

Improving Qualitative and Quantitative Performance for MS^E-based Label-free Proteomics

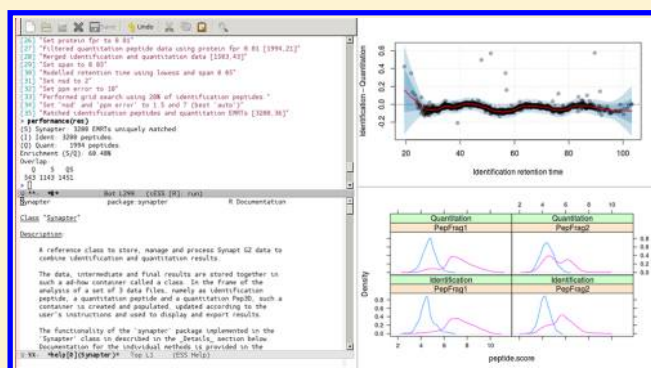
Nicholas J. Bond,^{†,‡} Pavel V. Shliaha,[†] Kathryn S. Lilley,[†] and Laurent Gatto^{*,†}

Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, United Kingdom

Supporting Information

ABSTRACT: Label-free quantitation by data independent methods (for instance MS^E) is growing in popularity due to the high technical reproducibility of mass spectrometry analysis. The recent introduction of Synapt hybrid instruments capable of incorporating ion mobility separation within mass spectrometry analysis now allows acquisition of high definition MS^E data (HDMS^E). HDMS^E enables deeper proteome coverage and more confident peptide identifications when compared to MS^E, while the latter offers a higher dynamic range for quantitation. We have developed synapter as, a versatile tool to better evaluate the results of data independent acquisitions on Waters instruments. We demonstrate that synapter can be used to combine HDMS^E and MS^E data to achieve deeper proteome coverage delivered by HDMS^E and more accurate quantitation for high intensity peptides, delivered by MS^E. For users who prefer to run samples exclusively in one mode, synapter allows other useful functionality like false discovery rate estimation, filtering on peptide match type and mass error, and filling missing values. Our software integrates with existing tools, thus permitting us to easily combine peptide quantitation information into protein quantitation by a range of different approaches.

KEYWORDS: data independent acquisition, data combination, missing values, false discovery rate, quantitation, identification transfer, HDMS^E



INTRODUCTION

The advancement of mass spectrometry over the past decade has greatly facilitated proteomics research. Initially, the goal of mass spectrometry based proteomics studies was to define the 'functional' genome by identifying as many protein constituents of a given sample as possible.¹ The introduction of stable isotope labeling strategies either *in vivo*, or *in vitro* have enabled mass spectrometry based quantitative assessments of protein abundances between samples. The utilization of stable isotope labels allows the differentiation of samples while permitting their coanalysis and affords numerous benefits. One of the most desirable of these benefits is that the proportion of the technical variation arising from sample preparation or LC–MS can be accounted for depending on the method employed.² In recent years, however, label-free protein quantitation has become popular alternative.³ Evans et al. estimated that label-free approaches were the most popular methods of protein quantitation in 2011.⁴ This can be attributed principally to the development of robust LC–MS platforms but may also be in response to the reported limitations of stable isotope labeling strategies and their associated cost.^{4,5}

In all label-free methods, samples under comparison are analyzed during separate runs, typically employing liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) where information is captured for precursor ions

(MS1) and their collision induced dissociation fragments (MS2). The simplest label-free method involves taking the number of spectra acquired and assigned to peptides from the same protein as a measure of abundance. This method, generally referred to as spectral counting, can be executed in several ways, including normalization of counts by the total number of peptides which can theoretically be generated from the protein.⁶ In an alternative approach, ion current recorded for a peptide ion is utilized as a measure of its abundance. It has been demonstrated that ion amount and signal are linearly correlated within the dynamic range of a mass spectrometer in simple and complex mixtures.^{7–9} Given the stochastic nature of intensity based tandem mass spectrometry, some peptides are not identified in some LC–MS/MS runs, giving rise to missing values.¹⁰ In the case of spectral counts, however, label-free quantitation does not necessarily require the same set of peptides to be identified across all LC–MS experiments. Instead, the number of spectra measured is correlated to the abundance of their protein of origin in each sample independently.³ Thus, if a protein has been identified in different samples by a different set of peptides, quantitative analysis on the protein level is still possible.

Received: August 17, 2012

Published: March 20, 2013



In label-free quantitation involving ion intensity measurements, the most common approach is to perform quantitation on peptides that have been consistently identified in all LC-MS/MS runs,¹¹ since direct comparison of integrated signals between different peptides is not possible given difference in sequence-specific ionization efficiency of peptides. Although these differences may be averaged out over multiple peptides from the same proteins,⁸ missing values can have significant impact on this approach reducing the number of ions common to all experiments which can be taken for quantification.³

Missing values are prevalent in acquisition methods where selection of a precursor ion is dictated by its intensity such as data dependent acquisition (DDA).¹⁰ DDA enables the sequential isolation and fragmentation of peptide ions, providing criteria (set *a priori*) about the precursors have been satisfied. Determining the frequency to switch between MS1 and MS2 modes is typically a trade-off between the optimal peak shape that is required for accurate quantification and generation of fragmentation data for identification purposes.¹² Assuming a peptide ion is exclusively isolated, the resulting spectrum constitutes fragment ions derived only from the selected parent ion. In reality this is not a trivial exercise, but a compromise is reached between completely isolating the ion of choice (selectivity) and transmission of this ion (sensitivity).^{13,14} A number of alternative, data independent acquisition (DIA) approaches have been introduced (parallel acquisition, MS^E,¹⁵ SWATH,¹⁶ PACIFIC,¹⁷ AIF,¹⁸ SWIFT parallel fragmentation¹⁹) that avoid sampling of peptide ions inherent of DDA experiments and their associated challenges.^{10,20}

It is also possible to reduce the number of missing values in DDA or DIA approaches by transferring identifications between LC-MS/MS runs. This involves matching features (ions or peptides) from different acquisitions, in one of which the feature has not been identified and is assigned the sequence from its matching pair in the other acquisition.²¹ Initially it was suggested that features could be matched solely on the basis of recorded *m/z* (accurate mass tag).²² However such matching would require sub ppm resolution²³ and later it was suggested that feature retention time can be used as additional discriminatory information to match features between runs^{24,25} (hence features started to be referred to as accurate mass retention time tags). The retention time of a peptide is more prone to systematic variation than its *m/z* between runs, hence a large number of retention time alignment algorithms have been suggested to address systematic retention time variability.²¹

The MS^E data independent acquisition method (initially suggested as parallel fragmentation approach for IMS-MS system²⁶), commercialized by Waters, is composed of alternate scans of low and high energy.¹⁵ In low energy scans, eluting peptide ions are detected and analyzed intact. In high energy scans, peptide ions are fragmented by employing a collision energy ramp and the resulting fragment ions are in turn detected. Postprocessing software allows the retrospective pairing of fragments with their precursors. In theory, issues surrounding the stochastic sampling of precursor ions are circumvented as fragment ions generated by all eluting peptide ions are detected. Instead, this approach is limited by the extent to which correct assignment of fragment ions to precursors is achieved. Waters have introduced the Synapt suite of hybrid mass spectrometers with ion mobility separation (IMS) capability that enables the drift time of a precursor to be

recorded within an MS^E experiment (referred to as high definition MS^E experiment, HDMS^E).²⁷ Since fragmentation occurs after IMS, fragments are expected to have the same mobility profile as their precursors.^{28,29} This additional discrimination of ions afforded by HDMS^E experiments enables more accurate assignment of fragments to their precursors and therefore results in higher number of confident peptide identifications (see ref 30).

ProteinLynx Global Server (PLGS) is a software developed by Waters, which is used to process MS^E and HDMS^E data.¹⁵ PLGS employs two algorithms to process the raw data. The first one, Apex3D, is used to subtract noise and integrate ion current signals across their chromatographic elution. Its output is a list of ions with intensities above a user specified threshold. The second one, Pep3D, collapses ions determined to be isotopes and charge states of common peptides into EMRTs. An EMRT (exact mass and retention time) is a peptide of unknown sequence, which is characterized by mass, retention time, intensity and mobility (in case of HDMS^E). The fragment ions are then tentatively associated with precursors based on the similarity of their elution and mobility profiles. In most cases, the same fragment is associated with multiple precursors at this stage. Prior to database searching which is performed by an Ion Accounting algorithm, a decoy database is concatenated to the forward database. The database search is protein centric and is performed in three stages, referred to as passes. During the first pass, PLGS iteratively cycles through the data, removing EMRTs determined to be peptides derived from the most confident protein identification. Fragments tentatively associated with the identified EMRT and corresponding to b and y fragment ions of the sequence assigned to it are also removed. The process terminates when the rate of decoy protein identifications exceeds a user specified FDR threshold. The FDR reported is therefore calculated at the protein level, even though peptides are given a score indicative of the strength of the spectrum:sequence match. Pass two of the database search continues to deplete data, but this time peptides are only identified for proteins which have been identified in the first pass. This time peptides can be subjected to missed cleavages, variable modifications and in-source fragmentation. During the third and final pass of the search, a fragment is allowed to have higher intensity than its precursor, a situation characteristic of in-source fragmentation of highly labile peptides. PLGS provides exhaustive output from all three algorithms (Apex3D, Pep3D and Ion Accounting) as comma-separated variable (csv) or XML format files.

Previously (see ref 30), we have demonstrated that, despite the clear advantages of using IMS for identification, there is an associated truncation of the dynamic range and a reduction in sensitivity on a Synapt G2 platform, an effect that appears to change on a peptide-by-peptide basis. We postulate, therefore, that the combination of data obtained with and without IMS will allow deeper coverage of the proteome than by MS^E alone and more accurate quantitation for high intensity peptides delivered by HDMS^E alone. The approach, although conceptually simple, requires specialized computational tools to accurately and efficiently combine experimental outputs.

Here, we present powerful and flexible software, synapter, for the thorough examination of data generated from DIA experiments performed on Waters mass spectrometers. Its central functionality is to transfer identifications between LC-MS/MS runs by transferring high confidence identifications between independent acquisitions, for example, from HDMS^E

Table 1. Dilution Series for UPS1 Experiment in *E. coli* Background

UPS1 amount per 10 μL injection (fmol)	UPS1 amount in 63 μL (fmol)	volume of 10 fmol UPS1 stock (μL)	volume of 50 fmol UPS1 stock (μL)	volume of 0.45 $\mu\text{g}/\mu\text{L}$ <i>E. coli</i> stock (μL)	volume of 100 fmol/ μL enolase (μL)	volume of 3% ACN, 0.1% FA (μL)
10	63	6.3	0	21	6.3	29.4
25	157.5	15.75	0	21	6.3	20
50	315	31.5	0	21	6.3	4.2
100	630	0	12.6	21	6.3	23.1
150	945	0	18.9	21	6.3	16.8
200	1260	0	25.2	21	6.3	10.5

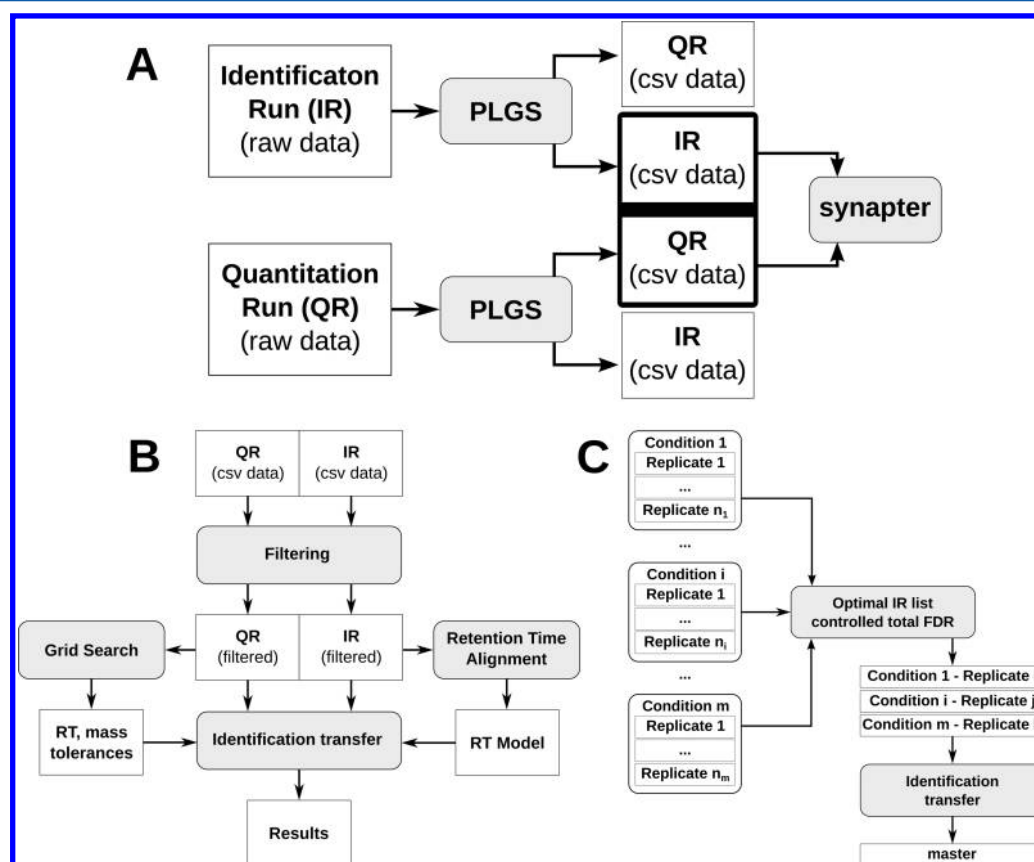


Figure 1. (A) Overview of synapter. (B) Synergize algorithm for identification transfer. Data is represented by white rectangles and computational procedures by gray boxes with rounded corners. (C) IR selection and identification transfer to create a master IR file.

to equivalent MS^E acquisitions. Synapter facilitates the reduction of missing values across an experimental set and by doing so maximizes the number of usable quantitative measurements thereby increasing both proteome coverage and quantitation accuracy. Unlike other packages,^{9,31,32} which provide similar functionality, synapter is capable of determining optimal thresholds for identification transfer and minimizes the effects of FDR increase that arise when transferring identifications across multiple runs. Furthermore, synapter also allows versatile filtering, including removal of non-proteotypic peptides and the ability to filter upon identification statistics. Finally, synapter is written as an add-on package for the statistical programming language R,³³ allowing direct analysis of its output through numerous statistical tools developed by the R community, including MSnbase,³⁴ a package specifically developed for the analysis of LC-MS data.

MATERIALS AND METHODS

Sample Preparation

Two standard samples were prepared and used for all analyses. Preparation of a six protein standard spike series in an *E. coli* background is described in the accompanying article (see ref 30). Three commercially available standards were used for preparation of a UPS1 spike series in *E. coli* background: UPS1 (Sigma), enolase (Waters) and *E. coli* (Waters) digest standard.

Two vials of *E. coli* digest standard from the same lot (specified digest amount of 100 μg) were each resuspended in 200 μL , 3% acetonitrile (ACN) 0.1% formic acid (FA), sonicated for 10 min in a water bath and pooled. The concentration of digest was approximated using a nanodrop-1000 (Thermo Scientific) according to its absorbance at 280 nm to be 0.45 mg/mL and 0.43 mg/mL for the two respectively, which is close to theoretical (0.5 mg/mL).

Enolase (specified digest amount of 1 nmol) was resuspended in 500 μL of 3% acetonitrile (ACN) 0.1% formic acid (FA) as described for *E. coli* digest standard. Enolase stock

concentration was estimated to be 1 pmol/ μ L (half the expected) by amino acid analysis performed in duplicate (PNAC, AAA Service, Dept. Biochemistry, University of Cambridge).

Proteins were digested to peptides as described in ref 30. The three components were mixed just prior to LC–MS analysis to generate 6 different peptide mixtures as described in Table 1; the concentration of *E. coli* and enolase digests was invariantly 150 ng/ μ L and 10 fmol/ μ L, respectively, and UPS1 either 1, 2.5, 5, 10, 15, or 20 fmol/ μ L. Sixty-three microliters of every mixture were prepared to allow six 10 μ L injections of every mixture (three for both modes)

LC–MS Configuration

See ref 30.

Data Processing

ProteinLynx Global Server (PLGS) version 2.5.2 was used to process .raw files and to perform the database search. Data was lock-mass corrected post acquisition. Identical background subtraction parameters were used for acquisitions made in both modes for consistency. Thresholds for low and high energy scan ions and peptide intensity combined across charge states and isotopes were fixed at 150, 10, and 750 counts, respectively. As shown in the accompanying manuscript,³⁰ HDMS^E is around 30% less sensitive than MS^E on a Synapt G2 platform. It is also expected that HDMS^E has lower levels of noise than MS^E. Thus different combinations of processing parameters for data processing and combination of acquisition made in MS^E and HDMS^E may need optimizing for different samples, although the thresholds described above were found to give consistently good performance on a variety of different sample types to date (data not shown). Database searches were performed at 100% protein FDR to maximize the number of reported decoy peptides to compute distribution of scores for decoy peptides. A protein was allowed to be identified by a single peptide. At least 1 fragment was required per peptide and three fragment ions required for protein level identification in accordance with the recommended parameters. The resulting pep3DAMRT.csv and final_peptide.csv files generated by PLGS provide a list of all EMRTs and peptide identifications, respectively.

■ RESULTS

Synapter is a multifunctional software package for post-PLGS DIA data analysis. We first describe the underlying algorithm in detail and then present results of its application. Two data sets of standard proteins, spiked into an *E. coli* background, are used to demonstrate ability of synapter to combine data from independent MS^E or HDMS^E acquisitions and its integration into a pipeline for differential protein expression analysis.

Algorithm Development and Applications

Algorithm Overview. The primary function of the synapter software is to combine quantitation and identification information from two separate data independent acquisitions, as demonstrated in Figure 1A. Throughout the article, the source of identification is referred to “Identification Run” (IR) and the source of quantitation data is labeled “Quantitation Run” (QR). The procedure can also be perceived as transferring identifications from IR to QR; hence, we refer to it as “identification transfer”.

The process of identification transfer is implemented by a function called *synergise* in the synapter package. The *synergise*

algorithm is summarized in Figure 1B. The default, albeit customizable pipeline is as follows. First, the PLGS output from IR and QR is loaded into the R environment and filtered on match type. Only peptides that are not subject to missed cleavages, variable modifications and in-source fragmentation from pass 1 (PepFrag1) and 2 (PepFrag2) are retained. Then data is filtered on peptide FDR (computed by synapter), protein FDR (computed by PLGS) and only proteotypic peptides are retained. Then, a group of high confidence peptides that have been identified in both IR and QR are selected to model retention time differences between the two acquisitions. Optimal retention time and precursor mass tolerances for best identification transfer are estimated by a grid search and used during identification transfer. The *synergise* algorithm thus results in a composite data set composed of quantitation information from QR and identification information from IR.

When dealing with large data sets, the procedure of transferring identifications from each run to all others may be time-consuming and suboptimal. In addition, transferring identifications from multiple acquisitions in series will accumulate false identifications and result in an increased and unknown final FDR, as opposed to the FDR controlled for individual runs. synapter controls for this effect by choosing an optimal combination of acquisitions to transfer identifications from. These acquisitions are used to generate a new, composite *in silico* IR³⁵ (Figure 1C). This composite or “master” IR is then used as the sole identification source when transferring identification to individual QRs.

The Synergise Algorithm. Data Filtering. Synapter requires three types of files produced by PLGS: two final_peptide.csv files (one for IR and one for QR) and a pep3DEMRT.csv file from the QR. The final_peptide files contain peptide identifications as determined from the PLGS database search, the pep3DEMRT file contains all recorded EMRTs from QR regardless of whether these EMRTs were identified by PLGS. Additionally, the fasta protein sequence database, as used in PLGS, is required in order to filter upon proteotypic peptides. After loading the data, synapter filters peptides based upon user specified preferences. Synapter allows a series of flexible filtering strategies which retain only confidently identified peptides prior to proceeding with the retention time alignment, identification transfer parameter estimation and the identification transfer.

Peptides identified from both regular and decoy databases that are subject to variable modifications, missed cleavages or in-source fragmentation are removed, since they are perceived as being unsuitable for label-free quantitation. Using the peptide scores determined during the PLGS database search, identification statistics are computed for fully tryptic, unmodified peptides identified in pass one (PepFrag1) and pass two (PepFrag2) database searches (peptides identified by pass three are in-source fragments and thus not used). Individual peptide *p*-values are computed as described in Käll et al.³⁶ Each *p*-value is empirically estimated by computing the percentage of decoy peptide scores that receive a score equal or higher than the target peptide. The respective regular and decoy (random) score distributions can be plotted (Figure 2) for quality assessment. The *p*-values are then adjusted by one of the following procedures: Bonferroni single-step adjustment for strong control of the family wise error rate, the Benjamini and Hochberg (BH, default) step-up FDR-controlling procedure³⁷ and Storey and Tibshirani’s *q*-value metric.^{36,38} The identi-

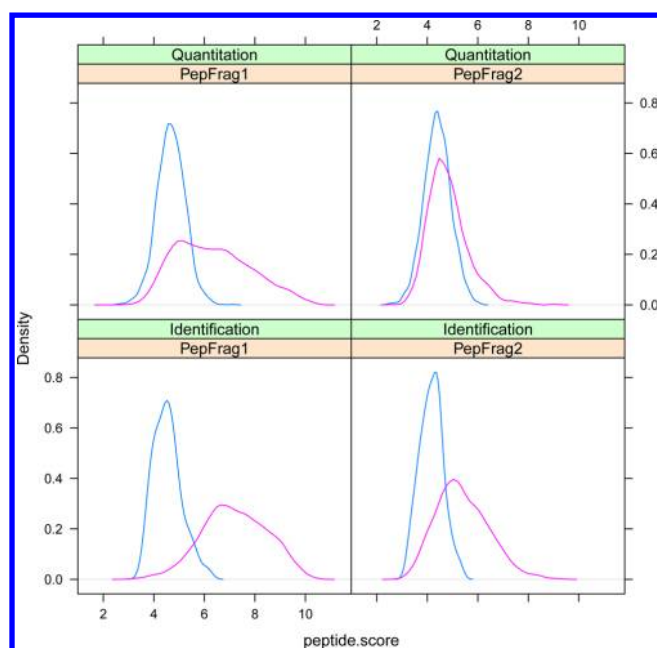


Figure 2. Distribution of PLGS scores for pass one (PepFrag1) and pass two (PepFrag2) unmodified fully tryptic peptides. Blue, decoy peptides; red, regular peptides. The scores for HDMS^E peptides from the regular database are on average higher than for MS^E peptides.

fication statistics are calculated for pass one and pass two peptides independently allowing user to perform filtering based on these statistics (Figure 3).

In addition to filtering upon the peptide identification reliability, it is recommended to filter the input on peptide mass tolerance (in ppm), and protein level false discovery rate (as computed during PLGS database search). Synapter also allows filtering on peptide uniqueness, removing any peptides that are not tryptic or proteotypic. When performing *in silico* digestion synapter uses generally accepted rules for trypsin specificity: cleavage at C-terminus of K and R, but not if these amino acids are followed by P.³⁹ It has been reported that miscleavages also occur where K, R, D or E follow a tryptic site and such peptides are used by PLGS for quantitation.⁴⁰ Thus synapter removes these peptides where filtering for database uniqueness is applied as the validity of using these peptides for label-free quantitation is questionable, and the reproducibility of these miscleavages has yet to be determined.

Retention Time Alignment. LC–MS is prone to variation in peptide retention time. Deviation in peptide retention time between runs has been characterized previously and shown to adopt either a common behavior whereby modest fluctuations in temperature cause the retention time of coeluting peptides to drift collectively, or to a lesser extent, peptide-specific deviations where by retention times for individual peptides change stochastically between runs. synapter uses locally weighted scatterplot smoothing⁴¹ (LOESS) to model collective retention time deviation behavior (Figure 4). LOESS application for this purpose has been described previously.^{9,42}

Grid Search. The identification transfer process involves finding EMRTs in QR which correspond to identified peptides in IR based on the expected similarity of their retention time and mass. In practice, retention time and the measured mass of a peptide do not replicate precisely between LC–MS runs. Thus, in order to transfer identifications, synapter needs to estimate the variability in retention time and mass measure-

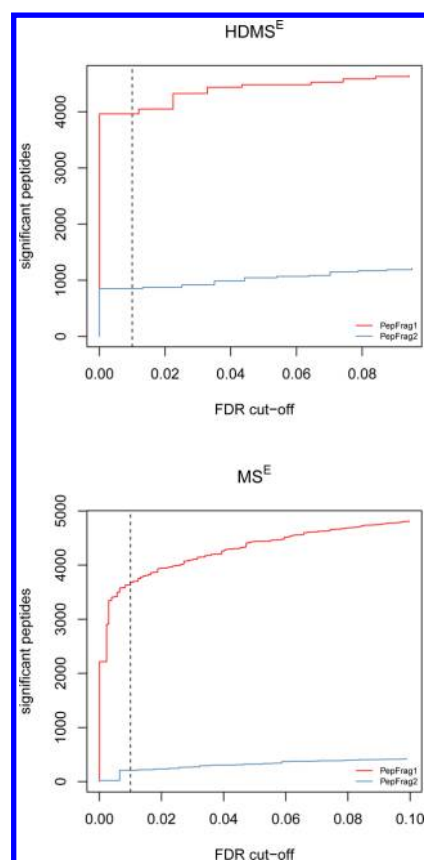


Figure 3. Cumulative plot of FDR for HDMS^E and MS^E data. The plot shows the number of identified peptides (y axis) should a user choose to accept a given FDR (x axis). The dashed line is the suggested FDR default of 0.01. Red and blue are fully tryptic unmodified peptides from database search pass one and two, respectively.

ments between QR and IR. This variability is expressed as mass and retention time tolerances during identification transfer. If two EMRTs in QR and IR differ in mass and retention time no more than the tolerances, they are assumed to represent the same peptide.

Synapter uses the number of standard deviations (nsd) to express the retention time tolerances during identification transfer. The mass tolerance is defined as the absolute deviation between the mass of the QR EMRT and the theoretical mass of the IR EMRT given its assigned peptide sequence. Thus the mass tolerance during identification transfer is based solely on the distribution of mass measurement errors in the QR.

The distributions of deviations from the retention time model and mass measurement errors can be visualized to assess their overall variability between runs. Although it is possible to manually select a specific tolerance, to achieve optimal identification transfer it is recommended to allow synapter to perform a grid search, during which a set of mass and retention time tolerance combinations are tested (Figure 5) on the full data set or a subset of data. For each pair of parameters, the percentage of total unique EMRT assignments is calculated and the combination that maximized this value is used for identification transfer.

Using a set of confidently identified peptides during the grid search allows synapter to compute the false positive and negative rate of identification transfer for all tolerance combinations tested. This is done by comparing which sequence a QR EMRT has been assigned during the database

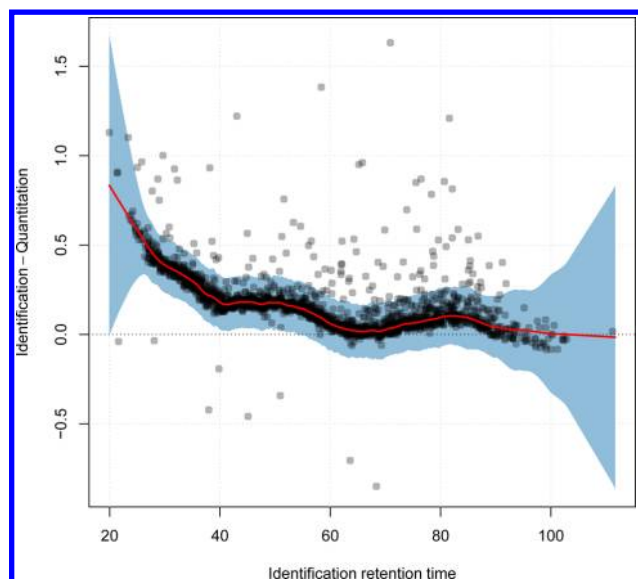


Figure 4. Retention time deviation and its model by synapter between IR (10 fmol of UPS1 in 1.5 μ g of *E. coli* background analyzed in HDMS^E mode, replicate 1) and QR (10 fmol of UPS1 in *E. coli* background analyzed in MS^E mode, replicate 1). Each point represents a single peptide identified in both QR and IR. Deviation in retention time between runs is plotted against peptide retention time in IR run. The red curve represents the LOESS fitted retention time model with span parameter of 0.05. Dark blue and light blue are areas of one and two standard deviations around the fitted model.

search and which sequence has been assigned to it during identification transfer. The best case scenario (a true positive) occurs when there is a single QR EMRT within the retention time and mass tolerance of the IR peptide and the QR EMRT has been assigned the same sequence during the PLGS database. In practice, however, there are several reasons why this may not happen, as illustrated in Figure 6. First, unique assignments can be either true positives (Figure 6A) or false positives (Figure 6C). Alternatively, no assignment might be possible within the defined tolerances (Figure 6B). False positives are possible if the QR EMRT of the peptide, whose identification is being transferred, has a higher deviation in retention time and/or mass measurement than the tolerances and a different QR EMRT falls within the defined tolerance window. A false positive assignment, can also be explained when a peptide was originally mis-identified in IR or QR. This situation seems to be prevalent among false positive assignments, due to the high proportion of PepFrag2 instances among the unique false assignment, especially for MS^E runs (Figure 2), which are known to be less reliable than PepFrag1. Finally, multiple assignments can include the true EMRT (Figure 6D) or not (Figure 6E). These situations highlight the importance of retention time and mass tolerance settings for optimal assignment. The distribution of assignment outcomes (Figure 6) varies depending on the combination of grid search tolerances. Small tolerances will result in less multiple QR EMRT assignments, but also in an increase of zero assignments. Using the set of common high confidence IR and QR peptides defined earlier, synapter systematically counts the number of unique, multiple, correct and incorrect assignments, to provide detailed figures of the identification transfer accuracy. Another scenario must be considered, in which more than one peptide sequence could be assigned to one QR EMRT. Such a situation is possible when two EMRTs are resolved and identified in the

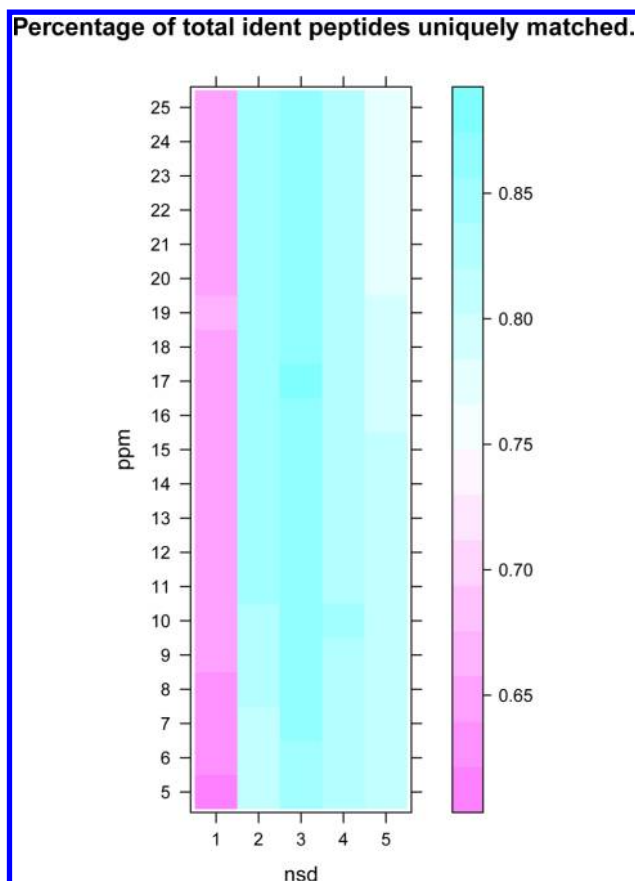


Figure 5. Graphical representation of a grid search result. Each combination of retention time (nsd) and mass tolerance (ppm) is represented by a cell in the table. Here, IR is 10 fmol of UPS1 in 1.5 μ g of *E. coli* background analyzed in HDMS^E mode, replicate 1 and QR is 10 fmol of UPS1 in *E. coli* background analyzed in MS^E mode, replicate 1. The color of each cell represents the proportion of successfully transferred IR identifications at particular tolerances. To the right is the conversion scale between color and proportion of successful identification transfers. Dark blue squares correspond to the best performing tolerance parameter pairs (in this case 12 ppm and 3 nsd).

IR but are not resolved (and hence recorded as a single EMRT) in the QR (Figure 6F). If the two IR identifications were independently transferred to the QR EMRT, the latter would be assigned two sequences. This can only occur when the IR has higher peak capacity than QR, for example when the IR and QR are acquired in HDMS^E and MS^E mode respectively. We estimated that, in our experiments, 0.6% of all MS^E EMRTs identified by identification transfer from HDMS^E were assigned two sequences. It is worth reinforcing that even when a recorded MS^E EMRT has been assigned a single sequence by identification transfer, this does not necessarily imply that it could not be composite of several EMRTs. Indeed, it has been estimated that at 25000 fwhm mass spectrometry resolution, around 35 and 20% of the EMRTs are composite in MS^E and HDMS^E, respectively.⁴³

Identification Transfer. Identification transfer is the process of assigning QR EMRTs with a sequence by pairing it with an IR peptide in order to retrieve quantitative measurement from QR. This process, automated by synapter, is controlled by retention time and mass error tolerances that were optimized in the grid search. Quantitative data is retrieved when a single QR

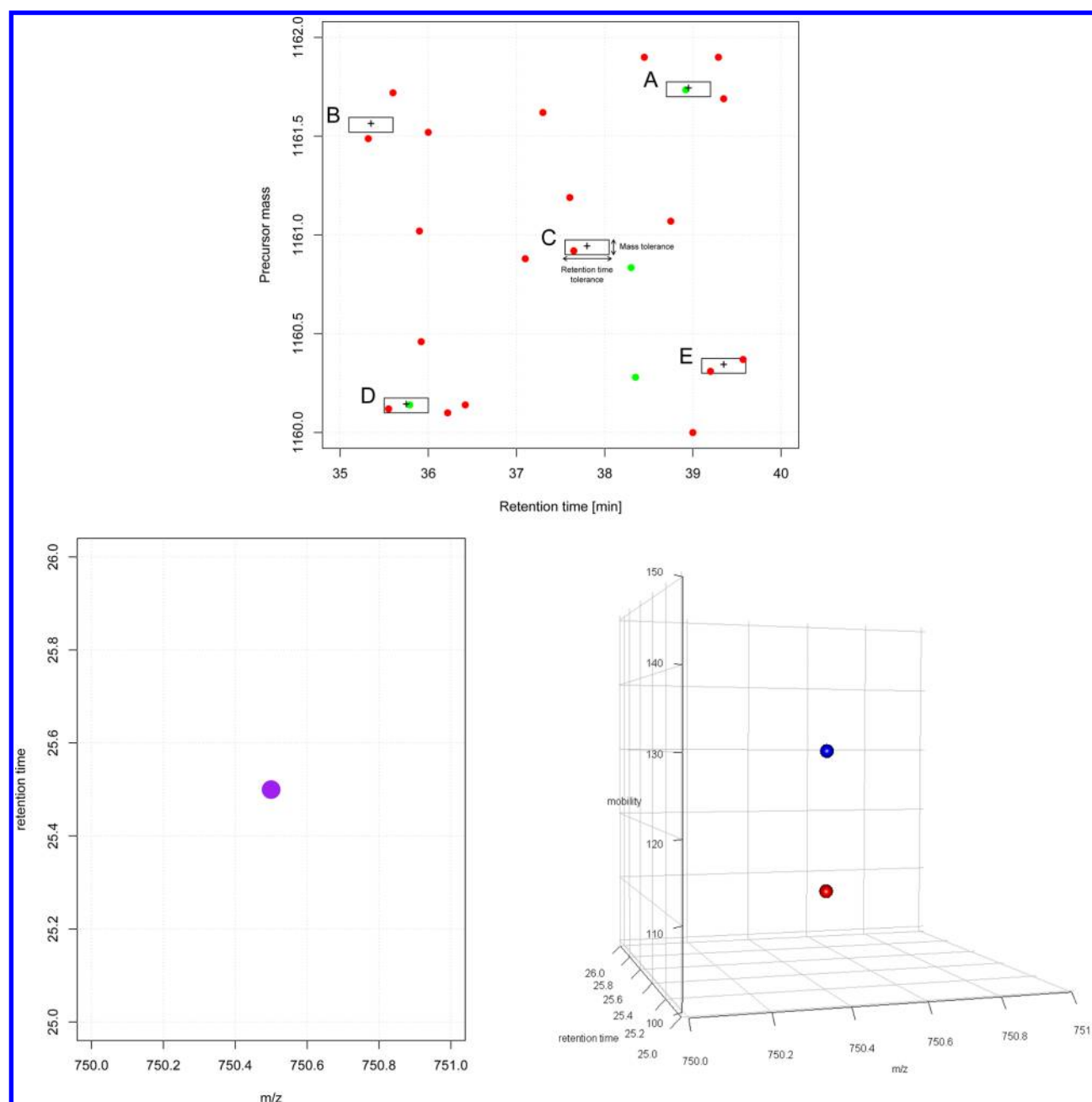


Figure 6. Potential outcomes of identification transfer. Each dot is a QR EMRT with a corresponding mass (y -axis) and retention time (x -axis). Green dots represent the correct QR EMRT identified during the database search for the given IR peptide under consideration, while red dots are other EMRTs. The rectangle height and width are mass and retention time tolerances respectively. Rectangle centers, represented by crosses, correspond to the theoretical mass and corrected retention time of the IR peptide. (A) Unique correct transfer (true positive). (B) No transfer. (C) Unique incorrect assignment (false positive). (D) Multiple assignment with the true result. (E) Multiple incorrect assignments. (F) Applicable only if IR has higher peak capacity and/or dimensionality than QR (e.g., MS^E and $HDMS^E$). A number of EMRTs can co-occupy the same m/z and retention time space, and hence will be recorded as a single EMRTs in MS^E (purple). They may however be resolved by IMS and hence be recorded as separate EMRTs (blue and red) in $HDMS^E$. If these EMRTs will then get identified in $HDMS^E$ and their identification transferred to MS^E , the same MS^E EMRT can be assigned several sequences.

EMRT falls within the defined mass and retention time tolerance limits (Figure 6B), in other cases (Figure 6A, C–E) NA is assigned.

Master Files. Choosing Runs to Transfer Identification From. Missing values is one of the primary limitations of label-free proteomics. Currently, a number of algorithms are available that simply transfer identifications between LC–MS runs under comparison. Since incorrect identifications are less likely to replicate between runs, the number of correct identifications

will accumulate to a lesser degree than the total number of incorrect identifications upon transfer between LC–MS runs. Thus the FDR for the whole analysis is always expected to be higher than the FDR for separate runs if identifications are transferred between runs. Assuming that incorrect identifications never replicate the total FDR for the analysis can be calculated using eq 1:

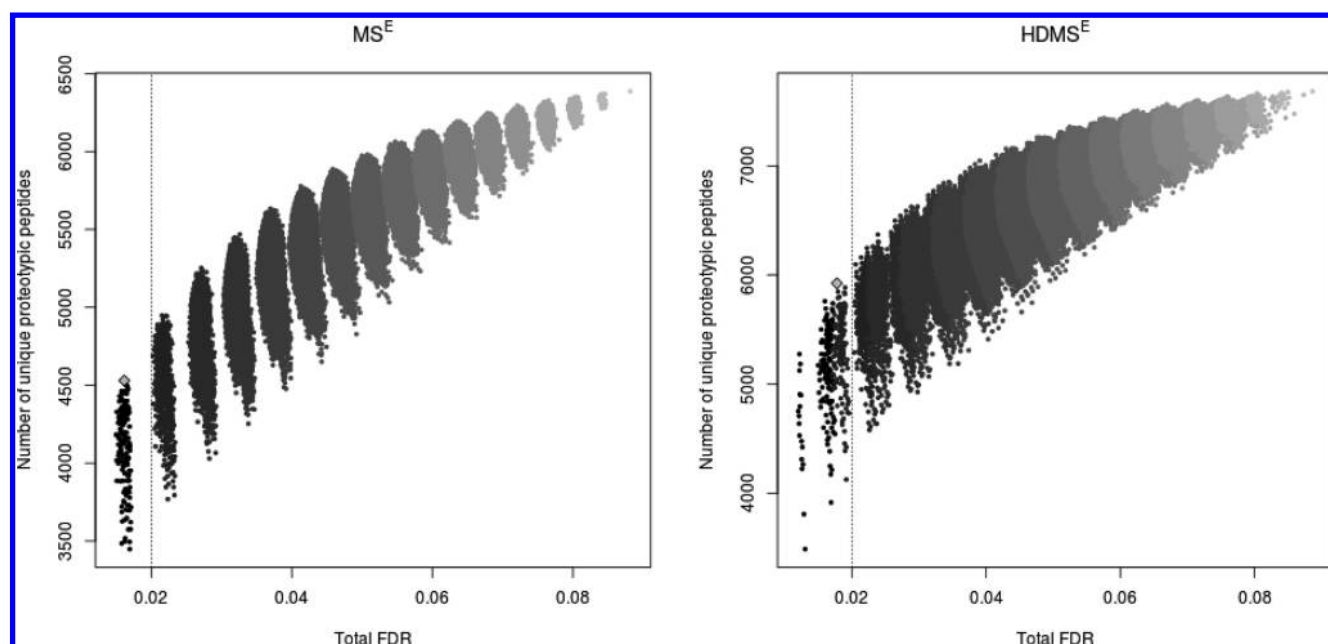


Figure 7. Visualization of synapter's output for the selection of runs to combine into a MS^E (A) and $HDMS^E$ (B) master for UPS1 spiked in *E. coli* background data set. Since there were 18 runs in MS^E and 18 runs in $HDMS^E$, 262125 combinations are possible for both modes. Each point represents a separate combination, that is characterized by the total FDR (x axis), the number of unique peptides (y axis) and the number of runs in the combination (from 2 runs on the left of the graph to 18 runs in the upper right corner). Light diamonds represent the best performing combination with FDR under 2%.

$$FDR_{\text{total}} = \frac{\sum_{i=1}^n IT_i \times FDR}{INU + \sum_{i=1}^n IU} \quad (1)$$

Where FDR_{total} is the false discovery rate of the analysis after identification transfer, FDR is the peptide level false discovery rate in individual runs, IT_i is the total number of identifications in run i , IU_i is the number of identifications unique to run i and INU is the number of (nonunique) identifications seen in more than one run. The numerator of the equation estimates the total number of incorrect identifications across all LC–MS acquisitions identifications are transferred from and the denominator represents the total number of peptide identifications in those acquisitions. This experiment-wide FDR calculation is based upon the premise that incorrect identifications never replicate and thus provides a stringent assessment of FDR, and as such reflects the maximal FDR within an experiment. As additional acquisitions are used as a source of identification there is a diminishing return of IU , while the numerator of eq 1 rises proportionally and hence increases the FDR_{total} . There is thus a compromise between increasing the number of identifications and increasing FDR.

Synapter enables a compromise by computing the total number of identifications and the new total FDR for every possible combination of runs using eq 1 (Figure 7). A user can then specify the total FDR for the analysis they are willing to accept, and synapter will return the optimal combination of runs that has the highest number of identifications at or below the specified total FDR.

Master File Construction. Synapter enables the construction of a master IR and subsequent identification transfer between the master IR and each QR. The master IR can be a composite of IRs that represents the maximal experiment-wide identification rate for a given FDR as described above, or be constructed from any relevant set of IRs.

The master IR is constructed by iteratively incorporating identifications from successive runs using a variation on the identification transfer protocol discussed above. The runs chosen for identification transfer are first ranked in order of the number of identifications. The identifications from the run with the highest number of peptides are used to initiate the master IR. Construction of the master IR then becomes an iterative process whereby successive runs are selected one at a time in order of their rank and new identifications are added to the master IR. Each time a new run is added, synapter models the new peptides' retention times against those in the first master IR data. Thus when a peptide is added to master its retention time is not taken from the run of its origin, but is modeled using LOESS.

Applications

Despite DIA approaches such as MS^E and $HSMS^E$ that do not employ precursor selection, not all recorded EMRTs are assigned a sequence, which results in different, but complementary sets of peptides being identified between related samples. The approach described above and automated by synapter exploits this complementarity by using an identification transfer protocol that maximizes the number of identifications within an experimental design while controlling for the FDR. This approach benefits both MS^E and $HSMS^E$ and depending upon the goals of the analysis can be used to increase the number of IDs, quantifiable measurements and quantitative accuracy of the analysis.

Increasing the Number of Identifications. Transferring identifications between runs allows for an increase in the number of peptide identifications in each sample. To demonstrate this, we mixed two commercially available complex samples: UPS1, an equimolar standard of 48 human proteins was spiked into an *E. coli* whole cell lysate background at increasing amounts to allow injection of 10, 25, 50, 100, 150, 200 fmols of standard in 1.5 μ g of *E. coli* lysate. For each UPS1

loading, triplicate acquisitions in both HDMS^E and MS^E were made in order to compare differences in their identification rate and the improvement when used in conjunction with synapter. We used this data set to demonstrate the three types of analysis possible as follows: MS^E only (for instruments with no IMS capability), HDMS^E only (when accurate quantitative measurements of higher abundance proteins are not crucial) and a combination of both modes.

Data from acquisitions in both modes were first processed by PLGS and then by synapter using default peptide identification statistics (BH, 0.01), precursor mass tolerance (20 ppm), protein false discovery rates (0.01) and retaining unique proteotypic peptides only. As expected, HDMS^E yielded, on average, more peptide identifications at all UPS1 loadings both for UPS1 proteins and the *E. coli* background (Table 2).

Table 2. Summary of UPS1 Spiked into *E. coli* Background Experiment^a

UPS1 loading (fmol)	total number of peptides identified		UPS1 peptides identified	
	MS ^E	HDMS ^E	MS ^E UPS1	HDMS ^E UPS1
10	3210 (±150)	4288 (±55)	32 (±3)	84 (±8)
25	3246 (±512)	4449 (±225)	141 (±48)	246 (±22)
50	3659 (±143)	4982 (±87)	313 (±5)	377 (±16)
100	3486 (±87)	4266 (±383)	443 (±9)	455 (±15)
150	3198 (±175)	3652 (±516)	470 (±13)	472 (±2)
200	3054 (±156)	4124 (±207)	495 (±5)	512 (±4)

^aColumns 2–3 represent total number of peptides identified, columns 4–5 UPS1 peptides that have passed synapter filtering. All entries are arithmetic means of three replicates injections.

Although the total number of peptides identified appeared to be more or less consistent across all UPS1 loadings, the number of UPS1 peptides gradually increased at the expense of *E. coli* background identifications indicative of ion suppression effects.

Two master files were created from MS^E runs and from HDMS^E runs respectively. Specific acquisitions were chosen by synapter, as described above, to provide the maximum number of identifications with total FDR for analysis less than 0.02 (Figure 7). It is noteworthy to highlight that if identifications from all of HDMS^E or MS^E runs were combined the FDR_{total} would increase to almost 0.1. Synapter selected acquisitions with the following UPS1 loadings to combine into the master: 25 fmol (3rd replicate) and 100 fmol (2nd replicate) for the MS^E master and 10 fmol UPS1 (3rd replicate), 50 fmol UPS1 (2nd replicate) and 200 fmol UPS1 (2nd replicate) for HDMS^E master. As predicted by eq 1, acquisitions selected for construction of the master IRs exhibit the most dissimilar sets of peptide identifications so to maximize overall peptide diversity. MS^E and HDMS^E masters contained 4433 and 5912 peptide identifications respectively.

Identifications were then transferred in turn from MS^E master to MS^E runs, HDMS^E master to HDMS^E runs and HDMS^E master to MS^E runs to simulate MS^E only, HDMS^E only and combined analysis types respectively. Optimal parameters for identification transfer were determined by grid search: mass tolerance from 5 to 25 ppm, retention time tolerance from 1 to 5 nsd, proportion of data 25%. Transferring identifications from MS^E master to MS^E runs increased identification rate by 12% on average and from HDMS^E master to HDMS^E by 5% on average. The most substantial increase of

36% was observed when HDMS^E master was used against MS^E runs (Table 3).

Table 3. Number of EMRTs that Have Been Assigned a Sequence by Identification Transfer^a

UPS1 loading (fmol)	MS ^E data MS ^E master	MS ^E data HDMS ^E master	HDMS ^E data HDMS ^E master
10	3578 (±121)	4341 (±36)	4193 (±181)
25	3425 (±557)	4296 (±656)	4816 (±30)
50	3981 (±21)	4696 (±92)	4441 (±590)
100	3787 (±234)	4421 (±138)	4379 (±15)
150	3739 (±55)	4523 (±66)	4450 (±44)
200	3681 (±214)	4379 (±204)	4373 (±345)

^aThe mode of data acquisition and the master file used are described in the column label. All entries are arithmetic means of three replicate injections.

Reducing the Number of Missing Values. A primary limitation of quantitative label-free proteomics is the accumulation of missing values that hinder downstream analysis. Implicit to using a master IR for identification is the reduction of missing values. MS^E and HDMS^E acquisitions at 25 and 50 fmol UPS1 loading were used to demonstrate the beneficial effect of identification transfer on missing data through a typical analysis for differential protein expression.

The master was created from 50 fmol first and second replicate HDMS^E runs, as suggested by synapter, and had 5612 peptides with a total FDR of under 0.02. The master identifications were then transferred to the MS^E acquisitions of 25 and 50 fmol UPS1 in *E. coli* background (three replicates for both loadings, six samples in total) and the quantitative measurements extracted were further analyzed within R. Figure 8 compares missing data distribution for the 6 samples utilized in the statistical analysis described below. Absence and presence of quantitative data are shown as white and black blocks respectively at the peptide (top) and protein (bottom) level for synapter (left) and MS^E (right) data. For peptide and protein levels respectively, MS^E data exhibits 33 and 22% of missing values, which are reduced to 14 and 2% with using the HDMS^E master IR. More striking than the reduction in percentage of missing values, however, is the positive impact this procedure has on the final complete, analyzable data set. Table 4 illustrates the number of features with 0, 1, ... to 5 missing values. If we consider that a peptide or a protein must have at least 2 out of 3 possible quantitation values for both 25 fmol and 50 fmol UPS1 sample groups to be deemed usable in the subsequent statistical test, the percentage of usable peptides increases from 55 to 87% and usable proteins from 71 to 98%.

Statistical Analysis of Differential Protein Expression.

Using MSnbase, peptide level quantitation for the 25 fmol and 50 fmol UPS1 samples described above was combined in order to perform protein level quantitation by the top 3 method⁸ and protein quantitative data was normalized using the *E. coli* background proteins as a common reference across samples. Forty-seven out of 48 UPS1 proteins were included in the final data set composed of a total of 705 proteins quantified in at least 2 out of 3 replicates in each group. A parametric *t* test assuming unequal variances was applied to find differentially expressed proteins. At a false discovery rate of 10%, 31 spiked in proteins were identified as differentially expressed, with no false positives (*E. coli* background protein).

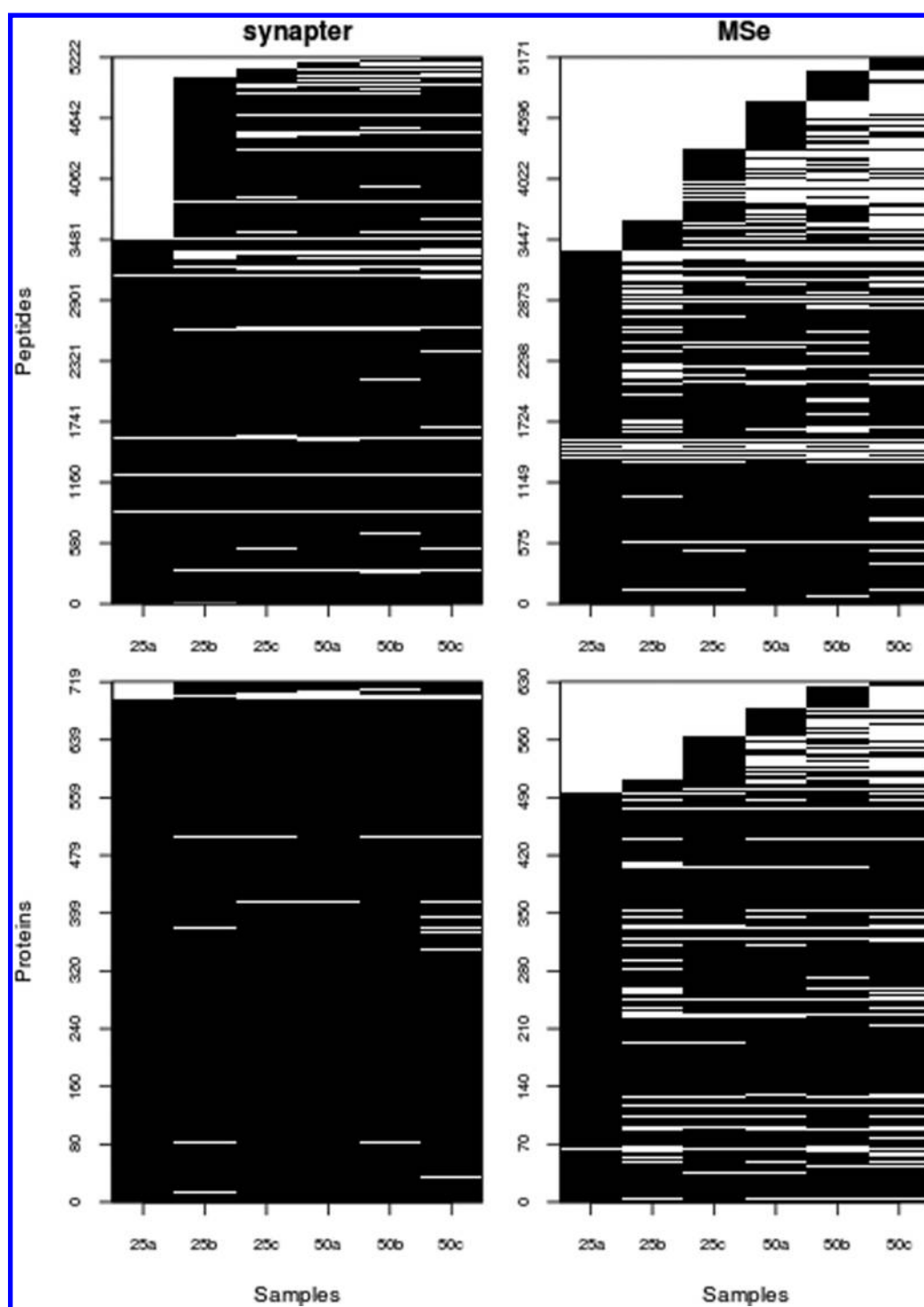


Figure 8. Illustration of missing data (white stripes) across an experiment containing 2 groups (25 and 50 fmols of spikes respectively) of 3 samples. The top and bottom figures represent peptide and protein data respectively. Synapter, on the left, substantially reduces the proportion of missing quantitation data compared to MS^E (right).

Figure 9 summarizes the results, showing that the majority of UPS1 proteins exhibit the expected fold-change and show significant q -values. The detailed description of the statistical analysis, including data and R code, is provided in the package documentation to allow other users to reproduce our findings or replicate this pipeline on their own data.

Increasing Quantitation Accuracy for High Intensity Peptides. Careful inspection of peptides quantified in MS^E and HDMS^E reveals that IMS, albeit beneficial to the identification

process, can lead to a combination of transmission loss and detector saturation, both of which occur in a peptide specific manner. To improve quantitative accuracy of higher intensity peptides, while maximizing the proteome coverage, it is beneficial to combine MS^E and HDMS^E data sets, using HDMS^E acquisitions as a source of peptide identifications (IR) and retrieve quantitative measurements from MS^E acquisitions (QR). To demonstrate the advantage of combining MS^E and HDMS^E data we spiked a six protein digest standard in *E. coli*

Table 4. Influence of the Identification Transfer on the Number of Missing Values at the Peptide and Protein Level for the Six Samples, Three for 25 fmol UPS1 Samples, Three for 50 fmol UPS1 Samples^a

	number of missing values out of a total of 6						total
	0	1	2	3	4	5	
Peptide MS ^E	2047	574	505	459	547	1039	5171
Peptide synapter	2850	1519	267	184	167	249	5236
Protein MS ^E	378	47	40	40	46	79	630
Protein synapter	674	26	6	5	3	8	722

Usable features Nonusable features

^aNote that one of the 6 usable synapter proteins with 2 missing values (last row), bears the two missing values in the same group (25 fmol UPS1 loading), effectively resulting in 705 usable proteins instead of 706.

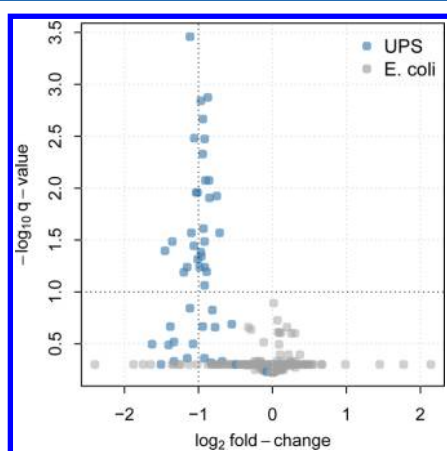


Figure 9. Volcano plot of the analysis for differentially expressed proteins at 25 and 50 fmol UPS1 loading. The quantitative information was acquired in MS^E mode and identifications were transferred from HDMS^E master. Blue and gray dots represent UPS1 and *E. coli* background proteins respectively. There are 31 UPS1 and 1 *E. coli* proteins that were found to be differentially expressed (q -value lower than 0.1).

lysate background at increasing amounts to allow injection of 25, 50, 75, 100, 250 fmol of standards and 1.5 μ g of *E. coli* lysate on column. Each spike loading was analyzed in triplicate in both modes. All acquisitions, from both modes were analyzed in PLGS and the resultant peptide identifications subjected to synapter filtering (peptide FDR = 0.01, protein FDR = 0.01, database uniqueness filter applied). Identifications were transferred to all MS^E QRs from the second replicate of 250 fmol loading HDMS^E acquisition, since this was the HDMS^E run that contained most (72) spike peptides suitable for quantitation.

Seventeen spike peptides were observed in at least two of three replicates at all loadings under all three experimental setups (MS^E, HDMS^E, HDMS^E identifications transferred to MS^E). In order to assess the linearity of response, peptide intensities were plotted against their loadings. Then a linear regression was extrapolated from the first four points (25, 50, 75, 100 fmol). If the 250 fmol measurement deviated from the value predicted by the regression model by more than 2 standard errors of the mean and at least 10% of the predicted value, the peptide was considered to exhibit a nonlinear response, characteristic of saturation with IMS. In total, 13 out

of 17 peptides exhibited saturation in HDMS^E mode and one peptide in MS^E mode. Figure 10 demonstrates two of such peptides chosen at random. Similar plots are provide in Supplementary Figure 1, Supporting Information, for all peptides. Although MS^E provides a higher dynamic range, its peak capacity is lower than that in HDMS^E.^{43,44} Thus, while MS^E is more accurate at high intensities, HDMS^E quantitation is more precise, especially for lower abundance peptides (see ref 30 for a detailed discussion of respective benefits).

DISCUSSION

Label-free approaches are becoming one of the most popular methods of quantitative proteomics.⁴ Although improvements in mass spectrometry and their DIA methods serve to improve both qualitative and quantitative analysis, large gains can be achieved by experimental design and careful treatment of data, postacquisition. To this end, we have developed software that serves to reduce the data to a set of confident peptide identifications and transfer these between acquisitions to maximize protein coverage and subsequent quantitation accuracy. Synapter seeks to optimize a number of key parameters during this process that have profound influence on data analysis, while it also provides comprehensive graphical representations of the data to provide the opportunity for the user to dictate these parameters. Development of the synapter package has allowed us to highlight some label-free quantitation subtleties. Several commercial and noncommercial software is available for label-free area under the curve quantitation on other platforms (for review see Vandenberg et al.²¹). Since synapter is designed for DIA acquisitions performed on Waters instruments, below we specifically compare its functionality with similarly targeted software.

First, synapter allows versatile filtering of the data. As mentioned earlier, PLGS performs the database search in a total of three stages. Only high confidence, unmodified fully tryptic peptides from stage one and two are used for label-free quantitation. Additional peptide identifications reinforce the confidence of protein identifications, playing no role in quantitation. Thus different criteria of stringency are applied to peptides identified during different stages of the database search.¹⁵ Figure 2 demonstrates that peptide identifications made in the second stage of the database search tend to have an overall lower score than identifications from first stage of database search. Synapter removes peptide types which cannot be used in quantitation (missed cleavage, in source fragment, etc.) and treats peptides from stages one and two of the database search separately when computing identification statistics. In comparison, Scaffold, a proprietary third party software that uses the Peptide Prophet⁴⁵ algorithm to estimate peptide FDR, splits peptides into four groups depending on their charge (from +1 to +4) and normal distributions are fitted through scores distribution of decoy and regular identifications. Thus Scaffold mixes peptides identified at different stages of PLGS's database search, which, given that different criteria of stringency is applied to different peptide identifications, can have an adverse effect on computing scores distributions.

Second, synapter, similarly to other software (Expression,⁹ Rosetta Elucidator,³¹ Progenesis³²) allows transferring identifications between data independent acquisitions performed on Waters instruments. A number of published algorithms were validated by comparing identifications of peptide ions acquired by MS/MS and identifications for the same EMRTs transferred from other acquisitions. In addition Prakash et al. demonstrated

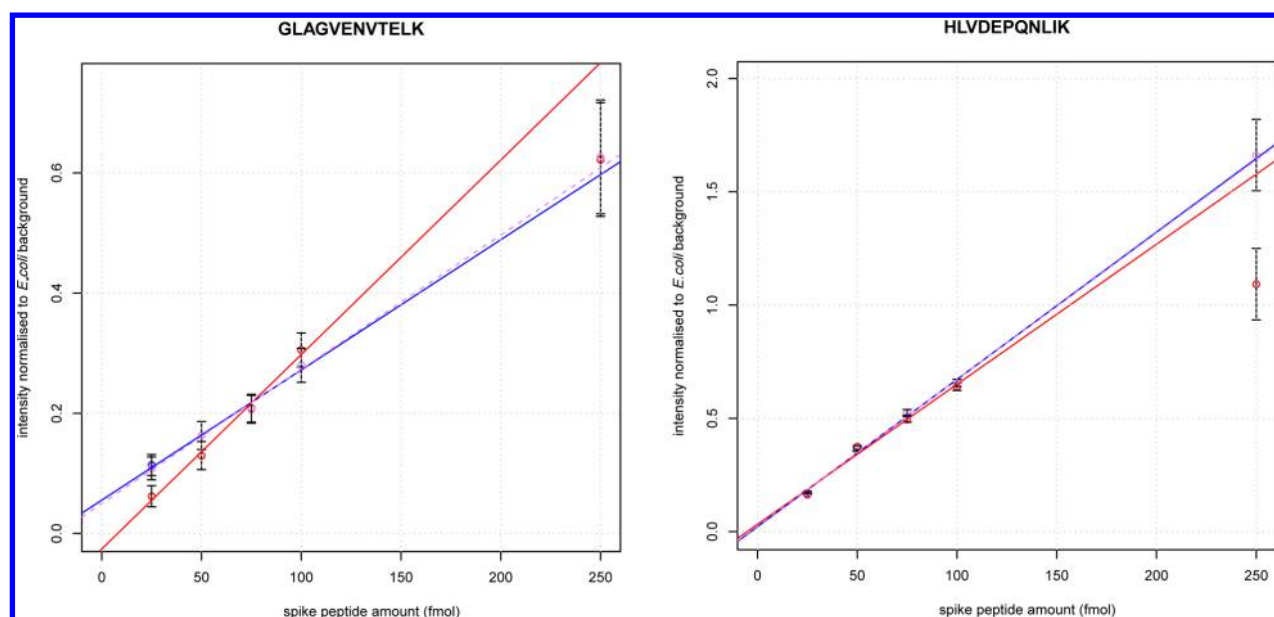


Figure 10. The intensity of two spike peptides at different loadings in HDMS^E (red dots), MS^E (blue dots) and MS^E for which identifications are transferred from HDMS^E (purple dots). Error bars represent 2 standard errors of the mean. The lines (red, HDMS^E; blue, MS^E; dashed purple, MS^E for which identifications are transferred from HDMS^E) are regressions through the first four points to approximate expected peptide intensity at 250 fmol. The intensity of peptides is normalized to a group of *E. coli* background proteins (sum of top three most intense peptides for each protein). As expected, no saturation is observed in MS^E runs or when identifications are transferred to MS^E runs from HDMS^E.

how implementing different tolerances affects the number of correct and incorrect identifications transferred.⁴⁶ Synapter models retention time and selects retention time and mass tolerances using peptides identified by MS/MS in both IR and QR, which allows it to compute the number and proportion of correct identification transfers at different tolerances during synapter analysis in an automated fashion. Other methods have been proposed to compute the proportion of incorrectly transferred identifications that do not rely on a set of commonly identified peptides.⁴⁷

As shown in Figure 5 and in Supplementary Tables 1–3, Supporting Information, selecting optimal tolerances has a profound effect on the success of identification transfer. The grid search employed by synapter provides an automated way to assess and select optimal tolerance values while simplifying the task for the user. Other packages compatible with Waters .raw data or PLGS output mostly require either a user to specify tolerances *a priori*, or simply use a set of common tolerances for every analysis. Rosetta Elucidator requires *a priori* specifications of tolerance parameters and Expression does not allow a user to specify tolerances through the graphical user interface, it initially uses 20 ppm mass and 5 min retention time tolerances and then iteratively refines LOESS retention time model.

Furthermore, transferring identifications from multiple runs in the analysis will cause a proportional increase of incorrect identifications (FDR) within the analysis and none of the software that we are aware of attempt to estimate the extent or minimize this effect. Synapter, however, enables a balance between increasing the number of successfully quantified peptides and increasing FDR, allowing the user to decide which is preferable within a large scale experiment.

One of the potential applications of identification transfer is combining qualitative and quantitative information acquired with and without IMS to benefit from deeper proteome coverage and higher dynamic range. Necessity to run samples in both modes, only marginally increases the instrument time

required for analysis, since replication is needed in MS^E for accurate quantitation. Indeed, replicating identical HDMS^E acquisitions leads to a marginal increase of identifications, and consecutively a high FDR increase (eq 1). For typical analyses we would recommend to run a number of biological replicates for each condition as QR MS^E acquisitions and a single pool of replicates for each condition as IR HDMS^E acquisition. Thus if analysis was performed in triplicate in MS^E, an additional 33% of instrument time would be required for a substantially improved proteome coverage and reduction in missing data by transferring identifications from IR HDMS^E. The proportion of HDMS^E IR acquisitions will subsequently decrease as the number of QR replication increases.

Third, synapter is developed for the R programming environment, which is specifically designed for robust statistical data analysis and allows efficient results visualization. While other commercial software (including those mentioned above), provide their own statistical tools, none can provide the flexibility and quality of R and the many packages that provide ready to use functionality highly relevant to sound high throughput data analysis. In our case, we have used the MSnbase package for downstream data manipulations, built-in statistical functionality to perform the statistical test and the q-value package⁴⁸ for FDR control. All these packages are directly interoperable and constitute a concise and consistent data analysis pipeline. Additional biologically relevant gene ontology and pathway analysis are readily available from the Bioconductor project⁴⁹

Finally, synapter comes with extensive documentation. It is distributed through the Bioconductor project (<http://www.bioconductor.org/packages/release/bioc/html/synapter.html>), benefitting of a facile installation framework and community support. It can be operated at different levels allowing maximal flexibility. A simple graphical user interface allows a new user to utilize the package with minimal R knowledge. A single and flexible function allows one to complete the synergise algorithm

for easy and reproducible batch processing. Experienced users and developers have access to low-level functionality to control every aspect of the pipeline. Supplementary File 1, Supporting Information, provides a brief practical overview of synapter to demonstrate its ease of use. Although a complete synapter pipeline can be executed without user intervention, we provide detailed logs and numerous quality and summary plots and tables as a comprehensive html report for careful inspection. While automated data analyses is essential in any contemporary high-throughput experiment, it is crucial that users are given the possibility to keep track of the processing and transformation applied to the data and the decisions that are made for them.

■ ASSOCIATED CONTENT

Supporting Information

Supplemental figures and overview of synapter. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: lg390@cam.ac.uk. Tel: +441223760255. Fax: +441223760241.

Present Address

[†]N.J.B.: MRC Human Nutrition Research, Elsie Widdowson Laboratory, 120 Fulbourn Road, Cambridge, CB1 9NL.

Author Contributions

[†]N.J.B. and P.V.S. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge Hans Vissers for useful comments about the manuscript. L.G. was supported by a seventh Framework Programme of the European Union (262067- PRIME-XS). P.V.S. was supported by The Darwin Trust of Edinburgh.

■ REFERENCES

- (1) Schulze, W. X.; Usadel, B. Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* **2010**, *61*, 491–516.
- (2) Gevaert, K.; Impens, F.; Ghesquière, B.; Van Damme, P.; Lambrechts, A.; Vandekerckhove, J. Stable isotopic labeling in proteomics. *Proteomics* **2008**, *8*, 4873–4885.
- (3) Neilson, K. A.; Ali, N. A.; Muralidharan, S.; Mirzaei, M.; Mariani, M.; Assadourian, G.; Lee, A.; van Sluyter, S. C.; Haynes, P. A. Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* **2011**, *11*, 535–553.
- (4) Evans, C.; Noirel, J.; Ow, S. Y.; Salim, M.; Pereira-Medrano, A. G.; Couto, N.; Pandhal, J.; Smith, D.; Pham, T. K.; Karunakaran, E.; Zou, X.; Biggs, C. A.; Wright, P. C. An insight into iTRAQ: where do we stand now? *Anal. Bioanal. Chem.* **2012**, *404* (4), 1011–1027.
- (5) Ting, L.; Rad, R.; Gygi, S. P.; Haas, W. MS3 eliminates ratio distortion in isobaric labeling-based multiplexed quantitative proteomics. *Nat. Methods* **2011**, *8*, 937–940.
- (6) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **2005**, *4*, 1265–1272.
- (7) Levin, Y.; Hradetzky, E.; Bahn, S. Quantification of proteins using data-independent analysis (MSE) in simple and complex samples: a systematic evaluation. *Proteomics* **2011**, *11*, 3273–3287.
- (8) Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. Absolute quantification of proteins by LCMSE: a

virtue of parallel MS acquisition. *Mol. Cell. Proteomics* **2006**, *5*, 144–156.

(9) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G.-Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **2005**, *77*, 2187–2200.

(10) Geromanos, S. J.; Vissers, J. P. C.; Silva, J. C.; Dorschel, C. A.; Li, G.-Z.; Gorenstein, M. V.; Bateman, R. H.; Langridge, J. I. The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics* **2009**, *9*, 1683–1695.

(11) Krishnamurthy, D.; Levin, Y.; Harris, L. W.; Umrana, Y.; Bahn, S.; Guest, P. C. Analysis of the human pituitary proteome by data independent label-free liquid chromatography tandem mass spectrometry. *Proteomics* **2011**, *11*, 495–500.

(12) Kennedy, J.; Yi, E. C. Use of gas-phase fractionation to increase protein identifications: application to the peroxisome. *Methods Mol. Biol.* **2008**, *432*, 217–228.

(13) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal. Chem.* **2007**, *79*, 5620–5632.

(14) Luethy, R.; Kessner, D. E.; Katz, J. E.; Maclean, B.; Grothe, R.; Kani, K.; Faça, V.; Pitteri, S.; Hanash, S.; Agus, D. B.; Mallick, P. Precursor-ion mass re-estimation improves peptide identification on hybrid instruments. *J. Proteome Res.* **2008**, *7*, 4031–4039.

(15) Li, G.-Z.; Vissers, J. P. C.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J. Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **2009**, *9*, 1696–1719.

(16) Gillet, L. C.; Navarro, P.; Tate, S.; Roest, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics* **2012**, *11* (6), No. O111.016717.

(17) Panchaud, A.; Scherl, A.; Shaffer, S. A.; von Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R. Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **2009**, *81*, 6481–6488.

(18) Geiger, T.; Cox, J.; Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Mol. Cell. Proteomics* **2010**, *9*, 2252–2261.

(19) Masselon, C.; Anderson, G. A.; Harkewicz, R.; Bruce, J. E.; Pasa-Tolic, L.; Smith, R. D. Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures. *Anal. Chem.* **2000**, *72* (8), 1918–1924.

(20) Michalski, A.; Cox, J.; Mann, M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10*, 1785–1793.

(21) Vandenbogaert, M.; Li-Thiao-Té, S.; Kaltenbach, H.-M.; Zhang, R.; Aittokallio, T.; Schwikowski, B. Alignment of LC-MS images, with applications to biomarker discovery and protein identification. *Proteomics* **2008**, *8*, 650–672.

(22) Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolić, L.; Smith, R. D. Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* **2000**, *72*, 3349–3354.

(23) Lipton, M. S.; Pasa-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarides, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Strittmatter, E.; Tolić, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K.-K.; Zhao, R.; Smith, R. D. Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049–11054.

(24) Strittmatter, E. F.; Ferguson, P. L.; Tang, K.; Smith, R. D. Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 980–991.

- (25) Pasa-Tolić, L.; Masselon, C.; Barry, R. C.; Shen, Y.; Smith, R. D. Proteomic analyses using an accurate mass and time tag strategy. *BioTechniques* **2004**, *37*, 621–624, 626–633, 636 passim.
- (26) Myung, S.; Lee, Y. J.; Moon, M. H.; Taraszka, J.; Sowell, R.; Koeniger, S.; Hilderbrand, A. E.; Valentine, S. J.; Cherbas, L.; Cherbas, P.; Kaufmann, T. C.; Miller, D. F.; Mechref, Y.; Novotny, M. V.; Ewing, M. A.; Spörleder, C. R.; Clemmer, D. E. Development of high-sensitivity ion trap ion mobility spectrometry time-of-flight techniques: a high-throughput nano-LC-IMS-TOF separation of peptides arising from a *Drosophila* protein extract. *Anal. Chem.* **2003**, *75*, 5137–5145.
- (27) Pringle, S. D.; Giles, K.; Wildgoose, J. L.; Williams, J. P.; Slade, S. E.; Thalassinou, K.; Bateman, R. H.; Bowers, M. T.; Scrivens, J. H. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass Spectrom.* **2007**, *261*, 1–12.
- (28) Hoaglund, C. S.; Valentine, S. J.; Clemmer, D. E. An ion trap interface for esi-ion mobility experiments. *Anal. Chem.* **1997**, *69*, 4156–4161.
- (29) Ibrahim, Y. M.; Prior, D. C.; Baker, E. S.; Smith, R. D.; Belov, M. E. Characterization of an ion mobility-multiplexed collision induced dissociation-tandem time-of-flight mass spectrometry approach. *Int. J. Mass Spectrom.* **2010**, *293*, 34–44.
- (30) Shliaha, P. V.; Bond, N. J.; Gatto, L.; Lilley, K. S. Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies. *J. Proteome Res.* **2013**, DOI: pr300775k.
- (31) Neubert, H.; Bonnert, T. P.; Rumpel, K.; Hunt, B. T.; Henle, E. S.; James, I. T. Label-free detection of differential protein expression by LC/MALDI mass spectrometry. *J. Proteome Res.* **2008**, *7*, 2270–2279.
- (32) Stoop, M. P.; Coulier, L.; Rosenling, T.; Shi, S.; Smolinska, A. M.; Buydens, L.; Ampt, K.; Stingl, C.; Dane, A.; Muilwijk, B.; Luitwieler, R. L.; Sillevius Smitt, P. A. E.; Hintzen, R. Q.; Bischoff, R.; Wijmenga, S. S.; Hankemeier, T.; van Gool, A. J.; Luiders, T. M. Quantitative proteomics and metabolomics analysis of normal human cerebrospinal fluid samples. *Mol. Cell Proteomics* **2010**, *9*, 2063–2075.
- (33) R Development Core Team R: *A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011.
- (34) Gatto, L.; Lilley, K. S. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **2011**, *28*, 288–289.
- (35) Katajamaa, M.; Oresic, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinform.* **2005**, *6*, 179.
- (36) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* **2008**, *7*, 29–34.
- (37) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **1995**, *57*, 289–300.
- (38) Storey, J. D.; Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9440–9445.
- (39) Olsen, J. V.; Ong, S.-E.; Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **2004**, *3*, 608–614.
- (40) Thiede, B.; Lamer, S.; Mattow, J.; Siejak, F.; Dimmler, C.; Rudel, T.; Jungblut, P. R. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 496–502.
- (41) Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836.
- (42) Mueller, L. N.; Rinner, O.; Schmidt, A.; Letarte, S.; Bodenmiller, B.; Brusniak, M.-Y.; Vitek, O.; Aebersold, R.; Müller, M. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **2007**, *7*, 3470–3480.
- (43) Geromanos, S. J.; Hughes, C.; Golick, D.; Ciavarini, S.; Gorenstein, M. V.; Richardson, K.; Hoyes, J. B.; Vissers, J. P. C.; Langridge, J. I. Simulating and validating proteomics data and search results. *Proteomics* **2011**, *11*, 1189–1211.
- (44) Geromanos, S. J.; Hughes, C.; Ciavarini, S.; Vissers, J. P. C.; Langridge, J. I. Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Anal. Bioanal. Chem.* **2012**, *404* (4), 1127–1139.
- (45) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383–5392.
- (46) Prakash, A.; Mallick, P.; Whiteaker, J.; Zhang, H.; Paulovich, A.; Flory, M.; Lee, H.; Aebersold, R.; Schwikowski, B. Signal Maps for Mass Spectrometry-based Comparative Proteomics. *Mol. Cell. Proteomics* **2006**, *5*, 423–432.
- (47) Stanley, J. R.; Adkins, J. N.; Slys, G. W.; Monroe, M. E.; Purvine, S. O.; Karpievitch, Y. V.; Anderson, G. A.; Smith, R. D.; Dabney, A. R. A statistical method for assessing peptide identification confidence in accurate mass and time tag proteomics. *Anal. Chem.* **2011**, *83* (16), 6135–6140.
- (48) Dabney, A.; Storey, J. D. and with assistance from Warnes, G. R. *qvalue: Q-value estimation for false discovery rate control*, version 1.35.0.
- (49) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y. H.; Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.