Published in final edited form as:

J Proteome Res. 2020 April 03; 19(4): 1863–1872. doi:10.1021/acs.jproteome.9b00766.

Accurate Identification of Deamidation and Citrullination from Global Shotgun Proteomics Data Using a Dual-Search Delta Score Strategy

Xi Wang,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States; Geomicrobiology Laboratory, State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Beijing 100083, China

Adam C. Swensen,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Tong Zhang,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Paul D. Piehowski,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Matthew J. Gaffrey,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Matthew E. Monroe,

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Ying Zhu,

Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Hailiang Dong,

Corresponding Author: Wei-Jun Qian – Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States; weijun.qian@pnl.gov.

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00766>

The authors declare no competing financial interest.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00766>.

Figure S1–S7: Same-sequence spectra with different positive or negative scores; Figure S8: A different-sequence spectrum with positive Delta Score; Figure S9: Selected spectra from a previous publication illustrating the utility of Delta Score; Figure S10: Relationship between Delta Score derived from MSGF+ and other searching algorithms (PDF)

Table S1: Final list of identified unique deamidated/citrullinated peptides; Table S2–S7: Modified peptides or PSMs identified from same-seq/diff-seq spectra; Table S8: Modified PSMs identified from a previously published citrullination data set (XLS)

Geomicrobiology Laboratory, State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Beijing 100083, China

Wei-Jun Qian

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States

Abstract

Proteins with deamidated/citrullinated amino acids play critical roles in the pathogenesis of many human diseases; however, identifying these modifications in complex biological samples has been an ongoing challenge. Herein we present a method to accurately identify these modifications from shotgun proteomics data generated by a deep proteome profiling study of human pancreatic islets obtained by laser capture microdissection. All MS/MS spectra were searched twice using MSGF+ database matching, with and without a dynamic +0.9840 Da mass shift modification on amino acids asparagine, glutamine, and arginine (NQR). Consequently, each spectrum generates two peptide-to-spectrum matches (PSMs) with MSGF+ scores, which were used for the Delta Score calculation. It was observed that all PSMs with positive Delta Score values were clustered with mass errors around 0 ppm, while PSMs with negative Delta Score values were distributed nearly equally within the defined mass error range (20 ppm) for database searching. To estimate false discovery rate (FDR) of modified peptides, a “target-mock” strategy was applied in which data sets were searched against a concatenated database containing “real-modified” (+0.9840 Da) and “mock-modified” (+1.0227 Da) peptide masses. The FDR was controlled to ~2% using a Delta Score filter value greater than zero. Manual inspection of spectra showed that PSMs with positive Delta Score values contained deamidated/citrullinated fragments in their MS/MS spectra. Many citrullinated sites identified in this study were biochemically confirmed as autoimmunogenic epitopes of autoimmune diseases in literature. The results demonstrated that in situ deamidated/citrullinated peptides can be accurately identified from shotgun tissue proteomics data using this dual-search Delta Score strategy. Raw MS data is available at ProteomeXchange (PXD010150).

Graphical Abstract

Keywords

deamidation; citrullination; Delta Score; dual search; shotgun proteomics

INTRODUCTION

The ability to accurately and confidently identify deamidated or citrullinated proteins from global shotgun proteomics data has been greatly desired in biomedical research due to the critical roles these modifications play in human diseases. Deamidation is the process by which an amide functional group is replaced by a hydroxyl functional group resulting in a mass increase of 0.9840 Da. These conversions can occur nonenzymatically on asparagine (N) and glutamine (Q) residues, changing them to aspartate (D) and glutamate (E), respectively.^{1,2} However, glutamine (Q) deamidation can also be mediated by tissue transglutaminase.³ These conversions dramatically influence protein conformation, degradation, and aggregation^{1,4–6} and may associate with immune response,³ cell aging,^{7,8} and neurodegenerative disorders.⁹ In a process similar to deamidation, citrullination is an enzymatic process by which an arginine (R) residue is transformed into citrulline by replacing the guanidine group into a ureido group through protein arginine deiminases (PAD),¹⁰ which also results in a mass increase of 0.9840 Da. Abnormal citrullination may generate new autoimmunogenic epitopes contributing to autoimmune diseases such as type 1 diabetes^{11,12} and rheumatoid arthritis.^{13,14}

Due to the nature of the small mass shifts of these modifications (0.9840 Da), it has been challenging to reliably identify deamidation and citrullination from mass spectrometry (MS)-based global shotgun proteomics proteome profiling data of complex samples. High-resolution MS (e.g., Orbitrap-based) has been used to identify deamidated/citrullinated peptides in global proteomics.^{15–18} However, the false discovery rate (FDR) tends to be high. Manual inspection of spectra was often necessary to confirm modified peptides in previous studies.¹⁸

The main challenge to overcome in order to correctly identify a deamidated/citrullinated peptide is differentiating peptides that have the incorporation of the naturally occurring isotopically heavy ¹³C, resulting in only a 0.0193 Da heavier shift in mass, from its ¹²C-deamidated/citrullinated counterpart. Thus, incorrect MS/MS monoisotopic ion selection during data acquisition will often produce a false positive identification of such modifications.^{15,16,19} Popular database searching tools such as SEQUEST,²⁰ Mascot,²¹ and Andromeda²² (used by MaxQuant) match an experimental spectrum with a theoretical spectrum within a given mass tolerance (typically 5–10 ppm for high resolution mass spectrum). Under this scenario, once a ¹³C-unmodified peptide ion is selected as the monoisotopic ion, the unmodified peptide will be identified as a deamidated/citrullinated peptide because the mass error between the ¹³C-unmodified peptide and its theoretical deamidated/citrullinated counterpart will often be within the given mass tolerance. The frequency of incorrect monoisotopic ion selections vary from 9% to 54% depending on the precursor-ion extracting software.²³

A mass error tolerance limited to 5 ppm for the precursor ion has been explored as an effective means to control the FDR of identified deamidated peptides in previous studies.^{15,16,19} Their rationale is that if the mono-¹³C peak of an *unmodified* peptide is selected as the precursor ion, its theoretical mass error would shift to a positive value.^{16,19} For instance, the theoretical mass error would be +6.4 ppm for a 3000 Da peptide (+0.0193 Da/3000 Da). However, real world experimental mass errors for high resolution MS can easily reach up to 5 ppm; therefore, this approach only partially alleviates the FDR challenge.

Herein, we present a dual-search Delta Score strategy for more accurate identification of deamidated/citrullinated peptides from global shotgun proteomics data acquired by high resolution MS. Our hypothesis is that genuine deamidated/citrullinated peptide fragments in high resolution MS/MS spectra would generate different confidence scores in MSGF+ database searches when allowing a dynamic modification of +0.9840 Da on amino acids asparagine, glutamine, and arginine (NQR), compared to MSGF+ database searches where these dynamic modifications are not allowed. Using this strategy, all spectra in our datasets were searched twice; one search was with +0.9840 Da mass shift as a dynamic modification on NQR, and another search without the modification parameter. For every putative deamidated/citrullinated PSM, a Delta Score was calculated based on the MSGF+ spectral *E*-values (the probability of random match).^{24,25} A “mock modification” search was introduced to evaluate the FDR of identified deamidated/citrullinated PSMs. The final FDR of deamidated/citrullinated PSMs identified could be controlled to ~2% with proper *positive* Delta Score cutoff.

MATERIALS AND METHODS

Datasets

The LC-MS/MS datasets analyzed in this study were described by Dou et al. and can be found using the ProteomeXchange data set identifier [PXD010150].²⁶ Briefly, 10 randomly thin sections of pancreatic islets (~1000 cells) obtained by laser microdissection (LMD) from a presymptomatic type 1 diabetic donor were analyzed by a nanowell-mediated 2DLC-based deep proteome profiling approach.²⁶ MS acquisition was performed on an Orbitrap Fusion Lumos platform (Thermo) in HMS-HCD-HMSn (high resolution MS-high energy collision-induced dissociation-high resolution MSn) mode. Detailed instrumental settings can be found in the publication by Dou et al.²⁶

Data Analysis

LC-MS/MS files (.raw) were first converted into peak list files (_dta.txt) using MSConvert,²⁷ a tool from the proteoWizard infrastructure.²⁸ These lists included extracted monoisotopic mass, charge state, and MS/MS peaks. The resulting peak list files were searched against a human Uniprot proteome (20 198 protein entries, April 2017 release) by MSGF+.²⁵ Methionine oxidation (M, +15.9949) was set as a dynamic modification in all searches. A mass shift of +0.9840 Da was used for identifying deamidation and citrullination on NQR in a dynamic mode. The mass-error tolerance was set to 20 ppm. Isotope error was set from -1 to 1 Da to account for incorrect monoisotopic peak selection. The maximum number of allowable modification sites was set to 3. The target-decoy approach was enabled for

estimating the overall false discovery rate (FDR) of peptide identifications.²⁵ The *A*-score algorithm was applied to determine the site localization of identified deamidated/citrullinated peptides and *A*-score 19 was used for unambiguous site localization.²⁹

FDR Estimation of Modified Peptides Using a Target-Mock Strategy

When a single ¹³C-containing peak of an unmodified peptide (+1.0034 Da shift compared to monoisotopic peak) is misidentified as a deamidated/citrullinated fragment, the mass difference between them is only 0.0193 Da. To estimate the FDR of modified peptides, we implemented a “target-mock” strategy by applying a “mock modification” with a mass shift of +1.0227 Da (equals to 0.0193 + 1.0034) on NQR, along with the +0.9840 Da mass shift for “true” deamidation or citrullination as dynamic modifications in a database search. Our rationale is that the number of modified peptides resulting from random matches should be similar for both the “mock modification” +1.0227 Da database and the “true modification” +0.9840 Da database because they are identical in terms of protein number and amino acid distribution. In addition, the probability of randomly identifying a mono-¹³C peak as a +0.9840 Da peak should be same as randomly recognizing a mono-¹³C peak as a +1.0227 Da peak since both have the same magnitude mass shifts, plus and minus. The FDR was calculated by $2 \times N_M/N$, where N_M and N were the number of mock-modification-containing (+1.0227 Da) PSMs and the total number of modified (+0.9840 Da or +1.0227 Da) PSMs, respectively. The concept of a “mock modification” database is similar to a decoy (e.g., reversed) protein database in a target-decoy strategy.^{30,31} The final Delta Score cutoff will be determined based on the FDR estimation.

RESULTS

The Dual-Search Delta Score Strategy

Figure 1 shows the workflow of the dual-search Delta Score strategy. Briefly, peak-list files from LC-MS/MS raw files (a total 12 datasets with ~34 000 MS/MS spectra per data set) were initially searched by MSGF+ either *with* or *without* + 0.9840 Da mass shift on NQR as a dynamic modification. In this first database search with the modification (Figure 1; step 1), putative deamidated/citrullinated PSMs were filtered *with* a mass-error restriction (20 ppm) to generate output A. In the second search (Figure 1; step 2), the same peak-list files were searched *without* the dynamic mass shift parameter on NQR to yield the output B, where all spectra were identified as unmodified peptides. In this case, no mass error restriction was applied to PSMs in output B to maximize the number of PSMs that can be linked to output A. There were 111 655 and 433 493 PSMs in output A and output B, respectively. The results from output A and B were then linked by PSM scan number (Figure 1; step 3) so that each putative deamidated/citrullinated PSM in output A had two distinct MSGF+ scores from the two searches. Herein, we define the MSGF+ score as $(-\log_{10}(\text{MSGF+ spectrum } E\text{-value}))$. The two MSGF+ scores were then used to compute the Delta Score, defined as the difference of the scores between the modified PSM and unmodified PSM from the same MS/MS spectrum. The resulting putative deamidated/citrullinated PSMs can be categorized into two basic groups: (1) spectra that result in identical sequences from both database searches with and without the +0.9840 Da modification, and (2) spectra that result in different sequences from the two searches. The number of same-sequence spectra and

different-sequence spectra containing putative modifications were 11 734 and 99 609, respectively. In addition, 312 putative deamidated/citrullinated spectra were not identified in the modificationnot-allowed search and a Delta Score value could not be calculated. It should be noted that these 312 spectra were unlikely to be correctly identified because of the poor spectrum *E*-values observed (median *E*-value: 4.04×10^{-3}). The final step (Figure 1; step 4) was to filter putative deamidated/citrullinated PSMs using a proper positive Delta Score cutoff along with mass error and MSGF+ spectrum *E*-value restrictions to generate results.

Delta Score Criteria for Same-Sequence Spectra

Genuine deamidated/citrullinated PSMs in the search with modifications should have mass error distributions centered at zero ppm. Figure 2 shows the mass error and Delta Score distributions of putative deamidated/citrullinated PSMs that produced identical sequences in both database searches. As in Figure 2A, two distinct PSM mass error distributions were separated by a Delta Score cutoff value around zero. For PSMs with Delta Scores >0 , their mass errors were distributed around zero within a narrow range (5 ppm). In contrast, For PSMs with Delta Scores ≤ 0 , their mass errors were randomly distributed over a much broader range.

Figure 2B shows that, after filtering with a more stringent spectral *E*-value cutoff ($<1 \times 10^{-10}$), the mass errors of PSMs whose Delta Scores >0 were still narrowly distributed around zero and the number of these PSMs remained mostly unchanged. By contrast, the number of PSMs whose Delta Score ≤ 0 dramatically decreased after filtering and the mass errors of remaining PSMs were systemically shifted to +7.33 ppm (median value).

Next, we performed an FDR assessment of modified peptides by implementing a “target-mock” strategy where the “mock modification” together with the “true” deamidation or citrullination were allowed as dynamic modifications in a database search (see Materials and Methods section). The influence of the Delta Score, MSGF+ spectrum *E*-value, and mass error on the FDR of identified deamidated/citrullinated PSMs was shown in Figure 2C. The FDRs were ~100% and ~60% when the mass errors >10 ppm and <10 ppm were applied, respectively. The observation of ~100% FDR for PSMs with mass errors >10 ppm is an indication of relative accurate estimation of FDR by the target-mock strategy since nearly all PSMs with >10 ppm were expected to be false IDs on high resolution MS. Adding the spectrum *E*-value ($<1 \times 10^{-10}$) criterion resulted in an FDR reduced to ~30%. Further restricting the mass error to 5 ppm decreased the FDR to ~15%. These results suggested that filtering using only mass error and spectrum *E*-value was insufficient to control the FDR to an acceptable low level. However, implementing a Delta Score value filter resulted in a dramatically reduced FDR overall. For example, a *positive* Delta Score value (>0) with a spectrum *E*-value ($<1 \times 10^{-10}$) and a mass error restriction (<5 ppm) used together controlled the FDR to less than 2.5%. An even lower 1% FDR can be obtained by increasing the Delta Score cutoff (e.g., >4). However, the number of true positives would decrease significantly if the FDR is controlled at 0.7% (Figure 2D). Thus, for the same-sequence spectra, the following final filtering criteria were applied: (1) Delta Score >0 and mass error <5 ppm; (2) Delta Score >4 and mass error <10 ppm, and spectrum *E*-values $<1 \times 10^{-10}$ for

both conditions. The results from each filtering criteria were combined to generate the final list.

After applying the combined criteria, a total of 458 unique deamidated/citrullinated peptides from 302 unique proteins was identified from same-sequence spectra, containing 415 N-deamidated, 34 Q-deamidated, and 9 R-citrullinated peptides (Supporting Information Table S2).

To further illustrate that *positive* Delta Scores originated from the modification-specific fragments in the MS/MS spectra, we manually inspected a subset of spectra. Figure 3 shows an example of confident identification with direct evidence of deamidated fragments. As illustrated, a typical Delta Score >0 for a same-sequence spectrum matched more fragments with a theoretical-deamidated peptide spectrum (Figure 3A) than with its unmodified counterpart (Figure 3B). Extra fragments matched in Figure 3A compared to Figure 3B were all deamidated fragments (labeled in red). This phenomenon was highlighted in Figure 3C where a deamidated fragment can be closely matched (0.0063 Th difference, in this case) with a theoretical-deamidated fragment (dotted line from top), while the same deamidated fragment did not match (-0.492 Th difference, in this case) with the theoretical-unmodified fragment (dashed line at bottom). A list of same-sequence spectra with Delta Scores from 0.04 to 4.01 (Figure S1–S6) were closely examined to see how the Delta Score value influences the identification of deamidated/citrullinated fragments. One deamidated fragment can be found when the Delta Score is 0.04 (Figure S1).

Spectra with negative Delta Score (Figure 4) were also manually inspected. In such cases, no deamidated fragments were identifiable (Figure 4, Figure S7), while corresponding unmodified fragments were detected (Figure 4, Figure S7 labeled in blue). MS1 spectrum showed that the ^{13}C -isotopic peak (labeled by a 4-point star in Figure 4, Figure S7) was incorrectly picked as the monoisotopic ion.

Delta Score Criteria for Different-Sequence Spectra

Besides same-sequence spectra, a subset of identified putative deamidated/citrullinated PSMs in output A (Figure 1) were from spectra that resulted in different peptide sequences from the two database searches. Figure 5 shows the distribution of mass errors and Delta Scores for these different-sequence spectra. Similar to same-sequence spectra, two distinct distributions of mass errors were also separated by a Delta Score cutoff except that the “border” of Delta Score was around $+2$ rather than 0 (Figure 5A). Following filtering with spectrum E-value ($<1 \times 10^{-10}$), the number of PSMs whose Delta Scores were <2 were dramatically decreased (Figure 5B). In contrast to same-sequence spectra (Figure 2B), mass errors of PSMs (from different-sequence spectra) with Delta Scores <0 were also clustered around zero rather than a positive value. Interestingly, we noticed that these spectra (with Delta Scores <0 and mass errors <5 ppm, in region of Figure 5B) were identified as identical peptides from closely related proteins from the two database searches, respectively. In these cases, deamidated asparagine (N# which is D) and deamidated glutamine (Q# which is E) were identified as belonging to a specific protein isoform in one search, could be identified as unmodified aspartate (D) and glutamate (E) in the other search belonging to another protein isoform with the sequence variation. For example, one spectrum was identified as

SDTSGHFQ#R (Q# which is E) which is from protein ANX11, but in the second modification-not-allowed search it was identified as SDTSGHFER from protein ANXA7 (Supporting Information Table S6). In these cases, we could not determine whether deamidation really occurred. It was observed that all PSMs containing N#Q#-to-DE matches have Delta Score less than 2 and majority of them have negative Delta Scores (95%). We also note that there are some PSMs with high Delta Score as well as high mass errors (Figure 5B). Given the high mass errors, these PSMs are most likely wrong identifications especially since ~24% of these PSMs are from decoy database (Supporting Information Table S7).

Thus, for the different-sequence spectra, the following final filtering criteria was applied: Delta Score >2, mass error <5 ppm, and spectrum E-value < 1×10^{-10} . After applying the criteria, 519 deamidated/citrullinated PSMs were identified, where 424 and 95 were 1-site and 2-site deamidated/citrullinated PSMs, respectively.

For 1-site modified PSMs, following the application of Delta Score criteria, 283 unique peptides were identified, among which 269 were N-deamidated, 9 were Q-deamidated, and 5 were R-citrullinated (Supporting Information Table S3). Unique peptides were selected from PSMs with the highest Delta Score. Again, when Delta Score is >2 for a different-sequence spectrum, deamidated/citrullinated fragments were detectable (Figure S8A). The spectrum matched better with the theoretical spectrum of modified peptides identified using the modification-allowed search (compare Figure S8A with Figure S8B). The overall FDR for different-sequence 1-site PSMs was 2.75%. In this case, the 2-site modified PSMs were excluded from our final results due to the apparent higher likelihood of false positives of peptides with multiple modification sites.

Summary of Final Identified Deamidated and Citrullinated Peptides

One of the common challenges of PTM identification is site localization. Herein, although we applied the A-score as a probability based algorithm for determining the PTM sites,²⁹ the exact localization of many modification sites was still ambiguous (e.g., A-score <19). In total, 689 unique deamidated/citrullinated peptides from 432 unique proteins were identified, and 400 of them have an A-score greater than 19. Overall, 632 N-deamidated, 43 Q-deamidated, and 14 R-citrullinated peptides were identified (Supporting Information Table S1). Interestingly, among the identified citrullinated proteins, many of them were reported as possible autoantigens. For example, PDIA1 was reported as a possible citrullinated autoantigen, and several others are biochemically confirmed autoantigens, including vimentin,¹³ Cytokeratin-2e (K22E),³² V-ATPase subunit A (VATA).³³ Two deamidated sites (DGQVIN#ETSQHDDLE, ETN#LDSLPLVDTHS) identified from vimentin were also biochemically confirmed as immunogenic epitopes (Immune Epitope Database (IEDB) epitope ID 772963 and 554545). Another immunogenic epitope containing deamidation (VVHVN#GYGK, IEDB epitope ID 860180) was identified on dihydrolipoyl dehydrogenase (DLDH). A number of deamidated/citrullinated proteins were also previously reported as possible autoantigens of type 1 diabetes, including the 60 kDa heat shock protein (CH60), islet-specific glucose-6-phosphatase-related protein (G6PC2), and 78

kDa glucose-regulated protein (GRP78), demonstrating deamidation and citrullination may play a role in autoimmunity.^{11,34–36}

Assessment of the Delta-Score Strategy Using a Published Citrullination Data Set

A recent publication¹⁷ reported one of the most advanced approaches for identifying citrullination by integrating deep proteomic profiling with manual annotation for the modification. To further demonstrate the utility of our dual-search Delta-Score strategy, we reanalyze this independent data set using our approach. As shown in Figure 6A, most “valid” PSMs annotated in the original report were also observed with a *positive* Delta Score (~95%) based on our reanalysis. Moreover, the majority PSMs have mass error less than 5 ppm (~98%). This again demonstrated the effectiveness of our strategy to identify “valid” citrullinated peptides which should have positive Delta Score and low mass error. However, our analysis does reveal a small portion of the original “valid” PSMs with *negative* Delta Score values, which we consider as likely “invalid” IDs (Figure 6A, PSMs in red circles). Furthermore, no citrullinated fragment ions were found in these “valid” PSMs with *negative* Delta Scores (examples shown in Supporting Information Figure S9).

When all same-sequence and different-sequence PSMs were presented (Figure 6B, 6C), it was observed that many PSMs with high Delta Score and low mass errors (<5 ppm) were previously annotated as “ambiguous or false” or “not inspected”. In these cases, these PSMs are likely to be either “valid” citrullinated or deamidated peptides. Finally, similar levels of FDRs were observed for this data set compared to our own data set (Figure 6D).

We should note that our Delta Score strategy offers clear advantages over the previous approach. The previous approach required multiple prefiltering steps based on an empirical assumption that “true” citrullinated sites should not be cleavage by trypsin. However, this is not accurate since the authors acknowledged that some “true” citrullinated sites may also be cleavable by trypsin.³⁷ Another prefiltering criterion was the detection of the loss of isocyanic acid, which is not specific to citrullination as acknowledged by the authors. Most importantly, the previous approach validated citrullinated peptides mainly through manual inspection,^{17,37} which is not feasible in many cases. In contrast, our strategy utilizes a simple Delta Score concept to differentiate true positives from false positives.

DISCUSSION

Due to the minimal difference in mass between a peptide that contains a deamidated/citrullinated amino acid and a peptide that contains the naturally occurring heavy isotope ¹³C, it can be difficult to accurately select and differentiate the monoisotopic mass of a modified peptide from MS1 spectra used for database searching. Previous reports suggest that controlling the mass error of the precursor ions to within 5 ppm should eliminate the possibility that an isotopically heavy ¹³C containing peptide is selected as the monoisotopic mass leading to a misidentification of deamidation.^{16,19} However, our data show that a 5-ppm cutoff is simply not enough to accurately identify such modified peptides (indeed, ~15% FDR observed in current datasets, Figure 2C). Instead, we demonstrated that a dual-search Delta Score strategy is effective for accurate identifications of deamidation and citrullination with a controlled FDR (~2%). Furthermore, a *positive* Delta Score is the result

of matching deamidated/citrullinated peptide fragments (true positive) with higher confidence scores in searches that allow modifications than in searches that do not allow modifications. Spectra with *positive* Delta Scores were likely correct identifications because the mass errors of these spectra were narrowly distributed around zero (Figure 2 and 5). In contrast, spectra with zero or negative Delta Score values were unlikely to be identified correctly. Interestingly, the mass error of peptides with Delta Scores <0 were shifted in the positive direction with a median mass error of 7.33 ppm (Figure 2B). This observed positive mass error shift is mostly likely due to the incorrect selection of a ^{13}C peak as the monoisotopic ion which results in a mass shift of 0.0193 Da higher than deamidation. If the average peptide mass is ~ 2500 , such monoisotopic peak selection errors will result in a mass error of ~ 7.7 ppm. One significant advantage of the Delta Score strategy is it can overcome the limitations of mass error filtering alone. Indeed, many spectra with negative Delta Scores have a mass error less than 5 ppm, which were less likely to be correct identifications. An example of such a case is shown in Figure 4, where a PSM was identified with a mass error <5 ppm, but was identified incorrectly and had a Delta Score <0 .

Our results suggest that there are three main types of errors in identifying deamidated/citrullinated peptides: (1) Low-quality spectra with spectrum *E*-values typically $>1 \times 10^{-10}$ resulting in peptide sequences that were not confident, and their mass errors were randomly distributed (Figure 2A). Implementing a MSGF+ spectrum *E*-value filtering (e.g., $<1 \times 10^{-10}$) can dramatically decrease this type of error (e.g., Figure 2B). (2) Incorrectly selecting ^{13}C peak as monoisotopic ion mass. The Delta Score strategy clearly differentiates this cluster of PSMs from true-modified peptides (Figure 2B). (3) Spectra identified as deamidated peptides but belonging to an unmodified variant of another closely related protein (deamidated NQ vs DE) in the proteome database. This type of error is inherent to the protein database. However, it can be avoided by increasing Delta Score cutoff (e.g., >2.0 in this case, Supporting Information Table S6). All three types of errors can be well controlled with a *positive* Delta Score cutoff (Figure 1C).

There are also several unique merits of MSGF+ that makes it a good choice for database searching with the Delta Score strategy. First, MSGF+ allows successful nonmodification searches. A spectrum identified as a deamidated/citrullinated peptide in a modification-allowed database can also be identified as an unmodified peptide in modification-not-allowed database search because MSGF+ provides options to decrease one or more 1 Da mass-shift masses of the precursor ion.²⁵ This means a spectrum is searched against a database multiple times with all possible precursor ion masses within a mass error restriction (20 ppm in this study) and a spectrum *E*-value is calculated for each search. Only $\sim 0.3\%$ (312 out of 111 655) putative deamidated/citrullinated spectra were not identified in modification-not-allowed database search and these spectra were very unlikely to be true because of their lowconfidence spectrum *E*-values (median 4.04×10^{-3}). Second, the statistical significance of an *individual* PSM is accurately estimated by MSGF+ spectrum *E*-value²⁴ that is *independent* from spectrum *E*-value of other PSMs. Other database searching algorithms with probability scores such as the X! Tandem³⁸ and Comet³⁹ are also applicable with our Delta Score strategy (Supporting Information Figure S10). However, these scores may be not as rigorous as MSGF+ spectrum *E*-value in characterizing an *individual* PSM's

statistical significance because the probability scores of these algorithms all depend on the score distribution of other spectra identified in their database search.

In summary, we have demonstrated a dual-search Delta Score strategy for improving identifications of deamidated/citrullinated peptides from a typical global shotgun proteomics data set. The FDR of deamidated/citrullinated peptides can be controlled to a reasonable level (~2%) by a proper positive Delta Score cutoff together with a mass error and MSGF+ spectrum *E*-value restriction. The data also show that such modifications are relatively rich in tissue proteomics datasets (e.g., human pancreatic islets in the case). The Delta Score strategy could be a powerful tool for discovering citrullination-related autoantigens in autoimmune diseases such as type 1 diabetes. Further validation of in situ PTMs can be pursued with synthetic peptides with MS. The strategy may also be modified to identify other kinds of challenging post-translational modifications in global shotgun proteomics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Portions of this work were supported by NIH Grants DP3 DK110844 and R01 DK122160 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and P41GM103493 from the National Institutes of General Medical Sciences (NIGMS). X.W. is also partially supported by China Scholarship Council (CSC). The experimental work was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE and located at Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the DOE under Contract DE-AC05-76RL0 1830.

REFERENCES

- (1). Tonie Wright H; Urry DW Nonenzymatic deamidation of asparaginy and glutaminyl residues in protein. Crit. Rev. Biochem. Mol. Biol 1991, 26 (1), 1–52. [PubMed: 1678690]
- (2). Peters B; Trout BL Asparagine deamidation: pH-dependent mechanism from density functional theory. Biochemistry 2006, 45 (16), 5384–5392. [PubMed: 16618128]
- (3). van de Wal Y; Kooy Y; van Veelen P; Peña S; Mearin L; Papadopoulos G; Koning F Cutting edge: selective deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity. J. Immunol 1998, 161 (4), 1585–1588. [PubMed: 9712018]
- (4). Geiger T; Clarke S Deamidation, isomerization, and racemization at asparaginy and aspartyl residues in peptides. Succinimide-linked reactions that contribute to protein degradation. J. Biol. Chem 1987, 262 (2), 785–794. [PubMed: 3805008]
- (5). Robinson AB; McKerrow JH; Cary P Controlled deamidation of peptides and proteins: an experimental hazard and a possible biological timer. Proc. Natl. Acad. Sci. U. S. A 1970, 66 (3), 753–757. [PubMed: 5269237]
- (6). Robinson NE; Robinson AB Deamidation of human proteins. Proc. Natl. Acad. Sci. U. S. A 2001, 98 (22), 12409–12413. [PubMed: 11606750]
- (7). Dilley KJ; Harding JJ Changes in proteins of the human lens in development and aging. Biochim. Biophys. Acta, Protein Struct 1975, 386 (2), 391–408.
- (8). Inaba M; Gupta KC; Kuwabara M; Takahashi T; Benz EJ Jr.; Maede Y Deamidation of human erythrocyte protein 4.1: possible role in aging. Blood 1992, 79 (12), 3355–3361. [PubMed: 1596575]
- (9). Adav SS; Gallart-Palau X; Tan KH; Lim SK; Tam JP; Sze SK Dementia-linked amyloidosis is associated with brain protein deamidation as revealed by proteomic profiling of human brain tissues. Mol. Brain 2016, 9 (1), 20. [PubMed: 26892330]

- (10). Clancy KW; Weerapana E; Thompson PR Detection and identification of protein citrullination in complex biological systems. *Curr. Opin. Chem. Biol* 2016, 30, 1–6. [PubMed: 26517730]
- (11). Rondas D; Crèvecoeur I; D'Hertog W; Ferreira GB; Staes A; Garg AD; Eizirik DL; Agostinis P; Gevaert K; Overbergh L; Mathieu C Citrullinated Glucose-Regulated Protein 78 Is an Autoantigen in Type 1 Diabetes. *Diabetes* 2015, 64 (2), 573. [PubMed: 25204978]
- (12). McGinty JW; Marré ML; Bajzik V; Piganelli JD; James EA T cell epitopes and post-translationally modified epitopes in type 1 diabetes. *Curr. Diabetes Rep* 2015, 15 (11), 90.
- (13). Bang H; Egerer K; Gauliard A; Lüthke K; Rudolph PE; Fredenhagen G; Berg W; Feist E; Burmester GR Mutation and citrullination modifies vimentin to a novel autoantigen for rheumatoid arthritis. *Arthritis Rheum.* 2007, 56 (8), 2503–2511. [PubMed: 17665451]
- (14). Schellekens GA; de Jong BAW; van den Hoogen FHJ; Van de Putte LB; van Venrooij WJ Citrulline is an essential constituent of antigenic determinants recognized by rheumatoid arthritis-specific autoantibodies. *J. Clin. Invest* 1998, 101 (1), 273–281. [PubMed: 9421490]
- (15). Hao P; Adav SS; Gallart-Palau X; Sze SK Recent advances in mass spectrometric analysis of protein deamidation. *Mass Spectrom. Rev* 2017, 36 (6), 677–692. [PubMed: 26763661]
- (16). Nepomuceno AI; Gibson RJ; Randall SM; Muddiman DC Accurate identification of deamidated peptides in global proteomics using a quadrupole orbitrap mass spectrometer. *J. Proteome Res* 2014, 13 (2), 777–785. [PubMed: 24289162]
- (17). Lee C-Y; Wang D; Wilhelm M; Zolg DP; Schmidt T; Schnatbaum K; Reimer U; Pontén F; Uhlén M; Hahne H; et al. Mining the human tissue proteome for protein citrullination. *Mol. Cell. Proteomics* 2018, 17 (7), 1378–1391. [PubMed: 29610271]
- (18). Bennike T; Lauridsen KB; Olesen MK; Andersen V; Birkelund S; Stensballe A Optimizing the identification of citrullinated peptides by mass spectrometry: utilizing the inability of trypsin to cleave after citrullinated amino acids. *J. Proteomics Bioinf* 2013, 6 (12), 288–295.
- (19). Hao P; Ren Y; Alpert AJ; Sze SK Detection, evaluation and minimization of nonenzymatic deamidation in proteomic sample preparation. *Mol. Cell. Proteomics* 2011, No. O111.009381.
- (20). Eng JK; McCormack AL; Yates JR An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom* 1994, 5 (11), 976–989. [PubMed: 24226387]
- (21). Perkins DN; Pappin DJC; Creasy DM; Cottrell JS Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, 20 (18), 3551–3567. [PubMed: 10612281]
- (22). Cox J; Neuhauser N; Michalski A; Scheltema RA; Olsen JV; Mann M Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res* 2011, 10 (4), 1794–1805. [PubMed: 21254760]
- (23). Hao P; Ren Y; Tam JP; Sze SK Correction of errors in tandem mass spectrum extraction enhances phosphopeptide identification. *J. Proteome Res* 2013, 12 (12), 5548–5557. [PubMed: 24147958]
- (24). Kim S; Gupta N; Pevzner PA Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res* 2008, 7 (8), 3354–3363. [PubMed: 18597511]
- (25). Kim S; Pevzner PA MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun* 2014, 5, 5277. [PubMed: 25358478]
- (26). Dou M; Zhu Y; Liyu A; Liang Y; Chen J; Piehowski PD; Xu K; Zhao R; Moore RJ; Atkinson MA; et al. Nanowell-mediated two-dimensional liquid chromatography enables deep proteome profiling of < 1000 mammalian cells. *Chem. Sci* 2018, 9 (34), 6944–6951. [PubMed: 30210768]
- (27). Adusumilli R; Mallick P Data conversion with ProteoWizard msConvert. In *Proteomics*; Springer, 2017; pp 339–368.
- (28). Kessner D; Chambers M; Burke R; Agus D; Mallick P Proteo Wizard: open source software for rapid proteomics tools development. *Bioinformatics* 2008, 24 (21), 2534–2536. [PubMed: 18606607]
- (29). Beausoleil SA; Villen J; Gerber SA; Rush J; Gygi SP A probability-based approach for high-throughput protein phosphor-ylation analysis and site localization. *Nat. Biotechnol* 2006, 24 (10), 1285–92. [PubMed: 16964243]

- (30). Elias JE; Gygi SP Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, 4 (3), 207. [PubMed: 17327847]
- (31). Qian W-J; Liu T; Monroe ME; Strittmatter EF; Jacobs JM; Kangas LJ; Petritis K; Camp DG; Smith RD Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res* 2005, 4 (1), 53–62. [PubMed: 15707357]
- (32). Wang Q; Drouin EE; Yao C; Zhang J; Huang Y; Leon DR; Steere AC; Costello CE Immunogenic HLA-DR-presented self-peptides identified directly from clinical samples of synovial tissue, synovial fluid, or peripheral blood in patients with rheumatoid arthritis or lyme arthritis. *J. Proteome Res* 2017, 16 (1), 122–136. [PubMed: 27726376]
- (33). Collado JA; Alvarez I; Ciudad MT; Espinosa G; Canals F; Pujol-Borrell R; Carrascal M; Abian J; Jaraquemada D Composition of the HLA-DR-associated human thymus peptidome. *Eur. J. Immunol* 2013, 43 (9), 2273–2282. [PubMed: 23719902]
- (34). Lillicrap MS; Duggleby RC; Goodall JC; Gaston JSH T cell recognition of a highly conserved epitope in heat shock protein 60: self-tolerance maintained by TCR distinguishing between asparagine and aspartic acid. *International immunology* 2004, 16 (3), 405–414. [PubMed: 14978014]
- (35). Yang J; Danke NA; Berger D; Reichstetter S; Reijonen H; Greenbaum C; Pihoker C; James EA; Kwok WW Islet-specific glucose-6-phosphatase catalytic subunit-related protein-reactive CD4+ T cells in human subjects. *J. Immunol* 2006, 176 (5), 2781–2789. [PubMed: 16493034]
- (36). Roep BO; Peakman M Antigen targets of type 1 diabetes autoimmunity. *Cold Spring Harbor Perspect. Med* 2012, 2 (4), No. a007781.
- (37). Bennike T; Lauridsen KB; Olesen MK; Andersen V; Birkelund S; Stensballe A Optimizing the identification of citrullinated peptides by mass spectrometry: utilizing the inability of trypsin to cleave after citrullinated amino acids. *J. Proteomics Bioinf* 2013, 6, 288–295.
- (38). Fenyö D; Beavis RC A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem* 2003, 75 (4), 768–774. [PubMed: 12622365]
- (39). Eng JK; Jahan TA; Hoopmann MR Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013, 13 (1), 22–24. [PubMed: 23148064]

Figure 1.

General workflow of Delta Score strategy. The peak list extracted from a raw MS file was first searched against protein database (1) with deamidation/citrullination (+0.9840 Da on amino acid residue NQ/R) as a dynamic modification with a mass error less than 20 ppm as output A. Another MSGF+ database search without the dynamic modification (2) was applied to the same peak-list file to generate output B. (3) PSMs in output B were linked to output A by scan number (or spectrum ID). Consequently, each spectrum (scan) in output A identified as a putative deamidated/citrullinated peptide will have two MSGF+ score ($-\log(\text{MSGF+ spectrum } E\text{-value})$). The Delta Score of each spectrum was then calculated as the difference of the two MSGF+ score. PSMs in Output A were then divided to two categories, same-sequence spectra or different-sequence spectra, which were spectra resulting in same-sequence peptides or different-sequence peptides in the two database searches, respectively. (4) The two categories of spectra were filtered by applying a positive Delta Score cutoff along with a spectrum *E*-value cutoff (typically less than 1×10^{-10}) and a mass error cutoff (typically less than 5 ppm), thereby yielding final PSMs likely to contain deamidated/citrullinated fragments in their MS/MS spectra.

Figure 2.

(A) Delta Score and mass error distributions of putative same-sequence deamidated/citrullinated spectra. Note that the observed systematic mass measurement error of +5.42 ppm was corrected to zero in the plot. (B) Delta Score and mass error distributions of more confident same-sequence spectra (spectrum E -value $< 1 \times 10^{-10}$). (C) Effect of Delta Score, mass error, and spectrum E -value on the FDR of modified PSMs. (D) Effect of Delta Score on the number of PSMs identified.

Figure 4.

A typical spectrum with negative Delta Score value. The spectrum was identified as a deamidated (Q#) peptide (A) and a same-sequence unmodified peptide (B). Mono- ^{13}C -ion (indicated by 4-point star) was incorrectly picked as monoisotopic ion as shown in MS1 (C).

Figure 5.

Delta Score and mass error distribution of putative different-sequence spectra. The mass error distribution of *all* different-sequence spectra (A) and more *confident* (spectrum *E*-value 1×10^{-10}) different-sequence spectra (B). Note that the systematic mass measurement error of +5.42 ppm was corrected to zero in the plot, which led to an apparent mass range of 15 ppm to -25 ppm for the ± 20 ppm database searching window. (C) Effect of Delta Score, mass error and spectrum *E*-value on the FDR of modified PSMs. Note that the observed FDR increases with higher Delta Scores are presumably due to the low number of positive IDs in this region. (D) Effect of Delta Score on the number of target and mock PSMs identified.

Figure 6.

Assessment of the Delta Score strategy using a recent published data set.¹⁷ (A) Distribution of "valid" PSMs. (B) Distribution of same-seq PSMs annotated with different categories. (C) Distribution of diff-seq PSMs annotated with different categories. (D) The relationship between Delta Score, PSMs, and FDR. Filtering criteria: mass errors < 5 ppm, spectrum E-value < 1×10^{-10} .