research articles Journal of proteome research

Precursor-Ion Mass Re-Estimation Improves Peptide Identification on Hybrid Instruments

Roland Luethy,[†] Darren E. Kessner,[†] Jonathan E. Katz,[†,‡] Brendan MacLean,[§] Robert Grothe,[†] Kian Kani,[†] Vitor Faça,[||] Sharon Pitteri,[||] Samir Hanash,[||] David B. Agus,[†] and Parag Mallick*[,†,‡]

Spielberg Family Center for Applied Proteomics, Cedars-Sinai Medical Center, Los Angeles, California 90048, Department of Chemistry & Biochemistry, University of California, Los Angeles, California 90095, Labkey Software, 312 N. 49th Street, Seattle, Washington 98103, and Fred Hutchinson Cancer Center, 312 N. 49th Street, Seattle, Washington 98103

Received April 22, 2008

Mass spectrometry-based proteomics experiments have become an important tool for studying biological systems. Identifying the proteins in complex mixtures by assigning peptide fragmentation spectra to peptide sequences is an important step in the proteomics process. The 1−2 ppm mass-accuracy of hybrid instruments, like the LTQ-FT, has been cited as a key factor in their ability to identify a larger number of peptides with greater confidence than competing instruments. However, in replicate experiments of an 18-protein mixture, we note parent masses deviate 171 ppm, on average, for ion-trap data directed identifications and 8 ppm, on average, for preview Fourier transform (FT) data directed identifications. These deviations are neither caused by poor calibration nor by excessive ion-loading and are most likely due to errors in parent mass estimation. To improve these deviations, we introduce msPrefix, a program to re-estimate a peptide's parent mass from an associated high-accuracy full-scan survey spectrum. In 18-protein mixture experiments, msPrefix parent mass estimates deviate only 1 ppm, on average, from the identified peptides. In a cell lysate experiment searched with a tolerance of 50 ppm, 2295 peptides were confidently identified using native data and 4560 using msPrefixed data. Likewise, in a plasma experiment searched with a tolerance of 50 ppm, 326 peptides were identified using native data and 1216 using msPrefixed data. msPrefix is also able to determine which MS/MS spectra were possibly derived from multiple precursor ions. In complex mixture experiments, we demonstrate that more than 50% of triggered MS/MS may have had multiple precursor ions and note that spectra with multiple candidate ions are less likely to result in an identification using TANDEM. These results demonstrate integration of msPrefix into traditional shotgun proteomics workflows significantly improves identification results.

Keywords: Precursor estimation • tandem MS • peptide identification • computational proteomics • FTMS • mass accuracy

Introduction

Over the past few years, a number of mass spectrometry (MS)-based quantitative proteomics methods have been developed that attempt to comprehensively identify the proteins present in complex samples. Using MS proteomics to characterize cellular functions and pathways affected by perturbations and disease,[1–5] and identifying new components and changes in the composition of protein complexes and organelles,[6–8] has led to the detection of putative disease biomarkers.[9,10]

In common MS-based proteomic pipelines,[11,12] protein samples are first partially purified or separated by chromatographic or electrophoretic methods and then digested with trypsin, resulting in highly complex peptide mixtures. These are further separated by liquid chromatography (one or more stages) and analyzed with a [tandem] mass spectrometer (LC [LC]-MS/MS). Last, computational tools attempt to assign one or more peptide sequences to each tandem spectra.

There are several tools available to identify peptides by correlating theoretical fragmentation spectra derived from known sequence databases to observed MS/MS spectra, the most prevalent being SEQUEST, Mascot and more recently, TANDEM.[13–15] Other tools attempt *de novo* interpretation of the observed spectra.[16,17] Both approaches have advantages and disadvantages,[18] but they both use the mass-to-charge ratio of the intact peptide-ion from the survey spectrum (the "precursor" or "parent ion") to limit the search space of possible reported peptide identifications. As such, more ac-

* Address correspondence to: Parag Mallick, Spielberg Family Center for Applied Proteomics, Cedars-Sinai Medical Center, 8750 West Beverly Blvd, Los Angeles, CA 90048. E-mail: parag.mallick@cshs.org. Phone: 310-423-7600. Fax: 310-423-8543.
[†] Cedars-Sinai Medical Center.
[‡] University of California.
[§] Labkey Software.
[||] Fred Hutchinson Cancer Center.

10.1021/pr800307m CCC: $40.75

Published on Web 08/16/2008

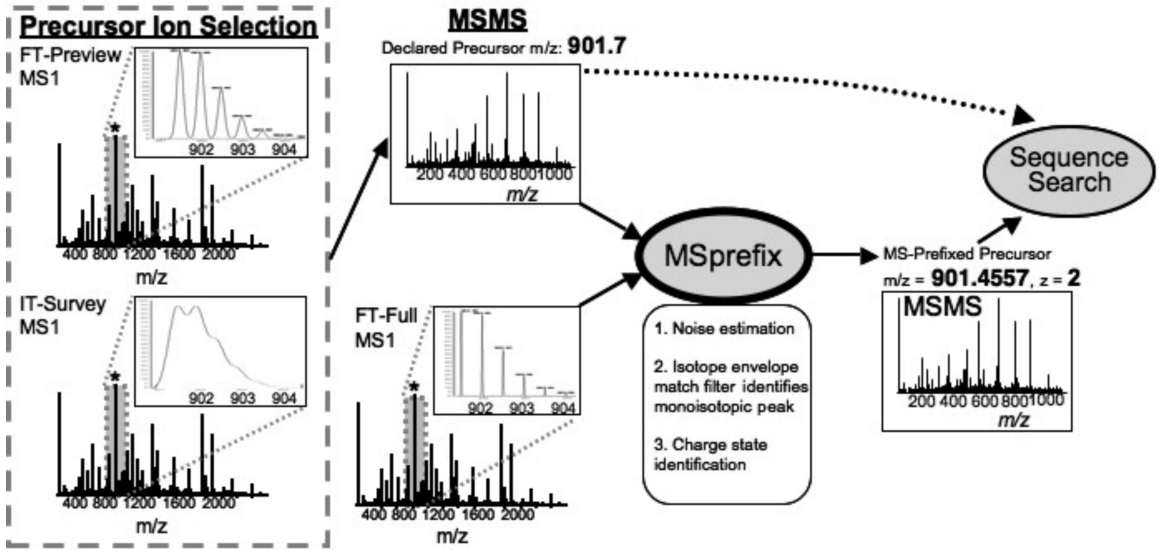

Figure 1. Overview of msPrefix data processing pipeline. Typically, the precursor or parent mass used for sequence searching is derived from an *m*/*z* triggering an MS/MS event. In data-dependent mode, this triggering *m*/*z* is determined from an FT preview scan or an IT survey scan. msPrefix refines this estimate of the precursor mass by attempting to correlate the signal detected in the low-resolution preview and IT spectra with a signal in the corresponding high-resolution FT spectrum. Signals in the high-resolution FT spectrum are extracted by searching the region neighboring the initial estimate with match filters for isotope envelopes for each of the likely peptide charge states. The best matching set of peaks is selected and used to estimate an adjusted precursor *m*/*z* and charge.

curate parent mass-to-charge estimation leads to more confident and faster peptide identifications.[19]

As shown in the left side of Figure 1, ions from a survey scan (large mass range, no isolation or fragmentation) can be selected for isolation and fragmentation. The determination of an *m*/*z* to isolate and fragment is called 'triggering' a fragmentation event. Ions may be selected for fragmentation by a process known as 'data-dependent' selection where predefined parameters, such as "most abundant ion" are measured in a survey scan and then used to assess which *m*/*z* to isolate and fragment.

The LTQ-FT instruments[20] used in this study have two mass spectrometers that can operate mostly in parallel: a slower Fourier transform-ICR (FT) and a faster ion trap (IT). Under our operation, a cycle consists of an FT MS scan in parallel with an IT MS scan and several IT MS/MS scans whose precursor ions are selected in a data-dependent manner. The determination of which ions are to be selected can use either the information from the preceding IT spectrum or from a spectrum ("preview scan") derived from the first quarter of the FT scan as the transient data are being acquired. Regardless of which spectrum is used to trigger an MS/MS, the isolation and fragmentation is performed in the IT, which isolates ions within a window of the triggering *m*/*z*. The triggering-*m*/*z*, an estimate of the intact parent *m*/*z* of the peptide(s) selected for fragmentation, is stored along with the MS/MS spectrum and is given to the identification program to aid in the inference of the likely peptide sequence(s) of the selected ion(s).

As shown in Figure 1, msPrefix intercedes between data collection and computational identification to improve the precision of the precursor mass by inspection of the preceding full-resolution FTMS survey-scan. Here, we compare how results of peptide identification are impacted by the re-estimation of the precursor mass by msPrefix. In particular, we note that using msPrefix can approximately double the number of high confidence identifications that can be made from a given data set.

Experimental Procedures

Protein Samples. A mixture of 18 proteins was used as described in ref 21.

Human blood plasma was obtained from Bioreclamation (Hicksville, NY). Blood was collected into vials containing K_3 EDTA as an anticoagulant, stored overnight at 4 °C, and then centrifuged at 2800*g* for 20 min at 4 °C; the resultant supernate from 20 normal males was pooled and used as our human plasma standard. Plasma was then prepared by mixing 1:9 with 3.3 M guanidine HCl in 100 mM phosphate buffer, pH 8 (Sigma-Aldrich, St. Louis, MO). The prepared plasma was then reduced by adding dithiothreitol (DTT, Sigma-Aldrich) to a final concentration of 10 mM and incubated for 1 h at 65 °C. Alkylation was then accomplished by a 45 min incubation at room temperature in the dark after the addition of iodoacetamide (IAA, Sigma-Aldrich) to a final concentration of 55 mM. Plasma samples were then incubated at 37 °C for 18 h after the 2:1 addition of a 55 mM ammonium bicarbonate solution containing a 1:20 (w/w) aliquot of sequencing grade trypsin (Promega, Madison, WI) to estimated plasma protein. Reaction was quenched by the addition of glacial acetic acid to a final concentration of 1% (v/v).

Human prostate carcinoma epithelial cells (22Rv1) were obtained from the American Type Culture Collection (ATCC, Manassas, VA). Cells were grown to 70% confluency and lysed in phosphate based saline (Invitrogen) with 1% *n*-octyl glucoside (Sigma-Aldrich) with protease inhibitors (Pierce, Rockford, IL) and sonicated. Lysates were centrifuged to remove insoluble material and protein concentration was determined by BCA assay (Pierce). The lysates were mixed 1:1 with 6 M Guanidine Hydrochloride (Sigma-Aldrich), reduced, alkylated, and subjected to whole protein reversed-phase fractionation as previously described.[22] The fractions were pooled with between 50 and 200 μg of protein per fraction prior to trypsinization (sequencing grade, Sigma-Aldrich) for 18 h at 37 °C in a water

bath. Reaction was quenched by the addition of glacial acetic acid to a final concentration of 1% (v/v).

Mass Spectrometry Experiments and Chromatography. Samples were analyzed on an LTQ-FT (Thermo Fisher Scientific, San Jose, CA) hybrid mass spectrometer. LC was performed on 100 mm × 1 mm columns filled with a Betasil C18 (3 μm bead size 100 Å pore size) resin (Thermo Fisher) using a stepwise linear gradient from 95% buffer A (0.1% formic acid in water, Burdick and Jackson) to 95% buffer B (0.1% formic acid in acetonitrile, Burdick and Jackson) as follows: hold for 20 min at 95% A, 23 min 90% A, 27 min 85% A, 80 min 65% A, 85 min 5% A, hold until 95 min, 98 min 95% A. During elution, one of two mass spectrometer configurations was used. In the first configuration, a Thermo-Fisher LTQ-FT running Tune v. 1.1b7 and Xcalibur v. 1.4SR1 was configured to repeatedly perform 6 separate measurements: a full ion-trap (IT) survey scan from 400 to 1800 m/z, a full FT survey scan from 400 to 1800 m/z and then 4 MS/MS IT scans on the 4 largest peaks from the preceding IT or FT survey scan (see Results). Potential precursor masses of FT triggered MS/MS scans were rejected if the precursor ion was identified as being singly charged. Additionally, dynamic exclusion was sometimes used as a means of limiting the repeated triggering based on the same precursor mass. Our exclusion criteria were defined to minimize collection of MS/MS for the same precursor mass (less 0.55 m/z to +1.55 m/z) in the same 45 s window. Our exclusion list size was limited to the last 100 triggered masses. Ion target values were set at 30 000, 10 000, and 1 000 000 ions for ion trap full scans, ion trap MS2 scans, and FT full scans, respectively. The first configuration was used only in Figure 7. The second configuration, used for all figures, consisted of a Thermo-Fisher LTQ-FT Ultra running Tune v. 2.2SP1 and Xcalibur v. 2.0.7 otherwise configured identically.

Following data collection, files were converted from RAW to mzXML using the ReAdW program from the NHLBI Proteome Center at the Institute for Systems Biology (sashimi.sourceforge.net). References to 'native' precursor masses refer to the value contained in the mzXML as extracted from the RAW file.

ISB Sample 3 data, collected on an LTQ-FT running Tune v 2.0 and Xcalibur v 2.0, was downloaded from the ISB (http://regis-web.systemsbiology.net/PublicDatasets/). The IPAS data set from the Hanash laboratory was generated on an Thermo-Fisher LTQ-FT running Xcalibur v 2.0 and Tune v 2.2.

msPrefix Algorithm Description. msPrefix re-estimates the precursor m/z value for each tandem mass spectrum in a data file (either native ".RAW" format or mzXML). In brief, the algorithm has four steps: (1) determine a window within the FT spectra to search for a precursor ion, (2) estimate the noise floor of the window, (3) find peaks in the window that lie significantly above the noise floor, (4) assess which of those peaks compose an isotope envelope. Source code for msPrefix is available at http://www.sfcap.cshs.org. Below we detail each of the steps msPrefix uses to 'fix' the native estimate of the precursor m/z.

For each tandem mass spectrum (either FT or IT triggered), we first determine which FT survey scan is associated. We define the search window for the true precursor ion m/z to be $[x - 3 < m/z < x + 1.6]$, where x is the native estimate of the precursor m/z value in atomic mass units. The window was selected to minimize false positives.

The msPrefix search algorithm operates in the frequency domain as frequency-domain data is regularly spaced and signals are of uniform shape, which facilitates efficient matched filter calculations. m/z-domain data is first converted to frequency-domain data by inverting the FT calibration function ($m/z = A/f + B/(f^2)$ for FT and $m/z = A/(f^2)$ for Orbitrap) that maps frequency to m/z; the calibration parameters are derived from constants stored in the data file. FT data must be collected in profile mode for matched filter calculations to effectively estimate m/z. If calibration parameters are not found in the file, we default to $A = 1.075 \times 10^8$, $B = 3.455 \times 10^8$ for FT and 4.753×10^{13} for Orbitrap.

Having defined the search window, we next estimate the noise floor by a simple two pass calculation: the first pass discards signal above one standard deviation from the mean; the second pass calculates the noise floor as one standard deviation above the mean of the remaining sample intensities.

For each point in the search window, we score the match between our Lorentzian kernel function (the Fourier transform of a decaying sinusoid) and the frequency domain data by computing the dot product of the kernel function with the frequency-domain data. An initial list of peaks is obtained by searching the dot product results for local maxima that exceed a 4× multiple of the noise floor. When the correlation between the matched filter and a local region of the spectrum exceeded this threshold, a peak was judged to be present. A matched filter for isotope envelopes of charge states (1+ to 6+) was constructed by combining multiple copies of the original filter. The spacing between copies is determined by the charge state; the relative intensities are determined by the averagine model[23] for a peptide of mass m ($m/z \times z$).

We discard peaks whose best isotope envelope score is less than 4× the noise floor score. As a particular monoisotope will be reported by each of the peaks in the envelope, we collapse the peak list. After finding monoisotopic peaks in the frequency domain, we convert the frequency values of the peaks back to m/z values using the calibration function specified above. For each isotope envelope, we report a monoisotopic mass, charge state and score for the peak closest to the original precursor m/z. In single precursor mode, alternate candidate masses are saved to a log file. In multiple precursor mode, each of the alternates is written to the mzXML file. We discuss the challenge of multiple candidate peaks in more detail in the Discussion.

Database Searches. Peptide searches were performed using TANDEM, and transproteomic pipeline tools (TPP).[24] The results were stored and visualized using the Computational Proteomics Analysis System (CPAS).[25] For searches on 18-protein mix experiments, the search database contained the sequences of the 18 proteins as well as approximately 67 000 decoys generated by reversing the sequences of the human IPI database and the 18 target proteins. Matches to one of the 18 proteins in the correct orientation were counted as true positives. Matches to a decoy sequence were considered false positives. For plasma samples and RV1 lysates in Figure 6, a database consisting of human IPI and 1.8 million decoys, generated by reversing the Uniref50 database, was searched. The searches were repeated with and without the msPrefix preprocessing and with different values of the precursor tolerance. Parameters for TANDEM searches were *K*-score pluggable scoring enabled,[26] Tryptic cleavage site, allowance for 1 missed-cleavage site. Refinement was allowed with potential modifications of asparagine and glutamine to capture deamidated residues. For results presented in Figure 9, a database of commonly observed human plasma proteins was

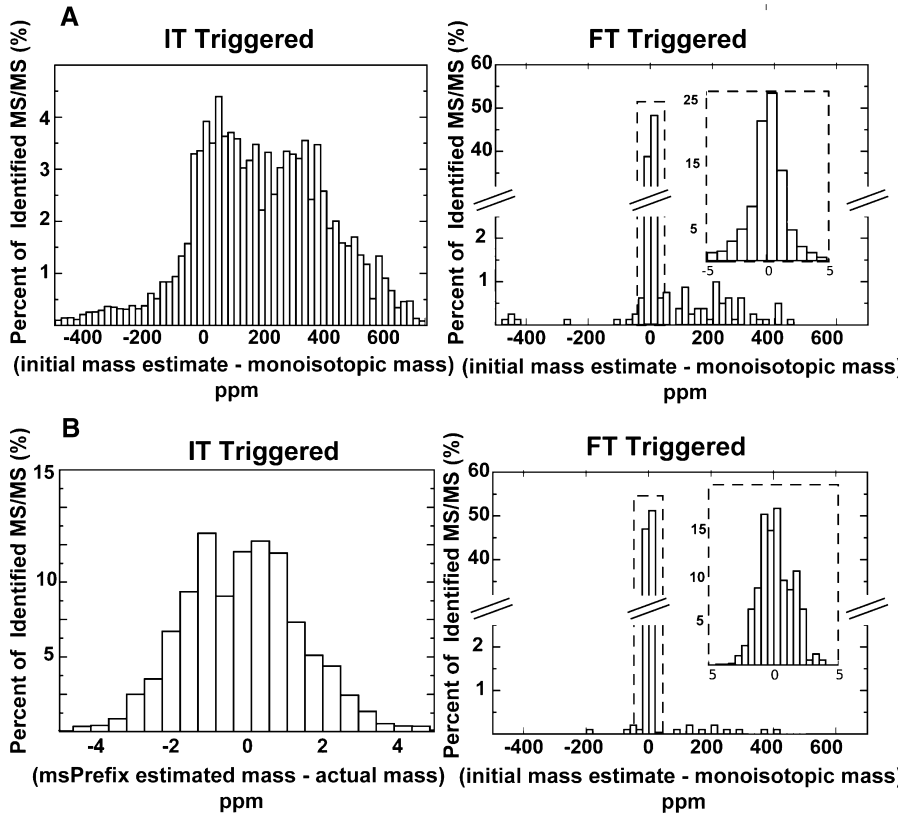

Figure 2. (A) Initial estimates of precursor mass significantly deviate from the true peptide masses. The delta-mass (the mass differences between initial estimates of the precursor mass and the calculated mass value for identified peptides) from the 18 known proteins was computed. The precursor tolerance for the database search was set to 0.9 Da. FTtriggered delta-masses were on average significantly smaller than IT-triggered. However, more than 11% were larger than 5 ppm. (B) msPrefix estimates of precursor mass deviate less from the true peptide masses Mass differences between msPrefix estimates of the precursor mass and the calculated mass value for matched peptides from the 18 known proteins were computed. The precursor tolerance for the database search was set to 0.9 Da. Following msPrefix, both FT-triggered and IT-triggered delta-masses were 98% of estimates and were within 4 ppm of the true mass.

constructed from the intersection of the HUPO-PPP[27] and Anderson et al.[28] lists.

Results

Native Estimates of Precursor Mass Have Large Mass Error. As noted in the Introduction, and depicted in Figure 1, the typical difference between the native estimate of the precursor mass and the actual peptide mass can be significant. To assess the extent of the difference, we first consider 21 replicates of a data set generated from a synthetic mixture containing 18 proteins.[21] For each identified peptide derived from one of the 18 proteins (see Experimental Procedures for details of TANDEM identification), we compute the delta-mass. Histograms of the delta-mass for the native estimates are plotted in Figure 2A and the msPrefix estimates in Figure 2B. The mean difference for ion trap data is 171 ppm with a standard deviation of 226 ppm. Even for FT-triggered data, the mean difference is 8 ppm with a standard deviation of 86 ppm. These deviations are significantly larger than the 1−2 ppm mass accuracy (as stated by Thermo-Fisher instrument specification) of the hybrid instruments. It is initially attractive to believe that the deviation is the result of operator error wherein the instrument was run extremely poorly calibrated or with ion-loading values orders of magnitude beyond suggested. However, as noted in Experimental Procedures, the instruments were recently calibrated and were run with target values of 10^6, quite within specification. Consequently, it seemed possible that the delta mass was instead large because of poor initial estimation of the precursor mass by the vendor software. Therefore, we hypothesized it may be possible to better estimate the precursor mass.

msPrefix Significantly Shifts Estimated Precursor Mass. Using the approach described in Experimental Procedures, msPrefix attempts to re-estimate the precursor mass. To determine the extent that msPrefix alters the initial estimate of the mass, we plot a histogram of the difference between the initial estimate and the msPrefix estimate as shown in Figure 3. Please note that, in some cases, msPrefix is shifting the estimated monoisotopic mass over 1000 ppm, potentially a shift of several Daltons. Even for FT-triggered precursors, msPrefix can shift the estimated precursor several hundred ppm.

msPrefix Shifted Precursor Masses Are More Accurate than Reported Precursor Masses. To determine the impact msPrefix has on peptide identifications, we next performed our TANDEM searches using a variety of precursor tolerance cutoffs both with and without msPrefix. As msPrefix also typically determines the charge state of the precursor, we also performed searches both considering and not-considering charge state to remove potential advantages and disadvantages that may have arisen from the charge state determination; searches without charge state information potentially have a higher rate of false identifications because TANDEM considers a larger hypothesis

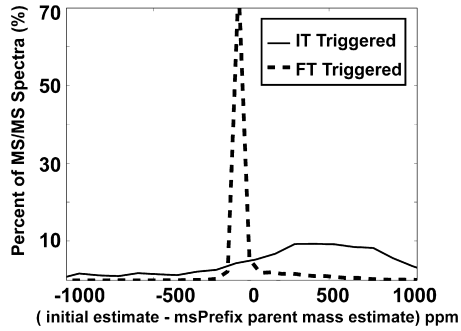

Figure 3. msPrefix precursor mass estimates are significantly different than native estimates. For both IT and FT-triggered precursor masses, msPrefix significantly alters the estimate of the precursor mass, in some cases more than 1000 ppm.

space. Although we have not observed incorrectly assigned charge states, it is possible that a small fraction of charge state determinations by msPrefix are wrong and lead to incorrect identifications. In Figure 4, we plot the percentage of peptide identifications made using the various precursor tolerance values and compare msPrefix with the native data. Note that, by using msPrefix, a majority (98%) of identifications are covered with a threshold below 4 ppm for both IT and FT-triggered experiments. To achieve a similar number of true positives without msPrefix, full 1000 ppm thresholds were required. We also note that charge has only minimal impact on the number of identifications. This data suggests msPrefix is significantly improving the estimate of precursor mass and thus the delta-mass. This data also validates the hypothesis that the instrument was functioning well, but providing poor precursor mass estimates. In the Discussion, we note that recent versions of vendor software (Tune, Xcalibur) partially improve the estimates. However, msPrefix is still highly relevant as it still has benefit on the most recent software. Furthermore, older data sets may greatly benefit from msPrefix.

To validate the hypothesis that msPrefix estimates are significantly improved, we plot the distribution of the differences of precursor m/z values after msPrefix and the calculated mass of the identified peptides derived from the 18 known proteins in Figure 2B. The delta-mass for IT-triggered data, which was previously mean 171 ppm and standard deviation 226, has now shifted to mean 0.7 ppm and standard deviation 35 ppm. For FT-triggered data, the mean delta-mass was originally 8 with a standard deviation of 86 and is now −0.5 ppm with a standard deviation of 53. These deviations are more consistent with the expected mass accuracy of the instrument.

msPrefix Improves the Confidence of Peptide Identifications. In theory, by improving the delta-mass, the rate of correct identifications should increase. To assess this difference, we explicitly compute the probability of false alarm by searching our data before and after msPrefix with only decoy sequences, neglecting the true positives and determining the distribution of delta-masses. Next, we use these distributions to compute the significance of a peptide identification as a function of delta-mass and generate ROC plots for standard data and msPrefixed data, as shown in Figure 5. Note that more true positives are detected for msPrefixed data than for native data. The total number of true positives is higher after using msPrefix by 10% for FT-triggered and 28% for the IT-triggered data.

msPrefix for Complex Samples. To test the performance of msPrefix for complex samples, we applied it to 6 unfractionated blood plasma samples and 5 fractions from RV1 whole cell lysates that had been fractionated into a total 20 fractions. Figure 6 compares the number of peptides from non decoy sequences identified with a PeptideProphet probability of greater than 0.5[29] with and without using msPrefix. In the case of the whole cell lysates, at its optimal precursor tolerance value of 50 ppm, msPrefix confidently identifies about 200% more peptides than using native data with the same tolerance, and 20% more peptides than searching native data at its best tolerance of 1000 ppm. Specifically, in RV1 lysate, 2295 peptides were identified in native data and 4560 peptides were identified in msPrefixed data. Likewise, in serum, 326 peptides where identified in native data and 1216 peptides were identified in msPrefixed data. Despite the large 50 ppm threshold, 79% of identifications in the RV1 lysate have delta-masses below 5 ppm in msPrefixed data, whereas only 69% of identifications have delta-masses below 5 ppm in native data. As we note in the Discussion, a significant number of MS/MS appear to have multiple peptides in their acquisition window leading to a larger than ideal search threshold. An analysis of the 21% of identifications with delta-masses greater than 5 ppm reveals that 54% have multiple candidate precursor ions.

As an additional test of msPrefix, an early version of the software was given to the Hanash Laboratory to execute on LTQ-FT data from their IPAS protocol.[30] Searches were performed on native files and msPrefixed files using the standard precursor error tolerance for the laboratory of 1.5 Da, and a narrower tolerance of 0.5 Da. The searches were performed against a FASTA sequence file created by appending a reversed UniRef50 to the IPI Human FASTA, yielding a file with 4% human and 96% decoy sequences. In Figure 8, we plot a receiver operator characteristic like (ROC) plot of false positives versus true positives for a set of identifications ranked by PeptideProphet score. An ideal result would score all true positives higher than any decoys and thus yield a horizontal line. With comparable allowed error tolerances, msPrefix halved the rate at which decoy sequences accumulate in the set of peptide identifications. In addition, searches using msPrefix and a 0.5 Da threshold identified more true positives than searches of native data using a larger 1.5 Da tolerance. This result highlights the potential of msPrefix to confer both accuracy and performance gains across data sets generated not just in our laboratory, but in other laboratories as well.

Discussion

Tools like PeptideProphet and ProteinProphet[29,31] have demonstrated the value of postprocessing in computational proteomic workflows. Although a variety of tools exist for postprocessing MS identifications, a smaller number of tools like msPrefix exist that interject postdata acquisition, but preanalysis. Many of these focus on improving computational efficiency by either removing low quality spectra or clustering related spectra.[32,33] We anticipate that msPrefix and other tools at this interface have the potential to impact the performance of downstream tools. In addition, tools at this interface may also enable researchers to explore alternate experimental procedures. For instance, although it has been suggested that any transit of ions from IT-to-FT may lead the ion trap to be more sensitive than the FT and consequently that IT directed triggering may be more sensitive than FT-based triggering, this option was previously not commonly used, as the ion trap derived estimates of precursor mass are significantly less precise than those derived from the FT-preview scan. As

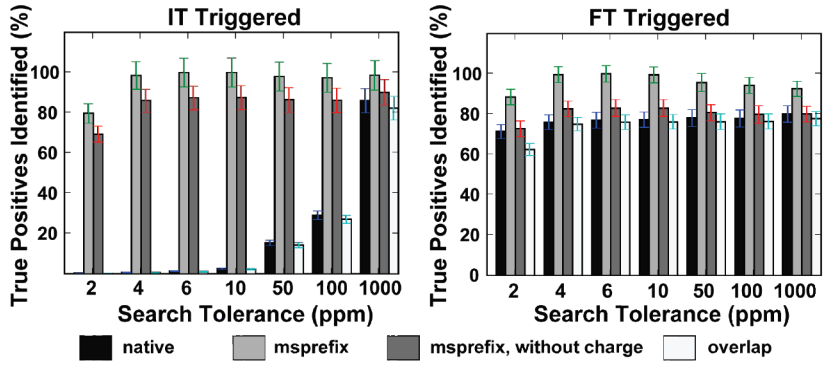

Figure 4. Effect of msPrefix on TANDEM search results for 18 protein mixture. Database searches were performed as described in Experimental Procedures with variations of the allowed precursor tolerance from 2 to 1000 ppm. Native = vendor provided precursor *m*/*z*; msPrefix = msPrefix precursor *m*/*z* and msPrefix determined charge; msPrefixwithout charge = msPrefix precursor *m*/*z* without charge information; overlap = positives found in both native and msPrefix. Following msPrefix, TANDEM is able to assign nearly 100% of positives with a precursor tolerance of 4 ppm.

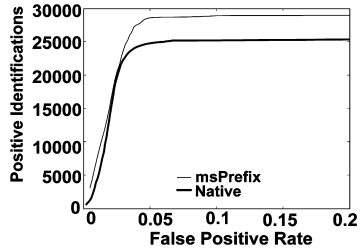

Figure 5. msPrefix greatly increases identification confidence. The rate of false positives was determined as a function of mass differences by using a database of decoys only. The number of positives, i.e., matches to a peptide of the 18 known proteins, is plotted as function of the false positive rate, for the FT-triggered data set. The total number of true positives is 10% higher after using msPrefix.

msPrefix uses the final high-resolution data to re-estimate the precursor mass; either portion of the instrument may drive ion selection.

msPrefix development has highlighted several interesting properties of certain hybrid instruments and precursor ion selection. For instance, as notable in Figure 2A, the most common errors in initial mass estimation are overestimation, rather than under-estimation (likely because of shape of isotope envelopes). Consequently, as illustrated in Figure 3, msPrefix is preferentially decreasing the estimate of the precursor mass. From manual inspection, there appear to be three causes for the typical overestimation of the precursor mass. First, the low resolution of the preview and ion-trap survey scans appears to blur the isotope envelope, which lies dominantly to the right of the monoisotopic peak. Second, for large peptides, the second-isotope peak is higher than the monoisotopic peak leading to the selected precursor to be the second-isotope. Third, the width of the isolation window in combination with overlapping species and dynamic exclusion may lead the isolation window to select a small peak adjacent to a larger peak (that has been excluded). However, because the larger peak is more intense, its ions may contribute more dominantly to the MS/MS spectrum.

One question raised in the analysis was how instrument method software may affect results as data sets may be collected across instruments running different versions of software. In addition, it is possible that older data files may suffer challenges less present in more recently collected data sets. To test this hypothesis, msPrefix analysis was performed on several data sets from our laboratory and other laboratories as described in ref 21. These data sets were each collected on similar FT instruments running different versions of tune. In Figure 7, we demonstrate the relationship between delta-mass and Tune version. While the most recent version confers significant advantages in delta-mass, there is still notable benefit to msPrefix. Furthermore, we note below that msPrefix had a significant positive impact on a human plasma data set collected using the newest vendor software.

msPrefix development also allowed us to test the hypothesis that the IT may have better sensitivity and thus may be able to trigger MS/MS events and identify peptides for which no signal was present in the FT. To test this hypothesis, we examined precursors from an IT triggered human blood plasma experiment to see if there were a significant number of instances where no signal was present in the FT full-scan data. If one restricts analysis to MS/MS events where a peptide was identified with PeptideProphet *p*-values greater than 0.9, there existed a likely precursor within 5 ppm for 98%, of the 5238 identifications. However, for the remaining 2%, msPrefix could not find any possible precursor, suggesting either that these identifications were incorrect or that the sensitivity of the IT was greater than the FT. As the number is close to expected false positive rate of PeptideProphet matches at a value of 0.9, we interpret that differential sensitivity may not significantly impact the rate of identifications. If we expand our search to consider all IT triggered MS/MS, many of which did not result in high confidence identifications, we were unable to find a likely precursor within a 4.6 *m*/*z* unit window for nearly 12% of spectra. Despite the significant number of events triggered by the IT that did not have corresponding FT signals, the majority of these events did not lead to identifications.

In addition to noting differential sensitivity, msPrefix development led us to note challenges in peptide identification, such as how frequently multiple candidate ions are likely to exist within a given MS/MS isolation window; specifically, there are numerous situations in complex mixtures where there are two or more possible sets of peaks corresponding to peptides that may have given rise a single MS/MS spectrum. In the 18-peptide mix experiments, multiple candidate ions were not a significant problem; consequently, searches within a unit window only uncovered a single likely entity in the majority of cases. As there was only a single candidate, it was uniformly correct and the average delta-mass following msPrefix dropped

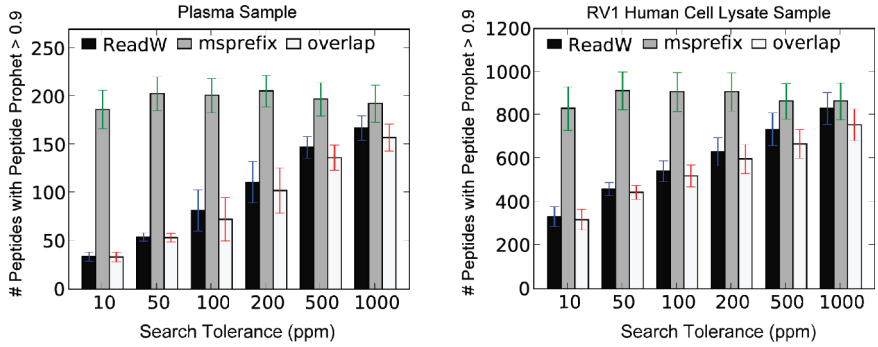

Figure 6. Effect of msPrefix on database search results for whole cell lysates and plasma using TANDEM and PeptideProphet. Using msPrefix allows more confident identifications with a lower search tolerance.

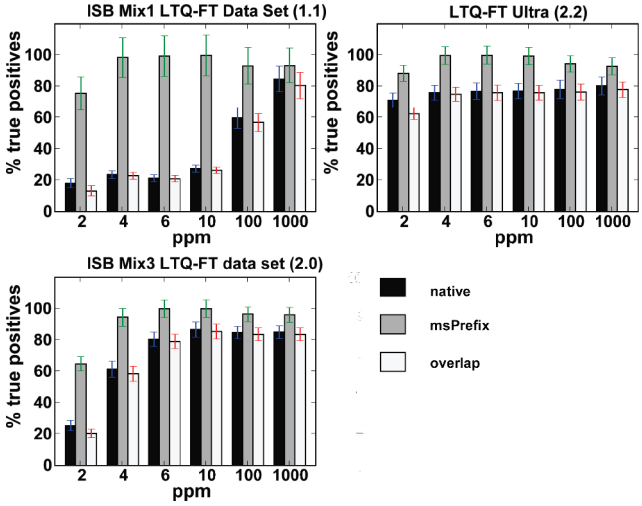

Figure 7. Changes in instrument software impacts precursor accuracy. The 18-mix data was analyzed to see if Xcalibur/Tune version impacted precursor accuracy. Though there was a significant improvement in accuracy in version Tune, version 2.2 msPrefix still improves identification performance.

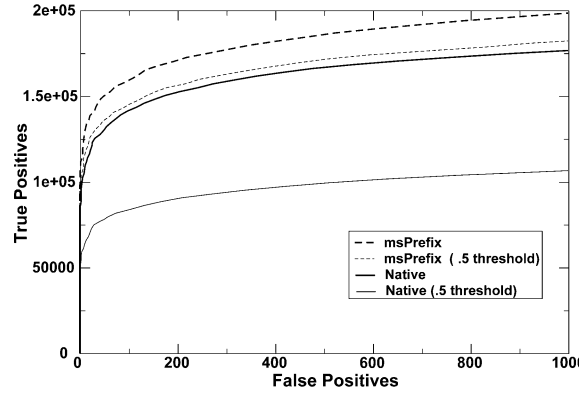

Figure 8. External validation of msPrefix reveals significant gains in performance. A data set of 96 LCMS runs of fractionated human plasma was searched with TANDEM using the standard precursor error tolerance for the laboratory of 1.5 Da, and a narrower tolerance of 0.5 Da both on native files and msPrefixed files. msPrefixed data resulted in more true positive identifications per false positive identification with a smaller search tolerance than native data.

from 8 to −0.5 ppm. However, in complex mixture experiments, multiple potential precursor ions were found in 40 and 57% of the selection windows in the unfractionated plasma and fractionated lysate experiments, respectively, leading to a larger average delta-mass of 5 ppm. For reference, an average of 194 proteins were identified per fraction in the lysate experiment. In some cases, there were as many as 12 likely precursor species in the msPrefix search window. In Figure 9A we plot the number of possible precursor masses for each acquired MS/MS of an LCMS experiment on an unfractionated plasma sample.

The implications of having multiple potential precursor masses for a single MS/MS spectrum suggest a number of possible future enhancements of search programs, such as TANDEM. For example, it is possible that additional identification performance could be gained by applying TANDEM with small delta-mass thresholds and multiple precursors instead of a single precursor and a large threshold. Alternately, knowledge that several, high intensity, precursor peptides may be present in a single MS/MS spectrum may help determine which spectra may need to be explained as the superposition of multiple fragmented species. To test the impact of this workflow, we adapted msPrefix to write multiple precursor tags to mzXML and also adapted TANDEM to consider multiple input precursors. As noted in Figure 9B, there appears to be a marginal benefit to this workflow. Also note in Figure 9, though a majority of acquired MS/MS have more than one likely precursor, a majority of the identifications are derived from MS/MS with only a single precursor. This is not entirely surprising, as TANDEM does not presently attempt to account for multiple peptide fragmentations contributing to a given MS/MS spectrum.

In addition to noting the frequency of multiple ions within our search window, we also note how critical search thresholds are on algorithm performance. For example, consider the overlap bar in Figures 4 and 6. Equal height of the overlap bar and the native bar would indicate that no matches are lost when using msPrefix. However, some matches are always lost with msPrefix. For searches with complex samples, the difference between the overlap and native can be as much as 10%. Examination of this phenomena reveals, as noted above, that when presented with several possible precursor options msPrefix sometimes selects an incorrect precursor. This incorrect precursor may be slightly further away in mass from the matching peptide than the initial estimate. Consequently, if the matching peptide was at the boundary of a search threshold, even a slight movement can result in the matching peptide being beyond the search threshold and thus remaining unidentified.

Another aspect of MS/MS search programs highlighted by msPrefix is the existence of optimum values for the precursor

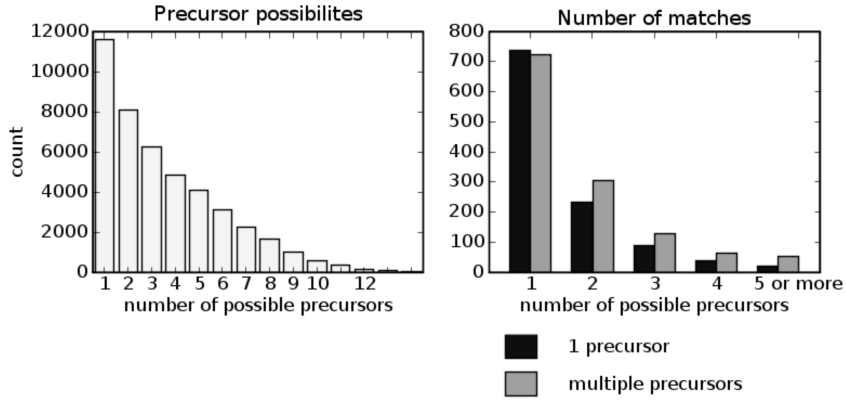

Figure 9. Number of possible precursor masses and resulting identifications on a human serum sample. msPrefix is able to detect if multiple parent ions are within our isolation window. Note that only about 1/3 of the acquired MS/MS spectra are derived from a single precursor ion, whereas 2/3 of the peptide identifications originate from spectra with a single precursor. Running msPrefix in multiple precursor mode is able to increase the number of peptides identified in spectra derived from multiple precursors.

search tolerance value. For instance, in Figures 4 and 6, the number of true positives increases to a point and then begins to decrease. When set too low, identifications are lost because of poor precursor precision. However, higher tolerances allow more decoy or false peptides to be selected for correlation, thus, increasing the number of false matches found by chance. Given the accuracy of the instrument, it should be possible to use tolerance values of 2−5 ppm. However, the multiple ion problem noted above makes this presently impractical for complex mixtures. As noted by Elias and Gygi, there is significant value in accurate estimation of precursor mass, as peptide masses occupy a limited number of locations along the mass axes, and the number of possible decoy precursor masses depends on the mass value and the precursor tolerance.[34] At settings less than 1 ppm, it would be possible to exclude certain mass values as being in unoccupied regions and therefore unlikely. Though msPrefix makes it possible to significantly decrease search tolerances, it is not yet possible to reduce tolerances for complex mixture searches to 1 ppm or lower. However, additional postacquisition, precomputational, or co-computational identification tools that better calibrate spectra[35] or better estimate which precursor masses best explain a tandem spectrum may have significant impact.

In addition to having an impact on qualitative proteomics studies, msPrefix may also impact quantitative studies. For instance, results obtained by XPRESS[36] can also be positively affected by msPrefix. Although XPRESS uses the calculated precursor m/z for the matched peptide when calculating peptide abundance, msPrefix increases the number of significant matches and therefore also the number of peptides and proteins with quantitation values. For example, in the RV1 data set, 763 proteins were quantifiable using native data, whereas 949 proteins were quantifiable following msPrefix analysis. Furthermore, the number of peptides identified for each protein may significantly impact quantification methods like spectral counting.

Conclusions

As noted above, trigger m/z typically deviate significantly more than 1 ppm, on average, data directed identifications using hybrid instruments. When msPrefix is used, one is able to decrease this to 1 ppm, thus, allowing smaller search tolerances and more confident identifications. msPrefix is available open source at http://www.sfcap.cshs.org/download.shtml and has been added to the ProteoWizard msConvert tool available at http://proteowizard.sourceforge.net.

Acknowledgment. The authors thank Kristen Lee and Sachin Gandhi for contributing towards data collection for this paper and Matthew Chambers for helpful input on the manuscript and the Hanash laboratory for additional benchmarking of msPrefix. This work has been supported by the NCI Center for Cancer Nanotechnology Excellence Focused on Therapy Response (CCNE)-NIH U54, the Wunderkinder Foundation and the Bennioff Foundation.

References

(1) Wright, M. E.; Eng, J.; Sherman, J.; Hockenbery, D. M.; Nelson, P. S.; Galitski, T.; Aebersold, R. Identification of androgen-coregulated protein networks from the microsomes of human prostate cancer cells. *GenomeBiology* **2003**, *5* (1), R4.

(2) Guina, T.; Purvine, S. O.; Yi, E. C.; Eng, J.; Goodlett, D. R.; Aebersold, R.; Miller, S. I. Quantitative proteomic analysis indicates increased synthesis of a quinolone by Pseudomonas aeruginosa isolates from cystic fibrosis airways. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (5), 2771–6.

(3) Bouwmeester, T.; Bauch, A.; Ruffner, H.; Angrand, P. O.; Bergamini, G.; Croughton, K.; Cruciat, C.; Eberhard, D.; Gagneur, J.; Ghidelli, S.; Hopf, C.; Huhse, B.; Mangano, R.; Michon, A. M.; Schirle, M.; Schlegl, J.; Schwab, M.; Stein, M. A.; Bauer, A.; Casari, G.; Drewes, G.; Gavin, A. C.; Jackson, D. B.; Joberty, G.; Neubauer, G.; Rick, J.; Kuster, B.; Superti-Furga, G. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat. Cell Biol.* **2004**, *6* (2), 97–105.

(4) Everley, P. A.; Krijgsveld, J.; Zetter, B. R.; Gygi, S. P. Quantitative cancer proteomics: stable isotope labeling with amino acids in cell culture (SILAC) as a tool for prostate cancer research. *Mol. Cell. Proteomics* **2004**, *3* (7), 729–35.

(5) Durr, E.; Yu, J.; Krasinska, K. M.; Carver, L. A.; Yates, J. R.; Testa, J. E.; Oh, P.; Schnitzer, J. E. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **2004**, *22* (8), 985–92.

(6) Ranish, J. A.; Hahn, S.; Lu, Y.; Yi, E. C.; Li, X. J.; Eng, J.; Aebersold, R. Identification of TFB5, a new component of general transcription and DNA repair factor IIH. *Nat. Genet.* **2004**, *36* (7), 707–13.

(7) Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **2003**, *33* (3), 349–55.

(8) Blagoev, B.; Kratchmarova, I.; Ong, S. E.; Nielsen, M.; Foster, L. J.; Mann, M. A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **2003**, *21* (3), 315–8.

(9) Aebersold, R.; Anderson, L.; Caprioli, R.; Druker, B.; Hartwell, L.; Smith, R. Perspective: a program to improve protein biomarker discovery for cancer. *J. Proteome Res.* **2005**, *4* (4), 1104–9.

(10) Alaiya, A.; Al-Mohanna, M.; Linder, S. Clinical cancer proteomics: promises and pitfalls. *J. Proteome Res.* **2005**, *4* (4), 1213–22.

(11) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198–207.
(12) Patterson, S. D.; Aebersold, R. H. Proteomics: the first decade and beyond. *Nat. Genet.* **2003**, *33*, 311–23.
(13) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–7.
(14) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
(15) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–67.
(16) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6* (3−4), 327–42.
(17) Frank, A.; Tanner, S.; Bafna, V.; Pevzner, P. Peptide sequence tags for fast database search in mass-spectrometry. *J. Proteome Res.* **2005**, *4* (4), 1287–95.
(18) Shadforth, I.; Crowther, D.; Bessant, C. Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics* **2005**, *5* (16), 4082–95.
(19) Haas, W.; Faherty, B. K.; Gerber, S. A.; Elias, J. E.; Beausoleil, S. A.; Bakalarski, C. E.; Li, X.; Villen, J.; Gygi, S. P. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell. Proteomics* **2006**, *5* (7), 1326–37.
(20) Peterman, S. M.; Dufresne, C. P.; Horning, S. The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing. *J. Biomol. Tech.* **2005**, *16* (2), 112–24.
(21) Klimek, J.; Eddes, J. S.; Hohmann, L.; Jackson, J.; Peterson, A.; Letarte, S.; Gafken, P. R.; Katz, J. E.; Mallick, P.; Lee, H.; Schmidt, A.; Ossola, R.; Eng, J. K.; Aebersold, R.; Martin, D. B. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **2008**, *7* (1), 96–103.
(22) Faca, V.; Pitteri, S. J.; Newcomb, L.; Glukhova, V.; Phanstiel, D.; Krasnoselsky, A.; Zhang, Q.; Struthers, J.; Wang, H.; Eng, J.; Fitzgibbon, M.; McIntosh, M.; Hanash, S. Contribution of protein fractionation to depth of analysis of the serum and plasma proteomes. *J. Proteome Res.* **2007**, *6* (9), 3558–65.
(23) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (4), 229.
(24) Keller, A.; Eng, J.; Zhang, N.; Li, X. J.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, 0017.
(25) Rauch, A.; Bellew, M.; Eng, J.; Fitzgibbon, M.; Holzman, T.; Hussey, P.; Igra, M.; Maclean, B.; Lin, C. W.; Detter, A.; Fang, R.; Faca, V.; Gafken, P.; Zhang, H.; Whiteaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M. W. Computational Proteomics Analysis System (CPAS): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *J. Proteome Res.* **2006**, *5* (1), 112–21.
(26) MacLean, B.; Eng, J. K.; Beavis, R. C.; McIntosh, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **2006**, *22* (22), 2830–2.
(27) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**, *5* (13), 3226–45.
(28) Anderson, N. L.; Polanski, M.; Pieper, R.; Gatlin, T.; Tirumalai, R. S.; Conrads, T. P.; Veenstra, T. D.; Adkins, J. N.; Pounds, J. G.; Fagan, R.; Lobley, A. The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **2004**, *3* (4), 311–26.
(29) 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* (20), 5383–92.
(30) Wang, H.; Hanash, S. Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry. *Mass Spectrom. Rev.* **2005**, *24* (3), 413–26.
(31) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–58.
(32) Tabb, D. L.; Thompson, M. R.; Khalsa-Moyers, G.; VerBerkmoes, N. C.; McDonald, W. H. MS2Grouper: group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (8), 1250–61.
(33) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **2007**, *4* (10), 787–97.
(34) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–14.
(35) Grothe, R. K. D.; Luthy, R.; Katz, J.; Agus, D.; Mallick, P. A physical model of peak shape for FT-ICR Mass Spectrometry. *Am. Soc. Mass Spec.*, in preparation.
(36) Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **2001**, *19* (10), 946–51.

PR800307M