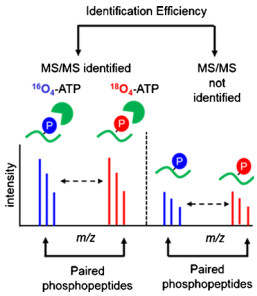
J. Am. Soc. Mass Spectrom. (2017) 28:1127–1135
DOI: 10.1007/s13361-017-1603-5

FOCUS: HONORING R. G. COOKS' ELECTION TO THE NATIONAL ACADEMY OF SCIENCES: RESEARCH ARTICLE

Estimating the Efficiency of Phosphopeptide Identification by Tandem Mass Spectrometry

Chuan-Chih Hsu,[1] Liang Xue,[1,2] Justine V. Arrington,[1] Pengcheng Wang,[3] Juan Sebastian Paez Paez,[1] Yuan Zhou,[1] Jian-Kang Zhu,[3] W. Andy Tao[1]

[1]Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA
[2]*Present Address:* Celgene Corporation, Cambridge, MA, USA
[3]Department of Horticulture and Landscape, Purdue University, West Lafayette, IN 47907, USA



Abstract. Mass spectrometry has played a significant role in the identification of unknown phosphoproteins and sites of phosphorylation in biological samples. Analyses of protein phosphorylation, particularly large scale phosphoproteomic experiments, have recently been enhanced by efficient enrichment, fast and accurate instrumentation, and better software, but challenges remain because of the low stoichiometry of phosphorylation and poor phosphopeptide ionization efficiency and fragmentation due to neutral loss. Phosphoproteomics has become an important dimension in systems biology studies, and it is essential to have efficient analytical tools to cover a broad range of signaling events. To evaluate current mass spectrometric performance, we present here a novel method to estimate the efficiency of phosphopeptide identification by tandem mass spectrometry. Phosphopeptides were directly isolated from whole plant cell extracts, dephosphorylated, and then incubated with one of three purified kinases—casein kinase II, mitogen-activated protein kinase 6, and SNF-related protein kinase 2.6—along with $^{16}O_4$- and $^{18}O_4$-ATP separately for in vitro kinase reactions. Phosphopeptides were enriched and analyzed by LC-MS. The phosphopeptide identification rate was estimated by comparing phosphopeptides identified by tandem mass spectrometry with phosphopeptide pairs generated by stable isotope labeled kinase reactions. Overall, we found that current high speed and high accuracy mass spectrometers can only identify 20%–40% of total phosphopeptides primarily due to relatively poor fragmentation, additional modifications, and low abundance, highlighting the urgent need for continuous efforts to improve phosphopeptide identification efficiency.

Keywords: Protein phosphorylation, Proteomics, Tandem mass spectrometry

Received: 3 October 2016/Revised: 8 January 2017/Accepted: 9 January 2017/Published Online: 10 March 2017

Introduction

It is estimated that there are over 500 protein kinases in mammals and over 1000 in plants, and an estimated half or more of all proteins are phosphorylated at certain points during their life span [1, 2]. Reversible phosphorylation of proteins is involved in the regulation of most, if not all, cellular processes [3]. Abnormal phosphorylation has been implicated in a number of diseases, most notably cancer [4]. The accurate determination of sites of phosphorylation and dynamics of this modification in response to extracellular stimulation is important for elucidating complex disease mechanisms and global regulatory networks. Development of methods for analyzing phosphorylated proteins, therefore, has been an active field of research in the signaling, mass spectrometry, and proteomics communities.

Advances in mass spectrometry (MS)-based proteomics have driven increasing efforts to identify reliable approaches for the large scale analysis of phosphoproteins (phosphoproteomics) that include both the identification of protein phosphorylation sites and the quantification of changes in phosphorylation at individual sites [5]. Phosphorylation is often a low stoichiometric event [6]. To

Dedicated to Professor R. Graham Cooks' achievements and his life-long devotion to nothing but mass spectrometry

Electronic supplementary material The online version of this article (doi:10.1007/s13361-017-1603-5) contains supplementary material, which is available to authorized users.

Correspondence to: W. Tao; *e-mail:* watao@purdue.edu

identify specific sites of phosphorylation, it is essential to have an efficient strategy for the selective enrichment of actual phosphopeptides. Current approaches include immobilized metal ion or metal oxide affinity chromatography (IMAC and MOAC) [7–10] and polymer-based metal ion affinity capture (polyMAC) for general phosphopeptide enrichment [11] and anti-phosphotyrosine antibodies for the isolation of phosphotyrosine-containing peptides [12]. High throughput analysis of phosphorylation using directed enrichment methods followed by MS has become a standard approach for phosphoprotein detection.

While phosphoproteomics has increasingly become an important, typically more informative, dimension in omics studies, its challenges have persisted [13]. The primary challenge in examining protein phosphorylation is its low stoichiometry. Phosphorylated proteins, especially those involved in signaling, are often expressed in relatively low amounts in a cell, and few of these proteins exist in a phosphorylated form at any one time. Furthermore, phosphopeptides have low ionization efficiency due to their negatively charged phosphate groups [14], and they can exhibit poor fragmentation in tandem mass spectra because of neutral loss of the phosphate groups [15]. Finally, informatics approaches for processing the results of mass spectrometry data for phosphopeptides are not yet mature [16].

There have been a number of attempts to improve phosphopeptide identification efficiency, particularly by alternative activation methods [17]. Faster and more accurate LC-MS systems have also made a significant contribution toward enhancing the coverage of the phosphoproteome, but it is not still clear what percentage of the phosphopeptide population is routinely identified by tandem mass spectrometry (MS/MS) and whether current phosphoproteomic strategies provide true representations of the phosphoproteome for systems biology analyses. Due to the dynamic nature of phosphorylation, low coverage of the phosphoproteome at certain cellular states might lead to biased or even incorrect conclusions. Here, we present a novel strategy to estimate the efficiency of phosphopeptide identification by tandem mass spectrometry. Instead of using a large pool of synthetic phosphopeptides, which is costly to generate [18] and still incomprehensive, we created a phosphopeptide pool directly from whole cell extracts. To generate phosphopeptides with distinctively recognizable features in the mass spectra, we introduced in vitro kinase reactions with $^{16}O_4$- and $^{18}O_4$-ATP to generate phosphopeptide pairs with similar intensity that are separated by 6 Da on mass spectra. Previous literature and our own data indicate that the ^{18}O atoms on the γ-phosphoryl group do not exchange with water during kinase reactions [19]. Thus, the efficiency of phosphopeptide identification can be estimated by comparing the phosphopeptides identified by MS/MS with the total number of phosphopeptide pairs that demonstrate the distinctive mass shift.

Experimental Methods

Plant Materials and Growth

The seeds of Col-0 wild type *Arabidopsis* were geminated on half-strength Murashige and Skoog (MS) medium (1% sucrose with 0.6% phytogel). Five d after germination, seedlings were transferred into 40 mL half-strength MS liquid medium with 1% sucrose at 22 °C in continuous light on a rotary shaker set at 100 rpm. For osmotic stress treatment, 12-d-old seedlings were transferred into fresh medium containing 800 mM Mannitol for 30 min. In parallel, the seedlings transferred into fresh medium were used as the control.

Protein Extraction and Digestion

Plant tissues were ground with mortar and pestle in liquid nitrogen, and the ground tissues were lysed in 6 M guanidine hydrochloride containing 100 mM Tris-HCl (pH = 8.5) with EDTA-free protease inhibitor cocktail (Roche, Madison, WI, USA) and phosphatase inhibitor cocktail (Sigma-Aldrich, St. Louis, MO, USA). Proteins were reduced and alkylated with 10 mM tris-(2-carboxyethyl)phosphine (TECP) and 40 mM chloroacetamide (CAA) at 95 °C for 5 min. Alkylated proteins were subjected to methanol-chloroform precipitation, and precipitated protein pellets were solubilized in 8 M urea containing 50 mM triethylammonium bicarbonate (TEAB). Protein amount was quantified by BCA assay (Thermo Fisher Scientific, Rockford, IL, USA). Protein extracts were diluted to 4 M urea and digested with Lys-C (Wako, Osaka, Japan) in a 1:100 (w/w) enzyme-to-protein ratio overnight at 37 °C. Digests were acidified with 10% trifluoroacetic acid (TFA) to a pH ~2 and desalted using a 100 mg Sep-Pak C18 column (Waters, Milford, MA, USA).

Stable Isotope Labeled In Vitro Kinase Reaction

The in vitro kinase reaction was performed based on previous reports [20, 21] with some modifications. The Lys-C digested peptides (200 μg) were treated with a thermosensitive alkaline phosphatase (TSAP) (Roche) in a 1:100 (w/w) enzyme-to-peptide ratio at 37 °C overnight for dephosphorylation, and the dephosphorylated peptides were desalted using Sep-Pak C18 column. The desalted peptides were resuspended in kinase reaction buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, and 1 mM DTT, pH 7.5) with either 1 mM ^{16}O-ATP or γ-[$^{18}O_4$]-ATP (Cambridge Isotope Laboratories, MA, USA). The suspended peptides were incubated with the recombinant SNF-related protein kinase 2.6 (SnRK2.6), mitogen-activated protein kinase 6 (MPK6), or casein kinase II (CK2) (500 ng) at 30 °C overnight. The kinase reaction was quenched by acidifying with 10% TFA to a final concentration of 1%, and the peptides were desalted by Sep-Pak C18 column. The light and heavy phosphopeptides were mixed and further digested by trypsin at 37 °C for 6 h. Tryptic phosphopeptides were desalted by Sep-Pak column, and then were enriched by PolyMAC-Ti

reagent, and the eluates were dried in a SpeedVac for LC-MS/MS analysis.

PolyMAC Enrichment

Phosphopeptide enrichment was performed according to the reported PolyMAC-Ti protocol [11] with some modifications. Tryptic peptides (200 µg) were resuspended in 100 µL of loading buffer [80% acetonitrile (ACN) with 1% TFA] and incubated with 25 µL of the PolyMAC-Ti reagent for 20 min. A magnetic rack was used to collect the magnetic beads to the sides of the tubes, and the flow-through was discarded. The magnetic beads were washed with 200 µL of washing buffer 1 (80% ACN, 0.2% TFA with 25 mM glycolic acid) for 5 min, and washing buffer 2 (80% ACN in water) for 30 s. Phosphopeptides were then eluted with 200 µL of 400 mM NH_4OH with 50% ACN and dried in a SpeedVac.

LC-MS/MS Analysis

The phosphopeptides were dissolved in 5 µL of 0.3% formic acid (FA) with 3% ACN and injected into an Easy-nLC 1000 (Thermo Fisher Scientific). Peptides were separated on a 45 cm in-house packed column (360 µm o.d. × 75 µm i.d.) containing C18 resin (2.2 µm, 100Å; Michrom Bioresources, Auburn, CA) with a 30 cm column heater (Analytical Sales and Services, Pompton Plains, NJ) set at 50 °C. The mobile phase buffer consisted of 0.1% FA in ultra-pure water (buffer A) with an eluting buffer of 0.1% FA in 80% ACN (buffer B) run over a linear 60 min gradient of 5%–30% buffer B at a flow rate of 250 nL/min. The Easy-nLC 1000 was coupled online with a LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific). The mass spectrometer was operated in the data-dependent mode in which a full MS scan (from *m*/*z* 350–1500 with the resolution of 30,000 at *m*/*z* 400) was followed by the five most intense ions being subjected to collision-induced dissociation (CID) fragmentation. CID fragmentation was performed and acquired in the linear ion trap (normalized collision energy (NCE) 30%, AGC 3e4, max injection time 100 ms, isolation window 3 *m*/*z*, and dynamic exclusion 60 s).

Data Processing

The raw files were searched directly against the *Arabidopsis thaliana* database (TAIR10) with no redundant entries using MaxQuant software (ver. 1.5.4.1) [22] with the Andromeda search engine. Initial precursor mass tolerance was set at 20 ppm, the final tolerance was set at 6 ppm, and the ITMS MS/MS tolerance was set at 0.6 Da. Search criteria included a static carbamidomethylation of cysteines (+57.0214 Da) and variable modifications of (1) oxidation (+15.9949 Da) on methionine residues, (2) acetylation (+42.011 Da) at the N-terminus of proteins, (3) phosphorylation (+79.996 Da), and (4) heavy phosphorylation (+85.979 Da) on serine, threonine, or tyrosine residues. The match between runs function was enabled with 1.0 min match time window. The searches were performed with trypsin digestion and allowed a maximum of two missed cleavages on the peptides analyzed from the sequence database. The false discovery rates for proteins, peptides, and phosphosites were set at 0.01. The minimum peptide length was six amino acids, and a minimum Andromeda score cut-off was set at 40 for modified peptides. A site localization probability of 0.75 was used as the cut-off for localization of phosphorylation sites. The MS/MS spectra can be viewed through the MaxQuant viewer. For the ProteomeDiscoverer searches, the raw files were searched directly against the same *Arabidopsis thaliana* database (TAIR10) with no redundant entries using the SEQUEST HT algorithm in Proteome Discoverer ver. 2.1 (Thermo Fisher Scientific). Peptide precursor mass tolerance was set at 10 ppm, and MS/MS tolerance was set at 0.6 Da. Search criteria included a static carbamidomethylation of cysteines (+57.0214 Da) and variable modifications of (1) oxidation (+15.9949 Da) on methionine residues, (2) acetylation (+42.011 Da) on protein N-termini, (3) phosphorylation (+79.996 Da), and (4) heavy phosphorylation (+85.979 Da) on serine, threonine, or tyrosine residues. Searches were performed with full tryptic digestion and allowed a maximum of two missed cleavages on the peptides analyzed from the sequence database. Relaxed and strict false discovery rates (FDR) were set to 0.05 and 0.01, respectively. All localized phosphorylation sites were submitted to Motif-X [23] to determine kinase phosphorylation motifs with the TAIR10 database as the background. The significance was set at 0.000001, the width was set at 13, and the number of occurrences was set at 20. The light and heavy phosphopeptide and peak pairs were identified through the LAXIC algorithm [24]. All of the light and heavy phosphopeptide and peak pairs are listed in the Supplementary Tables, and the raw data and analysis files for the proteomic analyses have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the jPOST partner repository (http://jpost.org) [25] with the data set identifier PXD005079.

Results and Discussion

The strategy to estimate the efficiency of phosphopeptide identification was devised based on our previous phosphoproteomic studies on kinase substrates [20, 21]. The general strategy to estimate the efficiency of phosphopeptide identification by MS/MS is illustrated in Figure 1. To generate a comprehensive pool of phosphopeptides, proteins were extracted from whole cell lysates such as whole cell extracts from plants. After digestion with Lys-C to generate peptides, the peptides were incubated with a thermosensitive alkaline phosphatase overnight to remove phosphate groups on the peptides and to generate a pool of peptide candidates for the in vitro kinase reactions. We chose three kinases, casein kinase 2 (CK2), mitogen-activated protein kinase 6 (MPK6), and SNF-related protein kinase 2.6 (SnRK2.6), for their known high specificity toward acidic, proline-directed, and basic motifs, respectively. These three kinases also have high enzymatic activity in vitro and can potentially phosphorylate hundreds of substrates. The most important feature of this strategy

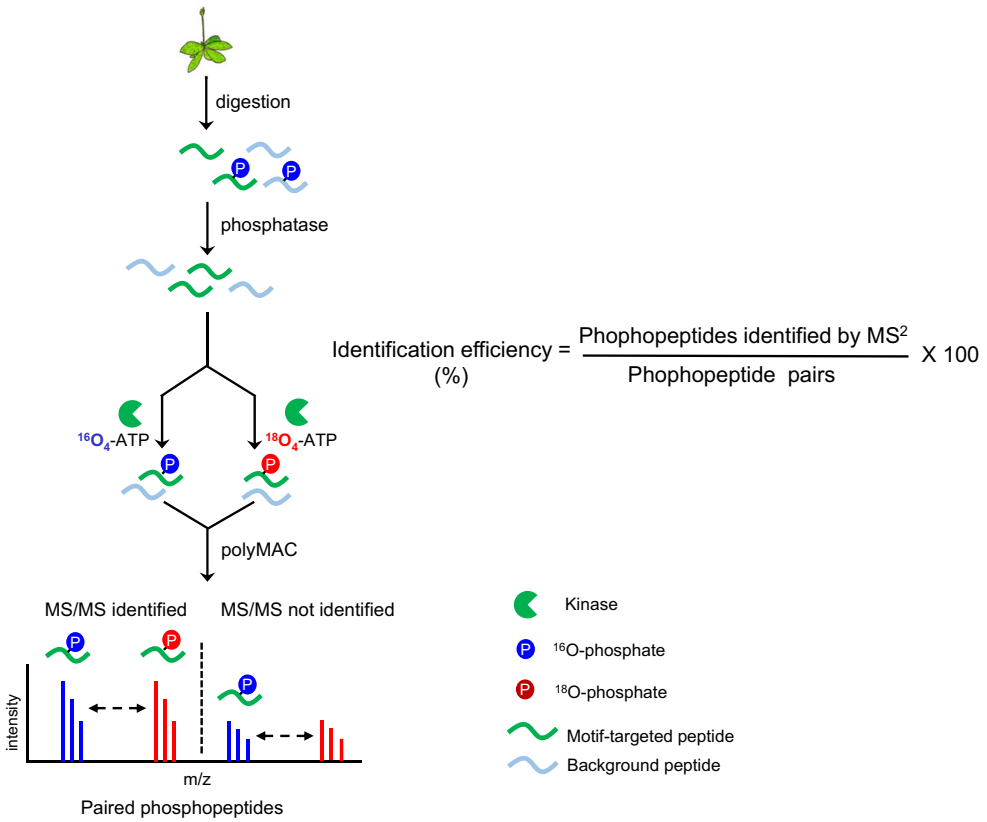


Figure 1. The workflow to estimate efficiency of phosphopeptide identification by tandem mass spectrometry. The ratio of phosphopeptides identified by MS/MS over total paired phosphopeptides identified in single-stage MS represents the phosphopeptide identification efficiency. See main text for details

is the generation of a large number of phosphopeptides that are invisible in MS/MS but have distinctive characteristics that can be unambiguously recognized even if their sequences are unknown. In doing so, we devised in vitro kinase reactions with γ-$^{16}O_4$- and γ-$^{18}O_4$-ATP in parallel. The kinase reaction transfers one or more γ-phosphate groups from ATP to substrate peptides, thus generating light- and heavy- phosphorylated peptides with similar intensities, assuming the same kinase has similar reactivity with γ-$^{16}O_4$- or γ-$^{18}O_4$-ATP. After the kinase reactions, samples were pooled together, and phosphopeptides were enriched with PolyMAC before LC-MS analyses. Data were searched against the appropriate protein database for sequence information. In-house LAXIC software was used to pick and quantify peptide pairs with the required characteristic features [24]. Finally, the efficiency of phosphopeptide identification can be estimated by comparing the phosphopeptides identified by MS/MS with the total number of phosphopeptide pairs.

We generated whole cell lysates from *Arabidopsis* seedlings in this study. The plant has over 1000 encoded kinases, and whole cell lysates likely contain tens of thousands of phosphorylation sites [2, 26]. Two plant kinases, MAPK6 and SnRK2.6, were recombinantly expressed and purified. Along with human recombinant CK2, the three kinases were incubated with *Arabidopsis* lysate and γ-$^{16}O_4$- or γ-$^{18}O_4$-ATP separately. Each kinase reaction can generate hundreds of phosphopeptides, which is sufficient for this study but still well within the capacity of a typical high resolution LC-MS system today. This strategy also minimizes the instrumentation factor. Although it is conceivable that the speed of mass spectrometers affects the identification rate of phosphopeptides, this factor would be minor with the current approach.

As anticipated, we observed multiple peak doublets in the mass spectra. With a high speed mass spectrometer such as the Orbitrap Velos, most of these precursor ions were selected for MS/MS. Figure 2 illustrates examples of three peptides phosphorylated by CK2, MPK6, or SnRK2.6, respectively. The extracted ion chromatogram (XIC) and MS/MS spectra of the paired light/heavy NUP50 phosphopeptide ApSDIEEGDEVDSK are shown in Figure 2a. The peptide was phosphorylated by CK2, and the doubly charged, heavy phosphate-labeled phosphopeptide (red line) has a 3.00 *m/z* shift from the doubly charged, light phosphate-labeled phosphopeptide (blue line). No significant retention time shift was observed as a result of heavy phosphate labeling. The MS/MS spectrum shows the identification of paired light/heavy phosphopeptide with the expected acidic phosphorylation motif. Similarly, the XIC and MS/MS spectra of the paired light/heavy CAD5 phosphopeptide DPSGILpSPYTYTLR phosphorylated by MPK6 are shown in Figure 2b. The phosphopeptide sequence has the characteristic proline-directed phosphorylation motif;

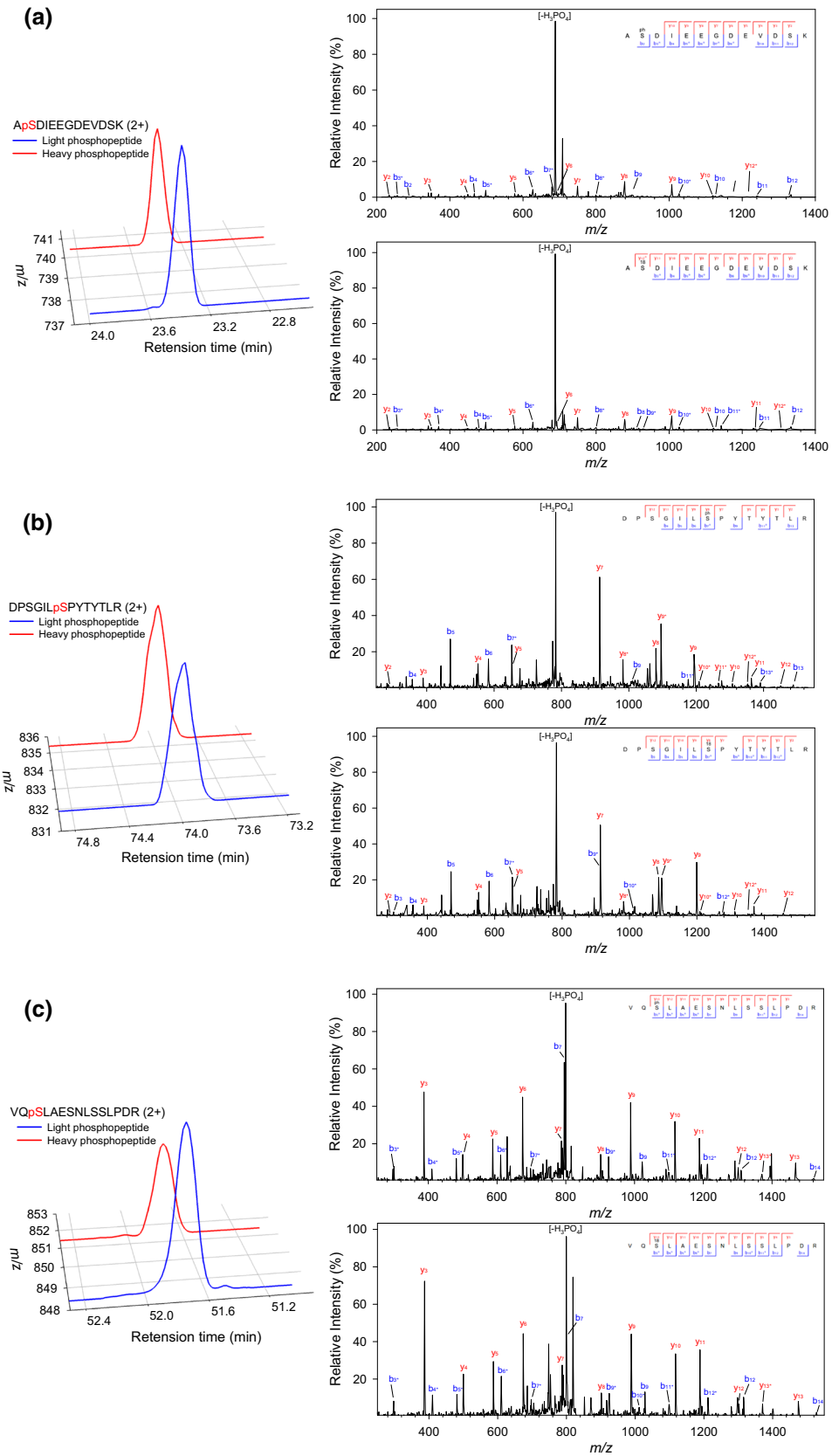


Figure 2. Selected examples of extracted ion chromatograms and MS/MS spectra of motif-targeted paired phosphopeptides from three in vitro kinase reactions. (**a**) NUP50 phosphopeptide ApSDIEEGDEVDSK phosphorylated by CK2. (**b**) CAD5 phosphopeptide DPSGILpSPYTYTLR phosphorylated by MPK6. (**c**) AT5G05600 phosphopeptide VQpSLAESNLSSLPDR phosphorylated by SnRK2.6

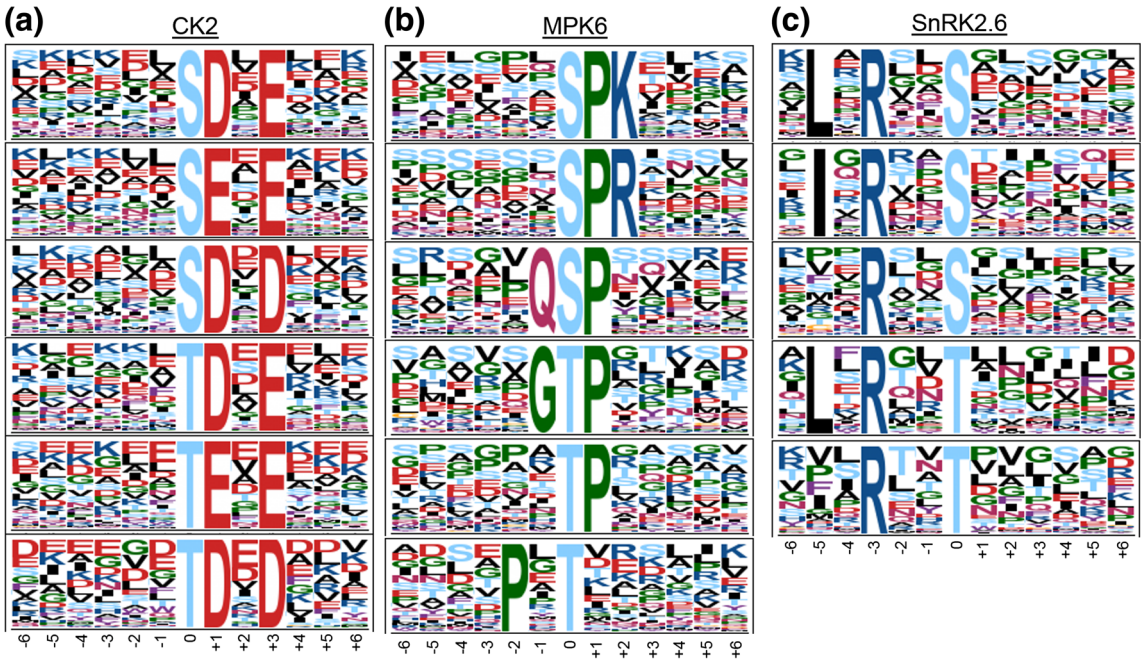


Figure 3. Motif analysis of identified light/heavy phosphopeptides. The phosphorylation motifs of identified light/heavy phosphopeptides were extracted by Motif-X from (**a**) CK2, (**b**) MPK6, and (**c**) SnRK2.6 kinase reactions

Figure 2c shows the XIC and MS/MS spectra of the paired light/heavy AT5G05600 phosphopeptide VQpSLAESNLSSLPDR phosphorylated by SnRK2.6. The MS/MS spectra show the identification of paired light/heavy phosphopeptide with the basic phosphorylation motif [-I-x-R-x-x-pS-]. Although the examples provided in Figure 2 show that the light- and heavy-labeled phosphopeptide pairs have similar intensity and were sequenced by MS/MS, not all phosphopeptide pairs have similar intensity. In many cases, the light-labeled phosphopeptide has higher intensity than its heavy-labeled counterpart (see Supplementary Tables S1–S3). The exact cause is not clear. While we expected similar kinase reactivity with light or heavy ATP, it is possible that γ-$^{18}O_4$-ATP has a bigger size, which might prevent it from fitting inside the ATP binding pocket perfectly. We will investigate this phenomenon in a separate study. All doublet peaks with appropriate mass difference and similar intensity were deconvoluted and counted as phosphopeptides, no matter whether they were sequenced by MS/MS or not.

Six in vitro kinase reactions, the result of three kinases with light and heavy ATP separately, generated thousands of phosphopeptides in vitro. We searched MS/MS spectra against the *Arabidopsis* proteome database. In total, 1498, 1472, and 1837 doublet phosphopeptides were identified by CK2, MPK6, and SnRK2.6, respectively. The data indicates high, specific in vitro kinase activity for all three kinases. Motif analyses of the identified phosphopeptides resulted in the acidic motif for the CK2 kinase reaction, [-(pS/pT)-(D/E)-x-(D/E)-], the proline-directed phosphorylation motif for MPK6, [-(pS/pT)-P-], and the basic phosphorylation motif for SnRK2.6, [-(I/L)-x-R-x-x-(pS/pT)-] (Figure 3). The results from these motif analyses are highly consistent with previous literature reports and known substrate specificity of the three kinases [27–29].

The advancement of high speed and high accuracy mass spectrometers, along with ultrahigh performance liquid chromatography (UHPLC), has greatly improved the coverage of phosphoproteomes. However, considering the high dynamics of protein phosphorylation, it is not clear whether current LC-MS technology can provide sufficient coverage of most phosphoproteomes. In our phosphopeptide samples prepared before LC-MS analyses, virtually all phosphopeptides were generated from in vitro kinase reactions. Assuming similar reactivity with light- or heavy-ATP, we expected to observe all phosphopeptides in doublets with similar intensities. We applied our in-house software LAXIC [24] to identify all peak pairs that were separated by ^{16}O and ^{18}O phosphoryl groups with similar intensities, and we calculated the successful phosphopeptide identification rate through three steps. First, the light and heavy peak pairs were identified from MS scans through two criteria: (1) the mass difference of 6 Da between the two peaks, and (2) the peaks were detected in the same full MS scan. Next, the light/heavy phosphopeptide pairs were selected from light and heavy phosphopeptides identified

Table 1. Number of Phosphopeptide Pairs in MS Spectra and Phosphopeptides Identified by MS^2. MQ is MaxQuant and PD is Proteome Discover

Kinase	Doublet	Doublet-MQ	Doublet-PD
CK2	4752	1498	2213
MPK6	4053	1472	1816
SnRK2.6	6749	1837	1405

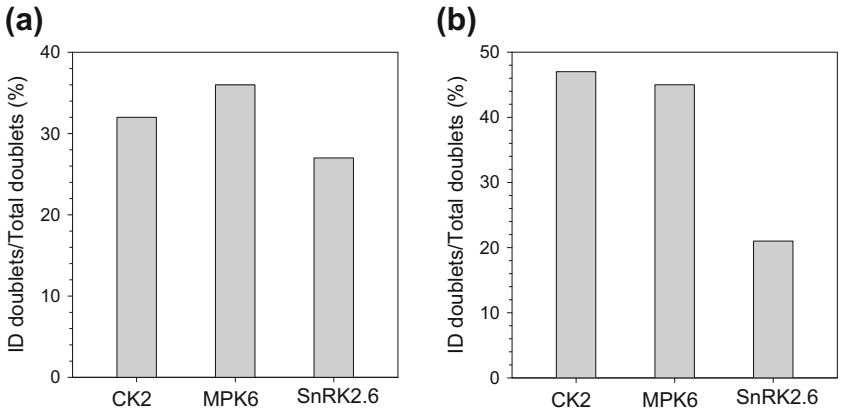


Figure 4. Comparison of phosphopeptide identification efficiencies between kinase reactions and search engines. The percentages of identification efficiency from (**a**) MaxQuant and (**b**) Sequest search engines were calculated by dividing phosphopeptides identified by MS/MS over total phosphopeptide pairs

through MS/MS, which fit the two criteria we mentioned above. Finally, the successful phosphopeptide identification rate was acquired via the ratio of the identified light/heavy phosphopeptide pairs over the identified light/heavy peak pairs. In total, we identified 4752, 4053, and 6749 pairs in samples related to the kinase reactions of CK2, MPK6, and SnRK2.6, respectively (Supplementary Tables S1–S1), which meet the criteria. Accordingly, we calculated the efficiency of phosphopeptide identification (Table 1 and Figure 4) by comparing the number of phosphopeptides identified by MS^2 (Supplementary Tables S1–S1) against the number of phosphopeptide peak pairs in the MS spectra. The percentages of successful phosphopeptide identification are 31%, 36%, and 27% for CK2, MPK6, and SnRK2.6, respectively. On average, only about 30% of phosphopeptides were identified by our current LC-MS instrument.

We also examined the effect of different search algorithms. Besides MaxQuant, we also searched the MS/MS spectra using Proteome Discoverer 2.1. MaxQuant is based on Andromeda whereas Proteome Discoverer uses Sequest HT as the search engine. Overall, with the same FDR cutoff value (FDR <1%), there are some obvious difference in the number of phosphopeptides identified by Andromeda (MaxQuant) or Sequest HT (Proteome Discoverer), but the efficiency of phosphopeptide identification is within a similar range (Table 1 and Figure 4).

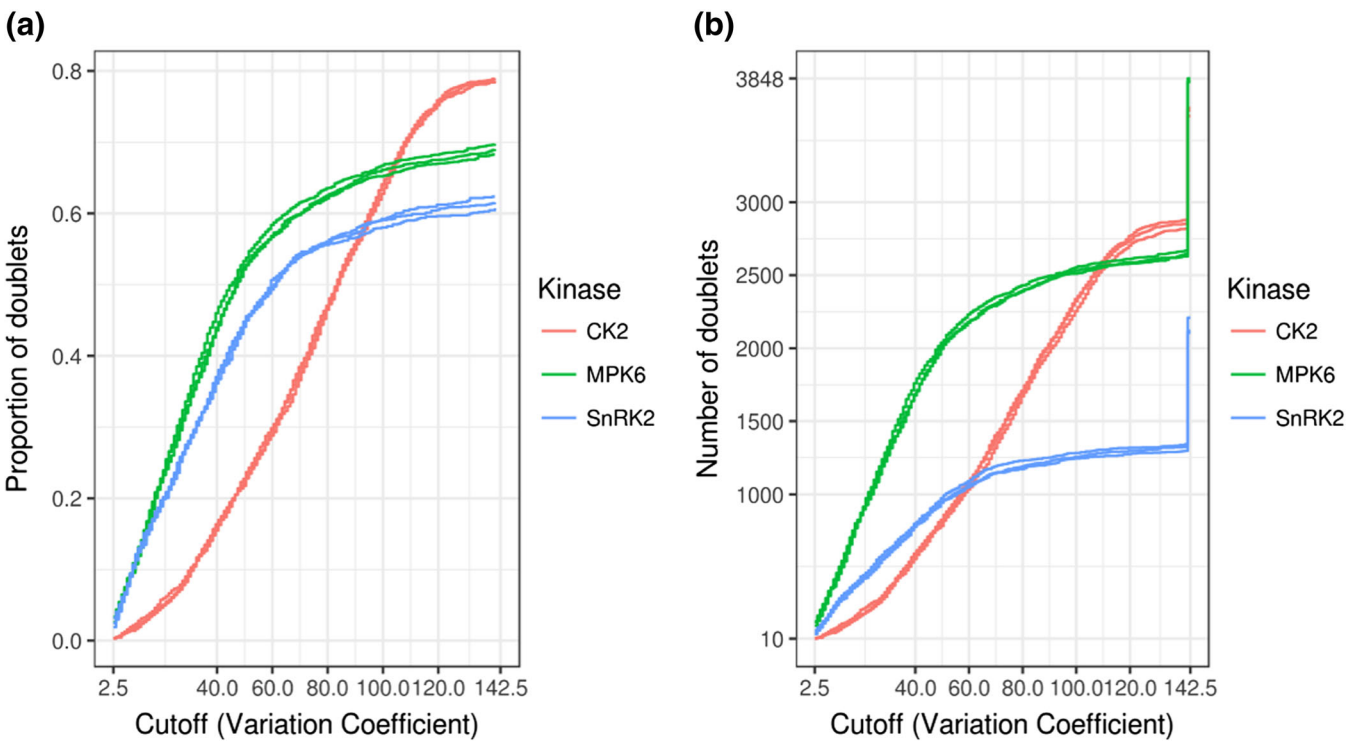


Figure 5. (**a**) Proportion of phosphopeptide doublets versus cut-off value; (**b**) number of phosphopeptide doublets versus cut-off value

There are multiple factors that may contribute to the relatively low efficiency of phosphopeptide identification (~30%). One obvious possibility is poor fragmentation of phosphopeptides in MS^2 spectra. We generated two plots to show the proportion and number of phosphopeptide doublets versus cut-off value (Figure 5a and b). The plots are quite informative. The maximum proportion of phosphopeptide doublets subjected to MS/MS is around 60%, indicating that 40% of the phosphopeptide doublets were not selected for MS/MS in our study. These phosphopeptide doublets that were not selected for MS/MS are likely of low abundance. When phosphopeptide abundance is low enough, the isotope pattern cannot be identified, and monoisotopic peaks cannot be recognized for MS/MS. Among the 60% of phosphopeptide pairs that were selected for MS/MS, the phosphopeptides in MPK6's samples have the highest identification efficiency. This is consistent with previous data that indicates proline-containing peptides have a high degree of peptide backbone fragment [30], which may facilitate identification, and the fact that most of MPK6's substrate peptides have the [(-pS/pT)-P-] motif. Moreover, other reasons such as additional modifications on the phosphopeptides [31] or variant isoforms not listed in the database may contribute to the high percentage of unassigned spectra.

Conclusion

Large scale analysis of protein phosphorylation, or phosphoproteomics, has become an important component of systems biology studies. While the advances of mass spectrometers in speed and accuracy, along with the introduction of ultra-high performance liquid chromatography, have greatly improved phosphoproteome coverage, it is critical to evaluate whether the phosphoproteomic data is comprehensive, especially considering that protein phosphorylation is highly dynamic. We have presented a novel method to estimate the efficiency of phosphopeptide identification by generating a large pool of phosphopeptides through direct isolation from cell lysates and in vitro kinase reactions. These phosphopeptides can be recognized according to specific features, though they may or may not be isolated for MS/MS. Examination of MS/MS data and MS features indicates that, on average, 30% of phosphopeptides were identified by MS/MS. Poor fragmentation and low abundance contribute to 70% of the phosphopeptides not being identified by MS/MS. This study highlights the need for additional efforts to increase the yield of phosphopeptides for MS analyses, possibly through better sample preparation, phosphopeptide enrichment, and LC resolution, and to further improve phosphopeptide fragmentation through alternative methods.

Acknowledgements

This study was partially supported by NIH grants 1R01GM111788 and 5R01GM088317, and NSF grant 1506752.

References

1. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., Sudarsanam, S.: The protein kinase complement of the human genome. Science **298**, 1912 (2002)
2. Schulze, W.X.: Proteomics approaches to understand protein phosphorylation in pathway modulation. Curr. Opin. Plant Biol. **13**, 280–287 (2010)
3. Cohen, P.: The origins of protein phosphorylation. Nat. Cell. Biol. **4**, E127–E130 (2002)
4. Hunter, T.: Signaling – 2000 and beyond. Cell **100**, 113–127 (2000)
5. Dephoure, N., Zhou, C., Villen, J., Beausoleil, S.A., Bakalarski, C.E., Elledge, S.J.: A quantitative atlas of mitotic phosphorylation. Proc. Natl. Acad. Sci. U.S.A. **105**, 10762–10767 (2008)
6. Mann, M., Ong, S.E., Gronborg, M., Steen, H., Jensen, O.N., Pandey, A.: Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. Trends Biotechnol. **20**, 261–268 (2002)
7. Stensballe, A., Andersen, S., Jensen, O.N.: Characterization of phosphoproteins from electrophoretic gels by nanoscale Fe(III) affinity chromatography with off-line mass spectrometry analysis. Proteomics **1**, 207–222 (2001)
8. Tsai, C.F., Hsu, C.C., Hung, J.N., Wang, Y.T., Choong, W.K., Zeng, M.Y.: Sequential phosphoproteomic enrichment through complementary metal-directed immobilized metal ion affinity chromatography. Anal. Chem. **86**, 685–693 (2014)
9. Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P., Jorgensen, T.J.D.: Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Mol. Cell. Proteom. **4**, 873–886 (2005)
10. Sugiyama, N., Masuda, T., Shinoda, K., Nakamura, A., Tomita, M., Ishihama, Y.: Phosphopeptide enrichment by aliphatic hydroxy acid-modified metal oxide chromatography for nano-LC-MS/MS in proteomics applications. Mol. Cell. Proteom **6**, 1103–1109 (2007)
11. Iliuk, A.B., Martin, V.A., Alicie, B.M., Geahlen, R.L., Tao, W.A.: In-depth analyses of kinase-dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble nanopolymers. Mol. Cell. Proteom. **9**, 2162–2172 (2010)
12. Rush, J., Moritz, A., Lee, K.A., Guo, A., Goss, V.L., Spek, E.J.: Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat. Biotechnol. **23**, 94–101 (2005)
13. Engholm-Keller, K., Larsen, M.R.: Technologies and challenges in large-scale phosphoproteomics. Proteomics **13**, 910–931 (2013)
14. Winter, D., Seidler, J., Ziv, Y., Shiloh, Y., Lehmann, W.D.: Citrate boosts the performance of phosphopeptide analysis by UPLC-ESI-MS/MS. J. Proteome Res. **8**, 418–424 (2009)
15. Leitner, A., Foettinger, A., Lindner, W.: Improving fragmentation of poorly fragmenting peptides and phosphopeptides during collision-induced dissociation by malondialdehyde modification of arginine residues. J. Mass Spectrom. **42**, 950–959 (2007)
16. Iliuk, A.B., Arrington, J.V., Tao, W.A.: Analytical challenges translating mass spectrometry-based phosphoproteomics from discovery to clinical applications. Electrophoresis **35**, 3430–3440 (2014)
17. Boersema, P.J., Mohammed, S., Heck, A.J.: Phosphopeptide fragmentation and analysis by mass spectrometry. J. Mass Spectrom. **44**, 861–878 (2009)
18. Marx, H., Lemeer, S., Schliep, J.E., Matheron, L., Mohammed, S., Cox, J.: A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. Nat. Biotechnol. **31**, 557–564 (2013)
19. Zhou, M., Meng, Z., Jobson, A.G., Pommier, Y., Veenstra, T.D.: Detection of in vitro kinase generated protein phosphorylation sites using gamma[18O4]-ATP and mass spectrometry. Anal. Chem. **79**, 7603–7610 (2007)
20. Xue, L., Wang, W.H., Iliuk, A., Hu, L.H., Galan, J.A., Yu, S.: Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates. Proc. Natl. Acad. Sci. U.S.A. **109**, 5615–5620 (2012)
21. Xue, L., Wang, P., Cao, P., Zhu, J.K., Tao, W.A.: Identification of extracellular signal-regulated kinase 1 (ERK1) direct substrates using stable isotope labeled kinase assay-linked phosphoproteomics. Mol. Cell. Proteom. **13**, 3199–3210 (2014)

22. Cox, J., Mann, M.: MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. **26**, 1367–1372 (2008)
23. Schwartz, D., Gygi, S.P.: An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. **23**, 1391–1398 (2005)
24. Xue, L., Wang, P., Wang, L., Renzi, E., Radivojac, P., Tang, H.: Quantitative measurement of phosphoproteome response to osmotic stress in arabidopsis based on Library-Assisted eXtracted Ion Chromatogram (LAXIC). Mol. Cell. Proteom. **12**, 2354–2369 (2013)
25. Okuda, S., Watanabe, Y., Moriya, Y., Kawano, S., Yamamoto, T., Matsumoto, M.: jPOSTrepo: an international standard data repository for proteomes. Nucleic Acids Res (2016)
26. Marx, H., Minogue, C.E., Jayaraman, D., Richards, A.L., Kwiecien, N.W., Siahpirani, A.F.: A proteomic atlas of the legume *Medicago truncatula* and its nitrogen-fixing endosymbiont *Sinorhizobium meliloti*. Nat. Biotechnol. **34**, 1198–1205 (2016)
27. Tsai, C.F., Wang, Y.T., Yen, H.Y., Tsou, C.C., Ku, W.C., Lin, P.Y.: Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. Nat. Commun. **6** (2015)
28. Umezawa, T., Sugiyama, N., Takahashi, F., Anderson, J.C., Ishihama, Y., Peck, S.C.: Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in *Arabidopsis thaliana*. Sci. Signal. **6** (2013)
29. Wang, P.C., Xue, L., Batelli, G., Lee, S., Hou, Y.J., Van Oosten, M.J.: Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action. Proc. Natl. Acad. Sci. U.S.A. **110**, 11205–11210 (2013)
30. Bleiholder, C., Suhai, S., Harrison, A.G., Paizs, B.: Towards understanding the tandem mass spectra of protonated oligopeptides. 2: The proline effect in collision-induced dissociation of protonated Ala-Ala-Xxx-Pro-Ala (Xxx = Ala, Ser, Leu, Val, Phe, and Trp). J. Am. Soc. Mass Spectrom. **22**, 1032–1039 (2011)
31. Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., Rad, R., Huttlin, E.L.: A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. **33**, 743–749 (2015)