Journal of Molecular Cell Biology Advance Access published February 26, 2015

Article

Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility

Yongbo Wang[1], Han Cheng[1], Zhicheng Pan[1], Jian Ren[2], Zexian Liu[1,*], and Yu Xue[1,*]

[1]Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

[2]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China

*Correspondence to: Zexian Liu, Tel/Fax: +86-27-87793172, E-mail: lzx@hust.edu.cn; Yu Xue, Tel: +86-27-87793903, Fax: +86-27-87793172, E-mail: xueyu@hust.edu.cn

Abstract

Large-scale sequencing has characterized an enormous number of genetic variations (GVs), and the functional analysis of GVs is fundamental to understanding differences in disease susceptibility and therapeutic response among and within populations. Using a combination of a sequence-based predictor with known phosphorylation and protein－protein interaction information, we computationally detected 9606 potential phosSNPs (phosphorylation-related single nucleotide polymorphisms), including 720 known, disease-associated SNPs that dramatically modify the human phosSNP-associated kinase－substratekinase－substrate network. Further analyses demonstrated that the proteins in the network are heavily associated in various signaling and cancer pathways, while cancer genes and drug targets are significantly enriched. We re-constructed four population-specific kinase－substrate networks and found that several inherited disease or cancer genes, such as IRS1, RAF1, and EGFR, were differentially regulated by phosSNPs. Thus, phosSNPs may influence disease susceptibility and be involved in cancer development by reconfiguring phosphorylation networks in different populations. Moreover, by systematically characterizing potential phosphorylation-related cancer mutations (phosCMs) in 12 types of cancers, we observed that both types of GVs preferentially occur in the known cancer genes, while a considerable number of phosphorylated proteins, especially those over-representing cancer genes, contain both phosSNPs and phosCMs. Furthermore, it was observed that phosSNPs were significantly enriched in amplification genes identified from breast cancers and tyrosine kinase circuits of lung cancers. Taken together, these results should prove helpful for further elucidation of the functional impacts of disease-associated SNPs.

Keywords: genetic variation, nsSNP, phosSNP, phosphorylation, phosSNP-associated kinase－substrate network

Introduction

In the post-genomic era, the large-scale identification of human genetic variations (GVs) is critically important for an understanding of the differences in both the susceptibility to diseases and response to therapeutic treatment (Manolio et al., 2008; Ren et al., 2010; Gunther et al., 2011; Gonzaga-Jauregui et al., 2012). After over a decade of effort, the HapMap Project, the 1000 Genomes Project, Cancer Genome Project, and other similar projects have identified an enormous number of individual GVs, such as germline single nucleotide polymorphisms (SNPs), somatic mutations and structural variants (Greenman et al., 2007; Manolio et al., 2008; 1000 Genomes Project Consortium, 2010; Forbes et al., 2010). Although most GVs may be incidental and not alter gene function (Collins et al., 1998), genome-wide association (GWA) studies have recently detected tens of thousands of GVs that are associated with a variety of diseases (Manolio et al., 2008; Stenson et al., 2009; Li et al., 2012b). SNPs in the coding region (cSNPs), especially non-synonymous SNPs (nsSNPs), are significantly associated with certain specific phenotypes and diseases (Cargill et al., 1999; Gunther et al., 2011). Additionally, a considerable proportion of somatic mutations have been predicted to be “cancer drivers” that advance disease progression (Greenman et al., 2007; Carter et al., 2009; Vandin et al., 2012; Reimand and Bader, 2013). In this regard, characterizing the functional consequences of GVs is of critical importance to understand individual genetic differences, thus supporting the concept of “personalized medicine” (Ren et al., 2010; Gonzaga-Jauregui et al., 2012).

It is well documented that nsSNPs generate deleterious effects by modifying protein structural conformation (Reumers et al., 2006; Yue and Moult, 2006). However, nsSNPs also influence the post-translational modifications (PTMs) of proteins (Savas and Ozcelik, 2005; Erxleben et al., 2006; Gentile et al., 2008; Radivojac et al., 2008; Yang et al., 2008a; Ryu et al., 2009; Li et al., 2010). For example, Savas and Ozcelik (2005) predicted that 15 nsSNPs potentially create or remove phosphorylation sites in cell-cycle and DNA repair proteins.The Armstrong group first suggested the concept of aberrant phosphorylation or “phosphorylopathy” caused by GVs in 2006 (Erxleben et al., 2006) and predicted 16 phosphorylopathies in human ion channel genes (Gentile et al., 2008). One of the predicted results, a K897T nsSNP in the human ether-a-gogo-related gene 1 (hERG1), was experimentally shown to generate a novel Akt phosphorylation site that inhibited

channel activity (Gentile et al., 2008). Subsequently, Yang et al. (2008a) mapped cSNPs to flanking regions of known phosphorylation sites and proposed that 64 of them were potentially disrupted by nsSNPs. Recently, systematic analysis of inherited disease- or cancer-associated GVs that potentially alter protein phosphorylation has emerged as an important undertaking (Radivojac et al., 2008; Ryu et al., 2009; Li et al., 2010; Reimand and Bader, 2013). Ryu et al. (2009) predicted that both nsSNPs and somatic mutations are implicated in various cancers and inherited diseases by changing phosphorylation. Also, Radivojac et al. (2008) showed that phosphorylation-associated mutations are significantly enriched in cancers and inherited diseases compared with germline nsSNPs, and further observed that loss of PTM sites might be an important mechanism in these diseases (Li et al., 2010). Furthermore, Liu et al. (2013) found that a cancer-patient-derived mutation R81T on Sin1 impairs the Sin1 phosphorylation, which leads to the hyper-activation of mTORC2 and facilitates tumorigenesis. In addition, Reimand and Bader (2013) developed a statistical approach termed ActiveDriver, and predicted 44 potential cancer driver genes with significant phosphorylation-associated single-nucleotide variants (pSNVs) from nine cancer data sets. Taken together, these studies showed that GVs, such as mutations and SNPs, might contribute to the development of cancers or inherited diseases through the reconfiguration of phosphorylation signaling.

Previously, we defined phosSNP (phosphorylation-related SNP) as an nsSNP that influences the protein phosphorylation state (Ren et al., 2010), with the concurrence of Dr David L. Armstrong in the form of personal communications. Using the kinase-specific predictor of GPS 2.0 (Xue et al., 2008), we directly predicted phosphorylation sites in both original and nsSNP-containing sequences (Ren et al., 2010). By comparison of the prediction results, we identified a total of ~70% of the nsSNPs as potential phosSNPs. However, because only a small proportion of serine, threonine, and tyrosine residues can be phosphorylated *in vivo*, *ab initio* prediction of phosphorylation sites merely from primary sequences is error-prone. Thus, more filters are needed to be included to reduce the false positive predictions.

In this work, we first predicted the exact kinase information for the original and nsSNP-containing protein sequences by manually forming links of the 407 human kinases with their corresponding predictors in GPS 2.1 (Song et al., 2012). To improve the accuracy of this process, we only considered potential phosSNPs that change known phosphorylation sites, using experimentally

identified phosphorylation sites as a highly efficient filter. Because the short motifs around phosphorylation sites do not provide full recognition specificity for kinases *in vivo*, the protein—protein interactions (PPIs) between kinases and substrates were adopted as an additional filter to remove false positive hits. With this procedure, a total of 9606 potential phosSNPs in 7946 proteins were identified with a high degree of confidence. Then we re-constructed the human phosSNP-associated kinase—substrate phosphorylation network and found the network topology to be dramatically changed by phosSNPs. Also, cancer genes and drug targets were significantly over-represented in the network, with the proteins in the network being heavily implicated in a number of signaling and cancer pathways. Furthermore, we observed that the known inherited disease- or cancer-associated mutations were significantly over-represented in the predicted phosSNPs, with a two-fold enrichment. The strong relation between phosSNPs and diseases suggested that phosSNPs may be associated with human disease by reconfiguring the phosphorylation network. Moreover, a number of disease- or cancer-associated genes were found to be differentially regulated by phosSNPs in different populations. Thus, phosSNPs may influence disease and cancer susceptibility by differentially altering the phosphorylation networks in different populations in different regions. In addition, we further systematically analyzed phosphorylation-related cancer mutations (phosCMs) in 12 types of cancer samples. Although both phosCMs and phosSNPs significantly occur more frequently in cancer genes, the known cancer genes were found to be significantly over-represented in phosSNP-containing proteins in most types of cancers. Thus, our results suggest that phosSNPs, along with phosCMs, may be involved in driving cancer progression. Taken together, our results provide help for further experimental analysis of the molecular mechanisms of known disease-associated phosSNPs as well as new leads for personalized medicine.

Results

Systematic characterization of phosSNPs with a high degree of confidence

In this work, three major improvements were introduced to improve the prediction accuracy of human phosSNPs (Figure 1). First, GPS 2.1 hierarchically clustered homologous kinases into the categories of group, family, and subfamily, but did not predict the exact kinase information for a given residue (Xue et al., 2011). Previously, we manually selected 407 human kinases for 56 serine/threonine kinase (STK) and 21 tyrosine kinase (TK) specific predictors (Song et al., 2012). Thus, from 1060070 nsSNPs, we directly predicted raw phosSNPs that exhibited changed phosphorylation patterns because of specific kinases (Figure 1). Second, PhosSNP 1.0 contained too many results, most phosphorylation sites of which were *ab initio* predicted merely from the primary sequences (Ren et al., 2010). Since recent phosphoproteomic experiments have identified ten thousands of phosphorylation sites (Song et al., 2012), this information was used to filter out the false positive hits (Figure 1). As previously described, we defined a phosphorylation site peptide PSP(15, 15) as a phosphorylation site flanked by 15 residues upstream and 15 residues downstream (Ren et al., 2010; Xue et al., 2011). We retrieved all PSP(15, 15) items from the phosphorylation data set, and exactly mapped them to the original and mutated RefSeq proteins, respectively. The PSP(15, 15) items that could not be mapped to any SNP were discarded. Then for a specific SNP, the phosphorylation states of original and mutated PSP(15, 15) were further compared. We considered a SNP as a phosSNP only if it could induce the phosphorylation change of the PSP(15, 15), such as creating or abolishing the phosphorylation site, or changing the kinase type of the phosphorylation site (Figure 1). Previously, our analyses suggested that the physical interaction between a kinase and a substrate efficiently improves the prediction accuracy of site-specific kinase—substrate relations (ssKSRs) (Song et al., 2012). Thus, both experimentally identified and pre-predicted PPIs (Exp. & STRING PPIs) were used as an additional filter. We only preserved phosSNPs that change the ssKSRs supported by the PPI information (Figure 1). These phosSNPs were classified into different types based on the definitions. In total, we computationally identified 9606 phosSNPs in 7946 RefSeq sequences (Figure 1).

In the absence of any filters, approximately 57.2% of the total nsSNPs (606321/1060070) were predicted to be phosSNPs by GPS 2.1 (Table 1). This result is moderately lower than the one in

our previous analysis (Ren et al., 2010). When known phosphorylation information was considered, only 40344 nsSNPs (~3.8%) were predicted to be phosSNPs (Table 1). Thus, it is evident that a large proportion of potentially false positive hits were filtered out. The predicted number was further reduced to 9606 (Exp. & STRING PPI, 0.91%) and 2496 (Exp. PPI, 0.24%) using the different PPI data sets (Table 1). Because the Exp. PPI filter was too stringent and only a small number of hits were predicted, we used the predicted results obtained with Exp. & STRING PPI as the core data set for further analyses.

PhosSNPs dramatically reconfigures the human kinase—substrate phosphorylation network

Kinases can phosphorylate substrates and be also modified by other kinases. Thus, a kinase—substrate phosphorylation network can be re-constructed from ssKSRs between kinases and substrates (Song et al., 2012). Because site-specific phosphorylation is the result of upstream regulatory kinases (Nilsson, 2012), phosSNPs reconfigure the human kinase—substrate network by either creating new or disrupting the original ssKSRs. However, because a substrate can be phosphorylated by a kinase at multiple sites, adding or removing one ssKSR does not disrupt the kinase—substrate relation (KSR) if multiple ssKSRs exist. In this regard, the impact of phosSNPs on the kinase—substrate network is classifiable into four types: (i) Added (+), one or multiple new ssKSRs were introduced into an unrelated kinase—substrate pairing; (ii) Removed (-), all existing ssKSRs in a kinase—substrate pair were disrupted; (iii) Changed (C), the KSR was not changed by either adding or removing one or multiple ssKSRs; (iv) Unchanged (N), all the ssKSRs were identical in the original and mutated proteins.

By comparing the predicted ssKSRs for the original and mutated proteins, we constructed a phosSNP-associated kinase—substrate phosphorylation network with 18056 KSRs for 374 kinases and 2270 substrates (Figure 2). For further comparison and analysis, the KSRs of both original and mutated proteins were integrated and visualized as a single network (Figure 2). It should be noted that 154 (~41.2%) predicted kinases did not have any phosSNPs in the network. However, their functions can still be influenced, because phosSNPs change their substrate profiles by reconfiguring the ssKSRs between these non-mutated kinases and substrates (Figure 2). In this regard, all proteins in the phosSNP-associated kinase—substrate network were taken together for further analysis.

From the network results, it is evident that a considerable proportion of KSRs were added (2989, ~16.6%) or removed (4322, ~23.9%) by phosSNPs (Figure 2). Also, although the KSRs were retained, a number of KSRs (6582, ~36.5%) were changed by adding or removing ssKSRs (Figure 2). Only a proportion of KSRs (4163, ~23.1%) were not influenced (Figure 2). In addition, statistical analysis suggested that the topological features of the kinase－substrate phosphorylation networks for the original and mutated proteins were significantly different (Table 2). Taken together, the results indicate that the human kinase－substrate network was dramatically reconfigured by the phosSNPs.

The phosSNP-associated kinase－substrate network is highly associated with various cancers

In a network, highly connected genes, known as hubs, usually play central roles in mediating the signal communication or exchange in multiple pathways (Jin et al., 2007; Wang et al., 2007; Zaman et al., 2013). From the phosSNP-associated kinase－substrate network, the top 10 substrates (Figure 3A) and kinases (Figure 3B) with most KSRs were shown, respectively. If the substrates are also kinases, only directed KSRs from other kinases to these kinases were considered. These genes act as hubs in the network, and their functions might be greatly influenced by the reconfiguration of KSRs. For example, EGFR, the substrate with the most KSRs, was implicated in signaling transduction from extracellular into appropriate cellular responses including regulated growth, proliferation, and survival (Hochgrafe et al., 2010) (Figure 3A). Previously, a structural modeling study suggested that a SNP of rs17290699 might disrupt the hydrogen and ion bonds between H988 and E690, and further influence EGFR dimer formation (Choura et al., 2010). Also, IRS1 is a signaling adapter involved in biological processes such as apoptosis, cell growth, and cell transformation through the regulation of insulin signaling (Chang et al., 2002) (Figure 3A), while mutations in IRS1 have been reported to be implicated in diabetes mellitus type 2 and cancers (Stenson et al., 2009; Hochgrafe et al., 2010). Among top 10 kinases, AKT1, which has the most KSRs, regulates various processes including metabolism, proliferation, and angiogenesis (Figure 3B). The aberrant signaling of AKT was involved in a variety of complex diseases (Manning and Cantley, 2007). Taken together, the KSRs of hub genes were preferentially changed by phosSNPs, which may contribute to the reconfiguration of signaling pathways and further contribute to the cancers or other diseases.

In addition, we found that 199 (~7.5%) and 219 (~8.3%) genes of the network are cancer

genes, using the different cancer gene data sets (Figure 3C). Nearly 20% of the network proteins are drug targets (Figure 3C). With the hypergeometric distribution, the statistical results suggested that both the cancer genes and drug targets are significantly enriched in the phosSNP-associated kinase－substrate network (Figure 3C, *p*-value << 0.01). Although the protein kinases are known to be highly associated with cancer (Haber and Settleman, 2007), excluding the kinases from the phosSNP-associated kinase－substrate network still resulted in a strong correlation between the phosSNP-containing proteins and cancer genes or drug targets (Supplementary Table S1, *p*-value << 0.01). Previously, using a combination of the sequence-based predictions and the PPI information, we predicted a total of 192756 ssKSRs at 25962 human phosphorylation sites, and constructed a human protein phosphorylation network (PPN) among 380 kinases and 4140 substrates (Song et al., 2012). Indeed, cancer genes and drug targets were shown to be significantly enriched in the PPN against the human proteome (*p*-value << 0.01), probably because the phosphorylation is highly involved in a variety of signaling processes and highly associated with cancers and diseases. However, using the human PPN as the background, it is evident that phosSNPs are able to further enrich cancer genes (Figure 3D, *p*-value < 0.01).

We statistically analyzed the distribution and diversity of Gene Ontology (GO) terms for proteins in the phosSNP-associated kinase－substrate network with the hypergeometric distribution (*p*-value < 1E-15). The results suggest that network proteins are significantly over-represented in a broad spectrum of biological processes and functions, such as protein phosphorylation, signal transduction, and apoptotic processes (Supplementary Table S2). Also, by mapping network proteins to the pathways of the Kyoto Encyclopedia of Genes and Genomes (KEGG), the statistical results indicated that these proteins were significantly associated with a number of cancer pathways, such as pathways in cancer (*p*-value = 5.69E-23), chronic myeloid leukemia (*p*-value = 7.56E-23), glioma (*p*-value = 3.07E-18), and prostate cancer (*p*-value = 2.30E-17) (Table 3). Again, we observed that cancer-associated pathways were still significantly enriched even after excluding protein kinases from the network proteins (Supplementary Table S3). Moreover, the statistical enrichment analysis of the Disease Ontology (DO) terms suggested that network proteins were significantly enriched in a number of diseases and cancers (Supplementary Table S4). In particular, the top 3 most significant DO terms were cancer (*p*-value = 2.53E-146), breast cancer (*p*-value = 1.8E-60), and prostate cancer (*p*-value = 1.31E-47) (Supplementary Table S4). In addition, the

protein complexes were computed from the network by MCODE v1.32 in Cytoscape (Bader and Hogue, 2003). The enrichment analysis of the KEGG pathways was performed for the proteins in each of the complexes (Supplementary Table S5). The top 4 complexes with the highest scores were shown, with three of them associated with cancers (Supplementary Figure S1). Taken together, several lines of evidence suggest that the proteins in the human phosSNP-associated kinase—substrate network are highly associated with cancer.

PhosSNPs are associated with known human diseases

We classified all predicted phosSNPs into four types (I, II, III, and IV) (Table 1), and several typical examples are shown in Figure 4. In total, it was found that up to 720 unique phosSNPs (~7.4%) were included in the ClinVar, a public resource of relations among GVs and human phenotypes (Landrum et al., 2014). With the hypergeometric distribution, the *p*-value was calculated to be 1.19E-81 (the enrichment ratio = 2.18), which suggests that phosSNPs are significantly involved in human diseases (Supplementary Table S6). In our results, the G691S nsSNP (rs1799939) in RET was reported to be associated with primary vesicoureteral reflux (pVUR) in the Quebec patients by removing the S691 phosphorylation site (Yang et al., 2008b). Here, we predicted it as a Type I (+) phosSNP to generate a new phosphorylation site for PKCA (Figure 4A). Also, it is known that NF-κB inhibitor α (NFKBIA) is phosphorylated by the IκB kinase at S32, while an S32I nsSNP (rs28933100) removes the phosphorylation site and enhances its inhibitory activity that is associated with autosomal-dominant ectodermal dysplasia with immunodeficiency (AD-EDA-ID) (Courtois et al., 2003). Consistent with the experimental observation (Courtois et al., 2003), we predicted it to be a Type I (-) phosSNP that removes the S32 phosphorylation site of IKKB (Figure 4B). By mapping the prediction results to the GWASdb (Li et al., 2012b), we also observed that 10 non-redundant phosSNPs in 10 genes are significantly associated with human diseases (Table 4). For example, MLXIPL, a carbohydrate-responsive element-binding protein, was previously identified as being associated with metabolic disorders, with an nsSNP of A358V (rs35332062) (Kettunen et al., 2012). Here, we predicted it to be a Type II (-) phosSNP as the result of inhibiting the PKACA-mediated phosphorylation at S361 (Figure 4D). The predictions for BCL10 (Figure 4C), SPTBN1 (Figure 4E), and EGFR (Figure 4F) still remain to be validated. Taken together, phosSNPs are highly associated with human diseases and useful for further experimental investigation.

PhosSNPs differentially reconfigure the phosphorylation network in different populations

By directly mapping, 1320 predicted phosSNPs were detected from the 1000 Genomes Project (Supplementary Table S7). Based on the allele frequency information, we first constructed four population-specific phosSNP-associated kinase－substrate networks for the Ad Mixed American (AMR), East Asian (ASN), African (AFR), and European (EUR), respectively (Table 5 and Supplementary Figure S2). PhosSNPs without any information on the allele frequencies were not considered in any network. In the four networks, there were 1525－2368 KSRs among 286－303 kinases and 265－392 non-kinase substrates (Table 5). Also, nearly half of the kinases in these networks were predicted not having any phosSNPs. This result is similar with to that from the analysis for the integrative phosSNP-associated kinase－substrate networks. In particular, only a limited proportion of KSRs (34.9%－38.0%) were not influenced (Table 5). By statistically analyzing the MCC and degree distributions for the original and mutated proteins, we observed that phosSNPs dramatically altered all population-specific networks (p-value << 0.01, Wilcoxon signed-rank test, Table 5).

To test whether phosSNPs differentially influenced the phosphorylation networks in different populations, we calculated the network proteins in a pairwise manner for the KSRs among the four networks. With Yates' chi-squared test, the KSRs of 30 proteins were determined significantly different in at least a pair of populations (Supplementary Table S8, p-value < 0.01), with the results shown in Table 6 (p-value < 1E-4). In our results, the gene with the greatest number of different KSRs is IRS1 (insulin receptor substrate 1) between the populations of AFR and EUR (Table 6, p-value = 3.74E-14). The KSRs of IRS1 are also differentially present for AMR-EUR (p-value = 1.42E-08), AFR-ASN (p-value = 6.54E-08), and AMR-ASN (p-value = 4.76E-04) (Table 6 and Supplementary Table S8). As a known drug target (Reimand and Bader, 2013), IRS1 plays a critical role in apoptosis, cell growth, and cell transformation by regulating insulin signaling (Chang et al., 2002), and a number of mutations in IRS1 have been shown to be associated with diabetes mellitus type 2 or noninsulin-dependent diabetes mellitus (NIDDM) from the ClinVar annotations (Landrum et al., 2014). From the SNPs of the 1000 Genomes Project, we predicted up to 9 nsSNPs in IRS1 to be phosSNPs (Supplementary Table S7). The two phosSNPs, rs148962208 (P1079S, allele frequency: 0.01 in ASN) and rs1801276 (A512P, allele frequency: 0.002 in AFR; 0.02 in EUR) (Supplementary

Table S7), were annotated as diabetes-associated mutations (Supplementary Table S6). The former phosSNP, rs148962208, was first identified in a Chinese population and the authors suggested that the hydrophobic to hydrophilic substitution of P1079S may impair the function of IRS1 by inducing a conformational change of its 3D structure (Zeng et al., 2000). However, our results predicted rs148962208 to be a Type II (-) phosSNP that potentially changes the phosphorylation motif pS-P into pS-S and disrupts the phosphorylation of its adjacent site S1078 by mTOR and CDK5. Because this SNP was not detected in other populations (Supplementary Table S7), its effect on IRS1 phosphorylation might be exclusive to the ASN population. The second phosSNP, rs1801276, was mainly identified in the EUR population, such as Danish and French Caucasians (Celi et al., 2000). It was proposed that the A512P SNP alters the secondary structure of IRS1 by disrupting the α-helix formation (Celi et al., 2000), whereas our predictions suggested that A512P inhibits the adjacent phosphorylation status of S503.

In addition, the well-characterized cancer gene EGFR was predicted to be differentially regulated by phosSNPs in different pairs of populations (Table 5). Interestingly, a structural modeling study suggested that rs17290699 (H988P, allele frequency: 0.03 in EUR, Supplementary Table S7) may influence EGFR dimer formation by disrupting the hydrogen and ion bonds between H988 and E690 (Choura et al., 2010). However, our results indicate that the SNP may also induce the phosphorylation of the adjacent S991 site that is potentially modified by MAPKs. Taken together, although in previous studies, the functional effect of nsSNPs has usually been attributed to their influence on protein structure, our results suggest that they also influence inherited diseases as well as cancer susceptibility by reconfiguring phosphorylation networks.

PhosSNPs as well as phosCMs may be involved in driving cancer progression

To investigate whether phosSNPs are involved in driving cancer progression, we systematically analyzed missense cancer mutations (CMs) in 12 types of cancers, including uterine corpus endometrial carcinoma (UCEC-US), colon adenocarcinoma (COAD-US), lung squamous cell carcinoma (LUSC-US), acute myeloid leukemia (LAML-KR), breast cancer (BRCA-US), rectum adenocarcinoma (READ-US), kidney renal clear cell carcinoma (KIRC-US), liver cancer (LINC-JP), brain glioblastoma multiforme (GBM-US), ovarian serous cystadenocarcinoma (OV-US), breast triple negative/lobular cancer (BRCA-UK), and pancreatic cancer (PACA-CA). Using the human proteome

as the background, we observed that known cancer genes (from the Cancer Gene Census) were significantly enriched in missense CM-containing proteins for all cancer types (Figure 5, *p*-value < 0.05). In particular, approximately 400 cancer genes were found to be mutated in the UCEC-US samples, and even in PACA-CA, there were still ~100 mutated cancer genes (Figure 5). Thus, the results supported the hypothesis that missense CMs can be “hotspots” of mutations that are drivers of cancer (Greenman et al., 2007; Haber and Settleman, 2007; Carter et al., 2009). Furthermore, by predicting potential phosCMs from missense CMs, we demonstrated that phosCM-containing proteins significantly enrich cancer genes, although different levels of significance were observed in different cancer types (Figure 5). For example, the *p*-value of the cancer genes over-represented in phosCM-containing proteins is < 1E-20 in UCEC-US, but only <0.05 in LUSC-US, LAML-KR, READ-US, GBM-US, OV-US, BRCA-UK, and PACA-CA (Figure 5). In this regard, our analyses suggest that mutated phosphorylation signaling pathways are involved in cancer progression (Reimand and Bader, 2013) and predicted phosCMs to be novel cancer drivers.

Interestingly, the comparison of population-specific phosSNPs and cancer type-specific phosCMs demonstrated that the *p*-values are more stringent in PhosSNP-containing proteins in the case of enriched cancer genes (Figure 5). It is only in UCEC-US and COAD-US that the number of the enriched cancer genes of phosCMs is greater than phosSNPs (Figure 5). Thus, these statistical analyses suggest that phosSNPs might also be involved in driving cancer progression. Moreover, we re-constructed the kinase－substrate network containing the regulatory kinases as well as the phosSNP- and phosCM-containing proteins for each cancer type based on the predicted KSRs (Figure 6). Clearly, these proteins are tightly connected in the networks, and there are a considerable proportion of substrates containing both phosSNPs and phosCMs (Figure 6). In addition, by comparing phosSNP- and phosCM-containing proteins, we found that many cancer genes contained both genetic variations (Supplementary Figure S3). Taken together, distinguishing phosSNP- and phosCM-containing proteins from the kinase－substrate networks is difficult, and our results suggest that both phosSNPs and phosCMs have a potential role in driving cancer progression.

PhosSNPs were significantly enriched in cancer-related amplification genes and tyrosine kinase circuits

Among various types of GVs, DNA copy number variations (CNVs) were frequently observed in a variety of tumors, and were considered to be involved in tumor evolution by altering the gene expression profile (Albertson, 2006). To investigate the potential role of GVs that “drive” the cancer development in different subtypes, Zaman et al. (2013) integrated a variety of distinct GVs including CNVs and missense mutations, and constructed cell line-specific survival networks in 16 specific breast cancer cell lines. Using the network approach, they revealed that the gene amplification functions as a part of driving regulators to affect different essential genes in different cancer cell lines. In this study, we tried to dissect the relationships between phosSNPs and gene amplifications. Based upon the previous rationales (Zaman et al., 2013), we obtained CNV data and gene expression information for 59 breast cancer tissues from Cancer Cell Line Encyclopedia (CCLE). The Genomic Identification of Significant Targets in Cancer (GISTIC 2.0) was used to calculate the G-scores for each gene (Mermel et al., 2011). As previously described (Mermel et al., 2011; Zaman et al., 2013), we only considered the amplified genes with G-score > 0.3 and expression level among top 50% in each cell line. Finally, we obtained 3409 amplified genes in 59 breast cancer cell lines, with 535 amplified genes containing phosSNPs. Using the human proteome as the background, the hypergeometric distribution based statistical analysis showed that the phosSNPs were significantly enriched in amplified genes of breast cancers (E-ratio = 1.78, *p*-value = 9.54E-45). This result demonstrate that genes with phosSNPs might have a relatively higher probability to be amplified in cancers.

In the phosSNP-associated kinase—substrate network, we observed that ~28.1% (5081 out of 18056 interactions) of KSRs belong to protein tyrosine kinase signaling. Among all TK signaling events, ~73.1% (3716 out of 5081 TK interactions) of KSRs were found to be influenced by phosSNPs, which demonstrated that the TK signaling was more likely to be associated with phosSNPs. Recently, Li et al. (2012a) performed systematic analyses on protein tyrosine signaling network in cancers. The results showed that cancer signaling preferentially employs phosphotyrosine (pTyr) substrates that contain kinase domains or SH2/PTB domains and pTyr sites that were detected in more tumor samples. In this regard, we tried to explore whether phosSNPs could influence the TK signaling in cancers. To address this question, we mapped the phosSNPs into

the dataset of TK circuits from lung cancers (Li et al., 2012a). While a TK circuit contains three parts including pTyr site, tyrosine kinase, and SH2/PTB protein, only TK circuits with phosSNP-influenced KSRs were reserved for further analysis.

Finally, we obtained 344 lung cancer-related TK circuits that were influenced by phosSNPs (Supplementary Table S10). Among the 344 TK circuits, 228 TK circuits were mediated by high frequency cancer (HFC) pTyr sites that were detected in more than one cancer samples. The other 116 TK circuits were linked to low frequency cancer (LFC) pTyr sites that were detected in only one cancer sample (Supplementary Table S10). The statistical analysis revealed that phosSNPs were significantly associated with TK circuits that linked to HFC pTyr sites rather than LFC pTyr sites (E-ratio =2.58, *p*-value = 4.52E-17; Yates' chi-squared test). Furthermore, Li et al. (2012a) have defined two types of pTyr substrates, dual-role substrates (DRSs) and single-role substrates (SRSs), based on the existence of one or multiple kinase domains or SH2/PTB domains in a substrate. The statistical analysis revealed that the HFC pTyr sites are significantly more enriched in DRSs than SRSs in lung cancers (Li et al., 2012a). Here, we also studied whether the phosSNPs are associated with DRSs or SRSs in lung cancers. Among HFC TK circuits, we obtained 88 TK circuits with DRSs and 140 TK circuits with SRSs (Supplementary Table S10). Statistical analysis demonstrated that phosSNPs preferentially occurred in DRS-containing TK circuits (E-ratio = 3.99, *p*-value = 2.16E-24; Yates' chi-squared test). This could be explained that proteins with multiple kinase domains or SH2/PTB domains were commonly involved in signaling pathway, while phosSNPs might alter the phosphorylation states of the substrates, rewire the phosphorylation signaling, and further affect the cancer susceptibility.

Discussion

Reversible phosphorylation plays an essential role in almost all biological processes and pathways. Recently, the rapid progress in phosphoproteomics using high-throughput mass spectrometry (HTP-MS) has enabled the identification of thousands of phosphorylated substrates from one sample (Olsen et al., 2006; Macek et al., 2009; Nilsson, 2012). Thus, an efficient means of retrieving useful information from the flood of data has emerged as a critical goal. In particular, systematic analysis of the phosphoproteomic data in the context of genetic content would advance our understanding of the molecular mechanisms and regulatory processes underlying phosphorylation, including that takes place in individual populations. At the same time, genomic studies with next-generation DNA sequencing (NGS) techniques have characterized an extremely large number of germline SNPs and somatic mutations (Greenman et al., 2007; Manolio et al., 2008; 1000 Genomes Project Consortium, 2010; Forbes et al., 2010). Systematic analysis of the GVs that reconfigure the phosphorylation network will form a link between phosphoproteomics and genomics, and thus will be helpful for understanding how GVs determine both disease susceptibility and therapeutic response in different populations.

Based on the hypothesis that the phosphorylation-associated mutation rate is significantly different to the gene-wide mutation rate, Reimand and Bader (2013) developed a gene-centric algorithm and identified 44 genes with unexpected pSNVs, including 15 known cancer genes, that locate in the kinase domains or flanking regions of phosphorylation sites. However, this statistical model is not applicable to the analysis of phosSNPs, because the mutation rate is difficult to calculate from SNP data sets. Also, SNVs located in the flanking regions of phosphorylation sites may only rarely change the phosphorylation status. For example, we observed that 1049291 nsSNPs located in at least one PSP(15, 15) region. However, only 9606 (~0.9%) of these nsSNPs were predicted to be potential phosSNPs. Thus, the prediction of exact nsSNPs that influence phosphorylation is much more helpful for the purpose of further experimental analysis. Moreover, although it is well known that kinase activity can be changed by GVs located in catalytic domains, it has remained unclear whether GVs alter the substrate specificity of kinases. In contrast, our results suggest that kinase specificity is dramatically changed by phosSNPs in substrates, even without any phosSNPs in the kinases (Figure 3B).

In this work, we developed a SNP-centric approach by directly predicting 9606 potential phosSNPs that change the phosphorylation pattern. To evaluate the prediction accuracy, we manually collected 11 experimentally identified phosSNPs from the scientific literature (Supplementary Table S9). In the absence of any filters, only sequence-based prediction with GPS 2.1 predicted all the phosSNPs as positive hits. However, such a prediction is of limited use because of the high number of false positive predictions. When the experimental phosphorylation information was taken into consideration, seven known phosSNPs were predicted (Supplementary Table S9). When the PPI information was added as well, there were still 4 phosSNPs recalled (Supplementary Table S9), whereas the number of false positive hits was greatly reduced (Table 1). In this regard, although the method is much simpler, without the utilization of any statistical models, the performance is still highly valuable. Unexpectedly, known disease-associated SNPs were found to be significantly present in these results, with a 2.18-fold enrichment (Supplementary Table S6). Based on the predicted phosSNPs, we also constructed a human phosSNP-associated kinase—substrate network and showed that cancer genes and drug targets were both over-represented. Further analysis demonstrated that the proteins in the network are highly involved in a number of signaling and cancer pathways. For example, there are 79 genes in the cancer pathway of chronic myeloid leukemia (KEGG ID: hsa05220), with 56 of them included in the phosSNP-associated network (Figure 7). Thus, the cancer pathways can be dramatically altered by phosSNPs. In particular, we observed that a number of disease- or cancer-associated genes were differentially regulated by phosSNPs in different populations. Thus, phosSNPs may influence inherited disease or cancer susceptibility by reconfiguring the phosphorylation networks in different populations.

Recently, the identification of potential “driver” mutations from the cancer genome sequencing data has emerged as an important topic. It is believed that a considerable proportion of somatic mutations are able to contribute to cancer progression by acting as “drivers” (Greenman et al., 2007; Haber and Settleman, 2007; Carter et al., 2009). Because cancers are complex diseases and attributed to multiple genes and/or somatic mutations, recent analyses have been focused on identifying “mutated driver pathways” rather than single genes or mutations (Wood et al., 2007; Vandin et al., 2012), e.g. functional mutations in phosphorylation signaling pathways (Reimand and Bader, 2013). In these studies, all of the known SNPs were removed and the remaining data were regarded as somatic mutations for the further detection of drivers. A hidden assumption underlying

these studies is that most SNPs are non-functional and do not contribute to cancer development. Indeed, only a small proportion of nsSNPs (0.91%) were predicted as phosSNPs in this study. However, the result is still a comparatively large number versus somatic mutations, because it was estimated that there are <15 driver mutations in an individual tumor (Wood et al., 2007). In particular, our results demonstrated that both phosSNPs and phosCMs preferentially occur in cancer genes, whereas known cancer genes were more significantly enriched in phosSNP-containing proteins than phosCM-associated proteins in most types of cancer (Figure 5). In addition, a considerable number of substrates, especially cancer genes, contain both phosSNPs and phosCMs (Supplementary Figure S3) and cannot be distinguished from the kinase—substrate networks (Figure 6). In this regard, it is possible that phosSNPs may influence cancer susceptibility by driving tumor progression together with somatic mutations. Further analyses showed that phosSNPs were significantly enriched in amplified genes of breast cancers, while the phosSNPs in tyrosine signaling of lung cancer were found to be significantly associated with TK circuits that linked to HFC pTyr sites and DRS-containing TK circuits.

Taken together, the results presented provide a systematic analysis of nsSNPs that influence protein phosphorylation and show that phosSNPs are significantly implicated in inherited diseases and cancers by dramatically reconfiguring the kinase—substrate phosphorylation network. These results not only highlight the potential roles of phosSNPs in driving cancer progression, but also provide a useful resource for further experimental consideration.

Experimental procedures

The SNP data sets

The human SNPs with map summaries were downloaded from the NCBI ftp server (dbSNP Build 141, RefSNP docsums in ASN.1 flat format) on September 17, 2014 (Sherry et al., 2001). In total there were 62386627 SNPs, including 1705956 cSNPs with 1060070 missense nsSNPs. In this work, we only considered nsSNPs that induce changes in amino acid residues. The nsSNPs in other forms of mutation, such as frameshift and stop-gained variations, were discarded.

To search for phosSNP—disease relations, we first downloaded 153839 disease-associated SNPs from the GWASdb (Li et al., 2012b). Also, the SNPs with annotated disease information were obtained from the ClinVar dataset (November 14, 2014) (Landrumet al., 2014) in the NCBI ftp server, including 260112 entries with 36371 unique missense nsSNPs.

For the population analysis, the cSNPs in the 1000 Genomes Project were downloaded on June 1, 2013 (Phase I Release v3) (1000 Genomes Project Consortium, 2010). In total, we obtained 325159 cSNPs along with reference allele frequencies from four ancestry-based super population groups (AMR, Ad Mixed American; ASN, East Asian; AFR, African; EUR, European).

The sequence data sets

We downloaded 72813 human protein sequences from the RefSeq database (NCBI Homo sapiens Annotation Release 66) on September 25, 2014 (Pruitt et al., 2007). These sequences were regarded as the benchmark sequence data set for identifying phosSNPs. Also, we obtained 101075 human protein sequences from the Ensembl database (Version 69, November 26, 2013) (Flicek et al., 2013) for the purpose of identifying phosCMs. As previously described (Ren et al., 2010), the redundancy in the RefSeq and Ensembl proteins was not cleared.

In a previous study (Song et al., 2012), we prepared a non-redundant sequence data set from the UniProt database (on April 6, 2010) (UniProt Consortium, 2013) with 81733 unique proteins in *Homo sapiens*. For the network analysis, we mapped all of the Refseq proteins to UniProt sequences by the BLAST program with a stringent threshold (E-value ≤ 1E-30, Identities ≥ 70%) (Johnson et al., 2008). For each RefSeq protein, the best hit was preserved when it met the threshold. In total, these RefSeq proteins were mapped to 28315 UniProt sequences. Analogously,

the Ensembl proteins were also mapped to UniProt sequences for the network analysis.

The data sets on the phosphorylation sites and PPIs

Previously, we collected 145646 experimentally identified phosphorylation sites in 28457 substrates for five eukaryotic species, whereas in this work, 60816 phosphorylation sites of 10253 human proteins were collected (Song et al., 2012). Moreover, we integrated 59481 experimentally identified PPIs (Exp. PPIs) among 12221 human proteins, and obtained 1212607 pre-calculated PPIs (STRING PPIs) for 16523 human proteins from the STRING database (Jensen et al., 2009; Song et al., 2012). The sequences of phosphorylated and interacting proteins were prepared in the UniProt FASTA format (Jensen et al., 2009; UniProt Consortium, 2013).

The data sets on somatic cancer mutations

The somatic CMs were downloaded from the data repository of the International Cancer Genome Consortium (ICGC) (http://dcc.icgc.org, Data Release 14, September 26, 2013) (Hudson et al., 2010). Somatic CMs of 26 types of cancers sequenced from the different populations were obtained. First, the missense CMs were extracted out and directly mapped to their corresponding protein sequences from Ensembl (Flicek et al., 2013). Because the cancer genome sequencing was still not completed, several types of cancers only contained very limited number of sequenced somatic CMs. Thus, ultimately, only 12 types of cancers with more than 10000 missense CMs were selected for further analyses. Because the corresponding SNPs were not available for these cancer samples, we selected population-specific SNPs for each cancer to compare phosSNPs and phosCMs in the different populations. We chose the AMR for UCEC-US, COAD-US, LUSC-US, BRCA-US, READ-US, KIRC-US, GBM-US, OV-US, and PACA-CA, the ASN for LAML-KR and LINC-JP, and the EUR for BRCA-UK.

Computational identification of phosSNPs and phosCMs

Based on the map summaries of human SNPs, we first picked out all RefSeq proteins with at least one nsSNP. Then we generated nsSNP-containing proteins (mutated proteins) from the benchmark sequences. As previously described (Ren et al., 2010), each of the mutated sequences contains only one nsSNP, whereas the combined effects of multiple nsSNPs were not analyzed in

this study.

GPS 2.1 (Xue et al., 2011) was chosen to predict ssKSRs for the original and mutated proteins, respectively. Because GPS only predicts kinase-specific sites at the PK cluster level, we manually formed links of 407 human kinases with their corresponding predictors in GPS 2.1 (Song et al., 2012). Then the exact kinases of the predicted phosphorylation sites were characterized. Furthermore, the known phosphorylation sites and PPIs were adopted as two filters to remove potentially false positive hits for the prediction of ssKSRs. By comparison of the prediction results for the original and mutated human proteins, the phosSNPs were identified.

Previously, we classified all of the predicted phosSNPs into five categories (Ren et al., 2010) as follows. (i) Type I: create (+) or remove (-) a phosphorylation site at a phosphorylatable position; (ii) Type II: create (+) or disrupt (-) one or multiple adjacent phosphorylation sites; (iii) Type III: change the kinase type for one or multiple adjacent phosphorylation sites; (iv) Type IV: induce a change in kinase type at a phosphorylatable position; (v) Type V: generate a stop codon to disrupt following phosphorylation sites. Because Type V phosSNPs occupied only ~0.7% of the total results (Ren et al., 2010) and truncated transcripts heavily influence protein function beyond phosphorylation, this type was not calculated or included thereafter. Using the same procedure, we also identified and classified potential phosCMs for each cancer type. The predicted datasets for phosSNPs and phosCMs were available at http://phossnp.biocuckoo.org/dataset.php.

Network construction and analysis

To avoid redundancy, all phosSNP- or phosCM-containing proteins were mapped to UniProt sequences. Because a kinase is able to phosphorylate a substrate at multiple sites, there can be multiple ssKSRs between the kinase and substrate. For the network construction, we only considered the KSR, and multiple ssKSRs between a kinase and a substrate were counted as a single KSR. Based on the predicted ssKSRs for the original and mutated proteins, a human phosSNP- or phosCM-associated kinase—substrate phosphorylation network was re-constructed and visualized by Cytoscape 2.8.3 (Shannon et al., 2003). In the network, the nodes are kinases or substrates, whereas the edges are KSRs. As previously described (Song et al., 2012), the network is directional and the three types of orientations were defined as Kinase -> Substrate (a kinase phosphorylates a substrate), Kinase *A* -> Kinase *B* (kinase *A* phosphorylates kinase *B*), and Kinase

A <-> Kinase *B* (kinase *A* phosphorylates kinase *B* and *vice versa*). Also, we used MCODE v1.32 (Bader and Hogue, 2003), a plugin of Cytoscape, to detect molecular complexes in the phosSNP-associated network. To analyze the impact of phosSNPs on the network topology, Cytoscape plugin cytoHubba v1.6 (Lin et al., 2008) was employed to separately calculate the topological features in original and mutated networks, including degree, maximal clique centrality (MCC), maximum neighborhood component (MNC), bottleneck (BN), and betweenness. As previously described (Jin et al., 2007; Wang et al., 2007; Zaman et al., 2013), we picked out top 200 genes (~10%) with most degrees as hub genes in original and mutated networks, respectively. Then the above five topological features for hub genes were also calculated, while non-parametric Wilcoxon signed-rank test was performed to statistically compare the topological changes of degree, MCC, MNC, BN, and betweenness (p-value < 0.01). Based on the annotation information of the 1000 Genomes Project, we also constructed and analyzed population-specific phosSNP-associated networks.

Identification of gene amplifications in breast cancer cell lines

Previously, the procedure for identifying gene amplification was described (Zaman et al., 2013). Following this method, the pre-segmented CNV dataset identified by Affymetrix SNP6.0 arrays and gene expression information for 59 breast cancer cell lines were downloaded from CCLE (http://www.broadinstitute.org/ccle/home, published on September 29, 2012). The corresponding marker file for Affymetrix SNP6.0 data, which was annotated with human reference genome 31, was downloaded from National Cancer Institute (NCI, https://wiki.nci.nih.gov). The gene amplification data processing software GISTIC 2.0 was downloaded from Cancer Cell Line Encyclopedia (Mermel et al., 2011).

The copy number variation dataset and the marker file were imported into GISTIC 2.0 to calculate the G-scores with the default parameters (Mermel et al., 2011). As previously described (Zaman et al., 2013), we identified amplified genes with the threshold of a G-score > 0.3 and expression level among top 50% in each cell line. The GISTIC results were further mapped to USCS human genome to obtain the gene names and UniProt accession numbers. Finally, we totally identified 3409 amplification genes in 59 breast cancer cell lines.

The statistical analysis of enrichment

From the Cancer Gene Census (Futreal et al., 2004), in total we obtained 474 well documented cancer genes. We also obtained 555 cancer genes as a secondary data set (Reimand's data) from a recently published analysis (Reimand and Bader, 2013). Only 442 and 467 cancer genes in the two data sets were successfully mapped to UniProt proteins, respectively. Furthermore, we downloaded the protein sequences of 4906 drug targets from the DrugBank database (Wishart et al., 2008), and 1919 of them were mapped to UniProt sequences. The proteins in the phosSNP-associated network (network proteins) were mapped to the three data sets for the purpose of statistical enrichment analyses, with a hypergeometric distribution (Liu et al., 2013).

Moreover, we retrieved the GO annotations of 28315 UniProt proteins (UniProt, 2013). A total of 18911 UniProt sequences, including 2375 network proteins, were annotated with at least one GO term. The KEGG was used to map UniProt proteins to the biological pathways (Kanehisa et al., 2004). There were 6824 UniProt proteins, including 1355 network proteins annotated with at least one KEGG pathway entry. In addition, because the FunDO (Functional Disease Ontology), an online service for exploring gene—disease relations, only adopts a Gene ID as the input (Schriml et al., 2012), we used the ID Mapping tool in UniProt database (UniProt, 2013) and obtained Gene IDs for 2544 network proteins. We directly submitted the Gene IDs to FunDO for the statistical enrichment analysis of the DO items (p-value < 1E-10) (Schriml et al., 2012).

Acknowledgements

The authors thank Dr David L. Armstrong (NIEHS/NIH) for the personal communications on phosphorylopathy or phosSNP. We thank Drs Francesca Diella and Toby J. Gibson (EMBL) for always providing the newly updated data set of Phospho.ELM database during the past ten years. We thank Dr Peter Hornbeck (Cell Signaling Technology, USA) for providing the PhosphoSitePlus data set on July 14, 2009. We thank Dr Edwin Wang and Dr Naif Zaman for providing the procedures on the identification of gene amplifications and the data set of human phosphotyrosine signaling. We also thank Drs Ruibin Xi and Ge Gao (Peking University.), and Drs Qi Liu and Xingming Zhao (Tongji University) for their helpful comments on the usage of GISTIC. We thank Meng Gao (NCU), Shuzhen Kuang, Drs Anyuan Guo and Xiaoping Miao (HUST) for their helpful comments on the network and SNP analysis.

Funding

This work was supported by grants from the National Basic Research Program (973 project) (2013CB933900, 2012CB910101, and 2012CB911201), the National Natural Science Foundation of China (31171263, 81272578, and 31071154), the International Science and Technology Cooperation Program of China (2014DFB30020), and China Postdoctoral Science Foundation (2014M550392).

References

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Albertson, D.G. (2006). Gene amplification in cancer. Trends Genet. *22*, 447-455.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics *4*, 2.

Cargill, M., Altshuler, D., Ireland, J., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. *22*, 231-238.

Carter, H., Chen, S., Isik, L., et al. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. *69*, 6660-6667.

Celi, F.S., Negri, C., Tanner, K., et al. (2000). Molecular scanning for mutations in the insulin receptor substrate-1 (IRS-1) gene in Mexican Americans with Type 2 diabetes mellitus. Diabetes Metab. Res. Rev. *16*, 370-377.

Chang, Q., Li, Y., White, M.F., et al. (2002). Constitutive activation of insulin receptor substrate 1 is a frequent event in human tumors: therapeutic implications. Cancer Res. *62*, 6035-6038.

Choura, M., Frikha, F., Kharrat, N., et al. (2010). Investigating the function of three non-synonymous SNPs in EGFR gene: structural modelling and association with breast cancer. Protein J. *29*, 50-54.

Collins, F.S., Brooks, L.D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. *8*, 1229-1231.

Courtois, G., Smahi, A., Reichenbach, J., et al. (2003). A hypermorphic IκBα mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. J. Clin. Invest. *112*, 1108-1115.

Erxleben, C., Liao, Y., Gentile, S., et al. (2006). Cyclosporin and Timothy syndrome increase mode 2 gating of CaV1.2 calcium channels through aberrant phosphorylation of S6 helices. Proc. Natl Acad. Sci. USA *103*, 3932-3937.

Flicek, P., Ahmed, I., Amode, M.R., et al. (2013). Ensembl 2013. Nucleic Acids Res. *41*, D48-55.

Forbes, S.A., Tang, G., Bindal, N., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. *38*, D652-657.

Futreal, P.A., Coin, L., Marshall, M., et al. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177-183.

Gentile, S., Martin, N., Scappini, E., et al. (2008). The human ERG1 channel polymorphism, K897T, creates a phosphorylation site that inhibits channel activity. Proc. Natl Acad. Sci. USA *105*, 14704-14708.

Gonzaga-Jauregui, C., Lupski, J.R., and Gibbs, R.A. (2012). Human genome sequencing in health and disease. Annu. Rev. Med. *63*, 35-61.

Greenman, C., Stephens, P., Smith, R., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153-158.

Gunther, T., Schmitt, A.O., Bortfeldt, R.H., et al. (2011). Where in the genome are significant single nucleotide polymorphisms from genome-wide association studies located? OMICS *15*, 507-512.

Haber, D.A., and Settleman, J. (2007). Cancer: drivers and passengers. Nature *446*, 145-146.

Hochgrafe, F., Zhang, L., O'Toole, S.A., et al. (2010). Tyrosine phosphorylation profiling reveals the signaling network characteristics of Basal breast cancer cells. Cancer Res. *70*, 9391-9401.

Hudson, T.J., Anderson, W., Artez, A., et al. (2010). International network of cancer genome projects. Nature *464*, 993-998.

Jensen, L.J., Kuhn, M., Stark, M., et al. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. *37*, D412-416.

Jin, G., Zhang, S., Zhang, X.S., et al. (2007). Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. PLoS One *2*, e1207.

Johnson, M., Zaretskaya, I., Raytselis, Y., et al. (2008). NCBI BLAST: a better web interface. Nucleic Acids Res. *36*, W5-9.

Kanehisa, M., Goto, S., Kawashima, S., et al. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res. *32*, D277-280.

Kettunen, J., Tukiainen, T., Sarin, A.P., et al. (2012). Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat. Genet. *44*, 269-276.

Landrum, M.J., Lee, J.M., Riley, G.R., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 42, D980-985.

Li, L., Tibiche, C., Fu, C., et al. (2012). The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. Genome Res. *22*, 1222-1230.

Li, M.J., Wang, P., Liu, X., et al. (2012). GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. *40*, D1047-1054.

Li, S., Iakoucheva, L.M., Mooney, S.D., et al. (2010). Loss of post-translational modification sites in disease. Pac. Symp. Biocomput. *15*, 337-347.

Lin, C.Y., Chin, C.H., Wu, H.H., et al. (2008). Hubba: hub objects analyzer--a framework of interactome hubs identification for network biology. Nucleic Acids Res. 36, W438-443.

Liu, P., Gan, W., Inuzuka, H., et al. (2013). Sin1 phosphorylation impairs mTORC2 complex integrity and inhibits downstream Akt signalling to suppress tumorigenesis. Nat. Cell Biol. 15, 1340-1350.

Liu, Z., Ren, J., Cao, J., et al. (2013). Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes. Brief. Bioinform. *14*, 344-360.

Macek, B., Mann, M., and Olsen, J.V. (2009). Global and site-specific quantitative phosphoproteomics: principles and applications. Annu. Rev. Pharmacol. Toxicol. *49*, 199-221.

Manning, B.D., and Cantley, L.C. (2007). AKT/PKB signaling: navigating downstream. Cell *129*, 1261-1274.

Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. *118*, 1590-1605.

Mermel, C.H., Schumacher, S.E., Hill, B., et al. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Nilsson, C.L. (2012). Advances in quantitative phosphoproteomics. Anal. Chem. *84*, 735-746.

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

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61-65.

Radivojac, P., Baenziger, P.H., Kann, M.G., et al. (2008). Gain and loss of phosphorylation sites in human cancer. Bioinformatics *24*, i241-247.

Reimand, J., and Bader, G.D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol. *9*, 637.

Ren, J., Jiang, C., Gao, X., et al. (2010). PhosSNP for systematic analysis of genetic polymorphisms that influence protein phosphorylation. Mol. Cell. Proteomics *9*, 623-634.

Reumers, J., Maurer-Stroh, S., Schymkowitz, J., et al. (2006). SNPeffect v2.0: a new step in investigating the

molecular phenotypic effects of human non-synonymous SNPs. Bioinformatics *22*, 2183-2185.

Ryu, G.M., Song, P., Kim, K.W., et al. (2009). Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases. Nucleic Acids Res. *37*, 1297-1307.

Savas, S., and Ozcelik, H. (2005). Phosphorylation states of cell cycle and DNA repair proteins can be altered by the nsSNPs. BMC Cancer *5*, 107.

Schriml, L.M., Arze, C., Nadendla, S., et al. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. *40*, D940-946.

Shannon, P., Markiel, A., Ozier, O., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498-2504.

Sherry, S.T., Ward, M.H., Kholodov, M., et al. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308-311.

Song, C., Ye, M., Liu, Z., et al. (2012). Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Mol. Cell. Proteomics *11*, 1070-1083.

UniProt Consortium. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. *41*, D43-47.

Vandin, F., Upfal, E., and Raphael, B.J. (2012). De novo discovery of mutated driver pathways in cancer. Genome Res. *22*, 375-385.

Wang, E., Lenferink, A., and O'Connor-McCourt, M. (2007). Cancer systems biology: exploring cancer-associated genes on cellular networks. Cell. Mol. Life Sci. *64*, 1752-1762.

Wang, M.H., Chang, J., Yu, D.K., et al. (2013). A missense SNP in the codon of ADD1 phosphorylation site associated with non-cardia gastric cancer susceptibility in a Chinese population. Zhonghua Zhong Liu Za Zhi *35*, 311-314.

Wishart, D.S., Knox, C., Guo, A.C., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. *36*, D901-906.

Wood, L.D., Parsons, D.W., Jones, S., et al. (2007). The genomic landscapes of human breast and colorectal cancers. Science *318*, 1108-1113.

Xue, Y., Liu, Z., Cao, J., et al. (2011). GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Eng. Des. Sel. *24*, 255-260.

Xue, Y., Ren, J., Gao, X., et al. (2008). GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol. Cell. Proteomics *7*, 1598-1608.

Yang, C.Y., Chang, C.H., Yu, Y.L., et al. (2008a). PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. Bioinformatics *24*, i14-20.

Yang, Y., Houle, A.M., Letendre, J., et al. (2008b). RET Gly691Ser mutation is associated with primary vesicoureteral reflux in the French-Canadian population from Quebec. Hum. Mutat. *29*, 695-702.

Yue, P., and Moult, J. (2006). Identification and analysis of deleterious human SNPs. J. Mol. Biol. *356*, 1263-1274.

Zaman, N., Li, L., Jaramillo, M.L., et al. (2013). Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. Cell Reports *5*, 216-223.

Zeng, W., Peng, J., Wan, X., et al. (2000). The mutation of insulin receptor substrate-1 gene in Chinese patients with non-insulin-dependent diabetes mellitus. Chin. Med. J. (Engl.) *113*, 80-83.

Figure legends

Figure 1 The procedure for the computational characterization of phosSNPs. With GPS 2.1 (Xue et al., 2011), we first predicted the ssKSRs for the original and nsSNP-containing proteins, respectively. The known phosphorylation sites and PPIs between kinases and substrates were adopted as two filters to reduce false positive predictions. By comparison of the results for the original and mutated human proteins, in total we identified 9606 phosSNPs in 7046 proteins with a high degree of confidence.

Figure 2 The human phosSNP-associated kinase—substrate phosphorylation network. Multiple ssKSRs between a kinase and a substrate were only counted as a single KSR, which was then classified into one of four types: added (+), removed (-), changed (C), and unchanged (N). In total, the network contains 18056 KSRs among 374 kinases and 2270 substrates, including 154 kinases (~41.2%) without any phosSNPs.

Figure 3 Cancer genes and drug targets are highly enriched in the human phosSNP-associated kinase—substrate network. (**A**) Top 10 substrates with the most KSRs. (**B**) Top 10 kinases with the most KSRs. (**C**) Statistical results suggested that 7.5%—8.3% and ~20% of genes in the network are cancer genes and drug targets, respectively. (**D**) Enrichment analysis of cancer genes and drug targets in the phosSNP-associated kinase—substrate network compared with the human PPN.

Figure 4 Examples of four types of phosSNPs. (**A**) Type I (+): the G691S (rs1799939) of RET generates a new phosphorylation site for PKCA. (**B**) Type I (-): the S32I (rs28933100) of NFKBIA removes a IKBKB phosphorylation site. (**C**) Type II (+): the L172P (rs113133553) of BCL10 creates a MAPK12 site at S171. (**D**) Type II (-): the A358V (rs35332062) of MLXIPL disrupts a PKACA site at S361. (**E**) Type III: the K2162R (rs142711774) of SPTBN1 induces a change in kinase type from CK2A1 to PRKCI at S2160. (**F**) Type IV: the S752Y of EGFR induces a change in kinase type from MOK to FAK2 at the same site. The phosSNPs in RET and NFKBIA were validated in previous experiments (Supplementary Table S9).

Figure 5 A histogram of the statistical results for cancer genes in somatic mutations, phosCMs, and phosSNPs from 12 types of cancers. The background is the human proteome, which contains 24080 proteins. The Cancer Gene Census was used to perform the mapping. The hypergeometric distribution was performed to calculate the significance of the cancer genes. For each cancer type, the left bar represents the somatic mutation, the middle bar represents the phosCMs, and the right bar represents the phosSNP. The Y axis of the histogram indicates the number of cancer genes in three variations. The colors of the bar represent the E-ratio. The star (*) number represents the degree of the *p*-value.

Figure 6 The phosSNP- and phosCM-associated kinase—substrate networks in different cancer types. Each network contains regulatory kinases (green), phosSNP- (yellow) or phosCM-containing (pink) proteins, and proteins with both phosSNPs and phosCMs (cyan).

Figure 7 An example of the crosstalk of the phosSNP-associated kinase—substrate network and cancer pathways. The cancer pathway for chronic myeloid leukemia (KEGG ID: hsa05220) contains 79 genes, and 56 of them are also included in the phosSNP-associated kinase—substrate network.

Tables

Table 1 The statistical analysis of the potential phosSNPs detected by different approaches.

PhosSNP	GPS 2.1 only[a]			Known Phos[b]			Exp. & STRING PPI[c]			Exp. PPI[d]		
	Pro.[e]	SNP[f]	Pro.	SNP	Pro.	SNP	Pro.	SNP	Pro.	SNP	Pro.	SNP
Type I (+)	61,609	93,076	15.35%	207	58	0.14%	91	27	0.28%	73	16	0.64%
Type I (-)	59,158	79,450	13.10%	9,321	3,262	8.09%	2,380	1,102	11.47%	898	388	15.54%
Type II (+)	62,133	122,682	20.23%	7,317	2,793	6.92%	2,913	1,497	15.58%	982	425	17.03%
Type II (-)	63,457	155,981	25.73%	11,388	5,022	12.45%	4,858	3,149	32.78%	1,766	1,169	46.83%
Type III	67,012	389,888	64.30%	26,257	33,344	82.65%	4,426	4,759	49.54%	793	650	26.04%
Type IV	17,489	6,293	1.04%	1,734	337	0.84%	198	58	0.60%	13	6	0.24%
Total	67,789	606,321		29,032	40,344		7,946	9,606		2,628	2,496	

[a] Only GPS 2.1 was used for the prediction;

[b] The known phosphorylation information was added as a filter to remove false positive hits;

[c] Both experimental and STRING PPIs were used;

[d] Only experimentally identified PPIs were used;

[e] Pro., the number of protein sequences;

[f] SNP, the number of phosSNPs.

To balance the false positive predictions and total predicted hits, the filter of known phosphorylation plus Exp. & STRING PPI information was adopted.

Table 2 The statistical analysis of topological features of original and mutated proteins in each network.

Topological features	Network[a]	Hub[b]
Degree	3.12E-102	1.90E-15
MCC	3.70E-22	4.29E-04
MNC	7.39E-49	2.42E-09
BottleNeck	1.97E-90	2.56E-08
Betweenness	2.03E-156	4.93E-40

The features, including degree, maximal clique centrality (MCC), maximum neighborhood component (MNC), bottleneck (BN), and betweenness, were calculated with CytoHubb v1.6 plugin in Cytoscape. Wilcoxon signed-rank test was used to calculate the significance.

[a] The statistical analysis for all proteins in original and mutated networks;

[b] The statistical analysis for hub proteins in original and mutated networks. The top 200 proteins with most KSRs were considered as hubs.

Table 3 The enrichment analysis of KEGG pathways for proteins in the phosSNP-associated kinase−substrate network.

KEGG ID	Description	Network[a]		Proteome		E-ratio[d]	*p*-value
		Num.[b]	Per.[c]	Num.	Per.		
The most over-represent pathway							
hsa04722	Neurotrophin signaling pathway	90	6.64%	136	1.99%	3.33	1.94E-32
hsa04660	T cell receptor signaling pathway	78	5.76%	124	1.82%	3.17	4.85E-26
hsa04012	ErbB signaling pathway	70	5.17%	105	1.54%	3.36	9.94E-26
hsa04910	Insulin signaling pathway	85	6.27%	148	2.17%	2.89	2.40E-24
hsa05200	Pathways in cancer	159	11.73%	388	5.69%	2.06	5.69E-23
hsa05220	Chronic myeloid leukemia	56	4.13%	79	1.16%	3.57	7.56E-23
hsa04010	MAPK signaling pathway	145	10.70%	342	5.01%	2.14	1.09E-22
hsa04510	Focal adhesion	104	7.68%	224	3.28%	2.34	5.80E-20
hsa04380	Osteoclast differentiation	79	5.83%	155	2.27%	2.57	2.10E-18
hsa05166	HTLV-I infection	121	8.93%	290	4.25%	2.10	2.38E-18
hsa05214	Glioma	47	3.47%	69	1.01%	3.43	3.07E-18
hsa04666	Fc gamma R-mediated phagocytosis	60	4.43%	104	1.52%	2.91	1.22E-17
hsa05162	Measles	72	5.31%	139	2.04%	2.61	2.25E-17
hsa05215	Prostate cancer	60	4.43%	105	1.54%	2.88	2.30E-17
hsa04914	Progesterone-mediated oocyte maturation	55	4.06%	92	1.35%	3.01	3.06E-17
hsa04520	Adherens junction	56	4.13%	96	1.41%	2.94	7.55E-17
hsa04110	Cell cycle	71	5.24%	139	2.04%	2.57	1.01E-16
hsa04664	Fc epsilon RI signaling pathway	51	3.76%	85	1.25%	3.02	3.60E-16
hsa05169	Epstein-Barr virus infection	94	6.94%	215	3.15%	2.20	5.14E-16
hsa04662	B cell receptor signaling pathway	51	3.76%	89	1.30%	2.89	5.15E-15
hsa04912	GnRH signaling pathway	58	4.28%	109	1.60%	2.68	6.46E-15
The most under-represent KEGG Pathway							
hsa01100	Metabolic pathways	113	8.34%	1265	18.54%	0.45	3.34E-31
hsa04740	Olfactory transduction	14	1.03%	393	5.76%	0.18	1.68E-22

The hypergeometric distribution was performed. *p*-value < 1E-14.

[a] Proteins in the network;

[b] The number of proteins annotated with the KEGG ID;

[c] The proportion of proteins annotated with the KEGG ID;

[d] E-ratio, the enrichment ratio as the proportion of proteins in the network divided by that in the proteome.

Table 4 Ten predicted phosSNPs in 10 genes are significantly associated with human diseases.

Gene	SNP	Disease	*p*-value
PPP1R12B	rs3881953	Amyotrophic Lateral Sclerosis (ALS)	0.000553
ERBB2	rs1058808	Asthma	8.33E-07
ERBB2	rs1058808	Breast Neoplasms	0
HNF4A	rs1800961	Coronary heart disease	8.00E-10
HNF4A	rs1800961	plasma HDL cholesterol (HDL-C) levels	0
HNF4A	rs1800961	HDL cholesterol	0
HNF4A	rs1800961	HDL cholesterol	8.00E-10
CFTR	rs74571530	Congenital Bilateral Absence of the Vas Deferens	0
ERG1	rs1805123	Acquired Long QT Syndrome (aLQTS); Long QT Syndrome	0
HNF4A	rs1800961	Triglycerides	1.99E-08
HNF4A	rs1800961	Triglycerides	4.19E-05
HNF4A	rs1800961	Triglycerides	6.65E-05
ICAM3	rs2230399	Soluble ICAM-1	4.40E-08
MLXIPL	rs35332062	Metabolite levels	1.18E-12
MLXIPL	rs35332062	Metabolite levels	5.92E-11
DDX20	rs197414	Obesity (extreme)	0.0009447
IL4R	rs1801275	IgE levels	1.00E-07
IL4R	rs1801275	IgE levels	1.54E-06
CCND3	rs1051130	Other erythrocyte phenotypes	1.70E-09
CCND3	rs1051130	Other erythrocyte phenotypes	2.55E-09
HNF4A	rs1800961	C-reactive protein	2.00E-09

Prediction results were mapped to different GWA studies (*p*-value < 0.001). The gene names along with the corresponding SNPs, diseases, and *p*-values were directly taken from the GWASdb (Li et al., 2012b).

Table 5 The statistical data of four population-specific phosSNP-associated kinase－substrate networks.

Population network	AMR	ASN	AFR	EUR
Kinase	193	193	202	198
Substrates	344	265	392	350
Kinase without phosSNP[a]	95	93	100	105
Added	291 (15%)	203 (13.3%)	417 (17.6%)	299 (14.7%)
Removed	717 (36.9%)	611 (40.1%)	790 (33.4%)	765 (37.6%)
Changed	209 (10.8%)	138 (9%)	335 (14.1%)	196 (9.6%)
Unchanged	725 (37.3%)	573 (37.6%)	826 (34.9%)	772 (38%)
Total KSRs	1942	1525	2368	2032
MCC	3.30E-48	1.29E-64	1.10E-35	1.13E-53
Degree	5.44E-64	1.68E-68	2.38E-46	4.69E-72

Four population-specific phosSNP-associated kinase－substrate networks were constructed for AMR, ASN, AFR, and EUR, respectively. The distributions of MCC and degree for the original and mutated proteins in each network were statistically analyzed by the Wilcoxon signed-rank test.

[a] Kinases without any predicted phosSNPs in the network.

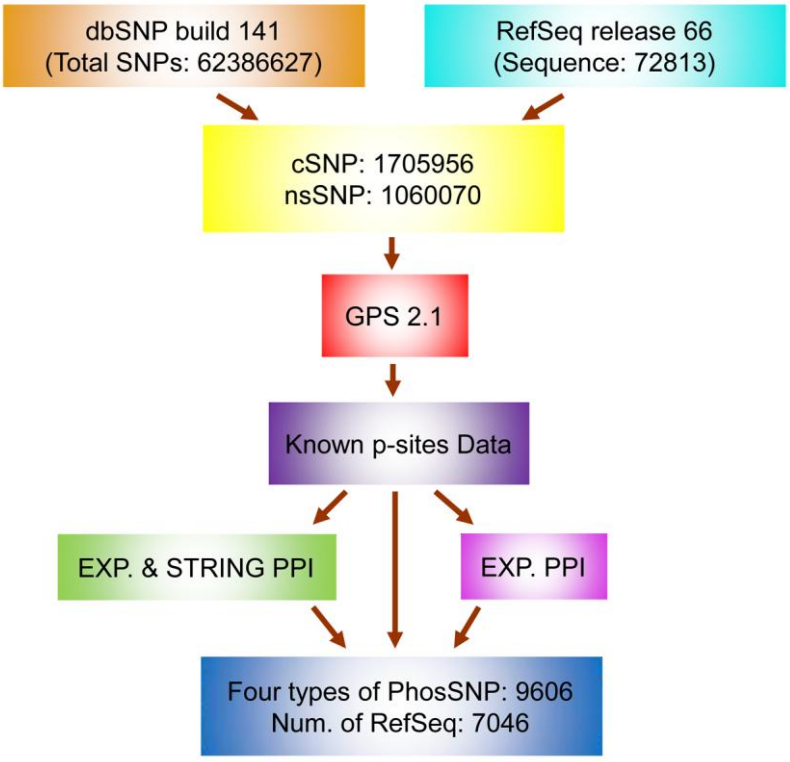
Table 6 The genes in the population-specific phosSNP-associated kinase－substrate networks with significantly different KSRs between two given populations.

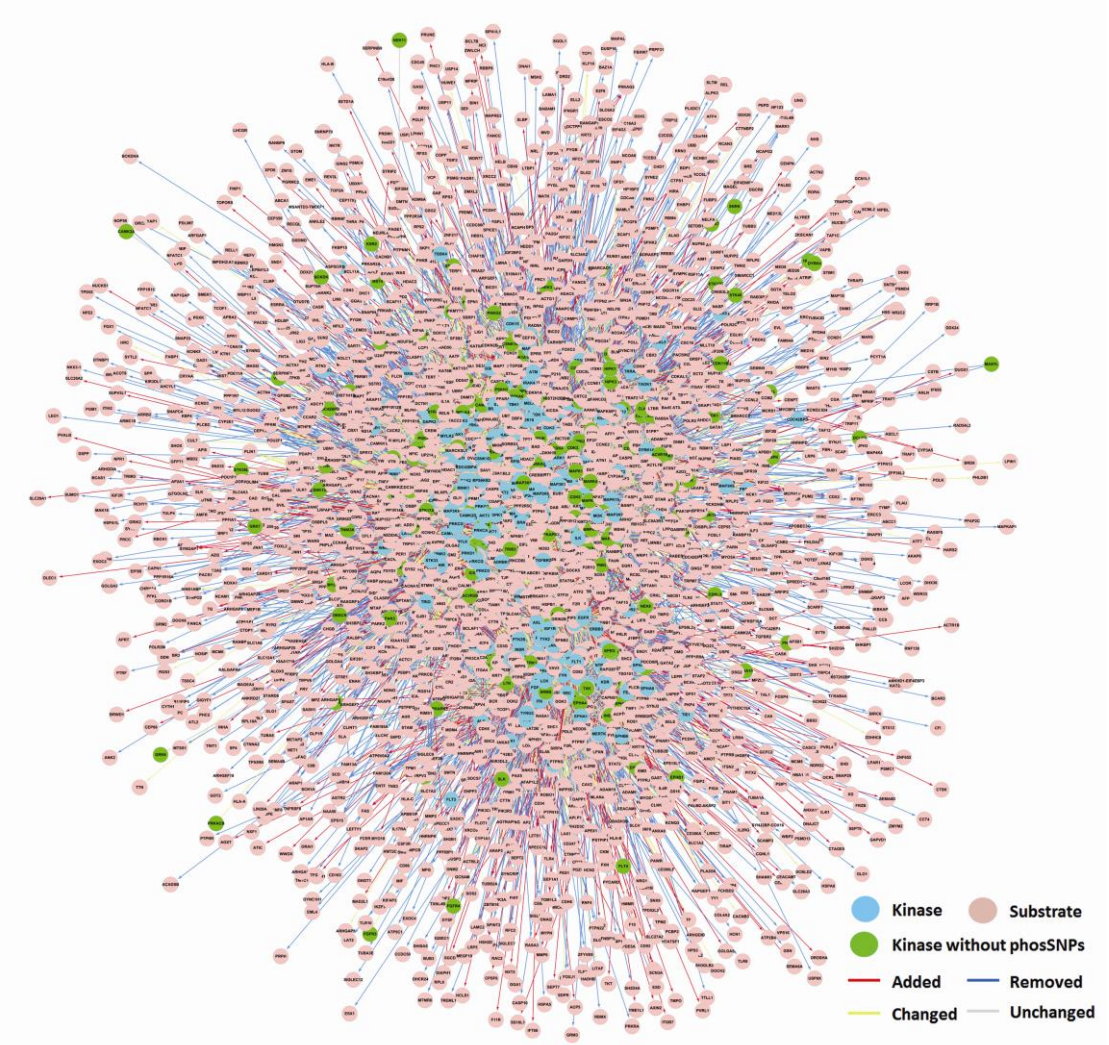
Gene	UniProt	A-B[a]	A population		B population		E-ratio	χ^{2b}	*p*-value
			Num.	Per.	Num.	Per.			
IRS1	P35568	AFR-EUR	77	3.25%	3	0.15%	22.02	57.30	3.74E-14
PXN	Q59GS5	AMR-AFR	39	2.01%	2	0.08%	23.78	39.89	2.68E-10
IRS1	P35568	AMR-EUR	40	2.06%	3	0.15%	13.95	32.16	1.42E-08
RAF1	P04049	AMR-EUR	36	1.85%	2	0.10%	18.83	30.48	3.37E-08
IRS1	P35568	AFR-ASN	77	3.25%	9	0.59%	5.51	29.20	6.54E-08
RAF1	P04049	AMR-AFR	36	1.85%	5	0.21%	8.78	28.84	7.88E-08
CDK14	O94921	AMR-EUR	38	1.96%	5	0.25%	7.95	25.58	4.25E-07
PXN	Q59GS5	AMR-ASN	39	2.01%	2	0.13%	15.31	24.17	8.80E-07
PXN	Q59GS5	AMR-EUR	39	2.01%	8	0.39%	5.10	20.79	5.13E-06
EGFR	P00533	AMR-EUR	20	1.03%	64	3.15%	0.33	20.55	5.80E-06
RAF1	P04049	AMR-ASN	36	1.85%	3	0.20%	9.42	19.62	9.43E-06
MBP	P02686	AFR-ASN	16	0.68%	36	2.36%	0.29	18.73	1.51E-05
GFAP	P14136	AFR-EUR	1	0.04%	19	0.94%	0.05	17.34	3.12E-05
CTTN	Q8N707	ASN-EUR	22	1.44%	4	0.20%	7.33	16.96	3.83E-05
CDK16	Q00536	AFR-EUR	30	1.27%	3	0.15%	8.58	16.93	3.88E-05
CDK14	O94921	AMR-AFR	38	1.96%	13	0.55%	3.56	16.90	3.94E-05
ERBB3	P21860	AFR-EUR	20	0.84%	48	2.36%	0.36	15.57	7.95E-05
CDK14	O94921	AMR-ASN	38	1.96%	6	0.39%	4.97	15.44	8.53E-05

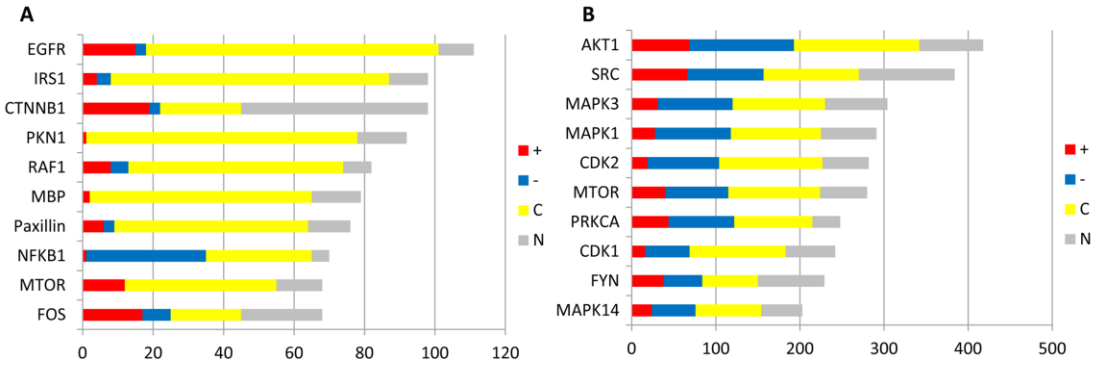
The Yates' chi-squared (χ^2) test was used (*p*-value < 1E-4). The full results (*p*-value < 0.01) were shown in Supplementary Table S8.

[a] A-B, a population pair;

[b] χ^2, the chi-squared test result.







C

	Network		Proteome		E-ratio	*P*-value
	m	M	n	N		
Cancer Gene Census	199	2644	442	28315	4.82	1.43E-87
Cancer Gene	219	2644	467	28315	5.02	7.19E-101
Drug target	501	2644	1919	28315	2.80	1.77E-110

D

	Network		Human PPN		E-ratio	*P*-value
	m	M	n	N		
Cancer Gene Census	199	2644	276	4234	1.15	3.31E-04
Cancer Gene	219	2644	309	4234	1.13	7.93E-04
Drug target	501	2644	770	4234	1.04	5.25E-02