EPSD: a well-annotated data resource of protein phosphorylation sites in eukaryotes

Shaofeng Lin†, Chenwei Wang†, Jiaqi Zhou, Ying Shi, Chen Ruan, Yiran Tu, Lan Yao, Di Peng and Yu Xue

Abstract

As an important post-translational modification (PTM), protein phosphorylation is involved in the regulation of almost all of biological processes in eukaryotes. Due to the rapid progress in mass spectrometry-based phosphoproteomics, a large number of phosphorylation sites (p-sites) have been characterized but remain to be curated. Here, we briefly summarized the current progresses in the development of data resources for the collection, curation, integration and annotation of p-sites in eukaryotic proteins. Also, we designed the eukaryotic phosphorylation site database (EPSD), which contained 1,616,804 experimentally identified p-sites in 209,326 phosphoproteins from 68 eukaryotic species. In EPSD, we not only collected 1,451,629 newly identified p-sites from high-throughput (HTP) phosphoproteomic studies, but also integrated known p-sites from 13 additional databases. Moreover, we carefully annotated the phosphoproteins and p-sites of eight model organisms by integrating the knowledge from 100 additional resources that covered 15 aspects, including phosphorylation regulator, genetic variation and mutation, functional annotation, structural annotation, physicochemical property, functional domain, disease-associated information, protein-protein interaction, drug-target relation, orthologous information, biological pathway, transcriptional regulator, mRNA expression, protein expression/proteomics and subcellular localization. We anticipate that the EPSD can serve as a useful resource for further analysis of eukaryotic phosphorylation. With a data volume of 14.1 GB, EPSD is free for all users at http://epsd.biocuckoo.cn/.

Key words: phosphorylation; phosphorylation site; post-translational modification; phosphoproteomics; PLK1
Introduction

Protein phosphorylation is one of the most indispensable and well-characterized posttranslational modifications (PTMs). It mainly occurs at specific serine, threonine and tyrosine residues of phosphoproteins, and participates in the regulation of almost all of biological processes and pathways in eukaryotes [1-12]. The regulation of phosphorylation events is highly dynamic and accurate in vivo, whereas aberrance in the phosphorylation system is closely associated with a variety of human diseases, such as cancer, neurodegenerative disease and diabetes [13-17]. Thus, the identification and functional analyses of phosphorylation sites (p-sites) in phosphoproteins provide the foundation for understanding the molecular mechanisms and regulatory roles of protein phosphorylation in eukaryotes.

Advances in chemical labeling strategies, phosphopeptide enrichment techniques and liquid chromatography-tandem mass spectrometry have resulted in rapid progress in high-throughput (HTP) phosphoproteomics, which enables thousands of p-sites to be quantitatively identified in a single run. For example, Humphrey et al. developed a sophisticated phosphoproteomics platform, EasyPhos, to generate time-course maps of insulin signaling dynamics from the large-scale quantification of 31 605 phosphopeptides in mouse liver tissues [1]. Additionally, Liu et al. used EasyPhos to quantify approximately 50 000 p-sites from five mouse brain regions and systematically investigated the in vivo signaling of the kappa opioid receptor, an important member of the G protein-coupled receptor superfamily [2].

In addition, using quantitative label-free technology, Drake et al. identified 8348 phosphopeptides from 27 samples of treatment-naive prostate cancer and metastatic castration-resistant prostate cancer patients and prioritized potentially therapeutic targets in a patient-specific manner [17]. Although many experimental studies have been conducted, the collection, curation, integration and annotation of the large number of p-sites remain to be performed.

In this manuscript, we first reviewed the mainstream p-site databases that were devoted to generate useful resources for academic community, including PhosphoELM/PhosphoBase [3-5, 18], PhosphoSitePlus [6, 7], dbPTM [8, 9], Phospho3D [19, 20], PhosphoPOINT [21], SysPTM [22, 23], PhosphoNET [24], iPTMnet [25, 26], PhOSIDA [26, 27], PhosphoPeP [28, 29], lymPHOS [30, 31], PlattPh [32, 33], PhosphoPhA [34, 35], PDB [36, 37], MPPD [38, 39], FPD [40], PTMfunc [41], UniProt [42], HPRD [43] and BioGRID/PhosphoGRID [44, 45] (Supplementary Table S1). Previously, we also designed two databases, dbPPT [11] and dbPAF [12], which contained 82 175 p-sites in 31 012 proteins of 20 plants, and 483 001 p-sites in 54 148 substrates of human, 4 animals and 2 fungi, respectively. At that time, less annotation information was integrated. To provide a more comprehensive and integrative resource, here, we developed the eukaryotic p-site database (EPSD), by recurating all entries in the dbPPT [11] and dbPAF [12] databases, as well as an additional collection of 1 451 629 known p-sites that were newly identified from HTP studies of 68 eukaryotic species (Figure 1A). We also integrated known p-sites from 13 publicly available p-site databases (Supplementary Table S2). In addition, we provided rich annotations for phosphoproteins and p-sites of eight model organisms by integrating the knowledge from 100 additional resources that covered 15 aspects (Figure 1A, Supplementary Table S2). In total, EPSD contained 1 616 804 known p-sites, including 1 085 095 phosphoserine (pS), 394 648 phosphothreonine (pT) and 13 061 phosphotyrosine (pY) residues in 209 326 phosphoproteins. The EPSD will be continuously maintained and updated and can serve as a useful resource for better understanding of eukaryotic phosphorylation.

Methods

Data collection was started from two previously constructed databases dbPPT [11] and dbPAF [12] which contained 565 176 experimentally identified p-sites of 85 160 phosphoproteins in 27 eukaryotes (Figure 1B). We carefully recurred each entry in the two databases to ensure the quality of the data. Then, we searched the PubMed database by using multiple keywords, including ‘phosphoproteome’, ‘phosphoproteomic’, ‘phosphoproteomics’, ‘large-scale phosphorylation’, ‘mass spectrometry phosphorylation’, ‘MS phosphorylation’, ‘phosphosite’ and ‘phospho-site’. We manually checked all retrieved manuscripts with corresponding supplemental materials and collected 2 509 188 phosphopeptides of 68 eukaryotic organisms from 722 HTP phosphoproteomic studies (Supplementary Table S3). For each study, the information regarding cell or tissue samples was also curated and reserved. In these studies, mass spectrometry spectral data sets were processed by a variety of tools, such as MaxQuant, to derive peptides and phosphopeptides, and a localization probability (LP) score was computationally assigned to each potential p-site in phosphopeptides that contained multiple serine, threonine or tyrosine residues, based on the cumulative binomial distribution [26]. LP scores range from 0 to 1, and a higher LP score represents a higher cumulative binomial probability for a site to be a real p-site [26]. However, LP scores of identified p-sites were only provided in 335 (46.40%) of the 722 studies (Supplementary Table S3). In this regard, as previously described [46], we classified the p-sites in original phosphopeptides of the 335 articles into four categories, including class I (<0.75), class II (<0.75 and >0.5), class III (>0.5 and ≤0.25) and class IV (≤0.25), based on the precalculated LP scores directly taken from these papers. Thus, class I p-sites had a probability of at least 0.75 to be phosphorylated, whereas the cumulative binomial probability of all other phosphorylatable residues was ≤0.25 [26]. To ensure the consistency with the literature, all HTP p-sites were reserved and no additional filters were applied.

To pinpoint the exact positions of p-sites in the protein sequences, we downloaded the reference proteomes containing canonical (protein fasta files) and isoform sequences (additional.fasta files) of the 68 eukaryotes from the UniProt database (Release version 2018-02, ftp://ftp.uniprot.org/) [42]. For each species, we first mapped all phosphopeptides to canonical sequences, and nonmappable phosphopeptides were mapped to isoform sequences. From the HTP studies, we obtained 1 451 629 nonredundant p-sites of 195 351 proteins (Figure 1B).

In addition to the database and literature curation, we further integrated the known p-sites of 13 additional public databases, including PhosphoELM [4, 5], PhosphoSitePlus [6, 7], dbPTM [8, 9], PhosphoPeP [28, 29], PhOSIDA [26, 27], PhosphoPhA [34, 35], PDB [36, 37], SysPTM [22, 23], BioGRID [44, 45], MPPD [38, 39], FPD [40], PTMfunc [41], UniProt [42], HPRD [43] and UniProt [42] (Figure 1B, Supplementary Table S2). Other databases were not considered because the known p-sites were not downloadable (Supplementary Table S1). The details regarding the processing of each phosphorylation database were provided in the Supplementary methods. To avoid any bias, we remapped all p-sites in the dbPPT [11], dbPAF [12] and 13 additional databases to canonical sequences and then to isoform sequences for each organism. We merged this data set with our manually collected p-sites. After redundancy clearing, the EPSD...
A database of phosphorylation sites

Figure 1. The procedure for the construction of the EPSD database. First, we merged the dbPPT [11] and dbPAF [12] databases and carefully re-curated each entry. Next, we searched PubMed to collect newly identified p-sites. Then, known p-sites in 13 additional phosphorylation databases were integrated. In addition to basic annotations, we further annotated eight model organisms, including *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. melanogaster*, *C. elegans*, *A. thaliana*, *S. pombe* and *S. cerevisiae*, by integrating the knowledge from 100 additional resources that covered 15 aspects, including phosphorylation regulator, genetic variation and mutation, functional annotation, structural annotation, physicochemical property, functional domain, disease-associated information, PPI, drug-target relation, orthologous information, biological pathway, transcriptional regulator, mRNA expression, protein expression/proteomics and subcellular localization. (B) The numbers of p-sites and phosphoproteins curated and integrated from dbPPT [11] and dbPAF [12], HTP studies and 13 additional databases.

Results

A summary of the mainstream p-site databases for eukaryotic phosphorylation

The past two decades have witnessed a rapid increase of numbers of p-sites experimentally identified from eukaryotic proteins. The collection, curation, integration and annotation of these known p-sites not only generated useful resources for further experimental analyses, but also provided high-quality benchmark data sets for various computational purposes, including but not limited to the development of p-site predictors [47], the reconstruction of kinase-substrate networks [10, 25, 48, 49], the analysis of genetic variations and cancer mutations that change p-sites [15, 50], the evolutionary analysis of phosphorylation regulation [51, 52], the prioritization of potentially functional p-sites [41] and the identification of PTM crosslinks between phosphorylation and other types of PTMs [9]. Besides our previously developed dbPPT [11] and dbPAF [12], there were at least 20 additional p-site databases, which were briefly reviewed in this manuscript (Supplementary Table S1).

PhosphoBase is the first phosphorylation database that was developed in 1998 [3]. Its 1.0 version only collected 398 p-sites of 156 proteins with from low-throughput (LTP) experiments, whereas its 2.0 release was considerably expanded to contain 414 phosphoproteins with 1052 p-sites [3, 18]. In 2004, Diella et al. manually collected 1703 LTP p-sites of 556 protein substrates from the literature and developed a high-quality database named Phospho.ELM [4]. Later, PhosphoBase was merged into Phospho.ELM, and its 9.0 version also curated known p-sites from various HTP experiments, with a total number of 43 673 p-sites in 11 462 proteins [5]. As one of the best curated and most widely used p-site resources, PhosphoSitePlus was initially established in 2004 and currently contains 381 305 experimental p-sites from both LTP and HTP experiments [6, 7]. In 2006, Lee et al. developed dbPTM for multiple types of PTM sites [8], and known p-sites were directly taken from other databases including Phospho.ELM [4] and UniProt/Swiss-Prot [43]. The 2019 update of dbPTM further collected experimental PTM sites from HTP studies and covered >130 types of PTMs including phosphorylation [9]. dbPTM also provides predicted PTM sites.
with high confidence [9]. Through database integration (DI), Phospho3D obtained 5387 p-sites that were mapped to protein 3D structures [19, 20]. By literature curation and DI, SysPTM 2.0 maintained known sites of ~50 types of PTMs, including 66 017 experimental p-sites in 18 333 proteins [22, 23]. Using a similar strategy, PhosphoNET integrated 22 698 phosphoryl proteins with 177 424 p-sites mainly in Homo sapiens [24]. More recently, through text mining and the integration of six public databases, Huang et al. constructed the iPTMNet database that contained 490 597 curated p-sites for PTM network discovery [10, 25]. It should be noted that the LTP p-sites only occupy a small proportion of totally reported p-sites. Due to rapid progresses in the MS-based phosphoproteomic profiling, a huge number of HTP p-sites has been detected and remained to be curated. In 2006, Olsen et al. developed the first MS-based p-site database, PHOSIDA, and its 2011 update contained 71 638 HTP p-sites of 18 262 proteins in five eukaryotes, including H. sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans and Saccharomyces cerevisiae [26, 27]. Also, PhosphoPep [28, 29] and LymphOS [30, 31] contained 81 335 and 15 566 MS-derived p-sites in four species and from primary human T cells, respectively. Beyond animals, the development of p-site databases for plants or fungi is also prevalent. In 2001, Grishkov et al. constructed the first plant-specific p-site database named PlantsP, which contained >300 HTP p-sites of about 200 plasma membrane proteins in Arabidopsis thaliana [32, 33]. Later, PhosPHat was established with a greatly enhanced number of 28 760 MS-based Arabidopsis p-sites in its 4.0 version [34, 35]. As one of the most comprehensive p-site databases for plants, P^D^IB integrated 41 746 HTP p-sites of 14 299 proteins in nine plants [36, 37]. In particular, MP^D^D was designed as a well-organized resource by collecting 2650 HTP p-sites in roots of Medicago truncatula [38, 39], whereas FP^D^P is the only fungus-specific database that contained 56 133 MS-derived p-sites for 11 fungi [40]. By compiling a large data set of multiple types of PTM sites from 11 eukaryotes, Beltrao et al. conducted a systematic prediction of potentially functional PTMs including phosphorylation, and related results were present in PTMFunc [41]. Besides professional databases that were mainly focused on eukaryotic phosphorylation or PTMs, a number of resources developed for more general purposes also curated and annotated p-sites. For example, as one of the most indispensable resource for protein annotations, UniProt also curated 18 466 phosphorylated proteins with 66 490 known p-sites [42]. HPRD, the best annotated resource for human proteins, collected 55 849 unique p-sites in 9541 human proteins release [43]. BioGRID, a widely used database of experimental protein–protein interactions (PPIs), also presented a subdatabase named PhosphoGRID to maintain 23 064 p-sites in 3264 budding yeast proteins [44, 45]. Undoubtedly, all these databases provided invaluable information for further analysis of eukaryotic phosphorylation.

A multilayer annotation of phosphoproteins and p-sites

In this study, we designed EPSD as a phosphoprotein-centered database. The EPSD ID (EP-) was automatically generated and assigned to each phosphoprotein as the primary accession, whereas the UniProt ID was used as the secondary accession for organizing the database. The basic annotations such as gene/protein names, GenBank Gene IDs, nucleotide and protein IDs, Ensembl gene, transcript and protein IDs, functional descriptions, protein sequences, nucleotide sequences of the coding region, keywords and gene ontology (GO) terms were taken from UniProt [42]. The annotations of major or minor isoforms were referenced from the canonical or isoform sequences downloaded from UniProt [42], respectively. By using 100 additional databases and computational tools that covered 15 aspects, we further annotated 97 871 phosphoproteins in eight model organisms, including H. sapiens, M. musculus, Rattus norvegicus, D. melanogaster, C. elegans, A. thaliana, Schizosaccharomyces pombe and S. cerevisiae (Figure 1A, Supplementary Table S2). The knowledge in these resources was carefully integrated, and the details regarding the processing of each resource are provided in the Supplementary methods. For example, to annotate upstream regulators of p-sites, we obtained experimentally identified kinase-substrate relations from PhosphoSitePlus [6, 7], Phospho.ELM [4, 5], PostMod [53], PSEA [54], PhosphoNetworks [55] and RegPhos [56], and we obtained known phosphatase-substrate relations in H. sapiens from HuPHO [57] and DEPOD [58]. We also used four tools, including GPS [59], NetworkIN [48], PKIS [60] and PhosphoPick [61], to predict potential protein kinases for modifying the p-sites. In total, 12 databases or tools were adopted for the annotation of protein kinases or protein phosphatases that potentially modify or demodify known p-sites in the eight species (Supplementary Methods). All data sets and annotations can be accessed at http://epsd.biocuckoo.cn/Download.php.

The data statistics of p-sites in the EPSD

In the EPSD, there are 1 616 804 experimentally identified p-sites of 209 326 nonredundant proteins integrated for 68 eukaryotes, including 18 animals, 7 protists, 24 plants and 19 fungi (Supplementary Table S4). The distribution of p-sites and unique proteins, as well as pS, pT and pY residues for each species, is illustrated in a heatmap (Figure 2A), whereas the detailed counts are shown in Supplementary Table S4. For all p-sites, there were 1 085 095 (67.11%) pS, 394 648 (24.41%) pT and 137 061 (8.48%) pY residues. We analyzed the distribution of the numbers of p-sites in phosphoproteins, and the results showed that up to 152 577 proteins (72.89%) are multiply phosphorylated with ≥2 known p-sites, indicating that multisite phosphorylation is the predominant mechanism for the regulation of protein substrates (Figure 2B). In particular, 4457 (2.13%) proteins are heavily phosphorylated with ≥50 p-sites, indicating a highly complicated phosphorylation code in these proteins (Figure 2B). From the analysis of the disorder propensities of p-sites in eight species, we observed that only pY residues had a strong preference (74.01%) to locate in protein ordered regions (Figure 2C). Moreover, the distribution of numbers of human p-sites in different cells or tissues was counted, and we found that 167 459 (37.57%) of known p-sites were only detected in one cell/tissue sample (Figure 2D). Whether these sites were cell- or tissue-specific remained to be validated. In addition, the major changes after re-curating entries in dbPPT [11] and dbPAF [12] were analyzed. From the 560 801 items reserved in EPSD, up to 525 134 (93.64%) p-sites were identical to the previous two databases without any changes (Figure 2E). There were 3661 (0.65%) p-sites with position changes, and 32 006 (5.71%) p-sites underwent changes in UniProt IDs (Figure 2F). Both changes were resulted from the updates of sequences or accession numbers in UniProt [42]. We deleted 4375 p-sites which were mistakenly collected into the two previous databases.

The usage of EPSD

The online service of the EPSD was developed in an easy-to-use manner. As an example, here, we used human PLK1
A database of phosphorylation sites

Figure 2. The data counts of p-sites in the EPSD. (A) A heatmap of the distribution of p-sites, protein entries and pS, pT and pY residues, for each species in the EPSD. More details are shown in Supplementary Table S4. (B) The distribution of numbers of p-sites in protein substrates. (C) The disorder propensity scores were calculated by IUPred [62] for p-sites in eight species. (D) The distribution of numbers of human p-sites in different cells or tissues. (E) The major changes after rechecking entries in dbPPT [11] and dbPAT [12], including identical, position changed (Pos. changed), UniProt ID changed and deleted p-sites.

protein (UniProt ID: P53350), an important serine/threonine kinase that plays important roles in mitosis [63], autophagy and cell death pathways [13], to describe the usage of the EPSD. From the home or search page (http://epsd.biocuckoo.cn/Search.php), by using the option of ‘Substrate Search’, users can select the ‘Protein Name’ and input ‘Serine/threonine-protein kinase PLK1’ to query the database (Figure 3A). The results will be shown in a tabular format with information for the EPSD ID, UniProt accession, gene name, protein name and species (Figure 3A). By clicking on ‘EP0017523’, the detailed information for human PLK1 can be viewed. In the EPSD, we mapped all integrated phosphoproteins to PDB [64] and found 5774 proteins containing at least one 3D structure. For proteins such as human PLK1, a representative 3D structure was visualized by 3Dmol.js (http://3dmol.csibaba.edu) [65] (Figure 3B). All p-sites or a specific p-site can be selected for visualization in the PLK1 structure (Figure 3B).

Since previous studies demonstrated that p-sites located in protein disordered regions evolve rapidly and are likely to result in nonfunctional phosphorylation events [52, 66], the EPSD used the program package of IUPred (https://iupred.elte.hu) [62], a tool for the prediction of disordered protein regions, to calculate the disorder propensity score for each residue of the phosphoproteins in the eight organisms. Such information will be helpful for biologists to prioritize potentially functional p-sites. For convenience, all p-sites of PLK1 were summarized in a schematic diagram of the protein sequence together with the calculated disorder propensity score for each residue, while the details of p-sites are shown in a tabular list (Figure 3C). For each known p-site, the position, phosphopeptide centered on the pS/pT/pY residues flanked by seven residues upstream and seven residues downstream, and references are provided (Figure 3C). By clicking on the ‘References’ of a known p-site, details regarding the original phosphopeptides, peptide IDs, classes, LP scores, cell/tissue samples or integrated sources and PMIDs of references can be viewed (Figure 3C).

More annotations of human PLK1 can be accessed by clicking on either the ‘Integrated Annotations’ button (Figure 3B) or the ‘Annotation’ option in the left column to reveal the annotation summary (Figure 3D). By clicking on the button ‘PhosphoSitePlus’ in the annotation list (Figure 3D), protein kinases that phosphorylate PLK1 curated in PhosphoSitePlus [6, 7] can be viewed (Figure 3D). The details of the integrative annotations for human PLK1 are shown in Figure 4. The PLK1 T210 residue is a critical p-site, which can be phosphorylated by Aurora kinase A (AurA), and its phosphorylation level positively correlates with the kinase activity [13]. Phospho.ELM only curated one regulatory kinase LOK of T210 [4, 5], whereas RegPhos [56] annotated T210 modified by AurA and LOK. Also, PSEA collected three kinases, including AurA, Cot and LOK for modifying T210 [53]. PhosphoSitePlus [6, 7] curated four kinases with an additional one as Aurora kinase B (AurB), which was reported to phosphorylate T210 during mitosis [67] (Figure 4). In regard to genetic variation and mutation, we found that cancer mutations annotated in The Cancer Genome Atlas (TCGA) [68] can alter two p-sites including S326 and Y582 in PLK1, while the functional consequences of these changes remain to be studied.
Figure 3. The ‘Substrate Search’ option of the EPSD. (A) One or multiple keywords such as ‘Serine/threonine-protein kinase PLK1’ can be inputted, and the results will be returned in a tabular format. (B) The protein entry of human PLK1. (C) The p-sites of PLK1 are shown in a schematic diagram of the protein sequence and in a tabular list. (D) The annotation summary of PLK1. (E) Protein kinases annotated in PhosphoSitePlus [6, 7] that phosphorylate PLK1.

For functional annotations, PLK1 was annotated as an autophagy and cell death regulator in our recently developed database THANATOS (40). There were 19 protein 3D structures of PLK1 maintained in PDB [64], and its molecular weight and Isoelectric point were calculated as 68 123.59 Da and 9.09, respectively (Figure 4). The functional domains in PLK1 were annotated. The phosphorylation level of PLK1 T210 is significantly up-regulated in pediatric acute lymphoblastic leukemia, and such a PTM-disease association (PDA) was curated in PTMD [69] (Figure 4). Moreover, 2940 interacting partners and 200 records of drug-target relations, as well as the orthologous information of PLK1, are present (Figure 4). PPIs were integrated from seven public databases, including STRING [70], IID [71], inBio Map [72], Mentha [73], HINT [74], iRefIndex [75] and PINA [76]. For each resource, we counted the number of PPIs (Supplementary Figure S1A) and interacting proteins (Supplementary Figure S1B) in *H. sapiens*. It was found that only 8885 PPIs (0.18%) and 8545 proteins (27.01%) were annotated by all databases. We also counted the number of interacting partners of human PLK1 and found only 28 proteins covered by all databases (Supplementary Figure S2A). In particular, Volasertib (Figure 4) has been approved by the Food and Drug Administration as the breakthrough therapy for the treatment of acute myeloid leukemia patients [16]. A number of biological pathways that involve PLK1 are also shown (Figure 4), and PLK1 is activated by E2F Transcription Factor 1 (E2F1) annotated by TRRUST [78]. From the TCGA [68], International Cancer Genome Consortium (ICGC) [79], BioXpress [80] and analogous databases, 36 476 records of mRNA expression in 48 cancer types were integrated for PLK1, which is significantly overexpressed in breast cancer (Figure 4). Meanwhile, EPSD integrated the protein expression levels of PLK1 measured in 47 tissues, which recorded that PLK1 is highly expressed in testis (Figure 4). As a multilocalized protein, human PLK1 can locate at a variety of subcellular compartments, such as centrosome, kinetochore and midbody (Figure 4).

Discussion

Protein phosphorylation is one of the most ubiquitous and well-studied PTMs, and recent advances in large-scale phosphopro-
The collection, biocuration, integration and annotation of experimentally characterized p-sites will provide highly useful resources for better understanding the molecular mechanisms and biological functions of eukaryotic phosphorylation. Although a dozen phosphorylation databases have been established, two issues require attention. First, the continuous update of the databases through literature curation is important for providing a more integrative data set of the known p-sites. For example, although dbPPT [11] and dbPAF [12] already contain 565,176 known p-sites, the update of the EPSD still expanded the number to 1,616,804, with a 2.86-fold increase. Second, multidimensional annotations of p-sites from various types of public resources will greatly enhance the usefulness of the databases. In fact, a number of phosphorylation databases have included other types of annotations. For example, dbPTM integrated PDAs and PTM crosstalk events in its 2019 update [9], whereas the PhosphoSitePlus 2014 update provided at least three highly quality annotation files, including ‘Kinase Substrate’ for manually curated kinase-substrate relations, ‘PTMVar’ for missense mutations occurring at or around p-sites and ‘Regulatory sites’ for functional p-sites [6]. In the EPSD, we integrated 15 types of annotations from 100 public resources for eight model organisms. Compared with dbPPT and dbPAF (~0.35 GB), the EPSD has a total data size of 14.1 GB, which represents a 40-fold increase in data volume (Supplementary Table S5).

Besides providing comprehensive annotations for individual phosphoproteins and p-sites, the data in EPSD could also be used for other purposes, e.g. cancer systems biology studies in the network or pathway level. Recently, cancer genome sequencing provided a great opportunity in profiling the molecular landscapes of human tumors, while the modeling of cancer-associated signaling networks was helpful for analyzing the cancer evolutionary dynamics, detecting early-warning signals, probing tumor heterogeneity and prioritizing potential agents to overcome drug resistance [81, 82]. How human signaling networks are rewired by cancer mutations remained to be dissected [83, 84]. Here, we focused on the phosphoinositide 3-kinase (PI3K)/protein kinase B (AKT)/mammalian target of rapamycin (mTOR) pathway, one of the most aberrantly activated pathways in driving cancer progression [81, 84, 85], and related annotations were directly retrieved from EPSD. We re-illustrated the simplified pathway and observed that all the 21 major components are phosphorylated, with at least three known p-sites changed by cancer mutations integrated from TCGA [68], ICGC [79] and COSMIC [86] (Supplementary Figure S3). In the pathway, class I PI3Ks phosphorylate phosphatidylinositol-4,5-bisphosphate (PIP2) into phosphatidylinositol-3,4,5-trisphosphate (PIP3) for promoting the activation of downstream signaling, whereas PTEN acts as a tumor suppressor and inhibits the pathway activation through dephosphorylating PIP3 to PIP2 [87] (Supplementary Figure S3). The enzyme activities of both PI3Ks and PTEN are regulated...
by phosphorylation [87, 88]. PIK3CB, a member of class I PI3Ks, is autophosphorylated at S1070 that inhibits its lipid kinase activity [88]. From EPSD, we found a missense mutation S1070Y from ICGC colon adenocarcinoma, and such a mutation will enhance the lipid kinase activity of PIK3CB to produce more PIP3 molecules, which might up-regulate the PI3K/AKT/mTOR pathway. Although it was well documented that the C-terminal phosphorylation of PTEN plays a critical role in regulating its phosphatase activity [87], EPSD contained an HTP p-site at S170, and a mutation of S170 N was demonstrated to significantly disrupt its activity [89]. In EPSD, the S170 N mutation could be found in multiple cancer types and might contribute to activate the PI3K/AKT/mTOR pathway through the accumulation of PIP3. Further studies need to be conducted to exploit whether other cancer mutations that change p-sites are also involved in the dysregulated activation of the PI3K/AKT/mTOR pathway.

For the future, we will continuously maintain and update the EPSD database, when newly identified p-sites are reported in the literature. More species will be added with experimentally identified p-sites. Additionally, we will include more annotations from other public resources to provide a more comprehensive resource for eukaryotic phosphorylation.

Key Points

• We reviewed the mainstream databases for the collection and annotation of protein p-sites in eukaryotes.
• We developed an integrative resource named EPSD which contained 1 616 804 known p-sites in 209 326 proteins from 68 eukaryotic species, by both literature curation and database integration.
• We provided rich annotations for phosphoproteins and p-sites of eight model organisms by integrating the knowledge from 100 additional resources that covered 15 aspects.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

The authors thank Drs Wankun Deng, Haodong Xu and Ying Zhang for their helpful comments during the database construction. American Journal Experts reviewed the manuscript prior to submission.

Funding

Funding for open access charge: Special Project on Precision Medicine under the National Key R&D Program (2017YFC0906600, 2018YFC09010500), the Natural Science Foundation of China (31930021, 31970633, 31671360, 81701567), the Fundamental Research Funds for the Central Universities (2017KFKJC001, 2019KfYRCY043), Changjiang Scholars Program of China, and the program for HUST Academic Frontier Youth Team.

References

64. Burley SK, Berman HM, Bhikadiya C, et al. RCSB protein data bank: biological macromolecular structures enabling research and education in fundamental biology,