC(15) acquisition uses gas-phase fractionation (GPF) (Figure 1-1F) to narrow DIA precursor isolation windows at the expense of multiple injections. While this strategy is not practical for large quantitative experiments, the approach can be used to generate chromatogram libraries that can improve detection rates when searching wide-window DIA experiments(16).

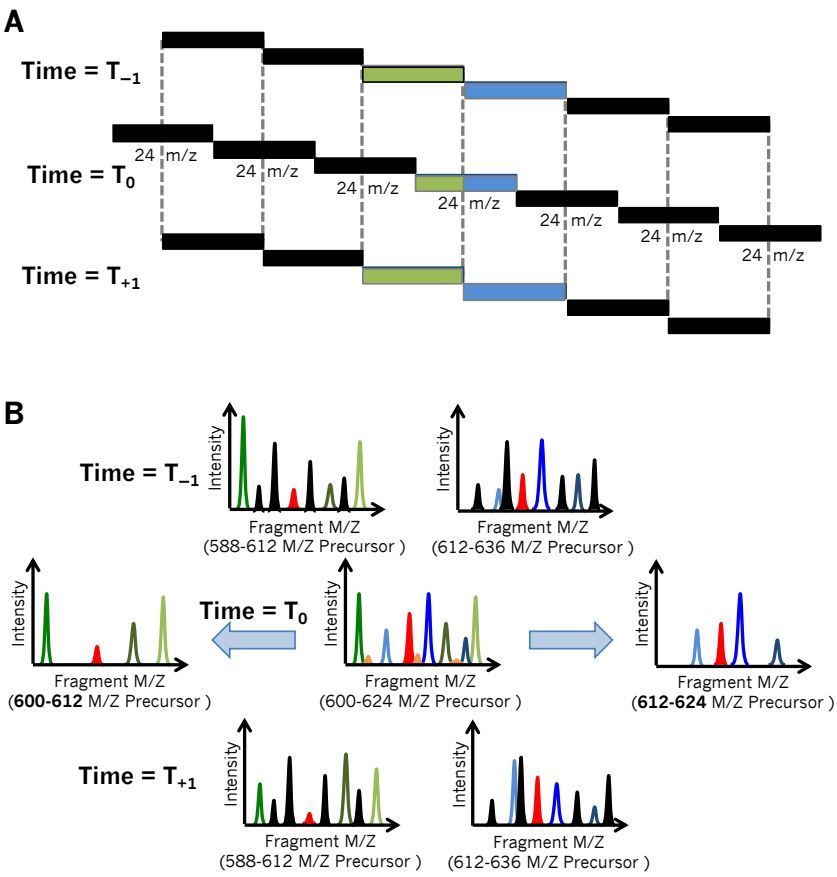


Figure 1-3: Schematic for overlapping window deconvolution.

(A) A schematic showing 3 overlapped DIA cycles where bars represent the precursor isolation window at staggered times. Precursor isolation windows are set to 24 m/z and overlapped by 12 m/z. (B) A schematic for representative fragmentation spectra from the red isolation windows. When deconvoluting the middle fragmentation spectrum (600-624 m/z, T_0), we refer to overlapping spectra from the previous and next spectra. Fragment ions in green are present in the lower precursor isolation window fragmentation spectra (588-612 m/z) for T_{-1} and T_{+1}, while fragment ions in blue are present in the upper precursor isolation window fragmentation spectra (612-636 m/z). Red fragmentation ions are present in both sets of windows, while orange fragmentation ions are likely to be noise because they are not present in any of the T_{-1} or T_{+1} spectra. For each ion in the 600-624 m/z, T_0 spectrum, we calculate the relative intensity proportion that that ion belonged to 600-612 m/z and 612-624 m/z.

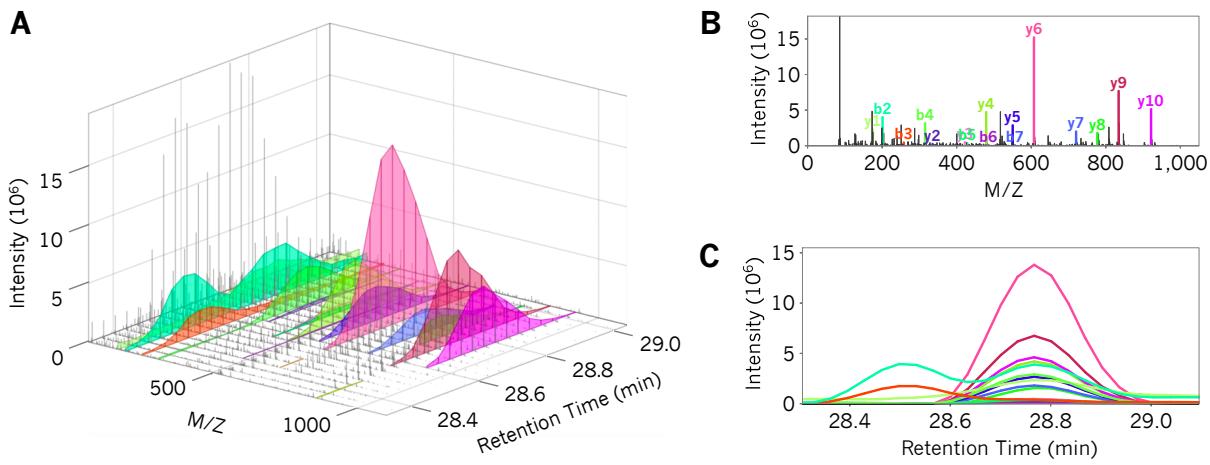


Figure 1-4: Interpretation of DIA peptide fragmentation. DIA acquires repeated fragmentation scans with the same precursor isolation window. (A) Sequential MS/MS scans across retention time for the precursor isolation window (478.1 to 497.4 m/z) corresponding to the +2H LSGGLGAGSCR peptide (489.2453 m/z). B/Y +1H fragment ions for this peptide are colored chromatograms, while ions not associated with the peptide are gray. (B) The individual MS/MS scan at the apex of the LSGGLGAGSCR peak (28.8 minutes) hides the fact that b2 (cyan), b3 (light red), and b4 (light green) are interfered with, while (C) the extracted ion chromatograms expose those interferences.

Early on, DIA data was analyzed using typical database search engines(15, 17) such as Sequest(18) but new approaches take advantage of the repetitive MS/MS measurements in DIA (Figure 1-4). Two major classes of tools have emerged to detect and quantify peptides from DIA experiments. Spectrum-centric analysis tools attempt to demultiplex several peptide signals from the same MS/MS spectra(19, 20) 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 data. In contrast, peptide-centric analysis(21) looks for specific peptides across all spectra in a precursor isolation window. PECAN(22) queries DIA data using peptide sequences and their predicted fragmentation models while spectrum library search tools(23–25) leverage previously acquired DDA data. Often DDA and DIA data are acquired in tandem(26, 27) such that detections can be performed using the DDA data sets and typical search engines, while quantitation can be performed using the DIA data.

DIA is gaining a resurgence because of improved data acquisition rates, routine high resolution and accurate mass measurement, and a paradigm shift in the analysis strategy from spectrum-centric to peptide-centric approaches that makes data interpretation easier. Here we report on the current state of DIA analysis considering both detection depth and quantitative accuracy. We outline some remaining problems with current DIA methodologies and indicate new directions explored in this dissertation.

1.2 Methods

Complete methods are discussed in Appendix A.

1.3 How deeply can DIA measure the human proteome?

The HeLa proteome is estimated to contain approximately 12,250 proteins(28). To investigate the benefits and shortcomings of various DIA strategies we serum starved S3 HeLa cells for either 2 or 16 hours to generate a proteome-wide perturbation that produced signals representative of biologically relevant changes. We pooled the two

samples and collected six GPF DIA experiments (Figure 1-1F) using the PAcIFIC approach. We searched this data set with the library search engine, EncyclopeDIA (presented in detail in Chapter 3), using a DDA-based spectrum library generated from in-house HeLa experiments containing 166,354 peptides filtered to a 1% peptide FDR threshold, corresponding to 9,947 parsimonious protein groups. While the serum starved HeLa proteome likely does not represent every protein the HeLa proteome is capable of producing, we were still able to recapitulate 102,971 peptides corresponding to 7,376 parsimonious protein groups filtered at a 1% protein FDR. Considering that the Bekker-Jensen et al(28) HeLa proteome corresponded to 46 individual MS/MS experiments, this represents a significant fraction of the proteome captured with only a fraction of the instrumentation time.

In addition, we collected triplicate experiments of the two serum-starved samples using single-shot DDA (Figure 1-1A) and three single-shot DIA strategies: “normal” 25x 24 m/z DIA (Figure 1-1C), 25x “variable” precursor isolation window DIA (Figure 1-1D), and 51x 24 m/z “overlap” DIA (Figure 1-1E). Searching the same DDA-based spectrum library, the three DIA strategies performed effectively equally well at detecting peptides, where the best strategy, overlapping 24 m/z DIA, performed only 13% better than the worst strategy, normal 24 m/z DIA (Figure 1-5). All three methods significantly outperformed DDA by approximately 20% using the same instrumentation at both peptide and protein detection.

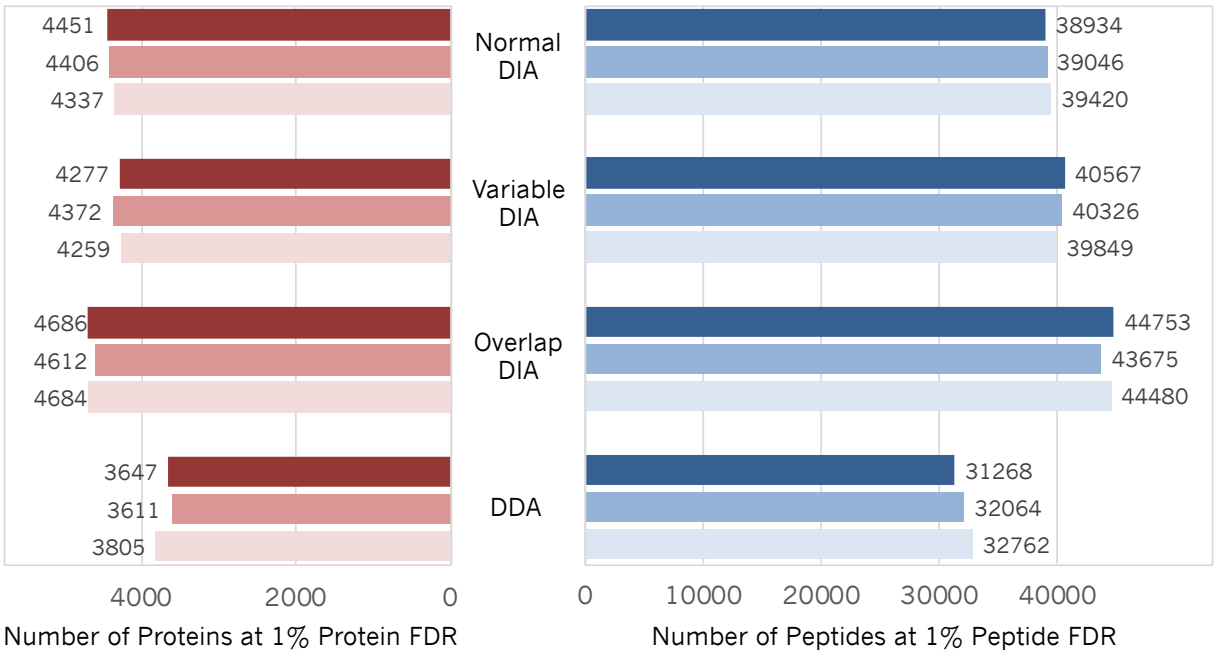


Figure 1-5: The number of peptides and proteins detected by various acquisition methods.

Using the same HeLa sample and the same hardware and chromatography setup, the number of human proteins (red) and peptides (blue) detected at a 1% protein and 1% peptide FDR, respectively. Three replicates for each method are shown as different shades. In general all three wide-window DIA windowing strategies perform similarly and outperform DDA by approximately 20% at the same FDR threshold.

1.4 Using DIA to build targeted assays

Similar to SRM and PRM, DIA quantification is performed using fragment ions. Given the consistency in instrumentation hardware, DIA is an excellent platform to develop targeted PRM assays. To demonstrate this, we picked over 600 peptides from 75 kinases detected in the six-injection GPF DIA experiment that we hypothesized to be changing in abundance as a result of our two serum starvation conditions based on prior

experiments. After overlap deconvolution, this analysis produces precursor isolation windows of 2 m/z, analogous to 1.4 m/z windows with typical PRM studies. Because we used the same instrumentation and chromatography set up for both experiments, we could set tight tolerances on both peptide retention times and fragmentation patterns. For each peptide, we scheduled PRM retention time windows that were +/- 60 seconds from the peak boundaries found with GPF DIA. Relatively tight retention time windows (less than 3% of a 90 minute linear gradient) enabled us to measure all 605 peptides with a minimum of 9 points across each peptide. For each peptide, we chose transitions that showed low interference in the DIA experiment as a starting point and further refined those transitions manually using Skyline(29).

We collected five technical replicates of each sample using this assay to determine a high precision “truth” set for quantitative changes between these samples. These peptides spanned over three orders of magnitude intensity (Figure 1-6A). It is commonly thought that serum starvation reduces the basal activity of cells. Others (30, 31) have reported changes in phosphorylation levels as a result of serum starvation, and we corroborate that observation by detecting changes in kinase levels. However, most of these changes are modest with a median fold change of 1.4x (Figure 1-6B). In this experiment 80% of peptides had less than 10% coefficient of variation (CV) across replicates and 94% of peptides had less than 20% CV (Figure 1-6C).

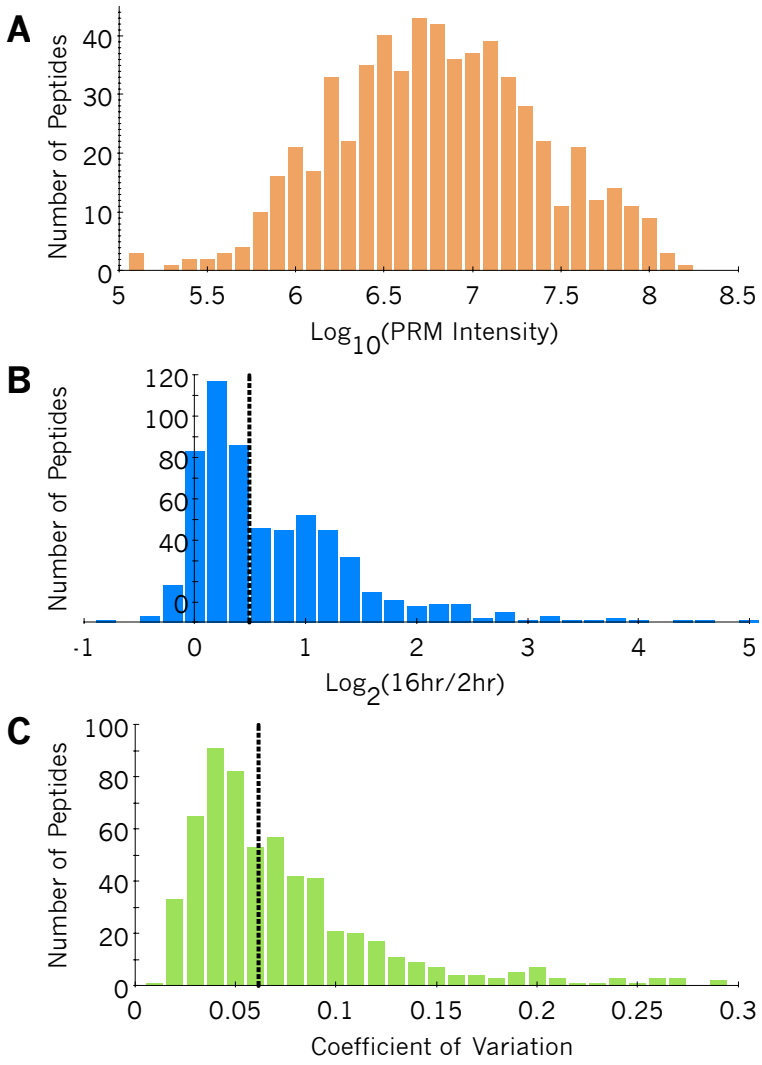


Figure 1-6: Histograms showing quantification statistics from the 605 peptide targeted PRM experiment.

(A) The median total fragment ion intensity for targeted peptides spans three orders of magnitude. (B) In general the targeted peptides in this experiment increased after serum starvation by 1.4x. These peptides were chosen because they had been observed to have changed in prior work. (C) The median coefficient of variation between replicates was 6.2%, where 94% of peptides exhibited less than 20% CV, a threshold commonly considered to indicate quantitatively robust peptides for targeted experiments.

The ease with which targeted assays can be prototyped with DIA in terms of a) target peptide selection, b) retention time windowing, and c) assay library development/transition refinement lowers the cost in time and sample for assay development. Chapter 2 explores an approach to extend this concept further by using DIA experiments to train a machine learning algorithm that predicts peptide response in SRM assays of completely different peptides.

Furthermore, GPF DIA can also be used to improve the interpretation of wide-window DIA. In a similar manner as with building PRM assays, narrow-window DIA of pooled samples can be used to produce retention time estimates tuned for a specific chromatography set up. These estimates are significantly more accurate than those from DDA libraries, which are typically combined across several days' worth of acquisitions. Chapter 3 discusses an approach to improve detection rates in wide-window DIA experiments by constructing DIA-based "chromatogram" libraries from GPF DIA runs.

1.5 How accurate are DIA quantitative measurements?

We used the PRM "truth" set to assess the quantitative accuracy of the single-shot DIA windowing strategies using fragment ion integration compared to more common precursor intensity integration used in DDA-based quantitation. For a peptide to be quantified with DIA we required at least three transitions with minimal interference in at least one replicate (on average 77% of 1% peptide FDR detections). Of the 605 peptides, DDA produced precursor quantification ratios for 57%, while DIA produced MS/MS quantitation ratios for between 50% and 58% percent (Table 1).

Table 1 Summary of quantification statistics for DIA and DDA.

(A) The percentage of PRM targeted peptides that were quantified. (B) The average bias in measurement. Since most of the targeted peptides increased in abundance with serum starvation, negative bias indicates compression in estimated ratios. The (C) interquartile range and (D) 90% range in ratio of ratios. (E) The percent of quantified peptides with fold change ratios that deviate by at least 2x from the targeted PRM ratios.

	Percent Quantified[A]	Bias[B]	50% Range[C]	90% Range[D]	Percent off by >2x[E]
Normal DIA	50%	-0.12	0.30	1.46	9%
Variable DIA	54%	-0.15	0.35	1.35	8%
Overlap DIA	58%	-0.15	0.33	1.51	9%
DDA	57%	-0.25	0.63	15.81	32%

However, DIA quantitative ratios tended to match ratios produced from the PRM experiments with both less bias and higher precision relative to DDA (Figure 1-7). This is strongly underlined when considering the number of measurements that incorrectly estimate the target PRM ratio by at least 2x fold. With this pool of peptides, 32% of DDA measurements are off by >2x fold, while that percentage is between 8% and 9% for the wide-window DIA windowing strategies (Table 1). Interestingly, while the different windowing strategies produced different numbers of quantified peptides, all of the strategies demonstrate the same level of quantitative error after interference is removed with automated transition refinement (Figure 1-8). While DIA methods show some degree of ratio compression due to signal interference, DDA has a bias indicating a 19% median reduction in ratios.

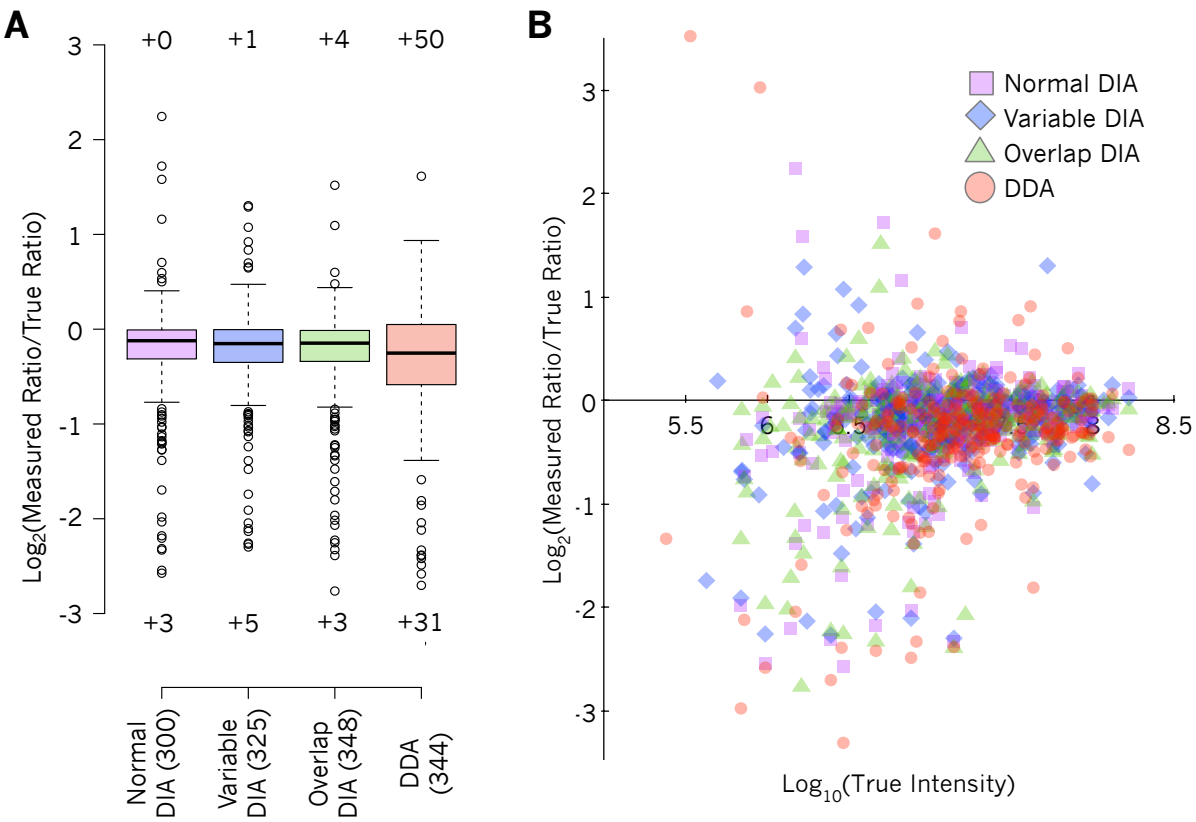


Figure 1-7: Accuracy of DDA and DIA global quantitative methods relative to targeted PRM.

The ratio of peptide intensities in HeLa after 16 hour versus 2 hour serum starvation was calculated for data collected with PRM, DDA, and three wide-window DIA windowing strategies. Assuming targeted PRM ratios as "true ratios", (A) boxplots showing the ratio of ratios deviation from ratios derived using global quantitative methods relative to those from PRM ratios. Here globally-derived ratios matching the PRM ratios would have a ratio of ratios equal to 0. The number of outliers outside of the plot range are shown above and below the plot. Boxes indicate the interquartile range (IQR) and Tukey-style whiskers are 1.5 * IQR away from the quartiles. (B) Scatterplot of ratio of ratios shows an intensity bias: lower scatter is seen in higher intensity peptides.

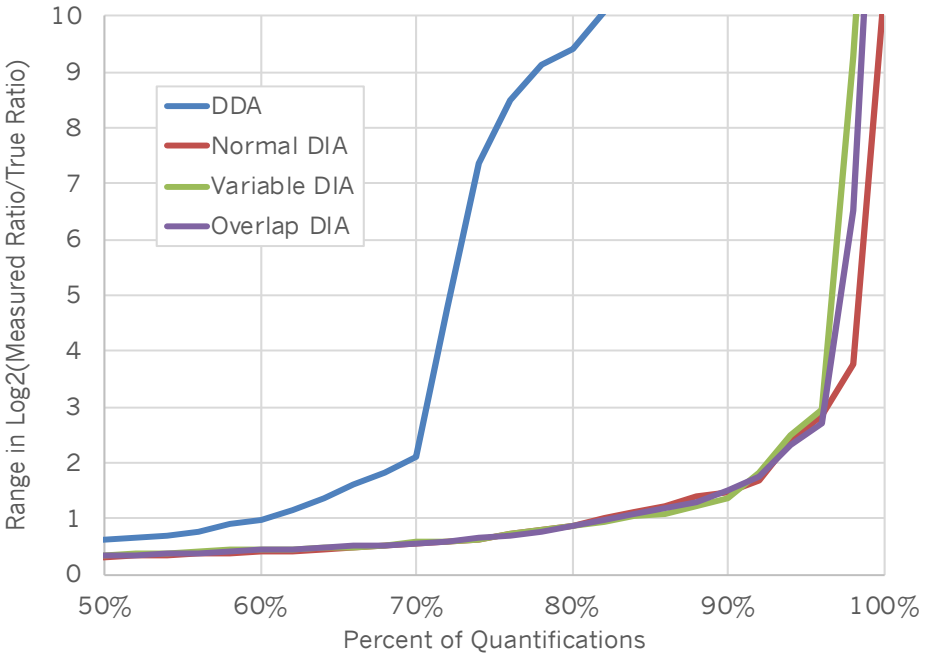


Figure 1-8: The range in ratio of ratios from DDA and DIA global quantitative methods compared to targeted PRM.
For reference, the interquartile ranges from Figure 1-7A are marked at 50% and the range of the all ratio of ratios for each method is marked at 100%. In general, only 70% of DDA measurements are consistent with PRM within +/- 2x fold change, while over 90% of transition-refined DIA measurements are consistent to the same thresholds.

To explore this phenomenon in greater depth, we used the peak boundaries from fragment ions in our PRM experiments to integrate the MS1 extracted ion chromatograms for monoisotopic, +1, and +2 ions. While there is almost always some MS1 signal around each peptide (only 14 of 605 had no signal), that signal does not always agree with the more selective fragment ion signal. Figure 1-9 shows that not only is there a high degree of scatter between MS1 and fragment ion ratios (R2=0.30), but the slope of that fit indicates MS1 measurements compress actual abundance by 1/3.

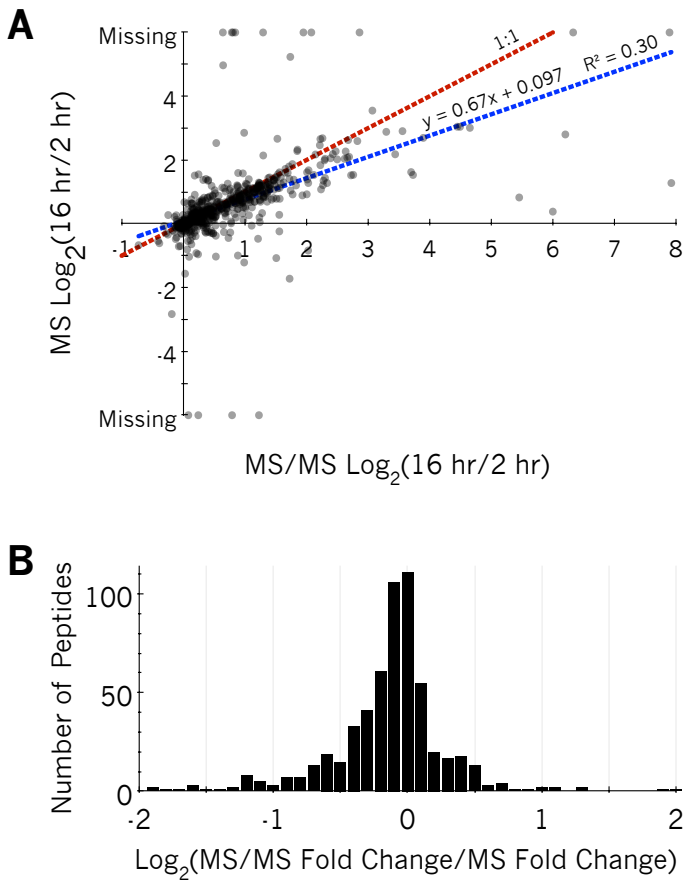


Figure 1-9: MS versus MS/MS quantitation within the same PRM experiments.

(A) A scatterplot showing the MS-based and MS/MS-based fold change estimates from 2 hour to 16 hour serum starvation conditions in 605 targeted HeLa peptides. If MS and MS/MS quantification methods produced the same results then points would fall on the dashed 1:1 red line. However, linear regression (blue line) indicates a 2/3rds compression in quantitative ratios derived using MS XIC integration. (B) A histogram showing the ratio of ratios between 16 hour/2 hour serum starvation fold changes estimated from MS and MS/MS signals.

This variation underlines how the reliance on precursor signal to trigger DDA scans is problematic. Not only does precursor signal come from interfering sources other than the

peptide of interest, full scan MS1 intensities are often less sensitive than MS/MS intensities due to dynamic range limitations. With ion traps (for example the Orbitrap used in this study) the trap capacity is limited to a fixed number of ions, which makes it possible for individual ion species to overwhelm the trap. Here sensitivity in MS/MS spectra benefits from automatic gain control as the instrument spends more time accumulating data for low abundance molecular species.

1.6 Shared fragment ions between peptides complicate detection confidence

Proteins are post-translationally modified in hundreds of ways(32), either through a natural aging process (e.g. deamidation) or enzymatically (e.g. acetylation, phosphorylation) in vivo, or as a byproduct of sample handling (e.g. oxidation, carbamylation) in vitro. Several key considerations must be factored into the analysis of post-translational modifications (PTMs) using DIA. Singly modified peptides implicitly share half of their B and Y ions with their unmodified counterparts. In some circumstances these two forms can be found within the same precursor window. For example, +2H singly oxidized peptides are precursor mass shifted by approximately 8 m/z from the unmodified form and 2/3 of the time fall in the same 24 m/z isolation window. This phenomenon grows at higher charge states and with lower mass PTMs (e.g. deamidation, methylation). Oftentimes, naturally occurring modified peptides exist at a small fraction of the modified form and can be misaligned to peaks where the unmodified peptide elutes because they share so many of the same ions (Figure 1-10).

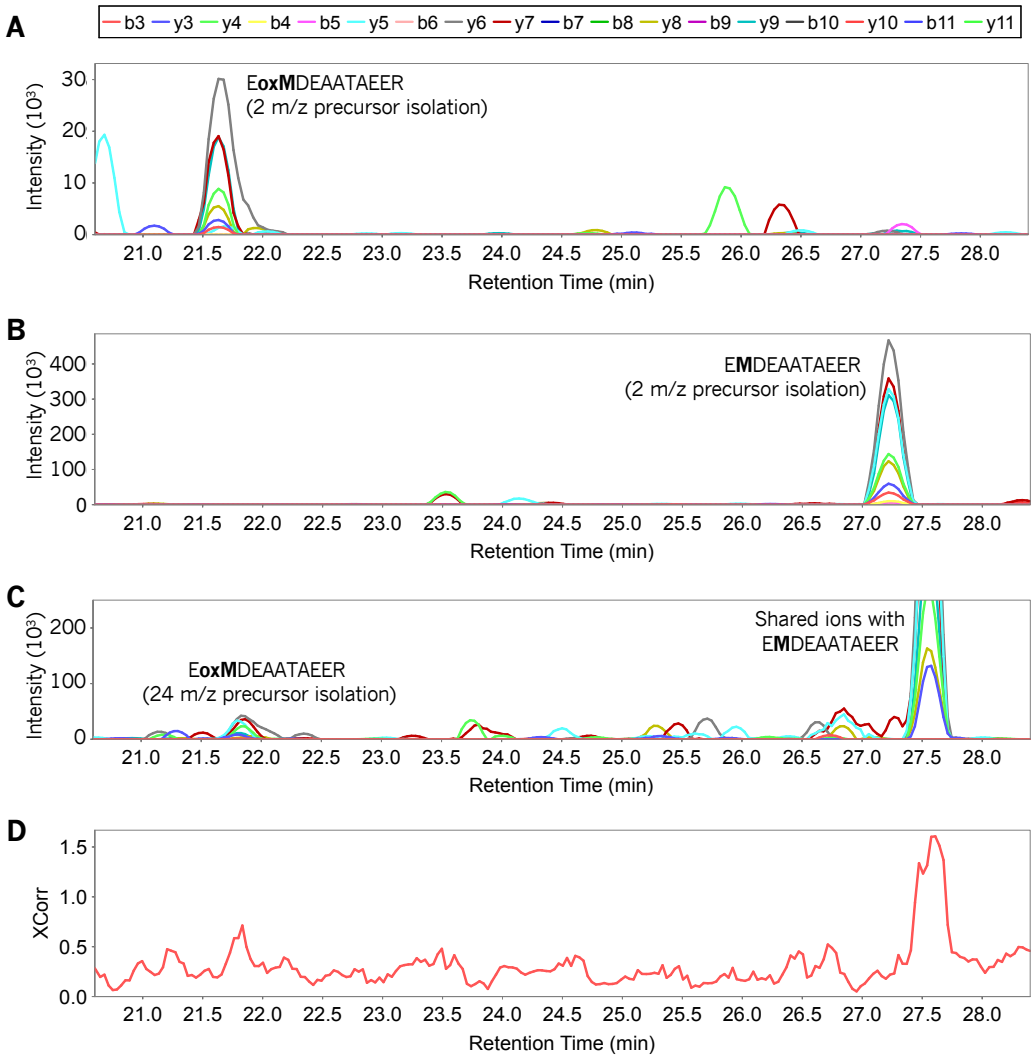


Figure 1-10: Retention time shift for oxidized EMDEAATAEER.

Fragment ion chromatograms for (A) methionine oxidized EoxMDEAATAEER (+2H, 640.54 m/z) and (B) unoxidized EMDEAATAEER (+2H, 628.54 m/z) from narrow window DIA indicate that the oxidized form elutes at 22.7 minutes, while the unoxidized form elutes at 27.5 minutes. While the mass shift between these peptides forces them to fall in different 2 m/z narrow windows, both forms fall in the same 24 m/z wide window. Here the fragment ion chromatogram for (C) EoxMDEAATAEER indicates Y-type ions shared with unoxidized EMDEAATAEER. (D) Despite missing several B-type ions, these Y-type ions are intense enough to cause the XCorr score to misassign the retention time for EoxMDEAATAEER as 27.5 minutes.

This can be readily apparent when mapping detections from narrow precursor window GPF DIA with effectively 2 m/z precursor isolation to normal DIA with 24 m/z precursor isolation (Figure 1-11).

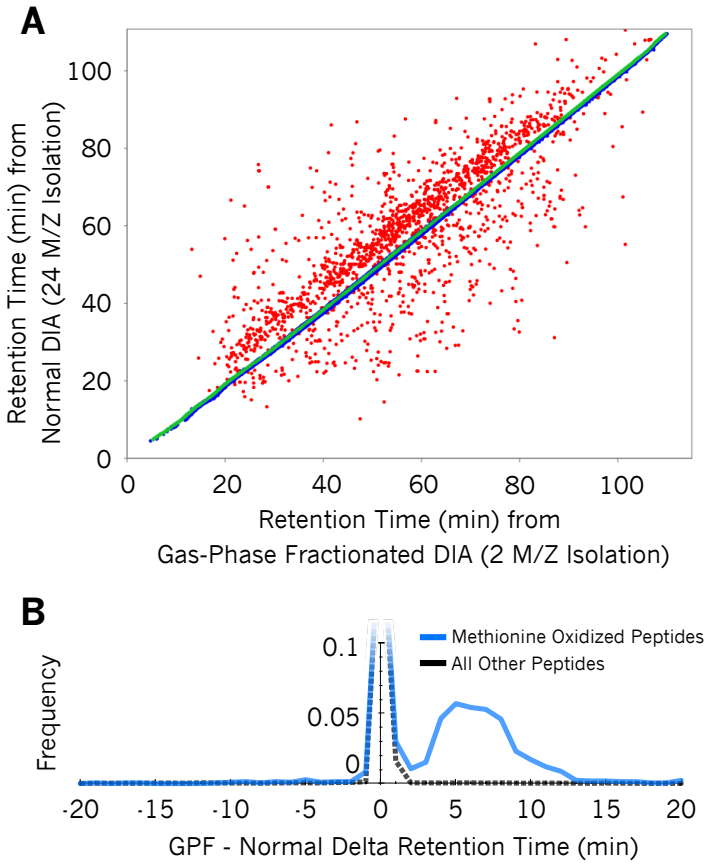


Figure 1-11: The effect of shared fragment ions on PTM detection.
(A) Retention time alignment between "Normal" DIA (with 24 m/z precursor window isolation) and gas-phase fractionated DIA (with 2 m/z effective precursor isolation) shows retention time shifted peptides despite being run using identical chromatography conditions. Peptides that fit the retention time model are blue, while those that do not are red. (B) Frequency plot showing that these shifted peptides are exclusively methionine oxidized. These detections are artifacts of the fact that normal and oxidized peptides often coisolate in the same precursor window.

Retention time alignment between these methods is very high since these were acquired using the same chromatography set up. However, there is a cloud of peptides that are shifted to appear as if they elute later by 4-8 minutes (more hydrophobic) in the wider window DIA experiment. These peptides are correctly detected in the GFP DIA with narrow precursor isolation (actually shifted earlier from unmodified form), but are incorrectly detected at the retention time when the unmodified peptide elutes because of shared fragment ions from the higher abundant unmodified peptide. Our observed retention time shift confirms previous work by Lao et al(33), who noted the decreased hydrophobicity of oxidized peptides. Regardless, these incorrect detections are insidious because they masquerade as “target” detections when computing false discovery rates using the target/decoy strategy(34, 35).

In addition, for any given PTM, in many cases it is possible for peptides to contain multiple potential sites of modification, or even to be present in multiple isomeric states. For example, within the human genome as many as 700,000 residues are suspected of being phosphorylatable(36) and accumulated phosphoproteomic data shows that phosphorylation sites cluster together in multi-phosphorylated proteins, where over half of phosphorylation sites are within four amino acids of each other(37). This, coupled with the observation that over 90% of the peptide sequences in the phosphopeptide database Phosphopedia(38) contain multiple serines, threonines, and tyrosines (Figure 1-12), underlines the importance of methods to determine the specific site of PTMs.

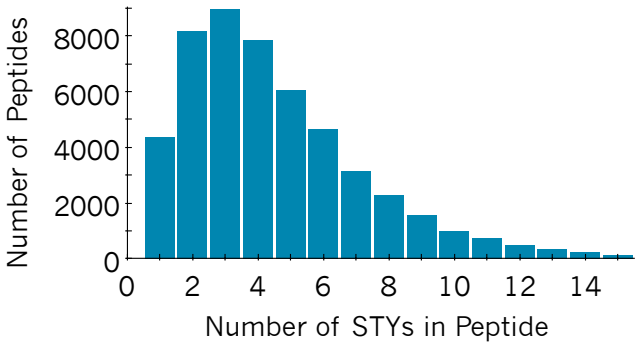


Figure 1-12: The distribution of number of serines, threonines, and tyrosines in phosphopeptides.

Previously detected peptide sequences were drawn from a large-scale phosphopeptide DDA spectrum library(38).

One major complication is that peptides that exist as multiple positional isomers (e.g. singly phosphorylated peptides where multiple residues are found to be phosphorylated) share the same precursor mass and many of the same fragment ions. While software tools that assign PTM sites using site-specific fragment ions are readily available to analyze DDA data(39–42), only recently have efforts been made to adapt these algorithms to handle the high interference levels in DIA data. One such effort is detailed in Chapter 4.

When considering DIA data sets, modified peptides are not the only peptides that share fragment ions. Many common sequence variants, either due to single nucleotide polymorphisms (SNPs), paralogs, or even orthologs (when considering multiple species) shift peptide masses by amounts small enough such that both peptides still coisolate in the same precursor isolation window. Similar to PTM localization, variants can be statistically confirmed using variant-specific fragment ions. An approach to this is briefly discussed in Chapter 5.

1.7 Conclusions

Throughout this work we demonstrate that despite limitations in precursor isolation, DIA is an effective tool to detect and quantify peptides. Here we find that all of the wide-window DIA windowing strategies tested, including the most basic 25x 24 m/z DIA method, outperformed DDA methods in detection rate and quantification accuracy. We found that the widely varying windowing schemes performed consistently with each other, underlining the reproducibility of the general DIA approach. In these experiments, the overlapping DIA windowing scheme tended to detect more peptides and proteins than either normal or variable width windowing schemes, but that those improvements were relatively small. As with detection rates, all three DIA strategies tested here exhibit relatively comparable performance on all quantitative metrics. This comparability suggests that any method using fragment ions for quantitation is more important for precision than the specific method that collected those fragment ions.

2 USING DATA INDEPENDENT ACQUISITION TO MODEL HIGH-RESPONDING PEPTIDES FOR TARGETED PROTEOMICS EXPERIMENTS

2.1 Summary

This chapter is based on the following published article: Searle BC, Egertson JD, Bollinger JG, Stergachis AB, MacCoss MJ. Using Data Independent Acquisition (DIA) to Model High-responding Peptides for Targeted Proteomics Experiments. Mol Cell Proteomics. 2015 Sep;14(9):2331-40.

Targeted mass spectrometry is an essential tool for detecting quantitative changes in low abundance proteins throughout the proteome. Although Selected Reaction Monitoring (SRM) is the preferred method for quantifying peptides in complex samples, the process of designing SRM assays is laborious. Peptides have widely varying signal responses dictated by sequence-specific physiochemical properties; one major challenge is in selecting representative peptides to target as a proxy for protein abundance. Here we present PREGO, a software tool that predicts high responding peptides for SRM experiments. PREGO predicts peptide responses with an artificial neural network trained using 11 minimally redundant, maximally relevant properties. Crucial to its success, PREGO is trained using fragment ion intensities of equimolar synthetic peptides extracted from data independent acquisition (DIA) experiments. Due to similarities in instrumentation and the nature of data collection, relative peptide responses from DIA experiments are a suitable substitute for SRM experiments because they both make

quantitative measurements from integrated fragment ion chromatograms. Using a SRM experiment containing 12973 peptides from 724 synthetic proteins, PREGO exhibits a 40-85% improvement over previously published approaches at selecting high responding peptides. These results also represent a dramatic improvement over the rules-based peptide selection approaches commonly used in the literature.

2.2 Introduction

Targeted proteomics using Selected Reaction Monitoring (SRM) and Parallel Reaction Monitoring (PRM) is increasingly becoming the gold-standard method for peptide quantitation within complex biological matrices(43, 44). By focusing on monitoring only a handful of transitions (associated precursor and fragment ions) for targeted peptides, SRM experiments filter out background signals, which in turn increases the signal to noise ratio. SRM experiments are almost exclusively performed on triple-quadrupole instruments. These instruments can isolate single transitions as an ion beam and measure that beam with extremely sensitive ion-striking detectors. As a result SRM experiments generally exhibit significantly more accurate quantitation when compared to similarly powered discovery based proteomics experiments, and frequently benefit from a much wider linear range of quantitation(45). SRM experiments often require less fractionation and can be run in shorter time on less expensive instrumentation. These factors allow researchers to greatly scale up the number of samples they can run, which in turn increases the power of their experiment.

However, the process of developing an effective SRM assay is often cumbersome, as subtle differences in peptide sequence can have a profound impact on the

physiochemical properties and subsequent SRM responses of a peptide. To successfully develop a SRM assay for a protein of interest, unique peptide sequences must be chosen that also produce a high SRM signal (e.g. high-responding peptides). Once identified, these high responding peptides are often synthesized or purchased, and independently analyzed to determine the most sensitive transition pairs. Finally, the selected peptides and transitions pairs must be tested in complex mixtures to screen for transitions with chemical noise interference and to validate the sensitivity of the assay within a particular sample matrix. Peptides and transitions that survive this lengthy screening process can then undergo absolute quantitation by calibrating the signal intensity against standards of known quantity.

While experimental methods have been developed to empirically determine a set of best responding peptides(46), these strategies can be time consuming and require analytical standards, which are not currently available for all proteins. More often than not, representative peptides are essentially chosen at random, using only a small number of criteria, such as having a reasonable length for detection in the mass spectrometer, a lack of methionine, and a preference for peptides containing proline(47). It is not uncommon for SRM assays to fail at the final validation steps simply because the peptides chosen in the first assay creation step happened to be unexpectedly poorly responding peptides.

In an effort to speed up the process of generating robust assays, several groups(48–51) have designed approaches to predict sets of proteotypic peptides using machine-learning algorithms. Proteotypic peptides are peptides commonly identified in shotgun proteomics experiments for a variety of reasons including high signal, low

interference, and search engine compatible fragmentation. Enhanced Signature Peptide (ESP) Predictor[7] was the first successful modification of this prediction approach to use proteotypic peptides as a proxy for high-responding peptides for SRM-based quantitation. In brief, Fusaro *et al* built a training data set from data-dependent acquired (DDA) yeast peptides and a proxy for their response was quantitated using extracted precursor ion chromatograms (XICs). The authors calculated 550 physiochemical properties for each peptide based on sequence alone and built a random forest classifier to differentiate between the high and low response groups. Other peptide prediction tools follow the same general methodology for developing training data sets. CONSeQuence[8] applies several machine learning strategies and a pared down list of 50 distinct peptide properties. Alternately, Peptide Prediction with Abundance[9] (PPA) uses a back-propagation neural network(52) trained with 15 distinct peptide properties selected from ESP Predictor's 550. The authors of CONSeQuence and PPA found that their approaches outperformed the ESP Predictor on a variety of data sets.

As with most machine learning-based tools, the generalizability of the training set to real-world data is key to the effectiveness of the resulting prediction tool. While MS1 intensities extracted from DDA data can be useful for predicting high-responding peptides(53, 54), several factors make them less than ideal for generalizing to SRM and PRM experiments. In particular, DDA peptides must be identified before being quantified, and key biochemical features beneficial for targeted analysis of transitions can reduce overall identification rates by producing fragment spectra that are difficult to interpret with typical search engines. By building training data sets on precursor intensities alone these tools ignore the fact that targeted assays actually use fragment ions for quantification. We

propose that constructing training sets from DIA fragment intensities(26) will produce machine-learning tools that are more effective at modeling peptides that produce detectible transitions, rather than just proteotypic peptides.

The use of digested proteins in training sets presents additional concerns. The observed variance in peptide intensities is confounded by variation in protein abundance. Converting peptide intensities to ranks can remove the dependence on varying protein levels at the cost of corrupting the training set with proteins that biochemically contain no high-responding peptides. PPA attempts to ease this concern by training with Intensity Based Absolute Quantitation (iBAQ) values(55) for DDA peptides estimated from XICs. We hypothesize that constructing a training set from equimolar synthetic peptides removes most adverse effects of digestion from the training set, making it possible to construct a more generalizable tool.

2.3 Methods

Complete methods are discussed in Appendix B.

2.4 Challenges in predicting peptide responses

Peptide response factors within proteins vary widely: on average by over three orders of magnitude between the highest and lowest responding peptides. Stergachis *et al* has previously presented an experimental method for determining the best responding peptides to monitor proteins in targeted experiments. This method was demonstrated by synthesizing over 700 human transcription factors *in vitro* and generating SRM assays for all singly charged, monoisotopic y_3 to y_{n-1} ions from virtually every tryptic peptide. Due to variations in translation, proteins in this experiment were not produced at the same

level. However, all peptides within a given protein were guaranteed to be present at equimolar levels, and using this knowledge the authors were able to determine which peptides produced the best SRM transitions for *in vivo* monitoring. In this work we use the Stergachis *et al* data set as an independent test set to validate our methods. Some potential limitations of this data set for benchmarking include that it was acquired only considering precursor charge state +2 peptides (which may bias against high basicity peptides and very long peptides), and that analyzed fragment ions were limited to only y-type ions. We feel that the benefits of the scale of this data set outweigh these limitations.

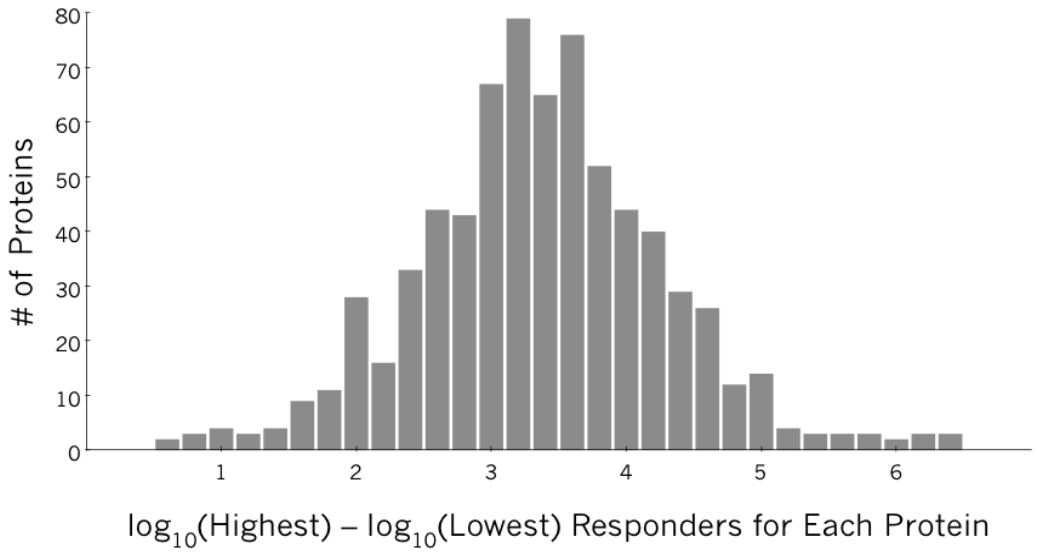


Figure 2-1: A histogram of the dynamic ranges calculated for 724 proteins. The dynamic range is estimated as the number of orders of magnitude separation for each protein. This value is calculated as the difference between the $\log_{10}$ intensities of the highest responding peptide and the lowest responding peptide. The median dynamic range is 3.4 orders of magnitude, with an interquartile range of 1.2 orders. All protein intensity data was drawn from the Stergachis *et al* SRM testing data set.

The Stergachis *et al* data set provides an excellent testing ground for understanding the challenges in predicting peptide responses. Figure 2-1 illustrates the

range of peptide transition responses in the Stergachis *et al* SRM data set. While the median dynamic range of peptide responses within a protein was 3.4 orders of magnitude, some rare proteins demonstrated response ranges of up to five or six orders of magnitude. An example distribution for CASZ1, a typical transcription factor with an apparent dynamic range of 4.1 orders of magnitude, is shown in Figure 2-2. This wide diversity of responses underlines the need for a robust mechanism for choosing peptides to target. In this work we leverage the Stergachis *et al* data set containing 12973 peptides from 724 proteins (with a median of 15 peptides per protein and a mode of 10) to test our approach for predicting peptide responses for SRMs and PRMs.

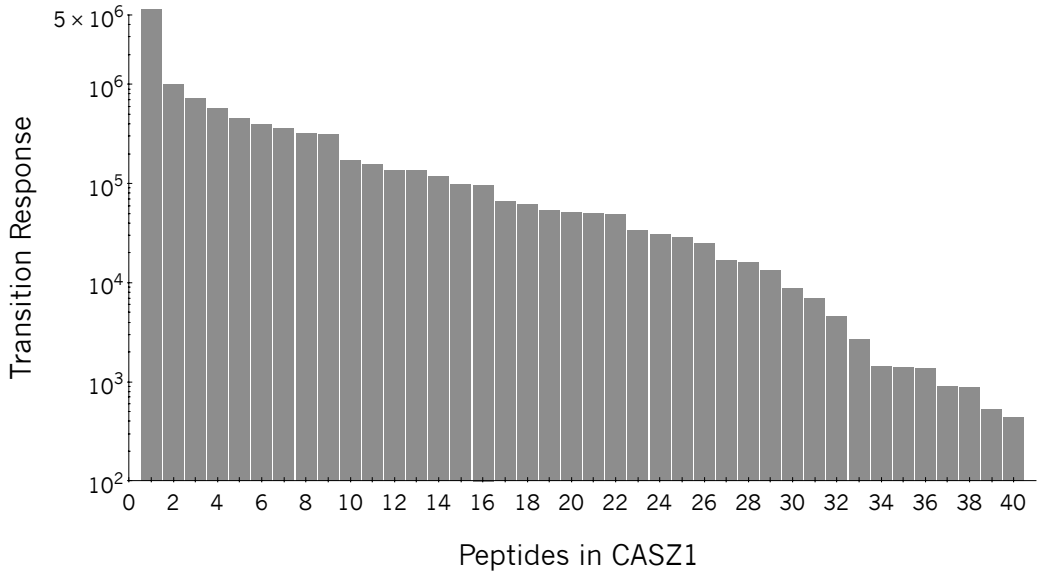


Figure 2-2: SRM transition responses for peptides in CASZ1.
The most intense y-type ion fragments intensities (a proxy for the best transition to monitor for each peptide) for the 40 tryptic peptides measured in CASZ1 span 4.1 orders of magnitude. CASZ1 represents a typical protein in the Stergachis *et al* data set, where transition responses of peptides in most proteins span an average of 3.4 orders of magnitude.

2.5 Training set preparation

Training data sets that are generalizable to real world applications are critical for effective machine learning. However, creating an exhaustive targeted data set of equimolar peptides for training a peptide response prediction model is extremely time consuming, as it would require very many SRM experiments to account for all potential transition ions for every peptide. We've developed a strategy for generating large-scale, realistic SRM- and PRM-like training sets using DIA MS/MS experiments acquired on a QExactive-HF (Thermo Scientific; Bremen) using HCD fragmentation. For the purposes of determining a training data set, DIA MS/MS has the advantage that all sequence specific fragments are measured, making it easy to identify the most promising transitions. Additionally, we used beam-type higher energy collisional dissociation (HCD) fragmentation to generate fragments, which is very similar to triple-quad fragmentation used in most SRM experiments(56). We derived the training set from the most intense singly-charged y-type fragment intensity for each of 1679 stable isotope labeled peptide detections made by Skyline, given certain restrictions. Only singly charged y-type fragments were used because b-type fragments can lose carbon monoxide to form a-type fragments, resulting in both lowered response and increased variability. Also, typically the b-ion series undergoes multiple collisions in beam-type instruments and fragments to smaller product ions until it stops at the b_2 ion. This fragment ion is frequently one of the most intense but least selective product ions in the spectrum. First we filtered our list of potential signature y-type fragment ions to remove non-specific y_2 fragments. Then, for each acquisition, we removed the 2.5% worst fragment ions by mass accuracy in both

directions. At this point we estimated the maximum y-type fragment for each peptide as a proxy for the maximum transition response.

Because peptide detections were made from two pairs of acquisitions at different amounts (approximately 45 fmol and 15 fmol on-column), we were able to use the distribution of parent-intensity quantitative ratios to indicate outlier peptides (Figure 2-3). Based on this analysis, from the initial 1,331 detected peptides we removed 69 stable-isotope labeled (SIL) peptides that eluted earlier than 30 minutes or later than 85 minutes. In our runs, early eluting peptides tended to saturate in ratio between 45 fmol and 15 fmol injections, suggesting that their intensities were unreliable. Peptides eluting after 85 minutes were excluded because our instrument tuning parameters made their intensities also unreliable. After removing these peptides, we recalculated the median ratio of the two pairs of acquisitions to be 2.45, slightly under the expected 45:15 fmol ratio. We estimated the overall intensity for each peptide as the average of the intensities from the 45 fmol acquisition and 2.45 times the 15 fmol intensities and removed the peptides with the 2.5% highest and 2.5% lowest ratios to compensate for peptides with unstable responses. This resulted in a final training data set of 1,186 well-behaved peptides. Finally we ranked the peptides in the training set based on these aggregate fragment ion intensities and linearly normalized the ranks to be between 0 and 1.

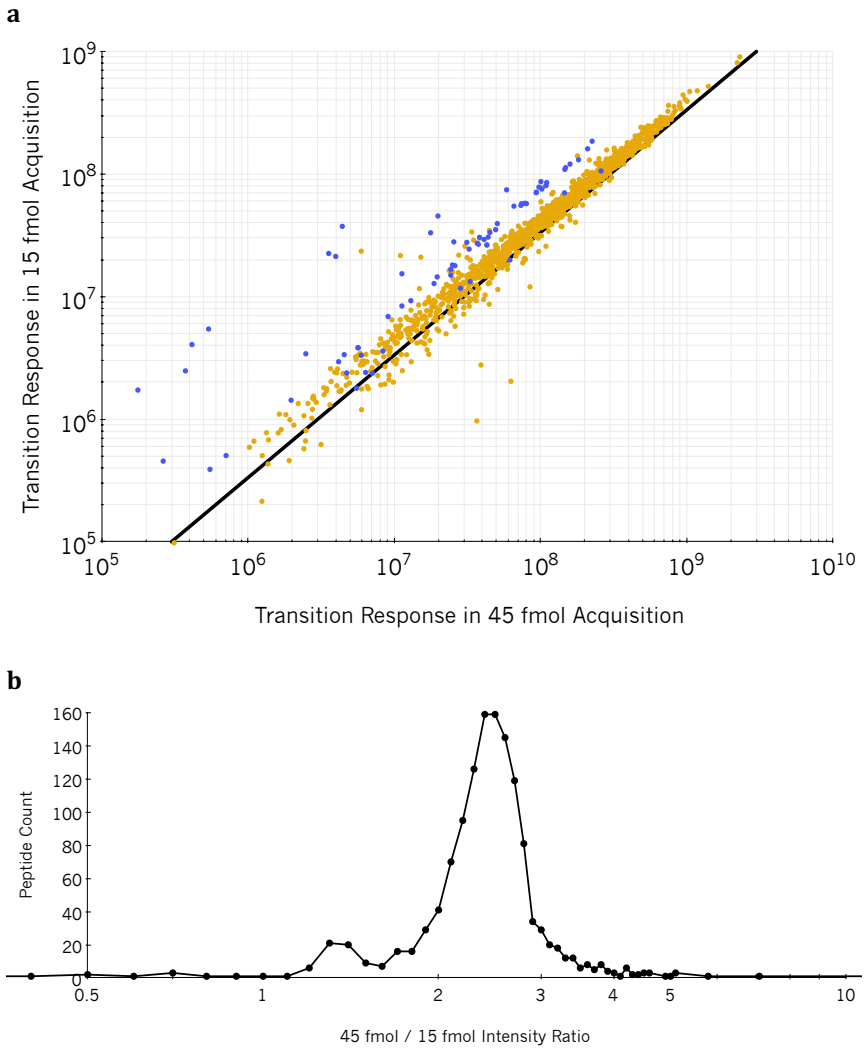


Figure 2-3: Distribution of transition responses in training data set.

(a) A scatter plot comparing the 45 fmol injections with 15 fmol injections. The majority of peptides (orange dots) produce responses that fall near the expected 3:1 ratio (black line). Peptides with retention times earlier than 30 minutes (blue) fall off that line (average of 1.35:1) and were excluded from the training data set. (b) The distribution of 45 fmol to 15 fmol peptide transition response ratios. The sample concentrations were generated simply by injecting 3x by volume using the relatively imprecise HPLC autosampler. It is not surprising that the true ratio of peptide transitions falls around 2.45:1. Consequently we used this ratio to normalize the sample intensities rather than 3:1.

2.6 Physiochemical property selection and artificial neural network training

For each peptide sequence we calculated 550 physiochemical properties used by ESP Predictor, the large majority of which were derived from the Amino Acid Index Database(57). We point out that one potential source of variability is that cysteines used in this work (and in proteomics generally) are alkylated, while the majority of the Amino Acid Index Database properties assume cysteines are unmodified. We normalized the values for these properties to be between 0 and 1. We selected meaningful physiochemical properties using a minimum redundancy, maximum relevance (mRMR) algorithm(58, 59). For each property, we calculated the Pearson's correlation coefficient of ranked peptides with the property values derived from their respective peptide sequences. The property with the highest correlation was selected as a meaningful feature and all other properties that correlate with that feature at an absolute Pearson's correlation coefficient of >0.3 are removed. This process is iterated using the remaining properties until all properties that have any positive correlation to the intensity ranks are either selected or removed.

The mRMR algorithm produced 11 most relevant physiochemical properties. These properties and their correlation to the ranked training intensities are listed in Table 2. As the mRMR algorithm chooses the most representative of several properties, the specific properties themselves are less important than their higher-level classification. Peptides with lower molecular weights correlated strongest with high transition intensities in our training set, followed by various structural and hydrophobicity properties.

The final training set consisted of the top 25% (high responders) and the bottom 25% (low responders) of peptides to promote differentiation between high and low responding peptides, where the expected output was the percentage intensity rank. We constructed a back-propagation neural network with 11 input neurons corresponding to the 11 mRMR-selected relevant physiochemical properties, 8 hidden neurons in a single layer, and a single output neuron. We configured the neural network for a 10% learning rate and trained it to reach a minimum recall error level of 1%. Neural networks typically produce a score between 0 and 1, indicating the classification of the input feature set. Instead of using the neural network score directly, the PREGO score was assigned to:

$$PREGO\ score = log_{10}\left(\frac{ANN\ score}{1-ANN\ score}\right) \quad (1)$$

in an effort to stratify scores that clump around 0 and 1. This score is analogous to the log-likelihood ratio statistic for comparing two classification models. Pseudo code of the PREGO algorithm is presented in Figure 2-4.

PREGO Algorithm Approach

[a] **Select Minimum Redundancy Maximum Relevance (mRMR) Features**

- Rank intensities from DIA training data set (1,186 well-behaved peptides)
- normalize intensity ranks to 0...1
- Calculate 550 physiochemical properties for each peptide
- Normalize properties individually to 0...1
- While there are still unconsidered properties:
  - Select property with highest Pearson's correlation to intensities
  - Remove all properties with >= 0.3 Pearson's correlation to selected property

[b] **Build Artificial Neural Network (ANN)**

- For i=1...1000
  - Assign peptides a percentage between 0 and 1 by intensity rank
  - High=top 25% responders (intensity percentage rank: 0 to 0.25)
  - Low=bottom 25% responders (intensity percentage rank: 0.75 to 1)
  - Construct ANN_i
    - Build X selected property input neurons
    - Build (X+1)*2/3 hidden neurons
    - Train ANN to differentiate High from Low intensity ranks to 1% recall error rate
  - Score ANN_i versus SRM cross validation data set (44 proteins), keep if best score

[c] **Test PREGO ANN**

- Test using Stergachis *et al* SRM data set (724 proteins)

Figure 2-4: Algorithmic outline of the PREGO method.

(a) Algorithmic outline describing feature selection using an mRMR style algorithm to identify non-redundant features with maximum relevance. Feature sets with low redundancy often decrease the potential for over-training in machine learning algorithms. (b) Algorithmic outline for neural network construction using the mRMR-selected feature set. (c) Testing of the algorithm was performed using the Stergachis *et al* SRM testing data set.

Table 2: Peptide properties used in PREGO

[a] Properties were iteratively selected from a pool of 550 total properties based on their Pearson's correlation with the intensity ranks in the training data set. Properties are sorted based on the absolute value of the correlation coefficient, which is an indication of their importance for classification. Negative correlations indicate inverse relationships. As each feature was selected, redundant features with inter-property correlation coefficients >0.3 were removed. [b] Peptide properties were loosely categorized into three types, those corresponding with peptide size, secondary structure, and hydrophobicity.

Rank	Correlation Coefficient[a]	Peptide Property	Property Type[b]
1	-0.53	Peptide Mass	Size
2	-0.36	Average Relative preference value at C1(60)	Structural
3	-0.33	Average Activation Gibbs energy of unfolding, pH7.0(61)	Hydrophobicity
4	-0.27	Average Hydrophobicity coefficient in RP-HPLC, C4(62)	Hydrophobicity
5	-0.2	Average Normalized frequency of zeta R(63)	Structural
6	0.2	Average Linker propensity from 1-linker dataset(64)	Structural
7	0.16	Average Hydrophobicity coefficient in RP-HPLC, C18	Hydrophobicity
8	0.15	Average AA composition of EXT2 of single-spanning proteins(65)	Structural
9	-0.14	Average Normalized frequency of alpha-helix in all-alpha class(66)	Structural
10	0.08	Average Relative population of conformational state A(67)	Structural
11	0.07	Average Surface composition of AAs in intracellular proteins of thermophiles(68)	Structural

There are many decisions to make when picking a supervised machine learning architecture. As with PPA and ConSEQuence, we chose to implement an artificial neural network because “deep architectures” (like ANNs) tend to perform better than “shallow architectures” (e.g. support vector machines) on “deep learning” tasks(69). However, unlike the support vector machine approach to gradient descent, back-propagation gradient descent is random in nature, causing artificial neural networks to often converge on local minima, rather than global minima. Consequently we trained 1000 different ANNs and validated them using 44 proteins selected from an exhaustive SRM data set (see Methods section) modeled after the Stergachis *et al* experiment. We selected the best model that maximized the area of the receiver operating characteristic (ROC) that compared the number of peptides picked per protein versus the number of proteins where at least one high responding peptide was picked. For each protein, peptides were considered high responders if they produced a single most intense y-type fragment ion for each peptide in the top 20% of peptides from that protein. This approach also provides a buffer against over-fitting since we trained using DIA data and validating the trained models with SRM data acquired in a completely different manner.

2.7 Evaluation of PREGO

We evaluated PREGO using the Stergachis *et al* data set, which describes experimental SRM transition responses acquired for almost 13,000 peptides found in over 700 proteins. For consistency with our current practice we reprocessed this dataset to quantitate using only the single most intense fragment ion (y_3 to y_{n-1}), while the original

publication used the sum of those ions. Figure 2-5 shows PREGO scoring for CASZ1, a representative protein in this data set.

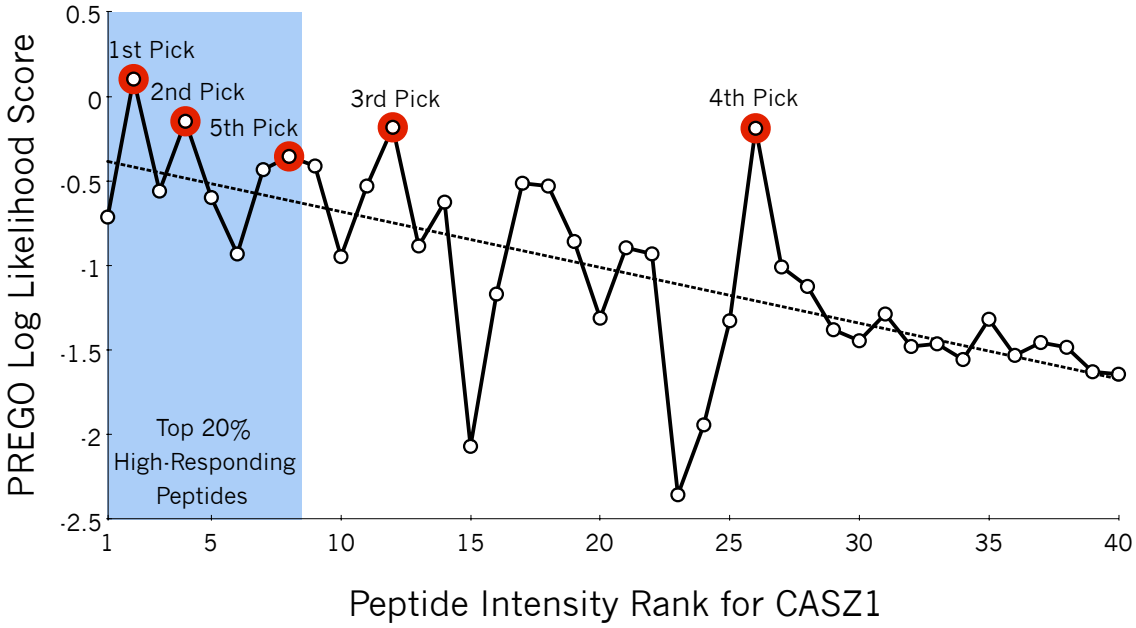


Figure 2-5: PREGO scores for peptides in CASZ1.

Peptides in CASZ1 (also known as cDNA FLJ20321) are ranked on their experimentally acquired transition fragment intensity from the Stergachis *et al* SRM testing data set where the peptide with the strongest response is awarded a rank of 1. The top 20% of peptides by intensity rank are considered "high-responding peptides" and are shaded in blue. The top five peptides chosen by PREGO are marked with red borders. While there is large variation in predicting response intensities for any given peptide (solid line), there is a definite trend (dashed line) to score first ranked peptides somewhat higher than worse ranked peptides. Consequently, the highest scoring peptides picked by PREGO are often also high-responding peptides. CASZ1 represents a "typical" protein with a correlation score of 0.65.

CASZ1 has a Pearson's correlation coefficient of 0.65 when compared with the experimental intensity ranks, the mode of the correlation distribution across all proteins in the data set (Figure 2-6).

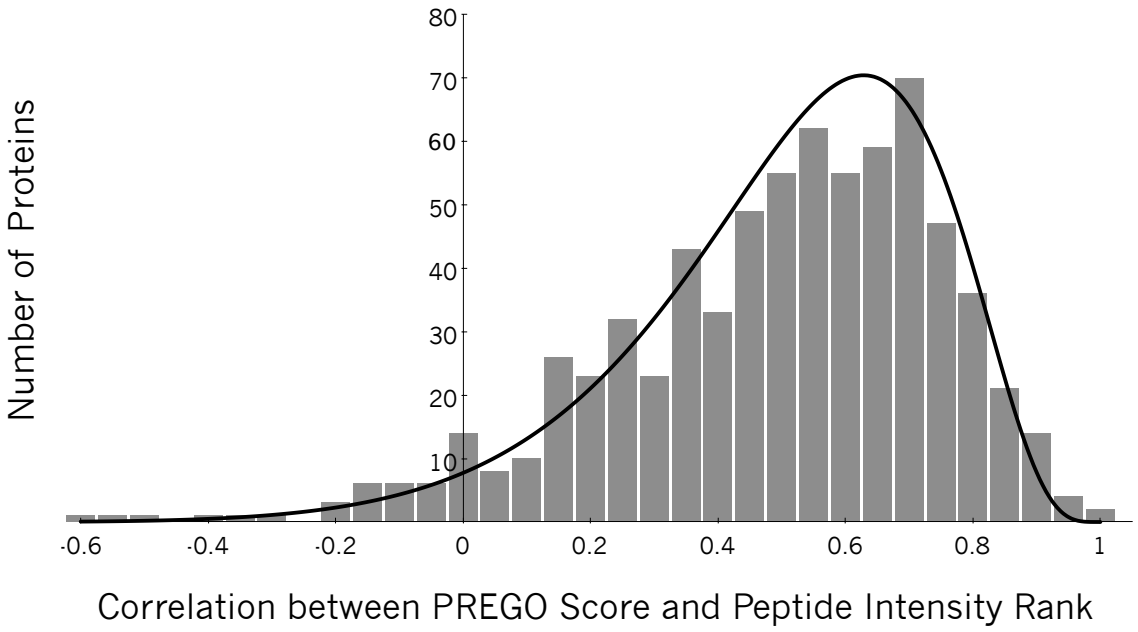


Figure 2-6: The distribution of correlation values between PREGO scores and ranked peptide intensities.

In general Pearson's correlation coefficients center between 0.4 and 0.8, but there is significant spread and the ranked peptide intensities of some proteins negatively correlate with PREGO scores. There is some expectation that poor correlation should occur by chance, since only a few peptides per protein are considered. The black line indicates the distribution of correlation coefficients predicted given a sample size of 10 (the mode of the number of peptides per protein) and a true correlation coefficient of 0.65.

Although there is significant deviation in any individual measurement, PREGO scores are generally high in cases of highly responding peptides, and low with less responsive peptides. Figure 2-7 illustrates the range of PREGO scores for a variety of proteins that show similar trends with correlation coefficients ranging from 0.9 to 0.2.

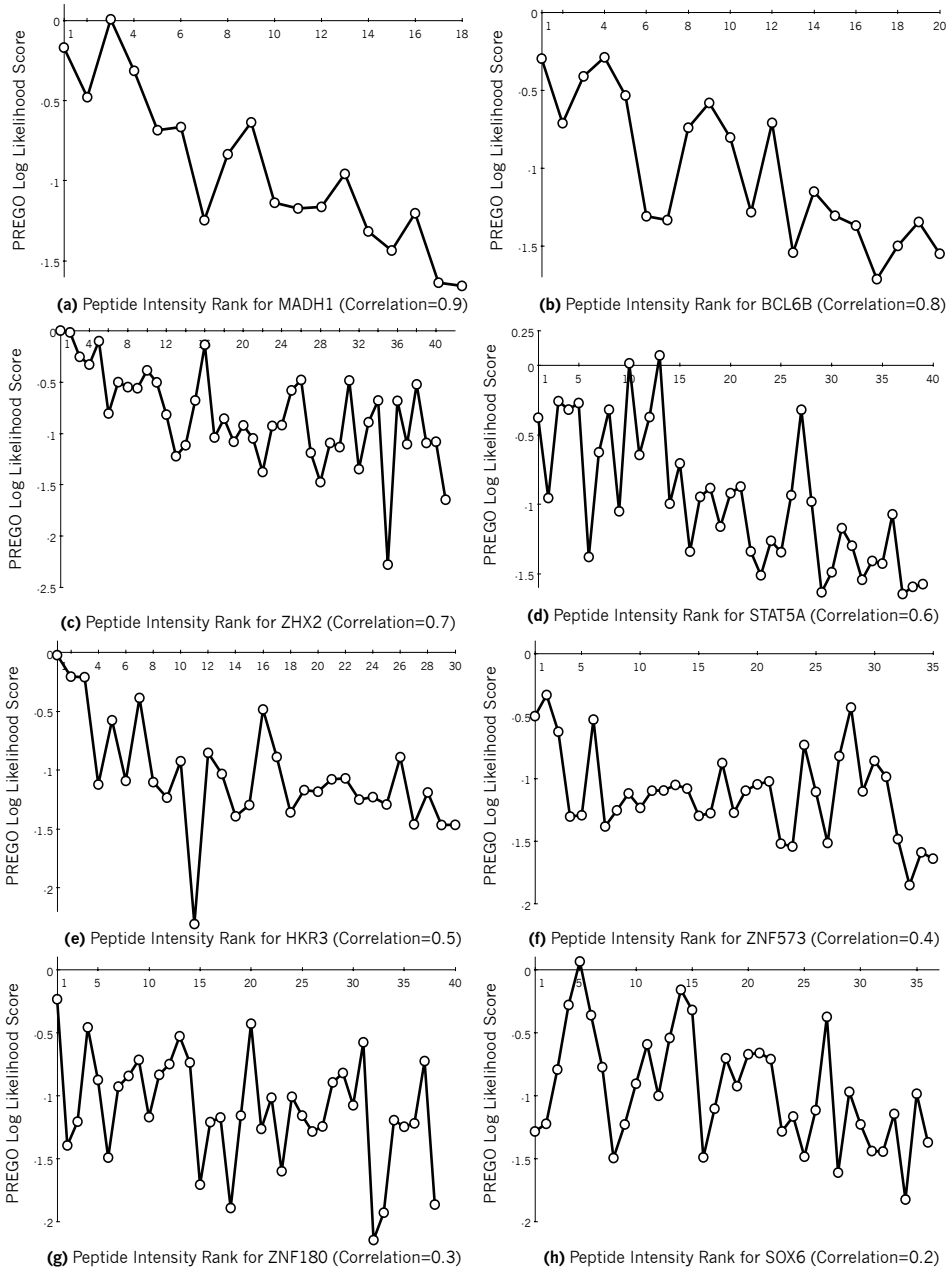


Figure 2-7: PREGO scores for peptides in proteins with a variety of correlation values.

PREGO scores for peptides are ranked on their experimentally acquired transition fragment intensity where the peptide with the strongest response is awarded a rank of 1. The proteins in this figure were chosen methodically: they were picked for having the highest number of peptides as long as the protein correlation value was within +/- 0.01 to the target correlation.

We combined traces like those shown in Figure 2-5 across all proteins in the Stergachis *et al* data set. Figure 2-8a depicts the distribution of PREGO scores for peptides at various ranks in all of the proteins, where the black line indicates the median score and the gray shaded area indicates the interquartile range.

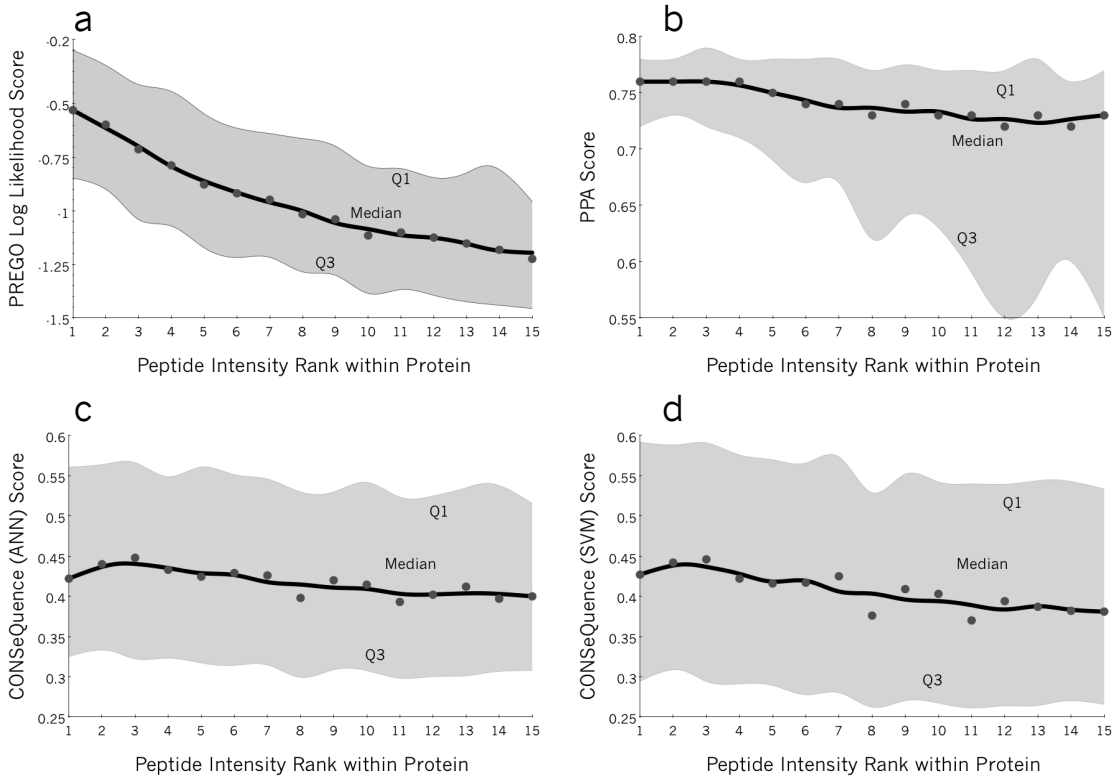


Figure 2-8: Score distributions for four scoring methods by peptide rank. (a) The PREGO score distribution for peptides of descending rank across the entire Stergachis *et al* SRM testing data set. The median ranks are annotated as dots, where the nearest-neighbor-smoothed trend is plotted as a black line. The interquartile range (Q1 to Q3) is shaded gray. In general, first ranked peptides with the highest responses tend to get higher scores than those of lower ranks, as indicated by the downward trend from left to right. The (b) PPA score distribution as well as the CONSeQuence (c) artificial neural network (ANN) and (d) support vector machine (SVM) score distributions all demonstrate weaker downward trends.

Following the trend demonstrated in Figure 2-5, there is wide scatter at each individual rank. However, the downward trend in scores as rank decreases suggests that PREGO is able to differentiate peptide responses in SRM experiments.

Figure 2-8b shows a similarly generated scoring profile for PPA on the same set of proteins. While there is a slight downward trend in the median, PPA assigns high scores to peptides at all ranks. The spreading shape of the distribution suggests that PPA is more likely to assign low scores to low responding peptides. For any given protein, PPA eliminates some of these low responding peptides from the pool of options and thus increases the odds for choosing a high responding peptide. CONSeQuence score distributions using both the artificial neural network option and the SVM option are depicted in Figures 2-8c and 2-8d, respectively. In this data set CONSeQuence produces a slight downward trend in scores with poorer responding ranks, although the scatter in the distributions overwhelms any major trends.

While it is important that response prediction scoring schemes correlate with experimental peptide intensities, these algorithms will mainly be used to select multiple peptides per protein to quantify in the hopes that at least one produces a strong response. The approaches need not identify the highest responding peptide every time; to be effective they must be able to select at least one relatively strong responding peptide in a handful of guesses. Figure 2-9a asks the question: “if we selected N peptides for any given protein, would at least one of those peptides demonstrate high response?” We defined high response as being in the top 20% of peptides for each protein by rank-response. Given these criteria, on average PREGO correctly selects a high responding peptide 57% of the time on the first selection. Similarly, if two peptides per protein are

selected, then at least one is a high responder 80% of the time, and on average selecting three peptides produces a high responder 90% of the time. At each of these three stages PREGO selects high responders approximately 40% to 85% more often than the best competing methods.

As a baseline, Figure 2-9a includes statistical calculations for selecting peptides entirely at random. However, typically scientists select peptides to build SRM and PRM assays by employing several simple selection rules and choosing randomly amongst the peptides that pass those rules. We built a simple scoring scheme to capture the Bereman *et al* rules strategy that has bonuses for prolines (which produce strong fragmentation signatures) and penalties for methionine (which can be oxidized), asparagine/glutamine (which can be deamidated), glutamine/glutamic acid in the n-terminal position (which can cyclize to form pyroglutamic acid), and carbamidomethyl-cysteine in the n-terminal position (which can also cyclize). The rules-based "score" is a summation of values across all of the *n* amino acids in a peptide:

$$Rules\ Based\ Score = \sum_{i=1}^{n} \left\{ \begin{array}{c|c} P_i & 5 \\ M_i & -10 \\ N_i, Q_i & -1 \\ Q_1 E_1 C_1 & -10 \\ other_i & 0 \end{array} \right\}$$

Not surprisingly this strategy performs somewhat better than the baseline of randomly guessing.

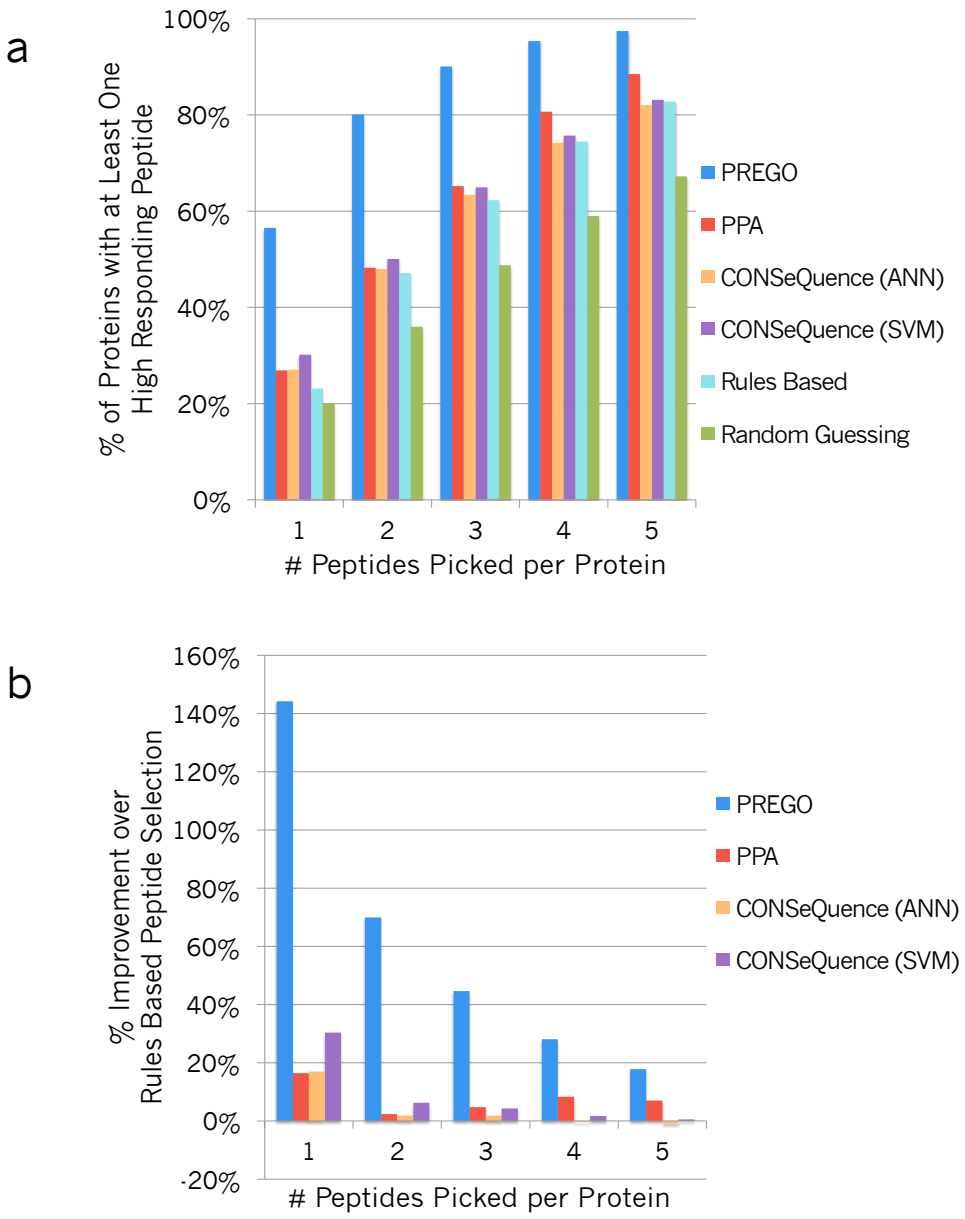


Figure 2-9: Percentage of proteins with at least one high-responding peptide, given N peptides picked.

(a) PREGO (blue), PPA (red), and CONSeQuence artificial neural network (ANN, orange) and support vector machine (SVM, purple) machine learning-based scorers are compared to randomly guessing to select peptides (green) and the simple scoring function described in Equation 2 (cyan) based on common rules in the literature. Scorers are graded based on the likelihood that

for any given protein, they could predict at least one high-responding peptide given N guesses. This is analogous to the strategy of picking N peptides to produce at least one useful peptide for each protein. For example in Figure 2-5 the top 1 - 5 peptides picked in CASZ1 have red borders and the high-responding peptides are shaded in blue. (b) The same four learning-based scorers as a percentage improvement over rules-based peptide selection. PREGO is dramatically better than the other approaches tested here at predicting high-responding peptides given five or fewer chances. All scoring data is based on the Stergachis *et al* SRM testing data set.

Figure 2-9b illustrates the relative improvement of PREGO and the other various trained approaches over the rules based approach. All of the trained approaches improve over the rules based approach when only considering the top peptide. However, it is rare that scientists choose only a single peptide per protein for targeted assays. As one chooses more peptides at random there is an increasing chance that at least one is a high responding peptide, which correspondingly makes it increasingly harder to do a better job. An unexpected result is that when choosing two or more peptides from the Stergachis *et al* data set, simply using the Bereman *et al* rules performs essentially equivalently to the PPA and CONSeQuence methods. PREGO, on the other hand, continues to show increased performance over the rules based approach when choosing a typical number of peptides for targeted assays.

2.8 Discussion

It is important to note that in the situation of predicting peptides for building SRM and PRM assays any level of success is still success. The factors that determine peptide response are largely unknown and are likely staggering in number and complexity. Consequently the vast majority of labs generating targeted assays do so by selecting peptides virtually at random using some variation of the rules described in Bereman *et al*. Improvement over these rules is the main measuring stick that peptide response prediction algorithms should be compared with.

Despite dramatically different training sets and machine learning architectures, PPA and both CONSeQuence scoring systems produce essentially identical success rates. We demonstrate that these software tools perform somewhat better than randomly selecting tryptic peptides for SRM assays, but not substantively better than using a rules-based random guessing approach for estimating peptide response characteristics. This suggests that there may be a glass ceiling for predicting SRM response behavior based on peptide responses in large-scale DDA data sets. Our results indicate that the PREGO algorithm produces a dramatic improvement over these other methods for building SRM assays.

Although the algorithmic improvements we propose likely provide some incremental improvement, we suspect that the large majority of PREGO's success stems from our training data set selection. In particular, we believe that training from DIA data sets using the Q-Exactive HF allows us to more closely represent data acquisition strategies employed by traditional SRM triple-quad instruments. In addition, DIA allows us to more accurately predict transition response directly from peptide fragmentation,

instead of assuming that precursor intensities equate with fragment intensities. We find that there is an order of magnitude variation between product and precursor intensities, which suggests that training using transition responses ought to be more accurate than training from precursors alone. Another key improvement is that PREGO ensures robust generalization by cross validating the DIA trained artificial neural network with SRM data. As different mass spectrometers and LC conditions can have a profound effect on peptide ionization, training using multiple diverse types of data from different sources is essential.

We also note that the underperformance of PPA and CONSeQuence may be partially driven by two aspects of our evaluation approach. First, data acquisition in the testing data set was restricted to only doubly charged precursor ions, and second, peptide response was evaluated using only the single most intense y-type fragment ion from each peptide. These aspects represent important practical considerations commonly employed in SRM assays and were incorporated into the training of PREGO but not in PPA or CONSeQuence.

Peptide response prediction can also be used to improve peptide-centric DIA search engines. Search engines that take this approach to querying DIA data sets can benefit from increased sensitivity using an SRM-like data analysis workflow. However, by individually considering every peptide for all proteins in a database, the peptide-centric approach suffers from a significantly increased false discovery rate that must be accounted for using multiple hypothesis testing corrections, which consequently decrease any sensitivity gains. Instead of looking for every possible peptide, PREGO can drastically help narrow down the search space by first considering only a handful of high-responding

peptides per protein. A peptide-centric DIA search engine then only needs to look for low-responding peptides if high-responders are seen.

2.9 Critical Evaluation

We make one major assumption in the construction of our DIA training data: we assume that crude peptides in our mixture are essentially at equimolar concentrations. We make this assumption because developing a training set from purified peptides would be prohibitively expensive. JPT estimates that these peptides are between 20% and 90% pure (personal communication), suggesting that there is somewhat less than 5x fold variation in their original concentrations. We believe that, while this variation is significant, the unknown level of variation in proteoforms present for each gene product would overwhelm it if we were to use biological samples, such as with the PPA or CONSeQuence methods. We also believe that the benefits of removing the assumption that high ranked peptides in each protein produce equivalently high fragment ion intensities outweighs any detriments in using crude peptides. On the other hand, training using the single most intense y-type fragment ion for each peptide might bias PREGO towards preferring peptides with dominant fragmentation pathways. Also, the most intense fragment ion by DIA might differ from the most intense fragment ion by SRM where collision energies can be tuned to produce the most reliable and easy to detect fragmentation on a peptide-by-peptide basis.

Similarly, varying efficiencies in tryptic digestion are also not accounted for with synthetic peptides. This may be an advantage from the standpoint of machine learning in that training goals are focused solely on identifying peptide sequences that produce

strong signals rather than being complicated by trying to interpret multiple layered sources of variation at the same time. The effects of incomplete digestion are difficult to ascertain in this experiment since the Stergachis *et al* SRM data set only assayed 1445 peptides with missed cleavages (1.2%). However, incomplete digestion can be a significant concern when interpreting particular classes of peptides, for example phosphopeptides. In the future additional layers of focused training or filtering may help account for digestion efficiency.

It is important to note that although PREGO performs better than alternative methods, there is still considerable variability between peptide scores and ranks within each protein. This is primarily due to the fact that peptide transition response is the product of many complex factors, only some of which can be captured using amino acid frequency-based physiochemical properties. The gold standard for predicted peptide response remains as experimental evidence derived from synthetic proteins. The utility of PREGO is primarily in situations where experimental data from controlled systems is expensive, time-consuming, or even impossible to generate. Considerable room for improvement still remains with future prediction methods to use more diverse training data sets and more complex properties crafted for modern proteomics methods that consider secondary and tertiary gas-phase structure and interactions.

2.10 Conclusions

We present a new method, PREGO, for predicting high responding peptides to aid in generating SRM and PRM assays. Our approach uses DIA experimental data of equimolar synthetic peptides to train an artificial neural network using 11 features

selected with a Pearson correlation-based minimum redundancy, maximum relevance algorithm. We have validated our software using a massive SRM data set measuring virtually every possible tryptic peptide from over 700 proteins.

We designed PREGO to make it easy to train new neural network models based on future data sets. We expect that as comprehensive DIA or PRM experiments of synthetic peptides are performed, the resulting data sets could be used to improve the accuracy of the approach. New models can be constructed based on specific experimental conditions; in particular we imagine designing models to predict PTM modified peptide responses, such as those of captured phosphopeptides using immobilized metal affinity chromatography (IMAC) or titanium dioxide enrichment. All that is required to retrain PREGO is a tab-delimited text file containing two columns: peptide sequences and experimental intensities. PREGO can score peptides for predicted response levels using a text file containing a single column of sequences.

While PREGO can be used for predicting the best responding SRM peptide, it makes no attempt to predict the best responding transition. Other modeling software, such as the thermodynamic peptide fragmentation model presented by Zhang(70, 71) will be required to make those predictions. Here we see inexpensive synthetic crude peptides as another answer. Due to the variability in actual abundance it is hard to estimate specific best responding SRM peptides from a massively parallel crude mixture. However, we intend to use PREGO to predict generally which peptides will be worth targeting and use inexpensively purchased synthetic crude peptides to identify preferred y-type ion transitions from MS/MS experiments. These issues are rendered moot with regards to PRM experiments because in that methodology all fragment ions are measured.

PREGO is written in Java and binaries will be available in an upcoming iteration of the Skyline software. We have released source code for PREGO on GitHub at https://github.com/briansearle/intensity_predictor/ under the Apache 2 license. The MS/MS data files used to train PREGO are available in mzML standard format at http://proteome.gs.washington.edu/SearleMCP/ and in RAW format at https://chorusproject.org/anonymous/download/experiment/-8935943952383739133. The exhaustive SRM training cross validation data is available on PanoramaWeb at https://panoramaweb.org/labkey/PREGO_manuscript.url.

3 COMPREHENSIVE PEPTIDE QUANTIFICATION FOR DATA INDEPENDENT ACQUISITION MASS SPECTROMETRY USING CHROMATOGRAM LIBRARIES.

3.1 Summary

Data independent acquisition (DIA) mass spectrometry is a powerful technique that is improving the reproducibility and throughput of proteomics studies. We introduce a new experimental workflow that uses this technique to construct chromatogram libraries that capture fragment ion chromatographic peak shape and retention time for every detectable peptide in an experiment. These coordinates calibrate information in spectrum libraries or protein databases to a specific mass spectrometer and chromatography setup, and enable sensitive peptide detection in quantitative experiments. We also present EncyclopeDIA, a software tool for generating and searching chromatogram libraries, and demonstrate the performance of our workflow by quantifying proteins in human and yeast cells. We find that by exploiting calibrated retention time and fragmentation specificity in chromatogram libraries, EncyclopeDIA can detect and quantify >50% more peptides from DIA experiments than with DDA-based spectrum libraries alone.

3.2 Introduction

Over the past two decades the continued refinement of proteomics methods using liquid chromatography (LC) coupled to tandem mass spectrometry (MS/MS) has enabled a deeper understanding of human biology and disease(2, 3). Recently data independent acquisition(11, 72) (DIA), in which the mass spectrometer systematically acquires MS/MS

spectra irrespective of whether or not a precursor signal is detected, has emerged as a powerful alternative approach to data dependent acquisition(4) (DDA) for proteomics experiments. In current DIA workflows, instrument cycle is structured such that the same MS/MS spectrum window is collected every 1 to 5 seconds, enabling quantitative measurements using fragment ions instead of precursor ions. This approach produces data analogous to targeted parallel reaction monitoring (PRM), except instead of targeting specific peptides, quantitative data is acquired across a predefined mass to charge (m/z) range. One trade-off is that to cover the m/z space where the majority of peptides exist, the mass spectrometer must be tuned to produce MS/MS spectra with wide precursor isolation windows that often contain multiple peptides at the same time. These additional peptides produce interfering fragment ions, and database search engines for DDA that rely on a precursor isolation window of at most a few daltons can struggle to detect the signal for a particular peptide from that background interference. The PAcIFIC approach(15) attempts to overcome this difficulty by using multiple gas-phase fractionated injections of the same sample to increase precursor isolation at the cost of both sample and instrument time.

Peptide-centric tools analyze DIA measurements for individual peptides across all spectra in a precursor isolation window. Spectrum library search tools for DIA data(23–25) use fragmentation patterns and relative retention times from previously collected DDA data. In contrast, other tools such as PECAN(22) query DIA data using just peptide sequences and their predicted fragmentation pattern without requiring a spectrum library. While library searching can achieve better sensitivity than PECAN, the approach is limited to detecting only analytes represented in the library. In addition, the quality of library-

based detections is only as strong as the quality of the library itself. Because mapping fragmentation patterns and retention times across instruments and platforms is difficult, many researchers prefer to simultaneously acquire both DDA and DIA data from their samples(26, 27). While this implicitly increases the acquisition time and sample consumption, it becomes possible to detect peptides using the DDA data while making peptide quantitation measurements using the DIA data. However, detection sensitivity is inherently limited to that of the DDA data.

Typically tens to hundreds of biological samples are processed and analyzed using LC-MS/MS in quantitative proteomics experiments. In DDA workflows each individual sample is informatically processed alone to account for stochastic variation in data acquisition. The regularity of DIA allows researchers to make peptide detections in one sample and transfer those detections to other samples(73). Here we extrapolate this concept by collecting certain runs where data acquisition is tuned to improve peptide detection rates, while collecting other runs with a focus on quantification accuracy and throughput. Results from runs dedicated to peptide detection are formed into a DIA-based chromatogram library. In a chromatogram library, we catalog retention time, precursor mass, peptide fragmentation patterns, and known interferences that identify each peptide on our instrumentation within a specific sample matrix.

We have developed EncyclopeDIA, a library search engine that takes full advantage of chromatogram libraries, and we demonstrate a substantial gain in sensitivity over typical DIA and DDA workflows. EncyclopeDIA also contains several new approaches to automate transition refinement to remove fragment ion interference,

improving the quality of quantification. This tool is instrument vendor neutral and available as an open source project with both a GUI and command line interface.

3.3 Methods

Complete methods are discussed in Appendix C.

3.4 Chromatogram library generation.

Chromatogram libraries differ from spectrum libraries in that they are generated from a small collection of narrow-window DIA experiments, rather than from DDA. We use a data acquisition scheme (Figure 3-1a) similar to PAcIFIC(15) for constructing chromatogram libraries. Briefly, for each experiment we create a representative sample, which pools subaliquots of each biological sample. We acquire six or more gas-phase fractionated runs from this pooled sample in an effort to comprehensively study all of the peptides in the pool. The gas-phase fractionated samples are injected into the mass spectrometer multiple times, with each injection staggered to acquire a fraction of the m/z space via overlapping 4m/z "narrow" precursor isolation windows. After overlap deconvolution these experiments effectively have 2 m/z precursor isolation (analogous to if we had conducted targeted PRM acquisition) except we are targeting all precursors between 400 to 1000 m/z. Previously we have shown that this type of DIA experiment can produce substantially richer peptide detection lists than similarly acquired DDA experiments(22).

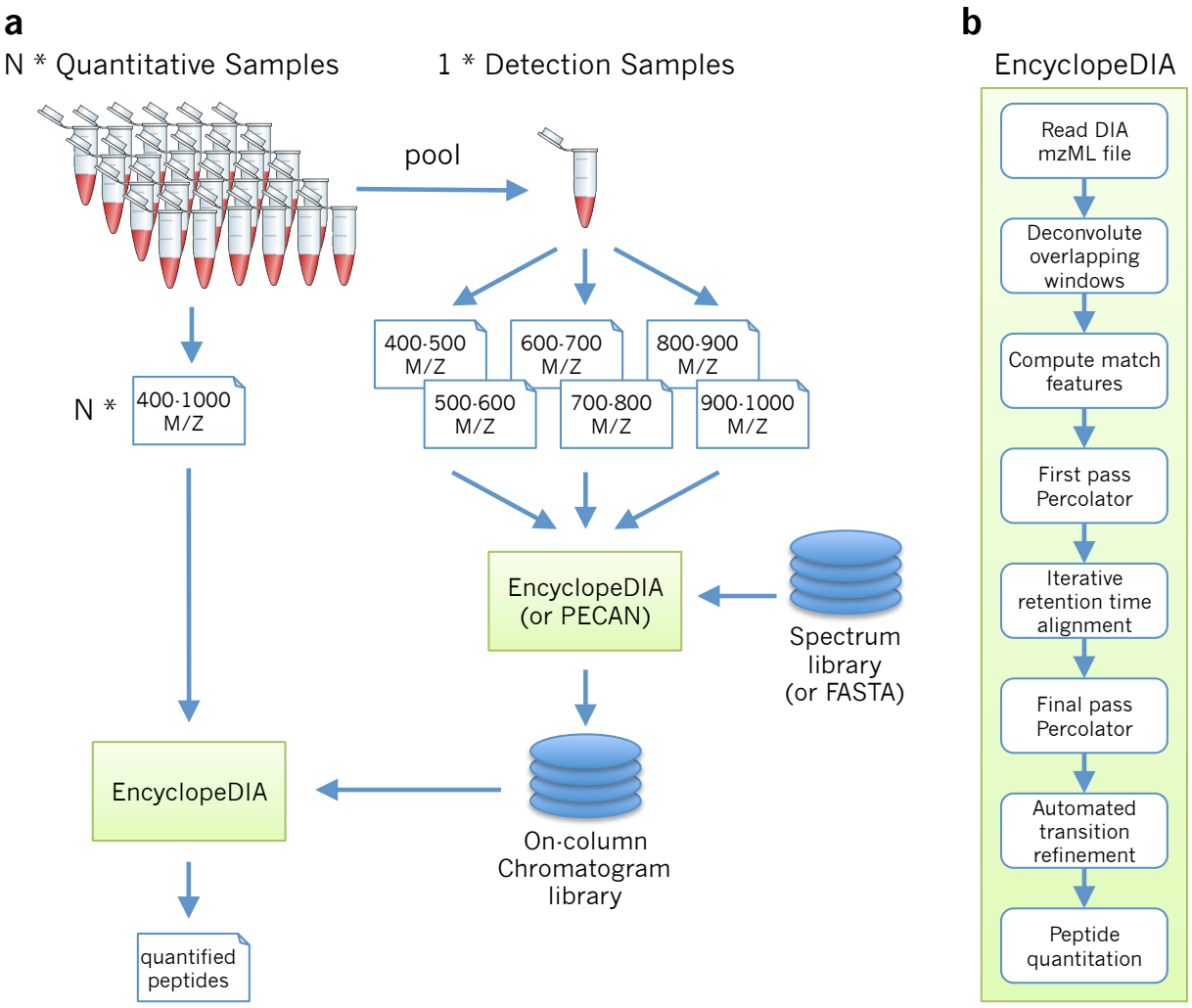


Figure 3-1: An approach for quantifying peptides with chromatogram libraries.

(a) The chromatogram library generation workflow. Briefly, in addition to collecting wide-window DIA experiments on each quantitative replicate, a pool containing peptides from every condition is measured using several staggered narrow-window DIA experiments. After deconvolution, these narrow-window experiments have 2 m/z precursor isolation, which is analogous to targeted parallel reaction monitoring (PRM) experiments, except effectively targeting every peptide between 400 and 1000 m/z. We detect peptide anchors from these experiments using either EncyclopeDIA (searching a DDA spectrum library) or PECAN (using a protein database) and chromatographic data about

each peptide is stored in a chromatogram library with retention times, peak shape, fragment ion intensities, and known interferences tuned specifically for the LC/MS/MS setup. EncyclopeDIA then uses these precise coordinates for m/z, time, and intensity to detect peptides in the quantitative samples. (b) The EncyclopeDIA algorithmic workflow for searching spectrum and chromatogram libraries. After reading and deconvoluting DIA raw files, EncyclopeDIA calculates several retention time independent feature scores for each peptide that are amalgamated and FDR corrected with Percolator. Using high confidence peptide detections, EncyclopeDIA retention time warps detections to the library, determines the retention time accuracy, and reconsiders outliers. After a second FDR correction with Percolator, EncyclopeDIA autonomously picks fragment ion transitions that fit each non-parametrically calculated peak shape and quantifies peptides using these ions.

While this data acquisition strategy would be impractical to perform for every biological sample, when applied to the pool it provides the mass, retention time, and fragmentation coordinates for virtually every detectable peptide in the experiment, which we use to lookup the peptides in quantitative samples.

3.5 The EncyclopeDIA workflow.

EncyclopeDIA is comprised of several algorithms for DIA data analysis (Figure 3-1b) that can search for peptides using either DDA-based spectrum libraries or DIA-based chromatogram libraries. The algorithms in this workflow are described in full detail in Appendix C. Briefly, the EncyclopeDIA workflow starts with reading raw MS/MS data in mzML files into an SQLite database designed for querying fragment spectra across precursor isolation windows. If fragment spectra are collected using overlapping windows,

they are deconvoluted on the fly during file reading. Libraries are read as DLIB (DDA-based spectrum libraries) or ELIB (EncyclopeDIA DIA-based chromatogram libraries). EncyclopeDIA determines the highest scoring retention time point corresponding to each library spectrum (as well as a paired reverse sequence decoy) using a scoring system modeled after the X!Tandem HyperScore(74). Fifteen auxiliary match features (not based on retention time) are calculated at this time point. These features are aggregated and submitted to Percolator 3.1(75), a semi-supervised SVM algorithm for interpreting target/decoy peptide detections, for a first pass validation. EncyclopeDIA generates a retention time model from peptides detected at 1% FDR using a non-parametric kernel density estimation algorithm that follows the density mode across time. Any target or decoy peptide in the feature set that does not match the retention time model is reconsidered up to 5 times until we find a highest scoring retention time point that matches the model. The retention time-curated feature sets are submitted to Percolator for final pass validation at 1% FDR.

3.6 Comparison between spectrum libraries and chromatogram libraries.

EncyclopeDIA can be used to query DIA data with DDA-based spectrum libraries. However, the benefit of the algorithms in EncyclopeDIA become transparent when using DIA-based chromatogram libraries. EncyclopeDIA can generate ELIB chromatogram libraries from gas-phase fractionated runs using DDA-based spectrum libraries if they are available, or using Walnut, which is a built-in, performance optimized re-implementation

of the PECAN algorithm(22) to search protein sequence FASTA databases. The resulting ELIB report can be fed back into EncyclopeDIA for chromatogram library searching. While this approach is inherently limited to the detectable proteins in the narrow-window pool, our perspective is that except for rare variants, very few quantitatively reliable peptides will be detectable in the wide-window data that are not also detectable in the narrow data. In cases where rare variants are important to a study or if samples are likely to represent very disparate proteomes, EncyclopeDIA can also generate chromatogram libraries from multiple batches of narrow-window acquisitions from different sample pools.

We evaluated the chromatogram library strategy using peptides derived from a HeLa S3 cell lysate as a representative high-complexity proteome. To this end we constructed a chromatogram library from six gas-phase fractionated DIA runs with 52 overlapping 4 m/z-wide windows, which produced 300 2 m/z-wide windows spanning 400.43 to 1000.70 m/z after deconvolution. Following the scheme in Figure 3-1a, we searched the narrow-window data against a HeLa-specific DDA spectrum library containing 166.4k unique peptides. This produced a chromatogram library containing 99.6k unique peptides where the retention times, fragmentation patterns, and interference likelihoods were calibrated to our mass spectrometer and HPLC setup. We performed an analogous approach using Walnut to detect peptides directly from the narrow-window DIA data using a Uniprot Human FASTA database, which generated a 53.2k peptide chromatogram library. The performance separation between these two library-generation methods is in part because the spectrum library represents a more targeted search space.

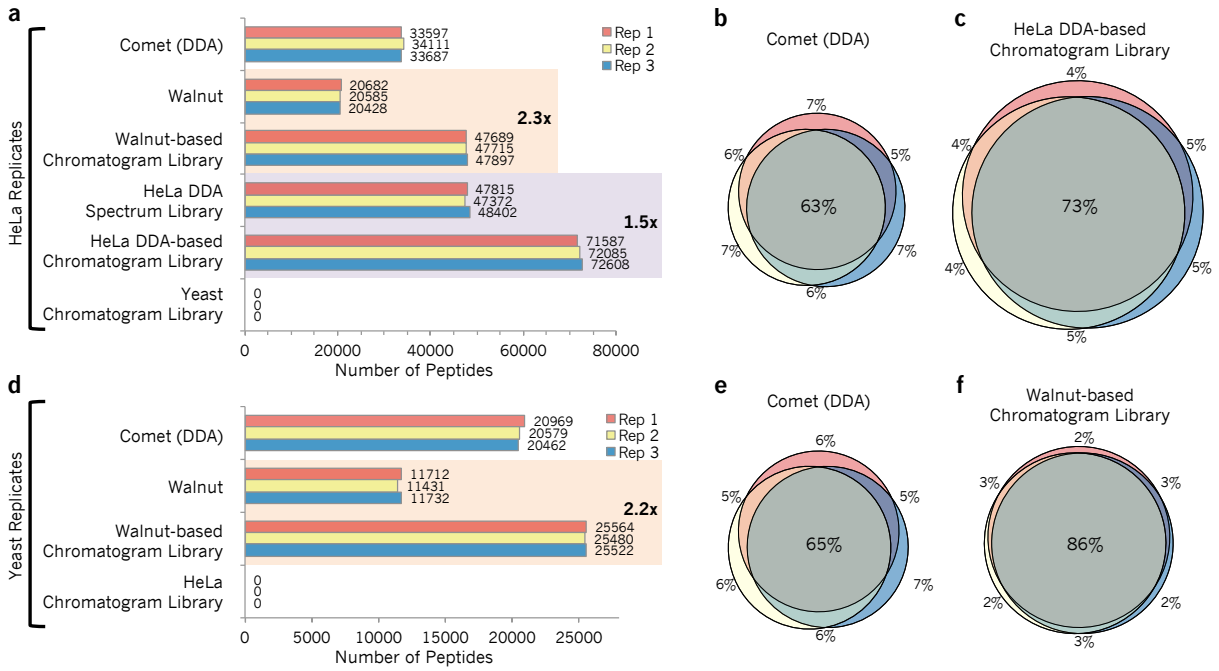


Figure 3-2: Untargeted peptide detection rates using DDA and DIA from human and yeast cell lysates.

We used EncyclopeDIA to search chromatogram and spectrum libraries, while we used Comet and Walnut to search DDA and DIA data directly using FASTA protein databases. (a) The number of peptide detections at 1% peptide FDR in triplicate HeLa injections. (b) The overlap in HeLa S3 peptide detections between replicates using DDA searched by Comet and (c) using DIA searched by EncyclopeDIA where the size of Venn diagram circles in HeLa analyses are consistent with the number of detections. (d) The number of peptide detections at 1% peptide FDR in triplicate BY4741 yeast injections. (e) The overlap in yeast peptide detections between replicates using DDA searched by Comet and (f) using DIA searched by EncyclopeDIA where the size of circles are consistent with the number of yeast peptide detections.

In addition to generating the library, we also collected triplicate wide-window DIA runs with 52 overlapping 24 m/z-wide windows from the same sample. From these runs we were able to detect an average of 20.6k peptides from the Uniprot Human FASTA

database using Walnut. In contrast, we found an average of 47.8k peptides (2.3x increase) when we searched the Walnut-based chromatogram library with EncyclopeDIA (Figure 2a). Requiring only an additional 6 injections, this search strategy found nearly an equal number of peptides compared to searching the SCX-fractionated, 36 injection DDA-based spectrum library (an average of 48.7k peptides). Finally, we found an average of 72.3k peptides when searching against the chromatogram library constructed using the DDA-based spectrum library. Here we detected over 2x more peptides than our benchmark top-20 DDA experiments (Figure 3-3). Despite this increased detection rate, we still find that DIA produces more consistent results compared to DDA, as indicated by the overlap in peptide detections between triplicate injections (Figure 3-2b and 3-2c).

Confirming these results, we performed the same analysis using a yeast cell lysate and found similar improvement rates when comparing Walnut versus EncyclopeDIA using a Walnut-based chromatogram library (2.2x increase, Figure 3-2d). Here we observe more modest gains over top-20 DDA experiments, which likely reflects the lowered proteomic complexity of yeast versus human cells and is echoed in the tight overlap (86%) between triplicate DIA injections versus DDA (Figure 3-2e and 3-2f). As is possible with any computational strategy that incorporates machine learning, we were concerned with the potential for overfitting that might manifest in over exaggerated peptide detection rates. To answer this question we searched the HeLa wide-window DIA data using the yeast chromatogram library (and vice versa) to verify that we see a negative result when searching the wrong library. As expected this result (Figure 2a and 2d) produced zero peptide detections that passed a 1% peptide FDR threshold.

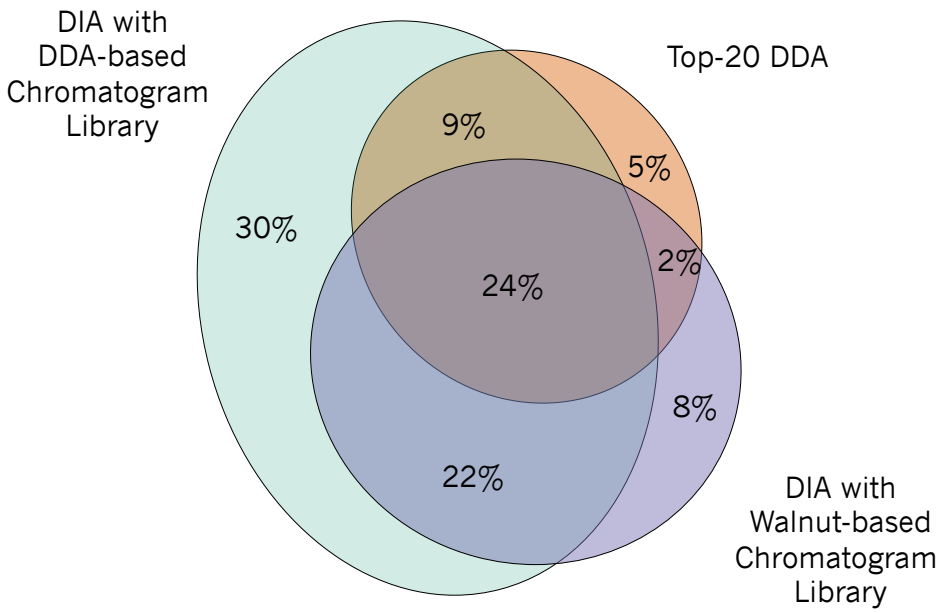


Figure 3-3: Peptide detection overlap between DIA and DDA.
An Euler diagram showing the overlap between unique peptide detections filtered at 1% FDR from a single HeLa replicate of either top-20 DDA or wide-window DIA searched with the DDA-based chromatogram library or the Walnut-based chromatogram library. EncyclopeDIA using a DDA-based chromatogram library detects 85% of peptides found using all three methods combined.

We also find that DIA analysis with chromatogram libraries is more sensitive at detecting low abundance proteins at a 1% protein FDR. Using tandem affinity purification tagging and quantitative Western blots, Ghaemmaghami *et al*(76) quantified 3868 yeast proteins with more than 50 estimated copies per cell. In this study we replicated strain and growing conditions as closely as possible to use their measurements as an independent benchmark. While both DDA and DIA confidently detect the majority of proteins at levels above 10^4 copies per cell, DIA outperforms DDA by 49% with proteins estimated to have between 10^3 and 10^4 copies per cell and by 2x with proteins estimated between 10^2 and 10^3 copies per cell (Figure 3-4).

3.7 Improved retention time and fragmentation pattern calibration in chromatogram libraries.

One of the primary reasons on-column chromatogram libraries enable such high performance is that they exploit within run retention time reproducibility. Accurate retention time filtering is an important consideration when analyzing high-complexity proteomes with DIA and virtually all DIA library search engines make use of this data. Retention times in aggregate spectrum libraries are typically derived by linearly interpolating multiple DDA data sets to a known calibration space (such as that defined by the iRT standard (77)), which enables retention times to be comparable from run to run, or even across platforms. However these measurements usually contain some wobble due to errors introduced by assuming a linear fit. Figure 3-5a shows a typical spread of retention times in EncyclopeDIA detected peptides using a DDA spectrum library, which is 95% accurate within a spread of 5.1 minutes (Figure 3-5c).

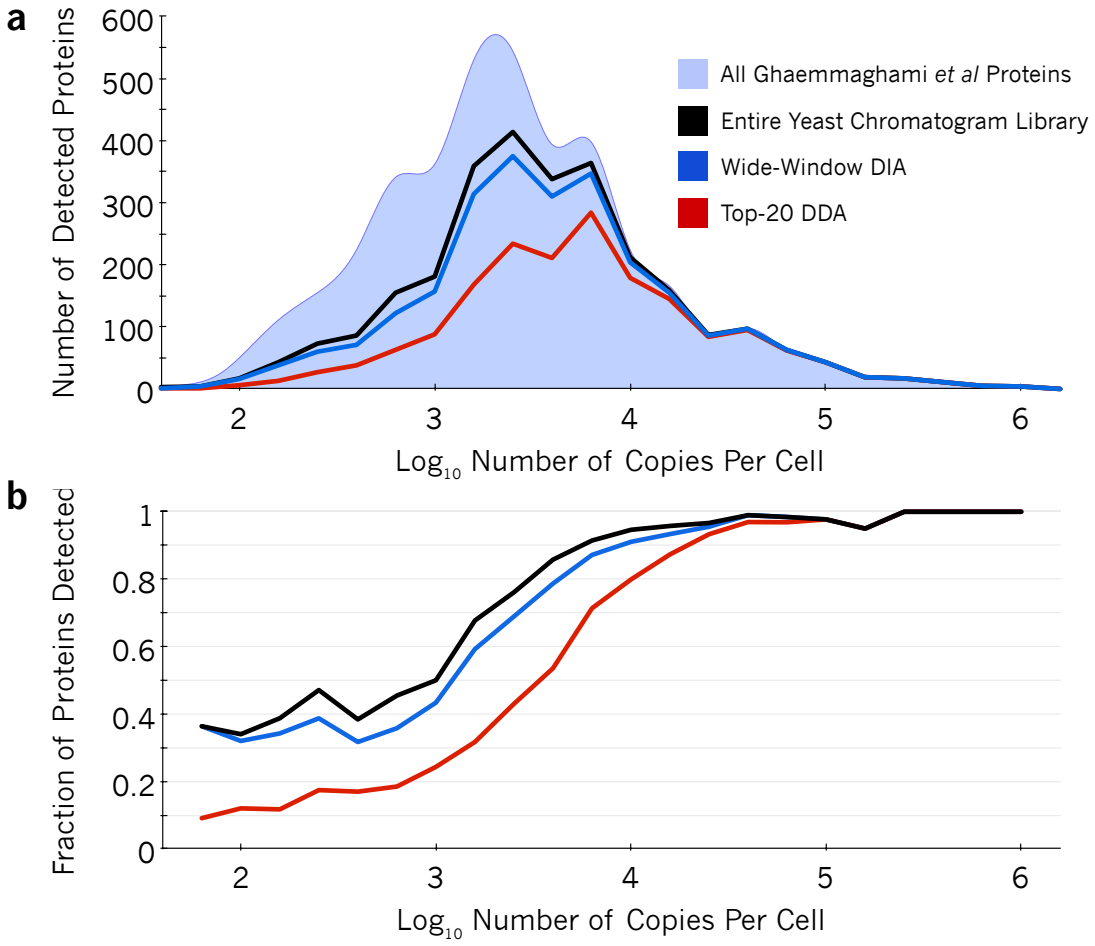


Figure 3-4: Protein detection rates scale with abundance.

The (a) number and (b) fraction of proteins detected in yeast at different orders of magnitude of abundance. Ghaemmaghami *et al* comprehensively estimated protein copies per cell in yeast (light blue area) using high-affinity epitope tagging. While top-20 DDA (red line) can measure some low abundant proteins at 1% protein-level FDR, the strategy only detected 48% of mid-range proteins with estimated copies per cell between 10^3 and 10^4. In contrast, at 1% protein-level FDR, wide-window DIA using a Walnut-based chromatogram library (blue line) detected 71% of these proteins and overall recapitulated 91% of proteins found in the entire Walnut-based chromatogram library (black line).

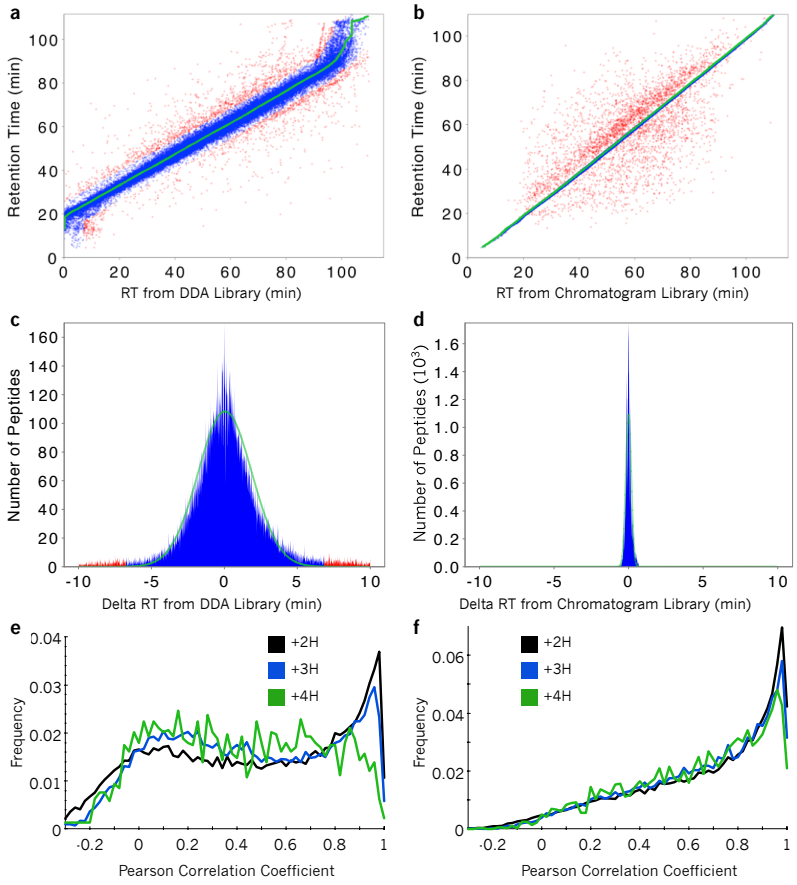


Figure 3-5: Retention time and fragmentation accuracy of the DDA spectrum library and the DIA chromatogram library.

Scatterplots comparing retention times from the (a) DDA spectrum library and the (b) DIA chromatogram library to those from in a single HeLa DIA experiment. Each point represents a peptide, where blue peptides fit the retention time trend (green) within a Bayesian mixture model probability of 5% and red peptides are outliers. (c) Retention times in the DDA spectrum library are 95% accurate to a window of 5.1 minutes, while (d) retention times in the chromatogram library are 95% accurate to 21 seconds. (e) The distribution of Pearson correlation coefficients between spectra in the DDA spectrum library and those detected from a single HeLa DIA experiment shows charge state bias, while (f) the distribution of correlation coefficients between spectra in the DIA chromatogram library and those from the same experiment shows much less bias.

In comparison, Figure 3-5b shows the typical spread of retention times in the chromatogram library, which is 95% accurate within 21 seconds (Figure 3-5d). This tightening of retention time accuracy is due to the fact that chromatogram libraries are collected on the same column as the wide-window acquisitions. Even if efforts are made to keep packing material, length, and gradient consistent, the dramatic gains in retention time accuracy with chromatogram libraries reflect variations that are difficult to control for, including packing speeds, pressures, and pulled tip orifice shapes. In addition, we find that DDA fragmentation patterns (Figure 3-5e) are often somewhat different than those collected in DIA experiments (Figure 3-5f). While DDA instrument methods usually tune MS/MS collision energies to the precursor charge and mass, some of this variation is likely due to fixed assumptions in charge states and precursor masses required by DIA methods when multiple precursors must be fragmented at the same time.

A subtle issue with DIA library searching when using generalized spectrum libraries is that many peptides generate the same fragment ions, either because of sequence variation, paralogs, or modified forms. While EncyclopeDIA attempts to control for this using background ion distributions to predict interference likelihoods, sequence variation due to homology or single nucleotide polymorphisms can be unintentionally detected as the wrong peptide sequence in certain circumstances. For example, a sequence variation of a valine to an isoleucine is relatively common, and the mass shift of a methyl group (+14/Z) will often place both peptides inside the same precursor isolation window when Z is 2 or greater. Using chromatogram libraries can provide some protection against these issues because the initial searches to generate the libraries are performed using narrow (2 m/z) precursor mass windows, and subsequent wide-window

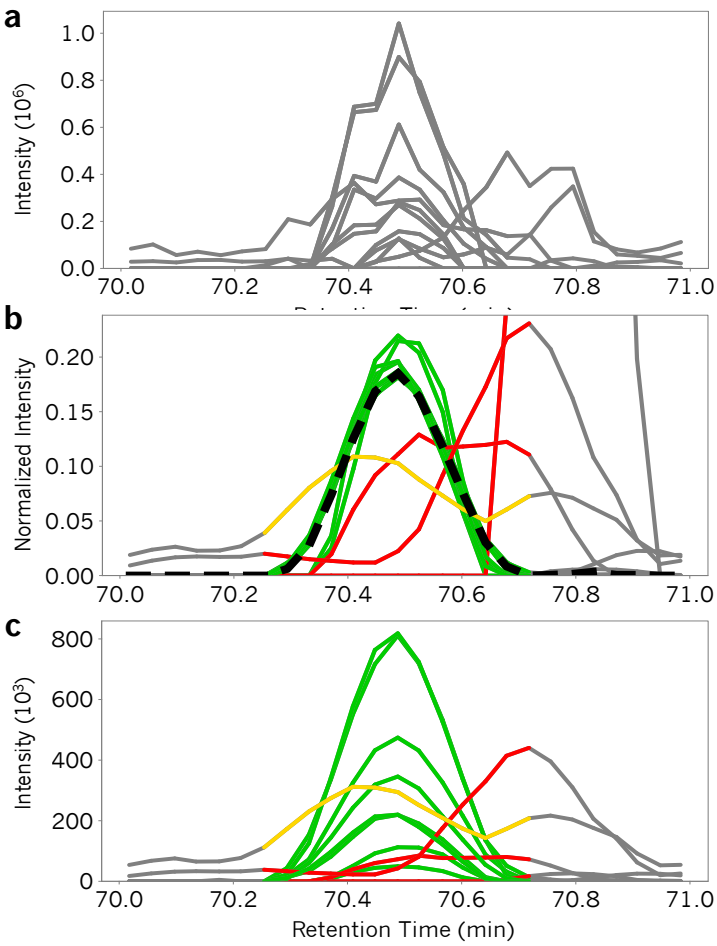


Figure 3-6: Schematic for automated transition refinement.

(a) After a retention time region for a peptide is detected using the primary EncyclopeDIA score, automated transition refinement is used to determine quantitative fragment ions. (b) Briefly, fragment ions are smoothed and normalized such that their area under the curve equals 1. At every retention time point, the median normalized intensity is calculated (dashed line). Normalized intensities that match this shape with a Pearson correlation coefficient >=0.9 are "quantitative" and labeled in green. Normalized intensities that fit with coefficients <0.9 and >=0.75 are labeled in yellow, and those with coefficients <0.75 are in red. (c) Quantitative ions can be integrated and summed to approximate the peptide intensity. In general, the median normalized intensity is robust to outliers (transitions with interference) and nonparametrically calculates peak shape without assuming a distribution.

searches benefit from precise retention time filtering. Additionally, EncyclopeDIA requires at least 25% of the primary score to come from ions that indicate the modified form to detect modified peptides when modified/unmodified peptide pairs fall in the same precursor isolation window (e.g. methionine oxidation).

3.8 Peptide and protein quantitation.

We present a novel algorithm for automated transition refinement to remove fragment ion interference and alleviate the need for manual curation (see Appendix C for more details). In short, after unit area normalizing all transitions assigned to a single peptide (Figure 3-6a), we determine the shape of the peak as the median normalized intensity at each retention time point (Figure 3-6b). Transitions that match this peak shape with Pearson's correlation scores >0.9 are considered quantitative (Figure 3-6c). We find that over 81% of peptides can be quantified with at least three transitions (Figure 3-4a) and that the transitions picked by our approach produce reproducible quantitative measurements between technical replicates in HeLa experiments (Figure 3-4b and 3-4c).

Combining peptide detections across multiple samples often increases false discoveries because false detections are usually found only in individual runs(78). To combat this, we recalculate global peptide FDR across all experiments in each study with Percolator and generate parsimonious protein detection lists that are also filtered to a 1% FDR. We use cross-sample retention time alignment(73) to help quantify peptides that are missing in specific samples. After filtering peptides based on coefficient of variance and measurement consistency we estimate protein quantities by summing fragment ion intensities across only sequence-unique peptides assigned to those proteins.

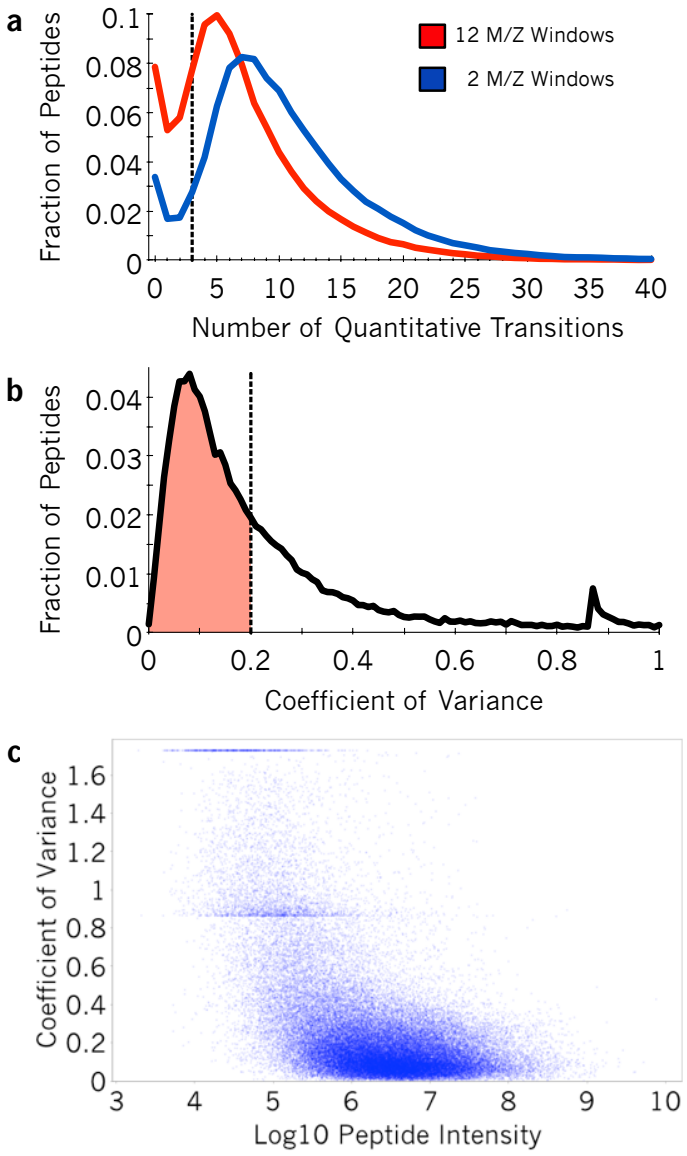


Figure 3-7: Quantitative reproducibility across replicates.

(a) The number of quantitative transition ions produced by each peptide is greatly affected by precursor isolation window. 93% of peptides detected in narrow-window experiments produce three or more transitions as compared to 81% of peptides in wide-window experiments. (b) The distribution of coefficient of variance (CV) in wide-window HeLa replicate experiments: 61% of peptides fit within a 20% CV. (c) Coefficient of variance is greatly affected by fragment ion intensity. CV streaks at 0.87 and 1.73 indicate 1 and 2 missing values, which are predominantly from low intensity peptides.

3.9 Determining global proteomic changes from serum starvation.

We used the chromatogram library approach to examine the effects of serum starvation in human cells. Serum starvation is a common step in signal transduction studies as serum contains several cytokines and growth factors that can confound signaling levels. It is commonly thought that serum starvation suppresses basal activity by reducing signaling activity that effectively resets cells to G0/G1 resting phase(79), although more recent experiments(30, 31) suggest otherwise. Serum starvation protocols vary widely from 2 to 24 hours, and this time frame is long enough to produce changes in protein levels resulting from transcriptional regulation. These changes are a source of variation that can have serious consequences when comparing between studies.

We designed a DIA quantitative experiment to map how the proteome of HeLa cells changes in response to serum starvation over time. We selected starvation times to match commonly used protocols. Of the 99.6k unique peptides in our chromatogram library, we recapitulated 93.5k unique peptides from 6,802 protein groups in at least one quantitative sample at a global protein FDR <0.01. Of these, 48.6k peptides (from 5,781 protein groups) produced at least three quantitative transition ions without interference, had <20% study-wide CVs, and were measured in every replicate of at least one time point. While at first these detection and quantification criteria may seem unusually stringent compared to typical proteomics experiments, narrowing our focus to confident measurements increased power in detecting subtle quantitative differences with high accuracy.

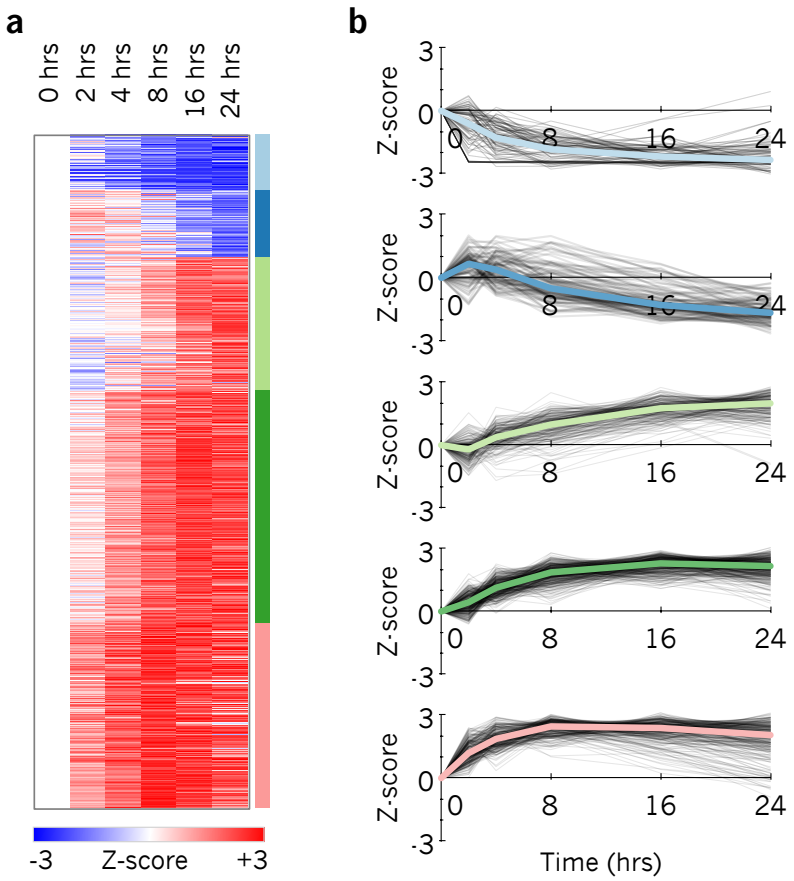


Figure 3-8: Protein quantification changes following serum starvation.

(a) Heatmap of 1097 proteins found to be quantitatively changing at a FDR corrected p-value<0.01 in HeLa. Colors are Z-score normalized and indicate the number of standard deviations away from the level at time 0. (b) Protein changes grouped into five K-means clusters showing separation between fast responding proteins (light blue, dark green, and pink) and delayed responses (dark blue, light green).

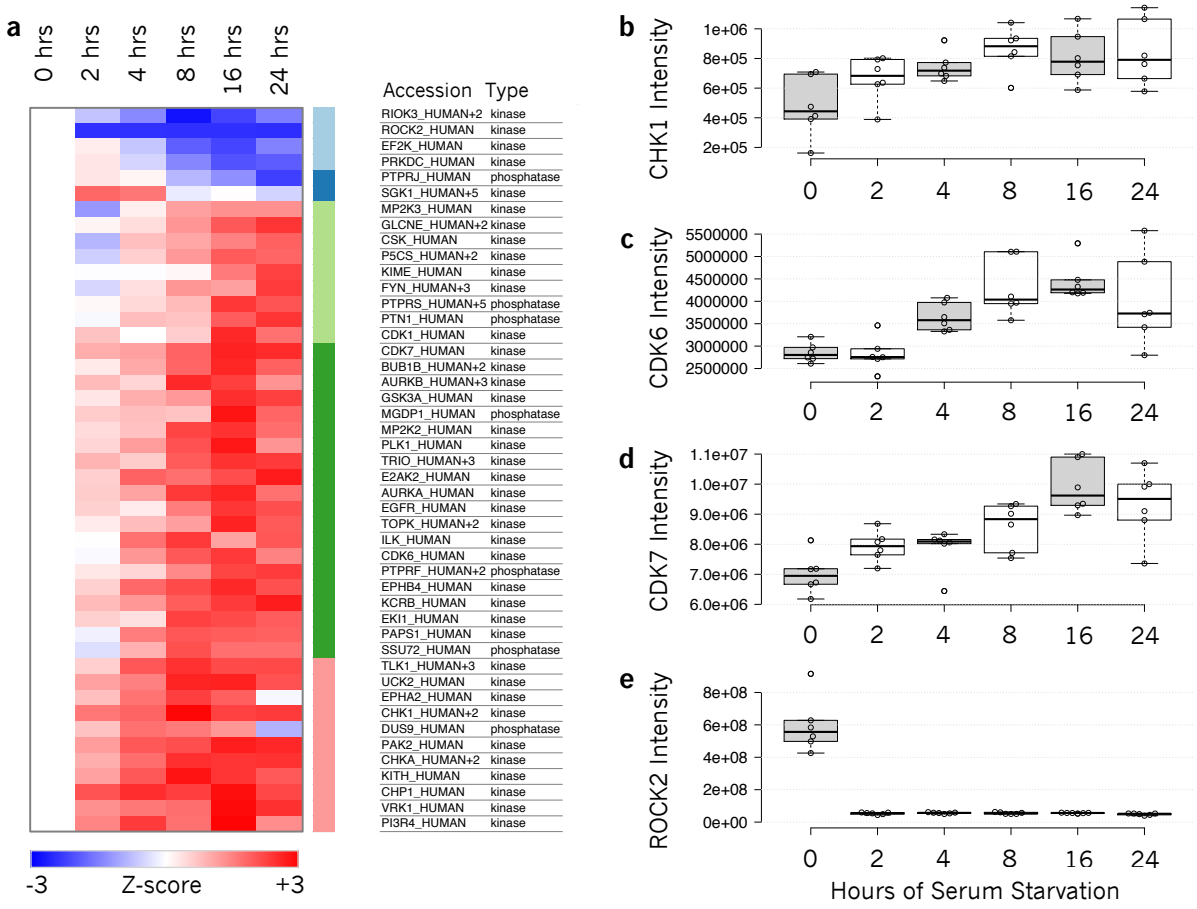


Figure 3-9: Changes in kinase expression following serum starvation.
(a) Heatmap of 39 kinases and 7 phosphatases found to be quantitatively changing at a FDR corrected p-value<0.01 in HeLa. Colors are Z-score normalized and indicate the number of standard deviations away from the level at time 0. K-means cluster groups are the same as from Figure 4. (b-e) Expression changes in known cell cycle regulator kinases (CHK1, CDK6, CDK7, and ROCK2) are out of sync over 24 hour serum starvation.

We found that 1097 protein groups in the HeLa proteome changed significantly over time at an FDR of 0.01. The temporal starvation profiles of these proteins fell into five groups (Figure 3-8) where the majority changing proteins increased in abundance. Several of these proteins are involved in expected pathways such as cell cycle regulation

(GO enrichment FDR=0.011), metabolism (GO enrichment FDR=0.011), and ubiquitination regulation (GO enrichment FDR=0.018). One advantage of our method is that quantitation is performed by summing peaks from several low interference fragment ions, which allows us to accurately quantify small changes. For example, we found that all eight of the observed components of the nuclear proteasome increased significantly by approximately 25%, which indicates nuclear maintenance consistent with G0/G1 resting phase.

We also observed significant regulation of the abundance of 39 kinases and 7 phosphatases (Figure 3-9). In particular, we found that EGFR levels increased by 30% over a 24 hour serum starvation time course (Figure 3-9), effectively sensitizing HeLa to the growth factor EGF. To confirm these experiments, we monitored relative changes in the phosphoproteome of HeLa after EGF stimulation at two common serum starvation times: 4 hours and 16 hours. We found that while phosphopeptide measurements at both time points directionally agreed, some phosphopeptide responses to EGF were stronger when cells were starved for 16 hours compared to when starving for only 4 hours (Figure 3-10). This increase corroborated our observation that EGFR protein levels increased from 4 to 16 hours of starvation. These protein and phosphopeptide-level changes underline a potentially significant source of variation when comparing phosphorylation signaling studies.

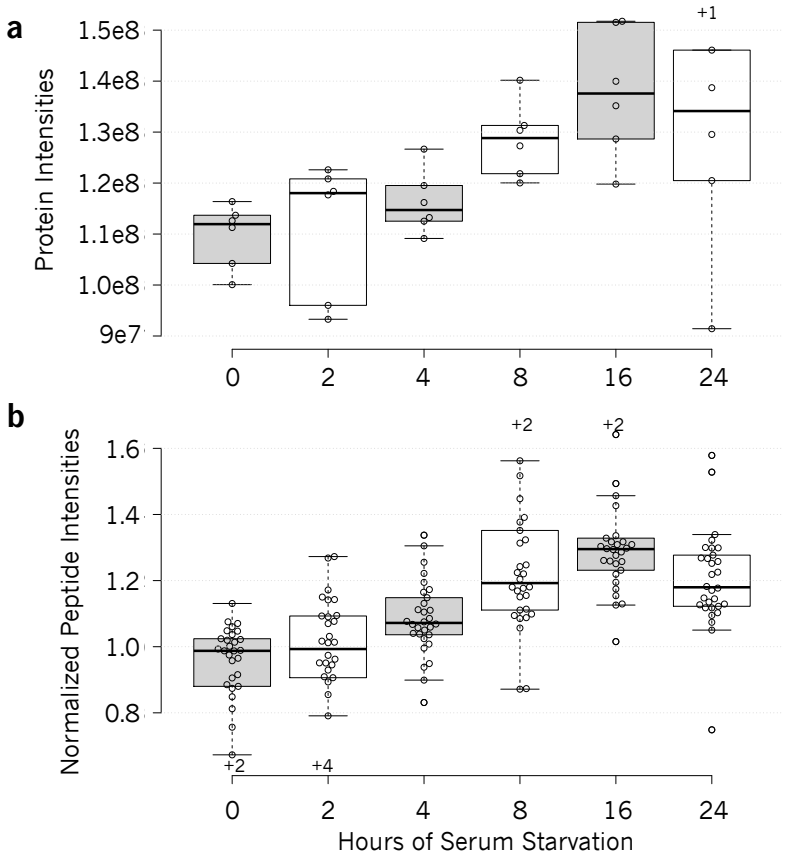


Figure 3-10: Boxplots showing intensity changes in EGFR following serum starvation.

(a) EGFR protein intensities for 6 replicates over 24 hours of serum starvation indicate an upward trend to a maximum 1.25x fold change at 16 hours. (b) Linear model normalized median peptide measurements across 6 replicates for all peptides assigned to EGFR indicate the same trend. Boxes indicate quartiles and bold lines indicate medians. Tukey-styled whiskers extend to data points that are less than 1.5x the interquartile range away from 1st/3rd quartile. All data points are plotted except where indicated on the plot with +X to indicate outliers. Alternating colors simply indicate every other time point.

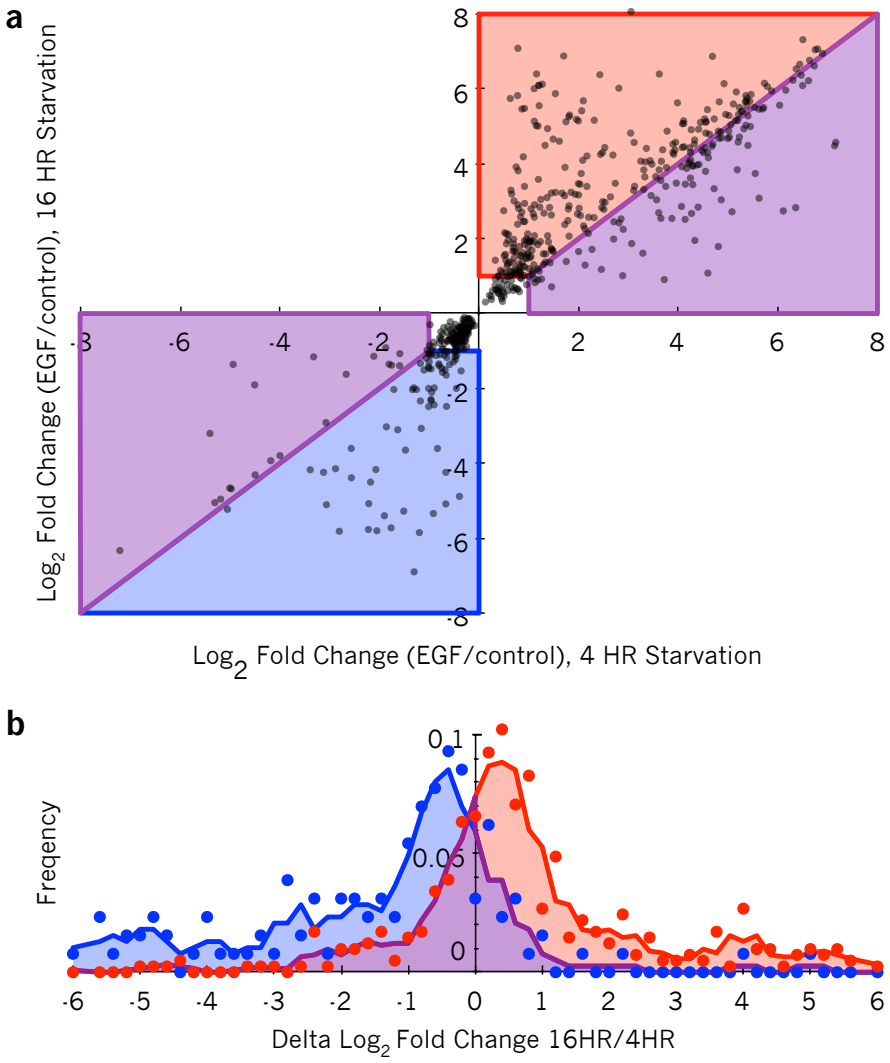


Figure 3-11: Changes in EGF phosphorylation regulation following different serum starvation protocols.

(a) A scatterplot of phosphopeptides that change significantly as a result of EGF stimulation (FDR<0.05) after 4 hours and 16 hours of serum starvation. If changes in serum starvation were an insignificant perturbation, phosphopeptide response to EGF would fall on a diagonal line. Corroborating our observation of an increase in EGFR, more phosphopeptides fall in the red and blue regions (stronger response after 16 hours starvation) rather than purple regions (stronger response after 4 hours starvation). (b) A histogram of delta changes in EGF response from the diagonal line. Dots indicate

frequencies binned at 0.2 Delta Log_2 fold change while lines are three-bin moving average smoothed values. The red distribution (matching the red and purple areas in the upper right quadrant) indicates that phosphopeptides that increase after EGF stimulation shows a median of 1.38x increased fold change when comparing a 16 hour starvation versus a 4 hour starvation, correlating with the 1.3x increase in EGFR observed at the proteome level. The blue distribution (matching the blue and purple areas in the lower left quadrant) indicate phosphopeptides that show a median decrease of 0.58x when comparing a 16 hour starvation to a 4 hour starvation and also indicate a sensitization of HeLa to EGF after a longer serum starvation.

3.10 Discussion

We have demonstrated an experimental strategy that enables comprehensive detection of peptides and proteins using chromatogram libraries. These libraries can be seeded either with a DDA spectrum library or generated in a DIA-only mode using Walnut for initial peptide searches. Finally, we showed that at the cost of only six additional narrow-window DIA runs, both of these strategies are more sensitive and reproducible relative to comparable DDA experiments. While this approach may be unrealistic for one-off experiments, we feel that in most quantitative proteomics studies the addition of these runs are a minor cost in exchange for a significant increase in sensitivity.

One important limitation of our method is that each chromatogram library is tuned for a specific mass spectrometer and chromatographic set up. In particular, we have observed that with the hand-pulled and packed columns used here, there is significant retention time variation between replicates run on different columns, even if effort is made to insure column consistency. We hypothesize that minor variations in packing speeds,

packing pressures, tip shapes, and column lengths can affect elution times and even peptide retention time ordering. This issue may be mitigated by acquiring a new library after a column change and retention time aligning the libraries to insure consistency. Future work remains to model these minor retention time shifts.

Another important consideration is library quality. All library searching strategies assume that entries in the library are correctly identified and consequently false positives in the library can be propagated as “true” positives by target/decoy analysis(80). This concern is potentially compounded in our approach, which can include up to two levels of library creation. Further work is necessary to improve FDR estimates for library searching in DIA experiments. In the meantime, we feel orthogonal filtering strategies are necessary to maintain conservative peptide detection lists. In addition to retention time fitting and 1% protein-level FDR filtering, in this work we require a minimum of three interference-free transitions and impose stringent measurement reproducibility requirements for peptides to be considered quantitative.

We have observed a complementarity of DDA and DIA through the use of building spectrum libraries to seed chromatogram libraries. Here the stochasticity of DDA sampling when coupled with offline peptide separation methods such as SCX fractionation can be exploited as a benefit in that only one observation of a peptide is necessary for inclusion in the library. With human samples, libraries constructed using previously recorded retention times and fragmentation patterns contained nearly twice the peptides as those constructed without prior knowledge. However, PECAN/Walnut can build on that knowledge by detecting peptide sequence variants illuminated by whole

exome sequencing(22), and we are exploring ways of generating chromatogram libraries that incorporate both pieces of data.

4 THESAURUS: QUANTIFYING PHOSPHOPEPTIDE POSITIONAL ISOMERS

4.1 Summary

Proteins can be phosphorylated at neighboring sites resulting in different functional states, and studying the regulation of these sites has been challenging. Here we present Thesaurus, a search engine that detects new positional isomers using site-specific fragment ions from parallel reaction monitoring and data independent acquisition mass spectrometry experiments. We apply Thesaurus to analyze phosphorylation events in the PI3K/AKT signaling pathway and show neighboring sites with distinct quantitative profiles, indicating regulation by different kinases.

4.2 Introduction

It is estimated that hundreds of thousands of residues from thousands of proteins are actively phosphorylated in every human cell(81). Many proteins are phosphorylated at neighboring sites(82) where over half of sites in multi-phosphorylated proteins are within four amino acids of each other(37). Several well-studied proteins make use of adjacent phosphorylation sites to act as switches (MAPK(83), CDC4(84)), timers (PER(85)) or as negative inhibition toggles (IRS1(86)) but global analysis of these phosphorylation clusters has remained impractical. Tandem mass spectrometry (MS/MS) of tryptic peptides is a key tool in discovering and quantifying sites of protein

phosphorylation. Typical phosphoproteomic workflows use data dependent acquisition (DDA) to collect MS/MS spectra based on peptide precursor m/z as peptides chromatographically elute. To increase the number of unique phosphopeptides that are sampled, DDA dynamically excludes peptides of the same m/z from being sampled twice within a narrow elution time. Software tools(39, 42, 87) detect specific positional isomers using site-specific fragment ions from those MS/MS spectra. However, phosphopeptides that exist as multiple positional isomers have 1) the same mass, 2) similar retention times, and 3) many of the same fragment ions, making them extremely difficult to differentiate. Thus MS/MS sampling using DDA and dynamic exclusion makes it challenging to sample multiple positional isomers.

Parallel reaction monitoring (PRM) and data independent acquisition (DIA) represent an alternative class of acquisition approaches that systematically collect MS/MS spectra across the chromatographic elution of the phosphopeptide, improving quantitative reproducibility. While PRM methods target specific phosphopeptide precursors(9), DIA methods iteratively sweep across m/z windows to acquire MS/MS spectra irrespective of m/z's that have been sampled previously(11). These methods are free of both: intensity biases during data collection and active exclusion of previously sequenced m/z, making it possible to detect closely eluting positional isomers. In addition, quantification can be performed on the more sensitive and selective product ion chromatograms, including those that are specific to each positional isomer.

4.3 Methods

Complete methods are discussed in Appendix D.

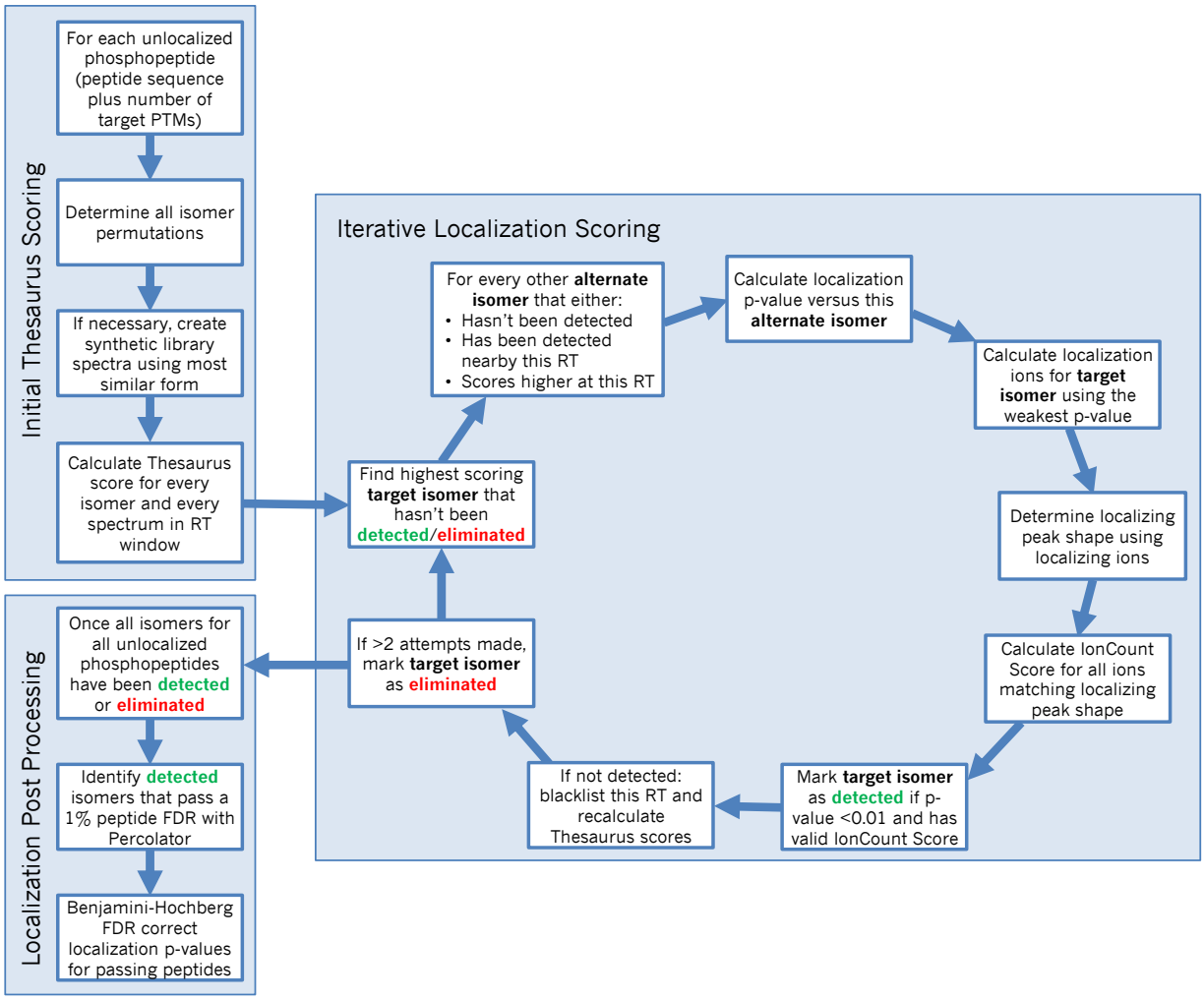


Figure 4-1: Thesaurus algorithmic workflow to search for phosphoisomers from DIA data

Briefly, for each unlocalized phosphopeptide in a spectrum library, Thesaurus iteratively detects positional isomers by finding the highest scoring isomer/retention time pair and calculating a p-value for localizing the isomer at that retention time. Up to two attempts are made to find the best retention time to localize each isomer. Positional isomers marked as detected are always saved for scoring. If all isomers for a unique peptide sequence were eliminated then the isomer with the highest Thesaurus score and best localization p-value is saved for scoring. Saved positional isomers are processed with Percolator, and localization p-values for passing isomers are also Benjamini-Hochberg

FDR corrected. Highlighted groups of steps refer to specific titled sections in Appendix D.

4.4 Detecting phosphopeptide positional isomers

Despite the strengths of these methods, mapping a phosphorylation site to a specific residue remains difficult. Rosenberger et al(88) proposed a method for determining the predominant modification site after library-search detection that determines the most likely positional isomer from fragment ions in a peak. An alternate approach(89) is to deconvolute DIA data with DIA-Umpire(19) for processing with PTMProphet(90), a site localizing tool originally designed for DDA. Positional isomers frequently co-elute; and a limitation of both of these approaches is that multiple isomers with similar retention times are competed against each other, where at most one will be detected. Here we extend these approaches and present a new DIA and PRM search engine named Thesaurus, which is designed to specifically look for and quantify all detectable positional isomers, including those that are not found in the library and isomers that co-elute.

Thesaurus detects phosphopeptides in spectrum libraries and uses those detections as retention time anchors to iteratively find new positional isomers that share many of the same fragment ions but differ in their site-specific ions (Figure 4-1). Thesaurus can detect multiple positional isomers at the same time point because it calculates localization probabilities directly using an interference distribution rather than by competing isomers against each other. For each unlocalized phosphopeptide (peptide sequence plus number of phosphorylations), Thesaurus determines every possible combinatorial positional isomer, extracts site-specific fragment ion signals for those

isomers, and calculates the probability that those ions would be observed by chance. Each ion has a unique likelihood of being interfered with across the experiment, and this probability of interference is highest with low M/Z ions (Figure 4-2). Thesaurus calculates this background frequency distribution for each precursor isolation window since this probability is heavily skewed by peptide mass and the fragment ion mass tolerance. If a combinatorial positional isomer is not present in our spectrum library, Thesaurus generates a synthetic spectrum using an approach similar to SpectraST(91) by shifting fragment ions from library spectra of known positional isomers to aid in the detection. Localization p-values are then FDR corrected using the Benjamini-Hochberg method.

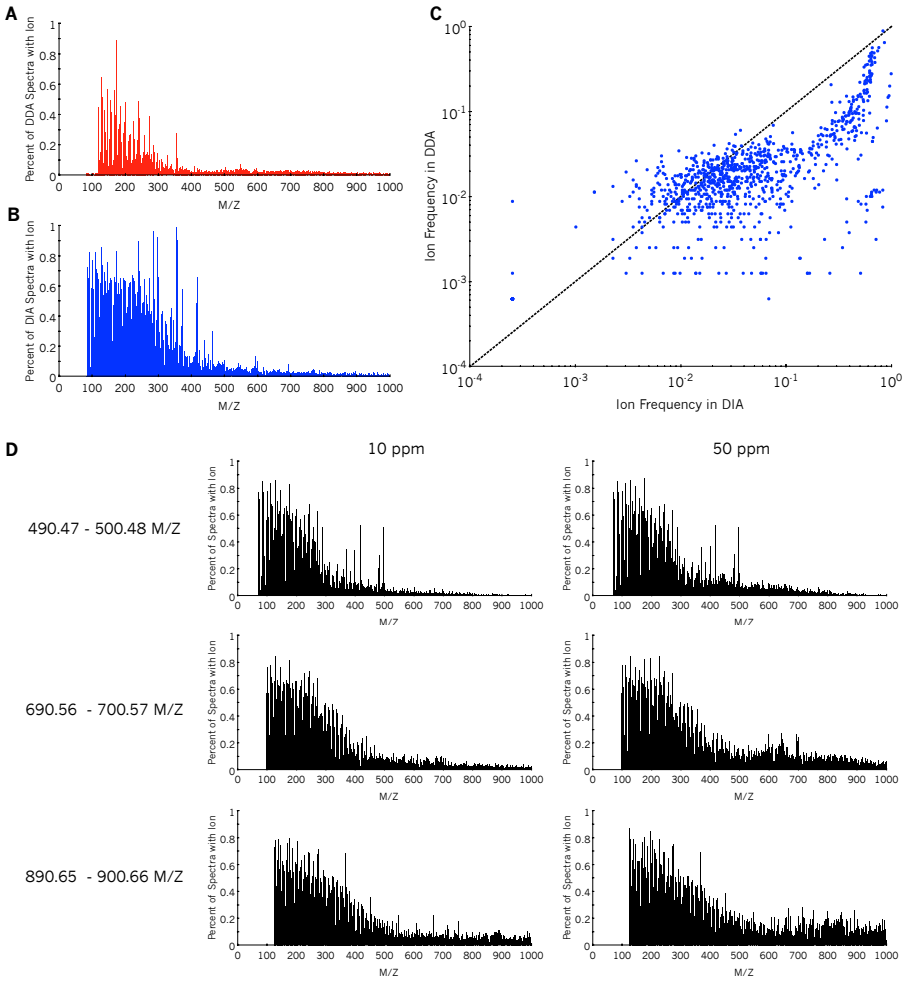


Figure 4-2: Increased interference in DIA experiments.

(A) The frequency in the DDA library that fragment ions are present (+/- 10 ppm) in each MS/MS with precursors of m/z from 590.5183 to 600.5229. (B) The frequency that fragment ions are present (+/- 10 ppm) in each MS/MS scan generated for the 590.5183 to 600.5229 m/z isolation window. (C) The relative ion frequency for the most common ion at each 1 m/z bin in both DIA and DDA. While many fragment ions found 1/100 or less are relatively consistent between the two data sets, many fragment ions show dramatically higher frequency in DIA spectra than their DDA counterparts. (D) Changes in the frequency distribution across precursor isolation windows and at different fragment mass tolerances. The frequency of higher mass fragment ions is increased with high mass precursor window and higher fragment mass tolerance.

4.5 Evaluation of Thesaurus

One major advantage of analyzing PRM and DIA phosphoproteomes with Thesaurus is that each positional isomer can be quantified using site-specific ions even if the precursor signals are convolved. Thesaurus quantifies phosphopeptides by automatically determining peak boundaries based on the site-specific ions and choosing other fragment ions that fit the same shape. We validated the false discovery estimation of our algorithm using the SWATH-MS DIA synthetic phosphopeptide reference set(88). Here Thesaurus compares favorably to other DIA localization methods, including IPF(88), PIQED(89), and DIA-Umpire(19) followed by Comet(92) and Ascore(39) (Figure 4-3), producing both more detections and a more accurate error estimate at a 5% localization FDR threshold. In addition to correctly localizing 240 phosphopeptides, Thesaurus was able to detect 11 positional isomer rearrangements (Figure 4-4). Gas-phase positional isomer rearrangement(93) is a chemical reaction that occurs in electrospray and can impede unambiguous site assignment. We confirmed that these rearrangements were a) detected at the same retention time as the expected isomer, and b) produced at least two site-specific fragment ions that fit the expected isomer peak shape without any interference. Thesaurus flags these sites and provides the first global proteomics method to explore the mechanism of gas-phase rearrangement, as well as the ability to separate chemical rearrangements from truly co-eluting phosphopeptides using quantitative fragment ion ratios.

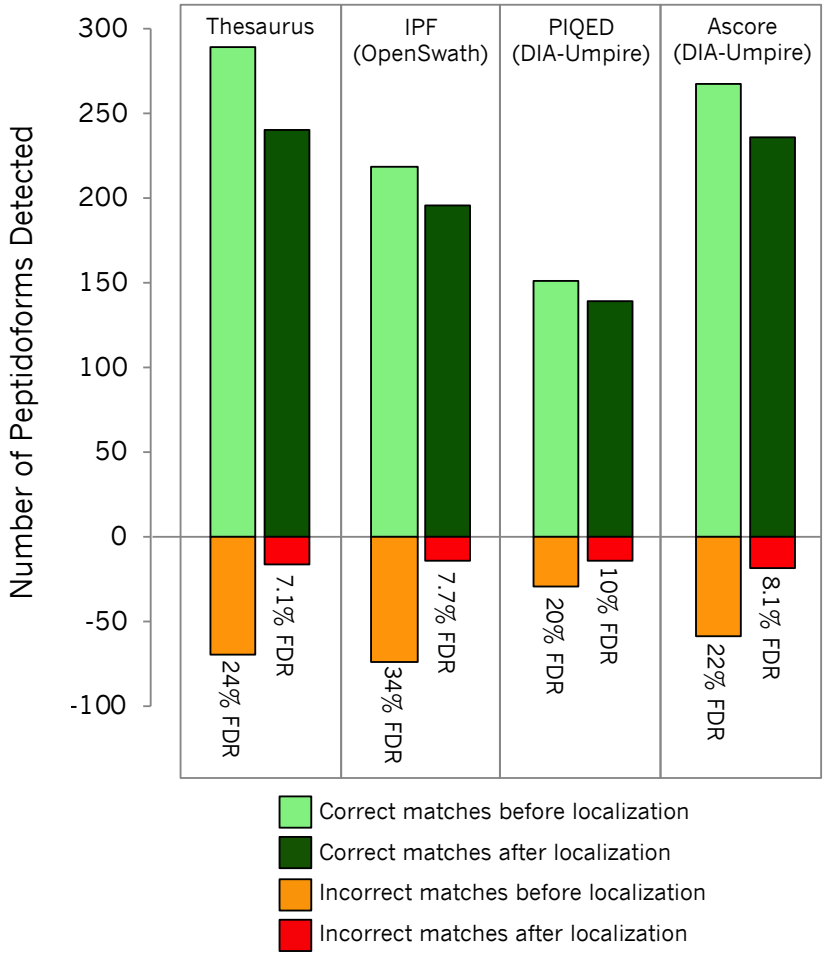


Figure 4-3: Validation using a known phosphopeptide standard.

The number of phosphopeptides detected from the Rosenberger et al SWATH-MS DIA synthetic phosphopeptide reference set(88) at an estimated 5% FDR. We performed replicate analyses of this data set using Thesaurus, OpenSwath/IPF (Inference of PeptidoForms), and two methods based on DIA-Umpire: PIQED (using Comet/PTMProphet), and Comet/Ascore. Library matches were marked as correct if they corresponded to expected synthetic positional isomers (579 of 1262 total). Thesaurus is able detect an additional 11 positional rearrangements. For the purposes of this comparison the detection of these manually curated isomers do not count as either a correct or an incorrect match to ensure a fair comparison. Manual curation of positional rearrangements required that the rearranged positional isomer was a) detected at the same retention time as the expected isomer, and b) contained at least

two site-specific fragment ions that fit the shape of the expected isomer without any interference. Comet is able to make detections to peptides not present in the library. Again, to ensure a fair comparison, these detections also do not count as either correct or incorrect.

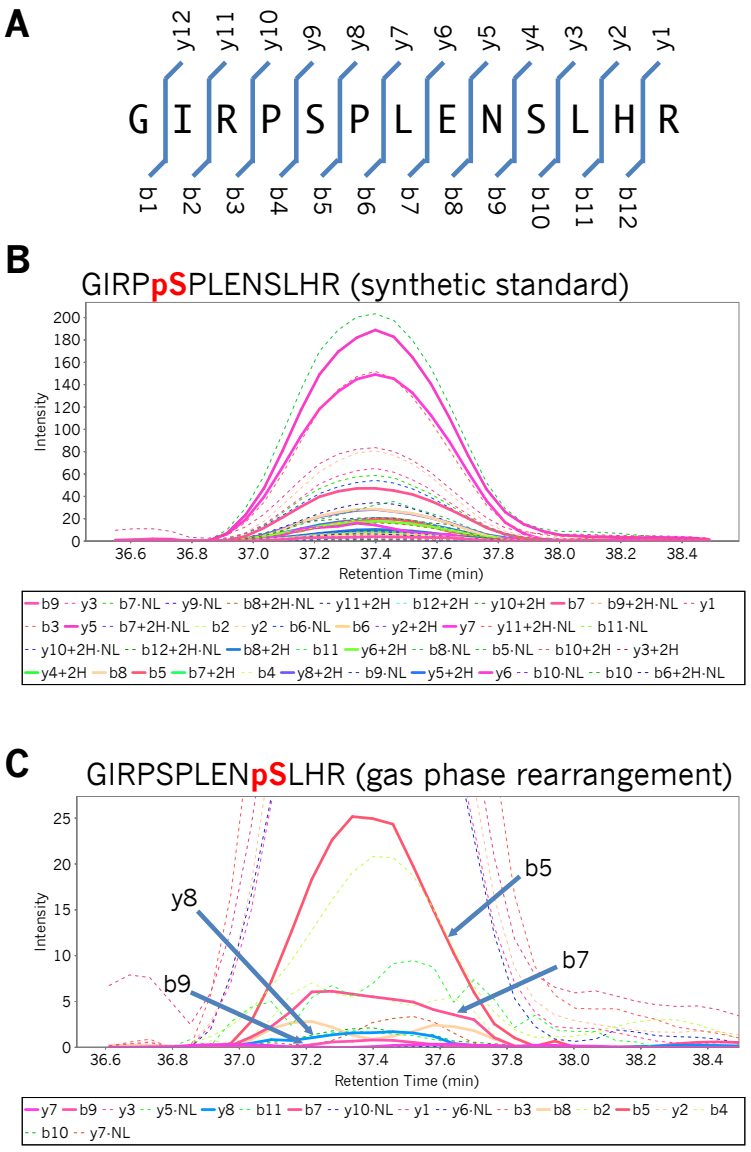


Figure 4-4: Phosphopeptide gas-phase rearrangement of the peptide GIRPpSPLENSHR.

(A) The synthetic phosphopeptide GIRPSPLENSHR produces 10 site localizing ions: b5-9 and y4-8. (B) All of the localizing ions are observed in the Rosenberger et al SWATH-MS DIA synthetic phosphopeptide reference set(88). (C) Four site localizing ions for the alternate isomer, GIRPSPLENpSHR, are also observable at the same elution time. Unlike localization algorithms that compete positional isomers against each other, Thesaurus is able to assign p-values to each variant independently.

To demonstrate the performance of our algorithm, we applied Thesaurus to phosphopeptides derived from serum-stimulated HeLa cells. Previously we reported a human phosphopeptide library based on nearly a thousand DDA experiments(38). In this work, we used a subset of this library that contains 82,029 phosphopeptides where 44% of phosphopeptides are represented with multiple isomers (Figure 4-5).

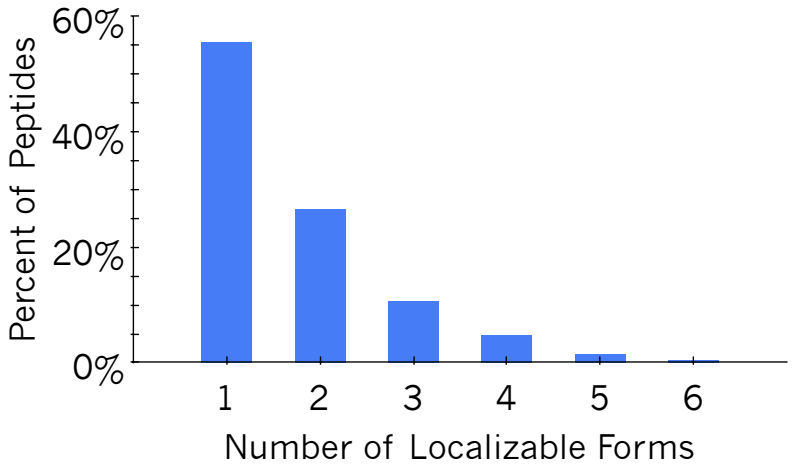


Figure 4-5: The number of confidently observed isobaric phosphopeptide forms for singly phosphorylated peptide species.

Previously we reported on the creation of a retention time database for phosphopeptides based on over one thousand DDA experiments from four labs(38). In this work we developed a subset of this spectrum library containing 82,029 distinct positional isomers detected at a 1% FDR level that were acquired in house on a Q-Exactive mass spectrometer. For this figure we required that each peptide was observed at least 50 times in this subset library to avoid undersampling issues, and each isomer was site localized to Ascore>13 (p-value<0.05).

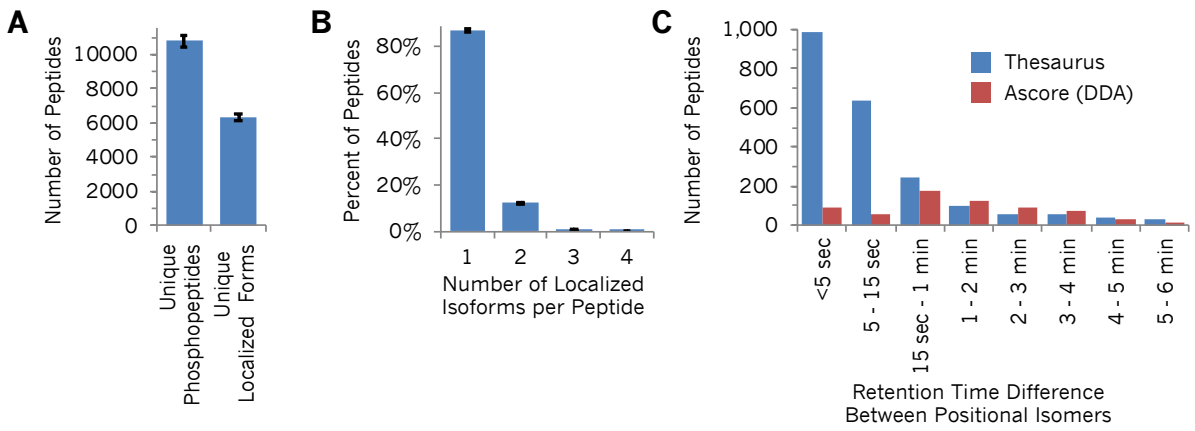


Figure 4-6: Statistics on phosphopeptide isomers

(A) The average number of unique phosphopeptide sequences and fully localized isomers (p-value<0.01) using Thesaurus across four HeLa technical DIA replicates. Error bars indicate 95% confidence intervals. (B) The percent of peptides that contained multiple serines, threonines, or tyrosines that could be fully localized to multiple isomers. (C) The percentage of fully localized, co-eluting phosphopeptides with retention time differences between positional isomers is highest below 60 seconds.

Thesaurus was able to detect an average of 10,780 unlocalized phosphopeptides across four technical replicate DIA experiments, corresponding to an average of 6,288 confidently localized positional isomer (Figure 4-6a). In this experiment we found that approximately 13% of the phosphopeptides that contained multiple serines, threonines, or tyrosines exist as multiple positional isomers (Figure 4-6b). Thesaurus was significantly better than DDA/Comet/Ascore at detecting isomer pairs when the difference in retention time was less than 60 seconds (Figure 4-6c). We found that while Thesaurus performed comparably to DDA/Comet/Ascore (Figure 4-7a), Thesaurus found 4x more phosphopeptides with multiple positional isomers (Figure 4-7b) and 30x more than DIA-Umpire/Comet/Ascore.

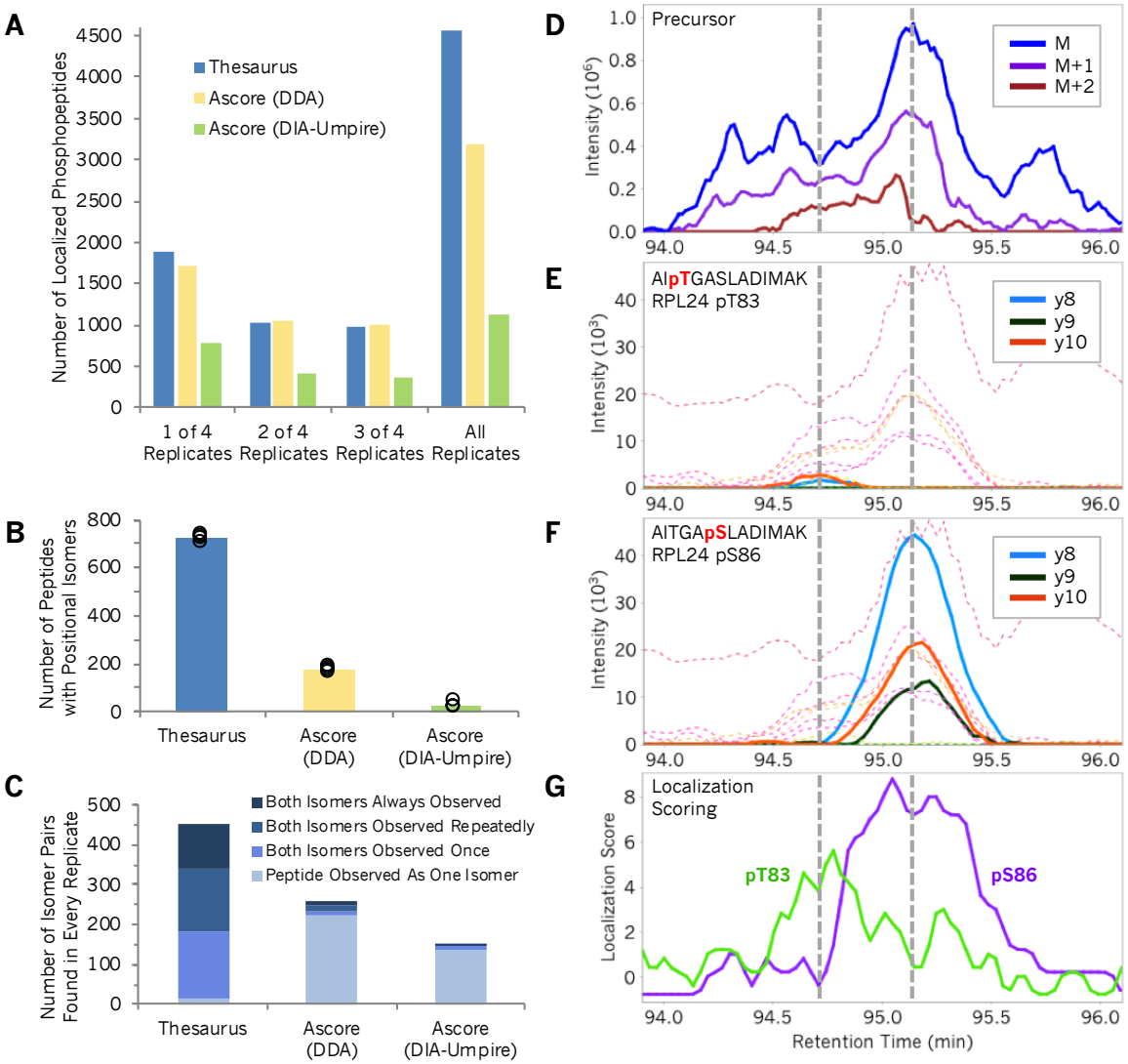


Figure 4-7: An approach for detecting phosphopeptides with Thesaurus. (A) The number of localized phosphopeptides across four technical replicates that were detected from DIA data with Thesaurus and DIA-Umpire/Ascore, or DDA data with Comet/Ascore. (B) The average number of peptides with multiple positional isomers that were detected from DIA data with Thesaurus and DIA-Umpire/Ascore, or DDA data with Comet/Ascore where error bars indicate 95% confidence intervals. (C) The number of singly phosphorylated peptides with two acceptor residues that were detected in all four technical replicates where both isomers were always observed or where an isomer was missing in least one replicate. To be included in this chart both isomers of the

phosphopeptide must have been observed in the same replicate by at least one analysis approach. (D) Precursor extracted ion chromatogram for the singly phosphorylated peptide AITGASLADIMAK from RPL24, which can be phosphorylated at both T83 and S86. Dashed grey lines indicate the retention time centers for the individual isomers. Site-specific y8, y9, and y10 ions (solid) and other y-ions (dashed) for (E) pT83 and (F) pS86. (G) Localization scores for pT83 (green) and pS86 (purple).

In particular, when considering phosphopeptides with only two acceptor sites, Thesaurus was markedly better at consistently detecting both positional isomer pairs across all four replicates (Figure 4-7c). For example, in every replicate Thesaurus consistently detected two isomers of the peptide AITGASLADIMAK from the 60S ribosomal protein RPL24 with single phosphorylation at either T83 or S86. These isomers co-elute within 25 seconds of each other, and while the precursor signal represents a mixture of both isomers (Figure 4-7d), Thesaurus confidently assigned them using site-specific ions (Figure 4-7e and 4-7f) to calculate localization scores (Figure 4-7g). While DDA reliably triggered MS/MS on the less intense early eluting isomer (pS86), in 3 of the 4 replicates MS/MS spectra corresponding to the more intense late eluting isomer (pT83) were never acquired due to dynamic exclusion. In these cases precursor quantification was unreliable because the signal from both isomers was assigned exclusively to the low abundance isomer because it eluted first. This result suggests that run-to-run variability between phosphoproteome experiments might be due in part to a fundamental aspect of the data acquisition method. Using Comet we confirmed that the site-specific fragment ions observed in DIA are the same as those observed in the single DDA replicate that collected an MS/MS for this phosphopeptide (Figure 4-8). In contrast, DIA-Umpire correctly generated a pseudo-MS2

for the higher intensity pT83 isomer in every DIA replicate but, due to interference, Comet scored the correct phosphopeptide only one of four times. In all four DIA replicates DIA-Umpire never generated a pseudo-MS2 for the lower intensity pS86 isomer.

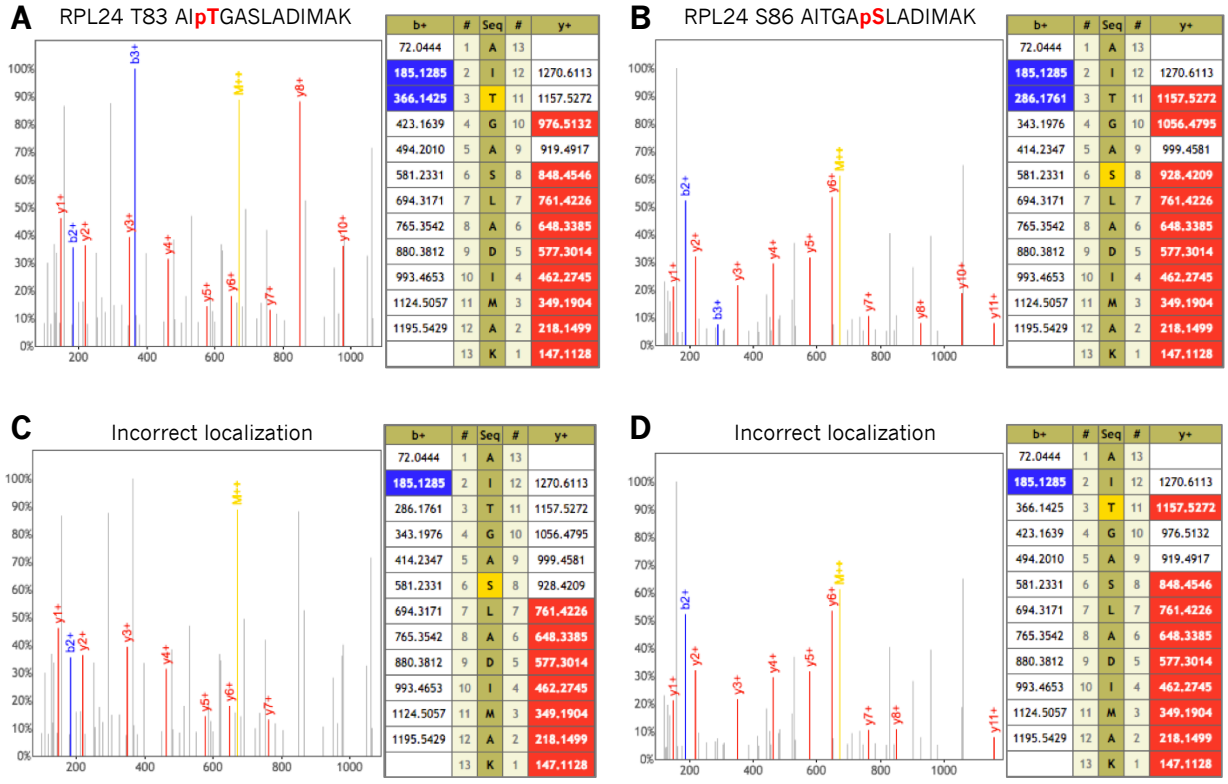


Figure 4-8: DDA spectra showing different localizations of singly phosphorylated AITGASLADIMAK.

We used DDA to confirm of the early and late eluting singly phosphorylated AITGASLADIMAK peptide from RPL24. Ion detections from (A) early eluting AIpTGASLADIMAK (Ascore=46.2) and (B) late eluting AITGApSLADIMAK (Ascore=30.7) are contrasted with the incorrect localization for (C) early eluting AITGApSLADIMAK and (D) late eluting AIpTGASLADIMAK.

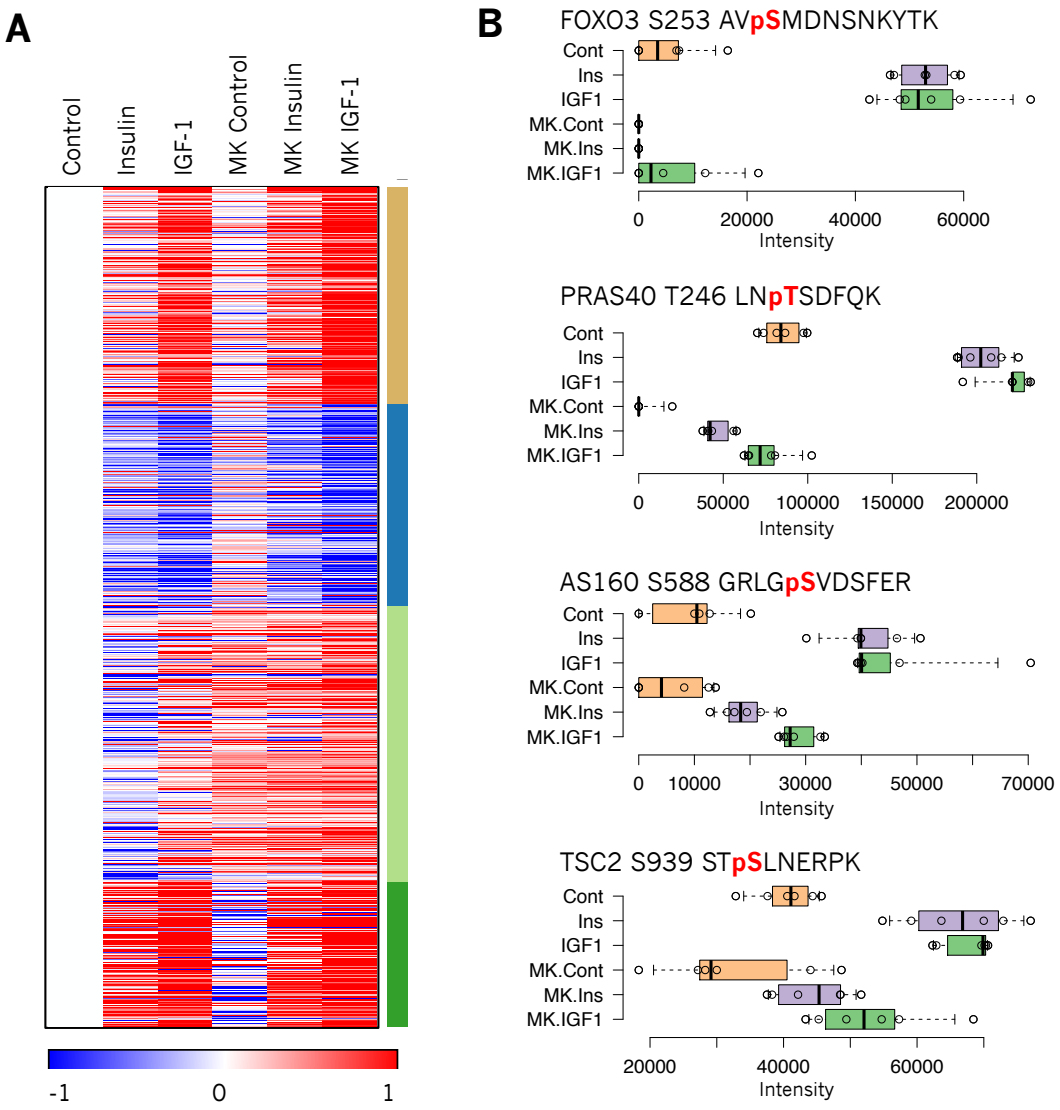


Figure 4-9: Global phosphoproteomic expression of insulin/iGF-1 stimulated MCF-7

(A) Heat map of 2273 significantly changing localized phosphopeptides at an ANOVA FDR <0.05. Red indicates >=2-fold up while blue indicates <=2-fold down. These peptides were K-means clustered into five groups. (B) Integrated intensities for FOXO3A pS253, PRAS40 pT246, AS160 S588, and TSC2 S939 all follow the canonical AKT RXRXX[pS/pT] motif and are known to be directly phosphorylated by AKT. Boxes indicate quartiles and medians, while whiskers indicate the estimated 5% and 95% ranges.

4.6 Quantifying phosphopeptides in the PI3K/AKT signaling network

Building on our new method to resolve positional isomers, we designed a DIA quantitative experiment to illuminate the PI3K/AKT signaling network in MCF-7 cells after stimulation with insulin and IGF-1.

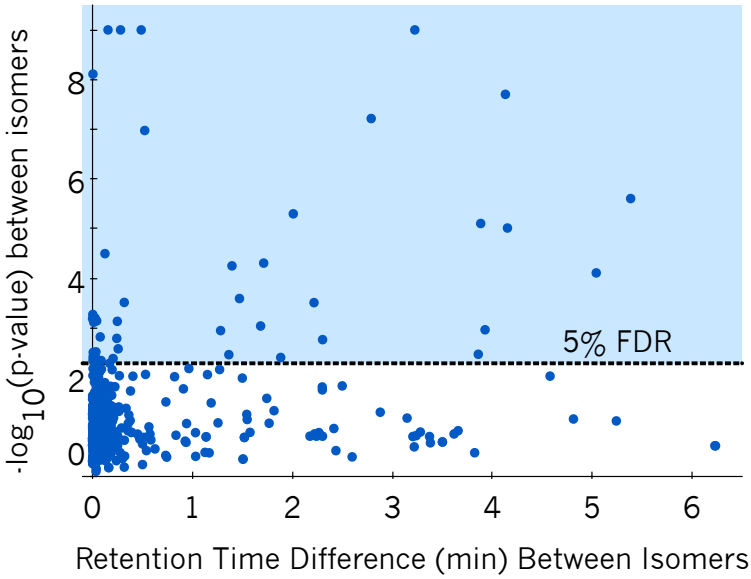


Figure 4-10: Scatterplot of retention time differences between phosphopeptide isomers

Here, both isomers are localized at <0.05% FDR and their delta retention time is compared to the p-value that the ratios of the isomer pairs after stimulation were pairwise consistent. As with HeLa, the majority of peptide isomers in MCF-7 elute within 60 seconds of each other. Most isomer pairs either do not significantly change after stimulus or do not differ from each other in relative expression profiles (white area). These sites are representative of either redundant function, background phosphorylation, or even indicate gas-phase rearrangement if the retention time differences are small enough. However, several positional isomers significantly change in different directions (light blue area), and these suggest differential function.

We found that 2,273 of the 7,434 localized and consistently measured phosphopeptides changed significantly at an FDR-corrected p-value <0.05, where the predominant group showed increased abundance in insulin/IGF-1 regardless of MK-2206 treatment (Figure 4-9a). As expected, some known AKT substrate sites showed significant response to MK-2206, such as FOXO3A S253 and PRAS40 T246 (Figure 4-9b).

In this experiment several phosphopeptides showed differential expression between individual isomers (Figure 4-10). For example, Figure 4-11a diagrams how the peptide KGSGDYMPMSPK from the insulin receptor scaffold protein IRS1 contains three residues that are putatively phosphorylated by three different kinases: Y632 by INSR (upstream of AKT), S636 by S6K1 (downstream of AKT), and S629 by PKA. Our library had no spectral representation of KGSGDpYMPMSPK with which to make a spectrum library detection of phosphorylated Y632. Despite this, Thesaurus was able to confidently detect, localize, and quantify all three singly phosphorylated isomers (Figure 4-11b and 4-11c). We confirmed these detections with scheduled PRM runs on the same samples (Figure 4-12) where the Thesaurus localization score can easily associate the retention times with each positional isomer. As expected from the model, after both insulin and IGF-1 stimulation phosphorylation of Y632 on INSR increased by >10-fold (Figure 4-11d). Similarly, S636 phosphorylation is also increased, but that effect was lower and significantly diminished when treated with the pan-AKT inhibitor MK-2206. As PKA is considered outside the AKT network, the phosphorylation state of S629 should stay unchanged and our measurements confirmed that.

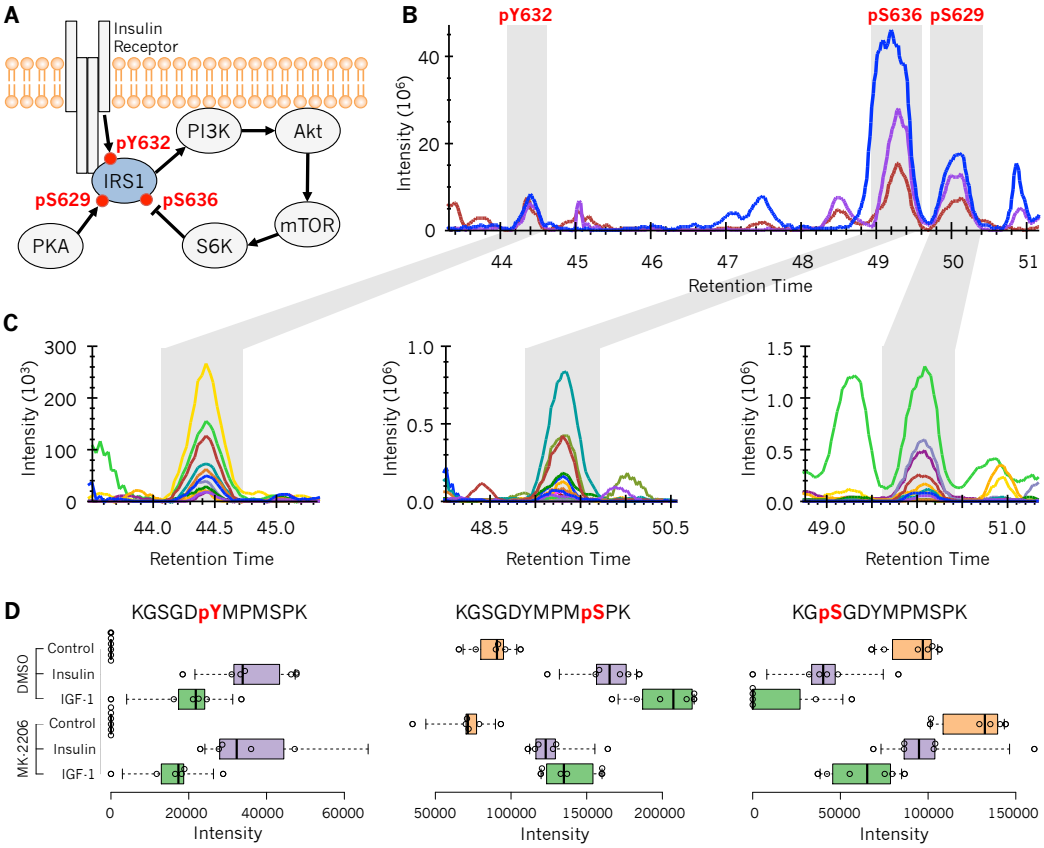


Figure 4-11: Detection and quantification of IRS1 phosphorylation.

(A) Diagram of IRS1 phosphorylation at sites S629, Y632, and S636 with respect to insulin/IGF-1 stimulation and treatment with the AKT inhibitor MK-2206. (B) Precursor traces for three singly phosphorylated positional isomers of peptide KGSGDYMPMSPK from IRS1 in insulin-stimulated MCF-7 cells, and (C) corresponding fragment ions indicating phosphorylation at Y632 by INSR, S636 by S6K1, and S629 by PKA. Thesaurus was able to detect, localize, and quantify pY632 despite that the pY632 positional isomer was absent from the searched spectrum library because it could use the pS629 and pS636 isomers as anchors. (D) Box plots and values indicating summed fragment ion intensities for the three phosphosites on IRS1 across six replicates after stimulation with insulin, IGF-1, or unstimulated (control); with and without MK-2206. Boxes indicate quartiles and medians, while whiskers indicate the estimated 5% and 95% ranges.

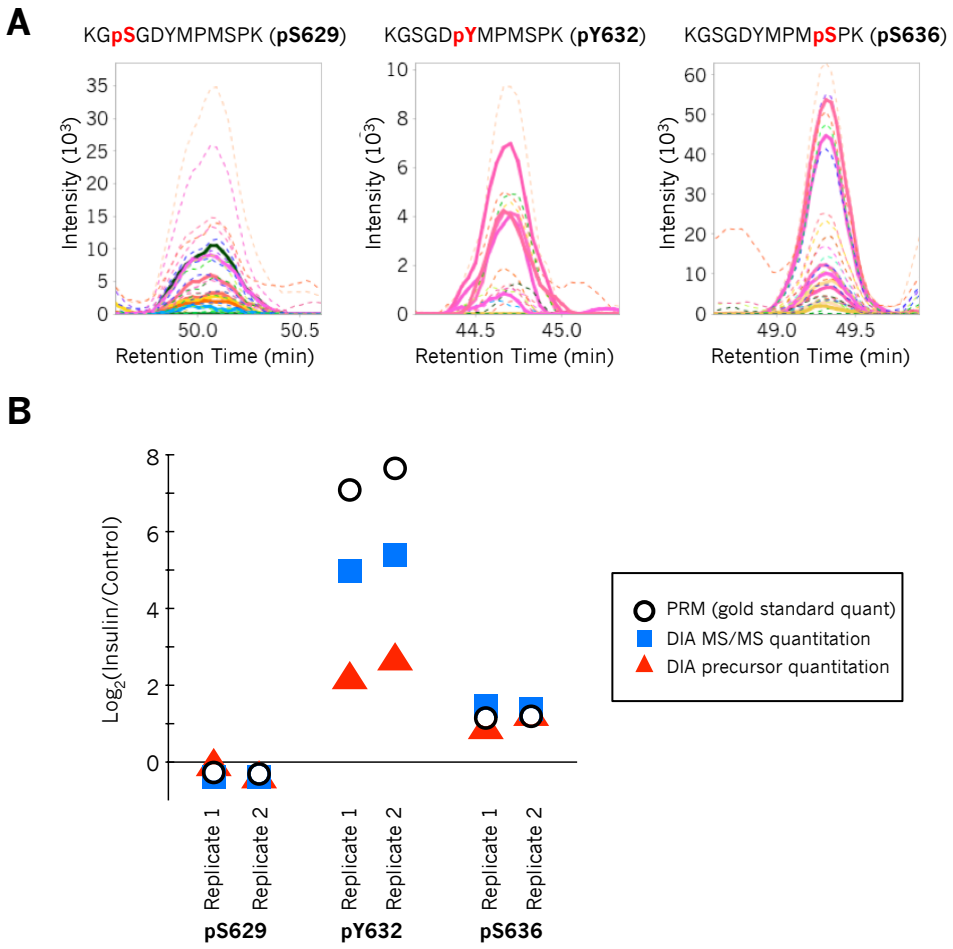


Figure 4-12: Validation of MS and MS/MS quantitation.

Validation of MS and MS/MS quantitation. (A) PRM localization scores and example fragment ion chromatograms of one replicate confirm the detection of three positional isomers of phosphorylated KGSGDYMPMSPK from IRS1 using +/- 0.7 m/z precursor isolation (as compared to +/- 10 m/z precursor isolation in the DIA experiments). (B) Phosphopeptide intensity ratios between insulin and control samples derived with targeted PRM (gold standard MS/MS measurements), and DIA with both MS and MS/MS-based quantitation. DIA MS and MS/MS quantitation for pS629 and pS636 are relatively accurate. While DIA-based quantitation for pY632 is suppressed, MS/MS quantitation is subject to less interference than MS quantitation and shows somewhat less ratio suppression.

All three phosphopeptides contained several shared fragment ions, and without statistical site localization it would have been extremely difficult and time consuming to manually determine which positional isomers were present and when they eluted. In addition to IRS1, we found numerous examples where only one positional isomers responded to insulin/IGF-1 stimulation or both responded with opposite directionality (Figure 4-13). Finally, some phosphopeptide isomers were completely indistinguishable by retention time, yet could be confidently localized and quantified using site-specific ions. For example, MARK3 positional isomers at S469 and S476 co-eluted under our chromatographic conditions. Using Thesaurus, we were able to detect that the S469 isomer responded to insulin/IGF-1 and AKT inhibition while the S476 isomer remained constant (Figure 4-14).

Positional isomers represent an important concept for understanding signaling biology where neighboring phosphosites can have profound biological impact. Our tool Thesaurus provides a new avenue to study positional isomers even if their precursor signals cannot be resolved. Now with a search engine specifically designed to analyze neighboring sites of phosphorylation it is possible to determine whether they have distinct functional implications, are redundant mechanisms for regulation, or are simply representative of a background phosphorylation state. The analysis of several other types of PTMs will also benefit from this analysis approach, so we have extended Thesaurus to support other modifications such as methylation, acetylation, and O-HexNAcylation and developed a robust, multi-threaded tool with a stand-alone graphical interface to enable wide adoption. Our results indicate that PRM and DIA strategies will be crucial in assessing the complex regulatory nature of the human phosphoproteome.

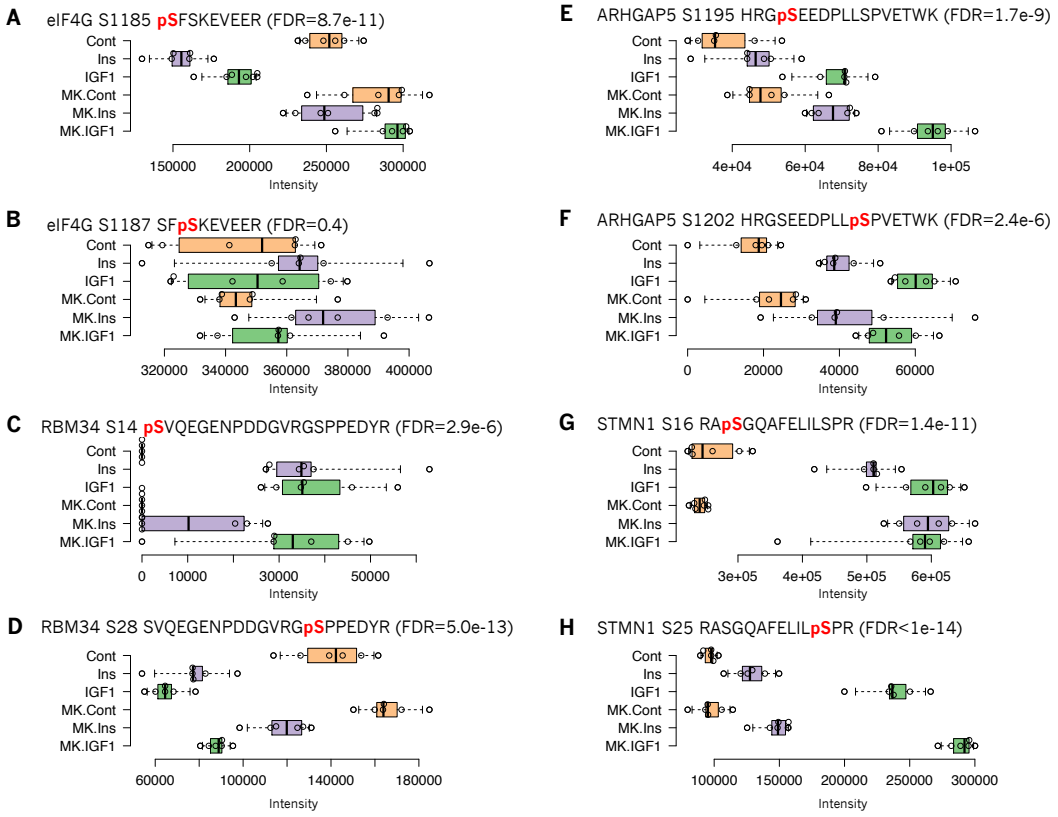


Figure 4-13: Dynamic response of site-specific phosphorylation in positional isomers.

Phosphopeptide signals can differ significantly between two positional isomers of the same peptide upon pathway stimulation or kinase inhibition, indicating differential regulation. Phosphorylation of eIF4G (A) S1185 by PKCα is thought to induce binding to MNK1, but the function of the more abundant (B) S1187 is unknown. The functional relevance of both phosphosites on RBM34 (C and D) are unknown, but behave in opposite directions in response to insulin/IGF-1 signaling. In the Rho GTPase activating protein ARHGAP5 (E and F), phosphorylation of both S1195 and S1202 increase with insulin/IGF-1, but only S1195 changes as a result of AKT inhibition. (G) S16 phosphorylation of STMN1 responds equally to insulin and IGF-1, while (H) S25 responds much more significantly to IGF-1. Boxes indicate quartiles and medians, while whiskers indicate the estimated 5% and 95% ranges. FDR is calculated as the Benjamini-Hochberg corrected ANOVA test between conditions.

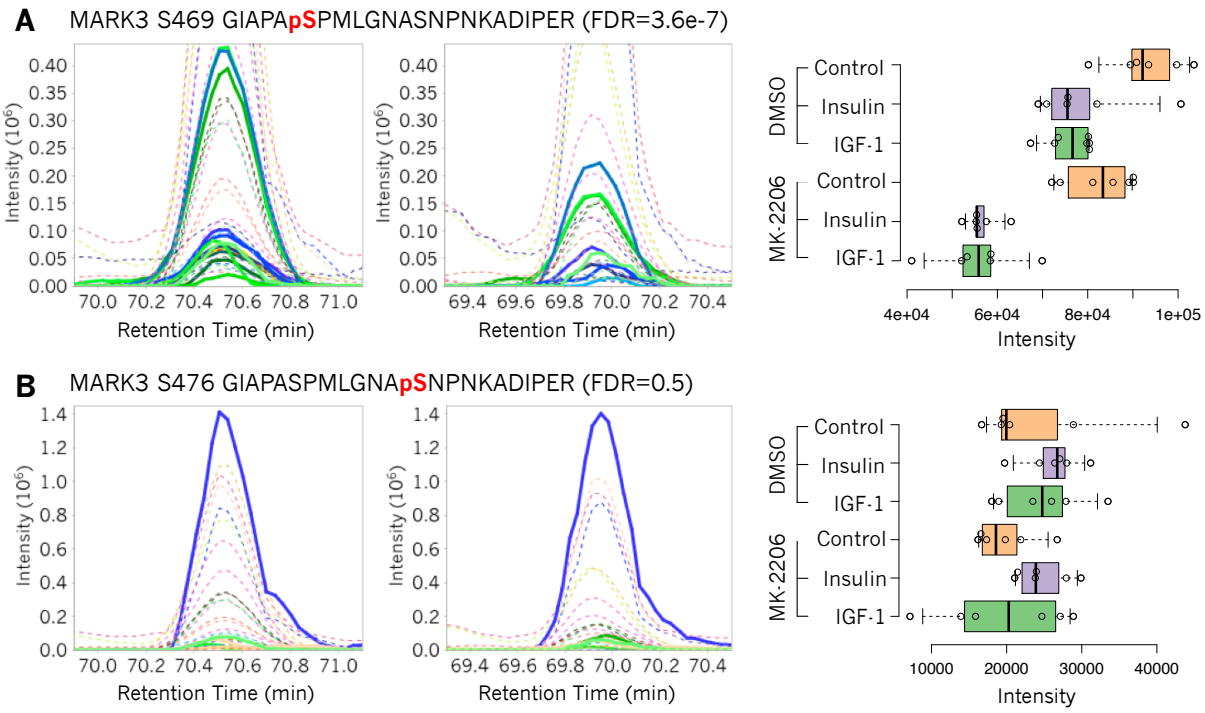


Figure 4-14: Several phosphopeptides are completely indistinguishable by retention time.

Thesaurus can detect and independently quantify both positional isomers using site-specific ions. Here we show two positional isomers of MARK3 with phosphorylation on S469 (A) or S476 (B). The extracted fragment ion chromatograms on the left are representative of the control group, while the right chromatograms are representative data from the IGF-1/MK-2206 treated cells. Solid lines are site-specific ions, and dashed lines are shared ions. On the right side we show the integrated intensities for those positional isomers in the different experimental conditions. Phosphorylation of S469 additively drops due to insulin/IGF-1 and AKT inhibition (FDR=3.6e-7), while phosphorylation of S476 stays constant (FDR=0.54).

5 DETECTING GENETIC VARIATION WITH DIA

5.1 Detecting genetic variation in amyloid fibrils using DIA

Single nucleotide polymorphisms and other genomic sequence variants can have profound impact on susceptibility to disease. Even still, most shotgun proteomics workflows focus on detecting canonical protein sequences found in FASTA databases. While proteogenomics methods that combine customized exome sequencing with mass spectrometry are emerging for data dependent acquisition (DDA), data independent acquisition (DIA) approaches frequently rely on curated spectrum libraries that lack sequence variants. Moreover, because most variants result in small M/Z shifts, these peptides often isolate and fragment together. Variant peptides produce many of the same fragment ions as canonical peptides, and confidently distinguishing variant forms is challenging. We present preliminary work on a new approach to detect and statistically validate peptide variants in DIA data even when the precursors co-fragment.

We developed a computational algorithm called XCorDIA to search DIA data directly using protein sequences. XCorDIA identifies peptides using a peptide-centric version of the XCorr(18) scoring system from Sequest. We validated XCorDIA against Walnut, another peptide-centric search engine for DIA based on the PECAN peptide detection algorithm(22) using HeLa samples. We found that not only was XCorDIA significantly faster, it outperformed Walnut at detecting peptides by greater than 20% (Figure 5-1).

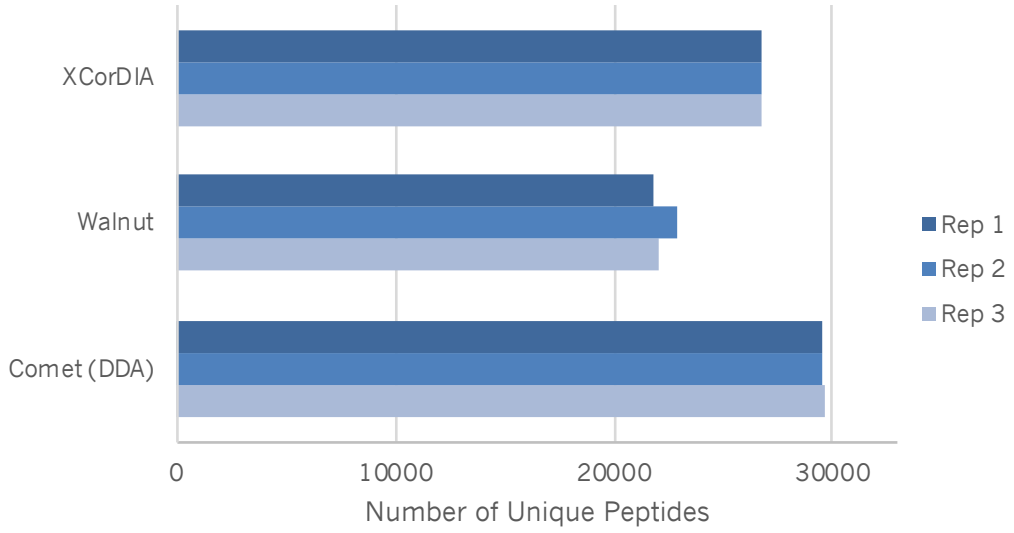


Figure 5-1: Numbers of peptides detected by XCorDIA and PECAN.
We analyzed three replicate injections of a HeLa proteome using DIA with XCorDIA and Walnut (with the Pecan algorithm), and DDA with Comet. On average XCorDIA detects over 20% more peptides than Walnut.

In addition to standard FASTA databases, XCorDIA can read PEFF-formatted databases(94) that contain known sequence variations. Peptides that share sequence homology (including both genomic variants and paralogs) and fall in the same precursor isolation window are grouped and searched simultaneously. Similar to the localization algorithm described in Chapter 4, we determine variant-specific ions and calculate the probability we would have seen those ions at random from a null distribution. Because peptide interference is common in DIA datasets, we calculate a new null distribution for every window as the percentage of MS/MS spectra in that window that contain each fragment ion.

Amyloidosis is a disease caused by the accumulation of misfolded proteins in tissue. Several sequence variants in commonly deposited proteins are known to increase

the risk of misfolding(95, 96), and we sought to discover other variants that were observable in amyloid fibrils. We sectioned formalin-fixed paraffin-embedded (FFPE) renal biopsy specimens from four controls and six amyloidosis cases with laser microdissection in three replicate tissue sections. We extracted and resolubilized proteins from each tissue section, followed by reduction, alkylation, and digestion for 18 hours. We then separated the resulting peptides using reversed-phase HPLC using a 60-minute gradient. We performed DIA experiments on a Q-Exactive HF using 20x overlapped 20 m/z windows, covering 500-910 m/z.

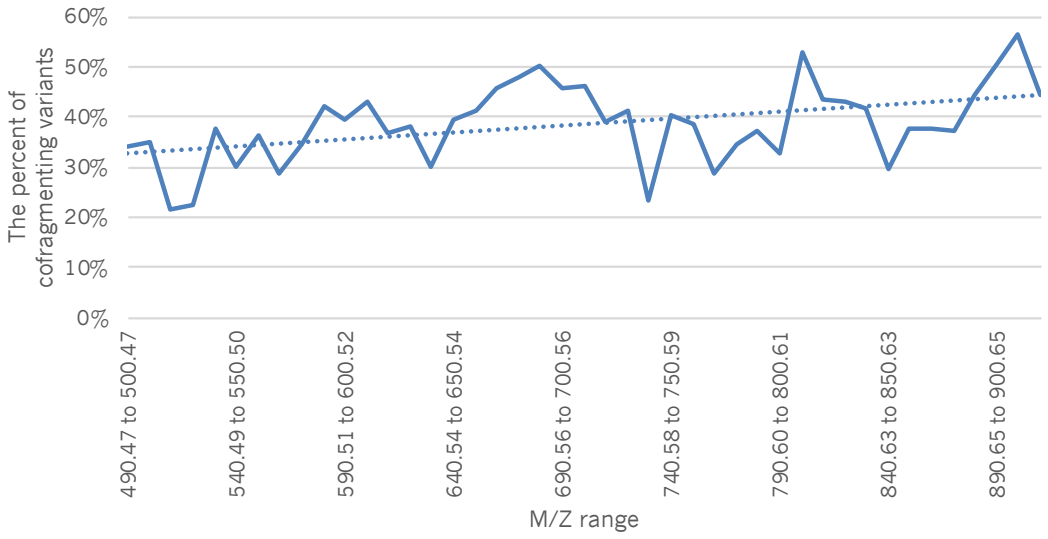


Figure 5-2: The percent of variants that co-fragment increases with increasing M/Z.

On average 39% of variants fall in the same 10 M/Z precursor isolation window.

We demultiplexed the overlapping windows using ProteoWizard, resulting in 40x windows with effectively 10 m/z precursor isolation. We searched the resulting files using XCorDIA, a new peptide-centric search engine specifically designed for detecting variant peptides.

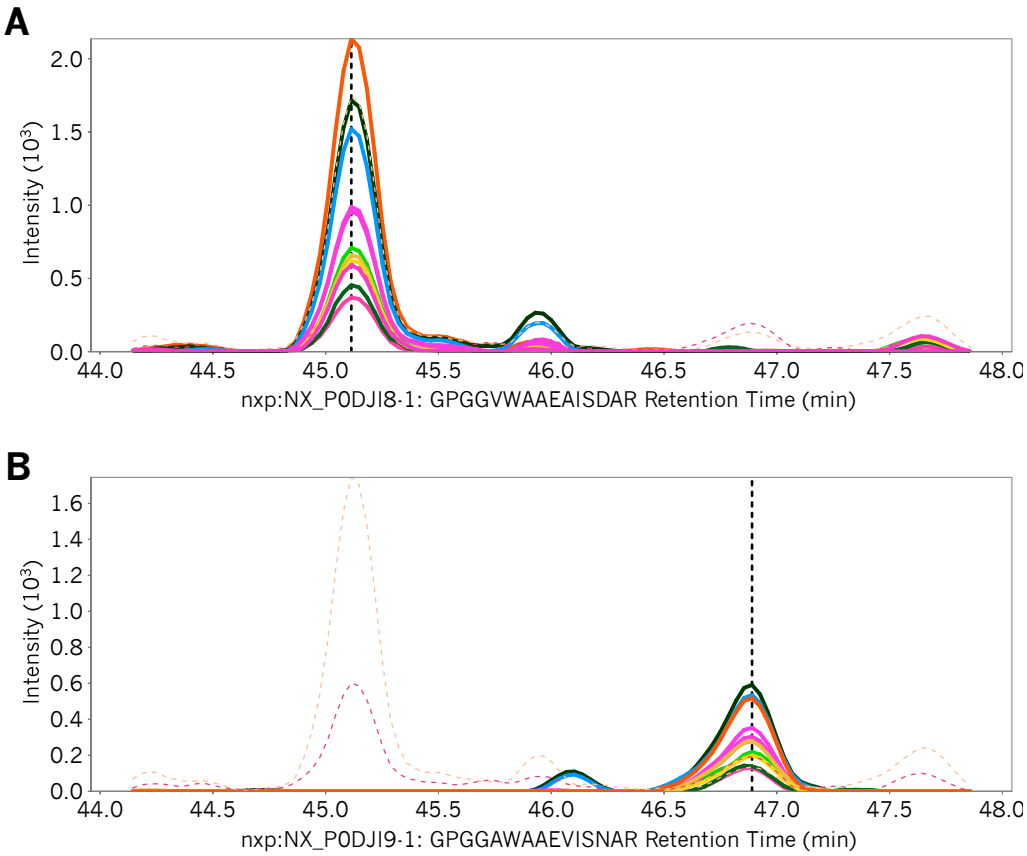


Figure 5-3: Detection of two coeluting peptides corresponding to serum amyloid paralogs.

Solid lines are variant-specific ions, and dashed lines are shared ions.

We searched triplicate tissue sections from four controls and six amyloidosis cases using an amyloidosis-specific PEFF database with genomic variants from ClinVar. We found that 39% of variant peptide groups from ClinVar fell into the same 10 m/z precursor isolation windows (Figure 5-2). Of 116 amyloidosis-specific peptides, we observed 3 variants of unknown pathogenicity in serum amyloid A-2, serum amyloid P, and transthyretin at an experiment-wide peptide FDR<0.05. In addition, we identified several peptides assigned to paralogous genes that fall within the same precursor windows. For example, the peptides GPGGVWAAEAISDAR (Serum amyloid A-1) and

GPGGAWAAEVISNAR (Serum amyloid A-2) only differ by 0.5 M/Z and both fall in the 720.57 to 730.58 M/Z precursor isolation window (Figure 5-3). While these findings are still exploratory in nature, they outline the strength of DIA for detecting allelic variation. In future work we aim to develop extensions of this method that pair the quantitative nature of DIA with allele specificity to develop new pQTL approaches(97, 98).

6 FUTURE DIRECTIONS

6.1 Concluding remarks

This dissertation explores a variety of new approaches to data independent acquisition. Based on its reproducibility, we believe that DIA will shine in large quantitative studies that require multiplexing beyond isobaric tag methods (e.g. iTRAQ, TMT) or ongoing long-term studies that require comparing results from experiments stretched over long time periods. While time will tell where the balance between DIA and DDA lies, each strategy has distinct strengths and they remain complementary to each other. In addition to use as a library building tool, DDA remains the method of choice for “one-off” experiments or pilot experiments without major quantitative requirements. Regardless, we hope the methods demonstrated here will enable new types of large scale quantitative studies, and in particular those involving phosphorylation and genetic variation.

APPENDIX A. SUPPLEMENTARY METHODS FOR CHAPTER 1

HeLa cell culture and sample preparation. HeLa S3 cervical cancer cells (ATCC) were cultured at 37°C and 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with L-glutamine, 10% fetal bovine serum (FBS), and 0.5% strep/penicillin. Cells were either serum starved for 2 hours or 16 hours prior to 3x washing with refrigerated phosphate-buffered saline and immediately flash freezing with liquid nitrogen. Cells were lysed in a buffer of 9 M urea, 50 mM Tris (pH 8), 75 mM NaCl, with a cocktail of protease inhibitors (Roche Complete-mini EDTA-free). Cells were sonicated for 2x 30 seconds after harvest, followed by 20 minutes of incubation on ice and 10 minutes of centrifugation at 21,000 x *g* and 4°C. The protein content of the supernatant was estimated using BCA. Proteins were reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature, and quenched with an additional 5 mM dithiothreitol for 15 minutes at room temperature. Proteins were then diluted to 1.8 M urea prior to digestion with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 12 hours at 37°C. The digestion was quenched by adding 10% trifluoroacetic acid to achieve approximately pH 2. Digested peptides were desalted with 100 mg tC18 SepPak cartridges (Waters) using vendor-provided protocols and dried with vacuum centrifugation.

Liquid chromatography mass spectrometry. Peptides were separated with a Waters NanoAcquity UPLC and electrosprayed into a Thermo Q-Exactive HF tandem mass spectrometer. Pulled tip columns were created from 75 um inner diameter fused

silica capillary in-house using a laser pulling device and packed with 3 µm ReproSil-Pur C18 beads (Dr. Maisch) to 270 mm. Trap columns were created from 150 um inner diameter fused silica capillary Kasil fritted with on one end and packed with the same C18 beads to 25 mm. Solvent A was 0.1% formic acid in water, while solvent B was 0.1% formic acid in 98% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted using a 90-minute gradient from 5 to 35% B. Data were acquired using either data-dependent acquisition (DDA) or data-independent acquisition (DIA) in an alternating mode to avoid bias.

DDA acquisition and processing. Precursor spectra (400-1600 m/z) were collected at 60,000 resolution to hit an AGC target of 3e6 with a maximum inject time of 100 ms. Fragment spectra were collected in a top-20 configuration at 15,000 resolution to hit an AGC target of 1e5 with a maximum inject time of 25 ms. The isolation width was set to 1.6 m/z with a normalized collision energy of 27. Only precursors charged between +2 and +4 that achieved a minimum AGC of 5e3 were acquired. Dynamic exclusion was set to “auto” and to exclude all isotopes in a cluster. Thermo RAW files were converted to mzXML format using ReAdW and searched against a Uniprot Human FASTA database (87613 entries) with Comet (version 2015.02v2), allowing for variable methionine oxidation, and n-terminal acetylation. Cysteines were assumed to be fully carbamidomethylated. Searches were performed using a 50 ppm precursor tolerance and a 0.02 Da fragment tolerance using fully tryptic specificity (KR|P) permitting up to two missed cleavages. Search results were filtered to either a 1% peptide-level or a 1%

protein-level FDR using Percolator (version 3.1). Quantification and protein grouping were performed using an in-house precursor XIC integration tool.

Table 3: Gas phase fractionated precursor isolation window center m/zs (Cycle T).

	Cycle Index	Narrow Window (400-500 M/Z)	Narrow Window (500-600 M/Z)	Narrow Window (600-700 M/Z)	Narrow Window (700-800 M/Z)	Narrow Window (800-900 M/Z)	Narrow Window (900-1000 M/Z)
Cycle T Precursor Isolation Window Centers	1	400.4319	500.4774	600.5229	700.5683	800.6138	900.6593
	2	404.4337	504.4792	604.5247	704.5701	804.6156	904.6611
	3	408.4355	508.4810	608.5265	708.5720	808.6174	908.6629
	4	412.4374	512.4828	612.5283	712.5738	812.6193	912.6647
	5	416.4392	516.4847	616.5301	716.5756	816.6211	916.6666
	6	420.4410	520.4865	620.5319	720.5774	820.6229	920.6684
	7	424.4428	524.4883	624.5338	724.5792	824.6247	924.6702
	8	428.4446	528.4901	628.5356	728.5811	828.6265	928.6720
	9	432.4465	532.4919	632.5374	732.5829	832.6284	932.6738
	10	436.4483	536.4937	636.5392	736.5847	836.6302	936.6756
	11	440.4501	540.4956	640.5410	740.5865	840.6320	940.6775
	12	444.4519	544.4974	644.5429	744.5883	844.6338	944.6793
	13	448.4537	548.4992	648.5447	748.5902	848.6356	948.6811
	14	452.4555	552.5010	652.5465	752.5920	852.6374	952.6829
	15	456.4574	556.5028	656.5483	756.5938	856.6393	956.6847
	16	460.4592	560.5047	660.5501	760.5956	860.6411	960.6866
	17	464.4610	564.5065	664.5520	764.5974	864.6429	964.6884
	18	468.4628	568.5083	668.5538	768.5992	868.6447	968.6902
	19	472.4646	572.5101	672.5556	772.6011	872.6465	972.6920
	20	476.4665	576.5119	676.5574	776.6029	876.6484	976.6938
	21	480.4683	580.5138	680.5592	780.6047	880.6502	980.6957
	22	484.4701	584.5156	684.5610	784.6065	884.6520	984.6975
	23	488.4719	588.5174	688.5629	788.6083	888.6538	988.6993
	24	492.4737	592.5192	692.5647	792.6102	892.6556	992.7011
	25	496.4756	596.5210	696.5665	796.6120	896.6575	996.7029
	26	500.4774	600.5229	700.5683	800.6138	900.6593	1000.7048

Table 4: Gas phase fractionated precursor isolation window center m/zs (Cycle T+1).

Cycle T+1 Overlap Shifted Precursor Isolation Window Centers	Cycle Index	Narrow Window (400-500 M/Z)	Narrow Window (500-600 M/Z)	Narrow Window (600-700 M/Z)	Narrow Window (700-800 M/Z)	Narrow Window (800-900 M/Z)	Narrow Window (900-1000 M/Z)
	27	398.4310	498.4765	598.5219	698.5674	798.6129	898.6584
	28	402.4328	502.4783	602.5238	702.5692	802.6147	902.6602
	29	406.4346	506.4801	606.5256	706.5711	806.6165	906.6620
	30	410.4364	510.4819	610.5274	710.5729	810.6183	910.6638
	31	414.4383	514.4837	614.5292	714.5747	814.6202	914.6656
	32	418.4401	518.4856	618.5310	718.5765	818.6220	918.6675
	33	422.4419	522.4874	622.5329	722.5783	822.6238	922.6693
	34	426.4437	526.4892	626.5347	726.5801	826.6256	926.6711
	35	430.4455	530.4910	630.5365	730.5820	830.6274	930.6729
	36	434.4474	534.4928	634.5383	734.5838	834.6293	934.6747
	37	438.4492	538.4947	638.5401	738.5856	838.6311	938.6766
	38	442.4510	542.4965	642.5419	742.5874	842.6329	942.6784
	39	446.4528	546.4983	646.5438	746.5892	846.6347	946.6802
	40	450.4546	550.5001	650.5456	750.5911	850.6365	950.6820
	41	454.4565	554.5019	654.5474	754.5929	854.6384	954.6838
	42	458.4583	558.5038	658.5492	758.5947	858.6402	958.6857
	43	462.4601	562.5056	662.5510	762.5965	862.6420	962.6875
	44	466.4619	566.5074	666.5529	766.5983	866.6438	966.6893
	45	470.4637	570.5092	670.5547	770.6002	870.6456	970.6911
	46	474.4656	574.5110	674.5565	774.6020	874.6475	974.6929
	47	478.4674	578.5128	678.5583	778.6038	878.6493	978.6947
	48	482.4692	582.5147	682.5601	782.6056	882.6511	982.6966
	49	486.4710	586.5165	686.5620	786.6074	886.6529	986.6984
	50	490.4728	590.5183	690.5638	790.6093	890.6547	990.7002
	51	494.4746	594.5201	694.5656	794.6111	894.6565	994.7020
	52	498.4765	598.5219	698.5674	798.6129	898.6584	998.7038

GPF DIA acquisition. Six acquisitions with 4 m/z DIA spectra (4 m/z precursor isolation windows at 30,000 resolution, AGC target 1e6, maximum inject time 55 ms) using an overlapping window pattern from narrow mass ranges using window placements

optimized by Skyline (i.e. 396.43 to 502.48 m/z, 496.48 to 602.52 m/z, 596.52 to 702.57 m/z, 696.57 to 802.61 m/z, 796.61 to 902.66 m/z, and 896.66 to 1002.70 m/z). See Tables 3 and 4 for the actual windowing scheme. A narrow MS spectrum was also acquired matching the range (i.e. 390-510 m/z, 490-610 m/z, 590-710 m/z, 690-810 m/z, 790-910 m/z, and 890-1010 m/z). These scans used an AGC target of 3e6 and a maximum inject time of 100 ms were interspersed every 18 MS/MS spectra.

]

DIA acquisition. Triplicate DIA experiments were acquired using three different MS/MS windowing schemes: 25x 24 m/z, 25x variable window m/z, and 51x overlapped 24 m/z. Variable windows were generated using the SWATH Acquisition Variable Window Calculator Excel spreadsheet from SCIEX (https://sciex.com/support/knowledge-base-articles/how-to-use-the-swath-variable-calculator-excel-sheet), while the other windowing schemes were optimized by Skyline. See Tables 5 for the actual windowing schemes. MS/MS scans were acquired at 30,000 resolution with an AGC target of 1e6 and a maximum inject time of 55 ms.

Table 5: Precursor isolation window center m/zs

Centers for normal windows (a), variable width windows (b) and deltas (c), and overlapping windows (d, e).

Scan	Normal Windows (a)	Variable Width Windows (b)	Variable Window Deltas (c)	Overlap Windows Cycle 1 (d)	Overlap Windows Cycle 2 (e)
1	412.44	415.70	31.4	400.43	412.44
2	436.45	444.05	25.3	424.44	436.45
3	460.46	467.40	21.4	448.45	460.46
4	484.47	487.75	19.3	472.46	484.47
5	508.48	506.75	18.7	496.48	508.48
6	532.49	524.90	17.6	520.49	532.49
7	556.50	542.20	17.0	544.50	556.50
8	580.51	558.65	15.9	568.51	580.51
9	604.52	574.60	16.0	592.52	604.52
10	628.54	590.30	15.4	616.53	628.54
11	652.55	605.70	15.4	640.54	652.55
12	676.56	621.65	16.5	664.55	676.56
13	700.57	638.15	16.5	688.56	700.57
14	724.58	655.45	18.1	712.57	724.58
15	748.59	673.85	18.7	736.58	748.59
16	772.60	693.10	19.8	760.60	772.60
17	796.61	713.20	20.4	784.61	796.61
18	820.62	734.10	21.4	808.62	820.62
19	844.63	756.65	23.7	832.63	844.63
20	868.64	780.85	24.7	856.64	868.64
21	892.66	807.25	28.1	880.65	892.66
22	916.67	836.70	30.8	904.66	916.67
23	940.68	870.25	36.3	928.67	940.68
24	964.69	911.20	45.6	952.68	964.69
25	988.70	967.00	66.0	976.69	988.70
26				1000.70	

DIA and GPF DIA processing. DIA experiments were analyzed using EncyclopeDIA (manuscript in preparation) using a HeLa-specific Bibliospec(99) HCD spectrum library was created from unpublished Thermo Q-Exactive DDA data using Skyline (version 3.1.0.7382). EncyclopeDIA was configured with default settings (10 ppm

precursor, fragment, and library tolerances, considering both B and Y ions, and trypsin digestion was assumed) and set to use Percolator version 3.1.

PRM acquisition and processing. Precursor spectra (395-1005 m/z) were collected at 30,000 resolution (AGC target 3e6, maximum inject time 100 ms) followed by N targeted fragment spectra. The isolation width was set to 0.7 m/z. Scans were collected at 30,000 resolution with an AGC target of 1e6 and a maximum inject time of 55 ms. In total 605 target peptides were analyzed in 5x replicates for each sample. RAW files were analyzed using Skyline Daily and transitions were manually refined, requiring at least three fragment ions with low interference. Skyline was also used to extract MS1 signal for comparison.

APPENDIX B. SUPPLEMENTARY METHODS FOR CHAPTER 2

Training set stable isotope peptides. A total of 1679 stable isotope labeled (SIL) peptides (C-terminal K*= Lys (U-13C6;U-15N2) or C-terminal R*= Arg (U-13C6;U-15N4)) were obtained as a crude (SpikeTide L) mixture from JPT Peptide Technologies GmbH (Berlin, Germany). All peptides are tryptic digestion products of human proteins that have been observed in previous shotgun DDA runs of human samples. This peptide selection may introduce a small bias towards peptides that can be interpreted with DDA, although significant fractionation was required to initially assign many of the peptides. Peptides were acquired with all cysteines alkylated to carbamidomethyl cysteine. In general the training peptides are representative of normal peptides with one exception: the training data set does not contain peptides with a methionine. One aliquot of the peptide mixture (~ 0.1 nmol of each peptide) was resuspended in 100 µL of 80% 0.1M ammonium bicarbonate and 20% acetonitrile. The mixture was bath sonicated for 5 minutes, vortexed at 37°C for 5 minutes. 1 µL of the ~1 picomole /µL solution was diluted in 99 µL of 0.1% formic acid for a 10 fmol/µL solution which was spun down prior to transferring to a sample vial for LC-MS/MS analysis.

Training set LC-MS/MS analysis. A 1.5 µL (15 fmol runs) or 4.5 µL (45 fmol runs) aliquot of the SIL mixture was loaded onto a 2 cm x 150 µm Kasil-fritted trap packed with 4 µm Jupiter C12 90A material (Phenomenex; Torrance, CA). The sample was loaded and desalted using 5 µL of a 0.1% formic acid, 2% acetonitrile solution. The trap was brought on-line with the analytical column. The analytical column was a fused-silica capillary (75 µm inner diameter) with a tip pulled using a CO_2 laser-based micropipette

puller (P-2000; Sutter Instrument Company; Novato, CA). The analytical column was packed with 15 cm of 3 µm Reprosil-Pur C18-AQ beads (Dr. Maisch GmbH, Germany). The analytical column was coupled in-line to a Waters nanoAcquity UPLC pump and autosampler (Waters Corp, Milford, MA). Peptides were eluted off of the column at a flow rate of 300 nL/min using a 90 minute gradient of 2-35% acetonitrile in 0.1% formic acid, followed by 35-60% acetonitrile in 0.1% formic acid over 5 minutes. Peptides were ionized by electrospray (2kV spray voltage) and emitted into a Q-Exactive HF mass spectrometer (Thermo Scientific; Bremen, Germany). Data were acquired using either of two acquisition methods: data-dependent acquisition (DDA) or data-independent acquisition (DIA).

Training set DDA acquisition. The DDA method acquires an MS scan analyzing 485 – 925 *m/z* with resolution 120,000 (@ 200 *m/z*), automated gain control (AGC) target 3 x 10^6 charges, and maximum injection time 50 ms. Next, up to 20 MS/MS scans were triggered from the top 20 most intense precursors detected in the MS master scan. The MS/MS scans have resolution 15,000 (@ 200 *m/z*), AGC target 1 x 10^5 charges, maximum injection time 25 ms, isolation width 1.5 *m/z*, normalized collision energy 27. Precursors with an intensity below 2 x 10^5, an unassigned charge state, charge state 1, or charge >5 were excluded. The dynamic exclusion time was 10 seconds, with isotope peaks of targeted precursors being excluded and the underfill ratio set to 5%.

Training set DIA acquisition. A full MS scan was acquired analyzing 495 – 905 *m/z* with resolution 60,000 (@ 200 *m/z*), AGC target 3 x 10^6 charges, and maximum inject time 100 ms. After the MS scan, 20 MS/MS scans were acquired, each with a 20 *m/z*

wide isolation window, resolution 30,000 @ 200 *m/z*, AGC target 1 x 10^6 charges, maximum injection time 55 ms, normalized collision energy 27, with the default charge state set to 2. The 20 MS/MS scans were contiguous and collectively cover the *m/z* range from 500 – 900 *m/z*. The cycle of 20 scans (center of isolation window) was as follows (*m/z*): 510.4819, 530.4910, 550.5001, 570.5092, 590.5183, 610.5274, 630.5365, 650.5456, 670.5547, 690.5638, 710.5729, 730.5820, 750.5911, 770.6002, 790.6093, 810.6183, 830.6274, 850.6365, 870.6456, and 890.6547. The entire cycle of MS and MS/MS scan acquisition takes roughly 2 seconds and was repeated throughout the LC-MS/MS analysis.

Training set data processing. The DDA data was searched using Comet 2014.02 rev. 2 against a database containing the heavy-labeled peptide sequences. Prior to searching with Comet, the MS/MS spectra had been processed using Hardklor(100) v. 2.16 and Bullseye(101) v. 1.30 to assign more accurate precursor matches based on analysis of MS spectra and remove MS/MS spectra without a matching MS1 precursor. The peptide-spectrum matches were processed with Percolator(35) v. 2.07 to assign q-values to peptide spectrum matches and peptide identifications. Bibliospec(99) v. 2.0 was used to combine the peptide-spectrum matches into a spectral library containing any spectrum with a spectrum-level q-value < 0.3. The score cutoff is extremely loose because the spectral library is simply used as an aide for manually choosing peaks during processing of the DIA data.

The DIA data were analyzed using the Skyline(29) software package. In Skyline, chromatograms were extracted for the +2 and/or +3 charged precursor of each peptide

that fell within the analyzed 500-900 *m/z* range. For each peptide precursor, chromatograms were extracted for the M, M+1, and M+2 precursor ions from the MS data, and chromatograms for the y-ion series (ion 2 to last ion – 1) were extracted from the MS/MS data. The chromatographic peaks for each peptide precursor were manually selected and integrated in each of the four DIA data sets acquired. The retention time of library matches from the DDA data were overlaid on the DIA data to aid in selecting the correct peak. Additionally, the mass measurement error (< 10 ppm), similarity in ratios of the area of the precursor peaks to the theoretical isotope distribution, and similarity in the ratios of the area of the extracted fragment ion chromatograms from the DIA data to matches in the spectral library were used to verify that the correct chromatographic peak was being integrated. In the vast majority of cases, there was a single, intense peak meeting all of these criteria. When this was not the case, the peptide precursor was discarded, resulting in a total of 1,331 confidently detected peptides remaining. Fragment ions showing interference were also discarded.

SRM testing set and training cross validation set. The data presented in Stergachis *et al.* was used as a primary testing data set. A new SRM training cross validation data set was constructed using the protocols presented in Stergachis *et al*. Briefly, clones for GST fusion proteins from the pANT7_cGST clone collection (https://dnasu.org/DNASU/Home.do) were synthesized in vitro using the Pierce 1-step Human Coupled in vitro protein synthesis kit (Thermo Scientific; Bremen, Germany). In instances where a cDNA clone was unavailable, recombinant proteins were purchased from a commercial source. GST tagged proteins were captured using glutathione

sepharose 4B beads (GE Healthcare Life Sciences; Pittsburgh, PA), and iteratively washed to remove nonspecific binders. Bead bound GST fusion proteins were individually denatured with 5 mM dithiothreitol (DTT) for 30 minutes at 60 °C and alkylated with 15 mM iodoacetamide for 30 minutes at room temperature. Proteins were then digested 1 ug of sequencing grade modified porcine trypsin (Promega; Madison, WI) for 2 hours at 37°C.

Protein digests were resolved on a 12 cm x 150 μm analytical column packed with ReproSil-Pur 3μm C18-AQ beads (Dr. Maisch GmbH, Germany). The analytical column was coupled in-line to a Waters nanoAcquity UPLC pump and autosampler (Waters Corp, Milford, MA). Peptides were eluted off the column at a flow rate of 0.75 μL/min using 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B) following this linear solvent schedule: 0 – 7 min, 95% – 60% A; 7.0 – 7.1 min, 60% – 32% A; 7.1 – 8.0 min, 32% A; 8.0 – 8.1 min, 32% – 5% A; 8.1 – 11.0 min, 5% A; 11.0 – 11.1 min, 5% – 95% A; 11.1-18.0 min, 95% A. Peptides are ionized by electrospray and emitted into a TSQ-Vantage triple quadrupole instrument (Thermo Scientific; Bremen, Germany). Doubly charged, fully tryptic peptides of length 7 to 23 for each protein were analyzed using the Skyline software package. Peptide fragment chromatograms for the y-ion series (ion 3 to last ion – 1) were extracted from the MS/MS data and quantified. 44 of the proteins were used for training cross validation to protect against over fitting. The 18 remaining proteins were reserved exclusively for a secondary testing data set and used only after training was complete.

Peptide response prediction. Peptide responses for peptides in the Stergachis *et al* SRM testing data set were predicted using PPA, CONSeQuence, and ESP Predictor. PPA RC4 (available online at http://software.steenlab.org/rc4/PPA.php) was run using the default parameters (peptide mass from 600 to 6000 and minimum peptide length of 5). The artificial neural network and linear support vector machine components of CONSeQuence (available online at http://king.smith.man.ac.uk/CONSeQuence/) were run independent of the consensus binary score. The consensus binary score was not used because it produces only four discrete values, which made it impossible to compare against the other scoring systems. ESP Predictor version 3 (available online at http://www.broadinstitute.org/cancer/software/genepattern/esppredictor) is parameter-free.

APPENDIX C. SUPPLEMENTARY METHODS FOR CHAPTER 3

HeLa cell culture and sample preparation. HeLa S3 cervical cancer cells (ATCC) were cultured at 37°C and 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with L-glutamine, 10% fetal bovine serum (FBS), and 0.5% strep/penicillin. Six cell culture replicates were grown to approximately a 50% density in 6-well plates prior to FBS starvation staggered for 24, 16, 8, 4, 2, and 0 hours (one time point in each well, one plate per replicate). At the 0 hour time point cells were quickly washed three times with refrigerated phosphate-buffered saline and immediately flash frozen with liquid nitrogen. Frozen cells were lysed in a buffer of 9 M urea, 50 mM Tris (pH 8), 75 mM NaCl, and a cocktail of protease inhibitors (Roche Complete-mini EDTA-free). After scraping, cells were subjected to 2x 30 seconds of probe sonication, 20 minutes of incubation on ice, followed by 10 minutes of centrifugation at 21,000 x *g* and 4°C. The protein content of the supernatant was estimated using BCA. The proteins were reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature, and quenched with an additional 5 mM dithiothreitol for 15 minutes at room temperature. The proteins were diluted to 1.8 M urea and then digested with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 12 hours at 37°C. The digestion was quenched by adding 10% trifluoroacetic acid to achieve approximately pH 2. Resulting peptides were desalted with 100 mg tC18 SepPak cartridges (Waters) using vendor-provided protocols and dried with vacuum centrifugation. Peptides were brought to 1 µg / 3 µl in 0.1% formic acid (buffer A) prior to mass spectrometry acquisition. For the reproducibility experiments and to build a chromatogram library we pooled aliquots from all six time points for three of the

replicates to ensure that the pool contained virtually every peptide present in the individual time points.

With the phosphoproteomics experiment, four replicates were performed for each of the four conditions: 20 minute EGF (100 ng/ml) or phosphate-buffered saline (PBS) stimulation following 4 hour starvation, and 20 minute EGF/PBS stimulation following 16 hour starvation. Sample generation and processing was performed in the same fashion with the following exceptions: 1) in addition to protease inhibitors, a cocktail of phosphatase inhibitors (50 mM NaF, 50 mM β-glycerophosphate, 10 mM pyrophosphate, and 1 mM orthovanadate) was also added to the lysis buffer, 2) proteins were digested for 14 hours, and 3) phosphopeptides were enriched using immobilized metal affinity chromatography (IMAC) using Fe-NTA magnetic agarose beads (Cube Biotech). Enrichment was performed with a KingFisher Flex robot (Thermo Scientific), which incubated peptides with 150 µl 5% bead slurry in 80% acetonitrile, 0.1% TFA for 30 minutes, washed them three times with the same solution, and eluted them with 60 µl 50% acetonitrile:1% NH_4OH. Phosphopeptides were then acidified with 10% formic acid and dried. Phosphopeptides were brought to 1 µg / 3 µl in 0.1% formic acid assuming a 1:100 reduction in peptide abundance from the IMAC enrichment. Again, to build a chromatogram library we pooled aliquots from all four conditions for three of the replicates to ensure that the pool contained virtually every peptide present in the individual conditions.

Yeast cell culture and sample preparation. Yeast strain BY4741 (Dharmacon) was cultured at 30°C in YEPD and harvested at mid-log phase. Cell pellets were lysed in

a buffer of 8 M urea, 50 mM Tris (pH 8), 75 mM NaCl, 1 mM EDTA (pH 8) using 7 cycles of 4 minutes bead beating with glass beads followed by one minute rest on ice. Lysate was collected by piercing the tube, placing it into an empty eppendorf, and centrifuging for 1 minute at 2000 rpm and 4C. Insoluble material was removed from the lysate by 15 min centrifugation at 14000 rpm and 4C. The protein content of the supernatant was estimated using BCA. The proteins were reduced with 5 mM dithiothreitol for 30 minutes at 55°C and alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature. The proteins were diluted to 1.8 M urea and then digested with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 16 hours at 37°C. The digestion was quenched using 5N HCl to achieve approximately pH 2. Resulting peptides were desalted with 30 mg MCX cartridges (Waters) and dried with vacuum centrifugation. Peptides were brought to 1 µg / 3 µl in 0.1% formic acid (buffer A) prior to mass spectrometry acquisition.

Liquid chromatography mass spectrometry. Peptides were separated with a Waters NanoAcquity UPLC and emitted into a Thermo Q-Exactive HF or a Thermo Fusion tandem mass spectrometer. Pulled tip columns were created from 75 µm inner diameter fused silica capillary in-house using a laser pulling device and packed with 3 µm ReproSil-Pur C18 beads (Dr. Maisch) to 300 mm. 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 25 mm. Solvent A was 0.1% formic acid in water, while solvent B was 0.1% formic acid in 98% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted using a 90-minute gradient from 5 to 35% B, followed by a 40 minute washing

gradient. Data were acquired using either data-dependent acquisition (DDA) or data-independent acquisition (DIA). Three DDA and DIA HeLa and yeast technical replicates were acquired by alternating between acquisition modes to minimize bias. Serum-starved HeLa acquisition was randomized within blocks to enable downstream statistical analysis.

DDA acquisition and processing. The Thermo Q-Exactive HF was set to positive mode in a top-20 configuration. Precursor spectra (400-1600 m/z) were collected at 60,000 resolution to hit an AGC target of 3e6. The maximum inject time was set to 100 ms. Fragment spectra were collected at 15,000 resolution to hit an AGC target of 1e5 with a maximum inject time of 25 ms. The isolation width was set to 1.6 m/z with a normalized collision energy of 27. Only precursors charged between +2 and +4 that achieved a minimum AGC of 5e3 were acquired. Dynamic exclusion was set to “auto” and to exclude all isotopes in a cluster. Thermo RAW files were converted to mzXML format using ReAdW and searched against a Uniprot Human FASTA database (87613 entries) with Comet (version 2015.02v2), allowing for variable methionine oxidation, and n-terminal acetylation. Cysteines were assumed to be fully carbamidomethylated. Searches were performed using a 50 ppm precursor tolerance and a 0.02 Da fragment tolerance using fully tryptic specificity (KR|P) permitting up to two missed cleavages. Search results were filtered to a 1% peptide-level FDR using Percolator (version 3.1).

DIA acquisition and processing. For each chromatogram library, the Thermo Q-Exactive HF was configured to acquire six chromatogram library acquisitions with 4 m/z

DIA spectra (4 m/z precursor isolation windows at 30,000 resolution, AGC target 1e6, maximum inject time 55 ms) using an overlapping window pattern from narrow mass ranges using window placements optimized by Skyline (i.e. 396.43 to 502.48 m/z, 496.48 to 602.52 m/z, 596.52 to 702.57 m/z, 696.57 to 802.61 m/z, 796.61 to 902.66 m/z, and 896.66 to 1002.70 m/z). Two precursor spectra, a wide spectrum (400-1600 m/z at 60,000 resolution) and a narrow spectrum matching the range (i.e. 390-510 m/z, 490-610 m/z, 590-710 m/z, 690-810 m/z, 790-910 m/z, and 890-1010 m/z) using an AGC target of 3e6 and a maximum inject time of 100 ms were interspersed every 18 MS/MS spectra.

For quantitative samples, the Thermo Q-Exactive HF was configured to acquire 25x 24 m/z DIA spectra (24 m/z precursor isolation windows at 30,000 resolution, AGC target 1e6, maximum inject time 55 ms) using an overlapping window pattern from 388.43 to 1012.70 m/z using window placements optimized by Skyline. Precursor spectra (385-1015 m/z at 30,000 resolution, AGC target 3e6, maximum inject time 100 ms) were interspersed every 10 MS/MS spectra. Phosphopeptide samples were analyzed in the same way using 20x 20 m/z DIA spectra in an overlapping window pattern from 490.47 to 910.66 m/z.

All DIA spectra were programed 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 3.0.7303) where they were peak picked using vendor libraries. A HeLa-specific Bibliospec(99) HCD spectrum library was created from unpublished Thermo Q-Exactive DDA data using Skyline (version 3.1.0.7382). This BLIB library and accompanying iRTDB normalized retention time database were converted into a ELIB library and used to search the mzMLs for peptides. EncyclopeDIA searches DIA

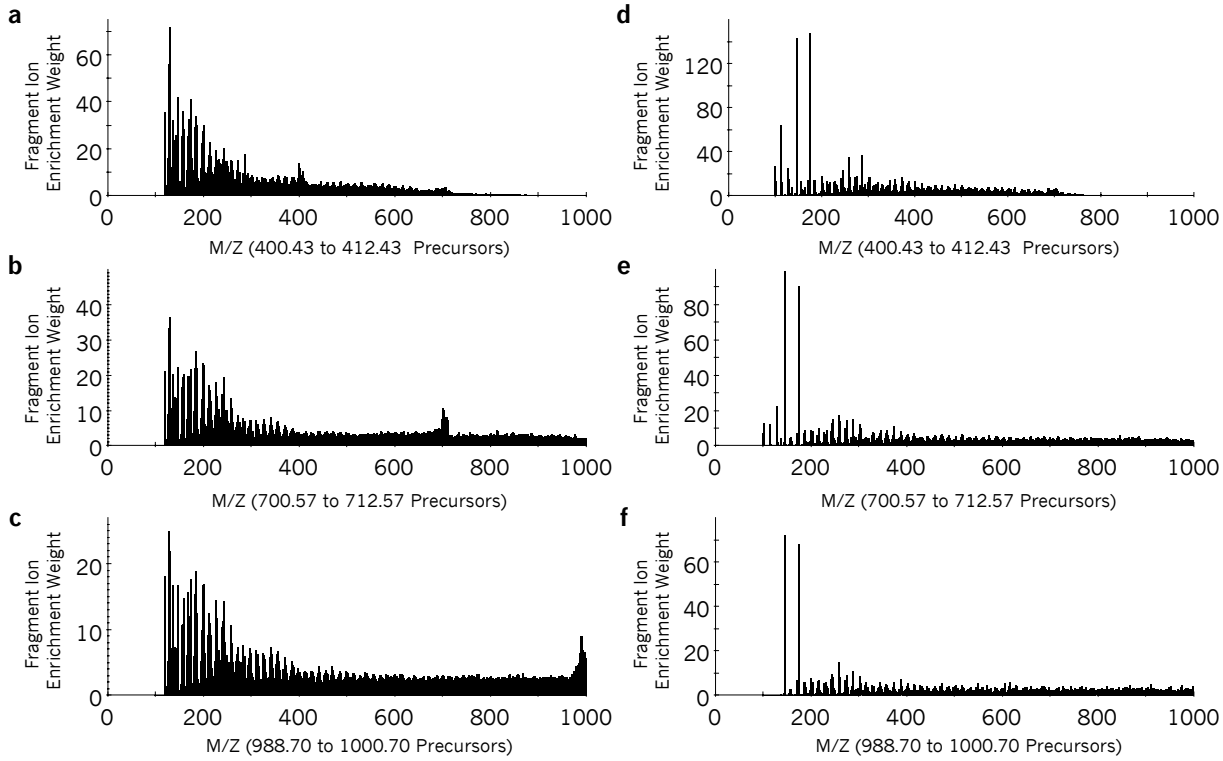
data using +1H and +2H b/y ion fragments that could be found in library spectra. EncyclopeDIA was configured with default settings (10 ppm precursor, fragment, and library tolerances, considering both B and Y ions, and trypsin digestion was assumed). EncyclopeDIA was configured to use Percolator version 3.1. Phosphopeptides were processed the same way except a HeLa-specific phosphopeptide HCD spectrum library was used(38) and phosphopeptides detected in EncyclopeDIA searches were localized using Thesaurus (see Chapter 4 for more details).

Overlapping DIA deconvolution. When using the overlapping DIA scheme, every spectrum in the entire raw file must be deconvoluted. In an effort to maintain consistency between analysis techniques, we used MSConvert to deconvolute RAW files in this study. However, we have also implemented a simple deconvolution algorithm in EncyclopeDIA that can be performed on-the-fly while reading spectra in a narrow I/O buffer. In a DIA data set, at each cycle (T) every MS/MS spectrum (S_{Ti}) comprises fragments from precursors within the precursor isolation window (i). Spectra in consecutive half cycles are overlapped by 50%, such that precursors from the lower 50% of the window in MS/MS spectrum S_{Ti} should also be present in the previous/next half cycles lower offset spectra ($S_{(T-1)(i-1)}$ and $S_{(T+1)(i-1)}$) while precursors from the upper 50% of the window should also be present in the corresponding upper offset spectra ($S_{(T-1)(i+1)}$ and $S_{(T+1)(i+1)}$). We divide these windows into two bins and attempt to determine which fragments were derived from precursors in the upper half or the lower half using previous and next half cycles. Fragment ions that are found exclusively on the lower previous/next spectra ($S_{(T-1)(i-1)}$ and $S_{(T+1)(i-1)}$) are assigned to the lower bin, while those found exclusively in the upper

previous/next spectra ($S_{(T-1)(i+1)}$ and $S_{(T+1)(i+1)}$) are assigned to the upper bin. Ions that are found in both sets of spectra are assigned proportionally to each bin where the proportion is set to the summed peak intensity for both spectra, e.g.: ($S_{(T-1)(i-1)}$ + $S_{(T+1)(i-1)}$) / ($S_{(T-1)(i-1)}$ + $S_{(T+1)(i-1)}$ + $S_{(T-1)(i+1)}$ + $S_{(T+1)(i+1)}$) for the lower bin. Peaks that are found in none of the previous and next overlapping spectra are assumed to be noise. New spectra are built from the deconvoluted peaks in both the lower and upper bins. Since this algorithm only needs to consider three half cycles at a time, deconvolution can happen quickly and in memory, with minimal impact on file reading speeds.

Decoy library entries. A decoy library entry is created for every target library entry. To generate a decoy, first the target peptide sequence (except for digestion enzyme-specific termini) is reversed, insuring that the decoy maintains its appearance as a tryptic peptide. Then fragment ions corresponding to amino acids (B/Y for CID, C/Z/Z+1 for ETD) or their expected neutral losses due to modifications (e.g. phosphorylation) are calculated for both target and decoy entries. If the precursor charge state is greater than +2, then +2 fragment ions are also considered. Uncommon neutral loss ions such as A-type ions or loss of water or ammonia are not considered to limit the likelihood of false detections. Fragment ions that correspond to target sequence m/zs are transferred to new decoy m/zs such that their ion type and index are kept consistent. Delta mass errors in each fragment ion are also maintained to preserve consistency, and all peaks corresponding to the fragment delta mass window are transferred if the library is collected in profile mode. Ions that cannot be assigned to amino acids (such as those

corresponding to precursor ions, background noise or interference) are not used by EncyclopeDIA.



Appendix Figure 1: Fragment ion enrichment weights.

Enrichment weights are normalized frequencies of fragment ion presence/absence (intensity independent) for library peptides in each precursor isolation window. DDA ion frequencies change depending on isolation window, as demonstrated in three different example ranges: (a) 400.43 to 412.43 m/z, (b) 700.57 to 712.57 m/z, and (c) 988.70 to 1000.70 m/z from the HeLa DDA spectrum library. Increased fragment ion frequency in the precursor isolation window is due to unfragmented precursors commonly found in MS/MS experiments. In contrast, DIA chromatogram libraries contain only expected B and Y ions, and consequently are more consistent (d) 400.43 to 412.43 m/z, (e) 700.57 to 712.57 m/z, and (f) 988.70 to 1000.70 m/z. In both cases, ion frequencies increase at lower m/z and are centered around common low m/z B and Y ions.

Ion weighting estimation. While searching, a unique background is calculated for each precursor isolation window using the prevalence of each fragment ion in the library spectra considered for that window (Figure A-1). This background helps estimate the interference frequency for any given ion and is used to weight some scores. This distribution is calculated as the frequency that any nominal m/z fragment ion (rounded by truncation) appears in entries from the library within the specified precursor window filter. m/z frequencies are calculated out to 4000 and a pseudocount is applied to every m/z bin to avoid “zero” frequency errors.

Primary scoring and feature scoring functions. The primary score in EncyclopeDIA conceptually draws on the X!Tandem HyperScore. Unlike scoring functions like XCorr in Sequest, the HyperScore does not attempt to account or penalize for ions that do not match the peptide in question, making it ideal for DIA analysis where coeluting peptides are common. The score function is the weighted dot product of the intensities in the acquired spectrum (*I*) and the library spectrum (*P*), weighted by a correlation score vector (*C*), which is discussed in detail in the Chromatogram Library ELIB Generation section. Again, any ions in the library spectrum that do not correspond to the amino acid sequence are not considered in this score. The dot product is multiplied by the factorial of the number of matching ions:

$$primary\ score = log_{10}\left(\left(\sum_{i=0}^{n} I_i \cdot P_i \cdot C_i\right) \cdot n!\right)$$

Sometimes modified peptides (for example, oxidized peptides) are present in the same precursor isolation window as their unmodified forms. Since often these often

peptides share several fragment ions in common, we require that at least 25% of the score contribution for modified peptides come from ions that exclusively indicate that modification in cases where any of up to four isotopic peaks from the modified/unmodified peptide pairs fall in the same window.

Several more computationally expensive secondary feature scores (Table 5) are calculated once peaks are assigned. Briefly, the scores are divided to cover various classes of features: overall scoring (deltaCN, eValue, logDotProduct, logWeightedDotProduct, xCorrLib, xCorrModel), fragment ion accuracy (sumOfSquaredErrors, weightedSumOfSquaredErrors, numberOfMatchingPeaks, averageAbsFragDeltaMass, averageFragmentDeltaMass), precursor ion accuracy (isotopeDotProduct, averageAbsPPM, averagePPM), and retention time accuracy (deltaRT). The deltaRT score is only used after retention time alignment has been performed. All of these scores are fed to Percolator 3.1 for target/decoy FDR analysis.

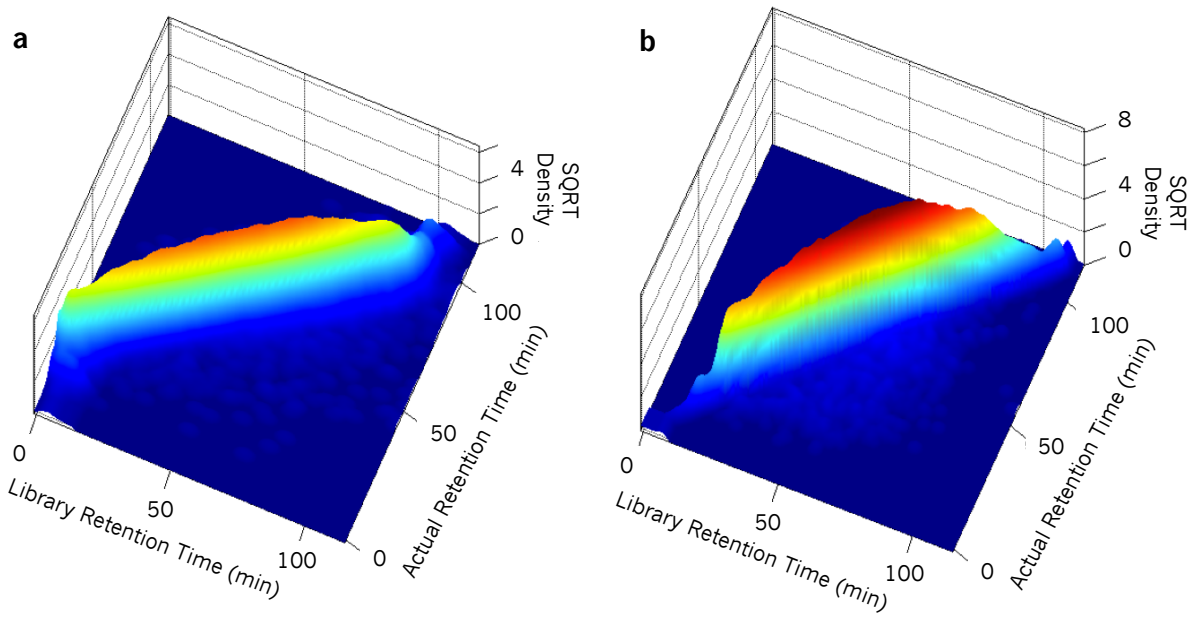
Table 6: EncyclopeDIA score features calculated for Percolator.

Index	Score	Definition
1	primary	the log10 dot product of intensities in the acquired spectrum, the library spectrum, and the correlation score vector, multiplied by the factorial of the number of matching ions
2	xCorrLib	the Sequest xCorr cross correlation score of the acquired spectrum and the library spectrum
3	xCorrModel	the Sequest xCorr cross correlation score of the acquired spectrum and the Sequest peptide model (only b and y ions)
4	LogDotProduct	the log10 dot product of the intensities in the acquired spectrum, the library spectrum, and the correlation score vector
5	logWeightedDot Product	the log10 dot product of the intensities in the acquired spectrum, the library spectrum, the correlation score vector, where each ion is weighted by the background frequency of ions in the library
6	sumOfSquared Errors	the sum of squared errors between target fragment ions in the sum normalized acquired spectrum and the sum normalized library spectrum
7	weightedSumOf SquaredErrors	the sum of squared errors between target fragment ions in the sum normalized acquired spectrum and the sum normalized library spectrum, where each ion is weighted by the background frequency of ions in the library
8	numberOf MatchingPeaks	the number of matching fragment ions between the acquired spectrum and the library spectrum
9	numberOf MatchingPeaks AboveThreshold	the number of matching fragment ions between the acquired spectrum and the library spectrum, where the intensity is greater than the acquired spectrum TIC/(1+N*N) where N is the total number of target fragment ions
10	averageAbs FragmentDeltaMass	the average absolute value of the fragment ion delta masses
11	averageFragment DeltaMasses	the average value of the fragment ion delta masses
12	isotopeDotProduct	the precursor isotope dot product between acquired and predicted for -1, +0, +1, and +2 neutrons
13	averageAbs ParentDeltaMass	the average absolute value of the precursor ion delta masses for -1, +0, +1, and +2 neutrons
14	averageParentDelta Mass	the average value of the precursor ion delta masses for -1, +0, +1, and +2 neutrons
15	eValue	the negative log10 e-value estimated using a log-linear fit

Retention time alignment. Accuracy and stability of retention time alignments is critical for EncyclopeDIA. Consequently, we designed a new algorithm that works analogous to how we visualize densities. This approach uses two dimensional kernel density estimates (KDE) that are much less prone to failure as compared to typical line fitting approaches such as LOESS in situations with grossly variable numbers of points and outliers. In this approach each X/Y coordinate is estimated as a symmetrical, two-dimensional kernel based on a cosine-based Gaussian approximation. Following Silverman's rule(102) the KDE bandwidth is set to:

$$bandwidth = N^{-\frac{1}{6}} \cdot \left(\frac{stdev(x)+stdev(y)}{2}\right)$$

where N is the number of matched peptides. The kernel's standard deviation is set to the bandwidth (analogous to full width at half max) divided by 2*sqrt(2*ln(2)). This distribution is stamped at every X/Y coordinate on a 1000 by 1000 grid mapping from the lowest and highest retention times in both the X and Y dimensions. Once the KDE is calculated, the optimal fit is traced using a ridge walking algorithm that traces the mode of the KDE across retention time (Figure A-2). In this algorithm the highest point in the KDE is identified and the line is fit in increasing retention time by moving to the highest local grid point to the north (increased sample retention time), east (increased library retention time), or northeast. If north and east are both the highest local point, then the line moves to the northeast. This is performed iteratively until the line is fit across the increasing retention time. Then the same ridge walk is performed in decreasing retention time by moving south, west, or southwest. This approach forces a monotonic line (it can never find a negative retention time change) that follows where the most number of X/Y coordinates lie.



Appendix Figure 2: Kernel Density Estimates for Retention Time Alignment.

Three dimensional density plots showing cumulative peptide piles for the DDA spectrum library (a) and DIA chromatogram library (b) retention time alignments in Figure 3.5a and 3.5b. Higher density (Z axis, square root normalized) indicates more peptides at a given retention time point. Treating the density estimate as a mountain range, the retention time alignment approach starts at the point of maximum density and traces the top of the density ridge using a non-parametric ridge walking algorithm to find the optimal alignment. In this approach outliers have relatively low density and generally do not affect the ridge walk algorithm, making it more robust than typical curve fitting strategies.

Retention time alignment mixture model. After the alignment is performed, we use the delta retention time data to produce a mixture model to determine outliers. We calculate a Gaussian distribution representing “correct” retention time matches using the median delta retention time as the Gaussian mean and interquartile range divided by 1.35 as the Gaussian standard deviation. We use a unit distribution to represent “incorrect”

retention time matches. Starting where the distribution priors are set to 0.5, we run 10 iterations of a PeptideProphet-like mixture model(103) to fit the two distributions to the delta retention time data using an Expectation Maximization algorithm(104). Peptide matches with posterior error probability estimations that are less than 5% likely to be in the "correct" retention time distribution are considered outliers.

Retention time alignment across experiments. For each passing peptide, we determine the experiment that produced the best scoring match and set that match aside as a "canonical" peptide representation. We chose the experiment with the most canonical peptides as an anchor and retention time align all of the experiments (and their canonical peptides) to that anchor. Mixture models (described above) for these retention time alignments are calculated and outliers are removed if the local-anchor delta retention time is less than 0.1% likely to fit the mixture model. New retention times for outlier-removed peptides and peptides that were only assigned globally are inferred using the anchor retention time.

FDR filtering peptide and protein detections across experiments. We concatenate peptide feature files from all experiments in a study and run Percolator 3.1 to perform global peptide FDR filtering at 0.01. Using this list of peptides, we generate a parsimonious list of protein groups using a greedy algorithm. Here peptides are assigned to protein groups with the highest protein score:

$$protein\ score(p) = N - \sum_{p \in P}^{N} (PEP_p)$$

where the the sum of the peptide (p) posterior error probabilities (PEPp) is subtracted from the number of peptides (N) assigned to that protein (P). Protein groups are sorted on the lowest PEPp assigned to them(75) and then stringently target/decoy filtered to 0.01 protein FDR.

Automated transition refinement. Fragment ion interference is common when analyzing wide-window data. While fragment ions that show interference may still be useful for detecting peptides, those ions must be screened prior to quantitation to ensure an accurate measurement. We designed a new non-parametric approach to selecting the best ions for quantitation. We first Savizky-Golay smooth(105) the fragment ion chromatograms and then normalize them to have unit integrated intensity. To simplify the smoothing mathematics, we make the assumption that cycle times are consistent within the time frame of a single peak, thus removing the need for interpolation over retention time. After normalization the chromatograms of quantitatively useful ions line up while those of interfered ions will have either higher or lower unit-normalized intensities at different retention times. We calculate the median normalized intensity at each retention time point as an approximation for the peptide peak shape. We then determine peak boundaries by tracing descent of the median peak shape from the maximum normalized intensity on either side of the peak. The boundaries are set to the minimum point at which the median peak trace starts increasing for >2 consecutive spectra or any point where the trace drops to less than 1% of the maximum. At that point we calculate a Pearson's correlation coefficient for the similarity between each fragment ion chromatogram with that of the median peak shape between those boundaries. Peaks that match with a

correlation coefficient of at least 0.9 are considered quantitative, while those that match with coefficients of at least 0.75 are considered useful for detection purposes.

Fragment ion quantification and background subtraction. We calculate trapezoidal peak areas across Savitsky-Golay smoothed chromatograms. Analogous to Skyline, peak intensities are background subtracted by removing a peak area rectangle with a height equal to the largest intensity of either of the boundary edges. If the area of the rectangle is larger than the area of the peak the intensity is set to zero.

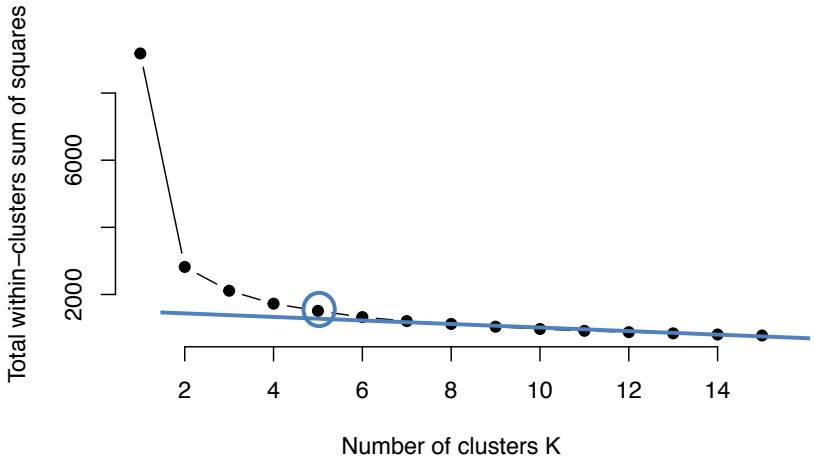
Peptide quantification and transition choice across experiments. Transition interference changes on a sample by sample basis. We rank quantitative transitions (>0.9 correlation) based on the sum of their correlation scores across all experiments (effectively counting the number of samples in which they are observed). In addition, for each transition we calculate a global interference score:

$$interference\ score(t) = \frac{\sum_s I_{s,t}[C_{s,t}<0.9]}{\sum_s I_{s,t}[C_{s,t}\geq 0.9]}$$

Which represents the sum of transition (t) intensities ($I_{t,s}$) across experiments (s) that show interference ($C_{t,s}$<0.9) over those that do not ($C_{t,s}$≥0.9). Transitions with interference scores > 0.2 are deemed untrustworthy for quantification and are dropped. Peptide quantities are set to the sum of the top 5 transitions that pass these criteria, where peptides with fewer than 3 quantitative transitions are not carried forward. We require additional stringent criteria for our time course study. Specifically, we required that each peptide be measured in every replicate of at least one time point, and that cross

experiment CVs (estimated using quantities from each time point corrected with a linear model) be less than 20%.

Protein quantification and statistical testing. Protein quantities were calculated as the sum of peptide quantities. We used Extraction of Differential Gene Expression (EDGE) 3.6(106) to statistically test for reproducible changes across the time course study. We performed k-means clustering of proteins that passed an EDGE q-value filter of 0.01 using 5 groups using 1,000 random starting points with 1,000 iterations. We estimated 5 groups by calculating the sum of within squared errors of each K model from 1 to 15 and estimating the first point where the change in the sum of within squared errors was flat (Figure A-3).



Appendix Figure 3: Choosing K in K-means clustering.
K-means clustering was performed for each K model from 1 to 15 using 1,000 random starting points with 1,000 iterations. We estimated 5 groups by calculating the sum of within squared errors and estimating the first point (highest K) where the change in the sum of within squared errors started to deviate from a flat line.

Gene Ontology enrichment. We performed Gene Ontology enrichment of significantly changing proteins using the online PANTHER Overrepresentation Test(107) (release 20170413) with the Homo sapiens Gene Ontology database (release 2017-10-24) using a background of all proteins consistently detected in our experiments. After removing terms with fewer than 20 proteins (to avoid weakly powered classes) and more than 1,000 proteins (to avoid vague classes), we applied Benjamini-Hochberg FDR correction and filtered enrichment tests to a FDR<0.05.

EncyclopeDIA implementation and software/data availability. EncyclopeDIA is implemented in Java 1.8 as both a command line and a stand-alone GUI application. EncyclopeDIA supports the HUPO PSI mzML standard for reading raw MS/MS data, and can construct DLIB DDA-based spectrum libraries from Skyline/Bibliospec BLIB files, NIST MSP files, or HUPO PSI TraML files. Additionally, EncyclopeDIA results can be imported into Skyline(29) to enable further visualization and downstream processing. EncyclopeDIA is heavily optimized and multi-threaded such that searches can be performed on conventional desktop computers with limited RAM and processing power. We have released source code and cross platform (Windows, Mac OS X, Linux) binaries for EncyclopeDIA on Bitbucket at: https://bitbucket.org/searleb/encyclopedia under the open source Apache 2 license. All mass spectrometry mzML and RAW data files are available on the Chorus Project.

Walnut implementation of PECAN. PECAN(22) (PEptide-Centric Analysis) is an algorithm for detecting peptides from DIA experiments directly without the use of

spectrum libraries. While high-mass accuracy spectrum libraries for human samples and some model organisms are commonly available, sometimes spectrum libraries are either impossible to gather (with precious samples) or not cost-effective to generate for uncommon organisms. PECAN can be used to help ease that burden by removing the need for DDA-based spectrum libraries.

We have implemented a desktop-friendly version of the PECAN scoring system, named Walnut, to enable chromatogram library generation from FASTA protein sequence databases when spectrum libraries are unavailable. We have made minor modifications to the scoring algorithm that heavily optimize it for speed and memory consumption. These modifications result in different scores for some types of peptides compared to the original Python implementation but in general do not affect the performance over all. Walnut is packaged with EncyclopeDIA and the source code is available as part of the EncyclopeDIA repository on Bitbucket at: https://bitbucket.org/searleb/encyclopedia under the open source Apache 2 license.

APPENDIX D. SUPPLEMENTARY METHODS FOR CHAPTER 4

Cell Culture: HeLa cervical cancer cells were cultured at 37°C and 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM) supplemented with L-glutamine, 10% FBS, and 0.5% streptomycin/penicillin. Cells were grown to an estimated 90% confluence in 10-cm plates, where one plate was used for each replicate/condition. Prior to harvest, cells were incubated for 4 hours under serum starvation conditions and then serum stimulated for 30 minutes. MCF-7 breast cancer cells were similarly cultured and starved, followed by stimulation with insulin (100 ng/ml) or IGF-1 (100 ng/ml) in phosphate-buffered saline (PBS) or unstimulated (control, added same volume of PBS) for 20 minutes. Some MCF-7 cells were additionally treated with DMSO or the pan-AKT inhibitor MK-2206 for 40 minutes before stimulation. After stimulation cells were quickly washed three times with refrigerated PBS and immediately flash frozen with liquid nitrogen. With the MCF-7 experiment, six replicates were performed for each of the six conditions: control/DMSO, insulin/DMSO, IGF-1/DMSO, control/MK-2206, insulin/MK-2206, and IGF-1/MK-2206. The six replicates were performed in three cell culture batches to simplify sample handling and ensure precise timing.

Sample Preparation: Frozen cells were lysed in a buffer of 9 M urea, 50 mM Tris (pH 8), and 75 mM NaCl, with a cocktail of protease inhibitors (Roche Complete-mini EDTA-free) and phosphatase inhibitors (50 mM NaF, 50 mM β-glycerophosphate, 10 mM pyrophosphate, and 1 mM orthovanadate). After scraping, cells were subjected to 2 cycles of 25 seconds of probe sonication each followed by 10 minutes of incubation on ice. Lysates were centrifuged for 10 minutes at 21,000 x *g* and 4°C to eliminate cell debris.

The protein content of the supernatant was estimated using BCA. For every condition/replicate, an estimated 850 µg of protein was reduced with 5 mM dithiothreitol for 30 minutes at 55°C, alkylated with 10 mM iodoacetamide in the dark for 30 minutes at room temperature, and the alkylation was quenched with an additional 5 mM dithiothreitol for 15 minutes at room temperature. The proteins were diluted to 1.8 M urea and then digested with sequencing grade trypsin (Pierce) at a 1:50 enzyme to substrate ratio for 4 hours at 37°C. The digestion was quenched by adding 10% trifluoroacetic acid (TFA) to achieve pH ~ 2. Resulting peptides were desalted with 100 mg tC18 SepPak cartridges (Waters) using vendor-provided protocols and dried with vacuum centrifugation. Phosphopeptides were enriched using immobilized metal affinity chromatography (IMAC) using Fe-NTA magnetic agarose beads (Cube Biotech). Enrichment was performed with a KingFisher Flex robot (Thermo Scientific), which incubated peptides with 150 µl 5% bead slurry in 80% acetonitrile, 0.1% TFA for 30 minutes, washed them three times with the same solution, and eluted them with 60 µl 50% acetonitrile:1% NH_4OH. Phosphopeptides were then acidified with 10% formic acid and dried. Phosphopeptides were brought to 1 µg / 3 µl in 0.1% formic acid assuming a 1:100 reduction in peptide abundance from the IMAC enrichment.

Liquid Chromatography Mass Spectrometry: Phosphopeptides were separated with a Waters NanoAcquity UPLC and emitted into a Thermo Q-Exactive HF or a Thermo Fusion tandem mass spectrometer. Pulled tip columns were created from 75 um inner diameter fused silica capillary in-house using a laser pulling device and packed with 3 µm ReproSil-Pur C18 beads (Dr. Maisch) to 300 mm. 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 25 mm. Solvent A was 0.1% formic acid in water, while solvent B was 0.1% formic acid in 98% acetonitrile. For each injection, 3 µl (approximately 1 µg) was loaded and eluted using a 90-minute gradient from 5 to 25% B, followed by a 40 minute washing gradient. Data were acquired using data-dependent acquisition (DDA), data-independent acquisition (DIA), or parallel reaction monitoring (PRM). Four DDA and DIA HeLa technical replicates were acquired in alternating mode to avoid bias. MCF-7 sample acquisition was randomized within blocks to enable downstream statistical analysis.

DDA Acquisition and Processing: The Thermo Q-Exactive HF was set to positive mode in a top 12 configuration. Full MS scans of mass range 400-1600 were collected at 60,000 resolution to hit an AGC target of 3e6. The maximum inject time was set to 100 ms. MS/MS scans were collected at 30,000 resolution, AGC target of 1e6, and maximum inject time of 55 ms. The isolation width was set to 1.5 m/z with a normalized collision energy of 27. Only precursors charged between +2 and +4 that achieved a minimum AGC of 1e4 were acquired. Dynamic exclusion was set to “auto” and to exclude all isotopes in a cluster.

Thermo .RAW files were converted to .mzXML format using ReAdW and searched against a Uniprot Human FASTA database (downloaded July 1 2014 to maintain consistency with Lawrence *et al*(38), 87,613 entries) with Comet (version 2015.02v2), allowing for variable methionine oxidation, protein N-terminal acetylation, and phosphorylation at serines, threonines, and tyrosines. Cysteines were assumed to be fully carbamidomethylated. Searches were performed using a 50 ppm precursor tolerance and

a 0.02 Da fragment ion tolerance using fully tryptic specificity (KR|P) permitting up to two missed cleavages. Search results were filtered to a 0.6% PSM-level (Peptide to Spectrum Match-level) FDR using Percolator (version 3.1), which we determined in this experiment to closely track to a 1% peptide-level FDR. Site localization was performed using an in-house implementation of Ascore that was modified to not compete positional isomers against each other in order to have a higher chance of detecting overlapping isomers. We set Ascore to use a 0.02 Da fragment ion tolerance and we filtered for phosphopeptides with at least one corresponding PSM that produced an Ascore value >= 20 (p-value<0.01).

DIA / PRM Acquisition and Processing: The Thermo Q-Exactive HF was configured to acquire 20 MS/MS scans at 30,000 resolution, AGC target 1e6, maximum inject time 55 ms, using overlapping precursor isolation windows of 20 m/z units and centered at: [500.4774, 520.4865, 540.4956, 560.5047, 580.5138, 600.5229, 620.5319, 640.541, 660.5501, 680.5592, 700.5683, 720.5774, 740.5865, 760.5956, 780.6047, 800.6138, 820.6229, 840.632, 860.6411, 880.6502, 900.6593, 490.4728, 510.4819, 530.491, 550.5001, 570.5092, 590.5183, 610.5274, 630.5365, 650.5456, 670.5547, 690.5638, 710.5729, 730.582, 750.5911, 770.6002, 790.6093, 810.6183, 830.6274, 850.6365, 870.6456, and 890.6547]. Full MS scans (mass range 485-925, resolution 30,000, AGC target 3e6, maximum inject time 100 ms) were interspersed every 18 scans. MS/MS scans were programed with normalized collision energy of 27 and an assumed charge state of +2.

For PRMs, the Thermo Fusion was configured to collect MS/MS scans corresponding to 62 precursor targets in the PI3K/AKT signaling pathway scheduled with 10-minute retention time windows using Phosphopedia. The large 10-minute window enables both scheduling from Phosphopedia without additional calibration runs and the detection of alternate positional isomers that may elute far away from the target. Full MS scans (mass range 400-1600) were collected at 60,000 resolution to hit an AGC target of 3e6. The maximum inject time was set to 100 ms. MS/MS scans were collected at mass range of 100-1600, resolution of 30,000, AGC target of 1e6, and maximum inject time of 55 ms. The isolation width was set to 0.7 m/z with a normalized collision energy of 27.

A Bibliospec(99) HCD spectrum library of tryptic phosphopeptides was created from the Thermo Q-Exactive data previously published in Lawrence *et al.*(38) using Skyline (version 3.1.0.7382)(29). This .BLIB library and accompanying .iRTDB normalized retention time database were used to search the .mzMLs for peptides. Thermo .RAW files were converted to .mzML and .mzXML formats using the ProteoWizard package (version 3.0.10922) where they were peak picked using vendor libraries and deconvoluted using Prism in “overlap_only” mode. We used EncyclopeDIA (https://www.biorxiv.org/content/early/2018/03/07/277822), a library search engine to detect peptides in a peptide-centric approach. Our engine searches DIA data using b- and y-ion fragments of charges +1 and +2, and includes phosphate neutral losses that can be found in library spectra. We applied the following settings: 30,000 resolution (effectively +/- 16.7 ppm tolerance) for precursor, fragment, and library. Detected features were assigned and corrected to <0.01 FDR using Percolator version 3.1. We also used DIA-Umpire to extract peptide signatures using the same DIA files after overlap-

deconvolution. DIA-Umpire was configured to use 10 ppm mass tolerances and to extract +2 to +4 charged peptides. DIA-Umpire produces three .MGFs for each .mzXML; all three were searched separately with Comet and the results were combined together for Percolator and Ascore interpretation. All Comet, Percolator and Ascore settings were identical to those used for the DDA experiments.

Initial Thesaurus Scoring: Site-specific fragment ions can distinguish positional isomers, but applying current site localizing tools originally designed for DDA(39) to DIA can be problematic because they assume a constant background noise level inconsistent with the high level of background interference ions found in DIA data. We account for this by calculating a background frequency distribution to estimate the likelihood of detecting an interfering signal as the frequency of each m/z in the raw file for a given precursor isolation window. Our approach allows us to quickly query the distribution of ions for every fragment ion individually, enabling the use tight mass tolerances (measured in ppm) to assess the likelihood of interference.

For each queried phosphopeptide, we determine the set of combinatorial permutations corresponding to the number of phosphorylations and the number of potential phosphoacceptor sites. If a positional isomer permutation is not present in the library, then a synthetic library spectrum is generated from the anchor by shifting fragment ion peak intensities for each B-type, Y-type, +2H, and neutral loss ions to the appropriate M/Zs, using an approach similar to that used in SpectraST(91).

First we extract chromatograms for fragment ions in a window +/- 10% of the total acquired chromatographic time from the retention time anchor. At every retention time

point in that window we calculate a score (Thesaurus score) based on the X!Tandem HyperScore where the function is the dot product of the intensities in the acquired spectrum (*I*) and the library spectrum (*P*) multiplied by the factorial of the number of matching ions:

$$Thesaurus\ Score = Log_{10}\left(\left(\sum_{i=0}^{n} I_i \cdot P_i\right) \cdot n!\right)$$

Iterative Localization Scoring: We begin the localization process by comparing the highest scoring isomer and retention time point to every other alternate isomer (*j*) that either a) hasn't been detected, b) has been detected nearby this RT, or c) scores higher at this RT. We calculate a p-value as the probability of finding all of the detected site-specific ions (*n*) by chance from the background frequency distribution (*m*) and the total number of site-specific ions considered (*N*). The final localization p-value is the max (least significant) of these values across alternate isomers (*j*):

$$localization\ p-value = \prod_{i=0}^{n} p(null|m_i)$$

The localization score is the -log10(p-value) to produce a positive score for higher confident localizations. This score is smoothed across time by Gaussian weighting, where the Gaussian standard deviation is estimated from the expected peak width (here we used 25 sec/6). Thesaurus then extracts the chromatographic shape of the localizing fragment ions (for the target isomer compared to the alternate isomer with the least significant p-value) to calculate the IonCount score, a measure of the number of co-eluting fragments. The peak shape of every co-eluting b/y ion (i) associated with the target isomer

is compared to the shape defined by the localizing fragment ions using Pearson's correlation (c_i). The IonCount score is calculated as:

$$IonCount\ score = \sum_{i=0}^{n} (c_i)^2$$

The sum of squares of correlation values is used to heavily downweight the impact of ions that poorly correlate with the target isomer. Only ions with positive correlation scores are used. The target isomer is considered detected if the apex p-value p <=0.01 and the IonCount score is >=3 (these thresholds are user adjustable). If the target isomer does not pass these thresholds, the retention time window is blacklisted for this isomer and up to one more attempt can be made to localize the peptide. The localization process is iterated until all isomers have been detected or ruled out.

Localization Post Processing: A parallel process is performed using decoy-generated spectra and the scoring features for both are fed into Percolator 3.1 to generate Q-values. Of peptides that pass the detection Q-value threshold, an additional localization Q-value is calculated using the Benjamini-Hochberg method and the top reported localization p-values for each isomer. Detected isomers are filtered to a user-settable Q-value (typically 0.01 or 0.05) using both thresholds to ensure high confidence detections.

Positional isomer searching can be performed in an "uncalibrated" manner (where retention times in the library are assumed to be precise) if the DIA data was searched directly with a DIA library search tool or if DDA experiments were run concurrently (SWATH). Alternatively, Thesaurus supports searching in a "calibrated" manner, which assigns relative retention time ordering by searching each peptide anchor across the entire experiment window using the Thesaurus score if retention times are unknown (e.g.

importing NIST libraries) or need to be calibrated (e.g. with spectrum libraries acquired on different platforms, gradients, or HPLC columns). Alternatively, searches can be performed across the entire acquisition window when analyzing targeted PRM data. We have also enabled options for only calculating localization scores and estimating FDR, skipping the detection of positional isomers that are not found in the library.

Quantitation and Statistical Analysis: We used strict criteria to consider a localized peptide quantifiable. In addition to the localization scoring requirements, we also required at least three quantitative fragment ions and that the localized isomer was observed in every replicate of at least one condition. Thesaurus uses the site specific fragment ions to determine the shape of the peak and assigns quantitative fragment ions as those that match that shape for quantification with Pearson's correlation coefficients greater than 0.9. Quantification was performed by summing the background-subtracted peak areas of site-specific fragment ions or all fragment ions, depending on the level of peptide separation. Background subtraction removes the trapezoidal area below the peak integration window. Integrated intensities were normalized within each replicate group, and across groups to the control intensity median. Statistical analysis was performed via 2-tailed t-tests paired within each replicate for comparing quantitative changes in positional isomers from the same peptide, or globally with Benjamini-Hochberg FDR corrected one-way ANOVAs.

PRM results were further validated with follow-up analysis in Skyline-Daily (version 3.6.1.10615). Skyline was configured to extract all +1 and +2 b- and y-ions, including neutral losses of phosphate, as well as precursor traces for the monoisotopic, first and

second isotopes. After initially importing the runs, peptides were hand-curated to match the retention time boundaries determined by site-localizing analysis. Fragment ions that appeared to be interfered with were removed from the analysis.

Software and Data Availability: Thesaurus is an open-source, cross-platform Java application available at http://bitbucket.org/searleb/thesaurus/. While Thesaurus can read and produce reports to enable Skyline analysis, it is a stand-alone tool and does not require any other software to run. Thesaurus is fully multi-threaded and designed to work on typical desktop computers where searches of individual mzML files in this experiment took on average of 10-12 minutes to complete with commodity PC hardware. All mass spectrometry files presented here have been deposited to the Chorus Project (https://chorusproject.org/) with the project identifier 1374.

Frequently asked questions:

How is reproducibility limited when analyzing phosphoproteomes with DDA?

The most common mass spectrometric method for large-scale phosphopeptide profiling uses DDA, in which MS survey scans are collected in one to three second intervals and MS/MS peptide sequencing scans are triggered for the top N most abundant ions. A narrow range around each precursor ion is selected for fragmentation, which generates an MS/MS spectrum containing several sequence specific ions. Once an MS/MS scan has been triggered, that ion is placed on an exclusion list, which is designed to limit wasted measurements on the same peptide. This process, called dynamic exclusion, is a key optimization that allows DDA to identify peptides from low abundance proteins.

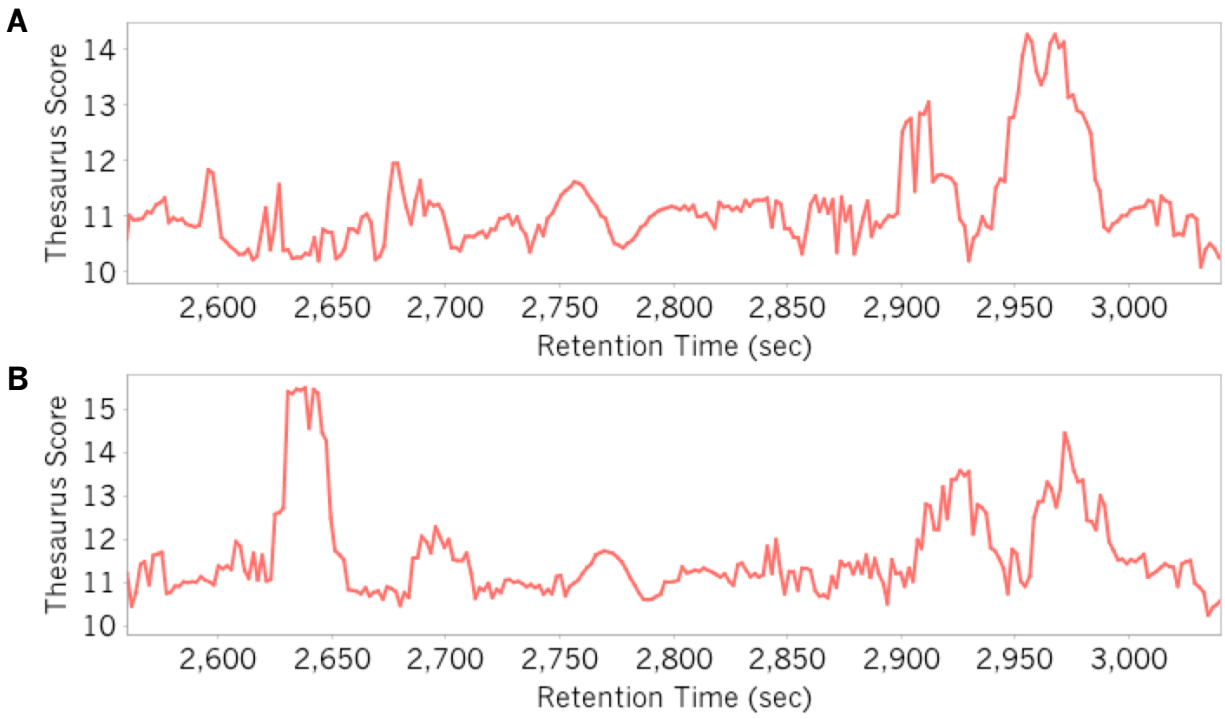
However, the trade off of deep sampling is that estimating quantitative phosphorylation levels is often hampered by a lack of consistent detection of low-level phosphorylation events due to stochastic MS/MS sampling. While the accuracy of site-specific phosphorylation calls can be estimated, the general approach is limited because localization is performed using data from a single fragmentation spectrum that is rarely collected near the apex of the peptide chromatographic peak. This can complicate the detection of position-specific fragment ions that do not always rise above the background signal. Even worse, dynamic exclusion actively fights against the ability to detect multiple positional phosphoisomers by intentionally excluding sampling of other isobaric peptides. Moreover, even if multiple species are detected by DDA, the precursor XIC's that are typically used to quantify these peptides are convoluted and cannot be distinguished.

How do fragmentation methods affect phosphopeptide fragmentation?

While CID is possible in modern Q-Orbitraps, in these instruments HCD is a faster fragmentation method that enables more reasonable cycle times in DIA experiments. HCD fragmentation may have additional benefits for phosphoproteomics as it yields fewer MS/MS spectra than CID that are dominated by precursor neutral loss peaks. Savitski *et al*(42) noted that ETD more often results in fragmentation along the phosphopeptide backbone, rather than fragmentation producing phosphate neutral losses. Fragment ions with neutral losses can be useful for peptide detection purposes, but cannot be used for localization because they overlap with normal loss-of-water peaks from unphosphorylated serines and threonines. Preliminary results of using DIA with ETD fragmentation(108) may provide even more specificity for positional isomer localization.

With peptide-centric searching I can look for specific positional isomers in PRM/DIA data. Why is this potentially dangerous without using localization tools?

While it is possible to use spectrum library search engines to detect phosphopeptides, we do not recommend it without using some sort of localization software. Unlike with spectrum-centric DDA searching where one implicitly competes phosphopeptide localizations against each other (e.g. Mascot Delta Score(42)), with peptide-centric DIA searching one only attempts to find the highest scoring spectrum across retention time as an indication for the presence of a peptide. While this works well for unmodified peptides, shared ions in modified peptides can lead to false identifications. As an illustration, consider the peptide KGSGDpYMPMSPK from Figure 4-11. In Figure A-4, panel A shows the Thesaurus score (which uses both site-specific and shared fragment ions) trace from a control sample, while panel B shows the same score trace from an IGF-1 stimulated sample. Despite that KGSGDpYMPMSPK should be below the level of detection in the control sample, a high score (>14) can be observed at the time point associated with another phosphorylated form, KGSGDYMPMpSPK, at 49.5 minutes because of a large number of shared peaks. Without considering site-specific ions, this score is high enough to be considered a detection at a FDR <0.01. Only when considering an IGF-1 stimulated sample is it possible to determine the true retention time for KGSGDpYMPMSPK at 44 minutes.



Appendix Figure 4: Thesaurus scores for KGSGDpYMPMSPK.
(A) Thesaurus scores from a control sample and (B) from a sample stimulated with IGF-1.

Is it possible to use DDA localization tools for PRM/DIA experiments?

Ascore(39), MaxQuant, and several other DDA site localization algorithms rely on probabilities generated from a binomial distribution, which enables those algorithms to ask the question “what is the likelihood of drawing a site-specific peak out of a random background?” This approach inherently assumes as unit-probability background, which is sensible for DDA experiments where interference is relatively low (Supplementary Figure 1A). While they may potentially work for PRM experiments, applying current site localizing tools originally designed for DDA to DIA data sets can be problematic because the assumed constant background noise level is inconsistent with the high level of

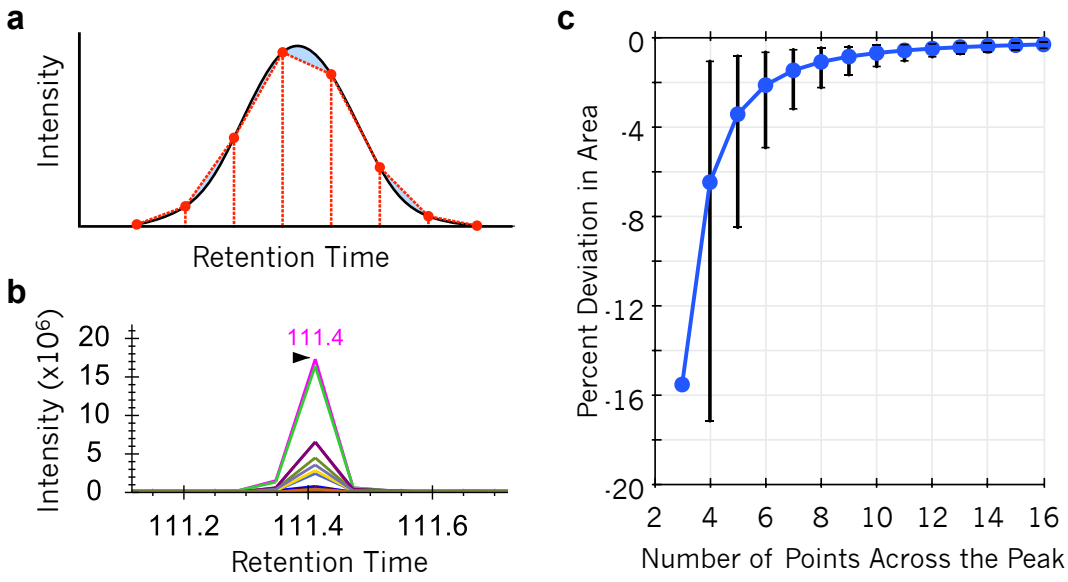
background interference ions found in DIA data. That said, this issue is somewhat alleviated by requiring peak shape consistency(89).

Why are there detections of multiple positional isomers at the same retention time point?

We find that approximately a quarter of positional isomers elute within five seconds of each other. Our retention time resolution is dependent on cycle time, and as our cycles are around two seconds each, these positional isomers are effectively chromatographically inseparable in our liquid chromatography (LC) conditions. In these cases, typical localization approaches that compete positional isomers found from the same chromatographic peak against each other either will detect at most only one form. Previously the only way to separate these forms was to use multi-phase LC separation or ion mobility devices that separate peptides under different constraints, but Thesaurus can detect and quantify these using site-specific ions. A subset of these isomer pairs suggests that there is a dominant isomer. Some of these cases might be indicative of phosphate rearrangement where the phosphate moves from one site to the other in the gas-phase(93). One way to conclusively separate true co-eluting peptides from gas-phase rearrangements is to determine quantitative changes between the forms across multiple conditions. In all cases we feel it is critical to quantify co-eluting phosphopeptides using site-specific ions only to avoid contaminated measurements.

APPENDIX E. DESIGNING A DIA METHOD

Background: Data independent acquisition (DIA) is a series of compromises between comprehensive detection and selective quantitation where the goal is to detect as many peptides as possible, but in a way that maintains quantitative rigor. With modern mass spectrometers you can only scan at a maximum rate of 10-20 Hz without sacrificing spectrum quality. An additional constraint is that at least 6-8 measurements are required to describe a quantitative peak (Figure A-5).



Appendix Figure 5: The effect of scan rate on peak integration accuracy. (a) Error (shaded in blue) in trapezoidal quantitation (red dashed lines) should cancel when measuring a Gaussian peak (black solid line) with 8 points across the peak. (b) However, trapezoidal quantitation can produce poor measurements with only 5 points (or fewer). (c) The average percent deviation with 95% error bars in actual/calculated area at different number of points across a Gaussian peak.

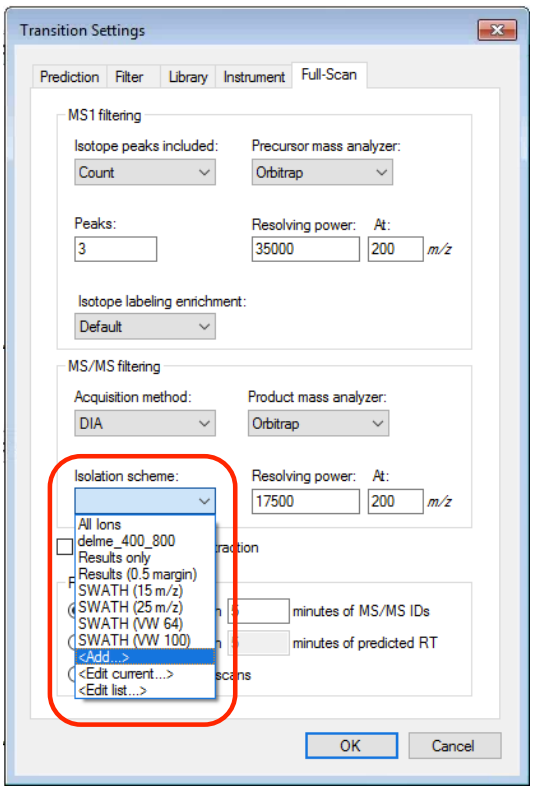
DIA methods make several compromises to manage these constraints. In particular, instead of trying to monitor all peptides, the measurement range is typically

limited to measure peptides in a restricted precursor window (e.g. 400-1000 m/z). In addition, precursor isolation windows are widened such that multiple peptides are usually isolated together and co-fragmented, resulting in higher interference.

In general we recommend structure windowing methods assuming 10 measurements across an average peak so that regions of the chromatogram with narrower peaks will not be underrepresented. We can calculate the necessary cycle time and windowing scheme from the average peak width and maximum scan rate. If we assume peak widths of 25 seconds and 10 Hz (typical for many quadrupole orbitraps such as the Thermo Fusion, QE, QE+, or QE-HF), we require cycle times of 2.5 seconds, resulting in 25 windows per cycle. This allows for 25x 24 m/z-wide precursor isolation windows to cover 600 m/z in windowing range (400-1000 m/z). However, if we use a faster scanning instrument that can achieve 20 Hz (typical for the Thermo Lumos, QE-HFX, and many TOFs), we can use up to 50 windows per cycle. This allows for 50x 12 m/z-wide precursor isolation windows to cover the same 600 m/z in windowing range.

Setting up a windowing scheme with Skyline: Skyline makes it easy to set up a regularly spaced windowing scheme. After downloading and installing Skyline 4[1] navigate to the “Settings/Transition Settings…” menu option to bring up the Transition Settings dialog (Figure A-6).

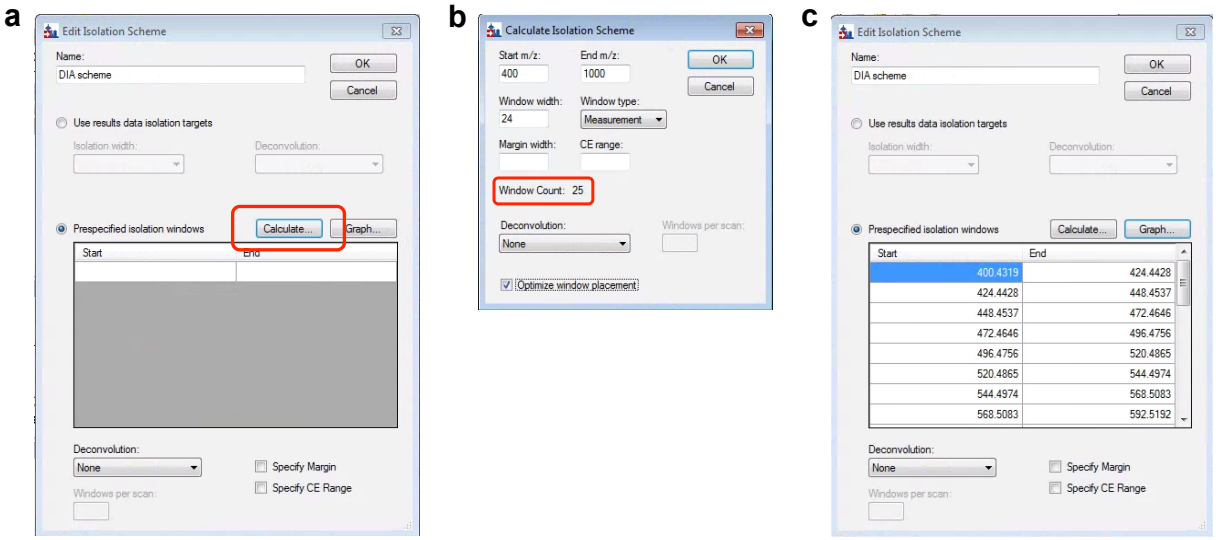
[1] https://skyline.ms/project/home/software/Skyline/begin.view



Appendix Figure 6: Adding a new isolation scheme with Skyline.

Set the "Acquisition method" to DIA, and "<Add…> a new "Isolation scheme". Here you can select "Prespecified isolation windows" and "Calculate…" a new windowing scheme (Figure A-7a). In the Calculate dialog (Figure A-7b) specify the start and stop m/z range (typically 500-900 m/z or 400-1000 m/z), and the calculated precursor isolation window width (e.g. 24 or 12 m/z). Skyline will automatically calculate the total window count. We recommend using "None" or "Overlap" deconvolution for normal and overlap window placement, respectively. We highly recommend selecting "Optimize window placement" to sit the windows in regions between nominal m/z values where peptides are unlikely to exist. Some older instruments (e.g. Thermo Fusion or QE) have quadrupole geometries that are not optimized for DIA. These instruments benefit from small (<0.5

m/z) margins added to the window width. At this time Skyline cannot be used to generate variable width windows.



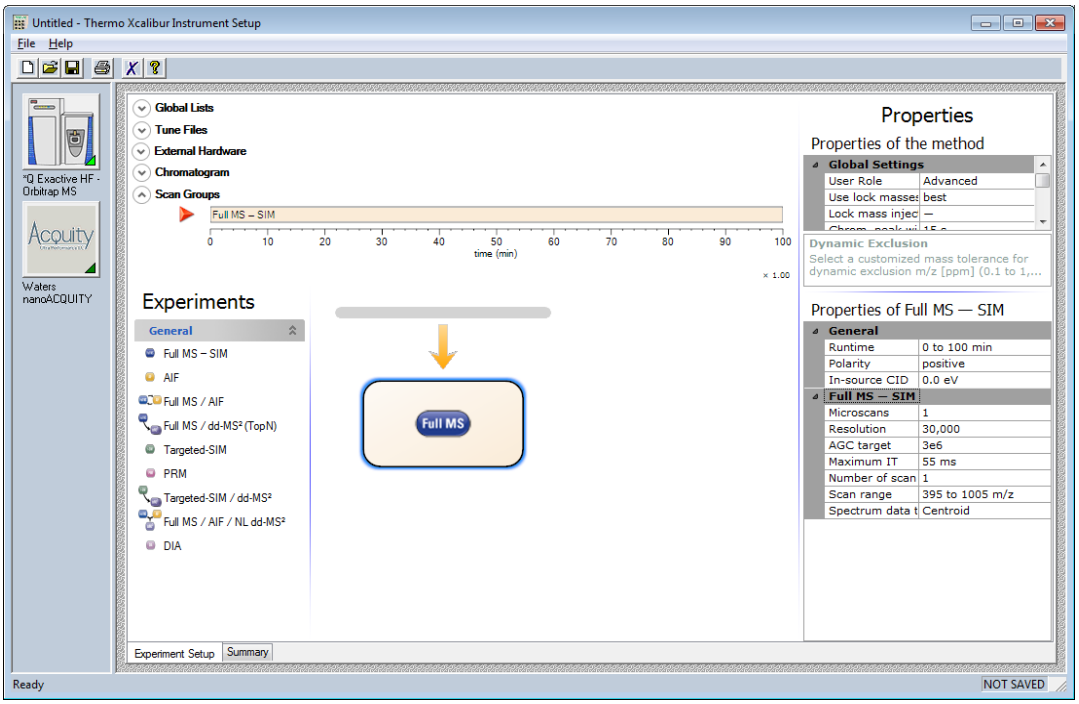
Appendix Figure 7: **Calculating a new isolation scheme with Skyline.**
(a) The Isolation Scheme dialog. (b) The Calculate Isolation Scheme dialog. (c) An automatically generated isolation scheme.

The resulting isolation scheme (Figure A-7c) can be exported or copy and pasted directly into method editors.

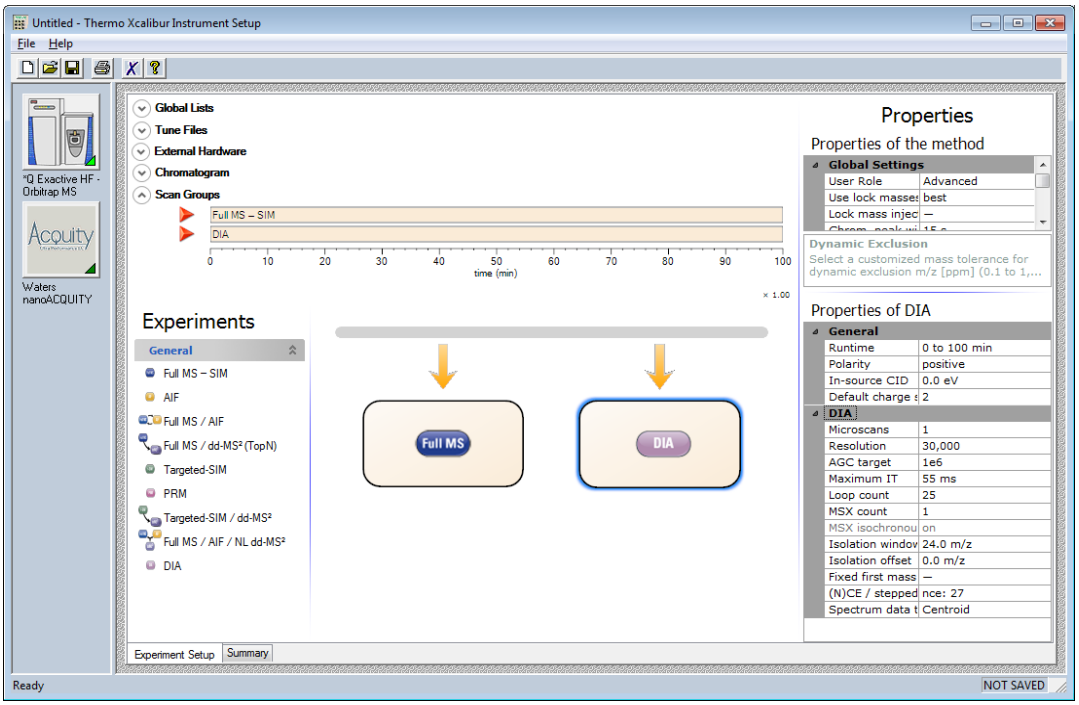
Setting up a DIA instrument method: Here we will use the Thermo QE-HF method editor, but most of these settings are transferable to other platforms. Starting with a new method, we recommend adding a precursor scan followed by DIA MS/MS scans (Figure A-8). While precursor scans are not required for detection or quantitation, we find they are useful for troubleshooting experiments while acquiring data. We recommend setting the resolution and maximum ion inject time to a lower value than for a typical DDA experiment to limit the time wasted on collecting MS scans. Also, we recommend only

scanning through the precursor isolation range of the experiment. Here for a 400-1000 m/z experiment we use a precursor scan of 395-1005 m/z. Since these windows are not being used for quant, we recommend collecting MS survey scans in centroid mode to help keep file sizes down.

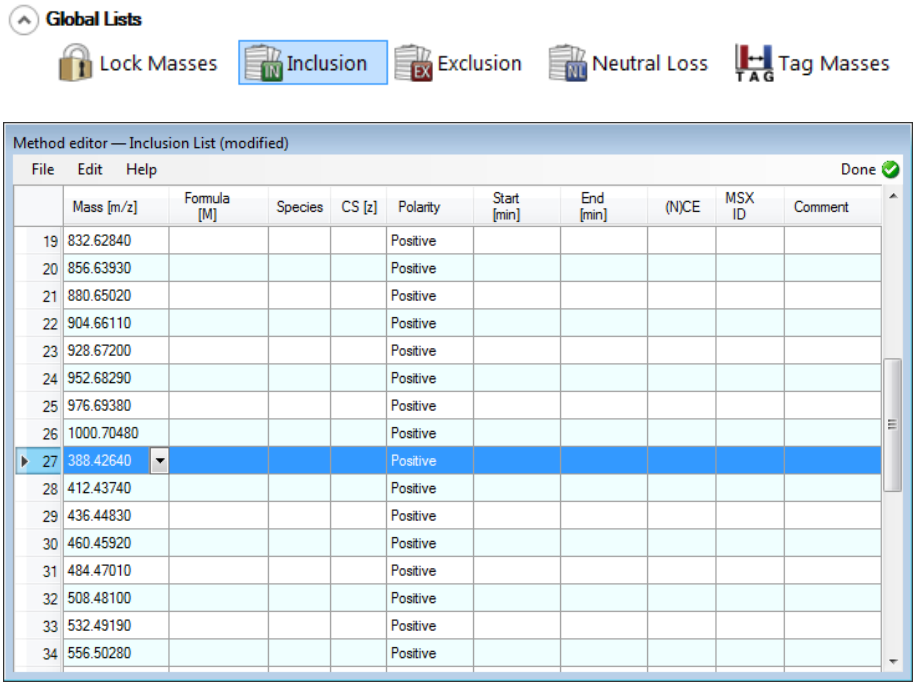
For DIA scans (Figure A-9), we set the default charge to 2 and a complementary collision energy setting (NCE) to what you typically use for DDA. In general we set automatic gain control (AGC) target intensity to a very high value so that every MS/MS scan is acquired to the maximum ion inject time (max IT) duration, ensuring regular scans time differences between cycles throughout the experiment. Again, collecting scans in centroid mode helps keep file sizes down. Finally, we add the inclusion list from Skyline (Figure A-10). When generating overlapping windows, a bug in Skyline causes it to add an extra window outside the range. This window should be deleted before saving the method.



Appendix Figure 8: Settings for a QE-HF "Full MS - SIM" scan.



Appendix Figure 9: Settings for QE-HF "DIA" scans.



	Mass [m/z]	Formula [M]	Species	CS [z]	Polarity	Start [min]	End [min]	(N)CE	MSX ID	Comment
19	832.62840				Positive					
20	856.63930				Positive					
21	880.65020				Positive					
22	904.66110				Positive					
23	928.67200				Positive					
24	952.68290				Positive					
25	976.69380				Positive					
26	1000.70480				Positive					
27	388.42640				Positive					
28	412.43740				Positive					
29	436.44830				Positive					
30	460.45920				Positive					
31	484.47010				Positive					
32	508.48100				Positive					
33	532.49190				Positive					
34	556.50280				Positive					

Appendix Figure 10: Adding the inclusion list from Skyline.

The blue highlighted line shows an unnecessary window (center at 388 m/z) added by Skyline that should be removed prior to method use.

REFERENCES:

1. Wu CC, MacCoss MJ. Shotgun proteomics: tools for the analysis of complex biological systems. Curr Opin Mol Ther. 2002;4:242-250.

2. Mertins P, Mani DR, Ruggles KV et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534:55-62.

3. Zhang B, Wang J, Wang X et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513:382-387.

4. Stahl DC, Swiderek KM, Davis MT, Lee TD. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. J Am Soc Mass Spectrom. 1996;7:532-540.

5. Wilhelm M, Schlegl J, Hahne H et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582-587.

6. Kim MS, Pinto SM, Getnet D et al. A draft map of the human proteome. Nature. 2014;509:575-581.

7. Richards AL, Hebert AS, Ulbrich A et al. One-hour proteome analysis in yeast. Nat Protoc. 2015;10:701-714.

8. Kondrat RW, McClusky GA, Cooks RG. Multiple reaction monitoring in mass spectrometry/mass spectrometry for direct analysis of complex mixtures. Analytical chemistry. 1978

9. Peterson AC, Russell JD, Bailey DJ, Westphall MS, Coon JJ. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. Mol Cell Proteomics. 2012;11:1475-1488.

10. Lawless C, Holman SW, Brownridge P et al. Direct and Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. Mol Cell Proteomics. 2016;15:1309-1322.

11. Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nat Methods. 2004;1:39-45.

12. Zhang Y, Bilbao A, Bruderer T et al. The Use of Variable Q1 Isolation Windows Improves Selectivity in LC-SWATH-MS Acquisition. J Proteome Res. 2015;14:4359-4371.

13. Egertson JD, Kuehn A, Merrihew GE et al. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013;10:744-746.

14. Soni MH, Cooks RG. Selective injection and isolation of ions in quadrupole ion trap mass spectrometry using notched waveforms created using the inverse Fourier transform. Analytical Chemistry. 1994;66:2488-2496.

15. Panchaud A, Scherl A, Shaffer SA et al. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. Anal Chem. 2009;81:6481-6488.

16. Searle BC, Pino LK, Egertson JD et al. Comprehensive peptide quantification for data independent acquisition mass spectrometry using chromatogram libraries. bioRxiv. 2018277822.

17. Venable JD, Yates JR. Impact of ion trap tandem mass spectra variability on the identification of peptides. Anal Chem. 2004;76:2928-2937.

18. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5:976-989.

19. Tsou CC, Avtonomov D, Larsen B et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods. 2015;12:258-64, 7 p following 264.

20. Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. Proteomics. 2009;9:1696-1719.

21. Ting YS, Egertson JD, Payne SH et al. Peptide-Centric Proteome Analysis: An Alternative Strategy for the Analysis of Tandem Mass Spectrometry Data. Mol Cell Proteomics. 2015;14:2301-2307.

22. Ting YS, Egertson JD, Bollinger JG et al. PECAN: library-free peptide detection for data-independent acquisition tandem mass spectrometry data. Nat Methods. 2017;14:903-908.

23. Weisbrod CR, Eng JK, Hoopmann MR, Baker T, Bruce JE. Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification. J Proteome Res. 2012;11:1621-1632.

24. Röst HL, Rosenberger G, Navarro P et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.[letter]. Nat Biotechnol 2014;32(3):219-223.

25. Bruderer R, Bernhardt OM, Gandhi T et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol Cell Proteomics. 2015;14:1400-1410.

26. Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ. Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat Protoc. 2015;10:887-903.

27. Schubert OT, Gillet LC, Collins BC et al. Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc. 2015;10:426-441.

28. Bekker-Jensen DB, Kelstrup CD, Batth TS et al. An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. Cell Syst. 2017;4:587-599.e4.

29. MacLean B, Tomazela DM, Shulman N et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26:966-968.

30. Levin VA, Panchabhai SC, Shen L, Kornblau SM, Qiu Y, Baggerly KA. Different changes in protein and phosphoprotein levels result from serum starvation of high-grade glioma and adenocarcinoma cell lines. J Proteome Res. 2010;9:179-191.

31. Pirkmajer S, Chibalin AV. Serum starvation: caveat emptor. Am J Physiol Cell Physiol. 2011;301:C272-9.

32. Jensen ON. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol. 2006;7:391-403.

33. Lao YW, Gungormusler-Yilmaz M, Shuvo S, Verbeke T, Spicer V, Krokhin OV. Chromatographic behavior of peptides containing oxidized methionine residues in proteomic LC-MS experiments: Complex tale of a simple modification. J Proteomics. 2015;125:131-139.

34. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods. 2007;4:207-214.

35. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4:923-925.

36. Ubersax JA, Ferrell JE. Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol. 2007;8:530-541.

37. Schweiger R, Linial M. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biol Direct. 2010;5:6.

38. Lawrence RT, Searle BC, Llovet A, Villén J. Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. Nat Methods. 2016;13:431-434.

39. Beausoleil SA, Villén J, Gerber SA, Rush J, Gygi SP. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat Biotechnol. 2006;24:1285-1292.

40. Lu B, Ruse C, Xu T, Park SK, Yates J. Automatic validation of phosphopeptide identifications from tandem mass spectra. Anal Chem. 2007;79:1301-1310.

41. Olsen JV, Blagoev B, Gnad F et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127:635-648.

42. Savitski MM, Lemeer S, Boesche M et al. Confident phosphorylation site localization using the Mascot Delta Score. Mol Cell Proteomics. 2011;10:M110.003830.

43. Liebler DC, Zimmerman LJ. Targeted quantitation of proteins by mass spectrometry. Biochemistry. 2013;52:3797-3806.

44. Marx V. Targeted proteomics. Nat Methods. 2013;10:19-22.

45. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012;9:555-566.

46. Stergachis AB, MacLean B, Lee K, Stamatoyannopoulos JA, MacCoss MJ. Rapid empirical discovery of optimal peptides for targeted proteomics. Nat Methods. 2011;8:1041-1043.

47. Bereman MS, MacLean B, Tomazela DM, Liebler DC, MacCoss MJ. The development of selected reaction monitoring methods for targeted proteomics via empirical refinement. Proteomics. 2012;12:1134-1141.

48. Muntel J, Boswell SA, Tang S et al. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment (PPA). Mol Cell Proteomics. 2015;14:430-440.

49. Eyers CE, Lawless C, Wedge DC, Lau KW, Gaskell SJ, Hubbard SJ. CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. Mol Cell Proteomics. 2011;10:M110.003384.

50. Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol. 2009;27:190-198.

51. Mallick P, Schirle M, Chen SS et al. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol. 2007;25:125-131.

52. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986;323:533.

53. Mead JA, Bianco L, Ottone V et al. MRMaid, the web-based tool for designing multiple reaction monitoring (MRM) transitions. Molecular & Cellular Proteomics. 2009;8:696-705.

54. Prakash A, Tomazela DM, Frewen B et al. Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. Journal of proteome research. 2009;8:2733-2739.

55. Schwanhäusser B, Busse D, Li N et al. Global quantification of mammalian gene expression control. Nature. 2011;473:337.

56. de Graaf EL, Altelaar AF, van Breukelen B, Mohammed S, Heck AJ. Improving SRM assay development: a global comparison between triple quadrupole, ion trap, and higher energy CID peptide fragmentation spectra. J Proteome Res. 2011;10:4334-4341.

57. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000;28:374.

58. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226-1238.

59. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005;3:185-205.

60. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science. 1988;240:1648-1652.

61. Yutani K, Ogasahara K, Tsujita T, Sugino Y. Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. Proceedings of the National Academy of Sciences. 1987;84:4441-4444.

62. Wilce MCJ, Aguilar M-I, Hearn MTW. Physicochemical basis of amino acid hydrophobicity scales: evaluation of four new scales of amino acid hydrophobicity coefficients derived from RP-HPLC of peptides. Analytical chemistry. 1995;67:1210-1219.

63. Maxfield FR, Scheraga HA. Status of empirical methods for the prediction of protein backbone topography. Biochemistry. 1976;15:5138-5153.

64. George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Engineering, Design and Selection. 2002;15:871-879.

65. Nakashima H, Nishikawa K. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS letters. 1992;303:141-146.

66. PALAU JAUME, ARGOS PATRICK, PUIGDOMENECH PERE. Protein secondary structure. Chemical Biology & Drug Design. 1982;19:394-401.

67. Vasquez M, Nemethy G, Scheraga HA. Computed conformational states of the 20 naturally occurring amino acid residues and of the prototype residue α-aminobutyric acid. Macromolecules. 1983;16:1043-1049.

68. Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria1. Journal of molecular biology. 2001;309:835-843.

69. Bengio Y, LeCun Y. Scaling learning algorithms towards AI. Large-scale kernel machines. 2007;34:1-41.

70. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. Anal Chem. 2005;77:6364-6373.

71. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem. 2004;76:3908-3922.

72. Gillet LC, Navarro P, Tate S et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics. 2012;11:O111.016717.

73. Röst HL, Liu Y, D'Agostino G et al. TRIC: an automated alignment strategy for reproducible protein quantification in targeted proteomics. Nat Methods. 2016;13:777-783.

74. Fenyö D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem. 2003;75:768-774.

75. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. J Am Soc Mass Spectrom. 2016;27:1719-1727.

76. Ghaemmaghami S, Huh WK, Bower K et al. Global analysis of protein expression in yeast. Nature. 2003;425:737-741.

77. Escher C, Reiter L, MacLean B et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics. 2012;12:1111-1121.

78. Rosenberger G, Bludau I, Schmitt U et al. Statistical control of peptide and protein error rates in large-scale targeted data-independent acquisition analyses. Nat Methods. 2017;14:921-927.

79. Pardee AB. G1 events and regulation of cell proliferation. Science. 1989;246:603-608.

80. Lam H, Deutsch EW, Eddes JS et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics. 2007;7:655-667.

81. Boersema PJ, Foong LY, Ding VM et al. In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. Mol Cell Proteomics. 2010;9:84-99.

82. Villén J, Beausoleil SA, Gerber SA, Gygi SP. Large-scale phosphorylation analysis of mouse liver. Proc Natl Acad Sci U S A. 2007;104:1488-1493.

83. Huang CY, Ferrell JE. Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci U S A. 1996;93:10078-10083.

84. Nash P, Tang X, Orlicky S et al. Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. Nature. 2001;414:514-521.

85. Chiu JC, Ko HW, Edery I. NEMO/NLK phosphorylates PERIOD to initiate a time-delay phosphorylation circuit that sets circadian clock speed. Cell. 2011;145:357-370.

86. Liu YF, Herschkovitz A, Boura-Halfon S et al. Serine phosphorylation proximal to its phosphotyrosine binding domain inhibits insulin receptor substrate 1 function and promotes insulin resistance. Mol Cell Biol. 2004;24:9668-9681.

87. Fermin D, Walmsley SJ, Gingras AC, Choi H, Nesvizhskii AI. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. Mol Cell Proteomics. 2013;12:3409-3419.

88. Rosenberger G, Liu Y, Röst HL et al. Inference and quantification of peptidoforms in large sample cohorts by SWATH-MS. Nat Biotechnol. 2017;35:781-788.

89. Meyer JG, Mukkamalla S, Steen H, Nesvizhskii AI, Gibson BW, Schilling B. PIQED: automated identification and quantification of protein modifications from DIA-MS data. Nat Methods. 2017;14:646-647.

90. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl. 2015;9:745-754.

91. Ma CW, Lam H. Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring. J Proteome Res. 2014;13:2262-2271.

92. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013;13:22-24.

93. Palumbo AM, Reid GE. Evaluation of gas-phase rearrangement and competing fragmentation reactions on protein phosphorylation site assignment using collision induced dissociation-MS/MS and MS3. Anal Chem. 2008;80:9735-9747.

94. Mayer G, Montecchi-Palazzi L, Ovelleiro D et al. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. Database. 2013;2013

95. Sikora JL, Logue MW, Chan GG et al. Genetic variation of the transthyretin gene in wild-type transthyretin amyloidosis (ATTRwt). Human genetics. 2015;134:111-121.

96. Kim YJ, Gallien S, El-Khoury V et al. Quantification of SAA1 and SAA2 in lung cancer plasma using the isotype-specific PRM assays. Proteomics. 2015;15:3116-3125.

97. Khan Z, Bloom JS, Amini S, Singh… M. Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. Molecular systems …. 2012

98. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi… K. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016

99. Frewen BE, Merrihew GE, Wu CC, Noble WS, MacCoss MJ. Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal Chem. 2006;78:5678-5684.

100. Hoopmann MR, Finney GL, MacCoss MJ. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Analytical chemistry. 2007;79:5620-5632.

101. Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. Comparison of database search strategies for high precursor mass accuracy MS/MS data. Journal of proteome research. 2009;9:1138-1143.

102. Silverman BW. Density estimation for statistics and data analysis. CRC press; 1986

103. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002;74:5383-5392.

104. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 19771-38.

105. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry. 1964;36:1627-1639.

106. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc Natl Acad Sci U S A. 2005;102:12837-12842.

107. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013;8:1551-1566.

108. Carvalho PC, Han X, Xu T et al. XDIA: improving on the label-free data-independent analysis. Bioinformatics. 2010;26:847-848.

VITA

Following an undergraduate degree in chemistry from Reed College, Brian was mentored in MS/MS-based proteomics by Ashley McCormack and software development by Mark Turner in Srinivasa Nagalla's lab at Oregon Health and Science University. In 2004 Brian co-founded Proteome Software with Mark and Ashley to produce and distribute cutting-edge data analysis software for proteomicists. As the owner of Proteome Software, Brian has produced numerous innovations in the analysis of MS/MS-based proteomics data. Brian is a member of the American Society for Mass Spectrometry (ASMS) and served on the board of directors for that organization. Brian also co-founded the Proteome Informatics Research Group of the Association of Biomolecular Resource Facilities (ABRF) society and is active in ABRF committees. Brian carried out his doctoral studies at University of Washington under the guidance of Michael J. MacCoss.