Bioinformatics, 38(15), 2022, 3785–3793
https://doi.org/10.1093/bioinformatics/btac406
Advance Access Publication Date: 22 June 2022
Original Paper

OXFORD

Systems biology

Functional characterization of co-phosphorylation networks

Marzieh Ayati [1,*,†], Serhan Yılmaz[2,†], Mark R. Chance[3,4,5] and Mehmet Koyuturk[2,4,5,*]

[1]Department of Computer Science, University of Texas Rio Grande Valley, Edinburg, TX 78531, USA, [2]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106, USA, [3]Department of Nutrition, Case Western Reserve University, Cleveland, OH, USA, [4]Center for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, OH, USA and [5]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on August 17, 2021; revised on March 16, 2022; editorial decision on June 11, 2022; accepted on June 18, 2022

Abstract

Motivation: Protein phosphorylation is a ubiquitous regulatory mechanism that plays a central role in cellular signaling. According to recent estimates, up to 70% of human proteins can be phosphorylated. Therefore, the characterization of phosphorylation dynamics is critical for understanding a broad range of biological and biochemical processes. Technologies based on mass spectrometry are rapidly advancing to meet the needs for high-throughput screening of phosphorylation. These technologies enable untargeted quantification of thousands of phosphorylation sites in a given sample. Many labs are already utilizing these technologies to comprehensively characterize signaling landscapes by examining perturbations with drugs and knockdown approaches, or by assessing diverse phenotypes in cancers, neuro-degerenational diseases, infectious diseases and normal development.

Results: We comprehensively investigate the concept of 'co-phosphorylation' (Co-P), defined as the correlated phosphorylation of a pair of phosphosites across various biological states. We integrate nine publicly available phosphoproteomics datasets for various diseases (including breast cancer, ovarian cancer and Alzheimer's disease) and utilize functional data related to sequence, evolutionary histories, kinase annotations and pathway annotations to investigate the functional relevance of Co-P. Our results across a broad range of studies consistently show that functionally associated sites tend to exhibit significant positive or negative Co-P. Specifically, we show that Co-P can be used to predict with high precision the sites that are on the same pathway or that are targeted by the same kinase. Overall, these results establish Co-P as a useful resource for analyzing phosphoproteins in a network context, which can help extend our knowledge on cellular signaling and its dysregulation.

Availability and implementation: github.com/msayati/Cophosphorylation. This research used the publicly available datasets published by other researchers as cited in the manuscript.

Contact: marzieh.ayati@utrgv.edu or mehmet.koyuturk@case.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein phosphorylation is a ubiquitous mechanism of post-translational modification observed across cell types and species. Recent estimates suggest that up to 70% of cellular proteins can be phosphorylated (Wilhelm *et al.*, 2014). Phosphorylation is regulated by networks composed of kinases, phosphatases, and their substrates. Characterization of these networks is increasingly important in many biomedical applications, including the identification of novel disease-specific drug targets, development of patient-specific therapeutics and prediction of treatment outcomes (Cohen, 2001; Rikova *et al.*, 2007).

Phosphorylation is particularly important in the context of cancer, as the down-regulation of tumor suppressors and the up-regulation of oncogenes (often kinases themselves) by dysregulation of the associated kinase and phosphatase networks are shown to have key roles in tumor growth and progression (Halim *et al.*, 2013). To this end, the characterization of signaling networks enables the exploration of the

interconnected targets (Yılmaz *et al.*, 2021) and the identification of causal pathways (Babur *et al.*, 2021), leading to the development of kinase inhibitors to treat a variety of cancers (Butrynski *et al.*, 2010; Zhou *et al.*, 2011). Disruptions in the phosphorylation of various signaling proteins have also been implicated in the pathophysiology of various other diseases, including Alzheimer's disease (Neddens *et al.*, 2018) and Parkinson's disease (Koyano *et al.*, 2014). As a consequence, there is increased attention to cellular signaling in biomedical applications, motivating researchers to study phosphorylation at larger scales (Hernandez-Armenta *et al.*, 2017).

In response to the growing need for large-scale monitoring of phosphorylation, advanced mass spectrometry (MS)-based phosphoproteomics technologies have exploded. These technologies enable simultaneous identification and quantification of thousands of phosphopeptides and phosphosites from a given sample (Yates III *et al.*, 2014). These developments result in the generation of data representing the phosphorylation levels of hundreds of thousands of phosphosites under various conditions across a range of biological contexts, including samples from human patients, cell lines, xenografts and mouse models (Liu and Chance, 2014). As compared to the widespread availability and sharing of genomic and transcriptomic data, public repositories of phosphoproteomic data are sparse, but growing. As a consequence, secondary or integrative analyses of phosphoproteomic data are less common. Despite tremendous advances in the last decade, a majority of the human phosphoproteome has not been annotated to date (Needham *et al.*, 2019). Technical issues such as noise, lower coverage, lower number of samples and low overlap between studies further complicate the analysis of phosphoproteomic data from a systems biology perspective (Liu and Chance, 2014).

In order to facilitate large-scale utilization of phosphoproteomic data, we introduced the notion of co-phosphorylation (Co-P) (Ayati *et al.*, 2019). The motivation behind this approach is to represent phosphorylation data in the form of relationships between pairs of phosphosites. Defining Co-P as the correlation between pairs of phosphosites across a range of biological states within a given study, we alleviate such issues as batch effects between different studies and missing identifications, while integrating phosphorylation data across multiple studies. Recently, we applied Co-P to the prediction of kinase-substrate associations (KSAs) (Ayati *et al.*, 2019) and unsupervised identification of breast cancer subtypes (Ayati *et al.*, 2020), showing that Co-P enables the effective integration of multiple datasets and enhances the reproducibility of predictions. In this article, we present a more comprehensive approach to investigate the functional information provided by Co-P, by focusing on multiple types of functional association among phosphorylation sites, considering a large number of datasets spanning multiple phenotypes, and assessing the value of integrating Co-P across different datasets.

Co-P is similar in spirit, but distinct and complementary to the notion of co-occurrence (Li *et al.*, 2017). Co-occurrence qualitatively assesses the relationship between the identification patterns of phosphosites in a broad range of studies. Co-P, on the other hand, quantitatively assesses the relationship between the phosphorylation levels of sites across a set of biological states (within a single study or by integrating different studies). Thus, co-occurrence captures high-level functional associations among phosphosites, whereas Co-P can also discover context-specific associations and provide insights into the dynamics of signaling interactions.

An important benefit of Co-P-based analysis is that it allows integration across datasets in a rather straightforward way as compared to direct integration. To directly integrate different datasets (where the unit of analysis is individual sites), normalization, standardization and correction for batch effects and other artifacts are needed, which significantly complicate the analysis. In contrast, by focusing on pairs of sites as units of analysis, we here accomplish cross-dataset integration by correcting for the number of different number of dimensions (samples) in different datasets.

In this article, we comprehensively characterize the relationship between Co-P and functional associations/interactions among protein phosphorylation sites. For this purpose, we systematically compare Co-P networks to networks that represent other functional relationships between proteins and phosphosites. These analyses serve two purposes: (i) validation of Co-P as a relevant and useful tool for inferring functional relationships between proteins and (ii) generation of knowledge on the basic principles of post-translational regulation of proteins and the functional relationships between them.

2 Materials and methods

2.1 Phosphoproteomic datasets

We analyze nine different MS-based phosphoproteomics data representing cancer and non-cancer diseases.

- **BC1 (breast cancer):** Huang *et al.* (2017) used the isobaric tags for relative and absolute quantification (iTRAQ) to identify 56 874 phosphosites in 24 breast cancer PDX models.
- **BC2 (breast cancer):** This dataset was generated to analyze the effect of delayed cold ischemia on the stability of phosphoproteins in tumor samples using quantitative LC-MS/MS. The phosphorylation level of the tumor samples was measured across three time points (Mertins *et al.*, 2014). The dataset includes 8150 phosphosites mapping to 3025 phosphoproteins in 18 breast cancer xenografts.
- **BC3 (breast cancer):** The NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) conducted an extensive MS-based phosphoproteomics analysis of TCGA breast cancer samples (Mertins *et al.*, 2016). After selecting the subset of samples to have the highest coverage and filtering the phosphosites with missing intensity values in those tumors, the remaining data contained intensity values for 11 018 phosphosites mapping to 8304 phosphoproteins in 20 tumor samples.
- **OC1 (ovarian cancer):** This dataset was generated by the same study as BC2, using the same protocol. The dataset includes 5017 phosphosites corresponding to 2425 phosphoproteins in 12 ovarian tumor samples.
- **OC2 (ovarian cancer):** The Clinical Proteomic Tumor Analysis Consortium conducted an extensive MS-based phosphoproteomic of ovarian HGSC tumors characterized by The Cancer Genome Atlas (Zhang *et al.*, 2016). We filtered out the phosphosites with missing data and also selected a subset of tumors to maximize the number of phosphosites. This resulted in a total of 5017 phosphosites from 2425 proteins in 12 tumor samples.
- **Colorectal cancer (CRC):** Abe *et al.* (2017) performed immobilized metal-ion affinity chromatography-based phosphoproteomics and highly sensitive pY proteomic analyses to obtain data from four different colorectal cancer cell line. The dataset included 5357 phosphosites with intensity values across 12 different conditions. These phosphosites map to 2228 phosphoproteins.
- **Lung cancer (LC):** Wiredja (2018) performed a time course label-free phosphoproteomics on non-small lung cancer cell lines across 1, 6 and 24 h after applying two different treatments of PP2A activator and MK-AZD, resulting in total of six samples. They reported phosphorylation levels for 5068 phosphosites, which map to 2168 proteins.
- **Alzheimer's disease (AD):** LC-MS/MS phosphoproteomics was performed on eight individual AD and eight age-matched control postmortem human brain tissues. The dataset contains 5569 phosphosites mapping to 2106 proteins (Dammer *et al.*, 2015).
- **Retinal pigmented epithelium (RPE):** MS-based phosphoproteomics was performed on three cultured human-derived RPE-like ARPE-19 cells which were exposed to photoreceptor outer segments (POS) for different time periods (0, 15, 30, 60, 90 and 120

min) (Chiang *et al.*, 2017). The dataset contains 1016 phosphosites mapping to 619 proteins in 18 samples.

The overlap between the sites identified in these nine datasets is shown in Supplementary Figure S1. As seen in the figure, the pairwise overlap between the breast cancer datasets is relatively higher as compared to the overlap between breast cancer datasets and other datasets, while one breast cancer dataset (BC3) also shares a high degree of overlap with an ovarian cancer dataset (OV2). This is expected as the mechanistic overlap between breast cancer and ovarian cancer is well established. Interestingly, however, we do not observe significant overlap between the two ovarian cancer datasets, demonstrating that the dropouts in proteomics may lead to a quite incomplete view of the phosphorylation events in the context of a specific phenotype.

2.2 Functional networks

To assess the functional relevance of Co-P, we use networks of functional relationships/associations between pairs of phosphorylation sites. For this purpose, we consider four types of functional networks:

Kinase-substrate associations. We use PhosphoSitePLUS (PSP) (Hornbeck *et al.*, 2015) as a gold-standard dataset for KSAs. PSP reports 9699 associations among 347 kinases and 6906 substrates. We use these associations to construct a 'shared-kinase network' of phosphorylation sites, in which nodes represent phosphosites and edges represent the presence of at least one kinase that phosphorylated both sites. The associations obtained from PSP lead to a shared-kinase network of 6906 phosphosite nodes and 881 685 shared-kinase edges.

Protein–protein interaction. Protein–protein interaction networks (PPI) encode physical and functional associations among proteins and thus have been used commonly for various inference tasks in cellular signaling. These tasks include the identification of signaling pathways (Wagner *et al.*, 2019), identification of pathways that are mutated in cancers (Ruffalo *et al.*, 2015), prediction of the effect of mutations on protein interactions (Rodrigues *et al.*, 2019) and prediction of KSAs (Horn *et al.*, 2014). Here, we use the PPIs that are provided in STRING database (Szklarczyk *et al.*, 2014) with high confidence (combined score $\geq$0.95). Overall, there are 61 452 high-confidence interactions among 8987 proteins. For each of the nine datasets, we use these PPIs to construct an interaction network among the sites identified in that dataset. In this network, each node represents a phosphosite and each edge represents an interaction between the two proteins that harbor the respective sites.

Evolutionary and functional associations. PTMCode is a database of known and predicted functional associations between phosphorylation and other post-translational modification sites (Minguez *et al.*, 2015). The associations included in PTMCode are curated from the literature, inferred from residue co-evolution, or are based on the structural distances between phosphosites. We utilize PTMcode as a direct source of functional, evolutionary and structural associations between phosphorylation sites. In the PTMcode network, there are 96 519 phosphosite nodes and 4 661 285 functional association edges between these phosphosites.

Phosphosite-specific signaling pathways. We use PTMsigDB as a reference database of site-specific phosphorylation signatures of kinases, perturbations and signaling pathways (Krug *et al.*, 2019). While PTMSigDB provides data on all post-translational modifications, we here use the subset that corresponds to phosphorylation. There are 2398 phosphosites that are associated with 388 different perturbation and signaling pathways. We represent these associations as a binary network of signaling-pathway associations among phosphosites, in which an edge between two phosphosites indicates that the phosphorylation of the two sites is involved in the same pathway. The resulting network contains 6276 edges between 2398 phosphosite nodes.

For each functional network, the number of nodes/edges that overlap with our nine phosphoproteomic datasets are shown in Table 1. Supplementary Figure S1 also shows the number of common phosphosites among the datasets.

2.3 Assessment of Co-P

For a given phoshoproteomic dataset, we define the vector containing the phosphorylation levels of a phosphosite across a number of biological states as the *phosphorylation profile* of a phosphosite. For a pair of phosphosites, we define the Co-P of the two sites as the statistical association of their phosphorylation profiles. To assess statistical association, we refer to the rich literature on the assessment of gene co-expression based on mRNA-level gene expression (Carter *et al.*, 2004), and consider Pearson correlation (Ballouz *et al.*, 2015), biweight-midcorrelation (Song *et al.*, 2012) and mutual information (Meyer *et al.*, 2008). Since our experiments suggest that the different measures of association lead to similar results (data not shown), we use Pearson correlation as a simple measure of statistical association in our experiments.

We use the datasets described in the previous section to characterize Co-P in relation to the functional, structural and evolutionary relationships between sites and proteins encoded in the functional networks. For this analysis, we investigate the correspondence between Co-P in each individual MS-based phosphoproteomics dataset and each functional network.

2.4 Predicting functional association using Co-P

2.4.1 Integration of Co-P networks across datasets

Since Co-P can potentially capture context-specific, as well as universal functional relationships among phosphorylation sites, we also investigate the functional relevance of Co-P across different datasets. While integrating Co-P across multiple datasets, the number of samples (i.e. the number of dimensions used to compute the correlation) in each dataset is different. For this reason, we use the adjusted R-squared (Miles, 2014) (denoted R^2_d) to remove the effect of number of dimensions from dataset-specific Co-P between pairs of phosphosites:

$$R^2_d(i,j) = 1 - \frac{n_d - 1}{n_d - 2}\left(1 - c_d(i,j)^2\right). \quad (1)$$

Here, $c_d(i,j)$ denotes the Co-P (measured by Pearson correlation) in dataset $d \in D$ with n_d samples.

In mass-spectrometry-based phosphoproteomics, the overlap between the phosphorylation sites that are identified across different studies is usually low and drop-outs can be common (Liu and Chance, 2014). Specifically, for the nine datasets we use in our computational experiments, there are only 17 phosphosites that are identified in all studies. Consequently, to preserve the scope of our cross-dataset analysis, we use all sites that are identified in at least one study. For this purpose, we develop a measure of cross-dataset Co-P that can integrate the Co-P scores computed on an arbitrary number of datasets. The principle behind our formulation of an integrated Co-P score is that Co-P witnessed by multiple datasets should be rewarded, but site pairs should not be penalized for lack of witnesses. Thus, to handle missing data without introducing bias, we set $R^2_d(i,j) = 0$ if phosphosite i or phosphosite j is not present in dataset d. Subsequently, we compute the integrated Co-P between sites i and j as follows:

$$c_{\text{integrated}}(i,j) = 1 - \prod_{d \in D}(1 - R^2_d(i,j)). \quad (2)$$

Observe that, $0 \leq c_{\text{integrated}}(i,j) \leq 1$, where the minimum value is realized if the two sites are never identified in the same dataset or their phosphorylation levels have zero correlation if they are identified together. If the phosphorylation levels of two sites exhibit perfect correlation in at least one dataset, then $c_{\text{integrated}} = 1$. Finally, as the number of datasets on which both sites are identified goes up, $c_{\text{integrated}}$ also tends to go up. Thus, $c_{\text{integrated}}$ can be thought of as a measure of both co-occurrence (Li *et al.*, 2017) and Co-P (Ayati *et al.*, 2019), since it captures both the tendency of the sites being identified in similar contexts, as well as the relationship between their dynamic ranges of phosphorylation.

Table 1. Descriptive statistics of the phosphoproteomic datasets used in our computational experiments and their overlap with functional networks

	Descriptive statistics			Overlap with functional networks			
Dataset	No. of samples	No. of phosphosites	No. of proteins	Shared kinase	PPI	PTMCode	PTMSigDB
BC1	24	15 780	4539	805	7632	4437	138
				27 791	142 077	15 335	2547
BC2	18	8150	3025	243	1639	1007	54
				2723	16 541	1811	429
BC3	20	11 472	3312	553	4491	3014	119
				13 123	45 911	9127	2226
OC1	12	5017	2425	414	2450	1318	74
				7174	17 584	2580	1032
OC2	12	4802	2230	157	818	510	32
				1114	4764	685	158
CRC	12	5352	2228	320	1663	1240	51
				6237	17 573	2715	421
LC	6	5068	2168	380	2036	1238	64
				6493	17 884	2919	588
AD	8	5569	1559	238	1743	941	44
				3637	19 075	3182	228
RPE	18	1016	619	120	371	193	31
				931	1667	320	216

Note: For each dataset, the number of samples, the number of phosphorylation sites that were identified and the number of proteins that are spanned by these sites are shown. For each dataset and functional network pair, the number in the first row shows the number of sites with at least one interaction in the functional network and the second row shows the number of interactions in the functional network with both sites present in the corresponding dataset.

2.4.2 Prediction of functional associations

While constructing the Co-P networks, we compute a Co-P score for each pair of phosphosites, namely $c_d(i,j)$ for individual dataset d and $c_{\text{integrated}}(i,j)$ for the integrated network. We then sort the pairs according to this Co-P score and apply a moving threshold to generate a series of Co-P networks with increasing number of edges. We use recall and precision to evaluate the prediction performance. In this context, recall is the defined as the fraction of edges in the corresponding functional network that also exist in the Co-P network, whereas precision is defined as the fraction of edges in the Co-P network that also exist in the functional network. To provide a baseline for the predictive ability of the Co-P network, we also visualize the mean precision and 95% confidence interval for given recall for a random ranking of phosphosite pairs across 20 runs.

3 Results and discussion

3.1 Statistical significance of Co-P

To understand whether the notion of Co-P is biologically relevant, we first investigate the distribution of Co-P levels across all pairs of phosphosites identified within a study. The results of this analysis for nine datasets are shown in Figure 1. As seen in the figure, Co-P follows a normal distribution with mean close to zero (as would be expected if phosphorylation levels were drawn from a normal distribution) and the distribution is narrower (and likely less noisy) if more biological states (dimensions) are available. Based on the premise that Co-P can capture functionally relevant relationships, we hypothesize that the distribution of Co-P on real datasets would contain more positively and negatively correlated phosphosite pairs than would be expected at random. To test this hypothesis, we conduct permutation tests by permuting phosphorylation levels across the entire data matrix and compute the Co-P distribution on these randomized datasets. As seen in the figure, Co-P is concentrated more on strongly positive or strongly negative correlation levels for all datasets. For all datasets, the Kolmogorov–Smirnov (KS) test P-values for the difference between the observed Co-P distribution and permuted Co-P of distribution are $\ll 1E-9$. Similarly, the t-test P-values for the difference between the means of these distributions are $\ll 1E-9$ for all datasets except CRC. The mean difference and the 95% confidence interval for each dataset are provided below the histograms in the figure.

Furthermore, for most datasets (BC2, BC3 and OC1), we observe that the mean Co-P is clearly shifted to the right, as also indicated by the effect size and the significance of the t-statistic. For other datasets (BC1 and CRC), the difference between the means is close to zero and the corresponding t-statistics are less significant. However, even for these datasets, the KS-test indicates that the difference between the distributions is significant, and visual inspection of the histograms suggests that the histogram for observed Co-P values is always more spread. This observation suggests that these datasets also contain a large number of site pairs with negatively correlated phosphorylation levels. Clearly, as with a positive correlation, a negative correlation can also be indicative of a functional relationship between two phosphorylation sites.

Taken together, for all studies considered, there are more pairs of phosphosites with (positively or negatively) correlated phosphorylation levels than would be expected at random—hence a large fraction of these strong correlations likely stem from functional or structural relationships between the phosphosites.

3.2 Co-P of intra-protein sites

Results of previous studies indicate that the phosphorylation of different sites of the same protein can lead to different functional outcomes (Nishi *et al.*, 2014, 2015). Here, with a view to characterizing the functional diversity of the phosphorylation sites on a single protein, we compare the Co-P distribution of pairs of phosphosites that reside on the same protein (intra-protein sites) against the Co-P distribution of pairs of phosphosites that reside on different proteins (inter-protein sites). We also investigate the effect of proximity between phosphorylation sites on the functional relationship between the sites. The results of this analysis are shown in Figure 2.

As seen in Figure 2a, the distribution of Co-P for pairs of intra- and inter-protein sites are significantly different for most of the datasets (the mean differences and confidence intervals are provided in the figure, the P-values for the t-test as well as the KS-test are $\ll 1E-9$ for all datasets except RPE). We consistently observe that the Co-P of intra-protein sites (orange histogram) is shifted towards

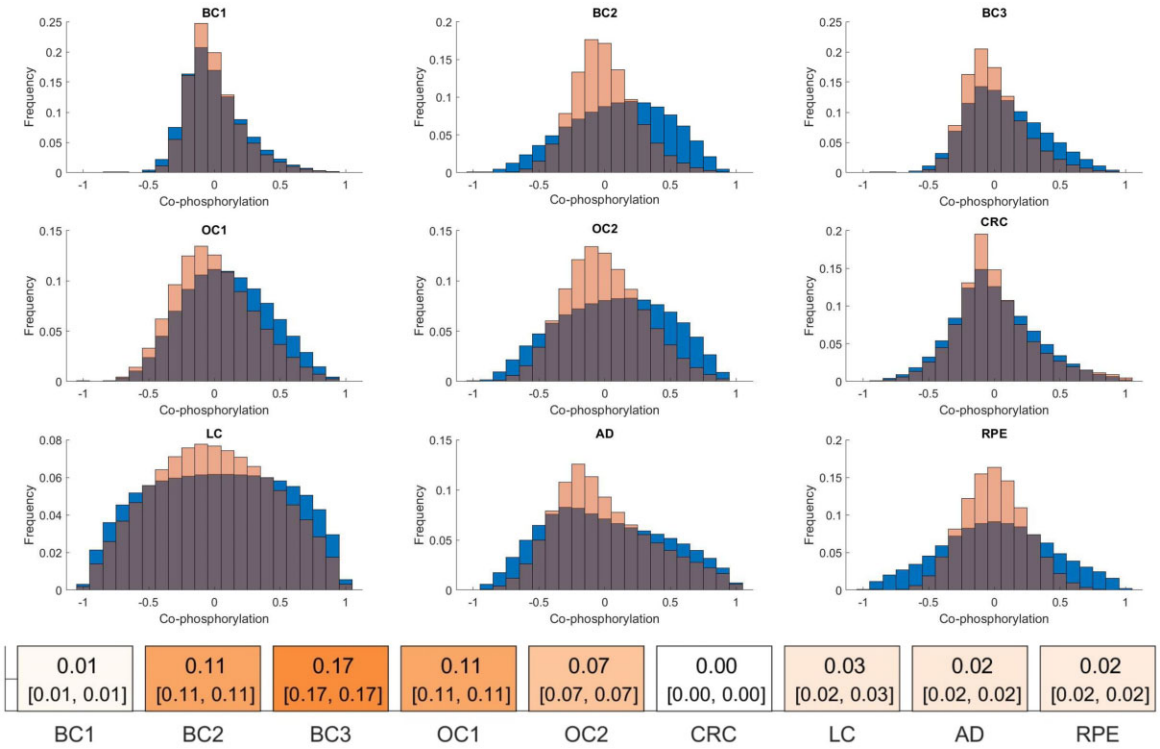

Fig. 1. Statistical significance of co-phosphorylation. Each panel compares the distribution of co-phosphorylation computed on a specific dataset against that computed on randomly permuted data for each dataset. The blue histogram shows the distribution of co-phosphorylation (the correlation between the phosphorylation levels) of all pairs of phosphosites identified in the corresponding study, the pink histogram in each panel shows the average distribution of co-phosphorylation of all pairs of phosphosites across 100 permutation tests. The permutation tests are performed by randomly permuting all entries in the phosphorylation matrix. The difference between the means of each pair of distributions is given on the colored boxes below. The 95% confidence intervals for the difference are provided in brackets (A color version of this figure appears in the online version of this article.)

high Co-P values. In other words, the phosphorylation levels of sites on the same protein are substantially more positively correlated as compared to the phosphorylation levels of sites on different proteins. While this observation can be partially explained by the impact of protein expression levels, a recent study showed that the protein abundance is overall not a strong indicator of phosphorylation fold changes (Arshad *et al.*, 2019). Thus, we hypothesize that intra-protein pairs exhibit higher Co-P because those pairs are more likely to be targeted by the same kinase/phosphates, or that they are more likely to be functionally associated by being part of the same signaling pathways.

Note that, the differences between the datasets in terms of the difference of intra- and inter-protein pairs are highly pronounced (e.g. we observe strong difference for BC1, BC3, OC1 while the difference is more modest for BC2, CRC and AD). While there can be biological reasons for this difference, it is important to note that each of these datasets come from different platforms, different sample types (e.g. patient-derived xenografts versus cell lines), different data collection procedures (e.g. protein degradation due to proteases in the sample) and are highly divergent in terms of availability of data (number of identified sites and number of samples). For this reason, the observed differences between the datasets can also be attributed to experimental, technological or statistical reasons. Further investigation is needed to elucidate potential biological differences between the systems that are represented by these datasets.

Next, we investigate whether the proximity on the protein sequence has any effect on the Co-P between two intra-protein sites. Since previous studies suggest that closely positioned sites tend to be phosphorylated by the same kinase (Schweiger and Linial, 2010), we expect a positive relation between sequence proximity and Co-P (i.e. we expect higher Co-P between close sites). To investigate this, we plot the relationship between the sequence proximity of intra-protein sites, and their Co-P. Figure 2b shows that the closely positioned intra-protein sites have higher Co-P. Thus, we observe that as the phosphosites get far away from each other, their Co-P typically reduces.

3.3 Co-P and functional association

Li *et al.* (2017) show that phosphorylation sites that are modified together tend to participate in similar biological processes. Here, focusing on the dynamic range of phosphorylation, we hypothesize that phosphosite pairs with correlated phosphorylation profiles are likely to be functionally associated with each other. To test this hypothesis, we investigate the relationship between Co-P and a broad range of functional associations. Since our results in Figure 2 suggest that there is a considerable difference between intra-protein and inter-protein sites in terms of their Co-P, we perform stratified analyses for intra- and inter-protein pairs. The results of this analysis are shown in Figure 3.

Shared-kinase pairs. First, we consider the Co-P of the substrates of the same kinase (i.e., shared-kinase pairs) as annotated by PhosphositePlus. As seen in Figure 3a, in all datasets, the Co-P distribution of shared-kinase pairs is significantly shifted upwards, i.e., sites that are targeted by the same kinase are likely to exhibit stronger correlation of phosphorylation as compared to arbitrary pairs. While this difference is more pronounced for intra-protein pairs, it is also evident for inter-protein pairs. AD and RPE have the largest positive shift in the Co-P distributions (0.37 and 0.43 respectively). This observation is also in line with previous findings in the literature (Arshad *et al.*, 2019; Ayati *et al.*, 2019). In Ayati *et al.* (2019), they used this characteristic to predict KSAs.

Phosphorylation sites on interacting proteins. It is well-established that proteins that are coded by co-expressed genes are likely to interact with each other (Ramani *et al.*, 2008). Here, we compare the PPI network and Co-P network to investigate the pattern of Co-P of pairs of phosphosites on interacting proteins. Note that, by definition, we only have this type of functional interaction

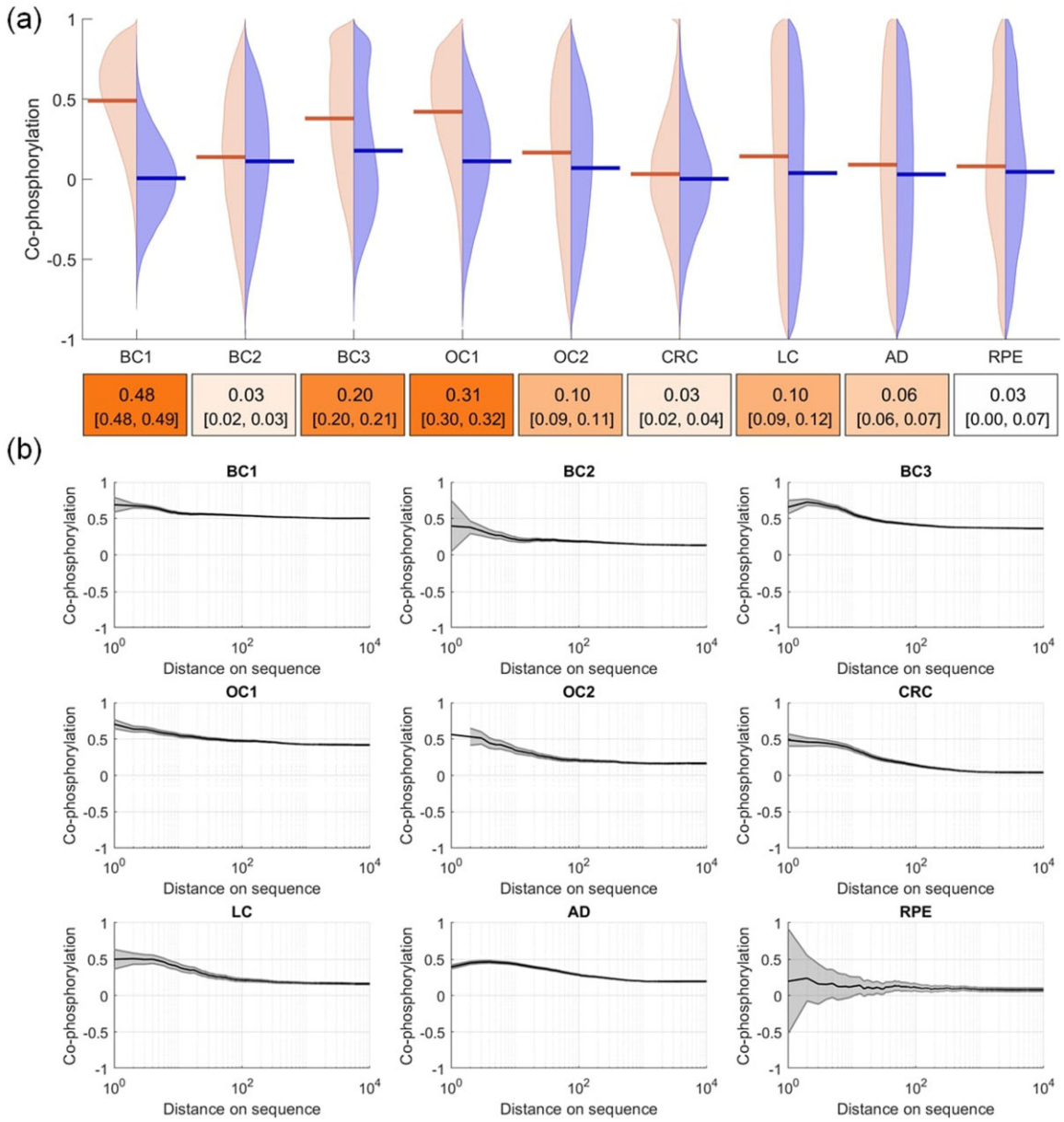

Fig. 2. Co-phosphorylation of phosphorylation sites on the same protein. (**a**) Comparison of the distribution of Co-P for all site pairs that are on the same protein (left histogram) versus Co-P for all pairs of sites on different proteins (right histogram). Each violin plot represents a different dataset. Colored boxes below indicate the mean difference between the intra-protein pairs and inter-protein pairs. Within brackets, 95% confidence intervals for the mean Co-P difference are provided. (**b**) The relationship between Co-P and sequence proximity for pairs of sites that reside on the same protein. Each panel shows a different dataset, the x-axis in each panel shows the distance between sites on the protein sequence (in terms of number of residues) and the y-axis shows the co-phosphorylation between pairs of sites in close proximity (up to the corresponding distance in x-axis). The curve and shaded area respectively show the mean Co-P and its 95% confidence interval (A color version of this figure appears in the online version of this article.)

for inter-protein sites. As seen in Figure 3b, in most of the datasets we consider (including BC1, BC3, OC1, OC2, LC and RPE), there is a clear upward shift of Co-P for sites on interacting proteins. OC1 and BC1, with 0.1 and 0.09 movement, have the largest shift among the datasets. This suggests that sites on interacting proteins are likely to be co-phosphorylated. Identification of the specific PPIs that are associated with Co-P can be potentially useful in elucidating the mechanisms of these PPIs.

Co-evolution of phosphorylation sites. The conservation status of the phosphosites has been used as a tool to measure PTM activity (Boekhorst *et al.*, 2008). It has been shown that co-evolving PTMs are likely to be functionally associated (Minguez *et al.*, 2012). Here, we investigate the relationship between co-evolution and Co-P of phosphosites. The results of this analysis are shown in Figure 3c. As seen in the figure, the association between co-evolution and Co-P is relatively weak compared to the association of Co-P with other functional networks. For some datasets such as BC1, OC1 LC and RPE, we observe that the co-evolving phosphosites on different proteins are more likely to have a higher Co-P. However, in some datasets such as BC2, CRC and AD the co-evolving phosphosites that are residing on the same proteins have higher Co-P.

Phosphorylation sites with common signaling pathways. Identifying the signaling pathways that are dysregulated in any perturbation and disease is crucial for understanding the underlying mechanism of diseases. Using PTMsigDB, we investigate the Co-P of phosphosites that are involved in the same pathway. As seen in Figure 3d, there is a considerable difference between the Co-P distribution of the phosphosites that are involved in the same signaling

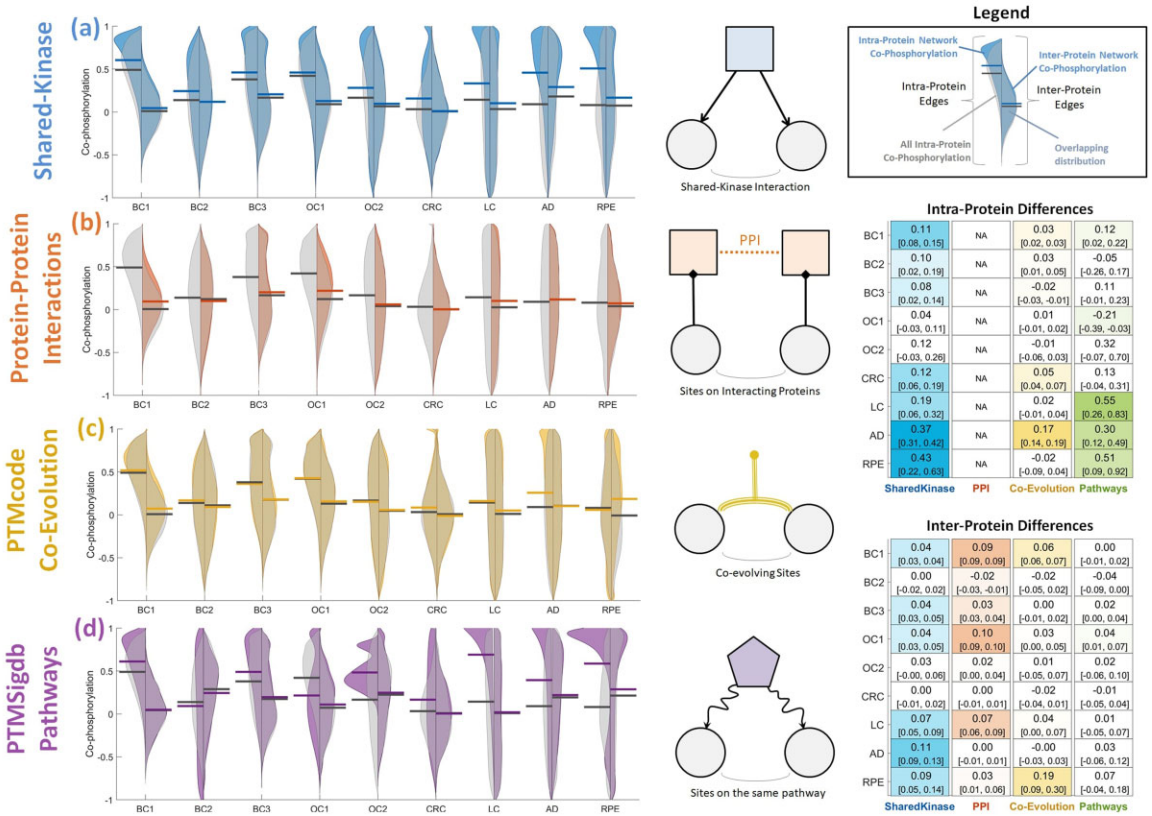

Fig. 3. The relationship between co-phosphorylation and functional association between pairs of phosphorylation sites. In each panel, the violin plots compare the distribution of Co-P for phosphosite pairs with an edge in the respective functional association network (colored histograms) against all phosphosite pairs (gray-colored histograms), across the nine datasets that are considered. For each dataset, the left/right violin plots respectively show intra-/inter-protein pairs. The black horizontal lines show the mean Co-P for all (intra- or inter-protein) phosphosite pairs, the colored horizontal lines show the mean Co-P for functionally associated pairs. The four type of functional association networks that are considered are illustrated on the right side of the corresponding violin plot. On the rightmost side, the colored tables show the mean difference between functionally associated pairs and all phosphosite pairs (corresponding to the gap between colored and black horizontal lines in the violin plots) for nine datasets and four functional networks. In each cell, the 95% confidence intervals for the mean difference are given within brackets (A color version of this figure appears in the online version of this article.)

pathway as compared to that of other phosphosite pairs. Similar to the results for shared-kinase pairs, this difference is more pronounced for intra-protein sites. PTMsigDB is a sparse database, and according to Table 1, there are not too many overlapping sites between PTMsigdb and these phosphoproteomics dataset. RPE and LC have the biggest shift, 0.51 and 0.55, respectively. However, there are just 588 and 216 phosphosites with annotations in the PTMsigDB for these datasets. BC1 and BC3 have the largest overlap with PTMsignDB with 2547 and 2226 phosphosites, and we observe a great positive shift among the intra-protein sites (0.12 and 0.11). It may suggest that the intra-protein phosphosites that are involving on the same pathways are more likely to have a higher Co-P compare to other phosphosites.

3.4 Predictive power of Co-P

Our results indicate that phosphosites involved in a common pathway or targeted by a common kinase are likely to be co-phosphorylated across different biological states. Motivated by this observation, we quantitatively assess the effectiveness of Co-P in predicting shared-kinase and shared-pathway associations between phosphorylation sites. While doing so, we also assess the contribution of Co-P evidence supported by multiple datasets to the reliability of predictions on functional association. For this purpose, we assess the predictive ability of Co-P computed using each individual dataset as well as the integrated Co-P computed using cross-dataset analysis. The results of this analysis are shown in Figure 4.

In the left panel of Figure 4, the precision-recall curves for the ability of the integrated network in predicting shared-kinase interactions (top-left panel) and shared-pathway interactions (bottom-left panel) are shown. As seen in the figure, the precision provided by the Co-P network is significantly higher than random ordering for both functional networks. We also observe that Co-P delivers higher precision for the shared-pathway network as compared to the shared-kinase network. This is likely because the information in PTMSigDB is sparser than the information in PhosphositePLUS.

The right panel of Figure 4 shows the odds ratio of a pair of sites being connected in the functional network as a function of the number of edges in the Co-P network. Namely, in these plots, a point on the x-axis corresponds to a Co-P network with a given number of edges. For this network, the value on the y-axis shows the odds ratio of the event that two sites are connected in the functional network given that they are connected in the Co-P network, as compared to a random pair of sites. As seen in the figure, for both shared-kinase and shared-pathway networks, the odds ratio provided by the integrated Co-P network is consistently higher than that provided by any individual network. While the odds ratio of sharing a kinase goes up to 100 and the odds ratio of being involved in the same pathway goes up to 30 for pairs of sites with Co-P, these odds ratios respectively converge to 4 and 2 as more edges are added to the integrated Co-P network. Overall, these results suggest that Co-P networks provide valuable information on the functional association of phosphorylation sites and this information becomes more reliable as Co-P information from more datasets are included in the Co-P network.

4 Conclusion

Mass-spectrometry techniques are advancing and more MS-based quantitative phosphoproteomics data are generated at high volumes.

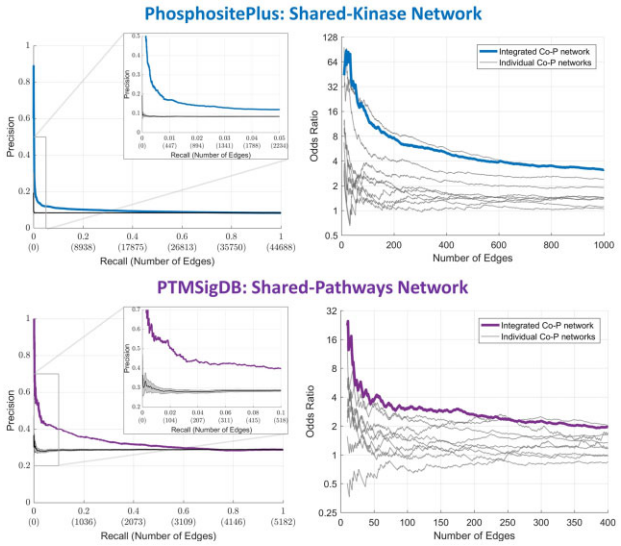

Fig. 4. The utility of Co-P in predicting the functional association of phosphorylation sites. (Left) Precision-recall curve showing the functional predictivity of the Co-P network obtained by integrating nine different phosphoproteomic datasets. The shaded gray area shows the 95% confidence interval for the mean precision-recall curve for permutation tests obtained by randomly ranking pairs of phosphosites (across 20 runs). (Right) Comparison of the predictive performance of the integrated Co-P network against the nine individual Co-P networks obtained using each dataset separately. The x-axis shows the number of pairs that are included in the Co-P network, the y-axis shows the odds ratio of being connected in the respective functional network given that the sites are connected in the Co-P network. (Top) Predicting shared-kinase associations. (Bottom) Predicting shared-pathway associations

However, integration of these data may be challenging since the data are generated in different labs and in different contexts. By focusing on the relationships between pairs of phosphosites as opposed to their individual phosphorylation levels, Co-P networks can alleviate the dependency of computational and statistical methods on these factors. In this article, we systematically investigated the relationship between Co-P and broad range of known functional associations between proteins and phosphorylation sites. Our results showed that the sites that are functionally associated tend to exhibit higher levels of Co-P. Our results also showed that the integration of Co-P networks across different datasets can improve the predictivity of Co-P, as compared to analyzing the datasets in isolation.

Although these results provide considerable novel insights, there are still some limitations to the power of Co-P. For example, Co-P can be reliably assessed for datasets with a relatively high number of samples (e.g. six) (Ayati *et al.*, 2019). Moreover, the limited overlap between LC/MS studies poses significant challenges to the integration of phosphorylation data from different studies. However, as our results show, Co-P can be a useful tool for integrating multiple datasets and increasing the reliability of functional association that is inferred from Co-P patterns. In addition, most of the datasets utilized in this study come from studies that aim to investigate a specific disease. While differential network analysis between Co-P and non-disease-specific Co-P may provide further insight into the rewiring of cellular signaling networks, additional non-disease-specific datasets are needed for this purpose. Therefore, as the scale and scope of LC/MC studies grow, the application of Co-P will become more valuable.

Funding

This work was supported by National Institutes of Health [R01-LM012980] from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

Abe,Y. *et al.* (2017) Deep phospho-and phosphotyrosine proteomics identified active kinases and phosphorylation networks in colorectal cancer cell lines resistant to cetuximab. *Sci. Rep.*, **7**, 1–12.

Arshad,O.A. *et al.* (2019) An integrative analysis of tumor proteomic and phosphoproteomic profiles to examine the relationships between kinase activity and phosphorylation. *Mol. Cell. Proteomics*, **18**(8 suppl 1), S26–S36.

Ayati,M. *et al.* (2020) Co-phosphorylation networks reveal subtype-specific signaling modules in breast cancer. *Bioinformatics*, **37**, 221–228.

Ayati,M. *et al.* (2019) Cophosk: a method for comprehensive kinase substrate annotation using co-phosphorylation analysis. *PLoS Comput. Biol.*, **15**, e1006678.

Babur,O. *et al.* (2021) Causal interactions from proteomic profiles: molecular data meet pathway knowledge. *Patterns*, **2**, 100257.

Ballouz,S. *et al.* (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, **31**, 2123–2130.

Boekhorst,J. *et al.* (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.*, **9**, R144.

Butrynski,J.E. *et al.* (2010) Crizotinib in ALK-rearranged inflammatory myofibroblastic tumor. *N. Engl. J. Med.*, **363**, 1727–1733.

Carter,S.L. *et al.* (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, **20**, 2242–2250.

Chiang,C. *et al.* (2017) Quantitative phosphoproteomics reveals involvement of multiple signaling pathways in early phagocytosis by the retinal pigmented epithelium. *J. Biol. Chem.*, **292**, 19826–19839.

Cohen,P. (2001) The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur. J. Biochem.*, **268**, 5001–5010.

Dammer,E.B. *et al.* (2015) Quantitative phosphoproteomics of Alzheimer's disease reveals cross-talk between kinases and small heat shock proteins. *Proteomics*, **15**, 508–519.

Halim,V.A. *et al.* (2013) Comparative phosphoproteomic analysis of checkpoint recovery identifies new regulators of the DNA damage response. *Sci. Signal.*, **6**, rs9.

Hernandez-Armenta,C. *et al.* (2017) Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, **33**, 1845–1851.

Horn,H. *et al.* (2014) KinomeXplorer: an integrated platform for kinome biology studies. *Nat. Methods*, **11**, 603–604.

Hornbeck,P.V. *et al.* (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.

Huang,K. *et al.* (2017) Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. *Nat. Commun.*, **8**, 1–17.

Koyano,F. *et al.* (2014) Ubiquitin is phosphorylated by PINK1 to activate parkin. *Nature*, **510**, 162–166.

Krug,K. *et al.* (2019) A curated resource for phosphosite-specific signature analysis. *Mol. Cell. Proteomics*, **18**, 576–593.

Li,Y. *et al.* (2017) Co-occurring protein phosphorylation are functionally associated. *PLoS Comput. Biol.*, **13**, e1005502.

Liu,Y., Chance,M.R. (2014) Integrating phosphoproteomics in systems biology. *Comput. Struct. Biotechnol. J.*, **10**, 90–97.

Mertins,P. *et al.*; NCI CPTAC. (2016) Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, **534**, 55–62.

Mertins,P. *et al.* (2014) Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics*, **13**, 1690–1704.

Meyer,P.E. *et al.* (2008) Minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Miles,J. (2014). *R Squared, Adjusted R Squared. Wiley StatsRef: Statistics Reference Online*.

Minguez,P. *et al.* (2015) PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins. *Nucleic Acids Res.*, **43**, D494–D502.

Minguez,P. *et al.* (2012) Deciphering a global network of functionally associated post-translational modifications. *Mol. Syst. Biol.*, **8**, **599**.

Neddens,J. *et al.* (2018) Phosphorylation of different tau sites during progression of Alzheimers disease. *Acta Neuropathol. Commun.*, **6**, 52.

Needham,E.J. *et al.* (2019) Illuminating the dark phosphoproteome. *Sci. Signal.*, **12, eaau8645**.

Nishi,H. *et al.* (2015) Crosstalk between signaling pathways provided by single and multiple protein phosphorylation sites. *J. Mol. Biol.*, **427**, 511–520.

Nishi,H. *et al.* (2014) Physicochemical mechanisms of protein regulation by phosphorylation. *Front. Genet.*, **5**, **270**.

Ramani,A.K. *et al.* (2008) A map of human protein interactions derived from co-expression of human mrnas and their orthologs. *Mol. Syst. Biol.*, **4**, 180.

Rikova,K. *et al.* (2007) Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell*, **131**, 1190–1203.

Rodrigues,C.H. *et al.* (2019) mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Res.*, **47**, W338–W344.

Ruffalo,M. *et al.* (2015) Network-based integration of disparate omic data to identify silent players in cancer. *PLoS Comput. Biol.*, **11**, e1004595.

Schweiger,R., Linial,M. *et al.* (2010) Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biol. Direct*, **5**, 6.

Song,L. *et al.* (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, **13**, 328.

Szklarczyk,D. *et al.* (2014) String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

Wagner,M.J. *et al.* (2019) Reconstructing signaling pathways using regular language constrained paths. *Bioinformatics*, **35**, i624–i633.

Wilhelm,M. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.

Wiredja,D. (2018). Phosphoproteomic characterization of systems-wide differential signaling induced by small molecule PP2A activation. PhD thesis, Case Western Reserve University.

Yates,J.R. III. *et al.* (2014) Phosphoproteomics. *Anal. Chem.*, **86**, 1313–1313.

Yılmaz,S. *et al.* (2021) Robust inference of kinase activity using functional networks. *Nat. Commun.*, **12**, 1–12.

Zhang,H. *et al.*; CPTAC Investigators (2016) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, **166**, 755–765.

Zhou,C. *et al.* (2011) Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer: a multicentre, open-label, randomised, phase 3 study. *Lancet Oncol.*, **12**, 735–742.