BIOINFORMATICS

Vol. 24 ECCB 2008, pages i241–i247
doi:10.1093/bioinformatics/btn267

Gain and loss of phosphorylation sites in human cancer

Predrag Radivojac[1,†], Peter H. Baenziger[2,†], Maricel G. Kann[3], Matthew E. Mort[2], Matthew W. Hahn[1,4] and Sean D. Mooney[2,*]

[1]School of Informatics, Indiana University, 901 East Tenth Street, Bloomington, IN 47408, [2]Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, 410 West Tenth Street, Suite 5000, Indianapolis, IN 46202, [3]Department of Biological Sciences, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250 and [4]Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

ABSTRACT

Motivation: Coding-region mutations in human genes are responsible for a diverse spectrum of diseases and phenotypes. Among lesions that have been studied extensively, there are insights into several of the biochemical functions disrupted by disease-causing mutations. Currently, there are more than 60 000 coding-region mutations associated with inherited disease catalogued in the Human Gene Mutation Database (HGMD, August 2007) and more than 70 000 polymorphic amino acid substitutions recorded in dbSNP (dbSNP, build 127). Understanding the mechanism and contribution these variants make to a clinical phenotype is a formidable problem.

Results: In this study, we investigate the role of phosphorylation in somatic cancer mutations and inherited diseases. Somatic cancer mutation datasets were shown to have a significant enrichment for mutations that cause gain or loss of phosphorylation when compared to our control datasets (putatively neutral nsSNPs and random amino acid substitutions). Of the somatic cancer mutations, those in kinase genes represent the most enriched set of mutations that disrupt phosphorylation sites, suggesting phosphorylation target site mutation is an active cause of phosphorylation deregulation. Overall, this evidence suggests both gain and loss of a phosphorylation site in a target protein may be important features for predicting cancer-causing mutations and may represent a molecular cause of disease for a number of inherited and somatic mutations.

Contact: sdmooney@iupui.edu

1 INTRODUCTION

Mutations in cancers likely alter a great number of molecular events. One of these events is protein phosphorylation, which when altered may result in system-wide disruption and deregulation of signal transduction (Lim, 2005). Indeed, because of their involvement in cancer, kinases remain important drug targets in several classes of human cancers. Currently, kinase inhibitors Gleevec and Herceptin represent the most powerful therapies (Garber, 2006; Lim, 2005; Moasser, 2007).

There are many known spontaneous or somatic amino acid substitutions (Shimizu *et al.*, 2007) and it is likely that some of these will have profound effects on protein function (Kaminker *et al.*, 2007). In addition, two large-scale studies recently identified amino acid substitutions linked to cancer (Greenman *et al.*, 2007; Sjoblom *et al.*, 2006) and others have developed computational methods to predict certain types of cancer mutation sites (Kaminker *et al.*, 2007). Within these sets of mutations, we expect to observe loss-of-function mutations that turn-off normal molecular function. We also expect to observe mutations that cause a gain of function; that is, mutations that cause a molecular function to have deregulated activation when compared to normal function. Finally, we expect to observe many mutations that do not participate in neoplastic development and progression (Sjoblom *et al.*, 2006). These so-called passenger mutations, in fact, may comprise an overwhelming majority of cancer-associated mutations (Futreal *et al.*, 2005). As we begin to discover inherited and spontaneous amino acid substituting mutations in cancer, we have an opportunity to hypothesize their effects and classify as loss-of-function, gain-of-function and no-contributing-function (passengers).

Phosphorylation of amino acid residues serine (S), threonine (T) and tyrosine (Y) is common in cancer-associated proteins (Iakoucheva *et al.*, 2004) and known to be deregulated in cancer (Lim, 2005). It is well understood that changes in phosphorylation signaling can be due to deregulation of kinase and phosphatase function, usually detected through altered gene expression (Blume-Jensen and Hunter, 2001; Stephens *et al.*, 2005). This could include amino acid substitutions on kinases or phosphatases that directly interrupt the stability and/or the function of the kinase or phosphatase, resulting in changes in target phosphorylation. Effects of kinase or phosphatase regulators can also lead to altered phosphorylation. Here, we hypothesize that kinase and phosphatase substrates, too, could be subject to the effects of mutation (Fig. 1). A test of this hypothesis is performed to understand how phosphorylation sites in proteins are altered in cancer.

There is evidence in the literature that disruptions of phosphorylation sites are associated with cancer, for instance, mutations of T286 in cyclin D1 (CCND1). Phosphorylation of T286 by GSK3B in the wild type form of cyclin D1 initiates its nuclear export and subsequent degradation in the cytoplasm (Alt *et al.*, 2000; Diehl *et al.*, 1998), while the loss of phosphorylation is causatively implicated in nuclear accumulation of cyclin D1 in esophageal cancer and generally increased oncogenic potential (Benzeno *et al.*, 2006). Interestingly, while overexpression of cyclin D1 has been linked to a wide range of human cancers, it is becoming increasingly evident that overexpression of its wild type form is not sufficient for oncogenesis (Benzeno *et al.*, 2006), further emphasizing the importance of understanding the functional effects of protein mutations.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

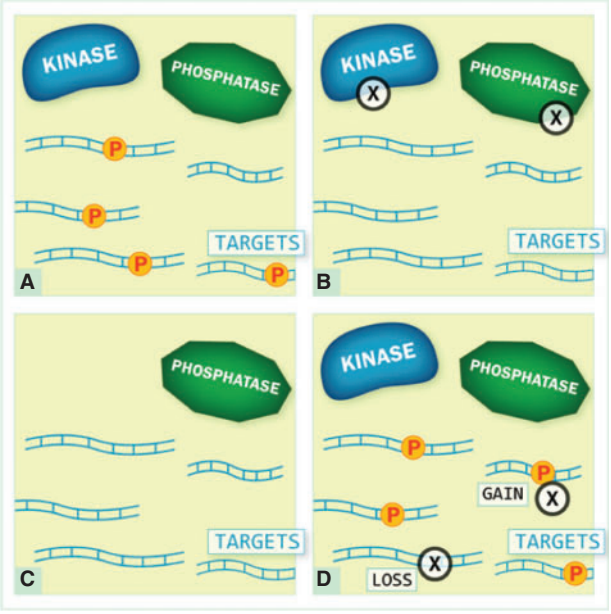

Fig. 1. Four scenarios for gain or loss of phosphorylation via a mutation mechanism. (**A**) The understood model of kinase and phosphatase adding and removing phosphate groups from targets. (**B**) The mutations on kinase or phosphates proteins as two possible mutation mechanisms for gain or loss of phosphorylation. (**C**) A situation when mutations occur on proteins activating or inhibiting kinases and phosphatases, thus effectively excluding these proteins from the signaling cascade. (**D**) The substrate mutations causing gain or loss of phosphorylation sites. Our study focuses on (**D**).

To test whether the hypothesis of target site disruption through spontaneous mutation is an important mechanism in cancer, we applied a high-throughput bioinformatics approach. Since no database of known phosphorylation disruption sites is available, development of computational techniques that directly predict such disruption is currently not possible. As an alternative, we use a less direct approach: the application of a phosphorylation site prediction algorithm to the mutated protein sequence before and after the mutation is present. These predictions are then analyzed for significant changes when compared to datasets of mutations unlikely to disrupt phosphorylation. To facilitate this analysis, three components are required: a method to predict phosphorylated residues from protein sequence, datasets of mutations unlikely to disrupt phosphorylation and a set of unbiased spontaneous mutations in cancer.

Assuming that our observations are not affected by the biases of the method or protein datasets, evidence of an enrichment of gain or loss of phosphorylation sites when comparing our disease sets to our control sets can be tabulated. We find that mutations from breast and colorectal cancers (Sjoblom *et al.*, 2006) are enriched in gain-of-phosphorylation sites and sequences from human kinases in cancer (Greenman *et al.*, 2007) contain significant enrichment of loss-of-phosphorylation sites. Undoubtedly, experimentally identifying whether the observed gain or loss sites are actually participating in some way in the cancer phenotype is a significant challenge. To begin to address this, we analyzed the pathways enriched in proteins with mutations predicted to disrupt phosphorylation sites. We identified Wnt/β-catenin as the most prominent pathway having proteins with mutations likely to disrupt phosphorylation.

2 METHODS

Datasets Mutation data included inherited disease amino acid substitutions from the Swiss-Prot database (Boeckmann *et al.*, 2003), two cancer tumor sequencing studies (Greenman *et al.*, 2007; Sjoblom *et al.*, 2006) and the SeattleSNPs resequencing project. The two cancer tumor sequencing projects and the Swiss-Prot dataset provide our 'disease' datasets. The breast and colorectal cancer dataset (herein referred to as the 'B&C Cancer set') originates from a sequencing project involving breast and colorectal cancer tumors (Sjoblom *et al.*, 2006). A subset of B&C Cancer limited to mutations in genes statistically overmutated is referred to as 'B&C Cancer—CAN'. The 'Kinase' dataset originates from a project that sequenced kinase genes from over 200 individual cancer tumors (Greenman *et al.*, 2007). For the analysis of Swiss-Prot, only disease-associated variants were included, while variants noted as 'in dbsnp' or having reference to collagen were excluded. We removed collagen variants due to overrepresentation in the set of inherited disease-associated mutations and the regularity of the collagen sequence that can bias the prediction process. To obtain population-relevant polymorphism data, we obtained all amino acid substitutions from the SeattleSNPs website. In total, 661 amino acid substitutions were analyzed in 189 genes and allele frequencies were determined from the SeattleSNPs project resequencing of African and European descent populations. For the purposes of this study, the wild type allele was defined as the most common allele for polymorphisms.

Additionally, several control sets were created. First, a model of 'neutral' mutations was constructed by collecting amino acid differences between human protein sequences and orthologous sequences from Ensembl's Compara database (Hubbard *et al.*, 2007), version 41. Neutral mutations were considered to be those where the orthologous sequence differed from both the wild type human protein and all considered mutants. They were limited to five closely related species: *Bos taurus*, *Canis familiaris*, *Macaca mulatta*, *Oryctolagus cuniculus* and *Rattus norvegicus*. Second, a random set of amino acid substitutions was constructed using Swiss-Prot data (version 45), where mutations were generated by randomly substituting the wild type amino acid by any of the remaining amino acids at a random position in wild type protein sequence. This set serves as a model of a computationally random comparison set. Lastly, two codon-based, random sets of mutations were created. These datasets were created by mutating the wild type sequence in our previously described cancer datasets (B&C Cancer and Kinase) and are therefore referred to as the 'B&C Cancer Control' set and the 'Kinase Control' set. The mutations were generated at the codon level and took into account measures of amino acid difference, codon frequency, transition-transversion ratio and gene variability (Goldman and Yang, 1994). Parameters for this mutating algorithm include 'μ' (dataset constant; value calculated on entire dataset to hold true to stipulations of Goldman and Yang), 'π' (frequency of codons; calculated for each dataset), 'k' (transition–transversion ratio; set to 1.45, as suggested by Goldman and Yang) and 'v' (gene variability; set to 43.99, as suggested by Goldman and Yang). Goldman and Yang provide a set of equations to calculate the probability that a given codon will mutate to the nine alternatives, limiting mutations to only one of the three nucleotides, as

$$Q_{ij} = \begin{cases} \mu\pi_j e^{-d_{ij}/v} & \text{(for transversions)} \\ \mu k\pi_j e^{-d_{ij}/v} & \text{(for transitions)} \end{cases}$$

where Q_{ij} is the probability that codon i will mutate to codon j and d_{ij} is the distance between amino acids i and j according to Li *et al.* (1985).

To summarize the dataset information, we list the number of mutations and proteins in each of the datasets: B&C Cancer (1099 mutations; 847 proteins), B&C Cancer—CAN (350; 171), B&C Cancer Control (7658; 847), Ensembl Orthologs (1181; 498), Kinase Cancer (695; 312), Kinase Cancer Control

(5442; 312), Swiss-Prot (12 614; 1142), Swiss-Prot Random (23 990; 2401) and SeattleSNPs (661; 189).

DisPhos application The DisPhos 1.3 predictor of S, T and Y phosphorylation sites is a discriminative method developed to classify unseen protein sites according to their probability to be phosphorylated (Iakoucheva *et al*., 2004). DisPhos utilizes position-specific amino acid compositions around phosphorylatable residues as well as several physicochemical and predicted properties correlated with phosphorylation (e.g. flexibility of the region). The predictor was trained using a set of experimentally determined phosphorylation sites obtained from Swiss-Prot and Phospho.ELM (Diella *et al*., 2004) databases. The balanced-sample accuracy of DisPhos is between 74% and 81% for the three phosphorylatable residues. DisPhos does not predict on any sequence that does not contain only the standard 20 amino acids. These sequences were excluded from the analysis.

DisPhos was trained to estimate the probability that a residue s_i can be phosphorylated, $P(s_i = s_i^P|s)$, given the protein sequence s and underlying distribution of training data $D(s)$. Assuming the new sequences are sampled from the same distribution $D(s)$, we can express the probability of a loss of phosphorylation at residue s_i as:

$$P(\text{loss of phosphorylation at } s_i|xjy) = P(s_i = s_i^P|s)\cdot(1 - P(s_i = s_i^P|s_{xjy}))$$

where s is the wild type protein sequence and s_{xjy} is the same sequence after undergoing a mutation of residue x into residue y at position j, where $x, y \in A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. Thus, the probability that position i will undergo a loss of phosphorylation depends on the probability that the site was phosphorylated in the wild type protein sequence s multiplied by the probability that the site is not phosphorylated in the sequence mutated at position j from residue x to residue y. Clearly, the greater the difference $|i - j|$, the less one expects mutation xjy to influence residue s_i. Therefore, if there is a direct mutation of the phosphorylation site s_i, (i.e. if $i = j$), the expected influence on phosphorylation will be maximal. In fact, the worst-case scenario is expected to occur when $y \notin \{S, T, Y\}$, in which case the probability of a loss of phosphorylation equals $P(s_i = s_i^P|s)$, because $P(s_i = s_i^P|s_{xjy}) = 0$.

In order to further explain this approach, consider the following example: if $P(s_i = s_i^P|s) = 0.9$ and $P(s_i = s_i^P|s_{xjy}) = 0.1$, then the probability of a loss of phosphorylation will be $0.9\cdot(1-0.1) = 0.81$. If the two probabilities are 0.9 and 0.8 or 0.2 and 0.1, the overall probability of the loss of phosphorylation will be 0.18 in both cases, although for a different reason. In the former case, the difference of 0.1 will indicate that the site is probably phosphorylated in both situations, while in the latter case the site is most likely not phosphorylated in both situations. On the other hand, if the two probabilities are 0.55 and 0.45 (also a difference in probabilities of 0.1), the probability of the loss of phosphorylation will be about 0.3.

Equivalent to the loss of phosphorylation, we can express the probability of a gain of phosphorylation at residue s_i as:

$$P(\text{gain of phosphorylation at } s_i|xjy) = (1 - P(s_i = s_i^P|s))\cdot P(s_i = s_i^P|s_{xjy})).$$

Here, the probability that position i will undergo a gain of phosphorylation is a product of the probability that the site s_i was not phosphorylated in the wild type sequence s and the probability that it is phosphorylated in the sequence mutated at position j after a mutation from residue x to residue y. We emphasize that a one-residue difference in target sequences can correspond to both small and large distances in the feature space due to the details of the construction of DisPhos classifier.

Note that, the definitions of phosphorylation loss and gain are also based on the assumption that DisPhos scores are independent of the kinase involved in the reaction. Given the size of DisPhos training data, we believe this is not a serious limitation to the validity of our results. However, it is unlikely that the DisPhos training distribution $D(s)$ and data from this study are identical, especially with respect to the prior probabilities of phosphorylated sites. Thus, to overcome a possibility of biased inference, most of this work addresses so-called 'high-confidence' gain or loss of phosphorylation sites, which at residue s_i we define as those where:

$$P(\text{loss/gain of phosphorylation at } s_i|xjy) > 0.75.$$

This definition is practically useful because it dictates that the DisPhos predictions have a 5–10% false positive rate (depending on the residue type) and that 10–20% of all predicted positives are false identifications (Iakoucheva *et al*., 2004). In reality, we believe that a lower percentage of all 'false positive' identifications are indeed false, because the set of 'negative' phosphorylation sites used to estimate false positive rate is believed to contain a number of actual phosphorylation sites that are yet unidentified or not stored in major databases.

Ingenuity pathway analysis Ingenuity pathway analysis (IPA) (Ingenuity® Systems, Redwood City, CA, USA, www.ingenuity.com) is a function and pathway exploration tool that uses text mining, an expertly curated knowledgebase, and association statistics to give the user the molecules, relationships, functions and pathways represented in a set of genes or proteins. Using IPA, we ran several 'core' analyses on protein sets from the breast and colorectal cancer study. We compared four subsets of the B&C Cancer dataset: all proteins, proteins with mutations yielding high-confidence loss or gain of phosphorylation, proteins with high-confidence loss of phosphorylation and proteins with high-confidence gain of phosphorylation sites.

3 RESULTS

3.1 Prevalence of phosphorylation site predictions in training set sequences

To verify the datasets were similar in the number of mutations involving a phosphorylatable amino acid mutating to a non-phosphorylatable amino acid or vice versa, we compared percentages of each dataset representing these types of mutations. Figure 2 shows that the datasets are closely gathered around an average of 12.2% of the data involving mutations from S, T or Y to another amino acid (Fig. 2; dark grey bars), and around 13.5% for the mutations to a phosphorylatable amino acid (Fig. 2; light grey bars). Interestingly, both datasets created with the biologically intelligent mutating algorithm have fewer mutations to phosphorylatable amino acids (Fig. 2; light grey bars). This likely occurs because of the consideration the algorithm gives to evolutionary preferred amino acid substitutions (mostly neutral).

Prediction was completed on 53 690 mutations from 5390 unique protein sequences. All insertions and deletions, all substitutions

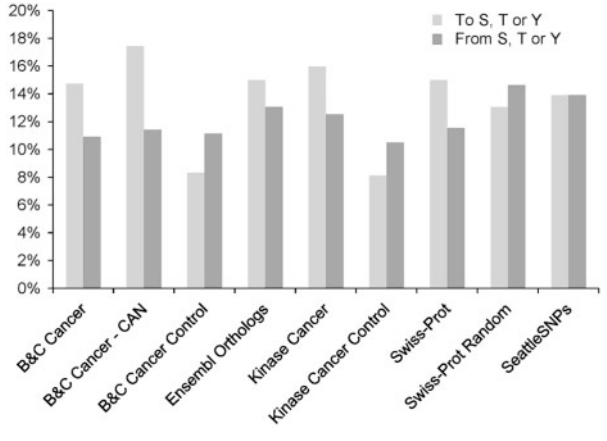

Fig. 2. Relative frequencies of mutations involving phosphorylatable residues in each dataset. Dark grey bars: the percent of mutations from a phosphorylatable amino acid (S, T or Y) to a non-phosphorylatable amino acid. Light grey bars: the percent of mutations from a non-phosphorylatable amino acid to a phosphorylatable amino acid (S, T or Y).

containing discrepancies between the sequence and mutations, and all protein sequences containing symbols not in *A* were excluded from the analysis, therefore resulting in smaller mutation datasets than those originally published. From the starting set of 53 690 mutations, 14 844 had changes of phosphorylation prediction (>0.001) on the mutation site. However, only 1033 of these mutations were considered high-confidence loss/gain-of phosphorylation sites. The original set of 53 690 mutations also caused phosphorylation score changes (>0.001) on 612 435 non-mutation sites, yet none of the non-mutation sites were of high confidence. Only 20 such mutations yielded a loss-/gain-of-phosphorylation score >0.5. This is expected from two standpoints: (i) functionally, a mutation around a phosphorylation site is more likely to change the efficiency of enzymatic reaction instead of fully disrupting it and (ii) statistically, a single change away from the modification site is less likely to be critically important due to an implicit assumption of statistical inference methods that similar sequence patterns should generally result in similar prediction scores, especially for robust inference models such as DisPhos.

3.2 Analysis of phosphorylation sites in germline polymorphic positions in two populations

In order to understand phosphorylation in the context of mutation frequency, we studied mutations from the SeattleSNPs resequencing project. We partitioned these amino acid substitutions into allele frequency ranges (low: frequency <0.05; medium: $0.25 <$ frequency < 0.75; and high: frequency > 0.95) based on data from the SeattleSNPs project and averaged the allele DisPhos scores. As Figure 3 shows, mutations with low allele frequency have a positive correlation with the propensity for phosphorylation. This result seems to support the recent finding that rare alleles tend to be mildly deleterious (Kryukov *et al.*, 2007).

3.3 Predicted gain and loss of phosphorylation sites in cancer compared to other datasets

In Figure 4A, we illustrate that the three cancer mutation sets show a greater rate of gain of phosphorylation than the non-disease-associated sets. For the highly confident predictions, the Kinase and B&C Cancer sets show gain of phosphorylation occurring on 1.87% ($n=13$) and 1.91% ($n=21$) of the mutations, respectively. These percentages are significantly larger than those in the control sets [Swiss-Prot Random: 0.81% ($n=194$), B&C Cancer Control: 0.86% ($n=66$), Kinase Control: 0.88% ($n=48$) and Swiss-Prot: 0.89% ($n=112$)]. The originating authors of the B&C Cancer set used more stringent statistical filters to isolate genes showing significantly higher mutation rates than expected. Limiting the B&C Cancer set to mutations from the overly mutated genes shows a gain of phosphorylation predicted on 2.29% ($n=8$) of the dataset (B&C Cancer—CAN). Finally, while the inclusion of data from orthologous sequences seems to have provided a good neutral model, we note that there are multiple examples where disease-associated mutations in humans correspond to the wild type sequences even in closely related species (Ostedgaard *et al.*, 2007; Rhesus Macaque Genome Sequencing and Analysis Consortium, 2007).

Figure 4B shows the rates of high-confidence loss of phosphorylation in each dataset. Again, cancer datasets show the greatest loss of phosphorylation and the Kinase dataset is particularly interesting because of its large percentage of loss of phosphorylation

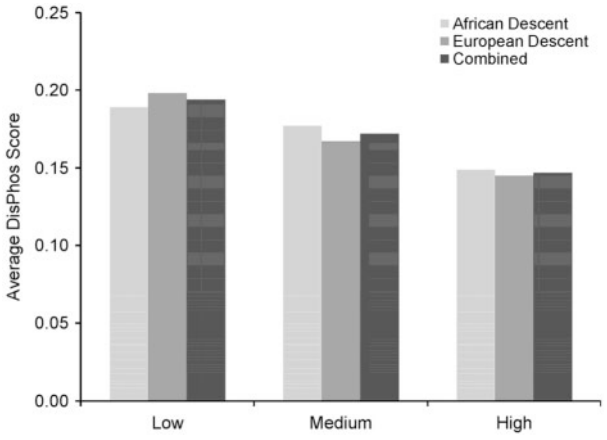

Fig. 3. Allele frequencies of nsSNPs and average phosphorylation scores. Human population sequences were partitioned into low (frequency <0.05), medium (0.25< frequency <0.75), and high (frequency >0.95) allele frequency groups. The phosphorylation scores were averaged for each group. Allele with lowest frequency show the highest average confidence of being phosphorylated. The difference between high and low allele frequency groups for the population of European descent is statistically significant ($P=0.04$; *t*-test).

sites: 3.17% ($n=22$). As mentioned earlier and shown in Figure 2, these results are not biased by the number of mutations to or from phosphorylatable amino acids.

3.4 Motif and pathway analysis of proteins affected by change of phosphorylation

Disease sets showing higher gain and/or loss of phosphorylation were analyzed to identify possible overrepresentation in known phosphorylation site motifs. We used Scansite (Obenauer *et al.*, 2003) to identify the frequency of motifs present in each dataset and compared highly disrupted motifs in the cancer datasets to the motifs disrupted in the cancer control sets. Several motifs showed large differences: Akt Kinase, Clk2 Kinase, 14-3-3 Mode 1, Nck 2, Cortactin SH3 and Grb2 SH3, Crk SH3, Amphiphysin SH3, Erk1 Kinase, DNA-activated Protein Kinase, Casein Kinase 1 and 2 and ATM Kinase. However, due to the further fragmentation of disease datasets, we were able to identify only two motifs that are disrupted in the cancer data with statistical significance (Bonferroni-corrected *t*-test): Cortactin SH3 and Grb2 SH3.

For pathway analysis, we created several subsets of proteins from the breast and colorectal cancer mutation set. By comparing the percent of each of these sets classified into specific pathways, we observe that gain of phosphorylation-associated mutations are enriched for several pathways (Fig. 5). Specifically, we see that these proteins are associated with signaling pathways such as Wnt/β-catenin, which has been shown to be associated with cancer and tumorigenesis (Mori *et al.*, 1992; Spink *et al.*, 2000).

3.5 Evidence for our predictions in the literature

A review of the literature provides evidence in support of our predictions. For example, mutation of T286, a known phosphorylation site, in cyclin D1, a known cancer-associated

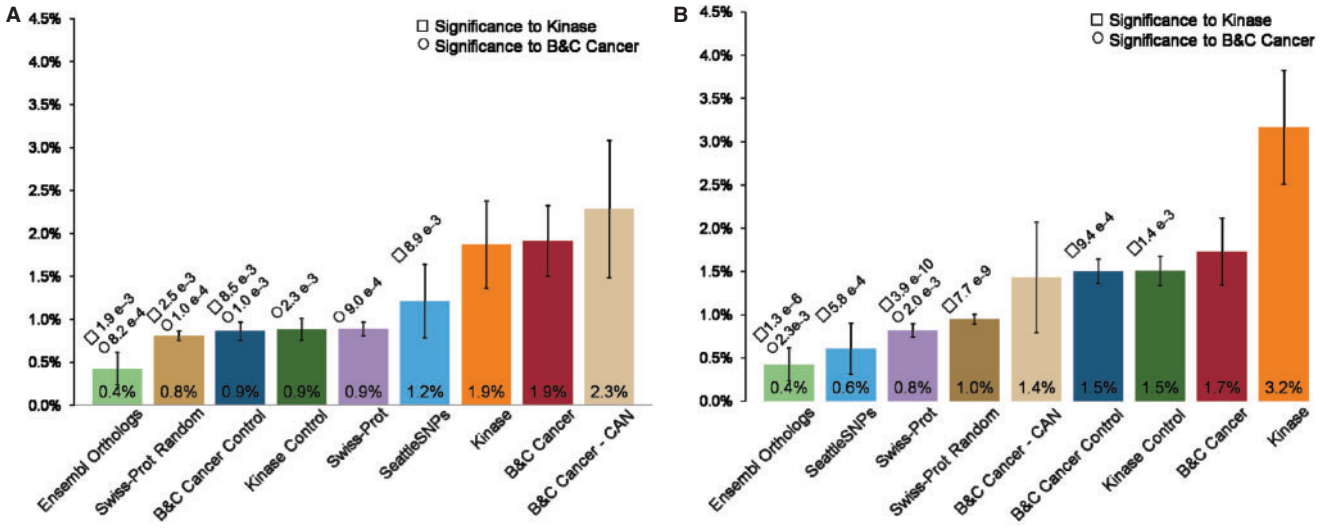

Fig. 4. Mutations that cause loss and gain of phosphorylation target sites are enriched in cancer. Relative frequencies of high-confidence gain-of-phosphorylation sites (**A**) and high-confidence loss-of-phosphorylation sites (**B**) in various datasets. *P*-values were calculated using a *t*-test and statistical significance determined using a Bonferroni-corrected threshold of 0.05.

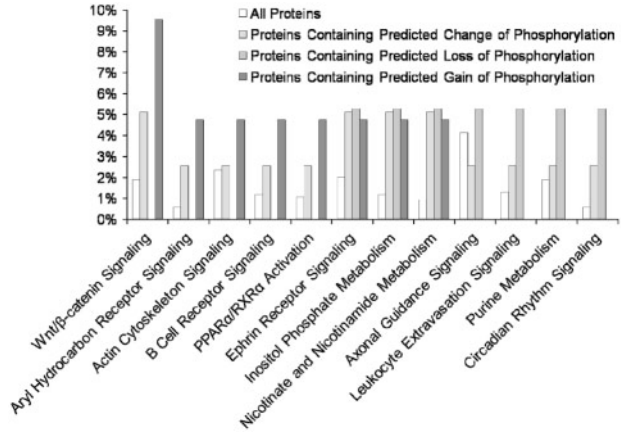

Fig. 5. IPA of proteins hosting mutations altering phosphorylation sites. After running IPA, the percent of proteins from each dataset classified into the pathways above was graphed for each subset of the B&C Cancer dataset.

protein (Benzeno *et al*., 2006), results in a loss-of-phosphorylation score >0.5, despite the absence of T286 from the DisPhos training data. Also, in catenin β-1, which is involved in hepatocarcinoma, phosphorylation is known to occur on amino acids T41 and S45 (Hagen and Vidal-Puig, 2002) and result in significant scores for loss-of-phosphorylation (>0.5 and >0.6, respectively).

Our results also complement those found by studying individual phosphorylation site mutations. In a review of EGFR and ErbB4's roles in tumorigenesis and signal transduction, several examples of phosphorylation site mutations are given which through creation of new docking sites lead to changes in regulation (Schulze *et al*., 2005; Yarden and Sliwkowski, 2001). Some interesting sites show gain and loss of phosphorylation, but have not been experimentally validated. In AXIN-1 (Spink *et al*., 2000) and BARD-1 (GenBank gi|20532378 and gi|116241265), two breast cancer-associated proteins, several specific mutation predictions suggest gain of phosphorylation sites. Mutant G650S yields a high-confidence gain-of-phosphorylation score (0.95). Similarly, mutant P24S causes the prediction to increase from 0 to 0.92.

Generalizing these specific examples, we observe an increased rate of gain-of-phosphorylation in entire disease-associated mutation sets (Fig. 4). This broad view of phosphorylation suggests that mutations introducing new phosphorylation sites are an important mechanism of disease. Results from analysis of prediction values on the SeattleSNPs dataset, in which larger number of high-confidence phosphorylation sites correspond to lower allele frequencies, support this hypothesis.

4 DISCUSSION

4.1 Phosphorylation target site mutation is likely a mechanism in cancer

Based on the results of this study, we believe both gain and loss of phosphorylation are an important mechanism causing deregulation of phosphorylation mediated signal transduction. We find that kinases in cancer are twice as likely to have mutations disrupting phosphorylation sites as compared to a kinase control set, but the increase compared to the set of Swiss-Prot mutations, human variation data (SeattleSNPs) and orthologous sequences is 4-, 5- and 8-fold, respectively. Similarly, gain of phosphorylation sites in cancer-associated mutations is about 2-fold as compared to Swiss-Prot and human variation data. This is an intriguing finding because it suggests that a mechanism of signal transduction deregulation in cancer is mediated by either removal or creation of phosphorylation sites thereby causing either a loss or a gain of phosphorylation function, depending on the role of the phosphorylated residue. This is further supported by the fact that an increase in number

of high-confidence phosphorylation sites is correlated with low allele frequencies in two human populations, suggesting purifying evolutionary selection against these mutations.

Perhaps most interesting of all is that cancer-associated mutations are predicted to be enriched in gain of phosphorylation sites. It can be speculated that mutations which create phosphorylation sites could destabilize proteins, interrupt protein interactions, or disrupt enzyme catalysis or other normal protein functions. They may also recruit kinases or phosphatases necessary for other cellular processes causing system-wide deregulation. When we examine the signaling pathways in which we observe phosphorylation disrupting mutations, we find that several pathways are enriched in mutations. Notably, the Wnt/β-Catenin pathway is enriched in both gain and loss of phosphorylation sites. This pathway is well known in cancer and several reviews convey its importance (Bienz and Clevers, 2000; Spink *et al*., 2000).

It is tempting to speculate on the percentages of known mutations likely causing gain or loss of phosphorylation sites. Using the high-confidence scores only, we predict that about 2% of disease-associated mutations in Swiss-Prot are caused by changes in phosphorylation patterns. For a less stringent threshold of 0.5 for the prediction of phosphorylation sites, our prediction rises to about 7%. However, these estimates should be interpreted with caution given by the recent analysis of Care *et al*. (2007).

The findings of our study that gain and loss of phosphorylation sites are linked to human cancer add to the increasing breadth of literature addressing post-translational modifications and disease, e.g. the role of glycosylation in inherited mutations (Vogt *et al*., 2005, 2007).

4.2 Machine learning methods that predict functional sites can be used to predict mutation disruption

One of the conclusions of this study is that residue functional prediction methods can give insight into the molecular mechanism of specific disease-causing mutations. Currently, prediction methods focus on utilizing sequence, protein structure and evolutionary features to predict whether an amino acid substitution is likely to affect protein function or be involved in disease (Bromberg and Rost, 2007; Ng and Henikoff, 2003; Ramensky *et al*., 2002; Yue *et al*., 2006). While these features are very informative for prediction, they do not provide insight into the underlying molecular cause of the disease. For example, knowledge that a specific conserved residue feature is used to make a prediction of a disruptive mutation does not provide direct insight into how the specific mutation disrupts function. Our results using DisPhos suggest that including features from function prediction methods may give direct evidence of the molecular causes. Here, we have investigated mutation sites highly likely to affect phosphorylation. Hypothesizing that the mutation of a specific residue disrupts phosphorylation is, in our opinion, more informative than knowledge that a conserved residue or a structural feature is disrupted. We believe the additional information about the functional effects of the mutation is an important factor to guide new experiments and will lead to a better understanding of the etiology.

ACKNOWLEDGEMENTS

Funding: This research is supported by NSF award DBI-0644017 (PI: Radivojac), NIH grants K22LM009135 (PI: Mooney) and R01LM009722 (PI: Mooney), and a grant from the IU Biomedical Research Council, Indiana University, the Showalter Trust and the Indiana Genomics Initiative (INGEN). INGEN is supported in part by the Lilly Endowment.

Conflict of Interest: none declared.

REFERENCES

Alt,J.R. *et al*. (2000) Phosphorylation-dependent regulation of cyclin D1 nuclear export and cyclin D1-dependent cellular transformation. *Genes Dev*., **14**, 3102–3114.

Benzeno,S. *et al*. (2006) Identification of mutations that disrupt phosphorylation-dependent nuclear export of cyclin D1. *Oncogene*, **25**, 6291–6303.

Bienz,M. and Clevers,H. (2000) Linking colorectal cancer to Wnt signaling. *Cell*, **103**, 311–320.

Blume-Jensen,P. and Hunter,T. (2001) Oncogenic kinase signalling. *Nature*, **411**, 355–365.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*., **35**, 3823–3835.

Care,M.A. *et al*. (2007) Deleterious SNP prediction: be mindful of your training data! *Bioinformatics*, **23**, 664–672.

Diehl,J.A. *et al*. (1998) Glycogen synthase kinase-3beta regulates cyclin D1 proteolysis and subcellular localization. *Genes Dev*., **12**, 3499-3511.

Diella,F. *et al*. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.

Futreal,P.A. *et al*. (2005) Somatic mutations in human cancer: insights from resequencing the protein kinase gene family. *Cold Spring Harb Symp Quant Biol.*, **70**, 43–49.

Garber,K. (2006) The second wave in kinase cancer drugs. *Nat Biotech*., **24**, 127–130.

Goldman,N. and Yang,Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol*., **11**, 725–736.

Greenman,C. *et al*. (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.

Hagen,T. and Vidal-Puig,A. (2002) Characterisation of the phosphorylation of [beta]-catenin at the GSK-3 priming site Ser45. *Biochem. Biophys. Res. Commun*., **294**, 324–328.

Hubbard,T.J. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Iakoucheva,L.M., *et al*. (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Res.*, **32**, 1037–1049.

Kaminker,J.S. *et al*. (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.*, **35**, W595–W598.

Kaminker,J.S. *et al*. (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.*, **67**, 465–473.

Kryukov,G.V. *et al*. (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet*., **80**, 727–739.

Li,W.H. *et al*. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.

Lim,Y.P. (2005) Mining the tumor phosphoproteome for cancer markers. *Clin. Cancer Res.*, **11**, 3163–3169.

Moasser,M.M. (2007) Targeting the function of the HER2 oncogene in human cancer therapeutics. *Oncogene*, **26**, 6577–6592.

Mori,Y. *et al*. (1992) Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Hum. Mol. Genet.*, **1**, 229–233.

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

Obenauer,J.C. *et al*. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

Ostedgaard,L.S. *et al*. (2007) Processing and function of CFTR-{Delta}F508 are species-dependent. *Proc. Natl Acad. Sci*. USA, **104**, 15370–15375.

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.

Rhesus Macaque Genome Sequencing and Analysis Consortium *et al*. (2007) Evolutionary and biomedical insights from the Rhesus Macaque genome. *Science*, **316**, 222–234.

Schulze,W.X. *et al*. (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol. Syst. Biol.*, **1**, 2005 0008, 1–13.

Shimizu,N. *et al*. (2007) MutationView/KMcancerDB: a database for cancer gene mutations. *Cancer Sci*., **98**, 259–267.

Sjoblom,T. *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.

Spink,K.E. *et al*. (2000) Structural basis of the Axin-adenomatous polyposis coli interaction. *EMBO J*, **19**, 2270–2279.

Stephens,B.J. *et al*. (2005) PRL phosphatases as potential molecular targets in cancer. *Mol. Cancer Ther.*, **4**, 1653–1661.

Vogt,G. *et al*. (2005) Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. *Nat. Genet*., **37**, 692–700.

Vogt,G. *et al*. (2007) Gain-of-glycosylation mutations. *Curr. Opin. Genet. Dev.*, **17**, 245–251.

Yarden,Y. and Sliwkowski,M.X. (2001) Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell Biol.*, **2**, 127–137.

Yue,P. *et al*. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.