Published in final edited form as:

Anal Chem. 2017 January 17; 89(2): 1244–1253. doi:10.1021/acs.analchem.6b03874.

Multiplexed Post-Experimental Monoisotopic Mass Refinement (*m*PE-MMR) to Increase Sensitivity and Accuracy in Peptide Identifications from Tandem Mass Spectra of Cofragmentation

Inamul Hasan Madar^{†,§}, Seung-Ik Ko^{†,§}, Hokeun Kim[†], Dong-Gi Mun[†], Sangtae Kim[‡], Richard D. Smith[‡], and Sang-Won Lee^{†,*}

[†]Laboratory of Gaseous Ion Chemistry, Department of Chemistry, Research Institute for Natural Sciences, Korea University, Seoul 136-701, South Korea

[‡]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, United States

Abstract

Mass spectrometry (MS)-based proteomics, which uses high-resolution hybrid mass spectrometers such as the quadrupole-orbitrap mass spectrometer, can yield tens of thousands of tandem mass (MS/MS) spectra of high resolution during a routine bottom-up experiment. Despite being a fundamental and key step in MS-based proteomics, the accurate determination and assignment of precursor monoisotopic masses to the MS/MS spectra remains difficult. The difficulties stem from imperfect isotopic envelopes of precursor ions, inaccurate charge states for precursor ions, and cofragmentation. We describe a composite method of utilizing MS data to assign accurate monoisotopic masses to MS/MS spectra, including those subject to cofragmentation. The method, “multiplexed post-experiment monoisotopic mass refinement” (*m*PE-MMR), consists of the following: multiplexing of precursor masses to assign multiple monoisotopic masses of cofragmented peptides to the corresponding multiplexed MS/MS spectra, multiplexing of charge states to assign correct charges to the precursor ions of MS/MS spectra with no charge information, and mass correction for inaccurate monoisotopic peak picking. When combined with MS-GF+, a database search algorithm based on fragment mass difference, *m*PE-MMR effectively

*Corresponding Author: sw_lee@korea.ac.kr. Fax: 82-2-3290-3121.

§Author Contributions

These authors made equal contributions to this work.

ORCID

Sang-Won Lee: 0000-0002-5042-0084

Notes

The authors declare no competing financial interest.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.anal-chem.6b03874. Experimental Section, iterative database strategy, and peptide identification files (PDF).

Supplemental Table S1 (PDF).

Supplemental Table S2 (PDF).

Supplemental Table S3 (PDF).

Supplemental Table S4 (PDF).

Supplemental Table S5 (PDF).

Supplemental Table S6 (PDF).

increases both sensitivity and accuracy in peptide identification from complex high-throughput proteomics data compared to conventional methods.

Graphical Abstract

Monoisotopic mass calculated from an isotopic envelope of a precursor ion selected for fragmentation is fundamental to peptide identification using database search engines. Typical search engines, such as SEQUEST¹ and MASCOT,² use precursor masses (i.e., monoisotopic masses) to find matches with theoretical peptides from a protein database within a specified mass tolerance. Each of the search engines subsequently yields a score only for the matched candidate peptides using its own scoring method (i.e., cross-correlation function in SEQUEST and probability score function in MASCOT, etc.). In the first “screening” step, which takes into account precursor mass, the number of matches for candidate peptides increases with wider mass tolerance. Modern high-resolution and high-speed mass spectrometric techniques allow mass measurement accuracies of low ppm in large scale bottom-up experiments, which greatly reduces the search spaces (i.e., the number of matched candidates for subsequent scoring) for peptide identification by using a tight tolerance (i.e., 10 ppm) for precursor mass, thereby increasing the identification accuracy of the resultant peptides. However, confidently assigning correct monoisotopic masses to the corresponding MS/MS spectra is often nontrivial and remains the subject of much research interest.^{3–5} The difficulties mainly arise from two issues: imperfect isotopic envelope of precursor ions and cofragmentation within the isolation window for fragmentation.

In order to overcome the under-sampling imposed by the complexities of a peptide mixture and insufficient analysis speed and sensitivity, precursor peptide ions are typically selected in a data-dependent MS/MS experimental scheme, in which ions are dynamically selected for fragmentation and excluded from an MS/MS event for a certain duration (the exclusion duration) in order to prevent reacquisition of MS/MS spectra of the same peptides. While this experimental method of dynamic selection of precursors with exclusion greatly alleviates the under-sampling of bottom-up experiments and increases the chances of MS/MS identification of low-abundance peptides in a complex proteome sample, it often leads to fragmentation of peptides at an early point in chromatographic elution. The situation entails obtaining monoisotopic mass information for peptides from very weak MS signals, in which the actual isotopic distributions tend to significantly deviate from the theoretical ones, often missing the true monoisotopic masses and thereby resulting in inaccurate calculation of monoisotopic masses. The resultant MS/MS spectra tend to contain precursor masses of

C13 isotopic peaks instead of the true monoisotopic peak (i.e., C13 isotope peak picking), and the use of a narrow precursor mass tolerance (e.g., 10 ppm) leads to false identification. A recent study showed that ~40% of precursor masses of conventionally processed MS/MS data contains incorrect monoisotopic masses.⁴

Even using state-of-the-art peptide separation techniques with high separation resolution, peptide ions of similar m/z and chromatographic retention times are likely present owing to the sheer complexity of a proteome sample, and they undergo cofragmentation, resulting in multiplexed (or chimeric or mixture) MS/MS spectra. Mallick and co-workers estimated that cofragmentation accounts for up to 50% of MS/MS spectra in a typical bottom-up experiment.⁶ The difficulty that cofragmentation imposes on correct peptide identification is 2-fold: inaccurate precursor charge determination and complex chimeric MS/MS spectra. When conventional methods are used to extract information on picked precursor m/z and charge state, the overlapping isotopic envelopes create difficulties in determining the charge state of the selected precursor ions owing to the interference of other isotopic peaks within the isolation window, and this situation results in a lack of charge state for the corresponding multiplexed MS/MS spectra. Conventional methods would then arbitrarily assign +2 and +3 charges for the MS/MS data with unknown charge, resulting in false identifications, especially when the true charge states are neither +2 nor +3. Even the case where either +2 or +3 is the true charge state, as conventional tools assume one peptide for one MS/MS spectrum, the presence of cofragmented peptides decreases the peptide identification rate owing to the presence of unannotated fragments, making the peptide scores lower.⁷ Various computational tools have been designed to address the unassigned charge state problem^{3,8–14} and to identify multiple peptides from multiplexed MS/MS spectra.^{12–24} The “Charger” algorithm uses a combination of signal processing and statistical machine learning to address charge determination of ETD spectra.⁹ The “Charge Prediction Machine (CPM)” assigns a charge state up to +7 for ETD data using Bayesian decision theory and reduces search time compared to all-hypothesis searches.¹⁰ “YADA” uses identification of isotopic envelopes by considering the charge states for a given peak up to +21 and by calculating matches between determined and average theoretical envelopes. It also considers multiple peptide ions for a given isolation window.¹¹ The “Hardklör” algorithm was developed to deconvolute overlapping peptide isotope distributions (PIDs) and determine charge states for the overlapping PIDs.¹² “MixGF” calculates a probability-based generation function for the identification of multiple peptides.²³ In the sequential peptide identification method based on the Andromeda/MaxQuant workflow, peptide MS features from MaxQuant were sequentially used for peptide identification of multiplexed MS/MS spectra by subtracting assigned fragments after each identification.¹⁸ “DeMix” workflow uses various previously reported methodologies with deconvolution of chimeric MS/MS spectra (spectral cloning) with the addition of a simple rescoring method for identification of cofragmented spectra.²⁴

There have been developments in database search engines that employ computing scores that are minimally affected by the presence of cofragmentation, such as MS-GF+²⁵ and MODa.²⁶ For example, when computing a SpecEval score, MS-GF+ considers a score distribution of all amino acid sequences matching the precursor mass; thus, the existence of several cofragmented peptides has only a minimal effect on SpecEval scores. It was previously demonstrated that MS-GF+ resulted in correct IDs for 98% of multiplexed MS/MS spectra

upon assigning correct precursor masses to a multiplexed MS/MS spectrum, even if the precursor isolation purity was as low as 50%.²⁷

In this paper, we report a technique for multiplexed post-experiment monoisotopic mass refinement, *m*PE-MMR, that utilizes peptide MS features (i.e., unique mass class or UMC) in LC-MS/MS data to decipher charge states and precursor monoisotopic masses of multiplexed MS/MS spectra. The technique performs charge multiplexing for precursor peaks, deconvolution of overlapping isotopic envelopes, and, finally, mass correction for “C13 isotope peak picking.” When combined with a state-of-the-art database search engine, MS-GF+, *m*PE-MMR provides more effective MS/MS data identification by allowing for identification of multiple peptides from multiplexed MS/MS spectra, along with a reduction in charge ambiguity for MS/MS spectra with unknown charge and thereby increases both the sensitivity and accuracy of peptide identification from complex high-throughput proteomics data compared to conventional methods.

EXPERIMENTAL SECTION

Detailed descriptions of the samples, LC-MS/MS analysis, and data analysis methods can be found in Supporting Information.

MS and MS/MS Data Acquisition

Briefly, full MS scans (m/z 400–2000 Th) were acquired at a resolution of 70000 with maximum ion injection times of 20 ms. The top 10 most abundant precursor ions were fragmented with an isolation window of ± 0.8 Th with an exclusion time of 30 s and at a normalized collision energy (NCE) of 30 for higher energy collisional dissociation (HCD) by data-dependent MS/MS experiments. The peptide match was disabled and the +1 charge was discarded. The MS/MS scans were acquired with a first fixed m/z of 100 Th, a resolution of 17500, and a maximum ion injection time of 60 ms. The automatic gain control (AGC) target value was set to 1.0×10^6 for both full MS and MS/MS scan.

MS Data Processing of *m*PE-MMR: UMC Generation

As indicated in Figure 1A, the RAPID algorithm was used to calculate monoisotopic masses from MS data.²⁸ RAPID has been shown to effectively identify overlapping isotopic envelopes even when the overlapping isotopic envelopes share one or more peaks. The monoisotopic masses obtained by RAPID deisotoping were subsequently grouped into unique mass classes (UMCs) using in-house software.²⁹ The MS spectra of a peptide are measured in series with the masses within a few ppm. An in-house software program performs clustering of sequential MS peaks with similar masses into unique mass classes (UMCs).^{29,30} Since monoisotopic masses are used for the UMC grouping, each UMC contains information from all measured charge states of a peptide. Each UMC contains other information such as MS intensity (measured at different elution times at different charges) and elution times for a given peptide. As previously described, an intensity-weighted average mass and a median elution time were calculated for each UMC and used as UMC mass and UMC elution time, respectively.²⁹ Ideally, a UMC represents a peptide by its UMC

mass and UMC elution time. All UMCs from an LC-MS/MS data were recorded in a UMC list (Figure 1A).

MS/MS Data Processing of *m*PE-MMR: Charge State Multiplexing and C13 Mass Correction

*m*PE-MMR uses the msconvert tool (ProteoWizard 3.0.7162)³¹ to generate mzxml files from experimental raw files. It subsequently uses MZXML2Search (TPP v4.5 RAPTURE rev 2)³² to create mgf files from the mzxml files (Figure 1A). *m*PE-MMR uses the targeted *m/z* information recorded in the mgf files to calculate the experimental isolation window (i.e., targeted $m/z \pm 0.8$ Th) for fragmentation. *m*PE-MMR discards the charge state information recorded in the raw file. *m*PE-MMR subsequently calculates monoisotopic masses for each *m/z* value (i.e., 621.79 Th in Figure 1B) by assuming charge states from +1 to +7 (“charge multiplexing”), along with mass subtraction of multiples of 1.00235 Da up to three times (i.e., -1, -2, and -3 Da) and mass addition of 1.00235 Da (i.e., +1 Da) for each calculated monoisotopic mass obtained for each of seven charges (“C13 mass corrections”), creating a total of 35 possible monoisotopic masses (Figure 1B). These considerations account for 99% of charge assignments to MS/MS data and 99.9% of picking isotopically substituted peaks. *m*PE-MMR uses all 35 masses to find matches with UMCs within 10 ppm and ± 5 MS scans (i.e., range scanning).²⁹ When no match is found, it discards the 35 monoisotopic masses, but not the MS/MS scan. When a match is found with a charge state, it generates MS/MS spectral data with the matched *m/z* value and charge. When two or more matches are found during charge multiplexing, it generates MS/MS spectral data of matched UMC masses for an MS-GF+ database search.

MS/MS Data Processing of *m*PE-MMR: Deconvolution of Overlapping Isotopic Envelopes by Removal of Matching Isotope Envelopes

When a match is found with the recorded *m/z* (e.g., +4 for 621.79 Th with no C13 isotope correction in Figure 1B), *m*PE-MMR removes all corresponding isotopic peaks (marked in red in Figure 1B) of the matched ion from the isolation window (i.e., ± 0.8 Th) by calculating the masses of the theoretical C13 isotopes using the UMC mass and the corresponding charge (“matching isotope envelope removal”). *m*PE-MMR subsequently performs charge multiplexing, C13 mass correction, and matching isotope envelope removal for each of the remaining peaks in the isolation window, starting from the lowest *m/z* peak within the isolation window (621.32 Th in Figure 1B). Every time a match between the calculated monoisotopic mass for a cofragmented ion and a UMC mass is found, the corresponding isotopic envelopes are removed from the isolation window before the next round of charge multiplexing/C13 mass correction/ matching isotope envelope removal is done for the next lowest *m/z* peak (621.35 Th in Figure 1B) until no peaks are left in the isolation window. Since *m*PE-MMR performs C13 mass correction for each charge, as shown Figure 1B, it correctly finds a match even if the monoisotopic peak is missing from the isolation window of the reference MS scan, as long as they appear in adjacent scans (i.e., during range scanning). After the multiplexed matching of precursor masses with UMC masses, *m*PE-MMR generates a modified MGF file (*m*PE-MMR MGF in Figure 1A).

Use of *m*PE-MMR MGF for Database Search and Target-Decoy Analysis

Graphical descriptions of the database search can be found in Supporting Information. In the conventional method, to process an MS/MS spectrum of unassigned charge, MS-GF+ identified two peptides, one assuming charge +2 and the other assuming charge +3, and selected only one with the higher score for target-decoy analysis. In the *m*PE-MMR method, we used two different database search schemes: a single-round database search and an iterative database search. In the single-round database search, we used the *m*PE-MMR MGF for a typical one-time MS-GF+ search. In the iterative database search (Figure S1), we first ran a MS-GF+ search with target precursor ion mass (1st round). After *i*-th round, we removed the peaks matched to the identified peptide within 0.01 Th from the MS/MS spectra. The subtracted MS/MS spectra were used for the (*i* + 1)-th round of database search using the next precursor masses (i.e., from the second round, precursor masses were ordered by their mass values). The multiround database search was continued until no precursor mass left for database search (e.g., up to 11th round for the data set used in this study). After the last round database search, all PSMs were combined for target-decoy analysis. Only PSMs that were selected by the following rules were used: (1) when one MS peak matches with more than one UMC during charge multiplexing and C13 mass correction, all clone MS/MS spectra of different monoisotopic masses are subjected to MG-GF+ search, but only the PSM with the highest peptide score is selected for target-decoy analysis, (2) when more than one MS peak in the isolation window, after deconvolution by matching isotope envelope removal, are matched to different UMCs of different monoisotopic masses, the resulting PSMs are treated as independent of one another and all of the PSMs are used for target-decoy analysis. In short, each of multiple monoisotopic precursor masses of a multiplexed MS/MS spectrum contribute one PSM to the target-decoy analysis. Target-decoy analysis was performed using the selected PSMs to obtain PSMs within FDR 1% (PSM-FDR).

RESULTS AND DISCUSSION

In LC-MS/MS data, MS spectra contain all MS information (experimental monoisotopic masses, intensities, and scan numbers or elution times) of the MS features of precursors on which MS/MS experiments are performed. Previously, PE-MMR was developed to determine correct precursor monoisotopic masses by referencing the five “candidate” masses (i.e., by considering the original mass, $-1/-2/-3$ Da subtracted masses, and +1 Da added mass) with UMC masses within 10 ppm and ± 5 MS scans (range scanning).²⁹ As PE-MMR utilizes UMC and UMC is a collection of monoisotopic masses for a peptide measured in multiple charge states along chromatographic elution (i.e., not only multiple charge states but also multiple scans), it was demonstrated that PE-MMR correctly identifies a peptide even when all of the MS data for the targeted *m/z* (i.e., of a certain charge state) failed to yield the correct monoisotopic mass owing to weak MS peak intensity; the precursor ions of different charge states having higher MS peak intensity allow for correct mass assignment and thereby lead to correct identification.⁴ PE-MMR, however, lacks the capability for both charge state determination of MS/MS spectra of unassigned charge state and deconvolution of cofragmented precursors for which it is necessary to assign multiple precursor masses to the multiplexed MS/MS spectra. As explained in the Experimental Section, we extend PE-

MMR using UMC information as confirmation criteria for both charge determinations for abundant MS/MS spectra of unassigned charge state and multiple precursor mass assignments for multiplexed MS/MS spectra.

Increased Sensitivity and Accuracy of Peptide Identification by *m*PE-MMR

Even in this study, in which we employed a narrow isolation window (± 0.8 Th) for fragmentation, up to 70% of the resultant MS/MS spectra were multiplexed MS/MS spectra, most of which failed to provide charge information. As shown in Table 1, about 70% of MS/MS spectra of LC-MS/MS experiments have no assigned charge states, irrespective of the use of sample fractionation (unfractionated or fractionated), making the calculation of monoisotopic masses difficult, if not impossible. One of the major causes of this problem is the presence of peaks of cofragmented precursors, making charge state determination difficult by simple calculation based on peak spacing between adjacent peaks (Figure 2A). In cases where “no charge” was determined, as shown in Figure 2B, a conventional method arbitrarily assigns both +2 and +3 charge states. This not only inevitably inflates the number of MS/MS data in an MGF file (i.e., 208,215 MS/MS data in MGF compared to 139,615 MS/MS scans, Table 1) for the subsequent database search, but can also result in false positive peptide identification when the true charge is neither +2 nor +3. As shown in Figure 2B, the conventional method assigned monoisotopic masses of 1241.5654 (+2) and 1862.3481 Da (+3), which in turn resulted in a decoy (+2) and a target hit (+3) with a poor score (cutoff at FDR 1%), respectively (Figure 2D).

*m*PE-MMR instead uses the m/z value of the picked precursor peak and considers seven different charge states, from +1 to +7. It then compares the resultant seven different monoisotopic masses with UMC masses around the MS/MS execution time. In this example, it found a match with a UMC mass (UMC #11594) when the charge state of +4 was considered (Figure 2C), while the masses of other charges resulted in no match within 10 ppm, generating one MS/MS data in *m*PE-MMR MGF with the matched UMC mass as the precursor mass. As explained in the Experimental Section, *m*PE-MMR continued to perform charge multiplexing, C13 mass correction, and matching isotope envelope removal subsequently for each of the remaining peaks within the isolation window and found matches with two additional UMC masses, resulting in a total of three different monoisotopic masses assigned to the corresponding MS/MS spectrum (Figure 2C). In this example, a total of 105 candidate masses (i.e., multiplexed matching of 35 masses (7 charges $\times$ 5 isotopic masses) for each of three isotopic envelopes) were compared with UMC masses within ± 5 MS scans (i.e., range scanning), resulting in three matched UMC masses assigned to this multiplexed MS/MS spectrum. *m*PE-MMR generated three MS/MS results using an identical MS/MS spectrum with three different precursor monoisotopic masses (three cloned MS/MS spectra). Despite the presence of fragments from cofragmentation, MS-GF+ successfully resulted in peptide identifications for all three precursors having +2, +3, and +4 charges within FDR 1%, as supported by the additively annotated fragments of the MS/MS spectra of the three peptides (Figure 2E).

Compared with the conventional method, the *m*PE-MMR method increased the numbers of PSMs and nonredundant peptides at FDR 1% (e.g., ca. 247% and 128%, respectively, for the

unfractionated data set; ca. 387% and 150%, respectively, for the one-fraction data set; and ca. 372% and 144%, respectively, for the combined 24-fraction data set) obtained from various LC-MS/MS data sets (Table 1 and supplementary Tables S1, S2, S3, S4, S5, and S6). It also increased the number of MS/MS scans having at least one peptide ID compared with the conventional method (i.e., >40% vs <20%). For example, the *m*PE-MMR method resulted in peptide identification within FDR 1% for more than half of the MS/MS spectra of the one-fraction data set, while only 16.1% of the MS/MS spectra resulted in peptide identification using the conventional method (Table 1). Figure 3A,B compares the target-decoy score distributions of PSMs obtained from the conventional method and the *m*PE-MMR method using the one-fraction data. The overall shapes and the cutoff values for FDR 1% for both distributions are similar; both effectively separate the decoy distributions from the target distributions. However, a significant increase in PSMs with a high score ($-\log(\text{SpecValue}) > 20$) was observed with the *m*PE-MMR method, suggesting that the *m*PE-MMR method resulted in identification of more PSMs with high confidence. When two PSMs at FDR 1% from the conventional and *m*PE-MMR methods were compared (Figure 3C), 38333 (74.3%) of a total of 51186 PSMs from *m*PE-MMR were *m*PE-MMR only, while only 384 PSMs (0.7%) were conventional method only. Furthermore, the experimental mass measurement accuracies (MMAs) of the *m*PE-MMR-only PSMs displayed a normal experimental MMA distribution (i.e., an offset of 0.93 ppm with a standard deviation of 1.15 ppm), while the MMAs of PSMs of the conventional-only displayed an essentially random distribution (i.e., they were equally distributed across the ± 10 ppm range), suggesting that many of the PSMs of the conventional-only were likely random peptide identifications.

The increased sensitivity of peptide identification by *m*PE-MMR can be attributed to two major improvements: charge assignments for charge-unassigned MS/MS spectra and multiple peptide identifications for multiplexed MS/MS spectra.

Increased Sensitivity of Peptide Identification by Accurate Charge State Assignments

As Figure 4 shows, the charge state distributions of precursor ions recorded in the conventional MGF (Figure 4A) and *m*PE-MMR MGF (Figure 4B) were using one-fraction data. In the conventional method, the charges of +2 and +3 comprised about 94% of MS/MS data in the corresponding MGF, while +4 or higher charges made up only about 5% (6496 MS/MS data). In contrast, the *m*PE-MMR method increased the charges of +4 or higher by about 13 $\times$ (85800 MS/MS data) after charge multiplexing. It is important to note that a single *m/z* peak can result in multiple UMC masses having different charge states by charge multiplexing, as redundancy was not removed when generating the *m*PE-MMR MGF file. The redundancy in the conventional method (i.e., +2 and +3 for unassigned charge peaks) was also not removed in the conventional MGF file. The redundancy, however, was removed after the database search and before the target-decoy analysis by using only one of the highest scores. As shown in Figure S2, *m*PE-MMR significantly increases the number of PSMs of high charge states at FDR 1% (from 1044 to 12840), which, in return, increases the number of nonredundant peptide identifications of high charge states (from 530 to 2230).

To maximize the benefit of charge multiplexing, it is critical to employ a database search engine that provides well-calibrated scores, that is, scores that are directly comparable across

different PSMs, even if the peptides or spectra are different. The choice of MS-GF+ is important here because it employs a dynamic programming-based, score calibration method. When poorly calibrated scores (e.g., SEQUEST Xcorr) are used, including spectra with high charge states, this decreases the number of identifications.³³

Increased Sensitivity of Peptide Identification by Deconvolution of Cofragmented Precursors

In addition to the significant increase in identification sensitivity for peptides with high charge states ($>+4$ charge), *m*PE-MMR also increased the number of PSMs with lower charge states ($+3$ charge). For example, the number of PSMs of charge states $+2$ and $+3$ using the *m*PE-MMR method and the conventional method were 38135 (of 8267 nonredundant peptides) and 12186 (of 6035 nonredundant peptides), respectively (Figure S2), and those of $+1$ charge were 221 (of 99 nonredundant peptides) and 7 (of 7 nonredundant peptides), respectively (Figure S2). The increase in the number of both PSMs and peptides for these lower charges is mainly due to additional precursor identification resulting from deconvolution of cofragmented precursors. For example, while the $+1$ charge state was excluded from the fragmentation, as indicated in the Experimental Section, some precursor ions of $+1$ charge can have their *m/z* value within the isolation window of other precursor ions of higher charges and undergo cofragmentation with the precursor ions of higher charges. *m*PE-MMR subsequently deconvoluted the precursor ions of $+1$ charge from other cofragmented precursors and resulted in PSMs (or peptide identification) of $+1$ charge.

As shown in Figure 5, only about 25% (18290) of MS/MS scans resulted in fragmentation of a single precursor ion after *m*PE-MMR analysis. *m*PE-MMR assigned more than two precursors for the rest of the MS/MS scans prior to the database search. Using the *m*PE-MMR MGF, MS-GF+ analysis followed by target-decoy analysis resulted in a total of 51186 PSMs at 1% FDR. A total of 48.4% (8856) of 18290 MS/MS scans with one precursor mass per scan resulted in PSMs within 1% FDR. As the number of precursor masses assigned to a single MS/MS scan increased, the portion of MS/MS scans resulting in at least one peptide identification increased (from 58.9% for two precursors per scan to 66.3% for five or more precursors per scan), resulting in more peptide information than was gained from the conventional method using the same LC-MS/MS scans.

While the overall percentage of MS/MS scans with at least one peptide identification increased significantly (from $>20\%$ with the conventional method to $>40\%$ with *m*PE-MMR), there are still significant MS/MS data in their corresponding MGF files (both conventional MGF and *m*PE-MMR MGF) left without peptide identification. The major cause for the lack of identification may be the incomplete use of protein variations (i.e., other PTMs, nonsynonymous mutations, and others) for database search, though there may be other unknown reasons for this phenomenon.

Effective Filtering of False Positive Peptide Identification by the “One Peptide for One UMC” Rule

One UMC corresponds to a peptide (or an ionized species) measured at multiple times in multiple charges (“one peptide for one UMC”). In other words, one UMC should not be linked to more than one peptide.

After obtaining PSMs at PSM-FDR 1%, we mapped the PSMs (both target and decoy PSMs) to UMCs. As discussed in the Experimental Section, every MS/MS data in the *m*PE-MMR MGF file were linked to UMCs and therefore, via the linked MS/MS information, PSMs of the MS/MS data could be linked back to UMCs after MG-GF+ search and target-decoy analysis. After mapping PSMs to UMCs, the PSMs with the highest scores within one UMC (when PSMs of two or more different peptides were mapped to one UMC) were selected for recalculation of the PSM-FDR. This recalculation resulted in 0.85% PSM-FDR, further decreasing false positive identification. This demonstrates that UMC information can be used as an effective filter for false positive identification.

The target-decoy distributions and the numbers of non-redundant peptides at 1% PepFDR obtained using *m*PE-MMR method with and without UMC filtering were compared in Figure S3A,B). Using the one-fraction data, *m*PE-MMR method resulted in 8289 (cutoff = 9.23) nonredundant peptides at 1% PepFDR. For UMC filtering, we mapped all PSMs (i.e., both target and decoy PSMs) after a database search (and before target-decoy analysis) to UMCs and collected only PSMs of a peptide with the highest score from one UMC and subsequently performed PepFDR analysis in the target-decoy setting. In the process of UMC filtering of PSMs, a total of 41454 peptides were removed from the target-decoy distribution (Figure S3C), but the difference in the final number of nonredundant peptides between the two analyses was only 21 peptides (8289 vs 8268 peptides) at 1% PepFDR. The removed target peptides after UMC filtering have very similar score distributions to those of the removed decoy peptides (Figure S3C). Therefore, the UMC filtering of PSMs based on the “one peptide for one UMC” rule effectively removed false positive PSMs and led to a slightly lower cutoff value for 1% PepFDR (from 9.23 to 9.18), resulting in an overall loss of only 21 peptides, despite filtering 41454 peptides. Figure 6 shows the annotated MS/MS spectrum for one of the “filtered-out” peptides, compared with that of the corresponding “filtered-in” peptide. The calculated monoisotopic masses of MANHPLLHR and AQLGPDESK are 1231.6729 and 1231.6650, respectively, which were matched to UMC #878 with a UMC mass of 1231.6649 within 10 ppm. Both PSMs were within PepFDR 1%, but only AQLGPDESK was selected, as it had the higher peptide score. Ambiguity associated with the two PSMs of different peptides with the same MS features was effectively alleviated with the use of the “one peptide for one UMC” rule.

Comparison of Results from Single-Round and Iterative Database Search after *m*PE-MMR

In the *m*PE-MMR process, MS/MS spectra are duplicated when it finds multiple precursors cofragmented and the duplicated MS/MS spectra are assigned with different precursor masses. When the resultant duplicated spectra are used for database search, some MS/MS peaks may be used multiple times in different peptide identifications, which can potentially result in false but high-scoring identifications.

In order to rule out this concern, an iterative database search was employed as described in the Experimental Section. The target-decoy score distributions of peptides using the results from the single-round database search and the iterative database search with UMC filtering, respectively, are near identical with very similar cutoff values for 1% PepFDR (i.e., 9.18 vs 9.17), which resulted in a similar number of nonredundant peptides (i.e., 8268 vs 8203; Figure S5). The peptide overlap between the two methods is 95.7% (Figure 7A), and the peptide scores of the two methods for the common peptides are also very similar (Figure 7B), suggesting that the fragments double counting in the single-round database search on *m*PE-MMR processed MS/MS data has negligible effect on the peptide identification results. As the total search time for the iterative method is nearly four times longer than that for single-round method (149 vs 563 min) with negligible increase in peptide identification, we propose the single-round database search using *m*PE-MMR $\oplus$ MS-GF+ as the method of choice.

Increased Sensitivity for Mutated Peptide Identification by *m*PE-MMR

As explained in the Experimental Section in the Supporting Information, a unified protein database, including a customized database and a reference database, was used for the protein database search to detect mutated peptides. Figure 8A compares the protein variation information obtained from the *m*PE-MMR method with that from the conventional method using the unified protein database. *m*PE-MMR increased the sensitivity of identification of mutated peptides, identifying 390 more mutated peptides. A total of 410 mutated peptides were exclusively detected using the *m*PE-MMR method. Of those, 402 peptides originated from germline SNVs, two originated from somatic SNVs, three resulted from insertions, one resulted from a deletion, and two were matched in the gene fusion region. Figure 8B shows an annotated MS/MS spectrum with the peptide sequence MQLQHLVEGEHI-TSDGLK, containing the V22I mutation. The conventional method failed to identify this peptide, as it assigned +2 and +3 charges for the corresponding MS/MS spectrum, while *m*PE-MMR correctly assigned a +4 charge. Figure 8C gives an example of the identification of a mutated peptide from multiplexed MS/MS spectrum. As shown in the inset, two precursor peptides were cofragmented and *m*PE-MMR successfully identified both DALSDLALHFLNK and IDGI-TIHQSLAIEYLEETRPTPR (containing the M42T mutation), while the conventional method only identified DALSDLA-LHFLNK.

CONCLUSIONS

We have developed a method for accurately assigning monoisotopic masses to their corresponding MS/MS spectra, including multiplexed MS/MS spectra. Utilizing the detailed information on MS data from LC-MS/MS data, *m*PE-MMR performs mass correction for C13 isotope peak picking, charge determination for MS/MS spectra of unassigned charge state, and finally deconvolution of cofragmented precursors. Compared to the conventional method, *m*PE-MMR resulted in increased sensitivity and accuracy of peptide identifications from the database search on high throughput LC-MS/MS data, which was mainly attributed to both reduced ambiguities in precursor charge determinations and multiple peptide identifications for multiplexed MS/MS spectra. Also, the use of UMC information for

filtering false positive peptide identification based on “one peptide for one UMC” rule was demonstrated.

The increased utilization of MS/MS scans for peptide identifications (i.e., ca. >40% of MS/MS scan resulted in at least one peptide identification and significant increase in the number of PSMs at FDR 1%) should increase the accuracy of label free quantitation based on spectral counts. Also *m*PE-MMR can be applicable for analyzing LC-MS/MS data from data independent acquisition, where a wide isolation window for fragmentation is used and thereby cofragmentation is highly abundant.

Even after *m*PE-MMR analysis, there are still many unidentified MS/MS data (i.e., “dark matter”), mainly because of the incomplete coverage of various protein variations in a database search. The unidentified MS/MS data may be used for further downstream analyses with a more complete understanding of protein modifications and different database search engines (i.e., multistage MS/MS data analysis). *m*PE-MMR may play an important role in such analysis, as it decreases the number of unidentified MS/MS scans, while providing correct charges and precursor masses for next-stage data analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Multiomics Research Program (NRF-2012M3A9B9036675), the Brain Research Program (NRF-2014M3C7A1046047), and the International Research and Development Program (NRF-2015K1A3A1A21000273) funded by the Korean Ministry of Science, ICT & Future Planning. Efforts at Pacific Northwest National Laboratory were supported by grants from the National Institute of General Medical Sciences (P41 GM103493) and the U.S. Department of Energy Office of Biological and Environmental Research Genome Sciences Program under the Panomics Program.

References

1. Eng JK, McCormack AL, Yates JR. J Am Soc Mass Spectrom. 1994; 5:976–989. [PubMed: 24226387]
2. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Electrophoresis. 1999; 20:3551–3567. [PubMed: 10612281]
3. Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD. Bioinformatics. 2008; 24:1021–1023. [PubMed: 18304935]
4. Jung HJ, Purvine SO, Kim H, Petyuk VA, Hyung SW, Monroe ME, Mun DG, Kim KC, Park JM, Kim SJ, Tolic N, Slys GW, Moore RJ, Zhao R, Adkins JN, Anderson GA, Lee H, Camp DG, Yu MH, Smith RD, Lee SW. Anal Chem. 2010; 82:8510–8518. [PubMed: 20863060]
5. Scherl A, Tsai YS, Shaffer SA, Goodlett DR. Proteomics. 2008; 8:2791–2797. [PubMed: 18655048]
6. Luethy R, Kessner DE, Katz JE, MacLean B, Grothe R, Kani K, Faca V, Pitteri S, Hanash S, Agus DB, Mallick P. J Proteome Res. 2008; 7:4031–4039. [PubMed: 18707148]
7. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. J Proteome Res. 2010; 9:4152–4160. [PubMed: 20578722]
8. Na S, Paek E, Lee C. Anal Chem. 2008; 80:1520–1528. [PubMed: 18247484]
9. Sadygov RG, Hao Z, Huhmer AFR. Anal Chem. 2008; 80:376–386. [PubMed: 18081262]
10. Carvalho PC, Cociorva D, Wong CCL, Carvalho MDD, Barbosa VC, Yates JR. Anal Chem. 2009; 81:1996–2003. [PubMed: 19203245]

11. Carvalho PC, Xu T, Han XM, Cociorva D, Barbosa VC, Yates JR. *Bioinformatics*. 2009; 25:2734–2736. [PubMed: 19684088]
12. Hoopmann MR, Finney GL, MacCoss MJ. *Anal Chem*. 2007; 79:5620–5632. [PubMed: 17580982]
13. Hsieh EJ, Hoopmann MR, MacLean B, MacCoss MJ. *J Proteome Res*. 2010; 9:1138–1143. [PubMed: 19938873]
14. Yuan ZF, Liu C, Wang HP, Sun RX, Fu Y, Zhang JF, Wang LH, Chi H, Li Y, Xiu LY, Wang WP, He SM. *Proteomics*. 2012; 12:226–235. [PubMed: 22106041]
15. Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. *Proteomics*. 2005; 5:4096–4106. [PubMed: 16196091]
16. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. *Nat Methods*. 2007; 4:923–925. [PubMed: 17952086]
17. Cox J, Mann M. *Nat Biotechnol*. 2008; 26:1367–1372. [PubMed: 19029910]
18. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. *J Proteome Res*. 2011; 10:1794–1805. [PubMed: 21254760]
19. Wang J, Bourne PE, Bandeira N. *Mol Cell Proteomics*. 2011; 10:M111.010017.
20. Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. *J Am Soc Mass Spectrom*. 2013; 24:301–304. [PubMed: 23292976]
21. Kryuchkov F, Verano-Braga T, Hansen TA, Sprenger RR, Kjeldsen F. *J Proteome Res*. 2013; 12:3362–3371. [PubMed: 23725413]
22. Tyanova S, Mann M, Cox J. *Methods Mol Biol*. 2014; 1188:351–364. [PubMed: 25059623]
23. Wang J, Bourne PE, Bandeira N. *Mol Cell Proteomics*. 2014; 13:3688–3697. [PubMed: 25225354]
24. Zhang B, Pirmoradian M, Chernobrovkin A, Zubarev RA. *Mol Cell Proteomics*. 2014; 13:3211–3223. [PubMed: 25100859]
25. Kim S, Pevzner PA. *Nat Commun*. 2014; 5:5277. [PubMed: 25358478]
26. Na S, Bandeira N, Paek E. *Mol Cell Proteomics*. 2012; 11:M111.010199.
27. Li H, Hwang KB, Mun DG, Kim H, Lee H, Lee SW, Paek E. *J Proteome Res*. 2014; 13:3488–3497. [PubMed: 24918111]
28. Park K, Yoon JY, Lee S, Paek E, Park H, Jung HJ, Lee SW. *Anal Chem*. 2008; 80:7294–7303. [PubMed: 18754627]
29. Shin B, Jung HJ, Hyung SW, Kim H, Lee D, Lee C, Yu MH, Lee SW. *Mol Cell Proteomics*. 2008; 7:1124–1134. [PubMed: 18303012]
30. Zimmer JSD, Monroe ME, Qian WJ, Smith RD. *Mass Spectrom Rev*. 2006; 25:450–482. [PubMed: 16429408]
31. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. *Nat Biotechnol*. 2012; 30:918–920. [PubMed: 23051804]
32. Keller A, Eng J, Zhang N, Li XJ, Aebersold R. *Mol Syst Biol*. 2005; 1:E1.
33. Howbert JJ, Noble WS. *Mol Cell Proteomics*. 2014; 13:2467–2479. [PubMed: 24895379]

Figure 1.

(A) Overall schematic of the *mPE-MMR* process. MS data were used for UMC generation. MS/MS data were used for original MGF generation. The MS/MS information on the original MGF were further processed by charge state multiplexing, C13 mass correction and deconvolution of cofragmented precursors of the dotted box. (B) Graphical workflow of the dotted box in (A). An MS/MS event targeting 621.79 Th led to cofragmentation of three different precursor ions within the isolation window. Among the 105 candidate masses that *mPE-MMR* calculated using three peaks (out of 10 peaks) in the isolation window, three different monoisotopic masses were found matches with UMC masses and the UMC masses were assigned to the corresponding MS/MS spectrum. More detailed information on the *mPE-MMR* procedures can be found in the text.

Figure 2.

(A) Expanded MS spectrum showing peaks within the isolation window (red arrow, targeted precursor ion). (B) Expanded MS spectrum with the two corresponding isotopic envelopes (+2 and +3 as assigned by the conventional method) indicated by purple and dark green filled circles. (C) Expanded MS spectrum with the three *mPE-MMR*-assigned isotopic envelopes indicated by dark brown, dark blue, and green filled circles. (D) Annotated MS/MS spectra using the precursor masses assigned by the conventional method. (E) Annotated MS/MS spectrum using the three precursor masses determined by *mPE-MMR*.

Figure 3.

(A) Target-decoy score distribution of PSMs using the conventional method on one-fraction data set. (B) Target-decoy score distribution of PSMs using the *m*PE-MMR method on the same data set. (C) Venn diagram of PSMs at 1% FDR in the conventional and *m*PE-MMR methods. (D) Distributions of mass measurement accuracies for the conventional method only (blue) and the *m*PE-MMR method only PSMs (red).

Figure 4.

(A) Charge state distribution of all MS/MS data in the conventional MGF using the one-fraction data set. (B) Charge state distribution of all MS/MS data in *mPE-MMR* MGF using the one-fraction data set. *mPE-MMR* increases not only the sensitivity of identification for the peptides of high charge states (+4), but also those of low change states (Figure S2).

Figure 5.

Bar graph showing the number of precursor masses per MS/ MS scan in the *m*PE-MMR MGF file. About 75% of MS/MS scans were assigned with two or more precursors. The stacked columns show the proportion of MS/MS scans that identified the indicated number of peptides.

Figure 6.
Example of a “filtered-out” peptide (A) by UMC filtering and its corresponding UMC peptide (B).

Figure 7.

(A) Venn diagram of peptides at 1% FDR in the single-round and iterative database search methods. (B) Distributions of peptide score differences between the single-round and iterative database searches for the common peptides. The MS-GF+ search time for the single-round and the iterative methods were 149 and 563 min, respectively.

Figure 8.

(A) Venn diagram of mutated peptides identified by the conventional and *mPE-MMR* methods. (B) Annotated MS/MS spectrum of MQLQHLVEGEHITS DGLK (+4) with the V22I mutation, which was only identified by the *mPE-MMR* method. (C) Annotated MS/MS spectrum of cofragmentation of DALSDLALHFLNK and IDGITIHQSLAIIIEYLEETRPTPR. The inset shows the expanded MS spectrum, illustrating two precursor ions within the isolation window.

Table 1Comparisons of Analysis Statistics between Conventional and *m*PE-MMR Methods

	unfractionated		fractionated (1 of 24 fractions)		fractionated (24 fractions combined)	
	conventional	<i>m</i> PE-MMR	conventional	<i>m</i> PE-MMR	conventional	<i>m</i> PE-MMR
No. MS/MS scans	139615	139615	82060	82060	2013338	2013338
No. MS/MS scans of unassigned charge	104508	104508	54981	54981	1383255	1383255
No. MS/MS data in MGF	208215	328085	130456	266853	3188899	6301694
No. PSMs at PSM-FDR 1%	27057	66766	13237	51186	279171	1037935
No. NR peptides at PSM-FDR 1%	17120	21987	6267	9368	108925	156577
No. NR peptides at PepFDR 1%	16671	20287	5928	8289	101600	135938
No. MSMS scan with ID	27057 (19%)	56264 (40%)	13237 (16%)	43469 (53%)	279171 (14%)	900334 (45%)
<i>m</i> PE-MMR process time ^a (min)		45		43		38 ^b
MS-GF+ search time ^a (min)	114	216	72	149	65.2 ^b	139 ^b

^a Intel Xenon E5-2630 v2 @ 2.60 GHz 16 GB.^b Average time per fraction data.