PhosphoSitePlus, 2014: mutations, PTMs and recalibrations

Peter V. Hornbeck*, Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham and Elzbieta Skrzypek

Cell Signaling Technology, 3 Trask Lane, Danvers, MA 01923, USA

Received November 14, 2014; Accepted November 19, 2014

ABSTRACT

PhosphoSitePlus® (PSP, http://www.phosphosite.org/), a knowledgebase dedicated to mammalian post-translational modifications (PTMs), contains over 330 000 non-redundant PTMs, including phospho, acetyl, ubiquityl and methyl groups. Over 95% of the sites are from mass spectrometry (MS) experiments. In order to improve data reliability, early MS data have been reanalyzed, applying a common standard of analysis across over 1 000 000 spectra. Site assignments with $P > 0.05$ were filtered out. Two new downloads are available from PSP. The ‘Regulatory sites’ dataset includes curated information about modification sites that regulate downstream cellular processes, molecular functions and protein-protein interactions. The ‘PTMVar’ dataset, an intersect of missense mutations and PTMs from PSP, identifies over 25 000 PTMVars (PTMs Impacted by Variants) that can rewire signaling pathways. The PTMVar data include missense mutations from UniPROTKB, TCGA and other sources that cause over 2000 diseases or syndromes (MIM) and polymorphisms, or are associated with hundreds of cancers. PTMVars include 18 548 phosphorylation sites, 3412 ubiquitylation sites, 2316 acetylation sites, 685 methylation sites and 245 succinylation sites.

INTRODUCTION

Discovery mode tandem mass spectrometry (MS) has transformed the landscape of cell biology and cell signaling research over the past 10 years. PhosphoSite® began its life just as this new era was in its infancy. When launched in 2003, it encompassed 1200 modification sites on 500 human and mouse proteins. Now, 11 years later, PhosphoSitePlus® (PSP) contains one-third of a million non-redundant modification sites on over 20 000 protein groups, and behind the scenes it hosts 1.5 million peptides. The high-throughput MS2 (HTP) data in PSP now dwarfs the high-quality data aggregated by manual curation of published low-throughput (LTP) scientific literature: only 4.5% of the sites in PSP are derived from LTP studies. While the number of sites in PSP has expanded nearly 350-fold since 2003, it is not the site count that matters most—data quality and reliability are top priority. In this paper we will review changes made to PSP since 2012, including the measures taken to ensure that the information in PSP is as reliable as possible.

Since 2003, over 340 000 non-redundant MS2 PTM sites have been curated into PSP from over 8000 separate experiments performed at CST and from 202 publications. The algorithms used for site assignment, as well as the mass accuracy of the commonly used spectrometers, have significantly improved over the last 10 years. Much of the HTP data from 2003 to 2008 came from less sensitive instruments using an older generation of software. Beginning in 2011–12, much of the older phosphorylation data in PSP has been re-analyzed using contemporary algorithms including Ascore (1).

Since PSP relies increasingly on MS2 data, and 68% of all sites are associated with only 1 or 2 HTP records, it is important to understand what the limits of interpretation are for MS data. Specifically, how reliable are PTMs that have been reported only once or twice by HTP MS? This question will be addressed in the Results and Discussion section.

On another front, the extraordinary advances in genome sciences and sequencing are uncovering a prodigious number of genetic variants and disease mutations. One of the most exciting challenges in this domain of Big Data is at the intersection of genetic variation and PTMs. The observation that missense mutations can rewire signaling pathways (2,3,4,5) has provided the inspiration for the production and monthly updating of PTMVar, a dataset that maps missense mutations onto the PTM sequence space. PTMVar is available on the PSP Download page (www.phosphosite.org/staticDownloads.do). This data is expected to be of value to genetic researchers in understanding molecular mechanisms underlying frank disease mutations and the association between polymorphic variants and genetic risks, and to cell biologists investigating the mechanisms through which mutations can rewire signaling networks.

*To whom correspondence should be addressed. Tel: +978 867 2368; Fax: +978 867 2400; Email: phornbeck@cellsignal.com

© The Author(s) 2014. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
A third focus of this paper is an examination of the in vitro and in vivo kinase-substrate interactions (KSIs) curated into PSP, and their use in generating sequence logos derived from reactions between protein kinases and intact protein substrates. Kinase specificity profiles generated using short peptides (6,7), methods that provide excellent profiles of the intrinsic affinities of kinase substrate binding pockets for peptides, cannot capture any additional information that might be generated when two surfaces collide, especially when the acceptor site is on the surface of a globular domain. The number of high-quality mammalian KSIs reported in PSP is 12 180 in 2014; 84% of these include human kinases, and many of the KSIs have been corroborated using kinases or substrates from multiple species. There are adequate numbers of KSIs in PSP to generate up to 100 sequence preference profiles of kinases interacting with substrate proteins. The utility of this data in generating kinase substrate sequence logos can be seen in Figure 5.

CONTENT: GROWTH AND CHANGES SINCE 2012

Site statistics

The total number of modification sites in PSP has increased from 128 943 to 338 948 over the past 3 years (Table 1). These include 274 396 modification sites on 20 021 human proteins, 139 066 sites on 15 389 mouse proteins and 42 465 sites on 7204 rat proteins.

User interfaces

The four main user interfaces (Homepage, Protein Page, Modification Site Page and Curated Information Page) have kept close to their original structure and function in order to provide a stable and recognizable milieu for user interactions. These were described in the 2012 article (8), are described in the online tutorial (http://www.phosphosite.org/staticTrainingTutorial.do) and will not be addressed here.

Protein group and site group—essential concepts

The concepts of Protein Group and Modification Site Group provide essential unique identifiers (UIDs) underlying the ontology and systematics of proteins and modification sites in PSP. A Protein Group includes all individual proteoforms and orthologs from other species. A Modification Site Group provides a critical classifier for the unambiguous identification of homologous and orthologous sites within proteoforms and between species. Other possible site-specific identifiers, including residue number and surrounding sequence for specific sites, are inadequate. Residue numbers cannot be used because they can vary between isoforms or orthologs. For example, in an alignment of all Tau protein group members curated into PSP (9 human isoforms of Tau plus mouse, rat and cow orthologs), many of the corresponding modification sites across all protein group members have different residue numbers, yet would be uniquely joined together under a single Site Group ID. Residue numbering can also vary over time in proteins from the same repository as sequences are refined or as terminal methionines come and go. Nor can the flanking sequences uniquely identify a site, as sequences surrounding a site can vary between proteoforms of the same individual (e.g. splice variants or somatic mutations), between individuals within a population (e.g. somatic or germline variants) and between orthologs. Additionally, sequences are often repeated throughout the same protein multiple times, rendering sequences alone unsuitable as modification site UID. For example, the tryptic peptide K.FS*MPGFKAEPEVDVNLPK is found five times in human protein AHNAK (Table 2). The systematics underlying PSP require each repeat to have a different Site Group ID—within any specific sequence, a site group can only have a one-to-one correspondence. This example illustrates a conundrum faced while curating such peptides—how to assign it? Any single assignment has a four-fifths chance of being incorrect. Perhaps some sort of fuzzy logic could handle such situations. In any case, throughout this paper, as in our previous 2012 article (8), the terms ‘protein’ and ‘site’ shall apply to a Protein Group and a Site Group, respectively, unless otherwise stated. Note that the Site Group ID is now included with each site in downloads from PSP, enabling the user to identify, sort and consolidate site-specific information when desired.

The evolution of sequence space

The processing of peptides from proteomic MS experiments has evolved over the years. We have always considered UniPROTKB SwissProt (9), followed by NCBI RefSeq NP (10), to be the most reliable protein sequence resource. Thus, we wanted to use them to define our sequence space, but we encountered many peptides reported in the literature associated with IPI, TrEMBL, RefSeq XP and Ensembl

<table>
<thead>
<tr>
<th>Table 1. Non-redundant numbers of proteins and post-translationally modified sites in PSP: 2011 vs. 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Proteins</td>
</tr>
<tr>
<td>Mod type</td>
</tr>
<tr>
<td>phospho-Ser</td>
</tr>
<tr>
<td>phospho-Thr</td>
</tr>
<tr>
<td>ubiquityl</td>
</tr>
<tr>
<td>phospho-Tyr</td>
</tr>
<tr>
<td>acetyl</td>
</tr>
<tr>
<td>mono-methyl</td>
</tr>
<tr>
<td>di-methyl</td>
</tr>
<tr>
<td>galNAc</td>
</tr>
<tr>
<td>di-methyl</td>
</tr>
<tr>
<td>tri-methyl</td>
</tr>
<tr>
<td>Total Sites</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Five internal sequence duplications within the human protein AHNAK. Each site requires a unique Site Group ID.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>S886</td>
</tr>
<tr>
<td>S1721</td>
</tr>
<tr>
<td>S2057</td>
</tr>
<tr>
<td>S2519</td>
</tr>
<tr>
<td>S2903</td>
</tr>
</tbody>
</table>

User interfaces

The four main user interfaces (Homepage, Protein Page, Modification Site Page and Curated Information Page) have kept close to their original structure and function in order to provide a stable and recognizable milieu for user interactions. These were described in the 2012 article (8), are described in the online tutorial (http://www.phosphosite.org/staticTrainingTutorial.do) and will not be addressed here.

Protein group and site group—essential concepts

The concepts of Protein Group and Modification Site Group provide essential unique identifiers (UIDs) underlying the ontology and systematics of proteins and modification sites in PSP. A Protein Group includes all individual proteoforms and orthologs from other species. A Modification Site Group provides a critical classifier for the unambiguous identification of homologous and orthologous sites within proteoforms and between species. Other possible site-specific identifiers, including residue number and surrounding sequence for specific sites, are inadequate. Residue numbers cannot be used because they can vary between isoforms or orthologs. For example, in an alignment of all Tau protein group members curated into PSP (9 human isoforms of Tau plus mouse, rat and cow orthologs), many of the corresponding modification sites across all protein group members have different residue numbers, yet would be uniquely joined together under a single Site Group ID. Residue numbering can also vary over time in proteins from the same repository as sequences are refined or as terminal methionines come and go. Nor can the flanking sequences uniquely identify a site, as sequences surrounding a site can vary between proteoforms of the same individual (e.g. splice variants or somatic mutations), between individuals within a population (e.g. somatic or germline variants) and between orthologs. Additionally, sequences are often repeated throughout the same protein multiple times, rendering sequences alone unsuitable as modification site UID. For example, the tryptic peptide K.FS*MPGFKAEPEVDVNLPK is found five times in human protein AHNAK (Table 2). The systematics underlying PSP require each repeat to have a different Site Group ID—within any specific sequence, a site group can only have a one-to-one correspondence. This example illustrates a conundrum faced while curating such peptides—how to assign it? Any single assignment has a four-fifths chance of being incorrect. Perhaps some sort of fuzzy logic could handle such situations. In any case, throughout this paper, as in our previous 2012 article (8), the terms ‘protein’ and ‘site’ shall apply to a Protein Group and a Site Group, respectively, unless otherwise stated. Note that the Site Group ID is now included with each site in downloads from PSP, enabling the user to identify, sort and consolidate site-specific information when desired.

The evolution of sequence space

The processing of peptides from proteomic MS experiments has evolved over the years. We have always considered UniPROTKB SwissProt (9), followed by NCBI RefSeq NP (10), to be the most reliable protein sequence resource. Thus, we wanted to use them to define our sequence space, but we encountered many peptides reported in the literature associated with IPI, TrEMBL, RefSeq XP and Ensembl.
that mapped to neither. We wanted to represent what the authors reported so we imported protein sequences from these other resources when necessary. These additional sequences proved unstable (11), making the curation process time consuming and unsustainable, especially as we moved further into the Big Data era. We are now transitioning to a stable human and mouse sequence space that will be exclusively populated with UniPROTKB SwissProt sequences updated once per month. Over 99.9% of the human proteins in PSP are from UniPROTKB, 97.1% from SwissProt and 2.8% from TrEMBL; 94.7% of the mouse sequences are from UniPROTKB, 88.8% of which are SwissProt. Sequences and accession numbers are synchronized with UniPROTKB every month.

Molecular structures
Since 2011, the number of PDB files (protein structural files from Protein Data Bank, http://www.rcsb.org/) corresponding to proteins in PSP has increased from 22 958 (4500 proteins) to 33 204 (5989 proteins). Nearly 27% of the 22 241 proteins in PSP have associated PDB files corresponding to some segment of the protein’s sequence. It is notable that 69 388 sites, 19.1% of all sites in PSP, are located within the structures included in PDBs. Thus, a significant fraction of post-translationally modified residues are amenable to some level of structural analysis, enabling, for example, research into the effect of phosphorylation on protein-protein interactions (12), or upon the binding of drug inhibitors to target proteins (13).

LTP data
The total number of curated papers that used LTP technology to characterize modification sites is nearly 17 000, an increase of 4000 over the past 3 years. The ordinal rank of the top four journals represented PSP is the same as reported in 2012 (J. Biol. Chem., Mol. Cell. Biol., Oncogene and Proc. Natl. Acad. Sci. U.S.A.) with the digital, open access journal PLoS now in fifth place. The use of I2E (14), a powerful language processing software application, to identify articles and highlight information for manual curation, has significantly increased the efficiency and throughput of our LTP curation efforts, and made recreation of selected information a realistic and less time-consuming option.

HTP data: peptides and sites
Peptides and sites from 98 new proteomic MS publications and 6560 curation sets from Cell Signaling Technology (CST) have been curated since 2012. A majority of sites in PSP have only been observed using MS: of 344 413 sites in PSP, only 15 596 (4.5%) have been reported in the LTP literature. Of the MS-only sites, 55% have been reported only once, 25% between 2 and 5 times, 9% between 6 and 25 times, and 2.4% more than 25 times. MS peptide sequences that lie behind the assignments in PSP are now curated into PSP: nearly 1 700 000 redundant and 315 000 non-redundant peptides have been incorporated into PSP from CST curation sets and publications. This data will be made available in the coming months via download and in the longer term via an API.

This dominance of HTP data has made it critical to understand what the limitations of MS data are, to improve the quality of information already in PSP, and to ensure that data going into PSP is of the highest quality. Towards this end, we have reprocessed previously curated papers and CST curation sets to exclude lower quality site assignments.

Recuration of HTP data: applying a higher standard
A new threshold was implemented in 2012 for inclusion of new sites into PSP, either from publications or from CST experiments: sites had to have localization scores of $P = / < 0.05$ or Ascore $> 13$ to qualify for inclusion in PSP. There were, however, thousands of curation sets from CST and close to 120 proteomic MS papers already curated into PSP prior to establishing this threshold. Indeed, much of this data was curated into PSP prior to the existence of the Ascore algorithm (1) or the PTM localization probability score (15). Papers already curated into PSP that included site assignment scores were recurated, excluding sites below the thresholds. The site assignments from 4002 phosphorylation curation sets already in PSP were removed and reanalyzed on the new CST MS2 server. Note that 94 000 sites meeting the new standard were curated into PSP. Many thousands of sites previously in PSP did not meet the new standard, and similar numbers of new sites were introduced. A little over 4800 new curation sets, containing 139 600 filtered sites, have been added since January 2012.

As of October 2014, none of the 168 ubiquityl or 551 acetyl CST curation sets curated prior to January 2012 has been rescored. This process has, however, been started and should be completed within 6 months.

New links to useful resources
A number of new links that take the user to corresponding pages in other outstanding resources have been added to PSP Protein Pages. These include: the Human Protein Atlas (http://www.proteinatlas.org/) for protein expression levels and subcellular localization based on immunohistochemistry, and transcript expression levels for a large number of human tissues, cancers and cell lines (16); BioGPS (http://biogps.org/) for tissue-specific pattern of mRNA expression (17); Wikipedia pages include information about protein function, RNA expression, interactions, clinical significance, references and further reading (e.g. http://en.wikipedia.org/wiki/ANXA2); neXtProt (18), a human protein resource from Swiss Institute of Bioinformatics and GeneBio that integrates information from the UniProtKB/Swiss-Prot Knowledgebase, the HUPO Human Proteome Project (19) and other post-proteomic resources; and Reactome for the visualization, interpretation and analysis of pathway knowledge (20). Also, the chromosomal location of the gene encoding each human protein in PSP is now included on Protein Pages and in many downloads.

Downloads and their content
Two types of download are available through PSP: those for downloading results of user specified searches initiated...
interactively from search pages (Figure 1, top), and those available via the download page (Figure 1, bottom). Downloads from Site Search Result Pages include the results of user-defined searches for modification sites that fulfill one or more of 12 criteria including sites that are: responsive to specific treatments; observed within specific domains, and proteins of a specific type; within proteins associated with specific functions, processes, cellular components, or molecular weight ranges; observed within a specific sequence or motif; and observed in specific cell lines, cell types or tissues. Since 2012, Site Group IDs, LTP and HTP counts have been added to all downloads containing site-specific information. This is in addition to standard names, accession numbers, molecular weights, modified residues and their flanking sequences. The number of associated LTP and HTP records may be of particular use in evaluating the strength of evidence associated with the modification site (for example, see the section "Evidence for neutral mutation..." in the Results and Discussion section).

Static downloads described in this section are available via the Download Page (www.phosphosite.org/staticDownloads.do), and are complete, pre-compiled datasets of various categories of information contained in PSP (Figure 1, bottom). While the information in the database is curated and updated daily, the downloads are prepared only once a month. In contrast, queries through the UI interact in real time with the PSP DB. Three datasets containing functional metadata are described in more detail below. ‘PTMVar’ and ‘Regulatory Sites’ are new to this version of PSP. The Kinase Substrate Dataset is analyzed in a later section to evaluate its information content and utility.

The download ‘PhosphoSitePlugin’ (Figure 2) is a Cytoscape plugin (21) that contains site-specific information abstracted from PSP. Initially, only KSIs were included in the plugin; subsequently, treatments that regulate the modification status of specific residues have been added to the graph. More than 1390 treatments and 29 700 downstream targets that they regulate have been curated into PSP and are available for incorporation into the Cytoscape plugin. Note, however, that the site-centric information in PSP is represented only at the protein level in Cytoscape, reducing graph complexity but also losing valuable information.

The ‘Regulatory Sites’ dataset provides curated information mainly from LTP papers about 7211 sites on 2374 proteins that regulate downstream cellular processes, molecular functions and protein-protein interactions: 3630 sites regulate 47 downstream cellular processes including 1743 that regulate transcription, 778 that influence cytoskeletal organization, 781 that regulate apoptosis, 252 that regulate cell adhesion and 539 that regulate cell motility; 6368 sites control 8 different molecular functions including 1788 that regulate intracellular localization, 1178 that induce enzymatic activity, 1045 that regulate protein stabilization or degradation and 399 that inhibit enzymatic activity; and 2845 sites on 1238 proteins that regulate protein-protein interactions with 821 binding partners. The top five binding partners are Src, Grb2, SHP-2, Shc1 and 14–3–3 β.
The ‘PTMVar’ dataset identifies modification sites that are either at or within five residues of a site that has been mutated, either in germline or somatically. PTMVar currently contains variant data from monthly downloads of ‘humsavar.txt’ from UniProtKB (http://www.uniprot.org/docs/humsavar). Cancer somatic mutations, identified in reference 4, are from TCGA (http://cancergenome.nih.gov), cBio (22), COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) and other sources (23). The current PTMVar dataset includes 17,551 modification sites associated with 24,592 mutations, of which 39% are classified as Disease mutations (including 1,250 discrete diseases and genetic disorders from Online Mendelian Inheritance in Man (http://omim.org/)), 44% as polymorphisms, and 17% as unclassified. Our goal is to include somatic mutations from cancer genomes only when observed in three to five independent cancer patient samples.

The current Kinase Substrate’ dataset contains 12,180 high quality manually curated KSIs, 10,194 of which have been demonstrated using human kinases. It is notable that 9% of the kinases and 8% of the substrate proteins have been corroborated using kinases and/or substrates from multiple species. While in vivo and in vitro results are curated into PSP, in vitro results outscore in vivo by 11,396 to 6,231. Note that 368 mammalian kinases have at least one experimentally reported substrate in PSP. A total of 232 have 5 or more non-redundant substrates, 170 have 10 or more and 100 have 20 or more and are listed in Table 3.

RESULTS AND DISCUSSION

Evidence for neutral mutation and the non-functional nature of rarely phosphorylated sites

Only 4.5% of the sites reported in PSP are associated with LTP literature records. All others are based on MS data alone. Of these, 55% have been reported in only one record, 13% in two, and 6% in three. Apparently none of these MS-only sites have yet been corroborated in publications using LTP methodologies. Are these ‘low hitters’ real sites or artifacts? Is there any evidence that they represent functional sites? This first question can be gauged by evaluating the phosphosites in PSP, many of which have been reanalyzed, keeping only site assignments with p < 0.05 or Ascore =/> 13. Although many of the ‘one hitters’ were removed during this process, a similar number were added back, keeping their overall numbers relatively constant, indicating that as many as 90-95% of the rescored ‘one-hit’ phosphosites from CST may be real, i.e. are not spurious artifacts of MS. Furthermore, the numbers associated with these ‘one hitters’ do not increase systematically over time.

In contrast, the number of hits for sites known to be functional role within cell signaling networks will diverge more quickly than sites that are functional nodes within cellular communication systems. Phosphosites were sorted into 5 bins according to the number of ‘hits’ (i.e., the number of records in which they were reported) from 1 to >100 (Figure 3). A total of 203,161 human/mouse and 148,441 mouse/rat phosphorylation sites were compared. The inset indicates the number of phosphosites included in each comparison. Blue, fraction of amino acid substitutions between human phosphosites and orthologous mouse residues; red, fraction of amino acid substitutions between mouse phosphosites and orthologous rat residues.

The amount of evolutionary divergence of phosphorylation sites (Ser, Thr and Tyr) between human and mouse, and between mouse and rat, reasoning that sites that play no functional role within cell signaling networks will diverge more quickly than sites that are functional nodes within cellular communication systems. Phosphosites were sorted into 5 bins according to the number of ‘hits’ (i.e., the number of records in which they were reported) from 1 to >100 (Figure 3). A total of 203,161 human/mouse and 148,441 mouse/rat pairs were tested. In both sets of pairs, the divergence is highest for sites with only 1 hit (20.6% and 8.8% for the human/mouse and the mouse/rat pairs, respectively), and lowest for the sites with >100 hits (5.4% and 3.0% for the human/mouse and the mouse/rat pairs, respectively).

These results demonstrate that the more frequently sites are observed, the more likely they are to be conserved.

It is notable that the divergence is on average 2.3-fold higher for the human-mouse pairs than for the mouse-rat pairings, numbers that are consistent with estimated divergence times between humans and mice and between mice and rats of about 96 and 33 million years ago, respectively (25). These results indicate that these low-scoring sites, while most are almost certainly ‘good’ assignments, should not be taken as strong evidence of any biological role for the site without additional evidence for functionality.

PTMVar: mutations that can rewire signaling circuits

Non-synonymous mutations are missense mutations that change one amino acid for another in a protein’s coding region. If the mutation is at or near the site of a PTM that participates in cell signaling, it can significantly impact signaling networks in which the affected site participates. Class I PTMVars are those in which a PTM site is lost by altering the relevant amino acid (Figure 4). In the Class I example in Figure 4, Ser-16 of phenylalanine hydroxylase (PAH) is
and related kinases, is mutated in 439 substrate sequences; Pro at +1, required for most CMGC kinases, is lost in 266 substrate sequences; and Gln at +1, essential for catalysis by the DNA damage kinases ATM, ATR and DNAPK, is lost in 66 sequences in the current release.

The current download includes 1622 Class I, 181 Class Ia and 23 178 Class II PTMVars. All modification types are affected by variants and include 18 548 phosphorylation sites, 3412 ubiquitylation sites, 2316 acetylation sites, 463 monomethylation sites and 245 succinylation sites (Table 4).

Searching the PTMVar dataset can be further enhanced by cross-dataset joins made by keying on Site Group IDs. This enables users to aggregate information about specific modsites from multiple datasets. For example, a join of the disease mutations from the PTMVar with and the Regulatory Site datasets identifies 272 PTMVars at sites known to regulate downstream cellular processes, molecular functions or protein-protein interactions. Disruptions of the phosphorylation of these ‘regulatory’ sites may well be expected to rewire the circuits in which they participate (Supplemental Table 1).

Generating sequence logos and PSSMs from whole protein KSIs

PSP contains enough high-quality, expert edited and curated KSIs to generate up to 100 substrate preference profiles based on intact protein-protein interactions alone. This stands in contrast to resources that generate KSIs based on in vitro reactions using short peptides only (6,7) and modified variously with predictions or adjustments of one sort or another (26). Figure 3 shows 10 sequence logos of kinases from different major groups built with data from PSP with tools available at www.phosphosite.org/sequenceLogoAction.do. Included are (i) three kinases with exceptionally strong and nearly absolute substrate preferences: ERK2, with a strong preference for P at +1, and secondary preference for P at −2; ATM, with a preference for Q at +1, and no observable secondary preference; and the basophilic kinase AKT with a strong preference for R at −3 and a secondary preference for R at −5; (ii) kinases with moderate or slight preferences: PKCε; [RK] at +2, −3 and −2; Src: E at −4 and −3, [DP] at −2, D at −1 and P at +3;
Table 4. The numbers of PTMVars for each of 18 modification types in the PTMVar download, September 2014

<table>
<thead>
<tr>
<th>Mod type</th>
<th>AA</th>
<th>N</th>
<th>Mod type</th>
<th>AA</th>
<th>N</th>
<th>Mod type</th>
<th>AA</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phospho</td>
<td>Ser</td>
<td>9871</td>
<td>Mono-methyl</td>
<td>Arg</td>
<td>382</td>
<td>Sumoyl</td>
<td>Lys</td>
<td>77</td>
</tr>
<tr>
<td>Phospho</td>
<td>Thr</td>
<td>4754</td>
<td>Mono-methyl</td>
<td>Lys</td>
<td>76</td>
<td>O-GlcNAcetyl</td>
<td>Ser</td>
<td>49</td>
</tr>
<tr>
<td>Phospho</td>
<td>Tyr</td>
<td>3923</td>
<td>Di-methyl</td>
<td>Arg</td>
<td>174</td>
<td>O-GlcNAcetyl</td>
<td>Ser</td>
<td>48</td>
</tr>
<tr>
<td>Ubiquityl</td>
<td>Lys</td>
<td>3412</td>
<td>Di-methyl</td>
<td>Lys</td>
<td>24</td>
<td>O-GlcNAcetyl</td>
<td>Thr</td>
<td>36</td>
</tr>
<tr>
<td>Acetyl</td>
<td>Lys</td>
<td>2316</td>
<td>Tri-methyl</td>
<td>Lys</td>
<td>22</td>
<td>Neddyl</td>
<td>Lys</td>
<td>35</td>
</tr>
<tr>
<td>Succinyl</td>
<td>Lys</td>
<td>244</td>
<td>O-GalNAcetyl</td>
<td>Thr</td>
<td>84</td>
<td>Caspase cleavage</td>
<td>Asp</td>
<td>26</td>
</tr>
</tbody>
</table>

Figure 5. Sequence logos of 10 different kinases from various kinase groups. The number of substrate sequences used to generate each logo: SRC, 505; Akt, 200; GRK2, 58; ATM, 177; PKCa, 472; MAPKAPK2, 53; PIM1, 26; ERK2, 538; CK2α, 506 and PAK1, 63. Logos generated at www.phosphosite.org/siteSearchSelectAction.do using the Frequency Change logo graph method (31).

Figure 6. Sequence logos generated using increasing numbers of input sequences. At 5 input Akt sequences, the dominance of R at −3 and −5 was already evident. The preference of PKCa for [RK] at +2 was evident using 10 input sequences, but the preference for R at −2 and −3 required between 10 and 50 input sequences. Logos were generated using the Frequency Change algorithm at http://www.phosphosite.org/sequenceLogoAction.do.

and (iii) preferences for a general quality but no outstanding single residue: GRK2, acidic residues throughout the kinase binding region.

How many sequences does it take to make a logo that reflects specificity? That depends on the topography and strength of the interactions between the residues surrounding the phosphate-accepting group on the substrate and the corresponding contact residues on the kinase. This is illustrated in Figure 6, where logos are built for Akt and PKCa from graded numbers of substrate sequences. In dozens of repetitions of randomly selected sequences, the preference of Akt for R at −3 and −5 showed up strongly for all 5-mers examined. In contrast, the preference of PKCa for [RK] at +2, −3 and −2 did not show up consistently until the input sequences numbered somewhere between 10 and 50. The conclusion drawn from this example is that as few as five input sequences can reveal strong preferences, but kinases with less selective requirements will require 20 or more sequences to reveal their substrate predilections; visualizing fine specificities would require significantly more than these numbers. A review of 100 kinase logos generated with data from PSP suggests that most kinases exhibit only moderate substrate sequence specificities (data not shown).

Domain location of various classes of PTMs

Serine and threonine phosphorylation sites modulate the structure and function of short linear motifs, regions that play important roles in cellular communication, and that are generally located between structured domains (27). We previously examined the location of various classes of PTMs relative to PFAM A domains and reported enrichment of pSer and pThr between domains (8). We repeated this analysis with substantially more data, and, in addition, analyzed the locations of arginine methylation sites (Figure 7). We confirmed the observations that 70–75% of all pSer/Thr sites occur in regions between structured do-
selective pressure on sites with only one or a few MS hits to the organism, and that neutral mutation is relatively unconstrained at these sites. This in turn should suggest to the experimental biologist that many of these 'low-hitter' sites are the result of stochastic processes and, in the absence of additional compelling evidence, may be risky targets for further research.

Lastly, PTMVar provides a proteome-wide snapshot of how thousands of disease-causing missense mutations can interact with the 300 000+ unique PTMs contained within PSP. There are almost certainly clues within this dataset that can provide actionable insights into the interplay between disease mutations and cell signaling mechanisms.

CONCLUDING REMARKS

The resoring of peptide and site assignments across archived CST experiments performed over the past nine years has increased the coherence and reliability of 85% of the phosphorylation sites from CST. The ongoing reanalysis of archived acetylation, methylation and ubiquitylation data in the coming months will similarly improve its coherence and accuracy. Users should not be alarmed when sites with 1 or a few hits disappear; they probably don’t meet our new standards, and are being replaced with more accurate data. They should also be aware that the MS data generated at CST and stored in PSP can be periodically reanalyzed using more advanced algorithms as they are developed, a capability shared by the PTM resource PHOSIDA (28) but few others.

The demonstration that pSer, pThr and pTyr sites with very low numbers of MS hits have undergone relatively rapid evolutionary divergence, and that amino acids frequently observed to be phosphorylated have diverged more slowly, indicates that evolutionary pressure conserves sites with lots of phosphorylation activity, but exerts little pressure against neutral mutations at sites with only one or a few phosphorylation hits. These results imply two things. First, the low divergence observed between orthologous sites with high phosphorylation activity suggests that these sites are being actively conserved, perhaps because they play positive roles in cellular life. Second, that the apparent lack of selective pressure on sites with only one or a few MS hits suggests that their phosphorylation is of little consequence to the organism, and that neutral mutation is relatively unconstrained at these sites. This in turn should suggest to the experimental biologist that many of these ‘low-hitter’ sites are the result of stochastic processes and, in the absence of additional compelling evidence, may be risky targets for further research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

With gratitude to Michael Comb for supporting this work since its inception.

FUNDING

National Institutes of Health [R43GM065768, R44AA014848m, R44CA126080 and U54-HL127624-01]. Funding for open access charge: Cell Signaling Technology.

Conflict of interest statement. None declared.

REFERENCES
