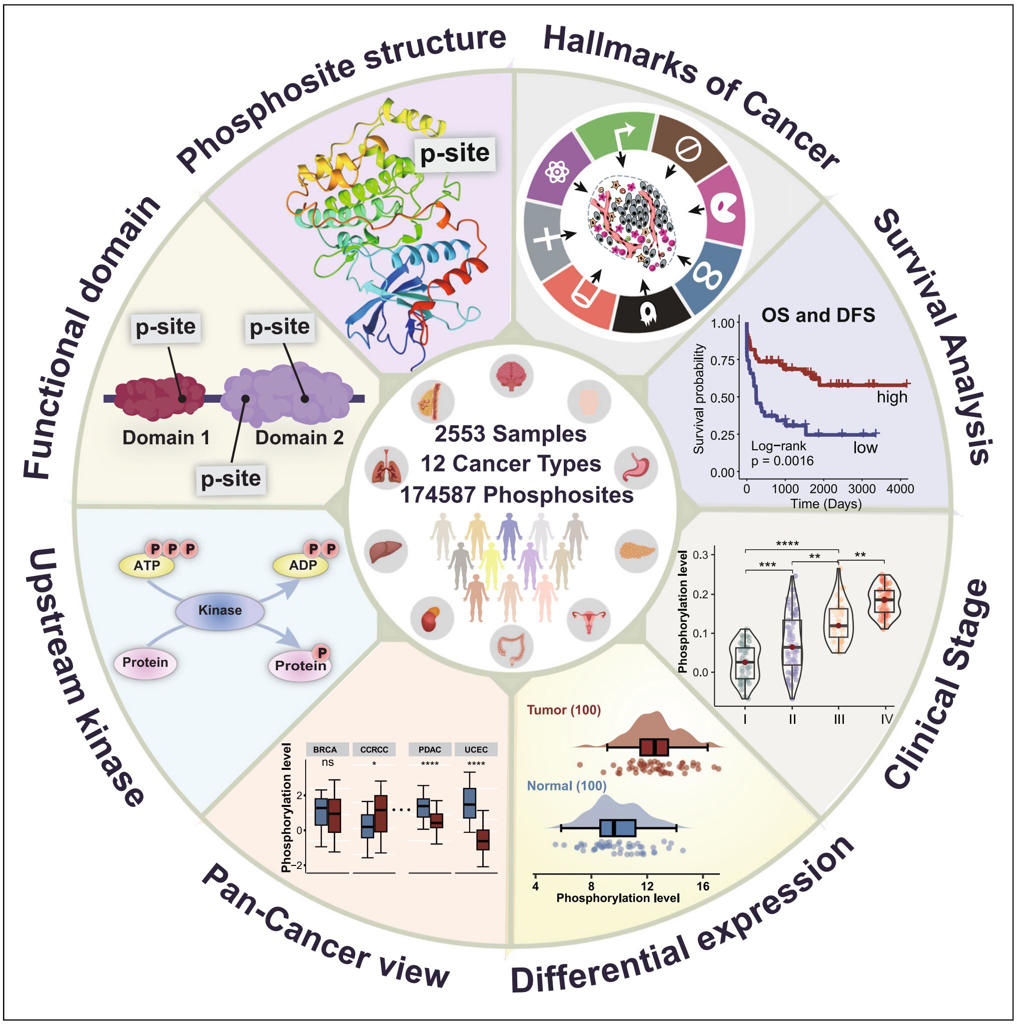
iScience

CellPress
OPEN ACCESS

Article

PhosCancer: A comprehensive database for investigating protein phosphorylation in human cancer



Qun Dong, Danqing Shen, Jiachen Ye, Jiaxin Chen, Jing Li

jing.li@sjtu.edu.cn

Highlights

Analyzed 174,587 phosphosites from 2,553 samples across 12 cancer types

Provided 3D structures, functional domains, kinases, and other essential annotations

Offered quantitative associations with nine clinical features and cancer hallmarks

Presented each quantitative analysis in both pan-cancer and cancer-specific views

Dong et al., iScience *27*, 111060
November 15, 2024 © 2024
Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111060

iScience

CellPress
OPEN ACCESS

Article

PhosCancer: A comprehensive database for investigating protein phosphorylation in human cancer

Qun Dong,[1] Danqing Shen,[1] Jiachen Ye,[1] Jiaxin Chen,[1] and Jing Li[1,2,*]

SUMMARY

Protein phosphorylation is a crucial post-translational modification implicated in cancer pathogenesis, offering potential diagnostic and therapeutic targets. Here, we developed PhosCancer, a user-friendly database for extracting biologically and clinically relevant insights from phosphoproteomics data. Leveraging data from the CNHPP and CPTAC, PhosCancer encompasses 174,587 phosphosites from 14 datasets spanning 12 cancer types. Through extensive statistical analyses and integration of annotations from external resources, PhosCancer serves as a convenient one-stop platform facilitating the exploration of phosphorylation profiles across different cancer types. Not only does PhosCancer encompass basic information, 3D structure, functional domains, and upstream kinases, but also provides quantitative associations with nine clinical features, and the relevance with hallmarks in both cancer-specific and pan-cancer views. PhosCancer is a valuable resource for cancer researchers and clinicians, promoting the identification of clinically actionable biomarkers and further facilitating the clinical applications of phosphoproteomic data.

INTRODUCTION

Protein phosphorylation is one of the most important post-translational modifications (PTMs), mainly occurring at specific serine (Ser/S), threonine (Thr/T), and tyrosine (Tyr/Y) residues.[1] This reversible modification acts as a dynamic molecular switch, regulating a broad spectrum of biological processes. Aberrant phosphorylation can induce alterations in enzymatic activity, subcellular localization, ligand binding, and interaction with other proteins, thereby affecting the dysfunctions of downstream pathways and ultimately contributing to various pathological conditions, especially carcinogenesis.[2,3] During recent decades, tremendous efforts have been made by researchers to assess the applicability of protein phosphorylation changes in cancer diagnosis, prognosis, and therapeutic strategies. Notable findings include the significant association between elevated phospho-Ser784-VCP levels and poor prognosis in breast cancer,[4] and Rb phosphorylation as a driver and potential therapeutic target in colon cancer.[5]

Advancements in mass spectrometry-based technologies enable the detection and profiling of phosphosites with unprecedented sensitivity, offering new avenues for investigating their biological functions. Cancer proteomics consortia, like the Chinese Human Proteome Project (CNHPP) and Clinical Proteomic Tumor Analysis Consortium (CPTAC), provide a wealth of phosphoproteomics data across various cancer types. The datasets hold significant value in enhancing our understanding of the molecular mechanisms underlying cancer and in identifying novel diagnostic and therapeutic biomarkers. However, the analysis of raw data from such consortia is labor-intensive, time-consuming, and requires specialized bioinformatics expertise. Moreover, information on the associations between phosphorylation and cancer is scattered across disparate phosphoproteomics studies, thus greatly limiting the clinical translation of these data. Although several databases—dbPTM,[6] EPSD,[7] Phospho.ELM,[8] and PhosphoSitePlus[9] provide the structural views and literature-based curation of phosphosites, they lack information on abundance patterns or quantitative annotation of phosphorylation in cancers. Other databases, such as CPPA,[10] iProPhos,[11] cProsite,[12] and UALCAN,[13] offer critical quantitative analyses for phosphorylation and have made significant contributions to cancer phosphorylation research, but each has certain limitations. CPPA relies on external databases and software predictions for upstream kinase data without quantitative kinase-substrate correlations and lacks functional analysis of phosphorylation. iProPhos offers data-based kinase-substrate correlations but lacks experimentally validated information and a pan-cancer perspective. cProsite only displays changes in phosphorylation levels between tumor and normal samples, and correlations among phosphorylation levels. UALCAN only focuses on analysis across race, age, subtype, and multiple hallmark pathways for phosphorylation. Overall, CPPA, iProPhos, cProsite, and UALCAN mainly offer certain quantitative analyses without qualitative annotations, which still results in fragmented phosphorylation information.

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Lead contact
*Correspondence: jing.li@sjtu.edu.cn

https://doi.org/10.1016/j.isci.2024.111060

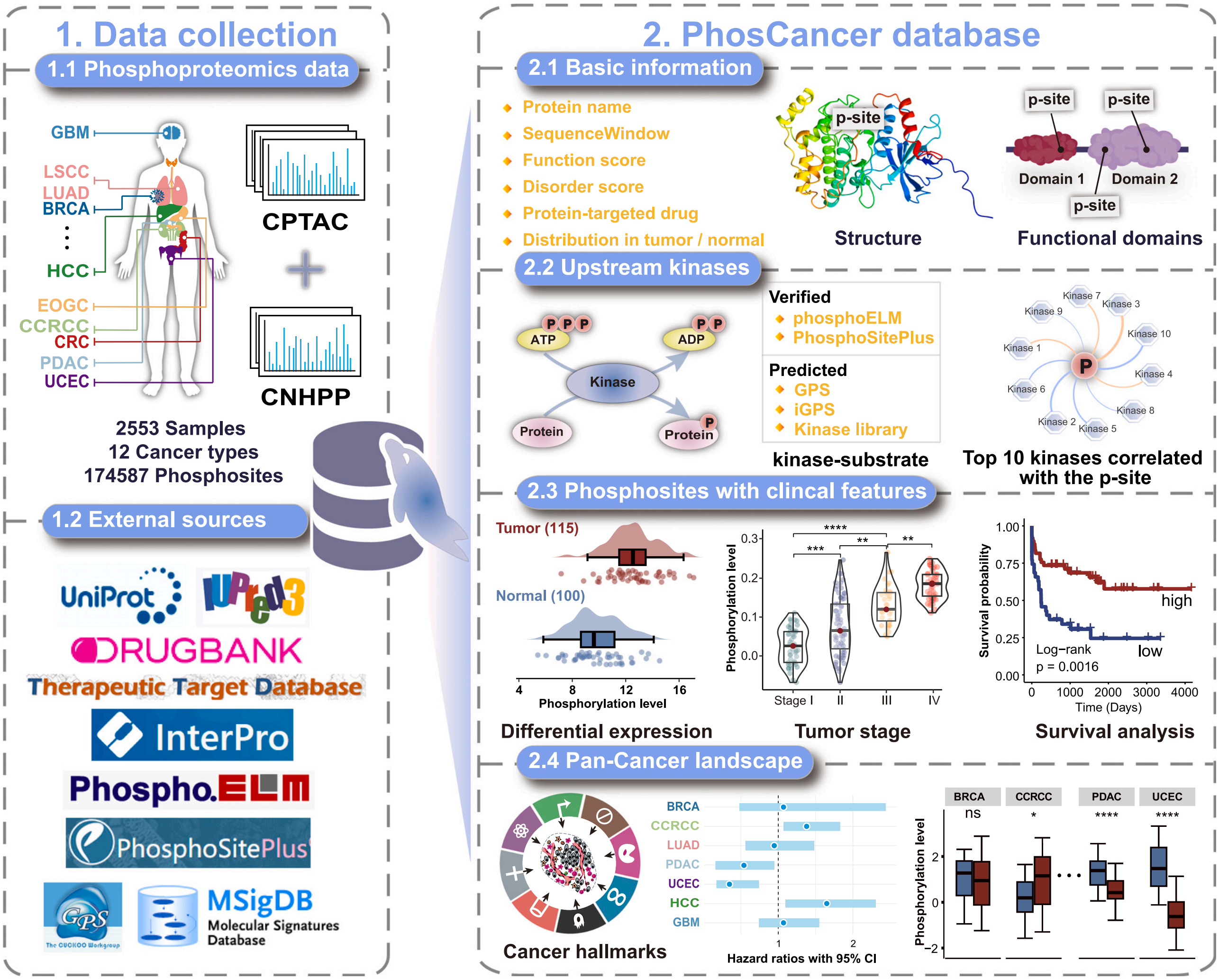


Figure 1. An overview of the PhosCancer database

First, we manually curated raw phosphoproteomics data from all available cancer types from CNHPP and CPTAC, encompassing 14 datasets, and preprocessed the data using standardized pipelines. Subsequently, we conducted detailed analyses and integrated various annotations from well-established resources to develop the PhosCancer database. PhosCancer provides multi-dimensional annotation information for each phosphosite.

Therefore, we developed PhosCancer, an integrative and user-friendly web server for exploring protein *Phos*phorylation in Human *Cancer*. We collected phosphoproteomics data from a wide array of cancer types sourced from the CNHPP and CPTAC, comprising 174,587 nonredundant phosphosites from 14 datasets across 12 cancer types. Through a series of statistical analyses and integration of essential annotations from multiple external resources, PhosCancer not only provides a rich knowledge repository but also comprehensively characterizes phosphorylation profiles within and across different cancer types. PhosCancer presents an integrated presentation of annotations, statistical analyses, and visualization through a flexible web interface. PhosCancer will be an invaluable resource for cancer biologists and clinicians, facilitating the identification of novel diagnostic and therapeutic targets and expediting the clinical translation of cancer phosphoproteomic data. PhosCancer is freely available at https://lilab.life.sjtu.edu.cn/PhosCancer.

RESULTS

Database construction and content

PhosCancer is a systematic resource designed for exploring protein phosphorylation in human cancer, depicted in Figure 1. To construct a comprehensive database, we collected raw phosphoproteomics and proteomics data from almost all available cancer types from the CNHPP and CPTAC data portals. Subsequently, the collected data were processed and analyzed using a standardized pipeline. Intending to serve as a convenient one-stop platform, we integrated key annotations from various resources alongside data-driven analyses to enrich functional and structural insights. Each phosphosite is linked to detailed annotations across seven categories: (i) basic information: provides essential details such as the gene name, protein identifier, and information on drugs targeting the protein. (ii) 3D structure: offers visualization of phosphosite into the amino acid sequence context and 3D structure to derive insights and hypotheses regarding biological functions and mechanisms. (iii) Functional domains: identifies protein functional domains containing the phosphosite, highlighting potential functional implications based on protein architecture. (iv) Upstream kinases: links phosphosites to known or predicted upstream kinases responsible for its phosphorylation, offering mechanistic insights into regulatory networks. Correlations between all protein kinases[14] and phosphorylation levels based on our 14 datasets were provided to enhance the identification of potential kinase-substrate interactions. (v) Associations with clinical features: investigates associations with tumor status (normal vs. tumor), stages, survival outcomes, age, gender, BMI, race, tumor size, and subtypes, aiding in the identification of clinically relevant biomarkers. (vi) Relevance with hallmarks: evaluate the correlation of phosphosites with hallmarks, inferring their potential roles in human cancer. (vii) Pan-cancer view: offers an integrative overview of the phosphorylation alterations across different cancer types, enabling cross-comparisons and facilitating the identification of phosphosites with shared effects on cancer, thereby offering potential implications for multi-cancer diagnosis and therapeutic strategies.

The current version of PhosCancer comprises 174,587 nonredundant phosphosites derived from 14 datasets, including breast invasive carcinoma (BRCA),[15] clear cell renal cell carcinoma (CCRCC),[16] colon adenocarcinoma (CRC),[5] early-onset gastric cancer (EOGC),[17] glioblastoma (GBM),[18] (HBV)-related hepatocellular carcinoma (HCC),[19] head and neck squamous cell carcinoma (HNSCC),[20] lung squamous cell carcinoma (LSCC),[21] lung adenocarcinoma (LUAD),[22] ovarian serous cystadenocarcinoma (OV),[23] pancreatic ductal adenocarcinoma (PDAC)[24] and uterine corpus endometrial carcinoma (UCEC)[25] in CPTAC, hepatocellular carcinoma (HCC_cnhpp)[26] and lung adenocarcinoma (LUAD_cnhpp)[27] in CNHPP. The distribution of phosphosites across these datasets is illustrated in Figure 2A, with counts ranging from 60,492 (LSCC) to 25,617 (LUAD_cnhpp). Phosphorylation primarily targets pS residues (66.59%), with lesser proportions observed on pT (22.75%) and pY (10.66%) residues, consistent with prior research.[28] These 174,587 phosphosites occur on only a subset of 16,034 proteins (Figure 2B). Among these proteins, up to 14205 (88.59%) exhibit multiple phosphorylation events with at least two phosphosites, highlighting multisite phosphorylation as the primary mechanism regulating protein substrates (Figure 2C). Remarkably, 477 proteins (2.97%) are extensively phosphorylated with fifty or more phosphosites, potentially indicating critical roles within complex cellular signaling networks. Phosphosites tend to be enriched in disordered protein regions, whereas pY residues show a preference for localization within ordered protein regions (Figure 2D). No apparent sequence preference in phosphorylation was observed (Figure 2E). These 14 datasets comprise 2553 samples including 1503 tumor samples and 1050 paired or unpaired normal samples (Figure 2F). Data-driven analyses primarily include correlation analysis of phosphorylation levels and kinase expression (KinaseCor), differential analysis between tumor and normal samples (DE), differential analysis across tumor stages (Stage), survival analysis (Survival), and correlation analysis results of phosphorylation levels with hallmarks (Hallmark). Due to high rates of missing values, certain phosphosites were excluded from statistical analyses within PhosCancer (Figure 2G), and qualitative information was provided for these phosphosites.

To further explore the clinical associations of phosphosites from a global perspective, we screened for associations including DE (|fold change|>1.2 & FDR<0.05), Survival ($p < 0.05$), and Stage ($p < 0.05$). Our analysis revealed a substantial proportion of phosphosites associated with multiple cancer types, with 63.05% in DE and 63.17% in Survival (Figure 3A). Among these, 6,442 (12.61%) phosphosites in DE and 1,657 (3.41%) in Survival were associated with seven or more cancer types. Specifically, seven phosphosites in DE were linked to 13 cancer types, including ABLIM3 pS503, HMGA1 pS36, NCL pS67, NUCKS1 pS181, DDX21 pS121, NOC2L pS49, and MCM2 pS27. The critical roles of certain phosphosites have been established; for instance, phosphorylation of the transcription factor NUCKS1 at S181 by CDK1 has been shown to attenuate its DNA binding ability.[29] To further explore the clinical significance of phosphosites on a global scale, we constructed a pan-cancer phosphosite association network. This network focused on phosphosites with differential phosphorylation in at least seven cancer types, applying stringent criteria (|fold change|>2, FDR<0.05, and a minimum sample size of 15 for both cancer and normal tissues). The network included 1,429 cancer-phosphosite associations, covering 13 cancer types and 192 phosphosites (Figure 3B). Among these, 136 phosphosites (70.83%) exhibited consistent differential expression across their respective cancer types. Specifically, 75 phosphosites were consistently upregulated in cancer tissues, while 61 were consistently downregulated compared to normal tissues. The top 20 phosphosites with the most associations across cancer types are highlighted in the network. Notably, some of these phosphosites have well-documented functions, such as TOP2A pS1106 associated with nine cancer types,[30] CDC20 pT106 associated with nine cancer types,[31] and ECT2 pT359 associated with ten cancer types.[32] The strong connections within this network suggest pan-cancer patterns of protein phosphorylation changes, underscoring the potential significance of these phosphosites in multi-cancer diagnosis and therapeutic strategies.

Database usage

The PhosCancer is a web-accessible and comprehensive open resource for interactive exploration, browsing, searching, visualization, and downloading. Within the Browse" section, users can access all database results. The "Quick Search" allows users to search easily for information using a single keyword of interest, gene symbol, or UniProt ID. For more specific queries, an advanced search option is available on the "Search" page. Here, users can retrieve desired phosphosites based on specific keywords including gene/protein, position, cancer type, protein correction status, missing value ratio, and different significance thresholds ($p < 0.05$, $p < 0.01$, FDR< 0.05, and FDR<0.01) for DE, Stage and Survival analysis. Users have the flexibility to perform keyword searches individually or in combination to refine their queries.

Detailed pages for each phosphosite can be accessed through browsing or searching (Figure 4). The page consists of six modules: "Basic info", "Upstream kinases", "DE", "Stage", "Survival", "Subtype", "More Clinical" and "Hallmark". Clicking on a module provides access to

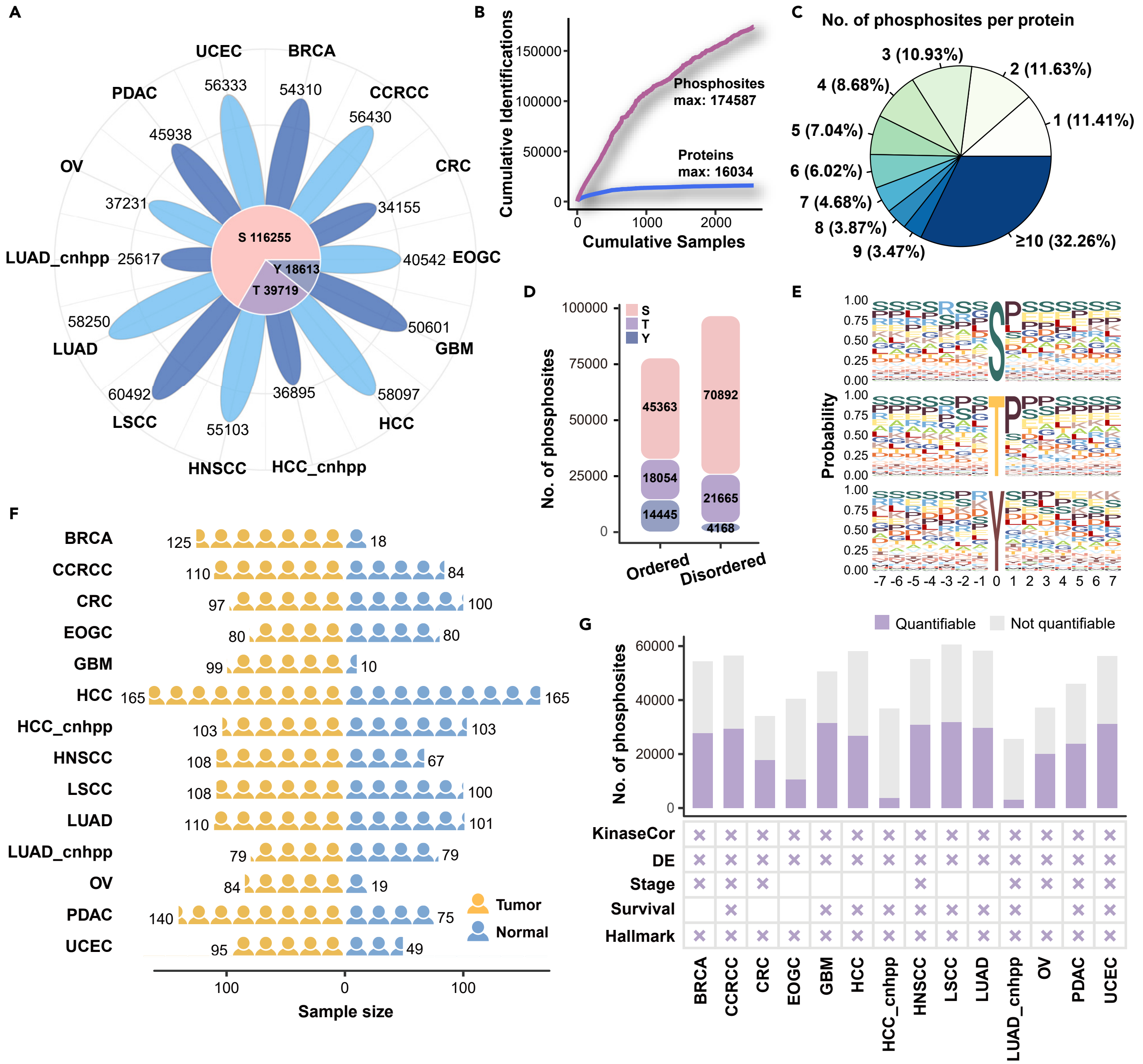


Figure 2. The data statistics of PhosCancer

(A) Phosphosite counts across 14 datasets and the distribution of residue types.

(B) Identification numbers of phosphosites (purple) and phosphoproteins (blue) with sample accumulation.

(C) Percentage of numbers of phosphosites in protein substrates.

(D) Distribution of phosphosites within or outside intrinsically disordered regions.

(E) Sequence preference analysis of phosphosites in homo sapiens, ordered by pS, pT, and pY residues.

(F) Tumor and normal sample sizes for each dataset.

(G) Numbers of quantifiable phosphosites with missing values below 80% and non-quantifiable phosphosites annotated only qualitatively in PhosCancer. Clinical feature data availability is summarized at the bottom.

the corresponding results. Here, we illustrate the utility of PhosCancer using phosphosite S27 on MCM2 (UniProt ID: P49736) as an example. The significant role of dysregulated MCM2 phosphorylation has been widely recognized and is implicated in the pathogenesis of various cancers.[33,34] The "Basic info" module provides basic information about the phosphosite, including UniProt ID, protein name, gene name, position, sequence window, function score, disorder score, drugs targeting the protein, protein subcellular localization, protein function, and functional domains where the phosphosite is located. PhosCancer reveals that MCM2 pS27 exhibits a functional score of up to 0.95 and a disorder score of 0.89 (Figure 4A). Additionally, the distribution of phosphorylation levels across all 14 cohorts, including tumor and normal

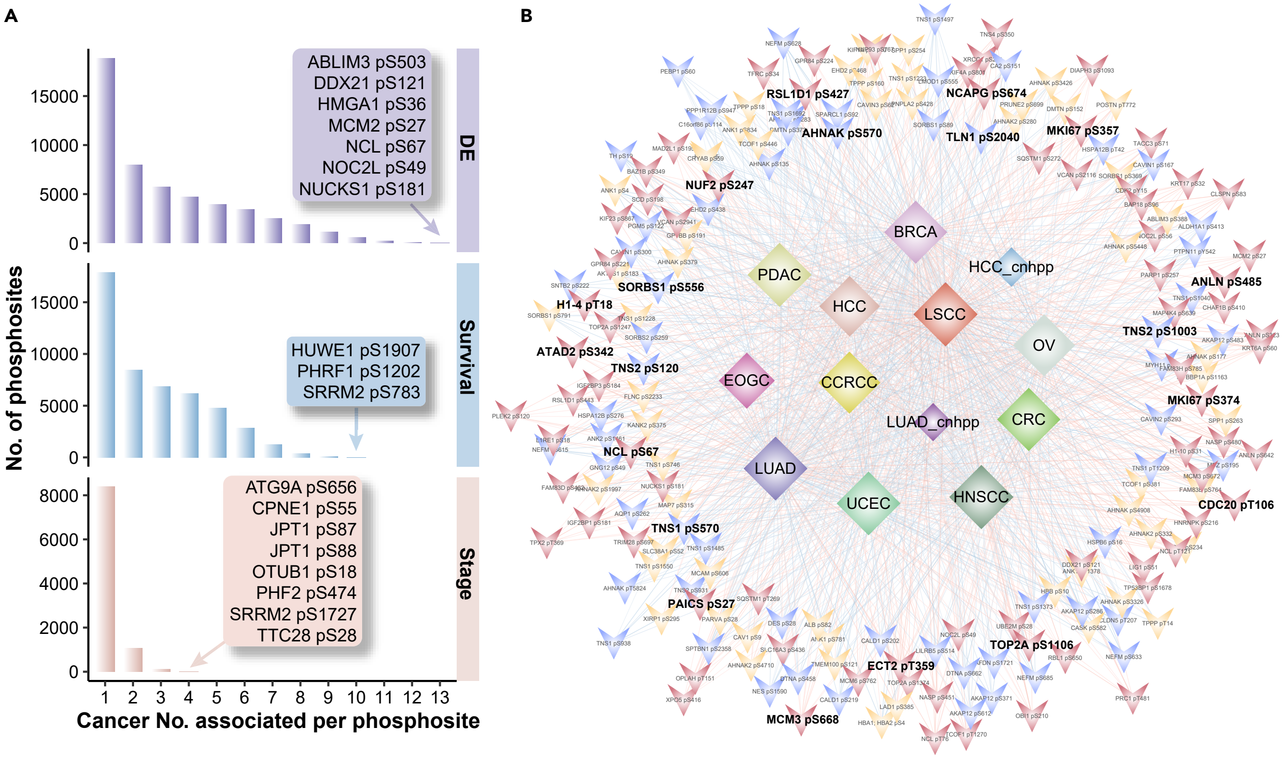


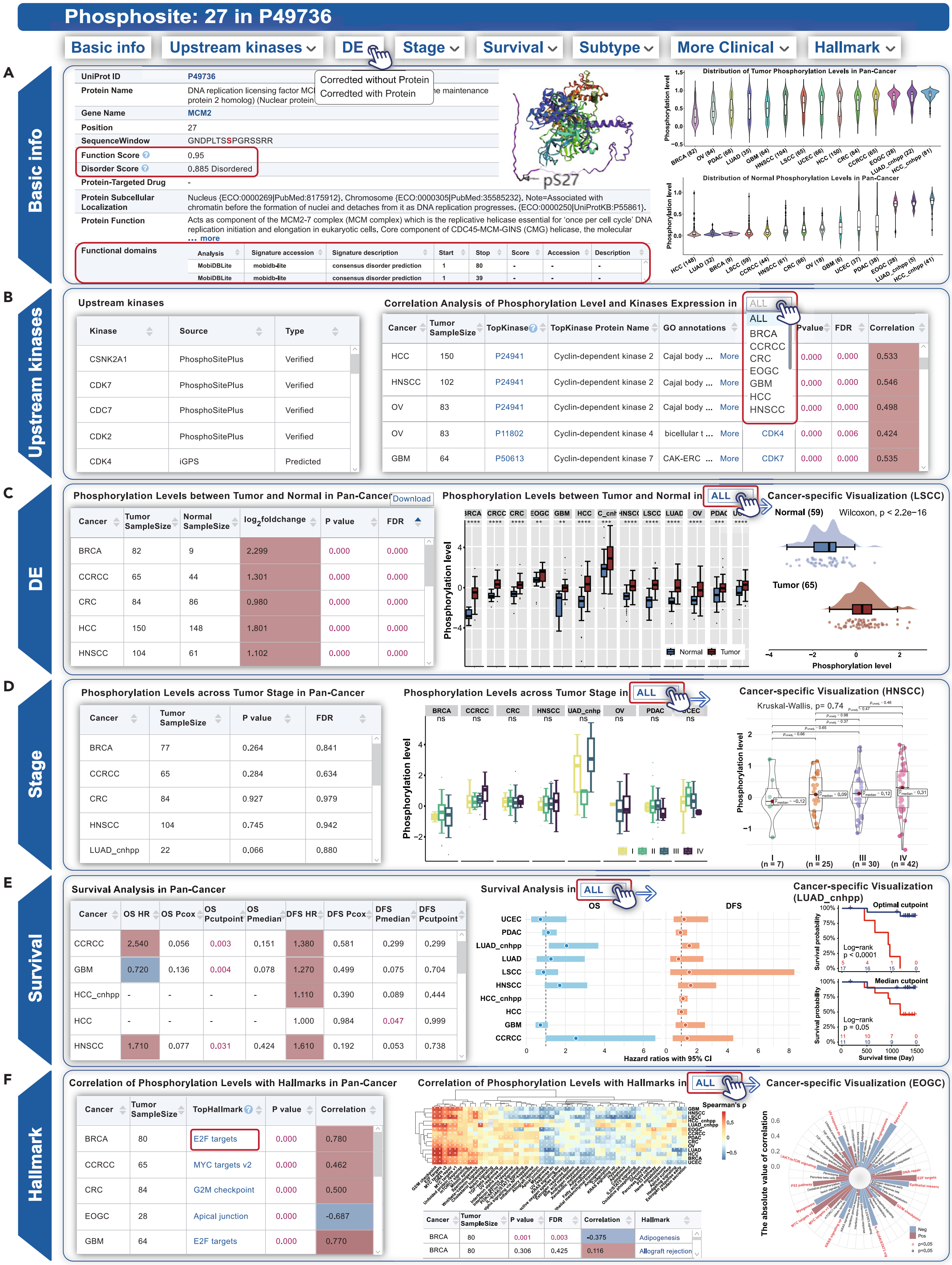
Figure 3. The landscape of cancer-associated phosphosites

(A) Distribution of phosphosites associated with varying numbers of cancer types.

(B) Network of differential phosphosites shared among at least seven cancer types. Edge colors indicate changes in phosphorylation levels: red denotes elevated phosphorylation in cancer tissues, while blue denotes elevated phosphorylation in normal tissues. Node colors represent the consistency of these differences across cancer types: red nodes indicate consistently elevated phosphorylation in cancer tissues, blue nodes indicate consistently decreased phosphorylation in cancer tissues, and yellow nodes indicate inconsistent phosphorylation patterns across different cancer types.

samples, is presented at the bottom. Phosphorylation levels at this site are the highest in the HCC_cnhpp cohort across both cancerous and normal tissues, whereas BRCA displays the lowest phosphorylation levels. Protein phosphorylation is a result of kinase regulation and kinases could serve as promising therapeutic targets in cancer.[5] It is thus particularly important to elucidate the regulatory relationship among kinases and phosphosites. The "Upstream Kinases" module includes two result tables. The first table lists verified upstream kinases from phosphoELM and PhosphoSitePlus, as well as predicted kinases from GPS, iGPS (high thresholds), and the study by Johnson et al.[35] The second table provides detailed statistical outcomes for the top 10 kinases from Spearman correlation analysis, including cancer type, sample size with non-missing values, kinases, GO annotations for kinases, *p* value, false discovery rate (FDR) and correlation coefficient. Several upstream kinases for MCM2 pS27 have already been verified previously, including CDK2, CDK7, CDC7, and CSNK2A1. Notably, CDK4, identified by iGPS and The Kinase Library,[35] ranks among the top 10 kinases with the highest correlation in the OV cohort (r = 0.424, FDR = 0.006), indicating its potential role as an upstream kinase for this phosphosite (Figure 4B). The "DE", "Stage", "Survival", "Subtype", "More Clinical" and "Hallmark" modules follow a similar structure: each module begins with a detailed statistical table at the top of the page, followed by visualizations at the bottom. The visualizations include a summarizing plot covering 14 cohorts and cancer-specific views are accessible via a dropdown menu. MCM2 pS27 consistently exhibits significantly elevated phosphorylation levels in cancerous tissues compared to normal tissues across 13 cohorts, with the highest fold change in BRCA. We also report the distribution of MCM2 pS27 phosphorylation across various tumor stages. However, the significance of *p* values did not reach statistical significance, potentially due to limited sample size (Figure 4D). Survival analysis utilizing a log rank test with optimal cutoff indicates statistical significance within the majority of cohorts (Figure 4E). MCM2, a core subunit of eukaryotic helicase, plays a vital role in DNA replication.[34] Indeed, the association analyses with hallmarks consistently confirm its functional relevance, revealing strong correlations between its phosphorylation level and hallmark pathways associated with proliferation across all cohorts, including E2F targets, G2M checkpoint, and MYC targets v2 (Figure 4F). Hyperlinks to UniProtKB, GeneCards, and the Molecular Signatures Database (MSigDB)[36] are provided for the UniProt ID, gene name, and hallmark, respectively, thereby facilitating users' access to additional supplemental information.

Finally, users can download generated plots in raster (PNG) formats, processed phosphoproteomics data, and detailed tables containing statistical analysis results as text files from the "Download" page. A detailed user guide is available on the "Help" page to assist users in navigating and utilizing the features of PhosCancer effectively.



UniProt ID	P49736
Protein Name	DNA replication licensing factor MCM2 … maintenance protein 2 homolog) (Nuclear protein …
Gene Name	MCM2
Position	27
SequenceWindow	GNDPLTSSPGRSSRR
Function Score	0.95
Disorder Score	0.885 Disordered
Protein-Targeted Drug	-
Protein Subcellular Localization	Nucleus {ECO:0000269|PubMed:8175912}. Chromosome {ECO:0000305|PubMed:35585232}. Note=Associated with chromatin before the formation of nuclei and detaches from it as DNA replication progresses. {ECO:0000250|UniProtKB:P55861}.
Protein Function	Acts as component of the MCM2-7 complex (MCM complex) which is the replicative helicase essential for 'once per cell cycle' DNA replication initiation and elongation in eukaryotic cells. Core component of CDC45-MCM-GINS (CMG) helicase, the molecular … more

Functional domains

Analysis	Signature accession	Signature description	Start	Stop	Score	Accession	Description
MobiDBLite	mobidb-lite	consensus disorder prediction	1	80	-	-	-
MobiDBLite	mobidb-lite	consensus disorder prediction	1	39	-	-	-

Upstream kinases

Kinase	Source	Type
CSNK2A1	PhosphoSitePlus	Verified
CDK7	PhosphoSitePlus	Verified
CDC7	PhosphoSitePlus	Verified
CDK2	PhosphoSitePlus	Verified
CDK4	iGPS	Predicted

Correlation Analysis of Phosphorylation Level and Kinases Expression in ALL

Cancer	Tumor SampleSize	TopKinase	TopKinase Protein Name	GO annotations		Pvalue	FDR	Correlation
HCC	150	P24941	Cyclin-dependent kinase 2	Cajal body … More		0.000	0.000	0.533
HNSCC	102	P24941	Cyclin-dependent kinase 2	Cajal body … More		0.000	0.000	0.546
OV	83	P24941	Cyclin-dependent kinase 2	Cajal body … More		0.000	0.000	0.498
OV	83	P11802	Cyclin-dependent kinase 4	bicellular t … More	CDK4	0.000	0.006	0.424
GBM	64	P50613	Cyclin-dependent kinase 7	CAK-ERC … More	CDK7	0.000	0.000	0.535

Phosphorylation Levels between Tumor and Normal in Pan-Cancer

Cancer	Tumor SampleSize	Normal SampleSize	log_2foldchange	P value	FDR
BRCA	82	9	2.299	0.000	0.000
CCRCC	65	44	1.301	0.000	0.000
CRC	84	86	0.980	0.000	0.000
HCC	150	148	1.801	0.000	0.000
HNSCC	104	61	1.102	0.000	0.000

Phosphorylation Levels across Tumor Stage in Pan-Cancer

Cancer	Tumor SampleSize	P value	FDR
BRCA	77	0.264	0.841
CCRCC	65	0.284	0.634
CRC	84	0.927	0.979
HNSCC	104	0.745	0.942
LUAD_cnhpp	22	0.066	0.880

Survival Analysis in Pan-Cancer

Cancer	OS HR	OS Pcox	OS Pcutpoint	OS Pmedian	DFS HR	DFS Pcox	DFS Pmedian	DFS Pcutpoint
CCRCC	2.540	0.056	0.003	0.151	1.380	0.581	0.299	0.299
GBM	0.720	0.136	0.004	0.078	1.270	0.499	0.075	0.704
HCC_cnhpp	-	-	-	-	1.110	0.390	0.089	0.444
HCC	-	-	-	-	1.000	0.984	0.047	0.999
HNSCC	1.710	0.077	0.031	0.424	1.610	0.192	0.053	0.738

Correlation of Phosphorylation Levels with Hallmarks in Pan-Cancer

Cancer	Tumor SampleSize	TopHallmark	P value	Correlation
BRCA	80	E2F targets	0.000	0.780
CCRCC	65	MYC targets v2	0.000	0.462
CRC	84	G2M checkpoint	0.000	0.500
EOGC	28	Apical junction	0.000	-0.687
GBM	64	E2F targets	0.000	0.770

Cancer	Tumor SampleSize	P value	FDR	Correlation	Hallmark
BRCA	80	0.001	0.003	-0.375	Adipogenesis
BRCA	80	0.306	0.425	0.116	Allograft rejection

Figure 4. A case study to explore phosphosite S27 on MCM2 using PhosCancer

Clicking the "Basic info", "Upstream kinase", "DE", "Stage", "Survival", "Subtype", "More Clinical" and "Hallmark" buttons in the toolbar displays the corresponding results.

(A) This section presents basic information, functional and disorder prediction scores of the phosphosite, along with 3Dmol visualization of the phosphosite on the tertiary structure. It also includes the landscape of phosphorylation levels at this phosphosite across 14 datasets in tumor or normal samples.

(B) The first table comprises experimental and predicted upstream kinases and the second table provides the correlation analysis between kinases and the phosphosite based on the 14 datasets. The DE (C), Stage (D), Survival (E), and Hallmark (F) sections each comprise three main components. Firstly, a detailed statistical outcome table for each analysis. Secondly, a pan-cancer visualization. Thirdly, a cancer-specific visualization is accessible via a dropdown menu.

DISCUSSION

The crucial role of protein phosphorylation in cancer leads to an urgent need for integrating associations between phosphorylation events and cancer biology. Benefiting from the extensive proteomics data deposited by cancer proteomics consortia, we assembled a comprehensive human phosphoproteome. Here, we present PhosCancer, a database aimed to translate phosphopeptide identification and quantification results into biological and clinical insights from a vast repository of phosphoproteomics data. PhosCancer allows cancer researchers and clinicians, regardless of their computational expertise, to access, analyze, visualize, and interpret the data with ease. Considering kinases as promising therapeutic targets for cancer,[5] PhosCancer not only provides experimental verified and predicted upstream kinases collected from external databases and tools, but also offers co-expression links between 518 protein kinases and the phosphosites across our 14 datasets to enhance the identification of potential kinase-substrate interactions. To prioritize phosphosites for further study regarding their potential oncogenic or tumor suppressor properties, one common strategy is to identify those associated with clinical features. PhosCancer comprehensively characterizes phosphosites with clinical features, including phosphorylation level differences across different cancer subsets as defined by clinicopathologic features, survival associations and hallmarks. To meet diverse user needs, survival analysis employs multiple methods such as Cox proportional hazards regression and log rank tests based on optimal cutoff or median value. Each analysis provides both cancer-specific results and pan-cancer visualizations, enabling users to explore the role of each phosphosite across different cancers simultaneously. To correct for differential underlying protein levels when performing certain phosphorylation-based analyses, a linear model was then fit (R function lm, PTM $\sim$ protein) to all matched PTM-protein data points by accession number. The residuals of each model served as protein-normalized phosphosite abundances.[3,21] Both the search page and each analysis module on the detailed page include a dropdown menu labeled "Corrected with Protein," indicating that the statistical analysis is based on phosphorylation data adjusted for protein expression. Moreover, critical annotations from external databases, such as whether the phosphosite falls within known protein functional regions, are integrated for comprehensive data interpretation. In PhosCancer, a flexible and user-friendly web interface offers an integrated presentation encompassing annotation, statistical analysis, and meticulously crafted images for visualization. The information about MCM2 pS27 underscores the database's ability to provide credible clues and functional insights. The information available in this database can be used by biology researchers to explore potential functional associations of phosphosites. Furthermore, PhosCancer enables rapid prioritization of critical phosphosites for in-depth investigation, resulting in significant time and effort savings, and aiding in the discovery of clinically relevant biomarkers.

PhosCancer will undergo quarterly maintenance and updates through ongoing monitoring of public resources and research articles. Additionally, we will integrate additional annotations from other public resources to rich annotations for phosphoproteins and phosphosites. We anticipate PhosCancer to serve as a valuable asset for future cancer proteomics research, significantly advancing our understanding of the molecular mechanisms underlying human cancer.

Limitations of the study

PhosCancer, as primarily a data exploration and hypothesis generation tool, has several limitations. For instance, while certain cancer-associated phosphosites in the database show promise as diagnostic and therapeutic targets, their translation into clinical applications remains to be experimentally validated. Moreover, although PhosCancer contains datasets from CNHPP and CPTAC, data from other sources necessitates further supplementation in PhosCancer.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jing Li (jing.li@sjtu.edu.cn).

Materials availability

This study did not generate new reagents.

Data and code availability

- All data used in this study were obtained from CPTAC and CNHPP. All generated data and analysis results of PhosCancer are accessible through the download page of PhosCancer (https://lilab.life.sjtu.edu.cn/PhosCancer/Download.html).
- Code for PhosCancer is available at https://github.com/Li-Lab-SJTU/PhosCancer.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the Key Project for Computational Biology of Shanghai (23JS1400800), the National Natural Science Foundation of China (32170664 and 31871329) and the Fundamental Research Funds for the Central Universities (YG2022QN065 and YG2023ZD11). The computations in this paper were run on the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

AUTHOR CONTRIBUTIONS

Q.D.: conceptualization, methodology, formal analysis, investigation, writing—original draft. D.-Q.S.: data curation, writing—review and editing. J.-C.Y.: software. J.-X.C.: software, writing—review and editing. J.L.: conceptualization, supervision, writing-review and editing, project administration, funding acquisition.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - ○ Data collection and processing
  - ○ Upstream kinase collection and prediction
  - ○ Associations with clinical features
  - ○ Relevance with hallmarks
  - ○ Integration of external biological resources
  - ○ Database implementation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111060.

Received: June 17, 2024
Revised: August 3, 2024
Accepted: September 24, 2024
Published: September 27, 2024

REFERENCES

1. Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int. J. Mol. Med. *40*, 271–280. https://doi.org/10.3892/ijmm.2017.3036.
2. Singh, V., Ram, M., Kumar, R., Prasad, R., Roy, B.K., and Singh, K.K. (2017). Phosphorylation: Implications in Cancer. Protein J. *36*, 1–6. https://doi.org/10.1007/s10930-017-9696-z.
3. Geffen, Y., Anand, S., Akiyama, Y., Yaron, T.M., Song, Y., Johnson, J.L., Govindan, A., Babur, Ö., Li, Y., Huntsman, E., et al. (2023). Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. Cell *186*, 3945–3967.e26. https://doi.org/10.1016/j.cell.2023.07.013.
4. Zhu, C., Rogers, A., Asleh, K., Won, J., Gao, D., Leung, S., Li, S., Vij, K.R., Zhu, J., Held, J.M., et al. (2020). Phospho-Ser(784)-VCP Is Required for DNA Damage Response and Is Associated with Poor Prognosis of Chemotherapy-Treated Breast Cancer. Cell Rep. *31*, 107745. https://doi.org/10.1016/j.celrep.2020.107745.
5. Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell *177*, 1035–1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.
6. Li, Z., Li, S., Luo, M., Jhong, J.H., Li, W., Yao, L., Pang, Y., Wang, Z., Wang, R., Ma, R., et al. (2022). dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res. *50*, D471–D479. https://doi.org/10.1093/nar/gkab1017.
7. Lin, S., Wang, C., Zhou, J., Shi, Y., Ruan, C., Tu, Y., Yao, L., Peng, D., and Xue, Y. (2021). EPSD: a well-annotated data resource of protein phosphorylation sites in eukaryotes. Brief. Bioinform. *22*, 298–307. https://doi.org/10.1093/bib/bbz169.
8. Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. (2011). Phospho.ELM: a database of phosphorylation sites–update 2011. Nucleic Acids Res. *39*, D261–D267. https://doi.org/10.1093/nar/gkq1104.
9. Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40*, D261–D270. https://doi.org/10.1093/nar/gkr1122.
10. Hu, G.S., Zheng, Z.Z., He, Y.H., Wang, D.C., and Liu, W. (2023). CPPA: A Web Tool for Exploring Proteomic and Phosphoproteomic Data in Cancer. J. Proteome Res. *22*, 368–373. https://doi.org/10.1021/acs.jproteome.2c00512.
11. Zou, J., Qin, Z., Li, R., Yan, X., Huang, H., Yang, B., Zhou, F., and Zhang, L. (2024). iProPhos: A Web-Based Interactive Platform for Integrated Proteome and Phosphoproteome Analysis. Mol. Cell. Proteomics *23*, 100693. https://doi.org/10.1016/j.mcpro.2023.100693.
12. Wang, D., Qian, X., Du, Y.C.N., Sanchez-Solana, B., Chen, K., Kanigicherla, M., Jenkins, L.M., Luo, J., Eng, S., Park, B., et al. (2023). cProSite: A web based interactive platform for online proteomics, phosphoproteomics, and genomics data analysis. J. Biotechnol. Biomed. *6*, 573–578. https://doi.org/10.26502/jbb.2642-91280119.
13. Chandrashekar, D.S., Karthikeyan, S.K., Korla, P.K., Patel, H., Shovon, A.R., Athar, M., Netto, G.J., Qin, Z.S., Kumar, S., Manne, U., et al. (2022). UALCAN: An update to the integrated cancer data analysis platform. Neoplasia *25*, 18–27. https://doi.org/10.1016/j.neo.2022.01.001.
14. Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934. https://doi.org/10.1126/science.1075762.
15. Krug, K., Jaehnig, E.J., Satpathy, S., Blumenberg, L., Karpova, A., Anurag, M., Miles, G., Mertins, P., Geffen, Y., Tang, L.C., et al. (2020). Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted

Therapy. Cell *183*, 1436–1456.e31. https://doi.org/10.1016/j.cell.2020.10.036.

16. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.S.M., Chang, H.Y., et al. (2019). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell *179*, 964–983.e31. https://doi.org/10.1016/j.cell.2019.10.007.
17. Mun, D.G., Bhin, J., Kim, S., Kim, H., Jung, J.H., Jung, Y., Jang, Y.E., Park, J.M., Kim, H., Jung, Y., et al. (2019). Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell *35*, 111–124.e10. https://doi.org/10.1016/j.ccell.2018.12.003.
18. Wang, L.B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., et al. (2021). Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell *39*, 509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.
19. Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., et al. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell *179*, 561–577.e22. https://doi.org/10.1016/j.cell.2019.08.052.
20. Huang, C., Chen, L., Savage, S.R., Eguez, R.V., Dou, Y., Li, Y., da Veiga Leprevost, F., Jaehnig, E.J., Lei, J.T., Wen, B., et al. (2021). Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer Cell *39*, 361–379.e16. https://doi.org/10.1016/j.ccell.2020.12.007.
21. Satpathy, S., Krug, K., Jean Beltran, P.M., Savage, S.R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M.H., Avanessian, S.C., et al. (2021). A proteogenomic portrait of lung squamous cell carcinoma. Cell *184*, 4348–4371.e40. https://doi.org/10.1016/j.cell.2021.07.016.
22. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell *182*, 200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.
23. McDermott, J.E., Arshad, O.A., Petyuk, V.A., Fu, Y., Gritsenko, M.A., Clauss, T.R., Moore, R.J., Schepmoes, A.A., Zhao, R., Monroe, M.E., et al. (2020). Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability. Cell Rep. Med. *1*, 100004. https://doi.org/10.1016/j.xcrm.2020.100004.
24. Cao, L., Huang, C., Cui Zhou, D., Hu, Y., Lih, T.M., Savage, S.R., Krug, K., Clark, D.J., Schnaubelt, M., Chen, L., et al. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell *184*, 5031–5052.e26. https://doi.org/10.1016/j.cell.2021.08.023.
25. Dou, Y., Kawaler, E.A., Cui Zhou, D., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Satpathy, S., et al. (2020). Proteogenomic Characterization of Endometrial Carcinoma. Cell *180*, 729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.
26. Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., Xing, B., Sun, W., Ren, L., Hu, B., et al. (2019). Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature *567*, 257–261. https://doi.org/10.1038/s41586-019-0987-8.
27. Xu, J.Y., Zhang, C., Wang, X., Zhai, L., Ma, Y., Mao, Y., Qian, K., Sun, C., Liu, Z., Jiang, S., et al. (2020). Integrative Proteomic Characterization of Human Lung Adenocarcinoma. Cell *182*, 245–261.e17. https://doi.org/10.1016/j.cell.2020.05.043.
28. Yu, K., Zhang, Q., Liu, Z., Zhao, Q., Zhang, X., Wang, Y., Wang, Z.X., Jin, Y., Li, X., Liu, Z.X., and Xu, R.H. (2019). qPhos: a database of protein phosphorylation dynamics in humans. Nucleic Acids Res. *47*, D451–D458. https://doi.org/10.1093/nar/gky1052.
29. Hume, S., Grou, C.P., Lascaux, P., D'Angiolella, V., Legrand, A.J., Ramadan, K., and Dianov, G.L. (2021). The NUCKS1-SKP2-p21/p27 axis controls S phase entry. Nat. Commun. *12*, 6959. https://doi.org/10.1038/s41467-021-27124-8.
30. Chikamori, K., Grabowski, D.R., Kinter, M., Willard, B.B., Yadav, S., Aebersold, R.H., Bukowski, R.M., Hickson, I.D., Andersen, A.H., Ganapathi, R., and Ganapathi, M.K. (2003). Phosphorylation of serine 1106 in the catalytic domain of topoisomerase II alpha regulates enzymatic activity and drug sensitivity. J. Biol. Chem. *278*, 12696–12702. https://doi.org/10.1074/jbc.M300837200.
31. Shevah-Sitry, D., Miniowitz-Shemtov, S., Teichner, A., Kaisari, S., and Hershko, A. (2022). Role of phosphorylation of Cdc20 in the regulation of the action of APC/C in mitosis. Proc. Natl. Acad. Sci. USA *119*, e2210367119. https://doi.org/10.1073/pnas.2210367119.
32. Chen, Y., Tian, P., and Liu, Y. (2017). P53 and Protein Phosphorylation Regulate the Oncogenic Role of Epithelial Cell Transforming 2 (ECT2). Med. Sci. Monit. *23*, 3154–3160. https://doi.org/10.12659/msm.905388.
33. Sun, Y., Cheng, Z., and Liu, S. (2022). MCM2 in human cancer: functions, mechanisms, and clinical significance. Mol. Med. *28*, 128. https://doi.org/10.1186/s10020-022-00555-9.
34. Fei, L., and Xu, H. (2018). Role of MCM2-7 protein phosphorylation in human cancer cells. Cell Biosci. *8*, 43. https://doi.org/10.1186/s13578-018-0242-2.
35. Johnson, J.L., Yaron, T.M., Huntsman, E.M., Kerelsky, A., Song, J., Regev, A., Lin, T.Y., Liberatore, K., Cizin, D.M., Cohen, B.M., et al. (2023). An atlas of substrate specificities for the human serine/threonine kinome. Nature *613*, 759–766. https://doi.org/10.1038/s41586-022-05575-3.
36. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.
37. Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., et al. (2012). A cross-platform toolkit for mass spectrometry and proteomics. Nat. Biotechnol. *30*, 918–920. https://doi.org/10.1038/nbt.2377.
38. Kim, S., and Pevzner, P.A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. *5*, 5277. https://doi.org/10.1038/ncomms6277.
39. Monroe, M.E., Shaw, J.L., Daly, D.S., Adkins, J.N., and Smith, R.D. (2008). MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. Comput. Biol. Chem. *32*, 215–217. https://doi.org/10.1016/j.compbiolchem.2008.02.006.
40. Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol. *24*, 1285–1292. https://doi.org/10.1038/nbt1240.
41. Wang, C., Xu, H., Lin, S., Deng, W., Zhou, J., Zhang, Y., Shi, Y., Peng, D., and Xue, Y. (2020). GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. Dev. Reprod. Biol. *18*, 72–80. https://doi.org/10.1016/j.gpb.2020.01.001.
42. Song, C., Ye, M., Liu, Z., Cheng, H., Jiang, X., Han, G., Songyang, Z., Tan, Y., Wang, H., Ren, J., et al. (2012). Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Mol. Cell. Proteomics *11*, 1070–1083. https://doi.org/10.1074/mcp.M111.012625.
43. Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. *14*, 7. https://doi.org/10.1186/1471-2105-14-7.
44. Erdos, G., Pajkos, M., and Dosztanyi, Z. (2021). IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. *49*, W297–W303. https://doi.org/10.1093/nar/gkab408.
45. Ochoa, D., Jarnuczak, A.F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A.A., Hill, A., Garcia-Alonso, L., Stein, F., et al. (2020). The functional landscape of the human phosphoproteome. Nat. Biotechnol. *38*, 365–373. https://doi.org/10.1038/s41587-019-0344-3.
46. Rego, N., and Koes, D. (2015). 3Dmol.js: molecular visualization with WebGL. Bioinformatics *31*, 1322–1324. https://doi.org/10.1093/bioinformatics/btu829.
47. Dinalankara, W., Ke, Q., Xu, Y., Ji, L., Pagane, N., Lien, A., Matam, T., Fertig, E.J., Price, N.D., Younes, L., et al. (2018). Digitizing omics profiles by divergence from a baseline. Proc. Natl. Acad. Sci. USA *115*, 4545–4552. https://doi.org/10.1073/pnas.1721628115.
48. Angel, P.W., Rajab, N., Deng, Y., Pacheco, C.M., Chen, T., Lê Cao, K.A., Choi, J., and Wells, C.A. (2020). A simple, scalable approach to building a cross-platform transcriptome atlas. PLoS Comput. Biol. *16*, e1008219. https://doi.org/10.1371/journal.pcbi.1008219.
49. McClure, M.B., Kogure, Y., Ansari-Pour, N., Saito, Y., Chao, H.H., Shepherd, J., Tabata, M., Olopade, O.I., Wedge, D.C., Hoadley, K.A., et al. (2023). Landscape of Genetic Alterations Underlying Hallmark Signature Changes in Cancer Reveals TP53

Aneuploidy-driven Metabolic Reprogramming. Cancer Res. Commun. *3*, 281–296. https://doi.org/10.1158/2767-9764.CRC-22-0073.

50. Freshour, S.L., Kiwala, S., Cotto, K.C., Coffman, A.C., McMichael, J.F., Song, J.J., Griffith, M., Griffith, O.L., and Wagner, A.H. (2021). Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. *49*, D1144–D1151. https://doi.org/10.1093/nar/gkaa1084.

51. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

52. Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., Qiu, Y., and Chen, Y. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res. *50*, D1398–D1407. https://doi.org/10.1093/nar/gkab953.

53. Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B.L., Salazar, G.A., Bileschi, M.L., Bork, P., Bridge, A., Colwell, L., et al. (2023). InterPro in 2022. Nucleic Acids Res. *51*, D418–D427. https://doi.org/10.1093/nar/gkac993.

54. Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., Zu, M., and Chen, W. (2018). ECharts: a declarative framework for rapid construction of web-based visualization. Visual Informatics *2*, 136–146.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
msConvert	Chambers et al.[37]	https://proteowizard.sourceforge.io/tools/msconvert.html
MS-GF+	Kim et al.[38]	https://github.com/MSGFPlus/msgfplus
MASIC	Monroe et al.[39]	https://github.com/PNNL-Comp-Mass-Spec/MASIC
Ascore	Beausoleil et al.[40]	https://github.com/PNNL-Comp-Mass-Spec/AScore
GPS 5.0	Wang et al.[41]	http://gps.biocuckoo.cn
iGPS	Song et al.[42]	http://igps.biocuckoo.org
R 4.0.2	R Core Team	https://www.r-project.org
impute	https://github.com/gangwug/impute	https://github.com/gangwug/impute
GSVA	Hänzelmann et al.[43]	https://bioconductor.org/packages/release/bioc/html/GSVA.html
IUPred3	Erdős et al.[44]	https://iupred3.elte.hu
InterProScan	https://github.com/ebi-pf-team/interproscan	https://github.com/ebi-pf-team/interproscan
funscoR	Ochoa et al.[45]	https://github.com/evocellnet/funscoR
HTML	https://developer.mozilla.org/zh-CN/docs/Web/HTML	https://developer.mozilla.org/zh-CN/docs/Web/HTML
CSS	https://developer.mozilla.org/zh-CN/docs/Web/CSS	https://developer.mozilla.org/zh-CN/docs/Web/CSS
JavaScript	https://developer.mozilla.org/zh-CN/docs/Web/JavaScript	https://developer.mozilla.org/zh-CN/docs/Web/JavaScript
MySQL	https://www.mysql.com	https://www.mysql.com
Bootstrap	https://v3.bootcss.com	https://v3.bootcss.com
Echarts	https://echarts.apache.org/zh/index.html	https://echarts.apache.org/zh/index.html
3Dmol	Rego et al.[46]	http://3dmol.csb.pitt.edu
Deposited data		
PhosCancer: Analysis scripts	This paper	https://github.com/Li-Lab-SJTU/PhosCancer
PhosCancer: Standardized Data for PhosCancer	This paper	https://lilab.life.sjtu.edu.cn/PhosCancer/Download.html
Other		
Resource website for the publication	This paper	https://lilab.life.sjtu.edu.cn/PhosCancer

METHOD DETAILS

Data collection and processing

We curated raw phosphoproteomics and proteomics data from nearly all available cancer types from the CNHPP and CPTAC data portals, encompassing 14 different cohorts with sample sizes exceeding 50 (Tables S1 and S2). These datasets include BRCA,[15] CCRCC,[16] CRC,[5] GBM,[18] HCC,[19] HNSCC,[20] LSCC,[21] LUAD,[22] OV,[23] PDAC[24] and UCEC[25] in CPTAC with TMT labeling, EOGC[17] in CPTAC with iTRAQ labeling, HCC_cnhpp[26] and LUAD_cnhpp[27] in CNHPP with label-free quantification. Clinical information was also downloaded with molecular data.

MS/MS spectra files obtained from the CNHPP in raw format were uniformly converted to mzML format along with data from CPTAC using the msConvert tool[37] (version 3.0.21128). Then, the MS/MS data in mzML format were searched against the UniProtKB/Swiss-Prot protein sequence database downloaded in January 2024 using the MS-GF+ search engine.[38] Specifically, the partially tryptic search used a ± 20 ppm parent ion tolerance, and the maximum missing cleavage site was set as 2. MS-GF+ considered static carbamidomethylation

(+57.0215 Da) on Cys residues, and dynamic oxidation (+15.9949 Da) on Met residues. Additional static modifications were set according to specific datasets, such as TMT modification (+229.1629 Da) on peptide N termini and Lys residue for TMT datasets, and iTRAQ4 modification (+144.1021 Da) on peptide N termini and Lys residue for iTRAQ4 datasets. All results were then filtered with a 1% false discovery rate (FDR) at the peptide-spectrum match (PSMs) level. Intensities of all TMT reporter ions and peak areas for the label-free dataset were extracted using MASIC software.[39] Phosphoproteomics data files underwent the same process with additional consideration for dynamic phosphorylation on Ser, Thr, or Tyr residues. Phosphosite localization was performed using the Ascore algorithm, and top-scoring sequences were reported.[40] Phosphorylation levels were determined based on the median intensity of all identified peptides at each specific site.[5] Subsequently, median-centered normalization and log2 transformation were applied to obtain final abundance values. Considering the variability introduced by different MS/MS platforms and experimental protocols, we employed percentile rank transformation to standardize the expression values of each MS/MS dataset to a uniform scale (0-1) when exploring the distribution of tumor/normal phosphorylation levels across various cancer types.[47],[48]

Tumor stage information in clinical data was standardized by consolidating subcategories, such as ‘stage IA’, ‘stage IB’, and ‘stage IC’, into a unified stage I group. Survival data included patient vital status (Dead/Alive), ‘Days to Last Follow Up’ for living patients, and ‘Days to Death’ for deceased patients.

Upstream kinase collection and prediction

We gathered experimentally verified human kinase-specific phosphosite relationships from the Phospho.ELM[8] and PhosphoSitePlus[9] (accessed in Apr. 2024). We employed the sequence-based predictor GPS 5.0[41] and network-based predictor iGPS[42] to predict kinase-substrate interactions with a high threshold. Furthermore, kinases ranking within the top 10 for the Ser/Thr kinases in The Kinase Library were considered as biochemically predicted according to Johnson et al.,[35] which predicts PTM regulators according to substrate specificity.

In addition, Spearman correlation analysis was conducted to explore the co-expression between 518 protein kinases[14] and the phosphosites across our 14 datasets. PhosCancer presents the detailed statistical results for the top 10 kinases exhibiting the highest correlation.

Associations with clinical features

The association analyses between phosphopeptide abundances and clinical features involved differential expression assessment between tumor and normal samples, differential expression across tumor stages, survival analyses and associations with age, gender, BMI, race, tumor size and subtypes. The subtype information comes from the data provided in the clinical documents, corresponding to the multi-omics classification results in the relevant literature. The Wilcoxon rank-sum test was employed to examine differential expression between tumor and normal samples, younger and older samples (separated by median values), and female and male samples. The Kruskal-Wallis test was utilized to evaluate differential expression across various tumor stages, races and tumor subtypes. Spearman correlation analysis was used to assess the correlation between phosphorylation and BMI/tumor size. For investigating associations with overall survival (OS) or disease-free survival (DFS), Kaplan-Meier survival analysis, log-rank test, and Cox proportional hazards regression were conducted. Using both Cox regression and log-rank tests provides a comprehensive understanding of survival differences and allows for cross-validation, enhancing the robustness of our conclusions. Notably, two cutoff methods based on median and optimal cutpoints were applied for the log-rank test. The median value is a commonly used statistical measure that divides the data into two equal parts, providing a straightforward and unbiased dichotomization for survival analysis. The optimal cutpoint, determined using the survminer R package (version 0.4.9) with the maxstat method, maximizes the difference in survival between groups. This method helps identify more potentially prognostic phosphosites, thereby improving the sensitivity of the analysis. FDR was controlled using the Benjamini-Hochberg method to address multiple comparisons.

To ensure statistical reliability, phosphosites with missing values exceeding 80% in samples were excluded.

Relevance with hallmarks

We explored the potential effects of phosphosites on hallmarks in cancer by conducting a Spearman correlation analysis between phosphorylation levels and the activities of hallmark pathways (h.all.v7.5.1.symbols.gmt). The 50 hallmark pathways, retrieved from the MSigDB[36], have been extensively employed to elucidate the fundamental mechanisms underlying cancer pathogenesis.[49] Pathway activities were inferred using single-sample gene set enrichment analysis (ssGSEA) applied to proteomics data with the R package GSVA.[43] Given the intolerance of the tool to missing data, we performed missing value imputation before analysis. Proteomics data underwent filtration to exclude proteins with more than 50% missing values in tumors, and any remaining missing values were imputed using k-nearest neighbor (kNN) imputation implemented in the impute R-package, with 5 nearest neighbors utilized.[22]

Integration of external biological resources

To provide more functional and structural annotations of the phosphosites and modified proteins, we integrated external biological databases and tools. For each phosphosite, the basic information about its protein has been obtained from UniProtKB, such as UniProt ID, protein name, gene name, sequence window (a peptide with 7 amino acids upstream and downstream of the modified residue), protein subcellular localization and protein function. Besides, we assigned a functional score to each phosphosite ranging from 0 to 1 using the funscoR R package, with higher scores indicating increased relevance for cell fitness.[45] To assess whether phosphosites reside in intrinsically disordered protein regions (IDRs), site-specific disorder scores were computed using IUPred3,[44] with scores exceeding 0.5 classified as disorder. We also investigated drugs targeting proteins where phosphosites are located, utilizing data from the Drug-Gene Interaction database (DGIdb) 4.0,[50] DrugBank,[51] and Therapeutic Target Database Therapeutic Target Database (TTD)[52] (downloaded in Apr. 2024). Lastly, to elucidate the functional roles of phosphosites, their presence within functional domains annotated in the InterPro database[53] was examined. InterProScan was employed with default settings to search protein sequences for potential matches against 13 InterPro member protein signature databases.

Database implementation

The PhosCancer database was implemented on an Apache HTTP server running on the Linux platform. The front-end of the interface was developed using HTML, PHP, CSS, and JavaScript, with Bootstrap with jQuery for responsiveness and enhanced user experience. MySQL server version 5.1.73 on a Linux-based system was utilized for back-end data management. Visualization of PDB tertiary structures utilized the 3Dmol.js plugin,[46] while interactive charts on the homepage were created using the ECharts plugin.[54] Besides, all figures displayed on the website were generated using R.

QUANTIFICATION AND STATISTICAL ANALYSIS

Protein expression and phosphorylation levels were quantified using MASIC software (v3.2.7465). Kinase-substrate interactions were predicted with the GPS 5.0 and iGPS 1.0. Disorder scores were computed with IUPred3, and functional scores for each phosphosite were assigned using the funscoR R package (v0.1.0). InterProScan (v5.61-93.0) was employed to match protein sequences against protein families and domains. Statistical analyses included Spearman correlation to evaluate relationships between phosphorylation and BMI/tumor size, hallmark activity, and protein kinases. Differential expression was assessed using the Wilcoxon rank-sum test for comparisons of tumor versus normal samples, age (younger versus older), and sex (female versus male). The Kruskal-Wallis test evaluated differential expression across tumor stages, races, and subtypes. The log-rank test and Cox proportional hazard regression were performed to investigate associations with survival. Specifically, '*': $p < 0.05$; '**': $p < 0.01$; '***': $p < 0.001$; '****': $p < 0.0001$; ns: not significant.