Bioinformatics, 33(12), 2017, 1845–1851
doi: 10.1093/bioinformatics/btx082
Advance Access Publication Date: 13 February 2017
Original Paper

OXFORD

Systems biology

Benchmarking substrate-based kinase activity inference using phosphoproteomic data

Claudia Hernandez-Armenta[1,†], David Ochoa[1,*,†], Emanuel Gonçalves[1], Julio Saez-Rodriguez[1,2] and Pedro Beltrao[1,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK and [2]RWTH Aachen University, Faculty of Medicine, Joint Research Center for Computational Biomedicine (JRC-COMBINE), Wendlingweg 2 D-52074, Aachen, Germany

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

Received on October 17, 2016; revised on January 19, 2017; editorial decision on February 6, 2017; accepted on February 9, 2017

Abstract

Motivation: Phosphoproteomic experiments are increasingly used to study the changes in signaling occurring across different conditions. It has been proposed that changes in phosphorylation of kinase target sites can be used to infer when a kinase activity is under regulation. However, these approaches have not yet been benchmarked due to a lack of appropriate benchmarking strategies.
Results: We used curated phosphoproteomic experiments and a gold standard dataset containing a total of 184 kinase-condition pairs where regulation is expected to occur to benchmark and compare different kinase activity inference strategies: Z-test, Kolmogorov Smirnov test, Wilcoxon rank sum test, gene set enrichment analysis (GSEA), and a multiple linear regression model. We also tested weighted variants of the Z-test and GSEA that include information on kinase sequence specificity as proxy for affinity. Finally, we tested how the number of known substrates and the type of evidence (*in vivo*, *in vitro* or *in silico*) supporting these influence the predictions.
Conclusions: Most models performed well with the Z-test and the GSEA performing best as determined by the area under the ROC curve (Mean AUC = 0.722). Weighting kinase targets by the kinase target sequence preference improves the results marginally. However, the number of known substrates and the evidence supporting the interactions has a strong effect on the predictions.
Availability and Implementation: The KSEA implementation is available in https://github.com/evocellnet/ksea. Additional data is available in http://phosfate.com
Contact: pbeltrao@ebi.ac.uk or ochoa@ebi.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The functional plasticity of the proteome is modulated by specific post-translational modifications (PTMs) such as phosphorylation, glycosylation or ubiquitination. Reversible modification of proteins controls signal transduction and information processing within the cell, mediating molecular processes such as protein enzymatic regulation, complex associations, protein degradation and changes in subcellular localization (Choudhary and Mann, 2010). Through the coordinated regulation of multiple parallel events, the cell is able to integrate all available signals to appropriately respond to environmental stimuli (Francavilla *et al.*, 2016). Reversible phosphorylation of individual residues remains one the most studied PTMs and its status results from the net activity of kinases and phosphatases. Kinases bind covalently a phosphate group onto specific amino-acids, most frequently serine, threonine or tyrosine, while phosphatases catalyze the reverse reaction.

Previous studies of cell decision-making mediated by protein phosphorylation have been limited by the reduced number of phosphosites and kinase activities that were possible to measure, most frequently using phospho-specific antibodies (Miller-Jensen *et al.*, 2007). However, recent developments in large-scale phosphoproteomics and mass spectrometry (MS) have fostered the system-wide identification and quantification of phosphorylated sites before and after controlled perturbation (Olsen *et al.*, 2006). These advances have not only allowed for the measurement of the response of thousands of individual phosphosites in a very rapid and granular timeframe (Humphrey *et al.*, 2015; Kanshin *et al.*, 2015), but also to estimate kinase regulation in large scale based on the changes in known kinase substrates (Casado *et al.*, 2013; Drake *et al.*, 2012). The emergence of methods to estimate the activities directly from high-throughput MS data and prior knowledge in kinase substrate interactions introduces a new perspective on understanding the signaling response. These have been used to study differences in tumors (Casado *et al.*, 2013; Drake *et al.*, 2012), to study the effects of drugs, to reconstruct signaling networks and to broadly survey the kinase signaling states of the cell (Ochoa *et al.*, 2016).

Substrate-based kinase activity prediction methods are founded on the assumption that the regulatory state of a kinase is reflected on the change of phosphorylation levels of its substrate sites. Drake *et al.* (2012) first reported this notion and performed an enrichment analysis of tyrosine kinase activities. The method used consisted of a permutation analysis to infer an enrichment score using the Kolmogorov Smirnov statistic, an algorithm similar to the popular Gene Set Enrichment Analysis (GSEA) (Drake *et al.*, 2012; Subramanian *et al.*, 2005). This approach has been applied to a large number of kinases and conditions in a recent study (Ochoa *et al.*, 2016). Alternatively, Casado *et al.* (2013) proposed an algorithm based on a one sample Z-test to measure for the first time the regulation of large number of kinases in label-free cancer samples. More recently, the IKAP machine learning method models the abundance of each phosphosite as the sum of all kinase effects acting on it (Mischnik *et al.*, 2016).

These methods depend on the aggregation of kinase substrate annotations. Comprehensive resources have aggregated experimentally verified *in vivo* and *in vitro* interactions between kinases and phosphorylation sites, including PhosphoSitePlus (Hornbeck *et al.*, 2015), Phospho.ELM (Dinkel *et al.*, 2011), HPRD (Peri *et al.*, 2004) and Signor (Perfetto *et al.*, 2016). This information is still limited to a fraction of all sites and kinases in very few model organisms. Computational approaches such as NetworKIN attempt to complete our knowledge on kinase targets by predicting new substrates from motif analysis and functional context derived from STRING (Kersten *et al.*, 2009; Mering, 2003).

Although the previously described methods have been benchmarked to different degrees by experimental and computational approaches (Casado *et al.*, 2013; Drake *et al.*, 2012; Ochoa *et al.*, 2016) they have not yet been systematically compared. The continuous development of new strategies requires a standardized framework to assess the substrate-based prediction of kinase activities. Therefore, we selected from our previous compilation effort (Ochoa *et al.*, 2016) a subset of 24 conditional phosphoproteomic studies, in which 30 different kinases are expected to display regulation. This gold standard was used to evaluate 5 different methods: Z-test, Kolmogorov-Smirnov test, Wilcoxon rank-sum test, GSEA algorithm and a multiple linear regression model (MLR). We also evaluated the effect of the kinase sequence specificity by weighting differently the target sites showing different sequence similarities to the binding motif. Moreover, we benchmarked the impact of the number of known target sites, as well as the effect of the evidence supporting the substrate interactions on the performance of the methods.

This gold standard and analysis provides a comprehensive comparison of the strengths and limitations of different methodologies and provides the necessary framework to evaluate future developments.

2 Methods

2.1 A benchmark dataset for kinase regulation predictions

To evaluate the performance of different kinase activity prediction methods, we obtained from a previous study (Ochoa *et al.*, 2016) a subset of 24 publications describing quantitative phosphoproteomic data reporting the response of a variable number of human phosphosites after perturbations linked to specific kinase activations or inhibitions. For example, the EGFR tyrosine kinase-receptor is activated when its ligand, the epidermal growth factor (EGF), binds the receptor (Oda *et al.*, 2005). It is therefore expected that experiments assaying the phosphoproteomic response after EGF stimulation present increased activities of EGFR. Similarly, the molecular response to DNA damage has been associated with up-regulation of the ATR and ATM kinases (Matsuoka *et al.*, 2007). Alternatively, drug treatments such as the addition of the kinase inhibitors Erlotinib or Torin 1 are expected to specifically inhibit their corresponding targets EGFR and mTOR. This gold standard positive set contains 184 pairs between 30 different kinases regulated in 62 different perturbations with publicly available phosphoproteomic data (Table 1). A more detailed description of the conditions in the gold standard can be found in Supplementary Table S1.

2.2 Human quantitative phosphoproteomic data preprocessing and normalization

Despite the heterogeneity of the biological samples, as well as the technical variation introduced by different experimental setups, high reproducibility is expected in the quantitative response of the direct kinase regulations included in the gold standard. As was done in our previous study (Ochoa *et al.*, 2016), we applied a set of filters and quality control steps: (i) we restricted the analysis to peptides mapped to the Ensembl canonical transcripts ignoring other modified isoforms; (ii) the log2-fold changes of phosphopeptides containing the same phosphosite position were averaged, even if differing in exact sequence [i.e. RLS(ph)PTK and LS(ph)PTK in the same protein]; (iii) the log2-fold changes of the same phosphopeptides in different replicates were averaged; (iv) to increase the interpretation of the fold changes as single PTMs, only monophosphorylated phosphopeptides were considered. Finally, after applying all filters, we excluded perturbations in the gold standard containing <1000 quantified phosphopeptides. Quantifications across conditions were quantile normalized to maximize reproducibility across samples. The quantification data for each phosphosite for the conditions in the gold standard can be obtained from our website (http://phosfate.com).

2.3 Substrate-based kinase activity inference methods

For each perturbation in the gold standard, we calculated the changes in kinase activities based on the quantitative phosphoproteomic profile and the set of known kinase substrates. A total of 2818 manually curated interactions for the 30 kinases under study were compiled from PhosphoSitePlus (Hornbeck *et al.*, 2015),

Table 1 Human kinases displaying specific regulation—up or down—in perturbations quantified using high-throughput phosphoproteomics

Experimental condition	Kinases	Regulation	References
AKTi (Akt inhibitor VIII), AKTi (MK-2206),PI3Ki (GDC-0941), PI3Ki (PI-103),PIK3CA activation (H1047R) + inhibitor	AKT1	down	Wilkes *et al.* (2015), Wu *et al.* (2014)
PIK3CA activation (E545K),PIK3CA activation (H1047R), RG7356	AKT1	up	Weigand *et al.* (2012) Wu *et al.* (2014)
Anti-CD3, anti-CD3 + anti-CD28, Differentiation (PMA)	PRKCA	up	Nguyen *et al.* (2016), Rigbolt *et al.* (2011) Salek *et al.* (2013)
PKCi (BIM-1),PKCi (Gö-6976)	PRKCA	down	Wilkes *et al.* (2015)
ATR inhibitor (VE-821)	ATR	down	Šalovská *et al.* (2014)
CAMK2i (KN-62),CAMK2i (KN-93)	CAMK2A	down	Wilkes *et al.* (2015)
Crizotinib	ALK	down	Oppermann *et al.* (2013)
Dasatinib (50 nM)	MAPK3,ABL,LCK	down	Pan *et al.* (2009)
DNA damage (Etoposide), DNA damage (Ionizing radiation), Early S (Thymidine), G1_S (Thymidine), Late S (Thymidine)	ATR, ATM	up	Beli *et al.* (2012), Dephoure *et al.* (2008), Olsen *et al.* (2010)
EGF	AKT1, BRAF, EGFR, MAP2K1, MAP2K2, MAPK1, MAPK3, RAF1	up	Engholm-Keller *et al.* (2011), Mertins *et al.* (2012), Pan *et al.* (2009)
EGFRi (PD-153035),EGFRi (PD-168393), Erlotinib,Gefitinib	EGFR	down	Weber *et al.* (2012), Wilkes *et al.* (2015)
EGF + SB202190	MAPK14	down	Pan *et al.* (2009)
EGF + U0126, MEKi (GSK-1120212),MEKi (U0126),Selumetinib (AZD6244)	MAP2K1, MAP2K2	down	Pan *et al.* (2009), Stuart *et al.* (2015), Wilkes *et al.* (2015)
ERKi (ERK inhibitor),ERKi (ERK inhibitor II),Dasatinib (50 nM)	MAPK1	down	Pan *et al.* (2009), Wilkes *et al.* (2015)
Iloprost (5 nM)	PRKACA	up	Beck *et al.* (2014)
Mitosis (AZD1152), Mitosis (MLN8054)	AURKA	down	Kettenbach *et al.* (2011)
Mitosis (AZD1152)	AURKB	down	Kettenbach *et al.* (2011)
Mitosis (BI2536),Mitosis (short Taxol + BI2536),BI 4834 (on Mitosis Nocodazole), PLK1 Inhibitor (G2r - BI2536)	PLK1	down	Grosstessner-Hain *et al.* (2011), Halim *et al.* (2013), Kettenbach *et al.* (2011)
mTORi (KU-0063794), mTORi (Torin-1),Rapamycin,Starved,Torin1	MTOR	down	Hsu *et al.* (2011), Wilkes *et al.* (2015)
PIK3CA activation (H1047R), PIK3CA activation (E545K)	PIK3CA	up	Wu *et al.* (2014)
P70S6K1i (DG2), P70S6K1i (PF-4708671)	RPS6KB1	down	Wilkes *et al.* (2015)
ROCKi (H-1152), ROCKi (Y-27632)	ROCK1	down	Wilkes *et al.* (2015)
Sorafinib	PDGFRB, PDGFRA, BRAF, RAF1, FLT1, FLT3, FLT4	down	Zhuang *et al.* (2013)
Quizartinib	PDGFRA, FLT3	down	Schaab *et al.* (2014)
LRRK2-IN-1 (Endogenous LRRK2),LRRK2-IN-1 (LRRK2 overexpression)	LRKK2	down	Luerman *et al.* (2014) Schaab *et al.* (2014)
VEGF	FLT4, FLT1	up	Zhuang *et al.* (2013)
Vemurafenib (PLX4032)	BRAF	down	Stuart *et al.* (2015)

Phospho.ELM (Dinkel *et al.*, 2011) and Human Protein Reference Database (Peri *et al.*, 2004). From this collection, 150 autoregulatory sites were excluded in order to prevent biases due to direct regulatory phosphorylations. The five predictors here explored can be classified according to their statistical properties as: parametric tests (Z-test), non-parametric tests (Kolmogorov-Smirnov test and Wilcoxon test), MLRs models and empirical computational approaches relying on data permutations (GSEA algorithm).

The parametric one-sample Z-test compares the mean fold change of all substrates of a given kinase to the mean and variance of all fold changes in the same condition. The non-parametric methods instead, assess the phosphorylation differences between the substrate and the non-substrate fold-change distributions using different metrics: the KS-test estimates the maximum distance between the cumulative probability distributions and the Wilcoxon test evaluates the rank differences among both distributions. The Z-test was implemented in R as previously described in (Casado *et al.*, 2013) and the non-parametric tests were calculated using the available R functions. The GSEA algorithm uses a modified weighted Kolmogorov-Smirnov test to look for the enrichment on substrates of each kinase within the top or bottom of the ranked list of phosphosites. To estimate the significance, the enrichment score was compared against an empirical distribution of scores derived from 10 000 random sets of substrates of the same size. To run the method, we used the in-house developed and freely available ksea R package (https://github.com/evocellnet/ksea).

Each of the aforementioned methods produces a P value that summarizes the significance of the observed kinase regulation. In order to get a kinase activity score that indicates whether the kinase is increasing or decreasing in activity, we calculated the $-\log10$ of the P-values and signed them based on the mean fold change of all substrate phosphosites. If in a given condition the target sites of a

kinase are increasing in phosphorylation, the kinase activity is expected to be also increasing and vice-versa.

In addition to the tests above we predicted kinase regulation using a linear regression approach:

$$Y = \beta X + \psi$$

where, the dependent variable Y represents the phosphosite measurements in a condition and X is the connectivity matrix representing the associations with kinases. X_{ij} equals 1 when phosphosite i is a known substrate of the kinase j, 0 otherwise. ψ represents the normally distributed error of the fit and β are the weights of the model features and also represent the scores of the kinases activity changes in the condition. Kinase activities were estimated by solving the regression model applying an L2-norm regularization, ridge regression, and using the Python machine learning module scikit-learn (Pedregosa *et al.*, 2011). Using a L2-norm regularization allows to ameliorate the impact of collinearity between features, namely different kinases can target similar sets of phosphosites. A regularization coefficient of 0.1 was used.

2.4 Performance evaluation of the methods

To evaluate the performance of the kinase activity scores, we created a set of kinase-condition pairs containing all predicted pairs in the gold standard, as well as the same number of negative pairs generated by randomly sampling kinases and perturbations from the positive set. We repeated this procedure 60 times to generate a diverse space of negative interactions. Due to the incomplete coverage of the MS detection, only substrates for 24 kinases in the gold standard were quantified. Using the ranked lists of positives and the randomly generated negatives, the methods were evaluated using Receiver Operating Characteristic (ROC) and Precision-Recall (PC) curves. Area Under ROC Curve (AUC) and the observed precision at recall 0.5 were systematically used as performance metrics. The benchmark robustness was evaluated by comparing the summary metrics obtained in each of the 60 negative set randomizations. Additionally, the technical variance introduced by each of the methods was assessed by creating 10 permuted sets of kinase-substrate relationships maintaining the same number of substrates per kinase.

2.5 Weighting kinase target sites by the kinase sequence specificity

The weighted Z-test and GSEA methods introduced in this study modify the original statistics to re-rank the substrate quantifications in order to include binding specificities as a proxy for their binding affinities. To calculate the estimated affinity, we constructed a library of position weight matrices (PWMs) from the amino-acid sequences surrounding the known kinase substrates (± 7 residues). Next, we calculated the matrix similarity scores (MSS) to measure the identity between each substrate flanking sequence and their corresponding kinase-PWM (Wagih *et al.*, 2015). The MSS were calculated using a re-implementation of the MATCH tool (Kel *et al.*, 2003) adapted to consider amino-acid sequences (https://github.com/omarwagih/matchtm). The MSS—ranging from 0 to 1—can be interpreted as an approximation of the kinase-substrate binding preferences. The kinase substrate fold changes were then multiplied by their corresponding MSS to obtain their weighted fold changes. As a consequence, differentially regulated substrates not showing affinity for the known kinase-binding motif would see their fold changes diminished. The weighted fold changes were then used to perform the Z-test and the GSEA algorithm as described previously. The fold changes of the sites whose phosphorylation is not catalyzed by the kinase under prediction remained unaltered. When benchmarking the weighted version of the methods we only considered kinases with at least 10 known substrates available to build the PWM. This decreases the benchmark set to 135 positive kinase-condition pairs including 21 kinases covering 57 different perturbations.

2.6 Benchmarking *in vivo, in vitro* and *in silico* substrates

Kinase activity prediction performances were evaluated separately depending on the support of the kinase-substrate relationships: *in vivo*, *in vitro* or *in silico*. The *in vivo* and *in vitro* sets were collected from the curated information available in PhosphositePlus (Hornbeck *et al.*, 2015) while the *in silico* predictions were derived from NetworKIN 3.0 (Linding *et al.*, 2007). NetworKIN integrates consensus substrate motifs based in NetPhorest (Miller *et al.*, 2008) with functional contextual information derived from STRING (Mering, 2003), in order to improve the kinase–substrate inference. A minimum of three substrates per kinase were required for the three sets of evidences. The negative sets in each case were generated using the same criteria as in the full set of substrates. To prevent the biases derived from the different number of available substrates in each category, the analysis was repeated by down-sampling the substrates to the size of the smallest set 25 times.

3 Results

3.1 The benchmark dataset of expected kinase regulation

The critical assessment of existing and forthcoming methods to infer changes in kinase regulation requires an appropriate benchmark datasets. Based on prior knowledge in kinase regulation, we obtained from a previous compilation (Ochoa *et al.*, 2016) a set of human quantitative phosphoproteomic datasets where the assayed perturbations are expected to trigger specific kinase responses. From all the assayed conditions, we defined a benchmark set of 184 positive kinase-condition pairs including 30 kinases and 62 experimental conditions. The selected perturbations include direct molecular regulation via kinase inhibitors or induction of cellular pathways as consequence of cellular processes or differentiation. The curated gold standard set of kinases regulated under MS-quantified conditions is shown in Table 1 along with the expected regulatory effect—up- or down-regulation. The total number of quantified phosphopeptides and more detailed information about the biological perturbations is available in Supplementary Table S1. Although the gold standard list of expected regulations is unlikely to be complete due to downstream and 'off-target' effects, it contains an enriched set of high confidence conditional kinase regulations where direct mechanisms of action have been reported in the literature.

3.2 Substrate-based inference of kinase activity changes

In order to benchmark the changes in kinase activities under different biological perturbations, we compared five different methodologies that integrate quantitative phosphoproteomics data and the network of known kinase substrates. All of the predictors tested have as a premise that known substrates of kinases under regulation display a significant change in phosphorylation over the background changes occurring in all phosphosites in the same condition. We had previously shown, using the same benchmarking strategy, that a GSEA algorithm can correctly predict kinase activity changes

(Ochoa *et al.*, 2016). Here, we evaluated and compared the following methodologies: Z-test, Kolmogorov-Smirnov test, Wilcoxon rank-sum test, a MLR model and the GSEA algorithm (described in Methods). For all methods except the MLR, the −log10 of the *P*-value was used as a proxy of the kinase regulatory response. These values were signed according to the mean fold change of the kinase substrate sites, in order to estimate the kinase regulatory effect—activation or inhibition. For the MLR method we used the beta values of the fitted model to quantify kinase regulation.

3.3 Benchmarking kinase activity changes predicted by different methodologies

For each method, we compared the absolute activity changes for kinases expected to display conditional regulation according to the gold standard, against the activities of the kinases in random conditions (Supplementary Fig. S1). For all methods, kinases under expected regulation display changes in activity significantly higher than random kinases (two-sample Wilcoxon-test, all *P*-values $< 1.33 \times 10^{-6}$). It is noteworthy that some kinases in the negative set also present significant regulation under specific conditions, potentially due to other regulatory events not included in the gold standard. These could include, for example, downstream activation/inhibition of kinases or off-target effects in the cases of drugs. All methods were benchmarked using ROC and PC analysis (Fig. 1 and Supplementary Fig. S2). The GSEA and the Z-test yield comparable results showing higher median AUCs (0.734 and 0.721, respectively) and median precision values at recall 0.5 (0.725 and 0.708) than other approaches. To elucidate the performance variance due to technical variability, we compared the AUCs and precisions derived from the predictions based on the known regulons with those obtained randomizing the network of kinase-substrates maintaining the same degree distribution (Supplementary Fig. S3). As expected, the randomization of the list of kinase targets results in near random predictions.

To evaluate the robustness of the methods to the number of quantified substrates, we compared the aforementioned performances with those obtained only for kinases with ≤ 5 quantified substrates and those with more than 5 (Fig. 1). Overall, activity predictions of kinases with fewer quantified substrates present lower performances. When comparing performances across methods, Wilcoxon, KS and MLR present significantly lower AUC performances than GSEA or Z-test (Wilcoxon-test, *P*-value $< 2.2 \times 10^{-16}$) if the number of quantifications remains below five substrates.

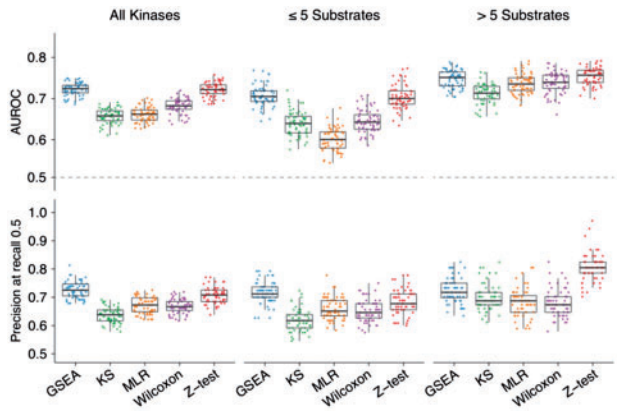

Fig. 1 Comparison of kinase activity prediction performances by methodology and number of known kinase substrates. The filtered gold standard of 149 kinase-condition pairs was assessed against the same number of 60 randomly generated negatives (dots)

3.4 Kinase target site information impacts on the kinase regulation predictions

We reasoned that not all kinase substrates might be equivalent proxies of the upstream kinase activity. We hypothesize that target sites that best fit the known kinase sequence preferences could better represent the regulation of the catalytic kinase. To test this, we obtained position specific scoring matrices that describe the binding specificity of the kinases in the positive set (see 'Methods' section). These were used to score each target site based on their sequence and the values incorporated into the Z-test and GSEA (see 'Methods' section). In these 'weighted' versions of the methodologies, phosphosites that best match the kinase preference were more strongly considered when predicting the regulation of the kinase. We observed a small but significant improvement in AUC and Precision at recall 0.5 whereby these changes tended to perform better than the initial implementations: Z-test-weighted versus Z-test (Wilcoxon-test, *P*-value $= 8.4 \times 10^{-10}$) and GSEA-weighted versus GSEA (Wilcoxon-test, *P*-value $= 1.26 \times 10^{-6}$) (Fig. 2 and Supplementary Fig. S4).

3.5 Activities inferred using *in vivo*, *in vitro* or *in silico* supported substrates

In addition to analyzing the effect of the substrate confidence based on their kinase specificity, we also compared the impact of the evidence supporting the kinase-substrate interactions. Candidate kinase substrates were separated into *in vivo*, *in vitro* and *in silico* supported substrates. Kinase activities predicted using GSEA or Z-test using *in vivo* characterized substrates perform better than predictions using *in vitro* supported substrates (AUC Wilcoxon-test, *P*-value $< 2.2 \times 10^{-16}$) (Fig. 3). Moreover, kinase activities based on substrates determined *in silico* display poor performances. To avoid potential biases derived from the different number of substrates in each set, we repeated the comparison by down-sampling the substrates to the size of the smallest set (Supplementary Fig. S5). Also in this case, *in vivo* characterized substrates present better

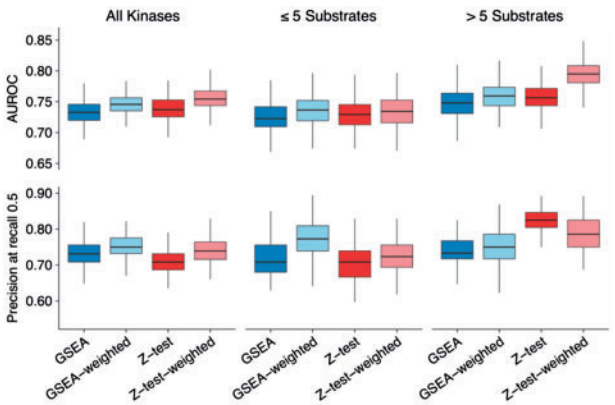

Fig. 2 Comparison of substrate-based kinase activity prediction methods weighting substrates by kinase binding specificity. Similar to Figure 1. 'Weighted' GSEA and Z-test include the similarity of the neighboring region to the kinase recognition motif to correct the relevance of each of the substrates (see 'Methods' section). Weighted and unweighted performances are based on the same set of kinases and conditions (135 positive kinase-condition pairs)

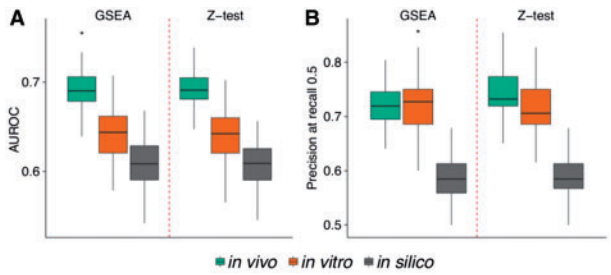

Fig. 3 Benchmark of kinase activity predictions using in vivo, *in vitro* or *in silico* supported substrates. **(A)** AUC performance and **(B)** precision at recall 0.5. Positive kinase-condition pairs are 82, 48 and 75, respectively. The trend is maintained when down-sampling the number of substrate sites to guarantee a similar number of target sites for all cases (Supplementary Fig. S5) or restricting to the same set of kinases and conditions (Supplementary Fig. S6)

performances. Similarly, we restricted the comparison to those positive pairs predicted in all sets observing a similar trend (Supplementary Fig. S6). The poor performance displayed by the *in silico* substrates can be improved by selecting a more stringent prediction score (Supplementary Fig. S7). However, this effect is mostly due to an enrichment of previously characterized substrates, rather than the finding of novel substrates relevant for the kinase activity prediction. Our results highlight the relevance of the curated *in vivo* substrates to accurately infer changes in kinases activities.

4 Discussion

The progress and developments in MS have led to the unbiased exploration of cell signaling changes via phosphoproteomics. Single experiments are now reporting how changes of thousands of phosphorylation sites occur across tens to hundreds of conditions (Abelin *et al.*, 2016; Mertins *et al.*, 2016) and it has been proposed that the activation status of kinases can be inferred by the measurement of its targets sites (Casado *et al.*, 2013; Drake *et al.*, 2012). This approach has been used in several contexts, including the stratification of cancer patients (Drake *et al.*, 2016) or to study kinase pathway changes occurring after specific stimulations (Terfve *et al.*, 2015). Although it is intuitive, that statistical analysis of changes in abundance of target phosphosites reflects the activation status of kinases such approaches have not yet been thoroughly benchmarked and compared. Here, we have selected a benchmark set of conditions where specific kinases are expected to be regulated and use it to compare different methodologies to predict kinase activity changes, as well as the quality of the list of substrates used. The required data is available through our website (at http://phosphate.com) allowing others to easily compare alternative approaches. We also provide an R package including one of the best performing methods (GSEA at http://github.com/evocellnet/ksea).

Overall, we see a strong performance of all methods tested. Some true kinase-conditions pairs were not predicted to be regulated (Supplementary Fig. S1). This could be due to technical issues in the MS data acquisition and quantification or other experimental problems derived from the treatments to stimulate and monitor the kinase activities: incorrect times selected to detect the changes in activity, inadequate stimulation of the control condition or defective concentration of the ligands to generate a cell response etc. However, it should also be considered that complex regulatory mechanisms are involved in the control of kinase activation like feedback regulation, changes in subcellular localization or binding to scaffold proteins. It is possible that different sets of substrates (e.g. in different cellular compartments) might be relevant to estimate the activity for different conditions. The performance of the methods has been judged on the capacity to predict kinase regulation in qualitative terms (i.e. regulated or not). We think the output of the described predictors should relate to the strength of activation but benchmarking these metrics as quantitative predictors would require also quantitative benchmarks of kinase activity in different conditions/states. This knowledge opens the door for the rational selection of kinase target sites that may serve as the best proxies for the activity of the effector kinase.

We observed dependence on the evidence supporting the kinase substrate interaction. This cautions against the use of predicted kinase interactions in such algorithms, at least with their current state of the art. Reversely, this suggests that the co-regulation patterns of phosphosites across conditions can be used as a signal for the prediction of new kinase targets.

Funding

PB acknowledges funding from an HFSP CDA award (CDA00069/2013).

Conflict of Interest: none declared.

References

Abelin,J.G. *et al.* (2016) Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. *Mol. Cell. Proteomics*, **15**, 1622–1641.

Beck,F. *et al.* (2014) Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood*, **123**, e1–e10.

Beli,P. *et al.* (2012) Proteomic investigations reveal a role for RNA processing factor THRAP3 in the DNA damage response. *Molecular Cell*, **46**, p212–225.

Casado,P. *et al.* (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal.*, **6**, rs6.

Choudhary,C., and Mann,M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev.. Mol. Cell Biol.*, **11**, 427–439.

Dephoure,N. *et al.* (2008) A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. USA*, **105**, 10762–10767.

Dinkel,H. *et al.* (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39(Database issue)**, D261–D267.

Drake,J.M. *et al.* (2012) Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proc. Natl. Acad. Sci. USA*, **109**, 1643–1648.

Drake,J.M. *et al.* (2016) Phosphoproteome Integration Reveals Patient-Specific Networks in Prostate Cancer. *Cell*, **166**, 1041–1054.

Engholm-Keller,K. *et al.* (2011) Multidimensional strategy for sensitive phosphoproteomics incorporating protein prefractionation combined with SIMAC, HILIC, and TiO(2) chromatography applied to proximal EGF signaling. *J. Proteome Res.*, **10**, 5383–5397.

Francavilla,C. *et al.* (2016) Multilayered proteomics reveals molecular switches dictating ligand-dependent EGFR trafficking. *Nat. Struct. Mol. Biol.*, **23**, 608–618.

Grosstessner-Hain,K. *et al.* (2011) Quantitative phospho-proteomics to investigate the polo-like kinase 1-dependent phospho-proteome. *Mol. Cell. Proteomics*, **10**, M111.008540.

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

Hornbeck,P.V. *et al.* (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43(Database issue)**, pD512–D520.

Hsu,P.P. *et al.* (2011) The mTOR-regulated phosphoproteome reveals a mechanism of mTORC1-mediated inhibition of growth factor signaling. *Science*, **332**, p1317–1322.

Humphrey,S.J. *et al*. (2015) High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat. Biotechnol*., **33**, 990–995.

Kanshin,E. *et al*. (2015) A cell-signaling network temporally resolves specific versus promiscuous phosphorylation. *Cell Rep*., **10**, 1202–1214.

Kel,A.E. *et al*. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

Kersten,B. *et al*. (2009) Plant phosphoproteomics: an update. *Proteomics*, **9**, 964–988.

Kettenbach,A.N. *et al*. (2011) Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Sci. Signal*., **4**, rs5.

Linding,R. *et al*. (2007) NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic Acids Res*., **36(Database issue)**, D695–D699.

Luerman,G.C. *et al*. (2014) Phosphoproteomic evaluation of pharmacological inhibition of leucine-rich repeat kinase 2 reveals significant off-target effects of LRRK-2-IN-1. *J. Neurochem*., **128**, p561–576.

Matsuoka,S. *et al*. (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, **316**, 1160–1166.

Mering,C. *et al*. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.

Mertins,P. *et al*. (2012) iTRAQ labeling is superior to mTRAQ for quantitative global proteomics and phosphoproteomics. *Mol. Cell. Proteomics*, **11**, M111.014423.

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

Miller-Jensen,K. *et al*. (2007) Common effector processing mediates cell-specific responses to stimuli. *Nature*, **448**, 604–608.

Miller,M.L. *et al*. (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal*., **1**, ra2.

Mischnik,M. *et al*. (2016) IKAP: a heuristic framework for inference of kinase activities from Phosphoproteomics data. *Bioinformatics*, **32**, 424–431.

Nguyen,T.D. *et al*. (2016) The phosphoproteome of human Jurkat T cell clones upon costimulation with anti-CD3/anti-CD28 antibodies. *J. Proteomics*, **131**, 190–198.

Ochoa,D. *et al*. (2016) An atlas of human kinase regulation. *Mol. Syst. Biol*., **12**, 888.

Oda,K. *et al*. (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol. Syst. Biol*., **1**, 2005.0010.

Olsen,J.V. *et al*. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.

Olsen,J.V. *et al*. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal*., **3**, ra3.

Oppermann,F.S. *et al*. (2013) Comparison of SILAC and mTRAQ quantification for phosphoproteomics on a quadrupole orbitrap mass spectrometer. *J. Proteome Res*., **12**, 4089–4100.

Pan,C. *et al*. (2009) Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Mol. Cell. Proteomics*, **8**, 2796–2808.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn Res*., **12**, 2825–2830.

Perfetto,L. *et al*. (2016) SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res*., **44**, D548–D554.

Peri,S. *et al*. (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*., **32(Database issue)**, D497–D501.

Rigbolt,K.T.G. *et al*. (2011) System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci. Signal*., **4**, rs3.

Salek,M. *et al*. (2013) Quantitative phosphoproteome analysis unveils LAT as a modulator of CD3ζ and ZAP-70 tyrosine phosphorylation. *PloS One*, **8**, e77423.

Šalovská,B. *et al*. (2014) Radiosensitization of human leukemic HL-60 cells by ATR kinase inhibitor (VE-821): phosphoproteomic analysis. *Int. J. Mol. Sci*., **15**, p12007–12026.

Schaab,C. *et al*. (2014) Global phosphoproteome analysis of human bone marrow reveals predictive phosphorylation markers for the treatment of acute myeloid leukemia with quizartinib. *Leukemia*, **28**, 716–719.

Stuart,S.A. *et al*. (2015) A phosphoproteomic comparison of B-RAFV600E and MKK1/2 inhibitors in melanoma cells. *Mol. Cell. Proteomics*, **14**, 1599–1615.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Terfve,C.D.A. *et al*. (2015) Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat. Commun*., **6**, 8033.

Wagih,O. *et al*. (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods*, **12**, 531–533.

Weber,C. *et al*. (2012) Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells. *J. Proteomics*, **75**, 1343–1356.

Weigand,S. *et al*. (2012) Global quantitative phosphoproteome analysis of human tumor xenografts treated with a CD44 antagonist. *Cancer Res*., **72**, 4329–4339.

Wilkes,E.H. *et al*. (2015) Empirical inference of circuitry and plasticity in a kinase signaling network. *Proc. Natl. Acad. Sci. USA*, **112**, 7719–7724.

Wu,X. *et al*. (2014) Activation of diverse signalling pathways by oncogenic PIK3CA mutations. *Nat. Commun*., **5**, 4961.

Zhuang,G. *et al*. (2013) Phosphoproteomic analysis implicates the mTORC2-FoxO1 axis in VEGF signaling and feedback activation of receptor tyrosine kinases. *Sci. Signal*., **6**, ra25.