SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
DATA MINING
DATA PROCESSING

Received
22 October 2013

Accepted
11 March 2014

Published
31 March 2014

Correspondence and requests for materials should be addressed to J.-D.Q. (jdqiu@ncu.edu.cn)

PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates

Sheng-Bao Suo[1], Jian-Ding Qiu[1,2], Shao-Ping Shi[1,3], Xiang Chen[1] & Ru-Ping Liang[1]

[1]Department of Chemistry, Nanchang University, Nanchang, 330031, China, [2]Department of Chemical Engineering, Pingxiang College, Pingxiang, 337055, China, [3]Department of Mathematics, Nanchang University, Nanchang, 330031, China.

Protein phosphorylation catalysed by kinases plays crucial regulatory roles in intracellular signal transduction. With the increasing number of kinase-specific phosphorylation sites and disease-related phosphorylation substrates that have been identified, the desire to explore the regulatory relationship between protein kinases and disease-related phosphorylation substrates is motivated. In this work, we analysed the kinases' characteristic of all disease-related phosphorylation substrates by using our developed Phosphorylation Set Enrichment Analysis (PSEA) method. We evaluated the efficiency of our method with independent test and concluded that our approach is reliable for identifying kinases responsible for phosphorylated substrates. In addition, we found that Mitogen-activated protein kinase (MAPK) and Glycogen synthase kinase (GSK) families are more associated with abnormal phosphorylation. It can be anticipated that our method might be helpful to identify the mechanism of phosphorylation and the relationship between kinase and phosphorylation related diseases. A user-friendly web interface is now freely available at http://bioinfo.ncu.edu.cn/PKPred_Home.aspx.

Protein phosphorylation is the most widespread and well-studied post-translational modification (PTM) in eukaryotes and plays a crucial role in the regulation of virtually every cellular behaviour, including DNA repair[1], regulation of transcription[2], apoptosis[3], immune response[4], metabolism[5] and cellular differentiation[6]. In addition, protein phosphorylation catalysed by kinase plays significant regulatory roles in intracellular signal transduction[7]. Therefore, annotation of phosphorylation in proteomes is a first-critical step toward decoding protein function and downstream regulatory networks. Historically, primarily though using low-throughout biological techniques, such as the site-directed mutagenesis, ^{32}P-labeling and degenerate peptide library screening, many novel and specific phosphorylation sites have been discovered. In recent years, the high-throughput studies-large scale mass spectrometry-of protein phosphorylation in different organisms have rapidly accelerated the identification of novel protein phosphorylation data[8–10]. For example, Wiśniewski et al.[11] identified nearly 12,035 unique phosphorylation sites in 4,579 mouse brain proteins using mass spectrometry. Unfortunately, both low-throughput and high-throughput biological techniques for identifying phosphorylation events are costly, time-consuming, and biased toward abundant proteins and proteotypic peptides and cannot provide information regarding the protein kinases that catalyse phosphorylation substrates in detail. Hence, *in silico* prediction of phosphorylation sites is potentially a useful alternative strategy for whole proteome annotation.

Until now, more than a dozen phosphorylation site prediction tools have been developed, such as NetworKIN[12] and recently developed Musite[13], GPS 2.1[14] and PKIS[15]. Musite combined three different features to extract the sequence information, and finally provided general prediction models for six organisms and kinase-specific prediction models for 13 kinases or kinase families. GPS 2.1 reserved the old version of GPS and integrated a novel approach of motif length selection. PKIS incorporated the composition of monomer spectrum (CMS) encoding strategy and support vector machines (SVMs) to identify kinase-specific phosphorylation sites and obtained good prediction results. In addition, to our knowledge, there are five review articles so far that have comprehensively and systematically discussed the methods of computational phosphorylation site prediction, and more information about these tools can be obtained from these five articles[16–20]. Although there are a number of computational methods of phosphorylation prediction and they have made great progress in prediction, some drawbacks of these models in this field still should be noted: (i) Many predictors are based on non-kinase-specific methods and could not give the predicted kinase information for users, such as PPRED[21] and PhosPhAt[22]. (ii) For some kinase-specific tools, the number of experimental phosphorylation substrates for certain kinase is too small

so that lacking of statistical significance, for example, GPS 2.1[14] and KinasePhos 2.0[23] considered the minimum phosphorylation substrates of only three and only one, respectively. (iii) The predicted coverage of kinase types for some kinase-specific tools is too narrow to provide comprehensive kinase prediction and the kinase hierarchical classification is not considered, for example, Li et al.[24] only considered eight kinase families for predictor construction. PKIS[15] only involved single kinases prediction but prediction for kinase family and kinase group were not provided. (iv) The provided stringency for some predictors is relatively high as they paid more attention to specificity than to sensitivity, although the prediction accuracy is promising, the true positive rate is low, such as Scansite[25] and Musite[13]. Therefore, new methods must be established and used for more effective and comprehensive identification of kinase-specific phosphorylation site.

In addition, protein phosphorylation regulates many aspects of cell life, whereas abnormal phosphorylation is a cause or consequence of disease[26]. As we all know, protein kinases are one of the most ubiquitous families of signalling molecules in the human cell, accounting for approximately 2% of the proteins encoded by the human genome[27]. The family of genes most frequently contributing to different diseases such as neurodegenerative disease and cancer is the protein kinase gene family[28] which is implicated in a huge number of tumorigenic functions including immune evasion, proliferation, antiapoptotic activity, metastasis and angiogenesis. It is now well-appreciated that many disease pathways involve abnormal regulation of phosphorylation events. And recently, increasingly more experimental observations have suggested that protein kinase could indirectly or directly influence the abnormal protein phosphorylation and further result in diseases[29]. For example, Bose et al.[30] have proven that Double-stranded RNA dependent kinase (PKR) could induce Glycogen Synthase Kinase Aβ (GSK-3β) activation, tau phosphorylation and apoptosis in human neuroblastoma cells exposed to tunicamycin or Ab peptide 1–42 and activated PKR is increased in brains with Alzheimer's disease (AD). Wong et al.[31] have found that Cyclin-dependent kinase 5 (Cdk5)-mediated phosphorylation of Endophilin B1 is essential for autophagy induction and neuronal loss in models of Parkinson's disease (PD). By analysing a unique randomized tamoxifen trial including breast cancer patients receiving no adjuvant treatment, Busch et al.[32] showed for the first time that patients low in ERK phosphorylation in cancer associated fibroblasts (CAFs) did not respond to tamoxifen treatment despite having estrogen-receptor alpha. Paccez et al.[33] in his recent review described that the receptor tyrosine kinase Axl has been implicated in the malignancy of different types of cancer. Saini et al.[34] described that the phosphatidylinositol 3-kinase (PI3K)/AKT/mammalian target of rapamycin (mTOR) and the Raf/mitogen-activated and extracellular signal-regulated kinase (MEK)/extracellular signal-regulated kinase (ERK) signalling pathways are critical for normal human physiology, and also commonly dysregulated in several human cancers, including breast cancer (BC). Ariadna et al.[35] demonstrated for the first time that the implication of DYRK1A overexpression in a developmental alteration of the central nervous system associated with Down syndrome (DS). In this regard, comprehensive analysis of abnormal phosphorylation associated with different types of kinases will be helpful to promote the understanding of how abnormal actions of the kinase are involved in regulating biological processes and how they affect susceptibility to diseases.

As described above that the kinase-specific methods could give the kinase information for predicted phosphorylation substrates and further help researches to explore the regulatory mechanism between kinase and abnormal phosphorylation substrates. So in this paper, we firstly constructed an efficient kinase-specific phosphorylation predictor and then used this tool to predict and analyse the types of kinases for all disease-related phosphorylation substrates.

Results

The ability of PSEA to recognize true phosphorylation sites. We first checked the ability of PSEA to correctly recognize the phosphorylation sites. To enlarge the prediction coverage of different types of kinase and to gain insights into function and evolution of kinase, we classified all kinases into a hierarchy of single kinase, kinase family and kinase group (for more information please see supporting information for data preparation and Tables S1–S4). A leave-one-out method was used to test the performance for each positive and negative set. For each test, a known peptide was picked and the others were treated as the predefined peptide set. The *P*-values for positive and negative peptides were then calculated by the PSEA method. The results are represented in Tables S5–S13. From these tables we can see that for all single kinases of phosphoserine, at the high stringency cut-off, there are 13 terms and 18 terms whose sensitivities and specificities are larger than 80%, respectively. When decreasing the stringency, the corresponding specificity reduces, but most of single kinases can still obtain high performance with both sensitivity and specificity larger than 80%. From the receiver operating characteristic (ROC) evaluation, we can find there are 15 kinases with areas under ROC curves (AUCs) larger than 80% and all kinases' AUCs are larger than 70% except the CaMK2-α, as shown in Figure 1. For single kinases of phosphothreonine, most of the predictors also can obtain promising performance except the PKCA (Table S6, Figure S1). The predictive sensitivity of single kinases of phosphotyrosine is a little lower compared with that of phosphothreonine and phosphoserine (Table S7 and Figure S2), mainly because the sequence conservation of phosphotyrosine kinase is lower than that of phosphothreonine and phosphoserine and the threshold setting for phosphotyrosine, which is the same as that of phosphothreonine and phosphoserine, is too stringent. Compared with single kinase method for phosphorylation prediction, the prediction performance of kinase family and kinase group methods is a little depressed, which could express that although some kinases in the same family or group have the similar structure and function, the slight substantial difference is still exist.

Test on independent data and comparison with other existing methods. To evaluate the performance of our method, we made

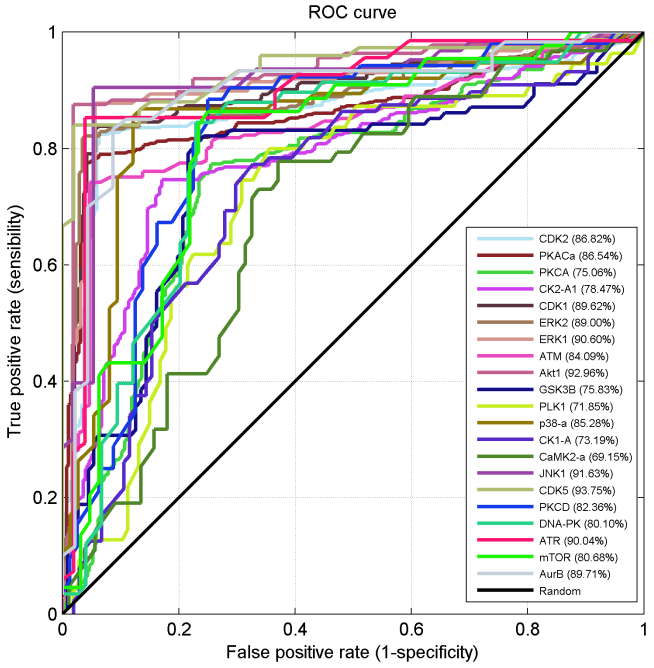

Figure 1 | The ROC curve and the corresponding AUCs for phosphoserine prediction of different single kinases.

comparisons with other existing predictors. Here we put our independent test set into three recently developed and online available methods: GPS 2.1[14], Musite[13] and NetworKIN[12]. GPS 2.1 provides three levels of stringency (High, Medium and Low) with different choices of threshold values. Musite and NetworKIN support continuous stringency adjustment to meet different confidence requirements for users. In order to compare our method with Musite and NetworKIN conveniently, we chose three appropriate levels of stringency (also called High, Medium and Low) for Musite and NetworKIN with specificity (or score) as 85%, 90% and 95% respectively through testing. All these stringency in GPS 2.1, Musite and NetworKIN are all relatively high as they all paid more attention to specificity than to sensitivity. The comparisons of predictive performance between our method and other prediction methods are shown in Figure 2. For different levels of stringency, although the Acc and MCC of the CK2 in GPS 2.1, the CDK and MAPK in Musite and the CDK in NetworKIN are slightly higher than those of our method, most of kinase families in our method are superior to other methods. For example, for the CK1 family, when the stringency level of these four methods is High, the MCC of our method reaches to 43.77%, which is about 19.58%, 28.20% and 44.09% higher than that in GPS 2.1, Musite and NetworKIN, respectively. Also for the Src family, the MCC of our method in these three levels of stringency is about 50%, which is about 25% ~ 50% higher than that of three other methods. The results show that the above three tools can achieve high specificity, but sacrificing sensitivity would therefore result in a low MCC. Our method offers good Sp as well as high Sn, which also illustrates that our method is superior to the current methods. Note that, when performing the comparisons, we used a prediction model that was trained from a dataset excluding the protein sequences in the independent test dataset. However, for GPS 2.1, Musite and NetworKIN, some of the test proteins might have been included in their training processes, and thus, the prediction performances may be biased favourably toward these tools in the comparisons. This possibility means that the performance improvement of our method over these tools might be underestimated. Compared with these existing methods, it is worth mentioning that the formula of our method is much more concise or at least comparable with previous established programs. More importantly, the reasonably good performance of our method reflects that our method can effectively evaluate the sequence similarity of phosphorylation substrates for different types of kinases.

Predicting the types of kinase for disease-related phosphorylation substrates. Protein kinases are a superfamily of proteins involved in crucial cellular processes such as cell cycle regulation and signal transduction. Accordingly, they play an important role in disease biology. To contribute to the study of the relation between kinases and diseases, we performed a prediction analysis by predicting corresponding kinases of all disease-related phosphorylation substrates which could result in different human diseases. To have large prediction coverage of protein kinases, we determined to use the kinase family predictors to predict the kinase families of all disease-related phosphorylation substrates one by one. The results are shown as orange bars in Figure 3 (we didn't consider the IKK because the prediction performance is not good enough). We could find that MAPK family could catalyse 63.02% of disease-related phosphoserine, while PLK family could only catalyse 32.45% of disease-related phosphoserine. We also checked the effects of these disease-related phosphorylation substrates from different databases (as described in the Methods), and found there are 464 substrates that have annotation information of kinases. From the known information, we found 111 (23.92%) disease-related phosphorylation substrates can be catalysed by the MAPK. After processing these substrates by using our predictors with High stringency, we predicted 103 (92.78%) substrates that can be catalysed by MAPK (all predicted and known kinase information, the source and other detailed information of all collected disease-related phosphorylation substrates can be downloaded from our web site).

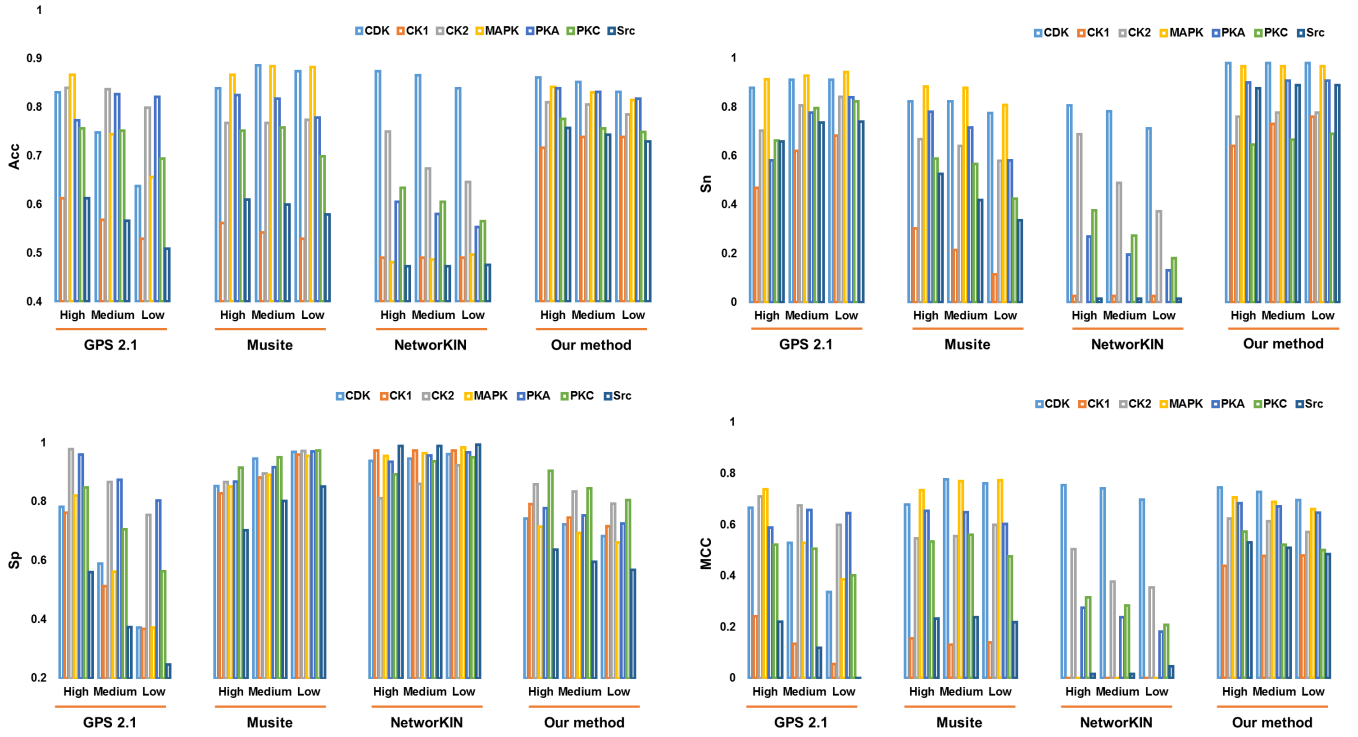

Figure 2 | Comparison of our method with other existing methods on independent set for different kinase families. (A), compared by accuracy (Acc); (B), compared by sensitivity (Sn); (C), compared by specificity (Sp); (D), compared by Matthews correlation coefficient (MCC).

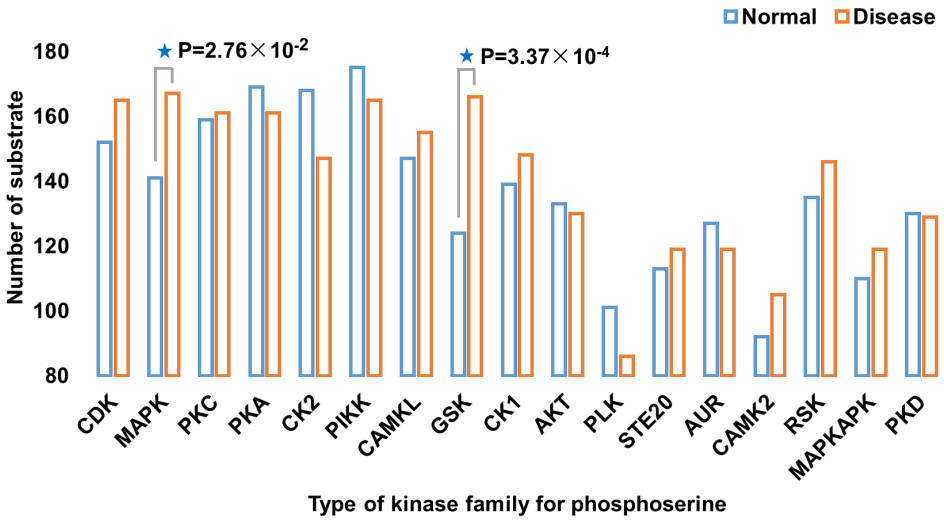

Figure 3 | The data statistics of predicted phosphoserine kinase family types for disease-related and normal phosphorylation substrates. Significant differences (*P*-value) refer to the Two-sided category. Statistical significance was calculated with a Fisher exact test.

Besides, the protein-protein interaction network in STRING database was used to analyse the relationship between the MAPK3 and PLK1 kinases (considering the limitation of the data of collected kinase-specific phosphorylation, we only analysed the largest quantity of MAPK3 and PLK1 kinases which contained in MAPK and PLK families, respectively) and all disease-related phosphorylation substrates, as shown in Figure 4. We could find that MAPK3 contacts much more disease-related phosphorylation substrates than that of PLK1 ($P = 4.66 \times 10e{-}8$). It only displayed the direct contacted interactions, considering the important roles in the regulation of phosphorylation, it is reasonable to believe that there must be many other indirect interactions which MAPK3 kinase reacts with disease-related phosphorylation substrates. From the above analysis, not only could we conclude that MAPK kinase family might occupy a relatively large proportion in abnormal phosphorylation and further result in different diseases but our method of kinase-specific prediction of phosphorylation can effectively predict the corresponding kinase type of phosphorylation substrates.

Significance analysis for the predicted kinase families of disease-related and normal phosphorylation substrates. To compare the difference of kinase families between disease-related and normal phosphorylation substrates, we also predicted the same size of normal phosphorylation substrates randomly selected from all collected phosphorylation sites. For kinase family of phosphoserine, the predicted results for disease-related and normal phosphorylation

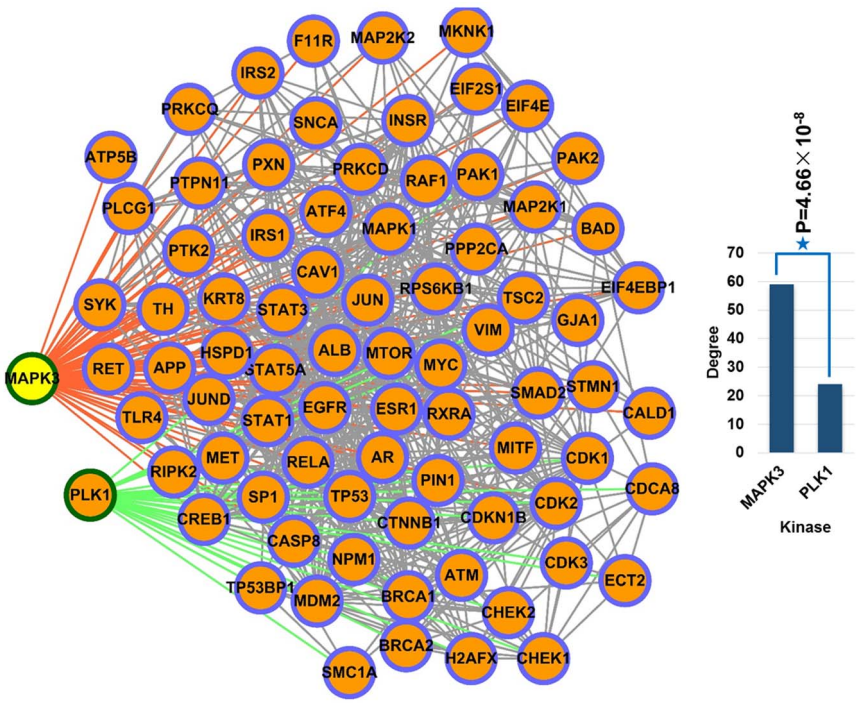

Figure 4 | The relationship between specified kinases (MAPK3 and PLK1) and disease-related phosphorylation substrates (only displayed directly contacted interactions). The nodes with green border line represent the specified kinases and the nodes with blue border line represent the diseased phosphorylation substrates. Bar plot represents the degrees of MAPK3 and PLK1.

substrates are shown in Figure 3. By using the Fisher exact test, we could clearly find that MAPK family ($P = 2.76 \times 10e{-}2$) and GSK family ($P = 3.37 \times 10e{-}4$) have significant difference in catalysing disease-related and normal phosphorylation substrates, which might infer that there two kinase families have more direct or indirect association compared with other kinase families. Although we cannot conclude that these two kinase families must involve in the pathological processes, it provided a useful resource and guidance for further experimental validation. Moreover, the analysis for phosphothreonine and phosphotyrosine was also carried out in the same way, but we could not find the significant difference ($P > 0.05$) between the disease-related and normal phosphorylation substrates among different types of phosphothreonine or phosphotyrosine kinase families, the results are shown in Figures S3–S4.

Function analysis of disease-related phosphorylation substrates. From the above results, we found the predicted MAPK and GSK kinase families have significant difference in catalysing disease-related and normal phosphorylation substrates. The question is whether our prediction results are reasonable and reliable enough. So in this part, we carefully analysed the characteristic of all disease-related phosphorylation proteins, and found whether there are connection between these proteins and the kinase families of MAPK and GSK.

To better understand the distribution of the disease-related phosphorylation substrates in function protein groups, we analysed our data to see if there is over- or under- representation (compared to the normal phosphorylation substrates) of function elements such as pathways and gene ontology (GO). Firstly, DAVID program[36,37] was used to analyse the pathway to further explore functional aspects of disease-related and normal phosphorylation substrates. Here the top 10 significant results ($P < 1.00 \times 10e{-}10$) were shown in Figure 5. We could find that 36.50% of disease-related phosphorylation substrates are involved in cancer pathways (containing pathways in cancer, prostate cancer, non-small cell lung cancer and pancreatic cancer) and 8.25% are involved in neurotrophin signalling pathway. After carefully analysing the characteristics of MAPK and GSK kinase families, we found that MAPK kinase family can regulate the related actions to induce the human colon, lung and breast cancers[32,38,39] and the glycogen synthase kinase 3β (GSK3β), one member of GSK kinase family, can catalyse Tau phosphorylation and plays an important roles in the genesis and maintenance of neurodegenerative changes associated with Parkinson's disease[40]. In addition, after carefully analysing the effects of all disease-related phosphorylation substrates we collected, we found those disease-related substrates could mainly cause neurodegenerative diseases (Parkinson's disease, Alzheimer's disease and Huntington's disease) and cancers (Carcinoma and Cancer), and the total numbers of those which could result in these two types of disease are 170 (the percentage of the whole disease-related phosphorylation substrates (806) which this disease (170) accounts for is 21.09%) and 424 (52.61%), respectively, which are also consistent with the results of pathway analyses.

As a result of the continuing advances made in previous studies, protein phosphorylation was found to target broad substrates in different biological processes. The collection of disease-related and normal phosphorylation substrates from databases provided an opportunity to analyse the functional abundance and diversity of protein phosphorylation. Here, we statistically analysed the enriched biological processes, molecular functions and cellular components with the gene ontology (GO) annotations and compared the differentiated GO terms with Fisher exact test (Two-sided category) for the human disease-related and normal phosphorylation substrates. The 10 most over-represented terms of these three criteria are shown in Table 1. We could find that chemical stimulus (GO:0070887, GO:0042221) and other stimulus (GO:0009719, GO:0009725, GO:0009605) in biologic processes are the most differentiated GO terms and they are all over-represented in diseased phosphorylation substrates. The results are consistent with the roles of MAPK kinase family, as Giuseppe et al.[41] have provided the evidence that p38(MAPK) and ERK1/2 dictate cell death/survival response to different pro-oxidant stimuli via p53 and Nrf2 in neuroblastoma cells SH-SY5Y, also, Min et al.[42] have discovered that NF kappa B and JNK/MAPK activation mediates the production of major macrophage- or dendritic cell-recruiting chemokine in human first trimester decidual cells in response to proinflammatory stimuli. For molecular functions, the binding functions such as receptor binding (GO:0005102), protein complex binding (GO:0032403) and identical protein binding (GO:0042802) are the most differentiated GO terms, which are also consistent with previous studies of MAPK and GSK families[43,44]. The diseased phosphorylation substrates are

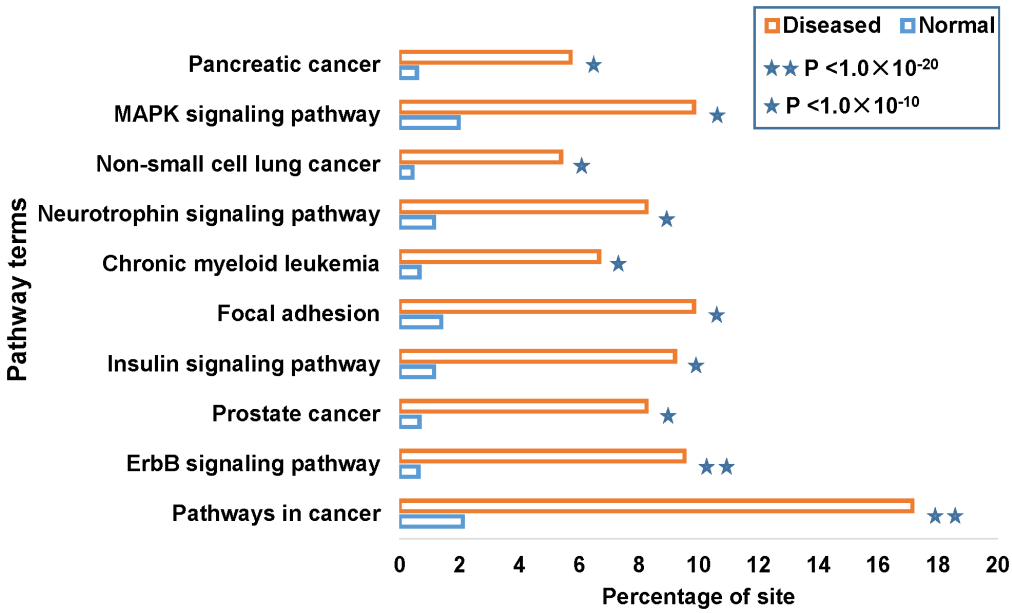

Figure 5 | The data statistics of pathway terms for disease-related and normal phosphorylation substrates. Significant differences (*P*-value) refer to the Two-sided category. Statistical significance was calculated with a Fisher exact test.

Table 1 | Statistical comparison of the GO terms of the disease-related and normal phosphorylation substrates

Description of GO term	Diseased phosphor. Num.[a]	Diseased phosphor. Per.[b] (%)	Normal phosphor. Num.	Normal phosphor. Per. (%)	FDR[c]	*P*-Value[d]	Over/Under[e]
The most different biologic processes							
Cellular response to chemical stimulus (GO:0070887)	161	51.27	1094	19.15	1.18e−30	7.87e−35	Over
Response to chemical stimulus (GO:0042221)	209	66.56	1852	32.42	3.80e−29	5.05e−33	Over
Response to organic substance (GO:0010033)	169	53.82	1306	22.86	1.14e−26	2.28e−30	Over
Response to endogenous stimulus (GO:0009719)	125	39.81	750	13.13	2.71e−26	7.22e−30	Over
Response to inorganic substance (GO:0010035)	83	26.43	370	6.48	8.58e−23	2.85e−26	Over
Response to hormone stimulus (GO:0009725)	107	34.08	622	10.89	1.42e−22	6.48e−26	Over
Response to external stimulus (GO:0009605)	134	42.68	951	16.65	2.06e−22	1.10e−25	Over
Regulation of multicellular organismal process (GO:0051239)	161	51.27	1337	23.41	6.90e−22	4.13e−25	Over
Cellular response to organic substance (GO:0071310)	129	41.08	908	15.90	1.44e−21	9.58e−25	Over
Cellular component movement (GO:0006928)	130	41.40	927	16.23	1.91e−21	1.40e−24	Over
The most different molecular functions							
Receptor binding (GO:0005102)	93	29.62	711	12.45	9.37e−13	8.41e−15	Over
Protein complex binding (GO:0032403)	62	19.75	356	6.23	1.10e−12	9.38e−15	Over
Identical protein binding (GO:0042802)	93	29.62	715	12.52	1.08e−12	1.02e−14	Over
Protein tyrosine kinase activity (GO:0004713)	32	10.19	105	1.84	2.02e−11	2.35e−13	Over
Protein kinase activity (GO:0004672)	63	20.06	407	7.13	5.41e−11	6.84e−13	Over
Molecular transducer activity (GO:0060089)	70	22.29	504	8.82	2.42e−10	3.33e−12	Over
Protein dimerization activity (GO:0046983)	75	23.89	584	10.22	1.29e−09	2.10e−11	Over
Phosphotransferase activity (GO:0016773)	64	20.38	463	8.11	2.61e−09	4.41e−11	Over
Kinase binding (GO:0019900)	69	21.97	534	9.35	5.84e−09	1.05e−10	Over
Protein kinase binding (GO:0019901)	65	20.70	487	8.53	6.76e−0-	1.25e−10	Over
The most different molecular components							
Cytosol (GO:0005829)	181	57.64	1705	29.85	4.43e−20	4.42e−23	Over
Cytoplasmic part (GO:0044444)	265	84.39	3474	60.82	1.65e−16	6.41e−19	Over
Cell periphery (GO:0071944)	176	56.05	1820	31.86	2.54e−15	1.30e−17	Over
Plasma membrane (GO:0005886)	170	54.14	1736	30.39	4.91e−15	2.78e−17	Over
Cytoplasm (GO:0005737)	294	93.63	4301	75.30	1.08e−14	6.77e−17	Over
Mitochondrion (GO:0005739)	90	28.66	697	12.20	4.49e−12	4.63e−14	Over
Cytoskeletal part (GO:0044430)	115	36.62	1036	18.13	7.61e−12	8.40e−14	Over
Cytoskeleton (GO:0005856)	134	42.68	1359	23.79	9.78e−11	1.27e−12	Over
Anchoring junction (GO:0070161)	46	14.65	237	4.149	1.23e−10	1.63e−12	Over
Adherens junction (GO:0005912)	45	14.33	234	4.097	2.79e−10	3.88e−12	Over

[a]The number of diseased phosphorylation substrate in different GO terms.
[b]The proportion of diseased phosphorylation substrate in different GO terms.
[c]The false discovery rate of Fisher exact test (Two-sided category).
[d]The P-value of Fisher exact test (Two-sided category).
[e]Over- or under- representation of diseased phosphorylation compared with normal phosphorylation in different GO terms.

mainly localized in cytoplasm and membrane (such as GO:0005829, GO:0044444 and GO:0005886) according to the cellular component analysis. The above analysis showed that MAPK and GSK families might have significant differences in biological processes, molecular functions and cellular components for diseased and normal phosphorylation substrates, which could be helpful to design the protein kinase inhibitors for abnormal phosphorylation related diseases. Therefore, from the pathway and GO analysis we could further confirm that the MAPK and GSK families might have some really regulatory effects in disease-related phosphorylation.

Computation programs construction and web server. Our main aim is to develop an open platform for computational analysis of phosphorylation of human proteins. We chose the C# programming language to execute all of calculation for its powerful computing capability and excellent portability. Besides that, we have constructed web service platform by using the Asp.net (C#). This web service of predicting kinase-specific phosphorylation site is freely accessible for academic researchers at: http://bioinfo.ncu.edu.cn/PKPred_Home.aspx. In download page, all of data used in this paper can be downloaded from this web site, such as all phosphorylation data, independent data and disease-related phosphorylation data. Bug fixing and minor changes of phosphorylation prediction model will be done. The improved phosphorylation prediction model will be constructed when the new phosphorylation sites data and kinase data become available.

With the web service of predictor, each query phosphorylation site (S/T/Y) in sequences can get a score. A higher score indicates a higher probability of the phosphorylation site by the selected kinase term. To control the false-positive predictions, we suggest users pay more attention to the predicted phosphorylation sites with *P*-values lower than the top 10%. In this case, the estimated specificity will be higher than 90%. Specific phosphorylation sites (S/T/Y) passing the suggested cut-off are highlighted by colour in the table of prediction results on the web site. The background set should contain unreported phosphorylation sites; as such, the specificity is very likely underestimated. In our opinion, this cut-off should be loosened once interaction between kinase and query protein occurs. In applications, users can adjust the cut-off values according to the trade-off between discovering more putative kinase-specific phosphorylation sites and making fewer false-positive predictions.

Discussion

Protein phosphorylation regulates most aspects of cell life, whereas abnormal phosphorylation is a cause or consequence of disease. A growing interest in developing orally active protein kinase inhibitors has recently culminated in the approval of the first of these drugs for clinical use. Protein kinases have now become the second most important group of drug targets, after G-protein-coupled receptors[26]. With increasingly more disease-related phosphorylation substrates were discovered in the clinic or in clinical trials, it is crucial to explore the relationship between the protein kinases and these disease-related phosphorylation substrates, such as whether the specific kinase has the potential relation to specific disease and which type of kinase can specifically result in the abnormal phosphorylation and further cause disease. On the basis of large amount of kinase-specific phosphorylation data and disease-related phosphorylation data, it has become both a possibility and a priority to determine what the functional implication of protein kinases are and how to use the abnormal regulatory information of phosphorylation to develop the corresponding protein kinase inhibitors for related diseases.

In this paper, we analysed the kinases' characteristic of all collected disease-related phosphorylation substrates for the first time on the basis of our kinase-specific prediction method and predicted that MAPK and GSK kinase families are enriched in the environment of disease-related phosphorylation, which could be helpful to design the corresponding protein kinase inhibitors to cure the diseases. Our findings about MAPK and GSK kinase families also have been confirmed by some experimental observations. For example, Noble et al.[45] have reported that activation of p38α MAPK can lead to increased activities of proinflammatory cytokines, such as tumor necrosis factor-α and interleukin 1β. This observation suggested that p38 selective inhibition could be a therapeutically useful route to treatment of a number of inflammatory and autoimmune diseases. Meijer et al.[43] have reviewed that more than 30 inhibitors of GSK have been identified to treat several diseases, including Alzheimer's disease and other neurodegenerative diseases.

Note that the capacity of disease-related phosphorylation data we collected is not lager enough, we cannot rule out the possibility that these data were gleaned from small-scale studies, which often aimed at studying only the better-known protein kinases, and not their less explored relatives, while the data in background set almost come from high-throughput studies. So the statistics analysis between disease-related and normal phosphorylation data might have bias. But to authors' knowledge, there is no efficient method for large-scale analysis of the relationship between kinases and phosphorylation related diseases with computational model. Although we are not completely sure that this statistics analysis is reliable, it provided a useful resource for further experimental medicinal considerations. In addition, with the rapid development of the biotechnology, more and more diseased-related phosphorylation data will be detected, and our idea of analysing the disease-related phosphorylation could give another choice for the pathogenesis research.

When personalized medicine is the next frontier of scientists, industry and the general population, it is important to develop computational approaches that can lead to a better understanding of the etiology of a disease. Considering the essential roles of phosphorylation in protein functions, integration of phosphorylation kinases and molecular information is a sensible step in this direction because it provides a structural and functional perspective to both the human protein kinase and abnormal phosphorylation. Our work can not only be used in pathophysiological diagnosis researches of abnormal phosphorylation but also in the selection of protein kinase inhibitors of clinical applications.

Based on the existing data, we constructed the kinase-specific predictors and carried out the systematic analysis of all disease-related phosphorylation and obtained some satisfactory result, but several issues must be solved in future researches. (i) The phosphorylation sites used in the predictors were mostly identified by high-throughput methods, which may have inherent bias in terms of representing the global phosphorylation events and consequently affect the performance of prediction. As techniques like electron transfer dissociation and alternative proteases are helping to resolve technology limitations, more complete phosphorylation data sets will be released. We will update our predictors as the new data become available. (ii) We have only labelled positive data, but we do not have labelled background and negative data (i.e. we still cannot ensure the non-phosphorylation substrates are truly negatives, although we have discarded the inaccessible substrate sites by using the structural filters). (iii) The types of kinase for disease-related phosphorylation substrates are only predicted from the aspect of kinase family due to the data limitation of single kinases (we only considered those kinases whose substrates is larger than 50), then the analysis for these diseased phosphorylation might not be elaborate enough. More kinases will be included in the system with the availability of more kinase-specific phosphorylation data and more detailed analysis for disease-related phosphorylation substrates will be carried out.

Methods

The PSEA method. Gene Set Enrichment Analysis (GSEA) was developed and used on DNA microarray data to detect coordinated expression changes in a group of functionally related genes and then was applied to find the putative functions of the

long non-coding RNAs[46–48]. Taking advantage of the idea of GSEA, we proposed a new method called PSEA (Phosphorylation Set Enrichment Analysis) to detect new sites phosphorylated by a specific kinase, kinase family and kinase group. For each term, we focused on finding sites which were similar in sequence with discovered ones. We treated the phosphorylation sites and their surrounding amino acids as phosphorylated peptides. Phosphorylated peptides from the above three levels of kinase hierarchical classification formed kinase specific peptides sets (see Tables S1–S3). To determine whether a query peptide could be phosphorylated or not, we just needed to judge whether this query peptide was similar to the phosphorylated peptides in that set. The PSEA method we developed can efficiently estimate this similarity and the significance of the similarity. The following calculation was based on the conception of the 15 amino acid long peptides. Compared with the original GSEA method several necessary changes have been made[49]. The details of the PSEA method (shown in Figure 6) are described as follows:

(i) *Prepare the input data*. Predefined each kinase group, kinase family and single kinase specific phosphorylation site peptide set S_P containing N peptides, for example, S_P (CDK2/S) containing 318 peptides and S_P (CDK2/T) containing 182 peptides, as shown in Table S1. Predefined background peptide set S_b containing 10,000 randomly selected peptides from whole background set.

(ii) *Calculate Similarity Score*. First, similarity scores between the query peptide (denoted as P_{query}) and each peptide in S_p and S_b were calculated according to local sequence similarity[13]. For example, for two local sequence fragments S_1 and S_2 (the window size is $2n + 1$), define the distance $D(S_1,S_2)$ between S_1 and S_2 as:

$$D(S_1,S_2) = \frac{\sum_{i=-n}^{n} Sim(S_1(i),S_2(i))}{2n+1} \quad (1)$$

$$Sim(a,b) = \frac{M(a,b) - min(M)}{max(M) - min(M)} \quad (2)$$

where, Sim is derived from the normalized amino acid substitution matrix. a and b are the two amino acids, M is the substitution matrix (BLOSUM62 was used in this paper). Then, all similarity scores were mixed together and ranked from high to low. According to the above steps, we know that if P_{query} is similar to the peptides in S_p for a specific kinase level (single kinase, kinase family or kinase group), these peptides in S_p should be enriched at the top of the ranked similarity score list, and P_{query} is likely a novel phosphorylation site for that kinase level.

(iii) *Calculate Enrichment Score* (*ES*). To test the enrichment of the peptides in S_p at the top of the ranked list, a running sum score was calculated by walking down the ranked list. Here, we denote d_i as the similarity scores between peptide P_{query} and peptide P_i (contained in S_p), and D as the sum of $|d_i|$ for all P_i. While walking down the ranked list, the running sum score increased $|d_j|/D$ when encountering a peptide P_j in S_p and decreased 1/10,000 when encountering a peptide in S_b. Finally, *ES* was defined as the maximum of the running sum score.

(iv) *Evaluate Significance of ES*. To evaluate the significance of the *ES* for a given peptide, a total of 999 peptide sets with the same size as S_p were randomly selected from the background peptides, and denoted as S_{b1} to S_{b999}. Then the *ES* for each set was calculated one by one by treating each set as predefined peptide set. Finally, all of these 1000 *ES* ($ES(S_{b1})$, $ES(S_{b2})$…$ES(S_{b999})$, plus $ES(S_p)$) were ranked from high to low. If the rank of $ES(S_p)$ is L, the nominal P-value of the given peptide was calculated as $L/1000$. The P-value should be ranked from 0.001 to 1, with the minimum interval of 0.001. The smaller the P-value is, the more significant the chance that the given peptides were phosphorylated by the specific kinase. In practice, the number of randomly selected peptide sets can be changed according to different needs.

For the PSEA method described above, there are several advantages for classification. Firstly, it does not require the balanced number of positive and negative datasets, as many machine learning methods did. Secondly, it directly calculates the sequence similarity between two single peptides. In most of the cases, peptides in the kinase set can be divided into several subsets which share little variation with each other. The PSEA method is particularly suitable for these cases, because the *ES* of a given peptide would be significant as soon as it is similar to some but not all peptides in the kinase set.

Predicting the disease-related phosphorylation substrates. The information about disease-related phosphorylation which could cause severe diseases were obtained from the PhosphoSitePlus[50], which is an online systems of biology resource providing comprehensive information and tools for the study of post-translational modifications (PTMs) of the protein, and providing MS/MS records for sets of modification sites observed in specified diseases, cell lines, and tissues. We collected 320 human disease-related phosphorylation proteins, which contain 806 disease-related phosphorylation terms. After collecting all the data, we further consulted the SwissVariant[51], UniProtKB/Swiss-Prot and PubMed databases about their effects, and the references to the variations, phosphorylation information and related diseases of corresponding proteins. The main alteration of these disease-related phosphorylation data are hyperphosphorylation and hypophosphorylation, which could further result in a series of diseases, such as Alzheimer's disease, breast cancer, and so on. As one disease-related phosphorylation site could cause two or more diseases and we only analyse the relationship between kinase and disease-related phosphorylation substrate, then we only count one time for these phosphorylation sites. So after removing the redundant disease-related phosphorylation terms, we lastly collected 265 disease-related phosphoserines, 63 disease-related phosphothreonines and 237 disease-related phosphotyrosines (the data can be

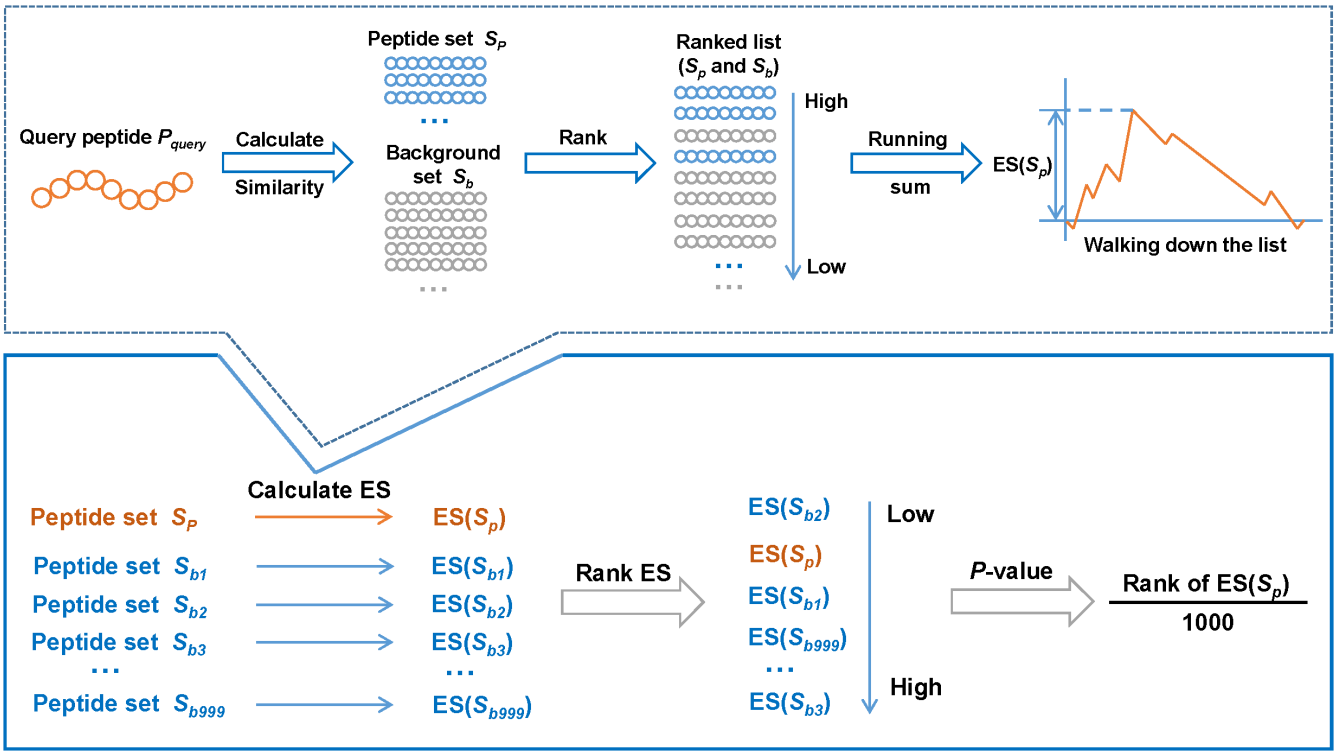

Figure 6 | Detailed processes of the PSEA method.

obtained from our online web site). With the kinase-specific predictor, PSEA, we predicted the corresponding kinase family for each disease-related phosphorylation site (because of the data limitation, the single kinase predictors only cover a small part of all kinase in PSEA, so we use kinase family predictors to predict the disease-related phosphorylation sites).

Enrichment analysis of normal and disease-related phosphorylation sites. Gene Ontology (GO) is a major bioinformatics initiative. It meets the need for consistent descriptions of gene products in different databases. It has been developed to manage the overwhelming mass of current biological data from a computational perspective and become a standard tool to annotate gene products for various databases[52]. Enrichment analysis was performed to identify over- or under-represented GO terms in the loss and gain of disease-related phosphorylation, compared to the normal phosphorylation in the entire human phosphorylation proteome. According to the two-sided category of Fisher exact test, the *P*-value of 0.05 was considered significant and was calculated. Besides, the KEGG pathway was analysed by using online DAVID program[36,37], which provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. We also used the STRING database (version 9.1) of protein-protein interaction network to analyse and explain the relationship between the kinase and the disease-related phosphorylation substrates.

Assessment of the performance. The performance of PSEA was evaluated with the positive and negative test sets, with the accuracy (Acc), specificity (Sp), sensitivity (Sn) and the Matthews correlation coefficient (MCC) measurements defined in the following way:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{5}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{6}$$

where, TP is the number of true positive predictions, TN is the number of true negative predictions, FN is the number of false negatives, and FP is the number of false positives. The accuracy denotes the percent of correct prediction in both the positive and negative sets. The sensitivity and specificity depend on the threshold used for the prediction. A highly stringent threshold will improve the specificity but reduce the sensitivity, whereas less stringent threshold will increase the sensitivity at the price of lower specificity. The MCC accounts for the true and false positives and negatives and is usually regarded as a balanced measure that can be used even if the classes are of very different sizes. Besides, receiver operating characteristic (ROC) curves were calculated and plotted based on Sp and Sn to evaluate the prediction performance of single kinases, and areas under ROC curves (AUCs) were also calculated based on the trapezoidal approximation.

Threshold setting. Threshold setting is also a difficult problem. In general, most of us chose different threshold for every protein kinase. Here we proposed that a uniform rule to choose cut-off values based on calculated *P*-values. For all single kinases, kinase families and kinase groups, the high, medium and low thresholds were established with *P*-values smaller than 0.002, 0.005 and 0.015, respectively. The high threshold is recommended to test a large-scale prediction of human phosphorylation sites. The medium threshold often reduces the stringency to be useful in small-scare experiments. Also, the low threshold reduces the specificity to improve sensitivity considerably which is very useful in extensively experimental identification of all potential phosphorylation sites in substrates.

1. Wood, C. D., Thornton, T. M., Sabio, G., Davis, R. A. & Rincon, M. Nuclear localization of p38 MAPK in response to DNA Damage. *Int. J. Biol. Sci.* **5**, 428–437 (2009).
2. Uddin, S. *et al.* Role of Stat5 in Type I interferon-signaling and transcriptional regulation. *Biochem. Biophys. Res. Commun.* **308**, 325–330 (2003).
3. Zhang, J. W. & Johnson, G. V. W. Tau protein is hyperphosphorylated in a site-specific manner in apoptotic neuronal PC12 cells. *J. Neurochem.* **75**, 2346–2357 (2000).
4. Kim, S. H. & Lee, C. E. Counter-regulation mechanism of IL-4 and IFN-alpha signal transduction through cytosolic retention of the pY-STAT6:pY-STAT2:p48 complex. *Eur. J. Immunol.* **41**, 461–472 (2011).
5. Bu, Y. H. *et al.* Insulin receptor substrate 1 regulates the cellular differentiation and the matrix metallopeptidase expression of preosteoblastic cells. *J. Endocrinol.* **206**, 271–277 (2010).
6. Lian, I. *et al.* The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Gene. Dev.* **24**, 1106–1118 (2010).
7. Steffen, M., Petti, A., Aach, J., D'Haeseleer, P. & Church, G. Automated modelling of signal transduction networks. *BMC Bioinf.* **3**, 34 (2002).
8. Boersema, P. J. *et al.* In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol. Cell. Proteomics* **9**, 84–99 (2010).
9. Sugiyama, N. *et al.* Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. *Mol. Syst. Biol.* **4**, 193 (2008).
10. Zhai, B., Villen, J., Beausoleil, S. A., Mintseris, J. & Gygi, S. P. Phosphoproteome analysis of Drosophila melanogaster embryos. *J. Proteome Res.* **7**, 1675–1682 (2008).
11. Wisniewski, J. R., Nagaraj, N., Zougman, A., Gnad, F. & Mann, M. Brain phosphoproteome obtained by a FASP-based method reveals plasma membrane protein topology. *J. Proteome Res.* **9**, 3280–3289 (2010).
12. Linding, R. *et al.* Systematic discovery of in vivo phosphorylation networks. *Cell* **129**, 1415–1426 (2007).
13. Gao, J. J., Thelen, J. J., Dunker, A. K. & Xu, D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteomics* **9**, 2586–2600 (2010).
14. Xue, Y. *et al.* GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel.* **24**, 255–260 (2011).
15. Zou, L., Wang, M., Shen, Y., Liao, J. & Li, A. PKIS: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC Bioinf.* **14**, 247 (2013).
16. Trost, B. & Kusalik, A. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* **27**, 2927–2935 (2011).
17. Xue, Y. *et al.* A summary of computational resources for protein phosphorylation. *Curr. Protein Pept. Sci.* **11**, 485–496 (2010).
18. Miller, M. L. & Blom, N. In Phospho-Proteomics, Vol. 527. (ed. de Graauw, M.) 299–310 (Humana Press Inc, 999 Riverview Dr, Ste 208, Totowa, Nj 07512–1165 USA, 2009).
19. Hjerrild, M. & Gammeltoft, S. Phosphoproteomics toolbox: Computational biology, protein chemistry and mass spectrometry. *FEBS Lett.* **580**, 4764–4770 (2006).
20. Kobe, B., Kampmann, T., Forwood, J. K., Listwan, P. & Worth, R. I. Substrate specificity of protein kinases and computational prediction of substrates. *BBA-Proteins Proteom.* **1754**, 200–209 (2005).
21. Biswas, A. K., Noman, N. & Sikder, A. R. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinf.* **11**, 273 (2010).
22. Heazlewood, J. L. *et al.* PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.* **36**, D1015–D1021 (2008).
23. Wong, Y. H. *et al.* KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* **35**, W588–W594 (2007).
24. Li, T. T., Du, P. F. & Xu, N. F. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS One* **5**, e15411 (2010).
25. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635–3641 (2003).
26. Cohen, P. Protein kinases - the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **1**, 309–315 (2002).
27. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
28. Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
29. Brognard, J. & Hunter, T. Protein kinase signaling networks in cancer. *Curr. Opin. Genet. Dev.* **21**, 4–11 (2011).
30. Bose, A. *et al.* Modulation of Tau phosphorylation by the kinase PKR: implications in Alzheimer's disease. *Brain Pathol.* **21**, 189–200 (2011).
31. Wong, A. S. L. *et al.* Cdk5-mediated phosphorylation of endophilin B1 is required for induced autophagy in models of Parkinson's disease. *Nat. Cell. Biol.* **13**, 568–U167 (2011).
32. Busch, S., Ryden, L., Stal, O., Jirstrom, K. & Landberg, G. Low ERK Phosphorylation in cancer-associated fibroblasts is associated with tamoxifen resistance in pre-menopausal breast cancer. *PLoS One* **7**, e45669 (2012).
33. Paccez, J. D., Vogelsang, M., Parker, M. I. & Zerbini, L. F. The receptor tyrosine kinase Axl in cancer: biological functions and therapeutic implications. *Int. J. Cancer* **134**, 1024–1033 (2014).
34. Saini, K. S. *et al.* Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer. *Cancer Treat. Rev.* **39**, 935–946 (2013).
35. Laguna, A. *et al.* Triplication of DYRK1A causes retinal structural and functional alterations in Down syndrome. *Hum. Mol. Genet.* **22**, 2775–2784 (2013).
36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
37. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

38. Ahn, E. H. & Schroeder, J. J. Sphinganine causes early activation of JNK and p38 MAPK and inhibition of AKT activation in HT-29 human colon cancer cells. *Anticancer Res.* **26**, 121–127 (2006).
39. Liu, Y. *et al*. Activation of ERK-p53 and ERK-Mediated Phosphorylation of Bcl-2 Are Involved in Autophagic Cell Death Induced by the c-Met Inhibitor SU11274 in Human Lung Cancer A549 Cells. *J. Pharmacol. Sci.* **118**, 423–432 (2012).
40. Duka, T., Duka, V., Joyce, J. N. & Sidhu, A. alpha-Synuclein contributes to GSK-3 beta-catalyzed Tau phosphorylation in Parkinson's disease models. *Faseb. J.* **23**, 2820–2830 (2009).
41. Filomeni, G., Piccirillo, S., Rotilio, G. & Ciriolo, M. R. p38(MAPK) and ERK1/2 dictate cell death/survival response to different pro-oxidant stimuli via p53 and Nrf2 in neuroblastoma cells SH-SY5Y. *Biochem. Pharmacol.* **83**, 1349–1357 (2012).
42. Li, M., Wu, Z. M., Yang, H. & Huang, S. J. NF kappa B and JNK/MAPK activation mediates the production of major macrophage- or dendritic cell-recruiting chemokine in human first trimester decidual cells in response to proinflammatory stimuli. *J. Clin. Endocrinol. Metab.* **96**, 2502–2511 (2011).
43. Meijer, L., Flajolet, M. & Greengard, P. Pharmacological inhibitors of glycogen synthase kinase 3. *Trends Pharmacol. Sci.* **25**, 471–480 (2004).
44. Veeranna. *et al*. Calpain mediates calcium-induced activation of the Erk1,2 MAPK pathway and cytoskeletal phosphorylation in neurons - relevance to Alzheimer's disease. *Am. J. Pathol.* **165**, 795–805 (2004).
45. Noble, M. E. M., Endicott, J. A. & Johnson, L. N. Protein kinase inhibitors: Insights into drug design from structure. *Science* **303**, 1800–1805 (2004).
46. Guttman, M. *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
47. Subramanian, A. *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P. Natl. Acad. Sci. Usa.* **102**, 15545–15550 (2005).
48. Mootha, V. K. *et al*. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
49. Li, T. T. *et al*. Characterization and prediction of lysine (K)-acetyl-transferase specific acetylation sites. *Mol. Cell. Proteomics* **11**, M111.011080 (2012).
50. Hornbeck, P. V. *et al*. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261–D270 (2012).
51. Yip, Y. L. *et al*. The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. *Hum. Mutat.* **23**, 464–470 (2004).
52. Harris, M. A. *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).

Acknowledgments

The authors thank Phospho. ELM, PhosphoSitePlus and UniProtKB/Swiss-Prot databases for supplying phosphorylation data on proteins. This work was supported by Program for New Century Excellent Talents in University (NCET-11-1002); and the National Natural Science Foundation of China (21175064, 21305062, 20605010).

Author contributions

Conceived and designed the experiments: J.D.Q. and S.B.S. Performed the experiments: J.D.Q. and S.B.S. Analysed the data: S.B.S. Contributed reagents/materials/analysis tools: J.D.Q. Wrote the paper: J.D.Q. and S.B.S. Responsible for the design development: J.D.Q., S.B.S., S.P.S. and R.P.L. Responsible for the computational modelling: J.D.Q., S.B.S., S.P.S. and X.C. Responsible for the web interface development: J.D.Q., S.B.S. and X.C.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Suo, S.-B., Qiu, J.-D., Shi, S.-P., Chen, X. & Liang, R.-P. PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates. *Sci. Rep.* **4**, 4524; DOI:10.1038/srep04524 (2014).