Large-scale phosphorylation analysis of mouse liver

Judit Villén, Sean A. Beausoleil, Scott A. Gerber*, and Steven P. Gygi†

Department of Cell Biology, Harvard Medical School, Boston, MA 02115

Communicated by Lewis C. Cantley, Harvard Institutes of Medicine, Boston, MA, November 6, 2006 (received for review September 25, 2006)

Protein phosphorylation is a complex network of signaling and regulatory events that affects virtually every cellular process. Our understanding of the nature of this network as a whole remains limited, largely because of an array of technical challenges in the isolation and high-throughput sequencing of phosphorylated species. In the present work, we demonstrate that a combination of tandem phosphopeptide enrichment methods, high performance MS, and optimized database search/data filtering strategies is a powerful tool for surveying the phosphoproteome. Using our integrated analytical platform, we report the identification of 5,635 nonredundant phosphorylation sites from 2,328 proteins from mouse liver. From this list of sites, we extracted both novel and known motifs for specific Ser/Thr kinases including a ''dipolar'' motif. We also found that C-terminal phosphorylation was more frequent than at any other location and that the distribution of potential kinases for these sites was unique. Finally, we identified double phosphorylation motifs that may be involved in ordered phosphorylation.

mass spectrometry | proteomics

In eukaryotes, protein phosphorylation is among the most important regulatory events in cells, guiding primary biological processes, such as cell division, growth, migration, differentiation, and intercellular communication (1, 2). Numerous efforts to study protein phosphorylation continue to provide the scientific community with an expanding phosphorylation database resource. Sequence alignment studies (3) have defined which genes encode protein kinases. In addition, *in vitro* reactions of kinases with synthetic peptide libraries have been used to define kinase specificities (4, 5). Finally, protein targets for individual kinases are being defined through elegant systematic studies (6–8). Notwithstanding, only MS-based proteomics currently provides the capability to identify at once and on a massive scale kinase substrates and the specific positions of their modification (9).

A major goal of systems biology is to integrate all *in vivo* phosphorylation events into the context of an organism. Phosphorylation analysis from primary tissue, in contrast to immortalized cell lines, best represents events that are occurring in the basal physiological state of an organism even though tissues often contain heterogeneous populations of cells. In the present study we chose mouse liver as a model tissue for establishing a pipeline for large-scale phosphorylation analysis. In addition, protein phosphorylation plays a critical role in normal liver development and function; therefore, the sites obtained here could be further characterized into physiological context. Several phosphorylation-related (PI3K and Akt signaling) liver phenotypes have been reported to be related to altered lipid and glucose metabolism via insulin control (10, 11) and liver regeneration (ribosomal protein S6) (12). Previously, only two studies have examined phosphorylation sites from liver tissue with 26 (13) and 339 (14) sites.

As the field of proteomics has matured, considerable attention has been focused on the development of strategies to facilitate the large-scale profiling of phosphorylated species. In a typical large-scale phosphorylation analysis, a preliminary enrichment step of phosphopeptides is essential to reduce sample complexity and increase their relative concentration. A wide variety of phosphopeptide enrichment strategies have been proposed, including chemical approaches using β-elimination or phosphoramidate chemistry (15–17), peptide immunoprecipitation with phospho-specific motif antibodies (18), affinity purification through metal complexation with the phosphate group [immobilized metal ion affinity chromatography (IMAC)] (19), acid–base interaction with TiO_2 (20), solution-charge-based enrichment by strong cation exchange (SCX) chromatography (9), and combinations of these (21, 22).

A primary limitation in conducting large-scale phosphorylation analysis concerns data processing and validation. There are three main issues. First, studies of posttranslational modifications cannot rely on redundant peptide identifications for correctness (i.e., multiple unique peptides assigned to the same protein). This limitation results in the net confidence of identification resting solely on single peptide identifications. Second, during fragmentation for sequence identification, pSer- and pThr-containing peptides can produce fragmentation patterns that are often dominated by products derived from neutral losses of phosphoric acid, which results in suppression of sequence-informative ions and consequently produces lower scores than unmodified peptides during database spectral matching. Third, the presence of multiple Ser, Thr, and Tyr candidate residues in a phosphopeptide can produce ambiguity when assigning the precise site of phosphorylation.

To address these issues, we and others (9, 23–25) have previously resorted to tedious manual validation after database searching, which can be subjective and has become impractical because data sets have grown in size. There is a clear risk in supplying a phosphorylation site without also establishing an associated probability of correct site localization, in particular when attempting to associate a function for that particular modification. We recently addressed this final issue by creating a probability-based score for evaluating ambiguity that provides an assessment of the likelihood that a site is correctly localized (26).

Here we present a combination of procedures for obtaining a large phosphorylation data set from mouse liver, with a defined error rate in phosphopeptide identification and probability assessment for correct site localization. Furthermore, we used this well curated data set to study phosphorylation motifs from singly and multiply phosphorylated peptides.

Results and Discussion

Generation of a Large Phosphorylation Repertoire by MS. The strategy used for large-scale phosphorylation analysis is shown in Fig. 1. Liver tissue from a 21-day-old mouse was lysed, and 90 mg of liver protein was digested with trypsin. To obtain a general phosphorylation data set, 10 mg of the resulting peptides were

Author contributions: S.P.G. designed research; J.V. performed research; S.A.G. contributed new reagents/analytic tools; J.V. and S.A.B. analyzed data; and J.V. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: IMAC, immobilized metal ion affinity chromatography; SCX, strong cation exchange; FP, false-positive; LSR, lipolysis stimulated receptor.

*Present address: Department of Genetics, Dartmouth Medical School, Lebanon, NH 03756.

†To whom correspondence should be addressed at: 240 Longwood Avenue, Harvard Medical School, Boston, MA 02115. E-mail: steven_gygi@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609836104/DC1.

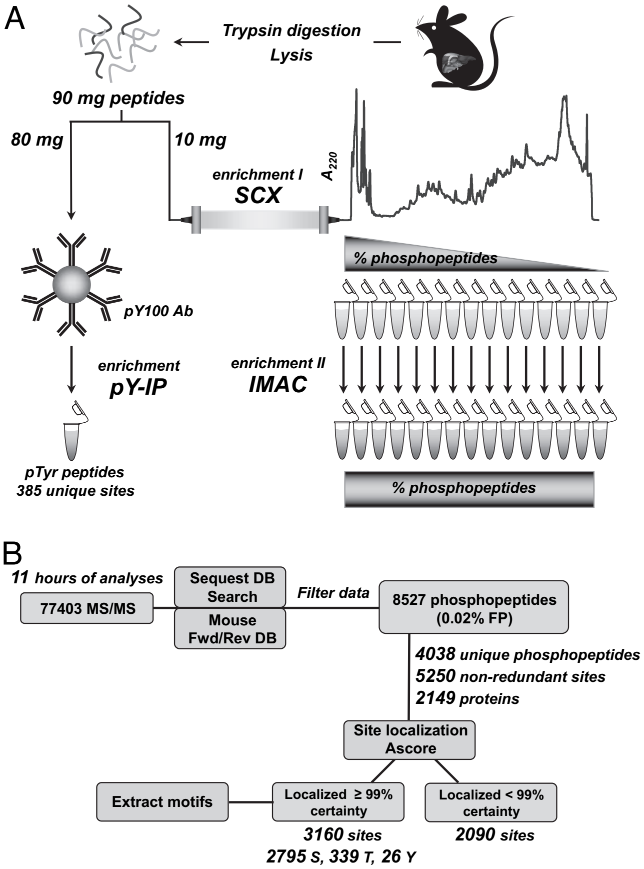

Fig. 1. Strategy used for the large-scale identification and characterization of phosphorylation sites from mouse liver. (*A*) Sample preparation. Tissue homogenization and lysis was followed by trypsin digestion. Tryptic peptides (10 mg) were subjected to a two-step phosphopeptide enrichment. SCX chromatography provided a substantial enrichment in early eluting fractions. Subsequent IMAC of each fraction provided additional selective capture. In addition, 80 mg of tryptic peptides were enriched for pTyr-containing peptides by immunoaffinity purification. (*B*) Data processing for the SCX/IMAC experiment. MS/MS spectra from 15 analyses were searched with Sequest against the mouse protein database (DB) containing both forward (target) and reversed (decoy) sequences for FP calculations. Importantly, the target/decoy database search strategy also provided a means to establish appropriate orthogonal filtering criteria (mass deviation, enzyme specificity, solution charge state at pH 2.65, and Sequest scoring). Only two reversed-sequence peptides were found in the final filtered list of 8,529 phosphopeptides (0.02% FP rate). In total, 5,250 nonredundant sites were identified. The Ascore algorithm (26) was used to determine a probability of correct localization for each individual site. Finally, phosphorylation motifs were extracted from the data set with the Motif-X algorithm (34).

subjected to a two-step phosphopeptide enrichment procedure (Fig. 1*A*) consisting of SCX chromatography followed by IMAC.

We collected 15 fractions along an SCX gradient expanded in the +1 charge state region, where a significant portion of phosphopeptides are expected to elute (9). After IMAC enrichment (19, 22), each fraction was analyzed by LC-MS/MS in a hybrid linear ion trap/Fourier transform ion cyclotron resonance mass spectrometer. High mass accuracy precursor ions were collected as Fourier transform ion cyclotron resonance master spectra, and >77,000 MS/MS spectra were generated by collision-activated dissociation in the linear ion trap. These MS/MS spectra were searched against a composite database of mouse proteins containing sequences first in the forward direction and then in the reverse direction (Fig. 1*B*). This target/decoy database strategy permits a false-positive (FP) rate to be estimated based on the number of reversed-sequence identifications populating the final data set (27, 28). Orthogonal filtering criteria (mass accuracy, tryptic state, solution charge state, and Sequest scoring) were used (26) to establish a final data set with 8,527 phosphopeptides from 2,149 proteins identified at a FP rate of 0.02% (only two reversed-sequence matches were contained within the final peptide list). These peptides contained 5,250 nonredundant phosphorylation sites.

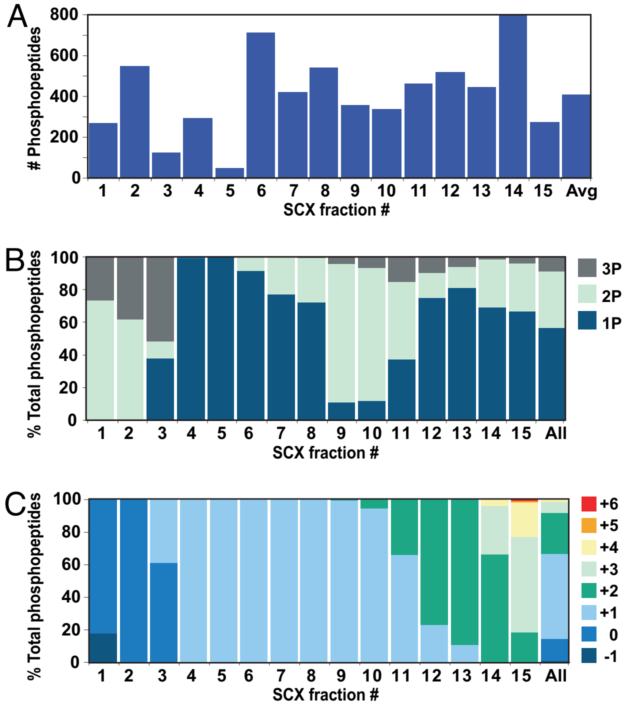

Fig. 2. Distribution of phosphopeptides and their properties across 15 SCX fractions. (*A*) Phosphopeptide distribution. Shown are data for phosphopeptides identified per fraction. (*B*) Phosphorylation sites per peptide. Shown are percentages of phosphopeptides in each fraction containing one, two, or three phosphorylation sites. (*C*) Net solution charge state (pH 2.65). Shown are percentages of phosphopeptides in each fraction with calculated solution charge states between −1 and +6. SCX chromatography separates primarily based on net solution charge state, and each phosphate group subtracts one net charge from a peptide.

As shown in Fig. 2*A*, the entire collection averaged >400 phosphopeptide identifications per fraction, which were primarily separated by solution charge state (Fig. 2*C*). The number of phosphates per peptide is shown in Fig. 2*B*. Overall, the 2-step purification successfully provided samples with >90% phosphopeptides relative to total peptide sequences (data not shown). Supporting information (SI) Table 2 contains the list of all identified phosphopeptides and phosphorylation sites, and all MS/MS spectra are available via hyperlink within this table.

As an example, Table 1 displays six phosphopeptides identified from a single protein, lipolysis stimulated receptor (LSR). Although a total of 23 redundant phosphopeptides were detected for LSR (SI Table 2), all 12 nonredundant sites were contained within this set of six phosphopeptides. LSR is thought to be involved in the clearance of triglyceride-rich lipoproteins (29), and the expression of LSR is critical for liver and embryonic development (30). LSR was highly phosphorylated and contained sites primarily suggestive of basophilic kinases (Rxxs) but also had acidiphilic (sxxE) and Pro-directed (sP) phosphorylation. In addition, two phosphorylation sites were found near the C terminus. To our knowledge, only two of these sites have been previously described (31).

Table 1. Example of phosphorylation sites and phosphopeptides identified from a single protein, LSR

Phosphorylation sites	Peptide	No. of sites	*m/z*	Mass error, ppm	Charge state	XCorr
S308, S313	(R)TSS*VGGHS*SQVPLLR	2	842.8795	4.1	2	2.127
S375, **S379**	(R)AMS*EVTS*LHEDDWR	2	926.3401	4.2	2	2.445
S407	(R)APALTPIRDEEWNRHS*PR	1	742.3697	15.3	3	2.351
S436, **S448**, **S451**, **S459**	(R)S*VDALDDINRPGS*TES*GRSSPPSS*GR	4	988.7249	13.4	3	2.906
S473	(R)SRS*RDDLYDPDDPR	1	893.8841	12.4	2	2.651
S588, S591[†]	(K)NLALS*RES*LVV-	2	680.8219	7.1	2	2.234
Y256[‡]	(K)CCCPEALY*AAGK	1	740.2803	4.9	2	2.354
Y478[‡]	(R)SRDDLY*DPDDPRDLPHSR	1	562.9973	2.7	4	3.136

LSR is also called Lisch7 (liver-specific basic helix–loop–helix-zip transcription factor) in the database. Phosphorylation sites were localized with the Ascore method (26). Sites with <99% certainties of correct site placement are bold. Fourteen sites on 25 redundant tryptic peptides were actually detected from this protein (see SI Tables 2 and 3). For space reasons, the minimal set of peptides (eight) is shown for the 14 detected sites. *, phosphorylation sites; -, C terminus of the protein. All Met residues were detected in their oxidized form. Additional residues from the database are shown within parentheses and help to visualize motifs (e.g., Rxxs). Mass accuracy values shown in ppm were not recalibrated. XCorr values were from the Sequest algorithm.

[†] Sites were taken from ref. 31.

[‡] Sites were extracted from the Phosphosite (www.phosphosite.org) database.

Because of the relatively low abundance of Tyr phosphorylation, we used an alternative enrichment strategy (Fig. 1*A*) that consisted of a peptide immunoprecipitation experiment with anti-pTyr antibody (18) and that used higher peptide starting amounts (80 mg). A large data set of phosphopeptides (385 sites) mostly containing pTyr (351 unique pTyr sites in 916 redundant peptides from 280 proteins) was obtained and is presented in SI Table 3. Twenty-two of these sites were also identified in the SCX/IMAC study.

Assessing the Certainty of Precise Site Localization for 5,635 Detected Sites. To precisely assign the phosphorylation site within a peptide, we used the Ascore (ambiguity score) algorithm (26), a probability-based metric that scans for site-determining ions and computes the likelihood of their detection by chance alone, allowing all possible site placements to be considered. Ascore values directly represent the probability (P) of detection due to chance as $-10 \times \log(P)$, with scores >19 corresponding to sites localized with near certainty ($P < 0.01$).

An Ascore value was calculated for every site from the 8,527 phosphopeptides in SI Table 2 and 916 pTyr-containing peptides in SI Table 3. An example is shown in SI Fig. 6. The score distribution for all accepted 5,635 nonredundant sites also is shown. Near certainty ($>$99%) of localization was achieved for 61% of the data set (3,439 of 5,635 sites) and an additional 18% (1,008 of 5,635 sites) could be localized with $\approx$90% confidence ($P < 0.1$, Ascore > 10). For 12% (670 of 5,635 sites), very few or no site-determining ions were detected.

Positional Analysis of Phosphorylation Sites in Proteins. Analysis of the position of phosphorylation sites along the protein sequences was performed by dividing the protein length into 1% bins and counting the number of sites within each division. As shown in Fig. 3*A*, a substantial number of sites fell into the two last bins (i.e., phosphorylation events located at 98–100% of the protein's length). In fact, a $>$2.5-fold increase in the number of phosphorylation sites was observed in this C-terminal region. One possible explanation could be that phosphorylation is more likely to occur in flexible and exposed regions of a protein. However, a similar trend was not observed near the N terminus. To account for all possible versions of N-terminal peptides, additional searches with N-terminal acetylation were performed also permitting cleavage of the initial Met residue. Even considering all four possibilities for each N terminus (included in Fig. 3*A*), the frequency of phosphorylation at the extreme N terminus of a protein was very similar to the central regions of the protein and distinctly different from what was obtained for the C terminus.

Moreover, this C-terminal preference was not specific to being acquired by using our methodology, given that the same distribution was obtained from plotting the complete contents of the

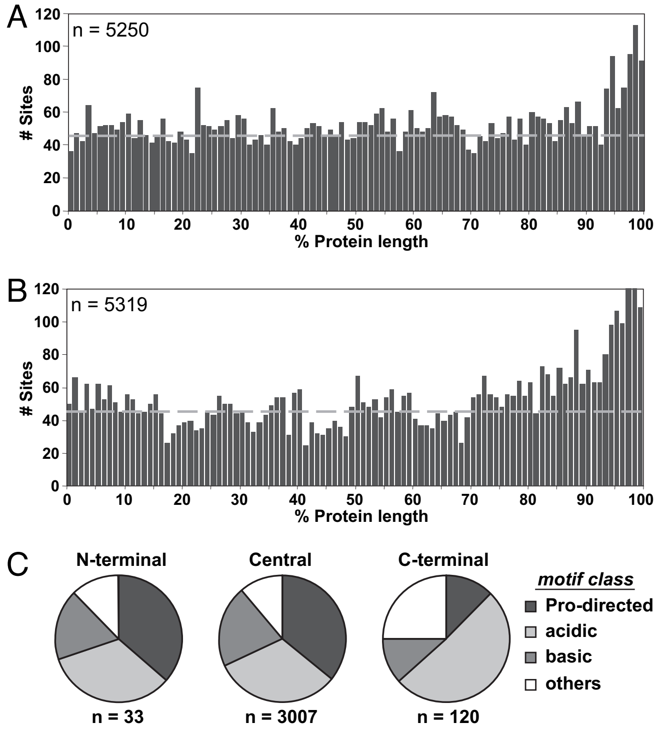

Fig. 3. Positional distribution of phosphorylation sites. (*A*) Frequency of site detection with respect to protein sequence position for this study. Protein sequences were divided into 1% bins and plotted by frequency. The dashed line shows the median value. A strong trend for C-terminal (98–100% of the protein length) phosphorylation was observed. (*B*) This same trend was observed for the distribution of sites from the Phospho.ELM database, a curated resource containing sites from the literature. (*C*) Classification of phosphorylation sites into the three most general motif classes based on their position within the protein. All localized sites were classified into one of three general kinase motifs or as ''other'' (see *Methods*), and their distributions at protein ends were determined. N-terminal, within first 10 aa; C-terminal, within last 10 aa; central, within all remaining residues. C-terminal sites had different distributions than N-terminal or central sites.

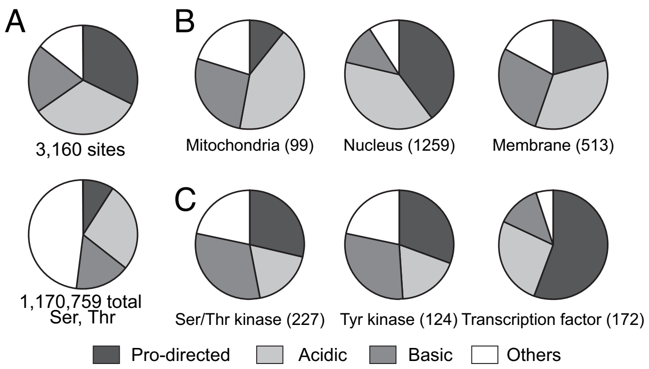

Fig. 4. Classification of localized sites ($P < 0.01$) into the most general kinase recognition sequence categories: acidic, basic, Pro-directed, and ''others.'' (*A*) Comparison of sequence category distributions for the sites detected in this study with the control set of all Ser and Thr residues from the same set of proteins. (*B*) Diverse phosphorylation patterns were observed for proteins belonging to different GO annotation cellular localization categories. (*C*) Examples of general phosphorylation kinase classification patterns for three functional GO categories.

Phospho.ELM database, which is similar in size to our data set (Fig. 3*B*). In contrast, constant numbers of Ser, Thr, and Tyr residues were observed along the entire protein sequence for these same proteins (SI Fig. 7), proving that this trend is directly associated with phosphorylation.

The Ser/Thr protein kinases fall into three major subgroups, Pro-directed, basophilic, and acidiphilic, on the basis of the types of substrate sequences preferred (32). To better understand the classes of kinases involved in these positional effects, we organized all localized phosphorylation sites into these three general sequence categories (see *Methods*). Whereas the pattern of sequence category frequencies for phosphorylation sites within the first 10 residues of a protein's sequence was identical to central sites, phosphorylation events within the final 10 residues were dramatically different. Pro-directed phosphorylation was greatly reduced at the C terminus, and acidic motifs were increased (Fig. 3*C*), whereas no variation for the frequencies of Pro, Asp, or Glu residues was observed (SI Fig. 7). An increase in acidic motifs might be explained if the C-terminal carboxylic acid itself was recognized by kinases preferring series of Glu and Asp residues. Finally, a control using and classifying all Ser and Thr from these proteins did not show a distinct sequence category distribution for C-terminal positions (data not shown).

These results may have implications for protein-tagging experiments for which specific epitope tags are incorporated into the sequence of proteins, often at the C terminus (33). It is possible that C-terminal phosphorylation events would not occur or be reduced under these conditions and therefore influence interactions with other proteins and protein function.

Correlation of General Ser/Thr Phosphorylation Motifs with Cellular Localization and Protein Function. Fig. 4*A* shows the distribution of all localized sites from 2,149 proteins within each of the three general motif classes. As a control, the background frequencies for all Ser and Thr residues from these proteins also are shown. From these data, most pSer- and pThr-containing sites (85%) were classified: Pro-directed (32%), acidic (33%), and basic (20%).

To investigate whether different cellular compartments, protein functions, or biological processes might exhibit idiosyncratic phosphorylation patterns, we looked for correlations between Gene Ontology (GO) annotation categories and phosphopeptide sequence characteristics in our list of 2,149 phosphoproteins from the SCX/IMAC study (SI Fig. 8). Only minor variations in the distribution of sites for the four sequence categories were observed between GO classes using a background of all Ser and Thr residues (SI Fig. 8*B*). In contrast, dramatic differences were found in the distributions for observed phosphorylation sites for both cellular location (Fig. 4*B* and SI Fig. 8*A*) and cellular function or process (Fig. 4*C* and SI Fig. 8*A*). For example, acidic phosphorylated sequences were frequently observed in extracellular and mitochondrial proteins but found less often in cytoskeletal proteins. Basic phosphorylated sequences were found with a lower relative frequency in nuclear proteins but were abundant in membrane, mitochondrial, and cytoskeletal proteins. Furthermore, Pro-directed phosphorylation events occurred with high frequency in proteins in the nucleus where many Pro-directed kinases are located, such as many cyclin-dependent kinase family members, but not in mitochondria or in the extracellular environment (Fig. 4*B* and SI Fig. 8).

Surprisingly, the same frequency distribution of general kinase categories was observed for Ser/Thr kinases and Tyr kinases (Fig. 4*C*), suggesting that kinase autophosphorylation may represent only one of many components in their overall regulation and is commonly accompanied by transphosphorylation by other kinases. Proteins involved in protein phosphorylation, signal transduction, and signaling cascades exhibited a distinct distribution with a low frequency of acidic phosphorylated sequences and higher numbers of basic sequences (SI Fig. 8*A*). Transcription factors showed a characteristic signature with a strong preference for phosphorylation at Pro-directed sites and relatively low numbers of basic sites (Fig. 4*C*). Conversely, low numbers of Pro-directed events were discovered in proteins involved in metabolism and electron transport and those with oxidoreductase activity (SI Fig. 8*A*).

Phosphorylation Motif Discovery. After we generally classified all phosphopeptides in our data set into three primary sequence categories, we sought to further refine these categories into specific, frequency-corrected phosphorylation motifs. To do this, peptide sequences for phosphorylation sites localized with >99% confidence (2,795 pSer, 341 pThr, and 283 pTyr) were all aligned, and their lengths were adjusted to ±6 aa from the central position and submitted to the Motif-X algorithm (34) (http://motif-x.med.harvard.edu). SI Table 4 lists the motifs generated containing a minimum of 50 pSer (≈2% of the total) and 6 pThr occurrences.

To graphically display each identified motif, logo-like representations were created. These logos included not only the residues strictly discovered to be part of the motif (SI Table 4), but also the frequencies of all additional adjacent amino acids (Fig. 5).

Certain motifs are commonly associated with specific kinases (35) and were prevalent in our data set. Basophilic kinase motifs such as RRxs (PKA), LxRxxs (CaM kinase family), and RxRxxs were identified. The latter has been associated with Akt or Akt-like kinase activity (36). We found 88 sites with this motif, including known Akt substrates (e.g., IRS1, ACINUS, CAHSP24). New potential substrates included SNIP1, CoREST, NDRG1, NDRG2, and NDRG3. In addition, several acidic casein kinase II motifs (e.g., sxDxExE, sxxEE, sDxE, and sDxD) and Pro-directed motifs recognized by MAP kinase (PxsP and PxtP) were well represented in our data, with substrates including MAPK2 and several Rsk family members. We had previously studied many of the Rsk1 sites as ordered phosphorylation events (37).

We also identified significant populations of peptides containing motifs that have not as yet been associated with a particular kinase. For example, there is no known kinase or interaction domain that specifically targets tPP sequences. This motif was frequently observed in our current data set (45 occurrences). It is possible that the second Pro residue could

APPLIED BIOLOGICAL SCIENCES

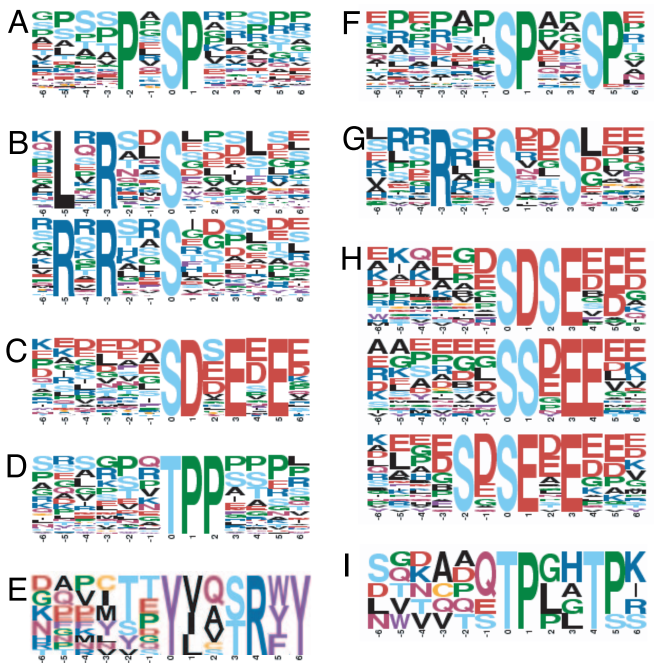

Fig. 5. Phosphorylation-specific motifs using the Motif-X algorithm (34). The data set included all sites with Ascore values of ≥ 19 ($n = 3{,}439$). The complete set of motifs is shown in SI Table 4. (*A*–*E*) Sequence logos for some examples of single-phosphorylation motifs where the phosphorylated residue (S or T) is centered. (*A*) Pro-directed motifs. (*B*) Basic motifs representative of CaM kinase and Akt kinase substrates, respectively. (*C*) Acidic motif with 56 occurrences in our data set of 630 in the entire mouse database. (*D*) Pro-directed motif centered on Thr with a strong preference for additional Pro residues C-terminal to the phosphate. (*E*) pTyr motif. (*F*–*I*) Examples of double-phosphorylation motifs. Secondary phosphorylated sites away from the central residue position also are phosphorylated. (*F*) Pro-directed with additional pSer at +4. (*G*) Basic upstream and acidic downstream. (*H*) Acidic motifs with two pSer residues. This category was the most highly represented for doubly phosphorylated peptides. (*I*) pThr-directed motif.

modulate kinase activity of a known Pro-directed kinase. Other examples of identified motifs were RxxsP and SPxxsP, with 126 and 51 occurrences, respectively. The latter could be a priming site for future GSK-3 phosphorylation at the −4 position.

As expected, fewer motifs were found for pThr and pTyr (5 and 1, respectively), because the total number of phosphorylation events identified on these residues was significantly less than that for Ser.

One noteworthy class of an undescribed motif exhibited properties of both acidic and basic motifs. To our knowledge, such "dipolar" motifs are previously uncharacterized and contain the minimum consensus sequence Rxxsxx[DE]. In addition to the locked positions, most residues on the N-terminal side of the pSer are basic in nature (K or R), whereas acidic (D and E) residues are abundant on the C-terminal half. The Phospho.ELM database contains 116 (44 with D and 72 with E) instances of such motifs, with associated kinases for 67 of them, usually directed to basic motifs. PKA was reported responsible for 21 sites, PKC and Akt were responsible for 6 sites, and CaM-II was responsible for 5 sites. Those sequences associated to acidic kinases were represented almost uniquely by nine sites assigned to CK2. However, this motif was not extracted when the Phospho.ELM database was run with Motif-X. Logo representations for these sequences were similar but not exact (SI Fig. 9). Moreover, of 125 sites with the dipolar motif identified in this study, only five (4%) had been reported in the Phospho.ELM database. It is conceivable that some sites may have evolved to respond to both acidiphilic and basophilic kinases.

Multiple Phosphorylations in a Short and Defined Distance: Double-Phosphorylation Motifs. In this study, 39% of the phosphopeptides identified were found to be multiply phosphorylated. Although we were able to establish a set of monophosphorylated motifs with high significance, we sought to find motifs involving more than one phosphorylated residue.

To find these motifs, we created an appropriate foreground for Motif-X that preserved multiple phosphorylation information by using all peptides with two, three, or four phosphorylations that were centered on each of those sites and by creating new amino acid notations for the additional phosphorylated sites that were observed outside the central position. Two different data sets were used as backgrounds (see *Methods* for details). In both cases, we obtained a similar subset of motifs. Twenty significant double-phosphorylation motifs at a minimum of 20 occurrences for pSer and 2 motifs with more than 4 instances for pThr were found (Fig. 5 *F–I* and SI Table 4).

In general, double phosphorylations were found more frequently in the context of acidic and Pro-directed motifs than in a basic environment. Not unexpectedly, some of the double-phosphorylation motifs identified could be deconvoluted into two separate motifs that were already identified in our previous general analysis. For example, the acidic motifs, ssxEE and sxsExE, can be both decomposed into two sxxE motifs.

Phosphorylation is known to often proceed in a step-wise fashion, where the first event serves as a priming event for the second (38). Well known examples of priming phosphorylation motifs were found. For example GSK-3 kinase, whose motif requires a phosphorylated Ser at position +4 (sxxxs), was found with 46 occurrences. We found several cases in which only one of the two specific loci in a double-phosphorylation motif was also identified as a singly phosphorylated species, which is suggestive of either ordered phosphorylation or dephosphorylation. As an example from the double-phosphorylation motif, sPxxsP (37 matches), only the single-phosphorylation motif, SPxxsP, was observed with significance, suggesting that the phosphorylation of the downstream Ser is a priming event for the upstream Ser at −4. Other doubly phosphorylated motifs were more frequent than their singly phosphorylated counterparts: sPsP was found with 23 occurrences, whereas only 6 and 12 examples were found for SPsP and sPSP, respectively. sPsP has been described previously as a motif for KIS kinase (39). Indeed, a known KIS substrate, SF1, was found among the 23 phosphopeptides containing this motif.

Methods

Tissue Preparation and Protein Digestion. Liver tissue from 21-day-old fed mice was homogenized and lysed by sonication in a buffer containing 8 M urea. Reduction and carboxyamidomethylation of Cys residues were performed on 90 mg of liver protein. Trypsin was added at 5 ng/μl and 1:250 enzyme/substrate. Digestion was stopped by the addition of TFA to 0.4%, and peptides were desalted through a C18 cartridge.

SCX Chromatography, IMAC, and pTyr-Immunoprecipitation. Preparative SCX separations were carried out on 10 mg of peptides with a 9.4- × 200-mm column packed with polysulfoethyl aspartamide (PolyLC, Columbia, MD) material similar to ref. 26. A total of 15 fractions were collected, acetonitrile was removed by evaporation, and all samples were desalted in C18 cartridges. Further details are provided in *SI Methods*.

Phosphopeptides were further enriched by IMAC with 10 μl of beads (Phos-Select iron affinity gel; Sigma, St. Louis, MO) per sample, essentially as described by the manufacturer. Details are given in *SI Methods*.

Immunoprecipitation of pTyr-containing peptides was performed by incubating 80 mg of desalted tryptic peptides with 40 μl of pY100 antibody beads (PhosphoScan pY100 Kit; Cell

Signaling Technologies, Danvers, MA) as described by the manufacturer.

MS and Database Searching. Samples were analyzed in the LTQ-FT or the LTQ-Orbitrap, essentially as described (40). Details are given in *SI Methods*.

MS/MS spectra were searched by using the Sequest algorithm against a database containing the mouse IPI protein database and its reversed complement. Search parameters included a static modification of 57.02146 Da (carboxyamidomethylation) on Cys; dynamic modifications of 79.96633 Da (phosphorylation) on Ser, Thr, and Tyr; and 15.99491 Da (oxidation) on Met. Additional details can be found in *SI Methods*.

Results were first filtered to contain only fully tryptic peptides, and then other cutoffs were established to achieve maximum sensitivity levels at $<0.1\%$ FP results using decoy matches as a guide. Sample-specific filters for the solution charge state and mass accuracy were used (26, 40). A ±3 SD mass tolerance window was applied separately to each analysis. After filtering by tryptic state, solution charge, and mass accuracy, only minimal filtering with Sequest scoring (XCorr and dCn′) values at the level of the entire data set was then required to achieve $<0.1\%$ FP rate. We define the dCn′ score as the dCn score to the first nonidentical sequence for a match.

Phosphorylation Site Localization. An Ascore was automatically calculated for every site using in-house software described elsewhere (26). A mass window setting of 100 *m/z* units and a fragment ion tolerance of ±0.3 *m/z* units were used. Ascores of >19 ($P < 0.01$) were considered to be confidently localized. For peptides with Ascores of <19, ambiguous sites were counted as only one site, regardless of the number of MS/MS spectra or potential site localizations. A conservative approach was also applied when counting the number of phosphopeptides such that different charge states, oxidized Met residues, misscleaved versions, and ragged ends did not add identifications to our nonredundant numbers.

Classification into General Motif Classes or Sequence Categories. Centered 13-mer sequences were assigned to a motif class sequentially by following a binary decision tree as follows: P at +1 (Pro-directed: P), 5 or more E/D at +1 to +6 (acidic: A), R/K at −3 (basic: B), D/E at +1/+2 or +3 (A), 2 or more R/K at −6 to −1 (B), otherwise (others: O).

Motif Analysis and Notation. Phosphorylation motifs present in our data set were extracted with the Motif-X algorithm (34). Only sites with Ascores of >19 were used. Phosphorylated residues were denoted by lowercase letters, and nonphosphorylated residues were uppercase. If two residues were significant in the same position, brackets were used, e.g., sxx[DE]. Variable positions were denoted by "x." Sequence logos were generated with Weblogo (41) (available at http://weblogo.berkeley.edu). Further details can be found in the *SI Methods*.

We thank B. Ballif for providing mouse liver tissue; D. Schwartz for designing the Motif-X algorithm; and D. Kirkpatrick, J. Elias, W. Haas, and B. Ballif for critically reading the manuscript. This work was supported in part by National Institutes of Health Grants GM67945 and HG3456 (to S.P.G.) and a postdoctoral fellowship from the Spanish Ministry of Education and Science (to J.V.).

1. Hunter T (2000) *Cell* 100:113–127.
2. Pawson T, Scott JD (2005) *Trends Biochem Sci* 30:286–290.
3. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) *Science* 298:1912–1934.
4. Songyang Z, Blechner S, Hoagland N, Hoekstra MF, Piwnica-Worms H, Cantley LC (1994) *Curr Biol* 4:973–982.
5. Hutti JE, Jarrell ET, Chang JD, Abbott DW, Storz P, Toker A, Cantley LC, Turk BE (2004) *Nat Methods* 1:27–29.
6. Dephoure N, Howson RW, Blethrow JD, Shokat KM, O'Shea EK (2005) *Proc Natl Acad Sci USA* 102:17940–17945.
7. Bishop AC, Ubersax JA, Petsch DT, Matheos DP, Gray NS, Blethrow J, Shimizu E, Tsien JZ, Schultz PG, Rose MD, *et al*. (2000) *Nature* 407:395–401.
8. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, *et al*. (2005) *Nature* 438:679–684.
9. Beausoleil SA, Jedrychowski M, Schwartz D, Elias JE, Villen J, Li J, Cohn MA, Cantley LC, Gygi SP (2004) *Proc Natl Acad Sci USA* 101:12130–12135.
10. Du K, Herzig S, Kulkarni RN, Montminy M (2003) *Science* 300:1574–1577.
11. Stiles B, Wang Y, Stahl A, Bassilian S, Lee WP, Kim YJ, Sherwin R, Devaskar S, Lesche R, *et al*. (2004) *Proc Natl Acad Sci USA* 101:2082–2087.
12. Ruvinsky I, Sharon N, Lerer T, Cohen H, Stolovich-Rain M, Nir T, Dor Y, Zisman P, Meyuhas O (2005) *Genes Dev* 19:2199–2211.
13. Jin WH, Dai J, Zhou H, Xia QC, Zou HF, Zeng R (2004) *Rapid Commun Mass Spectrosc* 18:2169–2176.
14. Moser K, White FM (2006) *J Proteome Res* 5:98–104.
15. Tao WA, Wollscheid B, O'Brien R, Eng JK, Li XJ, Bodenmiller B, Watts JD, Hood L, Aebersold R (2005) *Nat Methods* 2:591–598.
16. Oda Y, Nagasu T, Chait BT (2001) *Nat Biotechnol* 19:379–382.
17. Zhou H, Watts JD, Aebersold R (2001) *Nat Biotechnol* 19:375–378.
18. Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, Zhang H, Zha XM, Polakiewicz RD, Comb MJ (2005) *Nat Biotechnol* 23:94–101.
19. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM (2002) *Nat Biotechnol* 20:301–305.
20. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jorgensen TJ (2005) *Mol Cell Proteomics* 4:873–886.
21. Zhang Y, Wolf-Yadlin A, Ross PL, Pappin DJ, Rush J, Lauffenburger DA, White FM (2005) *Mol Cell Proteomics* 4:1240–1250.
22. Gruhler A, Olsen JV, Mohammed S, Mortensen P, Faergeman NJ, Mann M, Jensen ON (2005) *Mol Cell Proteomics* 4:310–327.
23. Collins MO, Yu L, Coba MP, Husi H, Campuzano I, Blackstock WP, Choudhary JS, Grant SG (2005) *J Biol Chem* 280:5972–5982.
24. Ballif BA, Villen J, Beausoleil SA, Schwartz D, Gygi SP (2004) *Mol Cell Proteomics* 3:1093–1101.
25. Trinidad JC, Specht CG, Thalhammer A, Schoepfer R, Burlingame AL (2006) *Mol Cell Proteomics* 5:914–922.
26. Beausoleil SA, Villen J, Gerber SA, Rush J, Gygi SP (2006) *Nat Biotechnol* 24:1285–1292.
27. Elias JE, Haas W, Faherty BK, Gygi SP (2005) *Nat Methods* 2:667–675.
28. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2003) *J Proteome Res* 2:43–50.
29. Yen FT, Masson M, Clossais-Besnard N, Andre P, Grosset JM, Bougueleret L, Dumas JB, Guerassimenko O, Bihain BE (1999) *J Biol Chem* 274:13390–13398.
30. Mesli S, Javorschi S, Berard AM, Landry M, Priddle H, Kivlichan D, Smith AJ, Yen FT, Bihain BE, Darmon M (2004) *Eur J Biochem* 271:3103–3114.
31. Kim JE, Tannenbaum SR, White FM (2005) *J Proteome Res* 4:1339–1346.
32. Pinna LA, Ruzzene M (1996) *Biochim Biophys Acta* 1314:191–225.
33. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B (1999) *Nat Biotechnol* 17:1030–1032.
34. Schwartz D, Gygi SP (2005) *Nat Biotechnol* 23:1391–1398.
35. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, *et al*. (2004) *Nucleic Acids Res* 32:D497–D501.
36. Alessi DR, Caudwell FB, Andjelkovic M, Hemmings BA, Cohen P (1996) *FEBS Lett* 399:333–338.
37. Ballif BA, Roux PP, Gerber SA, MacKeigan JP, Blenis J, Gygi SP (2005) *Proc Natl Acad Sci USA* 102:667–672.
38. Roach PJ (1991) *J Biol Chem* 266:14139–14142.
39. Manceau V, Swenson M, Le Caer JP, Sobel A, Kielkopf CL, Maucuer A (2006) *FEBS J* 273:577–587.
40. Haas W, Faherty BK, Gerber SA, Elias JE, Beausoleil SA, Bakalarski CE, Li X, Villen J, Gygi SP (2006) *Mol Cell Proteomics* 5:1326–1337.
41. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) *Genome Res* 14:1188–1190.