NIH Public Access
Author Manuscript
J Proteome Res. Author manuscript; available in PMC 2013 March 2.

Published in final edited form as:
J Proteome Res. 2012 March 2; 11(3): 1686–1695. doi:10.1021/pr200874e.

Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment

Surendra Dasari[1], **Matthew C. Chambers**[1], **Misti A. Martinez**[2], **Kristin L. Carpenter**[2], **Amy-Joan L. Ham**[2,3], **Lorenzo J. Vega-Montoto**[1,2], and **David L. Tabb**[1,2,3,4]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8575

[2]Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232-6840

[3]Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee 37232-0146

Abstract

Spectral libraries have emerged as a viable alternative to protein sequence databases for peptide identification. These libraries contain previously detected peptide sequences and their corresponding tandem mass spectra (MS/MS). Search engines can then identify peptides by comparing experimental MS/MS scans to those in the library. Many of these algorithms employ the dot product score for measuring the quality of a spectrum-spectrum match (SSM). This scoring system does not offer a clear statistical interpretation and ignores fragment ion m/z discrepancies in the scoring. We developed a new spectral library search engine, Pepitome, which employs statistical systems for scoring SSMs. Pepitome outperformed the leading library search tool, SpectraST, when analyzing data sets acquired on three different mass spectrometry platforms. We characterized the reliability of spectral library searches by confirming shotgun proteomics identifications through RNA-Seq data. Applying spectral library and database searches on the same sample revealed their complementary nature. Pepitome identifications enabled the automation of quality analysis and quality control (QA/QC) for shotgun proteomics data acquisition pipelines.

Introduction

Shotgun proteomics is a vital technology for clinical proteomics [1]. Hundreds of laboratories around the world routinely employ this technique for characterizing clinical samples. Although this popularity has led to many valuable findings, the shotgun technique has suffered from at least two critical weaknesses. First, repeated discovery of the same peptides by sequence database search is time consuming and error-prone. Second, searching for many peptide modifications simultaneously by database search is untenable [2]. In this context, spectral libraries are seen as a promising alternative to sequence database searching [3,4]. The spectral library method has been enabled by the collation of existing identifications with corresponding tandem mass spectra into a searchable library [5,6]. Peptides from new experiments can then be identified by matching their MS/MS to the library spectra.

The success of a library depends upon the completeness of its content; only those peptides contained in the library have any possibility of identification. This potential Achilles' heel

[4]Corresponding author: (voice) 615-936-0380, (fax) 615-343-8372, david.l.tabb@vanderbilt.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

has caused many potential users of library search to shy away from it. The recent explosion of shotgun proteomics data in public repositories, however, has substantially alleviated this concern [4]. By trawling well-vetted public datasets, researchers in multiple laboratories have created massive catalogues of peptide sequences along with the tandem mass spectra that match to them. The National Institutes of Standards and Technology (NIST) maintains a public library of proteomic tandem mass spectra that contains ion trap MS/MS scans for more than 190,539 distinct peptides from human samples [7]. Academic sites have also been busily collecting spectra; the Global Proteome Machine (GPM) contains more than 300 million peptide identifications [8]. The data stored in these libraries constitute a very powerful resource.

Multiple tools have been developed for identifying peptides using spectral libraries. A prescient paper by the Yates Laboratory demonstrated the use of cross-correlation for proteomic library search prior to the existence of substantial public libraries [9]. More recently, a second generation of dot-product library search tools has been developed; The GPM unveiled X!Hunter[5] to take advantage of their substantial libraries, and NIST introduced MSPepSearch[10] to complement their own. The Institute for Systems Biology (ISB) integrated their SpectraST[11] software into the Trans-Proteomic Pipeline, introducing this technology to its user base. The MacCoss laboratory created the BiblioSpec[12] software to field LC-MS/MS datasets against their own collections of spectra. The dot product scoring system shared by this generation of tools, however, has important limitations. First, dot products must be adapted to reflect that intense peaks vary in intensity more than minor peaks. Next, matching intensities for fragments from peptide bonds is more important than matching intensities for other ions. Dot products do not inherently evaluate discrepancies in the m/z values of ions being compared between spectra. Finally, this comparison strategy produces spectral contrast angles that lack a clear statistical interpretation.

Each implementation of these scorers addresses these weaknesses somewhat differently. Many have handled the first challenge by replacing each peak’s intensity with its square root, for example. With X!Hunter[5], Craig *et al.* produced expectation values from dot product scores to infer statistical significance from dot product distributions. The Spec2Spec tool from Yen *et al.* characterized improvements in scoring through use of intensity rank rather than intensity [13]. Compared with the palette of scoring approaches for database search algorithms, though, library search has been relatively monochromatic. New approaches such as an adaptation of the Kolmogorov-Smirnoff statistic to spectral comparison [14] have shown powerful discrimination of matches. Exploring novel means to score library matches has significant potential to improve library search sensitivity and specificity.

In this study, we describe “Pepitome,” a new algorithm for performing rapid and accurate searches of spectral libraries. The software employs probabilistic scoring metrics to evaluate the match between library and experimental MS/MS. Pepitome is integrated with a bioinformatics pipeline containing a high-performance database search engine and a flexible protein assembler. In this report, we compare the new software to the open-source SpectraST software in data sets from a variety of platforms. We validate the peptides gained through spectral library search in the context of complementary RNA-Seq data. For the first time, we show that spectral library-derived identifications are useful in performing quality analysis and quality control (QA/QC) of shotgun proteomics experiments.

Materials and Methods

Figure 1 illustrates the computational pipeline for identifying peptides from LC-MS/MS experiments. The standard peptide identification paradigm employs a database search engine (MyriMatch, in this case) to compare candidate peptides derived from a protein sequence

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

database to tandem mass spectra [15]. A protein assembler (IDPicker) filters the resulting identifications and assembles a protein list from the confident peptide identifications [16,17]. In this work, we substituted a spectral library search engine for the database search engine. Pepitome rapidly assigns peptide sequences to the experimental MS/MS by comparing them to pre-identified spectra stored in a library. IDPicker works equally well for either technique. The source code and binaries of all the software used in the workflow are available for download from our Web site: http://fenchurch.mc.vanderbilt.edu/.

Overview of Pepitome

We developed Pepitome (pronounced like "epitome") for rapid identification of peptides using spectral libraries. The software expects three types of input files: the raw MS/MS data, a spectral library, and a corresponding protein sequence database. Pepitome reads the raw files in a variety of instrument-native and derived formats (see Figure 1 for a partial list). The software reads spectral libraries in the MSP format of MSPepSearch [10] and the SPTXT format of SpectraST [11]. Both these formats are text, slowing repeated and random access. To remedy this, Pepitome indexes library files during their first use and refers to the index for subsequent analyses. The software writes peptide identifications in pepXML [18] and HUPO-PSI mzIdentML [19] formats. Pepitome was written in C++, and its multithreaded architecture can take advantage of multiple CPUs or multicore CPUs.

Pepitome starts by preprocessing the raw MS/MS spectra to remove the precursor ion and its associated neutral loss ions. Next, noise peaks are removed from the MS/MS using either a flexible total ion current (TIC) filter or a rigid peak count filter. The TIC filter computes the TIC available in the MS/MS and orders its fragment ions by decreasing order of their intensity. The filter walks down the sorted list and accepts the minimum number of ions required to meet a user-specified TIC threshold [15]. In contrast, the peak count filter accepts only the N (user-specified) most intense ions from the sorted list. After preprocessing, the intensities of the remaining fragment ions in the MS/MS are replaced by their ranks, where the most intense ion receives a low rank and the least intense ion receives a high rank.

When Pepitome detects that the mass of a library peptide matches an experimental precursor mass within a user-defined mass tolerance, the software loads the corresponding library spectrum into memory and applies identical preprocessing steps. Next, peak m/z positions in the library spectrum are matched to the peak m/z locations in the experimental MS/MS using a user-defined mass error window. If multiple library peaks match a single experimental peak or vice versa, the peak pair with the lowest m/z error is considered as a match. Pepitome computes three orthogonal sub-scores to assess the quality of a spectrum-spectrum match (SSM): a hypergeometric test (HGT), a Kendall-Tau T statistic [20], and an evaluation of m/z fidelity. Given a library spectrum and an experimental spectrum, the HGT score measures the probability of obtaining more than the observed number of peak matches by random chance, which follows a hypergeometric distribution and is described by the equation:

$$prob(X>k)=\sum_{i>k}^{m}\left\{\frac{\binom{m}{i}\binom{N-m}{m-i}}{\binom{N}{m}}\right\}$$

where *k* is the number of peak matches, *m* is the number of peaks in the library spectrum, and *N* is the total number of occupied and unoccupied m/z bins (of width equal to the fragment mass tolerance) in the experimental spectrum. The Kendall-Tau score measures the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

correspondence between the intensity ranks of the set of peaks matched between spectra. This score ranges from −1 (perfect inverse correlation) to 1 (perfect direct correlation). The raw Kendall-Tau score is converted into a probability of obtaining better than the observed intensity correlation by random chance, which is approximated using a normal distribution with μ=0 and $\sigma^2 = 2(2k+5)\ /9k\ (k-1)$[21] The mzFidelity[22] score measures the probability of obtaining the observed mass errors between the m/z locations of the matched peaks. The software reports all scores in negative logarithmic domain. An important distinction among the three scores is that HGT and Kendall-Tau scores are p-values, whereas mzFidelity is a point probability. Pepitome employs Fisher's Method [23] to combine the HGT and Kendall-Tau p-values into a single ranking score. This score is used as a primary sort order for the library matches, with mzFidelity acting as a tie-breaker.

At times, the peptide-protein associations present in a spectral library are incomplete. To remedy this, Pepitome can refresh all the peptide-protein associations against a user-supplied protein database, much as the Trans-Proteomic Pipeline does via the "RefreshParser" tool. In this step, the software builds a Wu-Manber [24] keyword tree with all the identified peptide sequences. Protein sequences from the database are scanned with the peptide tree and all peptide-protein associations are simultaneously identified and updated. Library peptides without a corresponding protein sequence in the database are removed from the results.

Data Sets

We employed five different data sets to demonstrate the utility of Pepitome in rapid protein identification. Four of these data sets (UPS1 [25], DLD1 Cell Lines [17], Plasma [17], and MMR Cell Lines [22]) were described in previous publications and briefly summarized here. The RAW data files associated with the data sets (except BSA QA/QC) are available for download from the website http://www.vicc.org/jimayersinstitute/data/.

1. UPS1—The universal protein standard (UPS1) obtained from Sigma-Aldrich (St. Louis, MO) was previously characterized at Vanderbilt University as part of the Clinical Proteomics Technology Assessment for Cancer (CPTAC) initiative [25]. In brief, proteins present in the samples were reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAM), and digested with trypsin. The resulting peptide mixtures were analyzed on either an LTQ or LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) using LC-MS/MS. A total of 216,901 MS/MS spectra were collected in twelve replicate analyses (six on each instrument). The resulting raw files were transcoded to mzXML format using the msConvert[26] tool of the ProteoWizard library.

2. DLD1 Cell Lines—This data set contains previously analyzed human colon adenocarcinoma cell lines (DLD1) [17]. In brief, proteins present in the samples were reduced with DTT, alkylated with IAM, and digested with trypsin. The resulting peptide mixtures were analyzed on a Thermo Fisher LTQ mass spectrometer. A total of 51,652 MS/MS spectra were collected in four replicate LC-MS/MS analyses. The resulting raw files were transcoded to mzXML format using the msConvert tool of the ProteoWizard library.

3. Plasma—Human plasma samples were previously analyzed at the Buck Institute for Age Research as a part of the CPTAC initiative [17]. In brief, a plasma sample was depleted of the 14 most abundant proteins; remaining proteins were reduced with DTT, alkylated with IAM, and digested with trypsin. The resulting peptide mixture was analyzed on a QSTAR Elite mass spectrometer (MDS SCIEX, Concorde, CA). A total of 35,289 MS/MS spectra were collected in five replicate LC-MS/MS analyses. The resulting raw files were transcoded to mzXML format using the msConvert tool of the ProteoWizard library.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

4. Mismatch Repair Cell Lines (MMR Cell Lines)—This dataset contains previously analyzed RKO (mismatch repair deficient) and SW480 (mismatch repair proficient) human adenocarcinoma cell line samples [22]. Proteins in the samples were reduced with DTT, alkylated with IAM, and digested with trypsin. The resulting peptide mixtures were separated into 10 fractions using isoelectric focusing (IEF). Each fraction from the two cell lines was analyzed in triplicate on a Thermo Fisher LTQ-Orbitrap mass spectrometer, and a total of 486,252 MS/MS spectra were collected. The resulting raw files were transcoded to mzXML format using the msConvert tool of the ProteoWizard library. Independently, expressed RNA transcripts from the cell lines were sequenced with the Illumina Genome Analyzer II (Solexa) instrument at the Vanderbilt University Genome Sciences Resource core. Detailed RNA sequencing methods are described in a separate publication [27].

5. BSA QA/QC—The Jim Ayers Institute for Precancer Detection and Diagnosis at Vanderbilt University employs a quality analysis protocol for ensuring the integrity of the LC-MS/MS data. A standard bovine serum albumin (BSA) stock solution (catalogue no. 23209) was obtained in bulk from Pierce Protein Research Products (Thermo Fisher Scientific, Waltham, MA) and batched into 100μL aliquots for storage. When needed, an aliquot of BSA (2mg/mL) was thawed and suspended in 600μL of 100mM ammonium bicarbonate buffer. Protein disulfide bonds were reduced by adding 100μL of 45mM DTT to the solution followed by a 20-minute incubation at 50°C. Reduced bonds were alkylated with 100μL of 100mM IAM followed by dark room incubation at room temperature for another 20 minutes. Proteins were digested into peptides by adding 40μL of 1% trypsin followed by an overnight incubation at 37°C. Thirty μL of the peptide mixture were resuspended in 970μL of 0.1% formic acid to make a 10X BSA (final concentration is 6.26 μg/mL). A separate 1X BSA solution was prepared by diluting 20μL of 10X BSA with 180μL of 0.1% formic acid (final concentration is 0.626μg/mL).

The 1X BSA peptide mixtures were analyzed on a Thermo Fisher LTQ mass spectrometer (manufacturer serial no. LTQ20585; alias: Amigo-2) equipped with an Eksigent nanoLC and autosampler (Dublin, CA). Peptides were resolved on a 100μm × 11cm fused silica capillary column (Polymicro Technologies, Phoenix, AZ) packed with 5μm, 300Å Jupiter C18 resin (Phenomenex, Torrance, CA) and an inline 100mm × 4cm solid phase extraction column packed with the same C18 resin [28]. Liquid chromatography was performed at room temperature at a flow rate of 0.6μL/min with a gradient mixture of 0.1% formic acid (FA) in water (solvent A) and 0.1% FA in acetonitrile (solvent B). Peptides were loaded on to the column and washed with a 100%-98% gradient A for 15 minutes. After the initial washing period, peptides were eluted with a 75 minute gradient (98% A to 75% A in 35 minutes, followed by a rapid descent to 10% A in 30 minutes, and held at 10% A for 10 minutes). Centroided MS/MS scans were acquired with an isolation width of 2 *m/z*, activation time of 30ms, activation *q* of 0.250. Both MS and MS/MS scans were acquired with a maximum ion time of 100ms. Fragmentation for the MS/MS scans was performed at 30% normalized collision energy. The AGC target value was set to 30,000 for a MS scan and 10, 000 for a MS/MS scan. A full MS scan of the eluting peptides was obtained with a mass range of 400–2000 m/z followed by data-dependent MS/MS scans of the top five most intense precursors. MS/MS scans were trigged with a minimum threshold of 1000. The MS/MS spectra were recorded using the dynamic exclusion feature of the instrument (precursor window width of 3 Daltons, exclusion time of 60 seconds, repeat count of 1, repeat duration of 1, and an exclusion list of length 50). A total of 10,737,907 (approximately 10.7 million) spectra were collected in 902 replicate LC-MS/MS analyses spanning eighteen months. The resulting raw files were transcoded to mzXML format using the msConvert tool of the ProteoWizard library.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Bioinformatics Methods

Spectral Library Preparation—We searched the experimental spectra against the NIST's tandem mass spectral reference libraries [7]. For this purpose, we obtained three different libraries (on 11/29/2010) from the NISTs' website: the ion trap library for humans, the QTOF library for humans, and a single protein spectral library containing bovine serum albumin (BSA). The libraries were augmented with spectra corresponding to common contaminant proteins, which were obtained from the GPM website [8]. Decoy spectra were generated in one-to-one correspondence with target spectra in the library using Henry Lam's fragment shuffling method [29]. Generated decoys were appended to the libraries for estimating false discovery rates (FDRs) using IDPicker [16].

Peptide Identification—The MS/MS scans present in the five datasets were identified using three different algorithms: SpectraST[11], Pepitome, and MyriMatch[15]. Pepitome and SpectraST are spectral library search engines, whereas MyriMatch is a database search engine. Table 1 summarizes the data sets, protein sequence databases, spectral libraries, and mass tolerances used in all searches. MyriMatch was configured to derive semitryptic peptides from the protein database while using carbamidomethylation of cysteine (+57.0125 Da), oxidation of methionine (+15.996 Da), formation of N-terminal pyroglutamine (−17.0265 Da), and N-terminal acetylation (+42.0123 Da) as variable modifications. Pepitome and SpectraST were configured to consider only fully tryptic and semitryptic peptides from the libraries. All search engines produced identifications in pepXML format. Detailed configuration parameters for all searches are listed in Supplemental File 1.

IDPicker[16,17] filtered peptide identifications from all search engines at a false discovery rate (FDR) of either 2% or 5%. For MyriMatch, the software automatically combined the MVH, mzFidelity, and XCorr scores for FDR filtering. SpectraST results were filtered using the "fval" score. IDPicker combined the HGT, Kendall-Tau, and mzFidelity scores for filtering Pepitome identifications. Peptides passing the FDR thresholds were assembled into protein identifications using parsimony rules [16]. Protein identifications with at least two distinct peptide identifications were considered for further analysis.

Assembling an "Expressed Proteome" from RNA Sequencing (RNA-Seq) Data —Library peptide identifications from the MMR Cell Lines sample were validated against the samples' "expressed proteome." We derived this proteome with a method developed by Wang *et al.* [27] for deriving protein sequence databases from RNA-Seq experiments. According to this method, expressed RNA transcripts detected in the MMR Cell Lines samples were mapped to the human reference genome (hg18, Ensembl 54) using the Tophat software package [30,31]. Independently, the expression level of each transcript was computed with the number of "reads per kilobase per million mapped reads" (RPKM) metric. The detected gene-transcript associations were filtered to remove entries with an RPKM value below 20. Protein sequences corresponding to the detected gene-transcript associations were selected from the Ensembl database and included in a custom database, which represents the "expressed proteome" of the sample. A separate publication from Wang *et al.* describes this method in greater detail [27].

Artificial Neural Networks (ANNs) for LC-MS/MS Quality Control—We developed ANNs for automated LC-MS/MS data quality assessment using spectral library searches. To accomplish this, we utilized a total of 902 shotgun analyses of BSA standards. Tryptic digests of these standards were regularly analyzed on a Thermo LTQ-XL mass spectrometer (Amigo-2) over a period of 18 months. The Pepitome and SpectraST search engines identified the MS/MS scans in the data set. From these analyses we derived three distinct data tables for developing the ANNs.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

First, we analyzed the SpectraST search results and corresponding RAW files with the NIST MS Metrics software following the protocol described in reference [32]. The software produced a table containing 40 LC-MS/MS data quality metrics for each raw file (labeled as “NIST MS w/ SpectraST”; Table 1 in Rudnick et al [32] describes this list of metrics). Next, we filtered Pepitome search results with IDPicker at 2% FDR threshold. Filtered results were analyzed with the NIST MS Metrics software modified to read IDPicker XML files [17]. This analysis produced a separate table of data quality metrics similar to that of SpectraST’s (labeled as “NIST MS w/ Filtered Pepitome”). Finally, we developed a separate computer program to infer data quality metrics directly from the Pepitome search results. For each pepXML file, the program extracted the HGT and Kendall-Tau score distributions from all top ranking SSMs. The following properties of the score distributions were reported as data quality metrics: median, median absolute deviation, kurtosis, interquartile range, skewness, and standard deviation. The search results were filtered with IDPicker at 2% FDR threshold and the total numbers of filtered spectral and peptide identifications were added to the metrics table (labeled as “Pepitome Metrics”). Independently, a panel of mass spectrometry experts classified each raw file as either high or low quality following the standard operating procedures for “Instrument BSA Acceptance Checks” developed at Vanderbilt University (Supplemental File 2). These quality assessments served as a gold standard while developing the ANNs.

We developed three separate ANNs for differentiating between the high and low quality LC-MS/MS analyses. The training and testing regimens for the ANNs were identical except they utilized different tables of data quality metrics (“NIST MS w/ SpectraST” or “NIST MS w/ Filtered Pepitome” or “Pepitome Metrics”). For training an ANN, we composed a training table of data quality metrics from a random selection of 300 high quality raw files and 300 low quality raw files. We further processed this training data table to remove constant rows (raw files) and constant columns (metrics). We created a two layer feed-forward pattern recognition network containing one sigmoid hidden neuron for each feature and one output neuron for each quality class. Network training was performed with Levenberg-Marquardt algorithm utilizing a scaled conjugate gradient back propagation method [33]. The training attempted to recapitulate the expert quality assessment of the raw file from its corresponding data quality metrics. The training algorithm was safeguarded against over-fitting by segregating 10% of the training data for cross-validation. In parallel, we created an independent testing data table with the remaining 302 quality-assessed raw files. This test data table was employed for estimating the generalizability of the trained neutral network. All ANNs were developed in the 2011 version of the Matlab programming environment.

Results and Discussion

Pepitome was designed for rapid identification of shotgun proteomics tandem mass spectra. We compared the new algorithm to SpectraST because it represents an accepted implementation of spectral library search engines. We also compared the performance of Pepitome to that of MyriMatch as an example of the more common database search approach. We initially characterized the software performance on known protein mixtures and then shifted to human samples to demonstrate the value of library searching in a real-world setting. To evaluate differential identifications, we employed RNA-Seq data to confirm or refute peptides produced through only one method. Finally, the value of library searching for rapid quality control was evaluated in hundreds of LC-MS/MS experiments from standard samples.

Standard Dot Products are Biased toward Spectra Containing more Peaks

The current generation of library search tools uses the dot product scoring system for matching an experimental spectrum to a set of library spectra. The spectra in the NIST

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

libraries, however, differ significantly in the number of peaks that they contain. Some have been generated from dozen of experimental tandem mass spectra, averaged together, while others have been identified only once from a low-concentration peptide. Ideally, all library spectra would have an equal opportunity to be matched successfully to an experimental spectrum. When the dot product scores for all comparisons to an experimental tandem mass spectrum are evaluated, though, the highest scores are clearly associated with the library spectra that contain the most peaks (Figure 2). We chose to move away from the dot product scoring system by implementing probabilistic scoring systems in Pepitome. The software replaces dot products with p-values derived from the hyper-geometric test (HGT) and Kendall-Tau statistic (T). The HGT score measures the probability of obtaining more peak matches than observed by random chance, whereas the Kendall-Tau score estimates the probability of a better intensity rank correlation among peaks matched between the spectra. As a result, the HGT and Kendall-Tau scores do not carry inherent bias towards high density library spectra (Figure 2). This independence of scores from spectral peak density guarantees that short peptides are as likely to attain high scores as long peptides, improving the chance of matching library spectra for which database search identification has been problematic.

Improved Score Combination for Filtering Spectral Library Search Results

Pepitome produces three orthogonal score metrics for each SSM: HGT, Kendall-Tau (T), and mzFidelity. The HGT and Kendall-Tau scores are the probabilistic replacements of the dot product scoring system. The mzFidelity score measures the significance of m/z errors associated with the matched peaks. We measured the relative contribution of these scoring metrics when filtering library search results for the DLD1 Cell Lines, Plasma, and MMR Cell Lines data sets. These data sets represent samples analyzed on three different mass spectrometry platforms. Pepitome was configured to match the MS/MS against the respective spectral libraries. IDPicker filtered the results at 5% FDR using either the HGT score or an optimal combination of HGT and Kendall-Tau metrics. When the combined score filtered the results, we observed a gain of 21%, 4%, and 17% in the identification rates of DLD1 Cell Lines, Plasma, and MMR Cell Lines samples, respectively (Figure 3). The reduced scoring difference observed for the Plasma set emphasizes that QTOF performance is limited by spectral library content rather than scoring. Adding the mzFidelity score to the other two metrics generated insignificant gains (1%, 1%, and 2%, respectively), despite the expectation that the well-resolved, mass accurate fragments of the QTOF data set would benefit from this type of metric.

Search Engine Performance Comparisons

We compared the performance of Pepitome, SpectraST, and MyriMatch when performing routine identification searches. Pepitome and SpectraST worked from a common spectral library to limit the differences to those from preprocessing and scoring. MyriMatch generated model spectra for peptides drawn from a FASTA protein sequence database.

First, we tested the protein identification accuracy of the search engines when analyzing known protein mixtures (UPS1). We configured MyriMatch to match the MS/MS against the latest UniProt complete proteome set for humans. Reversed entries of the protein sequences were added to the database for estimating false discovery rates (FDRs). Pepitome and SpectraST matched the experimental spectra against the NIST spectral libraries. Decoy spectra were added to the libraries for estimating the FDRs. IDPicker filtered the identifications from all search engines at 5% FDR. We measured the ability of Pepitome, SpectraST, and MyriMatch to recover known protein identifications from technical replicates using metrics developed to gauge the performance of information retrieval systems: recall, precision, and F_1-measure (Figure 4). Recall measures the percentage of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

known proteins that were identified; Precision measures the percentage of identified proteins that are true positives (known to be present in the sample); F_1-measure is a harmonic mean of recall and precision. In both samples, Pepitome recovered more known proteins from the sample at a higher precision than did SpectraST and MyriMatch. The database search slightly underperformed compared to spectral library search engines, likely due to the larger search space associated with this technique. In general, Figure 4 establishes Pepitome as highly reliable protein identification software.

Next, we measured the peptide identification sensitivity of Pepitome, SpectraST, and MyriMatch when analyzing data from complex samples. For this, we employed the DLD1 Cell Lines (LTQ), MMR Cell Lines (LTQ-Orbitrap), and Plasma (QTOF) data sets. We performed the identification searches following the described protocol. IDPicker filtered the identifications at a stringent 2% FDR. Figure 5 presents a sample-wide summary of peptides and spectra identified by the respective search engines. Overall, Pepitome identified more peptides and spectra than SpectraST and MyriMatch. Pepitome identified 10%–12% more peptides and 3%–15% more spectra when compared to SpectraST. A similar comparison to the MyriMatch database search engine produced mixed results. When identifying ion trap tandem mass spectra, Pepitome identified more peptides than MyriMatch. In terms of peptide counts, the QTOF data were more successfully identified by database search than by spectral library search; Pepitome identified 21% fewer peptides than MyriMatch. The NIST QTOF spectral library is more than an order of magnitude smaller than the corresponding ion trap library.

Comparing to the X!Hunter[5], BiblioSpec[12], and MSPepSearch[10] library search engines would have been desirable, but technical issues stood in the way. For a fair comparison, the tools needed to be able to read the SPTXT / MSP libraries used for Pepitome and SpectraST, but only MSPepSearch can do so. Complicating this matter, X!Hunter has been optimized for libraries that store only the twenty most abundant fragments for each peptide. The next challenge is to translate outputs from these identifiers into pepXML for ingestion by IDPicker. At present, only X!Hunter outputs meet this criterion.

Validation of Spectral Library Search Results with RNA-Seq Data

Are the identifications gained through spectral libraries real or artifacts? The MMR Cell Lines data set complemented LC-MS/MS experiments with RNA-Seq data, enabling validation. Peptides in the shotgun data were identified with Pepitome and SpectraST search engines from the NIST spectral libraries. IDPicker filtered the resulting identifications at 2% FDR. Proteins were required to be supported by two distinct peptides. In parallel, RNA-Seq data were filtered to remove noise, with the filtered transcripts mapped to their protein products. Amino acid sequences of the mapped proteins were included in a custom database. This database represents the "expressed proteome" of the sample. In this context, we assumed that the RNA-Seq data serves as a gold standard; library peptide identifications that do not map to transcripts are most likely false.

We mapped the peptide identifications from the shotgun data to the expressed proteome inferred from the RNA-Seq data. For this, we first merged the peptide identifications from Pepitome and SpectraST searches of the data set (modifications to peptides were ignored). Next, we constructed an Aho-Corasick keyword trie[34] with the merged list. Sequences of the expressed proteins were scanned through the peptide trie and all peptide-protein associations (RNA-Seq evidence) were detected. During this process some peptides failed to map to a protein sequence (no RNA-Seq evidence). Figure 6 shows the overlap in peptide identifications between the library search engines and the RNA-Seq evidence. We identified a total of 15,204 peptides identifications from the Pepitome and SpectraST searches of the sample. Seventy six percent (11,486) of these peptides were identified by both search

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

engines, 17% (2,606) were exclusively identified by Pepitome, and 7% (1,112) were exclusively identified by SpectraST (Figure 6). Overall, 96% (14,626) of the peptide identifications in the sample have RNA-Seq evidence. When both library search engines have identified a peptide, the rate of RNA-Seq support is higher. These results attest that the sensitivity gains of spectral library searches can be trusted.

Database and Library Search Engines Identify Different Peptides

Spectral library and sequence database searches are different paradigms for protein identification, but to what extent are their results complementary? The MMR Cell Lines data set enabled an investigation of the overlap between these technologies. Peptides in the data set were identified with Pepitome library search engine as well as the MyriMatch database search engine, with IDPicker combination handled as described in the section above. Figure 7 shows the overlap between peptide identifications produced by a spectral library search (Pepitome) and a database search (MyriMatch). Overall, we identified a total of 16,603 peptides in the sample. A modest 61% of these peptides were identified by both database and spectral library search engines, with 24% of the peptides exclusively identified by a library search engine and 15% exclusively identified by MyriMatch. The existence of RNA-Seq evidence for peptides falling in these three mutually exclusive categories is also shown in Figure 7. As expected, a majority of peptides (98%) identified by both search strategies have confirmatory RNA-Seq evidence. Large portions of peptides that were exclusively identified by only one search strategy also have confirmatory RNA-Seq evidence. This reveals the complementary nature of the database and spectral library search strategies.

We characterized the difference between the protein identification lists produced by spectral library and database search strategies. For this, we mapped the peptides identified by the spectral library search to its corresponding protein sequences in the latest Uniprot complete proteome set for humans (version 2011_03). We ignored parsimony and compiled a list of proteins matching to at least two unique peptide sequences. Following this protocol, we also generated a separate protein identification list from the MyriMatch search results. We collated the two identification lists and detected a total of 5,671 unique proteins from the sample. Overall, 75% of these proteins were identified by both the spectral library and database search strategies; 15% of the proteins were exclusively identified by a database search, and 10% of the proteins were exclusively identified by a spectral library search. A majority of proteins that were exclusively identified by database search did not have representative PSMs in the library. Proteins identified by spectral library search but missed by database search fell into two categories: a) proteins containing peptides with unanticipated modifications (such as dehydration, deamidation, methylation, and phosphorylation) b) low abundance proteins for which peptides are accidentally outscored by false candidate sequences.

Spectral Libraries for Rapid Analysis of LC-MS/MS Data Quality

Quality analysis and quality control (QA/QC) systems are routinely employed in many industries. Recently, Rudnick *et al* proposed the “NIST MS Metrics” system for performing QA/QC of shotgun proteomics data acquisition pipelines [32]. This system is composed of forty-two carefully selected metrics from six categories: chromatography, dynamic sampling, ion source, MS signal, MS/MS signal, and peptide identification. Ma *et al* developed identification free data quality metrics for an LC-MS/MS experiment [35]. All these quality metrics act as gauges for measuring the instrument’s performance. When a quality control sample reveals evidence of suboptimal performance, sample queues can be halted until instrumentation problems are resolved. Instrument operators are less likely to employ QC systems that bewilder them with complex sets of metrics rather than straightforward judgments about quality, and being forced to wait on a slow computation for each QC

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

sample may waste instrument time. The rapid performance of spectral libraries makes them a natural fit for QC evaluation. We paired Pepitome with an artificial neural network (ANN) to translate its metrics into a “go / no go” decision.

We targeted an LTQ-XL mass spectrometer (Amigo-2) located at Vanderbilt University for method development. The Amigo-2 instrument routinely analyzes tryptic digests of BSA standards for ensuring the integrity of the LC-MS/MS data. We selected 902 replicate analyses from the instrument spanning eighteen months. A panel of experts manually inspected each of the raw files and assigned a quality label (high or low), following the protocol described in the Materials and Methods section. The quality labels are a direct reflection of Amigo-2 performance at the time of data acquisition. Pepitome identified the peptides present in the raw files, and IDPicker filtered the identification results at 2% FDR. Figure 8A illustrates the distributions of peptide and spectral identifications obtained from low and high quality raw files. The partially separable distributions in Figure 8A suggest that simple identification count-based data quality metrics have limited effectiveness. Building upon these differences, we tested three different categories of quality metrics for automated QA of the raw file: a) twelve simple metrics characterizing score distributions from Pepitome results file (“Pepitome Metrics”) b) forty-two metrics produced by the NIST MS software analyzing the raw file and its corresponding SpectraST results (“NIST MS w/ SpectraST”) and c) forty-two metrics produced by the NIST MS pipeline modified to substitute filtered Pepitome results for SpectraST results (“NIST MS w/ Filtered Pepitome”). We trained and tested a pattern recognition ANN for recapitulating the expert quality assessment of the raw data file from its corresponding collection of the quality metrics. Figure 8B illustrates the receiver operating curves (ROCs) for both training and testing phases of the ANN when using different categorical collections of quality metrics as inputs. These curves illustrate that library-derived identification data are generally sufficient for classifying LC-MS/MS experiments for quality, though the data would be unable to reveal which element of analysis failed for failing experiments. Substituting the filtered Pepitome results for raw SpectraST results slightly improved the performance of NIST MS quality metrics by recognizing identified scans more accurately and sensitively. Overall, fusing the Pepitome search engine with the NIST MS system produced the best discrimination between the low and high quality raw files. The classification accuracy of the ANN trained on the updated NIST MS system was 92% for low quality raw files and 90% for high quality files.

Conclusion

The Pepitome spectral library search engine employs probabilistic scoring systems for assessing the quality of spectrum-spectrum matches. Pepitome outperformed SpectraST when identifying peptides from complex LC-MS/MS data sets. RNA-Seq data confirms the reliability of gains achievable through library searches. The rapid speed of library search enables low-latency quality assessment for shotgun proteomics data acquisition pipelines. Pepitome is encapsulated in an open-source pipeline that can be routinely employed for large-scale protein identification work.

John Yates introduced spectral library searching in 1998. The adoption of this technique by the research community has accelerated due to the assembly of comprehensive public spectral libraries. As the major workflows for proteome informatics gain support for libraries, this strategy will gain traction with increasing numbers of users. Continued research in library-based identification algorithms will enable more comprehensive identification of spectra from experimental collections.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

D.L. Tabb, S. Dasari, and M.C. Chambers were supported by NIH grants R01 CA126218, R01 CA159988, and U24 CA126479. We thank Bing Zhang and Xiaojing Wang from Vanderbilt University for providing access to the RNA-Seq data of the "MMR Cell Lines" sample. The "BSA QA/QC" samples are a generous contribution from the Jim Ayers Institute for Precancer Detection and Diagnosis. We thank Paul Rudnick and Stephen Stein from NIST for providing access to the "NIST MS Metrics" source code.

References

1. Rodriguez H, Tezak Z, Mesri M, Carr SA, Liebler DC, Fisher SJ, Tempst P, Hiltke T, Kessler LG, Kinsinger CR, Philip R, Ransohoff DF, Skates SJ, Regnier FE, Anderson NL, Mansfield E. Analytical validation of protein-based multiplex assays: a workshop report by the NCI-FDA interagency oncology task force on molecular diagnostics. Clin Chem. 2010; 56:237–243. [PubMed: 20007859]
2. Dasari S, Chambers MC, Codreanu SG, Liebler DC, Collins BC, Pennington SR, Gallagher WM, Tabb DL. Sequence tagging reveals unexpected modifications in toxicoproteomics. Chem Res Toxicol. 2011; 24:204–216. [PubMed: 21214251]
3. Lam H, Aebersold R. Using spectral libraries for peptide identification from tandem mass spectrometry (MS/MS) data. Curr Protoc Protein Sci. 2010; Chapter 25(Unit 25.5)
4. Deutsch EW. Tandem mass spectrometry spectral libraries and library searching. Methods Mol Biol. 2011; 696:225–232. [PubMed: 21063950]
5. Craig R, Cortens JC, Fenyo D, Beavis RC. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res. 2006; 5:1843–1849. [PubMed: 16889405]
6. Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R. Building consensus spectral libraries for peptide identification in proteomics. Nat Methods. 2008; 5:873–875. [PubMed: 18806791]
7. [accessed Jul 15 2011] NIST Peptide Mass Spectral Reference Data. http://peptide.nist.gov/
8. Fenyö D, Eriksson J, Beavis R. Mass spectrometric protein identification using the global proteome machine. Methods Mol Biol. 2010; 673:189–202. [PubMed: 20835799]
9. Yates JR 3rd, Morgan SF, Gatlin CL, Griffin PR, Eng JK. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. Anal Chem. 1998; 70:3557–3565. [PubMed: 9737207]
10. [accessed Jun 24 2011] MSPepSearch. http://peptide.nist.gov/software/ms_search/MS_Search.html
11. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics. 2007; 7:655–667. [PubMed: 17295354]
12. Frewen B, MacCoss MJ. Using BiblioSpec for creating and searching tandem MS peptide libraries. Curr Protoc Bioinformatics. 2007; Chapter 13(Unit 13.7)
13. Yen C-Y, Houel S, Ahn NG, Old WM. Spectrum-to-Spectrum Searching Using a Proteome-wide Spectral Library. Mol Cell Proteomics. 2011; 10:M111.007666. [PubMed: 21532008]
14. Cannon WR, Rawlins MM, Baxter DJ, Callister SJ, Lipton MS, Bryant DA. Large improvements in MS/MS-based peptide identification rates using a hybrid analysis. J Proteome Res. 2011; 10:2306–2317. [PubMed: 21391700]
15. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res. 2007; 6:654–61. [PubMed: 17269722]
16. Zhang B, Chambers MC, Tabb DL. Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency. J Proteome Res. 2007; 6:3549–3557. [PubMed: 17676885]
17. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: Improved protein assembly with high

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

discrimination peptide identification filtering. J Proteome Res. 2009; 8:3872–3881. [PubMed: 19522537]

18. Keller A, Eng J, Zhang N, Li X-jun, Aebersold R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol. 2005; 1:2005.0017. [PubMed: 16729052]
19. Eisenacher M. mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms. Methods Mol Biol. 2011; 696:161–177. [PubMed: 21063947]
20. Kendall MG. A new measure of rank correlation. Biometrika. 1938; 30:81–93.
21. [accessed Nov 3 2011] Kendall coefficient of rank correlation. http://eom.springer.de/K/k055200.htm
22. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJL, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res. 2010; 9:1716–1726. [PubMed: 20131910]
23. Mosteller F, Fisher RA. Questions and Answers. The American Statistician. 1948; 2:30.
24. Wu, S.; Manber, U. A fast algorithm for multi-pattern searching. Department of computer science, University of Arizona; 1994.
25. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJL, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res. 2010; 9:761–776. [PubMed: 19921851]
26. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics. 2008; 24:2534–6. [PubMed: 18606607]
27. Wang X, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. Protein identification using customized protein sequence databases derived from RNA-Seq data. Journal of Proteome Research. 2011 (EPub).
28. Licklider LJ, Thoreen CC, Peng J, Gygi SP. Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. Anal Chem. 2002; 74:3076–3083. [PubMed: 12141667]
29. Lam H, Deutsch EW, Aebersold R. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. J Proteome Res. 2010; 9:605–610. [PubMed: 19916561]
30. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]
31. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
32. Rudnick PA, Clauser KR, Kilpatrick LE, Tchekhovskoi DV, Neta P, Blonder N, Billheimer DD, Blackman RK, Bunk DM, Cardasis HL, Ham AJL, Jaffe JD, Kinsinger CR, Mesri M, Neubert TA, Schilling B, Tabb DL, Tegeler TJ, Vega-Montoto L, Variyath AM, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Carr SA, Fisher SJ, Gibson BW, Paulovich AG, Regnier FE, Rodriguez H, Spiegelman C, Tempst P, Liebler DC, Stein SE. Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. Mol Cell Proteomics. 2010; 9:225–241. [PubMed: 19837981]
33. Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM J Appl Math. 1963; 11:431.
34. Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Commun ACM. 1975; 18:333–340.
35. Ma ZQ, Chambers MC, Ham AJL, Cheek KL, Whitwell CW, Aerni HR, Schilling B, Miller AW, Caprioli RM, Tabb DL. ScanRanker: Quality Assessment of Tandem Mass Spectra via Sequence Tagging. J Proteome Res. 2011; 10:2896–2904. [PubMed: 21520941]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

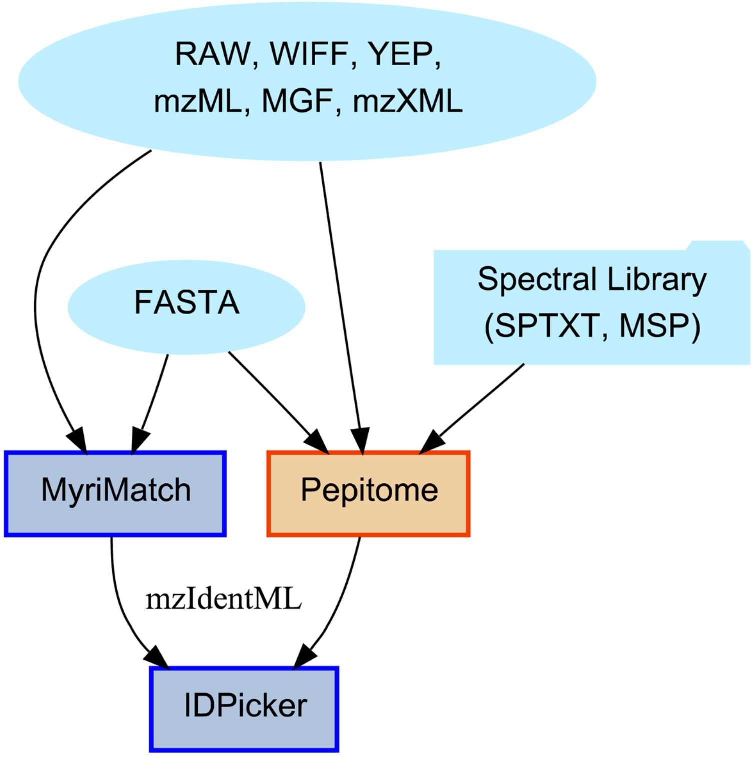


Figure 1. Peptide Identification Pipeline

MyriMatch is a database search engine. Pepitome is a spectral library search engine, which matches experimental MS/MS against library spectra. IDPicker is a parsimonious protein assembler, which filters peptide identifications using a target FDR.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

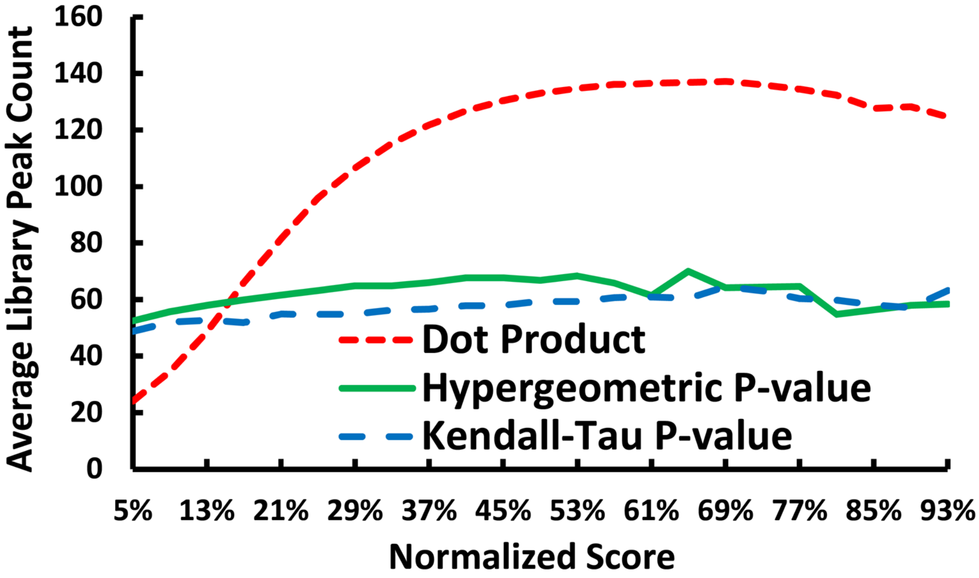


Figure 2. Probabilistic Scoring Systems Are Robust

One thousand MS/MS spectra were randomly selected from the DLD1 Cell Lines data set. Pepitome and SpectraST matched the spectra against the NIST ion trap library. Search engines were modified to make compact reports of all library comparisons made for each MS/MS. The top five matches by score were removed from each result set. The remaining matches were considered to be stochastic. This figure illustrates the functional relationship between the stochastic search scores and the peak density (average peak counts) of all the library spectra compared to the experimental spectra. Dot products have a positive bias towards high density library spectra. Probabilistic scores like hypergeometric test and Kendall-Tau statistic are resistant to changes in peak density of library spectra.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

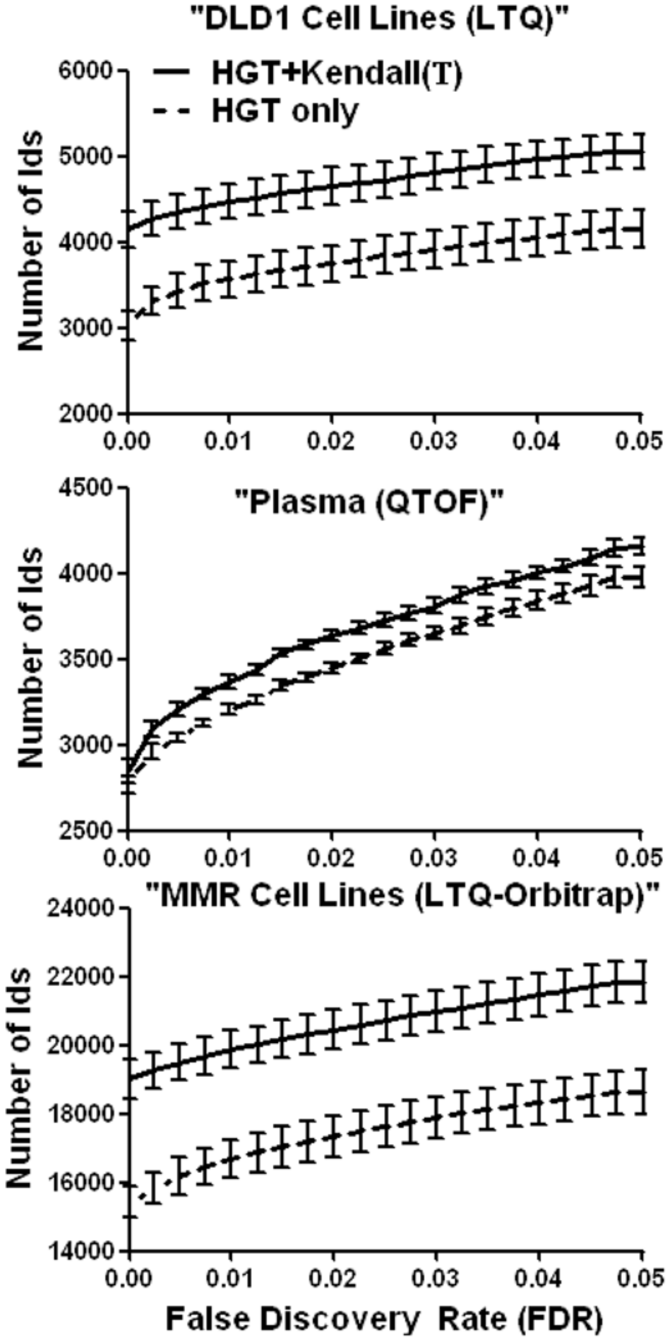


Figure 3. Intensity Correlation Metric Improves Identification Rates of Library Searches

Pepitome identified the MS/MS present in the samples. IDPicker filtered the results at 5% FDR using either HGT (peak presence or absence) score or an optimal combination of HGT and Kendall-Tau (intensity rank correlation) scores. In all samples, combining orthogonal scoring metrics improved peptide identification rates. Error bars in the figure represent standard error of the mean estimated from the replicate LC-MS/MS analyses.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

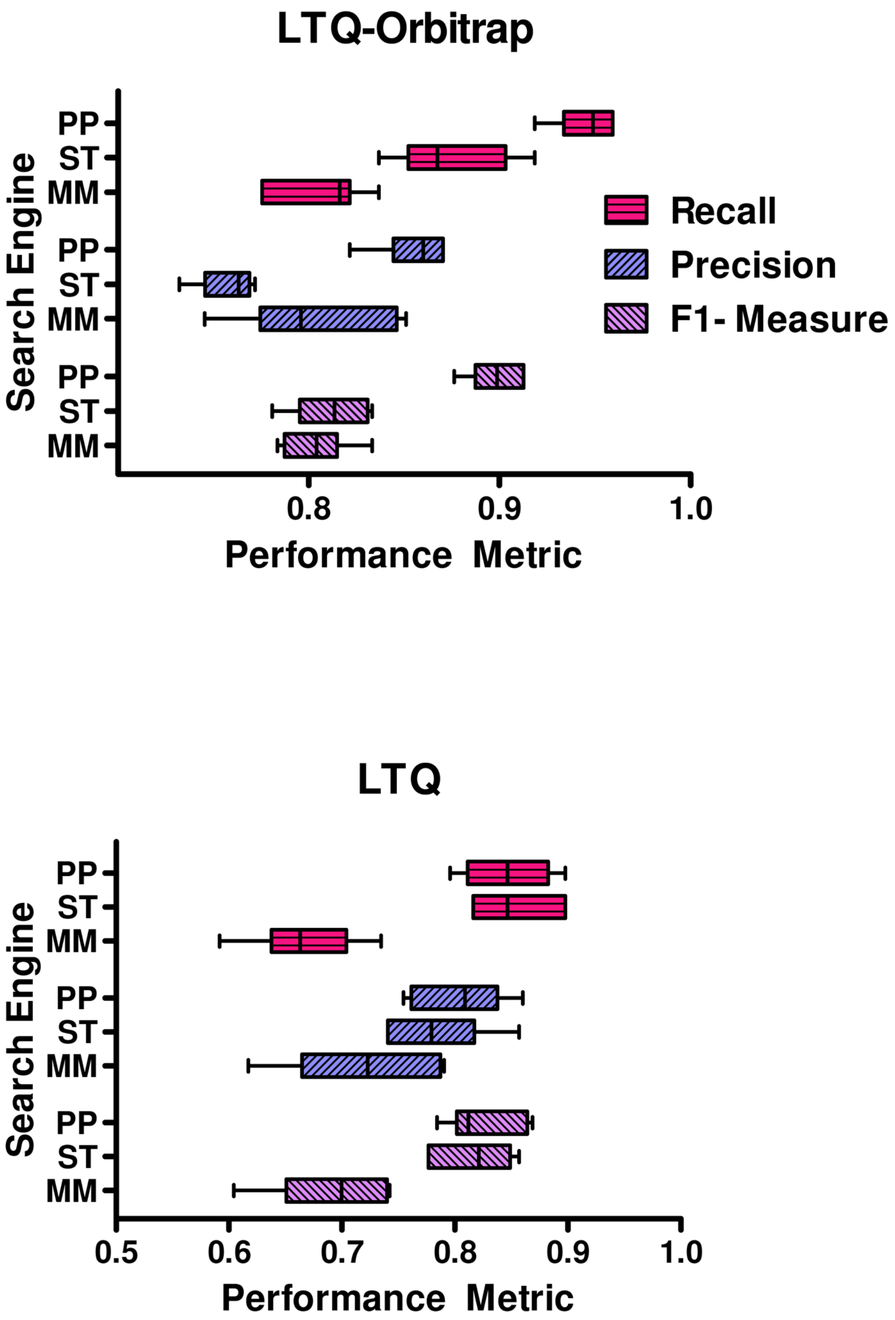


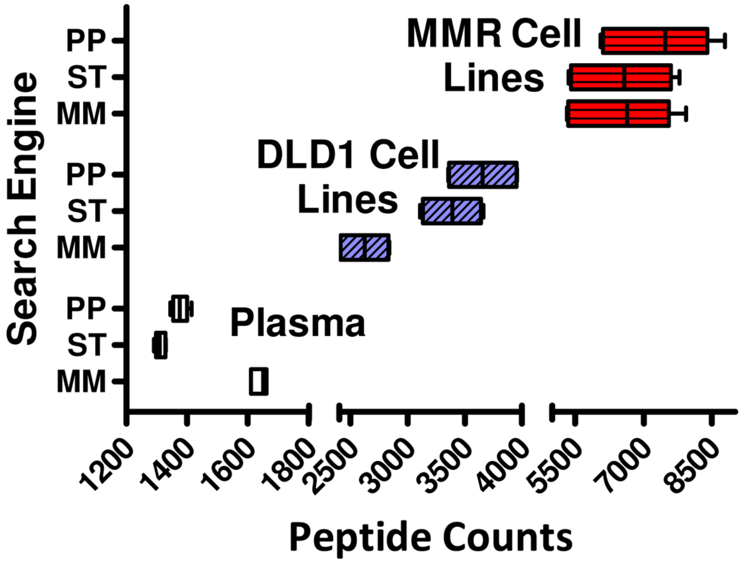
Figure 4. Performance on Universal Protein Standard (UPS1) Data Set

PP stands for Pepitome, ST stands for SpectraST, and MM stands for MyriMatch. IDPicker filtered the identifications at 5% FDR. Pepitome recovered more true positive proteins from the sample than any other search engine.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



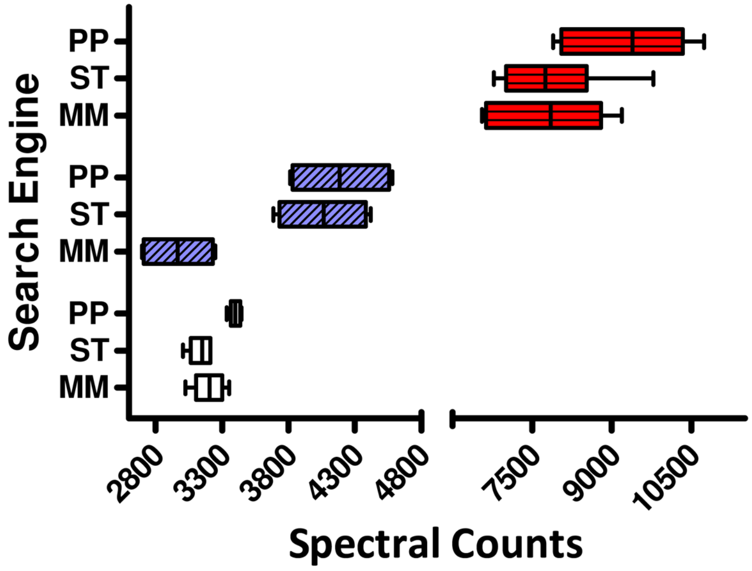


Figure 5. Performance on Real World Shotgun Proteomics Data Sets

Pepitome (PP) and SpectraST (ST) matched the experimental spectra against spectral libraries, whereas MyriMatch (MM) matched the MS/MS against a FASTA database. IDPicker filtered the identifications at 2% FDR. This figure illustrates the sample-wise summary of peptide and spectral identification numbers. Overall, Pepitome identified more peptides and spectra compared to other search engines.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

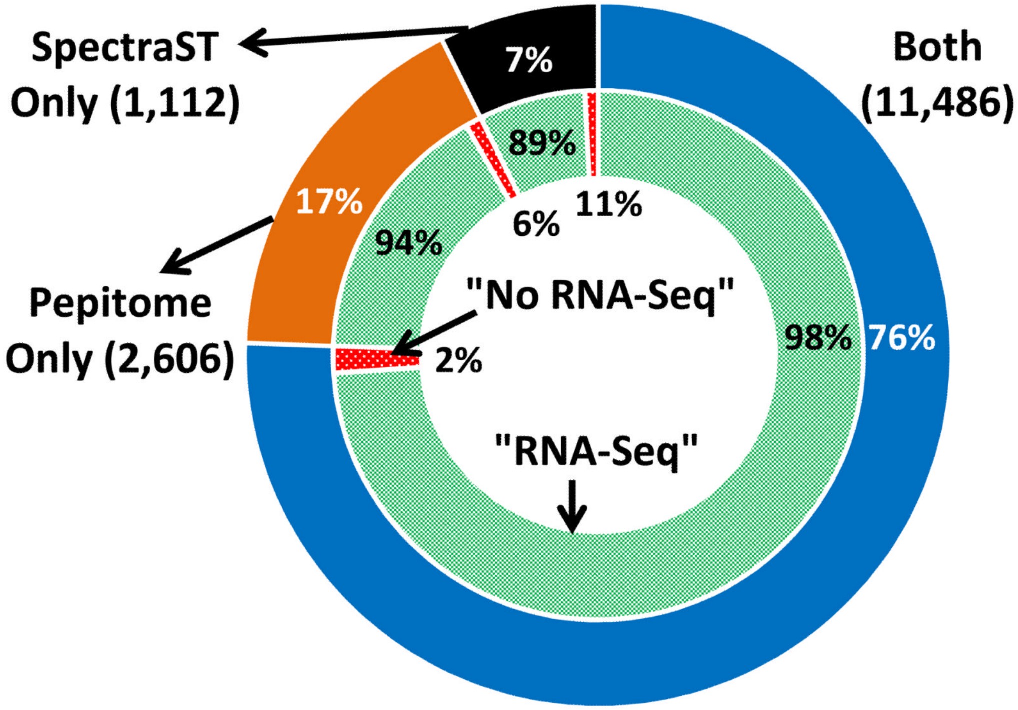


Figure 6. Gains from Spectral Library Searches are Trustworthy

MS/MS present in the MMR Cell Lines data set were identified with Pepitome and SpectraST. IDPicker filtered the results at 2% FDR. The peptide identification overlap between the search engines is shown in the figure. Identified peptide sequences were matched against the expressed proteome inferred from the RNA-Seq data. The percentage of peptides with corresponding RNA-Seq evidence is also presented in the figure.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

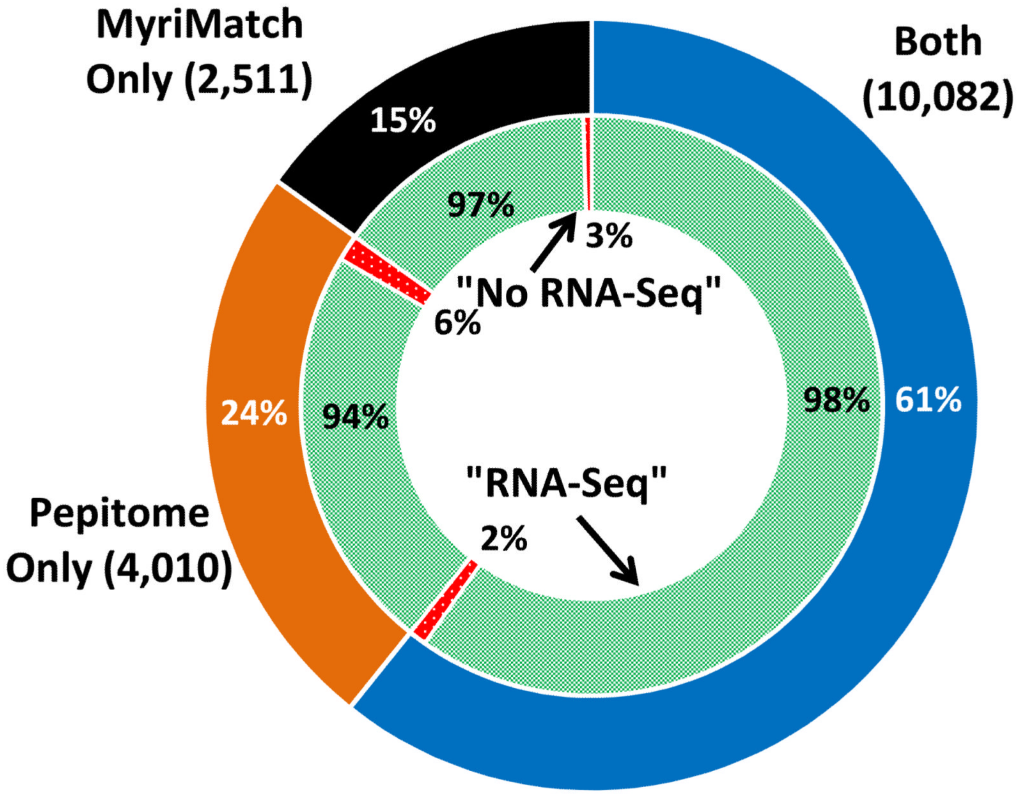


Figure 7. Spectral Library and Database Search Peptide Identification Overlap

MS/MS scans present in the MMR Cell Lines data set were identified with both a spectral library search engine (Pepitome) and a database search engine (MyriMatch). IDPicker filtered the results at 2% FDR. The peptide identification overlap between the spectral library and database searches is shown in the figure. Identified peptide sequences were matched against the expressed proteome inferred from the RNA-Seq data. The percentage of peptides with corresponding RNA-Seq evidence is also presented in the figure.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

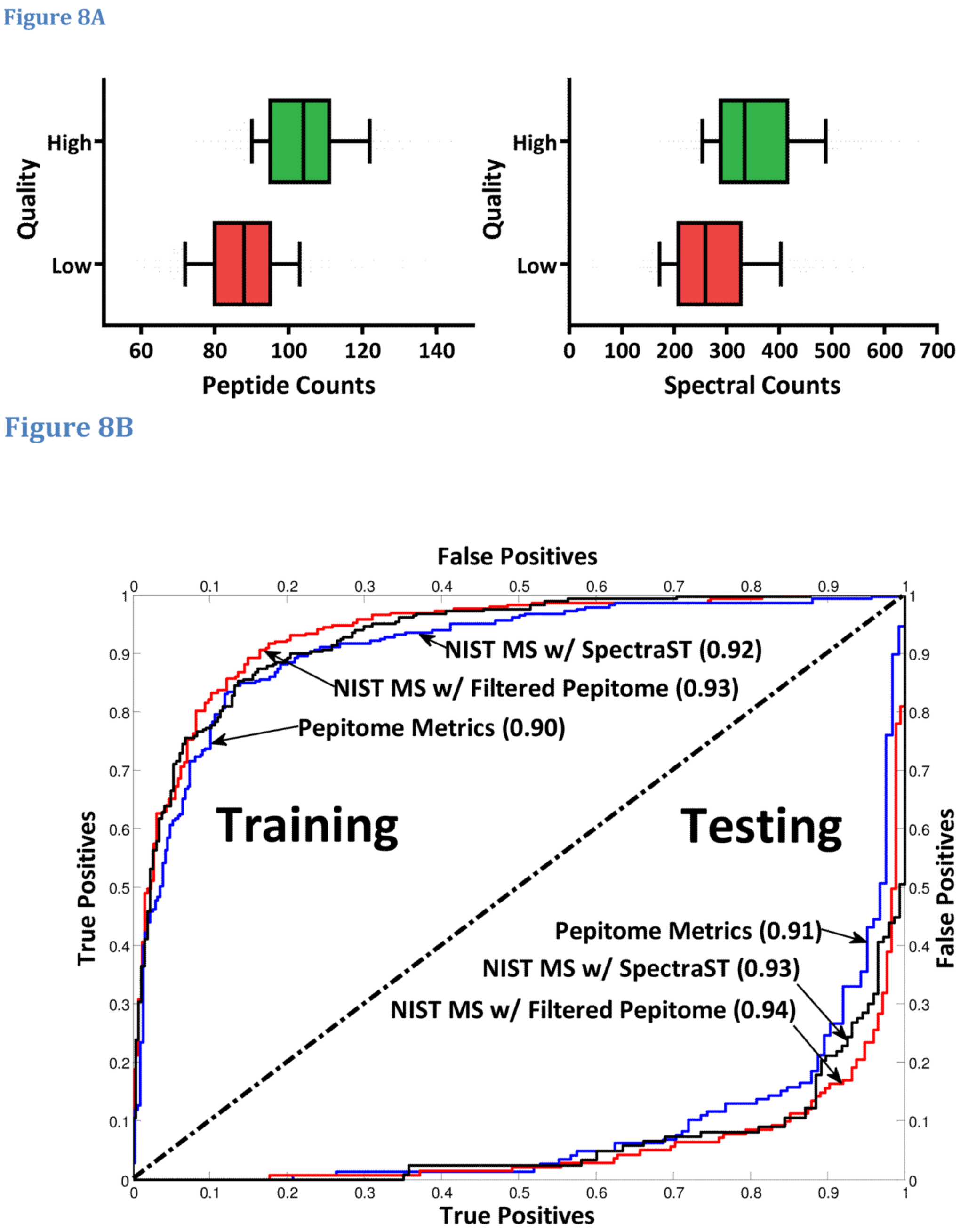


Figure 8. Quality Assessment (QA) Method for Shotgun Proteomics Data Sets

We performed 902 LC-MS/MS analyses of BSA standards on a LTQ-XL instrument spanning 18 months. Experts reviewed the raw files and assigned a quality label (low or high) to each file. Pepitome identified peptides from the samples and IDPicker filtered the identifications at 2% FDR **(a)** Peptide and spectral identification rates from quality assessed raw files. **(b)** We developed an artificial neural network (ANN) for recapitulating the expert QA of a raw file from a collection of quality metrics. This figure shows the training and testing receiver operating curves when the ANN is employing different categorical collections of quality metrics as inputs.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Data sets, Search Engines, Protein Sequence Databases, and Spectral Libraries Used in This Study.

					parent/fragment mass tolerances[d]		
data set	**instrument**	**# MS2 scans**	**sequence databases[a]**	**spectral libraries[c]**	**Pepitome**	**SpectraST**	**MyriMatch**
UPS1	LTQ	153,699	UniProt (2011_03)	NIST-Human-Iontrap	1.25/0.5	1.25	1.25/0.5
UPS1	LTQ-Orbitrap	63,202	UniProt (2011_03)	NIST-Human-Iontrap	10*/0.5	0.007	10*/0.5
DLD1 Cell Lines	LTQ	51,652	UniProt (2011_03)	NIST-Human-Iontrap	1.25/0.5	1.25	1.25/0.5
Plasma	QTOF	35,289	UniProt (2011_03)	NIST-Human-QTOF	0.1/0.1	0.1/0.2	20*/75*
MMR Cell Lines	LTQ-Orbitrap	486,252	UniProt (2011_03), ENSP-RKO[b], ENSP-SW480[b]	NIST-Human-Iontrap	10*/0.5	0.007	10*/0.5
BSA QA/QC	LTQ	10,737,907		NIST-BSA-Iontrap	1.25/0.5	1.25	

[a] Reversed protein sequences were appended to all sequence databases for estimating false discovery rates (FDRs).

[b] Customized sequence databases were assembled from the expressed transcripts (RNA-Seq) of the samples.

[c] All libraries were downloaded on 11/29/2010 and updated with spectra corresponding to common contaminant proteins. Decoy spectra were appended to the libraries for estimating FDRs.

[d] Asterisk denotes mass tolerances in parts-per-million.