An Atlas of the Human Kinome Reveals the Mutational Landscape Underlying Dysregulated Phosphorylation Cascades in Cancer

Aleksandra Olow1, Zhongzhong Chen2, R. Hannes Niedner1, Denise M. Wolf1, Christina Yau1, Aleksandr Pankov1, Evelyn Pei Rong Lee1, Lamorna Brown-Swigart1, Laura J. van ’t Veer1, and Jean-Philippe Coppé1

Abstract

Kinase inhibitors are used widely to treat various cancers, but adaptive reprogramming of kinase cascades and activation of feedback loop mechanisms often contribute to therapeutic resistance. Determining comprehensive, accurate maps of kinase circuits may therefore help elucidate mechanisms of response and resistance to kinase inhibitor therapies. In this study, we identified and validated phosphorylatable target sites across human cell and tissue types to generate Phospho-Atlas, a map of 1,733 functionally interconnected proteins comprising the human phospho-reactome. A systematic cura-
tion approach was used to distill protein phosphorylation data cross-referenced from 38 public resources. We demonstrated how a catalog of 2,617 stringently verified heptameric peptide regions at the catalytic interface of kinases and substrates could expose mutations that recurrently perturb specific phospho-
hubs. In silico mapping of 2,896 nonsynonymous tumor variants identified from thousands of tumor tissues also revealed that normal and aberrant catalytic interactions co-occur fre-
quently, showing how tumors systematically hijack, as well as spare, particular subnetworks. Overall, our work provides an important new resource for interrogating the human tumor kinome to strategically identify therapeutically actionable kinase networks that drive tumorigenesis. Cancer Res; 76(7): 1733–45. ©2016 AACR.

Introduction

Biological processes are largely mediated by fine-tuned cas-
cades of protein–protein interactions (PPI). Many PPIs are tran-
sient and enzymatic in nature, resulting in highly dynamic and plastic signaling networks capable of propagating information within and between cells. Protein kinases are phosphorylating enzymes that regulate the function of specific protein substrates by interacting with and chemically modifying them (1). Phos-
phorylation circuits govern most cellular processes essential to tissue homeostasis. Disruption of the activity of kinases or their substrates by mutations or aberrant signaling is associated with cancer and other diseases. As a pivotal cause of disease, disrupted phosphorylation circuits are frequent targets for therapeutic interven-
tions. Kinase inhibitors have proved successful in a number of clinical scenarios, including targeting HER2 in breast cancer, BCR-
ABL in chronic myeloid leukemia, and BRAFV600E in melanoma. However, adaptive reprogramming of kinase cascades and feed-
back loop mechanisms contributes to the development of treat-
ment resistance, the primary obstacle in achieving sustained responses to targeted therapies (2). Therefore, it is imperative not only to understand individual kinase–substrate protein relationships but also to acquire in-depth knowledge of the inter-
connectivity of these networks and the biochemical specificity of phospho-catalytic events in both healthy and diseased states. Such comprehensive kinome resource would greatly facilitate the elu-
cidation of disease mechanisms and help identify the network components responsible for therapeutic responses.

Decades of research and recent powerful advances in biotech-
nology and systems biology are producing an ever-increasing accumulation of biological information about proteins, phos-
pho-proteins, and proteomic-level regulation of cell processes and diseases (3–10). There are over 500 databases devoted to protein sequence, structure, molecular and chemical interactions, and signaling pathways (11–14). PPI databases such as STRING (15), PINA (16), MINT and IntAct (17), BioGRID (18), HPRD (19), IMEx (13), and numerous others collect and curate, to varying degrees, genetic and protein interaction data from numerous organisms. One limitation of these resources is that they merge known and predicted protein interactions, combine physical with indirect genomic interactions, and include data inferred from evolutionary conservation analysis. Moreover, PPI databases can contain several million interactions, but only about 50 thousand of these are found in human and even fewer relate to phosphorylation events.

1Department of Laboratory Medicine, Helen Diller Family Comprehen-
sive Cancer Center, University of California San Francisco, California.
2The State Key Laboratory of Genetic Engineering, Ministry of Edu-
cation (MOE) Key Laboratory of Contemporary Anthropology, Col-
laborative Innovation Center of Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

A. Olow and Z. Chen share first authorship of the article.

Corresponding Author: Jean-Philippe Coppé, University of California, San Francisco, 2340 Sutter Street, S441, San Francisco, CA 94115. Phone: 415-
846-7078; Fax: 415-502-2773; E-mail: Jean-Philippe.Coppe@ucsf.edu
doi: 10.1158/0008-5472.CAN-15-2325-T
©2016 American Association for Cancer Research.
Several major efforts integrate publicly available kinase–substrate phosphorylation data. A majority of these resources, such as KSD (20), KinBase (21), KinG (22), and Kinweb (23), provide valuable information about kinases, such as sequence alignment, phylogeny, or regulatory domains. These databases, however, often lack verified information on target residue sites within substrate proteins phosphorylated by kinases. Novel resources have begun to map the relationship between kinases and substrates. For example, PhosphoSitePlus (24), PhosphoPOINT (25), RegPhos 2.0 (26), or Phospho.ELM (27) provide information on phosphorylation events garnered from literature and derived from substrate motif scans and high-throughput mass spectrometry (MS) approaches. Although such resources are comprehensive, individual results may be compromised by integration systems that rely on highly automated computational pipelines. It often remains difficult for non-expert users to distinguish experimentally validated knowledge from potentially less reliable high-dimensional screens or prediction-based datasets. Finding information about the precise connectivity between a human kinase and its cognate substrates and specific target sites, or the reverse, can be cumbersome. Access to a well-characterized and experimentally validated, albeit smaller, dataset describing established phosphorylation circuits would enable efficient and reliable exploration of the molecular mechanisms of human diseases and generate new hypotheses.

Here, we present PhosphoAtlas, a highly curated map of the human phospho-reactome and its mutated, tumor kinase counterpart. We define the phospho-reactome as a complete catalog of verified phosphorylation reactions involving an effector/kinase and a receiver/substrate protein, and their network-level connectivity, including details of phosphorylatable residue sites and heptameric target sequences when known. This dataset was developed via a series of computational steps designed to extract protein information from multiple data sources and identify validated kinase–substrate interactions. We then utilized the PhosphoAtlas database to explore the mutational variability of phospho-protein networks in tumors. We identify nonsynonymous mutations that most directly and systematically affect the catalytic circuits of human kinase–substrate networks, potentially revealing which somatic mutations in cancer may have functional consequences and thus be therapeutically actionable.

Materials and Methods

Computational strategy to build PhosphoAtlas relational database

The Supplementary Extended Method section describes the computational approach for building a database where human protein kinases are associated with their downstream protein targets and exact heptameric peptide sequence (HPS) they phosphorylate.

Identification and analysis of variants in peptide target sequences

PhosphoAtlas’ compendium of HPR’s was analyzed using the Catalogue of Somatic Mutations (COSMIC) database (28). First, to establish the presence of reported variants within HPRs, their genomic locations were calculated using the GENCODE GRCh37/NCBI19 assembly (29) for all known substrate protein isoforms. Next, the variants within these locations were found by overlapping the variant GRCh37 genomic locations acquired from a complete COSMIC database with the genomic locations of our HPRs. The COSMIC database also provided additional information on type of the mutation, its consequence, and the histology associated with the particular alteration. The tools used for the database build and analysis included Bedtools, R programminglanguage (Vienna, Austria, http://www.R-project.org/), and R packages GenomicFeatures, Biostatistics, and RWebLogo.

PhosphoAtlas database access

The PhosphoAtlas dataset is a resource readily available to the public at http://cancer.ucsf.edu/phosphoatlas upon registration. Users may download a ZIP archive that includes CSV and XLS files of the database, and a pre-build Cytoscape CYS session file (Cytoscape version 3). Cytoscape (30) is a very well-adapted tool to visualize and explore the networks of kinase–substrate–residue target sites and their tumor variants. Cytoscape provides sophisticated means to search and filter the PhosphoAtlas dataset and access all nodes and edge attributes, and respective relationships for either the entire network or user-selected subnetworks. Once the CSV/XLS/CYS data files are respectively loaded in Excel or Cytoscape, data can be filtered and exported in a variety of additional formats that support further processing in other network analysis and statistics packages. All files can be used to study or query either normal phospho-circuits or tumor-associated networks.

Results

Strategy to establish a comprehensive map of the validated human phospho-reactome

We created a resource for the exploration of human phosphorylation circuits and its aberrant cancer variations. The flow diagram presented in Fig. 1A depicts the key steps of the processing pipeline we used to acquire, integrate, filter, curate, and analyze preexisting molecular data from publicly accessible resources (see Supplementary Extended Methods and Supplementary Fig. S1). This strategy enabled identifying experimentally verified kinase proteins, protein substrates, phospho-residue sites, and HPSs.

Integration of multiple data sources into one harmonized protein repository

First, we generated a ‘Protein Reference Index’ from HGNC and NCBI/Entrez to create a nonredundant inventory of human proteins defined by their standardized symbols, names, RefSeq protein and nucleotide IDs, which established a reliable ‘Primary Identifier’ record for each protein. Next, the Protein Reference Index served as a blueprint to systematically integrate and cross-reference the protein records extracted from other "external" data sources using a curation method analogous to the PPI integration pipeline from ref. 16 (Fig. 1A, left). Directly matching external records were imported and structured by their common, unique primary identifier. If no matching primary identifier was available in the record, identifiers from external sources were cross-referenced in curation 1 with the UniProtKB database (Supplementary Extended Methods; ref. 31). Once an external identifier was successfully cross-referenced, the related record was updated. Unmatched data were excluded. At each step of the assembly of the data, all additional available functional data, annotations and references were compiled as complementary content linked to their primary identifier.
PhosphoAtlas: A Resource to Explore the Tumor Kinome

To harmonize the heterogeneous content across the various sources of this initial human proteome resource, collected records were curated (curation 2). Pattern matching computer-implemented methods and sequence alignment algorithms was used to check, remove, or merge any redundant or ambiguous records, and flag discrepancies, as described in detail in Supplementary Extended Methods. This resulted in a highly curated, nonredundant, comprehensive dataset of known human proteins, referred to as the "Harmonized Proteome Index."

Building a relational database connecting kinases to substrates and phospho-residue sites

Next, we attempted to identify the most complete and rigorously verified phospho-reactome portion of the human proteome. We applied data-mining methods to extract protein records related to phosphorylation events to identify kinase-substrate catalytic interactions, and phosphorylatable amino acids from substrates’ phospho-residue sites. Records from the Harmonized Proteome Index were queried using natural language processing algorithms to identify proteins that function as kinases or phosphorylatable substrates, as described in detail in Supplementary Extended Methods using a curation method analogous to that in ref. 24 (Fig. 1A, Functional triage; Supplementary Fig. S1, bottom-left section).

Qualified records generated a preliminary repository of functionally interconnected kinase and substrate proteins that was then mined for phosphorylation modification information (Fig. 1A and Supplementary Fig. S1, curation 3). Each qualified substrate with an available phosphorylation (abbreviated as "phospho")-site or -peptide sequence was compared with the latest corresponding curated protein substrate sequence (RefSeq/NCBI) using string- and pattern-matching methods, as described in detail in Supplementary Extended Methods. Once the location of a phosphorylated residue was confirmed, or if the sequence alignment or peptide composition was conclusive, the amino acid sequence surrounding the validated phospho-site was extracted from the protein RefSeq. We collected HPSs around phospho-residue target sites based on previous biochemical observations (32, 33) and our data (Supplementary Fig. S3d). Both residue site location and heptameric peptide were indexed and assigned to their substrate (and kinases when reported; Fig. 1B). Because the repository of phospho-catalytic sites represents the critical cornerstone of all kinase–substrate functional connections, we rigorously ensured its accuracy by excluding all primary records providing candidate targets from data solely based on confidence-based approaches or not cross-referenced or not confirmed using complementary molecular techniques. We further distinguished validated phospho-residue sites with or without an identified upstream kinase, and subclassified them separately (Fig. 1B, two right groups).

The resulting database exclusively contains validated human kinases and substrates and their respective phosphorylation sites.
These interactions were integrated into a network of kinase–substrate phospho-catalytic circuits that constitute “PhosphoAtlas” (Fig. 1B and C). Flat database files and a Cytoscape session are publicly available via the web portal http://cancer.ucsf.edu/phosphoatlas to facilitate exploration and visualization of molecular networks (see Materials and Methods).

Overview of the human phospho-reactome
Analysis of PhosphoAtlas suggests that ~11% of all PPIs (Supplementary Fig. S2; ref. 16) represent potentially actionable phospho-catalytic circuits of human cells. A total of 4,758 unique edges connect kinases to a target. PhosphoAtlas catalogs this network as 3,641 kinases with known substrates and residue target sites, and 1,117 with known substrates only. A total of 292 kinases are reported as exclusively or redundantly phosphorylating 1,276 distinct substrate proteins via 2,617 unique heptameric peptide regions (HPR) that correspond to 2,492 distinct HPSs, indicating that some HPSs are "shared" across multiple HPRs (Fig. 2A).

The connectivity between kinases, substrates, and HPSs (Fig. 1B and C) can be categorized based on the number of cognate partners. For example, 300 kinases phosphorylate at least two substrates, and 220 kinases phosphorylate at least two HPSs (Fig. 2B, left; Supplementary Fig. S3). Most highly connected kinases and substrates are shown in Supplementary Fig. S4. Of the 1,276 substrates with identified kinase partners, 53% (n = 671) interact with a single unique kinase, while the remaining 47% (n = 605) interact with at least two kinases. Out of 1,276 substrates, 913 contain either one (29%, n = 363) or at least two (43%, n = 364) unique HPSs. And 39.9% of all substrates in PhosphoAtlas contain only one unique HPS. Of the verified 2,492 HPSs, 74% (n = 1,835) connect with a single unique kinase, and 26% (n = 667) are shared by at least two kinases. In contrast, 98% (n = 2,440) of the HPSs are associated with a single substrate, while the remaining 2% are found across multiple substrates. We also found that 58.9% of all kinase–substrate phosphorylation events are achieved via a unique peptide sequence; 24.7% of all kinases in PhosphoAtlas target one unique HPS; and 41.8% of all kinases in PhosphoAtlas target one unique substrate. Taken together, these data portray the human phospho-reactome as almost evenly divided between highly specific kinase–substrate pairs and more highly connected nodes. Combinatorial logic of signaling, whereby multiple kinases can phosphorylate a single substrate, is mainly achieved through multiple distinct HSPs on a single substrate (60.1%), while as few as 26.4% HSPs are recognized by multiple kinases. Only six kinase-specific HPSs are found on multiple substrates.

Grouping kinases by gene families (21) shows that 84% of validated substrates and 73% of validated peptides associate with either the AGC (e.g., PKA, PKC; 34.1% of substrates, 26.8% of HPSs; Table 1). The connectivity of kinases, substrates, and HPSs (Fig. 2B and C) can be categorized based on the number of cognate partners. For example, 300 kinases phosphorylate at least two substrates, and 220 kinases phosphorylate at least two HPSs (Fig. 2B, left; Supplementary Fig. S3). Most highly connected kinases and substrates are shown in Supplementary Fig. S4.
peptides), CMGC (e.g., MAPK, CDK; 27.4% substrates, 24.3% peptides) or TK (e.g., EGFR, SRC; 22.3% substrates, 22.1% peptides) kinase families (Fig. 2C and Supplementary Fig. S5). The kinase interaction edges for each kinase family can be categorized as fully validated (known kinase, substrate and target site) or partially known (known kinase and substrate only; Fig. 2C, gray bars).

Our curation process also revealed valuable knowledge gaps (Fig. 1B, 4 right groups). The database includes 275 kinase enzymes known to be upstream of 363 substrates, but for which no phospho-residue site was found or definitively established, and 277 identified kinases with no validated substrate. Reciprocally, 632 substrate proteins with 706 confirmed that HPSs have no conclusively established effector kinase, and 74 substrates have no phospho-residue site was found or de

B, Circos plots visualize the number of kinase–substrate interactions solely as 1,276 pairs masks the inherent complexity of phospho-catalytic circuits. We further resolved their connectivity using the available 3,641 phospho-residue site-specific distinct interactions, which yielded 4,758 unique kinase–substrate phospho-catalytic connections. Accordingly, Fig. 3 shows how incorporating phospho-residue sites increases the resolution of kinase–substrate catalytic interaction networks. Figure 3A and B displays kinases according to the number of identified substrates for each kinase and its number of HPs per substrate. There is a continuum from highly substrate-specific kinases with multiple HPS targets, such as MET (9 HPS on 1 substrate; Fig. 3A, top left; Fig. 3B, bottom), to broad-spectrum kinases that phosphorylate 10s to >100 individual substrates at an average of 1 to 2 HPS targets, such as PRKCA (210 HPS on 109 substrates, averaging 1.9 HSP per substrate; Fig. 3A, right; 3B, top). Most interaction networks are simpler; 91.4% kinases are

Target peptides provide insight into the diverse modes of kinase–substrate interactions

Target peptides provide insight into the diverse modes of kinase–substrate interactions. E, HPSs (green discs) are plotted based on the number of unique interacting kinases (y-axis) and substrate per kinase ratio (x-axis). All peptides in the plot are N-term to C-term amino acid sequence annotated in parentheses with the number of substrates it is found within per number of kinases that it is targeted by. F, Circos plots visualize the number of kinase–substrate and kinase–peptide–substrate connections for different entries.

Figure 3.

Biological peptide sequences collected in PhosphoAtlas provide a new depth and unique dimension into kinase–substrate catalytic interactions. A, individual kinases (red discs) are plotted based on the number of unique connecting substrates (x-axis) and related HPS target per substrate ratio (y-axis). Some kinases are annotated by name followed by the number of known HPS targets within it per number of substrates. B, Circos plots visualize the number of kinase–substrate (left) and kinase–HPS–substrate (right) connections for selected kinases. C, protein substrates (blue discs) are plotted based on the number of unique connecting kinases and peptide per kinase ratio. Annotated substrates are identified by name followed by the number of known HPS targets within it per number of kinases. D, Circos plots visualize the number of kinase–substrate (left) and kinase–peptide–substrate (left) connections for selected examples of substrates. E, HPSs (green discs) are plotted based on the number of unique interacting kinases (x-axis) and substrate per kinase ratio (y-axis). Every peptide in the plot is an N-term to C-term amino acid sequence annotated in parentheses with the number of substrates it is found within per number of kinases that it is targeted by. F, Circos plots visualize the number of kinase–substrate and kinase–peptide–substrate connections for different entries.
phosphorylate between 1 and 20 substrates, with a median of 2 unique HPS targets per substrate (mean = 2.804 HPS, SD = 2.813). Figure 3C and D represents a similar spectrum for protein substrates and their connections to kinases and identified HPS targets. At the extreme ends are substrates phosphorylated by multiple kinases at mostly unique phospho-sites on these substrates (e.g., IRS1 (phosphorylated by 13 kinases on 28 HPS, average 2.1 HPS per kinase), MAPT (34 HPS/14 kinases)), and substrates known to be targeted by few kinases, yet phosphorylation occurs at many target sites (e.g., RET (12 HPS/1 kinase). Similarly, when examining modes of action on HPS targets, there emerge sequences that are found in many substrates, but phosphorylated by single kinase, and those that are recognized by multiple kinases, yet are found only in few substrate proteins (Fig. 3E and F).

Ultimately, this increased resolution facilitates a more detailed understanding of regulatory signaling mechanisms, especially considering that many substrates have more than one phospho-residue site that are specifically and differentially phosphorylated by many kinases.

Exploring AKT cancer phospho-circuits using PhosphoAtlas

PhosphoAtlas supports the exploration of the human kinome at a global phospho-proteomic level, but also at the level of individual kinase proteins and their subnetworks, interactive partners, enzymatic modalities, and conservation patterns. For example, Fig. 4A represents the network of kinases and substrates directly upstream and downstream of the AKT1, AKT2, and AKT3 proteins. This subnetwork includes 102 kinase and substrate proteins (red and blue nodes, respectively) that interact via 184 unique kinase–substrate edges (all lines), of which 113 have known phospho-residue site information (green lines: Fig. 4A). Phospho-signaling cascades that funnel through AKTs can further propagate via downstream kinases (Fig. 4A, bottom left red) or affect non-kinase substrate proteins (Fig. 4A, bottom right blue) that regulate normal and tumor-associated cell processes such as growth, motility, transcription, or inflammation. The broad range of disease ontology terms associated with these 102 proteins and mainly associated with cancer demonstrates the diverse functional impact of AKT-related signals (Fig. 4A, pie chart; Supplementary Fig. S6; data mined using Ensembl, Biomart).

Considering the number of additional upstream kinases and distinct residue sites that exist per protein composition (AKT circuits) numbers specified within green and red squares above and below each protein, Fig. 4A), reinforces the complexity of AKT circuitry in particular and phospho-circuitry in general, and illustrates the utility of PhosphoAtlas for visualization and exploration.

Differences among the three AKT family members in their number of kinase–substrate edges with or without known phosphorylation sites are apparent (Fig. 4A and Supplementary Fig. S7). This in part reflects that, while AKT1 has been thoroughly studied, much less is known about AKT2, AKT3, and their targets. Such differences are repeatedly found across kinase subfamilies (e.g., CDKs or SRCs).

PhosphoAtlas also enables the identification of consensus peptide target sequences that could be used to scan the human proteome to predict new substrate phosphorylatable sites, or design probes for kinase activity assays. For instance, by comparing all HPS targets of AKT1, the identified the most conserved amino acid residues within heptameric peptides (Fig. 4B and Supplementary Fig. S8), which is supported by other studies. Beyond the required presence of a phosphorylatable Serine (S) or Threonine (T) at the center of all HPSs, the Arginine (R) at position –3 was highly conserved with a frequency of 83.2%. This result highlights the biochemical importance of Arginine’s presence at this site for AKT1 to phosphorylate its substrates.

Given that much of the human kinome is considered druggable, it is of interest to identify known (pathogenic and/or somatic) variants within each kinase subfamily. This represents a new group of 2,896 mutated sequences (Fig. 5C), which perturb 1,963 kinase–substrate catalytic interactions, with functional consequences on kinase circuits. PPI-perturbing mutations are significantly more likely to be deleterious than non-PPI-perturbing mutations (10, 34). To identify cancer mutations that directly impact kinase–substrate catalytic interactions, we designed a reverse-mining process that maps genetic variations onto phosphorylatable substrate regions (Fig. 5A; see Materials and Methods for details). We analyzed our compendium of HPSs using the COSMIC database (28), which includes curated information from The Cancer Genome Atlas (TCGA) project, the International Cancer Genome Consortium, and systematic screens of 1,000s of tumor genomes. We mapped the mutational landscape of cancer phospho-circuits in a tissue- or disease-agnostic fashion.

Protein variants identified through this process were classified based on the consequence of genetic alterations (Fig. 5B). Of the original 2,617 distinct substrates’ HPRs, 39.4% (n = 1,031) were not affected by any known mutation, 8.3% (n = 217) contained only silent variants (synonymous), and 52.3% (n = 1,369) contained at least one nonsynonymous variant that altered amino acid composition (nHPRv). The distinct variants generated a new group of 2,896 mutated sequences (Fig. 5C), which perturb 1,963 kinase–substrate interactions (i.e., 46.7% of all kinase circuits) with potential detrimental effects on phospho-catalytic circuits (10).

Among all HPRs affected by 3,703 synonymous and nonsynonymous variations, missense substitutions account for the majority (69.1%) of all variants, followed by 17.8% of coding but silent substitutions (Fig. 5D). The differential levels and pattern of co-occurrence of nonsynonymous versus synonymous mutations of HPRs (Fig. 5E and Supplementary Fig. S9c) imply that phospho-target sites are mutated in a selective fashion in tumors. Only 5% of all HPR variants originate from mutations qualified as germline single-nucleotide polymorphisms (Supplementary Fig. S10). Adenocarcinomas are the most prevalent pathology (20%) associated with all variants (Fig. 5F), while the highest number (14.4%) of phosphorylation-site variants originates from tumors of the large intestine (Supplementary Fig. S11).

We next proceeded to identify which substrates are most affected by cancer-associated mutations that alter their
phospho-targetability. To do so, each protein was plotted based on its individual ratio of mean number of variants per HPR (Fig. 5G, y-axis), against its total number of HPR (x-axis). Each dot, representing a protein, was subsequently color-coded as the mean percentage of nsHPRs per HPR, and averaged per substrate, to indicate whether genomic mutations mainly caused synonymous variants or nonsynonymous sequence alterations (nonsynonymous high%, red; low%, blue; Fig. 5G and Supplementary Fig. S9d and S9e). For instance, FLT3 contains few phosphorylatable HPRs associated with many cancer mutations that mostly result in alternative peptide sequences (high% nonsynonymous variants). Conversely, IRS1 display numerous HPRs that are overall subject to few variants, of which ~50% cause amino acid changes. In this spectrum, TP53 emerges as a substrate with a very high mean ratio of variants per HPR, associated with a large number of known
Figure 5.
Impact of nonsynonymous variants identified as genetic mutations found in tumors. A, reverse mapping approach to reveal the cancer-associated variants embedded within the phospho-reactome. The identification of cancer-associated genetic mutations as reported by COSMIC that exert specific effects on kinase–substrate catalytic interaction sites exposes an additional layer of disturbances caused by mutations found in patients’ tumors. B, bar graph depicting the number of HPRs in PhosphoAtlas that contain COSMIC-defined synonymous and nonsynonymous variants, or no mutation at all. The number of unique (kinase–HPR–substrate) entries affected by cancer mutations is shown on top. C, total number of cancer-associated, nsHPR derived from COSMIC mining analysis. The number of potential additional cancer-related predicted PPI entries is indicated on top. D, distribution of the protein-coding consequences of genetic mutations on phosphorylatable HPRs. E, distribution of co-occurrence of nonsynonymous versus synonymous variants with the highest proportion of nonsynonymous variants to total number of variants per HPR were selected, and finally all HPRs from each of these substrates were plotted. Proteins are sorted alphabetically on the x-axis. Each HPR is shown as a triangle and color-coded by the mean percentage of nonsynonymous variants out of total reported nonsynonymous and synonymous variants per HPR (color scale in G).
target sites and a high proportion of variants being reported as nonsynonymous (80.7%).

This analysis also identified proteins for which the COSMIC database does not report any variant in their phosphorylatable regions (Fig. 5G, separated bottom panel; Supplementary Fig. S9d and S9e). Out of these 214 proteins, IKBKB, CREB1, and STAT1 were especially interesting examples due to their known critical role in regulating phosphorylation cascades involved in cancer development.

When using HPRs as a way to report which kinases would be most affected by downstream nsHPRvs in substrates, the spectrum of kinases with most-to-least variable networks emerged (Supplementary Fig. S12). This highlighted kinases that are potentially more frequently affected by nonsynonymous mutations in their target sites (e.g., CHUK, GSK3A, or PRKDC), and kinases with genomically stable networks of downstream targets (e.g., MAP3K14 and EE2F2).

Next, we investigated the 50 HPRs with the highest number of reported unique nonsynonymous variants. These most diversely altered sequences identified 29 substrate proteins only (Fig. 5H). For example, TAPSLSG is a region from CTNNB1 that has the highest number of total variants per HPR, out of which 84.6% are amino acid coding, nonsynonymous alterations. Looking at individual proteins, we find not that all target sites show the same mutational susceptibility. Many of the listed proteins are known tumor suppressors or oncogenes, while some are recognized enzymes, including kinases or phosphatases, showing how mutations in their phospho-target sites could globally affect signaling cascades. For instance, target sites characterized by very high proportions of nsHPRvs were found in BRAF, EGFR, or FLT3, all of which contain activating kinase domain mutations that result in malignancies such as BRAFV600E in melanoma, or EGFRL858R in lung adenocarcinoma.

Subsets of distinct mutations may concentrate in pathways implementing cancer hallmark phenotypes (35). As mutations and resulting perturbations in kinase networks are key to cancer biology, we generated a network representation overlaying nonsynonymous heptameric peptide region variants (nsHPRvs), onto the phospho-reactome interaction map using Cytoscape (Fig. 6 and Supplementary Figs. S13–S17; purple lines; data file available at http://cancer.ucsf.edu/phosphoatlas upon registration; ref. 30). This network represents all distinct phospho variants in COSMIC associated with tumors of every type. As illustrated, kinase circuits most selectively disrupted (defined by width of purple lines and node margins) across tumor types emerge out of the background of predominantly invariant phospho-circuits; green lines). This mosaic of normal and aberrant catalytic interactions shows how tumors systematically hijack—and spare—particular subnetworks.

Discussion

Phosphorylation cascades intimately regulate cell behavior. The availability of a highly curated compendium of phosphorylation data accurately recapitulating the molecular wiring of cells could support elucidation of disease mechanisms and vulnerabilities to therapeutic intervention in cancer. We built PhosphoAtlas, a novel data resource to study human phosphorylation networks, which includes a map of the differential impact of tumor mutations on phospho-target sites and reveals the most variable phospho-signaling hubs of cancer. PhosphoAtlas can support the exploration of a wide range of research questions related to human phospho-circuitry and is meant to serve as a research hypothesis-generating platform.

To develop PhosphoAtlas, we systematically mined the profusion of available data existing in various formats and stages of verification or completion, found in a large number of state-of-the-art databases. We applied a stringent condensation-like process to filter, select, eliminate, cross-reference, qualify, and validate information related to protein phosphorylation in human cells. We did not use prediction algorithms, conditional selection, or inferred relationships, but rather relied on mining published data generated and validated by diverse experimental and computational techniques (16–18, 24). Although this approach restricted the number of kinase–substrate interactions represented in PhosphoAtlas, due to the exclusion of unvalidated phospho-catalytic interactions, it is meant to provide researchers with a still-large catalog of protein kinases, substrates, phosphorylatable sites, and their connectivities from only the most reliable, validated sources. The result of this effort is both a data resource and a coherent map of functionally interconnected proteins comprising the human phospho-reactome (http://cancer.ucsf.edu/phosphoatlas).

The content of PhosphoAtlas can be exploited for at least three broad purposes: integrative in silico analysis and visualization; experimental data analysis; and bioassay development. For all these purposes, a major contribution of PhosphoAtlas is its inclusion of information on the specific phospho-residue sites of substrates, which enables investigation of kinase–substrate interactions networks at a high level of resolution. As visualized and cataloged using PhosphoAtlas, the human phospho-reactome network appears to be roughly evenly divided between highly specific kinase–substrate pairs (33.2% proteins with single kinase–substrate connection, 19.1% proteins with two kinase–substrate connections), and more highly connected protein nodes (47.7% proteins with three or more kinase–substrate connections).

We identified kinases with preference to target many substrates via unique peptide target sequence versus those that target fewer substrates but can recognize a larger number of peptide target sequences. Such findings suggest, by the sheer number of alternative ways a kinase can phosphorylate a substrate, the essentiality of certain kinase–substrate interactions; in the event of a disruptive mutation in one region, a protein is likely to remain functional due to compensatory rerouting mechanisms. For example, the phosphorylation of GAB1 by MET kinase on nine different target sites suggests that this process may be essential for cell homeostasis. Moreover, kinases with the ability to phosphorylate many substrates on different target peptide sites may be active in multiple networks that regulate functionally similar cellular processes. If such a kinase’s first intended peptide target becomes disabled due to a mutation, a wider range of feedback mechanisms may be expected. Among such proteins, PRKACA can recognize and phosphorylate as many as 194 unique peptide sequences on 121 different substrates with prominent important functions in both the cytoplasm and the nucleus of cells. The diversity of connectedness of peptide targets at the regulatory interface of kinases and substrates emphasizes how the wide spectrum of network topologies of each substrate or kinase relates to their distinctive roles and functional impact on phospho-circuits.
Olow et al.

**Figure 6.**
The tumor kinome: differential impact of cancer-associated mutations on kinase–substrate phospho-catalytic networks. Data generated from cross-referencing PhosphoAtlas and COSMIC databases were used to render the differential impact—and potential convergence—of nonsynonymous peptide target variants onto substrates of kinases. The complex heterogeneity of mutations across tumor genomes was resolved via the unbiased aggregation of mutations that only and selectively alter substrates’ HPRs. Purple connectors represent nsHPRVs at the catalytic interface of kinases and substrates that are collapsed into more or less thick edge representing the prevalence of mutations occurring within one heptameric sequence. By overlaying both interconnectedness and mutational impact of tumors on the human phospho-reactome, the most variable nodes visually emerge from the otherwise rarely altered or non-mutated network. The absence of any reported mutation for a number of prevalent nodes with a high number of interactions is noticeable, such as CDK1, PRKACA, MAPK3, MAPK14, GSK3B, AKT3, or CSNK2A2. The phosphorylation networks were produced using the Cytoscape network analysis platform (30) and is available at http://cancer.ucsf.edu/phosphoatlas.
mutations across different post-translational modifications (PTMVar). Investigating the prevalence of these mutations and integrating them into comprehensive signaling networks is still needed to reveal the strategic, therapeutically actionable kinase circuits of tumors. Identifying which cancer mutations functionally impair biological networks remains a challenge, even when using large-scale differential mapping approaches (10, 36–40). Conceptually, the sparse and relatively nonrecurrent mutation profiles within and across tumor types may be more comprehensible at the pathway level than at the level of individual genes (35, 41–47). Mutations that directly perturb protein phosphorylation circuits may be especially pathologically relevant to cancer development and eventually guide targeted therapy.

Thus motivated, we used the heptameric peptide library created from our curated kinase–substrate repository as a functional blueprint to filter 100,000s of candidate somatic mutations found across 1,000s of different human tumors. We found that over half (54.9%) of the distinct target peptides (2,617) contain at least one nonsynonymous variant with altered amino acid composition. We identified which proteins are most affected by nonsynonymous variants in phosphorylatable peptide regions (Fig. 5G). We depicted the reported variant load for 29 proteins that are most affected by nonsynonymous alterations in the target peptide regions (Fig. 5H). Cancer phospho-circuits revealed widespread perturbations across the phospho-reactome, but also recurrent aberrations that selectively affect specific phospho-hubs, including both proto-oncogenes and tumor suppressor proteins. The diversity of subnetworks connectivity combined with the differential mutational status of target regions cataloged in PhosphoAtlas is a roadmap to explore the strategic kinase hotspots of cancer.

Because PhosphoAtlas is meant to be a hypothesis-generating platform, users may, for example, investigate the potential relationship between the impact of oncogenic mutations on kinases and their local signaling topology. For instance, upon examining the number of distinct variants of phosphorylation sites per kinase, users may notice that some of the kinases with a very high endogenous nsHPRv mutational load phosphorylate a relatively low number of substrates, and thus seem to have a low level of signaling pleiotropy (e.g., FLT3, KIT, KDR, PDGFRB, BRAF/RAF1; Supplementary Fig. S15). It is tempting to speculate that such kinases may be especially suitable targets for tumorigenesis, possibly not just because they achieve a particular advantageous hallmark of cancer but also because they minimize "collateral damage," which may otherwise lower the fitness of an emerging cancer cell. This speculation seems consistent with observations of the other end of the spectrum, whereby many kinases with a very large number of downstream substrates seem to show relatively few or no nonsynonymous variants in their phospho-sites (e.g., PRKCA/G/Z, PRKACA, MAPK1/3/4, CDK1/2/5, GSK3A/B, ATM/ATR, SRC/FYN/lyn; Supplementary Fig. S16). Systematic exploration of such hypothetical relationships could be valuable for understanding tumorigenic mechanisms and brainstorming new therapeutic strategies. More globally, modeling the cancer kinome by weighing in second-degree connectivity of upstream/downstream proteins, and the differential susceptibility of specific tumor tissues to particular nsHPRv mutations, may help identify likely driver mutations and/or functional dependencies of cancer. As such, PhosphoAtlas displays the distinctive cancerous perturbations on normal-phospho-catalytic networks with the resolution necessary to begin understanding the differential functional impact and integrative consequences of relevant tumor mutations on signaling networks.

Despite the many tumor genomes reported in COSMIC, we found that for a majority of kinase–substrate catalytic interactions no variants were reported (2,238 versus 1,963; Fig. 5B). Furthermore, there are significantly fewer mutational variants per amino acid in HPRs (mean = 0.20) than in the entire protein sequence of their substrate of origin and across all protein substrates cataloged in PhosphoAtlas database and recorded in COSMIC (mean = 0.36; mean difference [95%CI] = –0.16 variants/amino acid [–0.20–0.11]; t test P = 1.62e–13). Although it is possible that mutations in such phosphorylation sites could be fatal to most cells, these observations inspire the question: are non-mutated target sites not altered because they do not provide any selective advantage to tumors, or because their mutation would be so detrimental that cancer cells could not survive? If the latter, and if the node is not essential to normal cells, these observations could point to potential sources of synthetic lethality that may be exploited to design new, safe combinatorial targeted therapies. More generally, it will be important to investigate whether the spectrum of hyper-mutable-to-never-altered phospho-hubs can predict which oncogenic aberrations are most likely to be actionable in tumors.

The other two broad purposes of PhosphoAtlas, which we do not demonstrate in this study, are to assist proteomic data analysis and to support bioassay development and optimization. Projecting data obtained from RPMA/RPPA assays (3, 8) or MS approaches (5–7, 48–53) onto the curated map of kinase–substrate networks may improve assessment of which phospho-circuits are active in biological samples to identify adaptive reprogramming networks associated with cancer response to therapeutic interventions. Information in PhosphoAtlas, especially that from the phospho-residue site data, may also be leveraged for developing and optimizing epitope mapping for (phospho) antibody development, protein engineering, targeted mutagenic and genetic screens, synthetic phage display/yeast-two-hybrid systems, enzymatic assays to test inhibitor libraries, or utilize HPRs as barcodes of kinase–substrate pairs to infer levels of activity of relevant kinases in MS studies. We are currently in the process of implementing a high-throughput kinase activity-mapping platform that utilizes biological peptide sequences as experimental sensors to profile the activity of phosphorylation circuits in biological extracts.

In conclusion, the PhosphoAtlas database provides a rich resource for the exploration of the human kinome and its tumor variant and supports a range of applications from the functional mapping of modular networks for systems biology, to assisting in the development of physical or virtual interfaces for oncology and other diseases.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**Authors’ Contributions**

Conception and design: A. Olow, Z. Chen, R.H. Niedner, L.J. van ‘t Veer, J.-P. Coppé


Published OnlineFirst February 26, 2016; DOI: 10.1158/0008-5472.CAN-15-2325-T
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): A. Olow, L.J. van ’t Veer, J.-P. Coppé

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Olow, Z. Chen, R.H. Niedner, D.M. Wolf, C. Yau, A. Pankov, L.J. van ’t Veer, J.-P. Coppé

Writing, review, and/or revision of the manuscript: A. Olow, Z. Chen, R.H. Niedner, E.P.R. Lee, L. Brown-Swigart, L.J. van ’t Veer, J.-P. Coppé

Administrative, technical, or material support (i.e., reporting or organizing data, providing facilities, etc.): A. Olow, Z. Chen, D.M. Wolf, C. Yau, L. Brown-Swigart, L.J. van ’t Veer, J.-P. Coppé


Study supervision: L.J. van ’t Veer, J.-P. Coppé

Acknowledgments

The authors thank computational scientists, engineers, and specialists Erin Hayes, Chong Huang, Liyun Luo, Brent Nix, Xuqun Pan, Neil Park, Alexandra Solomon, Zoe Tribur, Changjun Wang, Kevin Wang, Amanda Wu, Zaidao Yu, Elad Ziv, and support teams of programmers and curators.

Grant Support

This work was supported by grants from the Friends for an Earlier Breast Cancer Treatment, UCSF Breast Oncology Program Developmental Research Program, T1 Catalyst Award (U11 TR000004), NIH UCSF SPORE project 2, National Institute of General Medical Sciences T32GM067547, the National Science Foundation 1144247, a gift from the “Atwaters,” and the Angela and Shu Kai Chan Endowed chair.

Received August 25, 2015; revised December 10, 2015; accepted January 4, 2016; published OnlineFirst February 6, 2016.

References


An Atlas of the Human Kinome Reveals the Mutational Landscape Underlying Dysregulated Phosphorylation Cascades in Cancer

Aleksandra Olow, Zhongzhong Chen, R. Hannes Niedner, et al.

Cancer Res 2016;76:1733-1745. Published OnlineFirst February 26, 2016.

Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-15-2325-T

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2016/04/13/0008-5472.CAN-15-2325-T.DC1

Cited articles
This article cites 53 articles, 5 of which you can access for free at:
http://cancerres.aacrjournals.org/content/76/7/1733.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.