Assessing the impact of single nucleotide variants on kinase–substrate phosphorylation

Master's Thesis

Submitted by: Omar Wagih
Supervised by: Gary D. Bader

Department of Computer Science, University of Toronto, Toronto, Canada
Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, Toronto, Canada

E-mail: omar.wagih@utoronto.ca

Acknowledgements

A profound gratitude and deep regards are due to my supervisor, Prof. Gary Bader, for his exemplary guidance and constant encouragement over the course of this thesis. The help and guidance given by him time and time again shall carry me far in the journey of life, on which I am about to embark.

This thesis would not have been possible without the support and guidance of my mentor and friend, Jüri Reimand, who has helped me through the tough times of this project. I would also like to thank colleagues and associates who have helped me in completing this task, and made my experience at Bader laboratory worthwhile.

Last but not least, a deep sense of gratitude is due to the loved members of my family: my wonderful parents, Nagat Mousa and Mohamed Wagih, my dear brother, Ramy, and my sister, Manar, for their cordial support and words of wisdom that kept me going.

Omar Wagih

The following thesis was partially published in the following article:

Reimand, J., Wagih, O., and Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).

Abstract

Phosphorylation is a prominent post-translation modification, which is carried our by protein kinases impacting a wide range of cellular processes. Kinases are among the most important groups of cancer drug targets in the 21st century. Aberrant phosphorylation of kinase substrates can have important downstream effects in signaling pathways and transcription factor regulation. Thus, identifying the functional impact of somatic cancer mutations on kinase binding sites can help decipher oncogenic signaling mechanisms and contribute to drug development. We recently found phospho-signaling sites to be enriched in cancer driver mutations. Here, we hypothesize that many phosphorylation-related single nucleotide variants (pSNVs) precisely modify kinase-binding sites and lead to phosphorylation network rewiring. Three types of rewiring events can occur: gain-of-signaling mutations introduce new kinase binding sites, loss-of-signaling mutations disrupt kinase binding at a site, and switch-of-signaling mutations cause a change of specificity from one kinase to another. We developed the computational pipeline MIMP to identify such rewiring events systematically. We tested MIMP on the TCGA pan-cancer collection of 250,000 mutations in $>$3,000 cancer genomes and 12 tumor types. We found that half of the 16,000 mutations that lie within phosphosites significantly alter kinase specificity. This effect is apparent in multiple known cancer genes such as CTNNB1 and TP53 as well as numerous novel cancer genes. Furthermore, we validated our predictions computationally with network and pathway analyses. Our approach provides experimentally testable signaling-associated hypotheses about known cancer driver mutations, and helps reveal novel cancer genes with oncogenic signaling mutations.

1 Introduction

Cellular signaling networks are complex systems of interacting proteins that are ultimately encoded in genomes. Computational analysis of protein-coding genomes shall therefore reveal blueprints of interaction networks and help interpret disease-associated genetic variants that affect signaling interfaces in proteins (*1*). Understanding how genetic variation induces changes in cellular signaling will reveal experimentally testable, mechanistic hypotheses, paving the way for precision medicine and improved drug target development.

While genomic analysis of interaction networks is generally difficult, a subset of protein–protein interactions (PPIs) mediated by peptide recognition modules (PRMs) can be feasibly modeled with computational methods. PRMs serve as adaptors, which bind certain linear motifs, usually in response to external or internal cues. In particular, protein phosphorylation is a system of post-translational modifications (PTMs) involving writers (protein kinases), readers (such as SH3 proteins) and erasers (phosphatases). Protein phosphorylation involves the reversible addition of a phosphate group, obtained from an ATP molecule, to serine (S), threonine (T), or tyrosine (Y) residues. This process can have multiple functional outcomes: it can modulate proteosomal degradation of proteins; affect the protein's electrochemical stability, leading to alternative folding; or inhibit or induce interactions with other proteins. Through such mechanisms, phosphorylation signaling is central to cellular processes such as transcription, proliferation, and apoptosis (*2–4*). As these processes are altered in human diseases, integration of kinase signaling and disease mutation data will lead to an improved understanding of disease mechanisms. Protein kinases have also become among the most important groups of cancer drug targets in the 21st century (*5–8*) and are one of the largest classes of proteins containing PRMs, with over 500 members in the human genome (*9*) and abundant datasets characterizing their sequence specificity (*10–12*).

Functional analysis of somatic mutations in cancer genomes and discovery of driver mutations are some of the most central goals of current cancer research (*13*). Driver mutations provide selective advantages to cells and are interesting for biomarker discovery and therapy development, while passenger mutations are functionally neutral and occur due to instability of cancer genomes. Each solid tumor is believed to contain two to eight driver mutations and tenfold more passengers (*14*), posing a challenge to cancer driver discovery. However, as the selective advantages of tumor cells such as proliferation, apoptosis, and metabolism are intrinsically mediated by signaling pathways involving phosphorylation, it is likely that cancer drivers involve precise modifications of kinase networks. We recently conducted a systematic analysis of phosphorylation-associated single nucleotide variants (pSNVs) in 800 tumor genomes, and found that phosphosite mutations occur in cancer genes and pathways and are informative of clinical outcome (*15*). In an extended analysis of >3,000 cancer genomes, we found that 90% of samples contained 16,000 pSNVs and predicted that 50% of these mutations severely affect kinase binding sites, including known cancer genes such as TP53 and CTNNB1 (*16*). These studies underline the importance of signaling-associated mutations in tumor biology and encourage further interpretation of cancer mutations in a signaling network context.

Indeed, the concept of single amino acid substitutions affecting kinase–substrate binding in diseases has been previously explored in small and large scales (*17–25*). However, due to a lack of phosphoproteomics data at the time, comprehensive characterization of individual kinase–substrate relationships was a difficult task. As a result, kinase–substrate specificities were modeled at the kinase group or family level (*9*). For example, Ren *et al.* studied four kinase groups (*18*) and Ryu *et al.* studied six kinase groups and 18 kinase families (*19*), none of which models kinase specificity on an individual basis. Other studies explored phosphorylation data in up to 8,400 proteins (*17, 20*). These numbers have significantly increased over the past few years, suggesting that a more comprehensive analysis may unveil unexplored biological

hypotheses. More importantly, studies developing a method will, at most, provide a database containing sites affected by a predefined set of mutations. These include PhosSNP (*18*) and PhosphoVariant (*19*), and are problematic for those wishing to analyze their own data.

To tackle these issues, we developed a systematic pipeline, MIMP, to characterize genetic missense variants such as cancer mutations that precisely and specifically alter kinase binding sites in proteins. MIMP makes use of phosphorylation data for over 360 kinases in over 10,500 proteins to comprehensively assess kinase rewiring. We assume that kinase binding to an experimentally determined protein site depends on the presence of several critical amino acids. As these residues are changed in disease mutations, alterations in kinase binding specificity potentially lead to rewiring (Fig. 1). We tested our approach on a recent large collection of cancer mutations and discovered cancer genes and pathways with frequent rewiring pSNVs. MIMP is freely available as an R package from `https://github.com/omarwagih/mimp` under the LGPL.

2 Results

2.1 Construction of high-confidence kinase binding specificity models from proteomics data

To study phosphorylation networks, we collected 88,253 phosphosites in 10,504 human proteins from four phosphoproteomic databases (*10–12, 26*). A minority of phosphosites (14%, 12,381) were assigned a binding kinase in earlier experiments, which we used to construct kinase binding specificity models. Many kinases had only a small number of sites. As such, we only considered kinases with 10 or more available phosphosites (Suppl. Fig. 1). In total, we generated kinase specificity models for 362 kinases, including 310 ST kinases and 52 Y kinases. Kinase specificity models were established as position weight matrices (PWMs), which contained a weight representing the likelihood of the kinase binding a particular amino acid for every position flanking the phosphosites ($\pm$7 residues). In order to improve the performance of our models, we designed an iterative refinement procedure, which discarded sequences used for PWM construction that did not correspond to the motif's general pattern (see Methods).

To validate the performance of our kinase models, we carried out a 10-fold cross-validation analysis, wherein binding models were constructed from a subset of phosphosite sequences and the remaining sequences were used for classification versus non-target sequences (see Methods). We computed the area under the receiver operating characteristic (AUC) curve and found that most of our models (333/362, 92%) performed as satisfactory classifiers of phosphosite sequences with AUCs $>$0.75 (Suppl. Fig. 2). We discarded 29 PWMs below that criterion. We also visually validated our models with that of other studies (*26, 27*) and found a good correspondence in the case of most models (Suppl. Fig. 3). This analysis provided a comprehensive set of kinase models for evaluating cancer mutations in a kinase-rewiring context.

2.2 MIMP—systematic prediction of network-rewiring mutations using kinase specificity models

We developed the statistical method MIMP to integrate kinase binding models and protein missense mutations for the prediction of network rewiring events. The method is based on the assumption that kinase binding specificity is determined by a small number of critical residues around the central phosphorylatable residue. Systematic analysis of all kinase models in our datasets confirms our assumption and shows that most specificity information is encoded in the proximal ± 3 residues around the central residue of the phosphorylation site (Fig. 3a). Thus, we propose that specific and precise disease mutations affecting phosphosite-flanking residues critically impact binding specificity, potentially altering kinase–substrate interactions (Fig. 1). Based on this assumption, three types of rewiring events can occur due to disease mutations. Gain-of-signaling, caused by a single mutation that introduces a new kinase binding site at a high-affinity position of a potential kinase target site; loss-of-signaling, caused by a single mutation that disrupts kinase binding at a site through mutating a residue critical for affinity; and switch-of-signaling, occurring if a single mutation causes a specificity change from one kinase to another, by replacing an amino acid critical for one kinase to that of another kinase (Fig. 1).

Our method is based on the matrix similarity score (MSS) that was previously developed for studying transcription regulatory motifs (*28*). Here, the MSS is used to quantify the similarity between a given phosphosite sequence and the kinase specificity model, reflecting the likelihood of the kinase to bind that sequence. The MSS ranges from 0 to 1, where 1 represents a perfect match and 0 represents no binding. In order for a rewiring event to occur, the MSS of the wildtype phosphosite must be significantly different from that of the mutant phosphosite. We defined a confidence interval for the MSS as $[\alpha-,\alpha+]$ using scores of the foreground and background sequences (see Methods). This helps in establishing gains and losses by allowing us to filter out dubious predictions, which may fall under either distribution (Fig. 2). There-

fore, we consider a loss-of-signaling event to occur if the MSS of the wildtype phosphosite is greater than $\alpha+$ and its corresponding MSS of the mutant phosphorylation is less than $\alpha-$, and vice-versa for a gain-of-signaling event. Switch-of-signaling occurs as a combination of both events.

2.3 MIMP predicts numerous network-rewiring pSNVs in the TCGA pan-cancer dataset

We tested our method on the TCGA pan-cancer dataset of 3,185 genomes and 236,367 missense SNVs from 12 tumor types. We used the collection of 16,891 pSNVs within ± 7 residues of phosphosites identified in our earlier pan-cancer analysis (*16*) to identify network-rewiring mutations. Our pipeline predicts that 9,545 of these pSNVs (55%) significantly impact 2,143 experimentally validated phsophorylation sites and lead to 53,784 network-rewiring events (14,216 gain-of-signaling events; 39,568 loss-of-signaling events), providing a resource of hypotheses for functional validation.

The majority of predicted kinase-rewiring pSNVs cause loss-of-signaling (7,595/9,545, 61%), as it is easier for pSNVs disrupting critical residues in binding to affect existing binding sites than to create novel ones. As a result, switch-of-signaling is the rarest event to occur. Furthermore, only 16 out of 9,545 (0.16%) rewiring pSNVs exist in > 4 of samples. Thus, the majority of rewiring pSNVs are infrequent, although highly influential in terms of predicted impact of multiple kinases. This suggests that many cancer mutations are rare, despite being highly specific to protein sites in signaling systems.

To construct a higher confidence kinase-rewiring network, we selected loss-of-signaling events where the kinase experimentally matches the binding site lost and gain-of-signaling events with the best-scoring kinase for mutated binding sites. The resulting dataset contained 1,285 rewiring events (674 gain-of-kinase events; 611 loss-of-kinase events) caused by 751

rewiring mutations, which were then clustered using Markov clustering (MCL) (*29*) and visualized as a network of rewiring events (Fig. 4).

2.4 Functional, pathway, and network analyses validate predicted rewiring mutations in kinase signaling

To computationally validate the rewiring mutations predicted by our pipeline, we carried out further analyses with complementary datasets. First, we investigated the properties of SNVs predicted to cause network rewiring. Expectedly, rewiring pSNVs mostly affect residues that flank ± 3 positions from the central site, likely because most information encoding kinase specificity is accumulated to these residues when averaging across all PWMs we use (Fig. 3b). In addition, on average, network-rewiring mutations appear to have a stronger functional bias than cancer mutations. We compared network-rewiring mutations to non-rewiring mutations using an ensemble of tools (phyloP (*30*), SIFT (*31*), PolyPhen2 (*32*), LRT (*33*), and MutationTaster (*34*)). Each mutation was assigned a value from 0 to 5, representing the number of tools that deemed it harmful. We found that rewiring mutations significantly enriched for high harmfulness (p-value < 0.0057; Binomial test) (Fig. 5a).

Next, we investigated the agreement of network-rewiring mutations with expression data, with the assumption that kinase–substrate interactions are more likely to be biologically valid if they are also co-expressed. We used the pan-cancer expression data, which measures expression in the samples where the mutations were studied. For each protein with a network-rewiring mutation, we counted the number of non-zero expression values of the predicted rewired kinase in samples containing the mutation. We found that pairs of kinases and network-rewiring mutations were co-expressed in 41,368 out of 43,415 cases (95%), which is a greater overlap than random chance would provide (Fig. 5b) (p-value $= 0$; Binomial test). Similarly, we used localization data to validate predictions: predicted rewired kinases and their target protein were

considered co-localized if they share at least one experimentally annotated localization. We found that co-localization occurred for 22,563 out of 37,941 cases (59%), which is significantly greater than random chance (Fig. 5b) (p-value $= 0$; Binomial test). This analysis suggests that many of our predictions are likely to occur *in vivo*.

Finally, we carried out a pathway enrichment analysis to characterize biological processes and pathways with rewiring mutations occurring more frequently than expected. This analysis revealed several functional themes that relate to hallmark cancer processes such as apoptosis, separation of sister chromatids, and splicing, as well as numerous cancer signaling pathways (Fig. 5c), validating the functional roles of network-rewiring mutations in cancer biology.

2.5 Cancer-relevant examples of kinase–substrate rewiring

2.5.1 Loss of TP53 phosphorylation by AURKB deregulates TP53 degradation

TP53 is a transcription factor and tumor suppressor that regulates apoptosis and cell cycle arrest. Mutations in TP53 are highly recurrent in cancer (50%) (*35, 36*). The phosphorylation site T284 is highly conserved across species (*37*) and exists in the DNA-binding domain of TP53. Phosphorylation of T284 by AURKB leads to negative regulation of TP53 (*37*). As a result, downstream pro-apoptotic targets of TP53 (p21, Bax, Puma, NOXA) are not activated (*38*). We predict the rewiring mutation TP53-R282W causes loss of AURKB (Fig. 7a), suggesting aberrant activation of downstream TP53 targets. While this may seem counterintuitive, as it implies continuous activation of proapoptotic genes, many biological processes rely on TP53 being tightly regulated. For example, germinal center B cells, a vital part of the immune system, which function to make antibodies against antigens, normally proliferate to maintain regular function. Suppression of TP53 is therefore crucial for B-cell proliferation (*39*). Thus, aberrant activation of TP53 may lead to disruptive effects on the proliferation of B cells.

2.5.2 Hotspot of mutations β-catenin disable GSK3β phosphorylation, required for degradation

CTNNB1, β-catenin, is a dual-function proto-oncogene, regulating the coordination of cell–cell adhesion and gene transcription. Mutations and overexpression of β-catenin are associated with many cancer types. CK1α phosphorylates β-catenin at S45, priming subsequent phosphorylation by GSK3β at residues 41, 37, and 33 (*40, 41*). β-catenin that is phosphorylated at residues 37 and 33 is ultimately recognized by the β-TrCP E3–ligase complex, ubiquitylated, and rapidly degraded by the 26S proteasome (*42, 43*).

Mutations S33P, S33Y, S33F, S37C, S37F, S37A, S37P, and S45C cause loss of phosphorylation by GSK3B at S29, S33, S37, and S41. While these mutations clearly disrupt the phosphorylation site itself, they also disrupt neighboring phosphorylation sites. For example, S37C mutation disrupts the phosphorylation site at S37 directly, while also disrupting S33 through its flanking region (Fig. 7b). Therefore, this cascade of phosphorylaton is clearly disrupted and thereby suppresses proteosomal degradation. This explains why high β-catenin levels act as a biomarker for many cancers. Previous studies have shown that S37C has been associated with abnormal β-catenin expression (*44*). Therapeutic targets that remedy the loss of phosphorylation would recover regular levels of β-catenin.

2.5.3 The role of PKC-ζ in NFE2L2 degradation

Nuclear factor (erythroid-derived 2)-like 2, also known as NFE2L2, is a transcription factor, which, along with the antioxidant response element (ARE), is responsible for antioxidant induction and the expression of several detoxifying genes (*45*), which are required for maintaining cell survival (*46*). Normally, NFE2L2 is bound to INrf2 (Keap1) in the cytoplasm, which maintains low levels of NFE2L2 through proteosomal degradation (*45, 47–49*). During oxidative stress, NFE2L2 releases itself from INrf2 and translocates to the nucleus, where it activates

its downstream targets along with ARE.

NFE2L2 S40 is highly conserved across species and lies within the INrf2 interaction domain (Nfe2). Phosphorylation of S40 by protein kinase C (PKC) (*45, 50–54*) is required for NFE2L2/Keap1 dissociation (*55*). Additionally, inhibition of PKC was shown to significantly repress expression of downstream NFE2L2 targets (*45*). Therefore, phosphorylation of S40 is crucial for the regulational activity of NFE2L2. We predict a loss of PKC-ζ through the recurrent mutation R34G (Fig. 7c), suggesting the loss of function of NFE2L2 through continuous degradation.

2.5.4 Loss of phosphorylation in CLIP1 may contribute to mitotic block

CLIP1 (CAP-GLY domain containing linker protein 1) has key roles in recruitment of dynactin to plus ends of microtubules (*56*), microtubule rescue (*57*) and mitosis (*58–60*). CLIP1 is also required in kinetochore–microtubule attachments (*60*). CLIP1-depleted cells show mitotic block due to the lack kinetochore–microtubule attachments (*59*). Phosphorylation at S1364 by CSNK2A1 is required for CLIP1 to bind to dynactin and localize to kinetochores during prometaphase (*61*). CLIP1-S1364A mutants show a loss of ability to bind dynactin and thus a loss of function (*61*). However, CLIP1 mutants containing the phosphomimetic substitution S1364D show only reduced binding with dynactin upon CSNK2A1 inhibition, suggesting that S1364 may not be the only site phosphorylated by CSNK2A1 (*61*).

S1009, upstream of S1364, is known from mass-spectrometry experiments to be phosphorylatable during mitosis (*62, 63*), but has no experimentally validated kinases. We predict phosphorylation of CSNK2A1 at S1009. More interestingly, we could predict a loss of CSNK2A1 through the mutation CLIP1-E1012K (Fig. 7d). Thus, we hypothesize that S1009 is a novel CSNK2A1 target and CLIP1-E1012K causes a loss of signaling, contributing to further reduced CLIP1–dynactin binding, which in turn causes mitotic block.

3 Methods

3.1 Data collection

3.1.1 Protein sequences and mutation data

The pan-cancer mutation dataset and their corresponding set of 18,671 protein sequences were obtained from Reimand *et al.* (*16*). The mutation dataset contained a total of 236,367 missense mutations in 3,185 samples for 12 tumour types: bladder urothelial carcinoma (blca), breast invasive carcinoma (brca), colon and rectum adenocarcinoma (coadread), glioblastoma multiforme (gbm), head and neck squamous cell carcinoma (hnsc), kidney renal clear cell carcinoma (kirc), acute myeloid leukemia (laml), lung adenocarcinoma (luad), lung squamous cell carcinoma (lusc), ovarian serous cystadenocarcinoma (ov) and uterine corpus endometrioid carcinoma (ucec).

3.1.2 Phosphosites

A total of 88,253 experimentally validated phosphosites used in this study were consolidated from four online databases (PhosphoSitePlus (*10*), PhosphoELM (*11*), HPRD (*12*), and PhosphoNetworks (*26*)), after filtering for duplicates and excluding phosphosites with no annotated literature reference. Phosphosites were mapped exactly to protein sequences and the ± 7 flanking residues were retained.

3.1.3 Expression, localization and pathway data

Expression data for 3,468 samples in 17,461 proteins was obtained from Synapse (syn1695373). Expression data for samples not existing in the pan-cancer mutations (354, 10%) was discarded. Localization data for 15,372 proteins was obtained from UniProt (*64*). Since localization data exists as a hierarchy, only the top level localization term was retained. For example, CDK1 is annotated the following hierarchy of terms: “cytoplasm”, “cytoskeleton”, “microtubule orga-

nizing center", "centrosome". Only "cytoplasm" is retained.

Functional gene lists for pathway and protein complexes were obtained from g:Profiler (*65*), and contained a total of 5,753 annotations from Reactome (*66*) and CORUM (*67*) in 4,580 proteins.

3.2 Pathway analysis

To test for enrichment of mutations in pathways and complexes, we mapped rewiring mutations to their respective pathways and carried out one-tailed Poisson tests. Resulting *p*-values were subject to multiple testing correction using the false discovery rate method (FDR). We considered pathways significantly enriched for network-rewiring mutations if FDR < 0.01. Pathways enriched with one mutation were discarded. Results were visualized using the Cytoscape (*68*) plugin, Enrichment Map (*69*). Clusters in the Enrichment Map output were defined by manual inspection.

3.3 Kinase specificity models

To model kinase specificities, we employed the PWM, which is often used to model specificities of linear motifs in biological sequences (*70*). A single PWM is constructed for each kinase using binding sites annotated it. Let S be a set of n binding sites of a kinase, each of length l, $s_1, \cdots, s_n$, where $s_k = s_{k1}, \cdots, s_{kl}$ and s_{kj} represents one of the 20 amino acids. A PWM $M_{20\times l}$ with weights p_{ij} as the relative frequency of each amino acid i at a particular position j is constructed as follows.

$$p_{ij} = \frac{1}{n}\sum_{k=1}^{n} f_i(s_{kj}) + \epsilon \qquad f_i(q) = \begin{cases} 1, & \text{if } i = q. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Where ϵ is an insignificant value to avoid any infinite values when computing the log of frequencies.

Given a potential phosphosite q of an interaction partner also of length l, $q_1, \cdots, q_l$, the relative frequencies p_{ij} are used to compute a score for the likelihood of binding. We adapted the MSS, originally developed in the MATCH algorithm (*28*) for DNA sequences. The MSS uses the information content of each position as well as normalization against the highest and lowest relative frequencies per position in the PWM to provide a score ranging from 0 to 1, where 0 represents no binding and 1 represents a perfect match. The MSS defined as:

$$\begin{aligned} MSS &= \frac{Current - Min}{Max - Min} \\ Current &= \sum_{j=1}^{l} I(j) p_{q_j,j} \\ Min &= \sum_{j=1}^{l} I(j) p_j^{min} \\ Max &= \sum_{j=1}^{l} I(j) p_j^{max} \\ I(j) &= -\sum_{i} p_{i,j} log(\frac{p_{i,j}}{p_b}) \end{aligned} \tag{2}$$

Such that q_j represents the amino acid at position j of the query sequence, p_j^{min} and p_j^{max} represent the minimum and maximum relative frequency at position j of the PWM, respectively, and p_b is the background frequency of a particular amino acid in the proteome.

Since kinases are typically known to target either an ST or a Y, but not both, scores for sequences with a central residue not matching that of the kinase are not scored. For example, a sequence with the central residue Y would not be scored against the PWM of CDK2, a known ST kinase. PWMs were constructed for kinases with 10 or more binding sites. Kinases with fewer than 10 binding sites do not provide sufficient variability for informed predictions.

3.3.1 Refining kinase specificity models

To refine our kinase specificity models, we employ a iterative method that discards sequences that did not correspond to the motif's general pattern. Here, the set of positive sequences, $S+$ is defined as the binding sequences of a particular kinase. The negative set of sequences, $S-$, is defined as all sites in the proteome with a center of ST or Y and does not exist in any experimental data. While some of the negative sequences may indeed be unexplored phosphorylation sites, not many negative phosphorylation sites exist in the literature, making this the best approximation for a negative set.

The $S+$ set is used to construct the initial PWM, M_0, which is used to score the $S+$ and $S-$ set. The distribution of $S-$ scores is used to define a threshold α as the score at the 90$^{\text{th}}$ percentile. Sequences in $S+$ which have a score below α are discarded. The remaining sequences are used to construct a new PWM, M_1. This process was repeated until there were no further sequences to discard (*i.e.* all sequence achieved a score greater than α), or, when discarding sequences, result in a retained set of less than 10 sequences.

3.3.2 Performance

To assess the performance of our PWMs, 10-fold cross-validation experiments was carried out. Here, $S+$ was randomly split into 10 equal groups, $g_1, \cdots, g_{10}$. The first group, g_1, was used as the test set. The remaining groups, $g_2, \cdots g_{10}$, were used to construct the PWM that was used to score the test set and the negative set, $S-$. Receiver operator (ROC) curves were then computed by representing the rate of true positives (TPR) versus the rate of false positives (FPR) as the cutoff varied.

$$TP = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{3}$$

This was repeated such that each group was used once as a test set. The area under the ROC

curve (AUC) was computed for each iteration and averaged after 10 runs. The average AUC provides an unbiased proxy of the PWM's prediction power.

3.4 Rewiring

3.4.1 Assessing the impact of mutations on phosphorylation binding

For a pSNV occurring in phosphosite, we score the wildtype and mutant phosphosite (*i.e.* the wildtype with the pSNV) against PWM M. Let the MSS of the wildtype and mutant be MSS_{wt} and MSS_{mt}, respectively. We score $S+$ and $S-$ using M, which form the foreground and background distributions of scores, respectively. We define thresholds $\alpha-$ and $\alpha+$ as the scores at the 90$^{\text{th}}$ and 10$^{\text{th}}$ percentiles of the background and foreground distributions, respectively, and consider a rewiring event to occur if $MSS_{wt} > \alpha+$ and $MSS_{mt} < \alpha-$ (loss of signaling) or vice-versa (gain of signaling).

3.5 Implementation of MIMP

MIMP was implemented as a freely-available R package with pre-computed kinase specificity models to allow users to analyze their data for network-rewiring events. MIMP requires two mandatory input files: mutation data containing single amino acid substitutions and the protein sequence data in FASTA format. The third input file is an optional file and is the phosphorylation data containing positions of the phosphorylated residues in the protein. If phosphorylation data is not supplied, MIMP will use all residues containing an S, T, or Y as probable phosphorylation sites. Users can also adjust the percentile values, which determine the $\alpha+$ and $\alpha-$ thresholds.

Users obtain results are returned in a table object containing rewiring pSNVs and their impact on the binding sites. This includes the position of the phosphosite affected, wildtype and mutant binding sites, scores before and after the mutation, etc. Users can chose to visually

display this table in their browser (Fig. 8). This allows for easier navigation of the results (through sorting columns, filtering rows, etc.), as well as visual inspection of the motifs of rewired kinases.

Instructions on how to install MIMP along with documentation and sample data can be obtained from `https://github.com/omarwagih/mimp`.

4 Discussion

Mutation data, particularly those recurrent in samples of cancer types, are of interest since they are found to be enriched in binding interfaces (*15, 71, 72*), which can ultimately impact physical protein binding, namely kinase phosphorylation, which acts as one the largest and most important regulatory mechanisms in eukaryotic cells. Thus, the identification of such mutations offers us a step forward in pinpointing and interpreting driver mutations. Despite this, the extent to which these mutations impact kinase binding is yet to be studied comprehensively.

To explore this, we developed a computational method, MIMP, which measures the impact of mutations existing around the flanking regions of phosphorylation sites. Our method assumes that the phosphorylation event, through any additional required cellular machinery, occurs in the tumor sample but also that other conditions of cellular context are fulfilled, both the kinase and the substrate are active in the cell (*i.e.* expressed), and that they are in the same sub-cellular compartment. Using genetic variation data from the TCAG pan-cancer dataset, we identified a network of gain and loss of kinase events and identified rewiring mutations in key cancer proteins, such as TP53, CTNNB1, and NFE2L2, as well as potential novel cancer proteins such as CLIP1. We show that mutations responsible for these events are enriched in key cancer themes and have a significant functional bias.

A major limitation to this type of analysis lies within the kinase–substrate predictions. To date, there still remains a relatively sparse number of experimentally validated phosphorylation sites, making computational prediction a challenging task for the majority of kinases. Currently, 300/501 (60%) of all kinases have at least one and less than 30 experimentally validated phosphorylation sites. This can be due to a number of factors. Firstly, the collection of phosphorylation sites requires curators to successfully extract them from scientific literature, which means that keeping up is an ongoing challenge. Secondly, there is a large bias towards the num-

ber of substrates available for well-studied kinases, whereas others are neglected. Lastly, some kinases interact with their substrate with a low affinity or in a transient manner, making their detection difficult. The 333 PWMs with satisfactory prediction power utilized in this study only represent a fraction of known kinases (64%) and for some kinases with low substrate counts, may not represent its global specificity. Furthermore, this analysis is limited for a majority of other kinases with little to no substrate data.

Most databases storing phosphorylation sites do not annotate the context in which they are phosphorylated (*e.g.* different tissues or conditions). Additionally, many kinase–substrate relationships can be due to a secondary effect, through the regulation of the kinase of interest on a downstream kinase. These issues raise a relatively large number of potential false positives or sequences, which do not match the known motif of the kinase and further overburdens computational predictors. MIMP attempts to overcome some of these issues by refining kinase specificities. This enriches prominent residues in substrate data, likely responsible for binding. MIMP is also maintained through annual updates of the latest phosphorylation data.

Our future direction includes effort to further improve kinase–substrate predictions through the use of phosphorylation data from closely related model organisms as well as incorporating genomic context into our prediction pipeline to allow for more biologically relevant predictions. While we applied MIMP of the pan-cancer pSNVs on to phosphorylation, the same principal applies to any motif-based interactions and PTMs (textite.g. transcription factor binding, SH3/SH2 domains, PDZ domains, and WW domains) with any disease-related mutations. Interestingly, interplay exists amongst many of such interactions and PTMs (*73*). Thus, we plan to utilize MIMP to carry out a comprehensive analysis of the mutational impact on linear-motif-based binding interfaces. We hope this will allow us to better understand how function of proteins is altered through PTMs in different disease conditions, either directly or indirectly and pave the way for personalized medicine.

References

1. Reimand, J., Hui, S., Jain, S., Law, B., and Bader, G. D. Domain-mediated protein interaction prediction: From genome to network. *FEBS Lett.* **586**(17), 2751–2763 (2012).

2. Whitmarsh, A. J. and Davis, R. J. Regulation of transcription factor function by phosphorylation. *Cell. Mol. Life Sci.* **57**(8-9), 1172–1183 (2000).

3. Ruvolo, P. P., Deng, X., and May, W. S. Phosphorylation of Bcl2 and regulation of apoptosis. *Leukemia* **15**(4), 515–522 (2001).

4. Joshi, K., Banasavadi-Siddegowda, Y., Mo, X., Kim, S.-H., Mao, P., Kig, C., Nardini, D., Sobol, R. W., Chow, L. M. L., Kornblum, H. I., Waclaw, R., Beullens, M., and Nakano, I. MELK-dependent FOXM1 phosphorylation is essential for proliferation of glioma stem cells. *Stem Cells* **31**(6), 1051–1063 (2013).

5. Mendelsohn, J. and Baselga, J. The EGF receptor family as targets for cancer therapy. *Oncogene* **19**(56), 6550–6565 (2000).

6. Sawyers, C. L. Rational therapeutic intervention in cancer: kinases as drug targets. *Curr. Opin. Genet. Dev.* **12**(1), 111–115 (2002).

7. Santarpia, L., Lippman, S. M., and El-Naggar, A. K. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opin. Ther. Targets* **16**(1), 103–119 (2012).

8. Sheppard, K., Kinross, K. M., Solomon, B., Pearson, R. B., and Phillips, W. A. Targeting PI3 kinase/AKT/mTOR signaling in cancer. *Crit. Rev. Oncog.* **17**(1), 69–95 (2012).

9. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**(5600), 1912–1934 (2002).

10. Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**(Database issue), D261–D270 (2012).

11. Diella, F., Cameron, S., Gemnd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontn, T., Blom, N., and Gibson, T. J. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79 (2004).

12. Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **37**(Database issue), D767–D772 (2009).

13. Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., Bader, G. D., Boutros, P. C., Muthuswamy, L., Ouellette, B. F. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., Dronov, S., Flicek, P., Shannon, N. B., Carter, H., Ding, L., Sander, C., Stuart, J. M., Stein, L. D., Lopez-Bigas, N., , I. C. G. C. M. P., and of the Bioinformatics Analyses Working Group, C. S. Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10**(8), 723–729 (2013).

14. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, Jr, L. A., and Kinzler, K. W. Cancer genome landscapes. *Science* **339**(6127), 1546–1558 (2013).

15. Reimand, J. and Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).

16. Reimand, J., Wagih, O., and Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).

17. Radivojac, P., Baenziger, P. H., Kann, M. G., Mort, M. E., Hahn, M. W., and Mooney, S. D. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* **24**(16), i241–i247 (2008).

18. Ren, J., Jiang, C., Gao, X., Liu, Z., Yuan, Z., Jin, C., Wen, L., Zhang, Z., Xue, Y., and Yao, X. PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. *Mol. Cell. Proteomics* **9**(4), 623–634 (2010).

19. Ryu, G., Song, P., Kim, K., Oh, K., Park, K., and Kim, J. H. Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. *Nucleic Acids Res.* **37**(4), 1297–1307 (2009).

20. Li, S., Iakoucheva, L. M., Mooney, S. D., and Radivojac, P. Loss of post-translational modification sites in disease. *Pac. Symp. Biocomput.* **1**, 337–347 (2010).

21. Savas, S., Taylor, I. W., Wrana, J. L., and Ozcelik, H. Functional nonsynonymous single nucleotide polymorphisms from the TGF-beta protein interaction network. *Physiol. Genomics* **29**(2), 109–117 (2007).

22. Riaño-Pachón, D. M., Kleessen, S., Neigenfind, J., Durek, P., Weber, E., Engelsberger, W. R., Walther, D., Selbig, J., Schulze, W. X., and Kersten, B. Proteome-wide survey of phosphorylation patterns affected by nuclear DNA polymorphisms in Arabidopsis thaliana. *BMC Genomics* **11**, 411 (2010).

23. Savas, S. and Ozcelik, H. Phosphorylation states of cell cycle and DNA repair proteins can be altered by the nsSNPs. *BMC Cancer* **5**, 107 (2005).

24. Cho, S., Savas, S., and Ozcelik, H. Genetic variation and the mitogen-activated protein kinase (MAPK) signaling pathway. *OMICS.* **10**(1), 66–81 (2006).

25. Hendriks, W. J. A. J. and Pulido, R. Protein tyrosine phosphatase variants in human hereditary disorders and disease susceptibilities. *Biochim. Biophys. Acta.* **1832**(10), 1673–1696 (2013).

26. Newman, R. H., Hu, J., Rho, H.-S., Xie, Z., Woodard, C., Neiswinger, J., Cooper, C., Shirley, M., Clark, H. M., Hu, S., Hwang, W., Jeong, J. S., Wu, G., Lin, J., Gao, X., Ni, Q., Goel, R., Xia, S., Ji, H., Dalby, K. N., Birnbaum, M. J., Cole, P. A., Knapp, S., Ryazanov, A. G., Zack, D. J., Blackshaw, S., Pawson, T., Gingras, A.-C., Desiderio, S., Pandey, A., Turk, B. E., Zhang, J., Zhu, H., and Qian, J. Construction of human activity-based phosphorylation networks. *Mol. Syst. Biol.* **9**, 655 (2013).

27. Miller, M. L., Jensen, L. J., Diella, F., Jrgensen, C., Tinti, M., Li, L., Hsiung, M., Parker, S. A., Bordeaux, J., Sicheritz-Ponten, T., Olhovsky, M., Pasculescu, A., Alexander, J., Knapp, S., Blom, N., Bork, P., Li, S., Cesareni, G., Pawson, T., Turk, B. E., Yaffe, M. B., Brunak, S., and Linding, R. Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* **1**(35), ra2 (2008).

28. Kel, A. E., Gssling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**(13), 3576–3579 (2003).

29. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**(7), 1575–1584 (2002).

30. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**(1), 110–121 (2010).

31. Ng, P. C. and Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**(13), 3812–3814 (2003).

32. Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7 20 (2013).

33. Chun, S. and Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**(9), 1553–1561 (2009).

34. Schwarz, J. M., Rdelsperger, C., Schuelke, M., and Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**(8), 575–576 (2010).

35. Vogelstein, B., Lane, D., and Levine, A. J. Surfing the p53 network. *Nature* **408**(6810), 307–310 (2000).

36. Vousden, K. H. and Lu, X. Live or let die: the cell's response to p53. *Nat. Rev. Cancer* **2**(8), 594–604 (2002).

37. Wu, L., Ma, C. A., Zhao, Y., and Jain, A. Aurora B interacts with NIR-p53, leading to p53 phosphorylation in its DNA-binding domain and subsequent functional suppression. *J. Biol. Chem.* **286**(3), 2236–2244 (2011).

38. Vousden, K. H. and Prives, C. Blinded by the Light: The Growing Complexity of p53. *Cell* **137**(3), 413–431 (2009).

39. MacLennan, I. C. Germinal centers. *Annu Rev Immunol* **12**, 117–139 (1994).

40. Liu, C., Li, Y., Semenov, M., Han, C., Baeg, G. H., Tan, Y., Zhang, Z., Lin, X., and He, X. Control of beta-catenin phosphorylation/degradation by a dual-kinase mechanism. *Cell* **108**(6), 837–847 (2002).

41. Yost, C., Torres, M., Miller, J. R., Huang, E., Kimelman, D., and Moon, R. T. The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in Xenopus embryos by glycogen synthase kinase 3. *Genes Dev.* **10**(12), 1443–1454 (1996).

42. Aberle, H., Bauer, A., Stappert, J., Kispert, A., and Kemler, R. beta-catenin is a target for the ubiquitin-proteasome pathway. *EMBO J.* **16**(13), 3797–3804 (1997).

43. Kitagawa, M., Hatakeyama, S., Shirane, M., Matsumoto, M., Ishida, N., Hattori, K., Nakamichi, I., Kikuchi, A., Nakayama, K., and Nakayama, K. An F-box protein, FWD1, mediates ubiquitin-dependent proteolysis of beta-catenin. *EMBO J.* **18**(9), 2401–2410 (1999).

44. Moreno-Bueno, G., Gamallo, C., Pérez-Gallego, L., de Mora, J. C., Suárez, A., and Palacios, J. beta-Catenin expression pattern, beta-catenin gene mutations, and microsatellite instability in endometrioid ovarian carcinomas and synchronous endometrial carcinomas. *Diagn. Mol. Pathol.* **10**(2), 116–122 (2001).

45. Huang, H.-C., Nguyen, T., and Pickett, C. B. Phosphorylation of Nrf2 at Ser-40 by protein kinase C regulates antioxidant response element-mediated transcription. *J. Biol. Chem.* **277**(45), 42769–42774 (2002).

46. Shelton, P. and Jaiswal, A. K. The transcription factor NF-E2-related factor 2 (Nrf2): a protooncogene? *FASEB J.* **27**(2), 414–423 (2013).

47. Zhang, D. D. and Hannink, M. Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress. *Mol. Cell. Biol.* **23**(22), 8137–8151 (2003).

48. Cullinan, S. B., Gordan, J. D., Jin, J., Harper, J. W., and Diehl, J. A. The Keap1-BTB protein is an adaptor that bridges Nrf2 to a Cul3-based E3 ligase: oxidative stress sensing by a Cul3-Keap1 ligase. *Mol. Cell. Biol.* **24**(19), 8477–8486 (2004).

49. Zhang, D. D., Lo, S.-C., Cross, J. V., Templeton, D. J., and Hannink, M. Keap1 is a redox-regulated substrate adaptor protein for a Cul3-dependent ubiquitin ligase complex. *Mol. Cell. Biol.* **24**(24), 10941–10953 (2004).

50. Bloom, D. A. and Jaiswal, A. K. Phosphorylation of Nrf2 at Ser40 by protein kinase C in response to antioxidants leads to the release of Nrf2 from INrf2, but is not required for Nrf2 stabilization/accumulation in the nucleus and transcriptional activation of antioxidant response element-mediated NAD(P)H:quinone oxidoreductase-1 gene expression. *J. Biol. Chem.* **278**(45), 44675–44682 (2003).

51. Baba, K., Morimoto, H., and Imaoka, S. Seven in absentia homolog 2 (Siah2) protein is a regulator of NF-E2-related factor 2 (Nrf2). *J. Biol. Chem.* **288**(25), 18393–18405 (2013).

52. Kim, W. D., Kim, Y. W., Cho, I. J., Lee, C. H., and Kim, S. G. E-cadherin inhibits nuclear accumulation of Nrf2: implications for chemoresistance of cancer cells. *J. Cell Sci.* **125**(Pt 5), 1284–1295 (2012).

53. Numazawa, S., Ishikawa, M., Yoshida, A., Tanaka, S., and Yoshida, T. Atypical protein kinase C mediates activation of NF-E2-related factor 2 in response to oxidative stress. *Am. J. Physiol. Cell Physiol.* **285**(2), C334–C342 (2003).

54. Li, Y., Paonessa, J. D., and Zhang, Y. Mechanism of chemical activation of Nrf2. *PLoS One* **7**(4), e35122 (2012).

55. Niture, S. K., Jain, A. K., and Jaiswal, A. K. Antioxidant-induced modification of INrf2 cysteine 151 and PKC-delta-mediated phosphorylation of Nrf2 serine 40 are both required for stabilization and nuclear translocation of Nrf2 and increased drug resistance. *J. Cell Sci.* **122**(Pt 24), 4452–4464 (2009).

56. Lansbergen, G., Komarova, Y., Modesti, M., Wyman, C., Hoogenraad, C. C., Goodson, H. V., Lemaitre, R. P., Drechsel, D. N., van Munster, E., Gadella, Jr, T. W. J., Grosveld, F., Galjart, N., Borisy, G. G., and Akhmanova, A. Conformational changes in CLIP-170 regulate its binding to microtubules and dynactin localization. *J. Cell Biol.* **166**(7), 1003–1014 (2004).

57. Komarova, Y. A., Akhmanova, A. S., Kojima, S.-I., Galjart, N., and Borisy, G. G. Cytoplasmic linker proteins promote microtubule rescue in vivo. *J. Cell Biol.* **159**(4), 589–599 (2002).

58. Dujardin, D., Wacker, U. I., Moreau, A., Schroer, T. A., Rickard, J. E., and De Mey, J. R. Evidence for a role of CLIP170 in the establishment of metaphase chromosome alignment. *J. Cell. Biol.* **141**(4), 849–862 (1998).

59. Wieland, G., Orthaus, S., Ohndorf, S., Diekmann, S., and Hemmerich, P. Functional complementation of human centromere protein A (CENP-A) by Cse4p from Saccharomyces cerevisiae. *Mol. Cell. Biol.* **24**(15), 6620–6630 (2004).

60. Tanenbaum, M. E., Galjart, N., van Vugt, M. A. T. M., and Medema, R. H. CLIP170 facilitates the formation of kinetochore-microtubule attachments. *EMBO J.* **25**(1), 45–57 (2006).

61. Li, H., Liu, X. S., Yang, X., Wang, Y., Wang, Y., Turner, J. R., and Liu, X. Phosphorylation of CLIP-170 by Plk1 and CK2 promotes timely formation of kinetochore-microtubule attachments. *EMBO J.* **29**(17), 2953–2965 (2010).

62. Shiromizu, T., Adachi, J., Watanabe, S., Murakami, T., Kuga, T., Muraoka, S., and Tomonaga, T. Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *J. Proteome Res.* **12**(6), 2414–2421 (2013).

63. Olsen, J. V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M. L., Jensen, L. J., Gnad, F., Cox, J., Jensen, T. S., Nigg, E. A., Brunak, S., and Mann, M. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.* **3**(104), ra3 (2010).

64. Magrane, M. and Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).

65. Reimand, J., Arak, T., and Vilo, J. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39**(Web Server issue), W307–W315 (2011).

66. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**(Database issue), D619–D622 (2009).

67. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. CORUM: the comprehensive resource of

mammalian protein complexes–2009. *Nucleic Acids Res.* **38**(Database issue), D497–D501 (2010).

68. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003).

69. Merico, D., Isserlin, R., and Bader, G. D. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. *Methods Mol. Biol.* **781**, 257–277 (2011).

70. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**(1), 16–23 (2000).

71. Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**(2), 159–164 (2012).

72. David, A., Razali, R., Wass, M. N., and Sternberg, M. J. E. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum. Mutat.* **33**(2), 359–363 (2012).

73. Woodsmith, J., Kamburov, A., and Stelzl, U. Dual coordination of post translational modifications in human protein networks. *PLoS Comput. Biol.* **9**(3), e1002933 (2013).