Protein phosphorylation is a major regulator of protein function and biological outcomes. This was first recognized through functional biochemical experiments, and in the past decade, major technological advances in mass spectrometry have enabled the study of protein phosphorylation on a global scale. This rapidly growing field of phosphoproteomics has revealed that more than 100,000 distinct phosphorylation events occur in human cells, which likely affect the function of every protein. Phosphoproteomics has improved the understanding of the function of even the most well-characterized protein kinases by revealing new downstream substrates and biology. However, current biochemical and bioinformatic approaches have only identified kinases for less than 5% of the phosphoproteome, and functional assignments of phosphosites are almost negligible. Notably, our understanding of the relationship between kinases and their substrates follows a power law distribution, with almost 90% of phosphorylation sites currently assigned to the top 20% of kinases. In addition, more than 150 kinases do not have a single known substrate. Despite a small group of kinases dominating biomedical research, the number of substrates assigned to a kinase does not correlate with disease relevance as determined by pathogenic human mutation prevalence and mouse model phenotypes. Improving our understanding of the substrates targeted by all kinases and functionally annotating the phosphoproteome will be broadly beneficial. Advances in phosphoproteomics technologies, combined with functional screening approaches, should make it feasible to illuminate the connectivity and functionality of the entire phosphoproteome, providing enormous opportunities for discovering new biology, therapeutic targets, and possibly diagnostics.

Reversible protein phosphorylation is a major posttranslational modification (PTM) for signal transduction that impinges on virtually all biological functions (1). Phosphorylation involves the kinase-catalyzed transfer of γ-phosphate from adenosine 5′-triphosphate (ATP) to amino acid residues, including serine, threonine, and tyrosine, and its removal by phosphatases (2). Phosphorylation of noncanonical amino acid residues is also increasingly being recognized as a major contributor to the cellular pool of protein phosphorylation. Histidine phosphorylation in particular has been reported to be a key cellular regulatory mechanism (3), with mass spectrometry (MS)—based methods revealing that it is remarkably widespread in bacteria, comprising as much as 10% of the Escherichia coli phosphoproteome (4). Phosphorylation modulates the function of proteins, for example, by altering their activity, localization, and interactions (5). MS-based phosphoproteomics has revealed that most proteins are phosphorylated, many on multiple sites. Dysregulated phosphorylation is a hallmark of many diseases, including numerous cancers (6–8), Alzheimer’s disease (9, 10), and diabetes (11, 12). Understanding phosphorylation and its effects on regulatory networks and effector proteins is therefore a major endeavor in the post-genomics era. The number of phosphosites identified has vastly expanded in the last decade because of advances in MS. The upstream regulation and downstream function of most of these sites are unknown, providing many opportunities to interrogate the phosphoproteome for new insights into cell behavior, as well as health and disease.

**MS-BASED PHOSPHOPROTEOMICS**

MS offers numerous advantages for studying protein phosphorylation, enabling its quantitative, sensitive, and site-specific measurement on a large scale. These capabilities have given rise to the field of phosphoproteomics—the global, quantitative study of protein phosphorylation—which emerged from pioneering studies at the turn of the century (13–17). A seminal study by Olsen *et al.* (18) combined multiple stable isotope labeling by amino acids in cell culture (SILAC) experiments through a common reference sample, together with global MS-based phosphoproteomics, to measure temporal changes in phosphorylation after acute treatment with epidermal growth factor (EGF). This study revealed the widespread and dynamic nature of signaling networks and highlighted that an unexpectedly large proportion of phosphorylation sites (14% or 883 sites for EGF) were modulated by an acute stimulus within minutes. Large-scale phosphoproteomics studies performed under varied biological contexts have continued to broaden our view of the boundaries of the phosphoproteome, revealing its extraordinary scale and complexity. The image emerging is that the phosphoproteome forms highly interconnected networks, far from the textbook view of simple linear pathways. Perturbation of kinases and phosphatases therefore often affects large portions of the phosphoproteome, rather than only their direct substrates (19). Phosphoproteomics provides a powerful tool to probe these intricately entangled networks, with major implications for therapeutically targeting them to treat disease and to manipulate cell behavior.

From a technical perspective, it has been predicted that ongoing developments leading to deeper sampling of the proteome will shift MS-based proteomics from a perpetual discovery mode to a remeasurement mode (20, 21). Arguably, this shift is already underway, with deep proteome measurements now able to measure nearly all proteins in model organisms (22, 23) and well over 10,000 proteins in single human cell lines (24–27). We predict that phosphoproteomics will undergo a similar transition, albeit with a delayed onset, due to the dynamic and contextual nature of phosphorylation and the resulting expansiveness of the phosphoproteome. In time, the focus of phosphoproteomics studies will shift from quantifying and categorizing various context-dependent repertoires of phosphorylation, toward more completely understanding the nature and functional
implications of signaling networks. Innovations in MS technologies, including sample preparation workflows, instrumentation, data acquisition methods, and software, are steadily increasing the scope of phosphoproteomics studies [reviewed in (28, 29)].

Improvements to mass spectrometers in the last decade have primarily focused on sensitivity and acquisition speed (30), which have greatly benefited phosphoproteomics experiments, but dynamic range remains a major limitation for current instruments. Acquisition over a wide dynamic range is particularly relevant to deep phosphoproteomics, because the reversible and substoichiometric nature of phosphorylation extends the dynamic range of peptides emanating from the underlying proteome by several orders of magnitude. The wide dynamic range of the phosphoproteome also presents a challenge to broadly popular “data-dependent” acquisition (DDA) modes. In such modes, the mass spectrometer selects ions from full scans (MS1) and isolates them for fragmentation and identification in consecutive “top-N” MS2 scans. In DDA methods, sequentially isolating low abundant peptide species for MS2 scans places considerable demands on maximizing ion transfer efficiencies and yet paradoxically also means discarding an increasingly large fraction of incoming ions in the process. This is because, during the isolation of peptides for fragmentation, nonselected ions are discarded. Therefore, when isolating low-abundance peptides, >99% of ions can be discarded. Dynamic range challenges on the MS1 level can be alleviated by the “BoxCar” data acquisition method. BoxCar uses a greater proportion of the incoming ions by filling multiple segmented windows to increase total MS1 fill times, thereby improving the signal-to-noise and dynamic range of MS1 scans (31). Faster instrumentation is also increasing the utility of data-independent acquisition (DIA) methods [reviewed in (32)], which attempt to combine the throughput and ease of implementation of DDA methods with the quantification robustness of targeted MS workflows (33, 34). In DIA methods, the mass spectrometer cycles through predefined mass/charge segments and acquires mixed fragment ion spectra of all precursors present in each segment. DIA methods promise to reduce the problem of missing values caused by stochastic MS2 sampling between measurements. It also promises excellent quantitative precision by enabling peptide quantification from multiple highly specific MS2 fragment ions, rather than from single MS1 precursor ions that are more likely to suffer from interference from coeluting peptides. Both of these attributes are particularly appealing for phosphoproteomics experiments for which only single peptides are available for quantification. DIA-MS has thus far been applied to relatively few phosphoproteomics studies (35–37); however, once current DIA-MS limitations such as automated assignment of phosphorylation site localizations are addressed in data analysis pipelines, we expect it to become a useful approach for future studies. The choice of peptide fragmentation method also greatly influences the identification and localization of phosphopeptides. Various fragmentation techniques have been evaluated for phosphopeptide analysis including electron transfer dissociation (ETD) (38), higher-energy collisional dissociation (HCD) (39), and ultraviolet photodissociation (UVPD) (40, 41), as well as combinations thereof [such as electron-transfer/higher-energy collision dissociation (ETheCD)] (42), with HCD-type fragmentation providing the greatest performance in typical large-scale phosphoproteomics experiments (43, 44).

Data generated from large-scale MS-based phosphoproteomics studies are compiled into databases that are increasingly becoming an indispensable resource for signal-transduction research. The largest and most active database, PhosphoSitePlus, currently comprises 233,295 distinct human phosphorylation sites, some of which are annotated with kinases and functional knowledge (45). Fewer than one-third of these sites (29%; 68,481) have been identified by more than one MS-based study, raising several issues. First is the equal consideration of phosphorylation sites identified under extraordinary conditions, such as cells treated with phosphatase inhibitors. Although phosphatase treatment facilitates the identification of vast numbers of phosphorylation sites, the widespread hyperphosphorylation induced by the inhibition of the cells’ phosphorylation “erasers” also likely gives rise to many sites that may not otherwise be phosphorylated to appreciable stoichiometry, even in disease contexts. Another issue is the accumulation of false positives, contributed by the amalgamation of many large-scale phosphoproteomic datasets, each containing a small number of incorrectly localized phosphorylation sites. In the case of PhosphoSitePlus, standards introduced in 2012 require high probability scores for PTM site localization (46). This requirement mitigates but does not resolve the problem, particularly as more phosphoproteomics datasets become available. Along with many other groups, we advocate for the deposition of raw data from MS-based phosphoproteomics studies to repositories such as ProteomeXchange (47), facilitating data reanalysis and re-estimation of false discovery and false localization rates. Last, the observation that fewer than one-third of phosphorylation sites have been identified in more than one study reflects the truly vast scale of the phosphoproteome. In all likelihood, we are some ways away from obtaining a comprehensive view of the phosphoproteome, and it certainly remains substantially under-studied in most biological contexts and perturbations. This view is further supported by the observation that functional phosphorylation sites not previously identified in large-scale phosphoproteomics studies continue to be reported. The phosphorylation of noncanonical amino acids has also been under-investigated, partly owing to their acid lability, and their importance in most contexts therefore remains unclear. Therefore, additional methods for the enrichment and detection of these modifications by MS need to be developed. More broadly, further developments on multiple technological fronts from sample preparation to instrumentation and software are critical, because these advances will increase the coverage of phosphoproteomics studies while also making the technology applicable to more biological contexts and accessible to wider audiences and will strengthen our view of both the breadth and impact of the phosphoproteome.

THE DARK PHOSPHOPROTEOME
Knowledge about the targets of kinases is extraordinarily unequally distributed, following an 80:20 rule: Just 20% of kinases are responsible for the phosphorylation of 87% of currently annotated substrates (Fig. 1 and data file S1). As others have argued (48), this imbalance likely reflects biases in the funding of biological research toward well-studied molecules. This imbalance is further compounded by the limited availability of high-quality small-molecule chemical probes (tool compounds) for most kinases, as well as limited guidance regarding their appropriate use in academic settings. Development of the Chemical Probes Portal (www.chemicalprobes.org/), a resource that crowdsources medicinal chemistry and pharmacology expertise and seeks to address some of these challenges, aiming to provide a concise point of reference guiding the use of chemical probes (49). Considering that around 80% of kinases have fewer than 20 substrates, and 30% are yet to be assigned a single substrate, we expect the actual allocation of
Fig. 1. The human kinases and their known substrates. The areas of circles correspond to the number of reported substrates recorded for each kinase in PhosphoSitePlus, accessed on March 2018. The circle color indicates the number of substrates that have an annotated function recorded in PhosphoSitePlus. TK, tyrosine kinase; TKL, tyrosine kinase–like; STE, sterile; CK1, casein kinase 1. AGC refers to a family that includes PKA (protein kinase A), PKB (protein kinase B; also known as AKT), and PKC (protein kinase C). CAMK, Ca$^{2+}$/calmodulin-dependent protein kinase. CMGC refers to a family that includes CDKs (cyclin-dependent kinases), MAPKs (mitogen-activated protein kinases), GSK (glycogen synthase kinases), and CDK-like kinases. The kinome tree dendrogram was obtained courtesy of Cell Signaling Technology Inc. (www.cellsignal.com) and was manually redrawn.
phosphorylation sites to their cognate kinases to be more evenly distributed than currently reported. Moreover, more than 95% of reported human phosphorylation sites have no known kinase or biological function. These vast numbers of unannotated phosphosites and understudied kinases represent opportunities to expand the current knowledge base of cell signaling and point to fertile ground for future therapeutic development. Unbiased MS-based phosphoproteomics provides a powerful toolkit for future exploration of this dark phosphoproteome.

As the expansiveness of the phosphoproteome continues to be uncovered, some of the greatest challenges now exist in understanding phosphorylation context. Phosphorylation site context can be considered from at least two levels. At a signal-processing level, the relationship between regulatory kinases or phosphatases and their substrates principally defines the network architecture and determines how information flows within and is processed by signaling networks. Signal propagation is shaped by overlapping (or discrete) kinase-substrate and phosphatase-substrate recognition [extensively reviewed in (50)], because phosphorylation sites can be targeted by a single, highly specific kinase or multiple kinases active under different contexts. Other factors contributing to the flow of information through signaling networks include multisite phosphorylation (51, 52), cross-talk between different PTMs (53, 54), feedback and feedforward mechanisms (55, 56), and the formation of scaffolding and adaptor protein complexes (57, 58). These factors collectively give rise to signal processing mechanisms capable of producing outputs seemingly more complex than the sum of their parts (59, 60). On a second level, phosphorylation site context can refer to the functional outcomes arising due to the modification of protein behavior. This change occurs by directly altering enzymatic activity, or indirectly by the formation of binding sites for modular phospho-binding domains, leading to altered protein location, activity, turnover, or association and dissociation of protein complexes. Ultimately, this influence on protein function dictates cell fate, in turn leading to a plethora of biological processes that can be influenced by phosphorylation. On each of these levels of phosphorylation site context, increasingly large phosphoproteomics datasets are giving rise to the concept of the “dark phosphoproteome” — the multitude of phosphosites that cannot currently be placed within known signaling networks and have no known upstream kinase or downstream functional consequence.

Looking Under the Streetlight
The in-depth characterization of the targets of even well-studied kinases has provided important insights into biology. For example, to identify new mTOR (mechanistic target of rapamycin) substrates, Hsu et al. (61) analyzed the phosphoproteomes of cells treated with insulin in the presence or absence of mTOR inhibitors and a TSC2 knockout cell line that displays hyperactivation of mTORC1. Their data revealed new mTORC1 substrates, including multiple phosphosites on the adaptor protein Grb10, informing a previously unknown mechanism by which mTORC1 can suppress insulin action. Likewise, we used global phosphoproteomics to identify substrates of adenosine monophosphate (AMP)–activated protein kinase (AMPK). Analysis of human skeletal muscle after a single bout of high-intensity exercise revealed more than 1000 phosphorylation sites with altered abundance, most of which were not previously implicated in exercise (62). By combining these data with phosphoproteomic analysis of cell lines treated with an AMPK-activating small molecule, we identified several high-confidence exercise-regulated AMPK substrates, including the scaffolding protein AKAP1 (A-kinase anchoring protein 1), which promotes AMPK-mediated phosphorylation that regulates mitochondrial respiration. Similarly, a detailed time-resolved analysis of insulin signaling in adipocytes together with inhibitors targeting AKT and phosphoinositide 3-kinase (PI3K)/mTOR, key nodes in the insulin signaling pathway, identified numerous potential substrates of these kinases. Biochemical characterization revealed that AKT directly phosphorylates SIN1, a component of the mTORC2 kinase complex, uncovering a long-sought-after mechanism by which growth factors can increase mTORC2 activity (55, 63). These studies emphasize that, even for well-studied kinases with many reported substrates, many more important substrates and functions likely remain to be discovered.

Exploring in the Dark
Turning to the ~400 kinases with fewer than 20 reported substrates and the array of phosphorylation sites with no annotated kinases, there are even greater opportunities to uncover kinases with important roles in diverse biological functions. Examining the substrates of these lesser-studied “dark kinases” will likely reveal major insights into cell function. This notion is supported by the observation that annotated kinase-substrate associations and the number of phenotypes reported for genomic lesions of kinases are not correlated: Equal numbers of kinases with fewer or more than 30 substrates have phenotypes identified from deletions or targeted mutations in the International Mouse Phenotyping Consortium (each group containing 32 kinases) (movie S1 and data file S1) (46, 64, 65). Moreover, this number may be even higher, because many kinases with fewer than 30 reported substrates have not yet been phenotypically evaluated in mice. The importance of dark kinases is also apparent from the findings of previous studies that have investigated them. For example, knockdown of the kinase TBK1 in lung cancer cells revealed its mediation of prosurvival signaling by direct phosphorylation and activation of Polo-like kinase 1 (PLK1) (66). In addition, in vivo knockout of NUAK1 (NUAK family SNF1-like kinase 1) in skeletal muscle coupled with phosphoproteomics revealed that it inhibits insulin signaling (67). Last, a screen followed by mechanistic studies showed that active DYRK3 disperses stress granules, releasing sequestered mTORC1 and activating it by phosphorylating PRAS40 and relieving its inhibitory effect (68). The numerous other kinases with limited known substrates and functions likely represent opportunities to gain substantial mechanistic insights into cellular function, and MS-based phosphoproteomics offers an ideal platform for such studies.

The promise of exploring dark kinases for health and disease is exemplified in a study by Steger et al. (69) that identified bona fide substrates of the Parkinson’s disease–related mutant leucine-rich repeat kinase 2 (LRRK2). Despite considerable efforts, high-confidence LRRK2 substrates had remained elusive (70). An activating Parkinson’s disease–related mutation provided an important tool that the authors used together with two LRRK2 inhibitors and MS-based phosphoproteomics. The kinase was also engineered to be resistant to an inhibitor to identify nonspecific drug effects, assisting the recognition of high-confidence targets. Combining mouse embryonic fibroblasts (MEFs) expressing these various mutant kinases with pharmacological manipulation revealed the Rab family of guanosine triphosphatases (GTPases) as direct substrates of LRRK2. In follow-up studies, the group identified the specific Rabs that are phosphorylated by LRRK2 by performing MS-based quantification of Rab immunoprecipitates...
from cells overexpressing >50 Rab proteins containing the consensus motif along with mutant LRRK2 and treated with an LRRK2 inhibitor (71). Coverage and sensitivity were increased with use of appropriate proteases and spike-in stable isotope–labeled peptides of the Rab phosphopeptides. Having successfully established kinase-substrate associations, the authors identified functional consequences of site-specific phosphorylation. Interactomics studies of phosphomimetic and phosphomutant versions of the 14 LRRK2-regulated Rabs identified phosphorylation-dependent interactions of RILPL1 and RILPL2 with Rab8A, Rab10, and Rab12, revealing LRRK2 as a regulator of ciliogenesis. These studies have markedly increased our understanding of the etiology of Parkinson’s disease and have provided potential new therapeutic targets, demonstrating the wealth of information encapsulated within the dark phosphoproteome.

**CHALLENGES IN UNCOVERING THE ARCHITECTURE OF SIGNALING NETWORKS**

Considering that around 97% of reported phosphosites currently have no known regulatory kinase (Fig. 2A), identifying new high-quality kinase-substrate relationships improves our understanding of the architecture and connectivity of signaling networks. The phosphorylation of these substrates can subsequently be used as a readout of kinase activity or to inform training sets for machine learning–based analytical methods. For dark kinases, the identification of substrates also implicates them in the control of specific biological functions. This substrate identification is particularly important because, for many kinases, there is limited knowledge of their cellular functions or their potential involvement in human disease. Establishing their substrate repertoire will therefore likely yield new therapeutic targets. In many cases, knowledge of the precise cellular contexts in which a kinase may be active, such as in the presence of a specific stress or after stimulation with a particular growth factor or cytokine, provides vital information that can assist in the design of experiments that connect the kinase with putative substrates in cells. Conversely, the absence of this knowledge represents a major hurdle for studying many kinases. Even with a basic understanding of conditions in which kinases are active, directly assigning kinases to phosphosites with high confidence is challenging and involves rigorous experiments, as illustrated by the LRRK2 studies described above.

A major issue also lies in the incorrect assignment of substrates to kinases that may not actually be responsible for their regulation in cells or may only phosphorylate them in rare contexts. Arguably, the correction of low-quality annotations is more difficult than the assignment of substrates to kinases in the first place, because compelling evidence is demanded in refuting prior knowledge. These challenges are underscored by the appreciation that many kinases share considerable structural homology in their substrate binding cleft, resulting in overlapping enzyme-substrate recognition. For example, many members of the AGC family of kinases (including PKA, PKC, and AKT) have a strong preference toward substrates with basic residues in specific positions N-terminal to the site of phosphorylation (Fig. 2B). The most common motif represented in the phosphoproteome comprises an acidic residue in the −1 position, fitting 76,265 sites (Fig. 2C and data file S2) (46, 72). Currently, casein kinase 2 (CK2) is the only kinase attributed to have such a preference. Although CK2 is implicated in several downstream biological functions such as adipocyte thermogenesis (73), it is unlikely that one kinase is responsible for such widespread phosphorylation, particularly because the most crucial specificity determinant for CK2 is an acidic residue in the +3 position. It is therefore likely that several dark kinases will display this sequence preference.

Reflecting a biochemical enzyme-substrate interaction, protein phosphorylation can also be context dependent, being influenced by kinase activity, the concentration of enzyme, substrate, and cofactors. This is particularly evident, for example, in the case of in vitro kinase assays, in which excess enzyme, substrate, and ATP are typically used, or after hyperactivating mutations in cells. Under such conditions, many proteins can be phosphorylated by kinases on sites that may not be predominant substrates in cellular contexts. Consequently, although it has been widely recognized that the conservation of specific amino acids surrounding a phosphorylation site (substrate “motifs”) can be a useful predictor of kinase identity (74, 75), few motifs provide sufficient resolving power alone to assign single kinases to substrates with high confidence. The ATR (ataxia telangiectasia mutated and Rad3-related) protein kinase has a particularly specific preference, and 91% of annotated ATR substrates have a glutamine in the +1 position (movie S2). However, motifs are often less specific, because fewer than 18% of sites match a single motif (Fig. 2, D and E). Motifs also share many common phosphosites, even between kinases from different families (movie S3). Even kinases with well-defined motifs can phosphorylate atypical sites in certain cellular contexts. Despite this limited resolving power, supplementing sequence information with additional discriminants, such as protein interactions (76, 77), or quantitative information from phosphoproteomics time series (78) or drug treatment (79, 80) studies is particularly powerful for identifying high-quality substrates. Given that some kinases such as the MAPK family target proteins with specific docking sites in a separate region to the phosphorylated site (81), including docking motif information could further improve kinase predictions. Last, considering that the most well-studied kinases are likely to be selected by researchers to validate potential kinase-substrate associations, the fact that well-studied kinases accumulate the most reported substrates may become somewhat of a self-fulfilling prophecy.

**THE FUNCTIONAL PHOSPHOPROTEOME**

A major challenge for phosphoproteomics, and indeed for the study of all major PTMs, lies in interrogating the function of modified sites. In many cases, the phenotypic consequence of phosphorylation sites can be obscured for many years and can require substantial efforts to uncover. Reflecting these challenges, fewer than 3% of identified human phosphosites currently have a reported function (Fig. 3, A and B), and yet, phosphorylation is frequently found to be essential for cell behavior and serves as a critical control point for life-or-death decisions. Phosphorylation can exert exquisite control over major cell fate decisions, as exemplified by the discovery that the phosphorylation of a single site in the pseudokinase MLKL by RIPK3 serves as a molecular switch that triggers tumor necrosis factor (TNF)–induced cell death in the necroptotic pathway (82). The topological location of phosphorylation sites within or between protein domains may occasionally provide insight into their potential functional consequences. Kinases themselves are commonly regulated by phosphorylation, and remarkably, of the ~80,000 phosphosites occurring in 5132 different pfam-defined protein domains, the largest fraction is found in protein kinase domains, specifically in the conserved catalytic domain rather than only in the activation loop (2789 sites) (Fig. 3C) (83). The second most frequently phosphorylated domain is a C2H2 zinc finger domain, which occurs in around 750 human genes. Fewer
than 1% of these sites have a reported function, but the most commonly reported biological process regulated by phosphorylation is transcriptional regulation (Fig. 3D and movie S4). Given the prominent role of these domains in binding to DNA, RNA, or protein, phosphorylation may frequently regulate such interactions. Similarly, the fifth most represented domain, the RNA recognition motif domain, is thought to function in binding single-stranded RNA during alternative splicing. Other highly represented domains include repetitive domains that may be commonly phosphorylated because of their abundant repetitions, such as in fibronectin type III (“fn3”) filament and immunoglobulin intermediate domains (“i-set”).

Fig. 2. Most phosphosites have no known upstream kinase. (A) The proportion of human phosphosites with a reported kinase in PhosphoSitePlus, accessed on March 2018. (B) Clustering of the surrounding sequence of AKT substrates that fit a previously defined AKT motif (RXRX[S/T][F/L]) and of all reported AKT substrates. Nodes converge if sites have the same amino acid at that position, moving outward. The size of a node indicates the number of sites on that amino acid in that position. (C) The frequency of phosphosites that match motifs defined in HPRD. The 25 most frequently occurring motifs are shown. (D) Specificity of motifs to phosphosites. (E) The number of motifs fit by all reported human phosphosites, ordered by number of motifs fit.
The disparity between the number of phosphosites and reported cognate kinases and functions has led some to speculate that a substantial proportion of the dark phosphoproteome may be without functional consequence ("silent phosphorylation") (84, 85). Despite potential promiscuity in vitro, kinases are typically highly specific in vivo, because of the structure of the active site, external regions that facilitate protein-protein interactions, and the localization and concentration of the kinase, substrate, and relevant scaffolding proteins (86). In cases in which kinases lack specificity, promiscuous phosphorylation events may be removed by phosphatases, providing the cell with another major control point for regulating the phosphoproteome [reviewed in (87)]. Given the enormity of the phosphoproteome, the extent to which it is functional is difficult to evaluate. Several large-scale methods have been developed enabling the global estimation of phosphorylation site occupancy (88–90), revealing that a considerable proportion of sites are phosphorylated to high stoichiometry. For example, during mitosis, half of all mitotic phosphorylation sites had a stoichiometry of more than 75% (89). This stoichiometry contrasts with lysine acetylation, because ~75% of acetylation sites in exponentially growing yeast had a stoichiometry of less than 0.5% (91), and similar observations have been reported in human cells. Although site stoichiometry alone is not sufficient to infer widespread functional relevance of phosphorylation, other evidence supports this view, including a growing number of examples of highly dynamic regulation of a substantial proportion of the phosphoproteome by diverse stimuli. Requiring only a fraction of the energy of transcriptional and translational control of protein abundance, global phosphorylation provides a highly efficient mechanism with which cells and tissues can respond to their environment. This ability to rapidly fine-tune responses is reflected in the circadian control of mouse liver, where both the extent and magnitude of rhythmic diurnal phosphorylation changes greatly surpass those at the level of proteome and transcriptome (92). The total energy demands of phosphorylation depends on the rate of phosphate turnover (cycles of phosphorylation and dephosphorylation). Although ATP consumption is estimated to be considerable for a few highly active and abundant metabolic enzymes, such as glycogen phosphorylase kinase in skeletal muscle (on the order of ~10% of the cellular ATP turnover) (93), it is likely to be diminishing for most of the kinases that

---

**Fig. 3. Phosphorylation site function.** (A) The proportion of human phosphosites with a reported function in PhosphoSitePlus. (B) Overlap of phosphosites with a reported function and reported kinase. (C) The top 13 domains in which reported phosphorylation sites occur. (D) and (E) Frequency of reported biological processes (D) and functions (E) controlled by phosphorylation, from the regulatory sites database in PhosphoSitePlus, accessed on March 2018.
participate in signal transduction. Although there remains no good estimate of the contribution of global protein phosphorylation to ATP consumption, an elegant study combining mathematical modeling with measurement of specific phosphorylation events enabled the authors to estimate that, in the case of receptor tyrosine kinases such as the EGF receptor (EGFR), tyrosine phosphorylation can over receptor tyrosine kinase autophosphorylation sites rapidly with half-lives of less than a minute (94). Methods capable of measuring site-specific phosphorylation turnover on a global scale would be desirable, because they would inform regulatory mechanisms and relative kinase and phosphatase activities and provide insight into the energy demands of maintaining cellular-wide phosphorylation.

Phosphorylation events may be silent, but they must not be deleterious, or they would have been selected against. Conversely, if changes in protein activity or behavior by phosphorylation provide an organism with an evolutionary advantage, then it is tempting to speculate that the positions of such residues should be conserved. Several studies have compared phosphosite conservation (95, 96), and in contrast to lysine acetylation, which is highly conserved across species, phosphorylated serine and threonine residues in most cases are only marginally more conserved than their nonphosphorylated counterparts (97). However, in contrast to acetylation, which commonly occurs in structured domains, phosphorylation frequently occurs in disordered regions, which have low conservation (85, 98). Combining global phosphoproteomics with genome alignment of 32 fungal species, Holt et al. (99) found that the precise position of most phosphorylation sites is not well conserved evolutionarily and that, although metabolic enzymes frequently contain highly positionally conserved phosphosites, less conserved sites regulate protein–protein interactions, particularly in disordered regions. Therefore, in contrast to the phosphorylation of key catalytic residues in enzymes, phosphorylation that regulates protein interactions could instead occur in “hotspots” because a relatively broad interface can affect protein interactions. The frequency of less conserved phosphosites is therefore not surprising given that regulation of molecular associations is currently the most frequently reported function affected by phosphorylation (Fig. 3E and movie S4).

The involvement of protein phosphorylation in disease states also exemplifies their functional importance. Kinases are frequently implicated in disease and are a major protein family targeted by therapeutics. Hundreds of phosphosites are known to be altered in abundance in disease contexts, and many of these are collated in the PhosphoSitePlus disease-associated sites collection (Fig. 4A). Functional roles of 453 of these sites have been found. For example, the activating site on Ser133 in CREB (cyclic AMP response element–binding protein) is required for heart function, and transgenic mice with an alanine mutation at this site experience heart failure (100).

Enrichment analysis is a popular tool for linking large-scale quantitative data with biological processes. If phosphosites are already known to have an inhibitory or activating function, then directional information can be incorporated in these analyses. As databases of site-specific functions become more comprehensive, these resources will have greater utility for analyzing both new and preexisting phosphoproteomics datasets, accelerating signal transduction research. Detailed annotation of phosphosite function, as well as the biological processes affected, can link signaling to outcomes and help inform mechanism. From the limited site-specific annotation currently available, trends are emerging, such as the frequent regulation of the process of exocytosis by phosphosites that control intracellular localization.

Considering the dark phosphoproteome in context with other PTMs can also help illuminate function. Protein domains may be particularly important if targeted by multiple modifications, and proteins harboring multiple PTMs are more likely to be dysregulated in disease (102), highlighting the functional importance of the theme of PTM cross-talk [reviewed in (103)]. Cross-talk can occur whereby PTMs promote or inhibit modification of other sites, either in close proximity, or even at distal locations on the protein, or through competition for the same residue. Prominent examples of the interplay between multiple PTMs decorating the same protein include those of histones (104) and p53 (105), whereas competition between O-GlcNAcylation and phosphorylation of Ser/Thr residues exemplifies reciprocal PTM cross-talk (106) [reviewed in (107) and (108)]. This paradigm is likely to extend to other “signal integrator” proteins, which will be illuminated by further experimental work in new biological contexts and with the study of previously unknown protein modifications (109). Studying PTM cross-talk on a global scale is challenging because the most effective strategies are currently specialized to target single PTMs. Moreover, the most broadly applied MS-based proteomics methods rely on digesting proteins into short peptides, thereby obscuring whether PTMs are colocalized on the same protein molecule. Top-down proteomics, in which intact proteins are analyzed (110), or “middle-down” proteomics (111), in which larger peptides (>50 amino acids) are analyzed, provide avenues to investigate PTM cross-talk. Innovative genome editing tools (112, 113) may also unlock methods to systematically explore the functional consequences of global phosphorylation. However, analogous to single-gene deletion screens in yeast, in which only about 20% of genes were found to be essential for growth (114), cooperativity (here multisite, rather than multicomponent), redundancy, and challenges in detecting context-dependent functions are likely to pose major challenges.

### TRADING THE TORCH FOR THE FLOODLIGHT

Current phosphoproteomics studies can identify tens of thousands of phosphorylation sites and frequently reveal many hundreds of changes among these. Because linking kinases with functional consequences of the phosphosites they target is challenging, most studies can only expect to characterize a small fraction of those identified. Therefore, elegant solutions are needed to explore and prioritize phosphoproteome function more broadly. Large-scale phosphoproteomics coupled with functional studies can help connect diverse biological processes and perturbations with underlying changes to signaling networks. Kinases and pathways found to be most active in these contexts can be prioritized for investigation, thereby linking potential regulatory kinases with the functional processes. For example, Lundby and colleagues (115) delineated β-adrenergic signaling...
in vivo by measuring the phosphoproteomes of mouse cardiac tissue with combinations of agonists and antagonists specific to β1- or β2-adrenergic receptors, highlighting the involvement of kinases not previously implicated in myocardial contractility. Wang and colleagues provided another elegant example of the effective combination of large-scale phosphoproteomics with functional studies to facilitate the identification of functionally implicated signaling activity. Forward-genetics analysis of sleep in randomly mutagenized mice identified a mutation in salt-inducible kinase 3 (SIK3), a member of the AMPK family, that causes a constitutive need for greater-than-normal amounts of sleep (116). This study provided a key genetic model (the Sleepy mice), which the group combined with behavioral models of sleep deprivation and global phosphoproteomics of whole mouse brain (117). The authors identified 80 hyperphosphorylated proteins whose changes in phosphorylation state mirror those of sleep need in both genetic and behavioral models. More than 86% of these were synaptic proteins, whereas only 20% of total phosphoproteins were synaptic, and mutations in 15% of these proteins are already known to cause sleep phenotypes in mice or humans, suggesting that these data are likely rich in functionally relevant targets for sleep regulation. Measuring the effect of a range of perturbations on the phosphoproteome in the context of a single biological process is also particularly powerful. For example, comparing the response of the phosphoproteome of pancreatic cells in response to glucose and in combination with seven different treatments that enhance glucose-stimulated insulin secretion highlighted targets common to multiple treatment conditions, revealing shared key regulatory nodes in this network (118).

To gain a global understanding of phosphosite function, orthogonal large-scale screens for the most common phosphosite functions could be coupled with phosphoproteomics. Many functional readouts are possible, including those for protein interactions, activity, localization, and stability. Screening for interactions specific to phosphorylation sites may inform function by providing clues as to signaling mechanisms and biological processes involved (119). Peptide interaction assays can use tagged, synthesized phosphopeptides and control nonphosphorylated peptides, enabling affinity purification coupled with liquid chromatography (LC)–MS to identify phosphorylation-specific interactions. These assays are powerful because they can be performed in an unbiased and large-scale manner (120, 121), providing opportunities to measure interactors of many different phosphosites in different cell types and perturbations (122, 123). As with in vitro kinase assays, care should be taken when performing in vitro peptide studies that potential artifacts are not introduced owing to high concentrations of peptides being exposed to non-natural cellular contexts. The creation of large libraries of synthesized peptides is aiding systematic peptide interaction screening and other applications, including evaluating data analysis pipelines (124). Sequence analysis of interacting proteins on a global scale could identify new phosphosite binding motifs. By collecting data on a broad range of contexts, we could approach a complete interaction annotation for all phosphosites. The regulation of protein interactions is a pervasive function of phosphorylation, with 31% of currently annotated functional sites serving this role (Fig. 3E), making this information valuable for the field. Integrating multiple “omics” datasets could prioritize the next most prevalent reported function: regulation of protein activity.
To link phosphorylation with metabolic enzymatic activity, phosphoproteomics could be coupled with metabolomics (125). Likewise, complementary transcriptomics studies could reveal regulatory phosphosites on transcription factors or coactivators. Such a model could be adopted for any class of proteins for which activity can be measured. Phosphosites that potentially alter localization may be identified by organelle-resolved phosphoproteomics (126). Protein stability can be assessed by monitoring protein thermal melting curves (127), which can be assessed on a global scale by MS-based proteomics in a method termed thermal proteome profiling (TPP) (128). In TPP, cell lysates are heated at a range of temperatures, with higher temperatures required to denature more stable protein forms. This technique has been used to globally investigate drug-binding targets (129), protein-protein interactions (130), and inter- and intraspecies thermostability (131). A broad survey of how phosphorylation alters the stability of proteins could also be possible using these same techniques. A global map of the association between phosphosites with protein stability would certainly be a valuable resource for evaluating widespread phosphorylation function.

Although phosphorylation within the activation loop of a kinase is often the best evidence for kinase activity, analysis with the Phomics activation loop tool (132) reveals that 26% of the 481 kinases containing a reported phosphosite do not have a reported phosphosite on an activation loop. This finding may be due to technical limitations prohibiting their identification by current methods or to these kinases requiring different modes of activation such as association with regulatory molecules or proteins or the release of autoinhibitory domains. Alternatively, kinases can be constitutively active and regulated by alternative means. For example, CSK (C-terminal Src kinase) is regulated by binding the scaffolding protein CBP (CSK-binding protein) rather than by activation loop phosphorylation (133). Directly studying the kinome to observe regulatory phosphorylation sites has been achieved with the aid of kinase enrichment using ATP probes or immobilized broad-spectrum kinase inhibitors followed by MS (134–136). Kinome-wide screens can also be applied to study processes or phenotypes that can be measured on a large scale. Global kinome knockdown screens by short interfering RNA have revealed biological insights such as discovery of the role of ULK1 in autophagy (137); interplay between PFKFB3, a regulator of glycolysis; and signal propagation through the insulin-like growth factor 1 (IGF1)–PI3K pathway (138). Phosphoproteomics can also effectively inform targeted loss-of-function screens to identify key determinants of function. Large-scale quantitative phosphoproteomics analysis of cells recovering from DNA damage highlighted 154 proteins harboring reproducibly perturbed phosphorylation sites, systematic depletion of which revealed that a substantial proportion of these phosphorylation sites are required for recovery (139). If phenotypes relevant to whole-body physiology are of interest, then Drosophila provides a particularly tractable system because loss-of-function screens can be readily performed at a whole-body or tissue-specific level (140, 141). Likewise, the formation of human-derived organoids provides opportunities for simplified functional screening with high relevance to human biology and disease (142, 143). We expect that such platforms when combined with phosphoproteomics will be particularly powerful for characterizing the interplay between signal transduction and physiological function.

Emerging techniques in cell biology, particularly the high-throughput application of the CRISPR/Cas9 system (144), have simplified large-scale knockout studies (145, 146). The use of CRISPR for identifying previously unknown substrates was demonstrated by CRISPR-directed knockout of PKA catalytic subunits in epithelial cells (147). Phosphoproteomics revealed that only around one-third of the sites decreased in abundance were possible PKA targets based on sequence. This revelation is not surprising because PKA, like many kinases, regulates the activity of other kinases and phosphatases, so removing this protein would have profound effects on the cell. Considerations for CRISPR knockouts include the use of multiple clonal cell lines for each knockout, the measurement of changes in global protein abundances that may occur in addition to phosphorylation, and suitable controls. Rescue experiments in which the deleted protein is restored provide the most convincing controls, because such experiments show that any effects induced solely by the generation of the cell line are adequately controlled for. Advances in genome editing are providing new tools with which dark kinases can now be genetically manipulated and can be particularly effective for shedding light on function when coupled with phenotypic screening. For example, monitoring inflammasome activation together with a genome-wide loss-of-function screen revealed a role for NEK7 (NIMA-related kinase 7) in this process (148), and lethality screens in patient-derived glioblastoma stem-like cells revealed a role for the kinase PKMYT1 (protein kinase, membrane-associated tyrosine/threonine 1) in mitosis (149). Coupling these technologies with global phosphoproteomics could assist in the identification of substrates responsible for these functional consequences.

**BIOCHEMICAL STRATEGIES FOR STUDYING SIGNALING NETWORK ARCHITECTURE**

Innovative biochemical and genetic approaches play an important role in pinpointing high-quality kinase-substrate relationships. In a series of pioneering studies, Shokat and colleagues (150, 151) elucidated a strategy that exploited some of the conserved properties of the adenosine binding pocket, enabling the engineering of mutant kinases to accept unnatural bulky ATP analogs, such as ATP-γ-S modified with a bulky group in the N6 position (A*TP-γ-S). A*TP-γ-S treatment of cells engineered to express these analog-sensitive kinases yields specific thio phosphorylation of targets of the mutant kinase. These targets can subsequently be tagged, immunopurified, and identified by MS, enabling the definition of direct protein kinase substrates (152). This approach has been used to identify high-confidence sub- strates of AMPK, illuminating roles for this kinase in mitosis and cell motility (153, 154). Similarly, a yeast strain in which the cyclin-dependent kinase CDK1 is replaced with an analog-sensitive kinase enabled specific inhibition by an N6-modified derivative of a protein kinase inhibitor. By coupling these chemical-genetic tools with phosphoproteomics, targets of CDK1 could be uncovered in a global and unbiased manner (99).

Optogenetics also provides powerful tools for unraveling complex signaling networks. A bulky photocaged lysine can be encoded at the conserved lysine residue to prevent ATP binding and hence kinase function, which can then be rapidly switched to lysine with light (155) using an alternate codon, a pyrrolysyl–transfer RNA (tRNA) synthetase, and pyrrolysine tRNA. Uncaging pyrrolysine can be used as an activation mechanism if the caged kinase is also mutated to be constitutively active, which can be achieved with phosphomimetic residue mutations or by deletion of an autoinhibitory domain. Optogenetics enables the activation of kinases with unprecedented temporal resolution, overcoming specificity issues facing pharmacological approaches, as well as network adaptation or
compensatory effects that may occur with longer-term genetic manipulations. Coupled with MS-based phosphoproteomics, such biochemical tools hold promise for discovering previously unknown kinase-substrate relationships and helping to unravel the complex architecture of cell signaling networks.

Approaches for dissecting context-specific signal transduction are also of interest. Interacting adaptor proteins alter the flow of information through signaling networks (156). Time-resolved interactomics of the receptor tyrosine kinase scaffold protein Shc1 complemented a phosphoproteomics time course in response to EGF, prioritizing targets and revealing Shc1-mediated signal flow (157). Photosensor domains, such as LOV2 from plant proteins, promote light-responsive protein-protein interactions (158) and could be harnessed to tether and release kinases with other proteins or organelles to explore the effects of adaptor proteins and localization on distinct substrate pools. Kinase isoforms also contribute to the tangled web of context-specific signaling, because they are often expressed at different levels between cells and tissue types and may be distinctly compartmentalized within cells. Genetic studies have revealed that, despite sharing considerable sequence homology, isoforms of kinases have functionally distinct roles, potentially owing to distinct pools of substrates. For example, the AKT2 isofrom is most abundant in metabolic tissues, and its deletion in mice causes an insulin resistance–like phenotype (159), whereas the deletion of AKT1 does not cause metabolic defects but is required for cell growth (160). Gonzalez and colleagues (161) used an elegant chemical genetic system to investigate isoform-specific AKT signaling by generating genetic mutants of AKT that are resistant to the potent allosteric AKT inhibitor MK–2206, pointing to overlapping and distinct functions. Using phosphoproteomics to study lung fibroblasts engineered to selectively express each of the three AKT isoforms, Sanidas et al. (162) also investigated isoform-directed signaling. This work not only revealed largely overlapping substrates but also identified specific targets of each AKT isofrom. Among the isoform-specific targets of AKT in lung fibroblasts was IWS1, which was phosphorylated by AKT1 and AKT3, but not by AKT2. Remarkably, this specificity was preserved under in vitro assay conditions, highlighting that isoform-specific differences can result from structural differences that directly confer enzyme-substrate recognition. Given the insights into the control of key biological processes that these and other similar studies have provided, future exploration of signaling specificity using innovative biochemical strategies is likely to be fruitful.

THE DRUGGABLE PHOSPHOPROTEOME

The phosphoproteome presents fertile ground for developing drugs to treat human disease (163), with the number of clinically approved drugs targeting protein kinases more than tripling to over 40 in the past 5 years (164). The first approved small-molecule kinase inhibitor developed for clinical use was imatinib (Gleevec; Novartis). Approved in 2001, imatinib inhibits the tyrosine kinase ABL and is used to treat chronic myeloid leukemia (CML) (165). In phase 2 trials, 95% of patients with CML who were resistant to interferon treatment—the previous first-line treatment—had complete hematological response with imatinib, and with a median follow-up of 18 months, the estimated progression-free survival was 89% (166). Continuing this trend, most approved drugs for cancer target tyrosine kinases, and more broadly, kinases are the most common targets for cancer currently in clinical trials (167). Interrogating kinase activity to better understand cancer treatment response has also been highly effective (168–170). However, despite these successes, more than half of the kinome remains untargeted, even when considering nonclinically approved small molecules. Moreover, considering the extremely high costs of developing completely new drugs, most new clinical trials are focused on repurposing already approved compounds toward new subindications, with relatively few focusing on new targets (171). Notable exceptions include the first inhibitors against MEK1/2 (trametinib), Bruton’s tyrosine kinase (BTK) (ibrutinib), CDK4/6 (palbociclib), and SYK (fostamatinib), all approved in the past 5 years. Many of these have been followed by additional compounds against the same target, emphasizing that target identification and validation, rather than lead discovery, are major hurdles in pharmaceutical development. Enormous opportunities therefore likely remain in targeting future therapeutics against a broader spectrum of kinases.

In addition to therapeutic development, it is also important to better understand the gamut of drug efficacy among the population. This requirement is underscored by findings that the response of patients with non–small cell lung cancer to the EGFR inhibitor gefitinib correlates with activating somatic EGFR mutations (172, 173), whereas in contrast, around half of patients with melanoma harboring BRAF mutations fail to respond to BRAF (Rapidly Accelerated Fibrosarcoma murine sarcoma viral oncogene homolog B) inhibition (174), and inhibition of human EGFR (HER) family shows only moderate results in HER2-driven breast cancers (175). Heterogeneity in treatment response may be in part due to the complex combinations of mutations, amplifications, and deletions in many cancers, confounding rational target selection from genomics data alone. Given the wide range of efficacy of various kinase inhibitors in treating human cancers, better understanding the activity of kinases and signaling network activity among different tumors will inform better treatment practices. Here, phosphoproteomics holds great promise, because the phosphoproteome is likely to be a more accurate predictor of kinase and signaling network activity than the genomic landscape in isolation. Numerous studies have demonstrated the utility of large-scale phosphoproteomics for characterizing and stratifying cancers (176–178) and for investigating mechanisms of drug resistance (179, 180). However, major hurdles remain in making this technology reliably applicable to clinical settings. In particular: achieving sufficient sensitivity to identify high coverage of key phosphorylation sites from the minimal amounts of protein material that can be retrieved from tumor biopsies; the throughput to process large numbers of patient samples; and the reproducibility to measure phosphorylation with high quantitative precision. Although several phosphoproteomics studies have demonstrated exquisite sensitivity (181) and the ability to measure hundreds of phosphosites in hundreds of samples (182), existing global phosphoproteomics workflows have not been sufficiently streamlined or robust to enable measurement of phosphoproteomes across very large numbers of samples at sufficient depth to cover key pathways and sites. To address this, we developed a phosphoproteomics workflow called “EasyPhos,” which simplifies sample preparation to such an extent that it enables the measurement of hundreds of phosphoproteomes in cells and tissues, and with very high reproducibility such that metabolic or chemical labeling is often not required (183). This workflow was originally limited to relatively large starting material requirements of at least a milligram of protein. We have made developments that now further improve the performance of these methods in high-sensitivity applications for which only minute starting materials are available, bringing this technology closer to the clinic (184).

The unbiased analysis of the selectivity of small-molecule inhibitors has repeatedly revealed that most commonly used compounds

in fact target many kinases, and perhaps surprisingly to many biologists, the same is true for many clinically approved drugs (185–188). Imatinib for example, a highly efficacious and relatively selective drug designed to target ABL, also inhibits the tyrosine kinases c-KIT and PDGFR with similar potency (189). The finding that a drug that was rationally designed against a single kinase and has subsequently displayed great success in the clinic in fact advantageously targets multiple kinases may be no coincidence. Here, it is worth emphasizing that the purpose and therefore most important characteristics of drugs and chemical probes are not the same. Chemical probes are primarily designed for revealing new biology and for exploring target function. Exquisite selectivity against close family member biological targets is therefore highly desirable for effective chemical probes [reviewed in (190)]. In the case of drugs, however, superior efficacy in treating disease phenotypes while also minimizing unwanted side effects is of far greater importance than is selectivity toward single biological targets. Given that multiple pathways often participate in disease processes and that phosphorylation networks are intrinsically plastic (191), it is unsurprising that most tumors can escape inhibition of a single kinase [see (192) for a comprehensive review of mechanisms underlying cancer drug resistance]. The specific design of drugs to intentionally target multiple kinases (“targeted polypharmacology”) (193) is therefore likely to become a particularly important approach in the future of rational drug discovery. Phosphoproteomics has already begun to show promise here, through the repurposing of existing drugs after the discovery of off-target activities pointing to noncanonical but efficacious targets (194). In the future, the availability of phosphoproteomics data against every compound synthesized in drug discovery programs would enable the optimization of drugs against complex signaling network patterns, rather than attempting to maximize drug specificity against a single target. Even without knowledge of the functional consequences, or even regulatory kinase of every phosphorylation site, we predict that such an approach could facilitate the discovery of highly effective target combinations that could not otherwise have been predicted.

Another emerging interface between the phosphoproteome and drug discovery is in the application of phosphoproteomics to gain mechanistic insights into the actions and side effects of drugs in vivo. This is particularly powerful when bridged by functional or behavioral studies. As the largest family of human membrane proteins, the G protein–coupled receptors (GPCRs) are a particularly important target for pharmaceutical development. Functionally selective agonists can activate GPCRs, such as the κ opioid receptor (KOR), in a pathway-specific manner, a phenomenon that could be exploited to achieve drugs with fewer side effects. Liu and colleagues (195) performed high-throughput phosphoproteomics using the EasyPhos platform to study the signaling downstream of the KOR after in situ administration of distinct KOR-specific agonists directly to the mouse brain. Analysis of more than 60,000 different phosphosites enabled dissection of anatomical brain region and time-resolved signaling in response to KOR agonists that elicit functionally selective outcomes, such as beneficial antinociceptive effects versus undesirable aversive effects. KOR agonists associated with aversion preferentially activated mTOR signaling in the striatum, and mTOR inhibition during KOR activation could specifically abolish aversive phenotypes during behavioral assays. This study is a powerful demonstration of systems pharmacology, because it links complex phenotypic responses to drugs (in this case, the response to KOR agonists) with an unbiased view of signaling network activity. It also demonstrates that rational manipulation of these networks can direct behavioral outcomes. The union of phosphoproteomics with pharmacology represents an alluring future application. We envision that continued developments in MS-based phosphoproteomics technologies will create new opportunities for better understanding drug responses and how signaling networks might be manipulated to direct drug responses to minimize side effects or enhance treatment efficacy.

**VISUALIZING THE PHOSPHOPROTEOME**

Effectively visualizing large-scale, multivariate data, particularly data that differ greatly in scale and qualitative features is a major challenge. Effective visualization provides a degree of data dimensionality reduction to enable its representation in a two-dimensional (2D) or 3D space while also helping to crystallize patterns in the data. The visualization tool Coral represents features of the kinome according to nodes and branches of a kinome tree or network (196), which could prioritize relevant dark kinases by representing quantitative values such as kinase enrichment, phosphorylation, and drug profiling or qualitative data such as associated diseases. The Minardo plot developed by O’Donoghue and colleagues (197) was designed to enable concise visual display of complex, multidimensional phosphoproteomics data for a focused set of molecules. Minardo can encode up to eight dimensions of data, including continuous time, subcellular location of proteins and complexes, phosphorylated residues, annotation of cellular functions, timing of reuptake and secretion of pathway products, association and disassociation of protein complexes, phosphorylation and dephosphorylation, and 2D thumbnails of protein structure. This approach has been applied to time-resolved phosphoproteomics data, providing clear and focused snapshots of dynamic signal transduction (197, 198). Broad exploratory visualization of phosphoproteomics data is also a challenge, which is particularly important considering the need to explore the dark phosphoproteome. Tools such as Cytoscape can map proteins onto protein-protein interaction networks to allow discovery of emergent patterns (199). Many plug-ins have been created for Cytoscape, with capabilities such as clustering and network analysis (200), pathway enrichment (201), kinase-substrate relationships, and time-series phosphoproteomics visualization (202). Although the analysis capabilities built into this interactive software are highly useful, large networks can quickly become overwhelming, limiting their interpretation.

We have developed a method for visualizing multivariate data in 3D (movies S1 to S4). The algorithm applies particle interaction rules on data points, such that nodes self-organize into layouts that enhance their representation of relationships in a multidimensional space. This self-organization allows nodes to cluster on the basis of multiple variables at one time, for example, the number of annotated substrates and kinase family. Clusters can be visualized by constellations of nodes with complementary density maps, superimposed network connectivity, and configurable node properties linked to extra dimensions. Nodes can be locked to a sphere to minimize occlusion. This approach is therefore applicable to exploring a range of data types, including phosphoproteomics data. It can also be used to visualize changes in time, amenable to signaling time courses. Compatibility of this visualization with virtual reality headsets is being developed and will facilitate improved interpretation with a full 3D view. This approach facilitates the visualization of multiple different omics datasets in the same space, in their relevant hierarchies. For example, regulatory transcription factors can reside above a layer of...
genes clustered by their transcriptional and proteomic changes. Combining multiple large-scale datasets in this way facilitates visualization and interpretation of the dark phosphoproteome.

THE FUTURE OF THE DARK PHOSPHOPROTEOME

Mapping substrates to kinases and assigning function to these substrates will be increasingly important as phosphoproteome coverage becomes saturated. These goals will require a transformation in the way phosphorylation function is investigated, and thoughtful experimental design will continue to be essential. Experiments that contribute insights into the regulation of the phosphoproteome under different cellular and organismal contexts, and particularly how this regulation drives phenotypes, will be key. Considering the complexity of signaling networks, such functional insights will rely on sophisticated experimental designs, often comprising many samples. Efforts to reduce the costs of performing phosphoproteomics experiments, through higher-sensitivity methods (184) and by increasing the throughput of phosphoproteomics workflows, will facilitate increasingly complex experimental designs comprising different pharmacological doses, time-series data, and positive and negative control samples. Throughput may also be improved by the ability to measure multiple samples in the same MS run (sample “multiplexing”), with isobaric peptide tags such as TMT (Tandem Mass Tags) enabling the simultaneous analysis of up to 11 samples. The EASI-tag (easily abstractable sulfoxide-based isobaric-tag) technology now enables multiplexed MS2-based peptide quantification that is free from interference (so-called ratio compression) normally resulting in impaired quantification accuracy with other reporter ion–based approaches (203). This method is attractive for phosphoproteomics studies relying on quantification of a single peptide harboring the site of interest and where precision and accuracy are therefore highly desirable.

Precision medicine promises to revolutionize health care through genomic sequencing or the direct measurement of blood-based protein biomarkers for which proteomics is the only applicable approach (204). In this setting, phosphoproteomics also promises to be an important diagnostic, by the measurement of large phosphorylation networks in patient-derived samples to identify systems-level changes that may be causal or predictive of disease states (205). Here, major challenges remain for the future, including further increasing the sensitivity, reproducibility, and throughput of phosphoproteome measurements to meet the demands of the clinic. From a data analysis perspective, with many measures and correlations within and between omics and clinical variables, advanced computational methods such as machine learning are improving the ability to extract insights from these data. New visualization approaches will facilitate exploratory analysis of large phosphoproteomics data, including in virtual 3D spaces. New methods will also facilitate the interactive sharing and exploration of data collaboratively, for example, between multiple laboratories and hospitals, to facilitate hypothesis generation and to identify patterns predictive of response. With nodes representing patients, multivariate clustering of clinical variables in space could group patients together with changes to their phosphoproteome in disease or in response to treatments. We anticipate that, in the future, patient stratification by their phosphorylation network status may be a powerful predictor of treatment response because it closely reflects cellular activity in the tissue from which it was measured. An advantage shared with proteome-level measurements is that the phosphoproteome incorporates a broad gamut of contextual inputs, such as genetic, environmental, and treatment variables. Continued acceleration in technologies that facilitate the measurement of high-quality and deep phosphoproteomics data will further expand our view of the phosphoproteome and will in turn provide analytical challenges. Rigorous experimental design, improved software for data analysis and visualization, and secondary screens and developments in cell biology will each contribute to illuminating the dark phosphoproteome. Future studies incorporating these features are poised to offer crucial new insights into biology, health, and disease.

SUPPLEMENTARY MATERIALS

www.sciencesignaling.org/cgi/content/full/12/565/eaau8645/DC1

Data file S1. Human kinases and annotations.

Data file S2. Motifs fit by the human phosphoproteome.


Movie S1. Many biologically important kinases have limited known substrates.

Movie S2. Sequences surrounding phosphoserine sites with reported kinases reveal patterns.

Movie S3. Many phosphosites share common sequence motifs.


REFERENCES AND NOTES

37.  

32.  

31.  

29.  

27.  

24.  

41.  

40.  

38.  

42.  


Programmable base editing of a•T to G•C in genomic DNA without DNA cleavage.
B.mutants are sensitive to the inhibition of activated RAS.

K. mutans are sensitive to the inhibition of activated RAS.

I. mutans are sensitive to the inhibition of activated RAS.

J. mutans are sensitive to the inhibition of activated RAS.

K. mutans are sensitive to the inhibition of activated RAS.

L. mutans are sensitive to the inhibition of activated RAS.

M. mutans are sensitive to the inhibition of activated RAS.

N. mutans are sensitive to the inhibition of activated RAS.

O. mutans are sensitive to the inhibition of activated RAS.

P. mutans are sensitive to the inhibition of activated RAS.

Q. mutans are sensitive to the inhibition of activated RAS.

R. mutans are sensitive to the inhibition of activated RAS.

S. mutans are sensitive to the inhibition of activated RAS.

T. mutans are sensitive to the inhibition of activated RAS.

U. mutans are sensitive to the inhibition of activated RAS.

V. mutans are sensitive to the inhibition of activated RAS.

W. mutans are sensitive to the inhibition of activated RAS.

X. mutans are sensitive to the inhibition of activated RAS.

Y. mutans are sensitive to the inhibition of activated RAS.

Z. mutans are sensitive to the inhibition of activated RAS.


Acknowledgments: We thank members of the Metabolic Systems Biology group for constructive feedback and T. Clark for narration of the accompanying videos. We acknowledge J. Cobcroft for funding support. Funding: This work was supported by the National Health and Medical Research Council (NHMRC) (grants GNT1061122 and GNT1086850 to D.E.J.). E.J.N. was supported by an Australian Government Research Training Program (RTP) Scholarship and the University of Sydney Val Street Scholarship, and S.J.H. was supported by a fellowship from the University of Sydney (G197569).

Submitted 23 July 2018
Accepted 3 January 2019
Published 22 January 2019
10.1126/scisignal.aau8645

Citation: E. J. Needham, B. L. Parker, T. Burykin, D. E. James, S. J. Humphrey, Illuminating the dark phosphoproteome. Sci. Signal. 12, eau8645 (2019).
Illuminating the dark phosphoproteome
Elise J. Needham, Benjamin L. Parker, Timur Burykin, David E. James and Sean J. Humphrey

DOI: 10.1126/scisignal.aau8645