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

Gil-Mi Ryu1,2, Pamela Song3, Kyu-Won Kim2, Kyung-Soo Oh1, Keun-Joon Park1 and Jong Hun Kim3,*

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

Received August 22, 2008; Revised December 1, 2008; Accepted December 2, 2008

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. Therefore, a comparison of phosphovariants between individuals or between species can give clues about phenotypic differences. 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.

INTRODUCTION

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 life span 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 (Figure 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.

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,

*To whom correspondence should be addressed. Tel: +82 2 3410 1426; Fax: +82 2 3410 1469; Email: jh7521@naver.com

© 2009 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
we tested two proven real data sets, which were the predictions at the kinase group level or at the family level. Therefore, to evaluate the performances of each model (Supplementary Table S3). The specificity (95%, 97%, 98% or 99%) according to the specificity of the variations. One variation can include more than one type of kinase motifs, and 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 Figure S2). 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 to sequences used for training PredPhospho (Table 4). 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 (8). 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 S3). The specificity options are described in the Supplementary Data.

**Sequence logos**

We obtained 562 phosphorylation site sequences recognized by the CMGC kinase group from Swiss-Prot and HPRD. We trimmed the sequences as six symmetric
results centered phosphorylation 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://phospho

variants.ngri.go.kr/seq_input_predphospho2.htm and http://phospho

variant.ngri.go.kr, respectively.

Bug fixing and minor changes of database will be done, whenever needed. The new version of Swiss-Prot and HPRD will be incorporated into our database, annually at least.

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 50 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 (HPRB, Table 1) (9,10). 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. (11) The cell line that they used had a Gly766Ser mutation in the probable ATP-dependent RNA helicase DDX27 (Swiss-Prot ID, Q96Q07), 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 (12). 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 (13). 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, nonreceptor 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 (14,15). The Pro387Leu substitution reduces 75% of the phosphorylation by CDC2 in vitro (16). 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 (17). 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 PTPN1a type III phosphovariant because it inhibits the recognition of Ser386 by CDC2 kinase but has little effect on its phosphorylation by CK2 (16,18).

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 3 (Panel b) are either type I or type III phosphovariants.
Table 1. Examples of type I phosphovariants

| Panel a: Type I(−) phosphovariants |
|------------------------------------|
| **Phosphovariants causing Mendelian inherited disease** |
| **Gene name** (Swiss-Prot ID) | **Variation site** (Swiss-Prot variant ID) | **Phosphorylation site** | **Local peptide sequence** | **Effect** | **Reference(s) for variation** | **Reference(s) for phosphorylation site** |
| EDNRB (P24530) | S305N (VAR_003472) | S305 | CEMLRKK S GMQIALN QSRD TLC S CLSPGLI | Hirschsprung disease type 2 | 8852659 | 14636059 |
| FANCA (O15360) | S858R (VAR_017498) | S858 | | Fanconi anemia | 11091222 | 17924679 |
| KCNJ1 (P48048) | S219R (VAR_019726) | S219 | RVANLRK S LLIGSHI AFGSSQP S LNDGIKP | Bartter syndrome type 2 | 8556302 | 17081983 |
| L1CAM (P32004) | S1194L (VAR_003947) | S1194 | AFGSSQP S LNDGIKP | Hydrocephalus due to stenosis of the aqueduct of Sylvius and mental retardation, aphasia, shuffling gait and adducted thumbs syndrome | 8556302 | 17081983 |
| MAPT (P10636) | S622N (VAR_010350) | S622 | KHVPGGG S VQIVYK | Frontotemporal dementia and parkinsonism linked chromosome 17 | 10208578 | 7706316 |
| MAPT (P10636) | S637F (VAR_019665) | S637 | VDLSKVT S KCGLGN | Pick disease | 11891833 | 11104762 |
| MAPT (P10636) | S669L (VAR_019667) | S669 | DFKDRVQ S KIGSLDN | Fatal respiratory hyperventilation | 14595660 | 11104762 |
| MITF (O75030) | S405P (VAR_010302) | S405 | QARAHGL S LIPSTGL | Waardenburg syndrome type IIa | 8589691 | 10587587 |
| NFKBIA (P25963) | S32I (VAR_034871) | S32 | LLDDRHD S GLDSMKD | Autosomal dominant anhidrotic ectodermal dysplasia with immunodeficiency | 14523047 | 9721103 |
| PER2 (O15055) | S662G (VAR_029080) | S662 | ALPGKAES V ASLTSQ | Familial advanced sleep-phase syndrome | 11232563 | 11232563 |
| PTPN11 (Q06124) | Y62D (VAR_015605) | Y62 | KIQNTGD Y YDLYGGE | Patients with Noonan syndrome 1 manifesting juvenile myelomonocytic leukemia | 11992261 | 11232563 |
| RAF1 (P04049) | S259F (VAR_037809) | S259 | SQRQRST S TPVHNMV | Noonan syndrome type 5 | 17603483 | 8349614 |
| RAF1 (P04049) | T491R (VAR_037819) | T491 | IGDFGLA T VKSRWSG | Noonan syndrome type 5 | 17603483 | 11447113 |
| RAF1 (P04049) | T491I (VAR_037818) | T491 | IGDFGLA T VKSRWSG | Noonan syndrome type 5 | 17603483 | 11447113 |
| RPS6KA3 (P51812) | S227A (VAR_006195) | S227 | DHEKKAY S FCITVEY | Coffin–Lowry syndrome | 12906218 | 15037656 |
| STAT3 (P40763) | Y657C (VAR_037381) | Y657 | FAEIIMG Y KIMDATN | Hyperimmunoglobulin E recurrent infection syndrome autosomal dominant | 17881745 | 15037656 |
| TGFB2 (P37173) | Y336N (VAR_022352) | Y336 | AKGNLQE Y LTRHVIS | Loeys–Dietz aortic aneurysm syndrome | 15731757 | 9169454 |
| TNN15 (P19429) | S166F (VAR_029454) | S166 | LGARAKE S LDLRAHL | Hypertrophic cardiomyopathy | 12974739 | 11211119 |
| TSC1 (Q92574) | S167F (VAR_009403) | S167 | LGARAKE S LDLRAHL | Tuberous sclerosis complex, could be a polymorphism | 10507911 | 14551205 |

**Phosphovariants found in cancer**

| **Gene name** (Swiss-Prot ID) | **Variation site** (Swiss-Prot variant ID) | **Phosphorylation site** | **Local peptide sequence** | **Effect** | **Reference(s) for variation** | **Reference(s) for phosphorylation site** |
| CDH1 (P12830) | S838G (VAR_001322) | S838 | LVFDYEG S GSEAASL | Ovarian cancer | 8075649 | 10671552 |
| CTNNB1 (P35222) | S23R (VAR_017612) | S23 | PDRKAAS S HWQQQSY | Hepatocellular carcinoma, no effect | 10435629 | 12027456 |
| CTNNB1 (P35222) | S33F (VAR_017617) | S33 | QQSYLD S GIHSGAT | Pilomatricoma, medulloblastoma and hepatocellular carcinoma | 10435629 | 12027456 |
| CTNNB1 (P35222) | S33L (VAR_017618) | S33 | QQSYLD S GIHSGAT | Hepatocellular carcinoma | 10435629 | 12027456 |
| CTNNB1 (P35222) | S37A (VAR_017624) | S37 | YLDGSIH S GATTAP | Medulloblastoma, hepatocellular carcinoma | 10435629 | 12027456 |
| CTNNB1 (P35222) | S37C (VAR_017625) | S37 | YLDGSIH S GATTAP | Pilomatricoma, hepatoblastoma | 9927029 | 11815847 |
| CTNNB1 (P35222) | S37F (VAR_017626) | S37 | YLDGSIH S GATTAP | Pilomatricoma | 10192393 | 11815847 |

(continued)
| Gene name (Swiss-Prot ID) | Variation site (Swiss-Prot variant ID) | Phosphorylation site | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|--------------------------|--------------------------------------|----------------------|-----------------------|--------|-----------------------------|-------------------------------------|
| CTNNB1 (P35222)          | T41A (VAR_017629)                    | T41                  | GHSGT T TAPLSG        | Hepatoblastoma and hepatocellular carcinoma, also in a desmoid tumor | 12051714, 10398436, 9927029, 12027456, 10655994, 10435629, 10192393, 10435629 | 12051714, 12114015, 11818547, 2000790 |
| CTNNB1 (P35222)          | T41I (VAR_017630)                   | T41                  | GHSGT T TAPLSG        | Pilomatrixoma and hepatocellular carcinoma                       | 12051714, 12114015, 11818547, 2000790 | 12051714, 12114015, 11818547, 2000790 |
| CTNNB1 (P35222)          | S45F (VAR_017631)                   | S45                  | GATTTAP S LSGKGNP     | Hepatocellular carcinoma                                         | 10435629 | 12051714, 12114015, 11818547, 2000790 |
| CTNNB1 (P35222)          | S45P (VAR_017632)                   | S45                  | GATTTAP S LSGKGNP     | Hepatocellular carcinoma                                         | 10435629 | 12051714, 12114015, 11818547, 2000790 |
| FAM10A4 (Q8IZP2)         | S71L (VAR_023644)                   | S71                  | DLKADEP S SEESDL       | B-cell leukemia, multiple myeloma, and prostate cancer           | 12079276, 17081983 | 12475979 |
| MET (P08581)             | Y1230C (VAR_006292)                 | Y1230                | FGLARM Y DKEYYSV      | Hereditary papillary renal carcinoma                             | 9140397 | 12051714, 12114015, 11818547, 2000790 |
| MET (P08581)             | Y1230H (VAR_006293)                 | Y1230                | FGLARM Y DKEYYSV      | Hereditary papillary renal carcinoma                             | 9140397 | 12051714, 12114015, 11818547, 2000790 |
| NME1 (P15531)            | S120G (VAR_004625)                  | S120                 | GRNIIHG S DSVESAE     | Neuroblastoma                                                    | 8047138 | 10671068, 2594029 |
| RB1 (P06400)             | S567L (VAR_005579)                  | S567                 | SLAWLS D PLFDLKL      | Retinoblastoma                                                   | 8810265 |
| TP53 (P04637)            | T155A (VAR_005901)                  | T155                 | DSTPPPG T RVRAIMAG    | Esophageal cancer                                                | 1868473 | 1262892 |

**Phosphovariants related with polymorphism**

| Protein name | Variation site | Amino acid change | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|--------------|----------------|-------------------|------------------------|--------|-----------------------------|--------------------------------------|
| BARD1 (Q97928) | S186G (VAR_038371) | S186 | SYEFSVP S PPAVDSV | Polymorphism (rs16852741) | 15855157 | 16696243 |
| C10orf11 (Q9H2I8) | S153F (VAR_033686) | S153 | SSEDVOS S PERHTP | Polymorphism (rs35349706) | 16696243 | 15952455 |
| CTNNB1 (O60716) | Y217C (VAR_020929) | Y217 | PDGYSRH Y EDGYPPG | Polymorphism (rs11570194) | 16696243 | 16212419 |
| CTPS (P17812) | S571L (VAR_027055) | S571 | RDTDYSR S GSSPDS | Polymorphism (rs17856308) | 15489334 | 16907034 |
| HIFIA (Q16665) | T796A (VAR_015854) | T796 | ESGLQPL T SYDCVEYN | Polymorphism (rs1802821) | 17081983 | 17382225 |
| INSR (P06213) | Y1361C (VAR_015933) | Y1361 | SYEEHIP Y THMNGGGK | Polymorphism (rs13306449) | 7657032 | 11401470 |
| KRT36 (O76013) | T315M (VAR_020306) | T315 | EIIELR T VNALEIE | Polymorphism (rs2301354) | 17081983 | 17081983 |
| MYH15 (Q9Y2K3) | T1125A (VAR_030238) | T1125 | KTVKELQ T QIKDLKE | Polymorphism (rs3900940) | 17081983 | 17081983 |
| PDLIM5 (Q9HC4) | S136F (VAR_023368) | S136 | PRPFGSV S SPKVTISI | Polymorphism (rs2452600) | 17081983 | 17287340 |
| PNN (Q9H307) | S671G (VAR_020306) | S671 | HKSSKGG S SRDTKGS | Polymorphism (rs13021) | 17081983 | 17287340 |
| RPS6KA3 (P04657) | S366A (VAR_023368) | S366 | PGGSRAH S SHLKSKK | Polymorphism (rs17136053) | 9183006 | 16889390 |

**Panel b: Type I(+) phosphovariants**

| Protein name | Variation site | Amino acid change | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|--------------|----------------|-------------------|------------------------|--------|-----------------------------|--------------------------------------|
| DDX27 (Q96QQ7) | G766S | G766 | ALKYQRA G PSFEERK | Unknown | 16565220 | 16565220 |

aLocations and amino acid changes of the variations in the proteins.
bPeptide sequences with 15-mer amino acids. The amino acids in the eighth position are phosphorylated residues.
cThe 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.
dPubmed ID for the references of the variations

Nucleic Acids Research, 2009, Vol. 37, No. 4 1301
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 S2). 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 Data for the modification of the specificity for each model), the specificities increased to 72.09% and

Table 3. Possible examples of type III phosphovariants

| Gene name (Swiss-Prot ID) | Variation site (Swiss-Prot variant ID) | Removed phosphorylation site (related kinases) | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|---------------------------|----------------------------------------|-----------------------------------------------|------------------------|--------|-----------------------------|-------------------------------------|
| DUT (P33316) | P100S (VAR_022314) | S99b (CDC2) | GPETPAI S PSKRARP | Polymorphism | 17081983 | 8631817 |
| GJA1 (P17302) | P283L (VAR_014101) | S282b (ERK1, ERK2 and MAPK7) | TAPLSPM S PPGYKLV | Polymorphism | 8631994 | 9535905 |
| PPARG (P37231) | P113Q (VAR_010724) | S112c (ERK2, JNK1 and MAPK8) | AIKVEPA S PPYYSEK | Obesity and polymorphism | 9753710 | 9030579 |
| RXRA (P19793) | P261L (VAR_014620) | S260b (ERK2 and MAPK7) | NMGLNPS S PNDPVNT | Polymorphism | 12048211 |

aThe variation sites are underlined and are marked with the bold style.
bThe 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.
cThe removal of the phosphorylation site by the variation has been confirmed by a experiment (12).

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.

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 2. Examples of type II(−) phosphovariants

| Gene name (Swiss-Prot ID) | Variation site (Swiss-Prot variant ID) | Removed phosphorylation site (related kinases) | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|---------------------------|----------------------------------------|-----------------------------------------------|------------------------|--------|-----------------------------|-------------------------------------|
| DUT (P33316) | P100S (VAR_022314) | S99b (CDC2) | GPETPAI S PSKRARP | Polymorphism | 17081983 | 8631817 |
| GJA1 (P17302) | P283L (VAR_014101) | S282b (ERK1, ERK2 and MAPK7) | TAPLSPM S PPGYKLV | Polymorphism | 8631994 | 9535905 |
| PPARG (P37231) | P113Q (VAR_010724) | S112c (ERK2, JNK1 and MAPK8) | AIKVEPA S PPYYSEK | Obesity and polymorphism | 9753710 | 9030579 |
| RXRA (P19793) | P261L (VAR_014620) | S260b (ERK2 and MAPK7) | NMGLNPS S PNDPVNT | Polymorphism | 12048211 |

aThe variation sites are underlined and are marked with the bold style.
bThe 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.
cThe removal of the phosphorylation site by the variation has been confirmed by a experiment (12).

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.

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

| Gene name (Swiss-Prot ID) | Variation site (Swiss-Prot variant ID) | Removed phosphorylation site (related kinases) | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|---------------------------|----------------------------------------|-----------------------------------------------|------------------------|--------|-----------------------------|-------------------------------------|
| PTPN1 (P18031) | P387L (VAR_022014) | S386 (CDC2 and CK2) | LRGQAQQ A S PAKGEP | Low glucose tolerance and polymorphism | 1591985 | 9600099 | 8491187 |
| BRCA1 (P38398) | S1217Y (VAR_020695) | S1217 | ESSENL S SEDEELP | Breast cancer and breast-ovarian cancer | 14722926 | 17081983 |
| CASP8 (Q14790) | S219T (VAR_025816) | S219 | PREQDSE S QTLDKVY | Polymorphism | 1396589 | 12912980 | 12972555 | 15144186 |
| CDK2 (P24941) | Y15S (VAR_016157) | Y15 | EKIGEPT Y GVVYKAR | Obesity and polymorphism | 1396589 | 12912980 | 12972555 | 15144186 |
| CTNNB1 (P35222) | S33Y (VAR_017619) | S33 | QQQSYLD S GIHGSGAT | Pilomatrixoma | 12027456 | 12000790 | 12114015 | 11818547 |
| CTNNB1 (P35222) | S37Y (VAR_017627) | S37 | YLDSGIH S GATTTAP | Hepatocellular carcinoma | 10435629 | 12000790 | 12114015 | 11818547 |
| TEK (Q02763) | Y897S (VAR_008716) | Y897 | GACEHRG Y LLYAEY | Dominantly inherited venous malformations | 10369874 | 11080633 |
| XRCC1 (P18887) | S485Q (VAR_014779) | S485 | QDNGAED S GDTEDEL | Polymorphism | 15066279 | 12027456 | 12000790 | 12114015 | 11818547 |

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 nonreceptor type 1, PTPN1; angiopoietin-1 receptor (precursor P, TEK; DNA repair protein XRCC1, XRCC1.
72.39%, respectively, for each data set, whereas the sensitivities decreased to 73.24% and 65.76%, respectively (Table 5). 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 (19). 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 (Table 6). 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 (Tables 5 and 6). 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 7. In the Supplementary Data, 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 5, 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

---

**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.

**Table 4.** The general performance test: the number of each data set

| Data I       | (+) | (−) | (+) | (−) |
|--------------|-----|-----|-----|-----|
| Ser/Thr      | 1860| 926 | 3017| 315 |
| Tyr          | 38  | 95  | 359 | 11  |

*aDifferent data sets compiled with mass spectrometer. See the text for the detail explanation for data set I and II.

*bThe 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.

**Table 5.** The general performance test of the PredPhospho for two real data sets

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

*aOptions 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).

Abbreviations: sensitivity, Sn; specificity, Sp.
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 5. Instead, the performance of the each kinase-specific prediction can be judged from the performances shown in Supplementary Table S2.

The proportion of phosphovariants predicted by PredPhospho is shown in Table 8. 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 Data, only phosphorylation sites with available kinase information were used as training models for PredPhospho. Hence, the phosphorylation sites in Table 8 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 ubiquitination. 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 (20). 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 (21). 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 (22,23).

Phosphorylation at Ser32 of NF-kB inhibitor alpha (Swiss-Prot ID, P25963) causes ubiquitination and results in the activation of NF-kB (24). The Ser32Ile substitution of NF-kB inhibitor alpha (Swiss-Prot ID, P25963) eliminates the phosphorylation site at Ser32. Consequently, the ubiquitination of NF-kB inhibitor alpha is inhibited by the Ser32Ile variant of NF-kB inhibitor alpha and NF-kB cannot be activated. The Ser32Ile variant of NF-kB inhibitor alpha causes autosomal dominant anhydrotic ectodermal dysplasia with immunodeficiency (25). 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 (12). The Pro387Leu substitution (dbSNP ID, rs16995309) of tyrosine-protein phosphatase nonreceptor, type 1 (Swiss-Prot ID, P18031) is associated with type II diabetes mellitus (16,18). 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 phospho moiety to an amino acid with a hydroxyl group. However, type II and III phosphovariants can significantly affect kinase kinetics without completely changing in the protein’s affinity for DNA, inducing hyperphosphorylation and the inhibition of ubiquitination.
### Table 8. Predicted phosphovariants whose phosphorylation sites were confirmed in human or orthologous proteins

| Gene name (Swiss-Prot ID) | Site (Swiss-Prot variant ID) | Related phosphorylation site (predicted kinase recognizing it) | Local peptide sequence | Effect | Reference(s) for variation | Reference(s) for phosphorylation site |
|---------------------------|-----------------------------|---------------------------------------------------------------|------------------------|--------|---------------------------|---------------------------------------|
| **Panel a: Type I (−) phosphovariants** | | | | | | |
| ACIN1 (Q9UVK3) | S478F (VAR_020233) | S478 (CDK, GSK, and MAPK) | VQLVGGL S PLSSPSD | Polymorphism (rs3751501) | 17242355 (mouse) | |
| MECP2 (P51608) | S229L (VAR_018200) | S229 (MAPK) | VKMPFQT S PGGAKEG | Polymorphism | 1076373 (mouse) |
| PAH (P00439) | S16P (VAR_000869) | S16 (RSK and CAMKL) | PGLGRKL S DFGQETS | Phenylethonuria | 1679029 (mouse) |
| **Panel b: Type II (−) phosphovariants** | | | | | | |
| GTSE1 (Q9NYZ3) | R506W (VAR_024154) | R506 (PKC) | GTSE1 R506W S504 (PKC) | Polymorphism (rs140054) | 10591208 (mouse) |
| LIG1 (P18858) | P52L (VAR_020194) | S52L (PKC) | GTSE1 R506W S504 (PKC) | Polymorphism (rs4987181) | 16964243 (mouse) |
| **Panel c: Type III phosphovariants** | | | | | | |
| ABCB11 (Q95342) | R698H (VAR_035352) | S698 (PKC) | ABCB11 R698H S698 (PKC) | Polymorphism (rs16703017) | 17242355 (mouse) |
| ABL1 (P00519) | P810L (VAR_032678) | S809 (GSK and MAPK) | ABL1 P810L S809 (GSK and MAPK) | Polymorphism (rs17344846) | 17081983 (mouse) |
| AQP2 (P41181) | P262L (VAR_015255) | S261 (IKK and MAPK) | ABCB11 R698H S698 (PKC) | Polymorphism (rs17344846) | 17081983 (mouse) |
| CASP9 (Q14790) | S219T (VAR_025816) | S219 (IKK and PKC) | CASP9 S219T S219 (IKK and PKC) | Polymorphism (rs3751501) | 17242355 (mouse) |
| CHGB (P05060) | R178Q (VAR_020287) | R178 (AKT and GRK) | CHGB R178Q S183 (AKT and GRK) | Polymorphism (rs3751501) | 17081983 (mouse) |
| EIF4G3 (O43432) | P496A (VAR_034009) | S495 (MAPK and RSK) | EIF4G3 P496A S495 (MAPK and RSK) | Polymorphism (rs93176330) | 16807684 (mouse) |
| MYBPC3 (Q14896) | G278E (VAR_019891) | G278 (CAMK2, PKA, PKC, RSK, and STE20) | MYBPC3 G278E LS4F GRG | Polymorphism (rs140054) | 16964243 (mouse) |
| PKMYT1 (Q99640) | R140C (VAR_019928) | R140 (CAMK2, PKA, PKC, and RSK) | PKMYT1 R140C S140 (CAMK2, PKA, PKC, and RSK) | Polymorphism (rs17344846) | 17192257 (mouse) |
| PPP1R12B (Q51608) | S1038 (VAR_023644) | S1038 (CAMK2, PKA, PKC, and RSK) | PPP1R12B S1038 S1038 (CAMK2, PKA, PKC, and RSK) | Polymorphism (rs17344846) | 17192257 (mouse) |
| PARK7 (Q99497) | E64D (VAR_020493) | E64 (Src) | PARK7 E64D RVGFLS (MAPK) | Polymorphism (rs17344846) | 17192257 (mouse) |
| PNN (Q99307) | S671G (VAR_023368) | S671 (AKT, GSK, and CAMKL) | PNN S671G S671 (AKT, GSK, and CAMKL) | Polymorphism (rs17344846) | 17192257 (mouse) |
| SH3PXD2A (Q51608) | S1038 (VAR_023644) | S1038 (CAMK2, PKA, PKC, and RSK) | SH3PXD2A S1038 S1038 (CAMK2, PKA, PKC, and RSK) | Polymorphism (rs17344846) | 17192257 (mouse) |
| WDR91 (A4D1P6) | L257P (VAR_033358) | L257 (IKK and PKC) | WDR91 L257P L257 (IKK and PKC) | Polymorphism (rs17344846) | 17192257 (mouse) |
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) (16).

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. First, we regarded nonannotated serine, threonine and tyrosine residues as nonphosphorylated sites. Although we tried not to select false nonphosphorylated sites by using real MS data sets, some of the nonphosphorylated sites might be determined to be phosphorylated sites, in the long run. Second, 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 5). 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 (26). 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.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

FUNDING
This work was supported by Health Fellowship Foundation. Funding for open access charge: intramural grants from Korea National Institute of Health, the Korea Center for Disease Control of the Republic of Korea (Project No.: 091-4845-301-210).

Conflict of interest statement. None declared.

REFERENCES
1. Hunter,T. (2000) Signaling 2000 and beyond. Cell, 100, 113–127.
2. Dash,P.K., Moore,A.N., Kobori,N. and Runyan,J.D. (2007) Molecular activity underlying working memory. Learn. Mem., 14, 554–563.
3. Hunter,T. and Karin,M. (1992) The regulation of transcription by phosphorylation. Cell, 70, 375–387.
4. Beausoleil,S.A., Villen,J., Gerber,S.A., Rush,J. and Gygi,S.P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol., 24, 1285–1292.
5. Olsen,J.V., Blagoev,B., Gnad,F., Macek,B., Kumar,C., Mortensen,P. and Mann,M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell, 127, 635–648.
6. Kecman,V. (2001) Learning and Soft Computing, The MIT Press, Cambridge.
7. Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. Science, 298, 1912–1934.
8. Yip,Y.L., Schib,I., Diemand,A.V., Gattiker,A., Fambigietti,L.M., Gasteiger,E. and Bairoch,A. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. Hum. Mutat., 23, 464–470.
9. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O’Donovan,C., Phan.I. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement Trembl in 2003. Nucleic Acids Res., 31, 365–370.
10. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, 13, 2363–2371.

11. Nousiainen, M., Sillje, H.H., Sauer, G., Nigg, E.A. and Korner, R. (2006) Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. USA*, 103, 5391–5396.

12. Ristow, M., Muller-Wieland, D., Pfeiffer, A., Krone, W. and Kahn, C.R. (1998) Obesity associated with a mutation in a genetic regulator of adipocyte differentiation. *N. Engl. J. Med.*, 339, 953–959.

13. Pinna, L.A. and Ruzzene, M. (1996) How do protein kinases recognize their substrates? *Biochim. Biophys. Acta*, 1314, 191–225.

14. Jung, E.J., Kang, Y.S. and Kim, C.W. (1998) 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, 238–242.

15. Flint, A.J., Gebbink, M.F., Franza, B.R., Hill, D.E. and Tonks, N.K. (1993) Multi-site phosphorylation of the protein tyrosine phosphatase, PTP1B: identification of cell cycle regulated and phorbol ester stimulated sites of phosphorylation. *EMBO J.*, 12, 1937–1946.

16. Echwald, S.M., Bach, H., Vestergaard, H., Richelsen, B., Kristensen, K., Drivsholm, T., Borch-Johnsen, K., Hansen, T. and Pedersen, O. (2002) 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.

17. Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S.G. and Pandey, A. (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, 25, 285–286.

18. Ukkola, O., Rankinen, T., Lakka, T., Leon, A.S., Skinner, J.S., Wilmore, J.H., Rao, D.C., Kesaniemi, Y.A. and Bouchard, C. (2005) Protein tyrosine phosphatase 1B variant associated with fat distribution and insulin metabolism. *Obes. Res.*, 13, 829–834.

19. Obenauer, J.C., Cantley, L.C. and Yaffe, M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, 31, 3635–3641.

20. Tassabehji, M., Newton, V.E., Liu, X.Z., Brady, A., Donnai, D., Krajewska-Walasek, M., Murday, V., Norman, A., Oberszyn, E., Reardon, W. et al. (1995) The mutational spectrum in Waardenburg syndrome. *Hum. Mol. Genet.*, 4, 2131–2137.

21. Taniguchi, T., Kawamata, T., Mukai, H., Hasegawa, H., Isagawa, T., Yasuda, M., Hashimoto, T., Terashima, A., Nakai, M., Mori, H. et al. (2001) Phosphorylation of tau is regulated by PKN. *J. Biol. Chem.*, 276, 10025–10031.

22. Nicholl, D.J., Greenstone, M.A., Clarke, C.E., Rizzu, P., Crooks, D., Crowe, A., Trojanowski, J.Q., Lee, V.M. and Heutink, P. (2003) An English kindred with a novel recessive tauopathy and respiratory failure. *Ann. Neurol.*, 54, 682–686.

23. Rosso, S.M., van Herpen, E., Deelen, W., Kamphorst, W., Severijnen, L.A., Willemens, R., Ravid, R., Niermeijer, M.F., Dooijes, D., Smith, M.J. et al. (2002) A novel tau mutation, S320F, causes a tauopathy with inclusions similar to those in Pick’s disease. *Ann. Neurol.*, 51, 373–376.

24. DiDonato, J., Mercurio, F., Rosette, C., Wu, L.J., Suyang, H., Ghosh, S. and Karin, M. (1996) Mapping of the inducible IkappaB phosphorylation sites that signal its ubiquitination and degradation. *Mol. Cell Biol.*, 16, 1295–1304.

25. Courtois, G., Smahi, A., Reichenbach, J., Doffinger, R., Cancrini, C., Bonnet, M., Puel, A., Chable-Bessia, C., Yamaoka, S. and Feinberg, J. (2003) A hypermorphic IkappaBalpha mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and T cell immunodeficiency. *J. Clin. Invest.*, 112, 1108–1115.

26. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V. and Roth, G.T. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452, 872–876.