Systematic Analysis of Protein Phosphorylation Networks From Phosphoproteomic Data*[S]

Chunxia Song‡‡, Mingliang Ye‡‡, Zexian Liu§, Han Cheng§, Xinning Jiang‡, Guanghui Han‡, Zhou Songyang¶, Yexiong Tan‖, Hongyang Wang‖, Jian Ren¶, Yu Xue§, and Hanfa Zou‡**

In eukaryotes, hundreds of protein kinases (PKs) specifically and precisely modify thousands of substrates at specific amino acid residues to faithfully orchestrate numerous biological processes, and reversibly determine the cellular dynamics and plasticity. Although over 100,000 phosphorylation sites (p-sites) have been experimentally identified from phosphoproteomic studies, the regulatory PKs for most of these sites still remain to be characterized. Here, we present a novel software package of iGPS for the prediction of *in vivo* site-specific kinase-substrate relations mainly from the phosphoproteomic data. By critical evaluations and comparisons, the performance of iGPS is satisfying and better than other existed tools. Based on the prediction results, we modeled protein phosphorylation networks and observed that the eukaryotic phospho-regulation is poorly conserved at the site and substrate levels. With an integrative procedure, we conducted a large-scale phosphorylation analysis of human liver and experimentally identified 9719 p-sites in 2998 proteins. Using iGPS, we predicted a human liver protein phosphorylation networks containing 12,819 potential site-specific kinase-substrate relations among 350 PKs and 962 substrates for 2633 p-sites. Further statistical analysis and comparison revealed that 127 PKs significantly modify more or fewer p-sites in the liver protein phosphorylation networks against the whole human protein phosphorylation network. The largest data set of the human liver phosphoproteome together with computational analyses can be useful for further experimental consideration. This work contributes to the understanding of phosphorylation mechanisms at the systemic level, and provides a powerful methodology for the general analysis of *in vivo* post-translational modifications regulating sub-proteomes. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.012625, 1070–1083, 2012.***

Protein kinase (PK)[1]-catalyzed phosphorylation is one of the most important and ubiquitous post-translational modifications (PTMs) of proteins. This process temporally and spatially modifies ~30% of all cellular proteins and plays a crucial role in regulating a variety of biological processes such as signal transduction and the cell cycle (1–3). The human genome encodes 518 PK genes (~2% of the genome), with different PKs showing distinct recognition specificities; each PK modifies only a limited subset of substrates, thereby guaranteeing the fidelity of cell signaling (1–3). It is accepted that short linear motifs (SLMs) around the phosphorylation sites (p-sites) provide primary specificity (2, 4–6), and a variety of additional contextual factors, including co-localization, co-expression, co-complex, and physical interaction of the PKs with their targets, contribute additional specificity *in vivo* (7–10). Aberrances of PKs or key substrates disrupt normal function, rewire signaling pathways, and are implicated in various diseases and cancers (3, 11). In this regard, the identification of kinase-specific p-sites and the systematic elucidation of site-specific kinase-substrate relations (ssKSRs) would pro-

From the ‡CAS Key Laboratory of Separation Sciences for Analytical Chemistry, National Chromatographic RandA Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, Liaoning 116023, China; §Hubei Bioinformatics and Molecular Imaging Key Laboratory, Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China; ¶State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China; ‖The International Cooperation Laboratory on Signal Transduction of Eastern Hepatobiliary Surgery Institute, Second Military Medical University, Shanghai, 200438, China

Received July 31, 2011, and in revised form, May 9, 2012

Published, MCP Papers in Press, July 13, 2012, DOI 10.1074/mcp.M111.012625

[1] The abbreviations used are: PK, protein kinase; PTM, post-translational modification; SLM, short linear motif; p-site, phosphorylation site; ssKSR, site-specific kinase-substrate relation; KSR, kinase-substrate relation; HTP-MS, high-throughput mass spectrometry; GPS, group-based prediction system; HPN, human phosphorylation network; iGPS, GPS algorithm with the interaction filter, or *in vivo* GPS; PPI, protein-protein interaction; PPN, protein phosphorylation network; RP-RPLC, reversed-phase-reversed-phase liquid chromatography; *P*, positive control; *N*, negative control; *Sn*, sensitivity; *Sp*, specificity; *Ac*, accuracy; *MCC*, Mathew correlation coefficient; *Kpr*, kinase precision; *Lpr*, large-scale precision; *FPR*, false positive rate; FDR, false discovery rate; STK, serine/threonine kinase; TK, tyrosine kinase; KTF, kiss-then-farewell; No PPI, without PPI; Exp. PPI, experimental PPI; KOW, Kyprides, Ouzounis, Woese; PAF, polymerase-associated factor; CTD, C-terminal repeat domain; HLPP, Human Liver Proteome Project; MPSS, massively parallel signature sequencing; CNHLPP, Chinese human liver proteome project; pS, phosphoserine; pT, phosphothreonine; pY, phosphotyrosine.

vide a fundamental basis for understanding cell plasticity and dynamics and for dissecting the molecular mechanisms of various diseases, whereas the ultimate progress could suggest potential drug targets for future biomedical design (8–10).

Conventional experimental identification of ssKSRs, performed in a one-by-one manner, is labor-intensive, time-consuming and expensive. There are only 3508 known kinase-specific p-sites in the 1390 proteins collected in the Phospho.ELM 8.2 database (released in April 2009) (12). In 2005, Ptacek *et al.* detected more than 4000 *in vitro* kinase-substrate relations (KSRs) in *Saccharomyces cerevisiae* using protein chip technology, although the exact p-sites were not determined (13). Recently, rapid advances in phosphoproteomics have provided a great opportunity to systematically assess phosphorylation (1, 14–21). State-of-the-art high-throughput mass spectrometry (HTP-MS) techniques have the ability to detect thousands of p-sites in cells or tissues in a single experiment (1, 14, 16, 22). We have collected 145,646 eukaryotic p-sites, primarily from these large-scale assays (supplemental Table S1); the regulatory PKs for 97.6% of these sites remain to be characterized.

Alternatively, the *in silico* prediction of ssKSRs can generate useful information for subsequent experimental manipulation. In 2001, Yaffe *et al.* developed the SLM-based software Scansite for the prediction of ssKSRs directly from protein primary sequences (7). Later, the strategy was employed in a variety of kinase-specific predictors (23), including our group-based prediction system (GPS) program (24). These tools may guarantee partially correct predictions for *in vitro* phosphorylation, but they are far from being adequate for *in vivo* hits because the contributions of various contextual factors cannot be neglected. To address this problem, Linding *et al.* developed a predictor of NetworKIN by combining an SLM-based approach with network contextual information to predict *in vivo* ssKSRs, and a potential *in vivo* human phosphorylation network (HPN) was modeled by annotating the phosphoproteomic data (8, 9).

In this work, we developed a software package of iGPS (GPS algorithm with the interaction filter, or *in vivo* GPS) mainly for the prediction of *in vivo* ssKSRs.Eukaryotic PKs were classified into a hierarchy with four levels: group, family, subfamily, and single PK (3). Based on the hypothesis that similar PKs recognize similar SLMs, we selected a predictor in GPS 2.0 (24) for each PK and directly predicted the potential PKs for the un-annotated p-sites from the phosphoproteomic studies. Consequently, protein–protein interaction (PPI) information was used as the major contextual factor to reduce over 95% potentially false-positive hits. The performance of iGPS was shown by critical evaluations and comparisons to be promising for the accurate prediction of *in vivo* ssKSRs. Based on the prediction results of iGPS, we modeled eukaryotic protein phosphorylation networks (PPNs) and observed that phosphorylation regulation changes dramatically over the course of evolution, with poor conservation at both the site and substrate levels. This observation is consistent with previous studies (17, 25). Furthermore, we combined a new multidimensional separation approach using reversed-phase-reversed-phase liquid chromatography (RP-RPLC) (22), with HTP-MS and a new data process platform of ArMone (26) to conduct a large-scale phosphorylation analysis of the human liver. Totally, 9719 p-sites of 2998 substrates were identified from 10,644 non-redundant phosphopeptides. The potential ssKSRs were predicted for the human liver phosphoproteome, whereas further statistical analysis suggested that 60 and 67 PKs preferentially regulate more or fewer p-sites in the human liver PPN (p value<0.01). A number of results are consistent with previous observations, whereas other predictions can be useful for further experimental manipulation.

EXPERIMENTAL PROCEDURES

Collection of Known p-sites—The experimentally identified p-sites were taken from several major databases, including PhosphoPep v2.0 (27), Phospho.ELM 8.3 (released in April 2010) (12, 28), SysPTM 1.1 (29), PhosphoSitePlus (30), and HPRD 9.0 (31). We also collected thousands of p-sites from several published articles (14, 17–21). The organism-specific information was distinguished from the database or data set comments. All p-sites with their protein sequences were mapped to the UniProt benchmark sequences (More details in supplemental Experimental Procedures). In total, the final phosphorylation data set contains 145,646 p-sites in 28,457 substrates, with 14,534, 5555, 15,622, 49,119, and 60,816 p-sites in *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens*, respectively (supplemental Table S1).

To evaluate the prediction performance of iGPS, we took 3508 experimentally verified kinase-specific p-sites in 1,390 proteins from Phospho.ELM 8.2 (12, 28) as the positive control (*P*) (supplemental Table S2), whereas all other Ser/Thr or Tyr residues in the same substrates were regarded as the negative control (*N*). Thus:

$$P = TP + FN \quad \text{(Eq. 1)}$$

$$N = TN + FP \quad \text{(Eq. 2)}$$

To compare iGPS with NetworKIN (8, 9), we collected 1701 kinase-specific p-sites in 830 substrates from Phospho.ELM 9.0 (released in September 2010) (12, 28) for twelve PK groups in iGPS, including AGC/AKT, AGC/PKA, Atypical/PIKK/ATM, CAMK/CAMK2, CMGC/CDK/CDC2, CMGC/MAPK, Other/AUR/AUR-A, Other/CK2, TK/ABL, TK/EGFR, TK/Src, and TK/Syk. Additionally, we manually collected 450 phosphoproteins with 1193 kinase-specific p-sites from the scientific literature. After redundant clearing, the testing data set contains 2894 kinase-specific p-sites in 1280 proteins (Table I, supplemental Table S3).

Performance Evaluation—As previously described (24), four standard measurements, including sensitivity (*Sn*), specificity (*Sp*), accuracy (*Ac*), and Mathew correlation coefficient (*MCC*) were defined as below:

$$Sn = \frac{TP}{TP + FN} \quad \text{(Eq. 3)}$$

$$Sp = \frac{TN}{TN + FP} \quad \text{(Eq. 4)}$$

TABLE I
A testing data set to compare iGPS with NetworKIN (8, 9) for twelve PK groups

PK clusters	Phospho.ELM 9.0		New data[a]		Total	
	Sub.[b]	Sites	Sub.	Sites	Sub.	Sites
AGC/AKT	64	90	55	90	119	180
AGC/PKA	242	385	12	25	254	410
Atypical/PIKK/ATM	29	58	19	35	48	93
CAMK/CAMK2	57	91	50	97	107	188
CMGC/CDK/CDC2	84	147	40	138	124	285
CMGC/MAPK	172	300	91	202	263	502
Other/AUR/AUR-A	13	29	22	26	35	55
Other/CK2	164	321	70	153	234	474
TK/ABL	36	51	20	39	56	90
TK/EGFR	23	55	5	19	28	74
TK/Src	98	156	42	86	140	242
TK/Syk	27	64	14	32	41	96
Total	830	1,701	450	1,193	1280	2,894

[a] The new data set was collected from the scientific literature.
[b] Sub., number of phosphorylated substrates.

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{(Eq. 5)}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \quad \text{(Eq. 6)}$$

To evaluate the proportion of correct hits among the total predicted PKs, we defined the kinase precision (*Kpr*). Given *m* real p-sites, if we predict Y_i PKs for the *i* site, with X_i being the experimentally verified PKs, the *Kpr* is defined as:

$$Kpr = \frac{\sum_{i=1}^{m} X_i}{\sum_{i=1}^{m} Y_i} \quad \text{(Eq. 7)}$$

As previously described (24), we adopted large-scale precision (*Lpr*) to estimate the proportion of correct hits from the large-scale prediction of the PK information. Given *n* potential p-sites, suppose that *p* positive hits are phosphorylated by a PK or PK cluster under a certain threshold with a calculated false positive rate (*FPR*) value. Then the theoretically maximal false positive hits will be $n * FPR$, if all of the *n* site are real negative sites. Thus, the minimal proportion of correct predictions can be calculated as:

$$Lpr = \frac{p - n \times FPR}{p} \quad \text{(Eq. 8)}$$

Furthermore, the above equation was extended to estimate the average precision for a large data set prediction. Given *n* potential p-sites, suppose that we predict p_1, p_2, …, and p_k positive hits for *k* PKs, respectively, the theoretically maximal false positive hits for each PK will be $n * FPR_j$ ($j = 1, 2, \ldots, k$). Then the total precision can be calculated as:

$$Lpr = \frac{\sum_{i=1}^{k} p_i - \sum_{j=1}^{k} (n \times FPR_j)}{\sum_{i=1}^{k} p_i} \quad \text{(Eq. 9)}$$

Although the *Lpr* does underestimate the accuracy of large-scale predictions, nevertheless, it still showed the power and utility of the iGPS algorithm.

Liver Sample Preparation and Phosphopeptide Enrichment—The study was approved by the Institutional Review Board of Eastern Hepatobiliary Surgery Hospital (Shanghai, China). Informed consent was obtained from patients enrolled in this study. The human liver tissues are the non-tumor liver tissues ≥2 cm outside the hepatic hemangiomas removed by surgical operation. The liver tissues have been checked by histopathological examination to exclude the presence of invading or microscopic metastatic tumor cells.

As previously described (15), human liver tissues were lysated in 20 ml ice-cold homogenization buffer consisting of 8 M urea, 1% Triton X-100 v/v, 65 mM DTT, 1 mM EDTA, 0.5 mM EGTA, 1 mM PMSF, 200 μl protease inhibitor mixture (Sigma), phosphatase inhibitors (1 mM sodium fluoride, 1 mM sodium orthovanadate, 1 mM β-glycerophosphate, 10 mM sodium pyrophosphate), and 40 mM Tris-HCl at pH 7.4. After being resuspended in the denaturing buffer containing 8 M urea and 50 mM Tris-HCl (pH 8.2), the proteins were reduced by DTT at 37 °C for 2 h and alkylated by iodoacetamide in the dark at room temperature for 40 min. Then the solutions were diluted to 1 M urea with 50 mM Tris-HCl and trypsin was added, with the weight ratio of trypsin to protein at 1/25, and incubated at 37 °C overnight. All of the resulting peptide solution was stored under −80 °C.

The phosphopeptides were enriched from the digest of human liver lysate by Ti^{4+}-IMAC microspheres (16). Briefly, peptide mixtures which were first incubated with the Ti^{4+}-IMAC microsphere suspension (10 mg ml^{-1} in 80% ACN, 0.1% TFA) for 30 min then were washed with a solution containing 50% ACN, 6% TFA and 200 mM NaCl, followed by washing with 30% ACN/0.1% TFA. Finally, the enriched phosphopeptides were eluted with 10% $NH_3 \cdot H_2O$ and dried by vacuum centrifugation.

Multidimensional Separation of Phosphopeptides and Mass Spectrometry Analysis—The enriched phosphopeptides were redissolved in mobile phase A (25 mM ammonium formate (NH_4FA) aqueous buffer, pH 7.5) and then were loaded onto the first dimensional separation column (250 mm × 4.6 mm I. D. column packed with 5 μm Hypersil GOLD aQ C18, Thermo). Mobile phase B was 25 mM NH_4FA in water/acetonitrile (1: 9) and the gradient elution was performed with 0%–10% B (0–80 min) and 10%–35% B (80–90 min). A total of 90 fractions (one fraction per each minute) from the first dimensional separation were collected and then divided into two groups: an early group (fractions 1–45) and later group (fractions 46–90). Then every two fractions with an equal time interval (fractions 1 and 46, 2 and 47, and so on) were mixed as described previously (22). Finally, half the

total number of 45 fractions were lyophilized and submitted to the second dimensional RP-RPLC separation.

The HPLC system consisted of a degasser and a quaternary surveyor MS pump (Thermo Finnigan, CA). The capillary separation column was prepared as previously described (32). Briefly, the capillary was manually pulled to a fine point of ~ 3 μm with a flame torch and then was packed with C18 aQ beads (5 μm, 120 Å) in a homemade pneumatic pressure cell using a slurry packing method. The lyophilized fractions from the first dimension were resuspended in 10 μL of 0.1% FA solution, and 2 μL of the sample were then manually loaded onto the column with each sample replicated three times. The mobile phase A was 0.1% FA in water and B was 0.1% FA in acetonitrile; gradient elution was performed with 3–25% B in 90 min at a flow rate after splitting 200 nL/min.

The MS analysis was performed on LTQ-Orbitrap mass spectrometer (Thermo, San Jose, CA) with a resolution of 100000 at *m/z* 400. The temperature of the ion transfer capillary was set at 200 °C. The spray voltage was set at 1.8 kV and the normalized collision energy was set at 35.0%. The detection of phosphopeptides was performed with the mass spectrometer set for a full scan MS followed by three data-dependent MS2 events. Subsequently, the MS3 spectrum was automatically triggered when a neutral loss event of 97.97, 48.99, or 32.66 Da (loss of H_3PO_4 for the +1, +2 and +3 charge ions, respectively) was detected among the three most intense peaks in MS2. The target ion setting was 5e5 for the Orbitrap, with a maximum fill-time of 500 ms. MS2 scans were acquired in the LTQ with a target ion setting of 3e4 and a maximum fill-time of 100 ms. The dynamic exclusion function was set as follows: repeat count 2, repeat duration 30 s, and an exclusion duration of 60 s.

Database Search and Data Analysis—The peak list files for MS2 and MS3 spectra were extracted by Extract_msn.exe in Bioworks 3.3 using default settings. The MS2 and MS3 spectra were searched with SEQUEST (version 2.8) against a composite database containing both the original human IPI protein database (ipi.HUMAN.v3.17.fasta, including 60234 entries, http://www.ebi.ac.uk/IPI/IPIhuman.html) and its reversed complement. Trypsin was set as the specific proteolytic enzyme with fully enzymatic and up to two missed cleavages were allowed. The mass tolerance for the precursor ion was set as 50 ppm and 0.8 Da for the fragment ion. Carbamidomethylation (+57.02146 Da) on cysteine was set as fixed modification, whereas oxidation (+15.99452 Da) on methionine, phosphorylation (+79.96633 Da) on serine, threonine, and tyrosine were set as variable modifications. For the searching with MS3 data, β-elimination of phosphoric acid (–18.010565 Da) on serine and threonine residues were also selected as variable modifications.

The database search results were processed with the software suite of ArMone, which was recently designed for the management and analysis of phosphoproteome data (26). The assignment of p-sites from identified phosphopeptides was determined by the Ascore algorithm (33), which was also implemented in ArMone (26). Based on the classification filtering strategy (34), the identified phosphopeptides were classified into four groups (supplemental Table S4, S5, S6, and S7). The mass spectra without significant neutral loss or without consecutive MS3 spectra were defined as the NoNeutral class (supplemental Table S5). The group of mass spectra with significant neutral loss was further separated into three classes, such as MS2/MS3 (MS2/MS3 pair can generate the same phosphopeptide assignment, supplemental Table S4), NeutralMS2 (Phosphopeptide exclusively generated from the MS2 spectra, supplemental Table S6) and NeutralMS3 (Exclusively generated from the MS3 spectra, supplemental Table S7) classes. Based on the different characteristics of four classes identifications, different filtering strategies were adopted to achieve the false discovery rate (FDR)<1% (FDR = $2N_d/N$, in which N is the number of peptide matches with scores above the cut-off and N_d was the number of matches to decoy sequences). For MS2/MS3 identifications, DeltaCn'm>0.1, Xcorr's>0.63; for NoNeutral, DeltaCn>0.1, Xcorr>2.6, 3.2, 4.2 and 4.8 for +1, +2, +3 and +4 charge states respectively; for NeutralMS2, DeltaCn>0.1, Xcorr>2.5, 3.8, 4.7 and 5 for +1, +2, +3 and +4 charge states respectively; for NeutralMS3, DeltaCn> 0.1, Xcorr>2.2, 3.5, 4.5 and 4.3 for +1, +2, +3 and +4 charge states respectively.

All of the mass spectra with matched ion information of identified unique phosphopeptides were generated by the batch drawing module of Armone (26) and exported in .html format with hyperlinks to the spectrum images. The annotated spectra can be accessed at the publicly accessible database Tranche (https://proteomecommons.org/tranche/), using the following hash: XnGGs9eaZCdxxw6gbI6RXpfe+0EKutSSQL7Ue3zcwrGG4IMY44oXJAr0B+mYzJmWUou2bfn0B6ojsAhSBvaRTAIX4G4AAAAAAAADnQ==.

RESULTS

Development of iGPS for the Prediction of in vivo ssKSRs—In a previous study (24) we developed the software GPS 2.0, which could predict the kinase-specific p-sites for 408 PKs in humans. We estimated the *FPR* by randomly generating PSP(7, 7) peptides based on the real frequencies of amino acids in the eukaryotic proteomes. The high, medium, and low cut-off values were chosen based on the FPRs of 2%, 6%, and 10% for serine/threonine kinases (STKs), and 4%, 9%, and 15% for tyrosine kinases (TKs). With these high thresholds, we directly predicted 170,593 ssKSRs for 12,219 unannotated p-sites from a total of 13,254 mammalian sites (~14 PK groups per site) (24). Although the coverage rate was very high at ~92.19%, it is strongly suspected that a large proportion of the predicted results might be false positive hits, because *in vivo* numerous contextual factors only permit a small number of PKs to specifically reach their substrates. Moreover, the GPS 2.0 program used a hierarchical structure with four levels, whereas only 39,540 kinase-substrate interactions for 7813 p-sites (a coverage rate of ~58.9%) were predicted at the single PK level (24). In addition, in our current data set, there are 35,711 nonmammalian p-sites (~24.5%). Annotation of the potential PK information for these sites continues to be a formidable challenge.

To address these issues, we first hypothesized that eukaryotic PKs classified in a same group, family or subfamily would recognize similar consensus motifs/patterns of substrate modification, although the recognition similarities would differ in extent for the different PK clusters. Thus, if one site was predicted to be phosphorylated by any PK group, family, or subfamily, we assumed that all the PKs in the same cluster would phosphorylate the site. Second, the "kiss-then-farewell" (KTF) model was adopted (24, 35). Thus, the PPI information was used as the major contextual factor to reduce false positive predictions (supplemental Table S19 and supplemental Fig. S1). Although some proportion of "kisses" might be slight and transient and thus not detected in standard PPI screenings, the interaction information would be expected to significantly enrich PK substrates, at least for Aurora-B (24) and PKA (35).

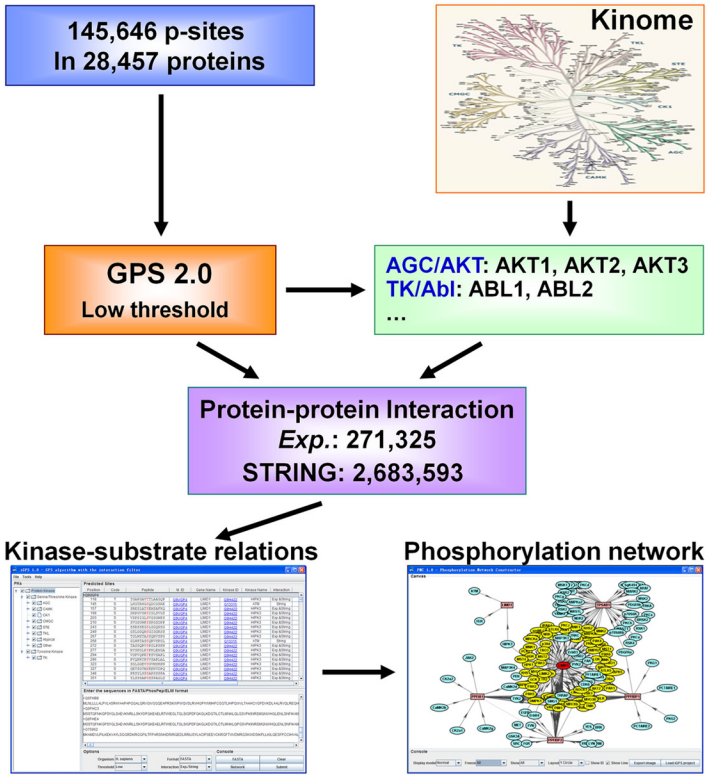

FIG. 1. **The computational procedure for predicting ssKSRs from eukaryotic phosphoproteomic data and visualizing potential PPNs.**

Based on the two hypotheses, we developed a novel software package of iGPS mainly for the prediction of *in vivo* ssKSRs from eukaryotic phosphoproteomic data (supplemental Fig. S2). The full computational procedure is shown in Fig. 1. First, we collected 28,457 phosphorylated substrates containing 145,646 p-sites from five eukaryotic organisms (Fig. 1 and supplemental Table S1). The GPS 2.0 algorithm was used for the prediction of ssKSRs, whereas a low threshold was chosen with a *FPR* of 10% for STKs, and 15% for TKs, respectively (24). The original GPS 2.0 software contained 213 subpredictors for 144 STK and 69 TK clusters (24). However, because of the training data limitation, not all predictors showed an accurate performance. To ensure the accuracy of our predictions, we selected 56 STK- and 21 TK-specific predictors in GPS 2.0 for most of the PKs (supplemental Table S8). To further reduce the number of false-positive hits at the substrate level, the predicted and experimentally identified physical interactions between PKs and substrates were combined and used as the major contextual filters (Fig. 1 and see supplemental Experimental Procedures). To visualize potential PPNs, the orientation was simply defined as Kinase -> Substrate. Because a number of substrates might be PKs, the orientation could also be Kinase *A* -> Kinase *B* (*A* phosphorylates *B*) or Kinase *A* <-> Kinase *B* (*A* and *B* reciprocally phosphorylate each other) (Fig. 1).

The PPI Information Can be an Efficient Filter to Reduce Potentially False Positive Hits—Although a number of contextual factors are believed to contribute additional specificity beyond the phosphorylation motif (7–10), here we only adopted one major factor of PPI information. Whether this single factor is able to significantly improve the prediction accuracy remains to be tested. Here, we collected a testing data set by obtaining from Phospho.ELM 8.2 (12), with 3508 kinase-specific p-sites in 1390 substrates (supplemental Table S20). The regulatory PKs of 3485 p-sites (99.3%) were identified in low-throughput experiments at a high level of confidence (supplemental Table S2). We calculated the prediction performances under three conditions: without PPI (No PPI), with both STRING and experimental PPI (STRING and Exp. PPI), and with experimental PPI information alone (Exp. PPI). The results for several typical predictors are shown in Table II, and the full performance results are available in supplemental Table S9. Clearly, the use of STRING and Exp. PPI or Exp. PPI moderately reduces the *Sn* but greatly enhances the *Sp* and *Kpr*. For example, without the PPI information, the *Sn*, *Sp*, and *Kpr* of AGC/AKT are 95.56%, 91.96%, and 12.45%, respectively. These values are 88.89%, 98.27%, and 25.45%, respectively, when STRING and Exp. PPI are used and 71.11%, 98.72%, and 51.74%, respectively, when Exp. PPI is used (Table II). For all of the PKs, the *Sn* decreased from 89.79% (No PPI) to 64.74% (both STRING and Exp. PPI) and 49.09% (Exp. PPI), and the *Kpr* increased from 23.25% (No PPI) to 40.69% (STRING and Exp. PPI) and 62.65% (Exp. PPI), 1.8- and 2.7-fold enhancements, respectively (Table II). This shows that the PPI information can greatly decrease the number of potential false-positive PKs for p-site annotation. In

TABLE II
Performance evaluation

For the testing data set, we took 1390 substrates with 3508 kinase-specific p-sites from Phospho.ELM 8.2 database (12). We calculated the prediction performances under three conditions, *i.e.* without any PPI (No PPI), STRING and experimental PPI (STRING & Exp. PPI), and only experimental PPI (Exp. PPI) information. The PPI filter moderately reduced the *Sn* but greatly enhanced the *Sp* and *Kpr*.

Predictor	No PPI			String & Exp. PPI			Exp. PPI		
	Sn	*Sp*	*Kpr*	*Sn*	*Sp*	*Kpr*	*Sn*	*Sp*	*Kpr*
AGC/AKT	95.56%	91.96%	12.45%	88.89%	98.27%	25.45%	71.11%	98.72%	51.74%
AGC/PKA	89.12%	92.11%	11.18%	47.75%	98.96%	25.09%	30.24%	99.28%	53.89%
CAMK/CAMK2	95.74%	90.19%	18.42%	59.57%	98.80%	35.81%	19.15%	99.31%	56.79%
CMGC/CDK	98.56%	85.41%	19.05%	64.27%	97.26%	36.57%	25.94%	98.46%	57.89%
CMGC/MAPK	92.41%	86.18%	17.03%	76.90%	97.37%	24.45%	60.07%	98.10%	56.11%
Other/AUR/AUR-B	100%	83.39%	14.99%	53.33%	94.57%	36.84%	30.00%	97.87%	46.30%
Other/CK2	85.98%	84.25%	10.39%	48.60%	98.02%	29.62%	41.12%	98.30%	63.64%
TK/Abl	82.00%	81.45%	16.94%	66.00%	91.21%	26.63%	32.00%	94.73%	36.00%
TK/InsR	93.94%	81.51%	16.91%	78.79%	93.84%	26.56%	77.27%	94.96%	44.32%
TK/Src	70.75%	82.91%	11.79%	58.16%	93.41%	21.81%	46.60%	94.18%	41.65%
STKs	90.94%	80.56%	24.79%	63.11%	97.56%	43.47%	46.67%	98.37%	68.69%
TKs	85.34%	80.27%	13.51%	71.09%	93.64%	31.72%	58.52%	94.80%	51.12%
All PKs	89.79%	80.55%	23.25%	64.74%	97.33%	40.69%	49.09%	98.16%	62.65%

addition, predictions suggest that regulatory PKs for the p-sites may not be fully identified. Thus, we mixed these sites together with other un-annotated p-sites for further analysis.

Extensive evaluations were performed for the prediction of 145,646 eukaryotic p-sites in 28,457 substrates (supplemental Table S10). Without the PPI information, we directly predicted 4,001,298 ssKSRs for 127,697 p-sites of all of the species, with a coverage rate of 87.7% (Table III and supplemental Table S11). When PPI information was used, although the coverage rates were reduced to 30.4% and 11.0% for the total PPIs and the experimental PPIs, respectively, the potential ssKSRs were also decreased to 186,922 (4.67%) and 34,873 (0.87%) (Table III and supplemental Table S11). In this regard, the potentially false positive predictions were greatly reduced through the PPI contextual filter. Because the coverage rate with the exclusively experimental PPIs is too low, both experimental and predicted PPIs were used for further analysis. Although the coverage rate can be enhanced if the threshold is relaxed, we adopted the current parameters for iGPS to ensure a high degree of confidence.

Additionally, the *Lpr* values were calculated (supplemental Table S10). For AGC/PKA, the *Lpr* was enhanced from 52.23% (No PPI) to 58.06% (STRING and Exp. PPI) and 59.82% (Exp. PPI), with a p value of 2.8e-7 and 2.6e-6 (Fisher's Exact Test, 2-Tail, http://www.langsrud.com/fisher.htm), respectively. However, the *Lpr* was not significantly increased for AGC/AKT (p value $>$ 0.05) (supplemental Table S10). Although not all of the *Lpr* values were increased, the total *Lpr* for all of the PKs was enhanced from 50.58% (No PPI) to 55.89% (STRING and Exp. PPI) and 61.08 (Exp. PPI) (both p value $<$ 3e-280), respectively (supplemental Table S10).

Comparison of iGPS with NetworKIN—In 2007, Linding *et al.* presented a pioneering study by introducing contextual factors to reduce potentially false positive hits for the prediction of ssKSRs (8, 9). The NetworKIN 1.1 contains 21 PK classifiers to predict kinase-specific p-sites for 108 PKs (8, 9), whereas the NetworKIN 2.0 beta version can predict ssKSRs for 123 PKs with 59 PK classifiers (unpublished). Here, we compared iGPS with both NetworKIN 1.1 and NetworKIN 2.0 beta for 12 PK groups, including 8 STK groups and 4 TK groups (Table I). To avoid any bias, the same testing data set was adopted for iGPS, NetworKIN 1.1 and NetworKIN 2.0 beta. Besides the known p-sites in Phospho.ELM 9.0 (12, 28), we additionally curated kinase-specific p-sites from the scientific literature. The nonredundant testing data set contains 1280 substrates with 2894 p-sites for the 12 PK groups (Table I).

We fixed the *Sp* values of iGPS so as to be similar to the two predictors, and then compared the *Sn* values (Table IV). In NetworKIN 1.1, the PK classifiers of CMGC/CDK/CDC2, Other/AUR/AUR-A, TK/ABL and TK/Syk were not available (8, 9). Thus, we chose the cdk5 predictor for CMGC/CDK/CDC2, whereas the performances of Other/AUR/AUR-A, TK/ABL and TK/Syk were not compared. Except AGC/AKT, Atypical/PIKK/ATM and Other/CK2, iGPS outperformed NetworKIN 1.1 for up to 6 PK groups (Table IV). For example, when the *Sp* value was 93.28%, the *Sn* values of iGPS and NetworKIN 1.1 for AGC/PKA were 57.56% and 32.20%, respectively (Table IV). Also, when the *Sp* value was 96.97%, the *Sn* of iGPS (56.76%) was much better than NetworKIN 1.1 (12.16%) for TK/EGFR (Table IV). For NetworKIN 2.0 beta, it only showed superiority for Other/CK2, whereas the performance of Atypical/PIKK/ATM was reduced and similar with iGPS (Table IV). The iGPS generated much better performances on the remaining 10 PK groups (Table IV). Taken together, we proposed the performance of iGPS is generally better than NetworKIN.

Computational Modeling and Analysis of Eukaryotic PPNs from the Phosphoproteomic Data—For the five eukaryotic phosphoproteomes, we predicted 186,922 ssKSRs among

TABLE III
The statistics of prediction results for five eukaryotic phosphoproteomes

Organism	String & Exp. PPI					No PPI				
	PK	Sub.[a]	Site	ssKSR	Ave[b]	PK	Sub.	Site	ssKSR	Ave
S. cerevisiae	91	1598	7041	20,909	3.0	91	2658	12,889	145,409	11.3
C. elegans	110	272	544	867	1.6	302	2153	5112	107,738	21.1
D. melanogaster	140	888	2697	6191	2.3	172	3896	13,656	236,780	17.3
M. musculus	358	2349	11,191	45,032	4.0	415	8219	43,131	1,588,383	36.8
H. sapiens	380	4140	22,817	113,923	5.0	407	9452	52,909	1,922,988	36.3
Total	1,079	9247	44,290	186,922	4.2	1387	26,378	127,697	4,001,298	31.3

[a] Sub., number of phosphorylated substrates.
[b] Ave, the average number of upstream PKs per p-site.

TABLE IV
Comparison of iGPS with NetworKIN 1.1 and NetworKIN 2.0 beta

We fixed the *Sp* values of iGPS so as to be similar to the two predictors, and then compared the *Sn* values. The performances with better values than those from iGPS are bold.

PK clusters	NetworKIN				iGPS			
	Ac	*Sn*	*Sp*	*MCC*	*Ac*	*Sn*	*Sp*	*MCC*
NetworKIN 1.1								
AGC/AKT	**98.87%**	**58.89%**	**99.44%**	**0.5865**	**98.81%**	**52.22%**	**99.47%**	**0.5445**
AGC/PKA	92.24%	32.20%	93.28%	0.1277	92.67%	57.56%	93.28%	0.2481
Atypical/PIKK/ATM	**97.50%**	**77.42%**	**97.83%**	**0.5204**	**97.47%**	**75.27%**	**97.83%**	**0.5080**
CAMK/CAMK2	98.43%	5.85%	99.91%	0.1671	98.51%	10.64%	99.91%	0.2580
CMGC/MAPK	93.45%	65.34%	94.06%	0.3312	93.60%	71.12%	94.09%	0.3614
CMGC/CDK/CDC2[a]	94.58%	46.67%	95.69%	0.2825	94.75%	54.04%	95.69%	0.3268
Other/CK2	**90.17%**	**55.49%**	**91.14%**	**0.2511**	**90.25%**	**54.85%**	**91.24%**	**0.2495**
TK/EGFR	85.26%	12.16%	96.97%	0.1554	91.42%	56.76%	96.97%	0.6059
TK/Src	90.04%	23.97%	95.40%	0.2141	90.69%	32.64%	95.40%	0.2956
NetworKIN 2.0 beta								
AGC/AKT	98.55%	42.78%	99.33%	0.4436	98.71%	54.44%	99.33%	0.5334
AGC/PKA	96.02%	32.44%	97.11%	0.2110	96.31%	50.24%	97.11%	0.3244
Atypical/PIKK/ATM	93.90%	91.40%	93.94%	0.4064	94.60%	90.32%	94.67%	0.4246
CAMK/CAMK2	91.18%	27.13%	92.20%	0.0881	91.56%	49.47%	92.23%	0.1866
CMGC/MAPK	96.80%	1.59%	98.87%	0.0063	97.43%	31.08%	98.87%	0.3280
CMGC/CDK/CDC2	97.29%	0.35%	99.54%	–0.0023	97.46%	7.72%	99.54%	0.1377
Other/AUR/AUR-A	91.61%	32.73%	92.66%	0.1244	92.15%	54.55%	92.82%	0.2291
Other/CK2	**80.06%**	**75.32%**	**80.19%**	**0.2202**	**79.65%**	**60.34%**	**80.19%**	**0.1619**
TK/ABL	89.00%	13.33%	96.72%	0.1449	90.54%	30.00%	96.72%	0.3323
TK/EGFR	84.70%	12.16%	96.32%	0.1362	91.42%	60.81%	96.32%	0.6162
TK/Src	90.88%	2.07%	98.09%	0.0029	92.18%	19.42%	98.09%	0.2611
TK/Syk	86.67%	2.08%	99.68%	0.0806	89.17%	20.83%	99.68%	0.4051

the 1079 PKs and 9247 substrates for the 44,290 p-sites, with an average of 4.2 upstream PKs per p-site and 57.6 substrates per PK (Table III and supplemental Table S11). The number of predicted phosphorylated proteins and p-sites for each PK were summarized in supplemental Table S12. For example, the PPNs in *S. cerevisiae* (Fig. 2*A*) and *H. sapiens* (Fig. 2*B*) were modeled and shown. The human PPN contains 113,923 ssKSRs among the 380 PKs and 4140 targets for the 22,817 p-sites, with an average of 5.0 upstream PKs per p-site and 98.7 substrates per PK (Table III and Fig. 2*B*). Using pairwise comparison we found only 653 (3.1%) and 1,291 (0.7%) conserved ssKSRs in *S. cerevisiae* and *H. sapiens*, respectively (supplemental Table S13). Thus, eukaryotic phosphorylation is poorly conserved at the site and substrate levels. And this result is consistent with a previous analysis (17, 25).

By comparison, we found only one KSR of the kinase BUR1/CDK9, which phosphorylates the transcription elongation factor protein Spt5, to be conserved in all five eukaryotic phosphoproteomes (Fig. 3). In *S. cerevisiae*, BUR1 might phosphorylate Spt5 at the single site of S136 in its N terminus, whereas its ortholog, CDK9, can phosphorylate up to 11 sites in humans (Fig. 3). We observed that fly Spt5 and S76, along with mouse and human Spt5 S19, are also phosphorylated (Fig. 3). Although the p-sites are not in the same position after sequence alignment, the Spt5 N terminus phosphorylation by BUR1/CDK9 might be a conserved mechanism. By sequence alignment, we observed that only the two p-sites of T806 and T814 in human are conserved in all five species (Fig. 4). From these results, it is proposed that segment around the last KOW (Kyprides, Ouzounis, Woese) domain might be the hotspot (from S666 to T814 in humans) for phosphorylation by CDK9 (Fig. 3). Furthermore, recent progress suggests that the transcription elongation factor Spt5 can be *in vivo* phosphorylated by BUR1 to recruit the polymerase-associated factor (PAF) complex in yeast (36, 37). The *in vitro* experiments

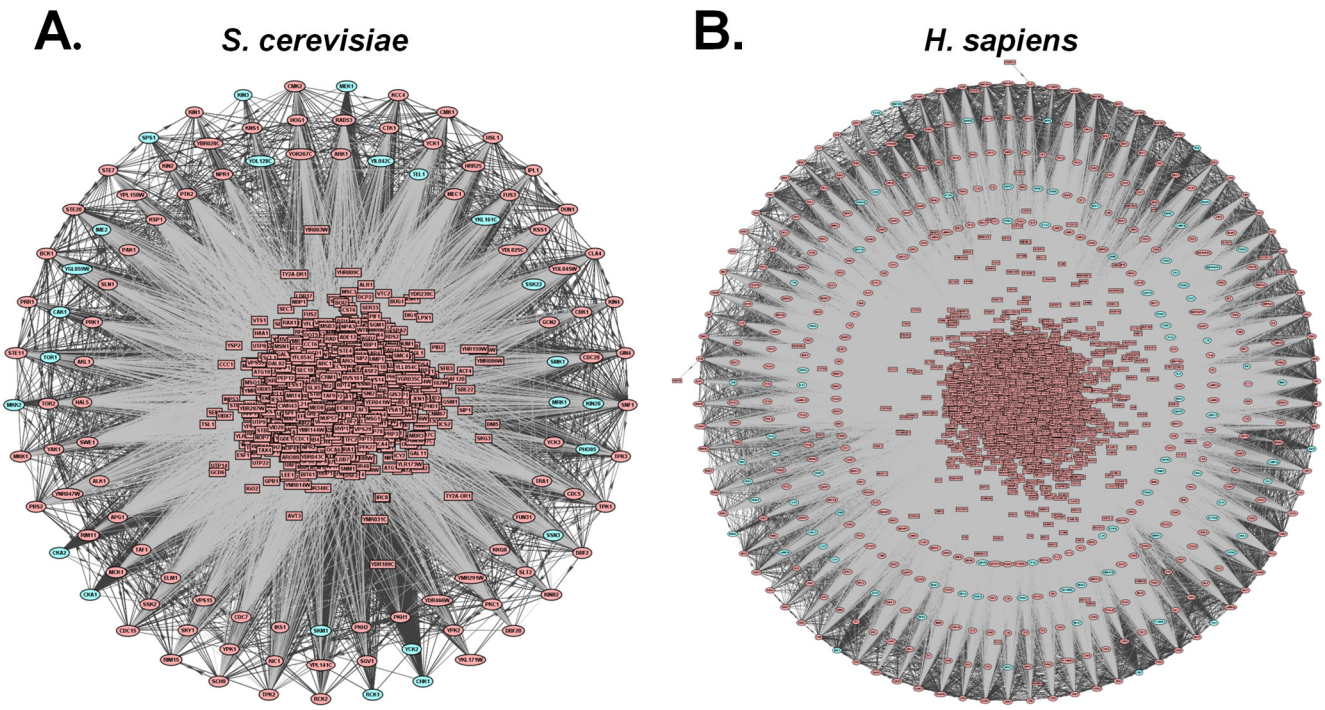

FIG. 2. **Eukaryotic PPNs were modeled and visualized based on predicted ssKSRs from the phosphoproteomic data.** *A*, The PPN of *S. cerevisiae* contains 20,909 ssKSRs among 91 PKs and 1,598 substrates for the 7,041 p-sites; *B*, The PPN of *H. sapiens* contains 113,923 ssKSRs among the 380 PKs and 4,140 targets for the 22,817 p-sites.

indicate that the C-terminal repeat domain (CTD) of yeast Spt5 is potentially responsible for modification, whereas the *in vivo* verification still remains to be obtained (36, 37). However, the yeast Spt5 CTD is not conserved in other species. Whether Spt5 CTD in higher eukaryotes is phosphorylated by CDK9 remains to be verified *in vivo*. In this regard, BUR1/CDK9 might phosphorylate Spt5 in a quite complicated manner. The predictions might be useful for further experimental investigation.

Systematic Analysis of the Human Liver Phosphoproteome and PPN—The liver is the largest internal organ in the human body. Beyond digestion, it plays a variety of essential roles in cell proliferation/differentiation, catabolism/metabolism, embryonic development, detoxification, drug pharmacokinetics, and so on (38). In 2002, the Human Liver Proteome Project (HLPP) was started as the first initiative for the understanding organ- or tissue-specific proteomes in humans (38). The recent completion of proteome, transcriptome, CHIP and massively parallel signature sequencing (MPSS) studies from the Chinese human liver proteome project (CNHLPP) revealed that tens of thousands of proteins/genes are regulated in the human liver (39, 40). Determining how these proteins/genes are temporally and spatially modulated is still a great challenge, and the functional roles of PTMs in the liver remain to be elucidated.

In this work, we conducted a global phosphorylation analysis of the human liver, and experimentally identified 10,644 nonredundant phosphopeptides (Fig. 5*A*). The excellent performance of the new RP-RPLC approach on the in-depth phosphorylation analysis was shown by the uniform distribution of the identified phosphopeptides throughout the first dimension (Fig. 5*B*). In contrast with our previous strategy that only phosphopeptides identified by both MS2 and MS3 in the MS2/MS3 class were retained (15, 16), the classification filtering strategy (34) generated a 31.8% increase in phosphopeptide identification by incorporating phosphopeptides identified exclusively from only the MS2 or MS3 spectra (Fig. 5*C*). In our results, there were 7214 singly (67.8%), 2403 doubly (22.6%), and 1027 triply (9.6%) phosphorylated peptides (Fig. 6*A*). Multiply phosphorylated peptides occupied over 30% of the total identifications (Fig. 6*A*), and this result is consistent with other large-scale phosphoproteomic studies (1, 14). The phosphopeptides were mapped to the UniProt benchmark sequences, and totally 9719 p-sites were identified from 2998 substrates. The distribution of phosphoserine (pS), phosphothreonine (pT) and phosphotyrosine (pY) sites is 85.3% (8,294), 12.9% (1,249) and 1.8% (176) respectively (Fig. 6*B*), and the result is similar with a previous study although different samples were used (1). More details on the phosphopeptide analysis are present in supplemental Results and supplemental Figs. S3-S6.

By comparing with 60,816 known p-sites collected from heterogeneous resources, we observed that 5818 (59.9%) p-sites with 5063 pS, 664 pT and 91 pY sites have been reported previously (Fig. 6*C*). The high percentage of known p-sites indicates that p-sites identified in this study are of high

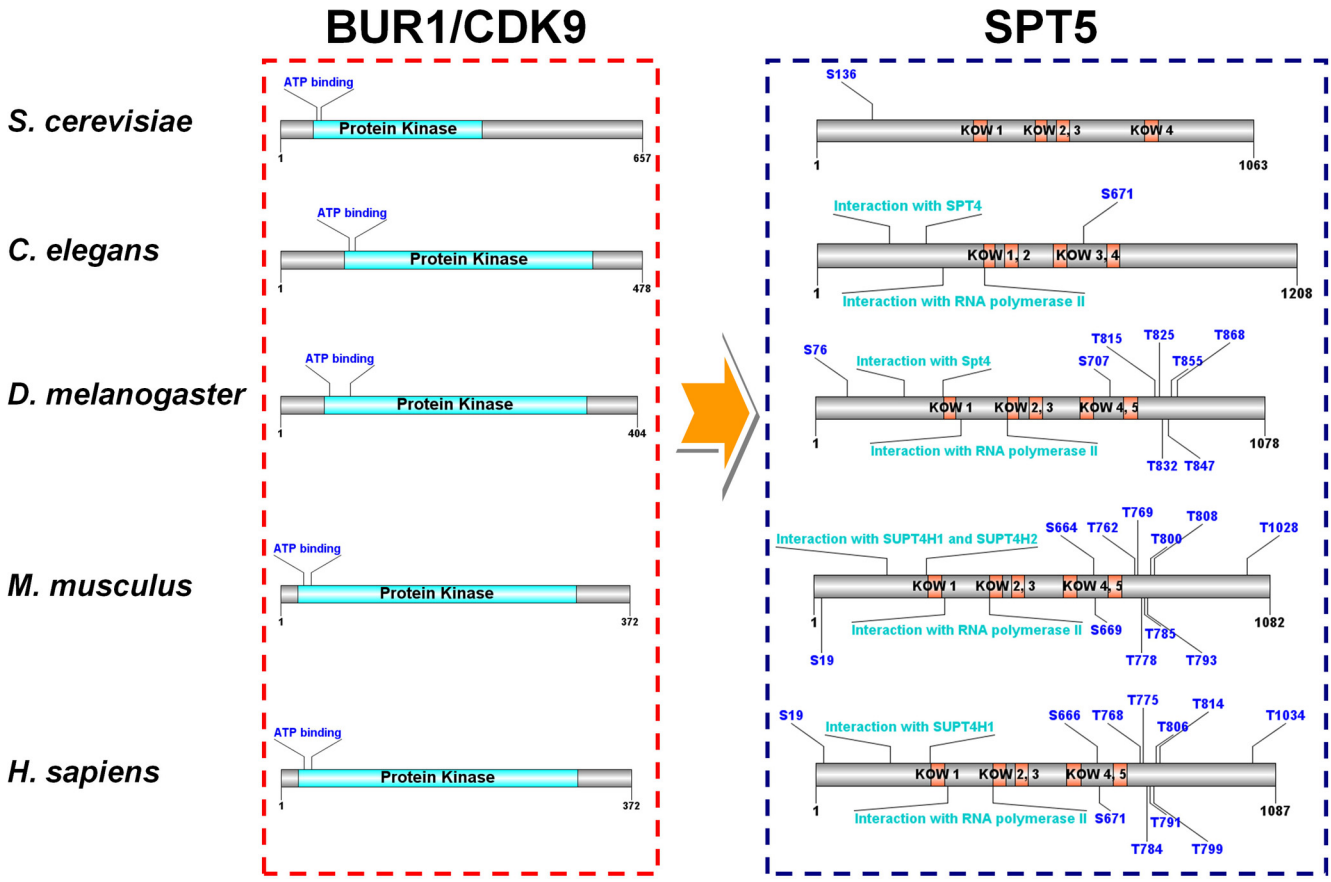

FIG. 3. **Only one KSR of the kinase BUR1/CDK9, which phosphorylates the transcription elongation factor protein Spt5, is conserved in all five eukaryotic PPNs.**

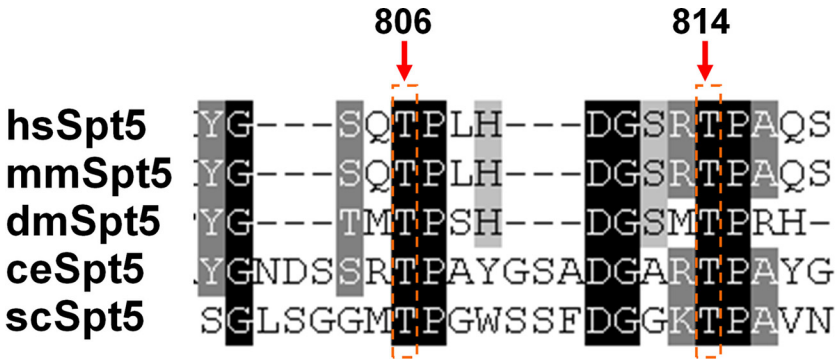

FIG. 4. **By sequence alignment, we observed that the two p-sites of T806 and T814 in human Spt5 are conserved in all five species.** Both sites were predicted to be phosphorylated by BUR1/CDK9.

confidence. Using iGPS, we predicted a potential PPN containing 12,819 potential ssKSRs among 350 PKs and 962 substrates for 2633 p-sites from the human liver phosphoproteome, with a coverage rate of 27.1% (supplemental Table S14). From the top 10 PKs with the most predicted p-sites, we observed that up to 6 PKs belong to the CMGC PK group, such as Erk1, CDK2, Erk2, and GSK3A, p38a and CDC2 (supplemental Table S14). Although these PKs were predicted to phosphorylate more p-sites, we speculated whether some PKs preferentially modify more or fewer p-sites in the liver PPN against the whole human PPN. To address this problem, we performed the Yates' chi-squared (χ^2) test with the 2×2 contingency table method (41) (for more details see supplemental Experimental Procedures). Totally, we observed that 60 PKs significantly modify more p-sites, whereas 67 PKs preferentially modify fewer p-sites (p value < 0.01, see supplemental Table S15). The top 10 PKs with significantly over- or under-represented p-sites in the human liver PPN were shown in Table V, respectively. Previously, experimental studies revealed that AKT family PKs play a predominantly regulatory role in regulating the hepatic gluconeogenesis (42, 43). In the results, we observed that the p-sites of AKT1 are significantly over-represented with an enrichment ratio of 1.55 (p value = 3.04E-19) (Table V). Also, it was reported that the

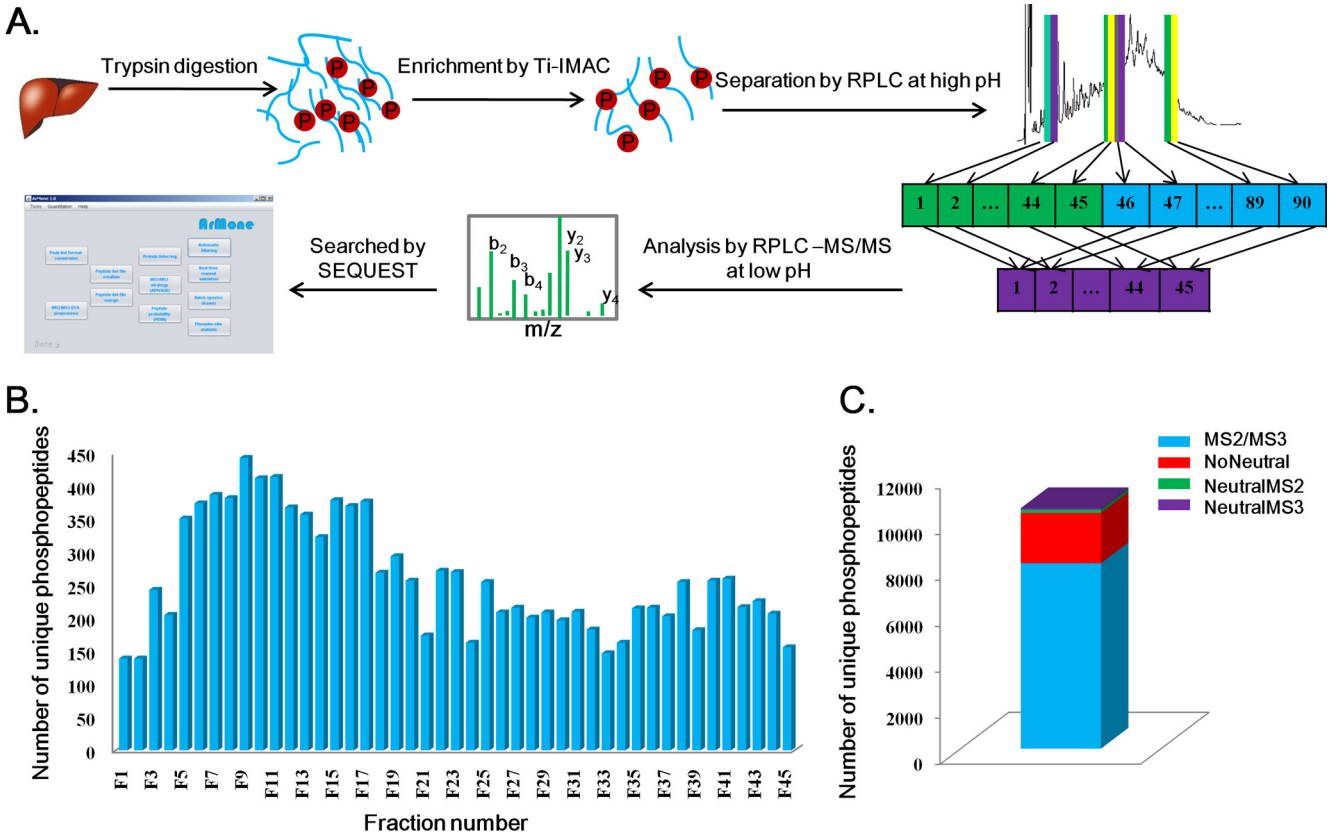

FIG. 5. **Overview of the strategy used for the large-scale phosphorylation analysis of the human liver.** *A*, Scheme for sample preparation and data processing of the human liver. *B*, Distribution of identified unique phosphopeptides from 45 pooled fractions. *C*, Distribution of identified unique phosphopeptides in four groups based on the classification filtering strategy (26).

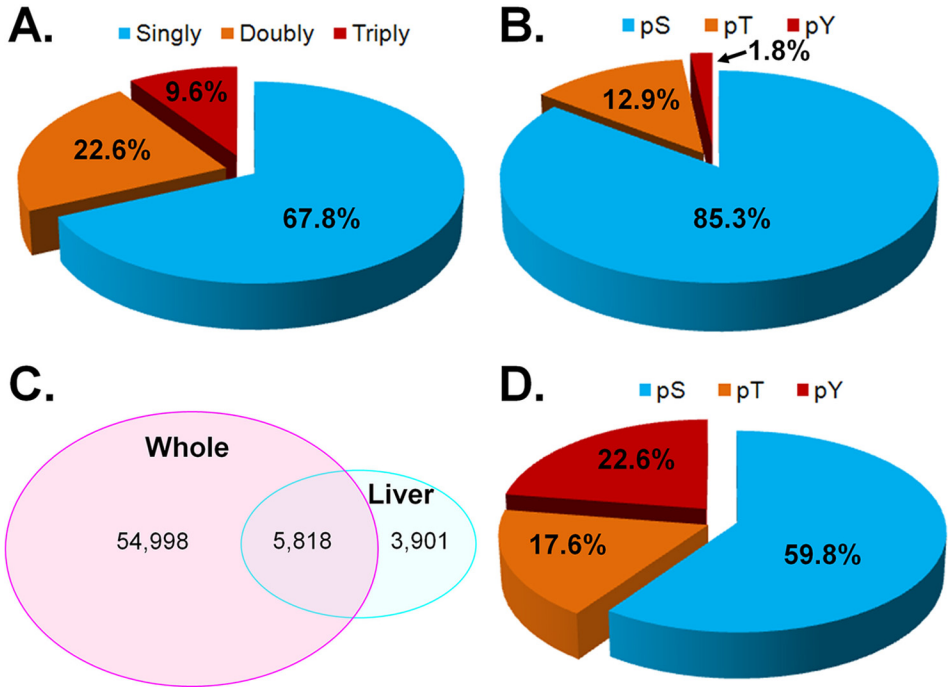

FIG. 6. **The data statistics of the human liver phosphoproteomic analysis.** *A*, The distribution of singly, doubly, and triply phosphorylated peptides; *B*, The distribution of pS, pT and pY sites in the human liver phosphoproteome; (*C*) By comparing with known information (Whole), up to 5818 p-sites have been reported previously; *D*, The distribution of pS, pT and pY sites in the whole human phosphoproteome.

TABLE V
Top 10 PKs with significantly over- or under-represented p-sites in human liver PPN, respectively

PK Name	Uniprot[a]	Predictor in iGPS	String & Exp. PPI				
			Protein	Site	E-ratio[b]	χ^{2c}	*p* value[d]
PKs with over-represented p-sites							
CK2a2	P19784	Other/CK2	123	420	1.87	154.92	1.46E-35
CLK1	B4DFW7	CMGC	21	141	2.82	127.27	1.63E-29
AKT1	P31749	AGC/AKT	244	428	1.55	80.41	3.04E-19
MSK1	O75582	AGC/RSK	44	96	2.17	49.46	2.02E-12
RSK4	Q9UK32	AGC/RSK	25	69	2.41	44.96	2.01E-11
MSK2	O75676	AGC/RSK	28	72	2.24	39.80	2.81E-10
PKN2	Q16513	AGC	43	93	1.93	34.87	3.52E-09
p70S6K	P23443	AGC/RSK	76	137	1.67	32.18	1.40E-08
CK1d	P48730	CK1	42	129	1.69	31.24	2.28E-08
RSK1	Q15349	AGC/RSK	27	68	2.07	30.51	3.31E-08
PKs with under-represented p-sites							
SRC	P12931	TK/Src	19	21	0.13	128.44	8.99E-30
EGFR	P00533	TK/EGFR	8	9	0.08	91.75	9.84E-22
FYN	P06241	TK/Src	9	11	0.11	79.07	5.98E-19
ErbB2	P04626	TK/EGFR	6	6	0.07	70.10	5.63E-17
LCK	P06239	TK/Src	8	10	0.13	61.82	3.76E-15
INSR	P06213	TK/InsR	6	6	0.09	54.70	1.41E-13
KIT	P10721	TK/PDGFR	4	4	0.07	54.19	1.82E-13
PDGFRb	P09619	TK/PDGFR	9	10	0.15	49.85	1.66E-12
IGF1R	P08069	TK/InsR	3	3	0.06	48.53	3.25E-12
ErbB3	P21860	TK/EGFR	2	2	0.04	48.29	3.68E-12

[a] Uniprot, Uniprot accession number.
[b] E-ratio, enrichment ratio, the liver PPN proportion divided by the whole human PPN proportion.
[c] The result of the Yates' chi-squared (χ^2) test.
[d] p value < 0.01.

CK2 activity increases after hepatectomy or laparatomy (44). Thus, the significantly enriched p-sites for CK2 might be attributed to the enhanced activity (E-ratio = 1.87, *p* value = 1.46E-35, Table V). Furthermore, ribosomal S6 kinases (RSKs) can be activated by hepatotoxin CCl4 on liver injury (45). In the results, the p-sites of up to five members of AGC/RSK group such as MSK1, RSK4, MSK2, p70S6K and RSK1 are significantly enriched (Table V). Because of the relatively low abundance of tyrosine phosphorylation, the number of identified pY sites from large-scale studies will be limited without specific enrichment through anti-pY antibodies (14). However, in our data set, the distribution of pY in the whole human phosphoproteome is 22.6% (Fig. 6*D*), with an order of magnitude higher than in liver (1.8%) (Fig. 6*B*). Thus, it can be expected that the p-sites of most TKs will be under-represented, because tyrosine-specific strategies were not used in our analysis. Indeed, no TKs were detected with over-represented p-sites, whereas 57 TKs (85.1%) preferentially modify fewer p-sites (Table V, supplemental Table S15).

DISCUSSION

In the post-genomic era, the dissection of the functional complexity and diversity of the proteome has emerged as an urgent challenge. In particular, proteins are transiently and dynamically regulated by hundreds of PTMs *in vivo*, which adds a dimension of functional complexity. As one of the most essential PTMs, phosphorylation has attracted considerable attention for its functional importance (1–3). Investigating phosphorylation at the systemic level can help in the understanding of its molecular mechanisms and regulatory activity (8–10). Rapid progresses in phosphoproteomics using phosphopeptide enrichment and HTP-MS techniques have detected tens of thousands of potential *in vivo* p-sites with high confidence (1, 14, 22). However, deeper analysis of these un-annotated p-sites to allow elucidation of ssKSRs in eukaryotes is lacking and at present is hampered by limited computational methods. In contrast with previous studies that focused exclusively on humans (8, 9, 17), here we designed a general and integrative approach to predict *in vivo* ssKSRs in five eukaryotic species. In iGPS 1.0, the GPS 2.0 algorithm was used to predict potential PKs for un-annotated p-sites (24), and both experimentally identified and pre-predicted PPI information was adopted for further filtering of false-positive hits. Extensive evaluations and comparisons suggest the prediction performance to be promisingly accurate and better than NetworKIN (8, 9) (Table IV).

With this powerful tool, we systematically predicted potentially ssKSRs and modeled PPNs from eukaryotic phosphoproteomic data. The total predictive coverage is 30.4% (44,290/145,646), which is a great amount of information for experimentalists. Among the top 10 PKs with the most p-sites in the five eukaryotic phosphoproteomes, we observed up to 33 PKs (66.7%) that belong to the CMGC group (supplemental Table S12). The PKs in the CMGC group are implicated in the cell cycle/cell division (*e.g.* CMGC/CDK) and signal transduction (*e.g.* CMGC/MAPK and CMGC/GSK) pathways (from the GO annotations in the UniProt database), which is con-

sistent with the major roles of phosphorylation (1–3, 17). Three potential hypotheses may be offered to interpret this observation. First, the prediction might be influenced by the GPS 2.0 algorithm at the p-site level such that better performance can generate more kinase-specific p-sites (24). In GPS 2.0, the *Ac*, *Sn*, *Sp,* and *MCC* of CMGC/MAPK are 86.05%, 91.21%, 85.94%, and 0.2950, respectively, whereas the performance of AGC/GRK is 92.46% (*Ac*), 94.05% (*Sn*), 92.37% (*Sp*), and 0.5999 (*MCC*) (24). Although the accuracy of AGC/GRK is much better than CMGC/MAPK, no GRK members are included in the top 10 PKs with the most p-sites in any organism. Also, although the *Ac*, *Sn*, *Sp,* and *MCC* of Atypical/PIKK/ATM are 94.47%, 100.00%, 94.38%, and 0.4451, respectively, the human ATM is not contained in the top 10 PKs with the most p-sites. Thus, this result is not caused by a bias from the GPS 2.0 prediction. Second, the number of PPIs might influence the prediction at the substrate level such that more PPIs lead to a greater number of predicted substrates. We counted the number of PPIs for each PK, and only 14 CMGC PKs (28%) belong to the top 10 PKs with the most PPIs in the five species (supplemental Table S16). Again, although the number of GSK3A interacting proteins in humans is only the 28th in rank (supplemental Table S16), it is one of the top 10 PKs with the most p-sites in this study (supplemental Table S12). Although the number of ATM binding proteins ranks 8th in humans (supplemental Table S16), it is not included in the top 10 PKs with the most p-sites (supplemental Table S12). In this regard, this observation is not caused by a bias from the PPI filter. Finally, this prediction might reflect the *bona fide* status that most of the p-sites were phosphorylated and regulated by the CMGC group PKs. In addition, by analyzing the human liver phosphoproteome, we observed a similar result that 6 of the top 10 PKs with the most p-sites belong to CMGC PKs (supplemental Table S14). Taken together, it is proposed that CMGC PKs play a predominant role in regulating cellular phosphorylation.

Although CMGC PKs play a general role for the phosphorylation, several PKs in a distinct sample might preferentially modify more or fewer p-sites to ensure the precise regulation. By the statistical analysis and comparison of predicted results of human liver and whole PPNs, we observed that a considerable number of PKs significantly regulate more or fewer p-sites in human liver PPN (supplemental Table S15). Beyond the results that are consistent with previous analyses, our study suggested that a number of PKs, such as CLK1, PKN2, and CK1d, also play a potentially important role in the human liver PPN (Table V). In 2007, Villen *et al.* experimentally identified thousands of p-sites from a 21-day-old mouse liver (14). By collecting 6089 p-sites in 2209 mouse liver proteins (14), we predicted 4502 ssKSRs among the 308 PKs and 543 proteins for 1176 p-sites, with a coverage rate of 19.3% (supplemental Table S17). However, we only detected 13 and 9 PKs with significantly over- or under-represented p-sites with the Yates' chi-squared (χ^2) test (p value < 0.01, supplemental Table S18) (41). And the statistical significance is much lower against the result in the human liver PPN (supplemental Table S15). In this regard, we proposed that our results might be more useful for further studying hepatic functions in *H. sapiens*.

Our approach can be generally used to identify potential *in vivo* ssKSRs in eukaryotes. The total predictive coverage is 30.4% (44,290/145,646) (Table III), which is a great amount of information for experimentalists. We anticipate that more efficient contextual filters will be integrated into this method over time to improve its prediction ability. This study can serve as a starting point for the general analysis of the various PTM-regulating proteomes, not limited to phosphorylation.

Acknowledgments—We thank Dr. Francesca Diella and Dr. Toby J. Gibson (EMBL) for always providing the new data set of the Phospho.ELM database during the past seven years. We thank Dr. Peter Hornbeck (Cell Signaling Technology, USA) for providing the PhosphoSitePlus data set on July 14, 2009. We thank Dr. Ralf Mrowka (Charité, Germany) for providing a Java applet for visualizing protein-protein interaction. We thank Drs. Hong Li and Guohui Ding (SIBS, China) for providing the SysPTM data set. We thank Dr. Rune Linding (ICR, UK) and Dr. Martin Lee Miller (Univ. of Copenhagen) for personal communications on computational phosphorylation. We thank Dr. Jing-Dong Jackie Han (IGDB), Dr. Edwin Wang (NRC, Canada), Dr. Xuegong Zhang (Tsinghua Univ.), Dr. Dong Li (BPRC), and Dr. Houjiang Zhou (The Heck Lab, Netherlands) for their helpful comments on network analysis. We thank Dr. Felix Cheung (Nature China) for his encouragement and helpful suggestions on presentation. Nature Publishing Group Language Editing (NPG Language Editing) and Pacific Edit reviewed the manuscript prior to submission. We also thank the anonymous reviewer, whose suggestions have greatly improved the presentation of this manuscript.

* This work was partly funded by the National Basic Research Program (973 project) (2012CB910101, 2010CB945401, 2012CB911201), the Creative Research Group Project by NSFC (21021004), the National Key Special Program on Infection diseases (2012ZX10002009-011), the Analytical Method Innovation Program of MOST (2009IM031800, 2010IM030500), National Natural Sciences Foundation of China (31171263, 20735004, 30830036, 30900835, 31071154, 91019020).

S This article contains supplemental Procedures, Results, Tables S1 to S20, and Figs. S1 to S6.

The authors have declared no conflict of interest.

** To whom correspondence should be addressed: Tel.: +86-411-84379610, Fax: +86-411-84379620, E-mail: hanfazou@dicp.ac.cn; Tel.: +86-27-87793903, Fax: +86-27-87793172, E-mail: xueyu@mail.hust.edu.cn; or Tel./Fax: +86-20-39943788, E-mail: renjian.sysu@gmail.com.

‡‡ Both authors contributed equally to this work.

REFERENCES

1. 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,** 635–648
2. Ubersax, J. A., and Ferrell, J. E., Jr. (2007) Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8,** 530–541
3. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science* **298,** 1912–1934
4. Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P., and Brinkworth, R. I.

(2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta* **1754,** 200–209
5. Kreegipuu, A., Blom, N., Brunak, S., and Järv, J. (1998) Statistical analysis of protein kinase specificity determinants. *FEBS Lett.* **430,** 45–50
6. Songyang, Z., Lu, K. P., Kwon, Y. T., Tsai, L. H., Filhol, O., Cochet, C., Brickey, D. A., Soderling, T. R., Bartleson, C., Graves, D. J., DeMaggio, A. J., Hoekstra, M. F., Blenis, J., Hunter, T., and Cantley, L. C. (1996) A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Mol. Cell. Biol.* **16,** 6486–6493
7. Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., and Cantley, L. C. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.* **19,** 348–353
8. Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A., Jorgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., and Pawson, T. (2007) Systematic discovery of in vivo phosphorylation networks. *Cell* **129,** 1415–1426
9. Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M. B., and Pawson, T. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* **36,** D695–D699
10. Tan, C. S., and Linding, R. (2009) Experimental and computational tools useful for (re)construction of dynamic kinase-substrate networks. *Proteomics* **9,** 5233–5242
11. Lahiry, P., Torkamani, A., Schork, N. J., and Hegele, R. A. (2010) Kinase mutations in human disease: interpreting genotype-phenotype relationships. *Nat. Rev. Genet.* **11,** 60–74
12. Diella, F., Gould, C. M., Chica, C., Via, A., and Gibson, T. J. (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.* **36,** D240–D244
13. Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., McCartney, R. R., Schmidt, M. C., Rachidi, N., Lee, S. J., Mah, A. S., Meng, L., Stark, M. J., Stern, D. F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P. F., and Snyder, M. (2005) Global analysis of protein phosphorylation in yeast. *Nature* **438,** 679–684
14. Villén, J., Beausoleil, S. A., Gerber, S. A., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 1488–1493
15. Han, G., Ye, M., Zhou, H., Jiang, X., Feng, S., Jiang, X., Tian, R., Wan, D., Zou, H., and Gu, J. (2008) Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography. *Proteomics* **8,** 1346–1361
16. Han, G., Ye, M., Liu, H., Song, C., Sun, D., Wu, Y., Jiang, X., Chen, R., Wang, C., Wang, L., and Zou, H. (2010) Phosphoproteome analysis of human liver tissue by long-gradient nanoflow LC coupled with multiple stage MS analysis. *Electrophoresis* **31,** 1080–1089
17. Tan, C. S., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M. O., Jørgensen, C., Bader, G. D., Aebersold, R., Pawson, T., and Linding, R. (2009) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci. Signal.* **2,** ra39
18. Xu, C. F., Lu, Y., Ma, J., Mohammadi, M., and Neubert, T. A. (2005) Identification of phosphopeptides by MALDI Q-TOF MS in positive and negative ion modes after methyl esterification. *Mol. Cell. Proteomics* **4,** 809–818
19. Steen, H., Jebanathirajah, J. A., Rush, J., Morrice, N., and Kirschner, M. W. (2006) Phosphorylation analysis by mass spectrometry: myths, facts, and the consequences for qualitative and quantitative measurements. *Mol. Cell. Proteomics* **5,** 172–181
20. Li, X., Gerber, S. A., Rudner, A. D., Beausoleil, S. A., Haas, W., Villén, J., Elias, J. E., and Gygi, S. P. (2007) Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J. Proteome Res.* **6,** 1190–1197
21. Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, E. R., 3rd, Hurov, K. E., Luo, J., Bakalarski, C. E., Zhao, Z., Solimini, N., Lerenthal, Y., Shiloh, Y., Gygi, S. P., and Elledge, S. J. (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316,** 1160–1166
22. Song, C., Ye, M., Han, G., Jiang, X., Wang, F., Yu, Z., Chen, R., and Zou, H. (2010) Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. *Anal. Chem.* **82,** 53–56
23. Xue, Y., Gao, X., Cao, J., Liu, Z., Jin, C., Wen, L., Yao, X., and Ren, J. (2010) A summary of computational resources for protein phosphorylation. *Curr. Protein Pept. Sci.* **11,** 485–496
24. Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics* **7,** 1598–1608
25. Boekhorst, J., van Breukelen, B., Heck, A., Jr., and Snel, B. (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.* **9,** R144
26. Jiang, X., Ye, M., Cheng, K., and Zou, H. (2010) ArMone: a software suite specially designed for processing and analysis of phosphoproteome data. *J. Proteome Res.* **9,** 2743–2751
27. Bodenmiller, B., Campbell, D., Gerrits, B., Lam, H., Jovanovic, M., Picotti, P., Schlapbach, R., and Aebersold, R. (2008) PhosphoPep–a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **26,** 1339–1340
28. Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5,** 79
29. Li, H., Xing, X., Ding, G., Li, Q., Wang, C., Xie, L., Zeng, R., and Li, Y. (2009) SysPTM - a systematic resource for proteomic research of post-translational modifications. *Mol. Cell. Proteomics* **8,** 1839–1849
30. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., and Zhang, B. (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4,** 1551–1561
31. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pander, A. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37,** D767–D772
32. Wang, F., Chen, R., Zhu, J., Sun, D., Song, C., Wu, Y., Ye, M., Wang, L., and Zou, H. (2010) A Fully Automated System with Online Sample Loading, Isotope Dimethyl Labeling and Multidimensional Separation for High-Throughput Quantitative Proteome Analysis. *Anal. Chem.* **82,** 3007–3015
33. Beausoleil, S. A., Villén, 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,** 1285–1292
34. Jiang, X., Ye, M., Han, G., Dong, X., and Zou, H. (2010) Classification filtering strategy to improve the coverage and sensitivity of phosphoproteome analysis. *Anal. Chem.* **82,** 6168–6175
35. Gao, X., Jin, C., Ren, J., Yao, X., and Xue, Y. (2008) Proteome-wide prediction of PKA phosphorylation sites in eukaryotic kingdom. *Genomics* **92,** 457–463
36. Zhou, K., Kuo, W. H., Fillingham, J., and Greenblatt, J. F. (2009) Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 6956–6961
37. Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W. H., Ranish, J., Shokat, K. M., and Hahn, S. (2009) Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. *Mol. Cell. Biol.* **29,** 4852–4863
38. He, F. (2005) Human liver proteome project: plan, progress, and perspectives. *Mol. Cell. Proteomics* **4,** 1841–1848
39. Sun, A., Jiang, Y., Wang, X., Liu, Q., Zhong, F., He, Q., Guan, W., Li, H., Sun, Y., Shi, L., Yu, H., Yang, D., Xu, Y., Song, Y., Tong, W., Li, D., Lin, C., Hao, Y., Geng, C., Yun, D., Zhang, X., Yuan, X., Chen, P., Zhu, Y., Li, Y., Liang, S., Zhao, X., Liu, S., and He, F. (2010) Liverbase: a comprehensive view of human liver biology. *J. Proteome Res.* **9,** 50–58
40. He, F., Chung, M. C., and Jordan, T. W. (2010) Chinese human liver proteome project: a pathfinder of HUPO human liver proteome project. *J. Proteome Res.* **9,** 1–2

41. Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J., and Xue, Y. (2011) GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Mol. Biosyst.* **7,** 1197–1204
42. Puigserver, P., Rhee, J., Donovan, J., Walkey, C. J., Yoon, J. C., Oriente, F., Kitamura, Y., Altomonte, J., Dong, H., Accili, D., and Spiegelman, B. M. (2003) Insulin-regulated hepatic gluconeogenesis through FOXO1-PGC-1alpha interaction. *Nature* **423,** 550–555
43. Du, K., Herzig, S., Kulkarni, R. N., and Montminy, M. (2003) TRB3: a tribbles homolog that inhibits Akt/PKB activation by insulin in liver. *Science* **300,** 1574–1577
44. Pancetti, F., Bosser, R., Itarte, E., and Bachs, O. (1996) Changes in the activity of nuclear protein kinase CK2 during rat liver regeneration. *Biochem. Biophys. Res. Commun.* **218,** 35–39
45. Buck, M., Poli, V., Hunter, T., and Chojkier, M. (2001) C/EBPbeta phosphorylation by RSK creates a functional XEXD caspase inhibitory box critical for cell survival. *Mol. Cell* **8,** 807–816