Computational Framework for Data-Independent Acquisition Proteomics

By

Chih-Chiang Tsou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2016

Doctoral Committee:

Associate Professor Alexey I. Nesvizhskii, Chair
Professor Philip C. Andrews
Assistant Professor Brent R. Martin
Associate Professor Kayvan Najarian
Associate Professor Maureen Sartor

Acknowledgements

First of all, I would like to thank my research mentor and dissertation advisor Dr. Alexey Nesvizhskii for the guidance towards the completion of this dissertation and giving me the opportunity to participate such an interesting project. Also, the work presented in this dissertation would not have been possible if we did not have all the supports and collaborations with Dr. Anne-Claude Gingras and her lab members. Her complementary insights on mass spectrometry and biological applications have tremendously helped the developments of the project. I am also grateful to have Dr. Philip Andrews, Dr. Brent Martin, Dr. Kayvan Najarian, and Dr. Maureen Sartor as the thesis committees. Their useful inputs and critical comments on the dissertation have been very important for me to finish the dissertation. I also would like to thank the guys in Nesvizhskii lab and all my friends, thank you for the constant supports and the company.

This dissertation is dedicated to my family, especially my lovely wife Wan-Hui Chang. I could not have finished the degree without her generous supports. Our two year old daughter, Annie, has been the greatest joy in my life and further motivated me to complete the degree and pursue the professional career in the next stage.

Table of Contents

List of Figures

List of Tables

List of Appendices

Appendix

Abstract

Mass spectrometry (MS) is one of the main techniques for high throughput discovery- and targeted-based proteomics experiments. For years, the most popular method for MS data acquisition has been the so-called data dependent acquisition (DDA) strategy which primarily selects high abundance peptide species for tandem mass spectrum sequencing. In order to reach low abundance peptides, most DDA strategies incorporate stochastic data acquisitions to avoid repetitive sequencing of same peptide over consecutive scan cycles, therefore resulting in relatively irreproducible qualitative and quantitative results for low abundance peptides between experiments. Data independent acquisition (DIA), in which peptide fragment signals are systematically acquired for all the peptides within a certain mass range, is emerging as a promising alternative to address the stochasticity of the conventional DDA. DIA by design results in more complex signals, posing a major computational challenge for complex sample and high-throughput analysis. As a result, targeted extraction which is dependent on pre-existing spectral libraries has been the most commonly used approach for automated DIA data analysis. However, building spectral libraries requires additional amount of analysis time and sample materials which are the major barriers for most proteomics research groups.

In my dissertation, I develop a computational tool called DIA-Umpire, which is comprised of multiple computational and signal processing algorithms to enable untargeted DIA identification and quantification analysis without relying on any prior spectral library. In the first study, a signal feature detection algorithm is developed to extract and assemble peptide precursor and fragment signals into pseudo tandem mass spectra which can be analyzed by the existing DDA untargeted analysis tools. This novel step enables direct and untargeted (spectral library-free) DIA identification analysis and we show the performance using complex samples including human cell lysate and glycoproteomics datasets. In the second study, a hybrid approach is developed to further improve the DIA quantification sensitivity and reproducibility. The performance of DIA-Umpire quantification approach is demonstrated using an affinity-purification mass spectrometry experiment for protein-protein interaction analysis. Lastly, in the third study, I improve the DIA-Umpire pipeline for data obtained from the Orbitrap family of mass spectrometers. Using several publicly available datasets, I show that the improved version of DIA-Umpire is capable of highly sensitive, untargeted and direct (spectral library-free) analysis of DIA data for the data generated using Orbitrap family of mass spectrometers. The dissertation work addresses the barriers of DIA analysis and should facilitate the adoption of DIA strategy for a broad range of discovery proteomics applications.

Chapter 1 Introduction

1.1 *Mass spectrometry-based proteomics*

Proteomics is the large-scale study of proteins, which includes protein sequence analysis, structural proteomics, interaction proteomics, post-translational protein modification, and etc. It has been greatly accelerated because of the achievements of genomics and mass spectrometry. Genome sequencing provided the blueprint of possible gene products which are comprised of the basis of proteomics, has shifted proteomics field from purely hypothesis-driven science to discovery science. The remarkable breakthrough in ionization techniques, including electrospray ionization (ESI) and matrix assisted laser desorption & ionization (MALDI), have enabled the detection of larger molecules such as proteins and peptides using mass spectrometers. As a result, a combination of liquid chromatography (LC) and mass spectrometry (MS), LC-MS, has rapidly evolved as a powerful technology for high-throughput proteomics analysis in a wide range of discovery-based biological applications. The most popular approach for high-throughput proteomics analysis is the so-called 'shotgun proteomics'. In a typical shotgun proteomics experiment, proteins are first digested into peptides using a proteolytic enzyme such as trypsin, and the resulting peptide samples are separated using LC coupled online to a tandem mass spectrometer. As peptides elute from the LC column, they are ionized as peptide ions and subjected to a survey scan (MS1) and further to tandem mass spectrometry analysis to obtain MS/MS (also called MS2) spectra. The spectral peaks in MS1 spectra, representing mass-to-charge ratio (m/z) and intensities of detected peptide ions, only indicate the observed molecular mass of peptide species and, are insufficient to uniquely identify them. In order to identify the detailed amino acid compositions of peptides, tandem mass spectrometry isolates ionized peptides with specific mass-to-charge ratio (*m/z*) by mass filter and breaks the isolated peptide ions into shorter fragments. For an isolated peptide ion signal, also called peptide precursor ion, the fragment signals are recorded as an MS/MS spectrum. Various computational strategies [1] including *de novo* sequencing and MS/MS database search algorithms can be applied to identify peptide amino acid composition given an MS/MS spectrum. Once peptide sequences are identified, protein inference strategies can further identify protein identities [1, 2].

1.2 *Data dependent acquisition*

In a typical LC-MS experiment with a complex proteomics sample, digested peptides are separated by LC and ionized into different charged forms, resulting in millions of ionized signals in a single LC-MS run (as the example of LC-MS image shown in Figure 1-1). The number of potential peptide precursor ions increases exponentially if we consider modified peptide species derived from chemical and post-translational modifications. With the huge number of ion signals, the current mass spectrometers do not have sufficient scan speed to acquire an MS/MS spectrum for each one of the precursor signals. As a result, most experiments adopt a compromised strategy called data dependent acquisition (DDA) [3]. In DDA, each scan cycle begins with a MS1 survey scan to detect peptide precursor ions. The *m/z* values of the top few most intense peptide precursor ions in the MS1 survey scan are then automatically isolated and fragmented by mass spectrometer to acquire corresponding MS/MS spectra in the following scans, as the illustration shown in Figure 1-2. Dynamic exclusion strategy [4, 5] is often applied with DDA to avoid repetitive sampling of same peptide precursor ions over a short period of time to increase identification coverage. As mentioned above, the MS/MS spectra can be used to identify the peptides and proteins by various computational methods, most commonly by MS/MS database search engines. In a single LC-MS run, DDA can effectively identify 6,000 - 10,000 proteins. For years, DDA has been the most popular approach for high-throughput proteomics experiments.

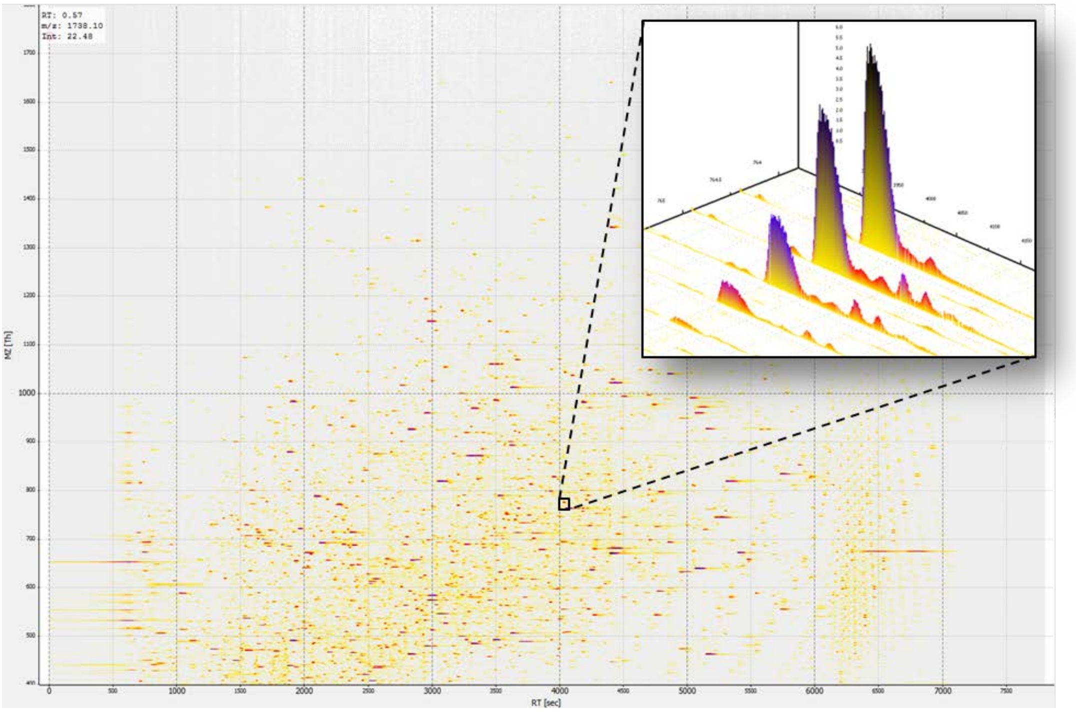


Figure 1-1 Illustration of LC-MS data and isotope peaks of a peptide precursor.

The *x*-axis is retention time, *y*-axis is *m/z*, and each dot represents a possible ion signal with the colors indicating its intensity. The image was exported using OpenMS 1.10 [6].

Despite the wide use of DDA strategy, its limitations are also well known and have been discussed in the literature. Even with a continuously increasing speed of data acquisition, the mass spectrometers are not able to reliably isolate and acquire high quality MS/MS spectra on all peptides present in a typical proteomics sample. Due to the dependency of ion selection on the precursor ion intensity, identification of low abundance peptides is more stochastic and less reproducible between replicate LC-MS analyses. In addition, the selection of a peptide ion for MS/MS sequencing during DDA is not guaranteed to always be at the LC elution peak apex, which may reduce the quality of the MS/MS spectrum and make its computational interpretation more difficult. These issues also affect the accuracy of protein quantification. The spectral-count based quantification strategies, while very robust and easy to use, are most affected by the stochastic nature of DDA. When the quantification is directly based on the number of acquired and identified MS/MS spectra for that protein, robust and sensitive detection of abundance changes across different samples for low abundance proteins becomes difficult due to their naturally small (and variable) spectral counts [7]. MS1 peak intensity-based quantitation approaches allow more sensitive quantification of proteins identified by one or several peptides. However, missing quantification remains a problem which is also caused by DDA's stochastic acquisition and not all MS1 peaks have MS/MS spectra to identify peptide identity.

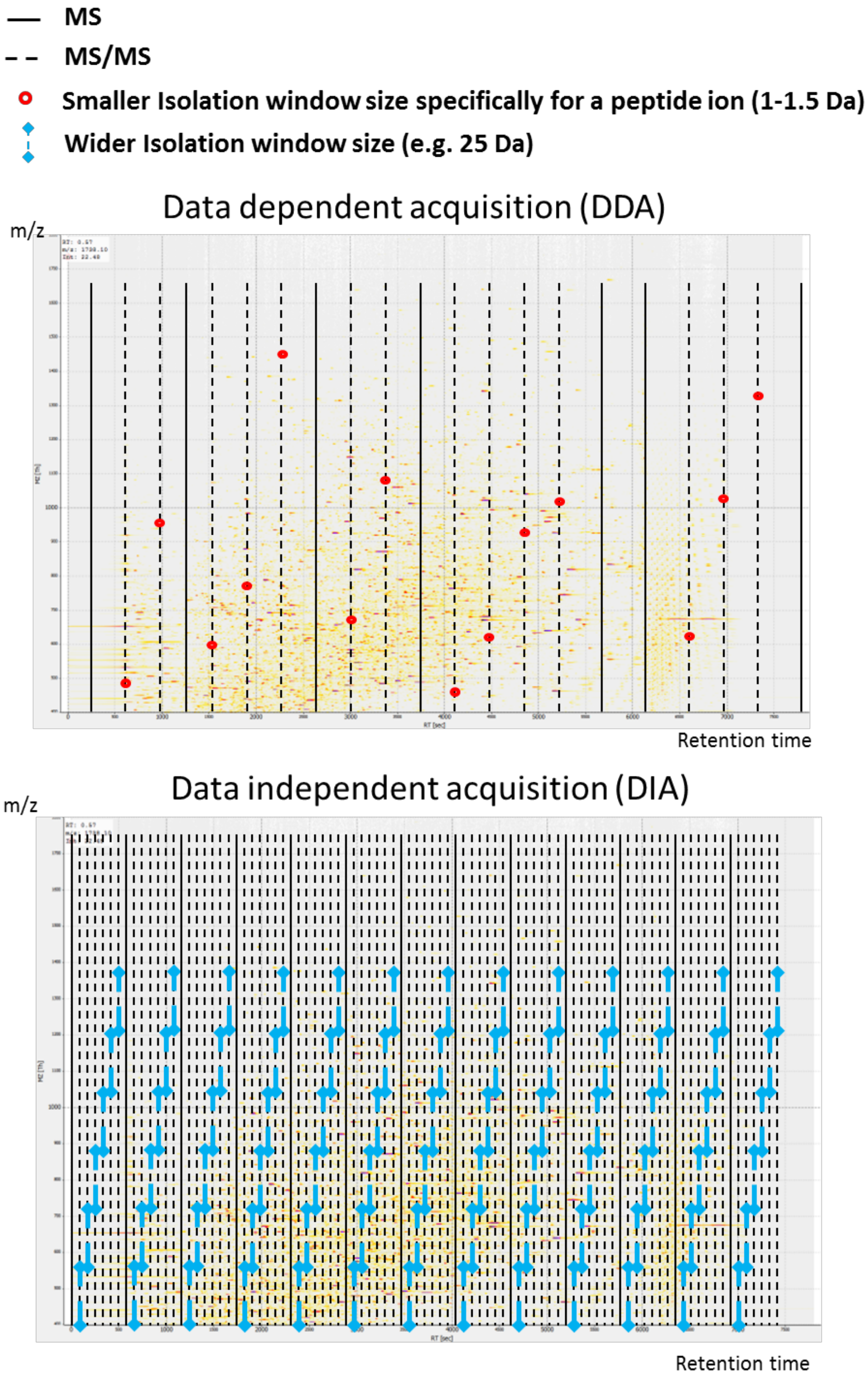


Figure 1-2 Illustration of the difference between DDA and DIA.

In each LC-MS image, the *x*-axis is retention time and *y*-axis is *m/z*. In each scan cycle, the solid line represents the MS1 survey scan and the dash lines represent the following MS/MS spectra. **DDA**: Each scan cycle begins with an MS1 survey scan followed by several MS/MS scans. MS/MS scans are acquired based on a smaller size of isolation window for a specific peptide precursor ion. Because the insufficient scan speed, DDA results in more stochastic identification and quantification performance. **DIA**: Each scan cycle also begins with an MS1 survey scan. In the following MS2 spectra, DIA uses a wider isolation window size to allow systematic acquisition of fragmentation signals across the entire mass and retention time ranges. The wider isolation window size used in DIA causes peptide co-fragmentation and results in more complex MS/MS spectra.

1.3 *Data independent acquisition*

The alternative to the DDA method are data-independent acquisition (DIA) methods [8-18], where the fragment ion information are acquired for all precursor ions within a certain range of *m/z* values. By design, DIA results in more complex MS/MS spectra because the signals are from co-fragmentation of multiple co-eluting peptides. The early DIA strategies had to use a very large precursor isolation window (e.g. the entire reliably measurable *m/z* range in the case of MS^E approach) to keep the duty cycle time within the limits established by the peptide LC elution peak width [8]. Alternatively, the isolation window could be significantly narrowed (e.g. to 1 - 4 Da), but at the expense of splitting the analysis of one sample into multiple LC-MS runs each covering a different range of *m/z* values [11, 15]. In the first case, the resulting MS/MS spectra were often too complex to be effectively analyzed, whereas in the second case the potential advantages of improved sensitivity of peptide identification were negated by the increase in the overall MS analysis time. As a result, the number of studies employing the DIA strategies for untargeted (discovery) proteomics has trailed significantly those based on the conventional DDA approach. Instead, a variant of DIA – multiple reaction monitoring (MRM) [19-21] – has gained a wider use for targeted protein quantification [22]. In targeted proteomics applications [21] using MRM, the analysis is restricted to a small number of predetermined peptide ions of interest. Selected fragment ions corresponding to these peptide ions are measured continuously over a period of time. This enables building an extracted ion chromatogram (XIC) specifically for these fragment ions, thereby allowing higher sensitivity of detection and better quantitation accuracy for selected peptides of interest.

Recent improvements in MS instrumentation have significantly widened the window of opportunity for applying DIA strategies in proteomics studies. Sequential Window Acquisition of all THeoretical Mass Spectra (SWATH) [16] is a variant of the DIA strategy that takes advantage of the increased scan speed and improved mass resolution available on newer instruments such as AB Sciex 5600 TripleTOF and Orbitrap mass spectrometers. In SWATH MS DIA (and related workflows [16]) an intermediate isolation window size (e.g. 25 Da in SWATH) is used instead of a narrow *m/z* window as in DDA or a very large window as in MS^E DIA. By sequentially stepping up the *m/z* windows across the wide mass range (e.g. 400 Da - 1200 Da), fragments of virtually all peptide ions from this range should be present in the corresponding MS/MS spectra (Figure 1-3). Due to the high scan speed resulting in a short cycle time (~ 3.4 seconds for AB Sciex 5600), the high resolution fragment signals in SWATH MS2 spectra can be viewed as high throughput MRM data. As a result, SWATH has been utilized as an alternative to MRM for targeted proteomics quantitation [23-26].

1.4 *Spectral library-dependent DIA targeted analysis*

Initially, DIA was considered as similar to high-throughput MRM and hence follows the same workflow as in MRM quantitation analysis. To quantify a target protein of interest, extensive prior information is needed to unambiguously locate its peptides in an LC-MS run. This information includes the choice of the peptides to monitor, their retention times and fragmentation patterns. The prior information is usually acquired in a separate analysis using DDA, and summarized in the form of a 'transition lists' or 'spectral library'. The fragment ion intensities for the target peptides in DIA data are then extracted from the data using a targeted extraction approach with the help of spectral library. The concept of targeted extraction was inspired by MRM analysis and has been adopted by different computational tools for DIA analysis including the commercial software PeakView and Spectronaut [27], and the open-source packages such as OpenSWATH [28] and Skyline [29].

To quantify a peptide ion in a DIA file using a spectral library, first the observed fragments, also called 'transitions', of the peptide ion are extracted from the library. Based on the observed fragment *m/z* values, the XICs of the fragments in the DIA MS2 spectra are then built for either the entire retention time range or for a smaller window around the peptide ion observed elution time (or calculated by computational algorithms) from the library. Each fragment XIC across the retention time extraction range is then split into individual peaks (each scattered at different retention time points) by peak detection algorithms, indicating potential retention time spots of the targeted peptide ion from the perspective of the single fragment. Combining all the detected peaks from all the fragment XICs, the targeted extraction tools mentioned above determine a group of co-eluting fragment peaks for the targeted peptide ion using the scoring models such as mProphet [30] and DIANA [31]. Finally, the fragment intensities from the DIA data are then extracted to quantify the peptide. Recent studies have further advanced such targeted extraction approaches to various proteomics applications [32-40] including post-translational modifications [34, 35], protein-protein interaction [35, 36], immunopeptidome [40].

1.5 *Drawbacks of spectral library-based targeted approaches*

The applications using spectral library-based DIA approach mentioned above require replicate analysis of the same samples using both DDA (to build the spectral library) and DIA (to quantify the target proteins). This essentially doubles the amount of the sample necessary for complete quantitative analysis as well as the MS analysis time. Furthermore, this strategy (as most targeted strategies) relies heavily on the precise knowledge of the peptide retention times, and thus ideally involves retention time calibration using peptide standards (e.g. iRT peptides) spiked-in each analyzed sample. The analysis of DIA data becomes further complicated if the DDA experiments used to build the spectral library were done using a different LC gradient or LC system, a different MS analyzer, or a different fragmentation method, which could lead to deviations in peptide retention times or fragmentation patterns between the DDA and DIA runs. In the analysis of complex samples such as human tissues, retention times of some peptides may vary significantly from sample to sample depending on the number of co-eluting peptides in each sample.

1.6 *Motivation*

The main motivation behind this dissertation work is to explore and develop computational strategies that would allow taking the advantage of the new generation of DIA strategies such as SWATH for untargeted protein identification and quantitation analysis without any extensive parallel DDA analysis of the same samples. As the conceptual workflow in Figure 1-3 shows the conventional identification analysis of DDA data often relies on MS/MS database search (Figure 1-3a). As mentioned, the analysis of DIA data mostly relies on targeted extraction using spectral library (Figure 1-3b). Therefore, the first aim of this dissertation work is to develop algorithms to transform DIA data into precursor-fragment group data (pseudo MS/MS spectra) which is fully compatible with DDA MS/MS database search engines (spectrum-centric search) so that one can perform untargeted identification analysis directly for DIA data (Figure 1-3c). The second aim of the dissertation is to extend the untargeted analysis to a complete DIA analysis including quantification analysis. When we consider quantification analysis in an experimental setting that includes multiple replicates / samples, retention time alignment and "internal spectral library" searching can be developed to further reduce missing quantification across multiple LC-MS experiments. Lastly, the third aim of this dissertation work is to further improve the algorithms of the pipeline and show its performance using DIA data obtained from different mass spectrometers.

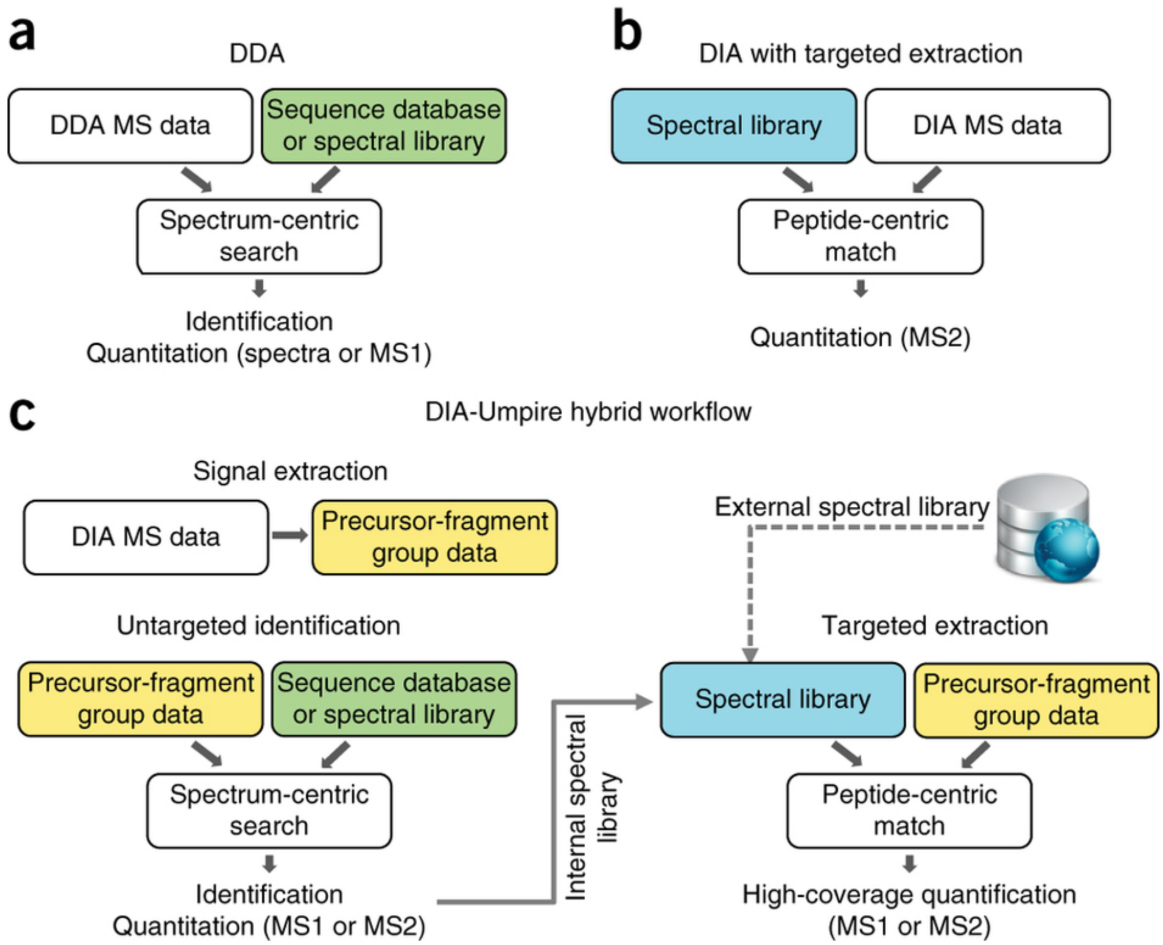

a
DDA
DDA MS data
Sequence database or spectral library
Spectrum-centric search
Identification
Quantitation (spectra or MS1)
b
DIA with targeted extraction
Spectral library
DIA MS data
Peptide-centric match
Quantitation (MS2)
c
DIA-Umpire hybrid workflow
Signal extraction
DIA MS data
Precursor-fragment group data
External spectral library
Untargeted identification
Precursor-fragment group data
Sequence database or spectral library
Spectrum-centric search
Identification
Quantitation (MS1 or MS2)
Internal spectral library
Targeted extraction
Spectral library
Precursor-fragment group data
Peptide-centric match
High-coverage quantification (MS1 or MS2)

Figure 1-3 Untargeted and targeted data analysis strategies and DIA-Umpire hybrid framework

(a) Conventional analysis of DDA data is based on matching MS/MS spectra against a proteome-wide sequence database or a spectral library (spectrum-centric search). Peptides (and then proteins) are quantified using MS1 signal intensity or spectral counts (label-free quantification) **(b)** Current methods for DIA analysis are based on targeted data extraction, in which peptide ions from a spectral library are queried against experimental data (peptide-centric search) to find the best matching fragment ion signals and their intensities (MS2 based quantification). **(c)** DIA-Umpire hybrid workflow performs signal extraction from DIA MS1 and MS2 spectra to construct precursor–fragment groups. Each precursor–fragment group is then analyzed using spectrum-centric searching to identify the peptides, as in (a). Peptide-centric matching is then performed to query unidentified precursor–fragment groups against a spectral library, as in (b). The spectral library can be built from the initial untargeted (spectrum-centric) results using the same DIA data, or can be combined (replaced) with an external spectral library built using DDA data. Quantification can be done from either MS1 precursor- or MS2 fragment-ion intensities.

Chapter 2 Untargeted proteomics identification analysis for data independent acquisition data

The content of this chapter was previously published by the author as a research article in *Nature Methods* [41].

2.1 *Background*

DIA data presents a potential for more comprehensive proteomics analysis because its unbiased acquisition. However, as mentioned in Chapter 1, the common approach for DIA analysis is dependent on spectral libraries. The aim of the first study is to develop a computational method which can enable untargeted analysis (spectral library-free) for DIA proteomics data. We develop a pipeline called DIA-Umpire that includes a series of optimized signal processing algorithms for detection of signal features from DIA MS1 and MS2 spectra. The detected MS1 and MS2 features represent all observed peptide precursor and fragment ions, respectively. All the detected features are then assembled into precursor-fragment groups. This strategy allows untargeted analysis of DIA data by means of converting the detected precursor-fragment feature groups into "pseudo" MS/MS spectra. These pseudo MS/MS spectra are fully compatible with the conventional DDA database search engines and statistical analysis tools for estimating the false discovery rates (FDR). This untargeted analysis method was first evaluated using four sets of samples of different complexity consisting of just the UPS (Universal Protein Standard) proteins, *E. coli* lysates, human cell lysates, and a public glycoproteomics dataset. We demonstrated that the algorithm can identify proteins in DIA data with similar numbers obtained from DDA data. We observed that, in our hands, DDA still outperforms DIA slightly for untargeted peptide and proteins identification in complex samples, especially in the low abundance range. We also performed a detailed comparison between the untargeted and targeted (exemplified by OpenSWATH [16]) analysis of DIA data using the complex *E. coli* and human cell lysate samples and the public glycoproteomics dataset.

2.2 *Methods*

2.2.1 *Sample preparation UPS2, E. coli, and human datasets*

Proteomics Dynamic Range Standard (UPS2) sample was acquired from Sigma-Aldrich (St. Louis, MO), the MassPREP *E. coli* Digest Standard was acquired from Waters (Milford, MA) and the MS compatible human protein extract digest was from Promega (Madison, WI). The UPS2 samples were reduced with 5 mM TCEP (tris(2-carboxyethyl)phosphine), alkylated with 50 mM iodoacetamide, and digested overnight with 1 µg trypsin (Promega, Madison, WI) in 100 mM Tris pH 8 at 37°C. UPS2, E. coli, and human peptides were acidified with formic acid and loaded at various concentrations, alone or in combination, onto an in-house made 75 µm x 12 cm analytical column emitter packed with 3 µm ReproSil-Pur C18-AQ (Dr. Maisch HPLC GmbH, Germany). A NanoLC-Ultra 1D plus (Eksigent, Dublin CA) nano-pump was used to deliver a 90 minute gradient from 2% to 35% acetonitrile with 0.1%

formic acid, followed by a 30 minute wash with 80% acetonitrile prior to re-equilibration to 2% acetonitrile with 0.1% formic acid.

2.2.2 *Mass spectrometric analysis*

Each sample was analyzed in duplicates (1 ug *E. coli* lysate, 500 ng Human lysate) or in triplicates (UPS2, UPS2 plus *E. coli*; affinity purified samples previously reported [26]) on a TripleTOF™ 5600 instrument (AB SCIEX, Concord, Ontario, Canada) once using DDA and once using DIA (SWATH) with an extended ion accumulation time of 250 ms for MS1 scans. UPS2 samples were also analyzed using SWATH with the previously-reported MS1 survey scan ion accumulation time of 50 ms [16, 26]. The DDA run consisted of one 250 ms MS1 TOF survey scan covering 400–1300 Da followed by ten data dependent 100 ms MS/MS scans (1 Da isolation window, scan range 100–2000 Da) with precursors excluded for 15 s after being selected for fragmentation once (dynamic exclusion option). The SWATH run consisted of one 250 ms or 50 ms MS1 TOF survey scan followed by 34 sequential MS2 windows of 25 Da covering a mass range of 400–1250 Da at 95 ms per each SWATH scan. The DIA run (Thermo Q Exactive Plus) consisted of one MS survey scan (17500 resolution, target 3e6, max fill time 50 ms) every 10 scans, and 24 sequential MS2 windows of 26 amu (17500 resolution, target 5e5, max fill time 80 ms) covering a mass range from 400–1000 Da. The DDA run (Thermo QE Plus) consisted of one MS survey scan (70000 resolution, target 1e6, max fill time 30 ms) followed by fifteen MS/MS scans (2 Da isolation, 17500 resolution, target 1e5, max fill time 125 ms), with former precursors excluded for 20 seconds after being selected once.

2.2.3 *Glycoproteomics SWATH dataset*

The .wiff raw files of the public glycoproteomics SWATH and DDA datasets [35] were downloaded from ProteomeXchange Consortium using the dataset identifier PXD000704.

2.2.4 *mzXML File conversion*

All the .wiff raw files from AB SCIEX 5600 TripleTOF were first converted to mzML format with the AB MS Data Converter (AB SCIEX version 1.3 beta) using "centroid" option, and the resulting mzML files were further converted into mzXML format by msconvert.exe from the ProteoWizard package (version 3.0.4462) [42] using the default parameters.

2.2.5 *Precursor and fragment ion 2D peak detection in DIA-Umpire*

A two-dimensional feature detection algorithm was developed to locate precursor and fragment ion signals in MS1 and MS2 data (Figure 2-1). Feature detection analysis starts with the LC elution profile ("peak curve") detection step. A peak curve represents a mass trace continuous in time, and a peak must be present in at least three consecutive scans (for data presented in this study, >9 second on average). It is stored as three vectors of *m/z* values MZ = (m_1, m_2, …, m_n), intensities $\mathrm{INT_{raw}}$ = (i_1, i_2, … i_n), and retention times $\mathrm{RT_{raw}}$ = (t_1, t_2, … t_n), where *n* is the number of consecutive scans and t_{i+1}>t_i. For detected features the algorithm reports *m/z* value, retention time span (elution start and end times, t_1 and t_n) and extracted ion chromatograms (XICs).

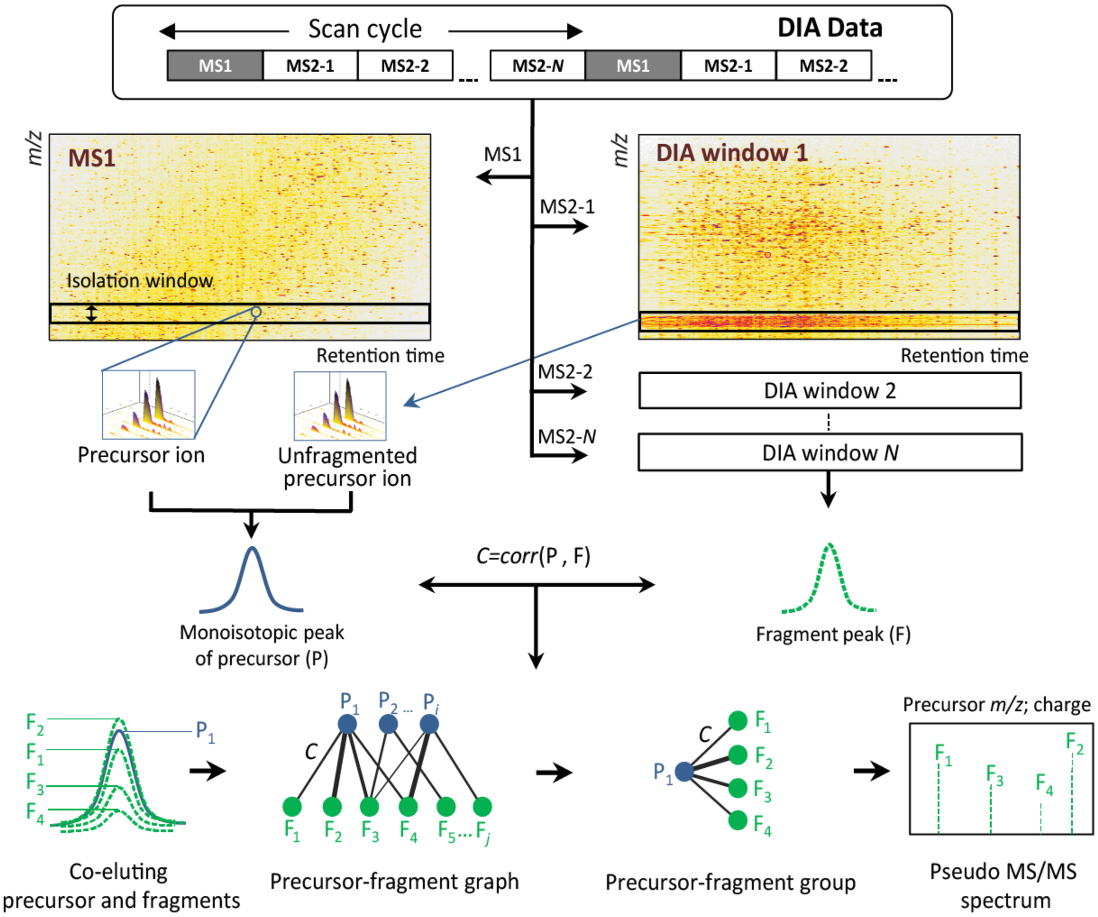


Figure 2-1 DIA-Umpire signal extraction algorithms.

The feature detection algorithm is applied to DIA MS1 and MS2 spectra to detect all possible MS1 peptide precursor ions and MS2 fragment signals. Each detected precursor feature is grouped with corresponding co-eluting fragment ion features based on Pearson correlation of LC elution peaks and retention times of peak apexes to form precursor-fragments groups. These precursor–fragment groups are used to construct pseudo MS/MS spectra (separated into different quality tiers based on the quality of detected precursor ion signal) for untargeted spectrum-centric database search and identification.

The m/z value M of a peak curve is calculated as a weighted average (by intensity) of detected m/z values in the retention time span,

$$M = \frac{\sum_{j=1}^{n} i_j m_j}{\sum_{j=1}^{n} i_j}.$$

Each peak curve is then smoothed by B-spline interpolation (using the 2nd degree basis function). XICs are represented as two vectors of interpolated retention times RT = (t_1, t_2, … t_k) and intensities INT = (i_1, i_2 , … i_k), where k is the total number of interpolated points per peak (we used 150 points per minute, making k = 150(t_n-t_1)). As a peak curve might have multiple maxima, we apply a Continuous Wavelet Transform (CWT)-based approach for splitting it into several separate peak curves using Mexican-hat wavelet (See Du et al [43] and Tautenhahn et al [44] for mathematical details of CWT) . For each unimodal peak curve, the apex intensity is determined as I_{max} = max(INT).

In MS1 data generated using high-resolution instruments, several isotope peaks for each peptide precursor ion can usually be detected (referred to as precursor ion features) helping to distinguish true precursor signals from noise. Single peak curves detected in MS1 scans are grouped together to form isotopic clusters based on RT apex distance and *m/z* spacing, which should fit the spacing for a given charge state (in this study, +2, +3, and +4 only).

In complex samples, however, the presence of multiple co-eluting peptides having similar *m/z* values results in overlapping signals, leading to multiple alternative possibilities for isotope peak grouping (see Figure 2-2 for an illustration). In such cases, the algorithm intentionally over predicts the number of precursor ion features by first considering the *m/z* of each peak curve as a possible monoisotope, and then attempting to find heavier isotope peaks for that presumed monoisotopic *m/z* value. In doing so, the algorithm maximizes the sensitivity with respect to finding true precursor ion features at the cost of introducing some redundant features with incorrectly assigned monoisotopic *m/z* values.

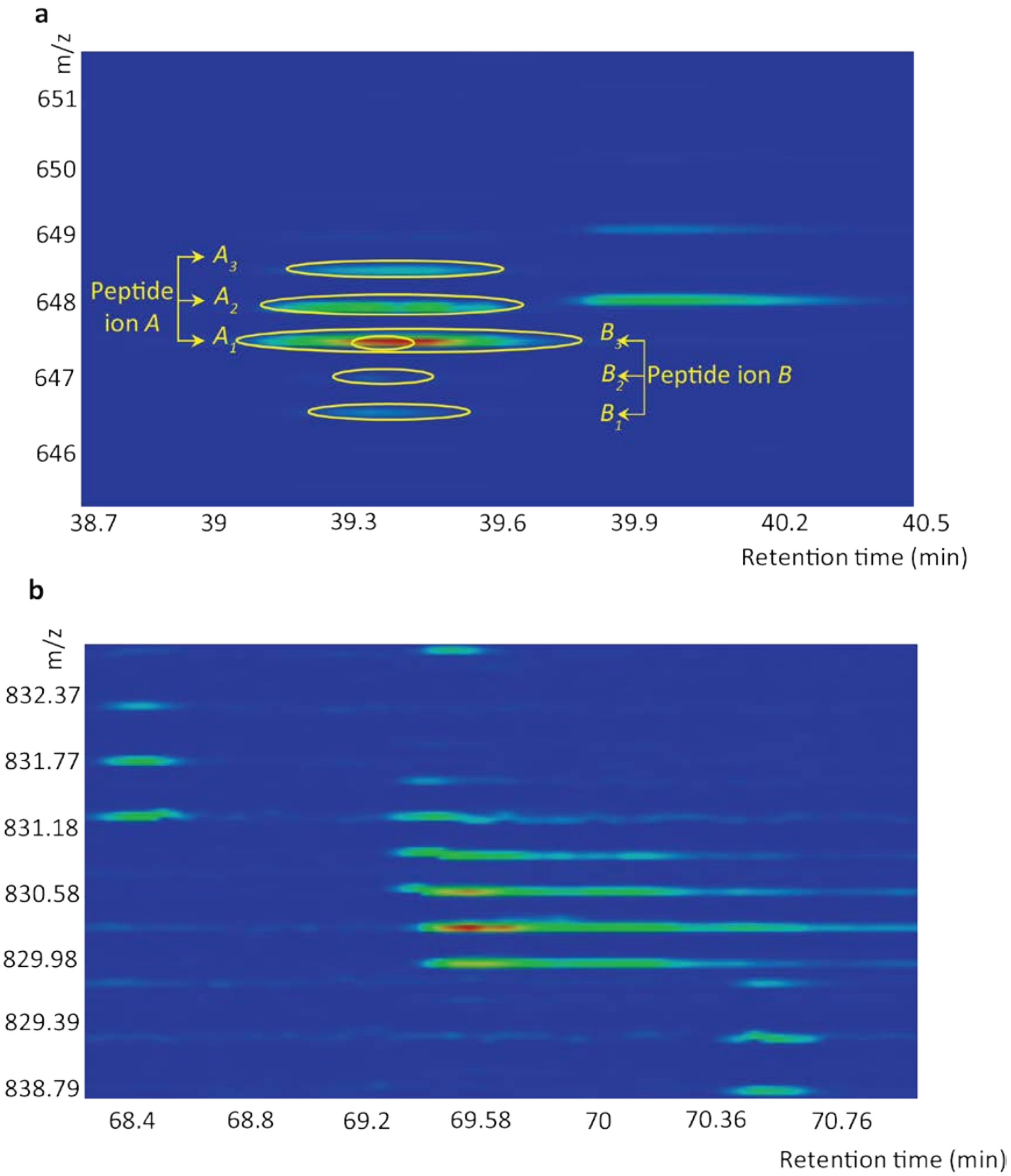


Figure 2-2 Examples of co-eluting peptide ions.

(a) Two co-eluted peptide ions *A* and *B* with the monoisotopic peak A_1 of peptide ion *A* overlapped with the third isotope peak B_3 of peptide ion *B*. The peak detection algorithms have a difficulty with detecting B_3 because it is completely buried by A_1 signal. **(b)** Another, more complicated example where co-elution of multiple peptide ions presents an ambiguity with the interpretation of different isotope peak groups. To effectively detect as many true precursor ions as possible, the signal detection algorithm of DIA-Umpire considers each peak curve as a possible monoisotopic peak, and then attempts to find higher isotope peak curves for the assumed monoisotopic peak.

In general, the higher the number of isotope peaks detected for an MS1 feature, the more likely it is to be a true precursor ion signal. Thus, the algorithm uses the number of isotope peaks as a measure of quality of precursor ion features. Features with three or more isotope peaks are labeled as Quality Tier 1 (QT = 1 or Q1) precursors, i.e. the precursors that are most likely to represent true precursor peptides with the correctly determined monoisotopic *m/z* values. MS1 features with only two detected isotope peaks are labeled as Quality Tier 2 (QT = 2 or Q2). All single peaks observed in MS1 scans (i.e. peaks with no isotopic envelope detected) are discarded.

In addition to detection of precursor ion features in MS1 scans, unfragmented precursor ions can sometimes be observed in DIA MS2 spectra. This is likely due to the collision energy not being universally suitable for complete fragmentation of all the precursor ions within a particular DIA isolation window. To take advantage of this, all peaks in MS2 spectra having *m/z* values within the corresponding DIA isolation window are considered as potential unfragmented precursors (see Figure 2-1). Unfragmented precursor ion features are detected as described above for MS1 data, requiring at least two isotope peaks. These features are added to the precursor list as Quality Tier 3 (QT = 3 or Q3). Note that some peptide precursor ions can be detected in both DIA MS1 and MS2 spectra, and their corresponding features thus may be included in both Quality Tier 3 and Quality Tier 1 (or 2) sets.

Fragment ion peak detection in MS2 data is performed similarly, with one modification. It is generally more difficult to detect multiple isotope peaks for low intensity fragment ions. Relaxed stringency of feature detection for fragment ions (compared to MS1 precursor ions feature detection described above) resulted in improved sensitivity of peptide identification and reduced the computational time. Thus, isotope peak grouping and charge state determination for fragment ions is not performed at this stage. Instead, each possible fragment peak is treated independently, and isotope detection and charge state determination is performed at a later stage (after the precursor–fragment grouping step described below).

2.2.6 *Precursor-fragment grouping*

"Co-elution" is an important characteristic of the data that reveals relationships between a precursor ion and its fragments [17]. The algorithm takes advantage of this characteristic by calculating the Pearson correlation coefficient and the retention time difference of LC elution peak apexes between all detected precursors (P) and all possible fragment ions (F) (see Figure 2-1). This pairing is naturally restricted to fragment ions in the DIA isolation window corresponding to the *m/z* value of the precursor. For a precursor P_q and a fragment F_r the Pearson correlation coefficient $C_{q,r}$ = corr(P_q, F_r) is computed using the LC profiles (XICs) of monoisotopic precursor and fragment ion features. All precursor-fragment pairs are represented as a bipartite graph (see Figure 2-1). In this representation, one fragment ion can have multiple precursors and several precursors can share the same fragment.

To better connect precursor ions to their most likely fragment ions, the following parameters are calculated based on the correlation scores for each possible P_q, F_r pair. First, given a fragment ion F_r, RP(P_q, F_r) score is calculated as the rank of the precursor ion P_q based on Pearson correlation $C_{q,r}$ between that fragment and all candidate precursors. Second, given a precursor ion P_q, RF(P_q, F_r) score is calculated as the rank of the fragment F_r based on Pearson correlation between that precursor and all possible fragments. For a precursor ion with many co-eluting fragments, a higher-ranking fragment is more likely to be derived from it. Similarly, for a fragment ion, a higher-ranking precursor ion is more likely to be its true precursor. These two metrics, as well as the retention time difference of LC profile apexes, ΔT(P_q, F_r), are used to assemble precursor-fragment groups (see Figure 2-1).

2.2.7 *Generation of pseudo MS/MS spectra*

To generate a pseudo MS/MS spectrum for a precursor ion P_q, the algorithm first detects the charge state of each fragment peak (if only a single isotope peak is detected, charge state +1 is assumed). It then detects all likely complementary *y*- and *b*-ions in the spectrum (detected as pairs of fragments summing up to the precursor peptide mass [45]). For non-complementary ion peaks, only those fragments F_r are kept that pass the following set of thresholds: $RF(P_q, F_r) \leq RF_{max}$, $RP(P_q, F_r) \leq RP_{max}$, and $\Delta T(P_q, F_r) \leq \Delta T_{max}$. These threshold parameters are implemented as user-specified options in the software, allowing re-evaluation and adjustment of the default thresholds (described below), if necessary.

Charge state and precursor *m/z* for each pseudo MS/MS spectrum are determined by precursor ion features. Fragment ion intensities are computed in three steps. For fragment F_r, the intensity is taken as LC apex intensity of the corresponding elution peak curve, I_r. Then for each complementary *b*-, *y*- fragment pair F_{r1}, F_{r2}, the intensity of the less intense fragment is boosted to match that of the more intense one, $I_{r1} = I_{r2} = \max(I_{r1}, I_{r2})$. At the last step, intensities are adjusted by weighting according to the square of correlation with the precursor peak curve, $I_r' = I_r \times C^2_{q,r}$. The presence of complementary ions is a positive sign of a connection between the precursor and fragment ions, and boosting the intensities of complementary ions has been shown to improve the sensitivity of peptide identification [46]. Note that this fragment intensity adjustment step can optionally be skipped for other applications, e.g. to use a spectral library search engine for searching pseudo MS/MS spectra or to build a spectral library from the pseudo MS/MS spectra. Also note that the adjusted (boosted) intensities are not used for quantitation, only for identification. An example of a pseudo MS/MS spectrum (before and after complementary ion boosting), the underlying precursor ion and fragment ion elution profiles in DIA MS1 and MS2 data, and the DDA MS/MS spectrum for the same peptide are shown in Figure 2-4.

The performance of the DIA-Umpire algorithm for different combinations of the threshold parameters described above was evaluated using a subset of the data. The results are shown in 2.3 and Table 2-1. When the pseudo MS/MS spectra extracted under different settings were searched using X! Tandem, the following threshold values resulted in the highest number of identifications (at 1% FDR) and were selected as default values in the software: allow the top 25 ranked precursors for each fragment (RPmax= 25), the top 300 ranked fragments for each precursor (RFmax = 300) and 0.6 minutes apex elution time difference (ΔTmax = 0.6). Note that the best performance was achieved by allowing the possibility of an MS2 fragment to be included in multiple MS/MS spectra (RPmax= 25). Because the algorithm takes the square of a peak shape correlation coefficient between the precursor and fragment signals as the weighting factors for calculation of adjusted fragment intensities in pseudo MS/MS spectra, true high intensity fragments can still contribute to the identification of their corresponding peptide even if they have a relatively poor correlation with the precursor (e.g., due to ion suppression effects affecting either the precursor ion or the fragment ion elution peak shape). The overall robustness of the pseudo MS/MS spectrum generation process was also evident from similar numbers of peptide ion identifications obtained by searching the spectra with three different database search engines (X! Tandem, Comet, and MSGF+, detailed results are shown in Table 2-2.

These results indicate that inclusion of more fragment ions in a pseudo MS/MS spectrum does not hamper the identification rate. On the contrary, by doing so the algorithm increases the chance of true fragments to be included, thus improving the number of confident identifications. An additional analysis was also carried out for *E. coli* and human cell lysate datasets by removing fragments from pseudo MS/MS spectra that were also matched in other pseudo MS/MS spectra identified with high confidence. Repeating X! Tandem search with those fragments removed did not change the number of identified peptide ions in either dataset.

2.2.8 *Peptide and protein identification using pseudo MS/MS spectra*

In this study, we used X! Tandem [47], Comet [48], and MSGF+ [49] as search engines to identify peptides from pseudo MS/MS spectra (however, any database search engine developed for searching DDA spectra can be used). Because of the similar characteristics of DDA and DIA pseudo MS/MS spectra, all downstream analysis of the database search results, including protein inference and estimation of posterior probabilities of correct identification and FDR, can also be performed using conventional strategies developed for DDA data. Database search output files were processed by PeptideProphet [50] via the Trans-Proteomic Pipeline (TPP) [51], followed by ProteinProphet [2] analysis to assemble peptides into proteins/protein groups and to determine protein probabilities. The final protein and peptide identification lists were filtered to achieve a desired FDR (here – 1%) estimated using the target-decoy approach [1]. The only modification was to compute posterior peptide probabilities by PeptideProphet separately for each of the three quality categories of MS/MS spectra (Quality Tiers QT = 1, 2 or 3) because of very different ratios of correct vs. incorrect identifications among them (see Figure 2-5).

Further analysis of the model parameters and the distributions of scores reported by PeptideProphet (see Figure 2-5) did not show any evidence indicating that pseudo MS/MS spectra extracted using DIA-Umpire behaved any different than conventional DDA spectra with respect to the basic assumptions in PeptideProphet or the target-decoy FDR estimation strategy.

2.2.9 *Peptide and protein identification parameters*

For UPS2, *E. coli*, and human cell lysate datasets, DDA MS/MS spectra and the DIA pseudo MS/MS spectra were searched by X! Tandem, Comet, and MSGF+ using the following parameters: allow tryptic peptides only, up to one missed cleavage, oxidation of methionine and cysteine alkylation as variable modifications. The glycoproteomics SWATH dataset was searched by X! Tandem only, with cysteine alkylation specified as a fixed modification and with deamidation of asparagine as a variable modification. The instrument-specific parameters – the precursor ion mass tolerance and the fragment ion mass tolerance – were set to 30 ppm and 40 ppm for AB SCIEX 5600 TripleTOF, respectively. In X! Tandem, the analysis was limited to 140 most intense peaks which gave the best results based on the same subset of the data that was used to select the parameters for the DIA-Umpire pseudo MS/MS extraction algorithm (see above). However, the search results were not very sensitive to the choice of this parameter (which is also evident from the fact that similar results were obtained using Comet and MSGF+ search tools that do not provide an option to restrict the number of peaks in the spectra). The sequence database for the UPS2 experiment was compiled from the UPS sequences (total 50 sequences: 48 UPS1 proteins and 48 UPS2 proteins, www.sigmaaldrich.com). For the *E. coli* experiments, *E. coli* proteome sequences (4,431 proteins) were extracted from UniProtKB. The non-redundant human protein sequence FASTA file from the

UniProt/SwissProt database (release of 09-Jan-2013), appended with common contaminant proteins, was used for the human cell lysate experiment and the glycoproteomics datasets. For all sequence databases, reversed sequences were added as decoys for target-decoy analysis. The initial search results from the search engines were first converted into pepXML format, followed by analysis using PeptideProphet [50] via the Trans-Proteomic Pipeline (TPP) [51] (v4.7). For DIA derived pseudo MS/MS spectra, PeptideProphet was run separately for each of the three quality categories of MS/MS spectra (Quality Tiers QT = 1, 2 or 3). The iProphet [52] tool was used when merging the search results from all three search engines. Unless noted otherwise, peptide ion identification lists for each DDA or DIA run were filtered at 1% FDR, estimated by target-decoy approach based on PeptideProphet probability for each search engine (or iProphet peptide ion probability when using iProphet).

Protein inference for different analyses was performed as follows. To report the numbers of protein identifications for individual DIA/DDA runs (Table 2-2), PeptideProphet output files (individual search engine analysis) or iProphet output files (when combining the search results) were analyzed by ProteinProphet [2] for protein inference. For the comparison between DDA and DIA or between DIA-Umpire and OpenSWATH results at the protein level, PeptideProphet output files (based on X! Tandem results) for both DIA and DDA were processed together by ProteinProphet. The final protein lists for each ProteinProphet analysis were determined by a 1% FDR threshold, estimated by target–decoy approach.

2.2.10 *Targeted extraction analysis using OpenSWATH*

The *E. coli* and human cell lysate experiments from AB SCIEX 5600 TripleTOF were also processed with OpenSWATH to identify proteins and peptides using the fully targeted approach. The two DDA replicates acquired for each sample were used to build the spectral library using SpectraST [53] with the following options: best replicate; union; 0 minimum peaks for exclusion; 0 minimum amino acids for exclusion. Only the DDA non-decoy identification spectra that passed 1% FDR threshold were used for building the library. The probability thresholds were: 0.6979 for DDA *E. coli* replicate 1; 0.7877 for DDA *E. coli* replicate 2; 0.8075 for DDA human replicate 1; 0.8233 for DDA human replicate 2. This resulted in a total of 12,820 and 17,402 peptide ions including decoys represented in the “transition lists” used by OpenSWATH for *E. coli* and human, respectively. For OpenSWATH analysis using DIA-derived libraries, the libraries were built with SpectraST using the pseudo MS/MS spectra (without complementary *b*- and *y*-ion boosting) from peptide ions identified by the DIA-Umpire’s untargeted workflow and filtered at an 1% FDR threshold (8,757 peptide ions for human and 6,364 for *E. coli* samples).

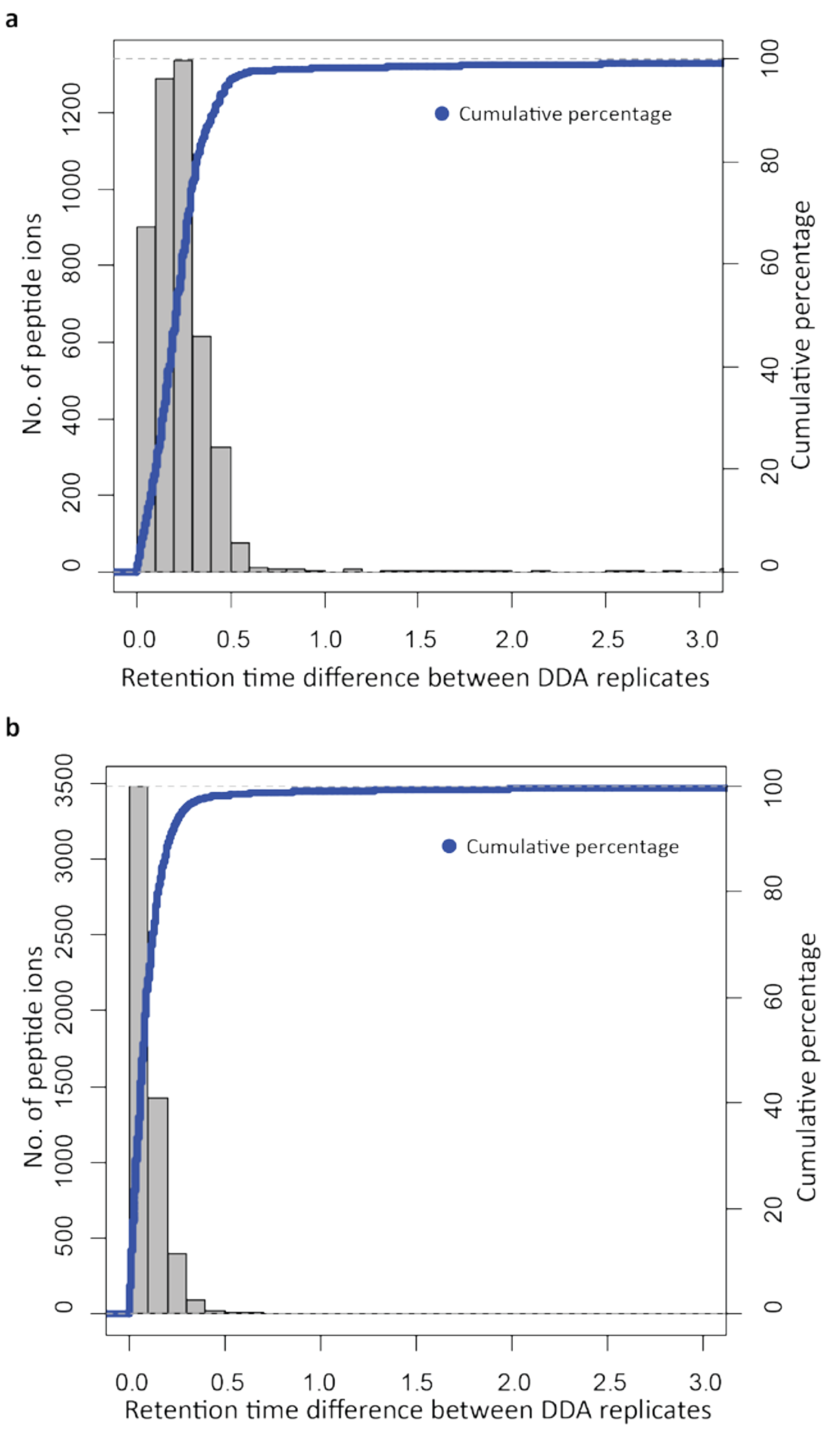


Figure 2-3 Retention time differences for peptide ions commonly identified between DDA replicates.

(a) *E. coli* cell lysate data **(b)** human cell lysate data.

OpenSWATH was run using the following parameters: extraction elution time window (seconds): 60; minimum transitions: 2; maximum transitions: 6; unique ion signature threshold: -1; retention time normalization factor (seconds): 7200 (i.e. the whole LC-MS run duration in our case). Our dataset did not contain iRT [54] peptides for retention time normalization because all the experiments were performed using the same instrumentation setup and the retention times were highly reproducible (within one minute) between the DDA/DIA runs (Figure 2-3). Peptide ion identification lists were filtered using mProphet [30] at 1% FDR. The number of candidate peptide ions used for scoring against the extracted peak groups in OpenSWATH analysis was estimated as the number of ions in the DDA-derived library falling within the corresponding 25 Da SWATH isolation window and within the specified retention time tolerance (1 minute).

2.2.11 *DIA-Umpire analysis using reduced database*

In order to demonstrate how search space affects peptide identifications, in addition to searching DIA pseudo MS/MS spectra against the proteome-wide sequence database (all *E. coli* or human proteome sequences plus decoys), we also used a smaller database of peptide sequences (5,997 and 8,784 sequences for *E. coli* and human cell lysate experiments, respectively) identified from the corresponding DDA data. Reverse versions of these sequences were also appended to the database for target-decoy analysis. All other search parameters and settings were the same as described above.

2.2.12 *Isotopic pattern validation of glycopeptide identifications*

Identification of *N*-linked glycopeptides relies on detection of asparagine deamidation due to PNGase F treatment which causes a small mass shift (0.984 Da). The mass shift is close to the mass difference between the isotope peaks which could lead to false identification of a peptide as deamidated if an "M+1" isotope peak is mis-recognized as a true monoisotopic peak. In another scenario, if there is a noise signal at "M-1" Da of a deamidated ion that is mis-recognized as the monoisotopic peak, the deamidated peptide would be mis-identified as an unmodified peptide ion. To remove these erroneous identifications, we applied a two-step filtering strategy. All confident identifications from DIA-Umpire were first grouped if their precursor features shared an isotope peak at same retention time (see Figure 2-2 for one such example). We then removed grouped precursor features if the observed MS1 isotope peak distribution did not fit the theoretical isotope pattern (chi-squared goodness of fit probability < 0.8). This first stage filtering was able to remove misidentifications in the second scenario. To remove the cases in the first scenario, the precursor masses of peptides identified in each group were compared, and only the identification with the smallest mass in the group was kept.

2.2.13 *Code and data availability*

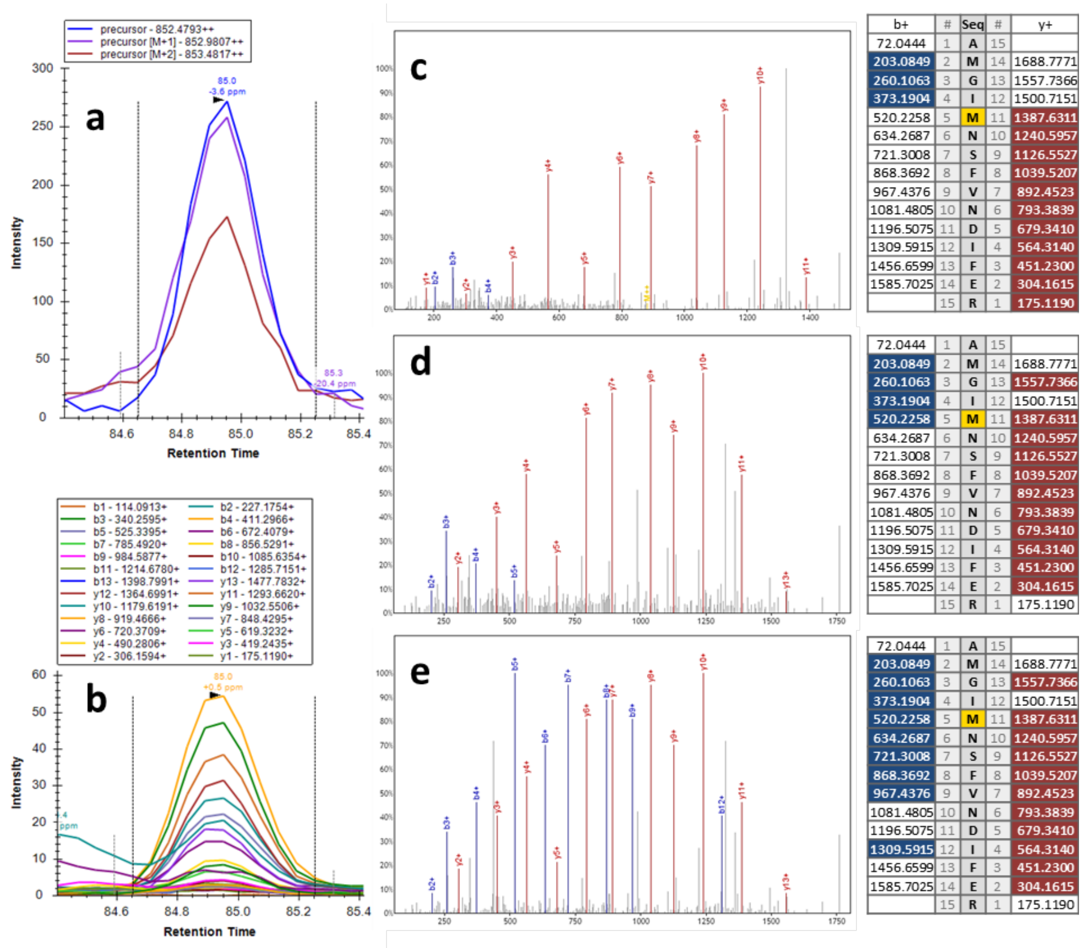
The program was developed in the cross-platform Java programming language (v1.7) and the executable files along with source codes are publically available at *http://diaumpire.sourceforge.net/*. All the spectrum files (Table A-1) along with DIA-Umpire results presented in this study have been deposited at the

ProteomeXchange Consortium [55] (*http://proteomecentral.proteomexchange.org*) via the PRIDE partner repository with the dataset identifier PXD001587.

2.3 *Results*

2.3.1 *DIA-Umpire untargeted identification workflow*

DIA-Umpire incorporates a number of computational algorithms for DIA analysis (see 2.2 for detail). It begins with a two dimensional (*m/z* - retention time) feature detection algorithm that discovers all possible precursor and fragment ion signals in DIA MS1 and MS2 data, respectively, and also possible unfragmented precursor ions in the MS2 data (Figure 2-1). Because DIA usually employs wider isolation *m/z* range (e.g. 25 Da) than DDA, co-eluting peptides are more frequently co-fragmented, generating complex MS2 spectra. In order to measure the likelihood that a detected fragment signal is derived from a particular precursor peptide ion, the algorithm calculates the Pearson correlation coefficient of LC elution peaks and retention time differences of LC elution peak apexes between all detected precursor features and all co-eluting fragment ions. Reflecting the complex nature of precursor-fragment relationships, all precursor-fragment pairs are represented as a bipartite graph (Figure 2-1). After filtering by a combination of thresholds, sets of fragment peaks are grouped with precursor features and stored as precursor-fragment groups (Figure 2-1).



b+	#	Seq	#	y+
72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

72.0444	1	A	15	
203.0849	2	M	14	1688.7771
260.1063	3	G	13	1557.7366
373.1904	4	I	12	1500.7151
520.2258	5	M	11	1387.6311
634.2687	6	N	10	1240.5957
721.3008	7	S	9	1126.5527
868.3692	8	F	8	1039.5207
967.4376	9	V	7	892.4523
1081.4805	10	N	6	793.3839
1196.5075	11	D	5	679.3410
1309.5915	12	I	4	564.3140
1456.6599	13	F	3	451.2300
1585.7025	14	E	2	304.1615
	15	R	1	175.1190

Figure 2-4 Example of precursor peptide ion and fragment ion LC elution signals and the corresponding pseudo MS/MS spectrum generated by DIA-Umpire.

(a) Elution profiles for the first 3 isotope peaks of a doubly charged precursor peptide ion AMGIM[Oxy]NSFVNDIFER extracted from MS1 data from a DIA (SWATH) run on a AB SCIEX 5600 instrument. **(b)** Elution profiles for fragments of this precursor peptide detected in the DIA MS2 data. **(c)** DDA MS/MS spectrum (from a DDA run generated on the same instrument and using the same sample) from which the same peptide was identified, with matched b- and y- ions highlighted. **(d)** Pseudo MS/MS spectrum extracted by DIA-Umpire from the DIA data (before complementary ion boosting). **(e)** Same pseudo MS/MS spectrum after complementary ion boosting. Note a larger number of b- ions matched in (e) compared to (d). (a) and (b) images exported from Skyline. (c), (d), and (e) are exported from TPP spectrum browser.

For direct untargeted analysis, DIA-Umpire generates a pseudo MS/MS spectrum (Figure 2-4) for each precursor-fragment group. The pseudo MS/MS spectra can be searched by any conventional DDA MS/MS database search engine. Here we used X! Tandem [47] Comet [48], and MSGF+ [49] followed by PeptideProphet [50] or iProphet [52] and ProteinProphet [2] analysis. The resulting peptide and protein identification lists were filtered using computed peptide and protein probabilities controlling the false discovery rate (FDR) via, e.g., the target-decoy approach [1].

2.3.2 *Analysis of experimental and computational parameters used in DIA-Umpire*

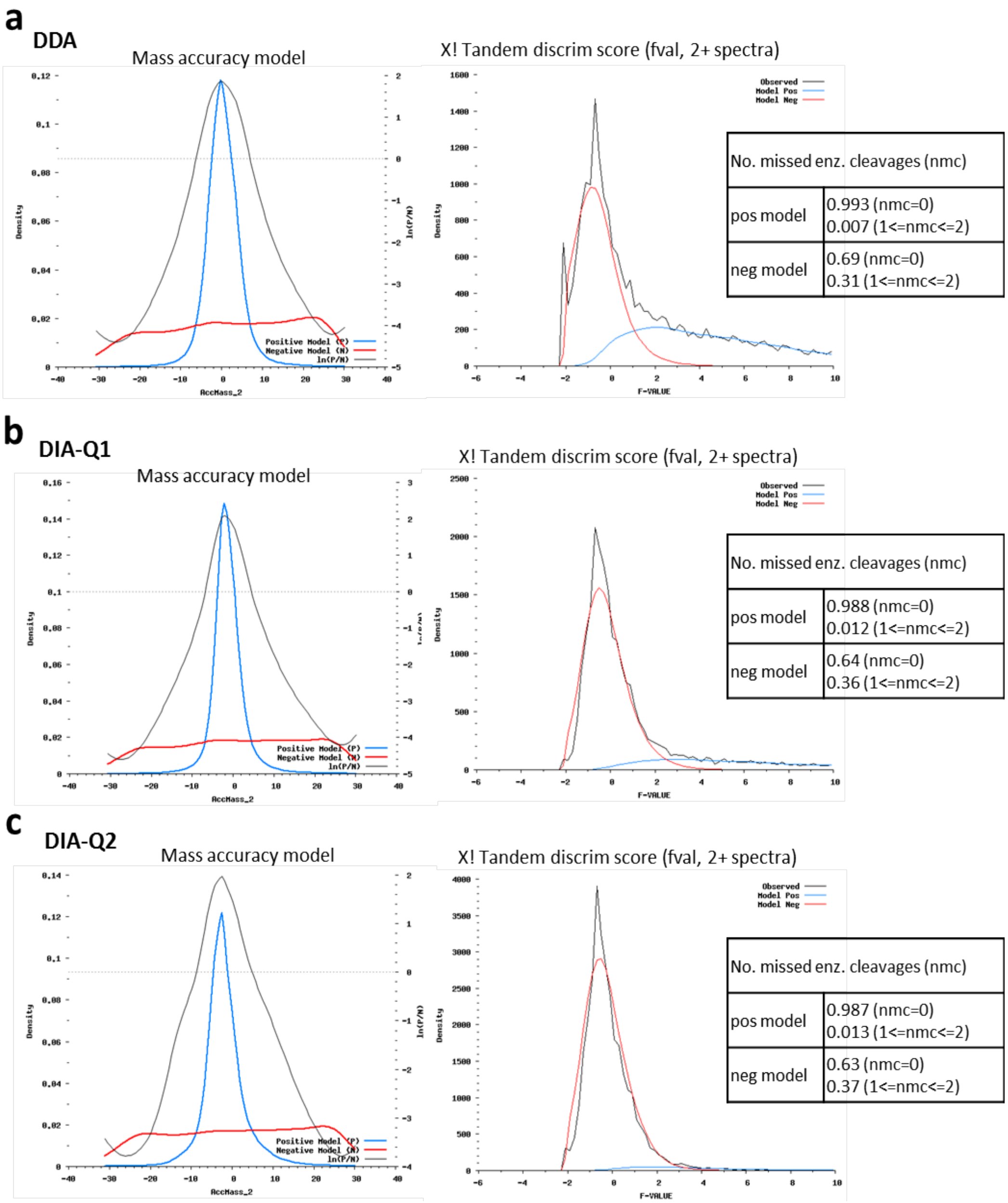
We first performed the analysis to understand how different parameters used in the precursor-fragment group step affects the results of untargeted peptide ion identifications. Number of identified peptide ions and proteins for a representative SWATH run (250 ms MS1 accumulation time, *E. coli* cell lysate mixed with UPS2 proteins sample) identified using different thresholds for precursor-fragment grouping. The search was done by X! Tandem search engine, and only DIA pseudo-MS/MS spectra from Q1 set were used. The numbers of peptide ions and protein identifications were filtered by an 1 % FDR threshold. The red-highlighted row is the parameter set chosen for this study. RPmax: maximum number of top precursors for a fragment ion considered for precursor-fragment grouping. Ranking is based on Pearson correlation between that fragment and all candidate precursors. RFmax: maximum number of fragments for a precursor feature considered for precursor-fragment grouping. ΔTmax: maximum retention time difference of LC peak apexes for a fragment to be considered as a precursor's fragment in precursor-fragment grouping. The results are shown in Table 2-1. It was clear that the parameters did not influence the numbers hugely, and generally including more fragments increased the number of identifications even many of low correlated fragments are noise peaks. The results indicated that the MS/MS database search engine is robust enough against the noise peaks. According to the results, we selected the following value as default setting in the software: allow the top 25 ranked precursors for each fragment (RPmax= 25), the top 300 ranked fragments for each precursor (RFmax = 300) and 0.6 minutes apex elution time difference (ΔTmax = 0.6).

Table 2-1 Analysis of precursor-fragment grouping parameters

RP_{max}	RF_{max}	ΔT_{max}	No. of identified peptide ions	No. of identified proteins
5	100	0.4	2088	391
5	100	0.5	2086	387
5	100	0.6	2089	388
5	300	0.4	2292	423
5	300	0.5	2282	421
5	300	0.6	2296	423
10	100	0.4	2215	407
10	100	0.5	2214	406
10	100	0.6	2216	407
10	300	0.4	2526	459
10	300	0.5	2542	462
10	300	0.6	2531	456
15	100	0.4	2236	391
15	100	0.5	2209	404
15	100	0.6	2209	393
15	300	0.4	2539	458
15	300	0.5	2548	455
15	300	0.6	2566	462
20	100	0.4	2239	391
20	100	0.5	2213	402

20	100	0.6	2213	392
20	300	0.4	2562	464
20	300	0.5	2563	466
20	300	0.6	2559	464
25	100	0.4	2243	391
25	100	0.5	2214	404
25	100	0.6	2214	393
25	300	0.4	2570	469
25	300	0.5	2571	469
25	300	0.6	2573	470
30	100	0.4	2244	391
30	100	0.5	2214	404
30	100	0.6	2214	393
30	300	0.4	2557	467
30	300	0.5	2570	468
30	300	0.6	2566	465

We next analyzed whether the score results obtained between DDA and DIA are different. Figure 2-5 shows the model distributions learned by PeptideProphet in the analysis of X! Tandem search results from identifications of doubly charged peptide ions for one replicate of the human cell lysate data. The learned distributions appear to be an accurate fit in both DIA and DDA data, demonstrating that the search results obtained using DIA pseudo MS/MS spectra can be satisfactory analyzed using PeptideProphet. The overall higher ratio of incorrect vs correct identification in the DIA QT=1 vs. DDA data (and similarly in DIA QT=2 vs. QT=1 data) simply reflects the higher number of pseudo MS/MS spectra extracted from the data compared to DDA data (and similarly, more noise in DIA QT=2 vs QT=1 data), which does not affect the accuracy of computed PeptideProphet probabilities or the subsequent FDR estimates for DIA data.



a

No. missed enz. cleavages (nmc)	
pos model	0.993 (nmc=0) 0.007 (1<=nmc<=2)
neg model	0.69 (nmc=0) 0.31 (1<=nmc<=2)

b

No. missed enz. cleavages (nmc)	
pos model	0.988 (nmc=0) 0.012 (1<=nmc<=2)
neg model	0.64 (nmc=0) 0.36 (1<=nmc<=2)

c

No. missed enz. cleavages (nmc)	
pos model	0.987 (nmc=0) 0.013 (1<=nmc<=2)
neg model	0.63 (nmc=0) 0.37 (1<=nmc<=2)

Figure 2-5 PeptideProphet analysis of X! Tandem search results using DDA and DIA pseudo MS/MS data

The figures shown here are model distributions learned by PeptideProphet in the analysis of X! Tandem search results (doubly charged peptide ions) for one replicate of the human cell lysate data. Left panels: mass accuracy distributions. Right panels: the distributions of the discriminant database search scores (computed from the X! Tandem Expect scores). Red and blue curves represent the models learned by PeptideProphet for correct and incorrect identifications, respectively. Also shown are the distributions for the number of missed cleavages parameter (nmc) between correct and incorrect identifications. **(a)** DDA data; **(b)** DIA data, QT=1 pseudo MS/MS spectra; **(c)** DIA data, QT=2 spectra.

DIA-Umpire relies on precursor signal features to be detected to generate corresponding pseudo MS/MS spectra. Therefore it is important to understand how MS1 signal quality would affect the performance of DIA-Umpire's untargeted identification. In addition to the standard MS1 accumulation time (50 ms) proposed by the original SWATH method [16], we also conducted experiment with 250 ms MS1 accumulation time to see if it improves DIA-Umpire's performance. The results are shown in Figure 2-6. The numbers shown are non-redundant contributions to the total number of peptide ion identifications in each replicate / condition from pseudo MS/MS spectra from three different quality tiers: QT = 1 (white bar), 2 (grey), and 3 (dark grey). The QT = 1 category represents pseudo MS/MS spectra that are linked to high quality MS1 precursor features (3 or more detected isotope peaks), QT = 2 represent lower abundance precursors (2 detected isotope peaks only), and QT = 3 represents unfragmented precursors which were detected in DIA MS2 scans.

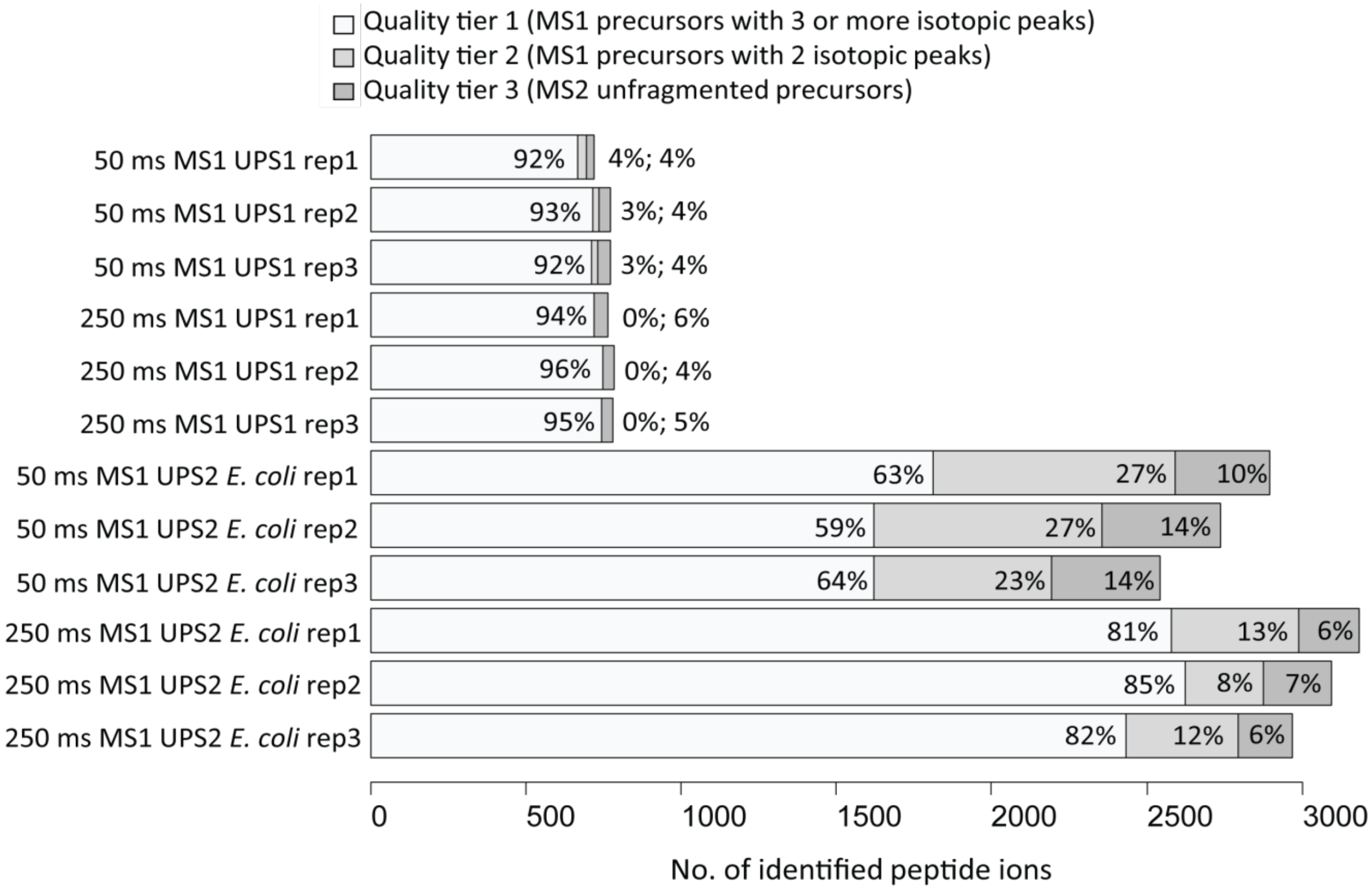


Figure 2-6 Effect of MS1 survey scan ion accumulation time on peptide identification using DIA-Umpire.

Experiments to assess the identification performance of DIA-Umpire on data generated using different MS1 ion accumulation times in DIA (SWATH) analysis using AB SCIEX 5600 instrument were carried out using two samples: UPS1 proteins, and UPS2 mixture spiked in with E. coli background. Two settings (50 ms and 250 ms MS1 ion accumulation times) were tested.

In a low complexity UPS1 sample, the dominant majority of peptide ions were identified from QT =1 spectra. Even with the using short MS1 accumulation time (50 ms), 92–93% of the peptides ions were identified from the QT = 1 spectral subset (this fraction increased slightly, to 94–96%, with the longer 250 ms accumulation time). Note that inclusion of unfragmented precursors detected in DIA MS2 data (QT = 3 subset) in the analysis contributed 4–6% of the total peptide ion identifications in UPS1 samples. In the more complex UPS2 plus *E. coli* samples, the effect of the accumulation time on the quality of MS1 signal was more pronounced. The longer DIA MS1 survey scan ion accumulation time resulted in more high quality (QT = 1) precursor peptide features detected, and thus more peptides identified from pseudo MS/MS spectra in the QT = 1 subset (81–85% for 250 ms vs. 59–64 % for 50 ms). Congruently, QT = 2 and QT = 3 spectral subsets contributed higher percentages to the total number of peptide ion identifications when using 50 ms accumulation time setting. The overall number of identifications (from all 3 QT sets) has improved with 250 ms vs. 50 ms acquisition time (~10%). Overall, this analysis indicates that longer MS1 accumulation time provides an advantage to DIA-Umpire algorithm with respect to the total number of identified peptide ions, especially peptide ions identified with a high quality MS1 precursor ion signals.

2.3.3 *Untargeted protein identification using DIA-Umpire*

We first evaluated the performance of DIA-Umpire for untargeted protein identification using samples ranging from low complexity (48 Universal Protein Standard (UPS) proteins) to high complexity (*E. coli* and human cell lysates) by performing parallel DDA and DIA runs in at least duplicates on an AB SCIEX TripleTOF 5600. Based on the results shown above, we acquired DIA data using 250 ms ion accumulation time for MS1 survey scans instead of the 50 ms SWATH setting used in earlier reports [16], which improved the MS1 signal quality and detectability of precursor ion signals in complex samples (See results in Figure 2-6). Three replicate runs were acquired for each sample / condition.

Table 2-2 Numbers of protein and peptide ion identifications from DDA MS/MS and DIA pseudo MS/MS spectra

The number of identifications obtained using three different search engines (X!Tandem, Comet, and MSGF+), as well as using all three search engines combined and analyzed by iProphet.

Sample	Acquisition	Replicate	File	Search engine	No. of proteins	No. peptide ions
UPS2	DDA	1	18185_REP2_4pmol_UPS2_IDA_1	X!Tandem	29	760
				Comet	26	330
				MSGF+	31	559
				Combined	30	858
		2	18187_REP2_4pmol_UPS2_IDA_2	X!Tandem	29	756
				Comet	26	333
				MSGF+	31	543
				Combined	30	835
	DIA	1	18186_REP2_4pmol_UPS2_SWATH_1	X!Tandem	30	822
				Comet	26	365
				MSGF+	32	737
				Combined	32	1220
		2	18188_REP2_4pmol_UPS2_SWATH_2	X!Tandem	31	794
				Comet	26	376
				MSGF+	33	665
				Combined	34	1527
E. coli	DDA	1	18483_REP3_1ug_Ecoli_NewStock2_IDA_1	X!Tandem	924	5821
				Comet	975	5564
				MSGF+	1023	5858
				Combined	1025	6532

		2	18485_REP3_1ug_Ecoli_NewStock2_IDA_2	X!Tandem	935	5460
				Comet	973	5522
				MSGF+	994	5774
				Combined	1020	6463
	DIA	1	18484_REP3_1ug_Ecoli_NewStock2_SWATH_1	X!Tandem	748	5313
				Comet	833	5198
				MSGF+	849	5235
				Combined	928	6903
		2	18486_REP3_1ug_Ecoli_NewStock2_SWATH_2	X!Tandem	774	5455
				Comet	855	5217
				MSGF+	857	5374
				Combined	928	6982
Human	DDA	1	18299_REP2_500ng_HumanLysate_IDA_1	X!Tandem	1450	7499
				Comet	1476	7512
				MSGF+	1423	7230
				Combined	1586	9364
		2	18301_REP2_500ng_HumanLysate_IDA_2	X!Tandem	1430	7089
				Comet	1509	7348
				MSGF+	1461	7099
				Combined	1578	9063
	DIA	1	18300_REP2_500ng_HumanLysate_SWATH_1	X!Tandem	1245	6934
				Comet	1346	7438
				MSGF+	1242	6945
				Combined	1448	9430
		2	18302_REP2_500ng_HumanLysate_SWATH_2	X!Tandem	1340	7344
				Comet	1269	7033
				MSGF+	1277	7158
				Combined	1457	9813

Database search results were processed using PeptideProphet and ProteinProphet (and iProphet when combining multiple search engines) and filtered to achieve 1% FDR at peptide ion or protein level.

We identified close numbers of peptide ions and proteins in the DDA and DIA runs for all samples and search engines tested (Figure 2-7a; Table 2-2). As shown in Table 2-2 for DDA data, combining DIA pseudo MS/MS search results from multiple search engines with iProphet [52] led to a consistent increase in the number of peptide ions and proteins identified at a given FDR (Table 2-2). However, for the sake of clarity and because single search engine analyses are still prevalent in the field, the remainder of the study, unless noted otherwise, is based on peptide and protein identifications using X! Tandem only (Figure 2-7a).

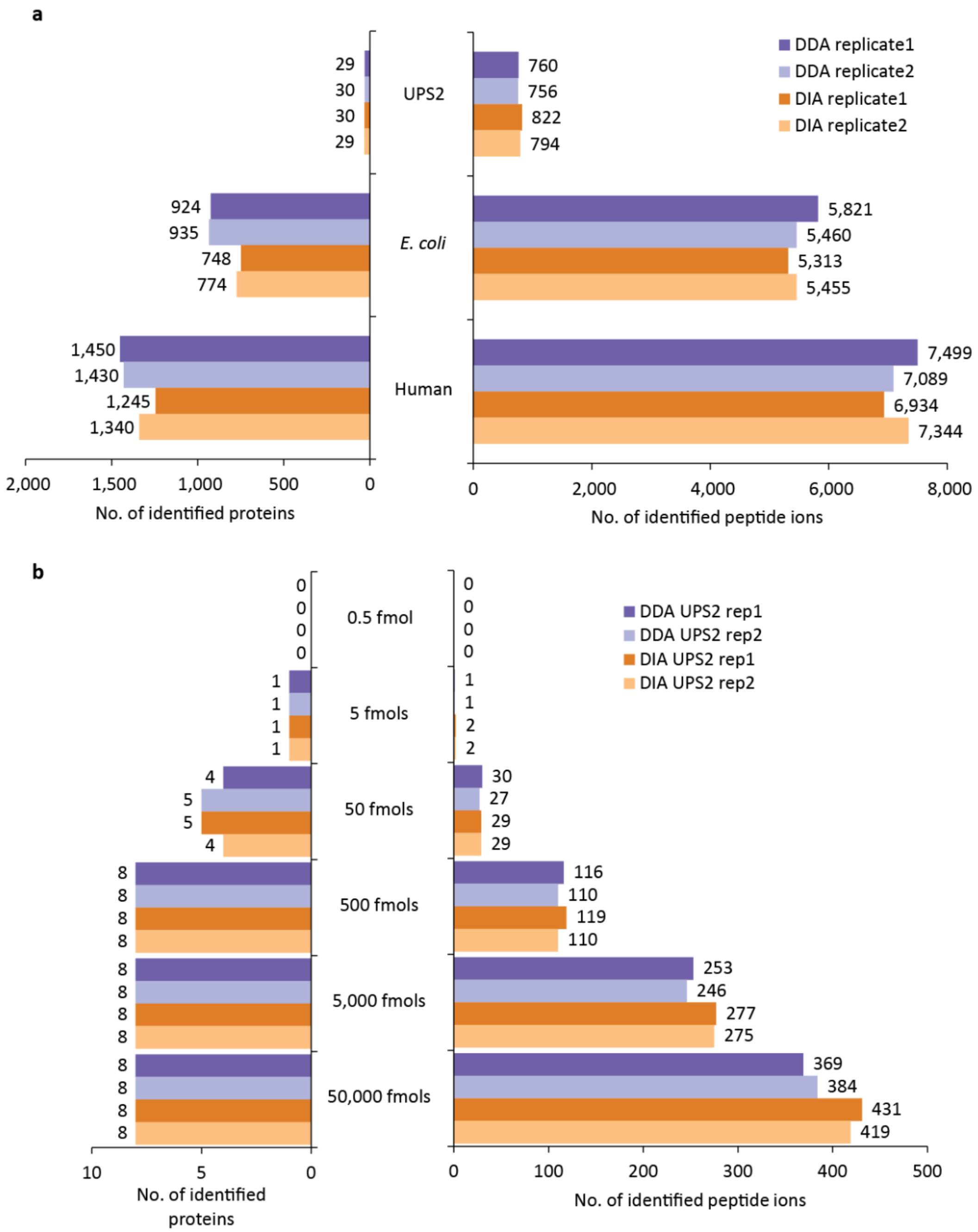


Figure 2-7 Untargeted peptide and protein identification using DDA and DIA data from UPS2, *E. coli*, and human cell lysate samples.

(a) The number of peptide ions and proteins identified by X! Tandem search engine at 1% FDR in DDA and in DIA (SWATH) data from UPS2, *E. coli*, and human cell lysate samples. **(b)** The number of peptide ions and protein identifications (X! Tandem) in each replicate of the UPS2 sample DDA and DIA data plotted separately for proteins of different abundance (in UPS2 samples 48 proteins span 5 orders of magnitude of abundance ranging from 0.5 to 50,000 fmoles with 8 proteins in each abundance range).

In low complexity UPS2 samples (48 proteins spanning 5 orders of magnitude in abundance), DIA and DDA identified similar numbers of peptide ions and proteins, with DIA identifying more peptide ions than DDA for higher abundance proteins (Figure 2-7b), and with the identification success depending on the abundance of each protein in the sample. In complex samples, such as human cell lysates (Figure 2-8a), DDA slightly outperformed DIA at both peptide ion (9,272 vs. 8,757) and protein levels (1,645 vs. 1,465). Interestingly, the overlap between the peptide ions identified with high confidence (1% FDR) by both methods was relatively low (42% compared to 78% overlap at the protein level). While some of these differences were simply due to a detected peptide ion not passing the 1% FDR threshold in one or the other approach, DIA was also able to identify peptide ions where no MS/MS spectrum was acquired in DDA (2,326 peptide ions). The lack of an acquired MS/MS spectrum in DDA was observed even for some high intensity ions, possibly due to a combination of dynamic exclusion settings and co-elution of a different (more abundant) peptide. On the other hand, DDA was more successful in identifying peptide ions for which the pseudo MS/MS spectra extracted by DIA-Umpire from DIA data did not contain enough fragment ions, many of which were of low intensity

(Figure 2-8b, c). The loss of fragment ions in DIA can be attributed to a number of factors, including suppression of fragment ions by higher intensity species in the same DIA window, which is further compounded by computational challenges such as the imperfect de-multiplexing of co-eluting peptide ions. Similar results and trends were observed when the results from all three search engines were combined (Figure A-1), and in the *E. coli* dataset (Figure A-3 and Figure A-4).

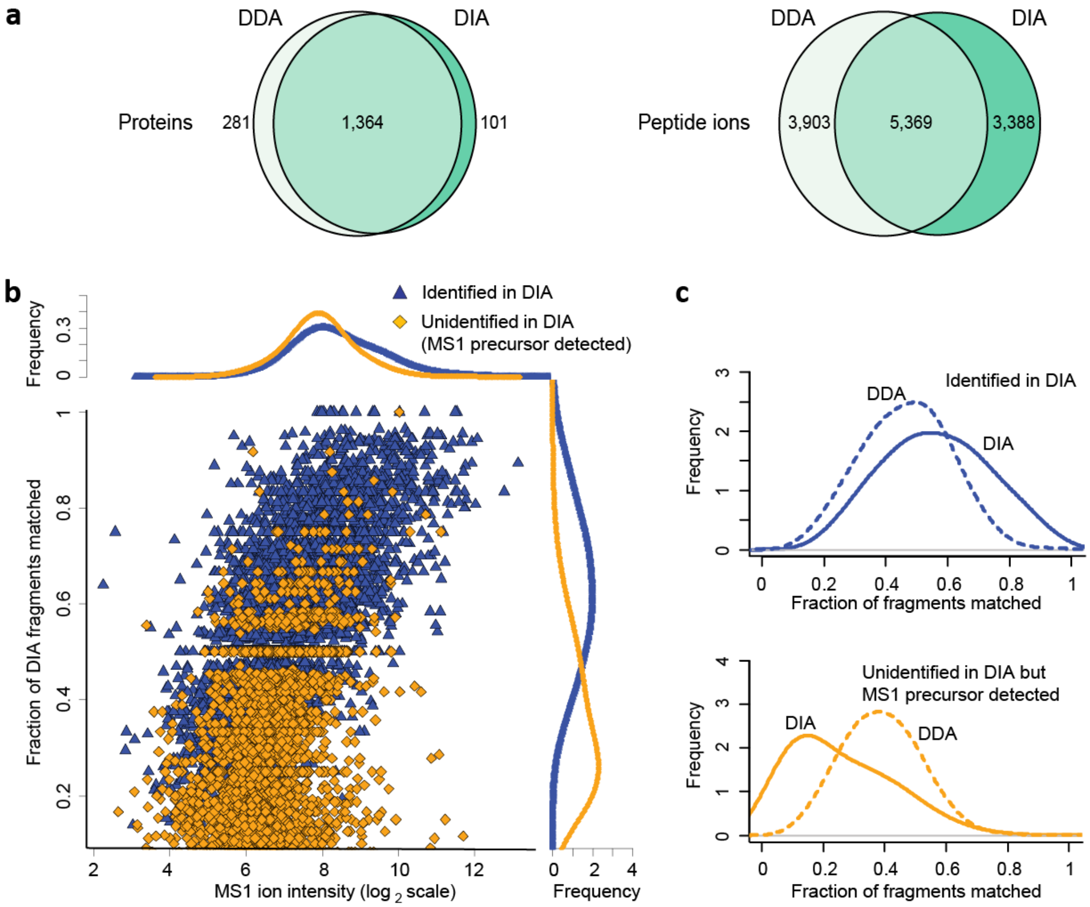


Figure 2-8 Comparative analysis of peptide identifications from DDA and DIA data from human cell lysate samples.

(a) The numbers of proteins and peptide ions identified at 1% FDR by X! Tandem search engine in DDA and in DIA (SWATH) data. Left: the number of protein identifications. Right: the number of peptide ion identifications (9,272 peptide ions identified from DDA data, 8,757 from DIA, 12,660 in total). Of the peptide ions identified by DIA and not DDA at 1% FDR (3,388), the majority were not identified by DDA because no MS/MS spectrum was acquired (2,326). Of the peptide ions identified from DDA data and not from DIA at 1% FDR (3,903), DIA-Umpire was able to detect precursor features for 3,338 of these peptide ions. **(b)** Fraction of fragment ions matched in pseudo MS/MS spectra extracted from DIA data as a function of MS1 peptide ion intensity in DDA data. Data points (peptide ions) and the summary density plots ("Frequencies") are colored according to the two categories of peptide ions: those identified from DIA data at 1% FDR (high scoring in DIA, blue), and unidentified in DIA (orange; these ions were found in DIA data as described in Online Methods). **(c)** Comparison between DDA and DIA in terms of fraction of fragments matched among the two categories of peptide ions described in (b), showing that peptide ions identified with confidence from DDA but not from DIA have fewer matched fragments.

2.3.4 *Comparison between untargeted and targeted DIA analysis*

To investigate the differences between the untargeted approach described above and targeted data extraction strategies previously applied to SWATH data, we processed the human and *E. coli* datasets using OpenSWATH [28]. We used SpectraST [53] to build spectral libraries by taking the union of DDA-identified spectra (9,272 peptide ions) from two replicates of human cell lysate data, and adding the same number of shuffled decoy spectra. In these data, OpenSWATH detected 7,372 peptide ions at 1% FDR according to mProphet [30] (Figure 2-9a). In comparison, the untargeted analysis using DIA-Umpire (i.e. searching against the whole proteome database) identified 8,757 peptide ions at the same FDR. OpenSWATH had a better overlap with the identifications from the target library than DIA-Umpire, 79% vs. 58% (Figure 2-9).

a

	DIA-Umpire		OpenSWATH
	Whole proteome (DB search)	Library peptide (DB search)	Library (Targeted extraction)
Total No. of candidate ions	68,344,142	584,721	18,544
Average No. of searched ions per spectrum	4,960	51	31
No. of identified peptide ions	8,757	8,230	7,372

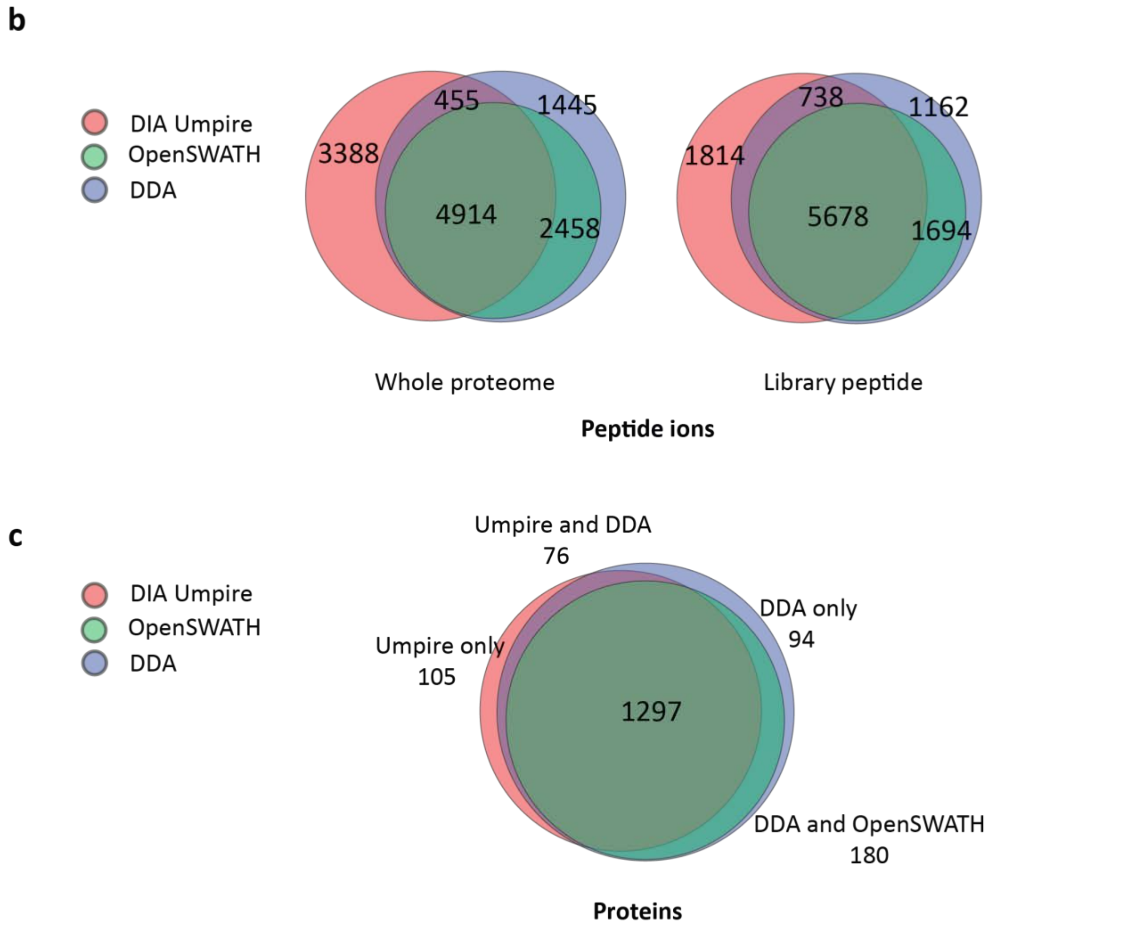


Figure 2-9 Comparison between untargeted DIA-Umpire analysis and OpenSWATH targeted extraction: effect of the search space. Human cell lysate data

(a) The pseudo MS/MS spectra extracted using DIA-Umpire were searched against two sequence databases: "Whole proteome" contains all proteins in the human proteome (plus decoy proteins); "Library peptide" database contains only the sequences of the DDA identified peptides (i.e. it is built using the same peptides as the spectral library used for targeted extraction with OpenSWATH). **(b)** Venn diagram of peptide ion identifications among the three analyses. **(c)** Venn diagram of protein identifications among the three analyses (Whole proteome sequence database was used for DIA-Umpire).

It is clear that the total search space between OpenSWATH (spectral library) and DIA-Umpire (whole proteome sequences) are significantly different. Therefore we then further compared the total search spaces between the two approaches and performed a "reduced search space" analysis to see if it improves DIA-Umpire's overlap to OpenSWATH and DDA identifications. The total numbers of candidate peptide ions considered for scoring of pseudo MS/MS spectra extracted by DIA-Umpire during database search against "Whole proteome" or "Library peptide" databases were estimated using the following parameters: 30 ppm precursor mass tolerance; peptide sequence length: 4 - 50 amino acids; one missed cleavage allowed; charge state considered: 2+, 3+, or 4+; *m/z* range: 350 - 1200 Da; variable modifications: oxidation of methionine, cysteine alkylation, conversion of pyroglutamate from glutamine or glutamic acid, and n-terminal acetylation (allowing less than six modifications on the same peptide). For OpenSWATH analysis, the following parameters were used to estimate number of candidate fragment groups in the experimental SWATH MS2 data considered for each target library peptide: mass range of the corresponding SWATH *m/z* isolation (25 Da wide) and ± 1 minute retention time window. The use of the precursor ion *m/z* value from MS1 or MS2 unfragmented precursors as a constraint during database search was the primary factor contributing to the significant reduction in the number of candidate peptide ions considered for scoring against each spectrum (from 68,344,142 peptides in the whole proteome database to 4,960 searched ions per spectrum on average, i.e. 13,779 fold reduction). Because targeted data extraction in OpenSWATH used the retention time of the peptide and wide (25 Da) *m/z* SWATH selection window (but not the precursor peptide *m/z*) for constraining the "search space", the reduction in the number of candidates was less significant (from 18,544 peptide ions in the library to 31 ions per ± 1 minute retention time slice of the corresponding 25 Da MS2 SWATH scan, i.e. 598 fold reduction). The order of magnitude difference in the search space reduction in DIA-Umpire/'Whole proteome' analysis (compared to OpenSWATH/Library analysis) explains why DIA-Umpire untargeted analysis performed well. DIA-Umpire/Whole proteome and OpenSWATH/Library identified a comparable number of peptide ions, but the two methods had only a moderate overlap.

OpenSWATH identified larger fraction of peptides in the library, 79% (i.e. (4,914 + 2,458 = 7,372) / 9,272) vs. 58% (i.e. (4914 + 455 = 5,369) / 9,272)) for DIA-Umpire. At the same time, DIA-Umpire was able to identify a large number of peptide ions not present in the library. DIA-Umpire / 'Library peptide' analysis had an effective search space similar to that of OpenSWATH, resulting in even closer performance (and better overlap) between the two methods: the overlap between the DIA-Umpire identified peptides and the DDA-identified peptides improved to 69% (or (5,678 + 738) / 9,272 peptide ions). Similar results were obtained for *E. coli* samples (Figure A-4). The use of the entire database however provided us with the opportunity to identify peptide ions not present in the DDA-constructed library.

Furthermore, when we built the spectral library for OpenSWATH from the pseudo MS/MS spectra confidently identified by DIA-Umpire (8,757 peptide ions for the human cell lysate dataset), OpenSWATH confidently identified 8,650 (98.8%) of the library peptides, providing additional validation of the peptides identified using untargeted DIA-Umpire approach. We observed similar results (96.2% confirmation rate) for *E. coli* samples. The small percentage of peptide ions not identified by OpenSWATH was in part due to OpenSWATH's internal filtering of spectra from the input library.

2.3.5 *Comparison between untargeted and targeted DIA analysis using an SWATH N-glycopeptide dataset*

We further assessed the performance of untargeted DIA-Umpire approach on a publicly available SWATH *N*-glycopeptide dataset [35] from prostate cancer, which was already processed by the authors with OpenSWATH using a spectral reference library containing deamidated asparagine peptides built from a large number of DDA runs. Identifications from DIA-Umpire were further filtered by isotopic distribution of precursor signal (See 2.2.12). The number of identified peptide ions and ambiguous identifications for each DIA run is shown in Figure 2-10.

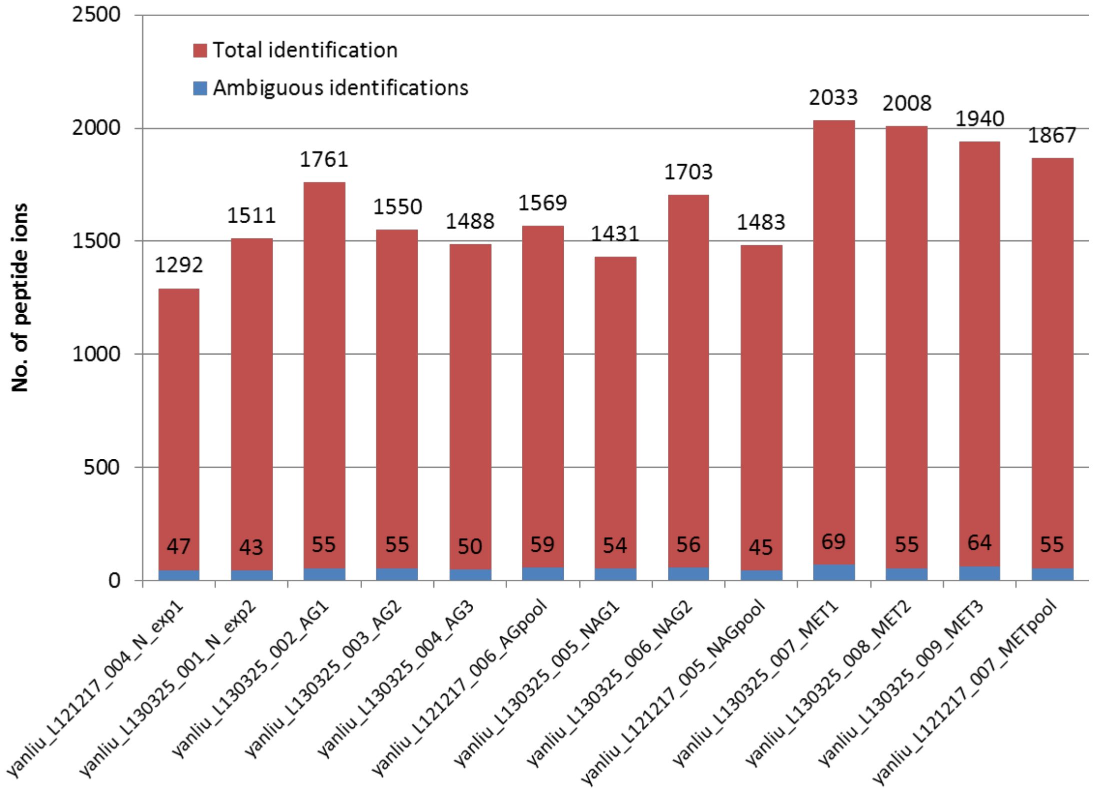


Figure 2-10 Number of peptide ion identifications and ambiguous identifications filtered by isotope pattern for each DIA run.

Red: The total number of identifications for each DIA file. **Blue**: The number of ambiguous identifications found by isotope pattern filtering.

At 1% FDR (mProphet computed m_score < 0.01 for OpenSWATH), DIA-Umpire and OpenSWATH identified 1,821 and 1,383 deamidated asparagine peptide sequences (2,933 and 1,537 peptide ions) respectively (Figure 2-11). Among the additional identifications introduced by DIA-Umpire, more than 80% had a *N*-glycosylation (NX-S/T) motif, indicating high site specificity of the additional identifications (non-consensus identifications could be due to standard deamidation of non-glycosylated peptides, as we have not restricted our analysis to a library enriched in glycosylation sites, in contrast to the OpenSWATH library).

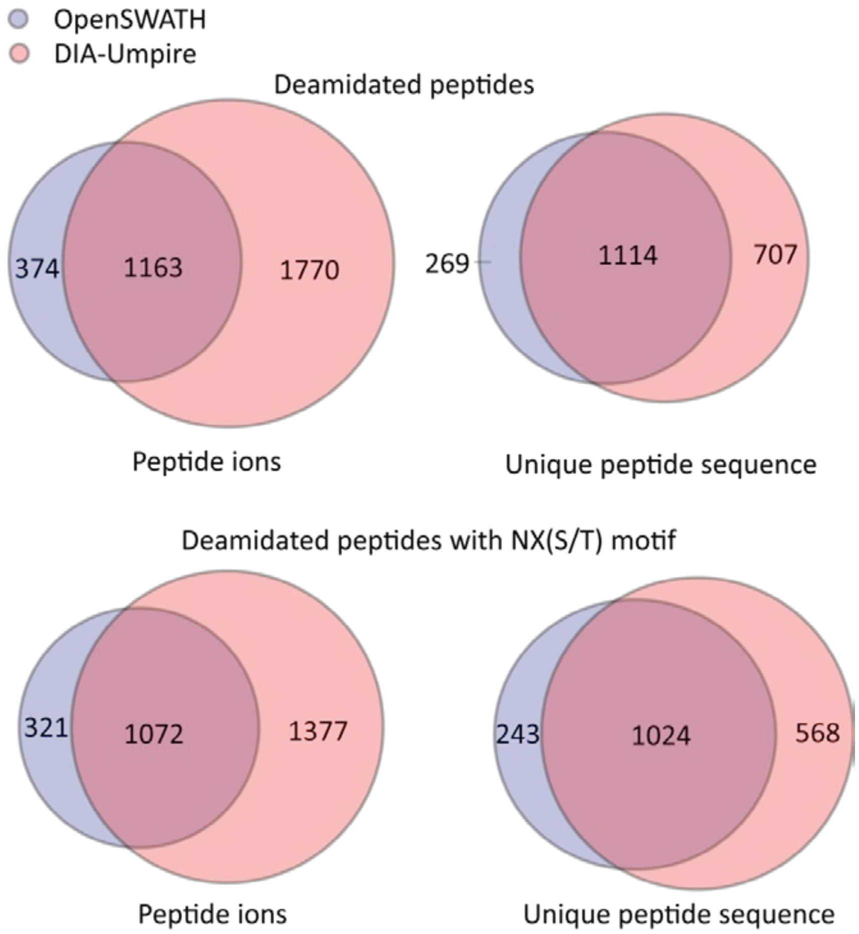


Figure 2-11 Deamidated peptide identifications

The number of peptide ions and unique peptide sequences from DIA-Umpire and OpenSWATH targeted search. Glycoproteomics data from Liu *et al* [35]). **Upper panel**: all peptides. **Lower panel**: peptides containing the NX(S/T) motif expected to be significantly enriched in these data.

An additional drawback of other targeted extraction approaches (e.g. OpenSWATH) for DIA analysis is that they have difficulties resolving ambiguities for peptide ions that share many MS2 fragments (e.g. unmodified and post-translationally modified peptides co-fragmenting in the same isolation window), especially since the exact precursor mass is not used for scoring. Although retention times of different modified / unmodified peptide species can help resolve the ambiguity [28], how different modifications influence retention time remains an open problem for computational prediction [56]. If both species exist in the library and only one of them is present in the sample, library spectra from both species might match the same fragment peak groups in the queried DIA MS2 data. Although the correct one should get a higher matching score, the score of the incorrect one is likely to be better than any decoy peptide and thus also deemed a confident identification. In the glycoproteomics application presented above, identification of *N*-linked glycopeptides depends on detection of asparagine deamidation in peptides due to PNGase F treatment, which causes only a small mass shift (0.984 Da), resulting in both modified and unmodified peptides often being co-fragmented. Therefore, we searched the OpenSWATH data for identifications of both deamidated and non-deamidated species reported as highly confident (m_score < 0.01), at same retention time (within 1 second), and of the same charge state. We present examples (Figure 2-12, and Figure A-5 - Figure A-8 for additional examples) in which OpenSWATH was not able to distinguish deamidated peptide ions from unmodified peptide ions (we manually checked these by using the exact precursor mass). More specifically, two separate identifications with different modification site compositions (with one and two deamidations; modification site shown in red) were reported by OpenSWATH. In contrast, DIA-Umpire constructs pseudo MS/MS spectra according to detected high mass accuracy precursor features, enabling better differentiation of peptide species.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
NTTFNVESTK	571.76	2	36.9	3.03E-05	N/A
NTTFNVESTK	571.27	2	36.9	1.07E-07	36.88

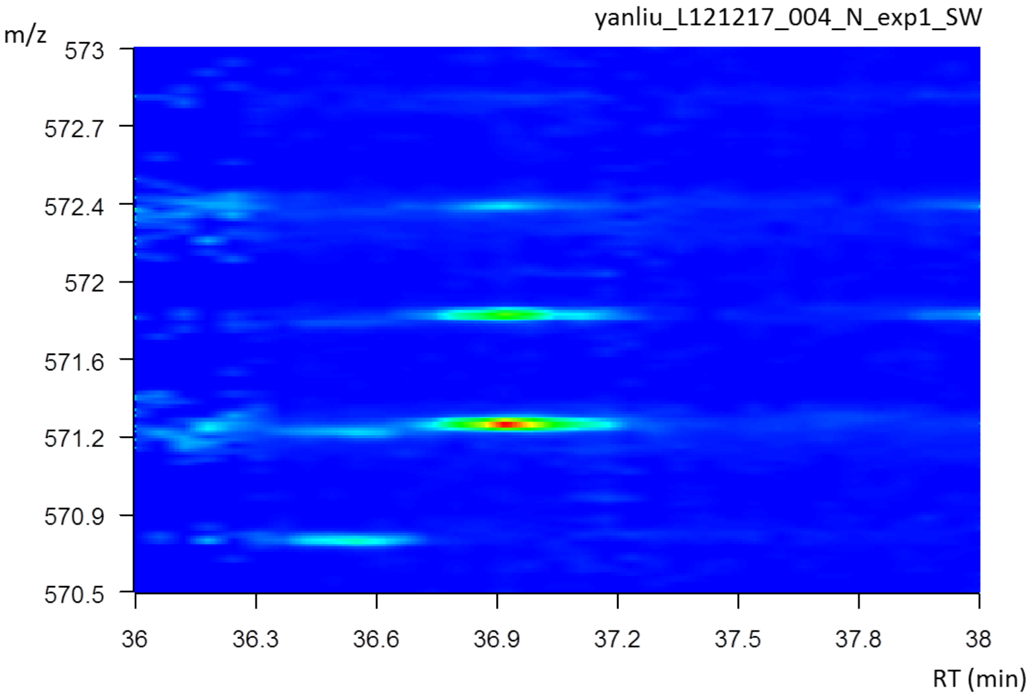


Figure 2-12 Example of an ambiguous identification involving the deamidated peptide NTTFNVESTK by OpenSWATH targeted search

The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 36.9 minutes (precursor *m/z* of 571.27 Da). DIA-Umpire reported only one (singly deamidated) form, further supported by the presence of NXS/T motif covering the reported site. This example demonstrates that the knowledge of the precursor mass can be very valuable for differentiating between different modification forms of the same peptide sequence in DIA experiments. Similar examples are shown from Figure A-5 to Figure A-8.

2.4 *Discussion*

We demonstrated in this study that by using untargeted identification analysis workflow in DIA-Umpire we could identify a comparable number of peptides and proteins from DIA and DDA data. However, we have also observed the complementary nature of these two data acquisition strategies. As both the DDA and DIA technologies and the underlying instrumentation are rapidly improving, future work should include a comprehensive comparative analysis of different workflows as they are applied to a variety of biological problems. DIA has a significant potential for further improvements, including removal of interferences by tuning and optimizing the DIA isolation window size to achieve a better balance between the scan cycle time and the level of co-fragmentation interferences, which should further boost the identification sensitivity of DIA-Umpire algorithm for low abundance ions. Importantly, DIA-Umpire is compatible with different DIA strategy variants, including implementations in different instruments. The highly flexible design of the DIA-Umpire computational framework allows us to quickly adopt the algorithms to take advantage of the new strategies and technological improvements.

Given previous reports discussing advantages of targeted data extraction strategies for DIA data, one may wonder why our untargeted DIA data analysis workflow would perform so well. First, we note that the targeted (peptide-centric matching) and untargeted (spectrum-centric matching) DIA data analysis strategies are not fundamentally different. Both approaches are trying to find the best matches between records in a sequence database or a spectral library and the signal in acquired experimental data (precursor and fragment ion features) that are statistically significant (i.e. scoring significantly higher than random matches). The main difference between the two is the search space, i.e. the total number of possible peptide ions (including decoys) that are considered as possible matches to the experimental signal. In peptide-centric matching of the targeted data extraction using tools such as OpenSWATH, the peptide “search space” is typically much smaller (i.e. peptides present in the spectral library vs. the entire protein sequence database), and the “search space” within the experimental data is restricted by knowledge of the peptide retention time contained in the spectral library. In the untargeted, spectrum-centric strategy of DIA-Umpire, pseudo MS/MS spectra extracted from experimental DIA data are searched against a much larger peptide sequence database. However, as with the conventional DDA MS/MS database search, the actual peptide search space for each pseudo MS/MS spectrum is very efficiently reduced using the precise knowledge of the precursor *m/z* value, therefore enabling sensitive, untargeted identification analysis for DIA data.

2.5 *Contributions*

The results presented in this chapter could not have been done without the efforts of the great collaboration. I am grateful and fortunate to have such a great team to participate the project.

Chih-Chiang Tsou: developed the algorithms, implemented the software, designed experiments, analyzed the data, and wrote the manuscript draft.

Dmitry Avtonomov: assisted with the OpenSWATH analysis and contributed to the algorithm and software development and reviewed the manuscript.

Brett Larsen: acquired mass spectrometry data, designed experiments, analyzed data and provided inputs for manuscript.

Monika Tucholska: acquired mass spectrometry data.

Hyungwon Choi: assisted with SAINT scoring and contributed to the development of protein quantification strategies.

Anne-Claude Gingras: designed experiments, analyzed data, supervised the project, and wrote the manuscript.

Alexey I Nesvizhskii: conceived the project, developed the algorithm, designed experiments, analyzed the data, supervised the project, and wrote the manuscript.

Chapter 3 Hybrid DIA quantification workflow using DIA-derived internal library

The content of this chapter was previously published by the author as a research article in *Nature Methods* [41].

3.1 *Introduction*

Quantitative proteomics experiments using mass spectrometry are often conducted with multiple biological and also technical replicates in order to estimate biological and technical variations for the downstream statistical analysis [57, 58]. In a typical label-free quantitative experiment using DDA with multiple LC-MS runs, a peptide ion identified in one replicate LC-MS run does not necessarily guarantee that it could be confidently identified in another replicate run. Such irreproducibility is attributed mainly by two factors. First is the stochastic precursor ion selection of DDA for MS/MS acquisition. A peptide ion with a good MS/MS identification in one LC-MS run may not have any MS/MS spectrum acquired in another run. Second, the variations of signal quality between LC-MS runs could also affect the identification reproducibility. A peptide ion having a confident identification in an LC-MS run may have a slightly poorer spectrum in another LC-MS run. Therefore the poorer spectrum may score just below the FDR threshold and be considered as an unconfident identification in the run. DIA addresses the issue of stochastic MS/MS acquisition, but it is still not immune to the variations of signal quality between LC-MS runs. The missing identifications, i.e. missing quantifications, make the statistical analysis less reliable and could mislead the biological interpretations. To alleviate the problem, label-free quantitative analysis obtained from DDA experiments performs a retention time alignment process to map peptide identifications between LC-MS runs based on detected MS1 precursor features. Several label-free quantification tools such as MaxQuant [59], OpenMS [60], and IDEAL-Q [61] have built-in retention time alignment algorithm in the quantification software tool.

In the second study, we extend DIA-Umpire's untargeted identification analysis further to quantification analysis. To address the missing identification issue, we developed a novel targeted re-extraction strategy which is based on DIA peptide ion retention time alignment and an internal spectral library matching. For quantification analysis based on MS2 fragment signals, we develop fragment and peptide ion selection algorithms to select high quality fragments and peptides for protein quantification. We use the datasets presented in the first study to show the improvements of the targeted re-extraction and the performance of the quantification. Furthermore, we perform a complete DIA quantification analysis on a public AP-SWATH dataset (Lambert *et al* [26]) to show that DIA-Umpire's quantification analysis can sensitively capture protein-protein interaction profile and achieve comparable results without using a pre-existing spectral library.

3.2 *Methods*

3.2.1 *Data processing for datasets*

The UPS and human datasets were the same as used in Chapter 2. The AP-SWATH dataset was downloaded from http://prohits-web.lunenfeld.ca/. All the raw .wiff files were converted into mzXML format. DIA data were processed by DIA-Umpire to generate pseudo MS/MS spectra, and all DDA and DIA pseudo MS/MS spectra were searched by database search engines using the same parameters as described in Chapter 2. For the AP-SWATH dataset, ProteinProphet analysis was performed by taking all PeptideProphet output files (X! Tandem results) from all SWATH runs, i.e. EIF4A2 and MEPCE bait data (biological triplicates for each bait) and the three GFP negative controls. The final protein lists for each ProteinProphet analysis were determined using an 1% FDR threshold, estimated by target–decoy approach.

3.2.2 *Quantification in DDA data*

We used elution apex intensity of the MS1 precursor feature when performing peptide quantification for DDA MS1 data. For each MS/MS spectrum identified in a DDA experiment, all precursor features observed in the MS1 data with close monoisotopic *m/z* (same precursor *m/z* tolerance as used in the database search), close retention time (within ±1 minute), and same charge state were considered as candidates. Among these candidates, the MS1 feature with the closest retention time was considered as the precursor ion for the identified MS/MS spectrum. As with DIA MS1 data in DIA-Umpire, peptide ion intensity and its retention time in DDA MS1 were determined from the intensity and the retention time at the LC apex of the monoisotopic peak.

3.2.3 *Comparison of ion intensities between DDA and DIA*

To compare the fragment ions observed for the same peptide in DDA and in DIA experiments, the compomics-utilities library [62] was used to generate theoretical peptide fragments. To find the signal of a peptide ion which was only identified in either DDA or DIA data, the retention time observed for that ion in the run where it was identified was used to detect the corresponding peptide ion feature in the other run. This was done without the need for retention time alignment between the runs because of the excellent retention time and ion intensity reproducibility between DDA and DIA runs on the same samples (see Figure B-1). An MS1 precursor feature *m/z* window of ±30 ppm and a retention time window of ±1 minute were used. For DDA data, the highest intensity candidate was selected among multiple possible ones. For DIA data, the best candidate (precursor-fragment group) was selected based on the number of matched fragments between the corresponding pseudo MS/MS spectrum and the DDA identified peptide sequence. The number of matched fragments was calculated as follows: for each DDA MS/MS spectrum, or pseudo MS/MS spectrum in DIA, only the top 140 highest intensity peaks were considered. The mass tolerance for peak matching was set to 40 ppm for AB SCIEX 5600 TripleTOF. The analysis was restricted to *b*- and *y*-fragment ions only. A peak in an experimental spectrum was allowed to be matched to only one theoretical fragment. The number of matched fragments for each spectrum was counted and then normalized by the total number of theoretical fragments for that peptide.

3.2.4 *Targeted extraction using internal spectral libraries*

The targeted extraction module of DIA-Umpire (peptide-centric matching) was developed as an optional second step in the DIA-Umpire workflow to increase the quantification coverage across multiple samples. Given a set of peptides identified by the initial untargeted database search, the algorithm builds an internal consensus spectral library using confident identifications from all DIA runs. In addition, DIA runs are aligned in retention time using commonly identified peptides between the DIA runs as pivot points for non-linear regression [61]. For a peptide ion not identified in a particular DIA run in the untargeted way, DIA-Umpire calculates its retention time (via retention time alignment) and *m/z* (via mass calibration model, described below), and performs targeted data re-extraction. This is achieved by matching the library spectrum of that peptide ion against precursor-fragment groups previously extracted from the experimental data within a narrow retention time window (in this work, ±1 minute of the calculated retention time) and a narrow precursor mass window (±30 ppm of the calibrated precursor mass). The details of this targeted data extraction algorithm are described below.

3.2.4.1 Internal spectral library generation

A consensus spectral library is built using confident identifications (here, 1% FDR at the peptide level) from the initial untargeted analysis of the DIA data. First, for each confident pseudo MS/MS spectrum match, the matched fragment intensities are normalized to the most intense matched fragment. For a peptide ion which has multiple spectra identified across samples, the intensity of a fragment in the consensus spectrum is computed as the average fragment intensity across all corresponding identified spectra. Decoy spectra are created by the "shuffle-and-reposition" method [63], and such a decoy is generated for each peptide ion in a consensus spectral library.

3.2.5 *Retention time prediction and mass calibration for target peptide ions*

DIA-Umpire adopts a previously described [61] nonlinear regression-based method for retention time calculation and a mass calibration model [64] for adjusting precursor *m/z* values of a peptide ion in a DIA run. To generate the retention time model for a pair of DIA-runs, retention times of commonly identified peptide ions from the initial untargeted identifications are used and a non-linear regression model is built based on these retention times. For mass calibration, mass errors are represented as a function of the retention time, and a non-linear LOWESS regression is done for calculation of peptide ion mass errors given the peptide retention time in a DIA run.

3.2.6 *Peptide-centric matching targeted re-extraction*

To find the best matching precursor-fragment group for a peptide ion from a spectral library, all precursor-fragment groups within the range of ± 30 ppm (user-defined parameter) of the calculated precursor *m/z* and ± 1 minute of calculated retention time are considered as candidates. A library spectrum S is denoted as

$$\mathrm{S} = \{(I_1^{\mathrm{S}}, M_1^{\mathrm{S}}), (I_2^{\mathrm{S}}, M_2^{\mathrm{S}}), \dots, (I_{\mathrm{NS}}^{\mathrm{S}}, M_{\mathrm{NS}}^{\mathrm{S}})\}$$

where NS is the number of fragment peaks in the spectrum, and I_r^{S} and M_r^{S} are the intensity and the theoretical *m/z* value, respectively, of each fragment F_r that belongs to spectrum S. A precursor-fragment group G is represented as

$$\mathrm{G} = \{(I_1^{\mathrm{G}}, M_1^{\mathrm{G}}, C_1^{\mathrm{G}}), (I_2^{\mathrm{G}}, M_2^{\mathrm{G}}, C_2^{\mathrm{G}}), \dots, (I_{\mathrm{NG}}^{\mathrm{G}}, M_{\mathrm{NG}}^{\mathrm{G}}, C_{\mathrm{NG}}^{\mathrm{G}})\}$$

where NG is the number of fragment peaks, I_r^G and M_r^G are the intensity and *m/z* value, respectively, of each fragment F_r that belongs to precursor-fragment group G, and C_r^{G} is the Pearson correlation coefficient between the fragment F_r and the precursor anchoring group G. Given a library spectrum S and a precursor-fragment group G, matched fragment peaks from the precursor-fragment group are extracted using a predefined mass tolerance (e.g. ± 40 ppm for AB SCIEX 5600 instrument). The algorithm then calculates five sub-scores for the match between S and G. In addition to the number of matched fragments (L), it calculates a spectral similarity score as follows. Consider an intensity vector $\mathrm{INT}^{\mathrm{G\text{-}S}} = (I_1^{\mathrm{G}}, I_2^{\mathrm{G}}, \dots, I_{\mathrm{NS}}^{\mathrm{G}})$ of length NS, with I_r^{G} taken as the intensity of the fragment peak F_r in G that matches to a fragment in S, and as zero if no fragment peak can be found in G within the specified mass tolerance window around M_r^{S}. The spectral similarity is then calculated by Pearson correlation between the vector $\mathrm{INT}^{\mathrm{G\text{-}S}}$ and the library spectrum intensity vector $(I_1^{\mathrm{S}}, I_2^{\mathrm{S}}, \dots, I_{\mathrm{NS}}^{\mathrm{S}})$.

Three more scores, Mass Error Score (MES), Intensity Score (IS), and Correlation Score (CS), are calculated using matched fragments F_j only as follows:

$$\mathrm{MES} = 1 - \frac{\sum_{j=1}^{L} \mathrm{PPM}(M_j^{\mathrm{G}}, M_j^{\mathrm{S}})}{40L}, \text{where } \mathrm{PPM}(m_a, m_b) = \frac{|m_a - m_b| \times 2 \times 10^6}{m_a + m_b}$$

$$\mathrm{IS} = \frac{\sum_{j=1}^{L} \left| I_j^{\mathrm{G}} \right|}{L}$$

$$\mathrm{CS} = \frac{\sum_{j=1}^{L} \left| C_j^{\mathrm{G}} \right|}{L}$$

The final match score (U-score) between S and G is calculated as a linear combination of these five sub-scores, with the score weights determined using the linear discriminant analysis (LDA) [30]. For LDA model training, 50% of all matches in which S is a decoy spectrum are randomly selected and labeled as negative training data (the other 50% are held away from the training; these can be used at the final stage to assess the quality of the model fitted using the mixture modeling algorithm described below). Positively labeled training dataset is composed of likely true matches, i.e. matches between S and G that were identified with high scores at the initial untargeted identification stage.

3.2.7 *Targeted re-extraction U-score probability and FDR*

The U-score is computed for all targets considered in a particular DIA run. Targets are defined here as peptide ions represented in the spectral library (created, as described above, from all DIA runs in the experiment) that were not identified in that particular DIA run by the untargeted search. The U-score distribution for these target matches computed as described above is assumed to be a bimodal distribution representing populations of correct and false matches (see Figure 3-8c). In the first study, this distribution is modeled as a mixture of two normal distributions and is de-convoluted using the expectation maximization (EM) algorithm[50]. The probability that a match is correct, given the U-score U, is determined as

$$P(\text{Correct}|U) = \frac{\pi_1 f_1(U)}{\pi_0 f_0(U) + \pi_1 f_1(U)}$$

Here $f_1(U)$ and $f_0(U)$ are Gaussian density functions (from the mixture model above) for correct and false matches, respectively. The parameters of the distributions and their mixing weights, π_0 and π_1, are determined directly from the data using the EM. FDR can then be estimated using computed probabilities [1, 50]. In this study, a probability threshold of 0.99 (estimated FDR of less than 1% in these data) was applied as the final filter.

3.2.8 *Quantification in DIA data using DIA-Umpire*

The quantification module of DIA-Umpire computes peptide and protein intensities estimated either from MS1 precursor ion intensities or from MS2 fragment ion intensities. We use the LC apex intensity of the smoothed MS1 monoisotopic peak to determine the MS1 precursor ion intensity. For MS2 fragment ion intensity, we use the raw LC apex intensity of the fragment signal. The MS2 fragment-based intensity for a protein can be computed by summing the intensities of all matched fragments of all identified peptide ions from that protein (or only using selected peptide ions and fragments, as described below). In rare cases, the same peptide ion is identified from multiple precursor ion features (i.e. at different retention times). Such peptides are excluded by default (optionally, such peptide ions can be used for quantification by selecting the precursor ion feature with the highest MS1 intensity). When computing protein-level intensities, the analysis can be based on all peptides or based only on peptides unique to a particular protein group (e.g. with ProteinProphet computed group weight above 0.9; default option).

For fragment-based quantification, DIA-Umpire computes two protein intensity measures. The MS2 iBAQ intensity is computed for each protein by summing the intensities of all matched fragments from all identified peptide ions divided by the number of expected tryptic peptides (similar to the iBAQ score [65] commonly used for DDA MS1 intensity data). This intensity measure can be computed for all proteins identified in the dataset. In addition, DIA-Umpire also computes a protein intensity measure using selected fragments and peptide ions consistently identified across multiple samples within the whole dataset, as described below.

3.2.8.1 Fragment selection

For a peptide ion which is identified in $\mathrm{N_{pep}}$ DIA runs within the experiment, only fragments detected in more than MinFreq × $\mathrm{N_{pep}}$ DIA runs are kept. For each remaining fragment F_r, the fragment quality score

$$\mathrm{FQ}_r = \sum_{j=1}^{\mathrm{N_{pep}}} C_r^j \times I_r^j,$$

is calculated using the Pearson correlation C_r^j between fragment F_r and its precursor peak in DIA run *j,* and the apex intensity of a fragment F_r in DIA run *j,* I_r^j. For a peptide ion, its top $\mathrm{T_F}$ best (i.e. with highest QF scores) fragments (e.g. $\mathrm{T_F}$ =6, denoted as Top6fra option) are selected for quantification. Peptide ion intensity in a DIA run is then determined by summing the intensities of all selected fragments.

Note that Teo et al recently proposed a downstream statistical analysis tool for DIA data called mapDIA [58], which includes a more sophisticated fragment selection based on fragment quantification profile correlation. The output file of DIA-Umpire output file is fully compatible with mapDIA to perform the downstream statistical analysis.

3.2.8.2 Peptide ion selection

To select peptide ions for protein quantification, only peptide ions identified in more than MinFreq × N_{prot} runs are kept, where N_{prot} is the number of DIA runs in which the protein was identified. The peptide ion intensity in each DIA run is computed using the intensities of its fragments selected as described above (e.g. “Top6fra” option). The peptide ion quality score is then computed as a sum of peptide ion intensities across all runs in the dataset. The protein intensity in then calculated in each DIA run by summing the intensities of the top T_P highest quality peptide ions (e.g. T_P =6, denoted as “Top6pep” option).

The thresholds described above were implemented as input parameter options. The following parameters were selected in this work based on the analysis of variability between two replicate human cell lysate runs: T_P = 6, T_F = 6, and MinFreq = 0.5 (denoted as “Top6pep/Top6fra Freq>0.5”). Note that this selection procedure may lead to a loss of a small number of identified proteins that cannot be quantified due to lack of reproducible peptide ions passing the filters described above. Out of 1,653 proteins identified in both replicates of the human cell lysate data, only 12 were not quantified by the “Top6pep/Top6fra Freq > 0.5” approach.

3.2.9 *MS1-based quantification*

DIA-Umpire also reports two MS1-based quantification scores. The MS1 iBAQ protein intensity is computed as previously described [65], with peptide ion intensities determined at the apex of the LC elution monoisotopic peak. Note that MS1-based peptide quantification is only available for peptide ions identified from QT = 1 or 2 category pseudo MS/MS spectra (no MS1 feature is detected for QT = 3 spectra). In the human cell lysate data, 1,324 out of 1,353 proteins were quantified by MS1 iBAQ in both replicates. In addition, DIA-Umpire reports a second MS1 quantification score computed as a sum of intensities of top T_P peptide ions, selecting peptide ions for quantification in a similar manner as described above for MS2-based quantification (e.g. top 6 most intense peptide ions, with an additional MinFreq = 0.5 filter: "Top6pep, Freq > 0.5" option).

To demonstrate the importance of selecting the most reliable peptide ions and fragments across all samples for quantification, we also implemented a selection procedure applied independently within each DIA run ("MS1 Top3pep (indep. selection)"**;** "MS2 Top3pep/Top2fra (indep. selection)"), which produced significantly worse results.

3.2.10 *SAINT interaction scoring for AP-SWATH interactome dataset*

AP-SWATH interactome dataset was processed using the entire DIA-Umpire pipeline including feature detection, untargeted identification, targeted re-extraction, peptide ion and fragment selection, and protein quantification. Protein and peptide identifications were filtered at 1% and 5% FDRs, respectively. Missing identifications across replicates and samples were re-extracted by the peptide-centric matching with a 0.99 probability threshold as the filter. For protein quantitation, we used the "Top6pep/Top6fra, Freq > 0.5" approach to determine protein intensity using only peptides unique to a particular protein group (ProteinProphet computed group weights above 0.9). Protein quantification data from EIF4A2 and MEPCE bait experiments with GFP negative controls were analyzed using SAINT (intensity model; v2.3.4) [66] to determine high confidence protein-protein interactions (here, SAINT probability above 0.95).

3.2.11 *Data availability*

All the spectrum files of the datasets (Table A-1) along with DIA-Umpire results presented in this study have been deposited at the ProteomeXchange Consortium [55] (*http://proteomecentral.proteomexchange.org*) via the PRIDE partner repository with the dataset identifier PXD001587.

3.3 *Results*

3.3.1 *DIA-Umpire's hybrid targeted re-extraction workflow*

DIA-Umpire pipeline includes a targeted quantification workflow in which the library is generated based on untargeted identifications directly from DIA data as described in Chapter 2. As the workflow shown in Figure 3-1, a targeted re-extraction approach that queries unidentified precursor-fragment groups against an internal spectral library built from untargeted database search results, allowing more consistent quantification of peptide ions across multiple experiments is developed in DIA-Umpire pipeline. Exact retention time is known for peptides identified in a given experiment, and the commonly identified peptides are used to perform retention time alignment across all the runs, negating the need for external retention time calibration peptides, e.g. iRT peptides. An additional advantage of this approach over previously-described targeted extraction strategies [16, 18, 26, 28] is that the precursor peptide *m/z* value is used to constrain the search space, enabling to distinguish between peptides with multiple shared fragments (e.g. modified and unmodified peptides).

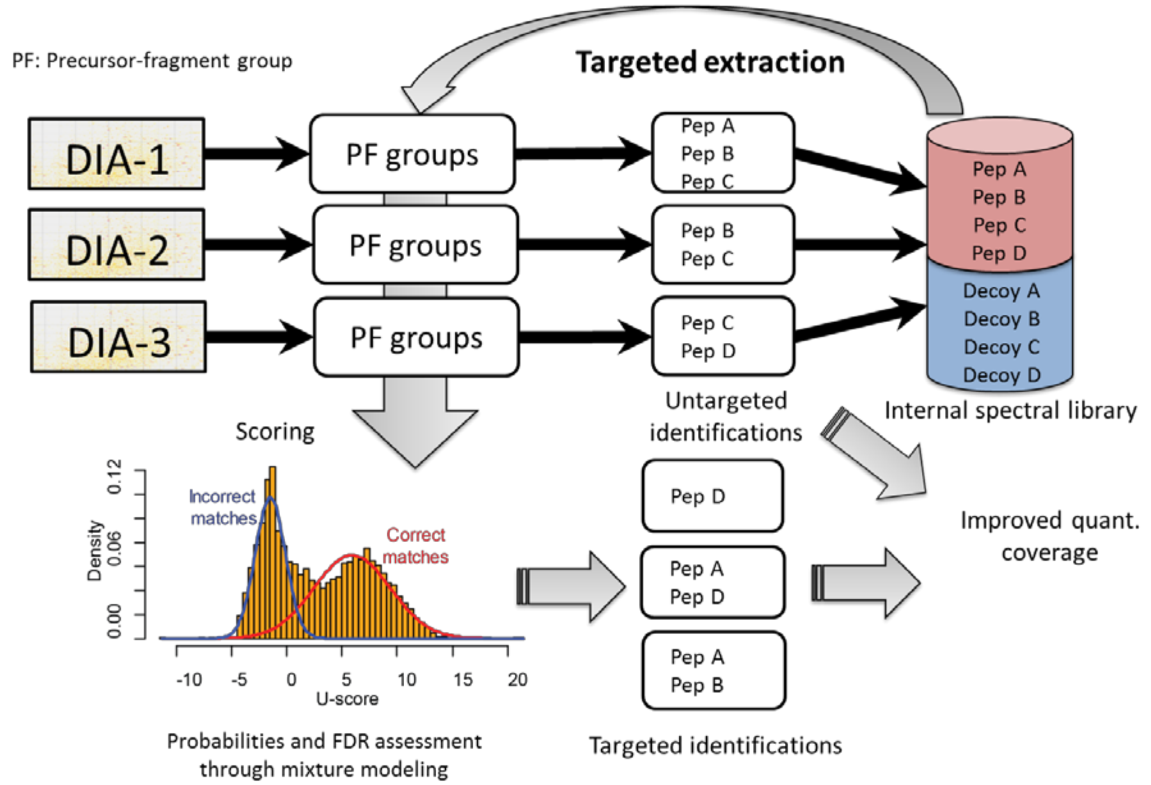


Figure 3-1 DIA-Umpire targeted re-extraction using internal library

3.3.2 *Targeted extraction and protein quantification*

Accuracy and coverage of protein quantification is of critical importance for downstream analysis of proteomic data. Following the initial untargeted analysis, DIA-Umpire fills in the missing peptide ion intensities across samples by creating an internal spectral library from all the identified peptides, followed by re-extraction of quantitative information across all precursor–fragment groups, including those which were not identified in the untargeted manner in some samples (Figure 3-1). In the human cell lysate data, targeted re-extraction improved the number of peptide ions and proteins identified across both replicate runs (estimated FDR less than 1%), with the overlap between the replicates at the peptide ion and protein levels increasing from 63% to 80% and from 84% to 93%, respectively, compared to the initial untargeted identification results (Figure 3-2).

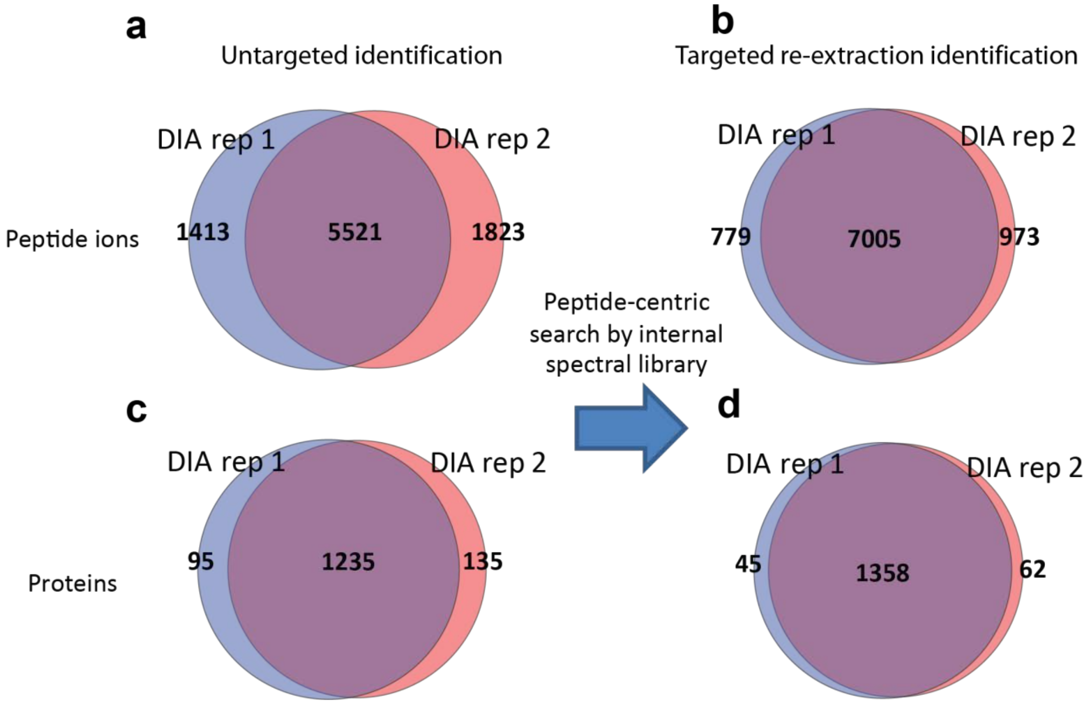


Figure 3-2 Increased identification coverage after targeted re-extraction in DIA-Umpire. Human cell lysate DIA data.

(a) Venn diagram of peptide ion identifications in two replicates of human cell lysate DIA (SWATH) data from untargeted X! Tandem search; **(b)** the number of peptide ions identified in both replicates increased after targeted re-extraction; **(c)** same as (a) at the protein level; **(d)** same as (b) at the protein level.

The same human cell lysate data was used to investigate the quantification performance of the algorithm. DIA-Umpire computes two iBAQ [65] protein abundance measures (from MS1 and MS2 data) as well as "Top *N* peptides" (MS1) [17] and "Top *N* peptides/ Top *M* fragments" (MS2) [67] metrics (Figure 3-3 and Figure 3-4).

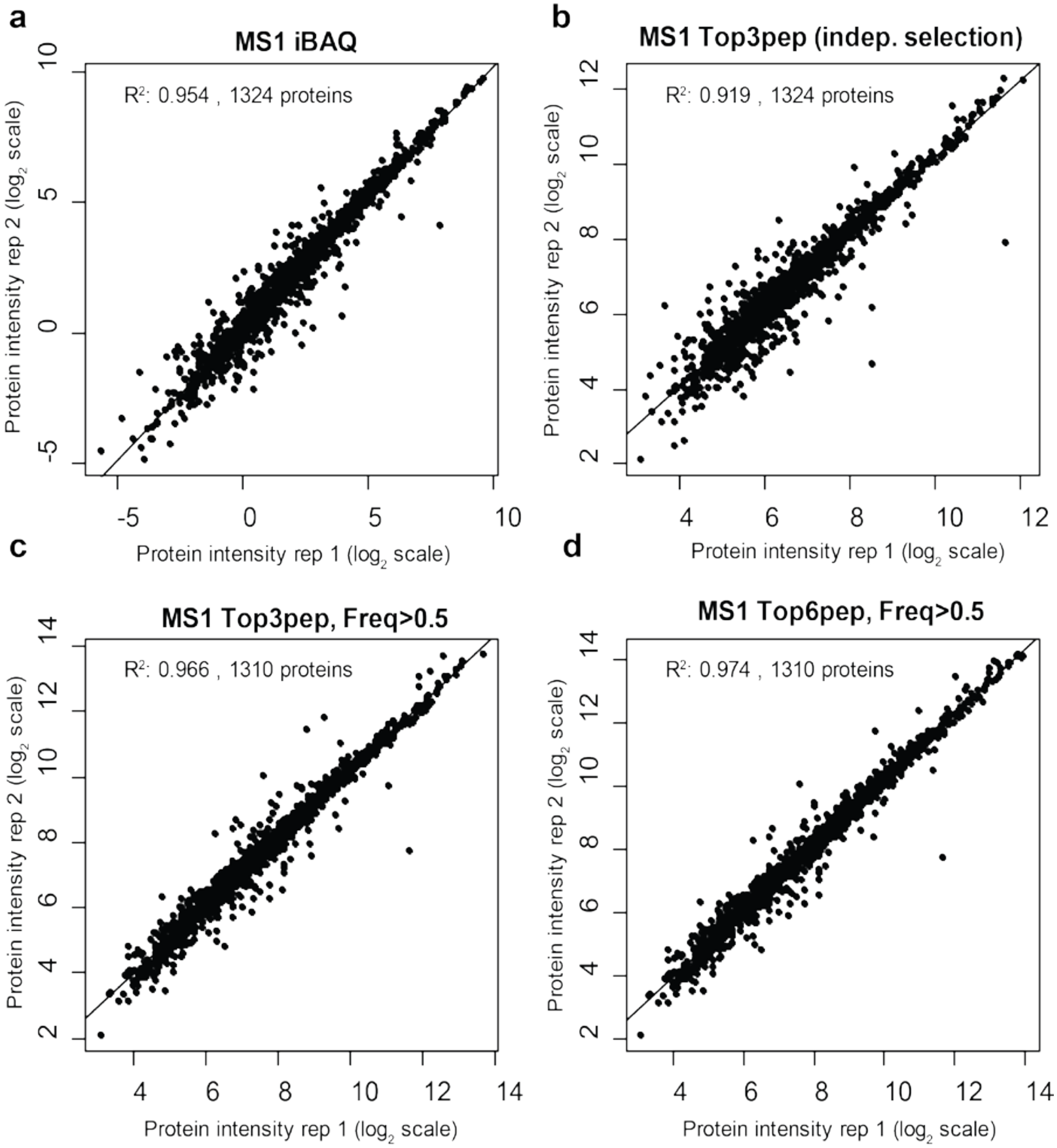


Figure 3-3 MS1-based protein quantification in DIA human cell lysate data.

(a) "MS1 iBAQ" intensity; **(b)** "MS1 Top3pep (indep. selection)": protein intensity estimated by summing the top three most intense peptide ions, independently in each DIA run; **(c)** "MS1 Top3pep, Freq>0.5": same as (b), with an additional requirement that the selected peptides are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein was identified; **(d)** "MS1 Top6pep, Freq>0.5": same as (c), but using six most intense peptides. Note that selection of consistently identified peptide ions (Freq > 0.5 filter) significantly improves the reproducibility of protein intensities between the replicates.

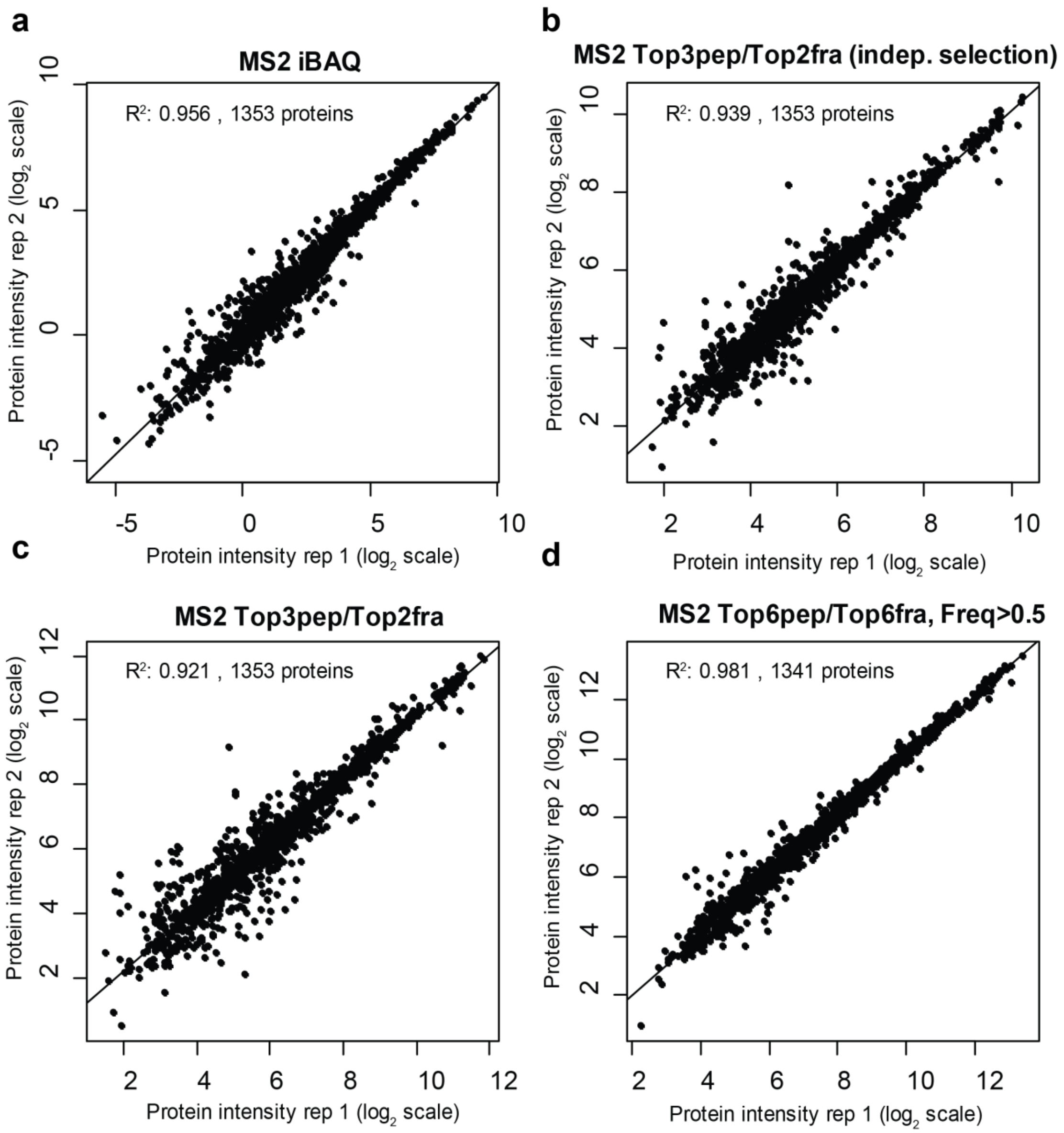


Figure 3-4 MS2-based protein quantification in DIA human cell lysate data.

(a) “MS2 iBAQ” intensity; **(b)** “MS2 Top3pep/Top2fra (indep. selection)”: protein intensity estimated by summing the top three most intense peptide ions. Peptide ion intensities are estimated by summing the intensities of their top 2 most intense matched fragments. Protein intensities are computed independently for each DIA run; **(c)** “MS2 Top3pep/Top2fra”: same as (b), but using peptide ions and fragments having the highest overall intensity across all runs (here, highest summed intensity across the two replicates); **(d)** same as (c), but using six selected peptide ions and fragments, with an additional requirement that the selected peptides and fragments are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein (peptide selection step) or peptide (fragment selection step) was identified. Note that selection of consistently identified peptide ions and fragments (Freq > 0.5 filter) significantly improves the reproducibility of protein intensities between the replicates.

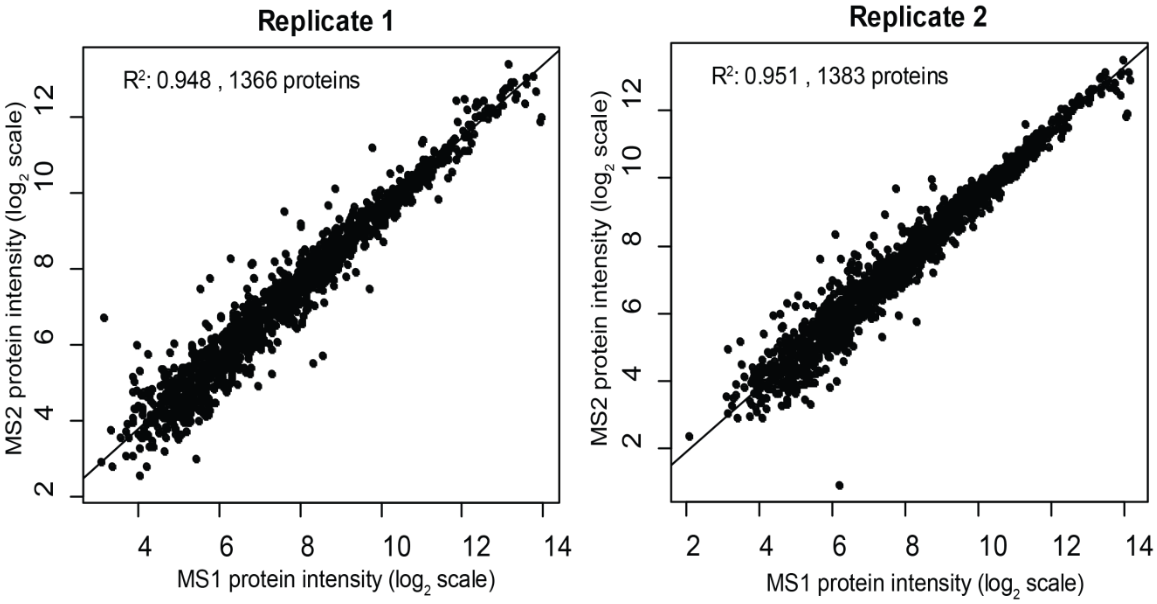


Figure 3-5 Comparison between MS1 and MS2-based protein quantification. Human cell lysate data.

MS1-based intensities are computed using the 'Top6pep, Freq>0.5' method. MS2-based protein intensities are computed using the "Top6pep/Top6fra, Freq>0.5" method.

Using the reproducibility of protein quantification across the two replicate runs as a benchmark measure, the MS2-based method with a stringent peptide and fragment selection procedure ("MS2 Top6pep/Top6fra, Freq > 0.5") outperformed the other methods considered (Figure 3-4). A similar MS1-based quantification metric ("MS1 Top6pep, Freq > 0.5") performed almost equally well, but with fewer (1,310 vs. 1,341) proteins quantified across both replicates (Figure 3-3). A good agreement was observed between these two (MS1 and MS2-based) abundance measures (Figure 3-5), further demonstrating the reliability of the feature detection and quantification algorithms in DIA-Umpire. In UPS2 standard protein sample, both

MS1 and MS2 quantification recovered the expected trend of differential abundance, suggesting that these measures are suitable for estimation of absolute protein abundances (Figure 3-6).

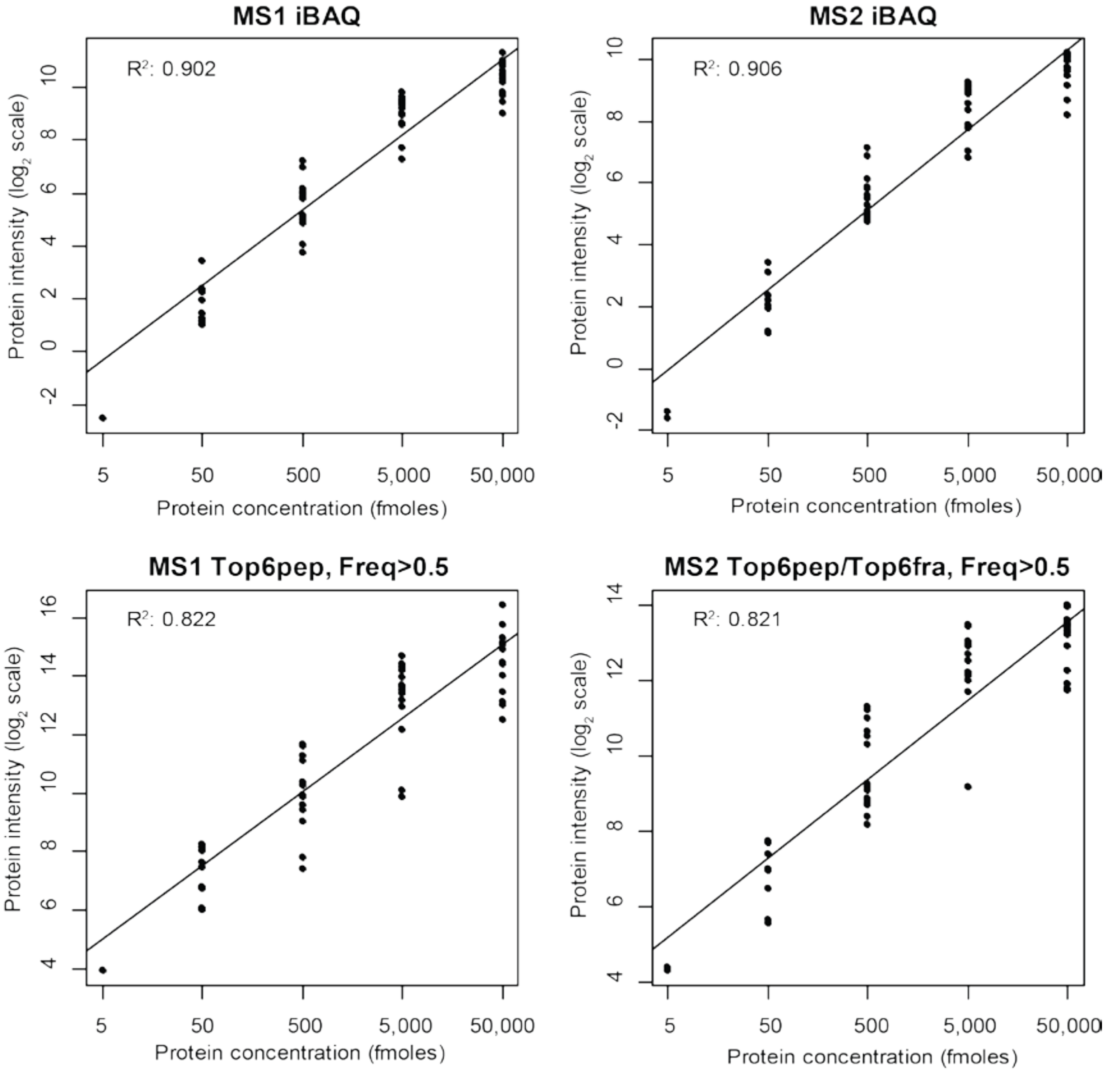


Figure 3-6 Protein quantification results in the UPS2 standard protein sample.

In UPS2 samples, protein concentrations are known and span five orders of magnitude (8 proteins in each of the five abundance bins; no proteins were quantified in the lowest abundance bin in this study). Proteins were quantified using four quantification methods. **(a)**"MS1 iBAQ" intensity; **(b)**"MS2 iBAQ" intensity; **(c)**"MS1 Top6pep, Freq>0.5": protein intensity estimated by summing the top six most intense peptide ions which are consistently identified (Freq > 0.5 filter); **(d)**"MS2 Top6pep/Top6fra, Freq>0.5" : protein intensity estimated by summing the top six most intense peptide ions with an additional requirement that the selected peptides and fragments are identified in more than 50% of the runs (Freq > 0.5) in which the corresponding protein (peptide selection step) or peptide (fragment selection step) was identified.

3.3.3 *Application of DIA-Umpire to interactome data*

A popular application of MS based proteomics is interactome analysis which involves in most cases the use of quantitative MS to monitor the relative abundance of a given protein in a bait purification experiment in comparison to negative controls. The coupling of affinity purification (AP) with targeted extraction strategies for SWATH analysis (AP-SWATH) was previously described [25, 26]. At the same time, a large number of scoring tools have been previously developed to assist identification of true interaction partners over background contaminants in DDA data [68], including Significance Analysis of INTeractome (SAINT) that performs such scoring using either spectral counts [69] or MS1 [66] intensity data. We reasoned that AP-SWATH data would provide a good test case for a complete analytical pipeline by demonstrating that DIA-Umpire in combination with SAINT can detect true interactors from DIA data, without the need for spectral libraries. We analyzed a dataset consisting of three biological replicates of the baits EIF4A2 and MEPCE and the negative GFP control analyzed by DIA [26] (Fig. 5a).

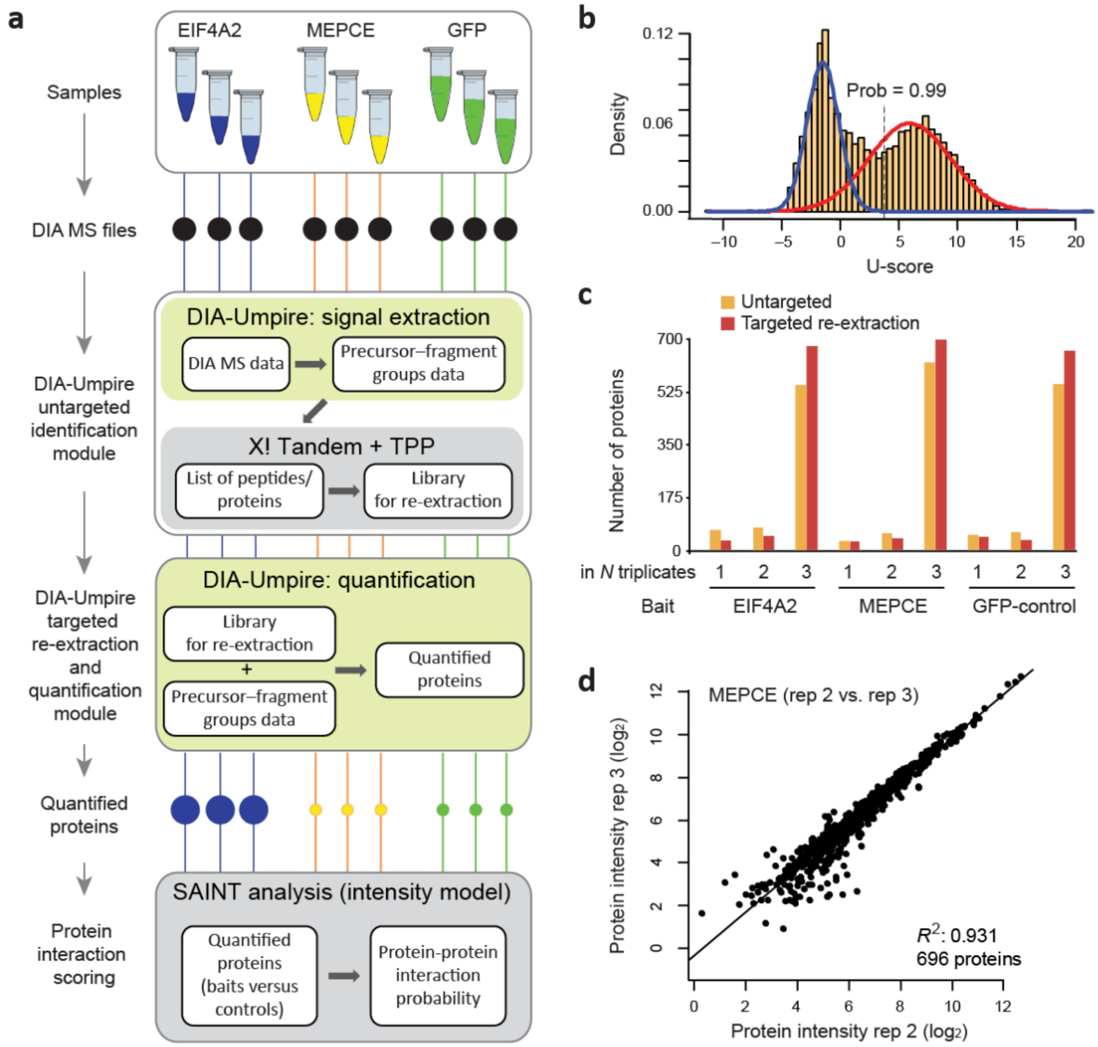


Figure 3-7 Application of the entire DIA-Umpire workflow to an AP-SWATH interactome data set.

(a) DIA-Umpire interatome analysis workflow. TPP, Trans-Proteomic Pipeline. **(b)** The distribution of scores (U-score) computed by the targeted re-extraction algorithm of DIA-Umpire. Data shown are from one biological replicate of an MEPCE AP-SWATH run. The observed distribution was modeled with the mixture modeling approach (blue curve, false identification model; red curve, correct identifications) to compute the posterior probability for each match. Peptide ions with a computed probability greater than 0.99 were considered confidently identified and contributed, together with the peptide ions identified at the initial, untargeted identification stage, to protein quantification for their corresponding proteins. **(c)** The numbers of proteins identified in one, two or all three biological replicates for each experiment after the initial, untargeted search and after targeted data re-extraction. **(d)** High reproducibility of protein intensities between two MEPCE AP-SWATH biological replicates computed by DIA-Umpire with the 'MS2 Top6pep/Top6fra, Freq > 0.5' quantification approach. rep, replicate.

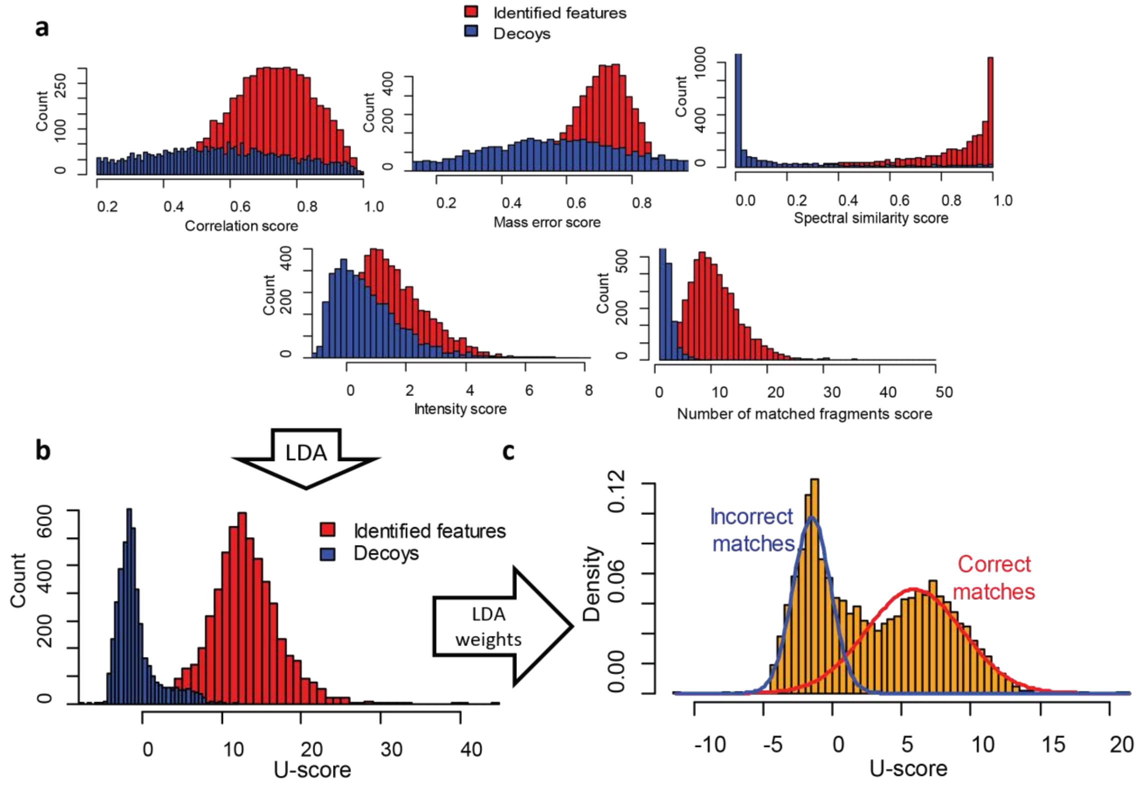


Figure 3-8 Distributions of scores computed by the targeted re-extraction algorithm of DIA-Umpire. AP-SWATH dataset, MEPCE bait (biological replicate 3).

(a) Distributions of the five sub-scores for peptide ions from the positive set ("Identified features", i.e. precursor-fragment group to spectrum matches that were confidently identified by the untargeted spectrum-centric search) and from the negative set (precursor-fragment groups matching to decoy spectra); **(b)** Linear discriminant analysis is used to train the weights in the linear combination used to combine the individual sub-scores to compute a single discriminant score (U-score). Shown are the resulting U-score histograms for positive and negative matches in the training set. **(c)** The final distribution of U-scores for all non-decoy matches. The observed distribution is model as a mixture of two underlying distributions representing high scoring, correct matches (red curve) and low scoring, incorrect matches (blue curve). The parameters of the distributions are learned using the expectation maximization mixture modeling algorithm. The posterior probability of a correct match is computed for a given non-decoy spectrum to precursor-fragments group match from the ratio of learned distributions among correct and incorrect matches. By default, peptide ions with a computed probability above 0.99 are considered confidently identified and contribute, together with the peptide ions identified at the initial untargeted identification stage, to protein quantification for their corresponding protein.

In the first step, we processed DIA data through DIA-Umpire in an untargeted manner, leading to identification and quantification of 3,900 - 4,900 peptide ions (600 - 700 proteins) in each AP-SWATH run. As expected, using targeted re-extraction with a stringent 0.99 peptide-centric identification probability threshold (Figure 3-8), we could identify and quantify additional peptide ions (1,300 - 2,300) and proteins (60 - 100) in each AP-SWATH run (Figure 3-9). Importantly, targeted re-extraction reduced the stochasticity issue, with an increase (by 19 - 23%) in the number of proteins quantified across all three replicates for the same bait (Figure 3-7c). Protein abundances were estimated using the "MS2 Top6pep/Top6fra, Freq>0.5" approach, with an excellent quantification reproducibility observed across the biological replicates for each bait and the GFP controls (Figure B-2).

Using SAINT [66] we recovered 45 significant interactors (SAINT probability above 0.95) for the EIF4A2 bait, a translation initiation factor implicated in the association of mRNAs to the ribosome. These proteins included 19 associated translation initiation factors (specifically the multi-subunit factors eIF3 and eIF4), the poly(A) binding protein which binds eIF4 and, as expected, several RNA helicases and RNA-binding proteins that are likely recruited via the mRNA. The ubiquitin protease USP10 (previously reported as an interaction partner for the eIF4A2 direct interactor eIF4G1) and casein kinase II subunits (which interact with eIF3) were also detected, alongside a known eIF4A inhibitor, PDCD4 [70]. A similar result was observed for the MEPCE bait – a protein that methylates the cap of the 7SK snRNA, leading to its stabilization [71]: 54 proteins were confidently scored as interactors, including well-characterized partners such as CDK9, Cyclin T, HEXIM1, METTL16, LARP7, SART1 and 3, several splicing components, and multiple components of the large, but not small, ribosomal subunit [26, 70, 71]. In summary, DIA-Umpire allows sensitive protein identification from DIA data, extraction of accurate quantitative information with less missing data, and is fully compatible with the existing interaction scoring methods such as SAINT, leading to the recovery of biologically-meaningful interactions.

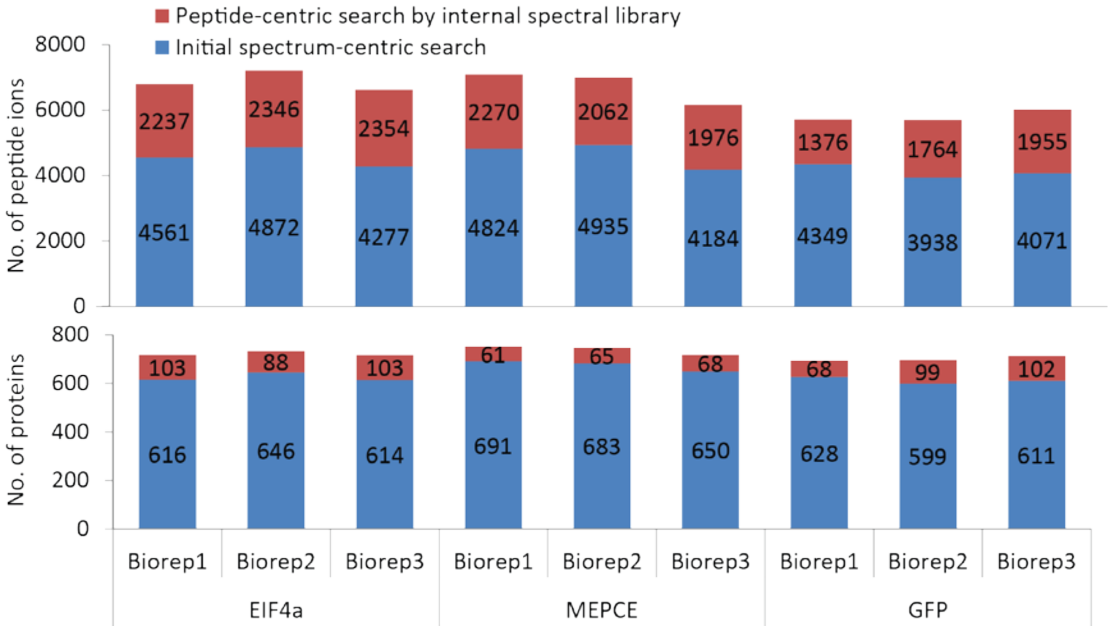


Figure 3-9 The numbers of identified proteins and peptide ions via untargeted spectrum-centric search and targeted re-extraction matching.

The numbers of identified peptide ions and proteins from the initial untargeted (spectrum-centric search) analysis using DIA-Umpire (blue bars) shown separately for 3 biological replicates (Biorep1…Biorep3) of the two bait proteins (EIF4A2 and MEPCE) and the GFP negative controls. Also shown (red bars) the numbers of additional peptide ions and proteins identified by the targeted re-extraction using the spectral library internally generated from the initial search results.

3.4 *Discussion*

In the second study we developed a targeted re-extraction approach for DIA quantification analysis. The targeted re-extraction searches DIA-Umpire preprocessed DIA data against an internal spectral library built from the initial identifications obtained by untargeted database search. We showed that using this hybrid approach, DIA-Umpire is able to perform more consistent identification and quantification across multiple DIA runs. In addition, we showed that such reproducible and reliable quantification is possible using both DIA MS1 and MS2 data with accurate protein quantification. We applied the whole DIA-Umpire pipeline to a publicly available AP-SWATH interatome DIA study data to show that the pipeline can capture sensitive protein-protein interaction profile without a need of prior spectral library for complete DIA analysis.

DIA-Umpire provides a complete pipeline for high throughput analysis of DIA data. The highly flexible design of the DIA-Umpire computational framework (importantly, with a full support for MS1 feature detection and quantification) should allow us to quickly adapt the algorithms to take advantage of new approaches and technological improvements, including emerging hybrid DIA/DDA strategies [72]. Finally, the pseudo MS/MS spectra generated by DIA-Umpire can be used to build spectral libraries for use with external tools, e.g. for visualization of spectra and precursor and fragment chromatograms in Skyline [29], or for targeted quantification using OpenSWATH, enabling an alternative solution to targeted re-extraction of quantitative information.

3.5 *Contributions*

The results presented in this chapter could not have been finished without the efforts of the great collaboration. I am grateful and fortunate to have such a great team to participate the project.

Chih-Chiang Tsou: developed the algorithms, implemented the software, designed experiments, analyzed the data, and wrote the manuscript draft.

Dmitry Avtonomov: assisted with the OpenSWATH analysis and contributed to the algorithm and software development and reviewed the manuscript.

Brett Larsen: acquired mass spectrometry data, designed experiments, analyzed data and provided inputs for manuscript.

Monika Tucholska: acquired mass spectrometry data.

Hyungwon Choi: assisted with SAINT scoring and contributed to the development of protein quantification strategies.

Anne-Claude Gingras: designed experiments, analyzed data, supervised the project, and wrote the manuscript.

Alexey I Nesvizhskii: conceived the project, developed the algorithm, designed experiments, analyzed the data, supervised the project, and wrote the manuscript.

Chapter 4 Improved DIA-Umpire pipeline for Untargeted, spectral library-free analysis of Orbitrap DIA data

The content of this chapter has been submitted to the special Issue of PROTEOMICS "Applications of targeted proteomics: from SRM to SWATH MS".

4.1 *Introduction*

Data independent acquisition (DIA) mass spectrometry is emerging as a promising alternative to data dependent acquisition (DDA) for quantitative proteomics analysis (for a recent review, see [73]), and is now available on most instrument platforms. As described in the previous chapters, DIA data is most commonly analyzed using spectral library-dependent tools such as OpenSWATH [28], Spectronaut [27], PeakView, and Skyline [29]. Ideally, a sample-specific spectral library is built from DDA experiments acquired in parallel with DIA data from the same or similar samples using same liquid chromatography system and same mass spectrometer. As we discussed, building a sample-specific spectral library requires additional sample materials and MS instrument time costs for the parallel DDA experiments. There have been continuing efforts to build comprehensive and publicly available spectral libraries based on a very huge number of DDA experiments on specific samples, e.g. the global spectral libraries deposited at SWATHAtlas repository (http://www.swathatlas.org/) which currently includes the global spectral libraries for different organisms [36, 74, 75], tissue-specific library [39], and human immunopeptidomes [40]. Therefore, alternatively one could use the global library instead of creating a sample-specific library for the DIA analysis.

In contrast to the spectral library-dependent tools, the alternative workflow, DIA-Umpire [41] is able to perform untargeted and direct (i.e. spectral library-free) analysis of DIA data using the existing MS/MS database search engines (e.g. X! Tandem [76], Comet [48], MSGF+ [77]) and peptide-spectrum match (PSM) statistical validation tools (PeptideProphet [50], Percolator [78], PeptideShaker [79]). We demonstrated that reliable quantification can be obtained from both fragment ion intensities and from MS1 data, and the targeted re-extraction workflow using internal spectral library allows more consistent identification and quantification analysis.

Because most of the recent studies used DIA (SWATH-MS) data generated using AB Sciex 5600 instruments, in this study we sought to test and improve the performance of DIA-Umpire using data generated using the Orbitrap family of mass spectrometers (Thermo Fisher Scientific) which also support acquisition of SWATH-like DIA data and other DIA variants [27, 80, 81] on the Orbitrap mass spectrometers. The Orbitrap mass analyzer, available in both Q Exactive and Fusion instruments, enables acquisition of tandem mass spectra with high mass accuracy and scan rate – the main prerequisites for successful interrogation of complex samples using DIA data [27, 80, 81]. Here we present DIA-Umpire v2, the new version of the software that enables analysis of complex DIA datasets generated using Orbitrap instruments without the need for a pre-existing spectral library. We describe the improvements made in the algorithms of DIA-Umpire, including the introduction of signal isotope pattern and fractional mass filters, the new targeted re-extraction scoring function, and the semi-parametric mixture modeling for computing the probabilities of correct identifications of peptide signals in DIA data using targeted re-extraction. Using the two series of Q Exactive DIA and DDA datasets published by Bruderer *et al*. [27], and a series of human HeLa cell line experiments on an Orbitrap Fusion mass spectrometer performed as part of this work we show that DIA-Umpire v2 enables highly sensitive analysis of DIA data. We also extend the improved targeted re-extraction module in DIA-Umpire v2 for external library search. We show that the external library search approach allows retention time alignment between DIA run and external spectral libraries without iRT peptides and the improved semi-parametric mixture modeling is able accurately calculate posterior probabilities for different libraries.

4.2 *Methods*

4.2.1 *Q Exactive datasets*

The raw files for two sets of Q Exactive DIA and DDA data described in [27] were downloaded from PeptideAtlas (http://www.peptideatlas.org; PASS00589). The first set was generated using HEK-293 cell lysates and the second set using human liver microtissue samples. All samples were analyzed using both DDA and DIA.

4.2.2 *Orbitrap Fusion datasets*

The MS system, Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose, CA), was coupled with an Ultimate 3000 RSLCnano system (Thermo Fisher Scientific). 1ug HeLa cell predigested using trypsin (Thermo Scientific, San Jose, CA) were directly loaded onto self-packed column. The 3 um ReproSil-Pur C18-AQ particles (Dr Maisch, Ammerbuch, Germany) were packed into a 30 cm self-pulled column with a 100 um inner diameter and 7 um opening to prepare an analytical column using "stone-arch" frit. The mobile phases consisted of (A) 0.1% formic acid and (B) 0.1% formic acid and acetonitrile. Peptides were separated through a gradient of up to 85% buffer B over 135 minutes at flow rate of 500 nL/min. The MS instrument was operated in the positive ion mode, with an electrospray through a heated ion transfer tube (250 °C), followed by a stacked ring ion guide (RF-lens) evacuated by a rotary vane pump to ~2 Torr. Full scan MS spectra were acquired in the Orbitrap mass analyzer (*m/z* range: 400–1250) with the resolution set to 60,000. Full scan target was 3e5 with a maximum fill time of 50 ms. All data were acquired in profile mode using positive polarity. MS/MS spectra of both DDA and DIA data were acquired in the Orbitrap as well with a resolution of 15,000 and higher-collisional dissociation (HCD) MS/MS fragmentation.

For DDA data, up to top 15 most intense ions were selected for MS/MS for each scan cycle. Target value for fragment scans was set at 1e5 with a maximum fill time of 35 ms and intensity threshold was kept at 2e4. Isolation width was set at 1.4 Th.

Two sets of independent DDA experiments (labeled DDA1 and DDA2) were acquired, each containing three replicate runs.

DIA experiments were run using different isolation window settings. A total of five DIA settings with 25, 20, 15, 10, and 5 Da SWATH-type fixed size isolation windows (resulting in 2.7, 3.3, 3.9, 6.2, and 13 seconds cycle time, respectively) were used to acquire the data. For each DIA experiment, the target value for fragment scans was set at 1e5 with a maximum fill time of 50 ms. Three replicates were acquired for each DIA experiment with one of the specified window sizes.

4.2.3 *Definition of datasets*

All DDA and DIA experiments were processed independently. False discovery rate (FDR) estimations at peptide ion or protein level, DIA internal library generation, and master protein list generation were done for each dataset separately. These datasets were defined as follows. The Q Exactive DIA (or DDA) datasets are referred to as 'HEK-293 DDA', 'HEK-293 DIA', 'Microtissue DDA', and 'Microtissue DIA' datasets. For Orbitrap Fusion DIA data, three replicates for each isolation window size setting were considered as part of the same dataset, referred to as 'DIA 5Da', 'DIA 10Da', 'DIA 15Da', 'DIA 20Da' and 'DIA 25Da'. The two independent sets of DDA data (each consisting of three replicates) were labeled 'DDA1' and 'DDA2' datasets.

4.2.4 *DIA-Umpire pseudo MS/MS extraction*

All .raw files were converted into mzXML format using msconvert.exe [82] with vendor (Xcalibur) peak picking option to generate centroid spectra. The DIA mzXML files were first processed by the signal extraction (SE) module of DIA-Umpire to generate pseudo MS/MS spectra in MGF format. For detection of precursor ion signal, the following parameters were used: 10 ppm mass tolerance for Orbitrap Fusion datasets and 15 ppm for Q Exactive datasets, charge state range from 1+ to 5+ for precursor ion detection in MS1 scans, and 2+ to 5+ for unfragmented precursor ion detection in MS2 scans. For detection of fragment ions in MS2 scans, 20 ppm mass tolerances for Orbitrap Fusion datasets and 25 ppm for Q Exactive datasets were used. Signal-to-noise ratio for both precursor and fragment signals was set to 1.1. The maximum retention time range was set to two minutes, and maximum of two consecutive gaps was allowed for detection of single *m/z* trace signals. Because the signal quality of the centroid spectra generated using Xcalibur library via msconvert.exe was deemed to be sufficiently high, no additional background detection and noise removal was used in DIA-Umpire_SE module. Because the MS2 scans in the resulting mzXML files contained isolation window ranges there was no need to specify isolation setting in the parameter file of DIA-Umpire_SE module.

4.2.5 *Filtering of detected features using fractional mass and isotope peak pattern*

The first step of DIA-Umpire analysis is extraction of precursor and fragment ion signals by the feature detection algorithm. DIA-Umpire v2 implements two new filters, the fractional mass filter and the isotopic pattern filter, to remove detected precursor ion and fragment features that are less likely to be true features.

Fractional mass filters have been used in a number of applications previously [83-85]. We adopted the fractional mass boundary equations described in Toumi *et al* [85] which was derived for human tryptic peptides. In order to allow modified peptides in the analysis, we extended the allowed fractional mass range by 2×d (d=0.1 used in this study). For each detected precursor ion or fragment ion feature with neutral mass M, the fractional mass $D(M)$ is calculated as

$$D(M) = M - \lfloor M \rfloor$$

The upper and lower bounds (the range of allowed fractional masses) of the fractional mass filter are derived according to the following equations, respectively:

$$H(M) = D\ (0.00052738 \times M + 0.066015 + d)$$

$$L(M) = D(0.00042565 \times M + 0.0003821 - d)$$

Finally, the binary classifier B(M) based on the fractional mass (1: accepted; 0: rejected) is determined as follows:

$$B(M) = \begin{cases} 1, \quad if\ H(M) \geq D(M) \geq L(M) \\ 1, \quad if\ H(M) < L(M) \wedge [D(M) \leq H(M) \vee D(M) \geq L(M)] \\ 0, \ \ \text{otherwise} \end{cases}$$

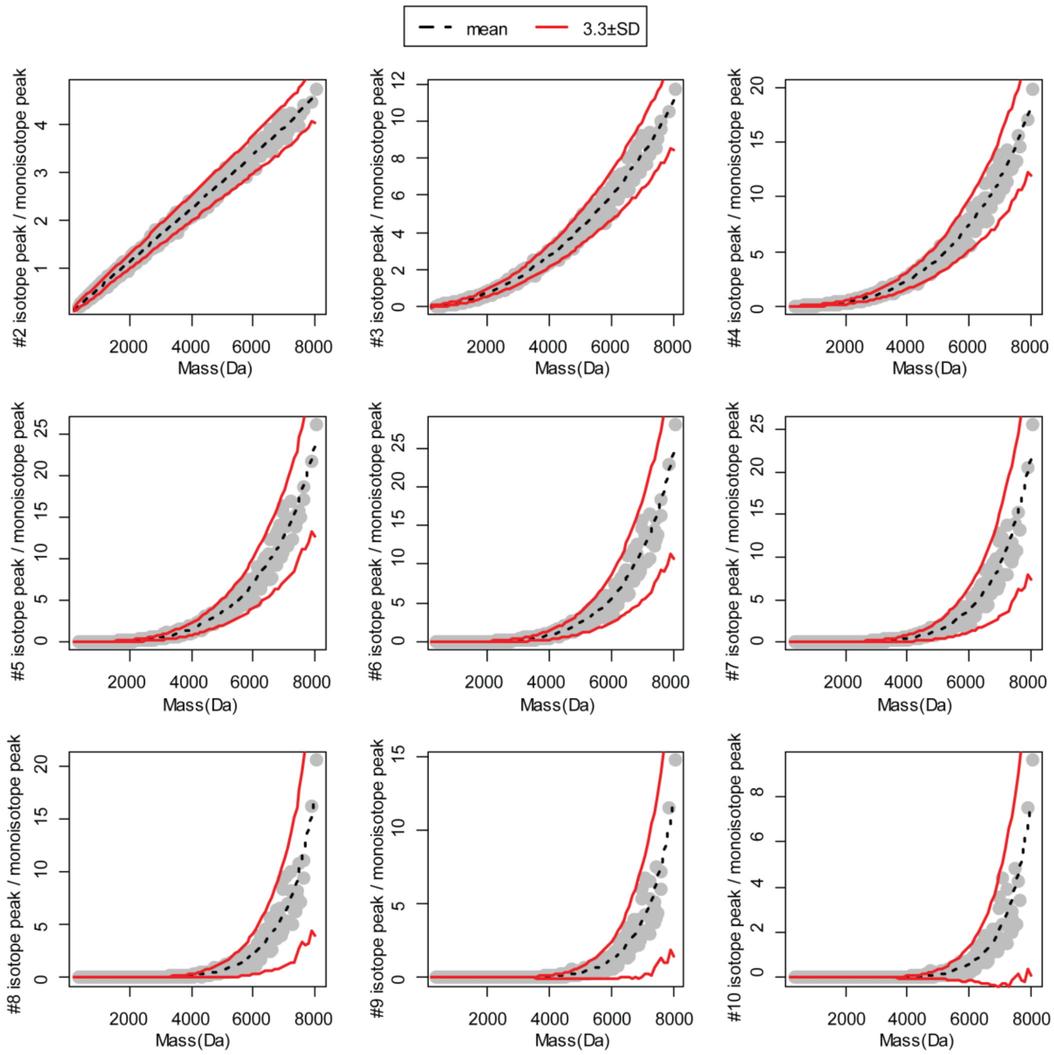


Figure 4-1 Theoretical intensity ratios of *i*th isotope peak over monoisotopic peak. Theoretical intensity ratios of *i*th isotope peak over monoisotopic peak. Grey dots represent isotope peak intensity ratio between *i*th isotope peak vs. monoisotopic peak for tryptic peptides generated from human proteome sequences. In each plot, the grey dots were partitioned into 100 Da mass bins and mean and standard deviation (SD) for each bin were calculated. The black dash lines are the mean values of each 100 Da mass bin, and red solid lines represent the boundary for each bin calculated by mean ± 3.3 standard deviations.

Second, an isotope pattern filter has been introduced to remove precursor features showing a poor fit between the observed and the theoretical isotope peak distributions. Theoretical isotope peak intensity ratios given peptide molecular weights calculated from all human tryptic peptides. The isotope peak ratios up to the 10th isotope peak were established in DIA-Umpire by generating 9 scatter plots (Figure 4-1).

To determine the boundary of the theoretical isotope ratios, the mean (μ) and standard deviation (σ) of each 100 Da bin in each plot were calculated. The 99.8% ($\pm 3.3\times\sigma$) confidence intervals were then selected to represent the boundaries for each bin (plotted in Figure 4-1). For a possible peak feature detected with peak intensities I = (I_1, I_2, ..., I_n) and neutral mass M, the observed peak ratios O = (O_2, ..., O_n), $O_i = I_i$ / I_1, were calculated, where n is the number of isotope peaks. Then the mean μ_i and the standard deviation σ_i of the closet mass bin for M from i^{th} scatter plot (corresponding to i^{th} isotope ratio) were extracted, and the boundary (H_i, L_i) of the expected peak ratio was calculated as follows: $H_{\mathrm{i}} = \mu_i + 3.3\times\sigma_i$ and L_i=μ_i-3.3×σ_i. Then the isotope pattern fitness probability score between the observed peak ratio and the theoretical peptide isotope distribution was estimated as $1 - C(X^2, n-1)$, where $C(X^2, n-1)$ is the standard Chi-Squared probability cumulative distribution function, and X^2 is Chi-Squared value calculated as follows:

$$X^2 = \sum_{i=2}^{n} \frac{(O_i - E_i)^2}{E_i^2}$$

$$E_i = \begin{cases} O_i, if\ O_i \geq L_i\ and\ \ O_i \leq H_i \\ H_i, if\ \ O_i > H_i \\ L_i, if\ \ O_i < L_i \end{cases}$$

In this study, all detected features with isotope pattern fitness probability score below 0.3 were removed.

4.2.6 *DDA and DIA (pseudo) MS/MS database search*

The DDA and DIA pseudo MS/MS spectra extracted using DIA-Umpire were searched using X! Tandem, Comet and MSGF+ search engines using the following parameters: allowing tryptic peptides only, up to one missed cleavage, methionine oxidation specified as variable modification, and cysteine carbamidomethylation as static modification. The precursor ion mass tolerance and the fragment ion mass tolerance were set, respectively, to 10 ppm and 20 ppm for Orbitrap Fusion data and to 15 ppm and 25 ppm, respectively, for Q Exactive data. The data were searched against a non-redundant human protein sequence FASTA file extracted from the UniProtKB/Swiss-Prot database (release date: June 19, 2015), appended with the corresponding reversed sequences as decoys for target-decoy analysis. The output files from each search engines were further analyzed by PeptideProphet, and the results were combined using iProphet [52] followed by ProteinProphet [2].

4.2.7 *FDR estimation independently for each DDA/DIA run*

The false discovery rate (FDR) for peptide ion (i.e. unique combination of peptide sequence, charge state, modification and modification site parameters) and protein identifications was first estimated independently for each individual run. For each individual run (e.g. Orbitrap Fusion DIA 5Da window, Replicate 1; denoted as ‘DIA 5DA R1’), FDR at the peptide ion level was estimated by sorting the identifications using the iProphet computed peptide ion probability followed by the selection of the probability threshold corresponding to 1% FDR based on the target-decoy strategy [1]. The number of peptide ions at 1% FDR determined independently for each run (column "Peptide ion IDs (1% FDR run level)") are shown in Table C-1 (Q Exactive HEK-293 data), Table C-2 (Q Exactive liver microtissue data), and Table C-3 (Orbitrap Fusion data). At the protein level, protein groups assembled by ProteinProphet for each run independently were sorted using the maximum peptide ion iProphet probability taken as the protein-level score, followed by target-decoy based FDR estimation. The numbers of protein groups determined independently for each run at 1% FDR are also shown in Table C-1, Table C-2, and Table C-3 (column "Protein IDs (1% FDR run level)").

4.2.8 *FDR for peptide ion identifications in DDA data at the dataset level*

In addition to estimating FDR at individual run level, FDR for DDA data was also estimated at the dataset level. In the dataset level FDR strategy, the list of peptide ions was filtered to achieve 1% FDR for the entire dataset (e.g. Orbitrap Fusion 'DDA1' dataset consisting of the three replicate runs 'DDA1 R1', 'DDA1 R2', and 'DDA1 R3'). If the peptide ion passed the desired FDR threshold (here 1%) at the dataset level, then all identifications of that peptide ion in each individual run within the same dataset were counted as identified in that run. Such a filtering strategy is useful for reducing the number of missing values in each individual run (which is important for achieving more complete quantification matrix across the dataset), while maintaining the desired FDR at the dataset level. It also allows more fair comparison of DDA numbers with DIA numbers after the second, targeted re-extraction step using the spectral library build from all identified spectra in the dataset (see below). The number of peptide ion identifications for each DDA runs determined using the dataset level FDR strategy is shown Table C-1, Table C-2, and Table C-3 (column "Peptide ion IDs (1% FDR dataset level)").

4.2.9 *FDR for protein identifications in DDA data at the dataset level*

To estimate protein FDR for DDA data at the dataset level, ProteinProphet [2] was used to assemble protein groups for each dataset taking pepXML files for all replicate runs from the same dataset as input. Protein FDR was estimated by the target-decoy approach based on the maximum peptide ion probability across the files within a dataset. At 1% FDR, a master protein list for each dataset was first generated. For each protein (representing a protein group) in the master list, that protein was considered identified in an individual run if it had at least one peptide ion identified in that run that was in the dataset level 1% FDR filtered list. The number of protein identifications for individual DDA runs counted using the dataset level FDR strategy is shown Table C-1, Table C-2, and Table C-3 (column "Protein IDs (1% FDR dataset level)").

4.2.10 *Generation of the spectral library for targeted re-extraction in DIA data*

The analysis of DIA data using DIA-Umpire includes an additional targeted data extraction step using the internal spectral library build from the peptides identified using the initial, untargeted analysis. In each DIA dataset, all peptide ion identifications passing 1% dataset-level FDR (estimated as described above for DDA data) were taken as input into the DIA-Umpire target re-extraction module (DIA-Umpire_Quant.jar) to generate an internal spectral library and perform targeted re-extraction analysis [41] to further reduce missing quantifications for each DIA dataset. For building consensus spectra in the internal spectral library, an option has been added in DIA-Umpire v2 to use the fragment selection algorithm described in Tsou *et al* [41]. With this option enabled, the consensus spectrum for each peptide ion is created using the *TopN* best fragments selected across all runs within the dataset (top six fragments in this study). The algorithms for building consensus spectra, retention time prediction, and mass calibration in DIA-Umpire v2 remained the same as described in Tsou *et al* [41].

4.2.11 *Targeted re-extraction scoring function*

Several components of the scoring function for the targeted re-extraction step were revised, and thus described here in more details. A precursor-fragment group G generated by DIA-Umpire, and a library spectrum S, represented as

$$\mathrm{S} = \{(I_1^{\mathrm{S}}, M_1^{\mathrm{S}}), (I_2^{\mathrm{S}}, M_2^{\mathrm{S}}), \dots, (I_{\mathrm{NS}}^{\mathrm{S}}, M_{\mathrm{NS}}^{\mathrm{S}})\}$$

$$\mathrm{G} = \{(I_1^{\mathrm{G}}, M_1^{\mathrm{G}}, C_1^{\mathrm{G}}, T_1^{\mathrm{G}}), (I_2^{\mathrm{G}}, M_2^{\mathrm{G}}, C_2^{\mathrm{G}}, T_2^{\mathrm{G}}), \dots, (I_{\mathrm{NG}}^{\mathrm{G}}, M_{\mathrm{NG}}^{\mathrm{G}}, C_{\mathrm{NG}}^{\mathrm{G}}, T_{\mathrm{NG}}^{\mathrm{G}})\}$$

where NS and NG are the numbers of fragment peaks in the library spectrum and in the precursor-fragment group, respectively (NS ≤ 6 in this study). $I_r{}^S$ and $M_r{}^S$ are the intensity and the theoretical *m/z* value, respectively, of a fragment *r* that belongs to the library spectrum S. Similarly, I_r^G and M_r^G are the intensity and *m/z* value, respectively, of a fragment *r* that belongs to the precursor-fragment group G. C_r^{G} and T_r^{G} are the Pearson correlation coefficient and peak apex retention time difference, respectively, between the peak profiles of a fragment *r* and the precursor anchoring group G. All negative Pearson correlation coefficients were set to 0. A matching intensity vector $\mathrm{INT}^{\mathrm{G\text{-}S}} = (I_1^{\mathrm{G}}, I_2^{\mathrm{G}}, \ldots, I_{\mathrm{NS}}^{\mathrm{G}})$ of length NS, with I_r^{G} taken as the intensity of the fragment peak *r* in G that matches to a fragment in S, and as zero if no fragment peak can be found in G within the specified mass tolerance (in ppm units) window $\mathrm{D_M}$ around M_r^{S}. Thus, $\mathrm{INT}^{\mathrm{G\text{-}S}}$ contains *L* non-zero values, where *L* is the total number of matched fragments between G and S. The following nine sub-scores are calculated during the spectral matching:

1. Spectral Similarity Score (SSS), in DIA-Umpire v2 calculated as the dot product [86] between the vector $\mathrm{INT}^{\mathrm{G\text{-}S}}$ and the library spectrum intensity vector $(I_1^{\mathrm{S}}, I_2^{\mathrm{S}}, \ldots, I_{\mathrm{NS}}^{\mathrm{S}})$.
2. Mass Error Score (MES):

$$\mathrm{MES} = 1 - \frac{\sum_{j=1}^{L} \mathrm{PPM}(M_j^{\mathrm{G}}, M_j^{\mathrm{S}})}{\mathrm{D_M} \times L}$$

$$\mathrm{PPM}(m_a, m_b) = \frac{|m_a - m_b| \times 2 \times 10^6}{m_a + m_b}$$

3. Correlation Score (CS):

$$\mathrm{CS} = \frac{\sum_{j=1}^{L} C_j^{\mathrm{G}}}{L}$$

The scores described above are essentially the same as described in the original DIA-Umpire manuscript, except that SSS is computed using the dot product instead of the Pearson correlation. In addition, the following six new scores are introduced:

4. Apex Delta Score (ADS):

$$\mathrm{ADS} = \frac{\sum_{j=1}^{L}\left|T_j^{\mathrm{G}}\right|}{L}$$

5. Weighted Number of matched Fragments (WNF):

$$\mathrm{WNF} = \sum\nolimits_{j=1}^{L} C_j^{\mathrm{G}} \times (1 - \frac{\mathrm{PPM}(M_j^{\mathrm{G}}, M_j^{\mathrm{S}})}{\mathrm{D_M}})$$

6. Retention time difference between the predicted retention time and the observed monoisotopic peak apex of the precursor peptide anchoring precursor-fragment group G.
7. Precursor isotope peak correlation score, computed as the Pearson correlation coefficient between the monoisotopic peak elution profile and the second isotope peak profile of the precursor anchoring group G (set to zero if the correlation is negative).
8. Precursor isotope pattern fitness probability score, calculated as described above.
9. Difference between the experimental mass of the precursor anchoring group G and the theoretical mass of the peptide ion in the internal library.

The final match score (U-score) between S and G is calculated as a linear combination of all the nine sub-scores described above. The linear combination coefficients were trained as described for DIA-Umpire previously [41].

4.2.12 *Posterior probabilities of correct identification at the targeted re-extraction step*

The probability calculation in DIA-Umpire v2 has been revised to implement a more robust semi-parametric mixture modeling. For each library spectrum S, let U be the best final match score (U-score described above) of all candidates in the searched range for S. The observed distribution of scores for all spectra in a particular run searched at the targeted extraction step, *f*(U), is a joint distribution of correct and incorrect identifications, i.e. $f(U) = \pi_0 f_0(U) + \pi_1 f_1(U)$, where f_0 and $f_1$are the respective distributions of incorrect and correct identifications, and π_0 and π_1 are the priors (proportions of true and false matches), where $\pi_0 + \pi_1 = 1$. To estimate the distributions f_0 and f_1, DIA-Umpire v2 implements the semi-parametric density estimation similar to that of Robin *et al* [87], which has been described for PSM validation by Choi *et al* [88] and implemented in PeptideProphet ('P' option) and in iProphet. The idea behind the semi-parametric mixture modeling is to use decoy identifications (that are known to be false) to first represent f_0, and f_1 can then be deconvoluted using the expectation maximization (EM) algorithm with a modified kernel density estimation. The first step of this mixture modeling is to estimate π_0 to avoid the over-fitting problem (maximum likelihood will be always at the point when $\pi_1$equals to 1 [87]) in the EM algorithm. $\pi_0 = {F(q)}/{F_d(q)}$, where

$F(.)$ and $F_d(.)$ are respective CDFs of empirical distributions of target and decoy identifications, and q is the mean score of decoys. The priors π_0 and π_1 estimated this way are then fixed throughout the EM algorithm. The kernel density estimation of distributions $f(U)$ and $f_0(U)$ are obtained by the following equations:

$$f(U|h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{U - U_i}{h})$$

$$f_0(U|h) = \frac{1}{n_d h}\sum_{i=1}^{n_d} K(\frac{U - U_i}{h})$$

where K is the Gaussian density function, and n and n_d are the numbers of identifications from all target library spectra and decoy spectra, respectively. The bandwidth parameter h is estimated using the Silverman's rule of thumb [89]. The initial estimation of $f_1(U)$ is done by the DIA-Umpire's original Gaussian mixture modeling approach [41]. In the E-step of the EM mixture modeling algorithm, the probability $p(U_i)$ of score U_i for spectrum S_i is calculated as

$$p(U_i) = \frac{\pi_1 f_1(U_i)}{f(U_i)}$$

Then in the M-step the kernel density estimation of correct distribution is updated as

$$f_1(U) = {\sum_{i=1}^{n}[p(U_i) \times K\left(\frac{U - U_i}{h}\right)]}\Big/{h\sum_{i=1}^{n} p(U_i)}$$

The EM algorithm iterates until the difference of log-likelihoods between two consecutive iterations is less than 0.00001 of the initial log-likelihood or the EM algorithm has reached 50 iterations. Once the EM algorithm is finished, the final π_0 and π_1 are updated by the following equations:

$$\pi_1 = \frac{1}{n}\sum_{i=1}^{n} p(U_i)$$

$$\pi_0 = 1 - \pi_1$$

Given a U-score U_i, the final probability is calculated as described above the updated priors.

4.2.13 *Combing untargeted and targeted re-extraction identification results*

DIA-Umpire v2 exports additional identifications obtained at the targeted re-extraction step into separate pepXML files. In order to be able to estimate FDR with inclusion of these additional identifications, decoy identifications and their probabilities are exported as well. Note that, for consistency, DIA-Umpire prints the corresponding reversed sequences in the resulting targeted re-extraction pepXML files for all decoy identifications, even though the actual spectra representing those decoys in the internal library were obtained using the shuffling approach. For each identification obtained at the targeted re-extraction step, DIA-Umpire v2 prints the U-score probability calculated as described above (which are labeled as iProphet probability in the generated pepXML files). These steps allow the protein inference algorithm of ProteinProphet to combine the results (pepXML files), including decoy identifications, from the initial untargeted database search step with the results generated using targeted re-extraction using the internal spectral library (see below).

4.2.14 *FDR for peptide ion identifications in DIA data at the dataset level*

As with DDA data, in addition to estimating FDR at individual run level, FDR for DIA data was also estimated at the dataset level. The list of peptide ions identified at the untargeted step was filtered to achieve 1% FDR for the entire dataset (e.g. Orbitrap Fusion 'DIA 5 Da' dataset consisting of the three replicate runs 'DIA 5Da R1', 'DIA 5Da R1', and 'DIA 5Da R3'). If the peptide ion passed the desired FDR threshold (here 1%) at the dataset level, then all identifications of that peptide ion in each individual run within the same dataset were counted as identified in that run. Peptides that were not identified in a particular run based on the untargeted analysis alone, but that were detected in that run using targeted re-extraction with a high probability (here, 0.99 or higher), were also counted as identified. It should be noted that inclusion of identifications from the targeted re-extraction step does not change the dataset level FDR, set to 1%. The number of peptide ion identifications for each DIA run is shown in Table C-1, Table C-2, and Table C-3 (column "Peptide ion IDs (1% dataset level FDR)").

4.2.15 *FDR for protein identifications in DIA data at the dataset level*

For estimating protein FDR at the dataset level for DIA data (after targeted re-extraction), ProteinProphet [2] was run for each dataset independently taking all pepXML from the untargeted (database search) step and from the targeted re-extraction step as input. FDR was then estimated using the target-decoy approach [1] based on the maximum peptide ion probability (iProphet probability from the untargeted database search step or the probability based on U-score from the targeted re-extraction step, also labeled as iProphet probability in the pepXML files as explained above). A master protein list corresponding to 1% FDR for each dataset was first generated. A protein in the master list was then considered identified in an individual run if it had at least one peptide ion identified in that run that was included in the 1% FDR set at the dataset level, or if there was a peptide ion identified with probability 0.99 or higher at the targeted re-extraction step. The number of protein identifications obtained this way is shown in Table C-1, Table C-2, and Table C-3(column "Protein IDs (1% dataset level FDR)").

4.2.16 *Targeted re-extraction analysis using external libraries*

The targeted re-extraction module in DIA-Umpire v2 is also compatible with external library search. As described above, the purpose of the internal library search is to reduce the number of missing identifications for individual DIA runs, and it does not increase the total number of peptide ion identifications for the entire dataset. If there is an external spectral library which contains additional peptide ions, using the sensitive targeted re-extraction scoring, additional peptides could be identified. In this study, we used the Q Exactive HEK-293 dataset to test the performance of external library search. We used the DDA sample-specific library (built from HEK-293 parallel DDA experiments), which was published with the original paper [27], and a human combined assay library (Human-CAL) [36] deposited at SWATHAtlas repository (phl004_canonical_s64_osw_decoys.TraML). The targeted re-extraction module of DIA-Umpire requires decoy spectra for semi-parametric mixture modeling and posterior probability calculation. But the DDA sample-specific library which the authors provided with the original manuscript did not contain decoy spectra. Therefore we used the "shuffle-and-reposition" method [63] to generate a decoy spectrum for each peptide ion for the DDA sample-specific library (29,243 peptide ions). For the Human-CAL library, because it contains decoy spectra generated by OpenSWATH pipeline, we directly used the decoy spectra provided in the TraML file.

The workflow of DIA-Umpire external library search is shown in Figure 4-2. For each DIA run, a peptide ion from an external library is classified as an "identified peptide ion" if it was confidently identified either by database search (in 1% FDR set at the dataset level) or by internal library search (U-score probability equal or higher than 0.99), otherwise the peptide ion is classified as a "target peptide ion". First, all the identified peptide ions are used to build a retention time alignment model using the nonlinear regression approach described previously [41]. The retention time alignment model represents the retention time mapping from normalized retention time scale in the external library to the observed retention time scale in the particular DIA run. For each target peptide ion, a predicted retention time and a corresponding prediction variance in the DIA run are calculated by the retention time alignment model given the normalized retention time of the target peptide ion in the external library (for examples of retention time alignment model, please see Figure 4-10B and Figure 4-10D). All the precursor-fragment groups within the retention time range (RT width determined by the prediction variance) and the precursor *m/z* range (15 p.p.m) are scored against the library spectrum of the target peptide ion. The calculated U-scores of all target peptide ions are then deconvoluted by the semi-parametric mixture modeling to compute posterior probabilities of true identifications. In this study we used 0.99 as the probability threshold to determine confident identifications from the external library searches.

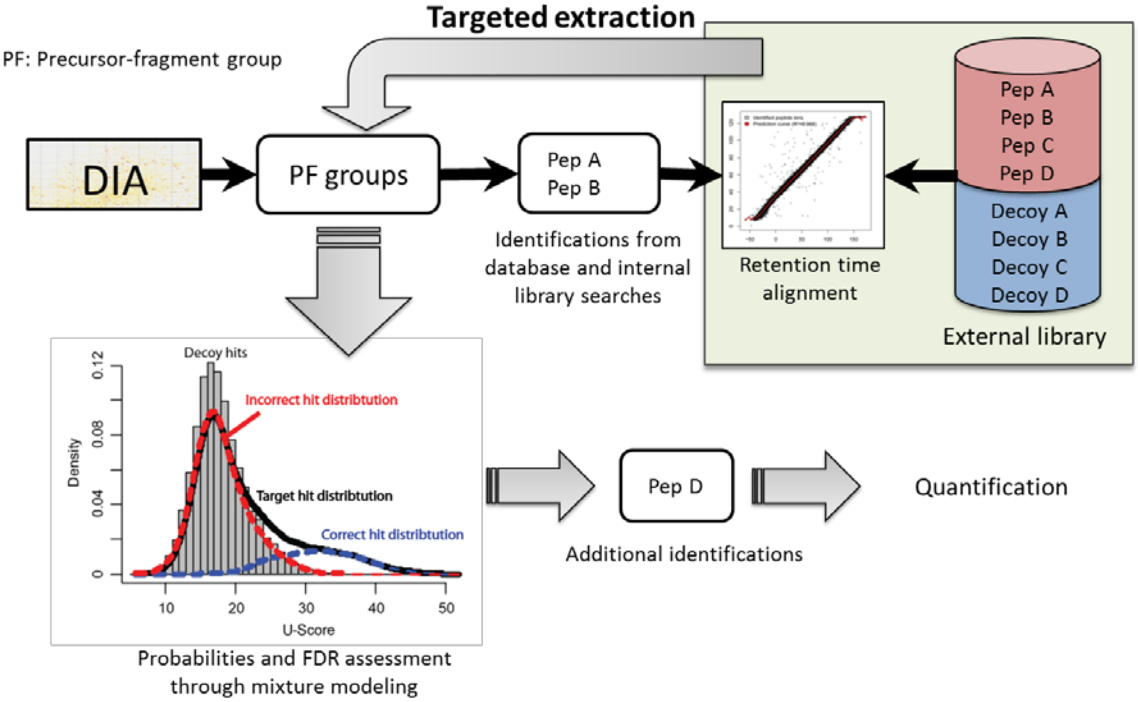

Targeted extraction
PF: Precursor-fragment group
DIA
PF groups
Pep A
Pep B
Identifications from database and internal library searches
Retention time alignment
Pep A
Pep B
Pep C
Pep D
Decoy A
Decoy B
Decoy C
Decoy D
External library
Decoy hits
Incorrect hit distribtution
Target hit distribtution
Correct hit distribtution
Density
0
0.04
0.08
0.12
10
20
30
40
50
U-Score
Probabilities and FDR assessment through mixture modeling
Pep D
Additional identifications
Quantification

Figure 4-2 Workflow of DIA-Umpire external library search

4.2.17 *Data availability*

All Orbitrap Fusion mass spectrometry data files and DIA-Umpire results for all the datasets presented in this paper have been deposited at the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/) via the PRIDE partner repository with the data set identifier PXD003179.

4.3 *Results*

4.3.1 *Improved feature detection using fractional mass and isotope pattern filters*

The DIA-Umpire workflow relies on detection of precursor and fragment ion signals by the feature detection algorithm. The sensitivity of the feature detection algorithm is a key factor for successful extraction of pseudo MS/MS spectra and subsequent untargeted peptide identification using database search. To increase the number of identifications, minimal filtering criteria can be applied to extract as many features as possible. In doing so, false (noise) features do not always negatively affect the results because MS/MS database search with FDR filtering can effectively eliminate randomly assembled pseudo MS/MS spectra. However, it is not always practical to consider all possible features because the overall computation costs (time and memory usage) increase with the number of features extracted from the data. In large datasets, this could become an issue, especially for the processing steps of DIA-Umpire's precursor-fragment grouping algorithm and for MS/MS database searching. Therefore, one challenge for the untargeted feature detection approach of DIA-Umpire is to find a reasonable balance between the number of missed signals and the total computation costs. To address this issue, we introduced two new filters in DIA-Umpire v2 to remove detected precursor ion and fragment features that are less likely to be true peptide features (see 4.2 for details).

We first investigated the effects of the feature detection filters using two DIA runs, one from the Orbitrap Fusion 10 Da isolation window width dataset generated as part of this work, and the other from the publicly available HEK-293 Q Exactive DIA dataset. We processed these two DIA runs through DIA-Umpire signal extraction module without any filtering to maximize the number of detected precursor features. The pseudo MS/MS spectra extracted by DIA-Umpire were then searched using X! Tandem, Comet, and MSGF+ search engines, and the results from all three search engines were further combined by iProphet analysis. Peptide ion identifications were filtered to achieve 1% peptide ion level FDR (see 4.2 for details regarding MS/MS database search and FDR calculations). All confidently identified peptide ion were linked to the corresponding detected precursor features.

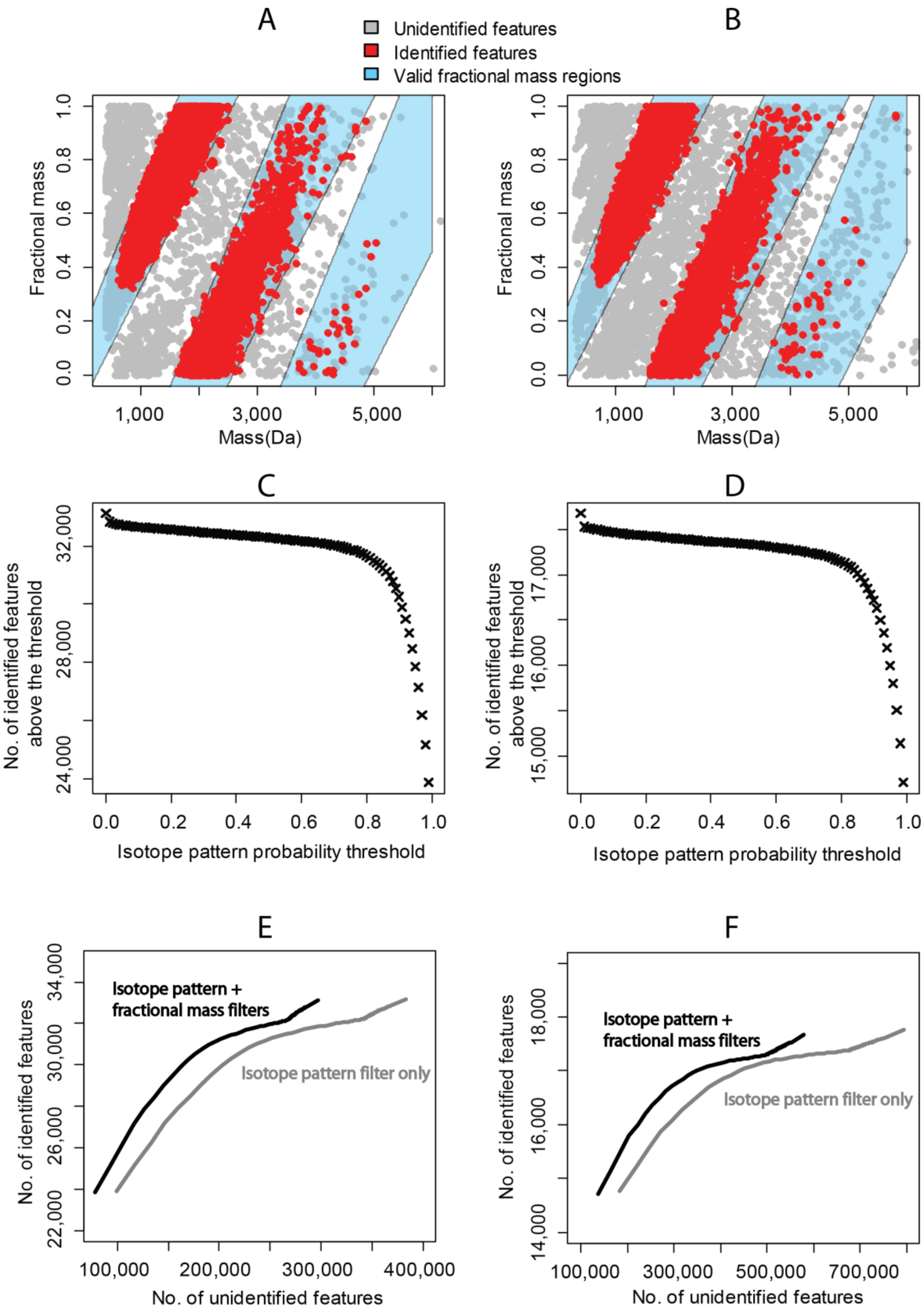


Figure 4-3 Effects of feature detection filtering.

(A) The fractional mass of detected precursor features from the first replicate of the Orbitrap Fusion DIA 10 Da dataset. The grey and red dots represent unidentified and identified features, respectively. Blue regions are the valid regions of the fractional mass filter. **(B)** Same as (A), results for the first replicate of HEK-293 Q Exactive dataset. **(C)** The number of identified precursor features at different isotope pattern fitness probability thresholds, the Orbitrap Fusion data. **(D)** Same as (C), the Q Exactive data. **(E)** The results of applying the isotope pattern filter alone or combination with the fractional mass filter, the Orbitrap Fusion data. **(F)** Same as (E), the Q Exactive data.

In total, there were 416,607 and 812,944 precursor features detected in the Orbitrap Fusion and Q Exactive runs, respectively. Of these, only 33,173 (7.9%) and 17,759 (2.1%) features were identified at 1% FDR threshold, respectively, in these two datasets. Figure 4-3A and Figure 4-3B plot the fractional masses of the identified and unidentified features in different mass ranges for the two DIA runs, with the valid fractional mass regions (*d*=0.1) highlighted in blue. Clearly, the fractional masses of almost all of the identified features were in the valid fractional mass regions. We then applied the fractional mass filter, which effectively removed 86,845 (22.6%) and 215,509 (27%) of the unidentified features for the Orbitrap Fusion and the Q Exactive run, respectively, at a loss of only 0.13% and 0.45% of true identifications for the Orbitrap Fusion run and the Q Exactive run, respectively.

As for the isotope pattern filter, Figure 4-3C and Figure 4-3D show the numbers of identified precursor features remained at different isotope pattern probability thresholds. The majority of the identified features had an isotope pattern probability of 0.8 or higher (95.6 % for the Orbitrap run and 96.9 % for the Q Exactive run). However, there were a small number of identified peptide ions which had extremely low isotope pattern probabilities. Some of these cases may be due to co-elution with other high abundance peptide ion signals, whereas others could be false identifications. Overall, the isotope pattern probability threshold was found to be useful for achieving the right balance between the identification sensitivity and the computational cost.

By combining the two filters, the fractional mass filter and the isotope pattern filter, it becomes possible to effectively reduce the number of extracted features without a significant reduction in the number of identified peptides. Figure 4-3E and Figure 4-3F show the receiver operating characteristic (ROC) curves of the detected features for the two DIA runs. Based on this analysis, and for the remainder of this study, we applied the fractional mass filter with d=0.1 and the isotope pattern probability threshold of 0.3. These parameters were also selected as the default settings in DIA-Umpire v2.

4.3.2 *Q Exactive DIA datasets*

We first evaluated the performance of DIA-Umpire untargeted identification analysis workflow using the complete Q Exactive DIA datasets published by Bruderer *et al* [32], including the HEK-293 cells and the human microtissue datasets (see 4.2). In the original publication, the authors used a conventional spectral library-based targeted extraction workflow using Spectronaut software. To build the spectral library, parallel DDA experiments were conducted using the same samples. Because DIA-Umpire allows library-free analysis, in this study we did not use the DDA-derived spectral library. Instead, the DDA data were used for comparing the number of identifications obtained using DIA and DDA strategies.

The DIA data were first processed using DIA-Umpire v2 signal extraction module (DIA_Umpire_SE.jar) to generate pseudo MS/MS spectra (see 4.2 for details). The spectra were searched using X! Tandem, Comet, and MSGF+ search engines. The results from the individual search engines were combined using iProphet, and protein lists were assembled using ProteinProphet. The corresponding DDA data were processed in a similar way. The results (peptide ion and protein identifications) were filtered at 1% FDR independently for each run (see 4.2, Table C-1 for HEK-293 cells, and Table C-2 for liver microtissue data). One average, the numbers of peptide ions identified per run at 1% FDR was slightly higher in DIA compared to DDA data (Table C-1 and Table C-2, columns "Peptide ion IDs (1% FDR run level)"). The number of proteins identified per run was comparable between DIA and DDA in HEK-293 data, and slightly less in DIA data than DDA data in the liver microtissue dataset (Table C-1 and Table C-2, "Protein IDs (1% FDR run level)" column).

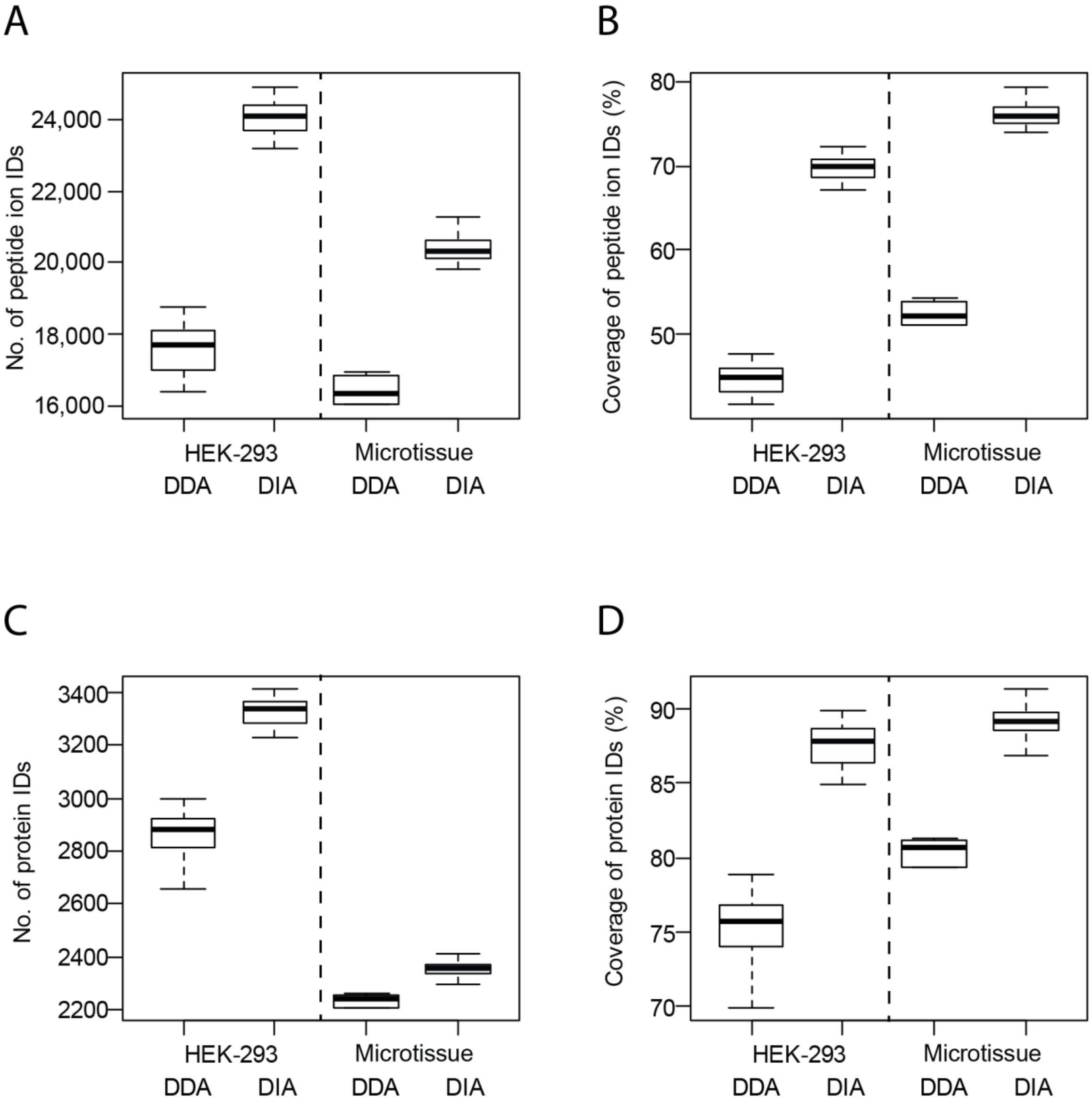


Figure 4-4 Identification numbers and reproducibility in the Q Exactive DIA and DDA datasets.

(A) The number of peptide ion identifications at individual run level in different datasets; **(B)** The coverage of peptide ion identifications (identification reproducibility across the dataset); **(C)** Same as (A), protein level; **(D)** Same as (B), protein level.

After the untargeted identification step, the DIA-Umpire's targeted re-extraction module was used to generate internal spectral libraries (from the spectra identified at 1% FDR at the dataset level), followed by targeted re-extraction with internal library to reduce the number of missing identifications across the replicates of the same dataset. Figure 3 shows that, after targeted re-extraction and with the data filtered at 1% FDR at the dataset level (see 4.2), DIA clearly outperformed DDA with respect to the number of peptide ions (Figure 4-4A) and proteins (Figure 4-4C) identified on average per run in both HEK-293 and liver microtissue datasets (individual run numbers are shown in Table C-1 and Table C-2, columns "Peptide ion IDs (1% FDR dataset level)" and "Protein IDs (1% FDR dataset level)"). Note that, for fair comparison, the number of identifications per run in DDA was counted after dataset level filtering as well.

Importantly, DIA data resulted in better identification coverage across different runs within the same dataset. Identification coverage for an individual run is defined here as the fraction of the total number of identifications in the dataset identified at 1% FDR (dataset level) that were detected in that run. The identification coverage was in the range of 63-79% at the peptide ion level and 82-91% at the protein level in DIA data, compared to 38-54% at the peptide ion level and 69-81% at the protein level in DDA data (Figure 4-4B and Figure 4-4D). These results were consistent with the original findings by Bruderer *et al* [27] for these data that demonstrated a very high completeness (i.e. low number of missing quantification values across different runs) that could be achieved using DIA.

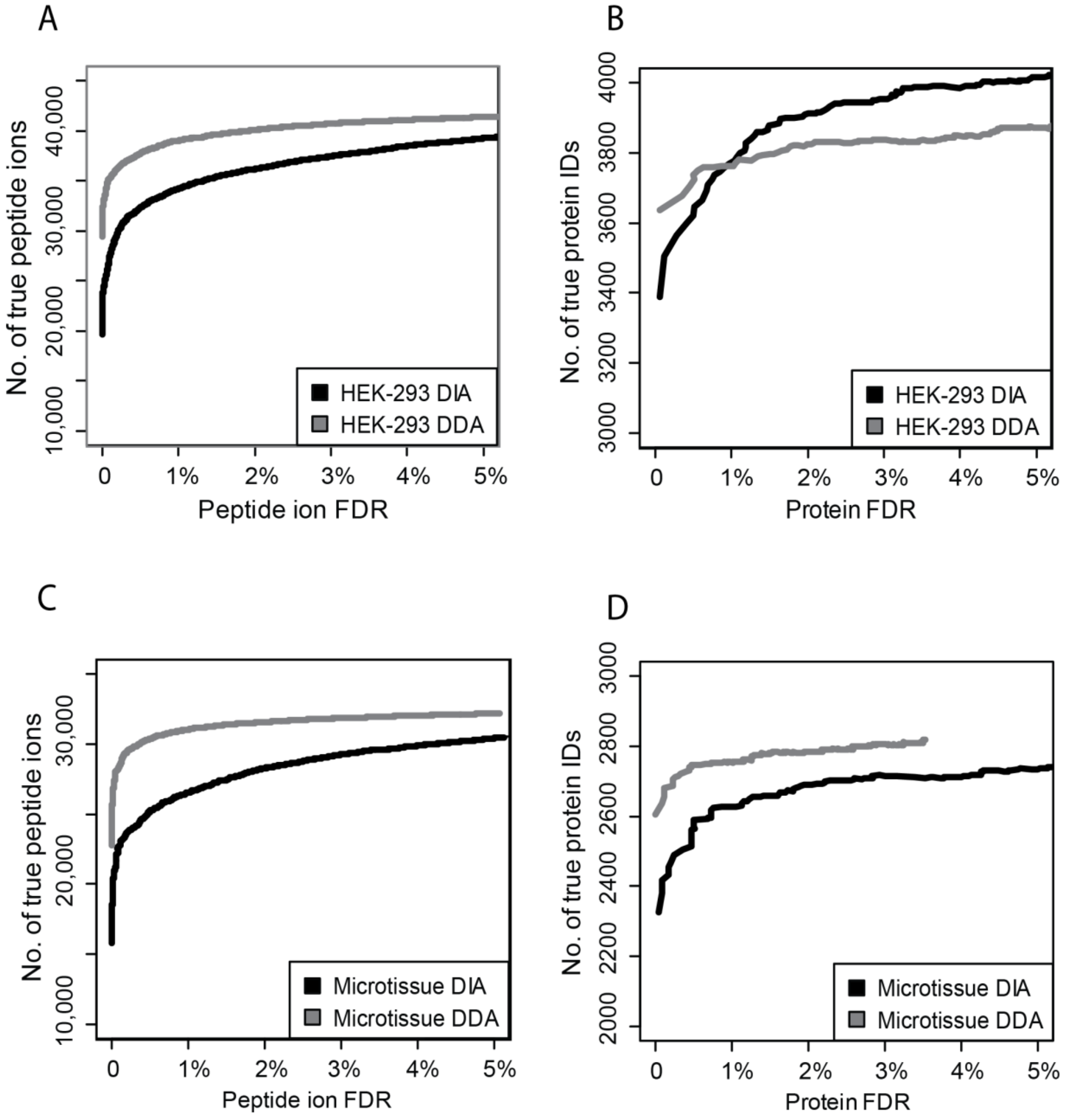


Figure 4-5 Number of identifications as function of FDR in the Q Exactive datasets. (A) Peptide ion identifications, HEK-293 Q Exactive DIA and DDA data. (B) Protein identifications, HEK-293 Q Exactive DIA and DDA data. (C) Same as (A), liver microtissue Q Exactive DIA and DDA data. (D) Same as (B), liver microtissue Q Exactive DIA and DDA data.

However, we also observed that the total number of peptide ion identifications per dataset (vs. individual run numbers discussed above) was higher in DDA than in DIA data, especially in the very low FDR range (below 1%). This is evident from Figure 4-5, which plots the ROC curves for the total number of peptide ion and protein identifications for each dataset. DIA data identified approximately 15% less peptide ions at 1% FDR in both datasets. At the protein level, the numbers were similar in the HEK-293 data, and DIA identified approximately 5% less proteins than DDA in the liver microtissue dataset. This shows that, using the spectral library-free workflow of DIA-Umpire, the main advantage of DIA versus DDA data remains to be better identification coverage (and thus quantification completeness) across the dataset, whereas DDA still provides a slight advantage in the total depth of the analysis. Note that the total number of peptide ions (and proteins) that can be quantified in DIA data using conventional targeted extraction methods is always less than that in DDA data since the extraction is limited to peptides present in the target spectral library built from the DDA data.

The original analysis of these data using targeted, spectral library-based software Spectronaut reported in [27] claimed almost no missing values (i.e. close to 100% identification coverage), compared to ~90% identification coverage (at the protein level) obtained here using DIA-Umpire. The original manuscript lacks sufficient details regarding FDR estimation in Spectronaut, and thus it is possible that such a high level of quantification completeness is achieved in part due to quantification of background (noise) signals instead of reporting them as missing values. Nevertheless, DIA-Umpire does have a limitation in that it relies on the detection of precursor ion signals. Peptides with MS1 precursor ion signals that not good enough to be detected using untargeted feature detection, but whose fragments have sufficiently strong signals in DIA MS2 spectra, are more likely to be identified using targeted extraction approaches based on fragment ion profiles alone. Although DIA-Umpire attempts to reduce the number of missing identification using targeted re-extraction as a second step, DIA-Umpire queries internal library spectra directly against the processed precursor-fragment groups, and not against the full raw data. Thus, the targeted re-extraction step of DIA-Umpire is still limited by the completeness of precursor-fragment signals assembled from detected MS1 and MS2 features. However, the untargeted identification workflow of DIA-Umpire is compatible with other targeted re-extraction and quantification analysis tools. When it is desirable to achieve as few missing identifications as possible, a spectral library can be built from DIA-Umpire-derived identifications to be used with targeted extraction tools like for more sensitive identification and quantification analysis. Although not discussed further in this manuscript, the use of DIA-Umpire results as input for targeted extraction analysis is already supported in the widely used tool Skyline.

4.3.3 *Orbitrap Fusion DIA datasets*

We next investigated the performance of DIA-Umpire on data from another advanced mass spectrometer from the Orbitrap family of instruments, Thermo Orbitrap Fusion, which brings high resolution, high mass accuracy, and high scan speed capabilities all together in a single instrument. It is capable of acquiring MS/MS spectra in either ion trap or in the Orbitrap, allowing implementation of conventional DDA, SWATH-like DIA, and several DDA/DIA workflows such as pSMART [81]. Here, we conducted five SWATH-like DIA experiments with different isolation windows of fixed widths (5, 10, 15, 20, and 25 Da), and two DDA experiments for comparison. Three replicate runs were performed for each experiment (see 4.2 for experimental details).

We processed the DIA and DDA data using same search parameters and FDR estimation as described above for Q Exactive data. Figure 4-6A and Figure 4-6C show the summary of peptide ion and protein identification numbers, respectively, for the DIA and DDA datasets (detailed numbers are shown in Table C-3). There were 30,000 - 32,000 peptide ions corresponding to 4,300 - 4,400 proteins identified by DDA per run (at 1% dataset level FDR). The best of DIA datasets identified similar or slightly higher number of peptide ions (33,000 - 34,000), corresponding to 4,000 - 4,200 proteins (slightly lower than DDA). Note that the experiments were conducted with only 135 minutes liquid chromatography (LC) gradient time and without any fractionation step – impressive numbers for both DDA and DIA. Similar to what was observed for Q Exactive dataset above, DIA allowed better identification coverage across the runs from the same dataset (Figure 4-6B and Figure 4-6D).

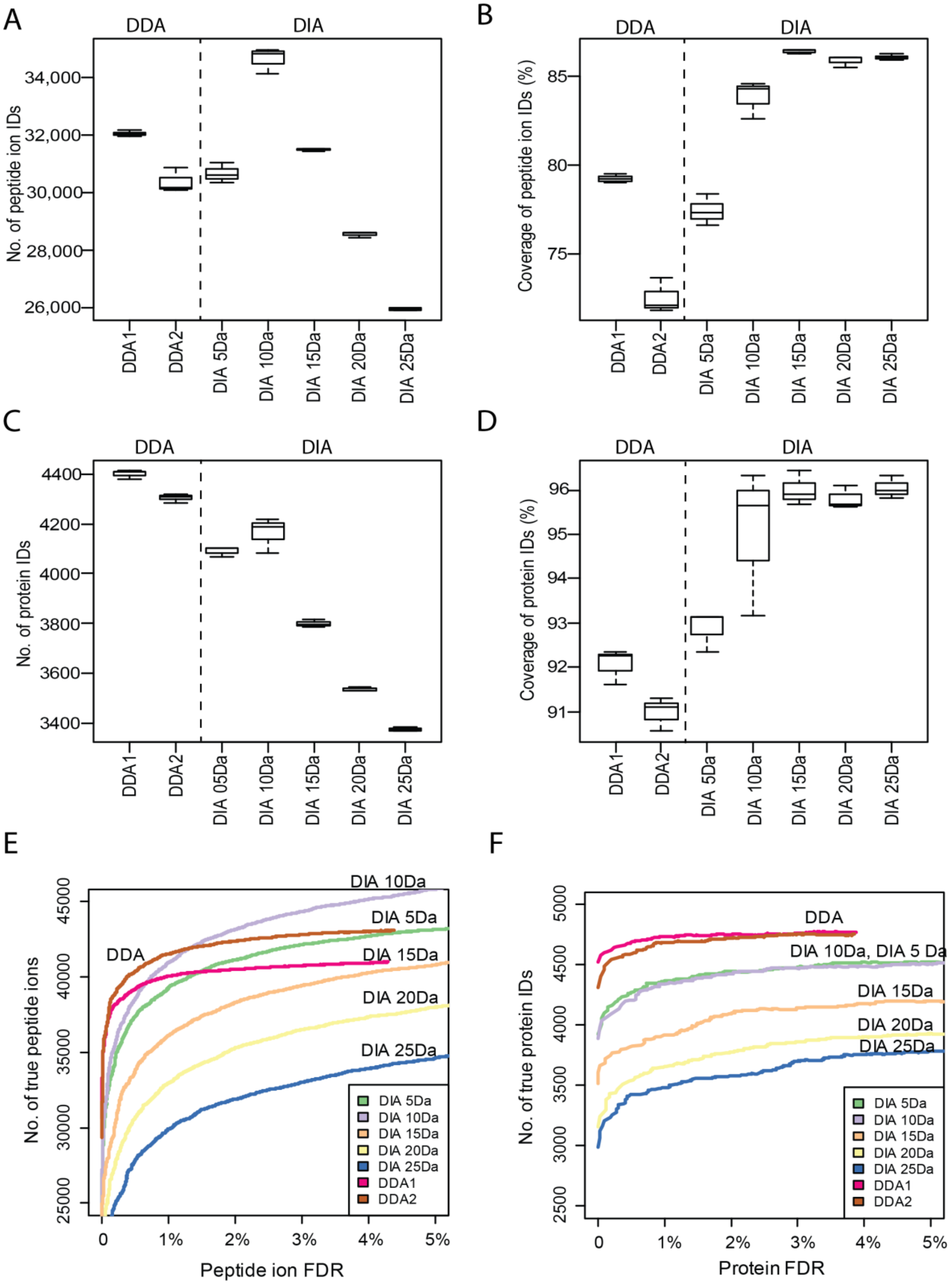


Figure 4-6 Identification numbers and reproducibility in the Orbitrap Fusion DIA and DDA datasets.

(A) The number of peptide ion identifications at individual run level in different datasets. **(B)** The coverage of peptide ion identifications (identification reproducibility across the dataset). **(C)** Same as (A), protein level. **(D)** Same as (B), protein level. **(E)** The number of peptide ion identifications as a function of FDR (dataset level, three replicates combined). **(F)** Same as (E), at the protein level.

In DIA data using Orbitrap Fusion, decreasing the isolation window widths from 25 Da (the window size used frequently to acquire SWATH-MS data on AB Sciex 5600 instruments) resulted in higher numbers of identifications per run. The best performance was observed at 10 Da isolation width, and the number of identification dropped slightly (more at the peptide ion than protein level) with 5 Da setting. At the same time, the identification reproducibility (identification coverage) was generally better for larger window sizes. Using smaller isolation windows reduces the number of co-fragmented peptides and therefore alleviates the difficulties of de-convoluting DIA MS/MS spectra using the approach of DIA-Umpire. However, using smaller isolation widths increases the number of required MS/MS scans to cover the same precursor *m/z* range, and therefore increases the cycle time. For example, narrowing the isolation window size from 10 Da to 5 Da, under the instrument settings used in this work, increased the cycle time from 6.2 to 13 seconds. Longer cycle times result in fewer measurement points acquired per peptide elution peak, making the measurement of peak shape correlation between precursor and fragment signals less reliable. This, in turn, makes it more difficult to detect low abundance and short eluting peptide ions (see Figure 4-7), thus lowering the identifications reproducibility (Figure 4-6B and Figure 4-6D).

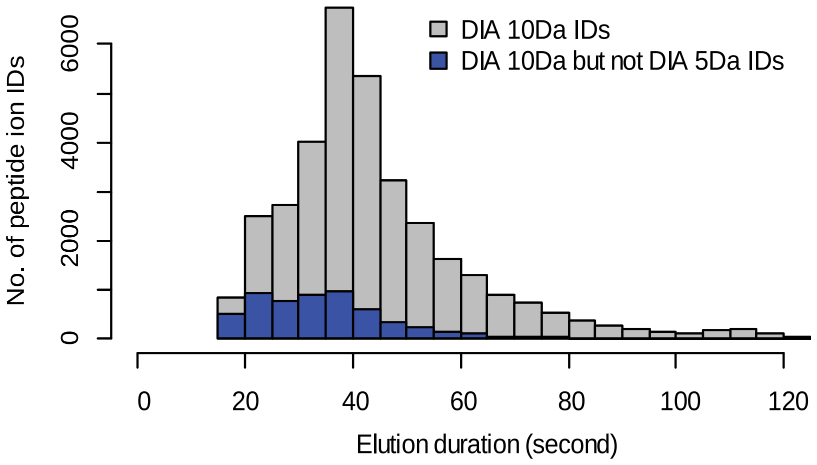


Figure 4-7 Elution time duration of peptide ions in the first replicate of DIA 10 Da Orbitrap Fusion dataset.

Grey: Histogram of identified peptide ion elution durations in the DIA run; **Dark Blue**: Histogram of the peptide ion elution durations which were identified in the replicate of DIA 10 Da dataset but not identified in any of DIA 5 Da replicates.

Investigating the total number of identifications per dataset (i.e. combining triplicate runs for each dataset) between DIA and DDA at various FDR levels data in more details, DDA had more peptide ions identified in the very low FDR range (below 0.5% FDR) than DIA with any window size (Figure 4-6E), even though the DIA numbers (5 and 10 Da windows) exceeded those of DDA in the FDR range of approximately 1% or higher. It is well known that, due to error rate inflation when going from peptide to protein level [1], achieving a certain low protein level FDR (e.g. 1%) requires peptide identifications passing lower than that FDR value at the peptide ion level. This explains why the number of protein identifications at 1% protein FDR was higher in DDA data (Figure 4-6F), even though the opposite was observed at 1% FDR at the peptide ion level. The reason why in DDA data there were more peptide ion identifications with very high confidence (FDR below 1%) is that MS/MS spectra acquired using DDA with a tighter isolation width of 1.4 Da were on average less noisy and contained more peptide-specific fragment ions than pseudo MS/MS spectra de-convoluted using DIA-Umpire.

4.3.4 *Performance of semi-parametric mixture modeling*

DIA-Umpire v2 implements an improved scoring function and a more robust strategy based on semi-parametric mixture modeling with kernel density estimation (replacing a parametric Gaussian mixture model) for computing posterior probabilities of true identifications (see 4.2 for details). We illustrate these improvements here by performing a comparison with the results obtained using DIA-Umpire v1.25 [41] on the Orbitrap Fusion and Q Exactive DIA datasets. Figure 4-8 shows an example of U-score histograms and mixture modeling results obtained using the two versions for a single DIA run from the Q Exactive microtissue dataset. The results from all the other DIA runs used in this work, including Orbitrap Fusion data, as the result from an example DIA run shown in Figure 4-8 shows a wider distribution of high scoring (i.e. potentially correct) identifications, while keeping the width of the decoy distribution unaffected. This results in a better discrimination between the correct and incorrect (decoy) identifications in these data. Combining the new scoring and the semi-parametric mixture modeling, DIA-Umpire v2 can extract more identification at different FDR ranges, especially in Q Exactive data.

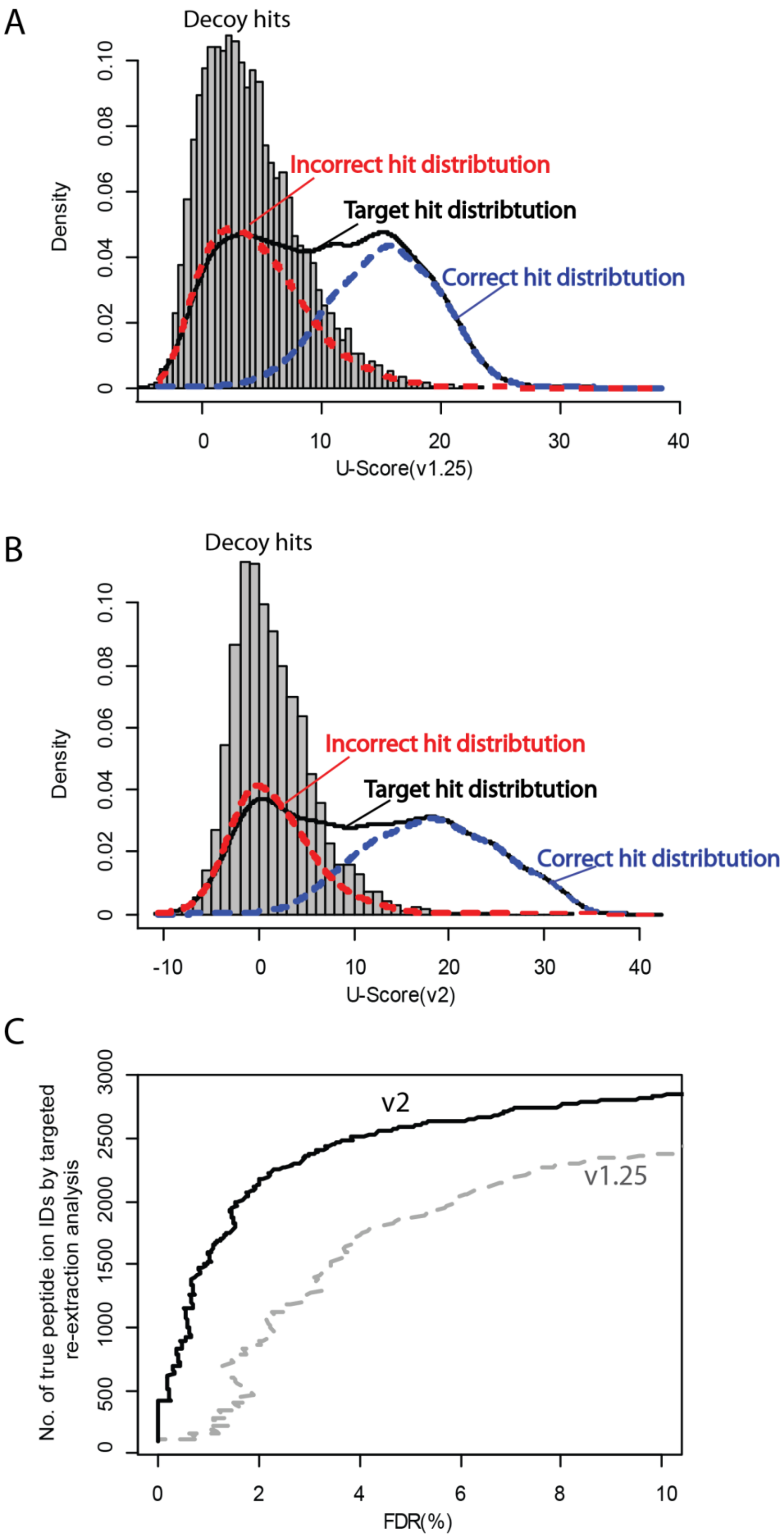


Figure 4-8 Score histograms and mixture modeling

(A) Score histograms and parametric Gaussian mixture modeling result obtained using DIA-Umpire v1.25. **(B)** Score histograms and semi-parametric mixture modeling result obtained using DIA-Umpire v2. **(C)** The number of targeted re-extraction identifications as a function of FDR obtained using DIA-Umpire v1.25 and v2. Data for one representative run from the Orbitrap HEK-293 Q Exactive dataset.

In addition, the flexible mixture modeling by the semi-parametric kernel density estimation provides a better fit for the correct distribution than that achievable under parametric (e.g. Gaussian shapes) assumptions. This ensures that the computed probabilities of correct identifications are more accurate [88]. This is a particularly important feature for the external library searches shown next, e.g. for combining the results of targeted extraction using the internal library built by DIA-Umpire with that using external DDA libraries. We also believe it is a preferred strategy of modeling the distributions of scores compared to that implemented in mProphet and currently used in targeted extraction tools such as OpenSWATH and Skyline.

4.3.5 *Performance of external library searches*

We performed DIA-Umpire external library searches on the Q Exactive HEK-293 DIA dataset. There were 34,475 peptide ion identifications and 3,827 proteins from DIA-Umpire untargeted identification analysis (See 4.2.6 for details, the numbers were determined by 1% FDR for both peptide ion and protein numbers at dataset level). The two external libraries, the sample-specific DDA spectral library [27] has 29,243 peptide ions and the Human-CAL library contains 205,320 peptide ions. The overlaps between the three sets of peptide ions are shown in Figure 4-9A. DIA-

Umpire identified majority of the peptide ions which were identified by DDA experiments. Among the 4,994 peptide ions which were identified only from DIA-Umpire's untargeted identification analysis, 2,242 (corresponding to 2,091 peptide sequences) of them have unique sequences which were not present in the two external libraries, and 1,471 of the remaining peptide ions have unique modified peptide species (peptide sequence plus modification type and site), and 1,281 peptide ions were due to the different charge states detected by DIA-Umpire's untargeted analysis. Note that among these 1,281 peptide ions, 67% of them were singly charged ions which are often ignored by DDA. This analysis again shows that the DIA-Umpire untargeted analysis is able to identify new peptides even compared with the global spectral library like Human-CAL.

After searching DIA data against external libraries, DIA-Umpire additionally identified 2,911 (2,222 + 689, see Figure 4-9B) peptide ions from the DDA sample-specific library, and 10,879 (8,126 + 2,753, see Figure 4-9C) peptide ions from the Human-CAL external library. Combining these two external library searches, we were able to identify in total 46,495 peptide ions (Figure 4-9D) which correspond to 4,540 proteins identified (determined by 1% protein FDR).

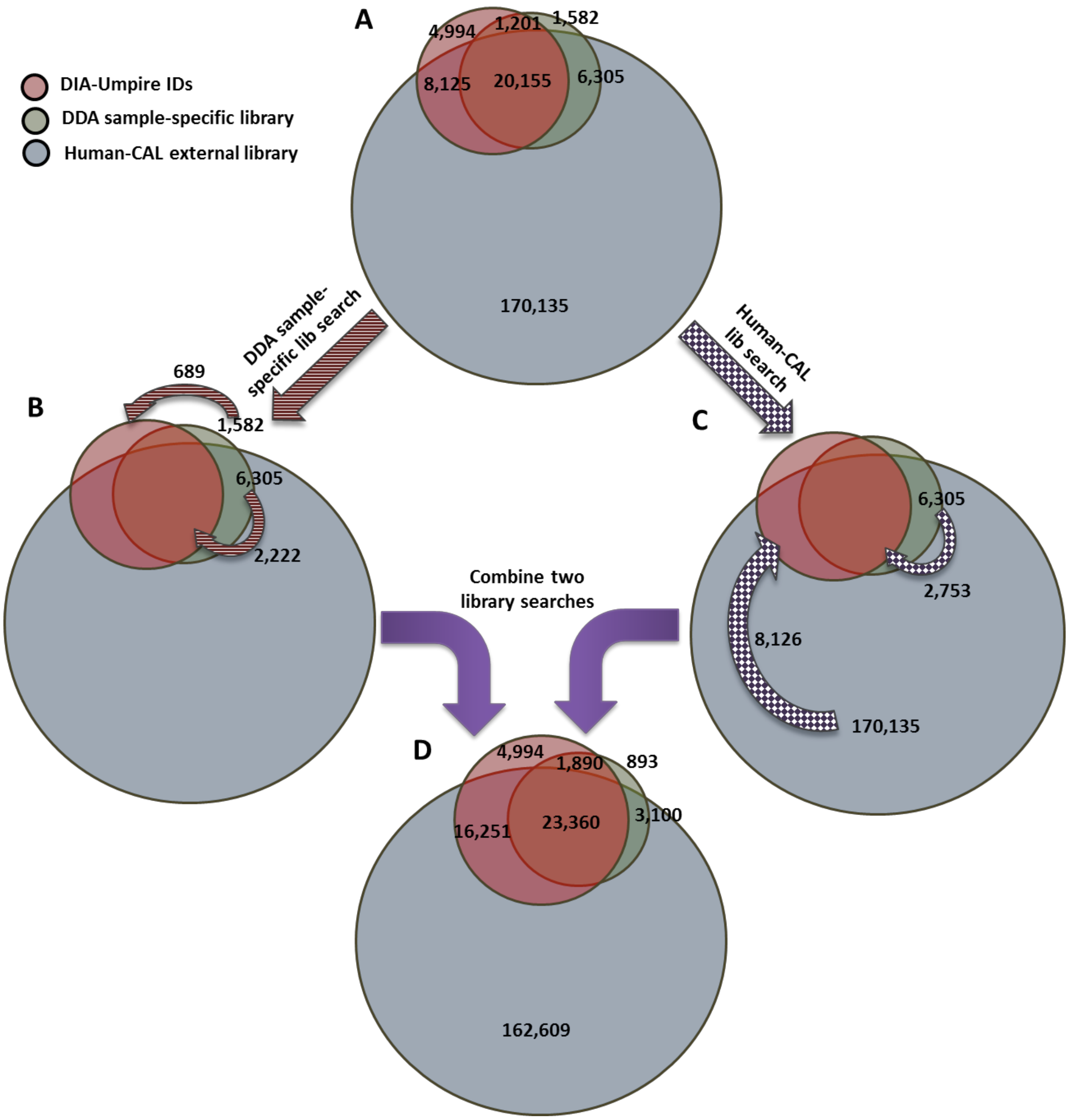


Figure 4-9 Venn diagrams of peptide ion identifications from DIA-Umpire analysis

(A) Overlaps between peptide ion identifications from DIA-Umpire untargeted analysis, DDA sample-specific library, and Human-CAL library; **(B)** After the DIA-Umpire external library search against DDA sample-specific library, additional 2,911 (689 + 2,222) peptide ions were considered confident identifications (equal or above 0.99 U-score probability); **(C)** After the DIA-Umpire external library search against Human-CAL library, additional 10,879 (2,753 + 8,126) peptide ions were considered confident identifications (equal or above 0.99 U-score probability); **(D)** Overlaps of peptide ion identifications after combining the two external library searches.

It is clear that the two external libraries are very different in terms of the number of peptide ions (29,243 vs. 205,320 peptide ions). In addition, the DDA experiments for building the two libraries were generated from different mass spectrometers and LC systems. The U-score histograms and the retention time alignment models of DIA-Umpire reflected these differences, as the examples shown in Figure 4-10. Figure 4-10A and Figure 4-10B are the U-score histogram and the retention time alignment model, respectively, obtained from the DDA sample-specific library search for one representative DIA file. Figure 4-10C and Figure 4-10D are the result figures obtained from the Human-CAL library search for the same DIA file. The two different U-score histograms (Figure 4-10A and Figure 4-10C) show that a higher proportion of target peptide ions from DDA sample-specific library were considered present (higher U-scores) in the DIA run, but most target peptide ions from Human-CAL library were not present in the DIA run (U-scores are similar to that of decoy hits). It is clear that the flexible semi-parametric mixture modeling was able to effectively deconvolute the U-score histograms to accurately calculate posterior probabilities for both scenarios. In addition, because there were sufficient peptide ions identified from the DIA-Umpire untargeted analysis, therefore the robust retention time alignment models (Figure 4-10B and Figure 4-10D) can be built without using the iRT peptides.

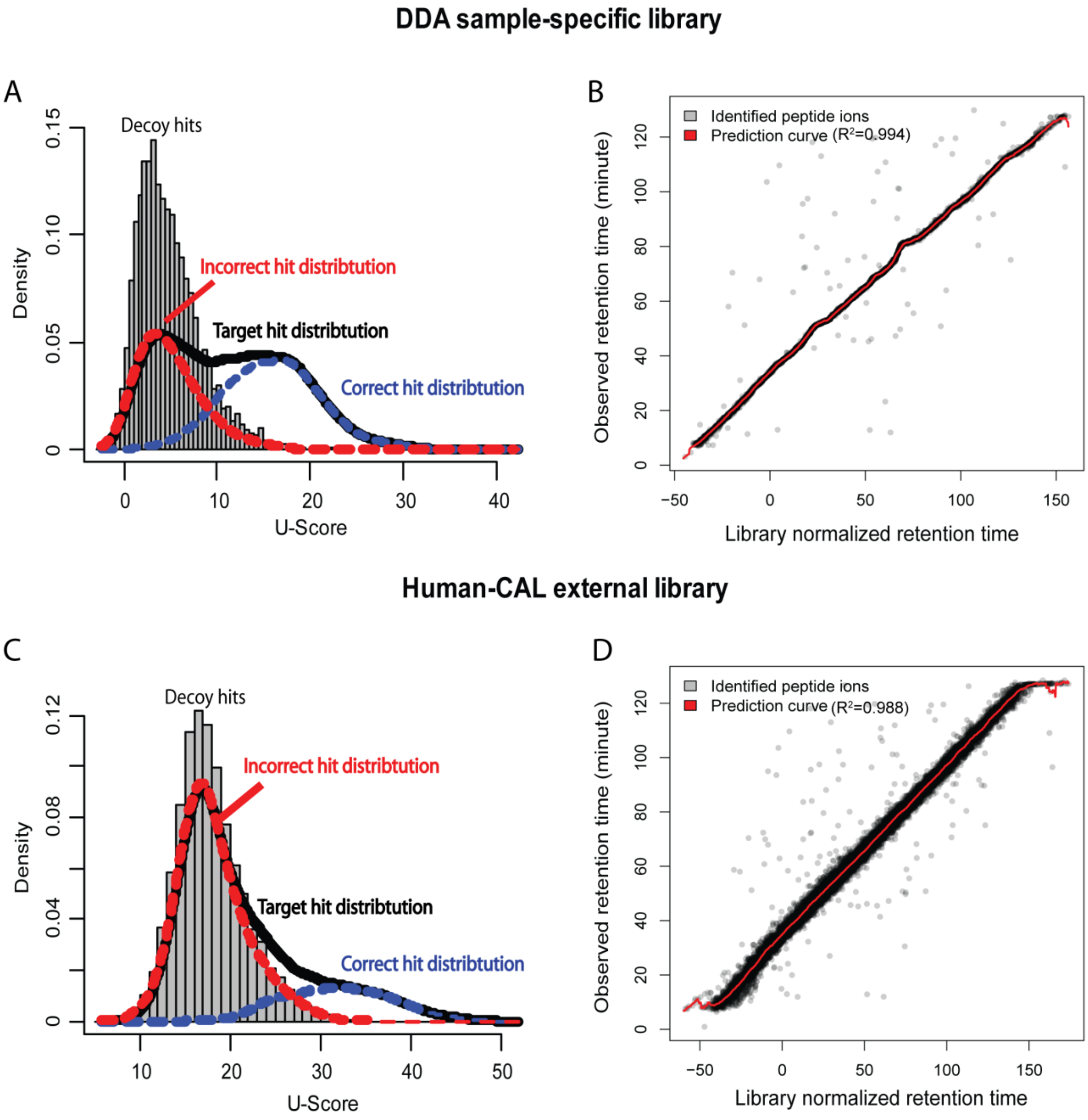


Figure 4-10 U-score histograms and retention time alignment for the two external library searches

The results shown in this figure were derived by a DIA run in HEK-293 Q Exactive dataset. **(A)** Score histograms and parametric Gaussian mixture modeling result obtained from DDA sample-specific library search; **(B)** Retention time alignment model obtained from DDA sample-specific library search; **(C)** Score histograms and parametric Gaussian mixture modeling result obtained from Human-CAL library search; **(D)** Retention time alignment model obtained from Human-CAL library search.

4.4 *Discussion*

In this study, we presented DIA-Umpire v2 and demonstrated that it is capable of untargeted complex proteome analysis using DIA data generated on Thermo Orbitrap mass spectrometers. In addition, the targeted re-extraction module is able to perform external library search. Using publicly available Q Exactive DIA data, and Orbitrap Fusion data acquired as part of this work, we showed that the DIA datasets achieved similar identification numbers and better identification reproducibility across samples and replicates than DDA data. With the smaller number of missing quantification values, DIA data should provide improved statistical power for the post-quantification analysis, e.g. using tools such as mapDIA [90]. Importantly, the workflow of DIA-Umpire does not require a spectral library, which should facilitate the adoption of DIA strategy for a broad range of discovery proteomics applications. DIA-Umpire is fully compatible to many existing DDA-type analysis pipelines, so the users can continue using the database search engines and post-processing tools they are familiar with to analyze the pseudo MS/MS spectra extracted using DIA-Umpire from DIA data.

DIA-Umpire's untargeted approach provides an alternative way to process DIA data. Unlike other targeted extraction software tools, DIA-Umpire extracts peptide precursor and fragment signals without any hypothesis or prior knowledge about the content of the samples. The untargeted detection has an advantage of finding new peptide ion signals in DIA data that may not be present even in a comprehensive spectral build from DDA data. One limitation of the untargeted DIA-

Umpire strategy is that it relies on good signal quality of precursor ions, most notably the MS1 signal quality. The challenge is more apparent for low abundance signals present at the noise level, where it becomes difficult to distinguish true signals from the noise. Although maximizing the sensitivity of feature detection (i.e. extracting as many features as possible) generally also maximizes the identification numbers, such a strategy is not always practical due to increase in the analysis time. To address this, we introduced in DIA-Umpire v2 two feature quality filters, the fractional mass and peak isotope pattern filters, which aim to increase the specificity of the feature detection process. Furthermore, because DIA-Umpire-derived identifications are compatible with other existing DIA targeted extraction tools such as Skyline, one can generate a DIA-derived spectral library to perform targeted extraction and quantification using those tools, potentially maximizing the amount of quantitative information that can be extracted from the data.

Experiments conducted using an Orbitrap Fusion instrument as part of this work demonstrated the high quality of DIA data with respect to the number of identifications and the identification reproducibility. Future work should also explore the accuracy of peptide and protein quantification that can be extracted from these data, either using the fragment ion intensities from MS2 data or MS1 precursor ion intensities (as both quantification options are supported in DIA-Umpire). It should also be noted that the quality of MS1 signal and good chromatography are very important for DIA-Umpire analysis, as these factors ensure accurate detection of precursor features and assembly of precursor-fragments groups. Evaluation of the Orbitrap Fusion data acquired using different window sizes showed noticeable differences in the numbers of identified peptides and proteins, with an overall preference for a 10 Da window size. However more comprehensive and consistent evaluation of different instrument settings should be performed in the future work. Finally, the analysis presented here was primarily concerned with the untargeted, spectral-library free workflow of DIA-Umpire. Thus, evaluation of the performance of targeted extraction tools on DIA data generated on an Orbitrap Fusion instrument, or comparison between different computational strategies for the analysis DIA data goes beyond the scope of this work. Nevertheless, we hope that the data generated as part of this work, which we make available via the ProteomeXchange consortium database (PXD003179), can be used for that purpose in the future.

4.5 *Contributions*

The results presented in this chapter could not have been done without the efforts of the great collaboration. I am grateful and fortunate to have them for the project.

Chih-Chiang Tsou: developed the algorithms, implemented the software, designed experiments, analyzed the data, and wrote the manuscript draft.

Chia-Feng Tsai: acquired mass spectrometry data, designed experiments, and provided inputs for manuscript.

Alexey I Nesvizhskii: provided inputs for algorithm development, supervised the project, and wrote the manuscript.

Chapter 5 Conclusions and future directions

The dissertation work presented a comprehensive computational platform called DIA-Umpire for proteomics data acquired by data-independent acquisition (DIA). Because of the challenges of DIA data analysis (See Chapter 1), spectral library-dependent approaches were most commonly used for DIA proteomics data analysis. As the drawbacks were discussed in 1.5, building a spectral library requires large amount of samples and huge MS analysis time on parallel DDA experiments. Therefore, there is a need for new developments of computational approaches which allow untargeted analysis for proteomics data acquired by DIA without a spectral library.

In Chapter 2, a signal feature detection algorithm was developed to extract meaningful peptide precursor and fragment signal features from DIA data to enable sensitive peptide and protein identifications. This workflow is fully compatible with most currently existing DDA-based peptide and protein identification tools, including database and spectral library search engines as well as FDR estimation methods. It allows identification of peptides directly from DIA data in an untargeted manner, i.e. without the need for building sample-specific spectral libraries using parallel DDA runs or relying on pre-existing libraries that may not accurately represent the peptides in the samples under investigation. We showed that this direct, untargeted workflow was able to identify similar number of identifications compared to the conventional DDA approach.

In Chapter 3, a targeted re-extraction approach for quantification analysis was developed and integrated into DIA-Umpire pipeline. It searches the DIA-Umpire preprocessed DIA data against an internal spectral library built from the initial identifications from database search for a dataset. Using this hybrid approach, DIA-Umpire is able to perform more consistent identification and quantification across multiple DIA runs. In addition, we showed that such reproducible and reliable quantification is possible using both DIA MS1 and MS2 data with accurate protein quantification. The whole DIA-Umpire pipeline was further demonstrated to be able to capture sensitive protein-protein interaction profile without a prior spectral library for complete DIA analysis.

In Chapter 4, several algorithms were further improved, including the feature detection algorithm, the targeted re-extraction scoring, and the mixture modeling for posterior probability calculation. In addition, we showed that the pipeline is capable of untargeted complex proteome analysis using DIA data generated from Thermo Orbitrap mass spectrometers and is compatible with external library searches. Using publicly available Q Exactive DIA data, and Orbitrap Fusion data acquired as part of this work, we showed that the DIA datasets achieved similar identification numbers and better identification reproducibilities across samples and replicates than DDA data. With the smaller number of missing quantification values, DIA data should provide improved statistical power for the post-quantification analysis.

In summary, DIA-Umpire provides comprehensive analysis including identification and quantification for proteomics data acquired using DIA. It accepts standard spectral data formats and supports various mass spectrometers. Most importantly, the workflow of DIA-Umpire does not require a spectral library, which should facilitate the adoption of DIA strategy for a broad range of discovery proteomics applications. DIA-Umpire is fully compatible with many existing DDA-type analysis pipelines, so the users can continue using the database search engines and post-processing tools they are familiar with to analyze the pseudo MS/MS spectra extracted using DIA-Umpire from DIA data. All the work presented in this dissertation is integrated as an open-source software package and written in Java programming language (v1.7) which can be executed in almost all operating systems including Windows, Linux and Mac OS. Example data and tutorial documents are publically available at http://diaumpire.sourceforge.net, and the source code is licensed under Apache 2.0 and available at https://github.com/cctsou/DIA-Umpire.

After the introduction of DIA-Umpire, the untargeted, spectral library-free approaches have been an emerging trend for DIA analysis. In contrast to DIA-Umpire which mainly focuses on single DIA run processing, Group-DIA [91] combines elution signal profiles of the same peptide ion across multiple DIA runs by retention time alignment to enhance the quality of precursor-fragment pairs. The authors showed that using this strategy Group-DIA was able to achieve better performance of untargeted identification and more reliable quantification compared to DIA-Umpire and OpenSWATH. Another newly published DIA analysis tool, MSPLIT-DIA [92], assumes peptide multiplicity in a single DIA MS2 spectrum, therefore allowing multiple peptide ions from spectral library to be identified by a single multiplexed DIA MS2 spectra. Using a DIA-specific scoring based on score profile which is similar to the concept proposed by Weisbrod *et al* [93], MSPLIT-DIA was shown to be a more sensitive DIA identification analysis tool compared to other competitors. MSPLIT-DIA is a spectral library-dependent method, although it was described as an untargeted identification approach as long as a global spectral library is provided, e.g. the comprehensive human SWATH combined assay library (Human-CAL) [35] mentioned in Chapter 4. Note that all spectral library-dependent tools described in 1.5 can be considered as untargeted analysis tool by such definition.

DIA-Umpire requires precursor ion signals to be detected to generate pseudo MS/MS spectra. The accurate mass of detected precursor signal is critical to restrict search space of peptide precursor masses. Most of the spectral library-dependent tools, such as OpenSWATH and MSPLIT-DIA, do not rely on precursor signals to restrict the precursor mass search space. Instead, they use isolation window *m/z* range to roughly restrict the search space (e.g. 25 Da in the standard SWATH). The advantage of not relying on precursor signals is that it allows identifications of low abundance peptide ions whose precursor signals may be too weak to be reliably detected by feature detection algorithm. On the other hand, however, as the examples shown in Figure 2-12, false identifications of modified peptide species can easily reach statistically significant scores because they have similar spectral (fragmentation) pattern with unmodified peptide species (or other different modification forms). We believe that the presence of precursor signal is a key to unambiguously distinguish peptide ion from its different modified or unmodified peptide species. Nevertheless, it is still important to further investigate how to unambiguously identify those low abundance peptide ions which do not have reliable precursor signals but do have fairly good fragment signals in DIA MS2 spectra. The future directions of this dissertation work could include modifying the feature detection algorithm described in 2.2.5 to detect and group fragment signals without requirements of parent precursor signals. The grouped fragment signals can be further classified into different signal quality tiers determined by the presence of possible precursor signals. For the quality tier of fragment groups where no reliable precursor signal is detected, a new scoring or a new target-decoy strategy needs to be specifically developed to make sure the identifications are reliable and not misidentified because of modified versions of other peptide species. The future directions shall also include more rigorous assessments on quantification performance of DIA-Umpire. OpenSWATH [28] was published with a Gold Standard (SGS) peptide dataset which includes ~300 peptides with various concentrations. Using the SGS dataset, we would be able to test quantification performance and also further understand the identification limit of DIA-Umpire.

The key feature of DIA data is the full MS2 fragment map for all peptide ions across entire LC range. This feature creates challenges for analysis but more importantly it opens up a lot of opportunities and allows one to analyze DIA data in different ways, as evidenced by the two newly published DIA tools [91, 92] which take advantage the additional information to improve DIA identification analysis. In the near future, as the new mass spectrometers continue to improve sensitivity and scan rate, more technological developments such as new DIA variants are expected to be on the horizon. The continuous development of novel algorithms and computational tools for addressing the new challenges derived from the forthcoming DIA methods are essential and will remain a major trend in computational mass spectrometry.

Appendix A Supplementary materials for Chapter 2

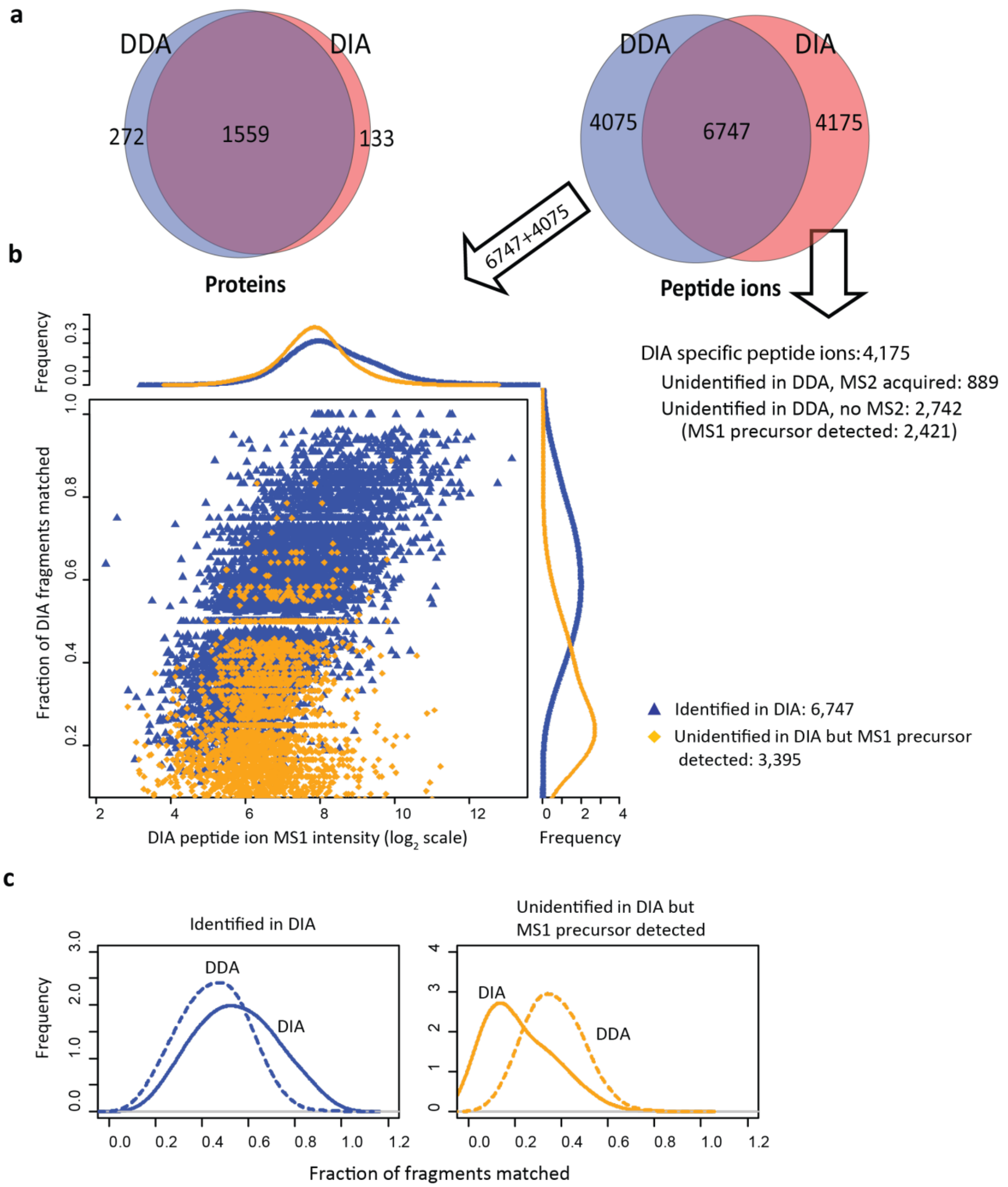


Figure A-1 Untargeted peptide identification using DDA and DIA data from human cell lysate samples using three search engines combined.

DIA pseudo MS/MS spectra were searched using X! Tandem, Comet, and MSGF+, and combined using iProphet. Protein and peptide ion identifications were then filtered at 1 % FDR using target-decoy approach. **(a)** The numbers of proteins and peptide ions identified at 1% FDR in DDA and in DIA data. Left: number of protein identifications in each experiment (1,831 proteins identified from DDA data, 1,692 from DIA, 1,964 in total). Right: Total number of peptide ion identifications from two replicates (10,822 peptide ions identified from DDA data, 10,922 from DIA, 14,997 in total). Compared to using X! Tandem only (main text, Figure 4) when the results from all three search engines were combined the number of identifications increased in both DDA (by 17% and 11% for peptide ions and proteins, respectively) and in DIA data (by 25% and 16% for peptide ions and proteins, respectively). However, the overlap between the DIA and DDA identified peptide ions and proteins increased only slightly, to 45% and 79%. Of the peptide ions identified by DIA and not DDA at 1% FDR (total 4,175 peptide ions), the majority of the remaining peptide ions were not identified by DDA because no MS/MS was acquired (2,742). **(b)** Percent of fragments ions matched in pseudo MS/MS spectra extracted from DIA data as a function of the MS1 peptide ion identity in DDA data. Data points (peptide ions) and the summary density plots are labeled according to the three categories of peptide ions: ions identified from DIA data at 1% FDR ("Identified in DIA"; blue), and unidentified in DIA (orange; these ions were located in DIA data as described in Online Methods). **(c)** Comparison between DDA and DIA in terms of numbers of fragments matched among two categories of peptide ions, showing that peptide ions identified with confidence from DDA but not DIA have fewer fragment ions that could be matched.

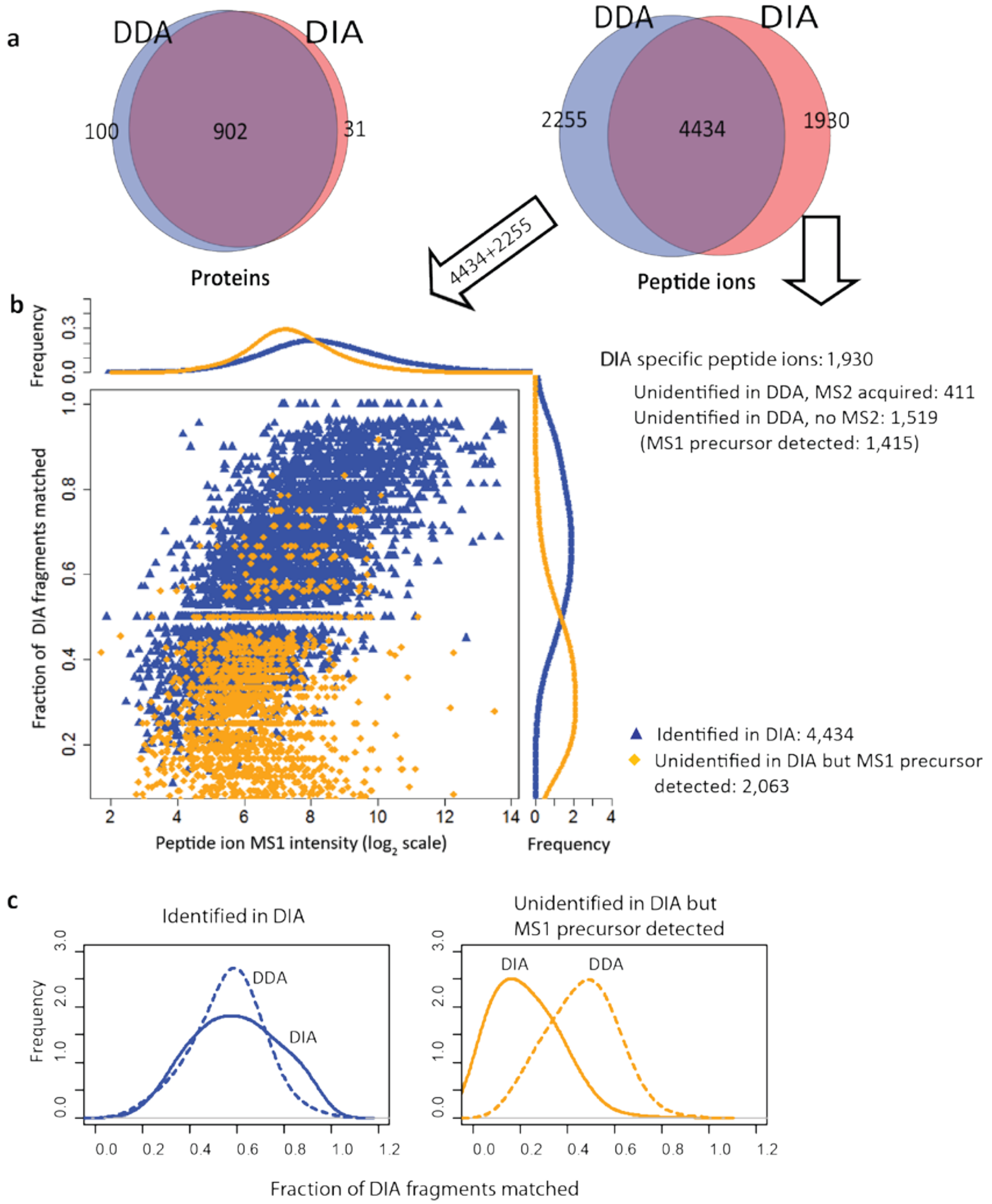


Figure A-2 Untargeted peptide identification using DDA and DIA data from E. coli cell lysate samples with X! Tandem search engine.

Results for *E. coli* data were similar to those obtained for human cell lysate data (see Figure 2-8).

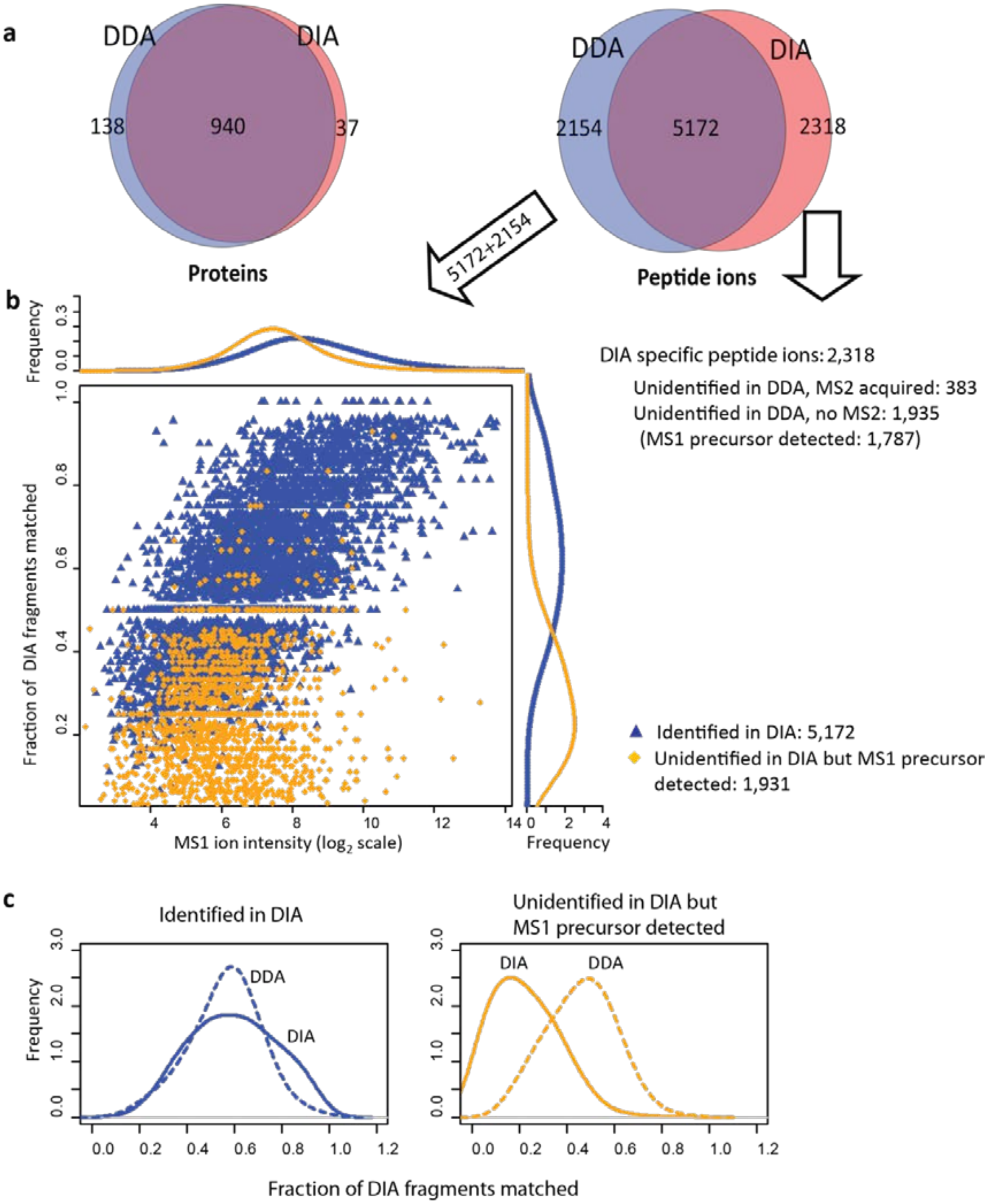
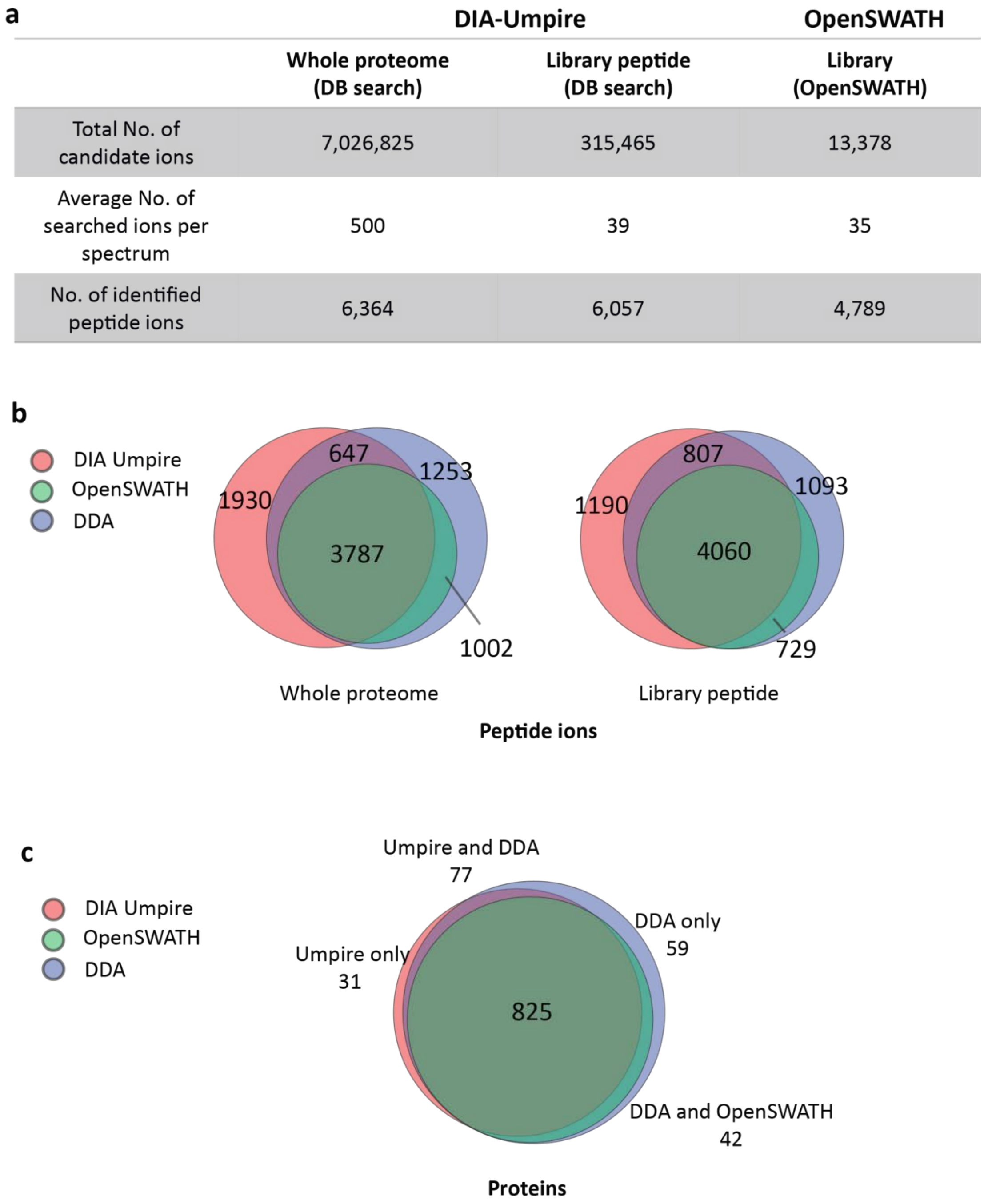


Figure A-3 Untargeted peptide identification using DDA and DIA data from *E. coli* cell lysate samples with three search engines combined.

Results for E. coli data were similar to those obtained for human cell lysate when using X! Tandem, Comet, and MSGF+ (combined using iProphet; see **Figure A-1**).



a

	DIA-Umpire		OpenSWATH
	Whole proteome (DB search)	Library peptide (DB search)	Library (OpenSWATH)
Total No. of candidate ions	7,026,825	315,465	13,378
Average No. of searched ions per spectrum	500	39	35
No. of identified peptide ions	6,364	6,057	4,789

Figure A-4 Comparison between untargeted DIA-Umpire analysis and OpenSWATH targeted extraction: effect of the search space. *E. coli* cell lysate data. Results similar to those presented from human cell lysate were obtained for E. coli data.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
NSPLDEENLTQENQDR	951.91	2	47.98	1.30E-06	N/A
NSPLDEENLTQENQDR	951.42	2	47.98	3.68E-08	48.0

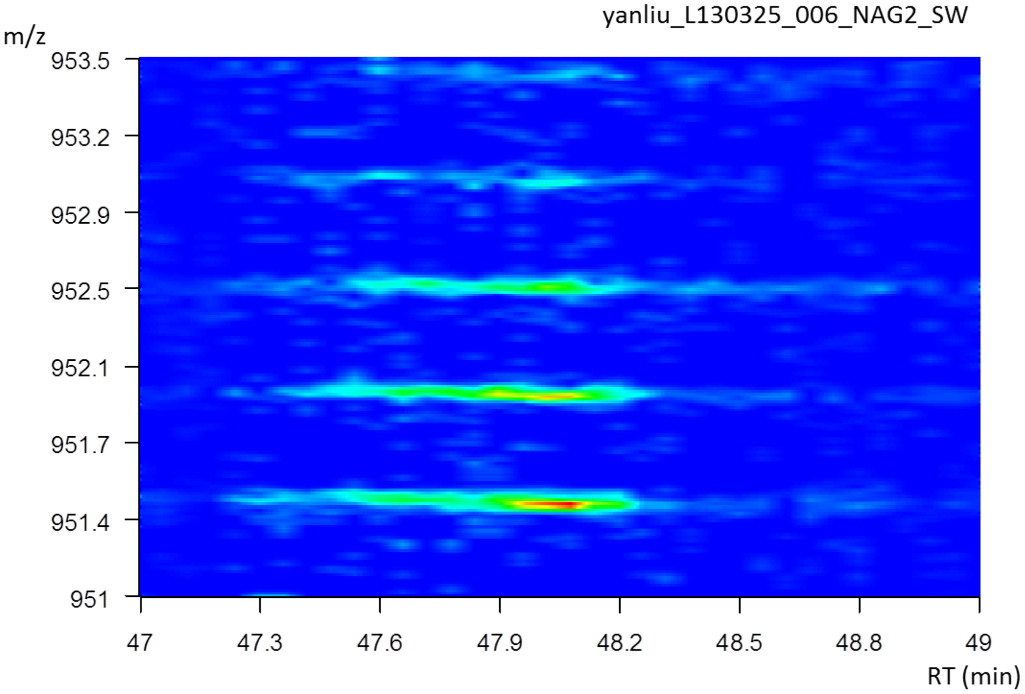


Figure A-5 Example of an ambiguous identification of the deamidated peptide NSPLDEENLTQENQDR by OpenSWATH targeted search.

Two separate identifications (in unmodified form and in a deamidated form; the site of the modification is shown in red) were reported by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 47.98 minutes (precursor *m/z* of 951.42 Da). DIA-Umpire reported only one (unmodified) form.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
DIENFNSTQK	599.26	2	37.41	0.000105616	N/A
DIENFNSTQK	598.77	2	37.39	1.67E-07	37.35

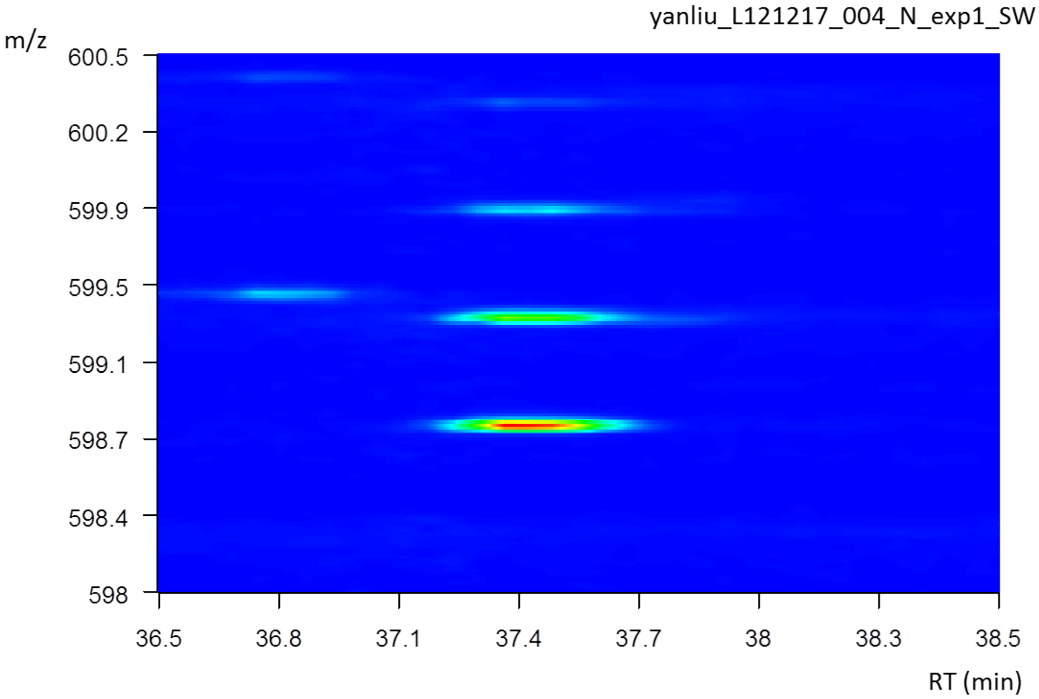


Figure A-6 Example of an ambiguous identification of the peptide DIENFNSTQK by OpenSWATH targeted search.

Two separate identifications with different modification site compositions (with one and two deamidations; modification site shown in red) were reported by OpenSWATH. The two identifications both had a small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had almost identical retention times (within 0.02 minute). The MS1 signal image shown above suggests there is only one peptide eluting at RT = 37.4 minutes (precursor *m/z* of 598.77 Da). DIA-Umpire reported only one (singly deamidated) form, further supported by the presence of NXS/T motif covering the reported site.

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
TGNGLFLSEGLK	618.82	2	67.58	3.04E-09	69.99
TGNGLFLSEGLK	618.33	2	67.58	3.42E-08	67.59

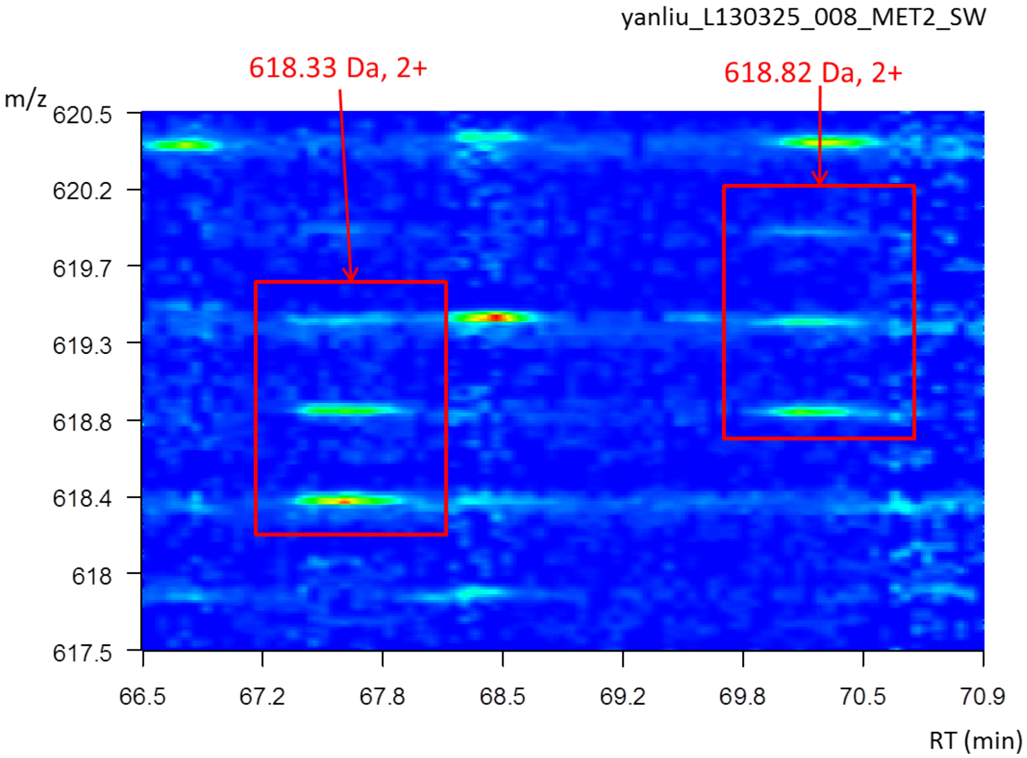


Figure A-7 Example of an ambiguous identification involving the deamidated peptide TGNGLFLSEGLK.

Two separate identifications were reported (in unmodified and in deamidated form) by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 67.6 minutes (precursor *m/z* of 618.33 Da), which was identified by DIA-Umpire as unmodified peptide. In addition, DIA-Umpire identified the deamidated form of the peptide at retention time of 69.99 minutes (also marked on the MS1 signal image).

Sequence	m/z	Charge	OpenSWATH RT	mProphet m_score	DIA-Umpire RT
VAPEEHPTLLTEAPLNPK	653.01	3	63.586	0.001410359	N/A
VAPEEHPTLLTEAPLNPK	652.686	3	63.591	3.67E-08	63.58

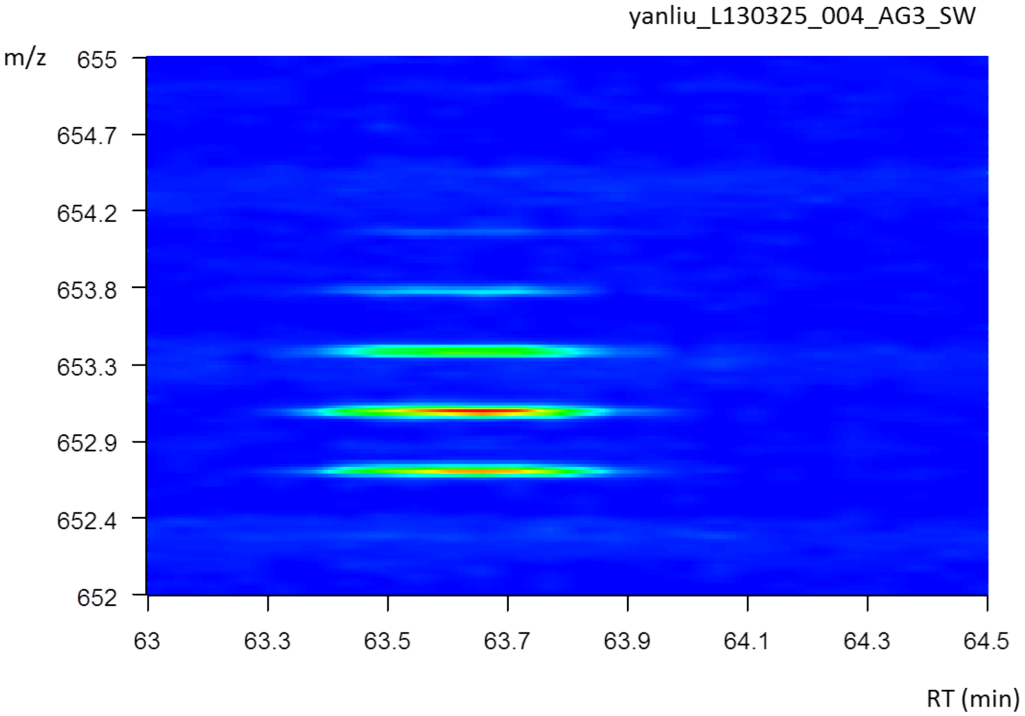


Figure A-8 Example of an ambiguous identification of the deamidated peptide VAPEEHPTLLTEAPLNPK by OpenSWATH targeted search.

Two separate identifications (in unmodified form and in a deamidated form; the site of the modification is shown in red) were reported by OpenSWATH. The two identifications both had an extremely small m_score (from mProphet), i.e. they both were reported as high confidence identifications. The two identifications had almost identical retention times. The MS1 signal image shown above suggests there is only one peptide eluting at RT = 63.58 minutes (precursor *m/z* of 652.68 Da). DIA-Umpire reported only one (unmodified) form.

Table A-1 List of the raw files deposited at ProteomeXchange.

UPS1 and UPS2 are Universal Protein Standards samples from Sigma; E. coli predigested lysate is from Waters; human predigested lysate is from Promega; the affinity-purified samples EIF4A2, MEPCE and the GFP negative control were described in Lambert et al., Nature Methods, 2013. The acquisition method (DDA or DIA/SWATH) is listed, along with the MS1 accumulation time (UPS2 plus E. coli samples). The "ProHits sample ID" refers to the unique identifier for the sample in the Lunenfeld-Tanenbaum LIMS, in project 94.

Short name	Sample	Instrument	Acquisition	DIA MS1 params	Raw file name (in MassIVE)	ProHits sample ID
UPS1_DIA_50ms_MS1	UPS 1 standard (The raw files' name were mislabeled, the corrected name is " Swath_UPS1_40fm-repX")	TripleTOF 5600	DIA	50 ms	Swath_UPS1_40fm_Ecoli_1ug-rep1	17276
					Swath_UPS1_40fm_Ecoli_1ug-rep2	17277
					Swath_UPS1_40fm_Ecoli_1ug-rep3	17278
UPS1_DIA_250ms_MS1	UPS 1 standard (The raw files' name were mislabeled, the corrected name is "LongSwath_UPS1_40fm-repX")	TripleTOF 5600	DIA	250 ms	LongSwath_UPS1_40fm_Ecoli_1ug-rep1	17279
					LongSwath_UPS1_40fm_Ecoli_1ug-rep2	17280
					LongSwath_UPS1_40fm_Ecoli_1ug-rep3	17281
UPS2_Ecoli_DIA_50ms_MS1	UPS2 in E. coli lysate	TripleTOF 5600	DIA	50 ms	Swath_UPS2_5pm_Ecoli_1ug-rep1	17238
					Swath_UPS2_5pm_Ecoli_1ug-rep2	17239
					Swath_UPS2_5pm_Ecoli_1ug-rep3	17240
UPS2_Ecoli_DIA_250ms_MS1	UPS2 in E. coli lysate	TripleTOF 5600	DIA	250 ms	LongSwath_UPS2_5pm_Ecoli_1ug-rep1	17241
					LongSwath_UPS2_5pm_Ecoli_1ug-rep2	17242
					LongSwath_UPS2_5pm_Ecoli_1ug-rep3	17243
UPS2_DDA	UPS2 standard	TripleTOF 5600	DDA	N/A	18185_REP2_4pmol_UPS2_IDA_1	18185
					18187_REP2_4pmol_UPS2_IDA_2	18187
UPS2_DIA_250ms_MS1	UPS2 standard	TripleTOF 5600	DIA	250 ms	18186_REP2_4pmol_UPS2_SWATH_1	18186
					18188_REP2_4pmol_UPS2_SWATH_2	18188
Ecoli_DDA	E. coli lysate (Waters)	TripleTOF 5600	DDA	N/A	18483_REP3_1ug_Ecoli_NewStock2_IDA_1	18483
					18485_REP3_1ug_Ecoli_NewStock2_IDA_2	18485
Ecoli_DIA_2	E. coli lysate (Waters)	TripleTOF	DIA	250 ms	18484_REP3_1ug_Ecoli_NewStock2_SWATH	18484

50ms_MS1		5600			_1	
					18486_REP3_1ug_Ecoli_NewStock2_SWATH_2	18486
Human_DDA	Human lysate (Promega)	TripleTOF 5600	DDA	N/A	18299_REP2_500ng_HumanLysate_IDA_1	18299
					18301_REP2_500ng_HumanLysate_IDA_2	18301
Human_DIA_250ms_MS1	Human lysate (Promega)	TripleTOF 5600	DIA	250 ms	18300_REP2_500ng_HumanLysate_SWATH_1	18300
					18302_REP2_500ng_HumanLysate_SWATH_2	18302
APMS_EIF4A2_DDA	Purification of EIF4A2	TripleTOF 5600	DDA	N/A	IDA_EIF4aJune7_Biorep1	17571
					IDA_EIF4aJune7_Biorep2	17572
					IDA_EIF4aJune7_Biorep3	17573
APMS_EIF4A2_DIA_250ms_MS1	Purification of EIF4A2	TripleTOF 5600	DIA	250 ms	LongSwath_EIF4aJune7_Biorep1	17577
					LongSwath_EIF4aJune7_Biorep2	17578
					LongSwath_EIF4aJune7_Biorep3	17579
APMS_GFP_DDA	Purification of GFP (control)	TripleTOF 5600	DDA	N/A	IDA_GFPJune7_Biorep1	17327
					IDA_GFPJune7_Biorep2	17328
					IDA_GFPJune7_Biorep3	17329
APMS_GFP_DIA_250ms_MS1	Purification of GFP (control)	TripleTOF 5600	DIA	250 ms	LongSwath_GFPJune7_Biorep1	17333
					LongSwath_GFPJune7_Biorep2	17334
					LongSwath_GFPJune7_Biorep3	17335
APMS_MEPCE_DDA	Purification of MEPCE	TripleTOF 5600	DDA	N/A	IDA_MEPCEJune7_Biorep1	17315
					IDA_MEPCEJune7_Biorep2	17316
					IDA_MEPCEJune7_Biorep3	17317
APMS_MEPCE_DIA_250ms_MS1	Purification of MEPCE	TripleTOF 5600	DIA	250 ms	LongSwath_MEPCEJune7_Biorep1	17321
					LongSwath_MEPCEJune7_Biorep2	17322
					LongSwath_MEPCEJune7_Biorep3	17323
Human_QE_DDA	Human lysate (Promega)	Q Exactive Plus	DDA	N/A	140120_Lysate90min_38	20135
Human_QE_DIA	Human lysate (Promega)	Q Exactive Plus	DIA	N/A	140120_Lysate_90min_DIA_53	20136

Appendix B Supplementary materials for Chapter 3

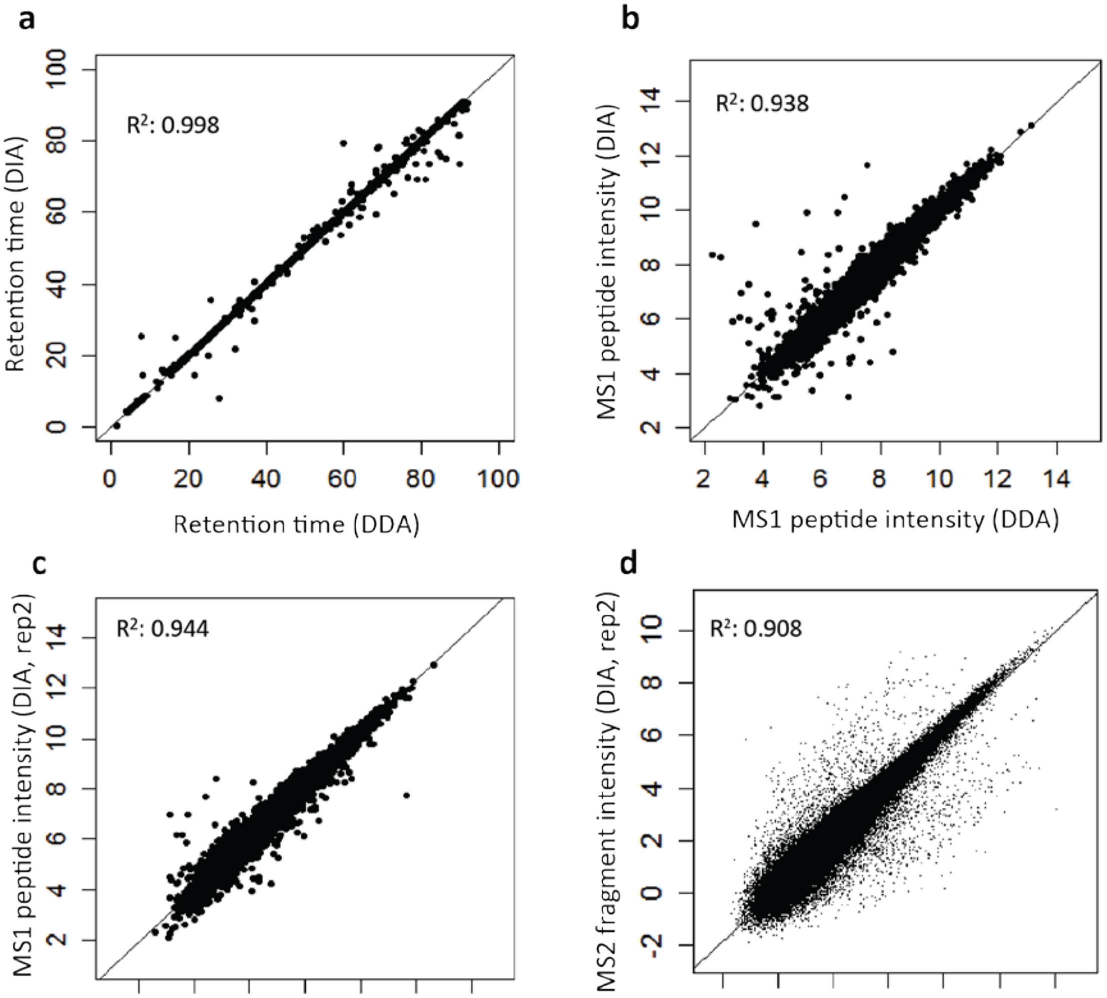


Figure B-1 Assessment of retention time and MS1 intensity reproducibility of identified peptide ions between DDA and DIA (SWATH) experiments.

Human cell lysate data. **(a)** LC retention time (LC peak apex) for peptide ions identified in both DDA and DIA experiments. **(b)** MS1 intensities (monoisotopic peak intensities at LC peak apex) of peptide ions identified commonly by DDA and DIA. **(c)** Reproducibility of DIA MS1 peptide ion intensities between two replicates of DIA data. **(d)** Reproducibility of DIA

MS2 fragment ion intensities (at the reconstructed LC peak apex) of peptide ions between two DIA replicates. Only matched b- and y-ion fragments were considered. Ion and fragment intensities are shown on log2 scale.

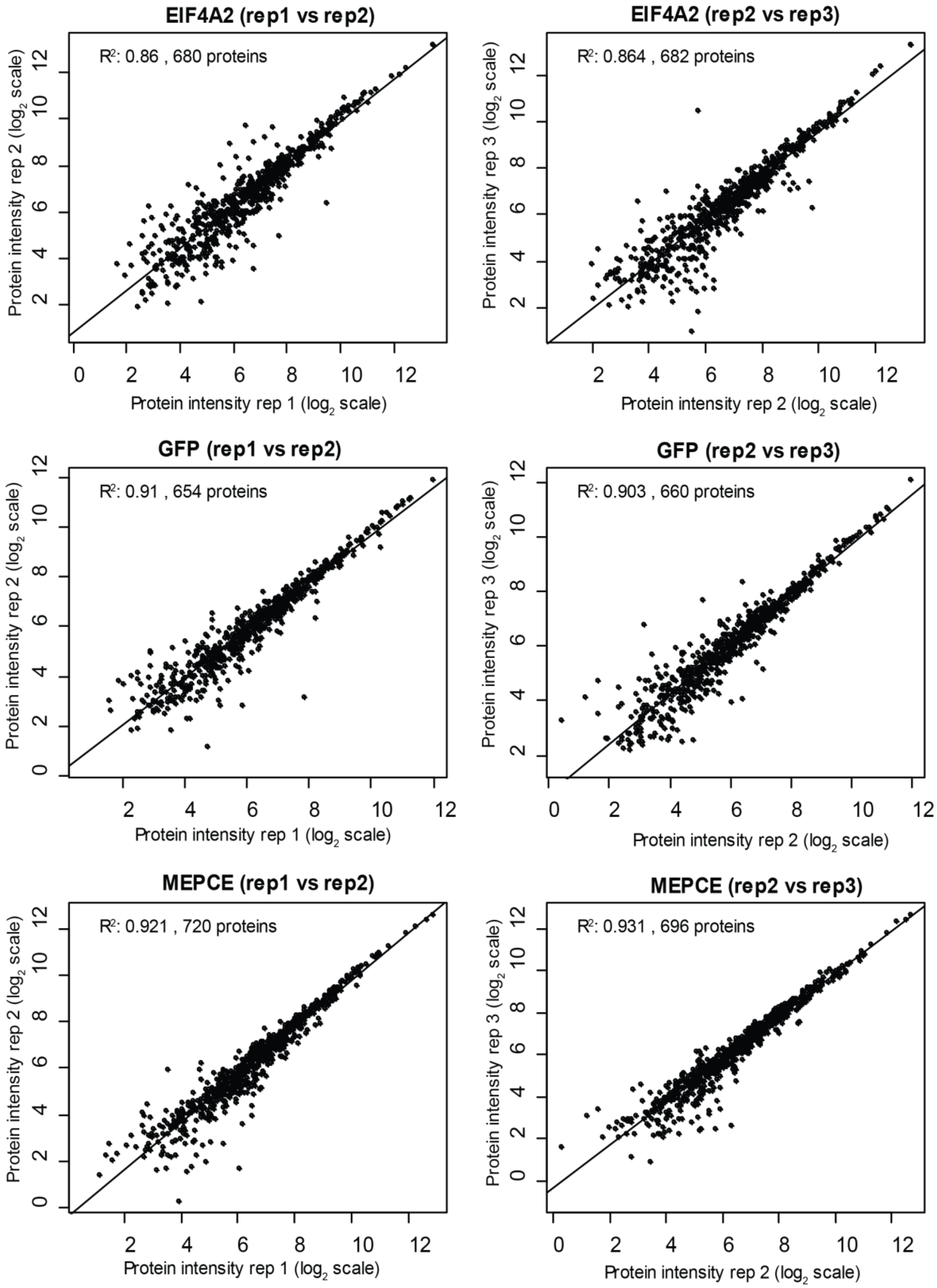


Figure B-2 Protein quantification in AP-SWATH data.

Protein intensities are computed using the “MS2 Top6pep/Top6fra, Freq>0.5’ approach. Each dot represents computed protein intensities for the same protein in two different biological replicates for the same bait (or GFP control).

Appendix C Supplementary materials for Chapter 4

Table C-1 Detailed identification results of HEK-293 Q Exactive dataset

Peptide ion IDs (1% Run level FDR): The number of peptide ion identifications determined at 1% individual run level FDR threshold for each run. Peptide ion IDs (1% Dataset level FDR): The number of peptide ion identifications at 1% dataset level FDR threshold for each run. For DIA datasets, the numbers include the additional IDs from targeted re-extraction (with a 0.99 probability threshold). Peptide ion ID coverage (Dataset level): Percent of peptide ion identifications from the 1% Dataset level FDR peptide ion list that were identified in that particular run. Protein IDs (1% Run level FDR): The number of protein identifications at 1% individual run level FDR threshold for each run. Protein IDs (1% Dataset level FDR): The number of protein identifications at 1% Dataset level FDR threshold for each run. Protein ID coverage (Dataset level): Percent of protein identifications from the 1% Dataset level FDR protein master list identified in that particular run.

	Peptide ion IDs (1% Run level FDR)	Peptide ion IDs (1% Dataset level FDR)	Peptide ion ID coverage (Dataset level)	Protein IDs (1% Run level FDR)	Protein IDs (1% Dataset level FDR)	Protein ID coverage (Dataset level)
S1_R1_DIA	19,945	24,216	70.2%	2,774	3,359	88.3%
S1_R2_DIA	19,836	24,440	70.9%	2,818	3,365	88.5%
S1_R3_DIA	19,075	23,413	67.9%	2,670	3,320	87.3%
S2_R1_DIA	20,271	24,592	71.3%	2,790	3,402	89.5%
S2_R2_DIA	19,548	24,277	70.4%	2,672	3,329	87.6%
S2_R3_DIA	18,650	23,621	68.5%	2,656	3,284	86.4%
S3_R1_DIA	19,673	23,881	69.3%	2,774	3,346	88.0%
S3_R2_DIA	19,386	24,336	70.6%	2,724	3,402	89.5%
S3_R3_DIA	18,693	24,098	69.9%	2,575	3,322	87.4%
S4_R1_DIA	20,491	24,614	71.4%	2,831	3,413	89.8%
S4_R2_DIA	19,748	24,702	71.7%	2,786	3,415	89.8%
S4_R3_DIA	18,662	23,863	69.2%	2,657	3,288	86.5%
S5_R1_DIA	20,864	24,913	72.3%	2,812	3,409	89.7%
S5_R2_DIA	19,749	24,258	70.4%	2,636	3,332	87.6%

S5_R3_DIA	17,611	24,093	69.9%	2,538	3,344	88.0%
S6_R1_DIA	20,037	23,844	69.2%	2,727	3,349	88.1%
S6_R2_DIA	19,893	24,297	70.5%	2,679	3,373	88.7%
S6_R3_DIA	17,831	23,295	67.6%	2,519	3,253	85.6%
S7_R1_DIA	20,279	24,484	71.0%	2,726	3,351	88.1%
S7_R2_DIA	18,703	23,765	68.9%	2,580	3,308	87.0%
S7_R3_DIA	18,292	23,173	67.2%	2,473	3,229	84.9%
S8_R1_DIA	19,710	23,733	68.8%	2,633	3,283	86.3%
S8_R2_DIA	19,270	23,328	67.7%	2,571	3,265	85.9%
S8_R3_DIA	16,343	21,827	63.3%	2,344	3,118	82.0%
S1_R1_DDA	17,823	18,194	46.2%	2,692	2,930	77.1%
S1_R2_DDA	17,459	17,821	45.3%	2,712	2,931	77.1%
S1_R3_DDA	17,109	17,446	44.3%	2,670	2,870	75.5%
S2_R1_DDA	17,625	17,952	45.6%	2,638	2,912	76.6%
S2_R2_DDA	17,074	17,440	44.3%	2,585	2,885	75.9%
S2_R3_DDA	16,608	16,972	43.1%	2,595	2,834	74.6%
S3_R1_DDA	17,319	17,599	44.7%	2,639	2,885	75.9%
S3_R2_DDA	17,938	18,134	46.1%	2,726	2,945	77.5%
S3_R3_DDA	16,536	17,057	43.3%	2,570	2,868	75.5%
S4_R1_DDA	18,543	18,776	47.7%	2,782	2,996	78.8%
S4_R2_DDA	18,231	18,316	46.5%	2,756	2,932	77.1%
S4_R3_DDA	16,496	16,959	43.1%	2,556	2,805	73.8%
S5_R1_DDA	17,938	18,276	46.4%	2,708	2,930	77.1%
S5_R2_DDA	17,162	17,390	44.2%	2,567	2,785	73.3%
S5_R3_DDA	16,703	16,944	43.0%	2,618	2,801	73.7%
S6_R1_DDA	17,645	18,088	45.9%	2,611	2,905	76.4%
S6_R2_DDA	18,030	18,243	46.3%	2,692	2,865	75.4%
S6_R3_DDA	15,940	16,388	41.6%	2,476	2,747	72.3%
S7_R1_DDA	17,539	17,951	45.6%	2,623	2,882	75.8%
S7_R2_DDA	17,688	17,891	45.4%	2,675	2,879	75.7%
S7_R3_DDA	16,283	16,847	42.8%	2,508	2,793	73.5%
S8_R1_DDA	17,893	18,021	45.8%	2,690	2,879	75.7%
S8_R2_DDA	17,198	17,414	44.2%	2,555	2,820	74.2%
S8_R3_DDA	14,813	15,262	38.8%	2,410	2,658	69.9%

Table C-2 Detailed identification results of the microtissue Q Exactive dataset

Peptide ion IDs (1% Run level FDR): The number of peptide ion identifications determined at 1% individual run level FDR threshold for each run. Peptide ion IDs (1% Dataset level FDR): The number of peptide ion identifications at 1% dataset level FDR threshold for each run. For DIA datasets, the numbers include the additional IDs from targeted re-extraction (with a 0.99 probability threshold). Peptide ion ID coverage (Dataset level): Percent of peptide ion identifications from the 1% Dataset level FDR peptide ion list that were identified in that particular run. Protein IDs (1% Run level FDR): The number of protein identifications at 1% individual run level FDR threshold for each run. Protein IDs (1% Dataset level FDR): The number of protein identifications at 1% Dataset level FDR threshold for each run. Protein ID coverage (Dataset level): Percent of protein identifications from the 1% Dataset level FDR protein master list identified in that particular run.

File	Peptide ion IDs (1% Run level FDR)	Peptide ion IDs (1% Dataset level FDR)	Peptide ion ID coverage (Dataset level)	Protein IDs (1% Run level FDR)	Protein IDs (1% Dataset level FDR)	Protein ID coverage (Dataset level)
S1_DIA_R1	16,678	20,060	74.9%	1,889	2,341	88.6%
S1_DIA_R2	17,254	20,160	75.3%	1,921	2,333	88.3%
S1_DIA_R3	17,339	19,994	74.7%	1,921	2,355	89.2%
S3_DIA_R1	16,550	20,408	76.2%	1,828	2,341	88.6%
S3_DIA_R2	16,945	20,612	77.0%	1,891	2,368	89.7%
S3_DIA_R3	16,791	20,191	75.4%	1,881	2,332	88.3%
S4_DIA_R1	16,639	20,030	74.8%	1,818	2,293	86.8%
S4_DIA_R2	17,561	21,038	78.6%	1,893	2,393	90.6%
S4_DIA_R3	17,633	20,644	77.1%	1,900	2,369	89.7%
S7_DIA_R1	17,841	21,264	79.4%	1,970	2,396	90.7%
S7_DIA_R2	18,093	21,227	79.3%	1,996	2,412	91.3%
S7_DIA_R3	17,778	20,574	76.9%	1,926	2,375	89.9%
S9_DIA_R1	17,068	19,810	74.0%	1,896	2,318	87.8%
S9_DIA_R2	17,507	20,227	75.6%	1,896	2,365	89.5%
S9_DIA_R3	17,380	20,307	75.9%	1,969	2,356	89.2%
pool_DDA_R1	16,514	16,607	53.0%	2,156	2,253	81.1%
pool_DDA_R2	17,027	16,979	54.2%	2,150	2,255	81.2%
S1_DDA	12,529	13,195	42.1%	1,787	2,014	72.5%
S3_DDA	15,966	16,034	51.2%	2,115	2,206	79.4%
S7_DDA	16,846	16,857	53.8%	2,187	2,258	81.3%
S9_DDA	15,941	16,093	51.4%	2,121	2,229	80.2%

Table C-3 Detailed identification results of Orbitrap Fusion dataset

Peptide ion IDs (1% Run level FDR): The number of peptide ion identifications determined at 1% individual run level FDR threshold for each run. Peptide ion IDs (1% Dataset level FDR): The number of peptide ion identifications at 1% dataset level FDR threshold for each run. For DIA datasets, the numbers include the additional IDs from targeted re-extraction (with a 0.99 probability threshold). Peptide ion ID coverage (Dataset level): Percent of peptide ion identifications from the 1% Dataset level FDR peptide ion list that were identified in that particular run. Protein IDs (1% Run level FDR): The number of protein identifications at 1% individual run level FDR threshold for each run. Protein IDs (1% Dataset level FDR): The number of protein identifications at 1% Dataset level FDR threshold for each run. Protein ID coverage (Dataset level): Percent of protein identifications from the 1% Dataset level FDR protein master list identified in that particular run.

File	Peptide ion IDs (1% Run level FDR)	Peptide ion IDs (1% Dataset level FDR)	Peptide ion ID coverage (Dataset level)	Protein IDs (1% Run level FDR)	Protein IDs (1% Dataset level FDR)	Protein ID coverage (Dataset level)
DIA 5Da R1	28,719	30,336	76.6%	3,846	4,066	92.3%
DIA 5Da R2	29,434	31,014	78.3%	3,854	4,101	93.1%
DIA 5Da R3	29,341	30,604	77.3%	3,858	4,101	93.1%
DIA 10Da R1	31,941	34,117	82.6%	3,691	4,082	93.2%
DIA 10Da R2	33,159	34,946	84.6%	4,009	4,220	96.3%
DIA 10Da R3	33,449	34,818	84.3%	3,962	4,190	95.6%
DIA 15Da R1	29,862	31,419	86.2%	3,545	3,788	95.7%
DIA 15Da R2	29,953	31,494	86.4%	3,598	3,797	95.9%
DIA 15Da R3	29,783	31,514	86.5%	3,616	3,818	96.4%
DIA 20Da R1	26,964	28,606	86.0%	3,342	3,547	96.1%
DIA 20Da R2	26,605	28,419	85.5%	3,348	3,530	95.6%
DIA 20Da R3	26,739	28,605	86.0%	3,330	3,532	95.7%
DIA 25Da R1	23,924	25,926	85.9%	3,125	3,373	96.0%
DIA 25Da R2	23,880	25,956	86.0%	3,052	3,367	95.8%
DIA 25Da R3	24,199	26,033	86.3%	3,101	3,385	96.3%
DDA1 R1	31,851	32,011	79.2%	4,256	4,378	91.6%
DDA1 R2	31,732	31,944	79.0%	4,257	4,409	92.3%
DDA1 R3	32,003	32,143	79.5%	4,314	4,413	92.3%
DDA2 R1	29,623	30,075	71.8%	4,102	4,284	90.6%
DDA2 R2	29,813	30,186	72.1%	4,123	4,310	91.1%
DDA2 R3	30,813	30,847	73.7%	4,227	4,319	91.3%

Bibliography

1. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* Journal of proteomics, 2010. **73**(11): p. 2092-123.
2. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Analytical chemistry, 2003. **75**(17): p. 4646-58.
3. Mann, M., R.C. Hendrickson, and A. Pandey, *Analysis of proteins and proteomes by mass spectrometry.* Annual review of biochemistry, 2001. **70**: p. 437-73.
4. Davis, M.T., et al., *Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses.* Proteomics, 2001. **1**(1): p. 108-17.
5. Kohli, B.M., et al., *An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments.* Rapid communications in mass spectrometry : RCM, 2005. **19**(5): p. 589-96.
6. Kohlbacher, O., et al., *TOPP--the OpenMS proteomics pipeline.* Bioinformatics, 2007. **23**(2): p. e191-7.
7. Zhang, Y., et al., *Effect of dynamic exclusion duration on spectral count based quantitative proteomics.* Analytical chemistry, 2009. **81**(15): p. 6317-26.
8. Purvine, S., et al., *Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer.* Proteomics, 2003. **3**(6): p. 847-50.
9. Venable, J.D., et al., *Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra.* Nature methods, 2004. **1**(1): p. 39-45.
10. Plumb, R.S., et al., *UPLC/MS(E); a new approach for generating molecular fragment information for biomarker structure elucidation.* Rapid communications in mass spectrometry : RCM, 2006. **20**(13): p. 1989-94.
11. Panchaud, A., et al., *Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean.* Analytical chemistry, 2009. **81**(15): p. 6481-8.
12. Geiger, T., J. Cox, and M. Mann, *Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation.* Molecular & cellular proteomics : MCP, 2010. **9**(10): p. 2252-61.
13. Bern, M., et al., *Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry.* Analytical chemistry, 2010. **82**(3): p. 833-41.

14. Carvalho, P.C., et al., *XDIA: improving on the label-free data-independent analysis.* Bioinformatics, 2010. **26**(6): p. 847-8.
15. Panchaud, A., et al., *Faster, quantitative, and accurate precursor acquisition independent from ion count.* Analytical chemistry, 2011. **83**(6): p. 2250-7.
16. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Molecular & cellular proteomics : MCP, 2012. **11**(6): p. O111 016717.
17. Silva, J.C., et al., *Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition.* Molecular & cellular proteomics : MCP, 2006. **5**(1): p. 144-56.
18. Egertson, J.D., et al., *Multiplexed MS/MS for improved data-independent acquisition.* Nature methods, 2013.
19. Lange, V., et al., *Selected reaction monitoring for quantitative proteomics: a tutorial.* Molecular systems biology, 2008. **4**: p. 222.
20. Doerr, A., *Mass spectrometry-based targeted proteomics.* Nature methods, 2013. **10**(1): p. 23.
21. Marx, V., *Targeted proteomics.* Nature methods, 2013. **10**(1): p. 19-22.
22. Aebersold, R., A.L. Burlingame, and R.A. Bradshaw, *Western Blots versus Selected Reaction Monitoring Assays: Time to Turn the Tables?* Molecular & cellular proteomics : MCP, 2013. **12**(9): p. 2381-2.
23. Liu, Y., et al., *Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS.* Proteomics, 2013. **13**(8): p. 1247-56.
24. Findlay, G.M., et al., *Interaction domains of Sos1/Grb2 are finely tuned for cooperative control of embryonic stem cell fate.* Cell, 2013. **152**(5): p. 1008-20.
25. Collins, B.C., et al., *Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system.* Nature methods, 2013.
26. Lambert, J.P., et al., *Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition.* Nature methods, 2013.
27. Bruderer, R., et al., *Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues.* Molecular & cellular proteomics : MCP, 2015. **14**(5): p. 1400-10.
28. Rost, H.L., et al., *OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.* Nature biotechnology, 2014. **32**(3): p. 219-23.
29. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.
30. Reiter, L., et al., *mProphet: automated data processing and statistical validation for large-scale SRM experiments.* Nature methods, 2011. **8**(5): p. 430-5.
31. Teleman, J., et al., *DIANA--algorithmic improvements for analysis of data-independent acquisition MS data.* Bioinformatics, 2015. **31**(4): p. 555-62.
32. Selevsek, N., et al., *Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry.* Molecular & cellular proteomics : MCP, 2015. **14**(3): p. 739-49.

33. Chang, R.Y., et al., *SWATH analysis of the synaptic proteome in Alzheimer's disease.* Neurochemistry international, 2015. **87**: p. 1-12.

34. Sidoli, S., et al., *SWATH Analysis for Characterization and Quantification of Histone Post-translational Modifications.* Molecular & cellular proteomics : MCP, 2015.

35. Liu, Y., et al., *Glycoproteomic analysis of prostate cancer tissues by SWATH mass spectrometry discovers N-acylethanolamine acid amidase and protein tyrosine kinase 7 as signatures for tumor aggressiveness.* Molecular & cellular proteomics : MCP, 2014. **13**(7): p. 1753-68.

36. Rosenberger, G., et al., *A repository of assays to quantify 10,000 human proteins by SWATH-MS.* Scientific data, 2014. **1**: p. 140031.

37. Collins, B.C., et al., *Quantifying protein interaction dynamics by SWATH mass spectrometry: application to the 14-3-3 system.* Nature methods, 2013. **10**(12): p. 1246-53.

38. Lambert, J.P., et al., *Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition.* Nature methods, 2013. **10**(12): p. 1239-45.

39. Guo, T., et al., *Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps.* Nature medicine, 2015. **21**(4): p. 407-13.

40. Caron, E., et al., *An open-source computational and data resource to analyze digital maps of immunopeptidomes.* eLife, 2015. **4**.

41. Tsou, C.C., et al., *DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics.* Nature methods, 2015. **12**(3): p. 258-64, 7 p following 264.

42. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nature biotechnology, 2012. **30**(10): p. 918-20.

43. Du, P., W.A. Kibbe, and S.M. Lin, *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.* Bioinformatics, 2006. **22**(17): p. 2059-65.

44. Tautenhahn, R., C. Bottcher, and S. Neumann, *Highly sensitive feature detection for high resolution LC/MS.* BMC bioinformatics, 2008. **9**: p. 504.

45. Nesvizhskii, A.I., et al., *Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides.* Molecular & cellular proteomics : MCP, 2006. **5**(4): p. 652-70.

46. Kryuchkov, F., et al., *Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry.* Journal of proteome research, 2013. **12**(7): p. 3362-71.

47. Craig, R., J.P. Cortens, and R.C. Beavis, *Open source system for analyzing, validating, and storing protein identification data.* Journal of proteome research, 2004. **3**(6): p. 1234-1242.

48. Eng, J.K., T.A. Jahan, and M.R. Hoopmann, *Comet: an open-source MS/MS sequence database search tool.* Proteomics, 2013. **13**(1): p. 22-4.
49. Kim, S., et al., *The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search.* Molecular & cellular proteomics : MCP, 2010. **9**(12): p. 2840-52.
50. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Analytical chemistry, 2002. **74**(20): p. 5383-92.
51. Deutsch, E.W., et al., *A guided tour of the Trans-Proteomic Pipeline.* Proteomics, 2010. **10**(6): p. 1150-9.
52. Shteynberg, D., et al., *iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates.* Molecular & cellular proteomics : MCP, 2011. **10**(12): p. M111 007690.
53. Lam, H., et al., *Building consensus spectral libraries for peptide identification in proteomics.* Nature methods, 2008. **5**(10): p. 873-5.
54. Escher, C., et al., *Using iRT, a normalized retention time for more targeted measurement of peptides.* Proteomics, 2012. **12**(8): p. 1111-21.
55. Vizcaino, J.A., et al., *ProteomeXchange provides globally coordinated proteomics data submission and dissemination.* Nature biotechnology, 2014. **32**(3): p. 223-6.
56. Moruz, L., et al., *Chromatographic retention time prediction for posttranslationally modified peptides.* Proteomics, 2012. **12**(8): p. 1151-9.
57. Choi, H., et al., *QPROT: Statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics.* Journal of proteomics, 2015. **129**: p. 121-6.
58. Teo, G., et al., *mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry.* Journal of proteomics, 2015. **129**: p. 108-20.
59. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nature biotechnology, 2008. **26**(12): p. 1367-72.
60. Sturm, M., et al., *OpenMS - an open-source software framework for mass spectrometry.* BMC bioinformatics, 2008. **9**: p. 163.
61. Tsou, C.C., et al., *IDEAL-Q, an automated tool for label-free quantitation analysis using an efficient peptide alignment approach and spectral data validation.* Molecular & cellular proteomics : MCP, 2010. **9**(1): p. 131-44.
62. Barsnes, H., et al., *compomics-utilities: an open-source Java library for computational proteomics.* BMC bioinformatics, 2011. **12**: p. 70.
63. Lam, H., E.W. Deutsch, and R. Aebersold, *Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics.* Journal of proteome research, 2010. **9**(1): p. 605-10.
64. Cox, J., A. Michalski, and M. Mann, *Software lock mass by two-dimensional minimization of peptide mass errors.* Journal of the American Society for Mass Spectrometry, 2011. **22**(8): p. 1373-80.

65. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.

66. Choi, H., et al., *SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments.* Journal of proteome research, 2012. **11**(4): p. 2619-24.

67. Ludwig, C., et al., *Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry.* Molecular & cellular proteomics : MCP, 2012. **11**(3): p. M111 013987.

68. Nesvizhskii, A.I., *Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments.* Proteomics, 2012. **12**(10): p. 1639-55.

69. Choi, H., et al., *SAINT: probabilistic scoring of affinity purification-mass spectrometry data.* Nature methods, 2011. **8**(1): p. 70-3.

70. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2013 update.* Nucleic acids research, 2013. **41**(Database issue): p. D816-23.

71. Jeronimo, C., et al., *Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme.* Mol Cell, 2007. **27**(2): p. 262-74.

72. Prakash, A., et al., *Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis.* Journal of proteome research, 2014.

73. Bilbao, A., et al., *Processing strategies and software solutions for data-independent acquisition in mass spectrometry.* Proteomics, 2015. **15**(5-6): p. 964-80.

74. Picotti, P., et al., *A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis.* Nature, 2013. **494**(7436): p. 266-70.

75. Schubert, O.T., et al., *Absolute Proteome Composition and Dynamics during Dormancy and Resuscitation of Mycobacterium tuberculosis.* Cell host & microbe, 2015. **18**(1): p. 96-108.

76. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.

77. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nature communications, 2014. **5**: p. 5277.

78. Kall, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nature methods, 2007. **4**(11): p. 923-5.

79. Vaudel, M., et al., *PeptideShaker enables reanalysis of MS-derived proteomics data sets.* Nature biotechnology, 2015. **33**(1): p. 22-4.

80. Egertson, J.D., et al., *Multiplexed MS/MS for improved data-independent acquisition.* Nature methods, 2013. **10**(8): p. 744-6.

81. Prakash, A., et al., *Hybrid data acquisition and processing strategies with increased throughput and selectivity: pSMART analysis for global qualitative and quantitative analysis.* Journal of proteome research, 2014. **13**(12): p. 5415-30.

82. Holman, J.D., D.L. Tabb, and P. Mallick, *Employing ProteoWizard to Convert Raw Mass Spectrometry Data.* Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 2014. **46**: p. 13 24 1-9.

83. Kirchner, M., et al., *Non-linear classification for on-the-fly fractional mass filtering and targeted precursor fragmentation in mass spectrometry experiments.* Bioinformatics, 2010. **26**(6): p. 791-7.

84. Tiller, P.R., et al., *Fractional mass filtering as a means to assess circulating metabolites in early human clinical studies.* Rapid communications in mass spectrometry : RCM, 2008. **22**(22): p. 3510-6.

85. Toumi, M.L. and H. Desaire, *Improving mass defect filters for human proteins.* Journal of proteome research, 2010. **9**(10): p. 5492-5.

86. Toprak, U.H., et al., *Conserved peptide fragmentation as a benchmarking tool for mass spectrometers and a discriminating feature for targeted proteomics.* Molecular & cellular proteomics : MCP, 2014. **13**(8): p. 2056-71.

87. Robin, S., et al., *A semi-parametric approach for mixture models: Application to local false discovery rate estimation.* Computational statistics & data analysis, 2007. **51**(12): p. 5483-5493.

88. Choi, H., D. Ghosh, and A.I. Nesvizhskii, *Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling.* Journal of proteome research, 2008. **7**(1): p. 286-92.

89. Silverman, B.W., *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability1998, Boca Raton: Chapman & Hall/CRC. ix, 175 p.

90. Teo, G., et al., *mapDIA: Preprocessing and statistical analysis of quantitative proteomics data from data independent acquisition mass spectrometry.* Journal of proteomics, 2015. **129**: p. 108-120.

91. Li, Y., et al., *Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files.* Nature methods, 2015. **12**(12): p. 1105-6.

92. Wang, J., et al., *MSPLIT-DIA: sensitive peptide identification for data-independent acquisition.* Nature methods, 2015. **12**(12): p. 1106-8.

93. Weisbrod, C.R., et al., *Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification.* Journal of proteome research, 2012. **11**(3): p. 1621-32.