Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3223−3234

pubs.acs.org/jpr

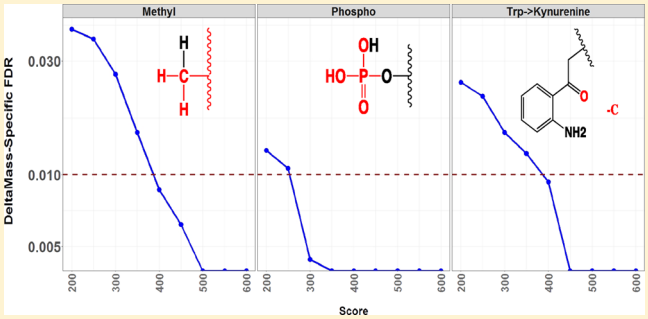
False Discovery Rate Estimation for Hybrid Mass Spectral Library Search Identifications in Bottom-up Proteomics

Meghan C. Burke,* Zheng Zhang, Yuri A. Mirokhin, Dmitrii V. Tchekovskoi, Yuxue Liang, and Stephen E. Stein

Mass Spectrometry Data Center, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899, United States

Supporting Information

ABSTRACT: We present a method for FDR estimation of mass spectral library search identifications made by a recently developed method for peptide identification, the hybrid search, based on an extension of the target-decoy approach. In addition to estimating confidence for a given identification, this allows users to compare and integrate identifications from the hybrid mass spectral library search method with other peptide identification methods, such as a sequence database-based method. In addition to a score, each hybrid score is associated with a "DeltaMass" value, which is the difference in mass of the search and library peptide, which can correspond to the mass of a modification. We explored the relation between FDR and DeltaMass using 100 concatenated random decoy libraries and discovered that a small number of DeltaMass values were especially likely to result from decoy searches. Using these values, FDR values could be adjusted for these specific values and a reliable FDR generated for any DeltaMass value. Finally, using this method, we find and examine common, reliable identifications made by the hybrid search for a range of proteomic studies.



KEYWORDS: *peptide mass spectral library, hybrid mass spectral library search, false discovery rate, target-decoy approach*

INTRODUCTION

Analysis of proteomic tandem mass spectral data with the hybrid mass spectral library search method,[1] referred to as the hybrid search, has been shown to extend the coverage of spectral libraries by identifying peptides differing by a single "inert" chemical group or modification to peptides represented in mass spectral libraries. This was accomplished by matching product ions in the query spectrum either directly to its corresponding product ion in the library spectrum or to an ion shifted by the difference in precursor mass, DeltaMass, for the two spectra. In this paper, we seek to develop a method to estimate false discovery rates (FDRs), or the percentage of incorrect assignments included among the accepted assignments,[2,3] for this search method. This must account for the fact that a given query spectrum may match multiple library sequences equally well since different library sequences may match a given sequence through different modifications. Moreover, we explore the possibility that certain specific DeltaMass values may be more subject to false identifications than others. We develop and apply an FDR estimation method using the target decoy method[4] based on reverse and random decoy mass spectral libraries.

Evidence has shown that the advantages of mass spectral library searching, which include a reduced search space and similarity-based scoring against real spectra, make it a valuable complement to sequence database searching.[5] However, the spectral match score, based on a modified cosine similarity, is not directly comparable to scores obtained from alternative search methods. Therefore, a method to estimate FDR for hybrid search identifications would make the confidence of a hybrid search result comparable to alternative search methods, such as sequence database searching.

Previous reports have established that the target-decoy approach can be used to estimate FDR for mass spectral library searching of low-resolution tandem mass spectra.[6−8] Decoy mass spectral library generation methods include the precursor swap[6] and shuffle and reposition[7,8] methods. The precursor swap method has been found to result in bias toward target spectral library identifications for high-resolution tandem mass spectra.[9] The method developed here is similar to the shuffle and reposition method in that the decoy spectrum is generated from reversing or randomizing of the target library peptide sequence and shifting of m/z values accordingly;[9] however, it expands on the previously established methods in the following ways.

First, as described in an earlier report, the decoy mass spectral libraries used are constructed from high mass accuracy target mass spectral libraries.[9] Second, the hybrid search produces a DeltaMass value for each peptide spectral match,

Received: December 6, 2018
Published: July 31, 2019

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

Table 1. Raw Tandem Mass Spectra Data Sources Used for Spectral Library Searching

test name	data source	matrix	organism	label	fractions	target spectral library	1% FDR score threshold
Tissue-iTRAQ	PRIDE: PXD002774[18]	tissue	human	iTRAQ 4-plex	18	Orbitrap-HCD iTRAQ-4 (2 parts)	385
Tissue-TMT	PRIDE: PXD004683[19]	tissue	human	TMT 6-plex	24	Orbitrap-HCD TMT (2 parts)	380
Cells-Unlabeled	PRIDE: PXD001468[20]	cells	human		24	Orbitrap-HCD (2 parts)	450
Plasma-Unlabeled	NIST SRM1950	plasma	human		24	Orbitrap-HCD (2 parts)	450

which can be used to further analyze the confidence of a spectral match in a modification-specific manner. It is important to note that the hybrid search does not restrict candidate spectra to be searched against a query spectrum based on precursor m/z, as done in a direct (nonhybrid) MS/MS search where candidates are selected based on the user-provided precursor m/z tolerance. Moreover, unlike other "blind" or "open" modification search methods,[10−15] all product ions, including those containing the modification, are employed in locating matching high-resolution tandem mass spectra. In addition to direct MS/MS spectral matches, the hybrid search also considers spectra for which increased spectral similarity is obtained after shifting product ions in the library spectrum by the observed DeltaMass normalized for charge. If we suppose that a true DeltaMass value must be restricted to a chemical difference between the query peptide and a peptide in the mass spectral library, rather than a random value or a value that is not chemically relevant, then we might expect that the DeltaMass value itself can also be used as a criterion to distinguish correct and incorrect hybrid search identifications.[16]

The hybrid mass spectral library search of raw data against both target and decoy mass spectral libraries will be used to determine if (1) both reverse and random decoy mass spectral libraries can be used to estimate FDR for hybrid search identifications, (2) whether the spectral match score threshold corresponding to 1% FDR is similar across multiple varieties of analysis, and (3) if observed DeltaMass values can be used to add an additional level of confidence to hybrid search identifications. Finally, with this ability to measure the confidence of identifications, we present a broad analysis of modifications found in a range of analyses.

EXPERIMENTAL SECTION

Hybrid Mass Spectral Library Search

The hybrid mass spectral library search was performed using NIST MSPepSearch[17] with a product ion tolerance of 40 ppm. Also, peaks within 2.5 m/z of the precursor m/z were ignored. Centroided tandem mass spectral data were searched against publicly available (peptide.nist.gov) target mass spectral libraries (Table 1) and the reverse or random decoy mass spectral libraries constructed from the target mass spectral library, as previously described.[9] Briefly, the reverse decoy library was constructed from peptides in the target mass spectral library, where each target peptide sequence was reversed while keeping the C-terminus fixed. The random decoy mass spectral library was also constructed from the target spectral library with the C-terminal residue fixed; however, the amino acid sequence was randomly assigned with the constraint that the overall amino acid distribution in the target mass spectral library was maintained. At this time, users may request the algorithm to generate reverse or randomized decoy libraries.

DeltaMass values in the hybrid search output were centroided with a precursor mass tolerance of 5 ppm, unless stated otherwise. Briefly, centroided DeltaMass values were computed by determining the DeltaMass value for which the greatest number of elements is within 5 ppm precursor mass tolerance. The median DeltaMass value for the bin is computed, and the calculation is repeated for the remaining DeltaMass values with the constraint that the DeltaMass value for a single hybrid identification is used only once. The centroiding methods used in this paper simply assumed a constant ppm error in the precursor mass. An optimal method would involve a more accurate model and optimization. Since this would not materially affect the ideas developed in the paper, it was not done here.

Three data sets[18−20] used to test the target decoy-based FDR estimation method were obtained from PRIDE,[21] for which the accessions, the target mass spectral libraries used, and the name used to refer to the test are shown in Table 1. In addition, raw data obtained from the 2D LC−MS/MS analysis of NIST SRM1950 (see Supplemental Methods), a pooled human plasma sample, was also used for testing.

FDR Estimation

Hits for which the original library peptide was semi-tryptic and/or the peptide length was <9 residues (as well as peptide length of <10 residues where DeltaMass is <−200 Da, which is approximately equivalent to a peptide length of <9 residues) were discarded. Additionally, only the highest scoring distinct peptides were kept, which was determined by the peptide sequence and DeltaMass value (rounded to the second decimal place). The FDR at a given score threshold was calculated as $(2 \times ND)/(ND + NT)$,[4,22] where ND and NT are the total number of decoy and target mass spectral library identifications above a given threshold, respectively. Briefly, the score that is computed for a given spectral match by MSPepSearch is based on a modified cosine similarity with a maximum score of 999; therefore, spectral matches with higher scores have a greater mass spectral similarity.

Peptide Sequence Homology

Peptide sequence homology was calculated as the number of identical amino acids in a pairwise alignment divided by the length of the known peptide sequence, where the number of identical amino acids was calculated using the BioPython Pairwise 2 module[23] with no match parameters and no gap penalties.

Assigning Proposed Chemical Formulas to DeltaMass Values

The NIST MS Interpreter[24] Chemical Formula Calculator was used to assign plausible chemical formulas to DeltaMass values that do not correspond to simple modifications (see Supplemental Methods). For this analysis, the following chemical composition was allowed for each centroided DeltaMass value (semicolon denotes from-to range): C,

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

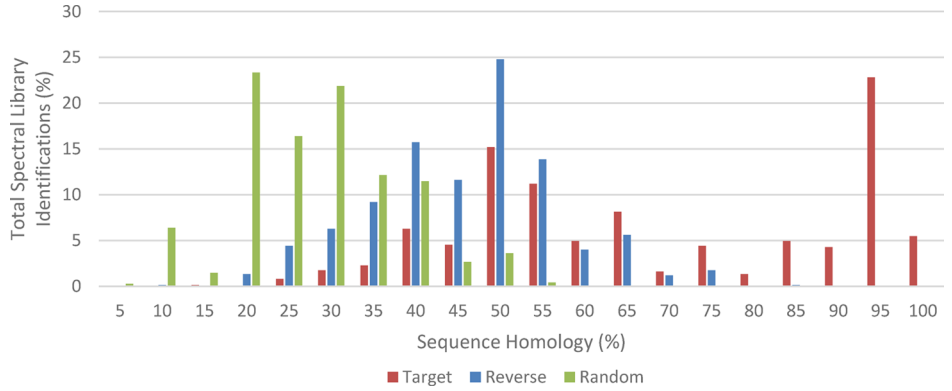
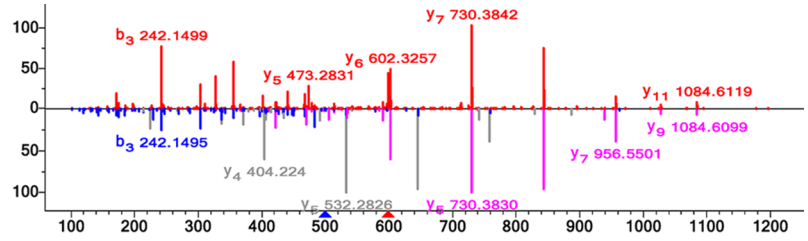


Figure 1. Comparison of sequence homology between synthetic peptide sequences that are not in the target spectral library and the highest scoring hybrid search identifications obtained from the target (red), reverse decoy (blue), and random decoy (green) spectral libraries for the top 5% of peptide spectra searched. Sequence homology was calculated as the number of matching amino acids, as determined by the BioPython Pairwise 2 module,[23] divided by the length of the known peptide sequence.



	Peptide	DeltaMass	Modification
Synthetic Peptide Sequence	LGALLQEAVGAR		
Reverse Decoy Spectral Match	LGALLQTGAR	198.100	T → EAV

Figure 2. Spectral match to the reverse decoy peptide sequence LGALLQTGAR (charge = 2; mods = 0; spectral match score = 578). The head-to-tail plot shows peaks present in the query spectrum (top) that either directly match the library spectrum (bottom, blue) or match after shifting by the DeltaMass of 198.100 Da (pink), which represents the difference in the peptide sequence between the decoy peptide sequence and the true identity of the synthetic peptide (highlighted in red). The original m/z values for the library spectrum peaks that match after shifting are shown as "ghost" peaks in gray.

−3:15; H, −4:25; O, −2:10; N, −2:10; S, 0:4; Na, 0:2; K, 0:1; I, 0:1.

RESULTS AND DISCUSSION

Part I: Evaluation of Decoy Mass Spectral Libraries

The first objective was to compare high scoring decoy spectral matches for different varieties of decoy searches. Publicly available spectra of synthetic peptides[25] not present in the target library served as search spectra, allowing the target library to also serve as a decoy library. Selecting peptide lengths from 11 to 25 residues generated 14,997 peptide mass spectra. The best spectrum for each peptide was selected as the one with the highest MSGF+ score (method described in ref 9.). Three decoy libraries were used; the first two were reverse and random[9] libraries derived from a target library (Orbitrap-HCD), and the third was the target library.

High scoring decoy spectral matches, which are expected to be representative of incorrect identifications, were selected as the 750 highest scoring peptide assignments (top 5% of peptide spectra searched) for each decoy spectral library. The score thresholds for each subset of spectral matches were 578, 496, and 483 for target, reverse, and random decoy mass spectral libraries, respectively. The relative score thresholds are consistent with those observed for the highest scoring (750 total peptide spectra) spectral matches obtained from a direct (nonhybrid) MS/MS search of the same synthetic peptide MS/MS spectra (228, 208, and 195 for target, reverse, and random, respectively).

The higher score thresholds for hybrid search identifications relative to a direct MS/MS search is partly a consequence of an increased chance of peak matching since a library peak can match a search spectrum peak directly or after mass shifting, as well as the expanded search space for the hybrid search. Score thresholds for other data sets show similar trends. For example, the score thresholds at 1% FDR for Cells-Unlabeled data are 475 and 440 for hybrid search with reverse and random decoy libraries (Supplemental Figure S1), respectively, and 360 for direct MS/MS search. Another measure of the expanded search space for hybrid searches is the difference in numbers of identifications for the hybrid and direct search at a given score level. Using data from the Plasma-Unlabeled data set, above a score of 400 (equivalent to 2.0% FDR for the hybrid search), we find that six times more decoy identifications were obtained from the hybrid search (data not shown).

Although it is possible that the higher score thresholds observed for the reverse decoy spectral library, relative to the random decoy, may be due to palindromic peptide sequences, these sequences only comprise 65 of the 390,009 distinct peptide sequences in the Orbitrap-HCD iTRAQ 4-plex

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

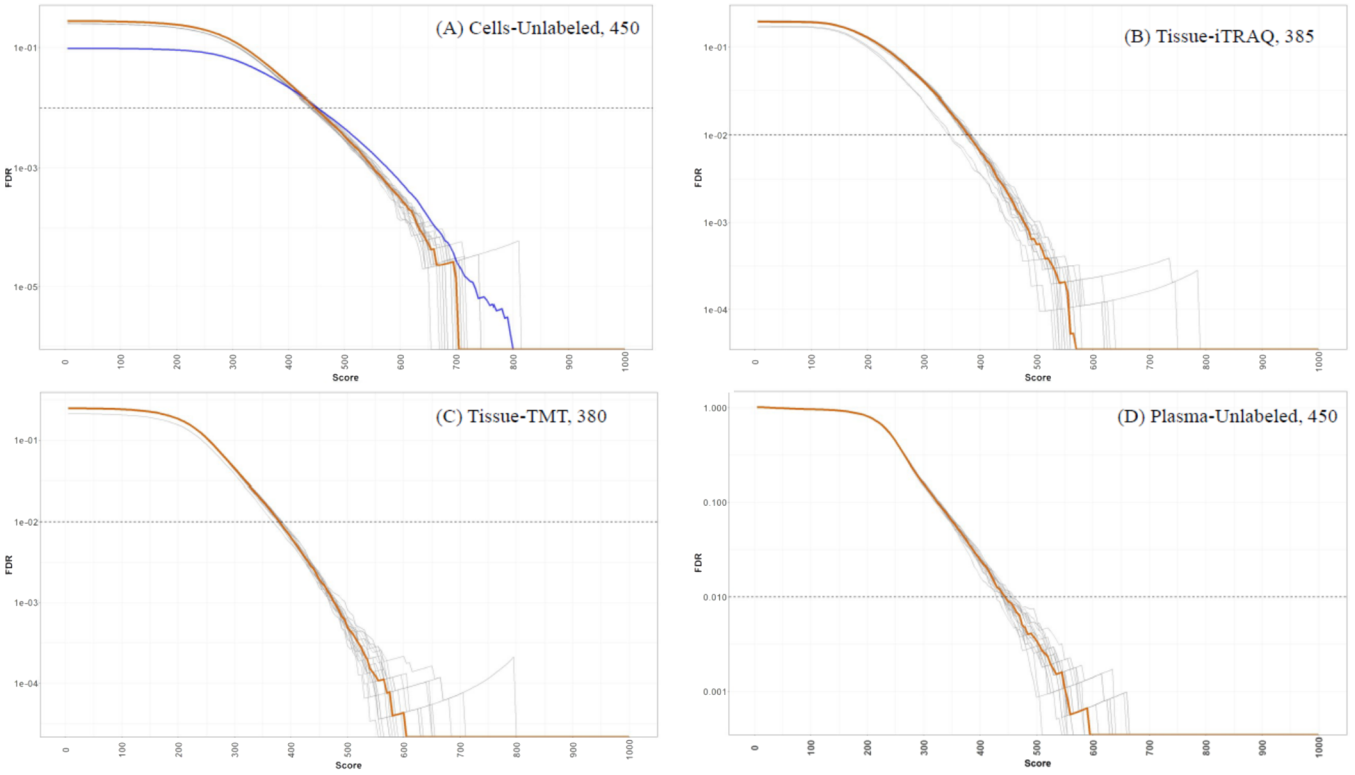


Figure 3. Comparison of observed false discovery rates, computed as the median of 20 separate, concatenated target-randomized decoy spectral library searches (shown in red) between (A) Cells-Unlabeled, (B) Tissue- iTRAQ, (C) Tissue-TMT, and (D) Plasma-Unlabeled. The curve shown in blue in (A) was taken as the highest scoring spectral assignment, followed by normalization for the increased size of the decoy spectral library, which was 20 times larger than the target spectral library. The target spectral library and corresponding random decoy spectral library used for each data source are listed in Table 1. Score thresholds corresponding to 1% FDR are labeled and shown as a dashed line.

library,[9] for example, and therefore are not expected to account for the difference in score thresholds. Next, to further explore the origin of the differences in threshold scores, sequences of high scoring decoy mass spectral library identifications were compared to those of the matching synthetic peptides. Because the synthetic peptides were derived from known sequences in the human proteome, some degree of homology with other human-derived peptides present is expected. This has been reported in a previous analysis of a cross-species mass spectral library search.[7] In the present study, for high scoring decoy matches, we use as a measure of sequence homology the number of matching amino acids following pairwise alignment.[23,26]

Using this measure of homology, Figure 1 shows the average sequence homology for the top 750 scoring spectral matches from target, reverse, and random mass spectral libraries to be 66.5, 43.9, and 25.9% homology, respectively. The large percentage (38%) of target mass spectral library identifications that share at least 80% sequence homology with the synthetic peptide sequence is not unexpected because, although the synthetic peptide sequences are absent from the target library, they are able match that of the reference library peptide through multiple modifications. Manual inspection of reverse decoy mass spectral library identifications with high sequence homology (>60%) revealed that such decoy library identifications may correspond to the correct peptide sequence. One such identification from the reverse decoy mass spectral library is shown in Figure 2 (LGALLQTGAR, DeltaMass = 198.100 Da, score = 578), where the calculated sequence homology with the known peptide sequence is 75% and the DeltaMass localized to Thr does, in fact, correspond to the true identity of the synthetic peptide (LGALLQEAVGAR).

Collectively, results show that both the reverse and random decoy spectral libraries generate similar score thresholds based on an average difference in identifications of 10.4%; however, the difference in observed sequence homology suggests that high scoring reverse decoy mass spectral library identifications may correspond to the correct sequence. We propose that hybrid search identifications from the decoy mass spectral library whose sequence and DeltaMass value correspond to the correct peptide sequence should be considered false negative identifications. Given that the random decoy spectral library resulted in fewer high scoring false negative identifications, the random decoy mass spectral library is better suited to estimate the error for the hybrid search. We further analyze whether identifications from the random decoy mass spectral library are representative of true false identifications from the target mass spectral library in the next section.

Part II: Does Sample Type or Sample Preparation Affect the 1% FDR Score Threshold?

Clearly, the number of peptide identifications and modifications found in any LC−MS/MS experiment will depend on the sample type and sample preparation methods. To examine the influence of these factors on score thresholds, we determined spectral match score thresholds at 1% FDR across four experiments (Table 1 and Figure 3), where the score threshold reported is taken as the median across 20 separate, concatenated searches of the target spectral library and a randomized decoy spectral library. Another possible method would be to select the highest scoring mass spectral assignment

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

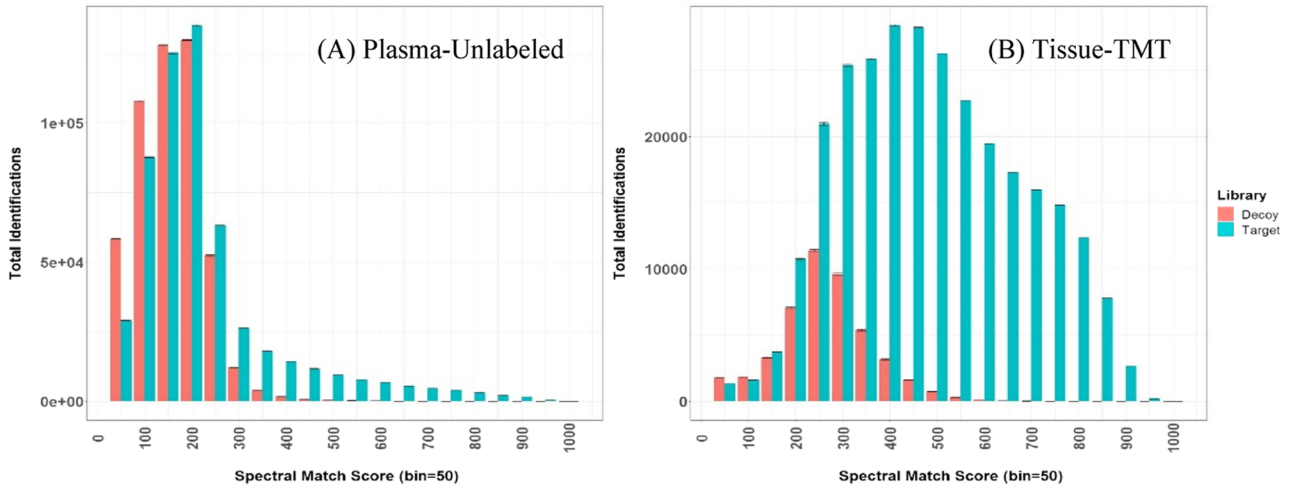


Figure 4. Comparison of the mean hybrid search spectral match score distribution for (A) Plasma-Unlabeled and (B) Tissue-iTRAQ identifications from the randomized decoy spectral library (red) and Orbitrap-HCD target spectral library (blue) computed for all 20 concatenated target-decoy searches where the error bars are the standard errors of the mean. The similar distribution at low score values indicates that decoy identifications are representative of false identifications. The difference in high scoring target library identifications between (A) and (B) reflects the difference in total distinct peptide sequences identified in Plasma-Unlabeled and Tissue-iTRAQ.

from the target library and a decoy spectral library that is 20 times larger, as has been suggested in the literature,[6] followed by normalization for the increased size of the decoy spectral library before computing the FDR; however, we find that this method underestimates the FDR at low spectral match scores and overestimates the FDR at high spectral match scores (Figure 3A) for the spectral library search results shown. The median spectral match score threshold avoids this bias while reducing the variability in computed FDR at low FDR values (Figure 3). Based on a comparison of median deviation across an increasing number of concatenated target-randomized decoy spectral library searches, the deviation in the computed spectral match score corresponding to a 1% FDR threshold does not significantly change after a total of five spectral library searches (Supplemental Figure S2).

The accuracy of the estimated error rate computed using the median across 20 concatenated target-randomized decoy spectral library searches was analyzed using the 14,997 tandem mass spectra obtained from synthetic peptides[25] for which the known peptide sequence was absent from the target spectral library, thereby evaluating the hybrid search only. Here, the resulting target library identifications below an estimated 1% FDR were inspected to determine if the difference between the target library peptide sequence, for identifications with a localized DeltaMass, was due to a discrete difference in the peptide sequence that could be explained by the DeltaMass value (true positive) or due to randomly overlapping m/z values (false positive). The resulting actual error rate corresponding to an estimated error rate of 1% (spectral match score > 670) is 3.57%. To determine if the total true and false positive identifications based on the estimated FDR of 1% and the actual FDR were significantly different, the Fisher's exact p-value was calculated using a 2 × 2 matrix composed of the total true and false positive identifications based on the actual and estimated FDR values.[27,28] The resulting p-value was 0.2137. Because the p-value is not less than 0.05, the difference between the estimated and actual FDR values is not statistically significant.

Score thresholds at 1% FDR were found to be highly similar for samples prepared in a similar manner. For example, both Cells-Unlabeled and Plasma-Unlabeled have spectral match score thresholds of 450 corresponding to 1% FDR. In addition, Tissue-iTRAQ and Tissue-TMT have spectral match score thresholds of 385 and 380, respectively. One possible cause for the observed difference in score thresholds is the increased stabilization of b-type ions for peptides bearing an N-terminal iTRAQ or TMT tag.

Next, we sought to determine if random decoy mass spectral library identifications are representative of the false positive identifications from the target spectral library. The target and random decoy library hybrid search identifications shown in Figure 4 show a similar score distribution at low spectral match score values, where true negative identifications are expected to occur, for both Plasma-Unlabeled and Tissue-TMT. However, a tail is observed at high spectral match scores for target library identifications in both Plasma-Unlabeled and Tissue-TMT, where the difference in target library score distributions between Figure 4A and Figure 4B reflects the difference in samples used in each analysis and is independent of the presence of reporter ions as they are also present in the reference library spectra in the Orbitrap-HCD iTRAQ 4-plex and TMT mass spectral libraries. Based on the observed similarity at low spectral match values, the random decoy hybrid search identifications appear to be representative of false positive identifications.[29] In addition, the average decoy fraction, taken across the top 10 ranking identifications for all search spectra, was 46.3% for Plasma-Unlabeled (Supplemental Figure S3). The slight bias toward the target mass spectral library is consistent with previously published reports using the target decoy approach to estimate FDR for mass spectral library searching.[7,9]

Part III: Are some DeltaMass Values more Likely to Occur than Others for High Scoring Decoy Library Spectrum Matches?

Each hybrid score is associated with a DeltaMass value, which can represent the difference in masses due to a difference in chemical formulas between the search and decoy library spectra. In this section, we examine the distribution of these values for decoy matches to find whether any DeltaMass values are more likely to be found than others. Since only high

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

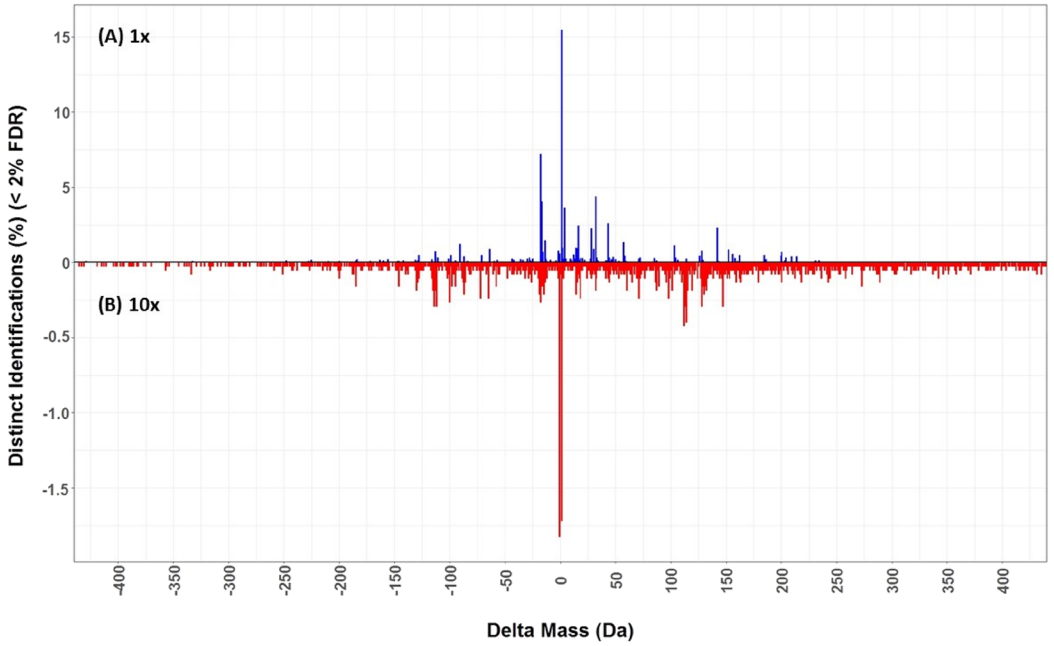


Figure 5. Head-to-tail comparison of the DeltaMass value distribution for Plasma-Unlabeled hybrid search identifications obtained from (A) the target spectral library shown in blue and (B) 100 randomized decoy spectral libraries, where the percent of distinct identifications has been normalized per spectral library searched (*y*-axis expanded 10×), shown in red at a score threshold of 400 (2% FDR). DeltaMass values were centroided with a precursor mass tolerance of 5 ppm.

scoring decoy spectra are relevant for FDR calculations, and there are relatively few at the most widely used 1% level, to derive statistically meaningful results, we created and searched 100 different random decoy libraries (112,797,000 total decoy spectra).

The centroided DeltaMass values obtained from the best scoring spectral match assignments for Plasma-Unlabeled raw data searched against the target and 100 randomized decoy mass spectral libraries, above a spectral match score of 400 (2% FDR), are shown as a head-to-tail plot in Figure 5 (see also Supplemental Methods). A FDR threshold of 2%, rather than 1%, was chosen to provide better statistics (1% FDR results are shown in Supplemental Figure S4). The target library DeltaMass distribution shown in Figure 5A contains maxima at values for many common modifications. Results from the random decoy libraries (Figure 5B with *y*-axis expanded 10× relative to that of Figure 5A) are far more uniformly distributed, although there are a relatively small number of preferred values (Table 2 shows the most common of these).

The distribution of DeltaMass identifications from the decoy libraries (Figure 5B) can be viewed as being composed of two classes of values. One comprises seemingly random DeltaMass values, and the other comprises a relatively small number of commonly occurring, preferred nonrandom values. To distinguish these two classes, we applied a χ^2 goodness-of-fit test[30] for all DeltaMass values within ±500 Da. This computed abundance threshold below which the percent of decoy identifications per DeltaMass is uniformly distributed is 0.053% (p = 1.0) of all decoy identifications (equivalent to two decoy identifications per DeltaMass) at a score threshold of 400, which represents 88% of all DeltaMass values for Plasma-Unlabeled at 2% FDR (Table 3). The fraction of uniformly distributed DeltaMass values decreases to 77% of values for Plasma-Unlabeled at 1% FDR (Table 3). The dependence on score of these two classes of DeltaMass values is shown in Supplemental Figure S5, which shows that the fraction of DeltaMass values classified as "nonrandom" increases significantly with increasing score, reaching 100% of false identifications above a score of 500 (Supplemental Table S1).

Table 2. Top 10 Most Frequently Observed Decoy DeltaMass Values and Corresponding Possible Modification(S) Obtained from the Hybrid Search of Plasma-Unlabeled against 100 Random Decoy Spectral Libraries (FDR < 2%; Score > 400)

rank	DeltaMass	% of decoy IDs	possible modification(s)
1	−1.014	1.82	isotope error (−1)
2	1.000	1.72	isotope error (+1)
3	−0.978	1.08	amidation (HNO-1)
4	−1.046	0.61	N-1H-3O-1S
5	112.073	0.42	add-Xle, isotope error (−1)
6	114.055	0.40	C3H6N4O
7	−114.089	0.29	loss-Xle, isotope error (−1)
8	−112.075	0.29	loss-Xle, isotope error (+1)
9	113.085	0.29	add-Xle
10	128.047	0.29	add-Glu, isotope error (−1)

Of the most commonly occurring decoy DeltaMass values shown in Table 2, most have readily interpretable assignments. For example, the two most frequently identified values correspond to the mass difference between ^{12}C and ^{13}C isotopes (−1.003355 Da), referred to here as isotope error (−1) and isotope error (+1) followed by amidation (−0.984016 Da) (Table 2 and Figure 6), all of which may match an isotopic peak present in the query spectrum given the allowed product ion tolerances at high m/z values. The decoy DeltaMass value of 113.085 Da (rank 9) is explained by the monoisotopic mass of the collectively most frequently occurring amino acid pair leucine and isoleucine (abbreviated Xle). Decoy DeltaMass values of −1.046 and 114.055 Da do

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

Table 3. Numbers of Total and Nonrandom Decoy DeltaMass Values between −500 Da and + 500 Da[a]

score	total nonzero decoy DeltaMass values	total nonzero decoy DeltaMass values (DeltaMass between −500 and +500 Da)	nonrandom decoy DeltaMass values
200	47,180	15,052	2746
250	26,907	9809	1418
300	10,339	5879	420
350	4869	3654	117
400	2384	2072	245
450	1086	1012	233
500	457	436	436
550	174	167	167
600	55	51	51

[a]Nonrandom values based on the χ^2 goodness-of-fit test at multiple score thresholds obtained from the hybrid search of Plasma-Unlabeled against 100 random decoy spectral libraries.

not correspond to a simple modification and therefore likely reflect multiple modifications or isotopic misidentifications. Combinations of modifications that may correspond to these DeltaMass values include loss of ammonia and thiocarboxy, as well as loss of Val and CAM, and addition of Arg for −1.046 and 114.055 Da, respectively.

Interestingly, isotope error (−1) and amidation are the most frequently identified decoy DeltaMass values for Plasma-Unlabeled and Cells-Unlabeled, respectively (Supplemental Tables S2 and S3). Excluding just the most frequent false identifications could be used to improve the confidence in hybrid search identifications. For example, simply excluding amidation from Cells-Unlabeled would decrease the number of false identifications and, therefore, the FDR by 25% at score 600; however, the total number of target library identifications would only be reduced by 1%.

For each centroided DeltaMass value, we may go one step further and examine the application of the hybrid search to compute DeltaMass-specific FDR values or the false discovery rate for a given DeltaMass value. The use of modification-specific FDR thresholds in proteomic studies has been previously discussed and implemented.[16,12] As an example, in the recent publication describing MSFragger,[12] the authors modeled the distribution of correct and incorrect identifications using the score obtained from a database search and mass shift values using a bin size of 1 Da. In this work, the DeltaMass values have been centroided using a precursor mass tolerance in ppm (as described in the Experimental Section) and a DeltaMass-specific FDR value calculated as $(2 \times ND_i)/(ND_i + NT_i)$, where ND_i and NT_i correspond to the total decoy and target identifications, respectively, for a given DeltaMass value ("i").

The DeltaMass-specific FDR distribution for 10 DeltaMass values (shown in Table 4) from the Cells-Unlabeled data set using the median values obtained from 20 concatenated searches of the target and a randomized decoy spectral library is shown in Figure 7. The resulting distributions for the selected DeltaMass values illustrate that most of the major false DeltaMass values correspond to rarely reported true modifications, such as isotope error (−1) and amidation. Furthermore, the difference in total decoy identifications per modification demonstrates that the prior probability of identifying a decoy spectrum with a given DeltaMass depends on the specific modification.

For targeted studies focused on selected modifications, using a suitably large number of randomized decoy libraries in concatenated target-decoy library searches, DeltaMass-specific FDRs may be derived, possibly leading to a substantial gain in identifications, especially at lower FDR levels. For example, using the actual number of decoy matches for FDR estimation for a DeltaMass value of 656.257 Da, which corresponds to the glycan composition Hex1HexNAc1NeuAc1, leads to five identifications (score threshold, 250; Supplemental Table S2) at 1% FDR for Plasma-Unlabeled, which is four more than the number obtained using the global count of decoy matches (score threshold, 450; Table 1). Moreover, all five peptides have been previously reported to contain a glycosylation site (Supplemental Table S4). Score threshold refinement based on the DeltaMass-specific FDR for the 10 DeltaMass values shown in Figure 7 results in a gain in identifications for four modifications (Cys(CAM) → cysteic acid, loss-Xle, methylation, and Trp → hydroxykynurenine) and no change in identifications for the DeltaMass values corresponding to add-Asp, hexose, loss-Ala, and loss-CAM/Gly (Table 4).

The actual rate per spectrum of false identifications is expected to depend on the nature and quality of the search spectra. For example, the number of target and decoy matches per library spectrum searched, above a score of 350 for the Cells-Unlabeled spectra, was, on average, 25- and 38-fold higher, respectively, than for Plasma-Unlabeled data (see Supplemental Table S5). This is not unexpected as the samples

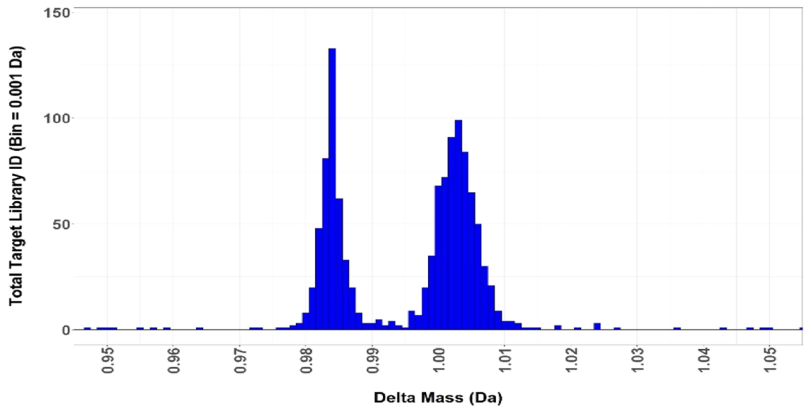


Figure 6. Total target library identifications per DeltaMass for Plasma-Unlabeled with a bin of 0.001 Da demonstrating sufficient resolution to distinguish deamidation (0.984016 Da) and isotope error (1.003355 Da).

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

Table 4. List of DeltaMass Values and the Corresponding Modifications Selected for DeltaMass-Specific FDR Analysis[a]

DeltaMass (Da)	modification	probable origin	score threshold (1% DeltaMass-specific FDR)	ID below 1% FDR
115.03	C4H5NO3	add-Asp	450	50 (+0)
−0.987	HNO(−1)	amidation	550	702 (−630)
−9.034	C-2H-3 N-1O + 2	Cys(CAM) → cysteic acid	400	123 (+17)
162.053	C6H10O5	hexose	450	4 (+0)
−1.015	isotope error (−1)	$^{12}C - {}^{13}C$	500	643 (−270)
−71.035	C-3H-5 N-1O-1	loss-Ala	450	182 (+0)
−57.018	C-2H-2 N-1O-1	loss-CAM/Gly	450	39 (+0)
−113.082	C-6H-11 N-1O-1	loss-Xle	400	347 (+66)
14.018	CH2	methyl	400	356 (+34)
19.99	C-1O2	Trp → hydroxykynurenine	350	18 (+6)

[a]The spectral match score threshold corresponding to a 1% DeltaMass-specific FDR and the corresponding number of identifications is provided with the change relative to the global 1% FDR threshold shown in parentheses.

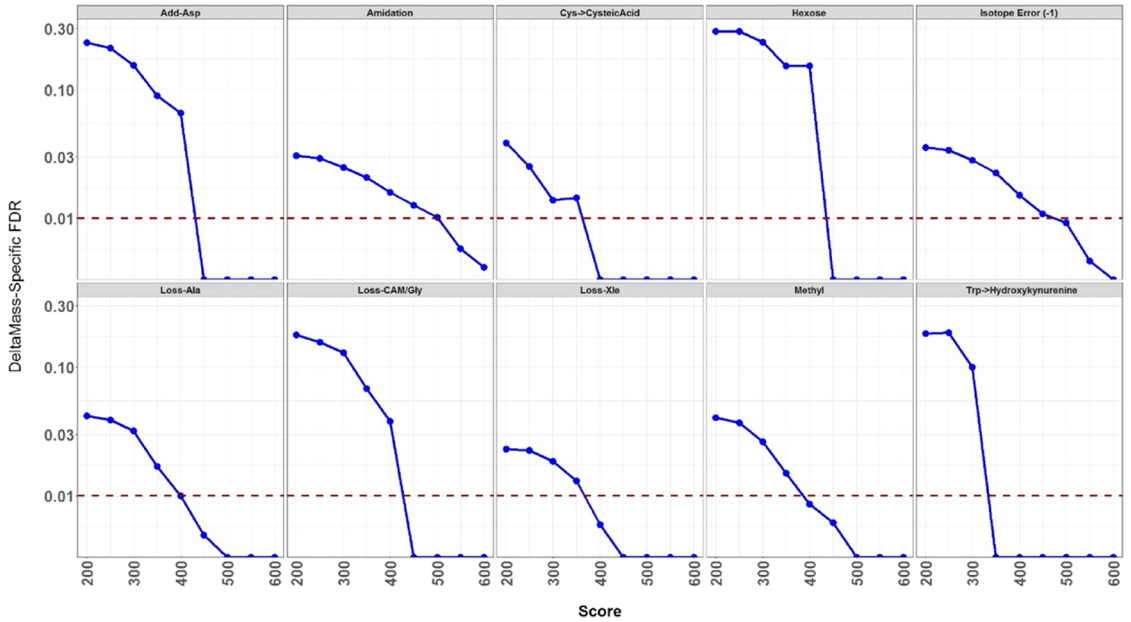


Figure 7. DeltaMass-specific FDR values, calculated as $(2 \times ND_i)/(ND_i + NT_i)$ versus score for DeltaMass values corresponding to add-Asp, amidation, cysteine → cysteic acid, hexose, isotope error (−1), loss-Ala, loss-CAM or loss-Gly, loss-Xle, methyl, and Trp → hydroxykynurenine from Cells-Unlabeled hybrid search identifications (see Table 4 for descriptions of these modifications).

that generated the mass spectra are themselves different. However, a comparison of false identifications per library spectrum searched, normalized for total search spectra, above a score of 400 for two different Tissue-iTRAQ_ data sources were, on average, within 1.95-fold of each other. This indicates that a similar rate per spectrum of false identifications for search spectra was obtained from similar sources and with similar sample preparation protocols.

In summary, using the decoy spectral libraries described here, a few and relatively rare DeltaMass values are associated with a large fraction of higher scoring decoy identifications (Table 3). The distribution of decoy DeltaMass values may be employed to further refine the FDR in two ways. The first method is a broad approach that excludes the most frequently identified decoy DeltaMass value. Applying this method to Cells-Unlabeled hybrid identifications at a score threshold of 500 would result in an FDR of 0.8 and 0.6% for the top excluded value and all remaining values, respectively. Because the contribution of preferred or nonrandom DeltaMass values increases with increasing score, the greatest effect of excluding the most preferred value occurs at high spectral match scores and, therefore, low FDR values (Supplemental Figure S5). For example, at score 600, the FDR for the most preferred value is 0.83%, which is more than 30-fold greater than that of the remaining values (0.03%). Given that the most common decoy DeltaMass value may be experiment-specific, we recommend that the most common decoy DeltaMass value be computed for each experiment. The second method, a targeted approach, is the DeltaMass-specific FDR, which has been shown to decrease the 1% FDR score threshold for select DeltaMass values that correspond to commonly occurring simple modifications. Although the targeted FDR method resulted in significantly more (4-fold increase) identifications for the DeltaMass value corresponding to Hex1HexNAc1NeuAc1 (Supplemental Table S2), the disadvantage of this approach is the requirement for the use of multiple decoy libraries to derive statistically significant DeltaMass-specific decoy score distributions (Figure 8). A third possible method to further refine FDR includes assigning "classes" of DeltaMass values and subsequently computing FDR for each class; however, this would require the analysis of a larger cohort of mass spectral data to develop rules for assigning classes.

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

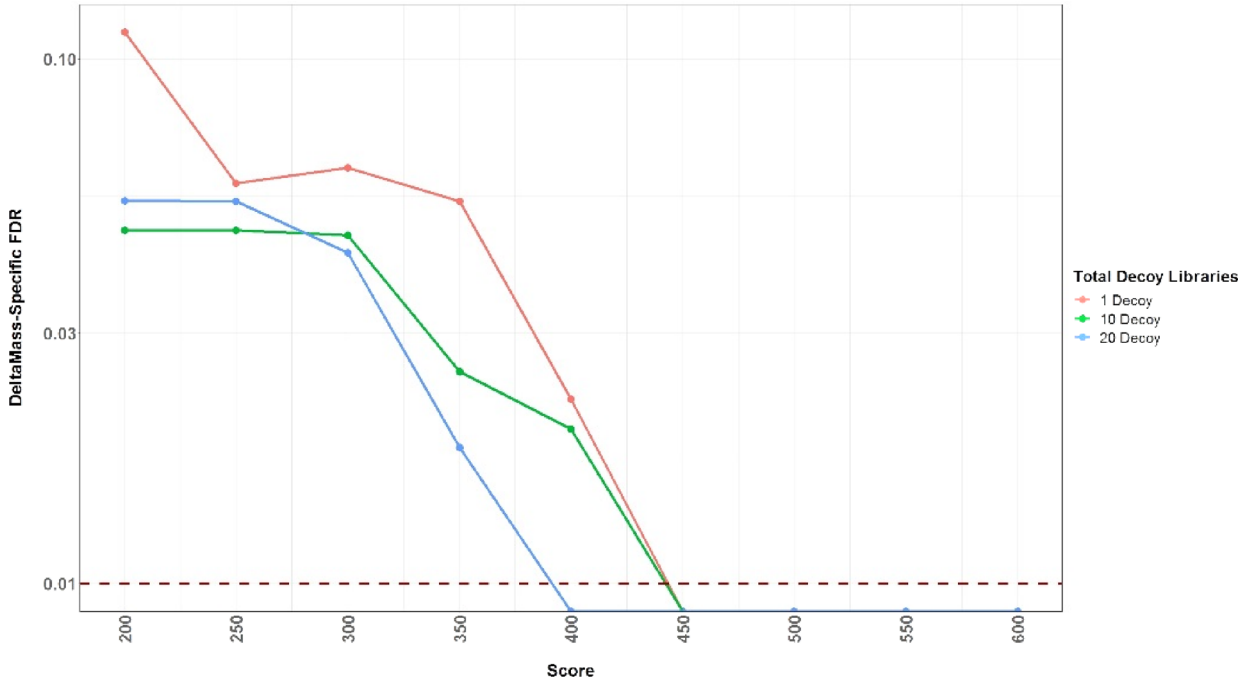


Figure 8. Comparison of the observed DeltaMass-specific FDR values from Cells-Unlabeled for the DeltaMass corresponding to acetylation obtained from a total of 1 (red), 10 (green), and 20 (blue) separate, concatenated target-randomized decoy library searches.

Part IV: Examination of Highly Confident DeltaMass Values

Using the methods described above, we examined modifications commonly found in the experiments given in Table 1 (Supplemental Tables S2, S3, S6, and S7). While most have been reported in UniMod[31] (see Supplemental Methods), there are 59 distinct DeltaMass values across the top 100 centroided DeltaMass values in Plasma-Unlabeled, Cells-Unlabeled, Tissue-iTRAQ, and Tissue-TMT that do not correspond to a combination of up to two common modifications, as defined in Supplemental Table S8, present in UniMod[29] (see Supplemental Tables S2, S3, S6, and S7). One possible source of these DeltaMass values is chemically feasible modifications (or a combination of modifications) not yet in the database. Another is an erroneous value possibly arising from incorrect precursor charge or precursor mass measurement. Below, we describe methods for assessing both cases.

In the first case, we employed an updated version of the NIST MS Interpreter[24] Chemical Formula Calculator, which can compute formulas that have positive and negative stoichiometric coefficients to find chemical formulas for unannotated DeltaMass values (see Supplemental Methods). This led to the assignment of 29 proposed chemical formulas for DeltaMass values not in Unimod. As an example, several high scoring hybrid search identifications for a DeltaMass of −116.061 Da (rank 28) were made in the Tissue-iTRAQ analysis for which the NIST MS Interpreter proposed a chemical formula of C-4H-8 N-2O-2 (Supplemental Table S6). Further inspection of the spectral matches suggests that it likely originates from an intrapeptide disulfide bond (mass error of 2 mDa), where the DeltaMass corresponds to the loss of (2) carbamidomethylations of cysteine and (2) hydrogens (Figure 9) relative to the library entry. Another example is −11.035 Da (rank 1), for which the NIST MS Interpreter proposed a chemical formula of C-1H-1 N-1O-1S (Supplemental Table S6). Here, the DeltaMass reflects a difference in reagents used for cysteine alkylation. The cysteine residues in the query spectra were alkylated with methyl methanthiosulfonate, a reversible cysteine alkylation reagent,[32] rather than iodoacetamide, which was used to produce the reference library entries. Additional possible elemental compositions for 29 of the 59 unannotated DeltaMass values obtained from MS Interpreter are shown in Supplemental Tables S2, S3, S6, and S7.

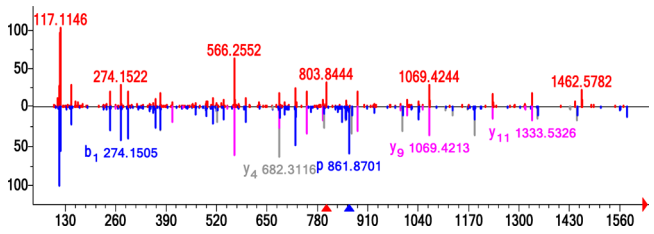


Figure 9. Spectral match of a query spectrum (top, red) to the peptide sequence ETYGEMADCCAK (charge = 2; mods = iTRAQ(N-term), iTRAQ(Lys), Cys(CAM), Cys(CAM); spectral match score = 699; bottom, blue) for which the DeltaMass of −116.057 Da does not correspond to a single known modification. Further inspection suggests the DeltaMass may correspond to an intrapeptide disulfide bond where the DeltaMass corresponds to the loss of (2) carbamidomethylation of cysteine and (2) hydrogens. Product ions that contain the modification are shown in pink with the original m/z values for the library spectrum peaks shown as ghost peaks in gray.

To estimate the level of unidentified DeltaMass values generated by faulty data, we examined values having implausible mass defects for peptides. We then selected a range of DeltaMass values having mass defects between 0.3 and 0.9, covering 60% of the possible range. On average, this comprised 13.25% of values with identifications below 1% FDR, and no identified modification mass defect values fell in this range. Inspection of high scoring hybrid identifications

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

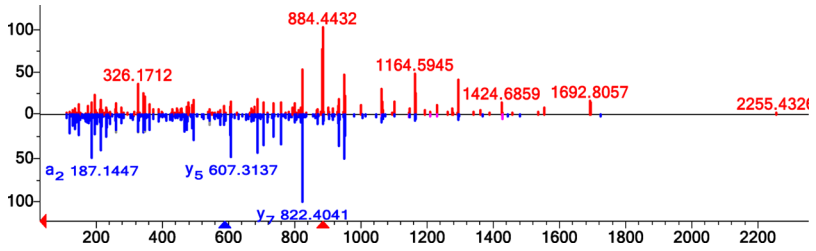


Figure 10. Spectral match of a query spectrum (top, red) to the peptide sequence LTEMETLQSQLMAEK (precursor m/z = 884.4324; charge = 3; mods = oxidation (Met, position 12); spectral match score = 595; bottom, blue) for which the DeltaMass value of 883.432 Da is the result of incorrect charge state assignment.

indicated that they were primarily generated from peptides with incorrect charge state assignments. An example is presented in Figure 10, which illustrates a high scoring hybrid identification (score = 595) with a DeltaMass of 883.432 Da and precursor m/z of 884.4324 m/z (z = 3). Together, these results demonstrate that mass defect assessment and MS Interpreter are useful methods to determine whether a given DeltaMass value is expected to correspond to a plausible chemical formula and, further, may be used to reject DeltaMass values that are likely due to incorrect precursor charge or precursor mass measurement. A third possible source of DeltaMass values that do not correspond to known modifications are those that arise from chimeric, or impure, tandem mass spectra, which have not been analyzed in this work.

In a recent publication, Kong et al.[12] described MSFragger, a new utility for finding blind modifications. They reported 408 mass shifts (FDR < 1%) between the query precursor and sequence database precursor for the Cells-Unlabeled data cohort.[12] We have subjected the same dataset to analysis by the hybrid search and found that 355 (87%) of these mass shift values were also identified within 9 mDa (FDR < 1%; centroided at 2 ppm). In addition, the top 10 distinct mass shift[12] or DeltaMass values are shown in Supplemental Table S9, of which only 15% of the distinct mass shift (MSFragger) and DeltaMass values (hybrid search) can be assigned to a single modification present in UniMod.[31]

An advantage of the hybrid search is its direct use of ions containing the modification in locating candidate matching spectra. For example, the hybrid search identified over 3-fold more peptides containing the modification add-Asp (50 peptide identifications, +115.0269 Da) than by MSFragger.[12] This primarily involves addition to the C-terminus, thereby shifting all y-ions (Supplemental Figure S6). Overall, numbers of modifications reported by the two methods are similar. For example, the total peptide identifications corresponding to the five most frequently observed modifications, isotope error (+1), carbamylation, (2) isotope error (+1), ammonia loss, and the unidentified DeltaMass of 301.987 Da, are within 9.5% of each other for both methods. However, different methods were used for estimating FDR, and more examination is needed to interpret these differences.

CONCLUSIONS

Decoy libraries can be effectively employed to estimate FDR for hybrid search identifications. Use of multiple concatenated target-randomized decoy searches to compute the median FDR enabled the more accurate determination of this threshold. The 1% FDR score threshold fell between 380 and 450 for a range of proteomic studies, roughly 23% higher on average than for direct library matching studies.[9] These higher values reflect the larger search space for the hybrid search (each search spectrum peak has two chances to match a library spectrum peak). The range of threshold scores for samples prepared in a similar manner is rather small (difference in score of 5 for Tissue-iTRAQ and Tissue-TMT) and would typically result in a difference in numbers of identifications of only 1.45%. Unexpectedly, a preference for a limited number of DeltaMass values was found, which was only apparent after the use of multiple random decoy libraries. This preference may be utilized to further refine the FDR in one of two ways. The first method, in which the top decoy DeltaMass value is excluded, results in a modest decrease in total false identifications. The second targeted method is DeltaMass-specific FDR, which has resulted in a significant decrease in score corresponding to 1% FDR for select DeltaMass values corresponding to simple, commonly occurring modifications. However, the increase in identifications for select DeltaMass values does come with a trade-off of additional search time as additional decoy mass spectral libraries may be required. Finally, an examination of the origin of many DeltaMass values was facilitated using a recently enhanced, freely available version of the NIST MS Interpreter[24] program that allows formulas to have negative as well as positive elemental compositions.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00863.

Figure S1, comparison of observed false discovery rates obtained from the hybrid search of Cells-Unlabeled using reverse and randomized decoy mass spectral libraries; Figure S2, median deviation in spectral match score thresholds at 1% estimated FDR across an increasing number of concatenated target-decoy searches; Figure S3, fraction of incorrect identifications, or decoy fraction, for the top 10 ranking hybrid identifications for all search spectra from Plasma-Unlabeled; Figure S4, head-to-tail comparison of the DeltaMass distribution for Plasma-Unlabeled decoy hybrid search identifications per library at 1% and 2% FDR thresholds; Figure S5, fraction of nonrandom and random decoy DeltaMass values as a function of score; Figure S6, distribution of DeltaMass localization relative to the C-terminus for hybrid search identifications with a DeltaMass corresponding to add-Asp (PDF)

Table S1, Plasma-Unlabeled decoy DeltaMass values obtained from 100 random decoy spectral libraries clas-

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

sified as non-random; Table S2, Plasma-Unlabeled centroided DeltaMass values with the total number of identifications from target or decoy spectral libraries; Table S3, Cells-Unlabeled centroided DeltaMass values with the total number of identifications from target or decoy spectral libraries; Table S4, Plasma-Unlabeled hybrid identifications with a DeltaMass corresponding to Hex1HexNAc1NeuAc1 (FDR < 1%); Table S5, comparison of the false identifications per spectrum for Plasma-Unlabeled and Cells-Unlabeled; Table S6, Tissue-iTRAQ centroided DeltaMass values with the total number of identifications from target or decoy spectral libraries; Table S7, Tissue-TMT centroided DeltaMass values with the total number of identifications from target or decoy spectral libraries; Table S8, monoisotopic masses and corresponding modifications allowed in annotation of DeltaMass values; Table S9, top 10 distinct DeltaMass values and mass shift values for Cells-Unlabeled (XLXS)

AUTHOR INFORMATION

Corresponding Author

*E-mail: meghan.burke@nist.gov. Tel.: +1 301 975 5631.

ORCID

Meghan C. Burke: 0000-0001-7231-0655

Zheng Zhang: 0000-0001-7058-4011

Funding

We acknowledge support from the NIH/NCI CPTAC program through an Interagency Agreement (ACO15005) with NIST.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Certain commercial equipment, instruments, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

ABBREVIATIONS

FDR, false discovery rate; hybrid search, hybrid mass spectral library search method; DeltaMass, mass difference (Da) between a query spectrum precursor and library spectrum precursor; 2D LC−MS/MS, two-dimensional liquid chromatography−tandem mass spectrometry;; iTRAQ, isobaric tags for relative or absolute quantitation; TMT, tandem mass tag; CAM, carbamidomethylation; N_T, number of target spectral library identifications; N_D, number of decoy spectral library identifications; Xle, leucine or isoleucine

REFERENCES

(1) Burke, M. C.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Markey, S. P.; Heidbrink Thompson, J.; Larkin, C.; Stein, S. E. The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. *J. Proteome Res.* **2017**, *16*, 1924−1935.

(2) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc., Ser. B (Methodol.)* **1995**, *57*, 289−300.

(3) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 9440−9445.

(4) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4*, 207−214.

(5) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **2011**, *11*, 1075−1085.

(6) Cheng, C.-Y.; Tsai, C.-F.; Chen, Y.-J.; Sung, T.-Y.; Hsu, W.-L. Spectrum-based Method to Generate Good Decoy Libraries for Spectral Library Searching in Peptide Identifications. *J. Proteome Res.* **2013**, *12*, 2305−2310.

(7) Lam, H.; Deutsch, E. W.; Aebersold, R. Artificial Decoy Spectral Libraries for False Discovery Rate Estimation in Spectral Library Searching in Proteomics. *J. Proteome Res.* **2010**, *9*, 605−610.

(8) Ahrné, E.; Ohta, Y.; Nikitin, F.; Scherl, A.; Lisacek, F.; Müller, M. An improved method for the construction of decoy peptide MS/MS spectra suitable for the accurate estimation of false discovery rates. *Proteomics* **2011**, *11*, 4085−4095.

(9) Zhang, Z.; Burke, M.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Markey, S. P.; Yu, W.; Chaerkady, R.; Hess, S.; Stein, S. E. Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches. *J. Proteome Res.* **2018**, *17*, 846−857.

(10) Ahrné, E.; Nikitin, F.; Lisacek, F.; Müller, M. QuickMod: A Tool for Open Modification Spectrum Library Searches. *J. Proteome Res.* **2011**, *10*, 2913−2921.

(11) David, M.; Fertin, G.; Rogniaux, H.; Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J. Proteome Res.* **2017**, *16*, 3030−3038.

(12) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14*, 513−520.

(13) Ma, C. W. M.; Lam, H. Hunting for Unexpected Post-Translational Modifications by Spectral Library Searching with Tier-Wise Scoring. *J. Proteome Res.* **2014**, *13*, 2262−2271.

(14) Ye, D.; Fu, Y.; Sun, R.-X.; Wang, H.-P.; Yuan, Z.-F.; Chi, H.; He, S.-M. Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010**, *26*, i399−i406.

(15) Wilhelm, T.; Jones, A. M. E. Identification of Related Peptides through the Analysis of Fragment Ion Mass Shifts. *J. Proteome Res.* **2014**, *13*, 4002−4011.

(16) Fu, Y.; Qian, X. Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry. *Mol. Cell. Proteomics* **2014**, 1359.

(17) *The NIST MSPepSearch Mass Spectral Library Search Program* (Version 0.95 Build May 17, 2016). http://chemdata.nist.gov/dokuwiki/doku.php?id=peptidew:mspepsearch.

(18) Danda, R.; Ganapathy, K.; Sathe, G.; Madugundu, A. K.; Ramachandran, S.; Krishnan, U. M.; Khetan, V.; Rishi, P.; Keshava Prasad, T. S.; Pandey, A.; Krishnakumar, S.; Gowda, H.; Elchuri, S. V. Proteomic profiling of retinoblastoma by high resolution mass spectrometry. *Clin. Proteomics* **2016**, *13*, 29.

(19) Stewart, P. A.; Fang, B.; Slebos, R. J. C.; Zhang, G.; Borne, A. L.; Fellows, K.; Teer, J. K.; Chen, Y. A.; Welsh, E.; Eschrich, S. A.; Haura, E. B.; Koomen, J. M. Relative protein quantification and accessible biology in lung tumor proteomes from four LC-MS/MS discovery platforms. *Proteomics* **2017**, 1600300.

(20) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33*, 743−749.

(21) Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; O'Kelly, G.; Schoenegger, A.; Ovelleiro, D.; Pérez-Riverol, Y.; Reisinger, F.; Ríos,

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234

D.; Wang, R.; Hermjakob, H. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, *41*, D1063−D1069.
(22) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 47−50.
(23) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(24) *Mass Spectrum Interpreter* (Version BETA 3.1c build 05/31/2017). https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:interpreter
(25) Zolg, D. P.; Wilhelm, M.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Delanghe, B.; Bailey, D. J.; Gessulat, S.; Ehrlich, H.-C.; Weininger, M.; Yu, P.; Schlegl, J.; Kramer, K.; Schmidt, T.; Kusebauch, U.; Deutsch, E. W.; Aebersold, R.; Moritz, R. L.; Wenschuh, H.; Moehring, T.; Aiche, S.; Huhmer, A.; Reimer, U.; Kuster, B. Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **2017**, *14*, 259−262.
(26) Poole, C. B.; Gu, W.; Kumar, S.; Jin, J.; Davis, P. J.; Bauche, D.; McReynolds, L. A. Diversity and Expression of MicroRNAs in the Filarial Parasite, Brugia malayi. *PLoS One* **2014**, *9*, No. e96498.
(27) Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification. *BMC Bioinf.* **2012**, *13*, S2.
(28) Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **1922**, *85*, 87−94.
(29) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* **2008**, *7*, 29−34.
(30) Snedecor, G.; Cochran, W. *Stadistical methods*; 8th ed. Iowa University Press: Ames, 1989.
(31) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4*, 1534−1536.
(32) Jaffrey, S.; Snyder, S. The biotin switch method for the detection of S-nitrosylated proteins. *Science's STKE: Signal Transduction Knowl. Environ.* **2001**, *2001*, pl1.

DOI: 10.1021/acs.jproteome.8b00863
J. Proteome Res. 2019, 18, 3223−3234