Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

PhosphoPICK-SNP: Quantifying the effect of amino acid variants on protein phosphorylation

Ralph Patrick [1,*], Bostjan Kobe [1,2,3], Kim-Anh Lê Cao [4] and Mikael Bodén [1,2,*]

[1] School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, 4072, Australia.
[2] Institute for Molecular Bioscience, The University of Queensland, St Lucia, 4072, Australia.
[3] Australian Infectious Diseases Research Centre, The University of Queensland, St Lucia, 4072, Australia.
[4] The University of Queensland Diamantina Institute, Translational Research Institute, QLD, 4102, Australia.

* To whom correspondence should be addressed.

Associate Editor: Prof. Alfonso Valencia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genome-wide association studies are identifying single nucleotide variants (SNVs) linked to various diseases, however the functional effect caused by these variants is often unknown. One potential functional effect, the loss or gain of protein phosphorylation sites, can be induced through variations in key amino acids that disrupt or introduce valid kinase binding patterns. Current methods for predicting the effect of SNVs on phosphorylation operate on the sequence content of reference and variant proteins. However, consideration of the amino acid sequence alone is insufficient for predicting phosphorylation change, as context factors determine kinase-substrate selection.

Results: We present here a method for quantifying the effect of SNVs on protein phosphorylation through an integrated system of motif analysis and context-based assessment of kinase targets. By predicting the effect that known variants across the proteome have on phosphorylation, we are able to use this background of proteome-wide variant effects to quantify the significance of novel variants for modifying phosphorylation. We validate our method on a manually curated set of phosphorylation change-causing variants from the primary literature, showing that the method predicts known examples of phosphorylation change at high levels of specificity. We apply our approach to data-sets of variants in phosphorylation site regions, showing that variants causing predicted phosphorylation loss are over-represented among disease-associated variants.

Availability: The method is freely available as a web-service at the website `http://bioinf.scmb.uq.edu.au/phosphopick/snp`

Contact: m.boden@uq.edu.au

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

1 Introduction

The identification of genetic variants linked to disease is transforming the biomedical research landscape. Genome wide association studies (GWAS) have been identifying numerous single nucleotide variants (SNVs) over-represented in patients within a wide variety of diseases including cancer. While many SNVs are being discovered, the precise effect that they have on resultant RNA or protein products is generally not known. One of the potential effects of non-synonymous SNVs (nsSNVs) on protein function is the disruption of post-translational modifications (Kim *et al.*, 2015). As phosphorylation is the most ubiquitous modification, the potential for phosphorylation sites to be affected by amino acid variants is high. For example, the PhosphoSitePlus® database (Hornbeck *et al.*, 2015) has identified numerous sequence variants that fall within the immediate vicinity of a phosphorylation site, and the recent analysis of cancer driver mutations has implicated phosphorylation as being a major factor in

understanding the disruption of signalling pathways caused by amino acid variations (Reimand *et al.*, 2013).

There have been numerous examples of disease-associated naturally occurring variants that impact the phosphorylation status of proteins. The majority of such examples have involved a variant disrupting a phosphorylation site in the reference protein, though there have been at least two examples of missense mutations found to introduce phosphorylation sites (Gentile *et al.*, 2008; Lagarde *et al.*, 2012). While there have been relatively few studies experimentally determining the effect of naturally occurring variants on phosphorylation, there are tens of thousands of nsSNVs that have the potential to impact phosphorylation. The PhosphoSitePlus® PTMVar dataset (Hornbeck *et al.*, 2015) , which is comprised of missense mutations cross-referenced to post-translational modifications, contains over 19,000 examples of variants falling within a 15-residue window surrounding a known phosphorylation site. Such variants have the potential to disrupt existing phosphorylation sites, but there will be many additional variants with the potential to introduce new phosphorylation sites. The PTM-SNP database collates variants that occur in the vicinity of a number of post-translational modifications, including phosphorylation (Kim *et al.*, 2015).

There have also been databases developed that catalogue the predicted effect of SNVs on potential phosphorylation sites. Ryu and colleagues defined the term "phosphovariant" to refer to a mutation that impacts the phosphorylation status of an amino acid (Ryu *et al.*, 2009). To predict examples of phosphovariants, they developed PredPhospho, a support vector machine model that predicts kinase-specific phosphorylation sites based on the amino acid motifs surrounding potential phosphorylation sites. Applying PredPhospho to missense mutations obtained from Swiss-Prot, they predicted examples of phosphovariants and incorporated them into the PhosphoVariant database (Ryu *et al.*, 2009). The PhosSNP database is another example of cataloging variants predicted to modify protein phosphorylation (Ren *et al.*, 2010). Ren and colleagues employed the GPS 2.0 software, a kinase-specific phosphorylation site predictor that uses optimised substitution matrices (Xue *et al.*, 2008). The GPS 2.0 predictor was applied to variants from the dbSNP database (Sherry *et al.*, 2001), with the variants predicted to cause a change in phosphorylation status or to cause a change in the kinase targeting the phosphorylation site compiled into the PhosSNP database.

Most recently, the MIMP (mutation impact on phosphorylation) method has been developed, which uses position weight matrices and Gaussian mixture models to score the probability that a variant will cause loss or gain of phosphorylation (Wagih *et al.*, 2015). In contrast to the other methods, MIMP provides a prediction service rather than a database. For the purpose of consistency with the most recently published work, we will consider two classes of "phosphovariants": phosphorylation-loss causing variants and phosphorylation-gain causing variants.

The current methods for predicting the effect of nsSNVs on phosphorylation, described above, operate on the sequence content surrounding a potential phosphorylation site. While methods based on linear motifs can predict the potential for a kinase binding site to be disrupted (Kobe and Bodén, 2012), the presence of a valid kinase-substrate binding motif on a protein is no guarantee that a kinase will come into contact with the protein (Zhu *et al.*, 2005). We have previously developed a method, PhosphoPICK, for predicting kinase substrates using protein-protein interaction networks and protein abundance across the cell cycle. The use of such context information can improve the prediction accuracy of kinase-specific phosphorylation site prediction from sequence by over two-fold at low false-positive levels (Patrick *et al.*, 2015). An approach that integrates cellular context information with sequence information should therefore be able to provide a more accurate assessment of the effect of SNVs on phosphorylation than methods that operate on sequence alone.

Building on the properties of PhosphoPICK, we present here a method for quantifying the effect of nsSNVs on protein phosphorylation status. Taking stock of known missense mutations across the proteome, as collected in UniProt, we use PhosphoPICK to build kinase-specific, proteome-wide sets of predicted variant effects on phosphorylation. These sets provide a "background distribution" that can be used to calculate a measure of significance for the predicted effect that a novel variant has on phosphorylation loss or gain.

In order to validate our approach, we searched the literature for naturally occurring variants causing phosphorylation loss or gain, identifying 24 such variants. By comparing the threshold at which our method detects true positives against that of the background, we demonstrate that our method is able to detect over 50% of the known phosphovariants within the first 2% of the background distribution. We further validate our method against three restricted sets of simulated negatives based on structural and network features, and show that our method maintains high prediction accuracy with an average AUC of 0.94 across all tests performed. These results demonstrate the method's reliability in detecting true examples of differential phosphorylation from the vast number of potential phosphovariants. Applying the method to variants in the vicinity of phosphorylation sites from the PhosphoSitePlus® PTMVar dataset (Hornbeck *et al.*, 2015), we find that the predicted phosphovariants are over-represented among the ones with disease annotations. These results support the conclusion that our method, named PhosphoPICK-SNP, is able to detect variants that have functional significance.

2 Methods

2.1 Data Resources

2.1.1 Missense mutation data

We obtained the UniProt index of protein altering variants (Consortium, 2015), which maps dbSNP variants (Sherry *et al.*, 2001) to proteins within the UniProt database (downloaded March, 2015). This file contained 752,857 variants mapped to amino acid variants in UniProt proteins. The variants covered 89,909 protein sequences in the UniProt database.

2.1.2 Phosphorylation sites affected by naturally occurring variants

Through a manual search of the literature, we compiled a list of naturally occurring variants that were found experimentally to either disrupt or introduce a phosphorylation site. For the purpose of this work we included variants that were shown either *in vivo* or *in vitro* to affect the phosphorylation of a specific site, or the ability of a kinase to bind to the site. Although there are examples of studies showing changing phosphorylation levels on the protein, we only recorded examples where the precise phosphorylation site was known. Table 1 contains the list of identified genes, with variant and phosphorylation site affected. We found 21 examples of phosphorylation loss and 3 examples of phosphorylation gain in response to nsSNVs. Of the 21 loss-causing variants, 6 of the mutations are on the phosphorylation site. The total 24 variants are across 21 unique proteins.

2.2 Building background distributions of variant effects

We built distributions of predicted variant effects on phosphorylation in a kinase-specific manner across all protein altering variants. PhosphoPICK employs two Bayesian network models to make predictions. The first model classifies kinase-substrate binding sites from sequence, and incorporates position-specific amino acid frequencies and counts of co-occurring neighbouring amino acids within some m length window surrounding a potential phosphorylation site (Patrick *et al.*, 2016). This

model is henceforth referred to as the *sequence model*. Separately, a Bayesian network model integrates the sequence model with protein-protein interaction and association data sourced from BioGRID (Chatr-aryamontri *et al.*, 2015) and STRING (Franceschini *et al.*, 2013), as well as protein abundance data across the cell cycle (Olsen *et al.*, 2010), in order to calculate the probability that a kinase ordinarily targets a given protein (Patrick *et al.*, 2016). This model is henceforth referred to as the *combined model*. When scoring the effect of a variant we use PhosphoPICK to generate three scores: (1) $R_{substrate}$, the prior probability based on the combined model that the kinase would be expected to target the reference protein, (2) R_{site} the probability according to the sequence model that the kinase will phosphorylate the site of interest on the reference protein, and (3) V_{site} the probability that the kinase will target the site of interest on the variant protein.

Kinases within PhosphoPICK contain different optimal binding site windows that are considered when making a prediction for a potential phosphorylation site. Therefore, given a query kinase, we checked for variants that fell within a window surrounding a potential phosphorylation site. For each potential phosphorylation site, we recorded a reference peptide and a variant peptide containing the missense mutation. We then used the sequence model to obtain the R_{site} and V_{site} scores from the reference and variant peptides respectively. If the central residue for a peptide is not a valid phosphorylation site (for example a threonine is mutated to an arginine) it will be scored 0. We defined a score difference,

$$D_{site} = V_{site} - R_{site} \tag{1}$$

where a negative value of D_{site} indicates the variant is predicted to cause *decreased* probability of phosphorylation, and a positive value represents an *increased* probability of phosphorylation.

We calculated distributions of D_{site} values in a kinase-specific manner across all potential phosphorylation sites that contained a missense mutation within the window for the query kinase. A potential phosphorylation site is defined as any serine (S) or threonine (T) residue for S/T kinases, any tyrosine (Y) residue for Y kinases, or any S/T/Y residue for dual specificity kinases.

2.3 Calculating variant significance

The significance of the effect on phosphorylation by a variant is calculated in a kinase-specific manner, as described by the following procedure. Given some kinase K, an m length window corresponding to K is centred on potential phosphorylation sites within the protein sequence, where if the variant falls within a window, m length reference and variant peptides are retained. D_{site} is then calculated from the reference and variant peptides using Equation 1. The difference is then compared to the background distribution and a P-value from both tails of the distribution is calculated – representing whether the difference is greater (increased probability of phosphorylation) or less (decreased probability of phosphorylation) than would be expected by chance. The P-values are calculated such that

$$P_{loss} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(D_i \leq D_{site}) \tag{2}$$

$$P_{gain} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(D_i \geq D_{site}) \tag{3}$$

where n is the number of variants contained in the background distribution for kinase K and $\mathbf{I}(\cdot)$ is the indicator function. The final P-value representing the site, P_{site}, is calculated as the minimum of P_{loss} and P_{gain}.

The $R_{substrate}$ context score for the query protein is then retrieved. As for the D_{site} scores, we have distributions of context scores across the proteome for each kinase. We therefore calculate an empirical P-value for the $R_{substrate}$ score, $P_{substrate}$, based on a count of the proteome-wide context scores that are greater than or equal to $R_{substrate}$, using the same form as Equation 3. We then use Fisher's method to combine the two P-values into a combined P-value that represents the confidence of the variant effect size given both the difference in sequence scores and likelihood that the reference protein would ordinarily be a substrate of the query kinase. Given the P-values P_{site} and $P_{substrate}$, we calculate:

$$X = -2(ln(P_{substrate}) + ln(P_{site})) \tag{4}$$

where X follows a Chi squared distribution with 4 degrees of freedom. The combined P-value, $P_{combined}$, can then be derived from X. As a single phosphovariant can be scored with all kinases available to PhosphoPICK (currently numbering 107), we correct the P-value for multiple testing using a Bonferroni multiple correction on P_{site} and $P_{combined}$ to obtain E_{site} and $E_{combined}$.

Table 1. Naturally occurring variants that have been shown through in vivo or in vitro experiments to affect the phosphorylation status of the proteins, by modifying the phosphorylateable residue or an adjacent residue. The effect can be to disrupt an existing phosphorylation site (loss), or introduce a new one (gain).

Gene	Variant	Phos. site	Effect	Reference
Cyclin D1	T286R	T286	loss	Benzeno *et al.* (2006)
hOG1	S326C	S326	loss	Luna *et al.* (2005)
TP53	P47S	S46	loss	Li *et al.* (2005)
TP53	R213Q	S215	loss	Wagih *et al.* (2015)
TP53	R282W	T284	loss	Wagih *et al.* (2015)
BDNF	V66M	T62	loss	Deng *et al.* (2013)
CDKN1A	D149G	S146	loss	Oh *et al.* (2007)
hERG1	K897T	T897	gain	Gentile *et al.* (2008)
PPARγ2	P113Q	S112	loss	Ristow *et al.* (1998)
PTP-1B	P387L	S386	loss	Echwald *et al.* (2002)
UBE3A	T485A	T485	loss	Yi *et al.* (2015)
PER2	S662G	S662	loss	Toh *et al.* (2001)
MeCP2	R306C	T308	loss	Ebert *et al.* (2013)
NKX3-1	R52C	S48	loss	Gelmann *et al.* (2002)
PLN	R14C	S16	loss	Ceholski *et al.* (2012)
ABCB4	T34M	T34	loss	Gautherot *et al.* (2014)
MAF	P59H	T58	loss	Niceta *et al.* (2015)
GLUT1	R223W	S226	loss	Lee *et al.* (2015)
AR	R405S	S405	gain	Lagarde *et al.* (2012)
Gab1	T387N	T387	loss	Ortiz-Padilla *et al.* (2013)
STAT1	L706S	Y701	loss	Dupuis *et al.* (2001)
CLIP1	E1012K	S1009	loss	Wagih *et al.* (2015)
CTNNB1	S37C	S33	loss	Wagih *et al.* (2015)
CTNNB1	G34R	S47	gain	Wagih *et al.* (2015)

2.4 Evaluating method accuracy on known variants

In order to calculate an estimate of the number of potential phosphorylation sites that were affected by the presence of a nearby variation, we used a 10-fold cross-validation approach to build a set of predicted background values. The proteins within the background set were split into 10 partitions, where 9 of the partitions were used to construct distributions for both the context scores and the D_{site} values. These distributions were then used to evaluate and obtain E-values for the variants in the remaining partition. For each variant the lowest E-value was retained as representing the greatest likelihood that the mutation resulted in a change in phosphorylation status.

To evaluate our method on its ability to detect the examples of differential phosphorylation recorded in Table 1, we evaluated the known

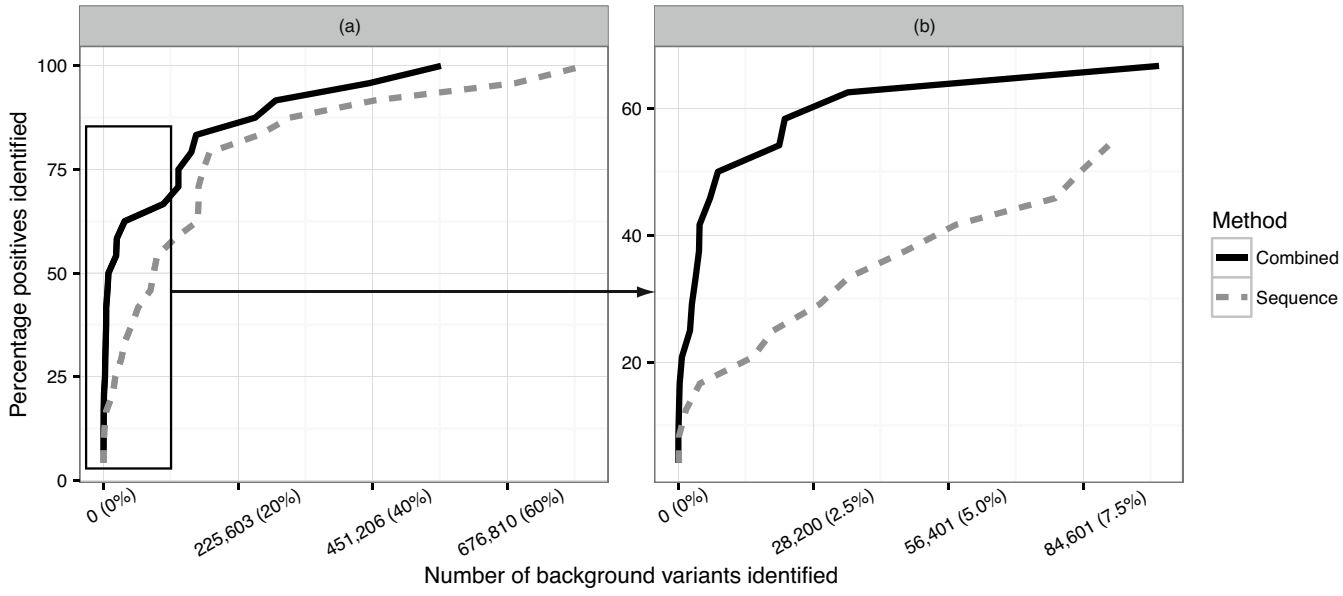

Fig. 1. Line-curves showing the tradeoff between the percentage of positive differential phosphorylation examples identified and the number of variants considered (as the E-value cut-off decreases). Comparison is made between predicting phosphorylation change using sequence alone, and combining sequence with context. Shown is the tradeoff until all positive examples are detected (a), as well as the tradeoff up until 10% of the background variants are detected (b).

variants on our method using each of the 10 partitions from the cross-validation test to construct the background distributions. For each variant we calculated the median of the E-values generated across the cross-validation runs; similar to the background, the final E-value assigned to a variant was the minimum of the E-values for all potential kinases. Due to the difficulty of defining a true negative set, we first compared to the full background set the E-value thresholds at which the true positives were identified; i.e. at each E-value threshold calculated for a true positive, we calculated the number and percentage of variants in the background set that were also identified at that threshold. We performed this test using both E_{site} and $E_{combined}$ values to understand the influence of context on predicting phosphorylation change.

We further evaluated our method on several sub-sets of the background set, chosen to simulate negatives based on the following assumptions. The primary assumption is that by defining criteria, based on structural or network data, that preclude the occurrence of phosphorylation at a given site, the occurrence of a variant will not alter the phosphorylation status of the site. For the purpose of this work, we define three non mutually-exclusive criteria: phosphorylation sites are less likely to occur (1) in solvent inaccessible/buried regions of a protein; (2) in transmembrane domains; (3) in proteins that do not interact either directly (first-order interactions) or through mediators (second-order interactions) with a kinase. Buried amino acids were predicted using the ACCpro 5 method (Magnan and Baldi, 2014) from the SCRATCH protein structure prediction suite of programs (Cheng *et al.*, 2005). Transmembrane domains were predicted using TMHMM 2.0 (Krogh *et al.*, 2001). Proteins that have no first or second-order interactions with a kinase in our set were identified based on a PPI network from BioGRID (Chatr-aryamontri *et al.*, 2015). We note that as our method includes BioGRID data in its predictions, this could introduce a positive bias; however, as our method does not include structural data, the first two sets will not exhibit such bias. We generated the three negative sets as follows. For the structural data, we first defined a 15-residue window around a potential phosphovariant. For set 1, we defined as negatives the potential phosphovariants whose 15-residue window fell in a predicted buried region; i.e. all 15 amino acids were required to be buried. For set 2, negatives were defined as potential phosphovariants with the 15-residue window falling in a predicted transmembrane helix. Set 3 was constructed by taking potential phosphovariants that occur in proteins with no first or second-order PPIs with a kinase in our set.

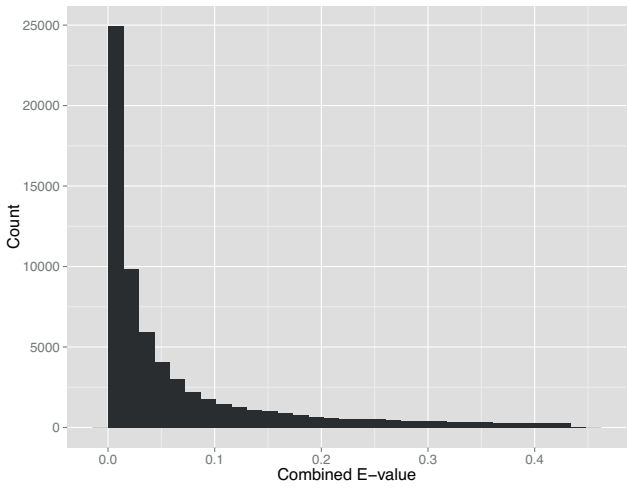

Fig. 2. Histogram showing the distribution of combined E-value ($E_{combined}$) scores for the subset of 65,203 variants predicted to be phosphovariants based on sequence alone ($E_{site} < 0.05$).

We compared our method's ability to detect the known variants against that of the MIMP predictor (Wagih *et al.*, 2015). We downloaded the local version of the software, and ran the background set of protein sequences and variants through it, specifying cutoff probability and log values of 0 to enable a comparison over all thresholds.

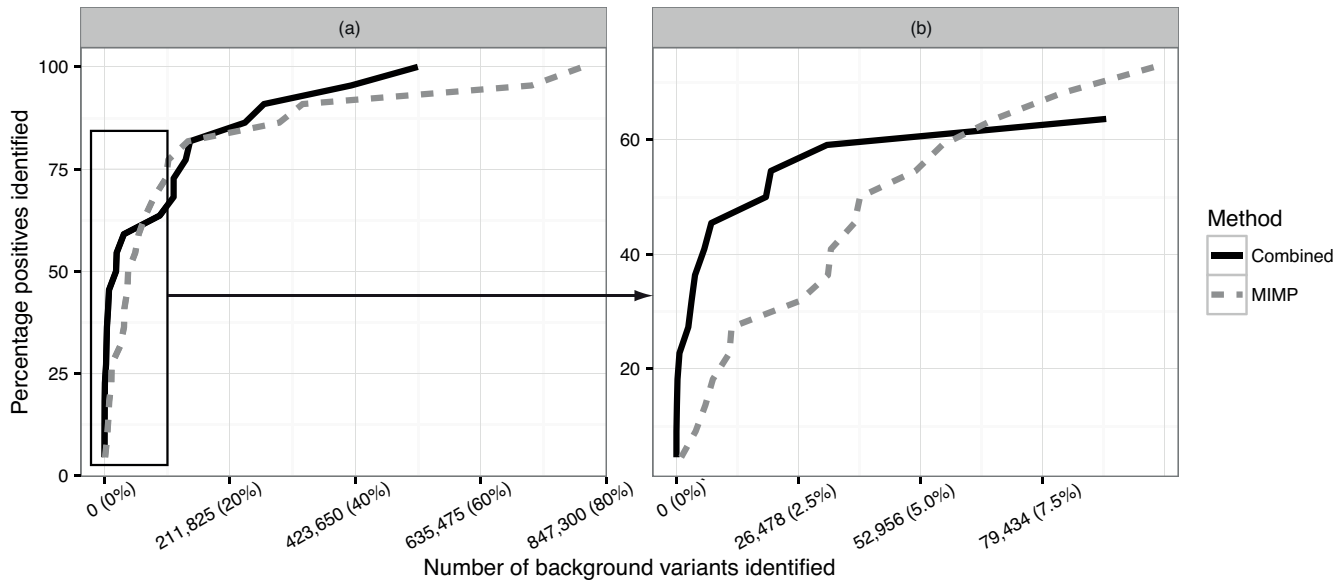

Fig. 3. Line-curves showing a comparison of detecting experimentally confirmed phosphovariants between the combined PhosphoPICK-SNP method and MIMP (Wagih et al., 2015). Shown is the tradeoff until all positive examples are detected (a), as well as the tradeoff up until 10% of the background variants are detected (b).

3 Results

The experimentally determined examples of differential phosphorylation listed in Table 1 were used to gauge how well our method performed in identifying real examples of phosphorylation gain and loss. Figure 1 shows a tradeoff between the percentage of known positives detected and the background at each E-value threshold a positive was discovered at, with the values for each threshold recorded in Supplementary Table 1. When predicting phosphovariants using the combined E-value, we found that the majority (over 50%) of the known positives could be identified within the first 2% of the background distribution. We were able to identify 75% of the experimental examples at an E-value threshold corresponding to 11% of the background. These results demonstrate that the method can identify true positive examples of phosphovariants at high levels of specificity, which represent candidates of real interest to biologists.

We also evaluated the use of sequence only for predicting phosphovariants (i.e. using the E_{site} value), in order to determine if the incorporation of context information was providing an increase in prediction accuracy. When using sequence alone, the majority of variants were not detected until 8% of the background distribution was reached (Figure 1). Given the combined method detected the majority of variants at 2% of the background, this represents a 4-fold increase when using the combined E-value. As can be seen from Figure 1(a), at the more liberal E-value thresholds there was less difference between sequence alone and the combined E-values. However, these results show that the approach of combining context and sequence information provides the greatest benefit for identifying true variants at higher levels of specificity.

We next compared our method's prediction accuracy for classifying the positive phosphovariant examples against three simulated negative sets. The negative sets were defined by the occurrence of a potential phosphovariant in (1) a solvent inaccessible/buried region, (2) a transmembrane helix and (3) a protein that has no first or second-order PPIs with a kinase in our set. We evaluated prediction accuracy using area under the receiver operating characteristic curve (AUC). Supplementary Table 2 shows AUC values for the various comparisons. We found prediction accuracy as measured by AUC was high on the test-sets of simulated negatives. The AUC for evaluating the predictions on PPI network-based simulated negatives was 0.94 (with 216,630 negatives), 0.95 for buried regions (35,312 negatives), and 0.95 for those based on transmembrane domains (12,741 negatives). These values were consistently higher than those of using sequence alone, with AUC values ranging from 0.86 to 0.87. These high AUC values across multiple test sets demonstrate the ability of our approach to distinguish true phosphovariants from high confidence negatives.

3.1 Estimating phosphorylation sites affected by SNVs

In order to investigate the effect of context on predicting differential phosphorylation, we used the methods for calculating E_{site} and $E_{combined}$ to estimate the number of putative phosphorylation sites affected by the nsSNVs contained in the UniProt index of protein altering variants. We performed two tests: firstly, we identified predicted differentially-phosphorylated sites on the basis of E_{site}, where if E_{site} fell below 0.05 the variant was considered to cause differential phosphorylation; i.e. a phosphovariant. In the second test, the $E_{combined}$ value was applied as a filter, where only variants with $E_{combined}$ and E_{site} falling below 0.05 were classified as a phosphovariant.

Based on our cross-validated analysis of the background distribution, we identified the variants that were predicted to be causing differential phosphorylation. In total we found 65,203 variants that were predicted, based on their E_{site} value, to cause differential phosphorylation. When requiring that a variant obtain an E-value < 0.05 for both E_{site} and $E_{combined}$, the number dropped to 41,075. Figure 2 shows a histogram of the $E_{combined}$ values calculated for all the variants that were found to be significant based on E_{site} alone. While the majority maintain a high level of significance when context is included, nearly 40% of the variants obtained an E-value > 0.05 after context is included. These results illustrate the effect that context has in filtering out spurious examples of phosphovariants where the kinase is unlikely to target the query protein.

3.2 Comparison with alternative method

We compared the ability of the MIMP method (Wagih *et al.*, 2015) to predict the set of positives out of the background to our combined method.

As MIMP was unable to make predictions for two of the phosphorylation gain sites (due to the centre residue of the reference protein being non-phosphorylatable) we performed the comparison using the remaining 21 phosphorylation loss-causing variants and 1 gain-causing variant. As can be seen from Figure 3 (the individual comparison values are also recorded in Supplementary Table 3), at stricter cut-off thresholds our method is able to detect greater numbers of the true positive examples. Within 2% of the background distribution our method is able to detect 55% of the 22 phosphovariants, however MIMP does not reach 55% until 4.9% of the background – this corresponds approximately to an additional 31,000 variants. We also performed comparisons using our 3 sets of simulated negatives. We found that our combined method consistently out-performed MIMP across the test sets, with an average AUC of 0.94 for our method compared to 0.91 for MIMP (Supplementary Table 2).

3.3 Phosphorylation loss in disease

We used our method to determine whether the variants that were most confidently predicted to result in a change in phosphorylation status were over-represented among disease-associated variants. We used the PTMVar database from PhosphoSitePlus®, which cross-references post-translational modification information from PhosphoSitePlus® with variant information from the UniProt human variation database. The PTMVar database annotates sites with the classification 'Disease', 'Polymorphism' or 'Unclassified'. Variants that were within the vicinity of phosphorylation sites and were annotated with either 'Disease' or 'Polymorphism' were selected. We then counted the number of times that a variant in each of these classes was predicted to be differentially phosphorylated with a decreased probability of phosphorylation (i.e. it obtained an $E_{combined}$ value < 0.05 in a test for decreased probability), and counted the number of times the variants in both classes were not predicted to be differentially down-phosphorylated.

Fisher's exact test was used to determine the over-representation. We found that variants annotated as disease-associated were significantly over-represented among the variants predicted to result in down-phosphorylation, with a P-value of 0.0002. This indicates that while the presence of a variant in the vicinity of a phosphorylation does not necessarily result in phosphorylation disruption, our method is able to detect the disease-associated variants that will have a strong impact on phosphorylation.

To further investigate our method's ability to identify disease-relevant examples of phosphorylation loss, we compared our predictions with a set of 58 genes previously found to have significant levels of cancer-associated mutations in phosphorylation site regions (Reimand and Bader, 2013). We found that 47 out of the 58 genes contained at least one predicted phosphovariant, with an $E_{combined}$ value under 0.05 in a test of phosphorylation loss. A permutation analysis based on random sampling of 58 proteins from our background set confirmed that finding 47 out of 58 proteins containing predicted phosphovariants is greater than would be expected by chance ($P = 1e^{-05}$). There was a total of 1025 predicted phosphovariants distributed across the 47 proteins. A second permutation test counting the number of predicted phosphovariants from randomly sampled sets of 58 proteins showed that this number of phosphovariants was also greater than expected by chance ($P < 1e^{-100}$).

3.4 Prediction of phosphorylation disruption in disease-associated sites

Given that our method is reliably able to detect phosphorylation loss events, we used it to identify the most likely examples of phosphorylation loss in the PhosphoSitePlus® PTMVar database that were associated with at least one of five cancer types: ovarian, breast, colorectal, liver and pancreatic. These variants were run through our method, and variants that obtained E-values below 0.05 for both E_{site} and $E_{combined}$ were retained. Table 2 lists top scoring variants with their disease associations, where the variant has been mapped to the vicinity of a phosphorylation site. The full list of variants is available in Supplementary Table 4. In total, we found 52 examples of predicted phosphorylation loss caused by variants related to ovarian cancer, 12 for breast cancer, 8 for colorectal cancer, 19 for liver cancer and 9 for pancreatic cancer. We found that the kinases most frequently associated with the phosphovariants were DNAPK, CK2A1, CaMK4 and GSK3B, which were each predicted to be targeting five of the phosphovariants (Supplementary Table 5).

We found several examples of predicted phosphorylation loss on the β-catenin protein (Uniprot accession number P35222), which was a top candidate for phosphorylation loss for both ovarian and liver cancer. The T41 phosphorylation site, which has been previously identified as a GSK3B target (van Noort *et al.*, 2002), is a known site mutated in cancers (Sagae *et al.*, 1999). It is predicted by PhosphopICK-SNP that the T41A mutation would abolish a GSK3B phosphorylation site at T41 (Table 2)

There was also an example of predicted phosphorylation loss on tumour suppressor protein p53 (Uniprot accession number P04637), which has been shown previously to lose phosphorylation as a consequence of a P47S mutation (Li *et al.*, 2005). The E271K variant, which was associated with both pancreatic and colorectal cancer (Sjöblom *et al.*, 2006), was found to have a significant likelihood of disrupting the phosphorylation site at S269. The phosphorylation site at S269 is known to be an important regulator of p53 transcriptional activity (Wu *et al.*, 2011).

4 Discussion

With increasing numbers of disease-associated variants being catalogued, the need for reliable functional annotations is only going to continue to grow. While there are many potential functional effects of gene-coding variants on protein function (Hecht *et al.*, 2013), including the perturbation of protein stability (Pires *et al.*, 2014) or the disruption of one of the many post-translational modifications that proteins undergo (Kim *et al.*, 2015), phosphorylation is a high-probability target of disruption due to the ubiquitous nature of this protein modification process (Hornbeck *et al.*, 2015). We have presented here a method for quantifying the expected effect of nsSNVs on protein phosphorylation, and have demonstrated that it detects experimentally confirmed examples of phosphovariants at high levels of specificity.

An advantage of our approach is the consideration of the cellular context that kinases and their substrates operate in. We have shown that by incorporating context into the prediction of phosphovariants, we can identify positive examples of phosphovariants at higher levels of specificity than if using sequence alone. There are examples of phosphovariants that represent a trivial loss of phosphorylation; the removal of a phosphorylated serine, threonine or tyrosine residue will by definition cause loss of phosphorylation. A method that operates only on sequence may be able to correctly predict such cases, but introduce false-positive predictions for cases where the mutation occurs on a residue adjacent to the phosphorylation site. Given the small number of trivial losses contained in our test set, the specificity increase gained by incorporating context into predictions indicates that our method is able to predict a broader spectrum of potential phosphovariants than by using sequence alone. In addition, when comparing our method to an alternative method of predicting the effect of variants on phosphorylation, MIMP, we found that we could predict positive examples of phosphorylation loss at stricter specificity levels than the MIMP method.

While there are over 19,000 examples of missense mutations in the vicinity of a phosphorylation site according to the PhosphoSitePlus® PTMVar dataset (Hornbeck *et al.*, 2015), we found that the mutations with the strongest propensity for causing phosphorylation loss were

Table 2. Cancer-associated variants predicted to cause loss of phosphorylation. Variants are listed according to the cancer or disease they are associated with. Each row contains protein name as UniProt accession, the location of the variant and phosphorylation site, the kinase predicted to target the site, the reference and variant scores for the peptide. The five most significant (according to $E_{combined}$ score) variants for each cancer type are shown, while the full list of significant variants for each cancer is shown in Supplementary Table 4. Where a variant is predicted to affect multiple kinases, or multiple phosphorylation sites, each prediction is listed separately.

Cancer type	Protein	Variant	Phos. site	Kinase	$R_{subst.}$	R_{site}	V_{site}	$E_{combined}$	Peptide
Ovarian	P35222	G555A	T551	Akt2	1.0	1.0	4.95E-05	9.36E-09	QDTQRR**T**$_p$SMG[G/A]TQ
	P26010	Y753H	Y753	FAK	1.0	1.0	0.0	2.23E-08	YRLSVEI[**Y**$_p$/H]DRREYSR
	Q7KZI7	S197N	S197	NEK6	1.0	1.0	0.0	2.3E-07	KIADFGF[**S**$_p$/N]NEFTFGN
	P51813	S212R	S212	GSK3B	0.998	1	0	1.33e-07	PPSSST[**S**$_p$/R]LAQYDS
	P46939	M1256R	T1259	MARK2	1	0.914	0.0005	4.47e-05	R[M/R]KS**T**$_p$EVLP
Breast	P14859	S88F	S88	DNAPK	1.0	1.0	0.0	8.6E-06	SQQPSQP[**S**$_p$/F]QQPSVQA
	P43355	K278T	Y276	Brk	0.998	1	0.000491	0.000103	RALAETS**Y**$_p$V[K/T]VLEYV
	P03372	H6Y	T2	VRK1	0.0492	0.0792	0.00421	0.00111	M**T**$_p$MTL[H/Y]TKA
	Q99490	D816Y	S818	P38B	0.0186	0.587	0.000173	0.00126	CTPSG[D/Y]L**S**$_p$PLSREPP
	P54646	S523G	S527	p90RSK	0.36	0.78	0.00501	0.00214	LTG[S/G]TLS**S**$_p$VSPRLGS
Colorectal	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN**S**$_p$F[E/K]VRVC
	Q9P253	A913S	S912	ERK5	0.495	0.848	0.0394	0.00475	APPPAKG**S**$_p$[A/S]RAKEAE
	Q9NPD5	I292M	S293	CaMK4	0.832	0.524	3.21e-08	0.00785	ERK[I/M]**S**$_p$LSLH
	Q6ZMN7	G784R	S783	CaMK4	0.792	0.454	1.52e-06	0.00954	TQSS**S**$_p$[G/R]QSS
	Q92953	V450I	S448	ROCK1	0.326	0.845	0.00479	0.012	RAKRNG**S**$_p$I[V/I]SMNL
Liver	P35222	T41A	T41	GSK3A	1	1	0	2.28e-09	GIHSGAT[**T**$_p$/A]TAPSLSG
	P35222	S37F	S37	GSK3A	1	1	0	2.28e-09	YLDSGIH[**S**$_p$/F]GATTTAP
	P35222	T41A	T41	IKKA	1	1	0	7.97e-09	GIHSGAT[**T**$_p$/A]TAPSLSG
	P35222	S37F	S37	IKKA	1	1	0	7.97e-09	YLDSGIH[**S**$_p$/F]GATTTAP
	P35222	T41A	T41	GSK3B	1	0.997	0	8.13e-06	IHSGAT[**T**$_p$/A]TAPSLS
Pancreatic	Q9BYV9	T519I	T519	p70S6K	1	1	0	1.14e-08	LETRTR[**T**$_p$/I]SSSCSS
	P04637	E271K	S269	CAMK2A	1	0.781	0.011	0.000131	NLLGRN**S**$_p$F[E/K]VRVC
	Q9BYV9	T519I	S525	p70S6K	1	0.84	0.0917	0.000294	[T/I]SSSCSS**S**$_p$YSYAED
	P56715	A135V	S137	MARK2	0.848	0.987	2.83e-06	0.000418	IS[A/V]H**S**$_p$PPHP
	P05129	P524R	Y521	Brk	0.356	0.227	0.000117	0.000629	TFCGTPD**Y**IA[P/R]EIIA

associated significantly with disease annotations. While computational analysis of variants has predicted both phosphorylation loss and gain to be associated with disease (Radivojac *et al.*, 2008), this study represents an analysis of the predicted effect of variants on experimentally determined phosphorylation sites. However, a greater availability of experimentally identified phosphorylation gain-causing variants would enable an analysis on the link between phosphorylation gain and disease. There are at least three experimentally identified examples of naturally occurring variants causing a gain of phosphorylation (Table 1). In the human ERG1 (hERG1) protein for example, a lysine to threonine mutation at position 897 (K897T) introduces a phosphorylation site at the mutated residue (Gentile *et al.*, 2008). Similarly, a mutation on the human androgen receptor gene, R405S, induces a phosphorylation site on residue 405 (Lagarde *et al.*, 2012).

There are key residues within a kinase-substrate binding motif that determine the ability of a kinase to catalyse a phosphorylation modification (Brinkworth *et al.*, 2003; Kobe *et al.*, 2005). The mutation of these key residues can disrupt the phosphorylation site, and specific effects will depend on the associated kinase. For example, the loss of a proline at the +1 position relative to a phosphorylation site in a proline-directed kinase-substrate binding motif will cause loss of phosphorylation (Ristow *et al.*, 1998; Li *et al.*, 2005; Niceta *et al.*, 2015). However, the mutation of alternative, non-key, residues within the motif would not be expected to disrupt the phosphorylation site the same extent. As a result, it is to be expected that many missense mutations, even if they are in the vicinity of a phosphorylation site, will not cause a loss of phosphorylation. Our results indicate that PhosphoPICK-SNP is able to detect the mutations that do have an impact on phosphorylation, and therefore have a greater likelihood of being associated with disease.

5 Availability

The PhosphoPICK-SNP web-service takes as input protein sequences in Fasta format, and information defining the mutation occurring in the proteins. This follows the format used for missense mutations in Tables 1 and 2, for example S523G. Users choose which kinases to make predictions for, and select an E-value threshold for returning results; results that obtain both E_{site} and $E_{combined}$ values below the threshold will be returned. The output is an interactive table of results which details the context score provided to the protein ($R_{substrate}$), the reference and variant scores (R_{site} and V_{site}) obtained from the potential phosphorylation peptide, the $E_{combined}$ value and the peptide itself. More comprehensive information regarding the variant is available in a downloadable tab-delimited text file of the results. In addition, we encourage researchers to inform us of phosphovariants that they publish. The phosphovariants listed in Table 1 are available online (linked from the PhosphoPICK-SNP web-site), and we plan to update this table with new examples of phosphovariants.

Funding

KALC was supported by the National Health and Medical Research Council (NHMRC) Career Development Fellowship (APP1087415). BK is a NHMRC Research Fellow (1003326 and 1110971).

References

Benzeno, S., Lu, F., Guo, M., Barbash, O., Zhang, F., Herman, J. G., Klein, P. S., Rustgi, A., and Diehl, J. A. (2006). Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. *Oncogene*, **25**(47), 6291–6303.

Brinkworth, R. I., Breinl, R. A., and Kobe, B. (2003). Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl. Acad. Sci. U. S. A.*, **100**(1), 74–79.

Ceholski, D. K., Trieber, C. A., Holmes, C. F. B., and Young, H. S. (2012). Lethal, hereditary mutants of phospholamban elude phosphorylation by protein kinase a. *J. Biol. Chem.*, **287**(32), 26596–26605.

Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M. S., Dolinski, K., and Tyers, M. (2015). The biogrid interaction database: 2015 update. *Nucleic Acids Res.*, **43**(D1), D470–D478.

Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P. (2005). Scratch: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**(suppl. 2), W72.

Consortium, T. U. (2015). Uniprot: a hub for protein information. *Nucleic Acids Res.*, **43**(D1), D204–D212.

Deng, F.-Y., Tan, L.-J., Shen, H., Liu, Y.-J., Liu, Y.-Z., Li, J., Zhu, X.-Z., Chen, X.-D., Tian, Q., Zhao, M., and Deng, H.-W. (2013). Snp rs6265 regulates protein phosphorylation and osteoblast differentiation and influences bmd in humans. *J. Bone Miner. Res.*, **28**(12), 2498–2507.

Dupuis, S., Dargemont, C., Fieschi, C., Thomassin, N., Rosenzweig, S., Harris, J., Holland, S. M., Schreiber, R. D., and Casanova, J.-L. (2001). Impairment of mycobacterial but not viral immunity by a germline human stat1 mutation. *Science*, **293**(5528), 300–303.

Ebert, D. H., Gabel, H. W., Robinson, N. D., Kastan, N. R., Hu, L. S., Cohen, S., Navarro, A. J., Lyst, M. J., Ekiert, R., Bird, A. P., and Greenberg, M. E. (2013). Activity-dependent phosphorylation of mecp2 threonine 308 regulates interaction with ncor. *Nature*, **499**(7458), 341–345.

Echwald, S. M., Bach, H., Vestergaard, H., Richelsen, B., Kristensen, K., Drivsholm, T., Borch-Johnsen, K., Hansen, T., and Pedersen, O. (2002). A p387l variant in protein tyrosine phosphatase-1b (ptp-1b) is associated with type 2 diabetes and impaired serine phosphorylation of ptp-1b in vitro. *Diabetes*, **51**(1), 1–6.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**(D1), D808–D815.

Gautherot, J., Delautier, D., Maubert, M.-A., Aït-Slimane, T., Bolbach, G., Delaunay, J.-L., Durand-Schneider, A.-M., Firrincieli, D., Barbu, V., Chignard, N., Housset, C., Maurice, M., and Falguières, T. (2014). Phosphorylation of ABCB4 impacts its function: Insights from disease-causing mutations. *Hepatology*, **60**(2), 610–621.

Gelmann, E. P., Steadman, D. J., Ma, J., Ahronovitz, N., Voeller, H. J., Swope, S., Abbaszadegan, M., Brown, K. M., Strand, K., Hayes, R. B., and Stampfer, M. J. (2002). Occurrence of NKX3.1 C154T polymorphism in men with and without prostate cancer and studies of its effect on protein function. *Cancer Res.*, **62**(9), 2654–2659.

Gentile, S., Martin, N., Scappini, E., Williams, J., Erxleben, C., and Armstrong, D. L. (2008). The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. *Proc. Natl. Acad. Sci. U. S. A.*, **105**(38), 14704–14708.

Hecht, M., Bromberg, Y., and Rost, B. (2013). News from the protein mutability landscape. *J. Mol. Biol.*, **425**(21), 3937 – 3948.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). Phosphositeplus, 2014: mutations, ptms and recalibrations. *Nucleic Acids Res.*, **43**(D1), D512–D520.

Kim, Y., Kang, C., Min, B., and Yi, G.-S. (2015). Detection and analysis of disease-associated single nucleotide polymorphism influencing post-translational modification. *BMC Med. Genomics*, **8**(Suppl 2), S7.

Kobe, B. and Bodén, M. (2012). Computational modelling of linear motif-mediated protein interactions. *Curr. Top. Med. Chem.*, **12**(14), 1553–1561.

Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P., and Brinkworth, R. I. (2005). Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta*, **1754**(1-2), 200 – 209.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of Molecular Biology*, **305**(3), 567 – 580.

Lagarde, W. H., Blackwelder, A. J., Minges, J. T., Hnat, A. T., French, F. S., and Wilson, E. M. (2012). Androgen receptor exon 1 mutation causes androgen insensitivity by creating phosphorylation site and inhibiting melanoma antigen-a11 activation of NH2- and carboxyl-terminal interaction-dependent transactivation. *J. Biol. Chem.*, **287**(14), 10905–10915.

Lee, E., Ma, J., Sacharidou, A., Mi, W., Salato, V., Nguyen, N., Jiang, Y., Pascual, J., North, P., Shaul, P., Mettlen, M., and Wang, R. (2015). A protein kinase C phosphorylation motif in GLUT1 affects glucose transport and is mutated in GLUT1 deficiency syndrome. *Mol. Cell*, **58**(5), 845 – 853.

Li, X., Dumont, P., Pietra, A. D., Shetler, C., and Murphy, M. E. (2005). The codon 47 polymorphism in p53 is functionally significant. *J. Biol. Chem.*, **280**(25), 24245–24251.

Luna, L., Rolseth, V., Hildrestrand, G. A., Otterlei, M., Dantzer, F., Bjørås, M., and Seeberg, E. (2005). Dynamic relocalization of hOGG1 during the cell cycle is disrupted in cells harbouring the hOGG1-Cys326 polymorphic variant. *Nucleic Acids Res.*, **33**(6), 1813–1824.

Magnan, C. N. and Baldi, P. (2014). Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**(18), 2592.

Niceta, M., Stellacci, E., Gripp, K., Zampino, G., Kousi, M., Anselmi, M., Traversa, A., Ciolfi, A., Stabley, D., Bruselles, A., Caputo, V., Cecchetti, S., Prudente, S., Fiorenza, M., Boitani, C., Philip, N., Niyazov, D., Leoni, C., Nakane, T., Keppler-Noreuil, K., Braddock, S., Gillessen-Kaesbach, G., Palleschi, A., Campeau, P., Lee, B., Pouponnot, C., Stella, L., Bocchinfuso, G., Katsanis, N., Sol-Church, K., and Tartaglia, M. (2015). Mutations impairing GSK3-mediated MAF phosphorylation cause cataract, deafness, intellectual disability, seizures, and a down syndrome-like facies. *Am. J. Hum. Genet.*, **96**(5), 816 – 825.

Oh, Y.-T., Chun, K., Park, B., Choi, J.-S., and Lee, S. (2007). Regulation of cyclin-dependent kinase inhibitor p21WAF1/CIP1 by protein kinase cδ-mediated phosphorylation. *Apoptosis*, **12**(7), 1339–1347.

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. (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.*, **3**(104), ra3.

Ortiz-Padilla, C., Gallego-Ortega, D., Browne, B. C., Hochgrafe, F., Caldon, C. E., Lyons, R. J., Croucher, D. R., Rickwood, D., Ormandy, C. J., Brummer, T., and Daly, R. J. (2013). Functional characterization of cancer-associated Gab1 mutations. *Oncogene*, **32**(21), 2696–2702.

Patrick, R., Lê Cao, K.-A., Kobe, B., and Bodén, M. (2015). PhosphoPICK: modelling cellular context to map kinase-substrate phosphorylation events. *Bioinformatics*, **31**(3), 382–389.

Patrick, R., Horin, C., Kobe, B., Cao, K.-A. L., and Bodén, M. (2016). Prediction of kinase-specific phosphorylation sites through an integrative model of protein context and sequence. *Biochim. Biophys. Acta, Proteins Proteomics*, **1864**(11), 1599 – 1608.

Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014). Duet: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res.*

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

Reimand, J. and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**(1).

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

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

Ristow, M., Müller-Wieland, D., Pfeiffer, A., Krone, W., and Kahn, C. R. (1998). Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *N. Engl. J. Med.*, **339**(14), 953–959.

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

Sagae, S., Kobayashi, K., Nishioka, Y., Sugimura, M., Ishioka, S., Nagata, M., Terasawa, K., Tokino, T., and Kudo, R. (1999). Mutational analysis of beta-catenin gene in japanese ovarian carcinomas: frequent mutations in endometrioid carcinomas. *Jpn. J. Cancer Res.*, **90**(5), 510–515.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.*, **29**(1), 308–311.

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**(5797), 268–274.

Toh, K. L., Jones, C. R., He, Y., Eide, E. J., Hinz, W. A., Virshup, D. M., Ptácek, L. J., and Fu, Y.-H. (2001). An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science*, **291**(5506), 1040–1043.

van Noort, M., van de Wetering, M., and Clevers, H. (2002). Identification of two novel regulated serines in the n terminus of β-catenin. *Exp. Cell Res.*, **276**(2), 264 – 272.

Wagih, O., Reimand, J., and Bader, G. D. (2015). MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Meth.*, **12**(6), 531–533.

Wu, L., Ma, C. A., Zhao, Y., and Jain, A. (2011). 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.

Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X. (2008). GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics*, **7**(9), 1598–1608.

Yi, J., Berrios, J., Newbern, J., Snider, W., Philpot, B., Hahn, K., and Zylka, M. (2015). An autism-linked mutation disables phosphorylation control of UBE3A. *Cell*, **162**(4), 795 – 807.

Zhu, G., Liu, Y., and Shaw, S. (2005). Protein kinase specificity: A strategic collaboration between kinase peptide specificity and substrate recruitment. *Cell Cycle*, **4**, 52 – 56.