Genome-wide analysis to predict protein sequence variations that change phosphorylation sites or their corresponding kinases

Authors

Gil-Mi Ryu[1,3], Pamela Song[2], Kyu-Won Kim[3], Kyung-Soo Oh[1], and Jong Hun Kim[2]*
*To whom correspondence should be addressed

Affiliation

[1]Center for Genome Science, 5 Nokbun-Dong, Eunpyung-Ku, Seoul, 122-701 Korea
[2]Department of neurology, Samsung Medical Center, 50 Ilwon-dong, kangnam-Ku, Seoul, 135-710 Korea
[3]Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, 599 Gwanak-ro Gwanak-Ku, Seoul, 151-742, Korea.

Address for correspondence:
Jong Hun Kim, M.D.
Department of Neurology, Samsung Medical Center
Sungkyunkwan University School of Medicine
50 Ilwon-Dong, Kangnam-Ku,
Seoul, 135-710 Korea
Phone: +82-2-3410-1426
Fax: +82-2-3410-1469
E-mail: jh7521@naver.com

Abstract

We define phosphovariants as genetic variations that change phosphorylation sites or their interacting kinases. Considering the essential role of phosphorylation in protein functions, it is highly likely that phosphovariants change protein functions and may constitute a proportion of the mechanisms by which genetic variations cause individual differences or diseases. We categorized phosphovariants into three subtypes and developed a system that predicts them. Our method can be used to screen important polymorphisms and help to identify the mechanisms of genetic diseases.

Protein phosphorylation is involved in various important processes: development and learning at the organism level, and the cell cycle, differentiation, and apoptosis at the cellular level[1,2]. Phosphorylation can change the subcellular localization of a protein, its lifespan, and its affinity for other proteins or DNA[3]. Therefore, the addition or deletion of phosphorylation sites through phosphovariants can lead to functional variations in proteins that can result in phenotypic variations or genetic diseases. By our definition, phosphovariants are variations that change phosphorylation sites or their interacting kinases. We propose three subtypes of phosphovariants. First, some variations occur directly at phosphorylation sites, and these sites will be removed if the phosphoreceptors are replaced with amino acids other than serine, threonine, or tyrosine. Conversely, replacement of another amino acid with a serine, threonine, or tyrosine may add a new phosphorylation site. Second, variations adjacent to phosphorylation sites can result in the removal or addition of phosphorylation sites. Third, variations may change the kinases that recognize phosphorylation sites, without changing the phosphorylation site itself. We divided phosphovariants into type I, II, and III, respectively, according to the above descriptions **(Fig. 1)**.

We developed PredPhospho (version 2), a Web-based computer program that predicts phosphorylation sites, and PhosphoVariant, a database for human phosphovariants. Even the advanced laboratory techniques used to analyze phosphorylation sites, such as mass spectrometry (MS), cannot analyze all types of proteins[4,5]. For example, peptides that are either too small or too large in mass can be easily missed. Moreover, membrane proteins cannot be obtained in sufficient quantities for analysis[5]. Even when proteins can be analyzed with MS, it is very time consuming and expensive to make thousands of variant proteins and select the phosphovariants. PredPhospho can predict the phosphorylation sites in kinase-specific ways, using the support vector machines (SVMs) derived from statistical learning theory proposed by Vapnik and Chervonenkis in 1995[6]. In our study, we searched for known phosphovariants and tried to predict other possible phosphovariants among human variations.

Results

Type I phosphovariants

The substitution of phosphoreceptor amino acids with amino acids other than serine, threonine, or tyrosine causes the elimination of phosphorylation sites and can be classified as type I (–) phosphovariants according to our classification. We found 52 type I phosphovariants by matching the locations of the variations and those of phosphorylation sites registered in the Swiss-Prot database and the Human Protein Resource Database[7,8] (HPRB, **Table 1**). Of these phosphovariants, 19 are known to cause Mendelian-inherited diseases and 18 are associated with cancers. Another 13 phosphovariants are polymorphisms.

Conversely, new phosphorylation sites can be created by variations and we defined these as type I (+) phosphovariants. We found an example in the study of Nousiainen *et al.*[9] The cell line that they used had a Gly766Ser mutation in the probable ATP-dependent RNA helicase DDX27 (Swiss-Prot ID, Q96GQ7), which was identified as phosphorylated. Similarly, the polymorphisms in **Tables 1**–**3** are good examples of the addition of phosphorylation sites by sequence variations. For example, if we postulate that the isoleucine at amino acid 571 of CTP synthase 1 (Swiss-Prot ID, P17812) is changed to serine, then it also represents a type I (+) phosphovariant, rather than a type I (–) phosphovariant (**Table 1**).

Type II phosphovariants

It is more difficult to find examples of type II phosphovariants than to find type I phosphovariants, because we cannot definitely say that a phosphorylation site is changed by a substitution near a phosphorylation site. However, when some kinases (although not all kinases) recognize phosphorylation sites, the specific amino acids near the phosphoreceptor are important. In **Figure 2**, we present sequence logos of phosphorylation site sequences for the CMGC kinase group. The proline residues at position +1 relative to the phosphorylation sites are important in the phosphorylation site sequences of the CMGC kinase group, especially the CDK kinase family and the MAPK kinase family. Most (84%) of the phosphorylation sites of the CMGC group registered in the Swiss-Prot database and the Human Protein Resource Database (91% of the CDK and 87% of the MAPK kinase families) have +1 proline residues. If the proline is substituted with another amino acid, it is highly probable that the adjacent phosphorylation site will be abolished. The phosphorylation site at Ser112 of peroxisome proliferator-activated receptor gamma protein (Swiss-Prot ID, P37231) is eliminated by the Pro113Gln substitution[10]. We found three other polymorphisms that abolish phosphorylation sites of the CMGC kinase group, but the removal of these phosphorylation sites has not yet been confirmed (**Table 2**). The presence of specific amino acids does not directly affect phosphorylation by kinases other than those of the CMGC kinase group, but sequences near the phosphorylation site must be considered. Kinases recognize the residues surrounding the target phosphorylation site, and the amino acids bordering phosphorylation sites are, in turn, affected by other nearby residues[11]. Hence, when the relevant kinases are not members of the CMGC kinase group, it is difficult to predict type II phosphovariants simply by database matching, without specific programs that predict phosphorylation sites.

Type III phosphovariants

Type III phosphovariants are those variations that change only the type of kinase involved, without affecting the phosphorylation site itself. For example, Ser386 of the tyrosine protein phosphatase, non-receptor type 1 (PTPN1, Swiss-Prot ID, P18031) is phosphorylated by cell division cycle 2 (CDC2) kinase, a member of the CMGC kinase group, and by casein kinase 2 (CK2), a member of the "Other" kinase group[12,13]. The Pro387Leu substitution reduces 75% of the phosphorylation by CDC2 in vitro[14]. However, it has not been confirmed that Pro387Leu inhibits the recognition of Ser386 by CK2. Only about 5% of the sites phosphorylated by CK2 that are registered in the Swiss-Prot database or HPRD have a proline residue at position +1 relative to the phosphorylation site. There is also no known consensus sequence for CK2 that contains proline at that location[15]. Therefore, we infer that the proline residue is not essential for phosphorylation by CK2 and that Pro387Leu will have little effect on phosphorylation by CK2. Therefore, we consider Pro387Leu of PTPN1 a type III phosphovariant because it inhibits the recognition of Ser386 by CDC2 kinase but has little effect on its phosphorylation by CK2[14,16].

Kinases that recognize serine and threonine differ from the kinases that recognize tyrosine. The substitution of phosphorylated serine or threonine for tyrosine, or vice versa, can remove a phosphorylation site or change the type of kinase that recognizes it. Changes between serine and threonine can also cause changes in the phosphorylation site and the responsive kinase. Therefore, the phosphovariants in **Table 3b** are either type I or type III phosphovariants.

Performance of PredPhospho and Scansite

We developed prediction models for six kinase groups and 18 kinase families. Their accuracy ranged from 70.80% to 94.67% at the kinase family level and from 71.77% to 91.18% at the kinase group level (**Supplementary Table 2** online). We tested our prediction models using two real laboratory data sets compiled with MS. For the six kinase group models, the sensitivities were 79.40% with data set I and 75.47% with data set II, but the specificities were as low as 60.62% for data set I and 61.04% for data set II because of the accumulation of false negatives by all six kinase group models. When we modified the specificity to > 95% (see Supplementary Material for the modification of the specificity for each model), the specificities increased to 72.09% and 72.39%, respectively, for each data set, whereas the sensitivities decreased to 73.24% and 65.76%, respectively (**Table 4**). At a specificity of > 99%, the specificities changed to 95.79% and 96.62%, respectively, and the sensitivities to 23.39% and 20.05%, respectively. Scansite is a widely used Web-based prediction software for phosphorylation sites[17]. We also tested Scansite with the same data sets. When we applied Scansite with the low-stringency option to the data sets, the specificities were 52.60% and 57.06%, respectively, and the sensitivities were 84.47% and 83.92%, respectively, whereas with the high-stringency option, the specificities were 96.77% and 95.71%, respectively, and the sensitivities were 16.39% and 13.60%, respectively. The performance of PredPhospho with no modification to the specificity is similar to that of Scansite used with the low-stringency option. The performance of PredPhospho with a specificity of > 99% was similar with that of Scansite used with the high-stringency option. The data sets were analyzed in the same way with the 18 kinase family models (**Table 4**). The family-wise prediction generally had greater sensitivity and lower specificity than the group-wise prediction.

Phosphovariants predicted with PredPhospho and Scansite

The numbers of phosphovariants predicted with PredPhospho and Scansite are shown in **Table 5**. In the supplementary data (online), we present the results for phosphovariants that were predicted with PredPhospho with the > 99% specificity option. The sensitivity and specificity of the prediction of type I phosphovariants will be the same as the result shown in **Table 4,** because we can predict type I phosphovariants simply by the location of phosphorylation sites, with no knowledge of the kinds of kinases involved. The notions of type II and type III phosphovariants include the kinds of kinases that recognize the phosphorylation sites. Therefore, the prediction of type II and III phosphovariants differs from that of type I. Not only the phosphorylation site but also the type of kinase must be identified, because important amino acid residues flanking the phosphorylation sites, which guide kinases to the site, may differ according to the kinase involved. Therefore, the general performance of our prediction of type II and type III phosphovariants will be somewhat different from those shown in **Table 4.** Instead, the performance of the each kinase-specific prediction can be judged from the performances shown in **Supplementary Table 2** (online).

The proportion of phosphovariants predicted by PredPhospho is shown in **Table 6**. These data were selected with the > 99% specificity option at the kinase family level, and are related to the confirmed phosphorylation sites in human or orthologous proteins, for which kinase information is not yet available. As described in the Supplementary Material (online), only phosphorylation sites with available kinase information were used as training models for PredPhospho. Hence, the phosphorylation sites in **Table 6** are predicted ones with PredPhospho, because the phosphorylation sites were not included in the training data for PredPhospho. If a specific site is shown to be a phosphorylation site in human or orthologous proteins, then the site is definitely located on the surface of the protein, and therefore, is accessible to a kinase. The predicted phosphovariants related to a proven phosphorylation site are more likely to be true than are those related to an unproven phosphorylation site. Numerous phosphorylation sites in humans or other species are constantly being identified. The priority for further research among predicted phosphovariants can be decided based on the confirmation of specific phosphorylation sites.

Discussion

Changes in phosphorylation sites cause various diseases by numerous mechanisms. Some proven mechanisms of the phosphovariants shown in **Tables 1–3** are related to changes in the protein's affinity for DNA, inducing hyperphosphorylation and the inhibition of ubiquitinization. For example, microphthalmia-associated transcription factor (MITF) activates the transcription of the tyrosinase gene. The Ser405Pro change in MITF eliminates the phosphorylation site at Ser405 and inhibits the binding of MITF to DNA. As a result, the mutation causes Waardenburg syndrome type IIa, the symptoms of which include depigmentation and sensorineural hearing loss[18]. Abnormal phosphorylation can also cause disease by increasing phosphorylation at other sites. The hyperphosphorylation of tau protein induces neurofibrillary tangles and the accumulation of these tangles can result in Alzheimer's disease and frontotemporal dementia (FTD). Paradoxically, serine threonine protein kinase N (PKN) interrupts the phosphorylation of other sites by phosphorylating Ser637 and Ser669 of tau protein[19]. The mutations Ser637Phe and Ser669Leu of tau protein eliminate the recognition sites for PKN and induce the hyperphosphorylation of tau protein. FTD and the respiratory failure with dementia are known to be related to Ser637Phe and Ser669Leu, respectively[20,21]. Phosphorylation at Ser32 of NF-κB inhibitor α (Swiss-Prot ID, P25963) causes ubiquitinization and results in the activation of NF-κB[22]. The Ser32Ile substitution of NF-κB inhibitor α (Swiss-Prot ID, P25963) eliminates the phosphorylation site at Ser32. Consequently, the ubiquitinization of NF-κB inhibitor α is inhibited by the Ser32Ile variant of NF-κB inhibitor α and NF-κB cannot be activated. The Ser32Ile variant of NF-κB inhibitor α causes autosomal dominant anhydrotic ectodermal dysplasia with immunodeficiency[23]. These are a few examples of phosphovariants. Considering the numerous functional roles played by phosphorylation in vivo, there must be many mechanisms by which phosphovariants can cause specific diseases, and these must be identified.

Some phosphovariants do not cause Mendelian-inherited diseases, but change an individual's susceptibility to disease. The Pro113Gln substitution (dbSNP id, rs1800571) of peroxisome proliferator-activated receptor gamma (Swiss-Prot ID, P37231), which eliminates the phosphorylation site at Ser112, is known to cause obesity[10]. The Pro387Leu substitution (dbSNP ID, rs16995309) of tyrosine-protein phosphatase non-receptor, type 1 (Swiss-Prot ID, P18031) is associated with type II diabetes mellitus[14,16]. We only found these two polymorphic phosphovariants that are related to disease susceptibility. As shown in **Tables 1–3**, 21 phosphovariants are polymorphisms, and the biological significance of 19 of these polymorphisms is not yet known. In addition, we do not know the biological significance of most of the polymorphic phosphovariants predicted in our study. Considering the importance of phosphorylation in protein function, polymorphic phosphovariants may well be involved in specific diseases or phenotypes.

Apart from the characteristics of the three types of phosphovariants already suggested, there are other fundamental differences between the type I and other phosphovariants. Type I phosphovariants completely add or remove a phosphorylation site because kinases can only donate a phosphor moiety to an amino acid with a hydroxyl group. However, type II and III phosphovariants can significantly affect kinase kinetics without completely changing the kinase's recognition site. For example, the Pro387Leu substitution of PTPN1 removes only 75% of the phosphorylation of Ser386 by CDC2 kinase in vitro, rather than 100% (**Table 3**)[14].

We can explain phenotypic variations and diseases in terms of phosphovariants in more cases than we have anticipated. Of the human proteins registered in the Swiss-Prot database, 25.5% are phosphoproteins and

60.9% of these phosphoproteins have multiple phosphorylation sites. The protein with the greatest number of confirmed phosphorylation sites is the serine/arginine repetitive matrix protein 2 (Swiss-Prot ID, Q9UQ35), with 195 phosphorylation sites. Although we could only determine 62 phosphovariants with a database search, many more phosphovariants must exist. Furthermore, if we count the mutations that add phosphorylation sites and the type II and III phosphovariants that cannot be found without specific programs, the number of mutations associated with changes in phosphorylation sites is much greater. We did not consider haplotypes in this study, for simplicity. However, if two or more nearby variations are frequently linked, a nearby phosphorylation site can be altered, although each variation individually does not affect the phosphorylation site. Therefore, phosphovariants may account for a much greater proportion of human variation than we have anticipated.

Several issues must be resolved in future studies. Type II and type III phosphovariants can be interchanged. For example, removed phosphorylation sites associated with type II (–) phosphovariants predicted with PredPhospho may be recognized by kinases that are not included in our prediction models. In such cases, these variations are type III phosphovariants, not type II (–). Conversely, if a phosphorylation site is falsely classified as a site recognized by multiple kinases, instead of by one true kinase, and if a variation affects only some of these kinases, this is a type II phosphovariant incorrectly predicted to be a type III phosphovariant. Other points that must be improved are the low sensitivity achieved with the high-specificity option and the low specificity achieved with the no-specificity option of PredPhospho (**Table 4**). Moreover, the number of types of kinases that we can predict must be increased and haplotypes should be considered in future studies.

Our method can be used in pathophysiological studies of mutations and in the selection of polymorphisms of clinical and phenotypical importance. Many of the papers that have described the variations, shown in **Tables 1–3,** did not mention that the variations could be related to changes in phosphorylation sites. This could be attributable to the lack of a specific database that connects mutations with phosphorylation sites, or the lack of a general understanding of the association between phosphorylation and mutation. The type I (+), II, and III phosphovariants we have defined cannot be identified simply by database analyses. Specific programs are required to identify these phosphovariants. Accordingly, many nonsense point mutations whose functional mechanisms are unknown can be reconsidered in terms of phosphovariant. Furthermore, if some mutations are predicted to be phosphovariants with our system, further research will clarify the cause of the associated disease or protein function. Our system can be used to select meaningful variations among endless numbers of newly identified polymorphisms. As sequencing techniques advance, a large number of genetic variations are emerging. At present, comparison of whole genomes of individuals is possible, because the human genome can be sequenced in two months[24]. A comparison of phosphovariants between individuals or between species can be undertaken before amino acid variations or nucleic acid variations are compared in whole genomes. A reverse genetic approach for unknown protein functions or phenotypic variations is possible with proven phosphovariants. The screening and prediction of phosphovariants can be a starting point for further research.

Methods

PredPhospho

We created classifiers of various kinases by training SVMs with phosphorylation site sequences and nonphosphorylated site sequences. In other words, our classifiers determine whether serine, threonine, or tyrosine residues within a sequence can be phosphorylated or not. "Phosphorylated site sequences" refers to peptide sequences with a serine, threonine, or tyrosine residue located at the center, and which are phosphorylated. Conversely, "nonphosphorylated site sequences" are sequences with a serine, threonine, or tyrosine residue located at the center, which cannot be phosphorylated. We obtained phosphorylated site sequences from public databases: the Swiss-Prot (release 54.8) and the Human Protein Resource Database (HPRD, release 7). Nonphosphorylated site sequences were taken from laboratory data confirmed by MS (see Supplementary Material online).

Manning et al. found 518 human protein kinase genes in the human genome sequence, using the hidden Markov model (HMM) profile, and confirmed the identities of more than 90% of the identified kinase genes using cDNA cloning[25]. They also classified the protein kinase superfamily into nine broad groups, and subdivided the groups into 134 families and 204 subfamilies, using sequence comparisons of the kinase catalytic domains. We classified the phosphorylated site sequences according to their kinases and created the classifiers in a kinase-specific manner. Because of the limitations of the phosphorylated sequence data presently available in public databases, we can make classifiers for only six kinase groups: AGC, CAMK, CK1, CMGC, STE, and TK; and 18 kinase families: AKT, CAMK2, CAMKL, CDK, CK1, CK2, GSK, IKK, JakA, MAPK, PDGFR, PIKK, PKA, PKC, RSK, Src, STE20, and Syk (all abbreviations are shown in the footnote to **Supplementary Table 1** online). The detailed algorithms and methods were described in the Supplementary Material (online).

Evaluation of the system

The performance status of the prediction for each kinase group and family is shown in **Supplementary Table 2** (online). The performance of the prediction with combinations of all the kinase group models or all the family group models is not the numerical multiplication of the performance of each model. Therefore, to evaluate the performances of the predictions at the kinase group level or at the family level, we tested two proven real data sets, which were compiled with MS experiments. Data set I was created by Olen et al.[5], who identified phosphorylation sites in proteins from HeLa cells and classified the phosphopeptides according to their definition. Four classes are based on their localization probabilities: < 0.25, $0.25 \leq < 0.75$ without kinase motifs, $0.25 \leq < 0.75$ with kinase motifs and ≥ 0.75. We used only monophosphopeptides that had localization probabilities of at least 0.75. Data set II was derived from the paper of Beausoleil et al.[4] and we used those of their phosphopeptides with a localization certainty of > 99.4% (see **Supplementary Fig. 2** online). We selected phosphorylation sites and nonphosphorylated sites from these two data sets. To avoid overestimating the performance of PredPhospho, we discarded sequences that were more than 70% identical to sequences used for training PredPhospho. We also avoided using false nonphosphorylated sites by omitting those sites that are listed as phosphorylation sites in Swiss-Prot or HPRD. We tested the two kinds of data sets not only with our PredPhospho, but also with Scansite.

Prediction of phosphovariants

We extracted information about human genetic variations from SwissVariant of the Swiss-Prot database. SwissVariant includes single amino acid polymorphisms and missense mutations[26]. The number of variations listed in SwissVariant was 33,651. We consulted the Swiss-Prot database and HPRD about their effects, and the references to these variations and phosphorylation sites. With PredPhospho and Scansite, we predicted the phosphorylation sites and related kinases for the original sequences and the variant sequences. The phosphovariants could be identified when the phosphorylation sites or interacting kinases were altered between the original sequence and the variant sequence. If the phosphorylation site is in the same location as the variation, it is type I. In type II phosphovariants, the variation is not in the same location as the phosphorylation site. We added the symbol (+) to types I and II when the phosphovariants added new phosphorylation sites, and (–) when the phosphovariants removed phosphorylation sites [e.g., type I (+] or type I (–)]. Type III phosphovariants are caused by changes in the types of kinases involved, rather than in the phosphorylation site itself, regardless of the locations of the variations. One variation can include more than one class of phosphovariant, because one variation can affect two or more phosphorylation sites. We predicted phosphovariants at the kinase group level and the family level. The predictions at the family level are more sensitive, but less specific, than those at the group level. To minimize false negatives, we varied the specificity options (95%, 97%, 98%, or 99%) according to the specificity of each model (**Supplementary Table 3** online). The specificity options are described in the Supplementary Material (online).

Sequence logos

We obtained 562 phosphorylation site sequences recognized by the CMGC kinase group from Swiss-Prot and HPRD. We trimmed the sequences as 6 symmetric residues centered phospohrylation sites. We aligned the sequences and obtained a sequence logo using the web program (http://weblogo.berkeley.edu/logo.cgi).

WWW programs

The PredPhospho (version 2) and PhosphoVariant were implemented using the PERL (version 5.8.8) programming language and MySQL (version 5.0.18). The PhosphoVariant is a database for the definite and possible human variants changing phosphorylation sites and their interacting kinases. They are available at http://phosphovariant.ngri.go.kr/seq_input_predphospho2.htm and http://phosphovariant.ngri.go.kr, respectively.

Acknowledgement

This study was supported by Health Fellowship Foundation.

Author contributions

This study was designed by J.H.K. (corresponding author), R.G.M. (first author), K.W.K. and O.K.S. Programming and web design were performed by J.H.K. and R.G.M. Literature search and data analysis were done by J.H.K., R.G.M., P.S., K.W.K. and O.K.S. Paper was written by J.H.K., R.G.M., and P.S.

Reference

1. Hunter, T. Signaling--2000 and beyond. *Cell* **100**, 113-27 (2000).
2. Dash, P.K., Moore, A.N., Kobori, N. & Runyan, J.D. Molecular activity underlying working memory. *Learn Mem* **14**, 554-63 (2007).
3. Hunter, T. & Karin, M. The regulation of transcription by phosphorylation. *Cell* **70**, 375-87 (1992).
4. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* **24**, 1285-92 (2006).
5. Olsen, J.V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-48 (2006).
6. Kecman, V. *Learning and Soft Computing*, (The MIT Press, Cambridge, 2001).
7. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-70 (2003).
8. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-71 (2003).
9. Nousiainen, M., Sillje, H.H., Sauer, G., Nigg, E.A. & Korner, R. Phosphoproteome analysis of the human mitotic spindle. *Proc Natl Acad Sci U S A* **103**, 5391-6 (2006).
10. Ristow, M., Muller-Wieland, D., Pfeiffer, A., Krone, W. & Kahn, C.R. Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *N Engl J Med* **339**, 953-9 (1998).
11. Pinna, L.A. & Ruzzene, M. How do protein kinases recognize their substrates? *Biochim Biophys Acta* **1314**, 191-225 (1996).
12. Jung, E.J., Kang, Y.S. & Kim, C.W. Multiple phosphorylation of chicken protein tyrosine phosphatase 1 and human protein tyrosine phosphatase 1B by casein kinase II and p60c-src in vitro. *Biochem Biophys Res Commun* **246**, 238-42 (1998).
13. Flint, A.J., Gebbink, M.F., Franza, B.R., Jr., Hill, D.E. & Tonks, N.K. Multi-site phosphorylation of the protein tyrosine phosphatase, PTP1B: identification of cell cycle regulated and phorbol ester stimulated sites of phosphorylation. *EMBO J* **12**, 1937-46 (1993).
14. Echwald, S.M. et al. A P387L variant in protein tyrosine phosphatase-1B (PTP-1B) is associated with type 2 diabetes and impaired serine phosphorylation of PTP-1B in vitro. *Diabetes* **51**, 1-6 (2002).
15. Amanchy, R. et al. A curated compendium of phosphorylation motifs. *Nat Biotechnol* **25**, 285-6 (2007).
16. Ukkola, O. et al. Protein tyrosine phosphatase 1B variant associated with fat distribution and insulin metabolism. *Obes Res* **13**, 829-34 (2005).
17. 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**, 3635-41 (2003).
18. Tassabehji, M. et al. The mutational spectrum in Waardenburg syndrome. *Hum Mol Genet* **4**, 2131-7 (1995).
19. Taniguchi, T. et al. Phosphorylation of tau is regulated by PKN. *J Biol Chem* **276**, 10025-31 (2001).
20. Nicholl, D.J. et al. An English kindred with a novel recessive tauopathy and respiratory failure. *Ann Neurol* **54**, 682-6 (2003).
21. Rosso, S.M. et al. A novel tau mutation, S320F, causes a tauopathy with inclusions similar to those in Pick's disease. *Ann Neurol* **51**, 373-6 (2002).

22. DiDonato, J. et al. Mapping of the inducible IkappaB phosphorylation sites that signal its ubiquitination and degradation. *Mol Cell Biol* **16**, 1295-304 (1996).
23. Courtois, G. et al. A hypermorphic IkappaBalpha mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. *J Clin Invest* **112**, 1108-15 (2003).
24. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-6 (2008).
25. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912-34 (2002).
26. 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**, 464-70 (2004).

Table 1 Examples of type I phosphovariants

(a) Type I(−) phosphovariants

Gene name	SWISS-PROT ID	Variation site[a] (SWISS-PROT variant ID)	Phosphor-ylation site	Effect[b]	Reference(s) for variation[c]	Reference(s) for phosphor-ylation site[d]
Phosphovariants causing Mendelian inheritant diseases						
EDNRB	P24530	S305N (VAR_003472)	S305	Hirschsprung disease type 2	8852659	14636059
FANCA	O15360	S858R (VAR_017498)	S858	Fanconi anemia	10094191 11091222	17924679
KCNJ1	P48048	S219R (VAR_019726)	S219	Bartter syndrome type 2	8841184	8621594
L1CAM	P32004	S1194L (VAR_003947)	S1194	Hydrocephalus due to stenosis of the aqueduct of Sylvius Mental retardation, aphasia, shuffling gait, and adducted thumbs syndrome	8556302 7881431	17081983
MAPT	P10636	S622N (VAR_010350)	S622	Frontotemporal dementia and parkinsonism linked chromosome 17	10208578	7706316
MAPT	P10636	S637F (VAR_019665)	S637	Pick disease	11891833	11104762 9199504
MAPT	P10636	S669L (VAR_019667)	S669	Fatal respiratory hypoventilation	14595660	11104762
MITF	O75030	S405P (VAR_010302)	S405	Waardenburg syndrome type IIa	8589691	10587587
NFKBIA	P25963	S32I (VAR_034871)	S32	Autosomal dominant anhidrotic ectodermal dysplasia with immunodeficiency	14523047	10882136 9721103 8601309 16319058 10723127 9214631
PER2	O15055	S662G (VAR_029080)	S662	Familial advanced sleep-phase syndrome	11232563	11232563
PTPN11	Q06124	Y62D (VAR_015605)	Y62	Patients with Noonan syndrome 1 manifesting juvenile myelomonocytic leukemia	11992261 12325025 12960218 12717436	15951569 15592455
RAF1	P04049	S259F (VAR_037809)	S259	Noonan syndrome type 5	17603483	8349614 11997508 11971957 10576742

RAF1	P04049	T491R (VAR_037819)	T491	Noonan syndrome type 5	17603483	11447113
RAF1	P04049	T491I (VAR_037818)	T491	Noonan syndrome type 5	17603483	11447113
RPS6KA3	P51812	S227A (VAR_006195)	S227	Coffin-Lowry syndrome	8955270	17192257
STAT3	P40763	Y657C (VAR_037381)	Y657	Hyperimmunoglobulin E recurrent infection syndrome autosomal dominant	17881745	15037656
TGFBR2	P37173	Y336N (VAR_022352)	Y336	Loeys-Dietz aortic aneurysm syndrome	15731757	9169454
TNNI3	P19429	S166F (VAR_029454)	S166	Hypertrophic cardiomyopathy	12974739	11121119
TSC1	Q92574	T417I (VAR_009403)	T417	Tuberous sclerosis complex, could be a polymorphism	10570911 10607950	14551205
Phosphovariants found in cancer						
CDH1	P12830	S838G (VAR_001322)	S838	Ovarian cancer	8075649	10671552
CTNNB1	P35222	S23R (VAR_017612)	S23	Hepatocellular carcinoma, no effect	10435629 12027456	12027456
CTNNB1	P35222	S33F (VAR_017617)	S33	Pilomatrixoma, medulloblastoma and hepatocellular carcinoma	10666372 10435629 10192393	12000790 12114015 11818547
CTNNB1	P35222	S33L (VAR_017618)	S33	Hepatocellular carcinoma	10435629	12000790 12114015 11818547
CTNNB1	P35222	S37A (VAR_017624)	S37	Medulloblastoma, hepatocellular carcinoma	12027456,10435629, 10666372	12000790 12114015 11818547
CTNNB1	P35222	S37C (VAR_017625)	S37	Pilomatrixoma, hepatoblastoma	9927029, 10192393	12000790 12114015 11818547
CTNNB1	P35222	S37F (VAR_017626)	S37	Pilomatrixoma	10192393	12000790 12114015 11818547
CTNNB1	P35222	T41A (VAR_017629)	T41	Hepatoblastoma and hepatocellular carcinoma, also in a desmoid tumor	12051714 10398436 9927029 12027456 10655994 10435629	12051714 12114015 11818547 12000790
CTNNB1	P35222	T41I (VAR_017630)	T41	Pilomatrixoma and hepatocellular carcinoma	10192393 10435629	12051714 12114015

						11818547 12000790
CTNNB1	P35222	S45F (VAR_017631)	S45	Hepatocellular carcinoma	10435629	12051714 12000790 11955436
CTNNB1	P35222	S45P (VAR_017632)	S45	Hepatocellular carcinoma	10435629	12051714 12000790 11955436
FAM10A4	Q8IZP2	S71L (VAR_023644)	S71	B-Cell leukemia, multiple myeloma, and prostate cancer	12079276	17081983
MET	P08581	Y1230C (VAR_006292)	Y1230	Hereditary papillary renal carcinoma	9140397	12475979
MET	P08581	Y1230H (VAR_006293)	Y1230	Hereditary papillary renal carcinoma	9140397	12475979
NME1	P15531	S120G (VAR_004625)	S120	Neuroblastoma	8047138	8810265
RB1	P06400	S567L (VAR_005579)	S567	Retinoblastoma	10671068 2594029	10207050
TP53	P04637	T155A (VAR_005901)	T155	Esophageal cancer	1868473	12628923
Phosphovariants related with polymorphism						
BARD1	Q99728	S186G	S186	polymorphism (rs16852741)		15855157
C10orf11	Q9H2I8	S153F (VAR_033686)	S153	polymorphism (rs35349706)		16964243
CTNND1	O60716	Y217C (VAR_020929)	Y217	polymorphism (rs11570194)		15592455 16212419
CTPS	P17812	S571I (VAR_027055)	S571	polymorphism (rs17856308)	15489334	16097034 17081983
HIF1A	Q16665	T796A (VAR_015854)	T796	polymorphism (rs1802821)		17382325
INSR	P06213	Y1361C (VAR_015933)	Y1361	polymorphism (rs13306449)	7657032	11401470
KRT36	O76013	T315M (VAR_020306)	T315	polymorphism (rs2301354)		17081983
MYH15	Q9Y2K3	T1125A (VAR_030238)	T1125	polymorphism (rs3900940)		17081983
PDLIM5	Q96HC4	S136F (VAR_023779)	S136	polymorphism (rs2452600)		17287340
PNN	Q9H307	S671G (VAR_023368)	S671	polymorphism (rs13021)	10095061	17287340
SUB1	P53999	S11G	S11	polymorphism (rs17850527)	15489334	17081983 16689930

SRRM2	Q9UQ35	S883C (VAR_027260)	S883	polymorphism (rs17136053)		17287340
TP53	P04637	S366A (VAR_022317)	S366	Polymorphism		9183006

(b) Type I(+) phosphovariants

Gene name	SWISS-PROT ID	Variation site	Phosphor-ylation site	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
DDX27	Q96GQ7	G766S	S766	Unknown	16565220	16565220

[a] Locations and amino acid changes of the variations in the proteins.

[b] The meanings or consequences of the variations. We referred to the feature tables of Swiss-Prot for these effects. If the polymorphisms are enrolled in dbSNP, the IDs of dbSNP are written in the parentheses.

[c] Pubmed ID for the references of the variations

[d] Pubmed ID for the references of the phosphorylation sites

Protein names which are abbreviated by their gene names: Catenin β-1, **CTNNB1**; Epithelial cadherin [Precursor], **CDH1**; Probable ATP-dependent RNA helicase DDX27, **DDX27;** Endothelin B receptor [Precursor], **EDNRB**; Protein FAM10A4, **FAM10A4**; Fanconi anemia group A protein, **FANCA**; ATP-sensitive inward rectifier potassium channel 1, **KCNJ1**; Keratin, type I cuticular Ha6, **KRT36**; Neural cell adhesion molecule L1, **L1CAM**; Microtubule-associated protein tau, **MAPT**; Hepatocyte growth factor receptor [Precursor], **MET**; Microphthalmia-associated transcription factor, **MITF**; NF-κ-B inhibitor α, **NFKBIA**; Nucleoside diphosphate kinase A, **NME1**; Period circadian protein homolog 2, **PER2**; Tyrosine-protein phosphatase non-receptor type 11, **PTPN11**; RAF proto-oncogene serine/threonine-protein kinase, **RAF1**; Retinoblastoma-associated protein, **RB1**; Ribosomal protein S6 kinase alpha-3, **RPS6KA3**; Signal transducer and activator of transcription 3, **STAT3**; TGF-beta receptor type-2 [Precursor], **TGFBR2**; Cardiac troponin I, **TNNI3**; Cellular tumor antigen p53, **TP53**; Hamartin, **TSC1**

Table 2 Examples of type II(−) phosphovariants.

Gene name	SWISS-PROT ID	Variation site (SWISS-PROT variant ID)	Removed phosphorylation site (related kinases)	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
DUT	P33316	P100S (VAR_022314)	S99[a] (CDC2)	Polymorphism		17081983 8631817
GJA1	P17302	P283L (VAR_014101)	S282[a] (ERK1, ERK2 and MAPK7)	Polymorphism (rs2228974)		8631994 9535905
PPARG	P37231	P113Q (VAR_010724)	S112[b] (ERK2, JNK1 and MAPK8)	Obesity and polymorphism (rs1800571)	9753710	9030579
RXRA	P19793	P261L (VAR_014620)	S260[a] (ERK2 and MAPK7)	Polymorphism (rs2234960)		12048211

If the variations substitute the proline residues at position +1 relative to the phosphorylation sites into other amino acids, the nearby phosphorylation sites recognized by the CMGC kinase group can be eliminated or the efficiency of phosphorylation in that site is significantly decreased.

[a] The removals of the phosphorylation sites by the variation have not been confirmed by experiments. However, the removals of the phosphorylation sites are highly possible because the nearby phosphorylation sites are proved to be recognized by the CMGC group.

[b] The removal of the phosphorylation site by the variation has been confirmed by a experiment[10].

Protein names which are abbreviated by their gene names: Deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial [Precursor], **DUT**; Gap junction α-1 protein, **GJA1**; Peroxisome proliferator-activated receptor γ, **PPARG**; Retinoic acid receptor RXR-α, **RXRA**

Table 3 Possible examples of type III phosphovariants.

(a) A possible example of type III phosphovariant

Gene name	SWISS-PROT ID	Variation site (SWISS-PROT variant ID)	Related phosphorylation site (kinase recognizing it)	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
PTPN1	P18031	P387L (VAR_022014)	S386 (CDC2 and CK2)	Low glucose tolerance and polymorphism (rs16995309)	15919835	9600099 8491187

(b) Possible examples of phosphovariants which can be classified as type I or III.

Gene name	SWISS-PROT ID	Variation site (SWISS-PROT variant ID)	Related phosphorylation site	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
BRCA1	P38398	S1217Y (VAR_020695)	S1217	Breast cancer and breast-ovarian cancer	14722926	17081983
CASP8	Q14790	S219T (VAR_025816)	S219	polymorphism (rs35976359)		17525332
CDK2	P24941	Y15S (VAR_016157)	Y15	polymorphism (rs3087335)		1396589 12912980 12972555 15144186
CTNNB1	P35222	S33Y (VAR_017619)	S33	Pilomatrixoma	12027456 10192393	12000790 12114015 11818547
CTNNB1	P35222	S37Y (VAR_017627)	S37	Hepatocellular carcinoma	10435629	12000790 12114015 11818547
TEK	Q02763	Y897S (VAR_008716)	Y897	Dominantly inherited venous malformations	10369874	11080633
XRCC1	P18887	S485Y (VAR_014779)	S485	polymorphism (rs2307184)		15066279

The reasons why these variations are classified as type III are detailed in the text.

Protein names which are abbreviated by their gene names: Breast cancer type 1 susceptibility protein; **BRCA1**; Caspase 8, **CASP8**; Cyclin dependent kinase 2, **CDK2**; Catenin β-1, **CTNNB1**; Tyrosine-protein phosphatase non-receptor type 1, **PTPN1**; Angiopoietin-1 receptor [Precursor], **TEK**; DNA repair protein XRCC1, **XRCC1**

Table 4 The general performance test of the PredPhospho and Scansite for two real data sets

(a) The number of each data set

	Data I[a]		Data II[a]	
	(+)[b]	(−)[b]	(+)	(−)
Ser/Thr[b]	1860	926	3017	315
Tyr[b]	38	95	359	11

(b) PredPhospho

Prediction at the group level					Prediction at the family level				
	Data I		Data II			Data I		Data II	
Specificity[c]	***Sn*** (%)	***Sp*** (%)	***Sn*** (%)	***Sp*** (%)	Specificity	***Sn*** (%)	***Sp*** (%)	***Sn*** (%)	***Sp*** (%)
No	79.40	60.62	75.47	61.04	No	95.36	29.29	93.60	30.98
95%	73.24	72.09	65.76	72.39	95%	92.68	42.80	88.95	46.63
97%	53.37	80.22	48.31	79.45	97%	88.46	56.81	83.50	57.36
98%	43.11	89.23	38.92	89.26	98%	82.03	66.80	75.44	62.88
99%	23.39	95.79	20.05	96.62	99%	73.18	72.67	66.73	72.39

(c) Scansite

	Data I		Data II	
Stringency[d]	***Sn*** (%)	***Sp*** (%)	***Sn*** (%)	***Sp*** (%)
Low	84.47	52.60	83.92	57.06
Medium	48.63	85.21	43.81	87.73
High	16.39	96.77	13.60	95.71

[a] Different data sets compiled with mass spectrometer. See the text for the detail explanation for data set I and II.

[b] The types of amino acids located at the center of peptides. We annotated the peptides as (+) if the Ser/Thr or Tyr at the center of the peptides is phosphorylated. On contrary, we designated the peptide as (−) if the center of the peptides is not phosphorylated.

[c] Options of the specificity. For example, '99%' specificity option mean cutoff value is adjusted for each model to have 99% specificity, and 'No' specificity option means each model has default cutoff value without adjustment of specificity (See supplementary material online).

[d] Scansite has three levels of stringency: high, medium and low. High stringency involves low sensitivity and high specificity, whereas low stringency involves high sensitivity and low specificity.

Abbreviations: sensitivity, ***Sn;*** specificity, ***Sp***

Table 5 The number of the phosphovariants

(a) The number of the phosphovariants predicted with PredPhospho at the kinase group level

	Type I(−)	Type I(+)	Type II(−)	Type II(+)	Type III
Specificity					
No	1729	2036	5455	4980	5299
95%	981	1195	1304	1070	986
97%	613	778	694	542	401
98%	314	409	329	213	151
99%	116	150	98	52	21

(b) The number of the phosphovariants predicted with PredPhospho at the kinase family level

	Type I(−)	Type I(+)	Type II(−)	Type II(+)	Type III
Specificity					
No	3039	3717	3969	3926	23955
95%	2379	2910	2882	2840	8113
97%	1720	2104	1439	1483	2390
98%	1268	1551	783	862	1213
99%	946	1180	539	548	638

(c) The number of the phosphovariants predicted with Scansite

	Type I(−)	Type I(+)	Type II(−)	Type II(+)	Type III
Stringency					
Low	1581	1852	4255	3773	7698
Medium	443	498	487	384	152
High	83	128	35	28	1

Table 6 Predicted[a] phosphovariants whose phosphorylation sites were confirmed in human or orthologous proteins

(a) Type I(−) phosphovariants

Gene name	SWISS-PROT ID	Site (SWISS-PROT variant ID)	Related phosphorylation site	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
ACIN1	Q9UKV3	S478F (VAR_022033)	S478	Polymorphism (rs3751501)		17242355 (mouse)[b]
MECP2	P51608	S229L (VAR_018200)	S229	Polymorphism	10767337 12872250	17046689 (rat)[a]
PAH	P00439	S16P (VAR_000869)	S16	Phenylketonuria	1679029 2246858 1301187	7387651 (rat)[b]

(b) Type II(−) phosphovariants

Gene name	SWISS-PROT ID	Site (SWISS-PROT variant ID)	Related phosphorylation site	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
CHGB	P05060	R178Q (VAR_020287)	S183	Polymorphism (rs910122)		16807684
GTSE1	Q9NYZ3	R506W (VAR_024154)	S504	Polymorphism (rs140054)	10591208	16964243
LIG1	P18858	P52L (VAR_020194)	S51	Polymorphism (rs4987181)		16964243

(c) **Type III phosphovariant**

Gene name	SWISS-PROT ID	Site (SWISS-PROT variant ID)	Related phosphorylation site[1]	Effect	Reference(s) for variant	Reference(s) for phosphor-ylation site
ABCB11	O95342	R698H (VAR_035352)	S701	Polymorphism	16763017	17242355 (mouse)[b]
ABL1	P00519	P810L (VAR_032678)	S809	Polymorphism	17344846	17081983
AQP2	P41181	P262L (VAR_015255)	S261	Nephrogenic diabetes insipidus	9550615 15509592	16641100 (rat)[b]
CASP8	Q14790	S219T (VAR_025816)	S219	Polymorphism (rs35976359)		17525332
EIF4G3	O43432	P496A (VAR_034009)	S495	Polymorphism (rs35176330)		17081983 16964243
MYBPC3	Q14896	G278E (VAR_019891)	S275	Familial hypertrophic	12707239	9784245 (chicken)[b]

				cardiomyopathy type 4		
PKMYT1	Q99640	R140C (VAR_019928)	S143	Polymorphism (rs4149796)		17192257
PPP1R12B	O60237	R836K (VAR_024177)	S839	Polymorphism (rs3881953)		17242355 (mouse)[b]
PARK7	Q99497	E64D (VAR_020493)	Y67	Autosomal recessive early-onset Parkinson disease 7	15365989 14607841	15592455
PNN	Q9H307	S671G (VAR_023368)	S667	Polymorphism (rs13021)	10095061	17287340
SH3PXD2A	Q5TCZ1	R1035Q (VAR_030782)	S1038	Polymorphism (rs3781365)		17525332
WDR91	A4D1P6	L257P (VAR_033358)	S256	Polymorphism (rs292592)	15489334 14702039	16964243

Eight of the type I(−) phosphovariants (VAR_006195, VAR_023368, VAR_023779, VAR_023644, VAR_030238, VAR_020306, VAR_033686, and VAR_027260), and a type III phosphovariant (VAR_020695) were also predicted. However, their detail information are already written in Table 1 and 3.

[a] The prediction was done with the 99% specificity option of PredPhospho at the kinase family level.

[b] The experiment was done in the proteins of other than human. The names of the species are written in the parenthesis.

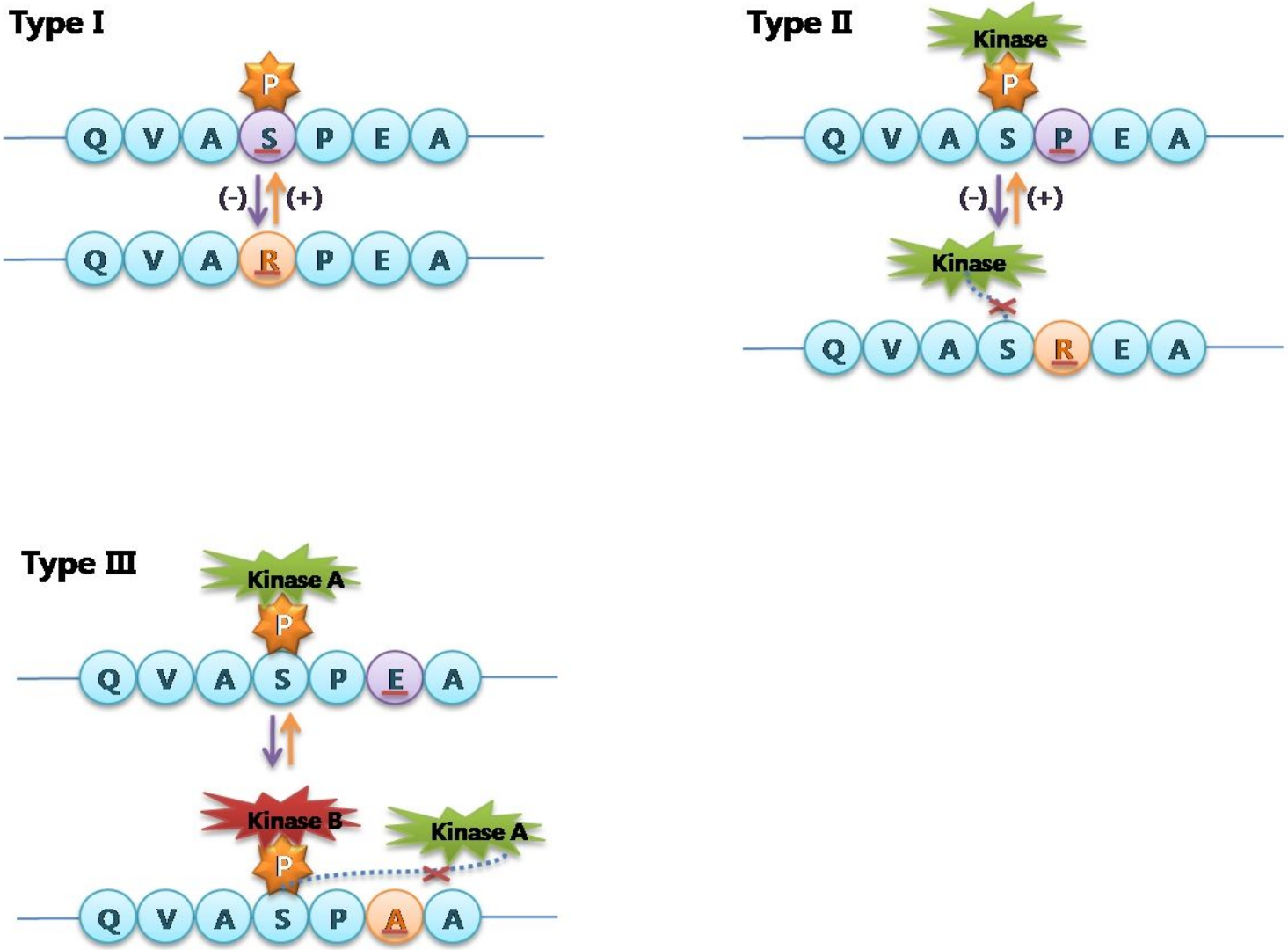
Protein names which are abbreviated by their gene names: Proto-oncogene tyrosine-protein kinase ABL1, **ABL1**; ATP-binding cassette sub-family B member 11, **ABCB11**; Apoptotic chromatin condensation inducer in the nucleus, **ACIN1**; Aquaporin-2, **AQP2**; Caspase-8 [Precursor], **CASP8**; Secretogranin-1 [precusor], **CHGB**; Eukaryotic translation initiation factor 4 gamma 3, **EIF4G3**; G2 and S phase-expressed protein 1, **GTSE1**; Zinc finger protein KIAA1802, **KIAA1802**; DNA ligase 1, **LIG1**; Methyl-CpG-binding protein 2, **MECP2**; Myosin-binding protein C, cardiac-type, **MYBPC3**; Phenylalanine-4-hydroxylase, **PAH**; Parkinson disease protein 7, **PARK7**; Membrane-associated tyrosine- and threonine-specific cdc2-inhibitory kinase, **PKMYT1**; Pinin, **PNN**; Protein phosphatase 1 regulatory subunit 12B, **PPP1R12B**; Ribosomal protein S6 kinase alpha-3, **RPS6KA3**; SH3 and PX domain-containing protein 2A, **SH3PXD2A**; WD repeat-containing protein 91, **WDR91**

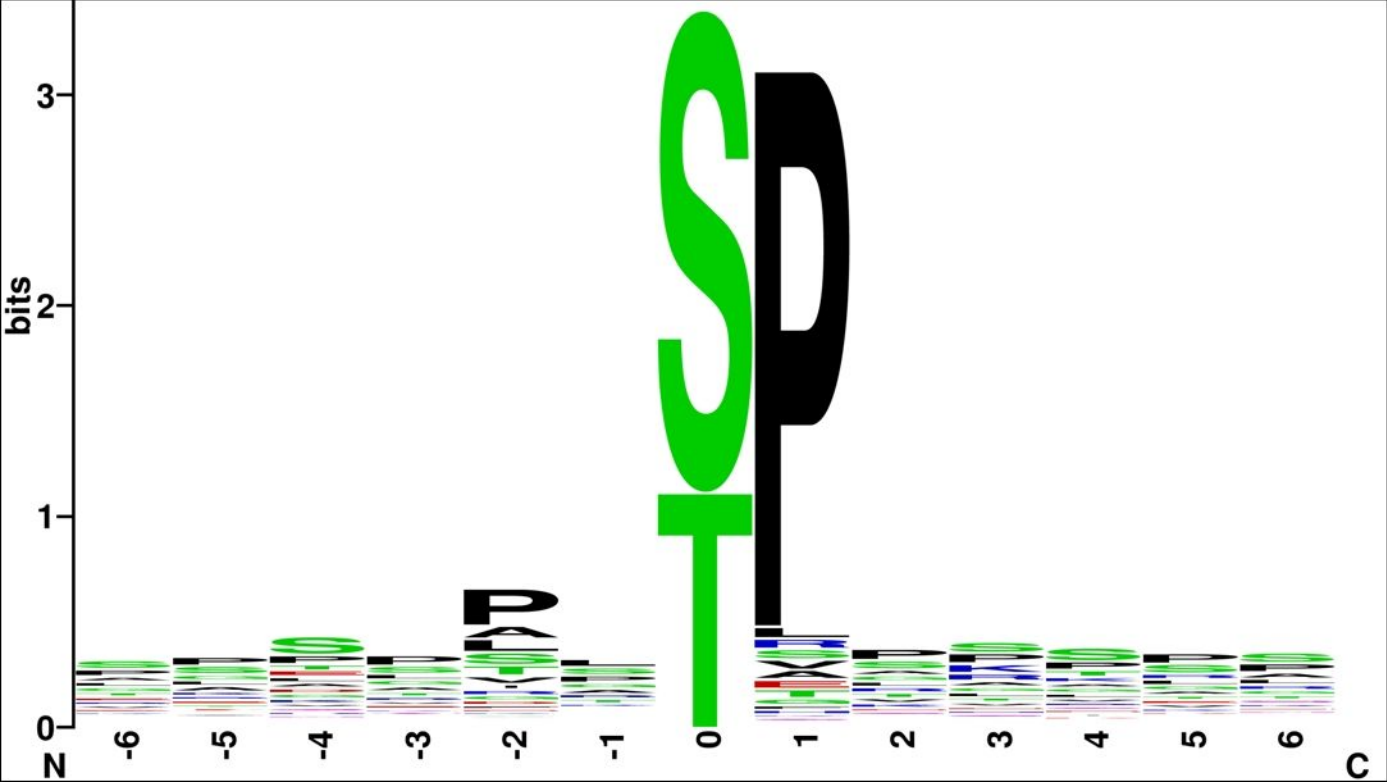
Figure legends

Figure 1. Schematic illustration of phosphovariants according to their types.

Figure 2. Sequence logos of amino acid sequences near phosphorylation sites recognized by the CMGC kinase group.

The horizontal axis represents sequential positions relative to the phosphorylation site. The vertical axis represents decreases in uncertainty. Each letter refers to an amino acid. As the frequency of an amino acid at a given position increases, its height increases.


Type I
P
Q V A S P E A
(-)
(+)
Q V A R P E A
Type II
Kinase
P
Q V A S P E A
(-)
(+)
Kinase
Q V A S R E A
Type III
Kinase A
P
Q V A S P E A
Kinase B
Kinase A
P
Q V A S P A A


bits
3
2
1
0
N
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
C