MCP Papers in Press. Published on October 31, 2013 as Manuscript M113.032813

iPRG 2012: A Study on Detecting Modified Peptides in a Complex Mixture

Robert J. Chalkley[1*], Nuno Bandeira[2], Matthew C. Chambers[3], Karl R. Clauser[4], John S. Cottrell[5], Eric W. Deutsch[6], Eugene A. Kapp[7], Henry H. N. Lam[8], W. Hayes McDonald[9], Thomas A. Neubert[10] and Rui-Xiang Sun[11]

[1] University of California, San Francisco, CA; [2] University of California San Diego, La Jolla, CA; [3] Vanderbilt University Medical Center, Nashville, TN; [4] The Broad Institute of MIT and Harvard, Cambridge, MA; [5] Matrix Science Ltd., London, UK; [6] Institute for Systems Biology, Seattle, WA; [7] Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia; [8]University of Science and Technology, Hong Kong, China; [9] Vanderbilt University School of Medicine, Nashville, TN; [10] New York University School of Medicine, New York, NY; [11] Chinese Academy of Sciences, Beijing, China

*Corresponding Author:

600 16th Street,

Genentech Hall Room N474A,

San Francisco, CA 94158-2517

Tel: 415 476 5189

Fax: 415 502 1655

E-mail: chalkley@cgl.ucsf.edu

Running Title: Detecting Modified Peptides in a Complex Mixture

Abbreviations

ABRF – Association of Biomolecular Resource Facilities

CID – Collision Induced Dissociation

FDR – False Discovery Rate

FLR – False Localization Rate

iPRG – Proteome Informatics Research Group of ABRF

sPRG – Proteome Standards Research Group of ABRF

PTM – Post-translational modification

Summary

The proteome informatics research group of the Association of Biomolecular Resource Facilities conducted a study to assess the community's ability to detect and characterize peptides bearing a range of biologically occurring post-translational modifications when present in a complex peptide background. A dataset derived from a mixture of synthetic peptides with biologically occurring modifications combined with a yeast whole cell lysate as background was distributed to a large group of researchers and their results were collectively analyzed. The results from the twenty-four participants, who represented a broad spectrum of experience levels with this type of data analysis, produced several important observations. First, there is significantly more variability in the ability to assess whether a results is significant than there is to determine the correct answer. Second, labile post-translational modifications, particularly tyrosine sulfation, present a challenge for most researchers. Finally, for modification site localization there are many tools being employed, but researchers are currently unsure of the reliability of the results these programs are producing.

Introduction

Nature uses a wide variety of protein post-translational modifications to regulate protein structure and activity[1] and tandem mass spectrometry has emerged as the most powerful analytical approach to detect these moieties[2]. However, modified peptides present special challenges for characterization. First, they are generally present at sub-stoichiometric levels, meaning that without enrichment strategies samples are dominated by unmodified peptides, so finding the modified peptides may be a challenge. Secondly, the modifications may have unique fragmentation behaviors that may need to be considered by database search engines. Finally, if there are multiple residues within a given peptide that could bear a particular modification type, then it is necessary to identify fragment ions that frame either side of the modification site in order to be able to localize the exact site of modification within the peptide[3].

Many factors can influence the ability of researchers to find and characterize post-translationally modified peptides. The choice of sample preparation method and instrumentation used to acquire data are major influencing factors. Once data are acquired, the software and analysis approaches are equally influential.

Many proteomic labs are performing post-translational modification analysis studies. It is well recognized that reliable identification of modified peptides by database searching is more challenging than unmodified, as the allowing for modifications, especially to commonly occurring amino acids, dramatically increases the number of possibilities that a search engine has to consider. Not only does this make searching of data slower, it can also lead to an increase in false negative identifications at a given false discovery rate (FDR) threshold; i.e. where correct answers are lost due to the quality of their matches no longer being deemed significantly better than random. One approach that is sometimes employed is a multi-stage searching strategy

where a protein composition list for the sample is initially derived by database searching considering a limited number of modifications, then a more extensive modification search where only modified versions of peptides derived from proteins ascertained to be present in the sample are considered[4]. This strategy can dramatically reduce the number of peptides considered and can improve sensitivity of analysis, but it is more difficult to determine the reliability of results achieved by this approach.

There are hundreds of known modifications, but it is computationally prohibitive and also counterproductive to perform searches with a very large number of potential modifications. An alternative approach that some search engines can employ to address this issue is to initially search allowing for any modification within a mass range; a strategy sometimes referred to as blind modification searching[5]. By studying a histogram of mass modifications reported from blind modification searching it is possible to determine modifications that frequently occur in the sample, which can then be specified in a subsequent conventional search of the data.

A final important aspect of PTM characterization is the ability to localize the site of modification. Most search engines do not perform this type of analysis, so it is common for secondary software to be employed for this step of analysis. The most popular tools have been A-score[6] and PTM score[7], although scoring systems derived directly from search engine results are growing in popularity due to the convenience of not needing to run an extra program in the analysis pipeline. A review of current status of site localization software has recently been published[3].

Assessing the relative effectiveness of different analysis approaches is problematic, as they are rarely broadly compared on the same sample or dataset. The Association of Biomolecular Resource Facilities (ABRF) sponsors research groups that perform studies that compare

participants' ability to perform a particular type of analysis. For the year leading up to the 2012 ABRF conference, two research groups investigated the research community's ability to identify and characterize peptides bearing a range of post-translational modifications of potential biological significance. The Proteomics Standards Research group (sPRG) produced a set of synthetic post-translationally modified peptides and supplied these to interested parties, then assessed a participant's ability to detect these by whatever sample preparation and data acquisition approach they chose. The modifications on these peptides included phosphorylation of serine, threonine or tyrosine; sulfation of tyrosine; nitration of tyrosine; acetylation of lysine; monomethylation or dimethylation of lysines or arginines; and trimethylation of lysines. The results of this study are being prepared by the sPRG for publication (manuscript in preparation).

The Proteome Informatics Research Group (iPRG) supplied an already acquired dataset to participants and assessed their ability to find post-translationally modified peptides by bioinformatic means. The results of this study are described in this manuscript. The sample for this dataset consisted of the same synthetic peptides as used in the sPRG study, but these were spiked into a tryptic yeast lysate background to simulate the challenge of finding modified peptides in a complex peptide background. The dataset consisted of roughly eighteen thousand high resolution and high mass accuracy tandem mass spectra acquired on an AB SCIEX TripleTOF 5600 mass spectrometer. The data were provided in several formats and participants were required to submit their results in a provided Excel spreadsheet template, then complete a short survey to document their approach to data analysis. All submissions were anonymous. Overall spectrum identification performance was assessed, but particular emphasis was placed on the ability to detect and localize modification sites on the synthetic peptides spiked in.

Experimental Procedures

Sample Preparation

Synthetic peptides containing a range of post-translational modifications were synthesized either in the Mechtler laboratory (Research Institute of Molecular Pathology, Vienna, Austria) or at Thermo Fisher Scientific (Bremen, Germany). Peptides were synthesized to between 60% to over 90% purity (crude synthesis) and validated by both LC-ESI-MS/MS and MALDI MS by Thermo and by members of the sPRG research committee, prior to mixing. Most peptides were greater than 90 % pure and the majority of the other components were either small molecules remaining from the peptide synthesis or the intended peptide modified with sample handling and storage modifications such as asparagine deamidation and methionine oxidation. A list of these modified peptides is provided as Supplemental Table 1.

Undigested *Saccharomyces cerevisiae* lysate, Reference Material (RM) 8323, was obtained from the National Institute of Standards and Technology (NIST) and is described at https://www-s.nist.gov/srmors/view_report.cfm?srm=8323. The S. cerevisiae, strain BY4741, was grown at Boston Biochem Inc. (Cambridge, MA) in rich (yeast peptone dextrose) medium and harvested by continuous-flow centrifugation. The cell pellet was then washed twice with ice-cold water, and lysed by incubation with ice-cold trichloroacetic acid (10 mL/L) in water for 1 h at 4 °C. The precipitate was collected by centrifugation, washed twice with 100 mL/L water in acetone, and pelleted again. The lyophilized yeast lysate was homogenized at NIST through manual grinding. The ground yeast lysate powder was suspended in 50 mmol/L ammonium bicarbonate containing 6 mol/L urea in water, pH 7.85. After gently stirring at 5 ºC overnight, the yeast lysate solution was filtered through a 0.22 µm cellulose acetate filter. To remove urea from the yeast lysate solution, the solution was thoroughly dialyzed (6,000 Da to 8,000 Da cutoff) at 5 °C using 50 mmol/L ammonium bicarbonate in water as the dialysis buffer. 40 µg of lysate was vacuum

centrifuged to almost dryness, then resuspended in 6M guanidine HCl / 25mM ammonium bicarbonate (AmBic), reduced with 2mM TCEP at 50C for 30 min, allowed to cool to room temperature then alkylated with 5mM iodoacetamide for an hour. After diluting to 1.5M guanidine HCl / 25mM AmBic, 1ug Trypsin (Promega) was added for overnight digestion. Digestion was stopped by addition of 5% formic acid, then desalted using a C18 sep-pak (Waters) cartridge. To create the spiked sample, 2 pmoles of synthetic peptides were added to 8ug of yeast lysate and dissolved in 0.1 % formic acid to a volume of 40 ul.

LC-MS/MS

5ul spiked sample was analyzed by an TripleTOF 5600 (AB SCIEX, Concord, ON) mass spectrometer interfaced with a Waters nanoAcquity UPLC system. For peptide separation a Waters Symmetry C18 180 µm x 20 mm trap column and a 1.7 µm, 75 µm x 150 mm nanoAcquity UPLC column (45ºC) was used. Trapping was performed at 5µl / min, 99% Buffer A (100% water, 0.1% formic acid) for 1 min. Peptide separation was performed at 500 nl / min with Buffer A: 100% water, 0.1% formic acid and Buffer B: 100% CH3CN, 0.075% formic acid. A linear gradient was run with 5% buffer B at initial conditions, 30% B at 70 minutes, and 85% B at 70.33 minutes. Data acquisition was performed with an AB SCIEX TripleTOF 5600 System fitted with a Nanospray III source (AB SCIEX, Concord, ON) and a pulled quartz tip as the emitter (New Objectives, Woburn, MA). Data were acquired using an ion spray voltage of 2.2 kV, curtain gas of 20 PSI. For data-dependent acquisition, survey scans were acquired in 250 ms and as many as 20 product ion scans were collected if exceeding a threshold of 150 counts per second and with a 2+ to 5+ charge-state. Total cycle time was fixed to 1.3 sec. Four time bins were summed for each scan at a pulser frequency value of 15.420 kHz through monitoring of the 40 GHz multichannel TDC detector with four-anode/channel detection. A sweeping collision energy setting of 35 ± 15 eV was applied to all precursor ions for collision-induced dissociation.

Creation of Files Required for Sequence Database Searching

The AB SCIEX MS Data Converter (AB SCIEX, Framingham, MA) was used to produce a de-isotoped peak list, containing only monoisotopic m/z values. For software that performs better with a complete peak list, containing an m/z value for every peak in each isotope distribution, the .wiff file was also processed in Mascot Distiller 2.3.2 (Matrix Science, London UK) using non-standard settings chosen to detect all features as individual peaks. This resulted in 16 additional spectra, missing from the de-isotoped peak list because they contained only very weak peaks that were probably noise. There was also a higher proportion of spectra with indeterminate precursor charge in the non-deisotoped peak list file, so the default charge settings considered when searching this file was more important.

The protein sequence database provided for all participants to use was created from a version of UniProtKB downloaded on 2012-01-05. This database was filtered to contain all yeast proteins, the bovine protein PDIA1, all human proteins, and TRYP_PIG, yielding a total of 42450 sequences. A set of decoy sequences were created by holding the location of any leading methionine and all internal lysines and arginines fixed, and scrambling the sequence of all amino acids between these fixed lysines and arginines. If a proline followed a lysine or arginine, its position was also fixed. The protein names of these scrambled sequences were the target protein name with a DECOY_ string prepended; the full protein description was set to “Decoy sequence”. Three sequence database files were made available: the target sequences file, the decoy sequences file, and both the target and decoy sequences combined into a single file, where the target and decoy sequences were interleaved.

The raw file, different processed versions of the data, protein database and Microsoft Excel Template in which to submit results were all made available for download from a file sharing site hosted by the World Design Group. All participants were requested to analyze the dataset

starting with the input data of their choice, using whatever methods they wished, to achieve two objectives: 1) Identify the CID spectra present in the sample with < 1% false discovery rate (FDR) at the peptide to spectrum match (PSM) level for matches to the target database. 2) For modified peptides, report which modification site assignments can be reliably localized. Participants were further requested to reformat the results to conform to the provided Excel template, to email the result file to an email address that was monitored by the anonymizer, and finally to fill out a survey hosted by Survey Monkey. The anonymizer, not a member of the iPRG working group, then placed the submitted files stripped of any identifying information except for a 5 digit code of the submitter's choosing in a file share for the iPRG members to analyze. The participants were also given the following information: "There are a wide variety of modifications present in this sample, both biological and chemical in nature. Naturally occurring modifications include, but are not limited to, acetylation, methylation, dimethylation, trimethylation, phosphorylation and sulfation. In the subsequent iPRG analysis of submitted results, special emphasis will be placed on the characterization of modifications not introduced by sample handling."

Perl Script for analyzing participants' identification results

Each of the submitted result files was manually reformatted with identical column headers and saved as a tab-delimited text file. The script used to read, analyze, and combine the participants results was written in the Perl programming language and executed on a computer running the Windows XP operating system. The starting content of each result submission was polished by the script to replace search engine specific terminology for spectrum identifier and the modifications present in a peptide sequence with consistent generic terminology. A subset of the spectra from 9/24 participants (71755v, 11211, 94158i, 40104i, 23117, 14151, 14152, 45511, 11821) included more than 1 identification/spectrum. The most common reason for multiple reporting was due to failure to conform to the template request to combine observations

of a single peptide occurring in multiple protein entries in the FASTA file. In all multiple reporting cases only the best scoring one was retained after sorting by score. After calculating FDR estimates but before assembling the consensus set, identifications to decoy sequences were removed. This polishing enabled the script to then compare the peptide identifications for each spectrum across all participants, create a consensus set of peptide spectrum matches, tabulate statistics related to modifications and precursor charge states, calculate consensus-set based surrogate measures of FDR for each participant, measure underperformance of individual participants by way of consensus PSM's reported with insufficient confidence to be above threshold, and calculate identification statistics and modification site false localization rates for the subset of PSMs derived from the spiked-in synthetic peptides. After polishing the submissions, modified residues were indicated by lowercase to track localization position. Sequence comparisons for consensus identification were done after converting to uppercase and replacing all isoleucines with leucines. Consequently, differences in modification localization and differences in assignment of asparagine (N) vs deamidated asparagine(n) due to precursor monoisotope assignments were treated as equivalent for determining consensus identifications. Modification nomenclature was unified for peptide N-termini and the key post-translational modifications included in this study to Arg, Lys, Ser, Thr, Tyr. The consensus identification set contains modifications to Arg (methyl, dimethyl), Lys (acetyl, methyl, dimethyl, trimethyl), Ser and Thr (phospho), Tyr (phospho, sulfo, nitro), peptide N-termini (Acetyl, Carbamyl, and Formyl). Other sample handling artifact modifications that were observed and are present in the consensus set include: Met (oxidation), Cys (pyrocarbamidmethyl, reduced sulfhydryl), Asn(deamidation), Gln (pyroglutamic acid, deamidation), Glu (pyroglutamic acid, sodiation), Asp (sodiation), Trp (oxidation), and Pro(dihydroxylation). The tabular output of the script was used with Microsoft Excel to create the charts presented in the figures. The script and participant's polished results files are available as supplementary data and can be freely downloaded through a link from the iPRG webpage[8].

Hierarchical clustering of consensus results

The 5 categories for each consensus peptide spectrum match in Supplementary Table 3 were assigned an integer value of 1 through 5, to partition the categories' relative significance (bad to good). Clustering was done with GENE-E configured to use a Pearson correlation distance metric and average linkage for both rows and columns. GENE-E can be obtained at http://www.broadinstitute.org/cancer/software/GENE-E/index.html.

Results

A total of 24 sets of results were submitted by participants. The majority of these people were from academic laboratories, although there was also representation from government and industrial laboratories as well as software vendors. Most of these participants described themselves as bioinformaticians / software developers, so one might predict that the participants were more knowledgeable than average about how to perform software analysis of mass spectrometric datasets. A total of 13 different search engines were employed by participants for peptide identification and a variety of search parameters and strategies were employed, which are summarized in the table under Figure 1, in Supplemental Table 2 and more complete details for each participant are available for download through a link from the iPRG webpage[8].

Spectrum Identification

Roughly eighteen thousand tandem mass spectra were acquired for analysis. A confident identification was reported by a range of 1 to 24 participants for each of just over ten thousand of these spectra, and a spreadsheet reporting and comparing all of these results is available in the supplementary data and through a link from the iPRG webpage[8]. The highest number of confident identifications reported by any single participant was 6669. For a total of 7840

spectra, at least 3 participants were all confident about the same interpretation of the spectrum. These answers were used as a consensus reference set to assess participant performance at spectrum identification. Thoroughly maximizing the size of the consensus set, which is primarily composed of identifications of unmodified yeast background peptides, was not a major objective of the study. While the sequences of the synthetic peptides spiked into the sample are known, there is no equivalent certainty of truth about the yeast peptides present in the sample. A threshold of 3 agreeing participants to achieve consensus was chosen for a few reasons. A threshold of 5 for consensus would have biased against the minority of participants who used the non-deisotoped peak list (discussed in greater detail below). A threshold of 2 for consensus would have biased in favor of participants 58288v and 33564 who used nearly the same software tools (PEAKS) and their results are highly correlated. Other pairs of participants using nearly the same tools include 97053i and 92653 (pFind), as well as 11211 and 14151 (Protein Pilot). In the absence of certainty of truth about the yeast peptides present in the sample it seems sensible to be concerned that a consensus-of-3 approach might inappropriately brand consensus false-positives as correct due to a combination of participants using similar methods that are not readily apparent. Consequently, we used hierarchical clustering to measure how similar participants' results are to each other. Clustering of all 7840 consensus PSM's shown in Supplemental Figure 1a illustrates the similarity of results from the pairs of participants described above and indicates that there is not a set of 3 participants that are inordinately similar enough to suggest an unreasonable bias in their favor. Clustering of the subset of 605 consensus PSM's identified by only 3 participants shown in Supplemental Figure 1b shows a significant cluster whose primary unifying attribute is the use of the non-deisotoped peak list. Despite the use of Mascot by 6 of the 19 participants who reported unmodified peptides, the clustering reveals only adjacent pairs of highly-correlated Mascot users. 3 of the 6 participants using Mascot did so in conjunction with other search engines.

Figure 1 shows a breakdown of each participant's reported results. A wide range in number of identifications were reported, although it should be noted that the 6 right-most participants in Figure 1 focused solely on the modified peptides. Hence, the results to the right of this chart do not necessarily represent these participants' ability to identify spectra in general. The number of identifications from participant 23117 is inordinately low because multiple spectra derived from repeat MS/MS of the same precursor were merged, while everyone else reported individual peptide spectrum matches. 13/24 participants provided decoy PSM's in their results, so we were able to independently compute their FDR. Based on comparing reported confident results to the consensus results, most participants had a PSM level FDR of between 1-2 %, so their estimates of spectrum identification reliability were generally fairly accurate. Hence, they drew their threshold at approximately the correct score/expectation value. While the participants with the tallest blue bars in Figure 1 might be considered the top performers from an identification perspective, the gaps between the top of the blue bars and the threshold of 7840 consensus identifications suggest that there is substantial room for improvement by all. In dissecting the limits of participants' abilities to report confident identifications, the gray, yellow, red and green portions of the bars in Figure 1 highlight potential target areas of improvement in each participant's search strategy or the underlying scoring mechanisms of the tools used. Supplemental Table 3 provides a spectrum level alignment of the consensus identifications across all participants. A table of all results, not thresholded to any consensus level, is also available in the Supplementary Data. These tables are intended resources to facilitate specific improvements, and include clickable links to visualize labeled spectra.

The gray portion of a bar in Figure 1 represents identifications (if provided) that differed from the consensus and were below the FDR threshold. These will tend to include examples where the spectral pre-processing or search engine parameters did not allow for a particular peptide to be matched (i.e. precursor monoisotope correction, unanticipated modifications).

The yellow portion of a bar in Figure 1 indicates spectra that a participant confidently identified, but for which there was no consensus agreement by 3 or more participants. This category will contain both false positives and examples where participants, their tools, or search parameters exhibit extraordinary skill or insight. Unfortunately, in the absence of more thorough analysis of the results, such as inspection of each spectrum, it is not possible to readily discern between fool and phenom.

The red portion of a bar in Figure 1 indicates spectra that a participant confidently identified, but where their identifications differed from the consensus. Although this category represents identifications that are highly likely to be incorrect, it contains identifications that are frequently highly homologous to the consensus identification. Comparison of sequences for determining the consensus identification allowed for indistinguishable differences between leucine and isoleucine, modification localization, differences in modification localization, and differences in assignment of asparagine (N) vs deamidated asparagine (n) due to precursor monoisotope assignment differences to be counted as equivalent. However, there were a variety of other homologous answers that were counted as different not only because handling each of them would be challenging but also because the consensus result is clearly better, in that it represents a simpler or more realistic identification that software tools should be expected to prioritize when reporting a representative single identification from multiple tie-scoring possibilities. These include peptides from two different isoforms of a protein where one peptide is modified and the other is not: (Asp vs deamidated Asn), Gln vs Lys (when the precursor mass error in the dataset can readily differentiate). For spectra where the consensus identification was a protein N-terminus with leading Met removed followed by acetylated serine participants 58288v and 33564 used nearly the same software tools and both consistently reported those peptides with the isobaric alternative of a serine to glutamic acid substitution. Also falling into this category are spectra that had different precursor m/z or charge due to differences in the

peak lists used by the participants (see Effect of Peak List section below). For participant 11211 this category is further complicated by the handling of multiple identifications reported for some spectra. In particular, sometimes an identification that was the best scoring was marked as not confidently identified and a lower scoring one was marked as confidently identified. The participant's tools may have made the choice by combining multiple scores and protein parsimony considerations rather than relying solely on the primary score submitted with the results. Nonetheless, for this study the confidently identified ones were chosen as the sole representative for those spectra. We tested a parsing variant of choosing the best scoring instead. This resulted in net changes of -153 spectra to the red category, +193 to the green category, -1 to the yellow category, and -3 to the blue category, a decrease of 11211's decoy based FDR estimate from 1.59% to 0.93%, and -1 to the overall number of identifications in the consensus set.

The green portion of a bar in Figure 1 represents results that agree with the consensus interpretation but the participant reported as being below the 1 % FDR threshold that they were asked to employ for their results. Overall, this shows that search engines were generally performing quite effectively at ranking the consensus answer as the top result, but there is significant room for improvement in separating correct results from incorrect. While the scoring schemes of particular search engines may be somewhat more adept than others, examination of the key aspects of search strategies noted in Supplementary Table 2 indicates that the largest green bars in Figure 1 appear to correlate with participants who used single pass search strategies with a large array of variable modifications allowed. Others used a multi-pass approach with an initial round that focused on common modifications searched against the entire sequence database and subsequent rounds, often limited to proteins confidently identified in the initial round. For the single-pass strategy users, the large effective size of the database search can be expected to yield higher score thresholds in order to meet the target FDR and

thus lower total numbers of confident identifications. Multi-pass strategy users can be expected both to have had a harder time calculating an accurate FDR and have been more reliant on manual examination of the results.

Effect of Peak List

Peak list data were supplied in several different formats, and among these there were two variants in the translation of the raw profile data into a list of centroided m/z. The most commonly used peak lists were those where the data had been deisotoped. Search engines are designed to search with a list of masses that correspond only to monoisotopic peaks. Hence, deisotoped peak lists would seem attractive. However, their limitation is that by removing the isotopes it is no longer possible to determine the charge state of the fragment ions. By throwing away this information, search engines typically must search data allowing for any peak to correspond to being either singly- or doubly-charged; effectively doubling the number of masses being considered. If a search engine can determine the charge state of fragment ions and then deisotope, then they should be able to perform better with a non-deisotoped peak list. Unfortunately, different software had to be used to create the deisotoped and non-deisotoped peak lists from the raw data. In both cases the software tried to re-determine the monoisotopic m/z of the precursor, with differing success. For 309 of the 7840 spectra with consensus identifications, the software that created the deisotoped peak list was unable to decide on the precursor charge state, whereas the charge state was undefined for 1147 of these spectra in the non-deisotoped peak list files. For 238 spectra the precursor charge state was not only unambiguously assigned in each peak list set but different between the two sets. The non-deisotoped peaklist's precursor charge matched the consensus identification for only 21% (49/238) of these cases. Furthermore, for 1014 of these spectra the precursor m/z differed by more than 0.02 Da between file types. Most of these differences were due to one peak list reporting the second isotope instead of the monoisotopic peak. The

precursor mass for the non-deisotoped peak list was consistent with the consensus sequence identification (+/- 20 ppm) in only 23% (234/1014) of the cases. Identifications from participant 87048i, who used only the non-deisotoped peak list, were amongst the consensus for only 103 of these 1013 spectra. That participant's search engine does not allow for ambiguous precursor monoisotopic assignments. When the two peak lists differed in precursor monoisotopic peak assignment the de-isotoped precursor was heavily favored to become the consensus identification due to its more frequent use. Clearly, the peak list generation software has substantial room for improvement in this area. However, there were a few examples where there were two or more precursors within the isolation window and one peak list reported one as the precursor and another was listed by the other peak list. Figure 2 shows an example where a different peptide was reliably identified depending on which peak list was searched. Other examples can be found in Supplemental Table 3 by noting where 94158i, 77777i, and 87048i agree on a confident assignment that differs from the consensus and the non-deisotoped peak list precursor disagrees with the de-isotoped one. As noted in Figure 1, these three participants were amongst the minority that used the non-deisotoped peak list.

Modified Peptide Identification

Of the 7840 consensus spectra results, 1723 (22 %) contained a modification. Protein N-terminal acetylation was the most common modification observed, followed by asparagine deamidation. However, there were also several hundred spectra of the spiked-in, modified synthetic peptides. Of the 70 spiked-in peptides (see Supplemental Table 1), only 63 were reported by at least one participant (no participant attempted to differentiate between symmetric and asymmetric arginine dimethylation, so there were only 69 possible answers). Figure 3 shows a plot representing which of the spiked-in peptides were identified by each participant, with peptides bearing the same type of modification grouped together. The most obvious observation from this plot is that most participants struggled with identifying the sulfated

peptides. Sulfation is almost isobaric with phosphorylation (they differ in mass by 9.5 mmu), so it can easily be mistaken for phosphorylation based on the mass. However, it does behave differently under CID fragmentation, as shown in Figure 4. Whereas the phosphate moiety is retained on some fragment ions and others lose phosphoric acid (-98 Da), sulfopeptides promptly lose sulfate (-80 Da) to produce fragmentation spectra that resemble that of the unmodified peptide.

As shown in Figure 3, participants 94158i and 87048i appear to have been particularly adept at correctly identifying and localizing the tyrosine-sulfated peptides. Participant 87048i provided the following feedback. “Although my searches were performed with a precursor mass tolerance of +/-20 ppm, the mass error was typically +/-5 ppm. Hence it was easy to see that the precursor mass errors relative to the unmodified peptide masses had 1 cluster with a large number of matches near +79.9663 (the expected shift due to phosphorylation) and another one with about a dozen matches near +79.9568 (the expected shift due to sulfation). Manual inspection of MS/MS spectra in the smaller cluster revealed a peak consistent with neutral loss of 80Da from the precursor mass and all b/y type ions consistent with an unmodified peptide. Hence, these couldn’t be phosphorylated and must instead be sulfated. Since each peptide had only a single tyrosine (where biological sulfation occurs), localization was trivial. Since the search engine used, Spectrum Mill, doesn’t specifically incorporate these features of sulfation into its current scoring, all these spectra would have otherwise been interpreted in fully automated mode as phosphorylated peptides with ambiguous localization of the modification on s, t, or y if the iPRG study instructions had not led me to anticipate the presence of sulfated peptides in the sample. I had not previously worked on sulfated peptides, but a web search for sulfation led me to a description of expected MS/MS spectral characteristics.” Participant 94158i had prior experience analyzing sulfated peptides and the search engine he used (Protein Prospector) assumes that the sulfate moiety is only observed as a neutral loss from all fragment

ions; i.e. none of the fragment ions are expected to contain the modification. This allowed reliable differentiation between sulfation and phosphorylation for the peptides in this study. However, because no modified fragment ions were considered, the participant reported all sulfation site localizations as ambiguous.

Figure 3 also shows that certain participants missed all spectra containing a particular modification type, suggesting they may not have considered it as a possibility in their search engine settings. In the study description it was explicitly stated that all of these modifications with the exception of tyrosine nitration were present in the sample. The number of people who did not identify any nitrotyrosine-containing peptides was no different to several other modifications. This may be partly explained by the fact that more than half of the participants either took part in the sPRG study or were aware of it, so they may have known from another source that this modification was present. The peptides that were not identified by any participant included the four quadruply-phosphorylated peptides, one of the sulfated peptides and one of the arginine dimethylated peptides. Upon examination of the raw MS data using extracted ion chromatograms, a peak for the dimethylated peptide could be observed (and an MSMS spectrum of this peak was reported as being this peptide, but it was below the 1 % FDR threshold). For the other 'missing' peptides there was no evidence for their precursor being present, so these were presumably lost during sample handling or separation prior to reaching the mass spectrometer.

Modification Site Localization

In their submitted results, participants were asked to indicate whether they could confidently localize all modifications within a given spectrum identification. For the spiked-in peptides the sites of modification are known. Hence, it is possible to accurately assess the participant's performance at site localization for this subset of the data. Figure 5 plots in the upper panel the

number of spectra matched to synthetic peptides, broken down by whether the site assignment was correct and whether the person indicated they thought that the localization was reliable. The lower panel reports a false localization rate (FLR) for the participant's results.

It is currently difficult using site localization software to produce an estimate of site localization reliability for a dataset as a whole. Hence the study guidelines did not specify a desired FLR threshold to apply. Based on the ratio of green to black bars for different participants in Figure 5 it can be seen that some participants were much more conservative than others with respect to whether modifications could be localized and two participants decided to not report any modifications as confidently localized, presumably as they were unfamiliar or not confident with the use of any tool that can assess site localization reliability. Fifteen different named programs were reported to be used for site localization and a further three participants used in-house tools. Interestingly, there was no correlation between the number of spectra for which site localization was confidently assigned and the participant's FLR. This suggests that there must be significant variability in the performance of the site localization software tools and that this is an area where there is room for improvement. When analyzing which peptides led to the most errors in site localization two peptides caused more errors than others: the peptides THILLFLPKS(Phospho)VSDYEGK and TVIDY(Sulfo)NGER. The former of these peptides contains two other potential phosphorylation sites nearby (Ser12 and Tyr14) and these other sites were often reported as the site of modification. The latter sulfopeptide was commonly reported as being phosphorylated on the N-terminal threonine. The sulfate moiety is promptly lost in CID to give a spectrum that contains only unmodified fragment ions. As quadrupole CID spectra of tryptic peptides predominantly contain y ions, by locating a phosphorylation on the N-terminal residue, then a complete unmodified y ion series will be matched, similarly to a sulfopeptide. Hence, if any software searched for sulfation assuming the modification would

remain on the fragment ions, then reporting phosphorylation of the threonine would likely score better than tyrosine sulfation.

Discussion

This study was able to broadly assess three attributes of database searching for peptide identification: how many spectra participants were able to identify when allowing for a range of post-translational modifications; how many of a specific set of spiked-in modified peptides could be identified; and how effective are people at assessing the reliability of modification site localizations.

Of those people who made an effort to identify and report all peptides (rather than only focusing on those modified), there was a variability of about two-fold between the most successful and least successful participants. This is a larger spread than one would expect in a typical CID peptide identification study. This could partly be an effect of searching quadrupole CID fragmentation spectra using tools that people are more used to analyzing ion trap CID data with. However, it is likely that the need to consider a wide range of modifications on the peptides was probably the major contributor to this variability: by significantly opening up the search space in order to allow for these PTMs it becomes more difficult to distinguish between correct and random results, and the large number of false negative identifications by many participants as shown in Figure 1 is evidence to support this as being a factor. One facet that has been significant in previous iPRG studies has been whether a participant considered peptides that are only partially tryptic in specificity. However, of the consensus results for this study only about 2 % are non-tryptic, so this was not an important parameter for success. A potential reason why this was less of a factor than in some previous studies is that this sample was analyzed unfractionated, whereas studies in previous years have employed strong cation exchange

chromatography and also phosphopeptide enrichment in the iPRG 2010 study. Non-tryptic peptides should be present at low stoichiometry, so are less likely to be detected without sample fractionation / simplification.

Of the 70 synthetic peptides introduced into the sample, 22 out of 24 people found thirty or more of them and half the submissions identified 44 or more. If one takes into consideration that for five of these peptides there was no MSMS spectrum acquired (and indeed there was no evidence for them at the MS level), then half the participants found two-thirds of those possible to be detected. Sulfated peptides presented a challenge to all participants; only 6 people found any of them. This is probably due to a lack of general experience in analyzing this PTM; tyrosine sulfation is an extracellular PTM, whereas the majority of PTM analysis is currently performed on intracellular proteins. Lack of knowledge of how this modification behaves in CID (either by software or by user) led to many people either completely missing these peptides or reporting them as phosphopeptides instead (half of the participants reported one or more spectra of a sulfopeptide as a phosphopeptide).

Site localization software is generally still in its infancy, and this is clearly an area where further work on tools is of immediate priority. Due to this, and in particular due to the lack of a simple, generic approach to estimate an FLR, the community has not settled on an appropriate reliability threshold that should be employed for site localization reporting. Most participants reported results with an FLR between 2-8 %; i.e. their site localization reporting was less reliable than their spectrum identification. It should be mentioned that for several of these peptides there was no possibility of incorrectly localizing the modification; e.g. all of the tyrosine nitrated peptides contained a single tyrosine residue. In a related situation, all of the tyrosine sulfated peptides also contained only a single tyrosine. However, the difference in behavior of these two modifications in the mass spectrometer makes site localization very different tasks: nitration is stable under CID conditions, so observation of modified fragment ions can make site

assignment straightforward. However, the extreme lability of the sulfate group means that modified fragments are unlikely to be observed, so if any participant claimed they had confidently localized the sulfation site, they probably did so on the basis of the assumption that tyrosine is the only residue that could have been modified, rather than based on any evidence in the MSMS spectrum. As sulfation has also been observed on serines, threonines [9] and cysteines[10], this could be a dangerous assumption.

This study also highlighted that the peak list generation software could have a significant effect on results. Both peak list generation software tools made many mistakes in their attempts to determine the precursor charge and monoisotopic peak for each spectrum. As a result it was not possible to draw any conclusions about the benefits/drawbacks of deisotoping peak lists.

Performing an unbiased comparison of different software analysis approaches and tools is a difficult task. Several published studies have tried to apply identical parameters for a selection of tools and comparing results[11-13]. However, a given set of parameters always slightly favors one tool over another. An alternative approach is to compare results when employing what are deemed optimal parameters for each program. However, this normally leads to a bias in results related to the experience level of the person with each software program. Hence, probably the most informative comparisons are where a large range of different people analyze the same dataset. This is where standard datasets, such as the one produced in this study, are so important to the research community[14].

Another issue with datasets is assessment of truth. It is practically impossible to model a sample of typical complexity for a proteomic experiment where all of the correct results can be known. This study provides an answer key with two different levels of confidence/reliability. Spectra of the synthetic modified peptides spiked into this sample can be matched with high confidence and reliability. In addition, an answer key was created based on consensus results.

Together, this provides an extensive set of results that can be used by software tool developers to benchmark performance or improvements in their tools[14].

All data, including raw file, peaklists in various formats, protein databases and Excel spreadsheet containing all participant's results including consensus assignments, are available for download through a link from the iPRG webpage[8]. Also made available are poster and Powerpoint summaries of the study, as well as scripts and instructions for how someone can add and compare their own results to those submitted during the study.

Acknowledgements

This work was supported by funding from ABRF. We would like to thank the sPRG committee of ABRF for producing the synthetic peptides; Chris Colangelo for acquiring the data used for the study; and Jeremy Carver for acting as the anonymizer of participant submissions. We would also like to thank all the participants, without which this study would not have been meaningful.

References

1. Walsh, C.T., Garneau-Tsodikova, S., and Gatto, G.J., Jr. (2005) Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl*. 44(45): 7342-72.
2. Zhao, Y. and Jensen, O.N. (2009) Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics*. 9(20): 4632-41.
3. Chalkley, R.J. and Clauser, K.R. (2012) Modification site localization scoring: strategies and performance. *Mol Cell Proteomics*. 11(5): 3-14.
4. Beavis, R.C. (2006) Using the global proteome machine for protein identification. *Methods Mol Biol*. 328: 217-28.
5. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P.A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol*. 23(12): 1562-7.
6. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*. 24(10): 1285-92.
7. Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*. 127(3): 635-48.
8. *iPRG webpage*. Available from: http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm.
9. Medzihradszky, K.F., Darula, Z., Perlson, E., Fainzilber, M., Chalkley, R.J., Ball, H., Greenbaum, D., Bogyo, M., Tyson, D.R., Bradshaw, R.A., and Burlingame, A.L. (2004) O-sulfonation of serine and threonine: mass spectrometric detection and characterization of a new posttranslational modification in diverse proteins throughout the eukaryotes. *Mol Cell Proteomics*. 3(5): 429-40.
10. Lim, A., Prokaeva, T., McComb, M.E., Connors, L.H., Skinner, M., and Costello, C.E. (2003) Identification of S-sulfonation and S-thiolation of a novel transthyretin Phe33Cys variant from a patient diagnosed with familial transthyretin amyloidosis. *Protein Sci*. 12(8): 1775-85.
11. Kapp, E.A., Schutz, F., Connolly, L.M., Chakel, J.A., Meza, J.E., Miller, C.A., Fenyo, D., Eng, J.K., Adkins, J.N., Omenn, G.S., and Simpson, R.J. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*. 5(13): 3475-90.
12. Balgley, B.M., Laudeman, T., Yang, L., Song, T., and Lee, C.S. (2007) Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics*. 6(9): 1599-608.
13. Kandasamy, K., Pandey, A., and Molina, H. (2009) Evaluation of several MS/MS search algorithms for analysis of spectra derived from electron transfer dissociation experiments. *Anal Chem*. 81(17): 7170-80.
14. Yates, J.R., 3rd, Park, S.K., Delahunty, C.M., Xu, T., Savas, J.N., Cociorva, D., and Carvalho, P.C. (2012) Toward objective evaluation of proteomic algorithms. *Nat Methods*. 9(5): 455-6.

Figure Legends

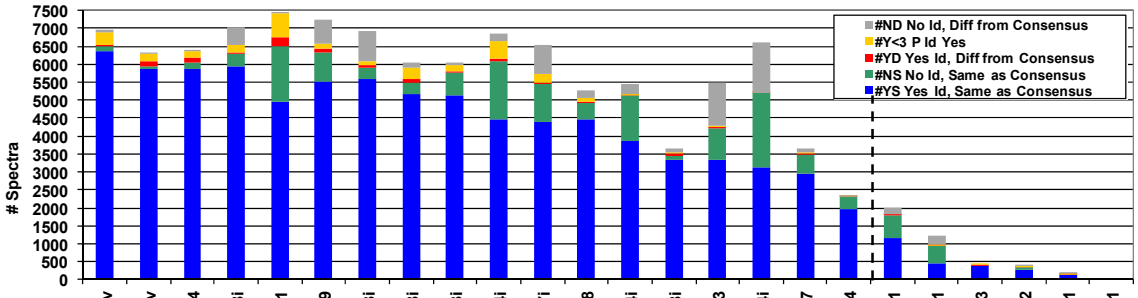
Figure 1: Summary of peptide identification results reported by participants. Submitted results were compared to consensus results. Participants were also asked to indicate whether a particular result was above or below their estimated 1% FDR threshold. The plot separates results based on whether assignment were deemed correct according to consensus results and whether a result was reported as significant by the participant: blue: consensus identification better than 1% FDR threshold; green: consensus identification worse than 1% FDR threshold; red: non-consensus result better than 1% FDR threshold; yellow: no consensus result (too few participants reporting a result to reach a consensus) better than 1% FDR; grey; non-consensus result worse than 1% FDR threshold. The five participants to the right of the dotted line reported only modified peptides. Participant 23117 merged in an additional 2257 spectra derived from repeat MS/MS of the same precursor (each merged set counts here as only one spectrum), while everyone else reported individual peptide spectrum matches.

Figure 2: Co-isolated precursor ions led to different peptide identifications for a single spectrum depending on the source of the pre-processed peak list used for searching. In the supplied deisotoped peak lists the spectrum was indicated as having the precursor m/z 465.19 2+; whereas in non-deisotoped peaklists the precursor was indicated as m/z 464.59 3+. A) In the MS1 scan that triggered the MS/MS spectrum the 3+ precursor is < 5% as abundant as the 2+ precursor. B)The resulting spectrum is able to match y2-y7 of the peptide SVSDY(Nitro)EGK for the 2+ precursor ion and C) y3-y8, as well as doubly charged ions of y10-y12, to the peptide LAAPENEKPAPVR for the 3+ precursor ion (the y-axis has been magnified relative to the base peak to allow easier visualization of the peaks).

Figure 3: Heat map plot reporting which peptides were identified by each participant. Each row represents one of the 70 spiked in synthetic peptides, which are grouped by modification type, whereas each column is a participant. Modification site localization was ignored for this plot.

Figure 4: Comparison of fragmentation spectra of a tyrosine phosphorylated and tyrosine sulfated peptide. The peptide DISLSDYK was synthesized with either a phosphorylation or sulfation on its tyrosine residue. The phosphopeptide spectrum contains ions y2-y6 where the phosphate group is retained, whereas the equivalent ions in the sulfopeptide spectrum are all observed 80 Da lower in mass due to prompt loss of SO3.

Figure 5: Summary of PTM site localization results by participants. For modified peptides, participants were required to indicate on which residue they believed the modification was localized and whether they were confident of the site localization. The top plot reports the number of spectra from the synthetic modified peptides that were identified by each participant. Green: correct site localization and confident assignment; red: incorrect site localization and confident assignment; grey: peptide contained additional modifications for which site localization is unknown but site localization was reported as confident; black: site localizations in peptide were reported as not confident. The lower panel reports a false localization rate (FLR) for reported results. This was calculated by dividing the number of confident incorrect site localization spectra (red portion of upper plot) by the total number of spectra for which the participant reported they were confident of site localizations. Values under the axis were participants' estimates of their FLR.



	71755v	58288v	33564	93128i	11211	58409	87133i	94158i	97053i	42424i	77777i	23068	40104i	87048i	92653	34284i	23117	74564	14151	52781	47603	14152	45511	11821
Peaklist	mgf	mzML	mzML	mgf	mgf	mgf mzML	mzML	mgf_nd	mgf	mgf_nd mzML WIFF	mgf mgf_nd	mzML	mzML	mgf_nd	mgf	mgf	mzML	mgf	mgf_nd	mgf	mgf	mgf_nd	mgf	mgf_nd
Spectral Pre-Processing		Pk PkDB	Pk		PPi					MDi PW		Ot		SM	pF		Pk PkDB						MDi	PD Sq
Peptide Identification	By	Pk PkDB	Pk PkDB	M	PPi	M O ST XT Ot	O ST XT	PPr	pF	M MM IH	MG	M O XT	P/PP TPP XT	SM	pF	M	Pk PkDB	M	PPi	M	O	M	M	PD Sq
Discovery of Unexpected Mods	By	Pk PkDB	Pk	M	PPi	ST		PPr	pF	IH	MO			SM	pF	M	Pk PkDB		PPi		O		M	
Site Localization		Pk PkDB	Pk	MDe	PPi	M O ST		PPr	pF	M MM IH	IH	M O		SM	pF	M	Pk PkDB	M		AS Sc	IH	An	MDe Ot	PD Sq
Results Filtering	By	Pk PkDB	Pk XL	IH Ot	PPi	P/PP TPP	P/PP TPP	PPr XL	pF	IP	IH	XL	IH	XL	pF	M	Pk PkDB	M			XL	Sc	M	PR
NTT	2	1	1	2	1	?	1	1	1	2	1	?	1	2	2	2	2	?	1	2	2	2	?	2
Experience	5-10 years	5-10 years	5-10 years	>10 years	< 1 year	3-4 years	5-10 years	>10 years	5-10 years	5-10 years	5-10 years	3-4 years	5-10 years	>10 years	1-2 years	>10 years	1-2 years	>10 years	>10 years	1-2 years	>10 years	>10 years	1-2 years	>10 years

An	Andromeda/MaxQuant		MG	MS-GFDB		pF	pFind	Sc	Scaffold
AS	A-Score		MM	MyriMatch		Pk	PEAKS	SM	Spectrum Mill
By	Byonics		MO	MODa		PkDB	PEAKSDB	Sq	Sequest
IH	In-house software		O	OMSSA		PPi	Protein Pilot	ST	SpectraST
IP	IDPicker		Ot	Other		PPr	Protein Prospector	TPP	TransProteomic Pipeline
M	Mascot		P/PP	Pep/Prot Prophet		PR	PhosphoRS	XL	Excel
MDe	Mascot Delta Score		PD	ProteomeDiscoverer		PW	ProteoWizard	XT	X!Tandem
MDi	Mascot Distiller								

Figure 1:

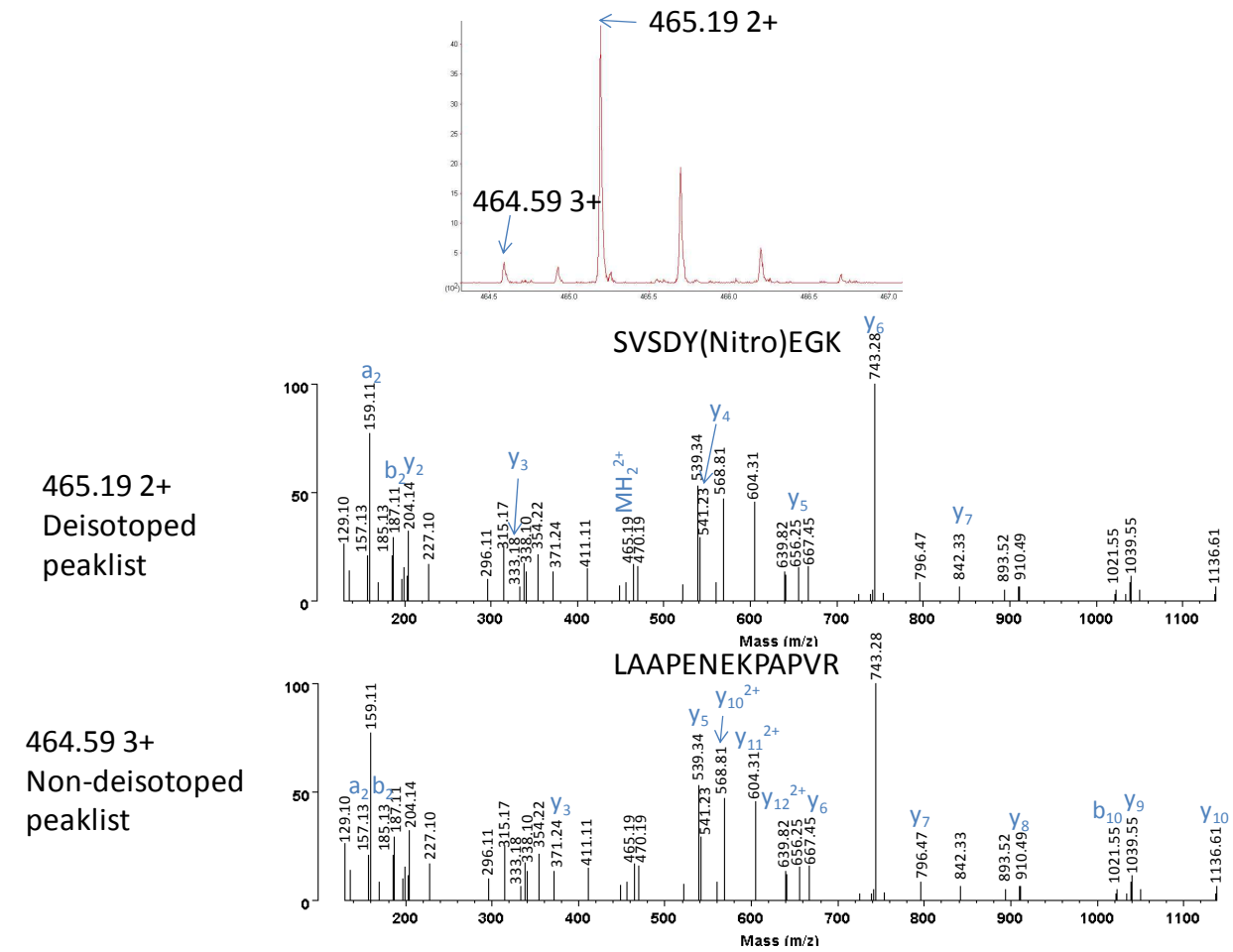

465.19 2+
464.59 3+
SVSDY(Nitro)EGK
465.19 2+
Deisotoped
peaklist
LAAPENEKPAPVR
464.59 3+
Non-deisotoped
peaklist
Mass (m/z)

Figure 2


71755v
58288v
33564
93128i
11211
58409
87133i
94158i
97053i
42424i
77777i
23068
40104i
87048i
92653
34284i
23117
74564
14151
52781
47603
14152
45511
11821
Acetyl (K)
Dimethyl (K)
Dimethyl (R)
Methyl (K)
Methyl (R)
Nitro (Y)
Phospho (STY)
Sulfo (Y)
Trimethyl (K)

Figure 3

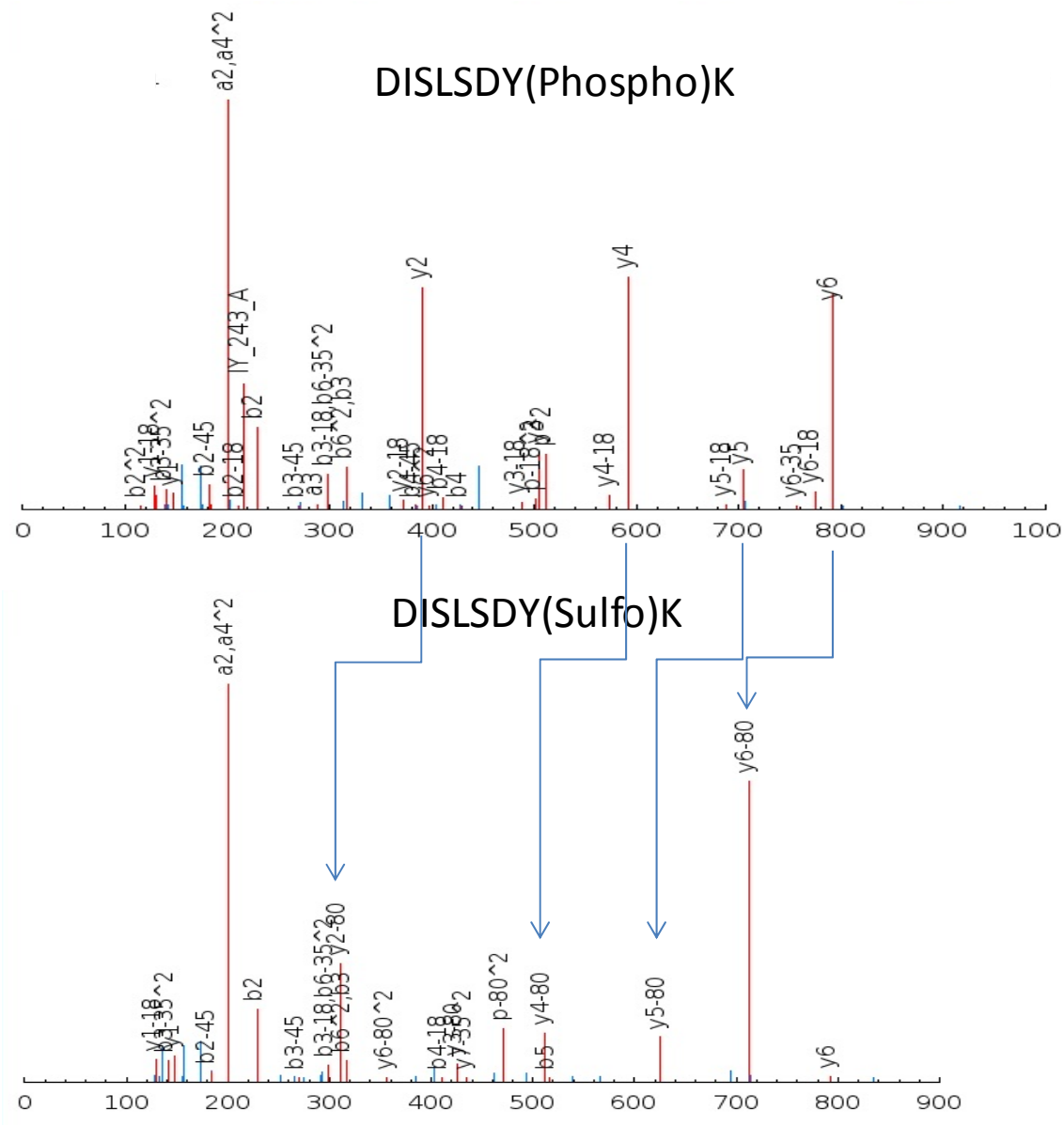


Figure 4

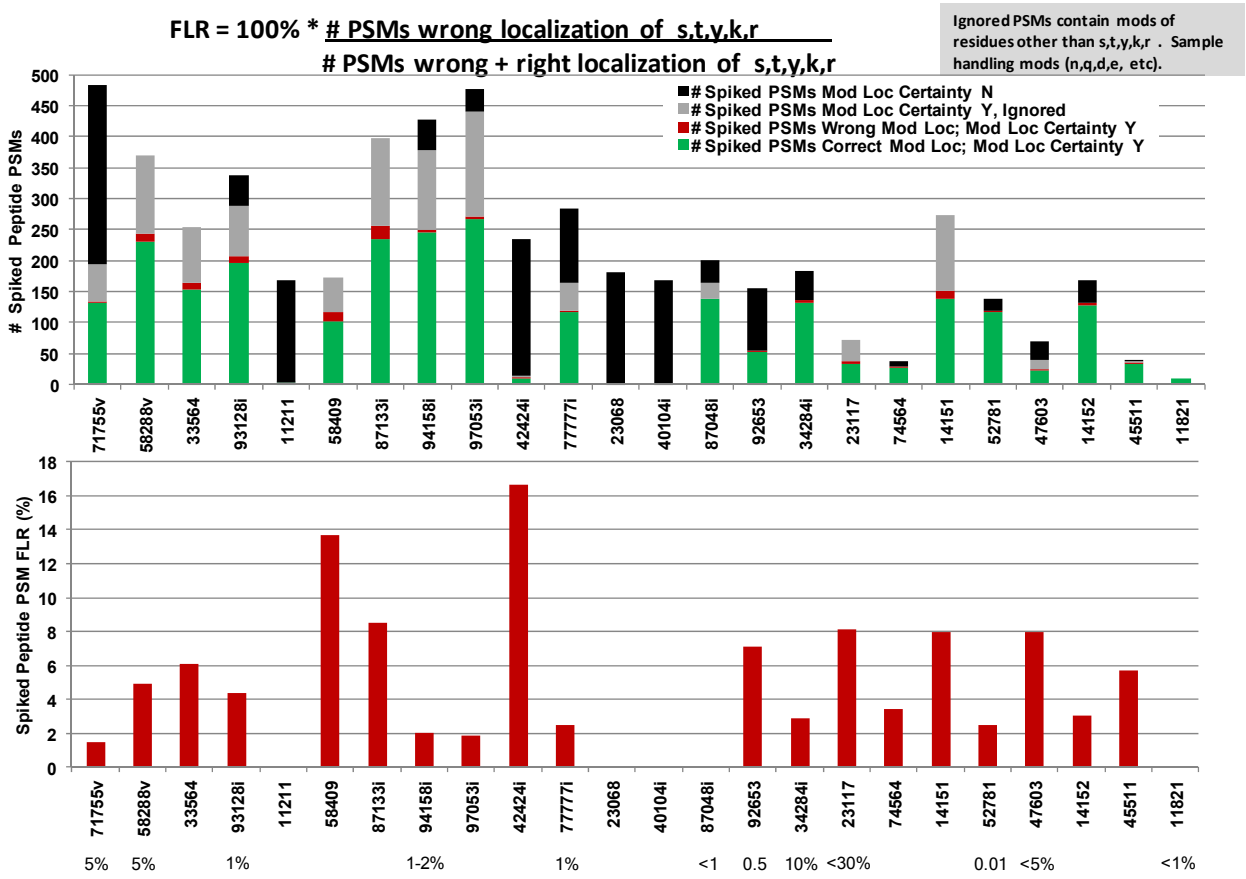

FLR = 100% * # PSMs wrong localization of s,t,y,k,r
# PSMs wrong + right localization of s,t,y,k,r
Ignored PSMs contain mods of residues other than s,t,y,k,r . Sample handling mods (n,q,d,e, etc).
Spiked Peptide PSMs
Spiked PSMs Mod Loc Certainty N
Spiked PSMs Mod Loc Certainty Y, Ignored
Spiked PSMs Wrong Mod Loc; Mod Loc Certainty Y
Spiked PSMs Correct Mod Loc; Mod Loc Certainty Y
Spiked Peptide PSM FLR (%)
71755v 58288v 33564 93128i 11211 58409 87133i 94158i 97053i 42424i 77777i 23068 40104i 87048i 92653 34284i 23117 74564 14151 52781 47603 14152 45511 11821
5% 5% 1% 1-2% 1% <1 0.5 10% <30% 0.01 <5% <1%

Figure 5