Published in final edited form as:
Chem Res Toxicol. 2011 February 18; 24(2): 204–216. doi:10.1021/tx100275t.

Sequence tagging reveals unexpected modifications in toxicoproteomics

Surendra Dasari[1], **Matthew C. Chambers**[1], **Simona G. Codreanu**[2], **Daniel C. Liebler**[2,3], **Ben C. Collins**[4], **Stephen R. Pennington**[4], **William M. Gallagher**[5], and **David L. Tabb**[1,2,3,6,7]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-0006 [2]Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee 37232-0146 [3]Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232-6840 [4]School of Medicine and Medical Science, UCD Conway Institute, University of Dublin, Ireland [5]School of Biomolecular and Biomedical Science, UCD Conway Institute, University of Dublin, Ireland [6]Mass Spectrometry Research Center, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8575

Abstract

Toxicoproteomic samples are rich in posttranslational modifications (PTMs) of proteins. Identifying these modifications via standard database searching can incur significant performance penalties. Here we describe the latest developments in TagRecon, an algorithm that leverages inferred sequence tags to identify modified peptides in toxicoproteomic data sets. TagRecon identifies known modifications more effectively than the MyriMatch database search engine. TagRecon outperformed state of the art software in recognizing unanticipated modifications from LTQ, Orbitrap, and QTOF data sets. We developed user-friendly software for detecting persistent mass shifts from samples. We follow a three-step strategy for detecting unanticipated PTMs in samples. First, we identify the proteins present in the sample with a standard database search. Next, identified proteins are interrogated for unexpected PTMs with a sequence tag-based search. Finally, additional evidence is gathered for the detected mass shifts with a refinement search. Application of this technology on toxicoproteomic data sets revealed unintended cross-reactions between proteins and sample processing reagents. Twenty five proteins in rat liver showed signs of oxidative stress when exposed to potentially toxic drugs. These results demonstrate the value of mining toxicoproteomic data sets for modifications.

Introduction

Posttranslational modifications (PTMs) of proteins are receiving heightened attention from many biologists. Identification of PTMs by shotgun proteomics, however, is a challenge. Database search engines originally designed for peptide identification have been adapted to identity PTMs. For instance, the Sequest algorithm can search for a small number of known modifications (provided as a list of known masses and sequence specificities) (1). The Mascot error-tolerant approach automatically searches for a comprehensive list of known PTMs (2). Even though the underlying algorithms are very effective, database searches fail to identify large numbers of tandem mass spectra (MS/MS). Some of these spectra are unidentifiable because they are produced from chemical noise, but in toxicoproteomics, many spectra fail identification because they contain unexpected chemical and

[7]Corresponding author: (phone) 615-936-0380; (fax) 615-343-8372; david.l.tabb@vanderbilt.edu.

posttranslational modifications. We believe that searching for unanticipated mass shifts in toxicoproteomic data sets will reveal a wide palette of modifications that are missed by a standard database search.

Many informatics approaches have been developed for detecting unanticipated (blind) modifications from clinical samples (3–12). The *de novo* sequencing method infers full length sequences directly from the MS/MS. Inferred sequences are reconciled against peptides in the protein database while interpreting any mass differences between the two sequences as potential modifications (3,13). This method is not sensitive because even contemporary *de novo* sequencers (14) fail to interpret large portions of identifiable spectra. The MS-alignment (4) method, employed by the InsPecT (15) software, introduces arbitrary mass shifts in a database peptide while matching its predicted spectrum to an MS/MS. During recent years, partial sequence tagging has emerged as a sensitive method for detecting mutations and PTMs. The GutenTag (5) software automated the inference of sequence tags from MS/MS, enabling the detection of unanticipated modifications. The Tabb laboratory introduced the DirecTag (16) software for highly accurate tag inference, followed by the TagRecon software for mutant peptide detection through tag reconciliation (17). The spectral clustering method, exemplified by the Bonanza (11) software, detected unanticipated PTMs by examining the mass shift differences between unmodified peptide identifications and unidentified spectra. The “fraglet” method, exemplified by the ByOnic (12) software, matches database peptides to the MS/MS based on matching fragment peaks without matching precursor masses. The mass difference between the candidate matches is interpreted as a modification. All these methods have potential to detect important, yet unanticipated, modifications of proteins. Blind PTM searching, however, remains an exotic concept for many biologists.

We perceive several challenges blocking the broader adaptation of PTM mining for toxicoproteomic data sets. The first is that searching for known PTMs with database search engines is prohibitively time consuming. Next is that blind PTM searches via sequence tagging detect a variety of mass shifts on all types of amino acid residues; some of the mass shifts correspond to real PTMs and others are search artifacts. Currently, there is no user-friendly infrastructure for detecting ubiquitous mass shifts. Finally, both commercially available and open-source blind PTM search engines take enormous amounts of time for processing a single LC-MS/MS file.

In this study, we describe a new version of TagRecon for detecting both known and unknown PTMs present in toxicoproteomic experiments. TagRecon is part of an integrated bioinformatics pipeline containing a high-performance database search engine, a flexible protein assembler, and a user-friendly PTM results reviewer. The pipeline produces HTML and text reports of protein, peptide, and PTM identifications. Here, we compare TagRecon’s performance to the open-source InsPecT blind PTM search software. We analyzed three complex toxicoproteomic data sets and uncovered large numbers of unexpected PTMs that were missed by an initial standard database search. We demonstrate the advantage of TagRecon in detecting large numbers of known PTMs from LTQ and QTOF data sets.

Materials and Methods

Figure 1 illustrates the 3-step workflow used to identify PTMs from biological samples. In the first step, we reduce the size of the protein database using a standard database search (Figure 1A). For this, we employ the MyriMatch (18) database search engine configured to identify peptides from a comprehensive protein database (no exotic PTMs are considered). IDPicker (19,20) filters the resulting identifications using false discovery rate (FDR) analysis, and prepares a parsimonious list of proteins (subset FASTA). In the second step, a

blind PTM search strategy interrogates the shortlisted proteins for unanticipated modifications (Figure 1B). This strategy starts with DirecTag (16) inferring short sequence tags from MS/MS scans. TagRecon reconciles the inferred sequence tags against the protein database while making allowances for unanticipated mass shifts in peptides. A PTM results reviewer (PTMDigger) flags the most confident mass shifts. In the third step, a directed PTM search gathers additional evidence for the confident PTMs detected in the blind PTM search (Figure 1C). This optional directed PTM search strategy has parallels to the X! Tandem (21) refinement approach and Mascot error-tolerant search (2), but it is more efficient than either approach. One may question the benefit of following up an expansive blind PTM search with a narrow, directed PTM search. Blind searches can effectively find unanticipated mass shifts at a global (sample) level; however, they lack the sensitivity associated with a narrow search space. The source code and binaries of all the software used in the workflow are available for download from our website: http://fenchurch.mc.vanderbilt.edu/.

Improvement of TagRecon

TagRecon software was developed to detect mutant peptides from complex samples (17). In short, TagRecon starts by accepting three types of inputs: the raw MS/MS data, short sequence tags inferred from the raw MS/MS, and a user-supplied protein sequence database. The peptide identifications found for each input MS/MS file are written to a pepXML file. When TagRecon detects that a sequence tag matches a database peptide sequence, the software checks the flanking regions of both the spectrum and the peptide sequence to determine whether the masses match within a user-defined mass tolerance (Figure 2). The database sequence is considered as a potential match to the spectrum if either of the flanking masses is a match. The software allows for only one terminus mismatch to occur during this process. When a mass mismatch is detected the software computes the delta mass (ΔM) between the corresponding database sequence and spectrum flanking mass. In mutation mode (17), the observed ΔM is interpreted as an amino acid substitution using the BLOSUM62(22) matrix.

In this study, we introduced a new modification search mode in TagRecon for detecting chemical and posttranslational modifications of peptides. When running in this mode, the software interprets the mass mismatch between a spectrum sequence tag and a database peptide sequence as a potential modification. The sequence location of the modification is determined by one of the two methods: directed or blind (Figure 2). In the directed method, the ΔM and the amino acids in the mismatch region are used to identify potential locations from a “Preferred Delta Masses” lookup table. This table contains different combinations of user-supplied modifications. For instance, when the user specifies a Met-oxidation (+16) and a Lys-methylation (+14) and allows up to two modifications per peptide, the lookup table contains one Met-oxidation (+16), one Lys-methylation (+14), two Met-oxidations (+32), two Lys-methylations (+28), and one Met-oxidation plus one Lys-methylation (+30). The modification combinations are indexed by their associated mass shift and the affected amino acids. The ΔM is attached to all permissible amino acids in the mismatch region and multiple decorations are generated (one per permissible combination). In contrast, the blind method attaches the full ΔM to every single amino acid in the mismatch region (Figure 2). Each resulting full-length decoration is compared to the observed spectrum using the scoring algorithms embodied in MyriMatch (18). The highest scoring decoration is stored as the best interpretation for the spectrum.

For each peptide-spectrum match (PSM), TagRecon predicts a list of m/z values at which the software expects to find fragment ion peaks in the MS/MS. The software examines each of these locations and computes two probabilistic sub-scores: an intensity-based MVH (18) score and a mass error-based mzFidelity (17) score. The MVH score appraises the intensity

classes of fragments found at the expected m/z locations. The mzFidelity score measures how well the predicted fragment ions match the experimental peaks in m/z space. Depending on the search options, the software adds a small probabilistic bonus to the MVH score if the peptide termini match the protease digestion rules. The software uses the adjusted MVH score as the primary sort order for sequences, with mzFidelity acting as a tie-breaker. After all identifications are made for a spectrum the software computes fast cross-correlation (23) (XCorr) scores to independently validate the five best PSMs ranked by the MVH score. The XCorr is a non-probabilistic metric that measures the correspondence of predicted and observed spectra in the frequency domain.

Data Sets

Four different shotgun proteomics data sets were used to demonstrate the utility of TagRecon in detecting chemical and posttranslational modifications of proteins. Detailed sample processing protocols and Internet links for RAW data download are presented in Supplemental File 1.

Human Lens—This data set contains previously analyzed human ocular lens samples (3,24). In brief, proteins from a series of human lens samples were reduced, alkylated, and digested with trypsin. Resulting peptide mixtures were fractionated with strong cation exchange (SCX) chromatography; individual fractions were analyzed on a Thermo LCQ mass spectrometer using LC-MS/MS, and a total of 426,120 MS/MS spectra were collected. Previous studies (3,24) analyzed the MS/MS with OpenSea (3) and InsPecT (15) software configured to find unanticipated PTMs. Detected PTMs were rigorously attested following stringent attestation guidelines (24). In this study, we collected all the MS/MS data of the above mentioned samples into MGF files and reanalyzed them with TagRecon configured to find unanticipated PTMs.

THP1 Cell Lines—Human acute monocytic leukemia cell lines (THP1) were exposed to varying concentrations (0 μM, 5 μM, 10 μM, 20 μM, and 50 μM) of alkynyl-hydroxynonenal (Alk-HNE). Cells from each treatment were lysed, and adducted proteins were captured with a click reagent that bears azido and biotin groups separated by a photocleavable linker (25). Captured proteins were reduced with dithiothreitol (DTT), alkylated with iodoacetamide (IAM), and separated into 10 regions of a 1-D SDS gel. Proteins in each gel region were digested with trypsin and analyzed on a Thermo LTQ mass spectrometer using LC-MS/MS. A total of 574,361 MS/MS spectra were collected from all gel bands. Binary spectral data present in raw files were converted to mzML and MGF formats using the msConvert (26) tool of the ProteoWizard library.

DNA-Histones—Loecken et al (27) previously studied the mutagenic properties of two bis-electrophile compounds (1,2-dibromoethane and diepoxybutane) on DNA-Histone complexes. In brief, human recombinant histones H2b and H3 were incubated with varying concentrations of the two bis-electrophiles. Sites involved in DNA cross-linking were analyzed by adding a 16-mer oligonucleotide to the reactions. Proteins were reduced with DTT, alkylated with IAM, and digested with trypsin. Peptides were subjected to LC-MS/MS analysis on a Thermo LTQ-Orbitrap mass spectrometer, and 239,571 MS/MS spectra were collected. Binary spectral data present in the raw files were converted to mzXML format using msConvert, configured to compute and report accurate masses for the MS/MS precursors whenever possible and to centroid the MS scans.

Rat Liver—This data set was generated by the InnoMed PredTox consortium project on pharmaceutical toxicology (28). A colony of 150 rats was treated with either a placebo or one of five drugs (troglitazone and four other proprietary compounds donated by the

pharmaceutical partners) suspected of hepatotoxicity for either 3 or 14 days (28). Liver samples from all subjects were pooled for shotgun proteomic analysis. Proteins were reduced with DTT, alkylated with IAM, and digested with trypsin. The peptide mixture was separated into either 23 off-gel electrophoresis (OGE) fractions or 10 strong cation exchange (SCX) fractions. All OGE peptide fractions were analyzed in duplicate on an Agilent 6520 QTOF mass spectrometer using LC-MS/MS, whereas the SCX peptide fractions were analyzed only once. A total of 652,868 MS/MS spectra were collected from all LC-MS/MS experiments. Binary spectral data present in the raw files were converted to mzML and MGF formats using msConvert.

Bioinformatics Methods

Peptide and PTM Identification—The MS/MS scans present in the four data sets were identified using three different algorithms: MyriMatch (18), TagRecon, and InsPecT (15). Table 1 summarizes the data sets, protein sequence databases, and mass tolerances used in all searches. Detailed configuration parameters for all search engines are listed in Supplemental File 2.

We employed a three-step strategy to identify the unanticipated PTMs present in each data set (Figure 1). In the first step, we reduced the size of the protein database using a standard MyriMatch (18) database search (Figure 1A). MyriMatch was configured to derive semitryptic peptides from a comprehensive protein database and look for the following *in vitro* mass shifts as variable modifications: alkylation of cysteine (+57.01 Da), oxidation of methionine (+15.99 Da), formation of N-terminal pyroglutamate (−17.02 Da), and N-terminal acetylation (+42.01 Da). IDPicker (19,20) filtered the resulting PSMs at a 2% FDR, and proteins with at least two confident peptide identifications were included in a subset FASTA database. Reversed versions of the protein sequences were added to the subset FASTA database for estimation of FDRs. This reduction in the database size is necessary for lowering the search times of the blind modification searches, which are computationally expensive (29).

In the second step, TagRecon and InsPecT utilized the subset FASTA databases to search for unexpected PTMs present in the data sets (Figure 1B). TagRecon was configured to use the same variable modifications as MyriMatch search engine (see above) and derive semitryptic peptides from the FASTA databases. DirecTag (16) generated partial sequence tags for MS/MS scans from each raw file. The software was configured to generate the best 50 tags of three amino acids from each spectrum. TagRecon reconciled the inferred sequence tags against the protein database while making allowance for unanticipated mass shifts in peptides. InsPecT was configured to use a static mass shift of 57.02 Daltons for alkylated cysteines and search for unrestrictive modifications in the “blind” mode following the recommendations in its manual. InsPecT automatically identifies semitryptic peptides whenever possible. All identifications were processed in pepXML format. Peptide identifications from InsPecT software were converted into pepXML format using InsPecTToPepXML.py (part of the InsPecT package), which was slightly modified to report the instrument assigned “NativeID” scan identifiers for the MS/MS spectra. The peptide-protein associations in the InsPecT’s pepXML files were corrected using the RefreshParser tool (version 200912031530, Trans-Proteomics Pipeline, Institute of Systems Biology, Seattle, WA). IDPicker filtered the peptide identifications from the blind PTM searches at either 2% or 5% FDR. The PTMDigger software attested the detected mass shifts following very stringent guidelines (see “Attestation of PTMs using PTMDigger”).

In the third step, we performed a directed PTM search for the confident modifications detected in the blind PTM search (Figure 1C). For this, we employed MyriMatch and TagRecon configured to use the subset FASTA database and search for the chosen mass

shifts as either variable modifications or "Preferred Delta Masses." IDPicker filtered the results at 2% FDR. Our three-step PTM identification workflow has some parallels with Ning K *et al*'s comprehensive MS/MS identification protocol (30), with two key exceptions. First, spectra identified in the first step (database search) are not excluded from the subsequent steps (blind PTM search). Second, we subject all spectra to a blind PTM search regardless of their quality. Both of these measures safeguard against erroneously removing modified peptide spectra from the workflow before they can be identified.

Results Filtering Using IDPicker—IDPicker (19,20) filtered peptide identifications from all search engines at either a 2% or 5% false discovery rate (FDR). For MyriMatch, IDPicker was configured to automatically combine the MVH and mzFidelity scores for FDR filtering (20). InsPecT searches were filtered using a combined MQScore ("Match Quality") and DeltaScore.

IDPicker generates a discriminant score (F-score) for efficient filtering of TagRecon search results. This score is a weighted summation of the following PSM features: MVH, mzFidelity, XCorr, number of enzymatic termini (NET), number of missed cleavage sites (NMC), and number of PTMs in the peptide sequence. The weights of these features are determined once per pepXML file using a semi-supervised machine learning method implemented in the Percolator (31) software. According to this method, feature vectors are extracted from all top ranking decoy PSMs and labeled as negative examples. Simultaneously, positive examples are generated from the top ranking target PSMs. The feature vectors are temporarily segregated by charge state and their values are normalized to a uniform distribution. After normalization, a support vector machine (32) (SVM) learns an optimal feature weighting scheme that maximizes the separation between the positive and negative training examples. A 2-fold cross-validation strategy prevents over fitting of the F-score to a particular subset of the results. IDPicker re-ranks the PSMs of each spectrum using the learned F-score. After re-ranking, the software collects all top ranking PSMs from the input file, segregates the identifications by charge state, and filters them at either 2% or 5% FDR using the F-score. We did not attempt to construct an F-score for filtering InsPecT's search results because the software already employs a similar method to generate the "MQScore," and these changes would be beyond the score of this study.

Peptides passing the FDR thresholds were assembled into protein identifications using parsimony rules (19). Protein identifications with at least two distinct peptide identifications were considered for further analysis.

Attestation of PTMs Using PTMDigger—Blind PTM searches are very flexible, and they produce thousands of identifications containing potential modifications. Some of these modifications reflect the underlying biology of the sample, while others are artifacts of the search process. In this study, we applied proven PTM attestation principles (24) for validating peptide modifications present in complex mixtures. We created the PTMDigger software (Supplemental File 3) and a few Microsoft Excel 2007 macros to enforce the following filtering guidelines:

1. Modified interpretation of an MS/MS must improve the XCorr by at least 10%,
2. Mass shifts located at the peptide termini are rejected if they present any possibility for misinterpretation (for instance, modified peptide I.M+113ETIFEQIGR.K can be "explained away" by expanding the N-terminus to include the isoleucine residue),
3. Mass shifts with ambiguous sequence locations and no prior art (Pubmed references) are rejected,
4. Modifications of lysine and arginine residues cannot be trypsin cutting sites,

5. Modifications found in common contaminant proteins (like trypsin, keratin, etc.) are ignored, and
6. A modification must be detected in at least two peptides and four different spectra (potentially from different precursor charges). Exceptions were made to the two peptide rule if the modification satisfied all the above criteria and passed manual inspection (33).

Directed PTM Searches for Rat Liver and THP1 Cell Lines Data sets—IDPicker filtered the blind PTM search results from "Rat Liver" and "THP1 Cell Lines" data sets at 2% FDR. Detected modifications were attested following the above protocol, and the top seven most frequent mass shifts (by peptide count) observed in each sample were chosen for a directed PTM search. The LC-MS/MS files from the SCX fractions of the "Rat Liver" sample and 1D gel bands from the 50 μM treatment of "THP1 Cell Lines" sample were chosen as input spectra for the searches. TagRecon was configured to derive either semitryptic or fully tryptic peptides from the subset database and search the chosen mass shifts as "Preferred Delta Masses." MyriMatch was configured to derive fully tryptic peptides and search the chosen mass shifts as variable modifications (see Supplemental File 2 for details). IDPicker filtered the results from all searches at 2% FDR. The software employed an optimal combination of MVH and mzFidelity scores for FDR filtering (20).

IDPicker reports for all the searches performed in this study are provided in Supplemental File 4. Each of these reports contains a complete list of protein identifications, sequence coverage, peptide identifications, corresponding peptide-spectrum matches, detected modifications, and input files employed in the PTMDigger analysis.

Results and Discussion

TagRecon was designed to detect mutant peptides from complex samples (17). In this study, we retooled TagRecon for detecting chemical and posttranslational modifications of peptides. We introduced secondary scoring of matches to improve the sensitivity of TagRecon. Modifications made to the false discovery rate (FDR) filtering method improved the sensitivity and specificity of blind PTM searches. We compared TagRecon to the InsPecT algorithm because it constitutes a standard platform for detecting unanticipated modifications. We initially characterized the performance on human lens samples that are rich in known PTMs. We later shifted to toxicoproteomic data sets to demonstrate the value of blind PTM searching in the toxicology context.

Secondary Scoring of Peptide Matches

Search engines employ peptide scoring systems to measure the quality of peptide-spectrum matches (PSMs). For instance, TagRecon employs the MVH (18) scoring system to measure the fragment ion intensity matched by a peptide. This system separates the fragment peaks of a MS/MS into three discrete intensity classes (high, medium, and low). Peptides are scored based on the number of peaks matched in each intensity class. The MVH score can be rapidly computed and its efficacy was demonstrated in the context of both database searching (18) and sequence-tag searching (17). Splitting the continuous distribution of intensities into discrete classes, however, results in loss of information. To remedy this, we modified TagRecon to compute fast cross-correlation (23) (XCorr) scores for the top ranking peptide matches of each MS/MS. Unlike MVH, XCorr is an analog system for measuring the correspondence of fragment ion intensity with a peptide sequence. We measured the contribution of XCorr to the peptide identification rate of TagRecon. For this, we utilized the raw files from all three toxicoproteomic data sets ("DNA-Histones," "Rat Liver," and "THP1 Cell Lines"). For each raw file, TagRecon matched MS/MS to a

respective subset FASTA database while making allowances for unanticipated mass shifts in peptides (blind PTM search). IDPicker (20) filtered the resulting identifications at 5% FDR using either the MVH score alone or an optimized combination of MVH+XCorr scores. When MVH and XCorr were conjointly employed for results filtering, we observed a gain of 33%, 15%, and 17% in the peptide identification rates of "DNA-Histones," "Rat Liver," and "THP1 Cell Lines" samples, respectively.

Rescuing Unmodified Peptides from Blind PTM Searches via False Discovery Rate (FDR) Correction

In this study, we employed a target/decoy-based false discovery rate (FDR) method for filtering blind PTM search results. This naïve FDR method assumes that both modified and unmodified peptides receive similar search scores and produce overlapping score distributions. A single score threshold meeting the target FDR is derived for both classes of peptides. Modifications, however, alter the fragmentation patterns of peptides and produce low-scoring identifications compared to the unmodified peptides (Figure 3A). As a result, modified peptides in a blind PTM search are at a higher risk of being false positive compared to unmodified peptides. Also, blind PTM searches produce numerous modified peptide identifications compared to unmodified identifications (compare the areas of the modified and unmodified peptide score distributions in Figure 3A). Both these factors force the user to filter the blind PTM search results at a stringent score threshold, which decreases the identification sensitivity of unmodified peptides. An obvious solution to this problem is to separate the modified and unmodified identifications into two classes. FDR filtering is applied on both classes separately. This "divide and filter" FDR method derives different score thresholds for filtering modified and unmodified peptides. This method, however, underestimates the FDRs of unmodified peptides because blind PTM searches produce small numbers of unmodified decoy identifications (Figure 3A).

We systematically investigated the reasons behind the paucity of unmodified decoy identifications in blind PTM searches. The results indicate that blind PTM searches make more mistakes when identifying modified peptides compared to unmodified peptides (Supplemental Table 1). As a result, the probability of an unmodified decoy identification passing a score threshold producing matching target identification is negligible in blind PTM searches (Supplemental Table 1). This is not taken into account by the naïve FDR method, which estimates the number of false positives associated with a score threshold by doubling the number of decoys passing the same threshold. We introduced a correction factor (K) in the FDR computation that takes into account the probabilities of modified and unmodified decoy peptides producing matching target matches above a score threshold:

$$FDR(P_S)=\frac{K(P_S)\times R}{(R+F)} \mid \mathrm{K(unmodified)}=\frac{(\mathrm{T_U}+\mathrm{D})}{\mathrm{D}} \; and \; \mathrm{K(modified)}=\frac{(\mathrm{T_M}+\mathrm{D})}{\mathrm{D}}$$

where P_S is a peptide passing a score threshold S; R and F are number of decoys and targets passing the score threshold S, respectively; D is total number of decoy comparisons (PSMs) made for all MS/MS in the raw file; T_U and T_M are the total number of unmodified and modified target comparisons made for all MS/MS in the raw file, respectively (Typical values for K (P_S) are given in Supplemental Table 1). We chose to compute the FDR correction factors at a global level (raw file) instead of at each MS/MS for two reasons. First, the FDR is defined over all top-ranking PSMs in a raw file. FDR of a PSM is expressed as the percentage of top-ranking decoy PSMs in a raw file that have a similar or better score. Second, local correction for FDR is only appropriate when the ratio of target to decoy comparisons varies significantly from spectrum to spectrum, which is not true for sequence tag searches. This revised method computes the FDR of both modified and

unmodified peptides using the same set of decoys, but the correction factor lowers the score threshold for filtering unmodified peptides. We emphasize, however, that the corrected method generates more liberal FDR estimates for unmodified peptides compared to the naïve method.

We wanted to see whether the correction factor introduced in the naïve FDR method improves unmodified peptide identification rates of blind PTM searches. For this test, we randomly chose a single RAW file from the "Rat Liver" data set and performed a blind PTM search with TagRecon. IDPicker employed a static MVH score for filtering the identifications at 5% FDR (Figure 3B). The naïve FDR method recovered more unmodified peptide identifications when employing the correction factor whereas modified peptide identification rates are unchanged (Figure 3B). Similar results were observed for other TagRecon scoring metrics and also all other raw files used in this study (data not shown).

Improved Score Combination for Filtering Blind PTM Search Results

TagRecon produces three orthogonal score metrics for each PSM: MVH, mzFidelity, and XCorr. IDPicker reads the score tuples of all top-ranking PSMs in a pepXML file, determines an optimal score combination via Monte Carlo simulation, and filters the results at a given FDR (20). This system works well in the context of standard protein identification searches and single amino acid mutation searches (17). Blind PTM searches, however, place undue burdens on peptide scoring systems, making it necessary to conscript auxiliary properties of PSMs (such as number of enzymatic termini, number of modifications, etc.) for results filtering. We modified IDPicker software to generate an F-score discriminant for filtering the TagRecon blind PTM search results. The F-score is a weighted summation of the following PSM properties: MVH, mzFidelity, XCorr, number of enzymatic termini (NET), number of missed cleavage sites (NMC), and number of PTMs. The software employs the Percolator (31) algorithm to learn the F-score discriminant from a subset of the search results. IDPicker re-ranks the best five peptide matches for each spectrum according to the learned discriminant, reads the new top-ranking PSMs from the pepXML file, and filters the results at a given FDR using the F-score.

We wanted to test whether the F-score can improve results filtering over the traditional score combination approach. For this, we utilized the raw files present in all three toxicoproteomic samples. TagRecon was configured to match the MS/MS against respective subset FASTA databases while making allowances for blind PTMs in peptides. IDPicker filtered the results at 5% FDR using either a static F-score or an optimal combination of MVH, mzFidelity, and XCorr metrics. When F-score filtered the results, we observed a net gain of 3.3%, 9.4%, and 13.4% in the identification rates of "DNA-Histones," "Rat Liver," and "THP1 Cell Lines" samples, respectively (Figure 4).

Testing PTM Recall with Human Lens Samples

We wanted to compute the fraction of known PTMs that are detectable by a TagRecon blind PTM search. For this test, we analyzed PTM-rich human lens data sets ("Human Lens") that were thoroughly characterized by multiple studies (3,4,24). We scanned the associated literature and derived a confident list of 125 PTM sites (Supplemental Table 2). More than half the sites in the PTM list are supported by at least two independent studies. We challenged the TagRecon's blind PTM search mode to find these well known PTM sites. TagRecon matched the MS/MS against semitryptic peptides derived from the respective subset FASTA database while making allowances for unanticipated mass shifts in peptides. IDPicker filtered the resulting identifications at 2% FDR. The blind PTM search confirmed 114 out of 125 known PTM sites in lens (Recall = 0.91). The software missed 11 sites, 9 of which are located close to the peptide termini. Semitryptic blind PTM searches often

encounter problems when identifying terminal modifications because the software can either grow or shrink the peptide sequence and change the associated mass shift. Supplemental Table 2 lists all the known PTM sites in human lens that were either confirmed or unconfirmed by TagRecon. The software also discovered 41 modification sites that fell beyond the confident list.

Search Engine Performance Comparisons

We compared the performance of TagRecon and InsPecT when searching for unexpected modifications present in the three toxicoproteomic samples. Both search engines were configured to match the spectra against respective subset FASTA databases while making allowances for unanticipated mass shifts in peptides. IDPicker filtered the resulting identifications at 2% FDR. Table 2 presents a sample-wise summary of proteins, peptides, and spectra identified by the respective search engines along with estimated global FDR for peptides containing unanticipated modifications. Overall, TagRecon recovered more proteins, peptides, and spectra from the samples than InsPecT (Table 2). Oddly, InsPecT reported half as many protein IDs as TagRecon in the “DNA-Histones” data set even though the tools identified similar numbers of peptides (Table 2). The source of this discrepancy stems from incomplete reporting of peptide-protein associations by the InsPecT analysis pipeline. Our attempts to track and fix the source of this error in the pipeline were unsuccessful. This issue underscores the importance of ensuring search engines compliance with industry standard output formats. While the peptide and spectral IDs reported by InsPecT were valid, the translation to pepXML lost some peptide/protein associations.

We computed the peptide overlap between TagRecon and InsPecT in the context of blind PTM searches. On average, 64% of peptides were identified by both search engines, 23% were identified by TagRecon but missed by InsPecT, and 12% were identified by InsPecT but missed by TagRecon. The peptide overlap between search engines improved when considering only unmodified peptides: 78% identified by both search engines, 18% identified exclusively by TagRecon, and 4% identified exclusively by InsPecT. However, an opposite effect was observed for modified peptides: only 42% were identified by both search engines, 33% by TagRecon only, and 25% by InsPecT only. A majority of this difference stemmed from the disagreements between the search engines about the sequence location of modifications, confirming the importance of modification localization for this area of research. The average global FDR for modified peptide identifications was slightly higher than the FDR threshold employed for initial filtering (2.5% for TagRecon and 2.1% for InsPecT) while unmodified peptide identifications had a much lower average global FDR (0.5% for both TagRecon and InsPecT). This highlights the fact that modified peptides are more prone to false discoveries and PTMs should be strictly attested with auxiliary evidence.

We compared the computational speed of TagRecon and InsPecT using a computer equipped with dual quad-core 2.4-GHz Intel processors, 12GB of RAM, and a CentOS Linux OS (kernel version 2.6). All timing statistics were gathered by confining TagRecon to a single core of the computer to match usage for InsPecT. For this test, we randomly selected 10K LTQ spectra from the “THP1 Cell Lines” data set and 5K protein sequences from the corresponding subset FASTA file. Both search engines derived semitryptic peptides from the proteins and searched for unanticipated modifications in the peptides. TagRecon required 53m43s to sequence tag and match the selected spectra, where as InsPecT took 9h55m43s. When searching for unanticipated modifications in this dataset, TagRecon is approximately 10 times faster than InsPecT. Computational times for regular protein identification searches were compared elsewhere (17).

Our attempts to include SIMS (10), MOD[i](9), Popitam (6), and ByOnic (12) in the comparisons were unsuccessful because of two reasons. First, the search tools produced non-standard output files and extracting modification sequence location information from these files was difficult. Second, some of the tools had limited accessibility via Internet websites.

Frequency and Abundance of Unanticipated Modifications in Toxicoproteomic Samples

Toxicology researchers often acknowledge the presence of unanticipated modifications in treated biological samples. The frequency and abundance of these variants, however, is not fully appreciated because they are invisible to a standard database search. We employed TagRecon to identify peptides with unanticipated modifications from “DNA-Histones,” “Rat Liver,” and “THP1 Cell Lines” samples. All these samples are expected to contain both unknown *in vivo* PTMs and also unanticipated *in vitro* modifications that are byproducts of toxicological treatment. IDPicker filtered the peptide identifications at 2% FDR. Resulting identifications were subjected to the strict PTM attestation guidelines outlined in the Materials and Methods section. Table 3 presents the percentage of peptides (frequency) and spectra (abundance) containing an attested unknown modification. The frequency and abundance of unanticipated modifications depends on the sample type. For instance, the “DNA-Histones” sample had the highest observed frequency of unanticipated modifications, whereas the “THP1 Cell Lines” sample had the lowest (Table 3). This is expected because, unlike the complex cell line sample, the “DNA-Histones” sample is comprised of a simple mixture of heavily modified proteins.

THP1 Cell Lines Data Set

This data set represents human acute monocytic leukemia (THP1) cell lines exposed to varying concentrations of alkynyl-hydroxynonenal (Alk-HNE). After the treatment, adducted proteins were captured with a click reagent that bears azido and biotin groups separated by a photocleavable linker. Captured proteins were subjected to GeLC-MS/MS analysis. We employed TagRecon to find unanticipated modifications present in the captured proteins. IDPicker filtered the resulting identifications at 2% FDR. Detected modifications were attested following the strict PTM attestation guidelines outlined in the Materials and Methods section (Table 4).

The 1D-gel images showed a direct correlation between the abundance of the captured proteins and Alk-HNE concentration used to treat the cells (Supplemental File 5). Hence, we anticipated seeing numerous HNE adducts (ΔM = 156 or 158 Da) on the free cysteines of the captured proteins. Contrary to our expectations, acrylamide (34) and carbamidomethyl +DTT (35) adducts emerged as primary reactants for free cysteines (Table 4). Four peptides, however, contained a HNE adduct that was attached to a linker-derived fragment (ΔM = 311 Da) (25). This suggests that very few free cysteines in the captured proteins react with HNE, leaving the rest as targets for other sample processing reagents. Also, the abundance (spectral count) of the HNE adducted peptides is miniscule compared to rest of the peptide matrix, making them relatively hard to detect (Table 4). We are conducting additional experiments to improve the visibility of HNE adducted peptides.

The most abundant modification discovered in this sample is an unexpected N-terminal adduct with a mass shift of 26±0.02 Daltons (Table 4). We observed a positive correlation (R^2=0.98) between the abundance of this unexpected adduct and the Alk-HNE concentration used to treat the cells (Supplemental File 6). The sample processing protocol, however, prevents us from explaining the unexpected adduct as acetaldehyde byproducts of HNE enhanced lipid peroxidation (36). We suspect that the observed mass shift corresponds to a derivative of some unknown compound employed in the 1D-gel electrophoresis. Further

experiments are needed to assign an identity for this adduct. A standard database search fails to detect such unanticipated chemistry. In contrast, “blind PTM” searches routinely detect unanticipated mass shifts in peptides, providing clues about undesirable sample chemistries.

Rat Liver Data Set

This data set contains liver proteins extracted from a colony of rats treated with potentially hepatotoxic drugs. The protein mixture was subjected to shotgun proteomic analysis. We employed TagRecon to find unanticipated modifications present in the sample. IDPicker filtered the resulting identifications at 2% FDR. Detected modifications were attested following strict attestation guidelines outlined in the Materials and Methods section (Table 5).

Acetone modification of peptides was the most abundant mass shift discovered in the “Rat Liver” sample (Table 5). This modification selectively targeted peptides containing a glycine residue at the second position (*XGXn* motif). One possible explanation is that the residual acetone from the protein precipitation step reacted with peptides containing a *XGXn* motif, forming a stable imidazolidinone ring (37). We advocate judicious use of acetone for precipitating proteins because of two reasons. First, unanticipated modifications confound the quantification methods by splitting the intensity signal across multiple peptide forms. Second, the acetone modification consistently suppressed the precursor fragmentation, producing low quality MS/MS that could be potentially misidentified (see Supplemental File 7A for an example).

We discovered a total of 33 cysteine modification sites in 25 proteins; 23 of the detected sites contained a mass shift of 134±0.04 Daltons and 10 sites had a mass shift of 25±0.03 Daltons (Supplemental File 8). Eighteen of the 25 proteins have known drug interactions in the Target Protein Database (38) (Supplemental File 8). An independent Gene Ontology (39) analysis also confirmed that a majority of the oxidized cysteines were located in mitochondrial and cytoplasmic proteins that are known to metabolize drugs and lipids (Supplemental File 8). Confirming previous reports (40, 41), we detected the oxidation of the Cys50 residue in microsomal glutathione s-transferase 1 (MGST1) protein, which is a known drug metabolite (Figure 5). All of this evidence suggests that the observed cysteine adducts correspond to reactive metabolites of the drug compounds given to the rats. As we do not have pertinent information regarding the existence and characterization of potential drug metabolites (28), the identities of the observed cysteine adducts remain unknown.

We also observed a −2±0.01 Dalton mass shift on Cysteine residues (approximately equal to the loss of two hydrogen atoms). Such low mass modifications are generally “explained away” as precursor mass misassignments. In this case, however, the Cys-2 modification was observed only in peptides containing two unalkylated cysteine residues, suggesting the possibility of an intra-peptide disulphide bridge (15). As further evidence to this hypothesis, MS/MS fragmentation of peptide bonds that lay between the two disulphide-bridged residues was suppressed (Supplemental File 7B) (15,42). We believe that the observed intra-peptide disulphide bridges are sample processing artifacts because fully alkylated forms of the same peptides were also observed (data not shown).

DNA-Histones Data Set

This shotgun data set was derived from purified DNA-Histone complexes that were incubated with two bis-electrophile compounds: 1,2-dibromoethane and diepoxybutane. We employed TagRecon to find unanticipated modifications present in the data set. IDPicker filtered the resulting identifications at 2% FDR. Detected modifications were attested following strict attestation guidelines outlined in the Materials and Methods section. As

anticipated, we detected several acetylated and carbamylated lysine residues in histone proteins (Table 6). Confirming previous reports (27), unprotected lysine residues in histones were prime targets of diepoxybutane modifications (Table 6). Also, the Lys80 residue in histone H3 cross-linked to DNA in the presence of 1,2-dibromoethane (Figure 6). Besides the known modifications, we also detected three unanticipated modifications in the sample: 70.013 Daltons on lysines, as well as 223.042 Daltons and 75.998 Daltons on cysteines.

Toxicology researchers often ignore the possibility of treatment agents cross-reacting with *in vivo* PTMs. Such reactions, however, can happen in real life samples. For instance, the unexpected Lys+70 modification was observed exclusively on lysine residues that were also found to be dimethylated (+27.989 Daltons) and only in samples treated with 1,2-dibromoethane (+44.026 Daltons). This suggests an unknown reaction between 1,2-dibromoethane and dimethylated lysine residues. Bandeira et al (43) observed the Cys+223 modification when analyzing antibodies raised against B- and T-cell lymphocyte attenuator molecule (BTLA), which suggests that the Cys+223 modification observed in this study is not related to the bis-electrophile treatment. The true source of the Cys+223 and Lys+76 modifications remains elusive.

What is The Best Strategy for Refinement Searches?

Refinement searches recover modified peptides from proteins that have already been identified by unmodified peptide forms. Traditionally, these searches are performed by a database search engine configured with a large palette of variable modifications (5–10 PTMs). This leads to an exponential increase in the number of sequences compared to the MS/MS, resulting in untenable search times. In this study, we introduced a “Directed PTM” search mode in TagRecon, which does not explicitly enumerate all possible modification variants of a peptide sequence (Figure 2).

We wanted to compare the performance of the “Directed PTM” search to that of a variable modification search. For this, we chose ten SCX-LC-MS/MS files from the “Rat Liver” data set (QTOF) and 10 GeLC-MS/MS files from “THP1 Cell Lines” data set (LTQ). For each data set, MyriMatch identified the proteins present in the sample with a standard database search. TagRecon interrogated the identified proteins (using a subset FASTA) for unanticipated modifications. Detected mass shifts were attested, and the top seven most frequent modifications were selected for a subsequent expanded modification search. MyriMatch was configured to search for the selected mass shifts as variable modifications while deriving fully tryptic peptides from the database. TagRecon searched the selected mass shifts as “Preferred Delta Masses” while deriving either fully tryptic or semitryptic peptides from the database (see Materials and Methods section for details). Table 7 scrutinizes the performance characteristics of the searches. Directed PTM searches identified more peptides and spectra than comparable variable modification searches (Table 7). As expected, directed PTM searches were faster than variable modification searches (Table 7). Even though we searched for seven mass shifts in both data sets, MyriMatch search times for “Rat Liver” data set (148K QTOF spectra against 4K proteins) are longer than the “THP1 Cell Lines” data set (160K LTQ spectra against 7.5K proteins). This is because two of the mass shifts chosen for the “Rat Liver” data set have multi-residue specificities. In contrast to variable modification searches, computational time for directed PTM searches is independent of the number of modifications employed in the search and their residue specificities (Supplemental File 9). With this evidence, we believe that sequence tag-based search engines are the future of PTM hunting.

Conclusion

Here we show that TagRecon can identify both known and unanticipated modifications from complex mixtures. When configured to identify known PTMs, TagRecon identified more unique peptides from the samples than the MyriMatch database search engine. TagRecon outperformed InsPecT when identifying unanticipated PTMs from samples. The advances made in the scoring system and results filtering improved the sensitivity of PTM searches. We developed user-friendly software for attesting and browsing the PTM search results. We encapsulated TagRecon in an open-source computational pipeline that can be used for routine, large-scale identification of modified peptides.

Sequence tagging was introduced in 1994 for detecting variant peptides. Biologists, however, perceive it as an exotic concept, and they still use standard database searching for identifying PTMs. We believe that recent advances made in automated tag inference and reconciliation have paved the way for sequence tagging to assume its role as a powerful PTM hunter. Although plenty of work remains to make PTM hunting routine, we believe that the software infrastructure we have made available is a solid step in the right direction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

D.L. Tabb and M.C. Chambers were supported by NIH grants R01 CA126218 and U24 CA126479. S. Dasari was supported by NIH grant R01 CA126218. We thank Phillip A. Wilmarth from the Oregon Health & Science University (Portland, OR) for providing the lens data set and also a list of known PTM sites. The “Rat Liver” data set was contributed courtesy of the InnoMed PredTox Consortium, which was funded by the European Union's Sixth Framework Program. We also thank Elisabeth M. Loecken and F. Peter Guengerich from Vanderbilt University (Nashville, TN) for providing the “DNA-Histones” data set. The NIH grant P01 ES013125 facilitated the collection of the “THP1 Cell Lines” data set. Finally, we would like to thank the anonymous reviewers for their insightful comments.

References

1. Yates JR, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. Anal Chem 1995;67:1426–1436. [PubMed: 7741214]
2. Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2002;2:1426–1434. [PubMed: 12422359]
3. Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, Nagalla SR. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. J. Proteome Res 2005;4:546–554. [PubMed: 15822933]
4. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nat Biotech 2005;23:1562–1567.
5. Tabb DL, Saraf A, Yates JR. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 2003;75:6415–6421. [PubMed: 14640709]
6. Hernandez P, Gras R, Frey J, Appel RD. Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. PROTEOMICS 2003;3:870–878. [PubMed: 12833510]
7. Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. Mol. Cell Proteomics 2006;5:935–948. [PubMed: 16439352]

8. Na S, Jeong J, Park H, Lee K, Paek E. Unrestrictive Identification of Multiple Post-translational Modifications from Tandem Mass Spectrometry Using an Error-tolerant Algorithm Based on an Extended Sequence Tag Approach. Mol. Cell Proteomics 2008;7:2452–2463. [PubMed: 18701446]
9. Kim S, Na S, Sim JW, Park H, Jeong J, Kim H, Seo Y, Seo J, Lee K, Paek E. MODi : a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. Nucl. Acids Res 2006;34:W258–W263. [PubMed: 16845006]
10. Liu J, Erassov A, Halina P, Canete M, Vo ND, Chung C, Cagney G, Ignatchenko A, Fong V, Emili A. Sequential Interval Motif Search: Unrestricted Database Surveys of Global MS/MS Data Sets for Detection of Putative Post-Translational Modifications. Anal Chem 2008;80:7846–7854. [PubMed: 18788753]
11. Falkner JA, Falkner JW, Yocum AK, Andrews PC. A Spectral Clustering Approach to MS/MS Identification of Post-Translational Modifications. J. Proteome Res 2008;7:4614–4622. [PubMed: 18800783]
12. Bern M, Cai Y, Goldberg D. Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. Anal Chem 2007;79:1393–1400. [PubMed: 17243770]
13. Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with de novo sequencing error. J Bioinform Comput Biol 2005;3:697–716. [PubMed: 16108090]
14. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem 2005;77:964–973. [PubMed: 15858974]
15. Tanner S, Shu H, Frank A, Wang L, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 2005;77:4626–4639. [PubMed: 16013882]
16. Tabb DL, Ma Z, Martin DB, Ham AL, Chambers MC. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. J. Proteome Res 2008;7:3838–3846. [PubMed: 18630943]
17. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AL, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. J. Proteome Res 2010;9:1716–1726. [PubMed: 20131910]
18. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J. Proteome Res 2007;6:654–661. [PubMed: 17269722]
19. 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]
20. Ma Z, 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 Discrimination Peptide Identification Filtering. J. Proteome Res 2009;8:3872–3881. [PubMed: 19522537]
21. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–1467. [PubMed: 14976030]
22. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A 1992;89:10915–10919. [PubMed: 1438297]
23. Eng JK, Fischer B, Grossmann J, MacCoss MJ. A Fast SEQUEST Cross Correlation Algorithm. J. Proteome Res 2008;7:4598–4602. [PubMed: 18774840]
24. Wilmarth PA, Tanner S, Dasari S, Nagalla SR, Riviere MA, Bafna V, Pevzner PA, David LL. Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? J. Proteome Res 2006;5:2554–2566. [PubMed: 17022627]
25. Kim HH, Tallman KA, Liebler DC, Porter NA. An Azido-Biotin Reagent for Use in the Isolation of Protein Adducts of Lipid-derived Electrophiles by Streptavidin Catch and Photorelease. Mol. Cell Proteomics 2009;8:2080–2089. [PubMed: 19483245]
26. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 2008;24:2534–2536. [PubMed: 18606607]

27. Loecken EM, Dasari S, Hill S, Tabb DL, Guengerich FP. The bis-electrophile diepoxybutane cross-links DNA to human histones but does not result in enhanced mutagenesis in recombinant systems. Chem. Res. Toxicol 2009;22:1069–1076. [PubMed: 19364102]
28. Adler M, Hoffmann D, Ellinger-Ziegelbauer H, Hewitt P, Matheis K, Mulrane L, Gallagher W, Callanan J, Suter J, Fountoulakis M, Dekant W, Mally A. Assessment of candidate biomarkers of drug-induced hepatobiliary injury in preclinical toxicity studies. Toxicol Lett 2010;196:1–11. [PubMed: 20362651]
29. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. Nat. Biotechnol 2005;23:1562–1567. [PubMed: 16311586]
30. Ning K, Fermin D, Nesvizhskii AI. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. Proteomics 2010;10:2712–2718. [PubMed: 20455209]
31. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Meth 2007;4:923–925.
32. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Adv Neur In 1996;9:155–161.
33. Tabb DL, Friedman DB, Ham AL. Verification of automated peptide identifications from proteomic tandem mass spectra. Nat. Protocols 2006;1:2213–2222.
34. Bordini E, Hamdan M, Righetti PG. Probing acrylamide alkylation sites in cysteine-free proteins by matrix-assisted laser desorption/ionisation time-of-flight. Rapid Commun. Mass Spectrom 2000;14:840–848. [PubMed: 10825247]
35. Chalkley RJ, Baker PR, Medzihradszky KF, Lynn AJ, Burlingame AL. In-depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types. Mol. Cell Proteomics 2008;7:2386–2398. [PubMed: 18653769]
36. Hartley DP, Petersen DR. Co-Metabolism of Ethanol, Ethanol-Derived Acetaldehyde, and 4-Hydroxynonenal in Isolated Rat Hepatocytes. Alcohol Clin Exp Res 1997;21:298–304. [PubMed: 9113267]
37. Simpson DM, Beynon RJ. Acetone Precipitation of Proteins and the Modification of Peptides. J. Proteome Res 2010;9:444–450. [PubMed: 20000691]
38. Hanzlik R, Koen Y, Theertham B, Dong Y, Fang J. The reactive metabolite target protein database (TPDB) - a web-accessible resource. BMC Bioinformatics 2007;8:95. [PubMed: 17367530]
39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25–29. [PubMed: 10802651]
40. Shin N, Liu Q, Stamer SL, Liebler DC. Protein Targets of Reactive Electrophiles in Human Liver Microsomes. Chem. Res. Toxicol 2007;20:859–867. [PubMed: 17480101]
41. Tzouros M, Päähler A. A Targeted Proteomics Approach to the Identification of Peptides Modified by Reactive Metabolites. Chem. Res. Toxicol 2009;22:853–862. [PubMed: 19317514]
42. Lioe H, O'Hair RA. A Novel Salt Bridge Mechanism Highlights the Need for Nonmobile Proton Conditions to Promote Disulfide Bond Cleavage in Protonated Peptides Under Low-Energy Collisional Activation. J. Am. Soc. Mass Spectrom 2007;18:1109–1123. [PubMed: 17462910]
43. Bandeira N, Clauser KR, Pevzner PA. Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins. Mol. Cell Proteomics 2007;6:1123–1134. [PubMed: 17446555]

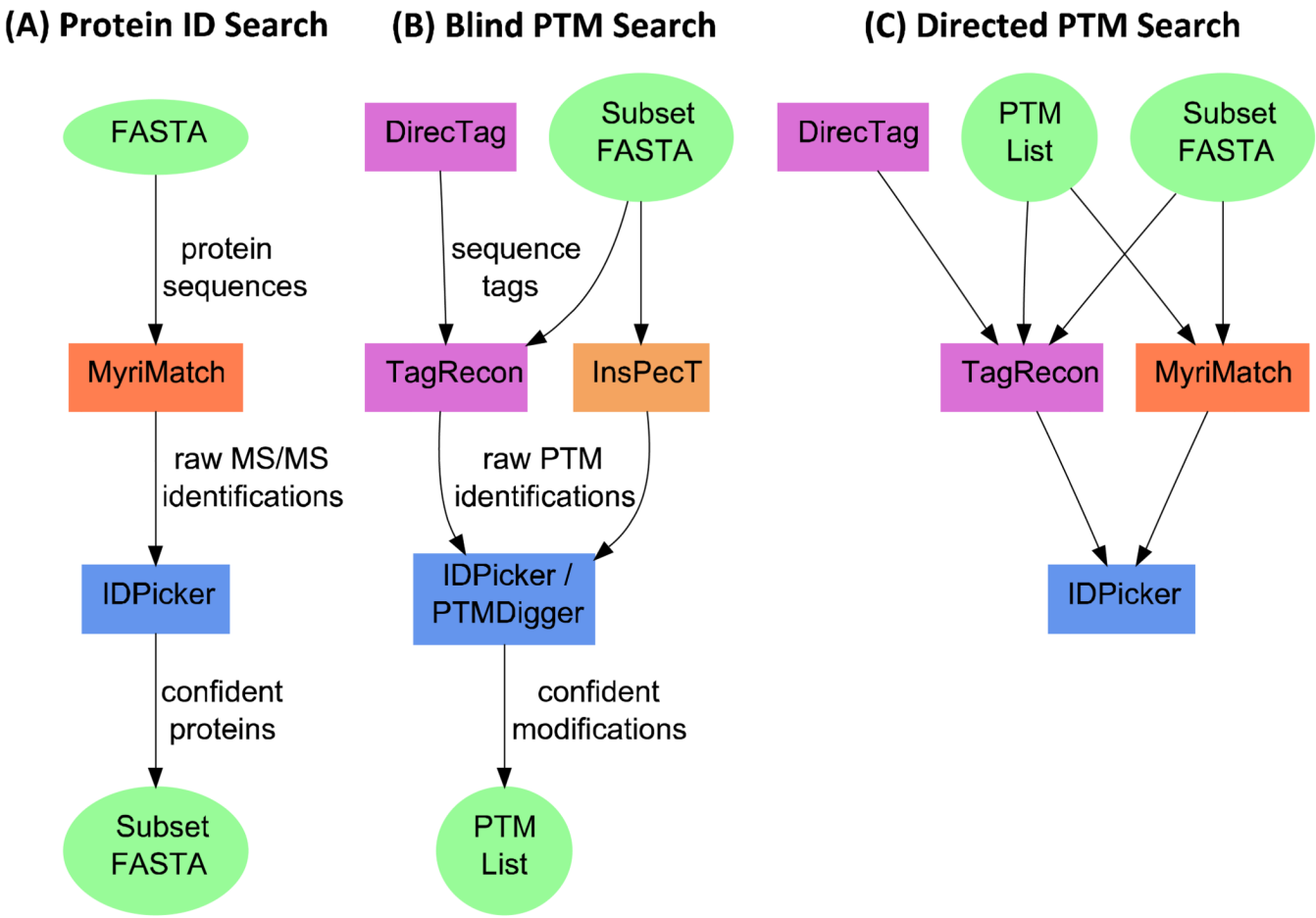


Figure 1. PTM Identification Workflow

(A) Protein database size is reduced with a standard database search (no exotic PTMs are allowed). **(B)** DirecTag software infers sequence tags from the MS/MS. TagRecon reconciles the inferred tags against the subset database while making allowances for unanticipated modifications in peptides. PTMDigger attests the resulting mass shifts and generates a list of confident modifications. **(C)** TagRecon and MyriMatch utilize the confident PTMs list for an expanded directed PTM search.

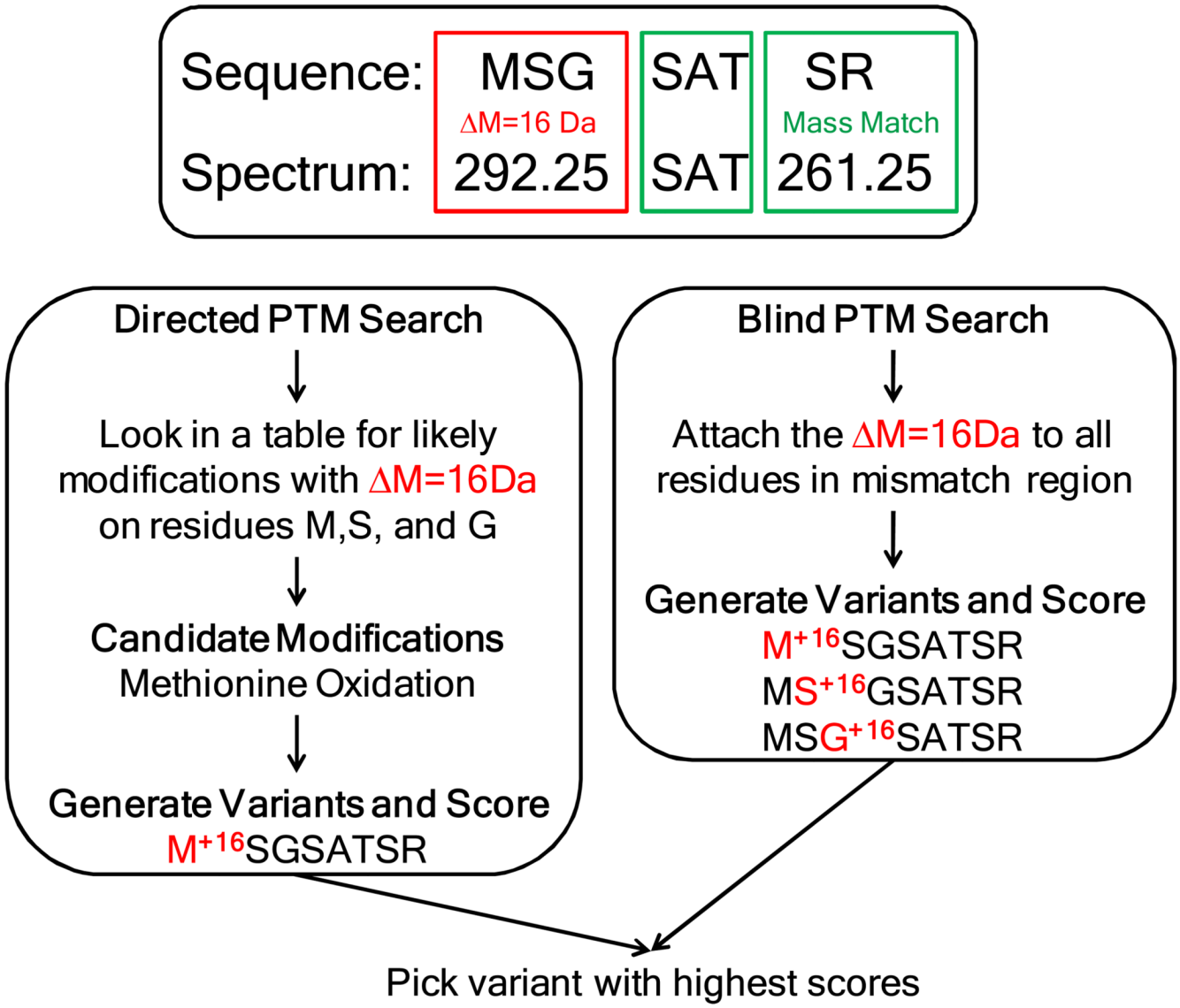
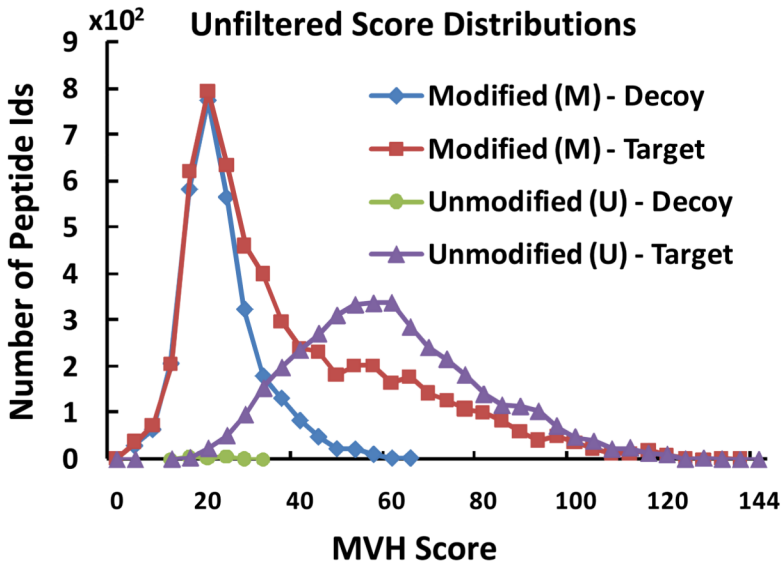


Figure 2. Modification Search Modes in TagRecon

This image illustrates two primary ways by which TagRecon interprets mass differences between peptides and spectrum sequence tags as chemical and posttranslational modifications.

3A



3B

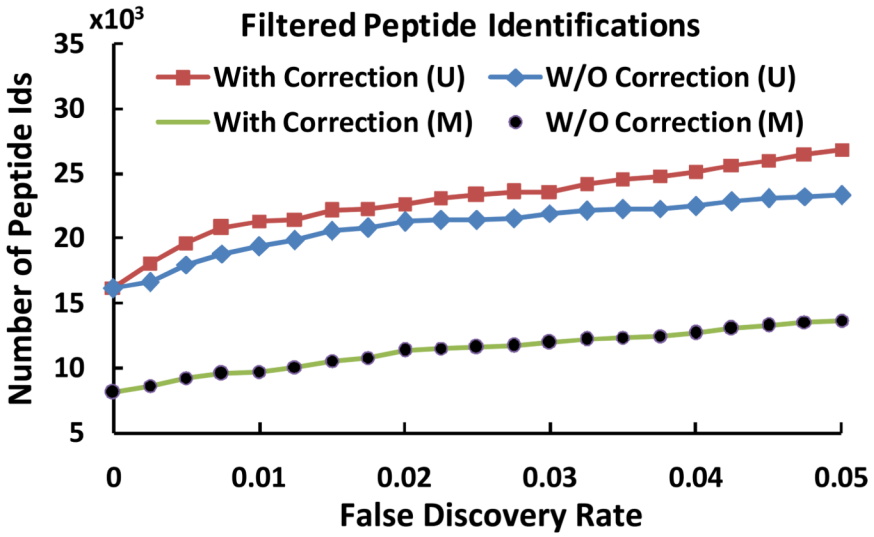


Figure 3. False Discovery Rate Correction Rescues Unmodified Peptides from "Blind PTM" Searches

(**A**) This figure illustrates the score distribution separation between modified (M) and unmodified (U) peptides. TagRecon was configured to match the MS/MS in the RAW file against a subset protein database while making allowances for unanticipated mass shifts in peptides. The scores of all top ranking peptide-spectrum matches (PSMs) were extracted and separated by peptide modification status and decoy status. Overall, modified peptides received lower search scores compared to unmodified peptides. (**B**) Correcting FDRs of the modified and unmodified peptides to offset distributional differences rescues more unmodified peptides from "blind PTM" searches.

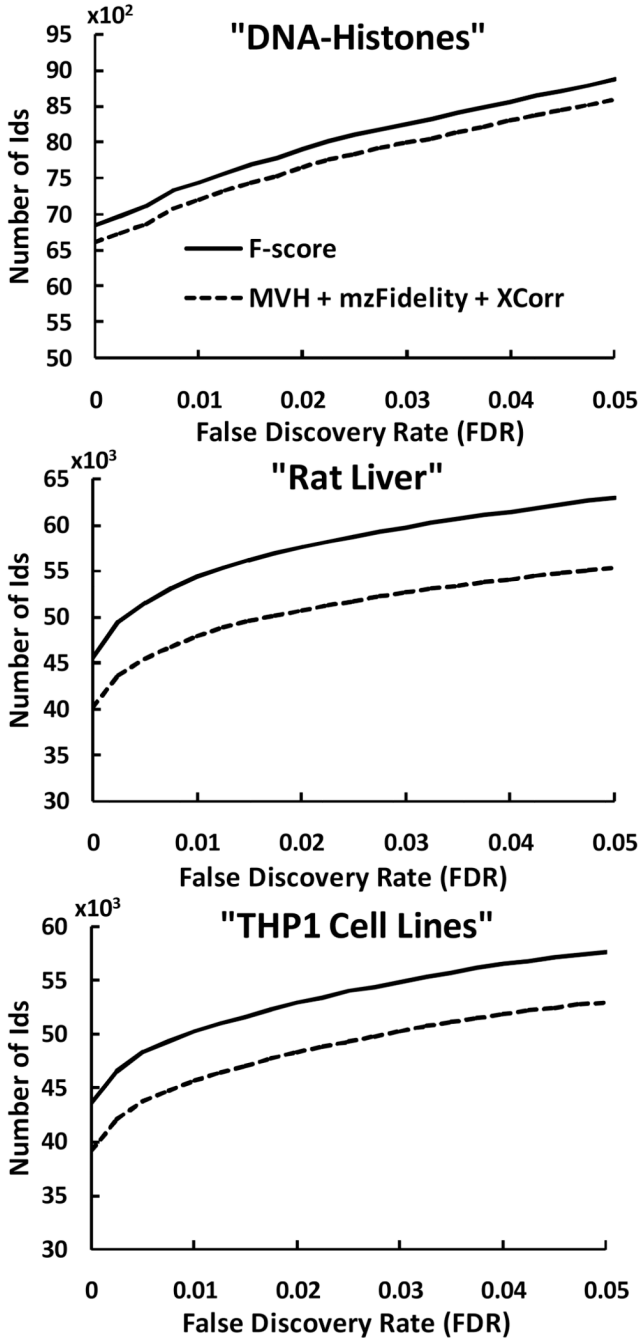


Figure 4. Combining Multiple Search Engine Metrics Improves Peptide Identification Sensitivity

TagRecon was configured to identify unexpected peptide modifications from all three toxicoproteomic samples. IDPicker employed a variety of score combinations for filtering the resulting identifications at 5% FDR. In all samples, combining orthogonal scoring metrics (MVH, mzFidelity, XCorr, number of enzymatic termini, number of missed cleavage sites, and number of sequence modifications) into a single discriminant (F-score) and re-ranking the results according to the learned score improved peptide identification rates.

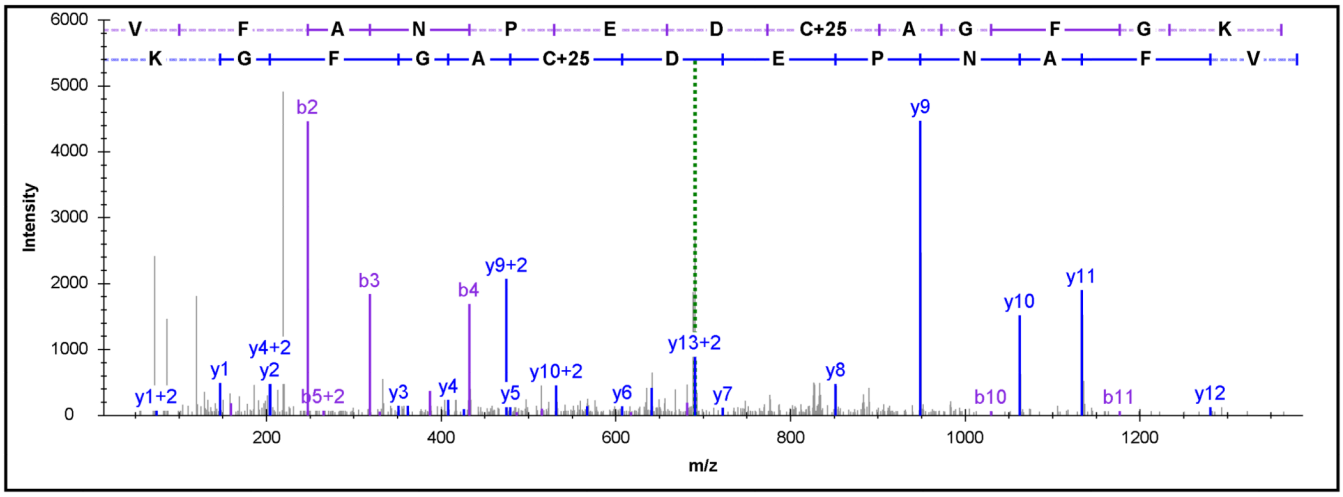


Figure 5. Unexpected Modification of Cys50 in Microsomal Glutathione S-transferase 1 (MGST1)

Posttranslational modification of this site is known to activate the enzyme. One may question the localization of the observed mass shift. We, however, note that oxidative properties of this site are previously studied. Furthermore, nine other peptides unambiguously localized similar mass shifts to cysteine residues.

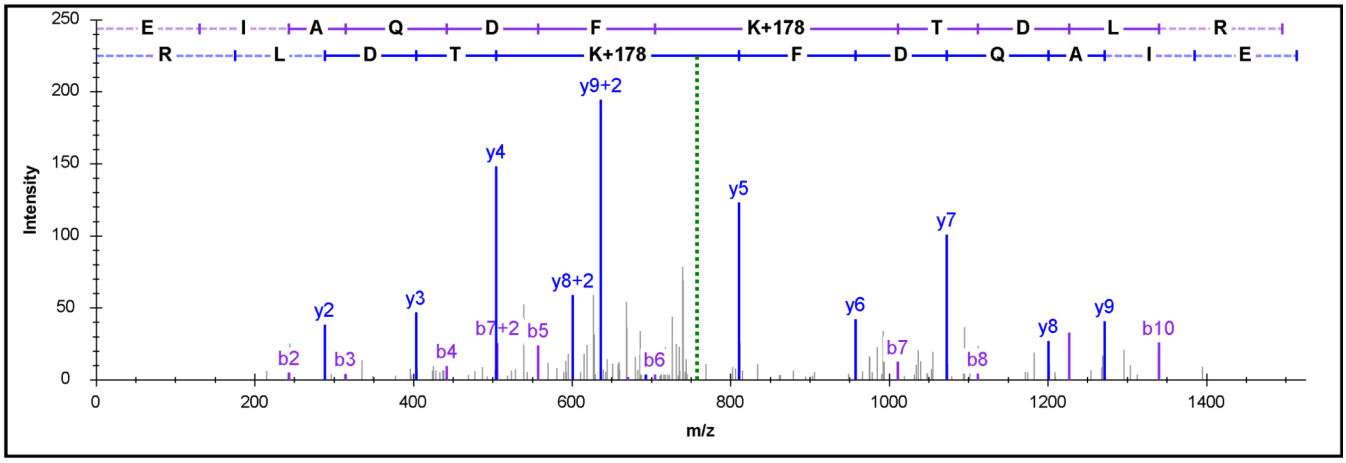


Figure 6. Modification of Histone H3

Lys80 cross-linked with DNA in the presence of 1,2-dibromoethane. Bold sections of the ion ladders indicate the fragment ion coverage for the peptide.

Table 1

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

				parent/fragment mass tolerances[c]		
data set	**instrument**	**# MS2 scans**	**sequence databases[a]**	**MyriMatch**	**TagRecon**	**InsPecT**
Human Lens	LCQ	486,120	uniprot_sprot_human_v56_5	1.25/0.5	1.25/0.5	
DNA-Histones	LTQ-Orbitrap	239,571	20100208_IPI_Human_Eco li[b]	10*/0.5	0.01/0.5	1.0/0.5
Rat Liver	Q-TOF	652,868	20100201_IPI_Rat	20*/75*	0.05/0.2	0.1/0.2
THP1 Cell Lines	LTQ	574,361	20090205_IPI_Human	1.25/0.5	1.25/0.5	2.5/0.5

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

(b) Escherichia coli proteins were included in the database to account for vector contamination of the sample.

(c) Asterisk denotes mass tolerances in parts-per-million.

Chem Res Toxicol. Author manuscript; available in PMC 2012 February 18.

Table 2

Comparison of TagRecon and InsPecT in the Context of “Blind PTM” Searching.

	no. of proteins[a]		no. of peptides[a]		no. of spectra[a]		modified peptide FDR[b]	
Sample	**TagRecon**	**InsPecT**	**TagRecon**	**InsPecT**	**TagRecon**	**InsPecT**	**TagRecon**	**InsPecT**
DNA-Histones	237	126	1,449	1,482	7,402	5,556	1.2%	1.1%
Rat Liver	1,533	1,210	9,994	7,534	50,921	32,504	3.0%	2.4%
THP1 Cell Lines	3,227	2,302	17,000	14,117	56,317	39,577	3.3%	2.9%

(a) TagRecon and InsPecT matched the MS/MS against respective subset FASTA databases. IDPicker filtered the PSMs at 2% FDR.

(b) Global false discovery rates of peptide identifications containing unanticipated modifications. Overall, TagRecon recovered more proteins, peptides, and spectra than InsPecT.

Table 3

Frequency and Abundance of Unanticipated Modifications in Samples.

sample	modified peptides	modified spectra	[a]modification frequency	[b]modification abundance
DNA-Histones	240	1,515	45.6%	30.6%
Rat Liver	1,033	4,613	23.8%	11.1%
THP1 Cell Lines	533	1,770	6.2%	3.8%

[a]Percentage of identified peptides containing an unanticipated modification.

[b]Percentage of identified spectra containing an unanticipated modification. The frequency and abundance of unexpected modifications depends on the type of the sample.

Table 4

Modifications Observed in “THP1 Cell Lines” Data set.

mass	residues	modification name[a]	peptides	spectra
26	N-terminus	*Unexpected*	469	816
209	C	Carbamidomethyl + DTT	196	308
−17	Q	Pyro-glutamic acid	85	185
42	N-terminus	Acetylation	85	98
16	F, P, W, Y	Oxidation	29	42
32	M	Double oxidation	22	31
22	D, E, T	Sodium	20	26
14	H, K, R	Methylation	11	23
54	D, E	*Unexpected*	15	21
28	N-terminus, K	Formylation or dimethylation	14	15
71	C	Acrylamide adduct	10	14
48	C	Cysteic acid	10	13
80	S	Phosphorylation	5	7
−17	N	N-succinimide	3	7
311	C	HNE + Biotinyl linker fragment	4	5
176	C	*Unexpected*	2	4

*(a)*UniMod annotation corresponding to the most frequently observed modification mass. Alternative modification names were included for ambiguous mass shifts.

Table 5

Modifications Observed in “Rat Liver” Data set.

mass	residues	modification name[a]	peptides	spectra
40	G at 2nd position	Acetone	482	1862
−17	Q	Pyro-glutamic acid	85	301
42	N-terminus	Acetylation	65	286
22	D, E, T, S	Sodium adduct	110	260
26	N-terminus, S, T	Acetaldehyde	37	252
28	D, K, S, T, N-terminus	Dimethylation or formylation	118	248
−18	D, E, S, T	Dehydration	22	214
134	C	*Unexpected*	23	117
1	N, Q	Deamidation	40	78
38	D, E	Potassium adduct	31	63
16	W	Oxidation	5	43
25	C	*Unexpected*	10	37
−2	C	Intra peptide C-C disulphide bridge	15	24
14	V	Subsititution Ile/Leu for Val	7	24
14	R, D	Methylation	3	21
60	S	*Unexpected*	2	21
40	C	Pyro-carbamidomethyl	9	20
30	R	Subsititution of Trp for Arg*	1	17
105	C	SelenoCys+Carbamidomethyl*	1	10
−17	N	N-succinimide	4	6

(a) UniMod annotation corresponding to the most frequently observed modification mass. Alternative modification names were included for ambiguous mass shifts. Asterisk denotes single peptide modifications with strong spectral evidence (Supplemental File 10).

Table 6

Modifications Observed in “DNA-Histones” Data set.

mass	residues	modification name[a]	peptides	spectra
−17	Q	Pyro-glutamic acid	6	189
22	E, D, S, T, G	Sodium adduct	47	128
12	N-terminal residues: W, H, S	N6-formyl	8	78
28	K, R, and N-terminus	Dimethylation or formylation	19	68
42	K and N-terminus	Acetylation	24	66
43	K and N-terminus	Carbamylation	7	30
48	C	Cysteic acid	9	21
86	K	Diepoxybutane adduct (Epoxide)	7	15
16	F	Oxidation	5	10
70	K	*Unexpected*	8	9
178	K	1,2-dibromoethane+DNA	2	9
76	C	*Unexpected*	4	6
14	D	Methylation	3	6
104	K	Diepoxybutane adduct (Triol)	3	5
223	C	*Unexpected**	1	5
32	W	Dioxidation	2	2

*(a)*UniMod annotation corresponding to the most frequently observed modification mass. Alternative modification names were included for ambiguous mass shifts. Asterisk denotes single peptide modifications with strong spectral evidence (see Supplemental File 10).

Table 7

Comparison of Variable Modification and Directed PTM Searches.

sample[a]	search type[b]	spectra	peptides	hh:mm:ss[c]
"Rat Liver" (QTOF)	TR-DPTM-FT	37864	9388	00:18:46
	TR-DPTM-ST	41076	10406	01:06:50
	MM-VAR-FT	29224	8745	27:21:10
"THP1 Cell Lines" (LTQ)	TR-DPTM-FT	20243	18141	00:16:51
	TR-DPTM-ST	21822	19602	00:59:39
	MM-VAR-FT	19654	17739	00:55:41

(a) Refinement searches were configured to match 148,528 “Rat Liver” spectra against 3942 proteins and 160,064 “THP1 Cell Lines” spectra against 7522 proteins, while looking for a total of seven possible modifications in each sample.

(b) Each row indicates an algorithm and configuration employed for refinement searches of the samples. TR reports the use of TagRecon and MM reports the use of MyriMatch. DPTM reports the use of directed PTM search mode and VAR reports the use of variable modification search mode. FT reports the use fully tryptic search as opposed to ST for semitryptic search.

(c) Time measurements were taken on a cluster equipped with 32 dual-core Intel 1.5GHz nodes.