Correction of errors in tandem mass spectrum extraction enhances phosphopeptide identification

Hao, Piliang; Ren, Yan; Sze, Siu Kwan; Tam, James P.

2013

Hao, P., Ren, Y., Tam, J. P., & Sze, S. K. (2013). Correction of Errors in Tandem Mass Spectrum Extraction Enhances Phosphopeptide Identification. Journal of Proteome Research, 12(12), 5548-5557.

https://hdl.handle.net/10356/103597

https://doi.org/10.1021/pr4004486

Correction of Errors in Tandem Mass Spectrum Extraction Enhances Phosphopeptide Identification

Journal:	*Journal of Proteome Research*
Manuscript ID:	pr-2013-004486.R3
Manuscript Type:	Article
Date Submitted by the Author:	20-Oct-2013
Complete List of Authors:	Hao, Piliang; Nanyang Technological University, ; Nanyang Technological University, Singapore Centre on Environmental Life Sciences Engineering Ren, Yan; Nanyang Technological University, Tam, James; Nanyang Technological University, School of Biological Sciences Sze, Siu Kwan; Nanyang Technological University, Chemical Biology and Biotechnology; Nanyang Technological University, Singapore Centre on Environmental Life Sciences Engineering

Correction of Errors in Tandem Mass Spectrum Extraction Enhances Phosphopeptide Identification

Piliang Hao[‡*1,2], Yan Ren[‡1], James P. Tam[1] and Siu Kwan Sze[*1,2]

[1] School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

[2] Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Running title: Correction of Errors in Phosphopeptide Spectrum Extraction

[‡] These authors contributed equally to this work.

***Corresponding author**

Dr. Siu Kwan SZE

Tel: (+65)6514-1006, Fax: (+65)6791-3856

Email: sksze@ntu.edu.sg

Piliang Hao

Tel: (+65)6316-2852, Fax: (+65)6791-3856

Email: haop0001@e.ntu.edu.sg

Correction of Errors in Phosphopeptide Spectra Extraction

Abbreviations: ERLIC, electrostatic repulsion-hydrophilic interaction chromatography; WAX, weak anion exchange; LC-MS/MS, liquid chromatography coupled to tandem mass spectrometry; PTM, post translational modifications; MS, mass spectrometry; LTQ, linear quadrupole ion trap; FA, formic acid; FDR, false discovery rate; MS^2, MS/MS; MS^3, MS/MS/MS; MGF, mascot generic format; HCD, higher-energy collisional dissociation;

Correction of Errors in Phosphopeptide Spectra Extraction

SUMMARY

The tandem mass spectrum extraction of phosphopeptides is more difficult and error-prone than that of unmodified peptides due to their lower abundance, lower ionization efficiency, the co-fragmentation with other high-abundance peptides, and the use of MS^3 on MS^2 fragments with neutral losses. However, there are still no established methods to evaluate its correctness. Here we propose to identify and correct these errors via the combinatorial use of multiple spectrum extraction tools. We evaluated 5 free and 2 commercial extraction tools using Mascot and phosphoproteomics raw data from LTQ FT Ultra, in which RawXtract 1.9.9.2 identified the highest number of unique phosphopeptides (peptide expectation value<0.05). Surprisingly, ProteoWizzard (v. 3.0.3476) extracted wrong precursor mass for most MS^3 spectra. Comparison of the top three free extraction tools showed that only 54% of the identified spectra were identified consistently from all three tools, indicating that some errors might happen during spectrum extraction. Manual check of 258 spectra not identified from all three tools revealed 405 errors of spectrum extraction with 7.4% in selecting wrong precursor charge, 50.6% in selecting wrong precursor mass and 42.1% in exporting MS/MS fragments. We then corrected the errors by selecting the best extracted MGF file for each spectrum among the three tools for another database search. With the errors corrected, it results in the 22.4% and 12.2% increase of spectrum matches and unique peptide identification, respectively, compared with the best single method. Correction of errors in spectrum extraction improves both the sensitivity and confidence of phosphopeptide identification. Data analysis on non-phosphopeptide spectra indicates that this strategy applies to unmodified peptides as well. The identification of errors in spectrum extraction will promote the improvement of spectrum extraction tools in future.

Correction of Errors in Phosphopeptide Spectra Extraction

KEYWORDS: phosphorylation, LC-MS/MS, MS^3, mass spectrum extraction, MGF

INTRODUCTION

Liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) has been widely used in the large-scale identification and quantification of phosphopeptides due to its ease of automation and sensitivity.[1-4] Generally, raw data from LC-MS/MS are encoded in a vendor-specific and closed format, and the vendors provide spectrum extraction tools to convert them into peak lists for database searches.[5] For example, extract_msn and Proteome Discoverer (Thermo Fisher, Waltham, MA) can be used in extracting peak lists from raw files of mass spectrometry from the same company. Generally, spectrum extraction includes several steps: 1) determination of the precursor ion charge; 2) extracting the precursor ion mass; 3) extracting the fragment ion mass. To improve peptide identification, some spectrum extraction tools can also include additional steps, such as charge state deconvolution of fragments, deisotoping of fragment ions and general noise reduction. It has been believed that the instrument vendors have taken necessary precautions to make the spectrum extraction valid and accurate, but it remains a black box since there are still no established ways to validate it. However, peptide identification will be adversely affected if the spectrum extraction results in some errors or the loss of some information.

In addition to the tools from vendors, some experts in the area of proteomics also develop their own tools to facilitate and improve spectrum extraction, e.g. DeconMSn[6], Raw2MSM[7], MaxQuant[8], ProteoWizzard[5] and RawXtract[9]. Most of the tools are claimed to have some advantages over those provided by vendors. For example, Raw2MSM improves the precursor

mass accuracy by intensity-weighting the measured masses over their LC elution profile.[7] ProteoWizard can process raw files from any mass spectrometry vendors and convert them into several different formats, e.g. MGF, mzML, mzXML, and so on. MaxQuant 1.3 integrates spectrum extraction with database searches, modification site localization and false discovery rate (FDR) calculation, greatly facilitating data analysis of proteomics data. Nine different spectrum extraction tools have been comprehensively compared in a recent paper, and their pros and cons are revealed. However, it neither evaluates the possible errors in spectrum extraction nor suggests a way to correct it.[10]

Phosphorylation is a reversible modification involved in the regulation of protein functions and many biological processes including cell division, signal transduction and enzymatic activity.[11-13] Spectrum extraction of phosphopeptides uses the same tools with those for unmodified peptides. However, the intensity of phosphopeptides is generally much lower than unmodified peptides with the same sequence due to their lower abundance, lower ionization efficiency, and the processing of weak signals is more difficult and error-prone because of the interference from noise and other co-eluted high-abundance peptides. In addition, MS^3 is usually used in dissociating the MS^2 fragments with the neutral loss of phosphate moiety since MS^2 spectra of phosphopeptides often do not contain sufficient information for identification, in which the labile phosphate group results in the dominance of MS^2 fragments with the neutral loss.[14, 15] The extraction of MS^3 spectra is more error-prone since it is slightly different from that of MS^2 spectra and needs more considerations.[16] The data analysis of MS^3 spectra is also more difficult than that of MS^2 spectra. For example, if spectrum extraction tools use the MS^2 fragments with neutral losses as the precursors for MS^3 spectra, the database search of MS^2 and MS^3 spectra has

to be done separately. When LC-MS/MS is done using hybrid mass spectrometry, such as LTQ FT Ultra and LTQ Orbitrap, the precursor mass tolerance of MS^3 spectra should be set to be 0.8 Dalton, and variable modification of -18 Da on S and T should be used due to the neutral loss of phosphate group, which is completely different from that of MS^2 spectra using the precursor mass tolerance of 10 ppm and the variable modification of phosphorylation on S, T and Y.[17] However, MS^3 spectra can be processed in the same way with MS^2 spectra when spectrum extraction tools use the MS precursor acquired in the full-scan spectrum as their precursors,[16] which results in the increase of phosphopeptide identification. [18]

Because of the above-mentioned difficulties in spectrum extraction and data analysis of phosphopeptides involving the use of MS^3, it is necessary to conduct a comprehensive comparison of the spectrum extraction tools, evaluate the possible errors in spectrum extraction and correct them if possible in order to improve the number and confidence of phosphopeptide identification. In this study, we evaluated seven of the above-mentioned spectrum extraction tools, i.e. DeconMSn, extract_msn 4.0, ProteoWizzard (v. 3.0.3476), MaxQuant 1.3.0.5, Raw2MSM (v. 1_10_2007_06_14), Proteome Discoverer 1.3 and RawXtract 1.9.9.2, using Mascot as search algorithm and 12 phosphoproteomics raw files from LTQ FT Ultra and proposed to solve these problems via the combinatorial use of multiple spectrum extraction tools.

MATERIALS AND METHODS

Sample Preparation

Correction of Errors in Phosphopeptide Spectra Extraction

MDA-MB-231 cells were obtained from American Type Culture Collection (ATCC, Manassas, VA) and cultured as recommended. Tryptic peptides of MDA-MB-231 cells were prepared as previously described.[19]

Phosphopeptide Enrichment using Electrostatic Repulsion-Hydrophilic Interaction Chromatography (ERLIC) and Titanium Dioxide

Peptides from 3 mg protein were fractionated using a PolyWAX LP weak anion-exchange column (4.6 × 200 mm, 5 μm, 300 Å, PolyLC, Columbia, MD) on a Shimadzu Prominence UFLC system. Seventeen fractions were collected with a 42 min gradient of 100% mobile phase A (70% ACN/1% FA) for 5 min, 0%–7% mobile phase B (10% ACN/2% FA) for 5 min, 7%–45% B for 18 min, 45%–100% B for 10 min, followed by 4 min at 100% B at a flow rate of 0.7 mL/min. Fraction 1-2, 3-4 and 5-6 were combined and enriched for phosphopeptides using titanium dioxide as described.[20] Fraction 7 to 17 were dried in vacuum and redissolved in 0.1% FA for LC-MS/MS analysis. Fractions 7-8 and 16-17 were combined, respectively.

LC-MS/MS

LC-MS/MS was done as previously described with minor modifications.[21] Briefly, peptides were separated and analyzed on a Dionex Ultimate 3000 RSLCnano system coupled to a LTQ FT Ultra (Thermo Electron, Bremen, Germany) using a 70 min gradient. Peptides were analyzed on LTQ FT Ultra with an ADVANCE™ CaptiveSpray™ Source (Michrom BioResources) at an electrospray potential of 1.5 kV. A gas flow of 2, ion transfer tube temperature of 180°C and collision gas pressure of 0.85 mTorr were used. A full MS scan (350-1600 m/z range) was acquired in the FT-ICR cell at a resolution of 100,000 and a maximum ion accumulation time of

1000 msec. The default AGC setting was used (full MS target at 3.0e+04, MS^n1e+04) in linear ion trap. The 10 most intense ions above a 500 counts threshold were selected for fragmentation in CAD, which was performed concurrently with a maximum ion accumulation time of 200 msec. Dynamic exclusion was activated for the process, with a repeat count of 1 and exclusion duration of 60 s. Single charged ion is excluded from MS/MS. For CAD, normalized collision energy was set to 35%, activation Q was set to 0.25, and activation time was 30 ms. A MS^3 scan was followed after each MS^2 scan when a neutral loss at 97.97 Da was detected. Isolation width of 2.50 and 5.00 are used in MS^2 and MS^3 scan, respectively.

Spectrum Extraction

The MS^2, MS^3 and MS^{All} spectra (a combination of MS^2 and MS^3 spectra) for 12 raw files from LTQ-FT were extracted separately using each of the 7 spectrum extraction tools (DeconMSn, extract_msn 4.0, ProteoWizzard (v. 3.0.3476), MaxQuant 1.3.0.5, Raw2MSM, Proteome Discoverer 1.3 and RawXtract 1.9.9.2) and submitted to Mascot database search in order to evaluate their performance in extracting MS^2 and MS^3 spectra.. Default settings were used for all extraction tools unless otherwise specified. For ProteoWizzard (v. 3.0.3476), Raw2MSM and Proteome Discoverer 1.3, MS^2 spectra were extracted for each raw file and combined into a single mascot generic format (MGF) file, and MS^3 and MS^{All} spectra were processed in the same way. For Raw2MSM, top 10 high-intensity MS^2 or MS^3 fragments are used every 100 Daltons; for Proteome Discoverer 1.3, precursor selection is set as "Use MS1 Precursor". For RawXtract 1.9.9.2, .ms2 and .ms3 files were generated and converted into MGF files using in-house made Perl scripts, and the MS^1 precursors of MS^3 spectra were extracted from .ms2 files based on their scan numbers since the MS^1 precursors are only provided with 2 decimals in .ms3 files, which is insufficient for the high mass accuracy database searches. The MS^2 and MS^3 MGF files for each

raw file were combined into a single MGF file, respectively. A combination of all MS^2 and MS^3 MGF files results in the MS^{All} MGF file for RawXtract 1.9.9.2. For DeconMSn and extract_msn 4.0, MS^{All} spectra are extracted for the 12 raw files and combined into a single MGF file, respectively. Since these two tools do not provide the option of processing MS^2 and MS^3 spectra separately, we extract MS^2 and MS^3 spectra for them using in-house made Perl scripts based on the scan number of MS^2 and MS^3 spectra provided by Proteome Discoverer 1.3. The MS^2 and MS^3 MGF files are combined into a single MGF file for each tool, respectively. For MaxQuant 1.3.0.5, the function of "Partial processing" was used with step 1 to 5. Several .apl files were generated for all of the 12 raw files and converted into MGF files using an in-house made Perl script. They were then processed in the same way with DeconMSn and extract_msn 4.0.

Database Searches and Data Analysis

The UniProt human protein database (release 2012_05, 87187 sequences) and its reversed complement were combined and used for database searches. The database search was performed using an in-house Mascot server (version 2.3.02, Matrix Science, Boston, MA, USA) with $\#^{13}C$ of 2 and MS/MS tolerance of 0.8 Da. Two missed cleavage sites of trypsin were allowed. Carbamidomethylation (C) was set as a fixed modification. For MS^3 data from DeconMSn and extract_msn 4.0, the MS tolerance of 0.8 Da was used since these two tools used the MS^2 fragments with neutral losses as the precursors for MS^3 spectra and the precursors were acquired with much lower mass accuracy using LTQ instead of FT, and oxidation (M), phospho_NL (S and T) and deamidation (N and Q) were set as variable modifications. Phospho_NL (S and T) is set as the loss of H_2O on S or T due to the neutral loss of phosphoric acid from modified S or T residues.[16] For all other database searches, the MS tolerance of 10 ppm was used, and oxidation

(M), phosphorylation (S, T and Y) and deamidation (N and Q) were set as variable modifications. For high confidence peptide identification, only peptides with an E-value of less than 0.05 were used for statistical calculation. The FDR of peptide identification was calculated based on the assigned spectra (FDR = 2.0 × decoy_hits/total_hits).[22] Since nonenzymatic deamidation occurs easily during proteomic sample preparation, phosphopeptides with difference only in the modification of deamidation are regarded as same unique phosphopeptides.

Determination of the Correctness of Monoisotopic Peak Selection for Precursor Ions

As we use 10 ppm as the MS tolerance during database searches using Mascot, it is regarded as the selection of the correct monoisotopic peak for precursor ions if the peptide is matched at an error of less than 10 ppm. Mascot also provides the match to the ^{13}C or $^{13}C_2$ peak of the precursor ions, but the error is about 1 or 2 Dalton. Thus, the correctness of monoisotopic peak selection can be determined based on the error between the measured and calculated masses of the precursor ions in database search results.

Identification of Errors in Spectrum Extraction Using MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2 by Manual Check

In order to confirm whether some errors really happen during the spectrum extraction, we manually checked phosphopeptide matches from one of the raw files, i.e. STNC06R, in the aspect of precursor charge, precursor mass and the exportation of MS/MS fragments by comparing the database search results with the raw file. If a phosphopeptide is identified consistently with the use of all three spectrum extraction tools, we do not check its correctness manually. If a phosphopeptide can only be matched with the use of one or two of the spectrum

extraction tools, we first check whether its determination of the precursor charge and precursor mass is correct in the extracted MGF file of the spectrum that cannot be matched. If both the precursor charge and precursor mass are correct, it is due to the difference in exporting MS/MS fragments.

Selection of the Best Extracted MGF files for Each Spectrum

First, we got the best scored peptide for each spectrum by comparing the database search results from MGF files extracted with MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2 using an in-house made Perl script. And then, we extracted the MGF files for all of the best scored peptides from MGF files extracted with these three tools using another in-house made Perl script and combined them into a single MGF file. It was then searched again using Mascot to generate the final protein and peptide list.

Correction of Errors in Phosphopeptide Spectra Extraction

RESULTS AND DISCUSSION

Comparison of the 7 Spectrum Extraction Tools in the Number of Extracted Spectra, Phosphopeptide and Unique Phosphopeptide Identifications

Table 1 summarizes the version, availability, URL and references of the 7 spectrum extraction tools evaluated in this study. These tools were compared in extracting MS^2, MS^3 and MS^{All} spectra from the 12 phosphoproteomics raw files from LTQ FT Ultra, respectively, in order to evaluate their performance in extracting MS^2 and MS^3 spectra and their overall performance in identifying phosphopeptides. As shown in Figure 1A and Supplemental Table 1, DeconMSn extracted 145,537 spectra from the 12 raw files, which was the highest among all 7 tools and 1.76 times of the average of extract_msn 4.0, ProteoWizzard (v. 3.0.3476), Proteome Discoverer 1.3 and RawXtract 1.9.9.2, i.e. 82,658±960. It is due to that two identical entries with difference only in the title are produced for each spectrum with 2 or 3 charges in DeconMSn, while the latter 4 tools generated only 1 entry for most spectra. MaxQuant 1.3.0.5 extracted the second highest number of spectra among the 7 tools, i.e. 138,824. It generated 58,747 MS^3 entries, which is 3.04 times of the average of extract_msn 4.0, ProteoWizzard (v. 3.0.3476), Proteome Discoverer 1.3 and RawXtract 1.9.9.2, i.e. 19,352±655. MaxQuant 1.3.0.5 generated 3 entries with difference in charges and/or peptide masses for each MS^3 spectra. Raw2MSM extracted the third highest number of spectra since an unresolved charge state or an uncertain precursor selection triggers multiple copies of the same spectrum with differences in precursor charge state and/or mass. This is consistent with the report from Mancoso et al.[10]

For high-confidence phosphopeptide identification, only identifications with peptide expectation value<0.05 are used for statistical analysis, and the FDR of phosphopeptide identification is

calculated to be between 0.1% and 0.4% for all 7 tools. As shown in Figure 1B and Supplemental Table 1, the highest number of phosphopeptides was identified from DeconMSn among the 7 tools due to the generation of two identical entries for each spectrum with 2 or 3 charges, but it identified the lowest number of unique phosphopeptides (Figure 1C and Supplemental Table 1). The number of phosphopeptide identification from MS^2 spectra is comparable for other 6 tools, indicating that all these tools perform well in extracting MS^2 spectra (Figure 1B). It is consistent with the report from Mancoso et al.[10] However, it is worthy of noticing that only 304 phosphopeptides were identified from the MS^3 spectra of ProteoWizzard (v. 3.0.3476), which is about 20% of that of other 6 tools, i.e. 1511±206. Manual check of the raw files and the extracted MGF files revealed that ProteoWizzard used the precursor of the first triggered MS^2 spectra after each full-scan spectrum as the precursors for all MS^3 spectra following the full-scan spectrum. It means that most of the precursor selections are wrong for MS^3 spectra in ProteoWizzard. Thus, it was only about 20% of other tools in the aspect of phosphopeptide and unique phosphopeptide identification from MS^3 spectra. This highlights the importance of evaluating spectrum extraction tools before applying them to biological studies. The 12 phosphoproteomics raw data files from LTQ FT Ultra and their database search results of MS^2, MS^3 and MS^{All} spectra can be downloaded from PeptideAtlas using the dataset identifier of PASS00263 (http://www.peptideatlas.org/PASS/PASS00263).

In the aspect of overall unique phosphopeptide identification, RawXtract 1.9.9.2 identified the highest among the 7 tools, which is 8% higher than the average of extract_msn 4.0, MaxQuant 1.3.0.5, Raw2MSM and Proteome Discoverer 1.3. For unique phosphopeptide identification from MS^2 spectra, the 6 tools except DeconMSn are comparable. However, for unique

phosphopeptide identification from MS^3 spectra, ProteoWizzard is the poorest due to the above-mentioned errors in precursor selection, and DeconMSn and extract_msn are also significantly poorer than other 4 tools possibly due to that they use the MS^2 fragments with neutral losses as the precursors for MS^3 spectra and a MS tolerance of 0.8 Da instead of 10 ppm is used during database searches, which may reduce the sensitivity of phosphopeptide identification.[17] For MaxQuant 1.3.0.5, Raw2MSM, Proteome Discoverer 1.3 and RawXtract 1.9.9.2, the use of MS^3 spectra leads to the increase of over 15% in the identification of unique phosphopeptides in comparison to only using MS^2 spectra (Figure 1C), indicating that MS^3 is more efficient than MS^2 in dissociating the MS^2 fragments with the neutral loss of phosphate moiety.[16] These four tools provide a convenient and efficient way to process raw data involving the use of MS^3 for phosphopeptide identification. For all seven evaluated tools, over 99.5% of the identified MS^3 spectra are phosphopeptides (Supplemental Table 1), indicating the power of MS^3 in dissociating the MS^2 fragments with the neutral loss of phosphate moiety. Mancuso et al. reported that the spectrum extraction of phosphopeptides had no significant differences with that of unmodified peptides, i.e. the number of unique phosphopeptide identifications is comparable among different extraction tools.[10] This is true for MS^2 analysis of phosphopeptides, but the difference among different tools becomes evident when MS^3 analysis is included due to the different processing of MS^3 spectra in these tools.

Mass Accuracy of Peptide Spectrum Matches and the Determination of Precursor Monoisotopic Peak

Mass accuracy means the relative difference between the measured and calculated masses of the precursor ions in database search results. As shown in Table 2, the average mass accuracy of all

identified peptides (E-value<0.05) for ProteoWizzard is 2.93 ppm, while that of other 6 tools are between -1.52 and -3.47 ppm. It indicates that the determination of precursor mass in ProteoWizzard is different from that of other tools. Our LTQ FT Ultra has a systematic error of about -3 ppm at the time of running the samples used in this study. Thus, MaxQuant 1.3.0.5, Raw2MSM, extract_msn 4.0 and Proteome Discoverer 1.3 produce the optimal mass accuracy for peptide precursors. The standard deviation of mass accuracy for MaxQuant 1.3.0.5 and Raw2MSM is significantly lower than that of other tools, possibly due to their specialized design for mass spectrometric data with high accuracy.[7, 8]

The assignment of monoisotopic peaks for precursor ions is crucial in peptide identification.[23] Thus, we evaluated the accuracy of the 7 tools in determining the monoisotopic peak for precursor ions. Since the MS^3 spectra extracted with DeconMSn and extract_msn 4.0 use MS^2 fragments with neutral losses as the precursors and they were acquired in LTQ with much lower mass accuracy, these data were excluded for determining the monoisotopic peak. As shown in Table 2, MaxQuant 1.3.0.5 extracts monoisotopic peaks for 91.4% of the peptide identifications, which is the best among all 7 tools. Raw2MSM ranks the second best, i.e. 90.0%. In comparison, DeconMSn and RawXtract 1.9.9.2 extract monoisotopic peaks for only about 46% of the peptide identifications, but it does not significantly affect phosphopeptide identification since Mascot can still identify the peptide even if ^{13}C or $^{13}C_2$ peaks are used. Please be noted that #^{13}C must be set to 2 in Mascot in order to correctly identify ^{13}C or $^{13}C_2$ peaks. Actually, RawXtract is designed to faithfully keep the precursor peak selected by the instrument for fragmentation, but not try to find the monoisotopic peak with some corrections.[9, 24] However, the ^{13}C or $^{13}C_2$ peaks are often misidentified as deamidated peptides for low-resolution MS/MS data.[25-27] Thus, MaxQuant

1.3.0.5 and Raw2MSM are recommended in studying protein deamidation in order to reduce the false positive identification of deamidated peptides. Alternatively, it can also be achieved by performing database searching with a wide mass tolerance window and then filtering with stringent delta mass.[28] This is very useful for low-resolution MS/MS data from hybrid mass spectrometers, such as LTQ FT Ultra and LTQ Orbitrap. Now, the introduction of high-resolution MS/MS using higher-energy collisional dissociation (HCD) in some newly released mass spectrometers, e.g. Q Exactive, may be good enough to differentiate the ^{13}C or $^{13}C_2$ peak from the corresponding deamidated peptides merely based on the high-resolution MS/MS.

The Overlap of Phosphopeptide and Spectrum identification among MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2

Our data analysis indicates that Proteome Discoverer 1.3 and RawXtract 1.9.9.2 have an overlap of 91.7% in unique phosphopeptide identification as spectrum extraction tools. Thus, the combinatorial use of these two tools does not lead to the much increase of unique phosphopeptide identification. We then studied the combinatorial use of MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2, which are the top 3 tools in unique phosphopeptide identification except for Proteome Discoverer 1.3 (Figure 1C). As shown in Figure 2, 58.4% of the unique phosphopeptides and 54% of the phosphopeptide spectra are identified consistently from MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2. As 41.6% of the unique phosphopeptides and 46% of the phosphopeptide spectra are identified only from one or two of the extraction tools, we assume that the spectrum extraction tools may make some mistakes during the processing of raw files, e.g. errors in determining precursor charge and precursor mass and loss of information during exporting MS/MS fragments.

Identification of Errors in Spectrum Extraction Using MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2 by Manual Check

For 511 phosphopeptide spectra identified from the raw file of STNC06R, 50% of them are identified consistently from all three tools; 25% of them are identified from two of the tools; 25% of them are identified only from one tool (Figure 3A). It indicates that some errors may happen during the spectrum extraction. We then manually checked the extraction of 258 spectra not identified from all three tools in the aspect of the determination of precursor charge, precursor mass and the exportation of MS/MS fragments. As shown in Figure 3B, 405 errors in spectrum extraction are revealed from the three tools with 7.4% in determining wrong precursor charge, 50.6% in determining wrong precursor mass and 42.1% in exporting MS/MS fragments. The details about the spectrum extraction errors can be found in Supplemental Table 2. RawXtract 1.9.9.2 performs the best in all of the three above-mentioned aspects among the three tested tools, which explains why it identifies the highest number of unique phosphopeptides among the 7 evaluated tools. In order to check whether these errors are specific to phosphopeptides, we also manually checked 50 non-phosphopeptide spectra from the raw file of STNC06R that were not identified consistently from all three tools of MaxQuant, Raw2MSM and RawXtract, and 72 errors are revealed with 9.7% in selecting wrong precursor charge, 63.9% in selecting wrong precursor mass and 26.4% in exporting MS/MS fragments. It indicates that these errors can happen on unmodified peptides as well. The details about the spectrum extraction errors can be found in Supplemental Table 3.

Correction of Errors in Phosphopeptide Spectra Extraction

Figure 3C illustrates an example in which MaxQuant 1.3.0.5 and RawXtract 1.9.9.2 select the wrong precursor mass but Raw2MSM selects the correct precursor mass. This situation happens on 8.5% (22/258) of the manually checked spectra. It is a typical situation of the co-fragmentation of several peptide precursor ions in complex samples at the isolation width of 2.50 in LTQ. Although all of the precursor ions indicated by arrows in Figure 3C are fragmented simultaneously in MS/MS, it prevents the identification of the phosphopeptide when the precursor ions with lower intensity, such 785.252 and 785.269, are extracted in MaxQuant 1.3.0.5 and RawXtract 1.9.9.2. The explanation is that Mascot only uses some of the MS/MS fragments with relatively high intensity in database searches and most of them are possibly derived from precursor ions with the highest intensity. MaxQuant 1.3.0.5 and RawXtract 1.9.9.2 determine the mass of the precursor targeted for MS/MS fragmentation, but not necessarily the precursor with the highest intensity within the isolation window. However, Raw2MSM improves the precursor mass accuracy by intensity-weighting the measured masses over their LC elution profile[7] so that it selects the precursor mass with the highest intensity, e.g. 785.306, which results in the confident identification of the phosphopeptide at a Mascot score of 68.2. This example shows that the fragment ions from the precursor with the highest intensity should be excluded in order that the co-fragmented precursors with lower intensity can be identified,[29, 30] and an algorithm that selects the precursor with the highest intensity within the isolation window should be used for chimera mass spectra. Figure 3D shows that a phosphopeptide is identified at a Mascot score of 52.0 (E-value=0.011) in RawXtract 1.9.9.2, but identified at a score of 32.6/30.3 (E-value=0.97/1.6) in MaxQuant 1.3.0.5/Raw2MSM due to the selection of different precursor ion mass. The same situation happens on 12.0% (31/258) of the manually checked spectra. It indicates that RawXtract 1.9.9.2 enhances phosphopeptide identification via selecting

the correct precursor ion mass when the monoisotopic peak of precursor ions is overlapped with isotopic peaks of other co-fragmented precursors. If a wrong precursor mass is determined for a peptide during the course of spectrum extraction, it can still be identified by using a broader mass tolerance during database searches, e.g. 1-2 Da. However, it will sacrifice the advantage of high-resolution MS. More importantly, it will cause some problems when Percolator is used in determining FDR since mass accuracy of peptide spectrum matches is also considered in Percolator.[31] For example, most correct matches are within the mass accuracy of the instrument, i.e. 5-10 ppm.

As shown in Figure 3B, 42, 43 and 85 spectra are not identified in RawXtract 1.9.9.2, MaxQuant 1.3.0.5 and Raw2MSM, respectively, due to the problems in exporting MS/MS fragments. RawXtract 1.9.9.2 exports the mass of MS/MS fragments directly, but MaxQuant 1.3.0.5 and Raw2MSM deisotope MS/MS fragments before exportation. In addition, Raw2MSM deconvolutes MS/MS fragments before exportation, and only top 10 high-intensity MS^2 or MS^3 fragments are used for database searches every 100 Daltons. The deisotoping and deconvolution of MS/MS fragments improve the Mascot score of mass spectra when they are done properly, and it is the reason why some spectra are identified while using MaxQuant 1.3.0.5 and Raw2MSM, but not identified while using RawXtract 1.9.9.2. However, some spectra are only identified while using RawXtract 1.9.9.2, indicating that information loss happens during the course of deisotoping and deconvolution. This is an obvious problem when low-resolution MS/MS spectra from hybrid mass spectrometry, such as LTQ FT Ultra and LTQ Orbitrap, are deisotoped and deconvoluted. It is reported that for high resolution MS/MS, the deisotoping of

fragment ions is favorable for MASCOT scoring, and charge state deconvolution is particularly useful for peptides with 3 or more charges.[10]

Correction of Errors in Spectrum Extraction of Phosphopeptides via the Combinatorial Use of Multiple Spectrum Extraction Tools

According to our manual check of 258 spectra not identified consistently from MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2, these tools make 81 to 180 mistakes in the spectrum extraction (Figure 3B), indicating that no single evaluated tool can handle the spectrum extraction of raw files from complex samples satisfactorily due to the complex situations mentioned in Figure 3C and 3D and other possible situations. It is of high possibility that a phosphopeptide fails to be identified in one spectrum extraction tool due to the errors in spectrum extraction, but it may be confidently identified while using another extraction tool due to its own advantage in selecting correct precursor charge, precursor mass and/or exporting MS/MS fragments. Thus, the combinatorial use of different spectrum extraction tools may be used in correcting errors in spectrum extraction and thus increase the confidence and sensitivity of phosphopeptide identification. It is easy to directly combine the database search results from several spectrum extraction tools together, but it will result in the redundant identification of the same spectrum, and it is also difficult to evaluate the confidence of the increased phosphopeptide identifications. We then proposed a new strategy to achieve it (Figure 4). First, the raw files were extracted using MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2, respectively, and combined into a single MGF file for each tool. Then, the 3 MGF files were submitted to Mascot database search, and the best peptide match for each spectrum is selected among the three tools from the database search results. At last, we extracted the best MGF file for each spectrum and

combine them into a single file, and submitted it to Mascot database search again in order to get the final phosphopeptide list. As shown in Figure 4, the combinatorial use of the three spectrum extraction tools leads to the increase of 22.4% and 12.2% in spectrum matches and unique phosphopeptide identification, respectively, compared with the best single method, i.e. RawXtract 1.9.9.2. Compared with the best commercial tool, i.e. Proteome Discoverer 1.3, the increase of spectrum matches and unique phosphopeptide identification are 34.2% and 17.6%, respectively. The increased sensitivity of phosphopeptide identification may be helpful in detecting low-abundance phosphopeptides with important functions. In addition to the improved sensitivity of phosphopeptide identification, the combinatorial use of multiple spectrum extraction tools also improves the confidence of phosphopeptide identification by increasing the peptide scores of many spectra with errors in spectrum extraction corrected. Before error correction, some of the spectra are identified with a relatively low score, and some of them are even assigned with a wrong peptide sequence. The combinatorial use of 3 spectrum extraction tools is tested in this study due to the limitation of computation capacity, but the combinatorial use of more tools may be possible in the future with the quick development of computation systems.

It is worthy of noticing that our strategy is based on the assumption that each MS^2 or MS^3 spectrum is derived from one precursor ion. Obviously, it is not true for data from complex samples.[32] Houel et al. reported that as high as 50% of the spectra from a typical LTQ-Orbitrap profiling of complex samples can be mixed spectra from at least two different precursors.[33] However, as discussed above, most spectrum extraction tools, such as extract_msn 4.0, ProteoWizzard (v. 3.0.3476), Proteome Discoverer 1.3 and RawXtract 1.9.9.2, extract only one

entry for most spectra. Although DeconMSn, MaxQuant 1.3.0.5 and Raw2MSM extract several entries for spectra with 2 or 3 charges, MS^3 spectra and a precursor with unresolved charge state, respectively, their aim is not to extract several correct entries for the spectra, but to provide one correct entry option. In addition, Mascot also does not support the identification of several different peptides from a mixed spectrum. Thus, it is assumed that each MS^2 or MS^3 spectrum is derived from one precursor ion in our proposed strategy. Certainly, we can easily upgrade our strategy to select two or more best extracted MGF files for each mixed spectrum once the spectrum extraction tools and database search software are to support the extraction and identification of mixed spectra.

CONCLUSIONS

In this study, 7 spectrum extraction tools were evaluated in spectrum extraction and phosphopeptide identification using phosphoproteomics data involving the use of MS^3, of which RawXtract 1.9.9.2 identified the highest number of unique phosphopeptides. Manual check of the spectra reveals that even RawXtract 1.9.9.2 makes 81 mistakes in extracting 258 spectra not identified consistently from three of the evaluated tools. The identification of errors in spectrum extraction facilitates the improvement of spectrum extraction tools in future. We then propose to correct the errors via the combinatorial use of multiple extraction tools, which results in the increase of 22.4% and 12.2% in spectrum matches and unique phosphopeptide identification, respectively, compared with RawXtract 1.9.9.2. Our proposed strategy currently bases on the assumption of each spectrum being derived from one precursor, but it can be upgraded to support the identification of mixed spectra once spectrum extraction tools and database search software can support the processing of mixed spectra. Since errors happen easily in spectrum extraction,

our proposed strategy can be applied to any peptide centric analysis, e.g. the analysis of peptides with any post translational modifications (PTMs), in order to improve the sensitivity and confidence of PTM characterization. The identification of unmodified peptides for protein identification can benefit from this strategy as well. With the quick development of computation systems, the combinatorial use of over 3 spectrum extraction tools may be feasible in the future.

ACKNOWLEDGEMENT

This work is in part supported by the Singapore National Research Foundation (NRF2011 NRF-CRP 001-109) and the Singapore Ministry of Health's National Medical Research Council (NMRC/CBRG/0004/2012). Piliang Hao is supported by research scholarship from Singapore Centre on Environmental Life Sciences Engineering.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org. Supplemental Table 1: Number of extracted spectra, phosphopeptide identifications and spectra identifications using the 7 tools; Supplemental Table 2: Statistics for errors in extraction of 258 spectra not identified from all three tools; Supplemental Table 3: Statistics for errors in extraction of 50 non-phosphopeptide spectra not identified from all three tools. The in-house made Perl scripts can be requested via emails.

REFERENCES

1. Villen, J.; Beausoleil, S. A.; Gerber, S. A.; Gygi, S. P., Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* **2007,** 104, (5), 1488-93.
2. Pan, C.; Gnad, F.; Olsen, J. V.; Mann, M., Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* **2008,** 8, (21), 4534-46.
3. Cantin, G. T.; Yi, W.; Lu, B.; Park, S. K.; Xu, T.; Lee, J. D.; Yates, J. R., 3rd, Combining protein-based IMAC, peptide-based IMAC, and MudPIT for efficient phosphoproteomic analysis. *J Proteome Res* **2008,** 7, (3), 1346-51.
4. Pinkse, M. W.; Mohammed, S.; Gouw, J. W.; van Breukelen, B.; Vos, H. R.; Heck, A. J., Highly robust, automated, and sensitive online TiO2-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster. *J Proteome Res* **2008,** 7, (2), 687-97.
5. Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P., ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008,** 24, (21), 2534-6.
6. Mayampurath, A. M.; Jaitly, N.; Purvine, S. O.; Monroe, M. E.; Auberry, K. J.; Adkins, J. N.; Smith, R. D., DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics* **2008,** 24, (7), 1021-3.
7. Olsen, J. V.; de Godoy, L. M.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M., Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **2005,** 4, (12), 2010-21.
8. Cox, J.; Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **2008,** 26, (12), 1367-72.
9. McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., 3rd, MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* **2004,** 18, (18), 2162-8.
10. Mancuso, F.; Bunkenborg, J.; Wierer, M.; Molina, H., Data extraction from proteomics raw data: an evaluation of nine tandem MS tools using a large Orbitrap data set. *J Proteomics* **2012,** 75, (17), 5293-303.
11. Hunter, T., Signaling--2000 and beyond. *Cell* **2000,** 100, (1), 113-27.
12. Pawson, T.; Nash, P., Assembly of cell regulatory systems through protein interaction domains. *Science* **2003,** 300, (5618), 445-52.
13. Gnad, F.; de Godoy, L. M.; Cox, J.; Neuhauser, N.; Ren, S.; Olsen, J. V.; Mann, M., High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* **2009,** 9, (20), 4642-52.
14. Tholey, A.; Reed, J.; Lehmann, W. D., Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J Mass Spectrom* **1999,** 34, (2), 117-23.
15. Olsen, J. V.; Mann, M., Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* **2004,** 101, (37), 13417-22.
16. Ulintz, P. J.; Yocum, A. K.; Bodenmiller, B.; Aebersold, R.; Andrews, P. C.; Nesvizhskii, A. I., Comparison of MS(2)-only, MSA, and MS(2)/MS(3) methodologies for phosphopeptide identification. *J Proteome Res* **2009,** 8, (2), 887-99.

17. Xu, H.; Wang, L.; Sallans, L.; Freitas, M. A., A hierarchical MS2/MS3 database search algorithm for automated analysis of phosphopeptide tandem mass spectra. *Proteomics* **2009,** 9, (7), 1763-70.
18. Timm, W.; Ozlu, N.; Steen, J. J.; Steen, H., Effect of high-accuracy precursor masses on phosphopeptide identification from MS3 spectra. *Anal Chem* **2010,** 82, (10), 3977-80.
19. Hao, P.; Guo, T.; Sze, S. K., Simultaneous analysis of proteome, phospho- and glycoproteome of rat kidney tissue with electrostatic repulsion hydrophilic interaction chromatography. *PLoS One* **2011,** 6, (2), e16884.
20. Thingholm, T. E.; Jorgensen, T. J.; Jensen, O. N.; Larsen, M. R., Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nat Protoc* **2006,** 1, (4), 1929-35.
21. Gan, C. S.; Guo, T.; Zhang, H.; Lim, S. K.; Sze, S. K., A comparative study of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) versus SCX-IMAC-based methods for phosphopeptide isolation/enrichment. *J Proteome Res* **2008,** 7, (11), 4869-77.
22. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
23. Yuan, Z. F.; Liu, C.; Wang, H. P.; Sun, R. X.; Fu, Y.; Zhang, J. F.; Wang, L. H.; Chi, H.; Li, Y.; Xiu, L. Y.; Wang, W. P.; He, S. M., pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics* **2012,** 12, (2), 226-35.
24. Xu, T.; Wong, C. C.; Kashina, A.; Yates, J. R., 3rd, Identification of N-terminally arginylated proteins and peptides by mass spectrometry. *Nat Protoc* **2009,** 4, (3), 325-32.
25. Robinson, N. E.; Lampi, K. J.; McIver, R. T.; Williams, R. H.; Muster, W. C.; Kruppa, G.; Robinson, A. B., Quantitative measurement of deamidation in lens betaB2-crystallin and peptides by direct electrospray injection and fragmentation in a Fourier transform mass spectrometer. *Mol Vis* **2005,** 11, 1211-9.
26. Segu, Z. M.; Hussein, A.; Novotny, M. V.; Mechref, Y., Assigning N-glycosylation sites of glycoproteins using LC/MSMS in conjunction with endo-M/exoglycosidase mixture. *J Proteome Res* **2010,** 9, (7), 3598-607.
27. Hao, P.; Ren, Y.; Alpert, A. J.; Sze, S. K., Detection, evaluation and minimization of nonenzymatic deamidation in proteomic sample preparation. *Mol Cell Proteomics* **2011,** 10, (10), O111 009381.
28. Hsieh, E. J.; Hoopmann, M. R.; MacLean, B.; MacCoss, M. J., Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J Proteome Res* **2010,** 9, (2), 1138-43.
29. Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R., ProbIDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005,** 5, (16), 4096-106.
30. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M., Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **2011,** 10, (4), 1794-805.
31. Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J., Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **2007,** 4, (11), 923-5.
32. Alves, G.; Ogurtsov, A. Y.; Kwok, S.; Wu, W. W.; Wang, G.; Shen, R. F.; Yu, Y. K., Detection of co-eluted peptides using database search methods. *Biol Direct* **2008,** 3, 27.

33. Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M., Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J Proteome Res* **2010,** 9, (8), 4152-60.

FIGURE LEGENDS:

Figure 1. Number of extracted spectra (A), identified phosphopeptides (B) and unique phosphopeptides (C) for 12 phosphoproteomics raw files with the use of the 7 evaluated spectrum extraction tools. For high confidence phosphopeptide identification, only phosphopeptides with an E-value of less than 0.05 were used for statistical calculation.

Figure 2. The overlap of unique phosphopeptide identification (A) and spectrum identification (B) for 12 phosphoproteomics raw files among the use of MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2. The low overlap of unique phosphopeptide identification and spectrum identification indicates that some errors may happen during the spectrum extraction using these tools.

Figure 3. Identification of errors in spectrum extraction using MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2 by manual check. (A) The overlap of 511 spectrum identification of phosphopeptides from the raw file of STNC06R among MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2; (B) Errors in extraction of 258 spectra not identified consistently from all three tools using MaxQuant 1.3.0.5, Raw2MSM and RawXtract 1.9.9.2 in determining precursor charge, precursor mass and exporting MS/MS fragments revealed by manual check; (C) A typical example in which Raw2MSM selects the precursor with the highest intensity from several co-fragmented precursors but MaxQuant 1.3.0.5 and RawXtract 1.9.9.2 fail to do so; (D)

Correction of Errors in Phosphopeptide Spectra Extraction

A typical example in which RawXtract 1.9.9.2 selects the correct precursor ion but MaxQuant 1.3.0.5 and Raw2MSM fail when the monoisotopic peak of the precursor ion is overlapped with isotopic peaks of other co-fragmented precursors.

Figure 4. The strategy for correction of errors in spectrum extraction of phosphopeptides via the combinatorial use of multiple spectrum extraction tools and its effect on the increase of unique phosphopeptide identification and spectrum identification.

Correction of Errors in Phosphopeptide Spectra Extraction

Correction of Errors in Phosphopeptide Spectra Extraction

TABLES:

Table 1. Summary of version numbers, availability, URL and references for the 7 spectrum extraction tools evaluated in this study

Tool	Availability	Description	URL	References
DeconMSn (v 2.2.2.2)	Free	DeconMSn creates spectrum files for tandem mass spectrometry data.	http://omics.pnl.gov/software/DeconMSn.php	6
Extract_msn 4.0 (imbedded in Bioworks 3.3)	Vendor	Windows console (DOS) utility for converting a raw file into a set of DTA format peak lists.	http://www.scientific-computing.com/products/product_details.php?product_id=63	
ProteoWizzard (v. 3.0.3476)	Free	The ProteoWizard Library and Tools are a set of modular and extensible open-source, cross-platform tools and software libraries that facilitate proteomics data analysis.	http://proteowizard.sourceforge.net/	5
MaxQuant 1.3.0.5	Free	MaxQuant is a quantitative proteomics software package designed for analyzing large mass-spectrometric data sets. It is specifically aimed at high-resolution MS data.	http://www.maxquant.org/	8
Raw2MSM (v. 1_10_2007_06_14)	Free	Raw2MSM creates MGF peak list files from Xcalibur raw files, and works best with high accuracy LC-MS/MS data, from an Orbitrap or FT instrument.	http://www.biochem.mpg.de/mann/publications/2006/0510_01/0510_01.html	7
Proteome Discoverer 1.3	Vendor	Proteome Discoverer software is a flexible, expandable platform for the analysis of qualitative and quantitative proteomics data.	http://www.thermoscientific.com/ecomm/servlet/productsdetail?productId=11961811&groupType=PRODUCT&searchType=0&storeId=11152	
RawXtract 1.9.9.2	Free	RawXtract can process raw files from Thermo Scientific into several different formats, e.g. MS1, MS2, MS3, MSzm, mzXML and DTA.	http://fields.scripps.edu/downloads.php	9

Table 2. Summary of mass accuracy of peptide spectrum matches (E-value<0.05) and the correct determination of precursor monoisotopic peak from the peak lists generated by the 7 spectrum extraction tools

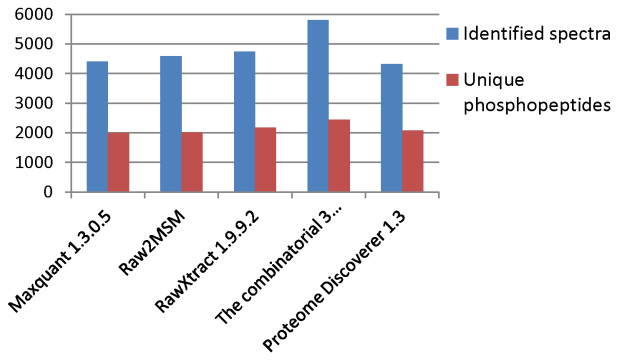
Names of spectrum extraction tools	Average precursor mass error [ppm]	Standard deviation [ppm]	Total peptide matches	Correct monoisotopic peak assignment	Percentage of correct monoisotopic peak assignment
DeconMSn	-1.61	3.34	11412	5241	45.9%
Extract_MSn 4.0	-2.47	2.63	8357	5977	71.5%
ProteoWizzard (v. 3.0.3476)	2.93	2.83	8326	5945	71.4%
MaxQuant 1.3.0.5	-3.47	1.95	10106	9241	91.4%
Raw2MSM (v. 1_10_2007_06_14)	-2.45	2.00	9768	8788	90.0%
Proteome Discoverer 1.3	-2.64	3.00	9748	8214	84.3%
RawXtract 1.9.9.2	-1.52	3.32	10393	4817	46.3%

Note: For DeconMSn and Extract_MSn 4.0, only MS^2 peptide matches are used for statistical calculation due to large error of MS^3 precursor acquired in LTQ.

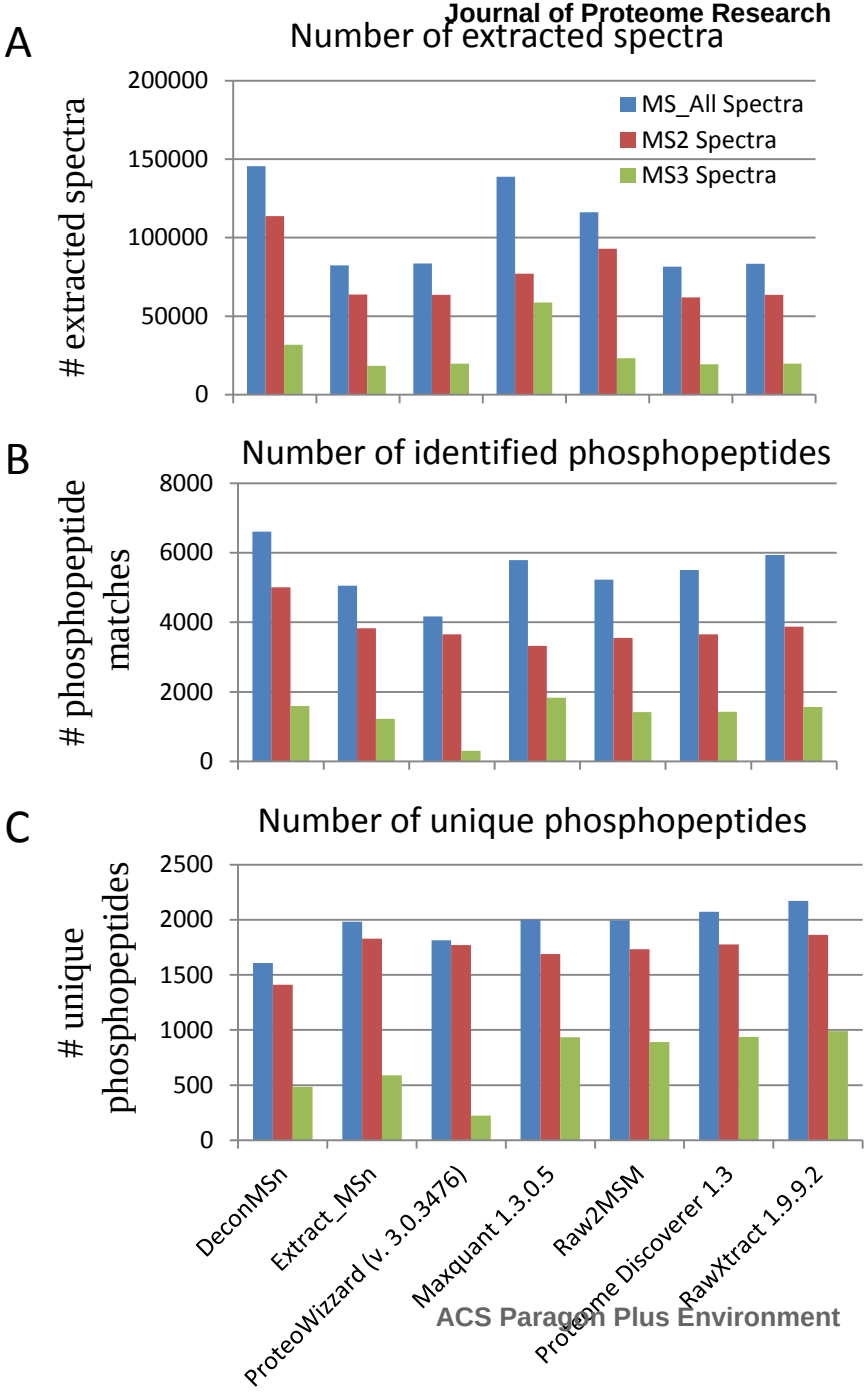
Table of Contents (TOC) Synopsis

Correction of Errors in Tandem Mass Spectrum Extraction Enhances Phosphopeptide Identification

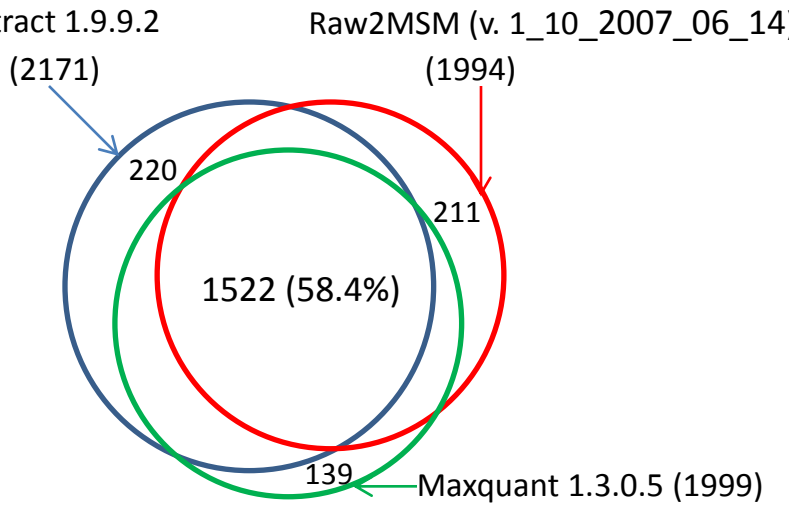
Piliang Hao, Yan Ren, and Siu Kwan Sze



Correction of Errors in Phosphopeptide Spectrum Extraction


A
Number of extracted spectra
extracted spectra
200000
150000
100000
50000
0
MS_All Spectra
MS2 Spectra
MS3 Spectra
B
Number of identified phosphopeptides
phosphopeptide matches
8000
6000
4000
2000
0
C
Number of unique phosphopeptides
unique phosphopeptides
2500
2000
1500
1000
500
0
DeconMSn
Extract_MSn
ProteoWizzard (v. 3.0.3476)
Maxquant 1.3.0.5
Raw2MSM
Proteome Discoverer 1.3
RawXtract 1.9.9.2
ACS Paragon Plus Environment

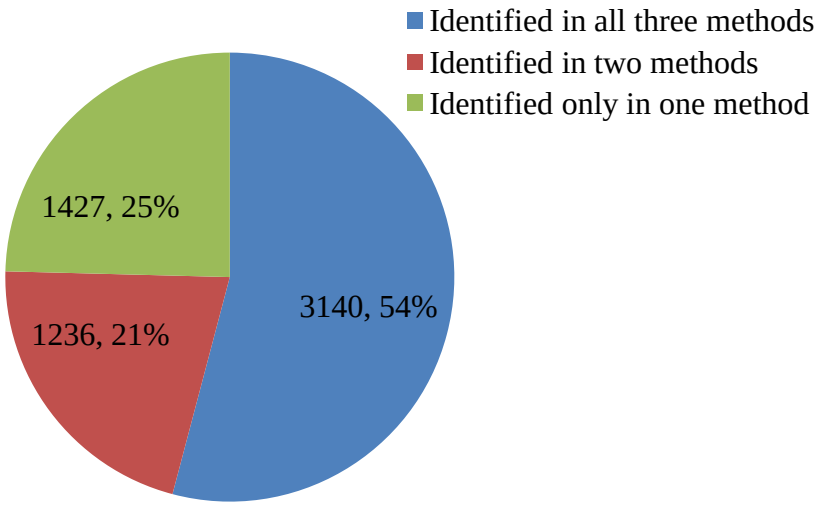
A



In total: 2606 unique phosphopeptides

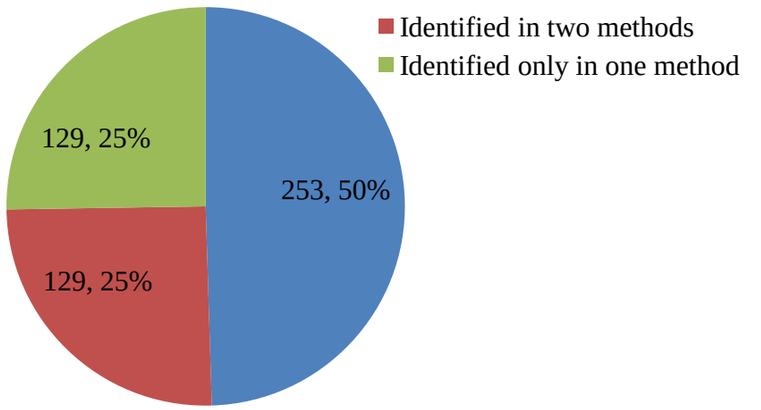
Overlap of unique phosphopeptide identification among the three extraction methods

B

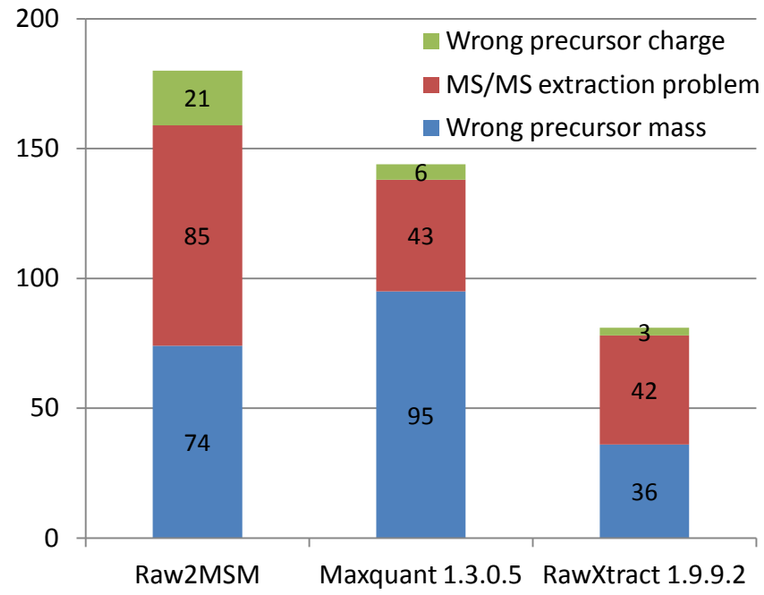


Overlap of spectrum identification of phosphopeptides (Expect value <0.05) among three extraction methods

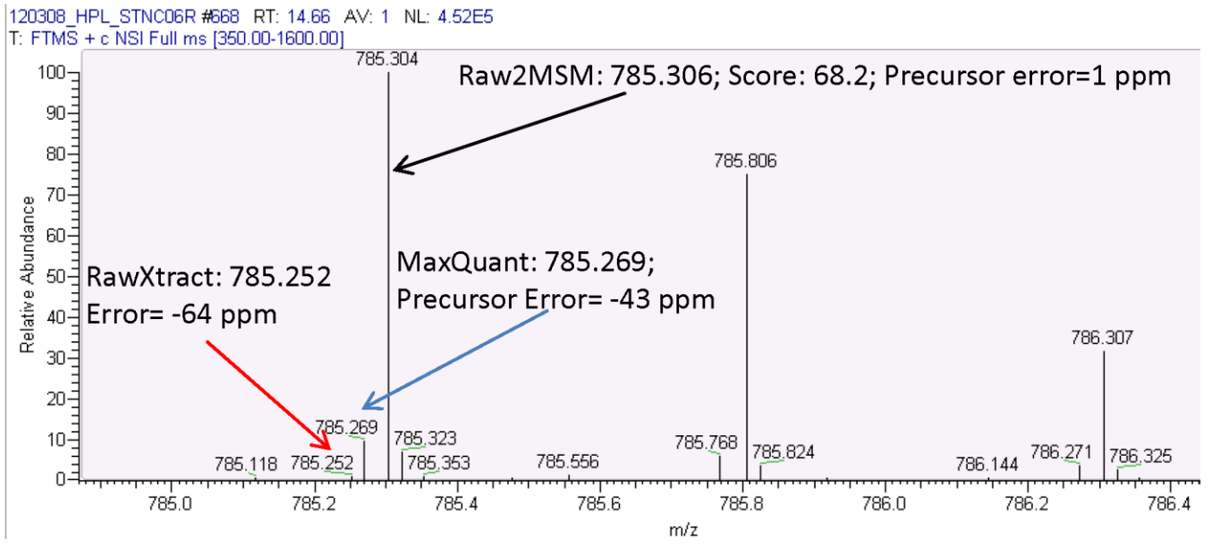
A


Identified in all three methods
Identified in two methods
Identified only in one method
253, 50%
129, 25%
129, 25%

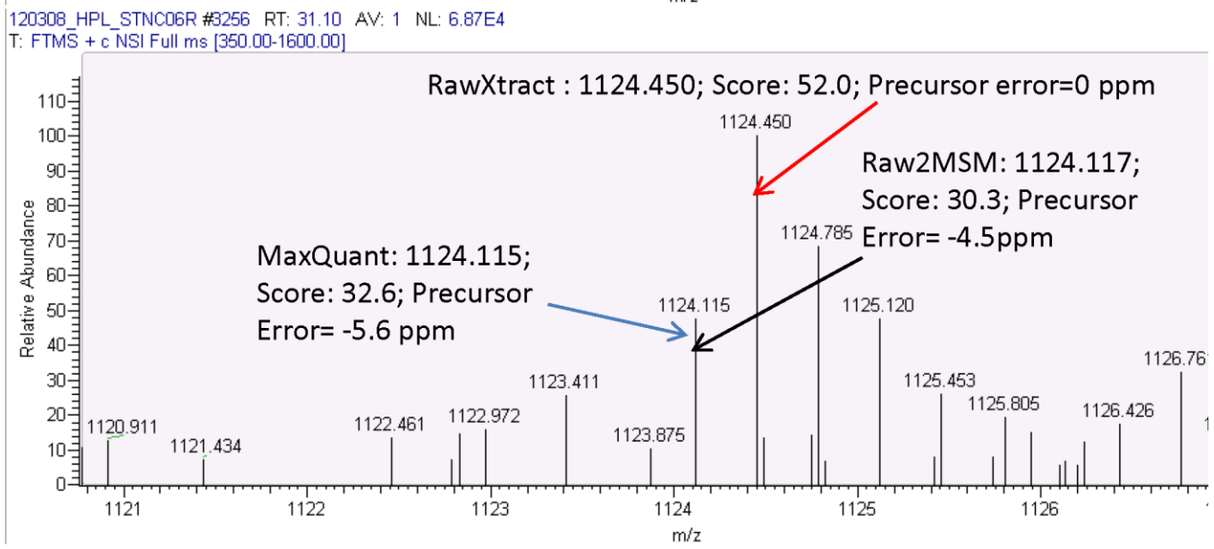
B


Wrong precursor charge
MS/MS extraction problem
Wrong precursor mass
200
150
100
50
0
21
85
74
6
43
95
3
42
36
Raw2MSM
Maxquant 1.3.0.5
RawXtract 1.9.9.2

C


120308_HPL_STNC06R #668 RT: 14.66 AV: 1 NL: 4.52E5
T: FTMS + c NSI Full ms [350.00-1600.00]
Raw2MSM: 785.306; Score: 68.2; Precursor error=1 ppm
RawXtract: 785.252
Error= -64 ppm
MaxQuant: 785.269;
Precursor Error= -43 ppm
Relative Abundance
m/z
785.304
785.806
786.307
785.269
785.323
785.118
785.252
785.353
785.556
785.768
785.824
786.144
786.271
786.325

D


120308_HPL_STNC06R #3256 RT: 31.10 AV: 1 NL: 6.87E4
T: FTMS + c NSI Full ms [350.00-1600.00]
RawXtract : 1124.450; Score: 52.0; Precursor error=0 ppm
Raw2MSM: 1124.117;
Score: 30.3; Precursor
Error= -4.5ppm
MaxQuant: 1124.115;
Score: 32.6; Precursor
Error= -5.6 ppm
Relative Abundance
m/z
1124.450
1124.785
1124.115
1125.120
1123.411
1125.453
1125.805
1126.426
1120.911
1121.434
1122.461
1122.972
1123.875

12 raw files from LTQ-FTUltra

Spectra extraction using Raw_Xtract 1.9.9.2, Maxquant 1.3.0.5 and Raw2MSM (v. 1_10), respectively

3 combined MGF files from each extraction method

Database search using Mascot

3 database search results from each extraction method

Selecting the best scored peptide for each spectrum from the three extraction methods based on peptide E-value

The best scored spectrum extraction for each spectrum

Copy the best extracted MGF file for each spectrum and combine them into a single file

A MGF file comprising of the best extracted MGF files for each spectrum

Database search using Mascot again

Final phosphopeptide list