Untargeted, spectral library-free analysis of data independent acquisition proteomics data generated using Orbitrap mass spectrometers

Chih-Chiang Tsou[1], Chia-Feng Tsai[2], Guoci Teo[3], Yu-Ju Chen[2], Alexey I. Nesvizhskii[1,3]

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

[2] Institute of Chemistry, Academia Sinica, Taiwan.

[3] Department of Pathology, University of Michigan, Ann Arbor, MI 48109, USA

Corresponding author:

Alexey I. Nesvizhskii

Department of Pathology, University of Michigan

4237 Medical Science I

Ann Arbor, MI, 48109

Email: nesvi@med.umich.edu

Tel: +1 734 764 3516

Abbreviations

DIA: Data Independent Acquisition

DDA: Data Dependent Acquisition

SWATH: Sequential Window Acquisition of all THeoretical Mass Spectra

MRM: Multiple Reaction Monitoring

SRM: Selected Reaction Monitoring

MS: Mass Spectrometry

PSM: Peptide-Spectrum Match

HCD: Higher-Collisional Dissociation

FDR: False Discovery Rate

EM: Expectation Maximization

ROC: Receiver Operating Characteristic

ABSTRACT

We describe an improved version of the data independent acquisition (DIA) computational analysis tool DIA-Umpire, and show that it enables highly sensitive, untargeted and direct (spectral library-free) analysis of DIA data obtained using the Orbitrap family of mass spectrometers. DIA-Umpire v2 implements an improved feature detection algorithm with two additional filters based on the isotope pattern and fractional peptide mass analysis. The targeted re-extraction step of DIA-Umpire is updated with an improved scoring function and a more robust, semi-parametric mixture modeling of the resulting scores for computing posterior probabilities of correct peptide identification in a targeted setting. Using two publicly available Q Exactive DIA datasets generated using HEK-293 cells and human liver microtissues, we demonstrate that DIA-Umpire can identify similar number of peptide ions, but with better identification reproducibility between replicates and samples, as with conventional data dependent acquisition (DDA). We further demonstrate the utility of DIA-Umpire using a series of Orbitrap Fusion DIA experiments with HeLa cell lysates profiled using conventional DDA and using DIA with different isolation window widths.

SIGNIFICANCE OF THE STUDY

As data independent acquisition mass spectrometry emerging as a promising technique, development of computational analysis tool for DIA data obtained from a wide range of mass spectrometers is the next critical step to facilitate its adoption for a board range of proteomics applications. The computational tool, DIA-Umpire v2, presented in this work is capable of

highly sensitive, untargeted analysis of DIA data from complex protein samples generated using the Orbitrap family of mass spectrometers. The tool supports various DIA strategies and mass spectrometers. Most importantly, the workflow is not completely dependent on a spectral library and is 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.

INTRODUCTION

Data independent acquisition (DIA) mass spectrometry (MS) [1-4] has recently emerged as a promising alternative to data dependent acquisition (DDA) for quantitative proteomics analysis (for a recent review, see [5]). The fundamental concept of DIA is to acquire fragment ion information for all precursor peptide ions within a certain window of m/z values, sequentially covering the entire range of relevant m/z values. This strategy is exemplified using the SWATH-MS [3] approach, and is now available on most instrument platforms. At present, DIA data is most commonly analyzed using targeted extraction tools such as OpenSWATH [6], Spectronaut [7], PeakView, and Skyline [8] for extraction of quantification information from DIA data, and tools for statistical scoring of extracted signals such as mProphet [9]. These tools are dependent on the availability of spectral libraries, typically built from DDA data acquired in parallel with DIA data from the same or similar samples. Recent studies have further advanced such targeted extraction approaches to various

proteomics applications [10-18] including post-translational modifications [12, 13], protein-protein interaction [13, 14], protein heritability analysis [19], and immunopeptidomics analysis [18].

We have recently described an alternative workflow, DIA-Umpire [20], for untargeted and direct (i.e. spectral library-free) analysis of DIA data. The feature detection algorithm of DIA-Umpire detects peptide and fragment ion features, and uses their peak elution similarities to group detected fragment and precursor signals. The detected m/z and intensity values of grouped signals are then assembled into pseudo MS/MS spectra that are fully compatible with any analysis tools developed for DDA data, including MS/MS database search engines (e.g. X! Tandem [21], Comet [22], MSGF+ [23]), peptide-spectrum match (PSM) statistical validation (PeptideProphet [24], Percolator [25], PeptideShaker [26]) and protein inference tools such as ProteinProphet [27]. We have demonstrated that reliable quantification can be obtained from both MS2 fragment ion intensities and from MS1 precursor peptide ion intensities. We have also demonstrated and implemented in DIA-Umpire an optional hybrid workflow, which builds an internal library from confident identifications from database search results when multiple DIA runs are available. This "internal" (i.e. DIA-derived) library can then be used to query preprocessed precursor-fragment groups using the second, targeted re-extraction step to reduce the number of missing identifications (quantifications) across all experiments from the same dataset. It should also be noted that DIA-Umpire-derived identifications are compatible with other targeted extraction tools, i.e. a DIA-derived spectral library can be built using tools such as SpectraST

[28], with the subsequent interrogation of the data using that library with targeted extraction tools such as Skyline or OpenSWATH.

Because most of the recent studies used DIA (SWATH-MS) data generated using AB Sciex 5600 instruments, we sought to evaluate the performance of the DIA-Umpire computational strategy on 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 [7, 29, 30]. The Orbitrap mass analyzer, available in both the Q Exactive and the Orbitrap Fusion instruments, enables acquisition of tandem mass spectra with high mass accuracy and scan rate – two of the main prerequisites for successful interrogation of complex samples using DIA data. Here we present DIA-Umpire v2, the new version of the software that enables analysis of complex DIA datasets generated using the Orbitrap instruments. We describe 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 approach for computing the probabilities of correct identifications of peptide signals in DIA data at the targeted re-extraction stage. Using two Q Exactive DIA and DDA datasets published by Bruderer *et al*. [7], and a series of human HeLa cell line experiments on an Orbitrap Fusion performed as part of this work, we show that DIA-Umpire v2 enables highly sensitive analysis of DIA data.

METHODS

Q Exactive datasets

The raw files for two sets of Q Exactive DIA and DDA data described in [7] 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.

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). HeLa cells (predigested using trypsin) were purchased from Thermo Scientific (San Jose, CA). 1ug of HeLa cells were loaded onto self-packed analytical column (300 mm length x 100 um i.d.) using 3 um ReproSil-Pur C18-AQ particles (Dr. Maisch, Ammerbuch, Germany). 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 gradient initially started from 1% B to 2% B for 2 mins and then was increased linearly to 25% B at 112 min, to 35% at 122 min, then to 90% B at 123 min, held for 6 mins. Finally, the gradient was decreased linearly to 1% at 130 min and held for 20 min for re-equilibration.

The MS instrument was operated in the positive ion mode, with an electrospray through a heated ion transfer tube (250 °C). Full scan MS spectra were acquired in the Orbitrap mass analyzer (m/z range: 400−1250 Da) with the resolution set to 60,000 (FWHM) at m/z 200 Da.

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 (FWHM) at m/z 200 Da 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 performed 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.

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 the 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.

DIA-Umpire pseudo MS/MS extraction

All .raw files were converted into mzXML format using msconvert.exe (version 3.0.6721) [31] with vendor (Xcalibur version 2.3-176401/2.3.0.1765) 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 the Orbitrap Fusion datasets and 15 ppm for the 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 the Orbitrap Fusion datasets and 25 ppm for the 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 the algorithm allowed missing peaks in up to two consecutive MS1 scans for detection of single m/z trace signals. Because the signal quality of the centroid spectra generated using Xcalibur library via msconvert.exe was manually inspected and deemed to be sufficiently high, no additional background detection

and noise removal was used in the DIA-Umpire_SE module. Furthermore, because the MS2 scans in the resulting mzXML files contained the isolation window ranges there was no need to specify these settings in the parameter file of the DIA-Umpire_SE module.

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 isotope 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 [32-34]. These studies have shown that the mass values of certain molecules (e.g. tryptic peptides and metabolites) distribute in specific fractional number regions. This characteristic can be used to detect false signals and to reduce the number of false positive peptide identifications. In this study, we adopted the fractional mass boundary equations described in Toumi *et al* [34] which were 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; parameter file option). For each detected precursor ion or fragment ion feature with neutral mass M, the fractional mass $D(M)$ is calculated as

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

$\lfloor M \rceil$ is the largest integer not greater than M. 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:

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

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 were calculated from all human tryptic peptides. The isotope peak ratios up to the 10th isotopic peak were established in DIA-Umpire by generating 9 (from the 2nd isotopic peak to the 10th isotopic peak) scatter plots (Supplementary Figure 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 Supplementary Figure 1). For a possible peak feature detected with peak intensities $I = (I_1, I_2, \ldots, I_n)$ and neutral mass M, the observed peak ratios $O = (O_2, \ldots, O_n)$, O_i

= I_i / I_1, were calculated, where n is the isotope peak number (n = 1 refers to the monoisotopic peak). 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: $H_i = \mu_i + 3.3\times\sigma_i$ and $L_i=\mu_i - 3.3\times\sigma_i$. 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 the 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.

DDA MS/MS 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 the Orbitrap Fusion data and to 15 ppm and 25 ppm, respectively, for the 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; 20,200 sequences), 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 [35] followed by ProteinProphet [27].

FDR estimation independently for each DDA/DIA run

The false discovery rates (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 [36]. The numbers of peptide ions at 1% FDR determined independently for each run (column "Peptide ion IDs (1% Run level FDR)") are shown in Supplementary Table 1 (Q Exactive HEK-293 data), in Supplementary Table 2 (Q Exactive liver microtissue data), and in Supplementary Table 3 (Orbitrap Fusion HeLa 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 number of protein groups determined independently for each run at 1% FDR are also shown in Supplementary Tables 1-3 (column "Protein IDs (1% Run level FDR)").

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 three replicate runs 'DDA1 R1', 'DDA1 R2', and 'DDA1 R3'). If a 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 fairer 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 run determined using the dataset level FDR strategy is shown in Supplementary Tables 1-3 (column "Peptide ion IDs (1% FDR dataset level)").

FDR for protein identifications in DDA data at the dataset level

To estimate protein FDR for DDA data at the dataset level, ProteinProphet [27] 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 using the target-decoy approach based on the maximum peptide ion probability across all files within the dataset. At 1% FDR, the 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 that individual run if it had at least one peptide ion identified in that run that was included in the 1% dataset level FDR list. The number of protein identifications for individual DDA runs counted using the dataset level FDR strategy is shown Supplementary Tables 1-3 (column "Protein IDs (1% Dataset level FDR)").

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

Analysis of DIA data using DIA-Umpire includes an additional targeted data extraction step using 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 extraction module (DIA-Umpire_Quant.jar) to generate an internal spectral library and perform targeted re-extraction analysis [20] to further reduce the number of 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 for quantification described in Tsou *et al* [20]. 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.

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 detail. A precursor-fragment group G generated by DIA-Umpire, and a library spectrum S, represented as

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

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

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 $INT^{G-S} = (I_1^G, I_2^G, \ldots, I_{NS}^G)$ of length NS, with I_r^G taken as the intensity of the fragment peak *r* in G that matches 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 D_M around M_r^S. Thus, INT^{G-S} contains *L* non-zero values, where *L* is the total number of matched fragments between G and S (*L* ≤ NS). The following nine sub-scores are calculated during the spectral matching:

1. Spectral Similarity Score (SSS), in DIA-Umpire v2 calculated using the Dot product scoring described in Toprak et al.[37] between the vector INT^{G-S} and the library spectrum intensity vector $(I_1^S, I_2^S, \ldots, I_{NS}^S)$.

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 earlier for DIA-Umpire [20], 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):

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

6. Retention time difference between the predicted retention time and the observed monoisotope peak apex retention time of the precursor peptide anchoring precursor-fragment group G.

7. Precursor isotope peak correlation score, computed as the Pearson correlation coefficient between the monoisotope 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 earlier in Methods.

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 are trained for each dataset as described for DIA-Umpire previously [20].

Posterior probabilities of correct identification at the targeted extraction step

The probability calculation in DIA-Umpire v2 has been revised to implement a more robust semi-parametric mixture modeling approach. 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* [38], which has been described for PSM validation by Choi *et al* [39] 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, so that 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 approach is to estimate π_0 to avoid the over-fitting problem (maximum likelihood will be always at the point when $\pi_1$equals 1 [38]) 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 [40]. The initial estimation of $f_1(U)$ is done by the DIA-Umpire's original Gaussian mixture modeling approach [20]. 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 the correct distribution is updated as

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

The EM algorithm iterates until the difference of log-likelihoods between two consecutive iterations is less than 0.00001 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 with the updated priors.

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 after 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 prints the U-score probabilities calculated as described above, which are labeled as iProphet probabilities 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 from the targeted re-extraction step.

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 5Da' dataset consisting of the three replicate runs 'DIA 5Da R1', 'DIA 5Da R1', and 'DIA 5Da R3'). If a 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%, because no new identifications are added at this step. The number of peptide ion identifications for each DIA run is shown in Supplementary Tables 1-3 (column "Peptide ion IDs (1% Dataset level FDR)").

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 [27] 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 [36] 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). The master protein list corresponding to 1% FDR for each dataset was 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 at 1% dataset level FDR in untargeted database search or with a probability 0.99 or higher at the targeted re-

extraction step. The number of protein identifications obtained this way is shown in Supplementary Tables 1-3 (column "Protein IDs (1% Dataset level FDR)").

Data availability

All Orbitrap Fusion mass spectrometry data files and DIA-Umpire results for all the datasets presented in this paper have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/) via the PRIDE partner repository with the data set identifier PXD003179. During the reviewing process, reviewers can access the data (https://www.ebi.ac.uk/pride/archive/login) using the following account details: Username: reviewer12096@ebi.ac.uk, and Password: mtlM1GSv

RESULTS AND DISCUSSIONS

Improved feature detection using fractional mass and isotope pattern filters

The DIA-Umpire workflow relies on accurate and sensitive detection of precursor and fragment ion signals. 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 necessarily 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 precursor-fragment grouping algorithm of DIA-Umpire 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 extracted features and the total computation costs. To address this issue, we introduced two new filters, the fractional mass and the isotope pattern filters**,** in DIA-Umpire v2 to remove detected precursor ion and fragment features that are less likely to be true peptide features (see Methods for details).

We first investigated the effects of these new feature detection filters using two DIA runs, one from the Orbitrap Fusion (10 Da isolation window) HeLa cell lysate dataset generated as part of this work, and the other from the publicly available Q Exactive HEK-293 cell lysate dataset [10] (see Methods for details regarding the experimental datasets). We processed these two DIA runs through the 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 combined using iProphet. Peptide ion identifications were filtered to achieve 1% peptide ion level FDR (see Methods for details regarding MS/MS database search and FDR calculations). All confidently identified peptide ions were linked to the corresponding detected precursor peptide ion features.

In total, there were 416,607 and 812,944 precursor ion 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. Figures 1A and 1B 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, Figures 1C and 1D show the number of identified precursor features at different isotope pattern fitness probability thresholds. The majority of the identified features had an isotope pattern fitness 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 fitness probabilities. Some of these cases may be due to co-elution with other high abundance peptide ion signals, whereas others could be false identifications. Additional analysis showed that the detected features with extremely low isotope pattern probabilities were mostly lower abundance signals (Supplementary Figure 2). Overall, the isotope pattern fitness probability threshold was found to be useful for more effective removal of false features.

By combining the two filters, the fractional mass filter and the isotope pattern filter, DIA-Umpire v2 reduced the number of extracted features more effectively and without a significant reduction in the number of identified peptides. Further analysis (Supplementary Table 4) showed that the filters were able to reduce the computation time for DIA-Umpire signal processing step and the number of pseudo MS/MS spectra generated, in turn reducing the MS/MS database search time. Figures 1E and 1F show the receiver operating characteristic (ROC) curves of the detected features for the two DIA runs. Based on this analysis, for the remainder of this study we applied the fractional mass filter with d=0.1 and the isotope pattern fitness probability threshold of 0.3. These parameters were also selected as defaults in DIA-Umpire v2. Note that these two filters were developed based on prior information available from the analysis of human tryptic peptides. They may not be applicable to data from other organisms or proteolytic enzymes, however the filtering thresholds can be adjusted (or the filters turned off altogether) in the DIA-Umpire_SE parameters file.

Application of DIA-Umpire v2 to AB Sciex TripleTOF 5600 datasets

We first evaluated the performance of DIA-Umpire v2 using the AB Sciex TripleTOF 5600 *E. coli* and Human datasets which were used as part of the original DIA-Umpire manuscript [20]. The derived pseudo MS/MS spectra were searched using X! Tandem, Comet, and MSGF+ search engines and combined by iProphet. Protein and peptide ion identifications were filtered at 1% FDR independently for each run. The number of identifications for each DIA run is shown in Supplementary Table 5. Using DIA-Umpire v2, we were able to identify

similar numbers of peptides and proteins in these data as previously reported using the earlier version (v 1. 25) of the software.

Q Exactive DIA datasets

We then evaluated the performance of DIA-Umpire v2 using the full Q Exactive DIA dataset [10], which included HEK-293 cell lysate and human liver microtissue data (see Methods). In the original publication, the authors used a spectral library-based targeted extraction workflow (Spectronaut). 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 the DIA-Umpire's signal extraction module (DIA_Umpire_SE.jar) to generate pseudo MS/MS spectra (see Methods 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 the same way as DIA pseudo MS/MS spectra. The results (peptide ion and protein identifications) were filtered at 1% FDR independently for each run (see Methods, Supplementary Table 1 for HEK-293 cells, and Supplementary Table 2 for liver microtissue data). On average, the number of peptide ions identified per run at 1% FDR was slightly higher in DIA compared to DDA data (Supplementary Tables 1 and 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 (Supplementary Tables 1-2, "Protein IDs (1% FDR Run level)" column).

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% dataset level FDR for each dataset. Then targeted re-extraction was performed to reduce the number of missing identifications across the runs from the same dataset (see Methods). Figure 2 shows that, after targeted re-extraction and with the data filtered at 1% dataset level FDR, DIA outperformed DDA with respect to the number of peptide ions (Figure 2A) and proteins (Figure 2C) identified on average per run in both HEK-293 and liver microtissue datasets (individual run numbers are shown in Supplementary Tables 1 and 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 using the dataset level FDR strategy as well (see Methods).

Importantly, DIA 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% dataset level FDR 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 2B and 2D). These results were

consistent with the original findings by Bruderer *et al* [7] 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.

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, especially in the very low FDR range (below 1%). This is evident from Figure 3B, which plots the ROC curves for the total number of peptide ion and protein identifications for each dataset. DIA 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 data. This shows that, using the spectral library-free workflow of DIA-Umpire, the main advantage of DIA versus DDA data remains a better identification coverage (and thus quantification completeness) across the dataset, whereas DDA still provides a slight advantage in the total depth of the analysis.

The original study in which these data were analyzed using targeted, spectral library-based software Spectronaut [7] reported fewer missing values than the results of DIA-Umpire. The details regarding FDR estimation in Spectronaut were unavailable in the original manuscript, and thus it is possible that the very high level of quantification completeness achieved using Spectronaut was in part due to forced quantification of background (noise) signals (instead of reporting them as missing values). Nevertheless, DIA-Umpire does have a limitation and dependence on the detection of precursor ion signals. Peptides with insufficient quality of

MS1 precursor ion signals to be detected using untargeted feature detection may have sufficiently strong fragment signals in DIA MS2 spectra, and thus can still be identified using targeted extraction approaches based on fragment ion profiles alone. Although DIA-Umpire attempts to reduce the number of missing quantifications using targeted re-extraction, it queries internal library spectra against the pre-assembled precursor-fragment groups, not against the raw data. Thus, the targeted re-extraction step of DIA-Umpire is still limited by the completeness of the precursor-fragment signals assembled from the detected MS1 and MS2 features at the first stage of the analysis. Thus, we also support alternative workflows by making the untargeted identification results of DIA-Umpire compatible with targeted extraction and quantification tools. To achieve as few missing quantification values across the dataset as possible, a spectral library can be built from DIA-Umpire derived identifications and used then by other targeted extraction tools (e.g. Skyline, OpenSWATH, and Spectronaut).

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, wiSIM, and hybrid DDA/DIA workflows such as pSMART [30]. Here, we conducted five SWATH-like DIA experiments with different isolation windows of fixed widths (5, 10, 15, 20, and 25 Da). Because the DIA-Umpire's feature detection algorithm was optimized for high mass accuracy

data (in both MS1 and MS/MS spectra), the DIA MS/MS spectra were acquired in the Orbitrap, and the alternative DIA methods in which MS/MS spectra are acquired in the ion trap such as wiSIM DIA were not explored in this work. The DDA experiments in this work were conducted for the purpose of providing a baseline number for comparison with DIA data, and thus a common Top 15 most intense ions DDA approach was used. Three replicate runs were performed for each DDA and DIA experiment (see Methods for experimental details).

We processed the DIA and DDA data using same search parameters and FDR estimation as described above for the Q Exactive data. Figures 4A and 4C show the summary of peptide ion and protein identification numbers, respectively, for the DIA and DDA datasets (detailed numbers are shown in Supplementary Table 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 the DIA datasets (5 and 10 Da isolation width 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 minute liquid chromatography (LC) gradient time and without any fractionation step. Similar to what was observed for the Q Exactive datasets discussed above, DIA allowed better identification coverage across the runs from the same dataset (Figures 4B,D).

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 number 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 the precursor and fragment signals less reliable. This, in turn, makes it more difficult to detect low abundant and short eluting peptide ions (see Supplementary Figure 3), thus lowering the reproducibility of identifications (Figures 4B and 4D). The increase in the cycle time can be avoided by decreasing the scan acquisition time and or by decreasing the number of MS/MS scans acquired in each cycle (i.e. by reducing the overall fragmentation mass range). However, these changes could lead to identification losses. The optimal settings are likely to vary depending on the nature of the biological samples under investigation.

Investigating the total number of identifications per dataset (i.e. triplicate runs from each dataset combined) between DIA and DDA at various FDR levels in more detail, DDA had more peptide ions identified in the very low FDR range (below 0.5% FDR) than DIA with any window size (Figure 4E), even though the DIA numbers (5 and 10 Da windows)

exceeded those of DDA in the FDR range of ~ 1% or higher. It is well known that, due to error rate inflation when going from peptide to protein level [36], achieving a certain low protein level FDR (e.g. 1%) requires peptide identifications with lower FDR value at the peptide level. This explains why the number of protein identifications at 1% protein FDR was higher in DDA data (Figure 4F), 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 extracted using DIA-Umpire.

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 at the targeted re-extraction step (see Methods for details). We illustrate these improvements here by performing a comparison with the results obtained using DIA-Umpire v1.25 [20] on the Orbitrap Fusion and Q Exactive DIA datasets. Figure 5 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 liver microtissue dataset. The results from all the other DIA runs used in this work, including the Orbitrap Fusion data, are shown in Supplementary Figure 4. Figure 5 shows a wider distribution of high scoring (i.e. likely correct) identifications, while the width of the decoy distribution is unaffected. This results in better discrimination between 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 threshold, especially in the Q Exactive data (Supplementary Figure 5).

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 [39]. This is a particularly significant feature for new applications we are currently exploring, e.g. for combining the results of targeted extraction using the internal library built by DIA-Umpire with that using external DDA libraries (built from sample specific DDA data, or using global libraries such as human SWATHAtlas spectral library).

CONCLUSIONS

In this paper, we presented DIA-Umpire v2 and demonstrated that it is capable of highly sensitive, untargeted analysis of DIA data from complex protein samples generated using the Orbitrap family of mass spectrometers. Using publicly available Q Exactive DIA data, and using Orbitrap Fusion data acquired as part of this work, we showed that the DIA can achieve similar identification numbers and better identification reproducibility across the datasets than DDA data. With fewer missing quantification values, DIA data should provide improved statistical power for post-quantification analysis, e.g. using tools such as mapDIA [41] developed specifically for DIA data. Importantly, the workflow of DIA-Umpire does not require a spectral library, which should facilitate the adoption of DIA 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.

The untargeted, spectral library-free approach of DIA-Umpire provides an alternative way to process DIA data. Unlike existing 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. It also alleviates the burden of building comprehensive, sample-specific libraries using DDA data in the first place. Furthermore, because DIA-Umpire-derived identifications are compatible with the targeted extraction tools (e.g. 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.

The Orbitrap Fusion experiments conducted 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 the Orbitrap Fusion DIA data generated in this work, or comparison between different computational strategies, go beyond the scope of this work. Nevertheless, we hope that the data presented here, which we make available via the ProteomeXchange consortium database (dataset identifier PXD003179), can be used for that purpose in the future.

ACKNOWLEGEMENTS

This work was supported in part by the US National Institutes of Health R01GM94231 and Taiwan Ministry of Science and Technology grant (104-2113-M-001-005-MY3 to Y.-J.C). We thank Dmitry Avtonomov for useful discussions and Felipe da Veiga Leprevost for the comments on the manuscript.

REFERENCES

1. 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.

2. 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.

3. 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.

4. 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.

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

6. Rost, H.L., et al., *OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data.* Nature biotechnology, 2014. **32**(3): p. 219-23.

7. 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.

8. MacLean, B., et al., *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.

9. Reiter, L., et al., *mProphet: automated data processing and statistical validation for large-scale SRM experiments.* Nature methods, 2011. **8**(5): p. 430-5.

10. 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.

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

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

13. 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.

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

15. 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.

16. 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.

17. 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.

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

19. Liu, Y., et al., *Quantitative variability of 342 plasma proteins in a human twin population.* Molecular systems biology, 2015. **11**(1): p. 786.

20. 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.

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

22. 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.

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

24. 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.

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

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

27. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Analytical chemistry, 2003. **75**(17): p. 4646-58.

28. Lam, H., et al., *Development and validation of a spectral library searching method for peptide identification from MS/MS.* Proteomics, 2007. **7**(5): p. 655-67.

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

30. 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.

31. 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.

32. 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.

33. 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.

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

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

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

41. 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.

FIGURE LEGENDS

Figure 1. 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.

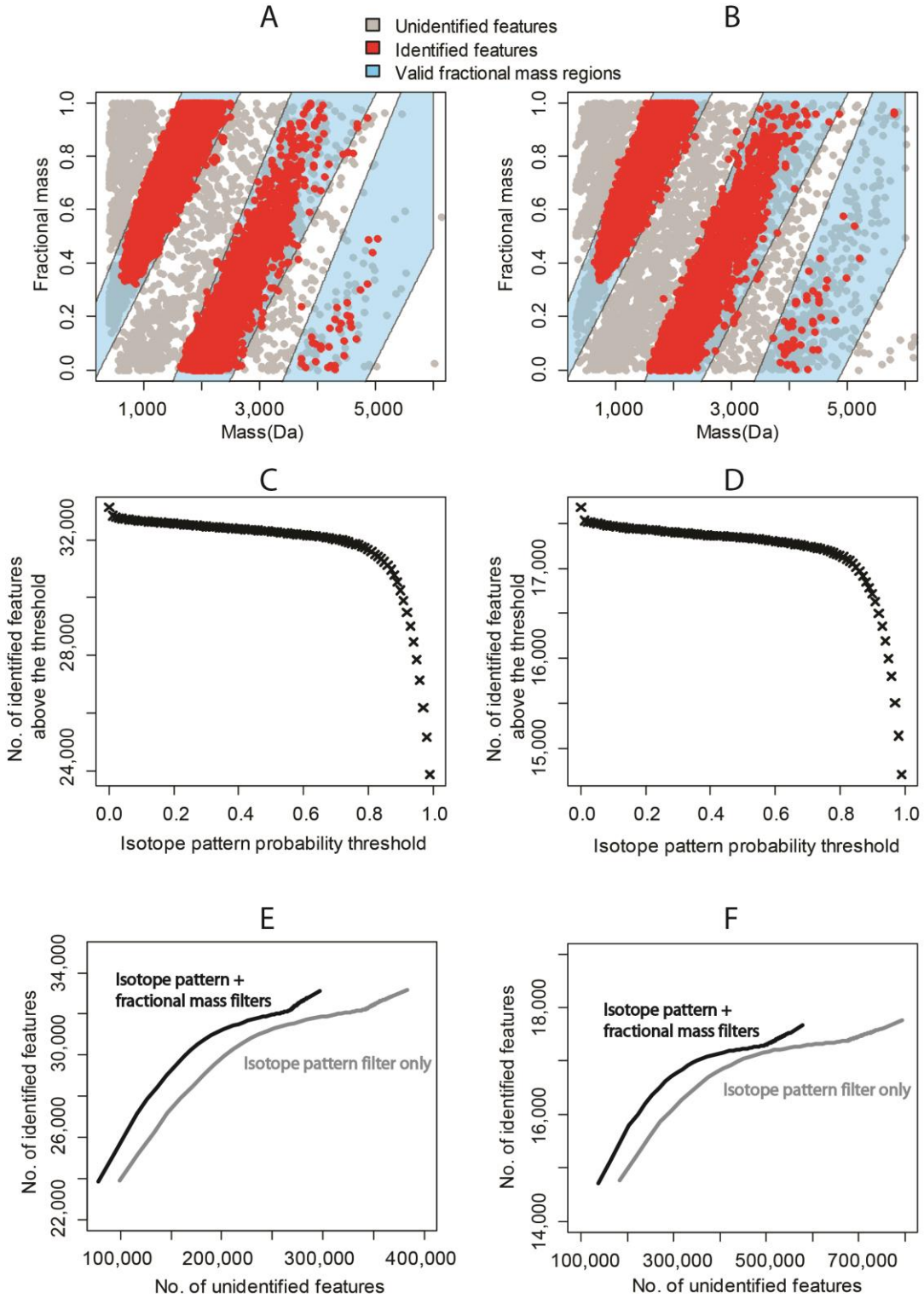


Figure 2. Identification numbers and reproducibility in the Q Exactive DIA and DDA datasets. (A) The number of peptide ion identifications at individual run level in different datasets. **(B)** The coverage of peptide ion identifications (identification reproducibility across the dataset). **(C)** Same as (A), protein level; **(D)** Same as (B), protein level.

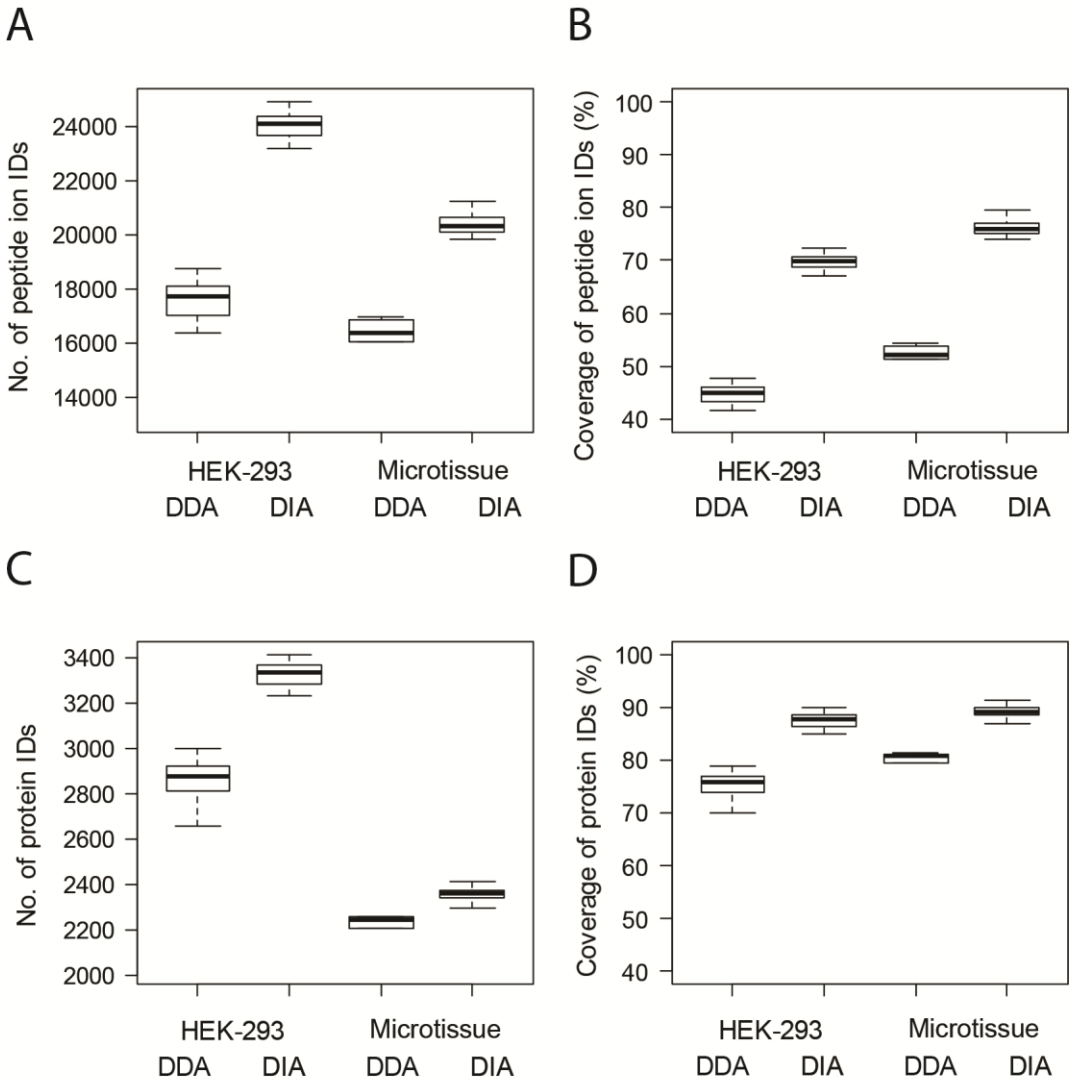


Figure 3. Number of identifications as function of FDR in the Q Exactive datasets. (A) Peptide ion identifications, HEK-293 Q Exactive DIA and DDA data. **(B)** Protein identifications, HEK-293 Q Exactive DIA and DDA data. **(C)** Same as (A), liver microtissue Q Exactive DIA and DDA data. **(D)** Same as (B), liver microtissue Q Exactive DIA and DDA data.

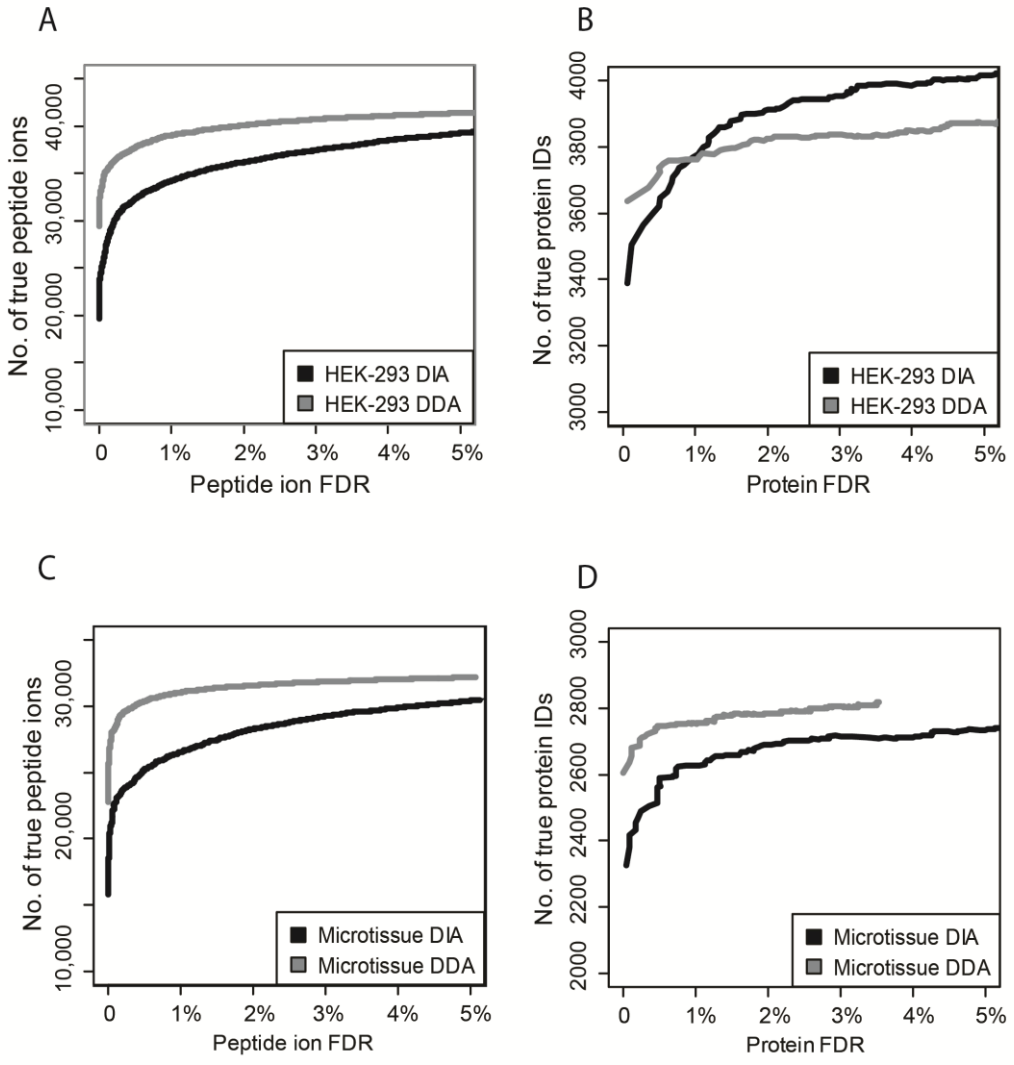


Figure 4. Identification numbers and reproducibility in the Orbitrap Fusion DIA and DDA datasets. (A) The number of peptide ion identifications at individual run level in different datasets. Red dot indicates the actual identification number from a replicate. **(B)** The coverage of peptide ion identifications (identification reproducibility across the dataset). The number was calculated as the number of identifications for each replicate divided by the total number of identification from all the replicates. Red dot indicates the actual value derived from a replicate. **(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). Solid line: DIA dataset. Dash line: DDA dataset. **(F)** Same as (E), at the protein level. Solid line: DIA dataset. Dash lines: DDA dataset.

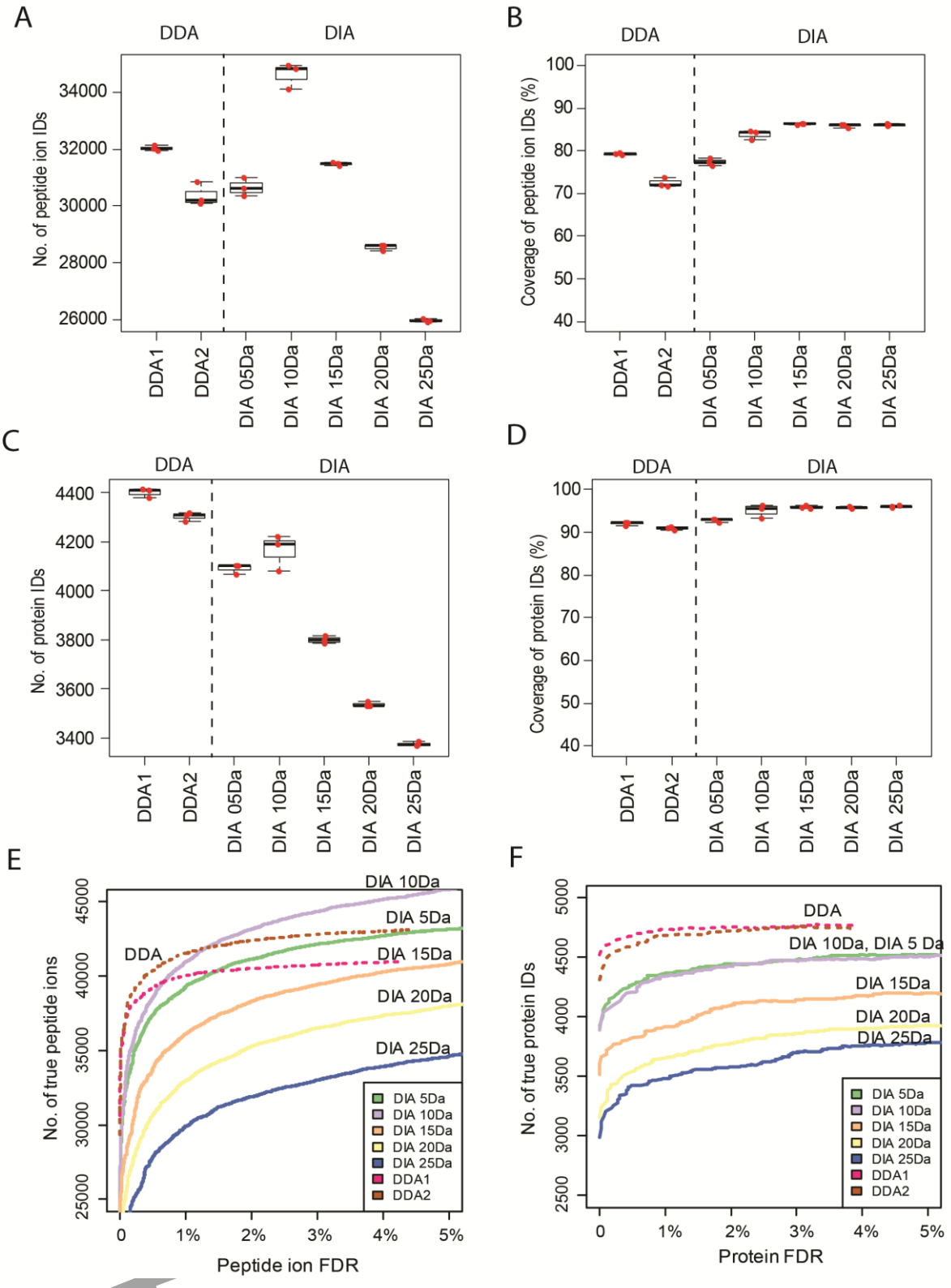
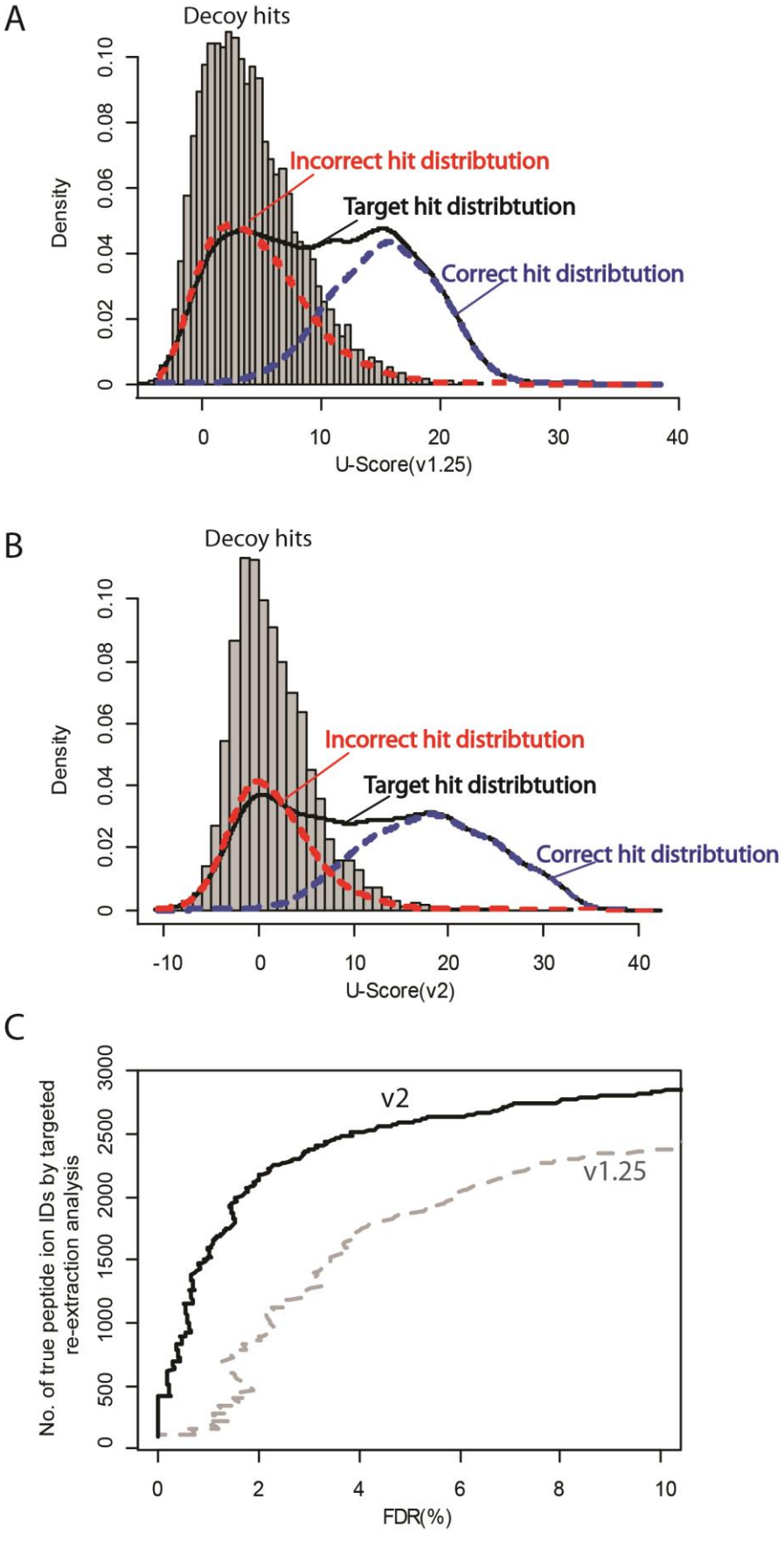


Figure 5. Score histograms and mixture modeling. (A) Score histograms and parametric Gaussian mixture modeling result obtained using DIA-Umpire v 1.25. **(B)** Score histograms

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 v 1.25 and v2. Data for one representative run from the Orbitrap HEK-293 Q Exactive dataset.


A
Decoy hits
Incorrect hit distribtution
Target hit distribtution
Correct hit distribtution
Density
0
0.02
0.04
0.06
0.08
0.10
0
10
20
30
40
U-Score(v1.25)
B
Decoy hits
Incorrect hit distribtution
Target hit distribtution
Correct hit distribtution
Density
0
0.02
0.04
0.06
0.08
0.10
-10
0
10
20
30
40
U-Score(v2)
C
v2
v1.25
No. of true peptide ion IDs by targeted re-extraction analysis
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
FDR(%)

ved.