Universal and Confident Phosphorylation Site Localization Using phosphoRS

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · NOVEMBER 2011

Impact Factor: 4.25 · DOI: 10.1021/pr200611n · Source: PubMed

Thomas Köcher

Research Institute of Molecular Pathology

51 PUBLICATIONS **3,583** CITATIONS

SEE PROFILE

Andreas Schmidt

Ludwig-Maximilians-University of Munich

19 PUBLICATIONS **778** CITATIONS

SEE PROFILE

Karl Mechtler

Research Institute of Molecular Pathology

166 PUBLICATIONS **16,035** CITATIONS

SEE PROFILE

Journal of proteome research

ARTICLE

pubs.acs.org/jpr

Universal and Confident Phosphorylation Site Localization Using phosphoRS

Thomas Taus,[†,#] Thomas Köcher,[*,†,#] Peter Pichler,[‡] Carmen Paschke,[§] Andreas Schmidt,[‡] Christoph Henrich,[§] and Karl Mechtler[*,†,||]

[†]Research Institute of Molecular Pathology (IMP), Dr. Bohrgasse 7, A-1030 Vienna, Austria

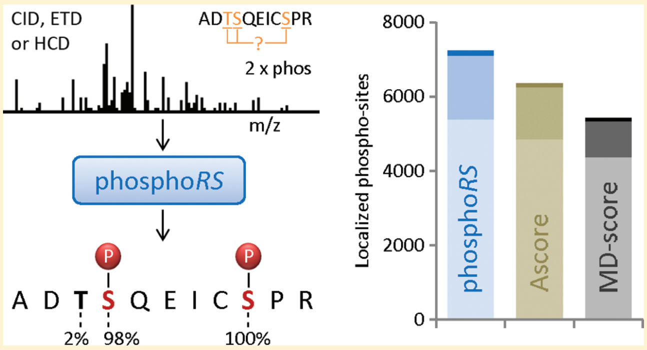
[‡]Christian Doppler Laboratory for Proteome Analysis, University of Vienna, Vienna, Austria

[§]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany

[||]Institute of Molecular Biotechnology (IMBA), Vienna, Austria

S Supporting Information

ABSTRACT: An algorithm for the assignment of phosphorylation sites in peptides is described. The program uses tandem mass spectrometry data in conjunction with the respective peptide sequences to calculate site probabilities for all potential phosphorylation sites. Tandem mass spectra from synthetic phosphopeptides were used for optimization of the scoring parameters employing all commonly used fragmentation techniques. Calculation of probabilities was adapted to the different fragmentation methods and to the maximum mass deviation of the analysis. The software includes a novel approach to peak extraction, required for matching experimental data to the theoretical values of all isoforms, by defining individual peak depths for the different regions of the tandem mass spectrum. Mixtures of synthetic phosphopeptides were used to validate the program by calculation of its false localization rate versus site probability cutoff characteristic. Notably, the empirical obtained precision was higher than indicated by the applied probability cutoff. In addition, the performance of the algorithm was compared to existing approaches to site localization such as Ascore. In order to assess the practical applicability of the algorithm to large data sets, phosphopeptides from a biological sample were analyzed, localizing more than 3000 nonredundant phosphorylation sites. Finally, the results obtained for the different fragmentation methods and localization tools were compared and discussed.



KEYWORDS: mass spectrometry, protein phosphorylation, phosphopeptides, tandem mass spectrometry, identification, localization

INTRODUCTION

The ability to characterize proteins in an unbiased fashion using mass spectrometry-based methods[1] has led to crucial discoveries in biological research. Ultimately, the technology will influence our understanding of pathological processes, with a potential role in enabling systems biology approaches in biomedical research.[2] In essence, this prediction is based on the tremendous and ongoing advances in chromatographic and mass spectrometric instrumentation. In recent years, the technology matured from a method capable of unambiguously identifying minute amounts of single proteins to a high-throughput technology enabling the identification and quantification of thousands of proteins in a single experiment.[3] In addition, proteins can be further characterized by identifying their interaction partners[4] and their post-translational modifications (PTMs).[5] In conjunction with sophisticated enrichment methods, mass spectrometry has become the method of choice to study phosphorylated proteins.[6,7] Large-scale studies of protein phosphorylation are of utmost importance because of the major role of this widespread PTM in many important cellular processes. In addition, aberrant phosphorylation can be a major cause of diseases such as cancer.[8]

Most approaches to the characterization of PTMs follow the basic principles developed for protein identification. Main steps of the workflow are the digestion of proteins with specific proteases, the online separation of the resultant proteolytic peptides by high performance liquid chromatography (HPLC), and their analysis by tandem mass spectrometry (MS/MS).[9] However, the identification and localization of PTMs such as phosphorylation sites is complicated by additional factors. In order to overcome these issues, a plethora of different strategies have been developed to facilitate the comprehensive identification of phosphorylation sites.[6,7]

Maybe one of the most significant complications is that proteins can be identified with any fragment ion spectrum with sufficient spectral quality and informational content, whereas the successful analysis of PTMs requires the detection of all peptides containing the PTM. Second, PTMs are often present at substoichiometric

Received: June 27, 2011
Published: November 10, 2011

levels. Consequently, chromatographic methods were invented to enrich phosphopeptides, improving the achievable limit of detection.[5] Third, intrinsic features of some PTMs such as the neutral loss of phosphoric acid from phosphopeptides upon collision induced dissociation (CID) make the interpretation of these spectra challenging. The alternative and complementary[10] fragmentation techniques electron capture dissociation (ECD)[11] or electron transfer dissociation (ETD)[12] maintain the phosphate group upon fragmentation, allowing the above-mentioned limitation to be overcome. Fourth, the fragment ions present in the tandem mass spectrum might not contain sufficient information to localize the PTM within the identified peptide. However, even if this information is present, the reliable and automatic localization of phosphorylation sites within the proteolytic peptides is still problematic, especially if multiple potential phosphorylation sites are present.[9]

In many cases, the search algorithms used for protein identification are also applied for the identification and localization of phosphorylation sites. The most commonly used search engines, Mascot[13] and Sequest,[14] are excellent tools for protein identification but are neither optimized nor score explicitly for a proper PTM site assignment. These algorithms do not assign site-specific probability values and frequently fail to correctly determine the actual phosphorylation site within the peptide sequence. In addition, search engines do not necessarily test all possible permutations in order to limit the search space of the identification process. Decoy strategies[15] for estimating a false discovery rate (FDR) can be also applied for the identification of phosphopeptides; however, the FDR is not indicative for the false localization rate (FLR).

Recently, software tools have been developed for the automatic data interpretation of MS/MS spectra, allowing precise localization and scoring of phosphorylation sites.[16−21] Ascore[16] and PTM score[17] use peptide sequences identified in the database search and generate a list of all possible isoforms of a phosphopeptide via permutation. The experimental MS/MS data are used for calculation of site-specific scores and probabilities. Scores are calculated based on binominal probabilities or as in the case of SLoMo[19] approximated by a Poisson distribution. Following a different strategy, the Mascot delta ion score (MD score), is calculated by the difference between the two best scoring phosphopeptide isoforms identified from the MS/MS spectrum.[9,20] Originally introduced in a normalized variant for benchmarking Ascore,[16] this simple metric was found to be suitable for differentiation between correctly and incorrectly assigned phosphorylation sites.[20] In all cases, phosphorylation site localization is highly dependent on the presence of site-determining ions.

Given the biological importance of PTMs in general and protein phosphorylation in particular, universally applicable tools correctly predicting their localization are as important as developing dedicated analytical methods for their enrichment and mass spectrometric analysis.[5] Here we present a new probability-based site localization software called phosphoRS, assigning and calculating individual site probabilities for phosphorylated peptides. The program can be used in conjunction with all commonly used fragmentation methods and data sets with high or low mass accuracy. The unique feature of phosphoRS is the combination of a novel and dynamic peak depth determination procedure, the assignment of individual site probabilities, and the optimization and validation for all fragmentation techniques and mass accuracies.

Using synthetic phosphopeptides with known phosphorylation sites and sequences, the algorithm was optimized and its performance was compared to MD score and Ascore, indicating favorable figures of merit for phosphoRS. Still phosphoRS did not overestimate site assignments, as calculated probabilities were consistently below the experimental values for all evaluated data sets. In addition, we applied the program to a titania-enriched[22] tryptic phosphopeptide mixture from a HeLa cell lysate evaluating its practical usability. Using data generated by the pseudo-MS/MS/MS method, called multistage activation (MSA),[23] we demonstrated higher efficiency for site localization than Ascore or MD score. We also investigated the number of localized phosphorylation sites with MSA, ETD, and higher energy collisional dissociation (HCD).[24]

■ MATERIALS AND METHODS

Sample Preparation and Tryptic Digestion

Phosphorylated peptides were chemically synthesized by solid-phase Fmoc-chemistry (Novabiochem) using an ABI 433a peptide synthesizer and dried down in vacuo. A list of the used phosphopeptides is given in Supplemental Table S1 (Supporting Information). Quality control was performed using a MALDI-TOF on a Reflex III (Bruker Daltonics) with dihydroxy benzoic acid as a matrix. For LC−MS/MS analysis, phosphopeptides were first dissolved in aqueous solution containing 30% acetonitrile and diluted to a concentration of 0.2 pmol/μL in 0.1% trifluoracetic acid (TFA) and 25 mmol/L NaH_2PO_4.

HeLa proteins were isolated and proteolytically digested as described.[25] Hela Kyoto cells were treated with nocodazole for 16 h. The cells were harvested with a scraper and washed two times with phosphate buffered saline. The cell pellet was suspended in an equal amount of lysis buffer, and the cells were disrupted by pulling the cell suspension through a thin needle. Soluble proteins were reduced with dithiothreitol, alkylated with methylmethanethiosulfonate, and digested with trypsin.[26] HeLa peptide mixtures were enriched using titanium dioxide essentially as previously described.[27] In brief, phosphopeptides were isolated with Titansphere 5 μm beads (GL Science) from the proteolytic digest by dissolving the peptide mixture from 3.5 mg of HeLa protein extract in 500 μL of 80% acetonitrile, 3.5% TFA, saturated with phthalic acid. The solution was added to 1.25 mg of TiO_2 material in a spin column, incubated for 30 min, washed, and eluted with 150 μL of 0.3 M NH_4OH. After elution, the solution was acidified with 10% TFA.

LC−MS/MS Analysis

Nano-HPLC−MS/MS analysis of synthetic peptides was performed on an UltiMate 3000 Nano LC (Dionex) online coupled to an LT-Orbitrap XL (Thermo Fisher Scientific), using mixtures of 20 different peptides. Peptide separation was carried out on a C18 column (Acclaim PepMap C18, 25 cm × 75 μm × 2 μm, 100 Å, Dionex) using the following a ternary solvent system: A: 5% acetonitrile, 0.1% formic acid (FA); B: 30% acetonitrile, 0.08% FA and C: 80% acetonitrile, 0.08% FA and 10% trifluoroethanol (TFE). Synthetic phosphopeptides were analyzed with a short gradient from 100% A to 100% B (10 min), followed by a gradient to 20% B and 80% C (5 min) and a gradient to 100% C (7 min). The MS survey scan was performed in the FT cell recording a window between 300 and 2000 m/z. The resolution was set to 60 000 and the automatic gain control (AGC) was set to

1 000 000 ions with a maximal acquisition time of 500 ms. Minimum MS signal for triggering MS/MS was set to 500. One microscan was recorded for both MS and MS/MS acquisition. The lock mass option was enabled for using polydimethylcyclosiloxane ions (protonated $(Si(CH_3)_2O))_6$; m/z 445.120025) for internal recalibration of the mass spectra. MS/MS was performed in a data-dependent acquisition scheme selecting the four most prominent ions detected in the single MS scan. For each of the selected precursor ions four different MS/MS experiments employing CID, ETD, HCD, and ETD recorded in the Orbitrap (ETD-high) were performed. CID and ETD recorded in the ion trap were performed with a target value of 50 000 in the linear ion trap and a maximal injection time 200 ms. For CID we used a collision energy of 35%, a Q value of 0.25, and an activation time of 30 ms. CID was performed in its variant multistage activation (MSA) activating neutral losses of phosphoric acid (m/z 98) from doubly and triply charged precursor ions. ETD was performed with supplemental activation, fluoranthene served as the electron donor for ETD, and the reaction time was dependent on the precursor charge state (3+, 100 ms). These settings were applied for recording ETD-generated MS/MS spectra, the linear ion trap and the Orbitrap; however in the latter case the target value was set to 300 000 ions and the maximum injection time to 500 ms. HCD was performed with a target value of 300 000 in the Orbitrap, a resolution of 7500, a maximum injection time of 500 ms, and a collision energy of 35%.

The titania-enriched HeLa peptide mixture was analyzed using an UltiMate 3000 RSLCnano (Dionex) online coupled to an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) operated in positive ionization mode. Peptide mixtures were separated with a dual gradient system A: 2% acetonitrile, 0.1% FA and B: 80% acetonitrile, 0.08% FA, 10% TFE. Here we applied a linear gradient ramping from 100% A to 40% B in 480 min, followed by a short gradient to 90% B (5 min). MSA, HCD, and ETD were performed in independent experiments. The MS survey scan was performed in the FT cell recording a window between 350 and 2000 m/z, mass resolution of 60 000, a target value of 1 000 000 ions, and a maximal injection time of 250 ms. Minimum MS signal for triggering MS/MS was set to 500, and m/z values triggering MS/MS were put on an exclusion list for 180 s. One microscan was recorded and the lock mass option was enabled. MS/MS was performed in a data-dependent acquisition scheme selecting the 12 most prominent ions detected in the MS scan. MSA was performed with a target value of 3000 in the linear ion trap, a maximal injection time of 200 ms, collision energy of 35%, Q value of 0.25, and an activation time of 10 ms. HCD was performed with a target value of 50 000 in the Orbitrap, resolution of 7500, maximum acquisition time of 250 ms, and a collision energy of 35%. ETD was performed with supplemental activation enabled fragmentation and using the same target value and injection time as for MSA.

Data Analysis

Raw MS/MS spectra were interpreted with Proteome Discoverer (v.1.3.0.117, Thermo Fisher Scientific) applying Mascot (v.2.2.04, Matrix Science) for peptide identification. The data generated from both the synthetic phospho-peptide mixtures and the HeLa sample were searched against an in-house generated database containing all human entries from Swissprot database (release August 10, 2010), synthetic phospho-peptides as individual sequence entries, a list of common contaminants and the reversed sequences of all entries. This concatenated target/decoy database was created using SequenceReverser.exe (v.1.0.13.13, Max Planck Institute of Biochemistry).[28] For all experiments, a precursor ion mass tolerance of 5 ppm was applied allowing up to three missed cleavage sites for trypsin. The fragment ion mass tolerance was set to either 0.5 Da for the linear ion trap data sets or 0.02 Da for HCD and ETD-high data recorded in the Orbitrap. Phosphorylation of serine, threonine, and tyrosine was defined as variable modification. For the analysis of the HeLa sample, oxidation of methionine was specified as additional variable modification and methylthio-cysteine as fixed modification. For further analysis, all peptide-spectrum matches (PSMs) were required to be a Mascot rank one identification. We considered synthetic phosphopeptides with a length of at least seven amino acids and HeLa-generated peptides with a minimum length of eight amino acids. Peptide sequences were required to have at least one phosphorylation site. An individual Mascot ion score cutoff was chosen for each analysis, leading to a FDR of 1% at the PSM level calculated by the target/decoy strategy.[15] FDR-based filtering at the peptide or protein level was not applied, leading to an average FDR of 1.8% at the peptide level and 4.4% at the protein level for the three activation techniques.

Phosphorylation sites were localized with MD score, Ascore, and phosphoRS. To calculate MD scores, the respective Mascot ion scores were extracted from the corresponding Mascot dat-file using Mascot Parser (v.2.3.1.0, Matrix Science). For obtaining Ascore values, mzIdentML-files were created using Scaffold Q+ (v.3.00.08, Proteome Software Inc.), these files were analyzed by Scaffold PTM (v.1.0.3, Proteome Software Inc.), and finally, xls-files were extracted containing Ascore values for each phosphorylation site. Applying an in-house developed Microsoft Excel macro, Ascore values were assigned to the respective PSMs.

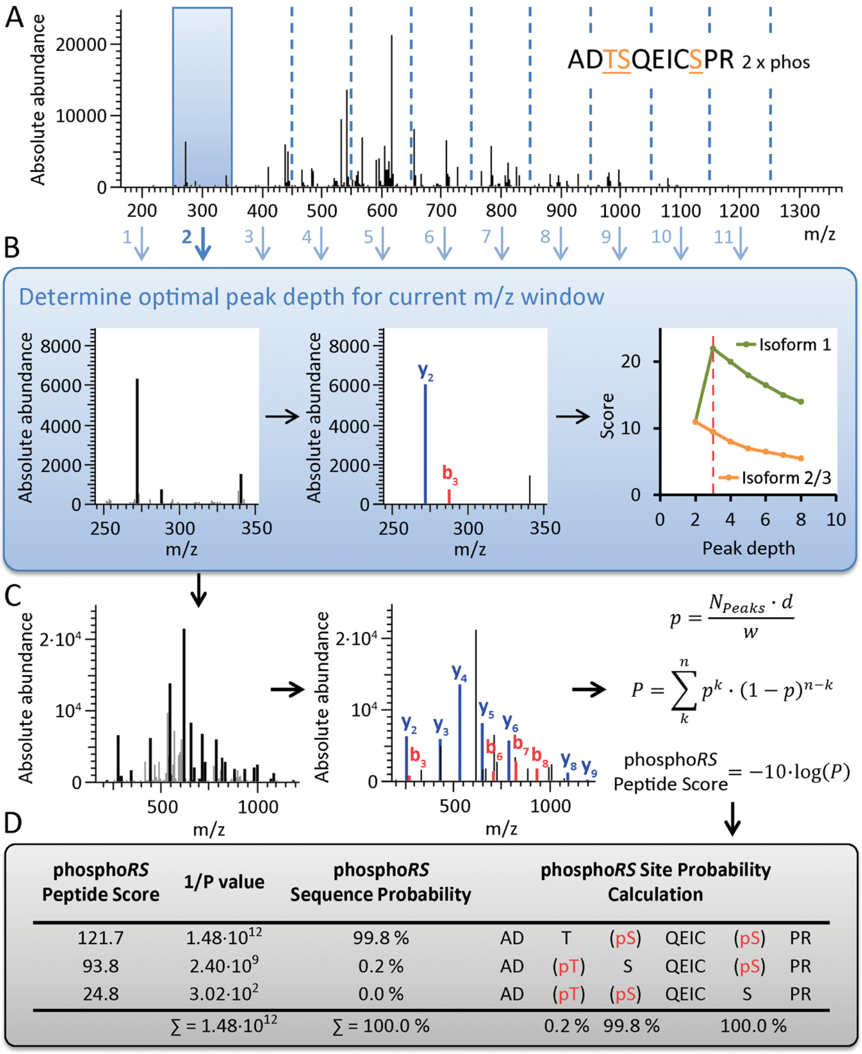
The described software, phosphoRS, is implemented in Java programming language (v.6.0.240, Oracle). For each positional isoform, the probability for the match between isoform and MS/MS spectrum being a random event is calculated using a cumulative binomial distribution. For a given PSM, individual site probabilities are calculated on the basis of probability-based phosphoRS peptide scores for the respective set of isoforms. The custom stand-alone version of phosphoRS used in this study extracts data from xml input files and writes the results as xml output files. Aiming at the facilitation of its ease of use, a preliminary version of the program is implemented in Proteome Discoverer software version 1.3 (Thermo Fisher Scientific) and will be implemented in the described form in a future version.

The data associated with this manuscript may be downloaded from ProteomeCommons.org Tranche using the following hash:

W0FetKH6001l0kNWKQ9C9r4+x0vTvjRgAg+McupehF/12TBlfy/gGCgPE46cGaZxDiEgbe5kSRRpg0JMcHY+JGCUyCgAAAAAAAAEAg==

The hash may be used to prove exactly what files were published as part of this manuscript's data set, and the hash may also be used to check that the data have not changed since publication.

In addition, the program is also available on our Web page (http://cores.imp.ac.at/protein-chemistry/download/).



phospho*RS* Peptide Score	1/P value	phospho*RS* Sequence Probability	phospho*RS* Site Probability Calculation					
121.7	$1.48{\cdot}10^{12}$	99.8 %	AD	T	(pS)	QEIC	(pS)	PR
93.8	$2.40{\cdot}10^{9}$	0.2 %	AD	(pT)	S	QEIC	(pS)	PR
24.8	$3.02{\cdot}10^{2}$	0.0 %	AD	(pT)	(pS)	QEIC	S	PR
	$\sum = 1.48{\cdot}10^{12}$	$\sum$ = 100.0 %		0.2 %	99.8 %		100.0 %	

Figure 1. Overview of the processing steps performed by phosphoRS. (A) The MS/MS spectrum is divided in windows of 100 m/z. (B) Dynamic peak extraction. For each m/z window the optimal peak depth is determined by calculating cumulative binomial probabilities for each isoform. (C) Cumulative binomial probabilities and peptide scores are calculated for each isoform, considering the optimal number of most intense peaks for each m/z window. (D) phosphoRS sequence and site probabilities are calculated, using the inverse probabilities for a random match obtained for each isoform.

■ RESULTS AND DISCUSSION

A number of different approaches to precise determination of phosphorylation sites exist;[16−20] however, these tools are optimized or even restricted to the analysis of a single fragmentation technique such as CID. Aiming at filling this gap, we developed a software tool for the automated localization of phosphorylation sites, integrating and extending existing approaches. The algorithm should enable localization of phosphorylation site in MS/MS spectra from all activation types, with optimized scoring for the various fragmentation methods and variable mass accuracy. The envisioned tool should work with peptide sequences identified with a specified FDR in conjunction with the respective raw data and lead to the proper identification of phosphorylation sites within the peptides. In addition, the algorithm should be capable of assigning individual probabilities to each potential phosphorylation site and should have comparable efficiencies for site localization as existing approaches such as Ascore[16] or the recently published MD score.[9,20]

Our approach is based on estimating the probability that the observed match between theoretical and acquired fragment ions is a random event. This probability P is calculated by applying the cumulative binominal distribution:

$$P = \sum_{k}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

In this formula n is the number of potentially existing theoretical fragment ions, k is the number of matches between theoretical and measured peaks, and p is the probability of randomly matching a single theoretical fragment ion. A peptide score for each isoform is then defined as 10 times the negative decadic logarithm of the random probability. Reflecting the quality of the match, this score can be also used for FDR-based filtering (data not shown).

In contrast to existing algorithms, we optimized phosphoRS for both high mass accuracy data such as HCD-generated fragment ion spectra as well as for low mass accuracy data acquired in ion traps. Consequently, we did not choose a uniform probability for

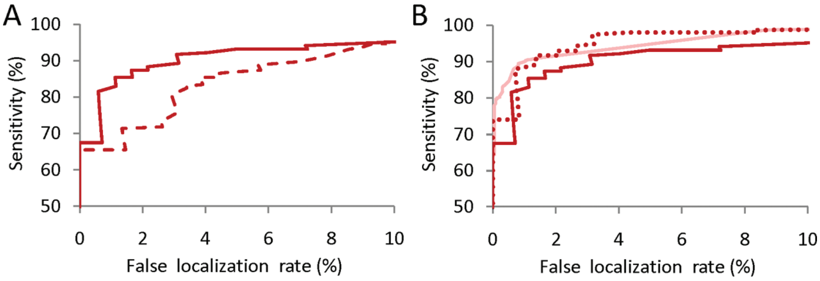


Figure 2. Optimization of peak extraction for phosphoRS. (A) The sensitivity of dynamic peak extraction (continuous line) compared to the conventional uniform approach (dotted line) is plotted against the FLR. (B) Sensitivity versus FLR curves are shown for PSMs (light red), unique sites (dotted red), and separately considering different charges states (red).

random matches between sequence-specific fragment ions with peaks in the spectrum. Instead, the probability p for a calculated fragment matching one of the experimental masses by chance is defined by

$$p = \frac{Nd}{w}$$

where N is the total number of extracted peaks, d is the specified fragment ion mass tolerance, and w is the full mass range of the MS/MS spectrum. Usually, approaches to peak extraction from peak lists divide the total m/z range of MS/MS spectra into windows of 100 m/z. Within each of these windows the same number of the most intense peaks is picked, representing the peak depth.[16,29,30] The peak depth can be either an a priori defined value or determined by an iterative process.[16] For current approaches to peak picking the number of extracted peaks is uniform for all windows over the entire m/z range. Predicted fragment ions for each isoform, representing the permutations of the phosphorylation sites, are then matched to the processed spectrum. Implementing a new algorithm allowed us to choose a novel and flexible approach, permitting a variable number of peaks extracted per 100 m/z window. The rationale behind this approach is that the different regions of MS/MS spectra can contain vastly different amounts of peaks (Figure 1A); therefore, the optimal peak depth might be different for the distinct m/z windows. We first calculated for each m/z window an optimal value for the peak depth by determining the cumulative binomial probability for each isoform, considering only peaks within the respective window (Figure 1B). The difference between the peptide score of the rank 1 and 2 isoforms is maximized. If not decisive, the difference to the rank 3 and, eventually, to the rank 4 isoform is considered. If an m/z window lacks any site determining ions, then the phosphoRS peptide score is maximized. Restriction of peak depths to a maximum of 8 peaks per 100 m/z window was necessary because gas-phase rearrangements of the phosphate group generated artifacts at low intensities leading to less sensitive or even wrong assignments.[31] Peaks are extracted from each 100 m/z window for the complete m/z range of the spectrum according to the individually determined optimal peak depths for calculating final probabilities (Figure 1C). In contrast to Ascore, all theoretical fragment ions are taken into account instead of scoring site determining ions only. If the sequence of amino acids assigned to the respective MS/MS spectrum is correct, then one of the putative isoforms has to be the true assignment.[17] Consequently, the sum of all sequence probabilities has to be equal to 100%. Reciprocal probabilities of observing the null hypothesis (random match) for each isoform were taken as relative weights to calculate the corresponding isoform confidence probabilities. These individual phosphoRS sequence probabilities were determined by normalization, dividing the reciprocal probabilities of the null hypothesis for each isoform by their sum (Figure 1D). Once sequence probabilities were calculated, individual site probabilities reflecting the probability for a specific phosphorylation site were estimated by summing up the sequence probabilities of those isoforms, in which the respective site is phosphorylated (Figure 1D).[17]

We evaluated our approach with MS/MS data sets of 179 chemically synthesized peptides with known phosphorylation sites. First, the MSA-generated data was used for plotting the sensitivity of localizing unique phosphorylation sites against the FLR, comparing dynamic peak extraction with the simpler approach of optimizing peak depth by iterating it over the complete spectrum. Using variable peak depths for each 100 m/z window led to a significantly higher number of correctly localized phosphorylation sites for a given FLR (Figure 2A). Applying a maximum mass deviation of 0.5 Da, the m/z values of the b- and y-ion series were considered, scoring singly and doubly charged fragment ions. If present, we counted the phosphorylation sites from doubly and triply charged precursor ions separately. We assumed this approach to be well suited for optimization of the algorithm because fragment ion spectra generated from different charge states of the same peptide have different characteristics. In addition, we also evaluated the respective performance of the algorithm at the level of unique sites and PSMs (Figure 2B). Analyzing the performance of phosphoRS by plotting unique sites considering different charges states separately proved to be more stringent than the other options (Figure 2B).

■ OPTIMIZATION OF SCORING

We then used LC−MS/MS data sets of synthetic phosphopeptide mixtures to determine the optimal parameters for scoring. The sensitivity of phosphorylation site localization was plotted versus the empirical FLR. We used MSA-, HCD-, or ETD-generated MS/MS spectra to identify the optimal conditions for each fragmentation technique. For MSA-generated data and using a maximal mass deviation of 0.5 Da, we tested various scoring options such as using b- and y-ions only, scoring for the additional neutral loss of phosphate moiety, accounting for doubly charged fragment ions and combinations of these settings. Confirming again the positive impact of dynamic peak extraction for all data sets (data not shown), we found that except for a few data points the variant considering singly and doubly charged ions but ignoring neutral loss peaks led to the best results for MSA generated spectra (Figure 3A).

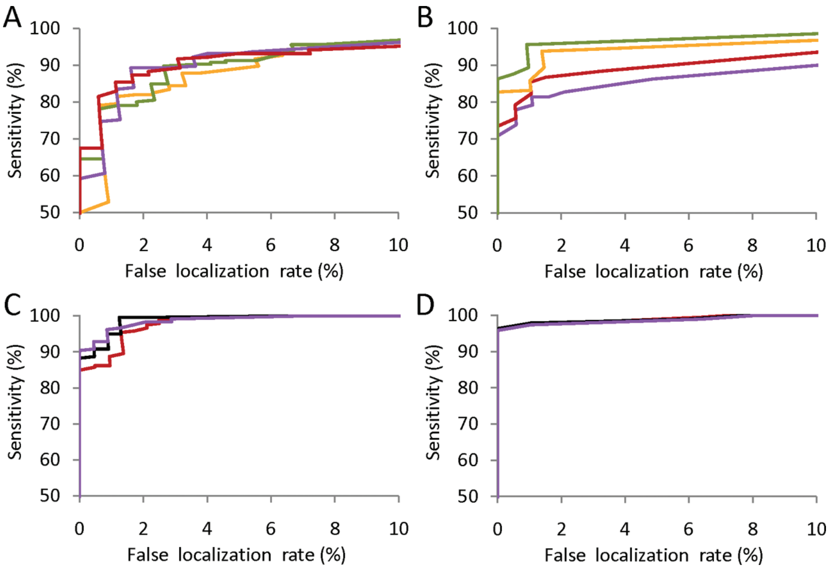


Figure 3. Optimization of the algorithm for the different fragmentation techniques, testing singly charged ions (violet), including neutral loss products (orange), singly and doubly charged ions (red), including neutral loss products in combination with scoring also for doubly charged ions (green). (A) For MSA-generated data considering singly and doubly charged ions but ignoring neutral loss peaks (red) preformed best. (B) For HCD-data, the best results were obtained when also scoring for the neutral loss of phosphoric acid (green). (C) For ETD-generated spectra c-and z-ions were scored instead of b- and y-ions. The inclusion of y-ions was also tested (black). The highest number of phosphorylation sites at a FLR of 1% was obtained when only scoring for singly charged c- and z-ions (violet). (D) For ETD-high all variants obtained similar results.

Next we used the same four parameter settings for the HCD-generated fragment ion spectra. It should be noted that in addition to a slightly different fragmentation regime and the possibility to detect fragment ions with low *m/z* values, HCD spectra were recorded with high mass accuracy. HCD data sets have proven to be superior for the analysis and localization of PTMs.[24] The data were analyzed with a maximum mass deviation of 20 mDa. Plotting again sensitivity versus FLR, we obtained a different result compared to the ion trap-generated MSA data (Figure 3B). The best results were obtained upon scoring in addition for the neutral loss products of the b- and y-ions, also considering doubly charged ions. Evidently, the high mass accuracy of the data marginalized random assignments to putative neutral loss derived product ions and the increased number of matching ions enhanced the sensitivity of the approach (Figure 3B).

Different scoring options were evaluated for ETD-generated MS/MS spectra. Naturally, we considered c- and z-ions but also tested the inclusion of y-ions and doubly charged fragment ions, although both less likely to occur in ETD spectra. We considered both z-radical ions and z-prime ions,[32] which are the even electron species after hydrogen transfer. Cleavage N-terminal to proline was not allowed.[10] Evaluating the various options, considering only singly charged ions of the c- and z-type resulted in the highest number of localized phosphorylation sites at a FLR of 1% (Figure 3C).

Finally we tried to optimize the algorithm for ETD generated fragment ions, recorded in the Orbitrap. In this case, almost all phosphorylation sites could be assigned at low FLRs, regardless of the applied scoring approach (Figure 3D).

■ RELATION BETWEEN PHOSPHORS SITE PROBABILITY AND FALSE LOCALIZATION RATE

Next, we investigated whether the experimentally observed precision of phosphorylation site localization and the calculated

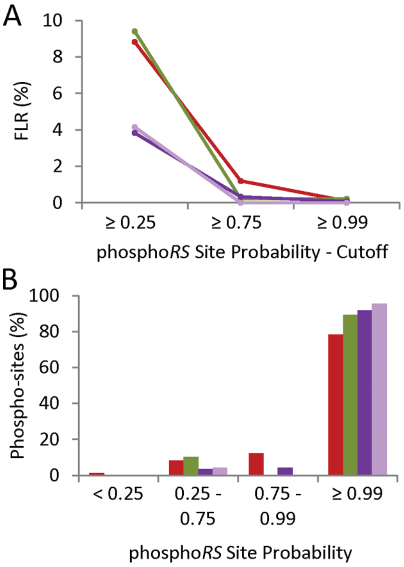


Figure 4. The data of the four different fragmentation techniques, MSA (red), HCD (green), ETD (violet), and ETD-high (light violet) were used to validate the results obtained with phosphoRS. (A) The FLRs obtained for the specified phosphoRS site probability cutoffs are shown. (B) Distribution of true phosphorylation sites on the level of PSMs for different phosphoRS site probability ranges and the different fragmentation techniques.

phosphoRS site probabilities were consistent with each other. If the statistical assumptions of phosphoRS are correct, then a site probability of 0.99 will correspond to a FLR of 1%. Plotting the relation between site probabilities and experimentally obtained FLRs for the four different MS/MS regimes studied (Figure 4A), we confirmed that the FLR of phospho-site localization was always better than or equal to the applied phosphoRS site probability cutoff. In all cases, the FLR converged to zero for phosphoRS site probabilities greater than 0.99. Plotting the fraction of true phosphorylation sites at the PSM level versus ranges of phosphoRS site probabilities (Figure 4B), the data confirmed that phosphoRS is able to efficiently localize phosphorylation sites with high sensitivity. As expected, HCD and ETD led to higher relative numbers of localized phosphorylation sites when compared to ion trap-generated MSA data sets.

■ COMPARISON OF PHOSPHORS WITH OTHER LOCALIZATION TOOLS

After optimization and validation of phosphoRS, we compared the algorithm to other approaches. For all PSMs with a FDR below 0.01, the respective values for their MD scores, Ascores, and phosphoRS site probabilities were obtained. At a FDR of 0.01, 2299 PSMs were identified from the MSA-generated data set. FLRs were determined on the basis of the known phosphorylation sites of the synthetic peptides. Plotting the sensitivity against the FLR, it became evident that phosphoRS performed slightly better on the training set at any given FLR (Figure 5). Again counting phosphorylation sites twice when identified in two different charge states and analyzing unique peptides instead of PSMs, phosphoRS localized 171 sites at a FLR of 0.01, 17, and 26 sites more than Ascore and MD score, respectively (Figure 5A). Analyzing the data at the level of PSMs, phosphoRS correctly localized 2454 sites, Ascore 2429 sites, and MD score 2302 sites. It should be noted that similar to FDR-based filtering of peptide identifications, FLR-based filtering at the level of PSMs is less stringent than for unique sites.

Using the HCD data set and a FLR of 1%, phosphoRS correctly localized 209 sites, 29 and 32 sites more than Ascore and MD score (Figure 5B). At the level of PSMs, phosphoRS could localize 1864 phosphorylation sites on 1701 PSMs, whereas Ascore localized 1741 sites and MD score 1882 sites.

Using ETD, we could localize 221 sites with the MD score and 231 with phosphoRS (Figure 5C). Ascore was not developed for ETD data sets. Consequently, Ascore did not perform equally well using ETD data sets, localizing only 159 sites at a FLR of 1%. Analyzing the results at the PSM level, phosphoRS localized 2734 sites in 2843 PSMs, Ascore 1741 and MD score 2754. Recording ETD-generated product ions in the Orbitrap, we could localize 188 phosphorylation sites with phosphoRS (Figure 5D). Using the MD score we could localize 176 sites, 12 sites less than with phosphoRS. Ascore did not perform equally well for ETD-high, localizing only 140 sites at a FLR of 1%. Interpreting the results obtained for the different localization tools, phosphoRS could robustly identify slightly more phosphorylation sites irrespectively of the used fragmentation technique.

■ APPLICATION TO A BIOLOGICAL SAMPLE

In order to test the practical applicability of the algorithm to a large data set, we applied phosphoRS to the analysis of titania-enriched phosphopeptides from a HeLa cell lysate. MSA, HCD, and ETD were used for fragmentation but owing to the sample complexity separate LC−MS/MS experiments were performed. Peptides and corresponding proteins were identified using Mascot applying a 1% FDR cutoff at the level of PSMs. Using MSA for fragmentation, 30 425 MS/MS spectra were recorded and 12 429 phosphorylation sites were identified in 10 641 PSMs by Mascot. Using phosphoRS, 7240 phosphorylation sites were localized using a phosphoRS site probability cutoff of 0.99. This is 58.3% of the sites identified by Mascot without applying any FLR-based filtering. We applied a cutoff value of 10 for MD score and 19 for Ascore, reflecting the published values required for a FLR of 1%.[16,20] Applying these cutoffs, we could localize 5418 phosphorylation sites applying MD score and 6348 phosphorylation sites with Ascore, significantly less than when using phosphoRS (Figure 6A). Upon analysis of the data set of chemically synthesized peptides, we obtained cutoff values of 15 for Ascore and 9.2 for MD score in order to reach a FLR of 1%. When we used these cutoffs, we obtained 7014 and 5730 phosphorylation sites, respectively, still fewer sites than with phosphoRS. A list of the

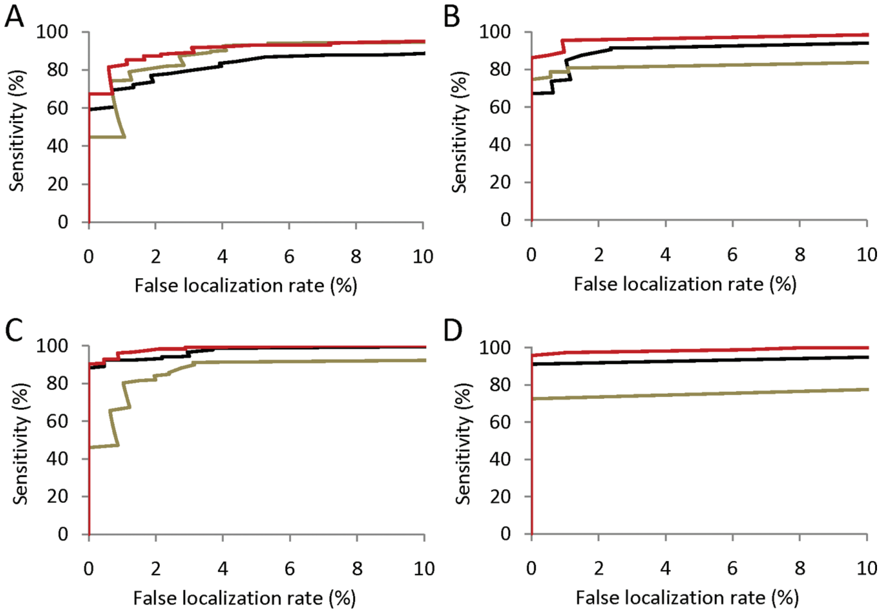


Figure 5. The performance of phosphoRS is compared to Ascore and MD score for the different fragmentation regimes: (A) MSA, (B) HCD, (C) ETD, and (D) ETD with high mass accuracy. Irrespective of the used fragmentation technique phosphoRS (red) performed at all FLR thresholds equally well or better than MD score (black) and Ascore (gold).

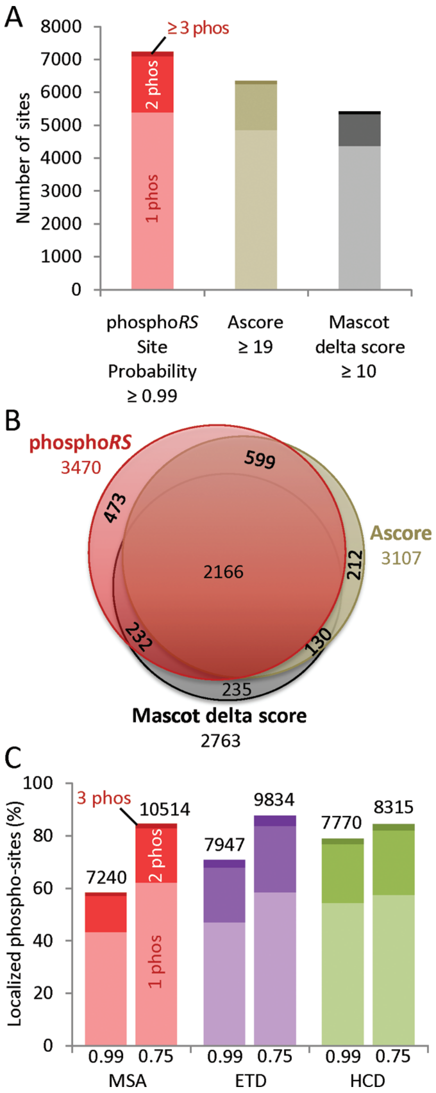


Figure 6. Analysis of a titania-enriched peptide mixture from a HeLa cell lysate. (A) The absolute numbers of phosphorylation sites at the PSM level localized with phosphoRS (red), Ascore (gold), and MD-score (black) from a MSA-generated data set are shown. (B) The numbers of nonredundant phosphorylation sites localized using the respective cutoff values for the different software tools are depicted. The overlap and the uniquely localized sites are visualized. (C) The percentage and absolute numbers of localized phosphorylation sites at the PSM level are shown for MSA (red), ETD (violet), and HCD (green) generated data sets applying phosphoRS site probability cutoffs of 0.99 and 0.75.

identified PSMs, the localized phosphorylation sites, and their respective scores generated by the three algorithms is provided in Supplementary Table S2 (Supporting Information).We then compared the total number of unique nonredundant phosphorylation sites identified with the three approaches. Here, we counted each phosphorylation site of a protein only once, irrespective of the number of corresponding phosphopeptides or PSMs. Using the three approaches, different numbers and sets of nonredundant phosphorylation sites were localized (Figure 6B). Interestingly, the common overlap of all three methods was only 2166 phosphorylation sites, about 50% of the unique 4047 sites identified with any of the three approaches. Each approach localized a significant fraction of additional sites. From the three methods, phosphoRS performed best by localizing 3470 unique sites, whereas Ascore localized 3107 and MD score 2763 phosphorylation sites.

Next we analyzed the number of phosphorylation sites localized with phosphoRS using replicates of the same sample comparing the three different fragmentation techniques, MSA, ETD, and HCD (Figure 6C). As already mentioned, we could localize 7240 phosphorylation sites using the MSA data set with phosphoRS applying a probability cutoff of 0.99. Using the ETD data set, we could confidently localize 7947 phosphorylation sites in 9163 PSMs. With the HCD-generated data we identified 7770 sites in 8140 PSMs. The efficiency of localizing identified phosphorylation sites was 70.8% and 78.9% for ETD and HCD, respectively.

Applying a phosphoRS site probability cutoff of 0.75, which should still lead to an acceptable FLR (Figure 4A), we could localize 10 514 phosphorylation sites, 84.6% of all phosphorylation sites identified from the MSA data set. Using the ETD and HCD data sets, we identified 9834 (88%) and 8315 (84%) phosphorylation sites (Figure 6C). A list of the identified PSMs, their predicted phosphorylation sites and phosphoRS site probabilities for the three activation techniques is provided in Supplementary Table S3 (Supporting Information). Interpreting the obtained results we conclude that although MSA-generated data led to more identified PSMs of phosphopeptides, the corresponding sites could be localized with less confidence compared to the HCD and ETD data sets.

■ CONCLUSIONS

We have shown that phosphoRS is capable of efficient and reliable localization of phosphorylation sites in peptides. Furthermore, our novel tool has been validated for all common fragmentation techniques. The combination of dynamic peak extraction and optimization of scoring parameters for the different fragmentation techniques resulted in higher numbers of localized phosphorylation sites compared to Ascore or MD score at a given FLR. We have demonstrated the practical utility of the software by determining phosphorylation sites in a titania-enriched phosphopeptide mixture generated from a HeLa cell lysate.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Supplementary Table S1: Sequences of synthetic phosphopeptides used. Supplementary Table S2: A list of the PSMs identified from the HeLa sample, their phosphorylation sites localized with the three localization tools, and the respective scores is provided. Supplementary Table S3: List of the PSMs identified from the HeLa sample, including their predicted phosphorylation sites and corresponding phosphoRS site probabilities for the three activation techniques used. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*(T.K.) Phone: 0043 1 79044 4283; fax: 0043 1 79044 110; e-mail: Thomas.Koecher@imp.ac.at. (K.M.) Karl.Mechtler@imp.ac.at.

Author Contributions

#These authors contributed equally to this work.

■ ACKNOWLEDGMENT

This work was funded by Boehringer Ingelheim, the Austrian Proteomics Platform (APP) within the Austrian GenomeResearch program (GEN-AU), the European Commission via the FP7 projects MeioSys and Prime XS, the Austrian Science Fund via the Special Research Program Chromosome Dynamics (SFB-F3402). The technical support of all other members of the Mechtler group, especially Mathias Madalinski, Susanne Opravil, Elisabeth Roitinger, Ines Steinmacher, and Johannes Fuchs, is gratefully acknowledged. We thank Bernard Delanghe and Torsten Ueckert for helpful discussions and critical reading of the manuscript.

■ REFERENCES

(1) Han, X. M.; Aslanian, A.; Yates, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **2008**, *12* (5), 483–490.

(2) Weston, A. D.; Hood, L. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *J. Proteome Res.* **2004**, *3* (2), 179–196.

(3) Nilsson, T.; Mann, M.; Aebersold, R.; Yates, J. R.; Bairoch, A.; Bergeron, J. J. M. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods* **2010**, *7* (9), 681–685.

(4) Kocher, T.; Superti-Furga, G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks. *Nat. Methods* **2007**, *4* (10), 807–815.

(5) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **2003**, *21* (3), 255–261.

(6) Zhao, Y. M.; Jensen, O. N. Modification-specific proteomics: Strategies for characterization of post-translational modifications using enrichment techniques. *Proteomics* **2009**, *9* (20), 4632–4641.

(7) Eyrich, B.; Sickmann, A.; Zahedi, R. P. Catch me if you can: Mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **2011**, *11* (4), 554–570.

(8) Harsha, H. C.; Pandey, A. Phosphoproteomics in cancer. *Mol. Oncol.* **2010**, *4* (6), 482–495.

(9) Boersema, P. J.; Mohammed, S.; Heck, A. J. R. Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom.* **2009**, *44* (6), 861–878.

(10) Zubarev, R. A. Electron-capture dissociation tandem mass spectrometry. *Curr. Opin. Biotechnol.* **2004**, *15* (1), 12–16.

(11) Stensballe, A.; Jensen, O. N.; Olsen, J. V.; Haselmann, K. F.; Zubarev, R. A. Electron capture dissociation of singly and multiply phosphorylated peptides. *Rapid Commun. Mass Spectrom.* **2000**, *14* (19), 1793–1800.

(12) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (26), 9528–9533.

(13) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(14) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(15) 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), 207–214.

(16) Beausoleil, S. A.; Villen, J.; Gerber, S. A.; Rush, J.; Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **2006**, *24* (10), 1285–1292.

(17) Olsen, J. V.; Blagoev, B.; Gnad, F.; Macek, B.; Kumar, C.; Mortensen, P.; Mann, M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **2006**, *127* (3), 635–648.

(18) Lu, B. W.; Ruse, C.; Xu, T.; Park, S. K.; Yates, J. Automatic validation of phosphopeptide identifications from tandem mass spectra. *Anal. Chem.* **2007**, *79* (4), 1301–1310.

(19) Bailey, C. M.; Sweet, S. M. M.; Cunningham, D. L.; Zeller, M.; Heath, J. K.; Cooper, H. J. SLoMo: Automated site localization of modifications from ETD/ECD mass spectra. *J. Proteome Res.* **2009**, *8* (4), 1965–1971.

(20) Savitski, M. M.; Lemeer, S.; Boesche, M.; Lang, M.; Mathieson, T.; Bantscheff, M.; Kuster, B. Confident phosphorylation site localization using the Mascot delta score. *Mol. Cell. Proteomics* **2011**, *10* (2), 12.

(21) Ruttenberg, B. E.; Pisitkun, T.; Knepper, M. A.; Hoffert, J. D. PhosphoScore: An open-source phosphorylation site assignment tool for MSn data. *J. Proteome Res.* **2008**, *7* (7), 3054–3059.

(22) Pinkse, M. W. H.; Uitto, P. M.; Hilhorst, M. J.; Ooms, B.; Heck, A. J. R. Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-nanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal. Chem.* **2004**, *76* (14), 3935–3943.

(23) Schroeder, M. J.; Shabanowitz, J.; Schwartz, J. C.; Hunt, D. F.; Coon, J. J. A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal. Chem.* **2004**, *76* (13), 3590–3598.

(24) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709–712.

(25) Poser, I.; Sarov, M.; Hutchins, J. R. A.; Heriche, J. K.; Toyoda, Y.; Pozniakovsky, A.; Weigl, D.; Nitzsche, A.; Hegemann, B.; Bird, A. W.; Pelletier, L.; Kittler, R.; Hua, S.; Naumann, R.; Augsburg, M.; Sykora, M. M.; Hofemeister, H.; Zhang, Y. M.; Nasmyth, K.; White, K. P.; Dietzel, S.; Mechtler, K.; Durbin, R.; Stewart, A. F.; Peters, J. M.; Buchholz, F.; Hyman, A. A. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **2008**, *5* (5), 409–415.

(26) Gregan, J.; Riedel, C. G.; Petronczki, M.; Cipak, L.; Rumpf, C.; Poser, I.; Buchholz, F.; Mechtler, K.; Nasmyth, K. Tandem affinity purification of functional TAP-tagged proteins from human cells. *Nat. Protoc.* **2007**, *2* (5), 1145–1151.

(27) Bodenmiller, B.; Mueller, L. N.; Mueller, M.; Domon, B.; Aebersold, R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat. Methods* **2007**, *4* (3), 231–237.

(28) 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), 1367–1372.

(29) Beer, I.; Barnea, E.; Ziv, T.; Admon, A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **2004**, *4* (4), 950–960.

(30) 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), 13417–13422.

(31) Palumbo, A. M.; Reid, G. E. Evaluation of gas-phase rearrangement and competing fragmentation reactions on protein phosphorylation site assignment using collision induced dissociation-MS/MS and MS3. *Anal. Chem.* **2008**, *80* (24), 9735–9747.

(32) Kjeldsen, F.; Haselmann, K. F.; Budnik, B. A.; Jensen, F.; Zubarev, R. A. Dissociative capture of hot (3−13 eV) electrons by polypeptide polycations: an efficient process accompanied by secondary fragmentation. *Chem. Phys. Lett.* **2002**, *356* (3−4), 201–206.