Published online 8 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D901–D910***
doi: 10.1093/nar/gkx973

ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins

Michal Krassowski[1,2], Marta Paczkowska[1], Kim Cullion[1], Tina Huang[1], Irakli Dzneladze[1,3], B. F. Francis Ouellette[1,4], Joseph T. Yamada[1], Amelie Fradet-Turcotte[5] and Jüri Reimand[1,3,*]

[1]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada, [2]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland, [3]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada, [4]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada and [5]Department of Molecular Biology, Medical Biochemistry and Pathology, Université Laval, Québec, Québec, Canada

Received August 15, 2017; Revised September 29, 2017; Editorial Decision October 08, 2017; Accepted October 18, 2017

ABSTRACT

Interpretation of genetic variation is needed for deciphering genotype-phenotype associations, mechanisms of inherited disease, and cancer driver mutations. Millions of single nucleotide variants (SNVs) in human genomes are known and thousands are associated with disease. An estimated 21% of disease-associated amino acid substitutions corresponding to missense SNVs are located in protein sites of post-translational modifications (PTMs), chemical modifications of amino acids that extend protein function. ActiveDriverDB is a comprehensive human proteo-genomics database that annotates disease mutations and population variants through the lens of PTMs. We integrated >385,000 published PTM sites with ∼3.6 million substitutions from The Cancer Genome Atlas (TCGA), the ClinVar database of disease genes, and human genome sequencing projects. The database includes site-specific interaction networks of proteins, upstream enzymes such as kinases, and drugs targeting these enzymes. We also predicted network-rewiring impact of mutations by analyzing gains and losses of kinase-bound sequence motifs. ActiveDriverDB provides detailed visualization, filtering, browsing and searching options for studying PTM-associated mutations. Users can upload mutation datasets interactively and use our application programming interface in pipelines. Integrative analysis of mutations and PTMs may help decipher molecular mechanisms of phenotypes and disease, as exemplified by case studies of *TP53*, *BRCA2* and *VHL*. The open-source database is available at https://www.ActiveDriverDB.org.

INTRODUCTION

DNA sequencing studies have enabled large-scale characterization of human genomes and revealed millions of single nucleotide variants (SNVs), copy number alterations, and other types of genetic variants. Identifying genotype-phenotype associations, molecular mechanisms, causal disease variants and cancer driver mutations remain major challenges of current biomedical research (1,2). Large catalogues of genetic variation comprising tens of thousands of individual and tumour genomes are now available from projects such as The Cancer Genome Atlas (TCGA) (3), the International Cancer Genome Consortium (ICGC) (4), the 1000 Genomes Project (5), The Exome Aggregation Consortium (ExAC) (6), and others. Open-access databases such as ClinVar (7) collect disease genes and mutations. Prediction of functional impact and prioritization of candidate variants primarily relies on evolutionary sequence conservation and other genomic features (8–10), however information about protein interactions and signaling is not routinely applied in such analyses.

Post-translational modifications (PTM) include more than 400 kinds of chemical modifications of amino acids that act as molecular switches and expand the functional repertoire of proteins (11,12). PTMs are carried out by modular reader–writer–eraser networks where specific enzymes induce PTMs in target proteins, remove PTMs, and interact with modified sites (13). Phosphorylation, ubiquitination, acetylation, and methylation are the most commonly characterized PTMs with nearly 400 000 experimentally determined sites in human proteins (14–16). PTMs are involved in various aspects of cellular organization includ-

*To whom correspondence should be addressed. Tel: +1 647 260 7983; Email: Juri.Reimand@utoronto.ca
Present address: B. F. Francis Ouellette, Génome Québec, Montréal, Québec, Canada.

© The Author(s) 2017. 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

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

ing protein activation and degradation, protein-protein interactions, chromatin organization, development, and signaling pathways associated with disease (17–20). Further, PTMs are increasingly drug targetable and used in precision cancer therapies (21–23). Thus PTM information helps interpret genetic variation, genotype-phenotype associations, and molecular disease mechanisms.

PTM sites are enriched in disease mutations and rare variants in the population (24–29). Such mutations often alter sequence motifs bound by PTM enzymes and may cause rewiring of signaling networks (27,30). Importantly, the functional impact of PTM mutations is often underestimated in standard annotation pipelines. We found that 15–30% of disease mutations in PTM sites are considered benign by tools such as PolyPhen2 (8), SIFT (9) and CADD (10), likely because PTM sites are located in disordered protein regions with lower evolutionary conservation (25). Thus PTMs remain understudied in the context of genetic variation and disease. The PhosphositePlus database maintains downloadable datasets with PTM site variation (14), however, a dedicated comprehensive database of genetic variation in PTM sites does not exist to our knowledge.

To address this limitation, we developed ActiveDriverDB, a proteo-genomics resource for interpreting human genome variation using PTM sites (24–27). The database integrates experimentally determined PTM sites with large genomics resources: cancer exomes from TCGA (3,31), known disease genes and mutations from the ClinVar database (7) and population variation from the 1,000 Genomes Project and ESP6500 (5,32). We also display the network context of PTM mutations by analyzing PTM-specific protein-protein interactions and the drugs targeting PTM enzymes that regulate the protein (33). Hundreds of thousands of amino acid substitutions in PTM sites are available in the database for browsing, visualization and interpretation. Datasets can be downloaded or analyzed using our application program interface (API). Users can also interactively upload, store and analyze their own custom datasets of mutations. Our open-source database can be downloaded for local use.

MATERIALS AND METHODS

Genomic and proteomic data in ActiveDriverDB

ActiveDriverDB includes two major types of human -omics data: genomics data on missense SNVs and proteomics data on PTMs (Figure 1, Table 1). Human genome variation datasets include disease-associated SNVs and those apparent in the human population. First, ActiveDriverDB includes somatic cancer mutations of nearly 9000 tumor samples of 34 types from exome sequencing by the TCGA compiled in the recent PanCanAtlas MC3 release (3,31). The TCGA dataset was further filtered to exclude non-passed mutations and hyper-mutated samples. Second, inherited disease mutations from the ClinVar database (7) are also available in ActiveDriverDB. Third, inter-individual genome variation of the human population includes the 1,000 Genomes Project (5) with >2500 genomes and the ESP6500 project (32) with >6500 exomes. Experimentally determined human PTM sites are retrieved from public databases PhosphositePlus (14), Phospho.ELM (15) and HPRD (16) and include primarily proteomics data on the four most frequently characterized PTM types (phosphorylation, acetylation, ubiquitination, methylation). Site-specific protein-protein interactions of substrate proteins and upstream enzymes (primarily kinases) are also included from these databases. We also integrate drugs that target upstream enzymes of PTMs using data from the DrugBank database (33).

Mapping mutations and PTM sites

Substitutions (SNVs) in PTM sites were mapped using our previously designed pipelines (24–27). Genomic coordinates of SNVs were mapped to protein amino acid substitutions using the Annovar software (34) and RefSeq genes (hg19/GRCh37). Peptide sequences corresponding to PTM sites were mapped to RefSeq proteins using exact sequence matching permitting multiple matches per sequence. PTM sites extended seven amino acids before and after of the modified protein residue, and multiple clustered PTM sites are merged into consecutive regions. Protein domains from the InterPro database (35) were mapped into non-redundant regions and combined with disorder predictions of the DISOPRED2 software (36). ActiveDriverDB provides information for 39 159 high-confidence isoforms of 19 062 human genes. We identify those by HGNC gene symbols (37) or RefSeq transcript IDs and show primary isoforms according to the Uniprot database (38) by default.

Impact of mutations on PTM sites

Amino acid substitutions (SNVs) in PTM sites are further annotated regarding their position relative to PTMs and potential impact on signaling networks. PTM mutations are considered direct if they substitute the central PTM amino acid residue while indirect mutations are classified either as proximal or distal (1–2 or 3–7 amino acid residues to the nearest PTM site, respectively). We distinguish variants that affect different types of PTM sites and mutations affecting multiple adjacent PTMs. To estimate the network impact of mutations, we performed sequence motif analysis with our machine learning method MIMP (27) using 476 models of sequence motifs of 322 kinases and families (14–16,39,40). MIMP analyses substitutions in PTM sites and predicts whether these cause loss of existing kinase-bound motifs or create new motifs, suggesting the impact of mutations on the rewiring of cellular signaling networks.

Software design and availability

The ActiveDriverDB website uses the Flask micro-framework and two relational databases: the first for constant biological data and the second for dynamic content and user-provided data. An additional key-value BerkleyDB database allows mapping of all potential missense SNVs in the human genome. Visualizations are implemented in the d3.js framework. Our needle-plots are inspired by the muts-needle-plot library (https://zenodo.org/record/14561). All code is available on terms of LGPL 2.1 license. Documentation is available at https://github.com/reimandlab/ActiveDriverDB.

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

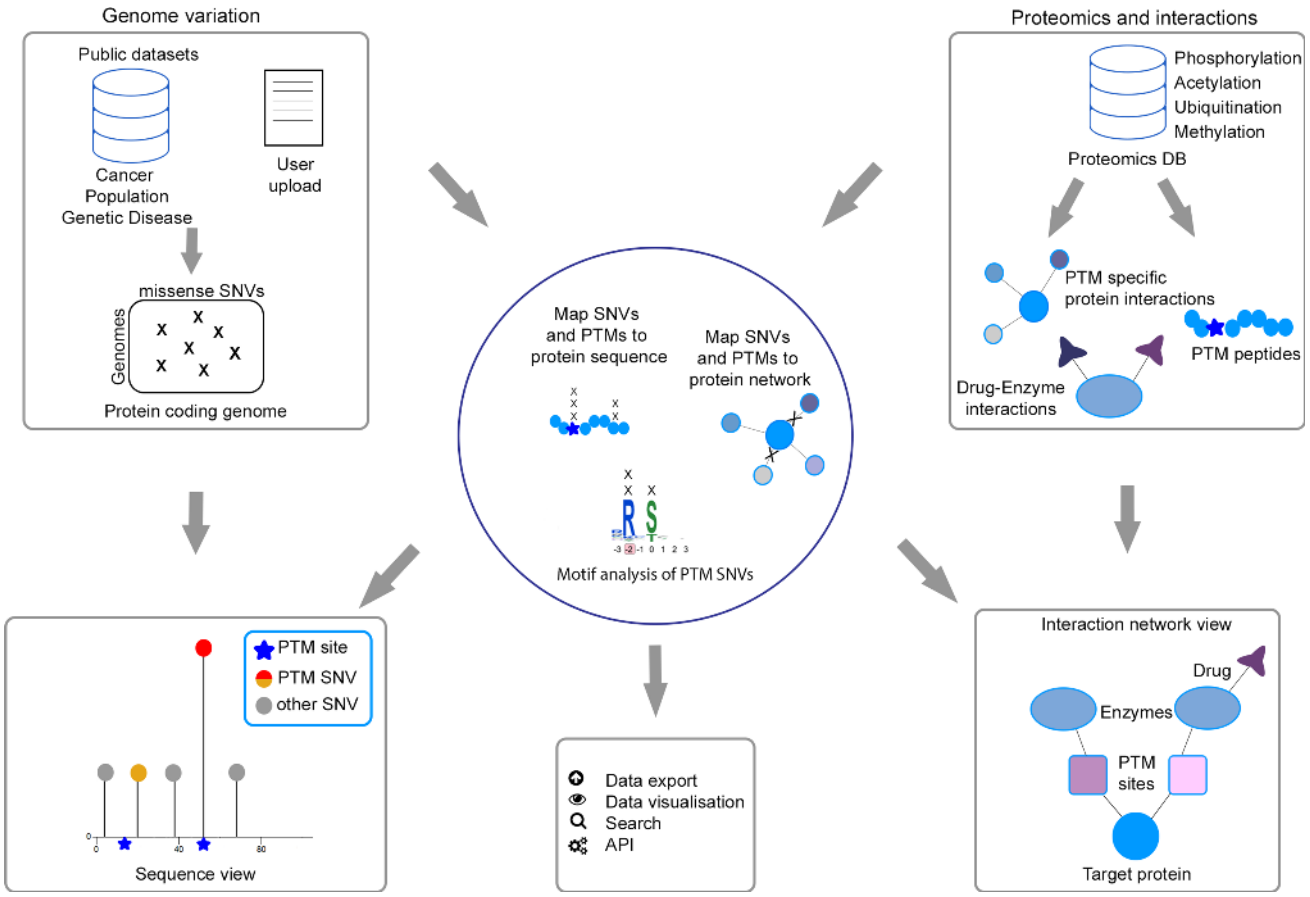

Figure 1. Overview and workflow of ActiveDriverDB. Our database integrates genomic and proteomic data for interpreting disease mutations and human inter-individual variation with PTMs. Genomics datasets include cancer exome sequencing studies (TCGA), disease genes and mutations (ClinVar), and human genome variation studies (ESP6500, the 1000 Genomes Project) (top-left panel). Proteomics datasets include PTM sites of four commonly studied PTM types, site-specific interactions of PTM enzymes and target proteins, and drug interactions with PTM enzymes (top-right panel). Our systematic analysis pipeline aligns PTM sites with missense SNVs, predicts the impact of amino acid substitutions on kinase-bound sequence motifs using the MIMP method, and organizes site-specific interaction networks of PTMs, upstream enzymes and drugs (middle panel). The protein sequence view shows the distribution of PTMs and substitutions along the protein sequence (bottom left panel), while the interaction network view shows site-specific interactions of mutated proteins with upstream PTM enzymes and their associated drugs (bottom right panel). The database also provides exporting, visualization, searching and automated analysis tools (bottom middle panel).

RESULTS

Thousands of disease mutations are enriched in PTM sites

The database is available at https://ActiveDriverDB.org. In total, ActiveDriverDB characterises 506 974 unique amino acid substitutions in PTM sites across high-confidence protein isoforms, including 221 472 in cancer genomes, 27 305 in inherited diseases and 143 489 and 185 982 in human genomes from the population sequencing projects 1000 Genomes and ESP6500, respectively. These substitutions affect the four types of most frequently characterized PTMs: phosphorylation sites (299 241), ubiquitination sites (67 933), acetylation sites (21 670) and methylation sites (5666), with 385 185 distinct sites in total across all protein isoforms. Among 558 high-confidence cancer genes of the Cancer Gene Census database (41), 9542 unique substitutions in the TCGA dataset (25%) are associated with PTM sites when considering primary isoforms of proteins (5773 expected from sampling of substitutions from the 1000 Genomes dataset, empirical $P < 10^{-5}$). Among disease genes annotated in the ClinVar database, 11 041 unique substitutions (21%) are associated with PTM sites (7963 PTM SNVs expected, $P < 10^{-5}$). Enrichment of disease-associated mutations in PTM sites is in agreement with our earlier studies (24–27). These statistics suggest that a large fraction of germline and somatic disease mutations can be interpreted using PTM information.

Visualization and analysis of mutations in PTM sites

The two primary pages of the database, *the protein sequence view* and *the interaction network view*, are focused on individual proteins (genes). Both views provide detailed visualizations of PTM-associated amino acid substitutions, tables with additional information, protein descriptions, and external links. The views permit filtering of mutations by dataset (inherited disease mutations, somatic cancer mutations, or inter-individual genome variation), disease types, pathogenicity and PTM types. All non-PTM mutations can be filtered as well. Both views display the primary isoform as default, while alternative isoforms can be selected.

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

Table 1. Overview of genome variation datasets and post-translational modifications included in ActiveDriverDB

	TCGA PanCanAtlas	ClinVar	1000 Genomes	ESP6500	Total
Dataset					
Size	8 856 exomes	494 059 records	2504 genomes	6503 exomes	-
Description	Cancer (somatic)	Inherited disease	General population		-
Mutations					
Total	1 595 400	137 860	1 066 906	1 318 972	3 588 280
in PTM sites	221 472	27 305	143 489	185 982	506 974
with network-rewiring effect	30 882	2 869	20 525	26 498	70 518
Annotated nucleotides (hg19/GRCh37)	1 865 173	155 824	1 206 968	1 486 067	4 124 041
PTM sites affected by mutations*					
Total	214 362	16 597	157 420	186 800	303 401
Phosphorylation sites	169 632	12 999	128 097	150 505	239 509
Acetylation sites	11 581	1 258	7 216	8 988	16 384
Ubiquitination sites	35 058	2 659	22 678	28 226	50 081
Methylation sites	3 377	272	2 293	2 577	4 416
Proteins with mutations affecting PTM sites					
Total	27 316	3 317	25 132	26 202	29 462
Kinases & PTM enzymes	613	115	580	594	624
Kinase families	127	58	127	126	127
PTM sites					
Total	-	-	-	-	385 185
Phosphorylation sites	-	-	-	-	299 241
Acetylation sites	-	-	-	-	21 670
Ubiquitination sites	-	-	-	-	67 933
Methylation sites	-	-	-	-	5 666

Counts of PTMs and amino acid substitutions reflect all high-confidence protein isoforms collected in the database.

Protein sequence view: mutation impact on PTMs, sequence features and network rewiring. The main components of this view include a needleplot with mutations and impact on PTM sites, sequence tracks with protein domains (35) and disorder predictions (36), and a detailed table of mutations. The needleplot represents the protein sequence and its PTM sites horizontally, while mutations extend vertically from the sequence according to their frequency (Figure 2A). Colored circles on top of needles represent mutation impact on PTM sites, and mouse-over motion shows information about the mutation, disease annotations, known PTM enzymes such as bound kinases, predictions of network rewiring with mutation-induced gains and losses of sequence motifs (Figure 2B), and drugs targeting the upstream PTM enzymes. The needleplot can be zoomed and searched by amino acid position. Mutations are also described in the table below (Figure 2C). The needleplot can be exported as a high-resolution PDF (Portable Document Format file) and the mutation table can be exported as a spreadsheet.

Interaction network view: PTM site-specific interactions, upstream enzymes and drug targets. This view displays the selected protein in a site-specific interaction network with upstream enzymes and associated drugs (Figure 2D). Two interaction networks are available: *the high-confidence experimental network* includes experimentally determined kinase–substrate interactions, and *the computationally predicted network* includes gained and lost kinase-substrate interactions derived from sequence motif analysis with MIMP. Most interactions comprise phosphorylation sites and associated kinases with largest body of experimental data. The network view uses an automatic layout algorithm that emphasizes the hierarchical network structure. It can be zoomed and arranged for clarity and exported as a high-resolution PDF.

Searching and browsing PTM mutations in proteins

The database provides a flexible graphical user interface for finding, visualizing and interpreting mutations in PTM sites and their potential impact on signaling networks.

Searching for genes, pathways, and diseases. The main search bar supports several options. First, the user can identify a gene (protein) of interest by either its HGNC symbol that retrieves the primary isoform (e.g.*TP53*) or a RefSeq transcript ID that retrieves a specific isoform (e.g. *NM_000546*). Second, genes associated with biological processes of Gene Ontology (42), molecular pathways of Reactome (43) (e.g. *Wnt signaling pathway*), and diseases in the ClinVar database (e.g. *Noonan syndrome*) can be looked up. These search options benefit researchers who are interested in specific genes, pathways, or disease mutations.

Searching for mutations. The user can search for a gene or protein using amino acid substitutions (e.g. *TP53 R282W*) or coordinates of missense mutations (e.g. *chr17 7577094 C T*) through our rapid indexing system that covers all potential missense SNVs in the human genome. Search for mutations is especially beneficial for genetics researchers who have identified interesting missense SNVs in a genome-wide association or DNA sequencing study.

Browsing top disease-associated genes and pathways. Users may browse sets of disease-associated genes and pathways with unexpectedly frequent mutations in PTM sites. Candidate gene lists are available for cancer mutations

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

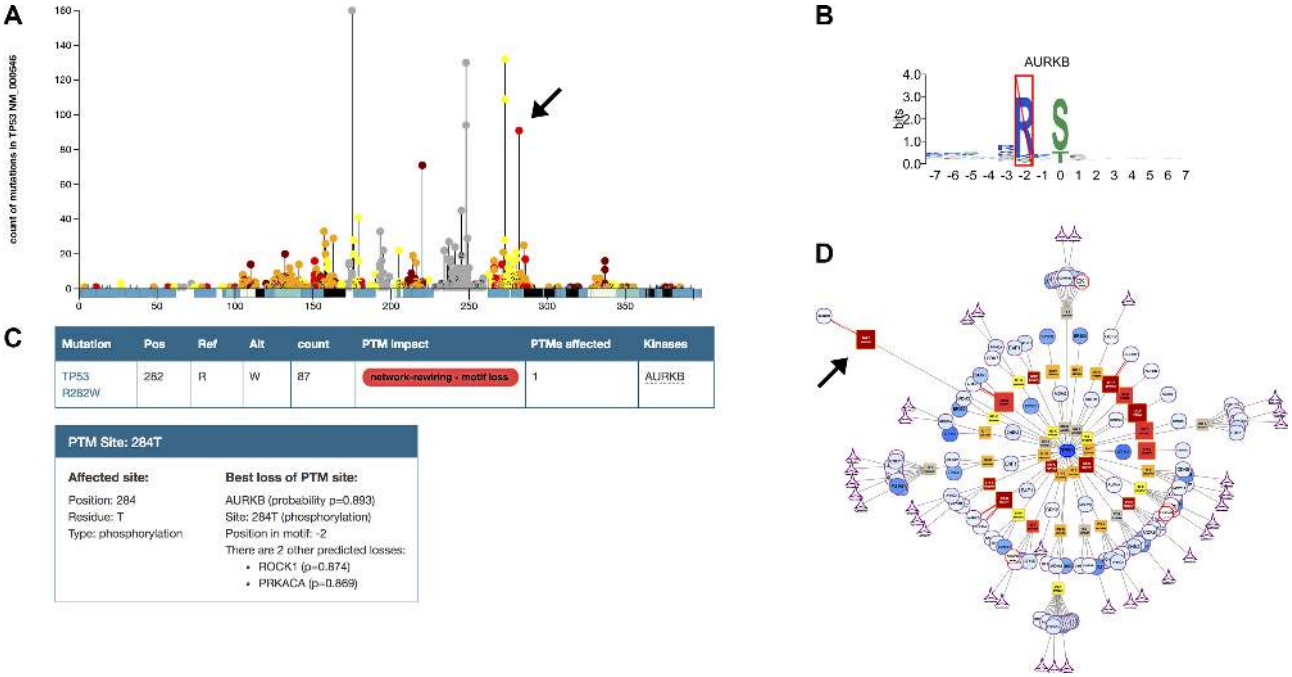

Figure 2. PTM-associated cancer mutations in the tumor suppressor protein TP53. (**A**) Needleplot in the protein sequence view shows the distribution of missense cancer mutations from TCGA (vertical bars) and their associations with PTM sites (blue boxes) with protein sequence on the x-axis and number of mutations on the Y-axis. (**B**) The substitution R282W disrupts the sequence motif bound by the Aurora Kinase B (AURKB). (**C**) Detailed table of mutations with disease associations and impact of substitutions on PTMs. (**D**) Experimentally determined interaction network shows the TP53 protein (middle node) and its PTM site-specific interactions with upstream enzymes, as well as approved drugs targeting these enzymes. Node shapes indicate types of interacting molecules and sites: protein of interest (oval), PTM sites (squares), enzymes interacting with PTM site (circles), and drugs targeting the enzyme (triangles). Arrows indicate the interaction of TP53 and Aurora kinase B at phosphosite T284 and the associated substitution R282W.

from TCGA and inherited disease mutations from ClinVar. Genes are ranked according to statistical significance of PTM mutations computed using our ActiveDriver method (26). These lists are useful for novice users of the database who are looking for an overview of the database through examples of genes and pathways.

Analysing custom datasets of mutations. Users can upload VCF or tab-separated files to analyze their datasets of variants with PTM information, using chromosomal or protein coordinates. A password-protected area is available for uploaded data.

Application programming interface (API). ActiveDriverDB includes an API allowing access to the database with programming languages like R and Python using the Representational State Transfer (REST) pattern. The API accepts chromosomal or protein coordinates of mutations, converts these appropriately and returns PTM annotations. Filters for mutation types (cancer, inherited disease, population), querying of mutations by gene symbol or RefSeq ID, and other options are also supported. Datasets of PTM sites and associated mutations are also available for download for advanced computational biology studies. We provide up-to-date input datasets for the ActiveDriver method (26) that reveals proteins with statistical enrichment of mutations in PTM sites.

Case studies of PTM-associated disease mutations

Frequent cancer mutations in PTM sites in the tumor suppressor protein TP53. Mutated in 50% of cancers, the transcription factor TP53 relies on its DNA-binding activity to perform its function as a tumour suppressor (44). Consistently, most mutations are found in the DNA-binding domain (DBD) of the protein with a third clustered in seven hotspot residues with high-frequency mutations (R175, G245, R248, R249, R273, R282) (Figure 2A) (45). Although most of the mutations in TP53 are associated with loss of function, mutations such as R282W lead to gain of function and TP53 with distinct oncogenic properties (reviewed in (45,46)). The mechanisms by which R282W leads to this transformation are still unclear (47). MIMP analysis of sequence motifs in TP53 predicts that the mutations R282W/G/Q rewire the phosphosite T284 by abolishing the sequence motif of the AURKB in the DBD of TP53 (Figure 2B and C) (27). This phosphosite is bound by AURKB kinase *in vitro* and in cells with AURKB ectopic expression (Figure 2D) (27,48,49). The substitution T284E inhibits the ability of TP53 to promote *CDKN1A* expression (48) highlighting a role of AURKB in TP53 signaling. More than 200 mutations in the TCGA dataset potentially interfere with the phosphosite T284 (Table 2), suggesting that these regulate a common function of TP53. As R282W is associated with early cancer development (50), the mutations should be further studied regarding their impact on the gene regulatory and tumour suppressive roles of TP53 (46). By highlighting clusters of PTM-associated mutations,

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

our database helps design experiments to understand the post-translational regulation of TP53.

PTM-associated disease mutations of BRCA2 and the DNA damage response pathway. Mutations in the tumor suppressor *BRCA2* are associated with elevated risk of breast and ovarian cancers as well as Fanconi Anemia, a rare chromosome instability syndrome characterized by aplastic anemia and susceptibility to childhood cancer (51,52). Consistently, disease-associated SNVs in *BRCA2* reported in the ClinVar database are associated with familial breast cancer and hereditary cancer-predisposing syndrome. BRCA2 is essential for DNA double-strand break (DSB) repair by homologous recombination and protects the stalled replication fork (53). To prevent genomic instability, BRCA2 relies on interactions with RAD51 mediated by cell cycle-dependent kinases (CDKs) (54–57). Using the ActiveDriverDB database, we found that a significant number of somatic and inherited cancer mutations of BRCA2 coincide with phosphosites (29 SNVs in ClinVar, FDR = 10^{-47}; 15 SNVs in TCGA, FDR = 10^{-6}) (Figure 3A). Interestingly, three phosphosites S3291, S3319 and T3323 occur in the C-terminus of BRCA2 whose deletion is associated with increased radiation sensitivity and early-onset breast and ovarian cancer (58–60). The C-terminal TR2 domain at 3265-3330 a.a. mediates the interaction of BRCA2 with nucleofilaments of RAD51 (55,57) and its phosphorylation by CDKs inhibits this interaction and is essential for mitotic entry (54,55,57,61). Substitutions that either abolish these phosphosites or the CDK consensus sites (P3292L/S, P3320H and P3324L) (Figure 4B–D) are associated with familial breast cancer and hereditary cancer-predisposing syndrome, suggesting that the mutations interfere with maintenance of genomic stability. Consistently, the substitution S3291A inhibits the interaction of BRCA2 with RAD51 filaments, a phenotype that abrogates the replication fork protection without affecting DNA repair (62). Whether the phosphorylation of S3319 and T3323 regulates BRCA2 is unknown, however mutant BRCA2 with glutamate substitutions in these amino acids still interacts with RAD51 filaments (54). This example illustrates the integration of PTM information and germline disease mutations to predict novel experimentally testable hypotheses of mechanisms.

Network-rewiring mutations in the tumour suppressor VHL alter putative CDK binding sites. The tumour suppressor VHL encodes a member of a E3 ubiquitin ligase complex that inhibits oncogenic substrates such as protein kinase C, retinol binding protein 1, and hypoxia-inducible transcription factors (HIF) (63,64). VHL is frequently inactivated in cancer and clear cell renal cell carcinomas (ccRCCs) harbour gene-silencing mutations including L169P and others in the p.157–172 subdomain (64) (Figure 4A). Phosphorylation of VHL at S168 by NIMA Related Kinase 1 (NEK1) has been associated with VHL degradation and cilliary homeostasis (65). The mutation L169P may impact VHL signaling as it flanks the phosphosites S168 and Y175 bound by the NEK1 kinase. ActiveDriverDB analysis suggests that three L169P substitutions observed in TCGA kidney cancers may induce gains of phosphosites of the cyclin dependent kinase 1 (CDK1) or related CDK and MAPK kinases (Figure 4B and C). While little is known about the interactions of VHL and CDKs, VHL inactivation has been linked to increased levels of CDK1 and CDK2 (66), and CDK1 stabilizes HIF transcription factors that are targets of VHL (67). Studying the L169P mutation in the context of VHL phosphorylation and upstream kinases may reveal details about disease mechanisms (Figure 4D). Since CDK1 is pharmaceutically targetable, drug assays using alsterpaullone and alvocidib (33) may also advance development of targeted therapies.

DISCUSSION

ActiveDriverDB is a comprehensive human proteogenomics resource that uses PTMs to interpret disease mutations and inter-individual variation. Although PTMs are important regulators of protein function and signaling pathways, genetic variant impact analysis pipelines usually neglect this information. Our database aims to advance analysis of missense mutations using PTMs. Novice users of our database can start from example queries of well-annotated genes and browse gene lists with PTM-enriched disease mutations. Basic and translational researchers can look up their favourite genes, upload candidate variants from DNA sequencing experiments, and export tables and publication-quality figures. Computational biologists can use the API to automatically analyze variants and download entire datasets for advanced studies.

Our collection of PTM sites is derived from many published studies that represent diverse cells and experimental conditions. While this large dataset allows us to maximise coverage of disease mutations and genetic variation, all PTM sites may be not directly comparable with one another. PTM sites observed in certain cell types may not be expressed as proteins or may be excluded due to alternative splicing in cells relevant to a disease of interest. Although we processed PTM sites uniformly across the entire collected dataset, false positives may have emerged from analyses in original studies, the databases reporting the data, or our multi-database integration pipeline. We recommend that users validate top PTM sites by looking these up in databases such as PhosphoSitePlus and the publications and supplementary materials that originally reported the PTM site.

We plan several important future developments of the database. Maintaining timely biomedical data resources is essential as new datasets accumulate rapidly and outdated resources hamper scientific advances (68). Thus we aim to provide at least annual updates of our database to include recent large-scale genomics and proteomics studies and molecular interaction networks. Recent proteomics technologies enable large-scale characterization of other PTM types such as glycosylation (69) and SUMOylation (70) and such datasets will be included in future releases. Additional species and genomes will be also considered, such as the most recent human genome assembly (GRCh38) and model organisms such as mouse and Arabidopsis with abundant genome variation and proteomics data (14,71).

Interpreting inter-individual genetic variation will become an increasingly important challenge as we enter the

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

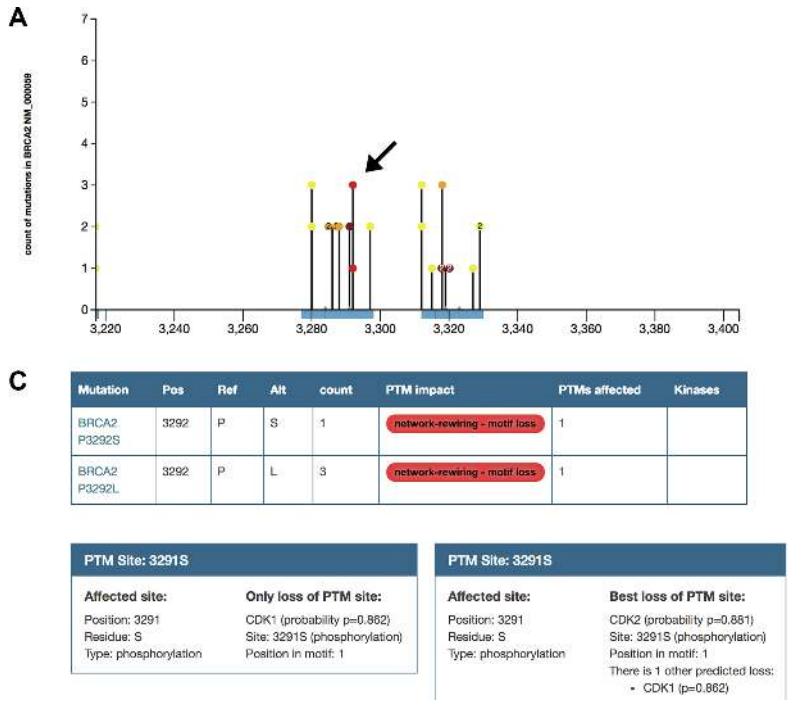

Figure 3. PTM-associated mutations in the BRCA protein involved in DNA repair and breast cancer. (**A**) Zoomed needleplot shows germline disease mutations located in three phosphorylation sites in the protein sequence at 3,200-3,400 residues. Only PTM-associated mutations are shown. (**B**) Mutations P3292L and P3292S are predicted to disrupt the sequence motif of the CDK2 kinase. (**C**) Table shows additional information on the two mutations. (**D**) The computationally derived PTM interaction network of BRCA2, kinases predicted to interact with mutant BRCA2, and drugs targeting the kinases. Arrows point to the mutations P3292L/S.

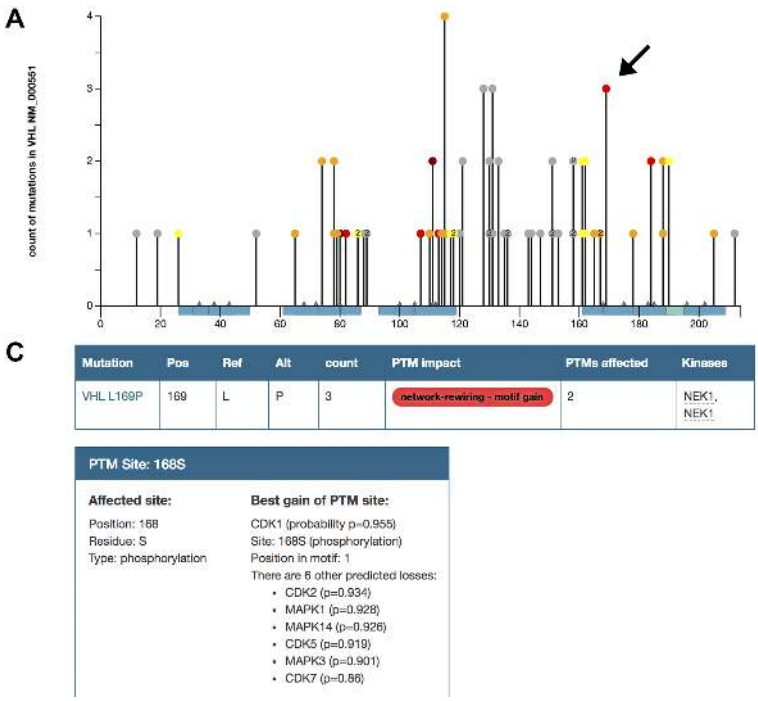

Figure 4. PTM-associated mutations in the tumor suppressor protein VHL. (**A**) Needleplot of mutations from the TCGA dataset with the VHL mutation L169P near two phosphsites. (**B**) The mutation L196P is predicted to disrupt the sequence motif of the CDK1 kinase. (**C**) Table shows additional information on the mutation. (**D**) The computationally predicted interaction network of VHL, its PTM sites, kinases predicted to interact with mutant VHL according to the MIMP method, and drugs targeting the kinases. Arrows indicate the mutation L169P.

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

Table 2. PTM-associated cancer mutations affecting the phosphosite T284 in the tumor suppressor protein TP53

Phosphosite	Reference a.a.	Mutated a.a	Number of mutations		Impact on PTM
			TCGA	ClinVar	
T284	R280	T	19	3	Distal
		I	4	3	Distal
		K	14		Distal
		S	3		Distal
		G	5		Distal
	D281	V	4	2	Distal
		G		2	Distal
		A	2		Distal
		H	3		Distal
		N	2		Distal
		Y	7		Distal
		E	6		Distal
	R282	W	74	5	Network-rewiring
		G	2	5	Network-rewiring
		Q	1	2	Network-rewiring
		L		2	Network-rewiring
	R283	H	1	2	Proximal
		P	8		Proximal
		C		3	Proximal
		S		3	Proximal
	T284	I		1	Direct
		P	1		Direct
		A			Direct
		E			Direct
	E285	V	4	2	Proximal
		K	23		Proximal
		Q	1		Network-rewiring
	E286	K	11	1	Network-rewiring
		A	1		Proximal
		G	4		Proximal
		V	1		Proximal
		Q	1		Proximal
Total	-	-	201	37	-

Somatic and germline disease mutations that potentially affect the phosphosite of Aurora kinase B.

era of personal genomics. Integration of proteomic and genomic information for deciphering the impact of variation on cellular systems and organism-level phenotypes is a powerful approach that will improve with future datasets of increased magnitude and complexity. We provide an integrated database resource to the research community to enable future discoveries.

ACKNOWLEDGEMENTS

We would like to thank Andrea Sabo for help with web design and members of the Reimand lab for useful comments. The results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/ as outlined in the TCGA publication guidelines. We are grateful to researchers and developers of databases PhosphoSitePlus, PhosphoELM, HPRD, ClinVar, DrugBank, Uniprot and others for providing high-quality and frequently maintained datasets.
Author contributions: M.K. developed the software. M.K. and J.R. designed the software. M.K., M.P., T.H., A.F.T. and J.R. analyzed the data. M.P., M.K., K.C., I.D., B.F.F.O., J.T.Y. and J.R. developed web design, documentation, and user stories. T.H., A.F.T. and J.R. compiled research case studies. All authors read and approved the final manuscript.

FUNDING

Investigator Award (to J.R.) from the Ontario Institute for Cancer Research (OICR) provided by the Government of Ontario; stipend to M.K. from the Google Summer of Code (GSoC) project; Canadian Cancer Research Society (CRS) Operating Grant to A.F.T. and J.R. [21428]. Funding for open access charge: internal funding from the Ontario Institute for Cancer Research.
Conflict of interest statement. None declared.

REFERENCES

1. MacArthur,D.G., Manolio,T.A., Dimmock,D.P., Rehm,H.L., Shendure,J., Abecasis,G.R., Adams,D.R., Altman,R.B., Antonarakis,S.E., Ashley,E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
2. Gonzalez-Perez,A., Mustonen,V., Reva,B., Ritchie,G.R., Creixell,P., Karchin,R., Vazquez,M., Fink,J.L., Kassahn,K.S., Pearson,J.V. *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.
3. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
4. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

5. Genomes Project, C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
6. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J., Cummings,B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
7. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
8. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
9. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
10. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
11. Montecchi-Palazzi,L., Beavis,R., Binz,P.A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
12. Mann,M. and Jensen,O.N. (2003) Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, **21**, 255–261.
13. Pawson,T. (1995) Protein modules and signalling networks. *Nature*, **373**, 573–580.
14. 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.
15. 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.
16. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
17. Pawson,T. and Scott,J.D. (2005) Protein phosphorylation in signaling–50 years and counting. *Trends Biochem. Sci.*, **30**, 286–290.
18. Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.
19. Welchman,R.L., Gordon,C. and Mayer,R.J. (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nat. Rev. Mol. Cell Biol.*, **6**, 599–609.
20. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
21. Hoeller,D. and Dikic,I. (2009) Targeting the ubiquitin system in cancer therapy. *Nature*, **458**, 438–444.
22. Gharwan,H. and Groninger,H. (2016) Kinase inhibitors and monoclonal antibodies in oncology: clinical implications. *Nat. Rev. Clin. Oncol.*, **13**, 209–227.
23. Jones,P.A., Issa,J.P. and Baylin,S. (2016) Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.*, **17**, 630–641.
24. Narayan,S., Bader,G.D. and Reimand,J. (2016) Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. *Genome Med.*, **8**, 55.
25. Reimand,J., Wagih,O. and Bader,G.D. (2015) Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genet.*, **11**, e1004919.
26. Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.
27. Wagih,O., Reimand,J. and Bader,G.D. (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods*, **12**, 531–533.
28. Li,S., Iakoucheva,L.M., Mooney,S.D. and Radivojac,P. (2010) Loss of post-translational modification sites in disease. *Pac. Symp. Biocomput.*, 337–347.
29. Wang,Y., Cheng,H., Pan,Z., Ren,J., Liu,Z. and Xue,Y. (2015) Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility. *J. Mol. Cell Biol.*, **7**, 187–202.
30. Creixell,P., Schoof,E.M., Simpson,C.D., Longden,J., Miller,C.J., Lou,H.J., Perryman,L., Cox,T.R., Zivanovic,N., Palmeri,A. *et al.* (2015) Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, **163**, 202–217.
31. Ellrott,K., Bailey,M.H., Saksena,G., Covington,K.R., Kandoth,C., Stewart,C., McLellan,M., Sofia,H.J., Hutter,C., Getz,G. *et al.* (2017) Automating somatic mutation calling for ten thousand tumor exomes. submitted.
32. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
33. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
34. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
35. Finn,R.D., Attwood,T.K., Babbitt,P.C., Bateman,A., Bork,P., Bridge,A.J., Chang,H.Y., Dosztanyi,Z., El-Gebali,S., Fraser,M. *et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.*, **45**, D190–D199.
36. Ward,J.J., McGuffin,L.J., Bryson,K., Buxton,B.F. and Jones,D.T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.
37. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
38. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
39. 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.
40. Newman,R.H., Hu,J., Rho,H.S., Xie,Z., Woodard,C., Neiswinger,J., Cooper,C., Shirley,M., Clark,H.M., Hu,S. *et al.* (2013) Construction of human activity-based phosphorylation networks. *Mol. Syst. Biol.*, **9**, 655.
41. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
42. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
43. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
44. Kastenhuber,E.R. and Lowe,S.W. (2017) Putting p53 in Context. *Cell*, **170**, 1062–1078.
45. Freed-Pastor,W.A. and Prives,C. (2012) Mutant p53: one name, many proteins. *Genes Dev.*, **26**, 1268–1286.
46. Muller,P.A. and Vousden,K.H. (2013) p53 mutations in cancer. *Nat. Cell Biol.*, **15**, 2–8.
47. Zhang,Y., Coillie,S.V., Fang,J.Y. and Xu,J. (2016) Gain of function of mutant p53: R282W on the peak? *Oncogenesis*, **5**, e196.
48. Wu,L., Ma,C.A., Zhao,Y. and Jain,A. (2011) Aurora B interacts with NIR-p53, leading to p53 phosphorylation in its DNA-binding domain and subsequent functional suppression. *J. Biol. Chem.*, **286**, 2236–2244.
49. Gully,C.P., Velazquez-Torres,G., Shin,J.H., Fuentes-Mattei,E., Wang,E., Carlock,C., Chen,J., Rothenberg,D., Adams,H.P., Choi,H.H. *et al.* (2012) Aurora B kinase phosphorylates and instigates degradation of p53. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E1513–E1522.
50. Xu,J., Qian,J., Hu,Y., Wang,J., Zhou,X., Chen,H. and Fang,J.Y. (2014) Heterogeneity of Li-Fraumeni syndrome links to unequal gain-of-function effects of p53 mutations. *Sci. Rep.*, **4**, 4223.
51. Lancaster,J.M., Wooster,R., Mangion,J., Phelan,C.M., Cochran,C., Gumbs,C., Seal,S., Barfoot,R., Collins,N., Bignell,G. *et al.* (1996)

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022

BRCA2 mutations in primary breast and ovarian cancers. *Nat. Genet.*, **13**, 238–240.
52. Wooster,R., Bignell,G., Lancaster,J., Swift,S., Seal,S., Mangion,J., Collins,N., Gregory,S., Gumbs,C. and Micklem,G. (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.
53. Fradet-Turcotte,A., Sitz,J., Grapton,D. and Orthwein,A. (2016) BRCA2 functions: from DNA repair to replication fork stabilization. *Endocr. Relat. Cancer*, **23**, T1–T17.
54. Esashi,F., Christ,N., Gannon,J., Liu,Y., Hunt,T., Jasin,M. and West,S.C. (2005) CDK-dependent phosphorylation of BRCA2 as a regulatory mechanism for recombinational repair. *Nature*, **434**, 598–604.
55. Esashi,F., Galkin,V.E., Yu,X., Egelman,E.H. and West,S.C. (2007) Stabilization of RAD51 nucleoprotein filaments by the C-terminal region of BRCA2. *Nat. Struct. Mol. Biol.*, **14**, 468–474.
56. Yata,K., Bleuyard,J.Y., Nakato,R., Ralf,C., Katou,Y., Schwab,R.A., Niedzwiedz,W., Shirahige,K. and Esashi,F. (2014) BRCA2 coordinates the activities of cell-cycle kinases to promote genome stability. *Cell Rep*, **7**, 1547–1559.
57. Davies,O.R. and Pellegrini,L. (2007) Interaction with the BRCA2 C terminus protects RAD51-DNA filaments from disassembly by BRC repeats. *Nat. Struct. Mol. Biol.*, **14**, 475–483.
58. Hakansson,S., Johannsson,O., Johansson,U., Sellberg,G., Loman,N., Gerdes,A.M., Holmberg,E., Dahl,N., Pandis,N., Kristoffersson,U. *et al.* (1997) Moderate frequency of BRCA1 and BRCA2 germ-line mutations in Scandinavian familial breast cancer. *Am. J. Hum. Genet.*, **60**, 1068–1078.
59. Donoho,G., Brenneman,M.A., Cui,T.X., Donoviel,D., Vogel,H., Goodwin,E.H., Chen,D.J. and Hasty,P. (2003) Deletion of Brca2 exon 27 causes hypersensitivity to DNA crosslinks, chromosomal instability, and reduced life span in mice. *Genes Chromosomes Cancer*, **36**, 317–331.
60. Morimatsu,M., Donoho,G. and Hasty,P. (1998) Cells deleted for Brca2 COOH terminus exhibit hypersensitivity to gamma-radiation and premature senescence. *Cancer Res.*, **58**, 3441–3447.
61. Ayoub,N., Rajendra,E., Su,X., Jeyasekharan,A.D., Mahen,R. and Venkitaraman,A.R. (2009) The carboxyl terminus of Brca2 links the disassembly of Rad51 complexes to mitotic entry. *Curr. Biol.*, **19**, 1075–1085.
62. Schlacher,K., Christ,N., Siaud,N., Egashira,A., Wu,H. and Jasin,M. (2011) Double-strand break repair-independent role for BRCA2 in blocking stalled replication fork degradation by MRE11. *Cell*, **145**, 529–542.
63. Moore,L.E., Nickerson,M.L., Brennan,P., Toro,J.R., Jaeger,E., Rinsky,J., Han,S.S., Zaridze,D., Matveev,V., Janout,V. *et al.* (2011) Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: associations with germline VHL polymorphisms and etiologic risk factors. *PLoS Genet.*, **7**, e1002312.
64. Rathmell,W.K. and Chen,S. (2008) VHL inactivation in renal cell carcinoma: implications for diagnosis, prognosis and treatment. *Expert Rev. Anticancer Ther.*, **8**, 63–73.
65. Patil,M., Pabla,N., Huang,S. and Dong,Z. (2013) Nek1 phosphorylates Von Hippel-Lindau tumor suppressor to promote its proteasomal degradation and ciliary destabilization. *Cell Cycle*, **12**, 166–171.
66. Hongo,F., Takaha,N., Oishi,M., Ueda,T., Nakamura,T., Naitoh,Y., Naya,Y., Kamoi,K., Okihara,K., Matsushima,T. *et al.* (2014) CDK1 and CDK2 activity is a strong predictor of renal cell carcinoma recurrence. *Urol. Oncol.*, **32**, 1240–1246.
67. Warfel,N.A., Dolloff,N.G., Dicker,D.T., Malysz,J. and El-Deiry,W.S. (2013) CDK1 stabilizes HIF-1alpha via direct phosphorylation of Ser668 to promote tumor growth. *Cell Cycle*, **12**, 3689–3701.
68. Wadi,L., Meyer,M., Weiser,J., Stein,L.D. and Reimand,J. (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706.
69. Moremen,K.W., Tiemeyer,M. and Nairn,A.V. (2012) Vertebrate protein glycosylation: diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.*, **13**, 448–462.
70. Hendriks,I.A. and Vertegaal,A.C. (2016) A comprehensive compilation of SUMO proteomics. *Nat. Rev. Mol. Cell Biol.*, **17**, 581–595.
71. Durek,P., Schmidt,R., Heazlewood,J.L., Jones,A., MacLean,D., Nagel,A., Kersten,B. and Schulze,W.X. (2010) PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res.*, **38**, D828–D834.

Downloaded from https://academic.oup.com/nar/article/46/D1/D901/4566599 by guest on 20 August 2022