MIMP: predicting the impact of mutations on kinase-substrate phosphorylation

Omar Wagih[1], Jüri Reimand[1,2] & Gary D Bader[1,2]

Protein phosphorylation is important in cellular pathways and altered in disease. We developed MIMP (http://mimp.baderlab.org/), a machine learning method to predict the impact of missense single-nucleotide variants (SNVs) on kinase-substrate interactions. MIMP analyzes kinase sequence specificities and predicts whether SNVs disrupt existing phosphorylation sites or create new sites. This helps discover mutations that modify protein function by altering kinase networks and provides insight into disease biology and therapy development.

Signaling networks mediate complex cellular logic, often via interactions of modular protein domains and short linear motifs[1]. The impact of genome variation on molecular interaction networks is difficult to predict in general; however, interactions involving short linear motifs[2] are well known enough to tackle. A prominent example is protein phosphorylation, a post-translational modification (PTM) of serine, threonine or tyrosine residues. Human proteins are phosphorylated by over 500 kinases[3] that bind motifs in flanking residues of modified sites. These motifs are inferred from networks of experimentally confirmed phosphorylation sites (phosphosites) and associated kinases. Phosphosites are evolutionarily constrained in human genomes and enriched in cancer driver mutations and causal variants of inherited disease, indicating their functional importance[4,5]. Phosphorylation-related SNVs (pSNVs) can disrupt existing phosphosites and create novel sites, rewiring kinase-substrate interactions and leading to disease phenotypes (**Fig. 1a**).

Earlier studies of disease mutations in kinase-binding sites[6–10] were limited to the analysis of mutations directly affecting the central phosphosite, whereas flanking sequence and the impact on kinase-specific phosphorylation was generally not considered[6,7,10]. Disease mutations were limited to few well-studied genes, resulting in static databases that have not kept pace with substantial growth in PTM data[9,10]. Methods to interpret mutations in kinase-substrate phosphorylation were developed[10,11], but no public tools are available for automated analysis. The PhosphoSitePlus database[12] provides a useful list of genome variants in phosphosites and other protein sites, but it does not provide methods to automatically predict the impact of mutations on kinase binding. Thus, updated methods are needed to leverage rapidly increasing genomic and phosphoproteomic data to interpret variation in signaling networks.

We developed a method called mutation impact on phosphorylation (MIMP) to predict the function of SNVs in phosphosites (**Fig. 1a**). MIMP is a machine learning method based on Bayesian statistics that builds on our previous analysis of phosphosite mutations[11] (**Supplementary Note**). We collected 7,004 kinase-associated phosphosite sequences from public databases[12–14] and constructed position weight matrix (PWM) models of amino acid specificities of kinases. We modeled 124 high-confidence kinases with at least ten known phosphosites, including 99 serine-threonine kinases and 25 tyrosine kinases (**Supplementary Data 1–3** and **Supplementary Fig. 1**). We compiled two further data sets: 58 models of kinase families[3] and 294 kinase models predicted from protein-protein interactions[15], totaling 476 models of 322 kinases (**Supplementary Note**). We removed outlier sequences with an iterative model-refinement procedure and discarded models with substandard classifier performance (**Fig. 1a**, **Supplementary Fig. 2** and **3** and Online Methods).

We quantify kinase-phosphosite interactions with the matrix similarity score (MSS) developed for DNA motifs[16] (Online Methods). We computed MSSs for true positive binding sequences (P) and randomly sampled negative sequences (N) and applied model-based clustering[17] to train Gaussian mixture models (GMMs) M_P and M_N for scores of each kinase. Phosphosites are classified with GMMs: a site is considered kinase bound if its score s is likely derived from the M_P distribution and considered not kinase-bound if s is likely derived from M_N. Binding is quantified with a Bayesian posterior probability P that comprises the likelihood of the score s derived from the distribution M_P, as well as our prior belief in M_N (Online Methods). Priors assign more confidence to kinases with higher-quality PWMs. We evaluate phosphorylation loss as the joint probability of the wild-type phosphosite sequence being kinase bound and the matched mutant sequence being unbound. Phosphorylation gain is evaluated analogously. We use a threshold of $p > 0.5$ for plausible rewiring hypotheses and maximum sensitivity, and we keep results with at least twofold change in wild-type and mutant scores. Method performance is tuned using these parameters.

We tested MIMP on 236,367 missense SNVs from The Cancer Genome Atlas (TCGA) pan-cancer data set of 3,185 cancer samples of 12 tumor types[18] (**Supplementary Data 4**). We analyzed ~190,000 experimentally derived phosphosites from public databases[12–14] (**Supplementary Data 5**) and found 37,996 pSNVs in the flanking sequences within seven residues of the sites. MIMP

[1]The Donnelly Centre, University of Toronto, Toronto, Canada. [2]These authors jointly supervised this work. Correspondence should be addressed to O.W. (wagih@ebi.ac.uk) or J.R. (juri.reimand@utoronto.ca) or G.D.B. (gary.bader@utoronto.ca).

RECEIVED 9 DECEMBER 2014; ACCEPTED 30 MARCH 2015; PUBLISHED ONLINE 4 MAY 2015; DOI:10.1038/NMETH.3396

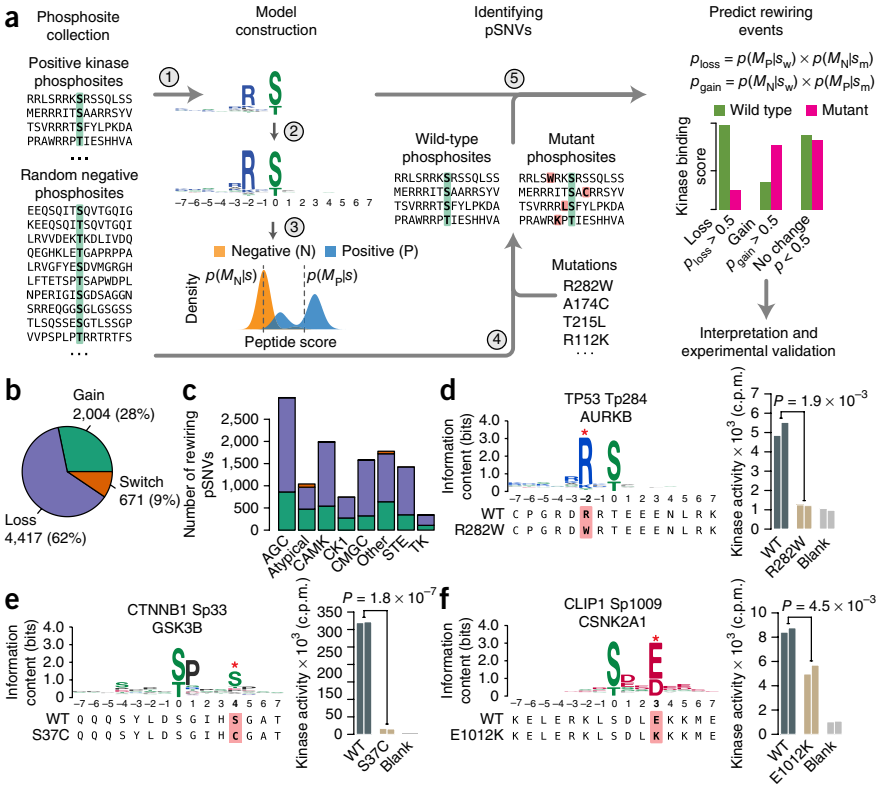

Figure 1 | MIMP workflow and analysis. (**a**) MIMP workflow. (1) kinase sequence specificity models are constructed from known binding sites as position weight matrices and (2) refined iteratively. (3) Scores of known positive kinase-binding sites P and random negative sites N are modeled as Gaussian mixture distributions. Bayesian posterior probabilities are computed to classify sites as positives or negatives. (4) Cancer mutations are mapped to phosphosites. (5) Rewiring events are predicted by classifying matched wild-type and mutant sequences to alternate distributions. Phosphorylation loss is predicted if the joint probability $p_{loss} > 0.5$ for the wild-type sequence being a positive kinase-binding site and the mutant sequence being a negative site. Phosphorylation gain is predicted similarly. (**b**) Analysis of mutations from the TCGA pan-cancer data set reveals numerous SNVs with predicted phosphorylation gain (green), loss (purple) and switch (orange). (**c**) Kinase families with the most frequent network-rewiring mutations. Colors as in **b**. (**d**–**f**) Experimental validation of three SNVs with predicted phosphorylation loss. Kinase sequence specificity models with wild-type (WT) and mutant phosphosites are shown on the left. Names of substrates, associated kinases, and phosphorylated residues are shown on top of sequence logos. Mutated residues are indicated with red asterisks and red shading. Bar plots quantify *in vitro* kinase activity in counts per minute (c.p.m.) for two replicates of wild-type and mutant sequences as well as negative controls (blank). FDR-corrected *P* values were computed with the naive Bayes modified *t*-test[23].

provides a functional hypothesis for 7,092 pSNVs that potentially disrupt 7,589 phosphosites in 4,355 genes, including a significant enrichment of cancer genes (**Fig. 1b**,**c** and **Supplementary Fig. 4**; $n = 193$, $P = 3.2 \times 10^{-13}$, Fisher's exact test). We predict that mutations induce 41,952 network-rewiring events, including 8,852 phosphorylation gains and 33,100 losses (**Supplementary Fig. 5**, **Supplementary Data 6** and **Supplementary Discussion**). Some pSNVs (n = 671; 9%) may cause phosphorylation switches by simultaneously disrupting existing phosphosites and introducing new sites (**Fig. 1b**). The most frequently mutated pSNV hotspots occur in cancer genes (*TP53*, n = 23 pSNVs; *BRAF*, n = 13; *CTNNB1*, n = 9), highlighting known[19] and predicted driver mechanisms of cancer.

We used pathways and genomic data for validation. Enrichment analysis revealed abundant network-rewiring pSNVs in cancer-related processes such as apoptosis, translation, cell cycle and DNA repair (**Supplementary Fig. 6** and Online Methods; false discovery rate (FDR) $P < 0.01$, Poisson exact test). We investigated matched gene expression data from TCGA and cell component annotations from UniProt, as predictions of coexpressed and colocalized proteins are more likely valid. We found that 90% of kinase-substrate pairs show mRNA coexpression in corresponding tumor samples and that 63% co-occur in cellular compartments, significantly exceeding corresponding rates in random kinase-substrate pairs (**Supplementary Fig. 7**; $P < 1.8 \times 10^{-3}$, Z-test). Thus, our predictions involve hallmark cancer processes and are likely compatible with cellular environment.

We experimentally tested 11 network-rewiring mutations (eight loss and three gain) of rewired kinases in five phosphosites of *TP53*, *CTNNB1* and *CLIP1*, corresponding to most frequent pSNVs (**Fig. 1d–f**, **Supplementary Table 1** and **Supplementary Results**). We exposed libraries of wild-type and mutated sequences to predicted kinase domains *in vitro* and quantified phosphorylation (**Supplementary Table 1** and Online Methods). We confirmed ten cases (91%) with significantly altered kinase activity in mutated sequences relative to wild types (FDR $P < 0.05$, naive Bayes *t*-test; **Supplementary Results** and **Supplementary Fig. 8**). We propose that substitutions R213Q and R282W in TP53 disrupt phosphorylation by aurora kinase B, leading to increased TP53 activity in mutated cancer samples in TCGA data (R282W, n = 23 samples; R213Q, n = 4; **Supplementary Figs. 9** and **10**). These phosphosites are associated with TP53 inhibition[20,21]. Thus, MIMP can detect functional mutations in kinase-binding sites and propose corresponding mechanisms.

MIMP is available as a user-friendly web server and R package (http://mimp.baderlab.org/). We aim to update data annually and extend MIMP to site-specific protein-protein interaction networks (14-3-3, SH3, SH2, PDZ[22], WW) as well as protein-DNA and protein-RNA interaction networks of transcriptional regulation and RNA splicing (**Supplementary Discussion**).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Moses for detailed comments that improved the method and the Kinexus Bioinformatics Corporation for conducting kinase assays. This work was supported by the Canadian Institutes of Health Research grant MOP-84324 to G.D.B.

AUTHOR CONTRIBUTIONS

O.W., J.R. and G.D.B. devised the method and designed the analysis. O.W. analyzed the data, implemented the method and developed the software. O.W. wrote the initial manuscript. All authors edited and approved the final manuscript. J.R. and G.D.B. jointly supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Pawson, T. *Nature* **373**, 573–580 (1995).
2. Reimand, J., Hui, S., Jain, S., Law, B. & Bader, G.D. *FEBS Lett.* **586**, 2751–2763 (2012).
3. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. *Science* **298**, 1912–1934 (2002).
4. Reimand, J. & Bader, G.D. *Mol. Syst. Biol.* **9**, 637 (2013).
5. Reimand, J., Wagih, O. & Bader, G.D. *PLoS Genet.* **11**, e1004919 (2015).
6. Riaño-Pachón, D.M. *et al. BMC Genomics* **11**, 411 (2010).
7. Savas, S. & Ozcelik, H. *BMC Cancer* **5**, 107 (2005).
8. Radivojac, P. *et al. Bioinformatics* **24**, i241–i247 (2008).
9. Ren, J. *et al. Mol. Cell. Proteomics* **9**, 623–634 (2010).
10. Ryu, G.M. *et al. Nucleic Acids Res.* **37**, 1297–1307 (2009).
11. Reimand, J., Wagih, O. & Bader, G.D. *Sci. Rep.* **3**, 2651 (2013).
12. Hornbeck, P.V. *et al. Nucleic Acids Res.* **40**, D261–D270 (2012).
13. Diella, F. *et al. BMC Bioinformatics* **5**, 79 (2004).
14. Keshava Prasad, T.S. *et al. Nucleic Acids Res.* **37**, D767–D772 (2009).
15. Newman, R.H. *et al. Mol. Syst. Biol.* **9**, 655 (2013).
16. Kel, A.E. *et al. Nucleic Acids Res.* **31**, 3576–3579 (2003).
17. Fraley, C. & Raftery, A.E. *J. Am. Stat. Assoc.* **97**, 611–631 (2002).
18. Weinstein, J.N. *et al. Nat. Genet.* **45**, 1113–1120 (2013).
19. Aberle, H., Bauer, A., Stappert, J., Kispert, A. & Kemler, R. *EMBO J.* **16**, 3797–3804 (1997).
20. Wu, L., Ma, C.A., Zhao, Y. & Jain, A. *J. Biol. Chem.* **286**, 2236–2244 (2011).
21. Gully, C.P. *et al. Proc. Natl. Acad. Sci. USA* **109**, E1513–E1522 (2012).
22. Gfeller, D., Ernst, A., Jarvik, N., Sidhu, S.S. & Bader, G.D. *PLoS ONE* **9**, e94507 (2014).
23. Smyth, G.K. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).

ONLINE METHODS

Data collection. We collected 190,428 experimentally validated human phosphosites from three online databases (PhosphoSitePlus[12], PhosphoELM[13], HPRD[14]) after excluding duplicates and sites without annotated literature reference. Phosphosites were matched with exact sequence to longest isoforms of CCDS proteins, and ±7 flanking residues were retained as previously described[11]. Nonmatching phosphosites were discarded. Somatic missense single-nucleotide variants of 3,185 cancer samples and 12 cancer types from the TCGA pan-cancer project[18] were retrieved from the Synapse database (http://www.synapse.org/; ID syn1729383). Matched expression data covering 3,468 samples and 17,461 genes from the TCGA were also obtained from Synapse (syn1695373). We discarded 354 gene expression samples (10%) that were not included in the mutation data set. Cell component annotations of 15,372 proteins were obtained from UniProt[24]. Only top-level terms of the localization hierarchy were retained for maximum coverage.

Kinase specificity models. Kinase sequence specificities were modeled as position weight matrices (PWMs). We initially studied sequence specificities of 151 kinases and filtered the set to obtain 124 high-confidence models (see "Performance"). A single PWM was constructed for each kinase using known binding sites. Let S be a set of n binding sites of a kinase, each of length m, $s_1,\ldots,s_n$, where $s_k = s_{k1},\ldots,s_{km}$ and s_{kj} represents one of the 20 amino acids. A PWM is a matrix M of size $20 \times m$ with weights f_{ij} as relative frequencies of amino acid i at position j

$$f_{ij} = \frac{1}{n}\sum_{k=1}^{n} \lambda_i(s_{kj}) + \varepsilon \qquad \lambda_i(q) = \begin{cases} 1, & \text{if } i = q \\ 0, & \text{otherwise} \end{cases}$$

The value ε is computed as the background probability of the amino acid multiplied by a pseudocount constant 0.01. Pseudocounts avoid infinite values and numerical problems when computing logarithms of frequencies, and they conservatively give higher preferences to the background model. Given a potential phosphosite sequence q of a kinase of length m, $q_1,\ldots,q_m$, the relative frequencies f_{ij} were used to compute binding scores. We adapted the matrix similarity score (MSS) originally developed for the analysis of DNA sequences in the MATCH algorithm[16]. MSS uses information content of each sequence position and normalizes against the highest and lowest relative frequencies per position in the PWM. The minimum MSS, 0, represents the lowest possible binding score, whereas the maximum MSS, 1, corresponds to a perfect match sequence. MSS is defined in the following formulas

$$MSS = \frac{\text{Current} - \text{Min}}{\text{Max} - \text{Min}}$$

$$\text{Current} = \sum_{j=1}^{m} I(j) f_{q_j,j}$$

$$\text{Min} = \sum_{j=1}^{m} I(j) f_j^{\min}$$

$$\text{Max} = \sum_{j=1}^{m} I(j) f_j^{\max}$$

$$I(j) = -\sum_{i} f_{i,j} \log\left(\frac{f_{i,j}}{f_b}\right)$$

The value q_j represents the amino acid at position j of the query sequence, $f_j^{\min}$ and $f_j^{\max}$ represent minimum and maximum relative frequency at position j of the PWM, respectively, and f_b is the background frequency of a particular amino acid in the proteome. The central residue is discarded from scoring, as it would provide the strongest signal in the PWM and would mask signals in the flanking sequence. We separated kinases into two classes (serine/threonine and tyrosine kinases) on the basis of the majority of their reported binding sites, and we discarded sequences where the central residue mismatched the kinase class. Sequences with a central residue mismatching the PWM class were not. PWMs were constructed for kinases with at least ten binding sites. Kinases with fewer sites were removed as their PWMs proved too variable for informed predictions.

Refining kinase specificity models. To refine our kinase specificity models, we iteratively discarded sequences with poor correspondence to the PWM. The set of positive sequences S^+ included all confirmed binding sites of a particular kinase, and the negative set S^- included 10,000 15-mer sequences with central (eighth) residue of serine/threonine or tyrosine. The negative set was uniformly sampled from the proteome and excluded experimentally confirmed phosphosites. Although S^- may include unexplored phosphorylation sites, it provides a proxy of true negative sites as only few experimentally confirmed nonphosphorylated sites are known. The S^+ set was compiled into the initial PWM M_0 used to score the sets S^+ and S^-. To refine PWMs, we compared positive sequences to the score distribution of negative sequences and discarded positives that had scores similar to negatives. The S^- distribution provided a threshold t as the 90th percentile, and sequences in S^+ with score below t were discarded. The remaining sequences in S^+ were summarized as the new PWM M_1. This process was repeated until no further sequences were discarded owing to all sequences in S^+ achieving a score greater than t or to the refinement surpassing the lower bound of ten sequences per positive set.

Performance. We used tenfold cross-validation to assess the performance of our PWMs. The S^+ set was randomly split into ten equal groups $g_1,\ldots,g_{10}$. The first group g_1 was used as the test set, and the remaining groups $g_2,\ldots,g_{10}$ were summarized as a PWM. The PWMs were used to score known kinase binding sites in the test set and true negative sequences in a conservative negative set $S^\#$. In model evaluation, the negative set $S^\#$ contained all experimental phosphosite sequences annotated to the other kinase families[3] excluding the family of the given kinase. This conservative approach selected kinase specificity models that discriminated kinase-specific phosphosites from other known phosphorylation sites. To evaluate model performance, we computed receiver operating characteristic (ROC) curves and area under curve

doi:10.1038/nmeth.3396

values (AUCs) from true positive rates (TPR) and false positive rates (FPR) as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}; \quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

This procedure was repeated ten times with each group as a test set, and the AUC values were averaged over ten iterations. We used an AUC of >0.6 to filter PWMs classifying kinase-specific sites and sites of other kinases. This corresponds to an AUC >0.65 of classifying true phosphosites from random nonphosphorylated sites (**Supplementary Fig. 3**). The filtering procedure resulted in 124 high-confidence PWMs.

Predicting kinase binding and impact of mutations. To quantify kinase interactions with phosphosite sequences, we trained two Gaussian mixture models (GMMs) M_P and M_N to reflect MSS scores of true positive kinase-bound sequences and randomly sampled nonphosphorylated sequences, respectively. Negative sequences were centered on S, T or Y residues and sampled from the proteome exclusive of experimentally confirmed phosphosites. Mixture models M were fitted using model-based clustering in the mclust R package[17] that infers an optimal number of components of the mixture $M_1, M_2, \ldots, M_n$ and their weights and parameters (means, s.d.) with maximum-likelihood estimation. For a given MSS score s, the probability density function of a GMM was computed as

$$\mathcal{L}(s \mid M) = \sum_{i=1}^{n} \mathcal{L}(s \mid M_i) w_i$$

where $\mathcal{L}(s|M_i)$ is the likelihood or density of s in the component M_i of the mixture and w_i is the weight of the ith component so that the mixture weights $w_1, w_2, \ldots, w_n$ sum to 1. Given the learned parameters, mean μ_i and s.d. σ_i of each component, the likelihood was computed using the normal distribution as

$$\mathcal{L}(s \mid M_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(s-\mu_i)^2}{2\sigma_i^2}}$$

To estimate probabilities of mutation-induced phosphorylation loss and gain, we computed Bayesian posterior probabilities of positive M_P and negative M_N distributions given wild type s_w and mutant s_m scores. The posterior probability of the positive distribution M_P given the wild-type sequence s_w was computed as

$$p(M_P \mid s_w) = \frac{\mathcal{L}(s_w \mid M_P) p(M_P)}{\mathcal{L}(s_w \mid M_P) p(M_P) + \mathcal{L}(s_w \mid M_N) p(M_N)}$$

and $p(M_N|s_w)$, $p(M_P|s_m)$ and $p(M_N|s_m)$ were derived similarly. The prior $p(M)$ reflects our belief in the distribution M. We defined the prior of the positive distribution $p(M_P)$ as equal to the AUC of the corresponding kinase PWM and the prior of the negative distribution $p(M_N)$ equal to 1. The AUC quantifies the performance of the kinase PWM in classifying known binding sites from sites of other kinases and thus serves as a measure of confidence in the model.

A loss-of-phosphorylation event was defined as the joint probability of two events, the positive distribution being representative of the wild-type sequence and the negative distribution being representative of the mutant sequence, and vice versa for phosphorylation gain

$$p_{\mathrm{loss}} = p(M_P \mid s_w) \times p(M_N \mid s_m)$$
$$p_{\mathrm{gain}} = p(M_N \mid s_w) \times p(M_P \mid s_m)$$

By default, we consider the joint probabilities greater than 0.5 to call network-rewiring mutations. Additionally, we compute fold changes in MSS scores of wild-type and mutant sequences $\Delta = \mathrm{abs}[\log_2(s_m/s_w)]$ and employ a minimum threshold of $\Delta \geq 1$ to filter events with small differences between s_w and s_m. These parameters can be altered to fine-tune method performance.

Pathway analysis. We carried out a pathway enrichment analysis to evaluate cancer mutations with predicted rewiring of phosphosites. We tested 5,753 protein groups representing pathways and protein complexes of 4,580 proteins from the databases Reactome[25] and CORUM[26] obtained from the g:Profiler[27] web server. Protein groups were filtered to exclude nonphosphorylated proteins, small ($n < 5$ proteins) and large groups ($n > 500$), groups with few pSNVs ($n < 2$) and groups where only one gene was mutated. One-tailed Poisson exact tests were applied to identify groups with excessive network-rewiring pSNVs, using global average number of network-rewiring pSNVs per protein as expected rate. Observed network-rewiring pSNVs per group and number of proteins in the group were used to determine the significance of enrichment. Resulting P values were corrected for multiple testing with the Benjamini-Hochberg method for false discovery rate (FDR). We considered pathways significantly enriched for network-rewiring mutations with FDR $P < 0.01$. Results were visualized with the Enrichment Map app[28] in Cytoscape (http://www.cytoscape.org/) and manually curated to identify functional themes.

MIMP implementation. MIMP is implemented as a web server and an R package with precomputed kinase specificity models for individual kinases and kinase families. MIMP requires two inputs to predict network-rewiring pSNVs in phosphorylation: mutation data of single amino acid substitutions and protein sequence data in FASTA format. The third optional input file includes positions of phosphorylated residues in protein sequences. If no phosphorylation data are provided, MIMP uses all S,T and Y residues as potential phosphorylation sites. Users can adjust the posterior probability cutoff and score fold change to tune results and to choose between kinase-specific and family-wide specificity models for site prediction. Results are returned in a table with lists of mutations and their impact on binding sites, phosphosite positions, and wild type and mutant sites. Browser visualization in the web server and R package enables easier navigation of results: for example, by sorting columns, filtering rows and visualizing motifs. MIMP is available at http://mimp.baderlab.org/. Installation instructions of the R package, documentation and sample data are available online.

Experimental validation. To validate predictions of network-rewiring mutations, *in vitro* kinase binding assays were carried out by the Kinexus Bioinformatics Corporation. Assay conditions for protein kinases were optimized to yield acceptable enzymatic activity and to give high signal-to-noise ratios. Recombinant protein kinases for substrate profiling were cloned, expressed and purified using proprietary methods. Quality control was carried out on all kinases to ensure compliance with acceptable standards. Gamma phosphate–labeled ATP ([γ-^{33}P]ATP) was purchased from PerkinElmer. All other materials were standard laboratory grade. Peptide substrates with 90–98% purity were synthesized by Kinexus (**Supplementary Table 1**). Kinase activities toward their substrates were profiled with radioisotope assays. Double replicates of assays were performed at ambient temperature for 20–40 min in a final volume of 25 µl, including 5 µl of diluted active protein kinase (~10–50 nM final protein concentration in the assay), 5 µl of assay solution of test substrate, 10 µl of kinase assay buffer, and 5 µl of [γ-^{33}P]ATP (250 µM stock solution, 0.8 µCi). Assays were initiated by adding [γ-^{33}P]ATP, which was followed by incubation of reaction mixture at ambient temperature for 20–40 min, depending on the protein kinase tested. Assays were terminated after the incubation period by spotting 10 µl of the reaction mixture onto a multiscreen phosphocellulose P81 plate. The plate was washed three times for 15 min each in a 1% phosphoric acid solution. Radioactivity of the plate was counted in the presence of scintillation fluid with a Trilux scintillation counter.

24. Magrane, M. *Database* **2011**, bar009 (2011).
25. Croft, D. *et al. Nucleic Acids Res.* **39**, D691–D697 (2011).
26. Ruepp, A. *et al. Nucleic Acids Res.* **38**, D497–D501 (2010).
27. Reimand, J., Arak, T. & Vilo, J. *Nucleic Acids Res.* **39**, W307–W315 (2011).
28. Merico, D., Isserlin, R. & Bader, G.D. *Methods Mol. Biol.* **781**, 257–277 (2011).

doi:10.1038/nmeth.3396