research articles Journal of proteome research

A Platform for Accurate Mass and Time Analyses of Mass Spectrometry Data

Damon May, Matt Fitzgibbon, Yan Liu, Ted Holzman, Jimmy Eng, C. J. Kemp, Jeff Whiteaker, Amanda Paulovich, and Martin McIntosh*

Fred Hutchinson Cancer Research Center, Seattle, Washington

Received March 15, 2007

We describe an integrated suite of algorithms and software for general accurate mass and time (AMT) tagging data analysis of mass spectrometry data. The AMT approach combines identifications from liquid chromatography (LC) tandem mass spectrometry (MS/MS) data with peptide accurate mass and retention time locations from high-resolution LC−MS data. Our workflow includes the traditional AMT approach, in which MS/MS identifications are located in external databases, as well as methods based on more recent hybrid instruments such as the LTQ-FT or Orbitrap, where MS/MS identifications are embedded with the MS data. We demonstrate our AMT workflow's utility for general data synthesis by combining data from two dissimilar biospecimens. Specifically, we demonstrate its use relevant to serum biomarker discovery by identifying which peptides sequenced by MS/MS analysis of tumor tissue may also be present in the plasma of tumor-bearing and control mice. The analysis workflow, referred to as msInspect/AMT, extends and combines existing open-source platforms for LC−MS/MS (CPAS) and LC−MS (msInspect) data analysis and is available in an unrestricted open-source distribution.

Keywords: CPAS • AMT • msInspect • LC • MS • MS/MS

Introduction

Comparing the protein constituents of two clinically relevant complex protein mixtures (such as plasma) traditionally makes use of either single liquid chromatography mass spectrometry (LC−MS) or tandem MS (LC−MS/MS) data analysis workflows. Each of these more traditional approaches has its own advantages and disadvantages. We present an open source and freely available analysis workflow for combining LC−MS and LC−MS/MS data to take advantage of the strengths of both approaches.

Quantitative comparison of protein mixtures based on LC−MS/MS data has the advantage of providing each peptide's amino acid sequence. However, whether using isotopic labeling or label-free methods, MS/MS-based methods can quantitate only a small subsample of the peptides which may be identified by LC−MS. Unfortunately, because the lower-abundance peptides are less likely to be identified in repeated experiments, the poor coverage and reproducibility of LC−MS/MS approaches limit their utility for comparing large numbers of complex mixtures such as plasma, which may be needed to identify clinically relevant findings.[1]

LC−MS data provides comparatively greater surveys of the constituent peptides in a complex mixture than LC−MS/MS approaches and has gained popularity recently due in part to advancements in high-resolution instrumentation (LTQ-FT, Orbitrap, and hybrid time-of-flight [TOF] platforms). Multiple samples may be compared using LC−MS data with recently developed advanced bioinformatics approaches (msInspect,[2] MapQuant,[3] etc.) to locate the mass and retention times of peptides in an experiment and associate the peptide locations across multiple experiments. These LC−MS approaches have comparatively greater capacity to survey the peptide constituents than do LC−MS/MS-based approaches, but peptide mass and retention times cannot on their own determine the chemical composition of a peptide.

One experimental approach to cope with the drawbacks of LC−MS data is to sequence the putative peptide in subsequent MS/MS interrogations that target the peptide's location.[4] A second strategy is to use computational approaches to synthesize related LC−MS and LC−MS/MS data and infer the amino acid sequence indirectly. The ability to infer LC−MS peptide sequence indirectly stems from the fact that both peptide mass and retention time are related to peptide amino acid composition. Even though the peptide location alone cannot determine the amino acid sequence with sufficient accuracy when interrogating the entire human proteome database, it may be sufficient when interrogating a smaller protein database, such as a database generated by interrogating related biosamples. Smith et al.,[5,6] who had early access to high-resolution LC−MS equipment, pioneered the approach referred to as the accurate mass and time (AMT) method. Their typical AMT approach combines LC−MS and LC−MS/MS data, acquired on separate instrumentation, by first generating a comprehensive organ-specific (e.g., serum or plasma) or organism-specific (e.g., bacteria) peptide sequence library using

* To whom correspondence should be addressed. E-mail: mmcintos@fhcrc.org.

10.1021/pr070146y CCC: $37.00

Published on Web 06/09/2007

extensive fractionation and interrogation by LC–MS/MS. This AMT database records each peptide's sequence (and so its mass) and observed retention time. Subsequent higher throughput interrogations with LC–MS then provide the quantitation and coverage needed to identify differences. The located peptides are matched to the AMT database to ascertain putative sequence assignment. Others have applied AMT techniques to instruments with lower mass accuracy.[7]

More recently, novel commercial instrumentation such as the LTQ-FT and Orbitrap, and the Q-TOF, have allowed simultaneous acquisition of both high-resolution LC–MS and LC–MS/MS data in a single experiment. Jaffe et al. recently described a workflow incorporated in PEPPeR (Platform for Experimental Proteomic Pattern Recognition)[8] that combines peptide locations with amino acid sequences acquired from a series of related runs with these platforms. This workflow is related to the classic AMT approach in that it combines LC–MS and LC–MS/MS data, but because it takes advantage of embedded MS/MS data, it is useful for combining data across a series of related experiments and does not involve the use of externally derived data sources.

The approach described by msInspect/AMT provides a general framework for combining LC–MS and LC–MS/MS data using either externally derived LC–MS/MS sequence databases (the classic AMT approach) or internal databases of LC–MS/MS data embedded with the LC–MS data (as with PEPPeR), or a combination of both. Both approaches might be needed for applications in which more general interrogation or reuse of data is warranted, such as interrogation of one's own high-resolution LC–MS data for peptides reported by published literature or constructed from data repositories, e.g., PRIDE.[9] When attempting to identify serum cancer biomarkers, for example, one might wish to determine which of many discriminating peptides may also be present in cancer tissue, since those may be more likely to have biological relevance or specificity.

To demonstrate the workflow we interrogate peptide locations identified in LC–MS experiments on normal and tumor-bearing mouse plasma, using an externally derived AMT database generated from LC–MS/MS interrogations of normal and tumor tissue from the same mouse model. We demonstrate that our AMT approach can identify which putative tissue-derived proteins are present in, and perhaps differentially abundant in, the plasma as well.

The algorithms and methods described here extend and combine two previously developed platforms for mining LC–MS[2] and MS/MS[10] data and also incorporate an open-source retention time prediction algorithm previously described.[11] All components described are made available as open-source platforms. Although our workflow builds on these specific previously described methods, we make extensive use of standardized, open file formats to make the AMT workflow usable with other fingerprinting or search engine results. For example, our software can construct an AMT database using data from any search engine whose results are represented in the pepXML file format[12] and use it to interrogate peptide locations encoded in a simple text file format. We also implement a retention time normalization procedure to allow the use of different retention time prediction algorithms.

Experimental Procedures

The msInspect/AMT workflow contains three logical components, shown in Figure 1 and summarized here. Detailed descriptions of the algorithms and components are also contained in PROCEDURES, parts I, II, and III below.

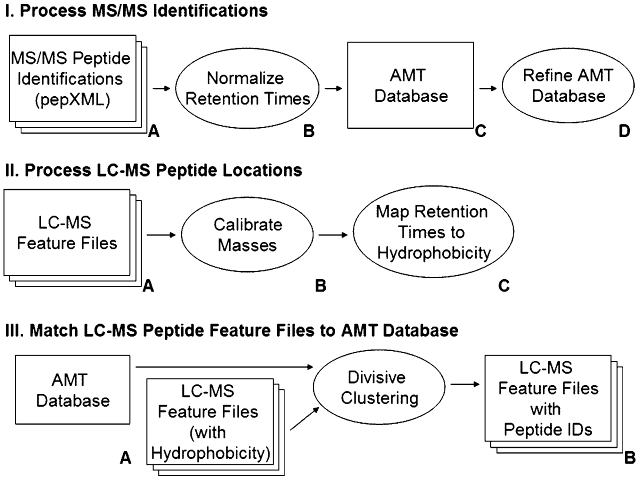
I. Process MS/MS Identifications
MS/MS Peptide Identifications (pepXML)
Normalize Retention Times
AMT Database
Refine AMT Database
A
B
C
D
II. Process LC-MS Peptide Locations
LC-MS Feature Files
Calibrate Masses
Map Retention Times to Hydrophobicity
A
B
C
III. Match LC-MS Peptide Feature Files to AMT Database
AMT Database
LC-MS Feature Files (with Hydrophobicity)
Divisive Clustering
LC-MS Feature Files with Peptide IDs
A
B

Figure 1. Three components of the AMT workflow. (Part I) Retention times from a sequence of MS/MS experiments normalized and combined/summarized in an AMT database. (Part II) Peptide locations from LC–MS data are filtered, mass calibration and filtering are applied, and retention times are normalized to the hydrophobicity scale. (Part III) Peptide locations (masses and normalized retention times) from LC–MS peptides are matched to the AMT database to assign peptide IDs. Following AMT assignment to any number of LC–MS peptide feature files, these feature files may be aligned into a peptide array, using linear or nonlinear chromatographic alignment algorithms.

Part I: AMT Database Generation. Beginning with a series of LC–MS/MS search results encoded in a pepXML file (A), we apply a normalization procedure to each file separately to transform the retention times to a normalized "hydrophobicity" scale independent of LC gradient length (B). All peptides identified in these experiments are combined into a consensus AMT database and summarized by their median normalized retention times and their standard deviations. The retention time normalization approach is performed on a per-experiment basis so that one may combine results from multiple different LC configurations into a single AMT database (C). Next, this AMT database is filtered to remove suspect peptide observations (D).

Part II: LC–MS Fingerprinting Normalization. We begin the LC–MS analysis with a series of peptide feature files (A): files encoding the masses, retention times, and intensities of all the peptides identified by software such as msInspect. Optional procedures may be applied to detect and correct potential mass calibration error in the LC–MS instrumentation using an approach described by Wolski et al.[13] (B). We use this same approach for further filtering to remove those peptide locations which are likely incorrectly identified (see Results for an example). Retention times are normalized to the same scale as the AMT database with a two-step procedure using quantile regression approaches (C). The resulting data represent a set of peptide locations with normalized retention times on the same scale as the AMT database.

Part III: Matching Normalized LC–MS Locations to the AMT Database and Estimating False Assignment Rate. The AMT database and normalized peptide feature files (A), now on a common mass and time scale, are then matched using two-dimensional divisive clustering approaches previously described.[2] The result is a set of peptide LC–MS feature files with amino acid sequences assigned to each matching peptide location (B). We then construct a decoy AMT database (as suggested by Smith et al.[5]) and perform the same matching procedures against the decoy database to approximate assignment error. Multiple feature files with peptide assignments may then be aligned using linear or nonlinear chromotographic alignment approaches into a single data set suitable for further analysis.

Platform Architecture and Integration with Other Software. The AMT workflow is built upon the msInspect suite of algorithms, which may also be used to generate the LC−MS peptide feature files. However, several efforts have been made to maximize the flexibilty of the msInspect software and to allow its use along with other platforms. For example, the use of search results encoded in the standardized pepXML format (supported by Tandem, Mascot, Sequest) allows the operator a choice of search engines. Also, the simple text-based format of the LC−MS feature files used by msInspect/AMT makes it easy to convert the results of other LC−MS analysis tools to the msInspect format. Moreover, we have implemented a standardization procedure (described in PROCEDURES, Part I), which allows replacement of Krokhin's open source algorithm[11] with any other retention time prediction algorithm—these algorithms are a key component of the AMT workflow, as they are used to normalize the retention times across experiments.

Integration with CPAS using caBIG Interface. The msInspect/AMT pipeline contains optional components allowing for, but not requiring, tight integration between the Computational Proteomics Analysis System[10] and msInspect.[2] Specifically, msInspect may be used to browse a CPAS site and select a set of LC−MS/MS data files for use in building the AMT database. This client application uses the caBIG client API provided by CPAS, which conforms to caBIG requirements. More detail is provided in the Supporting Information.

Software and Data Availability and Distribution. Software components of msInspect/AMT are distributed with the msInspect open-source distribution at http://proteomics.fhcrc.org. The majority of msInspect/AMT is written in Java, with some computational framework provided in the freely available R language.[14] CPAS is distributed at http://cpas.fhcrc.org. Both CPAS and msInspect are distributed under the Apache 2.0 open-source license. Information on Krokhin's hydrophobicity calculator can be found at http://hs2.proteome.ca/SSRCalc/SSRCalcHelp.htm. The hydrophobicity calculation algorithm is distributed under the Artistic open-source license.

Data presented in this manuscript, along with scripts and code to analyze that data to produce all results shown in this manuscript, are available by visiting www.proteomics.fhcrc.org/CPAS, navigating to the "published experiments" folder and selecting "AMT Manuscript". Available data include a sampling of the 84 MS/MS experiments used to construct the AMT database, the resulting AMT database, and the individual peptide locations from each of the LC−MS experiments which are disscussed below.

Procedures

Part I: Normalization of LC−MS/MS Experiments and Generation of AMT Databases. First, each LC−MS/MS experiment is processed individually using a multi-stage process that places the retention times for each experiment on a common scale. For each high-quality peptide identification in the experiment we use Krohkin's open source algorithm[11] to compute a predicted hydrophobicity value (H) from the amino acid sequence. H is a quantity that is linearly associated with actual retention time within an experiment. Using only the high-quality peptides (i.e., PeptideProphet score ≥ 0.95), we estimate a linear model associating H with retention time (RT) for that experiment: $\hat{\mathrm{H}}_i = \hat{\alpha} + \hat{\beta} \times RT_i$. RT is in clock-time or any scale linearly associated with clock-time (such as scan number). The linear model is estimated using robust regression techniques based on automated outlier detection to be robust with respect to peptide identification errors. We refer to the fitted value $\hat{\mathrm{H}}$ as the "normalized retention time" for that peptide. A summary schematic of the robust regression technique is shown in Supporting Information, and a demonstration of how this procedure normalizes retention times across experiments is discussed in the Results section and in Figure 2.

Many peptides are observed multiple times within each experiment and across multiple experiments, and we process the data to construct a consensus retention time for each peptide. Peptides observed multiple times in a single experiment are summarized with the time of the first occurrence. To combine peptides across multiple experiments we summarize with the median normalized retention time ($\hat{\mathrm{H}}$).

Before finalizing the database, however, we remove outliers to reduce the number of erroneous or aberrational peptide observations kept in the database. We do this in two ways. (1) We calculate the standard deviation of the amount by which observed $\hat{\mathrm{H}}$ for each peptide in the database differs from the peptide's algorithm-predicted values; any peptide observed only once with an observed $\hat{\mathrm{H}}$ that differs from its predicted H value by more than 2 standard deviations is removed. (2) For peptides with 3 or more observations, any single observation with an $\hat{\mathrm{H}}$ value more than 3 standard deviations from the median $\hat{\mathrm{H}}$ for that peptide is removed as well, and the peptide's median $\hat{\mathrm{H}}$ is recomputed. The resulting "refined" AMT database is stored in an XML document (see Supporting Information for details and schema). These refinement steps typically remove, respectively, roughly 2−3% of single-observation peptide entries and outlier observations from less than 1% of peptide entries. Both of these steps can reduce the number of false positive AMT matches quantifiably without a large reduction in the number of true positive matches.

To foster use of novel retention time prediction algorithms in the normalization procedure above, we standardize our hydrophobicity scale so that any prediction algorithm whose output is linearly associated with actual peptide retention time may replace Krohkin's algorithm. To standardize any such algorithm's output, we re-center and re-scale the algorithm's results so that when evaluating the tryptic digest of a specific IPI sequence database (July 13, 2006: ipi.HUMAN.fasta.20060713), the calculations average 0 and have a standard deviation of 1. This approach may be used generally to make hydrophobicity calculators less dependent on specific implementation and also to increase algorithm interpretability by allowing it to be compared with any other calculator similarly re-scaled.

Part II: Normalizing and Filtering Peptide Identifications from LC−MS Peptide Feature Files. Taking as input a peptide feature file containing the locations and intensities of peptides found in high-resolution LC−MS data, we normalize the mass and retention times to the same scale as the AMT database. To ensure that masses are reported as accurately as possible in the presence of calibration error, we first recalibrate masses using a procedure that is based on an approach described by Wolski.[13] This approach recognizes that, when calibrated correctly, distances between peptides will form distinct, equidistantly spaced clusters due to their similar elemental composition. This cluster spacing is due to the fact that peptide masses are largely composed of nucleons which weigh roughly, but not exactly, one Dalton (not exactly, because of the effect of the mass defect, which varies from element to element). We determine the distances between the centers of these theoretical clusters (the "wavelength") based on the average elemental distribution in peptides. A linear model associating mass deviation from theoretical wavelength (*D*) versus total mass separation (*M*) is estimated. If a significant nonzero slope or intercept is identified, the masses are scaled to remove the trend.

After this calibration, we also use these expected mass distributions to filter the observed peptides to remove incorrectly identified features with impossible (or highly unlikely) masses, a procedure described by Wolski[13] and related to that described by Peining et al.[15] Their work showed that the vast majority of peptide masses occur within a 200 ppm window (a theoretical window not related to instrumentation mass accuracy) surrounding the centers of the theoretical mass clusters. We remove all peptides whose mass falls outside of that window, because the probability that those identifications are correct is vanishingly small. A demonstration is shown below in the Results section, in Figure 3.

To normalize retention times between a peptide feature file and the AMT database, two procedures are available. When LC−MS data contain embedded MS/MS identifications (e.g., with LTQ-FT or Orbitrap instruments and the workflow suggested by PEPPeR), we apply the same normalization procedures described above for use in creating the AMT database to the embedded LC−MS/MS identifications to derive the correct normalization for the run. When such identifications are not available (e.g., the data described in the Data and Results sections of this paper, and the workflow suggested by Smith et al.), we first perform a crude match between the peptide feature locations and the AMT database using only the peptide masses, with a very tight mass tolerance selected to match instrumentation limits. The set of crude matches will contain a mixture of correct and incorrect identifications, and we estimate the underlying latent linear model associating the retention time (RT) of the peptide identifications to hydrophobicity ($\hat{H}$) using quantile regression approaches. A demonstration is shown in the Results section (see Figure 4 below). Prior to crude matching, we modify the database to anticipate static modifications by adjusting the mass of each peptide containing the modified amino acids. For anticipated variable modifications, we create a separate AMT entry for each possible combination of modified and unmodified residues.

Part III: Matching Normalized Peptides to the AMT Database and Approximating False Assignment Rate. With both the AMT database and peptide feature files on the same normalized scale, we apply a two-dimensional divisive clustering algorithm to assign peptide sequences from AMT database entries to their associated peptides in an LC−MS feature file. The basic clustering approach is based on the procedures currently used by msInspect to chromatographically align multiple peptide feature files into a single array.[2] The cluster diameters (tolerances) for the mass and $\hat{H}$ dimensions may be selected by the user. These diameters will have a great effect on the sensitivity and specificity of the matching. The mass diameter should depend on the mass accuracy of the instrumentation. The appropriate hydrophobicity diameter may be determined in a number of ways. When LC−MS data contain MS/MS identifications, performing clustering between the AMT database and the MS/MS peptide identifications to establish the desired true-positive and false-positive rates may refine the H cluster diameter. Otherwise (as described in Results, subsection "Matching to the AMT Database and Calculation of False Assignment Rate"), the appropriate cluster diameter in the H dimension may be determined based on the desired false assignment rate by repeated matching using a decoy AMT database.

We implement automatic procedures for approximating the false AMT assignment rate by producing a decoy AMT database (see Results for example) and comparing the peptide assignment rates between the actual and decoy databases. Following recommendations by Smith et al.,[5] the decoy AMT database is created by adding 11 Daltons to the peptide masses of the target AMT database, chosen because no post-translational modification is consistent with this mass shift. This gives us a decoy database of the same size and distribution as the original database, all of whose masses are incorrect but whose masses and hydrophobicities fall into a reasonable range for our data. Some of these shifted masses will coincide with actual masses from the database by chance; the decoy database represents a distribution of peptides throughout the detection range of the instrumentation, rather than only false matches. We then reapply the identical retention time normalization and clustering procedures used when matching to the target AMT database. The false assignment rate is defined as number of peptide assignments using the decoy database, divided by the number of peptide assignments using the target database plus the number of assignments to the decoy database.

Data

We interrogated LC−MS peptides identified in mouse plasma using an AMT database generated from a different biomaterial (normal and tumor mouse tissue). We constructed an AMT database from 84 total LC−MS/MS interrogations of murine mammary tissue. This tissue-derived AMT database was then used to infer peptide sequences for a large number of peptide features extracted from high-resolution LC−MS surveys of mouse plasma samples. Our example data was chosen to challenge all msInspect/AMT components: the LC−MS profiles are generated using instrumentation having lower resolution and mass accuracy than the most sophisticated commerical instruments now available. Moreover, the LC−MS interrogations require substantial mass recalibration.

Tissue Interrogations and MS/MS Data Processing. Detailed descriptions of the tissue processing and collection are given elsewhere[16] (manuscript in preparation). Briefly, pooled tissue lysates from 5 control and 5 tumor-bearing mice were denatured, reduced, alkylated, and digested by trypsin. A total of 42 interrogations each of normal and tumor mammary tissue lysate (84 interrogations total) were performed using an Agilent 1100 nanoHPLC in conjunction with an LTQ (Thermo). Separations were performed on a RP18 capillary monolithic column (Chromolith Caprod, Merck, Germany) by developing a linear gradient of 2−40% organic over 120 min. Raw data were converted to mzXML files and searched with X!Tandem[17] configured with the k-score scoring algorithm[12] and assigned probability values by PeptideProphet.[18]

Table 1. Summary of the LC−MS/MS Peptide Identifications from the Tissue Interrogations for Case and Control Experiments and All Experiments Together[a]

	normal	tumor	combined
runs	42	42	84
average peptides per run	838	1044	937
minimum peptides in a run	674	533	533
maximum peptides in a run	1140	1492	1492
total unique peptides	3405	5195	6412

[a] Peptides were selected based solely on PeptideProphet score (≥0.95).

Table 2. Peptide Locations Identified by LC−MS Interrogation of the Mouse Case and Control Plasma Including the Number of Peptide Locations in Each File and the Number of Peptide Features That Align Within the Cases and Within the Controls

	normal			tumor		
	all	filtered by mass	AMT matched	all	filtered by mass	AMT matched
3 pmol	4414	4282	1079	5898	5760	1276
4 pmol	5524	5370	1195	6839	6694	1415
5 pmol	6278	6098	1340	7638	7463	1593
Aligned across all 3		3217			3842	

For all analyses below, we retained only peptide identifications with PeptideProphet probability exceeding 0.95. We identified a total of 6412 unique peptides across all experiments. A greater number of unique peptide sequences were identified in tumor tissue (overall 5195, 1044 per run average) than in normal tissue (overall 3405, average 838) (see Table 1).

Serum Interrogations using LC−MS. A pool of healthy mouse plasma and a pool of plasma from tumor-bearing mice were depleted of the three most abundant proteins using a mouse plasma MARS column. The protein concentration was determined using Bradford QuickStart Assay (Bio-Rad Laboratories). Samples were denatured and reduced with 60% methanol and 10 mM dithiothreitol (DTT) at 60 °C for 1 h and alkylated with 50 mM iodoacetamide (IAM) at room temperature in the dark for 30 min. Ammonium bicarbonate (50 mM) was added to achieve a final methanol concentration of 20% and the samples were digested with Trypsin Gold (Promega, Madison, WI) at a protein to enzyme ratio of 50:1 (w/w) at 37 °C for 6 h. The samples were dried in a SpeedVac and resuspended in 50 mM ammonium bicarbonate prior to LC−MS analysis. Each pool was interrogated using an LCT-Premier (Waters) ESI-TOF three times, once each at three different protein loadings: 5 pmol, 4 pmol, and 3 pmol. Mass spectra were acquired over the range m/z 400−1600 every 1.0 s with a 0.05 s interscan delay time. Data were converted to mzXML format[19] and submitted for processing using msInspect, which produced one peptide feature file for each interrogation. Only peptide features with two or more identified isotopes and KL quality scores[2] less than 1 were considered. Table 2 shows the number of unique peptides identified at each concentration before and after mass filtering of peptide features (as described in Results, subsection "Calibrating and Filtering LC−MS Peptide Mass"). We then aligned the peptides from all feature files chromatographically. The final row of Table 2 indicates the number of confident peptides identified in all three of the cancer and normal plasma runs. Supporting Information contains the aligned peptide array and a script containing all commands and parameters used to produce the aligned peptide array.

For each loading concentration a larger number of peptides are identified in plasma from the tumor bearing mice (at 5 pmol 7638 before mass error filtering and 7436 after mass error filtering) than the normal plasma (at 5 pmol 6278 before mass error filtering and 6098 after mass error filtering). This is consistent with the LC−MS/MS interrogations. The consistency is maintained when examining only those peptides that, via chromatographic alignment, are found to exist across all three dilutions within the tumor-bearing (3842) and normal (3217) plasma: control plasma samples identify approximately 17% fewer peptides than the cases. Reproducibility of signal intensities was also high, having correlation coefficients of 0.991 in the cases and 0.993 in the controls.

Results

Generating an AMT Database. We downloaded all 84 LC−MS/MS experiments described above with msInspect using the caBIG interface to CPAS, applied normalization procedures, and evaluated the association between RT and H in all experiments (Figure 2). We found that two experiments (indicated by blue and yellow) have retention times (horizontal axis) dramatically different both from each other and from the remaining 82 experiments. However, within each experiment, a clear linear association between RT and H exists. Within each experiment, the average correlation coefficient is 0.991.

We show the reduction of variability of the normalized retention times ($\hat{H}$) compared to prenormalization clock times in Figure 2. For each peptide observed in two or more experiments, we plot normalized retention time within an experiment (vertical axis) versus median retention time across all experiments (horizontal axis). We see that the between-experiment differences seen in Figure 2A, which are dependent on the LC configuration, are dramatically reduced. The residual standard deviation of the points surrounding the major axis is 0.017H, which implies that the ability to predict normalized retention times between runs from previous retention times has R-squared = 0.9997. Thus, although actual retention time varies greatly from run to run, normalized retention times provide a good predictor of future retention times.

The normalization procedure is demonstrated in Figure 2A. Points 1A and 1B represent the same peptide sequence in two different experiments with greatly different clock times. Points 2A and 2B represent another peptide sequence observed in each run at the same time as peptide 1 but having different predicted H value. Because peptides 1 and 2 elute at the same time within each experiment, they clearly possess the same hydrophobicity, even though their calculated hydrophobicity (H) is in error. The normalized (fitted) hydrophobicity projects clock time onto the normalized time scale using the calculated hydrophobicity as an intermediate value for normalization. Projection to the vertical axis shows the hydrophobicity scale, which is less dependent on LC configuration (see Point X in Figure). Figure 2B shows the large degree of reliability of from run to run of this retention time normalization approach.

Calibrating and Filtering LC−MS Peptide Mass. We then constructed the AMT database combining all of these experiments, as described above, resulting in 6412 entries. Refinement procedures removed 357 individual observations that were more than 3 standard deviations away from the median observation for that peptide (for peptides with 3 or more observations). We also removed 96 peptide entries with only one observation which differed from their predicted hydrophobicity by more than two standard deviations. The resulting

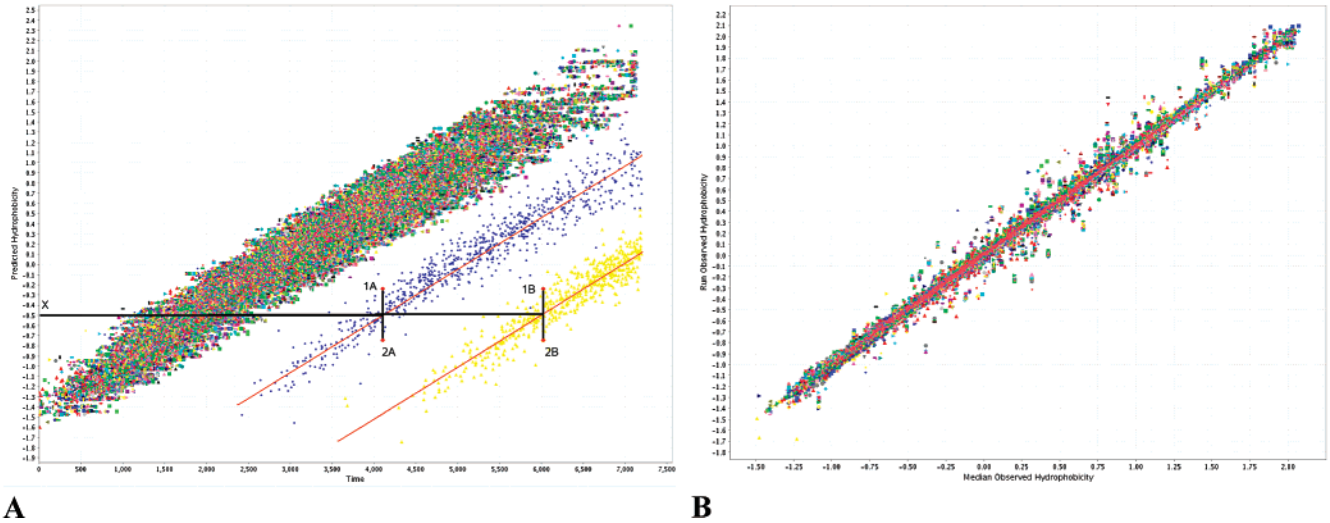

Figure 2. (A) Scatter plot of retention time (*x* axis) versus calculated hydrophobicity (*y* axis) for each of 42 LC−MS/MS experiments, each in its own color. The blue and yellow indicate two experiments with atypical LC−gradients. Black lines indicate RT and normalized RT ($\hat{H}$) for two peptides that are found across these two abnormal runs. (B) Scatter plot of median normalized retention time ($\hat{H}$) versus per-run normalized retention time ($\hat{H}$) for all peptides in the AMT database.

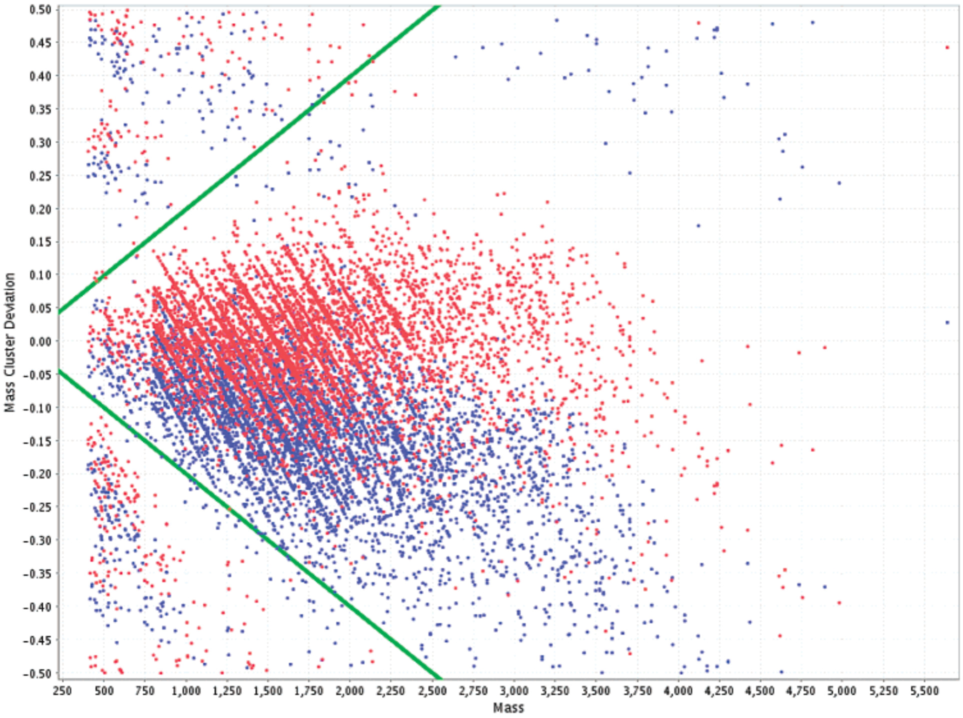

Figure 3. Demonstration of mass calibration of LC−MS peptides. *x* axis represents peptide mass; *y* axis represents deviation from the closest theoretical peptide mass cluster. Before calibration, all LC−MS interrogations in this dataset exhibited a significant, and consistent, downward shift of masses through the mass range. Mass calibration restores the masses to their proper values. Green lines represent a tolerance of 200 ppm around theoretical mass cluster centers. Peptides outside of those lines are excluded from the experiment.

AMT database of both case and control tissue peptide observations included 6316 unique peptide sequences.

We next processed the LC−MS peptide feature files by recalibrating their masses and filtering out low-quality peptide locations. For each peptide, feature file msInspect/AMT examined the association between each peptide's mass (as reported by the instrument) and distance to the nearest theoretical mass cluster.[13] For peptide feature files in which a linear relationship with nonzero slope is found (indicating calibration error), masses are readjusted. In this case, all six MS1 peptide feature files contained similar, highly significant deviations. The relationship between mass and cluster deviation for the 5 pmol/normal feature file is shown in Figure 3. The blue dots represent peptide masses as reported in the original peptide feature file, and the red dots represent the same features after recalibration. The linear association seen in the red points represents approximately a 75 ppm systematic error (a mass error of 0.15 Daltons for a peptide having 2000 Daltons), a magnitude that would have reduced the AMT matching rate dramatically (see Discussion).

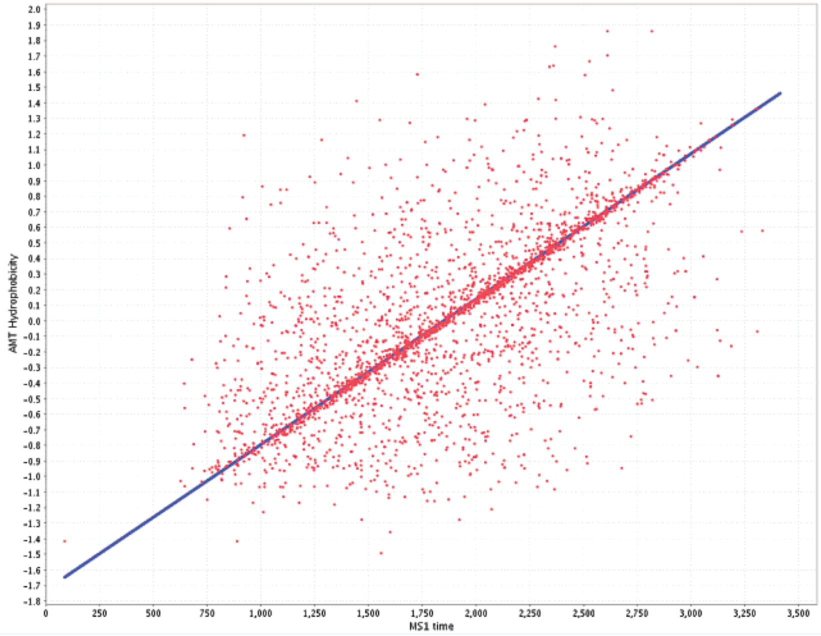

Figure 4. Results of a mass-only matching between and LC−MS set of peptides and the AMT database. *x* axis represents the RT of the LC−MS feature, and *y* axis represents the observed H of the AMT database entry. True matches stand out as a dense line against a roughly uniform background of false matches.

We next made use of the mass calibration method to identify and remove individual peptides that are likely incorrectly identified by the LC−MS peptide identification tool (in this case, msInspect). MsInspect/AMT generated the green lines in Figure 3, which identify peptides having individually estimated mass defect errors within 200 ppm, a tolerance recommended by Wolski.[13] We removed all peptides associated with points outside that region from each peptide feature file. Overall, we removed between 132 and 180 (mean: 154) peptides from each of the peptide feature files; see columns 3 and 6 of Table 1.

We next transformed RT to $\hat{H}$ for each of the peptide feature files, using a two-step process. The first step involves mass-only matching of features between the AMT database and the peptide feature file. Prior to matching, msInspect/AMT adjusted all peptides in the AMT database containing one or more Cysteine residues by expected 57.021-Dalton modification on that residue (C2H3NO, due to alkylation via carboxyamidomethylation), and AMT masses containing Methionine were duplicated (and the duplicate masses adjusted) to account for an expected possible 16-Dalton modification due to oxidation. All peptides from the feature file were then matched to the database by mass only, using a 10 ppm mass tolerance. The resulting scatter plot of retention time (LC−MS features) and hydrophobicity (AMT peptides) for these matched pairs is shown in Figure 4. Many points form an undifferentiated low-density background of matching errors, but the underlying linear model associating the (likely) correct matches, which defines the correct transformation of RT to $\hat{H}$, is clearly visible. To estimate this model, msInspect/AMT applied quantile regression procedures and iterated the model until the mode of the residuals maximized at zero (see Supporting Information for more details). The estimated linear model was then used to transform all RT measurements to $\hat{H}$.

Matching to the AMT Database and Calculation of False Assignment Rate. We next performed the two-dimensional divisive clustering procedure between each normalized peptide feature file and the AMT database to assign amino acid sequences to the peptide locations. We selected cluster diameters of 15 ppm in the mass dimension (derived from observed instrument capabilities) and .035 H in the hydrophobicity dimension. To choose the hydrophobicity cluster diameter, we performed AMT matching on the 5 pmol/normal experiment, varying the H tolerance between 0.01 and 0.07. For each H tolerance setting, msInspect/AMT generated a decoy AMT database and reapplied the same retention time normalization and clustering techniques, then computed the false assignment rate (FAR) as the ratio of the decoy matches over the forward matches. The results are summarized in Table 2, which relates the tolerances to the total number of forward and reverse matches. We also estimated the approximate number of true matches, as Total × (1 − FAR).

Table 3.

H tolerance	total matches	decoy matches	FAR	approximate true matches
0.01	488	39	0.079918033	449
0.015	692	55	0.079479769	637
0.02	903	72	0.079734219	831
0.025	1090	96	0.088073394	994
0.03	1224	122	0.099673203	1102
0.035	1335	132	0.098876404	1203
0.04	1414	148	0.10466761	1266
0.045	1469	158	0.107556161	1311
0.05	1516	174	0.114775726	1342
0.055	1569	187	0.119184194	1382
0.06	1600	203	0.126875	1397
0.065	1636	216	0.13202934	1420
0.07	1672	234	0.139952153	1438

As expected, the total number of matches increases with larger tolerances, as do the FAR and approximate true matches. On the basis of this table, we selected a tolerance of 0.035 H, which resulted in a total of 1335 LC−MS peptides from the 5 pmol/normal feature file matching the AMT database, with only 132 matching the decoy AMT database (FAR = 132/1335 = 9.8%) (Table 3).

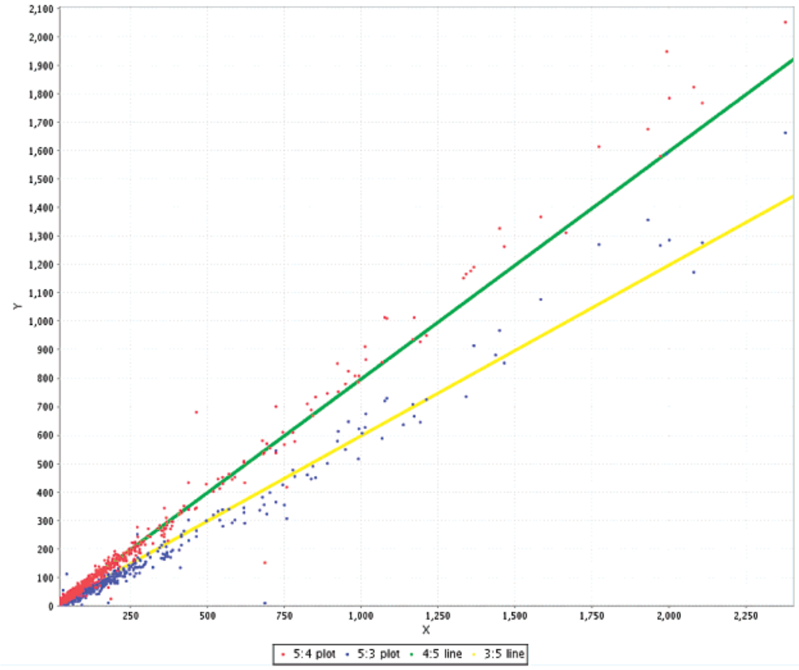

Figure 5. Comparison of peptide intensities matched to the AMT database in 5 pmol, 4 pmol, and 3 pmol normal plasma LC–MS features. The *x* axis represents intensity in the 5 pmol interrogation. The *y* axis represents intensity in the 4 pmol (red) and 3 pmol (blue) interrogations. The green line represents a 4:5 ratio; the yellow line represents a 3:5 ratio.

The error rate suggests that we are making accurate assignments overall. To further confirm this, we then investigated whether, across runs, these matched peptides have intensities consistent with the three different concentrations used (3 pmol, 4 pmol, and 5 pmol) of the cancer and control serum samples. We examined this association (Figure 5), which summarizes the performance of quantitation across runs using only those peptide features which match the same AMT peptide entry in two or more runs. Correct AMT assignments should follow the expected dilution pattern. The red and blue points represent comparisons of the 5 pmol intensities to the 4 pmol and 3 pmol intensities, respectively, for those peptides that match to the same AMT sequences. On the basis of the dilution assays, we would expect the peptide intensities for these two comparisons to have ratios of 5/4 = 1.2 (green line) and 5/3 = 1.66 (yellow). The actual median ratios for the peptides that are matched are 1.19 and 1.55, respectively. The correlation coefficient of the red (5/4) points is 0.995, implying an overall low level of assignment errors.

Discussion

We provided several demonstrations of the overall validity of our AMT workflow and algorithms. On the basis of a starting set of 84 MS/MS experiments interrogating tumor and normal mouse tissue, we generated an AMT database containing 6316 unique peptides. From thousands of LC–MS peptide features obtained from interrogations of plasma, we were able to identify a subset of approximately 2935 unique serum peptides that matched the tissue AMT database with a false assignment rate under 10%. Moreover, the relationship of intensities of those peptides across LC–MS interrogations is consistent with the dilutions of the designed experiment (as seen in Figure 5). Together, these findings demonstrate the overall validity of the algorithms and approaches.

Although our approach makes use of several component algorithms, including retention time normalizations and alignment algorithms, we have not made a comprehensive characterization of any specific component, nor have we demonstrated all the potential variations possible. For example, the results above made use of linear alignment schemes for retention time prediction and for AMT assignment, but msInspect/AMT also permits nonlinear alignments based on smooth-spline estimation, as described previously.[2] Nonlinear alignment procedures may provide better or worse performance depending on the relationship between gradients in a given set of experiments because of the potential for over-fitting, which may exist in the highly adaptive nonlinear methods. It is also difficult to compare our results to previously reported performance of AMT based approaches due to lack of availability of the software components or data to make that comparison and the lower resolution instrumentation we chose to demonstrate our proof of principle.

The workflow described here contains functionality similar to traditional AMT[5] and also PEPPeR,[8] but msInspect/AMT was designed to provide a highly flexible approach to combining data from different experimental and biological sources. For example, our AMT matching procedures described above were designed to function when only a potentially small number of peptide locations from an LC–MS experiment are present in the AMT database, such as may be the case when interrogating across dissimilar biospecimen types. Also, a unique feature of msInspect/AMT is the ability to account for potential mass calibration errors, which may be common when combining data across different instruments and data from different laboratories. As described above, all of the LC–MS experiments we aligned to the AMT database required re-calibration before matching was performed. AMT matching of the 5 pmol/normal experiment data without re-calibration, for instance, resulted

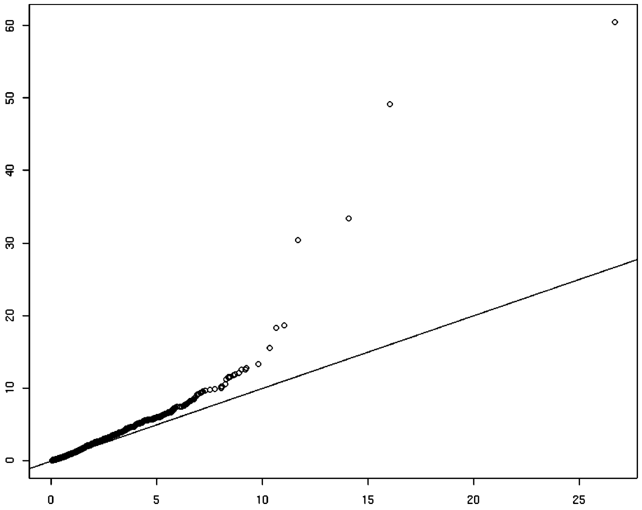

Figure 6. Plot of case to control intensity ratios (5 pmol) from LC−MS fingerprints for peptides that match the AMT database (vertical axis) versus those that do not match the AMT database (horizontal axis).

in only 55 matches, even fewer than the 132 we find with a reverse AMT database. After re-calibration, AMT matching produced 1203 matches. Thus, such mass re-calibration can be necessary to make use of AMT methodologies.

The example we chose demonstrated not only our specific algorithms but also the utility of AMT approaches in general for prioritizing serum biomarker candidates to identify higher-priority candidates. Ordinary LC−MS fingerprinting approaches are capable of identifying a large number of putative differential peptides, each of which may be equivocal in their empirical evidence (e.g., sensitivity or specificity). In practice, those which are tumor derived, as opposed to those that may result from inflammatory processes, may be the best candidates to pursue.[20] Although we have demonstrated that plasma and tissue share a large number of common peptides, we have not yet demonstrated that some of those serum peptides may have the tissue as their origin or otherwise may be used as a filter to enrich for the most promising biomarkers. A simple procedure for identifying biomarkers may be to select those LC−MS peptides having the highest case-to-control ratio. Figure 6 plots those case-to-control ratios for the 5 pmol LC−MS experiment, comparing those that match the AMT database (vertical axis) versus those that do not match the AMT database (horizontal axis). This plot clearly shows that the peptides that match the AMT database contain an overall enrichment toward those that are elevated in cancer. This is consistent with the hypothesis that we are able to use AMT methods to enrich for potential tissue-derived peptides in plasma interrogations and, thereby, to detect differentially expressed tissue-derived peptides in plasma.

Although a definitive statement would require more comprehensive experimentation, Figure 6 provides indirect evidence that those peptides which match the database may be enriched for putative biomarkers. Further support for this hypothesis may be demonstrated by interrogating for proteins that have been validated in these mouse models previously. Using the same mouse model, one of the proteins we have identified has been previously described by Zhang et al.[16] as overabundant in both tumor tissue and plasma in cancer compared to controls, with the plasma abundance confirmed by ELISA. In our analysis, four peptides from this protein were identified in MS/MS and are contained in the AMT database. All four of these peptides matched via AMT a peptide in the case plasma interrogation, with intensities 34.5, 183.7, 221.5, and 254.0. However, only one peptide (the one with intensity of 254.0 in the case) was also identified via AMT methods in normal plasma, and with an intensity of only 5.17. Thus, for this confirmed validated biomarker in these samples, we were able to validate our approach.

Conclusion

We have presented a suite of software to implement the AMT workflow based on the integration of many other open-source algorithms and software. The use of AMT methodologies may play a unique role in synthesizing data across different instrumentation and data types. In particular, the use of generalized AMT workflows that incorporate normalization of retention times and mass calibrations might help reduce the dependence of results on specific, local LC and mass spectrometer configuration. This can allow researchers to interrogate their data for peptides reported in the literature which they may not have sequenced by LC−MS/MS. The ability to reduce data comparability dependence on MS/MS identifications may assist in data reuse or sharing across laboratories.

Acknowledgment. This work was funded by National Cancer Institute contract # 23XS144A and Canary Foundation.

Supporting Information Available: Discussion of integration with the CPAS system for acquiring data files; detailed description of our robust regression methods for relating retention time to hydrophobicity; AMT database file schema description; explanation of quantile regression techniques. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Whiteaker, J. R.; Zhang, H.; Eng, J. K.; Fang, R.; Piening, B. D.; Feng, L. C.; Lorentzen, T. D.; Schoenherr, R. M.; Keane, J. F.; Holzman, T.; Fitzgibbon, M.; Lin, C.; Cooke, K.; Liu, T.; Camp, D. G., 2nd; Anderson, L.; Watts, J.; Smith, R. D.; McIntosh, M. W.; Paulovich, A. G. *J. Proteome Res.* **2007**, *6*, 828−836.

(2) Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. *Bioinformatics* **2006**, *22*, 1902−1909.

(3) Leptos, K. C.; Sarracino, D. A.; Jaffe, J. D.; Krastins, B.; Church, G. M. *Proteomics* **2006**, *6*, 1770−1782.

(4) Domon, B.; Aebersold, R. *Science* **2006**, *312*, 212−217.

(5) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513−523.

(6) Norbeck, A. D.; Monroe, M. E.; Adkins, J. N.; Anderson, K. K.; Daly, D. S.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1239−1249.

(7) 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. *Anal. Chem.* **2005**, *77*, 2187−2200.

(8) Jaffe, J. D.; Mani, D. R.; Leptos, K. C.; Church, G. M.; Gillette, M. A.; Carr, S. A. *Mol. Cell. Proteomics* **2006**, *5*, 1927−1941.

(9) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. *Proteomics* **2005**, *5*, 3537−3545.

(10) 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.; Whitaker, J.; States, D.; Hanash, S.; Paulovich, A.; McIntosh, M. W. *J. Proteome Res.* **2006**, *5*, 112−121.

(11) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3*, 908−919.

(12) Keller, A.; Eng, J. K.; Zhang, N.; Li, X. J.; Aebersold, R. *Mol. Sys. Biol.* **2005**, *1*, E1−E8.

(13) Wolski, W. E.; Farrow, M.; Emde, A. K.; Lehrach, H.; Lalowski, M.; Reinert, K. *Proteome Sci.* **2006**, *4*, 18.

(14) R. Development Core Team. A. language and environment for statistical computing. *Foundation for Statistical Computing*; Vienna, Austria, 2004.

(15) Piening, B. D.; Wang, P.; Bangur, C. S.; Whiteaker, J.; Zhang, H.; Feng, L. C.; Keane, J. F.; Eng, J. K.; Tang, H.; Prakash, A.; McIntosh, M. W.; Paulovich, A. *J. Proteome Res.* **2006**, *5*, 1527−1534.

(16) Zhang, H.; Zhao, L.; Wang, P.; Kelly-Spratt, K. S.; Ivey, R.; Piening, B.; Feng, L.-C.; Kasarda, E.; Gurley, K. E.; Eng, J.; Whiteaker, J. R.; Chodosh, L. A.; Kemp, C. J.; McIntosh, M.; Paulovich, A. G. Pending publication.

(17) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.

(18) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

(19) Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. *Nat. Biotechnol.* **2004**, *22*, 1459−1466.

(20) Hanash, S. *Nat. Rev. Cancer* **2004**, *4* 638−644.

PR070146Y